VILNIAUS UNIVERSITETAS

LAURA RINGIENĖ

HIBRIDINIS NEURONINIS TINKLAS DAUGIAMAČIAMS DUOMENIMS VIZUALIZUOTI

Daktaro disertacija
Technologijos mokslai, informatikos inžinerija (07 T)

Vilnius, 2014

Disertacija rengta 2008–2013 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinis vadovas:
prof. habil. dr. Gintautas Dzemyda (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T)

Padėka

Nuoširdžiai dėkoju darbo vadovui prof. habil. dr. Gintautui Dzemydai už vertingas mokslines konsultacijas, nuoseklų vadovavimą, pagalbą ir kantrybę rengiant šią disertaciją.

Esu dėkinga disertacijos recenzentams prof. dr. Daliui Navakauskui ir doc. dr. Olgai Kurasovai, o taip pat ir dr. Viktorui Medvedevui bei Robertui Juodaičiui atidžiai perskaičiusiems disertaciją ir pateikusiems vertingų pastabų bei patarimų, padėjusių pagerinti šio darbo kokybę. Taip pat nuoširdžiai dėkoju Janinai Kazlauskaitei už pagalbą rengiant disertacijos santraukos tekstą.

Dėkoju Vilniaus universiteto Matematikos ir informatikos instituto Sistemų analizės ir Atpažinimo procesų skyrių kolektyvams už bendradarbiavimą, pagalbą ir palaikymą.

Nuoširdžiai dėkoju vyrui, sūneliui, dukrytei ir tėvams už jų paramą, moralinį palaikymą, kantrybę ir supratingumą.

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

Laura Ringienė

Reziumė

Šio darbo tyrimų sritis yra duomenų tyryba remiantis daugiamačių duomenų vizualia analize. Tai leidžia tyrėjui betarpiškai dalyvauti duomenų analizės procese, geriau pažinti sudėtingus duomenis ir priimti geriausius sprendimus. Disertacijos tikslas yra sukurti metodą tokios duomenų projekcijos radimui plokštumoje, kad tyrėjas galėtų pamatyti ir įvertinti daugiamačių taškų tarpgrupinius panašumus/skirtingumus. Šiam tikslui pasiekti yra pasiūlytas radialinių bazinių funkcijų ir daugiasluoksnio perceptrono, turinčio „butelio kaklelio“ neuroninio tinklo savybes, junginys. Naujas tinklas naudojamas vizualiai daugiamačių duomenų analizei, kai atidėjimui plokštumoje arba trimatėje erdvėje taškai gaunami paskutinio paslėpto neuronų sluoksnio išėjimuose, kai į tinklo įėjimą paduodami daugiamačiai duomenys. Šio tinklo ypatybė yra ta, kad gautas vaizdas plokštumoje labiau atspindi bendrą duomenų struktūrą (klasteriai, klasterių tarpusavio artumas, taškų tarpklasterinis panašumas) nei daugiamačių taškų tarpusavio išsidėstymą.

Disertaciją sudaro 5 skyriai ir literatūros sąrašas. Bendra disertacijos apimtis yra 130 puslapių, 59 paveikslai ir 32 lentelės.

Tyrimų rezultatai publikuoti 3 periodiniuose recenzuojamuose moksliniuose leidiniuose.

Tyrimų rezultatai buvo pristatyti ir aptarti 5 nacionalinėse ir tarptautinėse konferencijose Lietuvoje.

Abstract

The area of research is data mining based on multidimensional data visual analysis. This allows researcher to participate in the process of data analysis directly, to understand the complex data better and to make the best decisions. The objective of the dissertation is to create a method for making a multidimensional data projection on the plane such that the researcher could see and assess the intergroup similarities and differences of multidimensional points. In order to achieve the target, a new hybrid neural network is proposed and investigated. This neural network integrates the ideas both of the radial basis function neural network and that of a multilayer perceptron, which has the properties of a "bottleneck" neural network. The new network is used for the visual analysis of multidimensional data in such a way that the output values of the neurons of the last hidden layer are the two-dimensional or three-dimensional projections of the multidimensional data, when the multidimensional data is given to the network. A peculiarity of the network is that the visualization results on the plane reflect the general structure of the data (clusters, proximity between clusters, intergroup similarities of points) rather than the location of multidimensional points.

The dissertation consists of 5 chapters and references. The scope of the work is 130 pages that include 59 figures and 32 tables.

The main results of the dissertation were published in 3 periodical scientific publications.

The main results of the work have been presented and discussed at 5 national and international conferences.

Žymėjimai

$\bar{a}_{K_j}$	didžiausias atstumas tarp klasterio K_j taškų (antrasis vizualizavimo kokybės kriterijus)
$\hat{a}$	mažiausias atstumas tarp gretimų klasterių (trečiasis vizualizavimo kokybės kriterijus)
α	konstanta, naudojama radialinių bazinių funkcijų pločio parametrui apskaičiuoti
k	klasterių skaičius
$K_1, K_2, \ldots, K_k$	klasteriai
κ_q	mažiausias atstumas tarp skirtingų klasterių taškų (antrasis atrankos kriterijus)
m	objektų skaičius duomenų rinkinyje
m_{K_j}	objektų klasteryje K_j skaičius
n	objektą apibūdinančių parametrų skaičius
n_v	neuronų skaičius v-ajame paslėptame sluoksnyje
$\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$	j-ojo klasterio centras, $\mu_j \in \mathbb{R}^n$
P^v	paslėptų neuronų sluoksnis
P^1	pirmasis paslėptas neuronų sluoksnis REGM tinkle
P^2	mažasis sluoksnis REGM tinkle
$\mathbb{R}^n$	n-matė erdvė
s	neuronų skaičius išėjimų sluoksnyje
σ	radialinių bazinių funkcijų pločio parametras
$T_i = (t_{i1}, t_{i2}, \ldots, t_{is})$	i-oji norima tinklo atsako reikšmė, $T_i \in \mathbb{R}^s$
τ	dispersijų vidurkis
u	iteracijos numeris
V	paslėptų neuronų sluoksnių skaičius
w_{ij}	svoris
x_{ij}	objektą X_i apibūdinančio j-ojo parametro reikšmė
$X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$	i-asis n-matis taškas, $X_i \in \mathbb{R}^n$
$\mathbf{X} = (X_1, X_2, \ldots, X_m)$	analizuojamų duomenų matrica, kurios i-oji eilutė yra n-matis taškas X_i

X	įėjimų sluoksnis
χ	klasterių išsaugojimas duomenyse po tinklo apmokymo (pirmasis atrankos kriterijus)
$\parallel X_i - X_j \parallel$	Euklidinis atstumas tarp taškų X_i ir X_j
Y	išėjimų sluoksnis
$Y_i = (y_{i1}, y_{i2}, \ldots, y_{is})$	tinklo išėjime gaunamas i-asis s-matis taškas
Z	radialinių bazinių funkcijų sluoksnis
$Z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$	po transformacijos gautas i-asis taškas, $Z_i \in \mathbb{R}^k$, $k < n$

Santrumpos

DNT	Dirbtiniai neuroniniai tinklai (angl. *Artificial neural networks*)
MDS	Daugiamačių skalių metodas (angl. *Multidimensional scaling*)
MLP	Daugiasluoksnis perceptronas arba daugiasluoksnis tiesioginio sklidimo neuroninis tinklas (angl. *Multilayer perceptron*)
RBF	Radialinių bazinių funkcijų neuroninis tinklas (angl. *Radial basis function network*)
REGM	Radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginys (angl. *Radial basis function, Eksponential function, Gaussian function, Multilayer perceptron*)

Turinys

1. Įvadas

1.1. Tyrimų sritis

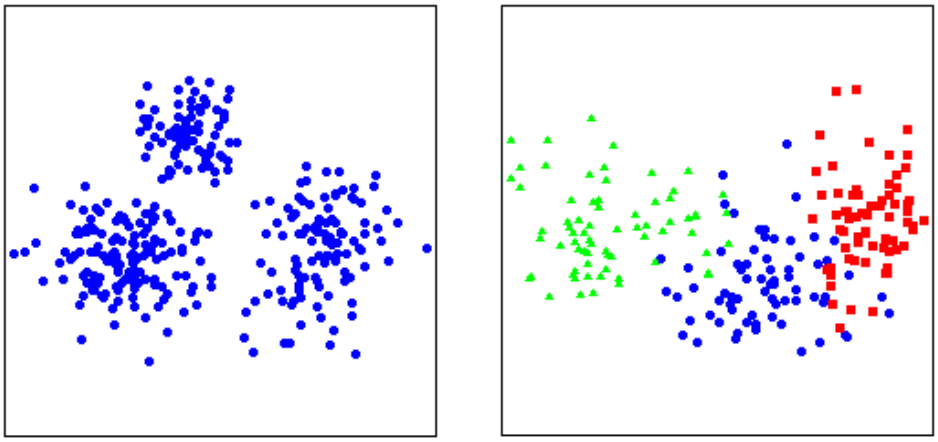
Sparčiai vystantis šiuolaikinėms technologijoms labai didėja kaupiamų duomenų apimtys įvairiose srityse: technikoje, ekonomikoje, medicinoje, ekologijoje ir daugelyje kitų. Duomenys kaupiami tam, kad vėliau iš jų būtų galima gauti naujų žinių, pavyzdžiui, prognozuoti būsimą veiklą, identifikuoti kritinius atvejus, apibendrinti. Tačiau, turimus labai didelės apimties duomenis (dažniausiai vadinamus daugiamačiais duomenimis) žmogui savarankiškai suvokti ir interpretuoti labai sudėtinga. Tam tikslui yra kuriami įvairūs duomenų tyrybos metodai, kurie sprendžia įvairius uždavinius: suskirsto duomenis į grupes, nustato duomenų struktūrą, randa tarpusavio ryšius ar net išskirtinumus, ir pan. Čia paminėtų uždavinių sprendimą padeda (palengvina) surasti daugiamačių duomenų vizualizavimas dvimatėje arba trimatėje erdvėje. Šio darbo tyrimų sritis yra duomenų tyryba remiantis daugiamačių duomenų vizualia analize.

1.2. Darbo aktualumas

Šioje disertacijoje tiriami tokie daugiamačiai duomenys, kurie aprašo objektų (žmonių, įrenginių, augalų, gamtos reiškinių ir kt.) rinkinius, kuriuos charakterizuoja tam tikri skaitiniai požymiai (parametrai, savybės). Objektų, sudarančių konkretų analizuojamą duomenų rinkinį, skaičius m yra baigtinis. Tam tikras požymių reikšmių rinkinys nusako vieną konkretų analizuojamo duomenų rinkinio objektą $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1, m}$, čia n yra požymių skaičius, i yra objekto numeris. Objektai X_i dar gali būti interpretuojami kaip n-mačiai taškai, o požymiai $x_1, x_2, \ldots, x_n$ – taškų koordinatėmis. Analizuojamų duomenų rinkinį galima aprašyti kaip matricą $\mathbf{X} = \{X_1, X_2, \ldots, X_m\} = \{x_{ij}, i = \overline{1, m}, j = \overline{1, n}\}$, kurios i-oji eilutė yra n-matės euklidinės erdvės taškas $X_i \in \mathbb{R}^n$ (Dzemyda ir kt., 2013).

Daugiamačių duomenų vizualizavimui jau yra sukurta nemažai metodų, bet jie ir toliau sparčiai vystomi siekiant lengvinti duomenų interpretavimą ir suvokimą (Dzemyda ir kt., 2013). Taip pat šie metodai yra realizuoti daugelyje programų sistemų: Orange (Podpečan ir kt., 2012), Matlab (R2009b, The MathWorks, http://www.mathworks.se/), Weka (Hall ir kt., 2009), ir kt. Vizualizavimo metodai turimus daugiamačius duomenis pateikia žmogui suvokiamoje erdvėje (dvimatėje arba trimatėje) perteikiant taškų išsidėstymą, t. y. išlaikant jų panašumus ir skirtingumus. Tačiau atsiranda poreikis vizualiai įvertinti duomenų rinkinio struktūrą ir

savybes: susidariusias grupes, žymiai išsiskiriančius objektus, objektų panašumus/skirtingumus, ir pan. Retame duomenų rinkinyje aiškiai atsiskiria objektų grupės, t. y. matoma riba tarp objektų grupių, kaip pateikta 1.1a paveiksle, kuriame į plokštumą vizualizuotas E.coli bakterijų duomenų rinkinys (angl. *ecoli data set*) (Horton ir Nakai, 1996). Duomenų rinkinį sudaro 336 bakterijos, kurios apibūdintos 7 požymiais. Matome tris objektų grupes, nors praktiškai jų yra daugiau. Dažniausiai skirtingų objektų grupės yra susiglaudusios arba net vienos grupės objektai pakliūna tarp kitos grupės objektų. Kaip pavyzdys pateikiamas Kviečių grūdų duomenų rinkinys (angl. *wheat seeds data set*) (Charytanowicz ir kt., 2010) vizualizuotas į plokštumą 1.1b paveiksle. Duomenų rinkinį sudaro 210 kviečių grūdų, kurie apibūdinti 7 požymiais. Vaizdumo dėlei skirtingų grupių objektai pavaizduoti skirtingomis spalvomis. Atsiranda poreikis atskirti vieną grupę nuo kitos arba išskirti objektų grupeles, kurios reikalauja nuodugnesnio tyrimo. Pavyzdžiui, gali kilti poreikis kiekvienoje Kviečių grūdų grupėje išskirti grūdus, kurie turi daugiausia panašumo su kitos grupės grūdais arba atvirkščiai – išgryninti konkrečios grupės grūdus.



(a) E.coli bakterijos **(b)** Kviečių grūdai

1.1 pav. Duomenų rinkinių vizualizavimo pavyzdžiai

Duomenų rinkinį papildžius nauju objektu ir norint jį pridėti turimame paveiksle tarp anksčiau atvaizduotų objektų, tenka arba iš naujo rasti visų duomenų projekcijas plokštumoje, jei duomenų vizualizavimas buvo atliktas klasikiniais vizualizavimo metodais, arba naudoti tam (naujų taškų atidėjimui) skirtus metodus, kurie yra netikslūs (pavyzdžiui, trianguliacijos metodas (Karbauskaitė ir Dzemyda, 2006)). Naujus objektus atitinkančių taškų atidėjimui plokštumoje sėkmingai taikomi ir dirbtiniai neuroniniai tinklai.

1.3. Darbo tikslas ir uždaviniai

Disertacijos tikslas yra sukurti metodą tokios duomenų projekcijos radimui plokštumoje, kad tyrėjas galėtų pamatyti ir įvertinti daugiamačių taškų tarpgrupinius panašumus/skirtingumus.

Šiam tikslui pasiekti buvo sprendžiami tokie uždaviniai:

1. Analitiškai apžvelgti su darbo tikslu susijusias duomenų tyrybos metodų grupes: vizualizavimo metodų, klasterizavimo metodų ir dirbtinių neuroninių tinklų, o taip pat sukurtus radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginius.

2. Išanalizuoti dirbtinių neuroninių tinklų galimybes daugiamačiams duomenims vizualizuoti.

3. Optimizuoti radialinių bazinių funkcijų pritaikomumą daugiamačių duomenų matmenų mažinimui remiantis gautų rezultatų vizualia analize.

4. Pasiūlyti ir ištirti radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginį (hibridinį tinklą REGM) daugiamačiams duomenims vizualiai tirti, siekiant įvertinti tarpgrupinius panašumus arba skirtingumus.

5. Pasiūlyti vizualizavimo kokybės kriterijus, kurie padėtų įvertinti gautus vizualizavimo rezultatus.

6. Pasiūlyti kriterijus kokybiškai apmokyto REGM tinklo atrankai.

1.4. Mokslinis naujumas

Šiame darbe pasiūlytas ir ištirtas naujas hibridinis neuroninis tinklas, kuris savyje integruoja ir radialinių bazinių funkcijų neuroninio tinklo, ir daugiasluoksnio perceptrono, turinčio „butelio kaklelio“ neuroninio tinklo savybes, idėjas. Tai ir yra disertacijos mokslinis naujumas. Toliau šis tinklas bus vadinamas REGM tinklu. Trumpai detalizuosime idėją.

REGM tinklas sudarytas iš dviejų dalių. Pirmojoje dalyje radialinės bazinės funkcijos, kurios atlieka tam tikrą n-matės erdvės $\mathbb{R}^n$ taškų transformavimą į norimo matmens erdvę $\mathbb{R}^k$, $k < n$. Radialinių bazinių funkcijų neuroniniuose tinkluose funkcijų reikšmių apskaičiavimui naudojamą pločio parametrą literatūroje siūloma parinkti pagal tinklo daromą paklaidą. Tačiau šiame darbe pasiūlytame REGM tinkle naudojamos tik

radialinės bazinės funkcijos, todėl tinkamam pločio parametro parinkimui tenka ieškoti kitokių būdų. Šioje disertacijoje pločio parametrą siūloma parinkti pagal objektų išsibarstymą klasteriuose ir vidutinį atstumą tarp tų klasterių centrų. Radialinių bazinių funkcijų pritaikomumas daugiamačių duomenų matmenų mažinimui optimizuotas remiantis gautų rezultatų vizualia analize.

Antroje sudedamojoje REGM tinklo dalyje yra specialios struktūros daugiasluoksnis perceptronas, kurio paskutinis paslėptas sluoksnis yra sudarytas iš nedidelio neuronų skaičiaus (2 arba 3). REGM tinklo paskirtis yra atlikti daugiamačių duomenų projekciją į dvimatę arba trimatę erdvę (projekcija gaunama būtent paskutiniame paslėptame sluoksnyje), kuomet objektus atitinkančius taškus galima stebėti vizualiai. Vizualizuotuose duomenyse atsiskleidžia ir juose esančių klasterių savybės, nes žinios apie klasterių sudėtį, objektus sudarančius klasterius, gaunamos prieš mokant REGM tinklą ir naudojamos to tinklo mokymo metu.

Po REGM tinklo apmokymo vizualiai pateiktos daugiamačių duomenų projekcijos yra įvertinamos šioje disertacijoje užsibrėžtais vizualizavimo kokybės kriterijais. Siekiant galimai geriausio vizualaus duomenų atvaizdavimo, tikslinga REGM tinklą apmokyti keletą kartų ir pasirinkti geriausią projekciją. Spartesniam geriausios duomenų rinkinio projekcijos radimui pagal užsibrėžtus vizualizavimo kokybės kriterijus yra pasiūlyti atrankos kriterijai.

1.5. Ginamieji teiginiai

1. Radialinių bazinių funkcijų neuroninio tinklo ir specialios struktūros daugiasluoksnio perceptrono idėjų apjungimas leidžia ieškoti tokios duomenų projekcijos plokštumoje, kad tyrėjas galėtų pamatyti ir įvertinti daugiamačių taškų tarpgrupinius panašumus/skirtingumus.

2. Radialinių bazinių funkcijų pločio parametrą REGM tinklui galima nustatyti pagal objektų išsibarstymą klasteriuose ir vidutinį atstumą tarp tų klasterių centrų.

3. Pasiūlyti trys vizualizavimo kokybės kriterijai įvertina apmokyto tinklo REGM vizualizavimo rezultatus.

4. Jei REGM tinklas apmokomas keletą kartų, geriausios duomenų rinkinio projekcijos pasirinkimą palengvina pasiūlyti du atrankos kriterijai, kuriuos naudojant atranka gali būti automatizuota.

1.6. Darbo rezultatų aprobavimas

Tyrimų rezultatai publikuoti 3 periodiniuose recenzuojamuose moksliniuose leidiniuose:

- **Ringienė, L.**, Dzemyda, G. Daugiamačių duomenų požymių mažinimas naudojantis eksponentine koreliacine funkcija. *Jaunųjų mokslininkų darbai*. Vilnius: Vilniaus universitetas. ISSN 2029-9958. 2013, Nr. 1, p. 152–158.
- **Ringienė, L.**, Dzemyda, G. Multidimensional data visualization based on the exponential correlation function. *Baltic Journal of Modern Computing*. Riga: University of Latvia. ISSN 2255-8942. 2013, Vol. 1, No. 1, p. 9–28.
- **Ringienė, L.**, Dzemyda, G. Specialios struktūros daugiasluoksnis perceptronas daugiamačiams duomenims vizualizuoti. *Informacijos mokslai*. ISSN 1392-0561. 2009, T. 50, p. 358–364.

Tyrimų rezultatai buvo pristatyti ir aptarti šiose nacionalinėse ir tarptautinėse konferencijose Lietuvoje:

1. „Kompiuterininkų dienos – 2009“. Kaunas, Lietuva. 2009 m. rugsėjo 25–26 d.
2. 15th International Conference Mathematical Modelling and Analysis. Druskininkai, Lietuva. 2010 m. gegužės 26–29 d.
3. 10th EUROPT Workshop on Advances in Continuous Optimization. Šiauliai, Lietuva. 2012 m. liepos 5–7 d.
4. Trečioji jaunųjų mokslininkų konferencija „Tarpdalykiniai tyrimai fiziniuose ir technologijos moksluose – 2012“. Vilnius, Lietuva. 2013 m. vasario 12 d.
5. 5th International Workshop „Data Analysis Methods for Software Systems“, Druskininkai, Lietuva. 2013 m. gruodžio 5–7 d.

1.7. Disertacijos struktūra

Disertaciją sudaro 5 skyriai ir literatūros sąrašas. Disertacijos skyriai: Įvadas, Duomenų tyrybos metodai susiję su darbo tikslu ir uždaviniais, REGM tinklas daugiamačiams duomenims vizualizuoti, Eksperimentiniai

tyrimai, Apibendrinimas ir bendrosios išvados. Papildomai disertacijoje pateikta: naudotų žymėjimų ir santrumpų sąrašas. Bendra disertacijos apimtis yra 130 puslapių, kuriuose pateikti 59 paveikslai ir 32 lentelės. Disertacijoje remtasi 101 literatūros šaltiniu.

2. Duomenų tyrybos metodai susiję su darbo tikslu ir uždaviniais

Sparčiai vystantis šiuolaikinėms technologijoms labai didėja kaupiamų duomenų apimtys įvairiose srityse: technikoje, ekonomikoje, medicinoje, ekologijoje ir daugelyje kitų. Duomenys kaupiami tam, kad vėliau iš jų būtų galima gauti naujų žinių, pavyzdžiui, prognozuoti būsimą veiklą, identifikuoti kritinius atvejus, apibendrinti. Šie kaupiami duomenys vadinami daugiamačiais duomenimis. Šioje disertacijoje tiriami tokie daugiamačiai duomenys, kurie aprašo objektų (žmonių, įrenginių, augalų, gamtos reiškinių ir kt.) rinkinius, kuriuos charakterizuoja tam tikri skaitiniai požymiai (parametrai, savybės). Objektų, sudarančių konkretų analizuojamą duomenų rinkinį, skaičius m yra baigtinis. Tam tikras požymių reikšmių rinkinys nusako vieną konkretų analizuojamo duomenų rinkinio objektą $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1,m}$, čia n yra požymių skaičius, i yra objekto numeris. Objektai X_i dar gali būti interpretuojami kaip n-mačiai taškai, o požymiai $x_1, x_2, \ldots, x_n$ – taškų koordinatėmis. Analizuojamų duomenų rinkinį galima aprašyti kaip matricą $\mathbf{X} = \{X_1, X_2, \ldots, X_m\} = \{x_{ij}, i = \overline{1,m}, j = \overline{1,n}\}$, kurios i-oji eilutė yra n-matės euklidinės erdvės taškas $X_i \in \mathbb{R}^n$ (Dzemyda ir kt., 2013).

Didelės apimties ir daug požymių turinčius daugiamačių duomenų rinkinius žmogui suvokti ir analizuoti yra sudėtinga, todėl būtina pasinaudoti duomenų tyrybos metodais, kurie palengvina duomenų rinkinio suvokimą ir interpretavimą. Universalaus daugiamačių duomenų tyrybos metodo, kuris palengvintų suvokti ir interpretuoti bet kokius (skaitinius ir (ar) tekstinius, mažesnės arba labai didelės apimties ir pan.) turimus daugiamačius duomenis, bei rastų sprendimą bet kokiam uždaviniui (identifikavimas, apibendrinimas, prognozavimas ir pan.), sukurti neįmanoma. Todėl sukurti ir dar kuriami duomenų tyrybos metodai skirti specialiems taikomiesiems uždaviniams spręsti. Esamus tyrybos metodus galima suskirstyti į grupes, pagal sprendžiamus uždavinius (Kantardzic, 2011):

- statistiniai metodai;
- klasterizavimo metodai;
- dirbtiniai neuroniniai tinklai;
- genetiniai algoritmai;
- vizualizavimo metodai;
- ir kt.

Dažniausiai yra taikomi kelių grupių metodai, kad geriau suvoktume ir įvairiapusiškai išanalizuotume turimą daugiamačių duomenų rinkinį. Šioje disertacijoje pasiūlytas radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginys apima vizualizavimo, klasterizavimo ir dirbtinių neuroninių tinklų grupėse pateiktus metodus. Todėl šias grupes aptarsime šiek tiek plačiau.

2.1. Daugiamačių duomenų vizualizavimas

Duomenų tyrybos tikslas – padėti žmogui suprasti ir interpretuoti turimus daugiamačių duomenų rinkinius. Duomenų rinkinių supratimą palengvina sugrupavimas į grupes, struktūros nustatymas, tarpusavio ryšių radimas ir pan. Šiuos tikslus pasiekti padeda daugiamačių duomenų vizualizavimas. Vizualizavimas – tai grafinis informacijos pateikimas. Grafiškai pateikta informacija daug lengviau ir greičiau suprantama (suvokiama) nei tekstinė. Taip pat ji palengvina naujų žinių atradimą (Dzemyda ir kt., 2008).

Daugiamačius duomenis vizualizuoti galima įvairiai: brėžti histogramas, prognozės grafikus arba pasinaudoti vizualizavimo metodais, kurie padeda nustatyti ar įvertinti daugiamačių duomenų struktūrą (susidariusias grupes, itin išsiskiriančius objektus, panašumus tarp analizuojamų objektų ar jų grupių ir pan.).

Vizualizavimo metodai yra plėtojami dviem kryptimis:

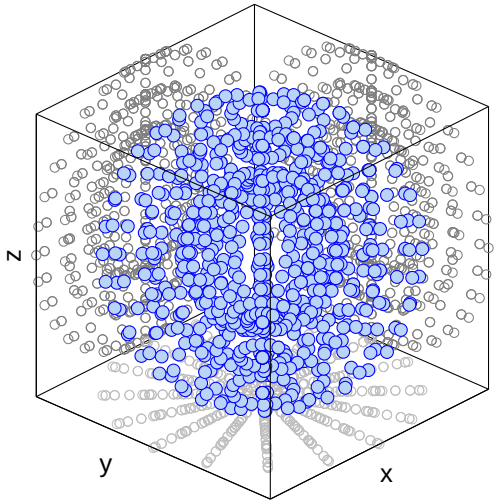
1. Tiesioginio vizualizavimo metodai, kuriuose kiekvienas objekto parametras pateikiamas tam tikra vizualia forma.
2. Projekcijos, dar vadinami matmenų skaičiaus mažinimo metodai, transformuoja turimą duomenų rinkinį iš n-matės erdvės $\mathbb{R}^n$ į d-matę erdvę $\mathbb{R}^d$, $d < n$.

Šioje disertacijoje naudojami tik daugiamačių duomenų projekcijos metodai, todėl tiesioginio vizualizavimo metodai aptarinėjami nebus. Apie tiesioginio vizualizavimo metodus informacijos galima rasti Medvedev (2007), Bernatavičienė (2008), Kantardzic (2011), Dzemyda ir kt. (2013) darbuose.

2.1.1. Projekcijos metodai

Duomenų rinkinio $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1, m}$, kuris išsidėstęs n-matėje erdvėje, kai $n > 3$, tiesiogiai pamatyti neįmanoma. Tačiau galima rasti šio duomenų rinkinio projekciją $Y_i = (y_{i1}, y_{i2}, \ldots, y_{id})$, $i = \overline{1, m}$, į dvimatę arba trimatę erdvę pasinaudojant projekcijos metodais. Šių metodų tikslas – pateikti daugiamačius duomenis mažesnio skaičiaus matmenų

erdvėje (R^2 arba R^3) taip, kad kiek galima tiksliau būtų išlaikyta pradinių duomenų struktūra (Dzemyda ir kt., 2013). Kaip pavyzdys, 2.1 paveiksle pateikiama sfera ir jos taškų projekcijos (xy, xz ir yz plokštumose) dvimatėje erdvėje. Sferos taškai buvo automatiškai sugeneruoti intervale $[-1;\ 1]$ trimatėje erdvėje ($m = 726$, $n = 3$), ir pažymėti ∘. Taškų projekcijos pažymėtos ∘.



2.1 pav. Sferos taškų projekcijos dvimatėse plokštumose

2.1 paveiksle matyti, kad projektuojant duomenis į skirtingas plokštumas gaunamos skirtingos projekcijos. xz ir yz plokštumose gautos trimačio duomenų rinkinio dvimatės projekcijos vizualiai atrodo panašiai (skritulys), o xy plokštumoje gauta projekcija, nuo šių dviejų skiriasi (iš centro į šonus eina spinduliai, tarsi nupiešta snaigė).

Atliekant daugiamačių duomenų projekciją siekiama įgyvendinti du svarbius tikslus: supaprastinti turimą duomenų rinkinį mažinant objektų požymių skaičių ir išlaikyti kiek galima daugiau originalios informacijos (Dzemyda ir kt., 2008).

Taigi pirmiausia reikia apsibrėžti artimumo matą, kuris bus reikalingas išlaikant duomenų struktūrą. Projekcijos metoduose duomenų struktūros artimumo matu dažniausiai yra naudojamas atstumas. Paprastai yra naudojami Minkovskio atstumai (angl. *Minkowski distance*):

$$d(X_i, X_j) = \left\{ \sum_{l=1}^{n} |x_{il} - x_{jl}|^q \right\}^{\frac{1}{q}}. \tag{2.1}$$

Vienas iš dažniausiai naudojamų Minkovskio atstumų vadinamas Euklidiniu atstumu:

$$d(X_i, X_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2}. \tag{2.2}$$

Literatūroje (Kruskal, 1964; Mao ir Jain, 1996; Žilinskas ir Žilinskas, 2008; Marcinkevičius, 2010; Jain, 2010; Dzemyda ir kt., 2013) yra pateikiama ir kitų atstumo skaičiavimo formulių. Pavyzdžiui:

- Mahalanobio atstumas (angl. *Mahalanobis distance*):

$$d(X_i, X_j) = \sqrt{(x_{il} - x_{jl})^T S^{-1} (x_{il} - x_{jl})}, \tag{2.3}$$

  čia S – kovariacinė matrica.

- Kanbera atstumas (angl. *Canberra distance*):

$$d(X_i, X_j) = \sum_{l=1}^{n} \frac{|x_{il} - x_{jl}|}{|x_{il} + x_{jl}|}. \tag{2.4}$$

Yra išskiriamos dvi projekcijos metodų grupės:

1. Tiesinės – ieškoma tiesinės analizuojamų duomenų transformacijos;
2. Netiesinės – ieškoma netiesinės analizuojamų duomenų transformacijos.

Tarkime, kad turime dvimačių taškų duomenų rinkinį $X_i = (x_{i1}, x_{i2})$, $i = \overline{1, m}$, kuriame tarp gretimų taškų atstumai yra vienodi. Šiuos taškus norime atvaizduoti vienmatėje erdvėje, t. y. išdėlioti juos ant tiesės. Tiesinės projekcijos atveju, atstumai tarp taškų projekcijų nebus išlaikyti, o netiesinės projekcijos atveju, atstumai tarp taškų projekcijų bus išlaikyti (Dzemyda ir kt., 2013).

Dažniausiai naudojami tiesinės projekcijos metodai:

1. Pagrindinių komponenčių analizė (angl. *principal component analysis*, PCA). Pagrindinė idėja yra sumažinti duomenų matmenų skaičių atliekant tiesinę transformaciją ir atsisakant dalies po transformacijos gautų naujų komponenčių, kurių dispersijos yra mažiausios (Pearson, 1901; Hotelling, 1933; Jolliffe, 2005; Abdi ir Williams, 2010).
2. Tiesinė diskriminantinė analizė (angl. *linear discriminant analysis*, LDA). Pagrindinė idėja yra n-matės erdvės duomenis transformuoti į mažesnę erdvę, tiesiogiai pasinaudojant žinomomis duomenų klasėmis taip, kad klasių atskiriamumo kriterijaus reikšmė būtų optimali (Duda ir Hart, 1973; Izenman, 2008).

3. Faktorinė analizė (angl. *factor analysis*). Šiame metode daroma prielaida, kad nagrinėjami požymiai priklauso nuo tam tikrų paslėptų faktorių. Metodo tikslas atskleisti tokius ryšius ir daugiamačių duomenų dimensijos mažinimui panaudoti tam tikrą faktoriaus modelį (Harman, 1976; Comrey ir Lee, 2013).

Dažniausiai naudojami netiesinės projekcijos metodai:

1. Daugiamatės skalės (angl. *multidimensional scaling*, MDS). Metodo tikslas – rasti duomenų rinkinio projekciją mažesnio skaičiaus matmenų erdvėje, siekiant išlaikyti analizuojamo rinkinio objektų panašumus. Gautuose vaizduose panašūs objektai išdėstomi arčiau vieni kitų, o skirtingi – toliau vieni nuo kitų (Kruskal, 1964; Borg ir Groenen, 2005; France ir Carroll, 2011).

2. Sammono algoritmas yra vienas iš MDS variantų. Šio algoritmo tikslas – minimizuoti atstumų skirtumus tarp taškų n-matėje erdvėje ir jų projekcijų d-matėje erdvėje (Sammon, 1969; Medvedev, 2007; Sun ir kt., 2012; Dzemyda ir kt., 2013).

3. Pagrindinės kreivės (angl. *principal curves*). Pagrindinė kreivė – tai glodžioji kreivė, brėžiama per duomenų centrinį tašką taip, kad vidutinis atstumas nuo duomenų taškų iki šios kreivės būtų minimalus, t. y. ši kreivė būtų kiek galima arčiau visų duomenų taškų (Hastie ir Stuetzle, 1989; Delicado, 2001; Ataer-Cansizoglu ir kt., 2013).

4. Izometrinis požymių vaizdavimas (angl. *isometric feature mapping*, ISOMAP). Taikant ISOMAP metodą, daroma prielaida, kad pradinėje erdvėje analizuojamus duomenis atitinkantys taškai yra išsidėstę ant mažesnio skaičiaus matmenų netiesinės daugdaros, ir todėl objektų panašumas vertinamas pagal geodezinius atstumus (Tenenbaum ir kt., 2000; Karbauskaitė, 2010).

5. Lokaliai tiesinis atvaizdavimas (angl. *locally linear embedding*, LLE). Šiuo metodu atvaizduojant n-mačius duomenų rinkinius į mažesnio skaičiaus matmenų erdvę, išlaikomi kaimynystės ryšiai tik tarp artimiausių taškų, bet atskleidžiama netiesinės daugdaros globali struktūra (Roweis ir Saul, 2000; Karbauskaitė ir Dzemyda, 2006; Li ir Zhang, 2011).

Disertacijoje tarpinių rezultatų peržiūrėjimui, jei toje vietoje požymių skaičius didesnis už tris, ir norimų tinklo atsako reikšmių nustatymui yra naudojamas daugiamačių skalių netiesinės projekcijos metodas. Todėl šis metodas yra aprašomas plačiau.

2.1.2. Daugiamačių skalių metodas

Daugiamačių skalių (angl. *multidimensional scaling*, MDS) metodas (Borg ir Groenen, 2005) plačiai naudojamas daugiamačių duomenų vizualizavimui. Daugiamačių skalių metodu, ieškoma taškų $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ projekcijų $Y_i = (y_{i1}, y_{i2}, \ldots, y_{id})$ į mažesnio skaičiaus matmenų erdvę R^d, $(d < n)$ (dažniausiai R^2 arba R^3), siekiant išlaikyti analizuojamos aibės objektų panašumus. Atlikus projekciją į mažesnio matmenų vaizdo erdvę, panašūs objektai išdėstomi arčiau vieni kitų, o skirtingi – toliau vieni nuo kitų (Dzemyda ir kt., 2013).

Atstumus tarp taškų X_i ir X_j pažymėkime $d(X_i, X_j)$, o atstumus tarp taškų Y_i ir Y_j pažymėkime $d(Y_i, Y_j)$, $i, j = \overline{1, m}$. Taigi, MDS metodas bando priartinti atstumus $d(Y_i, Y_j)$ prie atstumų $d(X_i, X_j)$. Galimi atstumo skaičiavimo variantai pateikiami 2.1.1. poskyryje. Skaičiuojama kvadratinė paklaidos funkcija, kuri yra minimizuojama. Literatūroje paprasčiausia kvadratinė paklaidos funkcija vadinama *raw Stress* ir užrašoma taip:

$$E_{\text{rawStress}} = \sum_{i<j} w_{ij}(d(Y_i, Y_j) - d(X_i, X_j))^2, \tag{2.5}$$

čia w_{ij} – svoris, kuris yra teigiamas skaičius (Borg ir Groenen, 2005). Dažnai naudojami tokie svoriai w_{ij}:

$$w_{ij} = \frac{1}{\sum_{i<j}(d(X_i, X_j))^2};$$

arba

$$w_{ij} = \frac{1}{d(X_i, X_j)\sum_{k<l} d(X_k, X_l)};$$

arba

$$w_{ij} = \frac{1}{md(X_i, X_j)}.$$

Paprasčiausias atvejis, kai $w_{ij} = 1$.

Kaip jau yra paminėta, kvadratinė paklaidos funkcija algoritmo veikimo metu yra minimizuojama. Pats paprasčiausias funkcijos minimizavimo būdas yra gradientinis nusileidimas. Toliau pateikiami MDS algoritmo žingsniai:

1. Skaičiuojami atstumai tarp turimo duomenų rinkinio objektų n-matėje erdvėje.
2. Atsitiktinai parenkamas rinkinys $\mathbf{Y}$ ($Y_i \in R^d$, $i = \overline{1, m}$).
3. Skaičiuojama kvadratinė paklaida pagal (2.5) formulę.

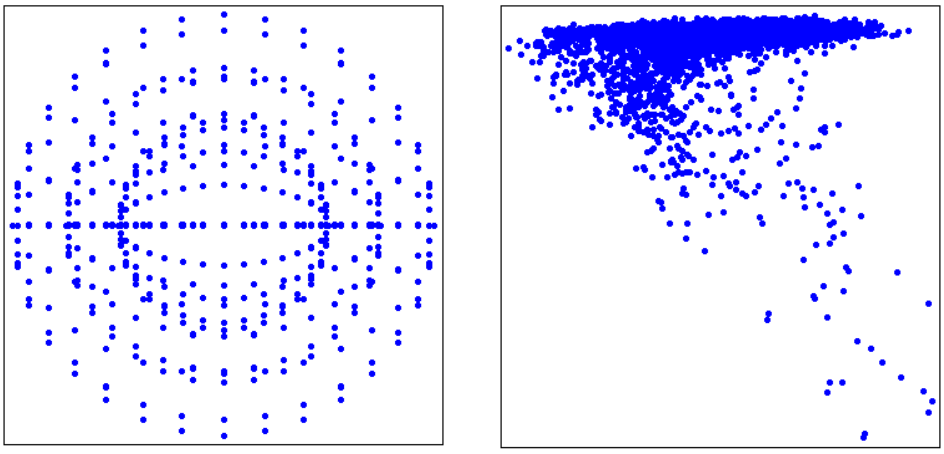
4. d-matės erdvės taškų Y_i komponentės atnaujinamos pagal formulę:

$$y_{i\tilde{j}}(u+1) = y_{i\tilde{j}}(u) - \eta\Delta(u), \text{ kur } \Delta = \frac{\partial E_{\text{rawStress}}}{\partial y_{i\tilde{j}}}.$$

Čia u – iteracijos numeris, η – optimizavimo žingsnio ilgį reguliuojantis parametras, $i = \overline{1,m}$, $\tilde{j} = \overline{1,d}$. Vienos iteracijos metu perskaičiuojamos visų m taškų $Y_i \in R^d$ komponentės.

5. Kartojama nuo 3 žingsnio, kol paklaidos reikšmė taps mažesnė už pasirinktą slenkstį arba bus viršytas nustatytas iteracijų skaičius.

Dviejų duomenų rinkinių (Sfera – atsitiktinai generuoti sferos taškai, $m = 726$, $n = 3$; Vystančių medžių duomenų rinkinys (angl. *Wilt data set*) (Johnson ir kt., 2013), $m = 4339$, $n = 5$. Detalesnis duomenų rinkinių aprašymas pateikiamas 4.1. poskyryje) projekcijos į dvimatę erdvę, gautos MDS metodu, pateiktos 2.2 paveiksle.



(a) Sfera **(b)** Vystantys medžiai

2.2 pav. Daugiamačių skalių metodu vizualizuoti duomenų rinkiniai

Galimos ir kitos MDS paklaidos funkcijos (Kruskal, 1964; Sammon, 1969; Borg ir Groenen, 2005; France ir Carroll, 2011):

- *Stress*-1 funkcija:

$$E_{\text{Stress-1}} = \sqrt{\frac{\sum_{i<j}(d(Y_i,Y_j) - d(X_i,X_j))^2}{\sum_{i<j}(d(Y_i,Y_j))^2}}. \tag{2.6}$$

- Sammono projekcija:

$$E_{\text{Sammon}} = \frac{1}{\sum_{i<j} d(X_i,X_j)} \sum_{i<j} \frac{(d(X_i,X_j) - d(Y_i,Y_j))^2}{d(X_i,X_j)}. \tag{2.7}$$

2.2. Klasterizavimo metodai

Vizualizavimo metodai turimus daugiamačius duomenis pateikia žmogui suvokiamoje erdvėje (dvimatėje arba trimatėje) perteikiant taškų išsidėstymą, t. y. išlaikant jų panašumus ir skirtingumus. Tačiau atsiranda poreikis vizualiai įvertinti duomenų rinkinio struktūrą ir savybes: susidariusias grupes, žymiai išsiskiriančius objektus, objektų panašumus/skirtingumus, ir pan. Patogu vizualizavimo metodus apjungti su kita duomenų tyrybos metodų grupe – klasterizavimu. Pirma atlikus duomenų rinkinio klasterizavimą, o po to vizualizavimą lengviau stebimos esančių objektų grupės. Klasterizavimas (angl. *clustering*) – tai toks duomenų rinkinį sudarančių objektų suskirstymas į skirtingas grupes, dar vadinamus klasterius (angl. *clusters*), kad grupės objektai būtų panašūs tarpusavyje, o objektai iš skirtingų grupių būtų nepanašūs (Dzemyda ir kt., 2013).

Klasterizavimo metodai yra taikomi daugelyje sričių: biomedicininiuose tyrimuose, atpažinimo procesuose, erdvinių duomenų analizėje, rinkos arba klientų skirstyme, dokumentų grupavime ir kt. Klasterizavimo metodai gali būti naudojami dvejopai: kaip atskiras duomenų tyrybos metodas arba kaip sudėtinė dalis kituose duomenų tyrybos metoduose (Han ir kt., 2011).

Pagrindiniai klasterizavimo bruožai (Dunham, 2002):

- Klasterių skaičius daugiamačiuose duomenyse nėra žinomas.
- Nėra jokių pradinių duomenų apie klasterius.
- Klasterių savybės gali kisti.

Duomenis suskirstyti į klasterius padeda įvairūs klasterizavimo metodai ir jų modifikacijos. Klasterizavimo metodų įvairovė yra labai didelė, todėl juos siūloma suskirstyti į grupes. Tačiau dalis sukurtų metodų gali priklausyti net kelioms grupėms. Skirtingi autoriai pateikia skirtingus klasterizavimo metodų grupavimus (Dunham, 2002; Gaur ir Gaur, 2013; Han ir kt., 2011). Vienas iš galimų klasterizavimo metodų grupavimų yra toks:

1. Dalijimo metodai (angl. *partitioning methods*) analizuojamą duomenų rinkinį padalina į pasirinktą klasterių skaičių. Dalinimo metu būdingas pakartotinis objektų perkėlimas iš vieno klasterio į kitą. Atlikus klasterizavimą patikrinama, ar tenkinamos dvi sąlygos:

   1) kiekvienas klasteris turi turėti bent vieną objektą;

   2) kiekvienas objektas turi priklausyti tik vienam klasteriui.

   Gerai suformuotame klasteryje objektai yra susiję vienas su kitu ir nelabai nutolę vienas nuo kito. Šiai grupei priklauso šie klasterizavimo metodai: k-vidurkių (angl. *k-means*) (MacQueen, 1967; Vesanto,

2001; Kanungo ir kt., 2002; Jain, 2010), k-medoidų (angl. *k-medoids*) (Kaufman ir Rousseeuw, 1990; Park ir Jun, 2009), CLARANS (angl. *Clustering Large Applications based upon RANdomized Search*) (Kaufman ir Rousseeuw, 1990; Ng ir Han, 2002; Liu ir Liu, 2006) ir jų modifikacijos (Gaur ir Gaur, 2013).

2. Hierarchiniai metodai (angl. *hierarchical methods*) formuoja duomenų rinkinio objektų hierarchiją. Hierarchiniams metodams būdinga savybė, kad jei duomenų klasteris išskirtas į du klasterius arba du klasteriai sujungti į vieną, tai negalima grįžti nei žingsnio atgal. Hierarchija gali būti formuojama dvejopai:
   1) Sujungimo principu (angl. *agglomerative*). Pradžioje kiekvienas objektas priklauso skirtingiems klasteriams. Vėliau objektai arba klasteriai, kurie yra panašūs, apjungiami tarpusavyje, kol visi klasteriai sujungiami į vieną didelį klasterį. Šiai grupei priskiriami klasterizavimo metodai: ROCK (angl. *RObust Clustering using linKs*) (Guha ir kt., 1999; Patidar ir kt., 2011), Chameleon (Karypis ir kt., 1999; Gaur ir Gaur, 2013).
   2) Išskaidymo principu (angl. *divisive*). Pradžioje visi objektai būna viename klasteryje. Vėliau klasteris skaidomas į mažesnius klasterius, atskiriant mažiau panašius objektus. Priskiriamas klasterizavimo metodas BIRCH (angl. *Balanced Iterative Reducing and Clustering using Hierarchies*) (Zhang ir kt., 1996; Horng ir kt., 2011).

3. Tankiu pagrįsti metodai (angl. *density-based methods*) – klasteris formuojamas pagal nurodytą objektų tankį. Šio metodo pagrindinė idėja yra baigti „auginti“ klasterį (nepriskirti jam daugiau objektų), kai pasiekiamas norimas tankis. Klasteris gali būti formuojamas dvejopai:
   1) Pagal atitinkamą kaimyninių objektų tankį (pavyzdžiui, DBSCAN (angl. *Density-Based Spatil Clustering of Applications with Noise*) (Ester ir kt., 1996; Liu ir kt., 2012)).
   2) Pagal tam tikrą tankio funkciją (pavyzdžiui, DENCLUE (angl. *DENsity-based CLUstEring*) (Hinneburg ir kt., 1998; Han ir kt., 2011; Gaur ir Gaur, 2013)).

4. Tinklu pagrįsti metodai (angl. *grid-based methods*) – turimo duomenų rinkinio n-matę erdvę sudalina į baigtinio skaičiaus vienodo dydžio ląsteles, kurios sudaro tinklo struktūrą. Tuomet turimi objektai

išdėliojami ant tinklo. Pagrindinis privalumas, kad metodo greitis priklauso nuo pasirinkto tinklo tankumo, o ne nuo duomenų rinkinio dydžio. Šios klasterizavimo grupės tipinis metodas yra STING (angl. *STatistical INformation Grid*) (Wu ir kt., 2012). WaveCluster (Yıldırım ir Özdoğan, 2011) ir CLIQUE (angl. *CLustering In QUEst*) (Agrawal ir kt., 1998; Zhang ir Liu, 2011) metodai yra priskiriami dviem klasterizavimo grupėms: tinklu pagrįstiems metodams ir tankiu pagrįstiems metodams.

5. Modeliu pagrįsti metodai (angl. *model-based methods*) – iškelia hipotezę apie modelį klasteriui ir ieško geriausiai tinkančių objektų pateiktam modeliui. Šiai grupei galima priskirti šiuos klasterizavimo metodus: EM algoritmas (Gupta ir Chen, 2011), neuroniniai tinklai (SOM (angl. *self-organizing maps*) (Kohonen, 2001)).

   Klasterizavimo metodai, kurie susideda iš kelių klasterizavimo metodų arba juose yra integruotos kitų klasterizavimo metodų idėjos, dažniausiai taip pat priskiriami šiai grupei.

Žinios apie klasterius (objekto priskyrimas konkrečiam klasteriui; klasterio centras) yra naudojamos šioje disertacijoje pasiūlytame metode (detalus metodo aprašymas yra pateiktas 3. skyriuje). Duomenis į klasterius galime suskirstyti bet kuriuo anksčiau paminėtu klasterizavimo metodu. Klasterių centrus, kuriuos žymėsime $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $\mu_j \in R^n$, $j = \overline{1,k}$, k – pasirinktas klasterių skaičius, galime nesunkiai apskaičiuoti, jei turime duomenis suskirstytus į klasterius. Tačiau vienas iš populiariausių ir paprasčiausių klasterizavimo metodų, kurio veikimo metu yra apskaičiuojami klasterių centrai, yra k-vidurkių metodas. Šio metodo populiarumą lemia tai, kad jis yra lengvai įgyvendinamas, paprastas ir veiksmingas (Jain, 2010; Kanungo ir kt., 2002). Dėl šių priežasčių disertacijoje taip pat yra naudojamas k-vidurkių klasterizavimo metodas.

Toliau trumpai pristatoma k-vidurkių klasterizavimo metodo idėja (MacQueen, 1967; Vesanto, 2001; Jain, 2010).

