Liet. matem. rink., **44**, *spec. nr.*, 2004, 670–674

Daugiamačių duomenų netolygumo analizės ypatybės

Vydūnas ŠALTENIS (MII, VPU)
el. paštas: saltenis@ktl.mii.lt

1. Įvadas

Daugiamačiai duomenys dažnai analizuojami siekiant nustatyti atskirų duomenų taškų ar jų grupių išsidėstymo netolygumus: rasti išsiskiriančius duomenų taškus ar grupuoti duomenis į tam tikru požiūriu artimas homogenines grupes – klasterius.

Išsiskiriantys duomenų taškai gali rastis kaip triukšmas, matavimų paklaidų rezultatas; kai kuriuose uždaviniuose išsiskiriantys taškai gali būti ypač įdomūs.

Išsiskiriančių duomenų taškų bei klasterių paieška tarpusavyje susijusi. Dažnai suprantama, kad taškas išsiskiria, jei jis nėra arti klasterio. Šiuo požiūriu klasterizavimo algoritmai generuoja išsiskiriančius taškus kaip pašalinį produktą (pavyzdžiui, [2]). Daugybė žinomų klasterizavimo bei išsiskiriančių taškų paieškos algoritmų labai skirtingai traktuoja duomenų klasterius, jų formas, remiasi skirtingomis prielaidomis. Jų darbo rezultatai priklauso nuo algoritmų parametrų parinkimo (pavyzdžiui, k-vidurkių metodui reikia užduoti klasterių skaičių).

Darbe įvedamas duomenų taškų išskirtinumo matas, kurio reikšmė būtų pagrindas priskirti tašką prie išsiskiriančių ar, priešingai, prie kurio nors klasterio. Panašus matas naudojamas [5], tačiau jis iš easmės priklauso nuo įtakos funkcijos bei jos parametrų parinkimo (pavyzdžiui, ši funkcija gali būti kvadratinė ar Gauso funkcija).

Įvestas duomenų taškų išskirtinumo matas grindžiamas taškų tarpusavio atstumų analize. Tam lyginamas atstumų tarp duomenų taškų pasiskirstymas su daugiamačiame kube tolygiai pasiskirsčiusių taškų tarpusavio atstumų pasiskirstymu.

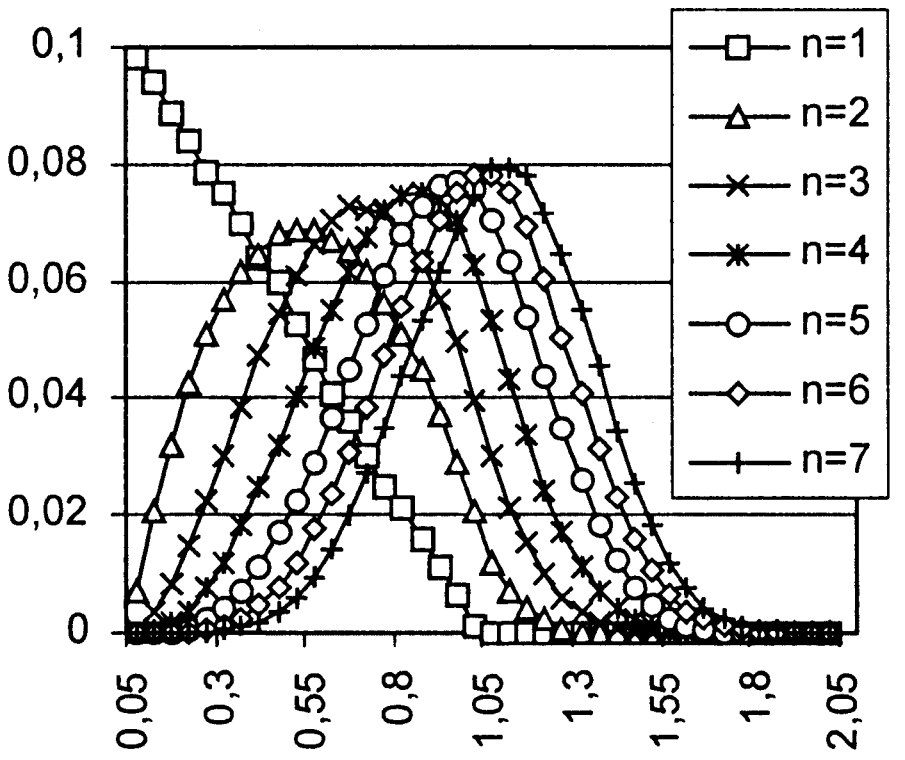
2. Atstumų tarp tolygiai pasiskirsčiusių taškų daugiamačiame kube pasiskirstymas

Analitiniai šių atstumų pasiskirstymai žinomi tik atskirais atvejais, esant mažam matavimų skaičiui n.

