SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1568

Primjena Fuzzy ARTMAP neuronske mreže za indeksiranje i klasifikaciju dokumenata

Stjepan Buljat

Zagreb, studeni 2005.

...mojoj obitelji, djevojci i prijateljima

Sažetak

Fuzzy ARTMAP – neuronska mreža natjecateljskog tipa s nadgledanim učenjem. Svrha – prepoznavanje kategorija i multidimenzionalnih mapa obzirom na ulazne vektore, koji mogu biti binarni ili analogni. Ovakva mreža je sinteza neizrazite (eng. *fuzzy*) logike i *Adaptive Resonance Theory – ART*, koju je smislio Stephen Grossberg. Koristi se komplementarno kodiranje ulaznih vektora kao metoda normalizacije, ali uz sačuvanje svojstava pojedinih ulaznih vektora. Koristi minimum-maksimum (eng. *minimax*) pravilo učenja kojim se minimizira greška predviđanja i maksimizira generalizacija – sustav uči minimalan broj kategorija koji će zadovoljiti preciznost klasifikacije. Donosi rješenje problemu stabilnost-plastičnost (eng. *stability-plasticity*), gdje je stabilnost sposobnost pamćenja starih uzoraka iako se uče novi uzorci, a plastičnost je sposobnost mreže da nauči nove uzorke jednako dobro kroz sve faze učenja.

Sadržaj

1. UVOD 1

2. AUTOMATSKA KLASIFIKACIJA TEKSTA 3

2.1. UVOD 3
2.2. KLASIFIKACIJA TEKSTA 4
2.2.1. DEFINICIJA 4
2.2.2. VRSTE KLASIFIKATORA 5
2.3. SKUP ZA UČENJE, SKUP ZA TESTIRANJE I VALIDACIJSKI SKUP 6
2.4. PRIKAZ TEKSTA 6
2.4.1. PRETPROCESIRANJE DOKUMENATA 8
2.5. INDEKSIRANJE TEKSTA 9

3. OCJENJIVANJE UČINKOVITOSTI KLASIFIKATORA 11

3.1. MJERE UČINKOVITOSTI KLASIFIKATORA 11
3.1.1. KOMBINIRANE MJERE 13

4. NEURO-RAČUNARSTVO 14

4.1. NEUROZNANOST 14
4.2. POVIJESNI RAZVOJ UMJETNIH NEURONSKIH MREŽA 15
4.3. TEMELJNI PRINCIPI UNM 17
4.3.1. VRSTE I STRATEGIJE UČENJA 19
4.3.2. TAKSONOMIJA UNM 20
4.3.3. UNM KAO AUTOMATSKI KLASIFIKATOR 21

5. ADAPTIVE RESONANCE THEORY 23

5.1. POVIJEST ART MODELA 23
5.2. POZADINA ART MODELA 24
5.3. TAKSONOMIJA ART ARHITEKTURA 25
5.3.1. NEIZRAZITA LOGIKA 26
5.3.2. FUZZY ART 28

5.3.2.1. Algoritam 29
5.3.2.2. Geometrijska interpretacija algoritma 30
5.3.3. FUZZY ARTMAP 33
5.3.3.1. Algoritam 34
5.3.3.2. Pojednostavljeni Fuzzy ARTMAP 36
5.3.3.3. Primjer rada pojednostavljenog FAM algoritma 37

6. ULAZNI PODACI 39

6.1. SKUPOVI DOKUMENATA 39
6.1.1. SKUP DOKUMENATA «AKADEMSKI PRIMJER» 39
6.1.2. SKUP DOKUMENATA «*CROATIA WEEKLY*» 42
6.1.3. SKUP DOKUMENATA «VJESNIK» 44
6.1.4. SKUP DOKUMENATA «*EUROVOC*» 46
6.2. PREDPROCESIRANJE PODATAKA 48

7. REZULTATI 51

7.1. «AKADEMSKI PRIMJER» 51
7.2. «CROATIA WEEKLY (EN)» 53
7.2.1. TEST – 1 53
7.2.2. TEST – 2 55
7.3. «CROATIA WEEKLY (HR)» 57
7.3.1. TEST – 1 57
7.3.2. TEST – 2 59
7.4. «VJESNIK» 61
7.4.1. TEST – 1 61
7.4.2. TEST – 2 63
7.5. «EUROVOC» 65
7.5.1. TEST – 1 66
7.5.2. TEST – 2 67
7.5.3. TEST – 3 68
7.6. RAZMATRANJE DOBIVENIH REZULTATA 69

8. ZAKLJUČAK 72

9. LITERATURA 73

Popis oznaka i kratica

Simboli

$A \cap B$	presjek skupova A i B
$A \cup B$	unija skupova A i B
$A \vee B$	neizrazita unija skupova A i B
$A \wedge B$	neizraziti presjek skupova A i B
$\lvert A \rvert$	L_1 norma vektora A
$a^c = 1 - a$	komplement vektora a

Notacija

$\mathbf{A}$	komplementarno kodirani ulazni vektor
C_J	*Match Criterion* funkcija pobjednika J
F_0	ulazni sloj
F_1	drugi sloj
F_2	izlazni sloj
F^{ab}	*Map Field*
$\mathbf{I}$	ulazni vektor
J	indeks neurona pobjednika
M	dimenzionalnost ulaznog vektora
N	broj neurona u sloju F_2
w^{ab}	težine *Map Field-a*
$\mathbf{w_j}$	težina neurona s indeksom *j* iz F_2 sloja
$\mathbf{X}^{ab}$	aktivacijski vektor *Map Field*-a
$\mathbf{x}$	aktivacijski vektor sloja F_1
$\mathbf{y}$	aktivacijski vektor sloja F_2
$\mathbf{Y^b}$	aktivacija izlaznog sloja ART_b modula
T_j	funkcija odabira za čvor j sloja F_2
ρ	parametar budnosti (eng. *vigilance parameter*)
$\rho_{baseline}$	početni parametar budnosti (eng. *baseline vigilance parameter*)
β	stopa učenja (eng. *learning rate*)
α	parametar odabira (eng. *choice parameter*)

Kratice

ART	*Adaptive Resonance Theory*
F-ART	*Fuzzy* ART
ARTMAP	*Adaptive Resonance Theory Map*
FAM	*Fuzzy* ARTMAP
NN	*Neural Network*
FNN	*Fuzzy Neural Network*
LTM	*Long-Term-Memory*
STM	*Short-Term-Memory*
MLP	*Multilayer Perceptron*
IRT	*Information Retrieval Techniques*
TC	*Text Categorisation*
KE	*Knowledge Engineering*
ML	*Machine Learning*
UI	Umjetna inteligencija
AI	*Artificial Intelligence*
UNM	Umjetne neuronske mreže
ANN	*Artificial Neural Networks*
DPC	*Document-Pivoted Categorization*
CPC	*Category-Pivoted Categorization*
TV	*Training(-and-validation) Set*
Te	*Test Set*
Tr	*Training Set*
Va	*Validation Set*
TP	*True Positives*
TN	*True Negatives*
FP	*False Positives*
FN	*False Negatives*
ADALINE	*ADAptive LINear Element*
MADALINE	*Multiple ADALINE*

Popis tablica

TABLICA 1. MATRICA ODLUKE ZA KLASIFIKACIJU TEKSTA 4
TABLICA 2. PRIKAZ ODLUČIVANJA KLASIFIKATORA I STRUČNJAKA ZA KATEGORIJU c_i 12
TABLICA 3. UNM TAKSONOMIJA S OBZIROM NA SMJER PROCESIRANJA INFORMACIJA I STRATEGIJE UČENJA 21
TABLICA 4. USPOREDBA GLAVNIH TIPOVA UNM-A [SAPOJNIKOVA2003] 22
TABLICA 5. POJEDNOSTAVLJENI FAM ALGORITAM (UČENJE I TESTIRANJE) 36
TABLICA 6. SKUP PODATAKA ZA UČENJE ZA POJEDNOSTAVLJENI FAM 37
TABLICA 7. SKUP ZA UČENJE «AKADEMSKOG PRIMJERA» U BINARNOM OBLIKU 41
TABLICA 8. SKUP ZA TESTIRANJE «AKADEMSKOG PRIMJERA» U BINARNOM OBLIKU 41
TABLICA 9. PARAMETRI KLASIFIKATORA ZA EKSPERIMENT «AKADEMSKI PRIMJER» 51
TABLICA 10. PARAMETRI KLASIFIKATORA ZA EKSPERIMENT 1 «CROATIA WEEKLY (EN)» 53
TABLICA 11. PARAMETRI KLASIFIKATORA ZA EKSPERIMENT 2 «CROATIA WEEKLY (EN)» 55
TABLICA 12. PARAMETRI KLASIFIKATORA ZA EKSPERIMENT 1 «CROATIA WEEKLY (HR)» 57
TABLICA 13. PARAMETRI KLASIFIKATORA ZA EKSPERIMENT 2 «CROATIA WEEKLY (HR)» 59
TABLICA 14. PARAMETRI KLASIFIKATORA ZA EKSPERIMENT 1 «VJESNIK» 61
TABLICA 15. PARAMETRI KLASIFIKATORA ZA EKSPERIMENT 2 «VJESNIK» 63
TABLICA 16. PARAMETRI INDEKSERA ZA EKSPERIMENT 1 «EUROVOC» 66
TABLICA 17. PARAMETRI INDEKSERA ZA EKSPERIMENT 2 «EUROVOC» 67
TABLICA 18. PARAMETRI INDEKSERA ZA EKSPERIMENT 3 «EUROVOC» 68

Popis slika

SLIKA 1. PRIKAZ BIOLOŠKOG NEURONA 15

SLIKA 3. PRIMJER JEDNOSTAVNE UMJETNE NEURONSKE MREŽE 18

SLIKA 4. STRATEGIJE UČENJA I PROBLEMI RASPOZNAVANJA UZORAKA [SAPOJNIKOVA2003] 20

SLIKA 6. SHEMA FUZZY ART MREŽE [BUSQUE1997] 28

SLIKA 7. (A) GEOMETRIJSKA INTERPRETACIJA FUZZY ART TEŽINSKOG VEKTORA (M=2). (B) GEOMETRIJSKA INTERPRETACIJA BRZOG UČENJA [CARPENTER1992] 31

SLIKA 8. FUZZY ARTMAP ARHITEKTURA [CARPENTER1992] 33

SLIKA 9. ULAZNI VEKTOR ZA SKUP DOKUMENATA "AKADEMSKI PRIMJER" 40

SLIKA 10. PODJELA ČLANAKA PO KATEGORIJAMA (SKUP ZA UČENJE) 43

SLIKA 11. PODJELA ČLANAKA PO KATEGORIJAMA (SKUP ZA TESTIRANJE) 43

SLIKA 12. PODJELA ČLANAKA PO KATEGORIJAMA 45

SLIKA 13. PRIKAZ HIJERARHIJSKE ORGANIZIRANOSTI EUROVOC POJMOVNIKA 47

SLIKA 14. DIJAGRAM TOKA PREDPROCESIRANJA DOKUMENATA 50

SLIKA 15. PRECIZNOST I ODAZIV PO KATEGORIJAMA 52

SLIKA 16. MAKROPROSJEČNA/MIKROPROSJEČNA PRECIZNOST I ODAZIV 52

SLIKA 17. MIKROPROSJEČNA F0.5 MJERA 52

SLIKA 18. PRECIZNOST PO KATEGORIJAMA 53

SLIKA 19. ODAZIVI PO KATEGORIJAMA 54

SLIKA 20. MAKROPROSJEČNA/MIKROPROSJEČNA PRECIZNOST I ODAZIV 54

SLIKA 21. MIKROPROSJEČNA F0.5 MJERA 54

SLIKA 22. PRECIZNOST PO KATEGORIJAMA 55

SLIKA 23. ODAZIV PO KATEGORIJAMA 56

SLIKA 24. MIKROPROSJEČNA/MAKROPROSJEČNA PRECIZNOST I ODAZIV 56

SLIKA 25. MIKROPROSJEČNA F0.5 MJERA 56

SLIKA 26. PRECIZNOST PO KATEGORIJAMA 57

SLIKA 27. ODAZIVI PO KATEGORIJAMA 58

SLIKA 28. MAKROPROSJEČNA/MIKROPROSJEČNA PRECIZNOST I ODAZIV 58

SLIKA 29. MIKROPROSJEČNA F0.5 MJERA 58

SLIKA 30. PRECIZNOSTI PO KATEGORIJAMA 59

SLIKA 31. ODAZIVI PO KATEGORIJAMA 60

SLIKA 32. MIKROPROSJEČNA/MAKROPROSJEČNA PRECIZNOST I ODAZIV 60

SLIKA 33. MIKROPROSJEČNA F0.5 MJERA 60

SLIKA 34. PRECIZNOSTI PO KATEGORIJAMA 61

SLIKA 35. ODAZIVI PO KATEGORIJAMA 62

SLIKA 36. MIKROPROSJEČNA/MAKROPROSJEČNA PRECIZNOST I ODAZIV 62

SLIKA 37. MIKROPROSJEČNA F0.5 MJERA 62

SLIKA 38. PRECIZNOSTI PO KATEGORIJAMA 63

SLIKA 39. ODAZIVI PO KATEGORIJAMA 64

SLIKA 40. MIKROPROSJEČNA/MAKROPROSJEČNA PRECIZNOST I ODAZIV 64

SLIKA 41. MIKROPROSJEČNA F0.5 MJERA 64
SLIKA 42. ISPRAVNI I GENERIRANI INDEKSI (PRIMJER DOBROG INDEKSIRANJA) 66
SLIKA 43. ISPRAVNI I GENERIRANI INDEKSI (PRIMJER PREKOMJERNOG PRIDJELJIVANJA INDEKSA) 67
SLIKA 44. ISPRAVNI I GENERIRANI INDEKSI (PRIMJER LOŠEG INDEKSIRANJA) 68
SLIKA 45. *FAM* APLIKACIJA (A) 75
SLIKA 46. *FAM* APLIKACIJA (B) 76
SLIKA 47. *FAM* APLIKACIJA (C) 77

1. Uvod

Danas je većini ljudi poznato da računalna tehnologija uvelike mijenja ljudski život, na ovaj ili onaj način. Stopa tehnološkog rasta računalne tehnologije je daleko ispred bilo koje druge, čovjeku poznate tehnologije. U zadnjih 60 godina dogodilo se mnoštvo čudesnih otkrića na ovom polju, od otkrića tranzistora[1], mikroprocesora[2], prvog operacijskog sustava[3] pa sve do uređaja za prepoznavanje govora (eng. *speech recognition devices*).

Ovaj diplomski rad govori o jednoj od dojmljivijih i očiglednijih promjena: o načinu na koji nam računalna tehnologija daje odgovore na pitanja koja ne «žele» dati odgovor. Radi se o tehnologiji koja se u zadnjih 15ak godina probila ispred ostalih – tehnologija analize podataka. Pod podacima mislimo na skup numeričkih vrijednosti koje predstavljaju veličine različitih svojstava objekata koje promatramo [Berthold2002]. Dakle, «analiza podataka» predstavlja procesiranje takvih podataka, s nekim predodređenim ciljem.

Tijekom posljednjih godina sami smo svjedoci sve većeg porasta količine informacija u digitalnom, elektronskom obliku. Nažalost, razvoj metoda za predstavljanje i interakciju s tolikim količinama podataka nije išao ukorak s porastom količine podataka. Unatoč tom nesrazmjeru, zbog potrebe za fleksibilnim pristupom tim podacima, u posljednjih 10 godina stasale su računalne tehnike za procesiranje tih podataka s ciljem dohvata informacija (eng. *information retrieval techniques - IRT*).

Konkretna tema ovog diplomskog rada je automatska klasifikacija i polu-automatsko indeksiranje teksta, što su jedne od tehnika dohvata informacija (IR). Klasifikacija teksta ili aktivnost predstavljanja tekstualnih zapisa «prirodnog» jezika s tematskim kategorijama, iz predefiniranog skupa, datira još iz ranih 1960.-ih, ali je tek početkom 1990.-ih postala «hit» u polju dohvata informacija (IR) zahvaljujući povećanom interesu i snažnijem hardveru [Sebastiani1999]. Indeksiranje je jedna

[1] 1947., John Bardeen, Walter Brattain i William Shockley

[2] 1971., Intel predstavlja prvi mikroprocesor na jednom čipu – Intel 4004.

[3] 1950.-ih, nekoliko OS-a : Fortran Monitor System, SAGE, SOS,...

od podvarijanti klasifikacije teksta, gdje se svaki dokument može predstaviti s više ključnih riječi iz zadanog kontroliranog rječnika.

Sve do kraja 80.-ih godina prošlog stoljeća najpopularniji pristup klasifikaciji teksta je bio «inženjerstvo znanja» (eng. *knowledge engineering - KE*), gdje su se ručno definirali skupovi pravila za predstavljanje znanja stručnjaka preko kojih se moglo klasificirati dokument pod određenu kategoriju. Početkom 90.-ih godina prošlog stoljeća ovaj je pristup izgubio na popularnosti i primat je predao tehnikama strojnog učenja (eng. *machine learning - ML*). U drugom pristupu nastoji se izgraditi automatski klasifikator koji će početno učiti na predefiniranom skupu dokumenata da bi kasnije mogao samostalno, automatski donositi odluke o klasifikaciji dokumenata. Prednosti drugog pristupa su očigledne, smanjenje troškova, praktičnost, podjednake razine preciznosti, ..., itd.

Postoje brojne tehnike strojnog učenja koje imaju istu primjenu, ali nemaju sve jednaka svojstva. Tako npr. statističke metode nemaju adekvatnu brzinu i kvalitetu analize podataka s obzirom na kontinuirano pojačavanje opterećenja s podacima. Ovakve metode imaju nekoliko nedostataka, nedostatak automatizma i fleksibilnosti, što zahtijeva uvođenje dodatnih ljudskih resursa. Stoga se javlja potreba za automatskim i inteligentnim metodama analize podataka i dohvata informacija.

U skladu s prethodnim zahtjevom, popularizirala se upotreba tehnika umjetne inteligencije (UI) (eng. *artificial intelligence – AI*), konkretno umjetnih neuronskih mreža (eng. *artificial neural networks*). Neuronske mreže (NM) su neovisne o modelu podataka i služe za višedimenzionalno nelinearno mapiranje ulaza i izlaza. U mnogim slučajevima predstavljaju idealno rješenje za probleme prepoznavanja uzoraka. Prednost NM je sposobnost učenja iako su podaci nekonzistentni, nedovršeni ili s šumom. Trenutno postoji jako puno tipova neuronskih mreža, koje imaju različite ili iste primjene, ali za probleme automatske klasifikacije teksta kao jedne od najboljih su one koje se temelje na *Adaptive Resonance Theory – ART* modulima koje je smislio Stephen Grossberg.

U nastavku slijedi detaljniji opis navedenih tehnologija, zajedno s konkretnom implementacijom Fuzzy ARTMAP neuronske mreže.

2. Automatska klasifikacija teksta

U posljednjih 10 godina interes za ovo polje računarstva je znatno porastao, prije svega zbog izrazitog rasta broja dokumenata u digitalnom obliku i potrebe da se ti dokumenti bolje razvrstaju, organiziraju, indeksiraju, klasificiraju,..., itd. Jedan od popularnijih pristupa rješavanju ovog problema je korištenje tehnika strojnog učenja, tj. gradnja automatskog klasifikatora teksta.

U ovom poglavlju riječ će biti o automatskoj klasifikaciji teksta i polu-automatskom indeksiranju teksta, zajedno s pregledom tehnika koje se koriste uz te dvije metode.

2.1. Uvod

Dakle, klasifikacija teksta (*TC*) nije ništa drugo nego predstavljanje pisanih tekstova (prirodnim jezikom) s tematskim kategorijama, iz predefiniranog skupa dokumenata [Sebastiani1999]. Prva istraživanja ovog polja računarstva sežu još u rane 1960.-te godine, ali tek se zapravo odnedavno počela posvećivati veća pozornost na mogućnosti klasifikacije teksta, pritom se misli na automatiziranu klasifikaciju teksta.

Kao što je rečeno postoje dva pristupa klasifikaciji teksta:

1) inženjerstvo znanja (*KE*)

2) strojno učenje (*ML*)

U prvom pristupu veći dio postupka se obavlja ručno, naime stručnjaci sami izrađuju skup pravila preko kojih se može pristupiti klasifikaciji teksta pod određenu kategoriju. Drugi pristup je puno jednostavniji i praktičniji za čovjeka, *ML* se temelji na generalnom induktivnom procesu koji automatski izrađuje automatski klasifikator teksta putem učenja određenog skupa dokumenata.

2.2. Klasifikacija teksta

2.2.1. Definicija

Klasifikacija je zadatak pridjeljivanja *Boolean*[4] vrijednosti svakom $(d_j, c_i) \in D \times C$ paru, gdje je $D = \{d_1, d_2, ..., d_n\}$ skup dokumenata koji se trebaju klasificirati, a $C = \{c_1, c_2, ..., c_n\}$ skup predefiniranih kategorija.

Tablica 1. Matrica odluke za klasifikaciju teksta

	d_1	...	d_j	...	d_n
C_1	a_{11}	...	a_{1j}	...	a_{1n}
...	...	...	...	...	...
c_i	a_{i1}	...	a_{ij}	...	a_{in}
...	...	...	...	...	...
c_m	a_{m1}	...	a_{mj}	...	a_{mn}

Vrijednost 1 za a_{ij} se interpretira kao odluka da se dokument d_j svrsta pod kategoriju c_i, dok vrijednost 0 znači da se taj isti dokument ne svrsta pod navedenu kategoriju [Sebastiani1999]. Dakle, zadatak klasifikatora je aproksimirati nepoznatu ciljnu funkciju $\Phi: D \times C \rightarrow \{1,0\}$ koja opisuje kako bi se dokumenti trebali klasificirati. Da bi klasifikator bio valjan on mora zadovoljavati dva svojstva:

1) kategorije su samo simboličke oznake i ne postoji nikakvo dodatno znanje o njima.

2) ne postoje vanjski podaci koji bi mogli utjecati na odluke klasifikacije, klasifikacija se mora temeljiti samo na podacima koje može dobiti iz dokumenata koje treba klasificirati.

[4] *Boolean* vrijednosti mogu poprimiti jednu od dvije vrijednosti, istina ili laž: {1,0}

2.2.2. Vrste klasifikatora

U osnovi postoje samo dvije vrste klasifikatora teksta, one su:

1) stožer-dokument klasifikacija (eng. *document-pivoted categorization*)

2) stožer-kategorija klasifikacija (eng. *category-pivoted categorization*)

Document-Pivoted Categorization (DPC) klasifikator za određeni dokument $d_j \epsilon D$ pronalazi sve kategorije $c_i \epsilon C$ pod koje bi se taj dokument mogao klasificirati, dok *Category-Pivoted Categorization (CPC)* nastoji za određenu kategoriju $c_i \epsilon C$ pronaći sve dokumente $d_j \epsilon D$ koji bi se mogli klasificirati pod tu kategoriju. *DPC* klasifikatori su jako korisni kad dokumenti koje treba klasificirati postaju dohvatljivi u različitim vremenskim periodima, npr. filtriranje elektronske pošte, dok *CPC* klasifikatori su pogodni kad skup kategorija nije unaprijed predodređen, nego se on može proširivati s dodatnim kategorijama. Ubuduće, pretpostavka je da se koristi *DPC* klasifikator.

S obzirom na odluku pridjeljivanja kategorije određenom dokumentu, klasifikator može donijeti konačnu odluku pripadnosti, ali može i generirati listu mogućih «kandidata», tj. kategorija, prema nekom unaprijed zadanom kriteriju. Ovaj drugi način koriste polu-automatski[5] klasifikatori, s namjerom da se olakša posao čovjeku pri konačnoj odluci o klasifikaciji.

[5] Polu-automatski klasifikatori ne obavljaju cjelokupan posao klasifikacije, oni zahtijevaju i ljudske resurse za potrebe konačne odluke o klasifikaciji.

2.3. Skup za učenje, skup za testiranje i validacijski skup

Tehnike strojnog učenja (*ML*) zahtijevaju posjedovanje inicijalnog skupa $\Omega = \{d_1, d_2, ..., d_{|\Omega|}\} \subset D$ dokumenata s poznatim pridijeljenostima kategorijama $C = \{c_1, c_2, ..., c_{|C|}\}$, tj. poznate su vrijednosti funkcije $\Phi: D \times C \rightarrow \{1,0\}$ za svaki par $(d_j, c_i) \in \Omega \times C$. Nakon što se klasifikator izgradi, potrebno je nekako ocijeniti njegovu učinkovitost, skup Ω se zato podijeli na dva dijela određene veličine:

1) skup podataka za učenje i validaciju $TV = \{d_1, ..., d_{|TV|}\}$, ovaj skup se koristi za učenje klasifikatora. Validacijski skup, kao podskup *TV* skupa, se koristi kod nekih metoda gdje je potrebno dodatno podesiti parametre klasifikatora prije samog testiranja.

2) skup podataka za testiranje $Te = \{d_{|TV|+1}, ..., d_{|\Omega|}\}$, koristi se za mjerenje učinkovitosti klasifikatora, tj. gleda se ispravnost odluka klasifikacije za pojedini dokument $d_j \in Te$.

Kad želimo podesiti parametre klasifikatora, kako bi sam klasifikator bio što učinkovitiji, tj. želimo pronaći optimalne parametre, onda se *TV* skup podataka dijeli na dva dijela, skup podataka za učenje $Tr = \{d_1, ..., d_{|Tr|}\}$ i validacijski skup $Va = \{d_{|Tr|+1}, ..., d_{|TV|}\}$. *Va* skup podataka se koristi za fino podešavanje parametara klasifikatora, na način da se ti podaci više puta predočavaju klasifikatoru kako bi dao što bolje rezultate. Ubuduće, pretpostavka je da se ne koristi validacijski skup podataka.

2.4. Prikaz teksta

Prije nego se dokumenti mogu automatski klasificirati, njih je potrebno dodatno preraditi ili predprocesirati u oblik koji automatski klasifikator «razumije». Stoga se nad dokumentima skupa za učenje, validaciju i testiranje provodi uniformni postupak izvlačenja značajki u oblik koji klasifikator zahtijeva. U praksi se tekstualni dokument d_j najčešće prikazuje kao vektor težina $d_j = (w_{1j}, ..., w_{|T|j})$, gdje je *T* skup izraza (značajki) koje se pojavljuju barem jedanput u barem jednom dokumentu skupa *Tr*, s tim da težine leže u intervalu $0 \leq w_{kj} \leq 1$. Težine predstavljaju koliko pojedini izraz t_k, doprinosi cjelokupnoj semantici dokumenta d_j [Sebastiani1999].

