Liet. matem. rink., **44**, *spec. nr.*, 2004, 628–633

Sammono projekcijos paklaidos minimizavimo strategijos

Gintautas DZEMYDA, Jolita BERNATAVIČIENĖ, Olga KURASOVA,
Virginijus MARCINKEVIČIUS (MII)
el. paštas: dzemyda@ktl.mii.lt, jolitab@ktl.mii.lt, kurasova@ktl.mii.lt, virgism@ktl.mii.lt

1. Įvadas

Dažnai duomenys aprašomi daugeliu parametrų, kurie vadinami daugiamačiais duomenimis. Būtina ieškoti būdų pateikti juos žmogui suvokiama forma, pvz., vizualiai. Šiam tikslui puikiai tinka daugiamačių duomenų projekcijos į mažesnės dimensijos erdvę metodai: principinių komponenčių analizė [9], daugiamatis vertinimas (MDS) [6], Sammono projekcija [8]. Jų tikslas pateikti n-mačius vektorius mažesnės dimensijos erdvėje $R^m (m < n)$ (dažniausiai vizualizavime $m = 2$) taip, kad būtų išlaikyta duomenų struktūra. Čia optimizuojami tam tikri matematiniai kriterijai. Tačiau projekcijos paklaidos yra neišvengiamos. Šio straipsnio tikslas – ieškoti tų projekcijos paklaidų minimizavimo strategijų.

2. Sammono projekcija

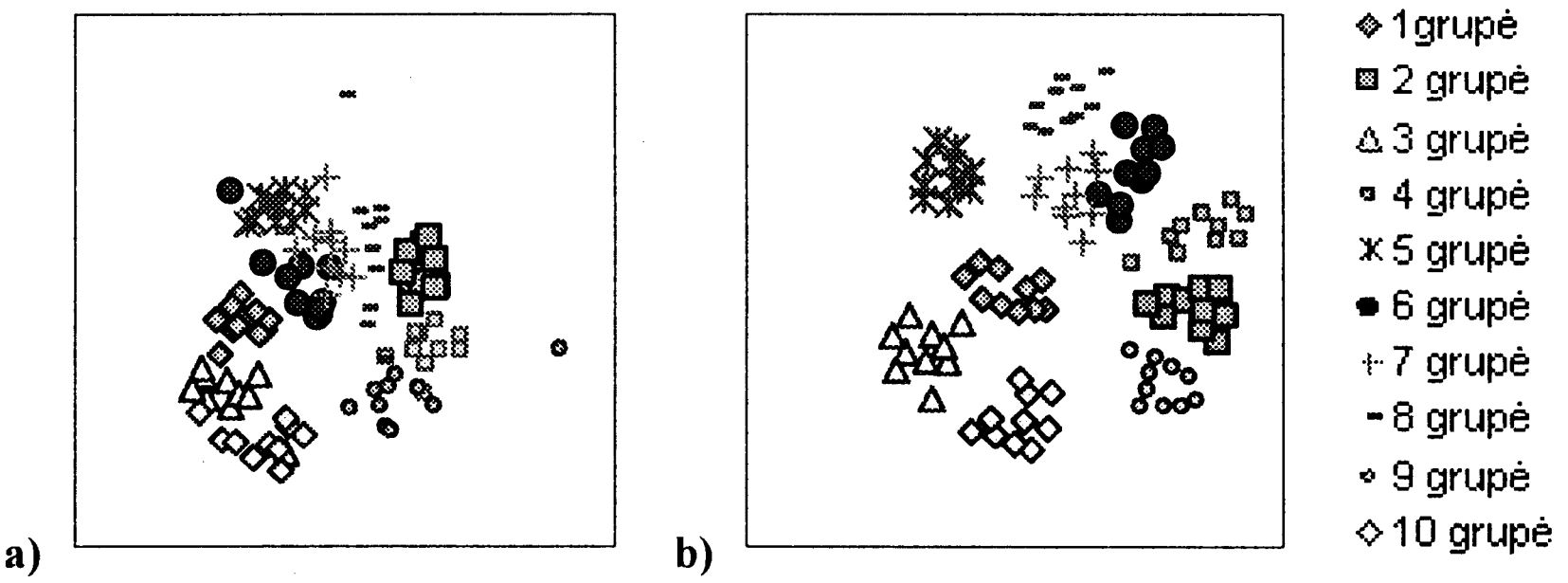
Sammono projekcija [8] yra netiesinis daugelio kintamųjų objektų (vektorių) atvaizdavimo žemesnio matavimo erdvėje metodas. Jis minimizuoja projekcijos paklaidą E_s:

