VILNIAUS UNIVERSITETAS
Gamtos mokslai, Informatika N 009

Kauno fakultetas

Atraminių vektorių mašinų parametrų derinimas grindžiamas dalelių spiečių optimizavimo euristika tekstinių duomenų klasifikavimui

Doktorantūros studijų metai

2015m. spalio mėn. 1d. – 2019m. rugsėjo mėn. 30d.

Darbo vadovas: prof. dr. Gintautas Garšva

Ataskaita

2015–2017 m. m. išlaikyti egzaminai:

	Dalyko pavadinimas	Kreditų skaičius ECTC	Atsiskaitymo data	Dalyko konsultantas	Įvertinimas
1	Informatikos ir informatikos inžinerijos tyrimo metodai ir metodika	9	2016.06 (2016.06.09)	Prof. dr. A. Čaplinskas	**8**
2	Skaitinis intelektas	7	2016.11 (2016.09.12)	Doc. dr. V. Rudžionis	**9**
3	Sistemų analizės technologijos	7	2017.03 (2016.10.20)	Prof. dr. S. Gudas	**9**
4	Duomenų analizės strategijos ir sprendimų priėmimas	7	2017.06 (2016.12.19)	Prof. habil. dr. G. Dzemyda Prof. dr. O. Kurasova dr. J. Bernatavičienė	**9**

Ataskaita

Publikacijos 2015–2019 m. m.:

Recenzuojami periodiniai mokslo leidiniai

1) Korovkinas, K., Danėnas, P., Garšva, G. *SVM and Naïve Bayes Classification Ensemble Method for Sentiment Analysis*. Baltic Journal of Modern Computing, 5(4), pp.398-409, 2017.

2) Korovkinas, K., Danėnas, P., Garšva, G. *SVM and k-Means Hybrid Method for Textual Data Sentiment Analysis*. Baltic Journal of Modern Computing, 7(1), pp. 47–60, 2019

3) Korovkinas, K., Danėnas, P., Garšva, G. *Support Vector Machine Parameter Tuning Based on Particle Swarm Optimization Metaheuristic*. Nonlinear Analysis: Modelling and Control, in Press (2020 m.).

Publikacijos recenzuojamuose konferencijų leidiniuose

4) Korovkinas, K., Garšva, G., 2018. *Selection of intelligent algorithms for sentiment classification method creation*. Proceedings of the International Conference on Information Technologies, Vol–2145, Kaunas, Lithuania, pp. 152–157, ISSN 1613-0073, CEUR. Available: http://ceur-ws.org/Vol-2145/p26.pdf

5) Vaitonis, M., Masteika, S., Korovkinas, K. 2018. *Algorithmic trading and machine learning based on GPU.* Proceedings of the Symposium for Young Scientists in Technology, Engineering and Mathematics, Vol–2147, Gliwice, Poland, pp. 49–54, ISSN 1613-0073, CEUR. Available: http://ceur-ws.org/Vol-2147/p09.pdf

6) Korovkinas, K., Danėnas, P., Garšva, G., 2018. *SVM accuracy and training speed tradeoff in sentiment analysis tasks*. In International Conference on Information and Software Technologies (pp. 227–239). Springer, Cham.

Ataskaita

Pranešimai konferencijose

1) Information Technologies (IT2018), The 23th Conference for Master and PhD students, Kaunas, Lithuania, 2018.

2) The Symposium for Young Scientists in Technology, Engineering and Mathematics (SYSTEM2018), The 23th Conference for Master and PhD students, Gliwice, Poland, 2018.

3) The 24th International Conference on Information and Software Technologies (ICIST 2018), Kaunas, Lithuania, 2018.

4) 10th International Workshop Data Analysis Methods for Software Systems (DAMSS 2018), Druskininkai, Lithuania, 2018.

Santraukos konferencijų leidiniuose:

Korovkinas, K., Garšva, G., 2018. *Large Scale Sentiment Analysis Using NLP Based Feature Extraction Technique and PSOLinearSVM*. Data analysis methods for software systems: 10th international workshop, Druskininkai, 2018. ISBN 978-609-07-0043-3. Available: https://www.mii.lt/datamss/files/DAMSS_2018_1.pdf

Ataskaita

Pranešimai seminaruose

1) Kauno fakulteto doktorantų tarpdisciplininis seminaras Palangoje. Pristatyta disertacija.

2) VU Duomenų Mokslo ir Skaitmeninių Technologijų Instituto organizuotas Informatikos inžinerijos problemų seminaras: "**Mašininis mokymasis ir nuotaikų analizė**". Pristatytas straipsnis "*Atraminių vektorių ir naivaus Bajeso klasifikavimo hibridinio metodo taikymas sentimentų analizei*".

3) VU Duomenų Mokslo ir Skaitmeninių Technologijų Instituto organizuotuose grupių seminaruose. Pristatyta disertacija.

Ataskaita

Kita veikla 2015–2018 m. m.:

2016m. Bakalauro kursinio darbo vadovas (darbas įvertintas 8) ir baigiamojo darbo vadovas (darbas įvertintas 10).

2017m. Bakalauro kursinio darbo vadovas (darbas įvertintas 7) ir baigiamojo darbo vadovas (darbas įvertintas 9).

2018m. Kauno fakultete dėstau „Kompiuterių architektūrą“ anglų ir lietuvių grupėms.

2019m. Kauno fakultete dėstau „Python programavimą” ir „Kompiuterių architektūrą“ anglų ir lietuvių grupėms.

Tyrimo objektas ir tikslas

Tyrimo objektas:

nuomonių klasifikavimas didelės apimties tekstiniuose duomenų masyvuose, paremtas linijinių atraminių vektorių mašinų (angl. Linear Support Vector Machines) klasifikavimo metodais.

Tyrimo tikslas:

pasiūlyti linijinių atraminių vektorių mašinomis grindžiamą metodą nuomonių klasifikavimui, naudojant didelės apimties tekstinius duomenis bei galimą šio metodo realizavimo scenarijų.

