INFORMACINĖS TECHNOLOGIJOS IR KALBA

Tekstinių dokumentų panašumų paieška naudojant saviorganizuojančius neuroninius tinklus ir *k* vidurkių metodą

Pavel Stefanovič

Vilniaus universiteto Matematikos ir informatikos instituto doktorantas
Vilnius University, Institute of Mathematics and Informatics, Doctoral student
Akademijos g. 4, LT-08663 Vilnius
El. paštas: pavel.stefanovic@mii.vu.lt

Olga Kurasova

Vilniaus universiteto Matematikos ir informatikos instituto vyresn. mokslo darbuotoja
Vilnius University, Institute of Mathematics and Informatics, Senior researcher
Akademijos g. 4, LT-08663 Vilnius
El. paštas: olga.kurasova@mii.vu.lt

Straipsnyje nagrinėjama dokumentų panašumų paieška naudojant du populiarius metodus: saviorganizuojančius neuroninius tinklus (SOM) ir k vidurkių metodą. Vienas iš šių metodų tikslų – suskirstyti duomenis į klasterius pagal jų panašumą. Analizuota tekstinių dokumentų matricos sudarymo faktorių įtaka gautiems rezultatams. SOM kokybei įvertinti pasiūlyti du nauji matai, skirti klasifikuotiems duomenims, kurių reikšmės parodo susidariusių klasterių išsidėstymą SOM žemėlapyje. Pirmasis matas parodo, kaip gerai tos pačios klasės duomenys išsidėsto žemėlapyje vienas šalia kito, antrasis matas – kaip toli yra skirtingų klasių centrai. K vidurkių metodu gautų rezultatų kokybei įvertinti skaičiuota suma nuo klasterio centro iki klasterio narių bei įvertintas klasių nesutapimas su klasteriais. Eksperimentiniams tyrimams atlikti pasirinkti tekstiniai dokumentai, paimti iš Lietuvos Respublikos Seimo dokumentų bazės.

Įvadas

Kelis dešimtmečius saviorganizuojantis neuroninis tinklas (SOM) buvo naudojamas įvairiems skaitiniams duomenims klasterizuoti, klasifikuoti ir vizualizuoti. Tačiau gausėjant įvairialypei informacijai, saviorganizuojantys neuroniniai tinklai pradėti taikyti tekstinei (Kohonen, Xing, 2011), grafinei (Sjoberg, Laaksonen, 2011), vaizdo (Maia, Barreto, Coelho, 2011), garso (Mayer, 2011) ir kitokio pobūdžio informacijai analizuoti.

Šis straipsnis skiriamas tekstinei informacijai nagrinėti. Internetinėje erdvėje gausu įvairaus pobūdžio tekstinės informacijos: dokumentų, tekstų svetainėse, mokslinių straipsnių, todėl natūralu, jog atsiranda poreikis šią informaciją apdoroti siekiant gauti naudingų žinių. Gausybė informacijos internete mums leidžia rasti vieną ar kitą dalyką labai greitai, tačiau ji dažnai būna nenaudinga, iškraipyta ar neesminė. Todėl vis dažniau atsiranda įvairiausių duomenų tyrybos metodų jai analizuoti ir sisteminti (Marinai, 2011; Alsmadi, Saleh, 2012; Sihag, Kumar, 2013).

Saviorganizuojantys neuroniniai tinklai yra sėkmingai taikomi įvairiems uždaviniams spręsti: dokumentų plagijavimui tikrinti (Chow, Rahman, 2009), internetinėse svetainėse pateiktiems straipsniams analizuoti (Mayer, Rauber, 2011) ir kt. Kitas būdas duomenims klaste-

rizuoti ir klasifikuoti yra *k* vidurkių metodas (MacQueen, 1967). Šis metodas taip pat sėkmingai naudojamas tekstinių duomenų analizei (Steinbach, Karypis, Kumar, 2000).

Darbe (Sihag, Kumar, 2013) analizuojamas *k* vidurkių metodas tekstiniams dokumentams klasterizuoti. Pasiūlytas grafais grįstas būdas pradiniams klasterių centrams skaičiuoti. Palyginti rezultatai, kai pradiniai centrai inicijuojami kaip įprasta ir pasiūlytuoju būdu. Rezultatai parodė naujojo būdo pranašumą. Hierarchinio klasterizavimo ir *k* vidurkių metodų, taikomų tekstiniams dokumentams klasterizuoti, lyginamoji analizė atlikta darbe (Steinbach, Karypis, Kumar, 2000). Rezultatai parodė, kad *k* vidurkių metodas yra pranašesnis už hierarchinį klasterizavimą. Saviorganizuojančiu neuroniniu tinklu grįstas metodas, taikytas tekstams iš senų šaltinių atkurti, tirtas darbe (Marinai, 2011). Atlikta tekstinių dokumentų panašumo analizė, siekiant nustatyti mokslinių straipsnių plagiatus (Alsmadi, Saleh, 2012). Dokumentų panašumui nustatyti naudoti kosinusų, Euklido, Dice ir miesto kvartalų atstumų matai.

Norint duomenis analizuoti saviorganizuojančiu neuroniniu tinklu, *k* vidurkių ar kitu klasterizavimo metodu, jie turi būti pateikiami skaitine išraiška, todėl iškyla problema, kaip tinkamai paversti tekstinę informaciją į skaitinę. Tinkamai parinkus faktorius, keičiančius duomenis iš įvairių tekstinių duomenų formatų į skaitinį, galima tiksliau vertinti ir analizuoti gautus rezultatus. Visuose minėtuose straipsniuose šių faktorių analizė neatlikta. Šiame straipsnyje tiriama minėtų faktorių įtaka rezultatams, kurie gauti saviorganizuojančiu neuroniniu tinklu ir *k* vidurkių metodu, siekiant nustatyti tekstinių dokumentų – kelių Lietuvos ministerijų įsakymų panašumus.

