Liet. matem. rink, **46***, spec. nr.*, 2006, 273–279

Neparametrinio tankio vertinimo panaudojant klasterizavimo metodus algoritmų tyrimas

Rasa ŠMIDTAITĖ, Tomas RUZGAS (KTU, MII)
el. paštas: rasuzele1984@takas.lt, tomas.ruzgas@ktu.lt

1. Įvadas

Statistinių metodų taikyme dažnai susiduriama su imties pasiskirstymo tankio vertinimo uždaviniu. Jei pasiskirstymo tankių šeima yra žinoma, tuomet uždavinio sprendimas susiveda į kelių nežinomų parametrų radimą. Tačiau ne visada būna žinoma parametrinė skirstinio forma. Tuo atveju naudojami neparametriniai tankių vertinimo metodai. Dažniausiai šie metodai reikalauja didelių skaičiavimų, dėl ko jų taikymas nėra labai patrauklus.

Straipsnyje [14] buvo apžvelgti populiarūs ir dažnai praktikoje sutinkami neparametriniai tankių vertinimo algoritmai. Pateikiamas darbas yra minėto straipsnio tęsinys – jame neparametriniai tankių vertinimo metodai skirstinių mišiniams taikomi kartu su klasterizavimo procedūromis, kurių pagalba atliekamas daugiamodalinio tankio analizės suvedimas į vienamodalinių tankių vertinimą. Įvertinus tankį ir taikant Bajeso metodologiją atliekamas perklasterizavimas, kuris leidžia patikslinus klasterius atlikti pakartotinį tankių vertinimą.

Šis straipsnis sudarytas taip: 2 dalyje aprašytos imties klasterizavimo procedūros; 3 dalyje pateiktas vienas populiariausių klasterių skaičiaus nustatymo algoritmų, skirtų geometriniam klasterizavimui; 4 dalyje trumpai apžvelgiami pasiskirstymo tankių įvertinių algoritmai; 5 dalis talpina skaitinio modeliavimo rezultatus; 6 dalyje pateiktos išvados.

2. Tirtos klasterizavimo procedūros

Klasterių formavimo metodų yra daug. Jie skirstomi pagal tai, kaip parenkami panašumo matai, atstumo tarp klasterių nustatymo kriterijai bei kokia skirstymo į klasterius strategija. Pagal skirstymo į klasterius strategiją išskiriamos dvi pagrindinės klasterinės analizės geometrinių metodų klasės – hierarchiniai ir nehierarchiniai metodai. Šiame darbe buvo nagrinėtos šios klasterizavimo procedūros:

1) hierarchinis jungimo algoritmas su įvairiai apibrėžtu atstumo matu tarp klasterių objektų;
2) k-vidurkių algoritmas;
3) k-artimiausių kaimynų algoritmas.

Hierarchinis jungimo algoritmas. Kai iš pradžių kiekvienas stebėjimas sudaro atskirą klasterį, o po to juos sujungiant galutiniame etape visi jie sudaro vieną klasterį, tai vadinama jungiančiąja hierarchija. Pažymėkime $S_i^{(k)}$ i-tąjį k-lygio klasterį, n_i – stebėjimų skaičių klasteryje; $\rho(S_l^{(k)}, S_m^{(k)})$ – atstumą tarp klasterių $S_l^{(k)}$ ir $S_m^{(k)}$. Jungiančiosios hierarchijos algoritmo pradinis skaidinys yra $S^{(0)} = (S_1^{(0)}, ..., S_n^{(0)})$, čia $S_i^0 = \{X_i\}$, k-lygio skaidinys $S^{(k)} = (S_1^{(k)}, ..., S_{n-k}^{(k)})$ gaunamas iš $S^{(k-1)}$ skaidinio apjungus klasterių porą (S_1^*, S_2^*):

$$(S_1^*, S_2^*) = \mathop{\arg\min}_{\substack{S_1 \neq S_2 \\ S_1, S_2 \in S^{(k-1)}}} \rho(S_1, S_2). \tag{1}$$

Galutinę hierarchiją sudaro įdėtų skaidinių sistema $S^{(0)} \subset S^{(1)} \subset ... \subset S^{(n-1)} \equiv X$, kurią galima vaizduoti grafiškai medžio formos diagrama, vadinama dendrograma. Literatūroje, pavyzdžiui [3], galima rasti daugybę įvairių hierarchijos formavimo būdų. Tyrime apsiribosime tolimiausio kaimyno, centroidų ir Ward metodais.