Tegul n-mačiai vektoriai tolygiai pasiskirstę vienetiniame daugiamačiame kube $[0..1]^n$. Tada vektoriaus v L_k-atstumas apibrėžiamas, kaip $\|v\|_k = \sqrt[k]{\sum_{i=1}^{n} |v_i|^k}$, o L_k-atstumas tarp dviejų vektorių v ir w, kaip $\|v - w\|_k$.

L_2-atstumų pasiskirstymų analizę galima rasti [8]. Kai $n = 1$ tankio funkcija lygi

$$\phi_{|V-V|}(x) = \begin{cases} \frac{d(2x-x^2)}{dx} = 2 - 2x, & \text{kai } 0 \leqslant x \leqslant 1, \\ 0, & \text{priešingu atveju,} \end{cases}$$



1 pav. L_2 atstumų tarp tolygiai pasiskirsčiusių daugiamačių taškų histograma, kai $n = \overline{1,7}$.

o atstumų vidurkis ir standartinis nuokrypis atitinkamai lygūs

$$E_{|V-V|} = \frac{1}{3} \quad \text{ir} \quad \sigma_{|V-V|} = \frac{1}{\sqrt{18}}.$$

$n > 1$ atvejams galime rasti tik apytikslius įverčius [8]:

$$E_{|V-V|} \xrightarrow[p]{} \sqrt{\frac{n}{6}}. \tag{1}$$

Modeliavimo eksperimentai patvirtina šį rezultatą didesniems n. Mažesnėms n reikšmėms (1) įverčiai skiriasi (pavyzdžiui, jei $n = 1$, tiksli vidurkio reikšmė lygi 1/3).

Standartinio nuokrypio tikslus radimas taip pat neįmanomas, kai $n > 1$. Modeliavimo eksperimentai rodo, kad didinant n jis artėja prie 0,24.

Kadangi siūloma metodika remiasi minėtų dydžių pasiskirstymu, nesant tikslių įverčių buvo panaudoti modeliavimo būdu gautas pasiskirstymas. Tam daugiamačiame vienetiniame kube buvo 1000000 kartų skaičiuotas L_2 atstumas tarp dviejų tolygiai pasiskirsčiusių taškų prie įvairių n. Rezultatai pateikti 1 pav. matavimų skaičiams 1–7.

Matome, kad pasiskirstymai netolygūs, ypač didesniems n.

3. Pagrindinė idėja ir duomenų taškų išskirtinumo matas

Norint analizuoti duomenų taškų netolygumą patogu eliminuoti aštrią tarpusavio atstumų pasiskirstymo kreivės viršūnę. Tam tikslinga nagrinėti ne analizuojamų duomenų tarpusavio atstumų pasiskirstymo funkciją $f^n(d)$, o jos skirtumą su analogiška funkcija tolygiai pasiskirsčiusiems taškams $f^u(d)$. Šių funkcijų skirtumas – atstumų dažnio funkcija (ADF)

$$f(d) = f^n(d) - f^u(d) \tag{2}$$

pasižymi tokiomis naudingomis savybėmis:

- jei duomenų taškai daugiamatėje srityje pasiskirstę tolygiai, ADF reikšmės bus artimos nuliui visame tarpusavio atstumų d intervale;

- jei duomenų taškai pasiskirstę netolygiai, didesnės teigiamos ADF reikšmės atitiks dažniau pasitaikantiems, tipiniams taškų tarpusavio atstumams d;
- mažoms ADF reikšmėms atitiks netipiniai taškų tarpusavio atstumai d.

Kiekvienam taškui i galima suskaičiuoti duomenų taško *išskirtinumo matą*:

$$R_i = 1/m \sum_{\substack{j=1 \\ j \neq i}}^{m} f\big(d(X_i, X_j)\big), \tag{3}$$

kur m – duomenų taškų skaičius, $d(X_i, X_j)$ – atstumas tarp i-jo ir j-jo taškų, X – jų koordinatės, $f(d)$ – ADF funkcija (2).

