VILNIAUS UNIVERSITETAS

PAVEL STEFANOVIČ

SAVIORGANIZUOJANČIŲ NEURONINIŲ TINKLŲ VIZUALIZAVIMAS IR JO KOKYBĖS NUSTATYMAS

Daktaro disertacija,
Fiziniai mokslai, informatika (09 P)

Vilnius, 2015

Disertacija rengta 2010–2014 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinė vadovė:
doc. dr. Olga Kurasova (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

PADĖKA

Nuoširdžiai dėkoju mokslinio darbo vadovei dr. Olgai Kurasovai už nuoseklų vadovavimą studijų metu, laiku išsakytas pastabas, vertingas mokslines konsultacijas, visapusišką pagalbą ir kantrybę, rengiant šią disertaciją. Be galo Jai dėkingas už padrąsinimus, supratingumą, optimizmą bei įkvėpimą.

Esu dėkingas disertacijos recenzentams prof. habil. dr. Antanui Žilinskui ir prof. habil. dr. Kaziui Kazlauskui atidžiai perskaičiusiems disertaciją ir pateikusiems vertingų pastabų bei patarimų, gerinant šio darbo kokybę.

Dėkoju VU Matematikos ir informatikos instituto direktoriui prof. habil. dr. Gintautui Dzemydai už sudarytas sąlygas tobulėti ir skleisti savo idėjas Lietuvoje ir užsienyje. Dėkoju Matematikos ir informatikos instituto Sistemų analizės skyriaus kolegoms už kritiką ir draugišką pagalbą, rengiant disertaciją.

Dėkoju Lietuvos mokslo tarybai už suteiktą finansinę paramą disertacijos rengimo metu.

Nuoširdžiai dėkoju žmonai, tėvams už jų paramą, moralinį palaikymą ir supratingumą.

Taip pat dėkoju visiems, kurie tiesiogiai ar netiesiogiai prisidėjo prie šio darbo.

Pavel Stefanovič

REZIUMĖ

Dažnai nagrinėjant daugiamačius duomenis iškyla būtinybė ištirti susidariusius duomenų klasterius, surasti panašumų arba skirtumų tarp duomenų, spręsti įvairius klasifikavimo uždavinius. Vis dažniau susiduriama su tekstiniais duomenimis, todėl būtina rasti efektyvių metodų šiems duomenimis nagrinėti. Šio darbo tyrimų sritis yra tekstinių ir skaitinių duomenų analizė, naudojant saviorganizuojančius neuroninius tinklus (SOM).

Disertacijoje nagrinėjant saviorganizuojančius neuroninius tinklus, didelis dėmesys skiriamas tekstinių duomenų vizualizavimui bei gauto SOM žemėlapio kokybės įvertinimui. Darbe pasiūlytas ir įgyvendintas SOM vizualizavimo būdas, leidžiantis pamatyti skirtingų klasių duomenis, patekusius į tą patį saviorganizuojančio neuroninio tinklo langelį. Taip pat pasiūlytos dvi naujos paklaidos, skirtos įvertinti gauto SOM žemėlapio kokybę, analizuojant klasifikuotus duomenis. Pirmoji paklaida įvertina, kaip arti SOM žemėlapyje yra tos pačios klasės duomenys. Kuo paklaidos reikšmė mažesnė, tuo analizuojamos klasės klasteris yra „stipresnis“, t. y. klasė sudaro klasterį. Antroji paklaida skirta įvertinti, kaip toli SOM žemėlapyje yra skirtingų klasių centrai. Didesnė paklaidos reikšmė reiškia, jog žemėlapyje skirtingų klasių klasteriai labiau atsiskyrę ir aiškiau matomi SOM žemėlapyje. Darbe sukurta nauja SOM sistema, kuriame įgyvendintas pasiūlytas SOM vizualizavimo būdas ir SOM kokybę nustatančios paklaidos. Atlikta populiariausių saviorganizuojančių neuroninių tinklų sistemų ir naujai pasiūlytos sistemos lyginamoji analizė. Eksperimentiškai nustatyta SOM mokymo faktorių įtaka gautiems rezultatams. Taip pat eksperimentiškai ištirta, kaip skirtingi teksto konvertavimo į skaitinę išraišką faktoriai daro įtaką gautiems SOM rezultatams.

ABSTRACT

Often in the context of multidimensional data, there is a need to analyze the clusters, find similarities or differences between the data and to deal with various challenges of the classification. Increasingly, we are surrounded by a various text data, so it is necessary to find effective methods to analyze these kind of data. The area of research is analysis of the text and numeric datasets by using the self-organizing maps (SOM).

In the dissertation, the biggest focus is on the text data visualization and evaluation of the quality of the resulting map, using self-organizing maps. In this work, there is proposed and implemented the SOM visualization way, which helps a researcher to see the different classes of data in the same self-organizing map cell. Also, the two new errors are proposed, which are helpful to estimate the SOM map quality. The first error shows how close the same class members in the SOM are. The smaller value of error means the better results, it means that all same class members are closer to each other, the clusters are "stronger". The second error shows how far the centers of different classes are. The bigger value of error means the better results, i.e. all the different class centers are far from each other, so they are separated on the map. Both errors are suitable when classified data are analyzed. In this work, the new SOM system is developed, in which we can use the purposed SOM visualization way and two new errors. A comparative analysis of the most popular SOM systems and the new system have been done. The dependence of the self-organizing maps learning parameters to SOM results has been defined experimental. Also, how different factors of text data conversation to numerical expression make influence to SOM results are experimentally investigated.

Žymėjimai

α	saviorganizuojančio neuroninio tinklo (SOM) mokymo parametras
c	duomenų klasės numeris
E_c	paklaida, įvertinanti, kaip arti SOM žemėlapyje yra c-osios klasės nariai
E_{center}	paklaida, įvertinanti, kaip toli SOM žemėlapyje yra skirtingų klasių centrai
E_{QE}	kvantavimo paklaida
E_{TE}	topografinė paklaida
h	kaimynystės funkcija
k	klasių skaičius
k_x	saviorganizuojančio neuroninio tinklo žemėlapio (lentelės) eilučių skaičius
k_y	saviorganizuojančio neuroninio tinklo žemėlapio (lentelės) stulpelių skaičius
m	duomenų aibės vektorių skaičius
M_{ij}	saviorganizuojančio neuroninio tinklo neuronas
M_w	neuronas-nugalėtojas su indeksu w
n	duomenų aibės vektoriaus (taško) požymių skaičius
n_c	SOM žemėlapio neuronų, į kuriuos pateko c-osios klasės vektoriai, skaičius
N_c	c-osios klasės vektorių skaičius
N_w	kaimyninių neuronų indeksų aibė aplink neuroną su indeksu w

η_{ij}^{w}	neurono M_{ij} kaimynystės eilės numeris neurono-nugalėtojo M_w atžvilgiu
$x_{i1}, x_{i2}, \dots, x_{in}$	i-tojo duomenų aibės vektoriaus (taško) požymių reikšmės
$X_1, X_2, \dots, X_m$	duomenų aibės vektoriai (taškai)
R_{ij}	dvimatis vektorius, atitinkantis neurono M_{ij} indeksus
R_w	dvimatis vektorius, atitinkantis neurono-nugalėtojo M_w indeksus
t	einamasis mokymo žingsnis (iteracija arba epocha)
T	bendras mokymo žingsnių skaičius (iteracijų arba epochų)
u_{ij}	unifikuotos matricos elemento reikšmė
Z_i^c	SOM žemėlapio langelių indeksai, į kuriuos pateko c-osios klasės neuronai
Y^c	c-osios klasės centro indeksas

Turinys

1. ĮVADAS 13
1.1. Tyrimo sritis ir problemos aktualumas 13
1.2. Tyrimo objektas 14
1.3. Darbo tikslas ir uždaviniai 15
1.4. Tyrimo metodai 15
1.5. Darbo mokslinis naujumas 15
1.6. Ginamieji teiginiai 16
1.7. Darbo rezultatų praktinė reikšmė 16
1.8. Darbo rezultatų aprobavimas 16
1.9. Disertacijos struktūra 17

2. SAVIORGANIZUOJANČIŲ NEURONINIŲ TINKLŲ APŽVALGA 19
2.1. Duomenų paruošimas saviorganizuojantiems neuroniniams tinklams 19
2.1.1. Tekstinių dokumentų konvertavimas į skaitinius duomenis 20
2.1.2. Duomenų klasifikavimas 24
2.2. Saviorganizuojantys neuroniniai tinklai 25
2.2.1. Saviorganizuojančių neuroninių tinklų mokymas 26
2.2.2. Mokymo taisyklės faktoriai 28
2.2.3. Kvantavimo ir topografinės paklaidos 33
2.3. Įvairūs SOM praplėtimai ir modifikacijos 33
2.4. Saviorganizuojančių neuroninių tinklų vizualizavimas 37
2.5. Saviorganizuojančių neuroninių tinklų programinės sistemos 39
2.5.1. Vizualizavimas SOM-Toolbox sistemoje 39
2.5.2. Vizualizavimas Databionic ESOM sistemoje 41
2.5.3. Vizualizavimas Viscovery SOMine sistemoje 43
2.5.4. Vizualizavimas NeNet sistemoje 45
2.5.5. Kitos SOM sistemos 46
2.6. Antro skyriaus apibendrinimas 47

3. NAUJAS SOM VIZUALIZAVIMO BŪDAS BEI JO KOKYBĘ ĮVERTINANČIOS PAKLAIDOS 49
3.1. Naujos SOM kokybę įvertinančios paklaidos 49
3.2. Pasiūlytas SOM vizualizavimo būdas 54
3.2.1. Nauja SOM sistema 56
3.2.2. SOM sistemos grafinė naudotojo sąsaja 57
3.3. Trečiojo skyriaus apibendrinimas 61

4. EKSPERIMENTINIŲ TYRIMŲ REZULTATAI 63
4.1. Tyrimuose naudojami duomenys 63
4.2. Saviorganizuojančių neuroninių tinklų sistemų lyginamoji analizė 66
4.3. Saviorganizuojančių neuroninių tinklų mokymo faktorių tyrimas 71
4.3.1. Mokymo faktorių įtaka tiriant tekstinius duomenis 71
4.3.2. Mokymo faktorių įtaka tiriant skaitinius duomenis 78
4.4. Teksto dokumentų konvertavimo į skaitinę išraišką faktorių įtaka 85
4.4.1. Rankinis teksto dokumento žodyno sukūrimas 86
4.4.2. Automatinis teksto dokumento žodyno sukūrimas 90
4.4.2.1. Dažniausiai vartojamų žodžių sąrašas 90
4.4.2.2. Kamieno išskyrimo algoritmas 94
4.5. Žodžių pasikartojimų skaičiaus įtaka SOM rezultatams 99
4.5.1. Rezultatai, gauti naudojant SOM tinklą 99
4.5.2. Apibendrinti žodžių pasikartojimo skaičiaus įtakos rezultatai 108
4.5.3. Klasterizavimo rezultatai, gauti naudojant ***k***-vidurkių metodą 110
4.6. Ketvirtojo skyriaus rezultatai ir išvados 113

BENDROSIOS IŠVADOS 117

LITERATŪRA IR ŠALTINIAI 119

AUTORIAUS PUBLIKACIJŲ SĄRAŠAS DISERTACIJOS TEMA 129
Straipsniai recenzuojamuose periodiniuose mokslo leidiniuose 129
Santraukos tarptautinių konferencijų santraukų rinkiniuose 130

1. Įvadas

1.1. Tyrimo sritis ir problemos aktualumas

Šių laikų technologijos leidžia kaupti didelius kiekius įvairialypės informacijos bei ją talpinti kompiuterio atmintyje, išorinėse laikmenose arba internete. Ilgą laiką kaupiant informaciją, saugyklos tampa dideliu šiukšlynu, kuriame dažnai tampa sunku rasti reikalingus duomenis ar kitą naudingą informaciją. Šiuolaikinės technologijos mums leidžia surasti iš gausybės informacijos vieną ar kitą norimą dalyką greitai, tačiau rasta informacija dažnai būna nenaudinga, iškraipyta ar neesminė. Todėl tai tampa didele problema ir iššūkiu kiekvienam naudotojui. Vienas iš šios problemos sprendimų būdų yra panaudoti duomenų tyrybos metodus (angl. *data mining*), kurie leidžia duomenis susisteminti juos klasterizuojant, klasifikuojant bei esant galimybei jų rezultatus pateikti vizualiai.

Vienas iš duomenų tyrybos metodų yra saviorganizuojantis neuroninis tinklas (SOM). SOM dažnai vadinamas saviorganizuojančiu žemėlapiu, o kartais pradininko pavarde – Kohoneno žemėlapiu (Kohonen, 2001). SOM tinklai gali būti naudojami duomenims klasterizuoti ir vizualizuoti. SOM gali pagelbėti ieškant daugiamačių duomenų projekcijų mažesnio skaičiaus matmenų erdvėje. Nors jau praėjo daugiau nei 40 metų nuo SOM tinklų atsiradimo, tačiau jie ir toliau intensyviai tiriami ir taikomi. Laikui bėgant atsirado daug įvairių SOM praplėtimų ir modifikacijų, pradedant nuo mokymo taisyklėje įvestų naujų pakeitimų iki skirtingų SOM vizualizavimo būdų. Tačiau pagrindinis mokymo principas išlieka tas pats. Daug metų SOM tinklai buvo taikomi įvairiems skaitinės išraiškos duomenims klasifikuoti ir klasterizuoti, bet šiuo metu taikymų sritis yra plečiama tiriant tekstinius ar kito tipo duomenis.

Vienas iš SOM tinklų privalumų, lyginant su kitais duomenų tyrybos metodais yra tai, kad gaunami ne tik skaitiniai įverčiai, kaip būna daugumoje kitų duomenų tyrybos metodų, bet ir jų rezultatai pateikiami vizualia forma, o vizualią informaciją žmogus suvokia greičiau nei tekstinę ar skaitinę. SOM

tinklai dažnai taikomi duomenims klasterizuoti. Lyginant su kitais klasterizavimo metodais, jie pasižymi tuo, kad čia nėra gaunami tiksliai apibrėžti klasteriai, t. y. duomenys nėra vienareikšmiškai priskiriami vienam ar kitam klasteriui. Klasterizavimo rezultatus gali įvairiai interpretuoti pats tyrėjas, stebėdamas vizualų SOM vaizdą. Tai leidžia pastebėti duomenų tarpusavio panašumą ir grupes, kurios iš anksto nėra žinomos, o tai gali būti privalumu prieš kitus klasterizavimo metodus. SOM tinklai gali būti taikomi ir duomenims, kurie jau yra priskirti klasėms, klasterizuoti. Tuomet tyrėjas gali matyti, ar klasės sutampa su SOM gautais klasteriais, ir aiškintis to nesutapimo priežastis, kurių viena gali būti susijusi su tuo, kad duomenys buvo netiksliai priskirti klasėms.

Šiuo metu yra sukurta įvairių programinių sistemų, kuriose įgyvendinti įvairūs SOM vizualizavimo būdai, tačiau trūksta sistemų, kuriose, vizualizuojant SOM tinklą, būtų matoma, kiek ir kokios klasės duomenų priskirta kiekvienam SOM tinklo langeliui. Problema yra ir ta, kad nėra skaitinių įverčių, parodančių duomenų klasių ir SOM gautų klasterių sutapimą.

Be to, SOM rezultatas labai priklauso nuo įvairių mokymo faktorių parinkimo, todėl iškyla problema, kokias faktorių reikšmes parinkti analizuojamiems duomenis. Taip pat svarbu ištirti, kokios reikšmės leidžia gauti tikslesnius rezultatus, kai analizuojami skirtingo tipo duomenys: tekstiniai ir skaitiniai.

Taigi šioje disertacijoje sprendžiamos dvi pagrindinės problemos:

1. Duomenų, priskirtų tam tikroms klasėms, vizualizavimas, taikant saviorganizuojančius neuroninius tinklus, ir gautų rezultatų kokybės vertinimas.
2. Gautų rezultatų priklausomybė nuo saviorganizuojančio tinklo mokymo faktorių reikšmių parinkimo.

1.2. Tyrimo objektas

Disertacijos tyrimo objektas – duomenų klasterizavimas, klasifikavimas ir vizualizavimas, naudojant saviorganizuojančius neuroninius tinklus, bei jų kokybės vertinimas.

1.3. Darbo tikslas ir uždaviniai

Darbo tikslas – sukurti saviorganizuojančių neuroninių tinklų vizualizavimo būdą, leisiantį vizualizuoti skaitinius ir tekstinius duomenis, kurių klasės iš anksto žinomos, ir stebėti gautų klasterių ir duomenų klasių sutapimą bei pasiūlyti ir ištirti šiuos sutapimus įvertinančias paklaidas.

Siekiant tikslo būtina spręsti šiuos uždavinius:

1. Atlikti esamų SOM vizualizavimo būdų analitinę apžvalgą.
2. Pasiūlyti paklaidas, įvertinančias SOM gautų klasterių ir duomenų klasių sutapimą.
3. Pasiūlyti SOM vizualizavimo būdą duomenims, kurių klasės yra žinomos, tirti.
4. Sukurti programinę sistemą, kurioje įgyvendintas pasiūlytas SOM vizualizavimo būdas, bei SOM kokybę įvertinančias paklaidas.
5. Eksperimentiškai ištirti pasiūlytą SOM vizualizavimo būdą, paklaidas, priklausomai nuo SOM mokymo faktorių reikšmių, tiriant skaitinius ir tekstinius duomenis.

1.4. Tyrimo metodai

Analizuojant mokslinius ir eksperimentinius pasiekimus saviorganizuojančių neuroninių tinklų srityje, buvo naudoti informacijos paieškos, sisteminimo, analizės, lyginamosios analizės ir apibendrinimo metodai. Remiantis eksperimentinio tyrimo metodu, atlikta statistinė duomenų ir tyrimų rezultatų analizė, kurios rezultatams įvertinti naudotas apibendrinimo metodas.

1.5. Darbo mokslinis naujumas

1. Pasiūlytas SOM vizualizavimo būdas, skirtas skirtingų klasių tiek tekstinių, tiek skaitinių duomenų, pakliuvusių į vieną SOM langelį, santykiui pavaizduoti.
2. Pasiūlytos naujos SOM kokybę įvertinančios paklaidos, kai analizuojami duomenys, priskirti iš anksto žinomoms klasėms.

3. Ištirta tekstinių dokumentų konvertavimo į skaitinę išraišką faktorių įtaka gautiems SOM žemėlapio rezultatams.

1.6. Ginamieji teiginiai

1. Pasiūlytas SOM vizualizavimo būdas leidžia pavaizduoti skirtingų klasių duomenų, pakliuvusių į tą patį SOM žemėlapio langelį, santykius.
2. Pasiūlytos SOM kokybės įvertinimo paklaidos leidžia įvertinti duomenų klasių ir SOM gautų klasterių atitikimą.
3. Tekstinių dokumentų konvertavimo į skaitinę išraišką tinkamas faktorių parinkimas pagerina gautus SOM rezultatus.

1.7. Darbo rezultatų praktinė reikšmė

Sukurta SOM programinė sistema, kurioje įgyvendintas ne tik pasiūlytas SOM vizualizavimo būdas bei SOM kokybę nustatančios paklaidos, bet ir yra galimybė pasirinkti įvairias kaimynystės funkcijas bei mokymo parametrus, kurių reikšmės gali keistis arba kiekvienoje iteracijoje, arba kiekvienoje epochoje. Taip pat yra galimybė išskaidyti nagrinėjamą duomenų aibę į du poaibius: mokymo ir testavimo. Dėl šių priežasčių sukurta SOM sistema gali būti naudojama ne tik duomenims analizuoti, bet ir SOM tinklui tirti. Dalis tyrimų rezultatų gauti vykdant Europos socialinio fondo finansuojamą projektą „Paslaugų interneto technologijų kūrimo ir panaudojimo našių skaičiavimų platformose teoriniai ir inžineriniai aspektai“ (Nr. VP1-3.1-ŠMM-08-K-01-010).

1.8. Darbo rezultatų aprobavimas

Tyrimų rezultatai publikuoti 7 moksliniuose leidiniuose: 5 periodiniuose recenzuojamuose mokslo žurnaluose, iš jų – 2 leidiniuose, referuojamuose „Thomson Reuters Web of Science“ duomenų bazėje ir turinčiuose citavimo indeksą; bei 2 straipsniai – konferencijų pranešimų medžiagoje. Taip pat publikuotos 2 santraukos tarptautinių konferencijų santraukų rinkiniuose. Tyrimų rezultatai buvo pristatyti ir aptarti šiose nacionalinėse ir tarptautinėse konferencijose Lietuvoje ir užsienyje:

1. XIV kompiuterininkų konferencija: Kompiuterininkų dienos – 2009. 2009 m. rugsėjo 25–26 d., Kauno technologijos universitetas, Kaunas, Lietuva.
2. 3-ioji Lietuvos jaunųjų mokslininkų konferencija (LOTD 2010): Operacijų tyrimai versle, inžinerijoje ir informacinėse technologijose. 2010 m. spalio 1 d., Mykolo Romerio universitetas, Vilnius, Lietuva.
3. 8th International Workshop on Self-organizing Maps (WSOM 2011). June 14 – June 17, 2011, Aalto university, Espoo, Finland.
4. XXV International Conference on Operational Research (EURO 2012). July 8 – July 11, 2012, Vilnius, Lithuania.
5. Konferencija „Fizinių ir technologijos mokslų tarpdalykiniai tyrimai“. 2013 m. vasario 12 d., Lietuvos mokslų akademija, Vilnius, Lietuva.
6. XXVI International Conference on Operational Research (EURO 2013). July 1 – July 4, 2013, Rome, Italy.
7. XVI kompiuterininkų konferencija: Kompiuterininkų dienos – 2013. 2013 m. rugsėjo 19–21 d., Šiaulių universitetas, Šiauliai, Lietuva.

1.9. Disertacijos struktūra

Disertaciją sudaro 5 skyriai ir literatūros sąrašas. Pirmas disertacijos skyrius yra „Įvadas“. Šioje dalyje yra pateiktos disertacijoje sprendžiamos problemos, tikslas, uždaviniai, naudoti tyrimo metodai, ginamieji teiginiai, mokslinis naujumas, disertacijos praktinė reikšmė bei darbo rezultatų aprobavimas. Antras skyrius „Saviorganizuojančių neuroninių tinklų apžvalga“ skirtas saviorganizuojančių neuroninių tinklų metodui bei jo mokymo faktoriams pristatyti. Aprašytas būdas, kaip yra konvertuojami tekstiniai duomenys į skaitinę išraišką. Apžvelgtos šio metodo modifikacijos bei praplėtimai. Pateiktos populiariausių saviorganizuojančių neuroninių tinklų sistemų vizualizavimo galimybės. Trečiame skyriuje „Naujas SOM vizualizavimo būdas bei jo kokybę nustatančios paklaidos“ pateiktas naujai pasiūlytas SOM vizualizavimo būdas, aprašytas jo pranašumas, lyginant su kitais antrame skyriuje aprašytais sistemų vizualizavimo būdais. Taip pat

pasiūlyti matai, kurie leidžia įvertinti gauto SOM žemėlapio kokybę, kai yra nagrinėjami duomenys, kurių klasės yra iš anksto žinomos. Skyriuje „Eksperimentinių tyrimų rezultatai“ pateikti eksperimentiniai tyrimai, jų aprašymai bei gauti rezultatai. Paskutiniame skyriuje „Bendrosios išvados“ yra disertacijos išvados bei rezultatai. Papildomai disertacijoje pateiktas naudotų žymėjimų sąrašas. Bendra disertacijos apimtis – 132 puslapiai, kuriuose – 49 paveikslai ir 27 lentelės. Disertacijoje remtasi 87 literatūros šaltiniais.

2. Saviorganizuojančių neuroninių tinklų apžvalga

Saviorganizuojantys neuroniniai tinklai (SOM) buvo pradėti kurti prieš 40 metų (Kohonen, 1982; 2001), tačiau nuo to praėjus netrumpam laikotarpiui šiuos tinklus vis dar tiria ir taiko daug pasaulio mokslininkų. Atsiranda vis naujų taikymo sričių, kuriose gali būti panaudoti ir SOM tinklai. SOM tinklai yra dažnai vadinami SOM žemėlapiais arba SOM lentelėmis. Didėjant duomenų kiekiams ir tobulėjant skaičiavimo resursams galima gauti tikslesnius rezultatus, todėl dirbtinių neuroninių tinklų, taip pat ir SOM tinklų, analizė ir taikymas tebelieka svarbia informatikos mokslo šaka. Laikui bėgant atsirado daug įvairių SOM modifikacijų ir praplėtimų, pradedant nuo mokymo taisyklės pakeitimo iki savitų vizualizavimo būdų, tačiau daugumos jų mokymas išlaiko pradinę idėją. Dauguma modifikacijų yra pritaikytos konkretiems specifiniams uždaviniams spręsti, pavyzdžiui, tekstinių, vaizdinių duomenų analizei ir kt. Šiame skyriuje aprašyti SOM tinklai ir jų mokymas bei apžvelgti SOM mokymo faktoriai. Taip pat pateikta esamų SOM sistemų lyginamoji analizė, apžvalgos, publikuotos autoriaus darbuose [A1], [A2], [B2].

2.1. Duomenų paruošimas saviorganizuojantiems neuroniniams tinklams

Sakykime, kad turime analizuojamų duomenų aibę $X = \{X_1, X_2, \dots, X_m\}$, kurią sudaro taškai (vektoriai) $X_1, X_2, \dots, X_m$, charakterizuojantys tam tikrą objektų aibę, kuri apibūdinama bendrais požymiais $x_1, x_2, \dots, x_n$; čia m – analizuojamų objektų (vektorių) skaičius, n – požymių skaičius. Dažniausiai požymiai gali įgyti skaitines reikšmes. Tuomet tokius duomenis vadinsime *skaitiniais*. Požymių reikšmės $x_{p1}, x_{p2}, \dots, x_{pn}$ yra vektoriaus $X_p, p \in \{1, \dots, m\}$ komponentės. Taigi, $X_1, X_2, \dots, X_m$ yra n-mačiai vektoriai, kurie interpretuojami kaip taškai n-matėje erdvėje R^n, čia n – erdvės matmenų skaičius. Įprastai duomenų aibė saviorganizuojančiam neuroniniam tinklui yra pateikiama kaip matrica (1).

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix}. \quad (1)$$

Čia x_{pl} yra vektoriaus X_p, $p = 1, \dots, m$, l-osios komponentės reikšmė, $l = 1, \dots, n$.

2.1.1. Tekstinių dokumentų konvertavimas į skaitinius duomenis

Tam, kad galėtume analizuoti tekstinius dokumentus įvairiais duomenų analizės metodais, būtina dokumentus paversti į skaitinę išraišką. Tuo tikslu turi būti sukurta vadinamoji teksto dokumentų matrica. Duomenis, kurie gauti iš tekstinių dokumentų, vadinsime *tekstiniais* duomenimis. Tekstiniai duomenys nuo skaitinių duomenų skiriasi tuo, jog tekstinių duomenų atveju gaunami vektoriai, kurių požymių skaičius įprastai būna didelis (gali siekti ir kelis tūkstančius, tai priklauso nuo žodžių skaičiaus žodyne), o duomenų matricos – išretintos, t. y. daugelis matricos reikšmių yra lygios 0. Teksto dokumentų konvertavimas į skaitinę išraišką gali būti vykdomas šiais etapais:

1. Tekstiniai dokumentai yra konvertuojami į atskirus tekstinius failus, kuriuose lieka tik tekstinė informacija, atmetant grafinę informaciją.
2. Pasirenkami faktoriai, į kuriuos bus atsižvelgiama, sudarant teksto dokumentų žodyną:
   - skaitmenų atmetimas;
   - žodžio ilgio parinkimas;
   - žodžių pasikartojimas dokumente;
   - dažniausiai vartojamų žodžių sąrašo įtraukimas;
   - kamieno išskyrimo algoritmo taikymas.
3. Pagal pasirinktus faktorius sudaromas *teksto dokumentų žodynas*, į kurį įtraukiami tik žodžiai, atitinkantys pasirinktus faktorius.
4. Atsižvelgiant į dokumentų žodyną, sudaroma *teksto dokumentų matrica*, kurios išraiška pateikta formulėje (1). Viena matricos eilutė atitinka vieną analizuojamos duomenų aibės dokumentą. Šiuo atveju x_{pl} yra l-ojo žodžio pasikartojimas p-ajame dokumente, $p = 1, \dots, m$. m – analizuojamų

dokumentų skaičius, o n – žodžių skaičius teksto dokumentų žodyne. Vektorių $X_p = (x_{p1}, x_{p2}, \dots, x_{pn})$ matmenų skaičius n visada priklauso nuo bendro žodžių skaičiaus teksto dokumentų žodyne.

Toliau detaliau aprašyti tekstinių dokumentų žodyno sudarymo faktoriai:

- **Skaitmenų atmetimas.** Beveik kiekviename teksto dokumente yra skaitmenų ar įvairios skaitinės informacijos, todėl nėra prasminga juos įtraukti, sudarant teksto dokumentų žodyną, kadangi ši informacija nėra esminė ir neatspindi dokumento prasmės.
- **Žodžių ilgis.** Žodžių ilgio apribojimas leidžia atmesti žodžius, kurie yra labai trumpi ar labai ilgi. Kiekviename dokumente galima rasti trumpų žodelių, tokių kaip prielinksniai ar jungtukai: „į“, „iš“, „o“ ir kt., todėl, sudarant žodyną, tikslinga juos atmesti. Jų dokumente yra daug, tačiau jie nesuteikia jokios naudingos informacijos.
- **Žodžių pasikartojimas dokumente.** Labai svarbu parinkti tinkamą žodžių pasikartojimų skaičių. Nurodant per mažą skaičių, į žodyną gali būti įtraukiami visai neesminiai žodžiai. Nurodant per didelį skaičių, dokumentai gali būti atmesti, kadangi gali atsitikti taip, kad juose nebus tiek kartų pasikartojančio žodžio.
- **Dažnai vartojamų žodžių sąrašas.** Sudarant tekstinių dokumentų žodyną, galima neįtraukti dažnai vartojamų ir sutinkamų visuose dokumentuose neesminių žodžių. Prieš tai reikia sudaryti tokių žodžių sąrašą. Pavyzdžiui, tokie trumpi žodeliai (jungtukai, prielinksniai, įvardžiai ir kt.): „čia“, „kur“, „kada“, „šie“, „ten“ ir t. t. Jie visai necharakterizuoja analizuojamo dokumento, tačiau kai kurie iš jų yra per ilgi, kad būtų atmesti vien dėl žodžio apribojimo ilgio. Be abejo, dažniausiai vartojamų žodžių sąrašas turėtų būti sudaromas ir pritaikomas priklausomai nuo analizuojamos srities. Tarkime, jeigu analizuojame panašaus pobūdžio dokumentus apie optimizavimą ir tikslas – parodyti, kokie skirtingi optimizavimo metodai yra aprašyti ar taikyti juose, būtina įtraukti į sąrašą žodžius: „optimizavimas“, „metodas“ ir kitus. Jie

greičiausiai bus kiekviename dokumente. Jeigu nežinome nieko apie analizuojamus dokumentus, patartina įtraukti bent pagrindinius jungtukus, prielinksnius ir kitus tyrėjo nuomone nesvarbius žodžius.

- **Kamieno išskyrimo algoritmas.** Dažnai svarbu iš žodžių išskirti kamieną, kuris vėliau yra įtraukiamas į teksto dokumentų žodyną, kadangi būtent kamienas nurodo pagrindinę žodžio prasmę. Kamieno išskyrimo algoritmai yra sukurti kelioms kalboms (Kamieniniai algoritmai, 2014). Anglų kalboje dažniausiai yra naudojamas Porterio kamieno išskyrimo algoritmas (Porter, 1980). Tarkime, dokumentas parašytas anglų kalba, kur yra žodžiai: „accepted“, „acception“, „acceptable“. Natūralu, jog šio žodžio reikšmė yra ta pati, todėl būtent dėl kamieno išskyrimo algoritmo į žodyną yra įtraukiamas tik vienas žodis „accept“. Lietuvoje taip pat galima rasti darbų, kurie nagrinėja lietuvių kalbos morfologiją. Kuriami įrankiai, padedantys nagrinėti žodžių sandarą, išskirti žodžio dalis ar skirtingas žodžių formas (Zinkevičius, 2004). Taip pat yra sukurtas kamieno išskyrimo algoritmas lietuvių kalbai (Krilavičius, Medelis, 2010) bei kiti įrankiai lietuvių kalbai nagrinėti (Krilavičius, Kuliešienė, 2010).

Yra sukurta įvairių įrankių, padedančių iš teksto dokumentų duomenų aibės sukurti teksto dokumentų matricą. Vienas iš jų yra „Text to Matrix Generator“ – papildomas įrankis Matlab aplinkai (Zeimpekis, Gallopoulos, 2005). Tokia populiari duomenų analizės sistema, kaip KNIME (Berthold ir kiti, 2007), taip pat yra naudojama teksto dokumentams į skaitinę išraišką konvertuoti.

Teksto dokumentų žodyno ir matricos sudarymas pademonstruotas paprastu pavyzdžiu. Tarkime, turime du tekstinius dokumentus (straipsnių santraukas):

1. Straipsnyje nagrinėjamos ir lyginamos tarpusavyje 3 saviorganizuojančių neuroninių tinklų (SOM) sistemos: NeNet, SOM-toolbox ir Databionic ESOM. Pagrindinis šių sistemų tikslas yra suskirstyti duomenis į klasterius pagal jų panašumą, pateikti juos SOM žemėlapyje. Sistemos viena nuo kitos skiriasi duomenų pateikimu, mokymo taisyklėmis, vizualizavimo galimybėmis, todėl čia aptariami sistemų panašumai ir skirtumai. SOM žemėlapiams mokyti ir vizualizuoti naudojamos 2 duomenų aibės: irisų ir stiklo duomenys.

2. Straipsnyje nagrinėjama ir lyginama dokumentų panašumų paieška naudojant du klasterizavimo metodus: saviorganizuojančius neuroninius tinklus (SOM) ir k-vidurkių metodą. Vienas iš šių metodų tikslų – suskirstyti duomenis į klasterius pagal jų panašumą. Analizuota dokumentų matricos sudarymo parametrų parinkimo įtaka gautiems rezultatams. SOM kokybei įvertinti naudoti du nauji matai, klasifikuotiems duomenims, kurie parodo susidariusių klasterių išsidėstymą SOM žemėlapyje. k-vidurkių kokybei įvertinti skaičiuota susidariusių klasterių atstumų suma nuo klasterio centro iki klasterio narių, bei skaičiuotas neteisingai priskirtų klasėms duomenų skaičius. Tyrimams atlikti naudoti dokumentai, paimti iš Lietuvos Respublikos Seimo dokumentų bazės.

Visų pirma, yra parenkami faktoriai, į kuriuos bus atsižvelgta, sudarant tekstinių dokumentų žodyną. Tarkime šiems tekstiniams dokumentams parenkami tokie faktoriai:

- skaitmenų atmetimas;
- žodžių ilgis ne mažesnis kaip trys raidės;
- žodžiai pasikartoja dokumente ne mažiau kaip du kartus;
- įtraukiamas dažniausiai vartojamų žodžių sąrašas, kuriame yra šie žodžiai: „čia“, „iki“, „juos“, „kurie“, „nuo“, „šių“, „tai“, „todėl“, „yra“;
- kamieno išskyrimo algoritmas nėra taikomas.

Atsižvelgiant į panaudotus faktorius, gautas tekstinių dokumentų žodynas, kuriame yra 11 žodžių: „dokumentų“, „duomenų“, „klasterių“, „kokybei“, „naudoti“, „sistemos“, „sistemų“, „som“, „susidariusių“, „vidurkių“, „įvertinti“. Tuomet pagal šį žodyną sudaroma tekstinių dokumentų matrica:

$$\begin{pmatrix} 0 & 2 & 0 & 0 & 0 & 2 & 2 & 4 & 0 & 0 & 0 \\ 3 & 0 & 2 & 2 & 2 & 0 & 0 & 3 & 2 & 2 & 2 \end{pmatrix}.$$

Čia pirma matricos eilutė atitinka pirmą tekstinį dokumentą, antra – antrą. Stulpeliai atitinka žodžius iš dokumentų žodyno (iš eilės), o matricų elementų reikšmės atitinka žodžio pasikartojimą dokumente. Tarkime pažvelgę į šią gautą matricą matome, jog 8-ame matricos stulpelyje reikšmės atitinka žodžio „som“ pasikartojimą dokumente, pirmame dokumente jis pasikartoja 4 kartus, o antrame – 3 kartus.

Jeigu šiai duomenų aibei būtų pritaikytas kamieno išskyrimo algoritmas, vietoje žodžių „metodas“, „metodų“, „metodai“ į žodyną būtų įtrauktas vienas žodis – „metod“, kuris atspindi visus juos tris. Nenaudojant kamieno išskyrimo algoritmo šie žodžiai laikomi skirtingais žodžiais.

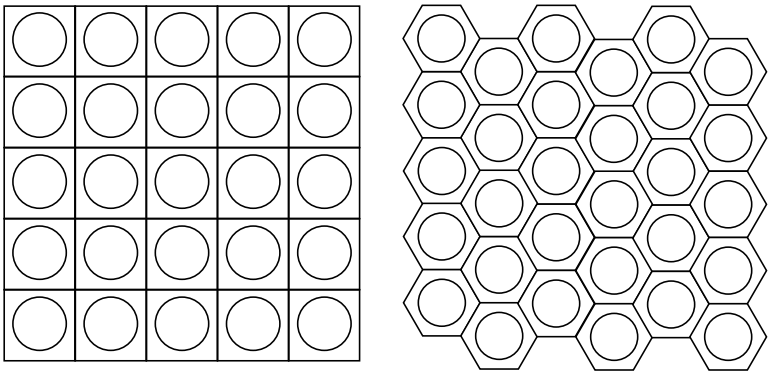
2.1.2. Duomenų klasifikavimas

Praktikoje įvairiose srityse dažnai tenka spręsti klasifikavimo ir klasterizavimo uždavinius. Klasifikavimo uždavinys yra specifinis atpažinimo uždavinys, kurio tikslas – duomenis priskirti vienai iš žinomų klasių. Duomenų klasterizavimo metu duomenys suskirstomi į grupes (klasterius) pagal jų panašumą taip, kad panašūs duomenys pakliūna į tą patį klasterį, o nepanašūs – į skirtingus. Šiame darbe didelis dėmesys skirtas klasifikuotiems duomenis analizuoti, t. y., kai duomenų aibių klasės yra iš anksto žinomos. Klasifikavimo uždaviniams spręsti taikomi įvairūs metodai: klasifikavimo medžiai, artimiausių kaimynų, atraminių vektorių klasifikatoriai, Naive Bayes ir kt. (Dunham, 2002). Kartais klasifikuoti duomenys dar yra klasterizuojami (Zeng ir kiti, 2003), (Guven, Cengizler, 2014), (Zhang, Xiao, 2012). Šiam tikslui naudojami ir saviorganizuojantys neuroniniai tinklai (Saarikoski, 2014), (Vasighi, Kompany-Zareh, 2013), (Yosob ir kiti, 2013).

Įprastai klasifikavimo metoduose naudojami mokymo duomenys, kurie jau priskirti tam tikroms klasėms, ir pagal juos duomenys, kurių klasės nėra žinomos, priskiriami vienai iš tų klasių. Norint pasiekti tikslų klasifikavimo rezultatų, būtina sąlyga – mokymo duomenys turi būti iš anksto tinkamai priskirti klasėms. Netinkamo priskyrimo priežasčių gali būti įvairių: matavimo paklaidos, žmogiškieji faktoriai ir kt. Tai ypač svarbu nagrinėjant, pavyzdžiui, medicininius duomenis, kadangi netinkamas mokymo duomenų priskyrimas gali iškreipti klasifikavimo rezultatus ir vėliau įtakoti ligos diagnostiką. L. Ringienės (2014) disertacijoje pasiūlytas dirbtinių neuroninių tinklų, pagrįstų bazinėmis radialinėmis funkcijomis, modelis galėtų būti taikomas siekiant nustatyti duomenis, kurie priskirti galimai netinkamoms klasėms. Tačiau tame darbe nėra skaitinių paklaidų, įvertinančių tokį priskyrimą.