k-vidurkių algoritmas. Šį algoritmą sudaro pradinių klasterių radimo metodas ir iteracinis algoritmas, kuris minimizuoja nuokrypių kvadratų sumą tarp klasterių vidurkių. Užduodami pradiniai taškai, kurie laikomi klasterių vidurkiais. Visi stebėjimai priskiriami laikiniems klasteriams pagal mažiausią atstumą iki užduotų klasterių vidurkių. Užduotų klasterių vidurkiai keičiami laikinų klasterių vidurkiais ir procesas kartojamas, kol klasteriai stabilizuojasi. Klasterizavimas yra paremtas Euklido atstumu ir stebėjimai, esantys arti vienas kito, priskiriami tam pačiam klasteriui, o stebėjimai, nutolę vienas nuo kito, – skirtingiems klasteriams [11].

k-artimiausių kaimynų algoritmas. Klasterizavimas pradedamas nuo to, jog kiekvienas stebėjimas priklauso atskiram klasteriui. Toliau jungiami du klasteriai į vieną:

- sudaromos visos įmanomos poros iš dviejų elementų;
- skaičiuojamas tankis kiekvienai porai:

$$f_i(x) = \frac{n_i(x)}{n \cdot V(x)}, \tag{2}$$

  čia $n_i(x)$ – i-tojo klasterio kaimynų (Euklido atstumo prasme artimiausių stebėjimų) skaičius kartu priskaičiuojant ir patį stebėjimą x, n – imties didumas, $V(x)$ – $n_i(x)$ sudarančių stebėjimų užimamos erdvės hipertūris;
- sujungiami du klasteriai, kurių bendras tankis didžiausias.

Toliau nagrinėjamas kiekvienas stebėjimas kartu su atitinkamu vieno ar kelių kaimynų pagalba įvertintu tankiu (2). Tariama, kad tiriamas stebėjimas gali priklausyti bet kuriam klasteriui ir tuo būdu randami jo k artimiausi kaimynai. Tiriamas stebėjimas priskiriamas tam klasteriui, su kuriuo jį apjungus tankis (2) yra didžiausias ir ne mažesnis nei apjungus su bet kuriuo kaimynu. Toks klasterių tikslinimas baigiamas, kai klasteriai nusistovi ir jų struktūra nebesikeičia [6].

3. Klasterių skaičiaus nustatymas

Viena iš problemų, su kuria dažnai susiduriama klasterinėje analizėje, – tai klasterių skaičiaus nustatymas. Iš populiariausių kriterijų labiausiai informatyviu laikomas Šarlio kubinis klasterizavimo kriterijus CCC [15]. Taikant šį kriterijų tikrinamos tokios hipotezės:

H_0: stebėjimų skirstinys daugiamatis tolygusis.

H_1: stebėjimų skirstinys daugiamatis Gauso.

Teigiamų CCC reikšmių atveju H_0 atmetama.

$$\mathrm{CCC} = \ln\left[\frac{1 - E(R^2)}{1 - R^2}\right]\frac{\sqrt{nd^*/2}}{(0.001 + E(R^2))^{1.2}}, \tag{3}$$

čia $R^2 = 1 - [d^* + \sum_{j=d^*+1}^{d} u_j^2]/\sum_{j=1}^{d} u_j^2$, n – stebėjimų skaičius, r $u_j = s_j/c$, kai $c = (v/q)^{\frac{1}{d}}$ ir $v = \prod_{i=1}^{d} s_i$, q – klasterių skaičius, s_j – hiperkubo kraštinės ilgis j-tos projekcijos kryptimi, d^* – didžiausias sveikasis skaičius mažesnis už q, bet toks, kad u_{d*} būtų nemažesnis už vienetą,

$$E(R^2) \cong 1 - \left[\sum_{j=1}^{d^*}\frac{1}{n + u_j} + \sum_{j=d^*+1}^{d}\frac{u_j^2}{n + u_j}\right]\left[\frac{(n-q)^2}{n}\right]\left[1 + \frac{4}{n}\right]\Big/\sum_{j=1}^{d} u_j^2.$$

Jei CCC reikšmės yra tarp 0 ir 2, tai tokia klasterinė struktūra yra galima. Jei reikšmės yra didesnės už 2, tai klasterių skaičius parinktas tinkamai.

4. Nagrinėti tankių vertinimo metodai

Šiame darbe buvo nagrinėti šie pasiskirstymo tankių statistiniai įvertiniai:

1) tikslinio projektavimo tankio įvertinys (PPDE);
2) adaptuotas branduolinis tankio įvertinys (AKDE);
3) pusiau parametrinis branduolinis pasiskirstymo tankio įvertinys (SKDE);
4) histosplaininis tankio įvertinys (HSDE).