Į pasirinktą skaičių k klasterių $K_1, K_2, \ldots, K_k$ suskirstomas turimas daugiamačių duomenų rinkinys $\mathbf{X} = \{X_1, X_2, \ldots, X_m\} = \{x_{ij},\ i = \overline{1,m}, j = \overline{1,n}\}$ ir apskaičiuojami klasterių centrai $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $\mu_j \in R^n$, $j = \overline{1,k}$:

$$\mu_j = \frac{1}{m_{K_j}} \sum_{X_i \in K_j} X_i, \tag{2.8}$$

čia K_j – j-asis klasteris, $j = \overline{1,k}$, $X_i \in K_j$, m_{K_j} – objektų klasteryje K_j skaičius, $\sum_{j=1}^{k} m_{K_j} = m$.

k-vidurkių klasterizavimo metodas duomenų rinkinį į klasterius suskirsto minimizuodamas kvadratinę paklaidą tarp klasterio centro μ_j ir tam klasteriui priklausančių objektų X_i. Kvadratinė paklaida yra atstumų (dažniausiai skaičiuojamas Euklidinis atstumas, bet gali būti skaičiuojami ir kiti atstumai, pateikti 2.1.1. poskyryje) tarp klasterių centrų μ_j ir tiems klasteriams priklausančių objektų X_i kvadratų suma:

$$E_{K_j} = \sum_{X_i \in K_j} \|X_i - \mu_j\|^2, \tag{2.9}$$

čia K_j – j-asis klasteris, $j = \overline{1,k}$, $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$ – klasterio K_j centras, $\mu_j \in R^n$.

Klasterizavimo metodo tikslas – minimizuoti visų klasterių kvadratinių paklaidų sumą:

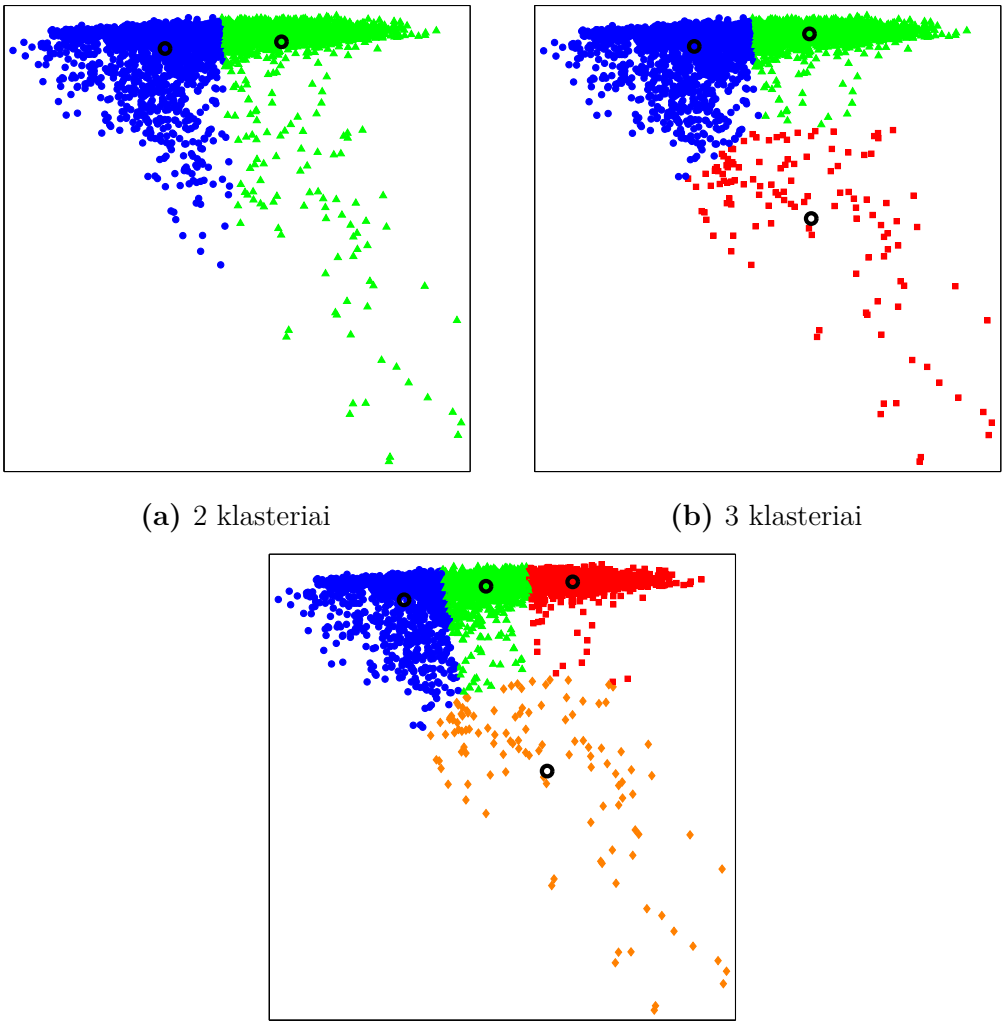
$$E_K = \sum_{j=1}^{k} \sum_{X_i \in K_j} \|X_i - \mu_j\|^2. \tag{2.10}$$

Pagrindiniai k-vidurkių algoritmo žingsniai:

1. Inicializuojami klasterių centrai μ_j.
2. Skaičiuojami atstumai nuo kiekvieno objekto X_i iki kiekvieno klasterio centro μ_j. Taškas X_i priskiriamas tam klasteriui μ_j, iki kurio atstumas yra mažiausias.
3. Perskaičiuojamas kiekvieno klasterio centras pagal (2.8) formulę.
4. Skaičiuojama kvadratinė paklaida pagal (2.10) formulę.
5. 2–4 žingsniai kartojami, kol pasiekiama norima paklaida arba objektai neperskirstomi kitiems klasteriams.

Į skirtingą klasterių skaičių k-vidurkių metodu klasterizuotas, o po to daugiamačių skalių metodu vizualizuotas Vystančių medžių duomenų rinkinys (duomenų rinkinio aprašymas pateiktas 4.1. poskyryje) pateiktas 2.3 paveiksle. Skirtingų klasterių objektus atitinkantys taškai pažymėti ●, ▲, ■ ir ♦. Klasterių centrai pažymėti ❍.

k-vidurkių klasterizavimo metodas turi ir keletą trūkumų: sunku nustatyti tinkamą klasterių skaičių k; randa kvadratinės paklaidos lokalų, o ne globalų minimumą; veikia tik su metriniais duomenimis.



(a) 2 klasteriai **(b)** 3 klasteriai

(c) 4 klasteriai

2.3 pav. *k*-vidurkių metodu klasterizuotas, o po to MDS vizualizuotas Vystančių medžių duomenų rinkinys

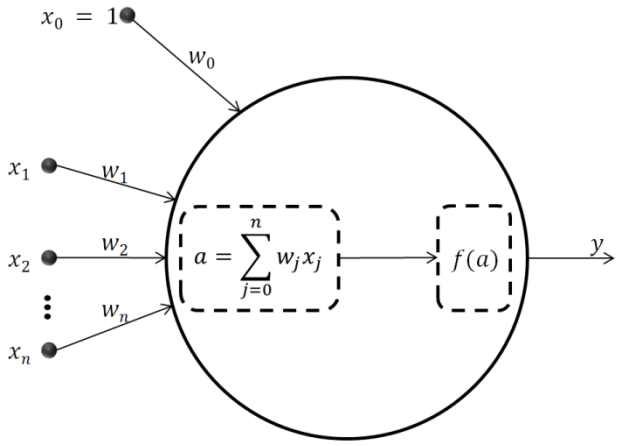
2.3. DNT daugiamačiams duomenims vizualizuoti

Trečioji duomenų tyrybos metodų grupė, kurios metodai naudojami disertacijoje, yra dirbtiniai neuroniniai tinklai. Dirbtiniai neuroniniai tinklai (DNT, angl. *artificial neural networks*) yra sukurti pagal biologinių neuroninių sistemų modelį (Verikas ir Gelžinis, 2008; Raudys, 2008; Haykin, 2009; Kantardzic, 2011; Dzemyda ir kt., 2013). Pagrindinis DNT tikslas yra išsiaiškinti ir pritaikyti biologinių neuronų sąveikos mechanizmus efektyvesnėms informacijos apdorojimo sistemoms kurti. Dirbtiniai neuroniniai tinklai yra naudojami diagnostikoje, modeliavime, vaizdų ir signalų atpažinime, kompiuterinės grafikos valdyme, intelektinėje paieškoje ir kitose sferose. Su jais atliekamas duomenų klasifikavimas, klasterizavimas, prognozavimas, optimizavimas, funkcijų aproksimavimas, matmenų skaičiaus

mažinimas ir vizualizavimas. DNT dažnai padeda atskleisti daugiamačių duomenų savybes, kurių negalima pastebėti klasikiniais daugiamačių duomenų vizualizavimo metodais (Dzemyda ir kt., 2013).

2.3.1. Dirbtinio neurono modelis

Dirbtinio neurono apibrėžimas buvo pasiūlytas dviejų amerikiečių mokslininkų (McCulloch ir Pitts, 1943). Remiantis biologinio neurono sandara buvo sukurtas dirbtinio neurono modelis, kuris pateiktas 2.4 paveiksle.



2.4 pav. Dirbtinio neurono modelis

Dirbtinio neurono modelį galima suskirstyti į tris pagrindines dalis (Haykin, 2009):

1. **Įėjimai**. Neuronas turi keletą įėjimų, kuriuos žymėsime x_l, $l = \overline{1,n}$. Kiekvienas įėjimas x_l turi savo perdavimo koeficientą (svorį) w_l, $l = \overline{1,n}$. Šalia įėjimų dar yra slenksčio reikšmė w_0 (angl. *bias*), kuri nurodo sustiprinti ar pasilpninti gaunamą signalą. Paprastai įėjimų ir svorių reikšmės yra realieji skaičiai.

2. **Sužadinimo signalas**. Skaičiuojama įėjimų ir svorių reikšmių sandaugų suma

$$a = w_1x_1 + w_2x_2 + \ldots + w_nx_n + w_0 = \sum_{l=1}^{n} w_lx_l + w_0. \tag{2.11}$$

Jei pridėtume nulinį įėjimą x_0, kuris yra visada pastovus, $x_0 = 1$, tai (2.11) formulę galėtume užrašyti:

$$a = \sum_{l=0}^{n} w_lx_l. \tag{2.12}$$

3. **Išėjimas**. Neurono išėjimą apibūdina aktyvavimo funkcija

$$y = f(a) = f\left(\sum_{l=0}^{n} w_l x_l\right). \tag{2.13}$$

Aktyvavimo funkcijų yra įvairių. Dažniausiai naudojamos aktyvavimo funkcijos (Kantardzic, 2011):

- Slenkstinė arba šuolinė

$$f(a) = \begin{cases} 1, \text{ jei } a \geq 0, \\ 0, \text{ jei } a < 0. \end{cases} \tag{2.14}$$

- Tiesinė

$$f(a) = a. \tag{2.15}$$

- Loginis sigmoidas

$$f(a) = \frac{1}{1 + \mathrm{e}^{-a}}. \tag{2.16}$$

- Tangento sigmoidas

$$f(a) = \frac{\mathrm{e}^{a} - \mathrm{e}^{-a}}{\mathrm{e}^{a} + \mathrm{e}^{-a}}. \tag{2.17}$$

Dirbtiniai neuronai, kurie yra sujungti vienas su kitu, vadinami dirbtiniu neuroniniu tinklu (DNT) arba tiesiog neuroniniu tinklu (NT). Kiekvienas neuroniniame tinkle esantis neuronas turi savo įėjimus ir išėjimą. Dažniausiai neuronai tinkle yra išdėstomi sluoksniais, nes taip paprasčiau juos modeliuoti. Pagal neuronų sujungimą vieną su kitu, DNT skirstomi į dvi dideles grupes (Haykin, 2009):

1. **Tiesioginio sklidimo DNT**. Šios grupės neuroniniuose tinkluose signalas iš įėjimų sklinda link išėjimo neuronų per visus paslėptus elementus. Šiai grupei priklausantys tinklai:
   - Vienasluoksnis perceptronas.
   - Daugiasluoksnis perceptronas.
   - Radialinių bazinių funkcijų neuroninis tinklas.
2. **Grįžtamojo ryšio arba rekurentiniai DNT**. Signalas sklinda ir atgalinėmis jungtimis iš vėlesniųjų į ankstesniuosius neuronus (Verikas ir Gelžinis, 2008). Šiai grupei priklausantys tinklai:
   - Konkurenciniai neuroniniai tinklai.
   - Saviorganizuojantys neuroniniai tinklai.

- Hopfieldo neuroniniai tinklai.
- Adaptyviojo rezonanso teorija paremti modeliai.

Sukonstruotą DNT būtina apmokyti, kad tinklas išspręstų jam skirtą užduotį. DNT mokymo proceso metu ieškoma slenksčio w_0 ir svorių w_j reikšmių, su kuriomis tinklas gautų tiksliausius rezultatus. Ieškomų parametrų reikšmės keičiamos atsižvelgiant į tinklo įėjimo ir išėjimo reikšmes, gautas ankstesniame mokymo žingsnyje. Procesas kartojamas, kol pasiekiamas norimas rezultatas (Dzemyda ir kt., 2013).

Skirtingos DNT architektūros apmokomos skirtingais jų algoritmais. Visus mokymo algoritmus galima suskirstyti į tris grupes:

1. Mokymo su mokytoju algoritmai (angl. *supervised learning*). Tinklo mokytojas – norimos tinklo atsako reikšmės $\mathbf{T} = \{T_1, T_2, \ldots, T_m\} = \{t_{ij}, i = \overline{1,m}, j = \overline{1,s}\}$. Tinklo mokymo metu ieškoma tokių svorių reikšmių, kad skirtumas tarp norimų tinklo atsako reikšmių t_j ir išėjimo reikšmių y_j būtų kiek galima mažesnis.

2. Mokymo be mokytojo algoritmai (angl. *unsupervised learning*). Svorių reikšmės keičiamos atsižvelgiant į koreliacijas ar panašumus tarp mokymo rinkinio įėjimų.

3. Skatinantis mokymas (angl. *reinforcement learning*).

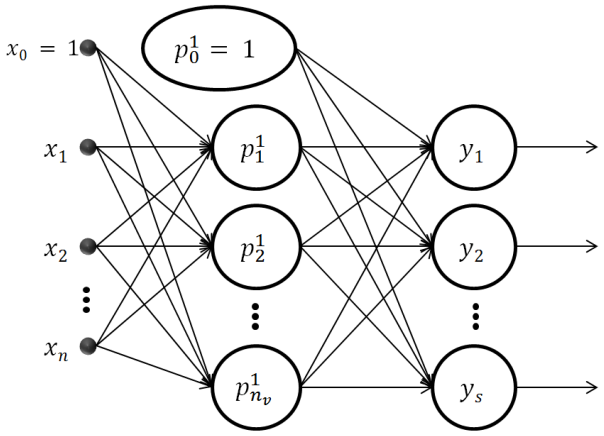
2.3.2. Daugiasluoksnio perceptrono naudojimas vizualizavimui

Kaip jau yra minėta DNT naudojami daugelyje sričių ir sprendžia labai įvairius uždavinius. Disertacijos tiksle yra užsibrėžta, kad bus ieškoma duomenų projekcija dvimatėje arba trimatėje erdvėje. Todėl toliau šiame skyriuje bus apžvelgiami tik DNT, pritaikyti daugiamačių duomenų projekcijos radimui ir vizualizavimui dvimatėje arba trimatėje erdvėje. Tokie DNT dažnai padeda atskleisti daugiamačių duomenų savybes, kurių negalima pastebėti klasikiniais daugiamačių duomenų vizualizavimo metodais. Klasikiniai daugiamačių duomenų vizualizavimo metodai turi trūkumą, jei norima plokštumoje atvaizduoti naujai atsiradusį tašką duomenų rinkinyje, tai tenka perskaičiuoti visų jau atvaizduotų taškų projekcijas. Šio trūkumo padeda išvengti DNT.

Pats paprasčiausias neuroninio tinklo tipas daugiamačių duomenų projekcijai nustatyti yra daugiasluoksnis perceptronas.

Dirbtinis neuroninis tinklas, kuriame neuronai išdėstyti keliais sluoksniais, vadinamas daugiasluoksniu tiesioginio sklidimo dirbtiniu neuroniniu tinklu (angl. *multilayer feedforward neural network*) arba tiesiog

daugiasluoksniu perceptronu (angl. *multilayer perceptron*, MLP). Kiekvienas daugiasluoksnis perceptronas susideda iš n įėjimų, V paslėptųjų neuronų sluoksnių, kuriuose yra po n_v neuronų ir s išėjimų. Paslėpto sluoksnio numerį pažymėkime v, tai $v = 0, 1, \ldots, (V+1)$, čia $v = 0$ žymi įėjimo sluoksnį, o $v = (V + 1)$ – išėjimų sluoksnį. Kiekviename paslėptame neuronų sluoksnyje v, $v = \overline{1, V}$, yra n_v neuronų. Daugiasluoksnio perceptrono schema su vienu paslėptu neuronų sluoksniu pateikta 2.5 paveiksle.



2.5 pav. Daugiasluoksnio perceptrono schema

2.5 paveiksle pateiktoje daugiasluoksnio perceptrono schemoje tinklo įėjimo rinkinys žymimas $X = (x_1, x_2, \ldots, x_n)$. Paslėptas neuronų sluoksnis žymimas $P = (p_1, p_2, \ldots, p_{n_v})$. Išėjimo rinkinys žymimas $Y = (y_1, y_2, \ldots, y_s)$. Vieno sluoksnio neuronai su kito sluoksnio neuronais (įskaitant ir įėjimo bei išėjimo sluoksnius) tarpusavyje sujungti svorių w_{jl} jungtimis (2.5 paveiksle pateiktoje schemoje jungtys žymimos rodyklėmis, kurios nurodo į kurią pusę sklinda signalas). Indeksai j ir l nurodo, kad signalas sklinda į j-ąjį neuroną v-ajame sluoksnyje iš l-ojo neurono $(v-1)$-ajame sluoksnyje.

Į tinklą paduotas įėjimo reikšmių rinkinys X sklinda palaipsniui per visus sluoksnius iki išėjimo sluoksnio. Pirmiausia apskaičiuojamos paslėpto sluoksnio neuronų p_j išėjimų reikšmės pagal formulę:

$$y_j = f(a_j) = f\left(\sum_{l=0}^{n} w_{jl} x_l\right), \tag{2.18}$$

čia w_{jl} yra jungties iš l-ojo įėjimo į j-ąjį neuroną svoris, įėjimo sluoksnyje $j = \overline{1, n}$, paslėptuose sluoksniuose $j = \overline{1, n_v}$.

Gautos neuronų reikšmės p_l yra išėjimų sluoksnio neuronų y_j įėjimų reikšmės. Indeksas j nurodo į kurį neuroną ateina signalas, o indeksas l – iš kurio neurono išeina signalas, t. y. kai skaičiuojami paslėpto sluoksnio išėjimai jie žymimi p_j, gavus paslėpto sluoksnio reikšmes, jų žymėjimas

pakeičiamas į p_l; $p_j = p_l$. Jeigu daugiasluoksniame neuroniniame tinkle yra daugiau paslėptų neuronų sluoksnių ($V \geq 2$, tada paslėptų neuronų sluoksnį žymėsime $P^v = (p_1^v, p_2^v, \ldots, p_{n_v}^v)$), tai gautos neuronų reikšmės p_l^v yra kito paslėpto sluoksnio neuronų p_j^{v+1} įėjimų reikšmės. Paslėptųjų sluoksnių arba išėjimo sluoksnio neuronų išėjimai apskaičiuojami pagal formulę:

$$y_j = f(a_j) = f\left(\sum_{l=0}^{n_v} w_{jl} p_l^v\right). \tag{2.19}$$

Kiekvienas neuronų sluoksnis gali turėti skirtingas aktyvavimo funkcijas arba net kiekvienas neuronas gali turėti skirtingas aktyvavimo funkcijas, bet tokiu atveju pasikeistų (2.18) ir (2.19) formulės (Dzemyda ir kt., 2013).

Gavus tinklo išėjimo reikšmes y_j apskaičiuojama paklaida $E(W)$. Paklaidos matas $E(W)$ yra apibrėžiamas kaip svorių matricos $W = \{w_{jl}, j = \overline{1,s}, l = \overline{0,n}\}$ funkcija. Dažniausiai naudojama paklaidos funkcija yra kvadratinių paklaidų suma (Haykin, 2009), kuri apskaičiuojama kiekvienam s-mačiam taškui išėjime:

$$E_i(W) = \frac{1}{2}\sum_{j=1}^{s}(y_{ij} - t_{ij})^2, \tag{2.20}$$

čia y_{ij} – j-tojo išėjimo reikšmė; t_{ij} – norima j-tojo išėjimo tinklo atsako reikšmė.

Bendra kvadratinė paklaidų suma visam duomenų rinkiniui:

$$E(W) = \sum_{i=1}^{m} E_i(W). \tag{2.21}$$

Daugiasluoksnio perceptrono mokymo tikslas yra minimizuoti paklaidos funkciją gradientiniu nusileidimo algoritmu. Algoritmas, leidžiantis minimizuoti paklaidos funkciją gradientiniu nusileidimo metodu daugiasluoksniam perceptronui, vadinamas „klaidos sklidimo atgal“ algoritmu (angl. *error back propagation learning algorithm*) (Rumelhart ir kt., 1986). Algoritmas taip vadinamas todėl, kad gautą paklaidą jis paskleidžia neuroniniu tinklu nuo išėjimo link įėjimo neuronų.

Visą algoritmo veikimą apibūdina du žingsniai:

1. Įėjimo reikšmių „sklidimas“ per visą neuroninį tinklą nuo įėjimų sluoksnio link išėjimų sluoksnio.

2. Gautos paklaidos „sklidimas“ atgal per visą neuroninį tinklą nuo išėjimų sluoksnio link įėjimų sluoksnio.

Algoritmo pirmojo žingsnio metu įėjimų reikšmės skleidžiamos palaipsniui per visus sluoksnius iki išėjimų sluoksnio. Gauta paklaida $E(W)$ rodo ar tinklas jau apmokytas. Jei paklaida nelygi nuliui arba nepasiekė norimo tikslumo, tai reikia keisti svorius w_{jl}. Svorių atnaujinimui reikia paskaičiuoti dalinę paklaidos išvestinę pagal svorius w_{jl}. Žymint įėjimo, svorinių sumų, išėjimo ir norimų reikšmių kintamuosius, įėjimo duomenų rinkinio indeksas i yra praleistas, kad nebūtų perkrauta žymėjimo sistema:

$$\frac{\partial E}{\partial w_{jl}} = \frac{\partial E}{\partial a_j}\frac{\partial a_j}{\partial w_{jl}}. \tag{2.22}$$

Iš (2.18) ir (2.19) formulių gauname:

$$\frac{\partial a_j}{\partial w_{jl}} = y_l. \tag{2.23}$$

Pažymėkime:

$$\delta_j = \frac{\partial E}{\partial a_j}. \tag{2.24}$$

Jei (2.23) ir (2.24) išraiškas įstatysime į (2.22) formulę, tai gausime:

$$\frac{\partial E}{\partial w_{jl}} = \delta_j y_l, \tag{2.25}$$

čia j-asis neuronas priklauso v-ajam sluoksniui, l-asis neuronas priklauso $(v-1)$-ajam sluoksniui.

Išėjimų sluoksnyje paklaidos kitimą pagal įėjimo reikšmių p_l ir svorių w_{jl} sandaugų sumą a_j nusako formulė:

$$\delta_j = \frac{\partial E}{\partial a_j} = f'(a_j)(y_j - t_j), \tag{2.26}$$

čia j-asis neuronas priklauso išėjimų sluoksniui.

Paslėptuose sluoksniuose esančių neuronų paklaidos kitimas $\frac{\partial E}{\partial a_j}$ užrašomas formule:

$$\delta_j = \frac{\partial E}{\partial a_j} = f'(a_j^v)\sum_{l=0}^{n_{v+1}} w_{jl}^{v+1}\delta_l^{v+1}, \tag{2.27}$$

čia n_{v+1} žymi $(v+1)$-ajame sluoksnyje esančių neuronų skaičių; j-asis neuronas priklauso v-ajam sluoksniui, o l-asis neuronas – $(v+1)$-ajam sluoksniui.

Paklaidos $E(W)$ sklidimo atgal metu δ_j reikšmės apskaičiuojamos palaipsniui visiems neuroninio tinklo sluoksniams pradedant nuo išėjimų

sluoksnio (δ_j apskaičiuojama pagal (2.26) formulę) ir baigiant įėjimų sluoksniu (pagal (2.27) formulę).

Apskaičiavus visas δ_j reikšmes, atliekamas svorių w_{jl} atnaujinimas. Svoriai atnaujinami pagal formulę:

$$\Delta w_{jl} = -\eta \delta_j y_l, \tag{2.28}$$

čia η yra mokymo greitis.

Svorių atnaujinimas galimas dviem būdais: po kiekvieno objekto pateikimo tinklui arba po viso objektų rinkinio pateikimo tinklui.

Aprašytasis daugiasluoksnis perceptronas gali būti pritaikytas daugiamačių duomenų projekcijos radimui ir jos vizualizavimui dvimatėje arba trimatėje erdvėje. Pats paprasčiausias būdas yra apmokyti tinklą su mokytoju (Mao ir Jain, 1996). Toliau bus pristatytas šis vizualizavimo būdas.

Tinklui mokyti naudojamas daugiamačių taškų duomenų rinkinys $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1, m}$. Prieš pradedant mokyti neuroninį tinklą šis duomenų rinkinys apdorojamas daugiamačių skalių metodu arba bet kuriuo kitu projekcijos metodu. Gaunamos taškų X_i projekcijos $T_i = (t_{i1}, t_{i2}, \ldots, t_{is})$ į $\mathbb{R}^s$ erdvę. Paprastai $s = 2$, jei norime tinklą mokyti projektuoti duomenis į plokštumą arbą $s = 3$, jei į trimatę erdvę. Gautosios projekcijos T_i ir bus norimos tinklo atsako reikšmės. Tinklas mokomas įprastiniu „klaidos sklidimo atgal“ algoritmu. Kai tinklas yra apmokytas, naują tašką X_{m+1} pateikus tinklui, išėjime gauname jo projekciją.

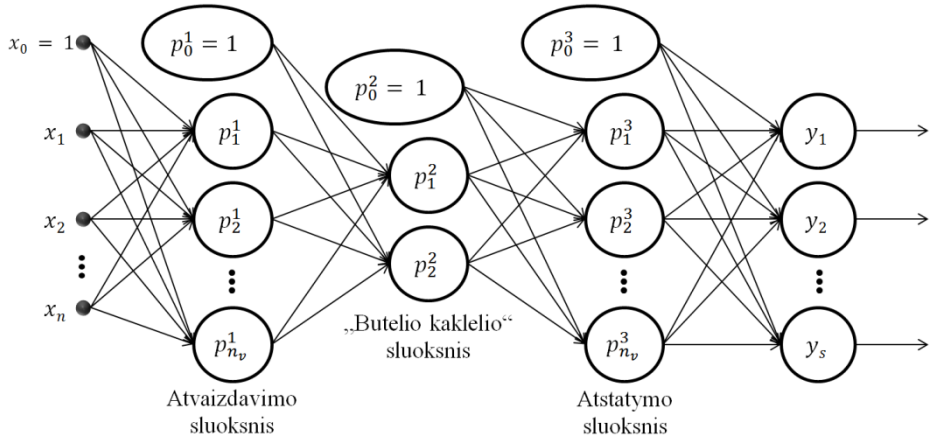
2.3.3. „Butelio kaklelio“ neuroninis tinklas

Kitas daugiasluoksnis perceptronas, mokomas su mokytoju ir skirtas daugiamačių duomenų projekcijos radimui bei vizualizavimui, vadinamas „butelio kaklelio“ neuroniniu tinklu (angl. *bottleneck neural network*) (Baldi ir Hornik, 1989; DeMers ir Cottrell, 1993; Araki ir kt., 2003). Šis tinklas priskiriamas autoasociatyviems neuroniniams tinklams (angl. *autoassociative neural network*). Šių tinklų ypatybė: išėjimuose tikimasi gauti tokias pačias reikšmes, kokios yra įėjimuose (t. y. norimos tinklo atsako reikšmės $T_i = X_i$, $i = \overline{1, m}$, $T_i \in \mathbb{R}^n$ ir $X_i \in \mathbb{R}^n$), o vidurinis paslėptas sluoksnis, sudarytas iš mažiau neuronų nei įėjimo ir išėjimo sluoksniai ($n_v^{Bk} < n$, čia n_v – neuronų skaičius v-ajame paslėptame sluoksnyje, Bk – „butelio kaklelio“ sluoksnis). Vidurinis paslėptas sluoksnis vadinamas „butelio kaklelio“ sluoksniu, nes jis tinkle suformuoja susiaurėjimą. Šiame sluoksnyje gaunama duomenų rinkinio projekcija norimoje erdvėje.

„Butelio kaklelio“ neuroninis tinklas sudarytas iš dviejų dalių, kurios yra simetrinės:

1. Atvaizdavimas – turimas duomenų rinkinys transformuojamas (projektuojamas) į mažesnio skaičiaus matmenų erdvę.
2. Atstatymas – rekonstruojamas (atstatomas) pradinis duomenų rinkinys iš gautų projekcijų (Thissen ir kt., 2001).

Šio tinklo schema pavaizduota 2.6 paveiksle.



2.6 pav. „Butelio kaklelio“ neuroninio tinklo schema

Iš 2.6 paveikslo matome, kad atvaizdavimo ir atstatymo dalys yra simetriškos, t. y. sudarytos iš vieno paslėpto neuronų sluoksnio, kuris susideda iš tiek pat neuronų, $n_1 = n_3$ (čia n_1 – P^1 paslėptame sluoksnyje esančių neuronų skaičius, n_3 – P^3 paslėptame sluoksnyje esančių neuronų skaičius), kurios sujungtos „butelio kaklelio“ sluoksniu. Bendru atveju atvaizdavimo ir atstatymo dalys savyje gali turėti po kelis paslėptus neuronų sluoksnius (Araki ir kt., 2003), tik būtina sąlyga, kad šios dvi dalys turi būti simetriškos. Pavyzdžiui, atvaizdavimo dalyje yra trys paslėpti neuronų sluoksniai P^1, P^2 ir P^3, kurie sudaryti iš $n_1 = 5$, $n_2 = 4$, $n_3 = 3$ neuronų, tai atstatymo dalyje taip pat turi būti trys paslėpti neuronų sluoksniai P^5, P^6 ir P^7 išdėstyti veidrodiniu variantu, t. y. sudaryti iš $n_5 = 3$, $n_6 = 4$, $n_7 = 5$ neuronų. Paslėptas sluoksnis P^4 yra „butelio kaklelio“ sluoksnis. „Butelio kaklelio“ sluoksnyje esančių neuronų skaičius priklauso nuo kokioje erdvėje ieškome projekcijos (dažniausiai $n_{Bk} = 2$, jei projekcijos ieškome $\mathbb{R}^2$ erdvėje, arba $n_{Bk} = 3$, jei projekcijos ieškome $\mathbb{R}^3$ erdvėje). „Butelio kaklelio“ neuroninis tinklas mokomas „klaidos sklidimo atgal“ algoritmu (žr. 2.3.2. poskyryje, 23 puslapyje).

2.3.4. SAMANN

Dar vienas dirbtinis neuroninis tinklas, skirtas daugiamačių duomenų projekcijai rasti, pavadintas SAMANN (Mao ir Jain, 1995; Dzemyda ir kt., 2013; Medvedev, 2007; Ivanikovas, 2010). Tai yra specialus tiesioginio sklidimo neuroninis tinklas, kuris realizuoja Sammono projekciją mokymo be mokytojo būdu. Tinklas apmokomas specifiniu „klaidos sklidimo atgal“ algoritmu.

Tinklas sudarytas iš dviejų identiškų daugiasluoksnių perceptronų. Vienu metu į tinklo įėjimus iš duomenų rinkinio $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}$ paduodami atsitiktinai parinkti du n-mačiai taškai $X_\mu = (x_{\mu 1}, x_{\mu 2}, \ldots, x_{\mu n})$ ir $X_\upsilon = (x_{\upsilon 1}, x_{\upsilon 2}, \ldots, x_{\upsilon n})$. Išėjimuose siekiama gauti jų projekcijas s-matėje erdvėje, t. y. taškus $Y_\mu = (y_{\mu 1}, y_{\mu 2}, \ldots, y_{\mu s})$ ir $Y_\upsilon = (y_{\upsilon 1}, y_{\upsilon 2}, \ldots, y_{\upsilon s})$, kur $s < n$. Bendru atveju SAMANN tinklas gali būti ir iš vieno daugiasluoksnio perceptrono, tik tuomet tinklas atmintyje turi saugoti daug daugiau informacijos (Medvedev ir Dzemyda, 2005; Ivanikovas ir kt., 2007).

Apmokant SAMANN tinklą „klaidos sklidimo atgal“ algoritmu, gautos projekcijos paklaida apskaičiuojama pagal formulę:

$$E_S = \frac{1}{\sum_{\mu=1}^{m-1} \sum_{\upsilon=\mu+1}^{m} d(X_\mu, X_\upsilon)} \sum_{\mu=1}^{m-1} \sum_{\upsilon=\mu+1}^{m} \frac{[d(X_\mu, X_\upsilon) - d(Y_\mu, Y_\upsilon)]^2}{d(X_\mu, X_\upsilon)}, \quad (2.29)$$

čia $d(X_\mu, X_\upsilon)$ yra atstumas tarp n-mačių taškų X_μ ir X_υ; $d(Y_\mu, Y_\upsilon)$ – atstumas tarp juos atitinkančių s-mačių taškų Y_μ ir Y_υ, $s < n$; m – duomenų rinkinį sudarančių taškų skaičius.

Tinklo idėja yra ta, kad pateikiami vienas paskui kitą du n-mačiai taškai X_μ ir X_υ, apskaičiuojami neuroninio tinklo atitinkami išėjimai Y_μ ir Y_υ, skaičiuojamas atstumas tarp taškų Y_μ ir Y_υ ir projekcijos paklaidos E_S, apibrėžtos (2.29) formule, reikšmė. Atsižvelgiant į ją, keičiami neuronų svoriai.

Sukurtojo tinklo trūkumas yra tai, kad ilgai trunka mokymas. Tačiau šis tinklas turi vieną išskirtinę savybę – galimybė iškart atvaizduoti naują n-matį tašką X_{m+1} neperskaičiuojant tinklo svorių. Taigi jeigu analizuojamame duomenų rinkinyje atsiranda naujas n-matis taškas X_{m+1} ir jis pateikiamas jau išmokytam SAMANN neuroniniam tinklui, tai tinklo išėjime gaunamos taško Y_{m+1}, kuris yra taško X_{m+1} projekcija, koordinatės. Žinoma, jeigu tų naujų taškų yra daug, po tam tikro laiko tinklą reikia mokyti iš naujo ir rasti naujas svorių reikšmes (Dzemyda ir kt., 2008).

2.3.5. Saviorganizuojantis neuroninis tinklas

Ankstesniuose skyreliuose aprašytieji DNT buvo skirti rasti duomenų rinkinio projekciją plokštumoje. Šiame poskyryje minimas DNT ne tik randa duomenų rinkinio projekciją plokštumoje, bet ir suskirsto turimus duomenis į klasterius. Toks yra saviorganizuojantis neuroninis tinklas (angl. *self-organizing maps*, SOM) (Kohonen, 2001; Vesanto ir Alhoniemi, 2000). SOM tinklo idėja – susikurti (organizuoti) save, naudojant turimą duomenų rinkinį. Mokymo metu yra išlaikoma duomenų topologija, t. y. taškai esantys arti įėjimo taškų erdvėje, yra atvaizduojami arti vieni kitų ir SOM tinkle. SOM tinklai gali būti naudojami siekiant vizualiai pateikti duomenų klasterius ir ieškant daugiamačių duomenų projekcijos į mažesnio skaičiaus matmenų erdvę, paprastai į plokštumą (Dzemyda ir kt., 2008). Saviorganizuojantis neuroninis tinklas mokomas mokymo be mokytojo būdu (Kohonen, 2001; Vesanto ir Alhoniemi, 2000; Kurasova, 2005; Dzemyda ir kt., 2013).

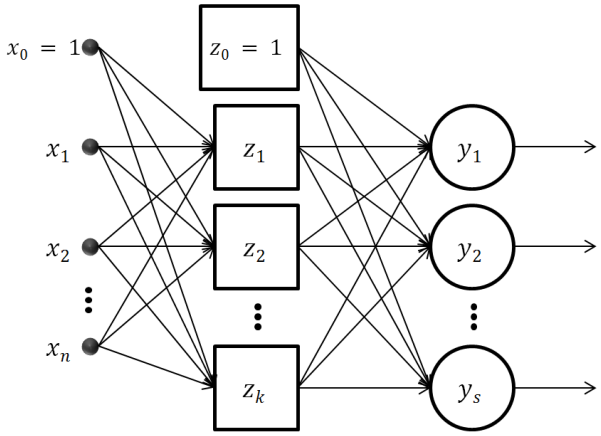
Apmokius tinklą ir jam padavus duomenų rinkinį, kiekvienam taškui yra randamas neuronas nugalėtojas. Pagal neuronus nugalėtojus yra sudaroma lentelė, kurios langeliuose surašyti analizuojamų taškų numeriai arba klasių pavadinimai. Tačiau gautoji lentelė nėra labai informatyvi, nes sunku įvertinti atstumus tarp taškų. Todėl buvo pasiūlyta vizualizuojant SOM tinklo rezultatus naudotis unifikuota atstumų matrica (angl. *unified distance matrix*, U-matrica). U-matricą sudaro atstumai tarp kaimyninių SOM neuronų. Remiantis U-matricos duomenimis vidutiniai atstumai tarp kaimyninių neuronų yra pateikiami kokios nors spalvos skalės atspalviais (pavyzdžiui, pilkos, žalios, mėlynos). Jei vidutiniai atstumai tarp kaimyninių neuronų yra maži, tuos neuronus atitinkantys tinklo langeliai spalvinami šviesia spalva; tamsi spalva reiškia didelius atstumus. Klasteriai yra nustatomi pagal šviesius atspalvius, o jų ribos – pagal tamsesnius (Kurasova, 2005; Dzemyda ir kt., 2008).

2.3.6. Vizualizavimas RBF tinklo paslėptame sluoksnyje

Kitas dirbtinis neuroninis tinklas, padedantis ne tik projektuoti duomenis plokštumoje, bet ir suskirstyti juos į klases, yra radialinių bazinių funkcijų neuroninis tinklas pritaikytas daugiamačių duomenų vizualizavimui. Toliau jis bus detaliai aprašytas.

Radialinių bazinių funkcijų neuroninis tinklas (angl. *radial basis function neural network*, RBF) (Broomhead ir Lowe, 1988; Chen ir kt., 1991; Buhmann, 2003) padeda išspręsti funkcijų aproksimavimo, laiko eilučių prognozavimo, klasifikavimo, sistemos kontroliavimo ir kitus uždavinius.

RBF tinklo modelis pateikiamas 2.7 paveiksle. Tinklas susideda iš n įėjimų, vieno paslėpto neuronų sluoksnio, kuris sudarytas iš k neuronų ir s išėjimų. Įėjimo duomenų rinkinys žymimas $X = (x_1, x_2, \ldots, x_n)$. Paslėptas neuronų sluoksnis žymimas $Z = (z_1, z_2, \ldots, z_k)$. Šiame sluoksnyje vietoj aktyvavimo funkcijų yra naudojamos radialinės bazinės funkcijos, todėl šis sluoksnis dar yra vadinamas radialinių bazinių funkcijų sluoksniu. Išėjimų sluoksnis žymimas $Y = (y_1, y_2, \ldots, y_s)$.



2.7 pav. Radialinių bazinių funkcijų neuroninio tinklo schema

Radialinių bazinių funkcijų sužadinimo lygis priklauso nuo atstumo tarp objekto $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1, m}$, ir radialinės bazinės funkcijos centro taško $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $j = \overline{1, k}$. Centras, tai taškas būtent nuo kurio yra skaičiuojamas atstumas. Bendra radialinių bazinių funkcijų išraiška užrašoma taip:

$$z_j = f(\|X - \mu_j\|), \tag{2.30}$$

čia $\|X - \mu_j\|$ – atstumas tarp objekto X_i ir centro taško μ_j, dažniausiai skaičiuojamas Euklidinis atstumas, bet gali būti skaičiuojamas ir bet kuris kitas atstumas, paminėtas 2.1.1. poskyryje; $f(\cdot)$ – tam tikra funkcija nuo anksčiau minėto atstumo.

Galimos radialinės bazinės funkcijos:

- Gausinė

$$z_j = \mathrm{e}\left(\frac{\|X - \mu_j\|^2}{2\sigma^2}\right), \tag{2.31}$$

  čia σ – pločio parametras.

- Multikvadratinė

$$z_j = \sqrt{\|X - \mu_j\|^2 + 1}. \tag{2.32}$$

- Multikvadratinė inversija

$$z_j = \frac{1}{\sqrt{\|X - \mu_j\|^2 + 1}}. \tag{2.33}$$

- Eksponentinė

$$z_j = \mathrm{e}\left(\frac{\|X - \mu_j\|}{2\sigma^2}\right). \tag{2.34}$$

- Splaininė (angl. *thin plate spline*)

$$z_j = \|X - \mu_j\|^2 \ln(\|X - \mu_j\|). \tag{2.35}$$

Dažniausiai RBF neuroniniuose tinkluose yra naudojama Gausinė radialinė bazinė funkcija, kuri yra apskaičiuojama pagal (2.31) formulę. Aprašant RBF neuroninio tinklo mokymą radialinės bazinės funkcijos yra Gausinės.

Radialinių bazinių funkcijų neuroniniai tinklai gali būti apmokomi dvejopai: visas tinklas iš karto arba skaidant mokymą į du etapus.

Visas tinklas iš karto yra apmokomas „klaidos sklidimo atgal“ algoritmu (šis algoritmas plačiau aprašytas 2.3.2. poskyryje, 23 puslapyje). Mokymo metu nustatomi pirmo paslėpto sluoksnio parametrai (radialinių bazinių funkcijų centro taškai μ_j ir pločio parametras σ) ir įvertinami išėjimų sluoksnio svoriai. Tačiau sprendimai gali būti gaunami neoptimalūs, kadangi pirmo paslėpto sluoksnio parametrų optimizavimo procedūra yra netiesinė (Verikas ir Gelžinis, 2008).

Kitas radialinių bazinių funkcijų neuroninių tinklų mokymo būdas yra tinklo apmokymas dalimis. Pirmojoje dalyje nustatomi radialinių bazinių funkcijų parametrai – bazinių funkcijų centro taškai μ_j ir pločio parametras σ. Nustačius parametrus radialinių bazinių funkcijų reikšmės tampa fiksuotos, todėl likusi tinklo dalis yra ekvivalentiška vienasluoksniam perceptronui (Verikas ir Gelžinis, 2008; Haykin, 2009). Antrosios tinklo dalies mokymas vyksta minimizuojant paklaidos funkciją gradientiniu nusileidimo algoritmu.

Mokant RBF neuroninį tinklą antruoju būdu daugiausia problemų kyla nustatant bazinių funkcijų parametrus. Centro taškai μ_j nurodo radialinių bazinių funkcijų vietą erdvėje. Juos reikia parinkti taip, kad apimtų visus duomenų rinkinio taškus. Pločio parametras σ apibūdina galimą taškų išsibarstymą aplink centro tašką μ_j. Idealiausiu atveju kiekvienai radialinei

bazinei funkcijai yra nustatomas atskiras pločio parametras. Tačiau paprasčiausias būdas yra imti visoms radialinėms bazinėms funkcijoms vienodą pločio parametro σ reikšmę (Lowe, 1989).

Galimi keli radialinių bazinių funkcijų centrų μ_j parinkimo būdai (Verikas ir Gelžinis, 2008):

1. Galimi bazinių funkcijų centrai μ_j atsitiktinai sutapatinami su įėjimo duomenų taškais. Tai nėra optimalus variantas centrams parinkti. Tačiau šis metodas dažnai naudojamas parenkant pradines centrų vertes, kai neuroninis tinklas apmokomas visas iš karto.

2. Daroma prielaida, kad visi duomenų taškai yra radialinių bazinių funkcijų centrai. Remiantis k artimiausių kaimynų (angl. *k-nearest neighbors*) metodu palaipsniui atsisakoma labiausiai nutolusių duomenų taškų taip, kad sistemos darbas kuo mažiau sutriktų.

3. Turimi daugiamačiai duomenys klasterizuojami k-vidurkių metodu (apie šį metodą plačiau pateikta 2.2. poskyryje) ir gauti k klasterių centrai laikomi radialinių bazinių funkcijų centrais μ_j.

Dauguma autorių pločio parametrą σ siūlo parinkti vienodą visoms radialinėms bazinėms funkcijoms. Vienas iš būdų yra pločio parametro σ parinkimas atsižvelgiant į klasterių centrų išsidėstymą (Haykin, 2009). Tuomet radialinė bazinė funkcija, kurios centras yra μ_j apibrėžiama taip:

$$z_j = \mathrm{e}\left(\frac{\|X-\mu_j\|^2}{2\sigma_{\mathrm{A}}^2}\right) = \mathrm{e}\left(-\frac{k}{d_{\max}^2}\|X-\mu_j\|^2\right), j = \overline{1,k}, \tag{2.36}$$

čia k – klasterių skaičius ir $d_{\max}$ – didžiausias atstumas tarp visų k klasterių centrų. Pločio parametras σ_{A} visoms Gausinėms radialinėms bazinėms funkcijoms yra fiksuotas:

$$\sigma_{\mathrm{A}} = \frac{d_{\max}}{\sqrt{2k}} = \alpha d_{\max}, \text{čia } \alpha = \frac{1}{\sqrt{2k}}. \tag{2.37}$$

Ši formulė užtikrina, kad individuali radialinė bazinė funkcija nėra per daug stati arba per daug lėkšta (plokščia); taip išvengta abiejų kraštutinių sąlygų.

Kitas būdas pločio parametro σ parinkimui yra vidutinis atstumas tarp klasterio centrų μ_j. Vidutinis atstumas nėra optimali pločio parametro σ reikšmė, todėl jį dar reikėtų padauginti iš konstantos, kuri parenkama eksperimentiškai. Prancūzų mokslininkų (Pierrefeu ir kt., 2006) pasiūlytas metodas:

1. Apskaičiuojamas vidutinis atstumas tarp centrų:

$$d_{\text{vid}} = \frac{\sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} \|\mu_i - \mu_j\|}{k(k-1)}, \tag{2.38}$$

čia $\|\mu_i - \mu_j\|$ – Euklidinis atstumas tarp centro taškų μ_i ir μ_j, k – klasterių skaičius.

2. Funkcijai

$$z_j(X) = \mathrm{e}\left(-\frac{\|X - \mu_j\|^2}{2\sigma_{\text{B}}^2}\right), \tag{2.39}$$

pločio parametras apskaičiuojamas taip:

$$\sigma_{\text{B}} = \alpha d_{\text{vid}}, \text{čia } \alpha = \frac{1}{\beta}. \tag{2.40}$$

Straipsnyje (Pierrefeu ir kt., 2006) reikšmė β eksperimentiškai parenkama iš intervalo [3,6; 0,05] prabėgant jį žingsniu 0,05, t. y. $\alpha \in$[0,28; 20]. Iš nurodyto intervalo imama ta parametro β reikšmė, su kuria apmokytas RBF tinklas gauna teisingus rezultatus (pavyzdžiui, teisingai suskirsto paveikslus į grupes).

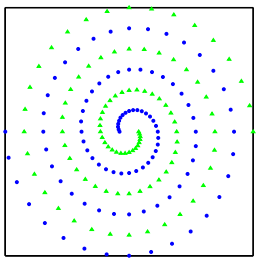
Aprašytasis radialinių bazinių funkcijų neuroninis tinklas gali būti pritaikytas daugiamačių duomenų klasifikavimui ir vizualizavimui paslėptame Z ir išėjimo Y sluoksniuose (Duch, 2004a,b). Vizualizavimas yra specifinis, nes paslėptame radialinių bazinių funkcijų sluoksnyje n-mačių taškų projekcijos dėliojamos ant hiperkubo (Saad ir Schultz, 1988) viršūnių. Hiperkubo dydis priklauso nuo duomenų rinkinyje esančių klasių skaičiaus ir nurodo radialinių bazinių funkcijų sluoksnyje esančių neuronų skaičių. Pavyzdžiui, jei turimą duomenų rinkinį sudaro keturios klasės, tai paslėptame Z sluoksnyje bus 4 neuronai, ir duomenys bus vizualizuojami ant 4-mačio hiperkubo. Paslėptame Z sluoksnyje naudojamos Gausinės (2.31) funkcijos. Į tinklą paduotas n-matis taškas X_i dedamas šalia tos hiperkubo viršūnės, kuri yra artimiausia. Kaip taškai bus išdėstyti ant hiperkubo priklauso ir nuo parinkto pločio parametro σ. Jei pločio parametras σ parinktas labai mažas, tai visi taškai bus sudėti šalia $(0, 0, \ldots, 0)$ viršūnės. Kitas kraštutinumas, kai parenkamas labai didelis pločio parametras σ, tada visi taškai dedami šalia $(1, 1, \ldots, 1)$ viršūnės. Tik tinkamai parinkus pločio parametrą σ taškai išdėstomi keliose viršūnėse ir tuo pačiu atskiriamos duomenų klasės (Duch, 2004b). Pločio parametras σ parenkamas atsitiktinai ir tik pagal gautą taškų išsidėstymą hiperkubo viršūnėse įvertinamas jo tinkamumas. Duomenų rinkinio taškų projekcijos taip pat gaunamos ir išėjimo sluoksnyje. Čia duomenys projektuojami į dvimatę erdvę.

