

Specialios struktūros daugiasluoksnis perceptronas daugiamatiams duomenims vizualizuoti

Pasiūlytas ir ištirtas radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginys daugiamatiams duomenims vizualizuoti. Siūlomas vizualizavimo būdas apima daugiamatį duomenų matmenų mažinimą naudojant radialines bazines funkcijas, daugiamatį duomenų suskirstymą į klasterius, klasterį charakterizuojančių skaitinių reikšmių nustatymą ir daugiamatį duomenų vizualizavimą dirbtinio neuroninio tinklo paskutiniame paslėptajame sluoksnyje.

1. Įvadas

Daugiamatiams duomenis analizuoti yra sukurta daug įvairių metodų: klasifikavimo, klasterizavimo, statistinės analizės ir kt. Jais galima nustatyti stebimų duomenų taškų ar jų grupių artumą, sudaryti taisykles, pagal kurias tokio tipo duomenys būtų rūšiuojami, vertinti atskirų parametrų įtaką daromam sprendimui. Svarbią vietą duomenų analizėje užima vizualizavimas. Didelio matmenų skaičiaus duomenų vizualizavimas leidžia geriau suvokti sudėtingas duomenų aibes, padeda nustatyti išskirtinius jų poaibius. Siekiant gauti kuo daugiau naujų žinių apie analizuojamus duomenis, bandoma net sujungti kelis skirtingais principais grindžiamus vizualizavimo metodus (Dzemyda, Kurasova, Medvedev, 2007).

Daugiamatiai duomenys – tai objektai (žmonės, įrenginiai, augalai, gamtos reiškiniai), kuriuos charakterizuoja faktiniai parametrai, dar vadinami požymiais, savybėmis, rodikliais, ypatybėmis. Objektų skaičius m yra baigtinis. Tam tikras parametrų reikšmių rinkinys nusako vieną konkretų analizuojamos aibės objektą $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $i = \overline{1, m}$, čia n yra parametrų skaičius, i yra objekto numeris. Objektai $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$ dar vadinami vektoriais ar taškais, o parametrai $x_1, x_2, \dots, x_n$ – komponentėmis ar požymiais. Analizuojamų duomenų aibę galima atvaizduoti kaip matricą $\mathbf{X} = \{X_1, X_2, \dots, X_m\} = \{x_{ij}, i = \overline{1, m}, j = \overline{1, n}\}$, kurios i -oji eilutė yra taškas $X_i \in R^n$ (Dzemyda, Kurasova, Žilinskas, 2008).