Uždaviniai

Tyrimo tikslui įgyvendinti iškelti šie uždaviniai:

1) Apžvelgti metodus naudojamus nuomonių klasifikavimui tekstiniuose duomenų masyvuose ir pasirinkti dažniausiai naudojamus, remiantis literatūros analize.
2) Lyginamosios analizės būdu palyginti atrinktus metodus, eksperimentams naudojant didelės apimties tekstinius duomenų masyvus.
3) Pasiūlyti nuomonių klasifikavimo metodą tekstiniuose duomenų masyvuose su geresniu klasifikavimo tikslumu ir vykdymo laiku.
4) Atlikti eksperimentinį tyrimą su pasiūlytu metodu ir atrinkti mašininio mokymo algoritmą, kuriam jis tiks labiausiai.
5) Suprojektuoti ir realizuoti pasiūlytą metodą realiai problemai – nuomonių ir rinkos tyrimams.

Mokslinis naujumas

1) Pasiūlytas metodas nuomonių klasifikavimui tekstiniuose duomenų masyvuose.
2) Skirtingai nuo kitų autorių pasiūlytas metodas turi keturias pagrindines dalis, kurios gali būti integruojamos į standartinį mašininio mokymo metodą nepriklausomai viena nuo kitos:
   a) spartinimo metodas;
   b) klasterių skaičiaus ir mokymo duomenų parinkimo metodas;
   c) parametrų derinimo metodas;
   d) metodų apjungimo (angl. ensemble) metodas.
3) Pasiūlytas metodas gali būti taikomas tekstinių duomenų klasifikavimui skirtingose srityse, taip pat darbui su dideliais tekstinių duomenų masyvais, tam nenaudojant galingų superkompiuterių; visi eksperimentai šioje disertacijoje yra atlikti naudojant nešiojamą kompiuterį.

Darbo praktinė reikšmė

Tekstinių nuomonių klasifikavimas yra iššūkių pilna sritis ir tuo pačiu labai aktuali praktikoje, plačiai naudojama:

- prekių apžvalgoms
- klientų lojalumo prognozėms
- sukčiavimo aptikimams
- rinkimams
- ir t.t.

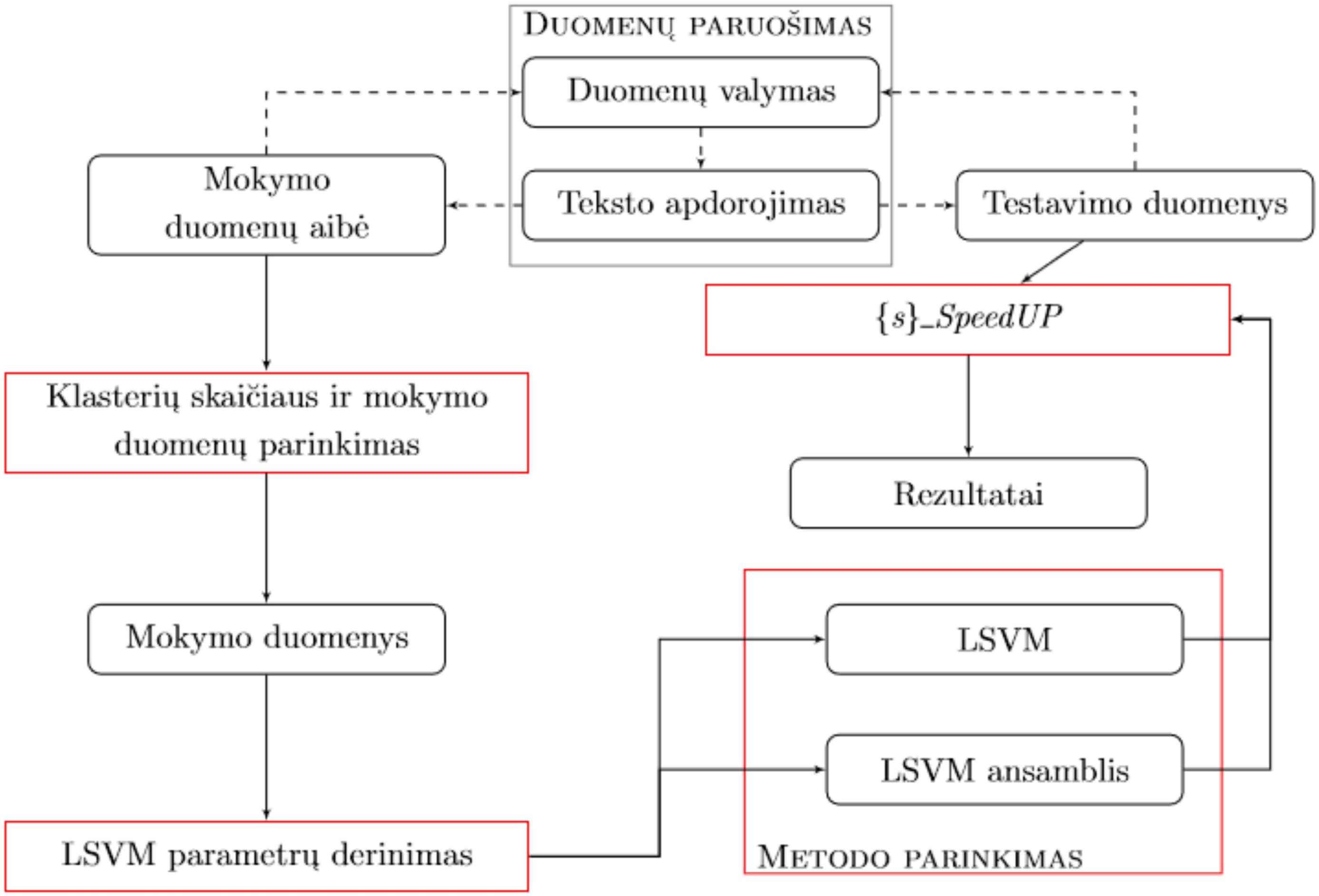
Pasiūlytas metodas gali būti sėkmingai pritaikytas šiose srityse naujų modelių kūrimui ar esamų tobulinimui.