2.4. Hibridiniai neuroniniai tinklai

Disertacijoje pasiūlytas metodas daugiamačiams duomenims tirti vizualiai, susideda iš radialinių bazinių funkcijų neuroninio tinklo ir daugiasluoksnio perceptrono. Todėl šiame poskyryje bus apžvelgiami įvairūs sukurti hibridiniai neuroniniai tinklai, kurie yra konstruojami sujungiant radialinių bazinių funkcijų tinklus su daugiasluoksnio perceptrono tinklais. Tačiau kiekvienas tinklas buvo konstruojamas specifiniam uždaviniui spręsti.

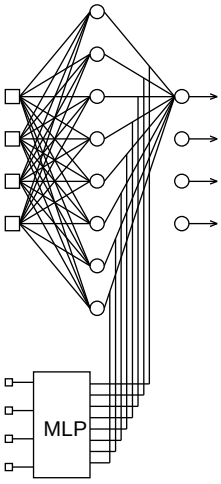
2.4.1. Hibridinis RBF-MLP neuroninis tinklas

Hibridinio radialinių bazinių funkcijų – daugiasluoksnio perceptrono (angl. *radial-basis function-multilayer perceptron*, RBF-MLP) idėją pasiūlė Tailando ir Didžiosios Britanijos mokslininkai (Chaiyaratana ir Zalzala, 1998; Zalzala ir Chaiyaratana, 2000). Šis tinklas skirtas sudėtingiems klasifikavimo uždaviniams spręsti. Vienas iš sudėtingo klasifikavimo pavyzdžių, tai Alexis P. Wieland iš Mitre korporacijos pasiūlytas dviejų spiralių uždavinys (Lang ir Witbrock, 1988; Fahlman ir Lebiere, 1990). Dviejų spiralių uždavinys – tai sudėtingo klasifikavimo uždavinys, kurio metu reikia atskirti dvi duomenų klases. Duomenys yra išdėstyti ant dviejų susipynusių spiralių plokštumoje. Vienos spiralės taškai priskiriami vienai klasei, o kitos – kitai klasei. 2.8 paveiksle pavaizduota, kaip yra išsidėstę duomenų taškai.



2.8 pav. Dviejų spiralių klasifikavimo uždavinys

Hibridinio RBF-MLP neuroninio tinklo architektūra pateikta 2.9 paveiksle. RBF-MLP neuroninis tinklas susideda iš radialinių bazinių funkcijų neuroninio tinklo (tinklas pateiktas 2.9 paveikslo viršuje) ir keleto daugiasluoksnių perceptronų (pateiktas tik vienas tinklas 2.9 paveikslo apačioje). Daugiasluoksnių perceptronų skaičius ir išėjimo sluoksnyje esančių neuronų skaičius priklauso nuo Gausinių bazinių funkcijų skaičiaus radialinių bazinių funkcijų tinkle. Pastebėsime, kad vienas MLP tinklas yra sujungiamas tik su vienu išėjimu. Kadangi 2.9 paveiksle pavaizduotas tik vienas MLP tinklas, tai likusiems išėjimams jungtys nepavaizduotos.



2.9 pav. Hibridinio RBF-MLP neuroninio tinklo schema

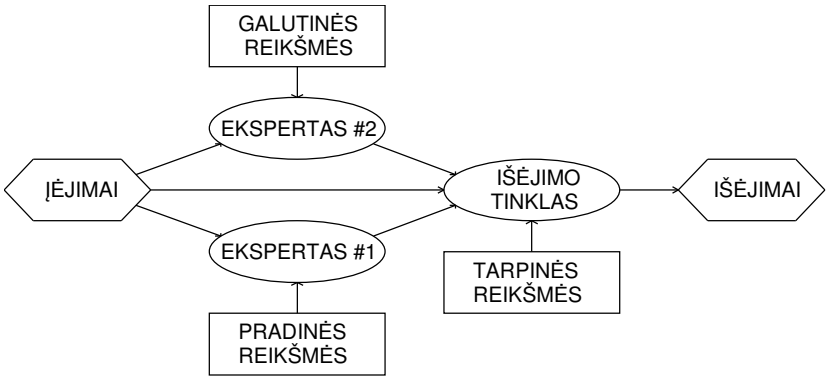
Iš 2.9 paveiksle pateiktos tinklo architektūros matome, kad hibridinis RBF-MLP neuroninis tinklas turi kelis įėjimus. Visiems tinklo įėjimams vienu metu yra paduodamos tos pačios reikšmės. Tinklas yra apmokomas genetiniu, mokymo su mokytoju ir mokymo be mokytojo algoritmais. Genetinis ir mokymo be mokytojo algoritmai yra naudojami Gausinių bazinių funkcijų centrams rasti. Daugiasluoksnis perceptronas yra apmokomas „klaidos sklidimo atgal“ algoritmu (žr. 2.3.2. poskyryje, 23 puslapyje). Kiekvienas daugiasluoksnis perceptronas sudarytas iš dviejų paslėptų neuronų sluoksnių. Neuronų perdavimo funkcija yra loginis sigmoidas. Daugiasluoksnio perceptrono išėjime yra tiesinė perdavimo funkcija. Išėjimo reikšmė yra radialinių bazinių funkcijų tinklo svorio reikšmė. Svorio parinkimas daugiasluoksnio perceptrono pagalba pagreitina radialinių bazinių funkcijų neuroninio tinklo apmokymą (Chaiyaratana ir Zalzala, 1998).

2.4.2. Neuroninio tinklo RBF/MLP modelis

Brazilų mokslininkai (Passos ir kt., 2006, 2007) pasiūlė kitokį radialinių bazinių funkcijų/daugiasluoksnio perceptrono (angl. *radial basis function/multilayer perceptron*, RBF/MLP) neuroninio tinklo modelį. Šis tinklas skirtas mikrobangų krosnelės įrenginiams modeliuoti. Pasiūlyto hibridinio RBF/MLP neuroninio tinklo modelio schema pateikta 2.10 paveiksle.

RBF/MLP modelis susideda iš trijų tiesioginio sklidimo neuroninių tinklų: du radialinių bazinių funkcijų neuroniniai tinklai (dar vadinami ekspertų tinklais) ir vienas daugiasluoksnis perceptronas (kitaip vadinamas išėjimo tinklas). Visi tinklai turi vieną paslėptą neuronų sluoksnį. Toks RBF ir MLP neuroninių tinklų išdėstymas pasirinktas dėl jų individualių charakte-

ristikų, kurias jie turi, kai yra atliekama funkcijos aproksimacija. Modelio netiesiškumas randamas atliekant lokalų tyrimą radialinių bazinių funkcijų neuroniniais tinklais. Modelinės struktūros apibendrinimas ir rezultatų išvedimas atliekamas globalaus tyrimo metu daugiasluoksniu perceptronu.



2.10 pav. RBF/MLP neuroninio tinklo modelis

RBF/MLP modelis suteikia galimybę turimą uždavinį išskaidyti į mažesnius ir paprastesnius uždavinius. Duomenys paimti iš hipotetinio įrenginio suskaidomi į tris dalis: „pradinės reikšmės“ (angl. *initial values*), „galutinės reikšmės“ (angl. *final values*) ir „tarpinės reikšmės“ (angl. *intermediate values*). RBF #1 ir #2 ekspertai yra apmokomi atitinkamai „pradinėmis reikšmėmis“ ir „galutinėmis reikšmėmis“; MLP išėjimo tinklas apmokomas visais duomenimis įskaitant ir „tarpines reikšmes“. RBF ir MLP tinklai yra apmokomi „klaidos sklidimo atgal“ algoritmu. Taip sukonstruotas modelis yra daug patikimesnis palyginus su pavieniais RBF ir MLP neuroniniais tinklais (Passos ir kt., 2006).

2.4.3. MLP-RBF tembro lygintuvas

Amerikiečių mokslininkai (Lu ir Evans, 1999; Lu, 2000) hibridinį daugiasluoksnio perceptrono-radialinių bazinių funkcijų neuroninio tinklo junginį (angl. *multilayer perceptron-radial basis function*, MLP-RBF) panaudojo tembro lygintuvo (angl. *equalizer*) kūrimui. MLP-RBF tinklo mokymas vyksta dviem etapais. Pirmiausia yra apmokomas daugiasluoksnis perceptronas „klaidos sklidimo atgal“ algoritmu. MLP neuroninio tinklo mokymo metu nuslopinamas triukšmas. Antrojo etapo metu yra apmokomas radialinių bazinių funkcijų neuroninis tinklas. RBF neuroninis tinklas turi tiek įėjimų, kiek yra MLP neuroninio tinklo išėjimų. RBF neuroninis tinklas atlieka tembro suvienodinimo funkciją. Pagal simbolio klaidų lygį MLP-RBF tembro lygintuvas lenkia atskirus MLP ir RBF tembro lygintuvus (Lu ir Evans, 1999).

2.4.4. MRHN tinklas

Ankstesniuose skyreliuose visi aptarti hibridiniai neuroniniai tinklai susideda iš dviejų dalių: radialinių bazinių funkcijų tinklo ir daugiasluoksnio perceptrono. Kiekvienos tinklo dalies apmokymas vyksta atskirai ir gauti rezultatai daro įtaką kitai tinklo daliai. Keli Taivano mokslininkai (Yeh ir kt., 2013) pasiūlė daugiasluoksnio perceptrono ir radialinių bazinių funkcijų neuroninių tinklų apjungimą į vientisą hibridinį neuroninį tinklą (angl. *MLP-RBF hybrid network*, MRHN), kuris atlieka erdvinę interpoliaciją. MRHN tinklas turi vieną paslėptą sluoksnį, kuris susideda iš loginio sigmoido (2.16) aktyvavimo funkcijų ir Gausinių (2.31) radialinių bazinių funkcijų. Paslėptame sluoksnyje neuronų skaičius yra lyginis, nes sigmoidinių perdavimo funkcijų ir Gausinių radialinių bazinių funkcijų turi būti po lygiai. Išėjimo sluoksnyje yra loginio sigmoido aktyvavimo funkcija. Tinklo mokymo metu mažinama kvadratinės paklaidos suma sukuriant mokymo su mokytoju taisykles visiems tinklo parametrams. Su pasirinktais testiniais duomenimis buvo palyginti RBF, MLP ir MRHN neuroniniai tinklai. Mažiausia paklaida buvo gauta naudojant MRHN (Yeh ir kt., 2013).

2.5. Antrojo skyriaus apibendrinimas ir išvados

Šiame skyriuje analitiškai apžvelgti daugiamačių duomenų vizualizavimo ir klasterizavimo metodai. Išanalizuoti dirbtiniai neuroniniai tinklai, kurie yra taikomi daugiamačiams duomenims vizualizuoti:

- Daugiasluoksnis perceptronas. Daugiamačiai duomenys vizualizuojami daugiamatėmis skalėmis, o po to tais rezultatais apmokomas daugiasluoksnis perceptronas. Rezultate – toks tinklas moka gauti naujų daugiamačių taškų, kurie nebuvo vizualizuoti naudojantis MDS, projekcijas į mažesnio matavimo erdvę.
- SAMANN tipo neuroninis tinklas. Tai yra specialus tiesioginio sklidimo neuroninis tinklas, kuris realizuoja Sammono projekciją mokymo be mokytojo būdu. Duomenų projekcijos mažesnio matavimo erdvėje gaunamos tinklo išėjime.
- „Butelio kaklelio“ tipo neuroninis tinklas. Jo idėja – kas paduodama į tinklo įėjimą, tai turi būti gaunama ir išėjime. Duomenų projekcija ieškoma viduriniame paslėptame neuronų sluoksnyje, kuris sudarytas iš dviejų arba trijų neuronų.
- Radialinių bazinių funkcijų neuroninis tinklas. Toks tinklas klasifikuoja duomenis ir paslėptame sluoksnyje ieško jų projekcijos hiperkube.

- Saviorganizuojantis neuroninis tinklas. Toks tinklas mokomas mokymo be mokytojo būdu. Šis tinklas ne tik randa duomenų rinkinio projekciją plokštumoje, bet ir suskirsto turimus duomenis į klasterius.

Paminėtųjų dirbtinių neuroninių tinklų, išskyrus RBF tinklą, veikimo strategijos yra orientuotos į tai, kad ieškant daugiamačių duomenų projekcijos plokštumoje siekiama išsaugoti atstumus tarp taškų. Priklausomai nuo optimizavimo kriterijaus, atstumai gali būti išlaikomi tarp labiau artimų arba labiau nutolusių taškų.

Atlikta analitinė hibridinių neuroninių tinklų apžvalga parodė, kad tokio tipo tinklai konstruojami labai įvairiose srityse ir specifiniams uždaviniams spręsti: sudėtingas (įvairiai susipynę klasteriai, pavyzdžiui, spirale) duomenų klasifikavimas, mikrobangų krosnelės įrenginių modeliavimas, ekvalaizerio sukūrimas, erdvinės interpoliacijos radimas. Hibridinių tinklų gaunami rezultatai yra tikslesnis palyginus su radialinių bazinių funkcijų neuroninių tinklų arba daugiasluoksnių perceptronų gaunamais rezultatais. Konkrečiam uždaviniui spręsti kuriamo hibridinio neuroninio tinklo struktūra pasirenkama pagal atskirų tinklų individualias charakteristikas.

Atlikta analitinė apžvalga parodė atskirų sprendimų privalumus ir specifiką:

1. Radialinių bazinių funkcijų neuroniniuose tinkluose realizuota galimybė įvertinti klasterius tiriamuose duomenyse, kai skirtingose radialinėse bazinėse funkcijose panaudojami atskirų klasterių centrai. Kiekviena radialinė bazinė funkcija yra „jautri“ konkrečiam vienam klasterio centrui.

2. „Butelio kaklelio“ tipo neuroniniame tinkle daugiamačių duomenų projekcija ieškoma paslėptame neuronų sluoksnyje.

3. Daugiasluoksnio perceptrono mokymui su mokytoju naudojamos žinios apie konkretų duomenų tašką.

Šiame skyriuje atlikta analizė parodė, kad ieškant duomenų projekcijos, kurioje tyrėjas galėtų pamatyti taškų tarpgrupinius panašumus/skirtingumus, reiktų bandyti apjungti skirtingų tipų neuroninių tinklų savybes, o ieškomoje projekcijoje nesistengti išlaikyti atstumų tarp taškų. Naujai konstruojamo tinklo mokymui labai svarbios žinios apie duomenų klasterius, kurios gali būti gaunamos duomenis suklasterizavus klasterizavimo metodais.

3. REGM tinklas daugiamačiams duomenims vizualizuoti

Šioje disertacijoje siekiama sukurti tinklą, kuris turėtų būti mokomas, jam pateikiant norimus vizualizuoti daugiamačius duomenis, o išėjime reikalaujant specifinės reakcijos, kuri susijusi su tam tikromis tų duomenų savybėmis, t. y. jų priklausymą klasteriams.

Šiame skyriuje pateiktas naujojo hibridinio radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginio (REGM) modelis, aptarti jo mokymo ypatumai ir pasiūlyti vizualizavimo kokybės kriterijai.

Pagrindiniai skyriaus rezultatai buvo pristatyti 5 konferencijose ir 3 straipsniuose, kurių sąrašai yra pateikti 1.6. poskyryje.

3.1. Prielaidos naujam vizualizavimo metodui kurti

Kaip jau yra minėta ankstesniame skyriuje, sukurtieji daugiamačių duomenų vizualizavimo metodai, ieškodami duomenų projekcijos plokštumoje, stengiasi išlaikyti atstumus tarp taškų. Šiame poskyryje bus pristatyta idėja, kaip transformuoti daugiamačius duomenis, kad gautoje projekcijoje labiau atsiskleistų tarpklasteriniai taškų panašumai.

Kyla idėja atlikti daugiamačių duomenų, kurie išreikšti n-matės erdvės duomenų taškais $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1,m}$, čia $X_i \in \mathbb{R}^n$, požymių skaičiaus n mažinimą, transformuojant $X_i \in \mathbb{R}^n$ į $Z_i \in \mathbb{R}^k$: $Z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$; čia $k < n$. n-mačio duomenų taško $X = (x_1, x_2, \ldots, x_n)$ dimensiškumas mažinamas naudojantis tam tikra radialine bazine funkcija, susieta su konkrečiu duomenų klasteriu. Gaunamas naujas k-matis duomenų taškas $Z = (z_1, z_2, \ldots, z_k)$, $k < n$, panaudojus šias formules:

1. Eksponentinė funkcija (Chen ir kt., 1993; Yaglom, 1986)

$$z_j(X) = \mathrm{e}(-\gamma \parallel X - \mu_j \parallel), j = \overline{1,k}, \gamma = \frac{1}{2\sigma^2}. \tag{3.1}$$

2. Gausinė funkcija (Haykin, 2009; Dzemyda ir kt., 2013)

$$z_j(X) = \mathrm{e}(-\gamma \parallel X - \mu_j \parallel^2), j = \overline{1,k}, \gamma = \frac{1}{2\sigma^2}. \tag{3.2}$$

Čia μ_j yra j-tosios funkcijos centro taškas, $\mu_j \in R^n$, $\|X - \mu_j\|$ – atstumas tarp X ir μ_j, σ – pločio parametras, nuo kurio priklauso funkcijos glotnu-

mas. Pastebėsime, kad $\|X - \mu_j\| > 0$ ir $\gamma > 0$. Eksponentinės funkcijos skirtumas nuo Gausinės yra tik tai, kad eksponentinėje funkcijoje naudojamas atstumas, o Gausinėje – atstumo kvadratas. Remiantis (3.1) ar (3.2) formule, iš duomenų rinkinio $\mathbf{X}$ gaunamas naujas duomenų rinkinys $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_m\} = \{z_{ij}, i = \overline{1,m}, j = \overline{1,k}\}$, t. y. atlikta netiesinė duomenų rinkinio $\mathbf{X}$ transformacija, kur atsižvelgiama į klasterius šio rinkinio duomenyse (Ringienė ir Dzemyda, 2013).

Paprastumo dėlei paimkime $n = 2$ ir $k = 2$. Paanalizuokime, kaip kinta taškų išsidėstymas plokštumoje atlikus duomenų transformaciją. Kaip pavyzdį imkime duomenų rinkinį $\mathbf{X}$, kuris sudarytas iš 6 duomenų taškų ($m = 6$). Duomenų rinkinys pateiktas 3.1 lentelėje.

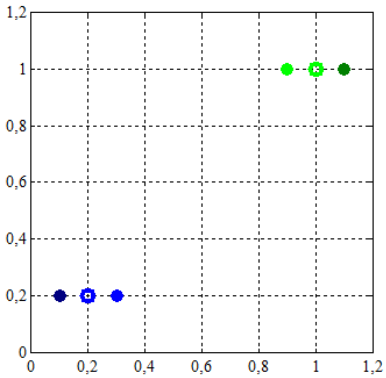
3.1 lentelė: Duomenų rinkinys ir po transformacijos gauti rezultatai

Nr.	Duomenų aibė $\mathbf{X}$		Eksponentinė transformacija		Gausinė transformacija	
	x_1	x_2	z_1	z_2	z_1	z_2
1	0,1	0,2	0,90	0,29	0,99	0,23
2	0,2	0,2	1	0,32	1	0,27
3	0,3	0,2	0,90	0,34	0,99	0,32
4	0,9	1	0,34	0,90	0,32	0,99
5	1	1	0,32	1	0,27	1
6	1,1	1	0,29	0,90	0,23	0,99

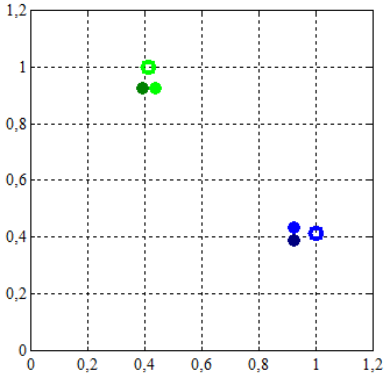
Tų duomenų išsidėstymas plokštumoje matomas 3.1a paveiksle. Iš 3.1a paveikslo matome, kad duomenų rinkinį sudaro 2 aiškūs klasteriai. Vienas klasteris pažymėtas mėlynai, o kitas – žaliai. Kiekvieno klasterio vidurinysis taškas yra klasterio centras μ_j, kuris pažymėtas mėlynu arba žaliu apskritimu.

Rezultatai gauti atlikus transformaciją eksponentine arba Gausine funkcija pateikti 3.1 lentelėje. Rezultatai vizualiai pateikti 3.1b ir 3.1c paveiksluose. Iš 3.1b paveikslo matome, kad eksponentinės funkcijos atveju, klasterių centrai atsiskiria nuo kitų klasterio taškų, o likę klasterio taškai suartėja. Taškai, turintys panašumo su gretimo klasterio taškais, atsiranda arčiau gretimo klasterio. Šviesesniais atspalviais (mėlynu ir žaliu) pažymėti taškai turi daugiau panašumo vienas su kitu, nei tamsesniais atspalviais pažymėti taškai.

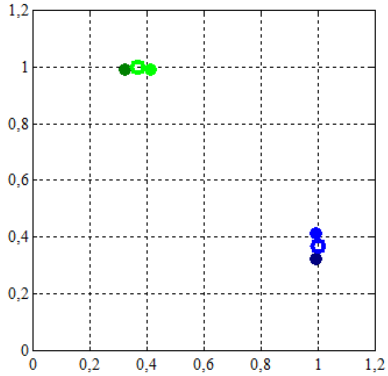
Gausinės funkcijos atveju (3.1c paveikslas) taip pat keičiasi klasterio taškų išsidėstymas klasterio centro aplinkoje. Stebime taškų, kurie nėra klasterio centrai, padėties pasikeitimą, panašų kaip ir eksponentinės funkcijos atveju, tačiau ne tokį ryškų.



(a) Duomenų rinkinys $\mathbf{X}$



(b) Transformacija atlikta eksponentine funkcija



(c) Transformacija atlikta Gausine funkcija

3.1 pav. Duomenų rinkinio vizualus pateikimas

3.2. REGM tinklo modelis

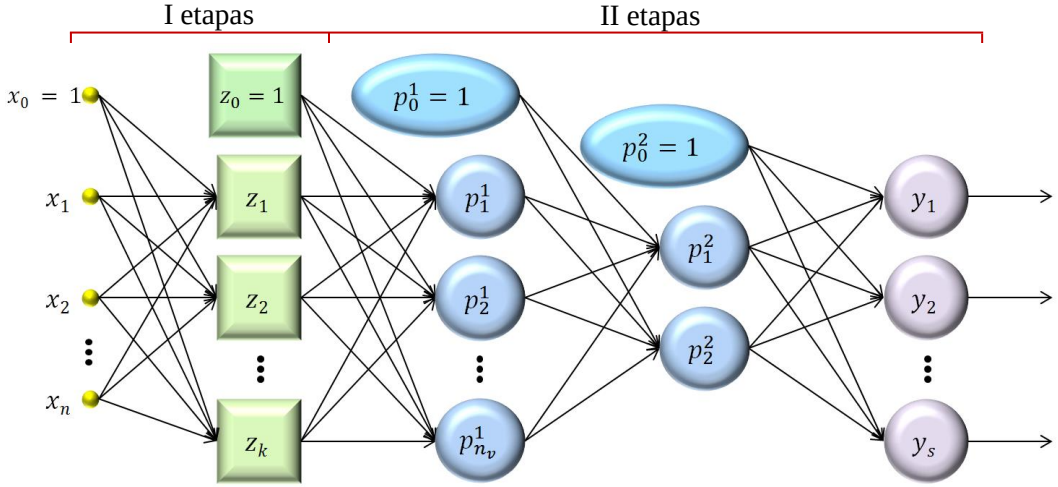
Šiame darbe pasiūlytas ir ištirtas hibridinis neuroninis tinklas, kuris savyje integruoja ir radialinių bazinių funkcijų neuroninio tinklo, ir daugiasluoksnio perceptrono, turinčio „butelio kaklelio“ neuroninio tinklo savybes, idėjas. Šis tinklas pavadintas REGM. Tinklo REGM pavadinimas sudarytas iš jį sudarančių neuroninių tinklų ir naudojamų transformacijos funkcijų angliškų pavadinimų pirmųjų raidžių (t. y. ***R****adial basis function neural network*, ***E****ksponential function*, ***G****aussian function*, ***M****ultilayer perceptron*).

Tinklas sudarytas iš dviejų dalių, kurios atitinka tokio tinklo atskirus mokymo etapus. Pirmoji dalis yra tam tikras n-matės erdvės $\mathbb{R}^n$ taškų transformavimas į norimo matmens erdvę $\mathbb{R}^k$, $k < n$. Antrojoje dalyje daugiasluoksnis perceptronas, kurio paskutinis paslėptas sluoksnis yra sudarytas iš nedidelio neuronų skaičiaus (2 arba 3). Kai išėjimo sluoksnyje pasirenkama daugiau neuronų, nei paskutiniame paslėptame sluoksnyje, tai tam tikra prasme primena „butelio kaklelio“ neuroninį tinklą. Tačiau tai tik labai tolima analogija, nes „butelio kaklelio“ neuroniniame tinkle vyrauja

simetrija ir mokymo metu išėjime stengiamasi gauti tai, kas paduodama į tinklą.

REGM tinklas naudojamas vizualiai daugiamačių duomenų analizei, kai atidėjimui plokštumoje arba trimatėje erdvėje taškai gaunami paskutinio paslėpto neuronų sluoksnio išėjimuose į tinklą padavus n-mačių analizuojamų duomenų rinkinį $\mathbf{X}$.

Šio tinklo ypatybė, yra ta, kad gautas vaizdas plokštumoje labiau atspindi bendrą duomenų struktūrą (klasteriai, klasterių tarpusavio artumas, taškų tarpklasterinis panašumas) nei daugiamačių taškų tarpusavio išsidėstymą. Pastebėsime, kad daugiamačių duomenų klasterizavimo rezultatai yra panaudojami ne tik apskaičiuojant radialinių bazinių funkcijų parametrus, bet ir pateikiant rezultatus plokštumoje. Skirtingų klasterių taškų plokštumoje dažymas skirtingomis spalvomis suteikia papildomų žinių tyrėjui, kas palengvina geriausio sprendimo priėmimą. Neuroninio tinklo REGM schema pateikta 3.2 paveiksle.



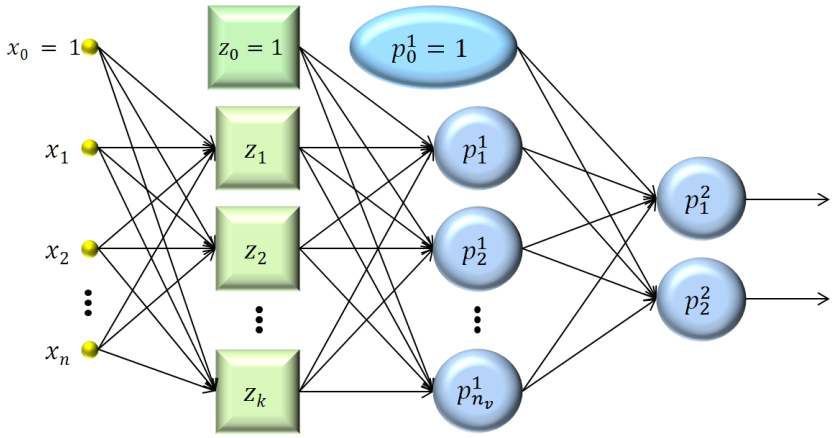
3.2 pav. Bendroji REGM tinklo schema

Tinklo REGM įėjimas žymimas $X = (x_1, x_2, \ldots, x_n)$. 3.2 paveiksle pateiktas REGM tinklas turi tris paslėptus neuronų sluoksnius. Pirmas paslėptas neuronų sluoksnis $Z = (z_1, z_2, \ldots, z_k)$ disertacijoje bus vadinamas radialinių bazinių funkcijų sluoksniu (3.2 paveiksle pažymėta žaliais kvadratais), o daugiasluoksnio perceptrono neuronų sluoksniai $P^1 = (p_1^1, p_2^1, \ldots, p_{n_v}^1)$ ir $P^2 = (p_1^2, p_2^2)$ – pirmuoju ir mažuoju (arba paskutiniu) paslėptais neuronų sluoksniais (3.2 paveiksle pažymėti mėlynais apskritimais). Radialinių bazinių funkcijų yra tiek, kiek spėjama daugiamačiuose duomenyse yra klasterių k. Neuronų skaičius n_v pirmame paslėptame sluoksnyje $P^1 = (p_1^1, p_2^1, \ldots, p_{n_v}^1)$ gali būti laisvai pasirenkamas. Mažajame sluoksnyje yra du ($n_v = 2$) arba trys ($n_v = 3$) neuronai. Neuronų skaičius priklauso

nuo erdvės, kurioje norime gauti daugiamačių duomenų projekciją ($\mathbb{R}^2$ arba $\mathbb{R}^3$). Tinklo REGM išėjimas žymimas $Y = (y_1, y_2, \ldots, y_s)$. Išėjimo sluoksnyje neuronų gali būti nuo vieno iki k (klasterių skaičiaus), $k \leq s$. Kai išėjimų sluoksnyje yra tiek neuronų, kaip ir klasterių skaičius, turime tam tikrą struktūrą panašią į „butelio kaklelio“ neuroninį tinklą, bet mokymas (plačiau apie tai 3.3. poskyryje) yra iš esmės kitoks. Tarkime, daugiamačiai duomenys turi penkis klasterius ($k = 5$). Tuomet radialinių bazinių funkcijų sluoksnyje bus penkios bazinės funkcijos, o išėjimo sluoksnyje pasirenkame vieną, du, tris, keturis arba penkis neuronus.

Paslėptuose sluoksniuose ir išėjimo sluoksnyje siūlomos naudoti loginio sigmoido (2.16) arba tiesinė (2.15) aktyvavimo funkcijos.

Faktiškai, kai REGM tinklas yra apmokytas, jis gali būti supaprastintas, atsisakant išėjimų sluoksnio, esančio 3.2 paveiksle. Toks naujas tinklas pateiktas 3.3 paveiksle.



3.3 pav. Po tinklo apmokymo daugiamačių duomenų vizualizavimui naudojamo REGM tinklo schema

Bendru atveju antrojoje REGM tinklo dalyje tarp pirmojo ir mažojo neuronų sluoksnių gali būti ir daugiau paslėptų neuronų sluoksnių. Tačiau disertacijoje nagrinėjamas tik 3.2 paveiksle parodytas neuroninis tinklas REGM.

3.3. REGM tinklo mokymas

Prieš pradedant mokyti REGM tinklą, pirmiausia turime kiekvienam tinklo įėjimo n-mačiui taškui $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1, m}$, nustatyti norimas tinklo atsako reikšmes $T_i = (t_{i1}, t_{i2}, \ldots, t_{is})$, $i = \overline{1, m}$ (čia s – tinklo išėjimo sluoksnyje esančių neuronų skaičius), kurios yra daugiamačių duomenų klasterių centrai $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $j = \overline{1, k}$. Toliau tai detalizuota.

Turimi daugiamačiai duomenys, kurie išreikšti n-matės erdvės duomenų taškais $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1, m}$, klasterizuojami į pasirinktą klasterių skaičių k k-vidurkių metodu (plačiau apie šį metodą 2.2. poskyryje). Taip nustatomi klasterių K_j centrai $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $j = \overline{1, k}$.

Norimai tinklo atsako reikšmei T_i priskiriamas tas klasterio centras μ_j, kuriam yra priskirtas įėjimo duomenų taškas X_i. Dėl šios priežasties bus vienodų norimų tinklo atsako reikšmių, esant skirtingiems įėjimo duomenų taškams. Pastebėsime, kad $n \neq s$, todėl yra siūlomas dvejopas norimų tinklo atsako reikšmių parinkimas:

1. k-vidurkių metodu gauti klasterių centrai $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$ daugiamačių skalių metodu (plačiau apie šį metodą 2.1.2. poskyryje) projektuojami iš $\mathbb{R}^n$ erdvės į mažesnio matavimo erdvę $\mathbb{R}^s$, $s < n$. Gauname klasterių centrų $\mu_j \in \mathbb{R}^n$ projekcijas $\mu_j^y \in \mathbb{R}^s$, $j = \overline{1, k}$. Norimos tinklo atsako reikšmės $T_i = \mu_j^y$, jei $X_i \in K_j$, $i = \overline{1, m}$. Pastebėsime, kad išėjimo sluoksnyje neuronų gali būti nuo 1 iki k (klasterių skaičiaus). Jei $s = k$, tai MDS metodu atliekant $\mu_j \in \mathbb{R}^n$ projekciją į $\mu_j^y \in \mathbb{R}^s$, $j = \overline{1, k}$, paskutinioji μ_j^y komponentė visada bus lygi 0.

2. Klasterių centrų $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$ transformacija iš $\mathbb{R}^n$ erdvės į $\mathbb{R}^k$ erdvę, $(k < n)$, atliekama radialine bazine funkcija (kaip atliekama daugiamačių duomenų ir klasterių centrų transformacija yra aprašyta 3.3.1. poskyryje). Klasterių centrai po transformacijos žymimi $\mu_j^z = (\mu_{j1}^z, \mu_{j2}^z, \ldots, \mu_{jk}^z)$. Jeigu $s < k$, tai transformuoti klasterių centrai μ_j^z, kad ir daugiamačių skalių metodu, projektuojami iš R^k į R^s erdvę. Gauname klasterių centrų $\mu_j^z \in \mathbb{R}^k$ projekcijas $\mu_j^y \in \mathbb{R}^s$, $j = \overline{1, k}$. $T_i = \mu_j^y$, jei $X_i \in K_j$, $i = \overline{1, m}$. Pastebėsime, kad jeigu $s = k$, tai $T_i = \mu_j^z$, t. y. projektavimas iš $\mathbb{R}^k$ į $\mathbb{R}^s$ nėra reikalingas.

REGM tinklo mokymas vyksta dviem etapais. 3.2 paveiksle pateiktoje tinklo REGM schemoje yra pažymėtas kiekvienas mokymo etapas.

I etapas. Atliekama daugiamačių duomenų transformacija į mažesnio matavimo erdvę naudojant radialines bazines funkcijas;

II etapas. „Klaidos sklidimo atgal“ algoritmu apmokomas daugiasluoksnis perceptronas.

Apžvelgsime kiekvieną mokymo etapą plačiau.

3.3.1. Pirmasis etapas

Pirmojo mokymo etapo metu atliekama daugiamačių duomenų transformacija su eksponentine (3.1) arba Gausine (3.2) funkcijomis.

Atliekant daugiamačių duomenų $X_i \in \mathbb{R}^n$, $i = \overline{1,m}$ transformaciją į $Z_i \in \mathbb{R}^k$, $i = \overline{1,m}$, naudojantis eksponentine arba Gausine funkcijomis, svarbu tinkamai parinkti funkcijų parametrus – centrus μ_j ir pločio parametrą σ. Centrus, kaip ir dauguma autorių (Pierrefeu ir kt., 2006; Chang ir kt., 2005; Benoudjit ir Verleysen, 2003), disertacijoje siūloma parinkti klasterizuojant duomenis k-vidurkių metodu. Tačiau eksponentinės ir Gausinės funkcijų rezultatai priklauso ne vien nuo tinkamai parinktų centrų μ_j, bet ir nuo pločio parametro σ.

Radialinių bazinių funkcijų neuroniniuose tinkluose pločio parametras σ gali būti parenkamas pagal tinklo daromą paklaidą (Pierrefeu ir kt., 2006; Chang ir kt., 2005; Benoudjit ir Verleysen, 2003). Anksčiau minėti autoriai RBF tinklą su tais pačiais daugiamačiais duomenimis, bet skirtingu pločio parametru σ apmoko keletą kartų. Tinkamiausia σ yra ta, su kuria tinklas daro mažiausią paklaidą. Tačiau pasiūlytame REGM tinkle toks pločio parametro σ parinkimas nėra tinkamas, nes hibridiniame tinkle yra panaudotos tik radialinių bazinių funkcijų neuroninio tinklo idėjos, o ne visas neuroninis tinklas.

Remiantis (Pierrefeu ir kt., 2006) rezultatais, pločio parametrą σ galima skaičiuoti pagal (2.40) formulę. Formulės autoriai parinkdami konstantą β, peržiūri rekomenduotinų β reikšmių intervalą [3,6; 0,05] (t. y. $\alpha \in$ [0,28; 20]) žingsniu 0,05. Su skirtingomis β reikšmėmis apmokomas RBF tinklas ir fiksuojama β reikšmė, su kuria tinklas daro mažiausią paklaidą. Bet kaip jau yra paminėta anksčiau, toks β nustatymo būdas mums nėra tinkamas. Todėl žemiau pasiūlytas kitas būdas, kaip skaičiuoti σ pagal (2.40) formulę. (2.40) formulės autoriams (Pierrefeu ir kt., 2006) patogu konstantą α pakeisti į $\frac{1}{\beta}$. Tačiau galima ir tiesiogiai naudoti konstantą α. Konstantos α reikšmė nustatoma pagal objektų išsibarstymą kiekviename klasteryje, t. y. skaičiuojama dispersija:

$$D_{K_j} = \frac{1}{km_{K_j} - 1} \sum_{X_i \in K_j} \sum_{\tilde{j}=1}^{k} \left(x_{i\tilde{j}}^{K_j} - \overline{x}_{K_j} \right)^2, \tag{3.3}$$

čia K_j – j-asis klasteris, $j = \overline{1,k}$; k – klasterių skaičius; m_{K_j} – objektų klasteryje K_j skaičius, $\sum_{j=1}^{k} m_{K_j} = m$; $x_{i\tilde{j}}^{K_j}$ yra K_j klasterio i-ojo objekto $\tilde{j}$-ojo požymio reikšmė; $\overline{x}_{K_j}$ – klasterio K_j objektų bendras požymių reikšmių vidurkis:

$$\overline{x}_{K_j} = \frac{1}{km_{K_j}} \sum_{X_i \in K_j} \sum_{\tilde{j}=1}^{k} x_{i\tilde{j}}^{K_j}.$$

Konstanta α parenkama iš tam tikro intervalo (kiekvienam duomenų rinkiniui ir skirtingoms radialinėms bazinėms funkcijoms intervalo rėžiai parenkami atskirai, jų parinkimas plačiau aprašytas 4.2.1 poskyryje), tą intervalą prabėgant žingsniu 0,01 ir kiekvienoje iteracijoje apskaičiuojant τ reikšmę pagal formulę:

$$\tau = \frac{1}{k} \sum_{j=1}^{k} D_{K_j}, \tag{3.4}$$

čia τ yra dispersijų (3.3) vidurkis.

Kiekviena gauta τ reikšmė yra lyginama su prieš tai gauta τ reikšme. Kai skirtumas tarp τ reikšmių pasiekia užsibrėžtą tikslumą $\epsilon = 0,0001$ ($0 < \tau^{u-1} - \tau^u \leq 0,0001$, čia u – iteracijos numeris), tai fiksuojama konstantos α reikšmė ir iteracinis procesas stabdomas. Gautoji konstantos α reikšmė įstatoma į (2.40) formulę ir apskaičiuojamas pločio parametras σ.

Radus tinkamas centrų μ_j ir pločio parametro σ reikšmes, radialinės bazinės funkcijos tampa pilnai apibrėžtos.

3.3.2. Antrasis etapas

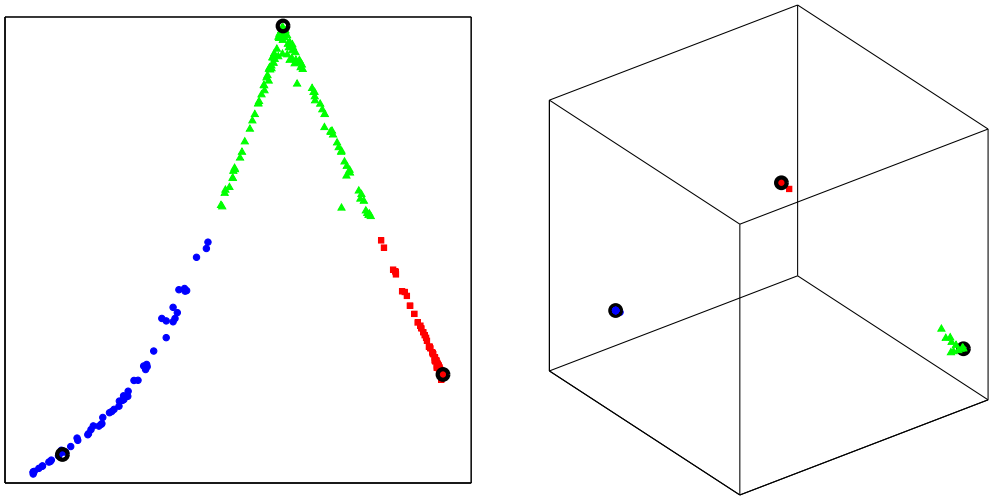
Šiame etape „klaidos sklidimo atgal“ algoritmu apmokomas daugiasluoksnis perceptronas. Daugiasluoksnis perceptronas yra apmokomas po transformacijos gautu nauju duomenų rinkiniu $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_m\} = \{z_{ij}, i = \overline{1,m}, j = \overline{1,k}\}$. Norimos tinklo atsako reikšmės T_i gali būti nustatomos pagal vieną iš dviejų būdų, aprašytų 3.3. poskyryje.

3.4. Gautų rezultatų vizualizavimo kokybės kriterijai

Kaip buvo minėta 3.2. poskyryje, po tinklo REGM apmokymo daugiamačių duomenų projekcija gaunama 3.2 paveiksle pateikto tinklo mažojo sluoksnio išėjime, ir 3.3 paveiksle pateikto tinklo išėjime, kai į tinklą yra paduodamas n-matis analizuojamų duomenų rinkinys $\mathbf{X}$. Vizualiai pateikta projekcija turėtų tyrėjui padėti atskleisti daugiamačiuose duomenyse esančių klasterių savybes.

Pastebėsime, kad 3.2 paveiksle pateikto tinklo išėjimų sluoksnyje gautų reikšmių vizualizavimas mums parodo ar tinklas kokybiškai apmokytas. Kadangi REGM tinklo norimos tinklo atsako reikšmės yra klasterių centrai, tai idealiu atveju išėjimų sluoksnyje turėtų gautis tik tiek skirtingų reikšmių, kiek duomenyse pasirinkta klasterių. Paprastumo dėlei po REGM tinklo apmokymo mažajame ir išėjimų sluoksniuose gautų rezultatų atvaizdavimą plokštumoje arba trimatėje erdvėje vadinsime mažajame (išėjimų) sluoksnyje gautų reikšmių vaizdu arba gautais vizualizavimo rezultatais.

REGM tinklo, apmokyto Širdies ligų duomenų rinkiniu (aprašymas pateiktas 4.1. poskyryje, $k = 3$), gauti vizualizavimo rezultatai mažajame neuronų sluoksnyje P^2 ir išėjimų sluoksnyje Y pateikti 3.4 paveiksle.

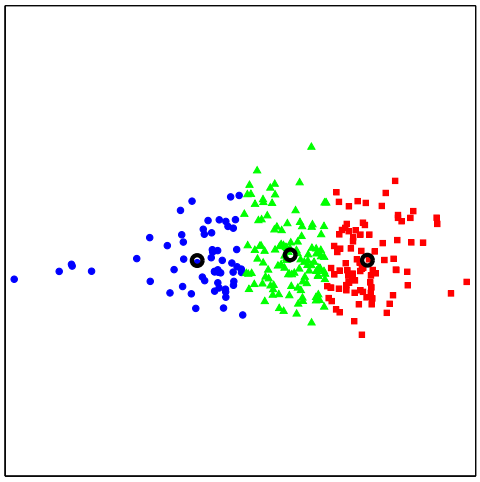


(a) Mažasis sluoksnis **(b)** Išėjimų sluoksnis

3.4 pav. REGM tinklo vizualizavimo rezultatai, $E(W) = 0{,}0006$

Disertacijoje pateiktuose paveiksluose nėra skalių žymėjimo, nes čia aktualus tik taškų tarpgrupinis išsidėstymas. 3.4 paveiksle skirtingų klasterių objektai pažymėti skirtingomis spalvomis ir ženklais (pirmas klasteris – ●; antras klasteris – ▲; trečias klasteris – ■). ○ žymi klasterių centrus. Pastebėsime, kad informacija apie objektų priskyrimą konkrečiam klasteriui yra gaunama k-vidurkių klasterizavimo metodu, kuris yra naudojamas radialinių bazinių funkcijų centrų μ_j nustatymui. Pateiktame 3.4b paveiksle matomos trys gana kompaktiškos taškų sankaupos. Tai rodo pakankamai gerą tinklo apmokymą, nes idealiu atveju turėtų būti tik trys taškai. Mažajame sluoksnyje gauta Širdies ligų duomenų projekcija buvo palyginta su daugiamačių skalių metodu gauta projekcija, kuri pateikta 3.5 paveiksle. Pastebėsime, kad atliekant projekciją daugiamačių skalių metodu stengiamasi išlaikyti atstumus tarp taškų prieš projekciją ir po jos. Tuo tarpu REGM tinkle pirmenybė teikiama tarpgrupiniams objektų panašumams/skirtingumams, o ne atstumo išlaikymui tarp skirtingų objektų.

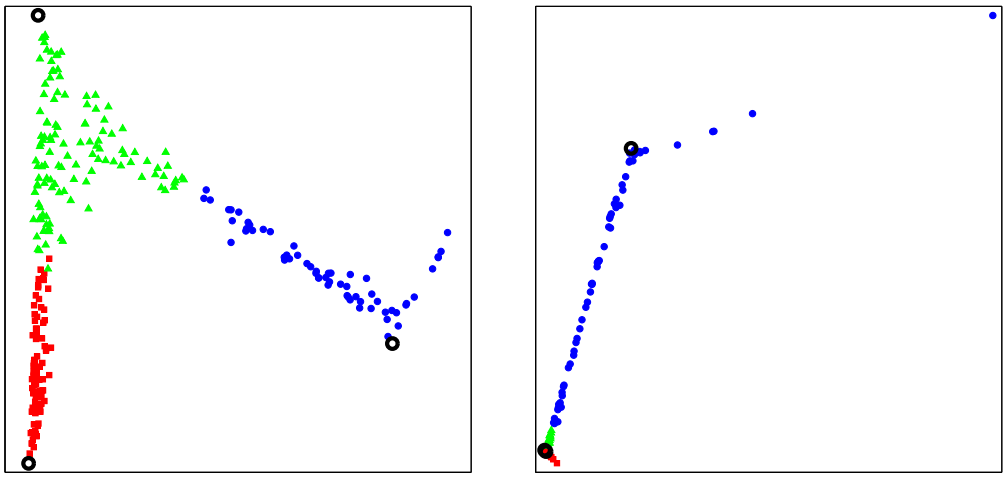
Abiejuose 3.4a ir 3.5 paveiksluose matoma daug taškų, sudarančių klasterį, t. y. taškai klasteriuose „nesusispietę“ aplink savo klasterio centrą, kaip išėjimo sluoksnyje (3.4b paveikslas) gautame vizualizavimo rezultate. Tačiau taškų išsidėstymas klasteriuose skiriasi. Daugiamačių skalių metodu rastoje projekcijoje, kiekvieno klasterio taškai „išsibarstę“ į visas puses nuo klasterio centro (3.5 paveikslas). Po REGM tinklo apmokymo gautoje projekcijoje, taškai klasteriuose išsidėsto kelių tiesių ar kreivių aplinkoje.



3.5 pav. Daugiamačių skalių metodu gauta duomenų projekcija

Tokio taškų išsidėstymo tiesių ar kreivių aplinkoje privalumas, kad išryškinami taškai, kurie turi panašumo su gretimų klasterių taškais arba išskiriami taškai, kurie būdingi tik konkrečiam klasteriui. Panašumo turintys taškai išdėstomi arčiau vieni kitų, o nepanašūs taškai išdėstomi toliau vieni nuo kitų. 3.4a paveiksle matomos ribos tarp klasterių, kai tuo tarpu 3.5 paveiksle aiškios ribos (tarpo) tarp klasterių nematyti.

Paprastai neuroninio tinklo mokymas pradedamas svoriams parenkant tam tikras atsitiktines reikšmes. Tad ir rezultatas priklauso nuo tų reikšmių. Todėl siekiant geriausio vizualaus duomenų atvaizdavimo, tikslinga REGM tinklą apmokyti keletą kartų. 3.6 paveiksle pateikiama keletas pavyzdžių, kokie dar gali būti gaunami vizualizavimo rezultatai.



(a) $E(W) = 0{,}0006$ **(b)** $E(W) = 0{,}0008$

3.6 pav. REGM tinklo vizualizavimo rezultatai mažajame sluoksnyje

Palyginus 3.4a su 3.6a paveikslu matyti, kad 3.6a paveiksle taškai yra labiau „pasibarstę“. ● pažymėtame klasteryje aiškiai išsiskiria taškai, kurie būdingi tik šiam klasteriui. Tačiau 3.4a paveiksle gauta projekcija yra informatyvesnė, nes ▲ pažymėtame klasteryje aiškiai išsidėsto taškai, kurie turi daugiau panašumo su ● pažymėto klasterio taškais, ir kurie – su ■ pažymėto klasterio taškais. Taip pat aiškiau matomos ribos tarp klasterių.

3.6b paveiksle pateikta gautoji projekcija neinformatyvi, nes ▲ ir ■ pažymėtų klasterių taškai yra išsidėstę trumpų tiesių aplinkoje ir sunku įžvelgti tarpgrupinius panašumus/skirtingumus.