$$E_s = \left(\sum_{\substack{i,j=1 \\ i<j}}^{s} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}} \right) \Big/ \sum_{\substack{i,j=1 \\ i<j}}^{s} d^*_{ij}. \qquad (1)$$

Čia d^*_{ij} – atstumas tarp daugiamačių vektorių $X_i, X_j \in R^n$, $i, j = 1, ..., s$, d_{ij} – tarp vektorius X_i, X_j atitinkančių dvimačių vektorių $Y_i, Y_j \in R^2$.

Net nedidelis paklaidos sumažinimas leidžia gauti iš esmės kitą taškų išsidėstymą plokštumoje. 1 pav. pateiktos 100 10-mačių vektorių, sudarančių 10 klasterių, projekcijos plokštumoje, fiksuojant skirtingas paklaidas. Esant mažesnei paklaidai, gaunamas tikslesnis vaizdas – tiksliau išskiriami duomenų klasteriai (1b pav).

Šiame darbe ištirtos trys projekcijos paklaidos funkcijos E_s(1) minimizavimo strategijos: (**S1**) klasikinis Sammono algoritmas [8]; (**S2**) Sammono projekcijos koordinatinė paieška Zeidelio tipo [5] metodu; (**S3**) koordinatinė paieška su triukšmu.



1 pav. Daugiamačių duomenų projekcijos plokštumoje: a) $E_s = 0,1204$, b) $E_s = 0,0694$.

2.1. Projekcijos paklaidos minimizavimo strategijos

Klasikiniame Sammono algoritme (**S1**) dvimačių vektorių $Y_i = (y_{i1}, y_{i2})$ koordinatės y_{ik}, $i = 1, ..., s$, $k = 1, 2$, randamos naudojantis iteracine formule:

$$y_{ik}(m'+1) = y_{ik}(m') - \alpha \frac{\partial E_s(m')}{\partial y_{ik}(m')} \Big/ \left| \frac{\partial^2 E_s(m')}{\partial y_{ik}^2(m')} \right|. \qquad (2)$$

Čia m' yra iteracijos numeris, o α – žingsnis, dar vadinamas „magiškuoju faktoriumi“, kadangi nuo jo priklauso projekcijos paklaida. Vienoje iteracijoje perskaičiuojamos s dvimačių vektorių Y_i, $i = 1, ..., s$, abi koordinatės atsižvelgiant į ankstesnėje iteracijoje gautas koordinates.

Sammono projekcijos paieškai tikslinga taikyti Zeidelio idėją, naudojamą ir tiesinių lygčių sprendime [7] ir optimizavime [5]. Ši idėja grindžiama koordinatine paieška. Sammono projekcijos koordinatinėje paieškoje (**S2**) dvimačių vektorių koordinatės y_{ik} skaičiuojamos atsižvelgiant ne tik į koordinates gautas ankstesnėje iteracijoje, bet taip pat ir į koordinates, gautas skaičiuojamoje iteracijoje, t.y. į koordinates $y_{jk}^{(m')}$, kai $j = i+1, \ldots, s$ ir į $y_{jk}^{(m'+1)}$, kai $j = 1, ..., i-1$. Perskaičiavus taško koordinates iš karto perskaičiuojami visi atstumai nuo jo iki visų kitų taškų.