Disertacijos ginamieji teiginiai

1) Pasiūlytas metodas turi keturias pagrindines dalis, kurios gali būti integruojamos atskirai nepriklausomai viena nuo kitos, tekstinių duomenų klasifikavimui.
2) Spartinimo metodas gali būti sėkmingai integruojamas į logistinę regresiją, linijines atraminių vektorių mašinas (LSVM), atsitiktinį mišką ir sprendimų medį.
3) k-vidurkių ir dalelių spiečiaus optimizavimo metodai gali padidinti klasifikavimo tikslumą logistinei regresijai ir linijinėms atraminių vektorių mašinoms.
4) Pilnas pasiūlytas metodas gali būti integruojamas į logistinę regresiją ir linijines atraminių vektorių mašinas, tačiau geriausias rezultatas pasiekiamas integruojant jį į LSVM.
5) Pasiūlytas metodas gali būti sėkmingai pritaikytas sprendžiant realias problemas su realiais duomenimis, netgi naudojant tekstines duomenų aibes lietuvių kalba.

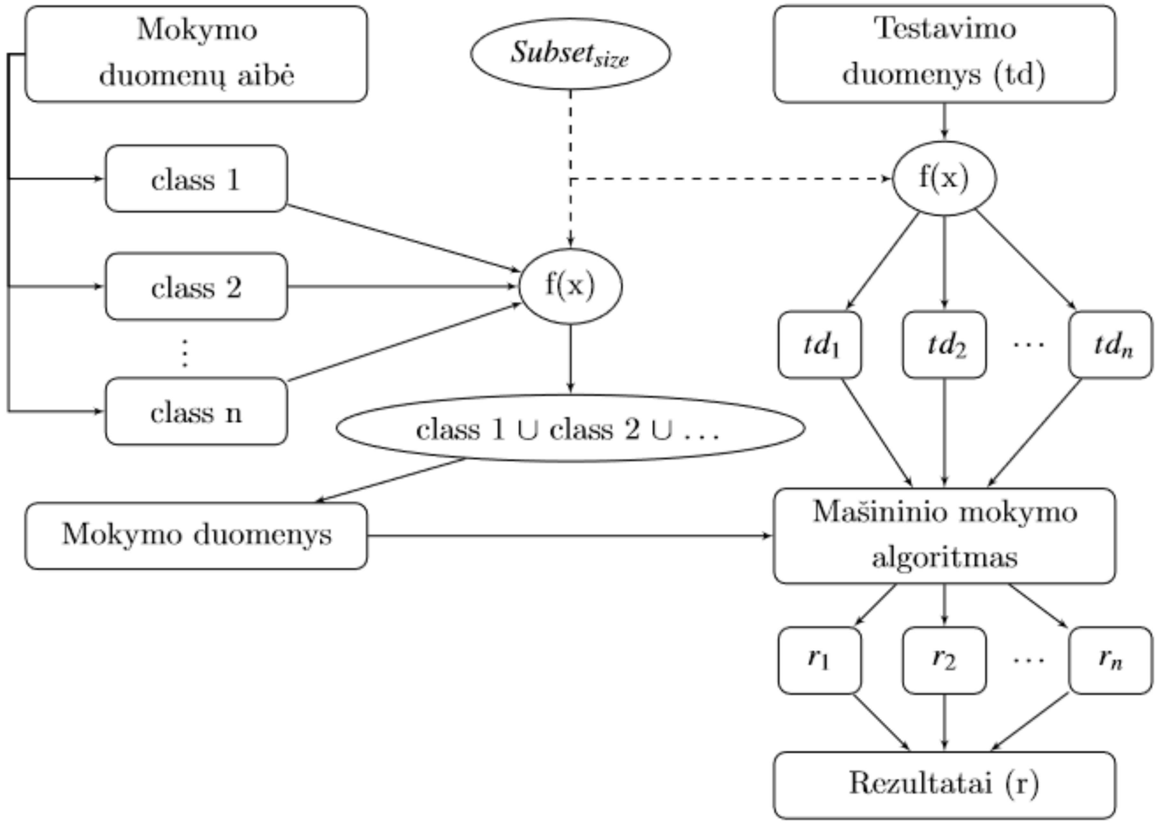
Siūlomas metodas

Siūlomo metodo diagrama



Siūlomas metodas

Spartinimo metodo diagrama



Siūlomas metodas

Klasterių skaičiaus ir mokymo duomenų parinkimas

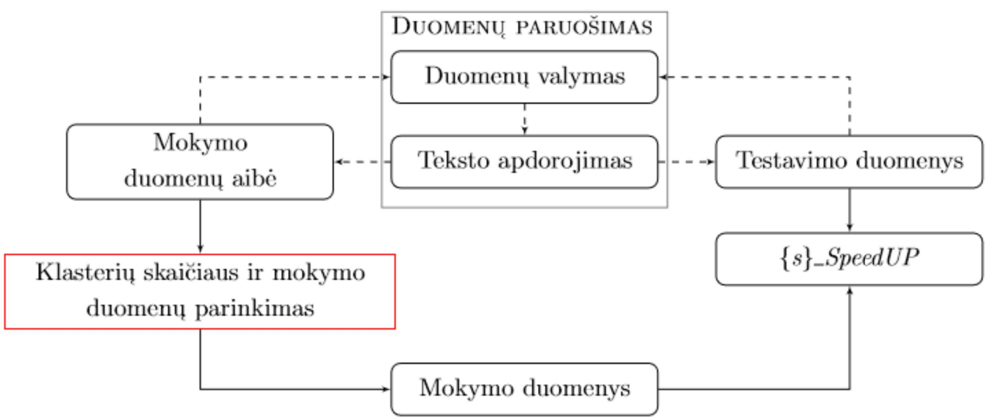
Pseudokodas *Klasterių skaičiaus parinkimas*

Require: *cluster_range* – maksimalus klasterių skaičius

$k_{opt} \leftarrow \arg\max(Performance(kmeans(k))),\ \ k \in cluster\_range$

Output : k_{opt} – optimalus klasterių skaičius.

Funkcija *Performance(function(.))* reiškia našumą, gautą vykdant tam tikrą funkciją; pagrindinis tikslas yra rasti klasterių skaičių optimalų laiko ir skaidymo į klasterius atžvilgiu.



