Hibridne tehnike kombinatorne optimizacije temeljene na genetskim algoritmima s primjenom na odabir atributa u ocjenjivanju kreditnog rizika građana

Oreški, Stjepan

Sveučilište u Zagrebu

FAKULTET ORGANIZACIJE I INFORMATIKE

STJEPAN OREŠKI

HIBRIDNE TEHNIKE KOMBINATORNE OPTIMIZACIJE TEMELJENE NA GENETSKIM ALGORITMIMA S PRIMJENOM NA ODABIR ATRIBUTA U OCJENJIVANJU KREDITNOG RIZIKA GRAĐANA

DOKTORSKI RAD

Varaždin, 2014.

PODACI O DOKTORSKOM RADU

I. AUTOR

Ime i prezime	Stjepan Oreški
Datum i mjesto rođenja	08.01.1964. Vukovoj
Naziv fakulteta i datum diplomiranja na VII/I stupnju	Fakultet organizacije i informatike, Varaždin, 1985.
Naziv fakulteta i datum diplomiranja na VII/II stupnju	Fakultet organizacije i informatike, Varaždin , 1997.
Sadašnje zaposlenje	Karlovačka banka d.d.

II. DOKTORSKI RAD

Naslov	Hibridne tehnike kombinatorne optimizacije temeljene na genetskim algoritmima s primjenom na odabir atributa u ocjenjivanju kreditnog rizika građana
Broj stranica, slika, tabela, priloga, bibliografskih podataka	159 stranica, 31 slika, 29 tablica, 2 priloga, 111 bibliografskih podataka
Znanstveno područje i polje iz kojeg je postignut doktorat znanosti	Društvene znanosti, informacijske i komunikacijske znanosti
Mentori ili voditelji rada	Prof. dr.sc. Božidar Kliček
Fakultet na kojem je obranjen doktorski rad	Fakultet organizacije i informatike, Varaždin
Oznaka i redni broj rada	*116*

III. OCJENA I OBRANA

Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena tema	17. 09. 2013.
Datum predaje rada	8. 07. 2014.
Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena pozitivna ocjena rada	25. 09. 2014.
Sastav povjerenstva koje je rad ocijenilo	Izv.prof.dr.sc. Diana Šimić Prof.dr.sc. Božidar Kliček dr.sc. Dragan Gamberger
Datum obrane doktorskog rada	14. 10. 2014.
Sastav povjerenstva pred kojim je rad obranjen	Izv.prof.dr.sc. Diana Šimić Prof.dr.sc. Božidar Kliček dr.sc. Dragan Gamberger
Datum promocije	

Sveučilište u Zagrebu

FAKULTET ORGANIZACIJE I INFORMATIKE

STJEPAN OREŠKI

HIBRIDNE TEHNIKE KOMBINATORNE OPTIMIZACIJE TEMELJENE NA GENETSKIM ALGORITMIMA S PRIMJENOM NA ODABIR ATRIBUTA U OCJENJIVANJU KREDITNOG RIZIKA GRAĐANA

DOKTORSKI RAD

Mentor:
Prof. dr. sc. Božidar Kliček

Varaždin, 2014.

University of Zagreb

FACULTY OF ORGANIZATION AND INFORMATICS

STJEPAN OREŠKI

HYBRID TECHNIQUES OF COMBINATORIAL OPTIMIZATION BASED ON GENETIC ALGORITHMS WITH APPLICATION TO FEATURE SELECTION IN RETAIL CREDIT RISK ASSESSMENT

DOCTORAL THESIS

Varaždin, 2014

ZAHVALE

Prvo želim zahvaliti prof. dr. sc. Božidaru Kličeku koji je prihvatio mentorstvo ovog rada i koji je svojim prijedlozima i savjetima pridonio njegovoj kvaliteti. Profesore bila mi je izuzetna čast i zadovoljstvo raditi s Vama.

Posebnu zahvalnost osjećam prema svojoj obitelji, osobito prema svojim roditeljima. Vaša podrška, uvijek i u svemu, bila mi je od neprocjenjivog značaja, omogućila mi je bezbrižno školovanje i sve ono što je došlo iza toga, a vi ste mi uvijek i u svemu ostavljali punu slobodu izbora. Žao mi je što radost završetka nisam u mogućnosti podijeliti s vama.

Međutim, najveće hvala zaslužuju Đurdica, Tina i Goran. Ja sam vrlo sretan suprug i otac jer u svakom trenutku imam najveću moguću podršku svoje obitelji. Vaša podrška mi je tako važna, ona me sve ovo vrijeme nosila u radu i istraživanju. Vi ste mi dali snage i motiva da ustrajem u svemu. Volim vas i veselim se svakom novom jutru i zajedništvu s vama.

SAŽETAK

Hibridne tehnike kombinatorne optimizacije predstavljaju rastuće područje istraživanja, namijenjeno za rješavanje složenih problema kombinatorne optimizacije. U prvom dijelu ove disertacije, usredotočiti smo se na metodološku pozadinu hibridnih tehnika kombinatorne optimizacije, usmjeravajući posebnu pozornost na važne koncepte u području kombinatorne optimizacije i računske teorije složenosti, kao i na strategije hibridizacije koje su važne pri razvoju hibridnih tehnika kombinatorne optimizacije. U skladu s prikazanim odnosima među tehnikama kombinatorne optimizacije, strategijama njihova kombiniranja kao i konceptima za rješavanje problema kombinatorne optimizacije, ova disertacija kreira nove hibridne tehnike za odabir atributa i klasifikaciju pri procjeni kreditnog rizika. Disertacija naglašava važnost hibridizacije kao koncepta suradnje među metaheuristikama i drugim tehnikama za optimizaciju. Važnost takve suradnje potvrđuju rezultati koji su predstavljeni u eksperimentalnom dijelu rada, koji su dobiveni na hrvatskom i njemačkom kreditnom skupu podataka korištenjem hibridnih tehnika kombinatorne optimizacije kreiranim u ovoj disertaciji.

Znanstveni doprinos disertacije:

•Kreirane hibridne tehnike selekcije atributa (GA-NN i HGA-NN), posebno prilagođene problemskoj domeni - temeljene na genetskim algoritmima i umjetnim neuronskim mrežama.

•Kreiran novi hibridni genetski algoritam uključivanjem rezultata filtarskih tehnika i a priori spoznaja u početnu populaciju genetskog algoritma.

•Kreiran novi operator selekcije kod genetskog algoritma, jedinstvena selekcija (engl. *unique selection*).

•Kreirani sofisticirani kreditni modeli koji omogućuju povećanje učinkovitosti alokacije kapitala.

Ključne riječi: hibridne tehnike, klasifikacija, odabir atributa, kreditni rizici, genetski algoritam, neuronske mreže

PROŠIRENI SAŽETAK

Svrha ovog doktorskog rada je temeljito istražiti ukupni skup podataka kojima raspolaže banka te utvrditi do koje mjere ti podaci mogu biti dobra osnovica za predviđanje sposobnosti tražitelja kredita da vrati kredit na vrijeme. Takvo predviđanje sposobnosti tražitelja kredita treba izvršiti bez traženja dodatnih podataka od klijenta uz pretpostavku da je tražitelj kredita već dulje vrijeme klijent banke te da je banka već u svojoj bazi podataka prikupila dovoljno podataka o klijentu. Pri tome se javlja problem. Problem je što u velikim količinama podataka i informacija koje su banke širom svijeta kumulirale u bazama podataka o svojim klijentima te njihovoj financijskoj i platežnoj povijesti obično ima i mnoštvo nebitnih podataka, odnosno atributa. U tom kontekstu nebitni atributi predstavljaju problem. Nebitni atributi u skupu podataka za učenje neće dovesti do preciznijih rezultata klasifikacijske analize, a povećavaju: (1) troškove prikupljanja podataka, (2) vrijeme potrebno za učenje i konstrukciju modela kao i (3) razumljivost samog modela. Stoga je potrebna priprema podataka za klasifikaciju kako bi se poboljšala kvaliteta konstruiranog modela, smanjila složenost modela te smanjio trošak korištenja modela. U toj pripremi podataka jednu od najvažnijih aktivnosti čini odabir atributa.

Krajnji ciljevi rada bili su dvojaki: (1) razviti vrlo efikasne, u skladu s najnovijim znanstvenim i tehničkim spoznajama, hibridne tehnike za odabir optimalnog podskupa atributa pri ocjenjivanju kreditne sposobnosti tražitelja kredita te (2) prikupiti dodatno znanje i iskustvo o specifičnim prednostima i nedostacima pojedinih tehnika te na smislene načine kombinirati te tehnike i za druge slične probleme.

Teorijski promatrano, odabir optimalnog podskupa atributa spada u razred kombinatorno optimizacijskih problema. Takvi problemi se najčešće rješavaju kombiniranjem egzaktnih i heurističkih algoritama ili više (meta) heurističkih algoritama. Novo nastali algoritmi, u ovom slučaju hibridi, pokušavaju na različite načine kombinirati prednosti dvaju ili više različitih tipova algoritama.

U radu su razmatrani različiti oblici hibridizacije. Od hibridizacije na niskoj razini gdje je rezultat jedna jedinstvena tehnika optimizacije kao funkcionalno nedjeljiva cjelina, do

hibridizacije na visokoj razini kod koje su različiti algoritmi samostalne cjeline, a oblik njihove suradnje je kooperacija. Kombinirane su različite tehnike optimizacije, od egzaktnih do heuristika, uz hipotezu da korist dolazi iz sinergije različitih tehnika. Kod toga je od najveće važnosti uspostaviti dinamičku ravnotežu između diversifikacije i intenziviranja radi brze identifikacije područja u prostoru pretraživanja s visokokvalitetnim rješenjima, ne gubeći previše vremena na područja u prostoru pretraživanja koja su već istražena ili ne daju kvalitetna rješenja.

Pored potencijalnih koristi, hibridne tehnike donose i neke neizbježne nedostatke kao što su: veća kompleksnost tehnike, traže više znanja i napora u dizajnu i implementaciji, rješenje (novi hibrid) može biti orijentirano na rješavanje samo određenog problema. Tu dolazi od izražaja teorem "No free lunch" (Wolpert i Macready, 1996), koji kaže da nema tehnike koja bi bila bolja od svih drugih u svim uvjetima.

U radu je razvijeno i prikazano više hibridnih algoritama. Prvi od njih je kombinacija genetskoga algoritma i umjetnih neuronskih mreža (engl. (i dalje): GA-NN, ***G****enetic* ***A****lgorithm with* ***N****eural* ***N****etworks*). Specifičnost ovog algoritma je da istovremeno izvodi odabir optimalnog podskupa atributa te u skladu s tim atributima i danim skupom podataka podešava parametre umjetnih neuronskih mreža. Drugi algoritam je kombinacija hibridnog genetskoga algoritma i umjetnih neuronskih mreža (engl. (i dalje): HGA-NN, ***H****ybrid* ***G****enetic* ***A****lgorithm with* ***N****eural* ***N****etworks*). Potonji je logički nastavak na prvi, i predstavlja proširenje GA-NN algoritma u pogledu preliminarnog ograničavanja atributa na samo one atribute koje izdvajaju brzi filtarski algoritmi kao i domenski eksperti. Isto tako, učinjena su određena poboljšanja genetskog algoritma kroz: (1) kreaciju inicijalne populacije i (2) uvođenje inkrementalne faze.

U trećem eksperimentu dan je poseban naglasak na probleme u klasifikaciji vezane uz neuravnotežene skupove podataka. Prezentiran je pregled glavnih značajki paradigmi koje se tradicionalno primjenjuju kod klasifikacije klasno neuravnoteženih podataka. Istražene su tehnike za ublažavanje problema vezanih uz troškovno osjetljivu klasifikaciju klasno neuravnoteženih podataka u kombinaciji s tehnikama temeljenim na genetskim algoritmima, GA-NN i HGA-NN. Kod toga su performanse mjerene različitim mjerama, s naglaskom na relativni trošak pogrešne klasifikacije. Istraživanje je provedeno na hrvatskom i njemačkom skupu podataka. Rezultati istraživanja su pokazali da su navedenim proširenjem, s troškovnog aspekta, rezultati HGA-NN ROS tehnike bolji u odnosu na rezultate prezentirane u literaturi.

Rezultati prezentiranih algoritama jasno ukazuju na njihov potencijal u rješavanju problema odabira atributa i ocjenjivanja kreditnog rizika građana te time opravdavaju veći napor u dizajnu i implementaciji. Potencijal prezentiranih algoritama u ocjenjivanju kreditnog rizika građana može se iskoristiti za poboljšanje načina na koji banke upravljaju kreditnim rizikom građana, što predstavlja promociju stabilnog i zdravog bankarstva. Potreba za boljim upravljanjem kreditnim rizicima i sofisticiranim kreditnim modelima je potaknula istraživanja prezentirana u ovom radu.

Ključne riječi: hibridne tehnike, klasifikacija, odabir atributa, kreditni rizici, genetski algoritam, neuronske mreže

Extended abstract

The purpose of this dissertation is to thoroughly investigate the overall data set available to the bank and to determine the extent to which these data can be a good basis for predicting the creditworthiness of the loan applicant. Such a prediction of the applicant's ability should be done without seeking additional information from the client, assuming that the loan applicant is a long-time customer of the bank and that the bank has collected sufficient data on the client in its database. Banks worldwide have accumulated large amounts of data and information about their clients, their financial solvency and payment history. The issue is usually in the multitude of irrelevant data or attributes contained in the accumulated data. In this context, irrelevant attributes are a problem. Irrelevant attributes in the training data set will not lead to more accurate results of classification analysis, but will: (1) increase the cost of data collection, (2) increase the time required for learning and constructing models as well as (3) decrease the user-friendliness of the model itself. Hence, there is the need for classification data preprocessing in order to: improve the quality of the constructed model, reduce the complexity of the model and to reduce the cost of usage. In the data preprocessing, one of the most important activities is the feature selection.

The ultimate objectives of the study were twofold: (1) to develop a highly efficient hybrid technique, in line with the latest scientific and technical knowledge, to select the optimal subset of features when assessing the credit worthiness of the loan applicant, and (2) to collect additional knowledge and experience about the specific advantages and disadvantages of individual techniques as well as combine these techniques to other similar problems in meaningful ways.

Theoretically speaking, the selection of the optimal features subset belongs to the class of combinatorial optimization problems. Such problems are usually solved by combining: exact and heuristic algorithms or more (meta) heuristic algorithms. The newly generated algorithms, in this case hybrids, are in various ways trying to combine the advantages of two or more different types of algorithms.

The paper discusses different forms of hybridization. From hybridization at a low level, where the result is one unique optimization technique which is a functionally indivisible whole, to the hybridization at a high level at which different algorithms are independent entities and their form of collaboration is cooperation. Various optimization techniques, from exact ones to heuristics, were combined with the hypothesis that the benefit

comes from the synergy of different techniques. It is of paramount importance to establish a dynamic balance between diversification and intensification for the quick identification of areas in the search space with high-quality solutions, without losing too much time in the search space that have already been explored or do not provide quality solutions.

In addition to the potential benefits, hybrid techniques bring some unavoidable disadvantages such as: the increased complexity of technique, the need for more knowledge and effort in the design and implementation of the solution, and the narrow orientation for solving specific problems only. Here the well-known theorem "No free lunch" (Wolpert and Macready, 1996) gains prominence. It says that there is no technique that would be better than all others in all conditions.

More hybrid algorithms are developed and shown in the paper. The first of these is a combination of genetic algorithms and artificial neural networks (GA-NN). The specificity of the mentioned algorithm is that it simultaneously performs the selection of the optimal subset of attributes, and accordingly to the attributes of a given set, adjusts the parameters of artificial neural networks. The second algorithm is a combination of the hybrid genetic algorithm and the artificial neural network (HGA-NN). The latter is a logical continuation of the first, and an extension of the GA-NN algorithm in terms of the preliminary restriction attributes to only those attributes that have been distinguished by fast filtering algorithms or domain experts. Also, some improvements have been made through the genetic algorithm: (1) the creation of the initial population and (2) the introduction of the incremental stage.

In the third experiment, special emphasis is given to the problems related to the classification of imbalanced datasets. An overview of the main paradigm characteristics was presented that is traditionally applied to the classification of imbalanced data. Techniques for mitigating problems related to the cost-sensitive classification of class imbalanced data in combination with techniques based on genetic algorithms, GA-NN and HGA-NN, are explored. Performance is measured by a variety of measures, focusing on the relative cost of misclassification. The study was conducted on Croatian and German data sets. The results showed that the specified extension, from the cost point of view, results in the HGA-NN ROS technique which is better compared to the results presented in the literature.

The results of the presented algorithms clearly indicate the potential in: solving the attributes selection problem and citizens' credit risk evaluation, thereby justifying a larger effort in the design and implementation. The presented algorithm's potential in evaluating citizens' credit risk may be used to improve the way in which banks manage the citizen's credit risk, which is the promotion of a stable and healthy banking. The need for better

management of credit risks and sophisticated credit models motivated the research presented in this paper.

Keywords: hybrid techniques, classification, feature selection, credit risk, genetic algorithm, neural networks

SADRŽAJ

POPIS SLIKA

POPIS TABLICA

POPIS KRATICA

ANN	umjetna neuronska mreža (engl. *Artificial Neural Network*)
AUC	površina ispod ROC krivulje (engl. *Area Under the ROC Curve*)
DA	diskriminacijska analiza
FS-NN	algoritam selekcije atributa (engl. *Forward Selection with Neural Networks*)
GA	genetski algoritam
GA-NN	algoritam za selekciju atributa i klasifikaciju (engl. *Genetic Algorithm with Neural Networks*)
HGA-NN	algoritam za selekciju atributa i klasifikaciju (engl. *Hybrid Genetic Algorithm with Neural Networks*)
IQR	interkvartil (engl. *Interquartile Range*)
IRB	sustav temeljen na internim rejtinzima (engl. *Internal Ratings-Based approach*)
KO	kombinatorna optimizacija
LR	logistička regresija
MDA	multivarijatna diskriminacijska analiza
ML	strojno učenje (engl. *Machine Learning*)
NNGM	algoritam za optimiranje parametara neuronskih mreža (engl. *Neural Network Generic Model*)
NP	problem za koji se rješenje može verificirati polinomijalnim algoritmom
O	Landauov simbol
P	problem rješiv polinomijalnim algoritmom
Q1	donji kvartil
Q3	gornji kvartil
ROC	krivulja (engl. *Receiver Operating Characteristic*)
ROS	tehnika uzorkovanja (engl. *Random Over-Sampling*)
RUS	tehnika uzorkovanja (engl. *Random Under-Sampling*)
SMOTE	tehnika uzorkovanja (engl. *Synthetic Minority Over-sampling TEchnique*)
UCI	University of California, Irvine

POGLAVLJE 1

UVOD

1.1 Uvodna razmatranja

Svakodnevni život stavlja pred nas beskonačni niz problema prepoznavanja uzoraka, od mirisa, slika, glasova do pisama i potpisa, prometnih situacija, mjesta i tako dalje. Većinu tih problema rješavamo na senzorskoj razini ili intuitivno, bez izričitih metoda ili algoritama. Čim smo u mogućnosti utvrditi odgovarajući algoritam, problem postaje trivijalan i možemo ga prenijeti na računalo. Računala danas doista pouzdano zamjenjuju ljude u mnogim nekad teškim ili nemogućim, sada samo zamornim zadatcima prepoznavanja uzoraka, kao što su: razvrstavanje pošte, očitanje rezultata medicinskih testova, prepoznavanje vojnih ciljeva, ovjera potpisa, meteorološka predviđanja, utvrđivanje DNK podudarnosti, prepoznavanje otiska prsta i tome slično. Većina navedenih zadataka rješava se sofisticiranim algoritmima prepoznavanja uzoraka temeljenim na umjetnoj inteligenciji. Stoga možemo reći da je područje prepoznavanja uzoraka i strojnog učenja sazrelo do točke gdje mnogi sofisticirani pristupi učenju mogu biti primijenjeni na rješavanje praktičnih problema (Japkowicz i Shah, 2011).

U fokusu ove disertacije je istraživanje i razvoj hibridnih tehnika kombinatorne optimizacije, temeljenih na umjetnoj inteligenciji, na osnovu kojih bismo mogli konstruirati odgovarajući model za procjenu kreditne sposobnosti tražitelja kredita, a da pri tome ne angažiramo kreditne stručnjake niti da klijent mora fizički dolaziti u banku radi procjene njegove kreditne sposobnosti.

U razvoju tehnika je poštivano načelo proporcionalnosti, uzimajući pri tome u obzir razmjer veličine kreditne izloženosti i obujma poslova koje kreditna institucija treba obaviti pri ocjeni rizika građana. Poštivanje načela proporcionalnosti znači da postupci procjene rizika moraju biti razmjerni vrsti, opsegu i složenosti rizika povezanih s poslovnim modelom i aktivnostima banke i klijenta. Navedeno načelo otvara vrata prijenosu poslova procjene kreditnih rizika građana na računala koja danas doista pouzdano zamjenjuju ljude u mnogim nekad nemogućim, sada samo zamornim zadatcima prepoznavanja uzoraka.

Do prije nekoliko godina istraživanja na području procjene kreditnih rizika građana bila su vrlo ograničena. Kvantitativni modeli za kredite građana su razvijeni puno kasnije nego za kredite poslovnim subjektima, uglavnom, zbog problema vezanih uz dostupnost podataka. Podaci su bili, i još uvijek jesu, ograničeni na baze podataka financijskih institucija. Javno dostupnih baza podataka vezanih uz građane, a upotrebljivih za procjenu kreditne sposobnosti građana ima vrlo malo. Takvi podaci se u pravilu svode na nekoliko skupova podataka javno objavljenih od strane znanstvenika iz pojedinih zemalja. Tako se uglavnom u literaturi spominju njemački i australski skup podataka za kredite građana na kojima su izrađeni i testirani neki modeli. Navedeno nikako ne znači da ne postoji potreba za takvim modelima ili da podaci ne postoje. Za potrebe ovog istraživanja je prikupljen kreditni skup podataka, podaci su transformirati u oblik prikladan za dubinsku analizu podataka, statistički su opisani, a detaljniji prikaz prikupljenih podataka kao i tijeka prikupljanja podataka dan je u trećem poglavlju. Na tom skupu podataka kreirane su nove hibridne tehnike. Prednosti i nedostaci konstruiranih modela za procjenu kreditne sposobnosti tražitelja kredita provjereni su i uspoređeni s rezultatima dosad korištenih tehnika i na njemačkom skupu podataka, kao svjetski priznatom referentnom skupu podataka u ovom području.

1.2 Motivacija

Kreditna kriza, koja je počela u srpnju 2007, potresla je financijska tržišta, potkopala povjerenje klijenata i ulagača, podigla ozbiljne zabrinutosti i strahove financijskih institucija u stabilnost financijskih tržišta te zaprijetila gospodarstvima diljem svijeta. Iako je ova kriza

imala mnogo uzroka, jasno je sada da banke, vlade i druge institucije mogu učiniti više kako bi se spriječili takvi problemi u budućnosti. U tom kontekstu Baselski odbor za superviziju banaka je dao odgovor na krizu kroz sveobuhvatni paket reformskih mjera za jačanje regulacije, nadzora i upravljanja rizicima u bankarskom sektoru. Navedene mjere čine novi međunarodni regulatorni okvir za banke - Basel III. Reformske mjere su usmjerene na (BIS, 2011):

- regulaciju na razini banaka ili mikrobonitet, čime će pomoći u podizanju otpornosti pojedinih bankarskih institucija u razdobljima stresa, i
- makroekonomsku razinu, čime se žele spriječiti rizici koji se mogu proširiti preko cijelog bankarskog sektora, kao i procikličko djelovanje tih rizika tijekom vremena.

Komplementarne mjere treba poduzeti na razini banaka i sustava kao cjeline, jer veća otpornost na razini pojedinačne banke smanjuje rizik od udara na razini sustava, i obrnuto. Uslijed navedenog rizici dolaze u centar zbivanja, a uprave banaka prisiljene su tražiti nova rješenja za svoje poslovanje na način da imaju, u isto vrijeme, veću fleksibilnost i osjetljivost prema riziku. Kako bi za unutarnje sustave procjene rizika osigurale cjelovitost, banke moraju prikupiti podatke iz mnogih izvora i koristiti ih u procjeni rizičnosti podnositelja kreditnih zahtjeva kao i za redovnu klasifikaciju svojih klijenata. Regulatorni zahtjev traži da banke koriste sofisticirane kreditne modele radi povećanja učinkovitosti alokacije kapitala. Izbor varijabli za model predstavlja ključni korak razvoja modela. U disertaciji će biti prikazan vlastiti hibridni sustav temeljen na genetskom algoritmu i nekim drugim tehnikama za pronalaženje optimalnog skupa varijabli i konstrukciju modela. Primjena navedenih tehnika može povećati prihode, smanjiti troškove, povećati portfelj inovativnih proizvoda, privući nove klijente i zadržati postojeće. To je tehnologija koja ima perspektivu za brzo širenje s obje strane, sa strane banaka kao davatelja usluga ali i pojedinaca kao korisnika tih usluga, s druge strane. To je ujedno i doprinos širem korištenju metoda dubinske analize podataka, a samim time i većem ulaganju u taj segment IT-a.

Iz navedenog se vidi da je istraživanje motivirano stvarnim problemima. Vjerujemo da nema bolje motivacije za istraživanje od spoznaje da koncepti koje proučavamo imaju stvarnu vrijednost u svakodnevnom životu. Kroz eksperimentalnu provjeru je potvrđeno da koncepti, koji se izlažu, imaju primjenu u vjerojatno najvažnijem segmentu bankarske industrije, u odobravanju kredita i procjeni kreditne sposobnosti tražitelja kredita.

1.3 Svrha i ciljevi istraživanja

Svrha ovog doktorskog rada je temeljito istražiti ukupni skup podataka kojima raspolaže banka te utvrditi do koje mjere ti podaci mogu biti dobra osnovica za predviđanje sposobnosti tražitelja kredita da vrati kredit na vrijeme. Takvo predviđanje sposobnosti tražitelja kredita treba izvršiti bez traženja dodatnih podataka od klijenta uz pretpostavku da je tražitelj kredita već dulje vrijeme klijent banke te da je banka već u svojoj bazi podataka prikupila dovoljno podataka o klijentu. U skladu s time potrebno je iz postojećih, trenutno raspoloživih, podatka izvršiti klasifikaciju klijenata prema njihovoj kreditnoj sposobnosti. Pri tome se pojavljuje, kao jedan od glavnih problema, činjenica da u velikim količinama podataka i informacija koje su banke širom svijeta kumulirale u bazama podataka o svojim klijentima te njihovoj financijskoj i platežnoj povijesti obično ima i mnoštvo nebitnih atributa. U tom kontekstu nebitni atributi predstavljaju problem. Nebitni atributi u skupu podataka za učenje neće dovesti do preciznijih rezultata klasifikacijske analize, a povećavaju troškove prikupljanja podataka, vrijeme potrebno za učenje i konstrukciju modela kao i razumljivost samog modela. Stoga je potrebna priprema podataka za klasifikaciju u kojoj jednu od najvažnijih aktivnosti čini odabir optimalnog skupa atributa. Odabirom optimalnog skupa atributa se poboljšava kvaliteta konstruiranog modela za klasifikaciju, smanjuje složenost modela te se smanjuje trošak korištenja modela. Odabir optimalnog skupa atributa je važna aktivnost u pripremi podataka za klasifikaciju, a samim time i važan problem u kreaciji metode za predviđanje kreditne sposobnosti tražitelja kredita.

U skladu s tako definiranom svrhom istraživanja postavljen je i glavni cilj istraživanja: razviti vrlo efikasne, u skladu s najnovijim znanstvenim i tehničkim spoznajama, hibridne tehnike za odabir optimalnog podskupa atributa i predviđanje sposobnosti tražitelja kredita da će vratiti kredit na vrijeme i u skladu s ugovorenim obvezama.

U skladu s prethodno definiranim općim ciljem istraživanja definirani su i znanstveni ciljevi istraživanja:

- Kreirati hibridne tehnike selekcije atributa (GA-NN i HGA-NN), posebno prilagođene problemskoj domeni - temeljene na genetskim algoritmima i umjetnim neuronskim mrežama.
- Kreirati novi hibridni genetski algoritam uključivanjem rezultata filtarskih tehnika i a priori spoznaja u početnu populaciju genetskog algoritma.

- Kreirati novi operator selekcije kod genetskog algoritma, jedinstvenu selekciju (engl. *unique selection*).
- Kreirati sofisticirane kreditne modele koji omogućuju povećanje učinkovitosti alokacije kapitala što predstavlja doprinos implementaciji Basella III.

Osim znanstvenih ciljeva istraživanja, očekuje se da će rad kroz rezultate istraživanja dati primjenjiv društveni doprinos:

- Kroz prikaz primjene tehnika umjetne inteligencije u rješavanju važnih praktičnih problema prezentirajući cjeloviti postupak po kojem je istraživanje i rješavanje problema provedeno.
- Primjena kreiranih tehnika može značajno doprinijeti korištenju tehnika dubinske analize podataka u bankarstvu uz evidentne uštede na strani banke i pojedinaca, odnosno društva.
- Predložene tehnike uspješno objedinjavaju odgovarajuće tehnike i algoritme za otkrivanje znanja u bazama podataka te dokazuju da velike količine podataka koje se kumuliraju u bazama podataka mogu biti pretočene u znanje.
- Prezentirani koncept može se vrlo lako primijeniti i na rješavanje druge klase praktičnih problema kombinatorne optimizacije.

1.4 Hipoteze istraživanja

U skladu s prethodno navedenim znanstvenim ciljevima istraživanja, definirane su slijedeće hipoteze istraživanja.

H1: GA-NN tehnika je statistički značajno točnija na razini statističke značajnosti $p<0,05$ u odnosu na šire primjenjivane tehnike selekcije atributa: omjer dobiti, Ginijev indeks, korelacija i tehnika glasanja.

H2: Uključivanje preliminarne selekcije atributa i inkrementalne faze u algoritam temeljen na GA, kombinirano s efektima nove strategije generiranja inicijalne populacije GA, rezultira statistički značajnim poboljšanjem prosječne klasifikacijske točnosti novog algoritma uz razinu statističke značajnosti $p<0,01$.

Hipoteza H1 se odnosi na prvi znanstveni cilj istraživanja, a njezino prihvaćanje predstavlja potvrdu uspješne realizacije tog cilja. Dokazivanje točnosti novo kreirane

hibridne tehnike za selekciju atributa GA-NN (engl. ***G**enetic **A**lgorithm with **N**eural **N**etworks*), posebno prilagođene problemskoj domeni, provesti će se statističkim testovima u odnosu prema šire primjenjivanim tehnikama selekcije atributa: omjer dobiti, Ginijev indeks, korelacija i tehnika glasanja.

Hipoteza H2 direktno se odnosi na drugi, a potom i sve ostale znanstvene ciljeve istraživanja, a njezino prihvaćanje interpretirati će se kao potvrda uspješne realizacije tih ciljeva. Hipotezom H2 obuhvaćena su poboljšanja učinjena na genetskom algoritmu, čija je kreacija postavljena kao drugi znanstveni cilj istraživanja. Hipoteza definira i statističke uvjete pod kojima će se prihvaćanje te hipoteze smatrati potvrdom kvalitete učinjenih poboljšanja na genetskom algoritmu (engl. ***G**enetic **A**lgorithm*, GA) te potvrdom uspješne realizacije tog cilja.

S obzirom da je novi hibridni genetski algoritam glavna komponenta HGA-NN (engl. ***H**ybrid **G**enetic **A**lgorithm with **N**eural **N**etworks*) tehnike time se ova hipoteza odnosi i na prvi cilj istraživanja. Isto tako, sastavni dio novog hibridnog genetskog algoritma je novi operator selekcije, jedinstvena selekcija, stoga se hipoteza H2 odnosi i na treći cilj istraživanja. Kao četvrti cilj istraživanja navedeno je: Kreirati sofisticirane kreditne modele koji omogućuju povećanje učinkovitosti alokacije kapitala što predstavlja doprinos implementaciji Basella III. Uvažavajući činjenicu da su ti modeli rezultat primjene ranije navedenih tehnika na konkretnim skupovima podataka, na taj način hipoteze H1 i H2 pokrivaju i realizaciju ovog cilja istraživanja.

1.5 Metodologija istraživanja

U radu je prikazano više istraživanja. Prikazana su onim redom kojim su i provedena te kao takva čine jednu cjelinu, mada se mogu promatrati i kao samostalne cjeline. Prvo provedeno istraživanje, a u radu prikazano u poglavlju 3 metodološki ima 4 glavne etape:

1. prikupljanje podataka
2. priprema podataka koja uključuje
    - čišćenje podataka kroz otklanjanje nedosljednosti,
    - otkrivanje i analiza ekstremnih (stršećih, engl. *outliers*) vrijednosti,
    - deskriptivna analiza podataka i
    - selekcija atributa

3. klasifikacija i evaluacija
    - klasifikacija je temeljena na neuronskim mrežama
    - unakrsna validacija s *k* preklapanja (engl. *k-fold cross-validation*)
4. komparacija rezultata
    - matrica grešaka (engl. *confusion matrix*)
    - t-test, Wilcoxonov test uparenih (ovisnih) parova
    - komparacija rezultata temeljena na različitim odnosima troškova (FN/FP).

Ostala dva istraživanja u suštini slijede istu strukturu, osim prve i dijela druge faze istraživanja. Naime, prikupljanje originalnog skupa podataka provedeno je u okviru prvog istraživanja te je taj skup podataka korišten u drugom i trećem istraživanju. Valja naglasiti da drugo i treće istraživanje koriste i podatke iz UCI-jeve baze referentnih podataka nad kojima je također rađena analiza i selekcija atributa.

Prikupljanje originalnog skupa podataka je provedeno u periodu od 7 godina, u razdoblju od rujna 2004 do rujna 2011. Nakon što je prikupljen kreditni skup podataka, izvršena je priprema podataka za analizu.

U procesu pripreme podataka su popunjene nedostajuće vrijednosti, identificirane ekstremne vrijednosti i ispravljene nedosljednosti u podacima, u skladu s preporukama (Han i Kamber, 2006). Čišćenje šuma u podatcima je izvedeno kao iterativni proces. U tu svrhu su pisane skripte koje su pronalazile nepotpune, netočne, nedosljedne i stršeće vrijednosti nad kojima je rađena analiza. Uočene nedosljednosti u podacima otklanjane su temeljem tumačenja internih eksperata. Tražili su se razlozi takvih nekonzistentnosti i temeljem toga je prihvaćan podatak kojem se više vjeruje. Stršeće vrijednosti za više od 1,5 IQR ispod donjeg kvartila ili iznad gornjeg kvartila, tj. izvan ranga vrijednosti [Q1 – 1,5(Q3 – Q1), Q3 + 1,5(Q3 – Q1)] i manje od 2% ukupnih podataka su korigirale u navedene granice.

Pri inicijalnom prikupljanju podataka nastojalo se prikupiti što više atributa. Kada je bila dilema da li je neki atribut povezan s predmetom istraživanja, odnosno ciljnim konceptom, onda je svakako uključen u skup, kako se u suprotnom ne bismo izložili opasnosti da isključimo neki bitan atribut.

Međutim, nebitni atributi u skupu podataka za učenje neće dovesti do preciznijih rezultata klasifikacijske analize, a mogu produžiti vrijeme potrebno za učenje i konstrukciju modela kao i razumljivost samog modela, a tijekom korištenja konstruiranog modela mogu stalno povećavati troškove prikupljanja podataka i time povećavati trošak eksploatacije modela. Stoga je u sljedećoj fazi otkrivanja znanja potrebna priprema podataka za

klasifikaciju kako bi se u konačnici poboljšala prediktivna sposobnost konstruiranog modela ali i minimizirali ostali potencijalno negativni učinci tako prikupljenog skupa podataka. U toj pripremi podataka jednu od najvažnijih aktivnosti čini odabir atributa.

U postupku selekcije atributa se utvrđuje koji su atributi nepotrebni kako bismo ih isključiti iz skupa podataka jer njihovo prikupljanje predstavlja trošak, usložnjavaju proces učenja, produžuju ga i model čine kompleksnijim. Nepotrebni atributi posebno predstavljaju problem u uvjetima malog broja primjera za učenje i slabe popunjenosti pretraživačkog prostora.

Selekcija atributa temeljena je na genetskom algoritmu kao optimizacijskoj tehnici i na njemu baziranim, u ovom radu kreiranim novim tehnikama: GA-NN, HGA-NN te njihovim proširenim i troškovno poboljšanim inačicama.

Specifičnost tih tehnika je u tome što istovremeno izvode selekciju atributa kao i selekciju parametara umjetnih neuronskih mreža i traže optimalno rješenje na oba segmenta. Navedena specifičnost se postiže izvođenjem jednog genetskog algoritma u okviru drugog. Konceptualno se baziraju na pretpostavci da određeni skup atributa daje najbolje rezultate uz određeni algoritam za klasifikaciju i uz točno određene parametre algoritma za klasifikaciju. Dakle, idu korak dalje od klasičnih tehnika omotača koje rade s pretpostavkom da određeni algoritam klasifikacije daje najbolje rezultate za određeni set atributa.

Hibridni genetski algoritam koji je kreiran i prikazan u radu bolje prati "umjetnu evoluciju", proces evolucije u koju je umješan čovjek sa svojim znanjima i ranijim iskustvom na predmetnom području. Uključivanje (1) znanja domenskih stručnjaka, (2) rezultata filtarskih tehnika i (3) rezultata ranijih izvođenja hibridnog genetskog algoritma (HGA) u inicijalnu populaciju HGA, poboljšava, pri istom broju generacija, prosječnu dobrotu funkcije cilja HGA u usporedbi s prosječnom dobrotom funkcije cilja standardnog GA.

Ove GA-bazirane tehnike istovremeno podešavaju i prilagođavaju parametre klasifikacijskog algoritma podacima i obrnuto. Radi finog podešavanja parametara klasifikacijskog algoritma odabranom skupu atributa kreirana je generička tehnika NNGM (engl. *Neural Network Generic Model*). Ova tehnika se može promatrati kao dio GA-tehnike, ali i kao samostalna tehnika.

Validacija rezultata klasifikacije provodi se unakrsnom validacijom s *k* preklapanja (engl. *k-fold cross-validation*), osim zadnjeg eksperimenta u kojem se evaluacija rezultata klasifikacije provodi na cijelom originalnom skupu podataka.

U komparaciji rezultata koristi se matricu grešaka (engl. *confusion matrix*) koja pokazuje koliko je primjeraka određenog razreda klasificirano točno u onaj razred kojem

pripadaju, a koliko ih je greškom klasificirano u ostale razrede. Temeljem matrice grešaka i temeljem različitih odnosa troškova greške tipa I i tipa II (FN/FP) izvršena je i komparacija rezultata s troškovnog aspekta. Za svaki odnos troškova odabran je model koji mu najviše odgovora.

Statistička komparaciju rezultata izvodi se pomoću *t*-testa za zavisne skupove i neparametarskog Wilcoxonovog testa za zavisne skupove (engl. *Wilcoxon matched-pairs signed rank test*). Također se za potrebe *t*-testa provodi provjera normalnosti distribucije temeljem D'Agostino - Pearson testa (engl. *D'Agostino & Pearson omnibus normality test*) i Shapiro-Wilksova testa normalnosti.

1.6 Struktura rada

Ostatak ovog rada organiziran je na sljedeći način. U sljedećem poglavlju dan je uvod u kombinatornu optimizaciju; definirani su važni pojmovi s područja kombinatorne optimizacije. Prikazan je i kratki pregled tehnika te su naglašeni glavni koncepti. Ukazano je na njihov značaj, kako s istraživačkog tako i s praktičnog aspekta. Poglavlje prikazuje one koncepte koji su bitni prilikom razvoja hibridnih tehnika kombinatorne optimizacije. Cilj poglavlja je prikazati veze i odnose među konceptima i tehnikama kombinatorne optimizacije te na taj način dati opći uvod za istraživanja u sklopu ove doktorske disertacije. Poglavlje zaključujemo ističući važnost hibridizacije kao koncepta suradnje meta-heurističkih tehnika i drugih tehnika za optimizaciju.

Premda se cijelo poglavlje 2 može promatrati, u širem kontekstu, kao svojevrstan uvod u disertaciju, ipak, svako poglavlje daje cjelovitu obradu određene teme. Pri tome svako poglavlje daje: svoj uvod, prethodno objavljene srodne radove, metodološku pozadinu problema, detalje predloženog rješenja, rezultate temeljitih testova, diskusiju rezultata sa zaključcima i smjernicama za potencijalna buduća istraživanja.

U poglavlju 3 je predložena hibridna tehnika za odabir optimalnog podskupa atributa GA-NN, koja poboljšava klasifikacijsku točnost umjetnih neuronskih mreža. U okviru te tehnike razvijen je odgovarajući genetski operator selekcije, jedinstvena selekcija (engl. *unique selection*). Klasifikacijska točnost predložene tehnike je uspoređena s rezultatima šire primjenjivanih tehnika selekcije atributa: omjer dobiti, Ginijev indeks, korelacija i tehnika glasanja.

U poglavlju 4 je prezentiran novi napredni hibridni algoritam, kombinacija hibridnog genetskog algoritma i neuronskih mreža HGA-NN, za pronalaženje optimalnog podskupa atributa te poboljšanje klasifikacijske točnosti i skalabilnosti neuronskih mreža prilikom ocjene kreditnog rizika građana. Algoritam se temelji na sljedećoj hipotezi: visoko-dimenzionalni prostor hipoteza (potencijalnih rješenja) se preliminarno ograničava samo na važne dimenzije. U preliminarnoj restrikciji su korišteni brzi egzaktni algoritmi za rangiranje atributa kao i iskustvo domenskih eksperata. Pored toga, unapređenje je napravljeno i u kreaciji inicijalne populacije kao i uvođenju inkrementalne faze u genetski algoritam. Performanse predloženog HGA-NN klasifikatora su procijenjene korištenjem stvarnog kreditnog skupa podataka prikupljenog u hrvatskoj banci, a rezultati su dodatno provjereni na referentnom skupu kreditnih podataka iz UCI baze podataka. Pored toga, klasifikacijska točnost je uspoređena s rezultatima prezentiranim u literaturi.

U poglavlju 5 je detaljno analiziran problem klasno neuravnoteženih skupova podataka. Manje zastupljene klase su obično značajnije za predmetni sustav, a njihovo netočno predviđanje je povezano s većim troškovima i/ili težim posljedicama pogreške toga tipa. Kod toga, odnos između većinske i manjinske klase nije jedini faktor koji utječe na težinu klasifikacijskog zadatka. Među utjecajnim faktorima su, prije svih, razina u kojoj se preklapaju klase od interesa kao i šum. Sve navedeno podupire ideju da se različitim uvjetima, kao što su različiti skupovi podataka i različiti ciljevi istraživanja, moraju prilagoditi i tehnike klasifikacije. U tom kontekstu su tehnike GA-NN i HGA-NN, proširene određenim tehnikama za ublažavanje negativnog utjecaja klasno neuravnoteženih skupova podataka na rezultate klasifikacije. Rezultati klasifikacije su mjereni pomoću više različitih mjera najčešće korištenih kod problema s klasno neuravnoteženim skupovima podataka.

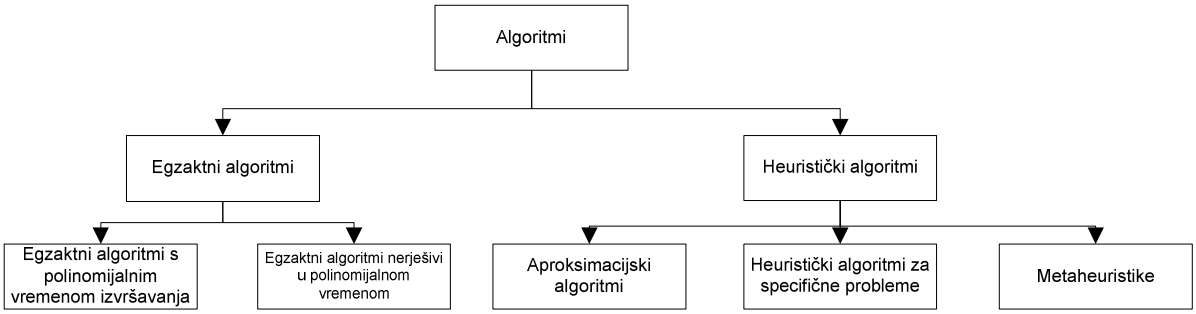
U zaključku su sagledani rezultati istraživanja te realizacija postavljenih ciljeva.

POGLAVLJE 2

TEHNIKE KOMBINATORNE OPTIMIZACIJE

Kombinatorna optimizacija (KO) (engl. *combinatorial optimization*) rješava optimizacijske probleme koji imaju naglašenu kombinatornu i diskretnu strukturu (Christofides, 1979). KO ima sve veći značaj zbog toga što veliki broj praktičnih problema može biti formuliran i riješen kao kombinatorno optimizacijski problem, kao i zbog činjenice da računala postaju sve performantnija te mogu kvalitetno riješiti probleme algoritmima kojima to ranije nije bilo moguće. Ti problemi su uglavnom NP-teški (engl. *NP-hard*), što znači da do sada nije poznat učinkovit algoritam za njihovo rješavanje u polinomijalnom vremenu, a malo je vjerojatno da će ga netko jednog dana naći, osim ako nije P = NP (Williamson i Shmoys, 2011). Za rješavanje spomenutih problema prvenstveno u obzir dolaze heuristički (uglavnom meta-heuristički) algoritmi. S njima se vrlo često mogu dobiti vrlo dobra ili čak (gotovo) optimalna rješenja u relativno kratkom vremenu, do kojih se, međutim, dolazi pod cijenu da nemamo nikakvo jamstvo o njihovoj kvaliteti (Papadimitriou, 2003).

Algoritme (tehnike) za rješavanje problema kombinatorne optimizacije možemo prikazati kao na slici 2.1.



Slika 2.1 Algoritmi za rješavanje problema kombinatorne optimizacije

Karakteristika egzaktnih algoritama u rješavanju takvih problema je da daju optimalna rješenja, ali u najgorem slučaju to postižu u eksponencijalnom vremenu izvršavanja. Unatoč mnogim istraživanjima i naporima znanstvenika i praktičara, egzaktni algoritmi su općenito primjenjivi samo na slučajeve ograničene veličine, iako je ta veličina definitivno porasla tijekom godina. Nasuprot tome, performanse heuristika su obično bolje s povećanjem veličine problema, što ih čini posebno atraktivnim za primjenu u praksi. Egzaktni i heuristički algoritmi su stoga aspekta nešto dijametralno suprotno i nitko ne može imati optimalnost i široku primjenjivost u isto vrijeme.

Heuristički algoritmi su razvijeni kao odgovor na nemogućnost egzaktnih algoritama u rješavanju mnogih problema. U slučaju NP-teških problema, mi žrtvujemo optimalnost u korist „dovoljno dobrih“ rješenja koja se mogu izračunati na efikasan način. Trgovanje optimalnošću u korist prilagodljivog rješavanja je paradigma heurističkih algoritama (Williamson i Shmoys, 2011).

Postoje brojni radovi koji se bave taksonomijom ovog vrlo dinamičnog područja. Neki se ograničavaju samo na određene aspekte kombinatorno optimizacijskih problema, dok drugi pokušavaju biti sveobuhvatni. Za potrebe ovog rada koristiti ćemo samo cjelovitije i/ili recentnije (Blum i Roli, 2003; Blum et al., 2011; Jourdan, Basseur i Talbi, 2009). Ova disertacija nastoji dati širi pregled područja na način da se u poglavlju 2.2 definiraju neki važni pojmovi s područja kombinatorno optimizacijskih problema. U poglavlju 2.3 se govori od računskoj složenosti problema, dok je poglavlje 2.4 posvećeno tehnikama rješavanja problema kombinatorne optimizacije i konceptima suradnje različitih tehnika pri razvoju hibridnih meta-heurističkih tehnika. Temeljni principi koje trebaju zadovoljiti tehnike za rješavanje kombinatorno optimizacijskih problema opisani su u poglavlju 2.5, iza čega slijede zaključna razmatranja ovog poglavlja.

2.1 Definiranje kombinatorno optimizacijskih problema

Optimizacijski problemi se u osnovi dijele u dvije kategorije: one kod kojih su rješenja kodirana realnim varijablama i one kod kojih su rješenja kodirana s diskretnim varijablama (Christofides, 1979). Među potonjima nalazimo klasu problema koju nazivamo problemi kombinatorne optimizacije (engl. *combinatorial optimization problems*). Kod njih smo u potrazi za objektom iz brojivog skupa (engl. *countable set*). Ovaj objekt je obično cijeli broj, podskup, permutacija ili struktura grafa (Blum i Roli, 2003).

Definicija 1. Kombinatorno optimizacijski problem $P = (S, f)$ može biti definiran pomoću:

- skupa varijabli $X = \{x_1, \ldots, x_n\}$;
- domena varijabli $D_1, \ldots, D_n$;
- ograničenja među varijablama C;
- ciljne funkcije f koja će biti minimizirana ili maksimizirana (u zavisnosti o problemu), gdje $f : D_1 x \ldots x D_n \rightarrow R^+$;

Skup svih mogućih pridruživanja je

$S = \{ s = \{(x_1, v_1), \ldots, (x_n, v_n)\} \mid v_i \in D_i, s$ zadovoljava sva ograničenja iz $C \}$.

S se obično naziva prostor pretraživanja (ili rješenja), pošto se svaki element skupa može promatrati kao kandidat za rješenje. Za rješavanje kombinatorno optimizacijskog problema mora se pronaći rješenje $s^* \in S$ sa minimalnom vrijednošću ciljne funkcije, tako da je, $f(s^*) \leq f(s)\ \forall s \in S$. s^* se zove globalno optimalno rješenje za (S, f), a skup $S^* \subseteq S$ se zove skup globalno optimalnih rješenja (Blum i Roli, 2003; Pirkwieser, 2012). U nastavku ovog rada ćemo, bez da to utječe na općenitost, raditi s minimizacijom vrijednosti ciljne funkcije *f,* pošto se maksimizacija može jednostavno transformirati u odgovarajuću minimizaciju uzimajući *–f*.

Važnost optimizacijskih tehnika, u današnjem društvu, proizlazi iz sveprisutnosti problema obrade velike količine podataka s ciljem donošenja racionalnih odluka koje, matematički gledano, predstavljaju globalno optimalno rješenje. U fokusu optimizacijskih

metoda je upravo kako doći do tog globalno optimalnog rješenja koje će posljedično ostvariti najbolje rezultate. Nažalost, kako je većina takvih problema NP-teška, nema efikasnog egzaktnog algoritma. Stoga vrijedi citirati (Williamson i Shmoys, 2011) stari inženjerski slogan koji kaže: "Brzo. Jeftino. Pouzdano. Izaberi dva." Prevedeno u kontekst KO, ako imamo $P \neq NP$, ne možemo istovremeno imati algoritme koji (1) daju optimalna rješenja (2) u polinomijalnom vremenu (3) za bilo koju instancu. U najmanju ruku, u bilo kojem pristupu kod NP-teških optimizacijskih problema, moramo popustiti na barem jednom od tih zahtjeva.

