Specialios struktūros daugiasluoksnis perceptronas daugiamačiams duomenims vizualizuoti

Laura Ringienė

Matematikos ir informatikos instituto, doktorantė
Institute of Mathematics and Informatics,
PhD student
Akademijos g. 4, LT-08663 Vilnius
Tel. (85) 210 93 22, fak. (85) 272 92 09
El. paštas: ringiene@ktl.mii.lt

Gintautas Dzemyda

Matematikos ir informatikos instituto
profesorius, habilituotas daktaras
Institute of Mathematics and Informatics,
Professor, Habil. Dr.
Akademijos g. 4, LT-08663 Vilnius
Tel. (85) 210 93 00, faks. (85) 272 92 09
El. paštas: dzemyda@ktl.mii.lt

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

Įvadas

Daugiamačiams duomenis analizuoti yra sukurta daug įvairių metodų: klasifikavimo, klasterizavimo, statistinės analizės ir kt. Jais galima nustatyti stebimų duomenų taškų ar jų grupių artimumą, 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 ir kt., 2007).

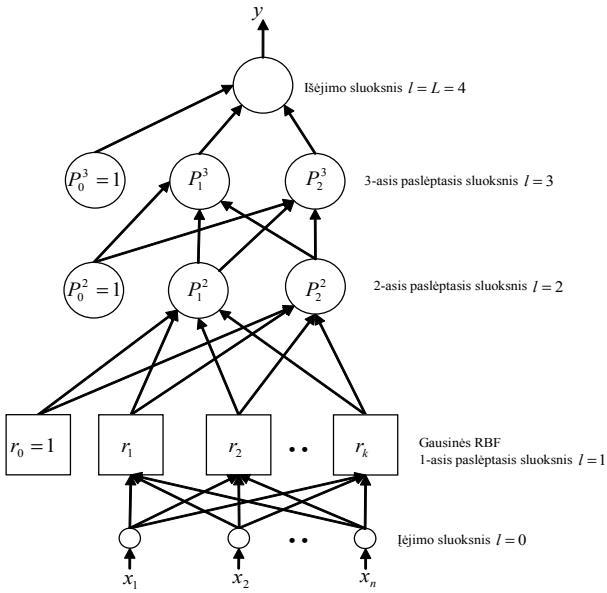
Daugiamačiai duomenys – tai objektai (žmonės, įrenginiai, augalai, gamtos reiškiniai), kuriuos charakterizuoja faktiniai parametrai, dar vadinami požymiais, savybėmis, rodikliais, ypatybėmis. Objektų skaičius *m* yra baigtinis. Tam tikras parametrų reikšmių rinkinys nusako vieną konkretų analizuojamos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

Daugiamačiams duomenims vizualizuoti plačiai naudojami ir dirbtiniai neuroniniai tinklai (DNT) (Kramer, 1991; Dzemyda ir kt., 2007; Vázquez Santacruz, Chakraborty, 2007). Šiame straipsnyje pasiūlytas ir ištirtas spacialios struktūros daugiasluoksnis 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.

1. Daugiamačių duomenų vizualizavimas naudojant DNT paslėptojo sluoksnio išėjimus

1.1. Tinklo struktūra ir idėja

Darbe konstruojamas tiesioginio sklidimo 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 daugiamačiuose 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, kai buvo trys paslėptieji 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, …, *L*; čia sluoksnis $l = 0$ žymi įėjimus, o $l = L$ – paskutinįjį (išėjimų) sluoksnį. Kiekviename sluoksnyje *l* gali būti n_l neuronų (Dzemyda ir kt., 2008).



1 p a v. ***Tiesioginio sklidimo DNT, naudotas tyrimams***

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

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

$$r_j(X_i) = \exp\left(-\frac{\| X_i - \mu_j \|^2}{2\delta^2}\right); \qquad (1)$$

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.

Antruoju etapu DNT mokomas klaidos skleidimo atgal algoritmu (angl. *error back-propagation*) (Hassoun, 1995). 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*-ojo sluoksnio *i*-ojo neurono ir *l*+1 sluoksnio *j*-ojo neurono) reikšmės ir siekiama gauti kuo mažesnę paklaidą (Dzemyda ir kt., 2008):

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

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

Pasiūlyto DNT taikymas duomenims vizualizuoti:

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 $X_i=(x_{i1},x_{i2},...,x_n)\in R^n$ į taškus $R_i=(r_{i1},r_{i2},...,r_k)\in R^k$.