Pseudokodas *Mokymo duomenų parinkimas*

Require: D, $Subset_{size}$, k_{opt}

$n \leftarrow 0$

$m \leftarrow Subset_{size}$

while $(len(D_{results}) <= len(D))$ **do**

EVALUATEKMEANS$(k_{opt}, random.sample(D, Subset_{size}))$

$$\textbf{k-Means}_{res} = \begin{cases} val1, \text{ reikšmė su } MAX \text{ atstumu iki klasterio centro} \\ val2, \text{ reikšmė su } MIN \text{ atstumu iki klasterio centro} \\ val3, \text{ reikšmė su } MEAN \text{ atstumu iki klasterio centro} \end{cases}$$

$\mathbf{D_{results}} \leftarrow \mathbf{D_{results}} \cup \textbf{k-Means}_{res}$

$n \leftarrow m + 1$

$m \leftarrow (n + Subset_{size}) - 1$

if $m >= len(D)$ **then**

$n \leftarrow 0$

$m \leftarrow Subset_{size}$

end if

end while

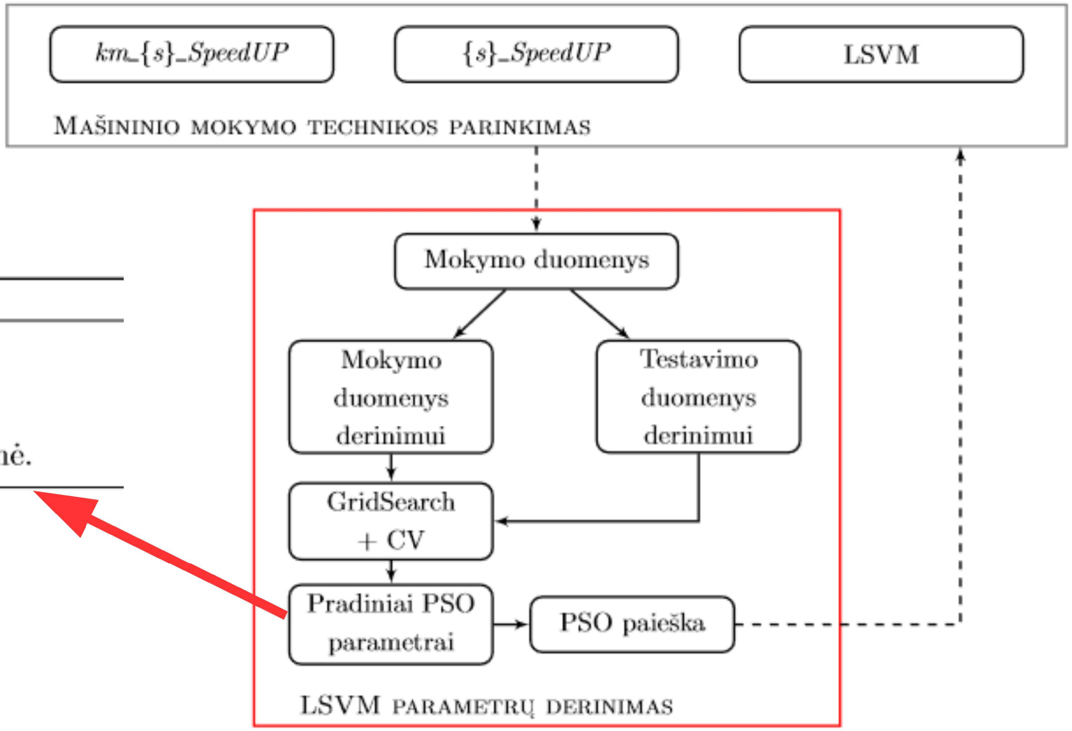
Output : $D_{results}$

Parametrai naudojami pseudokode:

D – mokymo duomenų aibė;
$\mathbf{D}_{results}$ – reprezentacinė duomenų aibė;
$\mathbf{Subset_{size}}$ – testavimo duomenų poaibio dydis į kuriuos dalinama testavimo duomenų aibė;
$\textbf{k-Means}_{results}$ – rezultatų aibė po k-vidurkių metodo vykdymo.

Siūlomas metodas

LSVM parametrų derinimas



Pseudokodas *SVM_Tuning*

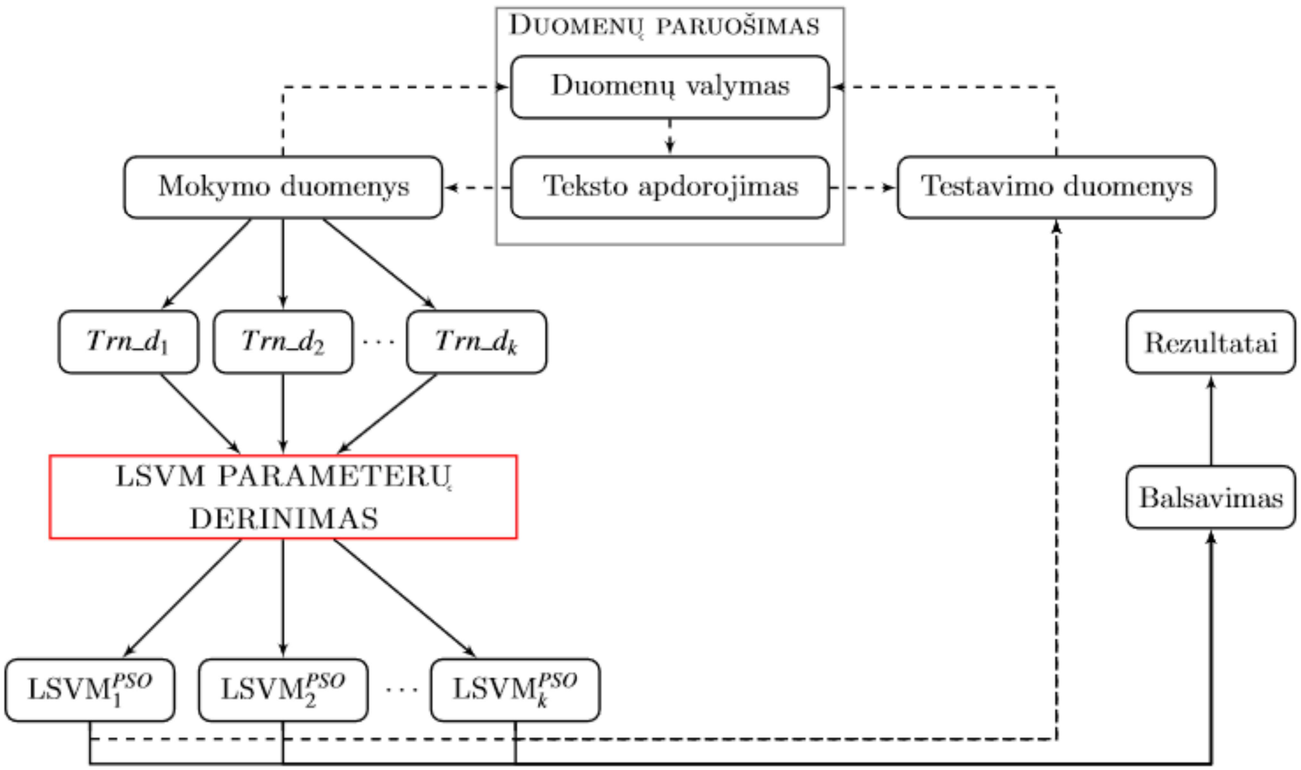
Require: mokymo duomenų aibė $-D$, C_{min}, C_{max}

