VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Vaidas BALYS

MOKSLINĖS TERMINIJOS MATEMATINIAI MODELIAI IR JŲ TAIKYMAS LEIDINIŲ KLASIFIKAVIME

DAKTARO DISERTACIJA

FIZINIAI MOKSLAI,
MATEMATIKA (01P)

VGTU
LEIDYKLA
Vilnius TECHNIKA 2009

Disertacija rengta 2004–2009 metais Matematikos ir informatikos institute.

Mokslinis vadovas

prof. habil. dr. Rimantas RUDZKIS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).

http://leidykla.vgtu.lt
VGTU leidyklos TECHNIKA 1650-M mokslo literatūros knyga

ISBN 978-9955-28-467-3

vbalys@gmail.com

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Vaidas BALYS

MATHEMATICAL MODELS FOR SCIENTIFIC TERMINOLOGY AND THEIR APPLICATIONS IN THE CLASSIFICATION OF PUBLICATIONS

DOCTORAL DISSERTATION

PHYSICAL SCIENCES,
MATHEMATICS (01P)

VGTU
LEIDYKLA
Vilnius TECHNIKA 2009

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2004–2009.

Scientific Supervisor

Prof Dr Habil Rimantas RUDZKIS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).

Reziumė

Disertacijoje nagrinėjamas mokslo publikacijų automatinio klasifikavimo uždavinys. Šis uždavinys sprendžiamas taikant tikimybinius diskriminantinės analizės metodus. Pagrindinis darbo tikslas – sukurti konstruktyvius klasifikavimo metodus, kurie leistų atsižvelgti į mokslo publikacijų tekstų specifiką.

Disertaciją sudaro įvadas, trys pagrindiniai skyriai, rezultatų apibendrinimas, naudotos literatūros ir autoriaus publikacijų disertacijos tema sąrašai ir vienas priedas.

Įvadiniame skyriuje aptariama tiriamoji problema, darbo aktualumas, aprašomas tyrimų objektas, formuluojamas pagrindinis darbo tikslas bei uždaviniai, aprašoma tyrimų metodika, darbo mokslinis naujumas, pasiektų rezultatų praktinė reikšmė, ginamieji teiginiai. Įvado pabaigoje pristatomos disertacijos tema autoriaus paskelbtos publikacijos ir pranešimai konferencijose bei disertacijos struktūra.

Pirmajame skyriuje matematiškai apibrėžtas ir detalizuotas sprendžiamas uždavinys, pateikta analitinė kitų autorių darbų apžvalga. Pasirinkti ir išanalizuoti keli populiarūs klasifikavimo algoritmai, kurie eksperimentinėje darbo dalyje lyginti su autoriaus pasiūlytaisiais.

Antrajame skyriuje sudarytas mokslo terminijos pasiskirstymo tekstuose tikimybinis modelis, išskirti atskiri atvejai, galiojant įvestoms prielaidoms apie terminų tarpusavio sąryšių formas, pasiūlytos modelio identifikavimo procedūros bei suformuluoti konstruktyvūs mokslo publikacijų klasifikavimo algoritmai.

Trečiajame skyriuje pateikti pagrindiniai rezultatai eksperimentinio tyrimo, kurio metu realių duomenų pagrindu tirti pasiūlytieji klasifikavimo metodai bei atlikta palyginamoji pasiūlytųjų bei keleto populiariausių kitų autorių algoritmų analizė.

Disertacijos tema paskelbti 7 straipsniai: vienas – ISI Web of Science duomenų bazėje referuotoje konferencijos medžiagoje, keturi – recenzuojamuose periodiniuose žurnaluose, 2 – recenzuojamose tarptautinių konferencijų medžiagose. Disertacijos tema perskaityta 11 pranešimų Lietuvos bei kitų šalių konferencijose.

Abstract

The dissertation considers the problem of automatic classification of scientific publications. The problem is addressed by using probabilistic methods of the discriminant analysis. The main goal of the dissertation is to create constructive classification methods that would allow to take into consideration specificity of scientific publication text.

The dissertation consists of Introduction, 3 chapters, Conclusions, References, list of author's publications, and one Appendix.

The introduction reveals the investigated problem, importance of the thesis and the object of research and describes the purpose and tasks of the paper, research methodology, scientific novelty, the practical significance of results examined in the paper and defended statements. The introduction ends in presenting the author's publications on the subject of the defended dissertation, offering the material of made presentations in conferences and defining the structure of the dissertation.

Chapter 1 presents a detailed mathematical formulation of the considered problem, reviews scientific papers on the subject, and analyses a few popular classification algorithms that in Chapter 3 are compared to the ones proposed in this paper.

Chapter 2 develops the probabilistic model for scientific terminology distribution over texts, discusses special cases of the model under specific assumptions on forms of terminology relations, suggests the model identification procedures, and formulates constructive scientific publication classification algorithms.

Chapter 3 reports the results of the real data based experimental evaluation and comparison of the methods formulated in the dissertation and the alternative popular methods chosen in Chapter 1.

7 articles focusing on the subject discussed in the dissertation are published: one article – in the conference material quoted by ISI Web of Science data base, four articles – in the reviewed journals, two articles – in material reviewed during international conferences. 11 presentations on the subject have been given in conferences at national and international level.

Žymėjimai

Simboliai

$\|A\|$	aibės A elementų skaičius;
$dim(x)$	vektoriaus x dimensija;
$rank(A)$	matricos A rangas;
O, X	atsitiktinis stebimas objektas ir jo požymių vektorius;
K_j, q	objektų klasės ir klasių skaičius;
H_j, p_j	atsitiktinis įvykis, kad stebėtas objektas priklauso klasei K_j, ir šio įvykio tikimybė;
$\eta(X)$	objekto su požymių vektoriumi X klasė (atsitiktinis dydis);
$\pi(j, x)$	tikimybė, kad objektas su požymių vektoriumi $X = x$ priklauso klasei K_j;
X_n, Y_m	mokymo ir testavimo imtis;
$a = (a_1, \ldots, a_d)$	atsitiktinio straipsnio projekcija (terminų vektorius);
A	generalinė straipsnių projekcijų aibė;
λ_i	papildomos informacijos, susijusios su i–ojo termino iš straipsnio projekcijos pozicija tekste, vektorius;

V, h	terminų žodynas ir jo dydis;
v, u	terminai iš V;
w	straipsnio klasė;
α	pasirinktas reikšmingumo lygis;
μ	minimalus reikalaujamas termino ar terminų poros stebėjimų skaičius mokymo imtyje;
$\sigma(\cdot)$, $\delta(\cdot,\cdot)$	terminų svorių funkcionalai;
Pr, Re, F_1, Pr_{avg}	algoritmų tikslumo matai.

Santrumpos

a.d.	atsitiktinis dydis;
a.v.	atsitiktinis vektorius;
BIM	binarinis nepriklausomumo modelis (*angl. Binary Independence Model*);
DF	požymių atrinkimo metodas (*angl. Document Frequency*);
IDC, IDCm	siūlomi identifikacinių debesėlių (*angl. Identification Clouds*) klasifikavimo algoritmai;
kNN	k kaimynų klasifikavimo algoritmas (*angl. k Nearest Neighbours*);
LLSF	tiesinis mažiausių kvadratų klasifikavimo algoritmas (*angl. Linear Least Squares Fit*);
MKM	mažiausiųjų kvadratų metodas;
MSC	matematikos temų klasifikacija (*angl. Mathematical Subject Classification*);
nB	naivaus Bajeso klasifikavimo algoritmas (*angl. naive Bayes*);
SVD	matricos skaidymas singuliariomis reikšmėmis (*angl. Singular Value Decomposition*);
SVM	atraminių vektorių klasifikavimo algoritmas (*angl. Support Vector Machines*).

Turinys

Padėka

Nuoširdžiai dėkoju darbo vadovui prof. habil. dr. Rimantui Rudzkiui už vadovavimą disertaciniam darbui, skirtą laiką ir energiją; recenzentams doc. dr. Marijui Radavičiui ir prof. habil. dr. Kęstučiui Kubiliui už pastabas, padėjusias patobulinti disertaciją; UAB „VTEX“ ir ypač jos vadovui dr. Rimui Maliukevičiui už neįkainojamą pagalbą gaunant duomenis tyrimams bei už sudarytas galimybes derinti darbą ir mokslą; visam Matematikos ir informatikos instituto Taikomosios statistikos skyriaus kolektyvui už moralinį palaikymą. Dėkoju Lietuvos valstybiniam mokslo ir studijų fondui už finansinę paramą doktorantūros studijoms. Ačiū visiems artimiesiems ir draugams už visokeriopą pagalbą, paramą ir supratimą.

Įvadas

Tiriamoji problema

Publikacijų klasifikavimas yra viena iš svarbių mokslo tekstų tvarkymo veiklų, sudarančių galimybes kaupti, ir, kas svarbiausia, esant poreikiui surasti bei panaudoti mokslinės informacijos ir žinių fragmentus. Rankinis šio darbo atlikimas ne tik neefektyvus ir sudėtingas, bet ir neprasmingas dabartinių techninių galimybių kontekste. Todėl šioje disertacijoje nagrinėjama specifinio mokslinio turinio automatinio klasifikavimo problema.

Darbo aktualumas

Per visą matematikos mokslo istoriją išleistų mokslinių darbų apimtys skaičiuojamos dešimtimis milijonų puslapių (Ewing, 2002), ir kiekvieną dieną šis kiekis dar padidėja. Šiuose darbuose sukauptos per daugelį šimtmečių įgytos žinios, kurių efektyvus valdymas yra ne šiaip svarbus uždavinys, bet viena pagrindinių mokslo kaip prasmingos veiklos sąlygų. Žinių valdymas, kitaip – žinių vadyba, paprastai suprantamas kaip rinkinys veiklų, kurios leidžia atpažinti, tvarkyti, kurti bei platinti žinias tam, kad jas būtų galima atkartoti, mokyti ir įsisavinti. Šios veiklos viena ar kita forma visada egzistavo mokslo pasauly-

je, tačiau paskutinių dešimtmečių pokyčiai kompiuterijos bei komunikacijų srityse atvėrė naujas galimybes, o tai sąlygojo naujų reikalavimų ir tikslų iškėlimą. Akademinėje visuomenėje formuojasi grupės, kurias jungia siekis pasinaudoti pakitusiu technologiniu kontekstu ir sukurti tobulesnes priemones bei įrankius aktualios informacijos bei žinių valdymui. Nuo 2001 metų reguliariai vykstančios *Matematinių žinių valdymo* (*Mathematical Knowledge Management, MKM*) konferencijos suburia tyrėjus, kurie domisi su matematiniu turiniu susijusiais uždaviniais. Šios bendruomenės internetinėje svetainėje `<http://www.mkm-ig.org>` pabrėžiama, kad toks matematinės informacijos išskyrimas ypač svarbus ir dėl to, kad ji dėl savo prigimties yra tinkamiausias pretendentas išbandyti inovatyviausius teorinius ir technologinius sprendimus.

Nors mokslo žinių valdymas yra labai plati sąvoka, tačiau galima išskirti kelis pagrindinius uždavinius, aktualius kiekvienam vienaip ar kitaip susiduriančiam su moksliniu turiniu: patogus ir patrauklus informacijos ir duomenų pateikimas, lanksti ir intuityvi paieška, logiškos ir aktyvios sąsajos tarp elementų ir kt. Šie uždaviniai itin svarbūs ir vienam pagrindinių mokslo pasaulio veikėjų – akademinės literatūros leidėjams. Tebevykstantis perėjimas nuo tradicinės popierinės leidybos prie elektroninės, augančios galimybės mokslinį turinį ar jo dalį gauti laisvai (nors dažniausiai ir ne paskutinės versijos) bei dėl tobulėjančių ir populiarėjančių mokslo tekstų redagavimo ir publikavimo priemonių didėjanti konkurencija įtakoja kintančius leidėjų prioritetus. Vis daugiau dėmesio skiriama naujų paslaugų naudotojams, tokių kaip automatinio recenzentų parinkimo, klasifikavimo, išsamios semantinės paieškos, kūrimui, tuo pat metu iš dalies nuskriaudžiant tradicinių veiklų (rinkimo, maketavimo, redagavimo) sektorių. Nepaisant įvardintų uždavinių aktualumo, tyrimų, atsižvelgiančių į mokslinės informacijos specifiką, trūksta. Tai iš dalies liudija ir aktyvus susidomėjimas šio darbo teoriniais ir praktiniai rezultatais, kurį išreiškė mokslo leidyklų *Elsevier B.V.* bei *Springer* atstovai.

Mokslo žinių valdymo problemų rate automatinis tekstų, ypač – mokslinių, klasifikavimas yra vienas aktualiausių uždavinių. Dabar tai, ką paprastai atlikdavo pats teksto autorius, pvz., nurodydamas raktinius žodžius ar išvardindamas mokslo sritį apibūdinančius standartinius klasifikatorių kodus, gali atlikti arba bent jau gali padėti atlikti automatinė sistema. Rezultatai nebeprivalo būti apriboti ir fiksuoti – jie gali kisti, priklausomai nuo konkrečių sistemos poreikių. Jei klasifikavimo algoritmai remiasi tam tikrų teksto elementų ryšių analize, jos rezultatus galima naudoti sprendžiant kitus mokslo tekstų apdorojimo uždavinius. O patys sąryšiai tarp mokslo srities elementų, pavyzdžiui, jos terminijos, savo ruožtu yra vertingas rezultatas, suteikiantis žinių apie nagrinėjamąją sritį.

Automatinis tekstų klasifikavimas, kaip ir didžioji dalis taikomojo pobūdžio uždavinių, apjungia bei skolinasi idėjas, metodus ir tyrėjus iš daugelio skirtingų mokslo krypčių bei sričių, tokių kaip tikimybių teorija ir matematinė statistika, statistinė duomenų analizė, dirbtinis intelektas, automatinis mokymasis (*machine learning*), duomenų gavyba (*data mining*), informacijos ištraukimas (*information retrieval*), kalbos apdorojimas (*natural language processing*) ir kt. Tačiau, nors bendras su tekstų klasifikavimu susijusių atliktų tyrimų skaičius yra labai didelis, darbų, nagrinėjančių būtent mokslo tekstams pritaikytus metodus, beveik nėra. Tiesiogiai taikant įprastų tekstų (elektroninio pašto žinutės, naujienų pranešimai) klasifikavimo algoritmus neatsižvelgiama į mokslo publikacijų specifiką: ilgą ir nehomogenišką tekstą, griežtą struktūrą, specifinę kalbą ir kt. Todėl iškyla natūralus poreikis sukurti matematinius metodus, kurie leistų į šią specifiką atsižvelgti ir ją panaudoti konstruojant tikslesnius ir geriau pritaikytus klasifikavimo algoritmus.

Tyrimų objektas

Darbo tyrimų objektas yra daugiamatės diskriminantinės analizės metodai.

Darbo tikslai

Pagrindinis darbo tikslas yra pasiūlyti ir ištirti matematinius klasifikavimo metodus, paremtus statistine mokslo terminijos pasiskirstymo tekstuose analize, kuriuos būtų galima taikyti taikomajam mokslo publikacijų klasifikavimo uždaviniui spręsti.

Darbo uždaviniai

Pagrindiniam darbo tikslui pasiekti suformuluoti šie uždaviniai:

1. Sudaryti mokslo terminijos pasiskirstymo publikacijų tekste tikimybinį modelį ir sukurti modelio identifikavimo procedūras;
2. Sukurti šiuo modeliu ir jo identifikavimo procedūromis pagrįstus konstruktyvius klasifikavimo algoritmus;
3. Pasiūlyti matematinius metodus, kaip klasifikavimo algoritmuose atsižvelgti į papildomą informaciją, susijusią su mokslo terminų pozicijomis

ir kontekstu tarp jų;

4. Realių duomenų pagrindu ištirti pasiūlytus sprendimus, atlikti sukurtų bei alternatyvių klasifikavimo algoritmų palyginamąją analizę.

Tyrimų metodai

Darbe taikomi šie tyrimų metodai: mokslinės literatūros disertacijos tema analizė, matematinis modeliavimas (tikimybinio modelio sudarymas bei jo identifikavimo procedūrų formulavimas) ir eksperimentas (pasiūlytųjų sprendimų analizė ir palyginimas su alternatyviais metodais, atliekant bandymus su realių publikacijų duomenų baze).

Darbo mokslinis naujumas ir jo reikšmė

Atlikto darbo rezultatai papildo ir praplečia kitų šioje bei giminiškose srityse atliktų tyrimų rezultatus. Nuo kitų autorių darbų skiriasi šiais nagrinėjamais klausimais, siūlomais sprendimais bei pasiektais rezultatais:

- nagrinėtas mokslo publikacijų tekstų klasifikavimo uždavinys, siūlomuose metoduose atsižvelgta į šių tekstų specifiką;
- sudarytas tikimybinis mokslo terminijos pasiskirstymo tekste modelis, sukurtos jo identifikavimo procedūromis, suformuluoti originalūs klasifikavimo algoritmai;
- pasiūlytas informatyviausių mokslo terminų nustatymo būdas, paremtas statistinių hipotezių tikrinimo teorijos metodų taikymu;
- sukurti papildomos informacijos, susijusios su terminų pozicijomis tekste bei kontekstu tarp jų, naudojimo klasifikavime metodai;
- atlikta išsami palyginamoji pasiūlytųjų bei keleto populiarių kitų autorių klasifikavimo metodų analizė realios mokslo publikacijų bazės pagrindu;
- tos pačios bazės pagrindu ištirti šie su mokslo publikacijų specifika susiję aspektai: atskirų teksto dalių naudojimo, mokslo terminijos žodyno parinkimo, pačios klasifikavimo sistemos (klasių rinkinio) parinkimo, ilgų tekstų naudojimo įtaka klasifikavimo tikslumui.

Darbo rezultatų praktinė reikšmė

Darbe suformuluoti konstruktyvūs mokslo publikacijų klasifikavimo algoritmai gali būti tiesiogiai realizuoti mokslo informaciją kaupiančiose ir pateikiančiose automatizuotose sistemose, o tai leistų naudotojams pasiūlyti patogias ir lanksčias duomenų paieškos ir navigacijos priemones. Siūloma mokslo terminijos pasiskirstymo statistine analize paremta klasifikavimo metodika gali būti pritaikyta sprendžiant ir kitus mokslo tekstų apdorojimo uždavinius. O suformuluoto modelio parametrų sąryšių išreikštinumas ir aiški interpretacija sudaro praktines galimybes į metodus įtraukti ekspertines žinias.

Ginamieji teiginiai

1. Pasiūlyta klasifikavimo metodika, pagrįsta statistine mokslo terminijos pasiskirstymo publikacijų tekstuose analize.
2. Sukurtos pasiūlyto terminijos pasiskirstymo modelio identifikavimo procedūros.
3. Pasiūlyti papildomos kontekstinės informacijos panaudojimo klasifikavime metodai.

Darbo rezultatų aprobavimas

Disertacijos rezultatai paskelbti 7 recenzuojamuose leidiniuose. Publikacijų sąrašas pateiktas disertacinio darbo pabaigoje.

Disertacinio darbo tematika skaityta 11 pranešimų Lietuvos ir tarptautinėse mokslinėse konferencijose:

- Tarptautinėje konferencijoje „Nordic Conference on Mathematical Statistics“ 2008 m. Vilniuje;
- Tarptautinėje konferencijoje „International Conference on Mathematical Knowledge Management“ 2006 m. Wokingham;
- Tarptautinėje konferencijoje „International Vilnius Conference on Probability Theory and Mathematical Statistics“ 2006 m. Vilniuje;
- Tarptautinėje konferencijoje „International Conference on Computer Data Analysis and Modeling“ 2004 ir 2007 m. Minske;

- Lietuvos matematikų draugijos konferencijoje 2003, 2005, 2007 m. Vilniuje ir 2004, 2006, 2008 m. Kaune.

Taip pat skaityti pranešimai Matematikos ir informatikos instituto Taikomosios statistikos skyriaus, Tikimybių teorijos ir statistikos skyriaus bei Vilniaus Gedimino technikos universiteto Matematinės statistikos katedros seminaruose.

Disertacijos struktūra

Disertaciją sudaro įvadas, trys pagrindiniai skyriai, išvados, naudotos literatūros sąrašas, autoriaus publikacijų disertacijos tema sąrašas ir vienas priedas.

Darbo apimtis yra 101 puslapis, tekste panaudotos 99 numeruotos formulės, 15 paveikslų ir 13 lentelių. Rašant disertaciją buvo pasinaudota 86 literatūros šaltiniais.

1

Tekstų klasifikavimo metodų analitinė apžvalga

1.1. Klasifikavimo teorijos elementai

1.1.1. Bazinės sąvokos

Tarkime, turime tiriamų objektų visumą $K = \{O_i, i = 1, 2, \ldots\}$, o kiekvienas objektas priklauso vienai ir tik vienai iš q klasių $K_i, i = \overline{1, q}$. Tada

$$K = K_1 \cup K_2 \cup \ldots \cup K_q. \tag{1.1}$$

Objekto O_i stebėjimo metu fiksuojamas skaitinių požymių vektorius $X(i)$.

Tarkime, stebime atsitiktinai parinkto K objekto O požymių vektorių X, $d = \dim(X)$. Apibrėžkime atsitiktinius įvykius

$$H_j = \{O \in K_j\}, \quad j = \overline{1, q} \tag{1.2}$$

ir apriorines šių įvykių tikimybes

$$p_j = \mathbb{P}\{H_j\}, \quad \text{čia } p_j > 0 \text{ ir } \sum_{j=1}^{q} p_j = 1.$$

Pažymėkime $F_j(x) = \mathbb{P}\{X < x \big| H_j\}$ sąlyginę atsitiktinio vektoriaus (a.v.) X pasiskirstymo funkciją prie sąlygos, kad stebėtas objektas priklauso klasei K_j. Tada besąlyginė X pasiskirstymo funkcija $F(x) = \mathbb{P}\{X < x\}$ apibrėžiama lygybe

$$F(x) = \sum_{j=1}^{q} p_j F_j(x). \tag{1.3}$$

Tegul stebėtas objektas O priklauso nežinomai (nestebimai) klasei η. *Bendras klasifikavimo uždavinys – nustatyti η pagal stebėtą požymių vektorių X.*

Klasifikavimas vadinamas griežtu, kai stebinys pagal tam tikrą taisyklę priskiriamas vienai klasei $\widehat{\eta}(X) \in \{1, \ldots, q\}$ ir negriežtu, kai vertinamos stebėto objekto priklausymo kiekvienai iš klasių tikimybės, t. y., skaičiuojami aposteriorinių tikimybių

$$\pi(j, x) = \mathbb{P}\{H_j \big| X = x\}, \quad j = \overline{1, q} \tag{1.4}$$

įverčiai. Iš negriežto klasifikavimo taisyklės visada galime gauti griežto klasifikavimo taisyklę:

$$\widehat{\eta}(x) = \arg\max_{j=\overline{1,q}} \widehat{\pi}(j, x). \tag{1.5}$$

Tegul $\phi_j(\cdot)$ yra sąlyginių pasiskirstymo funkcijų F_j pasiskirstymo tankiai. Jiems galioja (1.3) lygybei analogiška lygybė

$$\phi(x) = \sum_{j=1}^{q} p_j \phi_j(x), \tag{1.6}$$

čia $\phi(x)$ žymi besąlyginę a.v. X pasiskirstymo tankio funkciją.

Tada

$$\pi(j, x) = \frac{p_j \phi_j(x)}{\phi(x)}. \tag{1.7}$$

1.1.2. Diskriminantinės analizės metodai

Šiame darbe nagrinėjami diskriminantinės analizės metodai, kurie taikomi tuo atveju, kai stebimo objekto požymių vektoriaus X sąlyginės pasiskirstymo funkcijos F_j žinomos tik iš dalies, tačiau turima mokymo imtis

$$X_n = \{(X(1), \eta(1)), (X(2), \eta(2)), \ldots, (X(n), \eta(n))\}, \tag{1.8}$$

čia $\eta(j)$ žymi j–ojo imties objekto su požymių vektoriumi $X(j)$ tikrąjį klasės numerį.

Tegul sąlyginės pasiskirstymo funkcijos F_j turi žinomo pavidalo pasiskirstymo tankius $\phi_j(\cdot) = \phi_j(\cdot, \theta_j)$ su nežinomais bendru atveju daugiamačiais parametrais θ_j. Pažymėkime $\theta = (\theta_1, \ldots, \theta_q)$. Tada iš mokymo imties X_n gautą statistinį įvertį $\widehat{\theta}$ galime įstatyti į (1.7) vietoje nežinomo tikrojo θ:

$$\widehat{\pi}(j, x) = \frac{\widehat{p}_j \phi_j(x, \widehat{\theta}_j)}{\sum_{k=1}^{q} \widehat{p}_k \phi_k(x, \widehat{\theta}_k)}, \tag{1.9}$$

čia

$$\widehat{p}_j = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{\eta(k)=j\}}. \tag{1.10}$$

Parametras θ dažniausiai vertinamas maksimalaus tikėtinumo metodu:

$$\widehat{\theta}^{MT} = \arg\max_{\theta} l(\theta \big| X_n), \tag{1.11}$$

čia

$$l(\theta \big| X_n) = \sum_{k=1}^{n} \ln \phi(X(k), \theta). \tag{1.12}$$

(1.9) vadinamas tiesioginio įstatymo metodu. Paprastai (1.11) įvertiniui rasti naudojamas iteratyvus EM algoritmas (Behboodian, 1970; Dempster et al., 1977; Hasselbland, 1966).

Kai pasiskirstymo tankio funkcijų ϕ_j pavidalas nežinomas, tenka taikyti kitus metodus. Galima tiesiogiai apibrėžti parametrizuotą klasifikavimo funkciją

$$h_\theta : \mathbb{R}^d \to \{1, \ldots, q\}, \tag{1.13}$$

čia parametras θ vertinamas mokymo imtyje minimizuojant klaidų skaičių:

$$\widehat{\theta} = \arg\min_{\theta} \sum_{j=1}^{n} \mathbb{1}_{\{h_\theta(X(j)) \neq \eta(j)\}}.$$

Įstatę šį įvertinį į (1.13), gauname griežto klasifikavimo taisyklę $\widehat{\eta}(X) = h_{\widehat{\theta}}(X)$. Analogiškai galima apibrėžti negriežto klasifikavimo taisyklę $\widehat{\pi}(j, X) = h_{\widehat{\theta}}(j, X)$, čia $h_\theta : \mathbb{R}^{d+1} \to \mathbb{R}$ yra nuo parametro θ priklausanti žinomo pavidalo funkcija.

1.1.3. Klasifikavimo algoritmų vertinimo būdai

Klasifikavimo algoritmo kokybei vertinti naudojama keletas skirtingų būdų, kurių pasirinkimas priklauso nuo paties algoritmo, sprendžiamo uždavinio tikslų, duomenų prigimties ir pan.

Tarkime, turime klasifikavimo algoritmą A, nusakomą griežto klasifikavimo taisykle $\widehat{\eta}^A(\cdot)$ arba negriežto klasifikavimo taisykle $\widehat{\pi}^A(j,\cdot)$. Tegul vėl stebime atsitiktinį objektą ir fiksuojame jo požymių vektorių X.