1. Saviorganizuojantys neuroniniai tinklai

Pagrindinis saviorganizuojančio neuroninio tinklo tikslas – išlaikyti duomenų topologiją (Dzemyda, Kurasova, Žilinskas, 2008). SOM tinklai gali būti naudojami duomenims klasterizuoti ir vizualizuoti, taip pat ieškant daugiamačių duomenų projekcijų į mažesnio skaičiaus matmenų erdvę. Saviorganizuojantis neuroninis tinklas yra neuronų, išdėstytų dvimačio tinklelio (žemėlapio, lentelės) mazguose, masyvas $M = \{M_{kl}, k = 1,...,r, l = 1,...,s\}$ (Kohonen, 2001); čia r – tinklo eilučių skaičius, s – tinklo stulpelių skaičius, M_{kl} – vektorius, kurio matmenų skaičius n toks pat kaip mokymo duomenų vektorių. SOM tinklas mokomas mokymo be mokytojo būdu (angl. *unsupervised*). Neuroniniam tinklui daug kartų pateikiama skirtingų objektų, nusakomų n-mačiais vektoriais $X_1, X_2, \ldots, X_N$. Kiekviename mokymo žingsnyje (iteracijoje) vienas mokymo aibės vektorius $X_p \in \{X_1, X_2, ..., X_N\}$ pateikiamas į tinklą. Mokymo pradžioje neuronų (vektorių) M_{kl} komponenčių pradinės reikšmės dažniausiai nustatomos atsitiktinai (Stefanovič, Kurasova, 2009). Vektorius X_p palyginamas su visais neuronais M_{kl}. Dažniausiai skaičiuojamas Euklido atstumas $\| X_p - M_{kl} \|$ tarp šio vektoriaus X_p ir kiekvieno neurono M_{kl}. Randama, iki kurio neurono $M_{k_w l_w} \in \{M_{kl}, k = 1,...,r, l = 1,...,s\}$ atstumas yra mažiausias; rastas neuronas $M_{k_w l_w}$ vadinamas neuronu (vektoriumi) nugalėtoju (angl. *winner*). Visų neuronų komponentės keičiamos naudojantis iteracine formule (1).

$$M_{kl}(t+1) = M_{kl}(t) + \\ + h(k,l,k_w,l_w,t)\left(X_p - M_{kl}(t)\right). \quad (1)$$

Čia t yra iteracijos numeris, $h(k, l, k_w, l_w, t)$ – kaimynystės funkcija. Remiantis ankstesniais tyrimais, eksperimentiniuose tyrimuose naudota Gausso kaimynystės funkcija (Stefanovič, Kurasova, 2011):

$$h(k,l,k_w,l_w,t) = \\ = \alpha(t)\exp\left(-\left\| R_{k_w l_w} - R_{kl} \right\|^2 / 2(\eta(k,l,k_w,l_w,t))^2\right).$$

Čia $\alpha(t)$ – mokymo greitis, $\eta(k, l, k_w, l_w, t)$ yra neuronų M_{kl} ir $M_{k_w l_w}$ kaimynystės eilė, dvimačiai vektoriai R_{kl} and $R_{k_w l_w}$ sudaryti iš neuronų M_{kl} ir $M_{k_w l_w}$ indeksų, parodančių neuronų vietą SOM tinkle. Eksperimentiniuose tyrimuose naudotas mokymo greitis $\alpha(t) = (1 - t/T)$; čia T – iteracijų skaičius.

SOM kokybei nustatyti siūlomi du nauji matai, kurie naudojami tik klasifikuotiems duomenims analizuoti. Pirmasis matas (2) parodo, kaip arti žemėlapyje tos pačios klasės neuronai išsidėsto šalia vienas kito. Kuo mato reikšmė yra mažesnė, tuo rezultatas geresnis.

$$E_c = \frac{1}{N_c} \sum_{i=1}^{n_c-1} \sum_{j=1}^{n_c} \left(\| Z_i^c - Z_j^c \| k_i^c k_j^c + b \right). \quad (2)$$

Čia c – klasės numeris, N_c – c-osios klasės vektorių skaičius, n_c – žemėlapio neuronų, į kuriuos pateko c-osios klasės vektoriai, skaičius, Z_i^c – žemėlapio langelių indeksai, į kuriuos pateko c-osios klasės neuronai, k_i^c – c-osios klasės vektorių, patekusių į Z_i^c langelį, skaičius. Gali būti atvejų, kai į tuos pačius žemėlapio langelius patenka skirtingų klasių nariai, tuomet pagal (3) formulę skaičiuojama bauda b. Jei vienname langelyje yra tik tos pačios klasės vektoriai, tai $b = 0$.

$$b = \frac{l_i^{c'}}{k_i} + \frac{l_j^{c'}}{k_j}. \quad (3)$$

Čia k_i (k_j) – $Z_{i'}^c$ (Z_j^c) langelyje esančių vektorių skaičius, l_i^c (l_j^c) – kitų nei c-osios klasių vektorių, esančių Z_i^c (Z_j^c) langelyje, skaičius.

Antrasis pasiūlytas matas (5) įvertina, kaip toli yra išsidėstę skirtingų klasių centrai. Šiuo atveju kuo mato reikšmė yra didesnė, tuo rezultatas geresnis. Iš pradžių randami visų žemėlapyje pavaizduotų klasių centrų indeksai Y^c (4).

$$Y^c = \frac{1}{n_c} \sum_{i=1}^{n_c} Z_i^c. \quad (4)$$

Tuomet antrojo mato reikšmė skaičiuojama pagal formulę:

$$E_{center} = \frac{1}{n} \sum_{c=1}^{n-1} \sum_{d=c+1}^{n} \left\| Y^c - Y^d \right\|. \quad (5)$$

Čia n – klasių skaičius.

2. *K* vidurkių metodas

Naudojant k vidurkių metodą stengiamasi duomenų aibę padalinti į nesusikertančius klasterius (Molytė, 2011). Šiame klasterizavimo algoritme minimizuojama tam tikra kriterijaus funkcija. Ji turi būti tokia, kad minimizuojant klasterių panašumą, klasterių skirtumas būtų maksimizuojamas. Tai gali būti vidutinis atstumas tarp klasterių. *K* vidurkių metodo etapai yra šie: 1) atsitiktinai iniciuojami klasterių centrai; 2) kiekvienas analizuojamos duomenų aibės vektorius yra priskiriamas tam klasteriui, iki kurio atstumas nuo centro yra mažiausias; 3) perskaičiuojami kiekvieno klasterio centrai; 4) skaičiuojama kvadratinė paklaida tarp klasterio centro ir klasteriui priskirtų duomenų; 5) 2–4 etapai kartojami tol, kol analizuojami duomenys nebepasiskirsto kitiems klasteriams.