Iš aptartų paveikslų matome, kad ne visos vizualiai pateiktos duomenų projekcijos yra informatyvios ir atitinka užsibrėžtą disertacijos tikslą. Todėl mažajame sluoksnyje gautai duomenų projekcijai, pagal siekiamą disertacijos tikslą, buvo užsibrėžti vizualizavimo kokybės kriterijai, kurie įvertina gautą vizualizavimo rezultatą:

1. Taškų išsidėstymas tiesių ar kreivių aplinkoje.
2. Taškų „išsibarstymas“ klasteryje.
3. Riba tarp klasterių.

Pirmasis vizualizavimo kokybės kriterijus yra kokybinis, o kiti du kiekybiniai. Toliau bus pakomentuoti užsibrėžtieji vizualizavimo kokybės kriterijai plačiau.

Pirmasis vizualizavimo kokybės kriterijus nurodo taškų išsidėstymą projekcijoje. Vizualizuotuose duomenyse išsidėstę taškai turėtų sudaryti tieses arba kreives. Toks taškų išsidėstymas atskleidžia jų tarpgrupinius panašumus ir skirtingumus. Klasterių taškai, kurie turi panašumo tik su vieno gretimo klasterio taškais, išsidėsto vienos tiesės ar kreivės aplinkoje. Klasterių taškai, kurie turi panašumo su kelių gretimų klasterių taškais, idealiu atveju išsidėsto kelių tiesių ar kreivių aplinkoje. Pavyzdžiui, jei klasterio taškai turi panašumo su gretimų dviejų klasterių taškais, tai projekcijoje šio klasterio taškai turėtų išsidėstyti dviejų tiesių ar kreivių aplinkoje. Tačiau klasterio taškai gali išsidėstyti ir trijų tiesių ar kreivių aplinkoje, t. y. dviejų tiesių ar kreivių aplinkoje išsidėsto taškai turintys panašumo su gretimų klasterių taškais, o trečiosios tiesės ar kreivės aplinkoje išsidėsto taškai būdingi tik šiam klasteriui. Taškų išsidėstymas labiau primenantis „debesį“ tampa neinformatyvus, nes sudėtingiau išskirti tarpgrupinius taškų panašumus/skirtingumus.

Antrasis vizualizavimo kokybės kriterijus yra labai susijęs su pirmuoju vizualizavimo kokybės kriterijumi. Šis kriterijus nurodo, kad kiekviename klasteryje turi matytis kiek galima daugiau klasterį sudarančių taškų.

Kadangi taškai projekcijoje išsidėsto tiesių ar kreivių aplinkoje, tai taškų „pasibarstymą“ ant tiesių ar kreivių galima apskaičiuoti pagal didžiausią atstumą tarp klasterio K_j, $j = \overline{1,k}$, taškų. Šį atstumą žymėsime $\bar{a}_{K_j}$. 3.4a ir 3.6 paveiksluose esančiose projekcijose visų klasterių didžiausi atstumai $\bar{a}_{K_j}$ pateikti 3.2 lentelėje. Pastebėsime, kad pateiktose projekcijose didžiausias atstumas tarp taškų yra lygus 1. Toks normavimas padarytas siekiant turėti galimybę lyginti skirtingus vizualizavimo rezultatus.

3.2 lentelė: Antrojo vizualizavimo kokybės kriterijaus įverčiai

Paveikslas	● klasteris	▲ klasteris	■ klasteris
3.4a	0,57	0,41	0,29
3.6a	0,52	0,49	0,41
3.6b	0,96	0,04	0,03

Peržiūrėjus pateiktus vizualizavimo rezultatus (3.4a ir 3.6 paveiksluose) ir gautus taškų „pasibarstymo“ klasteriuose skaitinius įvertinimus (3.2 lentelė) matyti, kad kai $\bar{a}_{K_j}$ reikšmė yra mažesnė už 0,1, tai klasterio taškai yra susispietę į sankaupą. Tikslas yra pamatyti taškų tarpgrupinius panašumus/skirtingumus, todėl antrąjį vizualizavimo kokybės kriterijų atitinka tik tos projekcijos, kuriose visų klasterių didžiausi atstumai $\bar{a}_{K_j}$ yra didesni už 0,1 ($\bar{a}_{K_j} > 0{,}1$). Pagal 3.2 lentelėje pateiktus duomenis matome, kad antrojo vizualizavimo kokybės kriterijaus neatitinka 3.6b paveiksle pateiktoji projekcija.

Pirmieji du vizualizavimo kokybės kriterijai yra svarbiausi. Trečiasis vizualizavimo kokybės kriterijus yra pageidaujamas, bet neprivalomas. Šis kriterijus nurodo, kad turi būti riba tarp klasterių, t. y. tam tikras tarpas tarp skirtingų klasterių. Trečiasis vizualizavimo kokybės kriterijus – tai mažiausias atstumas tarp gretimų klasterių taškų – žymimas $\hat{a}$. Atstumu tarp dviejų (gretimų) klasterių disertacijoje vadiname mažiausią atstumą tarp skirtingiems klasteriams priklausančių taškų. 3.4a ir 3.6 paveiksluose esančiose projekcijose mažiausi atstumai $\hat{a}$ tarp skirtingiems klasteriams priklausančių taškų pateikti 3.3 lentelėje.

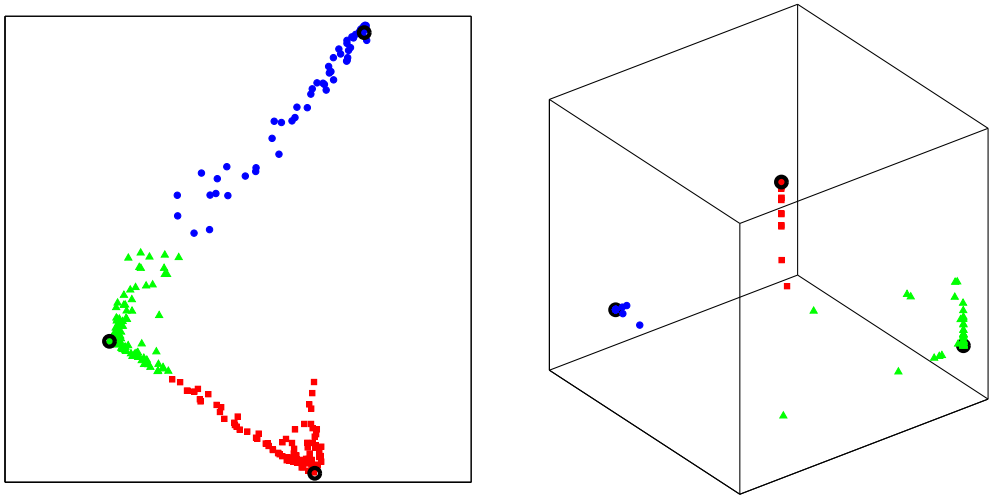
3.3 lentelė: Trečiojo vizualizavimo kokybės kriterijaus įverčiai

Paveikslas	● ir ▲ klasteriai	▲ ir ■ klasteriai
3.4a	0,07	0,05
3.6a	0,05	0,01
3.6b	0,01	0,00

Peržiūrėjus pateiktus vizualizavimo rezultatus (3.4a ir 3.6 paveiksluose) ir gautus mažiausius atstumus tarp skirtingiems klasteriams priklausančių taškų (3.3 lentelė) matyti, kad kai $\hat{a}$ reikšmė yra mažesnė už 0,05, tai

vizualizavimo rezultate riba tarp klasterių įžiūrima sunkiai. Pavyzdžiui, 3.6a paveiksle riba matoma tik tarp ● ir ▲ pažymėtų klasterių, o atstumas $\hat{a}$ lygus 0,05. Tuo tarpu tarp ▲ ir ■ pažymėtų klasterių ribos nesimato, nes $\hat{a} = 0{,}01$. Taigi trečiąjį vizualizavimo kokybės kriterijų atitinka tik tos projekcijos, kuriose mažiausias atstumas $\hat{a}$ tarp gretimiems klasteriams priklausančių taškų yra lygus arba didesnis 0,05 ($\hat{a} \geq 0{,}05$).

Kaip jau yra minėta, paprastai neuroninio tinklo mokymas pradedamas svoriams parenkant tam tikras atsitiktines reikšmes. Todėl siekiant geriausio vizualaus duomenų atvaizdavimo, tikslinga REGM tinklą apmokyti keletą kartų (pažymėkime c – tinklo apmokymų skaičius) ir parinkti geriausią. 3.7 ir 3.8 paveiksluose pateikti iš c tinklo apmokymų atrinkti du vizualizavimo rezultatai mažajame ir išėjimo sluoksniuose. Tinklas apmokytas Širdies ligų duomenų rinkiniu, kurio aprašymas pateiktas 4.1. poskyryje. Paprastumo dėlei tinklą, kuris po apmokymo gauna mažiausią paklaidą (3.7 paveikslas) pavadinkime a tinklu, o tinklą, kurio mažojo sluoksnio vizualizavimo rezultatai labiau atitinka užsibrėžtus vizualizavimo kokybės kriterijus (3.8 paveikslas) – b tinklu.



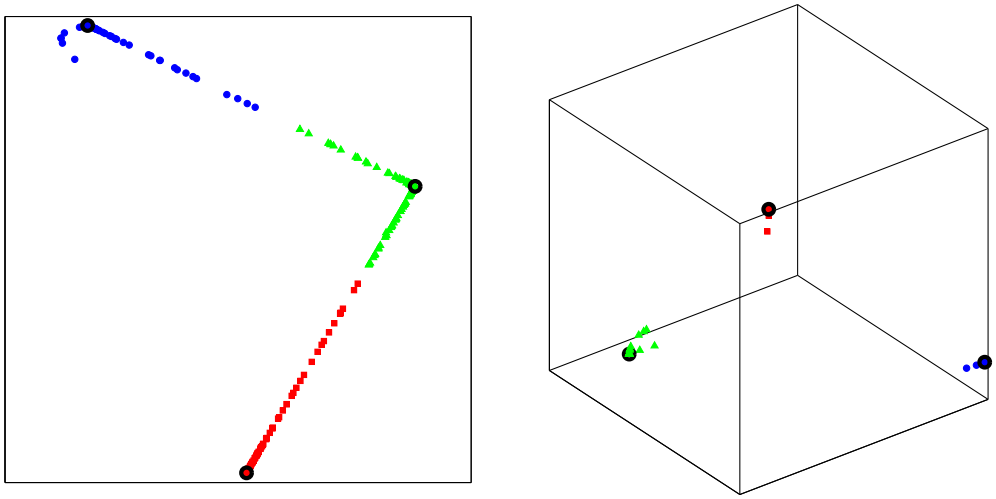
(a) Mažasis sluoksnis **(b)** Išėjimų sluoksnis

3.7 pav. a tinklas, $E(W) = 0{,}0006$

Vizualizavimo kokybės antrojo ir trečiojo kriterijaus įverčiai pateikti 3.4 ir 3.5 lentelėse.

3.4 lentelė: Antrojo vizualizavimo kokybės kriterijaus įverčiai Širdies ligų duomenų rinkiniui

Paveikslas	● klasteris	▲ klasteris	■ klasteris
3.7a	0,59	0,27	0,38
3.8a	0,43	0,32	0,46



(a) Mažasis sluoksnis **(b)** Išėjimų sluoksnis

3.8 pav. b tinklas, $E(W) = 0{,}0007$

3.5 lentelė: Trečiojo vizualizavimo kokybės kriterijaus įverčiai Širdies ligų duomenų rinkiniui

Paveikslas	● ir ▲ klasteriai	▲ ir ■ klasteriai
3.7a	0,06	0,02
3.8a	0,10	0,05

Palyginkime 3.7a ir 3.8a paveiksluose pateiktus po tinklo apmokymo gautus vizualizavimo rezultatus:

- Po apmokymo mažesnę paklaidą $E(W)$ (2.21) daro tinklas a.
- Pagal išėjimų sluoksnyje gautą vizualizavimo rezultatą matome, kad b tinklas apmokytas kokybiškiau, nes klasterių taškai labiau prigludę prie savo klasterių centrų. a tinkle ▲ ir ■ pažymėtų klasterių taškai yra labiau pasibarstę.
- Užsibrėžtus vizualizavimo kokybės kriterijus labiau atitinka b tinklas:
  1) b tinkle taškai aiškiai išsidėstę ant tiesių ar kreivių. Tuo tarpu a tinkle matyti didesnis taškų „pasibarstymas“, kuris apsunkina taškų tarpgrupinių panašumų įvertinimą.
  2) Antrąjį vizualizavimo kokybės kriterijų atitinka abi projekcijos, nes visi $\bar{a}_{K_j} > 0,1$. Tačiau a tinkle ▲ pažymėto klasterio taškai užima mažiau vietos, nes atstumo $\bar{a}_{K_j}$ reikšmė yra mažesnė už b tinkle gautą reikšmę.
  3) b tinkle matomos aiškesnės ribos tarp klasterių, nes atstumai tarp dviejų klasterių $\hat{a} \geq 0,05$, o a tinkle aiški riba matoma tik tarp klasterių, pažymėtų ▲ ir ●.

Taigi, pagal 3.7 ir 3.8 paveiksluose pateiktus vizualizavimo rezultatus matome, kad ne visada tinklas, darantis mažiausią paklaidą $E(W)$, duoda geresnius vizualizavimo rezultatus. Susiduriame su problema, kaip iš c tinklo apmokymų, kurie atlikti esant skirtingoms pradinėms sąlygoms (t. y. skirtingoms pradinėms tinklo svorių reikšmėms), atrinkti geriausią rezultatą – tinklą, kuriuo vizualizuoti rezultatai atitinka užsibrėžtus vizualizavimo kokybės kriterijus. Vienas iš būdų yra peržiūrėti visų tinklo c mokymų rezultatus ir išrinkti tinkamiausią. Tačiau vizualizavimas užima daug laiko ir jei tinklas apmokomas labai daug kartų (pavyzdžiui $c = 50$ arba $c = 100$), tai iš gausybės gautų vaizdų atrinkti vieną tinkamiausią žmogui labai sudėtinga. Todėl atsirado poreikis automatizuoti atranką. Šioje disertacijoje pasiūlyti du atrankos kriterijai:

1. Klasterių išsaugojimas duomenyse po tinklo apmokymo.
2. Išėjimų sluoksnyje gautų taškų išsibarstymas.

Dabar apie kiekvieną atrankos kriterijų plačiau.

Prieš pradedant mokyti tinklą REGM yra atliekamas daugiamačių duomenų klasterizavimas k-vidurkių metodu. Klasterizavimo rezultate kiekvienas n-matis taškas (objektas) X_i priskiriamas konkrečiam klasteriui K_j, kuris apima panašius objektus. Vizualizavimo naudojant REGM tinklą rezultatas yra s-mačių taškų rinkinys $\{Y_i, i = \overline{1,m}\}$. Kiekvienam X_i atitinka konkretus vienas Y_i. Po vizualizavimo norėtųsi, kad klasterį K_j sudarančius n-mačius taškus atitinkantys s-mačiai taškai taip pat sudarytų tokius pačius klasterius.

Klasterių išsaugojimo duomenyse kriterijaus χ reikšmė turėtų atspindėti, kiek s-mačių taškų, pakeitė savo klasterius lyginant su n-mačiu atveju.

Viena iš klasterio charakteristikų yra jo centras. Radialinėse bazinėse funkcijose yra naudojami duomenų rinkinio $\mathbf{X}$ klasterių svorio centrai μ_j, $j = \overline{1,k}$. Apmokius tinklą duomenų rinkiniu $\mathbf{X}$, ir tinklui pateikus klasterių svorių centrus μ_j, $j = \overline{1,k}$, išėjime gaunamos tų n-mačių centrų s-matės projekcijos μ_j^y, $j = \overline{1,k}$. Toliau s-mačių taškų Y_i, $j = \overline{1,m}$, klasterizavimas daromas taip:

1. Apskaičiuojami Y_i, $i = \overline{1,m}$, į tinklą paduodant X_i, $i = \overline{1,m}$.
2. Skaičiuojami atstumai tarp tinklo išėjime gautų s-mačių Y_i ir s-mačių centrų μ_j^y.
3. Taškas Y_i priskiriamas tam klasteriui K_j^y, iki kurio centro μ_j^y atstumas yra mažiausias.

Idealiu atveju klasterius K_j^y ir K_j turėtų sudaryti tais pačiais numeriais pažymėti taškai, t. y. jei $X_i \in K_j$, tai ir $Y_i \in K_j^y$. Tačiau bendru atveju gali nutikti, kad $X_i \in K_j$, o $Y_i \notin K_j^y$.

Klasterių išsaugojimo duomenyse kriterijaus χ reikšmė bus bendras skaičius s-mačių taškų Y_i visuose klasteriuose K_j^y, $j = \overline{1,k}$, kur galioja tokia sąlyga: $Y_i \notin K_j^y$, kai $X_i \in K_j$.

Buvo atlikti trys eksperimentai, kurie turėjo parodyti, ar tinklas, darantis mažiausią paklaidą, visada išsaugo tokius pat klasterius s-mačiuose duomenyse, kaip tai yra n-mačiuose duomenyse. Visi trys eksperimentai buvo atlikti su tokios pat sudėties neuroniniu tinklu, tik kiekvieną kartą buvo parenkami vis kiti pradiniai svoriai. Vieno eksperimento metu REGM tinklas buvo apmokytas 5 kartus. Eksperimentų rezultatai, kai tinklas buvo apmokytas Stuburo ligų duomenų rinkiniu (aprašymas pateiktas 4.1. poskyryje), pateikiami 3.6 lentelėje. Pasirinktas klasterių skaičius $k = 3$, todėl išėjimų sluoksnyje daugiausiai gali būti trys neuronai. 3.6 lentelėje pateikti duomenys surikiuoti pagal tinklo daromą paklaidą didėjimo tvarka.

3.6 lentelė: Eksperimentų rezultatai parodantys galimas kriterijaus χ reikšmes

Tinklo apmokymo numeris	1 eksperimentas		2 eksperimentas		3 eksperimentas	
	Paklaida	χ	Paklaida	χ	Paklaida	χ
1	0,00092	0	0,00093	1	0,00077	1
2	0,00096	1	0,00094	3	0,00087	11
3	0,00129	3	0,00125	0	0,00093	1
4	0,00133	2	0,00135	0	0,00096	3
5	0,00186	93	0,00202	81	0,00112	89

Iš 3.6 lentelėje pateiktų duomenų matome, kad pirmojo eksperimento metu iš atliktų $c = 5$ tinklo apmokymų, geriausias yra tas, kurio daroma paklaida $E(W)$ yra mažiausia ir duomenyse po transformacijos išsaugojami tokie patys klasteriai, t. y. $\chi = 0$. Antrojo eksperimento metu, mažiausios paklaidos atveju, kažkuris vienas s-matis taškas Y_i priskiriamas kažkuriam kitam klasteriui. Vadinasi, rezultatas su mažiausia paklaida nėra tinkamas, nes po tinklo apmokymo neišsaugojami klasteriai duomenyse. Pagal 3.6 lentelės duomenis, klasteriai duomenyse antrajame eksperimente išsaugojami tik trečiojo apmokymo metu, nors tinklo daroma paklaida yra didesnė, nei pirmuoju apmokymo atveju. Tai dar kartą įrodo, kad ne visada tinklas, darantis mažiausią paklaidą $E(W)$, duoda geresnius vizualizavimo rezultatus. Tačiau trečiojo eksperimento metu nėra nei vieno tinklo apmokymo atvejo, kai po transformacijos duomenyse išsaugojami tokie patys klasteriai, t. y. visiems penkiems tinklo apmokymo atvejams klasterių išsaugojimo

duomenyse kriterijus $\chi > 0$. Jei eksperimento metu, bent po vieno tinklo apmokymo klasterių išsaugojimo duomenyse kriterijus $\chi = 0$, tai kitų apmokymų rezultatai, kai $\chi > 0$, atmetami. Rezultatų vertinimui pagal antrąjį pasiūlytą atrankos kriterijų lieka $\tilde{c}$ tinklo apmokymo rezultatai. Pirmo eksperimento atveju $\tilde{c} = 1$, o antro eksperimento atveju $\tilde{c} = 2$. Jei visais atvejais $\chi > 0$, kaip trečiojo eksperimento atveju, tai paliekami tik tie variantai, kur χ reikšmė yra mažiausia. Pastebėsime, kad gali būti ne vienas rezultatas su tokia pat mažiausia reikšme. 3 eksperimento atveju tokių rezultatų yra $\tilde{c} = 2$.

Įvertinę atvejus pagal pirmąjį atrankos kriterijų taškai, netenkinę sąlygos $Y_i \in K_j^y$, kai $X_i \in K_j$ atmetami, ir pereinama prie antro atrankos kriterijaus skaičiavimo naudojantis likusiu taškų rinkiniu $\{\tilde{Y}_i, i = \overline{1, m - \chi}\}$. Pastebėsime, kad kai $\chi = 0$, tai $\{\tilde{Y}_i, i = \overline{1, m - \chi}\} = \{Y_i, i = \overline{1, m}\}$.

Pagal pirmąjį atrankos kriterijų atmetus tinklo apmokymo rezultatus dar lieka $\widetilde{c}$ tinklo apmokymo rezultatų, kur χ reikšmė yra mažiausia. Jei $\widetilde{c} > 1$, tai turime pasinaudoti antruoju atrankos kriterijumi. Kaip pavyzdys, po 3.6 lentelėje pateiktų rezultatų, apdorotų pirmuoju atrankos kriterijumi, lieka 3.7 lentelėje pateikti rezultatai. 3.7 lentelėje šalia kiekvieno tinklo apmokymo rezultato dar yra pateiktos antrojo atrankos kriterijaus reikšmės κ_q.

3.7 lentelė: Rezultatai, kurie tenkino pirmąjį atrankos kriterijų

	1 eksperimentas			2 eksperimentas			3 eksperimentas		
$q = \overline{1, \tilde{c}}$	Paklaida	χ	κ_q	Paklaida	χ	κ_q	Paklaida	χ	κ_q
1	0,00092	0	0,05	0,00125	0	0,02	0,00077	1	0,05
2				0,00135	0	0,03	0,00093	1	0,01

Paprastumo dėlei klasterių skaičius k ir tinklo išėjimų skaičius s eksperimentuose pasirinkti lygūs tarpusavyje ir lygūs 3. Toks išėjimų skaičiaus pasirinkimas leidžia vizualiai trimatėje erdvėje stebėti išėjime gautų s-mačių taškų Y_i, $i = \overline{1, m}$, išsidėstymą ir įvertinti, ar ir kaip tie taškai sudaro klasterius.

Iš pateiktų 3.7 ir 3.8 paveikslų pastebime, kad užsibrėžtus vizualizavimo kokybės kriterijus mažajame sluoksnyje ($P_i^2 \in \mathbb{R}^2$, $i = \overline{1, m}$) geriau atitinka tinklas, kurio išėjimų sluoksnyje ($Y_i \in \mathbb{R}^3$, $i = \overline{1, m}$) gautame vizualizavimo rezultate klasterių taškai labiau prigludę prie savo klasterių centrų ir nutolę nuo kitų klasterių. Taigi iš likusių $\widetilde{c}$ tinklo apmokymo rezultatų reikia rasti tinklą, kurio išėjimų sluoksnyje gauti taškai, priklausantys skirtingiems klasteriams, koncentruojasi apie savo klasterių centrus – idealiu atveju matoma tik tiek taškų, kiek yra klasterių. Pastebėsime, kad norint vizualiai įvertinti tinklo mokymo kokybę jei klasterių skaičius pasirenkamas $k > 3$, ir išėjimų

sluoksnyje pasirenkamas $3 < s \leq k$ neuronų skaičius, tai išėjimų sluoksnyje gautų rezultatų vaizdas plokštumoje gali būti gaunamas tik pasinaudojus projekcijos metodais (pvz. daugiamatėmis skalėmis), projektuojant $Y_i \in \mathbb{R}^s$, $i = \overline{1,m}$, į dvimatę arba trimatę erdvę.

Skirtingų klasterių taškų $\tilde{Y}_i \in K_j^y$, kai $X_i \in K_j$, tarpusavio išsidėstymą galima įvertinti skaičiuojant atstumą tarp skirtingų klasterių taškų. Pažymėkime didžiausią atstumą tarp skirtingų klasterių taškų visuose $\tilde{c}$ tinklo apmokymuose κ:

$$\kappa = \max_{q=\overline{1,\tilde{c}}} \kappa_q, \tag{3.5}$$

čia q – po pirmojo atrankos kriterijaus likusio apmokymo numeris; κ_q – mažiausias atstumas tarp skirtingų klasterių taškų q-ajame tinklo apmokyme, kuris apskaičiuojamas pagal formulę:

$$\kappa_q = \min_{1 \leq j_1 < j_2 \leq k} \min_{\substack{\tilde{Y}_{i_1} \in K_{j_1}^y \\ \tilde{Y}_{i_2} \in K_{j_2}^y}} \|\tilde{Y}_{i_1} - \tilde{Y}_{i_2}\|, \tag{3.6}$$

čia $\tilde{Y}_i$, $i = \overline{1, m-\chi}$, yra po q-tojo apmokymo tinklo išėjimuose gauti rezultatai. Paprastumo dėlei (3.6) formulės dešinėje indeksas q neįvedamas.

Jei κ_q reikšmė yra didelė, tai reiškia, kad taškai $\tilde{Y}_i \in K_j^y$ yra prigludę prie savo klasterių centrų ir gautame vaizde bus aiškiai matomos taškų sankaupos. Taigi, jei turime $\widetilde{c}$ skirtingų κ_q reikšmių, $q = \overline{1,\widetilde{c}}$, tai geriausias rezultatas bus tame tinklo apmokyme q, kur κ_q yra maksimalus. Pagal 3.7 lentelėje pateiktus rezultatus antrojo eksperimento atveju κ_q maksimali reikšmė gauta antrojo apmokymo metu ($\kappa = 0,03$), o trečiojo eksperimento atveju – pirmojo apmokymo metu ($\kappa = 0,05$).

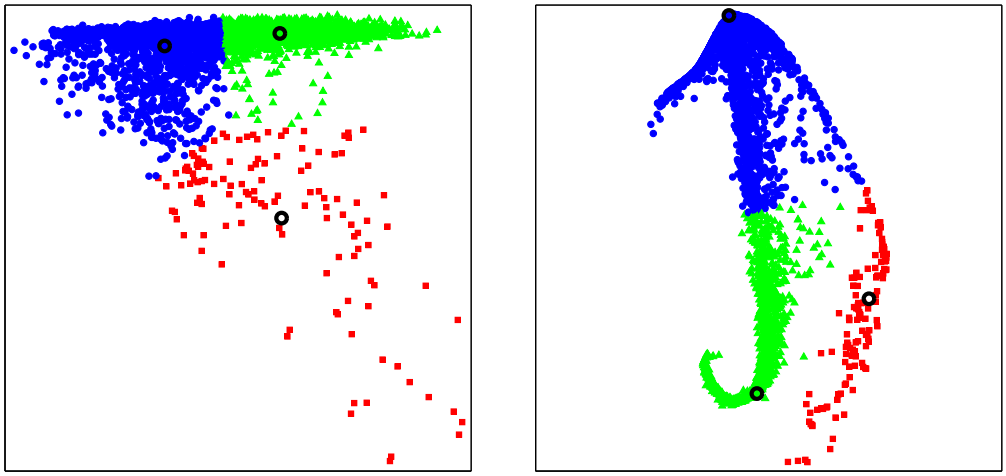
Apmokyto tinklo, atitikusio abu atrankos kriterijus (t. y. duomenyse išsaugojami klasteriai ir κ_q reikšmė yra maksimali iš $\widetilde{c}$ tinklo apmokymų) yra fiksuojamos svorių reikšmės. Duomenų rinkinį papildžius naujais objektais ir juos pateikus į REGM tinklą su fiksuotomis svorių reikšmėmis jų vieta projekcijoje parodoma neatliekant sudėtingų skaičiavimų (tinklo permokymo).

3.5. REGM tinklo praktinis pritaikymas

REGM tinklas naudojamas vizualiai daugiamačių duomenų analizei, kai atidėjimui plokštumoje arba trimatėje erdvėje taškai gaunami paskutinio paslėpto neuronų sluoksnio išėjimuose į tinklą padavus n-mačių analizuojamų duomenų rinkinį $\mathbf{X}$. Duomenų rinkinį $\mathbf{X}$ sudarančių taškų grupių skaičius yra nežinomas. Klasikiniai klasterizavimo metodai atlieka duome-

nų suskirstymą į grupes, tačiau neatskleidžia objektų tarpgrupinių panašumų/skirtingumų. Tuo tarpu REGM tinklo gautoje duomenų projekcijoje plokštumoje labiau atsispindi bendra duomenų struktūra (klasteriai, klasterių tarpusavio artumas, taškų tarpklasterinis panašumas) nei daugiamačių taškų tarpusavio išsidėstymas.

Panagrinėkime pavyzdį su Vystančių medžių duomenų rinkiniu (aprašymas pateiktas 4.1. poskyryje). Sprendžiamas uždavinys: atpažinti ir išskirti nuvytusius ir pradedančius vysti medžius. k-vidurkių metodu klasterizuoti ir daugiamačių skalių metodu vizualizuoti duomenys pateikti 3.9a paveiksle. Skirtingų klasterių objektai pažymėti skirtingomis spalvomis: vešantys medžiai pažymėti ●, vystantys medžiai pažymėti ■, o likęs žemės paviršius pažymėtas ▲. Iš pateikto 3.9a paveikslo galime pasakyti, kuris objektas, priklauso kuriam klasteriui, tačiau labai sunku įvertinti tarpgrupinius objektų panašumus/skirtingumus. O tai būtų labai svarbu, nes tarp vešančių medžių atrasti objektus, kurie turi panašumo su vystančiais medžiais, būtų galima nustatyti medžių vytimo priežastį (drėgmės trūkumas ar liga). Tarpgrupinius objektų panašumus/skirtingumus leidžia įvertinti po REGM tinklo apmokymo gauta projekcija, kuri yra pateikta 3.9b paveiksle.



(a) MDS **(b)** REGM

3.9 pav. Vystančių medžių duomenų projekcija

REGM tinklas, pateiktas 3.2 paveiksle buvo apmokytas 60 kartų. Remiantis atrankos kriterijais buvo nustatyta geriausia tinklo projekcija (3.9b paveikslas), kuri buvo įvertinta užsibrėžtais vizualizavimo kokybės kriterijais. Pirmasis vizualizavimo kokybės kriterijus nurodo, kad taškai turi būti išsidėstę tiesių ar kreivių aplinkoje. Iš 3.9b paveikslo matome, kad taškai yra „pasibarstę“ trijų kreivių aplinkoje. Taigi pirmasis vizualizavimo kriterijus yra tenkinamas. Antrasis vizualizavimo kokybės kriterijus nurodo

taškų „išsibarstymą“ klasteryje: ● pažymėtame klasteryje didžiausias atstumas $\bar{a}_{K_j}$ tarp klasterio taškų yra lygus 0,48 ($\bar{a}_{K_1} = 0{,}48$), ▲ pažymėtame klasteryje didžiausias atstumas yra $\bar{a}_{K_2} = 0{,}51$, o ■ pažymėtame klasteryje didžiausias atstumas yra $\bar{a}_{K_3} = 0{,}62$. Matome, kad visos $\bar{a}_{K_j}$ reikšmės yra didesnės už 0,1, tai vadinasi antrasis vizualizavimo kriterijus tenkinamas. Trečiasis vizualizavimo kokybės kriterijus (riba tarp klasterių) yra pageidautinas, bet nebūtinas. Šiuo atveju atstumas $\hat{a}$ tarp ● ir ▲ pažymėtų klasterių yra 0,006, o tarp ● ir ■ pažymėtų klasterių $\hat{a} = 0{,}02$. Atstumas tarp klasterių $\hat{a}$ turėtų būti didesnis arba lygus 0,05. 3.9b paveiksle pateikta projekcija netenkina trečiojo vizualizavimo kokybės kriterijaus.

Iš 3.9b paveikslo matome, kad ● pažymėto klasterio objektai tarsi suskirstomi į tris grupes. Kiekvienos grupės objektai išdėstomi atskirų tiesių ar kreivių aplinkoje. Vešančių medžių objektai (klasteris pažymėtas ●), turintys panašumo su vystančių medžių objektais (klasteris pažymėtas ■), išdėstomi ant kreivės artimiausios vystančių medžių klasteriui. Būtent į šiuos objektus reikėtų atkreipti didelį dėmesį, nes jie palengvins tyrėjui nustatyti medžių vytimo priežastį. Taip pat iš 3.9b paveiksle pateiktos projekcijos galime teigti, kad vystančių medžių klasterio objektai neturi jokio panašumo su likusio žemės paviršiaus klasterio objektais, nes tarp šių klasterių nėra jungiamosios tiesės arba kreivės.

3.6. Trečiojo skyriaus apibendrinimas ir išvados

Šiame skyriuje pasiūlytas hibridinis neuroninis tinklas REGM, kuris savyje integruoja ir radialinių bazinių funkcijų neuroninio tinklo, ir daugiasluoksnio perceptrono, turinčio „butelio kaklelio“ neuroninio tinklo savybes, idėjas. Tinklas sudarytas iš dviejų dalių. Pirmoji dalis yra tam tikras daugiamatės erdvės taškų transformavimas į norimo mažesnio matmens erdvę. Antroji dalis yra daugiasluoksnis perceptronas, kurio mažasis sluoksnis (paskutinis paslėptas sluoksnis) sudarytas iš nedidelio neuronų skaičiaus (2 arba 3). Hibridinio tinklo REGM paskirtis yra padėti atskleisti duomenyse esančių klasterių savybes, kai žinios apie šių klasterių sudėtį yra gaunamos prieš mokant REGM tinklą, atliekant daugiamačių duomenų klasterizavimą ir naudojamos to tinklo mokymo metu.

REGM tinklas naudojamas vizualiai daugiamačių duomenų analizei, kai atidėjimui plokštumoje arba trimatėje erdvėje taškai gaunami paskutinio paslėpto neuronų sluoksnio išėjimuose į tinklą padavus n-mačių analizuojamų duomenų rinkinį. Šio tinklo ypatybė yra ta, kad gautas vaizdas plokštumoje labiau atspindi bendrą duomenų struktūrą (klasteriai, klasterių tarpusavio artumas, taškų tarpklasterinis panašumas) nei daugiamačių

taškų tarpusavio išsidėstymą. Žinant šią specifiką, pasiūlyti trys vizualizavimo kokybės kriterijai, formalizuojantys gautų vizualizavimo rezultatų įvertinimą:

- taškų išsidėstymas tiesių ar kreivių aplinkoje;
- taškų „išsibarstymas“ klasteryje (didžiausias atstumas tarp klasterio taškų turi būti didesnis už 0,1);
- riba tarp klasterių (mažiausias atstumas tarp skirtingiems klasteriams priklausančių taškų turi būti didesnis arba lygus 0,05).

Siekiant geros vizualizavimo kokybės reikia:

- Tinkamai parinkti transformacijos funkcijos (eksponentinės arba Gausinės) parametrus – centrus μ_j ir pločio parametrą σ. Pločio parametras apskaičiuojamas pagal objektų išsibarstymą klasteriuose ir vidutinį atstumą tarp tų klasterių centrų. Daugiamačių duomenų transformaciją atlikus su taip apskaičiuotu pločio parametru, gaunamos reikšmės išsibarsto intervale $[0, 1]$, t. y. nesikoncentruoja šio intervalo kraštuose.
- Vykdyti keletą tinklo mokymų ir antrojo mokymo etapo metu atrinkti tinklo apmokymo rezultatą, kuris atitiktų užsibrėžtus vizualizavimo kokybės kriterijus. Tai atlieka šiame skyriuje aprašyti atrankos kriterijai.

4. Eksperimentiniai tyrimai

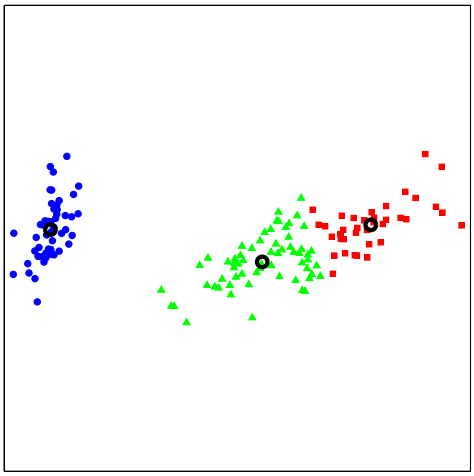
Šiame skyriuje pateikiami atlikti eksperimentai su hibridiniu neuroniniu tinklu REGM.

Pagrindiniai skyriaus rezultatai buvo pristatyti 5 konferencijose ir 3 straipsniuose, kurių sąrašai yra pateikti 1.6. poskyryje.

4.1. Tyrimuose naudojami duomenys

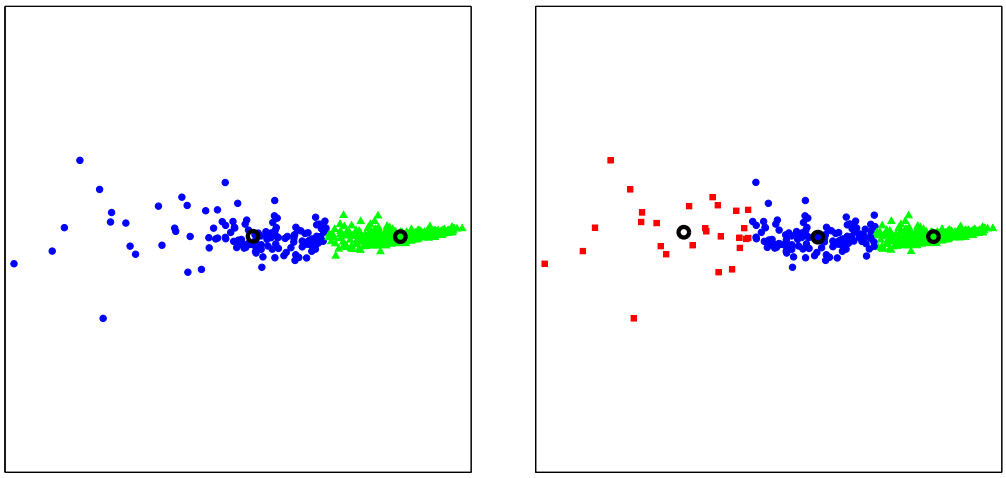
Eksperimentiniuose tyrimuose buvo naudoti 8 daugiamačių duomenų rinkiniai. Dauguma (5 iš pateiktųjų) duomenų rinkinių buvo parenkami taip, kad atlikus duomenų projekciją į plokštumą daugiamačių skalių metodu vizualiai nesimatytų aiškios ribos tarp klasterių. Duomenų rinkiniai į pasirinktą klasterių skaičių klasterizuoti k-vidurkių metodu. Šalia kiekvieno duomenų rinkinio nurodyta iš kelių klasių jis susideda, tačiau tyrimuose daroma prielaida, kad klasterių skaičius yra nežinomas. Vaizdumo dėlei visi duomenų rinkiniai klasterizuoti į pasirinktą klasterių skaičių ir vizualizuoti plokštumoje, kur skirtingų grupių objektai pavaizduoti skirtingomis spalvomis. Visi duomenų rinkiniai paimti iš duomenų bazės (Bache ir Lichman, 2013):

1. **Gėlių irisų duomenų rinkinys** (angl. *Iris Plants Database*) (Fisher, 1936). Šioje disertacijoje bus vadinamas Irisų duomenų rinkiniu. Duomenų rinkinį sudaro trijų rūšių irisai – „Setosa“, „Versicolour“ ir „Virginica“ ($k = 3$). Kiekvienos rūšies yra po 50 gėlių, iš viso 150 ($m = 150$). Kiekvieną irisą apibūdina keturi požymiai – taurėlapio ilgis, taurėlapio plotis, vainiklapio ilgis ir vainiklapio plotis ($n = 4$). Vizualizuotas Irisų duomenų rinkinys pateikiamas 4.1 paveiksle.



4.1 pav. Vizualizuotas Irisų duomenų rinkinys

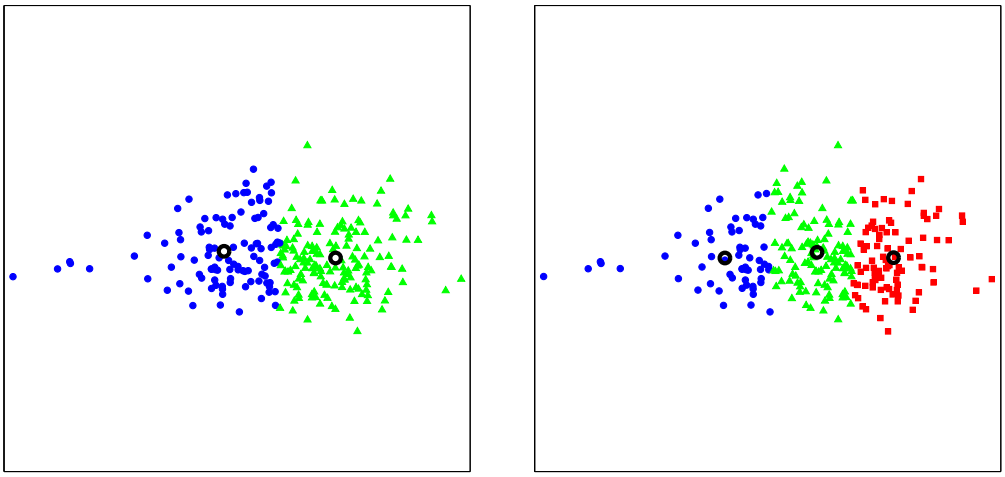
2. **Krūties vėžio duomenų rinkinys** (angl. *Breast Cancer Database*) (Street ir kt., 1993; Mangasarian ir kt., 1995). Duomenų rinkinys klasifikuojamas į 2 klases ($k = 2$) – piktybinis navikas (angl. *malignant*) ir gerybinis navikas (angl. *benign*). Visą duomenų rinkinį sudaro 569 navikai ($m = 569$). Kiekvieną naviką apibūdina 30 požymių: įvairūs naviko matavimai (spindulys, perimetras, plotis, kompaktiškumas ir kt.), vidurkis, standartinė paklaida ($n = 30$). Vizualizuotas Krūties vėžio duomenų rinkinys pateikiamas 4.2 paveiksle.



(a) 2 klasteriai **(b)** 3 klasteriai

4.2 pav. Vizualizuotas Krūties vėžio duomenų rinkinys

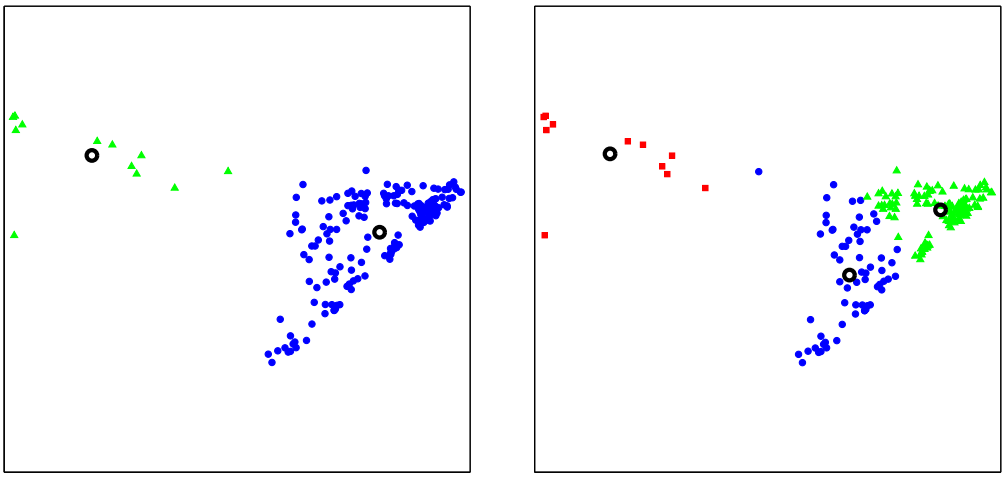
3. **Širdies ligų duomenų rinkinys** (angl. *Heart Database*). Duomenų rinkinys klasifikuojamas į 2 klases ($k = 2$) – sergantys širdies ligomis (angl. *presence of heart disease*) ir sveiki (angl. *absence of heart disease*). Visą duomenų rinkinį sudaro 270 pacientų ($m = 270$). Kiekvieną pacientą apibūdina 13 požymių: amžius, lytis, krūtinės skausmo tipas (angl. *chest pain type*), kraujo spaudimas ramybės būsenoje (angl. *resting blood pressure*), cholesterolio kiekis (angl. *serum cholesterol*), cukraus kiekis kraujyje nevalgius (angl. *fasting blood sugar*), elektrokardiograma ramybės būsenoje (angl. *resting electrocardiographic results*), maksimalus širdies susitraukimų dažnis (angl. *maximum heart achieved*), širdis darbo metu (angl. *exercise induced angina*), fizinio širdies darbo palyginimas su ramybės būsena (angl. *depression induced by exercise relative to rest*), širdies darbas krūvio mažėjimo metu (angl. *the slope of the peak exercise*), širdies sandara ($n = 13$). Vizualizuotas Širdies ligų duomenų rinkinys pateikiamas 4.3 paveiksle.



(a) 2 klasteriai **(b)** 3 klasteriai

4.3 pav. Vizualizuotas Širdies ligų duomenų rinkinys

4. **Parkinsono ligos duomenų rinkinys** (angl. *Parkinson's Database*) (Little ir kt., 2009). Duomenų rinkinys klasifikuojamas į 2 klases ($k = 2$) – sergantys parkinsono liga ir sveiki. Visą duomenų rinkinį sudaro 195 pacientai ($m = 195$). Kiekvieną pacientą apibūdina 22 požymiai, kurie aprašo Parkinsono ligą ($n = 22$). Vizualizuotas Parkinsono ligos duomenų rinkinys pateikiamas 4.4 paveiksle.

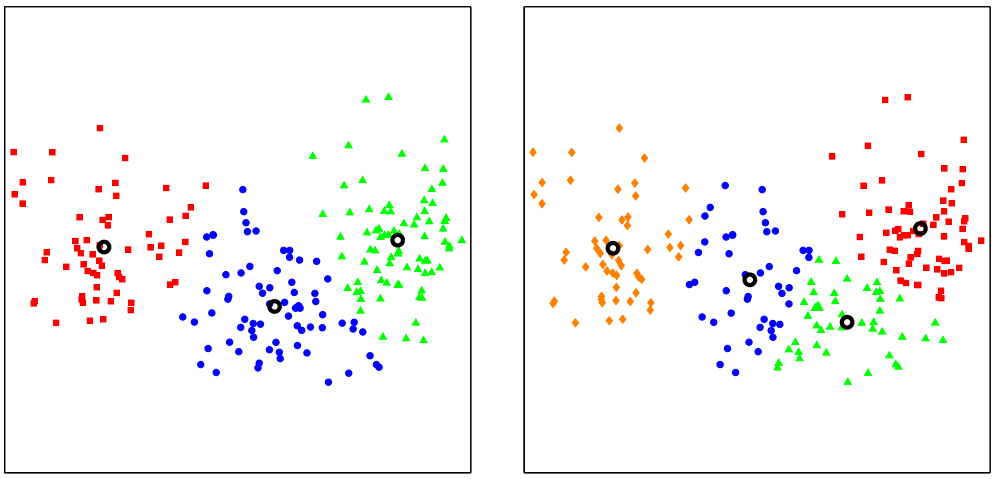


(a) 2 klasteriai **(b)** 3 klasteriai

4.4 pav. Vizualizuotas Parkinsono ligos duomenų rinkinys

5. **Kviečių grūdų duomenų rinkinys** (angl. *Wheat seeds data set*) (Charytanowicz ir kt., 2010). Duomenų rinkinį sudaro trijų rūšių kviečiai – „Kama“, „Rosa“ ir „Canadian“ ($k = 3$). Kiekvienos rūšies yra išmatuota po 70 kviečių, iš viso 210 grūdų ($m = 210$). Kiekvieną kviečio grūdą apibūdina septyni geometriniai požymiai: plotas,

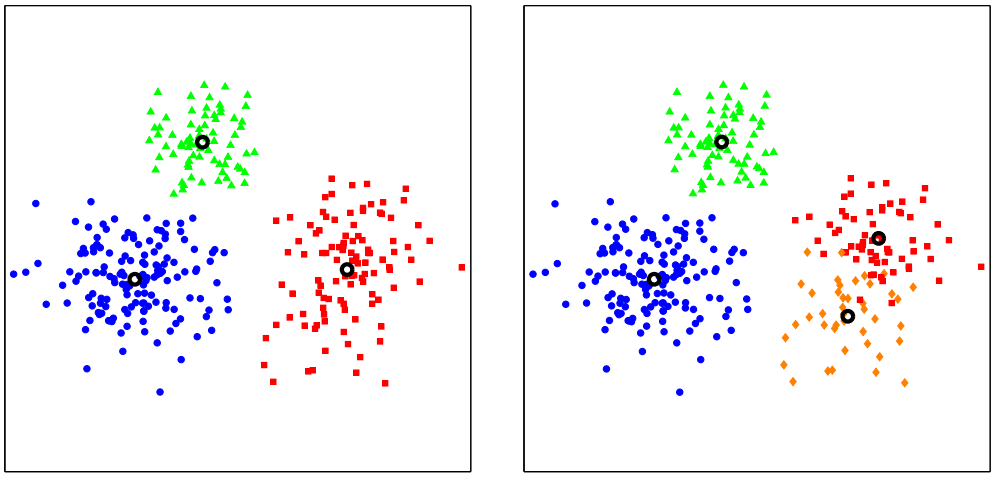
perimetras, kompaktiškumas, branduolio ilgis, branduolio plotis, asimetrijos koeficientas ir branduolio griovelio ilgis ($n = 7$). Vizualizuotas Kviečių grūdų duomenų rinkinys pateikiamas 4.5 paveiksle.