Neteisingo klasifikavimo tikimybė griežto klasifikavimo atveju, dar vadinama tikrąja arba generalizavimo klaida (*generalization error*), apibrėžiama lygybe

$$\epsilon_A = \mathbb{P}\{\widehat{\eta}^A(X) \neq \eta(X)\}. \tag{1.14}$$

Bendru atveju svarbu ne tik, ar algoritmas suklydo, bet ir kaip suklydo, todėl įvedama nuostolių sąvoka. Apibrėžkime funkciją $l(i,j)$, skaitiškai nusakančią nuostolius, kai objektas iš klasės K_i priskiriamas klasei K_j, tenkinančią natūralias sąlygas

$$l(i,j) \geq 0, \quad l(i,i) = 0, \quad 1 \leq i,j \leq q. \tag{1.15}$$

Tada galime apibrėžti vidutinius algoritmo A klasifikavimo nuostolius:

$$L(A) = \mathbb{E}\left[l(\eta(X), \widehat{\eta}^A(X))\right] = \int l(\eta(X), \widehat{\eta}^A(X)) dF(X), \tag{1.16}$$

čia $F(X)$ – a.v. X pasiskirstymo funkcija (1.3).

Akivaizdu, kad (1.14) yra atskiras (1.16) atvejis, kai $l(i,j) = \mathbb{1}_{\{i \neq j\}}$.

Diskriminantinės analizės atveju imtis, taigi ir pagal ją konstruojama klasifikavimo taisyklė, yra atsitiktinės. Tada vidutiniai algoritmo A su pagal imtį X_n sudaryta klasifikavimo taisykle $\widehat{\eta}^{(A,X_n)}$ nuostoliai apibrėžiami lygybe

$$L(A, X_n) = \mathbb{E}\left[l(\eta(X), \widehat{\eta}^{(A,X_n)}(X))\right]. \tag{1.17}$$

Vidurkindami pagal visas n dydžio mokymo imtis, gauname vidutinius algoritmo A nuostolius prie n dydžio imties (Aivazyan et al., 1989):

$$L_n(A) = \mathbb{E}L(A, X_n). \tag{1.18}$$

Negriežtos klasifikavimo taisyklės atveju analogiškai kaip ir (1.16) apibrėžiami nuostoliai

$$L(A) = \mathbb{E}\left[l(\pi(X), \widehat{\pi}^A(X))\right],$$

čia

$$\pi(X) = \pi(\cdot, X), \quad \widehat{\pi}^A(X) = \widehat{\pi}^A(\cdot, X).$$

Šiuo atveju nuostolių funkcionalas l apibrėžtas ne skaičių, o aibėje $\{1, \ldots, q\}$ apibrėžtų skirstinių poroms, pvz.:

$$l(f_1, f_2) = \sum_{j=1}^{q} |f_1(j) - f_2(j)|. \tag{1.19}$$

Praktinius tyrimus aprašančiuose darbuose dažniau vertinami ne algoritmų vidutiniai nuostoliai, o tam tikra prasme priešinga charakteristika – vidutinis algoritmų tikslumas. Šiuo atveju samprotavimai ir formulės išlieka tos pačios, tačiau skiriasi funkcionalai (1.15), o vietoje nuostolių minimizavimo atliekamas tikslumo maksimizavimas.

Praktiniuose taikymuose paprastai turima tik fiksuota ir, kaip taisyklė, nedidelė stebinių imtis, pagal kurią tenka daryti išvadas apie algoritmo vidutinius nuostolius ar tikslumą (toliau – efektyvumas). Konkrečios mokymo bei testavimo imties (imtis, kurioje bandoma pagal mokymo imtį sudaryta algoritmo klasifikavimo taisyklė) atsitiktinumas lemia tai, kad pagal šias imtis įvertintas empirinis efektyvumas gali būti paslinktasis tikrojo efektyvumo įvertinys su nežinoma variacija (Friedman, 1997). Siekiant pataisyti arba bent jau įvertinti empirinių rezultatų nuokrypius, taikomi specialūs metodai (Hastie et al., 2001). Naudojant AIC (Akaike, 1973) ar BIC (Schwarz, 1978) informacinius kriterijus, prie empirinio efektyvumo rezultato pridedama bausmė už modelio sudėtingumą, o naudojant struktūrinį rizikos minimizavimo principą (Vapnik, 1995) klasifikavimo taisyklės sudaromos specialiu būdu, leidžiančiu įvertinti tikrosios efektyvumo reikšmės viršutinį rėžį. Tokie metodai kaip kryžminis patikrinimas (*cross-validation*) (Allen, 1977; Stone, 1974) ar savirankos metodas (*bootstrap*) (Efron, 1979; Efron and Tibshirani, 1997) išnaudoja tą pačią imtį: tam tikru būdu atrinktuose imties poaibiuose atliekant mokymo-testavimo serijas vertinamos įvertinių paklaidos charakteristikos.

1.1.4. Objekto priklausymo kelioms klasėms atvejis

Kartais (taip pat ir šiame darbe) nagrinėjama situacija, kai objektas gali vienu metu priklausyti kelioms klasėms iš karto. Tokiu atveju lygybė (1.1) ir toliau galioja, tačiau klasės K_i nebėra tarpusavyje nesikertančios. Tarkime, kad vėl turime q klasių ir stebime atsitiktinį objektą O su požymių vektoriumi X.

Tegul objektas priskirtas atsitiktinėms klasėms $\eta = \{\eta_1, \ldots, \eta_k\}$, čia k taip pat atsitiktinis dydis. Kiekvienai aibei $J = \{j_1, \ldots, j_l\} \subset \{1, \ldots, q\}$ apibrėžkime atsitiktinį įvykį

$$H_J = H_{\{j_1,\ldots,j_l\}} = \{O \in K_{j_1}, \ldots, O \in K_{j_l}\}. \tag{1.20}$$

(1.2) formulėje apibrėžti įvykiai H_j susiję su (1.20) apibrėžtais H_J sąryšiu $H_j = \{H_{\{j\}} \big| |\eta| = 1\}$. Čia $|\cdot|$ kaip paprastai žymi aibės elementų skaičių. Laikysime, kad įvykiai H_j nepriklausomi. Tada

$$p_J = \mathbb{P}\{H_J\} = \prod_{j \in J} p_j \cdot \prod_{j \notin J} (1 - p_j).$$

Dažniausiai vietoje pilno modelio

$$\phi(x) = \sum_J p_J \phi_J(x),$$

čia $\phi(x)$ – a.v. X tankio funkcija, o $\phi_J(x)$ – atitinkamai sąlyginės tankio funkcijos, taikomas uždavinio transformavimas, vėl suvedantis į sąlyginai paprastą skirstinių mišinio modelio atvejį (1.6). Klasifikavimo taisyklė (1.5) apibendrinama:

$$\widehat{\eta}(X) = \{\widehat{\eta}_1(X), \ldots, \widehat{\eta}_{\widehat{k}}(X)\}, \tag{1.21}$$

čia

$$\widehat{\eta}_i(X) = \arg\max_{j \in Q_i} \widehat{\pi}(j, X), \quad i = \overline{1, \widehat{k}}, \tag{1.22}$$

$$Q_1 = \{1, \ldots, q\}, \quad Q_j = Q_{j-1} \setminus \widehat{\eta}_{j-1}(X), \quad j > 1. \tag{1.23}$$

$\widehat{\pi}(j, X)$ kaip ir anksčiau yra (1.4) statistinis įvertis, o $\widehat{k}$ yra nežinomo klasių skaičiaus k statistinis įvertis.

Mokymo imtis (1.8), kurioje $\eta(j) = \{\eta_1(j), \ldots, \eta_{k(j)}(j)\}$ dabar yra aibės, transformuojama į imtį X_N^* ($N \geq n$), kurioje kiekvienas elementas priskirtas tik vienai klasei. Tai atliekama pavyzdžiui pakartojant tą patį požymių vektorių su kiekviena klase atskirai:

$$(X(j), \eta(j)) \to \{(X(j), \eta_1(j)), \ldots, (X(j), \eta_{k(j)}(j))\}. \tag{1.24}$$

Vidutiniai pagal naująją imtį X_N^* sukonstruoto griežto klasifikavimo algoritmo A nuostoliai $L(A, X_N^*)$ apibrėžiami ta pačia (1.17) lygybe, tik šiuo atveju nuostolių funkcija $l(\cdot, \cdot)$ apibrėžiama aibių poroms.

1.2. Dokumento reprezentavimo skaitiniais požymių vektoriais būdai

Mokslo publikacijai, kaip ir bet kokiam tekstui, reikalinga transformacija, pervedanti jį į skaitinių požymių rinkinį, priimtiną algoritmams. Paprastai dokumentas pateikiamas kaip tam tikrų teksto elementų svarbą atspindinčių svorių vektorius. Dažniausiai naudojami du teksto elementų parinkimo ir reprezentavimo būdai: *žodžių krepšelis (bag of words)* ir *fiksuoto ilgio vektorius*. Pirmuoju atveju dokumentą reprezentuoja visų jo žodžių rinkinys, neretai ignoruojant jų tvarką. Atlikta nemažai tyrimų, pvz., (Dumais et al., 1998), kurie parodė, kad sudėtingesnes struktūras naudojantys pateikimo būdai dažnai neduoda apčiuopiamai geresnių rezultatų, tačiau ženkliai padidina skaičiavimų apimtis. Antruoju atveju fiksavus teksto terminų (teksto fragmentai, nebūtinai pavieniai žodžiai) žodyną $V = \{v_j, j = \overline{1,h}\}$, dokumentas pateikiamas kaip h ilgio vektorius $x = (x_1, \ldots, x_h)$, čia x_j žymi termino v_j svorį.

Vienos populiariausių svorių skaičiavimo schemų yra binarinė bei vadinamoji *tfidf (term frequency – inverse document frequency)* (Salton and Buckley, 1988). Tarkime, turime mokymo imtį (1.8), kurioje kiekvienas iš vektorių $X(i)$ atitinka stebėtą mokslinį dokumentą. Binarinės schemos atveju konkretaus termino svoris sutampa su įvykio „terminas aptinkamas tekste“ indikatoriumi. *Tfidf* schema termino $v_k \in V$ svorį konkrečiame mokymo imties dokumente $x = (x_1, \ldots, x_h)$ apskaičiuoja pagal formulę

$$tfidf(v_k, x) = \nu_x(v_k) \cdot \log \frac{|X_n|}{\mu_{X_n}(v_k)},$$

čia $\nu_x(v_k)$ žymi termino $v_k \in V$ pasirodymo dokumente x dažnumą, $\mu_{X_n}(v_k)$ – skaičių dokumentų iš X_n, kuriuose bent kartą sutinkamas v_k, o $|X_n|$ – aibės X_n elementų skaičių. Paprastai šie svoriai normalizuojami: galutinis v_k svoris dokumente x skaičiuojamas pagal formulę

$$w_k(x) = \frac{tfidf(v_k, x)}{\left(\sum_{i=1}^{h} tfidf(x_i, x)^2\right)^{\frac{1}{2}}}.$$

1.3. Informatyviausių požymių atrinkimo metodai

Klasifikavimo algoritmai paprastai remiasi didelės apimties skaičiavimais (ypač mokymo fazėje), o šios apimtys stipriai priklauso nuo dokumentus reprezentuojančių požymių vektorių ilgio. Informatyviausių požymių atrinkimo metodai (Liu and Yu, 2005) skirti optimaliai sumažinti šių vektorių ilgius. Šie metodai paprastai sumažina klasifikatorių persimokymo (*overfitting*) efektą, vadinasi, ir tikėtinus klasifikavimo nuostolius, gaunamus klasifikuojant mokyme nenaudotus duomenis. Vieni populiariausių požymių atrinkimo metodų yra IG (*information gain*), dar žinomas kaip santykinė entropija arba KL divergencija (Cover and Thomas, 1991; Kullback and Leibler, 1951), χ^2 ir DF (*document frequency*) (Sebastiani, 2002; Yang and Pedersen, 1997). Atlikta nemažai tyrimų, kurie praktiniuose eksperimentuose lygino šiuos bei kitus metodus, pvz., darbai (Forman, 2003; Rogati and Yang, 2002; Yang and Pedersen, 1997). Tyrimų rezultatai leidžia teigti, kad visi šie metodai leidžia ženkliai (90 % ir daugiau) sumažinti požymių aibę, neprarandant klasifikavimo tikslumo, o optimalaus metodo pasirinkimas priklauso nuo tokių faktorių, kaip duomenų apimtis, uždavinio tikslai ir pan.

1.4. Tekstų klasifikavimui taikomų algoritmų apžvalga

Didžioji dalis *mokslo tekstų* analizės, taip pat ir klasifikavimo, problemoms skirtų tyrimų orientuota į pastaruoju metu tokias itin aktyviai vystomas sritis, kaip biomedicina ar genų inžinerija, pvz., (Craven and Kumlien, 1999; Krallinger and Valencia, 2005; Shatkay et al., 2008), nors reikėtų pastebėti, kad juose dažniausiai apsiribojama tiesioginiu žinomų įprastų tekstų apdorojimo metodų taikymu. Tebesitęsiantis kompiuterinės technikos bumas bei praktinių rezultatų poreikis lemia, kad tokius tyrimus paprastai atlieka informatikai, kurie dažniausiai orientuojasi į didelių duomenų masyvų analizės pagrindu atliekamas optimalių žinomų sprendimų kombinacijų ar algoritmų parametrų paieškas.

Algoritmų, taikomų įprastų tekstų klasifikavimui, yra nemažai. Išsami jų apžvalga iš automatinio mokymosi (*machine learning*) perspektyvos pateikta darbe (Sebastiani, 2002). Tolesniuose skyreliuose panagrinėsime kelis pasirinktus algoritmus, kurie pasižymi dideliu tikslumu, ir kuriuos realių duomenų pagrindu lyginsime su mūsų siūlomais sprendimais. Algoritmai pasirinkti remiantis trimis pagrindiniais kriterijais: dažnas naudojimas kitų autorių tyrimuose, aukštas

tikslumo lygis (taip pat remiantis kitų autorių darbais) bei skirtingos prigimties pagrindinės idėjos.

Iš tų, kurių neapžvelgsime, derėtų išskirti tokias stambias grupes kaip neuroniniais tinklais paremti algoritmai (Ng et al., 1997; Ruiz and Srinivasan, 1999; Wiener et al., 1995), sprendimų taisyklių ir sprendimų medžių algoritmai (Cohen and Hirsh, 1998; Cohen and Singer, 1996; Dumais et al., 1998; Li and Yamanishi, 2002; Li and Jain, 1998; Weiss et al., 1999), genetiniai algoritmai (Bergström et al., 2000; Masand, 1994), algoritmų ansambliai (Breiman, 1998; Larkey and Croft, 1996; Schapire and Singer, 2000; Schapire et al., 1998) ir kt. Paminėtini tyrimai, kuriuose panašiai kaip ir šiame disertaciniame darbe vienu ar kitu būdu domimasi išreikštiniais klasių (kategorijų, raktinių žodžių) ir teksto elementų sąryšiais bei modeliais (Averin and Vassilevskaya, 2002; Widdows, 2003; Zaiane and Antonie, 2002), kontekstinės informacijos panaudojimo būdais (Metzler and Croft, 2005; Pickens and MacFarlane, 2006). Darbuose (Ghamrawi and McCallum, 2005; Yan et al., 2007) nagrinėjama situacija, kai klasės nėra tarpusavyje nepriklausomos (taip yra hierarchinių klasifikavimo schemų kaip MSC atveju). Atskiro paminėjimo vertas tematiškai disertaciniam darbui artimas tyrimas (Rehurek and Sojka, 2008), kuriame taip pat spręstas matematinių tekstų klasifikavimo uždavinys. Tiesa, šio darbo autoriai nenagrinėjo mokslinio teksto specifikos ir taikė standartinius metodus.

1.4.1. Naivaus Bajeso algoritmas

Tikimybiniai algoritmai tiesiogiai vertina aposteriorines tikimybes (1.4) ir šiuos įverčius naudoja klasių nustatymo taisyklėje (1.5). Tegul stebimas atsitiktinis dokumentas, kurį reprezentuojantis atsitiktinis požymių vektorius X įgyja konkrečią reikšmę x. Tegul šis dokumentas priklauso nestebimai klasei $\eta \in K$. Kaip ir anksčiau, $d = dim(X)$. Naudodamiesi Bajeso teorema, galime užrašyti

$$\mathbb{P}\{\eta = w|X = x\} = \frac{\mathbb{P}\{\eta = w\}\mathbb{P}\{X = x|\eta = w\}}{\mathbb{P}\{X = x\}}. \tag{1.25}$$

Naivaus Bajeso algoritmas remiasi prielaida, kad vektoriaus X komponentės yra sąlyginai nepriklausomos prie sąlygos $\eta = w$, todėl galioja

$$\mathbb{P}\{X = x|\eta = w\} = \prod_{i=1}^{d} \mathbb{P}\{X_i = x_i|\eta = w\}. \tag{1.26}$$

Čia dešinėje pusėje dauginamos tikimybės, kad atsitiktinai parinkto dokumento, priklausančio klasei w, atsitiktinai parinktoje pozicijoje i stebima konkreti reikšmė x_i.

Binarinių svorių fiksuoto ilgio vektoriumi reprezentuojamų dokumentų atveju, ($x = (x_1, \ldots, x_h), x_i \in \{0, 1\}$, čia $x_i = 1$, jei elementas $v_i \in V$ sutinkamas stebimo dokumento tekste) gauname vadinamąjį binarinį nepriklausomumo modelį (*BIM – Binary Independence Model*) (Lewis, 1998; Robertson and Jones, 1988). Pažymėję

$$p_w(j) = \mathbb{P}\{x_j = 1 | \eta = w\}, \quad j = \overline{1, h},$$

galime užrašyti

$$P\{X_j = x_j | \eta = w\} = p_w(j)^{x_j}(1 - p_w(j))^{1-x_j} = \left(\frac{p_w(j)}{1 - p_w(j)}\right)^{x_j}(1 - p_w(j)). \tag{1.27}$$

Įstatę (1.27) į (1.26), o pastarąją į (1.25), bei atlikę logaritmavimą, gauname

$$\log \mathbb{P}\{\eta = w | X = x\} = \log \mathbb{P}\{\eta = w\} + \sum_{i=1}^{h} x_i \log \frac{p_w(i)}{1 - p_w(i)} + \sum_{i=1}^{h} \log (1 - p_w(i)) - \log \mathbb{P}\{X = x\}. \tag{1.28}$$

Laikykime, kad jei dokumentas nepriskiriamas klasei w, tai jis priskiriamas klasei $\bar{w}$ (dviejų klasių atveju tokia klasė išties egzistuoja). Akivaizdu, kad $\mathbb{P}\{\eta = \bar{w} | X = x\} = 1 - \mathbb{P}\{\eta = w | X = x\}$. Turime:

$$\log (1 - \mathbb{P}\{\eta = w | X = x\}) = \log (1 - \mathbb{P}\{\eta = w\}) + \sum_{i=1}^{h} x_i \log \frac{p_{\bar{w}}(i)}{1 - p_{\bar{w}}(i)} + \sum_{i=1}^{h} \log (1 - p_{\bar{w}}(i)) - \log \mathbb{P}\{X = x\}. \tag{1.29}$$

Iš lygybės (1.28) atėmę lygybę (1.29), gauname

$$\log \frac{\mathbb{P}\{\eta = w | X = x\}}{1 - \mathbb{P}\{\eta = w | X = x\}} = \log \frac{\mathbb{P}\{\eta = w\}}{1 - \mathbb{P}\{\eta = \bar{w}\}} + \sum_{i=1}^{h} x_i \log \frac{p_w(i)(1 - p_{\bar{w}}(i))}{p_{\bar{w}}(i)(1 - p_w(i))} + \sum_{i=1}^{h} \log \frac{1 - p_w(i)}{1 - p_{\bar{w}}(i)}.$$

Gavome logistinės regresijos modelį aposteriorinėms tikimybėms, o kairėje

lygybės pusėje stovintis dydis tiesiškai priklauso nuo parametrų x_i ir nepriklauso nuo tikimybės $\mathbb{P}\{X = x\}$.

Paprasto *BIM* modelio trūkumai susiję su tuo, kad jame atsižvelgiama tik į termino egzistavimo dokumento tekste faktą, ir todėl ignoruojama informacija, kurią suteikia terminų pasikartojimo tekste dažnumas. Taip pat modelyje neatsižvelgiama į dokumento ilgį, todėl termino svoris ilgame dokumente laikomas lygus jo svoriui trumpame dokumente. Šie bei kiti trūkumai sąlygojo įvairiausių patobulinimų bangą (Kim et al., 2002; Lewis, 1998).

Viena pasiūlymų grupė susijusi su binarinio dokumento reprezentavimo būdo pakeitimu. Pavyzdžiui, galima laikyti, kad dokumentą reprezentuojančio vektoriaus $x = (x_1, \ldots, x_h)$ komponentės x_i yra sveikaskaitinių atsitiktinių dydžių, atitinkančių terminų pasikartojimo tekste dažnumus, stebėjimai. Naivaus Bajeso modelis laikys, kad šie dydžiai yra nepriklausomi, tačiau jie bus modeliuojami nebe Bernulio, o sveikųjų skaičių skirstiniais ar jų mišiniais. Tyrimuose dažniausiai pasirenkamas Puasono skirstinys, pvz., (Harter, 1975a,b; Katz, 1996). Darbe (Robertson and Walker, 1994) pasiūlytas metodas, kuris naudoja paprastą Puasono skirstinių mišinio modelio aproksimaciją.

Kiti autoriai siūlo pereiti nuo dokumento interpretavimo kaip atsitiktinio įvykio prie termino interpretavimo kaip atsitiktinio įvykio (vadinamasis polinominis modelis) (Guthrie et al., 1994; McCallum and Nigam, 1998). Tokiu būdu vietoje atsitiktinio vektoriaus, atitinkančio dokumentą, turime daug binominių atsitiktinių dydžių, atitinkančių dokumento terminus, o dokumento ilgis įkomponuojamas į modelį natūraliai. Šiuo atveju vėl įvedama sąlyginio terminų nepriklausomumo prielaida, tačiau reikia pastebėti, ji dabar galioja ne tik tarp skirtingų terminų, bet ir tarp to paties termino skirtingų egzempliorių tekste. Grįžę prie (1.25) ir (1.26) matome, kad iš mokymo imties reikia įvertinti apriorines klasių tikimybes $\mathbb{P}\{\eta = w\}$ ir sąlygines tikimybes $\mathbb{P}\{X_i = x_i | \eta = w\}$. Šio modelio atveju dokumentas paprastai reprezentuojamas *žodžių krepšelio* būdu, todėl natūraliai įvedama sąlyginio stacionarumo sąlyga (nepriklausomumo nuo konkrečios pozicijos tekste) terminų skirstiniams. Tada minėtąsias tikimybes galima paprastai įvertinti skaičiuojant atitinkamų elementų stebėjimų dažnumus mokymo imtyje. Siekiant išvengti terminų stebėjimo tekste sąlyginių tikimybių nulinių įverčių (terminas mokymo imtyje nestebėtas tam tikros klasės dokumentuose), kurie gali stipriai iškreipti aposteriorinių tikimybių įverčius, naudojami įvairūs glodinimo metodai, pvz., Laplaso (McCallum and Nigam, 1998), kurie užtikrina, kad nulių nebus.

Dar viena pasiūlymų grupė koreguoja pačią aiškiai neadekvačią nepriklausomumo prielaidą, jos susilpninimas naudojamas tokiuose metoduose kaip atidėtas

(tingus) Bajeso mokymasis (*Lazy Bayesian Learning*) (Wang and Webb, 2002; Zheng and Webb, 2000), įvairūs Bajeso tinklų modeliai (Cooper and Herskovits, 1992; Friedman et al., 1997; Peng and Schuurmans, 2003) ir kituose (Rennie et al., 2003; Webb et al., 2005).

(Langley and Sage, 1994) pasiūlytas selektyvus Bajeso metodas (*Selective Bayes Method*) yra pavyzdys, kaip galima pakeisti dokumento reprezentavimą, kad nepriklausomumo prielaida būtų teisingesnė, klasifikatorių mokymui naudojant ne visus terminus, bet tik jų dalį.

Iš kitos pusės, darbo (Domingos and Paffani, 1996) autoriai parodė, kad naivaus Bajeso metodui išties nėra būtinas nepriklausomumas, t. y., net ir esant akivaizdžioms priklausomybėms tikimybių įverčiai gali būti netikslūs, tačiau klasifikavimo rezultatas – tikslus.

Nors gauta nemažai rezultatų, patvirtinančių vienų ar kitų patobulintų ir sudėtingesnių metodų pranašumą prieš paprasčiausią naivaus Bajeso klasifikatorių, tačiau neretai šis pranašumas įgyjamas tik tam tikrose specifinėse situacijose, o stebimas tikslumo padidėjimas neatperka skaičiavimų sudėtingumo išaugimo. Todėl naivaus Bajeso klasifikatorius, dažniausiai – ganėtinai paprastas polinominis modelis su adityviu (Laplaso) glodinimu (McCallum and Nigam, 1998), ir toliau išlieka labai populiarus dėl savo universalumo, robastiškumo, skaičiavimų paprastumo bei dažniausiai itin gero klasifikavimo tikslumo. Būtent šis algoritmo variantas nagrinėjamas darbo eksperimentinėje dalyje.

1.4.2. Atraminių vektorių algoritmas

V. Vapnik su kolegomis (Boser et al., 1992; Cortes and Vapnik, 1995; Vapnik, 1995) pasiūlė neparametrinį klasifikavimo metodą *SVM (Support Vector Machines)*, kuris vietoje empirinės apsirikimo rizikos (klasifikavimo klaidų mokymo imtyse lygis) mažinimo konstruoja klasifikavimo taisyklę taip, kad būtų minimizuojama tikroji klasifikavimo klaida, t. y., tikimybė, kad klasifikatorius suklys, sutikęs nematytą testinį pavyzdį. Paprasčiausiame pavidale metodas yra tiesinis savo parametrų atžvilgiu, be to, skirtas spręsti paprastą vienos klasės uždavinį – teisingai atskirti duomenis, priklausančius klasei nuo nepriklausančių.

Tarkime, vėl turime mokymo imtį X_n, žr. (1.8). Čia $X(i) \in \mathbb{R}^h$: dokumentas reprezentuojamas fiksuoto ilgio požymių vektoriumi ir traktuojamas kaip taškas h–matėje euklidinėje erdvėje; $\eta(i) = 1$, jei objektas O_i priklauso klasei, ir $\eta(i) = -1$ priešingu atveju.

Sakysime, kad mokymo duomenys yra tiesiškai separabilūs, jei egzistuoja vektorius $w \in \mathbb{R}^h$, skaliarai b ir $C > 0$, kad galioja

$$\eta(i)(w \cdot X(i) + b) \geq C. \tag{1.30}$$

Pora (w, b), tenkinanti sąlygą (1.30), vienareikšmiškai identifikuoja lygybe $w \cdot X + b = 0$ apibrėžiamą hiperplokštumą H, vadinamą *sprendimo paviršiumi (decision surface)*, atskiriančią klasei priklausančius objektus nuo nepriklausančių. Kadangi tokių porų gali būti be galo daug, pasirenkama vadinamoji optimali hiperplokštuma (Vapnik, 1995) – ta, kuri skiria duomenis didžiausiu skirtumu. Toliau panagrinėsime šios hiperplokštumos nustatymą.

