Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
55 t., 2014, 137–142

ISSN 0132-2818
www.mii.lt/LMR/

Dimensijų mažinimo metodais gautų projekcijų įvertinimas

Kotryna Paulauskienė, Olga Kurasova

Vilniaus universitetas, Matematikos ir informatikos institutas
Akademijos g. 4, LT-08663 Vilnius
E. paštas: kotryna.paulauskiene@mii.vu.lt, olga.kurasova@mii.vu.lt

Santrauka. Šiame darbe nagrinėjami dimensijų mažinimo metodais gaunamų projekcijų kokybės įvertinimo matai. Tam pasirinkti du dimensijų mažinimo metodai (pagrindinių komponenčių analizė ir dalinai tiesinė daugiamatė projekcija), kurių projekcijos tikslumas vertinamas pagal kelis įvairias savybes atspindinčius projekcijos kokybės įvertinimo matus, naudojant realias ir dirbtinai sugeneruotas duomenų aibes.

Raktiniai žodžiai: projekcijos kokybės įvertinimas, dimensijų mažinimo metodai.

Įvadas

Dimensijos mažinimo metodai (angl. *dimension reduction techniques, projection techniques*) m-matės erdvės vektorius (taškus) transformuoja į d-matės erdvės vektorius (taškus), $d < m$. Jų tikslas – pateikti daugiamačius duomenis mažesnės dimensijos erdvėje taip, kad būtų kiek galima tiksliau išlaikyta tam tikra duomenų struktūra, taip būtų palengvintas didelės dimensijos duomenų apdorojimas ir interpretavimas. Sumažinus pradinių duomenų dimensiją būtina įvertinti gautos projekcijos kokybę.

Šio tyrimo objektas – struktūros išlaikymo tarp objektų pradinėje ir sumažintos dimensijos erdvėse kokybės įvertinimo matai. Tyrimo tikslas – parodyti, kad projekcija turi būti vertinama naudojant skirtingas savybes atspindinčius projekcijos kokybės įvertinimo matus. Šiam tikslui lyginamos dviem dimensijų mažinimo metodais gautos projekcijos.

1 Dimensijų mažinimo metodai

Šiame straipsnyje pradinių duomenų dimensijai sumažinti taikomi šie metodai: pagrindinių komponenčių analizė (angl. *Principal Component Analysis*, PCA) ir dalinai tiesinė daugiamatė projekcija (angl. *Part–Linear Multidimensional Projection*, PLMP).

Pagrindinių komponenčių analizės [8] esminė idėja yra sumažinti duomenų dimensiją atliekant tiesinę transformaciją ir atsisakant dalies po transformacijos gautų naujų komponenčių, kurių dispersijos yra mažiausios [2].

Straipsnyje [6] siūlomas naujas metodas, skirtas daugiamačių duomenų dimensijai mažinti – dalinai tiesinė daugiamatė projekcija. Siūloma pradžioje vertinti atstumus tik tarp dalies taškų (sudaroma imtis), pagal kuriuos vėliau randamos likusiųjų duomenų aibės taškų projekcijos.

Tarkime $X = \{X_1, \ldots, X_n\}$ yra duomenų aibė, sudaryta iš taškų $X_i \in R^m$. Reikia rasti tokią tiesinę transformaciją $\Phi : R^m \to R^d$, $d < m$, kuri tenkintų išraišką:

$$\Phi = \underset{\hat{\Phi} \in L_{m,d}}{\operatorname{argmin}} \left\{ \frac{1}{D} \sum_{ij} \left(d(X_i, X_j) - d\big(\hat{\Phi}(X_i), \hat{\Phi}(X_j)\big) \right)^2 \right\},$$

čia $L_{m,d}$ – tiesinės transformacijos iš erdvės R^m į R^d, $d(X_i, X_j)$ ir $d(\hat{\Phi}(X_i), \hat{\Phi}(X_j))$ yra atstumai tarp taškų pradinėje ir sumažintos dimensijos erdvėse, o $D = \sum_{ij} d(X_i, X_j)^2$.