Sammono projekcijos paklaida priklauso nuo pradinių dvimačių vektorių reikšmių. Darbe [2] pradinės reikšmės parinktos taip: $y_{i1} = i + 1/3$, $y_{i2} = i + 2/3$. Tiriant konvergavimo procesą pastebėta, kad, kai pradiniai vektoriai išdėstomi ant įstrižainės, konvergavimas labai lėtas. To priežastis yra apspręsta šia teorema.

Teorema. *Jei pradiniai projekcijos taškai* $Y_i = (y_{i1}, y_{i2})$, $i = 1, ..., s$, *yra išdėstyti ant tiesės* $y_{i1} = ay_{i2} + b$, ($a = \pm 1$, b – konstanta), *tai naujos taškų projekcijos, skaičiuojamos pagal* (2) *formulę, bus taip pat toje pačioje tiesėje.*

Iš teoremos seka, kad projekcijos taškai visada turėtų būti tiesėje. Bet po keletos iteracijų, kaupiantis skaičiavimo paklaidoms, taškai jau išsibarsto plokštumoje.

Analizuojant (**S1**) algoritmą, pastebėta, kad I eilės projekcijos paklaidos išvestinė formulėje (2) nusistovi monotoniškai, o II – svyruoja. Tai greitina taškų išbarstymą

nuo tiesės. (**S2**) metode tie svyravimai yra mažesni, todėl nuspręsta II eilės išvestinėje padaryti dirbtinius svyravimus – taikyti triukšmą (**S3**). Buvo bandyta naudoti atsitiktinį triukšmą, tačiau priklausomai nuo generuojamo triukšmo gaunamos skirtingos paklaidos. Todėl tikslinga triukšmą apibrėžti pagal konkrečią taisyklę:

$$\frac{\partial^2 E_s}{\partial y_{ik}^2} = \frac{\partial^2 E_s}{\partial y_{ik}^2}(1 - e^{-\lambda m'}) \left|\sin(\beta m')\right|, \quad \text{kai } m' < \frac{\tau}{n}, \tag{3}$$

čia λ, β – konstantos parenkamos eksperimentiškai, τ – skaičiavimams numatytų iteracijų skaičius, m' – einamosios iteracijos numeris.

3. Tyrimų rezultatai

Eksperimentai atlikti su atsitiktinai generuotais ir realiais duomenimis. Tirta paklaidos priklausomybė nuo laiko, iteracijų skaičiaus, „magiškojo faktoriaus“ reikšmių.

Duomenys analizei

Dirbtinai generuoti duomenys:

- atsitiktiniai (100 ir 500 10-mačių vektorių, kurių koordinatės yra atsitiktinai tolygiai pasiskirsčiusios intervale $[-1; 1]$);
- klasteriai (sugeneruota 10 atsitiktinių 10-mačių vektorių, kiekvieno iš jų aplinkoje sugeneruoti dar po devynis vektorius, pasiskirsčiusius pagal normalinį dėsnį);

