DAUGIAMAČIŲ DUOMENŲ POŽYMIŲ MAŽINIMAS NAUDOJANTIS EKSPONENTINE KORELIACINE FUNKCIJA

Laura Ringienė, Gintautas Dzemyda
Vilniaus universitetas, Matematikos ir informatikos institutas

Įvadas

Daugiamačiai duomenys paprastai aprašo objektų (žmonių, įrenginių, augalų, gamtos reiškinių...) rinkinius, kuriuos charakterizuoja tam tikri skaitiniai požymiai, dar vadinami parametrais ir kt. Objektų, sudarančių konkrečią analizuojamų objektų aibę, skaičius m yra baigtinis. Tam tikras požymių reikšmių rinkinys nusako vieną konkretų analizuojamos aibės objektą $X_i = (x_{i1}, x_{i2}, ..., x_{in})$, $i = \overline{1,m}$, čia n yra požymių skaičius, i yra objekto numeris. Objektai X_i dar gali būti interpretuojami kaip taškai ar vektoriai, o požymiai $x_1, x_2, ..., x_n$ – taškų ar vektorių komponentėmis. Analizuojamų duomenų aibę galima aprašyti kaip matricą $\boldsymbol{X} = \{X_1, X_2, ..., X_m\} = \{x_{ij}, i = \overline{1,m}, j = \overline{1,n}\}$, kurios i-oji eilutė yra n-matės euklidinės erdvės taškas $X_i \in R^n$ (Dzemyda, Kurasova, Žilinskas, 2013).

Dėl duomenų požymių gausos didelės apimties daugiamačius duomenis žmogui sunku suvokti: nustatyti struktūrą, tarpusavio ryšius, susidariusias grupes, prognozuojamus įverčius ir pan. Tuo tikslu daugiamačių duomenų vizualizavimui siūloma daug metodų. Vizualizavimas – tai grafinis informacijos pateikimas. Vizualizavimo sąvoka gana plati, bet mes nagrinėsime daugiamačių duomenų vizualizavimo metodus, padedančius nustatyti ar įvertinti daugiamačių duomenų, nusakančių objektų rinkinį, struktūrą (objektų grupių tarpusavio panašumus, išsiskiriančius objektus ir pan.). Yra dvi pagrindinės metodų grupės daugiamačiams duomenims vizualizuoti – tiesioginio vizualizavimo metodai ir projekcijos, dar vadinamieji matmenų skaičiaus mažinimo metodai. Projekcijos metodai transformuoja daugiamačių duomenų aibę $X = \{X_1, X_2, ..., X_m\}$ iš erdvės R^n į mažesnio matmenų skaičiaus vaizdo erdvę R^d, ($d < n$), kur duomenų aibės transformacijos $\boldsymbol{Y} = \{Y_1, Y_2, ..., Y_m\} = \{y_{ij}, i = \overline{1,m}, j = \overline{1,d}\}$ taškų išsidėstymą galima stebėti vizualiai, kai $d = 1$, 2 ar 3 (Dzemyda, Kurasova, Žilinskas, 2013).

Tikslas – sukurti metodą daugiamačių duomenų požymių skaičiui, sumažinti naudojantis eksponentine koreliacine funkcija, kai atsižvelgiama, jog daugiamačiuose duomenyse yra panašių duomenų klasteriai.

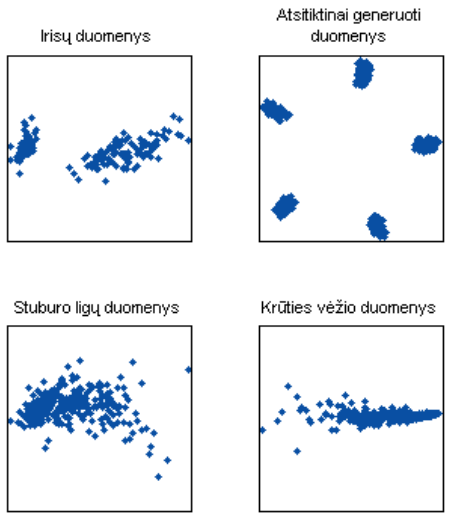
Metodas apima daugiamačių duomenų klasterizavimą į tam tikrą klasterių skaičių, duomenų transformavimą į k-matę erdvę R^k ir jau transformuotų k-mačių duomenų vizualizavimą, taikant projekcijos metodą. Čia gali būti vartojamas bet koks projekcijos (tiesinės ar netiesinės) metodas.

Tyrimo metodai – eksperimentiniai tyrimai.

Vizualizavimas daugiamatėmis skalėmis

Daugiamačių skalių (angl. *Multidimensional scaling*, MDS) metodas (Borg and Groenen 2005) – tai grupė metodų, plačiai taikomų daugiamačių duomenų vizualizavimui. Naudojantis MDS, ieškoma taško $X_i = (x_{i1}, x_{i2}, ..., x_{in})$ projekcijos $Y_i = (y_{i1}, y_{i2}, ..., y_{id})$ į mažesnio skaičiaus matmenų erdvę R^d, ($d < n$) (dažniausiai R^2 arba R^3), siekiant išlaikyti analizuojamos aibės objektų (taškų) panašumus. Atlikus projekciją į mažesnio matavimo vaizdo erdvę, panašūs objektai išdėstomi arčiau vieni kitų, o skirtingi – toliau vieni nuo kitų (Dzemyda, Kurasova, Žilinskas, 2013).

