Liet. matem. rink., T. 42, spec. nr., 571–576

Įvairių klasterizavimo algoritmų efektyvumo palyginimas

Tomas RUZGAS (MII, KTU)
el. paštas: tomas.ruzgas@fmf.ktu.lt

1. Įvadas

Pastaruoju metu vykstant sparčiam skaičiavimo technikos ir programinės įrangos vystymuisi galima apdoroti didelius duomenų masyvus, tai skatina sudėtingesnių matematikos metodų naudojimą. Kadangi duomenys dažnai yra daugiamačiai ir įvairialypiai, tai prieš atliekant jų analizę dažnai neaišku, kiek reikšmingas vienas ar kitas rodiklis konkretaus uždavinio sprendimui. Tokiu atveju vienas iš sprendimo būdų yra daugiamačių duomenų klasifikavimas į atskiras, homogenines grupes.

Šiame darbe nagrinėjamas Gauso skirstinių mišinio klasifikavimo efektyvumo uždavinys naudojant įvairius klasterizavimo metodus.

Tarkime, turime q nepriklausomų d-mačių atsitiktinių dydžių Y_i, kurių skirstinio tankiai φ_i su vidurkiais M_i ir kovariacinėmis matricomis R_i, $i = 1, 2, \ldots, q$. Tegul v yra atsitiktinis dydis, nepriklausomas nuo Y_i, ir įgyjantis reikšmes $i = 1, 2, \ldots, q$ su nežinomomis tikimybėmis $p_i > 0$, $i = 1, 2, \ldots, q$. Pažymėkime d-matį atsitiktinį dydį $X = Y_v$. Kiekvienas stebėjimas priklauso vienai iš q klasių, priklausančių nuo atsitiktinio dydžio v. Atsitiktinio dydžio X skirstinio tankis yra mišinio tankis

$$f(x) = \sum_{i=1}^{q} p_i \varphi_i(x) = f(x, \theta), \quad x \in \mathbf{R}^d, \tag{1}$$

čia $\theta = (p_i, M_i, R_i, i = 1, 2, \ldots, q)$ yra nežinomas daugiamatis parametras.

Pagrindinis klasifikavimo uždavinio tikslas yra stebėjimo rezultatų $\mathbf{X} = \{X_1, \ldots, X_n\}$ pagrindu nustatyti objektų su požymių vektoriumi X priklausomybės i-tajai klasei $i = 1, 2, \ldots, q$ aposteorines tikimybes $\pi_i(x) = P\{v = i | X = x\}$. Pasinaudoję įvestais pažymėjimais, galime užrašyti

$$\pi_i(x) = \frac{p_i \varphi_i(x)}{f(x)}, \quad i = 1, 2, \ldots, q. \tag{2}$$

Remiantis (2) imties $\mathbf{X}$ reikšmės priskiriamos grupėms

$$\hat{v}(X) = \arg\max_i \hat{\pi}_i(X), \quad i = 1, 2, \ldots, q. \tag{3}$$

2. Naudojami klasterizavimo algoritmai

Darbe naudojamus klasterizavimo algoritmus sąlyginai galima suskirstyti į „geometrinius“ ir „tikimybinius“. Pirmajai grupei yra priskiriamas hipersferinis, antrajai grupei – EM, k-vidurkių ir vienodų kovariacinių matricų algoritmai, jie dažniausiai taikomi naudojant Gauso skirstinių mišinio modelius.

EM algoritmas. Jeigu klasių skaičius q yra žinomas, tai maksimalaus tikėtinumo įvertis $\hat{\theta}_{MTM}$ yra efektyvus θ įvertis. Praktikoje maksimalaus tikėtinumo įverčio radimui dažniausiai taikomas EM algoritmas. Tegul $\pi_i(x)$, $i = 1, 2, \ldots, q$ yra duota apriorinė tikimybė imties $\mathbf{X}$ stebėjimams. Duotai $\pi_i(x)$ įvertinamas parametras $\theta = (p_i, M_i, R_i, i = 1, 2, \ldots, q)$ [5, 6, 7, 11]. Duotai pradinei $\theta^{(0)}$ yra skaičiuojamos $\hat{\pi}_i^{(0)}$ tikimybės. EM algoritmas yra rekurentinė procedūra, kuri pradeda skaičiuoti arba nuo duoto parametro θ, arba nuo duotos tikimybės $\pi_i(x)$ pradinių įverčių. EM algoritmas paprastai nutraukiamas po tam tikro, iš anksto užduotų, iteracijų skaičiaus. Įvertis $\hat{\theta}$ EM algoritme konverguoja į maksimalaus tikėtinumo įvertį $\hat{\theta}_{MTM}$, jeigu pradinis įvertis $\theta^{(0)}$ yra pakankamai arti $\hat{\theta}_{MTM}$ reikšmės.