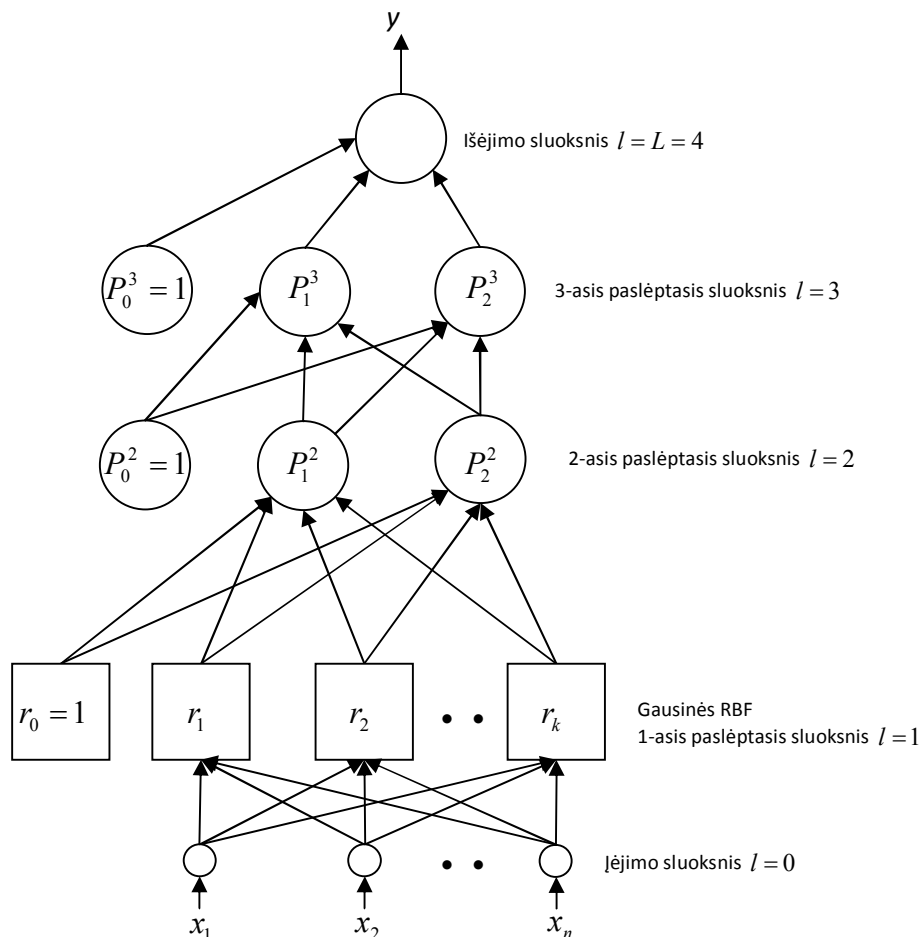
Darbe pasiūlytas ir ištirtas specialios struktūros daugiasluoksnis dirbtinis neuroninis tinklas (DNT), kurio pirmasis paslėptasis sluoksnis yra radialinės bazinės funkcijos, o likusioji tinklo dalis daugiasluoksnis perceptronas, kuris paprastai apmokomas klaidos skleidimo atgal algoritmu. Idėja: paskutiniojo paslėptojo sluoksnio išėjimų reikšmių vizualus pateikimas.

2. Daugiamatį duomenų vizualizavimas naudojant DNT paslėptojo sluoksnio išėjimus

2.1. Tinklo struktūra ir idėja

Darbe konstruojamas tiesioginio sklaidimo daugiasluoksnis DNT, skirtas mokymui su mokytoju, t. y., kai iš anksto žinomos norimos išėjimo reikšmės. Tinklo specifikacija: pirmasis paslėptasis sluoksnis susideda iš gausinių radialinių bazinių funkcijų (RBF), kurių yra tiek, kiek spėjama, kad daugiamatčiuose duomenyse yra klasterių; paskutinis paslėptasis sluoksnis susideda iš dviejų neuronų, jei norime gautus išėjimus pavaizduoti dvimatėje

erdvėje, arba iš trijų neuronų, jei norime gautus išėjimus pavaizduoti trimatėje erdvėje. 1 pav. pateikiamas tyrimams naudotas tinklo atvejis, kuomet buvo trys paslėpti sluoksniai, įskaitant ir pirmąjį su RBF, o paskutinio paslėptojo sluoksnio neuronų skaičius lygus 2. Bendru atveju paslėptųjų neuronų sluoksnių skaičius L yra neribojamas, $l = 0, 1, \dots, L$, čia sluoksnis $l = 0$ žymi įėjimus, o $l = L$ – paskutinįjį (išėjimų) sluoksnį. Kiekviename sluoksnyje l gali būti n_l neuronų (Dzemyda, Kurasova, Žilinskas, 2008).



1 pav. Tiesioginio sklido DNT, naudotas tyrimams

Tinklo mokymas vyksta dviem etapais: pirmo etapo metu mokomas RBF sluoksnis; antrojo etapo metu mokomas daugiasluoksnis perceptronas.