Absoliutus atstumas nuo taško $X \in \mathbb{R}^h$ iki hiperplokštumos H lygus

$$\rho_H(X) = \frac{1}{||w||}\eta(i)(w \cdot X + b).$$

Pažymėkime

$$\rho_H^-(X_n) = \min_{i:\eta(i)=-1} \rho_H(X(i)), \quad \rho_H^+(X_n) = \min_{i:\eta(i)=1} \rho_H(X(i)).$$

Tada skirtumas, kuriuo hiperplokštuma H skiria mokymo imties X_n taškus $X(i), i = \overline{1,n}$, apskaičiuojamas pagal formulę

$$\rho_H(X_n) = \rho_H^-(X_n) + \rho_H^+(X_n) \geq \frac{2C}{||w||}.$$

Tegul $\rho_H(X_n) > 2C/||w||$. Parinkime naujas reikšmes $C' = \rho_H(X_n)||w||/2$ ir $b' = b + ||w||(\rho_H^-(X_n) - \rho_H^+(X_n))/2$. Nauja hiperplokštuma H', apibrėžta to paties vektoriaus w ir naujo skaliaro b' pora, tenkins atitinkamai pakeistą sąlygą (1.30), be to, $\rho_{H'}(X_n) = 2C'/||w||$ ir $\rho_H^-(X_n) = \rho_H^+(X_n) = C'/||w||$. Todėl toliau nemažindami bendrumo laikysime, kad teisinga lygybė

$$\rho_H(X_n) = \frac{2C}{||w||}. \tag{1.31}$$

Skaičiuojant šią reikšmę atsižvelgiama tik į keletą hiperplokštumai artimiausių taškų – tuos, kuriems $\eta(i)(w \cdot X(i) + b) = C$. Šie vektoriai vadinami *atraminiais vektoriais (support vectors)*. Jeigu iš mokymo imties pašalintume visus vektorius, išskyrus atraminius, gautume lygiai tą patį sprendimo paviršių.

Iš (1.30) ir (1.31) gauname, kad optimali hiperplokštuma (w^*, b^*) randama

išsprendus optimizavimo uždavinį

$$w^*, b^* = \arg\max_{w,b} \{2C/||w|| \mid \eta(i)(w \cdot X(i) + b) \geq C\}.$$

Atlikę elementarias pertvarkas, gauname

$$w^*, b^* = \arg\min_{w,b} \{\frac{1}{2}||w||^2 \mid \eta(i)(w \cdot X(i) + b) \geq 1\}.$$

Šį optimizavimo uždavinį galime spręsti naudodami Lagranžo daugiklių metodą. Lagranžo funkcija lygi

$$L = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \lambda_i[\eta(i)(w \cdot X(i) + b) - 1], \quad \lambda_i \geq 0. \tag{1.32}$$

Diferencijuodami pagal w ir b bei gautąsias išvestines prilyginę nuliui, turime:

$$w = \sum_{i=1}^{n} \lambda_i \eta(i) X(i), \tag{1.33}$$

$$\sum_{i=1}^{n} \lambda_i \eta(i) = 0. \tag{1.34}$$

Įstatę (1.33) į (1.32) ir atsižvelgdami į (1.34) gauname funkciją

$$L = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \eta(i) \eta(j) X(i) \cdot X(j), \tag{1.35}$$

kurią reikia maksimizuoti prie sąlygos $\lambda_i \geq 0, i = \overline{1, n}$.

Šiam kvadratinio optimizavimo uždaviniui esama algoritmų, kurie randa globalų sprendinį $\{\lambda_i^*, i = \overline{1, n}\}$. Pagal (1.33) optimalios hiperplokštumos vektorius w^* gali būti užrašytas tiesine mokymo vektorių kombinacija

$$w^* = \sum_{i=1}^{n} \lambda_i^* \eta(i) X(i).$$

Čia tik atraminiams vektoriams $\lambda_i^* \neq 0$.

Skaliaras b^* ir vektorius w^* susieti lygybe

$$b^* = \frac{1}{2}(w^* \cdot X_+ + w^* \cdot X_-).$$

Čia X_+ ir X_- yra du atraminiai vektoriai iš imties X_n, kurių pirmasis priklauso klasei, o antrasis – nepriklauso.

Neseparabilių duomenų atveju ieškoma sprendimo paviršiaus, kuris atlikdamas duomenų atskyrimą derintų teisingo atskyrimo pločio maksimizavimą su baudos už neteisingai atskirtus objektus minimizavimu. Tegul

$$\Phi_\sigma = \sum_{i=1}^{n} \xi_i^\sigma, \tag{1.36}$$

$$\eta(i)(w \cdot X(i) + b) \geq C - \xi_i, \quad 1 \leq i \leq n, \tag{1.37}$$

$$\xi_i \geq 0, \quad 1 \leq i \leq n. \tag{1.38}$$

Mažam σ ieškodami hiperplokštumos (w, b), kuri minimizuotų (1.36) apibrėžtą Φ_σ prie sąlygų (1.37) ir (1.38), gauname optimalų rinkinį mokymo imties X_n vektorių $X(i)$, kuriems atskyrimas klaidingas ($\xi_i > 0$), o juos pašalinus iš mokymo imties, likusieji duomenys būtų separabilūs. Šiems likusiems vektoriams galima aukščiau aprašytais metodais surasti optimalią skiriančią hiperplokštumą. Tokiu būdu optimalaus atskyrimo idėja apibendrinama nebūtinai separabiliems duomenims:

$$w^*, b^* = \arg\min_{w,b} \left[\frac{1}{2}||w||^2 + A \cdot F(\Phi_\sigma) \right],$$

galiojant sąlygoms (1.37) ir (1.38). Čia A – konstanta, $F(u)$ – iškila monotoninė funkcija, skaliarui b^* galioja ta pati (1.4.2) formulė.

Tuo atveju, kai $\sigma = 1$, yra efektyvių metodų, sprendžiančių (1.4.2) optimizavimo uždavinį. Gautas sprendinys vadinamas *silpno skirtumo hiperplokštuma (soft margin hyperplane)*. Optimizavimo uždavinyje (1.35) skaliarines vektorių sandaugas pakeitus netiesiniais branduoliais gaunami metodai su netiesiniu duomenis skiriančiu paviršiumi (Boser et al., 1992; Joachims, 1998).

T. Joachims darbuose (Joachims, 1998, 1999) adaptavo SVM metodą tekstų klasifikavimui, apibrėždamas klasifikatorių $c : \mathbb{R}^h \to \{-1, 1\}$ lygybe

$$c(x) = sign\{w \cdot x + b\}, \tag{1.39}$$

čia x – stebėtą dokumentą atitinkantis požymių vektorius o (w, b) yra optimali

hiperplokštuma, gauta iš mokymo imties.

Kelių klasių atveju taikomas aprašytų metodų apibendrinimas. Visų pirma, kiekvienai klasei apmokomas atskiras binarinis klasifikatorius. Tada klasifikavimo fazėje vietoje griežto sprendimo „priklauso/nepriklauso klasei“ priėmimo vertinamos priklausymo kekvienai iš klasių tikimybės pagal tai, kiek toli nuo atitinkamos klasės skiriamosios hiperplokštumos yra nagrinėjamąjį dokumentą atitinkantis požymių vektorius.

Aptartoji atraminių vektorių metodika yra labai populiari ir aktyviai vystoma. Algoritmai naudojami sprendžiant ne tik klasifikavimo, bet ir daugelį kitų taikomųjų duomenų analizės uždavinių. Algoritmas pasižymi aukštu tikslumo lygiu, tačiau taip pat turi keletą šio tyrimo požiūriu didelių trūkumų. Visų pirma, optimalių parametrų parinkimas dažnai yra sudėtingas uždavinys, kuriam spręsti naudojami specialūs apytiksliai metodai. Antra, šis algoritmas yra vadinamosios „juodosios dėžės“ principo: vidiniai parametrai ir jų sąryšiai yra paslėpti nuo naudotojo, o jų aiški interpretacija būtų neįmanoma. Tai reiškia, kad negalima panaudoti ekspertų žinių apie mokslo sritį ar pritaikyti gautus rezultatus kitiems uždaviniams spręsti. Šie trūkumai lemia, kad nors algoritmas darbe nagrinėjamas kaip viena tiksliausių ir populiariausių alternatyvų, tačiau jis nėra laikomas pilnaverčiu kandidatu spręsti disertacijoje suformuluotą uždavinį.

1.4.3. Tiesinis mažiausių kvadratų algoritmas

Tiek naivaus Bajeso metodas, tiek atraminių vektorių algoritmas paprastu atveju yra tiesiniai savo parametrų atžvilgiu. Y. Yang ir C.G. Chute darbuose (Yang and Chute, 1992, 1993, 1994) pasiūlė dokumentų klasifikavimui naudoti vadinamąjį tiesinį mažiausių kvadratų metodą (*Linear least squares fit, LLSF*), kuris pagal savo apibrėžimą daro prielaidą apie tiesinę modelio parametrų priklausomybę.

Tegul turime sunumeruotą teksto terminų žodyną $V, |V| = h$ bei galimų klasių žodyną $K = \{K_i, i = \overline{1,q}\}$, o dokumentas gali būti priskirtas kelioms klasėms vienu metu. Mokymo duomenys (1.8) pateikiami kaip dvi matricos $T(n \times h)$ ir $C(n \times q)$, čia n yra mokymo imties X_n dydis. Matricos T eilutė atitinka vieną dokumentą – joje nurodyti terminų, kurių kiekvieną atitinka matricos stulpelis, svoriai tame dokumente. Atitinkama matricos C eilutė nurodo klasių svorius: vienetas, jei dokumentas priskirtas klasei ir nulis, jei ne.

Daroma prielaida, kad esama funkcinės priklausomybės tarp klasių ir terminų svorių $f : f(T^*) = C^*$, kuri, be to, yra tiesinė, t. y., $f(T^*) = T^*F$, $F(h \times q)$ – nežinoma matrica. Čia T^* ir C^* žymi generalinę dokumentų aibę

atitinkančius svorių matricų T ir C analogus. Kadangi stebinių imtis atitinka tik dalį generalinės aibės, be to, galimos stebėjimų klaidos, kaip paprastai matricai F rasti siūloma taikyti mažiausiųjų kvadratų metodą imties duomenims:

$$\hat{F} = \arg\min_{F} ||TF - C||^2_{HS}, \tag{1.40}$$

čia $||\cdot||_{HS}$ žymi matricos Hilberto–Šmito normą.

Vienas galimų (1.40) sprendimo būdų paremtas vadinamuoju skaidymu singuliariomis reikšmėmis (*Singular Value Decomposition, SVD*). Kiekviena stačiakampė $n \times h$ dydžio matrica A gali būti išreikšta matricų sandauga $A = U\Sigma V^T$, kurioje $m = \min\{n, h\}$, matricos $U(n \times m)$ ir $V(h \times m)$ turi ortogonalius stulpelius ($U^TU = V^TV = I_m$), o $\Sigma(m \times m)$ yra diagonalinė: $\Sigma = diag(\sigma_1, \ldots \sigma_m)$. Nemažinant bendrumo galima laikyti, kad galioja šios sąlygos:

$$\sigma_i \geq \sigma_{i+1}, \quad i = \overline{1, m-1}, \tag{1.41}$$

$$\sigma_i > 0, \quad 1 \leq i \leq r \quad \text{ir} \quad \sigma_i = 0, \quad i > r. \tag{1.42}$$

U ir V vadinamos kairiųjų ir dešiniųjų singuliarių vektorių, o Σ – singuliarių reikšmių matricomis. SVD yra kvadratinės matricos skaidymo tikrinėmis reikšmėmis (*eigen value decomposition*) apibendrinimas stačiakampėms matricoms, be to, nesunku pastebėti, kad $AA^T = (UV)\Sigma^2(UV)^T$ ir, analogiškai, $A^TA = (VU)\Sigma^2(VU)^T$, t. y., singuliariosios reikšmės iš matricos Σ yra matricų AA^T bei A^TA tikrinių reikšmių neneigiamos šaknys, o matricų UV bei VU stulpeliai yra atitinkamų matricų tikriniai vektoriai. Pagal Eckarto-Jango teoremą (Eckart and Young, 1936):

$$\arg\min_{rank(X)=k} ||A - X||^2_{HS} = A_k. \tag{1.43}$$

Čia $k < rank(A)$, $A_k = U\Sigma_k V^T$, o Σ_k gaunama iš matricos Σ, atlikus transformaciją $\sigma_i = 0, i > k$.

Tada (1.40) sprendinys gali būti užrašytas:

$$\hat{F} = T^*_k C = V\Sigma^{-1}_k U^T C. \tag{1.44}$$

Čia $T = U\Sigma V^T$ – matricos T SVD dekompozicija, Σ^{-1} yra matricai Σ atvirkštinė matrica, $T^* = V\Sigma^{-1}U^T$ – Mūro-Penrouzo pseudoatvirkštinė matrica matricai T, o T^*_k gaunama naudojant šioje pseudoatvirkštinės matricos išraiškoje Σ^{-1}_k vietoje Σ^{-1}.

Pastebėsime, kad (1.44) vietoje Σ naudodami Σ_k su k didžiausių singuliarių reikšmių atliekame požymių erdvės matavimo sumažinimą: tik šias reikšmes atitinkantys teksto terminai įtakoja klasifikavimo sprendimą.

Turėdami apskaičiuotą $\hat{F}$, naujai stebėtą dokumentą, kuris reprezentuojamas požymių vektoriumi $x = (x_1, \ldots, x_h)$, priskiriame toms klasėms, kurioms atitinkamos vektoriaus $x\hat{F}$ komponentės didžiausios.

Aptartasis tiesinis mažiausių kvadratų metodas pasižymi keletu įdomių savybių. Psuedoatvirkštinės matricos skaičiavimo metu papildomai atliekamas ir požymių erdvės matavimo mažinimas, eliminuojantis neinformatyvių terminų įtaką. Skirtingai negu ankstesniame skyrelyje aptartame atraminių vektorių algoritme, šiame parametrų (klasių svorių ir terminų svorių) sąryšiai yra išreikštiniai ir ganėtinai paprasti. Visgi, šių sąryšių interpretacija nėra aiški. Algoritmo mokymo fazėje atliekama daug skaičiavimo resursų reikalaujanti matricos skaidymo singuliariomis reikšmėmis operacija, tačiau klasifikavimo fazėje skaičiavimai yra labai paprasti ir greiti. Pastebėsime, kad dėl algoritmo pagrindą sudarančių ganėtinai įprastų tiesinės algebros operacijų, algoritmo realizavimas neturėtų būti sudėtingas, nes galima pasinaudoti stadartinėmis procedūromis ir paketais.

1.4.4. k kaimynų algoritmas

Didelę grupę sudaro vadinamieji pavyzdžiais paremti algoritmai (Aha et al., 1991; Creecy et al., 1992; Masand et al., 1992), kurią atstovauja populiarusis k artimiausių kaimynų metodas (Cover and Hart, 1967; Friedman; Yang, 1994), sprendimą apie stebėto objekto klasę ar klases priimantis pagal sprendimus tų mokymo imties elementų, kurie yra arčiausiai (apibrėžto atstumo arba panašumo mato prasme) nagrinėjamojo. Šie samprotavimai paremti akivaizdžia iš (1.4) išvada, kad stebinio su požymių vektoriumi $X = x$ klasifikavimui nereikia tiksliai žinoti pasiskirstymo tankių $\phi_j(x)$ – užtenka mokėti juos palyginti.

Paprasčiausiame pavidale algoritmas stebėto dokumento su požymių vektoriumi x negriežto klasifikavimo tikimybes (1.4) vertina pagal taisyklę

$$\widehat{\pi}(j, x) = \frac{1}{k} \sum_{i \in J_k(x)} \mathbb{1}_{\{\eta(i)=j\}}.$$

Čia $\eta(i)$ yra teisingas i–ojo dokumento iš mokymo imties (1.8) klasifikavimo sprendimas, o $J_k(x)$ – k elementų dydžio mokymo imties poaibis, kurį sudaro tie elementai, kurie yra arčiausiai nagrinėjamojo su požymių vektoriumi x.

Artimiausių kaimynų poaibis priklauso nuo pasirinkto atstumo arba panašu-

mo mato. Dažniausiai naudojamas euklidinis atstumas

$$d(X,Y) = \left(\sum_{i=1}^{h} (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

bei vektorių kampo kosinuso panašumo matas

$$\rho(X,Y) = \frac{\sum_{i=1}^{h} x_i y_i}{\left(\sum_{i=1}^{h} x_i^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{h} y_i^2 \right)^{\frac{1}{2}}}. \tag{1.45}$$

Čia $X = (x_1, \ldots, x_h)$ ir $Y = (y_1, \ldots, y_h)$ – du dokumentus reprezentuojantys fiksuoto ilgio vektoriai.

Kaimyninių dokumentų skaičius k, kuris turi didelės įtakos klasifikavimo tikslumui dažniausiai parenkamas eksperimentiškai.

Algoritmas gali būti apibendrintas, kad atsižvelgtų ir į artimiausių dokumentų panašumą (artumą):

$$\widehat{\pi}(j,x) = \frac{\sum_{i=1}^{n} g(\rho(x, X(i))) \mathbb{1}_{\{\eta(i)=j\}}}{\sum_{i=1}^{n} g(\rho(x, X(i)))}.$$

Čia $g : \mathbb{R} \to \mathbb{R}$ yra nemažėjanti funkcija, jei ρ yra panašumo matas.

Neparametrinis k kaimynų algoritmas skirtingai nuo kitų aptartųjų šiame darbe neturi įprastos išankstinio mokymo fazės. Iš kitos pusės, kiekvienam naujai klasifikuojamam dokumentui atliekamas atskiras apmokymas pagal tai, į kuriuos mokymo imties dokumentus jis panašiausias. Šis algoritmas turi tą patį trūkumą kaip ir atraminių vektorių algoritmas: nėra galimybių gauti ir interpretuoti sąryšių tarp klasių ir terminų.

1.5. Pirmojo skyriaus išvados ir disertacijos uždavinių formulavimas

- Tyrimų, nagrinėjančių specifinių mokslo tekstų klasifikavimo problemą, yra ganėtinai nedaug, specialių algoritmų nėra, o tiek tyrimuose, tiek praktiniuose taikymuose dažniausiai naudojami įprastų kasdienės kalbos tekstų klasifikavimo algoritmai.
- Naudojant dokumentų reprezentavimo skaitinių požymių vektoriais bū-

dus, taikomasis tekstų klasifikavimo uždavinys gali būti suvestas į daugiamačių duomenų klasifikavimo uždavinį, kuriam spręsti galima naudoti diskriminantinės analizės metodus.

- Naudotinų klasifikavimo metodų yra daug, tačiau yra ganėtinai nedidelė grupė algoritmų, kurie kitų autorių tyrimuose sutinkami dažniausiai, o jų tikslumo lygis paprastai būna labai aukštas. Keletas tokių algoritmų detaliau išanalizuoti šiame skyriuje.

- Visi išnagrinėti algoritmai dėl vienų ar kitų trūkumų negali būti tiesiogiai pritaikyti suformuluotam uždaviniui spręsti. Pagrindinis geometrinių metodų trūkumas yra vidinių parametrų ir jų sąryšių nepasiekiamumas bei aiškios jų interpretacijos nebuvimas. Tikimybiniai algoritmai šia prasme pranašesni, tačiau jiems trūksta sprendimų, kaip atsižvelgti į ilgų tekstų specifiką ar panaudoti kontekstinę informaciją.

Atsižvelgiant į literatūros analizės rezultatus, formuluojami šie pagrindiniai disertacijos uždaviniai:

- sukurti tikimybinius diskriminantinės analizės metodus, kurie atsižvelgtų ir išnaudotų mokslo publikacijų specifiką;

- realių duomenų pagrindu ištirti ir palyginti pasiūlytus ir populiarius kitų autorių algoritmus.

2

Terminijos modeliai ir jų taikymas klasifikavimui

Šiame darbo skyriuje pateikiama siūloma mokslo publikacijų klasifikavimo metodika, pagrįsta tikimybiniais mokslo terminijos pasiskirstymo tekstuose modeliais bei jų statistine analize. Metodikos pagrindas yra euristinė idėja, kad jei teksto fragmentas priskiriamas tam tikrai klasei, tai yra nedidelis rinkinys mokslo terminų, kuriuos labai tikėtina sutikti šiame fragmente. Vadinasi, klasę galime nustatyti nagrinėdami, kokie terminai sutinkami tekste ir kiek jie tikėtini tos klasės dokumentuose.

Skyriaus tematika paskelbti šie autoriaus straipsniai: (Rudzkis et al., 2006, 2007; Balys et al., 2003, 2004b).

2.1. Žymėjimai ir sąvokos

Tegul K žymi mokslo dokumentų klasifikavimo sistemą, kuri apibrėžiama visomis tą sistemą sudarančiomis klasėmis, tiksliau – tų klasių žymėmis, kurios gali būti priskirtos klasifikuojamam tekstui, žr. (1.1). Dokumentas gali būti priskirtas kelioms klasėms vienu metu, žr. 1.1.4 skyrelį. Toliau visus klasifikuojamus mokslinius dokumentus vadinsime tiesiog straipsniais.

Tegul V yra mokslo srities terminų, kurie svarbūs klasifikavimo požiūriu, žodynas. Darome prielaidą, kad visų straipsnių klasifikavimo rezultatai priklau-

so tik nuo šių terminų iš aibės V. Chronologiškai sunumeruotas straipsnio a elementų vektorius

$$a = (a_1, a_2, \ldots, a_d), \quad d = d(a), \quad a_i \in V, \tag{2.1}$$

kuriame nebūtinai $a_i \neq a_j$, vadinamas straipsnio a projekcija. Nuo čia dokumentą reprezentuojantį požymių vektorių-projekciją žymėsime raide a, kad pabrėžtume, jog tai mokslinis straipsnis (*angl. article*). Kartais patogu straipsnio a projekciją sutapatinti su begaline seka $(a_1, a_2, \ldots)$, kurioje $a_i = 0$ visiems $i > d(a)$. Čia $0 \in V$ žymi papildomą *nulinį terminą*, kuris realybėje neegzistuoja. Tegul A žymi visų mokslo srities straipsnių projekcijų aibę. Toliau žodį *projekcija* praleisime ir elementą $a \in A$ vadinsime tiesiog straipsniu.

Klasifikavimo požiūriu straipsnis nebūtinai yra homogeniškas tekstas – bendru atveju jį sudaro $m = m(a) \geq 1$ vientisų dalių, kurios klasifikuojamos skirtingai sistemoje K. Nepersikertantys indeksų intervalai $I_j(a) \subset \{1, 2, \ldots, d(a)\} = N(a)$ ir klasių žymės $w_j(a) \in K, j = \overline{1, m}$ atitinka kiekvieną iš šių dalių. Čia $\bigcup_{j=1}^{m} I_j(a) = N(a)$ ir $w_j \neq w_{j-1}, j = \overline{2, m}$: jei dvi gretimos teksto dalys priskiriamos tai pačiai klasei, jos apjungiamos į vieną.

Tegul N žymi natūraliųjų skaičių aibę. Tegul straipsnis $a \in A$ ir indeksų aibė $I \subset N$ parenkami atsitiktinai taip, kad straipsnio dalis $\{(a_\tau, \tau), \tau \in I\}$ yra homogeniška: $I \subset I_\nu(a), \quad \nu \in \{1, \ldots, m\}$. Ši dalis priskiriama klasei $\eta = w_\nu(a)$ sistemoje K. Kaip įprastai, klasifikavimo problema yra nustatyti nežinomą klasę η naudojantis stebėtu vektoriumi $a_I = (a_\tau, \tau \in I)$.

Pastaba: Akivaizdu, kad nagrinėjant tik (tam tikrus) mokslinius terminus iš teksto prarandama dalis svarbios informacijos – ne terminai, struktūrizavimo informacija ir pan. Skyrelyje 2.4 aptarta, kaip panaudoti dalį šios kontekstinės informacijos.

2.2. Tikimybiniai mokslo terminijos pasiskirstymo tekste modeliai

2.2.1. Bendras modelis

Kadangi (a, I, η) yra atsitiktinio eksperimento rezultatas, aibėje K apibrėžtas skirstinys

$$Q(w) = \mathbb{P}\{\eta = w\}, \quad w \in K. \tag{2.2}$$

$Q(w)$ yra apriorinė tikimybė, kad atsitiktinis tekstas priskiriamas klasei w.

Tegul Y žymi visų galimų a_I reikšmių aibę. Aibėje Y apibrėžti šie sąlyginiai tikimybiniai skirstiniai:

$$P(y) = \mathbb{P}\{a_I = y \big|\ |I| = d(y)\},$$

$$P(y|w) = \mathbb{P}\{a_I = y \big|\ |I| = d(y), \eta = w\}, \quad w \in K, \tag{2.3}$$

čia $d(y) = dim\ y, |I| = card\ I$.

Pažymėkime

$$Q(w|y) = \mathbb{P}\{\eta = w|a_I = y\}.$$

Jei η ir $|I|$ yra nepriklausomi, stebėjus a_I, aposteriorinė atsitiktinio įvykio $\{\eta = w\}$ tikimybė apibrėžiama lygybe

$$Q(w|a_I) = Q(w) \cdot \psi_w(a_I),$$

čia

$$\psi_w(y) = P(y|w)/P(y), y \in Y. \tag{2.4}$$

Funkcionalas ψ_w matematiškai apibrėžia klasės w identifikacinio debesėlio sąvoką (Hazewinkel, 2005, 1999). Jis atspindi, kaip tikimybė stebėti atsitiktinį tekstą priklauso nuo to, kuriai klasei šis tekstas priskiriamas. Akivaizdu, kad ψ_w taip pat priklauso nuo poros (a, I) skirstinio, todėl bendresnis identifikacinio debesėlio apibrėžimas būtų aibėje Y apibrėžtų funkcionalų šeima $\Psi_w = \{\psi_w(\cdot|H), H \subset N\}$, čia $\psi_w(y|H)$ apibrėžiamas analogiškai kaip $\psi_w(y)$ su sąlyga $I = H$. Toliau laikysime, kad (a, I) skirstinys yra fiksuotas.

Naudodami (2.3) ir (2.2) įvestus skirstinius, apibrėžiame Bajeso klasifikatorių, kuris minimizuoja vidutinius klasifikavimo nuostolius. Tegul $l(v, w)$ žymi nuostolių dydį, jei tekstas iš klasės v priskiriamas klasei w. Kaip paprastai, $l(\cdot, \cdot) \geq 0$ ir $l(w, w) = 0$. Jei klasifikuojama pagal stebėtą a_I, Bajeso klasifikavimo taisyklė nusakoma lygybe

$$\widehat{\eta} = \arg\min_{w \in K} \sum_{v \in K} Q(v) P(a_I|v) l(v, w). \tag{2.5}$$

Jei nuostolių funkcija triviali, t. y.,

$$l(v, w) = \begin{cases} c, w \neq v, \\ 0, w = v, \end{cases}$$

iš (2.5) išplaukia, kad

$$\widehat{\eta} = \arg\max_{w \in K} Q(w) P(a_I|w). \tag{2.6}$$

Toliau formules išrašinėsime tik trivialios nuostolių funkcijos atveju. (2.6) formulėje, $\psi_{(\cdot)}(a_I)$ gali būti įstatytas vietoje $P(a_I|\cdot)$:

$$\widehat{\eta} = \arg\max_{w \in K} Q(w) \psi_w(a_I). \tag{2.7}$$

2.2.2. Atskiri modelio atvejai, galiojant tam tikroms prielaidoms apie skirstinius

Norint taikyti klasių nustatymo formules (2.5), (2.6) ir (2.7) praktikoje tektų statistiškai įvertinti daugiamačius skirstinius $P(y|w), y \in Y$, o tai reikalautų labai didelės apimties mokymo duomenų. Toliau pateikiamos praktiškesnės skirstinių išraiškos, gaunamos galiojant tam tikroms prielaidoms apie straipsnių vektoriaus komponenčių tarpusavio priklausomybės charakteristikas.

Tegul indeksas $\tau \in I$ yra atsitiktinis dydis. Aibėje V apibrėžtas skirstinys

$$P(v) = \mathbb{P}\{a_\tau = v\}$$

bei sąlyginis skirstinys

$$P(v|w) = \mathbb{P}\{a_\tau = v|\eta = w\}, \quad w \in K.$$

Prielaida 1 (sąlyginis stacionarumas ir nepriklausomumas). Tegul visiems $y \in Y$ ir $w \in K$ galioja lygybė

$$P(y|w) = \prod_{i=1}^{d} P(y_i|w), \tag{2.8}$$

čia $d = d(y)$ kaip ir anksčiau yra vektoriaus y dimensija.