Tiesiogiai ieškoti Φ esant didelėms n reikšmėms yra sudėtinga, todėl yra ieškoma aproksimacija. Pradžioje iš aibės X parenkama dalis taškų $X' = \{X'_1, \ldots, X'_k\}$, $k \ll n$. Jie atvaizduojami sumažintos dimensijos erdvėje R^d. Tegu $\bar{X}'_i$ yra taško X'_i projekcija erdvėje R^d. Tada projekcija X'_i, minimizuojanti Φ, turėtų tenkinti lygybę:

$$\Phi\big(X'_i\big) = \bar{X}'_i, \quad i = 1, \ldots, k.$$

Ši lygybė leidžia apskaičiuoti aproksimuojančią Φ transformaciją nagrinėjant kiekvienos matricos Φ eilutės sandaugą su taškais iš aibės X'. Pilna Φ aproksimacija gaunama kartojant skaičiavimus kiekvienai Φ eilutei, kurių yra d. Šiame tyrime pasirinkta $d = 2$. Taip gaunamos tiesinių lygčių sistemos, kurios sprendžiamos jungtinių gradientų metodu. PLMP metode projekcijos kokybė priklauso nuo taškų $X' = \{X'_1, \ldots, X'_k\}$ parinkimo. Šio metodo autoriai naudoja atsitiktinį taškų parinkimą. Eksperimentiniais tyrimais yra nustata, kad geriausia pasirinkti $k = \sqrt{n}$ taškų, tai duoda pusiausvyrą tarp skaičiavimo kaštų ir taškų atvaizdavimo kokybės [6]. Šiame straipsnyje, vykdant eksperimentinį tyrimą, taškų $X' = \{X'_1, \ldots, X'_k\}$ skaičius yra $\sqrt{n}$, jų projekcija sumažintos dimensijos erdvėje randama daugiamačių skalių metodu [1].

2 Projekcijos kokybės įvertinimo matai

Projekcijos kokybės matai atspindi įvairias duomenų savybes, todėl yra tikslinga projekciją vertinti ne pagal vieną, o pagal kelis matus. Dimensijų mažinimo metodais gautos projekcijos kokybei įvertinti siūlomi šie kokybės matai: projekcijos paklaida (angl. *stress function*), Spirmeno rho koeficientas (angl. *Spearman's rho*), Konigo topologijos išlaikymo matas (angl. *Konig's topology measure*), silueto koeficientas (angl. *silhouette*), Renyi entropijos koeficientas (angl. *Renyi entropy*).

Projekcijos paklaida E – tai matas, kuris parodo, kaip tiksliai išlaikomi atstumai tarp taškų pereinant iš didesnės dimensijos erdvės į mažesnės dimensijos erdvę. Projekcijos paklaida apskaičiuojama pagal šią formulę [1]:

$$E = \frac{\sum_{ij} (d(X_i, X_j) - d(Y_i, Y_j))^2}{\sum_{ij} (d(X_i, X_j))^2},$$

čia $d(X_i, X_j)$ ir $d(Y_i, Y_j)$ yra atstumai tarp taškų pradinėje ir sumažintos dimensijos erdvėse.

Straipsnyje [5] projekcijos kokybei įvertinti naudojamas Spirmeno rho koeficientas, kuris apskaičiuojamas pagal šią formulę:

$$\rho = 1 - \frac{6}{(n')^3 - n'} \sum_{k=1}^{n'} \left(r'_X(k) - r'_Y(k) \right)^2,$$

čia r'_X ir r'_Y yra atstumų tarp taškų rangai pradinėje ir sumažintos dimensijos erdvėse, n – taškų skaičius, $n' = n(n-1)/2$. Spirmeno rho naudojamas siekiant įvertinti santykinį atstumų išlaikymą tarp taškų pereinant iš m-matės erdvės į d-matę erdvę. Spirmeno rho koeficientas yra skaičius, priklausantis intervalui $[-1, 1]$. Vertinant projekciją pagal šį matą, geriausia projekcija laikoma tokia, kai jo reikšmė yra 1.

Konigo topologijos išlaikymo matas yra pagrįstas atstumų tarp taškų rangų tvarkos išlaikymu m-matėje ir d-matėje erdvėse. Šis matas turi du valdymo parametrus, tai artimiausių kaimynų skaičiai: μ ir ν ($\mu < \nu$). Artimiausiems kaimynams nustatyti naudojamas Euklido atstumas. Konigo topologijos išlaikymas kiekvienam i-ajam taškui ir j-ajam kaimynui apskaičiuojamas pagal šią formulę [5]:

$$E_{KT}^{ij} = \begin{cases} 3, & \text{kai } r_X(i,j) = r_Y(i,j), \\ 2, & \text{kai } r_X(i,j) = r_Y(i,l),\ l \in (1,\dots,\mu),\ i \neq l, \\ 1, & \text{kai } r_X(i,j) = r_Y(i,t),\ t \in (\mu+1,\dots,\nu),\ \mu < \nu, \\ 0, & \text{kitais atvejais.} \end{cases}$$

Čia naudojami šie žymėjimai:

X_{ij}, $j = 1,\dots,\mu$ yra m-mačio taško X_i artimiausi kaimynai, kurie tenkina nelygybę $\|X_i - X_{ij_1}\| \leqslant \|X_i - X_{ij_2}\|$, $j_1 < j_2$, μ – kaimynų skaičius;

Y_{ij}, $j = 1,\dots,\nu$ yra d-mačio taško Y_i artimiausi kaimynai, ν – kaimynų skaičius;

$r_X(i,j)$ yra m-mačio taško X_i j-otojo kaimyno X_{ij} rangas;

$r_Y(i,j)$ yra d-mačio taško Y_i j-otojo kaimyno Y_{ij} rangas.