2.2. Saviorganizuojantys neuroniniai tinklai

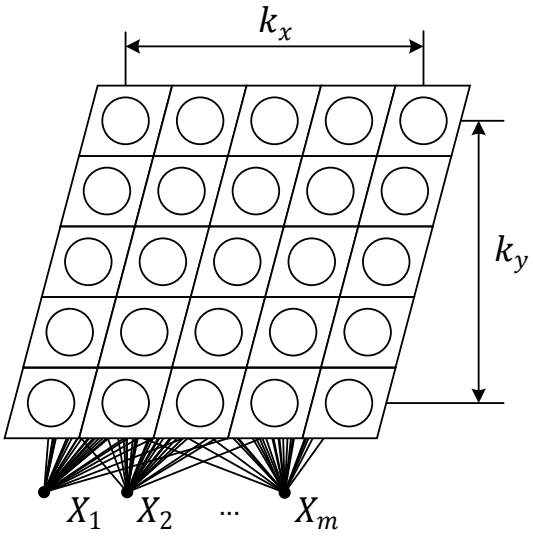
Pagrindinis saviorganizuojančio neuroninio tinklo tikslas – išlaikyti duomenų topologiją, t. y. taškai, esantys arti įėjimo vektorių erdvėje, turi būti atvaizduoti arti vieni kitų ir SOM tinkle. SOM tinklai gali būti naudojami duomenis klasterizuoti ir susidariusiems klasteriams vizualizuoti, taip pat ieškant daugiamačių duomenų projekcijų į mažesnio skaičiaus matmenų erdvę. SOM tinklas yra neuronų, paprastai išdėstytų dvimačio tinklelio, dar vadinamo *žemėlapiu* arba *lentele*, mazguose, masyvas $M = \{M_{ij}, i = 1, \dots, k_x, j = 1, \dots, k_y\}$ (Kohonen, 2001), čia M_{ij} – vektorius, kurio matmenų skaičius n yra toks pat kaip mokymo duomenų vektorių, k_x yra tinklo eilučių skaičius, k_y – tinklo stulpelių skaičius (visas SOM žemėlapio dydis $k_x \times k_y$). SOM žemėlapio dydis dažniausiai parenkamas eksperimentiškai, kadangi sunku iš anksto nustatyti, koks yra optimalus SOM dydis. Natūralu, kad kuo didesnė duomenų aibė, tuo žemėlapis turi būti parenkamas didesnis. Galima *stačiakampė* (angl. *rectangular*) arba *šešiakampė* (angl. *hexagonal*) tinklo struktūra (2.1 pav.).



a) **b)**

2.1 pav. SOM tinklo (žemėlapio) struktūra: a) stačiakampė, b) šešiakampė

Dažniausiai yra analizuojami dvimačiai SOM tinklai (2.2 pav.), nors galimi ir didesnio matmenų skaičiaus tinklai.



2.2 pav. Dvimatis SOM tinklas (žemėlapis)

2.2.1. Saviorganizuojančių neuroninių tinklų mokymas

SOM mokymas priskiriamas mokymui be mokytojo (angl. *unsupervised learning*). Mokymo pradžioje neuronų (vektorių) M_{ij} komponenčių pradinės reikšmės dažniausiai nustatomos atsitiktinai. Mokymo eigoje SOM tinklui daug kartų pateikiami n-mačiai vektoriai $X_1, X_2, \dots, X_m$. Kiekviename mokymo žingsnyje vienas mokymo aibės vektorius $X_p \in \{X_1, X_2, \dots, X_m\}$ pateikiamas į tinklą. Vektorius X_p palyginamas su visais neuronais M_{ij}: dažniausiai skaičiuojamas Euklido atstumas $\|X_p - M_{ij}\|$ tarp šio vektoriaus X_p ir kiekvieno

neurono $M_{ij} = \{m_{ij}^1, m_{ij}^2, \dots, m_{ij}^n\}$. Randama, iki kurio neurono $M_w \in \{M_{ij}, i = 1, \dots, k_x, j = 1, \dots, k_y\}$ atstumas yra mažiausias; rastas neuronas M_w vadinamas *neuronu (vektoriumi)-nugalėtoju* (angl. *neuron-winner* arba *best matching unit*). Visų tinklo neuronų komponentės keičiamos naudojantis formule:

$$M_{ij}(t+1) = M_{ij}(t) + h_{ij}^w(t)(X_p - M_{ij}(t)). \qquad (2)$$

Čia t yra mokymo žingsnio numeris, $h_{ij}^w(t)$ – *kaimynystės funkcija*. Žemiau yra pateiktas SOM mokymo algoritmo pseudokodas, kai mokymo žingsnis – epocha.

FOR $t = 1$ **TO** T // kiekviename mokymo žingsnyje
  FOR $p = 1$ **TO** m // kiekvienam mokymo aibės vektoriui
    FOR $i = 1$ **TO** k_x
      FOR $j = 1$ **TO** k_y
        $\|X_p - M_{ij}\| = \sqrt{\sum_{k=1}^{n}(x_{pk} - m_{ij}^k)^2}$ //skaičiuojamas Euklido atstumas
      END
    END
      $M_w = arg\min_{ij}\{\|X_p - M_{ij}\|\}$ // M_w – vektoriaus X_p neuronas-nugalėtojas
    FOR $i = 1$ **TO** k_x
      FOR $j = 1$ **TO** k_y
        $M_{ij}(t+1) = M_{ij}(t) + h_{ij}^w(t)(X_p - M_{ij}(t))$ // SOM mokymo taisyklė
      END
    END
  END
END

Yra įvairių SOM tinklo mokymo variantų, kurie vienas nuo kito skiriasi kaimynystės funkcijos $h_{ij}^w(t)$ išraiška. Tai yra euristinės funkcijos, todėl griežtų matematinių konvergavimo įrodymų nėra, ir skirtingų mokymo taisyklių rezultatuose gali būti šiek tiek kitokie žemėlapiai. Stabilūs analizuojamų duomenų klasteriai įprastai išlieka visuose žemėlapiuose, tačiau gali būti duomenų, kurie priskiriami vis prie kitų klasterių arba visai jų nesudaro. Tačiau tai yra savotiškas metodo privalumas, nes pagrindinis vizualizavimo tikslas – padėti suvokti analizuojamus duomenis, atskleisti jų struktūrą, kelti hipotezes

dėl analizuojamų duomenų aibės. Keli gauti vaizdai tai padaryti padeda daug efektyviau (Dzemyda ir kiti, 2008).

Baigus SOM tinklo mokymą, *mokymo* ar kita, vadinamoji *testavimo duomenų aibė*, pateikiama į tinką, randamas kiekvieno vektoriaus neuronas-nugalėtojas, jį atitinkančiame žemėlapio langelyje užrašomas vektoriaus eilės numeris arba klasės, kuriai priklauso šis vektorius, pavadinimas. Tarkime turime duomenų aibę, sudarytą iš 6 vektorių, kurių matmenų skaičius lygus 4: $X_1, X_2, X_3, X_4, X_5, X_6$. Tuomet išmokyto SOM tinklo langeliuose nurodomi vektorių eilės numeriai (2.3 pav.).

1,2		5
		3,6
	4	

2.3 pav. SOM tinklas (nurodyti vektorių eilės numeriai)

SOM tinkle galima atvaizduoti ir atskirus duomenų aibės vektorių požymius (angl. *component planes*). Šis atvaizdavimo būdas leidžia pamatyti, kuris konkretus požymis labiau daro įtaką gautiems SOM tinklo rezultatams.

2.2.2. Mokymo taisyklės faktoriai

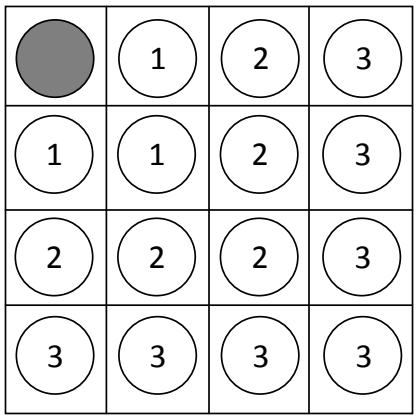
Turėdami SOM tinkle atvaizduotus neuronus-nugalėtojus, galime apibrėžti sąvoką „*kaimynystės eilė*“. Apie neuroną-nugalėtoją esantys neuronai yra vadinami pirmos eilės kaimynais. Toliau esantys neuronai yra antros eilės kaimynai ir t. t. 2.4 pav. neuronas-nugalėtojas pažymėtas pilku skrituliuku, kituose skrituliukuose pateikti skaičiai nurodo kaimynystės eilę neurono-nugalėtojo atžvilgiu.

SOM tinklo rezultatai priklauso nuo pasirinktų mokymo faktorių (Tan, George, 2004), todėl labai svarbu parinkti tinkamas jų reikšmes tam, kad gautume geresnius rezultatus. Daugiausiai tam įtakos turi *kaimynystės funkcijos* $h_{ij}^{w}(t)$ ir *mokymo parametrai* $\alpha(t)$. Paprastai literatūroje sutinkamos ir plačiai naudojamos dvi kaimynystės funkcijos: burbuliuko (3) ir Gauso (4).

$$h_{ij}^{w}(t) = \begin{cases} \alpha(t), (i,j) \in N_w \\ 0, (i,j) \in N_w \end{cases}, \quad (3)$$

$$h_{ij}^{w}(t) = \alpha(t) \cdot \exp\left(\frac{-\|R_w - R_{ij}\|^2}{2\left(\eta_{ij}^{w}(t)\right)^2}\right). \quad (4)$$

Čia N_w yra kaimyninių neuronų indeksų aibė aplink neuroną su indeksu w. Dvimačiai vektoriai R_w ir R_{ij} yra neuronų M_w ir M_{ij} indeksai. Indeksai parodo neurono vietą (eilutės ir stulpelio numerį) SOM žemėlapyje. Parametras η_{ij}^{w} yra neurono M_{ij} kaimynystės eilė numeris neurono-nugalėtojo M_w atžvilgiu.



2.4 pav. Kaimynystės eilė neurono-nugalėtojo atžvilgiu

Kaip buvo minėta prieš tai, mokymo parametras $\alpha(t)$ taip pat turi įtakos SOM tinklo rezultatams. Dažniausiai yra naudojami trys mokymo parametrų tipai:

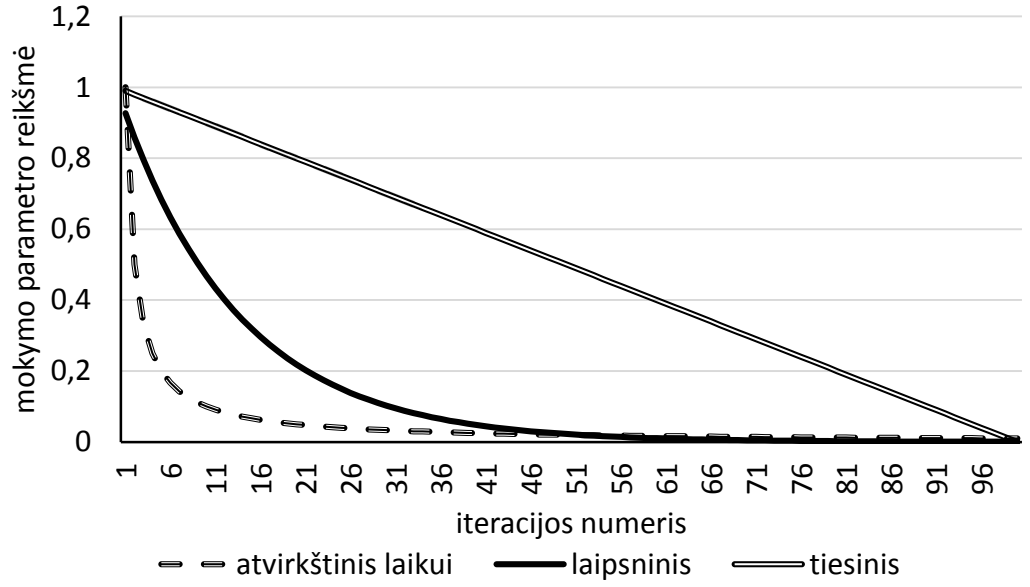
tiesinis: $\alpha(t) = \left(1 - \frac{t}{T}\right)$, (5)

atvirkštinis laikui: $\alpha(t) = \frac{1}{t}$, (6)

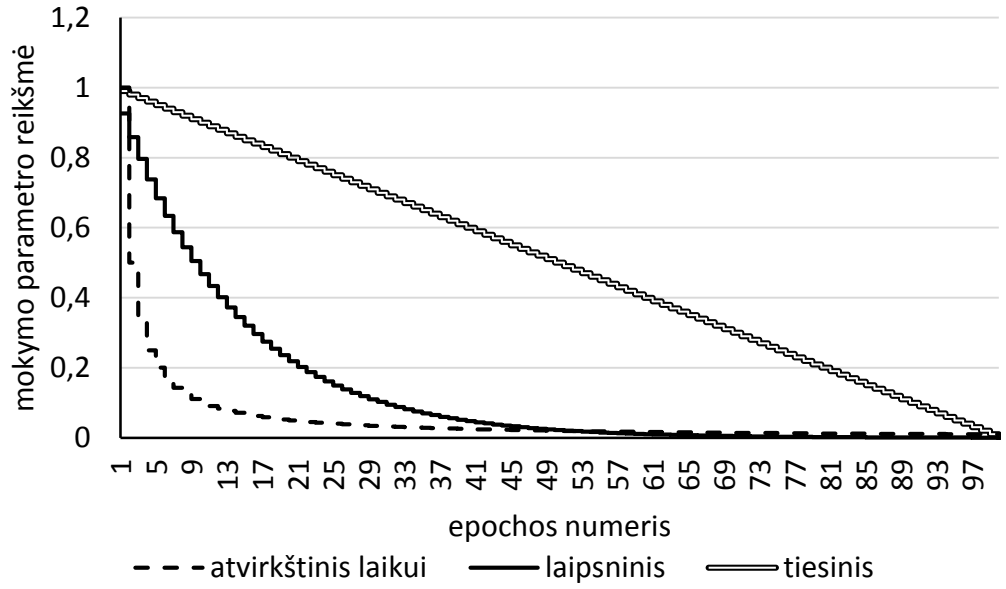
laipsninis: $\alpha(t) = (0{,}005)^{\frac{t}{T}}$. (7)

Šių mokymo parametrų reikšmės gali būti keičiamos arba kiekvienoje iteracijoje, arba kiekvienoje epochoje. Viena *iteracija* – tai mokymo dalis, kai SOM tinklui vienas mokymo aibės vektorius yra paduodamas į tinklą ir tuomet neuronai yra keičiami pagal mokymo taisyklę (2). Viena *epocha* – tai mokymo dalis, kai SOM tinklui visi mokymo aibės vektoriai paduodami į tinklą ir tuomet neuronai keičiami pagal mokymo taisyklę (2). Taigi, mokymo parametras $\alpha(t)$

gali priklausyti arba nuo iteracijų numerio (tuo atveju t – einamosios iteracijos skaičius, T – pasirinktas bendras iteracijų skaičius), arba nuo epochų numerio (tuo atveju t – einamosios epochos skaičius, T – pasirinktas bendras epochų skaičius).



a)



b)

2.5 pav. Mokymo parametrų reikšmės, a) keičiamos kiekvienoje iteracijoje, $T = 100$, b) keičiamos kiekvienoje epochoje, $T = 100$

2.5 pav. pateikta, kaip keičiasi mokymo parametrų reikšmės, jas keičiant kiekvienoje iteracijoje a) atveju, ir keičiant reikšmes kiekvienoje epochoje b) atveju, priklausomai nuo pasirinkto mokymo parametro tipo. 2.5 b) pav. vienoje epochoje mokymo parametro reikšmės išlieka nepakitusios visiems duomenų aibės vektoriams. Čia gaunamos „laiptuotos funkcijos“. Naudojant tiesinį mokymo parametrą, jo reikšmės mažėja ne taip greitai, kaip kitų parametrų atveju. Greičiausiai reikšmės mažėja, kai yra naudojamas atvirkštinis laikui mokymo parametras.

Be plačiai naudojamų burbuliuko ir Gauso kaimynystės funkcijų, galima rasti ir kitų euristinių kaimynystės funkcijų. Šioje disertacijoje buvo tiriama ir naudojama darbe (Dzemyda, 2001) pasiūlyta euristinė kaimynystės funkcija (8) ir euristinis mokymo parametras (9):

$$h_{ij}^{w}(t) = \frac{\alpha(t)}{\alpha(t)\cdot\eta_{ij}^{w}+1}, \qquad (8)$$

$$\alpha(t) = \max\left(\frac{T+1-t}{T}, \beta\right). \qquad (9)$$

Čia parametras η_{ij}^{w} yra neurono M_{ij} kaimynystės eilė neurono nugalėtojo M_w atžvilgiu, T – bendras mokymo žingsnių skaičius, t – einamojo žingsnio numeris, β – konstanta, kurios reikšmė įprastai lygi $0{,}01$. Neuronai M_{ij} yra perskaičiuojami tame mokymo žingsnyje, kuriame tenkinama nelygybė (10):

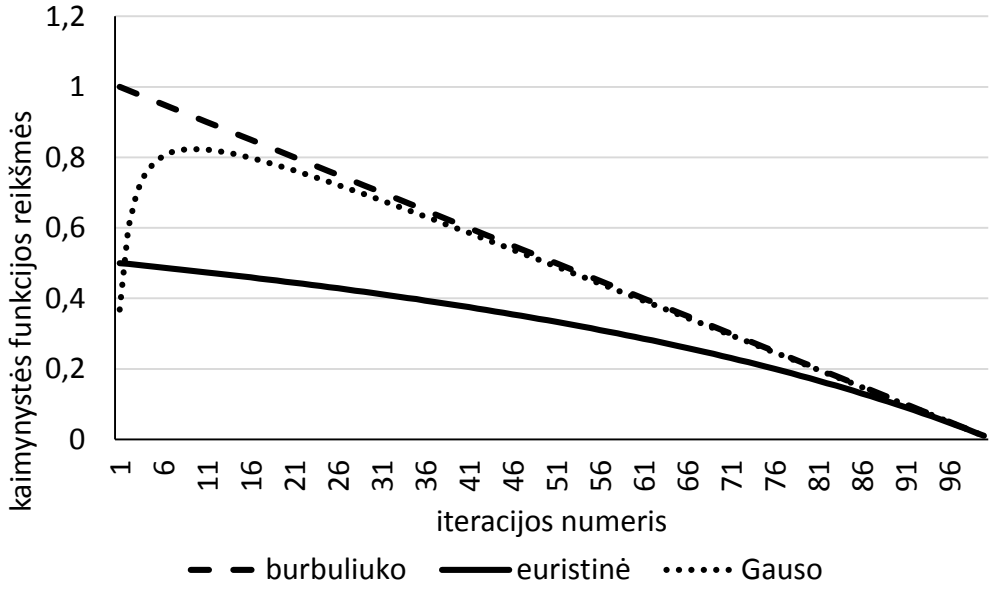
$$\eta_{ij}^{w} \le [\alpha \max(k_x, k_y), 1]. \qquad (10)$$

Darbe (Dzemyda, 2001) siūlyta kaimynystės funkcijos reikšmes keisti ne kiekvienoje iteracijoje, bet tik kiekvienoje epochoje, tačiau šioje disertacijoje nagrinėjami abu atvejai, siekiant atlikti išsamią mokymo faktorių analizę ir jų įtaką SOM rezultatams.

Ši euristinė kaimynystės funkcija ir mokymo parametras yra taip pat mažėjančios funkcijos. Mokymo parametras (9) yra tiesinis ir savo gaunamomis reikšmėmis panašus į tiesinį mokymo parametrą (5). Euristinio mokymo parametro atveju reikšmės mažėja tiesiškai, tačiau kai jos pasiekia tam tikrą pasirinktą konstantą (čia pasirinkta $\beta = 0{,}3$), reikšmė nebesikeičia (2.6 pav.).

2.6 pav. Tiesinio ir euristinio mokymo parametrų palyginimas, keičiant parametrų reikšmes kiekvienoje iteracijoje, $T = 1000$

Kaimynystės funkcijų Gauso (4), burbuliuko (3) ir euristinę (8) reikšmių kitimas pateiktas 2.7 pav. Čia pasirinktas tiesinis mokymo parametras (5), bei fiksuotos tokios parametrų reikšmės: $\eta_{ij}^{w} = t$ ($t = 1, \dots, 100$), $R_w = 3$, $R_{ij} = 1$.



2.7 pav. Burbuliuko, euristinės ir Gauso kaimynystės funkcijų palyginimas, keičiant mokymo parametrų reikšmes kiekvienoje iteracijoje

Kaip matome 2.7 pav., mokymo pabaigoje funkcijų reikšmės tampa lygios 0, o kitur didžiausios funkcijos reikšmės gaunamos burbuliuko kaimynystės

funkcijos atveju, o mažiausios, naudojant Gauso kaimynystės funkciją. Euristinės kaimynystės funkcijos reikšmės gaunamos didesnės nei burbuliuko, tačiau mažesnės, nei Gauso. Kaip matome iš pateiktų pavyzdžių tiek mokymo parametrai, tiek kaimynystės funkcijos SOM tinklo mokymo eigoje skiriasi, todėl tai gali įtakoti mokymo rezultatus.

2.2.3. Kvantavimo ir topografinės paklaidos

Kai tinklas išmokytas, būtina įvertinti jo kokybę. Tam dažniausia naudojamos dvi paklaidos: kvantavimo ir topografinė (Kohonen, 2001). *Kvantavimo paklaida* parodo, kaip tiksliai jau išmokyto tinklo neuronai prisiderina prie mokymo aibės vektorių. Kvantavimo paklaida E_{QE} – tai vidutinis atstumas tarp duomenų vektorių X_p ir jų vektorių nugalėtojų $M_{w(p)}$:

$$E_{\mathrm{QE}} = \frac{1}{m}\sum_{p=1}^{m}\left\|X_p - M_{w(p)}\right\|. \qquad (11)$$

Topografinė paklaida parodo, kaip gerai SOM tinklas išlaiko analizuojamų duomenų topografiją, t. y. tarpusavio išsidėstymą. Topografinė paklaida E_{TE} skaičiuojama pagal šią formulę:

$$E_{\mathrm{TE}} = \frac{1}{m}\sum_{p=1}^{m} u(X_p). \qquad (12)$$

Jeigu SOM žemėlapyje vektoriaus X_p neuronas-nugalėtojas yra šalia neurono, iki kurio atstumas nuo X_p yra mažiausias, neskaičiuojant iki neurono-nugalėtojo, tai (12) formulėje $u(X_p) = 0$, priešingu atveju $u(X_p) = 1$.

2.3. Įvairūs SOM praplėtimai ir modifikacijos

Nuo SOM sukūrimo praėjo daugiau nei 40 metų, todėl natūralu, jog atsirado įvairių jo praplėtimų bei modifikacijų, kurios apžvelgtos šiame poskyryje. Ilgą laiką buvo taikomi ir nagrinėjami rekurentinių saviorganizuojančių neuroninių tinklų klasės modeliai: laiko atžvilgio SOM žemėlapiai (angl. *temporal Kohonen map*) (Chappell, Taylor, 1993), rekurentinis saviorganizuojantis neuroninis tinklas (angl. *recurrent* SOM) (Koskela ir kiti, 1998a; 1998b), suliejamasis saviorganizuojantis neuroninis tinklas (angl. *merge* SOM) (Strickert, Hammer, 2004; 2005) ir rekursinis

saviorganizuojantis neuroninis tinklas (angl. *recursive* SOM) (Voegtlin, 2002). Šio tipo modeliai naudoja klasikinio SOM mokymo algoritmą, pritaikant įvairias modifikacijas.

Kai kurios SOM tinklų modifikacijos suteikia galimybę analizuoti struktūrinius duomenis: saviorganizuojantys neuroniniai tinklai struktūriniams duomenims (Hagenbuchner ir kiti, 2003) ir apibendrinti saviorganizuojantys neuroniniai tinklai struktūriniams duomenimis (Hammer ir kiti, 2004). Struktūriniai duomenys – tai duomenų tipas, susidedantis iš kitų elementų. Struktūriniai duomenys savyje gali laikyti metodus, konstantas, konstruktorius, kintamuosius, indeksus, operatorius ir panašiai.

Viena iš naujausių SOM modifikacijų yra grupinio mokymo (angl. *batch-learning*) saviorganizuojantys neuroniniai tinklai (BLSOM), kurie yra pritaikyti bioinformatikos sričiai (Iwasaki, 2013). Šiame metode saviorganizuojantis neuroninis tinklas yra specialiai modifikuotas taip, kad mokymo procesas ir rezultato atvaizdavimas būtų tinkamas genų analizei. BLSOM yra tinkamas metodas didelės apimties duomenims analizuoti, kuris leidžia klasifikuoti ir vizualizuoti didelių sekų aibes, gaunama iš milijono genų sekų.

Kitas naujas SOM praplėtimas yra aplinkos įtakos saviorganizuojantiems neuroniniams tinklams modelis (angl. *enviromental* SOM, EnvSOM) (Alonso ir kiti, 2011), pagrįstas aplinkos veiksnių įtaka SOM mokymui. Pirmajame šio algoritmo mokymo etape SOM mokomas klasikiniu algoritmu, naudojant visus duomenis. Šiame algoritme reikia žinoti, kurie duomenų požymiai yra aplinkos (angl. *enviromental*), kadangi būtent jie bus naudojami ieškant neurono-nugalėtojo. Kiti požymiai turi būti „paslėpti“ prieš ieškant neurono-nugalėtojo. Skirtumas nuo klasikinio SOM tinklo yra tas, jog čia naudojama dvejetainė žymė, nurodanti, kurie požymiai buvo naudojami ieškant neurono-nugalėtojo. Pirmojo etapo rezultatas bus žemėlapis, kuriame atvaizduoti tik aplinkos požymiai. Likusieji mokyme nedalyvavo, todėl ir nėra atvaizduoti. Šio etapo pagrindinis tikslas – sukurti modelį, kuris geriausiai atvaizduoja aplinkos požymius. Antrajame EnvSOM mokymo etape naudojamas naujas SOM mokymo algoritmas. Mokymui bus panaudoti pirmajame etape gauti duomenys.

Šiuo SOM inicijavimo būdu, gaunamas greitas algoritmo konvergavimas. Kadangi SOM tinklas apmokytas pirmoje fazėje naudojant tik duomenų aplinkos požymius, reikia tik tinkamai atnaujinti likusius požymius. Šios fazės tikslas yra gauti išmokytą SOM tinklą visai duomenų aibei, atsižvelgiant į aplinkos požymius.

Taip pat sukurta įvairių SOM praplėtimų, kurie turi savo savitą vizualizavimo būdą, nei yra įprasta klasikiniam saviorganizuojančiam neuroniniam tinklui. Pavyzdžiui, indukuoto vizualizavimo saviorganizuojantys neuroninis tinklas ViSOM (Hujun, 2002a, 2002b). Šis vizualizavimo praplėtimas leidžia pamatyti atstumus tarp visų atvaizduotų duomenų vektorių. Pagrindinis privalumas, jog atstumai žemėlapyje atitinka atstumus tarp įėjimo vektorių. Šiame metode naudojama dvimatė arba trimatė Dekarto koordinačių sistema. Dirbant su mažų matmenų žemėlapiais, ViSOM žemėlapiai tampa neaiškūs ir sunku įvertinti atstumus tarp klasių, todėl patartina ViSOM metode naudoti didesnius žemėlapius nei užtenka paprastam SOM tinklui. Kadangi vis dažniau kyla poreikis analizuoti didelės apimties duomenų aibes, todėl yra kuriami SOM praplėtimai, leidžiantys aiškiau atvaizduoti ir didelės apimties duomenų aibes (Prakash, 2013). Šio praplėtimo privalumas yra tas, jog vienu metu galima matyti kelias dešimtis požymių rezultatų. SOM žemėlapis pateikimas grafo pavidalu, kur kiekvienas grafas atitinka analizuojamą požymį.

Yra daugybė kitų SOM praplėtimų ir modifikacijų, kurios gali būti taikomos specifiniams uždaviniams spręsti. Teksto informacijai klasifikuoti yra pasiūlytas WEBSOM (Kaski ir kiti, 1998). Saviorganizuojantys neuroniniai tinklai taikomi ir socialinių tinklų analizei, pavyzdžiui, SOMSN (Ghaemmaghami, Manouchehri, 2013). Kiti SOM praplėtimai: SOM skirtas paveikslų nagrinėjimui (Laaksonen ir kiti, 2000), hierarchiškai augantys SOM tinklai (Rauber ir kiti, 2000; 2002) ir daugelis kitų (Ontrup, Ritter, 2001).

Daug metų saviorganizuojantys neuroniniai tinklai dažniausiai buvo taikomi tik įvairiems skaitiniams duomenims klasifikuoti ir klasterizuoti, tačiau šiuo metu atsiranda daug tyrimų, kuriuose nagrinėjami tekstiniai (Kohonen, Xing, 2011) arba struktūriniai duomenys. Gausėjant įvairialypiai informacijai,

saviorganizuojantys neuroniniai tinklai pradėti taikyti grafinei (Sjoberg, Laaksonen, 2011), vaizdo (Maia ir kiti, 2011) (Kanimozhi, Bindu, 2013) (Jordan, Angelopoulou, 2013) (Mahalakshmi ir kiti, 2014), garso (Mayer, 2011) (Cao ir kiti, 2013), binariniams duomenims (Almendra, Enachescu, 2014) ir kitokio pobūdžio informacijai apdoroti.

Internetinėje erdvėje gausu įvairaus pobūdžio tekstinės informacijos: dokumentai, tekstas svetainėse, moksliniai straipsniai, todėl natūralu, jog atsiranda poreikis šią informaciją apdoroti, siekiant neišnaudojant daug laiko, gauti kuo daugiau reikalingos informacijos. Naudodamiesi internetu galime greitai surasti vieną ar kitą dalyką, tačiau informacija dažnai būna nenaudinga, iškraipyta ar neesminė. Todėl vis dažniau atsiranda įvairiausių duomenų tyrybos metodų jai analizuoti ir susisteminti (Marinai, 2011), (Alsmadi, Saleh, 2012) (Sihag, Kumar, 2013) (Dobnikar ir kiti, 2011). Saviorganizuojantys neuroniniai tinklai taip pat sėkmingai taikomi įvairiems uždaviniams, susijusiems su teksto analize, spręsti: dokumentų plagijavimui tikrinti (Chow, Rahman, 2009), internetinėse žinių svetainėse pateiktiems straipsniams analizuoti (Mayer, Rauber, 2011) ir kt.

Dažnai saviorganizuojantys neuroniniai tinklai yra jungiami su kitais žinomais metodais, pavyzdžiui su k-vidurkių metodu, siekiant pagerinti gaunamų rezultatų kokybę (Ding ir kiti, 2013) (Dogan ir kiti ,2013) (Deligiorgi ir kiti, 2014) (Gorgonio, Costa, 2010).

Saviorganizuojančių neuroninių tinklų tematika yra nagrinėjama ir Lietuvoje. Per pastarąjį dešimtmetį sėkmingai apgintos kelios disertacijos, kuriose yra panaudoti, ar tirti saviorganizuojantys neuroniniai tinklai. O. Kurasovos (2005) disertacijoje nagrinėja saviorganizuojančių neuroninių tinklų ir daugiamačių skalių jungimo būdus. E. Merkevičiaus (2008) disertacijoje saviorganizuojantys neuroniniai tinklai taikyti kredito rizikos vertinimo sprendimų paramos sistemoje, kreditams klasifikuoti. Taip pat E. Merkevičiaus publikacijose (Merkevičius ir kiti, 2007) pasiūlytas modelis, pagrįstas SOM tinklais, kuris leidžia prognozuoti įmonių skolininkų finansinės būklės tendencijas, finansinius pokyčius. V. Marcinkevičiaus (2010) disertacijos

tyrimo objektas buvo daugiamačiai duomenys, jų atvaizdavimas netiesiniais daugiamačių skalių algoritmais ir saviorganizuojančiais neuroniniais tinklais, projekcijos kokybės vertinimas. A. Molytė (2011) disertacijoje nagrinėjo dirbtiniais neuroniniais tinklais grindžiamus vektorių kvantavimo metodus, taip pat ir saviorganizuojančius neuroninius tinklus ir daugiamačių duomenų vizualizavimo metodus, pagrįstus dimensijų skaičiaus mažinimu. R. Danilienė (2010) ir J. Pragarauskatė (2013) disertacijose taikė saviorganizuojančius neuroninius tinklus eksperimentinių bandymų rezultatams vizualizuoti. Detali vizualizavimo metodų apžvalga pateikta vadovėlyje (Dzemyda ir kiti, 2008), skirtame informatikos krypties doktorantams ir magistrantams, kuriame taip pat yra aprašyti ir saviorganizuojantys neuroniniai tinklai bei jų taikymas.

2.4. Saviorganizuojančių neuroninių tinklų vizualizavimas

Kai tinklas yra išmokytas, visi įėjimo vektoriai yra dar kartą paduodami į tinklą, taip randami neuronai-nugalėtojai. Kiekvienas neuronas-nugalėtojas atitinka vieną ar daugiau įėjimo vektorių, kurie yra atvaizduojami žemėlapyje. Pats paprasčiausias SOM vizualizavimo būdas – paprasta lentelė, kurioje yra atvaizduojami duomenų numeriai arba klasės žymės. 2.1 lentelėje irisų duomenys (duomenų aprašas pateiktas 4.1 poskyryje) atvaizduoti SOM tinkle.

2.1 lentelė. Paprasta SOM lentelė (atvaizduoti irisų duomenys)

2,3		2	2		2,3	2
2,3		2		2	2	2
3	2,3	2		2	2	2
3			2			
3		2	2			
3	3	2				1
3	2	2			1	1

Skaičiai lentelėje žymi duomenų klasių numerius: 1 – *iris setosa*, 2 – *iris versicolor* ir 3 – *iris virginica*. Jeigu keli tos pačios klasės duomenys patenka į

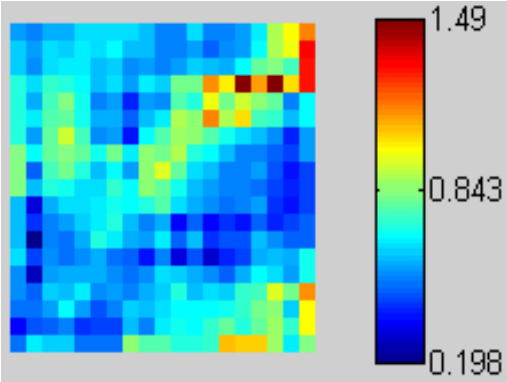
tą patį SOM tinklo langelį, klasės numeris užrašomas tik vieną kartą. Pavyzdžiui, į apatinio dešiniojo kampo tris langelius pateko visi 50 pirmos klasės duomenys. Jeigu į tą patį SOM tinklo langelį patenka skirtingų klasių nariai, nurodomi visų skirtingų klasių pavadinimai, tačiau lieka neaišku, kiek į tą patį langelį pakliuvo kiekvienos skirtingos klasės duomenų.

Gautoje SOM lentelėje (žemėlapyje) matome du klasterius, tarpusavyje atskirtus tuščiais SOM langeliais: pirmos klasės narių klasteris (dešiniajame apatiniame kampe) ir priešingoje pusėje – II ir III klasės klasteris. Turint tik tokią lentelę, sunku pasakyti, kaip arti n-matėje erdvėje yra duomenys, esantys šalia SOM lentelėje. Todėl svarbu rasti būdus, kurie pateiktų rezultatus kitokia forma ir suteiktų daugiau informacijos apie analizuojamus duomenis.

Vienas iš populiariausių SOM vizualizavimo būdų yra taip vadinamoji *unifikuota atstumų matrica* (angl. *u-matrix*) (Ulsch, Seemon, 1989). Ji parodo ryšius tarp kaimyninių neuronų:

$$u\text{-matrica} = \begin{pmatrix} u_{11} & u_{11|12} & u_{12} & u_{12|13} & \cdots & u_{1k_y} \\ u_{11|12} & & u_{12|22} & & \cdots & u_{1k_y||2k_y} \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ u_{k_x1} & u_{k_x1|k_x2} & u_{k_x2} & u_{k_x2|k_x3} & \cdots & u_{k_xk_y} \end{pmatrix}. \quad (13)$$

Čia u_{ij} atitinka SOM lentelės langelius, o $u_{ij|i(j+1)}$ – langelių rėmelius. $u_{ij|i(j+1)}$ ($u_{ij|(i+1)j}$) yra Euklidinis atstumas tarp kaimyninių neuronų M_{ij} ir $M_{i(j+1)}$ (M_{ij} ir $M_{(i+1)j}$). Reikšmės u_{ij} gali būti unifikuotos atstumų matricos kaimynių elementų vidurkis. Pavyzdžiui, jeigu u_{ij} turi keturis kaimynus, tuomet $u_{ij} = \left(u_{i(j-1)|ij} + u_{ij|i(j+1)} + u_{(i-1)j|ij} + u_{(i+1)j|i(j+1)}\right)/4$.
Apskaičiavus unifikuotos atstumo matricos reikšmes SOM lentelė gali būti nuspalvinama, atsižvelgiant į šias reikšmes. Naudojant pilkų atspalvių skalę, tamsesnis atspalvis reiškia didesnį atstumą, šviesesnis – mažesnį atstumą tarp kaimyninių neuronų. Gali būti naudojami ne tik pilki atspalviai, bet ir daugiau spalvų (2.8 pav.). SOM žemėlapio vizualizavimas pagal unifikuotą atstumų matricą yra įgyvendintas daugelyje SOM sistemų (detaliau apie tai 2.5 poskyryje).



2.8 pav. SOM vizualizuotas pagal unifikuotą atstumų matricą, gautą SOM-Toolbox sistema

2.5. Saviorganizuojančių neuroninių tinklų programinės sistemos

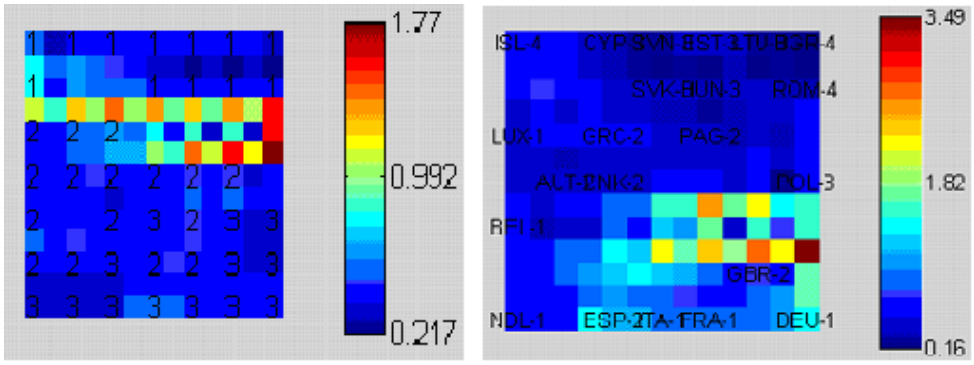
Šiuo metu yra sukurta įvairių programinių sistemų, kuriose įgyvendinti saviorganizuojantys neuroniniai tinklai. Kiekviena iš jų turi savitą naudotojo sąsają bei įvairių vizualizavimo būdų (Dzemyda, Kurasova, 2002) (Moehrmann ir kiti, 2011). Toliau apžvelgtos populiariausios sistemos. Du duomenų rinkiniai – irisai ir ekonominiai duomenys (apie juos detaliau 4.1 poskyryje) – atvaizduojami SOM žemėlapiuose siekiant pademonstruoti sistemose įgyvendintus vizualizavimo būdus. Irisų duomenys sudaro 150 vektorių, kurie suskirstyti į tris klases: I klasė – *iris setosa*, II klasė – *iris versicolar* ir III klasė – *iris virginica*. Ekonominius duomenis sudaro 31 vektorius, kurie suskirstyti į keturias klases: I klasė – šalys, kurios įkūrė Europos Sąjungą (ES), II klasė – šalys, kurios prisijungė prie ES 1957–1995 m., III klasė – šalys, kurios įstojo į ES 2004–2007 m., ir IV klasė – šalys, kurios siekia ES narystės.

