Liet. matem. rink, **45***, spec. nr.*, 2005, 442–446

Išsiskiriančių duomenų paieškos algoritmo tyrimai

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

1. Įvadas

Daugiamačių duomenų taškų išskirtinumo matas, autoriaus įvestas [9, 10], grindžiamas duomenų tarpusavio atstumų analize. Vertinant šio mato reikšmes siūloma lyginti tarpusavio atstumų tarp duomenų pasiskirstymą su daugiamačiame kube tolygiai pasiskirsčiusių taškų tarpusavio atstumų analogišku pasiskirstymu.

Daugiamačiai duomenys analizuojami siekiant nustatyti duomenų netolygumus: rasti išsiskiriančius duomenis ar grupuoti duomenis į klasterius. Preliminarūs taškų išskirtinumo matu pagrįsti taikymai, ieškant išsiskiriančių duomenų testiniuose duomenų rinkiniuose pademonstravo gerus rezultatus, pralenkiančius keletą žinomų metodų [3, 6, 7, 11].

Straipsnis skirtas tolesniam šio mato taikymo galimybių plėtimui ir tyrimui.

Kadangi analizuojamų taškų skaičius realiose daugiamačių taškų aibėse gali būti didelis, vertinamų tarpusavio atstumų skaičius gali kelti skaičiuojamųjų sunkumų. Todėl tikslinga ištirti tik dalies tarpusavio atstumų įvertinimo įtaką.

Taip pat aktualūs tyrimai, vertinantys papildomų (triukšmo) duomenų taškų įtaką algoritmo darbo kokybei. Tai algoritmo patikimumo tyrimai.

2. Duomenų išskirtinumo mato idėja ir preliminarių išskyrimo kokybės tyrimų rezultatai

Duomenų išskirtinumo matas pagrindžia, ar tašką priskirti prie išsiskiriančių ar, priešingai, prie kurio nors klasterio. Panašus matas naudojamas [5], tačiau jis esmingai priklauso nuo įtakos funkcijos bei jos parametrų parinkimo (pavyzdžiui, ši funkcija gali būti kvadratinė ar Gauso funkcija). Straipsnyje naudojamas matas nepriklauso nuo parametrų ir gali būti traktuojamas, kaip prisitaikantis prie tarpusavio atstumų struktūros.

Autoriai S. Brin [1] ir M.L. Steinbach [8] atkreipė dėmesį, kad daugiamačių duomenų pasiskirstymą patogu analizuoti pasitelkus taškų tarpusavio atstumų histogramą. Pavyzdžiui, esant duomenų klasteriams šioje histogramoje išryškės dvi viršūnės. Tiesa, tokias histogramas sunku analizuoti, nes ir tolygaus taškų pasiskirstymo atveju tarpusavio atstumai labai netolygiai pasiskirstę, ypač didesniam matavimų skaičiui. Eliminuojant aštrią tarpusavio atstumų pasiskirstymo histogramos viršūnę tikslinga nagrinėti ne 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)

$$f(d) = f^n(d) - f^u(d) \tag{1}$$

pasižymi naudingomis savybėmis:

- jei taškai pasiskirstę tolygiai, ADF reikšmės artimos nuliui visame atstumų d intervale;
- jei taškai pasiskirstę netolygiai, didesnės ADF reikšmės atitiks dažniau pasitaikantiems taškų tarpusavio atstumams d;
- mažoms ADF reikšmėms atitiks netipiniai taškų tarpusavio atstumai d.

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

$$R_i = 1/m \sum_{\substack{j=1 \\ j \neq i}}^{m} f\big(d(X_i, X_j)\big), \tag{2}$$

kur m – taškų skaičius, $d(X_i, X_j)$ – atstumas tarp i-jo ir j-jo taškų, X – jų koordinatės, $f(d)$ – ADF funkcija (1).

Išsiskiriančiam duomenų taškui atitiks mažiausios išskirtinumo mato R reikšmės, kadangi atstumai tarp jo ir likusių taškų bus retai pasitaikantys, netipiški, o tuo pačiu sumos (2) nariai – maži. Išsiskiriančių taškų radimo algoritmas randa taškus su mažiausiu išskirtinumo matu.

Algoritmo kokybė lyginta [10] su keturiais žinomais metodais: Donoho–Stahel [6], Hadi [3], MML klasterizavimo [7] ir replikacinių neuroninių tinklų (RNN) [11]. Naudotasi dviem paplitusiais daugiamačių duomenų rinkiniais: HBK [4] ir Wood [2] duomenimis. HBK duomenys (75 4-mačiai taškai) turi 14 išsiskiriančių duomenų taškų. Jų tarpe ypatingai išsiskiria 4 taškai.