Pirmojo etapo metu atliekamas daugiamačių duomenų $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $i = \overline{1, m}$, kur $X_i \in R^n$, matmenų mažinimas, transformuojant $X_i \in R^n$ į $R_i \in R^k$: $R_i = (r_{i1}, r_{i2}, \dots, r_{ik})$, kur $k < n$. Matmenys mažinami pasinaudojus gausine RBF, kuri apskaičiuojama pagal formulę:

$$r_j(X_i) = \exp\left(-\frac{\|X_i - \mu_j\|^2}{2\delta^2}\right),$$

čia μ_j yra radialinės bazinės funkcijos r_j centro taškas, $\mu_j \in R^n$, $\|X_i - \mu_j\|$ – atstumas tarp taškų X_i ir μ_j , $i = \overline{1, m}$, $j = \overline{1, k}$, δ – pločio parametras, nuo kurio priklauso funkcijos glotnumas.

Antrojo etapo metu DNT mokomas klaidos skleidimo atgal algoritmu (ang. *error back-propagation*). Neuronų aktyvacijos funkcija – netiesinis loginis sigmoidas $f(a) = \frac{1}{1+e^{-a}}$. Tinklo mokymo duomenys yra RBF išėjimuose

gauti taškai $R_i, i = \overline{1, m}$. DNT mokymo metu keičiamos neuronų perdavimo koeficientų (svorių) taškų

$W = \{w_{ij}^l, i = \overline{0, n_l}, j = \overline{1, n_{l+1}}\}$ (čia w_{ij}^l – sluoksnio jungtis tarp l -tojo sluoksnio i -tojo neurono ir $l+1$ sluoksnio j -tojo neurono) reikšmės ir siekiama gauti kiek galima mažesnę paklaidą (Dzemyda, Kurasova, Žilinskas, 2008)

$$E(W) = \frac{1}{2} \sum_{i=1}^m (y_i - t_i)^2,$$

čia y_i – gautos tinklo išėjimo reikšmės, t_i – norimos tinklo išėjimo reikšmės.

Pasiūlyto DNT taikymas vizualizavimui:

1. Pasirenkamas spėjamas skaičius k klasterių, kuriuos sudaro aibės X taškai.
2. Vykdomas aibės X taškų klasterizavimas į k klasterių $K_j, j = \overline{1, k}$.
3. Naudojantis klasterizavimo rezultatais: a) apskaičiuojami RBF parametrai; b) kiekvienam klasteriui priskiriama tam tikra skaitinė reikšmė $\gamma_j, j = \overline{1, k}$.
4. Transformuojami aibės X taškai $X_i = (x_{i1}, x_{i2}, \dots, x_{in}) \in R^n$ į taškus $R_i = (r_{i1}, r_{i2}, \dots, r_{ik}) \in R^k$.
5. Daugiasluoksnio perceptrono mokymas naudojant taškus $R_i = (r_{i1}, r_{i2}, \dots, r_{ik}), i = \overline{1, m}$, kaip įėjimo reikšmes ir γ_j , kaip norimą tinklo reakciją į $R_i (X_i \in K_j)$
6. Vizualizuojamos DNT $L-1$ sluoksnyje gautos reikšmės (1 pav. atveju, tai neuronų P_1^3 ir P_2^3 išėjimai).

2.2. Daugiamačių duomenų suskirstymas į klasterius ir klasterių charakterizuojančių skaitinių reikšmių γ_j nustatymas

Perceptrono tipo tinklo mokymui su mokytoju reikia žinoti norimas tinklo išėjimo reikšmes. Mūsų atveju tinklo išėjimo reikšmė yra konkretų klasterį K_j , kuriam priskiriamas duomenų taškas $X_i (X_i \in K_j)$, atitinkanti tam tikra skaitinė charakteristika.

Pradinius daugiamačius duomenis galima suskirstyti į nurodytą klasterių skaičių k , pasinaudojant kuriuo nors klasterizavimo metodu. Šiame straipsnyje naudotas k -vidurkių (ang. *k-means*) metodas. Šis klasterizavimo metodas dėl savo paprastumo ir greitumo yra dažnai naudojamas daugiamačiams duomenims skirstyti į klasterius (Frahling, Sohler, 2006). k -vidurkių metodas minimizuoja kvadratinę paklaidą (Kurasova, 2005)

$$E_K = \min \sum_{j=1}^k \sum_{X_i \in K_j} \|X_i - \mu_j\|^2,$$

čia $\|X_i - \mu_j\|$ atstumas tarp kiekvieno duomenų taško X_i ir klasterio, kuriam jis priklauso, centro μ_j , k – klasterių skaičius.