$C_{opt} \leftarrow \arg\max(SVM_{ACC}(C))$, $C_{min} <= C <= C_{max}$

Output : C_{opt} – optimali baudos parametro C reikšmė.

Siūlomas metodas

LSVM ansamblis



Duomenų aibės

Duomenų aibių aprašymas, pasiūlyto metodo testavimui

Duomenų aibė	Apžvalgų sk.	Klasių sk.
sentiment140	1,600,000	2
Amazon klientų apžvalgos	4,000,000	2

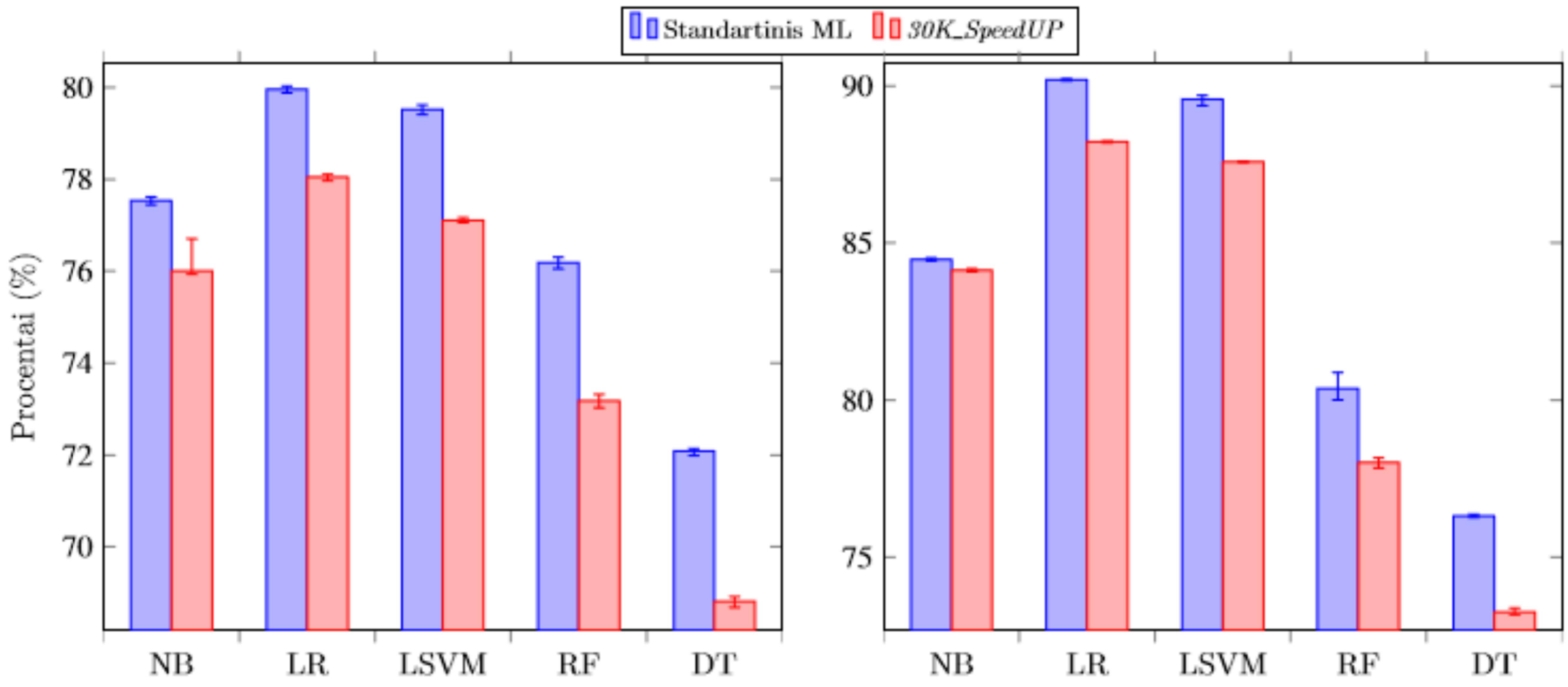
Nustatymai pasiūlytam metodui

Duomenų aibė	Testavimo duomenys (TDs)	Poaibio dydis (Ss)	Poaibių kiekis (SQ) trunc(TDs/Ss)	Liekana TDs-(Ss*SQ)	Apskaičiuotas mokymo duomenų dydis atsižvelgiant į Ss dydį
sentiment 140	480K	30K	16	0	70K
	480K	60K	8	0	140K
	480K	120K	4	0	280K
	480K	180K	2	120K	420K
Amazon klientų apžvalgos	1.2M	30K	40	0	70K
	1.2M	60K	20	0	140K
	1.2M	120K	10	0	280K
	1.2M	180K	6	120K	420K