Imajući na umu prethodnu definiciju, u osnovi postoje dvije stvari na koje se može utjecati kako bi dobili različite prikaze teksta:

1) postoje različite ideje kako možemo tumačiti što je izraz.

2) postoje različite metode kako izračunati vektor-težine w_j

Obično se za (1) uzima da je izraz jednak jednoj riječi, ovakav pristup predstavljanja tekstualnog dokumenta se još naziva «vreća riječi» ili «skup riječi» (eng. *bag of words, set of words* respektivno) ovisno da li su težine analogne[6] ili binarne. Uz ovakav pristup postoje i pristupi gdje se izrazi identificiraju s frazama, ali takvi pristupi obično ne daju bolje rezultate učinkovitosti [Fuhr et al. 1991.].

Što se tiče pristupa (2), za vrijednost težina se obično uzima interval [0, 1], uz poneke iznimke gdje se koriste binarne vrijednosti, gdje 1 predstavlja prisutnost, a 0 nedostatak izraza u dokumentu. U bilo kojem slučaju, za prikaz tekstulnog dokumenta koristi se vektor težina koje predstavljaju sve bitne značajke dokumenta, postoji cijeli niz metoda kako se izračunavaju težine, ali jedna od najčešćih i najpopularnijih je *TFIDF* metoda. *TFIDF* metoda koristi *tfidf* funkciju, definiranu ovako:

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|T_r|}{\#_{Tr}(t_k)} \qquad (1)$$

Gdje $\#(t_k, d_j)$ označava frekvenciju pojavljivanja izraza t_k u dokumentu d_j, $\#_{Tr}(t_k)$ označava broj dokumenata iz skupa *Tr* u kojima se izraz t_k pojavljuje [Salton1988]. Da bi vektor-težine zadovoljavale kriterij jednake duljine, tj. da su te težine normalizirane onda se *tfidf* funkcija mora normalizirati s kosinus normalizacijom:

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} (tfidf(t_s, d_j))^2}} \qquad (2)$$

Ova funkcija se temelji na pretpostavkama da što se više puta jedan izraz pojavljuje unutar jednog dokumenta onda on bolje predstavlja sadržaj tog

[6] Pod izrazom analogne težine misli se na sposobnost poprimanja bilo koje realne vrijednosti iz intervala [0, 1].

dokumenta, te što se više puta jedan izraz pojavljuje unutar više dokumenata onda njegovo značenje postaje manje bitno za klasifikaciju. Jedina mana ovakvih metoda je leksički pristup izvlačenja značajki iz dokumenata, naime ne posvećuje se nikakva pozornost na semantiku dokumenta jer poredak riječi, konkretno u *tfidf* funkciji, ne znači ništa. Unatoč tomu, brojni autori se i dalje priklanjaju tzv. leksičkoj-semantici[7] jer u protivnom bi se moralo potrošiti mnogo više ljudskih i računalnih resursa kako bi se eventualno postigli nešto bolji rezultati nego inače.

2.4.1. Predprocesiranje dokumenata

Prije nego se provede postupak izračunavanja vektor-težina poželjno je predprocesirati dokumente, u smislu da se iz njih izbace nepotrebne riječi, koje dodatno otežavaju postupak klasifikacije i povećavaju dimenzionalnost hiper-prostora u kojem su smješteni dokumenti. Postoje dva koraka koja se mogu provesti:

1) eliminacija stop-riječi
2) morfološka normalizacija (eng. *stemming*)

Prvi korak je skoro uvijek obvezan jer je implementacija jednostavna, a znatno doprinosi poboljšanju učinkovitosti klasifikatora. Cilj je izbaciti sve one riječi koje nemaju nikakav doprinos značenju (semantici) dokumenta, pritom se misli na prijedloge, zamjenice, priloge, veznike, punktuacijske znakove i slično. Mogu se dodatno izbaciti i riječi čija je frekvencija pojavljivanja ispod ili iznad nekog praga, jer jako frekventne ili slabo frekvente riječi ne doprinose značajno semantici dokumenta.

Drugi korak je malo složeniji za implementaciju, jer je ovisan o jeziku na kojem su dokumenti pisani. To je postupak u kojem se riječi koje dijele isti morfološki korijen grupiraju kako bi se smanjila dimenzionalnost hiper-prostora dokumenata. Kao što je poznato, neki jezici su morfološko bogatiji od drugih pa je za njih puno teže izgraditi kvalitetan morfološki normalizator.

[7] Pristup u kojem se semantika dokumenta promatra isključivo kroz neovisno promatranje izraza, bez interakcije među njima.

2.5. Indeksiranje teksta

Indeksiranje teksta je postupak stvaranja opisa dokumenta iz njegovog sadržaja ili ponekad teme koju obrađuje dokument. Nije isključeno ni da u tom formalnom indeksnom obliku dokumenta budu uključeni atributi poput imena autora, imena dokumenta, bibliografskog konteksta i sl. [Nordlie2001].

Analogno tomu, automatsko indeksiranje teksta je postupak koji se provodi automatski na dokumentima u digitalnom obliku, uz neznatan utjecaj čovjeka. Prvi postupci automatskog indeksiranja su provedeni još davnih 1950.-ih godina od izvjesnog gospodina Luhn-a, ali u operativnom smislu, automatsko indeksiranje je zaživjelo tek početkom 1990.-ih godina, pojavom Internet-a.

Uz ove dvije vrste indeksiranja, ručnog i automatskog, postoji još i polu-automatsko indeksiranje gdje indekser daje prijedloge ključnih riječi, a čovjek (indeksator) odabire od ponuđenog samo one ključne riječi za koje smatra da su bitne za taj dokument. Ova vrsta indeksiranja će se od sada primjenjivati u ovom diplomskom radu.

S obzirom na način odabira ključnih riječi (eng. *index*) postoje dvije vrste indeksatora [Cvitaš2005]:

1) ekstrakcija ključnih riječi (eng. *keyword extraction*)

2) dodjeljivanje ključnih riječi (eng. *keyword assignement*)

U slučaju (1) koriste se ključne riječi koje se već nalaze u tekstu dokumenta. To se može odnositi na samostalne riječi, ali i na složenije fraze. Problem predstavlja određivanje koji izrazi imaju veću važnost od drugih, u semantičkom smislu. U slučaju (2), za dodjeljivanje ključnih riječi koriste se vanjski izvori, taj vanjski izvor je najčešće kontrolirani rječnik (tezaurus) – skup pojmova koji su na neki način međusobno povezani. Postoji više vrsta tezaurusa, ali u ovom kontekstu koristi se tematski tezaurus. Tematski tezaurus ima tematsko ustrojstvo zajedno sa semantičkom organizacijom pojmova, koja se ostvaruje putem hijerarhijske strukture. Na taj način je jednostavno koristiti više pojmova u izgradnji nekog šireg koncepta. Dakle, cilj je pridjeljivati riječi[8] iz kontroliranog rječnika svakom

[8] Pridjeljene riječi se još nazivaju deskriptori.

dokumentu, na taj način je riješen problem višeznačnosti koji se javlja pri radu s višejezičnim skupovima podataka. Upotreba kontroliranog rječnika se u svakom slučaju preporučuje jer pruža daleko veće mogućnosti nego korištenje pojmova iz samih dokumenata, slučaj (1).

Tipičan postupak automatskog indeksiranja procesira tekstualne dokumente kroz nekoliko koraka [Nordlie2001]:

1) redukcija rječnika (eng. *vocabulary reduction*)

2) računanje težina (eng. *weighting*)

3) lingvistička analiza (eng. *linguistic analysis*)

U prvom koraku izbacuju se stop-riječi, radi se morfološka normalizacija, ..., itd. U drugom koraku promatraju se odnosi frekvencija pojavljivanja pojmova u pojedinom dokumentu i cijelom skupu dokumenata, promatra se duljina dokumenata, ..., itd. Treći korak je i najsofisticiraniji, nije često u uporabi, jer temelje vuče iz jezične analize teksta: prepoznavanje homonima[9], antonima[10], sinonima[11] i sl.

Navedeni postupak jako sliči postupcima automatske klasifikacije teksta, zapravo radi se o gotovo istom tipu problema. U oba slučaja nastoji se nekom tekstu pridijeliti simboličku vrijednost, kod klasifikacije je to ime kategorije, a kod indeksiranja je to ključna riječ. Jedina razlika je što se kod indeksiranja koristi višestruko pridjeljivanje ključnih riječi jednom dokumentu, premda i kod klasifikacije teksta postoje dokumenti koji istovremeno mogu pripadati nekolicini kategorija.

Dakle, uz sitne preinake klasifikator teksta se može pretvoriti u indeksatora teksta, potrebno je samo pojam ključnih riječi zamijeniti s pojmom kategorija.

[9] Homonim je jedna riječ s više značenja.

[10] Antonimi su riječi koje imaju suprotno značenje, npr. dan-noć.

[11] Sinonimi su različite riječi istog značenja.

3. Ocjenjivanje učinkovitosti klasifikatora

Kao i kod drugih sustava, ocjena rada klasifikatora teksta se provodi eksperimentalno, a ne analitički. Razlog tome jest što analitička procjena rada zahtijeva poznavanje cijelog formalnog karaktera klasifikatora. To predstavlja velik problem jer je jako teško objektivno formalizirati pojedini klasifikator.

Ocjenjivanje učinkovitosti klasifikatora provedeno je po naputcima autora članka [Sebastiani1999], gdje je navedeno više metoda provjere. Iako postoje brojne metode mjerenja učinkovitosti sustava, kako bi se što preciznije moglo uspoređivati različite metode, najbitnije je pri provjeri koristiti iste skupove podataka s istom podjelom skupa dokumenata za učenje/validaciju i skupa dokumenata za testiranje. Samo tako možemo reći da su metode ravnopravno uspoređene.

3.1. Mjere učinkovitosti klasifikatora

Standardne mjere učinkovitosti klasifikacije su preciznost i odaziv.

Preciznost Pr_i se definira kao uvjetna vjerojatnost $p(ca_{ix} = 1 \mid a_{ix} = 1)$, tj. vjerojatnost da je klasifikacija nasumično odabranog dokumenta d_x u kategoriju c_i, točna. Odaziv Re_i se definira kao uvjetna vjerojatnost $p(a_{ix} = 1 \mid ca_{ix} = 1)$, tj. vjerojatnost da se, ako dokument d_x stvarno pripada kategoriji c_i, odluka o toj klasifikaciji stvarno i dogodi.

Što su zapravo preciznost i odaziv? To su numeričke vrijednosti koje predstavljaju mjere očekivanja korisnika sustava za klasifikaciju da će sustav ispravno klasificirati slučajno odabrani dokument u kategoriju c_i. U tablici (2) mogu se vidjeti odluke za jednu kategoriju.

Tablica 2. Prikaz odlučivanja klasifikatora i stručnjaka za kategoriju c_i

kategorija c_i		**stručno mišljenje**	
		DA	NE
odluka klasifikatora	DA	TP_i	FP_i
	NE	FN_i	TN_i

Oznake u tablici predstavljaju:

1) TP_i = *True Positives*, broj dokumenata ispravno klasificiranih pod kategoriju c_i.

2) FP_i = *False Positivies*, broj dokumenata netočno klasificiranih pod kategoriju c_i.

3) TN_i = *True Negatives,* broj dokumenata koji nisu klasificirani pod kategoriju c_i, a trebali bi biti.

4) FN_i = *False Negatives*, broj dokumenata koji su ispravno neklasificirani pod kategoriju c_i.

Navedene mjere se koriste za izračunavanje preciznosti, odaziva i ostalih mjera učinkovitosti sustava za klasifikaciju, tako uz pomoć dobivenih rezultata tijekom faze testiranja klasifikatora možemo izračunati preciznost i odaziv po formulama:

$$\mathrm{Pr}_i^{est} = \frac{TP_i}{TP_i + FP_i} \qquad \mathrm{Re}_i^{est} = \frac{TP_i}{TP_i + FN_i} \qquad (3) \quad (4)$$

Za procjenu vrijednosti faktora preciznosti i odaziva koriste se dvije metode:

1) mikroprosjek – preciznost i odaziv se određuju globalnim sumiranjem po svim individualnim odlukama:

$$\mathrm{Re}^{\mu\,est} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} (TP_i + FN_i)} \qquad \mathrm{Pr}^{\mu\,est} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} (TP_i + FP_i)} \qquad (5) \quad (6)$$

2) makroprosjek – preciznost i odaziv se ocjenjuju lokalno za svaku kategoriju s traženjem srednje vrijednosti po svim rezultatima za različite kategorije:

$$\Pr^{M} \overset{est}{=} \frac{\sum_{i=1}^{m} \Pr_i}{m} \qquad \text{Re}^{M} \overset{est}{=} \frac{\sum_{i=1}^{m} \text{Re}_i}{m} \qquad (7) \quad (8)$$

Navedene metode mogu dati različite rezultate, posebno ako su različite kategorije nejednako popunjene. Još nema odgovora koja metoda je bolja.

3.1.1. Kombinirane mjere

Ni preciznost ni odaziv nemaju smisla ako su međusobno izolirane, tako da bi dobili klasifikator s 100% odazivom potrebno je postaviti svaki faktor ograničenja[12] na vrijednost 0, te bi tako dobili *trivijalnog prihvatioca*[13]. Poznati su slučajevi kad su preciznost i odaziv obrnuto proporcionalni. U praksi, finim podešavanjem faktora ograničenja reguliramo faktore preciznosti i odaziva, ali najčešće u obrnuto proporcionalnim odnosima. Stoga je potrebno naći metodu za podešavanje faktora ograničenja kako bi dobili nama odgovarajući omjer preciznosti i odaziva. Neke od metoda su:

1) *(interpolated) 11-point average precision* – svako ograničenje se postavlja na vrijednosti na kojima odaziv poprima vrijednosti 0.0, 0.1, ..., 0.9 i 1.0. Za ovih 11 ograničenja računa se preciznost.

2) *breakeven point* – vrijednost pri kojoj je *Pr* = *Re*

3) F_α funkcija, $0 \leq \alpha \leq 1$

$$F_\alpha = \frac{1}{\alpha \cdot \frac{1}{\text{Pr}} + (1-\alpha) \cdot \frac{1}{\text{Re}}} \qquad (9)$$

[12] Faktor ograničenja je prag koji klasifikator mora zadovoljiti kako bi pojedini dokument klasificirao pod neku kategoriju.

[13] Trivijalni prihvatioc klasificira sve dokumente pod sve kategorije.

4. Neuro-računarstvo

U ovom poglavlju bit će više riječi o umjetnim neuronskim mrežama (UNM), tj. o pokušaju modeliranja bioloških neuronskih mreža. Kroz nekoliko podpoglavlja obradit će se podrijetlo i povijest razvoja neuronskih mreža, različite vrste i tipovi, te u konačnici i formalna definicija Fuzzy ARTMAP (FAM) modela neuronske mreže.

4.1. Neuroznanost

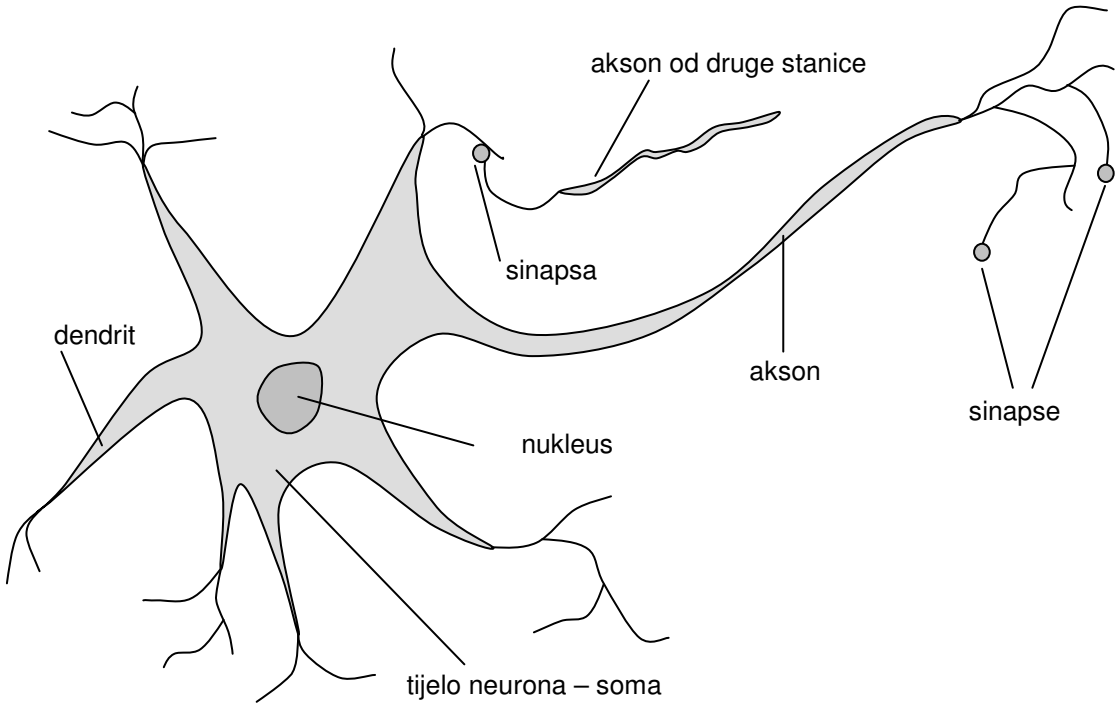
Neuroznanost je znanost koja proučava živčani sustav, osobito mozak. Neuro-računarstvo u osnovi vuče korijene iz neuroznanosti, iz proučavanja najvažnijih stanica živčanog sustava – neurona. Od davnina je poznato da mozak sudjeluje u postupcima ljudskog razmišljanja, ali tek se krajem 19. stoljeća, točnije 1861. godine, posvetila veća medicinska pozornost polju neuroznanosti od strane znanstvenika Paul Broca[14] (1824 - 1880). Nakon toga razvoj neuroznanosti nije više trebalo pogurivati, a neki od zaslužnijih istraživača su bili Camillo Golgi, S. Ramon y Cajal, Hans Berger i ostali.

Poznato je da se mozak sastoji od živčanih stanica ili neurona, na slici (1) nalazi se jedan takav neuron sa svim svojim dijelovima. Svaki neuron se sastoji od tijela stanice (soma), koja sadrži jezgru stanice (nukleus). Iz stanice se granaju brojne niti koje se nazivaju dendriti i jedna dugačka nit koja se naziva akson. Jedan neuron stvara veze s 10 do 100000 drugih neurona na spojnicama koje se nazivaju sinapse. Signali se propagiraju sa specifičnim elektrokemijskim reakcijama. Signali kontroliraju kratkotrajnu aktivnost mozga, ali i sudjeluju u dugotrajnim promjenama pozicije i veza neurona.

Polje umjetnih neuronskih mreža nastoji s jednostavnim aproksimacijama simulirati biološke neuronske mreže s ciljem rješavanja različitih vrsta problema i u tome uspijeva.

---

[14] Paul Broca je istraživao gubitke sposobnosti govora kod pacijenata s oštećenim mozgom.

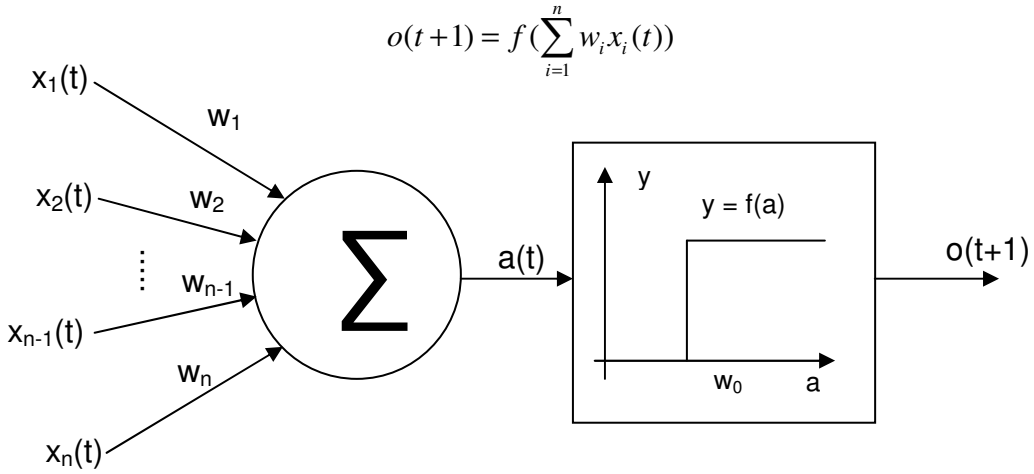


Slika 1. Prikaz biološkog neurona

4.2. Povijesni razvoj umjetnih neuronskih mreža

Mnogi bi se iznenadili kad bi shvatili da je razvoj UNM u blagom smislu te riječi bio kontroverzan, bilo je mnogo uspona i padova u tako malo godina (zadnjih 60ak godina). Stoga se povijesni razvoj može podijeliti na nekoliko cjelina:

1) **Prvi pokušaji** – 1943. neurofiziolog Warren McCulloch i matematičar Walter Pitts napisali su rad o funkcioniranju neurona temeljen na pretpostavkama. Izradili su jednostavne neuronske mreže gdje su neuroni bile binarne jedinice s fiksnim pragovima (eng. *threshold*). Uspjeli su modelirati jednostavne logičke funkcije poput **I**, **ILI** i **NE**. Na slici (2) se može vidjeti njihov model neurona. Nakon toga, 1949. Donald Hebb je napisao rad «*The Organization of Behavior*» gdje je objasnio model učenja neurona s promjenom težina veza među neuronima, takvo učenje se naziva **Hebbovo učenje** [Russel2003].



Slika 2. McCulloch - Pitts model umjetnog neurona [BertholdHand2002]

2) **Obećavajuća tehnologija** – ubrzo je ovo polje računarstva postalo trend pa su neurofiziolozi, inženjeri i mnogi znanstvenici nastojali doprinijeti razvoju UNM. 1958. Frank Rosenblatt je razvio **perceptron**, jednostavna NM arhitektura s tri sloja neurona od kojih se srednji sloj nazivao asocijativni sloj. Perceptron je mogao povezati dani ulaz s nekim slučajnim izlazom. 1959. Bernard Widrow i Marcian Hoff su razvili **ADALINE** i **MADALINE** sustav, analogni elektronski sustavi raspoznavanja binarnih uzoraka [StanfordProjects00].

3) **Mračno doba** – 1969. Marvin Minsky i Seymour Papert su napisali knjigu «*Perceptrons*» gdje su generalizirali ograničenja jednoslojnih perceptrona na višeslojne perceptrone [Stergiou1996]. Utjecaj navedene knjige je bio toliko značajan da su se ukinula sva financijska ulaganja u područje istraživanja UNM, nastupilo je tzv. mračno doba UNM-a koje je trajalo sve do 80.-ih godina prošlog stoljeća.

4) **Marginala neuroračunarstva** – unatoč nedostatku ulaganja, nekolicina istraživača je nastavila raditi na polju UNM-a. Neki od pionira neuroračunarstva su bili Stephen Grossberg koji je 1976. razvio ART model ljudskog shvaćanja da bi krajem 1980.-ih, zajedno s Gail Carpenter, razvio nekolicinu ART algoritama. Tu su bili i Anderson i Kohonen koji su

neovisno razvili asocijativne tehnike. 1974. Paul Werbos je razvio *back-propagation* algoritam učenja, primjenjuje se na višeslojni perceptron. Neki od inovatora su bili Amari Shun-Ichi, Fukushima Kunihiko, ..., itd. [Stergiou1996]

5) **Buđenje** – zahvaljujući prethodno navedenim istraživačima i njihovim uspjesima, krajem 1970.-ih započinje nova era neuroračunarstva koja traje sve današnjih dana. Ulagači su napokon prepoznali mogućnosti i možemo reći da je uloženi trud višestruko vratio uložena sredstva.

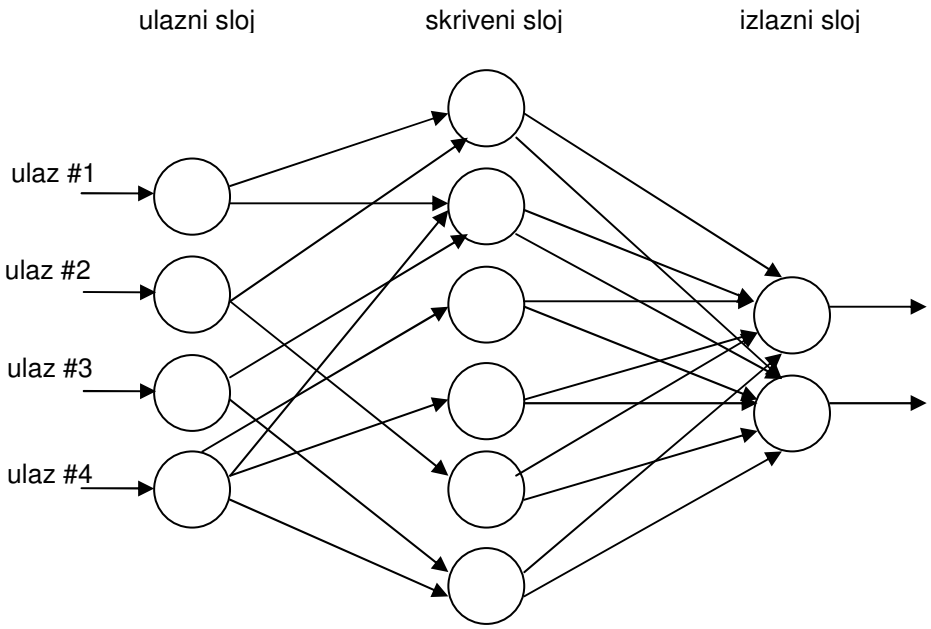
4.3. Temeljni principi UNM

S matematičkog gledišta, UNM imaju sposobnost adaptivnog predviđanja kontinuirane funkcije iz neobrađenih podataka. Točnije, UNM je računalni model za procesiranje informacija koji predstavlja usmjereni graf sastavljen od neurona i veza među njima (sinapse) [Sapojnikova2003]. Svaki čvor (neuron) *i* ima svoju staničnu aktivnost x_i, i svaka sinapsa između dva čvora *i* i *j* ima svoju težinsku vrijednost w_{ij}. Obično se težine nazivaju dugotrajna memorija (eng. *Long Term Memory – LTM*) jer se sporije mijenjaju nego izlazi čvorova, dok se stanična aktivnost naziva kratkotrajna memorija (eng. *Short Term Memory – STM*) jer ona predstavlja prolazne odgovore mreže na različite ulaze. Mreža obično ima slojevitu strukturu, gdje se dio čvorova naziva ulaznim, a dio izlaznim, a signal (ulaz) koji se dovodi na ulazne čvorove se propagira do izlaznih.