(a) 3 klasteriai **(b)** 4 klasteriai

4.5 pav. Vizualizuotas Kviečių grūdų duomenų rinkinys

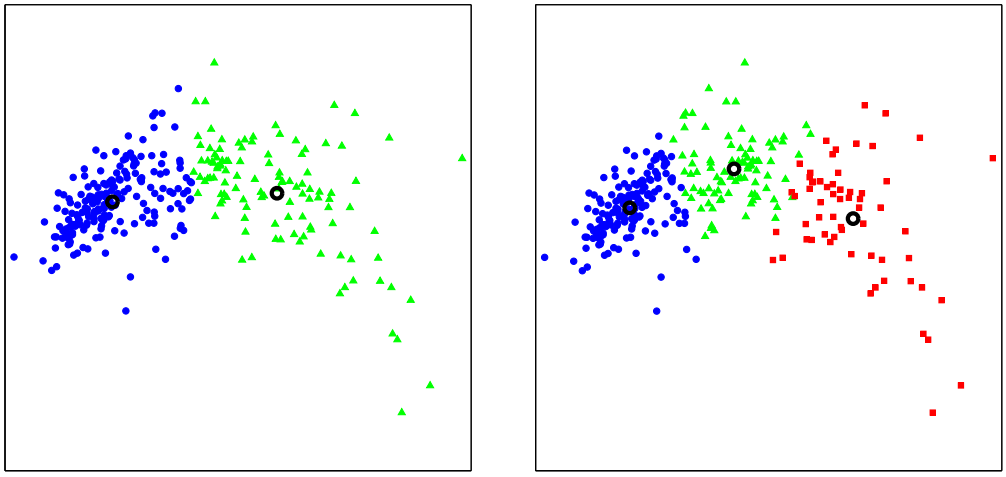
6. **E.coli bakterijų duomenų rinkinys** (angl. *Ecoli data set*) (Horton ir Nakai, 1996). Duomenų rinkinys sudarytas iš 8 skirtingose vietose išsidėsčiusių 336 E.coli bakterijų ($m = 336$). Į klases bakterijos suskirstytos pagal jų buvimo vietas, t. y. duomenų rinkinį sudaro 8 E.coli bakterijų klasės ($k = 8$). Kiekviena E.coli bakterija yra apibūdinta 7 požymiais ($n = 7$). Vizualizuotas E.coli bakterijų duomenų rinkinys pateikiamas 4.6 paveiksle.



(a) 3 klasteriai **(b)** 4 klasteriai

4.6 pav. Vizualizuotas E.coli bakterijų duomenų rinkinys

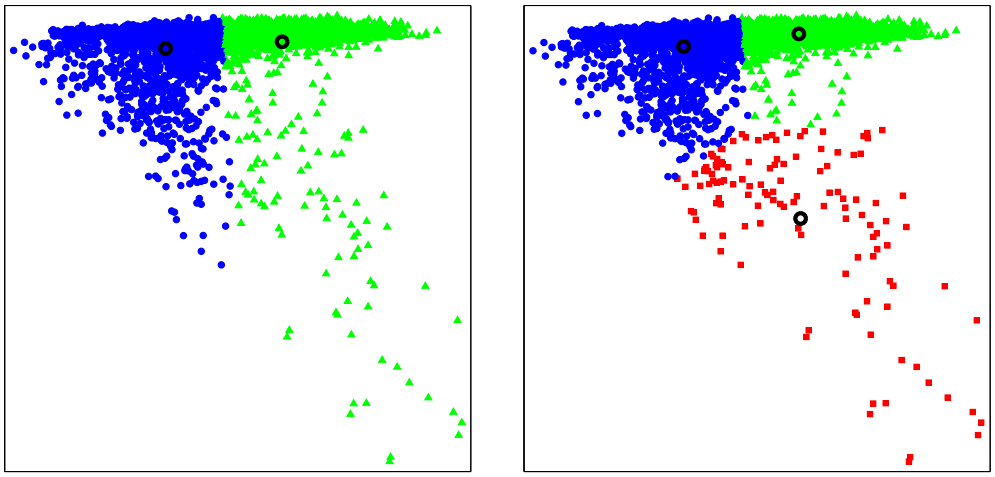
7. **Stuburo ligų duomenų rinkinys** (angl. *Vertebral Column Database*) (Rocha Neto ir kt., 2011). Duomenų rinkinį galima klasifikuoti į 3 klases ($k = 3$) – sveiki, stuburo disko išvarža, spondilolistezė (angl. *normal, disk hernia, spondilolysthesis*) – arba į 2 klases ($k = 2$) – sveiki, sergantys (angl. *normal, abnormal*). Visą duomenų rinkinį sudaro 310 pacientų ($m = 310$). Kiekvieną pacientą apibūdina šeši biomechaniniai požymiai: dubens dažnis (angl. *pelvic incidence*), dubens tentas (angl. *pelvic tilt*), juosmens kampas (angl. *lumbar lordosis angle*), sakraliniai nuolydžiai (angl. *sacral slope*), dubens spindulys (angl. *pelvic radius*) ir spondilolistezės klasė (angl. *the grade of spondylolisthesis*) ($n = 6$). Vizualizuotas Stuburo ligų duomenų rinkinys pateikiamas 4.7 paveiksle.



(a) 2 klasteriai **(b)** 3 klasteriai

4.7 pav. Vizualizuotas Stuburo ligų duomenų rinkinys

8. **Vystančių medžių duomenų rinkinys** (angl. *Wilt data set*) (Johnson ir kt., 2013). Mokymo duomenų rinkinys susideda iš palydovinio vaizdo 4339 segmentų ($m = 4339$). Kiekvienas segmentas sudarytas iš taškų (pikselių), todėl jam įvertinti buvo pasirinkti 5 požymiai ($n = 5$): vidutinės spalvų spektro vertės – R (raudona), G (žalia) ir NIR (angl. *Near-infrared*); du dažniausiai naudojami tekstūros rodikliai – standartinis nuokrypis ir pilko lygio matricos vidurkis (GLCM). Turimame palydoviniame vaizde, pagal matomas spalvas, reikia išskirti ligotus medžius (japoninius ąžuolus ir japonines pušis). Taigi duomenų rinkinys klasifikuojamas į 2 klases ($k = 2$) – ligoti medžiai (angl. *diseased trees*), kurių yra nedaug, ir kitas žemės paviršius (angl. *other land cover*). Vizualizuotas Vystančių medžių duomenų rinkinys pateikiamas 4.8 paveiksle.



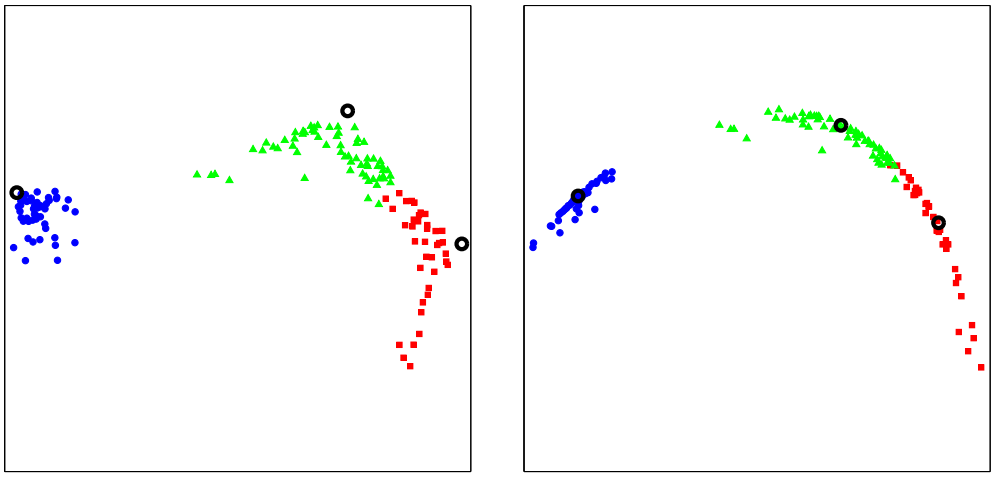
(a) 2 klasteriai **(b)** 3 klasteriai

4.8 pav. Vizualizuotas Vystančių medžių duomenų rinkinys

4.2. Daugiamačių duomenų transformacija

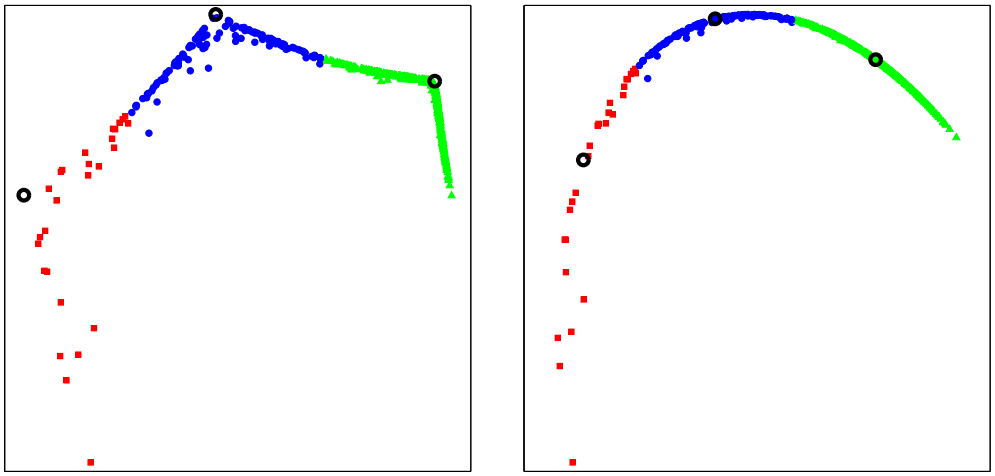
Daugiamačio duomenų rinkinio projekcija į dvimatę erdvę naudojantis MDS labai skiriasi nuo gautų rezultatų mažajame sluoksnyje po tinklo apmokymo (žr. 3.5 paveikslą ir 3.4a paveiksle pateiktus mažojo sluoksnio vizualius rezultatus). Atlikus turimų daugiamačių duomenų rinkinio projekciją į dvimatę erdvę naudojantis MDS, plokštumoje matomi taškų telkiniai, o po REGM tinklo apmokymo, taškai išsidėsto kelių tiesių ar kreivių aplinkoje. Kyla klausimas, kas įtakoja tokį daugiamačių duomenų vizualių rezultatų pasikeitimą ir kaip vizualiai atrodo tarpiniai tinklo rezultatai – radialinių bazinių funkcijų sluoksnio išėjimuose gaunami taškai $Z_i \in \mathbb{R}^k$, kai į įėjimą paduodami taškai X_i, $i = \overline{1,m}$? Tuo pačiu tikimasi, kad vizualizuotos Z_i reikšmės (t. y. transformuotos į $\mathbb{R}^2$ naudojantis kad ir MDS) palengvins radialinių bazinių funkcijų pločio parametro σ nustatymą.

Pirmame hibridinio neuroninio tinklo REGM mokymo etape atliekamas duomenų rinkinio $\mathbf{X}$ požymių skaičiaus n mažinimas, transformuojant $X_i \in \mathbb{R}^n$ į $Z_i \in \mathbb{R}^k$, čia $k < n$. Siekiant pažinti ir detaliau ištirti tą transformaciją, gautas duomenų rinkinys $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_m\} = \{z_{ij}, i = \overline{1,m}, j = \overline{1,k}\}$ vizualizuojamas į $\mathbb{R}^2$ erdvę. Akivaizdu, kad, jei klasterių skaičius $k > 2$, tai duomenys vizualizuojami į $\mathbb{R}^2$ erdvę projekcijos metodais. Savo tyrimuose taikome daugiamačių skalių projekcijos metodą. Siekiant dar giliau atskleisti eksponentinės (3.1) ar Gausinės (3.2) transformacijos savybes, vizualizuojamas ne tik duomenų rinkinys $\mathbf{Z}$, bet kartu ir klasterių centrai μ_j^z, $j = \overline{1,k}$ transformuoti naudojantis (3.1) ar (3.2) formulėmis. Vizualizuoti eksponentinių ir Gausinių radialinių bazinių funkcijų išėjimo rezultatai pateikti 4.9 ir 4.10 paveiksluose.



(a) Eksponentinė **(b)** Gausinė

4.9 pav. Daugiamačių skalių metodu gauta transformuotų Irisų duomenų rinkinio projekcija



(a) Eksponentinė **(b)** Gausinė

4.10 pav. Daugiamačių skalių metodu gauta Krūties vėžio duomenų rinkinio projekcija

Iš pateiktų paveikslų matome, kad skirtingomis funkcijomis (3.1) arba (3.2) atliktas daugiamačių duomenų požymių mažinimas duoda skirtingus vizualizavimo rezultatus. Pastebimi du skirtumai tarp vizualizavimo rezultatų naudojant eksponentines ir Gausines funkcijas:

1. Vizualizavimo būdas: eksponentinės funkcijos atveju vizualizuoti klasteriai yra labiau kampuoti (susidaro trikampio arba stačiakampio viršūnė, kurios pačiam kampe atsiduria klasterio centras), o Gausinės funkcijos atveju vizualizuoti klasteriai labiau užapvalinti (tarsi apskritimo arba elipsės kraštinė).

2. Centrų išsidėstymas klasteriuose: eksponentinės funkcijos atveju klasterių centrai yra išstumiami į klasterio šoną ir jie įgyja išskirtinę savybę būti tokiais taškais, kur keičiasi klasterių objektų ypatybės, ir tai matosi vizualiai, o Gausinės funkcijos atveju klasterių centrai lieka klasterių viduje.

Vizualizuotuose radialinių bazinių funkcijų išėjimo rezultatuose klasteriai išsidėsto dvejopai (žr. 4.9 ir 4.10 paveikslus):

1. *Izoliuotas klasteris.* Vizualiai matome, kad klasterio taškai sudaro atskirą grupę. Pavyzdžiui, Irisų duomenyse toks yra klasteris pažymėtas mėlynai. Jo taškai koncentruojasi aiškiai matomame atskirame klasteryje.

2. *Tarpusavyje artimi klasteriai* (panašūs klasteriai). Vizualizuoti klasterio taškai išsibarsto dviejų tiesių arba kreivių aplinkoje, kurios susijungia ties klasterio centru. Taškų išsidėstymą tiesių (eksponentinės funkcijos atveju) arba kreivių (Gausinės funkcijos atveju) aplinkoje geriausiai atspindi Krūties vėžio duomenys. Taškai, turintys panašumo su kaimyninio klasterio taškais, vizualizuojami arti tiesės arba kreivės, jungiančios kaimyninių klasterių centrus.

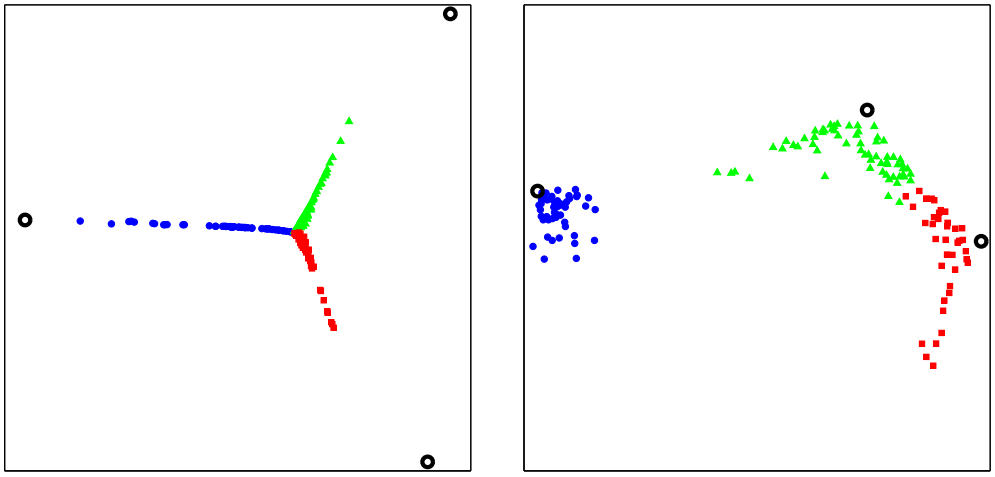
Atlikus daugiamačių duomenų transformaciją iš $X_i \in \mathbb{R}^n$ į $Z_i \in \mathbb{R}^k$ gauti vizualizavimo rezultatai kinta, kai yra keičiami radialinių bazinių funkcijų parametrai – centro taškai μ_j ir pločio parametras σ. Atliekant eksperimentus centrai buvo parenkami klasterizuojant duomenų rinkinį $\mathbf{X}$ k-vidurkių metodu. Siekiant gauti objektyvius rezultatus, atliktuose tyrimuose klasterizavimas buvo vykdomas keletą kartų, nes klasterizavimo paklaidą nusakanti (2.9) funkcija yra daugiaekstremė ir dažnai randa tik lokalų, o ne globalų funkcijos minimumą. Skaičiavimuose naudojami klasterizavimo rezultatai su mažiausiu lokaliu paklaidos minimumu. Po klasterizavimo radialinių bazinių funkcijų centrai μ_j tam pačiam duomenų rinkiniui yra fiksuojami. Vienintelis kintantis parametras, nuo kurio priklauso ir tolesni rezultatai yra pločio parametras σ.

Kadangi taškai išsidėsto skirtingai vizualizuojant eksponentinės ir Gausinės radialinių bazinių funkcijų išėjimo rezultatus, tai šios dvi funkcijos bus aptartos atskirai.

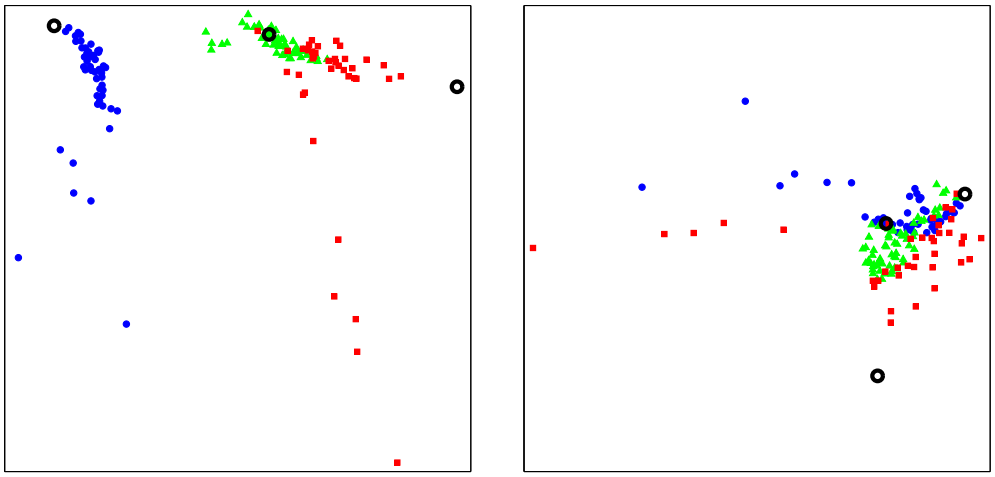
4.2.1. Eksponentinė funkcija

Eksperimentai buvo atlikti su keliais duomenų rinkiniais, tačiau vizualizavimo rezultatų pavyzdžiai parodyti tik su tais duomenų rinkiniais, kurie geriausiai atspindi pagrindinę esmę.

Eksperimentas, kuris parodo, kas vyksta vizualizavimo metu, kai kinta pločio parametras σ iliustruotas Irisų ir Širdies ligų duomenų rinkiniais 4.11 ir 4.12 paveiksluose.



(a) $\sigma = 0,4$ **(b)** $\sigma = 4$

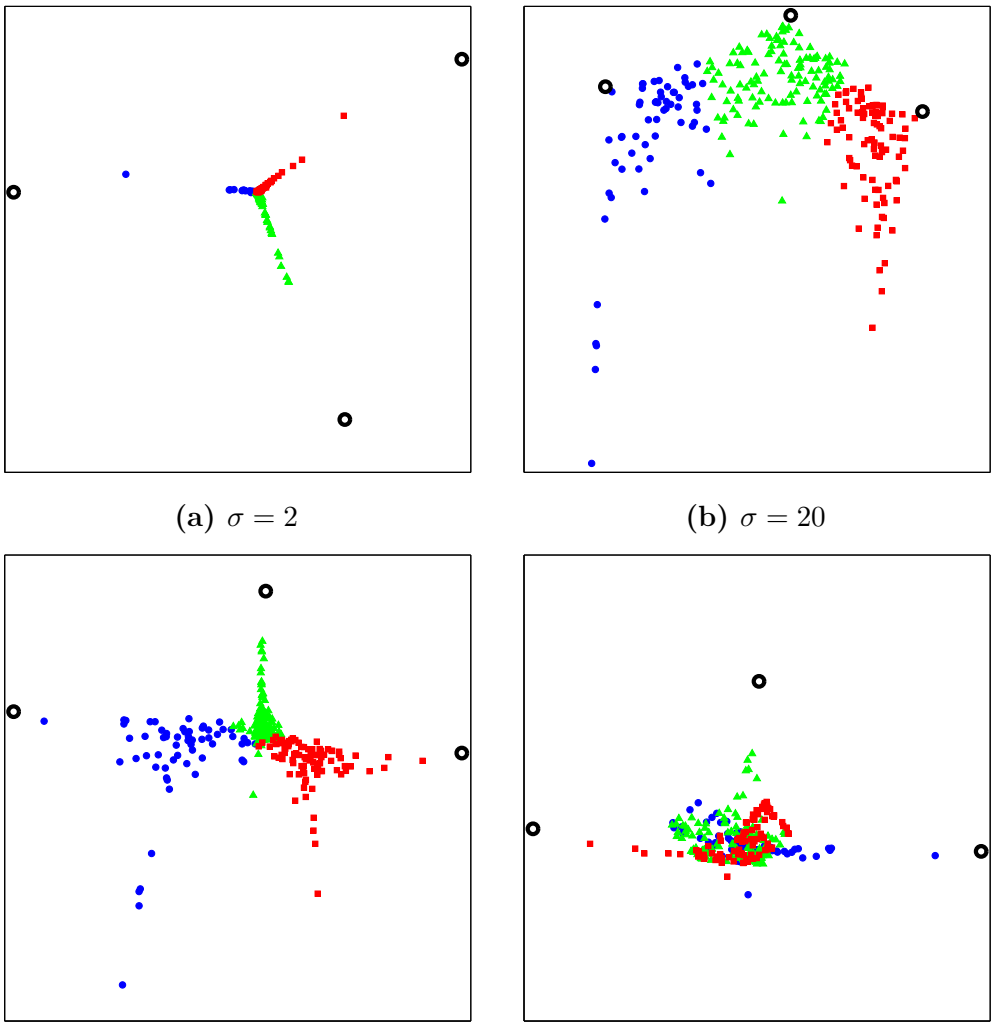


(c) $\sigma = 40$ **(d)** $\sigma = 100$

4.11 pav. Vizualizuota transformuoto Irisų duomenų rinkinio projekcija į plokštumą su skirtingomis pločio parametro reikšmėmis σ

Abiem duomenų rinkiniams pasirinktas klasterių skaičius $k = 3$, todėl duomenys į $\mathbb{R}^2$ erdvę vizualizuojami pasinaudojus projekcijos metodais. Šiame eksperimente naudotas daugiamačių skalių metodas. Irisų duomenų rinkiniui pločio parametras σ buvo parenkamas: a) $\sigma = 0,4$; b) $\sigma = 4$; c) $\sigma = 40$; d) $\sigma = 100$, o Širdies ligų duomenų rinkiniui – a) $\sigma = 2$; b) $\sigma = 20$; c) $\sigma = 200$; d) $\sigma = 2000$. Pastebėsime, kad parinkus labai mažą pločio parametrą σ, po transformacijos gaunami rezultatai artėja į 0, o jei σ parenkamas labai didelis, tai – artėja į 1. Jei konkrečiam duomenų rinkiniui parenkamas per mažas pločio parametras σ, tai atlikus duomenų rinkinio transformaciją visos reikšmės gaunamos labai mažos (t. y. beveik lygios

nuliui). Todėl atlikus transformuotų duomenų projekciją į dvimatę erdvę matomas tik vienas taškas arba šių duomenų projekcija negalima. Dėl šios priežasties Širdies ligų duomenų rinkinio pločio parametras σ parinktas didesnis nei Irisų duomenų rinkinio. Eksperimento tikslas pamatyti, kaip kinta vizualizavimo rezultatas, keičiantis pločio parametrui σ.



(a) $\sigma = 2$ **(b)** $\sigma = 20$

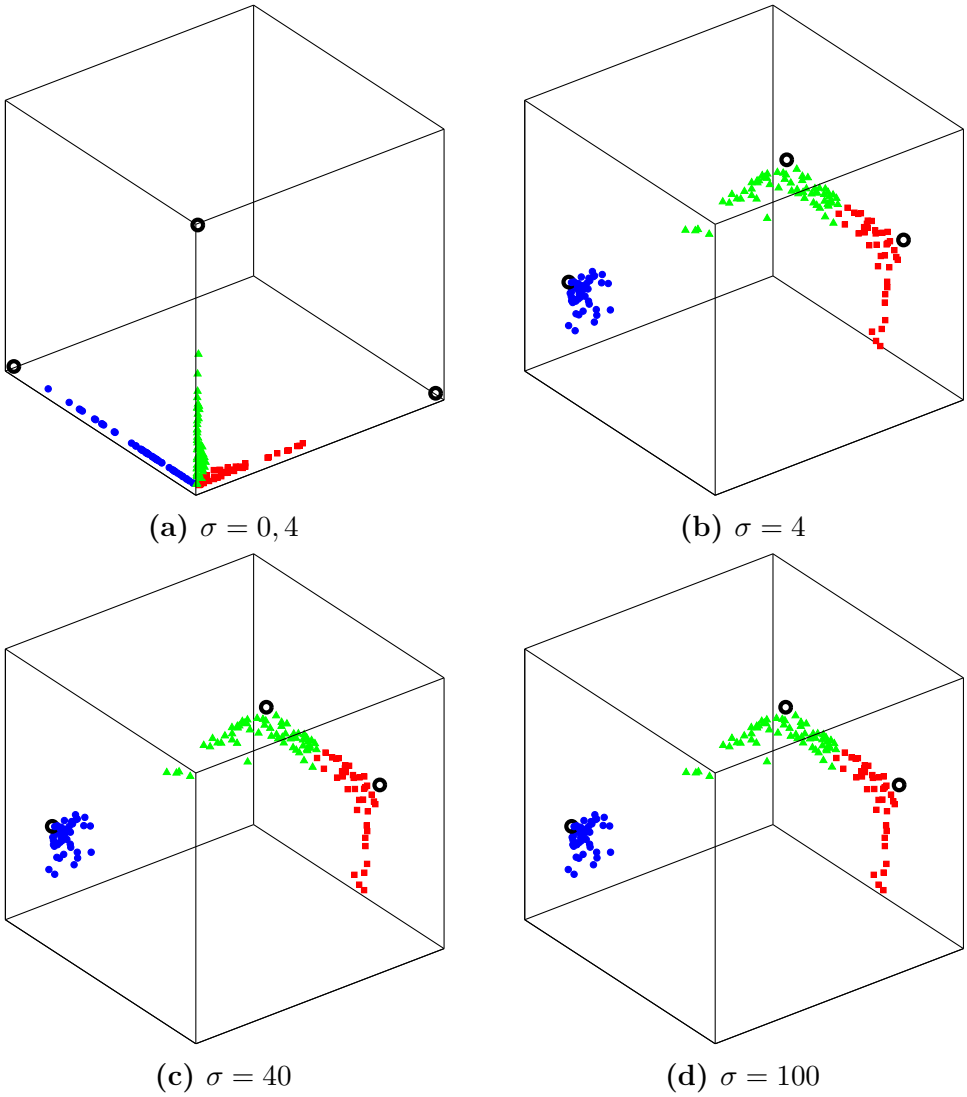
(c) $\sigma = 200$ **(d)** $\sigma = 2000$

4.12 pav. Vizualizuota transformuoto Širdies ligų duomenų rinkinio projekcija į plokštumą su skirtingomis pločio parametro reikšmėmis σ

Iš pateiktų paveikslų matome, kad duomenų rinkinių atvaizdavimas kinta keičiantis pločio parametrui σ. Apsibrėžkime pločio parametro σ įvertinimo kriterijus Eksponentinėje funkcijoje: *per maža pločio parametro σ reikšmė*, kai visų klasterių taškai sustumti į vieną visumą, o klasterių centrai yra išorėje (4.11a ir 4.12a paveikslai); *tinkama pločio parametro σ reikšmė* – duomenyse aiškiai išsiskiria klasteriai (klasterio taškai išsidėsto dviejų tiesių aplinkoje iš kurių susidaro trikampio arba stačiakampio viršūnė, kurios pačiam kampe atsiduria klasterio centras) (4.11b ir 4.12b paveikslai);

per didelė pločio parametro σ reikšmė – pastebimas gretimų klasterių persidengimas ir klasterio taškų „atsiplėšimas“ nuo klasterio centro (4.11c ir 4.12c paveikslai) arba klasteriai labai persidengia (4.11d ir 4.12d paveikslai).

4.11 ir 4.12 paveiksluose pateikta po transformacijos gautų rezultatų projekcija, kuri atlikta daugiamačių skalių metodu; buvo naudota $Stress-1$ kvadratinė paklaidos funkcija. Šiame eksperimente pasirinktas klasterių skaičius $k = 3$. Todėl po transformacijos gautų taškų $Z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$, $i = \overline{1, m}$ išsidėstymą erdvėje galime peržiūrėti ir trimatėje erdvėje, papildomai neatliekant duomenų projekcijos. Trimatėje erdvėje vizualizuoti Irisų ir Širdies ligų duomenų rinkiniai pateikti 4.13 ir 4.14 paveiksluose.

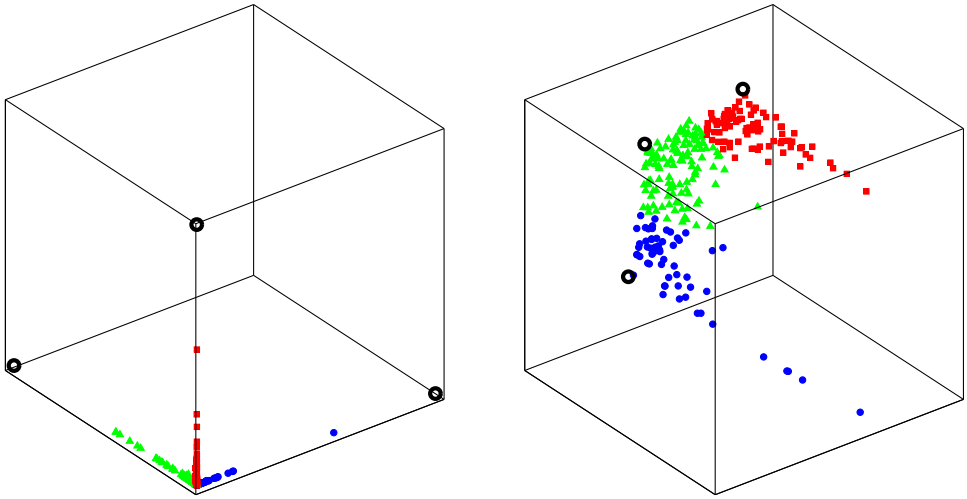


(a) $\sigma = 0,4$ **(b)** $\sigma = 4$

(c) $\sigma = 40$ **(d)** $\sigma = 100$

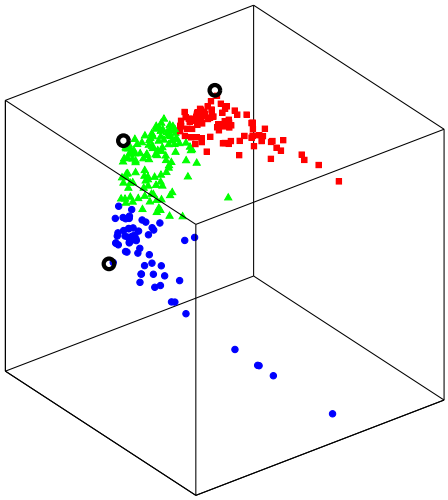
4.13 pav. Vizualizuota transformuoto Irisų duomenų rinkinio projekcija į plokštumą su skirtingomis pločio parametro reikšmėmis σ trimatėje erdvėje

Pagal gautą daugiamačių taškų Z_i išsidėstymą trimatėje erdvėje, galime teigti, kad pločio parametras σ gali būti mažas (4.13a ir 4.14a paveikslai)

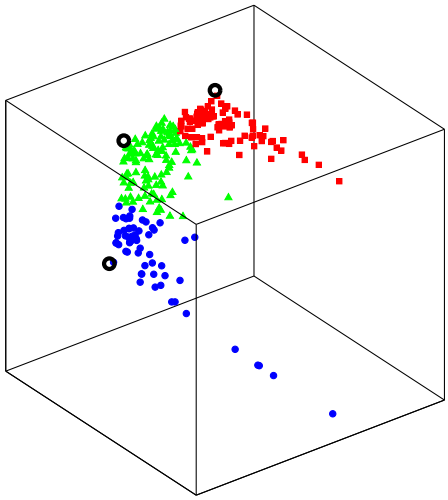


(a) $\sigma = 2$

(b) $\sigma = 20$



(c) $\sigma = 200$



(d) $\sigma = 2000$

4.14 pav. Vizualizuota transformuoto Širdies ligų duomenų rinkinio projekcija į plokštumą su skirtingomis pločio parametro reikšmėmis σ trimatėje erdvėje

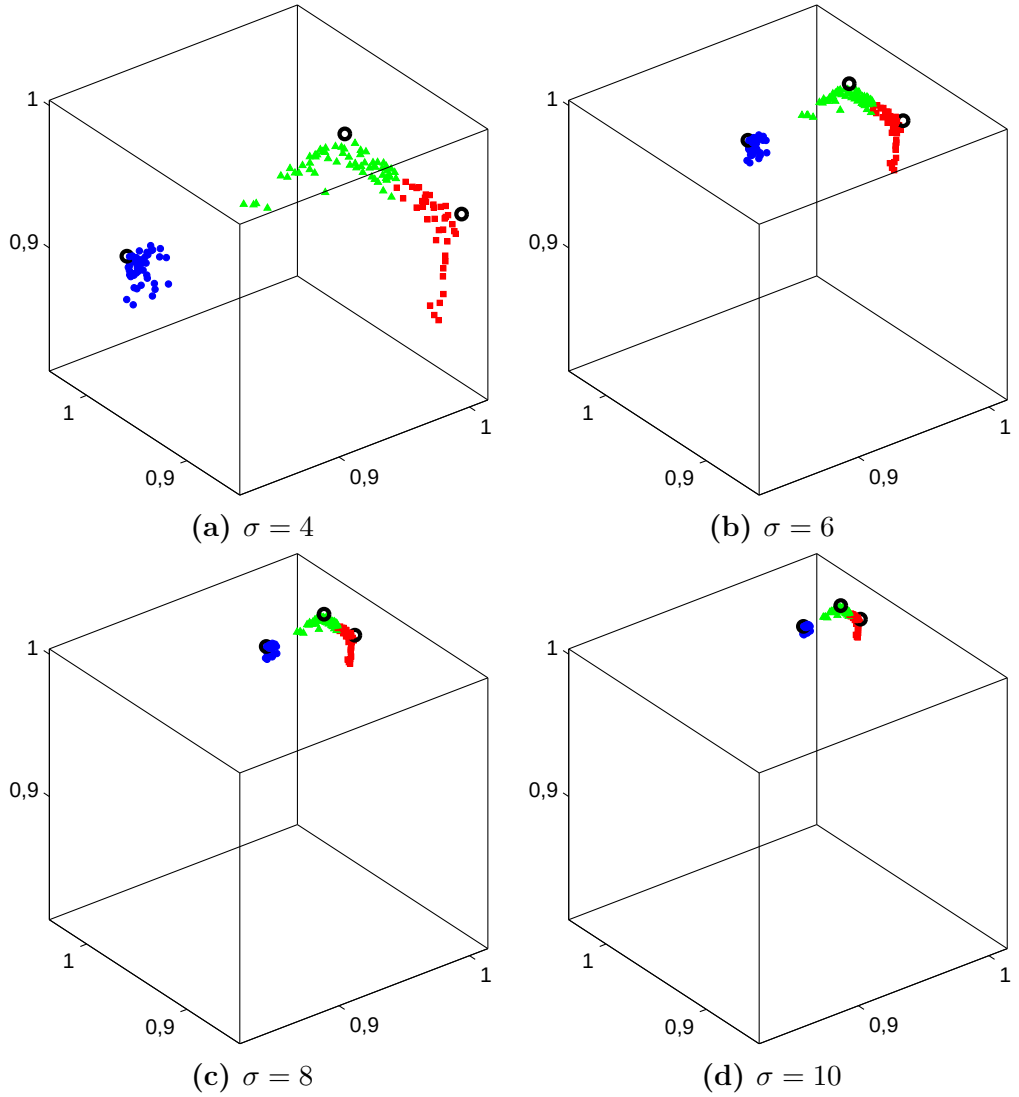
arba tinkamas (4.13b, 4.14b, 4.13c, 4.14c, 4.13d ir 4.14d paveikslai), duomenyse aiškiai išsiskiria klasteriai ir klasteriuose esantys taškai nepersidengia su gretimo klasterio taškais. Natūraliai kyla klausimai:

1. Kodėl daugiamačių duomenų projekcijoje į dvimatę erdvę MDS metodu atsiranda klasterių persidengimas (4.11c, 4.12c, 4.11d ir 4.12d paveikslai)?

2. Kuris pločio parametras σ yra geresnis iš tinkamų (4.13b, 4.14b, 4.13c, 4.14c, 4.13d ir 4.14d paveikslai)?

Pateiktuose paveiksluose nėra skalių žymėjimo, nes aktualus tik taškų tarpusavio išsidėstymas. Kiekvienam paveiksle skalės yra suvienodintos

pagal ilgiausią ašį (t. y. žiūrima kurioje ašyje taškai yra labiausiai išsibarstę ir pagal ją sulyginamos likusios ašys). Buvo atliktas toks eksperimentas. Paimtas Irisų duomenų rinkinio trimatėje erdvėje gautas vizualizavimo rezultatas, kai pločio parametras $\sigma = 4$ ir fiksuoti ašių ilgiai. Toliau atsitiktinai buvo parinkti pločio parametrai σ: $\sigma = 6$; $\sigma = 8$; $\sigma = 10$. Visiems gautiems vaizdams ašių ilgiai buvo nurodyti tokie patys, kaip pradiniame paveiksle. Gauti vizualizavimo rezultatai pateikti 4.15 paveiksle. 4.15a paveiksle patikrinus taškų išsibarstymą pagal ašis, nustatyta, kad labiausiai taškai išsibarstę x ašyje. Todėl y ir z ašių ilgiai nustatyti pagal x ašį. Tokie patys ašių ilgiai pritaikyti ir likusiuose 4.15b, 4.15c ir 4.15d paveiksluose.



(a) $\sigma = 4$ **(b)** $\sigma = 6$

(c) $\sigma = 8$ **(d)** $\sigma = 10$

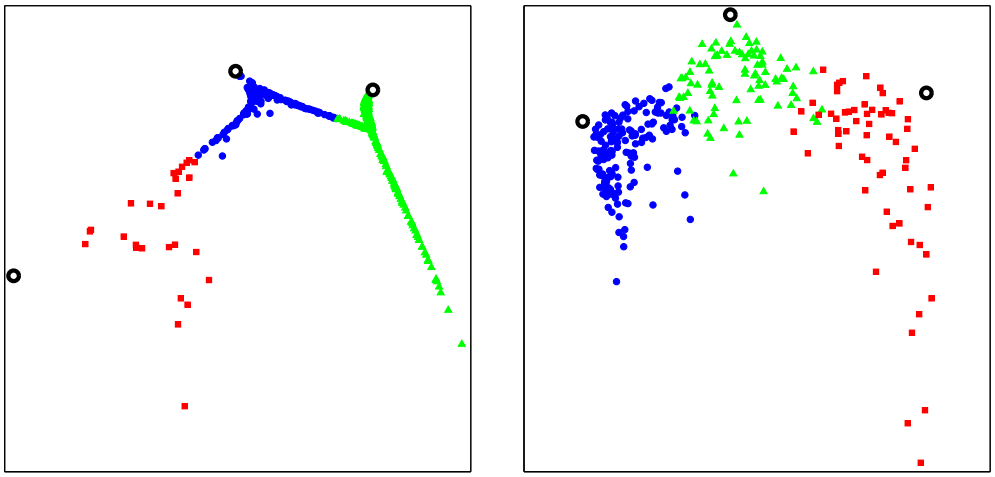
4.15 pav. Vizualizuota transformuoto Irisų duomenų rinkinio projekcija į plokštumą su skirtingomis pločio parametro reikšmėmis σ trimatėje erdvėje, kai skalės yra suvienodintos

Iš 4.15 paveikslo matome, kad didėjant pločio parametrui σ, mažėja taškų išsibarstymo plotas. Dėl šios priežasties projekcijos metodai projek-

tuojant daugiamačius duomenis Z_i į dvimatę erdvę iškraipo vaizdą, nes taškų Z_i išsibarstymo plotas tampa labai labai mažas, atstumai tarp jų supanašėja. Iš kitos pusės, tai yra privalumas. Laiku pastebėjus, kad po transformacijos gauti taškai Z_i užima mažai ploto, galima apsisaugoti nuo tolesniuose skaičiavimuose atsirasiančių netikslumų ir sumažinti atliekamų skaičiavimų apimtį. Taigi atlikus transformuotų duomenų Z_i projekciją į dvimatę erdvę MDS metodu ir peržiūrėjus gautus vizualizavimo rezultatus galime įvertinti transformuotų duomenų išsibarstymo plotą ir ar gautos reikšmės nėra labai mažos.

Atmetus mažas ir dideles pločio parametro σ reikšmes lieka dar nemažas intervalas, kuriame pločio parametras σ yra tinkamas. Tačiau norint toliau atlikinėti skaičiavimus su transformuotais duomenimis, reikia rasti geresnę pločio parametro σ reikšmę iš tinkamų.

Kaip jau minėta 2.3.6. poskyryje pločio parametras σ gali būti apskaičiuojamas pagal (2.37) arba (2.40) formules. Eksperimentas, kuris leidžia įvertinti pločio parametro σ reikšmės gerumą, iliustruotas Krūties vėžio ir Stuburo ligų duomenų rinkiniais 4.16 paveiksle. Abiem duomenų rinkiniams pasirinktas klasterių skaičius $k = 3$. Daugiamačių skalių metodu vizualizuotos transformuotų taškų Z_i projekcijos į dvimatę erdvę pateiktos 4.16 paveiksle. Projekcijos vizualiam pateikimui pasirinkta dvimatė erdvė, nes joje lengviau nustatyti ar pločio parametro σ reikšmė yra per didelė.



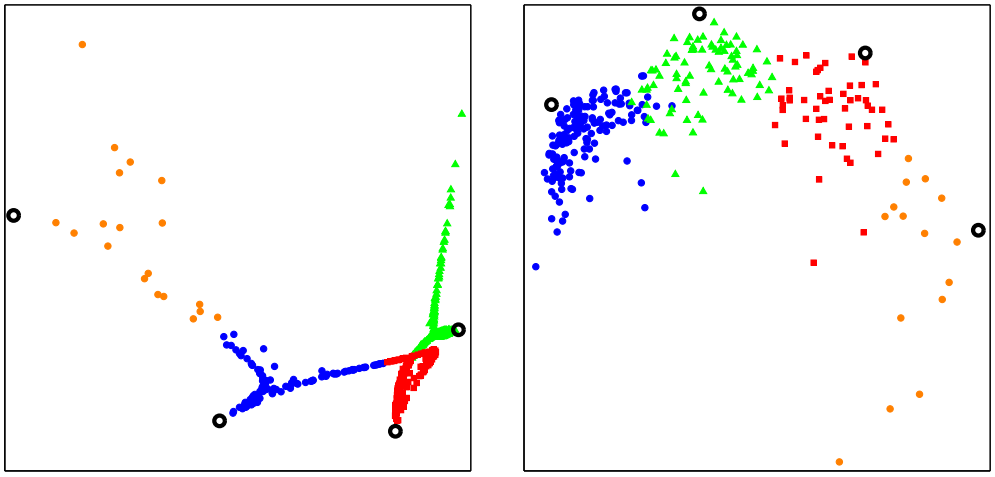
(a) Krūties vėžys, $\sigma = 918,57$ **(b)** Stuburo ligos, $\sigma = 36,69$

4.16 pav. Transformuoti duomenų rinkiniai suskirstyti į tris klasterius, kai pločio parametras σ apskaičiuojamas pagal (2.37) formulę

Peržvelgus gautus vizualizavimo rezultatus galime teigti, kad Krūties vėžio duomenų rinkiniui pločio parametro σ reikšmė apskaičiuota per didelė. Pastebimas akivaizdus klasterio taškų (pažymėtas ■) „atsiplėšimas“ nuo savo klasterio centro, bei klasterių centrai nebėra trikampio viršūnių

kampuose ● ir ▲ pažymėtuose klasteriuose. Stuburo ligų duomenims pločio parametras σ apskaičiuotas tinkamai. Duomenyse aiškiai išsiskiria klasteriai ir klasterių taškai yra prigludę prie savo klasterio centrų.

Kadangi mes nežinome, kiek yra iš tikrųjų daugiamačiuose duomenyse klasterių, tai pabandykime padidinti klasterių skaičių Krūties vėžio ir Stuburo ligų duomenų rinkiniams iki $k = 4$. Pločio parametrą σ apskaičiuosime pagal (2.37) formulę. Gauti vizualizavimo rezultatai pateikti 4.17 paveiksle.



(a) Krūties vėžys, $\sigma = 923,47$ **(b)** Stuburo ligos, $\sigma = 42,08$

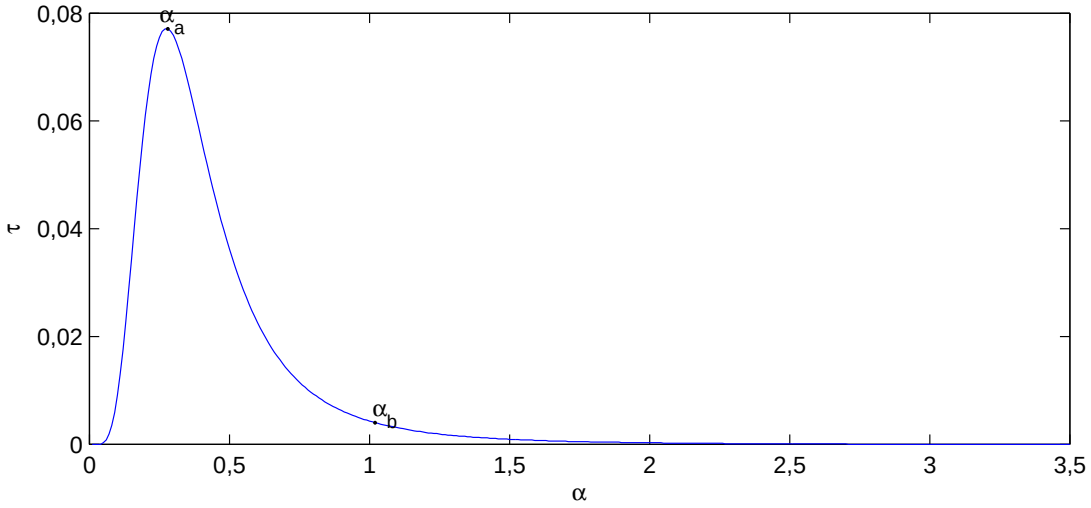
4.17 pav. Transformuoti duomenų rinkiniai suskirstyti į keturis klasterius, kai pločio parametras σ apskaičiuojamas pagal (2.37) formulę

4.17 paveiksle, kaip ir 4.16 paveiksle Krūties vėžio duomenų rinkiniui pločio parametras σ apskaičiuotas per didelis, o Stuburo ligų duomenų rinkiniui – tinkamas. Eksperimentas buvo atliekamas ir su kitais daugiamačių duomenų rinkiniais. Pagal visus gautus vizualizavimo rezultatus galima daryti išvadą, kad pločio parametro σ apskaičiavimas pagal (2.37) formulę tinkamas ne visiems duomenų rinkiniams. Daliai duomenų rinkinių pločio parametras σ apskaičiuojamas per didelis.

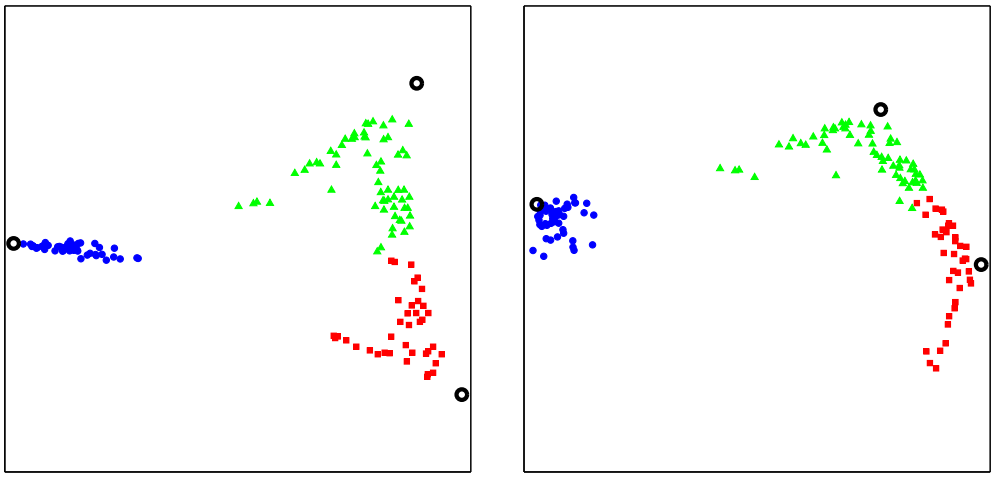
Toliau panagrinėsime pločio parametro σ apskaičiavimą pagal (2.40) formulę. Konstanta α parenkama iš nustatyto intervalo (kiekvienam duomenų rinkiniui intervalo rėžiai parenkami atskirai), tą intervalą prabėgant žingsniu 0,01 ir kiekvienoje iteracijoje apskaičiuojant τ reikšmę pagal (3.4) formulę. Kiekviena gauta τ reikšmė yra lyginama su prieš tai gauta τ reikšme. Kai skirtumas tarp τ reikšmių pasiekia užsibrėžtą tikslumą $\varepsilon = 0,0001$ ($0 < \tau^{u-1} - \tau^{u} \leq 0,0001$, čia u – iteracijos numeris), tai fiksuojama konstantos α reikšmė ir iteracinis procesas stabdomas. Ta vieta, kur τ reikšmių skirtumas pasiekia užsibrėžtą tikslumą $\varepsilon = 0,0001$, manome

kad yra tinkamo pločio parametro σ intervalo pradinė reikšmė. Eksperimento rezultatai iliustruojami Irisų ir Parkinsono ligos duomenų rinkiniais. Abiem duomenų rinkiniams pasirinktas klasterių skaičius $k = 3$.