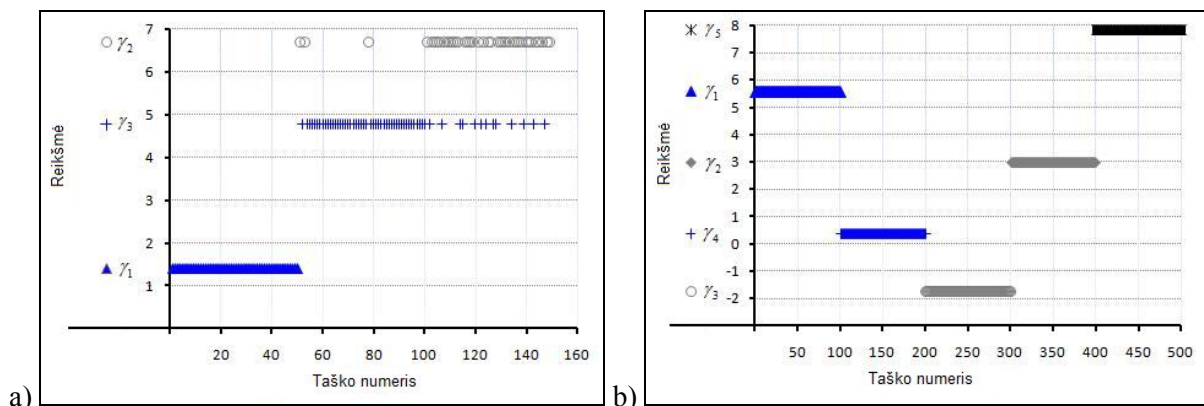
Klasterizavimo rezultate duomenų aibę $X_i, i = \overline{1, m}$ suskirstome į k klasterių $K_j, j = \overline{1, k}$. Viena kiekvieno klasterio skaitinė charakteristika yra jo centras μ_j , ji yra n -matė, t. y. $\mu_j \in R^n$. Kita skaitinė charakteristika – klasterio numeris.

1 pav. parodytame DNT yra vienas išėjimas. Naudojantis tinklu, į jį padavus $X_i \in R^n$, norėtusi išėjime gauti kažkokį skaičių $\gamma_j \in R^1$, kuris atspindėtų to taško priklausomybę j -tajam klasteriui. Klasterio numeris negali būti tokia charakteristika, nes jis neatspindi taškų X_i tarpusavio išsidėstymo n -matėje erdvėje. Klasterių centrai tam tikra forma atspindi taškų išsidėstymą, tačiau jie patys yra n -mačiai.

$\gamma_j \in R^1$ iš $\mu_j \in R^n$ galima gauti naudojantis projekcijos metodais (pagrindinių komponentų analize (ang. *principal component analysis*), tiesine diskriminantine analize (ang. *linear discriminant analysis*), daugiamačėmis skalėmis (ang. *multidimensional scaling*) ir kt.) (Dzemyda, Kurasova, Žilinskas, 2008).

Šiame straipsnyje taikomas daugiamačių skalių (MDS) metodas, kuriuo nustatomi atitikmenys tarp $X_i \in R^n$, $X_i \in K_j$, ir $\gamma_j \in R^1$.

2 pav. pateiktos $\gamma_j \in R^1$ reikšmės, atitinkančios n -mačius klasterių centrus $\mu_j \in R^n, j = \overline{1, k}$, o taip pat tas reikšmes atitinkančius taškus $X_i, i = \overline{1, m}$ numeriai, t. y. čia parodyta, koks turėtų būti neuroninio tinklo atsakas į tašką X_i .