MDS veikimas iliustruotas 1 paveiksle. Priede pateiktos keturios vizualizuotos duomenų aibės į R^2 erdvę, t. y. $d = 2$. Paveiksluose nepristatytas skalių žymėjimas, nes aktualu tik taškų tarpusavio išsidėstymas. Toks vizualus duomenų pateikimas leidžia tyrinėtojui „pajusti“ daugiamačių vektorių tarpusavio atstumus, o tai palengvina duomenų visumos pažinimą (Dzemyda, Kurasova, Žilinskas, 2013). Akivaizdu, kad Irisų duomenyse aiškiai atsiskiria vienas klasteris, o tarp kitų dviejų klasterių aiškios ribos nėra. Atsitiktinai generuotuose duomenyse aiškiai išsiskiria 5 klasteriai. Kitose dviejose duomenų aibėse aiškios skiriamosios ribos tarp klasterių nėra.



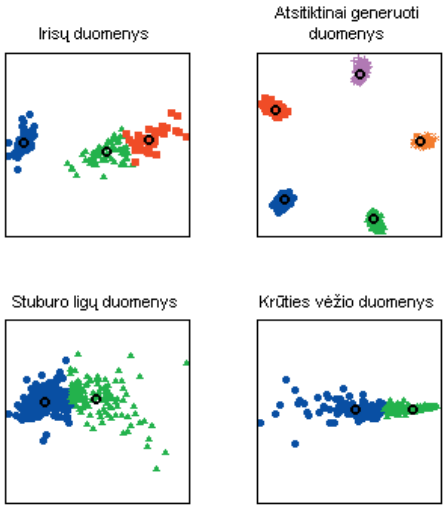
1 pav. Tyrimuose naudojami daugiamačiai duomenys, vizualizuoti MDS metodu

Daugiamačių skalių metodą duomenims vizualizuoti galima taikyti iš karto. Tačiau kyla idėja – gal pirma atlikti tam tikrą daugiamačių duomenų netiesinę transformaciją, paryškinančią klasterius tuose duomenyse, o po to tik vizualizuoti tuos klasterizuotus duomenis, siekiant dar geriau pamatyti objektų grupes.

Vienas iš vizualios duomenų analizės tikslų – atrasti ar net pamatyti duomenų klasterius. Bendru atveju norėdami atrasti klasterius duomenyse ir sužinoti tų klasterių centrus, turime naudotis specialiais, tam skirtais klasterizavimo metodais. Klasterizavimas (angl. *clustering*) – tai toks analizuojamų objektų suskirsty-

mas į skirtingas grupes, dar vadinamus klasterius (angl. *clusters*), kad grupės objektai būtų panašūs tarpusavyje, o objektai iš skirtingų grupių būtų nepanašūs. Duomenis suskirstyti į klasterius galima bet kuriuo pasirinktu klasterizavimo metodu (*k*-vidurkių (angl. *k-means*), klasifikavimo medžiu, *k* artimiausių kaimynų ar kt. (Han, Kambre, Pei, 2011; Dzemyda, Kurasova, Žilinskas, 2013). Šiame straipsnyje klasterizavimas yra vidinė siūlomo metodo procedūra. Taikytas *k*-vidurkių klasterizavimo metodas, kuriuo nustatomi klasterių centrai $\mu_j = (\mu_{j1}, \mu_{j2}, ..., \mu_{jn})$, $\mu_j \in R^n$, $j = \overline{1,k}$. Siekiant rezultatų objektyvumo, šiuose tyrimuose klasterizavimas vykdytas keletą kartų, nes klasterizavimo paklaidą nusakanti funkcija yra daugiaekstremė ir dažnai randamas tik lokalus, o ne globalus funkcijos minimumas.

Tolesniuose skaičiavimuose vartojami klasterizavimo rezultatai su mažiausiu lokaliu paklaidos minimumu, atliekant kelis kartus tų pačių duomenų klasterizavimą. 2 pav. pateikiami daugiamačių duomenų, kuriuos sudaro matrica *X* ir klasterizavimo metu gauti klasterių centrai $\mu_j = (\mu_{j1}, \mu_{j2}, ..., \mu_{jn})$, $\mu_j \in R^n$, $j = \overline{1,k}$, vizualizavimo MDS metodu rezultatai. Palyginus 1 ir 2 pav., matyti, kad taškų išsidėstymas plokštumoje faktiškai nepakito, o klasterių centrai μ_j, $j = \overline{1,k}$ yra savo klasterių viduryje.



2 pav. Klasterizuoti daugiamačiai duomenys, vizualizuoti MDS

Daugiamačių duomenų požymių skaičiaus mažinimas taikant eksponentinę koreliacinę funkciją

Ankstesniame skyriuje pateiktas daugiamačių duomenų požymių skaičiaus mažinimas MDS metodu, vykdant netiesinę duomenų projekciją į mažesnio matavimo erdvę.

Šiame skyriuje siūlomas metodas, kuris apima:

- daugiamačių duomenų klasterizavimą į tam tikrą klasterių skaičių *k*,
- duomenų transformavimą į *k*-matę erdvę R^k,
- jau transformuotų *k*-mačių duomenų vizualizavimą MDS metodu.