Generalni model umjetnog neurona je temeljen na McCulloch-Pitts modelu neurona. Takav neuron računa svoj izlaz S_j sumirajući vrijednosti ulaznih signala koji stižu putem sinapsi s određenom težinskom vrijednosti, da bi taj signal propustio kroz funkciju praga *f(x)* i propagirao ga neuronima u nastavku ako iznos prijeđe neki određeni prag postavljen funkcijom praga. Detaljniji prikaz takvog neurona se može vidjeti na slici (2). Dakle, umjetni neuroni obavljaju funkciju biološkog neurona, aktivnost neurona ovisi o izlaznim signalima neurona koji su direktno spojeni na njega i o težinama tih sinapsi.

Umjetna neuronska mreža (UNM) je skupina međusobno povezanih umjetnih neurona koja koristi matematički ili računalni model za procesiranje informacija

temeljen na konektivističkom[15] pristupu. Ne postoji predefinirani odgovor na pitanje što je to UNM, ali postoji generalno mišljenje da se pod UNM-om podrazumijeva mreža jednostavnih procesnih elemenata, gdje je globalno ponašanje mreže određeno sinapsama među procesnim elementima i njihovim parametrima [Wikipedia]. Na slici (3) nalazi se jedan tipičan primjer umjetne neuronske mreže.



Slika 3. Primjer jednostavne umjetne neuronske mreže

[15] Konektivizam je pristup u poljima umjetne inteligencije, kognitivne znanosti, neuroznanosti, ..., itd. On modelira mentalni fenomen putem jednostavnijih pravila, korištenjem mreže međusobno spojenih jednostavnih jedinica [Wikipedia].

4.3.1. Vrste i strategije učenja

Učenje ili adaptacija UNM se obično ostvaruje promjenama sinapsnih težina i ostalih parametara mreže. Prvi neurobiološki model učenja uveo je prethodno navedeni Donald Hebb. Nakon učenja mreža može dati odaziv (eng. *recall*) na postavljeni ulaz, tj. može dati izlazni signal kao odgovor na postavljeni ulazni signal. Kao što je već navedeno, potrebno je podatke razdvojiti u dva skupa, skup za učenje i skup za testiranje, kako bi mrežu mogli naučiti i testirati naučeno znanje.

Postoje dvije paradigme učenja UNM-a:

1) ***Off-line (batch)* učenje** – mreži se istovremeno daju svi podaci za učenje i dopušteno joj je koristiti te podatke u bilo kojem trenutku.

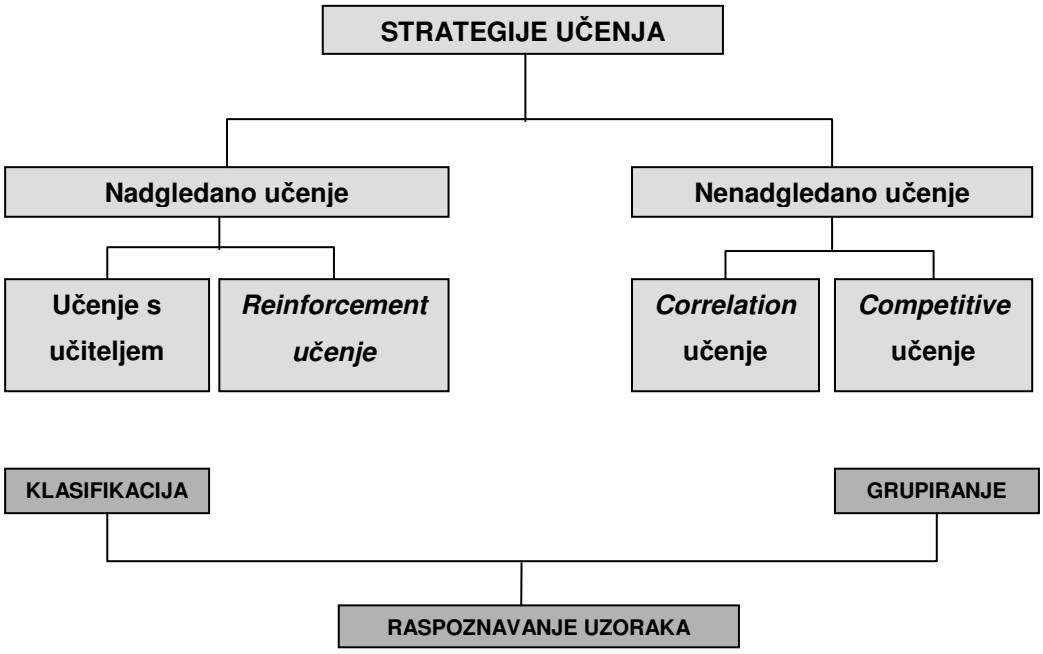
2) ***On-line* (inkrementalno) učenje** – mreža dobiva jedan po jedan uzorak za učenje, nakon učenja mreža više nema pristup tom uzorku.

Postupak *off-line* učenja završava kad se postignu optimalne performanse mreže nakon višestrukog predstavljanja skupa za učenje (epohe). Za podatke koji se vremenski mijenjaju potrebno je ponovno napraviti postupak učenja. Prednost *on-line* učenja je što se može prilagoditi dinamičkim okruženjima, npr. procesiranje signala, robotska kontrola, ..., itd.

Postoje i dvije vrste učenja, obzirom na brzinu učenja, **brzo** i **sporo učenje**. Za kontrolu brzine učenja koristi se parametar stope ili brzine učenja $\beta \in [0, 1]$, niže vrijednosti odgovaraju sporom učenju, a više bržem. Sporo učenje znači da *LTM* težine dolaze u ravnotežu postupno, ne naglo. Dok brzo učenje trenutno kodira ulazni uzorak u težinu sinapse, što omogućuje učenje u trajanju od svega par epoha. Mana sporog učenja je neosjetljivost na iznimke u podacima, rijetke slučajeve, dok brzo učenje bilježi takve slučajeve. Jako je važno dobro izbalansirati tehnike sporog i brzog učenja.

Strategije učenja se dijele na nadgledano (eng. *supervised*) i nenadgledano (eng. *unsupervised*) učenje. Kod nadgledanog učenja, uz ulazni uzorak koji se predstavlja mreži daje se još i nekakva informacija o ispravnosti odluke koju je mreža dala za taj uzorak, kod klasifikacijskog problema ta informacija bi bila ispravna kategorija ulaznog dokumenta. U slučaju nenadgledanog učenja, ne postoji *a priori* znanje o željenom izlazu, mreža mora samostalno analizirati ulazne

podatke kako bi dala smisleni izlaz. Ovakva vrsta učenja se obično koristi kod problema grupiranja podataka (eng. *clustering*). Na slici (4) prikazana je detaljna podjela strategija po vrstama učenja.



Slika 4. Strategije učenja i problemi raspoznavanja uzoraka [Sapojnikova2003]

4.3.2. Taksonomija UNM

Područje neuroračunarstva se jako brzo razvija i zbog toga danas postoji preko 20 algoritama umjetnih neuronskih mreža. Kako bi se omogućila usporedba pojedinih algoritama potrebno je napraviti taksonomiju, tj. podjelu po vrstama neuronskih mreža, ali problem je što još ne postoji generalna taksonomija, već postoji cijeli niz mogućih taksonomija [Sapojnikova2003]. Podjela se može ostvariti prema nekoliko kriterija: strategiji učenja, operacijskom svojstvu mreže, strukturi neurona, ..., itd. Pod operacijskim svojstvom mreže misli se na smjer procesiranja informacija i strategiji učenja mreže. S obzirom na smjer procesiranja informacija u mreži postoje **unaprijedne (eng. *feedforward*)** i **unazadne (eng. *feedback*)** mreže, a s obzirom na strategije učenja postoje ***supervised*** i ***unsupervised*** mreže. Unaprijedne mreže prenose signale u jednom smjeru, od ulaza prema izlazu, dok unazadne mreže mogu prenositi signale u oba smjera i one se češće

koriste. Prema [Sapojnikova2003] jedna od mogućih taksonomija s obzirom na operacijska operacijska svojstva mreže se nalazi u tablici (3).

Tablica 3. UNM taksonomija s obzirom na smjer procesiranja informacija i strategije učenja

STRATEGIJA UČENJA	UNAPRIJEDNO PROCESIRANJE	UNAZADNO PROCESIRANJE
NENADGLEDANO UČENJE	Linear Associative Memory (LAM)	Hopfield network Bidirectional Associative Memory (BAM)
NADGLEDANO UČENJE	Multilayer perceptron (MLP) Radial Basis Functions (RBF)	Recurrent MLP
NENADGLEDANO ILI NAGLEDANO UČENJE	Self-Organizing Map (SOM) Counterpropagation (CP)	Adaptive Resonance Theory (ART)

4.3.3. UNM kao automatski klasifikator

Prije primjene umjetnih neuronskih mreža kao automatskih klasifikatora teksta potrebno je analizirati svojstva određene UNM preko niza kriterija prema preporuci autora [Sapojnikova2003]. Kriteriji se mogu svrstati u dvije grupe:

1) kriteriji koji se tiču automatizacije UNM-a:
   a. **razina automatizacije** – ovo je najvažnija karakteristika umjetne
   b. neuronske mreže s obzirom na automatsku implementaciju, veća razina automatizacije se postiže smanjenom interakcijom korisnika i sustava (sudjelovanje u postupcima učenja i testiranja). Zbog toga je poželjno da mreža ima minimalan broj parametara, tj. mreža s topološkom samoorganizacijom.
   c. **on-učenje** – svojstvo koje mreži omogućuje dinamičko učenje, tj. individualno prikazivanje ulaznih uzoraka.
   d. **kratko vrijeme učenja** – ovo vrijeme se obično definira kao broj epoha koje su potrebne da mreža postigne optimalne klasifikacijske rezultate.

2) kriteriji koji tiču funkcionalnih svojstava mreže koja su korisna:

   a. **stabilnost učenja** – klasifikator bi trebao brzo učiti nove podatke, ali bez zaboravljanja starog znanja. Postoji ustupak za ustupak (eng. *trade-off*) između stabilnosti i plastičnosti, kojeg je formulirao Stephen Grossberg kao stabilnost-plastičnost dilema (eng. *stability-plasticity dilemma*) : «Kako sustav za učenje može biti dovoljno stabilan da zapamti naučene uzorke i dovoljno fleksibilan da nauči nove uzorke?».

   b. **otpornost na šum** – poželjna karakteristika automatskog klasifikatora.

   c. **kompaktna prezentacija znanja** – stvaranje prikladne prezentacije znanja, informacija je vrlo važno za automatski klasifikator. Cilj automatskog klasifikatora je minimizirati internu prezentaciju informacija uz maksimizaciju preciznosti predviđanja.

Trenutno ne postoji niti jedna umjetna neuronska mreža koja zadovoljava sve navedene kriterije. U tablici (4) prikazana je usporedba nekih sustava s obzirom na navedene kriterije.

Tablica 4. Usporedba glavnih tipova UNM-a [Sapojnikova2003]

SVOJSTVO	MLP	RBF	SOM	CP	ART
Topološka samoorganizacija	-/+	-/+	-/+	-	+
***On-line* učenje**	-/+	-	-	-	+
Kratko vrijeme učenja	-	-	-	-	+
Stabilnost učenja	-	-	-	-	+
Otpornost na šum	+	+	+	+	-
Kompaktna prezentacija znanja	+	+	-	-	-

5. Adaptive Resonance Theory

U ovom poglavlju riječ je skupini umjetnih neuronskih mreža koje se temelje na *ART* modelu spoznajne teorije. Kroz nekoliko podpoglavlja obrađena je povijest *ART* modela, taksonomija *ART* arhitektura te temeljni principi *Fuzzy ART* i *Fuzzy ARTMAP* neuronskih mreža.

5.1. Povijest ART modela

ART model je temeljen isključivo na neurofiziološkim podacima, uveo ga je istraživač Stephen Grossberg 1976. godine pri bostonskom sveučilištu (Boston, SAD). «*Adaptive Resonance Theory nastoji objasniti zašto, kao što su se filozofi pitali godinama, ljudi mogu biti 'namjerna' (eng. intentional) bića – planirajući buduća ponašanja i očekivajući posljedice. ART neuronska mreža postiže stabilnost učenjem očekivanja o 'svijetu' koja se konstantno uspoređuju s podacima iz 'svijeta'. Te namjere vode ka pozornosti (eng. attention) , tj. očekivanja se počinju fokusirati na podatke dostojna učenja.*» [Grossberg1995]

ART predstavlja kompleksan pristup aproksimaciji ljudskog spoznajnog (eng. *cognitive*) procesiranja informacija [Carpenter2002]. ART se prema brojnim autorima opisuje kao jedna od najkompleksnijih neuronskih mreža [Sapojnikova2003]. Zbog kompleksne pozadine ART-a, principi iz eksperimentalnog istraživanja govora, percepcije slike, korteksnog razvoja, mnogima ART model postaje teško razumljiv.

Do danas, teorija je evoluirala u cijeli niz *real-time* UNM modela koja obavljaju nadgledano ili nenadgledano učenje, raspoznavanje uzoraka i predviđanje.

5.2. Pozadina ART modela

Glavna osobina svih ART varijanti je proces uspoređivanja uzoraka (eng. *pattern matching process*), koji uspoređuje vanjski ulaz s internom memorijom jednog aktivnog kôda. Proces uspoređivanja vodi ka rezonantnom stanju, koje se održava dovoljno dugo da se omogući učenje, ili k paralelnom pretraživanju memorije. Ako pretraživanje završi na već uspostavljenom kôdu, memorijska prezentacija može ostati ista ili može unijeti nove informacije iz podudarenih dijelova trenutnog ulaza u mrežu. Ako pretraživanje završi na novom kôdu, memorijska prezentacija uči trenutni ulaz.

Navedeni proces je temelj stabilnosti kôda ART modela. Učenje temeljeno na uspoređivanju (eng. *match-based learning*) dopušta da se memorija mijenja samo kad je ulaz iz vanjskog svijeta dovoljno sličan internim očekivanjima, ili kad se dogodi nešto sasvim novo.

Ovakva vrsta učenja je komplementarna učenju temeljenom na greškama (eng. *error-based learning*[16]). Poznata umjetna neuronska mreža koja implementira učenje temeljeno na greškama je višeslojni perceptron s *backpropagation* algoritmom.

[16] *Error-based* učenje se odvija kad se dogodi krivo predviđanje, tada se mijenjaju memorijske prezentacije kako bi smanjile razlike između ciljnog izlaza i trenutnog ulaza.

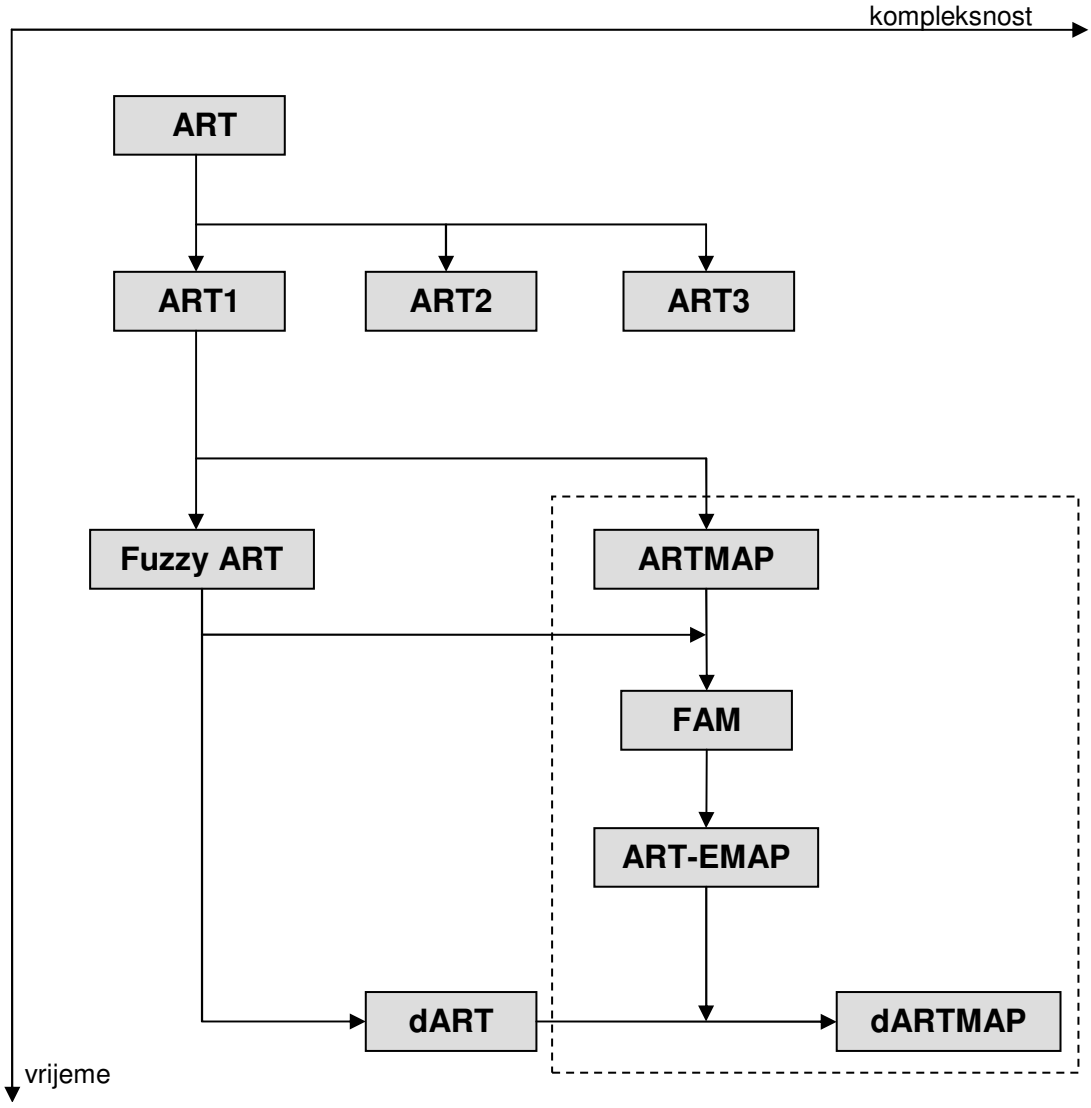
5.3. Taksonomija ART arhitektura

Još od davne 1976. godine, *ART* arhitekture su se razvijale neočekivanom brzinom, tako da danas postoji cijeli niz podvrsta *ART* arhitekture kojima je zajedničko svojstvo samoorganizacije s mogućnošću stabilnog, brzog ili sporog, učenja s obzirom na proizvoljne sekvence ulaznih uzoraka. Na slici (5) prikazana je evolucija i veze između pojedinih arhitektura, počevši od općenite *ART* arhitekture.

Prva kompletna *ART* mreža je bila *ART1* mreža, koristila je nenadgledano učenje binarnih uzoraka. Istovremeno su se razvijali modeli koji mogu učiti binarne i analogne uzorke (*ART2*), a *ART3* je kompleksna arhitektura koja se najviše približila modeliranju arhitekture i dinamike biološkog neurona. Sljedeći korak je bila arhitektura koja je implementirala nadgledano učenje spajanjem dvaju *ART1* modula (*ARTMAP*).

Nadogradnjom *ART1* arhitekture s neizrazitom (eng. *fuzzy*) logikom nastao je *Fuzzy ART*, može učiti binarne i analogne uzorke. Cijela skupina *ART* arhitektura koje se temelje na neizrazitoj logici se naziva *ART Fuzzy Networks* (ARTFNs). Na isti način je nastala i *Fuzzy ARTMAP* arhitektura, nadogradnjom *ARTMAP*-a s neizrazitom logikom. U ovom diplomskog radu, koristiti će se *Fuzzy ARTMAP* arhitektura za potrebe automatske klasifikacije i polu-automatskog indeksiranja teksta.

Uz navedene arhitekture postoji još cijeli niz arhitektura, od kojih je većina prikazana na slici (5), detalje o njima možete saznati u člancima: *dARTMAP* [Carpenter2003], *ART-EMAP* [Carpenter1995], *ARTMAP* [Carpenter1991], *μARTMAP* [Gomez-Sanchez2002].



Slika 5. Evolucija ART arhitektura [Sapojnikova2003]

5.3.1. Neizrazita logika

Kako bi mogli detaljnije pojasniti *Fuzzy ARTMAP* algoritam, potrebno se prije osvrnuti na teoriju koja leži iza neizrazite logike.

Neizrazita logika je ekstenzija Booleove logike koja uvodi koncept «djelomične istine». Dok klasična logika sve prikazuje u binarnom obliku (1,0; istina,laž), neizrazita logika uvodi stupnjeve istine. Dakle, neizrazita logika dopušta vrijednosti pripadnosti u intervalu [0, 1], vezana je uz neizrazite skupove i teoriju vjerojatnosti. Uveo ju je 1965. prof. Lofti Zadeh, pri kalifornijskom sveučilištu – Berkeley. U teoriji neizrazitih skupova (eng. *fuzzy set theory*) moguće je da jedan objekt

djelomično pripada nekom skupu. Kroz sljedećih nekoliko definicija preuzetih od autora [Zadeh1965] bit će opisana teorija neizrazitih skupova i operacija nad njima, od kojih se neke koriste u *Fuzzy ARTMAP* algoritmu.

Definicija 1. *Neka je U univerzalni skup, i neka je u generički element iz U. Onda se skup U predstavlja kao U = {u}.*

Definicija 2. *Neizraziti skup A je podskup univerzalnog skupa U i definira se kao uređeni par A = {u, $\mu_A(u)$} gdje je u ϵ U i 0 ≤ $\mu_A(u)$ ≤ 1. Funkcija pripadnosti $\mu_A(u)$ opisuje stupanj kojim element u pripada skupu A, gdje $\mu_A(u) = 0$ označava nikakvu pripadnost, a $\mu_A(u) = 1$ potpunu pripadnost.*

Definicija 3. *Potpora neizrazitog skupa A je podskup skupa U, svaki element koji stupanj pripadnosti skupu A različit od nule.*

Operacija nad neizrazitim skupovima su samo ekstenzija klasičnih operacija, jer su klasični skupovi specijalan slučaj neizrazitih skupova. Najčešće operacije za dva neizrazita skupa *A* i *B* definirana u istom univerzalnom skupu *U* su:

Unija: $\mu_{A \vee B}(u) = max(\mu_A(u), \mu_B(u)), \forall u \in U$ (10)

Presjek: $\mu_{A \wedge B}(u) = min(\mu_A(u), \mu_B(u)), \forall u \in U$ (11)

Komplement: $\mu_A^c(u) = 1 - \mu_A(u)$ (12)

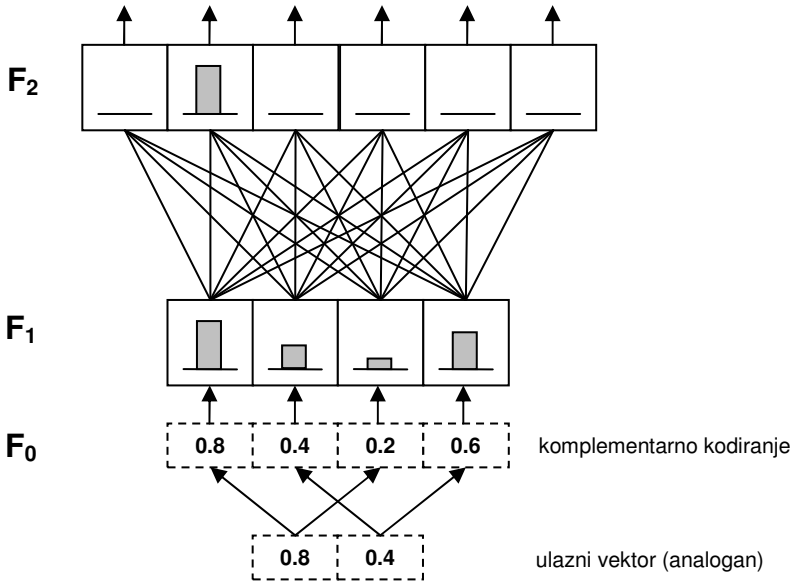
Ove operacije su asocijativne[17] i komutativne[18].

[17] Asocijativnost – $A \cup (B \cup C) = B \cup (A \cup C)$; $(A \cap B) \cap C = B \cap (A \cap C)$

[18] Komutativnost – $A \cap B = B \cap A$; $A \cup B = B \cup A$

5.3.2. Fuzzy ART

Fuzzy ART je modificirana verzija binarnog *ART1* modela, koja može učiti analogne ulazne uzorke, tj. vektore čije su komponente realni brojevi s vrijednostima između 0 i 1. To je UNM s nenadgledanim i inkrementalnim (*on-line*) učenjem. *Fuzzy ART* mreža se sastoji od 2 sloja neurona, ulaznog sloja F_1 i izlaznog sloja F_2 te od trećeg sloja F_0 gdje se obavlja komplementarno kodiranje ulaznog vektora, što će kasnije biti objašnjeno, kao što je to prikazano na slici (6).



Slika 6. Shema Fuzzy ART mreže [Busque1997]

Svi slojevi imaju vektore aktivnosti, na slici (6) prikazani kao vertikalni (zasjenjeni) stupci. Slojevi su potpuno međusobno povezani, a svaka veza ima težinu s vrijednostima iz intervala [0, 1]. Neuron u F_2 sloju predstavlja jednu kategoriju koju je stvorila mreža, a definirana je preko težinskog vektora *wj* (gdje je *j* indeks neurona). Veličina tog vektora je jednaka dimenzionalnosti sloja F_1, a inicijalno su sve vrijednosti težinskih vektora postavljene na 1. Dok se težine neurona ne promijene kažemo da je taj neuron neopredijeljen (eng. *uncommited*), nakon promjene težina za neuron kažemo da je opredijeljen (eng. *commited*).

Mreža koristi oblik normalizacije ulaznog vektora koja se naziva komplementarno kodiranje. To je operacija spajanja ulaznog vektora s

komplementom tog vektora (konkatenacija). Na ovaj način se vektor normalizira, ali uz sačuvanje informacija o amplitudama vektora.

5.3.2.1. Algoritam

Sljedeći algoritam je preuzet od autora [Carpenter1991].

Vektor aktivnosti F_0 sloja se definira kao $\boldsymbol{I} = (I_1, \ldots, I_M)$, gdje je svaka komponenta I_i u intervalu [0, 1]. Vektor aktivnosti sloja F_1 se definira kao $\boldsymbol{x} = (x_1, \ldots, x_M)$, a vektor aktivnost F_2 sloja je $\boldsymbol{y} = (y_1, \ldots, y_N)$. Broj čvorova u svakom polju (sloju) je proizvoljan.