k-vidurkių algoritmas. Šis algoritmas jungia pradinių klasterių radimo metodą ir iteracinį algoritmą, 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 [14]. Klasterizavimas yra paremtas Euklidiniu atstumu, ir stebėjimai esantys arti vienas kito priskiriami tam pačiam klasteriui, o stebėjimai nutolę vienas nuo kito – skirtingiems klasteriams.

Sugrupavus imties $\mathbf{X}$ stebėjimus į klasterius kiekviename klasteryje įvertinamas parametras $\theta = (p_i, M_i, R_i, i = 1, 2, \ldots, q)$. Įverčio $\hat{\theta}$ reikšmės $\hat{p}_i$, $\widehat{M_i}$, $i = 1, 2, \ldots, q$ k-vidurkių algoritme konverguoja į mažiausių kvadratų įverčio $\hat{\theta}_{MKM}$ reikšmes $\hat{p}_{iMKM}$, $\widehat{M}_{iMKM}$, $i = 1, 2, \ldots, q$, jeigu pradinio įveršio $\theta^{(0)}$ reikšmės $\hat{p}_i^{(0)}$, $\widehat{M}_i^{(0)}$, $i = 1, 2, \ldots, q$ yra pakankamai arti $\hat{\theta}_{MKM}$ reikšmių $\hat{p}_{iMKM}$, $\widehat{M}_{iMKM}$, $i = 1, 2, \ldots, q$.

Vienodų kovariacinių matricų algoritmas. Tegu $\mathbf{A} = (a_{jk})$ – kovariacinė matrica vienoda visuose klasteriuose, n_i – stebėjimų skaičius i-tame klasteryje ir

$$d'(t,h) = \\ = \begin{cases} \dfrac{1}{n_i}, & \text{jei } \sum_{j=1}^{d}\sum_{k=1}^{d} m^{(j)(k)}\big(X^{(j)}(t)-X^{(j)}(h)\big)\big(X^{(k)}(t)-X^{(k)}(h)\big) \leqslant u^2, \\ 0, & \text{jei } \sum_{j=1}^{d}\sum_{k=1}^{d} m^{(j)(k)}\big(X^{(j)}(t)-X^{(j)}(h)\big)\big(X^{(k)}(t)-X^{(k)}(h)\big) > u^2. \end{cases} \tag{4}$$

Matricos $\mathbf{A}$ elementai apibrėžiami kaip

$$a_{jk} = \frac{\sum_{t=1}^{n}\sum_{h=1}^{t-1} d'(t,h)\big(X^{(j)}(t)-X^{(j)}(h)\big)\big(X^{(k)}(t)-X^{(k)}(h)\big)}{2\sum_{t=1}^{n}\sum_{h=1}^{t-1} d'(t,h)}. \tag{5}$$

Vienodų kovariacinių matricų algoritmas yra rekurentinė procedūra, kuri pradeda skaičiuoti nuo duotos kovariacinės matricos **A** pradinio įverčio (atskiru atveju tai gali būti imties **X** kovariacinė matrica). Laikoma, kad matrica $\mathbf{M} = (m_{jk})$ lygi $\mathbf{A}^{-1}$. Pagal (4) ir (5) perskaičiuojama matrica **A**. Vienodų kovariacinių matricų algoritmas nutraukiamas kai įvertis stabilizuojasi [14]. Sugrupavus imties **X** stebėjimus į klasterius kiekviename klasteryje įvertinamas parametras $\theta = (p_i, M_i, R_i, i = 1, 2, \ldots, q)$.