Metodo privalumas – klasterių, esančių daugiamačiuose duomenyse, išryškinimas dvimatėje tų duomenų projekcijoje.

Suskirsčius duomenis į norimą klasterių skaičių *k*, atliekamas daugiamačių duomenų $X_i = (x_{i1}, x_{i2}, ..., x_{in})$, $i = \overline{1,m}$, kur $X_i \in R^n$, požymių skaičiaus *n* (dimensiškumo) mažinimas, transformuojant $X_i \in R^n$ į $Z_i \in R^k$: $Z_i = (z_{i1}, z_{i2}, ..., z_{ik})$; čia $k < n$. Objekto $X = (x_1, x_2, ..., x_n)$ dimensiškumas mažinamas naudojantis tam tikra koreliacine funkcija. Gaunamas naujas objektas $Z = (z_1, z_2, ..., z_k)$, $k < n$, naudojantis šiomis formulėmis:

A. Eksponentinė koreliacinė funkcija (Yaglom, 1986):

$$z_j(x) = \exp\left(-\gamma \left\| X - \mu_j \right\|\right),\ j = \overline{1,k},\ \gamma = \frac{1}{2\sigma^2}, \qquad (1)$$

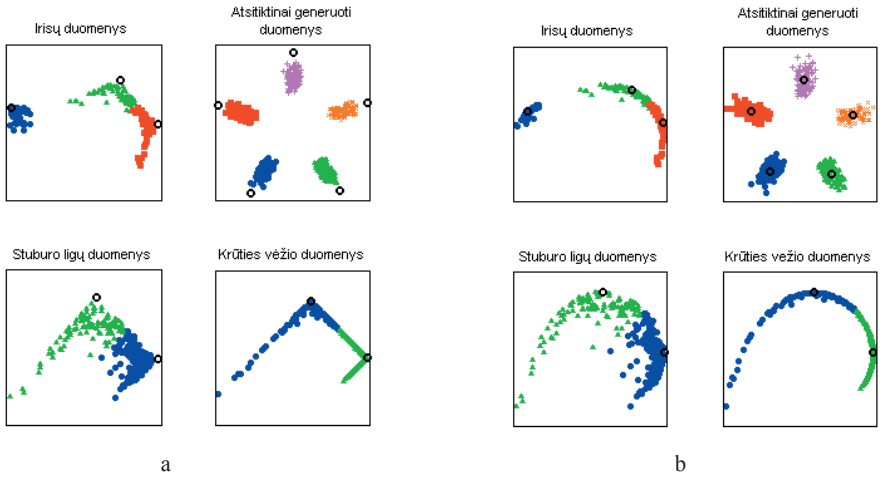
B. Gausinė koreliacinė funkcija (Stein, 1999):

$$z_j(x) = \exp\left(-\gamma \left\| X - \mu_j \right\|^2\right),\ j = \overline{1,k},\ \gamma = \frac{1}{2\sigma^2}, \qquad (2)$$

Čia μ_j yra koreliacinės *j*-tosios funkcijos centro taškas, $\mu_j \in R^n$, $\left\| X - \mu_j \right\|$ – atstumas tarp taškų *X* ir μ_j, σ – pločio parametras, nuo kurio priklauso funkcijos glotnumas. Pastebėsime, kad $\left\| X - \mu_j \right\| > 0$ ir $\gamma > 0$. Neretai funkcija (2) vadinama Gausine radialine bazine funkcija ir dažnai taikoma neuroniniuose tinkluose (Buhmann, 2003). Eksponentinės koreliacinės funkcijos skirtumas nuo Gausinės yra tik tai, kad eksponentinėje funkcijoje vartojimas atstumas, o Gausinėje – atstumo kvadratas. Remiantis (1) ar (2) formule, iš duomenų aibės *X* gaunama nauja duomenų aibė $Z = \{Z_1, Z_2, \ldots, Z_m\} = \{z_{ij}, i = \overline{1,m}, j = \overline{1,k}\}$, t. y. atlikta netiesinė duomenų aibės *X* transformacija, kur atsižvelgiama į klasterius šios aibės duomenyse.

Sumažinus daugiamačių duomenų požymių skaičių nuo *n* iki *k*, gauta duomenų aibė *Z* vizualizuojama į R^2 erdvę. Akivaizdu, kad jei klasterių skaičius $k > 2$, tai duomenys vizualizuojami į R^2 erdvę projekcijos metodai. Tolesniuose tyrimuose taikomas MDS metodas. Siekiant giliau atskleisti (1) ar (2) transformacijos savybes, vizualizuota ne tik aibė *Z*, bet kartu, naudojantis (1) ir (2) formulėmis, transformuoti centrai μ_j, $j = \overline{1,k}$.

Pažymėkime gautas svorių centrų transformacijas $\mu_j^z = \left(\mu_{j1}^z, \mu_{j2}^z, \ldots, \mu_{jk}^z\right) \in R^k$. Tad bendras vizualizuojamų objektų skaičius, koks ir ankstesniame skyriuje ir 2 pav., yra $m + k$. Rezultatai pateikiami 3 paveiksle. Skirtingų klasterių objektus atitinkantys taškai pažymėti •, ▲, ■, ×, +. Klasterių centrai pažymėti **O**.