Svaki F_2 neuron (čvor) j $(j = 1, \ldots, N)$ ima vektor $\boldsymbol{w_j} = (w_{j1}, \ldots, w_{jM})$ adaptivnih težina, ili LTM tragova. Inicijalno se postavlja:

$$w_{j1}(0) = \ldots = w_{jM}(0) = 1 \tag{13}$$

Fuzzy ART ima nekoliko parametara koji određuju dinamiku: **parametar odabira** $\alpha > 0$ (eng. *choice parameter*); **stopa učenja** $\beta \in [0, 1]$ (eng. *learning rate parameter*); **parametar budnosti** $\rho \in [0, 1]$ (eng. *vigilance parameter*).

Za svaki ulazni vektor $\boldsymbol{I}$ i F_2 čvor j računa se funkcija odabira T_j

$$T_j(I) = \frac{|\, I \wedge w_j \,|}{\alpha + |\, w_j \,|} \tag{14}$$

gdje se operator $\wedge$ definira prema formuli (11), a norma $|\cdot|$ je L_1 norma[19]. Za sustav kažemo da je donio odluku o kategoriji kad barem jedan F_2 čvor može postati aktivan u danom trenutku, odluka o kategoriji, neuronu pobjedniku, se donosi na sljedeći način:

$$T_J = \max(T_j : j = 1 \ldots N) \tag{15}$$

[19] L_1 norma se izračunava kao suma elemenata vektora, $|\, a \,| = \sum_i |\, a_i \,|$

Ako je istovremeno više neurona ima maksimalnu vrijednost T_j funkcije onda se odabire neuron s najmanjim indeksom. Kad se odabere *J*-ta kategorija onda vrijedi $y_J = 1$; *i* $y_j = 0$ za $j \neq J$. Vektor aktivnosti F_1 sloja zadovoljava jednadžbu:

$$x = \begin{cases} I; & \text{ako je } \mathrm{F}_2 \text{ neaktivan} \\ I \wedge w_J; & \text{ako je J - ti } \mathrm{F}_2 \text{ čvor odabran} \end{cases} \tag{16}$$

Rezonancija se događa ako funkcija preklapanja (eng. *match function*) zadovoljava sljedeći kriterij:

$$\frac{|\, I \wedge w_J \,|}{|\, I \,|} \geq \rho \tag{17}$$

Reset se događa kad se ne zadovolji navedeni kriterij, što znači da odabrani F_2 čvor nije dovoljno sličan ulaznom vektoru pa je potrebno pronaći sljedeći najbolji čvor po formuli (15) s tim da je potrebno zaustaviti odabiranje prethodno odabranog čvora njegovom inhibicijom, postavljanjem $T_J = 0$. Ovaj postupak se odvija sve dok se ne zadovolji uvjet (17).

Kad nastupi rezonancija može nastupiti proces učenja ulaznog uzorka, mreža uči ulazni uzorak promjenom vektora težine odabranog F_2 čvora po formuli:

$$w_J^{(novi)} = \beta(I \wedge w_J^{(stari)}) + (1-\beta) w_J^{(stari)} \tag{18}$$

Brzo učenje se može postići postavljanjem parametra $\beta = 1$. Praksa je da se koristi ***fast-commit slow-recode*** princip, što znači da se parametar β postavlja na vrijednost *1* kad je odabrani čvor neopredijeljen, a kad je odabrani čvor opredijeljen onda je $\beta < 1$. Ovaj princip je koristan jer mreža može brzo naučiti rijetke slučajeve direktnim kopiranjem ulaznog vektora u težinski vektor odabranog čvora, a nakon toga može koristiti sporo učenje koje je praktičnije za generalizaciju.

5.3.2.2. Geometrijska interpretacija algoritma

U ovom dijelu bit će opisano funkcioniranje *ART* algoritma, s komplementarnim kodiranjem, preko geometrijskog prikaza.

Redefiniranjem ulaznog vektora $\boldsymbol{I} = (\boldsymbol{a}, \boldsymbol{a}^C)$, koji se sastoji od dva dvodimenzionalna vektora $\boldsymbol{a} = (a_1, a_2)$ i njegova komplementa $\boldsymbol{a}^C$, u četverodimenzionalni vektor dobiva se sljedeći izraz za vektor $\boldsymbol{I}$:

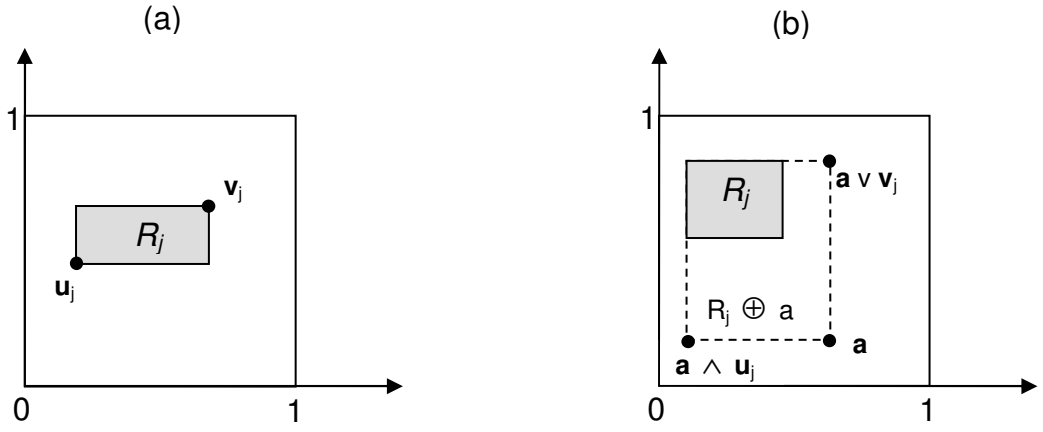
$$\boldsymbol{I} = (\boldsymbol{a}, \boldsymbol{a}^C) = (a_1, a_2, 1 - a_1, 1 - a_2) \tag{19}$$

U tom slučaju, svaka kategorija j ima geometrijsku interpretaciju kao pravokutnik R_j. Na isti način se može vektor težine $\boldsymbol{w}_j$ prikazati u obliku komplementarnog kodiranja, slika (7a):

$$\mathbf{w}_j = (\mathbf{u}_j, \mathbf{v}_j^C) \tag{20}$$

Gdje su $\boldsymbol{u}_j$ i $\boldsymbol{v}_j$ dvodimenzionalni vektori, neka $\boldsymbol{u}_j$ definira jedan ugao pravokutnika R_j, a vektor $\boldsymbol{v}_j$ drugi ugao. Veličina pravokutnika R_j se definira kao:

$$|R_j| \equiv |\boldsymbol{v}_j - \boldsymbol{u}_j| \tag{21}$$



Slika 7. (a) Geometrijska interpretacija Fuzzy ART težinskog vektora (M=2). (b) Geometrijska interpretacija brzog učenja [Carpenter1992]

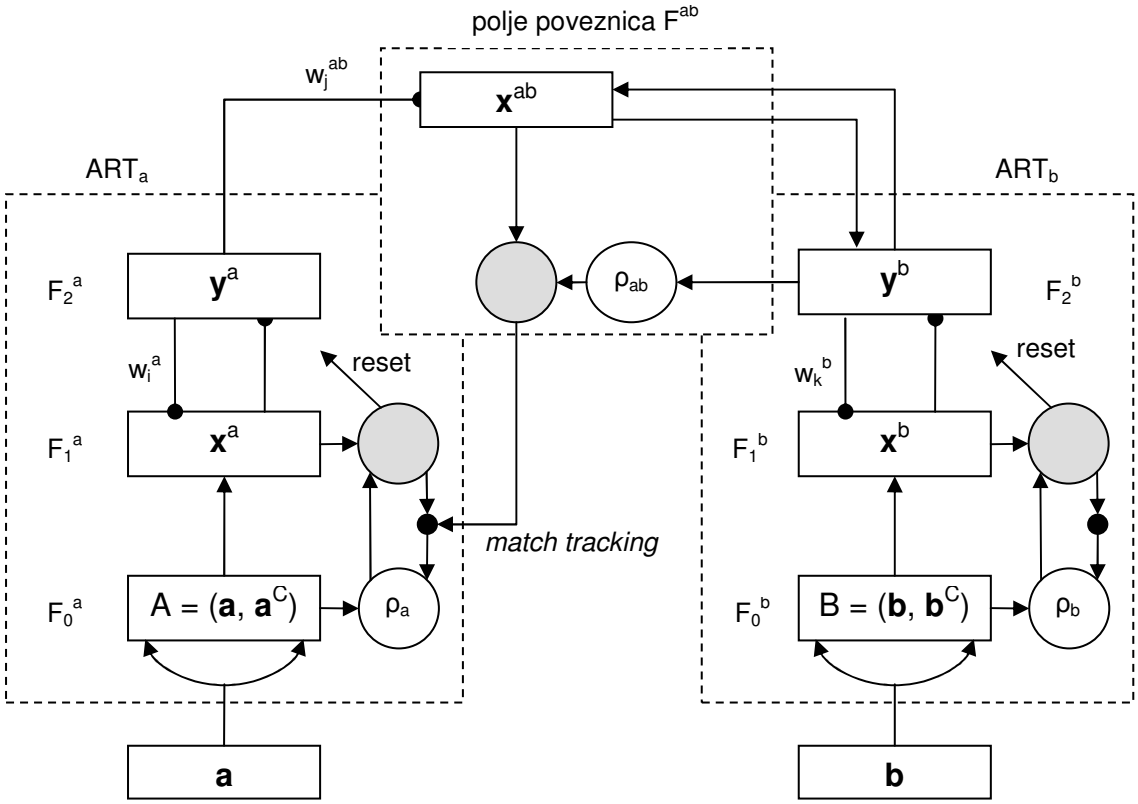
U sustavu s brzim učenjem, $\beta = 1$, vrijedi da je $\mathbf{w}_J^{(novi)} = \mathbf{I} = (\mathbf{a}, \mathbf{a}^C)$ kad je J neopredijeljeni čvor. Tada su uglovi novog pravokutnik $R_J^{(novi)}$ određeni s vektorima $\boldsymbol{a}$ i $(\boldsymbol{a}^C)^C = \boldsymbol{a}$, tj. $R_J^{(novi)}$ je točka $\boldsymbol{a}$. Zapravo, veličina pravokutnika R_j raste kako se smanjue veličina $\boldsymbol{w}_j$ tijekom učenja, a maksimalna veličina pravokutnika je određena s parametrom budnosti. Tijekom brzog učenja, nakon što čvor postane opredijeljen, pravokutnik R_J se proširuje na na $R_J \oplus \boldsymbol{a}$, minimalni pravokutnik koji sadrži R_J i $\boldsymbol{a}$, slika (7b). Općenito vrijedi, kod brzog učenja, svaki R_j je jednak

najmanjem pravokutniku koji sadrži sve vektore **a** koji «odabiru» kategoriju *j*, uz ograničenje $|R_j| \leq 2(1 - \rho)$.

Visoke vrijednosti parametra budnosti ($\rho \cong 1$) vode k manjim pravokutnicima R_j, dok niske vrijednosti ($\rho \cong 0$) vode k većim pravokutnicima [Grossberg1992].

5.3.3. Fuzzy ARTMAP

Fuzzy ARTMAP algoritam, uveden 1992. godine [Carpenter1992], je UNM s nadgledanim učenjem koja se sastoji od dvije samoorganizirajuće mreže, *Fuzzy ART$_a$* i *Fuzzy ART$_b$* koji su međusobno povezani asocijativnom memorijom koja se naziva polje poveznica (eng. *map field*) i označava se F^{ab}. Polje poveznica formira asocijacije između kategorija i predviđanja te realizira *match tracking* pravilo. To pravilo povećava parametar budnosti kod ART$_a$ modula ako dođe do krivo sparenog predviđanja *ART$_a$* i *ART$_b$* modula (eng. *predictive mismatch*). *Match tracking* reorganizira strukturu kategorija kako se ne bi ponovilo krivo predviđanje daljnjim prezentacijima istog ulaznog uzorka. Na slici (8) se nalazi shema *Fuzzy ARTMAP* mreže.



Slika 8. Fuzzy ARTMAP arhitektura [Carpenter1992]

5.3.3.1. Algoritam

Kroz nekoliko koraka bit će prikazan algoritam *Fuzzy ARTMAP* algoritma. Napomena, parametri i varijable pojedinog modula, ART_a i ART_b, se razlikuju po dodatnim oznakama a ili b, respektivno.

1) Inicijalizacija

U F_0^a sloju, M_a-dimenzionalni ulazni vektor $\boldsymbol{I}^a$, s komponentama $I_i^a \in [0, 1]$ se komplementarno kodira u vektor $\boldsymbol{A} = (\boldsymbol{a}, \boldsymbol{a}^C)$. Težinski vektori se početno postave na vrijednosti $w_{j1}^a(0) = \ldots = w_{j2Ma}^a(0) = 1$ za sve čvorove $j = 1, \ldots, N_a$. Analogno se postavljaju i vrijednosti ART_b modula: $\boldsymbol{I}^b = \boldsymbol{B} = (\boldsymbol{b}, \boldsymbol{b}^C)$, $w_{j1}^b(0) = \ldots = w_{j2Mb}^b(0) = 1$ za sve čvorove $j = 1, \ldots, N_b$. Težine polja poveznica se također postavljaju na vrijednosti $w_{jk}^{ab}(0) = 1$. Parametar budnosti se postavlja na početnu vrijednost (eng. *baseline vigilance*).

2) Traženje pobjednika

U slojevima F_2^a i F_2^b traži se neuron – pobjednik prema funkciji odabira definiranoj u formuli (14):

$$T_j^a(A) = \frac{|A \wedge w_j^a|}{\alpha^a + |w_j^a|} \qquad T_j^b(B) = \frac{|B \wedge w_j^b|}{\alpha^b + |w_j^b|} \qquad (22) \quad (23)$$

Funkcija odabira se računa za svaki neuron $(j = 1,\ldots, N^a; j = 1, \ldots, N^b)$ F_2 sloja ART_a i ART_b modula, odabiru se neuroni koji imaju najveću vrijednost T_j funkcije u pojedinom modulu i njima se pridjeljuju indeksi J^a i J^b, respektivno.

3) Uspoređivanje

Odabir neurona J^a i J^b se mora potvrditi provjerom kriterija preklapanja (eng. *match criterion*), tj. doza sličnosti između neurona pobjednika i ulaznog vektora mora zadovoljiti nekakav uvjet, parametar budnosti definira taj uvjet:

$$C_{J^a}^a = \frac{|A \wedge w_{J^a}^a|}{|A|} \geq \rho^a \qquad C_{J^b}^b = \frac{|B \wedge w_{J^b}^b|}{|B|} \geq \rho^b \qquad (24) \quad (25)$$

U slučaju da se ne zadovolji ovaj kriterij onda se odabire sljedeći najbolji neuron kao pobjednik i opet se provjerava kriterij preklapanja s ulaznim vektorom.

4) Učenje

Nakon što se pronašao neuron – pobjednik, koji zadovoljava kriterij preklapanja, onda može nastupiti proces učenja modifikacijom težinskog vektora pobjedničkog neurona u F_2 sloju kod ART_a i ART_b modula po formuli (18).

5) Aktivacija polja poveznica

Polje poveznica F^{ab} se aktivira kad je jedna od ART_a ili ART_b kategorija aktivna. Ako je odabran čvor s indeksom J^a $F_2{}^a$ polja onda pripadni težinski vektor $\boldsymbol{w}_J{}^{ab}$ aktivira F^{ab}, a ako je odabran čvor s indeksom J^b $F_2{}^b$ polja onda se aktiviraju 1-na-1 putevi između $F_2{}^b$ i F^{ab}. Ako su ART_a i ART_b aktivni onda je F^{ab} aktivan samo onda ako ART_a predviđa istu kategoriju kao ART_b. Stoga se aktivacijski vektor F^{ab} polja, $\boldsymbol{x}^{ab}$ definira ovako:

$$x^{ab} = \begin{cases} y^b \wedge w_J^{ab} & \text{ako je čvor J } \mathrm{F}_2^{\mathrm{a}} \text{ polja aktivan i } \mathrm{F}_2^{\mathrm{b}} \text{ je aktivan} \\ w_J^{ab} & \text{ako je čvor J } \mathrm{F}_2^{\mathrm{a}} \text{ polja aktivan i } \mathrm{F}_2^{\mathrm{b}} \text{ je neaktivan} \\ y^b & \text{ako je } \mathrm{F}_2^{\mathrm{a}} \text{ neaktivan i } \mathrm{F}_2^{\mathrm{b}} \text{ je aktivan} \\ 0 & \text{ako je } \mathrm{F}_2^{\mathrm{a}} \text{ neaktivan i } \mathrm{F}_2^{\mathrm{b}} \text{ je neaktivan} \end{cases} \qquad (26)$$

Po (26) $\boldsymbol{x}^{ab} = 0$ ako se predviđanje $\boldsymbol{w}_J{}^{ab}$ ne potvrđuje kod aktivacijskog vektora $\boldsymbol{y}^b$. Takav događaj uzrokuje ponovno odabiranje neurona – pobjednika u ART_a modulu, taj proces se naziva *match tracking*.

Prilikom prvog predstavljanja pojedinog ulaznog vektora, parametar budnosti se postavlja na osnovnu vrijednost (eng. *baseline vigilance*). Parametar budnosti polja poveznica se označava ρ_{ab} i ako vrijedi:

$$| x^{ab} | < \rho_{ab} | y^b | \qquad (27)$$

tada se ρ_a povećava dok ne bude za nijansu veći od iznosa $|\boldsymbol{A} \wedge \boldsymbol{w}_J{}^a|/|\boldsymbol{A}|^{-1}$. Kad se to dogodi onda će ART_a modul odabrati novi neuron iz $F_2{}^a$ sloja, povećanje parametra budnosti uzrokuje inhibiciju prethodno odabranog neurona preko kriterija preklapanja.

Učenje polja poveznica određuje kako se težine polja $w_{jk}{}^{ab}$ mijenjaju kroz vrijeme, inicijalno su te težine postavljene na vrijednost 1. Tijekom rezonancije s aktivnim neuronom J ART_a modula, $w_J{}^{ab}$ se približava aktivacijskom vektoru x^{ab}. S brzim učenjem, kad neuron J^a nauči predviđati ART_b kategoriju J^b, onda je ta asocijacija stalna, tj. $w_J{}^a{}_J{}^{b\,(ab)} = 1$.

5.3.3.2. Pojednostavljeni Fuzzy ARTMAP

Za potrebe automatske klasifikacije teksta, *FAM* algoritam se može pojednostaviti kompletnim izbacivanjem ART_b modula. Svrha ART_b modula je ostvarivanje nadgledanog učenja, njemu se prezentiraju vektori koje ART_b, zbog svojstva samoorganizacije, grupira u kategorije u F_2^b sloju, a te kategorije ART_a modul mora naučiti predviđati.

Budući da se kod problema klasifikacije unaprijed znaju kategorije koje ART_a modul treba naučiti predviđati onda se ART_b modul izbacuje, a vektori koji su se trebali prezentirati ART_b modulu se prezentiraju polju poveznica kao aktivacijski vektori $\boldsymbol{Y}^b$. Dakle, cijeli algoritam je gotovo identičan *FAM* algoritmu, osim što se više ne računaju funkcije ART_b modula. U tablici (5) su prikazane moguće situacije nakon odabiranja neurona u fazi učenja i testiranja mreže.

Tablica 5. Pojednostavljeni *FAM* algoritam (učenje i testiranje)

Odabrani neuron	Faza Učenja	Faza testiranja
J je opredijeljeni čvor	Predviđanje je točno; prilagodi $\boldsymbol{w}_J^a$ i $\boldsymbol{w}_J^{ab}$	Izlaz: ulazni vektor $\boldsymbol{I}^a$ pripada kategoriji koja se može očitati iz polja poveznica $\boldsymbol{w}_J^{ab}$
	Predviđanje je netočno; pokreni *match tracking*	
J je neopredijeljeni čvor	Kopiranje ulaznog vektora $\boldsymbol{I}^a$ u $\boldsymbol{w}_J^a$ (brzo učenje) i kopiranje kategorije $\boldsymbol{I}^b$ u $\boldsymbol{w}_J^{ab}$	Izlaz: ulazni vektor $\mathbf{I}^a$ pripada nepoznatoj kategoriji

Za potrebe polu-automatskog indeksiranja također se može iskoristiti pojednostavljeni *FAM*, ali u ovom slučaju aktivacijski vektor y^b može sadržavati nekoliko odabranih kategorija, jer se kod problema indeksiranja dokument može predstaviti s nekoliko indeksa, odnosno kategorija. Na ovaj način ART_a modul će istovremeno moći pridijeliti odabranom neuronu F_2 sloja nekoliko kategorija, što je zapravo cilj polu-automatskog indeksiranja. Kako bi smanjili pogrešku predviđanja moguće je u fazi testiranja za pojedini ulazni vektor pronaći nekoliko najboljih neurona u F_2^a sloju i onda napraviti *fuzzy* presjek pripadnih kategorija kako bi dobili što točniji skup kategorija.

5.3.3.3. Primjer rada pojednostavljenog FAM algoritma

Kako bi pokazali proces učenja pojednostavljenog *FAM* algoritma koristimo jednostavan primjer klasifikacije dvodimenzionalnih točaka. U tablici (6) nalazi skup točaka za testiranje zajedno s pripadnostima. Parametri mreže su sljedeći: $M = 2$, $\rho = 0.85$, $\alpha = 0.01$, $\varepsilon = 0.01$ (vrijednost za koju se povećava parametar budnosti u *match tracking* procesu), broj neurona u F_2^a sloju je 4.

Tablica 6. Skup podataka za učenje za pojednostavljeni FAM

Točka	**A_1**	**A_2**	**A_3**	**A_4**
a = (x, y)	(0.7, 0.7)	(0.2, 0.6)	(0.5, 0.1)	(0.8, 0.1)
$\mathbf{I^a} = (\mathbf{a}, \mathbf{a}^C)$	(0.7, 0.7, **0.3**, **0.3**)	(0.2, 0.6, **0.8**, **0.4**)	(0.5, 0.1, **0.5**, **0.9**)	(0.8, 0.1, **0.2**, **0.9**)
Kategorija	1	2	1	1
$\mathbf{I}^b$	(0, 1)	(1, 0)	(0, 1)	(1, 0)

1) Točka A_1:

Računa se funkcija odabira za svaki neuron u F_2^a sloju po formuli (22). Početno su svi neuroni neopredijeljeni pa su njihovi vektori težina postavljeni na jedinične vrijednosti, zbog toga su funkcije odabira svih neurona jednake, odabire se neuron s najmanjim indeksom. Funkcija odabira za prvi neuron iznosi:

$$T_1 = \frac{0.7 \wedge 1.0 + 0.7 \wedge 1.0 + 0.3 \wedge 1.0 + 0.3 \wedge 1.0}{0.01 + 4} \approx 0.49$$

Pobjednik $J = 1$, kriterij preklapanja C_J, formula (24):

$$C_{J=1} = \frac{0.7 \wedge 1.0 + 0.7 \wedge 1.0 + 0.3 \wedge 1.0 + 0.3 \wedge 1.0}{0.7 + 0.7 + 0.3 + 0.3} = 1 > 0.85$$

$\boldsymbol{X}^{ab} = \boldsymbol{I}^b = (0, 1)$, nema *match tracking* procesa.

Nakon učenja $\boldsymbol{w}_1^a = (0.7, 0.7, 0.3, 0.3)$, $\boldsymbol{w}_1^{ab} = (0, 1)$

2) Točka A_2:

Opredijeljeni neuroni: 1 →

$$T_1 = \frac{0.2 \wedge 0.7 + 0.6 \wedge 0.7 + 0.8 \wedge 0.3 + 0.4 \wedge 0.3}{0.01 + (0.7 + 0.7 + 0.3 + 0.3)} \approx 0.69$$

Neopredijeljeni neuroni: 2, 3, 4 →

$$T_2 = \frac{0.2 \wedge 1.0 + 0.6 \wedge 1.0 + 0.8 \wedge 1.0 + 0.4 \wedge 1.0}{0.01 + (1.0 + 1.0 + 1.0 + 1.0)} \approx 0.49$$

Pobjednik $J = 1$, kriterij preklapanja C_J:

$$C_1 = \frac{0.2 \wedge 0.7 + 0.6 \wedge 0.7 + 0.8 \wedge 0.3 + 0.4 \wedge 0.3}{2} = 0.7 < 0.85$$

Pobjednik se inhibira, postupak ponovnog odabiranja pobjednika. Budući da su ostali neuroni neopredijeljeni onda je pobjednik onaj s najmanjim indeksom, $J = 2$. Funkcija odabira $T_2 = 0.49$, a kriterij preklapanja $C_2 = 1 > 0.85$.

Nakon učenja $\boldsymbol{w}_2^a = (0.2, 0.6, 0.8, 0.4)$, $\boldsymbol{w}_2^{ab} = (1, 0)$.

3) Točka A_3:

Opredijeljeni neuroni: 1, 2 → $T_1 = 0.59$, $T_2 = 0.59$.

Neopredijeljeni neuroni: 3, 4 → $T_3 = 0.49$, $T_4 = 0.49$.

Pobjednik $J = 1$, kriterij preklapanja $C_1 = 0.6 < 0.85$. Pobjednik se inhibira i obavlja se novi postupak odabiranja neurona. Ponovo će se odabrati neuron 2, ali i on neće zadovoljiti kriterij preklapanja pa će se odabrati neopredijeljeni neuron 3, na kojem će se obaviti učenje. $T_3 = 0.49$, $C_3 = 1 > 0.85$, nakon učenja $\boldsymbol{w}_3^a = (0.5, 0.1, 0.5, 0.9)$, $\boldsymbol{w}_3^{ab} = (0,1)$

4) Točka A_4:

Opredijeljeni neuroni: 1, 2 ,3 → $T_1 = 0.65$, $T_2 = 0.44$, $T_3 = 0.84$

Neopredijeljeni neuroni: 4 → $T_4 = 0.49$

Pobjednik $J = 3$, zadovoljen kriterij preklapanja $C_3 = 0.85 \geq 0.85$, ali je predviđanje neurona 3 drugačije od kategorije točke A_4 pa se obavlja *match tracking* postupak kojim se parametar budnosti postavlja na iznos $C_3 + \varepsilon = 0.86$ i ponavlja se postupak odabiranja pobjednika. Pobjednik je $J = 3$, ali ne zadovoljava kriterij preklapanja $C_3 = 0.85 < 0.86$ pa se taj neuron inhibira i ponavlja se postupak. Ostali opredijeljeni neuroni također neće zadovoljiti kriterij preklapanja pa će se na kraju odabrati neopredijeljeni neuron $J = 4$: $T_4 = 0.49$, $C_4 = 1 > 0.86$. Nakon učenja $\boldsymbol{w}_4^a = (0.8, 0.1, 0.2, 0.9)$, $\boldsymbol{w}_4^{ab} = (1, 0)$.