Hipersferinis algoritmas. Apie kiekvieną stebėjimą formuojama r spindulio hipersfera $r = \left[\frac{2^{d+2}(d+2)\Gamma(\frac{d}{2}+1)}{nd^2}\right]^{1/(d+4)} \sqrt{\sum_{l=1}^{d} (s^{(l)})^2}$, $(s^{(l)})^2$ – mišinio empirinės dispersijos $l = 1, 2, \ldots, d$, ir randami artimiausi jo „kaimynai". Dvi šalimais esančios hipersferos yra apjungiamos, o jų taškai priskiriami vienam klasteriui, jei apjungtų hipersferų tankio įvertis yra didesnis už atskirose hipersferose esančių stebėjimų įvertinamą tankį [14]

$$\hat{g}_i = \frac{n_i}{nV_i}, \tag{6}$$

čia n_i – stebėjimų skaičius i-tame klasteryje, V_i – klasterio tūris, $i = 1, 2, \ldots, \hat{q}$.

Sugrupavus imties **X** stebėjimus į klasterius kiekviename klasteryje įvertinamas parametras $\theta = (p_i, M_i, R_i, i = 1, 2, \ldots, \hat{q})$.

3. Eksperimentinis tyrimas

Ankstesniame skyriuje aprašytų klasterizavimo algoritmų efektyvumo tyrimas atliktas Monte–Karlo metodu. Toks algoritmų palyginimo būdas sudarė galimybes išmatuoti tikrąsias stebėjimų grupes ir tuo būdu įvertinti algoritmų efektyvumą. Tyrimui buvo naudojami Gauso skirstinių mišiniai.

Klasterizavimo tikslumui vertinti skaičiuojamas padarytų klaidų priskiriant stebėjimus atskiroms grupėms santykinis dažnis

$$\Delta(\hat{v}) = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}_{\{\hat{v}(X(t)) \neq v(X(t))\}}. \tag{7}$$

Skaičiavimai atlikti su imties didumu $n = 1000$ keičiant mišinį sudarančių skirstinių vidurkius, kovariacines matricas ir atsitiktinio dydžio X matavimų skaičių.

Algoritmai: 1 – EM, 2 – automatizuotas EM, 3 – hipersferinis, 4 – k-vidurkių, 5 – vienodų kovariacinių matricų, 6 – apjungtas k-vidurkių ir EM, 7 – apjungtas vienodų kovariacinių matricų ir EM algoritmai.

Vienodų svorių ($p_i = 0,5,\ i = 1, 2$) Gauso mišiniai ($d = 5$):

Gauso hipersferiniai mišiniai (naudoti K. Fukunaga [2], O.J. Dunn [9])

I – $M_1 = (0, \ldots, 0)$, $M_2 = (1{,}68,\ 0,\ \ldots,\ 0)$,
$R_1 = I = \mathrm{diag}([1, \ldots, 1])$, $R_2 = I = \mathrm{diag}([1, \ldots, 1])$.

II – $M_1 = (0,\ldots,0)$, $M_2 = (2{,}56,\ 0,\ \ldots,\ 0)$,
R_1 =I=diag([1, ..., 1]), R_2 =I=diag([1,..., 1]).

III – $M_1 = (0,\ldots,0)$, $M_2 = (4{,}65,\ 0,\ \ldots,\ 0)$,
R_1 =I=diag([1, ..., 1]), R_2 =I=diag([1,..., 1]).

Duin mišinys (naudotas R.P.W. Duin [4], M. Skurichina [13])

IV – $M_1 = (0,\ldots,0)$, $M_2 = (3\sqrt{2}, 0,\ldots,0)$,

$$R_1 = R_2 = \begin{pmatrix} \frac{41}{80} & -\frac{39}{80} & 0 & \ldots & 0 \\ -\frac{39}{80} & \frac{41}{80} & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}.$$

J. Van Ness mišiniai [10]

V – $M_1 = (-\frac{2}{2\sqrt{2}}, 0,\ldots,0, -\frac{2}{4\sqrt{2}})$, $M_2 = (\frac{2}{2\sqrt{2}}, 0,\ldots,0, \frac{2}{4\sqrt{2}})$,
R_1 =I, R_2 =diag([1, 1, 1, 0,5, 0,5]).

VI – $M_1 = (-\frac{4}{2\sqrt{2}}, 0,\ldots,0, -\frac{4}{4\sqrt{2}})$, $M_2 = (\frac{4}{2\sqrt{2}}, 0,\ldots,0, \frac{4}{4\sqrt{2}})$,
R_1 =I, R_2 = diag([1, 1, 1, 0,5, 0,5]).

S. Marks ir O.J. Dunn mišiniai [9]

VII – $M_1 = (0,\ldots,0)$, $M_2 = (0,\ldots,0,1)$,
R_1 =I, R_2 =diag([8, 8, 8, 1, 1]).