Įprastai k vidurkių metodo kokybei įvertinti skaičiuojama gautų klasterių atstumų nuo klasterio centro iki jam priskirtų duomenų suma. Šiame straipsnyje analizuoti duomenys, priskirti kelioms klasėms, todėl svarbu įvertinti, ar gauti klasteriai atitinka klases. Pradžioje duomenys klasterizuojami į tiek klasterių, kiek yra klasių. Nustatomas klasių ir klasterių atitikimas, t. y. tariama, kad tam tikros klasės duomenys turi būti priskirti klasteriui, į kurį pateko dauguma tos klasės narių. Tuomet suskaičiuojama, kiek kiekvienos klasės narių pateko ne į savo klasterį. Šio įverčio taikyti SOM rezultatui negalima, nes čia nėra griežtai išreikštų klasterių, juos tik galima stebėti vizualiai.

3. Tekstinių dokumentų matricos sudarymas

Kad galėtume analizuoti tekstinius dokumentus įvairiais klasterizavimo metodais, būtina dokumentus paversti skaitine išraiška. Tuo tikslu sudaromas tekstinių dokumentų žodynas, o iš jo – tekstinių dokumentų matrica (6).

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \dots & x_{Nn} \end{pmatrix} \quad (6)$$

Čia x_{pl} – l-ojo žodžio pasikartojimų p-ajame dokumente skaičius, $l = 1,...,n$, $p = 1,...,N$. N – analizuojamų dokumentų skaičius, o n – žodžių skaičius dokumentų žodyne. Kiekviena matricos (6) eilutė atitinka vektorių, sudarytą iš n komponenčių. Šie vektoriai (duomenys) bus klasterizuojami SOM ir k vidurkių metodu. Pradžioje būtina sudaryti tekstinių dokumentų žodyną, atsižvelgiant į pasirinktus faktorius. Maksimalaus žodžių ilgio apribojimas leidžia atmesti žodžius, kurie yra per trumpi ar per ilgi. Kiekviename dokumente galima rasti trumpų žodelių, pavyzdžiui, prielinksnių ar jungtukų: į, iš, o ir kt., todėl sudarant žodyną juos reikia atmesti. Kad nebūtų iškraipomi rezultatai, būtina atmesti skaičius, numeraciją ir formules, nes tai neteikia esminės informacijos apie analizuojamą dokumentą. Labai svarbu parinkti tinkamą žodžių pasikartojimų dokumente skaičių. Nurodant per mažą skaičių, į žodyną gali būti įtraukiami visai neesminiai žodžiai. Nurodant per didelį skaičių, dokumentai gali būti atmesti, nes gali būti taip, kad juose nebus tiek kartų pasikartojančio žodžio.

Sudarant tekstinių dokumentų žodyną, galima neįtraukti dažnai vartojamus neesminius, bet visuose dokumentuose pasitaikančius žodžius iš sudaryto sąrašo. Pavyzdžiui, tokie trumpi žodeliai (jungtukai, prielinksniai, įvardžiai ir kt.): čia, kur, kada, šie, ten ir t. t. Jie visai necharakterizuoja analizuojamo dokumento. Teksto dokumentų matricai sudaryti naudotas „Text To Matrix Generator“ (TMG) įrankis (*Zeimpekis, Gallopoulos, 2005*).

4. Eksperimentinių tyrimų rezultatai

4.1. Tyrimų duomenys

Eksperimentiniuose tyrimuose nagrinėtos aštuonios dokumentų aibės, paimtos iš Lietuvos Respublikos Seimo duomenų bazės (LRS, 2013). Atsitiktiniu būdu pasirinkta po 15 panašaus ilgio skirtingų ministerijų įsakymų: Finansų, Kultūros, Susisiekimo, Sveikatos, Švietimo, Ūkio, Vidaus reikalų ir Žemės ūkio. Iš šių ministerijų įsakymų sudaryti trys duomenų rinkiniai: $X^1 = \{ X_1^1, X_2^1, ..., X_{60}^1 \}$, $X^2 = \{ X_1^2, X_2^2, ..., X_{60}^2 \}$ ir $X^3 = \{ X_1^3, X_2^3, ..., X_{60}^3 \}$.

Pirmąjį rinkinį sudaro: vektoriai $X_1^1, X_2^1, ..., X_{15}^1$ – I klasė (Sveikatos ministerija), vektoriai $X_{16}^1, X_{17}^1, ..., X_{30}^1$ – II klasė (Švietimo ministerija), vektoriai $X_{31}^1, X_{32}^1, ..., X_{45}^1$ – III klasė (Vidaus reikalų ministerija) ir vektoriai $X_{46}^1, X_{47}^1, ..., X_{60}^1$ – IV klasė (Žemės ūkio ministerija).

Antrąjį rinkinį sudaro: vektoriai $X_1^2, X_2^2, ..., X_{15}^2$ – I klasė (Finansų ministerija), vektoriai $X_{16}^2, X_{17}^2, ..., X_{30}^2$ – II klasė (Kultūros ministerija), vektoriai $X_{31}^2, X_{32}^2, ..., X_{45}^2$ – III klasė (Susisiekimo ministerija) ir vektoriai $X_{46}^2, X_{47}^2, ..., X_{60}^2$ – IV klasė (Ūkio ministerija).

Trečiąjį rinkinį sudaro: vektoriai $X_1^3, X_2^3, ..., X_{15}^3$ – I klasė (Finansų ministerija), vektoriai $X_{16}^3, X_{17}^3, ..., X_{30}^3$ – II klasė (Ūkio ministerija), vektoriai $X_{31}^3, X_{32}^3, ..., X_{45}^3$ – III klasė (Vidaus reikalų ministerija) ir vektoriai $X_{46}^3, X_{47}^3, ..., X_{60}^3$ – IV klasė (Žemės ūkio ministerija).