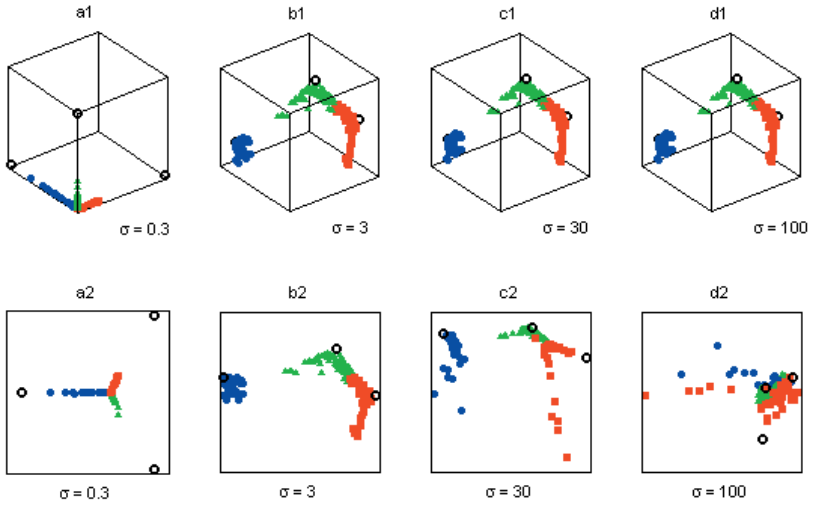


3 pav. Daugiamačiai duomenys, sumažinus požymius taikant a) eksponentinę koreliacinę funkciją, b) Gausinę koreliacinę funkciją

Eksponentinės koreliacinės funkcijos privalumas, lyginant su Gausine koreliacine funkcija, yra tai, kad klasterių centrai įgyja išskirtinę savybę būti tokiais taškais, kur keičiasi klasterių objektų ypatybės, ir tai matosi vizualiai (žr. 3a pav. Irisų, Stuburo ligų ir Krūties vėžio duomenų atvejus). Dėl to toliau straipsnyje bus išimtinai nagrinėjama eksponentinė koreliacinė funkcija.

3a pav. matyti, kad, lyginant su 2 pav., klasterių centrai μ_j^- pastumiami į šoną nuo atitinkamų klasterio taškų. Taip pat matyti, kad taškai, vizualizuoti po transformacijos eksponentine koreliacine funkcija, išsidėsto dvejopai:

1. *Izoliuotas klasteris*. Klasterio objektai vizualizuojami atskiroje grupėje. Pavyzdžiui, Irisų duomenyse toks yra klasteris pažymėtas •. Jo objektai koncentruojasi aiškiai matomame atskirame klasteryje. Taip pat labai aiškiai matomi atskiri (izoliuoti) klasteriai atsitiktinai generuotuose duomenyse.
2. *Tarpusavyje artimi klasteriai* (panašūs klasteriai). Vizualizuoti klasterio objektai išsibarsto dviejų tiesių aplinkoje, o tos tiesės susijungia ties klasterio centru. Objektų išsidėstymą tiesių aplinkoje geriausiai atspindi Krūties vėžio duomenys (3a pav.), taip pat gerai matyti ir Irisų bei Stuburo ligų duomenų atvejais. Arti tiesės, jungiančios kaimyninių klasterių centrus, vizualizuojami objektai, kurie turi panašumo su atitinkamo kaimyninio klasterio objektais.



4 pav. Irisų duomenys su skirtingais σ

Atliekant daugiamačių duomenų transformaciją iš $X_i \in R^n$ į $Z_i \in R^k$, $i = \overline{1,m}$, eksponentine koreliacine funkcija, svarbu tinkamai parinkti funkcijos parametrus – centrus μ_j ir pločio parametrą σ. Centrus, kaip ir dauguma autorių (Pierrefeu ir kt., 2006; Chang ir kt., 2005; Benoudjit, Verleysen, 2003) parenkame klasterizuodami duomenis k – vidurkių metodu. Tačiau koreliacinės funkcijos rezultatas priklauso ne vien nuo tinkamai parinktų centrų, bet ir nuo pločio parametro σ. Eksperimentai, atlikti su Irisų duomenimis, kai k = 3. 4 paveiksle pateikti Irisų duomenys vizualizuoti trimatėje ir dvimatėje erdvėje (trimačiu atveju Z_i, $i = \overline{1,m}$ vizualizuojami tiesiogiai, nes k = 3, o dvimačiu atveju vizualizuojama MDS metodu) su įvairiais σ parametrais (a) σ = 0,3; b) σ = 3;

c) σ = 30; d) σ = 100). Skirtingi vizualizavimo būdai pasirinkti tam, kad būtų galima nagrinėti rezultatus įvairiapusiškai. Peržvelgus visus vizualizavimo rezultatus po daugiamačių duomenų transformacijos, galima teigti: jei koreliacinės funkcijos pločio parametras parenkamas mažas (4 (a1, a2) pav.), tai visų klasterių taškai sustumiami į vieną visumą, o klasterių centrai yra išorėje. Tinkamai parinkus eksponentinės koreliacinės funkcijos pločio parametrą, duomenyse aiškiai išsiskiria klasteriai. Tačiau rezultatus vizualizavus į dvimatę erdvę MDS metodu, pastebima, kad pločio parametras gali būti per didelis (4 d2 pav.), nes vizualizuotų duomenų klasteriai „sulipa“ vienas ant kito. Iš paveikslų matyti, kad, vizualizavus rezultatus trimatėje erdvėje, pločio parametras σ didelės įtakos nedaro, nes gauti paveikslai (4 (b1, c1, d1) pav.) su skirtingomis σ reikšmėmis vizualiai atrodo panašūs. Tačiau po transformacijos gauti rezultatai, vizualizuoti dvimatėje erdvėje MDS metodu, taip pat labai vertingi (4 (b3, c3 d3) pav.), nes jie parodo, kad pločio parametras σ nuo tam tikros ribos gali tapti per didelis ir kad labai svarbu jį tinkamai parinkti.