2.5.1. Vizualizavimas SOM-Toolbox sistemoje

SOM-Toolbox sistema yra sukurta Matlab programavimo kalba, todėl jos naudojimui būtina Matlab aplinka. Šios sistemos kūrėjai su SOM tinklų pradininku T. Kohonenu – dirbančių mokslininkų grupė (Vesanto ir kiti, 2005), todėl SOM-Toolbox iki šiol yra populiariausia SOM sistema. Sistemoje yra įvairių funkcijų, leidžiančių pritaikyti SOM žemėlapio atvaizdavimą pagal savo poreikius bei pasirinkti įvarius išplėstinius parametrus (kaimynystės funkcijas, mokymo parametrus, mokymo žingsnį, skirtingus vizualizavimo būdus ir kt.).

Gautame SOM žemėlapyje langeliai yra atskiriami ne kraštinėmis aplink juos, o papildomais to paties dydžio langeliais. Tad, jeigu pasirinktas žemėlapio dydis 7×7, rezultate bus gautas žemėlapis, kurio dydis yra 13×13. SOM-Toolbox sistemoje žemėlapiai gali būti nuspalvinami pagal unifikuotos atstumų matricos reikšmes. Leidžiama pasirinkti skirtingas unifikuotos atstumų matricos reikšmių spalvų skales, kurios yra pateikiamos gauto žemėlapio dešinėje pusėje.

Išmokius SOM tinklą irisų duomenimis, matome (2.9a pav.), jog I klasės nariai yra atskirti nuo II ir III klasės narių ryškesnėmis spalvomis. Spalvų skalė parodo, kad šie duomenys yra skirtingi (atstumas tarp jų yra didelis). I klasės nariai tarpusavyje išsidėstę mėlynos spalvos langeliuose, vadinasi, šie duomenys yra tarpusavyje panašūs (atstumas tarp jų mažas).

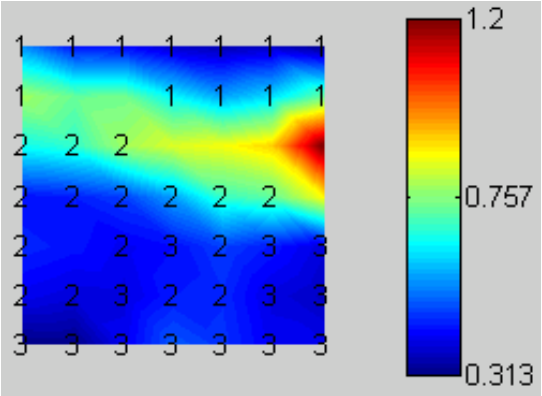


a) **b)**

2.9 pav. 7×7 SOM žemėlapiai, gauti SOM-Toolbox sistema: a) irisų duomenys, b) ekonominiai duomenys

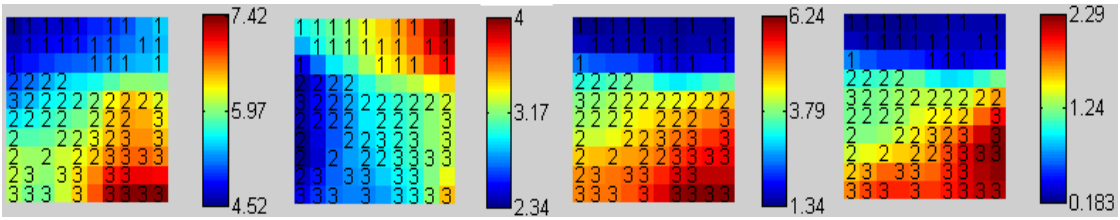
Ekonominių duomenų atveju (2.9b pav.) žemėlapio langeliuose pateikiami vektorius atitinkantys šalių sutrumpinimai ir jų klasės numeriai. SOM žemėlapio apačioje atsiduria I klasės nariai, kurie nuo kitų klasių yra atskiriami ryškesnių spalvų langeliais – suformuojamas atskiras klasteris. Tai parodo, kad I klasei priskirtos šalys (ES įkūrėjos) pagal tiriamus požymius atsiskiria nuo kitų šalių.

2.10 pav. yra pateiktas dar vienas SOM atvaizdavimo būdas, kurį leidžia gauti SOM-Toolbox sistema. Šiuo atveju yra spalvinami ne konkretūs SOM langeliai, o žemėlapio sritys.



2.10 pav. Irisų duomenys 7×7 SOM žemėlapyje, gautame SOM-Toolbox sistema

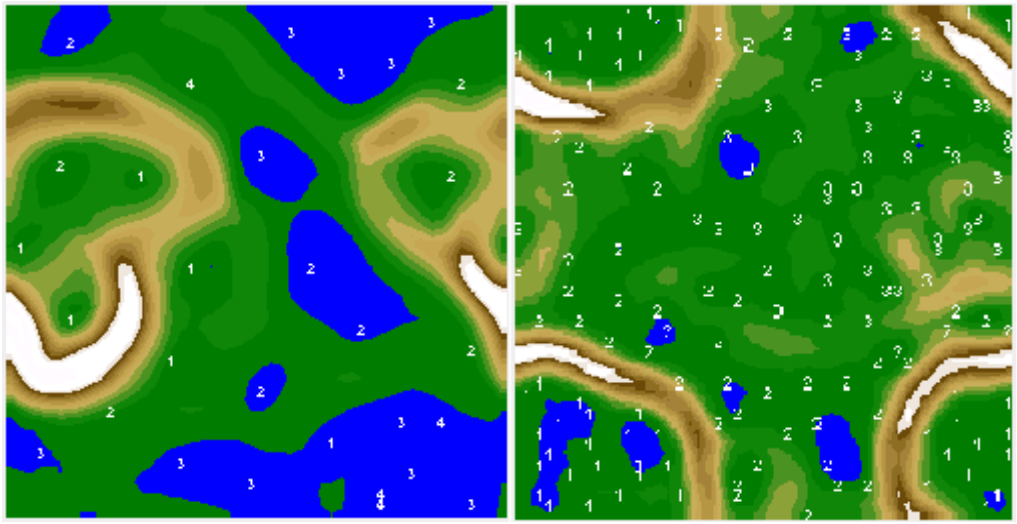
SOM-Toolbox sistemoje yra galimybė peržiūrėti atskirai atvaizduotas vektorių komponentes (angl. *component planes*). Irisų duomenų atveju (2.11 pav.) gaunamos keturios u-matricos, kadangi šios duomenų aibės vektoriai sudaryti iš keturių požymių. Pirma u-matrica atvaizduoja taurėlapio ilgio komponentę, antra – taurėlapio plotį, trečia – vainiklapio ilgį ir ketvirta – vainiklapio plotį. Šis vizualizavimo pasirinkimas leidžia nustatyti, kuri komponentė daro didžiausią įtaką galutiniam rezultatui.



2.11 pav. Irisų duomenys 7×7 SOM žemėlapyje (atskirų komponenčių atvaizdavimas), gautame SOM-Toolbox sistema

2.5.2. Vizualizavimas Databionic ESOM sistemoje

Dauguma saviorganizuojančių neuroninių tinklų sistemų turi trūkumą, jog nėra specialiai pritaikytos didelės apimties duomenims vizualizuoti. Databionic ESOM sistema naudoja itin didelius žemėlapius ir leidžia atvaizduoti dideles duomenų aibes (Ultsch, Moerchen, 2005). Šioje sistemoje galimi įvairūs vizualizavimo būdai: p-matrica (Ultsch, 2003), u-matrica, komponenčių atvaizdavimas, duomenų histogramos ir kt.

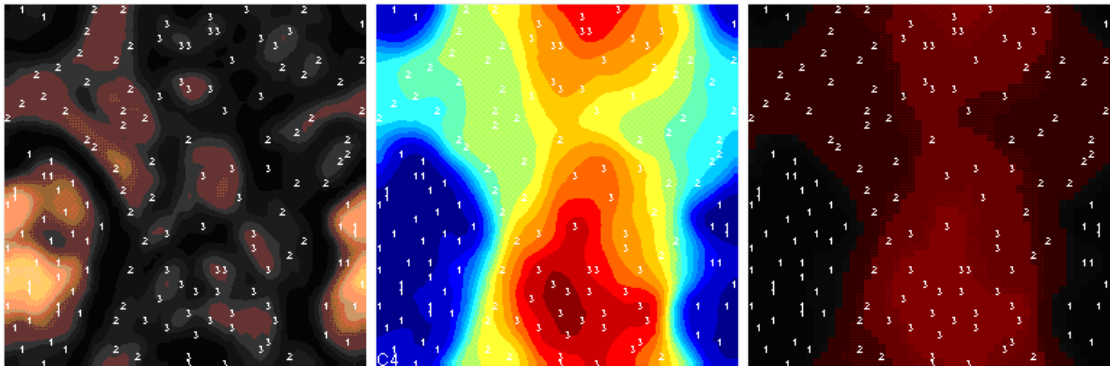


a) **b)**

2.12 pav. 50×50 SOM žemėlapiai, gauti Databionic ESOM sistema: a) irisų duomenys, b) ekonominiai duomenys

Apmokius Databionic ESOM irisų duomenimis (2.12a pav.), matome jog gauti panašūs klasteriai kaip ir SOM-Toolbox sistema, tačiau atvaizdavimo būdas čia visai kitoks. Sistemoje nėra langelių ir jų kraštinių, o duomenys yra atvaizduojami reljefo žymėjimo pavidalu.

Šioje sistemoje klasių pavadinimai būtinai turi būti rašomi skaitine išraiška, todėl, nagrinėjant ekonominių duomenų aibę, pateikiamas tik klasės numeris (be šalies sutrumpinimo). Kaip matome (2.12b pav.), ekonominių duomenų atveju I klasės nariai išsidėstę žemėlapio kampuose ir yra atskirti rudos spalvos lankais nuo II, III ir IV klasės narių. Kai kurie I klasės nariai atsiduria mėlynos spalvos srityse, tad būtent šie duomenys tarpusavyje yra daug panašesni nei kiti. Rudos sritys lyg kalnai skiria skirtingus klasterius, o mėlynos sritys (vanduo) parodo, jog vektoriai yra arti vienas kito, duomenys panašūs. Sistema leidžia pasirinkti ir kitus atvaizdavimo būdus, kurių rezultatai pateikti 2.13 pav.



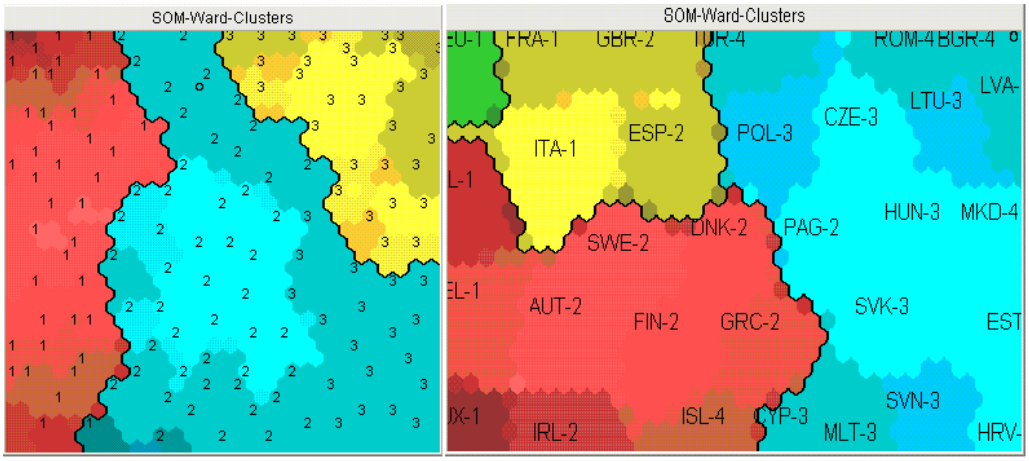
2.13 pav. Irisų duomenys, atvaizduoti Databionic ESOM sistemoje (įvairūs vizualizavimo būdai)

2.5.3. Vizualizavimas Viscovery SOMine sistemoje

Viscovery SOMine yra komercinė programa, kuri nuolat tobulinama, atsižvelgiant į vartotojų poreikius (Viscovery SOMine 6.0, 2014). Pagrindinis skirtumas nuo prieš tai aptartų sistemų – didelis sistemos funkcionalumas, tačiau jis dažnai apsunkina sistemos naudojimą. Tenka atlikti daug įvairių papildomų nustatymų. Išmokius SOM tinklą, sistema automatiškai parenka klasterius, kurie gali sutapti su analizuojamų duomenų klasių skaičiumi, tačiau yra galimybė ir pačiam tyrėjui pasirinkti norimą klasterių skaičių. Po saviorganizuojančio neuroninio tinklo apmokymo irisų duomenimis gautame žemėlapyje (2.14a pav.) matome, jog klasteriai pavaizduojami skirtingomis spalvomis. Raudonos spalvos srityje matome tik I klasės narius, geltonos – vien III klasės narius, o žydroje – II klasės narius ir dalį III klasės narių. Tai rodo, jog II ir III klasės vektoriai yra tarpusavyje panašūs pagal tam tikrus požymius. Viscovery SOMine sistema pati automatiškai nustato klasterių skaičių, tačiau, esant reikalui, yra galimybė patiems pasirinkti norimą klasterių skaičių. Čia taip pat yra atvaizduotos ribos tarp susidariusių klasterių.

Apmokius saviorganizuojantį neuroninį tinklą ekonominiais duomenimis (2.14b pav.), sistema automatiškai parinko septynis klasterius, tačiau kadangi šie duomenys suskirstyti į keturias klases, sistemai nurodyti 4 klasteriai. Tuomet dauguma III ir IV klasės narių atsiduria šviesiai mėlynos spalvos klasteryje, I ir II klasės nariai atitinkamai išsidėsto geltonos, raudonos ir žalios spalvos

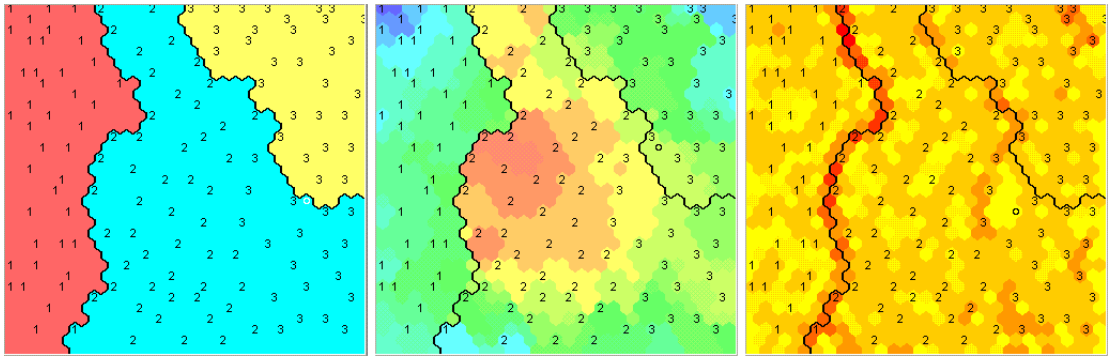
klasteriuose. Matome, jog šioje sistemoje ekonominiai duomenys sudarė ne tokius aiškius klasterius, skirtingai nei prieš tai aptartose sistemose.



a) **b)**

2.14 pav. SOM žemėlapiai, gauti Viscovery SOMine sistema: a) irisų duomenys, b) ekonominiai duomenys

Viscovery SOMine sistemoje leidžiama pasirinktinai naudoti ir kitus vizualizavimo būdus: šešėlinis klasterizavimas (angl. *shaded clustering*), lygusis klasterizavimas (angl. *flat clustering*), globalus šešėlių nustatymas (angl. *global shading*) ir unifikuota atstumo matrica (2.15 pav.).

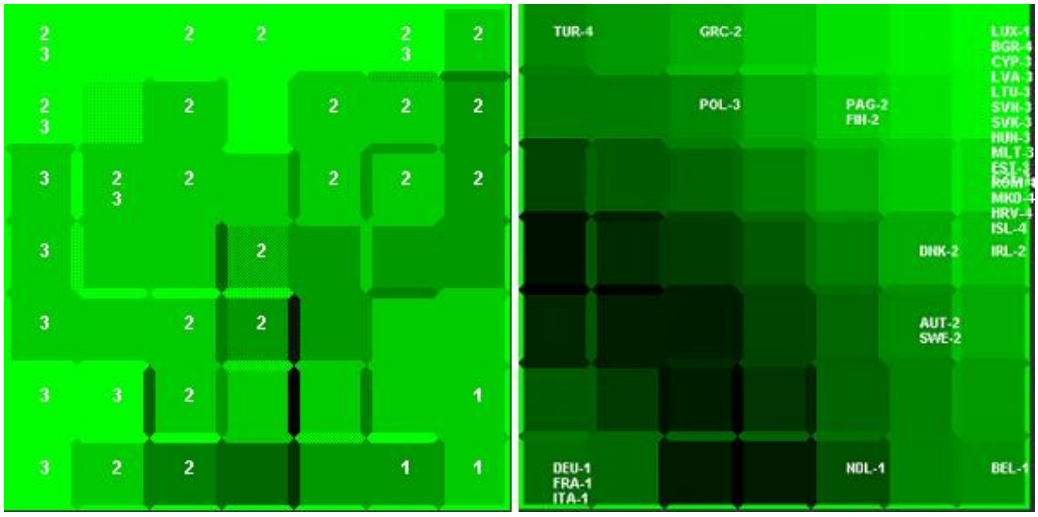


a) **b)** **c)**

2.15 pav. SOM žemėlapiai, gauti Viscovery SOMine sistema: a) lygusis klasterizavimas, b) globalus šešėlių nustatymas, c) u-matrica

2.5.4. Vizualizavimas NeNet sistemoje

Sistemoje NeNet vizualizavimo pagrindas yra unifikuotų atstumo matrica (Hassinen, 1999). Atvaizduotame SOM žemėlapyje atspalviai atitinka unifikuotos matricos reikšmes, kur šviesios langelių ir jų kraštinių spalvos žymi, jog gretimuose langeliuose esantys neuronus atitinkantys vektoriai yra arti vienas nuo kito n-matėje erdvėje, tamsios – vektoriai toli vienas nuo kito.

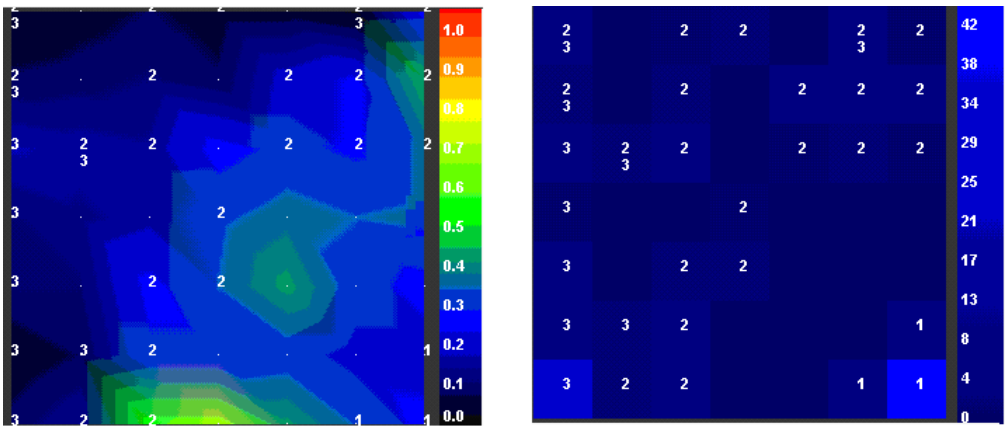


a) **b)**

2.16 pav. 7×7 SOM žemėlapiai, gauti NeNet sistema: a) irisų duomenys, b) ekonominiai duomenys

Išmokius saviorganizuojantį neuroninį tinklą irisų duomenimis, matome, jog susidaro du atskiri klasteriai (2.16a pav.). Pirmos klasės duomenis nuo II ir III klasės duomenų atskiria tamsesnės spalvos langeliai. Pastebime, jog kai kuriuose langeliuose atsiduria kelių klasių duomenys. Išmokius SOM tinklą ekonominiais duomenimis (2.16b pav.), matome, jog dešiniajame viršutiniame kampe atsiduria daug II, III ir IV klasės narių. Jie vėl tamsiais langeliais atskiria nuo I klasės narių.

Sistemoje NeNet galime pasirinkti dar du kitus SOM atvaizdavimo būdus (2.17 pav.). Kairiajame SOM žemėlapyje klasių numeriai yra atvaizduoti ne langelių viduje, o jų susikirtimo taškuose. Taip pat yra pateiktos unifikuoto atstumo matricos reikšmės ir spalvų paletė.

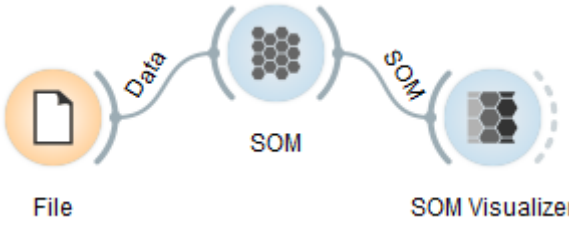


2.17 pav. Irisų duomenys, atvaizduoti NeNet sistemoje (įvairūs vizualizavimo būdai)

2.5.5. Kitos SOM sistemos

Taip pat yra sukurta įvairių kitų sistemų, kuriuos nėra plačiai paplitusios arba nėra specialiai sukurtos SOM tinklams nagrinėti ir taikyti, tačiau turi galimybę panaudoti SOM tinklus.

Orange yra nuolat tobulinama atvirojo kodo programa (Demšar ir kiti, 2013), kurioje įgyvendinti įvairūs duomenų tyrybos metodai. Joje yra galimybė taikyti saviorganizuojančius neuroninius tinklus duomenims analizuoti. Pagrindinis skirtumas nuo prieš tai apžvelgtų sistemų yra tas, jog čia duomenų įkėlimas, apmokimas ir vizualizavimas įgyvendinamas atskirais valdikliais (angl. *widget*), sudarant taip vadinamas darbų sekas (angl. *workflow*). Kiekvienas valdiklis atlieka tam tikrą funkciją. Tam, kad galėtume atvaizduoti duomenis, saviorganizuojančių neuroninių tinklų metodu užtenka įkelti tris valdiklius (duomenų failo, SOM algoritmo ir SOM vizualizavimo) bei nurodyti atitinkamus mokymo faktorius (2.18 pav.).



2.18 pav. Saviorganizuojančių neuroninių tinklų darbų seka Orange sistemoje

Nesenai sukurta TaxSOM sistema pasižymi tuo (Weber ir kt., 2010), kad yra prieinama vartotojams per interneto naršyklę. Sistemoje įgyvendinti dviejų tipų saviorganizuojantys neuroniniai tinklai: grupinio mokymo (angl. *batch-learning* SOM) ir augantys (angl. *growing SOM*) saviorganizuojantys neuroniniai tinklai. SOM mokymas vyksta keliais etapais: 1) įkeliamas duomenų failas, kurio dydis negali viršyti 50 megabaitų; 2) pasirenkamas vienas iš dviejų SOM tipų; 3) pasirinktais duomenimis mokomas SOM tinklas; 4) vizualiai pateikiamas gautas SOM žemėlapis bei skaitinės paklaidos, kurias galima atsisiųsti į savo kompiuterį.

Ilgą laiką daug mokslininkų kurdavo papildomus įrankius Matlab aplinkai, taip atsirado ir sistema SOM-Toolbox. Šiuo metu vis dažniau galima sutikti įvairių papildomų įrankių ir R paketui. Tai paketas, kuris leidžia laisvai kurti, dalintis savo sukurtais programiniais kodais. Pačioje sistemoje R yra siūlomi paketai (SOM, Kohonen), kurie skirti saviorganizuojantiems neuroniniams tinklams (Wehrens, Buydens, 2007), tačiau taip pat kuriami ir nauji specializuoti SOM paketai SOMbrero (Villa-Vialaneix ir kiti, 2013), leidžiantys analizuojant duomenis pasirinkti daugiau įvairių mokymo faktorių. Neskaitant paketo SOM-Toolbox, pati sistema Matlab taip turi neuroninių tinklų paketą „nnet“, kuriame dviem įrankiais (nntool, nctool) galima atvaizduoti paprasčiausius SOM žemėlapius. Pakete „nnet“ leidžiama keisti tik kelis mokymo nustatymus, todėl ši sistema nėra plačiai naudojama.

Saviorganizuojantys neuroniniai tinklai realizuoti ir kitose duomenų tyrybos sistemose: Weka (Witten ir kiti, 2011), RapidMiner (Hofmann, Klinkenberg, 2013), SOM-analyzer (SOM-analyzer, 2014), Java SOMToolbox (Mayer, Rauber, 2011), SOMVis (SOMVis, 2014) ir kiti. Tačiau dauguma išvardintų sistemų leidžia naudoti tik pagrindinius mokymo faktorius, sistemose nėra galimybės pamatyti į tą patį langelį patekusių visų duomenų aibės vektorių skaičiaus.

2.6. Antro skyriaus apibendrinimas

Šiame skyriuje yra atlikta saviorganizuojančių neuroninių tinklų praplėtimų ir modifikacijų apžvalga. Aprašytos dažniausiai sutinkamos SOM

modifikacijos, jų panaudojimas ir skirtumai, lyginant su klasikiniu SOM mokymo algoritmu. Saviorganizuojantis neuroninis tinklas gali apdoroti tik skaitinę informaciją, todėl pateiktas būdas, kaip galima konvertuoti tekstinę informaciją į skaitinę. Apžvelgti teksto konvertavimo į skaitinę išraišką faktoriai: žodžio ilgis, žodžio pasikartojimas, skaitmenų atmetimo reikšmė, kamieno išskyrimo algoritmas bei dažniausiai vartojamų žodžių sąrašas.

Taip pat apžvelgti pagrindiniai SOM mokymo taisyklės faktoriai: kaimynystės funkcijos (burbuliuko, Gauso) ir mokymo parametrai (tiesinis, atvirkštinis laikui, laipsninis). Įprastai naudojamos dvi kaimynystės funkcijos, burbuliuko ir Gauso, tačiau darbe papildomai naudota ir tirta euristinė kaimynystės funkcija ir euristinis mokymo parametras. Taip pat ištirta, kaip kinta mokymo parametro reikšmė, keičiant ją kiekvienoje iteracijoje arba epochoje. Išmokytą saviorganizuojantį neuroninį tinklą įprastai įvertina dvi paklaidos: kvantavimo ir topografinės paklaidos, kurios šiame skyriuje taip pat aprašytos.

Atlikta saviorganizuojančių neuroninių tinklų sistemų analitinė apžvalga bei pateikti sistemų (SOM-Toolbox, Databionic ESOM, Viscovery SOMine, NeNet) vizualizavimo būdai, nagrinėjant irisų ir ekonominius duomenis. Apžvalgoje matomi esamų sistemų mokymo, vizualizavimo privalumai bei trūkumai.

Esminis visų apžvelgtų SOM sistemų vizualizavimo būdų trūkumas yra tas, jog nėra galimybės pamatyti visų duomenų vektorių, patekusių į tą patį SOM žemėlapio langelį, skaičiaus, o tai yra svarbu, kadangi, atvaizduojant tik po vieną vienos klasės narį, nėra aišku, kiek skirtingų klasės narių yra tame langelyje. Taip pat, analizuojant gautus įvairių SOM sistemų žemėlapius, sunku pasakyti, kuriame žemėlapyje klasteriai yra „stipresni“, o kuriame – ne. Todėl būtina pasiūlyti naujas paklaidas, leidžiančias įvertinti gauto SOM žemėlapio kokybę. Taip pat yra svarbu įvertinti SOM žemėlapyje gautų klasterių sutapimą su duomenų klasėmis. Tam būtina sukurti naujas SOM kokybę įvertinančias paklaidas.

3. Naujas SOM vizualizavimo būdas bei jo kokybę įvertinančios paklaidos

Šiame skyriuje pasiūlytas saviorganizuojančio neuroninio tinklo vizualizavimo būdas, leisiantis pamatyti skirtingų klasių duomenų, pakliuvusių į tą patį SOM langelį, santykį. Taip pat pasiūlytos paklaidos, leisiančios įvertinti SOM gautų klasterių ir duomenų klasių panašumus. Be to, aprašyta sukurta SOM sistema, kurioje įgyvendintas pasiūlytas vizualizavimo būdas bei SOM kokybę įvertinančios paklaidos. Skyriuje pateikti rezultatai publikuoti darbuose [A2], [A4]. Eksperimentinių tyrimų rezultatai pateikti 4 skyriuje.

3.1. Naujos SOM kokybę įvertinančios paklaidos

Kaip jau buvo minėta, įprastai išmokius SOM tinklą skaičiuojamos kvantavimo E_{QE} (12) ir topografinės E_{TE} (13) paklaidos. Tačiau šios paklaidos neparodo, ar SOM žemėlapyje susidarę klasteriai atitinka duomenų klases. Dažnai duomenų tyrybos procese, analizuojant klasifikuotus duomenis klasterizavimo metodais, kyla poreikis įvertinti klasių ir klasterių sutapimą. Sutapimas nurodo, kad duomenys priskirti tinkamoms klasėms. Nesutapimo atveju duomenų tyrėjas turi ieškoti nesutapimo priežasčių. Viena galima priežasčių – netinkamas duomenų priskyrimas klasėms. Tokių paklaidų, įvertinančių klasių ir klasterizavimo rezultatų sutapimą, yra (Manning ir kiti, 2008), tačiau ten duomenys turi būti vienareikšmiškai priskirti vienam iš klasterių. SOM išskirtinumas, lyginant su kitais klasterizavimo metodais yra tas, kad čia nėra griežtai išreikštų klasterių, t. y. nėra nurodyta, kuriam klasteriui kuris duomenų objektas priskirtas, o duomenų grupavimąsi galima stebėti SOM žemėlapyje. Tyrėjas, matydamas SOM žemėlapį, gali vertinti klasių ir klasterių sutapimą (ir nesutapimą). Problema kyla tuomet, kai reikia peržiūrėti daug SOM žemėlapių. Kaip žinoma, rezultatai gali priklausyti nuo SOM mokymo ar tekstinių dokumentų konvertavimo į skaitines išraiškas faktorių (esant kitoms faktorių reikšmėms, gaunami skirtingi rezultatai). Todėl tyrėjui gali tekti peržiūrėti daug SOM žemėlapių. Be to, gali pasitaikyti, kad vizualiai rezultatų

skirtumai (klasių ir klasterių nesutapimai) nėra akivaizdūs. Tuomet yra be galo sunku nustatyti, kuriame SOM žemėlapyje gauti klasteriai toliau vienas nuo kito, kuriame arčiau. Dėl šių priežasčių saviorganizuojančio neuroninio tinklo kokybei nustatyti darbe siūlomos dvi euristinės paklaidos, kurios naudojamos duomenims, kurių klasės iš anksto žinomos. Paklaidos gali būti taikomos keliems SOM žemėlapiams tarpusavyje palyginti, kai jie gauti analizuojant tą patį duomenų rinkinį ir esant tam pačiam SOM žemėlapio dydžiui.

3.1.1. Pirmoji pasiūlyta paklaida – klasės narių įvertinimas

Nagrinėjant duomenis, kurių klasės iš anksto žinomos, svarbu patikrinti, kaip tos pačios klasės duomenys išsidėsto SOM žemėlapyje, todėl pirmoji paklaida parodo, kaip arti žemėlapyje tos pačios klasės duomenys išsidėsto šalia vienas kito, ar klasės atitinka SOM klasterius. Tai leidžia įvertinti, ar visi stebimos klasės duomenys yra tarpusavyje panašūs. Paklaidos reikšmė apskaičiuojama kiekvienai klasei atskirai. Siūloma paklaidą E_c apskaičiuoti pagal formulę:

$$E_c = \frac{1}{N_c}\sum_{i=1}^{n_c-1}\sum_{j=i+1}^{n_c}\left(\left\|Z_i^c - Z_j^c\right\|k_i^c k_j^c + b_{ij}\right). \quad (14)$$

Čia:

c – klasės numeris,

N_c – c-osios klasės duomenų vektorių skaičius,

n_c – žemėlapio neuronų, į kuriuos pateko c-tosios klasės vektoriai, skaičius,

Z_i^c – žemėlapio neuronų indeksai, į kuriuos pateko c-osios klasės vektoriai,

$Z_i^c \in R^2$.

k_i^c – c-osios klasės vektorių, patekusių į neuronų n_c, kurių indeksai Z_i^c, skaičius.

Gali būti atvejų, kai į tuos pačius žemėlapio neuronus patenka skirtingų klasių nariai, tuomet pagal formulę (15) skaičiuojama bauda b_{ij}, kuri naudojama formulėje (14). Jei viename žemėlapio neurone yra tik tos pačios klasės vektoriai, tai $b_{ij} = 0$.

$$b_{ij} = \frac{l_i^{c'}}{k_i} + \frac{l_j^{c'}}{k_j}. \quad (15)$$

Čia k_i (k_j) – duomenų vektorių, patekusių į neuroną, kurių indeksai Z_i^c (Z_j^c), skaičius, $l_i^{c'}$ ($l_j^{c'}$) – kitų nei c-osios klasės vektorių, patekusių į neuronus, kurių indeksai Z_i^c (Z_j^c), skaičius.

Tikslingai paklaidos formulėje suma nėra dalijama iš sumos narių skaičiaus $n_c(n_c - 1)/2$, kadangi tokia dalyba suvienodintų paklaidos reikšmes lyginant kelis SOM žemėlapius tarpusavyje, kai $b_{ij} = 0$, ir nebūtų įvertintas neuronų, į kuriuos pateko tos pačios klasės vektoriai, susigrupavimas. Tos pačios klasės vektoriams, plačiau pasklidus po SOM žemėlapį, skaičius n_c yra didesnis nei tuomet, kai tie vektoriai labiau susigrupavę vienoje vietoje. Todėl, vertinant kelių SOM žemėlapių rezultatus, tikslinga paklaidoje (14) apskaičiuotą sumą dalinti iš to paties skaičiaus, pavyzdžiui, c-osios klasės duomenų vektorių skaičius N_c.

Mažesnė paklaida E_c reiškia didesnį tos pačios klasės duomenų susigrupavimą SOM žemėlapyje, t. y. labiau koncentruotą klasterį, be to, į klasterį nėra (arba beveik nėra) įsimaišiusių kitos klasės duomenų. Tuo atveju galima teigti, kad SOM gautas klasteris sutampa su duomenų klase. Taigi, tyrėjas gali ne tik įvertinti klasterių ir klasių sutapimą vizualiai, tačiau ir stebėdamas paklaidos reikšmes.

3.1.2. Antroji pasiūlyta paklaida – klasės centrų įvertinimas

Pirmoji paklaida yra nustatoma kiekvienai klasei atskirai. Tačiau taip pat yra svarbu nustatyti ir SOM žemėlapyje susidariusių klasterių, atitinkančių duomenų klases, tarpusavio artumą (tolumą). Antra pasiūlyta paklaida įvertina, kaip toli SOM žemėlapyje išsidėstę skirtingų klasių centrai. Stebėdamas šios paklaidos reikšmę kartu su pirmosios paklaidos reikšmėmis, tyrėjas gali ne tik vizualiai įvertinti SOM žemėlapyje susidariusių klasterių sutapimą su duomenų klasėmis. Pradžioje randami kiekvienos klasės duomenų indeksų SOM žemėlapyje centrai Y^c:

$$Y^c = \frac{1}{n_c}\sum_{i=1}^{n_c} Z_i^c. \qquad (16)$$

Čia:

n_c – žemėlapio neuronų, į kuriuos pateko c-osios klasės vektoriai, skaičius,

Z_i^c – žemėlapio neuronų indeksai, į kuriuos pateko c-osios klasės vektoriai, $Z_i^c \in R^2$.

Tuomet paklaidos E_{center} reikšmė skaičiuojama pagal formulę:

$$E_{\text{center}} = \frac{1}{m'} \sum_{c=1}^{k-1} \sum_{d=c+1}^{k} \|Y^c - Y^d\|. \qquad (17)$$

Čia $m' = \frac{k(k-1)}{2}$, k – klasių skaičius.

Didesnė paklaida E_{center} reiškia, kad klasių centrai SOM žemėlapyje yra labiau nutolę vienas nuo kito, nei esant mažesnei paklaidai.

Šiuo atveju, kuo paklaidos E_{center} reikšmė yra didesnė, tuo rezultatas geresnis (atstumas didesnis).

Tiek paklaida E_{c}, tiek E_{center} gali būti taikomi siekiant įvertinti kelių to paties dydžio SOM žemėlapiuose susidariusių klasterių sutapimą su duomenų klasėmis, kai SOM žemėlapyje skirtingai atvaizduojami tie patys duomenys. Toliau pateiktas paprastas tą iliustruojantis pavyzdys. Daugiau pavyzdžių galima rasti 4 skyriuje pateiktų eksperimentų rezultatuose.

Tarkime, turime du tokio paties dydžio SOM žemėlapius, kuriuose yra skirtingai atvaizduoti tie patys duomenys, priklausantys vienai iš trijų klasių (3.1 pav.).

1 (3, 1)	(3, 2)	**3** (3, 3)
1, 1 (2, 1)	**2, 3, 3** (2, 2)	**3** (2, 3)
(1, 1)	(1, 2)	**2** (1, 3)

a)

1 (3, 1)	**1** (3, 2)	**3, 3** (3, 3)
1 (2, 1)	**2** (2, 2)	**3, 3** (2, 3)
(1, 1)	(1, 2)	**2** (1, 3)

b)

3.1 pav. SOM žemėlapiai: a) $E_1 = 0{,}67$, $E_2 = 1{,}04$, $E_3 = 1{,}63$, $E_{\text{center}} = 1{,}32$, b) $E_1 = 1{,}13$, $E_2 = 0{,}71$, $E_3 = 1$, $E_{\text{center}} = 1{,}48$

SOM žemėlapiuose paryškinti skaičiais 1, 2, 3 žymi įėjimo vektorių klasių numerius ($c = 1, 2, 3$). Langelių kampuose skaičių poros nurodo konkretaus langelio indeksą. Kaip matome 3.1 pav., a) atveju pirmos klasės nariai išsidėstę dvejuose langeliuose, o b) atveju – trijuose. Akivaizdu, jog pirmuoju atveju pirmos klasės duomenys sudaro ryškesnį klasterį. Tą patvirtina ir gautos paklaidų reikšmės ($E_1 = 0{,}67$ a) atveju ir $E_1 = 1{,}13$ b) atveju). Mažesnė šios paklaidos reikšmė reiškia, jog klasės nariai yra arčiau vienas kito. 3.1 pav. a) atveju II klasei apskaičiuota paklaida yra lygi $E_2 = 1{,}04$, o b) atveju $E_2 = 0{,}71$. Matome, kad pirmu atveju paklaida yra didesnė todėl, kad viename langelyje atsiduria ne tik II klasės nariai, tačiau ir du III klasės nariai. Vadinasi pagal formulę (15) yra pridedama bauda $b = \frac{2}{3}$. Apskaičiavus paklaidos reikšmes III klasei E_3, matome, jog a) atveju $E_3 = 1{,}63$, o b) atveju $E_3 = 1$. Taip pat kaip II klasės atveju prie a) atvejo rezultato yra pridedama bauda, kuri yra lygi $b = \frac{1}{3}$.