Eksperimentų metu vektorių matmenų erdvė kiekvienu atveju yra skirtinga, nes ji priklauso nuo susidariusio tekstinio dokumentų žodyno ilgio.

4.2. Tekstinių dokumentų matricos kūrimas eksperimentiniams tyrimams

Kaip jau minėta, svarbu parinkti tinkamus faktorius verčiant tekstinę informaciją į skaitinę. Dokumentuose pateikti skaičiai, numeracija nesuteikia jokios esminės informacijos, todėl sudarant tekstinių dokumentų matricą šie duomenys nebuvo įtraukiami. Tyrime analizuojama žodžių pasikartojimo ir dažniausiai vartojamų žodžių sąrašo naudojimo įtaka rezultatams. Atlikus pirminius eksperimentus pastebėta, kad tiriamai dokumentų aibei tinkamiausias maksimalus žodžių pasikartojimų skaičius dokumente yra penki. Pasirinkus didesnį skaičių sudarant tekstinių dokumentų žodyną, dokumentai buvo atmetami, nes juose tiek kartų pasikartojančių žodžių nebuvo. Tirtas žodžių pasikartojimas nuo 1 iki 5, naudojant dažniausiai vartojamų žodžių sąrašą arba ne. Iš viso kiekvienam duomenų rinkiniui

1 lentelė. ***SOM rezultatai pirmajam duomenų rinkiniui***

Pasikartojimas / Matas	Mokymo aibė					Testavimo aibė				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Nenaudojant dažniausiai vartojamų žodžių sąrašo										
E_1	17,84	17,78	19,32	20,34	18,94	0,75	1,22	2,70	2,44	5,17
E_2	28,50	28,29	28,25	26,76	29,69	2,59	3,75	3,24	3,28	4,95
E_3	15,17	17,62	22,20	17,99	19,68	2,35	2,50	4,46	2,82	4,51
E_4	24,30	30,13	26,08	31,94	30,63	2,02	1,60	1,80	2,59	1,68
E_{center}	6,25	5,36	5,42	4,77	5,06	6,14	6,87	5,85	5,45	5,26
Naudojant dažniausiai vartojamų žodžių sąrašą										
E_1	20,72	18,63	16,56	14,10	17,47	1,20	0,93	0,58	1,54	3,14
E_2	28,56	31,97	29,67	32,11	28,15	3,67	3,60	4,02	5,79	3,91
E_3	15,31	16,04	15,35	23,03	23,62	2,02	2,77	2,96	4,49	6,78
E_4	26,78	21,38	27,32	25,69	29,35	2,93	2,97	3,19	1,41	1,27
E_{center}	5,92	6,42	5,98	5,11	4,95	5,93	7,29	7,05	5,78	4,83

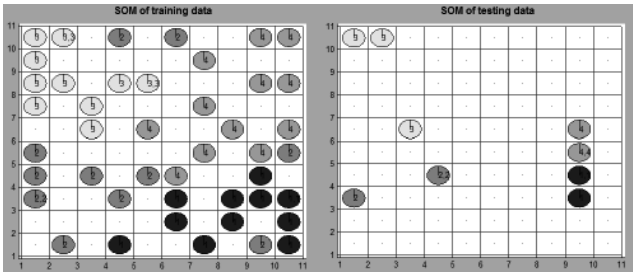
sudaryta po 10 tekstinių dokumentų matricų, kurios analizuotos saviorganizuojančių neuroninių tinklų ir *k* vidurkių metodu.

4.3. Rezultatai, gauti naudojant SOM tinklą

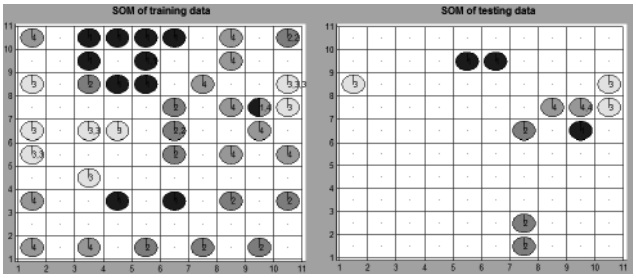
Pasirinktas 10×10 SOM žemėlapio dydis ($r = s = 10$). 80 proc. duomenų priskirti mokymo aibei, kiti 20 proc. testavimo aibei. Naudota sistema, pasiūlyta darbe (Stefanovič, Kurasova, 2011). Kiekvienas bandymas kartotas penkis kartus, esant skirtingoms žemėlapio neuronų pradinėms reikšmėms. Penkiais bandymais gautų matų reikšmių vidurkiai pateikiami 1–3 lentelėse. Kaip matome 1 lentelėje, pirmajam duomenų rinkiniui rezultatai yra gana panašūs naudojant dažniausiai vartojamų žodžių sąrašą ir jo nenaudojant. Mažiausi atstumai tarp klasių narių gauti I ir III klasėms (E_1 ir E_3), tai reiškia, kad jų nariai žemėlapyje yra išsidėstę arti vienas kito. Priešingai, II ir IV klasės atstumai (E_2 ir E_4) yra kur kas didesni ir tai rodo, jog šie duomenys pasklidę po žemėlapį plačiau. Didžiausias skirtumas tarp klasių centrų (E_{center}) gautas, kai sudarant žodyną naudojamas dažnai vartojamų žodžių sąrašas (6,42) ir žodžių pasikartojimų skaičius ne mažesnis nei 2. Testavimo aibėje atstumas siekia 7,29. Nurodant vis didesnį žodžių pasikartojimų skaičių tiek mokymo aibei, tiek testavimo aibei, atstumas tarp centrų E_{center} turi tendenciją mažėti. Tai reiškia, jog žemėlapyje pavaizduoti duomenys vis labiau susimaišo.

Pirmame paveiksle pateikiamas SOM žemėlapis pirmajam duomenų rinkiniui, iš visų bandymų išrinkus tą, kuriam matų reikšmės buvo geriausios, t. y. $E_1 - E_4$ reikšmė maža, o E_{center} – didelė. Žemėlapis, kai matų reikšmės blogiausios, vaizduojamas 2 paveiksle. SOM žemėlapis gautas (1 pav.), kai žodžių pasikartojimų skaičius yra 2 ir sudarant žodyną atmetami dažniausiai vartojami žodžiai. Paveiksle spalvos ir skaičiukai žymi klases.