Rezultatai

Spartinimo technikos rezultatai

Klasifikavimo tikslumas



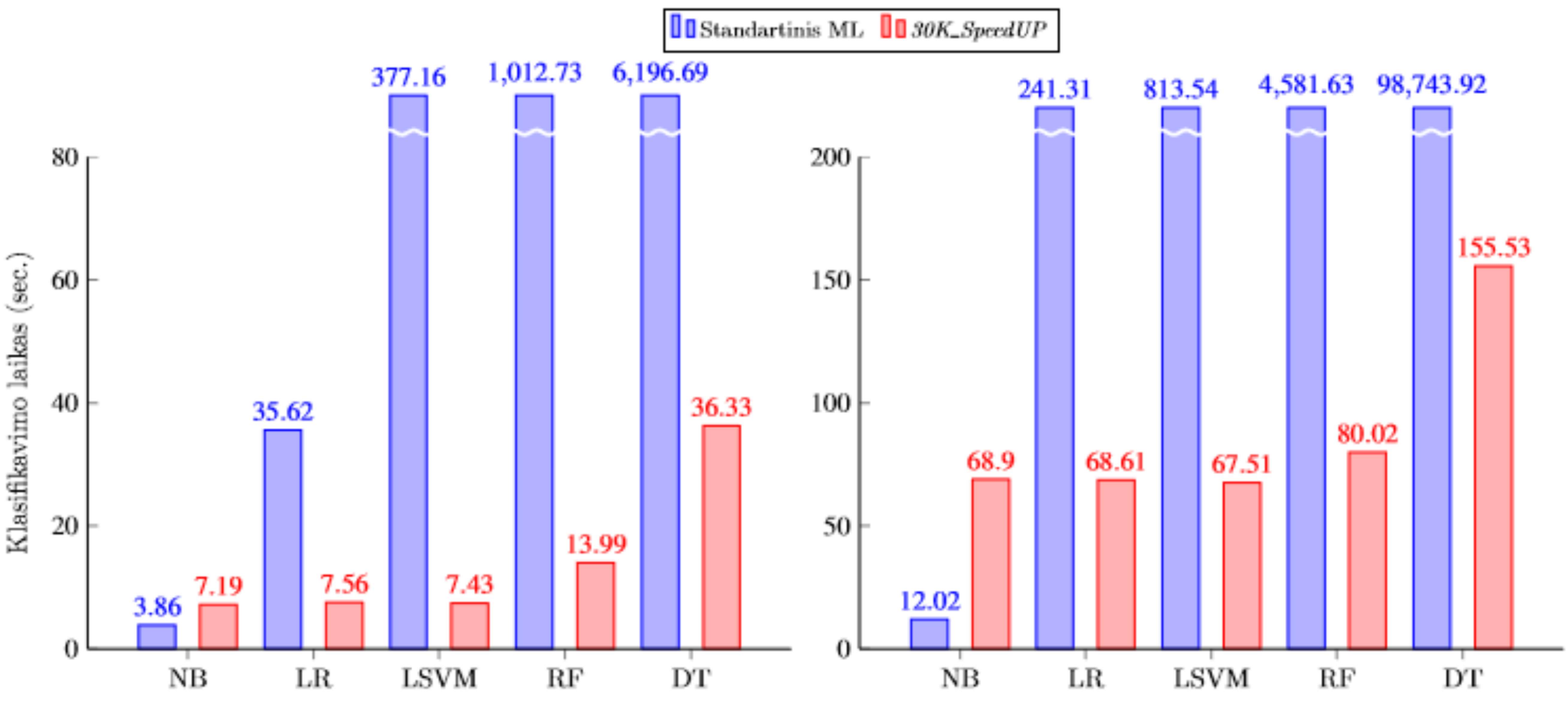
(a) sentiment140

(b) Amazon apžvalgos

Rezultatai

Spartinimo technikos rezultatai

Klasifikavimo laiko palyginimas



(a) sentiment140

(b) Amazon apžvalgos

Rezultatai

Rezultatai palyginimas

Metodas	ACC	PPV	NPV	TPR	TNR	$F_1 score$	AUC
Stanfordo “Twitter” nuomonių korpuso duomenų aibė							
LSVM *30K_SpeedUP*	77.10%	76.60%	77.62%	78.05%	76.16%	77.32%	85.36%
km_30K_SpeedUP	77.30%	76.78%	77.83%	78.26%	76.33%	77.51%	85.55%
$LSVM^{PSO}$*_km_30K_SpeedUP*	78.12%	77.55%	78.72%	79.17%	77.08%	78.35%	85.97%
*CL3_*$LSVM^{PSO}$*_km_30K_SpeedUP*	78.62%	77.98%	79.29%	79.76%	77.48%	78.86%	86.41%
*CL5_*$LSVM^{PSO}$*_km_30K_SpeedUP*	78.81%	78.16%	79.49%	79.96%	77.66%	79.05%	86.55%
LSVM	79.52%	78.83%	80.24%	80.71%	78.32%	79.76%	87.59%
LSVM su $LSVM^{PSO}$	79.90%	79.02%	80.82%	81.40%	78.39%	80.19%	87.82%
LR *30K_SpeedUP*	78.05%	77.54%	78.57%	78.96%	77.13%	78.25%	85.89%
km_30K_SpeedUP	78.14%	77.66%	78.64%	79.02%	77.26%	78.33%	85.98%
LR^{PSO}*_km_30K_SpeedUP*	78.12%	77.64%	78.62%	78.99%	77.24%	78.31%	85.95%
*CL3_*LR^{PSO}*_km_30K_SpeedUP*	78.66%	78.13%	79.20%	79.59%	77.73%	78.85%	86.42%
*CL5_*LR^{PSO}*_km_30K_SpeedUP*	78.81%	78.26%	79.39%	79.79%	77.83%	79.02%	86.51%
LR	79.96%	79.31%	80.63%	81.06%	78.85%	80.17%	87.84%
Amazon klientų apžvalgų duomenų aibė							
LSVM *30K_SpeedUP*	87.59%	87.50%	87.68%	87.71%	87.47%	87.60%	95.04%
km_30K_SpeedUP	87.74%	87.76%	87.72%	87.71%	87.76%	87.73%	95.11%
$LSVM^{PSO}$*_km_30K_SpeedUP*	88.45%	88.57%	88.33%	88.29%	88.61%	88.43%	95.24%
*CL3_*$LSVM^{PSO}$*_km_30K_SpeedUP*	88.88%	89.02%	88.74%	88.70%	89.06%	88.86%	95.49%
*CL5_*$LSVM^{PSO}$*_km_30K_SpeedUP*	89.03%	89.17%	88.89%	88.85%	89.21%	89.01%	95.56%
LSVM	89.58%	91.77%	87.60%	86.95%	92.20%	89.29%	96.33%
LSVM su $LSVM^{PSO}$	90.22%	90.20%	90.24%	90.24%	90.19%	90.22%	96.43%