Pristup rješavanju problema u mnogome zavisi o tome na kojem od zahtjeva ćemo popustiti. Popustimo li na zahtjevu „za bilo koju instancu" , algoritam će biti koristan samo za manje složene probleme što u praksi nije čest slučaj. Popuštanje na zahtjevu „u polinomijalnom vremenu" čini se logičnijim. Ovaj pristup koji ne ograničava vrijeme i traži optimalno rješenje često, za složenije slučajeve, ostaje bez ikakvog rješenja. Odnosno, algoritam ne završi izvođenje u nekom razumnom vremenu. Daleko najčešće se u algoritmima popušta na zahtjevu da „daju optimalna rješenja". Labavljenje tog kriterija nastoji biti što je moguće manje jer se nastoji naći rješenje što bliže optimalnom. Algoritmi koji daju takva rješenja su približni algoritmi i možemo ih podijeliti (Papadimitriou, 2003) na (1) aproksimacijske algoritme (engl. *approximation algorithms*) i (2) klasične heuristike (engl. *heuristics*). Karakteristika je aproksimacijskih algoritama da daju u polinomijalnom vremenu približno (sub-optimalno) rješenje, ali i garanciju da je to rješenje "blizu" optimalnom. Karakteristika je klasičnih heuristika da brzim i jednostavnim računanjem daju približno rješenje, koje se obično pokazuje zadovoljavajućim unatoč nedostatku garancije „dobrote". Aproksimacijski algoritmi su poželjna vrsta približnih algoritama. Nažalost, oni se relativno rijetko koriste jer za brojne probleme aproksimacija nije moguća. Heuristike su predmet razmatranja mnogih istraživača i praktičara jer se pomoću njih rješava većina problema kombinatorne optimizacije (Sergienko, Hulianytskyi i Sirenko, 2009). U relativno novije vrijeme se pojavila nova vrsta približnih algoritama, metaheuristički algoritmi, a odnosi se na heuristike koji nisu posebno izrađeni za određeni problem.

Približni algoritmi obično se dijele na konstruktivne (engl. *constructive*) i algoritme za lokalno pretraživanje (engl. *local search*). Konstruktivni algoritmi generiraju rješenje od nule (engl. *from scratch*) dodajući inicijalno praznom skupu komponente rješenja sve dok se ne dobije kompletno rješenje. Ovakvi algoritmi su obično brži, ali često daju inferiorna rješenja u komparaciji s algoritmima za lokalno pretraživanje. Algoritmi za lokalno pretraživanje počinju od nekog inicijalnog rješenja te kroz iteracije pokušaju zamijeniti

trenutno rješenje s boljim iz odgovarajuće definiranog susjedstva. Susjedstvo je formalno definirano na slijedeći način (Blum i Roli, 2003):

Definicija 2. Struktura susjedstva je funkcija $N : S \rightarrow 2^S$ koja pridružuje svakom $s \in S$ skup susjeda $N(s) \subseteq S$. $N(s)$ se naziva susjedstvom od s. Često je struktura susjedstva implicitno definirana specificiranjem promjena koje se moraju primijeniti na rješenje s kako bi se generirali svi susjedi. Primjena takvog operatora koji proizvodi susjedno rješenje $s \in N(s)$ se obično naziva potez.

Uvođenje strukture susjedstva omogućuje nam da definiramo koncept lokalno minimalnog rješenja.

Definicija 3. Lokalno minimalno rješenje (ili lokalni minimum), uzimajući u obzir strukturu susjedstva N, je rješenje s^* tako da $\forall s \in N(s^*) : f(s^*) \leq f(s)$. Mi s^* nazivamo strogo minimalnim rješenjem ako je $f(s^*) < f(s)\ \forall s \in N(s^*)$.

Definiranje kombinatorno optimizacijskih problema i lokalnog minimuma je važno za taksonomiju i ukupno sagledavanje tehnika kombinatorne optimizacije. Prije nego što se pristupi razmatranju tehnika za rješavanje problema KO potrebno je sagledati tipove (klase) problema s aspekta računske složenosti algoritama za njihovo rješavanje kao i identificirati težinu određenog problema i njegovu pripadnost određenoj klasi problema. Pored toga, važno je definirati i više pojmova iz područja računske složenosti, stoga su teme iz teorije računske složenosti predmet sljedećeg odjeljka.

2.2 Računska složenost

Do sada smo u ovom poglavlju govorili samo o problemima (zadacima) koji su algoritamski rješivi. Postoje i problemi koji su algoritamski nerješivi. To su uopćeni problemi koji se ne mogu riješiti na suvremenim računalima koja rade na principima koje je definirao von Neumann. Zbog te činjenice ni bitno poboljšanje karakteristika računala neće doprinijeti rješavanju ovakvih problema. U grupu algoritamski nerješivih problema, između ostalih, spadaju (Papadimitriou, 2003; Marić, 2008): Problem zaustavljanja Turingova stroja, Deseti Hilbertov problem, Problem ekvivalentnosti riječi u asocijativnom računu itd.

Za nas mnogo interesantniju klasu predstavljaju algoritamski rješivi problemi jer za njihovo rješavanje postoje odgovarajući algoritmi različite složenosti. Složenost podrazumijeva računalne resurse, odnosno procesorsko vrijeme i količinu memorijskog prostora, koji su potrebni za pronalaženje rješenja. Praktično, najčešće je kritični resurs procesorsko vrijeme, tako da u zavisnosti od potrebnog procesorskog vremena za njihovo rješavanje, algoritamski rješive probleme dijelimo na klase složenosti.

Prema Definiciji 1, svaki KO problem određen je popisom parametara (varijabli) koji predstavljaju ulazne podatke i uvjetima koje mora zadovoljiti rješenje. Vrijednosti parametara određuju dimenziju problema. Na primjer, kod prepoznavanja jezika to je obično duljina riječi koju treba prihvatiti ili odbiti. Ako neki problem ima više parametara onda se na određen način može izvršiti kodiranje parametara, tako da se za svaki individualni zadatak na jedinstven način odredi dimenzija. U svemu tome treba imati na umu da je problem općenita klasa, a kada su dane specifične ulazne vrijednosti, govorimo o primjeru (instanci) (engl. *individual, instance*) problema, dakle, problem je skup instanci.

Formalno, neki problem je algoritamski rješiv ako postoji algoritam koji se može primijeniti na svaki njegov individualni zadatak. Međutim, nije dovoljno da postoji algoritam, već je potrebno da taj algoritam bude efikasan, odnosno pri nalaženju rješenja ne smije zahtijevati nerealne računalne resurse. Efikasnost algoritma s aspekta procesorskog vremena potrebnog za rješavanje zadatka nastojimo izraziti kroz vrijeme složenosti algoritma.

Kod procjene vremenske složenosti uvijek analiziramo najgori slučaj. Vrijeme složenosti nekog algoritma za rješavanje zadatka $t(n)$ je maksimalno vrijeme koje je potrebno algoritmu za nalaženje rješenja bilo kog individualnog zadatka koji ima dimenziju n (Marić, 2008). Najčešće je točan zapis funkcije $t(n)$ kompliciran, pa se ne koristi takav zapis već funkcija koja ima općenitiji i jednostavniji oblik, ali pri tome i ograničava funkciju $t(n)$. Praktično, za mjerenje vremenske složenosti algoritma najčešće se koristi Landauov simbol O (veliko O).

Definicija 4. Funkcija vremenske složenosti algoritma $f(n)$ je $O(g(n))$ ako postoji konstanta $C > 0$ i broj $n_0 \in \mathrm{N}$ takvi da vrijedi: $0 \leq f(n) \leq C \cdot g(n)$ za sve vrijednosti $n \geq n_0$. Prema tome $C \cdot g(n)$ je gornja granica za $f(n)$.

Na primjer, vremenska složenost algoritma za problem pronalaženja neke vrijednosti po neuređenom nizu od n elemenata je $O(n)$, dok je vremenska složenost algoritma za problem

pronalaženja vrijednosti po uređenom nizu $O(\log n)$, složenost sortiranja niza od n elemenata Quick-sort algoritmom je $O(n \log n)$ itd. Vrijednost eksponencijalnih funkcija, ali i funkcije $n!$ mnogo brže raste od polinomijalnih, pa se zbog toga razlikuju problemi sa eksponencijalnom složenošću (kod kojih se vrijeme predstavlja eksponencijalnom funkcijom) i problemi sa polinomijalnom složenošću, kod kojih je vrijeme složenosti $O(n^k)$, gdje je k neka konstanta. Kod procjene vremena zanemarujemo konstante. Tako je svejedno ima li algoritam $5n^2$, $7n^2$ ili $1000n^2$ koraka, važno je samo da se vrijeme ponaša kao kvadrat od n. Kažemo da naš algoritam zahtijeva vrijeme $O(n^2)$.

KO problemi se mogu svesti na probleme odlučivanja. Tako dobivenom problemu odlučivanja se može odrediti pripadnost određenoj klasi problema odlučivanja.

Problem odlučivanja je problem koji kao izlaz na postavljeno pitanje daje odgovor u obliku da/ne. Svaki problem optimizacije se može svesti na problem odlučivanja (Goldreich, 2010). Problem odlučivanja pripada klasi P ako se može riješiti deterministčkim algoritmom s polinomijalnom složenošću. Ključna karakteristika determinističkih algoritama je da ista kombinacija ulaznih parametara uvijek dovodi do istih rezultata.

Ispitivanjem, odnosno "pretragom" svih mogućih rješenja može se riješiti vrlo mnogo zadataka optimizacije. Praktično, do rješenja se dolazi ispitivanjem predstavlja li svako potencijalno rješenje i stvarno rješenje. Obično se ispitivanje jednog potencijalnog rješenja može obaviti polinomijalnim algoritmom. U skladu s navedenim, može se definirati pojam NP problema, odnosno problema koji se mogu rješavati nedeterminističkim polinomijalnim algoritmima. Problem pripada NP klasi ako se za neko izabrano potencijalno rješenje može deterministčkim polinomijalnim algoritmom provjeriti je li i stvarno rješenje. Ključna razlika između klasa P i NP je u tome što se problemi iz klase P mogu efikasno riješiti, dok se u slučaju problema iz klase NP rješenje može samo verificirati u polinomijalnom vremenu izvršavanja (Goldreich, 2010). Također, jasno je da svaki problem iz klase P pripada i klasi NP, ali je pravo pitanje, da li je P prava podklasa klase NP. Praktički, pitanje je; postoji li neki problem koji pripada klasi NP, a nije u klasi P. Postoji dosta rezultata koji su vezani za NP probleme, a koji pomažu da se razgraniči odnos klasa P i NP.

Premda pitanje je li P = NP za sada nema odgovora[1], ipak možemo identificirati probleme u NP koji su u određenom smislu “najteži”. Takvi problemi zovu se NP-potpuni. Oni imaju sljedeće svojstvo: ako postoji polinomijalni algoritam za jednog od njih, tada postoji polinomijalni algoritam za bilo koji problem iz klase NP. Stoga traženje odgovora na

[1] vidi http://www.claymath.org/millennium/P_vs_NP/ (pristupano 22.12.2011.)

pitanje je li P = NP može biti usmjereno na traženje polinomijalnog algoritma za bilo koji od NP-potpunih problema.

Većina rezultata u razgraničavanju P i NP problema dobivena je svođenjem jednih problema na druge. Svođenje se izvodi redukcijom. Neka redukcija *f* je izračunljiva u polinomijalnom vremenu ako postoji konstanta k i algoritam T koji za argument duljine *n* računa vrijednost od *f* u vremenu $O(n^k)$. Problem A je svodljiv u polinomijalnom vremenu na problem B ako se A svodi na B preko funkcije *f* izračunljive u polinomijalnom vremenu, a to se obilježava kao $A \leq_p B$. Svodljivost u polinomijalnom vremenu je predstavljena teoremom o svodljivosti (Arora i Barak, 2009).

Teorem 1 (a) Ako je $A \leq_p B$ i B je u P, tada je također i A u P.
(b) Ako je $A \leq_p B$ i A je u NP, tada je također i B u NP.

Ako problem B pripada klasi P i problem A se može u polinomijalnom vremenu svesti na B, onda i problem A pripada klasi P. Također, ako problem A pripada klasi NP i problem se može u polinomijalnom vremenu svesti na problem B, onda i problem B pripada klasi NP.

Smatra se da je problem A, NP-potpun ako:

- pripada klasi NP,
- svaki drugi problem B iz NP je polinomski svodljiv na problem A, što se označava kao $B \leq_p A$.

Ova definicija zbog prethodnog teorema zapravo znači sljedeće:

- Ako je A NP-potpun i A je u P, tada je P=NP.
- Ako je A NP-potpun, B je u NP i $A \leq_p B$, tada je i B NP-potpun.

Početni problem za koji je neposredno dokazano da je NP-potpun je „Problem zadovoljivosti Booleove formule u konjunktivno normalnoj formi“ (SAT), poznat kao Cook-Levinov teorem. Cook je dokazao da je problem zadovoljivosti NP- potpun. Standardna metoda dokazivanja NP-potpunosti je da se uzme problem za koji se zna da je NP-potpun te da je svodljiv u polinomijalnom vremenu na problem iz NP za koji želimo dokazati da je NP-potpun. Nakon što je ustanovljeno da je SAT NP-potpun, dalje je SAT reduciran na razne druge probleme iz NP i time je dokazano da su i ti drugi problemi NP-potpuni. Za veliki broj problema je pokazano da predstavljaju NP-potpune probleme. Problem je NP-težak ako i samo ako postoji ekvivalentan problem odlučivosti koji je NP-potpun (Korte i Vygen, 2012).

Kada je dimenzija NP-teškog problema dovoljno velika, onda se takav problem najčešće rješava korištenjem tehnike koja se naziva heuristika. Dosta često se u praksi pomoću heuristika može dobiti optimalno rješenje, ali se ne može verificirati njegova optimalnost. Najčešće je verifikacija optimalnosti rješenja višestruko složenija od samog dobivanja takvog rješenja. Heuristike su veoma popularne i često su izbor za rješavanje NP–teških problema, prvenstveno zbog njihove efikasnosti, ali i mogućnosti primjene na probleme velikih dimenzija.

Nakon što smo definirali problem kombinatorne optimizacije i računsku složenost problema, u narednom odjeljku nešto detaljnije prikazujemo moguće tehnike rješavanja problema kombinatorne optimizacije.

2.3 Tehnike rješavanja problema kombinatorne optimizacije

Treba napomenuti da teorijska definicija „učinkovito rješiv" ne mora nužno biti jednoznačna s učinkovitom rješivošću u praksi. Ipak ćemo ponoviti, za neke NP-teške probleme malih dimenzija mogu se koristiti egzaktne tehnike, međutim, one najčešće daju loše rezultate kada se primjene na NP-teške probleme velikih dimenzija.

2.3.1 Egzaktne tehnike

Iz metodičkog stajališta možda najjednostavniji egzaktan pristup bio bi potpuno nabrajanje svih mogućih rješenja zadatka *S*, što se još naziva iscrpno ili brutalno pretraživanje (engl. *brute-force search*). Zbog inherentne kombinatorne eksplozije koja se događa s povećanjem dimenzije problema, ovaj pristup je jedino moguć, kao što je već navedeno ranije, za slučajeve problema malih dimenzija. Stoga sve praktične pristupe egzaktnom rješavanju treba promatrati tako da se što je moguće više prostora pretraživanja pretraži samo neizravno, isključivanjem područja gdje je zajamčeno da se ne može naći bolje rješenje od ranije nađenog. Često se ove tehnike temelje na stablima pretraživanja (engl. *tree search*), gdje je prostor pretraživanja rekurzivno podijeljen po principu podijeli-i-osvoji (engl. *divide-and-conquer*) u međusobno razdvojene podprostore, fiksiranjem određenih varijabli ili nametanjem dodatnih ograničenja. Skalabilnost stabla pretraživanja bitno ovisi o

učinkovitosti mehanizma obrezivanja (engl. *pruning mechanism*). U mehanizmu granaj-i-ograniči gornja i donja granice su određene za ciljne vrijednosti rješenja, a odbačeni su podprostori za koje niže granice prelaze gornje granice (Pirkwieser, 2012).

Od brojnih egzaktnih metoda najčešće se koriste granaj-i-ograniči (engl. *branch-and-bound*), granaj-i-odsjeci (engl. *branch-and-cut*), linearno programiranje (engl. *linear programming*), dinamičko programiranje (engl. *dynamic programming*) itd. Za egzaktne metode se zna da su vremenski skupe, tako da se ne mogu primijeniti za velike NP-teške probleme. Za NP-teške probleme velikih dimenzija se koriste heurističke tehnike.

2.3.2 Heurističke tehnike

Heurističke tehnike rješavanja KO problema pokrivaju područje od jednostavnih konstruktivnih tehnika, kao što su ad-hoc pohlepni algoritmi (engl. *ad-hoc greedy algorithms*) preko tehnika za lokalno pretraživanje (engl. *local search*) do različitih metaheurističkih tehnika (engl. *metaheuristics*) (Pirkwieser, 2012).

Općenito, heuristika je tehnika kojom se traži dobro rješenje zadatka za relativno kratko vrijeme, bez mogućnosti garantiranja njegove dopustivosti i optimalnosti, često čak ni njegove bliskosti optimalnom rješenju (Marić, 2008). S obzirom na optimalnost rješenja, odnosno bliskosti rješenja s optimalnim rješenjem, heuristike možemo podijeliti na one koje garantiraju određenu kvalitetu rješenja i one koje ne daju nikakvu garanciju.

Aproksimacijski algoritmi

Aproksimacijski algoritmi su specijalna vrsta heuristika koja, nasuprot svim ostalima, garantira određenu kvalitetu aproksimacijskog rješenja. Stoga možemo reći da je karakteristika aproksimacijskog algoritma da daje u polinomijalnom vremenu približno (sub-optimalno) rješenje, ali i garanciju da je to rješenje "blizu" optimalnom. Mada je teško pronaći aproksimacijske algoritme sa dobrim garancijama kvalitete rješenja, područje aproksimacijskih algoritama je važno i zbog toga što donosi matematičku strogost u proučavanje heuristika. To nam daje mogućnost da dokažemo koliko dobro heuristike rade na svim instancama problema ili na samo nekim tipovima instanci (Williamson i Shmoys, 2011).

U nastavku ovog odjeljka, prema (Korte i Vygen, 2012; Pirkwieser, 2012), neka je rješenje neke proizvoljne instance I optimizacijskog problema P pomoću algoritma A

obilježeno sa $A(I)$, a vrijednost optimalnog rješenja označena sa $Opt(I)$. Neka je k neka konstanta. Dane su definicije:

Definicija 5. Aproksimacijski algoritam A ima apsolutnu garanciju kvalitete rješenja k ($k > 0$), ako za svaku instancu I vrijedi $|Opt(I) - A(I)| \leq k$.

Apsolutni aproksimacijski algoritmi postoje, za relativno malo problema. Češći su sljedeći algoritmi, koji daju relativnu garanciju kvalitete rješenja:

Definicija 6. Aproksimacijski algoritam A za minimizacijski problem ima relativnu garanciju kvalitete rješenja k ($k > 1$), ako za svaku instancu I vrijedi da je $A(I) \leq k \cdot Opt(I)$.

Algoritam A se tada naziva k-aproksimacijski algoritam (Korte i Vygen, 2012; Pirkwieser, 2012), a k garancija kvalitete rješenja ili aproksimacijski odnos ili aproksimacijski faktor (engl. *performance guarantee, approximation ratio, approximation factor*). U ovom radu slijedimo konvenciju da je aproksimacijski faktor za minimizacijski problem broj veći od 1 ($k > 1$), a za maksimizacijski problem $k < 1$. Dakle, $1/2$-aproksimacijski algoritam je algoritam u polinomijalnom vremenu koji za maksimizacijski problem uvijek vraća rješenje čija je vrijednost najmanje polovina optimalne vrijednosti.

Promotrimo relativnu devijaciju za minimizacijski problem:

$$\frac{A(I)-Opt(I)}{Opt(I)} \leq \epsilon \iff A(I) \leq (1+\epsilon)Opt(I). \qquad (2.1)$$

Analogno relativnoj devijaciji za minimizacijski problem (2.1), za maksimizacijski problem se nastoji maksimizirati vrijednost aproksimacijskog algoritma:

$$\frac{Opt(I)-A(I)}{Opt(I)} \leq \epsilon \iff A(I) \geq (1-\epsilon)Opt(I). \qquad (2.2)$$

Aproksimacijski algoritam A za minimizacijski problem s relativnom devijacijom ϵ je $(1+\epsilon)$-aproksimacijski algoritam. Slučaj za maksimizacijski problem završava sa $(1-\epsilon)$-aproksimacijskim algoritmom. Pojavljuje se i pojam asimptotska garancija, koju dobijemo dodavanjem konstante d: $A(I) < k \cdot Opt(I) + d$ za minimizacijski problem.

Postoji i klasa algoritama (skup algoritama) koji uzimaju odstupanje ϵ kao ulaz u algoritam.

Definicija 7. Aproksimacijski shema A je obitelj algoritama $\{A_\epsilon\}$ gdje postoji algoritam za svaki $\epsilon > 0$ takvi da A_ϵ je $(1+\epsilon)$-aproksimacijski algoritam (za minimizacijski problem) ili $(1-\epsilon)$-aproksimacijski algoritam (za maksimizacijski problem).

Aproksimacijske sheme možemo razlikovati prema vremenu izvršavanja:

Definicija 8. Aproksimacijski shema A je polinomijalna (PTAS) ako je vrijeme izvršavanja A polinomijalno obzirom na veličinu problema.

Polinomijalne aproksimacijske sheme su najpoželjnija, fleksibilna, vrsta relativnih aproksimacija gdje odabirom parametra u samom algoritmu možemo postići da algoritam daje željenu točnost. U literaturi se najčešće spominje problem naprtnjače kao vrsta problema koja se može riješiti pomoću ove vrste približnih algoritama.

Klasične heuristike

Rezultati dobiveni primjenom klasičnih heurističkih tehnika (heuristički algoritmi za rješavanje specifičnih problema) su nepouzdaniji, ali se ovim tehnikama mogu rješavati i NP-teški problemi velikih dimenzija. Kod heuristika poseban naglasak je na relativno kratkom vremenu traženja rješenja, budući da je u većini slučajeva kada je dimenzija NP-teškog problema velika, nemoguće naći egzaktno rješenje u razumnom vremenu. Ove tehnike se koriste i kada nije moguće definirati matematički model za probleme koji nemaju lijepe karakteristike, poput konveksnosti, diferencijabilnosti itd. Kod ovih tehnika se eksperimentalnim putem zaključuje je li neka takva tehnika bolja od neke druge do tada poznate tehnike.

Klasične heuristike razvijane su u cilju rješavanja pojedinačnih problema i veoma su vezane za karakteristike tog problema. Napredak postignut na području: (1) struktura podataka, (2) organizaciji i načinu čuvanja podataka, kao i (3) poboljšanju samih karakteristika računala, doveo je do reprogramiranja klasičnih heuristika. Na taj način nastale su nove klase heuristika - tzv. moderne heuristike, odnosno metaheuristike. Za razliku od

klasičnih heuristika, metaheuristike sadrže pravila i načela koja se mogu primijeniti pri rješavanju velikog broja praktičnih zadataka iz različitih domena. Kao što će detaljnije biti objašnjeno u nastavku, svaka metaheuristika je općenita tehnika i može se uspješno primijeniti na široku klasu problema.

Metaheuristike

U posljednjih 30 godina razvija se vrsta približnih algoritama koja u osnovi kombinira osnovne heurističke tehnike na konceptualno višoj razini s ciljem efikasnog pretraživanja prostora potencijalnih rješenja. Te tehnike se danas općenito nazivaju metaheuristike. Pojam metaheuristike (engl. *metaheuristic*) prvi je uveo Glover 1986. god. Prije prihvaćanja toga pojma često se koristio pojam moderna heuristika (engl. *modern heuristics*) (Blum i Roli, 2003). Korijeni metaheuristika nalaze se u umjetnoj inteligenciji i operacijskim istraživanjima (Blum et al., 2011) .

Mnogi autori su dali svoju definiciju pojma metaheuristike. Osman (2002) definira metaheuristiku kao iterativni glavni proces koji vodi i modificira operacije podređenih heuristika s ciljem efikasnog stvaranja visoko kvalitetnih rješenja. Kod toga se mogu inteligentno kombinirati različiti koncepti pretraživanja prostora potencijalnih rješenja koristeći prilagodljive strategije učenja i strukturirane informacije. Blum i Roli (2003) rezimirajući različite definicije pojma, navode temeljna svojstva koja karakteriziraju metaheuristike:

- metaheuristike su strategije koje vode proces traženja rješenja,
- cilj je efikasno istražiti pretraživački prostor kao bi se pronašla (sub)optimalna rješenja,
- tehnike kojima se služe sežu od jednostavnih procedura za lokalno pretraživanje do složenih procesa učenja,
- algoritmi kojima se služe su aproksimativni i obično nedeterministički,
- koriste mehanizme za izbjegavanje upadanja u lokalne optimume pretraživačkog prostora,
- osnovni koncepti dozvoljavaju apstraktnu razinu opisa,
- nisu orijentirane na specifični problem,

- mogu koristiti specifična domenska znanja u obliku heuristika koja su kontrolirana strategijom na višoj razini,
- naprednije heuristike koriste iskustvo za vođenje pretrage.

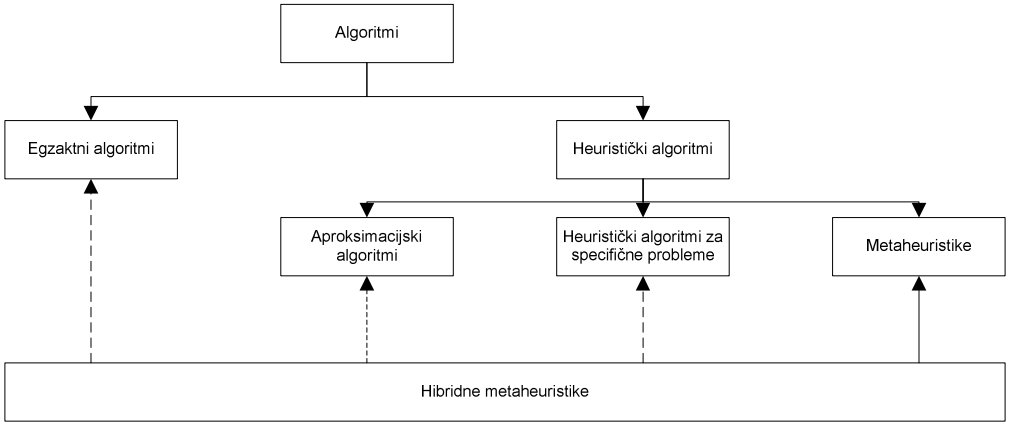
Možemo primijetiti da je glavna karakteristika metaheurističkih tehnika da se one, uz određene modifikacije, mogu primjenjivati na širok spektar problema kombinatorne optimizacije jer nisu orijentirane na specifični problem. Ove tehnike pretražuju skup dopustivih rješenja u cilju nalaženja što boljeg rješenja, pri čemu su dopušteni potezi: (1) kretanje ka lošijem rješenju od trenutnog, (2) proširivanje skupa na kojem se traži rješenje nedopustivim elementima, (3) traženje rješenja kombiniranjem postojećih itd. Postoji više podjela metaheuristika, a najčešće se spominje podjela koja dijeli metaheuristike, prema broju rješenja korištenih u isto vrijeme, u dvije kategorije (Jourdan, Basseur i Talbi, 2009): temeljene na jednom rješenju (engl. *single solution*) i temeljene na populaciji potencijalnih rješenja (engl. *population based*).

Tipični predstavnici metaheurističkih tehnika su:

- genetski algoritmi (engl. *Genetic algorithms*, GA),
- optimiranje rojem čestica (engl. *Particle swarm optimization*, PSO),
- optimiranje mravljom kolonijom (engl. *Ant colony optimization*, ACO),
- simulirano kaljenje (engl. *Simulated annealing*, SA),
- tabu pretraživanje (engl. *Tabu search*, TS),
- tehnika slučajnog prilagodljivog pretraživanja (engl. *Greedy Randomized Adaptive Search Procedure* - GRASP),
- tehnika promjenljivih okolina (engl. *Variable neighborhood search* - VNS)
- i brojne druge metaheurističke tehnike.

Svaka od tih metaheurističkih tehnika ima svoju pozadinu. Neke od tehnika su inspirirane prirodnim procesima, kao što je evolucija, druge su ekstenzija nekih manje sofisticiranih algoritama, kao što su pohlepni algoritmi i algoritmi za lokalno pretraživanje. Kada je postalo jasno da su čiste metaheuristike dosegle svoje granice, istraživači su se okrenuli prema kombinaciji različitih algoritama. Tijekom posljednjih godina, prilično impresivan broj algoritama više ne slijedi isključivo jednu paradigmu tradicionalnih metaheuristika. Naprotiv, oni kombiniraju različite algoritamske komponente, čiji korijeni često potječu iz

algoritama koji pripadaju različitim područjima optimizacije. Kao što prikazuje Slika 2.2, ovi algoritmi se nazivaju hibridnim metaheuristikama (Blum et al. , 2011).



Slika 2.2 Hibridna metaheuristika kao kombinacija algoritama

2.3.3 Hibridne metaheuristike

Glavna motivacija za križanje različitih algoritama je iskoristiti komplementarni karakter drugačijih optimizacijskih strategija, odnosno, kod hibrida se vjeruje u korist iz sinergije. Akkoc (2012) navodi da je prednost hibrida u tome što svaka tehnika donosi svoje vlastite prednosti. U posljednje vrijeme mnogi radovi (Jin et al., 2012; Chi i Hsu, 2012; Yuen et al., 2009; Lee i Ahn, 2011) nastoje iskoristi komplementarnost različitih tehnika. Pokazalo se da odabir adekvatne kombinacije komplementarnih algoritamskih koncepata može biti ključ za postizanje vrhunskih performansi u rješavanju mnogih teških optimizacijskih problema. Nažalost, razvoj učinkovite hibridne tehnike nije nimalo lak, naprotiv, to je općenito težak zadatak koji zahtijeva znanje iz različitih područja optimizacije.

U literaturi se često pojmovi hibridizacija i suradnja (engl. *hybridization and cooperation*) koriste u istom značenju, a označavaju algoritme koji kombiniraju različite tehnike optimizacije. U početku hibridizacija se uglavnom realizirala između više metaheuristika. Danas, se sve više predlaže suradnja između metaheuristika i drugih tehnika (Jourdan, Basseur i Talbi, 2009) . Takva strategija obično daje bolje rezultate jer je sposobna simultano iskoristiti prednosti obje vrste tehnika.

Blum i suradnici (Blum et al. , 2011) navode da se metaheuristike mogu hibridizirati s:

- drugim metaheuristikama,

- tehnikama programiranja s ograničenjima (engl. *constraint programming*),
- tehnikama temeljenim na stablima pretraživanja (engl. *tree search techniques*),
- tehnikama relaksacije problema (engl. *problem relaxation*) i
- tehnikama dinamičkog programiranja (engl. *dynamic programming*).

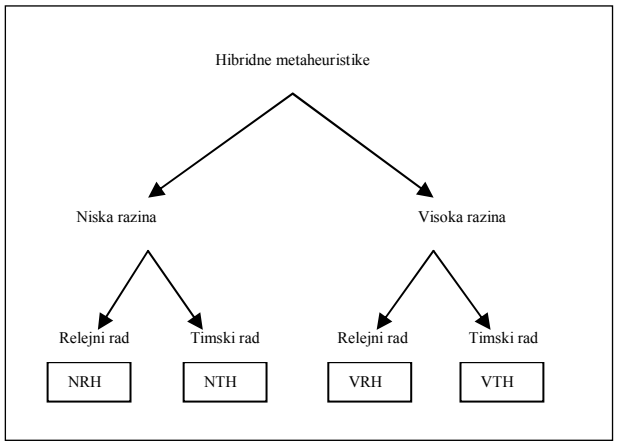
Suradnja različitih tehnika pri dizajnu metaheuristike može se promatrati s različitih aspekata (Jourdan, Basseur i Talbi, 2009). Prema razini na kojoj se odvija, suradnja može biti na:

- Niskoj razini: Rezultat takve suradnje je funkcionalno sastavljena jedna tehnika optimizacije. Određena funkcija unutar metaheuristike zamjenjuje se ili nadopunjuje drugom tehnikom.
- Visokoj razini: Različiti algoritmi su samodostatni.

Prema slijedu operacija, suradnja može biti:

- Relejni rad: Skup tehnika se primjenjuje jedna za drugom, svaka koristi izlaz prethodne kao svoj ulaz, djelujući kao cjevovod.
- Timski rad: Predstavlja kooperativne optimizacijske modele.

Slika 2.3 prikazuje 4 klase hibridnih metaheuristika koje dobijemo kada spojimo razinu na kojoj se odvija hibridizacija i podjelu s aspekta slijeda operacija.

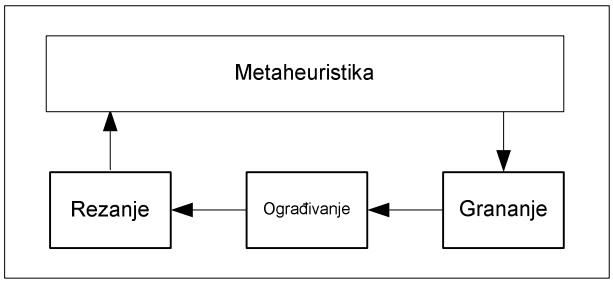


Slika 2.3 Klase hibridnih metaheuristika prema taksonomiji (Jourdan, Basseur i Talbi, 2009)

NRH klasa (hibrid niske razine, relejni rad)

Ova klasa odgovara algoritmima u kojima je određena tehnika ugrađena u drugu tehniku. Ugrađena tehnika se izvršava sekvencijalno, a izvođenje globalne tehnike ovisi o rezultatima dobivenim pomoću ugrađene tehnike. Ova vrsta suradnje je uobičajena kada je metaheuristička tehnika korištena za poboljšanje brzine egzaktne tehnike (Slika 2.4). U kontekstu suradnje među metaheuristikama, ova vrsta suradnje najčešće se realizira tako da se pokrene evolucijski algoritam koji tada pokrene lokalno pretraživanje, kako bi se intenziviralo pretraživanje područja najboljih rješenja (Jourdan, Basseur i Talbi, 2009).

U kontekstu suradnje između metaheuristike i egzaktnih tehnika može se razmotriti i sljedeći pristup: egzaktne tehnike daju potencijalna rješenja koja se koriste kako bi se definirao prostor pretraživanja za metaheurističku tehniku. Zatim se rezultati dobiveni metaheurističkom tehnikom mogu analizirati nekim drugim algoritmom. U ovu grupu hibridnih metaheuristika ulazi genetski algoritam hibridiziran nekim egzaktnim tehnikama, a korišten u hibridnoj GA-NN metaheuristici (Oreški i Oreški, 2014).

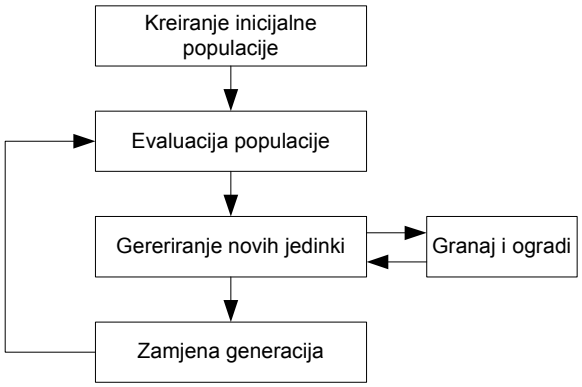


Slika 2.4. Skica NRH hibrida u kojem metaheuristička tehnika pretražuje širi prostor rješenja i optimira egzaktnu tehniku.

NTH klasa (hibrid niske razine, timski rad)

U ovoj klasi tehnika, dio jedne tehnike je zamijenjen ili nadopunjen s jednom ili više drugih tehnika. Ova vrsta suradnje može drastično poboljšati meta-heuristiku. Hibrid ove klase ima ugrađenu tehniku (tehnike) koja se može izvršiti paralelno s globalnom tehnikom. Kao klasični primjer NTH klase hibrida navodi se memetički algoritam koji predstavlja kombinaciju genetskog algoritma i tehnike lokalnog pretraživanja koja u genetskom algoritmu zamjenjuje transformacijski operator, najčešće mutaciju.

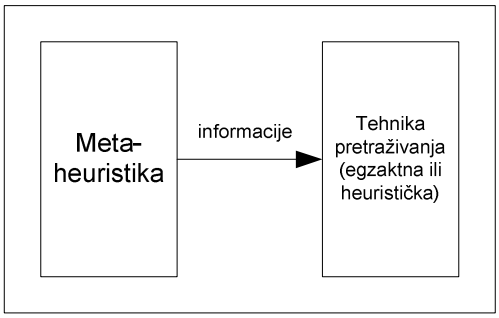
Značajno poboljšanje metaheuristike u NTH klasi prikazano je u radu (Muller et al., 2012). Hibrid se temelji na tehnici za lokalno pretraživanje i adaptivnom postupku koji iz složene strukture susjedstva odabire susjedno rješenje koje će biti analizirano u idućem koraku. U tu svrhu je u heuristiku ugrađena tehnika za mješovito cjelobrojno programiranje. U radu (Cotta et al., 1995) su prikazali hibridnu tehniku koja kombinira prednosti dviju vrlo različitih tehnika, genetskog algoritma i tehnike grananja i ograđivanja. Zbog različitosti tih tehnika, nova tehnika je razvijena tako da svaka od tehnika predstavlja alat druge tehnike (Slika 2.5).



Slika 2.5 Skica hibridnog genetskog algoritma (Cotta et al., 1995)

VRH klasa (hibrid visoke razine, relejni rad)

U VRH klasi hibrida različite metode su samodostatne i izvršavaju se u nizu. Ova shema suradnje je najzastupljenija kod hibridizacije. Kao i u drugim shemama suradnje, u ovoj klasi hibrida mogu postojati različite kombinacije rješavanja problema. Međutim, u cjelini, najprirodniji pristup je dizajnirati sekvencijalno izvršenje u kojem je metaheuristika pokrenuta prije egzaktne tehnike, tako da metaheuristika daje svoje rezultate egzaktnoj tehnici (Slika 2.6). Rezultati metaheuristika pomažu prilikom definiranja prostora pretraživanja egzaktnoj tehnici pretraživanja tako da joj reducira prostor pretraživanja prije samog pokretanja pretrage. U navedenom primjeru prostor pretraživanja bi mogao biti reduciran na susjedstvo oko predloženog rješenja.



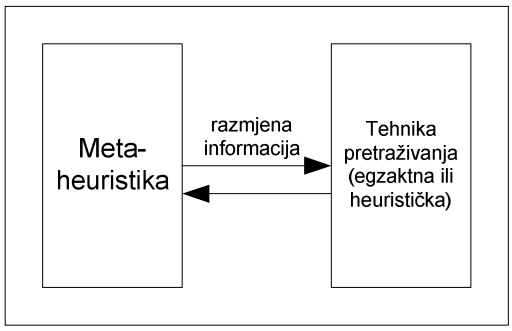
Slika 2.6 Metaheuristika predaje određena inicijalna rješenja i reducira granice pretraživanja

Primjer ove vrste hibridnog algoritma predlažu Bent i Van Hentenryck (2001) u svom dvostupanjskom algoritmu za transportni problem. Algoritam prvo minimizira broj vozila koristeći tehniku simuliranog kaljenja. Nakon toga minimizira troškove puta korištenjem tehnike velikih okolina (engl. *large neighborhood search*).

VTH klasa (hibrid visoke razine, timski rad)

Ova klasa sadrži algoritme u kojima samostalne tehnike izvode pretraživanje paralelno i na kooperativni način. Suradnja između metaheuristika, kao i metaheuristika i egzaktnih tehnika, odvija se kao među paralelnim otocima. Jedina razlika je što su to dvije različite vrste otoka; jedni su posvećeni točnom pretraživanju, a drugi izvode heurističko pretraživanje. Tijekom izvođenja, različiti algoritmi razmjenjuju informacije (Slika 2.7), a glavna poteškoća je utvrditi kada i kako razmijeniti informacije. Ovaj problem se nastoji riješiti parametarski.

Rad (Nwana, Darby-Dowman i Mitra, 2005) koji kombinira simulirano kaljenje i tehniku grananja i ograđivanja predstavlja primjer ove vrste hibridnog algoritma. Kod toga su korištena paralelna računala za istovremeno izvođenje tehnike simuliranog kaljenja i tehnike grananja i ograničavanja, koje su međusobno izmjenjivale informacije kako bi uzajamno utjecale na tijek i rezultate pretrage. Prezentirani rezultati, za širi spektar modela mješovito cjelobrojnog programiranja, su pokazali efikasnost korištenja ove hibridne tehnike.



Slika 2.7 Dva algoritma paralelno rade i razmjenjuju informacije.

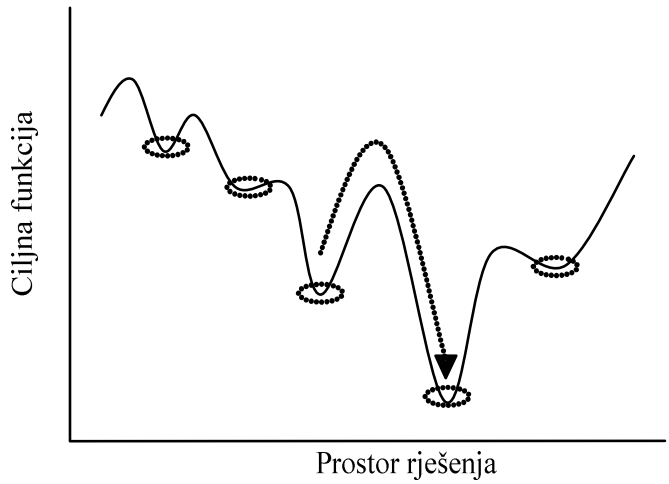
Među navedenim klasama hibrida vrlo je tanka granica. Ipak, svrha je postignuta jer prikazana podjela može poslužiti kao vodič i izvor ideja pri kreiranju novih hibridnih tehnika, ali i kao način sagledavanja i sistematizacije postojećih hibridnih tehnika. Kod svega toga valja zapaziti da prilikom kreiranja tehnika niske razine treba intervenirati u implementaciju same tehnike, dok kod tehnika viske razine u biti samo povezujemo „izvana“ dvije samostalne tehnike.

2.4 Diversifikacija i intenziviranje

Nakon kratkog prikaza tehnika rješavanja problema kombinatorne optimizacije, u kojem smo prikazali osnovne vrste tehnika, načine njihovog kombiniranja te potencijalne koristi od takvih kombinacija, u ovom dijelu ćemo prikazati koje temeljne principe pri rješavanju kombinatorno optimizacijskih problema trebaju zadovoljiti hibridne metaheuristike.

Hibridnu meta-heuristiku možemo promatrati kao strategiju visoke razine koja pretražuje prostor rješenja na različite načine. Kod toga je od najveće važnosti uspostaviti dinamičku ravnotežu između diversifikacije i intenziviranja (engl. *diversification and intensification*). Pod diversifikacijom se podrazumijeva snaga koja usmjerava proces pretraživanja potencijalnih rješenja na još neposjećena područja, generirajući time rješenja koja se na različite načine razlikuju od ranije viđenih. Intenziviranje fokusira pretragu na istraživanje susjedstva dotada najboljih rješenja (Glover i Laguna, 1997) .

Ravnoteža između diversifikacije i intenziviranja je važna radi brze identifikacije područja u prostoru pretraživanja s visokokvalitetnim rješenjima, ne gubeći previše vremena na područja u prostoru pretraživanja koja su već istražena ili ne daju kvalitetna rješenja. Postoje različite tehnike za postizanje te ravnoteže. Neke od tih tehnika predstavljaju vrlo inteligentna rješenja i pripadaju tehnikama lokalnog pretraživanja (Blum i Roli, 2003). Kod njih je cilj izbjeći lokalni minimum kako bi se nastavilo s pretraživanjem prostora rješenja s nadom pronalaska boljeg lokalnog minimuma, odnosno globalnog minimuma. U tu grupu ulaze npr. tehnika simuliranog kaljenja, tabu pretraživanje, tehnika slučajnog prilagodljivog pretraživanja – GRASP, tehnika promjenljivih okolina i njima slične.

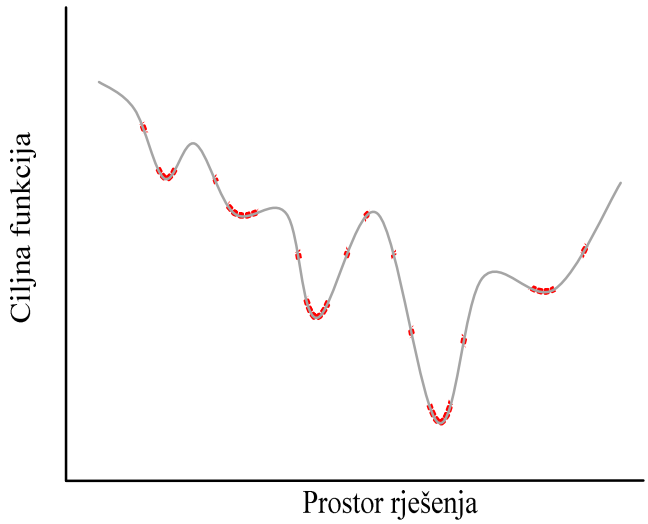


Slika 2.8 Temeljna ideja hibridnih metaheurističkih optimizacijskih tehnika

Drugu veliku grupu čine tehnike koje u svojoj filozofiji imaju komponentu učenja. One implicitno ili eksplicitno pokušaju naučiti korelaciju između nezavisnih i zavisne varijable i identificirati područja visoko kvalitetnih rješenja u prostoru pretraživanja. U tu grupu ulaze populacijski bazirane tehnike kao npr. genetski algoritmi, optimiranje rojem čestica (PSO), optimiranje mravljom kolonijom i njima slične.

Određene tehnike strateški različito rješavaju problem diversifikacije i intenziviranja i upravo u toj razlici pristupa postoji prostor za stvaranje kvalitetnih hibridnih tehnika. Hibridne tehnike mogu, a to i čine, kombinirati različite koncepte različitih tehnika i na taj način prevladati određene slabosti pojedinih metaheuristika, ili bolje rečeno, iskoristiti prednosti drugih kroz sinergizam. Ciljanim pristupom razvoju hibridnih metaheuristika može se postići, pored prednosti, i neželjena karakteristika koja se manifestira u tome da je rješenje (novi hibrid) strogo orijentirano na rješavanje samo određenog problema, pa nema širu primjenu. Tu dolazi od izražaja teorem "No free lunch" (Wolpert i Macready, 1996), koji kaže da nema tehnike koja bi bila bolja od svih drugih u svim uvjetima.

Strategiju pretraživanja rješenja moglo bi se opisati i na sljedeći način; atraktivna područja treba pretražiti temeljitije u usporedbi sa manje atraktivnim područjima. Stoga mnoge tehnike pamte kvalitetu prethodno ispitanih rješenja kako bi mogle ne samo na kraju odrediti najbolje rješenje nego i tijekom pretrage temeljitije istražiti susjedstvo boljih rješenja.



Slika 2.9 Ravnoteža između diversifikacije i intenziviranja

Na slici 2.8 područja boljih rješenja, područja lokalnih i globalnog minimuma, su obilježena elipsama. Slika 2.9 točkama prikazuje intenzitet pretraživanja pojedinih područja. Iz slika 2.8 i 2.9 se vidi da jedan od glavnih izazova pretraživanja područja rješenja predstavlja izlazak iz područja lokalnih minimuma prema globalnom minimumu ciljne funkcije. Stoga razmatrajući problem diversifikacije i intenziviranja Glover i Laguna (1997) navode da se cijeli proces pretraživanja odvija strateški u obliku oscilacija oko navedenih koncepata.

Ravnoteža između diversifikacije i intenziviranja u fokusu je mnogih radova u novije vrijeme. Navedimo samo neke tipične predstavnike. Rad Akpinar et al. (2012) prikazuje hibridni algoritam u kojem se kombinira optimiranje mravljom kolonijom sa genetskim algoritmom. Rama Mohan Rao i Shyju (2008) kombinirajući značajke simuliranog kaljenja i tabu pretraživanja prezentiraju učinkovitost kombiniranja dobrih osobina navedenih metaheuristika. Kako bi održali ravnotežu između intenziviranja i diversifikacije oni predlažu algoritam s više početaka. Mashinchi et al. (2011) predlažu hibrid temeljen na tabu pretraživanju i Nelder–Mead strategiji pretraživanja. Tabu pretraživanje je zaduženo za diversifikaciju dok je Nelder–Mead strategija fokusirana na intenziviranje pretrage. Radovi (Kao i Zahara, 2008; Yuan, Qian i Du, 2010; Al-Shihabi i Olafsson, 2010) također daju primjere kako postići ravnotežu navedenih koncepata.

2.5 Zaključci poglavlja

Temeljem svega navedenog možemo zaključiti da su hibridne tehnike kombinatorne optimizacije u novije vrijeme predmet interesa brojnih istraživača. Glavnu motivaciju istraživanjima ovog područja daju tisuće problema iz života koji mogu biti formulirani na apstraktnoj razini kao kombinatorno optimizacijski problemi. Upravo ti stvarni problemi razlog su promjene fokusa istraživanja, s isključivo algoritamski orijentiranih istraživanja prema problemski orijentiranim istraživanjima. U tom zaokretu došlo je do kombinacije komponenti iz različitih optimizacijskih tehnika. Poznavanje strateških koncepata kao i kvalitete pojedinih optimizacijskih tehnika preduvjet je za kreiranje novih efikasnih hibridnih tehnika. Jedino na taj način se može identificirati potencijalni doprinos različitih komponenti performansama buduće tehnike.

U ovom poglavlju su prezentirani određeni koncepti i generalni pristupi rješavanju problema kombinatorne optimizacije čije poznavanje je preduvjet za kreaciju dobre hibridne

tehnike za odabir atributa i klasifikaciju, a napose ako se klasifikacija temelji na velikim količinama podataka i informacija. Stoga ovo poglavlje u širem kontekstu predstavlja svojevrstan uvod u istraživanja prikazana u nastavku.

POGLAVLJE 3

TEHNIKA ZA ODABIR ATRIBUTA I PROCJENU RIZIKA TEMELJENA NA GENETSKIM ALGORITMIMA I NEURONSKIM MREŽAMA

3.1 Uvod

Konkurencija i regulativa, svaka sa svoje strane, pritišću banke na primjenu naprednih metoda u upravljanju kreditnim rizicima. Konkurencija stalnim smanjivanjem kamatne marže do razine kod koje banke više nemaju rezervi te mogu biti uspješne samo ako u poslovanju nisu izložene neočekivanim gubitcima. Konkurencija također sužava prostor za uzimanje dodatnih prvorazrednih kolaterala. Klijent je voljan dati dodatni prvorazredni kolateral samo ako ne može realizirati takav kredit kod neke druge banke bez takve vrste osiguranja. S druge strane, regulativa, temeljem zahtjeva Basel-a III, traži da banke koriste sofisticirane kreditne modele radi povećanja učinkovitosti alokacije kapitala. U takvim uvjetima menadžment banaka je prisiljen tražiti nove načine poslovanja, koji će u isto vrijeme biti fleksibilniji ali i osjetljiviji na rizik. Stoga, procjenu kreditnog rizika možemo promatrati kao

vjerojatno najvažniji i najteži segment bančinog poslovanja, a menadžment banaka kao najodgovorniji u prevenciji i usklađivanju kreditnih rizika s novim uvjetima.