Bendras Konigo topologijos išlaikymo matas apskaičiuojamas pagal formulę:

$$E_{KT} = \frac{1}{3\mu \times m} \sum_{i=1}^{\mu} \sum_{j=1}^{m} E_{KT}^{ij}.$$

Koeficientas E_{KT} yra intervale $[0, 1]$. Geriausiai topologija išlaikoma, kai koeficientas lygus 1. Šiame straipsnyje naudojamos $\mu = 3$ ir $\nu = 5$ reikšmės. Esant kitoms reikšmėms rezultatai šiek tiek skirtųsi, bet bendros tendencijos išliktų, lyginant dvi skirtingais metodais gautas projekcijas.

Silueto koeficientas buvo pasiūlytas klasterizavimo algoritmų kokybei nustatyti ir leido įvertinti sanglaudą ir atskyrimą tarp suklasterizuotų duomenų aibės taškų. Jis parodo, kaip gerai kiekvienas taškas priskirtas klasteriui. Darbe [4] šis koeficientas naudojamas projekcijos kokybei įvertinti. Silueto koeficiento vidurkis visai duomenų aibei apskaičiuojamas pagal šią formulę [7]:

$$S = \frac{1}{n} \sum_{k=1}^{n} \frac{b_{X_i} - a_{X_i}}{\max(a_{Xi}; b_{Xi})},$$

čia n – taškų skaičius aibėje. Taško X_i sanglauda a_{X_i} apskaičiuojama suvidurkinus taško X_i iki taškų, priklausančių tam pačiam klasteriui, atstumų skirtumus. Atskyrimas b_{X_i} yra mažiausias vidutinis atstumas tarp taško X_i iki taškų, priklausančių kitiems klasteriams. Silueto koeficiento reikšmės yra intervale $[-1, 1]$, kuo didesnė koeficiento S reikšmė, tuo geresnė sanglauda ir atskyrimas. Šiame darbe siūloma silueto koeficientą apskaičiuoti m-mačių ir d-mačių taškų duomenų aibėms, tada vertinti silueto koeficiento skirtumą. Kadangi šiame darbe nėra sprendžiamas klasterizavimo uždavinys, tai klasteriais laikomos duomenų aibių klasės.

Straipsnyje [3] analizuojant skaitmeninius vaizdus siekiant įvertinti informacijos kiekį, siūloma Euklido atstumų matricai taikyti Renyi entropijos koeficientą, kuris apskaičiuojamas pagal šią formulę:

$$H_\alpha(p) = \frac{1}{1-\alpha} \ln \sum_{i=1}^{n} p_i^\alpha,$$

čia $\alpha \geqslant 0$, p_i – tikimybė. Šiame darbe naudojama $\alpha = 2$ reikšmė. Renyi entropija apskaičiuojama m-mačių taškų ir sumažintos dimensijos d-mačių taškų duomenų aibėms. Tada siūloma vertinti entropijos skirtumą tarp taškų atstumų pradinėje ir sumažintos dimensijos erdvėse t. y. ar informacijos kiekis išlieka tas pats sumažinus duomenų aibės dimensiją.

3 Eksperimentinių tyrimų rezultatai

Analizei pasirinkti duomenų rinkiniai, kuriuos galima rasti duomenų bazėje „UCI Repository of Machine Learning Databases (http://archive.ics.uci.edu/ml/)“, taip pat naudotos dirbtinai sugeneruotos duomenų aibės. Duomenų aibių parametrai pateikiami 1 lentelėje.

Eksperimentams atlikti naudotas kompiuteris, kurio pagrindinės charakteristikos yra šios: operacinė sistema – Windows 8, operatyvioji atmintis (RAM) – 12 GB, procesorius – Intel i5-3317U, kurio taktinis dažnis – 1,7 GHz (Max Turbo dažnis 2,6 GHz). Skaičiavimai atlikti naudojant MATLAB R2012b.

