Izvirni znanstveni članek ■

Hiter in preprost algoritem za razdvoumljanje simbolov genov

Andrej Kastrin, Dimitar Hristovski

Izvleček. Razdvoumljanje simbolov genov je raziskovalno zelo aktualno področje. Največji problem predstavlja ločevanje med besednimi simboli, ki označujejo gene oz. njihove produkte ter simboli, ki se nanašajo na ostale biomedicinske koncepte (npr. CT, MR). V članku predstavimo hiter in preprost pristop k razdvoumljanju, ki temelji na razvrščanju MEDLINE® zapisov v genetsko in negenetsko domeno, ob predpostavki, da se simboli v genetski domeni najverjetneje nanašajo na gene. Algoritem sloni na statistični primerjavi domensko reprezentativnih korpusov zgrajenih na osnovi MeSH® deskriptorjev. Metoda je jasno razumljiva, preprosta za implementacijo ter računsko nezahtevna. Točnost razvrstitve, merjena na validacijski množici podatkov, je znašala 0,91. Algoritem je implementiran kot pomožni sistem za razdvoumljanje simbolov genov v sistemu za odkrivanje dejanskih ali potencialnih zakonitosti iz bibliografskih podatkov BITOLA.

A fast and simple document classification algorithm for gene symbol disambiguation

Abstract. Gene symbol disambiguation is an important problem for biomedical text mining systems. When detecting gene symbols in MEDLINE® citations one of the biggest challenges is the fact that many gene symbols also denote other, more general biomedical concepts (e.g. CT, MR). Our approach to this problem is first to classify the citations into genetic and non-genetic domains and then to detect gene symbols only in the genetic domain. We used ontological information provided by Medical Subject Headings (MeSH®) for this classification task. The proposed algorithm is fast and is able to process the full MEDLINE distribution in a few hours. It achieves predictive accuracy of 0,91. The algorithm is currently implemented in the BITOLA literature-based discovery support system.

■ **Infor Med Slov**: 2008; 13(1): 1-8

Instituciji avtorjev: Inštitut za medicinsko genetiko, Univerzitetni klinični center (AK), Inštitut za biomedicinsko informatiko, Medicinska fakulteta, Univerza v Ljubljani (DH).

Kontaktna oseba: Andrej Kastrin, Inštitut za medicinsko genetiko, Univerzitetni klinični center, Šlajmerjeva 3, 1000 Ljubljana. email: andrej.kastrin@guest.arnes.si.

Uvod

Redno sledenje novim znanstvenim spoznanjem, rezultatom raziskav in razvoju tehnologij je celo za izkušenega raziskovalca z dobro informacijsko podporo zahteven zalogaj. Bibliografske zbirke rastejo iz dneva v dan; na področju biomedicine najpogosteje uporabljena zbirka MEDLINE® trenutno obsega že več kot 16 milijonov zapisov. Za učinkovit priklic relevantnih zapisov zato potrebujemo ustrezno zmogljive iskalnike. Uveljavljenim načinom poizvedovanja po tekstovnih podatkovnih zbirkah se je v zadnjem desetletju pridružilo področje z literaturo podprtega odkrivanja zakonitosti iz podatkov (*angl*. literature-based discovery), katerega namen je iskanje novih in potencialno uporabnih zakonitosti na osnovi implicitnih relacij med posameznimi bibliografskimi zapisi. Oče ideje je Don R. Swanson, ki je na ta način odkril, da ribje olje lahko služi kot učinkovito zdravilo pri zdravljenju Raynaundovega sindroma.[1] Skupaj s sodelavci je ponudil tudi orodje za odkrivanje potencialno zanimivih relacij v obliki prosto dostopne spletne storitve Arrowsmith.[2] Inovativen koncept so povzeli tudi drugi raziskovalci in ga na različne načine implementirali v svojih iskalnih sistemih.

BITOLA je interaktivni sistem za podporo odkrivanju dejanskih ali potencialnih zakonitosti iz bibliografskih podatkov na področju biomedicine in je plod domačega znanja.[3] Relacije med biomedicinskimi koncepti so opisane s pomočjo asociacijskih pravil. Kljub temu, da omogoča posplošeno odkrivanje novega znanja znotraj celotne biomedicinske domene, je zlasti uporaben za opisovanje relacij znanja med posameznimi geni in fenotipom oz. boleznimi.