Basel II je već otvorio prostor za djelovanje menadžmentu, dozvolivši bankama da mjere kreditni rizik i određuju razinu potrebnog kapitala korištenjem sustava temeljenim na internim rejtinzima (engl. *internal ratings-based approach* - IRB). Bazelski odbor (BIS, 2006) je također postavio i detaljan set minimalnih zahtjeva koji trebaju osigurati integritet unutarnjih procjena rizika. Kako bi sustavi interne procjene rizika bili cjeloviti, banke moraju prikupiti podatke iz više izvora na dnevnoj osnovi i koristiti ih u procjeni zajmotražitelja, a to isto trebaju činiti i za redovnu klasifikaciju svojih klijenata s ciljem povećanja učinkovitosti alokacije kapitala (Khashman, 2010). Prema tome, model odobravanja kredita i model procjene već danih kredita su važni u ostvarenju IRB pristupa. Kod toga treba poštovati načelo proporcionalnosti, temeljem kojeg upotreba metode klasifikacije ovisi o složenosti institucije, veličini i vrsti kredita, te predstavlja važno područje istraživanja sustava kreditnih scoringa.

Trenutno korišteni modeli kreditnog scoringa mogu biti prema Li, Shiue i Huangu (2006) kategorizirani u jedan od sljedeća dva pristupa: specijalističke prosudbe i statističko modeliranje. Prvi pristup se oslanja na iskustvo i subjektivne procjene bankarskih specijalista, što dovodi zbog njihova umora do loših procjena i sporog odziva, pošto je proces procjene rizika obično dugotrajan i mukotrpan. Statističko modeliranje, s druge strane, može izbjeći takve scenarije zbog svoje prirodne objektivnosti i dosljednosti.

Do danas su financijske institucije i istraživači razvili različite kvantitativne modele za kreditni scoring. Šušteršič, Mramor i Župan (2009) su klasificirali kvantitativne modele za kreditni scoring na sljedeći način: temeljene na klasičnim statističkim metodama, i temeljene na umjetnoj inteligenciji. Klasične statističke metode su linearna diskriminacijska analiza, logistička regresija, logit, probit, tobit, binarno stablo i metoda minimuma. Dvije najčešće korištene su diskriminacijska analiza (DA) i logistička regresija. Malhotra i Malhotra (2003) navode da diskriminacijska analiza ima problema zbog pristranosti ekstremnim vrijednostima, pretpostavke o multivarijantnoj normalnosti i pretpostavke o jednakim kovarijancama. Niti jedan od tih ograničenja se ne odnose na modele neuronskih mreža. Šušteršič, Mramor i Župan (2009) navode da je slabost linearne diskriminacijske analize pretpostavka linearnog odnosa između varijabli, koji je najčešće nelinearni kao i osjetljivost na odstupanja od pretpostavke o multivarijantnoj normalnosti. Logistička regresija ne zahtijevaju multivarijantnu normalnost. Zbog pretpostavke linearnog odnosa između varijabli i DA i logistička regresija imaju nedostatak u točnosti predviđanja.

Postoje i sofisticiraniji modeli poznati kao modeli umjetne inteligencije: ekspertni sustavi, sustavi temeljeni na neizrazitoj logici, neuronske mreže i genetski algoritmi. Među njima su neuronske mreže moguća alternativa za DA i logističku regresiju zbog moguće složenih nelinearnih odnosa između varijabli. U literaturi, u većini slučajeva, kod problema kreditnog scoringa neuronske mreže su točnije od DA i logističke regresije (Šušteršic, Mramor i Zupan, 2009). Međutim, veliki broj parametara, kao što su topologija mreže, stopa učenja i algoritam učenja, moraju biti dobro prilagođene problemu prije nego što neuronske mreže mogu biti uspješno korištene. Nadalje, mogu se pojaviti nedostatci kao što je zaglavljivanje u lokalnom optimumu, pretreniranost mreže ili zahtjev za velikim vremenom za učenje (Malhotra i Malhotra, 2003).

Khashman (2010) zaključuje da su prilikom odobravanja kredita, kao i kod redovite procjene danih kredita, neuronske mreže bile učinkovite u posljednjih deset godina. Sposobnost neuronske mreže se temelji na (1) načinu na koji mreža djeluje; temeljem algoritma za propagaciju greške unatrag (engl. *error back-propagation algorithm*) te na (2) dostupnosti podataka za učenje. Prilikom dovođenja podataka o podnositelju kreditnog zahtjeva na ulaz neuronske mreže, varijable sa ulaza neuronske mreže se uzimaju kao linearna kombinacija s proizvoljnim ponderima. Te varijable linearno se kombiniraju i podliježu nelinearnoj transformaciji pomoću određene aktivacijske funkcije (najčešće sigmoidne funkcije), a zatim ulaze u sljedeći sloj na slične manipulacije. Konačna funkcija daje vrijednost koja će biti uspoređena sa željenom vrijednosti. Svaki slučaj za učenje je predstavljen mreži, konačni izlaz se usporedi s promatranom vrijednosti, a razlika, pogreška, se prenosi natrag kroz mrežu i ponderi se mijenjaju na svakom sloju prema doprinosu svakog pondera vrijednosti pogreške. U suštini, mreža uzima podatke, transformira ih pomoću pondera i aktivacijske funkcije u skrivenim slojevima i na kraju u izlaznom sloju koji daje linearno odvojiv rezultat.

U okviru akademske zajednice pojavljuje se sve više literature o primjeni različitih metoda za kreditni scoring i ocjenjivanje kredita. Ne postoji konsenzus o tome koju tehniku projektant modela treba usvojiti za navedeni problem. S obzirom na ovu činjenicu, nije neobično da praktičar izgradi nekoliko klasifikatora koristeći različite tehnike, a zatim odabere onaj koji donosi najbolje rezultate za njegov problem. Međutim, kada se uspoređuju klasifikatori, nije nužno da jedan (najbolji) klasifikator nadmašuje sve ostale po svim područjima problemske domene. Pri tome se postotak pogrešaka često može smanjiti kombiniranjem izlaza nekoliko klasifikatora. Istraživanje kombinacija klasifikatora je bogato

i u posljednje vrijeme ima dosta literature na tom području (Finlay, 2011; Twala, 2010), te predstavlja još jedno važno područje istraživanja sustava za kreditni scoring.

Može se reći da pojedinačni klasifikatori predstavljaju prvo, kombinacije klasifikatora predstavljaju drugo, a izbor varijabli predstavlja treće važno područje istraživanja sustava kreditnih scoringa. Istraživači često ne razmatraju izbor varijabli kao ključni korak u razvoju modela; vjerojatno zbog problema (ne)dostupnosti podataka. Performanse svojih klasifikatora kao i performanse kombinacija klasifikatora ispituju na skupovima podataka danim u raznim repozitorijima. Ipak, pri razvoju uspješnog modela za kreditni scoring pitanje izbora varijabli je bitan i izazovan problem koji u praksi treba riješiti kao i pitanje korištenja klasifikatora (Šušteršic, Mramor i Zupan, 2009). Kao što je poznato, različite tehnike izbora varijabli daju različite rezultate za isti skup. U ovom poglavlju, cilj je dizajnirati hibridni sustav u osnovi baziran na genetskim algoritmima i umjetnih neuronskim mrežama (GA-NN) za pronalaženje optimalnog podskupa atributa u svrhu procjene kreditnog rizika građana. Takav optimalni podskup atributa treba omogućiti visoku točnost klasifikacije kod neuronskih mreža kao klasifikatora. S tim ciljem su ispitane različite kombinacije ulaznih varijabli u smislu njihovog doprinosa što točnijoj klasifikaciji kreditnih zahtjeva s aspekta kreditnih rizika. Potrebno je biti svjestan činjenice da se neće moći točno klasificirati svaki analizirani klijent. Savršena klasifikacija je nemoguća, budući da i dobri i loši klijenti ponekad prilikom odobravanja kredita imaju iste ili slične karakteristike. Tijekom korištenja kredita različiti događaji utječu na njihovu sposobnost podmirivanja preuzetih obveza. Stoga je greška kod klasifikacije neotklonjiva jer ne zavisi od primijenjenog matematičkog aparata (modela) nego o budućim događajima. Zbog toga se pri modeliranju nastoji determinirati pravilo koje rezultira najmanjim mogućim brojem netočnih klasifikacija ili najmanjim troškovima za banku.

Preostali dijelovi ovog poglavlja su organizirani na sljedeći način. Odjeljak 2 opisuje problem potrošačkih kredita koji se proučava u ovom poglavlju i analizira prethodnu literaturu u vezi s problemom. Kratak pregled tehnika i koncepata koji se koriste u istraživanju je prikazan u trećem odjeljku. Odjeljak 4 opisuje dizajn eksperimenta u pogledu prikupljanja podataka, izbora atributa, klasifikacije, ocjenjivanja i usporedbe rezultata. Odjeljak 5 analizira i diskutira rezultate eksperimenta s fokusom na točnost procjene i na troškove pogrešne klasifikacije. Odjeljak 6 zaključuje ovo poglavlje i daje neke smjernice za daljnji rad.

3.2 Opis problema i pregled literature

Prema BIS (BIS, 2011), kreditni rizik se najjednostavnije definira kao potencijal da tražitelj kredita ili druga strana u kreditu neće uspjeti ispuniti svoje obveze u skladu s dogovorenim uvjetima. Točnost predviđanja dobrih ili loših klijenata u pogledu kreditnog rizika može se poboljšati: (1) dobrom pripremom ulaznih podataka u čemu je od posebne važnosti izbor atributa, (2) korištenjem najboljih metoda klasifikacije i (3) kombiniranjem rezultata različitih klasifikacijskih metoda. Do prije nekoliko godina, istraživanja kreditnih rizika građana su bila vrlo rijetka. Kvantitativni potrošački kredit scoring modeli razvijeni su mnogo kasnije od onih za poslovne kredite, uglavnom, zbog problema dostupnosti podataka. Podaci su ograničeni na vlastite baze podataka financijskih institucija. Danas su neki podaci javno dostupni, a znanstvenici su razvili više različitih kvantitativnih tehnika za kreditni scoring (Šušteršic,Mramor i Zupan, 2009).

Primarni cilj istraživanja prikazanog u ovom poglavlju je **razviti vlastitu tehniku za odabir atributa i klasifikaciju klijenata** te temeljem te i drugih tehnika istražiti u kojoj mjeri ukupni podaci s kojima raspolaže banka o svojim klijentima, mogu biti dobra osnovica za predviđanje sposobnosti tražitelja kredita da vrati kredit na vrijeme. U skladu s time je postavljena glavna hipoteza istraživanja H3.1:

H3.1: GA-NN tehnika, temeljena na genetskim algoritmima i umjetnim neuronskim mrežama, je statistički značajno točnija, na razini pouzdanosti od 95% u odnosu na najčešće korištene tehnike odabira značajki: informacijska dobit, omjer dobiti, Ginijev indeks i korelacija.

Također se pretpostavlja da iz postojećih podataka o klijentima banke možemo izabrati takav skup podataka (indikatora) koji pružaju dobru osnovu za predviđanje kreditne sposobnosti zajmoprimca. Pod skupom podataka, koji daju dobru osnovu za predviđanje kreditne sposobnosti zajmoprimca smatrati će se skup podataka na temelju kojeg će procijenjena točnost predviđanja biti iznad 80%.

U novijoj literaturi na temu tehnika za odabir atributa i tehnika klasifikacije pronađeno je da su Malhotra i Malhotra (2003) koristi objedinjeni skup podataka o kreditima, prikupljen iz 12 različitih kreditnih unija s ukupno 1078 instanci sa šest ulaznih varijabli:

1. vlasništvo stana/kuće ,
2. duljina vremena na trenutnoj adresi (godina),
3. kreditne kartice,

4. omjer ukupnih isplata i ukupnog prihoda (omjer 1),
5. omjer duga i ukupnog prihoda (omjer 2) i
6. kreditni rejting podnositelja

kao čimbenike koji mogu praviti razliku između dobrih i loših kredita, odnosno njihovih tražitelja. Oni su otkrili da su modeli neuronskih mreža dosljedno bolji od modela multivarijatne diskriminacijske analize (MDA) u identificiranju potencijalnih problematičnih kredita.

Moderne tehnike dubinske analize podataka koje su postigle značajan doprinos na polju informacijskih znanosti, mogu biti usvojene za izgradnju modela kreditnih scoringa. Prema prikazanim rezultatima (Huang, Chen i Wang, 2007) koje su dobili Tam i Kiang, neuronska mreža je bila najpreciznija u predviđanju propasti banaka, nakon čega slijede linearna diskriminacijska analiza, logistička regresija, stabla odlučivanja i tehnika *k*-najbližih susjeda. U usporedbi s drugim tehnikama, zaključili su da su modeli neuronskih mreža točniji, prilagodljiviji i robusniji.

Šušteršič, Mramor i Župan (2009) su dizajnirali scoring model temeljen na neuronskim mrežama za potrošačke kredite u financijskim institucijama kod kojih podaci obično korišteni u prethodnim istraživanjima nisu dostupni. Oni prvenstveno koriste opsežni skup knjigovodstvenih podataka o transakcijama i stanjima računa klijenata koji su na raspolaganju svakoj financijskoj instituciji. Baza podataka za potrebe njihove studije je izrađena u slovenskoj banci i sjedinjuje sve računovodstvene i nekoliko drugih internih bankovnih podataka dostupnih za 581 kratkoročnih potrošačkih kredita odobrenih postojećim i novim bančinim klijentima u razdoblju 1994-1998. Od 581 kredita 401 (69,0%) su bili dobri i 180 (31,0%) su bili loši krediti. Dobri krediti su nasumično odabrani iz baze svih dobrih kredita banke odobrenih u tom razdoblju, a isto vrijedi i za loše kredite, respektivno. Obilježja svakog klijenta su u izvornoj bazi podataka opisana s 84 varijable. Konačno je odabrana 21 varijabla za korištenje u daljnjem istraživanju jer su te varijable omogućile najvišu točnost u pred-testiranju.

Tsai, Lin, Cheng i Lin (2009) su konstruirali model za predviđanje neplaćanja kod potrošačkih kredita prilikom provođenja empirijske analize o kupcima neosiguranih potrošačkih kredita u jednoj financijskoj instituciji u Tajvanu te su pri tome uzeli demografske podatke dužnika i odnos prema novcu kao diskriminacijske informacije. Njihovo istraživanje je najprije uključivalo samo osnovne demografske varijable kao čimbenike koji utječu na neplaćanje kod potrošačkih kredita. Nakon uključivanja i odnosa prema novcu, stopa točnosti predviđanja modela je relativno porasla u odnosu na točnost

predviđanja samo na temelju osnovnih demografskih varijabli. Kao rezultat toga, osim promatranja demografskih varijabli dužnika, u pogledu odabira varijabli za predviđanje neplaćanja, navedena studija je potvrdila da odnos dužnika prema novcu predstavlja važan čimbenik za što točnije predviđanje rizika neplaćanja ugovornih obveza. Njihov uzorak je imao 281 slučaj, a svaki slučaj 14 prediktorskih varijabli. Unutar uzorka je bilo 207 dobrih i 74 loših zajmoprimaca.

Khashman (2010) u svom radu opisuje sustav ocjenjivanja kreditnog rizika koji koristi model neuronskih mreža temeljen na algoritmu učenja s propagacijom greške unatrag. Neuronske mreže su učene korištenjem kreditnih slučajeva iz stvarnog svijeta, iz njemačkog skupa za odobravanje kredita koji ima 1000 slučajeva, svaki slučaj ima 24 numerička atributa koji sadrže razne financijske i demografske podatke o tražiteljima kredita. Klasificirajući atribut opisuje klijente kao dobre (700 zapažanja) ili loše (300 zapažanja). Khashman istražuje različite sheme učenja kao i različite omjere podataka u skupu za učenje i skupu za testiranje kod jednostavne validacije (holdout). Dolazi do zaključka da je optimalan odnos učenje prema testu 40%:60%. Detaljniji opis njemačkog skupa za odobravanje kredita dan je u poglavlju 4.

Twala (2010) tretira predviđanje kreditnog rizika kao neku vrstu problema strojnog učenja (engl. *machine learning*, ML). Između ostalog, on je obradio dva seta podataka iz UCI repozitorija, njemački i australski. Njemački set podataka je, kao što je navedeno, detaljnije opisan u poglavlju 4. Australski skup podataka sa zahtjevima za kreditne kartice ima 690 primjera sa 15 atributa. Od atributa, devet je diskretnih s 2-14 vrijednosti, a šest je kontinuiranih atributa. U tom skupu podataka postoji 307 pozitivnih primjera i 383 negativnih primjera. U 37 slučajeva nedostaje jedna ili više vrijednosti atributa. Svi nazivi atributa i vrijednosti su promijenjeni u besmislene simbole radi zaštititi tajnosti podataka.

Finlay (2011) koristi dva skupa stvarnih podataka. Podaci iz skupa A su dobiveni od tvrtke Experian UK. Skup sadrži podatke o kreditnim zahtjevima građana prikupljenim u nekoliko kreditnih institucija između travnja i lipnja 2002. Podaci iz skupa A sadrže 88.789 primjera, od kojih je 75.528 slučajeva klasificirano pozitivno (kao dobri), a 13.261 slučaj je klasificiran kao loš. Skup ima 39 nezavisnih varijabli. Podaci iz skupa B su vezani uz ponašanje klijenata, a ustupio ih je davatelj revolving kredita. Slučajni uzorak računa postojećih klijenata je preuzet u jednom trenutku tijekom 2002. Skup podataka sadrži 120.508 dobrih i 18.098 loših slučajeva. Bile su dostupne 54 nezavisne varijable, koje su se odnosile na aktualne i povijesne podatke o stanju na računima, dugovima, plaćanjima te raznim omjerima tih varijabli tijekom 3, 6 i 12 mjeseci.

Finlay (2010) je koristio skup podataka vezan uz ponašanje klijenata, a ustupio mu ih je veliki britanski davatelj revolving kredita. Set podataka je sadržavao 54 prediktorske varijable. To su tipične varijable o ponašanju klijenata. Uzorak je sadržavao 105.134 dobrih i 17.109 loših slučajeva.

U skupovima podataka kod svih navedenih studija, osim Finlay-eve nije bilo više od 1.078 uzoraka. Finlay-evo se istraživanje ističe značajno u broju promatranih slučajeva, ali nauštrb vremenskog razdoblja koje nije bilo dulje od 12 mjeseci. Naš skup podataka je sa 1000 slučajeva i s vremenskim horizontom od 7 godina. Sve navedene studije koriste minimalno 6, do maksimalno 81 nezavisnih varijabli koje se odnose na demografske i financijske karakteristike, zatim na ponašanje klijenata kao i na njihov stav prema novcu. Međutim, ne postoji studija koja objedinjava sve vrste podataka o podnositeljima zahtjeva. Pri tome su mnogi autori zaključili da su modeli neuronskih mreža (NN) točniji, prilagodljiviji i robusniji, u usporedbi s klasičnim statističkim tehnikama, stoga su u istraživanju prikazanom u nastavku ovog poglavlja kod genetskih algoritama za izračun funkcije dobrote korištene NN.

3.3 Metodologija

Primarni cilj istraživanja prikazanog u ovom poglavlju je razviti vlastitu tehniku za odabir atributa i klasifikaciju klijenata te temeljem te i drugih tehnika istražiti u kojoj mjeri ukupni podaci o klijentima banke, kojima raspolaže banka, mogu biti dobra osnovica za predviđanje sposobnosti tražitelja kredita da vrati kredit na vrijeme. Glavni naglasak je na odabiru atributa i klasifikaciji klijenata. U tome su korištene, samostalno ili u kombinaciji, sljedeće tehnike: genetski algoritam, tehnika odabira najboljeg atributa za slijedeći korak (engl. *forward selection*), informacijska dobit, omjer dobiti, Ginijev indeks i korelacija. Osim toga, u predloženom, za klasifikaciju i optimiranje parametara, generičkom modelu neuronskih mreža (engl. *Neural Network Generic Model* - NNGM), kao i kod nove tehnike za odabir atributa i klasifikaciju klijenata GA-NN, centralno mjesto uz genetski algoritam pripada neuronskim mrežama. U ovom odjeljku ćemo ukratko opisati navedene tehnike i koncepte koji se koriste u istraživanju prikazanom u ovom poglavlju.

3.3.1 Genetski algoritam

Stvaranje umjetne inteligencije i umjetnog života su ciljevi računalstva od samih početaka, tj. od početka računalnog doba. Najraniji znanstvenici na području računalstva: Alan Turing, John von Neumann, Norbert Weiner i drugi, bili su u velikom dijelu motivirani vizijom prožimanja računalnih programa s inteligencijom, sposobnošću samostalne replikacije te adaptivnom sposobnosti da uče i da kontroliraju svoju okolinu. Ti rani pioniri informatike bili su jednako toliko zainteresirani za biologiju i psihologiju kao i za elektroniku, te su promatrali prirodne sustave kao metafore vodilje za ostvarenje svojih vizija. Stoga ne bi trebalo biti iznenađenje da se, od najranijih dana, računala primjenjuju ne samo za izračun putanja raketa i dešifriranje vojnih šifarskih sustava, nego i za modeliranje mozga, oponašanje ljudskog učenja i simuliranje biološke evolucije (Mitchell, 1995).

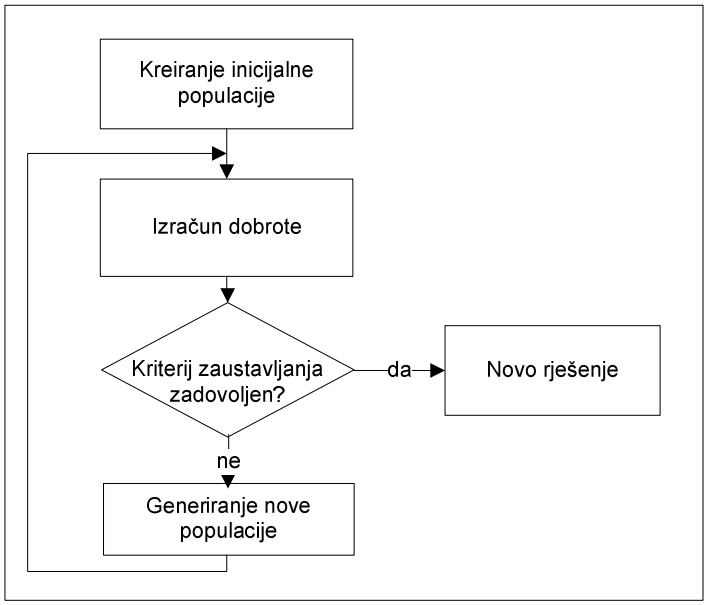
U skladu s navedenim ciljevima i nastojanjima pojavio se i genetski algoritam (GA) kao heuristička pretraga i optimizacijska tehnika inspirirana prirodnom evolucijom (McCall, 2005; Šušteršic, Mramor i Zupan, 2009) koja prenosi pojmove iz prirode u svijet računala i imitira prirodni razvoj. Pojam GA je inicijalno uveo John Holland za objašnjavanje procesa prilagodbe prirodnih sustava i za stvaranje novih umjetnih sustava koji rade na sličnim osnovama. U prirodi se novi organizmi prilagođavaju svojoj okolini kroz evoluciju. Genetski algoritam razvija rješenje za određeni problem na sličan način (Renner i Eka´rt , 2003).

U genetskom algoritmu, populacija nizova (zvanih kromosomi), koja kodira kandidate rješenja (zvani pojedinci, članovi ili fenotipovi) optimizacijskog problema, evoluira prema boljim rješenjima.

Prema tome, svaki GA operira na populaciji umjetnih kromosoma. To su nizovi izraženi konačnom abecedom, tradicionalno, u binarnom obliku kao nizovi jedinica (1) i nula (0). Svaki kromosom predstavlja rješenje problema i ima dobrotu (engl. *fitness*), realni broj koji izražava koliko je to rješenje dobro za određeni problem. Počevši s nasumično generiranom populacijom kromosoma, GA provodi proces odabira i rekombinacije, temeljen na dobroti, te proizvodi generacije jedinki. Tijekom rekombinacije se odabiru roditeljski kromosomi i njihov genetski materijal se rekombinira radi proizvodnje potomstva. Potomstvo potom čini sljedeću generaciju. Ponavljanjem ovog procesa razvija se slijed uzastopnih generacija, a prosječna dobrota kromosoma ima tendenciju poboljšanja sve dok se ne postigne neki kriterij zaustavljanja. Na taj način, GA razvija rješenje za određeni problem. Također je dokazano (McCall, 2005) da je GA jednostavan za implementaciju zbog svoje vrlo modularne prirode.

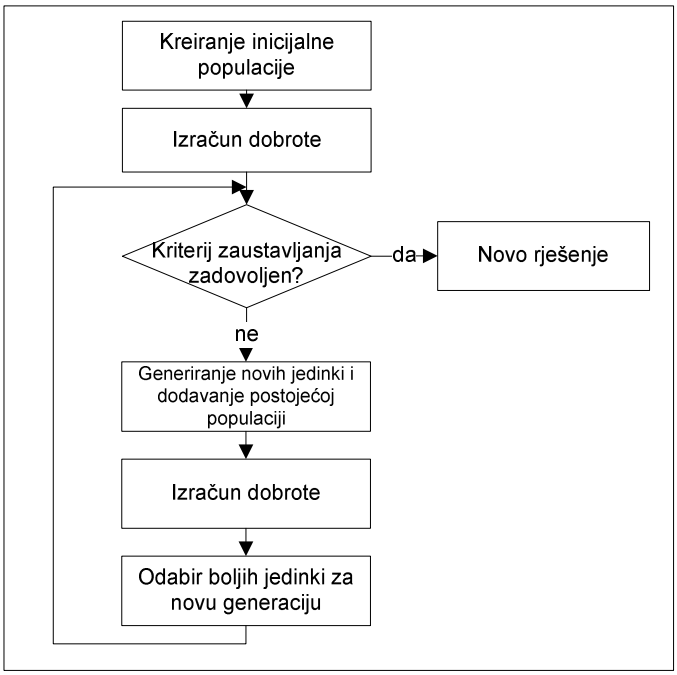
Kao posljedica toga, područje GA raste brzo, a tehnika je uspješno primijenjena na širok spektar problema u znanosti, inženjerstvu i industriji (McCall, 2005).

Implementacija genetskih algoritama može se značajno razlikovati prema načinu kreiranja nove populacije. Neke implementacije kreiraju zasebnu populaciju novih pojedinaca u svakoj generaciji primjenom genetskih operatora, kao što je prikazano na slici 3.1.



Slika 3.1 Algoritamski opis genetskog algoritma s kreacijom zasebne populacije

Ostale implementacije proširuju trenutnu populaciju dodavanjem novih pojedinaca, a zatim stvaraju novu generaciju izostavljajući najmanje sposobne pojedince, kao što je prikazano na slici 3.2. Postoje genetski algoritmi koji uopće ne koriste generacije, nego kontinuiranu zamjenu. Prema načinu stvaranja nove populacije, GA prilagođava druge operatore, posebice operator selekcije kao i mjesto izračuna dobrote pojedinaca.



Slika 3.2 Algoritamski opis genetskog algoritma s nadogradnjom postojeće populacije

Kao što se može vidjeti, s jedne strane postoji jednostavna jezgra genetskog algoritma, a s druge strane je problem koji želimo riješiti. Prezentacija, mehanizmi kontrole, funkcija dobrote, način inicijalizacije i genetski operatori trebaju biti na odgovarajući način prilagođeni problemu. GA ima posebnu snagu zbog toga što se standardne komponente mogu ponovno koristiti prilagodbom mnogim različitim situacijama, olakšavajući tako implementaciju. U nastavku ćemo se ukratko upoznati s osnovnim pojmovima genetskih algoritama: prezentacijom, funkcijom dobrote, križanjem, mutacijom i selekcijom.

3.3.1.1 *Prezentacija*

Pri projektiranju genetskog algoritma za određeni problem, odabir prikaza (prezentacije) je prvi korak. Bilo koji oblik kodiranja korišten za neki problem naziva se prezentacija problema. GA manipulira populacijama kromosoma koji su prezentacija rješenja određenog problema u obliku niza znakova. Kromosom je apstrakcija biološkog kromosoma (McCall, 2005), koji se može shvatiti kao niz slova abecede {A, B, D, O, S, T}. Određeno mjesto (engl. *locus*) na kromosomu nazivamo gen i slovo abecede koje se pojavi na tom mjestu u kromosomu nazivamo vrijednost alele ili jednostavno alela. Klasični GA koristi niz znakova u obliku bitova (bit-string) za kodiranje rješenja. Bit-string kromosomi se sastoje od niza

gena čije vrijednosti alela su znakovi iz abecede {0,1}. Kod problema u kojima se genetski algoritmi obično primjenjuju, mogući skupovi rješenja su konačni ali toliko veliki da iscrpna pretraga svih mogućih rješenja nije čak ni računalno izvediva. Nije neuobičajeno za GA da operira na bit-stringovima većim od 100 mjesta, što daje prostor rješenja veći od 2^{100}. Tumačenje tih nizova je u potpunosti ovisno o problemu.

Genetski algoritmi koriste zaseban prostor pretraživanja i prostor rješenja. Prostor pretraživanja je prostor kodiranih rješenja, odnosno genotipova ili kromosoma koji se sastoji od gena. Prostor rješenja je prostor stvarnih rješenja, odnosno fenotipova. Svaki genotip mora biti pretvoren u odgovarajući fenotip kod računanja dobrote. Na primjer, niz znakova duljine 32 bita može se koristiti za prikaz jednog cijelog broja (u standardnom binarnom zapisu) u jednom problemu, dok u drugom ovi bitovi mogu predstavljati prisutnost ili odsutnost 32 različitih faktora u složenom procesu. Snaga je genetskih algoritama u tome da se takva općenita prezentacija može koristiti za mnoštvo problema, omogućujući razvoj zajedničkih operatora i rutina za obradu, što ubrzava i olakšava primjenu GA-ova na nove situacije. S druge strane, posljedica je da će samo kodiranje kromosoma sadržavati samo ograničenu količinu problemu specifične informacije. Velik dio značenja određenog kromosoma za određeni problem je kodiran u funkciji dobrote.

3.3.1.2 ***Funkcija dobrote***

Funkcija dobrote (engl. *fitness function*) daje vrijednost kojom se ocjenjuje kvaliteta kromosoma kao rješenja određenog problema. To je proces prevođenja genotipa na fenotip, vrlo je važan i može biti prilično kompliciran. Vjerojatnost preživljavanja bilo koje jedinke određuje njezina dobrota; kroz evoluciju bolje jedinke istiskuju one manje sposobne. Kako bi se razvijala dobra rješenja, dobrota dodijeljena rješenju (kandidatu rješenja, jedinki) mora izravno odražavati njegovu kvalitetu, odnosno funkcija dobrote mora pokazati koliko dobro rješenje ispunjava zahtjeve određenog problema. Kriteriji dobrote se neprestano mijenjaju kako se fenotipovi razvijaju, evolucija je stalna potraga u skupu stalno mijenjajućih mogućnosti (Mitchell, 1996).

Pridruživanje vrijednosti dobrote može se obaviti na nekoliko različitih načina (Renner i Eka´rt , 2003):

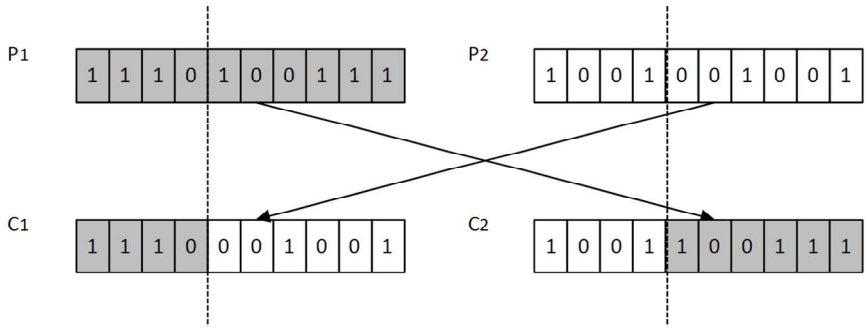
- Možemo definirati funkciju dobrote i ugraditi je u genetski algoritam. Pri ocjenjivanju bilo kojeg pojedinca (jedinke), funkcija dobrote se računa za pojedinca.

- Procjena dobrote se obavlja pomoću posebnog softvera za analizu. U takvim slučajevima procjena može biti dugotrajna, čime se usporava cijeli evolucijski algoritam.
- Ponekad ne postoji eksplicitna funkcija dobrote nego čovjek dodjeljuje vrijednost dobrote predstavljenom mu rješenju (kandidatu rješenja).
- Dobrota može biti dodijeljena i usporedbom jedinki u trenutnoj populaciji. Na primjer, ako je problem neka igra, a rješenje odražava strategiju igranja, njegova dobrota ovisi o nizu drugih rješenja u populaciji koje to rješenje može pobijediti.

Ukratko, cilj funkcije dobrote je osigurati smislenu, mjerljivu i usporedivu vrijednost za dani skup gena.

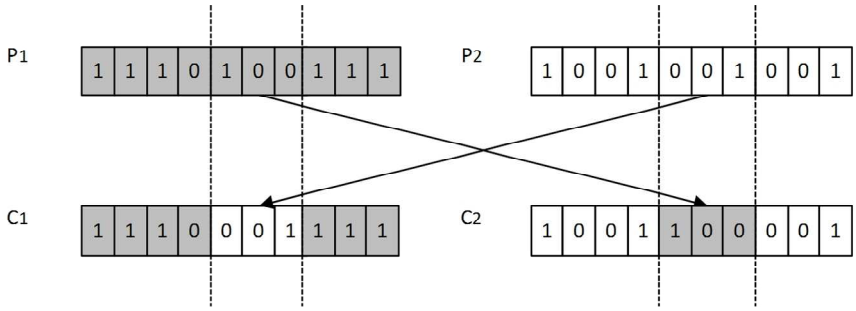
3.3.1.3 ***Križanje***

Operator križanja predstavlja miješanje genetskog materijala dvaju odabranih roditeljskih kromosoma radi stvaranja jednog ili dva kromosoma djeteta. Nakon što su dva roditeljska kromosoma odabrana za rekombinaciju, s uniformnom vjerojatnosti raspodjele generira se slučajni broj u intervalu [0,1] i uspoređuje s unaprijed zadanom stopom križanja (engl. *crossover rate*). Ako je slučajni broj veći od zadane stope križanja, križanje se ne obavlja. Ako je stopa križanja veća ili jednaka slučajnom broju, tada se križanje obavlja. Obično se koristi operator križanja s jednom točkom križanja (engl. *one-point crossover*). Točka križanja između 0 i *n* se odabire nasumce. Kromosomi potomaka su izgrađeni od karakteristika jednog roditelja prije točke križanja i karakteristika drugog roditelju poslije točke križanja. To je prikazano na slici 3.3 na nizu znakova duljine 10 bita.



Slika 3.3 Operator križanja s jednom točkom križanja

Postoji više oblika križanja. Križanje s jednom točkom uopćava se na dvije točke, kao što je prikazano na slici 3.4, i dalje na križanje u više točaka, gdje se niz točki križanja odabire uzduž duljine kromosoma, a kromosomi potomaka su izgrađeni od vrijednosti alela dvaju roditelja, izmjenjujući genetski materijal roditelja na svakoj točci križanja.



Slika 3.4 Operator križanja s dvije točke križanja

Uniformno križanje konstruira dijete ravnomjernim odabirom između vrijednosti alela roditelja na svakoj poziciji kromosoma. Algoritmi se također razlikuju s obzirom na to da li se kreira jedno ili više djece temeljem operacije križanja. Nakon križanja, dobiveni kromosom(i) bit će proslijeđeni u fazu mutacije. Operatori križanja i mutacije, mogu se promatrati kao način premještanja (evolucije) cijele populacije oko prostora definiranog funkcijom dobrote (Mitchell, 1996).

3.3.1.4 *Mutacija*

Mutacijom se stvara nova jedinka mijenjanjem genetskog materijala odabrane jedinke (unarni operator). Izmjena se može sastojati od promjene jedne ili više vrijednosti u prezentaciji jedinke ili dodavanja/brisanja dijelova prezentacije. U slučaju bit-string kromosoma, ovaj operater nasumično okreće neke od bitova u kromosomu (Mitchell, 1996). Na primjer, generira se slučajni broj u intervalu [0,1] s uniformnom vjerojatnosti raspodjele i uspoređuje s unaprijed zadanom stopom mutacije (engl. *mutation rate*). Ako je slučajni broj veći od stope mutacije, mutacija se ne primjenjuje na tom mjestu. Ako je stopa mutacije veća od ili jednaka slučajnom broju, tada se vrijednost alele okreće s 0 na 1 ili obrnuto. Stope mutacije su obično vrlo male, njihova vrijednost je često i manja od 1/*n*, gdje je *n* broj gena u kromosomima. U genetskim algoritmima mutacija je izvor varijabilnosti i provodi se nakon križanja.

3.3.1.5 *Selekcija*

GA koristi dobrotu kao diskriminator kvalitete rješenja predstavljenih kromosomima u populaciji GA. Kromosomi su stoga odabrani za reprodukciju na temelju dobrote. Oni s višim vrijednostima dobrote trebali bi imati veće šanse za izbor od onih s nižom dobrotom, stvarajući tako selekcijski pritisak prema boljim rješenjima. Uobičajen je izbor uz ponovnu mogućnost izbora, što znači da vrlo dobri kromosomi imaju priliku biti izabrani više od jednom ili čak da se križaju sa sobom, sa svojom kopijom. Tradicionalna metoda izbora je rulet izbor ili izbor proporcionalan dobroti. Ta metoda svakom kromosomu daje vjerojatnost da će biti odabran u omjeru njegove dobrote prema zbroju dobrota svih kromosoma u populaciji. Rulet kao proporcionalna selekcija ne jamči da najbolji član populacije prolazi u sljedeću generaciju, nego samo da ima vrlo dobre šanse da to učini. Proporcionalna selekcija funkcionira ovako: možemo zamisliti da je ukupna dobrota cijele populacije predstavljena strukturnim krugom (engl. *pie chart*) ili ruletnim kolom. Tada se dodijeli odsječak tog kruga svakom članu populacije. Veličina odsječka dodijeljena članu populacije je proporcionalna s ocjenom dobrote člana populacije. Bolji član populacije dobiva veći dio kruga, a time i veću šansu da bude odabran.

Postoji više različitih shema odabira (Renner i Eka´rt , 2003):

Odabir proporcionalan dobroti (engl. *fitness proportional selection*). Mehanizam opisan kao rulet izbor pripada ovaj shemi odabira. Kada se koristi ova shema odabira, potencijalno rješenje ima vjerojatnost odabira proporcionalno svojoj dobroti.

Odabir temeljen na rangu (engl. *ranked selection*). Problem je odabira proporcionalnog dobroti da se izravno temelji na dobroti. U većini slučajeva ne možemo definirati točnu mjeru kvalitete rješenja, tako da dodijeljena vrijednost dobrote ne izražava točnu kvalitetu rješenja. Ipak, jedinka s boljom vrijednosti dobrote je bolja jedinka. U izboru temeljenom na rangu jedinke su poredane prema njihovoj dobroti. Jedinke su zatim odabrane s vjerojatnošću razmjernom njihovom rangu koji se temelji na njihovoj vrijednosti dobrote.

Turnirski odabir (engl. *tournament selection*). U turnirskoj selekciji, iz populacije je najprije nasumce izabran skup od n jedinki. Tada se odabire najbolja jedinka iz tog skupa. Veća vrijednost n usmjerava odabir prema boljim jedinkama. Za $n = 1$, metoda je ekvivalentna slučajnog odabiru, za $n = 2$ relativno loši pojedinci imaju dobre šanse da prežive. Ako n dostigne veličinu populacije, preživjeti će samo najbolja jedinka.

Odabir odsijecanjem viška (engl. *cut selection*). Jednostavno, jedinke se poredaju prema njihovoj dobroti i odabire se fiksni broj (veličine populacije) najboljih kromosoma.

Jedinstveni odabir (engl. *unique selection*). Ova tehnika odabira jedinki iz populacije razvijena je u okviru našeg eksperimenta opisanog u nastavku ovog poglavlja. Konceptualno, ideja ove tehnike selekcije je da sačuva genetski materijal što je više moguće uz istovremenu konvergenciju prema sve boljim jedinkama. Tehnika radi na sljedeći način: kao i kod odabira odsijecanjem viška, jedinke se poredaju prema njihovoj dobroti i odabiru se najbolja rješenja, ali međusobno različita. Broj odabranih jedinki je manji ili jednak veličini populacije. Ako nema dovoljno različitih jedinki, populacija se dopunjuje s najboljim jedinkama. Cjeloviti pregled različitih shema selekcije koji se koriste u literaturi, kao i smjernice o tome kada je najprikladnije koristiti koju može se naći u (Bäck, Fogel i Michalewicz, 1997).

Selekcija članova populacije za reprodukciju i stvaranje novih generacija se nastavlja sve dok uvjet za završetak evolucijskog procesa nije zadovoljen. Obično algoritam završava kada se izvede maksimalni broj generacija ili kada se postigne zadovoljavajuća dobrota na razini cijele populacije. Ako algoritam (evolucijski proces) završava zbog dostignutog maksimalnog broja generacija, zadovoljavajuće rješenje može ali ne mora biti postignuto.

3.3.2 Tehnika odabira najboljeg atributa za slijedeći korak

Osnovna varijanta ovog algoritma obavlja odabir pod naivnom pretpostavkom da su atributi neovisni jedni od drugih. Ovaj pristup može donijeti dobre rezultate nakon kratkog vremena ako atributi doista nisu međusobno vrlo povezani. Algoritam počinje s praznim skupom atributa i, u svakom krugu, dodaje svaki neiskorišteni atribut iz zadanog skupa atributa. Za svaki dodani atribut, procjenjuje performanse unakrsnom validacijom. Samo atribut koji daje najveći porast performansi je dodan u tom krugu. Tada započinje novi krug izbora. Iteracije će se prekinuti kad se ispuni kriterij zaustavljanja, koji može biti: (1) kada nema daljeg povećanja učinkovitosti, (2) kada postoji povećanje učinkovitosti ali je manje od navedenog, bilo relativno ili apsolutno, (3) odabran je određen broj atributa. Poboljšana varijanta algoritama je opisano u nastavku:

1. Kreiraj početnu populaciju s n članova (kandidata rješenja), gdje je n broj atributa iz ulaznog skupa. Tako će svako inicijalno rješenje imati točno jedan atribut.
2. Ocijeni svako rješenje i odaberi samo najboljih k.

3. Za svako od *k* rješenja učiniti: Ako postoji *j* neiskorištenih atributa, učiniti *j* kopija tog rješenja i dodati točno jedan od prethodno neiskorištenih atributa postojećem rješenju.
4. Dokle god zaustavni kriterij nije zadovoljen, idi na korak 2.

Kad parametar *k* ima vrijednost 1 znači da se algoritam koristiti kao standardni algoritam izbora atributa. Korištenje drugih vrijednosti povećava vrijeme izvođenja, ali može pomoći da se izbjegne lokalni optimum u potrazi za globalnim optimumom.

3.3.3 **Informacijska dobit**

Informacijska dobit je tehnika odabira atributa koja daje rang za svaki atribut koji opisuje dane trening *n*-torke. Informacijska dobit temelji se na pionirskom radu Claude Shannona iz teorije informacija, koji se bavio proučavanjem vrijednosti poruke ili "informacijskim sadržajem" poruke. Atribut s najvišom informacijskom dobiti smanjuje količinu informacije potrebne za klasifikaciju *n*-torki u rezultirajućim particijama i odražava najniži stupanj slučajnosti ili "nečistoće" u tim particijama. Takav pristup smanjuje očekivani broj potrebnih testova u klasifikacijskom postupku.

Očekivana količina informacija potrebna za klasifikaciju *n*-torke u skupu D se izračunava pomoću jednadžbe:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (3.1)$$

gdje je p_i vjerojatnost da proizvoljna *n*-torka u D pripada klasi C_i, a procjenjuje se pomoću $|C_{i,D}| / |D|$. Koristi se log funkcija s bazom 2, jer su informacije kodirane binarno. *Info*(D) je samo prosječna potrebna količina informacija za identifikaciju oznake klase neke *n*-torke u skupu D. Imajmo na umu da se, u ovom trenutku, informacije koje imamo temelje isključivo na odnosu proporcije *n*-torki u svakoj klasi. *Info*(D) je također poznat kao entropija od D.

Sada, pretpostavimo da smo particionirali *n*-torke u D temeljem nekog atributa A koji ima *v* različitih vrijednosti, $\{a_1, a_2, ..., a_v\}$. Koliko informacija će još uvijek biti potrebno nakon particioniranja kako bi se došlo do točne klasifikacije? To se izračuna sljedećim izrazom:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j). \qquad (3.2)$$

Izraz $|D_j|/|D|$ djeluje kao ponder *j*te particije. $Info_A(D)$ je očekivana količina informacija potrebna za klasifikaciju *n*-torke iz D nakon particioniranja po A. Manje potrebno informacija, znači veću čistoću particija. Informacijska dobit se definira kao razlika između izvornog zahtjeva za informacijama (tj., onog koji se temelji samo na udjelu klasa) i novog zahtjeva (tj. dobivenog nakon podjele po A). To je,

$$Gain(A) = Info(D) - Info_A(D) \qquad (3.3)$$

Gain(A) nam govori koliko bi se dobilo podjelom po A. Očekivano je smanjenje informacijskih potreba uzrokovano poznavanjem vrijednosti od A. Atribut A s višim informacijskim dobitkom, *(Gain (A)),* je bolje rangiran po danim trening *n*-torkama.

3.3.4 **Omjer dobiti**

Tehnika informacijske dobiti preferira atribute s većim brojem vrijednosti. Ako uzmemo primjer atributa ID, podjela po njemu dovest će do onolikog broja podskupova koliko imamo primjera u skupu primjera za učenje, a svaki od podskupova imat će samo jedan element. Kako svaki element u takvim podskupovima pripada samo jednoj klasi, informacija koju možemo dobiti jednaka je nuli, tj. $Info_{ID}(D) = 0$. Dobit je u slučaju podjele po tom atributu maksimalna, taj atribut dobiva najviši rang, a mi od takve podjele nemamo nikakve koristi.

Nova tehnika koja pokušava prevladati pristranost iz navedenog slučaja naziva se omjer dobiti (engl. *gain ratio*). Ona to postiže uz svojevrsnu normalizaciju informacijske dobiti pomoću novodefinirane vrijednosti "*SplitInfo* " čija vrijednost je definirana analogno s *Info (D)* :

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right). \qquad (3.4)$$

Ova vrijednost uzima u obzir broj *n*-torki koje imaju određenu vrijednost atributa u odnosu na ukupan broj *n*-torki u D. Omjer dobiti se definira kao:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}. \quad (3.5)$$

Ako se sad vratimo primjeru sa ID-ovim sa početka, možemo vidjeti da će u tom slučaju *GainRatio* biti mali. Ako imamo *n* primjera za učenje, koji spadaju u *m* klasa, uz uvjet da imamo puno više primjera za učenje nego klasa, lako možemo vidjeti da će *GainRatio* biti malen, jer će informacijska dobit iznositi maksimalno $\log_2(m)$ dok će *SplitInfo* imati vrijednost $\log_2(n)$ koja je puno veća od $\log_2(m)$. Atribut s višim omjerom je bolje rangiran po danim trening *n*-torkama.

3.3.5 **Ginijev indeks**

Ginijev indeks mjeri nečistoću podataka, particije podataka ili skupa trening *n*-torki D, kao

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2, \quad (3.6)$$

gdje je p_i je vjerojatnost da *n*-torka iz D pripada klasi C_i, a procjenjuje se pomoću $|C_{i,D}| / |D|$. Zbroj se izračunava nad *m* klasa. Ginijev indeks za binarni atribut A koji dijeli D na particije D_1 i D_2 je

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2). \quad (3.7)$$

Smanjenje nečistoće nastalo po atributu A je

$$\Delta Gini(A) = Gini(D) - Gini_A(D). \quad (3.8)$$

Atribut koji maksimira smanjenje nečistoće (ili, ekvivalentno, ima minimalni Ginijev indeks) je najbolje rangiran po zadanim trening *n*-torkama.

Ginijev indeks kao i informacijska dobit su tehnike odabira atributa temeljene na stupnju nečistoće u rezultirajućim particijama. Za utvrđivanje ranga pojedinog atributa potrebno je usporediti stupanj nečistoće skupa (prije podjele) sa stupnjem nečistoće rezultirajućih particija (nakon podjele). Pošto je nečistoća skupa prije podjele jednaka za sve atribute,

maksimiziranje dobiti ekvivalentno je minimiziranju srednje vrijednosti mjera nečistoća rezultirajućih particija. Što je njihova razlika veća to je odabrani atribut bolji. Upravo zbog sličnosti postupka dolaženja do ranga atributa, ove dvije tehnike često daju slične ili gotovo identične rezultate.

3.3.6 **Korelacija**

Korelacija izražava statistički odnos između dviju slučajnih varijabli ili dva skupa podataka. Takav odnos između podataka često se kvantificira pomoću koeficijenta korelacije (Vořechovský, 2012). Za mjerenje stupanja korelacije postoji nekoliko koeficijenata korelacije, često se označavaju sa ρ ili r. Najčešći od njih je Pearsonov koeficijent korelacije, koji je osjetljiv samo na linearni odnos između dvije varijable. Pri tome je poznato da je Pearsonov koeficijent korelacije jako pod utjecajem rubnih vrijednosti, tj. stršećih ili ekstremnih podataka (engl. *outlier data*). Ostali koeficijenti korelacije su razvijeni kako bi bili više robusni na stršeće podatke od Pearsonova koeficijenta korelacije i osjetljiviji na nelinearne odnose.

Pearsonov produkt moment koeficijent korelacije (engl. *product moment correlation coefficient*), također poznat kao koeficijent linearne korelacije r, R, ili Pearsonov r, mjeri jačinu i smjer linearnog odnosa između dviju varijabli i definiran je pomoću kovarijance varijabli (uzorka) podijeljenih s njihovim standardnim devijacijama. Matematička formula za računanje je (Niven i Deutsch, 2012):

$$\mathrm{r} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{3.9}$$

koja također može biti napisana kao:

$$\mathrm{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{3.10}$$

gdje su $\overline{x}$ i $\overline{y}$ sredine uzoraka od X i Y, a n je broj parova podataka.

Vrijednost koeficijenta r je takva da je -1 ≤ r ≤ +1. Predznak koeficijenta korelacije je pozitivan ako su varijable izravno povezane i negativan ako su obrnuto povezane. Bliskošću s 1 ili -1 mjeri se jakost linearnog odnosa. Savršena korelacija ± 1 javlja se samo kad sve točke leže točno na ravnoj liniji. Ako je predznak koeficijenta korelacije pozitivan, nagib te linije je pozitivan i obrnuto. U nekim slučajevima svega nekoliko ekstremnih vrijednosti smanjuje inače visoku korelaciju. U ovom radu tehnika korištena za odabir atributa koristi Pearsonov koeficijent korelacije. Ovaj koeficijent korelacije nije mjerilo ukupnih odnosa jer ne daje informaciju o tome postoji li ili ne postoji nelinearni odnos između dviju varijabli. Stoga, korelacija 0 ne mora nužno značiti da su varijable nezavisne. Mora se naglasiti, da bi se shvatilo kako su varijable povezane ne treba se samo oslanjati na statistike (Ilakovac, 2009) kao što je koeficijent korelacije, nego podatke treba i grafički prikazati. U nekim kontekstima, čak i male korelacije mogle bi biti od velike praktične važnosti (Myers i Well, 2003).