Wood duomenys – 20 šešiamačių taškų, tarp kurių ryškiai išsiskiria keturių duomenų taškų grupė.

Pasiūlytas metodas be klaidų atskyrė visus išsiskiriančius taškus abiejose duomenų aibėse. Tuo tarpu kiti minėti metodai daugeliu atvejų klydo.

3. Išsiskiriančių taškų paieškos galimybių plėtimo tyrimai, klaidų įvertinimas

Preliminariuose algoritmo kokybės eksperimentiniuose tyrimuose [10] buvo pademonstruotas patikimas keturių taškų Wood duomenyse išskyrimas. Tai lėmė šių duomenų pasirinkimą šio tolesniems tyrimams, kuriuose išskyrimas vykdomas apsunkintomis sąlygomis, iššaukiančiomis išskyrimo klaidas.

Pirmoji šio darbo tyrimų grupė skirta išskyrimo metodo galimybėms plėsti, mažinant skaičiavimo laiką.

Įvestas išskirtinumo matas remiasi visų m duomenų taškų tarpusavio atstumų $d(X_i, X_j)$ analize, kur X_i, X_j – taškų koordinatės, $i, j = 1, \ldots, m$, o tarpusavio atstumų skaičius lygus

$$M = \frac{m(m-1)}{2}.$$

Išsiskiriančių duomenų paieškos algoritmo darbo laikas kvadratiškai priklauso nuo taškų skaičiaus m. Tai kliūtis analizuoti didesnės apimties duomenų aibes.

Buvo eksperimentiškai vertinama, kokią įtaką išsiskirančių taškų išskyrimo kokybei turi tarpusavio atstumų skaičiaus apribojimas. Analizuojamų tarpusavio atstumų dalis buvo atrenkama atsitiktinai parenkant analizuojamų taškų numerius. Siekiant patikimesnių tyrimų rezultatų duomenų analizė buvo kartojama daug kartų, klaidų skaičių vidurkinant.

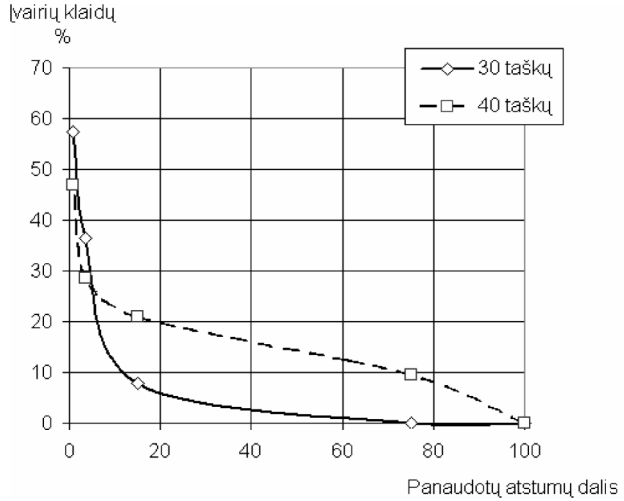
Antroji tyrimų grupė skirta papildomų (triukšmo) duomenų taškų įtakai algoritmo rezultatams įvertinti. Tai algoritmo patikimumo tyrimai, nes pašalinių duomenų taškų įtaka dažnai trikdo duomenų analizę. Šiuose tyrimuose taip pat naudoti Wood duomenys, papildyti tolygiai srityje pasiskirsčiusių taškų skaičiumi. Keičiant triukšminių taškų skaičių buvo matuojamas išskyrimo klaidų skaičius.

Abiejose tyrimų grupėse klaida buvo laikomas neteisingas vieno iš keturių išsiskiriančių taškų išskyrimas, t.y. nepatekimas į keturių taškų su mažiausiais išskirtinumo matais grupę. Tarp visų išskyrimo klaidų atskirai buvo fiksuojamos grubios klaidos, kai neteisingai išskiriama du ir daugiau išsiskiriančių duomenų taškų.