VIII – $M_1 = (0,\ldots,0)$, $M_2 = (0,\ldots,0,2)$,
R_1 =I, R_2 =diag([8, 8, 8, 1, 1]).

IX – $M_1 = (0,\ldots,0)$, $M_2 = (0,\ldots,0,4)$,
R_1 =I, R_2 = diag([8, 8, 8, 1, 1]).

W. Highleyman mišinys [3]

X – $M_1 = (1,1,0,\ldots,0)$, $M_2 = (2,0,\ldots,0)$,
R_1 =diag([1, 0,25, 1, 1, 1]), R_2 = diag([0,01, 4, 1, 1, 1]).

Įvertinus modeliuotų Gauso mišinių klasterizavimo paklaidas (1 lentelė) matosi, kad esant vienodoms kovariacinėms matricoms, o klasterių centrams nutolus efektyvūs yra k-vidurkių ir vienodų kovariacinių matricų algoritmai (**I**, **II**, **III**). Hipersferinis algoritmas neblogai veikia tik esant mažai išsibarsčiusiems ir toli vienas nuo kito esantiems klasteriams (**III**, **IV**). Naudojant apjungtus k-vidurkių ir EM, arba vienodų kovariacinių matricų ir EM algoritmus matosi, jog blogai parinkus parametro θ įvertį k-vidurkių arba vienodų kovariacinių matricų algoritmu, EM algoritmas klasterizuoja blogai, o jei pradinis klasterizavimas atliekamas gerai, tai vėliau taikyti EM algoritmą nėra efektyvu (**I**, **II**, **III**).

1 lentelė

Klasterizavimo tikslumo įvertinimas duomenis grupuojant skirtingais algoritmais

Gauso	*Algoritmai*						
mišiniai	1	2	3	4	5	6	7
I	0,217	0,306	0,511	0,217	0,219	0,219	0,221
II	0,122	0,124	0,511	0,121	0,120	0,124	0,121
III	0,019	0,019	0,023	0,019	0,016	0,020	0,019
IV	0,000	0,000	0,004	0,001	0,000	0,000	0,000
V	0,195	0,290	0,510	0,216	0,193	0,212	0,194
VI	0,039	0,044	0,081	0,059	0,044	0,053	0,042
VII	0,225	0,237	0,636	0,244	0,229	0,240	0,230
VIII	0,064	0,072	0,689	0,087	0,073	0,084	0,073
IX	0,003	0,003	0,758	0,004	0,003	0,003	0,003
X	0,059	0,058	0,509	0,253	0,204	0,223	0,184

Vienodų svorių ($p_i = 0,5,\ i = 1,2$) ir vidurkių ($\mathrm{M}_i = (0,\dots,0),\ i = 1,2$) Gauso mišiniai ($d = 5$):

$$\text{XI} \quad - \quad s_{1,jk} = 1 \cdot 1_{j=k}, \quad s_{2,jk} = 0,3 \cdot 1_{j=k}, \quad j,k = \overline{1;5};$$

$$\text{XII} \quad - \quad s_{1,jk} = 1 \cdot 1_{j=k}, \quad s_{2,jk} = 0,25 \cdot 1_{j=k}, \quad j,k = \overline{1;5};$$

$$\text{XIII} \quad - \quad s_{1,jk} = 1 \cdot 1_{j=k}, \quad s_{2,jk} = 0,2 \cdot 1_{j=k}, \quad j,k = \overline{1;5}.$$

2 lentelė

Klasterizavimo tikslumo įvertinimas, kai klasterių centrai sutampa

Gauso	*Algoritmai*						
mišiniai	1	2	3	4	5	6	7
XI	0,287	0,295	0,496	0,493	0,469	0,494	0,458
XII	0,281	0,293	0,495	0,494	0,477	0,494	0,461
XIII	0,193	0,223	0,497	0,494	0,489	0,493	0,487

Įvertinus modeliuotų Gauso mišinių klasterizavimo paklaidas (2 lentelė) matosi, kad klasterių centrams sutampant efektyvus yra EM algoritmas.