Kao alternativa Pearsonovu koeficijentu korelacije, neparametarski Spearmanov koeficijent korelacije ranga mjeri koliko dobro može biti opisan odnos između dviju varijabli monotonom funkcijom. Spearmanov koeficijent korelacije ranga r_s može biti izračunat kao Pearsonov koeficijent korelacije između rangiranih vrijednosti ili pomoću jednostavne formule za rangove bez jednakih vrijednosti:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\left(r_{x_i} - r_{y_i}\right)^2}{n(n^2-1)} \tag{3.11}$$

gdje su r_{x_i} i r_{y_i} rangovi za x_i i y_i, respektivno. Spearmanov koeficijent korelacije ne zahtjeva pretpostavku o linearnom odnosu između varijabli i općenito je otporniji na stršeće vrijednosti nego Pearsonov koeficijent korelacije. Spearmanov koeficijent korelacije ranga je još uvijek osjetljiv na stršeće vrijednosti, posebno u uvjetima ograničenog skupa podataka. Napomenimo da su *r* i r_s algebarski ekvivalentni u slučaju kada se vrijednosti za X i Y sastoje od dva skupa od *n* rangova. Ipak, r_s je jednako precizan kao *r* samo kad su rangovi u okviru X i Y uzastopni cijeli brojevi : 1, 2, 3, ...*n*, bez jednakih vrijednosti. S jednakim vrijednostima biti će razlike između r_s i *r*. Ako je udio jednakih vrijednosti za X i Y prilično velik , tada je bolje koristiti standardnu formulu za izračun *r*.

Postoje i druge mjere za korelaciju, kao što je Kendall tau koeficijent korelacije ranga koji mjeri udio rangova koji su upareni između dva skupa podataka ali to više nije predmet našeg interesa. U nastavku je dan kratki prikaz neuronskih mreža kojima, uz genetski algoritam, pripada centralno mjesto u okviru tehnike za odabir atributa i klasifikaciju klijenata GA-NN.

3.3.7 **Neuronska mreža**

Neuronska mreža (NN) je računalni sustav za obradu podataka koji koristi veliki broj jednostavnih povezanih umjetnih neurona koji simuliraju karakteristike biološke neuronske mreže (Tsai et al., 2009). Neuronsku mrežu se može jednostavno definirati kao skup povezanih ulazno-izlaznih jedinica kod čega svaka veza ima određeni ponder ili težinski faktor (Han i Kamber, 2006).

Postoji mnogo različitih vrsta neuronskih mreža i algoritama neuronskih mreža. Najreprezentativniji i najpopularniji algoritam neuronskih mreža je algoritam s propagacijom greške unatrag (engl. *backpropagation algorithm*). Višeslojna mreža s propagacijom unaprijed (engl. *multilayer feedforward network*) je tip neuronske mreže na kojoj se algoritam izvodi.

Bitni elementi umjetne neuronske mreže (ANN) su jedinice za obradu (neuroni ili čvorovi) i algoritam učenja koji se koristi za pronalaženje vrijednosti pondera ANN za određeni problem. Neuroni su povezani jedni s drugima, tako da izlaz iz jednog neurona može biti ulaz mnogim drugim neuronima. Svaki neuron pretvara višestruke ulazne vrijednosti u jednu izlaznu vrijednost pomoću predefinirane jednostavne aktivacijske funkcije. U našem slučaju se koristi sigmoidna funkcija.

U većini slučajeva oblik aktivacijske funkcije je identičan u svim neuronima, međutim svaki skup pondera koji predstavljaju ulaz u ovu funkciju za svaki neuron je različit. Vrijednosti pondera se određuju učenjem (engl. *training*) na podskupu podataka koji se sastoji od poznatih ulaza u mrežu i izlaza. Arhitektura mreže je definirana organizacijom neurona i vrstom dopuštenih veza. Neuroni su raspoređeni u niz slojeva s vezama prema neuronima u drugim slojevima, ali ne i između neurona u istom sloju. Sloj koji prima ulaze zove se ulazni sloj. Završni sloj, koji daje izlazni signal ili odgovor, se zove izlazni sloj. Svi slojevi između ta dva sloja se nazivaju skriveni slojevi (Šušteršic, Mramor i Zupan, 2009). Višeslojna neuronska mreža koja ima ulazni sloj, jedan skriveni sloj i izlazni sloj naziva se dvoslojna neuronska mreža (Han i Kamber, 2006). Ulazni sloj se ne ubraja jer služi samo za prihvaćanje ulaznih vrijednosti i prosljeđivanje na sljedeću razinu. U njemu se ne izvršava nikakva transformacija. Slično, mreža koja sadrži dva skivena sloja se naziva 3-slojna neuronska mreža, i tako dalje (Han i Kamber, 2006).

Prije samog učenja, moraju se donijeti odluke vezano uz topologiju mreže. Određuje se broj jedinica u ulaznom sloju (odgovara broju atributa u *n*-torci za učenje), broj skrivenih slojeva (ako ih je više), broj jedinica u svakom skrivenom sloju te broj jedinica u izlaznom

sloju. Normalizacijom ulaznih vrijednosti, za svaki atribut prezentiran u skupu za učenje, može se ubrzati faza učenja. Tipično se ulazne vrijednosti normaliziraju tako da padaju između 0,0 i 1,0. Nema jasnih pravila po pitanju "najboljeg" broja jedinica u skrivenom sloju. Dizajn mreže je proces tipa pokušaja i pogreške, a može utjecati na točnost mreže nakon učenja. Početne vrijednosti pondera također mogu utjecati na konačnu točnost. Nakon što je mreža naučena i njezina se točnost ne smatra prihvatljivom, uobičajeno je da se ponovi trenažni proces s različitom topologiju mreže ili drugačijim skupom početnih pondera ili s drugom stopom učenja (engl. *learning rate*) ili drugim momentom (engl. *momentum*). Za vrednovanje procijenjene točnosti mreže mogu se koristiti različite tehnike kako bi se odlučilo kada je pronađena prihvatljiva točnost mreže. Predloženo je niz automatiziranih tehnika za pronalaženje "dobrih" parametara mreže. U disertaciji se predlaže generički model za optimiranje parametara neuronskih mreža – NNGM koji se temelji na genetskom algoritmu.

Propagacija greške unatrag

Algoritam s propagacijom greške unatrag uči iterativnim procesiranjem *n*-torki za učenje, uspoređujući predviđanje mreže za svaku *n*-torku s pravom, unaprijed poznatom ciljnom vrijednošću. Za svaku *n*-torku iz skupa za učenje mijenjaju se mrežni ponderi s ciljem minimizacije sredine kvadrata greške između predviđanja mreže i prave ciljne vrijednosti. Promjene pondera se obavljaju u smjeru prema unazad, tj. od izlaznog sloja, kroz skrivene slojeve do prvog skrivenog sloja. Od tuda i naziv algoritma.

Postupak počinje inicijalizacijom pondera. Ponderi se inicijaliziraju na male slučajne vrijednosti (npr. od -0,5 do 0,5, ili -1,0 do 1,0). Isto tako, svaka jedinica (neuron) ima svoju pristranost (engl. *bias*) koja se također inicijalizira na male slučajne vrijednosti.

Nakon toga se svaka *n*-torka iz skupa za učenje obrađuje kroz slijedeće korake:

Propagacija ulaznih *n*-torki: *n*-torka dolazi na ulazni sloj mreže. Prolazi kroz ulazni sloj nepromijenjena. Za ulaznu jedinicu *j*, njezin izlaz O_j je jednak njezinom ulazu I_j. Nakon toga se izračunava ulaz i izlaz svake jedinice u skrivenom i izlaznom sloju. Mrežni ulaz jedinica u skrivenim slojevima i izlaznom sloju se računa kao linearna kombinacija njihovih ulaza. Da bi smo izračunali ulaz u jedinicu, svaki ulaz povezan na jedinicu je pomnožen s odgovarajućim ponderom i sumiran. Za određenu jedinicu *j* u skrivenom ili izlaznom sloju, ulaz I_j u jedinicu se računa (Han i Kamber, 2006):

$$I_j = \sum_i w_{ij} O_i + \Theta_j; \tag{3.12}$$

gdje je w_{ij} ponder veze iz jedinice i u prethodnom sloju prema jedinici j; O_i je izlaz iz jedinice i iz prethodnog sloja; i Θ_j je pristranost jedinice. Pristranost služi kao moment sile u odnosu na ranije pogreške jedince. Na tako sumiran ulaz u jedinicu se primjenjuje aktivacijska funkcija. Kao što je navedeno ranije, u svim istraživanjima ove disertacije se koristi sigmoidna funkcija:

$$O_j = \frac{1}{1 + e^{-Ij}}, \tag{3.13}$$

gdje je I_j mrežni ulaz u jedinicu j, a O_j je izlaz iz jedinice j. Logistička funkcija je nelinearana i diferencijabilna, čime omogućava algoritmu povratnog prostiranja (propagacije) da modelira klasifikacijske probleme koji su linearno nerazdvojivi (Han i Kamber, 2006).

Propagacija pogreške unatrag: pogreška se propagira unazad ažuriranjem pondera veza i pristranosti jedinica tako da odražavaju pogrešku predviđanja mreže. Za jedinicu j iz izlaznog sloja, pogreška Err se računa:

$$Err_j = O_j (1 - O_j) (T_j - O_j), \tag{3.14}$$

gdje je O_j stvarni izlaz jedinice j, a T_j je poznata ciljna vrijednost za danu n-torku. Pri tome je $O_j (1 - O_j)$ derivacija logističke funkcije. Dakle, derivacija logističke funkcije množi razliku između ostvarene i ciljne vrijednosti da bi se dobila pogreška jedinice iz izlaznog sloja.

Da bi smo izračunali pogrešku jedinice j u skrivenom sloju treba uzeti u obzir ponderiranu sumu pogrešaka jedinica u sljedećem sloju povezanih s jedinicom j. Prema tome, pogreška jedinice j u skrivenom sloju je:

$$Err_j = O_j (1 - O_j) \sum_k Err_k w_{jk}, \tag{3.15}$$

gdje je w_{jk} ponder veze jedinice j i jedince k u sljedećoj višem sloju mreže, a Err_k je pogreška jedinice k. Možemo primjetiti da se kod izračuna pogreške za neku jedinicu izlaznog sloja derivacija funkcije množi s razlikom između ostvarene i očekivane vrijednosti za danu jedinicu (izlazni neuron), dok se kod jedinica u skrivenom sloju derivacija množi s ponderiranom pogreškom svih jedinica u sljedećem višem sloju mreže (Han i Kamber, 2006).

Tako izračunatim pogreškama prilagođavaju se ponderi veza i pristranosti. Ponderi veza se prilagođavaju pomoću sljedećih jednadžbi (Han i Kamber, 2006):

$$\Delta w_{ij} = (l)Err_j O_i, \quad (3.16)$$
$$w_{ij} = w_{ij} + \Delta w_{ij}, \quad (3.17)$$

gdje je Δw_{ij} promjena pondera w_{ij}, a l je stopa učenja.

Stopa učenja je konstantna koja obično ima vrijednost od 0,0 do 1,0. Mreža temeljena na algoritmu s povratnom propagacijom greške uči pomoću metode gradijentnog silaska u potrazi za skupom pondera (modelom) koji odgovaraju skupu za učenje, na način da se minimizira srednja kvadratna udaljenost između predviđanja mreže i poznatih ciljnih vrijednosti skupa za učenje. Kod toga stopa učenja pomaže da se izbjegnu zapinjanja na lokalnom minimumu u prostoru pretraživanja.

Ponderi pristranosti jedinica se prilagođavaju pomoću sljedećih jednadžbi (Han i Kamber, 2006):

$$\Delta\Theta_j = (m)Err_j, \quad (3.18)$$
$$\Theta_j = \Theta_j + \Delta\Theta_j, \quad (3.19)$$

gdje je $\Delta\Theta_j$ promjena pristranosti jedinice Θ_j.

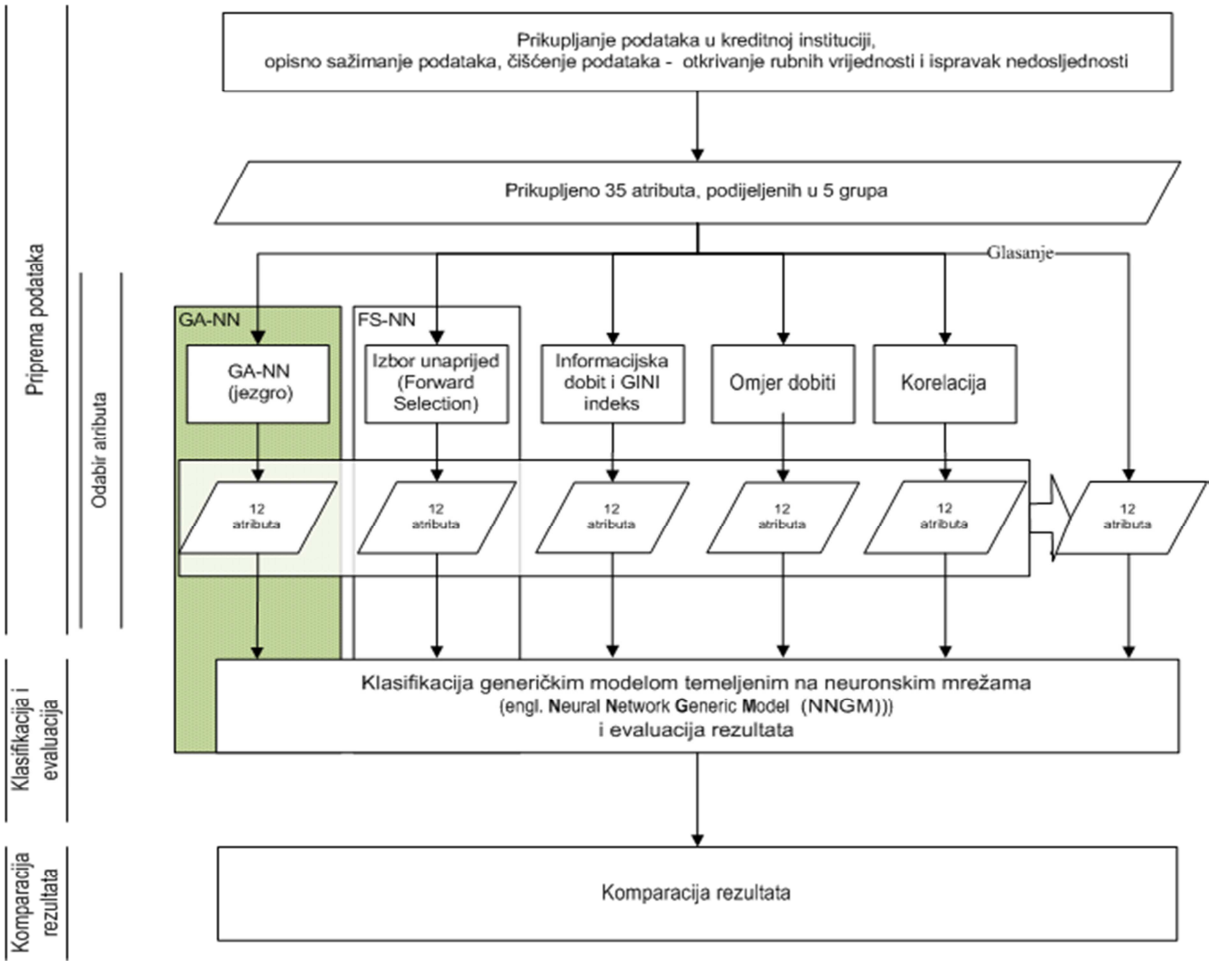
Promjena pondera (težinskih faktora) veza i pristranosti jedinica može se izvoditi nakon prezentacije svake pojedinačne n-torke, tj. nakon svakog primjera za učenje vrši se prilagodba težinskih faktora. Takvo učenje nazivamo pojedinačno učenje. Alternativno, promjene pondera veza i pristranosti jedinica može se akumulirati u varijablama tako da se promjene provode na pondere veza i pristranost jedinica nakon prezentacije svih n-torki iz skupa za učenje. Takvo učenje nazivamo skupno učenje ili učenje po epohama (engl. *epoch updating*). Naziv dolazi od toga što se jedna iteracija kroz cijeli skup za učenje naziva epoha.

Višeslojna mreža s propagacijom unaprijed s dovoljno jedinica u skrivenom sloju i dovoljnim brojem primjera za učenje može aproksimirati skoro svaku funkciju (Han i Kamber, 2006).

3.4 Razvoj modela

S najviše točke gledišta, cjelokupni proces konstrukcije modela za procjenu kreditnog rizika se sastoji od:

1) pripreme podataka s izborom atributa kao posebnim dijelom
2) klasifikacije i evaluacije
3) komparacije rezultata, kao što je prikazano na slici 3.5.



Slika 3.5 Dijagram procesa za procjenu kreditnih rizika

3.4.1 **Priprema podataka**

Nakon prikupljanja podataka u kreditnoj instituciji izvršeno je opisno sažimanje podataka. Opisno sažimanje podataka pruža analitičku podlogu za kasniju obradu podataka. U opisno sažimanje podataka uključene su osnovne statističke mjere, kao što su: srednja vrijednost te standardna devijacija i raspon vrijednosti kao korisne mjere za mjerenje disperzije podataka.

Nepotpuni, netočni i nedosljedni podatci vrlo su česta pojava u stvarnom svijetu velikih baza podataka i skladišta podataka. Postoji mnogo mogućih razloga za „šum“ u podacima, tj. za netočne i stršeće vrijednosti atributa. Stoga se rutinama za čišćenje podataka nastojalo: (1) popuniti nedostajuće vrijednosti, (2) izgladiti šum identificiranjem grubih pogreški i (3) ispraviti nedosljednosti u podacima. Čišćenje podataka se izvodilo kao iterativni proces koji se sastojao od otkrivanja različitih oblika nepodudarnosti (šuma) u podatcima i od transformacija podataka. U ovom koraku, pisane su razne skripte za pronalaženje grubih pogreški i nedosljednih vrijednosti za koje je bila potrebna istraga.

U sljedećem koraku, te vrijednosti su analizirane, od slučaja do slučaja, s nadležnim ekspertima u banci. Nakon analize i otkrivanja razloga grešaka definirana su pravila za njihovo otklanjanje, a podatci su transformirani u skladu s definiranim pravilima. Bitno je za naglasiti da su sve pogreške i nedosljednosti bile na nefinancijskim podatcima. Kao primjer takvih nedosljednosti možemo navesti sljedeće; klijent ima otvoren tekući račun u banci 25 godina, a stoji podatak da je klijent banke samo 6 godina. Razlozi greški takvog tipa su utvrđeni, a podatci ispravljeni.

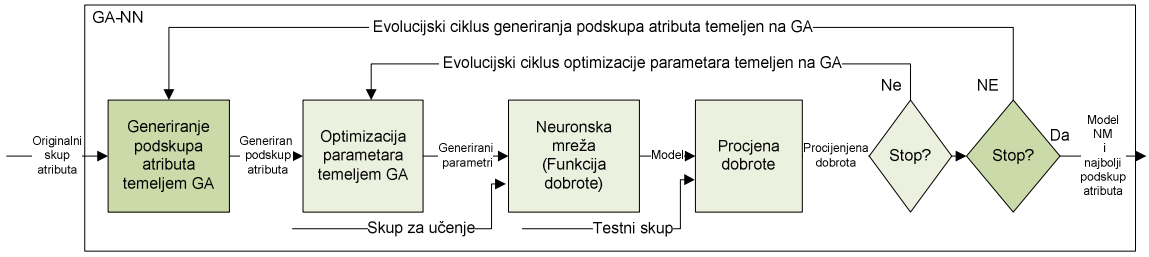
Financijske podatke banka strogo kontrolira na razini svake pojedinačne transakcije. Pored toga postoje redovne desetodnevne, mjesečne, kvartalne i godišnje kontrole. Stoga grešaka u toj vrsti podataka nije bilo. Među podatcima financijske prirode pojavile su se samo stršeće vrijednosti. Vrijednosti koje su za više od 1,5 x IQR (interkvartil raspon je definiran kao IQR = Q3 - Q1) iznad trećeg kvartila ili ispod prvom kvartila i unutar su od 2% svih vrijednosti bile su označena kao potencijalno zanimljiva odstupanja (Han i Kamber, 2006). Primjer takvog odstupanja je bio na uplatama. Analizom je utvrđeno da je riječ o prodaji dionica, koja je bila jednokratna i značajno odstupala od ostalih uplata. Takva odstupanja, za koja je utvrđeno da nisu pravilo i da je vrlo mala vjerojatnost njihova ponovnog pojavljivanja, su smatrana izuzetcima koje treba isključiti iz skupa, tj. iz uplata.

3.4.1.1 *Odabir atributa (značajki)*

Moguće je da su mnogi atributi irelevantni za klasifikacijski zadatak, ili jednostavno suvišni, redundantni. Odabir atributa se izvodi u fazi pripreme podataka radi poboljšanja učinkovitosti klasifikacijskog sustava u cjelini, s aspekta točnosti, brzine i skalabilnosti. Cilj toga postupka je pronaći, iz izvornih podataka, takav skup atributa koji će omogućiti da se izvede točna klasifikacija. U obavljanju ovog zadatka, korištene su sljedeće tehnike: Genetski algoritam, izbor prema naprijed, informacijska dobit, omjer dobiti, Ginijev indeks i korelacija, čije osnove su opisane u poglavlju o metodologiji.

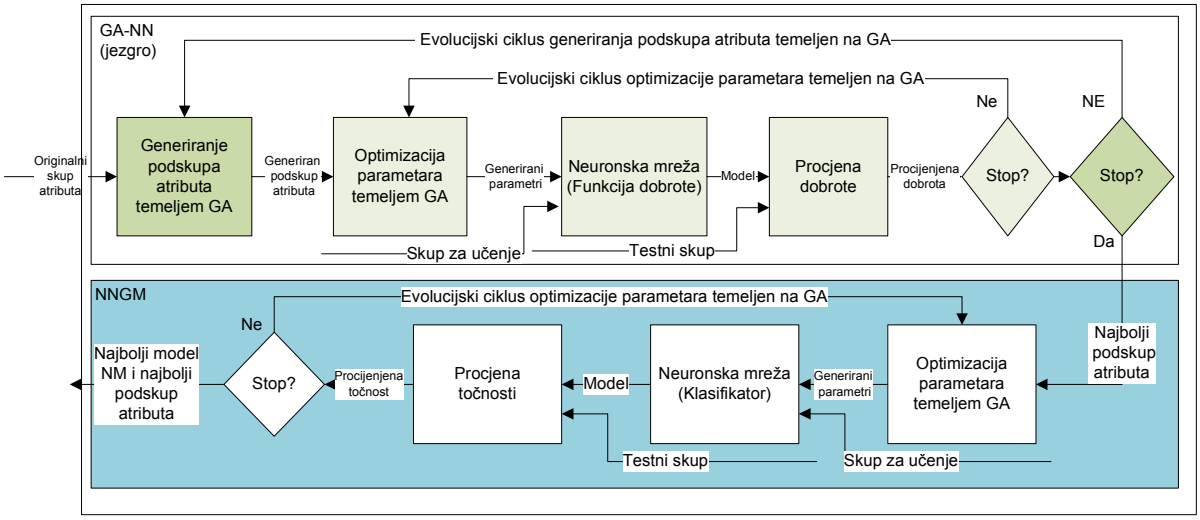
Optimalni podskup atributa ne mora biti jedinstven, jer može biti moguće postići istu točnost koristeći različiti podskup atributa (npr. kada su dva atributa u savršenoj korelaciji, jedan može biti zamijenjen drugim). Po definiciji, s aspekta dobivanja najveće moguće točnosti, optimalni podskup atributa je onaj podskup atributa koji algoritam za izbor atributa može odabrati, a s kojim se postiže najveća točnost.

Informacijska dobit, omjer dobiti, Ginijev indeks i korelacija kao tehnike za izbor atributa daju rang za svaki atribut koji opisuje dane trening *n*-torke. Njihova implementacija se temelji na jednadžbama opisanim ranije u ovom radu. Kohavi i John (1997) ove tehnike ubrajaju u filtarske tehnike za izbor podskupa atributa. Osnovna karakteristika filtarskih tehnika je da ne uzimaju u obzir pristranost klasifikacijskih algoritama, atributi su filtrirani neovisno o klasifikacijskom algoritmu. Filtarski pristup je pokušaj da se procjenjuju vrijednosti (kvalitete) atributa iz podataka, ignorirajući klasifikacijski algoritam. U radu je kreirana sofisticiranija tehnika za izbor atributa koja se temelji na genetskom algoritmu i umjetnoj neuronskoj mreži kao funkciji dobrote i pripada tehnikama omotača (engl. *wrapper approach*).



Slika 3.6 Dijagram toka GA-NN tehnike

Kao što je prikazano na slici 3.5, a detaljnije na slici 3.6, kombinirajući GA s NN, kao klasifikacijskim algoritmom, može se istovremeno obaviti zadatak izbora atributa i klasifikacija. Ipak, zbog poboljšanja točnosti kao i zbog potrebe usporedbe različitih tehnika, GA-NN tehnika je ugrađena u ukupni postupak procjene kreditnog rizika gdje je zadužena samo za izbor atributa, što je prikazano pod naslovom „Jezgro GA-NN-a“ na dijagramu toka na slici 3.7. Iz ovog dijagrama možemo vidjeti da klasifikacijski rezultati jezgre GA-NN tehnike nisu konačni, jer se u daljnjem tijeku postupka preuzimaju samo odabrani atributi kao konačni. NNGM, koji ima mnoge zajedničke karakteristike s GA-NN tehnikom, je odgovoran za optimiranje parametara konačnog klasifikatora (Slika 3.7) i za konačne rezultate klasifikacije. Na taj način je omogućena jednostavna komparacija učinkovitosti između nekih, ranije spomenutih, najčešće korištenih tehnika odabira atributa i GA-NN tehnike. Osim toga, kreirana je tehnika za odabir atributa, kodnog imena FS-NN (engl. *forward selection with neural networks*) koja također pripada skupini tehnika omotača. Algoritam za FS-NN tehniku je opisan u odjeljku o metodologiji u ovom poglavlju.



Slika 3.7 GA-NN tehnika procjene kreditnog rizika, modifikacija (Huang, Chen i Wang, 2007).

Model prikazan na slici 3.6, a detaljnije na slici 3.7, je realiziran uz pomoć alata Rapid Miner 5.1.15 uz postavke parametara prikazane u tablici 3.1. Rapid Miner je dostupan za preuzimanje (download) na adresi: http://rapidminer.com/, a izvorni kod na adresi http://sourceforge.net/projects/rapidminer/.

Tablica 3.1 Pregled parametara GA-NN tehnike

Parametar	**Postavke**
Initializacija populacije	
population size	30
initial probability for an feature to be switched on	0.7
minimum number of features	10
Reprodukcija	
Fitness measure	accuracy
Fitness function	neural network
the type of neural network	multilayer feed-forward network
network algorithm	backpropagation
activation function	sigmoid
the number of hidden layers	1
the size of the hidden layer	(number of features + number of classes) / 2 + 1
training cycles	[300;600]
learning rate	[0.3;1.0]
momentum	[0.2;0.7]
selection scheme	tournament
tournament size	0.25
dynamic selection pressure	Yes
keep best individual	Yes
mutation probability	1 / number of features
crossover probability	0.5
crossover type	shuffle
Uvjeti završetka	
maximal fitness	Infinity
maximum number of generations	6
use early stopping	Yes
generations without improvement	2

3.4.2 **Klasifikacija i evaluacija**

Učinkovitost selekcije varijabli se obično dokazuje točnošću klasifikacije. Kao klasifikator je korištena neuronska mreža zahvaljujući brojnim studijama (Zhang, Hu, Patuwo i Indro, 1999; Malhotra i Malhotra, 2003; Huang, Chen i Wang, 2007; Šušteršic, Mramor i Zupan, 2009) koje su zaključile da su neuronske mreže točnije, prilagodljivije i robusnije u odnosu na ostale klasične tehnike klasifikacije. Osim toga, modeli neuronskih mreža ne zahtijevaju zadovoljavanje pretpostavki kao što su: multivarijantna normalnost, eliminacija vršnih vrijednosti, linearno ovisnost zavisne varijable o nezavisnim varijablama, diskretne vrijednosti obilježja, jednake kovarijance i ostale pretpostavke koje su potrebne za neke druge metode.

Performanse umjetne neuronske mreže svakako ovise o topologiji mreže i o parametrima, stoga se kao najbolji vodič u većini slučajeva predlaže metoda pokušaja i pogreške (Malhotra i Malhotra, 2003). Ovaj pristup je vrlo dugotrajan, stoga se nakon mnogih pokusa predlaže pristup opisan u nastavku. Prema Li, Shiue i Huang (2006), Cybenko, Hornik, Stinchcombe i White su pokazali da je mreža s jednim skrivenim slojem dovoljna za simuliranje složenih sustava sa željenom točnošću. Stoga se predlaže struktura mreže s jednim skrivenim slojem. Broj neurona u skrivenom sloju je određen sljedećom jednadžbom:

Broj neurona u skrivenom sloju = (broj značajki + broj klasa) / 2 + 1.

Iako je u većini slučajeva, metoda pokušaja i pogreške najbolji vodič za optimiranje parametara mreže, genetski algoritam se može primijeniti za rješavanje problema parametrizacije umjetne neuronske mreže. Na taj način je kreiran generički model za optimiranje parametara umjetne neuronske mreže - NNGM. Ovaj generički model (slika 3.7) se koristi za procjenu učinkovitosti algoritama za izbor svojstava, a parametri tog modela su prikazani u tablici 3.2.

Tablica 3.2 Pregled parametara NNGM tehnike

Parametar	**Postavke**
Initializacija populacije	
population size	30
Reprodukcija	
Fitness measure	accuracy
Fitness function	neural network
the type of neural network	multilayer feed-forward network
network algorithm	back-propagation
activation function	sigmoid
the number of hidden layers	1
the size of the hidden layer	(number of features + number of classes) / 2 + 1
training cycles	[300;1000]
learning rate	[0.3;1.0]
momentum	[0.2;0.7]
selection scheme	tournament
tournament size	0.25
keep best individual	Yes
mutation type	Gaussian mutation
crossover probability	0.9
Uvjeti završetka	
maximal fitness	Infinity
maximum number of generations	10
use early stopping	Yes
generations without improvement	2

Kao što je prikazano u tablici 3.2, NNGM model se temelji na genetskom algoritmu, a točnost je mjera dobrote. Točnost klasifikacije izračunava se pomoću tehnike unakrsne validacije s 10 preklapanja. Ova tehnika procjenjuje performanse modela i testira utjecaj varijacija uzorkovanja na performanse modela. Ako se, u kreiranju modela, kandidati za kredit odabiru slučajnim odabirom (skup za učenje), mogu se izgubiti važne informacije s

obzirom na činjenicu da je postotak loših kredita obično mali u odnosu na one dobre. Za manje populacije, metoda slučajnog odabira u pravilu (Šušteršic, Mramor i Zupan, 2009) ne daje dobru distribuciju kao kod veće grupe, stoga je ta metoda odabira lošija. Iz tog razloga podskupovi podataka su kreirani tehnikom validacije s *k*-preklapanja koja koristi stratificirani uzorak. Stratificirani slučajni uzorak stvara podskupove i osigurava da je raspored klasa u podskupovima (gotovo) isti kao u cijelom skupu uzoraka.

Tehnika validacije s *k*-preklapanja je bolja od jednostavne tehnike validacije, jer jednostavna tehnika validacije dijeli skup podataka na skup s uzorcima za učenje i na skup s uzorcima za testiranje (engl. *holdout sample*) s kojima testira učinkovitost modela. S obzirom da se za najbolji model odabire onaj model koji najbolje klasificira jedan podskup, holdout, jednostavna tehnika validacije često procjenjuje pravu stopu pogreške preoptimistično (Malhotra i Malhotra, 2003). U postupku validacije s *k*-preklapanja, kreditni skup se dijeli na *k* nezavisnih skupova. Model se trenira korištenjem prvih *k*-1 skupova uzoraka, a trenirani model se testira na *k*-tom skupu. Ovaj postupak se ponavlja sve dok svaki od skupova ne bude korišten jednom kao skup za testiranje. Ukupna točnost skoring modela je prosječna točnost ostvarena kroz svih *k* skupova. Značajka validacije s *k*-preklapanja je da je model kreditnog scoringa razvijen na temelju velikog dijela svih raspoloživih podataka te da su svi podaci korišteni za testiranje konačnog modela.

3.4.3 **Komparacija rezultata**

Rezultati klasifikacije i validacije su prikazani u matrici konfuzije (ili matrici grešaka, engl. *confusion matrix,* CM), koja predstavlja koristan alat za analizu uspješnosti klasifikatora u prepoznavanju *n*-torki različitih klasa (Han i Kamber, 2006). Matrica konfuzije za dvije klase je prikazana u tablici 3.3. Za *m* klasa, matrica konfuzije je tablica veličine najmanje *m* x *m.* Pozicija $CM_{i,j}$ u prvih *m* redaka i *m* stupaca označava broj *n*-torki klase *i* koje su označene od strane klasifikatora kao klasa *j.* Klasifikator ima dobru točnost ako je većina *n*-torki prikazana na glavnoj dijagonali matrice konfuzije, tj. od pozicije $CM_{1,1}$ do pozicije $CM_{m,m}$ s ostatkom pozicija s vrijednostima blizu nule. Tablica može imati dodatne retke ili stupce za prikaz zbrojeva, preciznosti ili stopa prepoznavanja po klasama.

Tablica 3.3 Matrica konfuzije (grešaka)

		Rezultat klasifikacije		Stopa prepoznavanja (%)
		Rizičan	Nerizičan	
Stvarno stanje	Rizičan	true positives	false negatives (Pogreška tipa I)	Osjetljivost (%)
	Nerizičan	false positives (Pogreška tipa II)	true negatives	Specifičnost (%)
Ukupno				Točnost (%)

Broj primjera razvrstanih ispravno na cijelom uzorku, podijeljen s ukupnim brojem primjera iz cijelog uzorka daje točnost modela. Komparacija učinkovitosti ostalih, ranije navedenih, algoritama za izbor atributa i GA-NN algoritma biti će izvedena pomoću parnih *t*-testova. Tim testovima će se utvrditi da li se razlike procijenjenih srednjih vrijednosti točnosti mogu smatrati značajnim. Slučajni odabir uzorka, kao i njegova veličina (1000 opažanja) dobre su pretpostavke za provođenje ovog testa (Japkowicz i Shah, 2011). Osim toga, izračunati će se ukupni relativni trošak pogrešne klasifikacije (engl. *relative cost of misclassification*, RC) Prema formuli Swicegooda i Clarka (Sarlija, Bensic i Zekic-Susac, 2009):

$$RC = \alpha (P_I C_I) + (1- \alpha)(P_{II} C_{II}), \quad (3.20)$$

gdje je α vjerojatnost 'loših' klijenta, P_I je vjerojatnost pogreške tipa I, C_I je relativna cijena pogreške tipa I, P_{II} je vjerojatnost pogreške tipa II, i C_{II} je relativna cijena pogreške tipa II. RC svakog modela je izračunat za sedam scenarija, dok je najbolji model za svaki scenarij onaj s najnižom RC vrijednosti.

3.5 Empirijska analiza

Kreditni skorinzi se mogu podijeliti u dvije različite vrste. Prvi tip je skoring kod traženja (odobravanja) kredita. Zadatak tog skoringa je klasificirati kreditni zahtjev, odnosno klijenta, u "dobru" ili " lošu" rizičnu skupinu. U ovom radu, usredotočit ćemo se na ovu vrstu zadatka. Drugi tip skoringa se bavi s postojećim klijentima nakon što je klijentu već dan kredit i taj

skoring se odnosi na ponašanje klijenta, a naziva se bihevioralni skoring (Khashman, 2010). U kreditnom skoringu kod traženja potrošačkih kredita, atributi koji se obično koriste u različitim modelima uključuju: vrijeme koje je klijent proveo na sadašnjoj adresi, kućni status (vlasnik/podstanar), poštanski broj, telefon, godišnji prihod podnositelja zahtjeva, vlasništvo kreditnih kartica, vrste bankovnih računa, dob, vrsta zanimanja, svrha kredita, bračni status, dužina klijentskog odnosa s bankom, radni staž kod postojećeg poslodavca, kreditni rejting kod kreditnog ureda, redovne mjesečni obveze kao postotak redovnih mjesečnih prihoda, broj uzdržavanih članova (Sarlija, Bensic i Zekic-Susac, 2009). Prema Khashmanu (2010) podaci koji se koriste za modeliranje uglavnom se sastoje od financijskih i demografskih informacija o tražitelju kredita. Nasuprot tome, bihevioralni skoring se bavi s postojećim klijentima i uz ostale informacije, koristi informacije o povijesti plaćanja.

3.5.1 **Stvarni kreditni skup podataka**

Kreditna skup podataka za ovu disertaciju je prikupljen u jednoj hrvatskoj banci u razdoblju od rujna 2004 do rujna 2011 . U tom procesu prikupljeni su podaci o tekućim i štednim računima za 32.000 potencijalnih kandidata. Od potencijalnih kandidata u daljnje razmatranje su ušli oni slučajevi koji su uzeli kredit u iznosu manjem ili jednakom 100.000 kuna, a koji su imali otvoren tekući račun u banci najmanje 15 mjeseci prije datuma odobrenja kredita. Razdoblje od 15 mjeseci je vrijeme praćenja karakteristika tražitelja kredita (performansni period) i karakteristike iz tog razdoblja su korištene u razvoju skoring modela. Iz skupa kandidata, 1.000 slučajeva je nasumično odabrano, uključujući 750 koji su uspješno izvršili svoje kreditne obveze, odnosno koji su bili dobri kreditni klijenti, i 250 koji su kasnili u izvršavanju svojih kreditnih obveza i koji su zbog toga svrstani u skupinu loših kreditnih klijenata. Klijent je "loš" ako je kasnio s izvršavanjem svojih obveza 90 dana ili više u bilo kojem trenutku u životu kredita, što je u skladu s definicijom Baselskog sporazuma o kapitalu. Sporazum kaže da je netko loš ako 90 dana, ili u nekim zemljama 180 dana, kasni s plaćanjem svoje obveze ili ako vjerovnik pretpostavlja da klijent vjerojatno neće moći podmiriti svoju obvezu (De Andrade i Thomas, 2007) .

Iako je modeliran skoring odobravanja kredita, podaci koji se uobičajeno koriste kod odobravanja kredita ovdje su kombinirani s bihevioralnim podacima s tekućih računa kako bi se što točnije procijeniti zahtjevi za kredit i omogućio skoring bez dodatnih pitanja prema klijentu. To je razlog zašto je klijent morao imali otvoren tekući račun u banci najmanje 15

mjeseci prije datuma odobrenja kredita. Osim demografskih podataka, koje smo ranije naveli, u obzir su uzeti podaci o ponašanju klijenta kao što je povijest plaćanja, financijska pozicija, delikvencijska povijest te kreditna povijest, ako je takva postojala.

Karakteristike svakog klijenta u početku su bile opisane sa 37 varijabli. One su se odnosile na; spol i starost klijenta, svrhu kredita, iznos kredita, broj postojećih kredita u ovoj banci, kreditnu povijest s bankom prije nego što je kredit odobren, postotak raspoloživog dohotka u odnosu na redovne rate i detaljne podatke o računima i transakcijama u banci. Nakon smanjenja početnog skupa varijabli s obzirom na činjenicu da su neke varijable imale identičnu vrijednost u svim slučajevima ili izuzetno visoku korelaciju, konačan broj varijabli korištenih u istraživanju je bio 33 redovitih i 2 (id , labela) posebne.

Varijable su podijeljene u pet glavnih skupina : (i) osnovne karakteristike , (ii) povijest plaćanja (mjesečni prosjeci) , (iii) financijski uvjeti , (iv) delikvencijska povijest, i (v) prošla kreditna iskustava. Popis varijabli s njihovim obrazloženjem i deskriptivnom statistikom za podatke iz uzorka za razvoj modela dan je u dodatku A.

3.5.2 **Rezultati eksperimenta**

Testirano je sedam tehnika za izbor atributa: Genetski algoritam s neuronskim mrežama (GA-NN), tehnika odabira najboljeg atributa za slijedeći korak s neuronskim mrežama (FS-NN), informacijska dobit, omjer dobiti, Ginijev indeks, korelacija i glasovanje (atributi koji se preklapaju). Algoritam za informacijsku dobit i Ginijev indeks su dali iste rezultate stoga rezultati algoritma za informacijsku dobit nisu posebno navedeni u tablici s rezultatima. Pomoću navedenih tehnika odabrano je i prikazano u dodatku B prvih 12 atributa za svaku tehniku. Odabrano je 12 najboljih atributa zbog toga što je procijenjena točnost počela padati nakon što se smanjio broj atributa ispod 12. Nakon što su izabrani atributi za svaku tehniku i tehnikom glasanja su mogli biti izabrani najznačajniji atributi. Za potrebe ove tehnike relevantni su atributi koji se pojavljuju u više od polovice drugim tehnika za izbor atributa. Iz Dodatka B se može vidjeti da su za sve tehnike značajni atributi: RIDI, BCO i OINT. Značajni atribut za sve tehnike osim jedne je bio TWB, a atributi značajni za tri od pet tehnika su: ACAGE, LMM, TOUT, TITO, RII, BAL, INPO i CHD. Tih 12 atributa su najvažniji atributi u drugim tehnikama i oni predstavljaju atribute tehnike glasanja. Kako bi procijenili učinkovitost spomenutih tehnika selekcije atributa, s atributima svake tehnike je izvršena klasifikacija pomoću NNGM tehnike.

Broj ispravno klasificiranih objekata na cijelom uzorku također ovisi i o odabranoj graničnoj vrijednosti (engl. *cutoff value*). Ako je npr. cilj klasifikacije opisan s dvije vrijednosti - nula i jedan, a model vraća vrijednosti između nula i jedan, tada bi točnost trebala biti najviša na graničnoj vrijednosti 0,5 (Šušteršic, Mramor i Zupan, 2009). No, na toj graničnoj vrijednost troškovi pogrešne klasifikacije modela nisu nužno optimalni. To ovisi o odnosu troškova banke nastalim uslijed odobravanja kredita lošem klijentu (pogreška tipa I) i neodobravanja kredita (oportunitetni trošak) dobrom klijentu (pogrešaka tip II). Iz tog razloga su u ovom istraživanju za svaku tehniku prikazane procjene točnosti za dvije granične vrijednosti. Prvo, za graničnu vrijednost (prag za odluku davanju ili odbijanju kredita) u blizini 0,5 što se pokazalo, u smislu troškova, kao bolje nego 0,5, jer je pogreška tipa I znatno smanjena u odnosu na povećanje pogreške tipa II. Istodobno se s povećanim pragom gotovo nije promijenila prosječna točnost predviđanja modela. Za svaki odabrani skup atributa je napravljena i druga procjena, s takvom graničnom vrijednošću kojoj je bio cilj smanjiti grešku tipa I na vrijednost manju od 25%. Rezultati druge procjene za odabrane skupove atributa su također prikazani u tablicama. Kao što je već spomenuto, troškovno, u nekim okolnostima pogreška modela može biti optimalna pri manjoj ukupnoj točnosti modela, a ne pri najvišoj. To ovisi o omjeru između pogreške tipa I i pogreške tipa II, koji dalje ovisi o gospodarskim ciklusima, okolnostima unutar i oko banke kao i preferencijama banke. Rezultati su testirani na cijelom uzorku podataka pomoću opisane unakrsne validacije s *k* preklapanja

Da bi se izbjeglo mnoštvo tablica, pomoću matrice grešaka prikazani su rezultati samo za dvije ekstremne tehnike. Kao što je prikazano u tablici 3.4, ukupna prosječna stopa točnosti klasifikacije dobivena temeljem atributa odabranih pomoću GA-NN tehnike je 82,30% , granična vrijednost je 0,60. To je puno bolje u odnosu na ukupnu prosječnu stopu točnosti klasifikacije za isti skup atributa s graničnom vrijednošću 0,77, koja iznosi 73,90%. Pogoršanje točnosti je očekivano, s obzirom da je cilj bio podizanje granične vrijednosti radi smanjenja greške tipa I, što se i dogodilo. Greška tipa I je smanjena s 44,00% na 23,20%. Pogoršanje točnosti javlja se zbog činjenice što je u isto vrijeme pogreška tipa II značajno porasla s 8,93% na 27,07%. Je li cijena smanjenja pogreške tipa I previsoka? Odgovor na ovo pitanje može se dobiti izračunom troška pogrešne klasifikacije. U svakom slučaju, iz tablice 3.4 se može vidjeti da je poboljšanje od 52 lažna negativa slučaja u točne pozitivne plaćeno s pogoršanjem od 136 slučaja, s točnih negativnih u lažne pozitivne. Zanimljivi rezultati su dobiveni korištenjem atributa tehnike omjer dobiti. U usporedbi s drugim tehnikama, maksimalna točnost je relativno loša, samo 79,20%, međutim tehnika daje vrlo dobre

rezultate od 73,60%, pod uvjetom da je pogreška tipa I manja od 25%. To je zbog činjenice što je poboljšanje od 44 lažna negativa slučaja u točne pozitivne plaćeno, kao što je prikazano u tablici 3.5, rezultiralo pogoršanjem relativno manjeg broja slučajeva, 100, s točnih negativnih u lažne pozitivne. Slične promjene su se dogodile i s rezultatima kod drugih tehnika. Konsolidirani rezultati svih tehnika prikazani su u tablicama 3.6 i 3.7.

Tablica 3.4 Rezultati klasifikacije NNGM tehnike za atribute odabrane GA-NN tehnikom

		Rezultat klasifikacije		Stopa prepoznavanja (%)
		Rizičan	Nerizičan	
Granična vrijednost *(0.60)*				
Stvarno stanje	Rizičan	140	110	56.00
	Nerizičan	67	683	91.07
Točnost (%)				82.30
Granična vrijednost *(0.77)*				
Stvarno stanje	Rizičan	192	58	76.80
	Nerizičan	203	547	72.93
Točnost (%)				73.90

Tablica 3.5 Rezultati klasifikacije NNGM tehnike za atribute odabrane tehnikom omjera dobiti

		Rezultat kasifikacije		Stopa prepoznavanja (%)
		Rizičan	Nerizičan	
Granična vrijednost *(0.59)*				
Stvarno Stanje	Rizičan	147	103	58.80
	Nerizičan	105	645	86.00
Točnost (%)				79.20
Granična vrijednost *(0.76)*				
Stvarno Stanje	Rizičan	191	59	76.40
	Nerizičan	205	545	72.67
Točnost (%)				73.60

3.5.3 **Komparacija rezultata**

U istraživanju su korištene različite tehnike odabira atributa ali se za klasifikaciju i validaciju koristi ista tehnika za sve tehnike odabira atributa, kako bi bila podržana jednostavna izravna komparacija rezultata. S ciljem procjenjivanja učinkovitosti tehnika za izbor atributa u odnosu na GA-NN tehniku, uspoređujemo: (1) točnost tehnika, kao što je prethodno opisano, za dvije odabrane granične vrijednosti i (2) troškove klasifikacije kako bi pronašli model koji za banke najviše smanjuje troškove.

Za potrebe usporedbe točnosti tehnika, predviđanje s maksimalnom ukupnom prosječnom stopom točnosti je najtočnije predviđanje. Tablica 3.6 pokazuje da je GA-NN tehnika generirala najtočnije predviđanje s ukupnom prosječnom stopom točnosti od 82,30%, sa standardnom devijacijom od 1,85%. Na temelju parnog *t*-testa, prikazanog u tablici 3.8., u prosjeku, ukupna točnost GA-NN tehnike je bolja od ukupne prosječne točnosti svih drugih tehnika osim FS-NN, a razlika je statistički značajna s 95%-tnom razinom pouzdanosti u korist GA-NN tehnike. U slučaju FS-NN tehnike razlika nije statistički značajna.

Tablica 3.6 Komparacija rezultata maksimalnih prosječnih točnosti svih tehnika

Tehnika odabira atributa	Stopa prepoznavanja (%)		Prosječna točnost (%)	Std. dev. (%)
	Rizičan	Nerizičan		
GA-NN	56,00	91,07	82,30	1.85
Odnos dobiti	58,80	86,00	79,20	3.94
Gini indeks	59,20	84,67	78,30	3.20
Korelacija	59,20	87,07	80,10	2.55
FS-NN	56,80	88,00	80,20	4.83
Glasanje	60,80	86,67	80,20	2.23

Tablica 3.7 Komparacija rezultata svih tehnika uz graničnu vrijednost prilagođenu na grešku tipa I < 25%

Tehnika odabira atributa	Stopa prepoznavanja (%)		Prosječna točnost (%)	Std. dev. (%)
	Rizičan	Nerizičan		
GA-NN	76.80	72.93	73.90	3.83
Odnos dobiti	76.40	72.67	73.60	4.03
Gini indeks	76.00	70.53	71.90	4.50
Korelacija	75.60	72.00	72.90	5.45
FS-NN	76.40	71.20	72.50	5.97
Glasanje	76.00	75.20	75.40	4.92

Tablica 3.8. Zavisni (parni) t-test

Tehnika odabira atributa	Rezultati prosjek +/- std.dev.	p-vrijednost*
GA-NN	0.823 +/- 0.018	--
Odnos dobiti	0.792 +/- 0.039	0.042
Ginijev indeks	0.783 +/- 0.032	0.003
Korelacija	0.801 +/- 0.025	0.046
FS-NN	0.802 +/- 0.048	0.288
Glasanje	0.802 +/- 0.022	0.038

* alpha=0.050

Iz tablice 3.6 i 3.8 se može vidjeti da GA-NN tehnika postiže najbolje rezultate u odnosu na ukupnu prosječnu stopu točnosti. Pomicanjem granične točke (engl. *threshold point*) na vrijednosti gdje je pogreška tipa I <0,25 (Tablica 3.7), ne dobivamo tako dobre rezultate GA-NN tehnike u usporedbi s drugim tehnikama. Ovo je sasvim očekivano s obzirom da je optimizacija kod izbora značajki pomoću GA-NN tehnike provedena u odnosu na točnost, bez ikakvih dodatnih uvjeta. To je klasičan primjer u kojem, ukupno gledajući, najbolji skup atributa odabranih po jednoj vrsti uvjeta ne nadmašuje neki drugi skup atributa pod drugim uvjetima, u našem primjeru za neku drugu graničnu vrijednost. Ako je opisani postupak (algoritam) mogao pronaći najbolji skup atributa za određene uvjete, opravdano je pretpostaviti da će i za promijenjene uvjete opisani postupak pronaći najbolji skup atributa za te nove uvjete. To je zato što atributi, izabrani kao najbolja kombinacija za određene uvjete, ne moraju zadovoljiti uvjet optimalnosti na nekim drugim uvjetima. Ako je to istina da određeni atributi za različite uvjete daju rezultate različite kvalitete, a što je vidljivo u našem slučaju, onda tehnike omotača imaju prednost pred filtarskim tehnikama, jer tehnike omotača daju drugačiju kombinaciju atributa za različite ciljne uvjete, što nije slučaj s tehnikama filtra, one daju isti rang za određene atribute, bez obzira na metodu klasifikacije ili dodatne uvjete klasifikacije.

Kako bi tehnike omotača iskoristile svoje potencijalne prednosti u odnosu na filtarske tehnike, tehnike omotača podrazumijevaju obvezu ponovnog pronalaženja optimalne kombinacije atributa za svaku promjenu ciljne funkcije. To predstavlja dodatni napor i trošak koji će biti naplaćen nižim troškovima pogrešne klasifikacije.