5. Daugiasluoksnio perceptrono mokymas naudojant taškus $R_i=(r_{i1},r_{i2},...,r_k),\ i=\overline{1,m}$ 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).

1.2. Daugiamačių duomenų suskirstymas į klasterius ir klasterį 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*, naudojantis kuriuo nors klasterizavimo metodu. Šiame straipsnyje naudotas *k*-vidurkių (angl. *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):

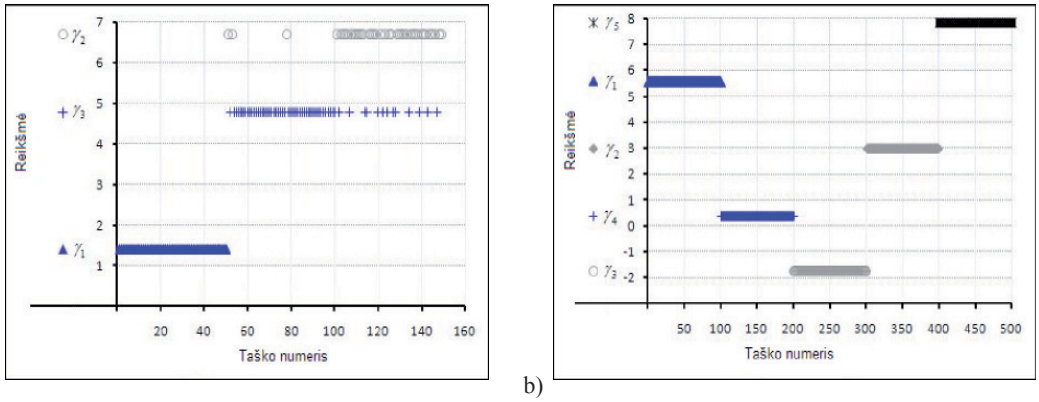
$$E_K=\min\sum_{j=1}^{k}\sum_{X_i\in K_j}\|\ X_i-\mu_j\ \|^2;\qquad(3)$$

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

Kaip klasterizavimo rezultatą 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.

Pirmame paveiksle parodytame DNT yra vienas išėjimas. Naudodamiesi tinklu ir į jį pateikę $X_i\in R^n$, išėjime norėtume gauti kokį nors skaičių $\gamma_j\in R^1$, kuris atspindėtų to taško priklausomybę *j*-ajam klasteriui. Klasterio numeris negali būti tokia charakteristika, nes jis nerodo taškų X_i tarpusavio išsidėstymo *n*-matėje erdvėje. Klasterių centrai tam tikra forma rodo 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 komponenčių analize (angl. *principal component analysis*), tiesine diskriminantine analize (angl. *linear discriminant analysis*), daugiamatėmis skalėmis (angl. *multidimensional scaling*) ir kt. (Dzemyda ir kt., 2008).



a) b)

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

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

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

2. Pasiūlyto dirbtinio neuroninio tinklo tyrimas

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

Gėlių irisų duomenys (*Iris Plants Database*) (Fisher, 1936). Duomenų rinkinį sudaro trijų rūšių irisai – *Setosa*, *Versicolor* ir *Virginica*. Kiekvienos rūšies yra po 50 irisų, iš viso 150 ($m = 150$). Kiekvienas irisą nusakantis duomenų taškas sudarytas iš keturių komponenčių – taurėlapio ilgio, taurėlapio pločio, vainiklapio ilgio ir vainiklapio pločio ($n = 4$).