Iš 4 pav. matyti, kad labai svarbu tinkamai parinkti pločio parametrą σ, bet vieningo koreliacinės funkcijos pločio parametro parinkimo būdo nėra. Jis priklauso nuo uždavinio.

Uždavinys – daugiamačių duomenų požymių skaičiaus mažinimas eksponentine koreliacine funkcija. Patirties, kaip parinkti tinkamą pločio parametrą σ, galime pasisemti ir iš radialinių bazinių funkcijų neuroninių tinklų uždavinių, kur kaip bazinė funkcija naudojama Gausinė koreliacinė funkcija. Tokio tipo neuroniniai tinklai plačiai taikomi vaizdams atpažinti, klasifikuoti, prognozuoti ir kitiems uždaviniams spręsti.

S. Haykin (1999) teigia, kad lengviausia atlikti skaičiavimus, kai visos bazinės funkcijos yra vienodos, t. y. izotropinė Gausinė funkcija, kurios standartinis nuokrypis (t. y. plotis σ_A) parenkamas atsižvelgiant į klasterių daugiamačiuose duomenyse centrų išsidėstymą. Konkrečiau, radialinė bazinė funkcija, kurios centras μ_j, apibrėžiama taip:

$$z_j(X) = \exp\left(-\frac{\|X-\mu_j\|^2}{2\sigma_A^2}\right) = \exp\left(-\frac{k}{d_{\max}^2}\|X-\mu_j\|^2\right), \quad j=\overline{1,k}, \tag{3}$$

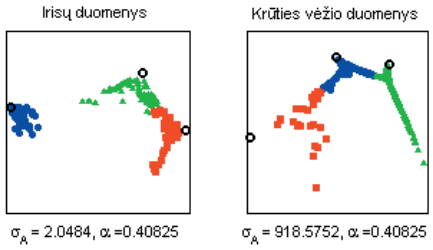
čia k – klasterių skaičius ir $d_{\max}$ – didžiausias atstumas tarp visų k klasterių centrų. Pločio parametras visoms Gausinėms radialinėms bazinėms funkcijoms fiksuotas:

$$\sigma_A = \frac{d_{\max}}{\sqrt{2k}} = \alpha d_{\max}, \text{ kur } \alpha = \frac{1}{\sqrt{2k}} \tag{4}$$

Tai sąlygoja šiuo analizuojamu atveju naudoti tokią eksponentinę koreliacinę funkciją:

$$z_j(X) = \exp\left(-\frac{\|X-\mu_j\|}{2\sigma_A^2}\right) = \exp\left(-\frac{k}{d_{\max}^2}\|X-\mu_j\|\right), \quad j=\overline{1,k}, \tag{5}$$

5 paveiksle pateikiami Irisų ir Krūties vėžio duomenys, gauti po transformacijos, kai σ apskaičiuojama pagal (4) formulę. Remiantis 4 paveiksle pateiktais rezultatais, galima daryti išvadą, kad Irisų duomenims apskaičiuotas pločio parametras σ_A yra tinkamas, tačiau Krūties vėžio duomenims pločio parametras σ_A per didelis, nes stebimas pirmojo (šio klasterio objektai pažymėti ■) ir antrojo (šio klasterio objektai pažymėti •) klasterių objektų judėjimas link trečiojo (jo objektai pažymėti ▲) klasterio centro. Taigi, pločio parametro σ_A apskaičiavimas pagal (4) formulę tinkamas ne visiems duomenims.



5 pav. Transformacijos rezultatai, gauti skaičiuojant σ pagal (4) formulę

S. Haykin (1999) pločio parametrui automatiškai parinkti naudoja didžiausią atstumą tarp klasterių centrų. Alternatyva yra vidutinis atstumas tarp jų (Pierrefeu ir kt., 2006). L. Pierrefeu ir kt. atliktų testų rezultatai, kai pločio parametras σ yra vidutinis atstumas tarp centrų, duoda gerus rezultatus ir lengvai pritaikomi. Faktiškai vidutinis atstumas nėra optimali reikšmė pločio parametrui. Geriausi rezultatai gaunami, kai pločio parametras yra apie 20 % mažesnis už vidutinį atstumą. Pasiūlytas metodas yra paprastas:

1. Apskaičiuojamas vidutinis atstumas tarp centrų:

$$d_{vid} = \frac{\sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k}\|\mu_i-\mu_j\|}{k(k-1)}, \tag{6}$$