V trenutni implementaciji sistem tekstovne simbole, ki se potencialno nanašajo na gene, ekstrahira iz naslovov in povzetkov posameznih MEDLINE zapisov. Problem pa predstavlja učinkovito ločevanje med simboli, ki dejansko označujejo gene oz. njihove produkte ter simboli, ki se nanašajo na ostale biomedicinske koncepte.

Ideja je, da najprej razvrstimo zapise na genetsko in negenetsko domeno,[4] pri čemer lahko upravičeno pričakujemo, da se bodo v genetski domeni simboli najverjetneje nanašali na posamezne gene. Z izrazom genetska domena označujemo množico MEDLINE zapisov, v kateri je verjetnost pojavljanja simbolov genov večja kot v katerikoli drugi množici zapisov.

Nalogo lahko opišemo kot problem razvrščanja MEDLINE zapisov na genetsko in negenetsko domeno na osnovi vsebine posameznega zapisa. Formalno imamo torej množico domen $C = \{c_1, c_2\}$ ter množico zapisov $D = \{d_1, d_2, ..., d_n\}$, kjer vsakemu paru $(c_i, d_j; i = 1,2 \text{ in } 1 \leq j \leq n)$ priredimo vrednost indikatorske spremenljivke 0 ali 1, odvisno od tega ali zapis d_j pripada domeni c_i ali ne:

$$F : C \times D \rightarrow \{0,1\} .$$

Funkcijo F, ki priredi vrednost indikatorski spremenljivki, imenujemo klasifikator.[5]

V literaturi najdemo vrsto različnih pristopov k razvrščanju podobnih problemskih domen (npr. metoda podpornih vektorjev, k najbližjih sosedov, naivni Bayesov klasifikator, itd.), med katerimi so ene bolj, druge manj uspešne. Večina do sedaj uporabljenih klasifikatorjev delno ali v celoti temelji na tehnologiji nadzorovanega učenja, kjer nov zapis razvrstimo na podlagi znanja, ki smo ga pridobili na osnovi množice učnih primerov.

V članku predstavimo hiter in preprost pristop k razvrščanju zapisov, ki temelji na metodi statistične primerjave dveh domensko reprezentativnih korpusov. Metoda je jasno razumljiva, preprosta za implementacijo ter računsko nezahtevna. Preizkusi kažejo, da lahko celotno distribucijo MEDLINE bibliografske zbirke s predlagano metodo sprocesiramo v pičlih nekaj urah.

Dosedanje raziskave

Razdvoumljanje večpomenskih besed (*angl*. word sense disambiguation) je proces iskanja ustrezne

interpretacije za pojavitev določene besede v danem kontekstu na osnovi množice možnih interpretacij, ki jih tej besedi lahko pripišemo.[6] Odličen pregled nad tem zelo širokim področjem ponujata Agirre in Edmonds.[7] Biomedicinska domena je z dvoumnimi besedami zelo bogata. Weber s sodelavci[8] npr. poroča, da metatezaver UMLS vsebuje kar 7400 dvoumnih konceptov.

V zadnjih letih je zelo aktualno, tako raziskovalno kot tudi aplikativno, področje razdvoumljanja simbolov genov (*angl.* gene symbol disambiguation).[9] Če za primer vzamemo poved[10] "The inverse association between MR and VEGFR-2 expression in carcinoma suggest a potential tumor-suppressive function for MR", je naloga sistema za razdvoumljanje, da poda oceno ali se simbol MR nanaša na gen za mineralokortikoidni receptor ali na slikanje z magnetno resonanco.

Obravava dvoumnih simbolov genov je z vidika računalniškega procesiranja kompleksen proces. Dosedanje raziskave na tem področju kažejo, da je dvoumnost problematična iz večih razlogov:[11] (i) simbol gena se lahko nanaša na različne organizme, (ii) simbol lahko označuje gen ali nek drug biomedicinski koncept, (iii) simbol lahko označuje gen, protein oz. nek drug biološki produkt ali (iv) pa se nanaša celo na različne gene.