PPDE, AKDE ir SKDE metodai nagrinėti [14] straipsnyje. Platesnis jų aprašymas yra [5, 8, 9] darbuose. Plačiau aprašysime HSDE metodą [1].

Pasiskirstymo tankio vertinimas histosplainu susideda iš dviejų etapų. Pirmame etape sudaroma d-matė histograma. Antrame etape, jau turint histograminį tankio įvertį, ieškoma daugiamačio splaino regresinės priklausomybės, leidžiančios patikslinti pirmame etape gautą tankio įvertį.

Stebėjimų x projekcijų į ašis $x^{(j)}$, $j = 1, \ldots, d$ kitimo intervalai padalinami į l dalinių intervalų ir jais apribotuose hiperkubuose randamas tankio įvertis:

$$f(c_k) = \frac{n(c_k)}{n \cdot h_1 \cdot h_2 \ldots \cdot h_d}, \tag{4}$$

čia c_k yra k-tasis hiperkubas, $n(c_k)$ yra į hiperkubą c_k patenkančių stebėjimų skaičius, o h_j, $j = 1, \ldots, d$ žymi hiperkubo kraštines. Hiperkubų skaičių rekomenduojama parinkti $r = 1 + 3.32\ln(n)$, kadangi $l = \sqrt[d]{r}$ turi būti sveikasis skaičius, tai r parenkamas $r = \lceil \sqrt[d]{1 + 3.32\ln(n)}\rceil^d$.

Hiperkubų centruose apskaičiuotus įverčius aproksimuojant d-mačiu splainu, pirmiausia d-matis baigtinis atsitiktinis histogramos vidurio taškų tinklelis $\mathbf{x}$ padalinamas atitinkamai į s ir $(d-s)$-mačius subtinklelius, **x**=(**y**,**z**), taip, kad **z** apibūdina regresijos tiesinę dalį, t.y., priklausomybė tarp **z** ir tankio įverčių histogramos hiperkubų centruose **w** yra tiesinė, o **y** apibūdina regresijos netiesinę dalį. (**y**,**z**) suskirstymas parenkamas naudojant Fišerio kriterijų hipotezės apie modelio netiesiškumą tikrinimui. Tada regresijos lygtis atrodys taip:

$$w_k = g(y_k) + z_k\beta + e_k, \tag{5}$$

čia g yra nežinoma tolydi funkcija, β – nežinomas $(d-s)$-matis parametrų vektorius, o $e_k, k = 1,\dots,r$ yra nepriklausomos atsitiktinės paklaidos, kurių vidurkis 0.

Aproksimavimo d-mačiu splainu procedūra susideda iš dviejų etapų. $z_k\beta$ yra tiesinė parametrinė modelio dalis, o z_k – tos regresijos nepriklausomi kintamieji. $g(y_k)$ yra neparametrinė modelio dalis.

Netiesinė regresijos dalies funkcija $g(y_k)$ apibrėžiama [2]:

$$g(y_k) = \theta_0 + \sum_{i=1}^{s}\theta_i y_{ki} + \sum_{j=1}^{r}\delta_j E_2(y_k - y_j), \tag{6}$$

čia $E_2(y_k - y_j) = \frac{1}{2^3\pi}\|y_k - y_j\|^2 \ln(\|y_k - y_j\|)$.

Taigi, regresijos lygties įvertinimas suvedamas į parametrų (β,δ,θ) radimą. Pažymėjus $\mathbf{K}=(K)_{kj} = E_2(y_k - y_j)$ ir $\mathbf{T}=(T)_{kj} = (y_{kj})$, koeficientai (β,δ,θ) randami mažiausių kvadratų metodu minimizuojant funkciją $S(\beta,\delta,\theta)$:

$$S(\beta,\delta,\theta) = \tfrac{1}{r}\,\|\mathbf{y} - \mathbf{T}\theta - \mathbf{K}\delta - \mathbf{z}\beta\|^2. \tag{7}$$