Šiuo atveju identifikacinio debesėlio apibrėžimas (2.4) gali būti pakeistas į

$$\psi_w(v) = P(v|w)/P(v), v \in V, w \in K, \tag{2.9}$$

o Bajeso klasifikavimo taisyklė dabar apibrėžiama lygybe

$$\widehat{\eta} = \arg\max_{w \in K} \left[Q(w) \prod_{\tau \in I} P(a_\tau|w) \right]. \tag{2.10}$$

(2.10) yra naivaus Bajeso klasifikavimo taisyklė, (žr. 1.4.1). Joje $\psi_{(\cdot)}(a_\tau)$ gali būti įstatytas vietoje $P(a_\tau|\cdot)$:

$$\widehat{\eta} = \arg\max_{w\in K}\left[Q(w)\prod_{\tau\in I}\psi_w(a_\tau)\right]. \tag{2.11}$$

Identifikacinio debesėlio apibrėžimas, paremtas prielaida (2.8) ignoruoja informaciją, kurią būtų galima gauti iš terminų tarpusavio išsidėstymo tekste. Todėl toliau įvesime silpnesnę prielaidą.

Prielaida 2 (sąlyginis stacionarumas ir markoviškumas). Tegul visiems $y \in Y$ ir $w \in K$ galioja lygybė

$$P(y|w) = P(y_1|w)\prod_{i=1}^{d-1}\left[P(y_i, y_{i+1}|w)/P(y_i|w)\right], \tag{2.12}$$

čia $P(v, u|w) = \mathbb{P}\{a_\tau = v, a_{\tau+1} = u|\eta = w\}$.

Taigi, laikoma, kad a_I yra Markovo grandinė, įgyjanti reikšmes iš V. Šiuo atveju identifikacinis debesėlis nusakomas dviem funkcionalais: $\psi_w(v)$, apibrėžtu (2.9) ir

$$\psi_w(v, u) = P(v, u|w)/P(v, u), v, u \in V, \tag{2.13}$$

čia $P(v, u) = \mathbb{P}\{a_\tau = v, a_{\tau+1} = u\}$.

Tegul $I = \{r, r + 1, \ldots, m\}$. Tada Bajeso klasifikavimo taisyklė gaunama pertvarkius lygybę (2.6) atsižvelgiant į lygybes (2.9), (2.12) ir (2.13):

$$\widehat{\eta} = \arg\max_{w\in K}\left[Q(w)P(a_r|w)\prod_{i=r}^{m-1}\left[P(a_i, a_{i+1}|w)/P(a_i|w)\right]\right].$$

Čia $\psi_{(\cdot)}(v)$ ir $\psi_{(\cdot)}(v, u)$ gali būti įstatyti vietoje $P(v|\cdot)$ ir $P(v, u|\cdot)$:

$$\widehat{\eta} = \arg\max_{w\in K}\left[Q(w)\psi_w(a_r)\prod_{i=r}^{m-1}\left[\psi_w(a_i, a_{i+1})/\psi_w(a_i)\right]\right]. \tag{2.14}$$

2.4 skyrelyje aptariami suformuluotų ganėtinai paprastų modelių apibendrinimai, kai atsižvelgiama į papildomą kontekstinę informaciją, susijusią su terminų pozicijomis straipsnio tekste ir kontekstu tarp jų.

2.3. Modelių identifikavimo metodai

Sprendžiant teksto klasifikavimo problemą skirstiniai P, Q ir funkcionalas $\psi_w(\cdot)$, naudojami (2.5), (2.6) ir (2.7), paprastai yra nežinomi. Čia gali būti panaudotas tiesioginio įstatymo metodas – nežinomi skirstinių parametrai pakeisti jų statistiniais įvertiniais.

Tarkime, turime stebėtus n homogeninių tekstų fragmentus bei jų klasifikavimo rezultatus. Paprastumo dėlei šiuos tekstų fragmentus vadinsime tiesiog straipsniais ir laikysime, kad stebėtos dalys yra nuoseklios, o jų pradžia sutampa su straipsnių pradžia. Todėl mokymo imtis X_n (plg. (1.8)) sudaryta iš n porų

$$X_n = (y(1), \eta(1)), \ldots, (y(n), \eta(n)), \quad \eta(i) \in K, y(i) \in Y. \tag{2.15}$$

Čia $Y = \{y = (y_1, \ldots, y_d) : y_i \in V, d \in N\}$.

2.3.1. Identifikavimas bendro modelio atveju

Norėdami klasifikavimui taikyti bendrą formulę (2.7), turime įvertinti funkcionalą $\psi_w(\cdot)$ bei skirstinį $Q(\cdot)$. Empirinis $Q(w)$ analogas apibrėžiamas lygybe

$$\widehat{Q}(w) = \frac{1}{n} \sum_{j=1}^{n} 1_{\{\eta(j)=w\}}. \tag{2.16}$$

Kur kas sudėtingiau įvertinti $\psi_w(y)$, $y \in Y, w \in K$, nes menkai tikėtina, kad kiekvienam stebėtam straipsnio terminų vektoriui y bus dar bent keletas stebinių su tokiu pat vektoriumi. Šioje vietoje galima panaudoti įprastą k kaimynų metodą. Pirma apibrėšime elementų iš Y atstumo matą: tegul visiems $y, z \in Y$, $\rho(y, z)$ yra neneigimas funkcionalas, kurio reikšmes vadinsime atstumu nuo elemento z iki elemento y. Fiksuotam $y \in Y$ pasirenkame k artimiausių kaimynų iš imties. Tegul $J_k(y) \subset \{1, \ldots, n\}$ žymi k elementų aibę, sudarytą iš indeksų tų stebinių, kuriems atstumas iki y yra mažiausias. Tada $\psi_w(y)$ įvertinys apibrėžiamas lygybe

$$\widehat{\psi}_w(y) = \frac{1}{\widehat{Q}(w) \cdot k} \sum_{j \in J_k(y)} \mathbb{1}_{\{\eta(j)=w\}}. \tag{2.17}$$

Čia $0/0 = 1$, o kintamasis $k = k(n)$ priklauso nuo imties dydžio ir jam galioja sąlyga $k \to \infty, k/n \to 0$, kai $n \to \infty$.

Natūralu apibrėžti atstumą tarp straipsnių y ir z tokį, kuris atsižvelgtų tiek į mokslo terminų dažnius, tiek į jų pozicijas. Siūlome atstumą $\rho(y, z)$ kiekvienam

$y = (y_1, \ldots, y_{d(y)}) \in Y$ ir $z = (z_1, \ldots, z_{d(z)}) \in Y$ apibrėžti lygybe

$$\rho(y, z) = \sum_{v \in V} [|\alpha_y(v) - \alpha_z(v)| + c|\beta_y(v) - \beta_z(v)|], \tag{2.18}$$

$$\alpha_x(v) = \frac{1}{d} \sum_{i=1}^{d} \mathbb{1}_{\{x_i = v\}},$$

$$\beta_x(v) = \frac{1}{\alpha_x(v) d^2} \sum_{i=1}^{d} i \cdot \mathbb{1}_{\{x_i = v\}}.$$

Čia $\alpha_x(v)$ yra termino v dažnis vektoriuje $x \in Y$, $\beta_x(v) \in [0, 1]$ yra standartizuotas termino v pozicijų vektoriuje x sunkio centras, $d = d(x)$ yra vektoriaus ilgis, o $c \geq 0$ yra funkcionalo β svoris.

Įstatę įvertinius (2.16) ir (2.17) vietoje atitinkamų nežinomų tikrų reikšmių į (2.7), gauname k artimiausių kaimynų klasifikavimo metodą.

2.3.2. Identifikavimas modelio su įvestomos prielaidomis atvejais

Norėdami taikyti klasifikavimo taisykles (2.11) ar (2.14), apibrėžtas galiojant atitinkamai sąlygoms (2.8) bei (2.12), vėlgi turime įvertinti skirstinį $Q(w)$ bei funkcionalus $\psi_w(\cdot)$ ir $\psi_w(\cdot, \cdot)$. $Q(w)$ vertinamas taip pat, kaip (2.16). Vertindami funkcionalus ψ pirma suskaičiuosime empirinius skirstinių $P(\cdot)$, $P(\cdot, \cdot)$, $P(\cdot|\cdot)$ ir $P(\cdot, \cdot|\cdot)$ įvertinius:

$$\widetilde{P}(v) = \sum_{j=1}^{n} \sum_{k=1}^{d(j)} \mathbb{1}_{\{y_k(j) = v\}} / \sum_{j=1}^{n} d(j), \tag{2.19}$$

$$\widetilde{P}(v, u) = \sum_{j=1}^{n} \sum_{k=1}^{d(j)-1} \mathbb{1}_{\{y_k(j) = v, y_{k+1}(j) = u\}} / \sum_{j=1}^{n} (d(j) - 1), \tag{2.20}$$

$$\widetilde{P}(v|w) = \sum_{j=1}^{n} \sum_{k=1}^{d(j)} \mathbb{1}_{\{y_k(j) = v, \eta(j) = w\}} / \sum_{j=1}^{n} d(j) \mathbb{1}_{\{\eta(j) = w\}}, \tag{2.21}$$

$$\widetilde{P}(v,u|w)=\sum_{j=1}^{n}\sum_{k=1}^{d(j)-1}\mathbb{1}_{\{y_k(j)=v,y_{k+1}(j)=u,\eta(j)=w\}}/\sum_{j=1}^{n}(d(j)-1)\mathbb{1}_{\{\eta(j)=w\}}. \tag{2.22}$$

Čia $d(j)=d(y(j))$ kaip ir anksčiau žymi vektoriaus $y(j)$ dimensiją.

Empirinius identifikacinius debesėlius $\widetilde{\psi}_w(v)$ ir $\widetilde{\psi}_w(v,u)$ gauname įstatę suskaičiuotus įvertinius vietoje atitinkamų tikrųjų reikšmių formulėse (2.9) ir (2.13):

$$\widetilde{\psi}_w(v)=\widetilde{P}(v|w)/\widetilde{P}(v), v\in V, w\in K, \tag{2.23}$$

$$\widetilde{\psi}_w(v,u)=\widetilde{P}(v,u|w)/\widetilde{P}(v,u), v,u\in V. \tag{2.24}$$

2.3.3. Įvertinių modifikavimas ir informatyviausių terminų atrinkimo būdai

Prieš naudodami įvertinius (2.23) ir (2.24) klasifikavimui atliksime tam tikras modifikacijas, susijusias su stebėjimų skaičiumi ir informatyvumu. Aptarsime šias modifikacijas funkcionalui $\psi_{(\cdot)}(\cdot)$ ($\psi_{(\cdot)}(\cdot,\cdot)$ – analogiškai).

Pirma modifikuosime tas (2.23) gauto empirinio $\widetilde{\psi}_w(v)$ reikšmes, kurios buvo gautos naudojant per mažai stebėjimų. Tegul $\nu(v)$ žymi termino v stebėjimų skaičių imtyje (2.15), $\nu_w(v)$ – atitinkamai termino v stebėjimų skaičių tuose imties vektoriuose, kurie priskirti klasei w, o $\mu\in N$ tegul žymi minimalų stebėjimų skaičių, reikalingą, kad galėtume skaičiuoti (patikimus) įvertinius $\widetilde{\psi}_w(v)$. Tegul

$$V_w=\{v:\nu(v)\geq\mu,\nu_w(v)\neq 0\}, \tag{2.25}$$

$$\underline{\psi}_w=\min_{v\in V_w}\widetilde{\psi}_w(v),$$

$$\overline{\psi}_w=\max_{v\in V_w}\widetilde{\psi}_w(v).$$

Funkcionalą

$$\widetilde{\psi}^*_w(v)=(\widetilde{\psi}_w(v)\vee\underline{\psi}_w)\wedge\overline{\psi}_w$$

toliau naudosime vietoje $\widetilde{\psi}_w(v)$. Čia $\vee$ ir $\wedge$ kaip įprastai žymi maksimumo ir minimumo operatorius.

Funkcionalas $\psi_w(\cdot)$ apibrėžia aibės V elementų išrikiavimą kiekvienai klasei $w\in K$: $V(w)=(v_1,\ldots,v_h)$, $h=|V|$ ir galioja nelygybės

$$\psi_w(v_1)\geq\psi_w(v_2)\geq\ldots\geq\psi_w(v_h). \tag{2.26}$$

Išrikiuokime aibės V elementus, naudodami įvertinį $\widetilde{\psi}^*_w(\cdot)$ vietoje nežinomo $\psi_w(\cdot)$. Kadangi h reikšmė gali būti labai didelė, o imčių dydis paprastai nėra didelis, kiekvienai klasei paliksime tik dalį informatyviausių terminų:

$$\widehat{\psi}_w(v_k) = \begin{cases} \widetilde{\psi}^*_w(v_k), \text{ jei } k \in L, \\ 1, \text{ kitu atveju.} \end{cases} \tag{2.27}$$

Čia, aišku, $L = L(w)$, t. y., informatyviausių terminų aibės skirtingos kiekvienai klasei. Identifikacinį debesėlį sudaro tik tie terminai v_k, $k \in L \subset \{1, \ldots, h\}$ kuriems $\widetilde{\psi}^*_w(v_k)$ reikšmė ženkliai skiriasi nuo 1. Aibę L sudaro du poaibiai, sudaryti atitinkamai iš indeksų tų terminų, kuriems $\widetilde{\psi}^*_w(v_k)$ ženkliai didesnis už 1 bei indeksų tų terminų, kuriems šis dydis ženkliai mažesnis už 1. Šiuos poaibius žymėsime atitinkamai $\overline{L}$ ir $\underline{L}$, tad

$$L = \overline{L} \cup \underline{L}. \tag{2.28}$$

Aibę L galima parinkti keletu būdu, toliau pateikiami siūlomi variantai.

Pirmasis ir pats paprasčiausias metodas *fixed*: parenkamas fiksuotas (nustatomas iš apriorinių samprotavimų arba atliekant eksperimentus) skaičius terminų, atitinkančių didžiausias ir mažiausias $\widetilde{\psi}^*_w(\cdot)$ reikšmes. Tada

$$\begin{aligned} \overline{L} &= \{1, \ldots, s\}, \\ \underline{L} &= \{h - l, \ldots, h\}, \end{aligned} \tag{2.29}$$

čia s ir l – fiksuoti skaičiai.

Toliau suformuluosime būdus, paremtus statistine hipotezių tikrinimo teorija. Nagrinėkime hipotezę

$$H_0 : \ \psi_w(v) = 1 \tag{2.30}$$

su alternatyva

$$H_1 : \psi_w(v) > 1$$

ir tegul $\overline{\alpha}(v)$ žymi p–reikšmę.

Analogiškai, nagrinėkime tą pačią (2.30) hipotezę, bet šįsyk su alternatyva

$$H_1 : \psi_w(v) < 1$$

ir tegul $\underline{\alpha}(v)$ žymi p–reikšmę.

Tegul α žymi pasirinktą reikšmingumo lygį. Informatyviausių terminų in-

deksų aibės L nustatymo metodas *hyp* apibrėžiamas lygybėmis

$$\overline{L} = \{k : \ \widetilde{\psi}^*_w(v_k) > 1, \overline{\alpha}(v_k) < \alpha\},$$

$$\underline{L} = \{k : \ \widetilde{\psi}^*_w(v_k) < 1, \underline{\alpha}(v_k) < \alpha\}. \tag{2.31}$$

Skaičiavimo prasme greitesnis, bet mažiau tikslus būdas *hyp/stop* apibrėžiamas (2.29) lygybėmis, kuriose s ir l reikšmės dabar nustatomos iš lygybių

$$s = \max\{k : \quad \overline{\alpha}(v_j) < \alpha, j = \overline{1,k}\}, \tag{2.32}$$

$$l = \max\{k : \quad \underline{\alpha}(v_{h-j}) < \alpha, j = \overline{1,k}\}. \tag{2.33}$$

Paskutinysis siūlomas informatyviausių terminų nustatymo metodas, kurį žymėsime *hyp/fixed*, derinantis *hyp* ir *fixed*, apibrėžiamas taip pat (2.29) lygybėmis, o s ir l reikšmės parenkamos

$$s = |\{k : \ \widetilde{\psi}^*_w(v_k) > 1, \overline{\alpha}(v_k) < \alpha\}|, \tag{2.34}$$

$$l = |\{k : \ \widetilde{\psi}^*_w(v_k) < 1, \underline{\alpha}(v_k) < \alpha\}|. \tag{2.35}$$

Šis metodas panašus į *hyp* metodą tuo, kad jie abu parenka tą patį skaičių informatyviausių terminų, tačiau patys parinktieji terminai gali skirtis.

Toliau nagrinėsime, kaip nustatyti $\overline{\alpha}(\cdot)$ ir $\underline{\alpha}(\cdot)$ reikšmes, naudojamas informatyviausių terminų nustatymo metoduose. Tarkime, H_0 teisinga ir indikatoriai $\mathbb{1}_{\{a_\tau = v\}}$, $\tau = 1, 2, \ldots, d$ yra sąlyginai nepriklausomi bei vienodai pasiskirstę, įgyjantys reikšmę 1 su tikimybe $P(v) \stackrel{def}{=} p$, jei galioja sąlyga $\eta = w$. Tada sąlyginis empirinės tikimybės $\widetilde{P}(v|w)$ skirstinys prie sąlygos $\sum_{j=1}^{n} d(j)\mathbb{1}_{\{\eta(j)=w\}} = m$ yra binominis, o kritinis reikšmingumo lygis $\overline{\alpha}(v)$ apibrėžiamas lygybe

$$\overline{\alpha}(v) = \sum_{m_0 \le k \le m} \binom{k}{m} p^k (1-p)^{m-k},$$

čia

$$m_0 = \sum_{j=1}^{n} \sum_{k=1}^{d(j)} \mathbb{1}_{\{y(j)_k = v, \eta(j) = w\}}.$$

Analogiškai $\underline{\alpha}(v)$ apibrėžiamas lygybe

$$\underline{\alpha}(v) = \sum_{0 \le k \le m_0} \binom{k}{m} p^k (1-p)^{m-k}.$$

p reikšmė paprastai nežinoma, todėl vietoje jos gali būti naudojama $\widetilde{P}(v)$.

2.3.4. Siūlomas parametrinio skirstinių vertinimo būdas

Neparametrinis skirstinių vertinimas yra labai jautrus mokymo imties dydžiui. Kai imtys ganėtinai mažos, įvertiniai bus nepatikimi, jei naudojamas k artimiausių kaimynų ar kitas neparametrinis vertinimo algoritmas. Todėl dabar pateiksime parametrinį funkcionalo $\psi_w(\cdot)$ vertinimą, kai galioja prielaida apie sąlyginį nepriklausomumą ir stacionarumą (2.8).

Naudosime tą patį (2.26) išdėstymą ir modifikuotuosius įvertinius (2.27) parametrinio modelio funkcionalui $\psi_w(\cdot)$ sudarymui. Pažymėkime $s = |\overline{L}|$ ir $l = |\underline{L}|$. Išrikiuojame $\overline{L}$ ir $\underline{L}$ elementus didėjimo tvarka:

$$\overline{L} = \{i_1, \ldots, i_s\}, \quad i_1 < \ldots < i_s,$$

$$\underline{L} = \{j_1, \ldots, j_l\}, \quad j_1 < \ldots < j_l.$$

Kadangi taikant informatyviausių terminų atrinkimo metodą *hyp* (2.31), šiuose išrikiuotose indeksų sąrašuose gali likti skylių (t. y., kai kurie indeksai prašokami), o mes norime apibrėžti parametrinį modelį, priklausantį tik nuo termino eilės numerio, apibrėšime numeruojančius atvaizdžius

$$\overline{e}(i_k) = k, \ 1 \le k \le s \quad \text{ir} \quad \underline{e}(j_k) = k, \ 1 \le k \le l. \tag{2.36}$$

Dabar norėdami įvertinti $\psi_w(v)$, pasirinkime parametrinių funkcijų klasę. Tegul

$$\gamma(k, \theta) = \theta_0 + \theta_1 k^{-\theta_2}, \quad \theta = (\theta_0, \theta_1, \theta_2). \tag{2.37}$$

Tada

$$\log \psi_w(v_k) = \begin{cases} \gamma(\overline{e}(k), \theta), & k \in \overline{L}, \\ \gamma(\underline{e}(k), \theta^*), & k \in \underline{L}. \end{cases} \tag{2.38}$$

Parametrų θ ir θ^* įverčiai randami iš lygybių

$$\widehat{\theta} = arg \min_{\theta} \sum_{k \in \overline{L}} \left| \log \widehat{\psi}_w(v_k) - \gamma(\overline{e}(k), \theta) \right|^{\beta}, \tag{2.39}$$

$$\widehat{\theta}^* = arg \min_{\theta} \sum_{k \in \underline{L}} \left| \log \widehat{\psi}_w(v_k) - \gamma(\underline{e}(k), \theta) \right|^{\beta}, \tag{2.40}$$

čia $\beta = 2$ arba $\beta = 1$.

Kai $\beta = 2$, turime įprastą mažiausių kvadratų metodą (MKM), kurio įverčių skaičiavimo procedūros realizuotos bet kuriame didesniame statistiniame pakete. Deja, MKM įverčiai yra gana jautrūs empirinių $\log \widehat{\psi}_w(\cdot)$ reikšmių nukrypimams nuo teorinių $\gamma(k, \theta)$, todėl kartais naudojamas robastiškesnis medianinis metodas ($\beta = 1$). Jei medianinis metodas nerealizuotas naudojamuose statistiniuose paketuose, θ gali būti vertinamas iteratyvia procedūra

$$\widehat{\theta}(j+1) = arg \min_{\theta} \sum_{k \in \overline{L}} \frac{(\log \widehat{\psi}_w(v_k) - \gamma(\overline{e}(k), \theta))^2}{|\log \widehat{\psi}_w(v_k) - \gamma(\overline{e}(k), \widehat{\theta}(j))|}, \quad j = 0, 1, \ldots. \tag{2.41}$$

Čia $\widehat{\theta}(0)$ yra MKM įvertis.

Kitas euristinis variantas, kuris buvo naudotas ir šiame darbe, – taikyti formalią MKM procedūrą, tačiau ne skirtumams $|\log \widehat{\psi}_w(v_k) - \gamma(\overline{e}(k), \theta)|$, o kvadratinėms jų šaknims.

Pastebėsime, kad parametrinio modelio naudojimas gali būti argumentuotas ne tik jau minėtu įvertinių stabilumo užtikrinimu. Praktikoje tikėtina situacija, kad klasifikavimo taisykles nusakančių identifikacinių debesėlių sąrašai bus konstruojami ir prižiūrimi ne tik automatinėmis procedūromis, bet ir įsikišant srities ekspertams. Turint parametrinį modelį, naujam terminui prijungti prie tam tikrą klasę identifikuojančių terminų sąrašo užtenka nurodyti tik jo poziciją pagal spėjamai nešamą diskriminantinę informaciją. Lygiai taip pat nesudėtinga jau turimą identifikacinį debesėlį perrikiuoti, kad naujieji terminų svoriai atitiktų specifinius poreikius. Taisyklės, kaip (automatiškai) perskaičiuoti svorius po tokių modifikacijų, kaip termino įterpimas, sukeitimas vietomis ar pašalinimas, šiame darbe nenagrinėtos.

2.4. Papildoma kontekstinė informacija ir jos naudojimas klasifikavimui

2.4.1. Argumentacija

Kaip minėta 2.1 skyrelio pabaigoje esančioje pastaboje, naudodami tik tam tikrus mokslo terminus iš dokumento teksto, prarandame dalį informacijos. Toliau panagrinėsime, kokią papildomą informaciją ir kaip būtų galima panaudoti konstruojant klasifikavimo algoritmus.

Pastebėsime, kad vertinant ir naudojant klasifikavimui stacionarų skirstinį $P(v|w)$ (žr. (2.10) ir (2.21)) neatsižvelgiama į termino v padėtį tekste. Prielaida, kad (ypač didelės apimties tekstuose) termino įtaka diskriminavimui priklauso nuo to, kur jis yra, atrodo visiškai natūrali. Logiška manyti, kad tam tikras terminas, vienos klasės mokymo dokumentuose sutinkamas santraukoje ir įvade, o kitos klasės dokumentuose – įrodymuose ar 15-ame puslapyje, yra stipresnis pirmosios klasės indikatorius. Lygiai taip pat natūralu manyti, kad tas pats pirmajai klasei labiau būdingas terminas, sutiktas klasifikuojamojo teksto pradžioje yra informatyvesnis diskriminavimo prasme, nei sutiktas straipsnio antroje pusėje.

Analogiškai, skirstiniai $P(v, u)$ ir $P(v, u|w)$ nepriklauso ne tik nuo atitinkamų terminų v ir u pozicijų straipsnyje, bet ir nuo konteksto, esančio tarp jų. Tai reiškia, kad situacija, kai šie gretimi terminai yra išties vienas šalia kito straipsnio tekste bei situacija, kai jie yra skirtinguose sakiniuose ar net skirtingose pastraipose traktuojamos visiškai vienodai. Iš kitos pusės, atsižvelgdami tik į gretimų terminų poras, susidursime su statistinio vertinimo problemomis. Tikėtina, kad klasifikuojant konkretų straipsnį, dalis jo tekste aptinkamų gretimų terminų porų bus nė karto nesutiktos mokymo imtyje, kas duotų nulines tikimybių įverčių reikšmes. Tokiu atveju galima taikyti analogiškas glodinimo procedūras, kaip ir nepriklausomumo atveju, tačiau jei nestebėtų porų bus ganėtinai daug, aposteriorinių priklausymo klasėms tikimybių įverčiai greičiausiai bus labai nestabilūs (toks efektas išties stebėtas eksperimentinio tyrimo metu). Todėl natūralu, kad tikėtis realaus griežtai suformuluotos Markovo savybės naudojimo įtakoto rezultatų pagerėjimo neverta. Remiantis euristiniais samprotavimais, Markovo savybę galima susilpninti – mokymo fazėje atsižvelgti ne tik į gretimų terminų poras, bet ir į nutolusias poras, skiriamas kitų terminų, taip padidinant stebėtų porų skaičių. Šis susilpninimas argumentuojamas tuo, kad tam tikroje samprotavimų grandinėje daugiau ar mažiau susiję didžioji dalis terminų (nebūtinai vienodai stipriai), o ne tik gretimi, be to, griežtos kaimynystės sąlyga itin jautri terminų žodynui – jį papildžius, anksčiau buvę gretimi terminai staiga tampa nebe gretimi

ir netenka tiesioginės įtakos klasifikavimui. Naujo straipsnio klasifikavimo metu taip pat logiška atsižvelgti ne tik į pačias poras, bet ir kontekstą tarp jų.

2.4.2. Papildomos informacijos apibrėžimas

Norėdami atsižvelgti į aukščiau įvardintus su kontekstu susijusius aspektus bei atitinkamai pakeisti skirstinių vertinimo bei klasifikavimo metodus, įsivesime papildomos informacijos duomenų struktūrą. Kiekvienam straipsnio moksliniam terminui apibrėšime jo poziciją tekste, t. y., greta jo eilės numerio straipsnio vektoriuje (viena iš pozicijos išraiškų) pridėsime tam tikrą informacijos elementą, kuris tiksliau nusakytų jo vietą kontekste. Poziciją tekste galima apibrėžti įvairiai, mes siūlome ją nusakyti termino, žodžio, sakinio ir pastraipos numeriais, skaičiuojant nuo teksto pradžios. Straipsnio projekciją (2.1) perapibrėžiame:

$$a = ((a_1, \lambda_1), (a_2, \lambda_2), \ldots, (a_d, \lambda_d)), \quad d = d(a), \quad a_i \in V.$$

Čia

$$\lambda_i = (\lambda_i^{(t)}, \lambda_i^{(w)}, \lambda_i^{(s)}, \lambda_i^{(p)}), \quad \lambda_i^{(\cdot)} \in N, \tag{2.42}$$

o $\lambda_i^{(t)}, \lambda_i^{(w)}, \lambda_i^{(s)}, \lambda_i^{(p)}$ žymi atitinkamai i–ojo straipsnio a vektoriaus elemento termino (*t –term*), žodžio (*w – word*), sakinio (*s – sentence*) ir pastraipos (*p – paragraph*) numerius. Akivaizdu, kad $\lambda_i^{(t)} = i$ ir $\lambda_{(\cdot)}^{(w)} \geq \lambda_{(\cdot)}^{(s)} \geq \lambda_{(\cdot)}^{(p)}$.