Kljub temu, da se s razdvoumljanjem simbolov genov ukvarja več raziskovalnih skupin, problem avtomatskega in predvsem učinkovitega procesiranja še vedno ostaja odprt. Weber[8] je sestavil obsežen korpus biomedicinskih zapisov namenjen vrednotenju novih algoritmov in metod. Savova in sodelavci[12] so problem dvoumnih simbolov poskusili rešiti z uporabo metode razvrščanja v skupine. Schijvenaars in sodelavci[13] so razvili hiter algoritem razvrščanja, ki zahteva zelo malo učnih podatkov. Humphreyeva[4,14] je problem dvoumnosti rešila z računanjem moči povezanosti med posameznimi besedami in širokopomenskimi deskriptorji MeSH® (Journal Descriptor Indexing). Liu je s sodelavci[15] predstavil dvostopenjsko metodo, kjer za vsak dvoumen simbol najprej zgradimo referenčni reprezentativni korpus, ki ga nato na drugi stopnji uporabimo za gradnjo klasifikatorja. Xu in sodelavci[11] so na osnovi različnih podatkovnih virov vsakemu simbolu gena priredili reprezentativni meta-profil, s čimer so dosegli boljšo diskriminativnost simbolov med seboj. Nedavno je Farkas[16] objavil zanimivo študijo, v kateri se je problema dvoumnosti lotil z analizo omrežij citiranosti posameznih avtorjev.

V članku najprej predstavimo teoretični okvir in gradnjo predlaganega klasifikatorja. Nato predstavimo rezultate eksperimentalnega dela. Končamo s kratko razpravo in idejami za nadaljnje delo.

Metoda

Gradnja reprezentativnih korpusov

Vsak MEDLINE zapis je označen z množico deskriptorjev MeSH, s katerimi opišemo vsebino posameznega zapisa. Iz obsežnega geslovnika MeSH, ki v letu 2008 obsega 24.767 deskriptorjev, izkušeni ocenjevalci vsakemu zapisu v povprečju priredijo 12 MeSH deskriptorjev.

Algoritem poleg korpusa genetske domene zahteva tudi podatke referenčnega korpusa. Referenčni korpus smo zgradili na osnovi polne distribucije bibliografske zbirke MEDLINE (MEDLINE Baseline Repository), ki je do konca leta 2007 vsebovala 16.880.015 zapisov. Distribucija je na voljo v XML zapisu, zato smo zaradi lažje obdelave podatkov s pomočjo razčlenjevalnika prebrali relevantne elemente ter jih prepisali v relacijski podatkovni model (t.j. ena vrstica za vsak deskriptor MeSH v vsakem MEDLINE zapisu). S preštevanjem smo potem izdelali frekvenčno porazdelitev deskriptorjev v referenčnem korpusu.

Korpus genetske domene smo pripravili na osnovi referenčnega korpusa, pri čemer smo upoštevali le tiste zapise, ki so reprezentativni za področje genetike. Pri tem smo si pomagali z datoteko "gene2pubmed" z Entrezovega[17] spletnega

skladišča, ki obsega seznam MEDLINE indentifikatorjev zapisov v katerih se pojavljajo simboli genov. S preštevanjem smo potem izdelali frekvenčno porazdelitev deskriptorjev v korpusu genetske domene.

Algoritem razvrščanja

Vsakemu deskriptorju MeSH smo priredili vrednost hi-kvadrat statistike ter testirali domnevo o statistično značilni povezanosti opazovanih frekvenc v domeni genetskega in referenčnega korpusa. Hi-kvadrat statistiko smo izračunali po enačbi:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}; E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

kjer je O_i opazovana frekvenca, E_i pričakovana frekvenca, N_i pa skupna frekvenca MeSH deskriptorjev, bodisi v korpusu genetske domene bodisi v referenčnem korpusu. Ničelna hipoteza je bila, da opazovani frekvenci danega deskriptorja v domeni genetskega in referenčnega korpusa nista statistično značilno soodvisni.

V naslednjem koraku smo vsakemu deskriptorju pripisali pozitiven predznak indikatorske spremenljivke (*Ind*), če je bila relativna opazovana frekvenca v korpusu genetske domene večja kot relativna opazovana frekvenca deskriptorja v referenčnem korpusu in obratno, deskriptor je dobil negativnen predznak indikatorske spremenljivke, če je bila relativna frekvenca opazovane frekvence v referenčnem korpusu večja kot v korpusu genetske domene.