4.18 paveiksle pateikiama Irisų duomenų rinkinio τ reikšmės priklausomybė nuo konstantos α. 4.19 paveiksle pateiktos transformuotų taškų Z_i vizualios projekcijos maksimaliame τ reikšmės taške, bei taške, kuriame fiksuojama α reikšmė. Irisų duomenų rinkinio konstantos α reikšmė buvo renkama iš intervalo [0,01; 3,5]. Intervalo rėžiai parenkami pagal τ reikšmės kitimo grafiką.



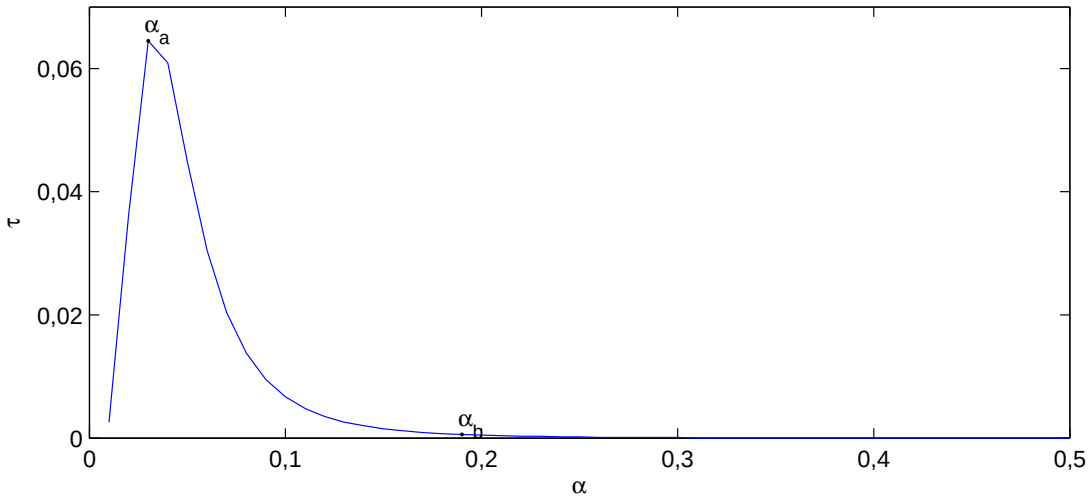
4.18 pav. τ reikšmės priklausomybė nuo konstantos α Irisų duomenų rinkiniui



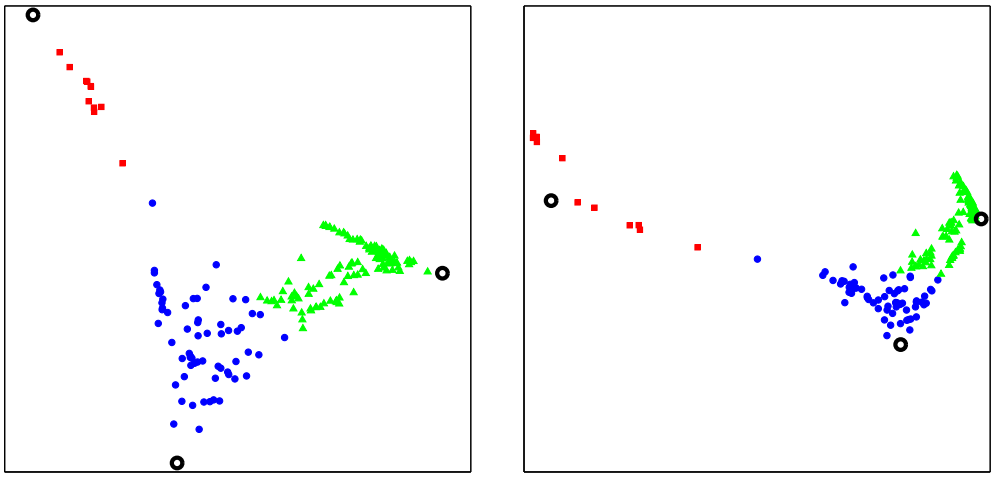
(a) $\sigma_B = 0,94$, $\alpha = \alpha_a = 0,28$ **(b)** $\sigma_B = 3,45$, $\alpha = \alpha_b = 1,02$

4.19 pav. Vizualios Irisų duomenų rinkinio projekcijos 4.18 grafiko α_a ir α_b taškuose

Visiems daugiamačių duomenų rinkiniams τ reikšmės kitimo grafikas vizualiai gaunamas labai panašus – turi matytis maksimalus τ reikšmės taškas ir τ reikšmės artėjimas prie nulio. Tam, kad įsitikinti τ reikšmės grafikų panašumu, 4.20 paveiksle papildomai pavaizduota τ reikšmės priklausomybė nuo konstantos α Parkinsono ligos duomenų rinkiniui. Parkinsono ligos duomenų rinkiniui konstantos α reikšmė buvo renkama iš intervalo [0,01; 0,5]. 4.21 paveiksle pateiktos transformuotų taškų Z_i vizualios projekcijos išskirtiniuose taškuose (maksimaliame τ reikšmės taške, bei taške, kuriame fiksuojama konstantos α reikšmė).



4.20 pav. τ reikšmės priklausomybė nuo konstantos α Parkinsono ligos duomenų rinkiniui



(a) $\sigma_B = 0,94$, $\alpha = \alpha_a = 0,28$ **(b)** $\sigma_B = 3,45$, $\alpha = \alpha_b = 1,02$

4.21 pav. Vizualios Parkinsono ligos duomenų rinkinio projekcijos 4.18 grafiko α_a ir α_b taškuose

Intervalo parinkimas individualus kiekvienam duomenų rinkiniui. Atsitiktinai parinkus intervalą stebima ar gautame grafike matosi maksimali τ reikšmė ir šios reikšmės artėjimas prie nulio. Pagal gautą grafiką intervalas koreguojamas jį praplečiant, susiaurinant arba pastumiant.

Pažiūrėję į 4.19 ir 4.21 paveiksluose pateiktus vizualizavimo rezultatus maksimumo taške α_a matome, kad pločio parametro σ reikšmė tuose taškuose dar per maža. Duomenų rinkiniuose dar stebimas klasterių taškų „judėjimas“ link klasterių centrų. τ reikšmės taškuose α_b, kuriuose fiksuojama α reikšmė, vizualizavimo rezultate klasterių taškai yra prigludę prie savo klasterio centro. Naudojantis (2.40) ir (3.4) formulėmis tinkami pločio parametrai σ buvo nustatyti visiems tyrimuose naudojamiems duomenų rinkiniams. Gauti rezultatai pateikiami 4.1 lentelėje. Iš 4.1 lentelėje pateiktų duomenų matome, kad pločio parametras σ priklauso nuo klasterių skaičiaus ir kiekvienam duomenų rinkiniui yra individualus.

4.1 lentelė: Eksponentinei funkcijai pagal (3.4) formulę rastos konstantos α ir su jomis gauti tinkami pločio parametrai σ

	2 klasteriai		3 klasteriai		4 klasteriai	
Duomenų rinkinys	α	σ	α	σ	α	σ
Irisai	0,9	3,53	1,02	3,45	1,02	2,88
Stuburo ligos	0,3	19,87	0,32	20,06	0,29	20,66
Krūties vėžys	0,1	132,6	0,09	135,02	0,1	145,61
Širdies ligos	0,26	20,82	0,29	23,41	0,3	22,23
Parkinsono liga	0,18	57,81	0,19	48,94	0,21	46,28
Vystantys medžiai	0,19	48,27	0,16	59,09	0,17	58,53
E.coli bakterijos	1,59	0,91	2,51	1,39	2,81	1,53
Kviečių grūdai	0,76	4,43	0,78	4,16	0,85	4,05

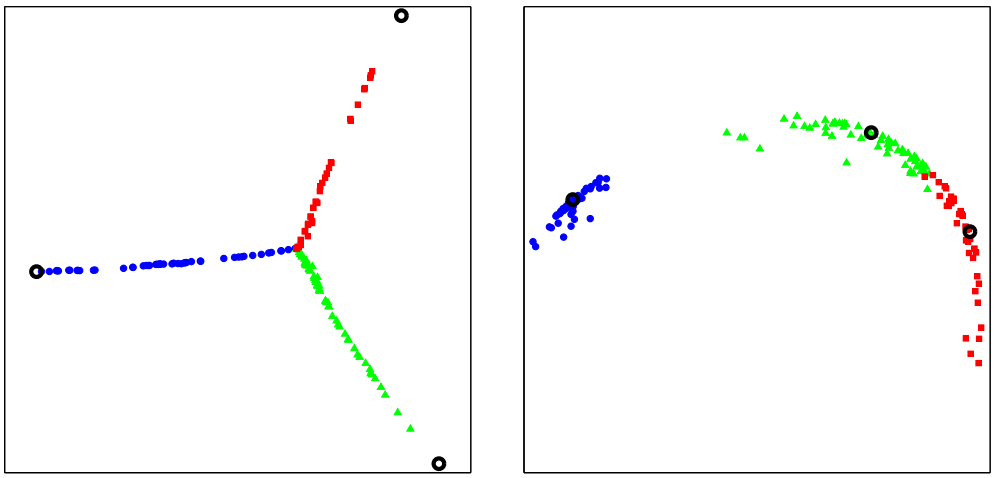
Apibendrinant su Eksponentine funkcija atliktus eksperimentus galima padaryti šias išvadas:

- Pločio parametro σ tinkamumą duomenų rinkiniui galime įvertinti pagal transformuotų taškų Z_i vizualią projekciją, atliktą MDS metodu.
- Pločio parametro σ nustatymas pagal maksimalų atstumą tarp klasterio centrų ir duomenų rinkinyje esančių klasterių skaičių k tinkamas ne visiems duomenų rinkiniams.
- Pasiūlytas konstantos α radimas, pagal taškų išsibarstymą kiekviename klasteryje, leidžia nustatyti tinkamą pločio parametro σ reikšmę. Su šia σ reikšme atlikus n-mačių taškų X_i dimensijos mažinimą, gauti taškai Z_i išsibarsto intervale [0; 1], t. y. nesikoncentruoja šio intervalo kraštuose.

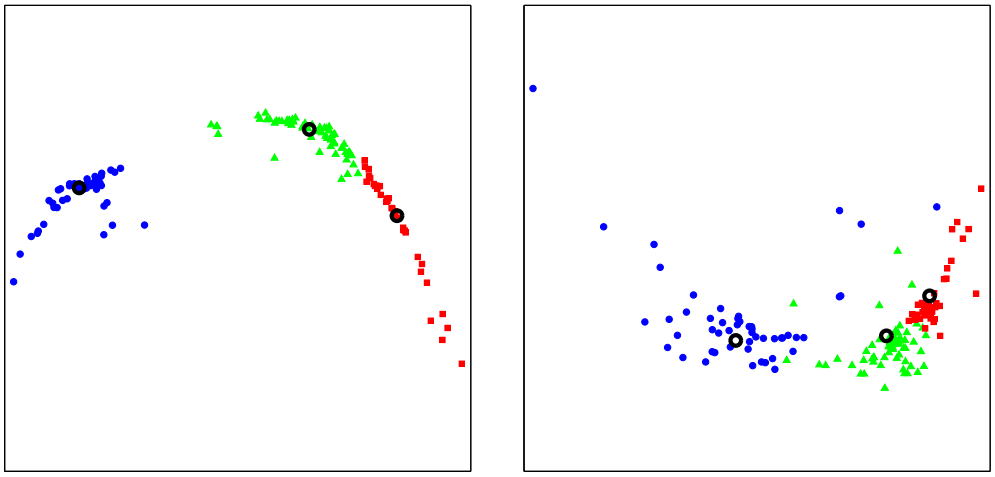
4.2.2. Gausinė funkcija

Prieš tai esančiame poskyryje buvo aptarti eksperimentai, kai duomenų rinkinio požymių mažinimas atliekamas su eksponentine bazine funkcija. Šioje dalyje bus pristatomi tie patys eksperimentai tik atlikti su Gausine bazine funkcija. Eksperimentai buvo atlikti su keliais duomenų rinkiniais, tačiau gauti rezultatai vizualiai parodyti tik su dviem duomenų rinkiniais.

Eksperimentas, parodantis vizualaus rezultato kitimą, kai keičiamas pločio parametras σ, iliustruotas Irisų ir Širdies ligų duomenų rinkiniais 4.22 ir 4.23 paveiksluose.



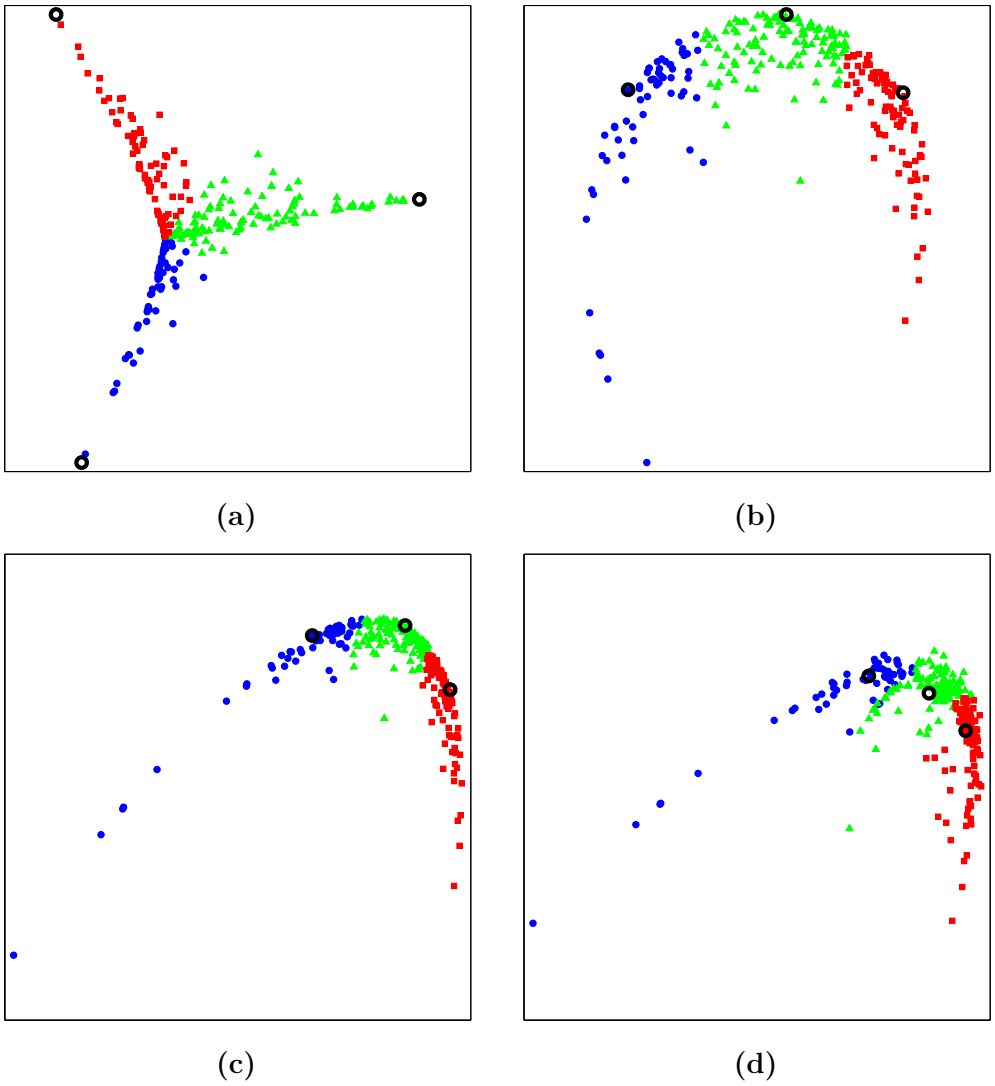
(a) $\sigma = 0,4$ **(b)** $\sigma = 4$



(c) $\sigma = 40$ **(d)** $\sigma = 100$

4.22 pav. Vizualizuota transformuoto Irisų duomenų rinkinio projekcija į plokštumą su skirtingomis pločio parametro reikšmėmis σ

Abiem duomenų rinkiniams pasirinktas klasterių skaičius $k = 3$. Irisų duomenų rinkiniui pločio parametras σ buvo parenkamas: a) $\sigma = 0,4$; b) $\sigma = 4$; c) $\sigma = 40$; d) $\sigma = 100$, o Širdies ligų duomenų rinkiniui –



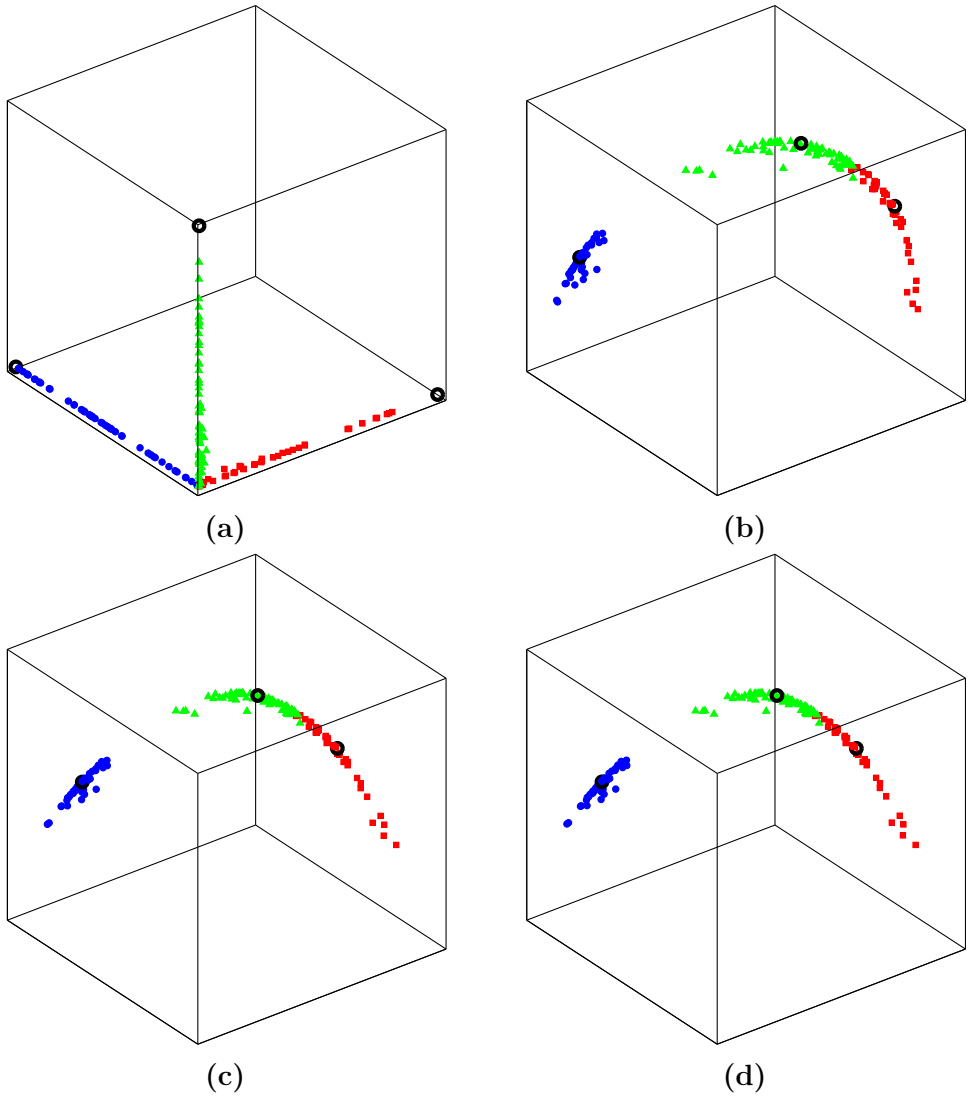
4.23 pav. Vizualizuota transformuoto Širdies ligų duomenų rinkinio projekcija į plokštumą su skirtingomis pločio parametro reikšmėmis σ

a) $\sigma = 20$; b) $\sigma = 100$; c) $\sigma = 500$; d) $\sigma = 2000$. Atliekant eksperimentus su Gausine radialine bazine funkcija Širdies ligų duomenų rinkiniui buvo parinktos didesnės pločio parametro σ reikšmės, nei eksperimentuose su eksponentine radialine bazine funkcija. Kintant pločio parametrui σ, po transformacijos gaunami rezultatai daug greičiau artėja į 0 arba į 1, nes Gausinės funkcijos atveju yra skaičiuojamas atstumo kvadratas.

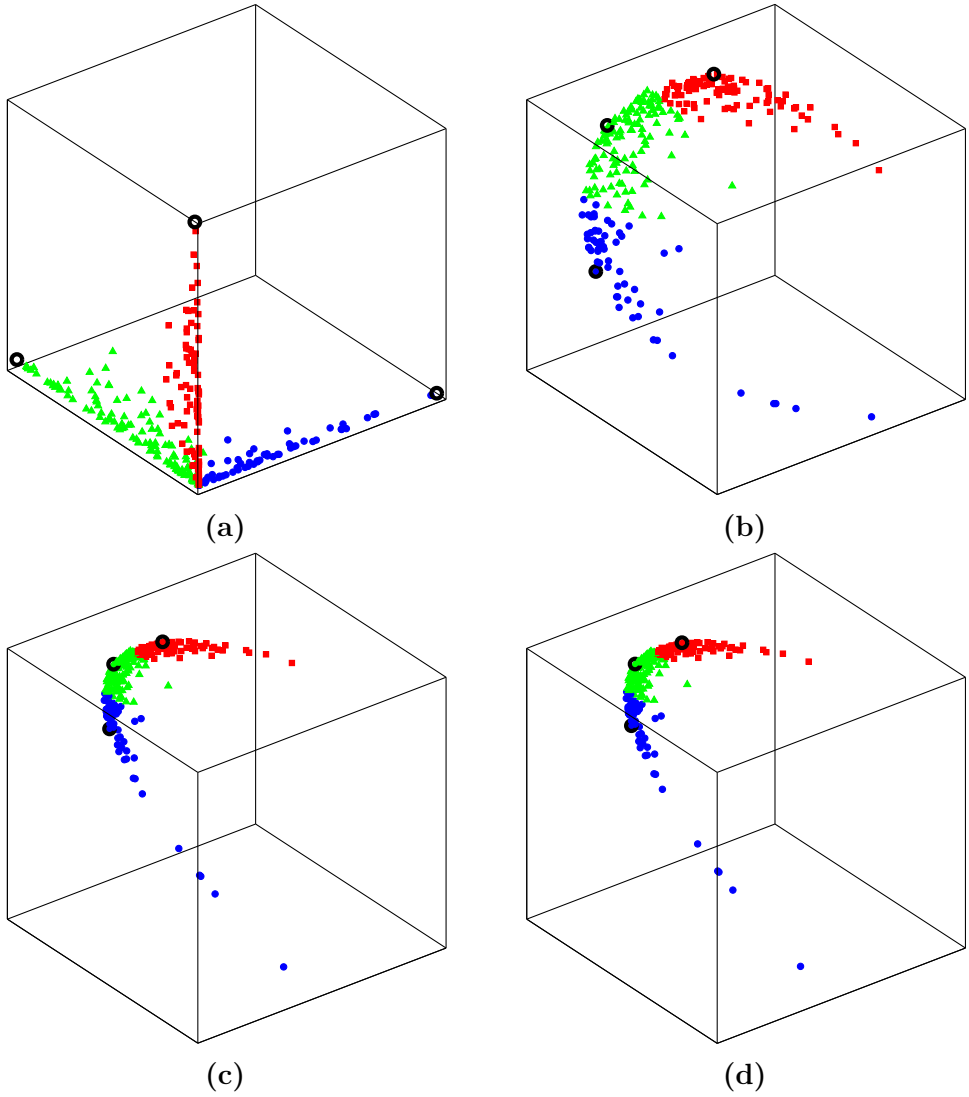
Iš 4.22 ir 4.23 paveikslų matome, kad gauti vizualizavimo rezultatai kinta keičiantis pločio parametrui σ. Apsibrėžkime pločio parametro σ įvertinimo kriterijus Gausinėje funkcijoje: *per maža pločio parametro σ reikšmė*, kai visų klasterių taškai sustumti į vieną visumą, o klasterių centrai yra išorėje (4.22a ir 4.23a paveikslai); *tinkama pločio parametro σ reikšmė* – visų klasterių taškai išdėstomi ant apskritimo arba elipsės kraštinių,

klasterių centrai yra klasterių viduje, bet prigludę prie išorinės klasterio ribos (4.22b ir 4.23b paveikslai); *per didelė pločio parametro σ reikšmė* – pastebimas didesnis klasterių taškų pasibarstymas ir viso duomenų rinkinio postūmis į kurią nors pusę (4.22c, 4.23c, 4.22d ir 4.23d paveikslai). Palyginus gautus vaizdus po duomenų rinkinio $\mathbf{X}$ transformacijos atliktos su eksponentine (3.1) (4.11 ir 4.12 paveikslai) ir Gausine (3.2) (4.22 ir 4.23 paveikslai) funkcijomis, galime teigti, kad eksponentinės funkcijos atveju per didelė pločio parametro σ reikšmė pastebima akivaizdžiau – vizualizavus po transformacijos gautus rezultatus atsiranda klasterių persidengimas.

Peržvelkime trimatėje erdvėje vizualizuotus Irisų ir Širdies ligų duomenų rinkinius, kurie pateikti 4.24 ir 4.25 paveiksluose.



4.24 pav. Vizualizuota transformuoto Irisų duomenų rinkinio projekcija į plokštumą su skirtingomis pločio parametro reikšmėmis σ trimatėje erdvėje

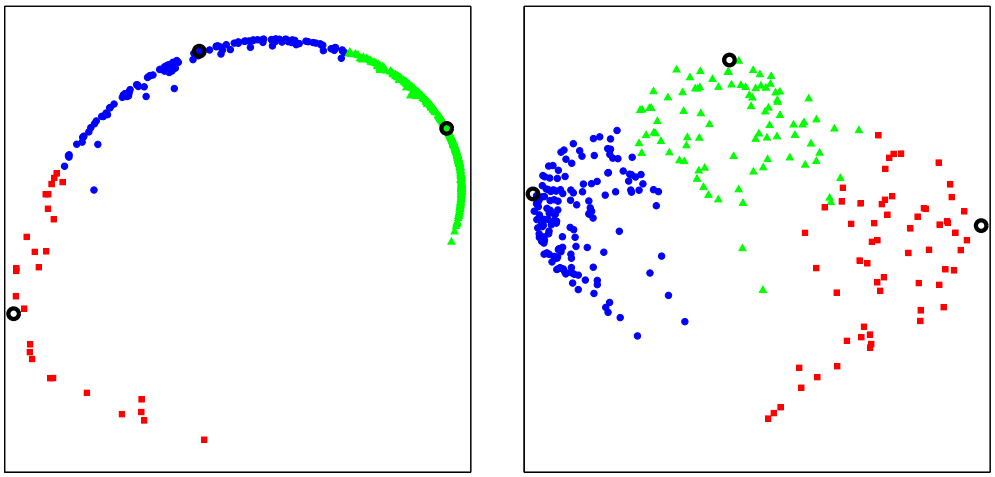


4.25 pav. Vizualizuota transformuoto Širdies ligų duomenų rinkinio projekcija į plokštumą su skirtingomis pločio parametro reikšmėmis σ trimatėje erdvėje

Pagal gautą transformuotų Irisų duomenų rinkinio $\mathbf{Z}$ vizualizavimą trimatėje erdvėje, galime teigti, kad pločio parametro σ reikšmė gali būti per maža (4.24a paveikslas) arba tinkama (4.24b, 4.24c ir 4.24d paveikslai). Tačiau vizualizuotame transformuotų Širdies ligų duomenų rinkinyje galima įžvelgti ir per dideles pločio parametro σ reikšmes (4.25c ir 4.25d paveikslai) – pastebimas taškų „suspaudimas“. Per didelę pločio parametro σ reikšmę trimačiame vaizde galime pastebėti tik tada, kai šalia turime vaizdą su tinkamu pločio parametru σ (4.25b paveikslas).

Eksperimentas, kuris leidžia įvertinti apskaičiuotos pagal (2.37) formulę pločio parametro σ reikšmės gerumą, iliustruotas Krūties vėžio ir Stuburo ligų duomenų rinkiniais 4.26 paveiksle. Abiem duomenų rinkiniams

pasirinktas klasterių skaičius $k = 3$. 4.26 paveiksle vizualizuotos transformuotų taškų Z_i projekcijos į dvimatę erdvę, naudojantis MDS metodu.



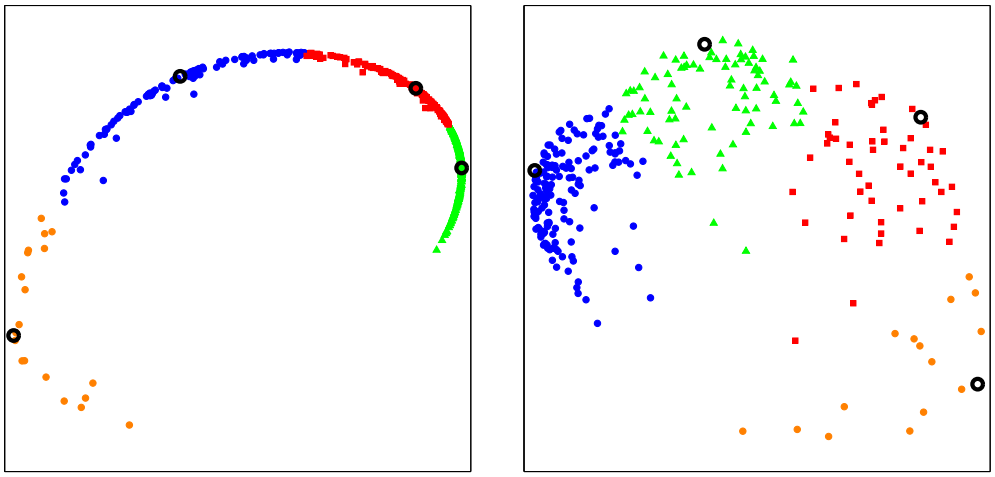
(a) Krūties vėžys, $\sigma = 923,47$ **(b)** Stuburo ligos, $\sigma = 42,08$

4.26 pav. Transformuoti duomenų rinkiniai suskirstyti į tris klasterius, kai pločio parametras σ apskaičiuojamas pagal (2.37) formulę

Peržvelgus gautus vizualizavimo rezultatus, galime teigti, kad Krūties vėžio duomenų rinkiniui pločio parametras σ apskaičiuotas tinkamas. Visų klasterių taškai išdėstomi ant elipsės kraštinių, klasterių centrai yra klasterių viduje. Stuburo ligų duomenų rinkiniui pločio parametras σ apskaičiuotas per mažas. Klasterių taškai dar tik artėja prie elipsės kraštinių, o klasterių centrai dar yra išskirtiniai. Pločio parametro σ apskaičiavimas pagal (2.37) formulę (σ priklauso nuo maksimalaus atstumo tarp klasterio centrų ir klasterių skaičiaus k) buvo pasiūlytas Gausinei, o ne eksponentinei funkcijai. Dėl šios priežasties su Gausine funkcija pločio parametras σ apskaičiuojamas tinkamesnis, nei su eksponentine funkcija.

Kaip ir eksponentinės funkcijos atveju, Krūties vėžio ir Stuburo ligų duomenų rinkiniams buvo pasirinktas kitas klasterių skaičius, $k = 4$. Pločio parametras σ apskaičiuojamas pagal (2.37) formulę. Vizualizuotos transformuotų duomenų rinkinių projekcijos pateiktos 4.27 paveiksle.

4.27 paveiksle, kaip ir 4.26 paveiksle pločio parametras σ Krūties vėžio duomenų rinkiniui apskaičiuojamas tinkamas, o Stuburo ligų duomenų rinkiniui – per mažas. Pagal visus gautus vizualizavimo rezultatus galima daryti išvadą, kad pločio parametro σ apskaičiavimas pagal (2.37) formulę tinkamas ne visiems duomenims. Daliai duomenų pločio parametras σ apskaičiuojamas per mažas.



(a) Krūties vėžys, $\sigma = 923,47$ **(b)** Stuburo ligos, $\sigma = 42,08$

4.27 pav. Transformuoti duomenų rinkiniai suskirstyti į keturis klasterius, kai pločio parametras σ apskaičiuojamas pagal (2.37) formulę

Paskutiniame eksperimente tinkamas pločio parametras σ apskaičiuojamas pagal (2.40) formulę, kur konstanta α apskaičiuojama pagal (3.4) formulę. Kaip ir eksponentinės funkcijos atveju, konstanta α parenkama iš nustatyto intervalo, tą intervalą prabėgant žingsniu 0,01. Gautos τ reikšmės priklausomybės nuo konstantos α formos yra labai panašios, kaip ir pateiktųjų 4.18 bei 4.20 paveiksluose. Kadangi gaunami grafikai vizualiai panašūs, tai pateikiamos tik po eksperimentų gautos konstantos α ir pločio parametro σ reikšmės 4.2 lentelėje.

4.2 lentelė: Gausinei funkcijai pagal (3.4) formulę rastos konstantos α ir su jomis gauti tinkami pločio parametrai σ

Duomenų rinkinys	2 klasteriai		3 klasteriai		4 klasteriai	
	α	σ	α	σ	α	σ
Irisai	1,64	6,44	1,95	6,61	1,99	5,63
Stuburo ligos	1,82	120,56	1,88	117,85	1,88	133,96
Krūties vėžys	1,78	2360,3	1,94	2787,4	2,08	3028,8
Širdies ligos	1,78	142,58	1,95	157,43	1,93	143,04
Parkinsono liga	1,68	539,64	1,88	484,25	1,98	436,43
Vystantys medžiai	1,82	462,37	1,89	698,00	1,95	671,44
E.coli bakterijos	1,63	0,93	1,69	0,93	1,71	0,93
Kviečių grūdai	1,67	9,74	1,89	10,10	1,93	9,21

Iš 4.2 lentelėje pateiktų duomenų matome, kad pločio parametras σ priklauso nuo klasterių skaičiaus k ir kiekvienam duomenų rinkiniui yra individualus. Jei palygintume eksperimentų rezultatus, atliktus su eksponentine (4.1 lentelė) ir Gausine (4.2 lentelė) funkcijomis, tai paste-

bėtume, kad tinkamos pločio parametro σ reikšmės Gausinei funkcijai yra kelis kartus didesnės, nei eksponentinei funkcijai.

Atlikus eksperimentus su Gausine funkcija išvados lieka tos pačios, kaip ir po eksponentinės funkcijos eksperimentų, tik dar galima pridurti, kad:

- Kintant pločio parametrui σ, po transformacijos, atliktos su Gausine funkcija, gaunami rezultatai daug greičiau artėja į 0 arba į 1.
- Pločio parametro σ tinkamumą pagal gautų rezultatų projekciją į dvimatę erdvę, lengviau įvertinti, kai daugiamačių duomenų požymių mažinimas atliekamas su eksponentine, o ne Gausine funkcija.
- Tinkama pločio parametro σ reikšmė Gausinei funkcijai yra kelis kartus didesnė, nei eksponentinei funkcijai.

Eksperimentuose atliktuose su eksponentine ir Gausine funkcija gauti pločio parametrai σ (4.1 ir 4.2 lentelės) bus naudojami tolesniuose tinklo REGM eksperimentuose.

4.3. REGM tinklas naudojamas eksperimentuose

Eksperimentai buvo atlikti su daugiamačių duomenų rinkiniais aprašytais 4.1. poskyryje.

Eksperimentai atlikti su tinklu REGM pateiktu 3.2 paveiksle. Pasirinktas klasterių skaičius $k = 3$, todėl radialinių bazinių funkcijų sluoksnyje Z yra trys radialinės bazinės funkcijos. Šiame sluoksnyje naudota eksponentinė radialinė bazinė funkcija. Eksperimente „Norimų tinklo atsako reikšmių parinkimas“ dar naudota ir Gausinė radialinė bazinė funkcija. Pirmojo paslėpto sluoksnio P^1 neuronų skaičius pasirinktas lygus penkiems. Mažajame sluoksnyje P^2 du neuronai, nes po tinklo apmokymo duomenis norima vizualizuoti plokštumoje. Išėjimų sluoksnio Y neuronų skaičius s lygus pasirinktam k klasterių skaičiui daugiamačiuose duomenyse, t. y. $s = k = 3$. Tik eksperimente „Neuronų skaičius išėjimo sluoksnyje“ neuronų skaičius parenkamas nuo vieno iki k. Mažajame sluoksnyje naudota tiesinė aktyvavimo funkcija. Pirmame paslėptame sluoksnyje ir išėjimo sluoksnyje – loginio sigmoido aktyvavimo funkcija. Atliekant eksperimentą „Antrosios REGM tinklo dalies aktyvavimo funkcijos“, mažajame ir išėjimo sluoksniuose naudotos loginio sigmoido arba tiesinės aktyvavimo funkcijos.

4.4. Norimų tinklo atsako reikšmių parinkimas

Atliekant eksperimentus su tinklu REGM labai svarbu tinkamai parinkti norimas tinklo atsako reikšmes, nes nuo jų priklauso apmokyto REGM

tinklo efektyvumas vizualizuojant daugiamačius duomenis. Kaip jau yra paminėta 3.3. poskyryje norimų tinklo atsako reikšmių $T_i = (t_{i1}, t_{i2}, \ldots, t_{is})$, $i = \overline{1,m}$ parinkimui siūlomos dvi strategijos:

1. Turimi klasterių centrai $\mu_j \in \mathbb{R}^n$ daugiamačių skalių metodu projektuojami į mažesnio matavimo erdvę $\mathbb{R}^s$, čia s – neuronų skaičius išėjimo sluoksnyje, $s < n$. Gauname klasterių centrų $\mu_j \in \mathbb{R}^n$ projekcijas $\mu_j^y \in \mathbb{R}^s$, $j = \overline{1,k}$. $T_i = \mu_j^y$, jei $X_i \in K_j$, $i = \overline{1,m}$. Pastebėsime, kad išėjimo sluoksnyje neuronų gali būti nuo 1 iki k (klasterių skaičiaus). Jei $s = k$, tai MDS metodu atliekant $\mu_j \in \mathbb{R}^n$ projekciją į $\mu_j^y \in \mathbb{R}^s$, $j = \overline{1,k}$, paskutinioji μ_j^y komponentė visada bus lygi 0.

2. Atliekamas turimų klasterių centrų μ_j požymių mažinimas eksponentine arba gausine funkcijomis iš $\mathbb{R}^n$ erdvės į $\mathbb{R}^k$ erdvę, čia k – klasterių skaičius, $k < n$. Jei $s < k$, tai transformuoti klasterių centrai μ_j^z, kad ir daugiamačių skalių metodu, projektuojami į $\mathbb{R}^s$ erdvę. Gauname klasterių centrų $\mu_j^z \in \mathbb{R}^k$ projekcijas $\mu_j^y \in \mathbb{R}^s$, $j = \overline{1,k}$. $T_i = \mu_j^y$, jei $X_i \in K_j$, $i = \overline{1,m}$. Pastebėsime, kad jeigu $s = k$, tai $T_i = \mu_j^z$, t. y. projektavimas iš $\mathbb{R}^k$ į $\mathbb{R}^s$ nėra reikalingas.

Paprastumo dėlei pirmąjį norimų tinklo atsako reikšmių pasirinkimo variantą pavadinkime *netransformuoti centrai*, o antrąjį – *transformuoti centrai*.

Atlikto eksperimento tikslas – nustatyti, kuria strategija (netransformuoti centrai arba transformuoti centrai) parinktos norimos tinklo atsako reikšmės leidžia REGM tinklą apmokyti kokybiškiau (idealiu atveju po apmokymo tinklas daro pakankamai mažą paklaidą; klasterių išsaugojimo duomenyse kriterijus $\chi = 0$; išėjimų sluoksnyje gautų reikšmių vaizde matoma tik tiek taškų, kiek duomenyse yra klasterių) ir mažajame sluoksnyje gautų reikšmių vaizdas atitinka užsibrėžtus vizualizavimo kokybės kriterijus, aprašytus 3.4. poskyryje.

Eksperimente tinklas buvo apmokytas 20 kartų tiek netransformuotų centrų atveju, tiek ir transformuotų centrų atveju. Abiem atvejais buvo naudojami tie patys 20 pradinių svorių rinkinių. Tai leidžia palyginti abi strategijas naudojant nedidelį kiekį ilgų skaičiavimų.

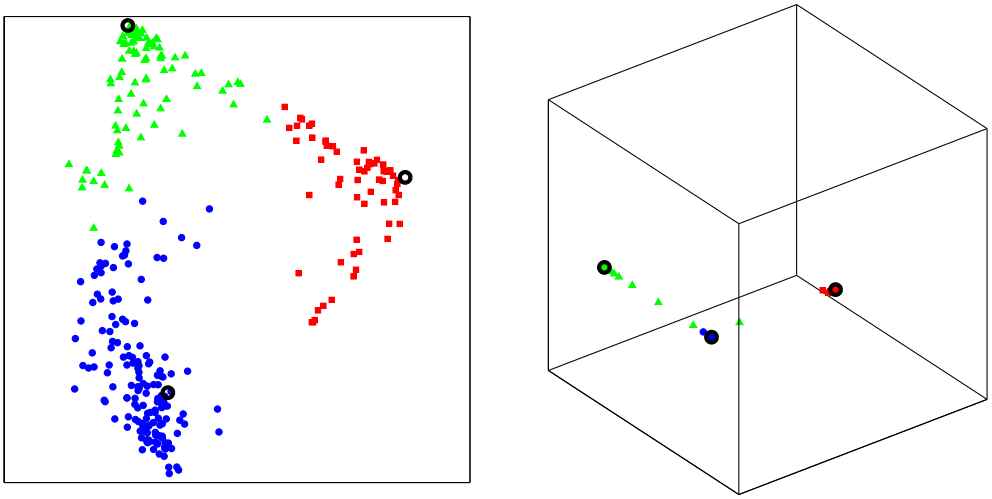
Eksperimento vykdymas, kai radialinių bazinių funkcijų sluoksnyje buvo naudota eksponentinė funkcija, iliustruotas Stuburo ligų duomenų rinkiniu 4.3 lentelėje. 4.3 lentelėje pateikti rezultatai surikiuoti pagal tinklo daromą paklaidą didėjimo tvarka.

4.3 lentelėje klasterių išsaugojimo duomenyse kriterijus χ (žiūrėti poskyrį 3.4.) parodo, kiek taškų po tinklo apmokymo išėjimo sluoksnyje

4.3 lentelė: Stuburo ligų duomenų rinkiniui norimų tinklo atsako reikšmių parinkimas, kai Z sluoksnyje naudojama eksponentinė funkcija

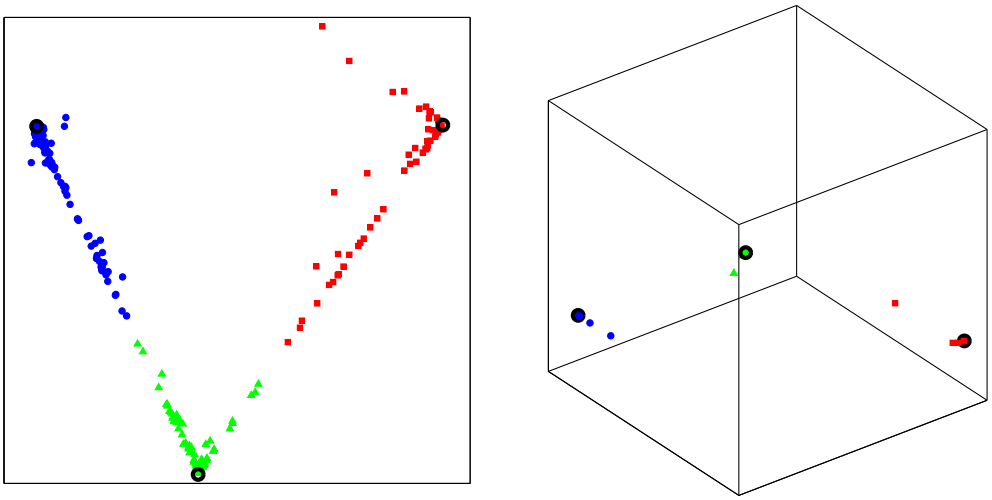
Numeris	Netransformuoti centrai			Transformuoti centrai		
	Paklaida	χ	κ	Paklaida	χ	κ
1	0,0450	65	0	0,000812	0	0,33
2	0,0452	7	0,12	0,000826	0	0,39
3	0,0454	19	0,12	0,000838	0	0,32
4	0,0455	160	0,99	**0,000853**	**0**	**0,52**
5	0,0472	160	1,00	0,000932	0	0,53
6	0,0472	150	0	0,000951	4	0,68
7	0,0472	102	0	0,000957	6	0,24
8	0,0477	164	0	0,001126	36	0
9	**0,0484**	**3**	**0,14**	0,001131	89	0,99
10	0,0484	49	0,002	0,001311	6	0,29
11	0,0495	160	0,99	0,001326	1	0,30
12	0,0496	140	0	0,001333	2	0,28
13	0,0502	166	0,97	0,001366	89	0,99
14	0,0612	221	0	0,001416	132	0
15	0,0620	221	0	0,001421	45	0
16	0,0621	221	0	0,001576	249	0
17	0,0669	221	0	0,001589	77	0,01
18	0,0699	221	0	0,001659	249	0
19	0,0718	202	0,001	0,001942	94	0,01
20	0,2001	221	0	0,002002	249	0

netenkino sąlygos $Y_i \in K_j^y$, kai $X_i \in K_j$, čia $i = \overline{1,m}$, $j = \overline{1,k}$ (t. y. bendras taškų Y_i skaičius per visus klasterius K_j^y, kur $Y_i \notin K_j^y$, kai $X_i \in K_j$). Kriterijaus κ reikšmė nurodo mažiausią atstumą tarp skirtingų klasterių taškų išėjimo sluoksnyje. Kuo taškai Y_i, $i = \overline{1,m}$ labiau prigludę prie savo klasterio K_j^y, $j = \overline{1,k}$ centro, tuo kriterijaus κ reikšmė didesnė. Kaip jau yra paminėta 3.4. poskyryje, apmokius tinklą tinkamiausias rezultatas (mažajame sluoksnyje gautų reikšmių vaizdas atitinka užsibrėžtus vizualizavimo kokybės kriterijus) yra tas, kuris tenkina abu atrankos kriterijus, t. y. klasterių išsaugojimo duomenyse kriterijaus reikšmė χ yra minimali (idealiu atveju $\chi = 0$) ir tarp minimalių χ reikšmių išėjimų sluoksnyje gautų rezultatų išsibarstymo kriterijaus reikšmė κ yra maksimali. Iš 4.3 lentelės matome, kad pirmojo eksfperimento metu, kai norimos tinklo atsako reikšmės yra netransformuoti centrai, tinkamiausias rezultatas gautas devintu tinklo apmokymo atveju, o antrojo eksperimento metu, kai norimos tinklo atsako reikšmės yra transformuoti centrai, – ketvirtu tinklo apmokymo atveju. Mažajame ir išėjimo sluoksniuose gauti vizualizavimo rezultatai pateikti 4.28 ir 4.29 paveiksluose. Mažojo sluoksnio įverčiai pagal antrąjį ir trečiąjį vizualizavimo kokybės kriterijus pateikti 4.4 ir 4.5 lentelėse.



(a) Mažasis sluoksnis **(b)** Išėjimų sluoksnis

4.28 pav. Z sluoksnyje naudojamos eksponentinės funkcijos, o norimos tinklo atsako reikšmės yra netransformuoti centrai, $E(W) = 0{,}0484$



(a) Mažasis sluoksnis **(b)** Išėjimų sluoksnis

4.29 pav. Z sluoksnyje naudojamos eksponentinės funkcijos, o norimos tinklo atsako reikšmės yra transformuoti centrai, $E(W) = 0{,}0009$

4.4 lentelė: 4.28 paveikslo antrojo vizualizavimo kokybės kriterijaus įverčiai

Paveikslas	● klasteris	▲ klasteris	■ klasteris
4.28a	0,61	0,46	0,48
4.29a	0,45	0,31	0,68

4.5 lentelė: 4.28 paveikslo trečiojo vizualizavimo kokybės kriterijaus įverčiai

Paveikslas	● ir ▲ klasteriai	▲ ir ■ klasteriai
4.28a	0,03	0,04
4.29a	0,07	0,11

Po apmokymo tinklas mažesnes paklaidas daro, kai norimos tinklo atsako reikšmės yra transformuoti centrai. Tinklo daromos paklaidos dydis mums suteikia informacijos apie bendrą tinklo apmokymą, t. y. tinklo išėjimo sluoksnyje gautų reikšmių Y_i, $i = \overline{1,m}$, artimumą norimoms tinklo atsako reikšmėms T_i, $i = \overline{1,m}$. Pastebėsime, kad šiame eksperimente REGM tinklo išėjimo sluoksnyje pasirinktas neuronų skaičius lygus pasirinktam klasterių skaičiui, $s = k = 3$. Todėl netransformuotų centrų atveju visų norimų tinklo atsako reikšmių T_i trečioji komponentė t_{i3} lygi 0. Pastebėsime, kad kaip jau yra minėta 3.4. poskyryje, ne visada tinklas, darantis mažiausią paklaidą $E(W)$, duoda geresnius vizualizavimo rezultatus. Šioje disertacijoje didžiausias dėmesys kreipiamas į n-mačių duomenų rinkinių projekcijos vizualizavimą plokštumoje, kuri gaunama vizualizuojant mažajame sluoksnyje gautus rezultatus. Todėl galutinę išvadą apie gaunamos projekcijos kokybę (atitikimą užsibrėžtiems vizualizavimo kokybės kriterijams) priimsime tik aptarę visus kriterijus.