6. Ulazni podaci

U ovom poglavlju bit će riječi o skupovima dokumenata, nad kojima će se obavljati eksperimenti klasifikacije i polu-automatskog indeksiranja, te o nekim konkretnim informacijama o programskoj implementaciji predprocesiranja podataka.

6.1. Skupovi dokumenata

Kroz nekoliko podpoglavlja navedene su sve informacije vezane uz korištene skupove dokumenata.

6.1.1. Skup dokumenata «Akademski primjer»

Ovaj skup dokumenata je preuzet od autora [Dobša2004], kao prijedlog testiranja funkcionalnosti algoritma za klasifikaciju teksta. Dokumenti su zapravo naslovi literature iz područja dubinske analize podataka i teksta (eng. *data/text mining*) i linearne algebre. Cijeli skup za učenje se sastoji od 14 naslova, od toga 9 iz područja *data mining-a* i 5 iz područja linearne algebre.

Mreži će se predstaviti cijeli skup podataka, a onda će se mreža testirati nad podacima koji će biti navedeni kasnije.

Podaci za učenje su sljedeći:

D1*. Survey of text mining: clustering, classification, and retrieval*

D2*. Automatic text processing: the transformation analysis and retrieval of information by computer*

D3. *Elementary linear algebra: A matrix approach*

D4. *Matrix algebra & its applications statistics and econometrics*

D5*. Effective databases for text & document management*

D6. *Matrix analysis and applied linear algebra*

D7. *Topological vector spaces and algebras*

D8. *Information retrieval: data structures & algorithms*

D9. *Vector spaces and algebras for chemistry and physics*

D10. *Classification, clustering and data analysis*

D11. ***Clustering of large data sets***

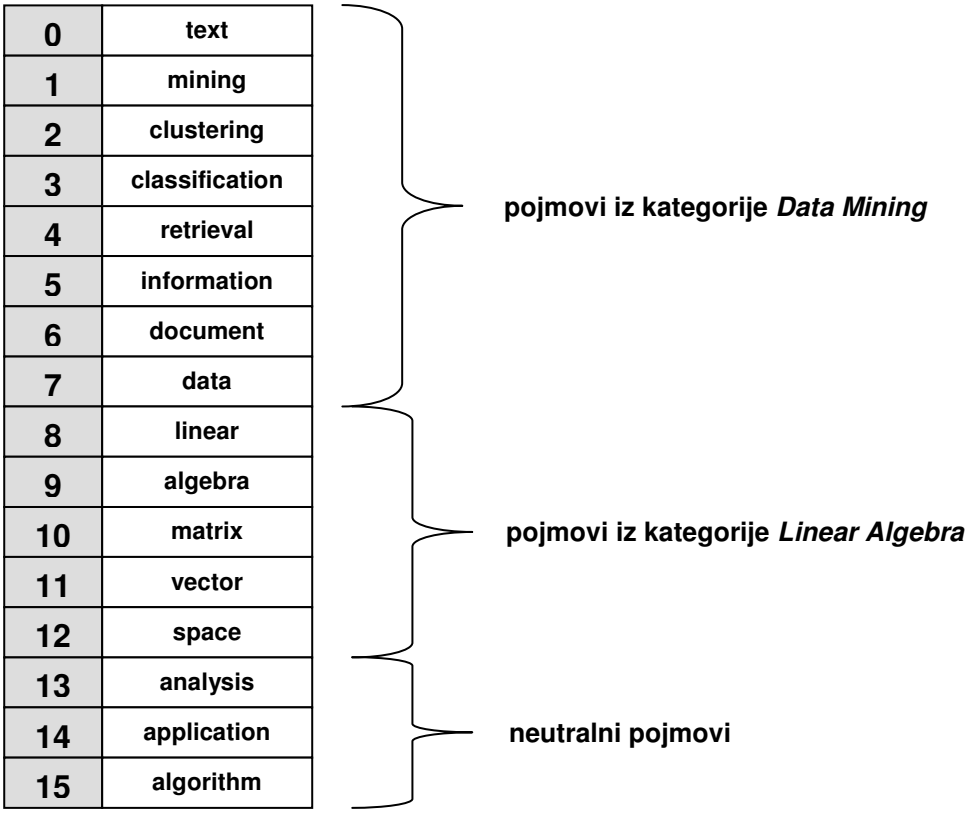
D12. ***Clustering algorithms***

D13. ***Document warehousing and text mining: techniques for improving business operations, marketing and sales***

D14. ***Data mining and knowledge discovery***

Prvoj kategoriji (*DATA MINING*) pripadaju sljedeći naslovi: D1, D2, D5, D8, D10, D11, D12, D13 i D14; dok drugoj kategoriji (*LINEAR ALGEBRA*) pripadaju naslovi: D3, D4, D6, D7 i D9.

U ovom slučaju predprocesiranje skupa podataka se neće obaviti automatski, zbog malog opsega problema, već će se to obaviti ručno koristeći standardne metode izbacivanja stop-riječi, morfološke normalizacije (eng. *stemming*). Dodatno se radi selekcija dobivenih riječi izbacivanjem riječi koje se pojavljuju samo u jednom naslovu. Nakon ovog postupka dobivaju se riječi koje tvore ulazni vektor veličine 16 elemenata, ulazni vektor je prikazan na slici (9).



Slika 9. Ulazni vektor za skup dokumenata "Akademski primjer"

Nakon predprocesiranja možemo svaki navedeni primjerak za učenje/testiranje pretvoriti u binarni vektor dimenzije 16, gdje **1** označava da je pojedina riječ na tom indeksu prisutna u naslovu, **0** inače. Ne koristi se *tfidf* metoda za izračunavanje ulaznih vektora jer je to nepotrebno za jednostavne primjere. Dobiveni vektori su prikazani u tablici (7).

Tablica 7. Skup za učenje «Akademskog primjera» u binarnom obliku

D1	**1111100000000000**
D2	**1000110000000100**
D3	**0000000011100000**
D4	**0000000001100010**
D5	**1000001000000000**
D6	**0000110000111000**
D7	**0000000001011000**
D8	**0000110100000001**
D9	**0000000001011000**
D10	**0011000100000100**
D11	**0010000100000000**
D12	**0010000000000001**
D13	**1100001000000000**
D14	**0100000100000000**

Skup podataka/naslova za testiranje se sastoji od 5 naslova, koji su jednako predtprocesirani kao i skup za učenje, ali riječi iz ovog skupa nisu uzete u obzir prilikom kreiranja ulaznog vektora različitih riječi jer bi to utjecalo na rezultate klasifikacije. Ti vektori su prikazani u tablici (8), *DM = Data Mining, LA = Linear Algebra*.

Tablica 8. Skup za testiranje «Akademskog primjera» u binarnom obliku

K1 (*Data Mining*)	***DM***	**0100000100000000**
K2 (*Matrix Algebra*)	***LA***	**0000000001100000**
K3 (*Application for Analysis of Clustering Space*)	***DM***	**0010000000001110**
K4 (*Algorithm for Vector Space Analysis*)	***LA***	**0000000000011101**
K5 (*Application for Linear Analysis*)	***LA***	**0000000010000010**

6.1.2. Skup dokumenata «*Croatia Weekly*»

Skup dokumenata «*Croatia Weekly*» je paralelna hrvatsko-engleska baza novinskog lista «*Croatia Weekly*», bazu je ustupio prof. dr. sc. Marko Tadić sa Zavoda za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu kao dio hrvatsko-engleskog paralelnog korpusa[20].

Baza je ustupljena za potrebe projekta prof. dr. Bojane Dalbelo-Bašić s Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu. Cilj projekta je obavljanje eksperimenata nad hrvatskim jezikom kako bi se utvrdile razlike u učinkovitostima klasifikatora nad engleskim i hrvatskim jezikom.

Dakle, izvor su novinski list «Croatia Weekly», brojevi 5 – 118, koji su izlazili od 1998. – 2000. Baza sadrži 4 kategorije:

po – politika

go – gospodarstvo

te – turizam i ekologija

ks – kultura i šport

Skup dokumenata se sastoji od **3582** članaka na hrvatskom i engleskom jeziku, od toga se **1792** članka koriste kao skup za učenje, a **1790** članaka kao skup za testiranje.

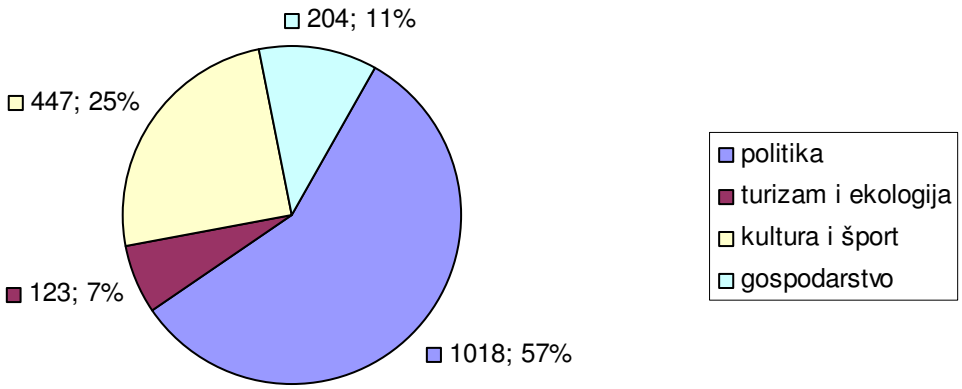
Predprocesiranje je pojednostavljeno, naime, izbačene su stop-riječi, ali nije obavljena morfološka normalizacija zbog preopsežnosti problema (normalizacija hrvatskog i engleskog jezika). Predprocesiranjem skupa za učenje engleskog korpusa dobiva se ulazni vektor različitih riječi dimenzije **31300**, dok se analognim postupkom za hrvatski korpus dobiva vektor dimenzije **72922**, što pokazuje koliko je hrvatski jezik morfološko bogatiji od engleskog jezika.

Nakon navedenog postupka koristila se normalizirana *tfidf* funkcija za izračunavanje elemenata ulaznih vektora skupa za učenje i skupa za testiranje.

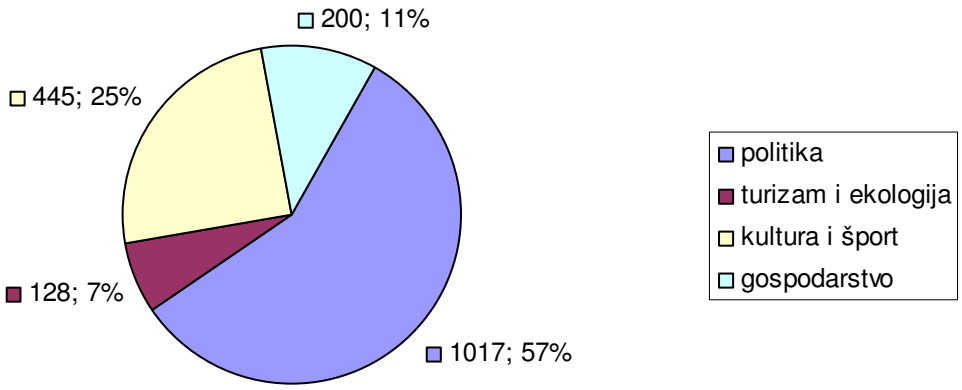
---

[20] Detaljnije informacije o hrvatsko-engleskom paralelnom korpusu se mogu pronaći na web stranici http://www.hnk.ffzg.hr/

Na slikama (10) i (11) prikazana je podjela članaka po kategorijama i skupovima za učenje i testiranje.



Slika 10. Podjela članaka po kategorijama (skup za učenje)



Slika 11. Podjela članaka po kategorijama (skup za testiranje)

Iz prikazanog se može vidjeti kako podjela po kategorijama nije ravnomjerna, vjerojatni uzrok bi mogao biti premalen broj kategorija koje pokrivaju širok spektar podkategorija, ali jedan od mogućih razloga je i specijaliziranost novinskog lista Croatia Weekly za političke članke. Navedene činjenice mogu dodatno otežati posao klasifikatoru.

6.1.3. Skup dokumenata «Vjesnik»

Skup dokumenata «Vjesnik» je baza od 90000 članaka novinskog lista «Vjesnik» koju je ustupio prof. dr. sc. Marko Tadić kao dio hrvatskog nacionalnog korpusa[21].

Od navedenog broja članaka preuzeto je samo 20000 članaka za potrebe učenja i testiranja klasifikatora, skupovi su međusobno ravnomjerno raspodijeljeni: 10000 članaka u skupu za učenje i 10000 članaka u skupu za testiranje. Cijeli skup dokumenata je podijeljen u 8 velikih kategorija, koje su po značenju imena dosta opširne jer pokrivaju širok spektar područja. Tih 8 kategorija su redom:

ck – crna kronika

ku – kultura

go – gospodarstvo

sp – šport

un – unutarnja politika

td – teme dana

zg – zagreb

vp – vanjska politika

Predprocesiranje članaka proveo je dipl. ing. Jan Šnajder s Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu. Predprocesiranjem su izbačene stop-riječi i obavljena je morfološka normalizacija[22]. Postupak morfološke normalizacije je jako složen pa trenutno postoji cijeli niz normaliziranih verzija skupa članaka – Vjesnik, u ovom diplomskom radu koristi se verzija 1.0, varijanta 1. U navedenoj verziji ulazni vektor različitih riječi ima dimenziju **81582**.

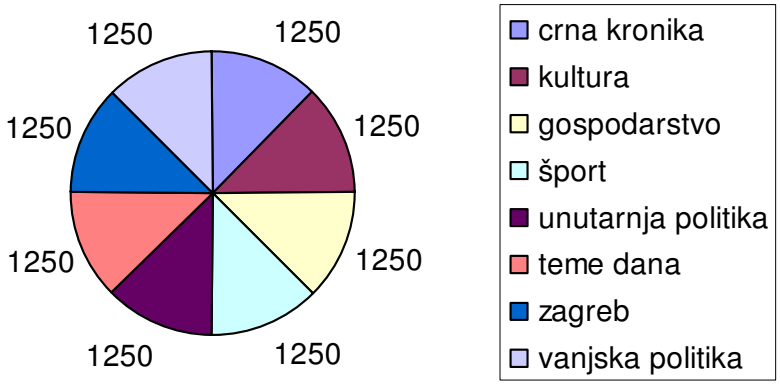
---

[21] Detaljnije informacije o hrvatskom nacionalnom korpusu možete pronaći na web stranici http://www.hnk.ffzg.hr/

[22] Detaljnije informacije o morfološkoj normalizacija skupa članaka «Vjesnik» možete pronaći na web stranici http://www.zemris.fer.hr/~jan/amn

Skupovi za učenje i testiranje su odvojeni u dvije različite datoteke. Članci unutar datoteka su zapisani na način da se u prvi red navode podaci o članku koji obuhvaćaju redni broj dokumenta, broj riječi u dokumentu i šifru kategorije kojoj dokument pripada. Zatim se u redovima ispod navode normalizirane *tfidf* vrijednosti za pojedinu riječ. U ovoj verziji predprocesiranog skupa članaka «Vjesnik» nije obavljena normalizacija *tfidf* vrijednosti pa je to bilo potrebno obaviti.

Podjela članaka po kategorijama je također ravnomjerna, 10000 članaka svakog skupa je ravnomjerno raspodijeljena na 8 kategorija, prikaz na slici (12).



Slika 12. Podjela članaka po kategorijama

6.1.4. Skup dokumenata «*Eurovoc*»

Pod pojmom Eurovoc misli se na opći višejezični pojmovnik namijenjen indeksiranju dokumenata iz djelatnosti Europskih zajednica dopunjen dodatkom hrvatskih specifičnih naziva[23] (HIDRA[24] - pojmovnik Eurovoc). Višejezični pojmovnici, poput Eurovoc-a, se koriste za indeksiranje i klasifikaciju različitih dokumenata na uniforman način preko univerzalnog, predefiniranog skupa pojmova koji su na neki način povezani.

Eurovoc pojmovnik je hijerarhijski organiziran i sadrži preko 6000 pojmova na 21 različitom jeziku, uključujući i hrvatski jezik. Eurovoc je strukturiran preko 8 razina koristeći konceptualne pojmove, što znači da su ti pojmovi dosta općeniti kako bi mogli opisati širok spektar koncepata za dano područje. Mnoge institucije u Europi koriste isti višejezični Eurovoc pojmovnik, s ciljem jednoznačnog indeksiranja službene dokumentacije. Korištenje istog višejezičnog pojmovnika od strane svih europskih zemalja je korisno za sve korisnike jer se bez problema, po potrebi, mogu pretraživati službeni dokumenti svih zemalja za jedan upit. Na slici (13) prikazana je hijerarhijska povezanost pojmova iz malenog djelića Eurovoc pojmovnika.

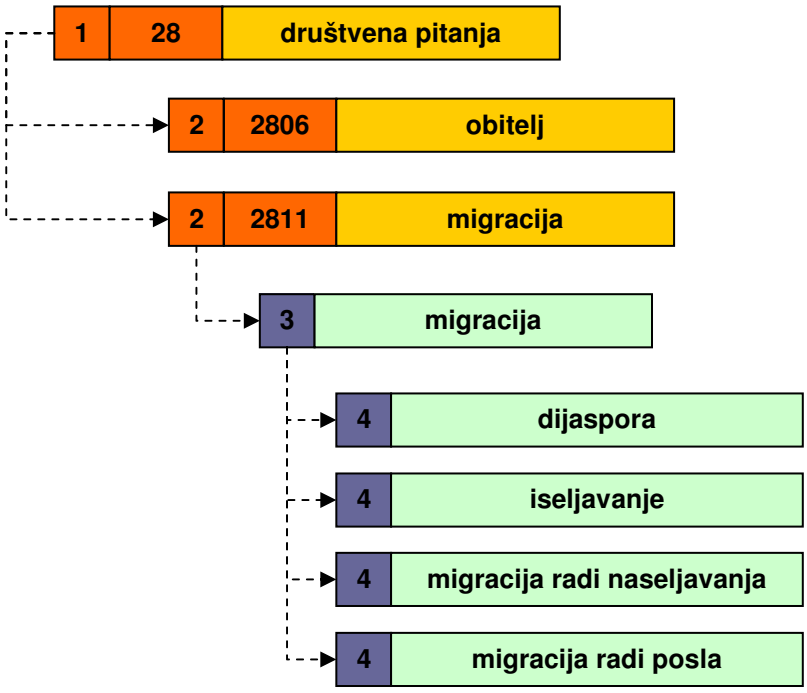
U sklopu ovog diplomskog rada pod pojmom Eurovoc misli se malen skup službenih dokumenata Vlade Republike Hrvatske (pravilnici, odluke, …) koji će se koristiti za potrebe testiranja funkcionalnosti programske implementacije polu-automatskog indeksera temeljenog na modificiranom *Fuzzy ARTMAP* algoritmu.

Taj skup dokumenata se sastoji od **2488** različitih dokumenata koji već imaju pridijeljene indekse od strane stručnjaka. Broj indeksa pridjeljenih dokumentu nije fiksan i isključivo ovisi o sadržaju dokumenta. Ovaj skup dokumenata je zapravo skup pravilnika, odluka i uredbi koje su donijela različita ministarstva Republike Hrvatske u periodu od 2000. – 2005. godine. Za potrebe testiranja polu-automatskog indeksera koristiti će se sažeti oblik skupa od **650** dokumenata za učenje i **300** dokumenata za testiranje. Skup pojmova korištenih za indeksiranje

[23] Detaljnije informacije možete pogledati web stranici www.hidra.hr

[24] Hrvatska informacijsko-dokumentacijska referalna agencija Vlade Republike Hrvatske

ovih dokumenata se sastoji od **760** pojmova, to je zapravo maleni pojmovnik koji je podskup Eurovoc pojmovnika. Dimenzija vektora, koji je numerička reprezentacija tekstualnog dokumenta, je **55933**.



Slika 13. Prikaz hijerarhijske organiziranosti Eurovoc pojmovnika

6.2. Predprocesiranje podataka

Kako bi se neuronskoj mreži mogli predstaviti ulazni dokumenti za učenje i testiranje, potrebno je prije tog čina obaviti predprocesiranje skupova dokumenata. Predprocesiranje je postupak svođenja svih dokumenata, nekog skupa dokumenata, na oblik koji je «razumljiv» klasifikatoru, obično je to *n*-dimenzionalni vektor-stupac, čiji su elementi popunjeni realnim ili cjelobrojnim numeričkim vrijednostima.

Po završetku predprocesiranja dobivaju se dvije matrice, jedna sadrži transformirane dokumente za postupak učenja mreže, a druga za postupak testiranja mreže.

Predprocesiranje se može podijeliti na nekoliko osnovnih koraka, na slici (14) je prikazan dijagram toka koji prikazuje postupak predprocesiranja kroz nekoliko koraka. Početno je potrebno razdijeliti ulazni skup dokumenata na skup za učenje i skup za testiranje, dokumente je potrebno odmah podijeliti na ta dva skupa jer se ne obavljaju iste računske operacije nad skupom za učenje i skupom za testiranje. Dokumenti su početno pohranjeni u nekom formatu, XML[25] ili običan tekstualni zapis, u datoteci na tvrdom disku. Postupkom čitanja i parsiranja dokumenti se prevode u memoriju računala gdje se im može brže pristupati pri obavljanju različitih računskih operacija nad njima. Nakon što su svi dokumenti parsirani potrebno je iz dokumenata za učenje generirati vektor svih različitih riječi koje se pojavljuju u tom skupu, pomoću tog vektora mogu se izgraditi numeričke reprezentacije dokumenata koje se predočavaju neuronskoj mreži.

Vektor raziičitih riječi se ne generira iz dokumenata za testiranje jer bi to indirektno utjecalo na rezultate klasifikacije, budući da bi neka informacija o skupu za testiranje bila ugrađena u skup za učenje (misli se na riječi koje se pojavljuju samo u skupu za testiranje).

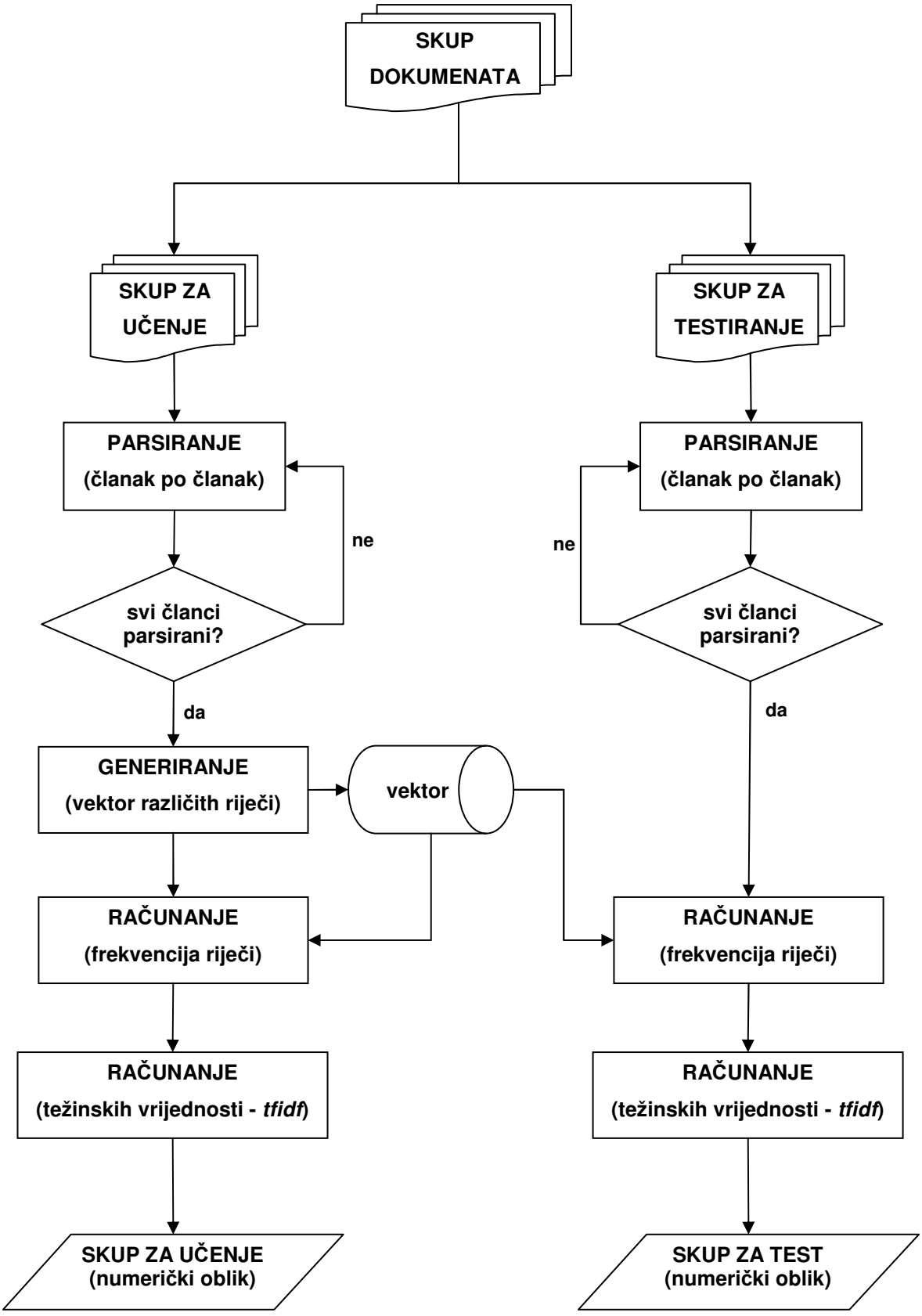
---

[25] E**X**tensible **M**arkup **L**anguage (XML) je strukturirani jezik namijenjen prezentaciji podataka, ima tekstualnu formu s korištenjem oznaka (eng. *tag*). XML pruža univerzalan način za razmjenu informacija. Vidi xml.coverpages.org/xml

Kad je vektor različitih riječi poznat, potrebno je obaviti sve *a priori* operacije kako bi se izračunali numerički vektori dokumenata s težinskim vrijednostima. Pod *a priori* operacijama misli se na računanje frekvencija pojavljivanja svake riječi unutar pojedinog dokumenta i cijelog skupa dokumenata. S tim podacima lako se izračunavaju težinske (*tfidf*) vrijednosti po formuli (2). Nakon ovog postupka dobiveni podaci se mogu predočavati neuronskoj mreži u postupcima učenja i testiranja mreže.

U sklopu ovog diplomskog rada bilo je potrebno obaviti eksperimente nad nekoliko skupova podataka, tj. dokumenata. Neke od tih skupova nije bilo potrebno predprocesirati, «Vjesnik» i «Akademski primjer». Skup «Vjesnik», verzija 1, je predprocesirana u vektore s težinskim vrijednostima pomoću *tfidf* funkcije, predprocesiranje je obavio Jan Šnajder s Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu. Skup «Akademski primjer» je malen pa je predprocesiranje obavljeno ručno, analogno slici (14).