2 pav. Klasterių centrų atvaizdavimas tiesėje: a) irisų duomenys b) atsitiktinai generuoti duomenys

3. Pasiūlyto dirbtinio neuroninio tinklo tyrimas

DNT tyrimai buvo atliekami eksperimentiškai. Eksperimentai atlikti su dviem daugiamačių duomenų aibėm:

1. Gėlių irisų duomenys (Iris Plants Database). Irisų duomenys buvo surinkti 1936 m. R. Fisher. Duomenų bazėje pateikiami trijų rūšių irisai – Setosa, Versicolour ir Virginica. Kiekvienos klasės yra po 50 duomenų taškų, iš viso 150 irisų duomenų taškų. Kiekvienas duomenų taškas sudarytas iš keturių parametų – taurėlapio ilgio, taurėlapio pločio, vainiklapio ilgio ir vainiklapio pločio. Ilgis ir plotis matuojamas centimetrais. (Fisher, 1936)
2. Atsitiktinai generuoti duomenys. Duomenys generuoti taip, kad sudarytų 5 klases. Kiekvienos klasės yra po 100 taškų, iš viso 500 duomenų taškų. Kiekvienas duomenų taškas sudarytas iš 10 parametų.

1 pav. pateiktas tinklas buvo mokomas aibės X taškais, išėjime siekiant gauti tuos taškus atitinkančias γ_j

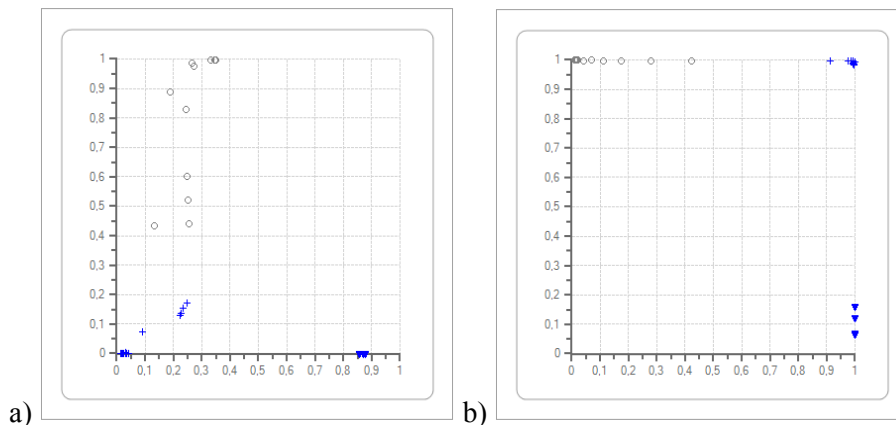
reikšmes. Vizualizavimo rezultatas – po mokymo gautos (P_1^3, P_2^3) reikšmės visiems aibės X taškams. Tyrime

buvo keičiamas RBF pločio parametras δ . Visos RBF naudojo tokį patį parametą δ . Kiekvieną kartą, prieš pradedant mokyti tinklą, atsitiktinai parenkami pradiniai svoriai, dėlto su kiekvienu δ buvo atlikta po 100 bandymų, iš kurių išrinktas geriausias tinklo apsimokymo rezultatas. Geriausiai apsimokiusiu tinklu vadinamas DNT, kurio daroma klaida iš atliktų eksperimentų po apsimokymo buvo mažiausia. 1 lentelėje pateiktos iš 100 atliktų bandymų su kiekvienu parametru δ atrinktos mažiausios daromos klaidos.