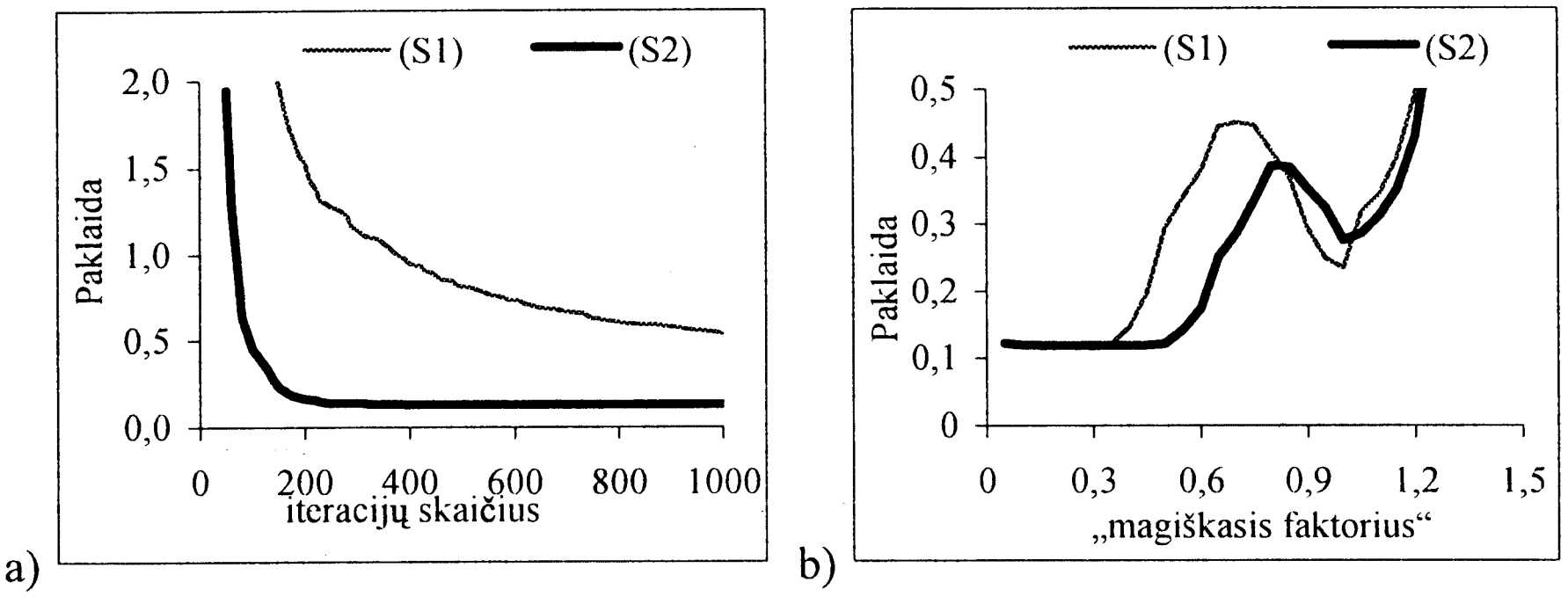
Naudojant atsitiktinius duomenis projekcijos paklaidos priklausomybės nuo įvairių faktorių tyrime, esant tom pačiom pradinėm sąlygom atlikta po 100 eksperimentų vis su kitu duomenų rinkiniu. Apskaičiuoti gautų rezultatų vidurkiai.

Realūs duomenys:

- klasikiniai Fišerio irisų duomenys [3] (150 4-mačių vektorių);
- Wood duomenys [1] (20 5-mačių vektorių, kuriuose yra 4 taškai-atsiskyrėliai);
- HBK duomenys [4] (75 4-mačiai vektoriai, kurie sudaro 3 atskirų taškų grupes: 1–10 taškai sudaro vieną grupę, 11–14 – antrą grupę ir likę taškai – trečią);
- hipersferos (sugeneruota 500 atsitiktinių šešiamačių taškų, kurie priklauso trims skirtingoms hipersferoms, įdėtoms viena į kitą).

Paklaida – laikas. Lyginant klasikinį Sammono algoritmą (**S1**) su algoritmu (**S2**) analizuoti atsitiktiniai 500 10-mačių vektorių duomenys ($\alpha = 0,25$). 2a pav. pateikti gautų projekcijų paklaidų vidurkiai. Algoritmu (**S2**) gaunama žymiai tikslesnė daugiamačių vektorių projekcija plokštumoje ir greitesnis paklaidos konvergavimas: mažai projekcijos paklaidai pasiekti reikia atlikti žymiai mažiau iteracijų, o tuo pačiu, sutaupomas skaičiavimo laikas. (**S3**) algoritmu projekcijos paklaidos konvergavimo greitis labai panašus į (**S2**), tik daugeliu atvejų gaunama mažesnė paklaida.