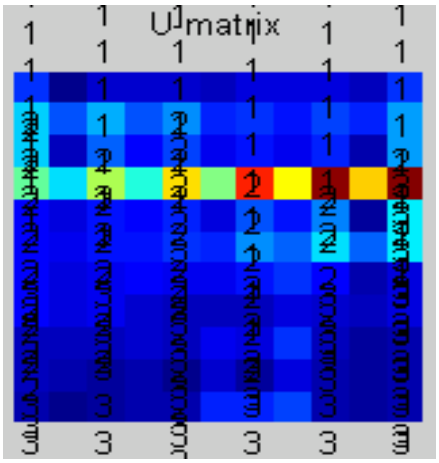
Paklaidos E_{center} reikšmė abiem atvejais yra panaši (a) atveju gauname $E_{\text{center}} = 1{,}32$, o b) atveju $E_{\text{center}} = 1{,}48$.), tačiau šiek tiek didesnė yra b) atveju, o tai reiškia, jog skirtingų klasių centrai yra toliau vienas nuo kito.

Kai SOM žemėlapiai yra dideli, sunku vizualiai įvertinti, kuriame SOM žemėlapyje tos pačios klasės nariai yra arčiau vienas kito, o pasiūlytos paklaidos leidžia tai padaryti.

Akivaizdu, kad toks SOM žemėlapio įvertinimas yra daugiakriterinis uždavinys, kadangi vienu metu reikia įvertinti kelių kriterijų reikšmes (pateikto pavyzdžio atveju vertintos 4 kriterijų reikšmės). Paprasčiausias tokio uždavinio sprendimo būdas – taikyti svertinės sumos metodą, t. y. sudėti paklaidų reikšmes, padaugintas iš norimų svorių reikšmių. Tačiau svorių parinkimas įprastai priklauso nuo sprendimų priėmėjo ir sprendžiamo uždavinio specifikos, pavyzdžiui, gali būti taip, jog vienos klasės paklaidos rezultatai yra svarbesni nei kitos. Šioje disertacijoje toks daugiakriterinis uždavinys nėra sprendžiamas, o gautos paklaidų reikšmės vertinamos atskirai.

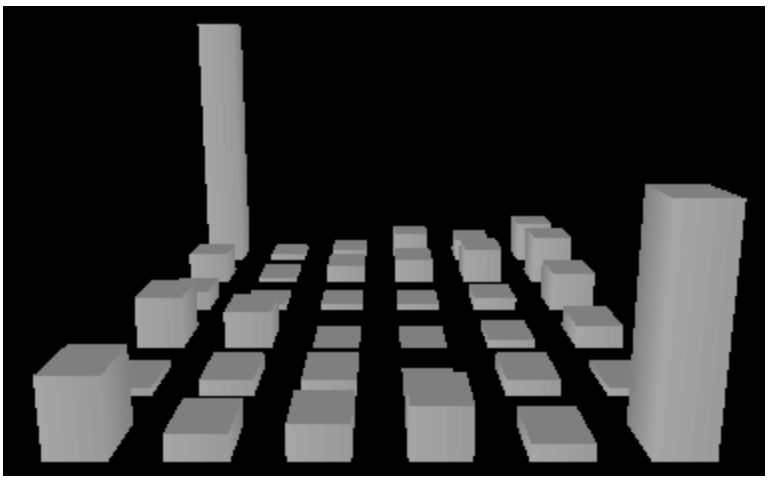
3.2. Pasiūlytas SOM vizualizavimo būdas

Analizuojant ir vizualizuojant duomenis naudojant SOM tinklą (žemėlapį), susiduriama su problema, jog visose prieš tai aptartose sistemose nėra galimybės matyti į vieną SOM langelį patekusių duomenų skaičiaus. Ypač tai aktualu tada, kai žemėlapyje sužymimi klasių pavadinimai ir gali būti atvejų, kad daug tos pačios klasės duomenų pateko į vieną langelį ir tik keli kitos klasės, tačiau langelyje nurodomos abi klasės ir nėra aišku, kiek duomenų yra kurios klasės. Pavyzdžiui, SOM-Toolbox sistemoje yra galimybė parodyti kiekvieno duomenų vektoriaus klasės pavadinimą, tačiau tuomet žemėlapis yra neaiškus ir sunku nustatyti, kuris vektorius kuriam langeliui priklauso. Didžiausia problema yra nagrinėjant duomenų aibę, kurioje duomenų vektorių skaičius yra pakankamai didelis. Pavyzdžiui, SOM-Toolbox sistemoje pasirinkus, kad būtų rodomi visi įėjimo vektorių klasių pavadinimai, žemėlapis tampa labai neaiškus (3.2 pav.).



3.2 pav. SOM-Toolbox sistemoje gautas žemėlapis, kuriame nurodyti visi analizuojamų vektorių pavadinimai

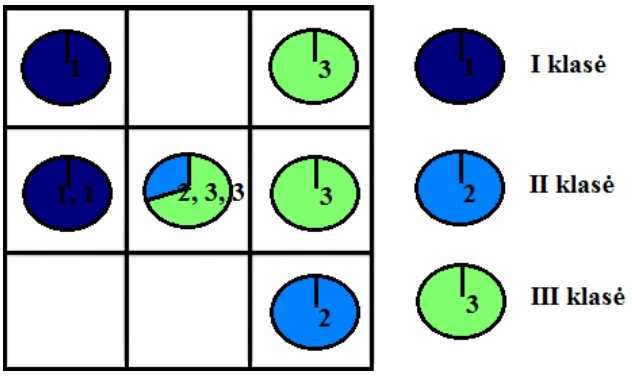
Sistemose Databionic ESOM ir Viscovery SOMine nėra galimybės atvaizduoti visų duomenų, pakliuvusių į tą patį SOM langelį, klasės pavadinimų. NeNet sistema leidžia atlikus testavimą pasižiūrėti gauto žemėlapio histogramą (3.3 pav.). Histogramoje stulpeliai atitinka vieną SOM žemėlapio langelį, o stulpelio aukštis dydis priklauso nuo į tą langelį patekusių duomenų aibės vektorių skaičiaus. Tačiau šiuo atveju vis tiek lieka neaišku, kiek ir kokios klasės narių pateko į vieną SOM žemėlapio langelį.



3.3 pav. NeNet sistema gauta žemėlapio histograma

Siekiant panaikinti komplikuoto duomenų vaizdavimo tame pačiame SOM žemėlapio langelyje trūkumą, šiame darbe pasiūlytas SOM vizualizavimo būdas, skirtas duomenims, kurių klasės yra žinomos, SOM tinkle vizualizuoti. Siūloma SOM langeliuose nubraižyti skritulinę diagramą, kurios dalys atitinka skirtingų klasių duomenų, patekusių į šį langelį, santykį su visu į tą langelį patekusių duomenų kiekiu. Be to, skirtingas klases tikslinga pavaizduoti skirtingomis spalvomis.

Tarkime turime SOM žemėlapį, pavaizduotą 3.1a pav.. Tuomet, pritaikius naują vizualizavimo būdą, turėtume 3.4 pav. pateiktą SOM žemėlapį.



3.4 pav. SOM žemėlapis, atvaizduotas pasiūlytu vizualizavimo būdu

Kaip matome, į vidurinį žemėlapio langelį patenka dviejų skirtingų klasių vektoriai (iš viso trys vektoriai), kur du vektoriai priklauso III klasei ir vienas II klasei. Skritulinės diagramos dalys nuspalvinamos atitinkamai atvaizduojant

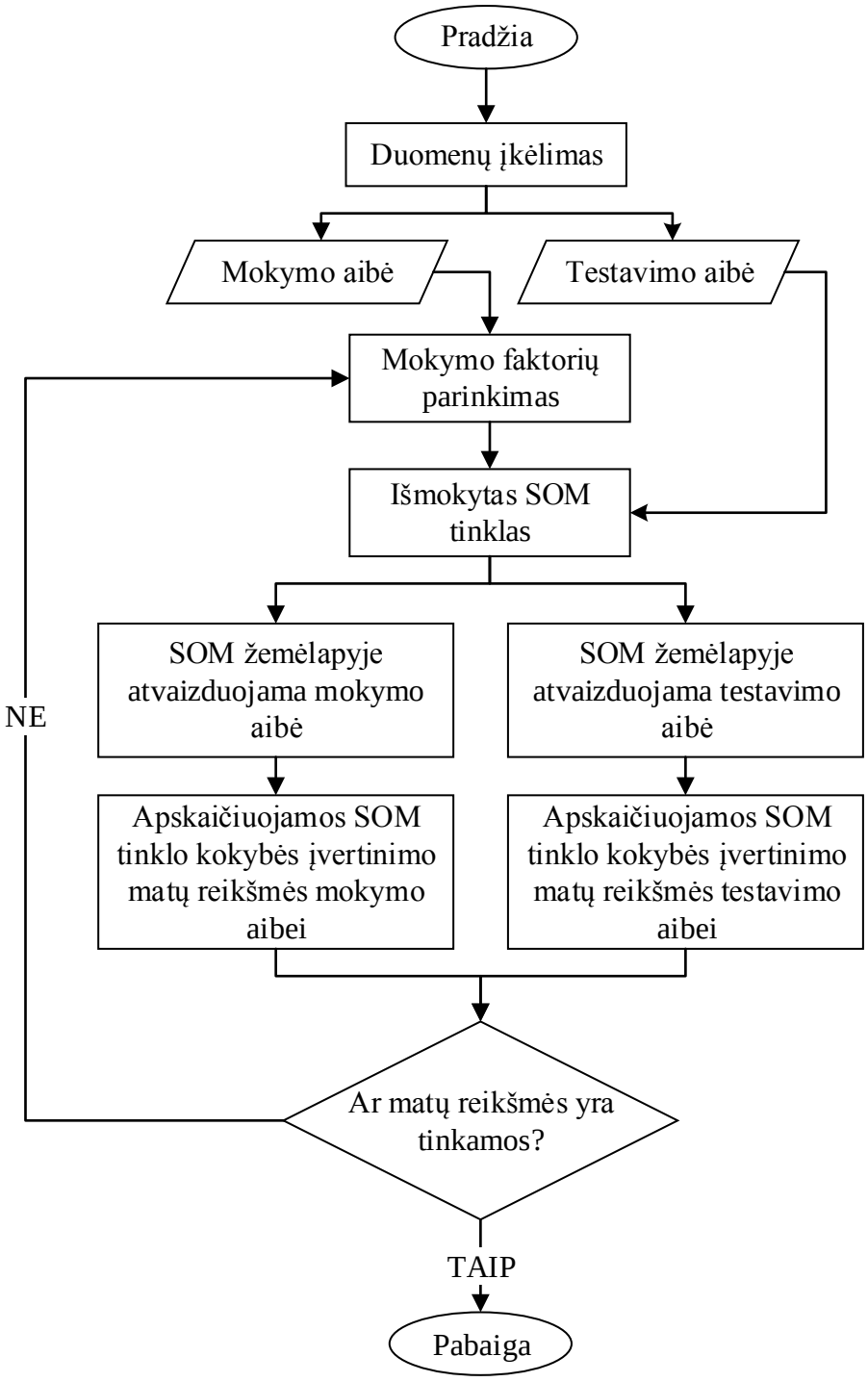
santykį: $\frac{1}{3}$ – II klasė (mėlyna) ir $\frac{2}{3}$ – III klasė (žalia). Diagramos, kuriuose yra vien tos pačios klasės nariai, nuspalvinamos tą klasę atitinkančia spalva.

Verta paminėti, kad šis SOM vizualizavimo būdas pasiūlytas 2011 metais disertacijos autoriaus darbe [A2]. Iki to laiko, remiantis žinoma informacija, nei vienoje SOM sistemoje nebuvo tokio SOM žemėlapio vizualizavimo būdo. Po kurio laiko Orange sistemoje (Demšar ir kiti, 2013) taip pat įgyvendintas panašus SOM vizualizavimo būdas.

3.2.1. Nauja SOM sistema

Įgyvendinant pasiūlytą SOM vizualizavimo būdą ir SOM kokybę įvertinančias paklaidas, darbe sukurta nauja SOM sistema. Sistemoje yra galimybė parinkti ne tik įprastas kaimynystės funkcijas (burbuliuko ir Gauso, kaip yra daugumoje sistemų), bet ir euristinę kaimynystės funkciją (8) ir euristinį mokymo parametrą (9). Taip pat sistemoje galima pasirinkti, ar mokymo parametro reikšmė turi būti keičiama kiekvieną iteraciją, ar kiekvieną epochą. Rezultate pateikiami SOM žemėlapiai ir mokymo, ir testavimo duomenų aibėms. Dėl šių savybių sistema yra tinkama ne tik konkrečių duomenų analizei, bet ir tolimesniam saviorganizuojančio neuroninio tinklo tyrimui.

3.5 pav. pateikta bendra SOM sistemos veikimo schema. Prieš pradedant vykdyti eksperimentą būtina įkelti duomenų aibę. Duomenų aibė yra išskaidoma į dvi aibes: mokymo ir testavimo, t. y. dalis duomenų bus naudojama tinklui mokyti, o kita dalis – jam testuoti. Pasirinkus norimas mokymo faktorių reikšmes, SOM tinklas mokomas naudojant mokymo aibę. Rezultate gaunamas SOM tinklas, į kurį paduodama testavimo aibė, siekiant įvertinti, kaip duomenų aibė, nenaudota mokyme, tinka gautam SOM tinklui. Taip pat yra apskaičiuojamos abiem duomenų rinkiniams SOM tinklo kokybę įvertinančių paklaidų reikšmės: kvantavimo paklaida (12), topografinė paklaida (13), paklaida tarp tos pačios klasės narių E_c (14) ir paklaida tarp skirtingų klasių centrų (17). Jeigu gautų SOM tinklų kokybė netenkina tyrėjo, jis gali grįžti prie mokymo faktorių naujų reikšmių parinkimo. Ciklas kartojamas tol, kol paklaidų reikšmės atitinka tyrėjo trokštamą gauti rezultatą.

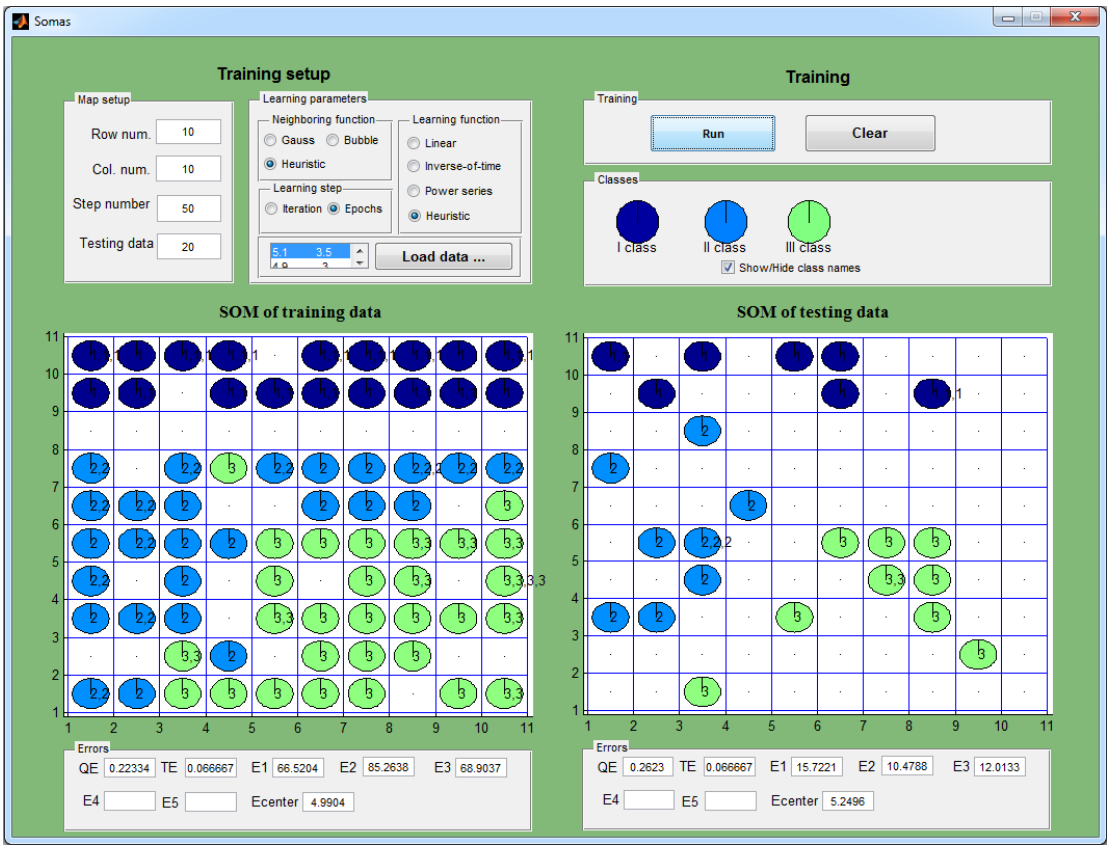


3.5 pav. Naujos SOM sistemos veikimo schema

3.2.2. SOM sistemos grafinė naudotojo sąsaja

Šiame skyrelyje aprašyta sukurtos SOM sistemos grafinė naudotojo sąsaja, kuri pateikta 3.6 pav. Sistema yra sukurta Matlab aplinkoje ir gali būti naudojama kaip papildomas Matlab įrankis. Norint įkelti duomenų failą reikia spausti mygtuką „Load data“. Duomenų failas turi būti pateiktas tekstinio failo

pavidalu (plėtiniu *.txt). Duomenų faile eilutės turi atitikti vektorius, o stulpeliai – vektorių požymius. Paskutiniame stulpelyje turi būti nurodytas vektoriaus klasės pavadinimas. Žemėlapio nustatymo srityje parenkamas SOM žemėlapio dydis („Row num" ir „Col num“), pasirenkamas mokymo žingsnių skaičius („Step number“) bei nurodomas procentinė duomenų dalis, kuri bus naudojama kaip testavimo aibė („Testing data“). Likusi duomenų dalis tampa mokymo aibe ir naudojama tinklui mokyti. Šioje sistemoje apmokant SOM tinklą galima pasirinkti vieną iš trijų kaimynystės funkcijų – burbuliuko („Bubble“) (3), Gauso („Gauss“) (4), euristinę („Heuristic“) (8) ir vieną iš keturių mokymo parametrų – tiesinį („Linear“) (5), atvirkštinį laikui („Inverse-of-time“) (6), laipsninį („Power series“) (7), euristinį („Heuristic“) (12). Taip pat yra galimybė pasirinkti, kad mokymo parametras būtų keičiamas kiekvienoje iteracijoje („Iteration“) ar kiekvienoje epochoje („Epoch“). Nustačius visus mokymo faktorius, paspaudus mygtuką „Run“ vykdomas SOM mokymas, testavimas, paklaidų apskaičiavimas bei rezultatų vaizdavimas. Rezultate gaunami du SOM žemėlapiai: mokymo aibei („SOM of training data“) ir testavimo aibei („SOM of testing data“). Vertinamos keturių paklaidų reikšmės, kurios yra pateikiamas po SOM žemėlapiais: kvantavimo (QE), topografinė (TE), paklaida, vertinanti kokybę tarp tos pačios klasės narių ($E_1, E_2, \dots, E_c$) ir paklaida, vertinanti kokybę tarp skirtingų klasių centrų (E_{center}). Sistemos dalyje „Classes“ yra pateikiamos vienspalvės skritulinės diagramos, kurios žymi duomenų klasių pavadinimus ir juos atitinkančią diagramų spalvą žemėlapyje. Taip pat galima pasirinkti žemėlapyje rodyti ar slėpti klasių numerių pavadinimus („Show/Hide class names“). Norint pradėti naują SOM mokymą, spaudžiamas mygtukas „Clear“.

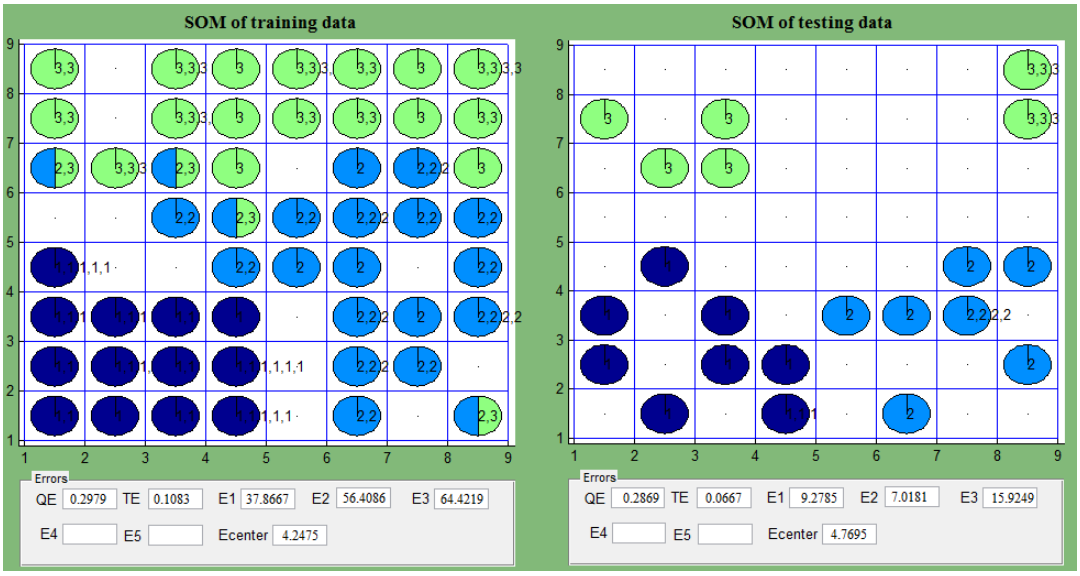


3.6 pav. SOM sistemos grafinė naudotojo sąsaja

Gautuose SOM žemėlapiuose skirtingų klasių duomenų santykis atvaizduojamas skritulinėse diagramose, atitinkamai nuspalvinant diagramos dalis. Patekus keliems skirtingų klasių nariams į tą patį SOM žemėlapio langelį, skritulinė diagrama yra nuspalvinama priklausomai nuo to, kokią dalį užima kiekviena klasė. Pasiūlyta SOM sistema pritaikyti duomenims, kurių klasių skaičius ne didesnis kaip 5, tačiau tai nėra didelis trūkumas, kadangi dažniausiai nėra sprendžiamas daugiau nei 5 klasių klasifikavimo uždavinys. Be to esant poreikiui, šią galimybę galima praplėsti.

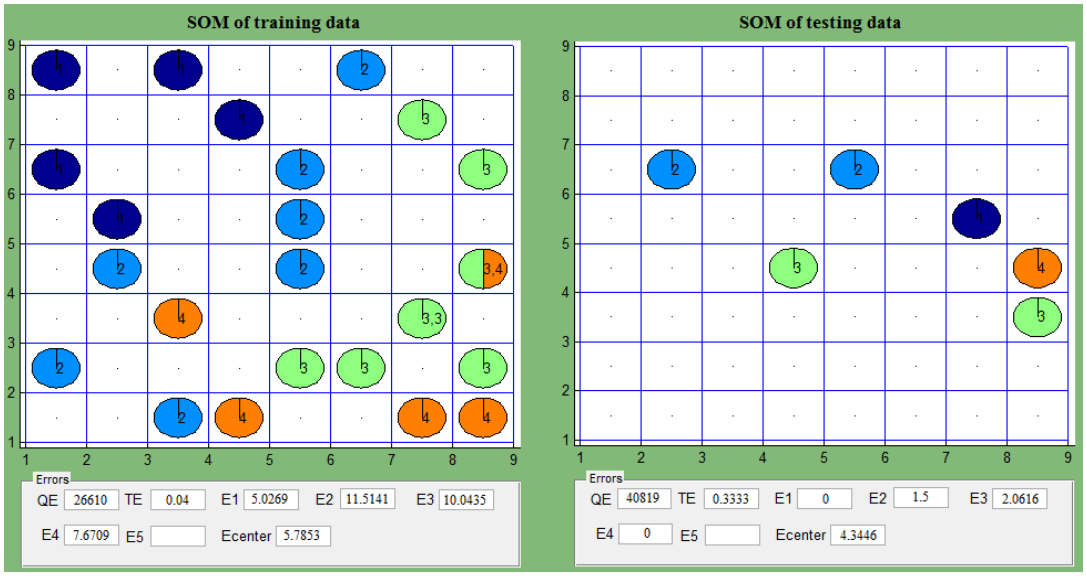
Nagrinėkime atvejį, kai SOM tinklas yra apmokomas irisų ir ekonominiais duomenimis. Irisų duomenų atveju (3.7 pav.) matome, jog kairėje pusėje žemėlapio apačioje išsidėsto vien tik I klasės nariai, kurie nuo II ir III klasės narių yra atskirti tuščiais langeliais, t. y. susidaro du atskiri klasteriai. II ir III klasės nariai kai kuriais atvejais yra „persipynę“, t y. skirtingų klasių nariai patenka į tą patį langelį, tais atvejais skritulinės diagramos dalys nuspalvinamos atvaizduojant santykį tarp iš viso patekusių duomenų į tą langelį ir konkrečios

klasės duomenų skaičiaus. Dešinėje yra pateikiamas testavimo aibės žemėlapis. Kaip matome testavimo duomenys prisitaiko prie apmokyto SOM tinklo ir atsiduria atitinkamose vietose kaip ir mokymo aibės duomenų atveju.



a) **b)**

3.7 pav. Irisų duomenys 8×8 SOM žemėlapyje, gautame nauja SOM sistema: a) mokymo aibė, b) testavimo aibė



a) **b)**

3.8 pav. Ekonominiai duomenys 8×8 SOM žemėlapyje, gautame nauja SOM sistema: a) mokymo aibė, b) testavimo aibė

Apmokius SOM tinklą ekonominių duomenų aibe (3.8 pav.) matome, kad visi I klasės nariai išsidėsto žemėlapio kairėje pusėje, II klasės nariai pasiskirsto į du mažesnius klasterius, kur vienas atsiduria žemėlapio apačioje, o kitas – žemėlapio viršuje. III ir IV klasės nariai susimaišo, kai kurie nariai patenka į tuos pačius langelius. Testavimo aibės žemėlapyje matome, jog I klasės narys atsiduria žemėlapio dešinėje pusėje, tai reiškia, jog šį neuroną atitinkanti šalis galbūt yra labiau panaši į III ir IV klasėms priskirtas šalis. Kitų klasių neuronai išsidėsto panašiose vietose kaip ir mokymo aibės atveju.

3.3. Trečiojo skyriaus apibendrinimas

Šiame skyriuje pasiūlytos naujos SOM žemėlapio kokybės įvertinimo paklaidos, skirtos keliuose SOM žemėlapiuose gautų klasterių sutapimui su klasėmis tarpusavyje palyginti. Pirmoji paklaida leidžia įvertinti, kaip arti vienas kito SOM žemėlapyje yra tos pačios klasės duomenys. Šiuo atveju, kuo paklaidos reikšmė yra mažesnė, tuo rezultatas yra geresnis, t. y. sudaryti klasteriai labiau atitinka duomenų klases. Antroji paklaida leidžia įvertinti, kaip toli SOM žemėlapyje yra skirtingų klasių duomenų centrai, todėl, kuo paklaidos reikšmė yra didesnė, tuo rezultatas yra geresnis, t. y. skirtingų klasių centrai yra toliau vienas nuo kito, todėl klases atitinkantys klasteriai žemėlapyje yra aiškiau matomi.

Atsižvelgiant į kitų SOM sistemų vizualizavimo trūkumą, pasiūlytas naujas vizualizavimo būdas, kur duomenys atvaizduojami skritulinėse diagramose. Privalumas yra tas, jog į tą patį langelį patekus skirtingų klasių duomenims, skritulinė diagrama parodo santykį, kiek yra vienos ar kitos klasės narių. Taip pat šiame skyriuje aprašyta sukurta nauja SOM sistema, kurioje realizuotos įprastai naudojamos kaimynystės funkcijos (burbuliuko ir Gauso), mokymo parametrai (tiesinis, atvirkštinis laikui ir laipsninis) bei euristinė kaimynystės funkcija ir euristinis mokymo parametras. Taip pat sistemoje yra galimybė keisti mokymo parametro reikšmes kiekvienoje iteracijoje arba epochoje.

4. Eksperimentinių tyrimų rezultatai

Šiame skyriuje yra pateikiami atliktų eksperimentinių tyrimų rezultatai. Pradžioje aprašytos duomenų aibės, naudotos eksperimentiniuose tyrimuose. Pateikta naują vizualizavimo būdą turinčios SOM sistemos ir 2.5 poskyryje nagrinėtų sistemų lyginamoji analizė. Eksperimentiškai ištirta, kaip skirtingi mokymo faktoriai įtakoja gautus SOM rezultatus, nagrinėjant tekstinius ir skaitinius duomenis. Taip pat eksperimentiškai ištirta, ar įvairūs teksto konvertavimo parametrai įtakoja gautus SOM rezultatus. Eksperimentuose gautiems SOM rezultatams vertinti naudota kvantavimo paklaida bei naujai pasiūlytos paklaidos E_c ir E_{center}. Atliktas eksperimentinis tyrimas siekiant palyginti, kaip tuos pačius duomenis apdoroja saviorganizuojantys neuroniniai tinklai ir k-vidurkių metodas. Tyrimų rezultatai publikuoti autoriaus darbuose [A3–A5], [B1], [C1], [C2].

4.1. Tyrimuose naudojami duomenys

Disertacijos eksperimentinėje dalyje yra naudojamos įvairios duomenų aibės, turinčios skirtingų specifinių savybių. Kelios duomenų aibės paimtos iš duomenų bazės „UCI Repository of Machine Learning Databases“ (Asuncion and Newman 2007):

1. **Fišerio irisų duomenys** (irisai) yra klasikiniai testiniai duomenys, naudojami daugiamačių duomenų analizei. Irisų duomenys – tai trijų veislių gėlės: iris setosa, iris versicolor ir iris virginica. Yra išmatuoti keturi gėlių parametrai: x_1 – taurėlapių ilgiai (sepal length), x_2 – taurėlapių pločiai (sepal width), x_3 – vainiklapių ilgiai (petal length), x_4 – vainiklapių pločiai (petal width). Sudaryti 4-mačiai vektoriai $X_1, X_2, \dots, X_{150}$, kur $X_p = (x_{p1}, x_{p2}, x_{p3}, x_{p4})$, $p = 1, \dots, 150$. Vektoriai $X_1, X_2, \dots, X_{50}$ atitinka I klasės irisus (iris setosa), vektoriai $X_{51}, X_{52}, \dots, X_{100}$ – II klasės (iris versicolor) ir vektoriai $X_{101}, X_{102}, \dots, X_{150}$ – III klasės (iris virginica).
2. **Stiklo duomenys** buvo surinkti mokslininkų, kurie norėjo padėti atpažinti kriminalistų rastas stiklo šukes, taip palengvinant pareigūnų darbą. Matuoti

šie parametrai: x_1 – lūžimo indeksas, x_2 – natris, x_3 – magnis, x_4 – aliuminis, x_5 – silikonas, x_6 – kalis, x_7 – kalcis, x_8 – baris ir x_9 – geležis. Taigi nagrinėjami 9-mačiai vektoriai $X_1, X_2, \dots, X_{214}$, čia $X_p = (x_{p1}, x_{p2}, x_{p3}, x_{p4}, x_{p5}, x_{p6}, x_{p7}, x_{p8}, x_{p9})$, $p = 1, \dots 214$. Visa duomenų aibė pagal savo savybes suskirstyta į penkias klases: I – pastatų stiklai, II – automobilių stiklai, III – taros stiklai, IV – stalo reikmenų stiklai ir V – žibintai.

3. **Zoologiniai duomenys** apibūdina tam tikras gyvūnų savybes. Šie duomenys pasižymi tuo, jog parametrų reikšmės įgyja 0 arba 1 (tenkina sąlyga arba ne). Matuoti šie parametrai: x_1 – turi plaukus, x_2 – turi plunksnas, x_3 – deda kiaušinius, x_4 – išskiria pieną, x_5 – skraido, x_7 – plėšrūnas, x_8 – turi dantis, x_9 – stuburinis, x_{10} – kvėpuoja per odą, x_{11} – nuodingas, x_{12} – turi pelekus, x_{13} – turi kojas, x_{14} – turi uodega, x_{15} – naminis ir x_{16} – didesnis nei 20 centimetrų. Taigi nagrinėjami 16-mačiai vektoriai $X_1, X_2, \dots, X_{92}$, čia $X_p = (x_{p1}, x_{p2}, \dots, x_{p16})$, $p = 1, \dots 92$. Visa duomenų aibė pagal savo savybes suskirstyta į penkias klases: I – žinduoliai, II – paukščiai, III – žuvys, IV – vabzdžiai ir V – bestuburiai.

Ekonominius duomenis, paimtus iš Eurostat duomenų bazės (Eurostat, 2010), sudaro Europos Sąjungos šalių bei siekiančių narystės 2009 metų ekonominiai rodikliai. Sudaryti 6-mačiai vektoriai $X_1, X_2, \dots, X_{31}$, čia $X_p = (x_{p1}, x_{p2}, x_{p3}, x_{p4}, x_{p5}, x_{p6})$, $p = 1, \dots 31$. Matuoti šie parametrai: x_1 – kompensacijos darbuotojams, x_2 – galutinio vartojimo išlaidos namų ūkio reikmėms ir ne pelno institucijų, teikiančių paslaugas namų ūkio reikmėms, x_3 – valdžios sektoriaus galutinio vartojimo išlaidos, x_4 – investicijos, x_5 – prekių ir paslaugų importas ir eksportas, x_6 – darbo našumas vienam dirbančiajam. Vektoriai $X_1, X_2, \dots, X_6$ atitinka valstybes, kurios įkūrė Europos Sąjungą (Belgija (BEL), Vokietija (DEU), Prancūzija (FRA), Italija (ITA), Liuksemburgas (LUX), Nyderlandai (NDL)), – I klasė. Vektoriai $X_7, X_8, \dots, X_{15}$ atitinka šalis, kurios prisijungė prie Europos Sąjungos 1957–1995 metais (Danija (DNK), Airija (IRL), Graikija (GRC), Ispanija (ESP), Austrija (AUT), Portugalija

(PAG), Suomija (FIN), Švedija (SWE), Anglija (GBR)), – II klasė. Vektoriai $X_{16}, X_{17}, \dots, X_{27}$ atitinka šalis, įstojusias į Europos Sąjunga 2004–2007 metais (Čekija (CZE), Estija (EST), Kipras (CYP), Latvija (LVA), Lietuva (LTU), Vengrija (HUN), Malta (MLT), Lenkija (POL), Slovėnija (SVN), Slovakija (SVK), Bulgarija (BGR), Rumunija (ROM)), – III klasė. Vektoriai $X_{28}, X_{29}, \dots, X_{31}$ atitinka valstybes, siekiančias narystės Europos Sąjungoje (Makedonija (MKD), Turkija (TUR), Islandija (ISL), Kroatija (HRV)), – IV klasė.

Tyrimuose taip pat naudotos skirtingo pobūdžio teksto dokumentų aibės, kur eksperimentų metu vektorių matmenų erdvė yra skirtinga, kadangi ji priklauso nuo tekstinio dokumentų žodyno ilgio.

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

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

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

   Trečią rinkinį sudaro: vektoriai $X_1^3, X_2^3, \dots, X_{15}^3$ – I klasė (Finansų ministerija), vektoriai $X_{16}^3, X_{17}^3, \dots, X_{30}^3$ – II klasė (Ūkio ministerija),

vektoriai $X_{31}^3,\ X_{32}^3, \dots, X_{45}^3$ – III klasė (Vidaus reikalų ministerija) ir vektoriai $X_{46}^3,\ X_{47}^3, \dots, X_{60}^3$ – IV klasė (Žemės ūkio ministerija).

2. **Moksliniai straipsniai I**. Atsitiktinai iš internete laisvai prieinamų duomenų bazių (SpringerLink, ScienceDirect ir kt.) pasirinkta 60 mokslinių straipsnių $X_1, X_2, \dots, X_{60}$. Vektoriai $X_1, X_2, \dots, X_{15}$ priskirti I klasei (atitinka straipsnius apie dirbtinius neuroninius tinklus), $X_{15}, X_{16}, \dots, X_{30}$ – II klasei (atitinka straipsnius, nagrinėjančius bioinformatikos sritį), $X_{31}, X_{32}, \dots, X_{45}$ – III klasei (atitinka straipsnius apie optimizavimo metodus) ir $X_{46}, X_{47}, \dots, X_{60}$ – IV klasei (atitinka straipsnius apie saviorganizuojančius neuroninius tinklus).
3. **Moksliniai straipsniai II**. Atsitiktinai iš internete laisvai prieinamų duomenų bazių (SpringerLink, ScienceDirect ir kt.) pasirinkta 45 straipsniai $X_1, X_2, \dots, X_{45}$, kurių tematika yra optimizavimas, tačiau skiriasi nagrinėjamos sritys. Vektoriai $X_1, X_2, \dots, X_{15}$ priskirti I klasei (Pareto optimizavimas), $X_{15}, X_{16}, \dots, X_{30}$ – II klasei (simpleksas), $X_{31}, X_{32}, \dots, X_{45}$ – III klasei (genetiniai algoritmai).

Tokį duomenų aibių skirstymą į rinkinius ir klasių priskyrimą atlikto disertacijos autorius.

4.2. Saviorganizuojančių neuroninių tinklų sistemų lyginamoji analizė

Darbe sukurta sistema, kurioje įgyvendintas pasiūlytas SOM vizualizavimo būdas (3.2 poskyris), palyginta su kitomis 2.5 poskyryje nagrinėtomis SOM sistemomis: NeNet, SOM-Toolbox, Databionic ESOM, Viscovery SOMine, Orange. Palyginę sistemas, galime pastebėti įvairias jų savybes, privalumus bei trūkumus. Pasirinkti sistemų vertinimo kriterijai bei jų pagrindimas pateikti 4.1 lentelėje.

4.1 lentelė. Sistemų vertinimo kriterijai

Kriterijaus Nr.	**Kriterijaus apibrėžimas**	**Kriterijaus pagrindimas**
K1	Yra galimybė analizuoti įvairaus dydžio duomenų aibes.	Nagrinėjant tekstinių dokumentų duomenis, duomenų aibės vektoriai turi daug požymių, todėl svarbu, jog SOM sistemos turėtų galimybę nagrinėti įvairaus dydžio duomenis ir nebūtų išankstinių apribojimų.
K2	Paprastas duomenų paruošimas sistemai, yra galimybė duomenis pateikti paprasčiausiu tekstinio failo pavidalu.	Ilgas ir sudėtingas duomenų failų paruošimas apsunkina darbą sistemoje, todėl dažnai tyrėjas dėl šios priežasties rinkasi paprastesnę sistemą.
K3	Yra galimybė duomenis išskaidyti į mokymo ir testavimo aibes.	Tai svarbu, nes išmokius SOM tinklą mokymo aibės duomenimis yra naudinga patikrinti, kaip prie jau išmokyto tinklo prisitaiko testavimo duomenų aibė, kuri nebuvo naudota tinklo mokyme.
K4	Yra galimybė naudoti daugiau nei du mokymo parametrus.	Norint išsamiai ištirti analizuojamą duomenų aibę, svarbu parinkti tinkamus mokymo parametrus, todėl

Kriterijaus Nr.	Kriterijaus apibrėžimas	Kriterijaus pagrindimas
		sistemoje turi būti įgyvendinti bent keli pagrindiniai mokymo parametrai, pavyzdžiui: tiesinis (5), atvirkštinis laikui (6), laipsninis (7) ar kt.
K5	Yra galimybė naudoti daugiau nei vieną kaimynystės funkciją.	Taip pat kaip ir mokymo parametrų atveju, sistemoje turi būti įgyvendintos bent dvi pagrindinės kaimynystės funkcijos: burbuliuko (3) ir Gauso (4).
K6	Yra galimybė naudoti skirtingas mokymo parametrų reikšmių keitimo taisykles (kiekvienoje epochoje arba kiekvienoje iteracijoje).	Skirtingos mokymo parametrų reikšmių keitimo taisyklės leidžia atlikti išsamesnę duomenų analizę, kadangi vieni mokymo parametrai pateikia geresnius SOM rezultatus, kai jų reikšmės keičiamos kiekvienoje iteracijoje, kiti – kiekvienoje epochoje.
K7	SOM žemėlapyje atvaizduojami visų į tą patį langelį patekusių duomenų klasių pavadinimai.	Matant visų duomenų patekusių į tą patį SOM žemėlapio langelį pavadinimus, tyrėjas gali daryti išvadas, kurie duomenų

Kriterijaus Nr.	Kriterijaus apibrėžimas	Kriterijaus pagrindimas
		vektoriai yra panašūs ir kurie klasteriai yra „stipresni“.
K8	SOM žemėlapyje atvaizduojamas santykis tarp skirtingų klasių duomenų, patekusių į tą patį žemėlapio langelį.	Tyrėjui tampa aišku, kiek ir kokios klasės duomenų yra tame pačiame SOM langelyje. Tuomet gali daryti išvadas apie galimai neteisingai priskirtas kai kurių duomenų klases arba tuos duomenis analizuoti detaliau.
K9	SOM žemėlapyje atvaizduojamas atstumas tarp neuronų.	Šis kriterijus yra svarbus tuo, jog yra matoma, kaip arti yra susidarę klasteriai ar atskiri duomenys. Tai tyrėjui leidžia spręsti apie gautų klasterių kokybę.
K10	Yra galimybė pasirinkti keletą skirtingų SOM vizualizavimo būdų.	Skirtingi sistemų SOM žemėlapio atvaizdavimo būdai leidžia tyrėjui pasirinkti jam priimtinesnį ir aiškesnį atvaizdavimą, tokiu būdu gauti daugiau informacijos apie analizuojamą duomenų aibę.