Kada pogledamo rezultate samo u smislu GA-NN tehnike i našeg primjera, za svaku usporedbu u kojoj se mijenjaju uvjeti i u kojoj želimo dobiti najbolje performanse iz

odabranih atributa, potrebno je provesti novi proces optimizacije pomoću GA-NN tehnike. Potencijalno za bilo koje nove uvjete može se dobiti novi skup atributa koji najbolje odgovara zahtjevima optimizacije. Kako je u ovom primjeru provedena optimizacija pomoću GA-NN tehnike samo sa stajališta maksimalne točnosti, logično je da atributi dobiveni u tom procesu daju najbolje rezultate samo sa stajališta maksimalne točnosti, a ne za neke druge uvjete. Općenito, GA-NN tehnika daje najbolje rezultate za funkciju cilja za koju je optimirana, a tehnika glasovanja je pokazala najbolju stabilnost što je u skladu s nalazima Schowe (2011). Isto tako se pokazalo da je relativno dugo vrijeme izvođenja najveće ograničenje GA-NN tehnike kao tehnike za odabir atributa; GA-NN je računski intenzivna tehnika.

Tablica 3.9 Komparacija troškova

Tehnika odabira atributa	Odnos troškova (C_I:C_{II})						
	1:1	2:1	3:1	4:1	5:1	8:1	10:1
Rezultati uz maksimalnu točnost							
GA-NN[1]	**0,1770***	**0,2870***	0,3970	0,5070	0,6170	0,9470	1,1670
Odnos dobiti	0,2080	0,3110	0,4140	0,5170	0,6200	0,9290	1,1350
Ginijev indeks	0,2170	0,3190	0,4210	0,5230	0,6250	0,9310	1,1350
Korelacija	0,1990	0,3010	0,4030	0,5050	0,6070	0,9130	1,1170
FS-NN	0,1980	0,3060	0,4140	0,5220	0,6300	0,9540	1,1700
Glasanje [1]	0,1980	0,2960	0,3940	0,4920	0,5900	0,8840	1,0800
Rezultati uz graničnu vrijednost (cutoff) prilagođenu na grešku tipa I < 25%							
GA-NN[2]	0,2610	0,3190	0,3770	0,4350	0,4930	0,6670	**0,7830***
Odnos dobiti	0,2640	0,3230	0,3820	0,4410	0,5000	0,6770	0,7950
Ginijev indeks	0,2810	0,3410	0,4010	0,4610	0,5210	0,7010	0,8210
Korelacija	0,2710	0,3320	0,3930	0,4540	0,5150	0,6980	0,8200
FS-NN	0,2750	0,3340	0,3930	0,4520	0,5110	0,6880	0,8060
Glasanje [2]	0,2460	0,3060	**0,3660***	**0,4260***	**0,4860***	**0,6660***	0,7860

U skladu s jednadžbom (3.20), ukupni relativni trošak pogrešne klasifikacije (RC) je izračunat za svaki model, za sedam scenarija, kod čega je najbolji model za svaki scenarij model s najnižom RC vrijednosti. Iz tablice 3.9 se može vidjeti da će najtočnije predviđanje biti najprihvatljivije za banke u slučaju da je trošak netočnog predviđanja koje lošeg klijenta proglašava dobrim (pogreška tipa I) jednak trošku netočnog predviđanja kod kojeg je dobar klijent proglašen lošim (tip II pogreške), i kada je omjer troškova 2:1 (C_I:C_{II}). Kao što se može vidjeti, za druge scenarije najtočnije predviđanje (GA-NN[1]) ne daje najnižu RC vrijednost. Ostali scenariji su vjerojatniji za banku. Tehnika Glasanja[2] daje najniže vrijednosti

za omjere troškova (pogreška tipa I / pogreška tipa II) 3:1, 4:1, 5:1 i 8:1, a GA-NN^2 tehnika za omjer 10:1. Opravdano je očekivati da banke neće optimirati izbor atributa temeljem funkcije točnosti, nego će vjerojatno htjeti optimirati troškovnu funkciju za neke sebi svojstvene omjere. To ovdje nije učinjeno, eksperiment za optimizaciju troškovne funkcije je izveden i prezentiran u poglavlju 5.

3.6 Zaključci poglavlja

Pri izradi modela za klasifikaciju klijenata banke iz ukupnog skupa podataka kojima raspolaže banka moraju se izabrati oni atributi koji su korisni u razvrstavanju klijenata, a oni koji su suvišni, oni koji unose šum u sustav, moraju se izostaviti. Teoretski, čak i ako prilikom klasifikacije nije postignuta bolja klasifikacijska točnost, postoje mnoge potencijalne prednosti koje se dobivaju odabirom atributa, kao: olakšavanje vizualizacije podataka, bolje razumijevanje podataka, ubrzavanje prikupljanja i obrade podataka kao i smanjenje troškova u fazi eksploatacije modela.

Za odabir atributa korišteno je nekoliko standardnih tehnika. Osim njih, kreirana je nova vlastita tehnika za izbor atributa i klasifikaciju klijenata, to je hibrid genetskih algoritama i neuronskih mreža. GA-NN tehnika istovremeno vrši selekciju atributa kao i selekciju parametara umjetnih neuronskih mreža i traži optimalno rješenje na oba područja. Onaj skup atributa koji daje najbolju točnost klasifikacije se odabire kao optimalan skup atributa. Za potrebe nedvosmislene usporedbe s drugim tehnikama odabira atributa, daljnji postupak je proveden na odabranom skupu atributa kao i za ostale tehnike. Klasifikacijski rezultati koji su postignuti GA-NN tehnikom, nisu uzeti kao konačni zbog potrebe daljnjeg postupka. Konačni rezultati klasifikacije potrebni za usporedbu s drugim tehnikama za odabir atributa su dobiveni stavljanjem tih atributa na ulaz NNGM tehnike, kao i kod ostalih tehnika za odabir atributa. Klasifikacijski rezultati za sve tehnike za izbor atributa dobiveni su na isti način, na taj način je osigurana njihova jednoznačna usporedivost.

Iz eksperimentalnih rezultata smo zaključili da je GA-NN tehnika značajno bolja pri izboru atributa za klasifikaciju u usporedbi s drugim tehnikama koje se često koriste za odabir atributa. Dobiveni rezultati potvrđuju hipotezu H3.1. Isti rezultati potvrđuju da se na temelju ukupnih podataka koje banka ima o svojim klijentima može klasificirati klijente u smislu njihove kreditne rizičnosti uz maksimalnu točnost iznad 80%, a što je preciznije od rezultata

pronađenih u literaturi o ovoj temi (Crook, Edelman i Lyn, 2007; Šušteršic, Mramor i Zupan, 2009; Zekić-Sušac, Šarlija i Benšć, 2004; Zekic-Sušac, Benšić i Šarlija, 2005).

Sve navedeno daje banci mogućnost da kreira takve proizvode, koji su istodobno u skladu s propisima s jedne strane, ali još značajnije, proizvode koji su konkurentni u odnosu na druge banke. Konkurencija prisiljava menadžment banke da traži nova rješenja za svoje poslovanje. Ta rješenja u isto vrijeme moraju imati veću fleksibilnost i osjetljivost prema riziku. Stoga se istraživanje prezentirano u ovom poglavlju može promatrati kao jedan korak prema većoj fleksibilnosti u procesu odobravanja kredita ali i prema većoj osjetljivosti na rizik. Pronalaženje mogućnosti da se procjeni kreditna sposobnost klijenta bez fizičke prisutnosti u banci, bez dolazaka u banku, temeljem čvrstih podataka kojima banka raspolaže o klijentu je korak u tom smjeru.

Premda je GA-NN tehnika statistički značajno bolja pri izboru atributa za klasifikaciju u usporedbi s drugim tehnikama, istraživanje ćemo nastaviti s ciljem daljeg poboljšanja tehnike. S obzirom da se kao glavni nedostatak tehnike nameće vrijeme izvršavanja, u nastavku se istražuju mogućnosti za ublažavanje tog nedostatka, a rješenje se traži u kombinaciji s nekim drugim tehnikama.

POGLAVLJE 4

TEHNIKA ZA ODABIR ATRIBUTA I PROCJENU RIZIKA TEMELJENA NA HIBRIDNIM GENETSKIM ALGORITMIMA I NEURONSKIM MREŽAMA

4.1 Uvod

Kreditni rizik jedan je od najvažnijih i sve izraženijih problema bankarske industrije i zbog toga je ocjenjivanje kreditnog rizika dobilo na važnosti tijekom nekoliko proteklih godina (Akkoç, 2012; Danenas et al., 2011; Finlay, 2011; Tsai, Lin, Cheng i Lin, 2009). Do prije par godina rijetka su bila istraživanja o ocjeni kreditnog rizika kod potrošačkih kredita. Kvantitativni modeli ocjene kreditnog rizika potrošača (građana) su razvijeni puno kasnije nego oni za kredite poslovnih subjekata, većinom zbog problema dostupnosti podataka. Podaci su bili ograničeni isključivo na baze podataka financijskih institucija. Trenutno postoje javno dostupni podaci iz nekoliko zemalja (Njemačka, Australija), a financijske institucije i istraživači su razvili različite kvantitativne tehnike ocjene kreditnog rizika (Šušteršic, Mramor i Zupan, 2009). Ipak, nema standardnog skupa atributa ili indikatora koji bi postojali u svim kreditnim institucijama i na temelju kojih bi se provodila klasifikacija građana u svrhu određivanja njihove kreditne sposobnosti. Zbog toga je nužno upotrebljavati sve dostupne podatke i informacije, metode i algoritme za selekciju atributa i preciznu klasifikaciju klijenata.

Banke su, pritisnute krizom i slabim vraćanjem kredita od zajmoprimaca, otkrile razinu rizika ugrađenu u bankarsko poslovanje. Prilikom kreditiranja građanstva, rizici su često preuzimani bez precizne procjene njihovog stupnja i mogućih posljedica. Veliki broj odluka koje se donose u kreditiranju građana, nužno nameću potrebu oslanjanja na modele i algoritme umjesto na ljudsku diskreciju te baziranje takvih algoritamskih odluka na "tvrdim" informacijama (Khandani, Kim i Lo, 2010). Tijekom visokog rasta ekonomije, banke su svojim vlasnicima donosile visoke razine profita i nisu bile izložene izazovima pronalaska novih načina poslovanja. Primorane krizom, banke su izložene izazovima pronalaska novih načina poslovanja koji moraju biti manje rizični, a efikasniji i profitabilniji.

Kriza je značajno smanjila razinu profita; mnoge banke su zapale u probleme, a neke su i bankrotirale. Ulagači su postali oprezniji i ne žele ulagati svoj kapital u takve problematične banke, a regulatori zahtijevaju od banaka jačanje njihovog kapitala (BIS, 2011), radi povećanja njihove otpornosti prema takvim izvanrednim okolnostima. Vlade mnogih država i međunarodne organizacije su uključene u spašavanje situacije i sprečavanje još većih posljedica krize.

Kriza je bila snažna, a jedan od osnovnih razloga krize je bio način na koji su banke funkcionirale. Malo je preciznih metoda upotrebljavano u ocjenjivanju kreditnog rizika, a uzimanje kolaterala za posuđena sredstva se upotrebljavalo kao glavni surogat (ne)ocjenjivanju kreditnog rizika. Kada je postalo očito da taj kolateral ne vrijedi onoliko kolika mu je procijenjena vrijednost i da kreditni rizik klijenata nije na odgovarajući način procijenjen, gubici su postali neizbježni. Zbog toga treba raditi na svim uzrocima krize uključujući i promjenu načina na koji banke posluju, prvenstveno što se tiče rizika prilikom posuđivanja sredstava. Poslovanje treba biti brže, manje rizično, egzaktnije i bazirano na podacima. Banke trebaju upotrebljavati ukupni raspoloživi kapital na bolji način. Taj kapital ne odnosi se samo na novac, nego i na prikupljene podatke u bazama podataka. S kapitalom u vidu podataka o klijentima, banke trebaju bolje upravljati, te ga transformirati u znanje i naposljetku u novac.

Podaci u bazama podataka mogu se upotrebljavati za procjenu kreditnog rizika, ali ti podaci su najčešće visoko dimenzionalni. Nebitni atributi unutar skupa podataka za učenje mogu dovesti do manje preciznih rezultata u klasifikacijskoj analizi (Lavrac, Gamberger i Turney, 1996). Selekcija atributa je potrebna da bi se izdvojili najbitniji atributi za povećanje brzine i preciznosti predviđanja, kao i skalabilnosti same konstrukcije modela. U problemima teškim za strojno učenje, kao što je ocjena kreditnog rizika, upotreba prikladnog skupa atributa je kritična za uspjeh procesa učenja i zbog toga je selekcija atributa, sama po sebi,

važan problem. Stoga u ovom poglavlju doktorskog rada istražujemo kombinacije komponenti iz različitih tehnika s ciljem postizanja veće preciznosti, brzine i skalabilnosti algoritma za odabir atributa te omogućavanja inkrementalnog odabira atributa. U nastavku je predstavljen novi i efikasan hibridni klasifikator.

Istraživanje je usmjereno na genetske algoritme (GA) i mogućnosti za njihovo poboljšanje. Genetski algoritam je vremenski zahtjevan i troši više procesorskog vremena od bilo koje druge metode optimiranja (Golub, 2001). Poboljšanje genetskog algoritma prezentirano u nastavku obuhvaća prevenciju trošenja vremena na istraživanje nebitnih područja istraživačkog prostora. Zbog toga, tema ovog poglavlja doktorskog rada je kreiranje naprednog, heurističkog, inkrementalno iterativnog algoritma i to hibridizacijom genetskog algoritma s nekim filtarskim tehnikama. Novi klasifikator, nazvan HGA-NN, sastavljen je od brzih filtarskih tehnika, hibridnog genetskog algoritma (HGA) i umjetne neuronske mreže. Istraživanje je provedeno na rješavanju problema selekcije atributa i klasifikacije prilikom ocjene kreditnog rizika građana.

Preostali dijelovi ovog poglavlja su organizirani na sljedeći način. Odjeljak 4.2 detaljnije opisuje problem selekcije atributa s naglaskom na primijenjene optimizacijske tehnike, predmet proučavanja u ovom poglavlju, te daje pregled literature povezane s tim problemom. Odjeljak 4.3 opisuje eksperimentalni dizajn i razvoj modela. Odjeljak 4.4 opisuje rezultate eksperimenata provedenih na dva stvarna skupa kreditnih podataka, jednom prikupljenom u hrvatskoj banci i drugom preuzetom iz UCI baze podataka. Također je dana evaluacija performansi i njihova komparacija s rezultatima prezentiranim u literaturi. Odjeljak 4.5 zaključuje ovo poglavlje te daje smjernice za dalji rad.

4.2 Opis problema i pregled literature

Selekcija atributa je tehnika pripreme podataka koja se najčešće koristi na visoko dimenzionalnim podacima, a njezina svrha uključuje; smanjenje broja dimenzija kroz uklanjanje nevažnih i redundantnih atributa, olakšavanje razumijevanja podataka, smanjenje količine podataka potrebnih za učenje, poboljšanje točnosti predviđanja algoritama i povećanje mogućnosti interpretacije modela (Oreski, Oreski, & Oreski, 2012). Selekcija atributa je problem izbora podskupa atributa koji je nužan ali ujedno i dovoljan za opis ciljnog koncepta (Kira & Rendell, 1992). Kada je selekcija atributa loše provedena, to može dovesti, između ostalog, do niza problema vezanih uz atribute s nepotpunim informacijama, nejasnim ili nebitnim informacijama, odnosno, do formiranja lošeg skupa atributa. Algoritam

učenja, koji se koristi u tom slučaju, nepotrebno se usporava zbog velikog broja dimenzija prostora pretraživanja, dok istodobno dolazi do manje klasifikacijske preciznosti zbog informacija nepotrebnih za učenje.

Problem selekcije *m* atributa iz skupa od *n* atributa može se riješiti različitim algoritmima. Iz perspektive procesorskog vremena potrebnog za rješavanje problema, računalna kompleksnost problema je $\binom{n}{m}$ i pripada klasi NP problema. Za veće dimenzije, takvi se problemi ne mogu rješavati iscrpnim pretragama ili jednostavnim heuristikama. U zadnjih nekoliko godina predloženi su razni algoritmi (tehnike) selekcije atributa. Neki od njih bit će spomenuti u nastavku.

Aha i Bankert (1996) izvještavaju o pozitivnim empirijskim rezultatima algoritma za sekvencijalni odabir atributa unaprijed i unazad. Oni su pokazali da selekcija atributa poboljšava performanse klasifikatora i dali empirijski dokaz da tehnike omotača nadmašuju filtarske tehnike. Danenas et al. (2011) su primijenili selekciju atributa na skupovima podataka koristeći algoritam selekcije podskupa atributa temeljen na korelaciji i Tabu pretraživanju u podskupu atributa. Jin et al. (2012) su predložili mjeru važnosti atributa i metodu izbora atributa temeljenu na rangiranju. U predloženoj metodi selekcije atributa, ulazno izlazna korelacija je korištena za računanje važnosti atributa, a potom su atributi sortirani padajućim redoslijedom. Predložen je i hibridni algoritam koji kombinira neuronsku mrežu s propagacijom greške unatrag (engl. *Back Propagation Neural Network* - BPNN) i optimizaciju rojem čestica (engl. *Particle Swarm Optimisation* - PSO). PSO je korišten za optimizaciju pondera i praga BPNN-a te za savladavanje nedostataka svojstvenih BPNN metodi. Njihovi eksperimentalni rezultati pokazuju da je predložena metoda selekcije atributa efikasna tehnika pripreme podataka.

Piramuthu (2006) razmatra alate za pomoć pri donošenju odluka iz perspektive strojnog učenja kod ocjenjivanja kreditnog rizika. Pritom analizira nekoliko načina poboljšanja performansi tih alata kroz pripremu podataka, posebno kroz odabir i konstrukciju atributa. Piramuthu jednostavno zaključuje da bi se dobili bolji rezultati u ocjenjivanja kreditnog rizika, u obzir treba uzeti podatke i/ili karakteristike problema isto kao i prikladnost odabranog algoritma problemu. Učinak, u tom kontekstu, ovisi o barem dva različita entiteta: o algoritmu i o skupu podataka.

Sve tehnike selekcije atributa se mogu razvrstati u 3 grupe: filtarske, tehnike omotača i hibridne tehnike. Filtarske tehnike se oslanjanju na opće karakteristike podataka te procjenjuju i izabiru podskupove atributa bez uključivanja klasifikacijskih algoritama. Jedna

prednost filtarskih tehnika je ta što su općenito brze jer ne upotrebljavaju klasifikacijske algoritme, a zbog toga su i prikladne za upotrebu na velikim skupovima podataka. Dodatno k tome, lako su uporabljive s raznim klasifikacijskim algoritmima. Tehnike omotača prvo koriste optimizacijski algoritam koji dodaje ili uklanja atribute da bi dobio različite podskupove atributa, a onda koriste klasifikacijski algoritam koji procjenjuje taj podskup atributa. Tehnike omotača su najčešće točnije od filtarskih tehnika, ali su i zahtjevnije za računalnu obradu. Zbog toga što se često poziva klasifikacijski algoritam, tehnike omotača su sporije od filtarskih i ne skaliraju dobro na velikim, visoko dimenzionalnim skupovima podataka. Hibridne tehnike pokušavaju unaprijediti filtarske i tehnike omotača iskorištavajući njihove komplementarne snage (Jin et al., 2012).

Hibridne tehnike su najčešće kombinacije filtarskih i tehnika omotača, dizajnirane tako da spajaju preciznost s brzinom izvršavanja na način da primjene tehnike omotača na samo one podskupove koji su pred-selektirani filtarskim tehnikama (Jin et al., 2012). Strategije korištene za pretraživanje prostora rješenja u hibridnim tehnikama su jako različite. Zbog vremenske kompleksnosti problema, često se koriste meta-heuristike. Jedna od meta-heuristika je GA. Prednost GA u odnosu na druge algoritme pretraživanja je mogućnost prihvaćanja većeg broja strategija za pronalaženje dobrih pojedinaca koji se tada dodaju u skup za uparivanje unutar GA okvira (engl. *framework*), i to kako u fazi kreacije inicijalne populacije tako i u fazi dinamičkog generiranja nove populacije (Pezzella, Morganti & Ciaschetti, 2007). U novije vrijeme su predložene razne varijante GA-a.

Yang, Li i Zhu (2011) opisuju poboljšani genetski algoritam za optimalnu selekciju podskupa atributa iz skupa atributa koji se sastoji iz više grupa atributa s kompletno različitim simbolima, abecedom (engl. *multi-character feature set*, MCFS). Oni dijele kromosome na nekoliko dijelova ovisno o broju grupa atributa u MCFS-u. Upotrebljavaju operator segmentiranog križanja i operator segmentirane mutacije za rad na tim segmentima kako bi izbjegli pogrešne kromosome. Vjerojatnost križanja i mutacije je dinamički prilagođena prema broju generacije i vrijednosti funkcije dobrote. Kao rezultat dobivaju snažne pretraživačke sposobnosti na početku evolucije i postižu ubrzanu konvergenciju tijekom evolucije.

Li et al. (2011) usredotočuju se na strategije generiranja inicijalne populacije genetskog algoritma i ispituju utjecaj takvih strategija na cjelokupne rezultate GA, gledano kroz kvalitetu rješenja i vrijeme izvršavanja. Njihov pristup inicijalno upotrebljava pohlepni algoritam koji brzo generira visoko kvalitetno rješenje s niskom računalnom kompleksnošću. Nakon toga upotrebljavaju to rješenje kao temelj za stvaranje skupa rješenja kao početne

populacije GA, koja se onda upotrebljava u hibridnom GA za ispitivanje učinaka predloženog pristupa. Da bi poboljšali kvalitetu početnog rješenja, Zhang et al. (2011) su dizajnirali novu metodu inicijalnog pridruživanja koja generira visoko kvalitetnu inicijalnu populaciju koja integrira različite strategije za poboljšanje brzine konvergencije i kvalitete konačnih rješenja. Maaranen et al. (2004) proučavaju upotrebu kvazi-slučajnih sekvenci u inicijalnoj populaciji genetskog algoritma. Oni zaključuju da različite inicijalne populacije genetskih algoritama imaju utjecaj na konačnu vrijednost funkcije cilja i na ukupan broj korištenih generacija.

Međutim, **ne postoji istraživanje koje bi generiralo inicijalnu populaciju genetskog algoritma na temelju znanja domenskih eksperata kao i na temelju rezultata filtarskih tehnika** za odabir podskupa atributa, kao što su: informacijska dobit (IG), omjer dobiti, Ginijev indeks i korelacija, a da ostatak inicijalne populacije generira slučajnim odabirom. Sukladno tim zapažanjima, u ovom poglavlju je definirana hipoteza H4.1 na slijedeći način:

H4.1: Uključivanje preliminarne selekcije atributa i inkrementalne faze u algoritam temeljen na GA, kombinirano s efektima nove strategije generiranja inicijalne populacije GA, rezultira statistički značajnim poboljšanjem prosječne klasifikacijske točnosti novog algoritma uz razinu pouzdanosti od 99%.

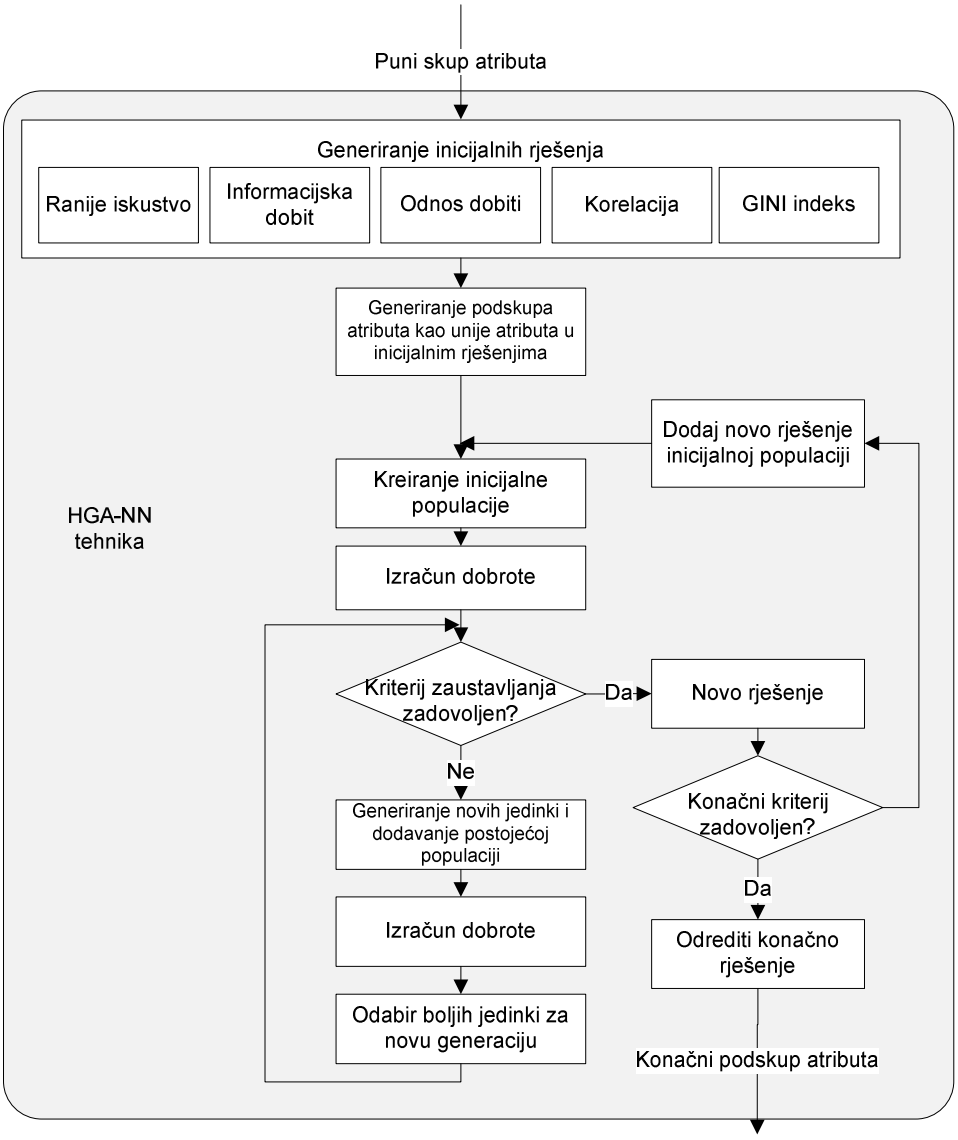
Bitno je istaknuti da je poboljšanje čak i za dio postotka dovoljno veliko da bi bilo znanstveno i praktično vrlo interesantno. Kako bi testirali našu hipotezu i ocijenili preciznost naše tehnike, eksperimenti su provedeni na dva stvarna skupa kreditnih podataka, jednom prikupljenom u hrvatskoj banci i drugom preuzetom iz UCI baze podataka.

4.3 Razvoj modela

GA kao evolucijska tehnika optimizacije počinje s početnom populacijom, tj. s inicijalnim rješenjima, kreće prema globalnom optimalnom rješenju i zaustavlja pretraživanje kada su zadovoljeni uvjeti za zaustavljanje. Standardni pristup počinje sa slučajnom inicijalnom populacijom i evoluira iz jedne generacije u drugu podvrgavajući jedinke križanjima i mutaciji (slike 3.1 i 3.2). Implementacija genetskog algoritma korištena u istraživanjima prezentiranim u ovoj disertaciji u osnovi prati algoritam opisan na slici 3.2. Detaljan opis genetskih algoritama s naputcima kada ih je najprikladnije koristiti može se naći u radovima Michalewicz (1998) i Mitchell (1996).

Pristup implementiran i testiran u nastavku ovog poglavlja je baziran na sljedećoj osnovnoj hipotezi: kada su dostupne a priori informacije o potencijalno atraktivnim područjima globalnog optimuma, tada početna populacija GA može biti generirana na taj

način da atraktivna područja (vjerojatnije regije globalnog optimuma) budu bolje pokrivena skupom točaka, a dimenzionalnost problema se može smanjiti na one atribute koji dobro opisuju atraktivna područja. Pri tome je atraktivno područje (atraktivna regija) globalnog minimuma definirano kao najveći skup točaka, takvih da za bilo koju početnu točku iz skupa vrijedi, da će iz nje, uz beskonačno mali pomak, algoritam stupnjevitog spusta konvergirati prema globalnom minimumu (Maaranen et al. 2004). Rezultati prijašnjih iskustava i rezultati brzih filtarskih tehnika smatraju se a priori informacijama o atraktivnim područjima. Zapravo su inicijalna rješenja koja su uključena u početnu populaciju GA te a priori informacije, a ostatak inicijalne populacije popunjava se slučajnim odabirom. Predloženi algoritam se može sagledati kao jedinstvena cjelina, kao što je prikazano na slici 4.1, ili može biti razmotren kroz faze izvedbe, kao što je prikazano na slici 4.2.



Slika 4.1 Dijagram tijeka hibridnog genetskog algoritma

Trorazinski hibridni algoritam HGA-NN se sastoji od sljedećih faza izvedbe:

1. redukcije pretraživačkog prostora,
2. pročišćavanja reduciranog podskupa atributa i
3. inkrementalne faze.

4.3.1 Redukcija prostora pretraživanja

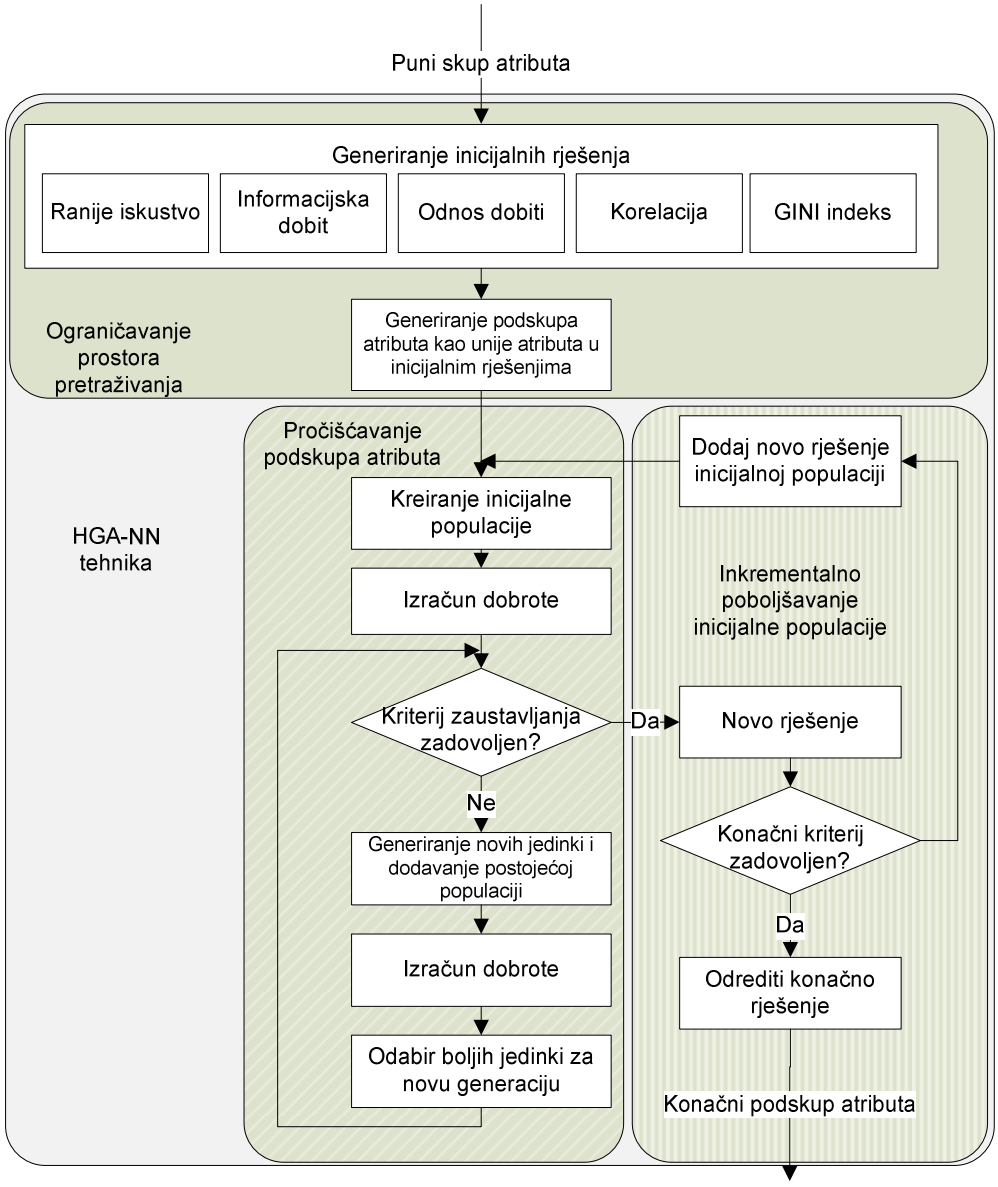
Cilj prve faze je preliminarno ograničiti područje pretrage. Prema slikama 4.1. i 4.2, ideja je spriječiti genetski algoritam da troši vrijeme na pretraživanje nebitnih područja pretraživačkog prostora. Redukcija prostora pretraživanja je izvedena u dvije podfaze: (1) generiranjem inicijalnih rješenja i (2) generiranjem reduciranog podskupa atributa.

Generiranje inicijalnih rješenja je izvedeno filtarskim tehnikama za rangiranje atributa i to: informacijskom dobiti, omjerom dobiti, Ginijevim indeksom i korelacijom (engl. *Information gain, Gain ratio, Gini index and Correlation*). Dodatno, u početna rješenja uključena su i rješenja predložena od domenskih stručnjaka na temelju njihovih prijašnjih iskustava. Ako u nekoj domeni taj tip znanja ne postoji, onda možemo uključiti ranije poznato rješenje problema, ako ono postoji. U drugoj fazi, inicijalna rješenja biti će ugrađena u početnu populaciju HGA.

Generiran reduciran podskup atributa predstavlja uniju svih atributa koji se pojavljuju u prijašnjim rješenjima i u rezultatima dobivenim brzim filtarskim tehnikama, tj. reducirani skup predstavlja uniju svih atributa koji se pojavljuju u inicijalnim rješenjima. Samo atributi koji se pojavljuju u rezultatima filtarskih tehnika i u prijašnjim rješenjima će se pojaviti na ulazu HGA i predstavljati reducirani skup atributa. Na taj način, restrikcija prostora pretraživanja može učiniti vjerojatan, odnosno, atraktivan prostor vrlo malim u odnosu na ukupan prostor pretraživanja. Reducirani skup atributa je nadalje filtriran u drugoj i trećoj fazi HGA-NN algoritma.

Preliminarna restrikcija atributa ima vrlo važnu ulogu u HGA-NN dizajnu. Kod toga korištenje velikog broja različitih tehnika često rezultira i odabirom velikog broja atributa, što može dovesti do problema. Jer ako postoji premali selekcijski pritisak, osnovni zadatak ove faze algoritma neće biti ispunjen. Tada neće biti poboljšana skalabilnost algoritma za one probleme odabira atributa koji su visoko dimenzionalni.

U suprotnom, upotreba malog broja i neadekvatnih brzo filtrirajućih tehnika može dovesti do izostavljanja važnih atributa. Uzimajući u obzir da karakteristike podataka suštinski utječu na performanse tehnika odabira atributa, dobro poznavanje karakteristika podataka je nužan preduvjet za dobru selekciju brzih tehnika u preliminarnoj fazi.



Slika 4.2 Hibridni genetski algoritam podijeljen u faze

4.3.2 Pročišćavanje reduciranog podskupa atributa

Pokazano je da jednostavna kombinacija najboljih individualnih atributa, izabranih pomoću filtarskih tehnika izbora, ne mora nužno dovesti do dobrih performansi klasifikacije (Kohavi i John, 1997). Drugim riječima, suvišnost atributa mogla bi uzrokovati kod algoritama za klasifikaciju degradaciju performansi klasifikacije. Zbog toga, u HGA-NN algoritmu reduciran podskup atributa je dodatno filtriran.

Ta filtracija je napravljena pomoću HGA. Standardni GA je poboljšan tako da može prihvatiti, kao dio svoje početne populacije, inicijalna rješenja dobivena drugim tehnikama kao i rješenja dobivena od eksperata u toj domeni. To poboljšanje je također omogućilo uvođenje treće faze algoritma.

Algoritam prikazan na slici 4.1 je napravljen korištenjem Rapid Miner 5.1.15 alata sa parametrima prikazanim u tablici 4.1. Kao što je prethodno rečeno, neka poboljšanja toga alata su bila nužna da bi se mogao izvršiti opisani algoritam. U tablici 4.1 su navedeni parametri za HGA i neki parametri za NN koji je komponenta GA-NNa i HGA-NNa. Parametri za NN nisu se mijenjali kroz eksperiment; na taj način, NN neće utjecati na razlike među rezultatima eksperimentalnih algoritama. Parametri koji su se mijenjali kroz izvođenje eksperimenta označeni su sa ✓ u drugom stupcu tablice, i svi se odnose na komponentu GA eksperimentalnih algoritama.

Tablica 4.1 Sažetak HGA-NN parametara

Parametar	**Promjena**	**Postavke**
Inicijalizacija populacije		
population size		50
initial probability for an feature to be switched on	✓	0.5
maximum number of features		12
minimum number of features		5
Reprodukcija		
Fitness measure		accuracy
Fitness function		neural network
the type of neural network		multilayer feed-forward network
network algorithm		backpropagation
activation function		sigmoid
the number of hidden layers		1
the size of the hidden layer		(number of features + number of classes) / 2 + 1
training cycles		[300;600]
learning rate		[0.3;1.0]
momentum		[0.2;0.7]
selection scheme	✓	roulette wheel
tournament size	✓	-
dynamic selection pressure		Yes
keep best individual		Yes
mutation probability		1 / number of features
crossover probability		0.9
crossover type	✓	uniform
Uvjeti završetka		
maximal fitness		Infinity
maximum number of generations		50
use early stopping		No

4.3.3 **Inkrementalna faza**

Kako navode Goldberg i Holland (1988), genetski bazirani klasifikatori počivaju na konkurenciji u rješavanju konflikata. To svojstvo omogućava inkrementalno izvođenje algoritma, pri čemu klasifikator testira nove strukture i hipoteze dok postupno poboljšava svoje performanse. Taj je pristup poglavito važan ako proces brzo konvergira ka lokalnom ekstremu (lokalnom optimumu). Zbog toga što su daljnje mogućnosti za poboljšanje rješenja standardnog genetskog algoritma vrlo male, potrebna je nova rekonstrukcija genetskog materijala. Tu ulogu preuzima inkrementalna faza. Inkrementalna faza omogućava rekonstrukciju genetskog materijala i doprinosi raznovrsnosti genetskog materijala.

Utjecaj početne populacije na brzinu konvergencije lokalnom optimumu može se promatrati na sljedeći način: Ako inicijalno izaberemo relativno dobro prilagođene jedinke koje pripadaju različitim atraktivnim područjima, križanje ne konvergira prebrzo lokalnom optimumu. Nasuprot tome, u uvjetima kada početno postoje relativno dobre jedinke koje pripadaju istom atraktivnom području, proces konvergira brzo lokalnom optimumu. Genetski materijal super jedinke brzo će prevladati u populaciji, te će nastati preuranjena konvergencija. Na taj način, određena količina genetskog materijala koja bi mogla biti upotrebljiva, a koja pripada različitim atraktivnim područjima je izgubljena jer se nalazi u lošim jedinkama.

Iz tog razloga, kao i zato što mi ne znamo unaprijed kojim područjima pripadaju naše relativno dobro prilagođene jedinke, nužno je imati mogućnost rekonstrukcije materijala i mijenjanja algoritamskih parametara da bi se poboljšale performanse modela. Važno je napomenuti da nakon svake faze hibridnog genetskog algoritma, možemo dinamički promijeniti uvjete za sljedeću fazu algoritma. Pa tako, prije svakog novog izvršenja faze 2 (pročišćavanje reduciranog podskupa atributa), možemo promijeniti početnu populaciju, shemu izbora, tip križanja, stopu mutacije ili bilo koji drugi parametar HGA i kontrolirati rezultate tih promjena.

4.4 Empirijska analiza

Performanse predloženog HGA-NN klasifikatora su vrednovane koristeći skup podataka stvarno odobrenih kredita u jednoj hrvatskoj banci, a zaključci su nadalje provjereni i potvrđeni na drugom skupu podataka stvarno odobrenih kredita koji je preuzet iz UCI baze podataka. Zbog toga je u ovome poglavlju doktorske disertacije, empirijska analiza podijeljena na dva dijela: (1) Eksperiment 1 sa hrvatskim skupom podataka i (2) Eksperiment 2 sa njemačkim skupom podataka.

4.4.1 **Eksperiment 1: Hrvatski skup podataka**

Empirijska analiza hrvatskog skupa podataka je podijeljena na četiri dijela. U odjeljku 4.4.1.1 dana je poveznica na opis skupa podataka. Odjeljak 4.4.1.2 prikazuje postavke parametara temeljem kojih je proveden eksperiment kao i postignute rezultate. Komparativni pregled performansi algoritama je prikazan u odjeljku 4.4.1.3, s osvrtom na iste. Statistička analiza dobivenih rezultata prezentirana je u odjeljku 4.4.1.4.

4.4.1.1 ***Opis skupa podataka***

Hrvatski skup kreditnih podataka je prikupljen u jednoj hrvatskoj banci i pokriva period od rujna 2004. do rujna 2011. Popis atributa s njihovim objašnjenjima kao i deskriptivnom statistikom te opisom cijelog procesa prikupljanja podataka za ovaj uzorak podataka nalazi se u odjeljku 3.5.1.

4.4.1.2 ***Eksperimentalni rezultati***

Kao što smo opisali u odjeljku 2 ovog poglavlja pod nazivom - Opis problema i pregled literature, **istraživanje u ovom poglavlju je koncentrirano**, a samim time i analiza rezultata, **oko sljedeća dva glavna pitanja**: (1) utjecaj redukcije prostora pretraživanja i (2) inkrementalne faze na cjelokupne performanse klasifikatora. Kombinirani efekti strategije

generiranja početne populacije GA na dobivene rezultate su također adresirani u kontekstu temeljna dva pitanja.

Tablica 4.2 Inicijalna rješenja i reducirani podskup atributa iz hrvatskog skupa podataka

Original skup atributa	**Tehnika odabira atributa**					**Reducirani skup atributa**
	Poznato rješenje	**GINI & IG**	**Odnos dobiti**	**Korelacija**	**Glaso-vanje**	
Grupa1						
att2		✓		✓		✓
att3						
att4						
att5	✓					✓
att6				✓		✓
att7	✓	✓		✓	✓	✓
att8			✓	✓		✓
att9						
att10	✓		✓	✓	✓	✓
att11						
att12		✓				✓
Grupa2						
att13		✓	✓			✓
att14						
att15		✓	✓			✓
att16	✓					✓
att17	✓	✓	✓		✓	✓
att18	✓		✓			✓
att19	✓					✓
att20						
att21						
att22						
att23	✓	✓	✓	✓	✓	✓
att24	✓	✓	✓		✓	✓
Grupa3						
att25						
att26						
att27		✓	✓	✓	✓	✓
att28	✓	✓	✓	✓	✓	✓
Grupa4						
att29				✓		✓
att30		✓	✓	✓	✓	✓
att31	✓	✓	✓	✓	✓	✓
Grupa5						
att32						
att33						
att34	✓			✓		✓

Na temelju niza eksperimenata, zaključeno je da procijenjena točnost klasifikacije počinje padati nakon smanjenja broja atributa ispod 12. Zbog toga je izabrano prvih 12 atributa, prikazanih u tablici 4.2, koristeći navedene tehnike. Dva posebna atributa, prvi (id) i zadnji (labela), su izostavljeni iz tablice. Jednom kada su izabrani atributi za svaku tehniku, atributi koji se pojavljuju u bilo kojoj selekciji ulaze u reducirani podskup atributa.

Atributi koji se pojavljuju u više od polovice drugih tehnika su važni za tehniku glasanja. Specifičnost tehnike glasanja je, u tom kontekstu, da ta tehnika daje samo jedno inicijalno rješenje i nema utjecaj na reducirani podskup atributa. Iz tablice 4.2 se može vidjeti da smo u fazi redukcije pretraživačkog prostora smanjili originalni skup atributa sa 33 na 21 atribut. Na taj način, kao što smo prethodno spomenuli, dobivena je reducirana verzija danog problema uz pretpostavku da: (1) su zadržani važni atributi, i da (2) rezultirajući smanjeni prostor pretraživanja može poboljšati skalabilnost tehnike i kvalitetu konačnog rješenja.

S ciljem usporedbe kvalitete i preciznosti, u ovom poglavlju opisane HGA-NN tehnike s u poglavlju 3 opisanom GA-NN tehnikom, izvedeni su paralelni testovi na istom skupu podataka. Temeljna razlika između navedenih tehnika je što GA-NN tehnika ne provodi redukciju pretraživačkog prostora i nema inkrementalnu fazu. Zapravo, druga faza u HGA-NN tehnici koja se zove „Pročišćavanje reduciranog podskupa atributa“ je izvedena iz GA-NN tehnike, tehnike s kojom uspoređujemo HGA-NN tehniku. Na taj način, komparacija će pokazati jesu li dodane faze, prva i treća, donijele poboljšanje. Opet napominjemo da su početni parametri koji se odnose na zajedničke komponente obiju tehnika, a to su genetski algoritam i neuronske mreže, postavljeni na iste vrijednosti u obje tehnike. Razlog takvoj izvedbi eksperimenta primarno leži u osiguranju jednostavne usporedbe rezultata.

Rezultati eksperimenata su prezentirani u tablicama 4.4 i 4.5. Svaka vrijednost u tablicama 4.4 i 4.5 predstavlja prosječnu točnost predviđanja spomenutih tehnika dobivenih na skupu od 1000 jedinki. Prosječna točnost predviđanja je izračunata koristeći deseterostruku unakrsnu validaciju zato što jednostavna tehnika validacije često procjenjuje pravu stopu pogrešaka preoptimistično (Malhotra i Malhotra, 2003). U našem slučaju, deseterostruka unakrsna validacija koristi stratificirani uzorak zato što je postotak loših kredita u odnosu na uredne kredite mali. Stratificirano uzorkovanje stvara slučajne podskupove i osigurava da je klasna distribucija u podskupovima (skoro) ista kao u cijelom skupu.

Nakon inicijalnog postavljanja parametara prikazanih u tablici 4.1, većina parametara se nije mijenjala od jednog do drugog izvođenja kako bi vertikalna komparacija rezultata imala smisla. Vertikalna komparacija rezultata je komparacija rezultata unutar iste tehnike koja se provodi s ciljem utvrđivanja utjecaja promjena vrijednosti parametra na rezultate svake tehnike. Promijenjeni parametri su naglašeni u tablici 4.3. Oni su postavljeni na iste vrijednosti u obje tehnike za isto izvođenje.

Tablica 4.3 Promjene parametara za različita izvođenja

Izvođenje	**Parametar**			
	p_initialise	**Shema selekcije**	**Udio u turniru**	**Tip križanja**
1	0.5	roulette wheel	-	uniform
2	0.5	Tournament	0.1	uniform
3	0.6	Tournament	0.05	uniform
4	0.6	Stochastic	-	one_point
5	0.6	Boltzmann	-	uniform
6	0.6	Cut	-	uniform
7	0.6	Unique	-	uniform
8	0.6	Unique	-	one point
9	0.6	Tournament	0.05	uniform
10	0.6	Tournament	0.05	one point
11	0.6	Cut	-	one point

Kao što se može vidjeti iz tablice 4.3, bilo je 11 izvođenja. Svako izvođenje odgovara jednom scenariju testiranja. U svakom scenariju, napravili smo neke promjene kako bi testirali puno različitih opcija. Značenja parametara kao i njihovi tipovi vrijednosti mogu se naći u Bäck et al. (1997) i Michalewicz (1998) s iznimkom jedinstvene (engl. *unique*) selekcije. Jedinstvena selekcija je vlastita shema selekcije. S jedinstvenom selekcijom, jedinke su poredane s obzirom na njihovu dobrotu u odnosu na funkciju cilja, a izabrane su one koje imaju najbolju dobrotu i koje su različite od prethodno izabranih. Broj izabranih jedinki je manji ili jednak veličini populacije. Ako nema dovoljno različitih jedinki, onda se populacija dopunjava s jedinkama najbolje dobrote.

Parametar nazvan „Udio u turniru“ ima vrijednosti u tablici samo kada je shema selekcije turnir; u ostalim shemama selekcije, vrijednost ovog parametra nema značenje. Dodatno, scenariji od 1 do 8 ne uključuju inkrementalne karakteristike HGA; ta karakteristika je uključena u scenarije od 9 do 11. Za svaki scenarij (izvođenje), rezultati su dani za različite generacije genetskog algoritma u tablici 4.4 i 4.5. Na taj način, dobivamo 99 prosječnih vrijednosti predviđanja, koje su izražene u postotku točnosti za svaku tehniku.

Tablica 4.4 Prosječna točnost predviđanja GA-NN tehnike izražena u % za hrvatski testni skup

Izvođenje	**Generacije GA-NNa**								
(scenario)	**1**	**2**	**3**	**5**	**8**	**10**	**20**	**30**	**50**
1.	79.7	80.5	80.5	80.8	81.6	81.8	81.8	81.8	81.8
2.	79.7	80.4	81.1	81.6	82.1	82.1	82.4	82.4	82.4
3.	80.5	81.3	81.3	81.3	81.4	81.8	81.9	82.5	82.5
4.	80.5	80.5	80.9	81.7	81.8	82.2	82.2	82.2	82.2
5.	80.5	80.6	80.9	80.9	81.1	81.1	81.9	81.9	81.9
6.	80.5	80.6	81.3	82.2	82.2	82.2	82.9	83.1	83.1
7.	80.5	80.6	81.3	81.5	82.2	82.2	82.2	82.3	82.7
8.	80.5	81.0	81.0	81.2	82.1	82.1	82.3	82.5	82.6
9.	80.5	81.3	81.3	81.3	81.4	81.8	81.9	82.5	82.5
10.	80.5	80.9	81.5	81.5	81.9	82.6	82.8	82.8	82.8
11.	80.5	81.1	81.3	81.5	81.5	81.5	82.2	82.2	82.5
Sred.vrij.	80.35	80.8	81.13	81.41	81.75	81.95	82.23	82.38	82.45
Std. dev.	0.324	0.332	0.283	0.383	0.378	0.406	0.364	0.371	0.378

Uspoređujući promjene parametara prikazane u tablici 4.3 s rezultatima izvođenja prikazanim u tablici 4.4 i 4.5, možemo vidjeti da je promjena u parametru pod nazivom p_initialize, koji označava inicijalnu vjerojatnost uključenja nekog atributa u inicijalno rješenje, imala izravan utjecaj na rezultat početne (prve) generacije trećeg i svakog sljedećeg izvođenja GA-NN tehnike. U isto vrijeme, ta promjena nije imala utjecaja na rezultate HGA-NN tehnike. Očito je da su inicijalna rješenja ugrađena u početnu populaciju HGA-NN tehnike bolja od bilo kojeg drugog rješenja u početnoj populaciji koje je generirano slučajno.