Žemėlapyje matyti (1 pav.), jog I ir III (Sveikatos ir Vidaus reikalų ministerijos) klasių duomenys sudaro du atskirus klasterius, o klasės nariai išsidėstę arti vienas kito, tai ir parodė 1 lentelėje pateikti rezultatai. II ir IV (Švietimo



1 pav. ***Geriausias SOM žemėlapis pirmajam duomenų rinkiniui***



2 p a v. ***Blogiausias SOM žemėlapis pirmajam duomenų rinkiniui***

ir Žemės ūkio ministerijos) klasių duomenys plačiau pasklidę po žemėlapį, todėl jų klasių atstumų matų reikšmės (E_2 ir E_4) gautos kur kas didesnės. Dešinėje pateikti testavimo rezultatai rodo, jog duomenys išsidėsto grupelėmis maždaug tose pačiose vietose, kur ir mokymo aibės žemėlapyje pateikti duomenys. SOM žemėlapis, kuriame matų reikšmės yra mažiausios, gautas, kai pasirinktas žodžių pasikartojimų skaičius 4 ir nenaudotas dažniausiai vartojamų žodžių sąrašas (2 pav.).

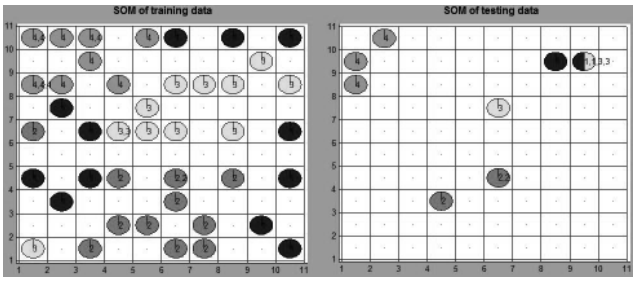
Šiuo atveju matome, jog duomenys plačiai išsibarsto po visą žemėlapį. Susiformavo tik nedideli I klasės klasteriai viršuje, III klasės kairėje, II klasės nariai apatiniame kampe ir IV klasės nariai viršutiniame dešiniajame kampe. Į vieną langelį patenka po vieną I ir IV klasių narį.

Atlikus antrojo duomenų rinkinio tyrimus matyti (2 lentelė), kad mažiausios pirmojo mato reikmės (E_2 ir E_4) gautos II ir IV klasei (Kultūros ir Ūkio ministerijos). Šių klasių duomenys žemėlapyje išsidėsto arčiau vieni kitų tiek mokymo, tiek testavimo aibei. I ir III (Finansų ir Susisiekimo ministerijos) klasių duomenų rezultatai gerokai blogesni. Kaip ir pirmojo duomenų rinkinio analizės atveju, didinant žodžių pasikartojimų skaičių, atstumai tarp centrų (E_{center}) turi tendenciją mažėti, duomenys pasiskirsto plačiau. Palyginus atstumus tarp centrų mokymo aibei matyti, kad geresni (didesnė mato reikšmė) rezultatai gaunami, kai sudarant dokumentų žodyną dažniausiai vartojamų žodžių sąrašas nenaudojamas.

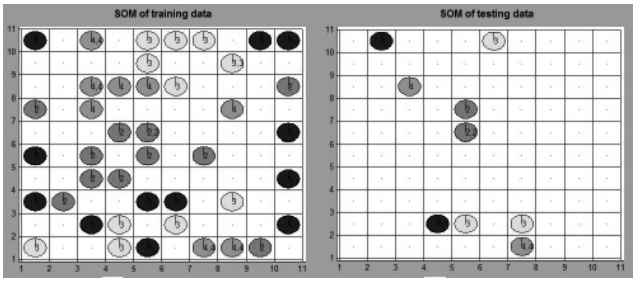
Geriausias SOM žemėlapis pagal pasiūlytus matus gautas, kai žodžių pasikartojimas lygus 1 ir sudarant žodyną naudojamas dažniausiai vartojamų žodžių sąrašas (3 pav.). Mokymo aibės kairiajame viršutiniame žemėlapio kampe išsidėsto visi IV klasės nariai (Ūkio ministerija), apačioje dauguma II klasės narių (Kultūros ministerija), per vidurį III klasės nariai (Susisiekimo ministerija), o I klasės nariai (Finansų ministerija) išsibarsto po visą SOM žemėlapį. Pagal 2 lentelėje pateiktus rezultatus, mažiausi atstumai (E_2 ir E_4) yra tarp IV ir II klasių narių ir tai matoma pateiktame žemėlapyje (3 pav.). Plačiausiai išsidėsto I klasės duomenys.

2 l e n t e l ė. ***SOM rezultatai antrajam duomenų rinkiniui***

Pasikartojimas	**Mokymo aibė**					**Testavimo aibė**				
Matas	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Nenaudojant dažniausiai vartojamų žodžių sąrašo										
E_1	28,02	34,40	32,36	28,82	33,38	3,45	4,33	3,44	4,17	4,26
E_2	19,86	20,88	19,91	22,88	22,52	1,56	2,06	1,56	0,67	0,83
E_3	28,35	25,84	25,14	26,40	26,62	5,06	4,79	5,10	4,48	5,02
E_4	14,95	15,12	16,98	22,24	21,26	1,49	2,30	2,88	4,12	3,59
E_{center}	5,98	4,92	4,99	4,61	3,81	7,01	7,11	6,29	6,06	4,96
Naudojant dažniausiai vartojamų žodžių sąrašą										
E_1	31,97	29,26	30,64	29,52	29,93	2,97	4,06	4,96	5,27	4,44
E_2	19,31	22,12	18,70	20,97	23,26	1,51	1,70	1,92	0,78	1,00
E_3	22,51	24,66	25,71	24,51	29,07	4,84	4,73	4,11	3,24	5,70
E_4	14,64	20,13	18,89	18,94	20,78	1,68	3,31	3,33	2,16	3,60
E_{center}	6,36	5,49	5,18	5,99	4,60	7,39	6,77	4,86	6,70	4,99