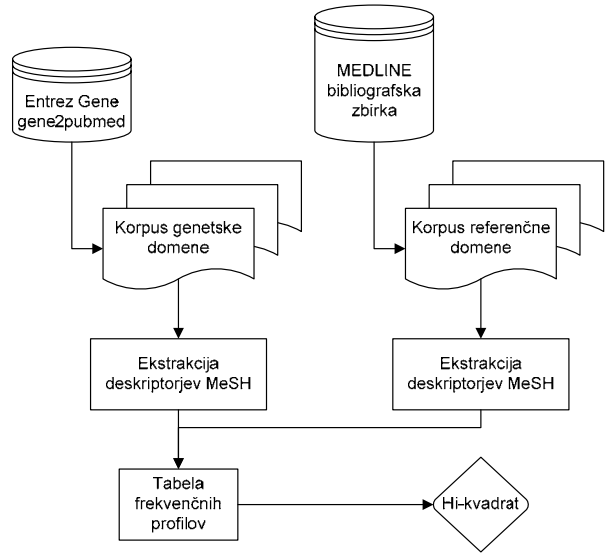
Algoritem za razvrščanja zapisov potrebuje dva vhodna podatka:

1. Tabelo frekvenčnih profilov z vsemi deskriptorji, pri katerih je razlika med korpusoma statistično značilno različna, z označenim predznakom indikatorske spremenljivke.
2. Množico zapisov, ki jih je potrebno razvrstiti.

Deskriptorje, katerih relativna zastopanost je v celotni bibliografski zbirki zelo visoka (npr. Humans, Animals, Mice, itd.) in s tega vidika ne pripomorejo k večji diskriminativnosti algoritma, smo predhodno odstranili iz procesa razvrščanja. Seznam praznih deskriptorjev smo povzeli po uradnem MEDLINE seznamu. Tabela frekvenčnih profilov indeksira frekvence deskriptorjev v korpusu genetske domene ter referenčnem korpusu, hi-kvadrat vrednosti razlik med njima ter vrednost indikatorske spremenljivke. Proces indeksiranja je prikazan na Sliki 1, primer tabele frekvenčnih profilov za množico vzorčnih deskriptorjev pa v Tabeli 1.

Algoritem začne z branjem MEDLINE zapisa *z* in piše vrednost skupnega odločitvenega dosežka po shemi:

Dosežek (z) = 0
za vsak deskriptor MeSH *m*
  če *Ind(m)* pozitiven indikator
    Dosežek (z) = *Dosežek (z)* + 1
  sicer *Ind(m)* negativen indikator
    Dosežek (z) = *Dosežek (z)* − 1



Slika 1. Diagram poteka priprave tabele frekvenčnih profilov na osnovi virov informacij.

Tabela 1. Struktura tabele frekvenčnih profilov.

MeSH Deskriptor	O_B	O_G	E_B	E_G	χ^2	*Ind*
Amyotrophic Lateral Sclerosis	6133	224	6196	161	25	+
Animals	3615845	133463	3654231	95077	15901	+
Computer Simulation	36544	326	35935	935	407	−
Elasticity	16272	48	15906	414	332	−
Genes, Recessive	10948	1269	11907	310	3047	+
Guanine Nucleotide Exchange Factors	1568	685	2196	57	7080	+
Humans	8544765	108109	8433449	219425	57941	−
Mice	691495	77905	749889	19511	179314	+
Models, Chemical	29588	608	29430	766	33	−
Models, Molecular	68781	5425	72324	1882	6845	+
Models, Statistical	19630	60	19191	499	397	−
Mutation	178885	23627	197377	5135	68316	+
Polymers	26724	170	26212	682	394	−
Stress, Mechanical	25706	320	25366	660	180	−
Surface Tension	2976	12	2912	76	55	−

Pojasnilo: O_B – opazovana frekvenca v korpusu referenčne domene; E_G – pričakovana frekvenca v korpusu genetske domene; χ^2 – hi-kvadrat statistika; *Ind* – indikatorska spremenljivka.

Na izhodu dobimo množico zapisov, katerim je pripisana vrednost odločitvenega dosežka. Višji odločitveni dosežek imajo zapisi, pri katerih obstaja večja verjetnost, da pripadajo genetski domeni.