Natūralu būtų atsižvelgti ir į aukštesnio lygio straipsnio teksto struktūrizaciją, išskiriant tokias dalis kaip įvadas, pagrindinių rezultatų formulavimas, įrodymai ir pan. Pasiūlytas papildomos informacijos apibrėžimas galėtų būti paprastai praplėstas, kad atsižvelgtų ir į šią struktūrą. Šiame darbe to nėra daroma dėl to, kad turimuose duomenyse atskirti tokius loginius elementus tiesiog nebuvo įmanoma.

2.4.3. Siūlomi papildomos informacijos naudojimo metodai

Apibrėžę papildomos informacijos elementą, galime grįžti prie šio skyrelio pradžioje įvardintų su kontekstu susijusių aspektų ir panagrinėti, kaip į juos atsižvelgti.

Apibrėžkime intervale $[0, 1]$ reikšmes įgyjančius funkcionalus $\sigma = \sigma(\lambda_{(\cdot)})$ bei $\delta = \delta(\lambda_{(\cdot)}, \lambda_{(\cdot)})$, kurių pirmasis apibūdina termino ar terminų poros *svorį* priklausomai nuo pozicijos, t. y., atstumo nuo teksto pradžios, o antrasis apibūdina terminų poros sąryšio *svorį* (stiprumą) priklausomai nuo jų pozicijų, t. y., konteksto tarp jų.

Tarkime, turime papildoma informacija praturtintą mokymo imtį (plg. (2.15))

$$X_n = (y(1), \lambda(1), \eta(1)), \ldots, (y(n), \lambda(n), \eta(n)), \quad \eta(i) \in K, y(i) \in Y, \quad (2.43)$$

čia $\lambda(j) = (\lambda_1(j), \ldots, \lambda_d(j))$, $d = d(y(j))$, $\lambda_{(i)}(j)$ – (2.42) apibrėžta papildoma informaciją apie i–ąjį j–ojo mokymo dokumento mokslo terminą.

Turėdami aukščiau įvestus funkcionalus ir praturtintą imtį, galime apibrėžti svertinius skirstinių $P(\cdot)$, $P(\cdot,\cdot)$, $P(\cdot|\cdot)$ ir $P(\cdot,\cdot|\cdot)$ įvertinius. Pažymėkime

$$z(j,k,l) = \sigma(\lambda_k(j)) \cdot \delta(\lambda_k(j), \lambda_l(j)),$$

$$S(j) = \sum_{k=1}^{d(j)} \sigma(\lambda_k(j)), \quad S_n = \sum_{j=1}^{n} S(j),$$

$$S_n^* = \sum_{j=1}^{n} S(j) \mathbb{1}_{\{\eta(j)=w\}},$$

$$Z(j) = \sum_{k=1}^{d(j)-1} \sum_{l=k+1}^{d(j)} z(j,k,l), \quad Z_n = \sum_{j=1}^{n} Z(j),$$

$$Z_n^* = \sum_{j=1}^{n} Z(j) \mathbb{1}_{\{\eta(j)=w\}}.$$

Tada apibrėžiame

$$\widetilde{P}^*(v) = \sum_{j=1}^{n} \sum_{k=1}^{d(j)} \sigma(\lambda_k(j)) \mathbb{1}_{\{y_k(j)=v\}} / S_n,$$

$$\widetilde{P}^*(v,u) = \sum_{j=1}^{n} \sum_{k=1}^{d(j)-1} \sum_{l=k+1}^{d(j)} z(j,k,l) \mathbb{1}_{\{y_k(j)=v, y_l(j)=u\}} / Z_n,$$

$$\widetilde{P}^*(v|w) = \sum_{j=1}^{n} \sum_{k=1}^{d(j)} \sigma(\lambda_k(j)) \mathbb{1}_{\{y_k(j)=v, \eta(j)=w\}} / S_n^*,$$

$$\widetilde{P}^*(v,u|w) = \sum_{j=1}^{n} \sum_{k=1}^{d(j)-1} \sum_{l=k+1}^{d(j)} z(j,k,l) \mathbb{1}_{\{y_k(j)=v, y_l(j)=u, \eta(j)=w\}} / Z_n^*.$$

Akivaizdu, kad jei $0 \leq \sigma(\cdot) \leq 1$ ir $0 \leq \delta(\cdot,\cdot) \leq 1$, $0/0 = 0$, bent vienai porai (j,k) galioja $\sigma(\lambda_k(j)) > 0$ bei bent vienam trejetui (j,k,l) galioja $z(j,k,l) > 0$, tai pateiktieji skirstinių įvertiniai yra korektiški, t. y., $0 \leq \widetilde{P}^*(\cdot) \leq 1$, $0 \leq \widetilde{P}^*(\cdot,\cdot) \leq 1$, $0 \leq \widetilde{P}^*(\cdot|\cdot) \leq 1$, $0 \leq \widetilde{P}^*(\cdot,\cdot|\cdot) \leq 1$, $\sum_v \widetilde{P}^*(v) = 1$, $\sum_{v,u} \widetilde{P}^*(v,u) = 1$, $\sum_v \widetilde{P}^*(v|w) = 1$, $\sum_{v,u} \widetilde{P}^*(v,u|w) = 1$. Paskutinės dvi lygybės galioja, jei yra mokymo duomenų, kurie priskiriami klasei w. Kitu atveju natūraliai $\sum_v \widetilde{P}^*(v|w) = 0$, $\sum_{v,u} \widetilde{P}^*(v,u|w) = 0$.

Pateiktieji įvertiniai apibendrina (2.19), (2.20), (2.21) ir (2.22) bei sutampa su jais trivialių funkcionalų $\sigma \equiv 1$ ir $\delta(\lambda_k(\cdot),\lambda_l(\cdot)) = \mathbb{1}_{\{l=\lambda_l^{(t)}(\cdot)=\lambda_k^{(t)}(\cdot)+1=k+1\}}$ atvejais. Formulėse (2.23) ir (2.24) įstatę naujuosius įvertinius, gausime identifikacinius debesėlius, kuriuose atsižvelgiama į kontekstinę informaciją.

Analogiškai modifikuojame ir klasių nustatymo formules, kad šiose būtų atsižvelgiama į kontekstinę informaciją. Todėl (2.11) ir (2.14) virsta atitinkamai

$$\widehat{\eta} = \arg\max_{w\in K} \left[Q(w) \prod_{\tau\in I} \psi_w^{\sigma(\lambda_\tau(a))}(a_\tau) \right], \tag{2.44}$$

$$\widehat{\eta} = \arg\max_{w\in K} \left[Q(w) \psi_w^{\sigma(\lambda_r(a))}(a_r) \prod_{i=r}^{m-1} \Psi_w(i) \right], \tag{2.45}$$

čia

$$\Psi_w(i) = (\psi_w(a_i,a_{i+1})/\psi_w(a_i))^{\sigma(\lambda_i(a))\delta(\lambda_i(a),\lambda_{i+1}(a))}.$$

Taigi, kuo $\sigma(\cdot)$ arčiau nulio, tuo atitinkamoje pozicijoje klasifikuojamame straipsnyje esantis terminas neša mažiau diskriminantinės informacijos. Lygiai taip pat $\delta(\cdot,\cdot)$ reikšmė, artima nuliui, sumažina atitinkamose pozicijose esančios poros informatyvumą.

2.4.4. Svorių funkcionalai

Norint naudoti aukščiau pateiktas modifikuotas vertinimo ir klasifikavimo formules, reikia apsibrėžti konkrečius funkcionalus $\sigma(\cdot)$ bei $\delta(\cdot,\cdot)$. Trivialūs apibrėžimai $\sigma \equiv 1$ ir $\delta(\lambda_k(\cdot),\lambda_l(\cdot)) = \mathbb{1}_{\{l=\lambda_l^{(t)}(\cdot)=\lambda_k^{(t)}(\cdot)+1=k+1\}}$ veda prie lygybių ir metodų, nenaudojančių papildomos informacijos.

Funkcionalas $\sigma(\lambda) = \sigma_\theta(\lambda)$, kuriame $\lambda = (\lambda^{(t)},\lambda^{(w)},\lambda^{(s)},\lambda^{(p)})$, nurodo mokslo termino ar jų poros *svorį*, priklausomai nuo atstumo nuo teksto pradžios. Natūralu manyti, kad šis svoris turėtų mažėti didėjant atstumui. Be to, ties tam tikra riba svoris turėtų virsti nuliu, taip sumažinant analizuojamo teksto

apimtį bei išvengiant problemų su tokiose teksto dalyse kaip įrodymų dėstymas pasipilančiais tiesiogiai su straipsnio tematika nesusijusiais terminais.

Paprastas variantas apibrėžiamas rekursyvia formule

$$\sigma^*_\theta(i) = \begin{cases} 1, & i = 0, \\ 0, & i > p_k, \\ \max\{0, \sigma^*_\theta(i-1) + \alpha_j\}, & i \in (p_{j-1}, p_j] \end{cases} \tag{2.46}$$

su daugiamačiu parametru

$$\theta = (\alpha_1, p_1, \alpha_2, p_2, \ldots, \alpha_k, p_k).$$

Čia $0 = p_0 \leq p_1 < p_2 < \ldots < p_k$, $\alpha_{(\cdot)} \leq 0$ bei $\alpha_i \neq \alpha_{i+1}$.

Tada pasirinktam parametrų vektoriui $\theta^{(\sigma)}$ apibrėžiame

$$\sigma(\lambda) = \sigma_{\theta^{(\sigma)}}(\lambda) = \sigma^*_{\theta^{(\sigma)}}(\lambda^{(t)}). \tag{2.47}$$

Galima apibrėžti alternatyvius svorio funkcionalus, kurie atsižvelgtų į žodžio, sakinio ar pastraipos numerius, (2.47) $\lambda^{(t)}$ keičiant atitinkamai į $\lambda^{(w)}$, $\lambda^{(s)}$ ar $\lambda^{(p)}$.

Funkcionalą $\delta(\lambda_k, \lambda_l) = \delta_\theta(\lambda_k, \lambda_l)$, nusakantį priklausomybės tarp dviejų tekste esančių terminų svorį, apibrėšime naudodamiesi parametrine funkcija

$$\delta^*_\theta(i,j) = \begin{cases} 0, \text{ jei } i \geq j, \\ 0, \text{ jei } j - i > \theta_0, \\ \min\{1, \max\{0, 1 + \theta_1 + \theta_2 \cdot (j-i)\}\}, \text{ kitu atveju,} \end{cases} \tag{2.48}$$

čia $\theta = (\theta_0, \theta_1, \theta_2)$ yra trimatis parametrų vektorius, $\theta_0 \geq 0$. Komponentė θ_0 apibrėžia atstumą tarp teksto elementų, kurį viršijus laikoma, kad šie elementai nepriklausomi. Komponentės θ_1 ir θ_2 apibrėžia priklausomybės stiprumą aprašančią tiesinę funkciją nuo atstumo tarp elementų.

Pasirinkę parametrą $\theta^{(\delta)} = (\theta_0, \theta_1, \ldots, \theta_{11})$, apibrėžiame

$$\begin{aligned} &\delta(\lambda_k, \lambda_l) = \delta_{\theta^{(\delta)}}(\lambda_k, \lambda_l) \\ &= \delta^*_{(\theta_0,\theta_1,\theta_2)}(\lambda_k^{(t)}, \lambda_l^{(t)}) \cdot \delta^*_{(\theta_3,\theta_4,\theta_5)}(\lambda_k^{(w)}, \lambda_l^{(w)}) \\ &\cdot \delta^*_{(\theta_6,\theta_7,\theta_8)}(\lambda_k^{(s)}, \lambda_l^{(s)}) \cdot \delta^*_{(\theta_9,\theta_{10},\theta_{11})}(\lambda_k^{(p)}, \lambda_l^{(p)}). \end{aligned} \tag{2.49}$$

(2.49) pateiktoje $\delta(\cdot,\cdot)$ išraiškoje atsižvelgiama į visus kontekstinės informacijos elementus (termino, žodžio, sakinio ir pastraipos numerius).

Pavyzdžiai.

Parametrų vektorius $\theta^{(\delta)} = (1, 0, 0, \infty, 0, 0, \infty, 0, 0, \infty, 0, 0)$ atitinka paprastą Markovo savybę, kai nagrinėjamos gretimų terminų poros, neatsižvelgiant į kontekstą tarp jų.

Vektorius $\theta^{(\delta)} = (10, 0, 0, 51, 0{,}01, -0{,}01, 2, 0, -0{,}3, 0, 0, 0)$ atitinka kitą hipotetinę situaciją. Nagrinėjami terminai, kurie nutolę ne daugiau kaip per 10 pozicijų straipsnio projekcijoje. Už kiekvieną tarp dviejų terminų esantį žodį *svoris* sumažinamas 1 %. Jei tarpe daugiau nei 50 žodžių – pora laikoma nesusijusi, t. y., jos *svoris* lygus nuliui. Jei terminai yra skirtinguose sakiniuose, tai už kiekvieną tašką (sakinių skirtuką) papildomai nubraukiama dar 30 % *svorio*. Jei terminus skiria daugiau, nei 1 sakinys (du taškai) – pora nesusijusi. Atitinkamai jei terminai yra skirtingose pastraipose, pora taip pat laikoma nesusijusi.

Akivaizdu, kad ieškant optimalių parametrų $\theta^{(\sigma)}$ ir $\theta^{(\delta)}$, reikia remtis euristiniais samprotavimais, nes perrinkti net ir nedidelę dalį galimų parametrų reikšmių būtų labai sudėtinga ir neabejotinai neprasminga.

2.5. Siūlomi klasifikavimo algoritmai

Šiame skyrelyje suformuluotų modelių bei jų identifikavimo procedūrų pagrindu siūlomi konstruktyvūs mokslo publikacijų klasifikavimo algoritmai.

Kiekvieną algoritmą sudaro dvi dalys: apmokymas ir naujo straipsnio klasifikavimo procedūra. Visų algoritmų klasifikavimo procedūros paremtos formule (2.7) ar jos atmainomis (2.11), (2.14), (2.44), (2.45). Kadangi dokumentas gali būti priskirtas kelioms klasėms vienu metu, taikomas apibendrinimas (1.21): klasifikavimo procedūra sudaro ranginį klasių sąrašą pagal reikšmę maksimizuojamo dydžio ir po to pagal pasirinktą taisyklę iš šio sąrašo išrenka tam tikrą skaičių aukščiausio rango klasių. Išrinkimo strategija šiame darbe nenagrinėjama, o tyrimuose naudojamos paprastos fiksuoto klasių skaičiaus parinkimo taisyklės.

Algoritmas, apibrėžiamas formulėmis (2.7), (2.16) ir (2.17), yra įprastas k kaimynų algoritmas, bet su mūsų pasiūlytu atstumu (2.18).

Toliau laikykime, kad turime papildoma informacija praturtintą mokymo imtį, apibrėžtą (2.43).

IDCm algoritmas. Šis algoritmas apibrėžtas galiojant prielaidai apie mokslo terminų skirstinių sąlyginį stacionarumą ir Markovo savybę (2.12).

Parametrai. $\mu_1 \in N$ – minimalus reikalaujamas termino stebėjimų skaičius, $\mu_2 \in N$ – minimalus reikalaujamas terminų poros stebėjimų skaičius, α_1 – reikšmingumo lygis informatyvių terminų atrinkimui, α_2 – reikšmingumo lygis informatyvių terminų porų atrinkimui, $\theta^{(\sigma)}$ – daugiamatis parametras funkcionalui $\sigma(\cdot)$, $\theta^{(\delta)}$ – dvylikamatis parametras funkcionalui $\delta(\cdot,\cdot)$.

Mokymasis.

- **Vertinimas.** Visų pirma iš imties įvertinamos apriorinės klasių tikimybės $Q(\cdot)$ pagal (2.16). Pagal (2.4.3), (2.4.3), (2.4.3) ir (2.4.3) skaičiuojami skirstinių $P(\cdot)$, $P(\cdot,\cdot)$, $P(\cdot|\cdot)$ ir $P(\cdot,\cdot|\cdot)$ įverčiai. $\sigma(\cdot)$ skaičiuojama pagal (2.47), naudojantis rekursyvia formule (2.46) su parametru $\theta^{(\sigma)}$. $\delta(\cdot,\cdot)$ skaičiuojama pagal (2.49) formulę su parametru $\theta^{(\delta)}$.
- **Modifikavimas.** Gauti skirstinių $P(\cdot)$, $P(\cdot|\cdot)$, $P(\cdot,\cdot)$ ir $P(\cdot,\cdot|\cdot)$ įverčiai statomi į (2.23) ir (2.24) ir taip gaunami įverčiai $\widetilde{\psi}_{(\cdot)}(\cdot)$ ir $\widetilde{\psi}_{(\cdot)}(\cdot,\cdot)$. Naudojant parametrus μ_1 ir μ_2 šie funkcionalų įverčiai glodinami pagal (2.3.3) formulę (poroms analogiškai) – gaunami įverčiai $\widetilde{\psi}^*_{(\cdot)}(\cdot)$ ir $\widetilde{\psi}^*_{(\cdot)}(\cdot,\cdot)$. Pastarieji darsyk modifikuojami pagal (2.27) taisyklę (ir analogišką poroms). Informatyvių terminų aibė L nustatoma iš (2.28) ir (2.31), įstatant užduotą parametrą α_1. Informatyvių terminų porų aibė L_2 nustatoma analogiškai su užduotu parametru α_2. Tokiu būdu gaunami įverčiai $\widehat{\psi}_{(\cdot)}(\cdot)$ ir $\widehat{\psi}_{(\cdot)}(\cdot,\cdot)$.
- **Parametrizavimas.** Taikome parametrinį modelį (2.38), kuriame $\gamma(\cdot,\cdot)$ apibrėžtas (2.37), o parametrai θ ir θ^* parenkami pagal (2.39) ir (2.40). Funkcionalui $\psi_{(\cdot)}(\cdot,\cdot)$ parametrizacija netaikoma.

Klasifikavimas. Stebėtam straipsnio fragmentui a_I kiekvienam $w \in K$ skaičiuojamas lygybės (2.45) dešinėje pusėje esantis dydis, tačiau jis nemaksimizuojamas. Vietoje nežinomų $Q(\cdot)$, $\psi_w(\cdot)$ ir $\psi_w(\cdot,\cdot)$ statomi jų įverčiai, apskaičiuoti algoritmo mokymosi žingsniuose. Visos klasės išrikiuojamos pagal suskaičiuotą dydį mažėjančia tvarka ir pagal tam tikrą taisyklę atrenkama dalis sąrašo pradžioje esančių, kurios ir yra klasifikavimo algoritmo rezultatas.

IDC algoritmas. Algoritmas apibrėžtas galiojant prielaidai apie mokslo terminų skirstinių sąlyginį stacionarumą ir nepriklausomumą (2.8) ir yra *IDCm* modifikacija, kai neatsižvelgiama į terminų poras.

Parametrai. $\mu \in N$ – minimalus reikalaujamas termino stebėjimų skaičius, α – reikšmingumo lygis informatyvių terminų atrinkimui, $\theta^{(\sigma)}$ – daugiamatis parametras funkcionalui $\sigma(\cdot)$.

Mokymasis. Kaip *IDCm*, neatliekami su terminų poromis susiję žingsniai.

Klasifikavimas. Stebėtam straipsnio fragmentui a_I kiekvienam $w \in K$ skaičiuojamas lygybės (2.44) dešinėje pusėje esantis dydis, tačiau jis nemaksimizuojamas. Vietoje nežinomų $Q(\cdot)$ ir $\psi_w(\cdot)$ statomi jų įverčiai, apskaičiuoti algoritmo mokymosi žingsniuose. Toliau analogiškai kaip *IDCm*.

Algoritmų variantai. Apibrėžtieji *IDC* ir *IDCm* algoritmai pateikti bendriausioje savo formoje, tačiau juose galimos įvairios variacijos, pvz.: informatyviausių terminų atrinkimą atlikti ne *hyp*, o kitu metodu, netaikyti parametrizacijos, nenaudoti neigiamą diskriminantinę informaciją nešančių identifikacinių debesėlių dalių ($\psi_w(\cdot) < 1$) ir pan. Šie variantai čia neišskirti kaip atskiri algoritmai – jie aptariami ir lyginami eksperimentinį tyrimą aprašančioje darbo dalyje.

2.6. Antrojo skyriaus išvados

- Šiame skyriuje suformuluotas stochastinis mokslo terminijos pasiskirstymo straipsnių tekstuose modelis, įvestos prielaidos apie terminijos sąlyginių skirstinių tarpusavio priklausomybes, supaprastinančios modelių identifikavimo procedūras.
- Pateikta intuityvios mokslo termino aplinkos, dar vadinamos identifikaciniu debesėliu, sąvokos formalizacija, apibrėžta pasiūlytuose modeliuose įvestų terminijos skirstinių funkcionalais.
- Suformuluotos modelių identifikavimo procedūros, derinančios neparametrinius ir parametrinius statistinius metodus.
- Skyriuje pasiūlyta papildomos informacijos, susijusios su terminijos pozicijomis tekste bei kontekstu tarp jų, struktūra bei pateikti šios informacijos panaudojimo klasifikavimui sprendimai.
- Pasiūlytųjų modelių ir jų identifikavimo procedūrų pagrindu suformuluoti konstruktyvūs klasifikavimo algoritmai.

3

Klasifikavimo algoritmų eksperimentinis tyrimas

Skyriaus tematika paskelbti šie autoriaus straipsniai: (Balys et al., 2004a, 2005, 2008).

3.1. Eksperimentinė sistema ir tyrimo metodika

3.1.1. Naudota publikacijų duomenų bazė

Eksperimentai atlikti su duomenų baze, sudaryta iš 14497 straipsnių iš tikimybių teorijos ir matematinės statistikos sričių, gauta iš Matematinės statistikos instituto (*Institute of Mathematical Statistics, IMS, USA*). Bazę sudarantys straipsniai turi du svarbius metaduomenų elementus: raktinių žodžių (frazių) sąrašą ir MSC klasifikatorių sąrašą, greičiausiai priskirtus autoriaus. Algoritmų analizės tikslais šiuos raktinius žodžius ir MSC klasifikatorius laikome esant teisingais. Tyrime naudotos dvi dokumentų aibės – atvejui, kai klasės yra MSC klasifikatoriai bei atvejui, kai klasės yra raktiniai žodžiai. Šios straipsnių aibės toliau atitinkamai žymimos A^{MSC} bei A^{KWD}.

44 MSC klasifikatoriai (24 iš 60XXX pomedžio ir 20 iš 62XXX pomedžio) buvo atrinkti eksperimentams, o kiekvienas klasifikatorius priskirtas bent 100-ui skirtingų straipsnių (klasifikatorių sąrašas pateiktas darbo priede). Surinkus šiuos

straipsnius gauta 5338 straipsnių pirmoji aibė A^{MSC}.

Atitinkamai atrinkti 44 raktiniai žodžiai, kurie yra priskirti bent 50-iai skirtingų straipsnių (raktinių žodžių sąrašas pateiktas darbo priede). Surinkus šiuos straipsnius gauta 2977 straipsnių antroji aibė A^{KWD}.

3.1 ir 3.2 lentelėse pateikta klasių (klasifikatorių bei raktinių žodžių), priskirtų straipsniams, skaičiaus statistika. *Pastaba:* laikoma, kad tie raktiniai žodžiai ir klasifikatoriai, kurie nepakliuvo į atrinktųjų sąrašus, straipsniams nėra priskirti, todėl jie statistikoje neatsispindi.

3.1 lentelė. MSC klasifikatorių skaičiaus straipsniuose statistika

	Straipsniui priskirtų klasifikatorių skaičius						
	1	2	3	4	5	6	7
Straipsnių skaičius	2617	1905	639	161	15	0	1
Straipsnių dalis	0,49	0,36	0,12	0,03	~ 0	0	~ 0
						Vidurkis:	1,46
						Mediana:	2

3.2 lentelė. Raktinių žodžių skaičiaus straipsniuose statistika

	Straipsniui priskirtų raktinių žodžių skaičius					
	1	2	3	4	5	6
Straipsnių skaičius	2094	697	146	35	4	1
Straipsnių dalis	0,70	0,23	0,05	0,01	~ 0	~ 0
					Vidurkis:	1,37
					Mediana:	1

3.1.2. Mokslo terminų žodynai

Straipsnių tekstai paversti į požymių vektorius, kuriuose vektoriaus komponentės atitinka terminus iš pasirinkto terminų žodyno. Tyrime naudoti trys skirtingi terminų žodynai:

- Žodynas V_1 sudarytas išrinkus visus raktinius žodžius, tiksliau – raktines frazes, iš duomenų bazę sudarančių straipsnių (iš straipsnio autoriaus sudaryto raktinių žodžių sąrašo). Kiekviena frazė suskaidyta į atskirus žodžius ir šie žodžiai (atmetus jungtukus ir kitus neprasminius žodžius)

taip pat sudėti į žodyną. Taip gautas 17632 unikalių terminų sąrašas. Dažnas terminas iš V_1 yra ne vienas žodis, o frazė – žr. 3.4 lentelę, kurioje pateikta terminų ilgio (matuojamo žodžių skaičiumi) statistika.

- Žodynas V_2 sudarytas iš atskirų žodžių iš V_1, be frazių. Taip gautas 5196 unikalių terminų sąrašas. Šio žodyno naudojimas leido ištirti frazių naudojimo įtaką algoritmų tikslumui.
- Žodynas V_3 sudarytas iš visų straipsnių tekstuose aptiktų žodžių (atmetus jungtukus ir kitus neprasminius žodžius). Taip gautas 16808 unikalių terminų sąrašas. Šio žodyno naudojimas leido palyginti algoritmų rezultatus naudojant visus kalbos žodžius bei naudojant tik mokslinių terminų sąrašą.

3.3 lentelėje pateikta terminų iš šių žodynų aptikimo skirtingose straipsnių dalyse statistika.

3.3 lentelė. Unikalių terminų aptikimo straipsnių dalyse statistika

Straipsnių aibė	Terminų žodynas	Straipsnio dalis		
		Pavadinimas	Santrauka	Tekstas
A^{MSC}	V_1	2419	3832	5072
	V_2	1169	1718	2065
	V_3	1302	3234	6123
A^{KWD}	V_1	1721	3001	4508
	V_2	883	1455	1937
	V_3	993	2755	6046

3.4 lentelė. Terminų iš žodyno V_1 ilgio statistika

	Termino ilgis (žodžių skaičius)					
	1	2	3	4	5	> 5
Terminų skaičius	5196	7271	3848	998	235	84
Dalis žodyne V_1	0,30	0,41	0,22	0,06	0,01	~ 0

3.1.3. Nagrinėtos tekstų dalys

Trys dokumentų dalys buvo prieinamos, t. y., jas buvo galima vienareikšmiškai išskirti: pavadinimai, santraukos (abstraktai) ir pilni tekstai (toliau vadinsime tiesiog tekstais). 3.5 lentelėje pateikta pagrindinė statistinė informacija apie šių dalių ilgius, matuojamus terminų iš V_1 skaičiumi.