4.2 lentelėje nurodyta, kokius kriterijus atitinka nagrinėtos sistemos. Paskutiniame stulpelyje nurodyta, kiek iš viso kriterijų kiekviena sistema atitinka. „+“ ženklas reiškia, jog sistema tenkina kriterijų, „–“ – netenkina.

Įvertinus sistemas matome, jog mažiausiai kriterijų (3 kriterijai) atitinka sistemos NeNet ir Databionic ESOM. Šiose sistemose yra įgyvendintos tik pagrindinės SOM funkcijos ir valdymo parametrai, todėl, norint atlikti detalesnį duomenų tyrimą, susiduriama su vienais ar kitais apribojimas. Sistema Viscovery SOMine (4 kriterijai) turi įvairių vizualizavimo būdų, tačiau taip pat nėra įgyvendintos galimybės pasirinkti įvairius mokymo faktorius. Sistemoje Orange taip pat buvo įgyvendintas panašus SOM vizualizavimo būdas į disertacijoje pasiūlytą (Demšar ir kiti, 2013), tačiau sistemoje nėra kitų SOM mokymui svarbių nustatymų, todėl sistema atitiko iš viso tik 6 kriterijus. Daugiausiai kriterijų atitiko sistemos SOM-Toolbox (9 kriterijai) ir naujai pasiūlyta SOM sistema (8 kriterijai). SOM-Toolbox sistemą yra sukūrę SOM tinklų pradininko T. Kohoneno komanda, todėl joje yra realizuota labai daug įvairių funkcijų, kurios leidžia atlikti išsamią duomenų analizę. Tačiau sistemoje SOM-Toolbox nėra galimybės žemėlapyje matyti santykio tarp skirtingų duomenų klasių, patekusių į tą patį SOM langelį. Ši galimybė yra įgyvendinta naujoje SOM sistemoje.

4.2 lentelė. SOM sistemų palyginimas

Kriterijai / **Sistema**	**K1**	**K2**	**K3**	**K4**	**K5**	**K6**	**K7**	**K8**	**K9**	**K10**	**Iš viso**
NeNet	–	+	–	–	–	–	–	–	+	+	3
SOM-Toolbox	+	+	+	+	+	+	+	–	+	+	9
Databionic ESOM	+	–	–	–	–	–	–	–	+	+	3
Viscovery SOMine	+	+	–	–	–	–	–	–	+	+	4
Orange	+	+	–	–	+	–	–	+*	+	+	6
Nauja SOM sistema	+	+	+	+	+	+	+	+	–	–	8

* – panašus SOM žemėlapio vizualizavimo būdas, įgyvendintas vėliau nei buvo pasiūlytas disertacijos autoriaus.

4.3. Saviorganizuojančių neuroninių tinklų mokymo faktorių tyrimas

Šio tyrimo tikslas – išsiaiškinti, kaip skirtingos kaimynystės funkcijos, mokymo parametrų tipai bei mokymo parametro reikšmių keitimas kiekvienoje iteracijoje arba epochoje daro įtaka gautiems SOM rezultatams. SOM kokybei įvertinti naudota kvantavimo paklaida (11) ir pasiūlytos paklaidos E_c (14) ir E_{center} (17).

4.3.1. Mokymo faktorių įtaka tiriant tekstinius duomenis

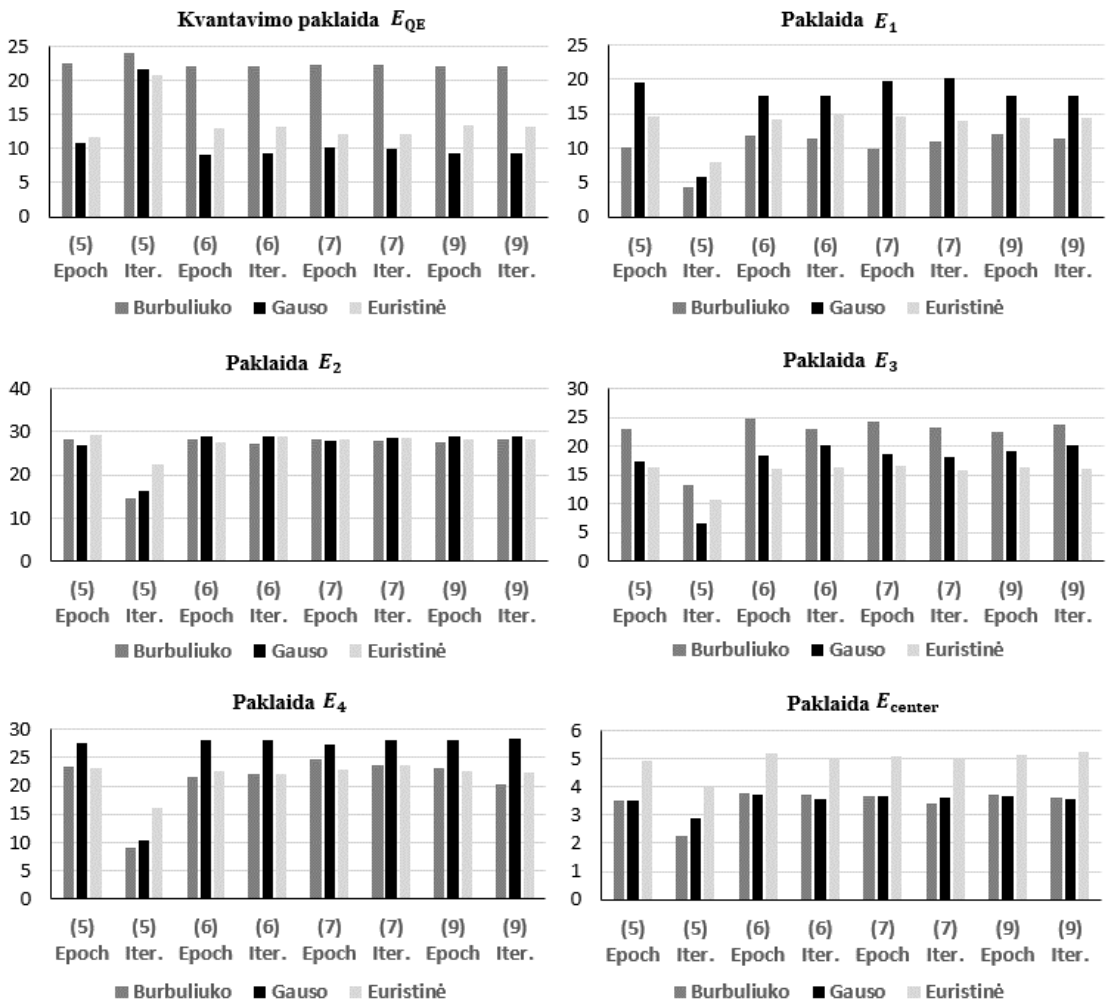
Toliau aprašytas eksperimentinis tyrimas atliktas naudojant duomenų aibės „Ministerijų įsakymai“ pirmąjį dokumentų rinkinį. Visi 60 dokumentų konvertuoti į skaitines išraiškas. Sudarant teksto dokumentų matricą, buvo atmesti skaitmenys, kadangi jie neteikia jokios esminės informacijos apie nagrinėjamus dokumentus. Pirminis eksperimentas parodė, jog šiam duomenų rinkiniui tikslingą žodžių pasikartojimų skaičių parinkti nedidesnį už 5, kadangi pasirinkus didesnį pasikartojimų skaičių, kai kuriuose dokumentuose nėra tokių žodžių, kurie kartotųsi daugiau nei 5 kartus, ir tuomet, sudarant teksto dokumentų žodyną, šie dokumentai yra tiesiog atmetami. Taigi, naudojant pirmąjį rinkinį, yra sukuriamos penkios teksto dokumentų matricos (žodžių pasikartojimas kinta nuo 1 iki 5), kurias sudaro 60 eilučių ($m = 60$, tiek kiek yra dokumentų) ir skirtingas stulpelių skaičius n:

- 3812 (kai žodžių pasikartojimas nemažesnis 1);
- 1494 (kai žodžių pasikartojimas nemažesnis 2);
- 769 (kai žodžių pasikartojimas nemažesnis 3);
- 446 (kai žodžių pasikartojimas nemažesnis 4);
- 287 (kai žodžių pasikartojimas nemažesnis 5).

Pirminiai tyrimai parodė, jog saviorganizuojančio neuroninio tinklo žemėlapio dydis bei parinktas didesnis mokymo žingsnio skaičius (tiriant nagrinėjamas duomenų aibes) iš esmės neturi įtakos gautiems SOM rezultatams, skirtumas yra nežymus. Kai kuriais atvejais padidinus arba sumažinus SOM žemėlapio dydį, paklaidų reikšmių rezultatai pagerėja, kai kuriais – pablogėja. Todėl atliekant šį tyrimą pasirinktas žemėlapis, kurio dydis 10×10 ($k_x = k_y =$

10). Saviorganizuojantis neuroninis tinklas apmokytas naudojant 80 % duomenų aibės, o likusieji 20 % sudaro testavimo aibę. Ji naudojama siekiant patikrinti, kaip testavimo aibės duomenys prisitaiko prie jau išmokyto SOM tinklo. Kiekvienas eksperimentas kartotas 10 kartų, kaskart parenkant skirtingas pradines žemėlapio neuronų M_{ij} $(i = 1, \dots, k_x, j = 1, \dots, k_y)$ reikšmes. Taigi rezultate bus apskaičiuojamos ir pateikiamos 10 bandymų kvantavimo paklaidos (11) bei naujai pasiūlytų paklaidų (14), (17) vidurkiai.

Pirmame eksperimente naudotos penkios teksto dokumentų matricos, gautos, kai žodžių pasikartojimas kinta nuo 1 iki 5. Apskaičiuoti gautų rezultatų vidurkiai, kurie yra pateikti 4.1 pav. Po pateiktomis stulpelinėmis diagramomis skaičiai skliausteliuose žymi mokymo parametro formulės numerį, pateiktą 2.2.2 skyrelyje. Pateikti dviejų atvejų rezultatai: kai mokymo parametro reikšmė yra keičiama kiekvienoje iteracijoje (Iter.) arba kiekvienoje epochoje (Epoch). Naudojamos trys kaimynystės funkcijos (burbuliuko, Gauso ir euristinė), aprašytos 2.2.2 skyrelyje. Kaip matome 4.1 pav. mažiausios kvantavimo paklaidos E_{QE} gaunamos, kai naudojama Gauso kaimynystės funkcija ir atvirkštinis laikui (6) bei euristinis mokymo parametrai (9). Naudojant atvirkštinį laikui mokymo parametrą kvantavimo paklaida yra šiek tiek mažesnė, kai parametro reikšmė keičiama kiekvienoje epochoje, o euristinio mokymo parametro atveju, – kai parametro reikšmė keičiama kiekvienoje iteracijoje. Vieninteliu atveju kvantavimo paklaida yra mažesnė, kai naudojama euristinė kaimynystės funkcija ir tiesinis mokymo parametras, keičiant jo reikšmes kiekvienoje iteracijoje. Didžiausios E_{QE} reikšmės gautos, kai naudojama burbuliuko kaimynystės funkcija. Naudojant euristinę kaimynystės funkciją, gauti kvantavimo rezultatai yra daugeliu atvejų mažesni, lyginant su rezultatais, gautais naudojant burbuliuko kaimynystės funkciją, bet didesni, lyginant su Gauso kaimynystės funkcijos atvejų rezultatais.



4.1 pav. SOM kokybę įvertinančių paklaidų vidutinės reikšmės mokymo aibei (naudota ministerijų įsakymų duomenų aibės)

Pažvelgus į paklaidų E_1, E_2, E_3, E_4 rezultatus, matome, jog visais I klasės atvejais mažiausia E_1 reikšmė yra gauta naudojant burbuliuko kaimynystės funkciją. Tai reiškia, jog šiuo atveju I klasės nariai yra išsidėstę arčiau vienas kito (kuo paklaidos E_c reikšmė yra mažesnė, tuo rezultatas geresnis). Taip pat E_1 reikšmės pakankamai mažos, naudojant euristinę kaimynystės funkciją. Mažiausia E_1 reikšmė yra gauta, naudojant tiesinį (5) mokymo parametrą, keičiant jo reikšmę kiekvienoje iteracijoje. II klasės atveju rezultatuose sunku įžiūrėti kokias nors tendencijas, kadangi E_2 reikšmės yra gana panašios, naudojant skirtingas kaimynystės funkcijas ir mokymo parametrus. Labiausiai rezultatai išskiria tik tuomet, kai naudotas tiesinis (5) mokymo parametras, keičiant jo reikšmes kiekvienoje iteracijoje. III klasės atveju, mažiausios E_3

reikšmės gaunamos, naudojant euristinę kaimynystės funkciją (išskyrus atvejį, kai naudojamas tiesinis (5) mokymo parametras ir jo reikšmės keičiamos kiekvienoje iteracijoje). IV klasės E_4 reikšmės yra labai panašios, naudojant burbuliuko ir euristinę kaimynystės funkcijas. Mažiausia E_4 reikšmė gauta, naudojant tiesinį (5) mokymo parametrą keičiant jį kiekvienoje iteracijoje. Didžiausia paklaidos E_{center} reikšmė yra gauta, naudojant euristinę kaimynystės funkciją ir euristinį mokymo parametrą (9), keičiant jo reikšmę kiekvienoje iteracijoje. Tai reiškia, jog šiuo atveju skirtingų klasių centrai yra nutolę vienas nuo kito labiau, todėl klasės atsiskiria ryškiau. Gauti rezultatai, naudojant burbuliuko ir Gauso kaimynystės funkcijas, yra panašūs. 4.3 lentelėje pateikti skaičiai nurodo, kiek kartų (iš 8 galimų) kuri kaimynystės funkcija leido gauti geriausias paklaidų reikšmes. Kaip matome, vertinant rezultatus geriausi, rezultatai gauti, naudojant euristinę kaimynystės funkciją (22 kartai), o blogiausi, – naudojant Gauso kaimynystės funkciją (11 kartų). Iš anksčiau pateikto grafiko (2.7 pav.) matyti, kad euristinės funkcijos reikšmės yra mažesnės nei Gauso, todėl galbūt tai lemia gautus rezultatus.

4.3 lentelė. Apibendrinti 4.1 pav. rezultatai, vertinant kaimynystės funkcijas

Paklaida / **Kaimynystės funkcija**	E_{QE}	E_1	E_2	E_3	E_4	E_{center}	Iš viso
Burbuliuko	0	8	4	0	3	0	15
Gauso	7	0	2	1	1	0	11
Euristinė	1	0	2	7	4	8	22

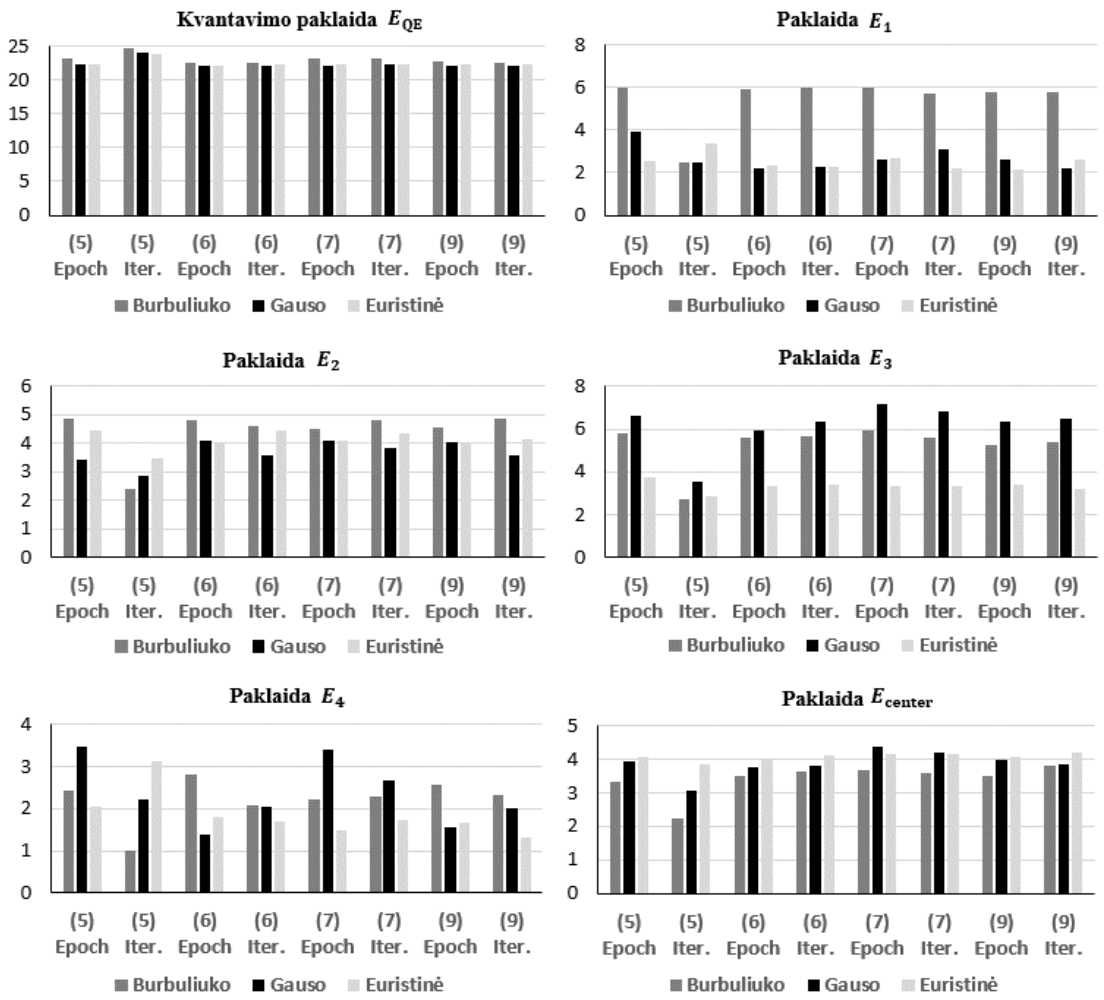
Siekiant apibendrinti, kuris mokymo parametras leidžia gauti geresnius SOM rezultatus, apskaičiuoti eksperimento metu gautų trijų kaimynystės funkcijų paklaidų reikšmių vidurkiai ir gauti rezultatai įvertinti balais nuo 1 iki 8: geriausias iš visų 8 eksperimentų rezultatas įvertintas 8 balais, blogiausias – 1. Paskutiniame 4.4 lentelės stulpelyje pateikta balų suma. Kuo ji didesnė, tuo eksperimentas laikomas geresniu.

4.4 lentelė. Apibendrinti 4.1 pav. rezultatai, vertinant mokymo parametrus

Paklaida / **Mokymo parametras**		E_{QE}	E_1	E_2	E_3	E_4	E_{center}	Iš viso
Tiesinis (5)	Epochos	2	2	7	7	2	2	22
	Iteracijos	1	8	8	8	8	1	34
Atvirkštinis laikui (6)	Epochos	8	6	5	4	5	8	36
	Iteracijos	5	5	3	2	6	4	25
Laipsninis (7)	Epochos	4	3	6	3	1	6	23
	Iteracijos	7	1	2	6	4	3	23
Euristinis (8)	Epochos	3	4	4	5	3	7	26
	Iteracijos	6	7	1	1	7	5	27

Kaip matome, 4.4 lentelėje daugiausiai balų surinko atvirkštinis laikui mokymo parametras (36 balai), kai jo reikšmės yra keičiamos kiekvienoje epochoje, šiuo atveju gaunami geriausi SOM rezultatai. Blogiausi rezultatai gaunami, naudojant tiesinį mokymo parametrą (22 balai), keičiant jo reikšmes kiekvienoje epochoje. Taip pat kaip ir anksčiau pažiūrėjus į 2.2.2 skyrelyje pateiktus 2.5–2.6 pav. matome, kad atvirkštinio laikui mokymo parametro reikšmės kur kas greičiau mažėja nei tiesinio mokymo parametro reikšmės, todėl galbūt tai lemia, kad geresnius rezultatus gauname tuomet, kai yra naudojamas atvirkštinis laikui mokymo parametras, o blogesnius, – naudojant tiesinį mokymo parametrą.

Nagrinėjamų paklaidų vidutinės reikšmės testavimo duomenų aibėms pateiktos 4.2 pav. Kvantavimo paklaidos E_{QE} rezultatai yra tarpusavyje panašūs nepriklausomai nuo kaimynystės funkcijos, todėl sunku įžvelgti esminę mokymo parametrų įtaką. Paklaidų E_1 ir E_2 reikšmės yra panašios, naudojant Gauso ir euristinę kaimynystės funkcijas, tačiau euristinė kaimynystės funkcija leidžia gauti šiek tiek geresnius rezultatus. Mažiausios paklaidos E_3 reikšmės gaunamos, naudojant euristinę kaimynystės funkciją (išskyrus atvejį, kai naudojamas tiesinis (5) mokymo parametras, keičiant jo reikšmes kiekvienoje iteracijoje). Paklaidos E_4 reikšmės yra gana įvairios, todėl sunku pasakyti, kuri kaimynystės funkcija leidžia gauti geresnius rezultatus.



4.2 pav. SOM kokybę įvertinančių paklaidų vidutinės reikšmės testavimo aibei (naudota ministerijų įsakymų duomenų aibės)

Mažiausia paklaidos E_4 reikšmė gauta, naudojant burbuliuko kaimynystės funkciją ir tiesinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje. Beveik visais atvejais paklaidos E_{center} reikšmė yra didesnė (skirtingų klasių centrai yra nutolę vienas nuo kito), naudojant euristinę kaimynystės funkciją. Geriausias rezultatas gautas, naudojant euristinę kaimynystės funkciją ir euristinį mokymo parametrą keičiant jo reikšmes kiekvienoje iteracijoje.

4.5 lentelėje matome, jog taip pat kaip yra mokymo aibės rezultatų atveju, naudojant euristinę kaimynystės funkciją, geresni rezultatai gaunami dažniau (25 kartai), nei burbuliuko ir Gauso kaimynystės funkcijų atvejais. Blogiausi rezultatai gauti naudojant burbuliuko kaimynystės funkciją (4 kartai).

4.5 lentelė. Apibendrinti 4.2 pav. rezultatai, vertinant kaimynystės funkcijas

Paklaida / **Kaimynystės funkcija**	E_{QE}	E_1	E_2	E_3	E_4	E_{center}	Iš viso
Burbuliuko	0	1	1	1	1	0	4
Gauso	6	4	5	0	2	2	19
Euristinė	2	3	2	7	5	6	25

4.6 lentelė. Apibendrinti 4.2 pav. rezultatai, vertinant mokymo parametrus

Paklaida / **Mokymo parametras**		E_{QE}	E_1	E_2	E_3	E_4	E_{center}	Iš viso
Tiesinis (5)	Epochos	2	1	3	2	1	3	12
	Iteracijos	1	8	8	8	4	1	30
Atvirkštinis laikui (6)	Epochos	8	7	2	7	5	2	31
	Iteracijos	6	6	5	4	6	4	31
Laipsninis (7)	Epochos	3	2	4	1	2	8	20
	Iteracijos	4	3	1	3	3	6	20
Euristinis (8)	Epochos	5	5	6	6	7	5	34
	Iteracijos	7	4	7	5	8	7	38

4.6 lentelėje matome, jog testavimo duomenų aibės atveju daugiausiai balų surinkta, naudojant euristinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (38 balai). Mažiausiai balų surinkta, taip pat kaip ir mokymo aibės atveju naudojant tiesinį mokymo parametrą, keičiant jo reikšmes kiekvienoje epochoje (12 balų).

Apibendrinant eksperimento rezultatus, naudojant teksto dokumentų aibes, galima pasakyti, jog mažiausia kvantavimo paklaida yra gaunama, naudojant Gauso kaimynystės funkciją ir bet kurį iš nagrinėtų mokymo parametrų, keičiant jo reikšmes tiek kiekvienoje epochoje, tiek kiekvienoje iteracijoje. Kai kuriais atvejais kvantavimo paklaida šiek tiek mažesnė, kai mokymo parametro reikšmės keičiamos kiekvienoje epochoje. Mažiausi paklaidų E_1, E_2, E_3, E_4 rezultatai daugeliu atvejų yra gaunami, kai naudojama euristinė kaimynystės funkcija su įvairiais mokymo parametrais, keičiant jų reikšmes tiek kiekvienoje iteracijoje, tiek kiekvienoje epochoje. Iš gautų rezultatų sunku vienareikšmiškai pasakyti, kurį mokymo parametrą naudojant gaunami geresni rezultatai.

Paklaidos E_{center} reikšmės gautos didžiausios, naudojant euristinę kaimynystės funkciją ir euristinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje. Kaip jau buvo pateikta prieš tai (2.2.2 skyrelis, 2.7 pav.), euristinės kaimynystės funkcijos reikšmės yra mažesnės nei burbuliuko ir Gauso kaimynystės funkcijų, todėl galbūt tai turi įtakos gautiems SOM rezultatams.

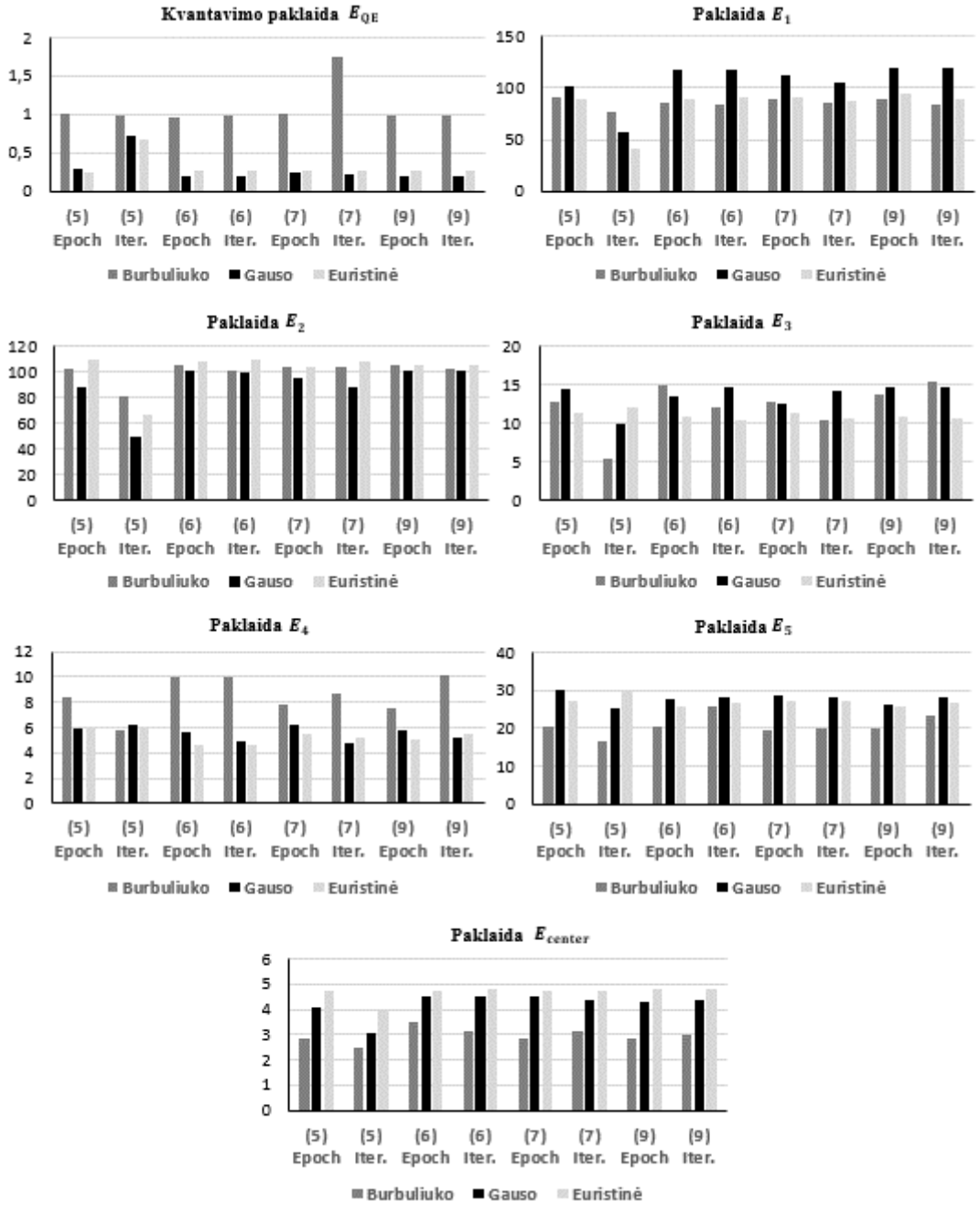
4.3.2. Mokymo faktorių įtaka tiriant skaitinius duomenis

Panašus tyrimas atliktas naudojant dvi skaitines duomenų aibes: stiklo ir zoologinius duomenis. Apskaičiuoti gautų kvantavimo paklaidų ir pasiūlytų paklaidų rezultatų vidurkiai, kurie pateikti 4.3 pav.

Beveik visais atvejais mažiausia kvantavimo paklaida E_{QE} taip pat yra gauta, naudojant Gauso kaimynystės funkciją (išskyrus atvejus, kai naudojamas tiesinis mokymo parametras (5), keičiant jo reikšmes kiekvienoje iteracijoje ir kiekvienoje epochoje). Didžiausios kvantavimo paklaidos gautos, kai yra naudojama burbuliuko kaimynystės funkcija.

I klasės atveju paklaidos E_1 reikšmės yra mažesnės, naudojant burbuliuko ir euristinę kaimynystės funkcijas, lyginant su reikšmėmis, gautomis naudojant Gauso kaimynystės funkciją. Naudojant burbuliuko ir euristinę kaimynystės funkcijas, gaunamos paklaidos reikšmės yra tarpusavyje labai panašios. Paklaidos E_2 mažiausios reikšmės yra gautos, naudojant Gauso kaimynystės funkciją. Mažiausia reikšmė gauta, naudojant tiesinį mokymo parametrą (5), keičiant jo reikšmes kiekvienoje iteracijoje. Naudojant kitas kaimynystės funkcijas, paklaidų reikšmės yra didesnės. Mažiausia trečiosios paklaidos E_3 reikšmė yra gauta, naudojant euristinę kaimynystės funkciją (išskyrus atvejį, kai naudotas tiesinis mokymo parametras (5), keičiant jo reikšmes kiekvienoje iteracijoje). Burbuliuko ir Gauso kaimynystės funkcijų atvejais paklaidos reikšmės yra didesnės. Ketvirtosios paklaidos E_4 reikšmės yra mažiausios, naudojant burbuliuko ir euristinę kaimynystės funkcijas. Mažiausios reikšmės gautos, kai naudotas atvirkštinis laikui mokymo parametras (6), keičiant jo reikšmes kiekvienoje iteracijoje arba kiekvienoje epochoje. Gautos E_5 paklaidos reikšmės yra mažiausios, naudojant burbuliuko kaimynystės funkciją. Reikšmės

tarp skirtingų klasių centrų E_{center} yra didžiausios, naudojant euristinę kaimynystės funkciją. Didžiausia reikšmė gauta, kai naudotas euristinis mokymo parametras (9), keičiant jo reikšmes kiekvienoje iteracijoje. Mažiausios reikšmės gautos, naudojant burbuliuko kaimynystės funkciją.



4.3 pav. SOM kokybę įvertinančių paklaidų vidutinės reikšmės mokymo aibei (stiklo ir zoologinių duomenų aibės)

4.7 lentelėje pateikti skaičiai nurodo, kiek kartų (iš 8 galimų), kuri kaimynystės funkcija leido gauti geriausias paklaidų reikšmes. 4.7 lentelėje matome, jog dažniausiai geresni rezultatai gauti, kai naudota euristinė kaimynystės funkcija (21 kartas), tačiau skirtumas yra nežymūs, lyginant su kitomis kaimynystės funkcijomis.

4.7 lentelė. Apibendrinti 4.3 pav. rezultatai, vertinant kaimynystės funkcijas

Paklaida / **Kaimynystės funkcija**	E_{QE}	E_1	E_2	E_3	E_4	E_5	E_{center}	Iš viso
Burbuliuko	0	7	0	2	1	8	0	18
Gauso	6	0	8	0	3	0	0	17
Euristinė	2	1	0	6	4	0	8	21

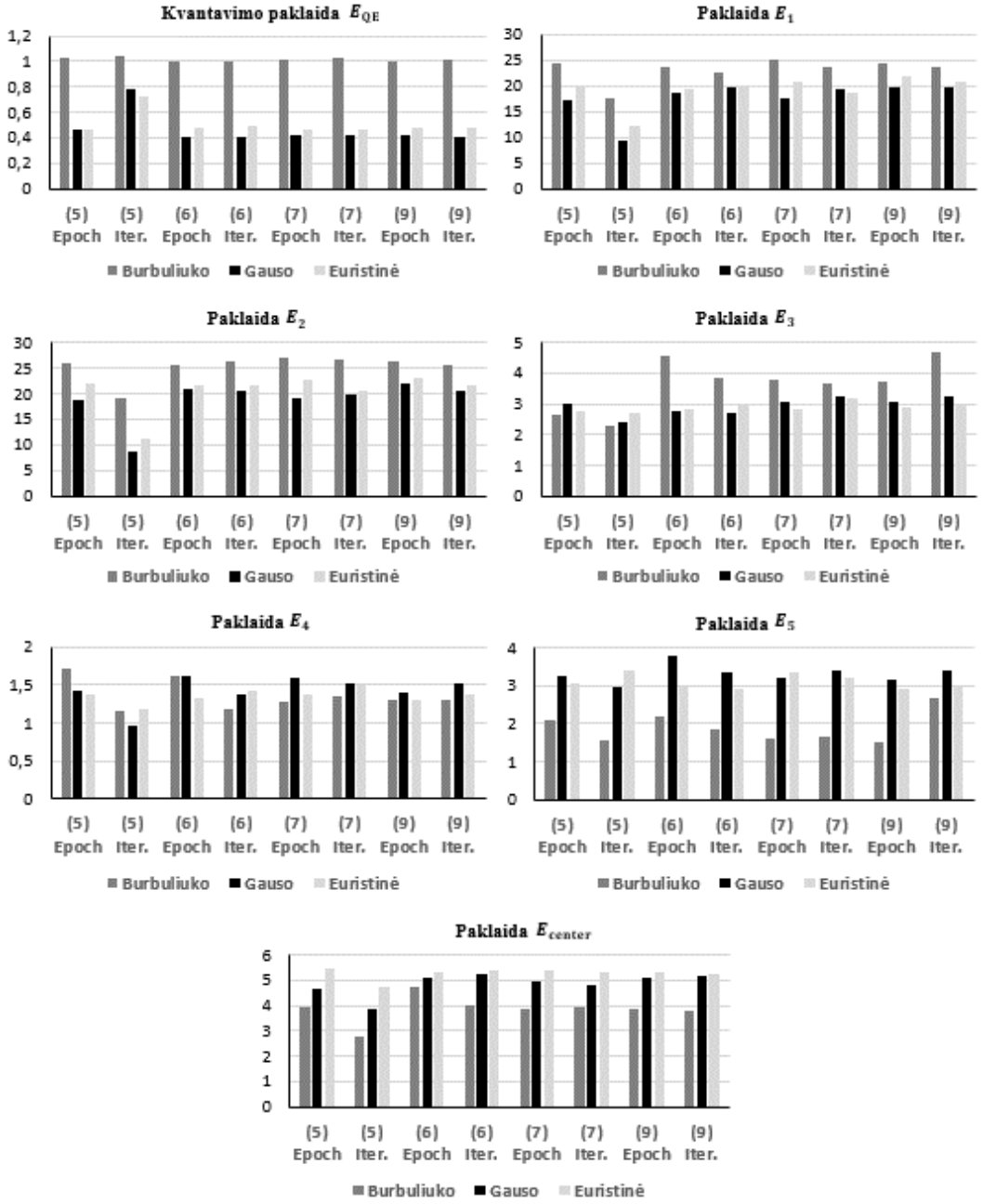
4.8 lentelėje matome, jog skaitinių duomenų atveju daugiausiai balų surinkta, naudojant tiesinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (41 balai). Panašūs rezultatai yra gauti, naudojant laipsninį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (40 balų). Mažiausiai balų surinkta, naudojant euristinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (23 balai).

4.8 lentelė. Apibendrinti 4.3 pav. rezultatai, vertinant mokymo parametrus

Paklaida / **Mokymo parametras**		E_{QE}	E_1	E_2	E_3	E_4	E_5	E_{center}	Iš viso
Tiesinis (5)	Epochos	2	6	7	4	3	3	2	27
	Iteracijos	1	8	8	8	8	7	1	41
Atvirkštinis laikui (6)	Epochos	8	2	1	3	2	6	8	30
	Iteracijos	6	4	3	5	5	1	7	31
Laipsninis (7)	Epochos	4	5	5	6	4	4	4	32
	Iteracijos	3	7	6	7	6	5	6	40
Euristinis (8)	Epochos	5	1	2	2	7	8	3	28
	Iteracijos	7	3	4	1	1	2	5	23

Nagrinėjamų paklaidų vidutinės reikšmės testavimo duomenų aibėms pateikti 4.4 pav. Kvantavimo paklaida E_{QE} daugeliu atveju taip pat mažiausia,

kai naudojama Gauso kaimynystės funkcija. Naudojant burbuliuko kaimynystės funkciją, gaunamos didžiausios E_1, E_2, E_3 reikšmės ir mažiausia E_{center} reikšmė. Naudojant Gauso kaimynystės funkciją, gaunamos mažiausios E_1 ir E_2 reikšmės, tačiau didžiausios E_4 (išskyrus atvejus, kai naudojamas tiesinis (5) ir atvirkštinis laiko (6) mokymo parametrai, keičiant jų reikšmes kiekvienoje epochoje) ir E_5 (išskyrus atvejį, kai naudojamas laipsninis mokymo parametras (7), keičiant jo reikšmę kiekvienoje epochoje).



4.4 pav. SOM kokybę įvertinančių paklaidų vidutinės reikšmės testavimo aibei (stiklo ir zoologinių duomenų aibės)

Naudojant euristinę kaimynystės funkciją, gaunamos didžiausios paklaidos E_{center} reikšmės nepriklausomai nuo pasirinkto mokymo parametro. Kaip ir prieš tai nagrinėtais atvejais, naudojant šią kaimynystės funkciją, gaunami geresni vizualizavimo rezultatai, kadangi skirtingų klasių centrai yra labiau nutolę vienas nuo kito.