3 p a v. ***Geriausias SOM žemėlapis antrajam duomenų rinkiniui***



4 p a v. ***Blogiausias SOM žemėlapis antrajam duomenų rinkiniui***

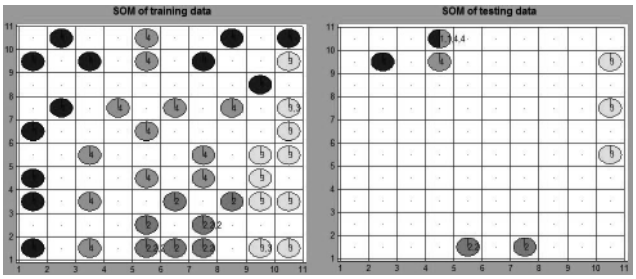
Blogiausias SOM rezultatas pagal tiriamus matus gautas, kai pasikartojimų skaičius 5 ir nenaudojamas dažniausiai vartojamų žodžių sąrašas. Matome (4 pav.) tik nedidelius skirtingų klasių klasterius, tačiau iš esmės visų klasių duomenys plačiai išsidėsto įvairiose žemėlapio vietose.

Atlikus eksperimentus naudojant trečiąjį duomenų rinkinį matyti, kad mažiausios matų reikšmės (E_2 ir E_3) gautos II ir III klasėms (Ūkio ir Vidaus reikalų ministerijos). I ir IV klasėms (Finansų ir Žemės ūkio ministerijos) matų reikšmės (E_1 ir E_4) yra didžiausios. Didinant pasikartojimo skaičių mokymo ir testavimo aibėms, analogiškai kaip antrajam duomenų rinkiniui, atstumai tarp centrų E_{center} turi tendenciją mažėti. Šiuo atveju, kai pasikartojimų skaičius nuo 1 iki 3, didesni atstumai tarp centrų E_{center} gaunami nenaudojant dažniausiai vartojamų žodžių sąrašo, o kai pasikartojimų skaičius 4 ir 5, naudojant sąrašą rezultatai yra geresni.

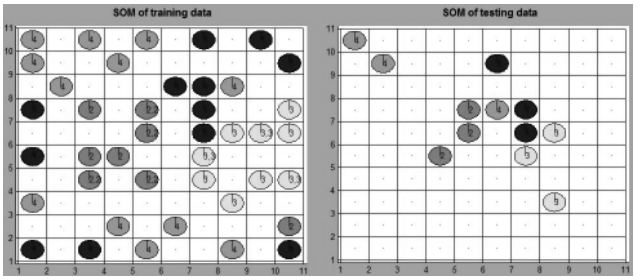
Šiuo atveju geriausias pagal pasiūlytų matų reikšmes SOM žemėlapis gautas, kai žodžių pasikartojimų skaičius 1 ir dažniausiai vartojamų žodžių sąrašas nenaudotas. Aiškiai matomi mokymo aibės žemėlapio apačioje išsidėstę II klasės nariai (Ūkio ministerija), dešinėje – III klasė (Vidaus reikalų ministerija), per vidurį išilgai nuo apačios iki viršaus – IV klasės nariai (Žemės ūkio ministerija) ir kairėje pusėje – dauguma I klasės narių (Finansų ministerija).

3 l e n t e l ė. ***SOM rezultatai trečiajam duomenų rinkiniui***

Pasikartojimas	**Mokymo aibė**					**Testavimo aibė**				
Matas	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Nenaudojant dažniausiai vartojamų žodžių sąrašo										
E_1	29,98	31,65	30,11	30,22	32,77	1,93	5,13	4,26	3,96	4,36
E_2	11,49	18,47	17,41	21,78	21,45	1,65	3,08	2,96	4,03	4,24
E_3	17,18	22,67	19,75	19,63	23,60	2,57	3,06	3,25	2,66	5,84
E_4	26,00	26,29	29,64	30,11	27,87	1,54	4,85	2,69	3,22	1,28
E_{center}	6,68	5,32	5,14	4,62	4,33	7,86	6,06	4,89	3,84	4,93
Naudojant dažniausiai vartojamų žodžių sąrašą										
E_1	30,63	33,42	33,44	32,65	30,43	4,80	3,94	2,75	4,32	3,99
E_2	12,52	14,46	15,82	16,86	18,46	1,95	2,30	1,84	2,94	3,85
E_3	14,48	17,47	21,25	21,76	26,09	2,32	1,98	3,22	4,23	6,35
E_4	27,92	30,09	30,52	29,43	30,21	3,08	1,26	3,27	2,85	1,78
E_{center}	6,15	5,15	3,89	4,83	4,41	7,33	6,38	5,03	5,00	3,98



5 p a v. ***Geriausias SOM žemėlapis trečiajam duomenų rinkiniui***



6 p a v. ***Blogiausias SOM žemėlapis trečiajam duomenų rinkiniui***

Blogiausi rezultatai pagal pasiūlytus matus gauti, kai žodžių pasikartojimų skaičių 3 ir naudojamas dažniausiai vartojamų žodžių sąrašas. Kadangi II klasės nariai atitinka Ūkio ministerijos dokumentus, o IV klasės – Žemės ūkio ministerijos, galima pamatyti, kad šios klasės tiek 5, tiek 6 paveiksluose yra išsidėsčiusios greta, tai rodo šių dokumentų panašumą. Lygiai taip pat I klasės (Finansų ministerija) ir III klasės (Vidaus reikalų ministerija) dokumentuose nagrinėjami panašūs klausimai, todėl natūralu, jog I ir III duomenys yra greta.

4.4. Rezultatai, gauti naudojant k vidurkių metodą

Tyrimams *k* vidurkių metodu atlikti naudota *Matlab* sistemos funkcija *k-means*. Imant vis kitas pradines klasterių centrų koordinates, bandymai buvo kartoti po 10 kartų. Kaskart buvo vertinamas klasių priskyrimas klasteriams (4 lentelėje tai nurodoma „Neteisingai priskirti“) bei atstumų tarp klasterių centrų iki jiems priskirtų duomenų sumos (4 lentelėje – AKS). Dešimties bandymų metu gautų įverčių vidurkiai pateikiami 4 lentelėje.