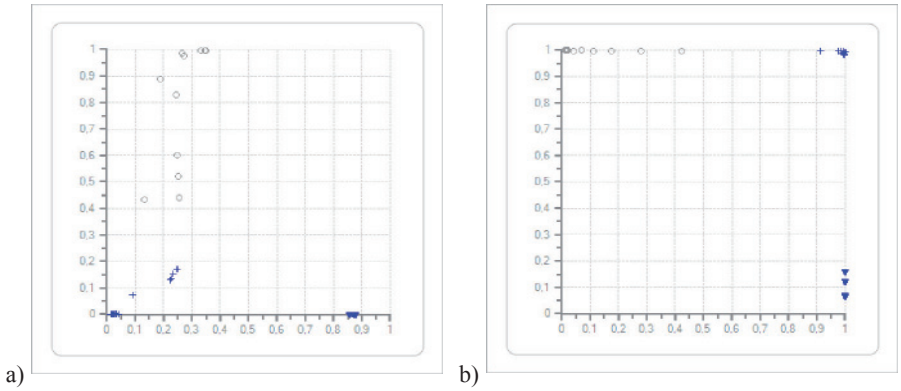
Atsitiktinai generuoti duomenys. Duomenys generuoti taip, kad sudarytų penkis klasterius po 100 taškų kiekviename klasteryje, iš viso 500 duomenų taškų ($m = 500$). Kiekvienas duomenų taškas sudarytas iš 10 komponenčių ($n = 10$). Klasterio, kuriam turi priklausyti generuojamas taškas, numeriu pažymėtos komponentės reikšmė generuojama intervale [3, 5], o kitų komponenčių reikšmės – intervale [-1, 1], t. y. $x_j \in$ [-1, 1], ir tik jei $X_i \in K_j$, tai $x_j \in$ [3, 5].

DNT buvo apmokytas visa duomenų aibe, nedalijant jos į mokymo ir testavimo poaibius, nes tikslas yra vizualizuoti visus tos aibės taškus. Pasiūlyto metodo testavimas grindžiamas tuo, kad žinoma tirtų duomenų aibių struktūra – klasterių skaičius.

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 parametrą δ. Kiekvieną kartą, prieš pradedant mokyti tinklą, atsitiktinai parenkami pradiniai svoriai, dėl to 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 apmokymo buvo mažiausia. Lentelėje pateiktos iš 100 atliktų bandymų su kiekvienu parametru δ atrinktos mažiausios daromos klaidos.

Lentelė. ***Tinklo, mokyto irisų ir atsitiktinai generuotais duomenimis, eksperimentų rezultatai***

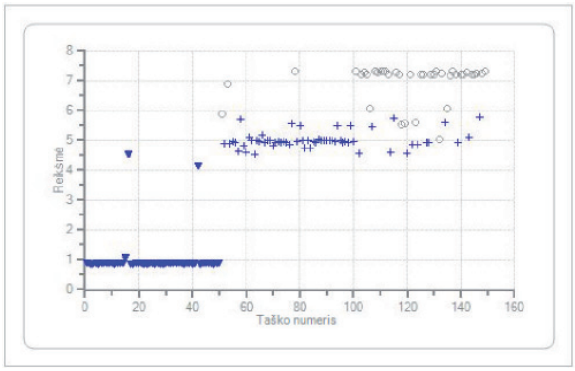
Irisų duomenys		**Atsitiktinai generuoti duomenys**	
Sigma	**Klaida**	**Sigma**	**Klaida**
2	0,00139	3,3	0,00996
1	0,00042	2	0,00345
0,67	0,00037	1	0,00292
0,5	0,00030	0,67	0,00813
0,4	0,00155	0,5	0,08028
0,3	0,00820	0,4	0,14905



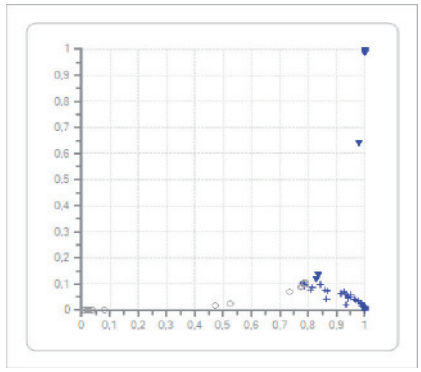
a) b)

3 pav. ***Paslėptajame L-1 sluoksnyje gautos reikšmės: a)*** δ= ***0,5; b)*** δ= ***0,67***

a)