Ocenjevanje natančnosti razvrščanja

Pražno vrednost, ki ločuje med zapisi obeh domen, ocenimo na testni množici zapisov, natančnost razvrstitve pa potem preverimo na neodvisni validacijski množici zapisov. Iz bibliografske zbirke MEDLINE smo najprej naključno izbrali vzorec 200 zapisov; prvo polovico zapisov smo uvrstili v testno, drugo polovico pa v validacijsko množico. Vsak zapis sta ocenila dva kompetentna ocenjevalca in mu pripisala indikator domene, ki je označeval reprezentativnost zapisa, bodisi za genetsko bodisi za negenetsko domeno. Skladnost med ocenjevalcema smo ovrednotili s pomočjo kappa (κ) koeficienta.[18] V testni množici je znašalo ujemanje med ocenjevalcema $\kappa = 0{,}81$, v validacijski množici pa $\kappa = 0{,}91$. Zapise, pri katerih se je ocena razlikovala, sta ocenjevalca ponovno pregledala in končno oceno podala s soglasjem. Obe množici sta prosto dostopni pri avtorjih.

Empirično pražno vrednost odločitvenega dosežka, ki ločuje med zapisi genetske in negenetske domene smo določili z optimizacijo točnosti razvrstitve (*angl.* accuracy). Točnost (*Acc*) razvrstitve izračunamo kot razmerje med vsoto pravilnih razvrstitev ($TP + TN$) in skupnim številom zapisov, ki jih razvrščamo ($TP + FP + FN + TN$). *TP* (*angl.* true positive) označuje število pravilno razvrščenih pozitivnih zapisov (zapisi poročajo o genetiki in so razvrščeni v domeno genetike); *FP* (*angl.* false positive) označuje število napačno razvrščenih negativnih zapisov (zapisi so razvrščeni v domeno genetike, dejansko pa poročajo o nečem drugem); FN (*angl.* false negative) se nanaša na število napačno razvrščenih pozitivnih zapisov (zapisi so napačno razvrščeni v domeno, ki ni reprezentativna za genetiko), *TN* (*angl.* true negative) pa označuje število pravilno razvrščenih negativnih zapisov (zapisi so pravilno razvrščeni v domeno, ki ni reprezentativna za genetiko). Pražno vrednost lahko določimo tudi poljubno in s tem spreminjamo razmerje med napačno pozitivnimi in napačno negativnimi zapisi. Poleg natančnosti razvrstitve smo izračunali še priklic (*angl.* recall), natančnost (*angl.* precision) ter povprečno mero priklica in natančnosti *F*. Priklic predstavlja razmerje med pravilno razvrščenimi zapisi in

množico vseh zapisov v testni domeni, ki pripadajo domeni genetike. Natančnost pa predstavlja razmerje med pravilno razvrščenimi zapisi in vsemi zapisi, ki jih je algoritem razvrstil v domeno genetike. Priklic (*Rec*) in natančnost (*Pre*) izračunamo po naslednjih enačbah:

$$Rec = \frac{TP}{TP + FN} \text{ in } Pre = \frac{TP}{TP + FP} .$$

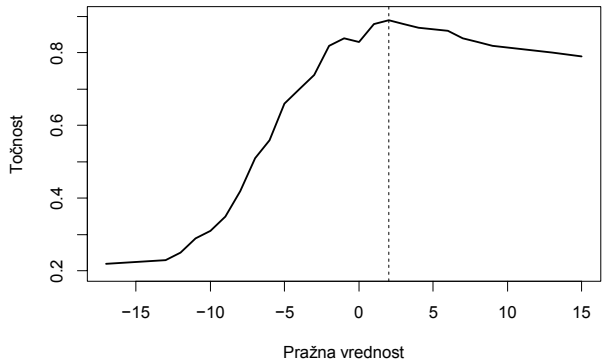
F statistiko izračunamo kot harmonično sredino priklica in natančnosti po enačbi:

$$F = \frac{2 \times Rec \times Pre}{Rec + Pre}.$$

Razpon vseh mer je v intervalu [0,1], pri čemer 1 označuje maksimalno kvaliteto klasifikatorja.