Iš pateiktų rezultatų (4 lentelė) matyti, kad nagrinėjant visus tris duomenų rinkinius, didinant žodžių pasikartojimų skaičių tekste ir sudarant žodyną nenaudojant dažniausiai vartojamų žodžių sąrašo, pastebimas neteisingai priskirtų duomenų pagal klases skaičiaus didėjimas, tačiau klasterizavimo rezultatai pagal AKS kaskart gerėja, neskaitant vieno atvejo antrajam duomenų rinkiniui, kai pasikartojimų skaičius lygus 4. Todėl naudojant *k* vidurkių metodą sunku pasakyti, kuris rezultatas yra geriausias, nes reikia atsižvelgti tiek į klasterizavimo kokybę, tiek į klasių suskirstymą klasteriams. Atlikus eksperimentus, kai sudarant žodyną naudojamas dažniausiai vartojamų žodžių sąrašas, matoma, kad klasterizavimo rezultatai gerėja didinant pasikartojimų skaičių, tačiau klasių priskyrimo klasteriams rezultatai yra įvairūs.

4 l e n t e l ė. ***K vidurkių metodu gauti eksperimentiniai rezultatai***

Pasikartojimas	**Nenaudojant dažniausiai vartojamų žodžių sąrašo**					**Naudojant dažniausiai vartojamų žodžių sąrašą**				
Matas	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Pirmajam duomenų rinkiniui										
Neteisingai priskirti	18,3	23,6	26,6	30,7	33,2	22,8	20,1	21,5	27,3	32,1
AKS	43608	40513	35885	31221	26923	40112	37259	32887	29115	24972
Antrajam duomenų rinkiniui										
Neteisingai priskirti	21,9	23,3	24,7	28,8	28	25	25,2	23,9	26,5	28,9
AKS	36796	34209	30580	26469	22719	35732	32302	28056	24934	21585
Trečiajam duomenų rinkiniui										
Neteisingai priskirti	19,9	21,9	27,6	32,5	35,1	21,7	27,7	26,8	32,7	34,2
AKS	39383	37593	33090	28965	24125	37124	34765	32004	26021	21524

Lyginant rezultatus, kai sudarant žodyną naudojamas dažniausiai vartojamų žodžių sąrašas ir ne, matoma, kad analizuojant visus tris duomenų rinkinius gaunama mažesnė mato AKS reikšmė, kai vartojami dažniausi žodžiai. Gaunami geresni klasterizavimo rezultatai. Nagrinėjant klasių priskyrimus klasteriams matyti, jog pirmajam duomenų rinkiniui, nurodant pasikartojimų skaičių nuo 2 iki 5, klasių priskyrimas tinkamiems klasteriams yra geresnis, kai naudojamas dažniausiai vartojamų žodžių sąrašas. Antrojo ir trečiojo duomenų rinkinių atvejais klasių priskyrimo klasteriams rezultatai yra gana panašūs.

Išvados

Straipsnyje nagrinėjama tekstinių dokumentų panašumo paieška naudojant du populiarius metodus – saviorganizuojantį neuroninį tinklą ir *k* vidurkių metodą. Tirta kelių tekstinių dokumentų matricos sukūrimo faktorių įtaka gautiems rezultatams. SOM kokybei įvertinti pasiūlyti du nauji matai. *K* vidurkio metodo kokybei įvertinti skaičiuotos atstumų nuo klasterių centrų iki klasterių narių sumos ir nustatomas klasių ir klasterių atitikimas. Klasių ir klasterių atitikimą SOM tinkle vienareikšmiškai įvertinti neįmanoma, nes čia negalima duomenų vienareikšmiškai priskirti klasteriams, kaip įmanoma *k* vidurkių metodu, klasterius galima tik stebėti vizualiai. Tekstinių duomenų panašumo analizė taikant saviorganizuojantį neuroninį tinklą yra informatyvesnė nei taikant *k* vidurkių metodą, nes naudojant SOM gaunami ne tik skaitiniai įverčiai, bet ir žemėlapiai, kuriuose matyti tekstinių dokumentų išsidėstymas, o tai leidžia vizualiai vertinti dokumentų panašumus.

Atlikti tyrimai parodė, jog geresni rezultatai pagal pasiūlytus matus gaunami, kai sudarant žodyną naudojamas dažniausiai vartojamų žodžių sąrašas. Tiek SOM, tiek *k* vidurkių metodas pateikė geresnius rezultatus. Teksto žodžių pasikartojimų skaičius dokumente taip pat yra svarbus. Tyrimai parodė, jog per didelis ar per mažas žodžių pasikartojimų skaičius rezultatus blogina. Nagrinėjant tekstinius duomenis, apie kuriuos turima mažai informacijos, sudarant tekstinę dokumentų matricą rekomenduotume naudoti dažnai vartojamų žodžių sąrašą ir žodžių pasikartojimų skaičių parinkti ne mažesnį nei 3. Naudojant saviorganizuojančius neuroninius tinklus, pirmojo mato reikšmės ($E_1 - E_4$) labiau nesiskiria priklausomai nuo žodžių pasikartojimų skaičiaus, tačiau didinant šį skaičių atstumas tarp klasių centrų E_{center} mažėja, o tai reiškia, kad rezultatai gaunami blogesni. Kai žodžių pasikartojimų skaičius didėja, klasterizavimo *k* vidurkių metodu rezultatai AKS mato prasme gaunami geresni, t. y. mažesnės sumos tarp klasterių centrų ir jų narių. Be to, naudojant dažniausiai vartojamų žodžių sąrašą gaunami geresni rezultatai. Tačiau, didėjant žodžių pasikartojimų skaičiui, klasių priskyrimų tinkamiems klasteriams rezultatai blogėja.

Atlikus eksperimentus naudojant trečiąjį dokumentų rinkinį matyti, kad nors dokumentai yra skirtingų klasių, tačiau pagal sudarytą tekstinių dokumentų matricą Finansų ir Vidaus reikalų ministerijų tekstiniai dokumentai yra panašūs. Taip pat panašumų matoma ir tarp Ūkio ir Žemės ūkio ministerijų dokumentų. Naudojant pirmąjį ir antrąjį duomenų rinkinius, aiškių panašumų tarp skirtingų klasių negauta.