4.9 lentelėje matome, jog geresni rezultatai testavimo aibei gaunami, kai naudojama Gauso kaimynystės funkcija (24 kartai), blogiausi, – naudojant burbuliuko kaimynystės funkciją (14 kartų).

4.10 lentelėje matome, kad daugiausiai balų surinkta, naudojant tiesinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (41 balas), o mažiausiai balų surinkta, naudojant euristinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (24 balai).

4.9 lentelė. Apibendrinti 4.4 pav. rezultatai, vertinant kaimynystės funkcijas

Paklaida / **Kaimynystės funkcija**	E_{QE}	E_1	E_2	E_3	E_4	E_5	E_{center}	Iš viso
Burbuliuko	0	0	0	2	4	8	0	14
Gauso	6	7	8	2	1	0	0	24
Euristinė	2	1	0	4	3	0	8	18

4.10 lentelė. Apibendrinti 4.4 pav. rezultatai, vertinant mokymo parametrus

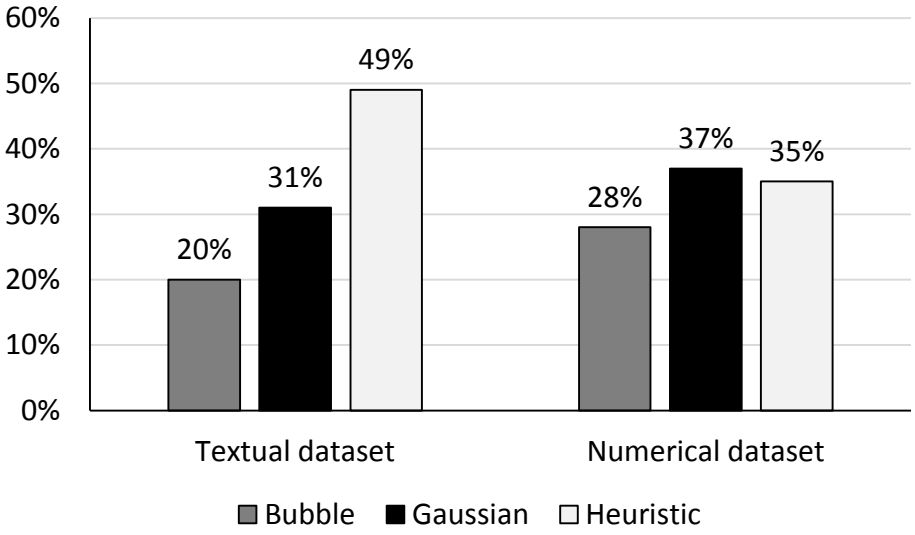
Mokymo parametras / **Paklaida**		E_{QE}	E_1	E_2	E_3	E_4	E_5	E_{center}	Iš viso
Tiesinis (5)	Epochos	2	7	7	7	2	3	3	31
	Iteracijos	1	8	8	8	8	7	1	41
Atvirkštinis laikui (6)	Epochos	8	5	4	2	1	2	8	30
	Iteracijos	5	4	2	6	7	6	7	37
Laipsninis (7)	Epochos	3	3	1	4	4	5	5	25
	Iteracijos	4	6	6	3	3	4	2	28
Euristinis (8)	Epochos	7	1	3	5	6	8	6	36
	Iteracijos	6	2	5	1	5	1	4	24

Gauti SOM paklaidų rezultatai, tiek stiklo ir zoologiniams duomenims, tiek tekstiniams duomenims parodė, jog mažiausia kvantavimo paklaida yra

gaunama, naudojant Gauso kaimynystės funkciją. Ir skaitiniams, ir tekstiniams duomenimis euristinė kaimynystės funkcija leidžia gauti didžiausias paklaidos E_{center} reikšmes. Analizuojant pakaidos E_c reikšmes, sunku pasakyti, kurie mokymo parametrai leidžia gauti mažesnes paklaidos reikšmes, kadangi skirtingai pasirinkti mokymo parametrai pateikia įvairius rezultatus.

4.3.3. Apibendrinti mokymo faktorių rezultatai skaitiniams ir tekstiniams duomenims

Apibendrinti skaitinių ir tekstinių mokymo ir testavimo aibių rezultatai, vertinant SOM mokymo faktorių įtaką, pateikti 4.5 ir 4.6 pav. Procentai, pateikti diagramose, apskaičiuoti susumavus mokymo ir testavimo aibių rezultatus, pateiktus 4.3–4.10 lentelėse.

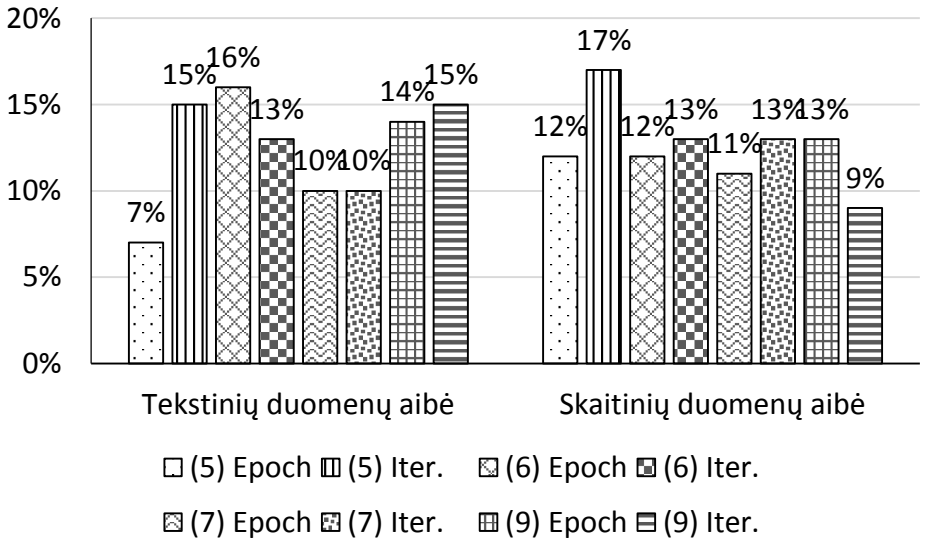


4.5 pav. Apibendrinti 4.1–4.4 pav. rezultatai, vertinant kaimynystės funkcijas

Matome, kad tekstiniams duomenims, 49 % atvejų mažiausios paklaidų reikšmės gautos, naudojant euristinę kaimynystės funkciją, ir tik 20 % atvejų – burbuliuko kaimynystės funkciją. Galima daryti prielaidą, jog mažesnės euristinės kaimynystės funkcijos reikšmės (2.2.2 skyrelis, 2.7 pav.) leidžia gauti geresnius SOM rezultatus nagrinėjant tekstinius duomenis, o didesnės kaimynystės funkcijos reikšmės (burbuliuko kaimynystės funkcija) rezultatus pablogina. Skaitinių duomenų atveju apibendrintos paklaidų reikšmės yra gana

panašios, šiek tiek geresni rezultatai (37 % atvejų) gauti, naudojant Gauso kaimynystės funkciją.

4.6 pav. matome, jog, vertinant SOM rezultatus, atsižvelgiant į naudotą mokymo parametrą (bendri mokymo ir testavimo aibės rezultatai), geriausi rezultatai tekstinių duomenų aibei gauti, naudojant atvirkštinį laikui mokymo parametrą (6), keičiant jo reikšmes kiekvienoje epochoje (16 %), o blogiausi, naudojant tiesinį mokymo parametrą (5), keičiant jo reikšmes kiekvienoje epochoje (7 %). Kaip ir kaimynystės funkcijų atveju, geriausi rezultatai, nagrinėjant tekstinius duomenis, gaunami tuomet, kai naudojamas mokymo parametras, kurio įgyjamos reikšmės yra mažiausios (2.2.2 skyrelis, 2.5 pav.), o blogiausi, – naudojant tiesinį mokymo parametrą (mokymo parametro reikšmės didžiausios, lyginant su kitais mokymo parametrais). Skaitinių duomenų atveju geriausi rezultatai gauti, naudojant tiesinį mokymo parametrą (5), keičiant jo reikšmes kiekvienoje iteracijoje (17 %), o blogiausi, – naudojant euristinį mokymo parametrą (9), keičiant jo reikšmes kiekvienoje iteracijoje (9 %). Nagrinėjant skaitinius duomenis atvirkščiai nei tekstinių duomenų atveju, geresni rezultatai gaunami tuomet, kai mokymo parametro reikšmės yra didžiausios (tiesinis mokymo parametras).



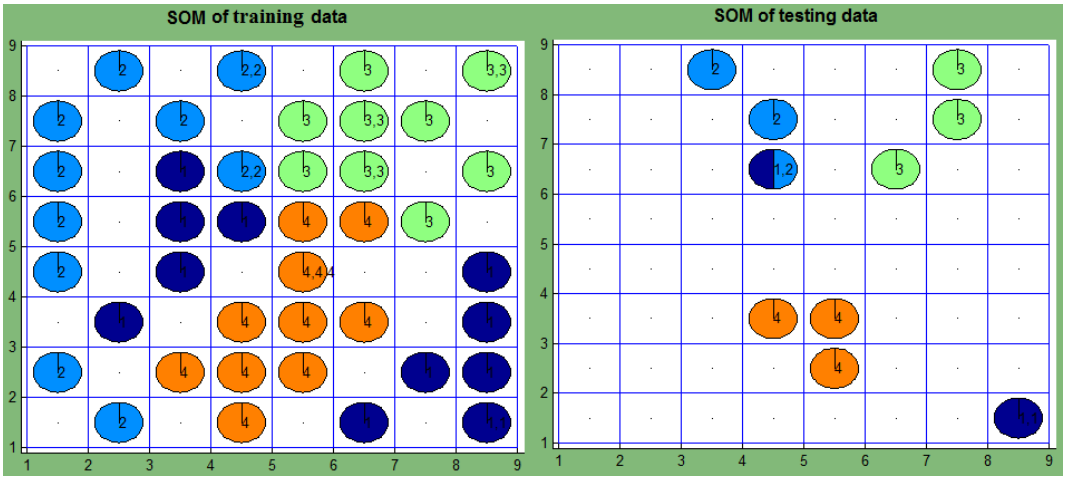
4.6 pav. Apibendrinti 4.1–4.4 pav. rezultatai, vertinant mokymo parametrus

4.4. Teksto dokumentų konvertavimo į skaitinę išraišką faktorių įtaka

Teksto analizė gali būti taikoma įvairiose srityse, tokiose kaip bankininkystė, draudimo įmonės, finansinių tyrimų institucijos, užklausų internete semantinė analizė ir pan. Taip pat šiais laikais internete talpinami dideli kiekiai mokslinių straipsnių. Interneto paieškos įrankiai leidžia rasti reikiamus straipsnius gana greitai, tačiau neretai informacija juose būna ne tokia naudinga, kokios tikėjomės. Įprastai mokslinių straipsnių yra ieškoma pagal straipsnio pavadinimą arba raktinius žodžius, tačiau problema iškyla tuomet, kai norime rasti kitų straipsnių, panašių į tam tikrą konkretų straipsnį. Kitas galimas būdas (nėra dažnai naudojamas) straipsniams grupuoti — klasterizavimo metodai, kurie panašius straipsnius priskiria vienam klasteriui. Tuo tikslu šiame tyrime panašiems moksliniams straipsniams klasterizuoti ir vizualizuoti panaudotas saviorganizuojantis neuroninis tinklas. Kaip jau buvo minėta, didelę įtaką teksto dokumentų žodynui ir teksto dokumentų matricai sukurti turi pasirinkti pradiniai faktoriai. Tolimesniame tyrime mokymo aibę sudaro 80 % visų duomenų, likusi dalis – testavimo aibė. Pasirinktas SOM dydis: 8 eilutės ir 8 stulpeliai ($k_x = k_y = 8$).

Daroma prielaida, kad, tiriant skirtingų sričių mokslinius straipsnius, juos atitinkantys klasteriai turi nepersidengti (klasterius turi sudaryti tik tos pačios klasės duomenys) ir būti aiškiai matomi SOM žemėlapyje. Tai puikiai iliustruoja eksperimentas, atliktas naudojant duomenų aibę „Moksliniai straipsniai I“. Duomenų aibę konvertuojant į skaitinę išraišką buvo pasirinkti tokie faktoriai, kurie yra fiksuoti ir nekeičiami tyrimo metu: atmesti skaitmenys, minimalus žodžių ilgis ir žodžių pasikartojimas nemažesnis kaip 3 ir įtrauktas „Text to Matrix Generator“ (TMG) sistemos siūlomas dažniausiai naudojamų žodžių sąrašas. Tokiu atveju matricą sudaro 60 eilučių ir 2368 stulpeliai ($m = 60, n = 2368$). Kai matome 4.7 pav., dauguma tos pačios klasės duomenų sudaro atskirus klasterius, tik kai kurie I ir II klasės testavimo aibės duomenys persidengia. Visi III klasės duomenys (straipsniai apie optimizavimą) ir IV klasės duomenys (straipsniai apie SOM) sudaro atskirus klasterius, ypač tai

matosi testavimo duomenų aibės atveju. Kai kurie I klasės (straipsniai apie dirbtinius neuroninius tinklus) ir II klasės (straipsniai apie bioinformatiką) nariai tarpusavyje susimaišo, t. y. dalis II klasės yra arčiau I klasės narių, tačiau tai yra natūralu, kadangi šios sritys tarpusavyje iš dalies susijusios, nes abejose galima rasti informacijos apie neuronus (natūralius ar dirbtinius).



a) **b)**

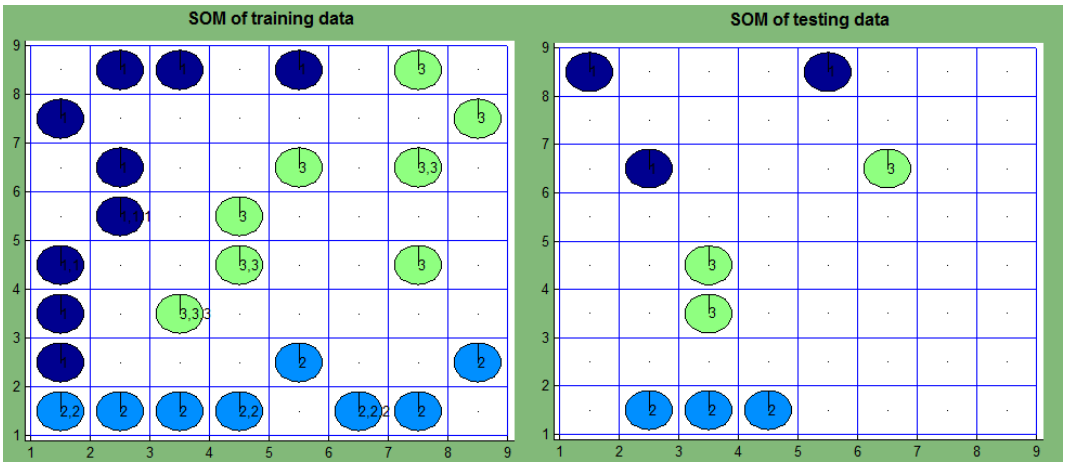
4.7 pav. SOM žemėlapiai, gauti naudojant duomenų aibę „Moksliniai straipsniai I“: a) mokymo aibė, b) testavimo aibė

Tam, kad pamatytume tam tikras tendencijas, kaip pasirinkti teksto konvertavimo į skaitinę išraišką faktoriai lemia gautus rezultatus, detalesniam tyrimui buvo naudoti duomenys, kurie yra tarpusavyje panašūs (panašios tyrimo sritys). Panašių dokumentų analizė leidžia pamatyti, kurie konkrečiai pasirinkti faktoriai turi didesnę įtaką gautiems SOM rezultatams. Tyrimas atliktas naudojant duomenų aibę „Moksliniai straipsniai II“. Kaip jau buvo minėta, teksto dokumentų žodyną galima sukurti rankiniu ir automatiniu būdu, todėl tolimesniame tyrime buvo nagrinėjami abu būdai.

4.4.1. Rankinis teksto dokumento žodyno sukūrimas

Pirmu atveju sudaryta teksto dokumentų matrica atsižvelgiant į žodyną, kuriame yra tik trys žodžiai: „simplex“, „Pareto“ ir „genetic“ (I rankiniu būdu sukurtas žodynas). Šiuo atveju matricą sudaro 45 eilutės ir 3 stulpeliai ($m = 45, n = 3$). Kaip matome 4.8 pav., SOM gerai atskiria skirtingų klasių narius: a) žemėlapyje pavaizduoti mokymo aibės rezultatai, o b) – testavimo aibės

rezultatai. I klasės nariai (straipsniai apie simplekso metodą) išsidėstę žemėlapio kairėje pusėje (tamsiai mėlyna spalva), apačioje – II klasės nariai (straipsniai apie genetinius algoritmus, šviesiai mėlyna spalva) ir dešinėje III klasės nariai (straipsniai apie Pareto, šviesiai žalia spalva).



a) **b)**

4.8 pav. SOM žemėlapiai (teksto dokumentų matrica sukurta naudojant rankinį žodyną I): a) mokymo aibė, b) testavimo aibė

4.11–4.12 lentelėse pateiktos gautų SOM žemėlapių (4.7–4.15 pav.) skaitiniai įverčiai. Pateikta kvantavimo paklaida E_{QE} ir naujai pasiūlytų paklaidų reikšmės E_1, E_2, E_3 ir E_{center}. Mokymo aibės rezultatai pateikti 4.11 lentelėje, o 4.12 lentelėje – testavimo aibės rezultatai.

4.11 lentelė. Eksperimentinio tyrimo rezultatai mokymo aibei

Nr.	**Eksperimentas**	E_{QE}	E_1	E_2	E_3	E_{center}
1	Rankinis žodynas I, $n = 3$	2,26	2,84	2,81	3,11	4,01
2	Rankinis žodynas II, $n = 15$	7,55	4,25	2,30	2,38	4,02
3	Nenaudojant dažniausiai vartojamų žodžių sąrašo, $n = 3441$	96,18	4,49	4,13	4,74	1,57
4	Naudojant TMG įrankio sudarytą dažniausiai vartojamų žodžių sąrašą, $n = 3198$	77,12	2,86	4,67	4,68	1,64
5	Naudojant naują dažniausiai vartojamų žodžių sąrašą, $n = 3157$	69,19	3,25	4,00	2,91	3,40

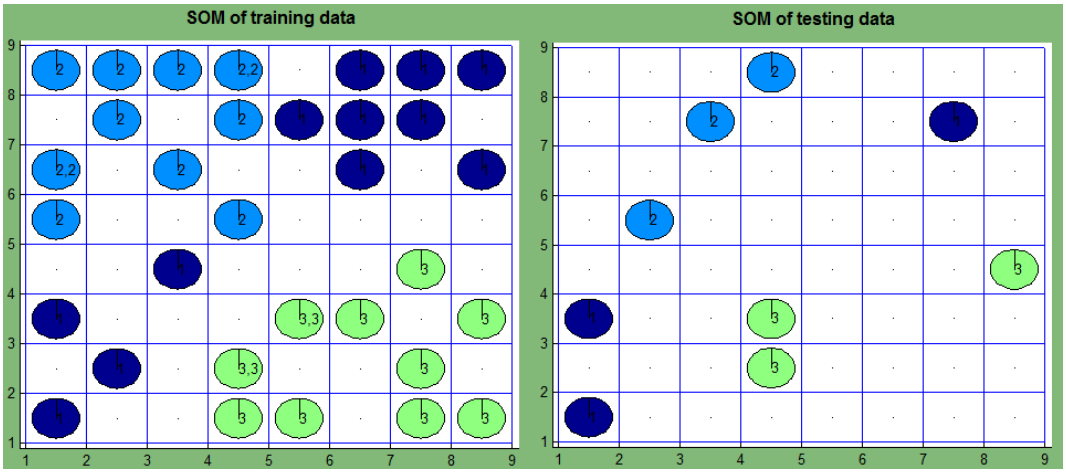
6	Nenaudojant dažniausiai vartojamų žodžių sąrašo, bet įtraukus kamieno išskyrimo algoritmą, $n = 2685$	105,76	4,59	3,81	3,71	2,32
7	Naudojant TMG įrankio sudarytą dažniausiai vartojamų žodžių sąrašą ir kamieno išskyrimo algoritmą, $n = 2486$	88,47	3,57	4,07	4,84	2,05
8	Naudojant naują dažniausiai vartojamų žodžių sąrašą ir kamieno išskyrimo algoritmą, $n = 2471$	82,70	3,48	3,95	4,60	2,75

4.12 lentelė. Eksperimentinio tyrimo rezultatai testavimo aibei

Nr.	**Eksperimentas**	E_{QE}	E_1	E_2	E_3	E_{center}
1	Rankinis žodynas I, $n = 3$	2,92	3,28	1,33	2,95	4,37
2	Rankinis žodynas II, $n = 15$	14,45	5,90	2,42	3,20	3,26
3	Nenaudojant dažniausiai vartojamų žodžių sąrašo, $n = 3441$	143,68	2,75	3,24	4,90	1,76
4	Naudojant TMG įrankio sudarytą dažniausiai vartojamų žodžių sąrašą, $n = 3198$	122,06	1,33	0,83	1,94	2,40
5	Naudojant naują dažniausiai vartojamų žodžių sąrašą, $n = 3157$	117,46	1,05	0,67	3,48	3,13
6	Nenaudojant dažniausiai vartojamų žodžių sąrašo, bet įtraukus kamieno išskyrimo algoritmą, $n = 2685$	155,64	3,70	3,41	4,47	1,63
7	Naudojant TMG įrankio sudarytą dažniausiai vartojamų žodžių sąrašą ir kamieno išskyrimo algoritmą, $n = 2486$	137,27	0,67	1,80	5,34	2,69
8	Naudojant naują dažniausiai vartojamų žodžių sąrašą ir kamieno išskyrimo algoritmą, $n = 2471$	134,03	0,83	3,82	1,14	2,27

Mažesnės pasiūlytų paklaidų E_1, E_2, E_3 reikšmės reiškia geresnius SOM rezultatus (tais atvejais tos pačios klasės nariai SOM žemėlapyje yra arčiau vienas kito), o paklaidos E_{center} rezultatai geresni, kai reikšmė yra didesnė (tais atvejais skirtingų klasių centrai yra toliau vienas nuo kito).

Kitame tyrime teksto dokumentų žodynas papildytas žodžiais, apibūdinančiais skirtingas klases. Į žodyną įrašyti šie žodžiai: „simplex“, „programming“, „convex“, „corner“, „vertices“, „genetic“, „mutation“, „crossover“, „chromosome“, „fitness“, „Pareto“, „multiobjective“, „front“, „dominate“, „decision“. Šiuo atveju matricą sudarys 45 eilutės ir 15 stulpelių ($m = 45, n = 15$). Gautame SOM žemėlapyje (4.9 pav.) matome, jog klasės gerai atsiskiria ir sudaro kelis aiškius skirtingus klasterius. Tačiau šiuo atveju I klasės nariai susiskirsto į du mažesnius klasterius, galbūt parinkti žodyno žodžiai ne visiškai tiksliai atspindi straipsnių esmę. Palyginti SOM žemėlapių, pateiktų 4.8–4.9 pav., rezultatus neįmanoma tik kvantavimo paklaidos prasme, kadangi kiekvienu atveju vektorių matmenų erdvė yra skirtinga (priklauso nuo žodžių žodyne skaičiaus). Kuo didesnis žodžių skaičius n, tuo didesnė E_{QE} reikšmė. Todėl toliau SOM rezultatai lyginami pagal pasiūlytas paklaidas E_{c} ir E_{center}.



a) **b)**

4.9 pav. SOM žemėlapiai (teksto dokumentų matrica sukurta naudojant rankinį žodyną II): a) mokymo aibė, b) testavimo aibė

Palyginus dviejų gautų SOM žemėlapių (4.11–4.12 lenteles, Nr. 1–2) rezultatus matome, jog, naudojant II rankiniu būdu sudarytą žodyną, mokymo aibei paklaidos E_1 reikšmė ir testavimo aibei E_1, E_2, E_3 reikšmės yra didesnės, lyginant su rezultatais, gautais atliekant eksperimentą su I rankiniu būdu sukurtu žodynu, o paklaidos E_{center} reikšmė yra mažesnė testavimo aibei, tačiau nežymiai didesnė mokymo aibei. Galima daryti išvadą, jog abu žodynai yra tinkami, kadangi rezultatai taikytų paklaidų prasme yra panašūs, o skirtumai nėra reikšmingi.

4.4.2. Automatinis teksto dokumento žodyno sukūrimas

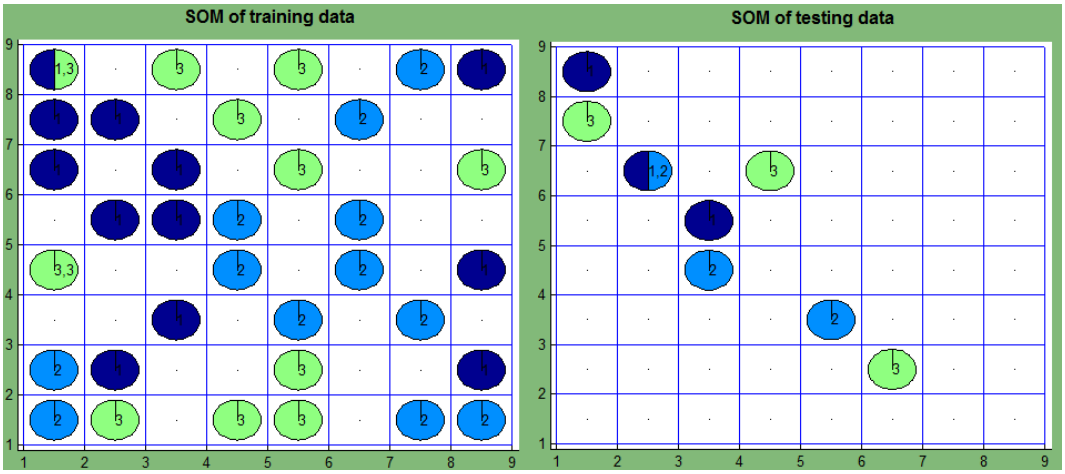
Rankinis žodyno sudarymo būdas gali būti naudojamas tuomet, kai iš anksto turime pradinės informacijos apie analizuojamą teksto duomenų aibę. Kitas būdas sudaryti žodyną – jį kurti automatiškai, atsižvelgiant į iš anksto pasirinktus faktorius, kurie aprašyti 2.1 poskyryje. Į teksto dokumentų žodyną įtraukiami tik tiek žodžiai, kurie tenkina pradinius faktorius. Žodynas sudaromas taip, kad būtų tenkinami visi pasirinkti faktoriai. Kitame tyrime nagrinėjama, kaip dažniausiai vartojamų žodžių sąrašo įtraukimas ir kamieno išskyrimo algoritmas įtakoja SOM rezultatus.

4.4.2.1. Dažniausiai vartojamų žodžių sąrašas

Dažniausiai vartojamų žodžių sąrašas – tai rinkinys žodžių, kurie neturi būti įtraukti į teksto dokumento žodyną. Įtraukus į šį sąrašą nereikšmingus žodžius, sumažinamas teksto dokumentų žodyno dydis, kartu sumažėja ir teksto dokumentų matricos stulpelių skaičius n. Taip pat išvengiama tų žodžių, kurie nėra esminiai ir neatspindi dokumentų prasmės. Todėl svarbu tinkamai parinkti šiuos žodžius. Analizuojant skirtingų sričių dokumentus, žodžių sąrašas turėtų būti parenkamas atsižvelgiant į duomenų aibę. Šiame tyrime taip pat atmesti skaitmenys bei žodžių pasikartojimas ne mažesnis kaip 3 kartai. Tirti keli atvejai:

1. **Dažniausiai vartojamų žodžių sąrašas nėra naudojamas.** Visų pirma, atliktas eksperimentas, kai automatiškai sudarant teksto dokumento žodyną dažniausiai vartojamų žodžių sąrašas nėra naudojamas. Šiuo

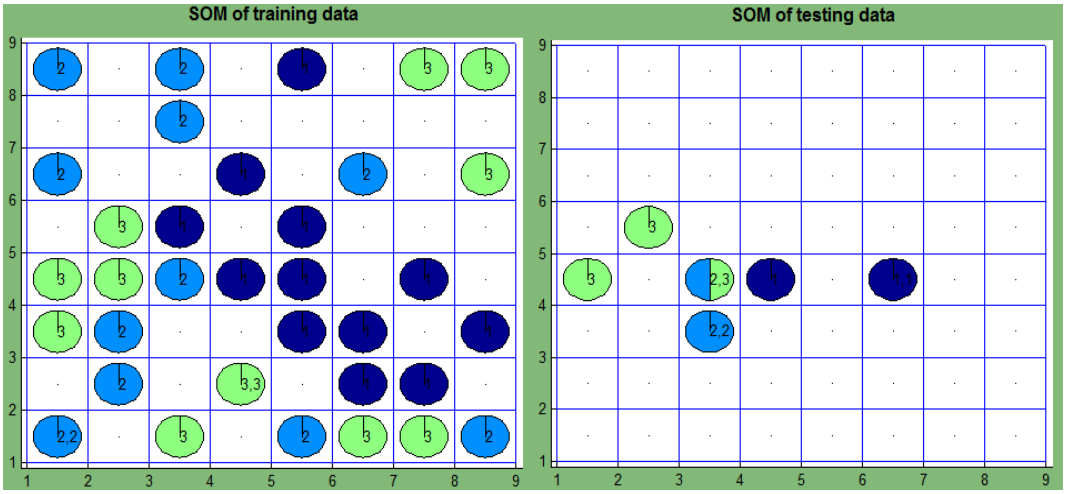
atveju teksto dokumentų matricą sudaro 45 eilutės (teksto dokumentų aibėje esantis dokumentų skaičius) ir 3441 stulpelis (teksto dokumento žodyne esantis žodžių skaičius) ($m = 45,\ n = 3441$) (naudojami duomenys „Moksliniai straipsniai II“). Gautame SOM žemėlapyje (4.10 pav.) sudėtinga įžvelgti aiškius klasterius, kadangi vektoriai, atitinkantys straipsnius, yra „išsimėtę“ po visą SOM žemėlapį. Akivaizdu, jog sudarant žodyną, nenaudojant dažniausiai vartojamų žodžių sąrašo, į teksto dokumento žodyną įtraukiama daug neesminių žodžių, kurie bendri ir visiškai neatspindi dokumento esmės, todėl tokį sąrašą sudaryti ir į jį atsižvelgti būtina. 4.11–4.12 lentelėse (Nr. 3) matome, jog kvantavimo paklaida E_{QE} yra žymiai didesnė nei gauta atlikus eksperimentus Nr. 1–2, kadangi šiuo atveju vektorių matmenų skaičius yra didesnis. Akivaizdu, kad pasiūlytų paklaidų reikšmės blogesnės (didesnės E_1, E_2, E_3, mažesnės E_{center}), lyginant su atvejais, kai teksto dokumento žodynas sudarytas rankiniu būdu, išskyrus tik reikšmę E_1 testavimo aibei (4.12 lentelę, Nr. 1–3), kadangi šiuo atveju dažniausiai vartojamų žodžių gali būti visose tekstinių dokumentų (mokslinių straipsnių) klasėse, o tai neleidžia klasėms tinkamai susigrupuoti SOM žemėlapyje.



a) **b)**

4.10 pav. SOM žemėlapiai (žodynas sukurtas neįtraukiant dažniausiai vartojamų žodžių sąrašo): a) mokymo aibė, b) testavimo aibė

2. **Atsižvelgiama į TMG įrankiu sudarytą dažniausiai vartojamų žodžių sąrašą.** Kitame eksperimente sudarant žodyną įtrauktas dažniausiai vartojamų žodžių sąrašas, sudarytas TMG įrankiu. Šis sąrašas turi daugiau nei 300 žodžių (įtraukti tokie trumpi žodžiai kaip: „such“, „there“, „where“, „here“, „some“ ir t. t.). Šie žodžiai nėra įtraukiami į teksto dokumento žodyną. Šiuo atveju matricą sudaro 45 eilutės ir 3198 stulpeliai ($m = 45$, $n = 3198$). Gautame SOM žemėlapyje (4.11 pav.) duomenys yra labiau susigrupavę nei 4.10 pav., tačiau ryškiausias klasteris yra tik I klasės narių. 4.11–4.12 lentelėse (Nr. 3–4) matome, jog rezultatai pasiūlytų paklaidų prasme yra geresni, kai naudojame TMG įrankio dažniausiai vartojamų žodžių sąrašą, nei kai joks žodžių sąrašas visai nėra įtraukiamas sudarant teksto dokumentų žodyną. Rezultatas blogesnis tik mokymo aibei paklaidos E_2 prasme. Atsižvelgus į dažniausiai vartojamų žodžių sąrašą, sudarytame žodyne yra atskiras dokumentų klases charakterizuojančių žodžių, todėl tai leidžia duomenų klasėms labiau susiklasterizuoti SOM žemėlapyje.

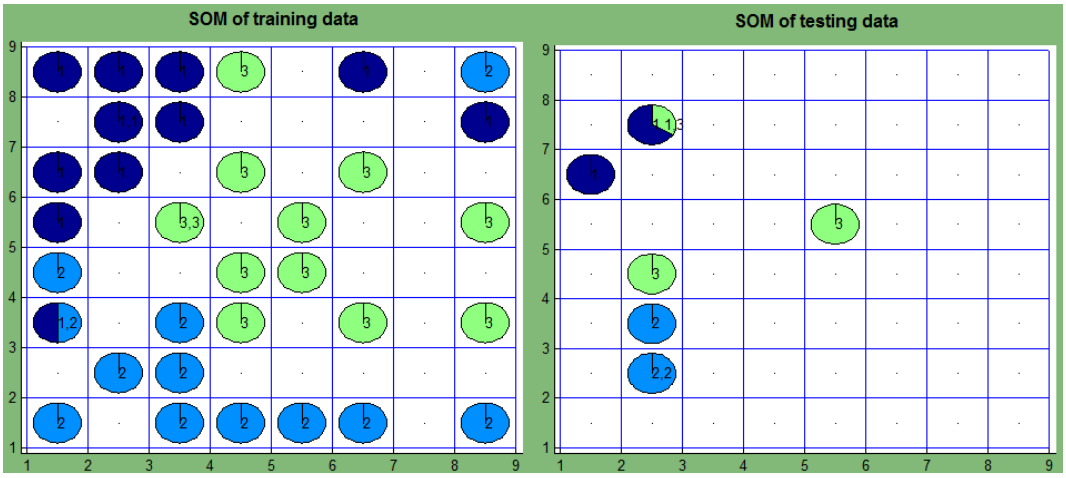


a) **b)**

4.11 pav. SOM žemėlapiai (teksto dokumentų matrica sukurta įtraukiant TMG dažniausiai vartojamų žodžių sąrašą): a) mokymo aibė, b) testavimo aibė

3. **Naujo dažniausiai vartojamų žodžių sąrašo įtraukimas.** TMG įrankio sudarytame dažniausiai vartojamų žodžių sąraše yra tik bendri žodžiai ar jungtukai, tačiau jis nėra pritaikytas moksliniams straipsniams. Taigi,

atsižvelgiant į dažnai pasikartojančius žodžius moksliniuose straipsniuose, sąrašas papildytas naujais žodžiais: „function“, „fig“, „table“, „formula“, „optimization“, „present“, „minimum“, „maximum“, „function“, „variable“ ir pan. Atsižvelgiant į žodyną, sudaryta matrica, kurios dydis 45 eilutės ir 3157 stulpeliai ($m = 45, n = 3157$). Gautame SOM žemėlapyje (4.12 pav.) klasteriai matomi kur kas aiškiau nei 4.10–4.11 pav. Žemėlapio centre išsidėstė III klasės nariai, dauguma II klasės narių atsidūrė kairiajame apatiniame kampe, o I klasės nariai išsidėstė kairiajame viršutiniame kampe. Keletas I ir II klasės narių atsidūrė dešiniajame viršutiniame kampe. Pažiūrėję į gautus pasiūlytų paklaidų rezultatus (4.11–4.12 lentelėse, Nr. 4–5) matome, jog rezultatai yra geresni, nei kai naudojame TMG įrankio sąrašą, išskyrus paklaidos E_3 reikšmes testavimo aibės duomenims. Vadinasi, dažniausiai vartojamų žodžių sąrašą papildžius žodžiais, kurie yra dažni moksliniuose straipsniuose, tačiau jų necharakterizuoja, sudarytas žodynas labiau atspindi dokumentų klases, todėl SOM žemėlapyje jos labiau klasterizuojasi.



a) **b)**

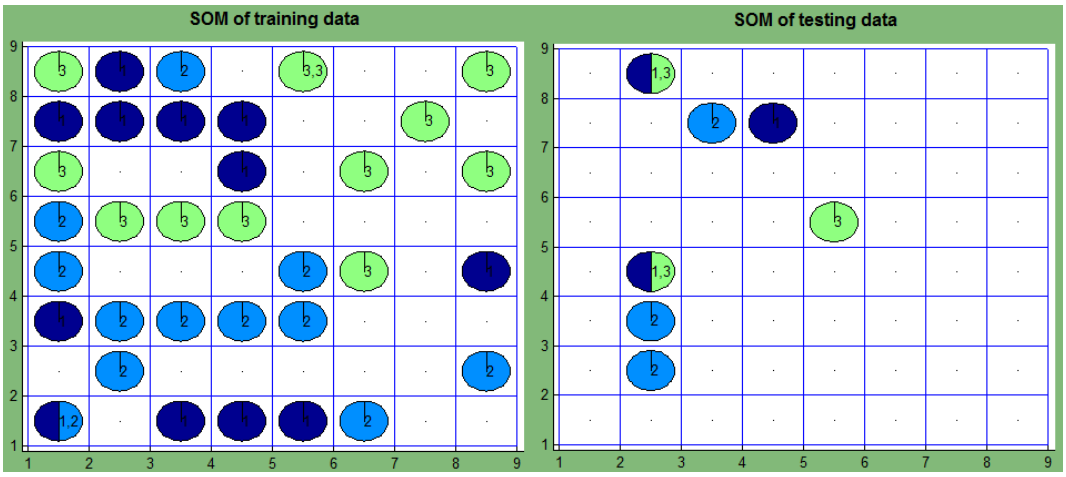
4.12 pav. SOM žemėlapiai (teksto dokumentų matrica sukurta įtraukiant naują dažniausiai vartojamų žodžių sąrašą): a) mokymo aibė, b) testavimo aibė

4.4.2.2. Kamieno išskyrimo algoritmas

Kamieno išskyrimo algoritmas leidžia atskirti žodžio kamieną ir tik jis yra įtraukiamas į žodyną. Šiame tyrime naudotas Porterio kamieno išskyrimo algoritmas, skirtas anglų kalbai (Porter, 1980). 4.13–4.15 pav. pavaizduoti SOM žemėlapiai, kai kamieno išskyrimo algoritmas yra panaudotas teksto dokumentų žodynui sudaryti. Paklaidų įverčiai yra pateikti 4.11–4.12 lentelėse, Nr. 6–8. Palyginę rezultatus, kai kamieno išskyrimo algoritmas yra panaudotas ir kai nenaudotas, matome:

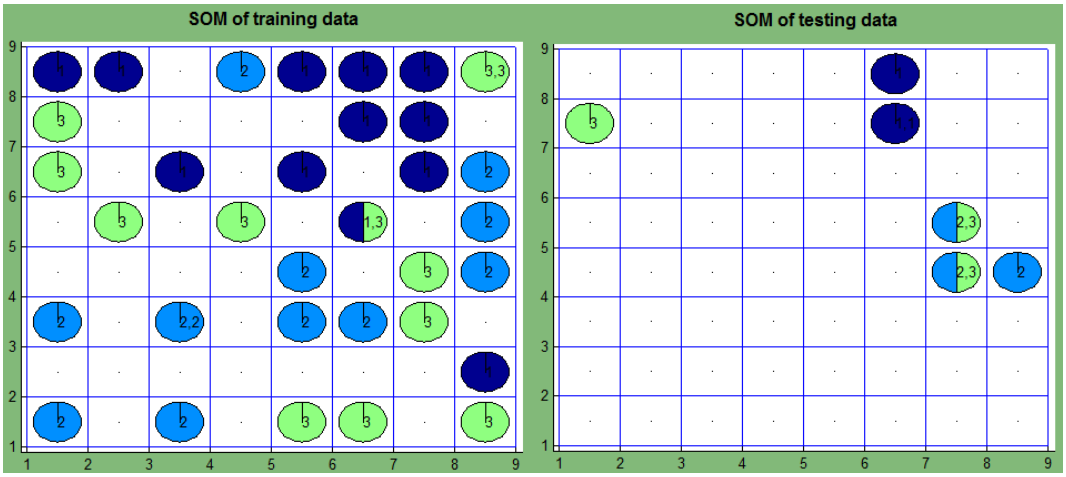
- Nors naudojant kamieno išskyrimo algoritmą vektorių matmenų skaičius n yra mažesnis, lyginant su atveju, kai šis algoritmas nėra naudojamas, tačiau kvantavimo paklaida E_{QE} yra didesnė, tai reiškia, kad šiuo atveju rezultatas gaunamas blogesnis, kai yra naudojamas kamieno iškyrimo algoritmas.
- Jeigu nėra įtraukiamas nei vienas dažniausiai vartojamų žodžių sąrašas, tuomet kamieno išskyrimo algoritmas mokymo aibei pagerina pasiūlytų paklaidų rezultatus, išskyrus paklaidos E_1. Testavimo aibės atveju paklaidos reikšmė E_3 yra geresnė, kai naudojamas dažniausiai vartojamų žodžių sąrašas ir kamieno išskyrimo algoritmas, bet kitais atvejais $E_1, E_2, E_{\mathrm{center}}$ rezultatas blogesnis. Vadinasi, šiuo atveju kamieno išskyrimas nepagerina SOM kokybės nagrinėjamų paklaidų prasme.
- Jeigu įtraukiamas TMG įrankio sudarytas dažniausiai vartojamų žodžių sąrašas, tuomet pusė rezultatų pagerėja tiek mokymo, tiek testavimo aibei, tačiau kita dalis pablogėja.
- Jeigu įtraukiamas naujas dažniausiai vartojamų žodžių sąrašas, tuomet kamieno išskyrimo algoritmas pablogina visus SOM rezultatus, išskyrus paklaidos E_2 mokymo aibei ir paklaidų E_1, E_3 testavimo aibei.