Ostali skupovi, «Eurovoc» i «Croatia Weekly (CRO/EN)», su predprocesirani na način koji je prethodno objašnjen. Jedina razlika je što se za «Eurovoc» skup dokumenata koristila jednostavna težinska funkcija, umjesto *tfidf* funkcije, koja generira binarne vektore. Dakle, elementi vektora mogu poprimiti samo binarne vrijednosti, 1 ako je riječ na tom indeksu prisutna u dokumentu, 0 inače.



Slika 14. Dijagram toka predprocesiranja dokumenata

7. Rezultati

U ovom poglavlju grafički su prikazani rezultati eksperimenata nad svim skupovima podataka. U zadnjem podpoglavlju razmotrene su prednosti i mane obavljenih eksperimenata.

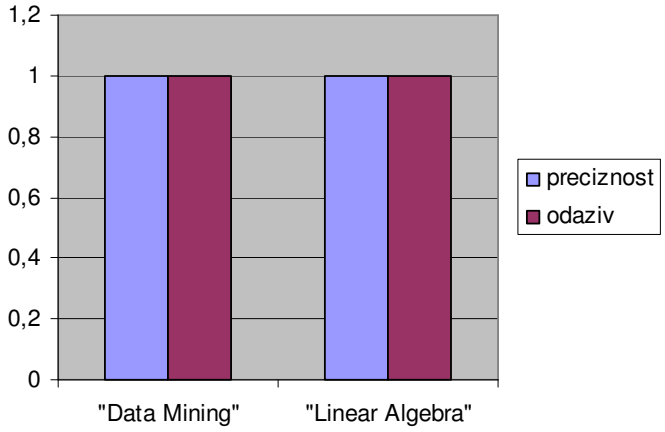
7.1. «Akademski primjer»

Parametri klasifikatora, za eksperimente nad skupom «Akademski primjer», su navedeni u tablici (9).

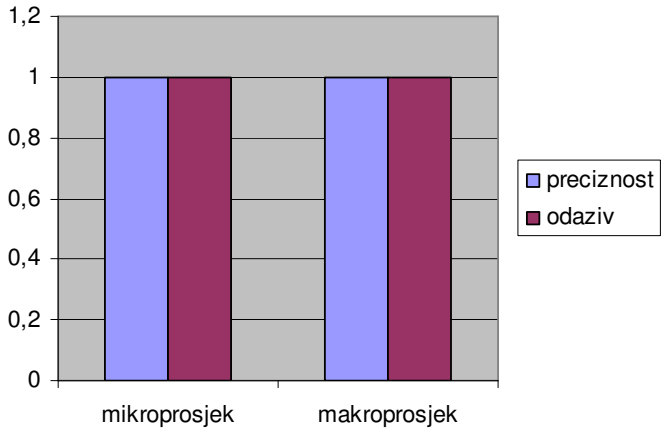
Tablica 9. Parametri klasifikatora za eksperiment «Akademski primjer»

veličina skupa za učenje	**14**
veličina skupa za testiranje	**5**
broj epoha učenja	**1**
parametar budnosti	**0.5**
stopa učenja	**0.7**
parametar odabira	**0.5**
epsilon parametar	**0.01**
broj kategorija	**2**
dimenzija ulaznog vektora (F_0 sloj)	**16**
broj generiranih neurona (F_2 sloj)	**3**

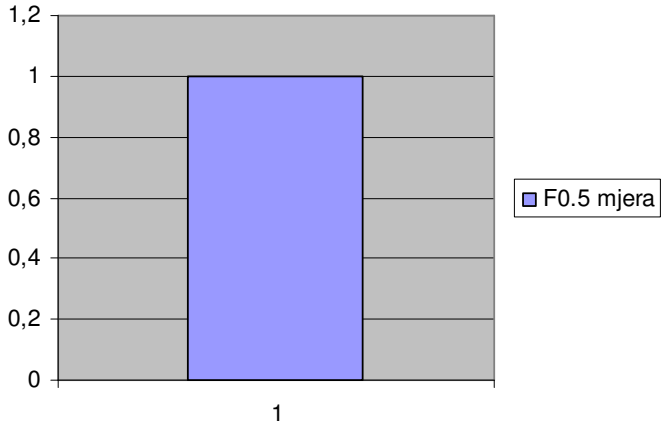
Kroz nekoliko sljedećih grafikona prikazani su dobiveni rezultati preko različitih mjera uspješnosti: preciznost i odaziv po kategorijama, slika (15); makroprosječna i mikroprosječna preciznost i odaziv, slika (16); $F_{0.5}$ mjera, slika (17).



Slika 15. Preciznost i odaziv po kategorijama



Slika 16. Makroprosječna/Mikroprosječna preciznost i odaziv



Slika 17. Mikroprosječna F0.5 mjera

7.2. «Croatia Weekly (EN)»

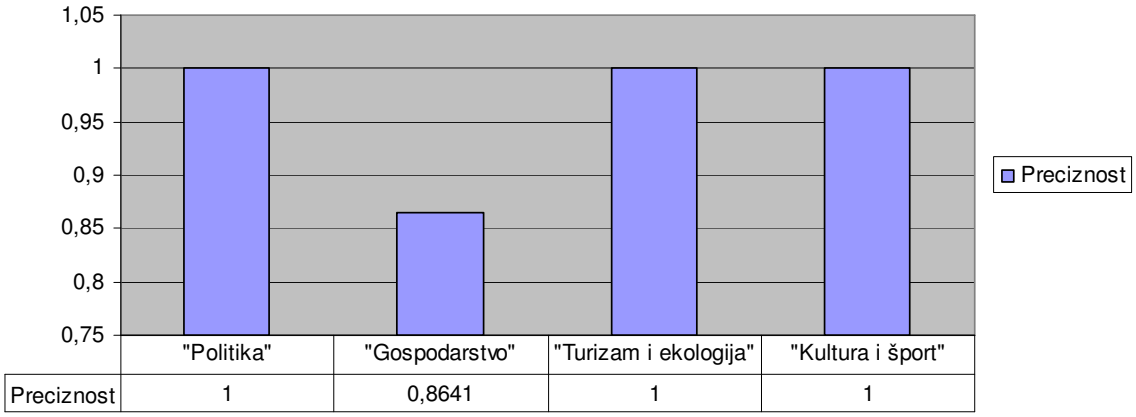
U ovom poglavlju obavljeni su eksperimenti nad «Croatia Weekly» skupom podataka u engleskoj verziji. Obavljena su dva eksperimenta, razlikuju se po broju dokumenata za testiranje i učenje.

7.2.1. Test – 1

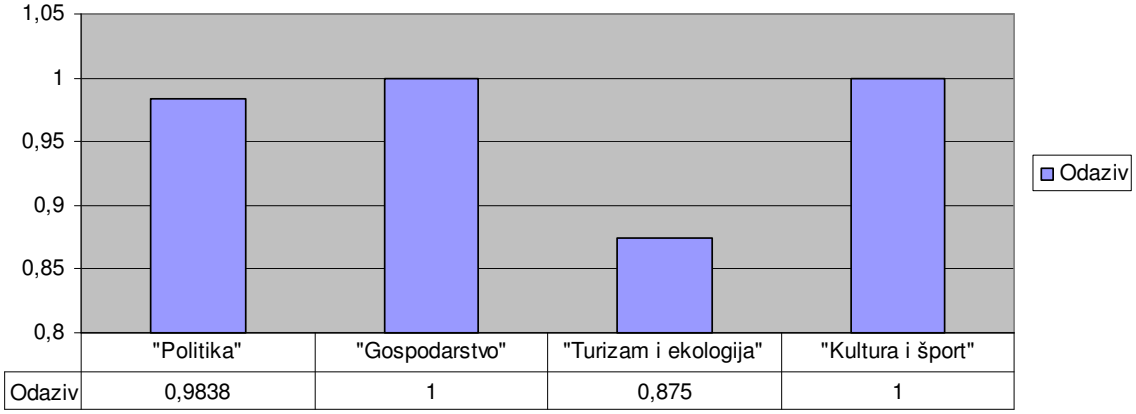
Parametri klasifikatora se nalaze u tablici (10), a rezultati eksperimenta su prikazani na slikama (18) – (21).

Tablica 10. Parametri klasifikatora za eksperiment 1 «Croatia Weekly (EN)»

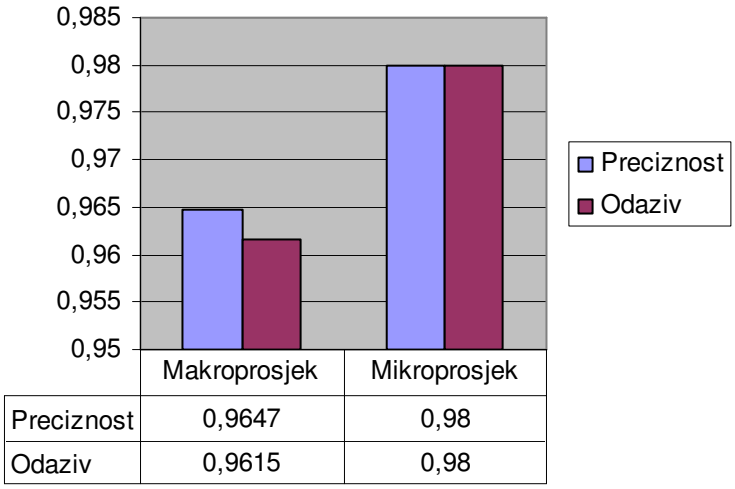
veličina skupa za učenje	**200**
veličina skupa za testiranje	**100**
broj epoha učenja	**1**
parametar budnosti	**0.5**
stopa učenja	**0.7**
parametar odabira	**0.5**
epsilon parametar	**0.01**
broj kategorija	**4**
dimenzija ulaznog vektora (F_0 sloj)	**31300**
broj generiranih neurona (F_2 sloj)	**119**



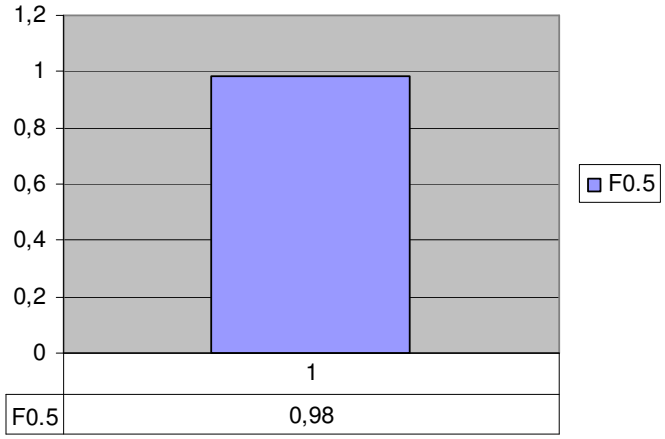
Slika 18. Preciznost po kategorijama



Slika 19. Odazivi po kategorijama



Slika 20. Makroprosječna/mikroprosječna preciznost i odaziv



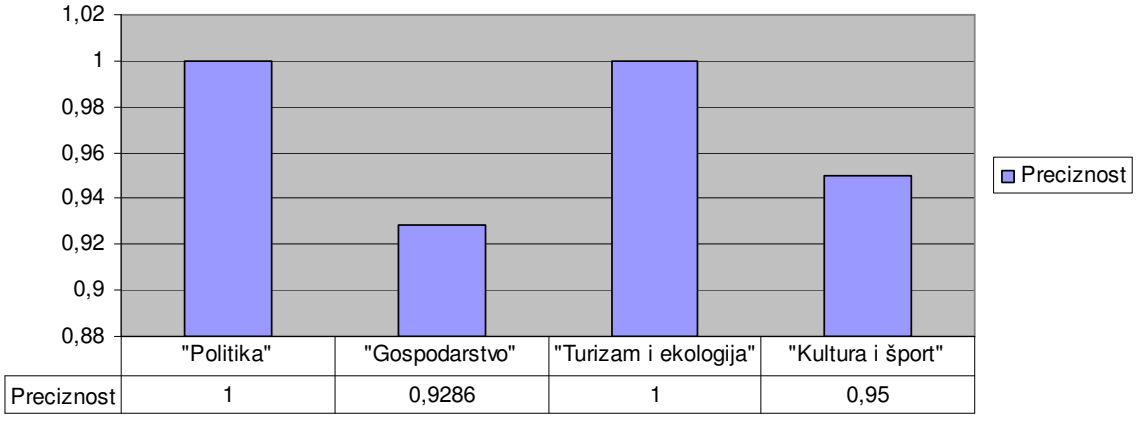
Slika 21. Mikroprosječna F0.5 mjera

7.2.2. Test – 2

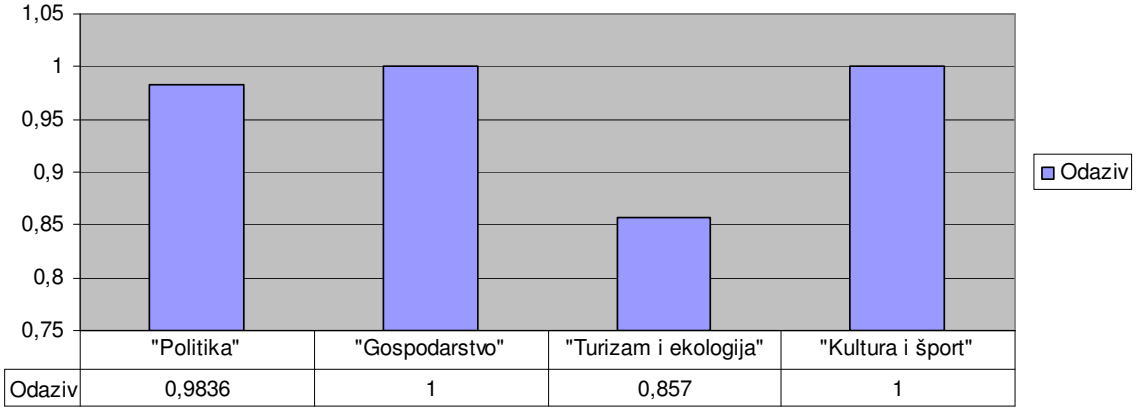
Parametri klasifikatora se nalaze u tablici (11), a rezultati eksperimenta su prikazani na slikama (22) – (25).

Tablica 11. Parametri klasifikatora za eksperiment 2 «Croatia Weekly (EN)»

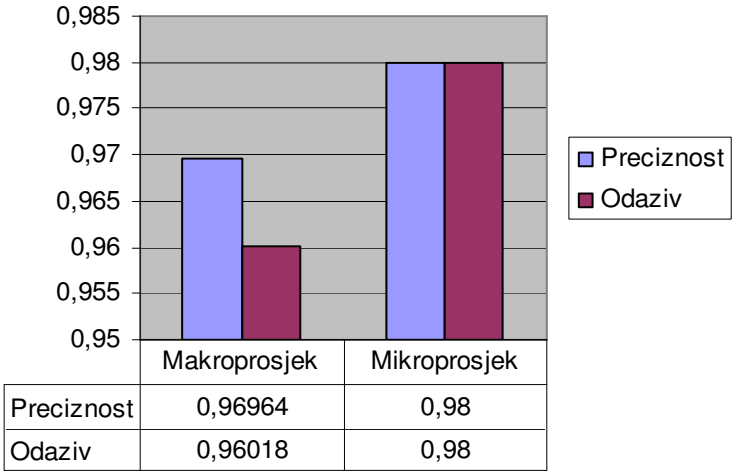
veličina skupa za učenje	**500**
veličina skupa za testiranje	**100**
broj epoha učenja	**1**
parametar budnosti	**0.5**
stopa učenja	**0.4**
parametar odabira	**0.01**
epsilon parametar	**0.01**
broj kategorija	**4**
dimenzija ulaznog vektora (F_0 sloj)	**31300**
broj generiranih neurona (F_2 sloj)	**402**



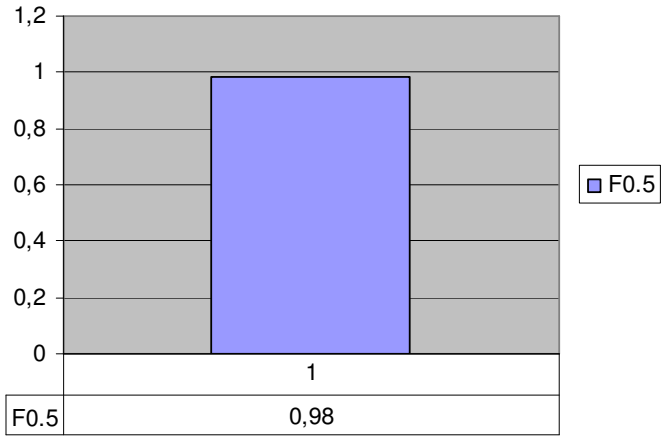
Slika 22. Preciznost po kategorijama



Slika 23. Odaziv po kategorijama



Slika 24. Mikroprosječna/makroprosječna preciznost i odaziv



Slika 25. Mikroprosječna F0.5 mjera

7.3. «Croatia Weekly (HR)»

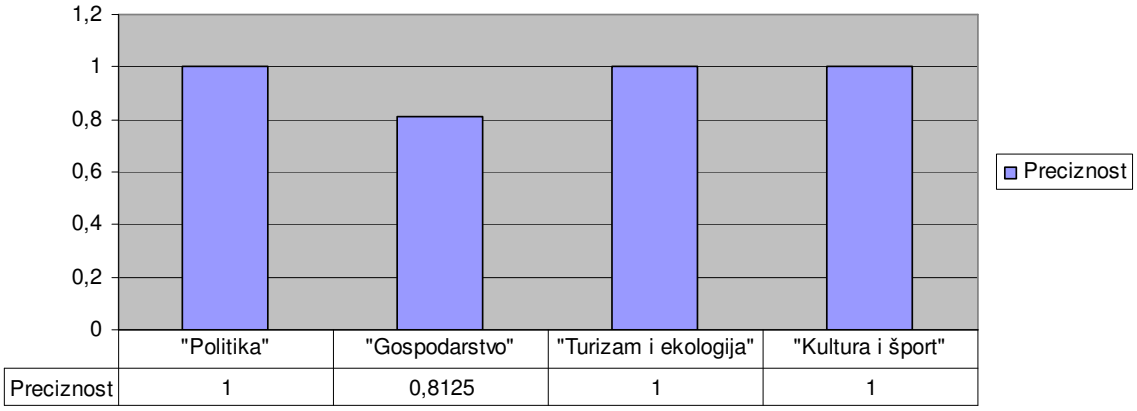
U ovom poglavlju obavljeni su eksperimenti nad «Croatia Weekly» skupom podataka u hrvatskoj verziji. Obavljena su dva eksperimenta, razlikuju se po broju dokumenata za testiranje i učenje.

7.3.1. Test – 1

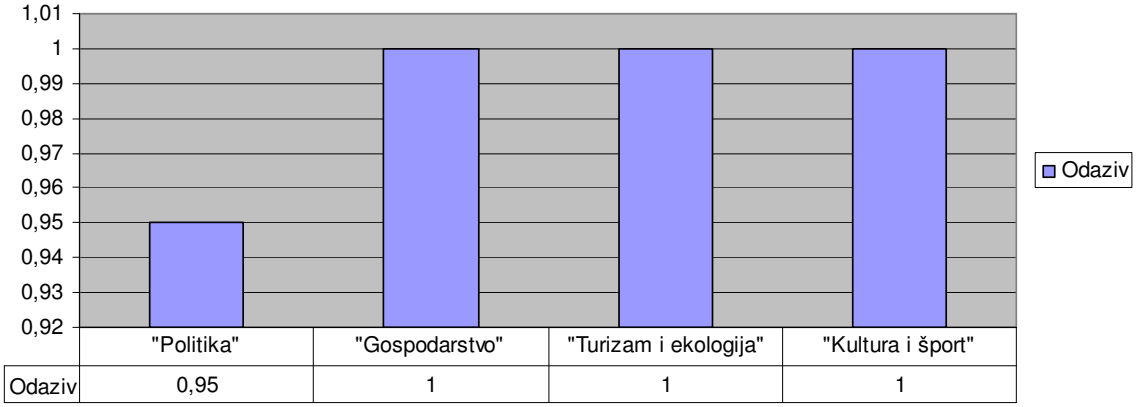
Parametri klasifikatora se nalaze u tablici (12), a rezultati eksperimenta su prikazani na slikama (26) – (29).

Tablica 12. Parametri klasifikatora za eksperiment 1 «Croatia Weekly (HR)»

veličina skupa za učenje	**200**
veličina skupa za testiranje	**100**
broj epoha učenja	**1**
parametar budnosti	**0.5**
stopa učenja	**0.7**
parametar odabira	**0.5**
epsilon parametar	**0.01**
broj kategorija	**4**
dimenzija ulaznog vektora (F_0 sloj)	**34705**
broj generiranih neurona (F_2 sloj)	**121**

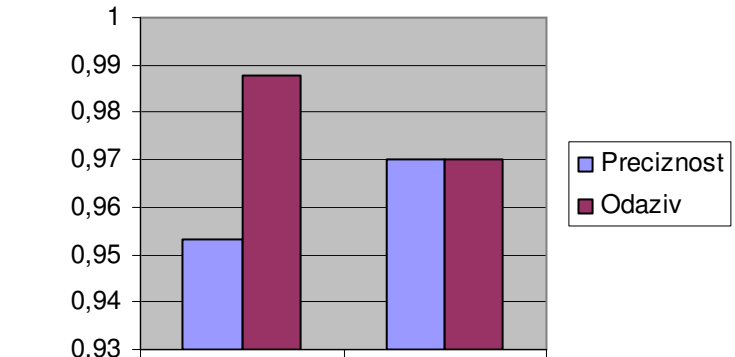


Slika 26. Preciznost po kategorijama



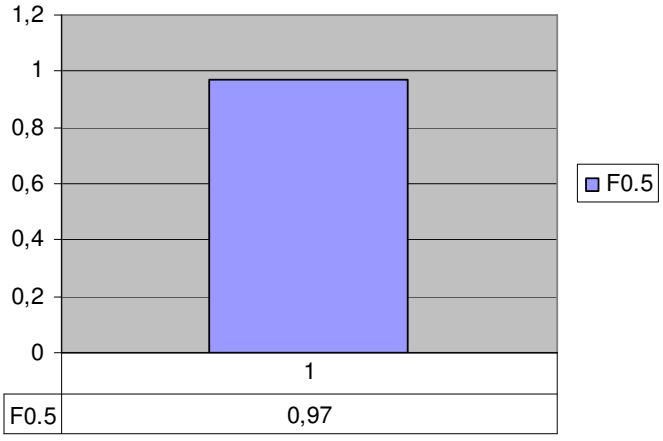
	"Politika"	"Gospodarstvo"	"Turizam i ekologija"	"Kultura i šport"
Odaziv	0,95	1	1	1

Slika 27. Odazivi po kategorijama



	Makroprosjek	Mikroprosjek
Preciznost	0,9531	0,97
Odaziv	0,9877	0,97

Slika 28. Makroprosječna/mikroprosječna preciznost i odaziv



	1
F0.5	0,97

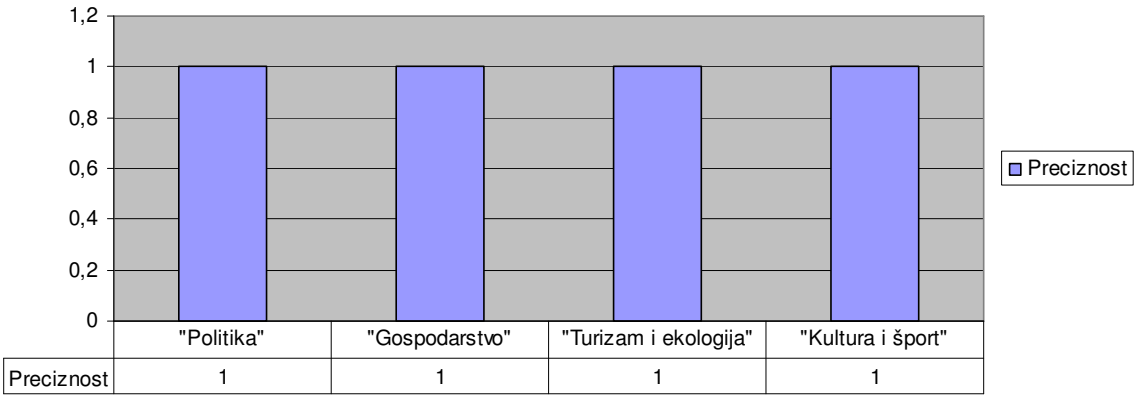
Slika 29. Mikroprosječna F0.5 mjera

7.3.2. Test – 2

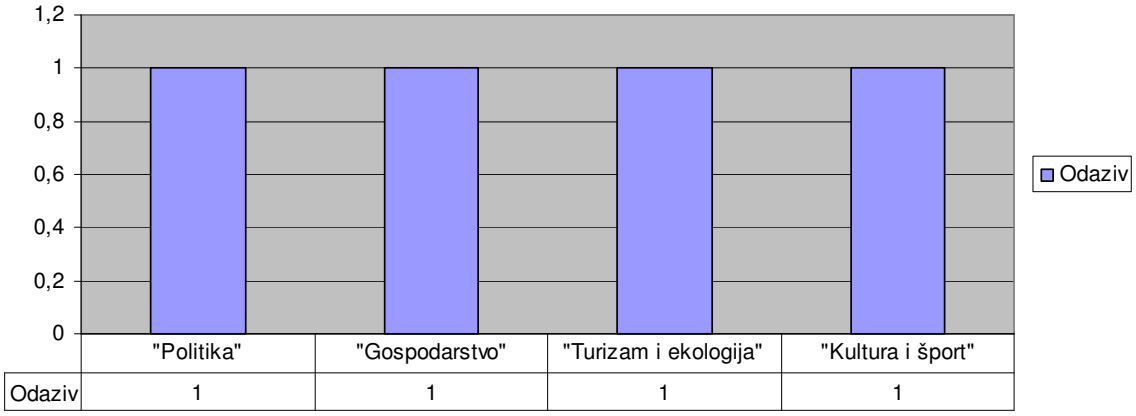
Parametri klasifikatora se nalaze u tablici (13), a rezultati eksperimenta su prikazani na slikama (30– (33).