Padėka. Šis tyrimas atliktas Europos socialinio fondo finansuojamo projekto „Paslaugų interneto technologijų kūrimo ir panaudojimo našių skaičiavimų platformose teoriniai ir inžineriniai aspektai“ (Nr. VP1-3.1-ŠMM-08-K-01-010) lėšomis. Straipsnio autoriai dėkoja recenzentams už pastabas ir pasiūlymus, kurie leido iš esmės pagerinti straipsnį.

LITERATŪRA IR ŠALTINIAI

ALSMADI, Izzat; SALEH, I. Zakaria (2012). Documents Similarities Algorithms for Research Papers Authenticity. *ICCIT*, p. 210–214.

CHOW, W. S. Tommy; RAHMAN, M. K. M. (2009). Multilayer SOM With Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection. *IEEE Transactions on Neural networks*, vol. 20, no. 9.

DZEMYDA, Gintautas; KURASOVA, Olga; ŽILINSKAS, Julius (2008). *Daugiamačių duomenų vizualizavimo metodai*, p. 108–127.

KOHONEN, Teuvo (2001). *Self-organizing Maps*. 3rd ed. Springer series in information sciences. Berlin: Springer-Verlag. 506 p. ISBN 3540679219.

KOHONEN, Teuvo; XING, Hongbing (2011). Contextually Self-Organized Maps of Chinese Words. In: J. Laaksonen, T. Honkela, eds. Advances in Self-Organizing Maps – WSOM 2011, *Lecture Notes in Computer Science*, vol. 6731, Heidelberg: Springer Verlag, p. 16–29.

LIETUVOS RESPUBLIKOS SEIMAS (LRS) [interaktyvus] [žiūrėta 2013 m. gegužės 2 d.]. Prieiga per internetą: <http://www3.lrs.lt/dokpaieska/forma_l.htm>.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In: Le Cam, L. M. and Neyman, J., editors. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Statistics.* Berkeley and Los Angeles: University of California Press, p. 281–297.

MAIA, E. B. Jose; BARRETO, A. Guilherme; COELHO, V. Andre Luis (2011). Evolving a Self-Organizing Feature Map for Visual Object Tracking. In: J. Laaksonen, T. Honkela, eds. Advances in Self-Organizing Maps – WSOM 2011, *Lecture Notes in Computer Science*, vol. 6731, Heidelberg: Springer Verlag, p. 121–130.

MARINAI, Simone (2011). Text from early printed books. *International Journal on Document Analysis and Recognition – IJDAR*, vol. 14, no. 2, p. 117–129.

MAYER, Rudolf (2011). Analysing the Similarity of Album Art with Self-organizing maps. In: J. Laaksonen, T. Honkela, eds. Advances in Self-Organizing Maps – WSOM 2011, *Lecture Notes in Computer Science*, vol. 6731, Heidelberg: Springer Verlag, p. 357–366.

MAYER, Rudolf.; RAUBER, Andreas (2011). On Wires and Cables: Content Analysis of WikiLeaks Using Self-Organising Maps. In: J. Laaksonen, T. Honkela, eds. Advances in Self-Organizing Maps – WSOM 2011, *Lecture Notes in Computer Science*, vol. 6731, Heidelberg: Springer Verlag, p. 238–246.

MOLYTĖ, Alma (2011). *Vektorių kvantavimo metodų jungimo su daugiamatėmis skalėmis analizė*: Daktaro disertacija. Vilniaus universitetas.

SIHAG, K. Vikas; KUMAR, Subhash (2013). Graph based Text Document Clustering by Detecting Initial Centroids for k-means. *International Journal of Computer Applications* (0975–8887), vol. 62, no. 19, p. 1–4.

SJOBERG, Mats; LAAKSONEN, Jorma (2011). Analysing the Structure of Semantic Concepts in Visual Databases. In: J. Laaksonen, T. Honkela, eds. Advances in Self-Organizing Maps – WSOM 2011, *Lecture Notes in Computer Science*, vol. 6731, Heidelberg: Springer Verlag, p. 338–347.

STEFANOVIČ, Pavel; KURASOVA, Olga (2009) Saviorganizuojančių neuroninių tinklų sistemų lyginamoji analizė. *Informacijos mokslai*, t. 50, p. 334–339.

STEFANOVIČ, Pavel; KURASOVA, Olga (2011). Visual analysis of self-organizing maps. *Nonlinear Analysis: Modeling and Control*, vol. 16(4), p. 488–504.

STEFANOVIČ, Pavel; KURASOVA, Olga (2011). Influence of Learning Rates and Neighboring Functions on Self-Organizing Maps. In: J. Laaksonen, T. Honkela, eds. Advances in Self-Organizing Maps – WSOM 2011, *Lecture Notes in Computer Science*, vol. 6731, Heidelberg: Springer Verlag, p. 141–150.

STEINBACH, Michael; KARYPIS, George; KUMAR, Vipin (2000). Comparison of document clustering techniques. *KDD Workshop on Text Mining.*

ZEIMPEKIS, Dimitrios; GALLOPOULOS, Efstratios (2005). *TMG: A Matlab Toolbox for Generating Term-Document Matrices from Text Collections.* Technical Report HPCLAB-SCG 1/01-05, University of Patras, GR-26500, Patras, Greece.

SIMILARITY ANALYSIS OF TEXT DOCUMENTS BY SELF-ORGANIZING MAPS AND *K*-MEANS

Pavel Stefanovič, Olga Kurasova

S u m m a r y

In this paper, we try to find similarities of different text documents by the self-organizing map (SOM) and *k*-means method. One of the main goals of these methods is to cluster a dataset. Using SOM, the similarities of documents can be observed visually. Both methods can be used only for numerical information, so we analyse the different options by converting text data on to numerical in order to get better results. To estimate the SOM quality, when the classified data are analysed, we propose two new measures: distances between SOM cells, corresponding to data items assigned to the same class, and the distance between centres of SOM cells, corresponding to different classes. We also analyse the results of visualization by self-organizing maps. In order to estimate the *k*-means quality, we calculate the sum of distances between cluster centres and class members and also we estimate assignment of the data from particular classes to the clusters. The experiments have been carried out using three datasets ocquired from the document database of Seimas of the Republic of Lithuania.