Galima padaryti išvadą, kad neįmanoma aiškiai nusakyti, ar kamieno išskyrimo algoritmas pagerina ar pablogina rezultatus šiai nagrinėjamai duomenų aibei, kadangi kai kuriais atvejais rezultatai pagerėja, tačiau neretai ir pablogėja, galbūt reikia sukurti naujus arba panaudoti kitus kamieno išskyrimo algoritmus.



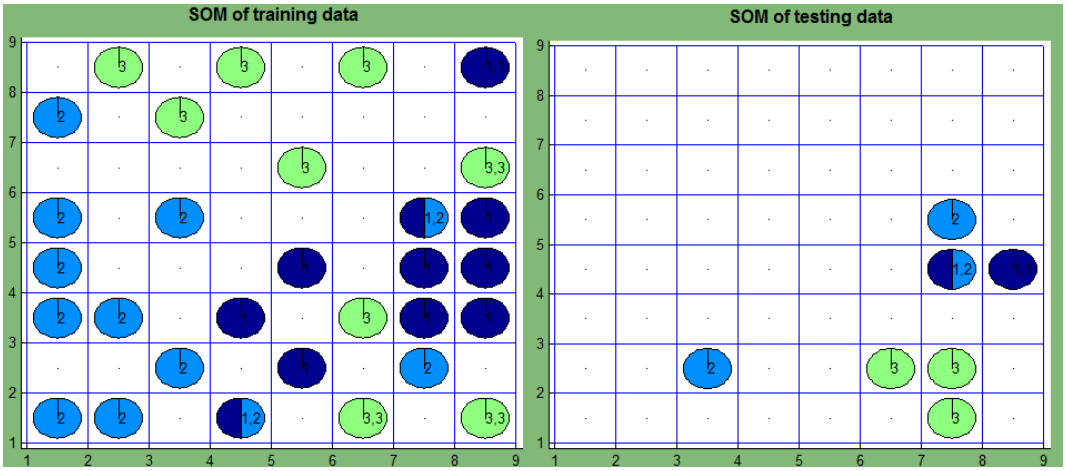
a) **b)**

4.13 pav. SOM žemėlapiai (teksto dokumentų matrica sukurta neįtraukiant dažniausiai vartojamų žodžių sąrašo, bet naudojant kamieno išskyrimo algoritmą): a) mokymo aibė, b) testavimo aibė



a) **b)**

4.14 pav. SOM žemėlapiai (teksto dokumentų matrica sukurta įtraukiant TMG įrankio dažniausiai vartojamų žodžių sąrašą ir naudojant kamieno išskyrimo algoritmą): a) mokymo aibė, b) testavimo aibė



a) **b)**

4.15 pav. SOM žemėlapiai (teksto dokumentų matrica sukurta įtraukiant naują dažniausiai vartojamų žodžių sąrašą ir naudojant kamieno išskyrimo algoritmą): a) mokymo aibė, b) testavimo aibė

4.4.3. Apibendrinti teksto konvertavimo į skaitinę išraišką įtakos rezultatai

Apibendrinant gautus eksperimentinio tyrimo rezultatus, kiekvieno eksperimento metu (4.13–4.14 lentelės) gautos paklaidų reikšmės yra įvertintos balais nuo 1 iki 8: geriausias iš visų 8 eksperimentų rezultatas įvertintas 8 balais, blogiausias, – 1. Kvantavimo paklaida nėra vertinama taškais, kadangi skirtingos duomenų aibių matmenų skaičius neleidžia gautų kvantavimo paklaidų vertinti tarpusavyje (kuo duomenų matmenų skaičius mažesnis, tuo kvantavimo paklaida mažesnė). Paskutiniuose lentelių stulpeliuose pateikta balų suma. Kuo ji didesnė, tuo eksperimentas laikomas geresniu.

Kaip matome 4.13 lentelėje, geriausi rezultatai mokymo aibei gauti, naudojant rankinį žodyną I (28 balai) ir rankinį žodyną II (27 balai), kadangi į žodyną įtraukiami tik straipsnius charakterizuojančius žodžius, todėl SOM žemėlapyje klasės gerai klasterizuojasi. Palyginus rezultatus, kai teksto dokumentų žodynas sudaromas automatiškai, geriausias rezultatas gautas, kai yra naudotas naujas dažniausiai vartojamų žodžių sąrašas (23 balai), blogiausias – kai nėra naudotas nei dažniausiai vartojamų žodžių sąrašas, nei kamieno išskyrimo algoritmas (7 balai).

4.13 lentelė. Apibendrinti 4.11 lentelės rezultatai

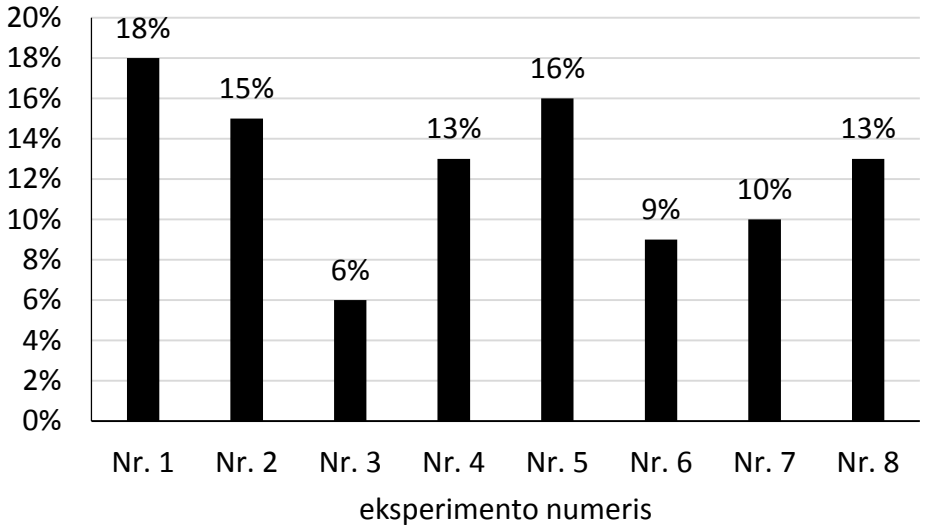
Nr.	**Eksperimentas**	E_1	E_2	E_3	E_{center}	Iš viso
1	Rankinis žodynas I, $n = 3$	8	7	6	7	28
2	Rankinis žodynas II, $n = 15$	3	8	8	8	27
3	Nenaudojant dažniausiai vartojamų žodžių sąrašo, $n = 3441$	2	2	2	1	7
4	Naudojant TMG įrankio sudarytą dažniausiai vartojamų žodžių sąrašą, $n = 3198$	7	1	3	2	13
5	Naudojant naują dažniausiai vartojamų žodžių sąrašą, $n = 3157$	6	4	7	6	23
6	Nenaudojant dažniausiai vartojamų žodžių sąrašo, bet įtraukus kamieno išskyrimo algoritmą, $n = 2685$	1	6	5	4	16
7	Naudojant TMG įrankio sudarytą dažniausiai vartojamų žodžių sąrašą ir kamieno išskyrimo algoritmą, $n = 2486$	4	3	1	3	11
8	Naudojant naują dažniausiai vartojamų žodžių sąrašą ir kamieno išskyrimo algoritmą, $n = 2471$	5	5	4	5	19

Testavimo aibei geriausi rezultatai gauti (4.14 lentelė), kai naudotas naujas dažniausiai vartojamų žodžių sąrašas (24 balai). Panašus rezultatas gautas, naudojant rankinį žodyną I (23 balai). Blogiausias rezultatas, – nenaudojant dažniausia vartojamų žodžių sąrašo ir įtraukus kamieno išskyrimo algoritmą (8 balai). Bendri mokymo aibės (4.13 lentelė) ir testavimo aibės (4.14 lentelė) rezultatai pateikti 4.16 pav.

4.14 lentelė. Apibendrinti 4.12 lentelės rezultatai

Nr.	**Eksperimentas**	E_1	E_2	E_3	E_{center}	Iš viso
1	Rankinis žodynas I, $n = 3$	3	6	6	8	23
2	Rankinis žodynas II, $n = 15$	1	4	5	7	17
3	Nenaudojant dažniausiai vartojamų žodžių sąrašo, $n = 3441$	4	3	2	2	11

4	Naudojant TMG įrankio sudarytą dažniausiai vartojamų žodžių sąrašą, $n = 3198$	5	7	7	4	23
5	Naudojant naują dažniausiai vartojamų žodžių sąrašą, $n = 3157$	6	8	4	6	24
6	Nenaudojant dažniausiai vartojamų žodžių sąrašo, bet įtraukus kamieno išskyrimo algoritmą, $n = 2685$	2	2	3	1	8
7	Naudojant TMG įrankio sudarytą dažniausiai vartojamų žodžių sąrašą ir kamieno išskyrimo algoritmą, $n = 2486$	8	5	1	5	19
8	Naudojant naują dažniausiai vartojamų žodžių sąrašą ir kamieno išskyrimo algoritmą, $n = 2471$	7	1	8	3	19



4.16 pav. Bendri 4.13–4.14 lentelių rezultatai

4.16 pav. diagramose pateikti procentai apskaičiuoti susumavus 4.13 ir 4.14 lentelėse pateiktus rezultatus. Kaip matome, geriausi rezultatai gaunami, naudojant rankinį žodyną I (18 %). Panašūs rezultatai yra gaunami, naudojant naują dažniausiai vartojamų žodžių sąrašą (16 %) bei rankinį žodyną II (15 %). Blogiausi rezultatai, – nenaudojant dažniausiai vartojamų žodžių sąrašo (6 %).

Taigi, kai yra žinoma išankstinė informacija apie analizuojamus duomenis, geriausia sudaryti teksto dokumentų žodyną rankiniu būdu. Sudarant teksto dokumentų žodyną automatiškai, patartina sukurti dažniausiai vartojamų žodžių sąrašą, į jį įtraukiant pagrindinius analizuojamos duomenų aibės srities žodžius, kurie gali iškraipyti gautus galutinius rezultatus, ir į teksto dokumentų žodyną neįtraukti žodžių iš šio sąrašo.

4.5. Žodžių pasikartojimų skaičiaus įtaka SOM rezultatams

Šiame tyrime nagrinėjama žodžių pasikartojimų skaičiaus įtaka SOM rezultatams, sudarant teksto dokumento žodyną. Tyrimams atlikti naudota duomenų aibė „Ministerijų įsakymai“. Atlikus eksperimentus pastebėta, kad tiriamai dokumentų aibei tinkamiausias maksimalus žodžių pasikartojimų skaičius dokumente nemažesnis nei 5. Pasirinkus didesnį skaičių, sudarant tekstinių dokumentų žodyną, dalis dokumentų buvo atmetami, kadangi juose tiek kartų pasikartojančių žodžių nebuvo. Tirtas žodžių pasikartojimas nuo 1 iki 5, naudojant dažniausiai vartojamų žodžių sąrašą arba ne, todėl kiekvienam duomenų rinkiniui sudaryta po 10 tekstinių dokumentų matricų, kurios analizuotos saviorganizuojančiu neuroniniu tinklu ir k-vidurkių metodu.

4.5.1. Rezultatai, gauti naudojant SOM tinklą

Pasirinktas 10×10 SOM žemėlapio dydis ($k_x = k_y = 10$). 80 % duomenų priskirti mokymo aibei, likusieji 20 % – testavimo aibei. Kiekvienas bandymas kartotas 5 kartus (pirminis tyrimas parodė, jog rezultatai iš esmės nesikeičia, todėl 5 kartų pakanka), esant skirtingoms žemėlapio neuronų pradinėms reikšmėms. 4.15–4.16 lentelėse pateikti penkių bandymų metu gautų paklaidų reikšmių vidurkiai. Pirmas bandymas atliktas naudojant duomenų aibės „Ministerijų įsakymai“ pirmąjį duomenų rinkinį (60 įsakymų, I klasės duomenys – Sveikatos apsaugos ministerija, II klasė – Švietimo ir mokslo ministerija, III klasė – Vidaus reikalų ministerija, IV klasė – Žemės ūkio ministerija). Kaip matome 4.15 lentelėje, pirmo duomenų rinkinio rezultatai yra gana panašūs, naudojant ir nenaudojant dažniausiai vartojamų žodžių sąrašą.

4.15 lentelė. SOM rezultatai, naudojant pirmojo tekstinių duomenų rinkinio mokymo aibę

Pasikartojimas / Paklaida	1	2	3	4	5
Nenaudojant dažniausiai vartojamų žodžių sąrašo					
E_1	17,84	17,78	19,32	20,34	18,94
E_2	28,50	28,29	28,25	26,76	29,69
E_3	15,17	17,62	22,20	17,99	19,68
E_4	24,30	30,13	26,08	31,94	30,63
E_{center}	4,17	3,57	3,61	3,18	3,37
Naudojant dažniausiai vartojamų žodžių sąrašą					
E_1	20,72	18,63	16,56	14,10	17,47
E_2	28,56	31,97	29,67	32,11	28,15
E_3	15,31	16,04	15,35	23,03	23,62
E_4	26,78	21,38	27,32	25,69	29,35
E_{center}	3,95	4,28	3,99	3,41	3,30

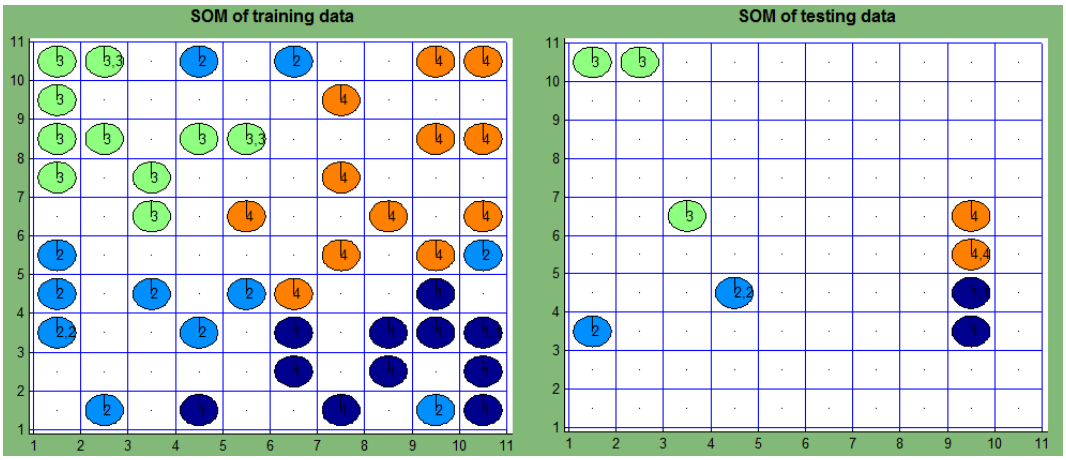
4.16 lentelė. SOM rezultatai, naudojant pirmojo tekstinių duomenų rinkinio testavimo aibę

Pasikartojimas / Paklaida	1	2	3	4	5
Nenaudojant dažniausiai vartojamų žodžių sąrašo					
E_1	0,75	1,22	2,70	2,44	5,17
E_2	2,59	3,75	3,24	3,28	4,95
E_3	2,35	2,50	4,46	2,82	4,51
E_4	2,02	1,60	1,80	2,59	1,68
E_{center}	4,09	4,58	3,90	3,63	3,51
Naudojant dažniausiai vartojamų žodžių sąrašą					
E_1	1,20	0,93	0,58	1,54	3,14
E_2	3,67	3,60	4,02	5,79	3,91
E_3	2,02	2,77	2,96	4,49	6,78
E_4	2,93	2,97	3,19	1,41	1,27
E_{center}	3,95	4,86	4,70	3,85	3,22

Mažiausi vidutiniai atstumai tarp klasių narių gauti I ir III klasėms (E_1 ir E_3). Tai reiškia, jog jų nariai žemėlapyje yra išsidėstę arti vienas kito. Priešingai II ir IV klasės atstumai (E_2 ir E_4). Jie kur kas didesni ir tai parodo, jog šie duomenys pasklidę po žemėlapį plačiau. Didžiausias skirtumas tarp klasių centrų (E_{center}) gautas, kai sudarant žodyną naudojamas dažniausiai vartojamų žodžių sąrašas ($E_{\text{center}} = 4{,}28$) ir žodžių pasikartojimų skaičius nemažesnis

nei 2. Testavimo aibėje atstumas $E_{\text{center}} = 4{,}86$ tuomet, kai naudojamas dažniausiai vartojamų žodžių sąrašas ir žodžių pasikartojimas nemažesnis nei 2.

4.17 pav. pateiktas SOM žemėlapis pirmajam duomenų rinkiniui, iš visų bandymų išrinkus tą, kurio paklaidų reikšmės buvo geriausios, t. y. $E_1 - E_4$ reikšmė maža, o E_{center} – didelė. 4.18 pav. pavaizduotas žemėlapis, kai paklaidų reikšmės blogiausios. Geriausias SOM žemėlapis gautas, kai žodžių pasikartojimų skaičius nemažesnis kaip 2 ir sudarant žodyną atmetami dažniausiai vartojami žodžiai. Paveiksluose spalvos ir skaičiai žymi klases. SOM žemėlapis, kurio paklaidų reikšmės yra mažiausios, gautas, kai pasirinktas žodžių pasikartojimų skaičius nemažesnis kaip 4 ir nėra naudotas dažniausiai vartojamų žodžių sąrašas (4.17 pav.).



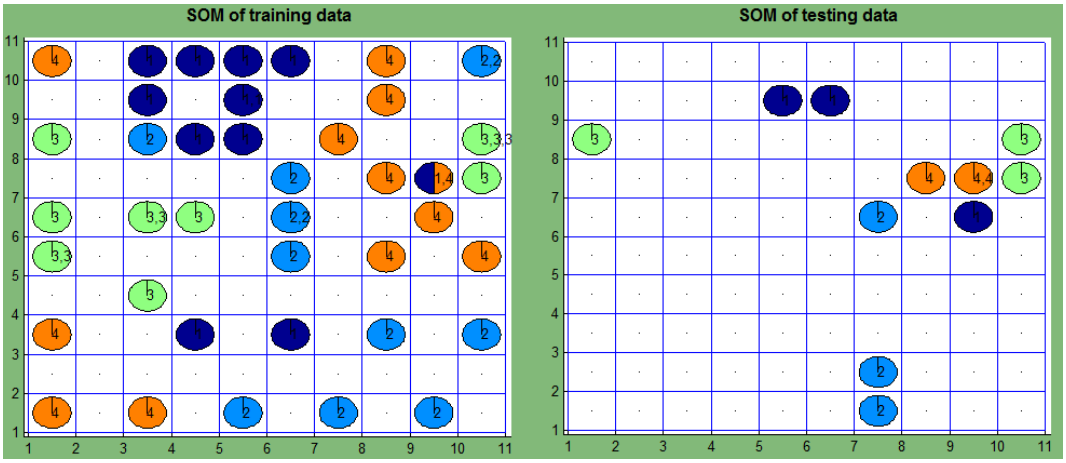
a) **b)**

4.17 pav. Geriausias SOM žemėlapis pirmajam duomenų rinkiniui:
a) mokymo aibė ($E_1 = 15{,}15, E_2 = 29{,}33, E_3 = 14{,}08, E_4 = 18{,}63, E_{\text{center}} = 5{,}36$),
b) testavimo aibė ($E_1 = 0{,}67, E_2 = 2{,}11, E_3 = 3{,}20, E_4 = 0{,}67, E_{\text{center}} = 5{,}89$)

4.17 pav. pateiktame žemėlapyje matyti, jog I ir III (Sveikatos apsaugos ir Vidaus reikalų ministerijos) klasių duomenys sudaro du atskirus klasterius, o klasės nariai išsidėstę arti vienas kito, ką ir parodė 4.15 lentelėje pateiktos paklaidų E_1 ir E_3 reikšmės. Taigi Sveikatos apsaugos ir Vidaus reikalų ministerijų įsakymų dokumentai yra tarpusavyje susiję. II ir IV (Švietimo ir mokslo ir Žemės ūkio ministerijos) klasių duomenys plačiau pasklidę po žemėlapį, todėl jų klasių atstumų paklaidų reikšmės (E_2 ir E_4) kur kas didesnės.

4.17 ir 4.18 pav. dešinėje pateikti testavimo rezultatai parodo, jog duomenys išsidėsto grupelėmis maždaug tose pačiose vietose, kur ir mokymo aibės žemėlapyje pateikti duomenys.

4.18 pav. pateiktame žemėlapyje mokymo aibės atveju matome, jog duomenys išsibarsto plačiai po visą žemėlapį. Susiformavo tik nedideli I klasės klasteriai viršuje, III klasės – kairėje, II klasės – apatiniame kampe, IV klasės – viršutiniame dešiniajame kampe. Į vieną langelį patenka po vieną I ir IV klasių narį. Šis eksperimentas dar kartą patvirtina, kad SOM žemėlapių kokybei (duomenų klasių atitikimui SOM klasteriams) vertinti tikslinga naudoti darbe pasiūlytas paklaidas.



a) **b)**

4.18 pav. Blogiausias SOM žemėlapis pirmajam duomenų rinkiniui:
a) mokymo aibė ($E_1 = 20{,}15, E_2 = 27{,}93, E_3 = 27{,}28, E_4 = 32{,}31, E_{center} = 2{,}19$),
b) testavimo aibė ($E_1 = 3{,}41, E_2 = 3{,}33, E_3 = 6{,}35, E_4 = 0{,}67, E_{\text{center}} = 3{,}09$)

Antras bandymas atliktas naudojant duomenų aibės „Ministerijų įsakymai“ antrąjį duomenų rinkinį (60 įsakymų, I klasė – Finansų ministerija, II klasė – Kultūros ministerija, III klasė – Susisiekimo ministerija, IV klasė – Ūkio ministerija). 4.17–4.18 lentelėse matome, kad mažiausios pirmosios paklaidos reikšmės (E_2 ir E_4) gautos II ir IV klasei (Kultūros ir Ūkio ministerijos). Tai rodo, kad šių ministerijų įsakymų dokumentai yra skirtingi. Šių klasių duomenys žemėlapyje išsidėsto arčiau vieni kitų tiek mokymo, tiek testavimo aibei. I ir III (Finansų ir Susisiekimo ministerijų) klasių duomenų rezultatai gerokai blogesni.

4.17 lentelė. SOM rezultatai, naudojant antrojo tekstinių duomenų rinkinio mokymo aibę

Pasikartojimas / Paklaida	1	2	3	4	5
Nenaudojant dažniausiai vartojamų žodžių sąrašo					
E_1	28,02	34,40	32,36	28,82	33,38
E_2	19,86	20,88	19,91	22,88	22,52
E_3	28,35	25,84	25,14	26,40	26,62
E_4	14,95	15,12	16,98	22,24	21,26
E_{center}	3,99	3,28	3,33	3,07	2,54
Naudojant dažniausiai vartojamų žodžių sąrašą					
E_1	31,97	29,26	30,64	29,52	29,93
E_2	19,31	22,12	18,70	20,97	23,26
E_3	22,51	24,66	25,71	24,51	29,07
E_4	14,64	20,13	18,89	18,94	20,78
E_{center}	4,24	3,66	3,45	3,99	3,07

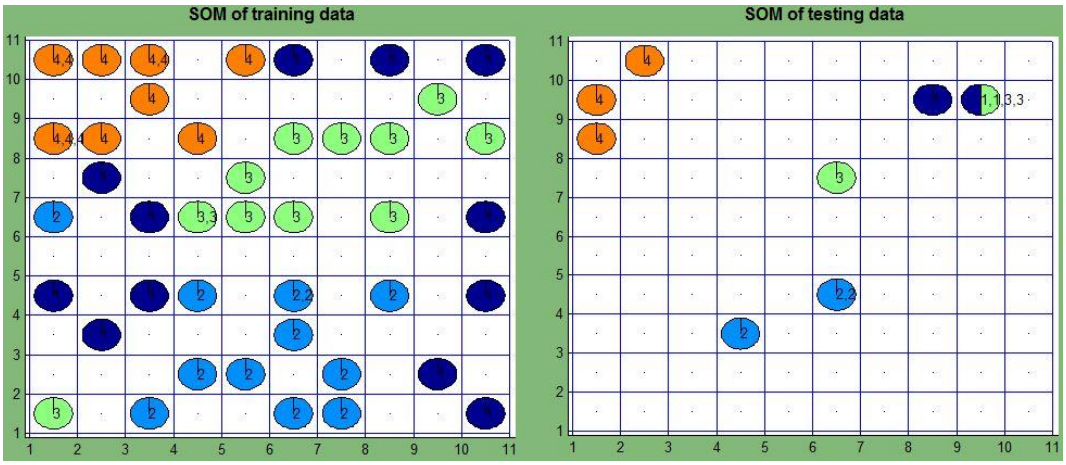
4.18 lentelė. SOM rezultatai, naudojant antrojo tekstinių duomenų rinkinio testavimo aibę

Pasikartojimas / Paklaida	1	2	3	4	5
Nenaudojant dažniausiai vartojamų žodžių sąrašo					
E_1	3,45	4,33	3,44	4,17	4,26
E_2	1,56	2,06	1,56	0,67	0,83
E_3	5,06	4,79	5,10	4,48	5,02
E_4	1,49	2,30	2,88	4,12	3,59
E_{center}	4,67	4,74	4,19	4,04	3,31
Naudojant dažniausiai vartojamų žodžių sąrašą					
E_1	2,97	4,06	4,96	5,27	4,44
E_2	1,51	1,70	1,92	0,78	1,00
E_3	4,84	4,73	4,11	3,24	5,70
E_4	1,68	3,31	3,33	2,16	3,60
E_{center}	4,93	4,51	3,24	4,47	3,33

Kaip ir pirmojo duomenų rinkinio analizės atveju, didinant žodžių pasikartojimų skaičių, atstumai tarp centrų (E_{center}) mažėja, duomenys SOM žemėlapyje pasiskirsto plačiau. Matome, kad geresni (didesnė paklaidos reikšmė) rezultatai gaunami, kai dažniausiai vartojamų žodžių sąrašas nėra naudojamas sudarant dokumentų žodyną.

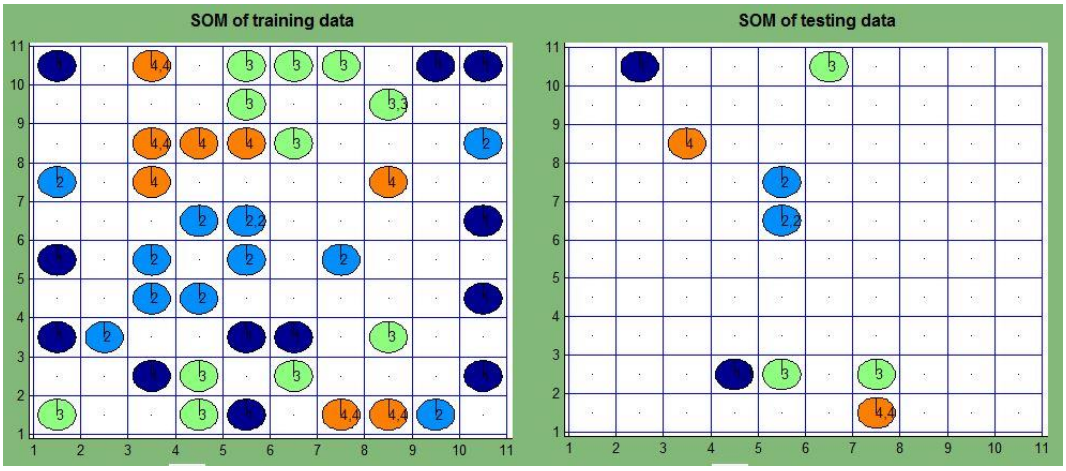
Geriausias SOM žemėlapis gautas, kai žodžių pasikartojimas nemažesnis kaip 1 ir sudarant žodyną naudojamas dažniausiai vartojamų žodžių sąrašas (4.19 pav.). Mokymo aibės kairiajame viršutiniame žemėlapio kampe išsidėsto visi IV klasės nariai (Ūkio ministerijos), apačioje – dauguma II klasės narių (Kultūros ministerijos), per vidurį – III klasės nariai (Susisiekimo ministerijos), o I klasės nariai (Finansų ministerijos) išsibarsto po visą SOM žemėlapį. Pagal 4.17 lentelėje pateiktus rezultatus mažiausi atstumai (E_2 ir E_4) yra tarp IV ir II klasių narių, ir tai matosi pateiktame žemėlapyje (4.19 pav.). Plačiausiai išsidėsto I klasės duomenys.

Blogiausias SOM rezultatas tiriamų paklaidų prasme gautas, kai pasikartojimų skaičius 5 ir nenaudojamas dažniausiai vartojamų žodžių sąrašas. 4.20 pav. matome tik nedidelius skirtingų klasių klasterius, tačiau iš esmės visų klasių duomenys plačiai išsidėsto įvairiose žemėlapio vietose.



a) **b)**

4.19 pav. Geriausias SOM žemėlapis antrajam duomenų rinkiniui:
a) mokymo aibė ($E_1 = 31{,}18, E_2 = 17{,}65, E_3 = 20{,}85, E_4 = 15{,}98, E_{\text{center}} = 4{,}59$),
b) testavimo aibė ($E_1 = 2{,}33, E_2 = 1{,}33, E_3 = 5{,}33, E_4 = 1{,}61, E_{\text{center}} = 4{,}55$)



a) **b)**

4.20 pav. Blogiausias SOM žemėlapis antrajam duomenų rinkiniui:
a) mokymo aibė ($E_1 = 35{,}93, E_2 = 21{,}86, E_3 = 29{,}85, E_4 = 28{,}13, E_{\text{center}} = 1{,}08$),
b) testavimo aibė ($E_1 = 5{,}50, E_2 = 0{,}67, E_3 = 6{,}04, E_4 = 5{,}37, E_{\text{center}} = 2{,}35$)

Atlikus eksperimentus, naudojant trečiąjį duomenų rinkinį (60 įsakymų, I klasė – Finansų ministerija, II klasė – Ūkio ministerija, III klasė – Vidaus reikalų ministerija, IV klasė – Žemės ūkio ministerija), matome (4.19–4.20 lentelės), kad mažiausios paklaidų reikšmės (E_2 ir E_3) gautos II ir III klasės (Ūkio ir Vidaus reikalų ministerijos), o I ir IV klasėms (Finansų ir Žemės ūkio ministerijos) paklaidų reikšmės (E_1 ir E_4) yra didžiausios.

4.19 lentelė. SOM rezultatai, naudojant trečiojo tekstinių duomenų rinkinio mokymo aibę

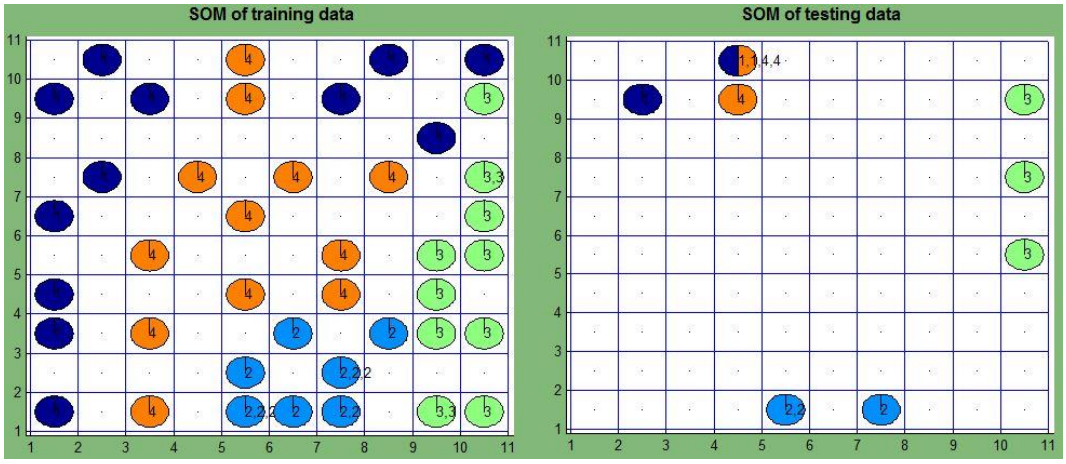
Pasikartojimas / Paklaida	**1**	**2**	**3**	**4**	**5**
Nenaudojant dažniausiai vartojamų žodžių sąrašo					
E_1	29,98	31,65	30,11	30,22	32,77
E_2	11,49	18,47	17,41	21,78	21,45
E_3	17,18	22,67	19,75	19,63	23,60
E_4	26,00	26,29	29,64	30,11	27,87
E_{center}	4,45	3,55	3,43	3,08	2,89
Naudojant dažniausiai vartojamų žodžių sąrašą					
E_1	30,63	33,42	33,44	32,65	30,43
E_2	12,52	14,46	15,82	16,86	18,46
E_3	14,48	17,47	21,25	21,76	26,09
E_4	27,92	30,09	30,52	29,43	30,21
E_{center}	4,10	3,43	2,59	3,22	2,94

4.20 lentelė. SOM rezultatai, naudojant trečiojo tekstinių duomenų rinkinio testavimo aibę

Pasikartojimas / Paklaida	1	2	3	4	5
Nenaudojant dažniausiai vartojamų žodžių sąrašo					
E_1	1,93	5,13	4,26	3,96	4,36
E_2	1,65	3,08	2,96	4,03	4,24
E_3	2,57	3,06	3,25	2,66	5,84
E_4	1,54	4,85	2,69	3,22	1,28
E_{center}	5,24	4,04	3,26	2,56	3,29
Naudojant dažniausiai vartojamų žodžių sąrašą					
E_1	4,80	3,94	2,75	4,32	3,99
E_2	1,95	2,30	1,84	2,94	3,85
E_3	2,32	1,98	3,22	4,23	6,35
E_4	3,08	1,26	3,27	2,85	1,78
E_{center}	4,89	4,25	3,35	3,33	2,65

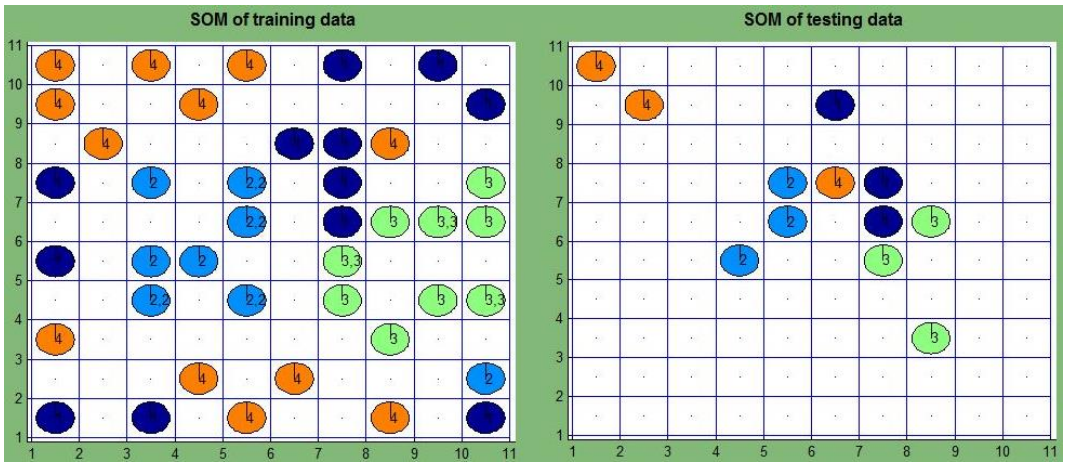
Didinant pasikartojimų skaičių mokymo ir testavimo aibėms, analogiškai antrajam duomenų rinkiniui, atstumai tarp centrų E_{center} mažėja. Šiuo atveju, kai pasikartojimų skaičius nuo 1 iki 3, didesni atstumai tarp centrų E_{center} gaunami, kai nenaudojamas dažniausiai vartojamų žodžių sąrašas, o kai pasikartojimų skaičius 4 ir 5, naudojant sąrašą, rezultatai geresni.

Šiuo atveju geriausias pagal pasiūlytas paklaidas SOM žemėlapis gautas (4.21 pav.), kai žodžių pasikartojimų skaičius nemažesnis kaip 1 ir dažniausiai vartojamų žodžių sąrašas nėra naudotas. Aiškiai matomi mokymo aibės žemėlapio apačioje išsidėstę II klasės nariai (Ūkio ministerijos), dešinėje – III klasė (Vidaus reikalų ministerija), per vidurį iš ilgai nuo apačios iki viršaus – IV klasės nariai (Žemės ūkio ministerijos) ir kairėje pusėje – dauguma I klasės narių (Finansų ministerija).



a) **b)**

4.21 pav. Geriausias SOM žemėlapis trečiajam duomenų rinkiniu:
a) mokymo aibė ($E_1 = 26{,}30, E_2 = 11{,}04, E_3 = 17{,}45, E_4 = 25{,}41, E_{\text{center}} = 4{,}88$),
b) testavimo aibė ($E_1 = 0{,}67, E_2 = 1{,}55, E_3 = 3{,}04, E_4 = 4{,}27, E_{\text{center}} = 4{,}85$).



a) **b)**

4.22 pav. Blogiausias SOM žemėlapis trečiajam duomenų rinkiniui:
a) mokymo aibė ($E_1 = 32{,}33, E_2 = 14{,}21, E_3 = 28{,}29, E_4 = 30{,}31, E_{\text{center}} = 1{,}87$),
b) testavimo aibė ($E_1 = 2{,}33, E_2 = 1{,}33, E_3 = 4, E_4 = 7{,}07, E_{\text{center}} = 2{,}87$)

Blogiausi rezultatai pasiūlytų paklaidų prasme gauti, kai žodžių pasikartojimų skaičius nemažesnis kaip 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 4.21 pav., tiek 4.22 pav. yra išsidėsčiusios greta, tai rodo, kad šie dokumentai yra panašūs. Tikėtina, kad I klasės (Finansų ministerija) ir III klasės (Vidaus

reikalų ministerija) dokumentuose nagrinėjami panašūs klausimai, todėl I ir III klasių duomenys yra greta.