5. Eksperimentinis tyrimas

Aukščiau aprašytų neparametrinių tankių vertinimo algoritmų tikslumo tyrimas atliktas Monte Karlo metodu. Toks algoritmų palyginimo būdas sudarė galimybes išmatuoti tikrąsias tankių reikšmes kiekviename stebimame taške ir taip įvertinti algoritmų tikslumą. Tyrimą sudaro trys dalys:

a) tankių vertinimas, kai duomenys nėra klasterizuojami;
b) atliekamas pradinis duomenų suskaidymas į klasterius ir tuomet kiekviename klasteryje atskirai įvertinamas tankis;
c) antroje dalyje gauti tankio įverčiai panaudojami duomenų perklasterizavimui taikant Bajeso principą ir aposteriorinių tikimybių $\pi_i(x) = \mathbf{P}\{v = i | X = x\}$ įverčius $\hat{\pi}_i(x)$, čia $\hat{v}(t) = \arg\max_{k=1,\dots,q}\hat{\pi}_k(X(t))$ interpretuojamas kaip klasės, kuriai priklauso stebimas objektas, numeris. Duomenys perklasterizuojami tam, kad būtų patikslintas pradinis suskaidymas ir dar kartą, iš naujo, atliekamas tankio vertinimas. Ši klasterizavimo procedūra rekurentiškai atliekama keletą kartų.

Naudojami trijų tipų (vienos modos, dviejų modų mažai persidengiantis ir dviejų modų stipriai persidengiantis) daugiamačiai ($d = 2$, $d = 5$) Gauso ir Koši skirstinių su nepriklausomomis komponentėmis mišiniai. Duomenų skirstinių tankių mišiniai:

Gauso mišinys: $\sum_{i=1}^{q} p_i f_N(x, m_i, \sigma_i)$, Koši mišinys: $\sum_{i=1}^{q} p_i f_C(x, m_i, u_i)$

su apribojimais

$$\sum_{i=1}^{q} p_i = 1, \quad p_i > 0, \quad i = 1, \ldots, q,$$

$$f_N(x, m_i, \sigma_i^2) = \frac{1}{\prod_{j=1}^{d} \sqrt{2\pi}\sigma_{ij}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{d} \frac{1}{\sigma_{ij}^2}(x_j - m_{ij})^2 \right\},$$

$$f_C(x, m_i, u_i) = \prod_{j=1}^{d} \frac{u_{ij}}{\pi[u_{ij}^2 + (x_j - m_{ij})^2]}.$$

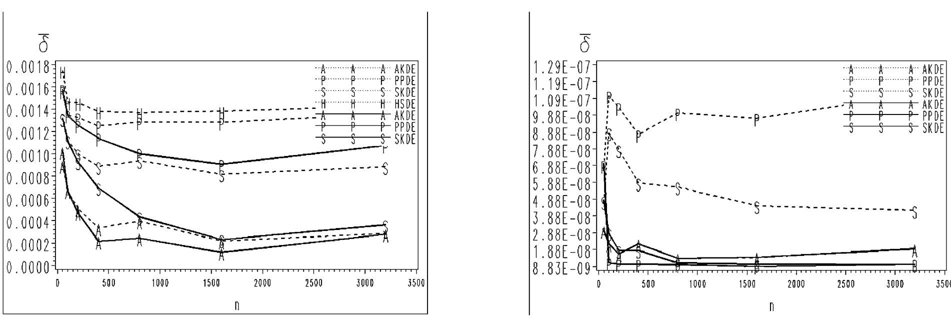
Skirstinių parametrų reikšmės parinktos tokios pat kaip ir J.N. Hwang, S.R. Lay ir A. Lippman [9] darbe. Kiekvieno tipo duomenims (tiek Gauso, tiek Koši mišiniams) skirtingiems matavimams ($d = 2, d = 5$) generuotos įvairaus dydžio imtys (50, 100, 200, 400, 800, 1600, 3200).

Tankių vertinimo tikslumui išreikšti skaičiuojama vidutinė kvadratinė paklaida:

$$\delta = \frac{1}{n} \sum_{t=1}^{n} \left(f(X(t)) - \hat{f}(X(t)) \right)^2.$$

Kiekvienu atveju apskaičiuoti paklaidų δ aritmetiniai vidurkiai $\bar{\delta}$ gauti sugeneravus 100 nepriklausomų imčių.

Tyrimo rezultatai. Atlikto kompiuterinio eksperimento rezultatai pilnai patvirtino darbo [14] išvadą apie klasterizavimo tikslingumą. Paklaidos δ priklausomybė nuo imties dydžio ir atstumo tarp vertinamo tankio viršūnių buvo panaši (kokybiniu požiūriu). Duomenų klasterizavimas įvairiais metodais, bendru atveju, labiausiai rezultatus pagerino dažniausiai mažoms imtims, kadangi mažo didumo imtys dažniau buvo



1 pav. Paklaidų priklausomybė nuo imties dydžio (Koši skirstinys, k-artimiausių kaimynų klasterizavimo procedūra).

skirstomos į didesnį klasterių skaičių nei didesnio didumo imtys. Koši tipo skirtiniams rezultatai pagerėjo ženkliau lyginant su Gauso tipo skirstiniais. Tankių įverčių tikslumui buvo stebėta didžiausia k-artimiausių kaimynų klasterizavimo procedūros įtaka.