Paklaida – „magiškasis faktorius“. Tiriant projekcijos paklaidos priklausomybę nuo „magiškojo faktoriaus“ α reikšmės analizuoti atsitiktiniai duomenys, esant įvairioms α reikšmėms $(0,1; 0,11; \ldots; 1,45; 1,5)$. 2b pav. pateikti gautų projekcijos paklaidų vidurkiai. Matyti, kad naudojant koordinatinės paieškos algoritmą (**S2**) priklausomybė nuo α sumažėja. Triukšmo taikymas koordinatinės paieškos algoritmui – (**S3**) metodas – ypatingos įtakos priklausomybei nuo α nedaro.



2 pav. Projekcijos paklaidos priklausomybė
(a) nuo iteracijų skaičiaus, (b) nuo „magiškojo faktoriaus“ reikšmės.

1 lentelė. Projekcijos paklaidos, gautos skirtingais algoritmais

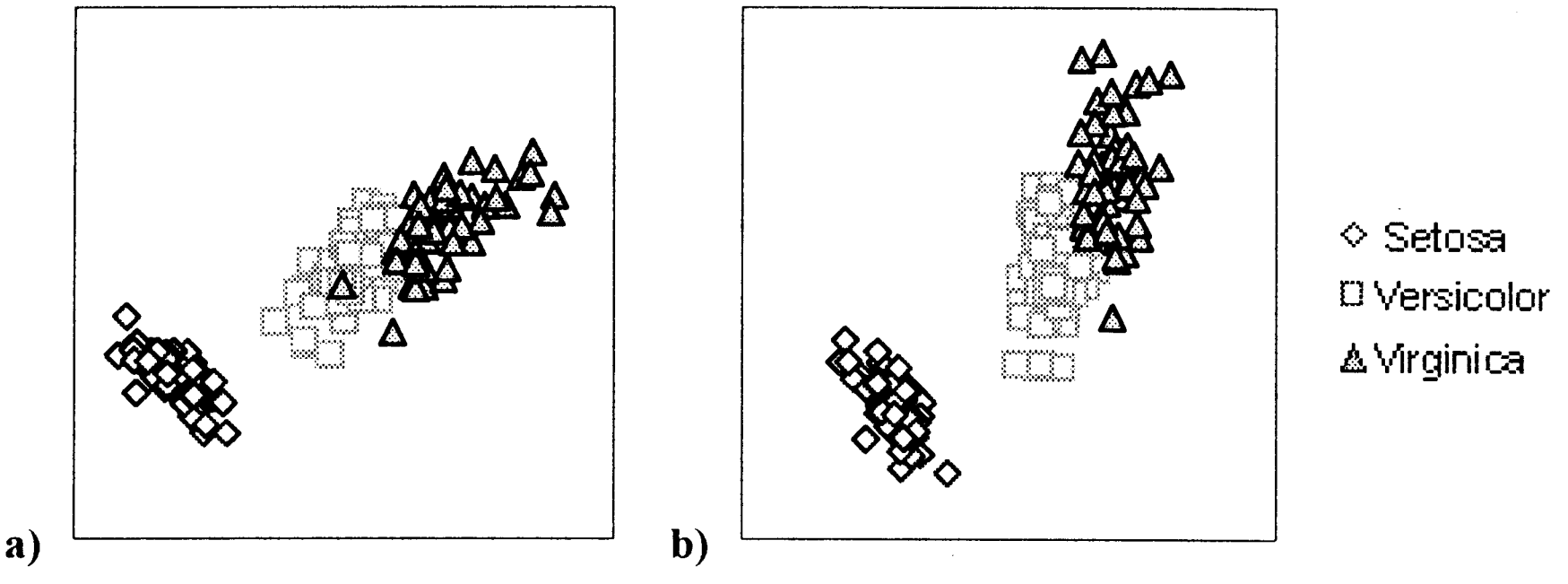
Sammono paklaidos minimizavimo strategijos			Duomenys
(S1)	**(S2)**	**(S3)**	
0,1209452	0,1201307	0,1203316	atsitiktiniai
0,0710200	0,0711317	0,0695346	klasteriai
0,0058476	0,0045259	0,0040088	irisai
0,0243263	0,0257550	0,0256691	Wood
0,0112111	0,0113962	0,0049657	HBK
0,1093391	0,1090868	0,1090809	Hipersferos