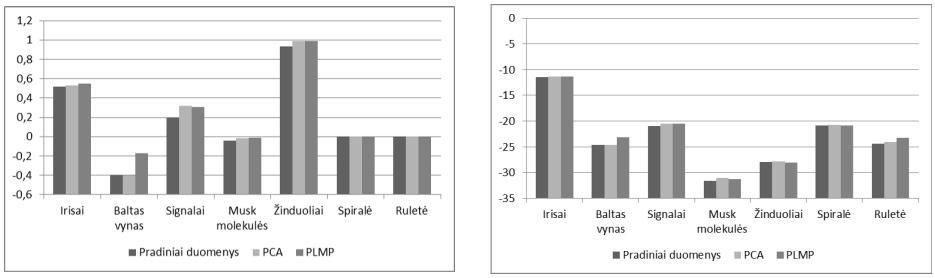
2 lentelėje pateikiamos projekcijos paklaidos, Spirmeno rho ir Konigo topologijos išlaikymo koeficientų reikšmės analizuojant įvairias duomenų aibes. Geresnės matų reikšmės pateikiamos paryškintu šriftu. Pagrindinių komponenčių metodu gaunama mažesnė projekcijos paklaida nei dalinai tiesinės daugiamatės projekcijos metodu, analizuojant keturias (*Irisų, Balto vyno, Žinduolių, Ruletės*) iš septynių duomenų aibių. Dalinai tiesinės daugiamatės projekcijos metodu gauta projekcijos paklaida mažesnė nagrinėjant tris (*Signalų, Musk molekulių, Spiralės*) iš septynių duomenų aibių. Tuo tarpu pagrindinių komponenčių metodu gaunamos projekcijos Spirmeno rho koeficientas, nors ir nežymiai, išskyrus *Balto vyno* duomenų aibę, tačiau yra didesnis nei dalinai tiesinės daugiamatės projekcijos, analizuojant visas duomenų aibes. Pagrindinių komponenčių metodu gaunamos projekcijos Konigo topologijos koeficiento reikšmės nežymiai yra didesnės naudojant keturias (*Irisai, Baltas vynas, Signalai, Ruletė*) iš septynių duomenų aibių, o dalinai tiesinės daugiamatės projekcijos metodu

1 lentelė. Duomenų aibių parametrai.

Duomenų aibė	Duomenų aibė anglų k.	Duomenų aibės tipas	Objektų skaičius (n)	Požymių skaičius (n)	Klasių skaičius (l)
Irisai	*Iris*	Reali	150	4	3
Baltas vynas	*White wine*	Reali	4 898	11	10
Signalai	*Waveform*	Dirbtinė	5 000	21	3
Musk molekulės	*Musk*	Dirbtinė	6 598	166	2
Žinduoliai	*Mammals*	Dirbtinė	15 000	72	4
Spiralė	*Helix*	Dirbtinė	15 000	3	2
Ruletė	*Swiss roll*	Dirbtinė	15 000	3	2

2 lentelė. Dimensijų mažinimo metodais gautų projecijų kokybės matų reikšmės, naudojant skirtingas duomenų aibes.

	Projekcijos paklaida (E)		Spirmeno rho (ρ)		Konigo topologijos išlaikymas (E_{KT})	
Duomenų aibė	PCA	PLMP	PCA	PLMP	PCA	PLMP
Irisai	**0,0018**	0,005	**0,9935**	0,9927	**0,5533**	0,5467
Baltas vynas	**0,0005**	0,5039	**0,9989**	0,2899	**0,4237**	0,2318
Signalai	0,0852	**0,0078**	**0,9571**	0,9516	**0,0128**	0,0127
Musk molekulės	0,1387	**0,1089**	**0,9262**	0,9146	0,1811	**0,1841**
Žinduoliai	**0,0016**	0,0052	**0,9932**	0,9928	0,3032	**0,3038**
Spiralė	0,0115	**0,0097**	**0,9788**	0,9785	0,2942	**0,2946**
Ruletė	**0,0568**	0,2692	**0,8667**	0,6021	**0,3826**	0,3669



(a) (b)

1 pav. Kokybės matų priklausomybė nuo duomenų aibės: (a) Silueto koeficientas, (b) Renyi entropija.

tris (*Musk molekulių, Žinduolių, Spiralės*) duomenų aibes. Blogiausios kokybės matų reikšmės gaunamos dalinai tiesinės daugiamatės projekcijos metodu nagrinėjant *Balto vyno* duomenų aibę.