1 lentelė Tinklo mokyto irisų ir atsitiktinai generuotais duomenimis eksperimentų rezultatai

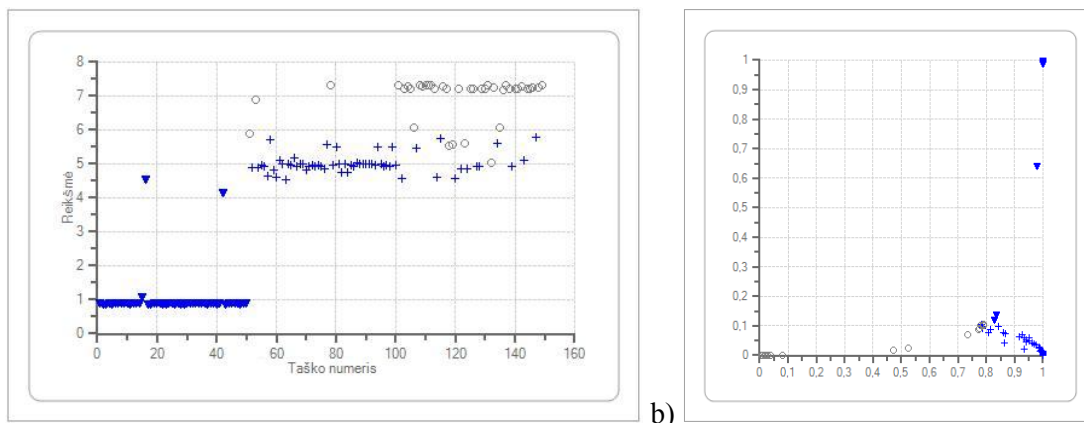
Irisų duomenys		Atsitiktinai generuoti duomenys	
Sigma	Klaida	Sigma	Klaida
2	0.00139	3.3	0.00996
1	0.00042	2	0.00345
0.67	0.00037	1	0.00292
0.5	0.00030	0.67	0.00813
0.4	0.00155	0.5	0.08028
0.3	0.00820	0.4	0.14905

Pagal 1 lentelėje pateiktus duomenis matyti, kad irisų duomenimis gerai apsimokė tinklas, kuriam parinktas $\delta = 0.5$. Nuo jo nedaug atsilieka tiklas, kuriam parinktas $\delta = 0.67$. Po tinklo apmokymo išėjimuose gauti taškai atitinka γ_j reikšmes (2a pav.). 3 pav. parodyti vizualizuoti $L-1$ sluoksnyje gauti išėjimai po tinklo apmokymo. Šiame paveiksle, kaip ir kituose, kur pateikiami vizualizavimo rezultatai, abscisių ašyje atidėtos P_1^3 reikšmės, o ordinačių ašyje atidėtos P_2^3 reikšmės, gautos pateikiant tinklui visus taškus X_i , $i = \overline{1, m}$.



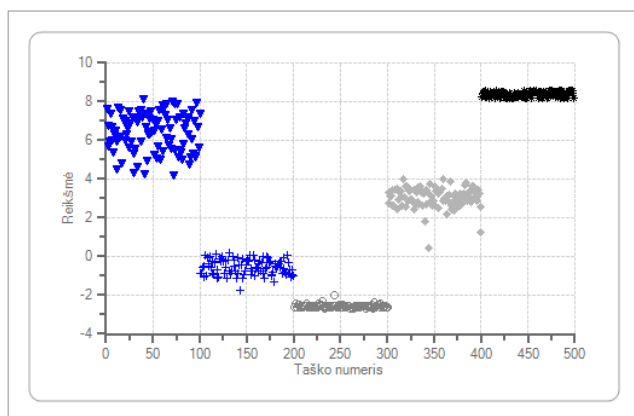
3 pav. Paslėptajame $L-1$ sluoksnyje gautos reikšmės: a) $\delta = 0.5$ b) $\delta = 0.67$

Iš 3 pav. pateiktų grafikų matyti, kad irisų duomenys aiškiai susiskirsto į tris klasterius. Taip atsitinka, tik gerai apsimokius tinklui. Pagal 1 lentelėje pateiktus tinklo apmokymo irisų duomenimis gautus rezultatus matome, kad prasčiausiai apsimokė tinklas, kuriam parinktas $\delta = 0.3$. 4a pav. matome, kad po DNT apmokymo tinklo išėjimo reikšmės yra susimaišiusios – dalis taškų priklausančių pirmajam ir antrajam klasteriui perėjo į trečiąjį klasterį. 4b pav. parodo, kad blogai parinkus parametą δ duomenys nesusiskirsto į klasterius.