Rezultati

Na Sliki 2 je prikazana točnost razvrstitve v odvisnosti od pražne vrednosti oz. porazdelitve odločitvenega dosežka. V našem primeru smo kot pražno vrednost definirali $\theta = 2$, pri kateri je bila točnost razvrstitve maksimalna. Pražne vrednosti med poskusi nismo več spreminjali. Vse zapise, pri katerih je bila vrednost odločitvenega dosežka $\theta \geq 2$, smo uvrstili v genetsko domeno.



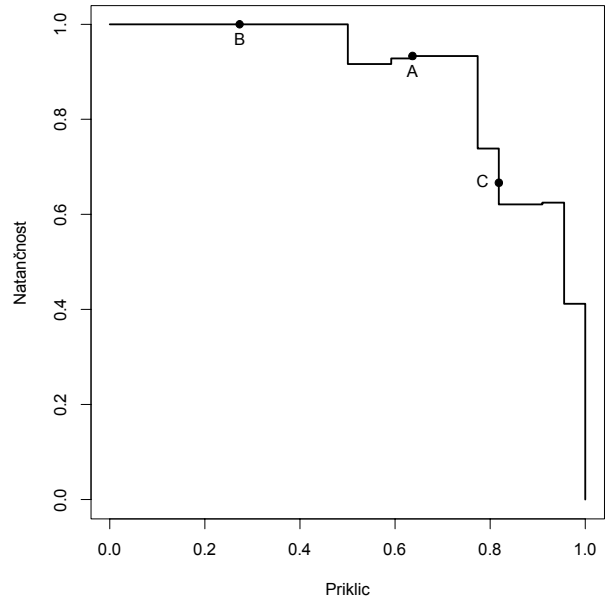
Slika 2. Točnost razvrstitve zapisov v odvisnosti od pražne vrednosti na testni množici zapisov.

Proces razvrščanja si oglejmo na primeru. Izbrali smo dva MEDLINE zapisa, katerih naslova sta "Strain-dependent localization, microscopic deformations, and macroscopic normal tensions in model polymer networks" (*PMID*: 15697942) in "Recessive motor neuron diseases: mutations in the ALS2 gene and molecular pathogenesis for the upper motor neurodegeneration" (*PMID*: 15651293). Prvi zapis vsebuje osem, drugi pa sedem deskriptorjev MeSH. Trije deskriptorji v drugem zapisu so prazni (Aminals, Humans in Mice), zato jih izločimo iz nadaljnjega procesiranja. Glede na vsak deskriptor povečamo oz. zmanjšamo vrednost odločitvenega dosežka skladno z vrednostjo indikatorske spremenljivke (Tabela 1). Končni rezultat sta dva odločitvena dosežka:

Dosežek (*PMID*: 15697942) $= -6$
Dosežek (*PMID*: 15651293) $= 4$.

Glede na zgoraj definirano pražno vrednost lahko prvi zapis uvrstimo v negenetsko domeno, drugi zapis pa v genetsko domeno.

Učinkovitost klasifikatorja smo preverili na validacijski množici zapisov. Na Sliki 3 so prikazane vrednosti priklica in natančnosti za različne pražne vrednosti parametra θ.



Slika 3. Graf odnosa med priklicem in natančnostjo razvrstitve na validacijski množici zapisov.

Povedena metoda je dosegla $Acc = 0{,}91$ točnost razvrstitve, pri čemer je priklic znašal $Rec = 0{,}64$, natančnost pa $Pre = 0{,}93$. Harmonična sredina priklica in natančnosti je znašala $F = 0{,}76$.

Graf poteka skozi dve točki. Točka (0,1) označuje klasifikator, ki ne prepozna genetsko relevantnih zapisov. Vsi negenetski zapisi so v tem primeru razvrščeni pravilno, genetski zapisi pa napačno. V točki (1,0) klasifikator vse zapise razvrsti kot genetsko relevantne. Genetski zapisi so v tem primeru razvrščeni pravilno, vsi negenetski zapisi pa napačno. Optimalno točnost oz. optimalno razmerje med priklicem in natančnostjo je v našem primeru klasifikator dosegel v točki A. Če pražno vrednost povečamo ($\theta_B = 5$), bomo zapise razvrstili z večjo natančnostjo ($Pre = 1{,}00$), na račun katere pa se bo zmanjšal priklic ($Rec = 0{,}27$). Do obratnega učinka pride v točki C, kjer pražno vrednost zmanjšamo ($\theta_C = -1$). Dokumenti so v tem primeru razvrščeni z manjšo natančnostjo ($Pre = 0.67$), priklic pa je višji ($Rec = 0.82$).