Nagrinėjant silueto koeficiento priklausomybę nuo duomenų aibės, pastebima, kad pradinių duomenų aibių silueto koeficiento reikšmės yra panašios į sumažintos dimensijos duomenų aibių silueto koeficiento reikšmes (1 pav. (a)). Labiausiai skiriasi *Balto vyno* pradinių duomenų aibės silueto koeficiento ($S = -0{,}3984$) ir dalinai tiesinės daugiamatės projekcijos metodu gautos projekcijos silueto reikšmės ($S = -0{,}1753$). *Signalų* duomenų aibės silueto koeficientas ($S = 0{,}1972$) taip pat skiriasi nuo projekcijos silueto koeficientų (PCA: $S = 0{,}3188$, PLMP: $S = 0{,}307$).

Renyi entropija pakinta nežymiai lyginant pradinių duomenų entropijos koeficientą su dimensijų mažinimo metodais gautos projekcijos entropijos koeficientais, nagrinėjant visas duomenų aibes (1 pav. (b)). Galima teigti, kad sumažinus duomenų aibės dimensiją, tiek pagrindinių komponenčių metodu, tiek dalinai tiesinės daugiamatės projekcijos metodu, informacijos kiekis išlieka beveik nepakitęs.

4 Išvados

Atliktas tyrimas parodė, kad projekcijos kokybės matų reikšmės priklauso nuo nagrinėjamų duomenų aibių ir taikomo dimensijų mažinimo metodo. Nustatyta, kad pagrindinių komponenčių metodu geresnės projekcijos paklaidos ir Konigo topologijos

išlaikymo reikšmės gaunamos analizuojant keturias iš septynių duomenų aibių, o dalinai tiesinės daugiamatės projekcijos metodu atitinkamai tris iš septynių. Tuo tarpu Spirmeno rho koeficiento reikšmės, apskaičiuotos pagrindinių komponenčių metodu gautai projekcijai, buvo didesnės naudojant visas duomenų aibes. Renyi entropijos koefiecientas parodė, kad abiem metodais gautos projekcijos išlaiko panašų informacijos kiekį, kaip ir pradinėje duomenų aibėje. Kalbant apie taškų priskyrimą klasteriams sumažinus duomenų dimensiją, nustatyta, kad gautų projekcijų ir pradinės duomenų aibės silueto koefieciento reikšmės daugeliu atveju yra panašios. Apibendrinus rezultatus, galima teigti, kad sprendžiant dimensijos mažinimo uždavinį, būtina taikyti kelis projekcijos kokybės įvertinimo matus, siekiant nustatyti, ar dimensijos mažinimas išlaiko duomenų savybes.

Literatūra

[1] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, 2005. ISBN 978-0-387-28981-6.

[2] G. Dzemyda, O. Kurasova and J. Žilinskas. *Multidimensional Data Visualization: Methods and Applications*. Springer, New York, 2012. ISBN 978-1-4419-0236-8.

[3] A. Gupta and R. Bowden. Evaluating dimensionality reduction techniques for visual category recognition using Renyi entropy. In *19th European Signal Processing Conference (EUSIPCO 2011)*, pp. 913–917, 2011.

[4] P. Joia, F.V. Paulovich, D. Coimbra, J.A. Cuminato and L.G. Nonato. Local affine multidimensional projection. *IEEE Trans. Visual. Comp. Graph.*, **17**(12):2563–2571, 2011.

[5] O. Kurasova and A. Molytė. Quality of quantization and visualization of vectors obtained by neural gas and self-organizing map. *Informatica*, **22**(1):115–134, 2011.

[6] F.V. Paulovich, C.T. Silva and L.G. Nonato. Two-phase mapping for projecting massive data sets. *IEEE Trans. Visual. Comp. Graph.*, **16**(6):1281–1290, 2010.

[7] P. Tan, M. Steinbach and V. Kumar. *Introduction to Data Mining*. Addison–Wesley, Boston, 2005. ISBN 0321321367.

[8] L.P.J. van der Maaten, E.O. Postma and H.J. van den Herik. Dimensionality reduction: a comparative review. *Technical Report TiCC TR 2009-005*, 2009.

SUMMARY

Evaluation of projections, obtained by dimensionality reduction techniques
K. Paulauskienė, O. Kurasova

In this paper, the projection evaluation measures such as stress function, Spearman's rho, Konig's topology preservation, silhouette and Renyi entropy have been analyzed. The principal component analysis (PCA) and part–linear multidimensional projection (PLMP) techniques are used to reduce the dimensionality of the initial data set. The experiments have been carried out with seven real and artificial datasets. The experimental investigation has shown that several quality evaluation measures have to be used when dimension reduction problem is solved.

Keywords: projection quality evaluation, dimensionality reduction.