Tablica 13. Parametri klasifikatora za eksperiment 2 «Croatia Weekly (HR)»

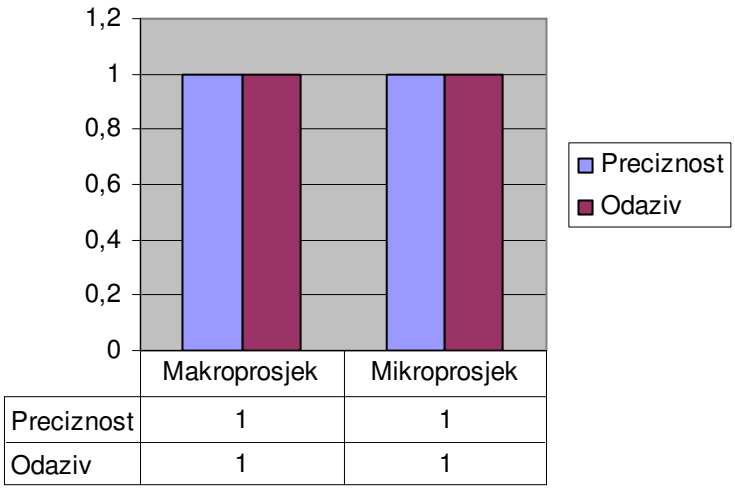
veličina skupa za učenje	**500**
veličina skupa za testiranje	**100**
broj epoha učenja	**1**
parametar budnosti	**0.5**
stopa učenja	**0.4**
parametar odabira	**0.01**
epsilon parametar	**0.01**
broj kategorija	**4**
dimenzija ulaznog vektora (F_0 sloj)	**34705**
broj generiranih neurona (F_2 sloj)	**227**



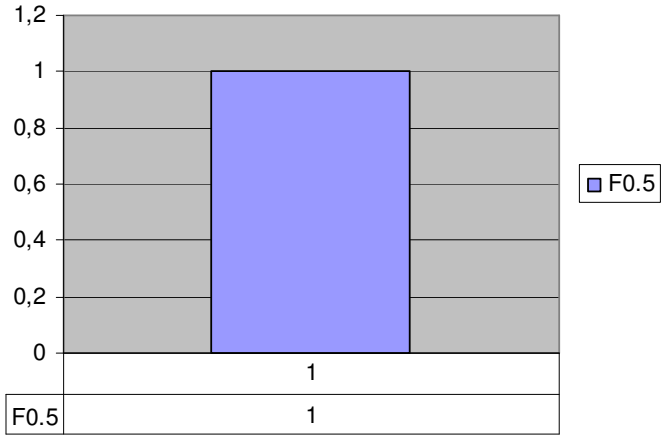
Slika 30. Preciznosti po kategorijama



Slika 31. Odazivi po kategorijama



Slika 32. Mikroprosječna/makroprosječna preciznost i odaziv



Slika 33. Mikroprosječna F0.5 mjera

7.4. «Vjesnik»

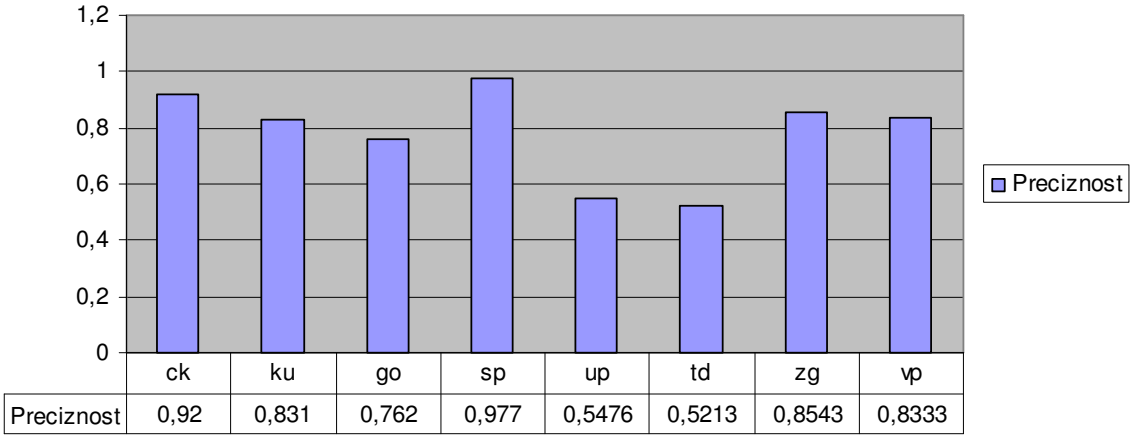
U ovom poglavlju obavljeni su eksperimenti nad verzijom 1 (varijanta 1) «Vjesnik» skupom podataka. Obavljena su dva eksperimenta koja se razlikuju po broju dokumenata za učenje i testiranje.

7.4.1. Test – 1

Parametri klasifikatora se nalaze u tablici (14), a rezultati eksperimenta su prikazani na slikama (34) – (37).

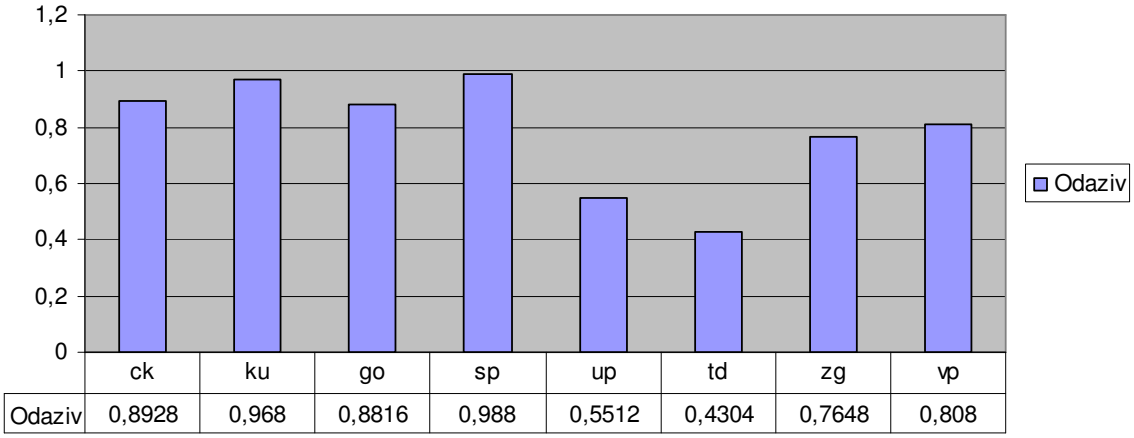
Tablica 14. Parametri klasifikatora za eksperiment 1 «Vjesnik»

veličina skupa za učenje	**10000**
veličina skupa za testiranje	**10000**
broj epoha učenja	**1**
parametar budnosti	**0.6**
stopa učenja	**0.7**
parametar odabira	**0.01**
epsilon parametar	**0.01**
broj kategorija	**8**
dimenzija ulaznog vektora (F_0 sloj)	**81582**
broj generiranih neurona (F_2 sloj)	**229**



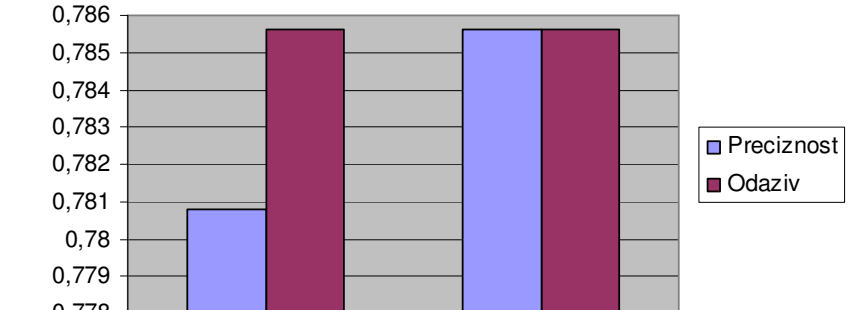
	ck	ku	go	sp	up	td	zg	vp
Preciznost	0,92	0,831	0,762	0,977	0,5476	0,5213	0,8543	0,8333

Slika 34. Preciznosti po kategorijama



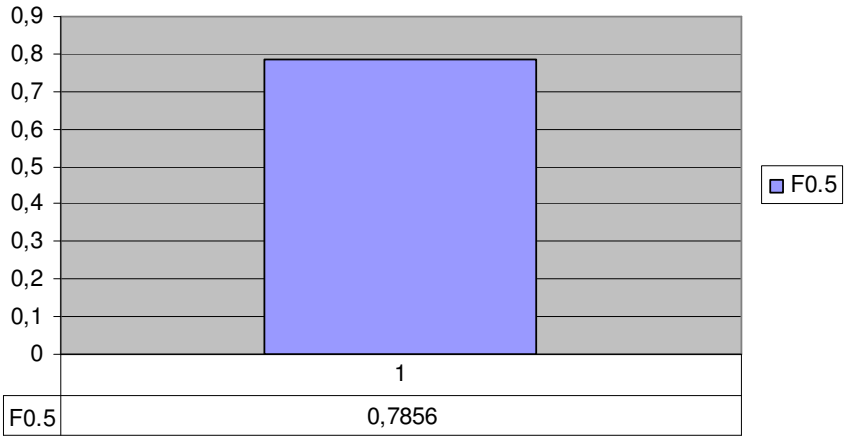
	ck	ku	go	sp	up	td	zg	vp
Odaziv	0,8928	0,968	0,8816	0,988	0,5512	0,4304	0,7648	0,808

Slika 35. Odazivi po kategorijama



	Makroprosjek	Mikroprosjek
Preciznost	0,7808	0,7856
Odaziv	0,7856	0,7856

Slika 36. Mikroprosječna/makroprosječna preciznost i odaziv



	1
F0.5	0,7856

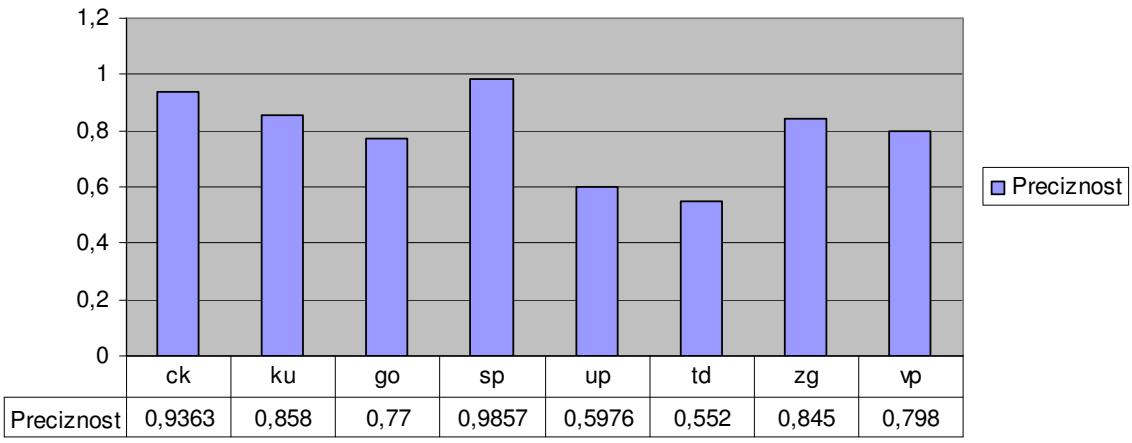
Slika 37. Mikroprosječna F0.5 mjera

7.4.2. Test – 2

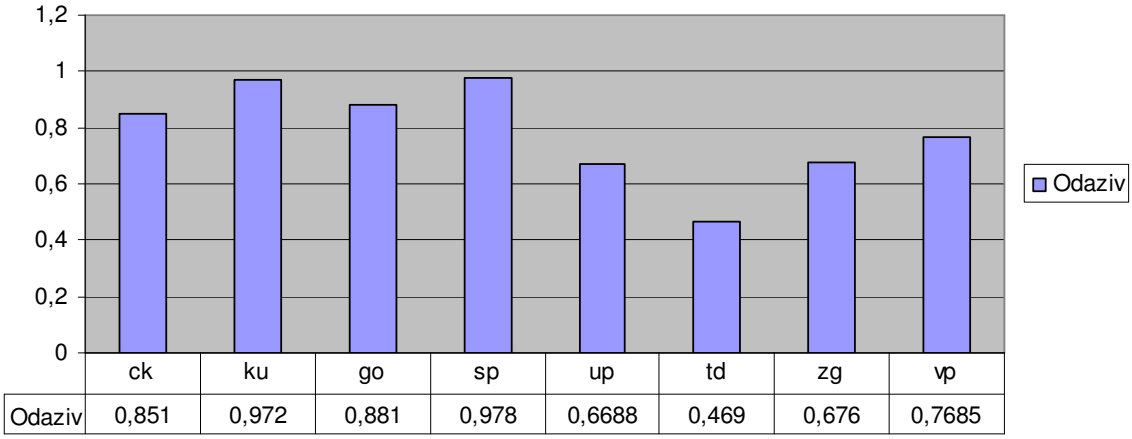
Parametri klasifikatora se nalaze u tablici (15), a rezultati eksperimenta su prikazani na slikama (38) – (41).

Tablica 15. Parametri klasifikatora za eksperiment 2 «Vjesnik»

veličina skupa za učenje	**2000**
veličina skupa za testiranje	**1000**
broj epoha učenja	**1**
parametar budnosti	**0.6**
stopa učenja	**0.7**
parametar odabira	**0.01**
epsilon parametar	**0.01**
broj kategorija	**8**
dimenzija ulaznog vektora (F_0 sloj)	**81582**
broj generiranih neurona (F_2 sloj)	**158**

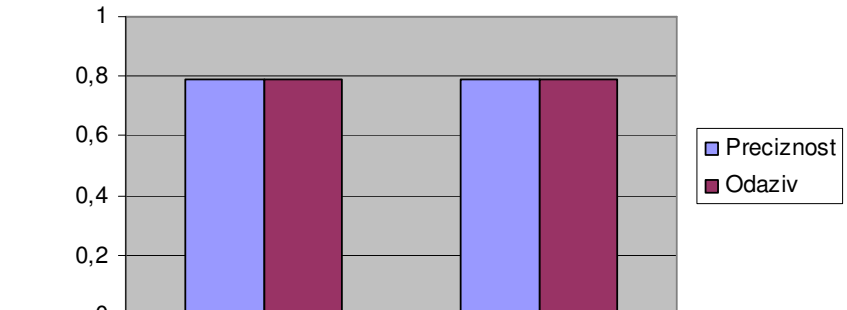


Slika 38. Preciznosti po kategorijama



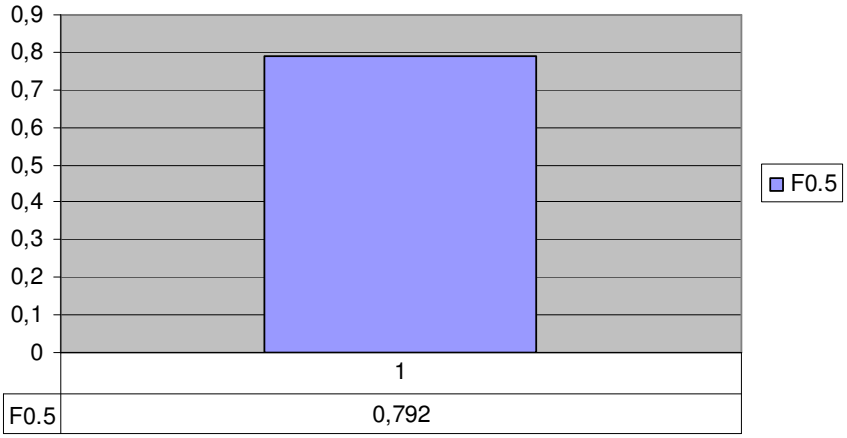
	ck	ku	go	sp	up	td	zg	vp
Odaziv	0,851	0,972	0,881	0,978	0,6688	0,469	0,676	0,7685

Slika 39. Odazivi po kategorijama



	Makroprosjek	Mikroprosjek
Preciznost	0,792	0,792
Odaziv	0,792	0,792

Slika 40. Mikroprosječna/makroprosječna preciznost i odaziv



	1
F0.5	0,792

Slika 41. Mikroprosječna F0.5 mjera

7.5. «Eurovoc»

Zbog specifičnosti problema rezultate za ovaj skup dokumenata nije moguće numerički, odnosno grafički, uobličiti. Budući da svaki dokument može imati dva i više pridjeljenih indeksa, čiji se broj može razlikovati od drugih dokumenata, onda se ne mogu koristiti prethodno navedene mjere uspješnosti klasifikatora.

Ovaj polu-automatski indekser radi na istom principu kao i klasifikator osim što ovdje dokumenti istovremeno mogu pripadati nekolicini kategorija, indeksa. Nakon postupka učenja, mreži se predočavaju testni primjerci (jedan po jedan) i onda mreža generira izlaz u obliku indeksa za koje smatra da opisuju testni primjerak. U programskoj implementaciji je moguće mijenjati broj neurona pobjednika koji se uzimaju u obzir pri generiranju indeksa za testni primjerak dokumenta, korištenje aplikacije je objašnjeno u dodatku A.

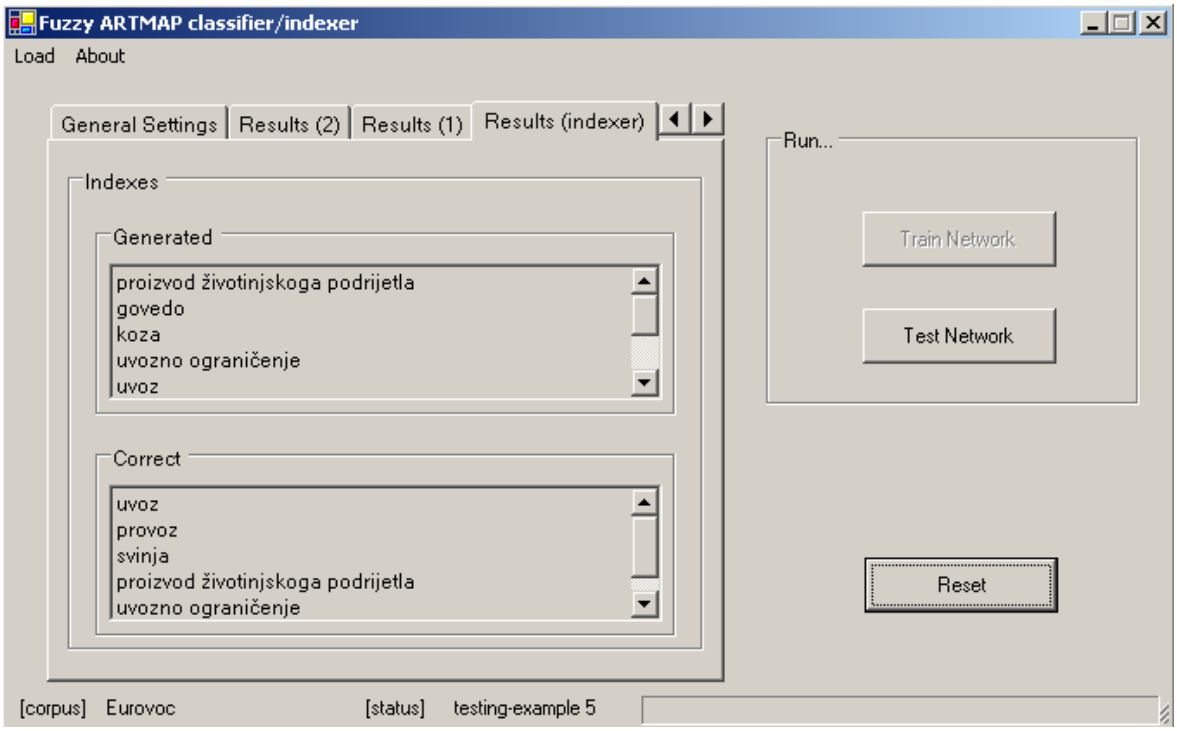
Budući da se rezultati ne mogu numerički prikazati zbog složenosti problema, onda su rezultati prikazani kroz nekoliko slika programske implementacije polu-automatskog indeksera temeljenog na *Fuzzy ARTMAP* algoritmu. U tablicama (16) – (18) su prikazani parametri mreže za pojedine eksperimente, a na slikama (42) – (44) su prikazani izlazi koje je mreža generirala i izlazi koje je mreža trebala generirati, točne izlaze.

Treba napomenuti da ovakva prezentacija rezultata ni u kom slučaju nije službena metodologija prezentiranja rezultata eksperimenata, već je samo pokazatelj mogućnosti korištene neuronske mreže pri problemima polu-automatskog indeksiranja.

7.5.1. Test – 1

Tablica 16. Parametri indeksera za eksperiment 1 «Eurovoc»

veličina skupa za učenje	**60**
veličina skupa za testiranje	**20**
broj epoha učenja	**1**
broj neurona pobjednika	**1**
fuzzy presjek ili *fuzzy* unija?	**-**
parametar budnosti	**0.6**
stopa učenja	**0.8**
parametar odabira	**0.01**
epsilon parametar	**0.01**
broj kategorija	**760**
dimenzija ulaznog vektora (F_0 sloj)	**55933**
broj generiranih neurona (F_2 sloj)	**58**

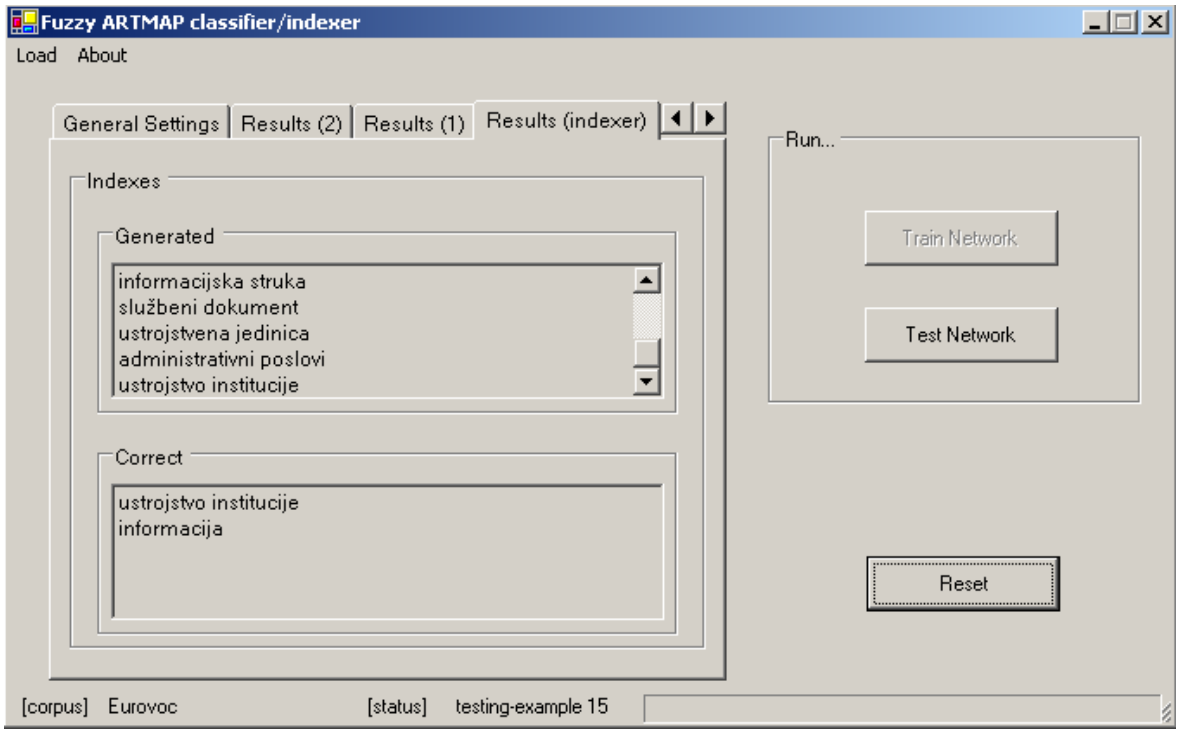


Slika 42. Ispravni i generirani indeksi (primjer dobrog indeksiranja)

7.5.2. Test – 2

Tablica 17. Parametri indeksera za eksperiment 2 «Eurovoc»

veličina skupa za učenje	**60**
veličina skupa za testiranje	**20**
broj epoha učenja	**1**
broj neurona pobjednika	**2**
fuzzy presjek ili *fuzzy* unija?	**unija**
parametar budnosti	**0.6**
stopa učenja	**0.8**
parametar odabira	**0.01**
epsilon parametar	**0.01**
broj kategorija	**760**
dimenzija ulaznog vektora (F_0 sloj)	**55933**
broj generiranih neurona (F_2 sloj)	**157**

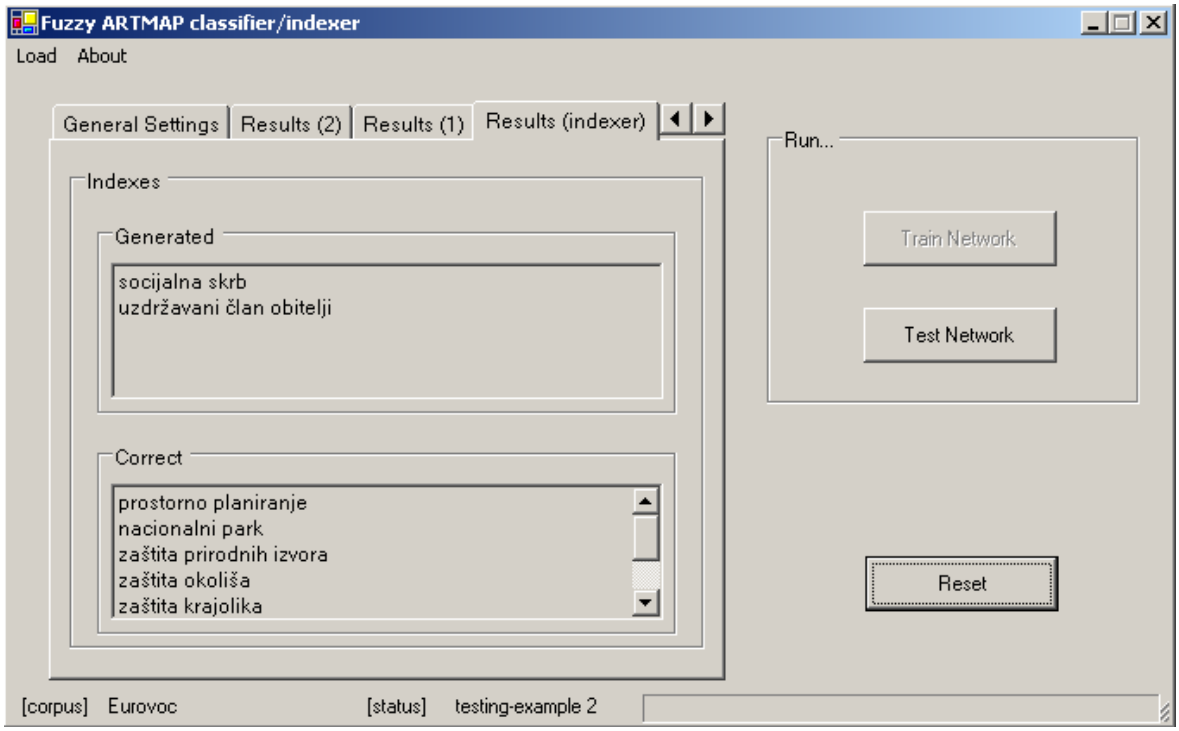


Slika 43. Ispravni i generirani indeksi (primjer prekomjernog pridjeljivanja indeksa)

7.5.3. Test – 3

Tablica 18. Parametri indeksera za eksperiment 3 «Eurovoc»

veličina skupa za učenje	**10**
veličina skupa za testiranje	**10**
broj epoha učenja	**1**
broj neurona pobjednika	**1**
fuzzy presjek ili *fuzzy* unija?	**-**
parametar budnosti	**1.0**
stopa učenja	**1.0**
parametar odabira	**0.7**
epsilon parametar	**0.01**
broj kategorija	**760**
dimenzija ulaznog vektora (F_0 sloj)	**55933**
broj generiranih neurona (F_2 sloj)	**10**



Slika 44. Ispravni i generirani indeksi (primjer lošeg indeksiranja)

7.6. Razmatranje dobivenih rezultata

Treba primijetiti da su korištene veličine skupova dokumenata za učenje i testiranje nedovoljno velike za iscrpno testiranje mogućnosti *Fuzzy ARTMAP* neuronske mreže, to ne vrijedi za «Vjesnik» i «Akademski» skup dokumenata. Skupovi dokumenata su se koristili u sažetom obliku iz jednog jedinog razloga, veći skupovi zahtijevaju ogromne računalne resurse s velikim vremenom izvođenja. Stoga, rezultate eksperimenata ne treba olako shvatiti, ali dobiveni rezultati mogu jako dobro pokazati kvalitetu navedene neuronske mreže.