Kaip matome iš 4.3 lentelės, pirmame eksperimente, kai norimos tinklo atsako reikšmės yra netransformuoti centrai, mažiausia χ reikšmė yra 3 ir tokia reikšmė yra vienintelė. Visos kitos χ reikšmės yra daug didesnės. Todėl jei tokiam tinklui paduotume mokyme nedalyvavusį n-matį tašką X, tai negalėtume būti tikri, ar rezultate gauta to taško projekcija tikrai atspindės nurodyto klasterio savybes. Kadangi mažiausia χ reikšmė yra vienintelė, tai geriausias tinklo apmokymo rezultatas atrinktas tik pagal pirmąjį atrankos kriterijų. Pažvelkime į gautų reikšmių vaizdus po devintojo apmokymo 4.28a paveiksle. Išėjimų sluoksnyje gautame vizualizavimo rezultate matome klasterių taškų pasibarstymą. Mažajame sluoksnyje klasterių taškai išsidėstę „debesėliuose“, o ne tiesių ar kreivių aplinkoje. Pagal 4.4 ir 4.5 lentelėse pateiktus įverčius matome, kad gautas vaizdas atitinka tik antrąjį vizualizavimo kokybės kriterijų.

Iš 4.3 lentelėje pateiktų rezultatų, kai tinklo atsako reikšmės yra transformuoti centrai, matome, kad pirmasis atrankos kriterijus χ net penkiais tinklo apmokymo atvejais lygus 0, t. y. tinklas apmokytas idealiai, nes tenkinama sąlyga: $Y_i \in K_j^y$, kai $X_i \in K_j$. Tinkamiausias galutinis rezultatas buvo atrinktas pagal antrąjį atrankos kriterijų, t. y. kur κ reikšmė maksimali. Pažvelkime į gautus vizualizavimo rezultatus po ketvirtojo apmokymo 4.29a paveiksle. Išėjimų sluoksnyje matomi tik trys taškai, nes duomenų rinkinyje yra tik 3 klasteriai. Mažajame sluoksnyje gautas vaizdas tenkina visus tris užsibrėžtus vizualizavimo kokybės kriterijus.

Apibendrinus abiejų eksperimentų rezultatus galime teigti, kad tinklo REGM mokyme kaip norimas tinklo atsako reikšmes tikslingiau imti transformuotus centrus, nes:

- labiau tikėtina, kad REGM tinklas bus apmokytas idealiai, t. y. $\chi = 0$;
- mažajame sluoksnyje gauti vizualizavimo rezultatai labiau atitinka užsibrėžtus vizualizavimo kokybės kriterijus.

Pakomentuosime, kokių žinių mums suteikia mažajame sluoksnyje gautas vizualizavimo rezultatas. Kaip jau yra minėta, REGM tinklas buvo apmokytas Stuburo ligų duomenų rinkiniu. Duomenų rinkinyje išskirti trys objektų klasteriai: sveikų pacientų klasteris pažymėtas ●, pacientų, turinčių stuburo disko išvaržą, klasteris pažymėtas ▲ ir pacientų, sergančių spondilolisteze, klasteris pažymėtas ■. Iš 4.29a paveikslo matome, kad nuo ● ir ▲ pažymėtų klasterių yra atsiskyrusios taškų grupės, kurios turi tarpusavyje panašumo, nors priklauso skirtingiems klasteriams. Šiose grupėse esantys objektai, tyrėjui gali padėti atkreipti dėmesį į galimus pakitimus (ankstyvą ligos stadiją) arba ieškoti priežasčių, dėl kurių atsiranda pakitimai. Taip pat iš 4.29a paveikslo stebimi ir tarp pačių stuburo ligų esantys panašumai.

4.6 lentelėje pateikiami su visais disertacijoje naudojamais daugiamačių duomenų rinkiniais atliktų eksperimentų rezultatai. Radialinių bazinių funkcijų sluoksnyje naudota eksponentinė funkcija.

4.6 lentelė: Visų duomenų rinkinių norimų tinklo atsako reikšmių parinkimas, kai Z sluoksnyje naudojama eksponentinė funkcija

Duomenų rinkinys	Netransformuoti centrai			Transformuoti centrai		
	Paklaida	χ	κ	Paklaida	χ	κ
Irisai	0,0254	0	0,34	0,002229	0	0,73
Stuburo ligos	0,0484	3	0,14	0,000853	0	0,52
Krūties vėžys	0,0039	1	0,04	0,000261	0	0,71
Širdies ligos	0,0194	2	0,02	0,000824	0	0,63
Parkinsono liga	0,0103	1	0,37	0,000710	0	0,57
Vystantys medžiai	0,0691	16	0,03	0,000301	0	0,25
E.coli bakterijos	0,0589	2	0,55	0,002967	0	0,71
Kviečių grūdai	0,0315	1	0,21	0,002833	0	0,80

Iš 4.6 lentelėje pateiktų rezultatų matome, kad eksperimente, kai norimos tinklo atsako reikšmės yra netransformuoti centrai, tik Irisų duomenų rinkinio atveju klasterių išsaugojimo duomenyse kriterijaus reikšmė χ yra lygi 0. Visais kitais atvejais $\chi > 0$. Eksperimente, kai norimos tinklo atsako reikšmės yra transformuoti centrai, visada klasterių išsaugojimo duomenyse kriterijaus reikšmė $\chi = 0$. Transformuotų centrų atveju antrojo kriterijaus reikšmė κ visiems duomenų rinkiniams yra didesnė nei netransformuotų centrų atveju, o tai parodo, kad REGM tinklas apmokytas kokybiškiau. Pastebėsime, kad objektai, kurie po tinklo apmokymo pakliūna į gretimą klasterį, antrojo atrankos kriterijaus reikšmės skaičiavimuose nenaudojami.

Pagal 4.6 lentelėje pateiktus duomenis galima teigti, kad norimas tinklo atsako reikšmes tikslingiau imti transformuotus centrus.

Pirmasis vizualizavimo kokybės kriterijus labiau tenkinamas transformuotų centrų atveju, nes taškai aiškiau išsidėsto tiesių ar kreivių aplinkoje. Visų duomenų rinkių gautų projekcijų kiekybiniai vizualizavimo kokybės kriterijai (antrasis ir trečiasis) pateikti 4.7 ir 4.8 lentelėse.

4.7 lentelė: Antrojo vizualizavimo kokybės kriterijaus įverčiai, kai Z sluoksnyje naudojama eksponentinė funkcija

Duomenų rinkinys	Netransformuoti centrai			Transformuoti centrai		
	●	▲	■	●	▲	■
Irisai	0,17	0,14	0,72	0,14	0,04	0,86
Stuburo ligos	0,61	0,46	0,48	0,45	0,31	0,68
Krūties vėžys	0,12	0,57	0,37	0,39	0,34	0,53
Širdies ligos	0,47	0,32	0,22	0,37	0,55	0,44
Parkinsono liga	0,003	0,40	0,54	0,58	0,20	0,19
Vystantys medžiai	0,43	0,58	0,69	0,61	0,35	0,69
E.coli bakterijos	0,29	0,57	0,51	0,24	0,36	0,62
Kviečių grūdai	0,38	0,49	0,25	0,40	0,37	0,38

4.8 lentelė: Trečiojo vizualizavimo kokybės kriterijaus įverčiai, kai Z sluoksnyje naudojama eksponentinė funkcija

Duomenų rinkinys	Netransformuoti centrai		Transformuoti centrai	
	● ir ▲	▲ ir ■	● ir ▲	▲ ir ■
Irisai	0,32	0,12	0,29	0,11
Stuburo ligos	0,03	0,04	0,07	0,11
Krūties vėžys	0,04	0,01	0,05	0,09
Širdies ligos	0,01	0,02	0,08	0,06
Parkinsono liga	0,01	0,08	0,13	0,16
Vystantys medžiai	0,002	0,04	0,002	0,05
E.coli bakterijos	0,04	0,04	0,08	0,19
Kviečių grūdai	0,09	0,13	0,12	0,18

Iš 4.7 lentelės matome, kad antrąjį vizualizavimo kokybės kriterijų ($\bar{a} > 0,1$) ir netransformuotų centrų ir transformuotų centrų atveju atitinka beveik visi duomenų rinkiniai. Netransformuotų centrų atveju antrojo vizualizavimo kokybės kriterijaus neatitinka Parkinsono ligos duomenų rinkinys, nes vieno klasterio didžiausias atstumas tarp klasterio taškų $\bar{a} = 0,003$. Transformuotų centrų atveju antrojo vizualizavimo kokybės kriterijaus neatitinka Irisų duomenų rinkinys, nes vieno klasterio didžiausias atstumas tarp klasterio taškų $\bar{a} = 0,004$.

Trečiasis vizualizavimo kokybės kriterijus yra pageidautinas, bet nebūtinas. Jo reikšmė $\hat{a}$ turi būti didesnė arba lygi 0,05. Iš 4.8 lentelės matome,

kad netransformuotų centrų atveju, šį kriterijų tenkina tik Irisų ir Kviečių grūdų duomenų rinkiniai. Transformuotų centrų atveju nepilnai tenkina trečiąjį vizualizavimo kokybės kriterijų tik Vystančių medžių duomenų rinkinys. Visiems kitiems duomenų rinkiniams trečiasis vizualizavimo kokybės kriterijus yra tenkinamas.

Pagal 4.7 ir 4.8 lentelėse pateiktus duomenis taip pat galima teigti, kad norimas tinklo atsako reikšmes tikslingiau imti transformuotus centrus.

Taigi, kai REGM tinklo radialinių bazinių funkcijų sluoksnyje naudojama eksponentinė funkcija, tai norimas tinklo atsako reikšmes tikslingiau imti transformuotus centrus.

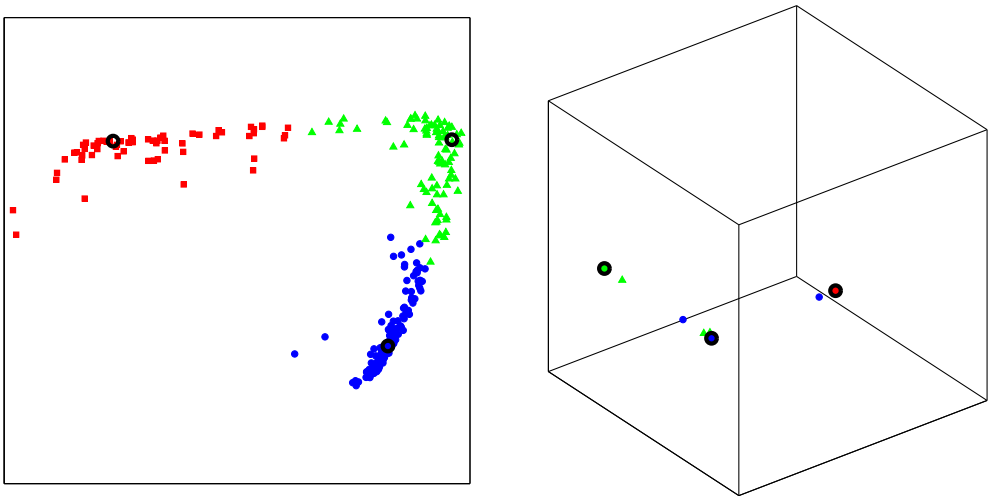
Pažiūrėkime, kokias norimas tinklo atsako reikšmes tikslingiau naudoti, kai radialinių bazinių funkcijų sluoksnyje naudojama Gausinė funkcija. Eksperimento vykdymas iliustruotas Stuburo ligų duomenų rinkiniu 4.9 lentelėje. 4.9 lentelėje pateikti rezultatai surikiuoti pagal tinklo daromą paklaidą didėjimo tvarka.

4.9 lentelė: Stuburo ligų duomenų rinkiniui norimų tinklo atsako reikšmių parinkimas, kai Z sluoksnyje naudojama Gausinė funkcija

Numeris	Netransformuoti centrai			Transformuoti centrai		
	Paklaida	χ	κ	Paklaida	χ	κ
1	0,0437	160	0,99	0,0039	2	0,18
2	0,0450	160	0,85	0,0039	4	0,11
3	**0,0452**	**3**	**0,12**	0,0041	2	0,21
4	0,0455	160	1	**0,0041**	**0**	**0,23**
5	0,0459	162	0,94	0,0041	2	0,17
6	0,0459	160	0,85	0,0041	2	0,19
7	0,0471	42	0	0,0041	2	0,28
8	0,0472	31	0	0,0042	4	0,08
9	0,0481	32	0,01	0,0043	1	0,07
10	0,0481	32	0	0,0043	1	0,08
11	0,0482	161	0,99	0,0043	8	0,06
12	0,0486	160	0,93	0,0043	23	0,03
13	0,0491	26	0	0,0045	9	0,13
14	0,0504	163	0,98	0,0052	4	0,06
15	0,0510	160	0,89	0,0054	56	0
16	0,0520	30	0	0,0072	15	0,10
17	0,0540	190	1	0,0073	2	0,11
18	0,0540	164	0,91	0,0074	8	0,08
19	0,0549	73	0	0,0077	221	0
20	0,0626	93	0	0,0081	39	0,03

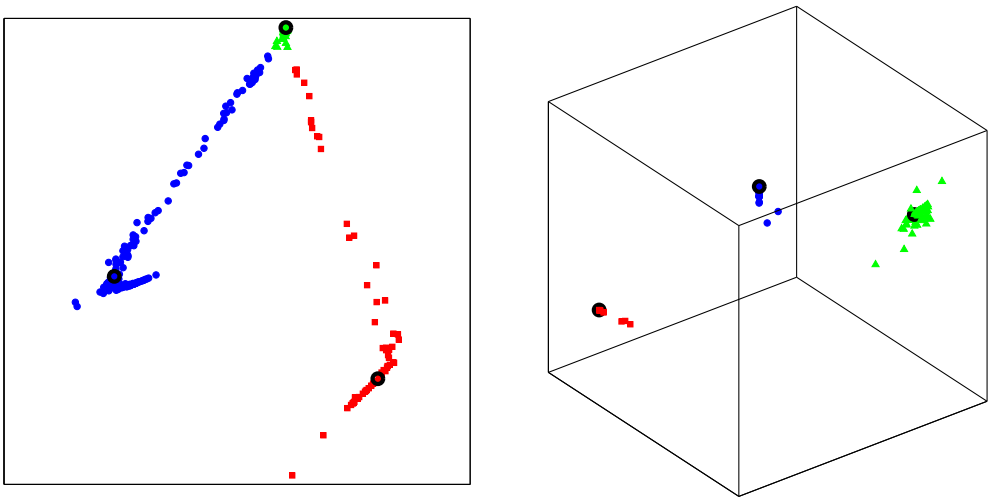
Iš 4.9 lentelės matome, kad pirmojo eksperimento metu, kai norimos tinklo atsako reikšmės yra netransformuoti centrai, tinkamiausias rezultatas gautas trečiu tinklo apmokymo atveju, o antrojo eksperimento metu, kai

norimos tinklo atsako reikšmės yra transformuoti centrai, – ketvirtu tinklo apmokymo atveju. Mažajame ir išėjimo sluoksniuose gauti vizualizavimo rezultatai pateikti 4.30 paveiksle. Mažojo sluoksnio įverčiai pagal antrąjį ir trečiąjį vizualizavimo kokybės kriterijus pateikti 4.10 ir 4.11 lentelėse.



(a) Mažasis sluoksnis **(b)** Išėjimų sluoksnis

4.30 pav. Z sluoksnyje naudojamos Gausinės funkcijos, o norimos tinklo atsako reikšmės yra netransformuoti centrai, $E(W) = 0{,}0484$



(a) Mažasis sluoksnis **(b)** Išėjimų sluoksnis

4.31 pav. Z sluoksnyje naudojamos Gausinės funkcijos, o norimos tinklo atsako reikšmės yra transformuoti centrai, $E(W) = 0{,}0009$

4.10 lentelė: 4.30 paveikslo antrojo vizualizavimo kokybės kriterijaus įverčiai

Paveikslas	● klasteris	▲ klasteris	■ klasteris
4.30a	0,36	0,38	0,64
4.31a	0,70	0,05	0,91

4.11 lentelė: 4.30 paveikslo trečiojo vizualizavimo kokybės kriterijaus įverčiai

Paveikslas	● ir ▲ klasteriai	▲ ir ■ klasteriai
4.30a	0,02	0,05
4.31a	0,03	0,05

Kaip ir eksponentinės funkcijos atveju, po apmokymo tinklas mažesnes paklaidas daro, kai norimos tinklo atsako reikšmės yra transformuoti centrai. Peržvelgę 4.9 lentelėje pateiktus eksperimentų rezultatus matome, kad abiejuose eksperimentuose tinkamiausi rezultatai atrenkami tik pagal pirmąjį atrankos kriterijų (t. y. klasterių išsaugojimą duomenyse). Pirmojo eksperimento metu, kai norimos tinklo atsako reikšmės yra netransformuoti centrai, pirmojo atrankos kriterijaus reikšmė χ yra didesnė už 0, t. y. $\chi > 0$, visais tinklo apmokymo atvejais. Mažiausia χ reikšmė yra 3, vadinasi 3 taškai Y_i netenkino sąlygos $Y_i \in K_j^y$, kai $X_i \in K_j$, $i = \overline{1,m}$, $j = \overline{1,k}$. Pirmojo atrankos kriterijaus reikšmė $\chi = 3$ per visus tinklo apmokymus yra tik vienintelė, todėl antrąjį atrankos kriterijų taikyti nėra tikslo, nes visos kitos χ reikšmės yra didesnės nei 3. Lygiai taip pat ir antrojo eksperimento metu, kai norimos tinklo atsako reikšmės transformuoti centrai, yra tik vienintelė minimali pirmojo atrankos kriterijaus reikšmė χ lygi 0 ir ji rodo idealų tinklo apmokymą. Taigi pagal klasterių išsaugojimo duomenyse kriterijų REGM tinklas geriau apmokomas, kai norimos tinklo atsako reikšmės yra transformuoti centrai.

Peržvelkime gautus vizualizavimo rezultatus mažajame ir išėjimo sluoksniuose, kurie pateikti 4.30 paveiksle. Išėjimo sluoksnyje gautuose vaizduose matomas taškų pasibarstymas. 4.31a paveiksle taškai labiau koncentruojasi aplink savo klasterio centrą, o 4.30a paveiksle klasterių, pažymėtų ▲ ir ●, taškai pasibarstę, net iki gretimų klasterių. Tiek 4.30a, tiek 4.31a paveiksluose mažajame sluoksnyje gautų rezultatų vaizdai tik iš dalies tenkina užsibrėžtus vizualizavimo kokybės kriterijus. Matome, kad taškai labiau yra pasibarstę, kai norimos tinklo atsako reikšmės netransformuoti centrai. Todėl pirmąjį vizualizavimo kokybės kriterijų labiau tenkina 4.31a paveiksle pateiktas vizualizavimo rezultatas. Pagal 4.10 lentelėje pateiktus duomenis matome, kad antrąjį vizualizavimo kokybės kriterijų tenkina tik 4.30a paveiksle pateikti vizualizavimo rezultatai. Trečiojo vizualizavimo kokybės kriterijaus netenkina nei vienas vizualizavimo rezultatas. Tačiau šis kriterijus nėra būtinas. Apibendrinus 4.10 ir 4.11 lentelių bei 4.30 paveikslo rezultatus galime teigti, kad nei vienas vizualizavimo rezultatas pilnai netenkina užsibrėžtų vizualizavimo kokybės kriterijų. Tačiau remiantis atrankos rezultatais, kurie pateikti 4.9 lentelėje, galime teigti, kad REGM

tinklas kokybiškiau bus apmokytas, kai norimos tinklo atsako reikšmės yra transformuoti centrai.

4.12 lentelėje pateikiami su visais šioje disertacijoje aprašytais daugiamačių duomenų rinkiniais atliktų eksperimentų rezultatai. Radialinių bazinių funkcijų sluoksnyje naudota Gausinė funkcija.

4.12 lentelė: Visų duomenų rinkinių norimų tinklo atsako reikšmių parinkimas, kai Z sluoksnyje naudojama Gausinė funkcija

	Netransformuoti centrai			Transformuoti centrai		
Duomenų rinkinys	Paklaida	χ	κ	Paklaida	χ	κ
Irisai	0,0280	0	0,12	0,0033	0	0,27
Stuburo ligos	0,0452	3	0,12	0,0041	0	0,23
Krūties vėžys	0,0617	5	0,02	0,0045	0	0,28
Širdies ligos	0,0218	3	0,05	0,0033	0	0,26
Parkinsono liga	0,0633	1	0,11	0,0054	0	0,18
Vystantys medžiai	0,0134	1	0,06	0,0022	0	0,02
E.coli bakterijos	0,0328	1	0,32	0,0039	0	0,21
Kviečių grūdai	0,0277	1	0,09	0,0055	0	0,38

Iš 4.12 lentelėje pateiktų rezultatų matome, kad ir su kitais duomenų rinkiniais gauname, kad norimas tinklo atsako reikšmes yra tikslingiau imti transformuotus centrus. Eksperimente, kai norimos tinklo atsako reikšmės yra netransformuoti centrai, tik Irisų duomenų rinkinio atveju klasterių išsaugojimo duomenyse kriterijaus reikšmė χ yra lygi 0. Visais kitais atvejais $\chi > 0$. Antrojo eksperimento metu, kai norimos tinklo atsako reikšmės yra transformuoti centrai, visada klasterių išsaugojimo duomenyse kriterijaus reikšmė χ yra lygi 0, $\chi = 0$.

Pirmasis vizualizavimo kokybės kriterijus labiau tenkinamas transformuotų centrų atveju, nes taškai aiškiau išsidėsto tiesių ar kreivių aplinkoje. Visų duomenų rinkių gautų projekcijų kiekybiniai vizualizavimo kokybės kriterijai (antrasis ir trečiasis) pateikti 4.13 ir 4.14 lentelėse.

Iš 4.13 lentelės matome, kad antrąjį vizualizavimo kokybės kriterijų ($\bar{a} > 0{,}1$) netransformuotų centrų atveju atitinka penki duomenų rinkiniai, o transformuotų centrų atveju atitinka tik du duomenų rinkiniai.

Trečiasis vizualizavimo kokybės kriterijus yra pageidautinas, bet nebūtinas. Jo reikšmė $\hat{a}$ turi būti didesnė arba lygi 0,05. Iš 4.14 lentelės matome, kad netransformuotų centrų atveju, šį kriterijų tenkina tik Irisų duomenų rinkinys. Transformuotų centrų atveju trečiąjį vizualizavimo kokybės kriterijų tenkina tik Irisų, E.coli bakterijų ir kviečių grūdų duomenų rinkiniai. Visiems kitiems duomenų rinkiniams trečiasis vizualizavimo kokybės kriterijus yra netenkinamas.

4.13 lentelė: Antrojo vizualizavimo kokybės kriterijaus įverčiai, kai Z sluoksnyje naudojama Gausinė funkcija

Duomenų rinkinys	Netransformuoti centrai			Transformuoti centrai		
	●	▲	■	●	▲	■
Irisai	0,03	0,41	0,47	0,43	0,07	0,05
Stuburo ligos	0,36	0,38	0,64	0,70	0,05	0,91
Krūties vėžys	0,33	0,34	0,30	0,01	0,40	0,62
Širdies ligos	0,52	0,006	0,47	0,81	0,01	0,97
Parkinsono liga	0,33	0,28	0,32	0,01	0,41	0,57
Vystantys medžiai	0,56	0,19	0,65	0,75	0,16	0,89
E.coli bakterijos	0,59	0,24	0,53	0,44	0,24	0,65
Kviečių grūdai	0,04	0,56	0,34	0,09	0,77	0,82

4.14 lentelė: Trečiojo vizualizavimo kokybės kriterijaus įverčiai, kai Z sluoksnyje naudojama Gausinė funkcija

Duomenų rinkinys	Netransformuoti centrai		Transformuoti centrai	
	● ir ▲	▲ ir ■	● ir ▲	▲ ir ■
Irisai	0,72	0,12	0,44	0,32
Stuburo ligos	0,02	0,05	0,03	0,05
Krūties vėžys	0,03	0,01	0,02	0,02
Širdies ligos	0,01	0,002	0,04	0,02
Parkinsono liga	0,04	0,20	0,01	0,04
Vystantys medžiai	0,001	0,02	0,001	0,05
E.coli bakterijos	0,04	0,11	0,05	0,13
Kviečių grūdai	0,01	0,04	0,06	0,10

Pagal 4.13 ir 4.14 lentelėse pateiktus duomenis galime teigti, kad gauti vizualizavimo rezultatai netenkina užsibrėžtų vizualizavimo kokybės kriterijų, kai daugiamačiams duomenims transformuoti naudojama Gausinė radialinė bazinė funkcija.

Apibendrinus norimų tinklo atsako reikšmių parinkimo eksperimentų rezultatus galime padaryti išvadą, kad REGM tinklas kokybiškiau apmokomas ir mažajame sluoksnyje gauti vizualizavimo rezultatai atitinka užsibrėžtus vizualizavimo kokybės kriterijus, kai norimomis tinklo atsako reikšmėmis imami transformuoti centrai ir radialinių bazinių funkcijų sluoksnyje naudojama eksponentinė funkcija.

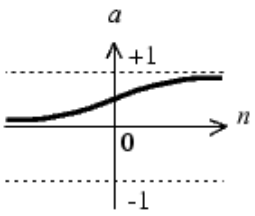
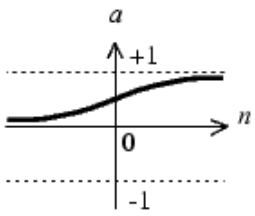
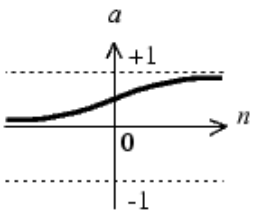
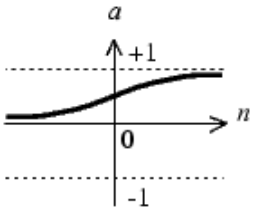
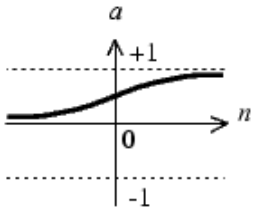
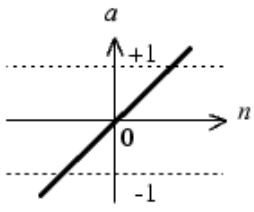
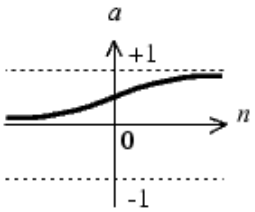
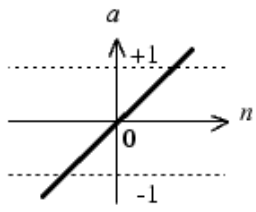
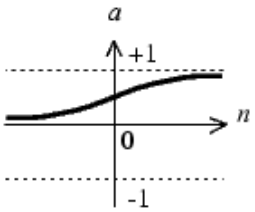
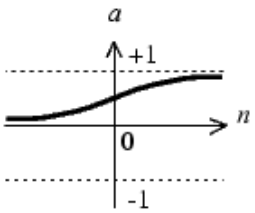
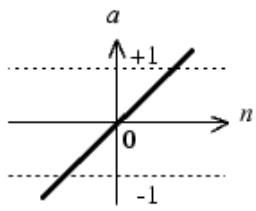
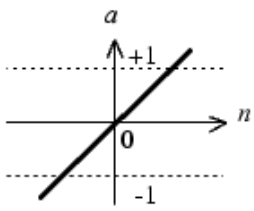
4.5. Antrosios REGM tinklo dalies aktyvavimo funkcijos

Kitas labai svarbus REGM tinklo apmokymo faktorius ir nuo to priklausantys mažajame sluoksnyje gaunami vizualizavimo rezultatai, tai antrojoje REGM tinklo dalyje esančiuose paslėptuose neuronų sluoksniuose ir išėjimų sluoksnyje naudojamos aktyvavimo funkcijos. Kaip yra paminėta

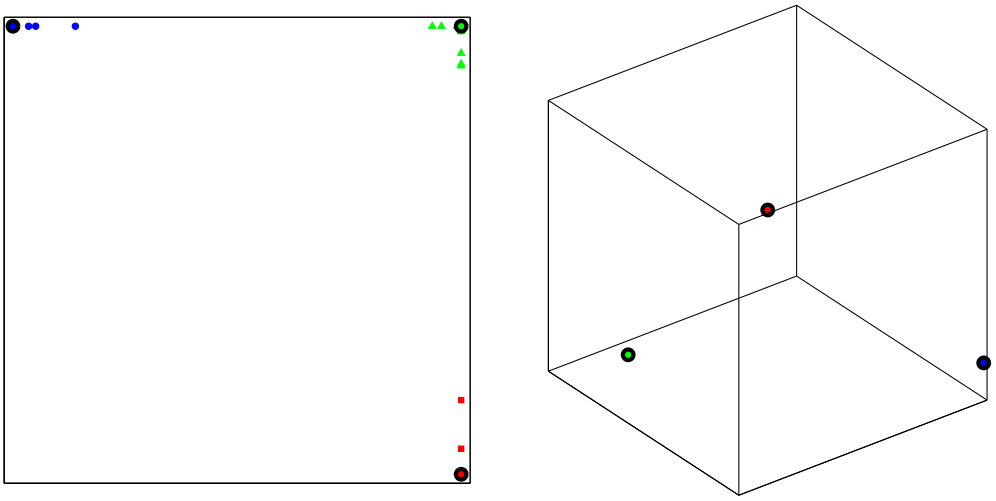
3.2. poskyryje, šiuose sluoksniuose naudojamos loginio sigmoido (2.16) arba tiesinė (2.15) aktyvavimo funkcijos. Tik kyla vienintelis klausimas, kuriame sluoksnyje ir kokią aktyvavimo funkciją geriausia naudoti, kad apsimokiusio REGM tinklo mažajame sluoksnyje gautas vaizdas atitiktų užsibrėžtus vizualizavimo kokybės kriterijus?

Buvo atlikti keturi eksperimentai. Kiekvienas eksperimentas skyrėsi pagal naudojamas aktyvavimo funkcijas mažajame ir išėjimų sluoksnyje. Aktyvavimo funkcijų naudojimas eksperimentuose pateikiamas 4.15 lentelėje. Paprastumo dėlei pirmąjį eksperimentą pasižymėkime *2L*, antrąjį – *LT*, trečiąjį – *TL*, o ketvirtąjį – *2T*. Pirmajame paslėptame sluoksnyje visuose eksperimentuose buvo naudojama loginio sigmoido aktyvavimo funkcija. Tinklas buvo apmokytas 30 kartų kiekviename eksperimente. Šių keturių eksperimentų atvejais buvo naudojami tie patys 30 pradinių svorių rinkinių. Tai leidžia palyginti po eksperimentų gautus rezultatus naudojant nedidelį kiekį ilgų skaičiavimų.

4.15 lentelė: Paslėptuose ir išėjimo sluoksnyje esančių aktyvavimo funkcijų parinkimas

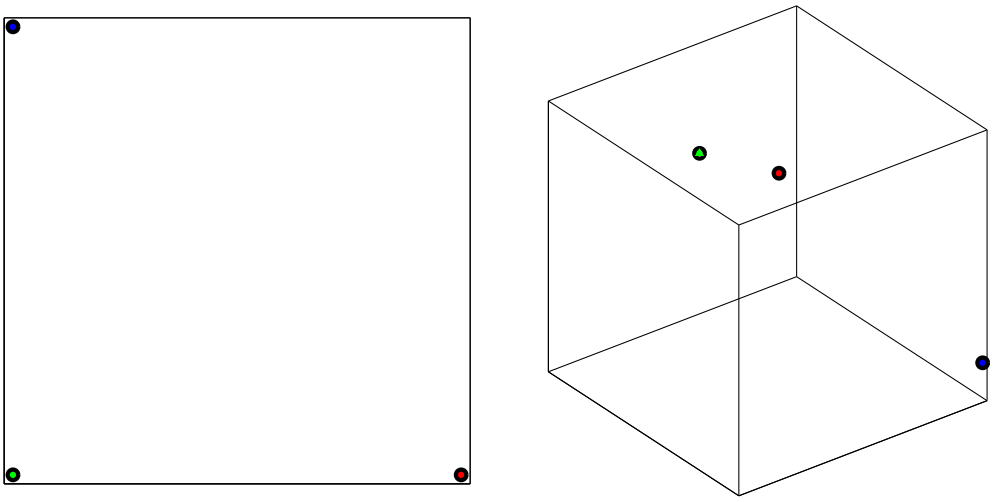
	P^1	P^2	Y
2L			
LT			
TL			
2T			

Atliktų eksperimentų rezultatai iliustruoti Širdies ligų duomenų rinkiniu ir pateikti 4.32, 4.33, 4.34 ir 4.35 paveiksluose.



(a) Mažasis sluoksnis **(b)** Išėjimo sluoksnis

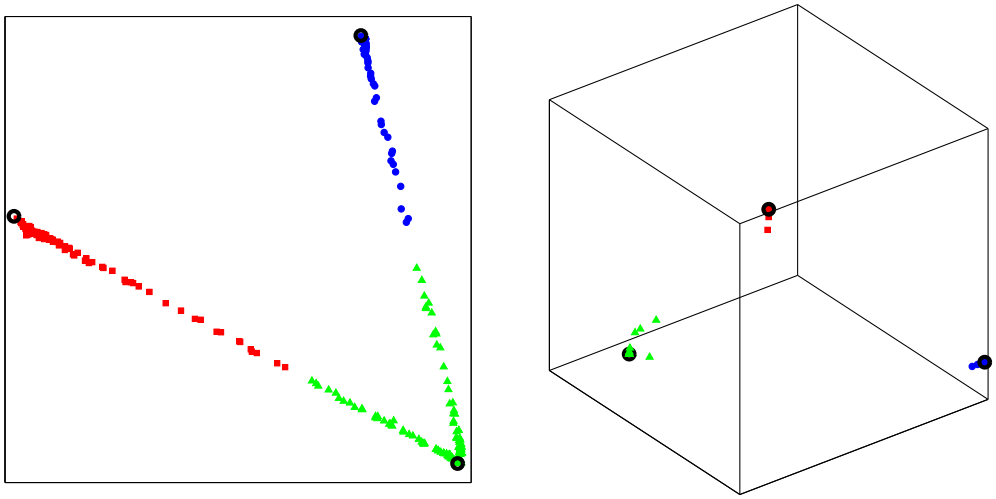
4.32 pav. *2L* eksperimente gauti vizualizavimo rezultatai, $E(W) = 0{,}00059$



(a) Mažasis sluoksnis **(b)** Išėjimo sluoksnis

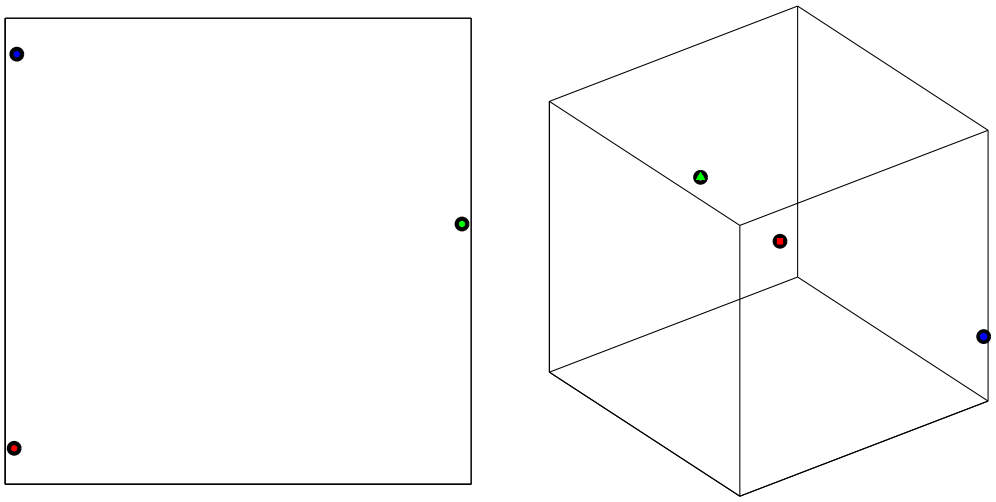
4.33 pav. *LT* eksperimente gauti vizualizavimo rezultatai, $E(W) = 1{,}09 \times 10^{-14}$

Iš 4.32a, 4.33a, 4.34a ir 4.35a paveiksluose pateiktų mažojo sluoksnio vizualizavimo rezultatų matome, kad informatyviausias ir užsibrėžtus vizualizavimo kokybės kriterijus atitinka vaizdas, kuris gautas *TL* eksperimente. Aptarkime po kiekvieno tinklo apmokymo gautus vizualizavimo rezultatus (mažajame ir išėjimo sluoksniuose). Priminsime, kad gautas vizualizavimo rezultatas atitiktų užsibrėžtus vizualizavimo kokybės kriterijus turi būti patenkinti pirmi du kriterijai ir pageidautinas, bet neprivalomas trečiasis kriterijus. Vizualizavimo rezultatų kiekybiniai antrojo ir trečiojo kriterijų įverčiai pateikti 4.16 ir 4.17 lentelėse.



(a) Mažasis sluoksnis **(b)** Išėjimo sluoksnis

4.34 pav. *TL* eksperimente gauti vizualizavimo rezultatai, $E(W) = 0{,}00063$



(a) Mažasis sluoksnis **(b)** Išėjimo sluoksnis

4.35 pav. *2T* eksperimente gauti vizualizavimo rezultatai, $E(W) = 2{,}19{\times}10^{-14}$

4.16 lentelė: Antrojo vizualizavimo kokybės kriterijaus įverčiai

Paveikslas	● klasteris	▲ klasteris	■ klasteris
4.32a	0,09	0,08	0,12
4.33a	0,000009	0,000064	0,000003
4.34a	0,38	0,39	0,61
4.35a	0,00003	0,00033	0,00041

Po *2L* eksperimento išėjimo sluoksnio vizualizavimo rezultate matome tik tiek taškų, kiek yra duomenyse klasterių, t. y. trys. Vadinasi REGM tinklas apmokytas labai gerai, nes nėra taškų Y_i, $i = \overline{1, m}$, pasibarstymo. Mažajame sluoksnyje gautas vaizdas neatitinka labai svarbaus antrojo vizualizavimo kokybės kriterijaus.

4.17 lentelė: Trečiojo vizualizavimo kokybės kriterijaus įverčiai

Paveikslas	● ir ▲ klasteriai	▲ ir ■ klasteriai
4.32a	0,56	0,53
4.33a	0,70	0,70
4.34a	0,09	0,06
4.35a	0,95	0,99

Po *LT* ir *2T* eksperimentų išėjimo sluoksnyje gauti vizualizavimo rezultatai labai panašūs. Abiem atvejais matome tik tiek taškų, kiek duomenyse yra klasterių, t. y. trys. Du klasteriai, kurių taškai pažymėti ■ ir ▲, yra artimi, o klasteris, pažymėtas ●, yra labiau nutolęs. Mažajame sluoksnyje gautų rezultatų vaizde taip pat matoma tik tiek taškų, kiek duomenyse yra klasterių, todėl šie vaizdai nesuteikia daugiau žinių nei išėjimo sluoksnyje pateiktas vaizdas.

Po *TL* eksperimento pagal išėjimų sluoksnyje gautą vizualizavimo rezultatą matome, kad tinklas dar nėra labai gerai apmokytas, nes taškai Y_i pasibarstę aplink klasterius K_j^y. Tačiau mažajame sluoksnyje gautas vaizdas yra informatyvus ir atitinka visus tris užsibrėžtus vizualizavimo kokybės kriterijus:

1. taškai išsidėstę tiesių ar kreivių aplinkoje;
2. taškai „išsibarstę“ klasteriuose (didžiausi atstumai tarp klasterių taškų yra didesni už 0,1);
3. matomos ribos tarp klasterių (mažiausias atstumas tarp skirtingiems klasteriams priklausančių taškų yra didesnis už 0,05).

Apibendrinus atliktus keturis eksperimentus su Širdies ligų duomenų rinkiniu galime daryti išvadą, kad hibridinis neuroninis tinklas REGM mažajame sluoksnyje gauna informatyvesnius ir atitinkančius užsibrėžtus vizualizavimo kokybės kriterijus vizualizavimo rezultatus, kai mažajame sluoksnyje naudojama tiesinė aktyvavimo funkcija, o pirmame paslėptame ir išėjimo sluoksnyje naudojama loginio sigmoido aktyvavimo funkcija.

Peržiūrėkime ir palyginkime su kitais (Stuburo ligų, Krūties vėžio, Parkinsono ligos ir Vystančių medžių) daugiamačių duomenų rinkiniais atliktų eksperimentų rezultatus, kurie pateikti 4.18, 4.19, 4.20 ir 4.21 lentelėse. Pirmasis vizualizavimo kokybės kriterijus buvo tenkinamas tik *2L* ir *TL* eksperimentuose.

Iš 4.18 lentelėje pateiktų duomenų matome, kad antrasis vizualizavimo kokybės kriterijus netenkinamas visiems duomenų rinkiniams, nes $\bar{a}_{K_j}$ turi būti didesnis už 0,1 visiems klasteriams. Trečiasis vizualizavimo kokybės

4.18 lentelė: *2L* eksperimente gautų vizualizavimo rezultatų antrojo ir trečiojo vizualizavimo kokybės kriterijų įverčiai

Duomenų rinkinys	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● ir ▲	▲ ir ■
Stuburo ligos	0,07	0,32	0,06	0,56	0,33
Krūties vėžys	0,05	0,29	0,11	0,38	0,57
Parkinsono liga	0,20	0,49	0,04	0,38	0,49
Vystantys medžiai	0,29	0,12	0,04	0,25	0,21

kriterijus yra tik pageidautinas, todėl jei netenkinamas bent vienas iš pirmųjų vizualizavimo kokybės kriterijų, tai į šį kriterijų yra neatsižvelgiama.

4.19 lentelė: *LT* eksperimente gautų vizualizavimo rezultatų antrojo ir trečiojo vizualizavimo kokybės kriterijų įverčiai

Duomenų rinkinys	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● ir ▲	▲ ir ■
Stuburo ligos	0,00001	0,00002	0,00050	0,71	0,71
Krūties vėžys	0,07000	0,00001	0,00002	0,71	0,63
Parkinsono liga	0,00001	0,00007	0,00002	0,71	0,71
Vystantys medžiai	0,02317	0,01549	0,01639	0,54	0,69

Iš 4.19 lentelėje pateiktų duomenų matome, kad antrasis vizualizavimo kokybės kriterijus taip pat netenkinamas visiems duomenų rinkiniams.

4.20 lentelė: *2T* eksperimente gautų vizualizavimo rezultatų antrojo ir trečiojo vizualizavimo kokybės kriterijų įverčiai

Duomenų rinkinys	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● ir ▲	▲ ir ■
Stuburo ligos	0,00004	0,00009	0,00077	0,46	0,84
Krūties vėžys	0,00004	0,00005	0,00001	0,99	0,52
Parkinsono liga	0,00002	0,00028	0,00003	0,96	0,44
Vystantys medžiai	0,09249	0,08390	0,04963	0,38	0,82

Iš 4.20 lentelėje pateiktų duomenų matome, kad antrasis vizualizavimo kokybės kriterijus taip pat netenkinamas visiems duomenų rinkiniams.

4.21 lentelė: *TL* eksperimente gautų vizualizavimo rezultatų antrojo ir trečiojo vizualizavimo kokybės kriterijų įverčiai

Duomenų rinkinys	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● ir ▲	▲ ir ■
Stuburo ligos	0,35	0,40	0,51	0,06	0,08
Krūties vėžys	0,36	0,29	0,39	0,03	0,06
Parkinsono liga	0,36	0,15	0,65	0,04	0,22
Vystantys medžiai	0,28	0,68	0,77	0,003	0,01

Pagal 4.21 lentelėje pateiktus duomenis matome, kad visiems duomenų rinkiniams yra tenkinamas antrasis vizualizavimo kokybės kriterijus. Trečiasis vizualizavimo kokybės kriterijus tenkinamas ne visiems duomenų rinkiniams, tačiau jis nėra būtinas.

Apibendrinus po keturių eksperimentų gautus rezultatus galime daryti išvadą, REGM tinklas kokybiškiau apmokomas ir mažajame sluoksnyje gauti vizualizavimo rezultatai atitinka užsibrėžtus vizualizavimo kokybės kriterijus, kai mažajame sluoksnyje naudojama tiesinė aktyvavimo funkcija, o pirmame paslėptame ir išėjimo sluoksnyje naudojama loginio sigmoido aktyvavimo funkcija.

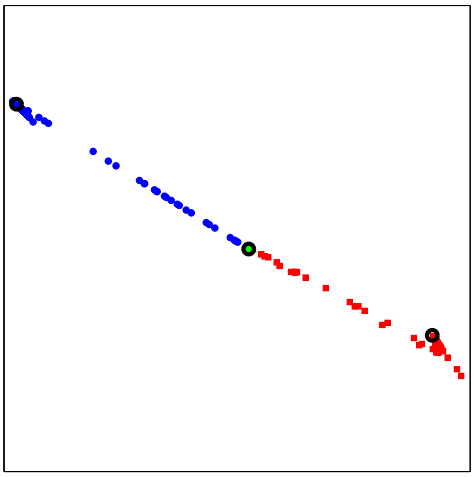
4.6. Neuronų skaičius išėjimo sluoksnyje

Šiame tinklo REGM eksperimente buvo stebėta, kaip kinta mažajame sluoksnyje gauti vizualizavimo rezultatai, kai kiekvieną kartą mokant tinklą parenkamas skirtingas neuronų skaičius išėjimo sluoksnyje. Išėjimo sluoksnyje neuronų gali būti nuo vieno iki k (klasterių skaičiaus). Tarkime, duomenų rinkiniui pasirinktas klasterių skaičius $k = 3$. Tuomet išėjimo sluoksnyje gali būti vienas, du arba trys neuronai.

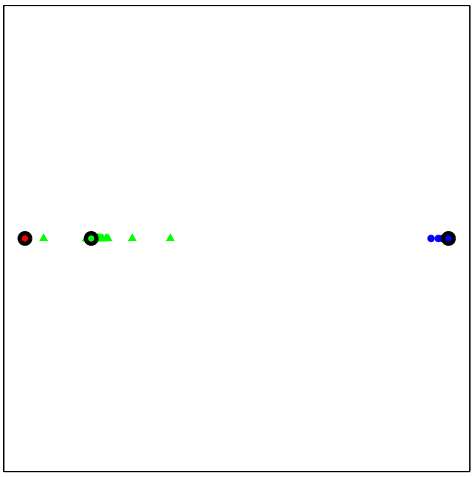
Eksperimento metu buvo apmokomas tinklas aprašytas 4.3. poskyryje. Išėjimų sluoksnio Y neuronų skaičius pasirenkamas nuo vieno iki k (pirmame eksperimente $s = 1$; antrame eksperimente $s = 2$; trečiame eksperimente $s = 3$). Norimos tinklo atsako reikšmės yra transformuoti klasterių centrai. Pastebėsime, kad pirmajame ir antrajame eksperimente transformuoti klasterių centrai $\mu_j^z \in \mathbb{R}^k$ daugiamačių skalių metodu buvo projektuojami į $\mathbb{R}^s$ erdvę, nes $s < k$, o trečiajame eksperimente transformuotų duomenų projektavimas iš $\mathbb{R}^k$ į $\mathbb{R}^s$ nėra reikalingas, nes $s = k$. Vieno eksperimento metu REGM tinklas buvo apmokytas 30 kartų.

Eksperimentai buvo atlikti su daugiamačių duomenų rinkiniais, aprašytais 4.1. poskyryje. Po eksperimentų gaunami vizualizavimo rezultatai mažajame ir išėjimų sluoksnyje iliustruoti Stuburo ligų, Krūties vėžio ir Širdies ligų duomenų rinkiniais.

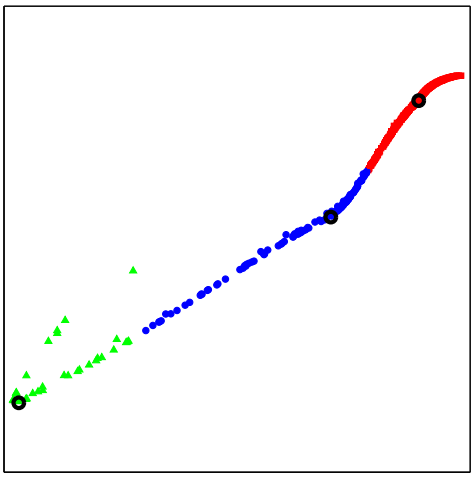
Pirmojo eksperimento, kai $s = 1$, rezultatai pateikiami 4.36 paveiksle. Skirtingi klasteriai pažymėti ●, ▲, ■, o klasterių centrai ⭘. Stuburo ligų duomenų rinkiniu apmokyto REGM tinklo, kuris atrinktas pagal du atrankos kriterijus iš 30 tinklo apmokymų, gaunama paklaida lygi $E(W) = 0{,}0012$. Krūties vėžio duomenų rinkiniu apmokyto REGM tinklo gaunama paklaida lygi $E(W) = 0{,}0014$. Širdies ligų duomenų rinkiniu apmokyto REGM tinklo gaunama paklaida lygi $E(W) = 0{,}0019$.