Duomenų aibės palyginimui

Duomenų aibių nustatymai palyginimui su kitai autoriais

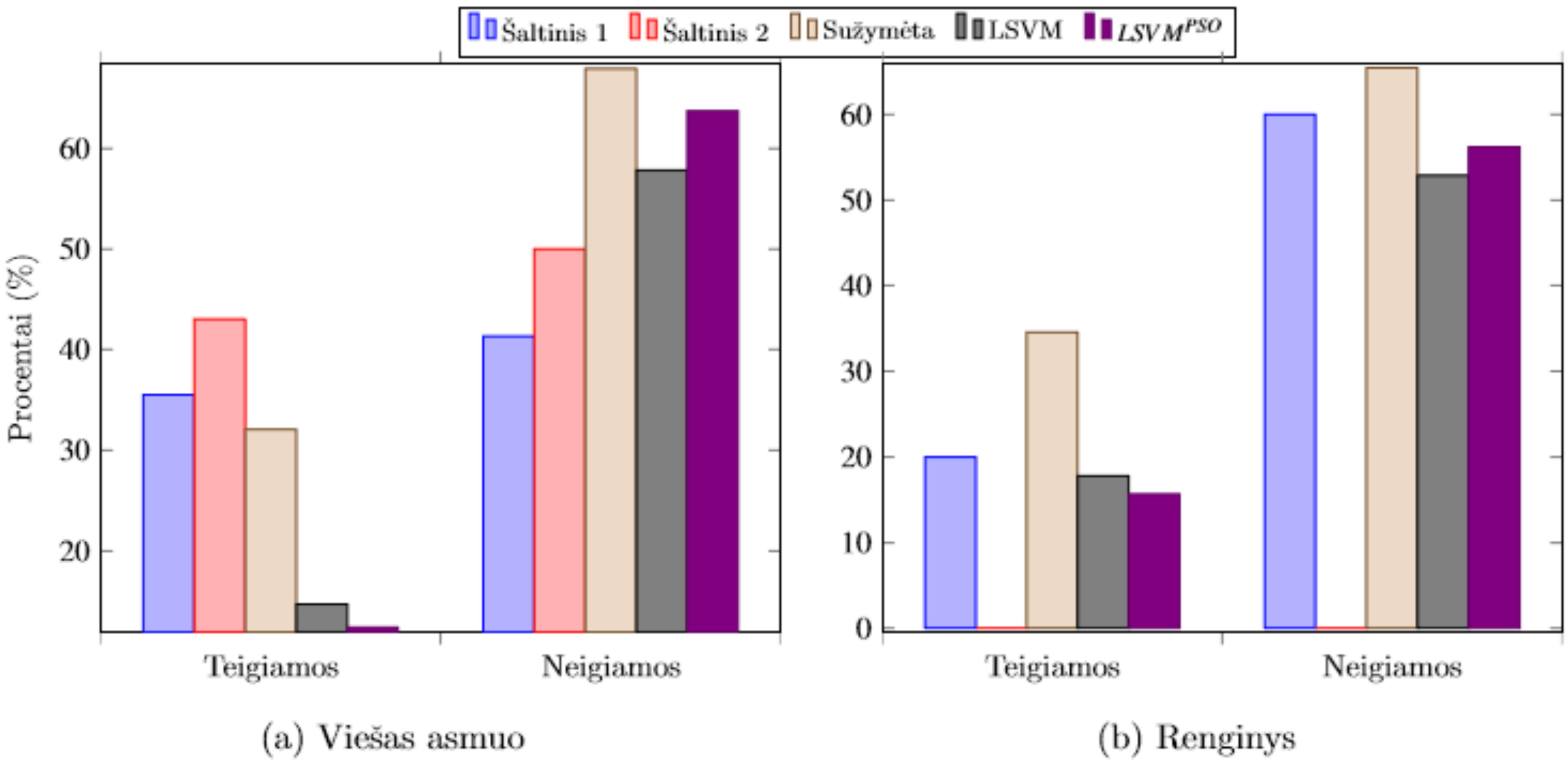
Duomenų aibė	Klasių sk.	Testavimo duomenys (TDs)	Poaibio dydis (Ss)	Poaibų sk. (SQ) trunc(TDs/Ss)	Liekana TDs-(Ss*SQ)	Mokymo duomenys
Knygos	2	20,037,414	30,000	667	27,414	70,000
Electronika	2	7,117,716	30,000	237	7,716	70,000
Kindle parduotuvė	2	2,828,300	30,000	94	8,300	70,000
Mobilaus ryšio telefonai ir jų priedai	2	3,025,090	30,000	100	25,090	70,000

Rezultatai

Palyginimo rezultatai

Autorius	Duomenų aibė	ML metodas	Tikslumas
Rain C. [9] (2013)	Knygos	Naivus Bajesas	84.50%
Shaikh and Deshpande [10] (2016)		Naivus Bajesas	80.00%
Pasiūlytas metodas		$LSVM^{PSO}$_30K_SpeedUP	**89.50%**
		$CL3_LSVM^{PSO}$_30K_SpeedUP	**89.86%**
Rain C. [9] (2013)	Kindle parduotuvė	Naivus Bajesas	87.33%
Pasiūlytas metodas		$LSVM^{PSO}$_30K_SpeedUP	**91.27%**
		$CL3_LSVM^{PSO}$_30K_SpeedUP	**91.50%**
Haque et al. [11] (2018)	Elektronika	linijinis SVM	**93.52%**
Pasiūlytas metodas		$LSVM^{PSO}$	90.14%
		$CL3_LSVM^{PSO}$_30K_SpeedUP	90.52%
		LSVM su $LSVM^{PSO}$	93.17%
Haque et al. [11] (2018)	Mobilaus ryšio telefonai ir jų priedai	linijinis SVM	**93.57%**
Wang et al. [12] (2018)		Konvoliucinis neuronų tinklas (CNN-S(+))	85.9%
		Kontekstinės faktorizacijos mašina (CFM)	83.5%
		Padėties suvokimo faktorizacijos mašina (PFM)	84.2%
Pasiūlytas metodas		$LSVM^{PSO}$_30K_SpeedUP	90.57%
		$CL3_LSVM^{PSO}$_30K_SpeedUP	90.83%
		LSVM su $LSVM^{PSO}$	93.22%