Paklaidos priklausomybę nuo imties dydžio iliustruoja 1 pav. Jame vaizduojami Koši vienamodalinių tankių tyrimo rezultatai, čia punktyrine linija žymima neklasterizuotų duomenų pasiskirstymo tankio įverčių paklaidos, o ištisine – k-artimiausių kaimynų procedūra atlikus pradinį duomenų klasterizavimą. Tiriant kitus modelius nustatytos tos pačios tendencijos.

6. Išvados

1. Koši dvimačius mišinius geriausiai vertina adaptuotas branduolinis ir pusiau parametrinis branduolinis metodai, o penkiamačius – tikslinio projektavimo metodas.
2. Gauso dvimačius mišinius geriausiai vertina adaptuotas branduolinis metodas, o penkiamačius – pusiau parametrinis branduolinis ir tikslinio projektavimo metodai.
3. Histosplaininis tankių vertinimo metodas yra konkurencingas su kitais esant mažo matavimo duomenims. Didesnio matavimo duomenims šis metodas duoda blogesnius rezultatus lyginant su kitais metodais.
4. Klasterizavimas tirtais metodais bei perklasterizavimas pagerino tankių vertinimo rezultatus tik Koši mišinių atveju.

Literatūra

1. P. Delicado, M. del Rio, A generalization of histogram type estimators, *Journal of Nonparametric Statistics*, **15**(1), 113–135 (2003).
2. J. Duchon, Fonctions-spline et esperances conditionnelles de champs Gaussiens, *Ann. Sci. Univ. Clermont Ferrand II Math.*, **14**, 19–27 (1976).
3. B. Everitt, S. Landau, M. Leese, *Cluster Analysis*, Oxford University Press, NY (2001).
4. L.L. Fleiss, J. Zubin, On the methods and theory of clustering, *Multivariate Behavioral Research*, **4**, 235–250 (1969).
5. J.H. Friedman, Exploratory projection pursuit, *Journal of the American Statistical Association*, **82**(397), 249–266 (1987).
6. I. Gitman, An algorithm for nonsupervised pattern classification, *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-**3**, 66–74 (1973).
7. W. Härdle, M. Müller, *Multivariate and Semiparametric Kernel Regression*, Wiley Publishers, 357–391 (2000).
8. L. Holmström, F. Hoti, Application of semiparametric density estimation to classification, *ICPR*, **3**, 371–374 (2004).
9. J.N. Hwang, S.R. Lay, A. Lippman, Nonparametric multivariate density estimation: a comparative study, *IEEE Transactions on Signal Processing*, **42**(10), 2795–2810 (1994).
10. D. Yeo, Applied clustering techniques, in: *Course Notes*, SAS Institute Inc., NC (2003), p. 341.
11. J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1 (1967), pp. 281–297.
12. R. Rudzkis, M. Radavicius, Statistical estimations of a mixture of gaussian distributions, *Acta Applicandae Mathematicae*, **38**, 37–54 (1995).
13. T. Ruzgas, Įvairių klasterizavimo algoritmų efektyvumo palyginimas, *Liet. matem. rink.*, **42**(spec. nr.), 571–576 (2002).
14. T. Ruzgas, M. Kavaliauskas, Daugiamačių Gauso skirstinių mišinio modelio panaudojimas neparametrinių tankių vertinime, *Liet. matem. rink.*, **45**(spec. nr.), 369–374 (2005).

15. W.S. Sarle, The Cubic Clustering Criterion, *SAS Technical Report A-108*, SAS Institute, Cary, NC (1983).

SUMMARY

R. Šmidtaitė, T. Ruzgas. Research of nonparametric density estimation algorithms by applying clustering methods

One of the ways to improve the accuracy of probability density estimation is multi-mode density treating as the mixture of single-mode one. In this paper we offer to use data clustering in the first place and to estimate density in every cluster separately. To objectively compare the performance, Monte Carlo approximation is used. While using various methods to evaluate the accuracy of probability density estimations we tried to use clustered and not clustered data. In this paper we also tried to reveal the usefulness of using clustering for data generated by single-mode and multi-mode distributions.

Keywords: nonparametric density estimation, sample clustering, Monte-Carlo method.