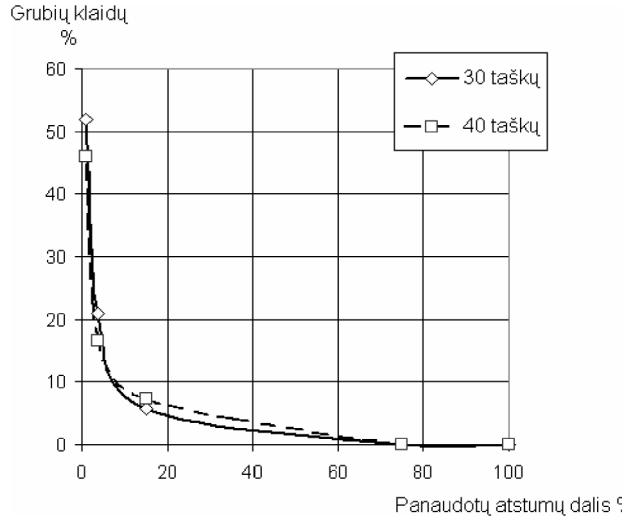
4. Tyrimų rezultatai

Dalies tarpusavio atstumų tarp taškų panaudojimo įtakos tyrimų rezultatai iliustruojami 1 ir 2 pav. Eksperimentuota su 30 ir 40 taškų aibėmis (tai Wood duomenys, papildyti iki reikiamo skaičiaus tolygiai pasiskirsčiusiais taškais).

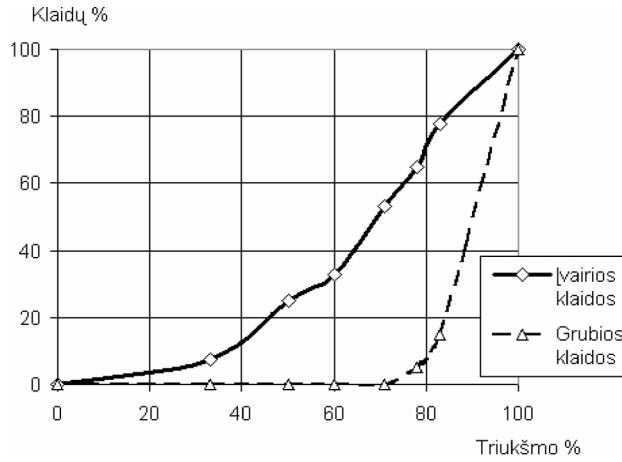
Triukšmo įtakos išsiskiriančių duomenų paieškos kokybei tyrimų rezultatai iliustruojami 3 pav.



1 pav. Įvairių klaidų procento priklausomybė nuo analizėje panaudotų tarpusavio atstumų dalies analizuojant 30 ir 40 taškų duomenų aibes.



2 pav. Grubių klaidų procento priklausomybė nuo analizėje panaudotų tarpusavio atstumų dalies analizuojant 30 ir 40 taškų duomenų aibes.



3 pav. Klaidų procento priklausomybė nuo papildomų duomenų taškų (triukšmo) dalies analizuojant Wood duomenų aibes.

5. Išvados

Straipsnyje pateikti autoriaus pasiūlyto išsiskiriančių daugiamačių duomenų paieškos algoritmo tyrimų rezultatai (1 ir 2 pav.) leidžia teigti, kad ribojant tarpusavio atstumų panaudojimą iki 20% klaidų skaičius išauga nežymiai, iki 5–10%.

Papildomų taškų pridėjimas iki 70% nesukelia žymesnio grubių klaidų skaičiaus, tuo tarpu nežymių klaidų skaičius auga daug greičiau (žiūr. 3 pav.).

Literatūra

1. S. Brin, Near neighbor search in large metric spaces, in: *Proceedings of the 21st International Conference on Very Large Databases* (*VLDB-1995*), Morgan Kaufmann, Zurich, Switzerland (1995), pp. 574–584.
2. N.R. Draper, H. Smith, *Applied Regression Analysis*, John Wiley and Sons, New York (1966).
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. M.L. Steinbach, M.L. Ertoz, V. Kumar, *Challenges of Clustering High Dimensional Data*, New Vistas in Statistical Physics. Applications in Econophysics, Bioinformatics, and Pattern Recognition, Springer-Verlag (2003).
9. V. Šaltenis, Daugiamačių duomenų netolygumo analizės ypatybės, *Liet. matem. rink.*, **44** (spec. nr.), 670–674 (2004).
10. V. Šaltenis, Outlier detection based on the distribution of distances between data points, *Informatica*, **15**(3), 399–410 (2004).
11. 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. Investigation of outlier detection algorithm

The proposed outlier factor was used to analyze the multidimensional data sets regarding outlier detection. The paper describes two kinds of investigation: the influence of omitting some part of distances between data points, and the influence of additional (noisy) points to outlier detection quality. The results demonstrate the possibilities to improve the performance of computation and the stability of the outlier detection algorithm.

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