Vaizdų analizė. Tiriant atsitiktinius duomenis trimis minėtais algoritmais apskaičiuoti gautų mažiausių projekcijų paklaidų vidurkiai. Naudojant realius duomenis rastos mažiausios projekcijos paklaidos. Analizuojamiems duomenims $\alpha = 0,25$. Gauti rezultatai pateikti 1 lentelėje. Klasikiniu Sammono algoritmu visais nagrinėtais atvejais gaunamas didesnis projekcijos iškraipymas. Daugumoje atveju (**S3**) metodu gauta mažiausia projekcijos paklaida.

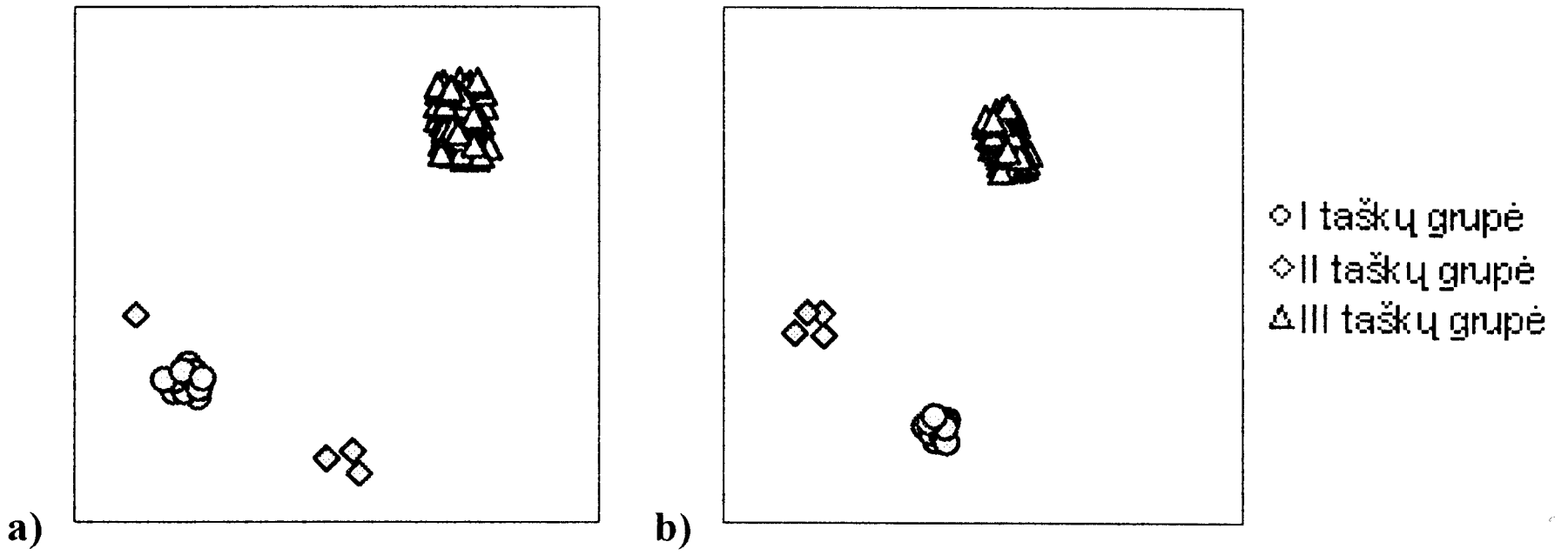
3–4 pav. pateikiami realių duomenų vaizdai, gauti klasikiniu Sammono algoritmu bei mažiausią projekcijos paklaidą radusiu modifikuotu metodu (šiuo atveju **S3**). Matyti, kad mažesnė projekcijos paklaida leidžia labiau išlaikyti duomenų struktūrą. Pvz., analizuojant irisų duomenis labiau išsiskiria trys žiedų klasės, naudojant (**S3**) algoritmą (3b pav.); tiriant HBK duomenis (**S1**) algoritmas taškų grupių išskirti nepajėgus (4a pav.), o algoritmu (**S3**) visos trys grupės tiksliai išskiriamos (4b pav.).