Atlikti skaičiavimai parodė, kad priklausomai nuo atsitiktinio dydžio X dimensijos d mišinio klasterizavimo tikslumas kinta ne monotoniškai. Bendrai paėmus, atsižvelgiant į skaičiavimo rezultatus, galima teigti, kad EM, k-vidurkių ir vienodų kovariacinių matricų klasterizavimo algoritmai duoda panašius rezultatus nepriklausomai nuo dimensijos, kai mišinį sudarančių klasterių kovariacinės matricos sutampa, o skiriasi tiktai jų vidurkiai.

Išvados

1. Atlikta EM, hipersferinio, k-vidurkių ir vienodų kovariacinių matricų klasterizavimo algoritmų analizė. Modeliuojant atsitiktinius dydžius apytiksliai įvertintos šiais klasterizavimo algoritmais gaunamos klaidos.

2. Panaudojus taikomosios statistikos metodus sukurtas Gauso skirstinių mišinio klasifikavimo efektyvumo tyrimo modelis, kuris realizuotas programiškai panaudojus SAS programavimo priemones.
3. Gauso mišinio imitacinis tyrimas parodė, kad klasterizuojant daugiamačius duomenis tiksliausi rezultatai gaunami taikant EM algoritmą, kiti naudoti algoritmai duoda tikslų rezultatą, kai tiriami vienas nuo kito nutolę klasteriai. Gauti rezultatai rodo, kad klasterių centrams sutampant, o skiriantis tiktai kovariacinėms matricoms hipersferinio, k-vidurkių ar vienodų kovariacinių matricų klasterizavimo algoritmų taikymas yra visiškai neefektyvus. Kai Gauso skirstinių vidurkiai skiriasi, efektyvu atlikti pradinį duomenų grupavimą naudojant k-vidurkių arba vienodų kovariacinių matricų klasterizavimo algoritmą ir įvertinus parametrą $\theta = (p_i, M_i, R_i, i = 1, 2, \ldots, q)$ toliau jį naudoti EM algoritme.

Literatūra

[1] R.P. Cody, J.K. Smith, *Applied Statistics and the SAS Programming Language*, Fourth edition, Prentice Hall, New Jersey (1997).

[2] K. Fukunaga, *Statistical Pattern Recognitions*, Second Edition, Academic Press, Boston (1990).

[3] W. Highleyman, The design and analysis of pattern recognition experiments, *Bell System Technical Journal*, **41**, 723–744 (1962).

[4] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, *IEEE Transactions on PAMI*, **22**(1), 4–37 (2000).

[5] G. Jakimauskas, J. Sushinskas, *Computational Aspacts of Statistical Analysis of Gaussian Mixture Combining EM Algorithm with Non-parametric Estimation (One-dimensional Case), Preprint No. 96-6*, Institute of Mathematics and Informatics, Vilnius (1996).

[6] G. Jakimauskas, Efficiency analysis of one estimation and clusterization procedure of one-dimensional Gaussian mixture, *Informatica*, **8**(3), 331–343 (1997).

[7] G. Jakimauskas, R. Krikstolaitis, Influence of projection pursuit on classification errors: computer simulation results, *Informatica*, **11**(2), 115–124 (2000).

[8] T. Marill, D.M. Green, On the ectiveness of receptors in recognition system, *IEEE, Transactions on Information Theory*, **9**, 11–17 (1963).

[9] S. Marks, O.J. Dunn, Discriminant functions when the covariance matrices are unequal, *Journal of the American Statistical Association*, **69**(346), 555–559 (1974).

[10] J. Van Ness, On the dominance of non-parametric Bayes rule discriminant algorithms in high dimensions, *Pattern Recognition*, **12**, 355–368 (1980).

[11] R. Rudzkis, M. Radavicius, Statistical estimatios of a mixture of Gaussian distributions, *Acta Applicandae Mathematicae*, **38**, 37–54 (1995).

[12] M.E. Stokes, C.S. Davis, G.G. Koch, *Categorical Data Analysis Using the SAS System*, SAS Institute Inc., Cary (1995).

[13] M. Skurichina, R.P.W. Duin, Boosting in linear discriminant analysis, in: *First International Workshop in Multiple Classifiers Systems*, Cagliari (2000).

[14] *SAS/STAT® User's Guide*, Version 8, Second Edition, Volume 1 and 2, SAS Institute Inc., Cary, NC (2001).

Comparison of various clustering algorithms efficiency

T. Ruzgas

This article illustrates the problem of clustering efficiency of Gaussian mixture models using various clustering methods. The results of investigation by simulation are discussed.