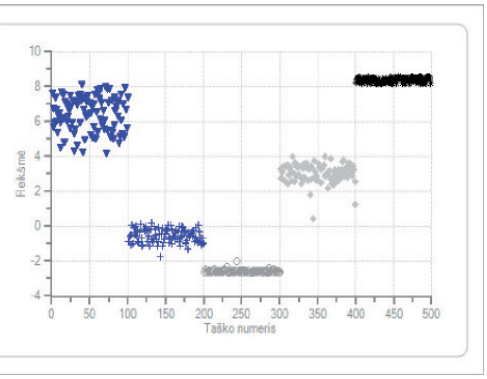
b)



4 pav. ***δ= 0,3: a) tinklo išėjime gautos reikšmės; b) paslėptajame L-1 sluoksnyje gautos reikšmės***

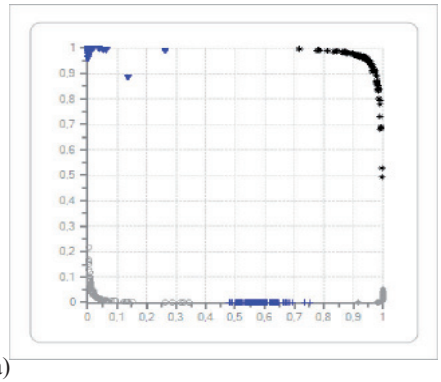
Iš lentelėje pateiktų duomenų matyti, kad irisų duomenimis gerai apmokytas tinklas, kuriam parinktas $\delta = 0{,}5$. Nuo jo nedaug atsilieka tiklas, kuriam parinktas $\delta = 0{,}67$. Po tinklo apmokymo išėjimuose gauti taškai atitinka γ_j reikšmes (2a pav.). 3 pav. parodyti vizualizuoti *L*-1 sluoksnyje gauti išėjimai apmokius tinklą. Šiame paveiksle, kaip ir kituose, kur pateikiami vizualizavimo rezultatai, abscisių ašyje atidėtos P_1^3 reikšmės, o ordinačių ašyje – P_2^3 reikšmės, gautos pateikiant tinklui visus taškus X_i, $i = \overline{1,m}$.

Iš 3 pav. pateiktų grafikų matyti, kad irisų duomenys aiškiai susiskirsto į tris klasterius. Taip atsitinka tik gerai apsimokius tinklui. Pagal lentelėje pateiktus tinklo, apmokyto irisų duomenimis, rezultatus matome, kad prasčiausiai apmokytas 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 parametrą δ duomenys nesusiskirsto į klasterius.

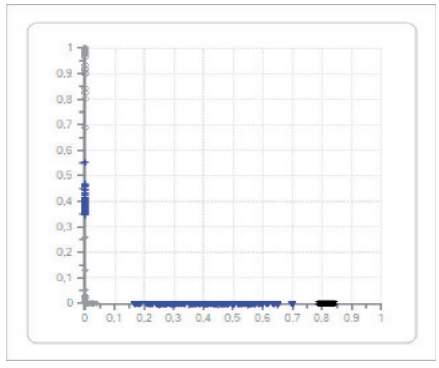


5 pav. ***Tinklo išėjimo reikšmės, kai δ= 1***

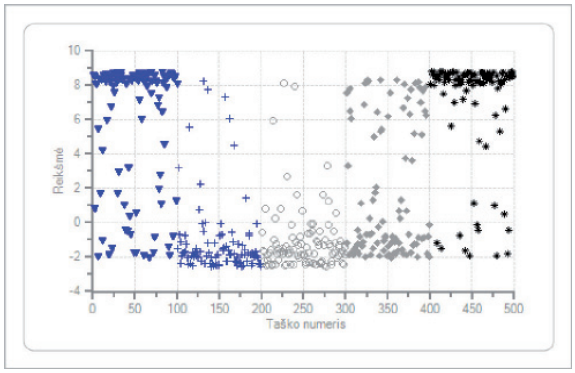
a)