4. Išvados

Šiame straipsnyje atskleistos naujos galimybės daugiamačių duomenų projekcijos paklaidai minimizuoti, kurios ne tik užtikrina mažesnių paklaidų radimą, bet ir sudaro pagrindą tolesniems tyrimams šioje duomenų analizės srityje. Klasikinis Sammono



3 pav. Irisų duomenų projekcijos plokštumoje: a) (**S1**) ($E_s = 0,0059$); b) (**S3**) ($E_s = 0,0040$).



4 pav. HBK duomenų projekcijos plokštumoje: a) (**S1**) ($E_s = 0,0112$); b) (**S3**) ($E_s = 0,0050$).

algoritmas (**S1**) palygintas su algoritmais (**S2**) ir (**S3**), kuriuose buvo realizuotos naujos Sammono projekcijos paklaidos minimizavimo strategijos.

Naudojant (**S2**) ir (**S3**) algoritmus gaunama mažesnė projekcijos paklaida lyginant su klasikiniu Sammono algoritmu (**S1**) ir žymiai greičiau. Todėl vaizdų projekcijos tikslesnės, be to gaunama mažesnė jų priklausomybė nuo „magiškojo faktoriaus“ α reikšmės.

Pradinis dvimačių vektorių inicijavimas ant įstrižainės lėtina vizualizavimo konvergavimą, todėl pirmiausia taškus reikia išmėtyti. Taškų išbarstymas turi būti apibrėžtas. Vienas iš būdų – taikomas mažėjantis triukšmas II eilės išvestinei pirmosiose iteracijose – (**S3**) metodas.

Literatūra

1. N.R. Draper, H. Smith, *Applied Regression Analysis*, John Wiley and Sons, New York (1966).
2. G. Dzemyda, O. Kurasova, Visualization of multidimensional data taking into account the learning flow of the self-organizing neural network, *Journal of WSCG*, **11**(1), 117–124 (2003).
3. R.A. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179–188 (1936).
4. D.M. Hawkins, D. Bradu, G.V. Kass, Location of several outliers in multiple regression data using elemental sets, *Technometrics*, **26**,197–208 (1984).

5. A. Karbowski, Direct method of hierarchical nonlinear optimization – reassessment after 30 years, in: *Proceedings of III International Conference on Decision Support for Telecommunications and Information Society* (DSTIS 2003), Warsaw (2003).
6. S. Kaski, *Data Exploration Using Self-Organizing Maps*, PhD thesis, Helsinki University of Technology, Department of Computer Science and Engineering (1997). http://www.cis.hut.fi/~sami/thesis/
7. B. Kvedaras, M. Sapagovas, *Skaičiavimo metodai*, Mintis (1974).
8. J.W. Sammon, A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers*, **C-18**, 401–409 (1969).
9. P. Taylor, *Statistical Methods. Intelligent Data Analysis: an Introduction*, edited by M. Berthold, D.J. Hand, Springer-Verlag, 69–129 (2003).

SUMMARY

G. Dzemyda, J. Bernatavičienė, O. Kurasova, V. Marcinkevičius. Strategies of minimization of Sammon's mapping error

The classic algorithm for Sammon's projection and two new its modifications are examined in details. All the algorithms are oriented to minimize the projection error. The discovered new ways for minimization of the projection error makes a background for the further research in this field.

Keywords: visualization, projection error, minimization, Sammon's mapping.