Zaključki

V članku smo predstavili preprost in hiter algoritem za razvrščanje MEDLINE zapisov na osnovi deskriptorjev MeSH. Trenutno je algoritem implementiran kot pomožni sitem za razdvoumljanje simbolov genov v interaktivnem sistemu za podporo odkrivanju zakonitosti iz bibliografskih podatkov BITOLA. Eksperimentalni rezultati potrjujejo razmeroma visoko napovedno točnost klasifikatorja ($Acc = 0{,}91$). V odvisnosti od namena poizvedovanja ter željenega razmerja med priklicem in natančnostjo, lahko pražno vrednost tudi spreminjamo.

V nadaljnjih raziskavah bomo preizkusili klasifikatorje z vključitvijo kompleksnejših prediktorskih spremenljivk: (i) kvalifikatorjev MeSH, ki podrobneje omejijo vsebinski obseg posameznega deskriptorja, (ii) parov deskriptor/kvalifikator ter (iii) prostim besedilom naslova in povzetka MEDLINE zapisa. V gradnji je tudi obsežnejši označeni korpus zapisov, ki bo omogočal bolj zanesljivo in veljavno vrednotenje uporabljenih klasifikatorjev.

Zahvala

Avtorja se zahvaljujeva Susanne M. Humphrey in Thomasu C. Rindfleschu za koristne napotke in vzpodbudne komentarje. Raziskovalna sredstva je zagotovila Javna agencija Republike Slovenije za raziskovalno dejavnost (J3-7411).

Literatura

1. Swanson DR: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986; 30(1): 7-18.
2. Swanson DR , Smalheiser NR: An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery. *Artif Intell* 1997; 91(2): 183-203.
3. Hristovski D, Stare J, Peterlin B, et al.: Supporting discovery in medicine by association rule mining in MEDLINE and UMLS. Medinfo 2001; 10(2): 1344-1348.
4. Hristovski D, Peterlin B, Mitchell JA, et al.: Using literature-based discovery to identify disease candidate genes. *Int J Med Inform* 2005; 74(2-4): 289-298.
5. Feldman R, Sanger J: *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge 2006: Cambridge University Press.
6. Manning CD, Schuetze H: *Foundations of statistical natural language processing*. Cambridge 2003: MIT Press.
7. Agiree E, Edmonds P (ur.): *Word sense disambiguation: Algorithms and applications*. Berlin 2006: Springeer.
8. Weeber M, Schijvenaars BJ, Van Mulligen EM, et al.: Ambiguity of human gene symbols in LocusLink and MEDLINE: creating an inventory and a disambiguation test collection. *Proc AMIA Symp* 2003; 704-708.
9. Chen L, Liu H, Friedman C: Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics* 2005; 21(2): 248-256.
10. Di Fabio F, Alvarado C, Majdan A, et al.: Underexpression of mineralocorticoid receptor in colorectal carcinomas and association with

VEGFR-2 overexpression. *J Gastrointest Surg* 2007; 11(11): 1521-1528.
11. Xu H, Fan JW, Hripcsak G, et al.: Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics* 2007; 23(8): 1015-1022.
12. Savova G, Pedersen T, Purandare A, et al.: Resolving ambiguities in biomedical text with unsupervised clustering approaches. Research Report UMSI 2005/80 and CB Number 2005/21; Minneapolis, Minnesota 2005: University of Minessota Supercomputing Institute.
13. Schijvenaars BJ, Mons B, Weeber M, et al.: Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics* 2005; 6: 149.
14. Humphrey SM, Rogers WJ, Kilicoglu H, et al.: Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: preliminary experiment. *J Am Soc Inform Sci Tech* 2006; 57(1): 96-113.
15. Liu H, Lussier YA, Friedman C: Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *J Biomed Inform* 2001; 34(4): 249-261.
16. Farkas R: The strength of co-authorship in gene name disambiguation. *BMC Bioinformatics* 2008; 9: 69.
17. Maglott D, Ostell J, Pruitt KD, et al.: Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2007; 35(Database issue): D26-31.
18. Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20(1): 37-46.