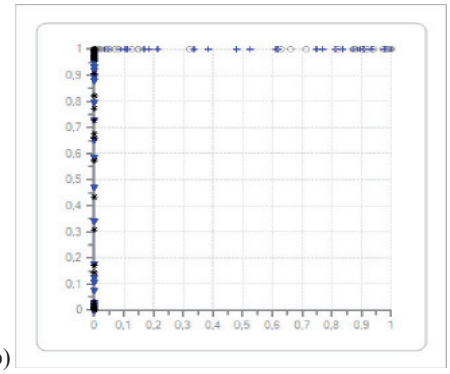


b)



6 pav. ***Paslėptajame L-1 sluoksnyje gautos reikšmės: a) δ= 2; b) δ= 1***

a) 

b) 

7 pav. ***δ= 0,4: a) tinklo išėjime gautos reikšmės; b) paslėptajame L-1 sluoksnyje gautos reikšmės***

Iš lentelėje pateiktų atsitiktinai generuotų duomenų gautų rezultatų matome, kad gerai apmokyti tinklai, kurių $\delta = 2$ ir $\delta = 1$. Tačiau visais atvejais tinklas neapmokytas taip gerai, kaip su irisų duomenimis. Iš 5 pav. matome, kad tinklo išėjimuose gautos reikšmės išsibarsčiusios (palyginti su 2b pav.), bet klasteriai tarpusavyje nesusimaišę.

6 pav. parodyti vizualizuoti *L*-1 sluoksnyje gauti išėjimai apmokius tinklą, kai $\delta = 2$ ir $\delta = 1$.

Nors po tinklo apmokymo išėjimo reikšmės yra išsisklaidžiusios, bet iš 6 pav. pateiktų grafikų matyti, kad atsitiktinai generuoti duomenys susiskirsto į klasterius. Iš lentelės duomenų matome, kad didžiausią klaidą tinklas daro, kai $\delta = 0{,}4$. Tinklo išėjime gautos reikšmės pateikiamos 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 nė vienas klasteris.

Išvados

Darbe pasiūlytas metodas daugiamačiams duomenims vizualizuoti naudojant radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginį. Apmokius tokį tinklą, paskutinio 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 du neuronai. Bendru atveju neuronų gali būti ir daugiau. Atlikti eksperimentai leidžia daryti šią išvadą, kad dirbtinio daugiasluoksnio neuroninio tinklo apmokymo kokybė labai priklauso nuo parinkto RBF parametro δ. Nustatyta, kad parinkus tinkamą δ reikšmę tinklo paskutinio paslėptojo sluoksnio išėjimų reikšmės tinkamai susiskirsto į vaizdžiai matomus klasterius.

LITERATŪRA:

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 Word 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. 204 p. ISBN 9789986680420.

FISHER, R. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics,* vol. 7, p. 179–188.

FRAHLING, G.; SOHLER, Ch. (2006). A fast k-means implementation using coresets. Iš: *Proceedings of the twenty-second annual symposium on Computational geometry (SoCG)*. New York: The Association for Computing Machinery, p. 135–143. ISBN 1595933409.

HASSOUN, M. H. (1995). *Fundamentals of Artificial Neural Networks. A Bradford Book.* Cambridge, Masachusetts: MIT Press. 511 p. ISBN 026208239X.

KRAMER, M. A. (1991). Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE Journal*, vol. 37, no. 2, p. 233–243.

KURASOVA, O. (2005). *Daugiamačių duomenų vizuali analizė taikant savireguliuojančius neuroninius tinklus* (daktaro disertacija). Vilnius: Matematikos ir informatikos institutas.

VÁZQUEZ SANTACRUZ, E. F.; CHAKRABORTY, D. (2007). A Modified Bottleneck Neural Network for Dimensionality Reduction. *Research in Computing Science. Special issue on Neural Networks and Associative Memories*, vol. 28, p. 127–136.

SPECIAL MULTILAYER PERCEPTRON FOR MULTIDIMENSIONAL DATA VISUALIZATION

Laura Ringienė, Gintautas Dzemyda

S u m m a r y

In this paper a special feed forward neural network, consisting of the radial basis function layer and a multilayer perceptron is presented. The multilayer perceptron has been proposed and investigated for multidimensional data visualization. The proposed visualization approach includes data clustering, determining the parameters of the radial basis function and forming the data set to train the multilayer perceptron. The outputs of the last hidden layer are assigned as coordinates of the visualized points.