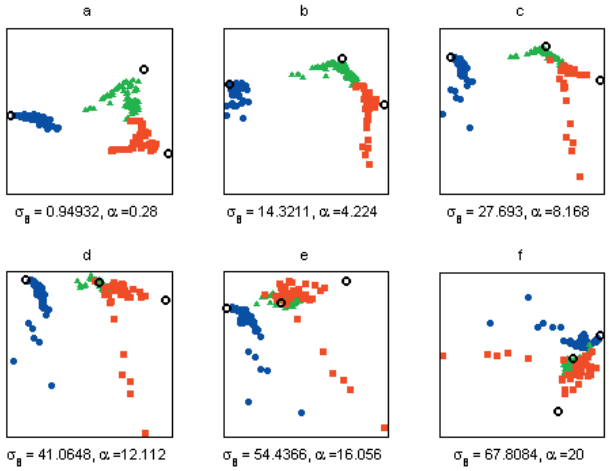
čia $\|\mu_i-\mu_j\|$ – euklidinis atstumas tarp centro taškų μ_i ir μ_j, k – klasterių skaičius.

2. Funkcijai $z_j(X) = \exp\left(-\frac{\|X-\mu_j\|}{2\sigma_B^2}\right)$ pločio parametras apskaičiuojamas taip:

$$\sigma_B = \alpha d_{vid}, \text{ kur } \alpha = \frac{1}{\beta}. \tag{7}$$

Straipsnyje (Pierrefeu ir kt., 2006) siūloma reikšmę keisti nuo 3,6 iki 0,05 kas 0,05, t. y. $\alpha \in [0.28,20]$.

Daugiamačių duomenų požymių skaičiaus mažinimo rezultatai, kai σ_B apskaičiuojama pagal (7) formulę, naudojant skirtingas α reikšmes (α kinta nuo 0,28 iki 20 kas 3,94), pateikti 6 pav.



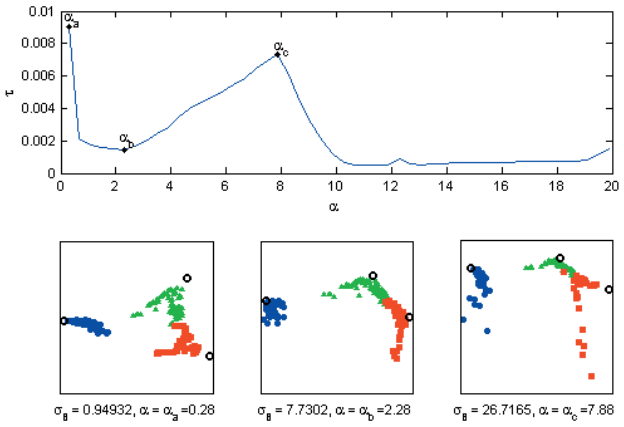
6 pav. Irisų duomenys po transformacijos su skirtingom σ_B reikšmėmis

Remiantis 4 paveikslo duomenimis, galima teigti, kad Irisų duomenims (6 a pav.) pločio parametras σ_B apskaičiuojamas šiek tiek per mažas, nes dar stebimas klasterių objektų judėjimas link klasterių centrų. 6 (e ir f) paveikslo rezultatai labai panašūs į 4 d3 paveikslo rezultatus, todėl galima teigti: kai $\alpha > 16$, tai pločio parametras σ_B apskaičiuojamas per didelis, nes, vizualizuojant duomenis, klasteriai „sulipa“ vienas ant kito. 6 (c ir d) paveiksle σ_B taip pat šiek tiek per didelis, nes stebimas klasterių (šių klasterių objektai pažymėti • ir ■) objektų judėjimas link viduriniojo klasterio centro. Atmetę po transformacijos gautus rezultatus, kai pločio parametras σ_B Irisų duomenims apskaičiuojamas per mažas arba per didelis, pastebime, kad tinkamai apskaičiuojamas eksponentinės koreliacinės funkcijos pločio parametras σ_B, kai $\alpha = 4{,}224$.

(7) formulėje α parinkimui siūlomas gana platus intervalas, o išsirinkti tinkamą α vizualizuojant kiekvieną variantą reikia daug laiko. Taigi, kad rastume tinkamą α, kad pločio parametras σ_B būtų tinkamas transformacijai, randamas maksimalus atstumas iš k minimalių atstumų tarp klasterių centrų $\mu_j^z = \left(\mu_{j1}^z, \mu_{j2}^z, \ldots, \mu_{jk}^z\right)$, $j = \overline{1,k}$ plokštuminių projekcijų $\mu_j^y = \left(\mu_{j1}^y, \mu_{j2}^y\right)$, $j = \overline{1,k}$ ir objektų $Z_i = \left(z_{i1}, z_{i2}, \ldots, z_{ik}\right)$, $i = \overline{1,m}$ plokštuminės projekcijos $Y_i = \left(y_{i1}, y_{i2}\right)$, $i = \overline{1,m}$.

$$\tau = \max_{j=\overline{1,k}} \left\{ \min_{X_i \in K_j} \left\| Y_i - \mu_j^y \right\| \right\}. \qquad (8)$$

Irisų duomenims tinkamo parametro α ieškoma intervale [0,28, 20]. τ priklausomybė nuo α pateikta 7 pav. Akivaizdu, kad didėjant α, τ reikšmė mažėja iki α pasiekia, po to didėja iki α pasiekia α_c, o po to vėl mažėja. 7 pav. pateikiami vizualizavimo rezultatai išskirtiniuose taškuose ($\alpha = \alpha_a = 0{,}28$, $\alpha = \alpha_b = 2{,}28$, $\alpha = \alpha_c = 7{,}88$). Nors iš grafiko matyti, kad τ yra ir mažesnių reikšmių, esant didesniems α, bet tikslinga fiksuoti rastą pirmą lokalų τ minimumą, nes $\alpha > \alpha_c$ yra jau per didelis, tai rodo 6 ir 7 paveikslai.