3.5 lentelė. Teksto dalių ilgio statistika

Teksto dalis	Terminų žodynas	Ilgio statistika		
		Maksimumas	Vidurkis	Mediana
Pavadinimas	V_1	12	3,8	4
	V_2	10	3	3
	V_3	10	2,9	3
Santrauka	V_1	88	22	19
	V_2	74	18	16
	V_3	92	22	19
Tekstas	V_1	3809	602	482
	V_2	3495	547	457
	V_3	4505	775	620

Pastaba: naudojant terminų žodyną V_1 iškyla klausimas, ką daryti, kai tekste randama frazė, nes tą frazę sudarantys pavieniai žodžiai taip pat yra šiame žodyne. Šioje situacijoje galimi du sprendimai: laikyti, kad rasta tik frazė arba laikyti, kad rasta ir frazė, ir ją sudarantys žodžiai. 3.5 lentelėje tekstų ilgių statistika pateikta antruoju atveju, nes atliekant tyrimus nustatyta, kad jo naudojimas leidžia pasiekti aukštesnius tikslumo rezultatus.

Tyrimuose naudoti du pilnų tekstų variantai – tekstai, apimantys 110 pirmų terminų, ir visiškai pilni tekstai. Šis atskyrimas susijęs su labai didelėmis kai kurių algoritmų skaičiavimų apimtimis bei su labai ilgų tekstų specifika. Skaičius 110 parinktas atlikus empirinius bandymus ir nustačius, kad tiek terminų užtenka, kad visi algoritmai pasiektų 90–95 % savo maksimalaus tikslumo. Toliau „tekstai“ pagal nutylėjimą reikš pirmo tipo, t. y., sutrumpintus tekstus, jei nenurodyta kitaip.

Tyrime taikyti keli dokumentų teksto dalių panaudojimo mokymui ir testavimui scenarijai, siekiant įvertinti, kaip naudojamos dokumentų dalys ir jų ilgiai įtakoja klasifikavimo rezultatus. Nagrinėtos penkios kombinacijos: pavadinimas/pavadinimas, santrauka/santrauka, santrauka/tekstas, tekstas/santrauka ir tekstas/tekstas. Čia pirmas elementas nurodo, kuri dokumentų dalis naudota mokymo fazėje, o antrasis – kuri dokumentų dalis naudota testavimo fazėje.

3.1.4. Nagrinėti klasifikavimo algoritmai

Eksperimentiniame tyrime nagrinėti šie algoritmai:

- *IDC*, *IDCm* algoritmai ir jų variantai, apibrėžti 2.5 skyriuje. Toliau *IDC* ir *IDCm* žymėsime algoritmus, nenaudojančius papildomos informacijos, o $IDC^{(a)}$ ir $IDCm^{(a)}$ – naudojančius;
- *nB (polinominis naivaus Bajeso algoritmas su adityviu glodinimu)* – tikimybinis algoritmas, kuris remiasi teksto elementų nepriklausomo išsidėstymo tekste prielaida ir naudoja Bajeso taisyklę aposteriorinėms klasių tikimybėms skaičiuoti, žr. 1.4.1 skyrelį;
- *kNN* – įprastas pavyzdžiais grįstas k artimiausių kaimynų algoritmas, kuris praleidžia mokymosi fazę, o sprendimus priima analizuodamas teisingus ekspertų sprendimus dokumentuose iš mokymo imties, kurie yra artimiausi nagrinėjamam, žr. 1.4.4 skyrelį. Algoritmo parametrai: atstumo arba panašumo matas bei kaimynų skaičius k;
- *SVM* – atraminių vektorių algoritmas, kuris bando su didžiausiu galimu pločiu hiperplokštuma atskirti taškus, atitinkančius dokumentus, priklausančius klasei nuo nepriklausančių, žr. 1.4.2 skyrelį.
- *LSSF* – algoritmas, darantis prielaidą apie tiesinę priklausomybę tarp klasių svorių ir teksto elementų svorių ir naudojantis mažiausių kvadratų metodą sąryšio parametrams įvertinti, žr. 1.4.3 skyrelį. Algoritmas turi vieną parametrą k, nurodantį po SVD dekompozicijos (skaidymo singuliariomis reikšmėmis) paliekamų singuliarių reikšmių skaičių.

Visų nagrinėtų algoritmų, išskyrus *kNN*, veikimo principas paremtas teigiamų ir neigiamų klasių pavyzdžių analize, kurios pagrindu konstruojamos atskyrimo taisyklės, suklasifikuojančios mokymo duomenis kuo tiksliau.

Algoritmai realizuoja rangavimo procedūrą: kiekvienam dokumentui pateikiamas sąrašas klasių, kurioms būtų galima jį priskirti, kartu su jų atitinkamais svoriais. Tada, priklausomai nuo pasirinktos strategijos, išenkama dalis pirmųjų klasių, kurioms dokumentas ir priskiriamas.

3.1.5. Algoritmų tikslumo vertinimo metodika

Algoritmai lyginami pagal jų vidutinius klasifikavimo nuostolius (1.17), tiksliau, kaip minėta, pagal vidutinį klasifikavimo tikslumą, kuris apibrėžiamas lygiai taip pat, tik skiriasi funkcionalas l, o minimizavimas keičiamas maksimizavimu. Kadangi tiksliai suskaičiuoti šio dydžio neįmanoma, skaičiuojamas

empirinis atitikmuo

$$\widehat{L}(A, X_n, Y_m) = m^{-1} \sum_{j=1}^{m} l(\eta(Y_j), \widehat{\eta}^{(A,X_n)}(Y_j)). \tag{3.1}$$

Čia $Y_m = \{(Y(1), \eta(1)), \ldots, (Y(m), \eta(m))\}$ yra dar viena, vadinamoji testavimo, imtis.

Apmokymui ir testavimui naudoti tų pačių duomenų nerekomenduojama, nes taip būtų gaunami iškreiptai optimistiniai įverčiai. Kadangi kaip paprastai duomenų kiekis ribotas ir juos dalinti į atskiras mokymo ir testavimo imtis būtų per didelė prabanga, tikslumo vertinimui naudojamas vadinamas k kryžminio patikrinimo metodas (*k-fold cross-validation*) (Burman, 1989; Zhang, 1993). Turima imtis X_n skaidoma į k apylygių dalių: $X_n = X_n^1 \cup \ldots \cup X_n^k$. Pažymėkime $X_n^{(-i)} = X_n \setminus X_n^i$. Tada (1.17) keičiama į

$$\widehat{L}_{CV}(A, X_n) = k^{-1} \sum_{i=1}^{k} \widehat{L}(A, X_n^{(-i)}, X_n^i), \tag{3.2}$$

čia sumuojamasis narys apibrėžtas (3.1).

Darbe naudotas k kryžminio patikrinimo metodas su $k = 5$.

3.1.6. Naudoti klasifikavimo tikslumo matai

Tarkime, straipsnis iš testavimo imties (autoriaus) priskirtas klasėms iš aibės K_T, o algoritmas straipsnį priskyrė klasėms iš aibės K_A. Norėdami taikyti formulę (3.2), turime apibrėžti funkcionalą $l(K_T, K_A)$. Populiariausi tyrimuose naudojami tikslumo matai yra vadinamieji *precision* (Pr) ir *recall* (Re) (Salton and McGill, 1986):

$$Pr(K_T, K_A) = \frac{|K_T \cap K_A|}{|K_A|}, \quad Re(K_T, K_A) = \frac{|K_T \cap K_A|}{|K_T|}. \tag{3.3}$$

Pr yra dalis algoritmo spėtų klasių, kurios išties yra tarp žinomai teisingai priskirtų, o Re yra dalis teisingai priskirtų klasių, kurios yra ir tarp spėtųjų. Šie dydžiai tam tikra prasme vienas kitam priešingi, todėl naudojami atskirai vienas nuo kito suteikia nepilną informaciją. Dažnai naudojamas vadinamasis

F-matas (van Rijsbergen, 1979), kuris apjungia šiuos matus su pasirinktu svoriu:

$$F_\beta = \frac{(1+\beta^2)\cdot(Re\cdot Pr)}{\beta^2\cdot Pr + Re}. \tag{3.4}$$

Populiariausias yra harmoninis Pr ir Re vidurkis $F_1 = 2\cdot Pr\cdot Re/(Pr+Re)$.

Akivaizdu, kad aukščiau aprašyti matai priklauso nuo algoritmo parenkamų klasių skaičiaus. Šiame darbe klasių skaičiaus nustatymo uždavinys nenagrinėtas, todėl vertinant ir lyginant algoritmus naudota paprasta fiksuoto tikėtiniausių klasių skaičiaus (1–5) išrinkimo strategija. Žemiau pateiktas matas, panašus į *11-point average precision* (Salton and McGill, 1986), kuris leidžia įvertinti algoritmo tikslumą, nepriklausomai nuo klasių parinkimo strategijos, atsižvelgiant vien į jų ranginį sąrašą. Tegul algoritmas paeiliui straipsniui priskyrinėja po vieną klasę iš savo susidaryto ranginio klasių sąrašo. Fiksuojami žingsniai k_i, kai eilinė priskiriama klasė pasirodo besanti tarp žinomai teisingai priskirtų (K_T), taip padidinant Re reikšmę dydžiu $1/|K_T|$. Kai visos klasės atspėtos (Re = 1), procedūra stabdoma ir suskaičiuojamas Pr reikšmių vidurkis užfiksuotais Re padidėjimo momentais k_i:

$$Pr_{avg} = |K_T|^{-1}(1/k_1 + 2/k_2 + \ldots + |K_T|/k_{|K_T|}). \tag{3.5}$$

Įsivestuosius matus įstatę vietoje l į (3.1) išraišką, o šią į (3.2) gauname dydžius, kuriais remdamiesi lyginsime algoritmus. Toliau juos atitinkamai žymėsime Pr^A, Re^A, F_1^A ir Pr_{avg}^A, čia A – nagrinėjamas algoritmas.

3.1.7. Naudota programinė įranga

Tyrimams naudota autoriaus sukurta programinė įranga, parašyta Python programavimo kalba (<http://www.python.org>). Skaičiavimams naudotos Python kalbos SciPy bibliotekos (<http://www.scipy.org>) matematinės funkcijos. Paveikslų braižymui naudota Python kalbos matplotlib biblioteka (<http://matplotlib.sourceforge.net>). LLSF metode naudojama SVD dekompozicija atlikta su SciPy *linalg.svd* funkcija, mažiausių kvadratų metodui naudota SciPy *optimize.minpack.leastsq* funkcija. SVM algoritmui naudota SVM^{light} programa (<http://svmlight.joachims.org>).

Tyrimų metu sukurta programinė įranga panaudota demonstracinėje internetinėje automatinio klasifikavimo sistemoje.

3.2. Eksperimento rezultatai

3.2.1. Terminų žodyno įtaka klasifikavimo tikslumui

3.7 ir 3.8 lentelėse pateikti algoritmų tikslumo vertinimo rezultatai (Pr_{avg} reikšmės) prie skirtingų žodynų bei skirtingų mokymo ir testavimo imčių porų.

Akivaizdu, kad aibės A^{KWD} atveju, kai klasės yra raktiniai žodžiai, rezultatai yra žymiai prastesni (10–20 %), nei aibės A^{MSC} atveju, kuriame klasės yra MSC klasifikatoriai. Skirtumas greičiausiai gali būti paaiškintas tuo, kad pirmojoje aibėje mokymui ir testavimui turėta vidutiniškai perpus mažiau duomenų, be to, raktiniai žodžiai dėl savo prigimties nėra tiksliai ir vienareikšmiškai apibrėžti klasifikatoriai. Toliau aibės A^{KWD} atvejo nebenagrinėsime, apsistodami ties turtingesne A^{MSC}.

Įvairioms mokymo ir testavimo imčių poroms bei kiekvienam algoritmui atlikti bandymai su visais trimis žodynais V_1, V_2 ir V_3. Visais atvejais taikyta DF žodyno sumažinimo strategija (žr. 1.3 skyrelį) – tie terminai, kurie mokymo imtyje sutinkami mažiau nei tam tikrame fiksuotame skaičiuje skirtingų dokumentų, nenagrinėti. Nustatyta, kad optimali dažnumo slenksčio DF reikšmė yra tarp 2 ir 5, priklausomai nuo tekstų ilgio (ilgesniems tekstams – didesnė). Iš pateiktų rezultatų matome, kad naudojant žodyną V_1 visada gauname tikslesnius rezultatus, nei naudojant V_3. Priklausomai nuo algoritmo ir nuo duomenų, šis skirtumas svyruoja maždaug ties 10 %. Žodynas V_2 beveik visada duoda rezultatus per vidurį tarp V_3 ir V_1, išskyrus keletą išskirtinių atvejų. Iš 3.6 lentelės matome, kad žodynas V_2 ženkliai mažesnis (50–70 %) už žodyną V_3, o žodynas V_1 panašaus dydžio ar mažesnis už V_3 prie ilgesnių mokymo tekstų. Todėl darome išvadą, kad naudoti vienažodžių mokslo terminų žodyną V_2 vietoje įprasto visų kalbos žodžių žodyno V_3 yra verta. Naudoti didesnį mokslo terminų (tiek žodžių, tiek ir frazių) žodyną V_1 yra taip pat prasminga, nes gaunamas tikslumo padidėjimas kompensuoja pakankamai nedidelį apimties prieaugį.

3.2.2. Algoritmų tikslumo palyginimas

3.7 lentelėje ir 3.1–3.5 paveiksluose pateikti algoritmų tikslumo rezultatai, kuriuos dabar ir aptarsime.

- kNN algoritmas tiksliausias binarinio teksto reprezentavimo atveju, naudojant kampo tarp vektorių kosinuso panašumo matą (1.45). Optimalus kaimynų skaičius k priklausomai nuo mokymo ir testavimo duomenų imčių svyruoja apie 25–50 (žr. 3.11 paveikslą).

3.6 lentelė. Žodynų dydžio statistika. Skliaustuose nurodytas santykis su atitinkama žodyno V_3 reikšme

Mokymo duomenys	Žodynas	Žodyno dydis		
		Pradinis	Po DF sumažinimo	Sumažinimo koef.
Pavadinimas	V_1	2419 (1,85)	929 (1,51)	0,38
	V_2	1169 (0,89)	572 (0,93)	0,49
	V_3	1302	615	0,47
Santrauka	V_1	3832 (1,18)	2012 (1,03)	0,52
	V_2	1718 (0,53)	1127 (0,57)	0,66
	V_3	3234	1962	0,61
Tekstas	V_1	5072 (0,83)	3458 (0,72)	0,68
	V_2	2065 (0,34)	1650 (0,34)	0,8
	V_3	6123	4833	0,79

3.7 lentelė. $Pr_{avg}^{(\cdot)}$ priklausomybė nuo terminų žodyno bei mokymo ir testavimo imčių poros (A^{MSC} atveju)

Mokymo ir testavimo imtys	Žodynas	Algoritmas				
		nB	kNN	SVM	LLSF	IDC
Pavadinimas/Pavadinimas	V_1	0,512	0,485	0,503	0,500	0,502
	V_2	0,481	0,460	0,489	0,493	0,477
	V_3	0,466	0,443	0,443	0,459	0,461
Santrauka/Santrauka	V_1	0,584	0,540	0,590	0,580	0,592
	V_2	0,561	0,520	0,557	0,540	0,567
	V_3	0,539	0,491	0,532	0,522	0,545
Santrauka/Tekstas	V_1	0,598	0,562	0,602	0,594	0,604
	V_2	0,591	0,532	0,541	0,548	0,592
	V_3	0,573	0,504	0,537	0,553	0,570
Tekstas/Santrauka	V_1	0,618	0,568	0,629	0,555	0,617
	V_2	0,580	0,536	0,595	0,529	0,577
	V_3	0,572	0,509	0,589	0,535	0,568
Tekstas/Tekstas	V_1	0,653	0,586	0,661	0,628	0,658
	V_2	0,618	0,557	0,621	0,601	0,632
	V_3	0,601	0,534	0,607	0,592	0,616

- $LLSF$ algoritmas optimalus su parametro k reikšme tarp 500 ir 600 (žr. 3.12 paveikslą). Dėl didelių skaičiavimų apimčių nagrinėtos tik kelios fiksuotos parametro reikšmės $k = 200, 400, 500, 600, 800$.

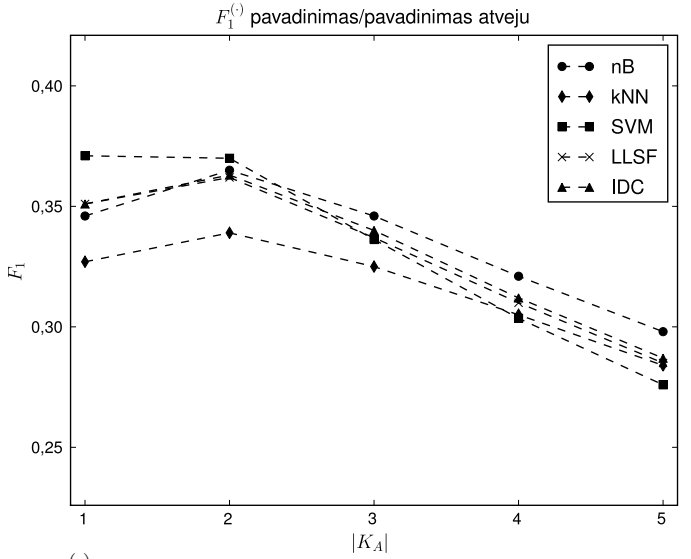
3.8 lentelė. $Pr_{avg}^{(\cdot)}$ priklausomybė nuo terminų žodyno bei mokymo ir testavimo imčių poros (A^{KWD} atveju)

Mokymo ir testavimo imtys	Žodynas	Algoritmas				
		nB	kNN	SVM	LLSF	IDC
Pavadinimas/Pavadinimas	V_1	0,438	0,450	0,428	0,417	0,428
	V_2	0,402	0,406	0,379	0,399	0,395
	V_3	0,389	0,397	0,364	0,395	0,376
Santrauka / Santrauka	V_1	0,510	0,482	0,534	0,553	0,515
	V_2	0,470	0,442	0,461	0,482	0,475
	V_3	0,455	0,427	0,456	0,473	0,464
Santrauka/Tekstas	V_1	0,510	0,461	0,523	0,5	0,516
	V_2	0,493	0,421	0,448	0,483	0,486
	V_3	0,475	0,400	0,435	0,491	0,475
Tekstas/Santrauka	V_1	0,542	0,501	0,492	0,529	0,542
	V_2	0,491	0,454	0,429	0,474	0,491
	V_3	0,487	0,440	0,441	0,504	0,491
Tekstas/Tekstas	V_1	0,572	0,514	0,572	0,541	0,579
	V_2	0,538	0,478	0,506	0,531	0,552
	V_3	0,521	0,455	0,507	0,528	0,537

- SVM algoritmas tiksliausias nenaudojant branduolio funkcijos Visgi, kadangi žinoma, jog šio algoritmo optimalių parametrų paieška yra itin kebli, galutinės griežtos išvados apie optimalumą daryti negalime.
- IDC algoritmo parametrai ir rezultatų priklausomybė nuo jų išsamiau aptariami 3.2.4 skyrelyje, 3.7 lentelėje pateikti rezultatai paprasto algoritmo varianto, nenaudojančio papildomos informacijos.

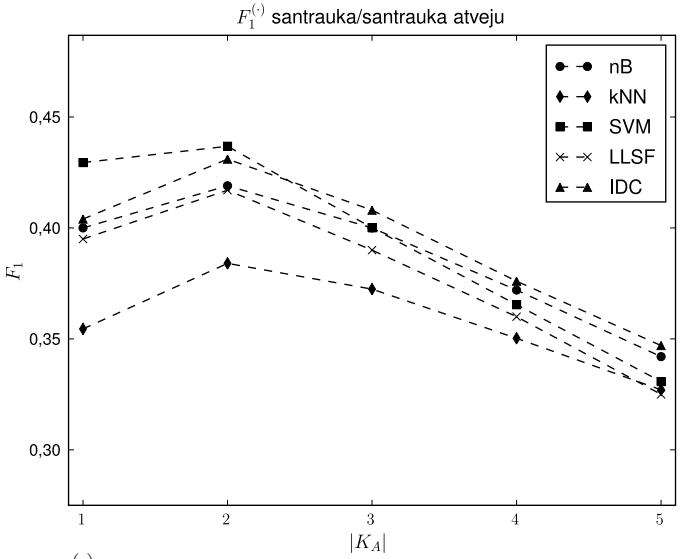
Iš gautų rezultatų darome šias išvadas:

- kNN algoritmo tikslumas mažiausias. Nors pavadinimas/pavadinimas situacijoje jis atsilieka nuo konkurentų ganėtinai nedaug, tačiau pereinant prie realistiškesnių situacijų, kai naudojami ilgesni tekstai, skirtumas išauga iki maždaug dešimties procentų, lyginant su tiksliausiais algoritmais. Tekstas/santrauka atveju kNN yra tikslesnis už $LLSF$, tačiau šioje vietoje išties stebime pastarojo algoritmo „duobę“, kurią įtakoja disproporcija tarp mokymo ir testavimo dokumentų ilgio, o ne kNN realų pranašumą, kas akivaizdu lyginant su kitų algoritmų rezultatais.



3.1 pav. $F_1^{(\cdot)}$ priklausomybė nuo parenkamų klasių skaičiaus pavadinimas / pavadinimas atveju)

- $LLSF$ algoritmo tikslumas mažesnis, nei nB, IDC ir SVM, be to, jis kaip ir SVM pasižymi didelėmis skaičiavimų apimtimis. Tiesa, sudėtingas skaidymas singuliariomis reikšmėmis atliekamas tik mokymo fazėje, o pats klasifikavimas yra labai greitas, nes jis remiasi paprasta algebrine matricų sandauga.
- nB algoritmas patvirtino savo kaip vieno iš patikimiausių algoritmų reputaciją – jo tikslumas yra itin aukštas ir ganėtinai nedaug atsilieka nuo žymiai sudėtingesnio SVM algoritmo rezultatų. Be to, šis metodas pats paprasčiausias skaičiavimų prasme.
- SVM algoritmo tikslumas didžiausias iš visų bandytųjų, nors paprastesnius nB ir IDC lenkia nedaug. 3.1–3.5 paveiksluose stebimas SVM tikslumo kritimas parenkant daugiau klasių, ypač trumpesnių tekstų atveju. Parinkdamas vieną klasę algoritmas yra žymiai tikslesnis, nei konkurentai, tačiau parenkant vis daugiau klasių, tikslumas ($F_1^{(\cdot)}$ reikšmė) krenta, kai kuriais atvejais nusileisdamas net kNN algoritmui.
- IDC algoritmo (paprastas variantas, be papildomos informacijos) rezultatai panašūs į nB. Šiuos rezultatus galima pagerinti, žr. 3.2.4 skyrelį.



3.2 pav. $F_1^{(\cdot)}$ priklausomybė nuo parenkamų klasių skaičiaus (santrauka / santrauka atveju)

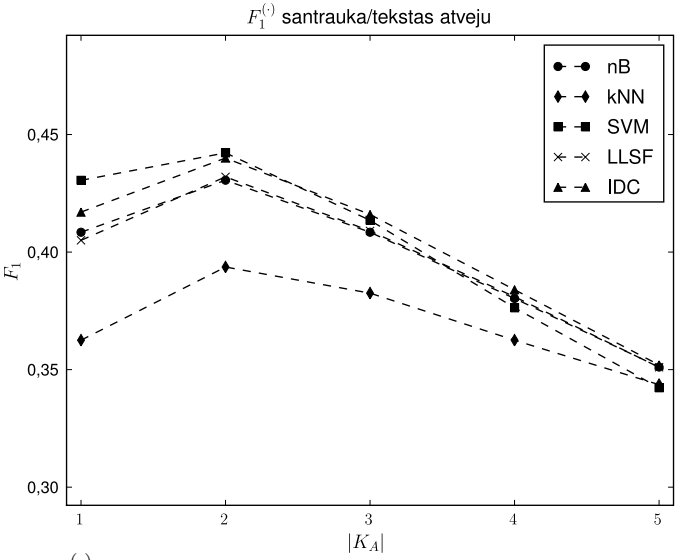
Apibendrinant rezultatus galima teigti, kad tikimybiniai algoritmai (nB ir IDC) nežymiai atsilieka nuo sudėtingesnio ir daugiau skaičiavimų reikalaujančio SVM algoritmo ir lenkia kNN bei tiesinį $LLSF$. Taip pat jie yra stabilesni parenkant daugiau klasių ir juose figūruoja išreikštiniai klasių ir terminų sąryšiai.

Toliau, 3.2.6 skyrelyje pateiktas apibendrintas neformalus nagrinėtų algoritmų palyginimas.

3.2.3. Teksto dalių naudojimo įtaka klasifikavimo tikslumui

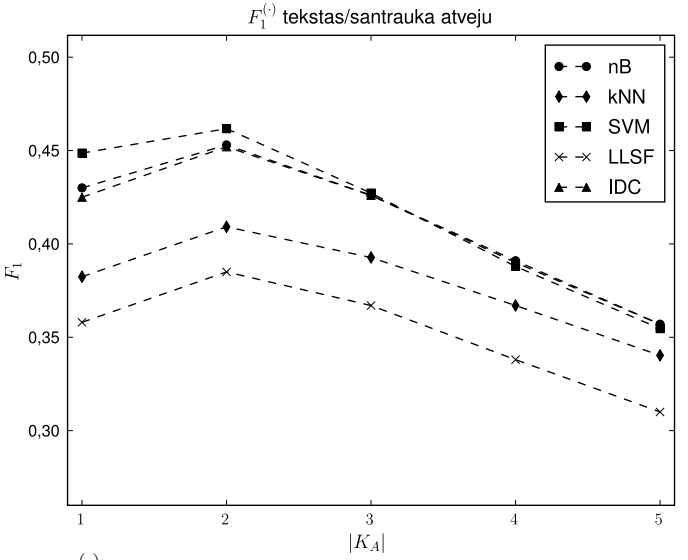
3.7 lentelėje ir 3.6–3.10 paveiksluose pateikti atskirų teksto dalių naudojimo mokymo ir testavimo fazėse įtakos klasifikavimo tikslumui tyrimo rezultatai. Iš jų darome šias išvadas:

- pavadinimas/pavadinimas atveju gaunami prasčiausi tikslumo rezultatai, tačiau ne tiek, kaip būtų galima tikėtis. nB, kNN ir IDC algoritmams jie panašūs į santrauka/santrauka ar santrauka/tekstas atvejus aibėje A^{KWD} (žr. 3.8 lentelę). Šioje situacijoje kNN algoritmas nenusileidžia arba tik nežymiai nusileidžia kitiems algoritmams. Kadangi poreikis apsiriboti pavadinimais mažai tikėtinas, šis atvejis praktinės reikšmės neturi;



3.3 pav. $F_1^{(\cdot)}$ priklausomybė nuo parenkamų klasių skaičiaus (santrauka / tekstas atveju)

- santrauka/santrauka ir santrauka/tekstas atvejais gaunami panašūs rezultatai. Nors antrasis variantas, klasifikavimui naudojantis pilnus tekstus, šiek tiek tikslesnis, tačiau dėl nedidelio skirtumo (iki 5 %) galima teigti, kad praktiniuose taikymuose dažniausias santrauka/santrauka atvejis yra visiškai priimtinas;
- perėjimas nuo santrauka/santrauka prie tekstas/santrauka atvejo duoda didesnį teigiamą efektą, nei aukščiau aptartas perėjimas prie santrauka/tekstas atvejo. Vienintelio $LLSF$ algoritmo atveju gaunami prastesni rezultatai, nei santrauka/santrauka ar santrauka/tekstas atveju. Šis variantas kaip ir santrauka/santrauka yra ganėtinai įprastas praktiniuose taikymuose ir kaip matyti iš rezultatų, leidžia pasiekti aukštesnius tikslumo rezultatus;
- tekstas/tekstas atveju, kaip ir buvo galima tikėtis, gaunami geriausi tikslumo rezultatai: lyginant su įprastu santrauka/santrauka atveju stebimas apie 10 % rezultatų pagerėjimas. Iš kitos pusės, šis variantas labiausiai imlus skaičiavimams dėl apdorojamų duomenų apimčių.