Praktinio pritaikymo rezultatai



Bendrosios išvados

1) Susijusių darbų analizė tekstinių duomenų klasifikavimo srityje parodė, kad:
   a) daugiausiai nadojami mašininio mokymo metodai yra naivus Bajesas, logistinė regresija, atraminių vektorių mašinos, atsitiktinis miškas ir sprendimų medis;
   b) atraminių vektorių mašinos yra dažniausiai naudojamas metodas nuomonių klasifikavimui, tačiau klasifikavimo laikas labai pailgėja klasifikuojant didelius duomenų masyvus. LSVM yra labai jautrus parametrų derinimui, tačiau tai vis dar yra viena iš aktualiausių problemų susijusių su praktiniais atraminių vektorių mašinų tyrimais;
   c) k-vidurkiai yra vienas iš populiariausių, labiausiai žinomų ir efektyvus metodas didelių duomenų klasterizavimui;

Bendrosios išvados

d) PSO charakterizuojamas kaip geriausias sekmės lygio ir sprendimo kokybės prasme, o taip pat tai yra daug žadantis pasirinkimas naudojimui su kitais metodais;

e) naudojant klasifikatorių ansamblius gaunami geresni klasifikavimo rezultatai palyginus su pavieniais metodais.

2) Pasiūlytas linijinėmis atraminių vektorių mašinomis grindžiamas metodas tekstinių duomenų klasifikavimui dideliuose duomenų masyvuose, turintis šias ypatybes:

a) metodas turi keturias pagrindines dalis, kurios gali būti integruojamos į standartinius mašininio mokymo metodus nepriklausomai viena nuo kitos;

b) spartinimo metodas leidžia dirbti su didelėmis tekstinių duomenų aibėmis, pagerindamas standartinio metodo klasifikavimo vykdymo laiką;

Bendrosios išvados

c) klasterių skaičiaus ir mokymo duomenų parinkimo metodas padeda sumažinti mokymo duomenų aibę bei pasirinkti daugiau skirtingų duomenų.

d) SVM Tuning metodas minimizuoja baudos parametro C paieškos intervalą dalelių spiečiaus optimizavimo euristikai.

e) ansambliai iš vienodų klasifikatorių yra žymiai efektyvesni, nes yra naudojamos skirtingos mokymo duomenų aibės kiekvienam klasifikatoriui. Tokiu būdu padidinama mokymo duomenų įvairovė tam pačiam klasifikatoriui ir taip padidinamas jo klasifikavimo tikslumas.

3) Eksperimentų rezultatai parodė, kad:

a) spartinimo metodas integruotas į mašininio mokymo metodą gerokai sumažina klasifikavimo laiką visų eksperimentuose naudotų standartinių klasifikatorių išskyrus naivų Bajesą, kuris ir taip yra pakankamai greitas.

Bendrosios išvados

b) klasterių skaičiaus ir mokymo duomenų parinkimo metodas gali padidinti klasifikavimo tikslumą iki 0.2% - 0.25% integruojant jį į LSVM, o logistinės regresijos atveju -tik 0.09%.

c) pilnas pasiūlytas metodas geriausiai tinka integracijai į LSVM, kurio klasifikavimo tikslumą gali padidinti 0.86% - 1.02%; jis taip pat gali būti integruotas ir į LR, tačiau rezultatai nėra labai reikšmingi – klasifikavimo tikslumas padidėjo tik 0.07%.

d) LR ir LSVM ansambliai taip pat gali padidinti klasifikavimo tikslumą. LSVM ansambliai iš 3-jų padidina 30K SpeedUP metodo tikslumą 1.29% - 1.52%, o LR atveju – 0.61%. LSVM ansambliai iš 5-ių tikslumą padidina 1.44% - 1.71%, o LR atveju 0.76% lyginant su 30K SpeedUP metodu.

e) Pasiūlytas metodas gali būti lengvai integruojamas į bet kokį LSVM teksto klasifikavimui dideliuose duomenų masyvuose ir tai atlikti greičiau nei standartinis LSVM.

Bendrosios išvados

- naudojant tik 70K mokymo duomenų pasiūlytas metodas gali būti konkuruojantis su moderniais metodais tokiais kaip konvoliucinis neuronų tinklas, kontekstinės faktorizacijos mašina ir padėties suvokimo faktorizacijos mašina.
- tekstinių duomenų klasifikavimui pakanka 16GB vidinės atminties, kadangi nereikia tiek daug duomenų apmokymui. Pvz. naudojant standartinį dalinimą į 70% mokymui ir 30% testavimui bei turint 20 mln. testavimo duomenų apmokymui reikia 50 mln. duomenų, tuo tarpu naudojant pasiūlytą metodą – 70000.

f) LSVM^{PSO} integruotas į standartinį LSVM taip pat pagerino pastarojo klasifikavimo tikslumą 0.48% - 0.64%.

Bendrosios išvados

4) Sukurtas viešosios nuomonės tyrimų sistemos karkasas LSVM pagrindu su integruotu $LSVM^{PSO}$:

   a) pateikta diagrama praktiniam pasiūlyto metodo pritaikymui.

   b) eksperimentai atlikti su realiais duomenimis ir rezultatai palyginti su duomenimis pateiktais dviejų institucijų, kurios užsiima visuomenės nuomonės ir rinkos tyrimais.

Ačiū už dėmesį