Tablica 4.5 Prosječna točnost predviđanja HGA-NN tehnike izražena u % za hrvatski testni skup

Izvođenje	**Generacije HGA-NNa**								
(scenario)	**1**	**2**	**3**	**5**	**8**	**10**	**20**	**30**	**50**
1.	81.3	81.3	81.3	81.3	82.2	82.3	82.3	82.3	82.3
2.	81.3	81.4	82.0	82.6	82.8	82.8	82.8	82.8	82.8
3.	81.3	81.3	82.4	82.4	82.4	82.4	82.7	83.0	83.1
4.	81.3	82.0	82.0	82.2	82.2	82.2	82.4	82.5	82.5
5.	81.3	81.3	81.3	81.3	81.5	81.5	81.7	81.8	81.8
6.	81.3	81.3	82.5	82.5	82.5	82.5	82.5	82.5	82.9
7.	81.3	81.3	81.3	81.7	81.7	81.7	82.4	82.4	83.0
8.	81.3	81.3	81.4	81.8	81.8	82.5	82.5	82.7	83.1
9.	81.7	82.0	82.2	82.3	82.9	82.9	82.9	82.9	83.4
10.	82.0	82.5	82.5	82.8	82.9	82.9	83.4	83.4	83.4
11.	82.0	82.1	82.1	82.3	82.4	82.5	82.9	83.0	83.4
Sred.vrij.	81.46	81.62	81.91	82.11	82.30	82.38	82.59	82.66	82.88
Std. dev.	0.291	0.442	0.495	0.511	0.48	0.451	0.432	0.43	0.508

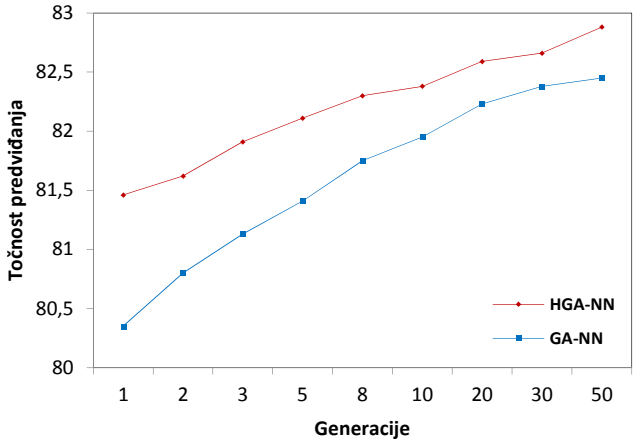
Prva promjena u prvoj generaciji HGA-NN-a se dogodila u devetom izvođenju. Promjena se dogodila zato što je rezultat (rezultirajuća kombinacija atributa) osmog izvođenja HGA-NN algoritma dodan u inicijalno rješenje devetog izvođenja. Drugim riječima, u devetom izvođenju HGA-NN-a, funkcija (karakteristika) zvana inkrementalna faza se aktivirala. Uz utjecaj na rezultate prve generacije, inkrementalna faza je utjecala i na konačni rezultat HGA-NN tehnike. To je vidljivo iz toga što je u devetom izvođenju prvi put preciznost HGA-NN-a dosegla vrijednost 83.4%. Iako smo dodali rezultat prethodnog izvođenja inicijalnoj populaciji desetog i jedanaestog izvođenja HGA-NNa, konačni rezultati se nisu popravili. Nakon dva generiranja bez ikakvog poboljšanja u preciznosti, proces je zaustavljen. Usporedivši rezultate prije i poslije aktivacije inkrementalne faze, iz tablice 4.5 može se vidjeti da postoje korisni efekti aktivacije te faze na klasifikacijsku preciznost algoritma, iako je HGA-NN tehnika već ranije postigla u prosjeku bolji rezultat od GA-NN tehnike.

4.4.1.3 *Komparacija i diskusija rezultata*

Da bi procijenili točnost HGA-NN algoritma i izmjerili utjecaj redukcije prostora pretraživanja i inkrementalne faze na sveukupne performanse klasifikatora, uspoređujemo njegovu točnost s točnosti GA-NN algoritma. Jednostavna komparacija te dvije tehnike odabira atributa je moguća zato što su korišteni isti početni parametri i iste tehnike klasifikacije i validacije. Sažeti rezultati ovog iscrpnog istraživanja su prikazani u tablici 4.6. Iz tih rezultata, može se primijetiti, da u svih devet mjerenih generacija, od 1 do 50, HGA-NN algoritam postiže bolje prosječne rezultate od GA-NN algoritma.

Tablica 4.6 Komparacija prosječnih točnosti predviđanja izražena u % za hrvatski testni skup

Tehnika	**Generacije GA**								
	1	**2**	**3**	**5**	**8**	**10**	**20**	**30**	**50**
GA-NN	80.35	80.80	81.13	81.41	81.75	81.95	82.23	82.38	82.45
HGA-NN	81.46	81.62	81.91	82.11	82.30	82.38	82.59	82.66	82.88

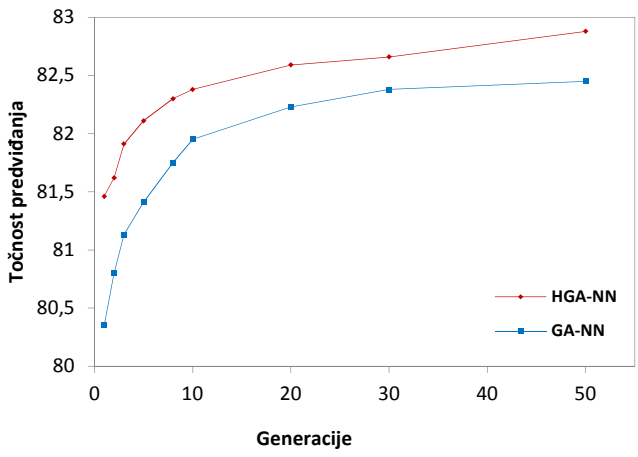


Slika 4.3 Komparacija rezultata za hrvatski skup podataka sa intervalom na abscisi prema izvedenim mjerenjima

Trendovi su još jasniji na slikama 4.3 i 4.4, koje prikazuju iste rezultate promatrano s dvije različite pozicije, s dva različita aspekta. Stoga su, mjerne skale na osi X različite.

Mjerna ljestvica je na slici 4.3 uravnotežena u odnosu na očekivane rezultate istraživanja. Ravnoteža je postignuta jer su rezultati mjerenja prikazani gotovo ravnom linijom, kao što se može vidjeti na slici 4.3. Dodatno, tablica 4.6 i slika 4.3. pokazuju da se razlika između prosječne točnosti predviđanja dviju tehnika u prvih 10 generacija GA smanjuje, dok između 10-te i 50-te generacije genetskog algoritma razlika u točnosti predviđanja ostaje ista. Ovaj trend pripisujemo činjenici da je HGA-NN tehnika dobila nekoliko dobrih inicijalnih rješenja u početnu populaciju od kojih nije odmah pronašla puno bolje kombinacije. U isto vrijeme, u GA-NN tehnici, genetski algoritam je počeo sa slučajno generiranom inicijalnom populacijom i pronašao je kombinacije gena koje su poboljšale vrijednost funkcije cilja relativno bolje nego kod HGA-NN algoritma. Unatoč inicijalno sporijem poboljšanju u vrijednosti objektivne funkcije, HGA-NN algoritam je kroz sve generacije postigao bolje rezultate.

Izraženo u postotcima, vidljivo je da se prosječna točnost predviđanja GA-NN tehnike u prvih 10 generacija poboljšala za 1.6 postotnih bodova, dok se u isto vrijeme, prosječna točnost predviđanja HGA-NN tehnike poboljšala za 0.92 postotnih bodova. Prosječna točnost predviđanja za obje tehnike između 10-te i 50-te generacije se poboljšala za 0.5 postotnih bodova.



Slika 4.4 Komparacija rezultata za hrvatski skup podataka s jednakim intervalima na abscisi

Iz slike 4.4 jasno je da vrijednosti ciljne funkcije poslije desete generacije rastu vrlo polako, što je u skladu s prijašnjim opažanjem da vrijednosti objektivne funkcije u zadnjih 40 generacija, od 10-te do 50-te, rastu manje nego u samo prve četiri generacije.

Iz rezultata prikazanih u tablici 4.6 i na slikama 4.3 i 4.4, možemo zaključiti da su, u rješavanju problem odabira atributa za skup kreditnih podataka iz hrvatske banke, rezultati HGA-NN tehnike bolji od rezultata GA-NN tehnike u pogledu prosječnoj točnosti svih generacija, uključujući i točnost konačnog rješenja. Ta tvrdnja je podržana i statističkim testovima.

4.4.1.4 *Statistička analiza*

Testirali smo novo predloženu HGA-NN tehniku s GA-NN tehnikom koristeći udvojeni *t*-test za njihove sumarne prosječne rezultate točnosti prikazane u tablici 4.6. Nulta hipoteza je da ne postoji razlika između prosječne točnosti te dvije tehnike. U usporedbi udvojenog *t*-testa s t-vrijednosti od 6.8136 i s 8 stupnjeva slobode, odbacili smo nultu hipotezu zato što je izračunata t-vrijednost veća od kritične tablične vrijednosti za dvostrani test koja je 3.3554 s odabranom razinom signifikantnosti od 0.01. Udvojeni *t*-test pretpostavlja da smo uzorkovali naše parove vrijednosti iz populacije parova u kojoj razlika između parova slijedi Gausovu distribuciju (Myers i Well, 2003). Zbog toga smo testirali tu pretpostavku testom normalnosti. S obzirom da je p-vrijednost D'Agostino i Pearson skupnog normalnog testa jednaka 0.6137 i veća je od a=0.05 te zbog toga što je p-vrijednost Shapiro-Wilk normalnog testa jednaka 0.59 i veća od a= 0.05, nema dovoljno dokaza koji bi podržali tvrdnju da smo uzorkovali naše parove vrijednosti iz populacije parova u kojoj su razlike između parova različite od Gausove distribucije.

Tablica 4.7 Statistički testovi razlika rezultata HGA-NN i GA-NN tehnike za hrvatski skup podataka

Test	**Razina signifikantnosti (alfa)**	p-vrij.
t-Test za zavisne uzorke (dvostrani)	0.05	0.000
D'Agostino & Pearson omnibus test normalnosti	0.05	0.614
Shapiro-Wilks test normalnosti	0.05	0.590
Wilcoxonov test ekvivalentnih parova (Wilcoxon matched-pairs signed rank test)	0.05	0.004

Osim s udvojenim *t*-testom, testirali smo novo predloženu HGA-NN tehniku s GA-NN tehnikom koristeći Wilcoxonov test uparenih parova (engl. *Wilcoxon matched-pairs signed rank test*). Taj test pruža dobru alternativu udvojenom *t*-testu kad je populacija razlika rezultata simetrično raspoređena. Taj test je malo manje snažan od *t*-testa kada su podaci raspoređeni po normalnoj distribuciji, a može biti znatno bolji kada su razlike rezultata simetrično (ali ne nužno i normalno) distribuirane s jakim repovima (Myers i Well, 2003).

Wilcoxon Signed Rank test uparenih parova testira nultu hipotezu da je medijan razlike jednak 0.0 nasuprot alternativne hipoteze da je medijan razlike različit od 0.0, tj. nulta hipoteza je da su oba algoritma jednako dobra. Prema tablici egzaktnih kritičnih vrijednosti za Wilcoxon Signed Rank test uparenih parova (Myers i Well, 2003), za razinu signifikantnosti od $\alpha = 0.01$ i $N = 9$, razlika između klasifikatora je signifikantna ako je manja suma manja od 2. Suma rangova za pozitivne razlike je $R_{pos} = 45$, a suma rangova za negativne razlike je $R_{neg} = 0$. Iz razloga što je manja suma rangova $R_{neg} = 0$, možemo odbaciti nultu hipotezu s 99.0% pouzdanosti i odbaciti ideju da je razlika slučajna, stoga zaključujemo da populacije imaju različite medijane u korist novo predložene HGA-NN tehnike. Taj zaključak je u skladu s *t*-testom i hipotezom H4.1 iz ovog poglavlja doktorske disertacije.

4.4.2 **Eksperiment 2: Njemački skup podataka**

S ciljem dalje validacije iznesenih rezultata HGA-NN algoritma, izveden je Eksperiment 2 na javno dostupnom njemačkom skupu podataka. Rezultati su dodatno uspoređeni s drugim rezultatima objavljenima u literaturi.

4.4.2.1 ***Opis skupa podataka***

Njemački skup kreditnih podataka je dostupan u UCI Repository of Machine Learning bazi podataka (Bache i Lichman, 2013), i sastoji se od 700 instanci dobrih klijenata i 300 instanci loših kreditnih klijenata. Originalni skup podataka, u obliku koji je dao prof. Hofmann, sadrži, za svakog klijenta, 20 ulaznih atributa koji opisuju kreditnu povijest, salda

računa, namjene kredita, iznose kredita, status zaposlenja, osobne informacije, dob, stambeni status i naziv posla.

Tablica 4.8 Transformirani njemački skup kreditnih podataka s deskriptivnom statistikom

Atr.	**Šifra**	**Opis**	**Statistika**	**Raspon**
at1	**Id**	redni broj instance	avg = 500.500 +/- 288.819	[1; 1000]
at2	chk_acct	status tekućeg računa	avg = 1.577 +/- 1.258	[0; 3]
at3	duration	trajanje kredita u mjesecima	avg = 20.903 +/- 12.059	[4; 72]
at4	history	kreditna povijest	avg = 2.545 +/- 1.083	[0; 4]
at5	new_car	namjena kredita	avg = 0.234 +/- 0.424	[0; 1]
at6	used_car	namjena kredita	avg = 0.103 +/- 0.304	[0; 1]
at7	furniture	namjena kredita	avg = 0.181 +/- 0.385	[0; 1]
at8	radio/tv	namjena kredita	avg = 0.280 +/- 0.449	[0; 1]
at9	education	namjena kredita	avg = 0.050 +/- 0.218	[0; 1]
at10	retraining	namjena kredita	avg = 0.097 +/- 0.296	[0; 1]
at11	amount	iznos kredita	avg = 3271.258 +/- 2822.737	[250; 18424]
at12	sav_acct	prosječni saldo štednje	avg = 1.105 +/- 1.580	[0; 4]
at13	employment	trajanje sadašnjeg zaposlenja	avg = 2.384 +/- 1.208	[0; 4]
at14	install_rate	rata u odnosu na raspoloživ dohodak	avg = 2.973 +/- 1.119	[1; 4]
at15	male_div	rastavljeni muškarac	avg = 0.050 +/- 0.218	[0; 1]
at16	male_single	neoženjen muškarac	avg = 0.548 +/- 0.498	[0; 1]
at17	male_mar_wid	oženjen muškarac ili udovac	avg = 0.092 +/- 0.289	[0; 1]
at18	co-applicant	ima sudužnika	avg = 0.041 +/- 0.198	[0; 1]
at19	guarantor	ima jamaca	avg = 0.052 +/- 0.222	[0; 1]
at20	present_resident	isto prebivalište - godina	avg = 2.845 +/- 1.104	[1; 4]
at21	real_estate	posjeduje nekretninu	avg = 0.282 +/- 0.450	[0; 1]
at22	prop_unkn	ne posjeduje nikakvu imovinu (ili nepoznato)	avg = 0.154 +/- 0.361	[0; 1]
at23	age	starost u godinama	avg = 35.546 +/- 11.375	[19; 75]
at24	other_install	ima drugih kredita	avg = 0.186 +/- 0.389	[0; 1]
at25	rent	podstanar	avg = 0.179 +/- 0.384	[0; 1]
at26	own_res	vlasnik stana	avg = 0.713 +/- 0.453	[0; 1]
at27	num_credits	broj kredita u ovoj banci	avg = 1.407 +/- 0.578	[1; 4]
at28	job	vrsta posla	avg = 1.904 +/- 0.654	[0; 3]
at29	num_dependents	broj osoba koje uzdržava	avg = 1.155 +/- 0.362	[1; 2]
at30	telephone	ima telefon na svoje ime	avg = 0.404 +/- 0.491	[0; 1]
at31	foreign	strani radnik	avg = 0.037 +/- 0.189	[0; 1]
at32	**LABELA**	**Kreditni rejting**	mode = 1 (700), least = 0 (300)	1 (700), 0 (300)

Originalni skup podataka je imao 13 kategorijskih atributa, od kojih su nominalni pretvoreni u seriju binarnih atributa kako bi mogli biti primjereno korišteni u neuronskim

mrežama. Nekoliko ordinarnih kategorijskih atributa su ostavljeni u originalnom stanju i tretirani su kao brojčani podaci. Transformirani njemački skup kreditnih podataka sadrži 30 redovnih atributa tipa integer i dva (id, labela) posebna atributa, te se do njega može doći na <http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/download-course-materials/>. Svi atributi s deskriptivnom statistikom su prikazani u tablici 4.8.

4.4.2.2 *Eksperimentalni rezultati*

Inicijalna rješenja su generirana prema proceduri definiranoj u odjeljku razvoj modela ovog poglavlja te su korištene iste selekcijske tehnike kao i za hrvatski skup podataka. Nakon generiranja inicijalnih rješenja, reducirani podskup atributa je generiran kao unija svih atributa iz inicijalnih rješenja. Iz tablice 4.9 se može vidjeti da smo u fazi redukcije prostora pretraživanja reducirali originalni skup atributa s 30 na 20 atributa.

Tablica 4.9 Inicijalna rješenja i reducirani podskup atributa za njemački skup kreditnih podataka

Original skup atributa	Tehnika odabira atributa					Reducirani skup atributa
	Poznato rješenje	Informa. dobit	Omjer dobiti	Korelacija	Glasanje	
att2	✓	✓	✓	✓	✓	✓
att3	✓	✓	✓	✓	✓	✓
att4	✓	✓	✓	✓	✓	✓
att5	✓					✓
att6			✓	✓		✓
att7						
att8		✓		✓		✓
att9						
att10						
att11		✓	✓	✓	✓	✓
att12	✓	✓	✓	✓	✓	✓
att13		✓		✓		✓
att14	✓					✓
att15	✓					✓
att16	✓					✓
att17						
att18						
att19	✓					✓
att20						
att21	✓	✓	✓	✓	✓	✓
att22	✓	✓	✓	✓	✓	✓
att23		✓	✓			✓
att24		✓	✓	✓	✓	✓
att25						
att26		✓	✓	✓	✓	✓
att27						
att28						
att29	✓					✓
att30						
att31	✓		✓			✓

Inicijalni parametri genetskog algoritma su postavljeni na iste vrijednosti kao i u Eksperimentu 1, osim maksimalnog broja atributa koji je postavljen na 16 iz razloga što je na manjem broju atributa počela padati preciznost predviđanja. Nadalje, inicijalni parametri neuronske mreže su postavljeni na iste vrijednosti kao i u Eksperimentu 1, osim broja ciklusa treninga koji je postavljen na interval [30; 100]. Svi inicijalni parametri, iz izvođenja u izvođenje, nisu se mijenjali osim onih parametara koji su označeni u tablici 4.3.

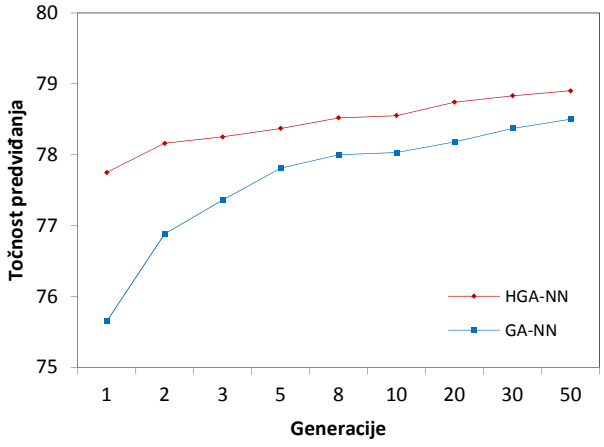
U skladu s navedenim, za njemački skup kreditnih podataka napravili smo 99 izmjera, koje su izražene u postotcima točnosti za svaku tehniku. HGA-NN tehnika je u osmom izvođenju dostigla najbolju prosječnu točnost predviđanja od 79.4%, s opisanom jedinstvenom shemom selekcije (engl. *unique selection scheme*) i jednostrukim križanjem (engl. *one_point crossover*). GA-NN tehnika dostigla je svoj najbolji rezultat, koji je bio 79.0% u sedmom izvođenju, s jedinstvenom shemom selekcije i jednostrukim križanjem. Najbolja rješenja HGA-NN i GA-NN tehnike su postignuta s odabranih 12 i 16 atributima, respektivno.

Da bismo dodatno provjerili valjanost rezultata Eksperimenta 1, sumarni komparativni rezultati Eksperimenta 2 su prikazani u tablici 4.10 i na slikama 4.5 i 4.6.

Tablica 4.10 Komparacija prosječnih točnosti predviđanja izražena u % za njemački testni skup

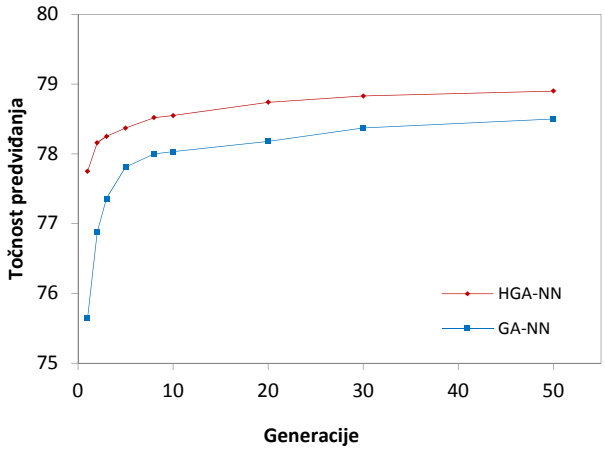
Tehnika	**Generacije GA**									
	1	**2**	**3**	**5**	**8**	**10**	**20**	**30**	**50**	**p-vrij.**[a]
GA-NN	75.65	76.88	77.36	77.81	78.00	78.03	78.18	78.37	78.50	
HGA-NN	77.75	78.16	78.25	78.37	78.52	78.55	78.74	78.83	78.90	0.004*

[a] p-vrijednost je za Wilcoxonov test ekvivalentnih parova (alpha=0.05)



Slika 4.5 Komparacija rezultata za njemački skup podataka sa intervalom na abscisi prema izvedenim mjerenjima

Iz tablice 4.10 i slike 4.5 je vidljivo da se razlika između prosječne točnosti predviđanja dviju tehnika u prvih 8 generacija GA smanjuje, dok između osme i pedesete generacije GA razlika u točnosti predviđanja ostaje ista.



Slika 4.6 Komparacija rezultata za njemački skup podataka
s jednakim intervalima na abscisi

Slika 4.6 jasno prikazuje da vrijednost ciljne funkcije poslije osme generacije ulazi u fazu zasićenosti, što je u skladu s nalazima iz Eksperimenta 1. Vrijednost ciljne funkcije u zadnjih 40 generacija, od 10-te do 50-te generacije, raste manje nego u samo prve 4 generacije.

Rezultati prikazani u tablicama 4.6 i 4.10 pokazuju jasnu vezu između HGA-NN i GA-NN tehnike i njihovih performansi. Prema ovim rezultatima možemo zaključiti da utjecaj redukcije prostora pretraživanja i inkrementalne faze, kombinirano s efektima strategije generiranja inicijalne populacije genetskog algoritma na ukupne performanse klasifikatora, postoji i da je pozitivan. Taj rezultat je konzistentan na oba skupa podataka, hrvatskom i njemačkom, a potvrđuju ga i statistički testovi.

Tablica 4.11 Komparacija rezultata s drugim tehnikama (njemački kreditni skup)

Tehnika	Odabrani atributi		Točnost predviđanja	
	Art.sred.	Std. dev.	Art.sred. (%)	Std. dev.(%)
HGA-NN	12,37	1,57	**78,90**	0,297
GA-NN	13,91	1,81	78,50	0,402
SVM+GA[1]	13.30	1.41	77.92	3.970
GP[1]	13.20	2.10	78.10	4,120
NN[2]	Svi	-	74.67	-
SVM[3]	Svi	-	76.59	0.328

[1] Huang, Chen i Wang, 2007.

[2] Khashman, 2010.

[3] Wang et al. 2009.

Tablica 4.11 uspoređuje rezultate dobivene predloženom HGA-NN tehnikom s rezultatima dobivenim drugim postojećim tehnikama, uključujući GA-NN tehniku, na njemačkom skupu kreditnih podataka. Prikazane prosječne stope točnosti predviđanja odnose se na rezultate dobivene na validacijskom skupu podataka. Rezultati dobiveni na skupu podataka za učenje nisu prikazani. Komparativni pregled ukazuje da je predloženi GA-baziran algoritam prihvatljiva alternativa za optimiranje podskupa atributa i parametara neuronske mreže kod ocjene kreditnog rizika, i kroz to, za optimiranje skupa mrežnih pondera koji daju najbolju prosječnu točnost predviđanja.

U njemačkom kreditnom skupu podataka odnos loših i dobrih instanci kredita je 30:70. Kod takvih klasno neujednačenih klasifikacijskih problema detekcija instanci manjinske klase od posebnog je značaja. U takvim slučajevima pored točnosti predviđanja često se koriste i druge mjere za vrednovanje kvalitete klasifikatora. Stoga su u tablici 4.12 uz procijenjene prosječne točnosti predviđanja prikazane i vjerojatnosti pogrešaka tipa I i tipa II za njemački kreditni skup podataka, a u tablici 4.13 su prikazani usporedni rezultati relativnog troška pogrešne klasifikacije HGA-NN tehnike i tehnika čiji su rezultati dostupni u literaturi za različite odnose greški tipa I i II .

Tablica 4.12 Komparacija točnosti i vjerojatnosti pogrešaka tipa I i II HGA-NN algoritma u odnosu na druge tehnike za njemački skup podataka

Tehnika	Vjerojatnost pogreške		Točnost
	P_1 (%)	P_2 (%)	Art.sred. (%)
HGA-NN	46.61	10.17	**78.90**
SVM(Han, Han i Zhao, 2013)	37.00	18.00	77.00
LogR(Zhu et al., 2013)	50.66	11.69	76.62
Bagging/MLP(Nanni i Lumini, 2009)	49.40	24.60	75.33
Logit(Lieli i White, 2010)	18.33	44.00	63.70

Za izračun ukupnih relativnih troškova pogrešne klasifikacije (RC) koristiti smo formulu (3.20), podatke o vjerojatnosti grešaka tipa I i II iz tablice 4.12 te relativne odnose troškova pogrešne klasifikacije prikazane u zaglavlju tablice 4.13. Za svaku tehniku su izračunate RC vrijednosti za 7 scenarija, pri čemu je troškovno gledajući za svaki scenarij najbolja tehnika koja ima najmanju RC vrijednost.

Tablica 4.13 Komparacija troškova pogrešne klasifikacije za njemački skup podataka

Tehnika	Odnos troškova (C_I:C_{II})						
	1:1	2:1	3:1	4:1	5:1	8:1	10:1
HGA-NN	0,2161*	0,3559	0,4958	0,6356	0,7754	1,1949	1,4746
SVM(Han, Han i Zhao, 2013)	0,2370	0,3480*	0,4590*	0,5700	0,6810	1,0140	1,2360
LogR(Zhu et al., 2013)	0,2338	0,3858	0,5378	0,6898	0,8417	1,2977	1,6016
Bagging/MLP (Nanni i Lumini, 2009)	0,3204	0,4686	0,6168	0,7650	0,9132	1,3578	1,6542
Logit(Lieli i White, 2010)	0,3630	0,4180	0,4730	0,5280*	0,5830*	0,7479*	0,8579*

* Najbolja tehnika za svaki scenarij

Iz tablice 4.13 se može vidjeti da će HGA-NN tehnika, koja daje najtočnije predviđanje, biti najprihvatljivija za banke u slučaju da je trošak netočnog predviđanja koje lošeg klijenta proglašava dobrim (pogreška tipa I) jednak trošku netočnog predviđanja kod kojeg je dobar klijent proglašen lošim (pogreška tipa II). Za druge scenarije HGA-NN tehnika, koja daje

najtočnije predviđanje, ne daje najnižu RC vrijednost. Kod toga je uočljiv jedan naizgled paradoksalni slučaj, da tehnika Logit, koja daje ukupno najnetočnije predviđanje, daje najniže vrijednosti ukupnih troškova pogrešne klasifikacije za omjere troškova (pogreška tipa I / pogreška tipa II) 4:1, 5:1, 8:1 i 10:1. Kao što smo već ranije naglasili, opravdano je očekivati da banke neće optimirati izbor atributa temeljem funkcije točnosti, nego će vjerojatno htjeti optimirati troškovnu funkciju za neke sebi svojstvene omjere, što ovisi o sklonosti svake banke za preuzimanje rizika.

U prezentiranoj HGA-NN tehnici smo i dalje optimirali prosječnu točnost predviđanja kao ciljnu funkciju, što je najčešća benchmark mjera. Pri tome nismo koristili niti jednu tehniku za ublažavanje negativnih utjecaja klasno neuravnoteženih podataka i različitih odnosa troškova pogrešne klasifikacije. Postignuti su najbolji rezultati u smislu ukupne prosječne točnosti klasifikacije kao i najbolji rezultati kod relativnih troškova klasifikacije 1:1, ali ne i za sve ostale omjere. Zbog toga je u HGA-NN tehniku potrebno implementirati odgovarajuće tehnike za ublažavanje negativnih utjecaja klasno neuravnoteženih podataka i različitih odnosa troškova pogrešne klasifikacije kako bi smo je učinili konkurentnom u svim uvjetima.

4.5 Zaključci poglavlja

S rastućom konkurencijom i rizikom, kao i znatno smanjenom razinom profita u industriji odobravanja kredita građanima, poslovi bi trebali biti brži, manje rizični, egzaktniji i bazirani na podacima. Da bi udovoljile tim zahtjevima, kreditne institucije trebaju više koristiti sofisticirane metode za ocjenu kreditnih rizika. Selekcija atributa je jedan od izazovnih problema u stvaranju takve sofisticirane metode. U ovome poglavlju predložena je nova, napredna hibridna tehnika selekcije atributa koja rješava navedeni izazov.

Nova tehnika koristi ranija iskustva eksperata i efikasnost brzih algoritama za rangiranje atributa kao i optimizacijske mogućnosti GA. Dizajniran je trorazinski hibridni algoritam koji se sastoji od (1) redukcije pretraživačkog prostora, (2) pročišćavanja reduciranog podskupa atributa, i (3) inkrementalne faze. Redukcijom područja pretraživanja, nepotrebni atributi mogu biti na brzinu uklonjeni. Pročišćavanje reduciranog podskupa atributa nadalje ispituje i filtrira reducirani skup atributa. Inkrementalna faza dodatno poboljšava performanse modela. Da bismo procijenili njegovu efikasnot, hibridni algoritam je primijenjen na dva skupa stvarnih kreditnih podataka: hrvatski i njemački.

Ostvarena prosječna točnost klasifikacije je uspoređena pomoću parametarskog udvojenog *t*-testa i neparametarskog Wilcoxon Signed Rank testa uparenih parova s prosječnom stopom točnosti ostvarenom pomoću GA-NN tehnike. Rezultati statističkih testova pokazuju da podaci potvrđuju hipotezu istraživanja na razini signifikantnosti $p < 0,01$.

Zbog toga što je problem odabira atributa sveprisutan u aktivnostima dubinske analize podataka, tema budućih istraživanja može biti primjena opisanog algoritma na skupove podataka iz drugih područja. Na taj način, ovisno o karakteristikama podataka, mogu se koristiti različite brze tehnike u redukciji prostora pretraživanja. Konačno, iz usporedbe rezultata na njemačkom skupu podataka, jasno je da rezultati primjene ove tehnike nadmašuju rezultate objavljene u literaturi. Rezultati dodatno potvrđuju hipotezu da redukcija područja pretraživanja i inkrementalna faza, kombinirano s efektima strategije generiranja inicijalne populacije GA, rezultiraju poboljšanjem performansi klasifikatora. To poboljšanje je dovoljno veliko da bi bilo znanstveno i praktično interesantno. Stoga se HGA-NN heuristika može smatrati obećavajućim dodatkom postojećim tehnikama dubinske analize podataka. U nastavku rada želimo empirijski potvrditi da je predložena nova tehnika primjenjiva u općem slučaju, odnosno, u bilo kojoj banci neovisno o njezinoj politici preuzimanja kreditnog rizika, stoga će tema sljedećeg poglavlja biti optimizacija klasno i troškovno neuravnoteženih podataka.

POGLAVLJE 5

UČENJE TROŠKOVNO OSJETLJIVE KLASIFIKACIJE IZ NEURAVNOTEŽENIH PODATAKA

5.1 Uvod

U mnogim primjerima nadziranog učenja postoji značajna razlika među apriornim vjerojatnostima različitih klasa, to jest među apriornim vjerojatnostima po kojima neki primjer pripada klasama (klasifikacijskim kategorijama) danog klasifikacijskog problema. Ova situacija je poznata kao problem klasne neravnoteže. Skup je klasno neuravnotežen ako klasifikacijske kategorije nisu približno jednako zastupljene (Chawla et al., 2002; Chawla, Japkowicz i Kotcz, 2004; Menardi, Tedeschi i Torelli, 2011; Naganjaneyulu i Kuppa, 2013).

Takva neravnoteža je uobičajena u mnogim stvarnim problemima iz financija, medicine, telekomunikacija, ekologije, biologije, i ne samo njih, a može se smatrati danas jednim od glavnih problema u dubinskoj analizi podataka. Nadalje, valja istaknuti da je, prilikom učenja klasifikatora, manjinska klasa obično ta za kojom postoji najviši interes jer, kada nije dobro klasificirana, predstavlja najveći trošak (López et al., 2013; Ling, 2004). Stoga bismo mogli reći da je u klasno neuravnoteženim skupovima podataka vrlo važno što točnije klasificirati instance manjinske klase.

Većina objavljenih znanstvenih radova o ponašanju standardnih klasifikacijskih algoritama na neuravnoteženim skupovima podataka ukazuje da zbog klasne neravnoteže dolazi do značajnog pogoršanja performansi. Neravnoteža se izražava kroz omjer neravnoteže (engl. *imbalance ratio* , IR), koji je definiran kao omjer broja slučajeva u većinskoj klasi prema broju primjera u manjinskoj klasi (López et al., 2013).

Istraživanja ukazuju i na druge faktore koji pridonose degradaciji performansi kod neuravnoteženih skupova podataka. Tako se navodi da je vrlo utjecajni faktor razina u kojoj se preklapaju klase od interesa (Dal Pozzolo et.al., 2013), kao i da šum (engl. *noise*) značajno utječe na performanse (Van Hulse i Khoshgoftaar, 2009), kod čega je posebno značajno u kojoj se klasi šum pojavljuje. Gamberger, Lavrac i Dzeroski (2000) navode da se šum pojavljuje u onim primjerima koji iz bilo kojeg razloga odstupaju od karakteristika većine ostalih primjera. Uzroci mogu biti ili u procesu prikupljanja podataka, zbog grešaka bilo koje vrste, ili zato što su ti primjeri stvarni izuzeci kao predstavnici vrlo rijetkih ali realnih pojava. Otkrivanje šuma definiraju kao postupak otkrivanja onih primjera koji omogućuju smanjenje veličine minimalnog skupa literala nužnih za konstrukciju potpuno točnog rješenja. Utjecaj šuma na performanse klasifikatora vrlo značajno zavisi i o korištenom algoritmu za učenje (Van Hulse i Khoshgoftaar, 2009), kod čega su jednostavniji algoritmi kao što su naivni Bayesov algoritam i algoritam najbližeg susjeda često robusniji nego znatno složeniji algoritmi kao algoritam s potpornim vektorima (engl. *Support Vector Machines,* SVM) i algoritam slučajne šume (engl. *Random Forests*).

Kod neuravnoteženih skupova podataka standardnim algoritmima klasifikacije je svojstveno da su često pristrani u korist većinske klase, poznate još kao "negativna" klasa, stoga postoji veća stopa klasifikacijske pogreške za instance manjinske klase, poznate kao "pozitivna" klasa. Tijekom posljednjih nekoliko godina, predložene su mnoge tehnike za rješavanje ove vrste klasifikacijskih problema, bilo kroz (i) podatkovni, (ii) algoritamski ili (iii) hibridni (kombinirani) pristup. Podatkovni pristup obično se temelji na različitim tehnikama uzorkovanja, uključujući: slučajno nad-uzorkovanje (engl. *random over-sampling*) manjinske klase s mogućnošću ponovnog izbora (engl. *with replacement*), slučajno pod-uzorkovanje (engl. *random under-sampling*) većinske klase kao i upravljano (engl. *directed*) nad i pod uzorkovanje te njihove kombinacije (Chawla, Japkowicz i Kotcz, 2004). Algoritamski pristup uvodi: (i) nejednake težine za instance klasa kako bi usmjerio klasifikator da obrati više pozornosti na manjinsku klasu , (ii) promjene u klasifikacijskom pragu ili (iii) promjene funkcije cilja. Navedene tehnike mogu se koristiti i u međusobnoj kombinaciji, a pritom se govori o hibridnom pristupu. Sve navedene tehnike na sebi svojstven

način adresiraju problem klasne neravnoteže i pridonose poboljšanju performansi standardnih klasifikatora kada se oni primjenjuju nad klasno neuravnoteženim podatcima (Oreski i Oreski, 2014).

Klasifikatori se u pravilu temelje na algoritmima strojnog učenja, a njihove performanse obično se procjenjuju prediktivnom točnošću konstruiranog modela. Međutim, ta mjera nije prikladna kada su podaci neuravnoteženi i/ili se troškovi različitih pogrešaka uvelike razlikuju (Mazurowski et al., 2008). U prilog toj tezi Chawla, Bowyer, Hall i Kegelmeyer (2002) navode da tipični mamografski skup podataka može sadržavati 98% normalnih piksela i 2% abnormalnih piksela. Jednostavna strategija pogađanja koja uvijek odabire većinsku klasu dala bi prediktivnu točnost od 98%. Međutim, da bi se postigao cilj klasifikacije, priroda stvari zahtijeva prilično visoku stopu prediktivne točnosti manjinske klase te dozvoljava malu stopu pogreške u većinskoj klasi. Stoga, standardna prediktivna točnost kao mjera očito nije prikladna u takvim situacijama.

ROC krivulja (engl. *Receiver Operating Characteristic curve*) je standardna tehnika za prikaz performansi klasifikatora koja daje uravnoteženu sliku odnosa točno klasificiranih primjera pozitivne klase u odnosu na ukupan broj pozitivnih primjera (engl. *true positive rate,* TPR) i netočno klasificiranih primjera negativne klase u odnosu na ukupan broj negativnih primjera (engl. *false positive rate,* FPR). Stoga je površina ispod ROC krivulje (engl. *Area Under the ROC Curve,* AUC) tradicionalno prihvaćena kao mjera performansi za neuravnotežene skupove (Chawla et al., 2002; Mazurowski et al., 2008). U literaturi se predlažu i mnoge druge metrike za prikaz performansi klasifikatora. Njihov odabir treba biti u skladu sa svrhom i ciljevima klasifikacije.

U ovom uvodu smo naveli samo neka pitanja vezana uz klasifikaciju kod klasno i troškovno neuravnoteženih skupova podataka. U nastavku ovog poglavlja su prezentirane glavne značajke paradigmi koje se tradicionalno primjenjuju prilikom klasifikacije kod takvih skupova podataka. **Cilj ovog poglavlja je istražiti utjecaj predstavljenih paradigmi na performanse klasifikatora** pri ocjeni rizika kod kredita građana na temelju analize provedene na dva skupa podataka. **U analizi performansi** GA-NN i HGA-NN klasifikatora sa ili bez primjene kompenzacijskih metoda za ublažavanje negativnih efekata klasne neravnoteže **naglasak je na analizi troškova pogrešne klasifikacije**.

Poglavlje je organizirano na sljedeći način. Odjeljak 5.2 detaljnije opisuje problem klasne neravnoteže te daje pregled literature povezane s tim problemom. Odjeljak 5.3 opisuje tehnike rješavanja problema klasne neravnoteže. Posebno su naglašene specifičnosti evaluacijskih mjera te tehnika validacije koje se koriste u uvjetima klasne neravnoteže.

Odjeljak 5.4 opisuje proširenja postojećih tehnika za selekciju atributa i klasifikaciju. U odjeljku 5.5 su prikazani rezultati eksperimenata te je dana evaluacija performansi i njihova komparacija s rezultatima prezentiranim u literaturi. Odjeljak 5.6 zaključuje ovo poglavlje.

5.2 Opis problema klasne neravnoteže i pregled literature

U području klasifikacije učestalo se pojavljuju podaci s klasnom neravnotežom. Glavna karakteristika ove vrste klasifikacijskih problema je da slučajevi jedne klase značajno nadmašuju u brojnosti slučajeve druge klase (López et al., 2013; Estabrooks, Jo i Japkowicz, 2004). U većini slučajeva, problem klasne neravnoteže je povezan s binarnom klasifikacijom, ali se pojavljuju i višeklasni problemi, kod čega može biti više manjinskih klasa što predstavlja još veći problem kod klasifikacije (López et al., 2013). Važno je napomenuti da je manjinska klasa obično ta za koju je najteže pribaviti instance iz stvarnih podataka te je prikupljanje tih podataka povezano s značajnim troškovima. Kod medicinskih dijagnoza to su pacijenti s određenom bolešću, kod čega drugu klasu predstavljaju zdravi pacijenti, kod kontrole kvalitete to su proizvodi s greškom nasuprot onih bez grešaka, u financijskim transakcijama, transakcije kod kojih ima elemenata prevara i pronevjera spram normalnih transakcija ili konačno, kod kredita to su loši krediti čiji je udio daleko manji od dobrih.

Pošto većina standardnih algoritama za strojno učenje pretpostavlja uravnotežen skup podataka za učenje s približno ujednačenim troškovima netočne klasifikacije, kod neuravnoteženih skupova ti algoritmi generiraju suboptimalne klasifikacijske modele, kod čega dobro klasificiraju većinske slučajeve, a manjinske instance učestalije klasificiraju netočno. Stoga, oni algoritmi koji postižu dobre rezultate kod standardne klasifikacije ne postižu nužno i najbolje performanse kod neuravnoteženih skupova za klasifikaciju (López et al., 2013). Postoji više razloga za takvo ponašanje:

- Standardni algoritmi učenja koriste točnost kao globalnu mjeru performansi koja usmjerava učenje tako da daje prednost većinskoj klasi.
- Klasifikacijska pravila koja predviđaju pozitivnu klasu su često vrlo specijalizirana i pokrivaju vrlo malo slučajeva te su odbačena u korist općenitijih pravila koja predviđaju negativnu klasu.
- Vrlo male skupine slučajeva manjinske klase mogu biti identificirane kao šum u podatcima te ih također algoritam za klasifikaciju može pogrešno odbaciti.
- Samo nekoliko primjera podataka sa šumom na manjinskoj klasi može značajno degradirati identifikaciju manjinske klase jer je manje zastupljena.

Kao što je u uvodu ovog poglavlja već naglašeno, u literaturi je prezentiran niz znanstvenih studija kojima je predmet istraživanja rješavanje navedenih problema. Ovdje će biti spomenuta neka istraživanja koja ilustriraju glavne karakteristike najpoznatijih i najviše primjenjivanih metoda za ublažavanje negativnih efekata klasne neravnoteže. Pored njih, biti će ukazano i na istraživanja koja koriste drugačije pristupe problemu i/ili daju drugačije zaključke.

Chawla, Bowyer, Hall i Kegelmeyer (2002) u svom radu zaključuju da njihova tehnika SMOTE, koja nad-uzorkuje manjinsku klasu, u kombinaciji s pod-uzorkovanjem većinske klase može postići bolju učinkovitost klasifikatora (u ROC prostoru) nego što se postiže promjenom omjera gubitka kod RIPPER tehnike ili promjenom prioriteta klasa kod naivnog Bayesovog klasifikatora.

Brennan (2012) istražuje, na skupovima podataka koji su klasno neuravnoteženi, različite oblike prijevara te navodi da su dva osnovna načina za prevladavanje problema klasne neravnoteže: podatkovne metode i algoritamske metode. Podatkovne metode uključuju pod-uzorkovanja, nad-uzorkovanja ili kombinirane pristupe uzorkovanju. Također istražuje SMOTE tehniku, koja koristi k-NN algoritam za umjetno sintetiziranje uzoraka manjinske klase. Algoritamske metode koje istražuje vezane su uz troškove netočne klasifikacije, a provodi ih pomoću Metacost metode i troškovno osjetljivih pragova (engl. *metacost thresholds*). Zaključuje da su podatkovne metode pokazale veću učinkovitost (u smislu F-vrijednosti), na klasno neuravnoteženim podatcima analiziranih skupova podataka.

Istraživanje na neuravnoteženim podatcima proveli su Dal Pozzolo i koautori (2013). Oni predlažu strategiju utrke (engl. *racing strategy*) za ubrzano i automatsko pronalaženje najbolje strategije za neuravnotežene podatke. Temeljna logika njihove tehnike je u paralelnom testiranju alternativnih strategija samo na podskupu podataka i progresivnom odustajanju od alternativa koje su značajno lošije. Testiranja alternativnih strategija se obavljaju statističkim testovima uz određenu razinu signifikantnosti. Zaključuju da najbolja strategija ovisi od primijenjenom algoritmu kao i korištenom skupu podataka.

Hulse i Khoshgoftaar (2009) prikazuju tehnike uzimanja uzoraka, koje se često koriste za ublažavanje štetnih utjecaja neuravnoteženih podataka. Specifičnost je njihove studije u tome što navodi da su njihovi zaključci kontradiktorni u odnosu na ranija istraživanja i zaključuje da su jednostavne tehnike uzorkovanja poput slučajnog pod-uzorkovanja općenito najučinkovitije za poboljšanje performansi modela konstruiranih na temelju podataka sa šumom i neuravnoteženih podataka.

Učenje iz neuravnoteženih skupova podataka predstavlja zamršen problem u pogledu modeliranja ali i s troškovnog, kao i aspekta optimalnog odnosa među klasama. Chawla, Cieslak, Hall i Joshi (2008) analiziraju utjecaj odnosa među klasama na troškove pogrešne klasifikacije. Za optimizaciju odnosa među klasama prije konstrukcije modela predlažu tehniku omotača koja pronalazi potrebni (optimalni) odnos među klasama ponavljanjem uzorkovanja i optimizacijom evaluacijske funkcije kao što je F-mjera, AUC, troškovi, krivulja troška i troškovno zavisna F-mjera.

Skupovi koji imaju visoko neuravnoteženu klasnu distribuciju predstavljaju temeljni izazov u strojnom učenju, ne samo s aspekta konstrukcije modela, nego i u pogledu načina procjene kvalitete konstruiranih modela. Kod toga postoji više različitih mjera vrednovanja modela koje se koriste u uvjetima klasne neravnoteže, svaka s vlastitim pristranostima. Također postoje različite strategije unakrsne validacije. Njihove karakteristike treba dobro poznavati, a odabrane evaluacijske mjere i strategije unakrsne validacije moraju biti u skladu s problemom koji se analizira (Raeder, Forman i Chawla, 2012).

Pored svega navedenog, klasno neuravnotežene skupove podataka u stvarnosti često prati još jedan problem - visoka dimenzionalnost. Kod tih skupova podataka posebno je važno odabrati atribute koji će dovesti do veće odjeljivosti među klasama. Odnosno, važno je odabrati atribute koji će dobro „uhvatiti“ visoku asimetriju u distribuciji klasa i omogućiti dobru diskriminaciju (Chawla, Japkowicz i Kotcz, 2004).

Zheng,Wu i Srihari (2004) ukazuju da postojeće tehnike koje se koriste za odabir atributa kod klasno uravnoteženih skupova podataka nisu vrlo prikladne za neuravnotežene skupove podataka. Oni predlažu okvir za izbor atributa koji odabire atribute za pozitivne i negativne klase zasebno, a zatim ih eksplicitno kombinira. Tim okvirom je objedinjeno nekoliko standardnih metoda te je predloženo niz novih metoda koje optimalno kombiniraju atribute za svaku klasu prema karakteristikama podataka i mjerama performansi. Autori također ukazuju da dobra tehnika za odabira atributa treba uzeti u obzir: (1) skup podataka, (2) metodu klasifikacije i (3) mjeru performansi.

Premda ima različitih prijedloga tehnika za rješavanje problema klasne neravnoteže, iz pregleda literature može se uočiti da glavnina tehnika koristi neki oblik podatkovnog ili algoritamskog pristupa. I dok većina studija zaključuje da podatkovni pristup daje bolje rezultate od algoritamskog, nema izrazito većinskog zaključka koja od podatkovnih tehnika je bolja. Većina autora navodi da je naduzorkovanje dalo bolje rezultate. Japkowicz (2000) navodi da na velikim domenama bolje rezultate daje poduzorkovanje. Sličan zaključak su dali ranije spomenuti Hulse i Khoshgoftaar (2009).

Ipak možemo navesti da je prednost podatkovnog pristupa pri rješavanju problema neravnoteže u podacima u tome što je to: (1) vanjski postupak koji je lako prenosiv, (2) primjenom tog postupka ne treba mijenjati ništa u implementaciji klasifikacijskog algoritma, a može se provesti čak i nekim drugim alatom i tada samo prenijeti rezultate. Upravo na taj način izvedeno je, u nastavku ovog poglavlja prezentirano, uzorkovanje SMOTE tehnikom u alatu Weka, a rezultati (uravnotežen skup) su preneseni u Rapid Miner. Kao glavni nedostatak podatkovnog pristupa se spominje problem podešavanja optimalnog odnosa među klasama i načina kako postići taj odnos jer nije poznato koji odnos nad ili pod uzorkovanja je optimalan (Estabrooks, Jo i Japkowicz, 2004; Chawla et.al., 2008; Dal Pozzolo et.al., 2013).

Zbog velike različitosti prikazanih tehnika i algoritama, kao i različitosti domena primjene, teško je jednostavno klasificirati sve postojeće pristupe ublažavanju problema vezanih uz troškovno osjetljivu klasifikaciju klasno neuravnoteženih podataka. Iz pregleda literature se može zaključiti da postoji relativno veliki interes za istraživanjem ovog problema, da su u istraživanjima uglavnom korišteni podatkovni i algoritamski pristup, rjeđe njihove kombinacije, a nismo pronašli istraživanja koja ovaj problem sagledavaju u kombinaciji s odabirom atributa i troškovima pogrešne klasifikacije. Stoga **istraživanje prezentirano u nastavku nastoji pokriti uočenu prazninu istražujući utjecaj tehnika uzorkovanja u kombinaciji s tehnikama selekcije atributa na rezultate klasifikacije i to prije svega na troškove pogrešne klasifikacije**. U istraživanju se kombiniraju tehnike uzorkovanja i ranije prezentirane tehnike selekcije atributa i klasifikacije, temeljene na genetskim algoritmima, GA-NN i HGA-NN u domeni odobravanja kredita građana.

5.3 Metodološke osnove problema klasne neravnoteže

Tipični klasifikatori kao što su stabla odlučivanja i neuronske mreže su dizajnirani tako da optimiraju ukupnu točnost klasifikacije bez obzira na relativnu distribuciju svake klase (Estabrooks, Jo i Japkowicz, 2004). U literaturi se navodi da neravnoteža odnosa između većinske i manjinske klase utječe na težinu klasifikacijskog zadatka, a osjetljivost na neravnotežu također raste rastom složenosti domene (Japkowicz, 2000b). Kao utjecajni faktori na složenost domene često se navode razina u kojoj se preklapaju klase od interesa (Dal Pozzolo et.al., 2013) te šum u podacima (Van Hulse i Khoshgoftaar, 2009).