Išsiskiriančiam duomenų taškui atitiks mažiausios išskirtinumo mato R reikšmės, kadangi atstumai tarp jo ir likusių taškų bus retai pasitaikantys, netipiški, o tuo pačiu sumos (3) nariai – maži.

4. Eksperimentinis algoritmų galimybių įvertinimas

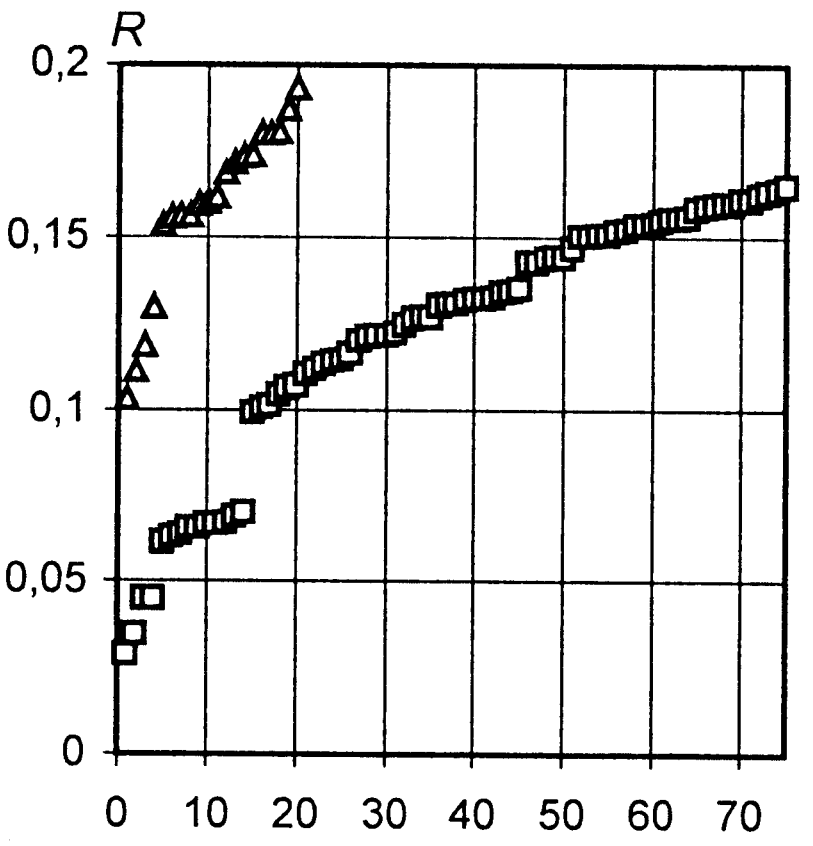
Išsiskiriančių taškų radimo algoritmas randa taškus su mažiausiu išskirtinumo matu. Metodo veikimo kokybė lyginta su keturiais žinomais metodais: Donoho–Stahel [6], Hadi [3], MML klasterizavimo [7] ir replikacinių neuroninių tinklų (RNN) [9]. Naudotasi dviem paplitusiais daugiamačių duomenų rinkiniais.

HBK duomenys [4] (75 4-mačiai taškai) pasižymi 14 pirmųjų išsiskiriančių duomenų taškų. Jų tarpe ypatingai išsiskiria 11–14 taškai. Pasiūlyto metodo ir kitų keturių metodų rezultatai analizuojant HBK duomenis, iliustruojami 1 lentelėje.

Duomenų taškų numeriai lentelėje pateikti išskirtinumo mažėjimo tvarka. 1–14 išsiskiriantieji taškai spausdinti kursyvu, o ypatingai išsiskiriantieji taškai 11–14 dar ir pabraukti. Pirmųjų keturių stulpelių duomenys palyginimui paimti iš [9].

1 lentelė. Labiausiai išsiskiriantys duomenų taškai, rasti įvairiais metodais

Donoho–Stahel	Hadi	MML klasterizavimo	RNN	Pasiūlytas metodas
14	*14*	*12*	*14*	*11*
4	*4*	*14*	*12*	*14*
3	*5*	*13*	*13*	*13*
5	*3*	*11*	*11*	*12*
9	*9*	*4*	*7*	*5*
7	*7*	53	*6*	*3*
10	*12*	*7*	*8*	*1*
6	*10*	47	*3*	*7*
2	*6*	68	*1*	*8*
8	*2*	62	*2*	*10*
12	*8*	60	*10*	*2*
13	*13*	34	*5*	*9*
1	*11*	43	16	*6*
11	*1*	27	49	*4*



2 pav. Išskirtinumo mato reikšmės, išdėstytos didėjimo tvarka:
HBK duomenims (kvadratinės žymės) ir Wood duomenims (trikampės žymės).