3.4 pav. $F_1^{(\cdot)}$ priklausomybė nuo parenkamų klasių skaičiaus (tekstas / santrauka atveju)

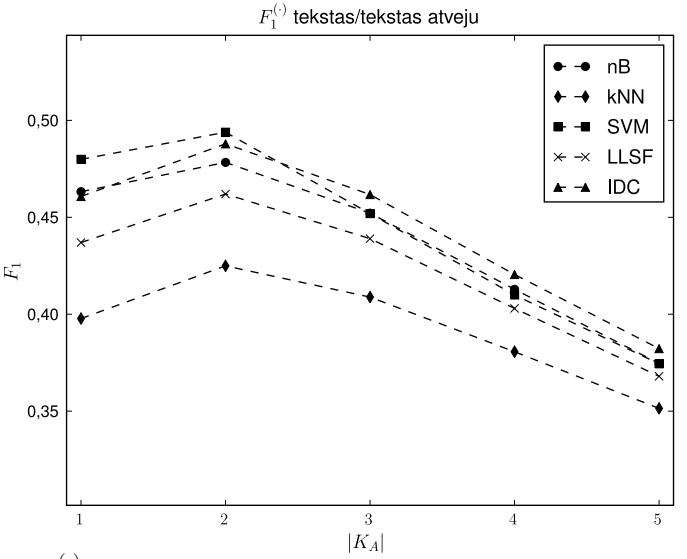
Apibendrinant galima teigti, kad pilniausias tekstas/tekstas atvejis leidžia pasiekti patį aukščiausią tikslumo lygį, o praktikoje dažnesni santrauka/santrauka ar tekstas/santrauka atvejais stebimi maždaug 5–10 % prastesni rezultatai.

3.2.4. Pasiūlytųjų algoritmų detali analizė

Skyriuje 2.5 apibrėžtas IDC algoritmas priklauso nuo šių elementų:

- tikimybių įvertinių glodinimo parametro μ, žr. (2.25);
- informatyviausių terminų indeksų aibės L atrinkimo metodo (žr. (2.29), (2.31), (2.32), (2.33), (2.34) ir (2.35)) bei nuo jo parametrų;
- parametrizacijos (2.37).

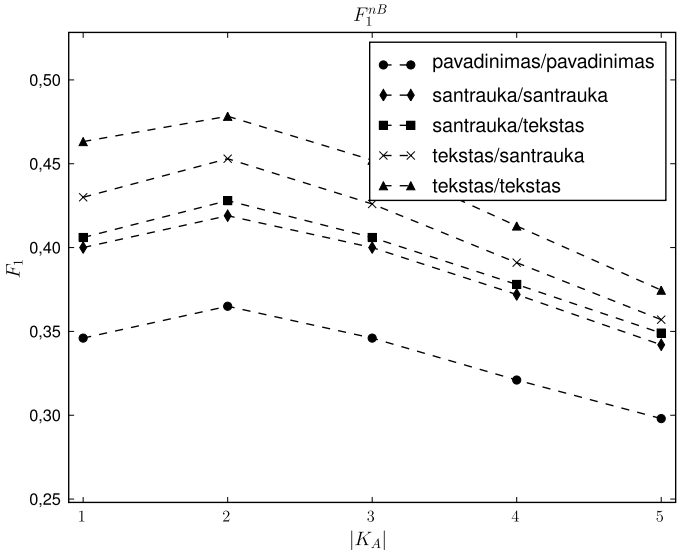
Atliekant bandymus su parametro μ reikšmėmis 1, 2, 3, 4, 5, 6 santrauka/santrauka ir tekstas/tekstas atvejais, kai kiti parametrai fiksuoti ($\alpha = 0{,}001$, $0{,}005$, $0{,}01$, $0{,}05$, hyp informatyvių terminų atrinkimo būdas, parametrizacija neatliekama) nustatyta, kad optimalus μ dydis yra tarp 2 ir 4 (žr. 3.9 lentelę, kurioje duomenys pateikti atvejui $\alpha = 0{,}01$).



3.5 pav. $F_1^{(\cdot)}$ priklausomybė nuo parenkamų klasių skaičiaus (tekstas / tekstas atveju)

Atliekant bandymus su parametro α reikšmėmis $0{,}001$, $0{,}005$, $0{,}01$, $0{,}05$ santrauka/santrauka ir tekstas/tekstas atvejais, kai kiti parametrai fiksuoti ($\mu = 2, 4$, *hyp* informatyvių terminų atrinkimo būdas, parametrizacija neatliekama) nustatyta, kad optimalus α dydis yra $0{,}001$ (žr. 3.10 lentelę).

Siekiant palyginti statistinių hipotezių tikrinimo teorija paremtas informatyvių terminų parinkimo procedūras *hyp*, *hyp/stop* ir *hyp/fixed*, atlikti bandymai santrauka/santrauka ir tekstas/tekstas atvejais su parametrais $\alpha = 0{,}001$, $0{,}005$, $0{,}01$, $0{,}05$, fiksuotais $\mu = 2, 3, 4$ ir neatliekant parametrizacijos. 3.13 ir 3.14 paveiksluose pateikti palyginimo rezultatai atvejui $\alpha = 0{,}01, \mu = 2$, kitais atvejais vaizdas panašus. Matyti, kad *hyp/stop* metodo tikslumas prasčiausias, o *hyp* – geriausias. *hyp/fixed* metodui priklausomai nuo naudojamų straipsnio dalių gaunami ganėtinai geri rezultatai. Ilgų tekstų (tekstas/tekstas) atveju jo ir *hyp* rezultatai beveik nesiskiria. Tai gali būti paaiškinta tuo, kad didelių mokymo imčių atveju svoriai įvertinami pakankamai tiksliai, todėl atvejų, kai labai didelis ar labai mažas svoris pasirodo nereikšmingai tesiskiriantis nuo vieneto, pasitaiko retai. Tokiu atveju *hyp/fixed* metodas, paliekantis pačius didžiausius ir pačius mažiausius svorius, beveik nesiskiria nuo *hyp* metodo.



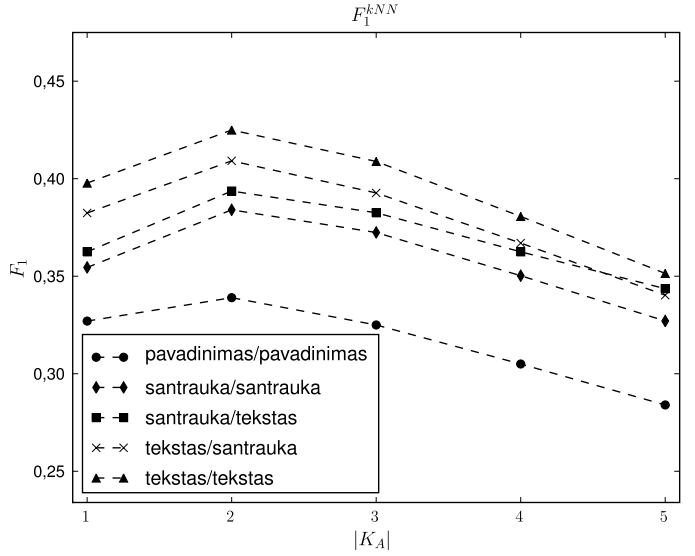
3.6 pav. F_1^{nB} priklausomybė nuo duomenų aibių

Pastebėsime, kad jis naudoja *hyp* metodo nustatytą paliekamų terminų su didžiausiais ir mažiausiais svoriais skaičių (skirtingas kiekvienai klasei), kuris kontroliuojamas reikšmingumo lygiu α. Bandant šiuos skaičius parinkti empiriškai (metodas *fixed*), susiduriama su problemomis dėl to, kad metodas nėra adaptyvus ir visoms klasėms naudoja tas pačias parametrų reikšmes, dėl ko tikslumas yra labai žemas (panašus, kaip ir *hyp/stop*).

Informatyviausių terminų atrinkimo etape, naudojant *hyp* metodą ir optimalias α bei μ reikšmes, realiai naudojamo (informatyviausių) terminų žodyno apimtis sumažėja dar 20–30 % nuo likusio po atlikto bendro sumažinimo *DF* metodu. Pastebėsime, kad šis sumažintas informatyviausių terminų žodynas yra skirtingas kiekvienai klasei.

(2.37) apibrėžta parametrizacija priklausomai nuo situacijos nežymiai pablogina tikslumo rezultatus. 3.11 lentelėje pateikti tikslumo priklausomybės nuo parametrizacijos prie skirtingų mokymo ir testavimo imčių bei parametrų α ir μ, atliekant informatyvių terminų atrinkimą *hyp* metodu. Visais atvejais parametrizuota tik informatyvių terminų aibės poaibio $\overline{L}$ dalis, nes atliekant $\underline{L}$ parametrizavimą stebimas ženklus tikslumo sumažėjimas.

Kaip ir tikėtasi, mažiausių kvadratų metodas ((2.39), $\beta = 2$) yra nestabilus. Rezultatus pavyko ženkliai pagerinti taikant euristinį patobulinimą, aprašytą 2.3.4 skyrelio paskutinėje pastraipoje.

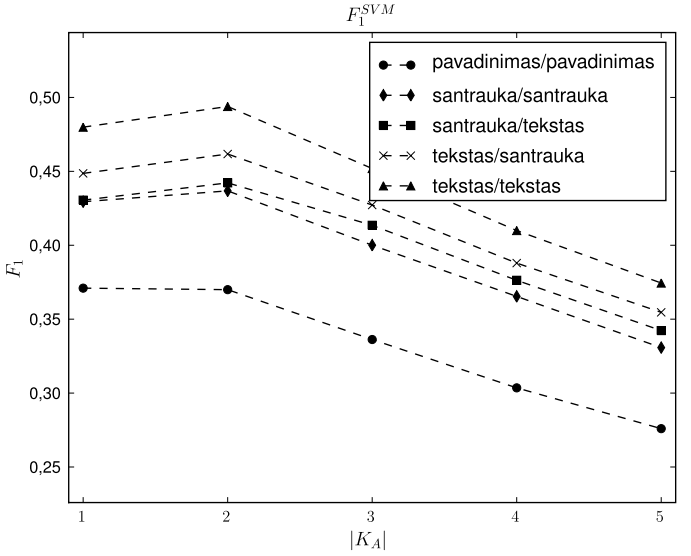


3.7 pav. F_1^{kNN} priklausomybė nuo duomenų aibių

3.9 lentelė. F_1^{IDC} priklausomybė nuo μ, $\alpha = 0{,}01$

Duomenys	μ reikšmė	Parenkamų klasių skaičius $\lvert K_A\rvert$				
		1	2	3	4	5
Santrauka/Santrauka	$\mu = 1$	0,397	0,423	0,402	0,375	0,344
	$\mu = 2$	0,408	0,429	0,407	0,377	0,348
	$\mu = 3$	0,407	0,430	0,406	0,376	0,347
	$\mu = 4$	0,406	0,429	0,406	0,376	0,347
	$\mu = 5$	0,403	0,427	0,404	0,375	0,341
	$\mu = 6$	0,400	0,422	0,401	0,372	0,338
Tekstas/Tekstas	$\mu = 1$	0,453	0,479	0,451	0,412	0,376
	$\mu = 2$	0,453	0,479	0,454	0,413	0,378
	$\mu = 3$	0,451	0,481	0,451	0,413	0,377
	$\mu = 4$	0,451	0,479	0,451	0,414	0,377
	$\mu = 5$	0,450	0,478	0,450	0,413	0,376
	$\mu = 6$	0,451	0,478	0,452	0,413	0,376

$IDCm$ algoritmas skiriasi nuo IDC tuo, kad jame atsižvelgiama ir į mokslo terminų poras. Algoritmas turi dar vieną parametrą, analogišką μ parametrui IDC atveju, atsakingą už glodinimą poroms (darbe naudota reikšmė 4).



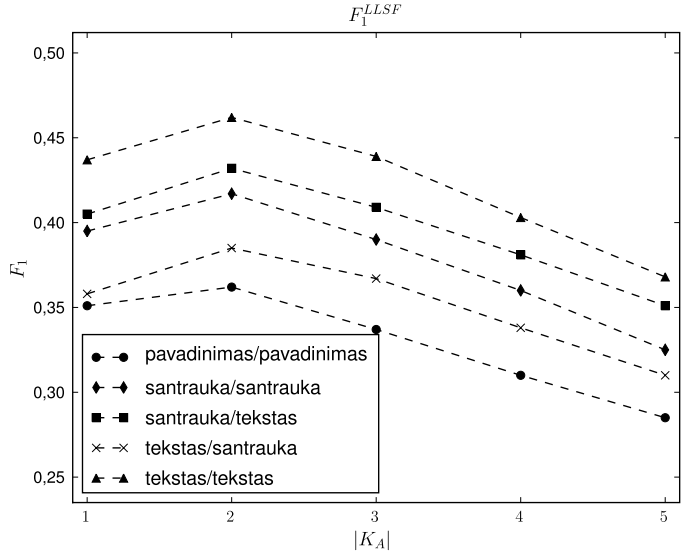
3.8 pav. F_1^{SVM} priklausomybė nuo duomenų aibių

3.10 lentelė. F_1^{IDC} priklausomybė nuo α, $\mu = 2$

Duomenys	α reikšmė	Parenkamų klasių skaičius $\|K_A\|$				
		1	2	3	4	5
Santrauka/Santrauka	$\alpha = 0{,}001$	0,409	0,427	0,406	0,376	0,346
	$\alpha = 0{,}005$	0,408	0,428	0,406	0,376	0,347
	$\alpha = 0{,}01$	0,408	0,429	0,407	0,377	0,348
	$\alpha = 0{,}05$	0,404	0,431	0,408	0,376	0,347
Tekstas/Tekstas	$\alpha = 0{,}001$	0,459	0,481	0,453	0,414	0,377
	$\alpha = 0{,}005$	0,454	0,480	0,453	0,412	0,377
	$\alpha = 0{,}01$	0,453	0,479	0,454	0,413	0,378
	$\alpha = 0{,}05$	0,451	0,479	0,455	0,414	0,377

Atlikti eksperimentai parodė, kad Markovo savybe paremto sudėtingesnio algoritmo $IDCm$, nenaudojančio papildomos informacijos naudojimas neprasmingas, nes klasifikavimo tikslumas sumažėja, lyginant su IDC algoritmu, o skaičiavimų sudėtingumas ir apimtys ženkliai (keletą kartų) išauga.

Naudojant papildomą kontekstinę informaciją poroms stebimas tikslumo padidėjimas, tačiau patikimoms išvadoms suformuluoti reikalingi papildomi išsamūs tyrimai su didesnėmis mokymo imtimis.



3.9 pav. F_1^{LLSF} priklausomybė nuo duomenų aibių

3.11 lentelė. Pr_{avg}^{IDC} priklausomybė nuo parametrizacijos naudojimo

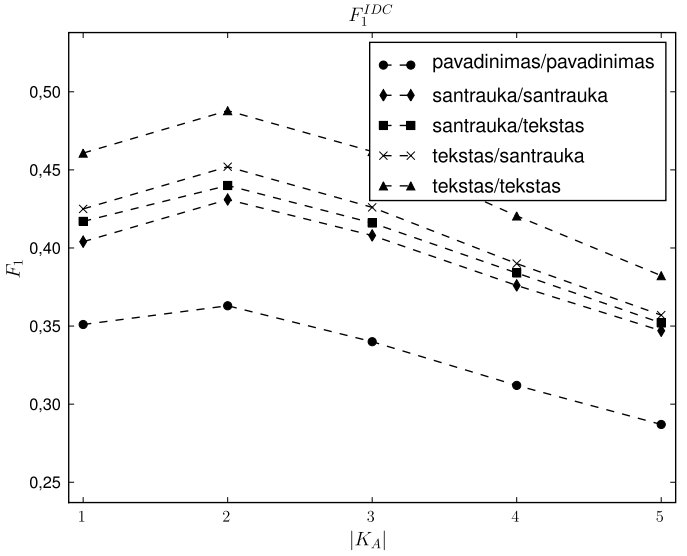
Duomenys	Parametrai	IDC variantas	
		Be parametrizacijos	Su parametrizacija
Santrauka/Santrauka	$\alpha = 0{,}001, \mu = 2$	0,592	0,590
	$\alpha = 0{,}001, \mu = 3$	0,592	0,587
	$\alpha = 0{,}001, \mu = 4$	0,589	0,583
Tekstas/Tekstas	$\alpha = 0{,}01, \mu = 2$	0,658	0,654
	$\alpha = 0{,}001, \mu = 3$	0,656	0,651
	$\alpha = 0{,}001, \mu = 4$	0,657	0,652

3.2.5. Pilnų tekstų ir papildomos informacijos naudojimo įtaka klasifikavimo tikslumui

Iki šiol „tekstas“ reiškė ne pilną tekstą, bet tik jo dalį, sudarytą iš pirmų 110 terminų. Toliau pateiksime rezultatus, gautus naudojant ilgesnius tekstus.

3.15 paveiksle pateikti algoritmų tikslumo rezultatai prie vis didėjančių tekstų ilgių, iš kurių galima padaryti šias išvadas:

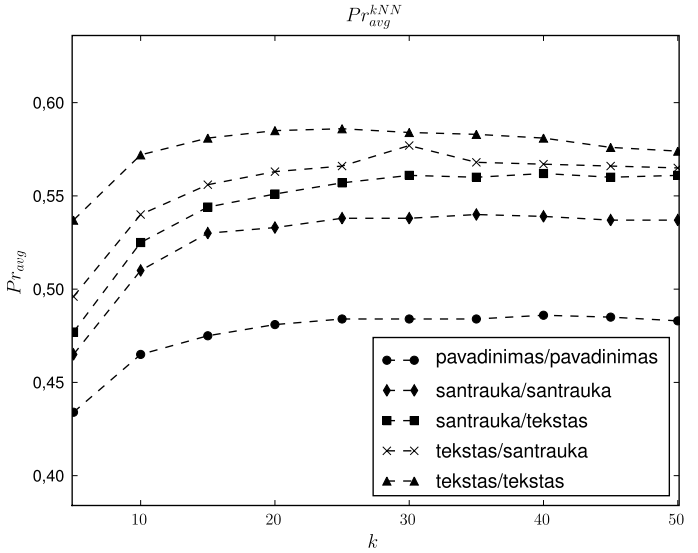
- naudojant vis ilgesnius tekstus netikimybinių algoritmų *SVM*, *LLSF* ir *kNN* bei tikimybinio *IDC* tikslumas nežymiai auga;



3.10 pav. F_1^{IDC} priklausomybė nuo duomenų aibių

- kai tekstų ilgis pasiekia maždaug 120 terminų, *nB* algoritmo tikslumas staigiai krenta. Taip nutinka dėl to, kad klasifikuojamame dokumente sutinkama pernelyg daug terminų, kurie buvo retai stebėti arba išvis nestebėti mokymo dokumentuose, jų svoriai yra labai arti nulio, ir tai iškraipo aposteriorinių tikimybių įverčius;
- *IDC* algoritmui ilgų tekstų atveju nuo tam tikrų ilgių ($\approx$ 200 terminų) stebimas rezultatų nebegerėjimas arba net blogėjimas.

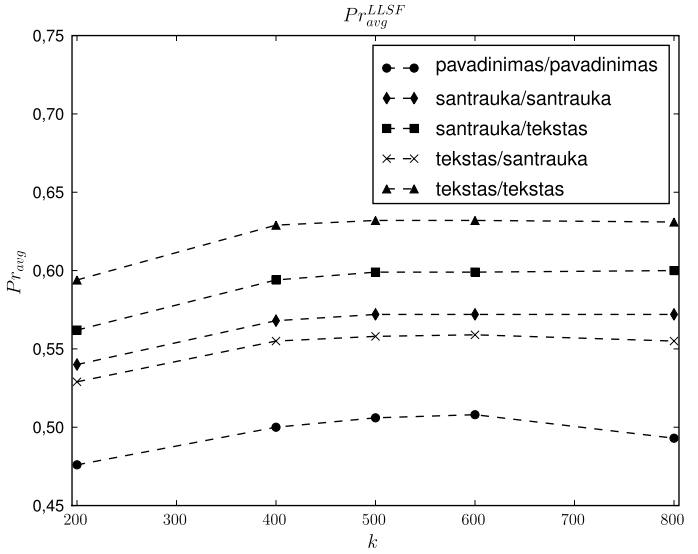
Toliau panagrinėsime 2.4 skyrelyje pasiūlytus *IDC* algoritmo apibendrinimus, leidžiančius atsižvelgti į papildomą kontekstinę informaciją, susijusią su terminų pozicijomis tekste. Papildomos informacijos įtaka apibrėžiama svorių funkcionalu (2.47) su parametru $\theta^{(\sigma)}$. Atliktas *IDC* algoritmo varianto, naudojančio ganėtinai paprastus papildomos informacijos naudojimo modelius (nukirtimas ties tam tikra pozicija, ženklus svorio sumažinimas ties tam tikra pozicija, palaipsniui mažėjantis svoris), tikslumo tyrimas. 3.12 lentelėje pateikti šio tyrimo rezultatai tekstas/tekstas atvejui. Pirmame lentelės stulpelyje pateiktos Pr_{avg}^{IDC} reikšmės, gautos atlikus „nukirtimą“, t. y., mokymui ir testavimui naudojant tik tą dalį teksto, kurią atitinka terminai su nenuliniais svoriais, apibrėžiamą $\theta^{(\sigma)}$ reikšmės paskutine komponente.



3.11 pav. Pr_{avg}^{kNN} priklausomybė nuo parametro k

Paskutiniajame stulpelyje pateikta atitinkama Pr_{avg}^{SVM} reikšmė tokiam pačiam „nukirtimo“ atvejui. Iš tyrimo rezultatų darome šias išvadas:

- neilgų tekstų (santrauka/santrauka ar tekstas/tekstas su mažiau kaip 150–200 pirmų terminų) atveju papildomos informacijos naudojimas įtakos klasifikavimo tikslumui neturi, pilnų tekstų atveju stebimas tikslumo padidėjimas;
- papildomos informacijos naudojimas tik mokymo fazėje rezultatų nepagerina, naudojimas testavime rezultatus šiek tiek pagerina, o naudojimas tiek mokyme, tiek testavime (su tuo pačiu parametru $\theta^{(\sigma)}$) duoda didžiausią tikslumo padidėjimą, siekiantį iki 2 %;
- iš tirtų papildomos informacijos panaudojimo modelių didžiausią papildomą tikslumą duoda tokio tipo modeliai: vienas–du ženklūs svorio sumažinimai kas 100–200 terminų, visiškas „nukirtimas“ ties 500 ar tolimesniais terminais. Nuoseklus svorio mažinimas nedideliais žingsneliais tikslumo beveik neįtakoja.



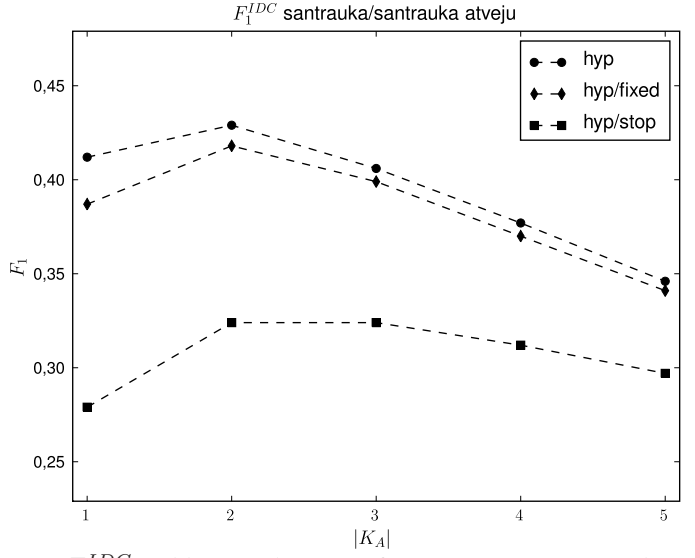
3.12 pav. Pr_{avg}^{LLSF} priklausomybė nuo parametro k

- svorių funkcionalo apibrėžime 2.47 vietoje $\lambda^{(t)}$ galima naudoti $\lambda^{(w)}$, $\lambda^{(s)}$ ar $\lambda^{(p)}$, t. y., termino poziciją skaičiuoti žodžio, sakinio ar pastraipos tikslumu. Tyrimas parodė, kad visais šiais atvejais gaunamos panašios tikslumo reikšmės, pagrindinis skirtumas – optimalaus parametro $\theta^{(\sigma)}$ parinkimas, kuris yra paprastesnis grubesnių $\lambda^{(s)}$ bei $\lambda^{(p)}$ atvejais.

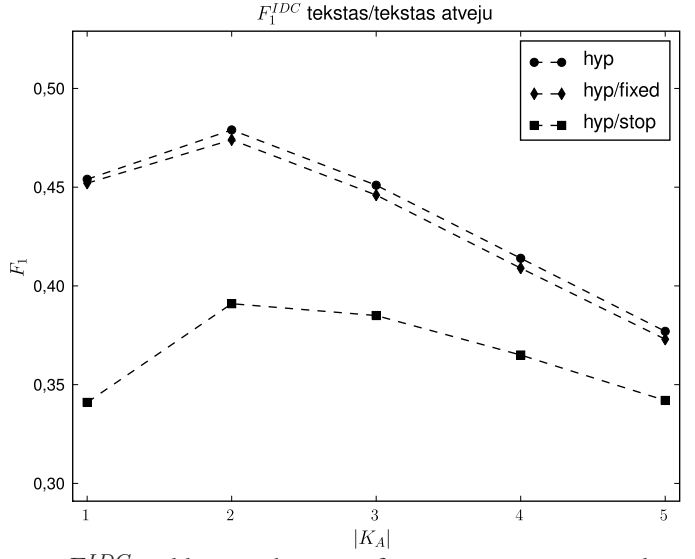
3.2.6. Neformalus algoritmų palyginimas

3.13 lentelėje pateiktas neformalus tirtų algoritmų palyginimas pagal keletą kriterijų: realizacijos paprastumą, tikslumą, mokymo ir klasifikavimo greitį, tekstų ilgių įtaką paskutiniosioms dviems savybėms bei rezultatų interpretuojamumą. Kuo daugiau žvaigždučių (nuo 1 iki 5), tuo algoritmas stipresnis toje srityje.

IDC žymi algoritmą, nenaudojantį papildomos informacijos, o *IDC*$^{(a)}$ – naudojantį. Jei „realizavimo paprastumo“ stulpelyje pateikti du vertinimai, pirmasis atitinka pilną realizavimą, remiantis tik algoritmo aprašymu, antrasis – realizavimą šiame tyrime, naudojantis jau sukurtomis procedūromis ir programomis. „Jautrumas tekstų ilgiui“ atspindi, kaip atitinkama charakteristika reaguoja pereinant prie vis ilgesnių tekstų: a) ar nenukenčia greitis ir b) ar išauga tikslumas. Paskutinioji „interpretuojamumo“ savybė atspindi, kiek algoritmo vidinės sąryšių struktūros yra matomos, prasmingos ir galbūt panaudojamos kitiems tikslams.