7 pav. Irisų duomenys: τ priklausomybė nuo α

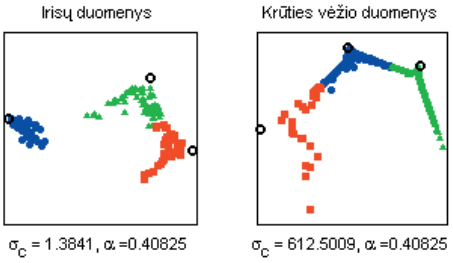
Aukščiau aprašyti skaičiavimai reikalauja nemažai sąnaudų. Mat, norint rasti minimalų τ, tenka atlikti daug skaičiavimų, kurie savo viduje turi ir MDS metodą. Norėdami sutaupyti laiko skaičiavimams, pabandėme paimti fiksuotą α iš (4) formulės, kur α priklauso nuo klasterių skaičiaus, o pločio parametrą skaičiuoti pagal (7) formulę:

$$\alpha = \frac{1}{\sqrt{2k}},\ \sigma_C = \alpha d_{vid} = \frac{d_{vid}}{\sqrt{2k}}, \qquad (9)$$

čia d_{vid} – vidutinis atstumas tarp klasterių centrų, k – klasterių skaičius, $\alpha \in (0,05]$, kai $k \geq 2$.

Daugiamačių duomenų transformacijos rezultatai, kai σ_c apskaičiuojama pagal (9) formulę, pateikti 8 paveiksle. Palyginus 5, 6 ir 8 paveiksluose pateiktus Irisų duomenų transformavimo rezultatus, pastebima, kad pločio parametras σ_c yra šiek tiek per mažas, nes dar vyksta objektų trauka link centrų.

Krūties vėžio duomenims σ_c yra truputį per didelis, nes stebimas pirmojo (šio klasterio objektai pažymėti ■) ir antrojo (šio klasterio objektai pažymėti •) klasterių objektų judėjimas link trečiojo (jo objektai pažymėti ▲) klasterio centro. Matyti, kad rezultatai pagal (9) formulę nėra patys geriausi, tačiau šiuo atveju vaizdas geresnis nei 5 pav.



8 pav. Transformavimo rezultatai, gauti skaičiuojant σ_c pagal (9) formulę

Išvados

1. Didelės apimties ir daug požymių turinčius daugiamačius duomenis žmogui suvokti sunku. Vienas iš būdų, palengvinančių duomenų suvokimą, yra daugiamačių duomenų vizualizavimas, kurio metu daugiamačiai duomenys transformuojami į mažesnio matavimo erdvę. Transformacijos rezultatas – atsiradę nauji požymiai, kurių skaičius yra mažesnis nei pradinių požymių skaičius.
2. Daugiamačių duomenų požymių skaičių galima sumažinti naudojantis eksponentine koreliacine funkcija, kai atsižvelgiama, jog daugiamačiuose duomenyse yra panašių duomenų klasteriai.
3. Aprašytasis daugiamačių duomenų požymių skaičiaus mažinimo eksponentine koreliacine funkcija metodas apima:
   - daugiamačių duomenų klasterizavimą į tam tikrą klasterių skaičių k,
   - duomenų transformavimą į k-matę erdvę R^k,
   - jau transformuotų k-mačių duomenų vizualizavimą daugiamačių skalių metodu.
4. Ištyrus metodą eksperimentiškai, nustatyta:
   1) vizualizavimo rezultatai po daugiamačių duomenų transformavimo į mažesnio matavimo erdvę labai priklauso nuo:
      a) faktinio klasterių skaičiaus duomenyse;
      b) pasirinkto pločio parametro σ. (Aptarti trys pločio parametro σ pasirinkimo būdai. Tiksliausias σ apskaičiavimas yra pagal σ_B formulę, bet jis reikalauja daugiausia resursų ir laiko sąnaudų).
   2) objektai klasteryje išrūšiuojami pagal panašumą į kaimyninių klasterių objektus ir būdingumą tik konkrečiam klasteriui.
5. Viena iš tolimesnio tyrimo pusių – eksponentinės koreliacinės funkcijos taikymo lyginimas su rezultatais, gaunamais taikant įvairias radialines bazines funkcijas.