Uz klasnu neravnotežu podataka, na ukupne performanse klasifikacije kod neuravnoteženih skupova podataka utječu i druge unutarnje karakteristike podataka. López i

koautori (López et al., 2013) navode šest značajnih problema, vezanih uz unutarnje karakteristike podataka, koji se moraju uzeti u obzir kako bi se osigurala bolja rješenja i pravilno identificirale klase kod klasifikacijskih problema:

1. Identifikacija područja s malim skupovima podataka (engl. *small disjuncts*).
2. Nedostatak gustoće (engl. *lack of density*) u podatcima za učenje.
3. Problem preklapanja između klasa (engl. *overlapping between the classes*).
4. Utjecaj šuma u podatcima (engl. *noisy data*) u neuravnoteženim domenama.
5. Značaj graničnih slučajeva (engl. *borderline instances*) za provođenje dobre diskriminacije između pozitivne i negativne klase, i njihov odnos s šumom u podacima.
6. Moguće razlike u raspodjeli podataka za trening i test, problem poznat kao pomak u podacima (engl. *dataset shift*).

Navedena pitanja predstavljaju izvor problema kod klasifikacije uravnoteženih podataka, a kod neuravnoteženih skupova podataka dobivaju još više na značajnosti. Stoga se kod dizajna tehnika za rješavanje problema klasne neravnoteže naročito mora voditi računa o navedenim karakteristikama podataka.

5.3.1 Tehnike rješavanja problema klasne neravnoteže

U dubinskoj analizi podataka i strojnom učenju se pitanju klase neravnoteže uglavnom pristupa na dva načina. Jedan je da se primjerima za učenje dodijele različiti ponderi, a drugi je ponovno uzorkovanje originalnih podataka (Chawla et al., 2002). Zaključci o tome koji je pristup bolji nisu uvijek isti, ali je u većini slučajeva učinkovitost podatkovnih tehnika bila bolja (Brennan, 2012; Van Hulse i Khoshgoftaar, 2009; Chawla et al., 2002). Podatkovni pristup je istraživaniji zbog toga što su tehnike tog pristupa nezavisne od korištenog klasifikatora i mogu biti lakše primijenjene na bilo koji problem (García, Marqués i Sánchez, 2012). Najpopularnije strategije podatkovnog pristupa sastoje se od ponovnog uzorkovanja podataka kako bi se dobila drugačija distribucija klasa. Uzorkovanje se u pravilu provodi sve dotle dok klase nisu približno jednako zastupljene.

Tehnike za ublažavanje problema u radu s neuravnoteženim skupovima podataka svi autori ne promatraju na isti način. Primjer drugačije klasifikacije iznesen je u studiji Dal Pozzolo et.al. (2013). Ta studija tehnike za rad s neuravnoteženim skupovima podataka

svrstava u sljedeće kategorije: temeljene na uzorkovanju, temeljene na udaljenostima i na metode ansambla.

Ipak, velika većina autora tehnike za ublažavanje problema klasne neravnoteže klasificira u sljedeće kategorije: tehnike uzorkovanja, tehnike temeljene na algoritamskom pristupu te na njihove kombinacije, odnosno, tehnike s hibridnim pristupom. Stoga u nastavku slijedi prikaz tehnika u skladu s navedenom podjelom. Počinjemo s tehnikama uzorkovanja jer većina studija koristi tehnike uzorkovanja za ublažavanje problema klasne neravnoteže (Cieslak i Chawla, 2008) uravnoteženjem broja uzoraka u manjinskoj i većinskoj klasi (Naganjaneyulu i Kuppa, 2013).

5.3.1.1 *Tehnike uzorkovanja*

Različite studije, čiji je predmet istraživanja klasifikacija neuravnoteženih skupova podataka, koriste različite varijante nad i pod uzorkovanja, a prezentiraju ponekad i proturječne stavove o korisnosti naduzorkovanja u odnosu na poduzorkovanje (Chawla, 2005).

Slučajno naduzorkovanje (engl. *random over-sampling,* ROS) manjinske klase s mogućnošću ponovnog izbora je najjednostavnija strategija koja slučajnim odabirom i replikacijom pozitivnih primjera uravnotežuje distribuciju klasa (García, Marqués i Sánchez, 2012). Ova tehnika povećava vjerojatnost prejake prilagodbe (engl. *overfitting*) primjerima manjinske klase jer pravi egzaktne kopije tih primjera.

Kod slučajnog naduzorkovanja manjinske klase može doći, kao u slučaju Brennan (Brennan, 2012), do izostavljanja određenih primjera manjinske klase iz originalnog skupa jer ih procedura slučajnog odabira nužno ne mora obuhvati sve. To dodatno smanjuje granice skupa manjinske klase što može dodatno povećati negativni utjecaj na klasifikacijske performanse. Stoga kod primjene ove tehnike treba voditi računa da se originalni skup primjera manjinske klase najprije uključi u cijelosti, a tek nakon toga da se pristupi naduzorkovanju potrebne razlike.

S ciljem prevladavanja prejake prilagodbe primjerima manjinske klase, Chawla i koautori (2002) su predložili upravljanu tehniku SMOTE (engl. ***S****ynthetic* ***M****inority* ***O****ver-sampling* ***TE****chnique*). Ta tehnika generira umjetne pozitivne primjere interpolirajući vrijednosti atributa postojećih najbližih primjera. To čini na način da najprije pronalazi *k* (autori tehnike u svom radu eksperimentiraju s parametrom *k* postavljenim na 5) najbližih susjeda iste (manjinske) klase, a nakon toga generira nove sintetičke primjere u smjeru nekog ili svih

najbližih susjeda što zavisi o količini zahtijevanih novih primjera. Pseudo kod SMOTE tehnike (Chawla et al., 2002) dan je u nastavku.

Algoritam SMOTE(T, N, k) opisan pesudo kodom

Ulaz: broj primjera manjinske klase ***T***;
SMOTE vrijednost ***N%***;
broj najbližih susjeda ***k***
Izlaz: (*N*/100) * *T* generiranih primjera manjinske klase

```
// Ako je N manji od 100%, slučajnim odabirom odaberi samo N% primjera manjinske klase.
if N < 100
// Slučajnim odabirom odaberi T primjera manjinske klase.
then
        T = (N/100) ∗ T
        N = 100
endif

// Za SMOTE vrijednost se očekuje da je cijeli broj pomnožen sa 100.
N = (int)(N/100)
k = broj najbližih susjeda
numattrs = broj atributa

Sample[ ][ ]: polje originalnih primjera manjinske klase
newindex: evidentira broj generiranih sintetičkih primjera, inicijaliziran na 0

// Prva dimenzija je primjer, druga je atribut.
Synthetic[ ][ ]: polje sintetičkih primjera

// Računa k susjeda za svaki primjer manjinske klase.
for i ← 1 to T
        Računa k najbližih susjeda za primjer i, i sprema njihove indekse u polje nnarray

        //Popuni N primjera za i-ti primjer, prema tablici indeksa nnarray.
        Popuni(N, i, nnarray)
endfor

// Funkcija generira sintetičke primjere.
Popuni (N, i, nnarray)
while N ≠0

        Odaberi slučajni broj između 1 and k, nazovi ga nn.
        Ovaj korak odabire jednog od k najbližih susjeda od i.
        for attr ← 1 to numattrs

                // nnarray[nn] je indeks susjeda.
                Izračunaj: dif = Sample[nnarray[nn]][attr] − Sample[i][attr]
.               Izračunaj: gap = slučajni broj između 0 i 1
                Synthetic[newindex][attr] = Sample[i][attr] + gap ∗ dif
        endfor
        newindex++
        N = N − 1
endwhile

// Kraj Popuni.
return
Kraj Pseudo koda.
```

U literaturi je prezentirano više modifikacija SMOTE tehnike. Među najpoznatije se ubraja granični SMOTE (engl. *borderline-SMOTE*) (Han, Wang i Mao, 2005). Ova tehnika odabire primjere manjinske klase za koje se utvrdi da su na granici manjinske regije u prostoru odlučivanja i izvodi SMOTE naduzorkovanje samo za te primjere, umjesto generiranja sintetičkih primjera za sve manjinske primjere.

Slučajno poduzorkovanje (engl. *random under-sampling*, RUS) većinske klase je tehnika uzorkovanja koja uklanja primjere većinske klase iz populacije sve dotle dok manjinska klasa ne postigne definiran postotak većinske klase.

Slučajno poduzorkovanje i naduzorkovanje imaju različite nedostatke. Metodom slučajnog poduzorkovanja potencijalno se može ukloniti neke važne primjere većinske klase, a slučajno naduzorkovanje može dovesti do prejake prilagodbe primjerima manjinske klase (Chawla, 2005). Japkowicz (2000) je istakla da su oba pristupa uzorkovanju učinkovita, ali je također primijetila da uzorkovanje pomoću sofisticiranih tehnika nije dalo jasnu prednost u odnosu na jednostavne tehnike u promatranoj domeni. Ipak valja naglasiti da tada nije bilo SMOTE tehnike. López i koautori (2013) također zaključuju da su sofisticiranije tehnike manje općenite, ciljano su razvijane za određenu grupu problema, a kada se uspoređuju u velikom broju referentnih problema u prosjeku mogu dati slabije performanse.

Tehnike uzorkovanja jesu najučestalije korištene tehnike za rješavanje problema vezanih uz klasnu neravnotežu i različite troškove netočne klasifikacije, ali ne i jedine. Za rješavanje ove vrste klasifikacijskih problema koriste se i tehnike koje se svrstaju u algoritamske.

5.3.1.2 *Algoritamski pristup*

Algoritamski pristup je orijentiran na prilagodbu temeljnih metoda učenja tako da budu više usredotočene na pitanja vezana uz klasnu neravnotežu (López et al., 2013). Centralno pitanje kod algoritamskog pristupa troškovno osjetljivom učenju je; kako u fazi konstrukcije modela uzeti u obzir troškove netočne klasifikacije. Pored faze konstrukcije modela postoje slučajevi kada u fazi korištenja modela treba prilagođavati klasifikaciju promjenama (zahtjevima) korisnika.

Algoritamski pristup utječe na rezultate klasifikacije na neki od sljedećih načina, tj. kroz:

1) promjene u temeljnom algoritmu,

2) nejednake težine za instance klasa kako bi usmjerio klasifikator da obrati više pozornosti na manjinsku klasu ili
3) promjene u klasifikacijskom pragu.

Promjena u temeljnom algoritmu podrazumijeva direktnu modifikaciju algoritma za učenje. Na primjer, kod neuronskih mreža s propagacijom greške unazad možemo u algoritam za izračun greške uključiti i tablicu troška netočne klasifikacije, a time direktno utjecati na pondere u mreži. Isto tako kod stabla odlučivanja prilikom konstrukcije stabla možemo minimizirati trošak netočne klasifikacije odabirom atributa koji s troškovnog aspekta najbolje dijeli podatke ili uvažiti tablicu troška prilikom obrezivanja stabla.

Promjena svih temeljnih algoritama dostupnih u literaturi da budu troškovno osjetljivi bio bi vrlo dugotrajan i složen zadatak. Stoga je Domingos (Domingos, 1999) došao na ideju da se napravi jedna procedura koja će konvertirati široki spektar algoritama temeljenih na minimizaciji greške u troškovno osjetljive. Proceduru je nazvao MetaCost, a zasniva se na omotaču oko algoritma temeljenog na minimizaciji greške na takav način da klasifikator minimizira trošak. Ideja MetaCost algoritma kao i ideja cjelokupne troškovno osjetljive klasifikacije nije načiniti što manje grešaka u klasifikaciji, odnosno postići maksimalnu točnost, nego konstruirati takav model koji će producirati što manji trošak netočne klasifikacije. Znači griješiti tako da je najmanje skupo. Ciljevi, minimizacija greške i minimizacija troška, nisu nužno isti, štoviše, kod neuravnoteženih skupova najčešće nisu.

MetaCost tehnika procjenjuje pripadnost određene instance nekoj klasi učenjem većeg broja klasifikatora te njihovim glasanjem određuje klasu za svaki primjer, tj. koristi jednu varijantu bagging tehnike.

Promjena u klasifikacijskom pragu se može, kao pristup, koristiti prilikom konstrukcije modela. Tako se kod klasifikatora konstruiranog pomoću genetskog algoritma učenje usmjerava uz prilagodbu određenom klasifikacijskom pragu (Oreški, Oreški i Oreški, 2012). Ipak, promjena u klasifikacijskom pragu se više koristi kod već konstruiranog modela radi prilagodbe ponašanja modela potrebama korisnika.

5.3.1.3 *Hibridne tehnike*

Ove tehnike uključuju pristupe na podatkovnoj i algoritamskoj razini, kao i njihove kombinacije tako da pri tome minimiziraju troškove netočne klasifikacije (López et al., 2013). Razlog korištenja hibridnih tehnika može biti optimizacija klasifikacije u odnosu na bilo koju mjeru performansi.

Uključivanjem hibridnih tehnika, proces konstrukcije klasifikacijskog modela postaje kompleksniji. Kao najjednostavniji oblik korištenja hibridnih tehnika za ublažavanje negativnog utjecaja klasne neravnoteže možemo navesti scenarij u kojem se za konstrukciju modela koristi neka od tehnika ponovnog uzorkovanja radi ujednačavanja klasne zastupljenosti, a kod eksploatacije modela se dodatno korigira klasifikacijski prag.

5.3.2 **Evaluacijske mjere**

Kod odabira atributa za klasifikaciju, kvaliteta tehnike za odabir atributa mjeri se kvalitetom konstruiranog modela. Pored kvalitete podataka, kvaliteta konstruiranog modela ovisi o kvaliteti algoritma za klasifikaciju kao i o odabranoj evaluacijskoj mjeri. Naime, nećemo dobiti isti pokazatelj kvalitete nekog modela mjerimo li tu kvalitetu mjerom A ili B. U nastavku ovog odjeljka proširujemo skup evaluacijskih mjera prezentiranih u poglavlju 3 onim mjerama koje se standardno koriste kod klasifikacije na neuravnoteženim skupovima podataka.

Koja od mjera za vrednovanje rezultata će se koristiti u određenom slučaju ovisi o prirodi problema i ciljevima klasifikacije. Točnost je dobra integralna evaluacijska mjera ako je podjednak broj pozitivnih i negativnih primjera te ako je trošak netočne klasifikacije pozitivnih i negativnih primjera podjednak. Točnost klasifikacije je mjera performansi koja najčešće usmjerava algoritme strojnog učenja, a prema matrici konfuzije (Tablica 3.3), definira se kao:

$$\text{Točnost (\%)} = (TP + TN) / (TP + FP + TN + FN) * 100.$$

Točnost predviđanja razborito je koristiti kao mjeru kvalitete konstruiranog modela u kontekstu uravnoteženih skupova podataka i jednakih troškova grešaka (Chawla et al., 2002). Kada nisu zadovoljeni navedeni uvjeti, tada se često (López et al., 2012) kao mjera kvalitete modela koristi površina ispod ROC krivulje, AUC (López et al., 2013), koja se računa:

$$AUC = (1 + TPR - FPR) / 2.$$

Omjeri TPR i FPR se računaju prema sljedećim izrazima:

$$TPR = TP / (TP + FN) \text{ i}$$

$$FPR = FP / (TN + FP).$$

Za ROC krivulju se može reći da najbolje prikazuje odnos među omjerima: TPR i FPR. X-os predstavlja FPR, a Y-os predstavlja TPR. Za omjer TPR često se koristi naziv odziv.

Idealna točka na ROC krivulji je (0;1,00), u toj točci su svi pozitivni primjeri klasificirani ispravno (pozitivno) i nema krivo klasificiranih negativnih primjera. Linija y = x predstavlja scenarij slučajnog pogađanja klase. Površina ispod ROC krivulje (AUC) je korisna mjera kvalitete klasifikatora jer je neovisna o odabranom kriteriju odlučivanja i o apriornoj vjerojatnosti (Chawla et al., 2002). Nezavisnost o apriornoj vjerojatnosti i nezavisnost od kriteriju odlučivanja nije uvijek poželjna. Dapače, želimo li konstruirati klasifikator koji će biti najbolje prilagođen konkretnom kriteriju odlučivanja onda nam treba takva mjera koja će taj kriterij najbolje izraziti i temeljem koje ćemo donijeti ispravnu odluku o najboljem klasifikatoru. Za određene distribucije klasa i troškova, klasifikator s najboljom površinom ispod ROC krivulje može biti suboptimalan. Zbog toga, pored AUC, kao dodatne mjere kvalitete klasifikatora u nastavku koristimo F-vrijednost i funkciju troškova.

Dok ROC krivulja grafički predstavlja odnos među omjerima TPR i FPR, F-vrijednost predstavlja odnos između vrijednosti TP, FP i FN. F-vrijednost je mjera koja balansira odnos između preciznosti i odziva, a rezultat je broj koji odražava "dobrotu" klasifikatora kod manjinskih primjera. Izraz za F-vrijednost je kao što slijedi (Han, Kamber i Pei, 2006; Han, Wang i Mao, 2005)::

$$\text{F-vrijednost} = (2 * \text{preciznost} * \text{odziv}) / (\text{preciznost} + \text{odziv}),$$

kod čega je izraz za izračun odziva jednak izrazu za TPR, a izraz za izračun preciznosti je

$$\text{preciznost} = TP/(TP+FP).$$

Postoji i ponderirana mjera odnosa preciznosti i odziva:

$$F\text{-}_{\beta} = ((1+ \beta^2) * \text{preciznost} * \text{odziv}) / (\beta^2 * \text{preciznost} + \text{odziv}),$$

gdje β odgovara relativnoj važnosti odziva nasuprot preciznosti.

5.3.3 **Tehnike validacije**

Kao što smo u prethodnom odjeljku naglasili, kod klasno neuravnoteženih skupova podataka važno je koristiti primjerene mjere za mjerenje kvalitete konstruiranog modela. Jednako tako je važno odabrati i odgovarajuće tehnike validacije performansi modela. Za ocjenu performansi konstruiranog modela u pravilu se koristi testni skup podataka, a ne skup za učenje. Skup za testiranje dati će pravu, nepristranu procjenu klasifikacijske greške ako je on reprezentativan uzorak populacije i ako nije korišten za učenje. Budući da su nam rijetko kad dostupna dva različita skupa podataka o istoj populaciji, nužno je odlučiti kako podijeliti dostupni skup podataka na skup za učenje i skup za testiranje.

Najpoznatije tehnike za validaciju performansi klasifikatora su (Kohavi, 1995):

- podjela skupa na dva dijela (engl. *holdout procedure*),
- unakrsna validacija s *k* preklapanja (engl. *k-fold cross-validation*)
- uzorkovanje s ponavljanjem (engl. (i dalje): *bootstrap*)

Zbog toga što je skup podataka kojim raspolažemo samo uzorak podataka na kojemu treba konstruirati i testirati model, koji će se zatim primijeniti na čitavoj domeni, iznimno je važno da je uzorak reprezentativan, što znači da na najbolji mogući način predstavlja čitavu populaciju.

Ako je dostupnih podataka mnogo, više nego ih je moguće učinkovito obraditi, što je rjeđe slučaj, potrebno je izdvojiti uzorak podataka koji će se analizirati. Najbolja tehnika odabira dijela podataka u uzorak je slučajno uzorkovanje. Slučajno uzorkovanje se koristi i za podjelu uzorka podataka na skup za učenje i testiranje.

Kao i cijeli uzorak, tako i skup za učenje i skup za testiranje moraju biti reprezentativni, odnosno, moraju na najbolji mogući način predstavljati čitavu populaciju.

Podjela skupa na dva dijela kod manjih uzoraka se nije pokazala najboljom tehnikom (Kohavi, 1995; Malhotra i Malhotra, 2003). U praksi je skup kojim raspolažemo ograničene veličine i obično manji nego što bismo željeli da bude. Tehnika podjele skupa na dva dijela neefikasno koristi podatke jer se dio podataka skupa ne koristi za učenje algoritma za klasifikaciju (Kohavi, 1995), a pokazala se i previše optimističnom. Tehnika unakrsne validacije s *k* preklapanja se pokazala, u većini slučajeva, najboljom tehnikom validacije performansi modela (Kohavi, 1995; Malhotra i Malhotra, 2003).

Težina odabira odgovarajuće tehnike validacije u uvjetima klasne neravnoteže proizlazi iz činjenice da je na tim skupovima, radi ublažavanja problema vezanih uz klasnu neravnotežu, najčešće korištena neka od tehnika ponovnog uzorkovanja. Nakon njihove

primjene, tehnike unakrsne validacije s *k* preklapanja za procjenu performansi klasifikatora treba koristiti s oprezom (Brennan, 2012) i to zbog toga što je, prije primjene algoritma za indukciju modela, skup za učenje i testiranje promijenjen. Stoga nije reprezentativan, odnosno, ne predstavlja na najbolji mogući način čitavu populaciju. Malhotra i Malhotra (Malhotra i Malhotra, 2003) navode da se za cjelovito testiranje ukupne prediktivne sposobnosti nekog modela za neku nepoznatu populaciju treba koristiti cijeli originalni uzorak podataka. Upravo kod neuravnoteženih uzoraka podataka, nakon primjene različitih tehnika ponovnog uzorkovanja za uravnoteženje uzorka podataka, uzorak više nije reprezentativan u odnosu na čitavu populaciju stoga je tada najbolji predstavnik cijele populacije originalni uzorak podataka, prije uravnoteženja. U ovom poglavlju se koristi tehnika unakrsne validacije s 10 preklapanja kod učenja, a za provjeru valjanosti prediktivne sposobnosti modela koristi se cijeli originalni uzorak podataka.

Kao drugi način prevladavanja razlike u raspodjeli podataka između originalnog uzorka podataka i uzorka koji je rezultat primijene neke od tehnika ponovnog uzorkovanja može se koristiti podjela skupa prije nad ili pod uzorkovanja, na dio za učenje i dio za test, u odnosu 70:30 (Brennan, 2012). Nakon toga se uravnoteženje primjeni samo na skup za učenje. Ovakvim pristupom eksperimentu, još uvijek je dopušteno da se za konstrukciju modela, odnosno učenje, koristi unakrsna validaciju s 10 preklapanja, a 30% podataka koji se ne koriste za indukciju modela može se koristiti za procjenu performansi modela (Brennan, 2012). Dakako, u tom slučaju manifestiraju se nedostatci svojstveni holdout metodi validacije.

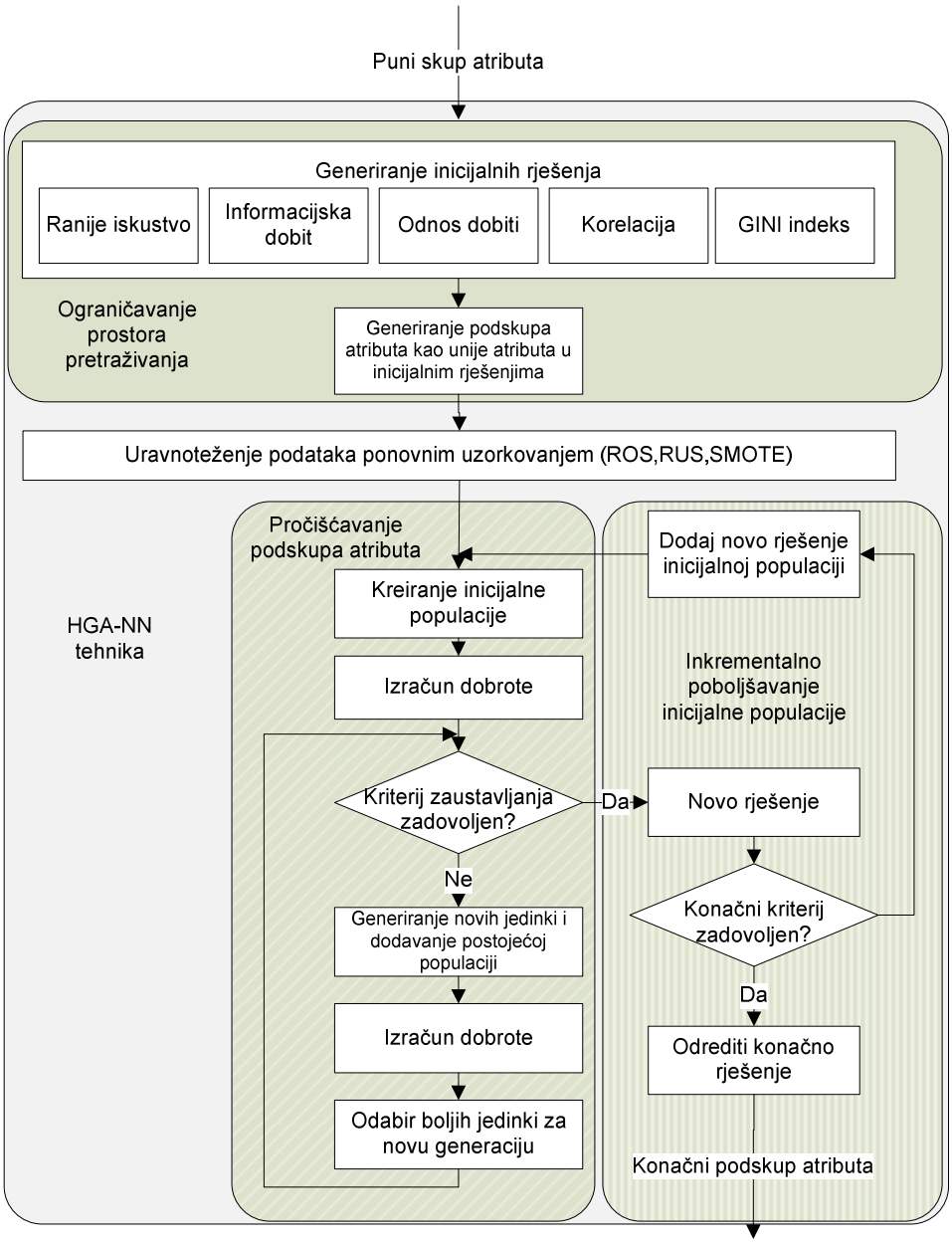
5.4 Razvoj modela

Klasna neravnoteža i različiti troškovi netočne klasifikacije predstavljaju značajne izazove za samu klasifikaciju te se posljedično manifestiraju i na algoritam za izbor atributa i klasifikaciju, HGA-NN. To je vidljivo iz rezultata prezentiranih u prethodnom poglavlju; HGA-NN algoritam postiže vrlo kvalitetne rezultate u pogledu preciznosti klasifikacije, ali ne ostvaruje jednako kvalitetne rezultate s osnova relativnog troška pogrešne klasifikacije. Stoga je u ovaj algoritam, temeljen na genetskim algoritmima, potrebno ugraditi tehnike koje će poboljšati performanse tog algoritma u odnosu na relativni trošak pogrešne klasifikacije, kao i u odnosu na neke druge mjere kvalitete klasifikacije svojstvene klasnoj neravnoteži.

Utvrđeno je da ponovno uzorkovanje općenito poboljšava performanse induciranog modela kod troškovno osjetljive klasifikacije nad neuravnoteženim podacima (López et al., 2013; Japkowicz, 2000). Uvažavajući upravo navedeno, kao i sve do sada rečeno u ovom

poglavlju, podatkovne tehnike koje ublažavaju problem klasne neravnoteže ugrađujemo u HGA-NN algoritam.

U nastavku ovog poglavlja je prikazana tehnika za odabir atributa i klasifikaciju podataka HGA-NN, proširena određenim tehnika ponovnog uzorkovanja s ciljem poboljšanja performansi te tehnike u odnosu na mjere kvalitete klasifikacije svojstvene klasnoj neravnoteži. Istovremeno su izvedena proširenja i na GA-NN tehnici radi usporedbe rezultata.



Slika 5.1 HGA-NN tehnika proširena ponovnim uzorkovanjem

Iz slike 5.1 je vidljivo da je ranije prezentirana tehnika HGA-NN proširena dodatnom fazom, ponovnim uzorkovanjem radi uravnoteženja podataka. Ponovno uzorkovanje nekom od ranije opisanih tehnika uzorkovanja, ROS, RUS ili SMOTE, izvodi se prije pročišćavanja skupa atributa. Na taj način se algoritmu za klasifikaciju, koji se koristi za izračun dobrote jedinki u populaciji, prezentira uravnotežen skup podataka. Time se postiže okolina u kojoj se minimizira slaba strana većine algoritama za klasifikaciju, a to je njihova pristranost većinskoj klasi.

S obzirom na činjenicu da GA-NN tehnika nema preliminarnu redukciju prostora pretraživanja, kod nje se ROS, RUS ili SMOTE izvode kao prva faza proširene tehnike. Pri tome se bez obzira na primijenjenu tehniku uzorkovanja, kao i bez obzira na to dali je riječ o GA-NN ili HGA-NN tehnici, uzorkovanjem nastoji dobiti klasna ravnoteža na podacima za učenje. U prvom redu zbog toga što troškovno osjetljivo učenje pretpostavlja da je matrica troškova poznata po različitim vrstama pogrešaka. Međutim, često matrica troškova nije poznata (Sun et al., 2007), što je i ovdje slučaj. Stoga se u radu u svim slučajevima najprije pokušava dobiti ravnoteža u podacima za učenje u blizini odnosa 50:50, što je teorijski gledano najbolji odnos za većinu algoritama za klasifikaciju, a nakon toga se testira niz troškovnih odnosa. Uspostava odnosa različitog od 50:50, bez da je unaprijed točno definirana matrica troškova za koju se želi postići optimum, s konceptualnog aspekta nije opravdana.

Kroz eksperiment nam je cilj pokazati da primjenom opisane tehnike možemo postići, kod algoritma za klasifikaciju, ravnopravni tretman različitih klasa što se manifestira kroz mjere za vrednovanje rezultata, prije svega AUC. Nakon konstrukcije modela se može, za određeni odnos troškova, kontrolirano mijenjati klasifikacijski prag kako bi se polučili najbolji rezultati za ciljni odnos troškova. Takvim pristupom možemo relativno brzo, kada u konkretnom slučaju dobijemo zadatak da treba izvršiti optimizaciju rezultata u skladu s određenom matricom troškova, to i postići.

Algoritam prikazan na slici 5.1 je napravljen korištenjem Rapid Miner 5.1.15 alata s parametrima prikazanim u tablici 5.1. U tablici 5.1 su prikazani parametri za (H)GA i neki parametri za NN, komponente GA-NNa i HGA-NNa. Parametri se nisu mijenjali kroz eksperiment, zadržani su najbolji parametri postignuti u eksperimentima prezentiranim u prethodnom poglavlju na navedenim skupovima podataka i s navedenim algoritmima.

Tablica 5.1 Sažetak (H)GA i NN parametara

Parametar	**Promjena**	**Postavke**
Initializacija populacije		
population size		50
initial probability for an feature to be switched on	✓	0.6
maximum number of features		12
minimum number of features		5
Reprodukcija		
Fitness measure		accuracy
Fitness function		neural network
the type of neural network		multilayer feed-forward network
network algorithm		backpropagation
activation function		sigmoid
the number of hidden layers		1
the size of the hidden layer		(number of features + number of classes) / 2 + 1
training cycles	✓	500
learning rate		0.6
momentum		0.2
selection scheme	✓	tournament
tournament size	✓	0.05
dynamic selection pressure		Yes
keep best individual		Yes
mutation probability		0.1
crossover probability		0.9
crossover type	✓	one_point
Uvjeti završetka		
maximal fitness		Infinity
maximum number of generations		10
use early stopping		No

Na razlike rezultata između eksperimenata će, zadržavanjem istih parametara, utjecati samo pomoćne tehnike koje ublažavaju utjecaj klasne neravnoteže i različitih troškova klasifikacije na rezultate klasifikacije, a njihov će se doprinos lako moći kvantificirati i to za oba algoritma, GA-NN i HGA-NN.

Uspoređujući parametre s onima iz prethodnog poglavlja za (H)GA i NN vidimo da je promijenjen maksimalni broj generacija genetskog algoritma s 50 na 10. To je učinjeno s ciljem smanjenja potrebnog vremena za eksperiment te je u skladu s zaključcima prethodnog poglavlja da je poboljšanje točnosti algoritama u prvih samo 4 generacije veće nego u 40 generacija, odnosno od 10-te do 50-te generacije. S obzirom da je istraživanje utjecaja pomoćnih tehnika za ublažavanje utjecaja klasne neravnoteže i različitih troškova netočne klasifikacije na rezultate klasifikacije glavni cilj ovog eksperimenta, taj cilj je moguće postići i s manjim brojem generacija GA.

5.4.1 Evaluacija i komparacija rezultata

Kao što smo ranije naveli, nakon primjene različitih tehnika uzorkovanja, uzorak podataka je uravnotežen, ali više nije reprezentativan u odnosu na čitavu populaciju. Na tako uravnoteženom skupu izvodi se konstrukcija modela, tj. učenje, uz korištenje tehnike unakrsne validacije s 10 preklapanja. Na kraju se za provjeru valjanosti prediktivne sposobnosti konstruiranog modela koristi cijeli originalni uzorak podataka, prije uravnoteženja.

Komparacija rezultata, dobivenih tehnikama GA-NN i HGA-NN u odnosu na rezultate dobivene proširenim tehnikama GA-NN i HGA-NN s ciljem utvrđivanja utjecaja uravnoteženja skupova na rezultate, je moguća zbog toga što se parametri nisu mijenjali kroz eksperiment, samo su dodane pomoćne tehnike koje ublažavaju utjecaj klasne neravnoteže i različitih troškova klasifikacije na rezultate klasifikacije. Iz navedenog proizlazi da su razlike u rezultatima isključiva posljedica proširenja navedenih tehnika.

5.5 Empirijska analiza

U empirijskom dijelu, cilj je istražiti i dati procjenu efekata opisanih proširenja tehnika HGA-NN i GA-NN na performanse tih tehnika pri ocjeni rizika kod kredita građana na dva skupa podataka, hrvatskom i njemačkom. Pri tome se procjenjuju efekti djelovanja

kompenzacijskih tehnika, mjereno različitim mjerama performansi primjerenih primjeni u uvjetima klasne neravnoteže s naglaskom na relativni trošak netočne klasifikacije.

Općenito uzevši, klasifikatori se u pravilu temelje na algoritmima strojnog učenja, a njihove performanse obično se procjenjuju prediktivnom točnošću konstruiranog modela. Prediktivna točnost klasifikatora je također korištena u prethodnim eksperimentima u ovom radu. Međutim, koja od mjera za vrednovanje rezultata će se koristiti u određenom slučaju ovisi o prirodi problema i ciljevima klasifikacije. Ako je cilj klasifikacije optimiranje troška pogrešne klasifikacije, prediktivna točnost kao mjera je manje prikladna ako se troškovi različitih pogrešaka uvelike razlikuju (Mazurowski et al., 2008). Površina ispod ROC krivulje, AUC, je tradicionalno prihvaćena kao mjera performansi za neuravnotežene skupove (Chawla et al., 2002; Mazurowski et al., 2008), stoga je i mi koristimo u nastavku. Pored nje koristimo i F-vrijednost, koja predstavlja odnos između vrijednosti TP, FP i FN. Također koristimo F-β vrijednost koja predstavlja ponderiranu mjeru odnosa preciznosti i odziva gdje β odgovara relativnoj važnosti odziva nasuprot preciznosti. Sumjervljiva mjera F-β vrijednosti je relativni trošak netočne klasifikacije koji koristimo kao glavnu mjeru kvalitete klasifikacijskog algoritma s obzirom da nam je cilj optimirati trošak pogrešne klasifikacije.

U nastavku su prikazani rezultati eksperimenata, a skupovi podataka, hrvatski i njemački, korišteni u ovom poglavlju, korišteni su i opisani u ranijim poglavljima.

5.5.1 **Rezultati na hrvatskom skupu podataka**

Iz rezultata prikazanih u tablici 5.2 kao i na slikama 5.2, 5.3 i 5.4 se vidi da kada se ne koristi nijedna tehnika uzorkovanja podataka, tj. ne koriste se proširene tehnike, dobivaju se modeli s većom točnošću, ali na ostalim mjerama performansi se dobivaju lošiji rezultati.

Kada promatramo rezultate po primijenjenim tehnikama uzorkovanja, vidimo da ROS uzorkovanje daje, osim po točnosti, najbolje rezultate po svim ostalim mjerama performansi. Također je vidljivo da obje tehnike HGA-NN i GA-NN, na hrvatskom skupu podataka, proširene tehnikama uzorkovanja SMOTE daju vrlo slične rezultate po svim mjerama performansi. Rezultati dobiveni proširenjem RUS uzorkovanjem pokazuju najveće oscilacije po pitanju kvalitete konstruiranih modela.

Promatrajući rezultate po osnovnim tehnikama, rezultati HGA-NN tehnike ukupno su bolji u odnosu na GA-NN.

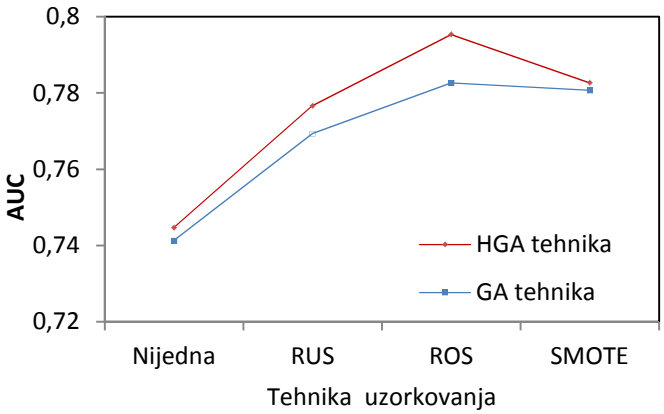
Uspoređujući krivulje sa slika 5.2, 5.3 i 5.4 s ukupnom usporedbom relativnog troška HGA-NN i GA-NN tehnike sa i bez korištenja tehnika uzorkovanja na slici 5.5, možemo zaključiti da su F-β krivulje vrlo slične krivuljama troškovnih funkcija. Iz toga proizlazi da navedene mjere daju konzistentnu sliku kvalitete klasfikatora s troškovnog aspekta. Sličnost se očituje u tome što su krivulje HGA-NN i GA-NN tehnike međusobno vrlo blizu, dok istovremeno iskazuju vrlo veliku razliku u odnosu na primjenu/neprimjenu tehnika uzorkovanja.

Tablica 5.2 Rezultati HGA-NN i GA-NN tehnike nakon različitih tehnika uzorkovanja na hrvatskom skupu podataka

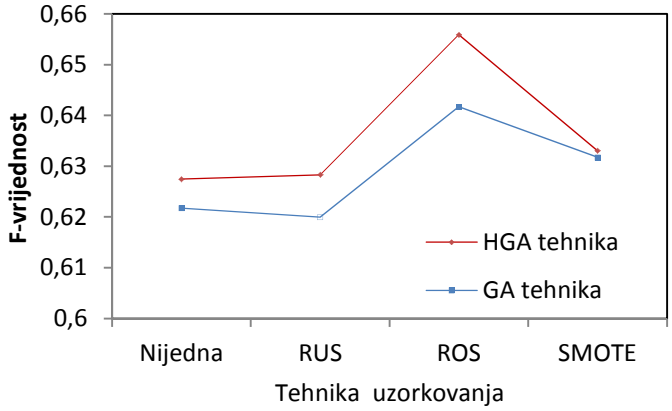
Tehnika uzorkovanja	Točnost	AUC	F-vrij.	F-β	TP	FN	FP	TN
Rezultati GA-NN tehnike								
Nijedna	82,6	0,7413	0,6217	0,5909	143	107	67	683
RUS	75,6	0,7693	0,6199	0,7148	199	51	193	557
ROS	78,0	0,7827	0,6417	0,7221	197	53	167	583
SMOTE	76,1	0,7807	0,6317	0,7327	205	45	194	556
Rezultati HGA-NN tehnike								
Nijedna	**82,9***	0,7447	0,6274	0,5955	144	106	65	685
RUS	76,1	0,7767	0,6283	0,7251	202	48	191	559
ROS	78,7	**0,7953***	**0,6559***	**0,7414***	203	47	166	584
SMOTE	76,0	0,7827	0,6330	0,7372	207	43	197	553

* označava najbolji rezultat za svaku od mjera

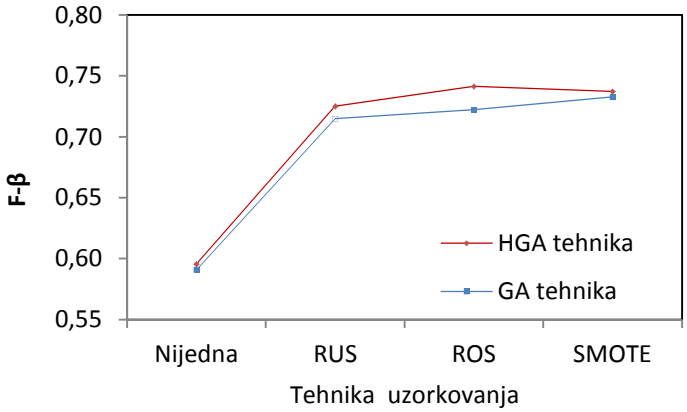
Rezultati prikazani u tablici 5.2 su u nastavku prikazani linijskim dijagramima na slikana 5.2 do 5.4.



Slika 5.2 Komparacija AUC rezultata HGA-NN i GA-NN tehnike za različite tehnike uzorkovanja na hrvatskom skupu podataka



Slika 5.3 Komparacija F-vrijednosti HGA-NN i GA-NN tehnike za različite tehnike uzorkovanja na hrvatskom skupu podataka



Slika 5.4 Komparacija F-$_\beta$ vrijednosti, tehnika HGA-NN i GA-NN, za različite tehnike uzorkovanja na hrvatskom skupu podataka

Kao što smo već ranije naveli, kod klasno neuravnoteženih skupova podataka trošak netočne klasifikacije manjinske klase je obično veći od troška netočne klasifikacije većinske klase. Zbog toga najviša točnost klasifikacije ne znači i troškovno optimalni rezultat. U takvim uvjetima troškovi netočne klasifikacije mogu biti optimalni kod manje točnosti. Troškovni optimum zavisi o odnosu između troškova tipa I i tipa II, a taj odnos kod kredita građana ovisi o ekonomskim ciklusima, dostupnosti kapitala na tržištu, nekim drugim okolnostima te sklonostima banke. Pri tome se trošak tipa I pojavljuje kao posljedica odobravanja kredita lošim klijentima, a trošak tipa II kao oportunitetni trošak zbog neodobravanja kredita dobrim klijentima. Troškovno gledano pogoršanje točnosti predviđanja je opravdano sve dotle dok smanjenje troška zbog smanjenja greške tipa I ima za posljedicu manje povećanje troška zbog povećanja greške tipa II.

Prema formuli 3.20 izračunati su i prikazani u tablici 5.3 ukupni relativni troškovi netočne klasifikacije (engl. *relative cost of misclassification*, RC) svakog modela za sedam odnosa troška tipa I i tipa II, kod čega je za svaki odnos najbolji model s najmanjom vrijednošću relativnog troška. Kao što smo i u ranijim analizama utvrdili, iz tablice 5.3 je također vidljivo da je najtočnija klasifikacija za banku najbolja samo u slučaju kada je odnos troška greške tipa I i II jednak. Već kod odnosa troška greške tipa I i II od 2:1, najtočnija klasifikacija ne mora biti i troškovno najpovoljnija. Što se više odnos troška povećava, to više u ovom eksperimentu najtočnija klasifikacija zaostaje u troškovnom pogledu u odnosu na modele koji su koristili neku vrstu tehnika uzorkovanja.

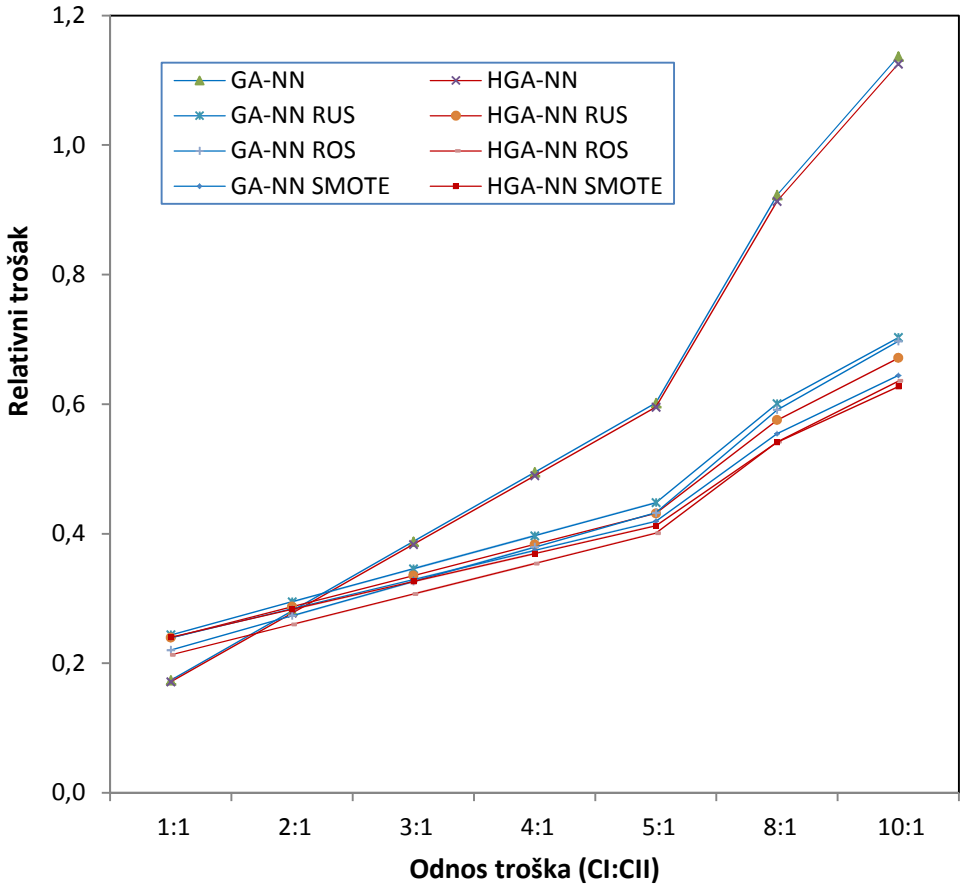
Tablica 5.3 Komparacija relativnih troškova HGA-NN i GA-NN tehnike nakon primjene različitih tehnika uzorkovanja na hrvatskom skupu podataka

Tehnika uzorkovanja	Troškovni odnos (C_I:C_{II})						
	1:1	2:1	3:1	4:1	5:1	8:1	10:1
	Rezultati GA-NN tehnike						
Nijedna	0,1740	0,2810	0,3880	0,4950	0,6020	0,9230	1,1370
RUS	0,2440	0,2950	0,3460	0,3970	0,4480	0,6010	0,7030
ROS	0,2200	0,2730	0,3260	0,3790	0,4320	0,5910	0,6970
SMOTE	0,2390	0,2840	0,3290	0,3740	0,4190	0,5540	0,6440
	Rezultati HGA-NN tehnike						
Nijedna	**0,1710***	0,2770	0,3830	0,4890	0,5950	0,9130	1,1250
RUS	0,2390	0,2870	0,3350	0,3830	0,4310	0,5750	0,6710
ROS	0,2130	**0,2600***	**0,3070***	**0,3540***	**0,4010***	0,5420	0,6360
SMOTE	0,2400	0,2830	0,3260	0,3690	0,4120	**0,5410***	**0,6270***

* označava najbolji rezultat u svakom troškovnom odnosu

Želi li banka optimirati odobravanje svojih kredita uz neki odnos troška različit od 1:1, iz rezultata ovog eksperimenta proizlazi da će vjerojatno za konstrukciju modela koristiti neku od tehnika proširenu pomoćnim tehnikama uzorkovanja. U našem slučaju na hrvatskom skupu podataka rezultati pokazuju da HGA-NN tehnika s ugrađenim tehnikama uzorkovanja daje s troškovnog aspekta bolje rezultate nego ta ista tehnika bez proširenja s tehnikom za dodatno uzorkovanje. Kod toga se ROS tehnika pokazala kao najbolja za omjere troškova 2:1, 3:1, 4:1 i 5:1, a SMOTE kao najbolja za omjere troškova 8:1 i 10:1.

Uspoređujući rezultate HGA-NN i GA-NN tehnike zaključujemo da HGA-NN tehnika daje bolje rezultate, konzistentno sa i bez dodatnog uzorkovanja.



Slika 5.5 Komparacija relativnog troška HGA-NN i GA-NN tehnike sa i bez korištenja tehnika uzorkovanja na hrvatskom skupu podataka

Na slici 5.5 je prikazana komparacija relativnog troška HGA-NN i GA-NN tehnike za sve konstruirane modele na jednom dijagramu. Premda bismo mogli prigovoriti da je dijagram nepregledan, ipak je vrlo koristan jer zorno prikazuje da bez korištenja pomoćnih tehnika uzorkovanja oba algoritma postižu najbolju točnost predikcije što je vidljivo iz najmanjeg relativnog troška za odnos troškova 1:1. Da najtočniji modeli postaju granično dobri s troškovnog aspekta kod odnosa troška 2:1, a da za sve odnose troškova iznad 2:1 modeli koji su konstruirani bez korištenja tehnika uzorkovanja na hrvatskom skupu podataka daju lošije rezultate, tj. viši relativni trošak. Linije troškova za te modele su najstrmije, što

znači da uz povećanje relativnog odnosa troškova pokazuju najbrži rast ukupnih relativnih troškova.

5.5.2 **Rezultati na njemačkom skupu podataka**

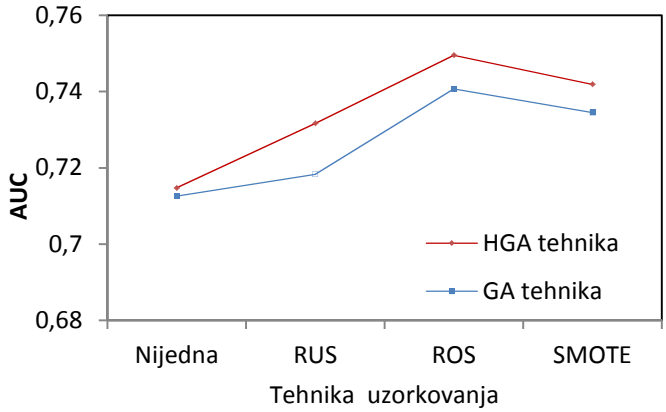
Iz rezultata prikazanih u tablici 5.4 kao i na slikama 5.6, 5.7 i 5.8 se vidi, kao i kod hrvatskog skupa podataka, da se dobiva model s većom točnošću kada se ne koristi nijedna tehnika uzorkovanja podataka, ali ostale mjere performansi iskazuju lošije vrijednosti. Promatrajući pojedine tehnike uzorkovanja, vidimo da ROS uzorkovanje i ovdje daje, osim po točnosti, najbolje rezultate po svim ostalim mjerama performansi. Isto tako se vidi da su rezultati HGA-NN tehnike ukupno bolji u odnosu na GA-NN tehniku.