Matome, kad pasiūlytas metodas geriausiai atskiria tiek pirmuosius 14 išsiskiriančių taškų, tiek keturis ypač išsiskiriančius taškus. Tuo tarpu kiti metodai daugeliu atvejų klysta.

Wood duomenyse [1] (20 6-mačių taškų) išsiskiria 4 duomenų taškai. Jų išskyrimo eksperimentai taikant tuos pačius keturis metodus, taip pat pademonstravo tikslų pasiūlyto metodo veikimą. Tuo tarpu kiti metodai veikė netiksliai.

2 pav. iliustruojamos išskirtinumo mato reikšmės, išdėstytos didėjimo tvarka. Matome, kad išsiskiriančių taškų reikšmės žymiai mažesnės abiems duomenims.

Klasterizavimo algoritmas naudoja ne išskirtinumo matą (3), o jo modifikaciją

$$R(X) = 1/m \sum_{j=1}^{m} f\big(d(X, X_j)\big).$$

Algoritmo idėja ta, kad šios funkcijos lokalūs maksimumai sutampa su klasterių centrais, todėl klasterizavimui pakanka ieškti lokalaus maksimumo, pradinėmis koordinatėmis pasirenkant analizuojamų duomenų taškų koordinates. Jei randami tie patys lokalūs maksimumai, ir duomenų taškai priklauso tam pačiam klasteriui.

Preliminarūs klasterizavimo eksperimentai rodo, kad algoritmas atsparus papildomų triukšminių duomenų taškų įvedimui. Taip pat algoritmas gerai skiria įvairių geometrinių formų klasterius.

5. Išvados

Straipsnyje įvestas duomenų taškų išskirtinumo matas, leidžiantis naujai ir efektyviai vykdyti daugiamačių duomenų analizę: tiek išsiskiriančių duomenų taškų paiešką, tiek duomenų klasterizavimą. Pasiūlytiems algoritmams nebūtinas parametrų parinkimas – metodas natūraliai prisitaiko prie analizuojamų duomenų struktūros.

Eksperimentai pasiūlytu algoritmu analizuojant plačiai paplitusius testinius išsiskiriančių duomenų taškų paieškos uždavinius parodė jo geresnį veikimą lyginant su žinomais metodais.

Literatūra

1. N.R. Draper, H. Smith, *Applied Regression Analysis*, John Wiley and Sons, New York (1966).
2. M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (1996), pp. 226–231.
3. A.S. Hadi, A modification of a method for the detection of outliers in multivariate samples, *Journal of the Royal Statistical Society*, B, **56**(2), 393–396 (1994).
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. Hinneburg, D. Keim, An efficient approach to clustering large multimedia databases with noise, in: *Proceedings of the 4th ACM SIGKDD*, New York, NY (1998), pp. 58–65.
6. E.M. Knorr, R.T. Ng, R.H.Zamar, Robust space transformations for distance-based operations, in: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD01)*, San Francisco, California (2001), pp. 126–135.
7. J.J. Oliver, R.A. Baxter, C.S. Wallace, Unsupervised learning using MML, in: *Proceedings of the Thirteenth International Conference (ICML 96)*, Morgan Kaufmann Publishers, San Francisco (1996), pp. 364–372.
8. L. Schmitt, *Nearest Neighbor Search in High Dimensional Space by Using Convex Hulls*, Preprint No. 6/01, Fakultat fur Informatik, Universitat Magdeburg, 1–30 (2001).
9. G. Williams, R. Baxter, H. He, S. Hawkins, L. Gu, *A Comparative Study of Replicator Neural Networks for Outlier Detection in Data Mining*, CSIRO Technical Report CMIS-02/102, Canbera, Australia, 1–16 (2002).

SUMMARY

V. Šaltenis. The properties of nonuniformity analysis of high dimensional data

A novel approach to outlier detection and clustering on the ground of the distribution of distances between multidimensional points is presented. The basic idea is to evaluate the outlier factor for each data point.

A comparison with some popular outlier detection and clustering methods shows the superiority of our approach.

Keywords: outlier detection, high-dimensional data, distribuion of distances.