4.5.2 Apibendrinti žodžių pasikartojimo skaičiaus įtakos rezultatai

Apibendrinant 4.15–4.20 lentelėse pateiktus rezultatus, apskaičiuoti trijų duomenų aibių eksperimentų mokymo aibės ir testavimo aibės rezultatų vidurkiai, naudojant ir nenaudojant dažniausiai vartojamų žodžių sąrašą (4.21–4.22 lentelės).

4.21 lentelė. Bendri trijų duomenų rinkinių SOM rezultatai mokymo aibei

Pasikartojimas / **Paklaida**	**1**	**2**	**3**	**4**	**5**
Nenaudojant dažniausiai vartojamų žodžių sąrašo					
E_1	25,28	27,94	27,26	26,46	28,36
E_2	19,95	22,55	21,86	23,81	24,55
E_3	20,23	22,04	22,36	21,34	23,3
E_4	21,75	23,85	24,23	28,10	26,59
E_{center}	4,20	3,47	3,46	3,11	2,93
Naudojant dažniausiai vartojamų žodžių sąrašą					
E_1	27,77	27,10	26,88	25,42	25,94
E_2	20,13	22,85	21,40	23,31	23,29
E_3	17,43	19,39	20,77	23,1	26,26
E_4	23,11	23,87	25,58	24,69	26,78
E_{center}	4,10	3,79	3,34	3,54	3,10

4.22 lentelė. Bendri trijų duomenų rinkinių SOM rezultatai testavimo aibei

Pasikartojimas / **Paklaida**	**1**	**2**	**3**	**4**	**5**
Nenaudojant dažniausiai vartojamų žodžių sąrašo					
E_1	2,04	3,56	3,47	3,52	4,60
E_2	1,93	2,96	2,59	2,66	3,34
E_3	3,33	3,45	4,27	3,32	5,12
E_4	1,68	2,92	2,46	3,31	2,18
E_{center}	4,67	4,45	3,78	3,41	3,37
Naudojant dažniausiai vartojamų žodžių sąrašą					
E_1	2,99	2,98	2,76	3,71	3,86
E_2	2,38	2,53	2,59	3,17	2,92
E_3	3,06	3,16	3,43	3,99	6,28
E_4	2,56	2,51	3,26	2,14	2,22
E_{center}	4,59	4,54	3,76	3,88	3,07

4.21–4.22 lentelių reikšmės įvertintos balais nuo 1 iki 5: geriausias iš visų gautas rezultatas įvertintas 5 balais, blogiausias – 1. Gauti rezultatai pateikti 4.23–4.24 lentelėse.

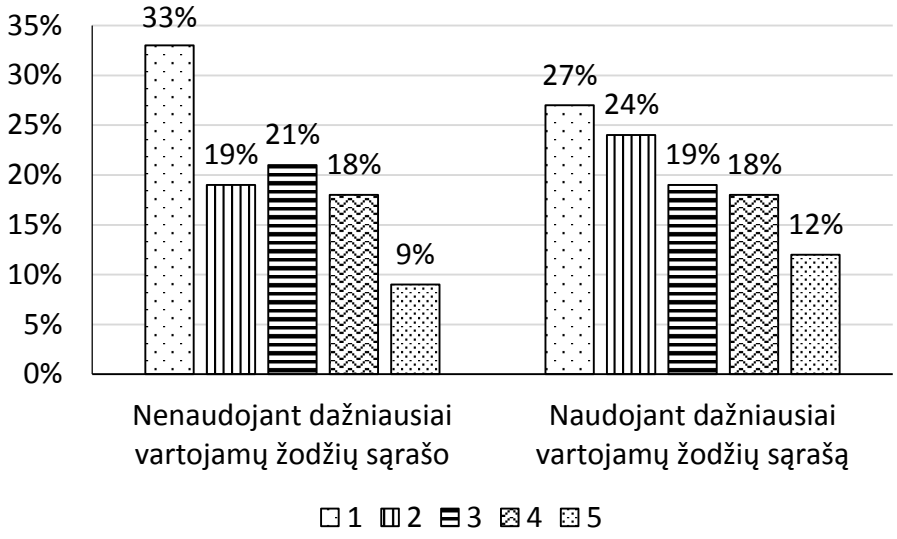
4.23 lentelė. Apibendrinti 4.21–4.22 rezultatai, nenaudojant dažniausiai vartojamų žodžių sąrašo

Pasikartojimas / Paklaida	Mokymo aibė					Testavimo aibė				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
E_1	5	2	3	4	1	5	2	4	3	1
E_2	5	3	4	2	1	5	2	4	3	1
E_3	5	3	2	4	1	4	3	2	5	1
E_4	5	4	3	1	2	5	2	3	1	4
E_{center}	5	4	3	2	1	5	4	3	2	1
Iš viso	**25**	**16**	**15**	**13**	**6**	**24**	**13**	**16**	**14**	**8**

4.24 lentelė. Apibendrinti 4.21–4.22 rezultatai naudojant dažniausiai vartojamų žodžių sąrašą

Pasikartojimas / Paklaida	Mokymo aibė					Testavimo aibė				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
E_1	1	2	3	5	4	3	4	5	2	1
E_2	5	3	4	1	2	5	4	3	1	2
E_3	5	4	3	2	1	5	4	3	2	1
E_4	5	4	2	3	1	2	3	1	5	4
E_{center}	5	4	2	3	1	5	4	2	3	1
Iš viso	**21**	**17**	**14**	**14**	**9**	**20**	**19**	**14**	**13**	**9**

Diagramoje (4.23 pav.) procentais parodyti bendri mokymo aibės ir testavimo aibės rezultatai. Tiek nenaudojant, tiek naudojant dažniausiai vartojamų žodžių sąrašą, geriausi rezultatai gauti, kai žodžių pasikartojimas dokumente yra pasirinktas mažiausias (33 % ir 27 % atvejų). Didėjant žodžių pasikartojimui, rezultatai kiekvieną kartą pablogėja, vertinant gautas paklaidų E_1, E_2, E_3, E_4 ir E_{center} reikšmes. Blogiausi rezultatai gauti, kai, sudarant teksto dokumentų žodyną nenaudotas dažniausiai vartojamų žodžių sąrašas, t. y. įtraukti visi žodžiai (9 % atvejų).



4.23 pav. Bendri 4.23–4.24 lentelių rezultatai

4.5.3. Klasterizavimo rezultatai, gauti naudojant *k*-vidurkių metodą

Tikslinga SOM klasterizavimo rezultatus palyginti su rezultatais, gautais vienu populiariausiu klasterizavimo metodu – k-vidurkių (angl. *k-means*). Naudojant k-vidurkių metodą, stengiamasi duomenų aibę padalinti į nesusikertančius klasterius (MacQueen, 1967). Šiame metode minimizuojama tam tikra kriterijaus funkcija. Ji turi būti tokia, kad, minimizuojant klasterių panašumą, klasterių skirtumas būtų maksimizuojamas. Tai gali būti vidutinis atstumas tarp klasterių. Tarkime, kad klasteriui K^i priskirta objektų aibė $X = \{X_1^i, X_2^i, \dots, X_\mu^i\}$, μ – objektų klasteryje K^i skaičius, $X_j^i = \{x_{j1}^i, x_{j2}^i, \dots, x_{jn}^i\}, j = 1, \dots, \mu$. Tuomet kvadratinė paklaida vienam klasteriui K^i yra Euklido atstumų tarp kiekvieno klasterio elemento ir klasterio centro C^i kvadratų suma $E_{K^i} = \sum_{j=1}^{\mu} \left\| X_j^i - C^i \right\|^2$. 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 (18) tarp klasterio centro ir klasteriui priskirtų duomenų.

$$E_{AKS} = \sum_{i=1}^{K} E_{K^i}. \quad (18)$$

5. 2–4 etapai kartojami tol, kol analizuojami duomenys nebepasiskirsto kitiems klasteriams.

Įprastai k-vidurkių metodo kokybei įvertinti skaičiuojama gautų klasterių atstumo suma nuo klasterio centro iki jam priskirtų duomenų E_{AKS}. Šiame tyrime analizuoti duomenys priskirti klasėms, todėl svarbu įvertinti, ar gauti klasteriai atitinka klases. Pradžioje duomenys klasterizuojami į tiek klasterių, kiek yra klasių. Nustatomas klasių atitikimas klasteriams, 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.

Tyrimams atlikti naudota Matlab sistemos funkcija „kmeans“. Imant vis kitas pradines klasterių centrų koordinates, bandymai buvo kartoti po 10 kartų. Kas kartą buvo vertinamas klasių priskyrimas klasteriams (4.25–4.26 lentelėse tai nurodoma „Neteisingai priskirti“) bei atstumų tarp klasterių centrų iki jiems priskirtų duomenų sumos E_{AKS}. Lentelėse pateikiamas 10 bandymų metu gautų įverčių vidurkiai.

Iš 4.26–4.26 lentelėse pateiktų rezultatų matome, 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 E_{AKS} kaskart gerėja, neskaitant vieno atvejo antrajam duomenų rinkiniui, kai pasikartojimų skaičius lygus 4. Tačiau vertinti gautus rezultatus pagal E_{AKS} nėra tikslinga, kadangi nagrinėjamų duomenų aibių matmenų skaičius yra skirtingas, ir mažesnė E_{AKS} reikšmė nenurodo klasterizavimo tikslumo. Vertinant k-vidurkių metodo rezultatus pagal klasių priskyrimo paklaidą „Neteisingai priskirti“, didinant žodžių pasikartojimų skaičių, matoma tendencija didėti ir šios paklaidos reikšmei. Daugelių atvejų, didžiausios paklaidos reikšmės gautos, kai pasikartojimų skaičius lygus 5, o mažiausios, – kai 1 arba 2. Lentelėse didžiausios reikšmės pateiktos pasviruoju šriftu, mažiausios – paryškintu.

Nagrinėjant klasių priskyrimą klasteriams, matome, jog pirmajam duomenų rinkiniui, nurodant pasikartojimų skaičių nuo 2–5, klasių priskyrimas tinkamiems klasteriams yra geresnis, kai yra naudojamas dažniausiai vartojamų žodžių sąrašas, nei tuo atveju, kai dažniausiai vartojamų žodžių sąrašas nėra įtrauktas. Antrojo ir trečiojo duomenų rinkinių atvejais klasių priskyrimo klasteriams rezultatai yra gana panašūs. Tyrimas parodė, jog duomenų aibės analizė, naudojant saviorganizuojančius neuroninius tinklus, yra išsamesnė nei k-vidurkio metodo atveju, kadangi SOM rezultate gaunami ne tik skaitiniai įverčiai, bet rezultatus galima stebėti ir vizualiai.

4.25 lentelė. k-vidurkio metodo rezultatai, nenaudojant dažniausiai vartojamų žodžių sąrašo

Pasikartojimas / **Paklaida**	**1**	**2**	**3**	**4**	**5**
Pirmajam duomenų rinkiniui					
Neteisingai priskirti	**18,3**	23,6	26,6	30,7	*33,2*
E_{AKS}	43608	40513	35885	31221	26923
Antrajam duomenų rinkiniui					
Neteisingai priskirti	**21,9**	23,3	24,7	*28,8*	28
E_{AKS}	36796	34209	30580	26469	22719
Trečiajam duomenų rinkiniui					
Neteisingai priskirti	**19,9**	21,9	27,6	32,5	*35,1*
E_{AKS}	39383	37593	33090	28965	24125

4.26 lentelė. k-vidurkio metodo rezultatai, naudojant dažniausiai vartojamų žodžių sąrašą

Pasikartojimas / **Paklaida**	**1**	**2**	**3**	**4**	**5**
Pirmajam duomenų rinkiniui					
Neteisingai priskirti	22,8	**20,1**	21,5	27,3	*32,1*
AKS	40112	37259	32887	29115	24972
Antrajam duomenų rinkiniui					
Neteisingai priskirti	25	25,2	**23,9**	26,5	*28,9*
AKS	35732	32302	28056	24934	21585
Trečiajam duomenų rinkiniui					
Neteisingai priskirti	**21,7**	27,7	26,8	32,7	*34,2*
AKS	37124	34765	32004	26021	21524

4.6. Ketvirtojo skyriaus rezultatai ir išvados

Šiame skyriuje atlikta saviorganizuojančių neuroninių tinklų sistemų lyginamoji analizė pagal nustatytus kriterijus parodė, kad daugiausia kriterijų atitinka T. Kohoneno vadovaujamos mokslininkų grupės sukurta SOM-Toolbox sistema, tačiau ir joje nėra galimybės žemėlapyje matyti santykio tarp skirtingų duomenų klasių, pakliuvusių į tą patį SOM langelį, o naujoje SOM sistemoje šis vizualizavimo būdas yra įgyvendintas. Be to, panašus vizualizavimo būdas yra ir naujausiose Orange sistemos versijoje, tačiau jis įgyvendintas vėliau, nei buvo pasiūlytas disertacijos autoriaus.

Tiriant įvairių mokymo faktorių reikšmę gautiems SOM rezultatams, buvo naudoti ir tekstiniai, ir skaitiniai duomenys. Įprastai tekstiniai duomenys skiriasi nuo skaitinių tuo, jog juos atitinkantys vektoriai turi labai daug požymių. Gauti SOM rezultatai vertinti ne tik pagal kvantavimo paklaidą E_{QE}, tačiau taikytos ir darbe pasiūlytos SOM žemėlapio kokybės įvertinimo paklaidos E_{c} ir E_{center}. Atlikti tyrimai parodė, kad tirtiems duomenims:

- Pasiūlytos SOM žemėlapio kokybę įvertinančios paklaidos tinkamai įvertina duomenų klasių atitikimą SOM klasterius. Pagal pasiūlytas paklaidas galima lyginti tarpusavyje keliuose to paties dydžio SOM žemėlapiuose susidariusių klasterių sutapimą su duomenų klasėmis.
- Daugeliu atvejų mažiausia kvantavimo paklaida yra gaunama, kai naudojama Gauso kaimynystės funkcija nepriklausomai nuo mokymo parametro tipo.
- Analizuojant tekstinius duomenis, 49 % tirtų atvejų mažiausios paklaidų reikšmės gautos, naudojant euristinę kaimynystės funkciją, 31 % atvejų – Gauso ir tik 20 % atvejų – burbuliuko. Analizuojant skaitinius duomenis, vertintų paklaidų reikšmės gana panašios, naudojant bet kurią kaimynystės funkciją. 37 % tirtų atvejų mažiausios paklaidos gautos, naudojant Gauso kaimynystės funkciją, 35 % atvejų – euristinę, 28 % atvejų – burbuliuko.

- Vertinant SOM rezultatus atsižvelgiant, į naudotą mokymo parametrą, geriausi rezultatai tekstinių duomenų aibei gauti, naudojant atvirkštinį laikui mokymo parametrą, keičiant jo reikšmes kiekvienoje epochoje (16 % tirtų atvejų), o blogiausi, – naudojant tiesinį mokymo parametrą, keičiant jo reikšmes kiekvienoje epochoje (7 % tirtų atvejų). Skaitinių duomenų atveju geriausi rezultatai gauti, naudojant tiesinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (17 % tirtų atvejų), o blogiausi, – naudojant euristinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (9 % tirtų atvejų).

Tiriant teksto dokumentų konvertavimo į skaitinę išraišką faktorių įtaką gaunamiems SOM rezultatams, paaiškėjo, kad:

- Tiksliausi SOM rezultatai, vertinant paklaidas gauti tuomet, kai teksto dokumentų matrica sudaroma, atsižvelgiant į rankiniu būdų sudarytus žodynus (18 % ir 15 %) arba į automatiniu būdu sudarytą žodyną, į jį neįtraukus dažniausių žodžių iš sąrašo, sudaryto atsižvelgiant į dokumente pateikiamą informaciją (16 %).
- Blogiausi rezultatai gauti (8 %), kai žodynas sudarytas neatsižvelgiant į dažniausiai vartojamų žodžių sąrašą, t. y. į žodyną įtraukti žodžiai, kurie dokumentuose kartojasi ne mažiau nei nustatytas pasikartojimų skaičius, o dažnai pasikartojantys žodžiai neatspindi analizuojamo dokumento turinio.
- Tyrimuose taikytas Porterio kamieno išskyrimo algoritmas nelemia SOM rezultatų kokybės, vertintų paklaidų prasme.
- Tinkamiausias žodžių pasikartojimų skaičius dokumente yra ne daugiau kaip 5, kadangi pasirinkus didesnį skaičių, sudarant tekstinių dokumentų žodyną, dalis dokumentų buvo atmesti, nes juose tiek kartų pasikartojančių žodžių nebuvo.
- Didėjant žodžių pasikartojimų skaičiui dokumente, paklaidų E_1, E_2, E_3, E_4 reikšmės gautos didesnės, o paklaidos E_{center} – mažesnės abiem tirtais atvejais (kai sudarant žodyną atmesti žodžiai iš dažniausiai vartojamų žodžių sąrašo ir kai jie neatmesti); geriausi rezultatai gauti, kai

minimalus pasikartojimų skaičius lygus 1 ir sudarant žodyną nenaudotas dažniausiai vartojamų žodžių sąrašas (33 % atvejų); blogiausias rezultatas gautas nenaudojant dažniausiai vartojamų žodžių sąrašo, kai žodžių pasikartojimų skaičius nurodytas nemažesnis nei 5 (9 % atvejų).

- Vertinant k-vidurkių metodo rezultatus pagal klasių atitikimą klasteriams, didinant minimalų pasikartojimų žodžių skaičių, didėja ir šios paklaidos reikšmė; daugumoje atvejų tiksliausi rezultatai gauti, kai minimalus pasikartojimų skaičius lygus 1 ar 2, o blogiausi rezultatai, – kai pasikartojimų skaičius lygus 5.

Bendrosios išvados

Tiriant saviorganizuojančius neuroninius tinklus, gauti šie rezultatai: sukurtas naujas SOM vizualizavimo būdas; pasiūlytos paklaidos, leidžiančios įvertinti (tarpusavyje palyginti) keliuose SOM žemėlapiuose susidariusių klasterių atitikimą su duomenų klasėmis; sukurta nauja SOM sistema, kurioje įgyvendintas pasiūlytas SOM vizualizavimo būdas, SOM kokybę įvertinančios paklaidos bei galimybė rinktis įvairius SOM mokymo faktorius; ištirta SOM mokymo faktorių bei tekstinių dokumentų konvertavimo į skaitinius duomenis įtaka gautiems SOM rezultatams.

Atlikti tyrimai atskleidė darbe pasiūlytų SOM rezultatų kokybę, vertintų paklaidų bei pasiūlyto SOM vizualizavimo būdo naudą, tiriant duomenis, kurių klasės iš anksto žinomos. Remiantis eksperimentinių tyrimų rezultatais, padarytos šios išvados:

1. Pasiūlytos SOM žemėlapio kokybę įvertinančios paklaidos tinkamai parodo duomenų klasių ir klasterių atitikimą žemėlapyje.
2. Pasiūlytas SOM vizualizavimo būdas leidžia pavaizduoti skirtingų klasių duomenų, pakliuvusių į tą patį SOM žemėlapio langelį, santykius.
3. Tiriant tekstinius duomenis, pasiūlytų SOM kokybę įvertinančių paklaidų prasme, euristinės funkcijos naudojimas, leidžia gauti tikslesnius rezultatus – 49 % tirtų atvejų, Gauso funkcijos – 31 %, burbuliuko – 20 %; tiriant skaitinius duomenis geriausi SOM rezultatai taikytų paklaidų prasme gauti, naudojant Gauso kaimynystės funkciją (37 % atvejų), tačiau jie mažai skiriasi nuo rezultatų, gautų naudojant euristinę funkciją (35 % atvejų).
4. Atsižvelgiant į naudotą mokymo parametrą, geriausi SOM rezultatai vertintų paklaidų prasme tekstinių duomenų aibei gauti, naudojant atvirkštinį laikui mokymo parametrą, keičiant jo reikšmes kiekvienoje epochoje (16 % tirtų atvejų), o blogiausi, – naudojant tiesinį mokymo parametrą, keičiant jo reikšmes kiekvienoje epochoje

(7 % tirtų atvejų); Skaitinių duomenų atveju geriausi rezultatai gauti, naudojant tiesinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (17 % tirtų atvejų), o blogiausi, – naudojant euristinį mokymo parametrą, keičiant jo reikšmes kiekvienoje iteracijoje (9 % tirtų atvejų).

5. Tiriant žodyno, kuris naudojamas konvertuojant tekstinius dokumentus į skaitinius duomenis, sudarymo būdus, tiksliausi SOM rezultatai naudotų paklaidų prasme gauti žodyną sudarant rankiniu būdu (18 % ir 15 %), t. y. į jį įtraukiant norimus raktinius žodžius; Tiriant automatinius žodyno sudarymo būdus, tiksliausi SOM rezultatai gauti (16 %), kai sudarant žodyną atmetami dažniausiai vartojami žodžiai iš sąrašo, sudaryto atsižvelgiant į dokumente pateikiamą informaciją ir nenaudojamas joks žodžių kamieno išskyrimo algoritmas.
6. Tiriant žodžių pasikartojimo dokumente skaičiaus įtaką SOM rezultatams, didinant minimalų žodžių pasikartojimų skaičių, bendras SOM rezultatų tikslumas mažėja; tiksliausi rezultatai gauti, kai minimalus pasikartojimų skaičius lygus 1 (vidutiniškai 30 %), o blogiausi rezultatai, – kai minimalus pasikartojimų skaičius lygus 5 (vidutiniškai 10,5 %).

Literatūra ir šaltiniai

1. Alonso, S., Sulkava, M., Prada, M. A., Domínguez, M., and Hollmén, J. (2011). EnvSOM: A SOM Algorithm Conditioned on the Environment for Cluster ing and Visualization. WSOM 2011, LNCS 6731, pp. 61–70. Springer-Verlag Berlin Heidelberg.
2. Asuncion, A., Newman, D. J. (2007). UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer. Prieiga internete: <http://www.ics.uci.edu/~mlearn/MLRepository.html>
3. Almendra, V., Enachescu, D. (2014). Using Self-organizing Maps for Binary Classification with Highly Imbalanced Datasets. International Journal of Numerical Analysis and Modeling, series b, volume 5, number 3, 238–254.
4. Alsmadi, I., Saleh, I. Z. (2012). Documents Similarities Algorithms for Research Papers Authenticity. ICCIT, pp. 210–214.
5. Berthold, M., Cebron, N., Dill, F., Kotter, T., and Meinl, T. (2007). KNIME: The Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization (GFKL 2007). Springer.
6. Cao, M., Li, A., Fang, Q., Kroger, B. J. (2013). Growing Self-Organizing Map Approach for Semantic Acquisition Modeling. 4th IEEE International Conference on Cognitive Infocommunications, pp. 33–38.
7. Chappell, G., Taylor, J. (1993). The Temporal Kohonen Map. Neural Networks, 6: 441–445.
8. 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.
9. Danilienė, R. (2010). Research of Intelligent Computer Literacy Test Design Method. Daktaro disertacija. Vilniaus universitetas.

10. Deligiorgi, D., Philippopoulos, K., Kouroupetroglou G. (2014). An Assessment of Self-Organizing Maps and k-means Clustering Approaches for Atmospheric Circulation Classification. Recent advances in environmental science and geoscience, Proceedings of ESG'14, pp. 17–23.
11. Demšar, J., Curk, T., & Erjavec, A. (2013). Orange: Data Mining Toolbox in Python; Journal of Machine Learning Research 14(Aug): 2349–2353.
12. Ding, J., Lu, Y-Z., and Chu, J. (2013). Self-Organizing Map-based k-means Clustering for Stability Analysis of Product Quality in Packaging in Semiconductor Manufacturing. Journal of Theoretical and Applied Information Technology, vol. 50, no. 2, pp. 435–442.
13. Dobnikar, A., Lotrič, U.,Šter B. (2011). ICANNGA 2011. Part I, LNCS 6593, pp. 260–269.
14. Dogan, Y., Birant, D., Kut, A. (2013). SOM++: Integration of Self-Organizing Map and K-Means++ Algorithms. Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science Volume 7988, pp. 246–259.
15. Dzemyda, G. (2001). Visualization of a Set of Parameters Characterized by their Correlation Matrix. Computational Statistics and Data Analysis, 36(1): 15–30.
16. Dzemyda, G., Kurasova, O., Žilinskas, J. (2008) Daugiamačių duomenų vizualizavimo metodai, pp. 108–127.
17. Dzemyda, G., Kurasova, O. (2002). Comparative Analysis of the Graphical Result Presentation in the SOM Software. Informatica, vol. 13(3), pp. 275–286.
18. Dunham, M. H. (2002). Data Mining: Introductory and Advanced Topics. Prentice Hall PTR, Upper Saddle River, NJ, USA.
19. Eurostat, (2010). Prieiga internete: <http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home>

20. Ghaemmaghami, F., Manouchehri, R. S. (2013). SOMSN: An Effective Self Organizing Map for Clustering of Social Networks. International Journal of Computer Applications, vol. 84, no. 5, pp. 7–12.
21. Gorgonio, F. L., Costa, J. A. F. (2010). PartSOM: A Framework for Distributed Data Clustering Using SOM and k-means. Self-Organizing Maps, George K Matsopoulos (Ed).
22. Guven, M., Cengizler, C. (2014). Data Cluster Analysis-Based Classification of Overlapping Nuclei in Pap Smear Samples. BioMedical Engineering OnLine, vol. 13.
23. Hagenbuchner, M., Sperduti, A., and Tsoi, A. C. (2003). A Self-Organizing Map for Adaptive Processing of Structured Data. IEEE Transactions on Neural Networks 14 (3), 491–505.
24. Hammer, B., Micheli, A., Sperduti, A., and Strickert, M. (2004). A general framework for unsupervised processing of structured data. Neurocomputing 57, 3–35.
25. Hassinen, P., Elomaa, J., Rönkkö, J., Halme, J., Hodju, P. (1999). Neural Networks Tool – NeNet. Prieiga internete: <http://koti.mbnet.fi/~phodju/nenet/Nenet/General.html>
26. Hofmann, M., Klinkenberg, R. (2013). RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series). CRC Press.
27. Hujun, Y. (2002a). ViSOM – A Novel Method for Multivariate Date Projection and Structure Visualization. IEEE Transactions of Neural Networks.
28. Hujun, Y. (2002b). Data Visualisation and Manifold Mapping Using the ViSOM. Neural Networks.
29. Iwasaki, Y., Abe, T.,Wada, Y.,Wada, K., Ikemura, T. (2013). Novel Bioinformatics Strategies for Prediction of Directional Sequence Changes in Influenza Virus Genomes and for Surveillance of Potentially Hazardous Strains. BMC Infectious Diseases 13(386).

30. Jordan, J., Angelopoulou, E. (2013). Hyperspectral Image Visualization With a 3-D Self-Organizing Map. IEEE 5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS).
31. Kamieno išskyrimo algoritmai įvairioms užsienio kalboms. Prieiga internete: http://snowball.tartarus.org/texts/stemmersoverview.html
32. Kanimozhi, M., Bindu, CH. (2013). Brain MR Image Segmentation Using Self Organizing Map. International Journal of Advanced Research in Computer and Communication Engineering Vol. 2.
Issue 10, pp. 3968–3973.
33. Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1998). WEBSOM – Self-Organizing Maps of Document Collections. Neurocomputing 21:101–117.
34. Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics, 43: 59–69.
35. Kohonen, T. (2001). Self-Organizing Maps, 3rd ed., Springer Series in Information Sciences. Berlin: Springer-Verlag.
36. Kohonen, T., Xing, H. (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, Springer Verlag, Heidelberg, pp. 16–29.
37. Koskela, T., Varsta, M., Heikkonen, J., and Kaski, K. (1998a). Temporal sequence processing using recurrent SOM. In Proceedings of the 2nd International Conference on Knowledge-Based Intelligent Engineering Systems, volume 1, Adelaide, Australia.
38. Koskela, T., Varsta, M., Heikkonen, J., and Kaski, K. (1998b). Time series prediction using recurrent SOM with local linear models. International Journal of Knowledge-Based Intelligent Eng. Systems 2 (1), 60–68.
39. Krilavičius, T., Kuliešienė, D. (2010). Soundex for Lithuania Language. Tech. Rep. TokenMill.

40. Krilavičius, T., Medelis, Ž. (2010). Porter Stemmer for Lithuania Language. Tech. Rep. TokenMill.
41. Kurasova, O. (2005). Daugiamačių duomenų vizuali analizė taikant savireguliuojančius neuroninius tinklus (SOM). Daktaro disertacija. Matematikos ir informatikos institutas.
42. Laaksonen, J., Koskela, M., Laakso, S., and Oja, E. (2000). Picsom – Content-Based Image Retrieval with Self-Organizing Maps. Pattern Recognition Letters 21(13-14), pp. 1199–1207.
43. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, In Le Cam, L. M. and Neyman, J., editors. In Proccedings of the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Statistics. I, 281–297. Berkeley and Los Angeles: University of California Press.
44. 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, Springer Verlag, Heidelberg, pp. 121–130.
45. Mahalakshmi, B., Duraiswamy, K., Gnanasuganya, P., Aruldhevi, P., Sundarapandiyan, R. (2014). Face Verification across Ages Using Self Organizing Map. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 2. Special Issue 1, pp. 1482–1487.
46. Manning, D. C, Raghavan, P. and Schütze, H. (2008). Introduction to Information Retrieval, Cambridge University Press.
47. Marcinkevičius, V. (2010). Netiesinės daugiamačių duomenų projekcijos medų savybių tyrimas ir funkcionalumo gerinimas. Daktaro disertacija. Matematikos ir informatikos institutas.
48. Mayer, R. (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, Springer Verlag, Heidelberg, pp. 357–366.

49. Mayer, R., Rauber, A. (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, Springer Verlag, Heidelberg, pp. 238–246.
50. Mayer, R., Rauber, A. (2011). Data Mining with the Java SOMToolbox. Prieiga internete: <http://www.ifs.tuwien.ac.at/dm/somtoolbox/index.html>
51. Marinai, S. (2011). Text from Early Printed Books. International Journal on Document Analysis and Recognition – IJDAR, vol. 14, no. 2, pp. 117–129.
52. Merkevičius, E. (2008). Savitvarkių neuroninių tinklų-diskriminantinio modelio tyrimai kredito rizikos vertinimo sprendimų paramos sistemoje. Daktaro disertacija. Vilniaus universitetas.
53. Merkevičius, E., Garšva, G., Simutis, R. (2007). Neuro-discriminate Model for the Forecasting of Changes of Companies Financial Standings on the Basis of Self-organizing Maps. Lecture Notes in Computer Science Volume 4488, pp. 439–446.
54. Moehrmann, J., Burkovski, A., Baranovskiy, E., Heinze, G. A., Rapoport, A., Heidemann, G. (2011). A Discussion on Visual Interactive Data Exploration using Self-Organizing Maps. In: J. Laaksonen, T. Honkela (Eds.) Advances in Self-Organizing Maps – WSOM 2011, Lecture Notes in Computer Science, vol 6731, Springer Verlag, Heidelberg, pp. 238–246.
55. Molytė, Alma (2011). Vektorių kvantavimo metodų jungimo su daugiamatėmis skalėmis analizė. Daktaro disertacija. Vilniaus universitetas.
56. Ontrup, J., Ritter, H. (2001). Hyperbolic Self-Organizing Maps for Semantic Navigation. NIPS 2001: 1417–1424.

57. Porter, M. F. (1980). An Algorithm for Suffix Stripping. Program, 14: 130–137.
58. Pragarauskaitė, J. (2013). Dažnų sekų analizė sprendimų priėmimui labai didelėse duomenų bazės. Daktaro disertacija. Vilniaus universitetas.
59. Prakash, A. (2013). Reconstructing Self Organizing Maps as Spider Graphs for Better Visual Interpretation of Large Unstructured Datasets. Infosys Lab Briefings 11 (1).
60. Rauber, A., Merkl, D., Dittenbach, M. (2000). The Growing Hierarchical Self-Organizing Map. In: Proceedings of the International Joint Conference on Neural Networks 2000 (IJCNN'2000), 24–27.
61. Rauber, A., Merkl, D., Dittenbach, M. 2002. The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High Dimensional Data. IEEE Transactions on Neural Networks 13 (6), 1331–1341.
62. Ringienė, L. (2014). Hibridinis neuroninis tinklas daugiamačiams duomenims vizualizuoti. Daktaro disertacija. Vilniaus universitetas.
63. Saarikoski, J. (2014). On text document classification and retrieval using self-organising maps. Doctoral dissertation. The School of Information Sciences of the University of Tampere. Finland.
64. Seimas of the Republic of Lithuania (2013). Prieiga internete: http://www3.lrs.lt/dokpaieska/forma_l.htm
65. Sihag, K., Kumar, S. V. (2013). Graph based Text Document Clustering by Detecting Initial Centroids for k-means. International Journal of Computer Applications (0975–8887), vol. 62, no. 19, pp. 1–4.
66. Sjoberg, M., Laaksonen, J. (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, Springer Verlag, Heidelberg, pp. 338–347.
67. SOM-analyzer (2014). Prieiga internete: <https://sites.google.com/site/somanalyzer/>

68. SOMVis (2014). Prieiga internete:
<http://www.ifs.tuwien.ac.at/dm/somvis-matlab/index.html>
69. Strickert, M., Hammer B. (2004). Self-Organizing Context Learning. In European Symposium on Artificial Neural Networks, pp. 39–44.
70. Strickert, M., Hammer, B. (2005). Merge SOM for temporal data. Neurocomputing 64: 39–72.
71. Tan, H. S., George, S. E. (2004). Investigating Learning Parameters in a Standard 2-D SOM Model to Select Good Maps and Avoid Poor OnesAI 2004, Lecture Notes in Artificial Intelligence, vol. 3339,
pp. 425–437.
72. Ultsch, A., Siemon, H. (1989). Exploratory Data Analysis: Using Kohonen Networks on Transputers. Technical Report 329, Univ. of Dortmund, Dortmund, Germany.
73. Ultsch, A. (2003). Maps for the Visualization of High Dimensional Data Spaces. In Proc. WSOM'03, Japan.
74. Ultsch, A., Moerchen, F. (2005). ESOM-Maps: Tools for Clustering, Visualization, and Classification with Emergent SOM, Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, no. 46. Prieiga internete:
<http://databionic-esom.sourceforge.net>
75. Vasighi, M., Kompany-Zareh, M. (2013). Classification Ability of Self Organizing Maps in Comparison with Other Classification Methods. MATCH Commun. Math. Comput. Chem. 70, pp. 29–44.
76. Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J. (2005). SOM Toolbox for Matlab 5. Prieiga internete:
<http://www.cis.hut.fi/projects/somtoolbox/about.shtml>
77. Villa-Vialaneix N., Bendhaiba L., Olteanu M. (2013) SOMbrero: SOM Bound to Realize Euclidean and Relational Outputs. *R* package
version 0.4.
78. Viscovery SOMine 6.0 (2014). Prieiga internete:
<http://www.viscovery.net/self-organizing-maps.>

79. Voegtlin, T. (2002). Recursive Self-Organizing Maps. Neural Networks 15 (8-9), 979–992.
80. Weber, M., Teeling, H., Huang, S., Waldmann, J., Kassabgy, M., Fuchs, BM., Klindworth, A., Klockow, C., Wichels, A., Gerdts, G., Amann, R., Glockner, FO. (2010). Practical Application of Self-Organizing Maps to interrelate biodiversity and functional data in NGS-based environmental metagenomics.
81. Wehrens, R., Buydens, L. M. C. (2007). Self- and Super-organizing Maps in R: The kohonen Package. Journal of Statistical Software. Vol. 21 Issue 5.
82. Witten, I. H., Frank, E., Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition.
83. Yusob, B., Mariyam Shamsuddin S. H., Nuzly, H., Hamed, A. (2013). Spiking Self-organizing Maps for Classification Problem. Procedia Technology 11, pp. 57–64.
84. Zeimpekis, D., Gallopoulos, E. (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.
85. Zeng, H. J., Wang, X. H., Chen, Z., Lu, H., Ma, W. Y. (2003). CBC: clustering based text classification requiring minimal labeled data. Third IEEE International Conference on Data Mining, ICDM 2003, pp. 443–450.
86. Zhang, H., Xiao, W. X. (2012). Clustering Based Two-stage Text Classification Requiring Minimal Training Data. International Conference on Systems and Informatics (ICSAI 2012), pp. 2233–2237.
87. Zinkevičius, V. (2004). Lemo – lietuvių kalbos leksika ir morfologija. Prieiga internete: <http://donelaitis.vdu.lt/~vytas/lemo/liet/lemo_pasiimt.htm>

Autoriaus publikacijų sąrašas disertacijos tema

Straipsniai recenzuojamuose periodiniuose mokslo leidiniuose:

A 1. Stefanovič P., Kurasova O. (2009). Saviorganizuojančių neuroninių tinklų sistemų lyginamoji analizė. Informacijos mokslai. ISSN 1392-0561. T. 50, pp. 334–339.

A 2. Stefanovič, P., Kurasova, O. (2011). Visual analysis of self-organizing maps. Nonlinear Analysis: Modelling and Control. Vol. 16, no. 4. ISSN 1392-5113 pp. 488–504 (Impact Factor 2013: 0,914).

A 3. Stefanovič, P., Kurasova, O. (2013). Tekstinių dokumentų panašumų paieška naudojant saviorganizuojančius neuroninius tinklus ir k vidurkių metodą. Informacijos mokslai. T. 65, ISSN 1392-0561 pp. 24–33.

A 4. Stefanovič, P., Kurasova, O. (2014). Creation of text document matrices and visualization by SOM. Information Technology and Control / Kauno technologijos universitetas. Vol. 43, no. 1. ISSN 1392-124X pp. 37–46 (Impact Factor 2013: 0,813).

A 5. Stefanovič, P., Kurasova, O. (2014). Investigation on learning parameters of self-organizing maps. Baltic Journal of Modern Computing. Vol. 2, no. 2. ISSN 2255-8942 pp. 45–55.

Straipsniai konferencijų medžiagoje:

B 1. Stefanovič, P., Kurasova, O. (2011). Influence of Learning Rates and Neighboring Functions on Self-Organizing Maps. In: J. Laaksonen, T. Honkela (Eds.). Advances in Self-Organizing Maps: 8th International Workshop, WSOM 2011, Espoo, Finland, June 13–15, 2011: Proceedings. Book Series: Lecture Notes in Computer Science. Vol. 6731. ISBN 9783642215 pp. 141–150.

B 2. Kurasova, O., Marcinkevičius, V., Medvedev, V., Rapečka, A., and Stefanovič, P. (2014). Strategies for Big Data Clustering. Proceedings

of 26th IEEE International Conference on Tools with Artificial Intelligence, ISSN 1082-3409 pp. 740–747.

Santraukos tarptautinių konferencijų santraukų rinkiniuose:

C 1. Stefanovič P., Kurasova O. (2012). Text mining and visualization with self-organizing maps. EURO 25: 25th European Conference on Operational Research: Abstracts Book, Vilnius, 8–11 July, 2012. pp. 252.

C 2. Stefanovič, P. (2013). Finding scientific article similarities by selforganizing maps. EUROINFORMS: 26th European Conference on Operational Research: Abstract Book, Rome, 1–4 July, 2013, pp. 155.

Pavel Stefanovič

SAVIORGANIZUOJANČIŲ NEURONINIŲ TINKLŲ VIZUALIZAVIMAS IR JO KOKYBĖS NUSTATYMAS

Daktaro disertacija

Fiziniai mokslai,
Informatika (09 P)

Redaktorė Agnė Bolytė

Pavel Stefanovič

VISUALIZATION OF SELF-ORGANIZING MAPS AND ESTIMATION OF THEIR QUALITY

Doctoral Dissertation

Physical sciences,
Informatics (09 P)

Editor Agnė Bolytė