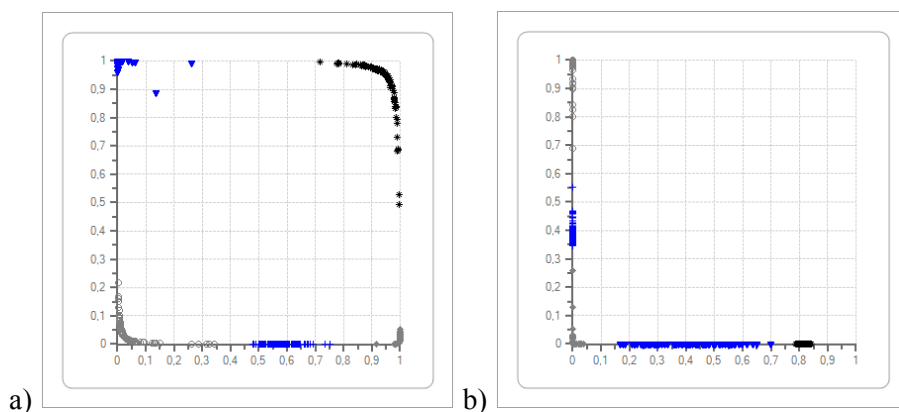
4 pav. $\delta = 0.3$: a) tinklo išėjime gautos reikšmės b) paslėptajame $L-1$ sluoksnyje gautos reikšmės

Pagal 1 lentelėje pateiktus atsitiktinai generuotų duomenų gautus rezultatus matome, kad gerai apsimokė tinklai, kurių $\delta = 2$ ir $\delta = 1$. Tačiau visais atvejais tinklas neapsimokė taip gerai kaip su irisų duomenimis. Iš 5 pav. matome, kad tinklo išėjimuose gautos reikšmės išsibarsčiusios (palyginus su 2b pav.), bet klasteriai tarpusavyje nesusimaišę.



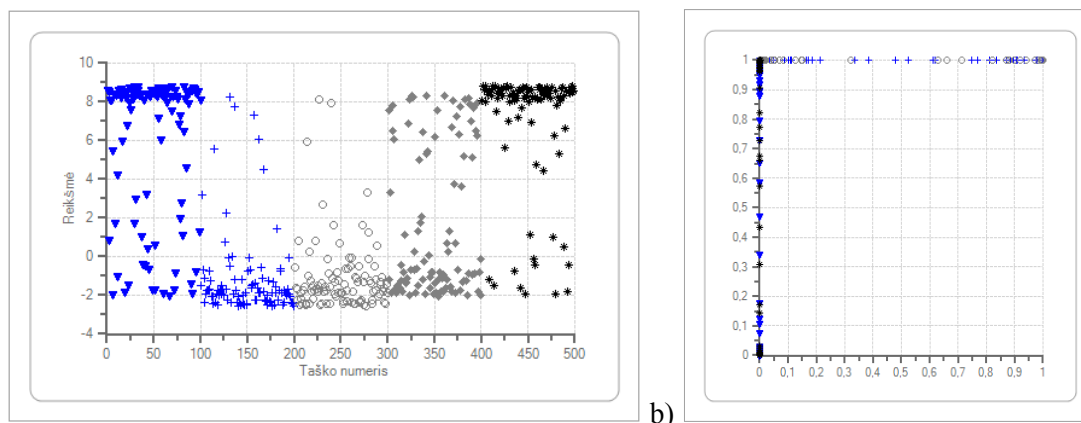
5 pav. Tinklo išėjimo reikšmės, kai $\delta = 1$

6 pav. parodyti vizualizuoti $L-1$ sluoksnyje gauti išėjimai po tinklo apmokymo, kai $\delta = 2$ ir $\delta = 1$.