Prvi eksperimenti su se obavili nad skupom dokumenata «Akademski primjer», koristi se za testiranje funkcionalnosti klasifikatora. *FAM* klasifikator je očekivano dao izvanredne rezultate, savršeno je naučio sve primjere za učenje, a da se pritom nije specijalizirao samo za te primjere. Zadržao je sposobnost generalizacije i na taj način je savršeno klasificirao sve testne primjerke dokumenata. Za ovaj eksperiment parametri klasifikatora nisu toliko bitni jer pri gotovo svim vrijednostima parametara (budnost, stopa učenja, …) klasifikator uspijeva savršeno klasificirati sve testne primjerke, što nam govori da je skup dokumenata vrlo vjerojatno linearno separabilan[26]. U samo tri neurona mreža je uspjela pohraniti sve karakteristike skupa za učenje.

Sljedeći eksperiment je obavljen nad «Croatia Weekly» skupom dokumenata, u engleskoj i hrvatskoj verziji. Iako su veličine skupova dokumenata za učenje malene (500, 100) klasifikator je dao odlične rezultate, što se može objasniti na nekoliko načina. Malen broj kategorija, četiri, dobro opisuju dokumente koji im pripadaju što klasifikatoru omogućuje da dobro nauči karakteristike svake kategorije. Budući da je ovaj skup sastavljen od članaka novinskog lista «Croatia Weekly» onda postoji mogućnost da taj list objavljuje članke koji su u pojedinoj kategoriji jako slični jedni drugima što olakšava problem klasifikacije. Rezultati za englesku i hrvatsku verziju skupa su podjednaki, malo bolji za englesku verziju što je i logično jer je hrvatski jezik morfološko bogatiji, što dodatno otežava klasifikaciju.

[26] Kategorije skupa dokumenata se mogu u hiperprostoru odvojiti hiperravninom

Eksperiment nad skupom dokumenata «Vjesnik» nije dao rezultate ni slične prethodnim eksperimentima, iako je ovdje velik skup dokumenata za učenje neuronska mreža nije uspjela sasvim dobro naučiti karakteristike svake kategorije, a time i cijelog skupa dokumenata. U drugom testiranju rezultati su neznatno bolji, ali i testni skup je manji pa je to mogući uzrok boljih rezultata. «Vjesnik» je specifičan skup jer je podijeljen u 8 kategorija od kojih su neke, poput «teme dana», «unutarnja politika» i «vanjska politika», neodređene i preopširne. Naime, postoji cijeli niz članaka koji bi se mogli svrstati pod kategoriju «teme dana», a da nemaju nikakve veze člancima koji su već pridijeljeni toj kategoriji, npr. današnja tema dana može biti «Početak pregovora s EU», a sutrašnja «ptičja gripa». Da su navedene kategorije dodatno podijeljene na nekoliko podkategorija klasifikator bi sigurno dao puno bolje rezultate, kao što ima odlične rezultate u kategorijama «sport», «crna kronika» i «kultura». Iz ovih promatranja vidi se da je *FAM* neuronskoj mreži dovoljna jedna zajednička karakteristika svim dokumentima u pojedinoj kategoriji kako bi dao dobre rezultate, problem se javlja s semantički i leksički različitim dokumentima koji su svrstani pod istu kategoriju, uvođenje podkategorija bi poboljšalo rezultate.

Sljedeći eksperiment je nezahvalan jer se ne može metodološki provjeriti učinkovitost indeksera, a da ne koristimo stručnjaka za to područje. Upravo zbog većeg broja indeksa koji mogu biti pridijeljeni jednom dokumentu ne može se odrediti da li je indekser dobro indeksirao ako je ispravno dodijelio 3 od 4 indeksa, 2 od 3 indeksa ili 16 od 17 indeksa. Mjerenje učinkovitosti temeljeno samo na postotku ispravno pridijeljenih indeksa nije dovoljno jer ponekad jedan nepridijeljeni indeks može biti važniji od pridijeljenih indeksa, npr. dokumentu s indeksima «pomorsko dobro», «granica» i «vlasništvo» indekser pridijeli samo indekse «granica» i «vlasništvo», učinkovitost indeksera je 66%, ali i dalje se ne zna bitna stvar: dokument govori o granici na pomorskom dobru! Problem kod ovog, sažetog «Eurovoc» skupa dokumenata, je što ima nedovoljno velik skup primjeraka za učenje koji mogu poprimiti velik broj indeksa, najviše 760. Ti dokumenti mogu biti ponekad jako slični, ali mogu imati različite indekse što zbunjuje klasifikator. *FAM* neuronska mreža radi na principu povezivanja neurona pobjednika u F_2^a sloju s pripadnom kategorijom preko asocijativne memorije, u slučaju da se pripadna kategorija neurona pobjednika ne podudara s kategorijom

primjerka za učenje onda se postupkom *match tracking* pronalazi novi pobjednički neuron koji će imati istu kategoriju kao i primjerak ili će se stvoriti novi neuron koji će savršeno naučiti taj primjerak. Zbog te činjenice *FAM* neuronska mreža neće dobro naučiti karakteristike skupa dokumenata koji imaju širok raspon pripadajućih kategorija, potrebno je obaviti testiranje na jako velikom skupu kako bi se dobili kvalitetni rezultati, koji u tom slučaju ne bi izostali uzevši u obzir prethodno dobivene rezultate.

8. Zaključak

Pisanjem ovog diplomskog rada rastumačile su mi se brojne teme vezane uz probleme klasifikacije i indeksiranja teksta. Nadam se da će se ovim diplomskim radom posvetiti veća pozornost primjeni neuronskih mreža na klasifikaciju teksta, posebno *Fuzzy ARTMAP* neuronske mreže, i uspjeti zainteresirati studente budućih generacija da nastave istraživanje ovog područja.

Važnost ove teme očituje se u pronalaženju kvalitetnog algoritma koji će zadovoljavati potrebe klasifikacije tekstualnih sadržaja, na engleskom i hrvatskom jeziku. Klasifikacija hrvatskog jezika je kompliciranija zbog njegovog morfološkog bogatstva te bi se u tom smjeru trebala vršiti nova istraživanja. Budući da je područje dubinske analize teksta (eng. *text mining*) još u začetku razvoja, posebno u Hrvatskoj, trebalo bi istaknuti pionire istraživače, prof. dr. sc. Bojana Dalbelo Bašić s Fakulteta elektrotehnike i računarstva i prof. dr. sc. Marka Tadića s Filozofskog fakulteta Sveučilišta u Zagrebu.

Tijekom razrade ove teme otvorila su mi se brojna pitanja vezana uz principe rada *FAM* neuronske mreže te se nadam će neka od mojih daljnjih istraživanja ići u tom smjeru.

9. Literatura

1. Stergiou, Chris: «What is a Neural Network», s Interneta, http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/cs11/article1.html , 1996.
2. Stanford Projects, s Interneta, http://www-cse.stanford.edu/classes/sophomore-college/projects-00/neural-networks , 2000.
3. Wikipedia, s Interneta, http://en.wikipedia.org , 2005.
4. «History of Operating Systems», s Interneta, http://www.osdata.com/kind/history.htm , 2005.
5. Bellis, Mary: «Intel 4004 - The World's First Single Chip Microprocessor», *Inventors of The Modern Computer*, s Interneta, http://inventors.about.com , 2005.
6. Bellis, Mary: «The History of the Transistor - John Bardeen, Walter Brattain, and William Shockley», *Inventors of The Modern Computer*, s Interneta, http://inventors.about.com , 2005.
7. Grossberg, Stephen: «The Attentive Brain», *The American Scientist*, vol. 83, pp. 438-449, 1995.
8. Carpenter, Gail; Grossberg, Stephen: «Adaptive Resonance Theory», *The Handbook of Brain Theory and Neural Networks*, 2nd Edition, Massachusetts, MIT Press, 1998.
9. Carpenter, Gail: «Default ARTMAP», *CAS/CNS Technical Report TR-2003-008*, Portland, 2003.
10. Carpenter, Gail; Ross, William: «ART-EMAP: A Neural Network Architecture for Object Recognition by Evidence Accumulation», *IEEE Transactions on Neural Networks*, vol. 6, no. 4, Srpanj, 1995.
11. Carpenter, Gail; Grossberg, Stephen; Reynolds, John: «ARTMAP: Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network», *Neural Networks*, vol. 4, pp. 565-588, 1991.
12. Gomez-Sanchez, Eduardo; Dimitiriadis, Yannis; Cano-Izquierdo, Jose Manuel; Lopez-Coronado, Juan: «µARTMAP: Use of Mutual Information for Category Reduction in Fuzzy ARTMAP», *IEEE Transactions on Neural Networks,* vol. 13, no. 1, Siječanj, 2002.
13. Zadeh, Lofti: «Fuzzy Sets», *Information and Control*, vol. 8, no. 4, pp. 383-353, lipanj, 1965.
14. Busque, Martin; Parizeau, Marc: «A Comparison of Fuzzy ARTMAP and Multilayer Perceptron for Handwritten Digit Recognition», *Universite Laval*, listopad, 1997.
15. Carpenter, Gail; Grossberg, Stephen; Rosen, David: «Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system», *Neural Networks,* vol. 4, pp. 759-771, 1991.
16. Carpenter, Gail; et al: «Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps», *IEEE Transactions on Neural Networks*, vol. 3, no. 5, rujan, 1992.
17. Dobša, Jasminka: «Comparison of Information Retrieval Techniques: Latent Semantic Indexing (LSI) and Concept Indexing (CI)», *Fakultet organizacije i informatike*, Varaždin, 2004.
18. Nordlie, Ragnar: «Automated or Human Indexing», *European Library Automation Group: Integrating Heterogeneous Resources*, 25th Library Systems Seminar, Prag, lipanj, 2001.

19. Sebastiani, Fabrzio; «A Tutorial on Automated Text Categorisation», *Proceedings of ASAI-99, First Argentinian Symposium on Artificial Intelligence*, 1999.

20. Salton, G.; Buckley, C: «Term-weighting approaches in automatic text retrieval», *Information Processing and Management,* 24(5), pp. 513-523, 1998.

21. Fuhr, N; et al: «AIR/X – Rule-based Multistage Indexing System for Large Subject Fields», *In proceedings of RIAO'91, 3rd International Conference «Recherche d'Information Assistee par Ordinateur»*, pp. 606-623, Barcelona, 1991.

22. Rauber, Andreas; Merkl, Dieter: «Text Mining in the SOMLib Digital Library System: The Representation of Topics and Genres», *Applied Intelligence*, vol. 18, pp. 271-293, 2003.

23. Berthold, Michael, Hand, J., David: *Intelligent Data Analysis – An Introduction*, 2nd Edition, Springer, 2002.

24. Russell, Stuart; Norvig, Peter: *Artificial Intelligence – A Modern Approach*, 2nd Edition, Prentice Hall, 2003.

25. Sapojnikova, Elena: «ART-based Fuzzy Classifiers: ART Fuzzy Networks for Automatic Classification», doktorska dizertacija, Eberhard-Karls-Universität Tübingen, Njemačka, 2003.

26. Schiel, Ulrich; Sodré Ferreira de Sousa, Ianna; Ferneda, Edberto: «Semi-Automatic Indexing of Multilingual Documents», Universidade Federal da Paraíba, Campina Grande, Brazil

27. Cvitaš, Ana: «Automatsko indeksiranje dokumenata u modelu vektorskog prostora», diplomski rad, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, srpanj, 2005.

28. Buršić, Dajana: «Primjena koncepta jezgrenih funkcija u dubinskoj analizi teksta», diplomski rad, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, srpanj, 2005.

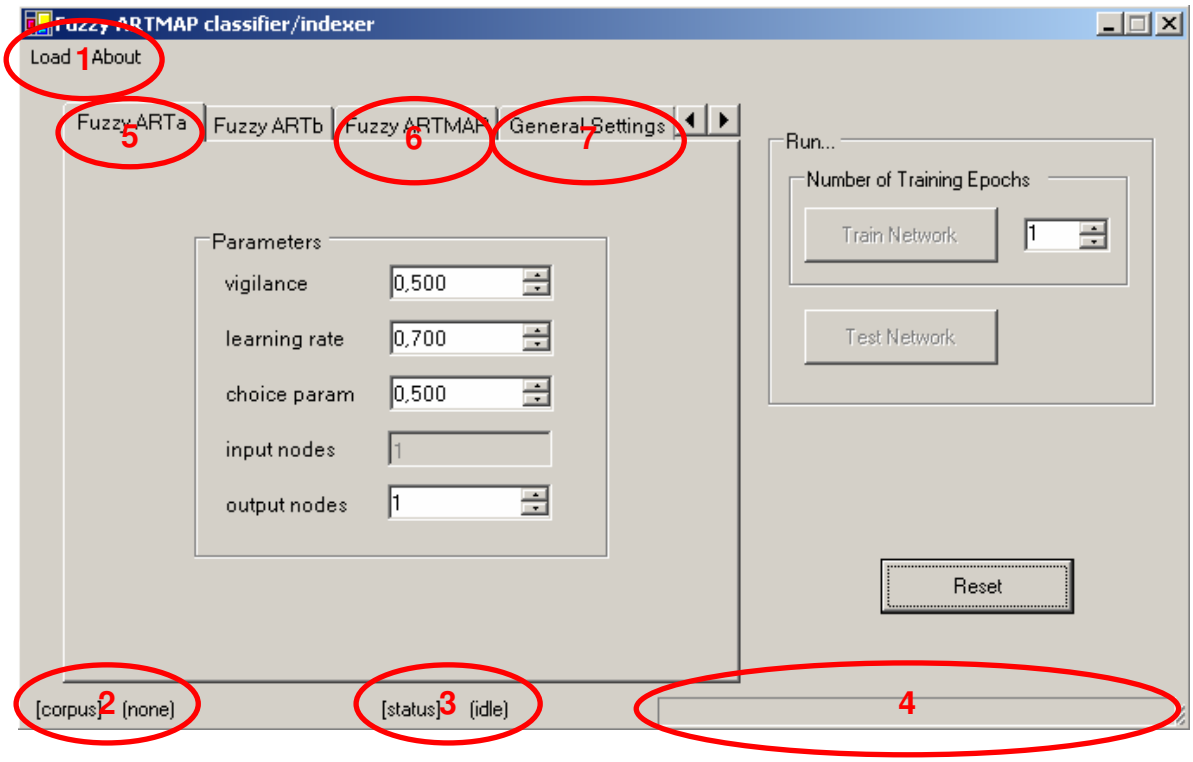
29. Malenica, Mislav: «Primjena jezgrenih metoda u kategorizaciji teksta», diplomski rad, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, rujan, 2004.

30. Ahel, Renee: «Postupak klasifikacije teksta temeljen na k-nn metodi i naivnom Bayesovom klasifikatoru», diplomski rad, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, rujan, 2003.

Dodatak A: Korištenje programskog ostvarenja

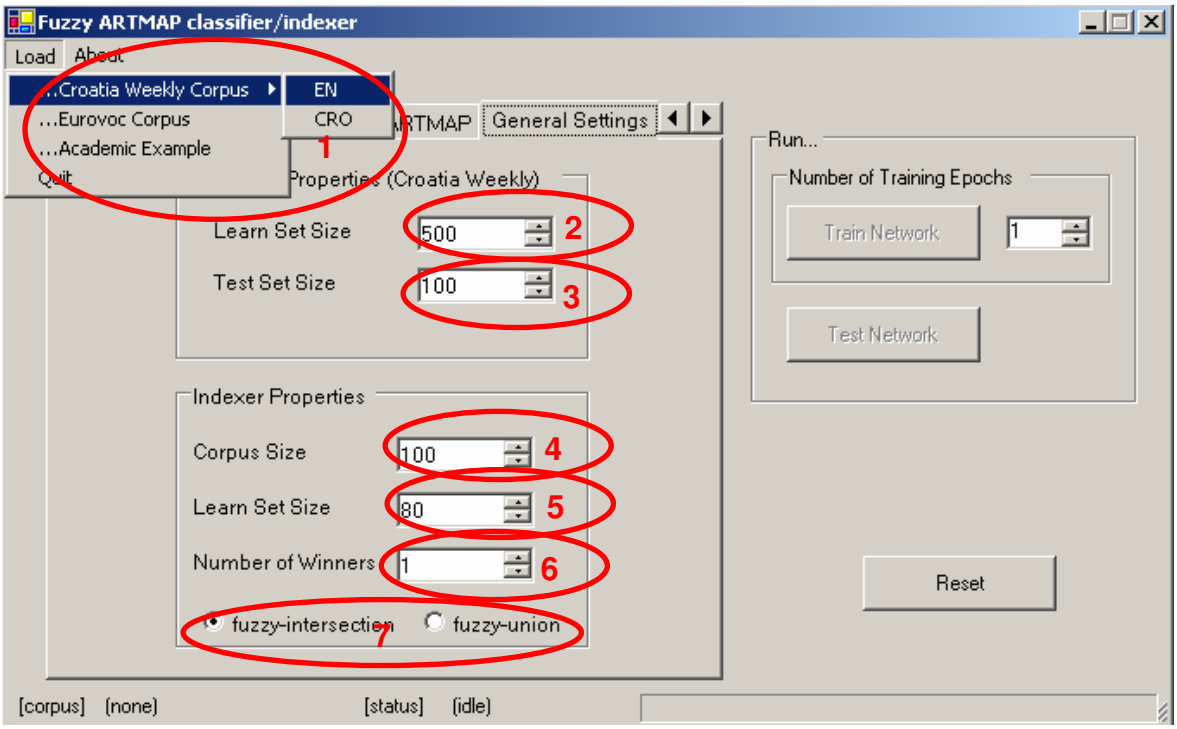
Kroz sljedećih nekoliko stranica bit će objašnjeno korištenje aplikacije ***Fuzzy ARTMAP classifier/indexer*** koja je izrađena za potrebe diplomskog rada. Uz ovu aplikaciju postoji i ekvivalentna aplikacija u C programskom jeziku, konzolna aplikacija (bez sučelja), ona se koristi samo za potrebe klasifikacije «Vjesnik» skupa dokumenata. Aplikacija u C-u se koristi zbog preopsežnosti skupa «Vjesnik» jer se time povećava vrijeme izvođenja aplikacije, a C aplikacije se kompiliraju u strojni kôd pa su nešto brže od njihovih ekvivalenata u C# programskom jeziku.

FAM classifier/indexer aplikacija je napisana u C# programskom jeziku, a za format ulaznih podataka i međurezultata korišten je XML. Budući da *FAM* algoritam zahtijeva interaktivnost s korisnikom, unos parametara, aplikacija ima i grafičko korisničko sučelje (eng. *Graphical Users Interface – GUI*). Na slici (45) je prikazan prozor aplikacije nakon što se pokrene.



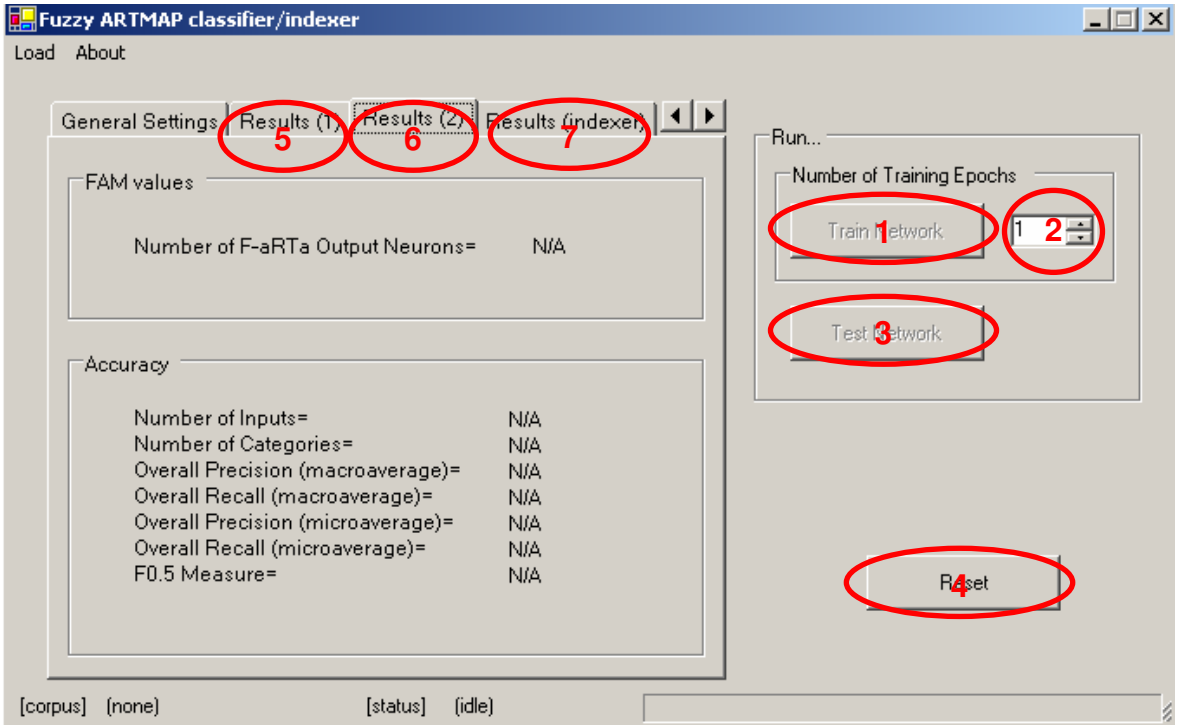
Slika 45. *FAM* aplikacija (a)

Korisničko sučelje se sastoji od nekoliko dijelova, zaokruženi su na slici (45). Dio (1) je izbornik putem kojeg se odabire skup dokumenata za učenje i testiranje, ti skupovi su već predprocesirani i nalaze se u istom direktoriju gdje je izvršna datoteka aplikacije, na slici (46) je prikazan sadržaj izbornika. (2) je pokazatelj trenutno učitanog skupa dokumenata, a (3) je pokazatelj operacije koju aplikacija u danom trenutku obavlja, učenje ili testiranje. (4) je pokazatelj statusa nekih procesa, poput učitavanja dokumenata, učenja mreže ili testiranja mreže. (5) je skup formi za unos parametara *Fuzzy ART*$_a$ modula: parametar budnosti, stopa učenja, parametar odabira i početni broj neurona u F_2^a sloju. U dijelu (6) se nalazi forma za unos parametra epsilon, koristi se kod *match tracking* procesa za povećavanje parametra budnosti *F-ART*$_a$ modula. (7) sadrži dvije grupe formi za unos parametara, jedna se koristi za «Croatia Weekly», a druga za «Eurovoc» skup dokumenata, na slici (46) se nalazi detaljni prikaz tih formi.



Slika 46. *FAM* aplikacija (b)

Na slici (46) prikazane su ostale dijelovi *FAM* aplikacije. Detaljni prikaz izbornika se nalazi na dijelu (1), prije nego što se započne s učenjem mreže potrebno je učitati dokumente preko izbornika. (2) i (3) se odnose na «Croatia Weekly» skup dokumenata i preko tih formi se određuje veličina skupa za testiranje i učenje, prije nego se učitaju «Croatia Weekly» dokumenti potrebno je podesiti navedene parametre. Ostali dijelovi se odnose na «Eurovoc» skup dokumenata i parametre koji reguliraju postupak testiranja kod indeksera. (4) i (5) određuju veličinu «Eurovoc» skupa za učenje i testiranje, njih je potrebno podesiti prije učitavanja skupa. (6) se koristi za postupak testiranja, određuje koliko se pobjedničkih neurona uzima u obzir kad se mreži predoči testni primjerak, a (7) određuje na koji način će se pripadni indeksi pobjedničkih neurona spojiti kako bi se generirao zajednički skup indeksa. Koristi se *fuzzy* presjek ili *fuzzy* unija.



Slika 47. *FAM* aplikacija (c)

Na slici (47) prikazan je ostatak dijelova *FAM* aplikacije, najbitniji su dijelovi (1), (2) i (3). Nakon učitavanja dokumenata i podešavanja parametara *FAM* mreže može se pristupiti učenju mreže, što se radi preko gumba (1). S (2) se može regulirati broj epoha učenja, tj. broj predstavljanja primjeraka za učenje. Napredak učenja mreže se može pratiti preko (4) sa slike (45). Nakon učenja, mrežu je potrebno testirati na primjercima koje još nikad nije «vidjela», koristi se gumb (3). Napredak procesa testiranja se također može pratiti preko (4) sa slike (45).

Nakon testiranja rezultati se mogu vidjeti u dijelovima (5), (6) i (7), s tim da se u (7) nalaze rezultati samo za «Eurovoc» skup, generirani i ispravni indeksi. (5) prikazuje preciznosti i odazive po kategorijama za sve skupove, osim «Eurovoc». (6) prikazuje makroprosječnu preciznost i odaziv, mikroprosječnu preciznost i odaziv, mikroprosječnu $F_{0.5}$ mjeru, broj kategorija, broj testnih primjeraka i broj generiranih neurona u F_2^a sloju.

Budući da je za potrebe diplomskog rada dovoljna pojednostavljena verzija *Fuzzy ARTMAP* algoritma, onda su forme za unos parametara *Fuzzy* ART_b modula i neke forme za unos parametara *FAM* modula onemogućene i ostavljene su za budući rad.