Uspoređujući rezultate prikazane na slikama 5.6, 5.7 i 5.8 s ukupnom usporedbom relativnog troška HGA-NN i GA-NN tehnike, sa i bez korištenja tehnika uzorkovanja, na slici 5.9, dolazimo do zaključka da su F-β vrijednosti slične troškovnim funkcijama te da izrazito naglašavaju razliku rezultata u korist tehnika uzorkovanja, što je također konzistentno s rezultatima kod hrvatskog skupa kreditnih podataka.

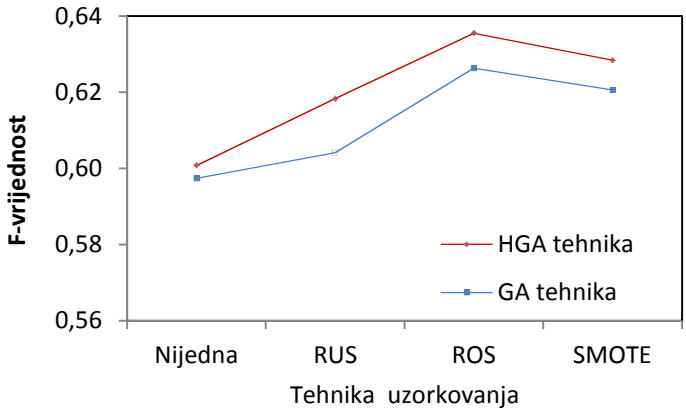
Tablica 5.4 Rezultati HGA-NN i GA-NN tehnike nakon primjene različitih tehnika uzorkovanja na njemačkom skupu podataka

Tehnika uzorkovanja	Točnost	AUC	F-vrij.	F-β	TP	FN	FP	TN
Rezultati GA-NN tehnike								
Nijedna	78,3	0,7126	0,5974	0,5594	161	139	78	622
RUS	67,5	0,7183	0,6041	0,7205	248	52	273	427
ROS	68,5	0,7407	0,6263	0,7573	264	36	279	421
SMOTE	68,7	0,7345	0,6206	0,7420	256	44	269	431
Rezultati HGA-NN tehnike								
Nijedna	**78,6***	0,7148	0,6008	0,5606	161	139	75	625
RUS	69,5	0,7317	0,6183	0,7269	247	53	252	448
ROS	69,6	**0,7495***	**0,6355***	**0,7641***	265	35	269	431
SMOTE	69,6	0,7419	0,6284	0,7480	257	43	261	439

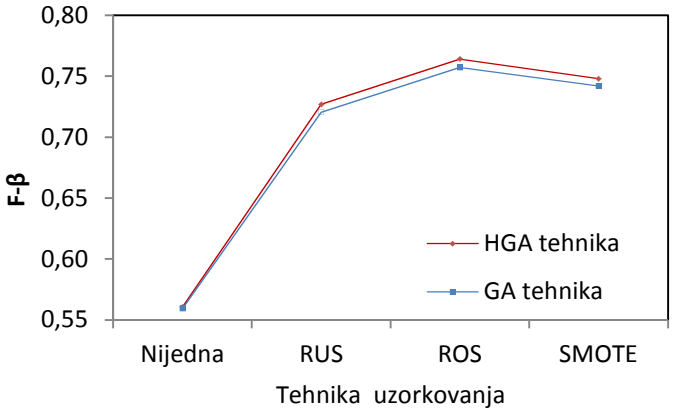
* označava najbolji rezultat za svaku od mjera



Slika 5.6 Komparacija AUC rezultata HGA-NN i GA-NN tehnike za različite tehnike uzorkovanja na njemačkom skupu podataka



Slika 5.7 Komparacija F-vrijednosti HGA-NN i GA-NN tehnike za različite tehnike uzorkovanja na njemačkom skupu podataka

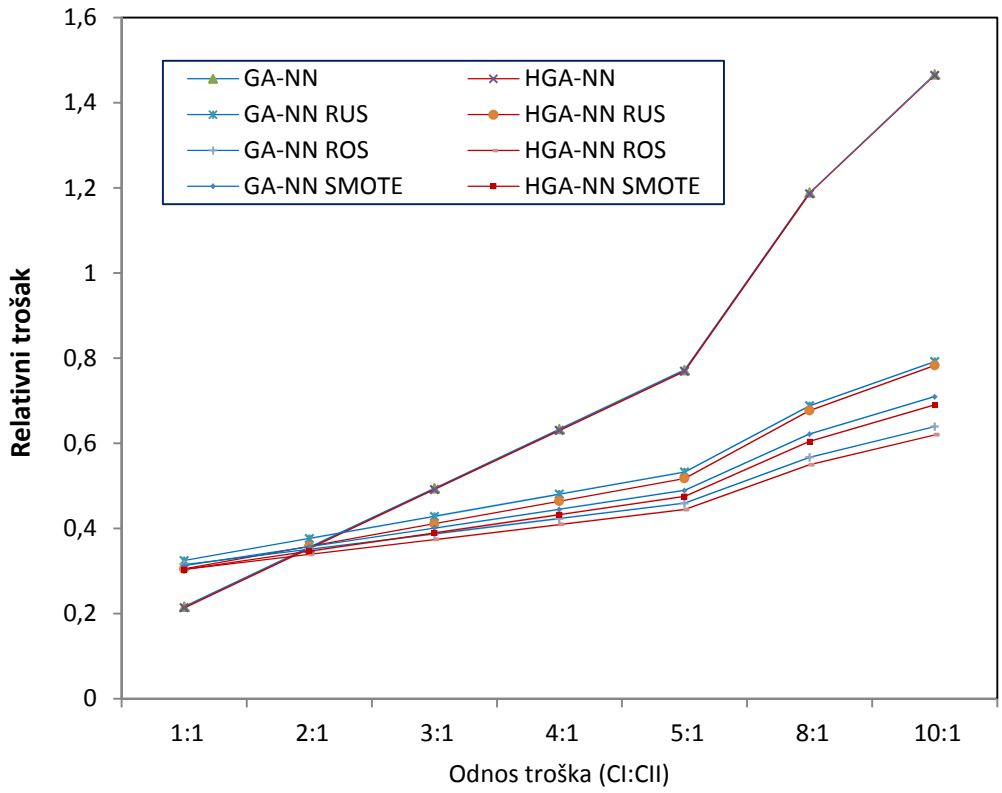


Slika 5.8 Komparacija F-$_\beta$ vrijednosti, tehnika HGA-NN i GA-NN, za različite tehnike uzorkovanja na njemačkom skupu podataka

Tablica 5.5 Komparacija relativnih troškova HGA-NN i GA-NN tehnike nakon primjene različitih tehnika uzorkovanja na njemačkom skupu podataka

Tehnika uzorkovanja	Troškovni odnos (C_I:C_{II})						
	1:1	2:1	3:1	4:1	5:1	8:1	10:1
Rezultati GA-NN tehnike							
Nijedna	0,2170	0,3560	0,4950	0,6339	0,7729	1,1899	1,4679
RUS	0,3250	0,3770	0,4290	0,4810	0,5330	0,6889	0,7929
ROS	0,3150	0,3510	0,3870	0,4230	0,4590	0,5670	0,6390
SMOTE	0,3130	0,3570	0,4010	0,4451	0,4891	0,6211	0,7091
Rezultati HGA-NN tehnike							
Nijedna	**0,2140***	0,3530	0,4919	0,6309	0,7699	1,1869	1,4649
RUS	0,3050	0,3580	0,4110	0,4640	0,5171	0,6761	0,7821
ROS	0,3040	**0,3390***	**0,3740***	**0,4091***	**0,4441***	**0,5491***	**0,6191***
SMOTE	0,3040	0,3470	0,3900	0,4330	0,4760	0,6050	0,6909

* označava najbolji rezultat u svakom troškovnom odnosu



Slika 5.9 Komparacija relativnog troška HGA-NN i GA-NN tehnike sa i bez korištenja tehnika uzorkovanja na njemačkom skupu podataka

Na slici 5.9 je prikazana, na jednom dijagramu, komparacija relativnog troška HGA-NN i GA-NN tehnike za sve konstruirane modele. Iz dijagrama se jasno vidi da obje tehnike postižu najbolju točnost predikcije tada kada ne koriste dodatne tehnike uzorkovanja. Navedeno proizlazi iz činjenice što za odnos troška 1:1 obje tehnike, kada ne koriste dodatne tehnike uzorkovanja, postižu najniži relativni trošak. Iz dijagrama se također vidi da najtočniji modeli postaju granično dobri, s troškovnog aspekta, kod odnosa troška 2:1, a da za sve odnose troškova iznad 2:1 modeli koji su konstruirani uz korištenje dodatnih tehnika uzorkovanja daju bolje rezultate.

5.5.3 **Diskusija rezultata**

Promatrajući rezultate AUC mjere na hrvatskom i njemačkom skupu podataka vidljivo je da primjena bilo koje od tehnika; ROS, RUS ili SMOTE daje bolje rezultate u odnosu na slučaj kada se ne koristi niti jedna tehnika ujednačavanja klasne ravnoteže. Rezultati su konzistentni na obadva skupa podataka.

Rezultati F-mjere su distribuirani gotovo na isti način kao i rezultati AUC mjere na njemačkom skupu podataka. Na hrvatskom skupu podataka značajnije bolje rezultate daje ROS tehnika u odnosu na RUS i SMOTE, kao i u odnosu na slučaj kada se ne koristi niti jedna tehnika uzorkovanja.

Komparacija F-β vrijednosti tehnika HGA-NN i GA-NN, za različite tehnike uzorkovanja na hrvatskom i njemačkom skupu podataka, pokazuje značajno poboljšanje rezultata primjenom tehnika ROS, RUS ili SMOTE u odnosu na originalne skupove. Tako značajna razlika proizlazi iz činjenice da smo u izrazu po kojem se računa F-β vrijednost uvrstili vrijednost 2 za parametar β te tako naglasili važnost odziva u odnosu na preciznost. Odnosno, značajno smo sankcionirali grešku tipa I u odnosu na grešku tipa II. U takvim okolnostima primjena dodatnih tehnika, daje više nego evidentan doprinos kvaliteti rezultata klasifikacije. Može se primijetiti da su međusobne razlike rezultata unutar tehnika ROS, RUS i SMOTE značajno manje u odnosu na rezultate dobivene konstrukcijom modela na originalnom skupu podataka.

Dobiveni rezultati pokazuju da F-β vrijednost u suštini daje rezultate vrlo slične funkciji relativnog troška pogrešne klasifikacije. Stoga su dobiveni rezultati vrlo usporedivi s rezultatima funkcije relativnog troška, u ovom slučaju za troškovni odnos 5:1. U ovom eksperimentu je korišten odnos 5:1 kao očekivani odnos greške tipa I i tipa II. Odnos 5:1 je definirao prof. Hofmann, koji je dao njemački skup kreditnih podataka dostupan u UCI repozitoriju podataka za strojno učenje, kao očekivani odnos u normalnim okolnostima. Temeljem toga smo, kao i temeljem praktičnih spoznaja, parametar β postavili na 2, a funkciju relativnog troška prikazali za relativne odnose troškova od 1:1 do 10:1.

Iz tablica 5.3 i 5.5 kao i iz grafikona (slika 5.5 i 5.9) koji prikazuju usporedbu relativnog troška krive klasifikacije HGA-NN i GA-NN tehnike sa i bez korištenja tehnika uzorkovanja se jasno vidi da navedene tehnike uzorkovanja daju pozitivan doprinos rezultatima temeljnih tehnika u smanjenju troška pogrešne klasifikacije. Kod toga je taj doprinos značajniji što je relativni odnos troška veći.

Rezultati eksperimenata prikazanih u ovom poglavlju podupiru hipotezu da ROS, RUS i SMOTE uzorkovanja poboljšavaju rezultate GA-NN i HGA-NN tehnike kod troškovno osjetljive klasifikacije nad neuravnoteženim podacima kredita građana mjereno: AUC mjerom, F-mjerom, F-β mjerom i funkcijom relativnog troška. Pri tome, ukupno uzevši, ROS uzorkovanje u kombinaciji sa HGA-NN tehnikom daje najbolje rezultate. Stoga će se u usporedbi rezultata s rezultatima prezentiranim u literaturi za njemački skup kreditnih podataka koristiti, u skladu s tablicom 5.4, rezultati HGA-NN i HGA-NN ROS tehnike. HGA-NN tehnika daje najtočniju predikciju, a rezultati HGA-NN ROS tehnike su najbolji po svim ostalim mjerama performansi.

Mnogi autori, među kojima su Tsai i Cheng (2012), navode da je njemački kreditni skup izazovna referentna mjera (engl. *benchmark*) za predviđanje bankrota. Stoga, kako bi se pouzdano i učinkovito ispitala uspješnost predviđanja nekog modela, treba barem ispitati i usporediti rezultate za njemački skup podataka. U skladu s tim, kreditni skup podataka koji se koristi u ovom eksperimentu kao referentna mjera je njemački kreditni skup podataka.

Tablica 5.6 Komparacija točnosti klasifikacije HGA-NN i HGA-NN ROS tehnike s rezultatima iz literature na njemačkom skupu podataka

Tehnika	**Vjerojatnost greške**		**Točnost**
	P_1 (%)	P_2 (%)	Sred.vrij. (%)
HGA-NN	46,61	10,17	**78,90**
HGA-NN ROS	11,67	38,43	69,60
SVM[Han, Han i Zhao, 2013]	37,00	18,00	77,00
LogR[Zhu et al., 2013]	50,66	11,69	76,62
Bagging/MLP[Nanni i Lumini, 2009]	49,40	24,60	75,33
Logit[Lieli i White, 2010]	18,33	44,00	63,70

Tablica 5.7 Komparacija relativnih troškova pogrešne klasifikacije HGA-NN i HGA-NN ROS tehnike s rezultatima iz literature na njemačkom skupu podataka

	Troškovni odnos (C_I:C_{II})						
Tehnika	1:1	2:1	3:1	4:1	5:1	8:1	10:1
HGA-NN	**0,2110***	0,3509	0,4907	0,6305	0,7703	1,1898	1,4695
HGA-NN ROS	0,3040	**0,3390***	**0,3740***	**0,4091***	**0,4441***	**0,5491***	**0,6191***
SVM(Han, Han i Zhao, 2013)	0,2370	0,3480	0,4590	0,5700	0,6810	1,0140	1,2360
LogR(Zhu et al., 2013)	0,2338	0,3858	0,5378	0,6898	0,8417	1,2977	1,6016
Bagging/MLP (Nanni i Lumini, 2009)	0,3204	0,4686	0,6168	0,7650	0,9132	1,3578	1,6542
Logit(Lieli i White, 2010)	0,3630	0,4180	0,4730	0,5280	0,5830	0,7479	0,8579

* označava najbolji rezultat u svakom troškovnom odnosu

Iz usporednih rezultata prezentiranih u literaturi te rezultata HGA-NN tehnike i njezinog proširenja tehnikom za ublažavanje negativnog utjecaja klasno neuravnoteženih podataka ROS, može se zaključiti da HGA-NN tehnika uz navedena proširenja potiže takve rezultate koji opravdavaju dodatne napore u njezinoj implementaciji. Komparacija izvedena na njemačkom skupu podataka po točnosti klasifikacije (Tablica 5.6), kao i po relativnim troškovima pogrešne klasifikacije (Tablica 5.7) pokazuje da u ovoj disertaciji prezentirane tehnike ostvaraju odlične rezultate. Rezultat je to temeljnog algoritma po kojem tehnika istovremeno radi prilagodbu: (1) podataka algoritmu klasifikacije (selekcija atributa) kao i (2) parametara algoritma podacima. Kod prve verzije algoritma su uočene određene slabosti, a očitovale su se u vremenu potrebnom za izvršavanje. Uočeni nedostatci su minimizirani uvođenjem predselekcije atributa brzim fitarskim tehnikama te uvođenjem inkrementalne faze. Nakon toga je tehnika dodatno proširena određenim tehnikama za ublažavanje negativnog utjecaja klasno neuravnoteženih podataka s ciljem postizanja ne samo najbolje točnonosti, nego i troškovne efikanosti što je od posebnog značaja u domeni za koju je tehnika razvijana.

5.6 Zaključci poglavlja

Iz rezultata eksperimenata u kojima je istraživan utjecaj tehnika za ublažavanje negativnog utjecaja klasno neujednačenih skupova podataka u kombinaciji s tehnikama selekcije atributa na rezultate klasifikacije možemo zaključiti:

- Da je točnost klasifikacije najbolja kada se ne primjenjuju dodatne tehnike i to kod oba klasifikatora na oba skupa podataka.
- Da su rezultati klasifikacije mjereno AUC mjerom, F-mjerom i F-β mjerom bolji kada se primjenjuju dodatne tehnike.
- Da su rezultati mjereni troškovima netočne klasifikacije bolji bez primjene dodatnih tehnika samo kada je odnos troška tipa I i tipa II jednak. Kod svih ostalih odnosa troškova bolji rezultati se dobivaju korištenjem dodatnih tehnika.
- Da se od tehnika za ublažavanje negativnog utjecaja klasno neuravnoteženih podataka u najvećem broju slučajeva (kriterija) ROS tehnika pokazala kao najbolja, konzistentno na hrvatskom i njemačkom skupu podataka.

Uspoređujući rezultate klasifikacije, mjereno relativnim troškovima netočne klasifikacije, HGA-NN ROS tehnike s ostalim rezultatima prezentiranim u literaturi na njemačkom kreditnom skupu podataka utvrdili smo da ovdje prezentirana HGA-NN ROS tehnika ostvaraju obećavajuće rezultate.

POGLAVLJE 6

ZAKLJUČAK

Svrha ovog doktorskog rada bila je temeljito istražiti ukupni skup podataka kojima raspolaže banka te utvrditi do koje mjere ti podaci mogu biti dobra osnovica za predviđanje sposobnosti tražitelja kredita da vrati kredit na vrijeme. Takvo predviđanje sposobnosti tražitelja kredita trebalo je izvršiti bez traženja dodatnih podataka od klijenta uz pretpostavku da je tražitelj kredita već dulje vrijeme klijent banke te da je banka već u svojoj bazi podataka prikupila dovoljno podataka o klijentu.

U skladu s tako definiranom svrhom istraživanja bio je postavljen glavni cilj istraživanja; razviti vrlo efikasne, u skladu s najnovijim znanstvenim i tehničkim spoznajama, tehnike za odabir optimalnog podskupa atributa i predviđanje sposobnosti tražitelja kredita da će vratiti kredit na vrijeme i u skladu s ugovorenim obvezama.

Realizacija glavnog cilja operacionalizirana je kroz sljedeće detaljnije definirane ciljeve istraživanja:

- Kreirati hibridne tehnike selekcije atributa posebno prilagođene problemskoj domeni (GA-NN i HGA-NN) - temeljene na genetskim algoritmima i umjetnim neuronskim mrežama.
- Kreirati novi hibridni genetski algoritam uključivanjem rezultata filtarskih tehnika i a priori spoznaja u početnu populaciju genetskog algoritma.

- Kreirati novi operator selekcije kod genetskog algoritma, jedinstvena selekcija.
- Kreirati sofisticirane kreditne modele koji omogućuju povećanje učinkovitosti alokacije kapitala što predstavlja doprinos implementaciji Basella III.

Pojedinačni ciljevi su realizirani provođenjem niza istraživanja koja su prikazana u radu kao zaokružene cjeline:

- Prva cjelina, kao svojevrstan uvod u navedena istraživanja, definira važne pojmove s područja kombinatorne optimizacije te daje kratki pregled tehnika i glavnih koncepata bitnih prilikom razvoja hibridnih tehnika kombinatorne optimizacije. Također prikazuje veze i odnose među konceptima i tehnikama kombinatorne optimizacije te ističe važnost hibridizacije kao koncepta suradnje meta-heurističkih tehnika i drugih tehnika za optimizaciju.
- Temeljem niza istraživanja je kreirana, pod nazivom GA-NN, hibridna tehnika za odabir optimalnog podskupa atributa, koja istovremeno radi odabir atributa te optimira parametre umjetnih neuronskih mreža što poboljšava njihovu klasifikacijsku točnost. U okviru te tehnike razvijen je odgovarajući genetski operator selekcije, jedinstvena selekcija. Klasifikacijska točnost predložene tehnike je uspoređena s rezultatima šire primjenjivanih tehnika selekcije atributa: omjer dobiti, Ginijev indeks, korelacija i tehnika glasanja.
- Uočene slabosti genetskog algoritma, kao osnovne komponente GA-NN tehnike nastojalo se prevladati kreiranjem novog, hibridnog genetskog algoritma i to uključivanjem rezultata filtarskih tehnika i a priori spoznaja u početnu populaciju genetskog algoritma kao i uvođenjem inkrementalne faze u algoritam. Na taj način je kreirana HGA-NN tehnika.
- Na kraju su GA-NN i HGA-NN tehnika proširene određenim tehnikama za ublažavanje negativnog utjecaja klasno neuravnoteženih podataka na rezultate klasifikacije, kako bi kreditni modeli koje generiraju navedene tehnike omogućili dodatno povećanje učinkovitosti alokacije kapitala banaka.

Iz svega navedenog se vidi da se u disertaciji prikazuju teorijski koncepti kao i rezultati empirijskih istraživanja s područja hibridnih tehnika kombinatorne optimizacije temeljenih na umjetnoj inteligenciji s primjenom na procjenu rizika pri odobravanju kredita građanima.

Realizacija znanstvenih ciljeva istraživanja i očekivanog znanstvenog doprinosa rada analizira se prihvaćanjem ili odbacivanjem hipoteza istraživanja.

Hipoteze istraživanja

H1: GA-NN tehnika je statistički značajno točnija na razini statističke značajnosti $p<0,05$ u odnosu na šire primjenjivane tehnike selekcije atributa: omjer dobiti, Ginijev indeks, korelacija i tehnika glasanja.

Temeljem provođenja niza istraživanja te prezentiranih rezultata tih istraživanja u poglavlju 3 smo zaključili da je GA-NN tehnika značajno bolja pri izboru atributa za klasifikaciju u usporedbi s drugim tehnikama koje se često koriste za odabir atributa. Dobiveni rezultati podupiru hipotezu H3.1 koja je istovjetna hipotezi istraživanja H1. Stoga možemo zaključiti da se

prihvaća prva hipoteza istraživanja.

H2: Uključivanje preliminarne selekcije atributa i inkrementalne faze u algoritam temeljen na GA, kombinirano s efektima nove strategije generiranja inicijalne populacije GA, rezultira statistički značajnim poboljšanjem prosječne klasifikacijske točnosti novog algoritma uz razinu statističke značajnosti $p<0,01$

Iz rezultata prezentiranih istraživanja u poglavlju 4 i njihove usporedbe s rezultatima prezentiranim u literaturi na njemačkom skupu podataka, jasno je da rezultati primjene ove tehnike nadmašuju rezultate objavljene u literaturi. Rezultati dodatno potvrđuju hipotezu H4.1 koja je istovjetna hipotezi H2, a koja kaže da redukcija područja pretraživanja i inkrementalna faza, kombinirano s efektima strategije generiranja inicijalne populacije GA, rezultira poboljšanjem performansi klasifikatora. To poboljšanje je dovoljno veliko da bi bilo znanstveno i praktično interesantno. Stoga možemo zaključiti da se

prihvaća druga hipoteza istraživanja.

Prihvaćanje hipoteza istraživanja H1 i H2 predstavlja potvrdu uspješne realizacije postavljenih znanstvenih ciljeva istraživanja.

Osim znanstvenog doprinosa istraživanja, rezultati rada daju i primjenjiv društveni doprinos. Društveni doprinos sagledavamo na sljedeći način:

- Uspješno korištenje rezultata istraživanja u procesu odobravanja kredita, za ocjenu rizičnosti i kreditne sposobnosti tražitelja kredita, može značajno doprinijeti korištenju tehnika dubinske analize podataka u bankarstvu uz evidentne uštede na strani banke i pojedinaca, odnosno društva, čime se dinamizira tržište i otvara nova stranica u pogledu načina odobravanja kredita. Tržište i društvo doživljavaju time značajne promjene.
- Disertacija prikazuje i proširuje područje primjene hibridnih tehnika umjetne inteligencije. U bogatu kolekciju klasifikacijskih problema donosi: (1) nove analize, (2) originalni set podataka i (3) sveobuhvatan skup eksperimentalnih rezultata.
- Doprinos je rada i u tome što prikazuje kako primijeniti tehnike umjetne inteligencije u rješavanju važnih praktičnih problema jer prikazuje cjeloviti postupak po kojem je istraživanje provedeno. Pored toga, postoji doprinos jer je kreiran opći model koji je u određenoj mjeri primjenjiv za unapređenje procesa odobravanja kredita u svim bankama. Svaka banka može, primjenom modela na vlastitim podatcima unaprijediti proces odobravanja kredita.

Ova doktorska disertacija jeste teorijska studija poduprta empirijska analizama te ima doprinos u oba područja, znanstvenom i društvenom. Ona daje ocjenu postojećih metoda i tehnika, osvjetljavajući snage i slabosti tih metoda i tehnika te stvaralačkom primjenom navedenih spoznaja dolazi do novih hibridnih tehnika koje evaluira na praktičnim primjerima. Modeli i koncepti izloženi u radu imaju potencijala da radikalno utječu na to kako će građani i tvrtke koristiti informacijsku tehnologiju u svom svakodnevnom životu.

Rezultati prezentiranih algoritama jasno ukazuju na njihov potencijal pri rješavanju problema odabira atributa i ocjenjivanju kreditnog rizika građana te time opravdavaju veći napor u dizajnu i implementaciji. Potencijal prezentiranih algoritama u ocjenjivanju kreditnog rizika građana može se iskoristiti za poboljšanje načina na koji banke upravljaju kreditnim rizikom građana, što predstavlja promociju stabilnog i zdravog bankarstva. I konačno, potreba za boljim upravljanjem kreditnim rizicima i sofisticiranim kreditnim modelima je potaknula istraživanja prezentirana u ovom radu.

Istraživanja i dobiveni rezultati prezentirani u poglavljima 2, 3 i 4 su već objavljeni u uglednim međunarodnim časopisima. Rezultati istraživanja prezentirani u poglavlju 5 prvi puta su objavljeni u ovom doktorskom radu.

LITERATURA

[1] Aha D.W. & Bankert R.L. (1996). A Comparative Evaluation of Sequential Feature Selection Algorithms. Learning from Data (pp. 199-206). Springer New York.

[2] Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. European Journal of Operational Research, 222(1), 168-178.

[3] Akpınar, S., Mirac Bayhan, G., & Baykasoglu, A. (2013). Hybridizing ant colony optimization via genetic algorithm for mixed-model assembly line balancing problem with sequence dependent setup times between tasks. Applied Soft Computing, 13(1), 574-589.

[4] Al-Shihabi S., Olafsson S. (2010). A hybrid of nested partition, binary ant system, and linear programming for the multidimensional knapsack problem. Computers and Operations Research, 37, 247–255.

[5] Arora, S., & Barak, B. (2009). Computational complexity: a modern approach. Cambridge University Press.

[6] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[7] Bäck T., Fogel D.B. & Michalewicz Z. (1997). Handbook of Evolutionary Computation, IOP Publishing.

[8] Bent R., Van Hentenryck P. (2001). A two-stage hybrid local search for the vehicle routing problem with time windows. Technical Report, CS-01-06, Department of Computer Science, Brown University, USA.

[9] BIS. Basel III: A global regulatory framework for more resilient banks and banking systems. (2011). Basel Committee on Banking Supervision, Bank for International Settlements, Basel. ISBN print: 92-9131-859-0 (http://www.bis.org/publ/bcbs189.pdf)

[10] BIS. International convergence of capital measurement and capital standards: A revised framework. (2006).Basel Committee of Banking Supervision, Bank for International Settlements, Basel.

[11] Blum C. et al. (2011). Hybrid metaheuristics in combinatorial optimization: A survey. Applied Soft Computing, 11 : 4135–4151.

[12] Blum C., Roli A. (2003). Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. ACM Computing Surveys, 35(3): 268–308.

[13] Brennan, P. (2012). A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection, Disertacija, Blanchardstown Dublin, Irska.

[14] Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 853-867). Springer US.

[15] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16 , 321–357.

[16] Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. Data Mining and Knowledge Discovery, 17(2), 225-252.

[17] Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter, 6(1), 1-6.

[18] Chi B.-W., Hsu C.-C. (2012). A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. Expert Systems with Applications, 39, 2650–2661.

[19] Christofides, N. (1979). Combinatorial optimization. In A Wiley-Interscience Publication, Based on a series of lectures, given at the Summer School in Combinatorial Optimization, held in Sogesta, Italy, May 30th-June 11th, 1977, Chichester: Wiley, 1979, edited by Christofides, Nicos (Vol. 1).

[20] Cieslak, D. A., & Chawla, N. V. (2008). Learning decision trees for unbalanced data. In Machine Learning and Knowledge Discovery in Databases (pp. 241-256). Springer Berlin Heidelberg.

[21] Cotta C. et al. (1995). Hybridizing genetic algorithms with branch and bound techniques for the resolution of the tsp, in: D.W. Pearson, N.C. Steele, R.F. Albrecht (Eds.), Artificial Neural Nets and Genetic Algorithms 2, Springer-Verlag, pp. 277–280.

[22] Crook, J. N., Edelman, D. B.,& Lyn C. (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183, 1447–1465.

[23] Dal Pozzolo, A., Caelen, O., Waterschoot, S., & Bontempi, G. (2013). Racing for unbalanced methods selection. In Intelligent Data Engineering and Automated Learning–IDEAL 2013 (pp. 24-31). Springer Berlin Heidelberg.

[24] Danenas P. et al. (2011). Credit Risk Evaluation Model Development Using Support Vector Based Classifiers. Procedia Computer Science 4, 1699–1707.

[25] De Andrade, F.W.M., & Thomas, L.C. (2007) . Structural models in consumer credit. European Journal of Operational Research, 183, 1569–1581.

[26] Domingos, P. (1999, August). Metacost: A general method for making classifiers cost-sensitive. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 155-164). ACM.

[27] Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. Computational Intelligence, 20(1), 18-36.

[28] Finlay, S. (2010). Credit scoring for profitability objectives. European Journal of Operational Research, 202, 528–537.

[29] Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, 210 (2), 368-378.

[30] Gamberger, D., Lavrac, N., & Dzeroski, S. (2000). Noise detection and elimination in data preprocessing: experiments in medical domains. Applied Artificial Intelligence, 14(2), 205-223.

[31] García, V., Marqués, A. I., & Sánchez, J. S. (2012, January). Improving risk predictions by preprocessing imbalanced credit data. In Neural Information Processing (pp. 68-75). Springer Berlin Heidelberg.

[32] Glover, F. and Laguna, M. (1997). Tabu Search. Kluwer Academic Publishers.

[33] Goldberg D. E., & Holland J. H. (1988). Genetic algorithms and machine learning. Machine Learning 3: 95-99.

[34] Goldreich, O. (2010). P, NP, and NP-Completeness: The basics of computational complexity. Cambridge: Cambridge University Press.

[35] Golub M. (2001). Poboljšavanje djelotvornosti paralelnih genetskih algoritama. PhD thesis. Zagreb: University of Zagreb.

[36] Han J, & Kamber M. (2006). Data Mining: Concepts and Techniques (Second Edition). 500 Sansome Street, Suite 400, San Francisco, CA 94111:Morgan Kaufmann Publishers.

[37] Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Advances in intelligent computing (pp. 878-887). Springer Berlin Heidelberg.

[38] Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann.

[39] Han, L., Han, L., & Zhao, H. (2013). Orthogonal support vector machine for credit scoring. Engineering Applications of Artificial Intelligence, 26(2), 848-862.

[40] http://web.studenti.math.pmf.unizg.hr/~manger/tr/TR-VI.pdf, dostupano 20.08.2012.

[41] http://www.bis.org/bcbs/basel3.htm, dostupano 21.11.2011.

[42] Huang C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines, Expert Systems with Applications 33, 847–856.

[43] Ilakovac, V. (2009). Statistical hypothesis testing and some pitfalls. Biochemia Medica, 19(1), 10-16.

[44] Japkowicz, N. (2000). Learning from imbalanced data sets: a comparison of various strategies. In AAAI workshop on learning from imbalanced data sets (Vol. 68).

[45] Japkowicz, N. (2000b). The class imbalance problem: Significance and strategies. In Proc. of the Int'l Conf. on Artificial Intelligence.

[46] Japkowicz, N., Shah, M., Evaluating learning algorithms: A classification perspective, Cambridge University Press, New York, 2011.

[47] Jin C. et al. (2012). Attribute selection method based on a hybrid BPNN and PSO algorithms. Applied Soft Computing 12, 2147–2155.

[48] Jourdan L., Basseur M., Talbi E.-G. (2009). Hybridizing exact methods and metaheuristics: A taxonomy. European Journal of Operational Research, 199 : 620–629.

[49] Kao Y.-T., Zahara E. (2008). A hybrid genetic algorithm and particle swarm optimization for multimodal functions. Applied Soft Computing, 8, 849–857.

[50] Khandani A.E., Kim A.J., & Lo A.W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance 34, 2767–2787.

[51] Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Systems with Applications, 37, 6233–6239.

[52] Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Proceedings of the ninth international workshop on Machine learning (pp. 249-256). Morgan Kaufmann Publishers Inc.

[53] Kohavi, R, & John, G.H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97, 273-324.

[54] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI (Vol. 14, No. 2, pp. 1137-1145).

[55] Korte, B. B. H., & Vygen, J. (2012). Combinatorial optimization (Vol. 21). Springer.

[56] Lavrac, N., Gamberger, D., & Turney, P. (1996). Cost-sensitive feature reduction applied to a hybrid genetic algorithm.

[57] Lee S., Ahn H. (2011). The hybrid model of neural networks and genetic algorithms for the design of controls for internet-based systems for business-to-consumer electronic commerce. Expert Systems with Applications. 38, 4326–4338.

[58] Li S.-T., Shiue, W., & Huang, M.-H. (2006). The evaluation of consumer loans using support vector machines. Expert Systems with Applications 30, 772–782.

[59] Li X. et al. (2011). Initialization strategies to enhancing the performance of genetic algorithms for the p-median problem. Computers & Industrial Engineering 61, 1024–1034.

[60] Lieli, R. P., & White, H. (2010). The construction of empirical credit scoring rules based on maximization principles. Journal of Econometrics, 157(1), 110-119.

[61] Ling, C. X., Yang, Q., Wang, J., & Zhang, S. (2004). Decision trees with minimal costs. In Proceedings of the twenty-first international conference on Machine learning (p. 69). ACM.

[62] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113-141.

[63] López, V., Fernández, A., Moreno-Torres, J. G., & Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. Expert Systems with Applications, 39(7), 6585-6608.

[64] Maaranen H. et al. (2004). Quasi-Random Initial Population for Genetic Algorithms. Computers and Mathematics with Applications 47, 1885-1895.

[65] Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. Omega, 31(2), 83–96.

[66] Marić M. (2008). Rešavanje nekih NP – teških hijerarhijsko-lokacijskih problema primenom genetskih algoritama. Doktorska disertacija. Matematički fakultet, Univerzitet u Beogradu.

[67] Mashinchi M.H. et al. (2011). Hybrid optimization with improved tabu search. Applied Soft Computing, 11, 1993–2006.

[68] Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks, 21(2), 427-436.

[69] McCall J. (2005). Genetic algorithms for modeling and optimization. Journal of Computational and Applied Mathematics 184, 205–222.

[70] Menardi, G., Tedeschi, F., & Torelli, N. (2011). On the Use of Boosting Procedures to Predict the Risk of Default. In Classification and Multivariate Analysis for Complex Data Structures (pp. 211-218). Springer Berlin Heidelberg.

[71] Michalewicz, Z. (1998). Genetic algorithms+ data structures= evolution programs. Springer.

[72] Mitchell, M. (1995). Genetic algorithms: an overview. Complexity, 1(1), 31-39.

[73] Mitchell, M. (1996) An introduction to genetic algorithms. MIT Press/Addison-Wesley, Cambridge, MA.

[74] Muller L.F. et al. (2012). A hybrid adaptive large neighborhood search heuristic for lot-sizing with setup times. European Journal of Operational Research, 218, 614–623.

[75] Myers, J. L., & Well, A. (2003). Research design and statistical analysis. Lawrence Erlbaum Associates, Inc., Publishers. Mahwah, New Jersey, USA.

[76] Naganjaneyulu, S., & Kuppa, M. R. (2013). A novel framework for class imbalance learning using intelligent under-sampling. Progress in Artificial Intelligence, 2(1), 73-84.

[77] Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications, 36(2), 3028-3033.

[78] Niven E.B., Deutsch C.V. (2012). Calculating a robust correlation coefficient and quantifying its uncertainty. Computers & Geosciences 40, 1–9. – nema ga nigdje

[79] Nwana V., Darby-Dowman K., Mitra G. (2005). A co-operative parallel heuristic for mixed zero-one linear programming: Combining simulated annealing with branch and bound. European Journal of Operational Research, 164, 12–23.

[80] Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Systems with Applications, 41(4), 2052-2064.

[81] Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert Systems with Applications 39, 12605–12617.

[82] Osman I.H. (2002). Preface: Focused issue on applied meta-heuristics. Computers & Industrial Engineering, 44 : 205–207.

[83] Papadimitriou, C. H. (2003). Computational complexity. John Wiley and Sons Ltd.

[84] Pezzella, F., Morganti, G., & Ciaschetti, G. (2007). A genetic algorithm for the flexible job-shop scheduling problem. Computers and Operations Research, 35(10), 3202–3212.

[85] Piramuthu S. (2006). On preprocessing data for financial credit risk evaluation. Expert Systems with Applications 30, 489–497.

[86] Pirkwieser S. (2012). Hybrid Metaheuristics and Matheuristics for Problems in Bioinformatics and Transportation. PHD thesis, Faculty of Informatics, University of Technology, Vienna.

[87] Raeder, T., Forman, G., & Chawla, N. V. (2012). Learning from imbalanced data: evaluation matters. In Data Mining: Foundations and Intelligent Paradigms (pp. 315-331). Springer Berlin Heidelberg.

[88] Rama Mohan Rao A., Shyju P.P. (2008). Development of a hybrid meta-heuristic algorithm for combinatorial optimization and its application for optimal design of laminated composite cylindrical skirt. Computers and Structures, 86, 796–815.

[89] Renner G., Eka´rt A. (2003). Genetic algorithms in computer aided design. Computer-Aided Design 35, 709–726.

[90] Sarlija, N., Bensic, M., & Zekic-Susac, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. Expert Systems with Applications 36, 8778–8788.

[91] Schowe B. (2011).Feature Selection for high-dimensional data with RapidMiner. Dortmund :Technical University of Dortmund.

[92] Sergienko, I. V., Hulianytskyi, L. F., & Sirenko, S. I. (2009). Classification of applied methods of combinatorial optimization. Cybernetics and Systems Analysis, 45(5), 732-741.

[93] Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 40(12), 3358-3378.

[94] Šušteršic, M., Mramor, D., & Zupan J. (2009). Consumer credit scoring models with limited data. Expert Systems with Applications 36, 4736–4744.

[95] Tsai, C. F., & Cheng, K. C. (2012). Simple instance selection for bankruptcy prediction. Knowledge-Based Systems, 27, 333-342.

[96] Tsai, M.-C., Lin, S.-P., Cheng, C.-C., & Lin, Y.-P. (2009). The consumer loan default predicting model – An application of DEA–DA and neural network. Expert Systems with Applications 36, 11682–11690.

[97] Twala B. (2010). Multiple classifier application to credit risk assessment. Expert Systems with Applications, 37, 3326–3336.

[98] Van Hulse, J., & Khoshgoftaar, T. (2009). Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering, 68(12), 1513-1542.

[99] Vořechovský M. (2012). Correlation control in small sample Monte Carlo type simulations II: Analysis of estimation formulas, random correlation and perfect uncorrelatedness. Probabilistic Engineering Mechanics 29, 105–120.

[100] Wang, S. J., et al. (2009). Empirical analysis of support vector machine ensemble classifiers. Expert Systems with Applications, 36(3), 6466-6476.

[101] Williamson D. P., Shmoys D. B. The Design of Approximation Algorithms. Cambridge University Press, 2011.

[102] Wolpert D.and Macready. W. (1996). No free lunch theorems for search. The Santa Fe Institute, Santa Fe.

[103] Yang W., Li D. & Zhu L. (2011). An improved genetic algorithm for optimal feature subset selection from multi-character feature set. Expert Systems with Applications 38, 2733–2740.

[104] Yuan Q., Qian F., Du W. (2010). A hybrid genetic algorithm with the Baldwin effect. Information Sciences, 180, 640–652.

[105] Yuen C.W.M. et al. (2009). A hybrid model using genetic algorithm and neural network for classifying garment defects. Expert Systems with Applications, 36, 2037–2047.

[106] Zekic-Sušac, M., Benšić, M., & Šarlija, N. (2005). Selecting neural network architecture for investment profitability predictions. Journal of Information and Organizational Sciences, 29, 83-95.

[107] Zekić-Sušac, M., Šarlija, N., & Benšić, M. (2004). Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network and Decision Tree Models. In Proceedings of the 26th International Conference on Information Technology Interfaces, June 7-10., Cavtat/Dubrovnik, Croatia, pp.265-270.

[108] Zhang, G. et al. (2011). An effective genetic algorithm for the flexible job-shop scheduling problem. Expert Systems with Applications 38, 3563–3573.

[109] Zhang, G., Hu, M.Y., Patuwo, B.E., & Indro, D.C. (1999). Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. European Journal of Operations Research, 116, 16–32.

[110] Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter, 6(1), 80-89.

[111] Zhu, X., Li, J., Wu, D., Wang, H., & Liang, C. (2013). Balancing accuracy, complexity and interpretability in consumer credit decision making: A C-TOPSIS classification approach. Knowledge-Based Systems, 52, 258-267.

PRILOZI

Dodatak A

Tablica A.1 Ulazni atributi s deskriptivnom statistikom

Attributi	Tip	Kod	Opis	Statistika	Raspon
att1	integer	ID	*Id - redni broj zapisa*	avg = 500.5 +/- 288.819	[1; 1000]
Grupa1		**G1**	**Osnovne karakteristike**		
att2	integer	AGE	*Starost*	avg = 46.198 +/- 14.097	[20; 80]
att3	integer	GENDER	*Spol*	avg = 0.506 +/- 0.500	[0; 1]
att4	integer	POST	*Pošta (selo,grad)*	avg = 0.581 +/- 0.494	[0; 1]
att5	integer	TLF	*Telefon (0-nema, 1-ima)*	avg = 0.906 +/- 0.292	[0; 1]
att6	integer	TAPA	*Vrijeme na sadašnjoj adresi (izraženo u godinama)*	avg = 2.722 +/- 1.594	[0; 6]
att7	integer	TWB	*Klijent banke (izraženo u godinama)*	avg = 14.039 +/- 8.347	[1; 50]
att8	integer	ACAGE	*Starost tekućeg računa(izraženo u godinama)*	avg = 9.825 +/- 7.566	[1; 34]
att9	integer	MOA	*Mjesec izdavanja kredita (u toku godine)*	avg = 6.400 +/- 3.265	[1; 12]
att10	integer	LMM	*Rok povrata kredita (u mjesecima)*	avg = 39.378 +/- 19.562	[11; 61]
att11	integer	POL	*Namjena kredita (svaka vrsta kredita ima svoj broj)*	avg = 0.874 +/- 0.508	[0; 5]
att12	integer	CRAM	*Iznos kredita (u HRK)*	avg = 25057.85 +/- 18075.76	[2000; 100000]
Grupa2		**G2**	**Povijest uplata/isplata za posljednjih 12 mjeseci**		
att13	integer	TIN	*Ukupne uplate na tekući račun*	avg = 4373.09 +/- 3351.98	[1; 32522]
att14	real	CSHT	*Gotovinske uplate / Ukupne uplate*	avg = 0.071 +/- 0.158	[0; 1]
att15	real	RPT	*Redovne uplate (plaće, mirovine) / Ukupne uplate*	avg = 0.884 +/- 0.206	[0; 1]
att16	real	OTIN	*Dozvoljen minus / Ukupne uplate*	avg = 1.754 +/- 1.272	[0; 5]
att17	integer	TOUT	*Ukupne isplate sa tekućeg računa*	avg = 4665.159 +/- 4967.562	[1; 117664]
att18	real	TITO	*Ukupne uplate / Ukupne isplate*	avg = 0.991 +/- 0.238	[0.010 ; 3]
att19	real	PSTW	*EFTPOS isplate / Ukupne isplate*	avg = 0.164 +/-	[0;

				0.170	0.950]
att20	real	TMTW	*ATM isplate / Ukupne isplate*	avg = 0.398 +/- 0.303	[0; 1]
att21	real	SSTW	*Samoposlužne isplate / Ukupne isplate*	avg = 0.563 +/- 0.322	[0; 1]
att22	real	COTW	*Redovne isplate s računa (rate,trajni nalozi) / Redovne uplate (plaće, mirovine)*	avg = 1.680 +/- 1.235	[0; 5]
att23	real	RIDI	*Odnos isplata u tromjesečju odobravanja kredita / isto tromjesečje godinu ranije*	avg = 0.296 +/- 0.323	[0.030 ; 2]
att24	real	RII	*Odnos uplata u tromjesečju odobravanja kredita / isto tromjesečje godinu ranije*	avg = 1.388 +/- 1.000	[0.010 ; 5]
Grupa3		**G3**	**Financijski uvjeti**		
att25	integer	COD	*Dozvoljen minus*	avg = 7921 +/- 7283.530	[0; 29000]
att26	integer	TDB	*Oročeni depoziti (saldo na dan odobravanja kredita)*	avg = 1248.602 +/- 7892.941	[0; 107590]
att27	integer	BAL	*Saldo svih računa na dan odobravanja kredita (Potr.-Dug.)*	avg = -5353.451 +/- 7085.791	[-30632; 31338]
att28	real	BCO	*Saldo tekuće računa / Dozvoljen minus*	avg = -0.448 +/- 1.473	[-9; 9]
Grupa4		**G4**	**Podaci za posljednjih 12 mjeseci**		
att29	integer	TECO	*Koliko puta je klijent imao negativnu kamatu*	avg = 2.683 +/- 3.237	[0; 12]
att30	real	INPO	*Pozitivna kamata/ Negativna kamata*	avg = 2.157 +/- 3.092	[0; 10]
att31	integer	OINT	*Iznos negativne kamate*	avg = 7.514 +/- 25.154	[0; 369]
Grupa5		**G5**	**U vrijeme traženja kredita, koliko je klijent imao:**		
att32	integer	CHG	*Odobrenih kredita većeg iznosa od trenutnog*	avg = 0.212 +/- 0.491	[0; 3]
att33	integer	CHLE	*Odobrenih kredita manjeg ili jednakog iznosa od trenutnog*	avg = 0.504 +/- 0.750	[0; 4]
att34	integer	CHD	*U koliko je postojećih kredita klijent bio u zakašnjenju većem od 90 dana*	avg = 0.031 +/- 0.173	[0; 1]
att35	**binominal**	**IR**	**Urednost klijenta (0-neuredan, 1–uredan)**	mode = 1 (750), least = 0 (250)	0 (250), 1 (750)

Dodatak B

Tablica B.1 Odabrani atributi

Kod	Tehnika selekcije atributa					
	GA-NN	FS-NN	GINI	Gain Ratio	Korelacija	Voting
G1						
AGE			✓		✓	2
GENDER						0
POST						0
TLF	✓	✓				2
TAPA					✓	1
TWB	✓	✓	✓		✓	4
ACAGE		✓		✓	✓	3
MOA		✓				1
LMM	✓			✓	✓	3
POL						0
CRAM			✓			1
G2						
TIN			✓	✓		2
CSHT						0
RPT			✓	✓		2
OTIN	✓					1
TOUT	✓		✓	✓		3
TITO	✓	✓		✓		3
PSTW	✓					1
TMTW						0
SSTW						0
COTW		✓				1
RIDI	✓	✓	✓	✓	✓	5
RII	✓		✓	✓		3
G3						
COD						0
TDB		✓				1
BAL			✓	✓	✓	3
BCO	✓	✓	✓	✓	✓	5
G4						
TECO		✓			✓	2
INPO			✓	✓	✓	3
OINT	✓	✓	✓	✓	✓	5
G5						
CHG						0
CHLE						0
CHD	✓	✓			✓	3

Životopis

Stjepan Oreški rođen je 08. siječnja 1964. godine u Vukovoju. Diplomirao je 1985. na Sveučilištu u Zagrebu, na Fakultetu organizacije i informatike Varaždin. Na tom je fakultetu 1997. obranio magistarski rad, a 2014. i disertaciju. Akademske godine 1986./1987. stekao je Diplomu o dodatnom pedagoško-psihološkom obrazovanju na Filozofskom fakultetu u Zagrebu.

Oženjen je i ima dvoje djece.

U periodu od 1985.- 1989. godine zaposlen kao predavač informatičke grupe predmeta u Gimnaziji Karlovac i Tehničkoj školi Karlovac.

Od 1989.-1999. godine obnaša dužnost rukovoditelja Odjela za programsku i sistemsku podršku u Karlovačkoj banci. Stječe znanja i vještine kroz dodatno obrazovanje i osposobljavanje iz sljedećih područja:

- vođenje projekata,
- izrada poslovnih planova,
- analiza financijskih izvješća,
- SWOT analiza,
- administracija operativnog sustava AIX-Unix,
- uklanjanje problema na OS-u AIX ,
- administriranje i uklanjanje problema na Windows 2003 OS-u,
- razvoj aplikacija uz pomoć Oracle designera i developera te
- administriranje VMS-a.

Od 1999.- 2008. godine direktor je Sektora informatike u Karlovačkoj banci. U tom periodu glavne odgovornosti su mu:

- rukovođenje radom Sektora,
- izrada poslovnih planova za Sektor (godišnjih i višegodišnjih),
- vođenje Upravljačkog odbora za IT,
- vođenje ključnih projekata u Banci,
- komunikacija s poslovnim partnerima,
- rad u Nadzornom odboru MBU-a i
- organizacija IT edukacije u Banci.

Godine 2008. upisuje doktorski studij na Fakultetu organizacije i informatike Varaždin te istim danom preuzima mjesto Pomoćnika direktora Sektora informatike na kojem radi do danas.

Glavninu radnog vijeka provodi u bankarskoj industriji gdje unapređuje mnoge poslovne procese, a posebni interes iskazuje za područje kredita i kreditnih rizika. S tog područja je objavio četiri znanstvena rada, od kojih dva u jednom od vodećih časopisa na području umjetne inteligencije. Recenzent je u više znanstvenih časopisa s istog područja od kojih su neki po faktoru odjeka među 10 najboljih u svijetu.

Popis radova

Radovi u CC časopisima:

1. **Oreški, S**., Oreški, D., Oreški, G., Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment, Expert systems with applications, 39, 2012., str. 12605–12617.
2. **Oreški, S**., Oreški, G., Genetic algorithm-based heuristic for feature selection in credit risk assessment, Expert Systems with Applications, 41, 2014., str. 2052-2064.

Znanstveni radovi u drugim časopisima:

1. **Oreški S.,** Hybrid techniques of combinatorial optimization with application to retail credit risk assessment, Artificial Intelligence and Applications, Volume 1, Number 1, pp.21-43, 2014.

Znanstveni radovi u zbornicima skupova s međunar.rec.:

2. Oreški, G., **Oreški, S**., An experimental comparison of classification algorithm performances for highly imbalanced datasets, Central European Conference on Information and Intelligent Systems, Varaždin, 2014, , rad je nagrađen kao najbolji na konferenciji.

Magistarski rad

1. **Oreški, S**., Reinženjerstvo bankarskog informacijskog sustava, Fakultet organizacije i informatike, Varaždin, 1997.