6 pav. Paslėptajame $L-1$ sluoksnyje gautos reikšmės: a) $\delta = 2$ b) $\delta = 1$

Nors po tinklo apmokymo išėjimo reikšmės yra išsisklaidę, bet iš 6 pav. pateiktų grafikų matyti, kad atsitiktinai generuoti duomenys susiskirsto į klasterius. Pagal 1 lentelėje pateiktus duomenis matome, kad didžiausią klaidą tinklas daro, kai $\delta = 0.4$. Tinklo išėjime gautos reikšmės pateiktos 7a pav. Jos yra išsibarsčiusios intervale $[-3, 9]$ ir taškus sunku vizualiai priskirti klasteriams. 7b pav. pateiktos vizualizuotos gautos (P_1^3 ir P_2^3) reikšmės, čia taškai tarpusavyje susimaišę ir neišskiriamas nei vienas klasteris.



7 pav. $\delta = 0.4$: a) tinklo išėjime gautos reikšmės b) paslėptajame $L-1$ sluoksnyje gautos reikšmės

4. Išvados

Darbe pasiūlytas metodas daugiamačių duomenų vizualizavimui naudojant radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginį. Apmokius tokį tinklą, paskutiniojo paslėptojo sluoksnio neuronų išėjimo reikšmės yra laikomos įėjimo taško atvaizdu žemesnio matavimo erdvėje. Kiek tame sluoksnyje yra neuronų, į tokio matavimo erdvę galime projektuoti daugiamačius duomenis. Šiame straipsnyje eksperimentiškai nagrinėjamas atvejis, kai paslėptajame sluoksnyje yra 2 neuronai. Bendru atveju neuronų gali būti ir daugiau. Atlikti eksperimentai leidžia padaryti šias išvadas, kad dirbtinio daugiasluoksnio neuroninio tinklo apsimokymo kokybė labai priklauso nuo parinkto RBF parametro δ . Nustatyta, kad parinkus tinkamą δ reikšmę, tinklo paskutiniojo paslėptojo sluoksnio išėjimų reikšmės teisingai susiskirsto į vaizdžiai matomus klasterius.

Literatūros sąrašas:

- DZEMYDA, G.; KURASOVA, O.; MEDVEDEV, V. (2007). Dimension Reduction and Data Visualization Using Neural Networks. *Frontiers in Artificial Intelligence and Applications. Emerging Artificial Intelligence Applications in Computer Engineering – Real World AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, 2007, Vol. 160, p. 25-49.
- DZEMYDA, G.; KURASOVA, O.; ŽILINSKAS, J. (2008). Daugiamačių duomenų vizualizavimo metodai. Vilnius: Mokslo aidai, 2008. 204 p. ISBN 9789986680420
- FISHER, R. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 1936, Vol. 7, p. 179-188.
- FRAHLING, G.; SOHLER, Ch. (2006). A fast k-means implementation using coresets. *Proceedings of the twenty-second annual symposium on Computational geometry (SoCG)*. New York: The Association for Computing Machinery, 2006. p. 135-143. ISBN 1595933409
- KURASOVA, O. (2005). Daugiamačių duomenų vizuali analizė taikant savireguliuojančius neuroninius tinklus (daktaro disertacija). Vilnius: Matematikos ir Informatikos institutas, 2005
- MEDVEDEV, V.; DZEMYDA, G. (2004). Lygiagreti SAMANN vizualizavimo algoritmo realizacija. *Informacinės technologijos 2004, konferencijos pranešimo medžiaga*. Kaunas: Technologija, 2004. p. 344-350. ISBN 9955095881

Special multilayer perceptron for multidimensional data visualization

In this paper present, the special feed forward neural network which consisting of radial basis function layer and multilayer perceptron. This multilayer perceptron has been proposed and investigated for multidimensional data visualization. The proposed visualization approach includes data clustering, determining the parameters of radial basis function and forming the data set for training multilayer perceptron. The outputs of the last hidden layer considered as coordinates of visualized points.