3.13 pav. F_1^{IDC} priklausomybė nuo informatyvių terminų atrinkimo metodo ir parenkamų klasių skaičiaus (santrauka/santrauka atveju)



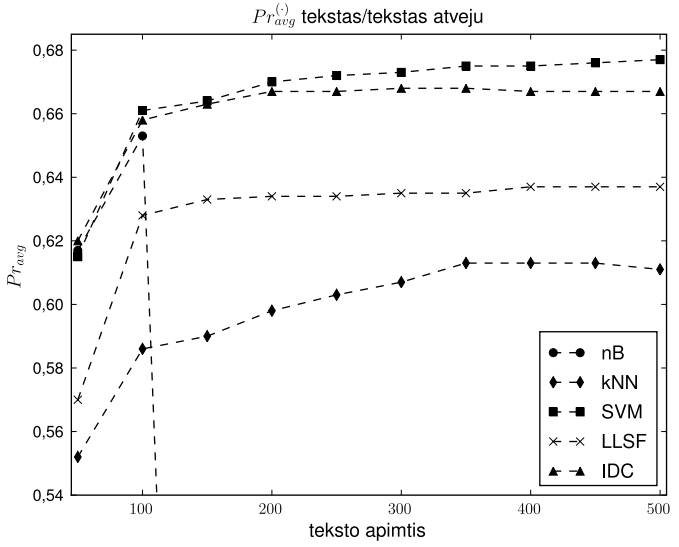
3.14 pav. F_1^{IDC} priklausomybė nuo informatyvių terminų atrinkimo metodo ir parenkamų klasių skaičiaus (tekstas/tekstas atveju)

3.12 lentelė. Pr_{avg}^{IDC} priklausomybė nuo papildomos informacijos naudojimo

Papildoma informacija naudojama...				Atitinkama Pr_{avg}^{SVM} reikšmė
nukirtimui	mokyme	testavime	visur	
$\theta^{(\sigma)} = (0,60,-0{,}01,110)$				
0,658	0,659	0,659	0,660	0,661
$\theta^{(\sigma)} = (0,110,-0{,}005,200)$				
0,667	0,667	0,668	0,668	0,670
$\theta^{(\sigma)} = (0,110,-0{,}5,111,0,200)$				
0,667	0,668	0,669	0,671	0,670
$\theta^{(\sigma)} = (0,110,-0{,}5,111,0,300)$				
0,668	0,668	0,671	0,673	0,673
$\theta^{(\sigma)} = (0,110,-0{,}5,111,0,200,-0{,}2,201,0,300)$				
0,668	0,669	0,672	0,675	0,673
$\theta^{(\sigma)} = (0,110,-0{,}5,111,0,400)$				
0,667	0,668	0,671	0,673	0,675
$\theta^{(\sigma)} = (0,110,-0{,}5,111,0,200,-0{,}2,201,0,400)$				
0,668	0,670	0,673	0,676	0,675
$\theta^{(\sigma)} = (0,110,-0{,}5,111,0,200,-0{,}3,201,0,400)$				
0,668	0,670	0,674	0,678	0,675
$\theta^{(\sigma)} = (0,110,-0{,}5,111,0,500)$				
0,667	0,669	0,673	0,676	0,677
$\theta^{(\sigma)} = (0,110,-0{,}5,111,0,200,-0{,}2,201,0,500)$				
0,667	0,670	0,675	0,679	0,677
$\theta^{(\sigma)} = (0,110,-0{,}5,111,0,200,-0{,}2,201,0,350,-0{,}2,351,0,500)$				
0,667	0,670	0,677	0,681	0,677

3.13 lentelė. Neformalus metodų palyginimas

Metodas	Savybė						
	realizacijos paprastumas	tikslumas	greitis		jautrumas tekstų ilgiui		interpr.
			mokymo	klasifikavimo	tikslumo	greičio	
nB	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆
kNN	⋆⋆⋆⋆	⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆
SVM	⋆ (⋆⋆⋆⋆)	⋆⋆⋆⋆⋆	⋆	⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆
$LLSF$	⋆ (⋆⋆⋆⋆)	⋆⋆⋆	⋆	⋆⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆
IDC	⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆
$IDC^{(a)}$	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆



3.15 pav. $Pr_{avg}^{(\cdot)}$ priklausomybė nuo ilgų tekstų naudojimo

3.3. Trečiojo skyriaus išvados

- Šiame skyriuje atliktas pasiūlytų bei alternatyvių algoritmų eksperimentinis tyrimas, naudojant 15000 tikimybių teorijos ir matematinės statistikos sričių mokslo publikacijų duomenų bazę.
- Tyrimas parodė, kad vietoje įprasto visų kalbos žodžių, aptinkamų straipsnių tekstuose, žodyno naudojant pavienių mokslo terminų žodyną (darbo autoriaus surinktą iš duomenų bazės straipsnių) gaunamas iki 5 % didesnis klasifikavimo tikslumas ir iki 2–3 kartų mažesnė žodyno apimtis. Prie pavienių terminų prijungus frazes, gaunamas iki 10 % didesnis tikslumas ir panaši žodyno apimtis, lyginant su tuo pačiu visų kalbos žodžių žodynu.
- Tyrimas parodė, kad IDC, nB ir SVM algoritmai tiksliausi, skirtumai tarp jų rezultatų dažniausiai nesieka 1 %. Tiesinio $LLSF$ algoritmo tikslumas mažesnis apie 5 %, o k kaimynų algoritmo tikslumas mažesnis apie 10–15 %.
- SVM algoritmas paprastai tikslesnis už konkurentus, kai straipsniui parenkama tik viena klasė, tačiau parenkant vis daugiau tikimybiniai nB ir IDC tampa tikslesni.
- Tikimybiniai nB ir IDC ne tik yra vieni tiksliausių, tačiau jie yra greitesni už sudėtingesnį SVM. Taip pat iš šių dviejų algoritmų galima ištraukti aiškiai interpretuojamus klasių ir terminų sąryšius. Tokie sąryšiai yra ir $LLSF$ algoritme, tačiau jie neturi aiškios interpretacijos.
- Tyrimas parodė, kad naudojant pilnus tekstus ir mokyme, ir klasifikavime, gaunamas maždaug 10–12 % didesnis algoritmų tikslumas, negu tuo atveju, kai naudojamos tik santraukos. Naudojant pilnus tekstus tik mokymo fazėje gaunamas apie 5–7 % tikslumo padidėjimas, o naudojant tik klasifikavimo fazėje, gaunamas iki 3–5 % tikslumo padidėjimas. Naudojant tik pirmus 110 pilno teksto terminų pasiekiama iki 90–95 % maksimalaus tikslumo (kuris pasiekiamas naudojant ilgesnius tekstus).
- Pasiūlytoji statistinių hipotezių tikrinimo teorija pagrįsta informatyviausių terminų atrinkimo metodika leidžia sumažinti naudojamą terminų žodyną apie 20–30 % (papildomai po atlikto bendro sumažinimo įprastais metodais), neprarandant arba net padidinant pasiūlytojo algoritmo klasifikavimo tikslumą.

- Autoriaus pasiūlytasis klasifikavimo metodas tiksliausias, kai remiamasi prielaida apie terminijos skirstinių stacionarumą ir nepriklausomumą. Prielaidos apie nepriklausomumą pakeitimas prielaida apie Markovo savybę lemia kelis kartus išaugusias skaičiavimų apimtis ir sumažėjusį tikslumą.
- Papildomos kontekstinės informacijos, susijusios su terminų pozicijomis tekste, panaudojimo tyrimas parodė, kad labai ilgų tekstų atveju naudojant žingsneliais mažėjančių svorių sistemas pasiekiamas iki 2 % didesnis pasiūlytojo algoritmo tikslumas, lyginant su algoritmo variantu, neatsižvelgiančiu į papildomą informaciją. Naudojant papildomą informaciją pasiūlytojo algoritmo tikslumas nenusileidžia ir net nežymiai (iki 1 %) lenkia alternatyvius tiksliausius algoritmus.

Bendrosios išvados

1. Moksliniams tekstams klasifikuoti skirtų specialių algoritmų nėra, paprastai tyrimuose ir praktiniuose taikymuose naudojami įprasti bendrinės kalbos tekstų klasifikavimo metodai.
2. Pritaikius dokumentų reprezentavimo skaitiniais požymių vektoriais metodus ir turint dideles teisingai suklasifikuotų dokumentų aibes, taikomąjį mokslo tekstų klasifikavimo uždavinį galima spręsti naudojant daugiamatės diskriminantinės analizės metodus.
3. Darbe suformuluotas tikimybinis mokslo terminijos pasiskirstymo publikacijų tekstuose modelis bei jo identifikavimo procedūros. Sukurti matematiniai metodai, kaip į klasifikavimą įtraukti papildomą kontekstinę informaciją, susijusią su terminų pozicijomis tekste ir kontekstu tarp jų. Sukurti konstruktyvūs klasifikavimo algoritmai, kurie gali būti taikomi mokslo publikacijų klasifikavimo uždaviniui spręsti.
4. Atliktas eksperimentinis tyrimas parodė, kad teksto elementų (žodžių, frazių), kuriuos atitinkantys skaitiniai požymiai reprezentuoja dokumentus, aibės parinkimas stipriai įtakoja algoritmų rezultatus. Naudojant iš tyrimams naudotos publikacijų duomenų bazės sudarytą mokslo terminijos (pavienių žodžių ir frazių) žodyną gautas apie 5–10 % didesnis tirtų klasifikavimo algoritmų tikslumas, nei naudojant įprastą visų kalbos

žodžių, sutinkamų straipsnių tekstuose, žodyną.

5. Pilnų tekstų (ne tik santraukų) naudojimas ir mokymo, ir klasifikavimo fazėse iki 10–12 % padidina tirtųjų algoritmų tikslumą. Apie 5–7 % tikslumo padidėjimas stebimas naudojant pilnus tekstus vien tik mokymo fazėje. Pilnų tekstų naudojimas tik klasifikavimo fazėje duoda iki 3–5 % tikslumo padidėjimą.

6. Pasiūlytos terminų svorių nustatymo bei informatyvių terminų atrinkimo procedūros leidžia apie 20–30 % sumažinti realiai naudojamo žodyno, jau sumažinto įprastu *DF* metodu, apimtis, neprarandant algoritmo tikslumo.

7. Prielaidos apie sudėtingesnes terminijos skirstinių tarpusavio sąryšio formas (negu sąlyginis nepriklausomumas ir stacionarumas) kelis kartus padidina tikimybinių algoritmų skaičiavimo sąnaudas, tačiau nepadidina tikslumo. Pasiūlytiems metodų apibendrinimams, leidžiantiems atsižvelgti į nutolusias terminų poras bei kontekstą tarp jų, ištirti reikalingos didesnės mokymo imtys.

8. Papildomos kontekstinės informacijos, susijusios su terminų pozicijomis tekste, panaudojimo metodų eksperimentinis tyrimas parodė, kad labai ilgų tekstų atveju naudojant žingsneliais mažėjančių svorių sistemas pasiekiamas iki 2 % didesnis pasiūlytojo algoritmo tikslumas, lyginant su algoritmu, nenaudojančiu papildomos informacijos. Šiuo atveju pasiūlytasis algoritmas nenusileidžia ir net nežymiai (iki 1 %) lenkia alternatyvius tiksliausius algoritmus.

9. Pasiūlytasis algoritmas turi svarbių aiškios rezultatų ir modelio parametrų sąryšių interpretacijos lemiamų taikomojo pobūdžio privalumų. Metodus galima papildyti ekspertinėmis žiniomis, o sukurti algoritmai ir terminijos sąryšių analizės rezultatai gali būti panaudoti sprendžiant kitus mokslo tekstų apdorojimo uždavinius.

Literatūra ir šaltiniai

Aha, D. W.; Kibler, D.; Albert, M. K. 1991. Instance-based learning algorithms, *Machine Learning*, 6(1): 37–66.

Aivazyan, S. A.; Buchstaber, V. M.; Yenukov, I. S.; Meshalkin, L. D. 1989. *Applied Statistics. Classification and Reduction of Dimensionality*. Moscow: Finansy i statistika. ISBN 5-279-00054-X.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Infomation Theory* 267–281.

Allen, D. 1977. The relationship between variable selection and data augmentation and a method of prediction, *Technometrics*, 16: 125–127.

Averin, A. V.; Vassilevskaya, L. A. 2002. *An Approach to Automatic Indexing of Scientific Publications in High Energy Physics for Database SPIRES HEP*.

Behboodian, J. 1970. On a mixture of normal distributions, *Biometrika*, 57(1): 215–217.

Bergström, A.; Jaksetic, P.; Nordin, P. 2000. Enhancing information retrieval by automatic acquisition of textual relations using genetic programming, in *Proc. of the 5th International Conference on Intelligent User Interfaces* 29–32.

Boser, B. E.; Guyon, I. M.; Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers, in *COLT '92: Proc. of the 5th Annual Workshop on Computational Learning Theory* 144–152.

Breiman, L. 1998. Arcing classifiers, *The Annals of Statistics*, 26(3): 801–824.

Burman, P. 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods, *Biometrika*, 76(5): 503–514.

Cohen, W. W.; Hirsh, H. 1998. Joins that generalize: text classification using Whirl, in *In Proc. of the 4th International Conference on Knowledge Discovery and Data Mining* 169–173.

Cohen, W. W.; Singer, Y. 1996. Context-sensitive learning methods for text categorization, in *Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 307–315.

Cooper, G. F.; Herskovits, E. 1992. A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, 9(4): 309–347.

Cortes, C.; Vapnik, V. 1995. Support-vector networks, *Machine Learning*, 20(3): 273–297.

Cover, T.; Hart, P. 1967. Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, 13(1): 21–27.

Cover, T. M.; Thomas, J. A. 1991. *Elements of Information Theory*. New York: Wiley-Interscience. ISBN 0-471-06259-6.

Craven, M.; Kumlien, J. 1999. Constructing biological knowledge bases by extracting information from text sources, in *Proc. of the 7th International Conference on Intelligent Systems for Molecular Biology* 77–86.

Creecy, R. H.; Masand, B. M.; Smith, S. J.; Waltz, D. L. 1992. Trading MIPS and memory for knowledge engineering, *Communications of the ACM*, 35(8): 48–64.

Dempster, A. P.; Laird, N. M.; Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–38.

Domingos, P.; Paffani, M. J. 1996. Beyond independence: Conditions for the optimality of the simple Bayesian classifier, in *International Conference on Machine Learning* 105–112.

Dumais, S.; Platt, J.; Heckerman, D.; Sahami, M. 1998. Inductive learning algorithms and representations for text categorization, in *Proc. of the 7th International Conference on Information and Knowledge Management* 148–155.

Eckart, C.; Young, G. 1936. The approximation of one matrix by another of lower rank, *Psychometrika*, 1(3): 211–218.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife, *The Annals of Statistics*, 7(1): 1–26.

Efron, B.; Tibshirani, R. 1997. Improvements on cross-validation: The .632+ bootstrap method, *Journal of the American Statistical Association*, 92(438): 548–560.

Ewing, J. 2002. Twenty centuries of mathematics: digitizing and disseminating the past mathematical literature, *Notices of the American Mathematical Society*, 49(7).

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research*, 3: 1289–1305.

Friedman, J. H. 1994. Flexible metric nearest neighbor classification. Technical report, Dept. of Statistics, Stanford University.

Friedman, J. H. 1997. On bias, variance, 0/1—loss, and the curse-of-dimensionality, *Data Mining and Knowledge Discovery*, 1(1): 55–77.

Friedman, N.; Geiger, D.; Goldszmidt, M. 1997. Bayesian network classifiers 131–163.

Ghamrawi, N.; McCallum, A. 2005. Collective multi-label classification, in *CIKM '05: Proc. of the 14th ACM International Conference on Information and Knowledge Management* 195–200.

Guthrie, L.; Walker, E.; Guthrie, J. 1994. Document classification by machine: theory and practice, in *Proc. of the 15th Conference on Computational Linguistics* 1059–1063.

Harter, S. P. 1975a. A probabilistic approach to automatic keyword indexing. Part I. On the distribution of specialty words in a technical literature, *Journal of the American Society for Information Science and Technology*, 26(4): 197–206.

Harter, S. P. 1975b. A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing, *Journal of the American Society for Information Science and Technology*, 26(5): 280–289.

Hasselbland, V. 1966. Estimation of parameters of for a mixture of normal distributions, *Technometrics*, 8: 431–444.

Hastie, T.; Tibshirani, R.; Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. ISBN 0-387-95284-5.

Hazewinkel, M. 2005. Dynamic stochastic models for indexes and thesauri, identification clouds, and information retrieval and storage, in Baeza-Yates, R.; Glaz, J.; Gzyl, H.; Hüsler, J.; Palacios, J. L. (eds.), *Recent Advances in Applied Probability* 181–204. Springer US.

Hazewinkel, M. 1999. Topologies and metrics on information spaces, *CWI Quarterly*, 12: 93–110.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features, in *Proc. of ECML-98, 10th European Conference on Machine Learning* 137–142.

Joachims, T. 1999. Transductive inference for text classification using support vector machines, in *Proc. of the 16th Int. Conf. on Machine Learning* 200–209.

Katz, S. M. 1996. Distribution of content words and phrases in text and language modelling, *Nat. Lang. Eng.*, 2(1): 15–59.

Kim, S.-B.; Rim, H.-C.; Yook, D.; Lim, H. 2002. Effective methods for improving naive Bayes text classifiers, in *PRICAI '02: Proc. of the 7th Pacific Rim International Conference on Artificial Intelligence* 414–423.

Krallinger, M.; Valencia, A. 2005. Text-mining and information-retrieval services for molecular biology, *Genome Biology*, 6(7).

Kullback, S.; Leibler, R. A. 1951. On information and sufficiency, *Annals of Mathematical Statistics*, 22: 49–86.

Langley, P.; Sage, S. 1994. Induction of selective Bayesian classifiers, in *Proc. of the 10th Conference on Uncertainty in Artificial Intelligence* 399–406.

Larkey, L. S.; Croft, W. B. 1996. Combining classifiers in text categorization, in *SIGIR '96: Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 289–297.

Lewis, D. D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval, in *ECML '98: Proc. of the 10th European Conference on Machine Learning* 4–15.

Li, H.; Yamanishi, K. 2002. Text classification using ESC-based stochastic decision lists, *Information Processing and Management*, 38(3): 343–361.

Li, Y. H.; Jain, A. K. 1998. Classification of text documents, *The Computer Journal*, 41(8): 537–546.

Liu, H.; Yu, L. 2005. Toward integrating feature selection algorithms for classification and clustering, *IEEE Transactions on Knowledge and Data Engineering*, 17(4): 491–502.

Masand, B. 1994. Optimizing confidence of text classification by evolution of symbolic expressions, *Advances in Genetic Programming* 445–458.

Masand, B.; Linoff, G.; Waltz, D. 1992. Classifying news stories using memory based reasoning, in *SIGIR '92: Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 59–65.

McCallum, A.; Nigam, K. 1998. *A comparison of event models for Naive Bayes text classification*.

Metzler, D.; Croft, W. B. 2005. A Markov random field model for term dependencies, in *SIGIR '05: Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 472–479.

Ng, H. T.; Goh, W. B.; Low, K. L. 1997. Feature selection, perceptron learning, and a usability case study for text categorization, in *SIGIR '97: Proc. of the 20th annual international ACM SIGIR conference on Research and development in information retrieval* 67–73.

Peng, F.; Schuurmans, D. 2003. Combining naive Bayes and n-gram language models for text classification, in *Advances in Information Retrieval*, volume 2633 of *Lecture Notes in Computer Science* 335–350.

Pickens, J.; MacFarlane, A. 2006. Term context models for information retrieval, in *CIKM '06: Proc. of the 15th ACM International Conference on Information and Knowledge Management* 559–566.

Rehurek, R.; Sojka, P. 2008. Automated classification and categorization of mathematical knowledge, 5144: 543–557.

Rennie, J. D. M.; Lawrence, S.; Teevan, J.; Karger, D. R. 2003. Tackling the poor assumptions of naive Bayes text classifiers, in *Proc. of the 20th International Conference on Machine Learning* 616–623.

Robertson, S. E.; Jones, K. S. 1988. Relevance weighting of search terms, in *Document Retrieval Systems* 143–160.

Robertson, S. E.; Walker, S. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, in *SIGIR '94: Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 232–241.

Rogati, M.; Yang, Y. 2002. High-performing feature selection for text classification, in *CIKM '02: Proc. of the 11th International Conference on Information and Knowledge Management* 659–661.

Ruiz, M. E.; Srinivasan, P. 1999. Hierarchical neural networks for text categorization, in *SIGIR '99: Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 281–282.

Salton, G.; Buckley, C. 1988. Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, 24(5): 513–523.

Salton, G.; McGill, M. J. 1986. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, Inc. ISBN 0-070-54484-0.

Schapire, R. E.; Singer, Y. 2000. BoosTexter: A boosting-based system for text categorization, *Machine Learning*, 39(2-3): 135–168.

Schapire, R. E.; Freund, Y.; Bartlett, P.; Lee, W. S. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods, *The Annals of Statistics*, 26(5): 1651–1686.

Schwarz, G. 1978. Estimating the dimension of a model, *The Annals of Statistics*, 6(2): 461–464.

Sebastiani, F. 2002. Machine learning in automated text categorization, *ACM Computing Surveys*, 34(1): 1–47.

Shatkay, H.; Pan, F.; Rzhetsky, A.; Wilbur, W. J. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users, *Bioinformatics*, 24(18): 2086–2093.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society: Series B*, 36: 111–147.

Rijsbergen, C. van. 1979. *Information Retrieval*. Newton: Butterworth-Heinemann. ISBN 0-408-70929-4.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag. ISBN 0-387-94559-8.

Wang, Z.; Webb, G. I. 2002. Comparison of lazy Bayesian rule and tree-augmented Bayesian learning, in *ICDM '02: Proc. of the 2002 IEEE International Conference on Data Mining* 490.

Webb, G. I.; Boughton, J. R.; Wang, Z. 2005. Not so naive Bayes: aggregating one-dependence estimators, *Machine Learning*, 58(1): 5–24.

Weiss, S. M.; Apte, C.; Damerau, F. J.; Johnson, D. E.; Oles, F. J.; Goetz, T.; Hampp, T. 1999. Maximizing text-mining performance, *IEEE Intelligent Systems*, 14(4): 63–69.

Widdows, D. 2003. A mathematical model for context and word-meaning, in *4th International and Interdisciplinary Conf. on Modeling and Using Context*.

Wiener, E. D.; Pedersen, J. O.; Weigend, A. S. 1995. A neural network approach to topic spotting, in *SDAIR-95: Proc. of the 4th Annual Symposium on Document Analysis and Information Retrieval* 317–332.

Yan, R.; Tesic, J.; Smith, J. R. 2007. Model-shared subspace boosting for multi-label classification, in *KDD '07: Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 834–843.

Yang, Y. 1994. Expert network: effective and efficient learning from human decisions in text categorization and retrieval, in *SIGIR '94: Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 13–22.

Yang, Y.; Chute, C. G. 1992. A linear least squares fit mapping method for information retrieval from natural language texts, in *Proc. of the 14th Conference on Computational Linguistics* 447–453.

Yang, Y.; Chute, C. G. 1993. An application of least squares fit mapping to text information retrieval, in *Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 281–290.

Yang, Y.; Chute, C. G. 1994. An example-based mapping method for text categorization and retrieval, *ACM Trans. Inf. Syst.*, 12(3): 252–277.

Yang, Y.; Pedersen, J. O. 1997. A comparative study on feature selection in text categorization, in *ICML '97: Proc. of the 14th International Conference on Machine Learning* 412–420.

Zaiane, O. R.; Antonie, M.-L. 2002. Classifying text documents by associating terms with text categories, in *ADC '02: Proc. of the 13th Australasian Database Conference* 215–222.

Zhang, P. 1993. Model selection via multifold cross validation, *Annals of Statistics*, 21(1): 299–313.

Zheng, Z.; Webb, G. I. 2000. Lazy learning of Bayesian rules, *Machine Learning*, 41(1): 53–84.

Autoriaus publikacijos disertacijos tema

Straipsniai recenzuojamuose mokslo žurnaluose

Balys, V.; Rudzkis, R. 2008. Classification of Publications Based on Statistical Analysis of Scientific Terms Distributions, *Austrian Journal of Statistics* 37(1): 109–118.

Rudzkis, R.; Balys, V.; Hazewinkel, M. 2006. Stochastic Modelling of Scientific Terms Distribution in Publications, in *Proceedings of 5th International Conference on Mathematical Knowledge Management*, *Lecture Notes in Artificial Intelligence* 4108: 152–164 (ISI Web of Science).

Balys, V.; Rudzkis, R. 2005. Stochastinių mokslo terminų aplinkų modelių tyrimas, *Lietuvos matematikos rinkinys, spec. nr.* 45: 329–334.

Balys, V.; Rudzkis, R. 2004a. Mokslo terminų aplinkų modelių taikymas straipsnių klasifikavime, *Lietuvos matematikos rinkinys, spec. nr.* 44: 537–541.

Balys, V.; Rudzkis, R. 2003. Mokslinių terminų statistinio pasiskirstymo taikymas straipsnių klasifikavime, *Lietuvos matematikos rinkinys, spec. nr.* 43: 463–467.

Straipsniai kituose mokslo leidiniuose

Rudzkis, R.; Balys, V. 2007. On Statistical Classification of Scientific Texts, in *Proceedings of 8th International Conference on Computer Data Analysis and Modeling* 1: 100–103.

Balys, V.; Rudzkis, R. 2004b. Stochastic Models for Keyphrase Assignment, in *Proceedings of 7th International Conference on Computer Data Analysis and Modeling* 2: 118–122.

Priedas

Eksperimente naudoti MSC klasifikatoriai ir raktiniai žodžiai

MSC klasifikatoriai

60B10, 60E05, 60E15, 60F05, 60F10, 60F15, 60F17, 60F99, 60G10, 60G15, 60G17, 60G40, 60G50, 60H05, 60H10, 60J10, 60J15, 60J25, 60J55, 60J60, 60J65, 60J80, 60K25, 60K35, 62C10, 62C15, 62E10, 62E20, 62F05, 62F10, 62F12, 62F15, 62F35, 62G05, 62G07, 62G10, 62G20, 62G30, 62G35, 62H99, 62J05, 62K05, 62L10, 62M10

Raktiniai žodžiai

admissibility, asymptotic distribution, asymptotic efficiency, asymptotic expansion, asymptotic normality, bootstrap, brownian motion, censored data, central limit theorem, confidence interval, consistency, contact process, convergence rate, density estimation, edgeworth expansion, efficiency, empirical process, estimation, exponential families, gaussian processes, invariance principle, large deviation, law of iterated logarithm, local time, markov chain, markov process,

martingale, maximum likelihood, maximum likelihood estimation, nonparametric regression, optimal stopping, order statistic, percolation, poisson process, random walk, rate of convergence, regression, regular variation, robustness, stationary process, stochastic integral, stopping time, time series, weak convergence

Vaidas BALYS

MOKSLINĖS TERMINIJOS MATEMATINIAI MODELIAI
IR JŲ TAIKYMAS LEIDINIŲ KLASIFIKAVIME

Daktaro disertacija

Fiziniai mokslai,
matematika (01P)

Vaidas BALYS

MATHEMATICAL MODELS FOR SCIENTIFIC TERMINOLOGY
AND THEIR APPLICATIONS IN THE CLASSIFICATION OF PUBLICATIONS

Doctoral Dissertation

Physical Sciences,
Mathematics (01P)

2009 07 17. 9,5 sp. l. Tiražas 20 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“,
Saulėtekio al. 11, 10223 Vilnius,
http://leidykla.vgtu.lt
Spausdino UAB „Biznio mašinų kompanija“,
J. Jasinskio g. 16A, 01112 Vilnius