Literatūra

1. Benoudjit N., Verleysen M., 2003, On the Kernel Widths in Radial-Basis Function Networks. *Neural Processing Letters*. Nr. 18. P. 139–154.
2. Borg I., Groenen P., 2005, *Modern Multidimensional Scaling*. 2nd ed. New York: Springer.
3. Buhmann M. D., 2003, *Radial Basis Functions: Theory and Implementations*. United Kingdom: Cambridge University Press.
4. Chang Q., Chen Q., Wang X., 2005, *Scaling Gaussian RBF kernel width to improve SVM classification. International Conference on Neural Networks and Brain* (ICNN&B `05). Oct 13–15. P. 19–22. Beijing, China.
5. Dzemyda G., Kurasova O., Žilinskas J., 2013, Multidimensional Data Visualization: Methods and Applications. *Springer Optimization and Its Applications*. Vol. 75. New York: Springer.
6. Haykin, S. 1999. Neural Networks and Learning Machines, 2rd ed. New York: Prentice Hall.
7. Han J., Kamber M., Pei J., 2011, Data mining: concepts and techniques. *Morgan-Kaufman Series of Data Management Systems*. Amsterdam: Elsevier.
8. Yaglom A. M., 1986, Correlation Theory of Stationary and Related Random Functions I: Basic Results. *Springer Series in Statistics*. New York: Springer.
9. Pierrefeu L., Jay J., Barat C., 2006, Auto-adjustable method for Gaussian width optimization on RBF neural network. Application to face authentication on a mono-chip system. *The 32nd Annual Conference of the IEEE Industrial Electronics Society* (IECON 2006). November 7–10. P. 3481–3485. Paris.
10. Stein M. L., 1999, Interpolation of Spatial Data: Some Theory for Kriging. *Springer Series i Statistics*. New York: Springer.
11. Tiago C. M., Leitão V. M. A., 2006, Application of Radial Basis Functions to Linear and Nonlinear Structural Analysis Problems. *Computers and Mathematics with Appli*cations. Vol. 51. Nr. 8. P. 1311–1334.

Priedas

Eksperimentuose naudotos 4 daugiamačių duomenų aibės. Trijų daugiamačių duomenų aibių duomenys paimti iš duomenų bazės „UCI Repository of Machine Learning Databases“ (http://archive.ics.uci.edu/ml/):

1. Gėlių irisų duomenų aibė (angl. *Iris Plants Database*). Duomenų rinkinį sudaro trijų rūšių irisai – Setosa, Versicolour ir Virginica (k = 3). Kiekvienos rūšies yra po 50 gėlių, iš viso 150 (m = 150). Kiekvieną irisą apibūdina keturi požymiai – taurėlapio ilgis, taurėlapio plotis, vainiklapio ilgis ir vainiklapio plotis (n = 4).

2. Atsitiktinai generuotų duomenų aibė (angl. *Random Generated Database*). Duomenys generuoti taip, kad sudarytų penkis klasterius (k = 5) 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 $x_j \in [3,5]$.
3. Stuburo ligų duomenų aibė (angl. *Vertebral Column Database*). Duomenų rinkinį galima klasifikuoti į 3 klasterius (k = 3) – sveiki, stuburo disko išvarža, spondilolistezė – arba į 2 klasterius (k = 2) – sveiki, sergantys. Visą duomenų rinkinį sudaro 310 pacientų (m = 310). Kiekvieną pacientą apibūdina šeši biomechaniniai požymiai: dubens dažnis, dubens tentas, juosmens kampas, sakraliniai nuolydžiai, dubens spindulys ir spondilolistezės klasė (n = 6).
4. Krūties vėžio duomenų aibė (angl. *Breast Cancer Database*). Duomenų rinkinys klasifikuojamas į 2 klasterius (k = 2) – piktybinis navikas ir gerybinis navikas. Visą duomenų rinkinį sudaro 569 navikai (m = 569). Kiekvieną naviką apibūdina 30 požymių: įvairūs naviko matavimai (spindulys, perimetras, plotis, kompaktiškumas ir kt.), vidurkis, standartinė paklaida (n = 30).

APPLICATION OF EXPONENTIAL CORRELATION FUNCTION FOR REDUCTION OF ATTRIBUTES OF MULTIDIMENSIONAL DATA

Laura Ringienė, Gintautas Dzemyda

Summary

Multidimensional data of large quantity and with many attributes are often difficult to understand for a human. Multidimensional data visualization where multidimensional data are transformed into lower dimension space is one of the ways to facilitate data comprehension. The transformation results in appearance of new attributes the number of which is lower than that of initial attributes. In this paper we discuss the way to reduce the number of attributes of multidimensional data by using exponential correlation functions by taking into account that multidimensional data have clusters of similar data.

Keywords: exponential correlation function, data clustering, multidimensional scales, visualization.

DAUGIAMAČIŲ DUOMENŲ POŽYMIŲ MAŽINIMAS NAUDOJANTIS EKSPONENTINE KORELIACINE FUNKCIJA

Laura Ringienė, Gintautas Dzemyda

Santrauka

Didelės apimties ir daug požymių turinčius daugiamačius duomenis žmogui sunku suvokti. Vienas iš būdų, palengvinančių daugiamačių duomenų suvokimą, yra jų vizualizavimas, kurio metu atliekama daugiamačių duomenų transformacija į mažesnio matavimo erdvę. Transformacijos rezultatas – nauji požymiai, kurių skaičius mažesnis nei pradinių požymių skaičius. Straipsnyje aptariama, kaip galima sumažinti daugiamačių duomenų požymių skaičių eksponentinėmis koreliacinėmis funkcijomis, kai atsižvelgiama, jog daugiamačiuose duomenyse yra panašių duomenų klasteriai.

Prasminiai žodžiai: eksponentinė koreliacinė funkcija, duomenų klasterizavimas, daugiamatės skalės, vizualizavimas.

Įteikta 2013-04-05