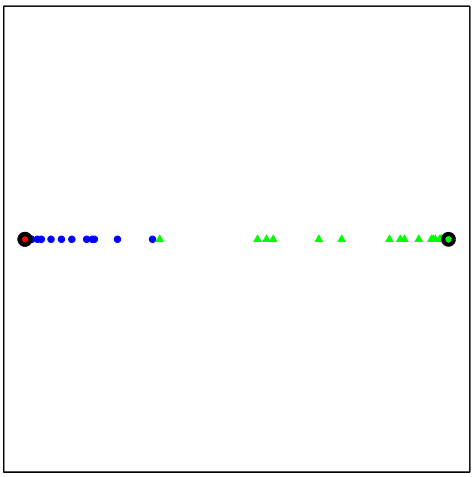
(a) SL, mažasis sluoksnis



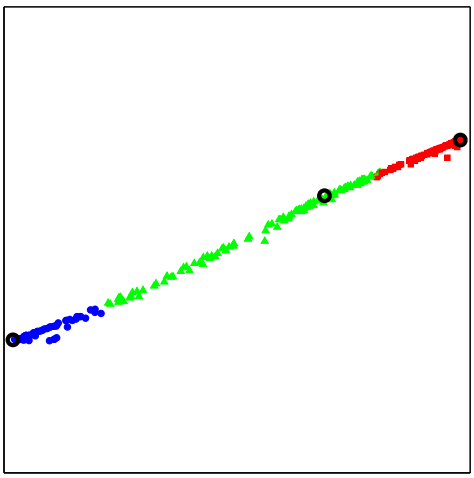
(b) SL, išėjimo sluoksnis



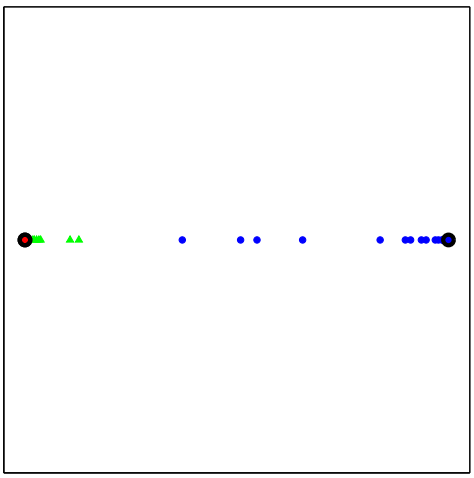
(c) KV, mažasis sluoksnis



(d) KV, išėjimo sluoksnis



(e) ŠL, mažasis sluoksnis



(f) ŠL, išėjimo sluoksnis

4.36 pav. REGM tinklas apmokytas (SL – stuburo ligų, KV – krūties vėžio, ŠL – širdies ligų) duomenų rinkiniais, kai išėjimo sluoksnyje pasirinktas vienas neuronas

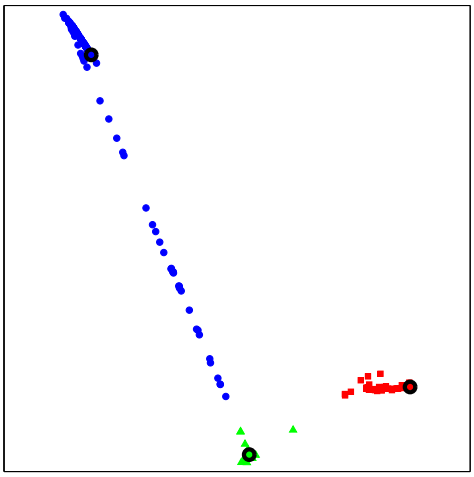
4.36 paveikslo išėjimų sluoksnio vaizduose matome, kad taškai yra išsidėstę ant tiesės, nes šiame sluoksnyje buvo pasirinktas vienas neuronas. Kaip jau yra minėta, tinklo išėjimų sluoksnyje gautų reikšmių vizualizavimas parodo, ar tinklas kokybiškai apmokytas, t. y. idealiu atveju turėtų matytis tik tiek taškų, kiek duomenų rinkinyje pasirinkta klasterių. Iš 4.36 paveiksle pateiktų išėjimo sluoksnyje gautų vizualizavimo rezultatų matome, kad tik Stuburo ligų duomenų rinkinyje (4.36b paveikslas) išsiskiria pasirinktų trijų klasterių centrai, o Krūties vėžio (4.36d paveikslas) ir Širdies ligų (4.36f paveikslas) duomenų rinkiniuose, išsiskiria tik dviejų klasterių centrai. Pagal skirtingomis spalvomis pažymėtus klasterius, galime daryti prielaidą, kad dviejų klasterių centrai susiprojektavę į tą patį tašką (Krūties vėžio duomenų rinkinyje (4.36d paveikslas) ● ir ■ pažymėti klasteriai, o Širdies ligų duomenų rinkinyje (4.36f paveikslas) ▲ ir ■ pažymėti klasteriai). Taip pat pastebimas ir nemažas taškų (objektų) pasibarstymas. Pagal išėjimo sluoksnyje gautus vizualizavimo rezultatus galime daryti išvadą, kad REGM tinklas apmokytas nekokybiškai.

Mažajame sluoksnyje gauti vizualizavimo rezultatai bus pakomentuoti vėliau.

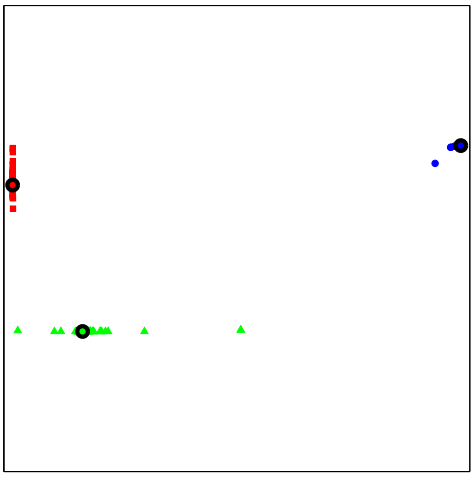
Antrajame eksperimente, išėjimo sluoksnyje buvo pasirinkti du neuronai, t. y. $s = 2$. Po eksperimento gauti vizualizavimo rezultatai pateikiami 4.37 paveiksle. Stuburo ligų duomenų rinkiniu apmokyto REGM tinklo, kuris atrinktas pagal du atrankos kriterijus iš 30 tinklo apmokymų, gaunama paklaida lygi $E(W) = 0{,}0006$. Krūties vėžio duomenų rinkiniu apmokyto REGM tinklo gaunama paklaida lygi $E(W) = 0{,}0007$. Širdies ligų duomenų rinkiniu apmokyto REGM tinklo gaunama paklaida lygi $E(W) = 0{,}0011$.

Iš 4.37 paveiksle pateiktų išėjimo sluoksnyje gaunamų vaizdų matome, kad kokybiškai apmokytas tik tas tinklas, kuris buvo mokomas Krūties vėžio duomenų rinkiniu (4.37d paveikslas), t. y. po tinklo apmokymo matoma tik tiek taškų, kiek duomenyse pasirinkta klasterių. Tinklą apmokius kitais dviem duomenų rinkiniais (4.37b ir 4.37f paveikslai), dar stebimas šioks toks taškų pasibarstymas.

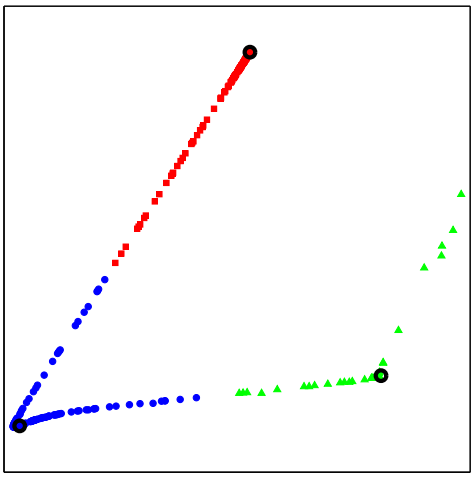
Trečiojo eksperimento, kai $s = 3$, rezultatai pateikiami 4.38 paveiksle. Stuburo ligų duomenų rinkiniu apmokyto REGM tinklo, kuris atrinktas pagal du atrankos kriterijus iš 30 tinklo apmokymų, gaunama paklaida lygi $E(W) = 0{,}0008$. Krūties vėžio duomenų rinkiniu apmokyto REGM tinklo gaunama paklaida lygi $E(W) = 0{,}0002$. Širdies ligų duomenų rinkiniu apmokyto REGM tinklo gaunama paklaida lygi $E(W) = 0{,}0005$.



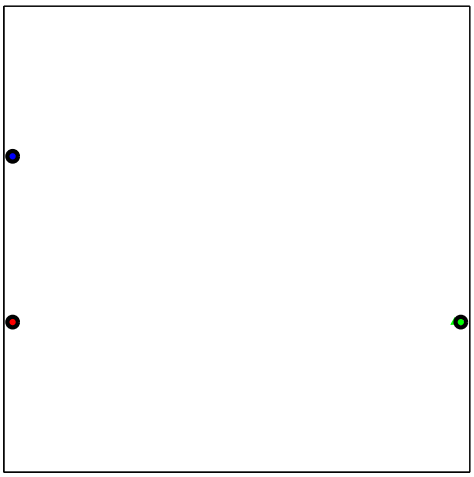
(a) SL, mažasis sluoksnis



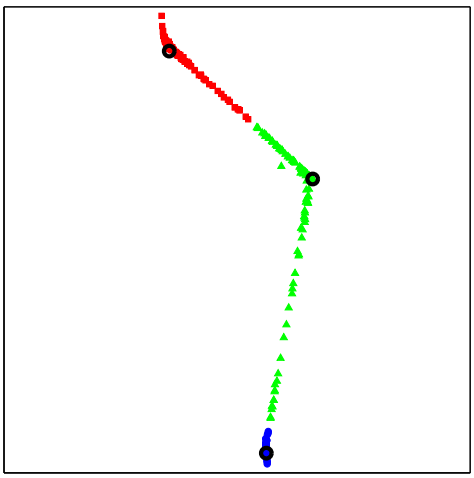
(b) SL, išėjimo sluoksnis



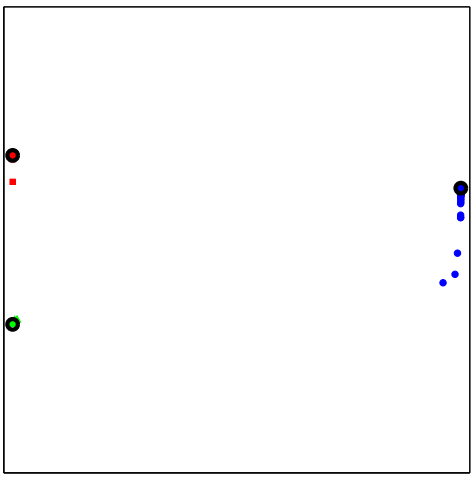
(c) KV, mažasis sluoksnis



(d) KV, išėjimo sluoksnis

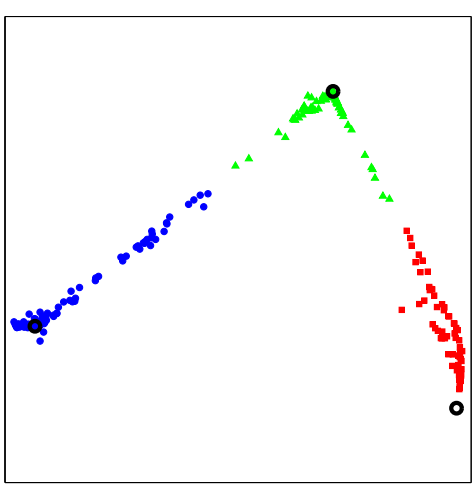


(e) ŠL, mažasis sluoksnis

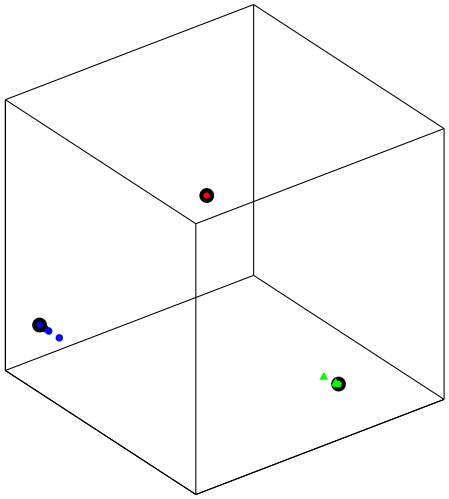


(f) ŠL, išėjimo sluoksnis

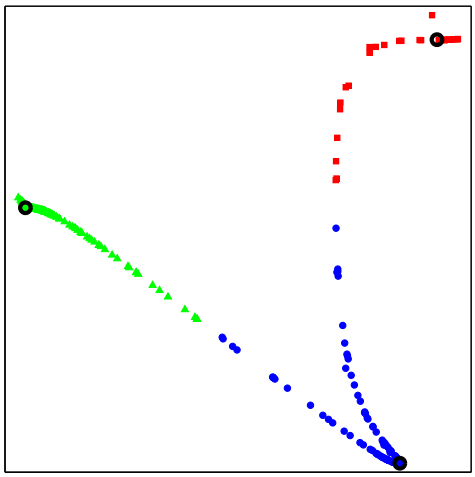
4.37 pav. REGM tinklas apmokytas (SL – stuburo ligų, KV – krūties vėžio, ŠL – širdies ligų) duomenų rinkiniais, kai išėjimo sluoksnyje pasirinkti du neuronai



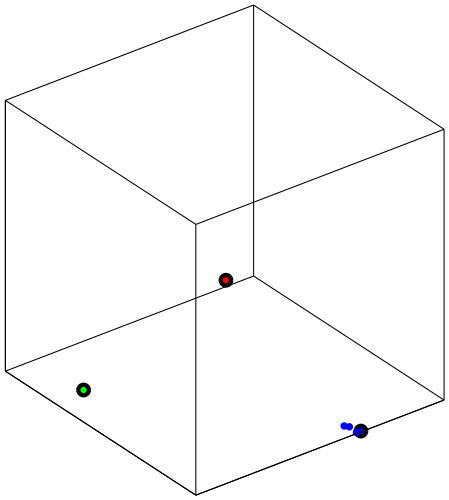
(a) SL, mažasis sluoksnis



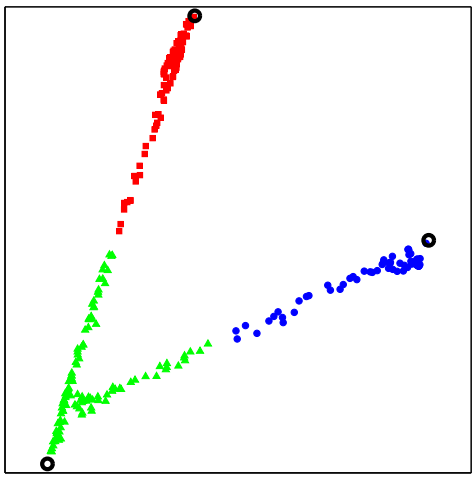
(b) SL, išėjimo sluoksnis



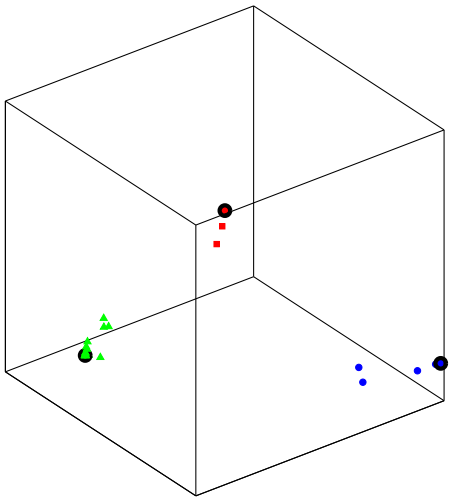
(c) KV, mažasis sluoksnis



(d) KV, išėjimo sluoksnis



(e) ŠL, mažasis sluoksnis



(f) ŠL, išėjimo sluoksnis

4.38 pav. REGM tinklas apmokytas (SL – stuburo ligų, KV – krūties vėžio, ŠL – širdies ligų) duomenų rinkiniais, kai išėjimo sluoksnyje pasirinkti trys neuronai

Iš 4.38 paveiksle pateiktų išėjimo sluoksnyje gautų vizualizavimo rezultatų matome, kad tinklą apmokius Stuburo ligų (4.38b paveikslas) ir Krūties vėžio (4.38d paveikslas) duomenų rinkiniais, REGM tinklas apmokytas kokybiškai, nes gautuose vaizduose matome tik tiek taškų, kiek duomenyse pasirinkta klasterių. Širdies ligų duomenų rinkinio (4.38f paveikslas) atveju, stebimas nedidelis taškų pasibarstymas.

Pateiktųjų 4.36, 4.37 ir 4.38 paveikslų mažojo sluoksnio įverčiai pagal antrąjį ir trečiąjį vizualizavimo kokybės kriterijus pateikti 4.22, 4.23 ir 4.24 lentelėse.

4.22 lentelė: Stuburo ligų duomenų rinkiniui gautų projekcijų antrojo ir trečiojo vizualizavimo kokybės kriterijų įverčiai

Išėjimo sluoksnis	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● ir ▲	▲ ir ■
$s = 1$	0,50	0,01	0,44	0,02	0,02
$s = 2$	0,82	0,12	0,13	0,08	0,12
$s = 3$	0,52	0,35	0,37	0,09	0,08

4.23 lentelė: Krūties vėžio duomenų rinkiniui gautų projekcijų antrojo ir trečiojo vizualizavimo kokybės kriterijų įverčiai

Išėjimo sluoksnis	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● ir ▲	● ir ■
$s = 1$	0,49	0,32	0,24	0,04	0,003
$s = 2$	0,37	0,59	0,49	0,09	0,04
$s = 3$	0,52	0,46	0,41	0,07	0,10

4.24 lentelė: Širdies ligų duomenų rinkiniui gautų projekcijų antrojo ir trečiojo vizualizavimo kokybės kriterijų įverčiai

Išėjimo sluoksnis	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● ir ▲	▲ ir ■
$s = 1$	0,18	0,62	0,19	0,03	0,003
$s = 2$	0,07	0,63	0,29	0,03	0,02
$s = 3$	0,46	0,45	0,49	0,06	0,05

Iš 4.36 paveikslo matome, kad pirmasis vizualizavimo kokybės kriterijus yra nepatenkinamas, nes taškai turėtų išsidėstyti kelių tiesių arba kreivių aplinkoje. Stuburo ligų ir Širdies ligų duomenų rinkiniams gautose projekcijose matoma tik viena kreivė. Krūties vėžio duomenų rinkiniui gautoje projekcijoje yra matomos dvi kreivės. Tačiau kelių tiesių ar kreivių aplinkoje turėtų išsidėstyti viduriniojo klasterio taškai, kad atsiskleistų šio klasterio taškų panašumas su gretimų klasterių taškais. Jei pirmasis vizualizavimo

kokybės kriterijus yra nepatenkinamas, tai į kitus vizualizavimo kokybės kriterijus galime ir neatsižvelgti.

Pagal 4.37 paveiksle pateiktus vizualizavimo rezultatus matome, kad gautiems vaizdams pirmasis vizualizavimo kokybės kriterijus yra tenkinamas. Pagal 4.22, 4.23 ir 4.24 lentelėse pateiktus duomenis matome, kad antrasis vizualizavimo kokybės kriterijus netenkinamas tik Širdies ligų duomenų rinkiniui. Nors pažiūrėjus į 4.37a paveikslą, norėtųsi, kad ▲ ir ■ pažymėtuose klasteriuose matytųsi daugiau taškų. Trečiasis vizualizavimo kokybės kriterijus tenkinamas tik Stuburo ligų duomenų rinkiniui.

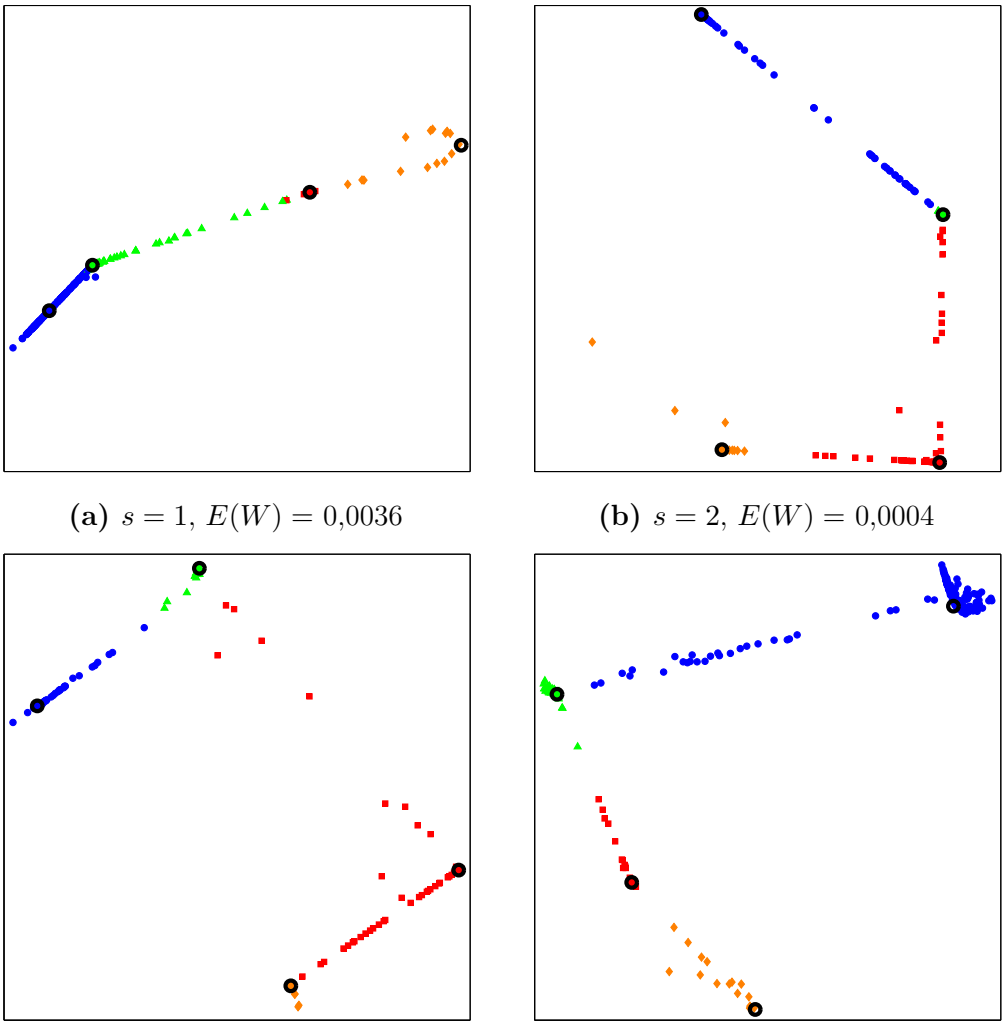
Pagal 4.38 paveiksle pateiktus vizualizavimo rezultatus ir 4.22, 4.23, 4.24 lentelėse pateiktus vizualizavimo kokybės kriterijų įverčius matome, kad visi trys vizualizavimo kokybės kriterijai yra tenkinami visiems duomenų rinkiniams.

Aptarus pateiktus paveikslus ir lenteles galime daryti išvadą, kad tinklas kokybiškiau apmokomas ir mažajame sluoksnyje gauti vizualizavimo rezultatai tiksliau atitinka užsibrėžtus vizualizavimo kokybės kriterijus, kai išėjimo sluoksnyje parinktas neuronų skaičius lygus pasirinktam klasterių skaičiui, t. y. $s = k$. Tačiau šie trys eksperimentai buvo atlikti, kai pasirinktas klasterių skaičius lygus 3. Pasirinkime kitą klasterių skaičių, $k = 4$, ir atlikime dar keturis eksperimentus (pirmas eksperimentas $s = 1$, antras eksperimentas $s = 2$, trečias eksperimentas $s = 3$ ir ketvirtas eksperimentas $s = 4$). Eksperimentų rezultatai iliustruoti Stuburo ligų duomenų rinkiniu 4.39 paveiksle. Kadangi mus labiau domina mažajame sluoksnyje gautų rezultatų vaizdai, tai 4.39 paveiksle pateikiame tik juos. Skirtingi klasteriai pažymėti ●, ▲, ■, ♦, o klasterių centrai ⊙. Mažojo sluoksnio įverčiai pagal antrąjį ir trečiąjį vizualizavimo kokybės kriterijus pateikti 4.25 lentelėje.

4.25 lentelė: Stuburo ligų duomenų rinkiniui gautų projekcijų antrojo ir trečiojo vizualizavimo kokybės kriterijų įverčiai

Išėjimo sluoksnis	$\bar{a}_{K_j}$				$\hat{a}$		
	●	▲	■	♦	● ir ▲	▲ ir ■	■ ir ♦
$s = 1$	0,23	0,43	0,06	0,24	0,0001	0,0007	0,068
$s = 2$	0,59	0,02	0,51	0,37	0,02	0,02	0,14
$s = 3$	0,34	0,11	0,80	0,05	0,06	0,08	0,03
$s = 4$	0,82	0,15	0,19	0,23	0,07	0,11	0,11

Pirmasis vizualizavimo kokybės kriterijus nurodo, kad taškai turėtų išsidėstyti kelių tiesių ar kreivių aplinkoje. Peržvelgę vizualizavimo rezultatus, pateiktus 4.39 paveiksle matome, kad pirmojo vizualizavimo kokybės kriterijaus netenkina tik po pirmojo eksperimento, kai $s = 1$, gautoji projekcija. Antrasis vizualizavimo kokybės kriterijus nurodo, kad visų klasterių



(a) $s = 1$, $E(W) = 0{,}0036$ **(b)** $s = 2$, $E(W) = 0{,}0004$

(c) $s = 3$, $E(W) = 0{,}0003$ **(d)** $s = 4$, $E(W) = 0{,}0007$

4.39 pav. REGM tinklas apmokytas Stuburo ligų duomenų rinkiniu, kai išėjimo sluoksnyje pasirinktas skirtingas neuronų skaičius

didžiausi atstumai $\bar{a}_{K_j}$ turi būti didesni už 0,1. Remiantis 4.25 lentelėje pateiktais rezultatais galime teigti, kad antrąjį vizualizavimo kokybės kriterijų tenkina tik ketvirtame eksperimente, kai $s = 4$, gautoji projekcija. Trečiasis vizualizavimo kokybės kriterijus nurodo, kad mažiausias atstumas tarp gretimų klasterių taškų $\hat{a}$ turi būti lygus arba didesnis už 0,05. Remiantis 4.25 lentelėje pateiktais rezultatais galime teigti, kad trečiasis vizualizavimo kokybės kriterijus taip pat tenkinamas tik po ketvirto eksperimento gautuose vizualizavimo rezultatuose.

Iš atliktų eksperimentų ir gautų rezultatų galime padaryti išvadą, kad tinklas kokybiškiau apmokomas ir mažajame sluoksnyje gauti vizualizavimo rezultatai tiksliau atitinka užsibrėžtus vizualizavimo kokybės kriterijus, kai išėjimo sluoksnyje parinktas neuronų skaičius lygus duomenų rinkiniui pasirinktam klasterių skaičiui, t. y. $s = k$.

4.7. Ketvirtojo skyriaus apibendrinimas ir išvados

Atlikus daugiamačių duomenų transformacijos eksperimentinius tyrimus gautos šios išvados:

- Pločio parametro σ tinkamumą duomenų rinkiniui galime įvertinti vizualiai pagal transformuotų taškų Z_i projekciją, gautą MDS metodu.
- Pločio parametro σ nustatymas pagal maksimalų atstumą tarp klasterio centrų ir duomenų rinkinyje esančių klasterių skaičių k tinkamas ne visiems duomenų rinkiniams.
- Pasiūlytas konstantos α radimas pagal taškų išsibarstymą kiekviename klasteryje leidžia nustatyti tinkamą pločio parametro σ reikšmę. Su šia σ reikšme atlikus n-mačių taškų X_i dimensijos mažinimą, gauti taškai Z_i išsibarsto intervale [0; 1], t. y. nesikoncentruoja šio intervalo kraštuose.
- Tinkama pločio parametro σ reikšmė Gausinei funkcijai yra kelis kartus didesnė nei eksponentinei funkcijai.

Atlikus REGM tinklo eksperimentinius tyrimus nustatyta, kad REGM tinklas kokybiškiau apmokomas ir mažajame sluoksnyje gauti vizualizavimo rezultatai atitinka užsibrėžtus vizualizavimo kokybės kriterijus, kai:

- norimomis tinklo atsako reikšmėmis imami radialinėmis bazinėmis funkcijomis transformuoti klasterių centrai ir radialinių bazinių funkcijų sluoksnyje naudojama eksponentinė, o ne Gausinė funkcija;
- mažajame sluoksnyje naudojama tiesinė aktyvavimo funkcija, o pirmame paslėptame ir išėjimo sluoksnyje naudojama loginio sigmoido aktyvavimo funkcija;
- išėjimo sluoksnyje parinktas neuronų skaičius yra lygus pasirinktam klasterių skaičiui.

5. Apibendrinimas ir bendrosios išvados

Atlikta analitinė hibridinių neuroninių tinklų (įvairūs radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginiai) apžvalga parodė, kad tokio tipo tinklai konstruojami labai įvairiose srityse ir specifiniams uždaviniams spręsti. Hibridinių tinklų gaunami rezultatai yra tikslesni palyginus su radialinių bazinių funkcijų neuroninių tinklų arba daugiasluoksnių perceptronų gaunamais rezultatais. Konkrečiam uždaviniui spręsti kuriamo hibridinio neuroninio tinklo struktūra pasirenkama pagal atskirų tinklų individualias charakteristikas.

Disertacijoje yra pasiūlytas hibridinis neuroninis tinklas REGM, kuris savyje integruoja ir radialinių bazinių funkcijų neuroninio tinklo, ir daugiasluoksnio perceptrono, turinčio „butelio kaklelio“ neuroninio tinklo savybes, idėjas. Tinklas sudarytas iš dviejų dalių. Pirmoji dalis yra tam tikras daugiamatės erdvės taškų transformavimas į norimo mažesnio matmens erdvę. Antroji dalis yra daugiasluoksnis perceptronas, kurio mažasis sluoksnis (paskutinis paslėptas sluoksnis) sudarytas iš nedidelio neuronų skaičiaus (2 arba 3). REGM tinklo paskirtis yra padėti atskleisti duomenyse esančių klasterių savybes.

REGM tinklas naudojamas vizualiai daugiamačių duomenų analizei, kai atidėjimui plokštumoje arba trimatėje erdvėje taškai gaunami paskutinio paslėpto neuronų sluoksnio išėjimuose į tinklą padavus n-mačių analizuojamų duomenų rinkinį. Šio tinklo ypatybė yra ta, kad gautas vaizdas plokštumoje labiau atspindi bendrą duomenų struktūrą (klasteriai, klasterių tarpusavio artumas, taškų tarpklasterinis panašumas) nei daugiamačių taškų tarpusavio išsidėstymą.

Iš atliktų tyrimų buvo padarytos tokios išvados:

1. REGM tinklas yra nauja efektyvi priemonė daugiamačiams duomenims vizualiai tirti, nes atsiranda galimybė geriau pažinti bendrą duomenų struktūrą. Daugiamačių duomenų klasterizavimo rezultatai gali būti panaudojami ne tik apskaičiuojant radialinių bazinių funkcijų parametrus, bet ir vizualiai pateikiant rezultatus plokštumoje.

2. Jei radialinių bazinių funkcijų (RBF) pločio parametras apskaičiuojamas pagal objektų išsibarstymą klasteriuose ir vidutinį atstumą tarp tų klasterių centrų, tai RBF išėjime gaunamos reikšmės išsibarsto intervale [0; 1], t. y. nesikoncentruoja šio intervalo kraštuose.

3. REGM tinklą apmokius keletą kartų, geriausios duomenų rinkinio projekcijos pasirinkimą palengvina pasiūlyti du atrankos kriterijai, kuriuos naudojant atranka gali būti automatizuota:

- klasterių išsaugojimas duomenyse kriterijus, kurio reikšmė yra sveikas skaičius ir idealiu atveju lygus 0;
- išėjimų sluoksnyje gautų taškų išsibarstymo kriterijus, kurio reikšmė yra didžiausias atstumas tarp skirtingiems klasteriams priklausančių taškų.

4. Mažajame sluoksnyje gauta daugiamačių duomenų projekcija yra įvertinama trimis vizualizavimo kokybės kriterijais:
   - taškų išsidėstymas tiesių ar kreivių aplinkoje;
   - taškų „išsibarstymas“ klasteryje (didžiausias atstumas tarp klasterio taškų turi būti didesnis už 0,1);
   - riba tarp klasterių (mažiausias atstumas tarp skirtingiems klasteriams priklausančių taškų turi būti didesnis arba lygus 0,05).

5. Disertacijoje visi eksperimentiniai tyrimai su REGM tinklu atlikti naudojant praktinę svarbą turinčius realius duomenų rinkinius, kurių apimtis siekia 4500 objektų. Gautose projekcijose matomi ne tik duomenų rinkinį sudarantys klasteriai, bet ir tarpklasteriniai objektų panašumai/skirtingumai. Skirtinguose klasteriuose esantys, bet panašumų turintys objektai, tyrėjui padeda atkreipti dėmesį į galimus esminius pakitimus objektų savybėse (pavyzdžiui, ankstyvą ligos stadiją arba rūšių panašumus) arba ieškoti priežasčių, dėl kurių atsiranda pakitimai.

6. Hibridinis neuroninis tinklas REGM kokybiškiau apmokomas ir mažajame sluoksnyje gauti vizualizavimo rezultatai atitinka užsibrėžtus vizualizavimo kokybės kriterijus, kai:
   - norimomis tinklo atsako reikšmėmis imami radialinėmis bazinėmis funkcijomis transformuoti klasterių centrai ir radialinių bazinių funkcijų sluoksnyje naudojama eksponentinė, o ne Gausinė funkcija;
   - mažajame sluoksnyje naudojama tiesinė aktyvavimo funkcija, o pirmame paslėptame ir išėjimo sluoksnyje naudojama loginio sigmoido aktyvavimo funkcija;
   - išėjimo sluoksnyje parinktas neuronų skaičius yra lygus pasirinktam klasterių skaičiui.

Literatūra

Abdi, H. and L. Williams (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics 2*(4), 433–459.

Agrawal, R., J. Gehrke, D. Gunopulos, and P. Raghavan (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 94–105.

Araki, Y., T. Ohki, D. Citterio, M. Hagiwara, and K. Suzuki (2003). A new method for inverting feedforward neural networks. In *IEEE International Conference on Systems, Man and Cybernetics, 2003*, Volume 2, pp. 1612–1617.

Ataer-Cansizoglu, E., E. Bas, J. Kalpathy-Cramer, G. Sharp, and D. Erdogmus (2013). Contour-based shape representation using principal curves. *Pattern Recognition 46*(4), 1140–1150.

Bache, K. and M. Lichman (2013). UCI machine learning repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Baldi, P. and K. Hornik (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks 2*(1), 53–58.

Benoudjit, N. and M. Verleysen (2003). On the kernel widths in radial-basis function networks. *Neural Processing Letters 18*(2), 139–154.

Bernatavičienė, J. (2008). *Vizualios žinių gavybos metodologija ir jos tyrimas*. Daktaro disertacija, VGTU, MII.

Borg, I. and P. Groenen (2005). *Modern Multidimensional Scaling: Theory and Applications, 2nd edn.* Springer, New York.

Broomhead, D. and D. Lowe (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. *Complex System 2*, 321–355.

Buhmann, M. (2003). *Radial Basis Functions: Theory and Implementations*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Chaiyaratana, N. and A. M. S. Zalzala (1998). Evolving hybrid RBF-MLP networks using combined genetic/unsupervised/supervised learning. In *UKACC International Conference on Control '98. (Conf. Publ. No. 455)*, Volume 1, pp. 330–335.

Chang, Q., Q. Chen, and X. Wang (2005). Scaling gaussian RBF kernel width to improve svm classification. In *International Conference on Neural Networks and Brain, 2005. ICNN B '05*, Volume 1, pp. 19–22.

Charytanowicz, M., J. Niewczas, P. Kulczycki, P. Kowalski, S. Łukasik, and S. Żak (2010). Complete gradient clustering algorithm for features analysis of X-ray images. In *Information Technologies in Biomedicine*, pp. 15–24. Springer.

Chen, S., C. Cowan, and P. Grant (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks 2*(2), 302–309.

Chen, S., B. Mulgrew, and P. Grant (1993). A clustering technique for digital communications channel equalization using radial basis function networks. *IEEE Transactions on Neural Networks 4*(4), 570–590.

Comrey, A. and H. Lee (2013). *A First Course in Factor Analysis*. Psychology Press.

Delicado, P. (2001). Another look at principal curves and surfaces. *Journal of Multivariate Analysis 77*(1), 84–116.

DeMers, D. and G. Cottrell (1993). Non-linear dimensionality reduction. *Advances in Neural Information Processing Systems 5*, 580–580.

Duch, W. (2004a). Visualization of hidden node activity in neural networks: I. visualization methods. In *Artificial Intelligence and Soft Computing-ICAISC 2004*, pp. 38–43. Springer.

Duch, W. (2004b). Visualization of hidden node activity in neural networks: II. application to RBF networks. In *Artificial Intelligence and Soft Computing-ICAISC 2004*, pp. 44–49. Springer.

Duda, R. and P. Hart (1973). *Pattern Recognition and Scene Analysis*. Wiley, New York.

Dunham, M. H. (2002). *Data Mining: Introductory and Advanced Topics*. Prentice Hall PTR.

Dzemyda, G., O. Kurasova, ir J. Žilinskas (2008). *Daugiamačių duomenų vizualizavimo metodai*. Mokslo Aidai.

Dzemyda, G., O. Kurasova, and J. Žilinskas (2013). *Multidimensional Data Visualization: Methods and Applications*. Springer Optimization and Its Applications, Vol. 75. Springer.

Ester, M., H. Kriegel, J. Sander, and X. Xu (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In E. Simoudis, J. Han, and U. Fayyad (Eds.), *Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231. AAAI Press.

Fahlman, S. E. and C. Lebiere (1990). The cascade-correlation learning architecture. In *Advances in Neural Information Processing Systems 2*, pp. 524–532. Morgan Kaufmann.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2), 179–188.

France, S. and J. Carroll (2011). Two-way multidimensional scaling: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 41*(5), 644–661.

Gaur, D. and S. Gaur (2013). Comprehensive analysis of data clustering algorithms. In H.-K. Jung, J. T. Kim, T. Sahama, and C.-H. Yang (Eds.), *Future Information Communication Technology and Applications*, Volume 235 of *Lecture Notes in Electrical Engineering*, pp. 753–762. Springer Netherlands.

Guha, S., R. Rastogi, and K. Shim (1999). ROCK: a robust clustering algorithm for categorical attributes. In *15th International Conference on Data Engineering, 1999*, pp. 512–521.

Gupta, M. and Y. Chen (2011). *Theory and Use of the EM Algorithm*. Now Publishers Inc.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter 11*(1), 10–18.

Han, J., M. Kamber, and J. Pei (2011). *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)* (3rd ed.). Morgan Kaufmann.

Harman, H. (1976). *Modern Factor Analysis*. University of Chicago Press.

Hastie, T. and W. Stuetzle (1989). Principal curves. *Journal of the American Statistical Association 84*(406), 502–516.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Prentice Hall.

Hinneburg, A., E. Hinneburg, and D. Keim (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In *4th International Conference in Knowledge Discovery and Data Mining (KDD 98)*, pp. 58–65.

Horng, S., M. Su, Y. Chen, T. Kao, R. Chen, J. Lai, and C. Perkasa (2011). A novel intrusion detection system based on hierarchical clustering and support vector machines. *Expert Systems with Applications 38*(1), 306–313.

Horton, P. and K. Nakai (1996). A probabilistic classification system for predicting the cellular localization sites of proteins. In *Fourth International Conference on Intelligent Systems for Molecular Biology*, Volume 4, pp. 109–115.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*(6), 417–441.

Ivanikovas, S. (2010). *Lygiagrečių skaičiavimų taikymo daugiamačiams duomenims vizualizuoti problemos*. Daktaro disertacija, MII.

Ivanikovas, S., V. Medvedev, and G. Dzemyda (2007). Parallel realizations of the SAMMAN algorithm. In *Algorithm, International Conference on Adaptive and Natural Computing Algorithms – ICANNGA 2007*, Volume 4432 of *Lecture Notes in Computer Science*, pp. 179–188. Springer.

Izenman, A. (2008). *Linear Discriminant Analysis*. Springer.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31*(8), 651–666.

Johnson, B., R. Tateishi, and N. Hoan (2013). A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. *International Journal of Remote Sensing 34*(20), 6969–6982.

Jolliffe, I. (2005). *Principal Component Analysis*. Wiley Online Library.

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons.

Kanungo, T., D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*(7), 881–892.

Karbauskaitė, R. (2010). *Daugiamačių duomenų vizualizavimo metodų, išlaikančių lokalią struktūrą, analizė*. Daktaro disertacija, VDU, MII.

Karbauskaitė, R. and G. Dzemyda (2006). Multidimensional data projection algorithms saving calculations of distances. *Information Technology and Control 35*(1), 57–64.

Karypis, G., E. Han, and V. Kumar (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer 32*(8), 68–75.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley and Sons.

Kohonen, T. (2001). *Self-Organizing Maps, 3rd edn.* Springer Series in Information Science, Vol. 30. Springer.

Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*(1), 1–27.

Kurasova, O. (2005). *Daugiamačių duomenų vizuali analizė taikant savireguliuojančius neuroninius tinklus*. Daktaro disertacija, MII.

Lang, K. and M. Witbrock (1988). Learning to Tell Two Spirals Apart. In D. Touretzky, G. Hinton, and T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School*, pp. 52–59. Morgan Kaufmann.

Li, B. and Y. Zhang (2011). Supervised locally linear embedding projection (SLLEP) for machinery fault diagnosis. *Mechanical Systems and Signal Processing 25*(8), 3125–3134.

Little, M., P. McSharry, E. Hunter, J. Spielman, and L. Ramig (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering 56*(4), 1015–1022.

Liu, Q., M. Deng, Y. Shi, and J. Wang (2012). A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Computers & Geosciences 46*, 296–309.

Liu, X. and H. Liu (2006). A new CLARANS algorithm based on particle swarm optimization. In *The Sixth IEEE International Conference on Computer and Information Technology, 2006. CIT '06*, pp. 12–12.

Lowe, D. (1989). Adaptive radial basis function nonlinearities, and the problem of generalisation. In *First IEE International Conference on Artificial Neural Networks, 1989. (Conf. Publ. No. 313)*, pp. 171–175.

Lu, B. (2000). *Wireline Channel Estimation and Equalization*. Ph. D. thesis, University of Texas at Austin.

Lu, B. and B. Evans (1999). Channel equalization by feedforward neural networks. In *International Symposium on Circuits and Systems (ISCAS 1999)*, pp. 587–590. IEEE.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In L. Lecam and J. Neyman (Eds.), *The Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 281–297. University of California Press.

Mangasarian, O., W. Street, and W. Wolberg (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research 43*(4), 570–577.

Mao, J. and A. Jain (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks 6*(2), 296–317.

Mao, J. and A. Jain (1996). A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Transactions on Neural Networks 7*(1), 16–29.

Marcinkevičius, V. (2010). *Netiesinės daugiamačių duomenų projekcijos metodų savybių tyrimas ir funkcionalumo gerinimas*. Daktaro disertacija, VDU, MII.

McCulloch, W. and W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics 5*(4), 115–133.

Medvedev, V. (2007). *Tiesioginio sklidimo neuroninių tinklų taikymo daugiamačiams duomenims vizualizuoti tyrimai*. Daktaro disertacija, VGTU, MII.

Medvedev, V. ir G. Dzemyda (2005). Vizualizavimui skirto neuroninio tinklo mokymosi greičio optimizavimas. *Lietuvos matematikos rinkinys 45*, 426–431.

Ng, R. and J. Han (2002). CLARANS: a method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering 14*(5), 1003–1016.

Park, H. and C. Jun (2009). A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications 36*(2), 3336–3341.

Passos, M., H. Fernandes, and P. d. F. Silva (2007). Applications of modular RBF/MLP neural networks in the modeling of microstrip photonic bandgap structures. *PIERS Online 3*(5), 695–700.

Passos, M., P. d. F. Silva, and H. Fernandes (2006). A RBF/MLP modular neural network for microwave device modeling. *International Journal of computer science and network security 6*(5A), 81–86.

Patidar, A., R. Joshi, and S. Mishra (2011). Implementation of distributed ROCK algorithm for clustering of large categorical datasets and its performance analysis. In *3rd International Conference on Electronics Computer Technology (ICECT), 2011*, Volume 2, pp. 79–83.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 2*(11), 559–572.

Pierrefeu, L., J. Jay, and C. Barat (2006). Auto-adjustable method for gaussian width optimization on RBF neural network. Application to face authentication on a monochip system. In *32nd Annual Conference on IEEE Industrial Electronics, IECON 2006*, pp. 3481–3485.

Podpečan, V., M. Zemenova, and N. Lavrač (2012). Orange4WS environment for service-oriented data mining. *The Computer Journal 55*(1), 82–98.

Raudys, Š. (2008). *Žinių išgavimas iš duomenų*. Klaipėdos universiteto leidykla.

Rocha Neto, A., R. Sousa, G. Barreto, and J. Cardoso (2011). Diagnostic of pathology on the vertebral column with embedded reject option. In *Pattern Recognition and Image Analysis*, pp. 588–595. Springer.

Roweis, S. and L. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science 290*(5500), 2323–2326.

Rumelhart, D., G. Hintont, and R. Williams (1986). Learning representations by back-propagating errors. *Nature 323*(6088), 533–536.

Saad, Y. and M. Schultz (1988). Topological properties of hypercubes. *IEEE Transactions on Computers 37*(7), 867–872.

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers 18*(5), 401–409.

Street, W., W. Wolberg, and O. Mangasarian (1993). Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pp. 861–870. International Society for Optics and Photonics.

Sun, J., C. Fyfe, and M. Crowe (2012). Extending SAMMON mapping with bregman divergences. *Information Sciences 187*, 72–92.

Tenenbaum, J., V. De Silva, and J. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science 290*(5500), 2319–2323.

Thissen, U., W. Melssen, and L. Buydens (2001). Nonlinear process monitoring using bottle-neck neural networks. *Analytica Chimica Acta 446*(1), 369–381.

Verikas, A. ir A. Gelžinis (2008). *Neuroniniai tinklai ir neuroniniai skaičiavimai.* Technologija, Kaunas.

Vesanto, J. (2001). Importance of individual variables in the k-means algorithm. In D. Cheung, G. Williams, and Q. Li (Eds.), *Advances in Knowledge Discovery and Data Mining*, Volume 2035 of *Lecture Notes in Computer Science*, pp. 513–518. Springer Berlin Heidelberg.

Vesanto, J. and E. Alhoniemi (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks 11*(3), 586–600.

Wu, M., B. Chen, B. Gao, X. Cheng, and Z. Yan (2012). Dimensionality reduction method of training sample set for SVDD based on statistical information. *Applied Mechanics and Materials 220*, 2097–2101.

Yaglom, A. (1986). *Correlation Theory of Stationary and Related Random Functions I: Basic Results.* Springer Series in Statistics. Springer.

Yeh, I.-C., K.-C. Huang, and Y.-H. Kuo (2013). Spatial interpolation using MLP–RBFN hybrid networks. *International Journal of Geographical Information Science 27*(10), 1884–1901.

Yıldırım, A. and C. Özdoğan (2011). Parallel wavecluster: A linear scaling parallel clustering algorithm implementation with application to very large datasets. *Journal of Parallel and Distributed Computing 71*(7), 955–962.

Zalzala, A. M. S. and N. Chaiyaratana (2000). Myoelectric signal classification using evolutionary hybrid RBF-MLP networks. In *The 2000 Congress on Evolutionary Computation, 2000*, Volume 1, pp. 691–698.

Zhang, H. and X. Liu (2011). A CLIQUE algorithm using DNA computing techniques based on closed-circle DNA sequences. *Biosystems 105*(1), 73–82.

Zhang, T., R. Ramakrishnan, and M. Livny (1996). BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Record 25*(2), 103–114.

Žilinskas, A. and J. Žilinskas (2008). A hybrid method for multidimensional scaling using city-block distances. *Mathematical Methods of Operations Research 68*(3), 429–443.

Autorės publikacijų sąrašas disertacijos tema

Ringienė, L., Dzemyda, G. Daugiamačių duomenų požymių mažinimas naudojantis eksponentine koreliacine funkcija. *Jaunųjų mokslininkų darbai*. Vilnius: Vilniaus universitetas. ISSN 2029-9958. 2013, Nr. 1, p. 152–158.

Ringienė, L., Dzemyda, G. Multidimensional data visualization based on the exponential correlation function. *Baltic Journal of Modern Computing*. Riga: University of Latvia. ISSN 2255-8942. 2013, Vol. 1, No. 1, p. 9–28.

Ringienė, L., Dzemyda, G. Specialios struktūros daugiasluoksnis perceptronas daugiamačiams duomenims vizualizuoti. *Informacijos mokslai*. ISSN 1392-0561. 2009, T. 50, p. 358–364.

Laura Ringienė

HIBRIDINIS NEURONINIS TINKLAS DAUGIAMAČIAMS DUOMENIMS VIZUALIZUOTI

Daktaro disertacija
Technologijos mokslai,
informatikos inžinerija (07 T)

Laura Ringienė

HYBRID NEURAL NETWORK FOR MULTIDIMENSIONAL DATA VISUALIZATION

Doctoral Dissertation
Technological Sciences,
Informatics Engineering (07 T)