سیزدهمین کنفرانس آمار ایران
۴-۲ شهریور ۱۳۹۵

انتخاب مدل در خوشه‌بندی بلوکی به‌کمک انتگرال درستنمایی رده‌بندی

حمیده سادات فاطمی‌قمی [۱] ، موسی گل‌علی‌زاده [۲]، منصور رزقی [۳]

[۲،۱]گروه آمار دانشگاه تربیت‌مدرس

[۳]گروه علوم کامپیوتر دانشگاه تربیت‌مدرس

چکیده: هدف خوشه‌بندی بلوکی، یافتن ساختاری به‌صورت بلوک‌های همگن است که موجب افراز همزمان سطر و ستون‌ها در ماتریس داده می‌شود. به‌کارگیری مدل بلوکی پنهان برای این منظور به‌عنوان یک مدل خوشه‌بندی مدل مبنا است که درواقع یک گسترشی از مدل آمیخته متناهی است. همان‌طور که در خوشه‌بندی یک سویه، انتخاب مدل از نظر یافتن بهترین تعداد خوشه‌ها یک موضوع حیاتی به‌شمار می‌رود؛ در خوشه‌بندی همزمان نیز یافتن بهترین تعداد بلوک‌ها باید مورد بررسی قرار بگیرد. در این مقاله، ضابطه مبتنی بر تقریب انتگرال درستنمایی رده‌بندی برای مدل بلوکی تعمیم داده می‌شود و یک معیار اطلاع بیزی نیز به کمک آن بدست می‌آید. این ضابطه‌ها روی داده‌های شبیه‌سازی شده و با اندازه‌های مختلف از ماتریس و با چند دسته از تداخل خوشه‌ها مقایسه خواهند شد.

واژه‌های کلیدی: خوشه‌بندی بلوکی، مدل بلوکی پنهان، انتخاب مدل، انتگرال درستنمایی رده‌بندی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H30، 07 – 62.

۱ مقدمه

در سال‌های اخیر یکی از روش‌های مورد علاقه محققان در زمینه خوشه‌بندی، به‌کارگیری خوشه‌بندی همزمان[۱] در یک ماتریس داده است که در زمینه‌های مختلفی مانند تشخیص الگو، ژنتیک و تحلیل متن کاربرد دارد. خوشه‌بندی بلوکی[۲] به کمک یک مدل احتمالی به‌نام مدل

[۱] حمیده سادات فاطمی‌قمی : h.sadat.fatemi@gmail.com

[۱] Biclustering

[۲] Block clustering

بلوکی پنهان[۳] یکی از این روش‌ها است که در یک ماتریس با افراز همزمان سطر و ستون‌ها، بلوک‌های همگن (خوشه‌ها) را تشخیص می‌دهد (گووارت و ندیف (۲۰۰۸)). اما پیاده‌سازی این نوع خوشه‌بندی با ابعاد مختلفی از بلوک امکان‌پذیر است بنابراین انتخاب تعداد مناسب خوشه‌ها یک مسئله مهم شناخته شده در بحث خوشه‌بندی از جمله خوشه‌بندی همزمان به‌شمار می‌رود. تاکنون روش‌های مختلفی برای انتخاب مدل برای خوشه‌بندی همزمان به‌کار گرفته شده است. اولین رویکرد ایده خاصی از ذات مسئله خوشه‌بندی همزمان نمی‌گیرد و تنها ضابطه ارزیابی برای خوشه‌بندی یک‌سویه (نظیر ضابطه سیل‌هات) را با دوبار محاسبه نمودن برای سطر وستون‌ها ترکیب می‌کند (اسچپر و همکاران (۲۰۰۸) و چارد و همکاران (۲۰۱۰)). روسی و ویچی (۲۰۰۸) در روشی دیگر، واگرایی بین بلوک‌ها را به واگرایی خوشه‌بندی بلوکی تقسیم کردند. همچنین برخی ضابطه‌های اطلاع برای انتخاب مدل، برای خوشه‌بندی بلوکی به‌کار گرفته شده اند. ون دیجک و همکاران (۲۰۰۹) یک ضابطه آکائیک اصلاح شده به‌این منظور معرفی نمودند. اما در سال‌های اخیر نیز ضابطه مبتنی بر انتگرال درستنمایی رده‌بندی (بیرناکی و همکاران (۲۰۰۰)) تعمیم داده شده است که در این مقاله روی این ضابطه متمرکز خواهیم شد. ساختار مقاله حاضر به‌صورت زیر تدوین شده است. در بخش ۲ مدل بلوکی پنهان تشریح می‌شود. انتخاب مدل براساس انتگرال درستنمایی رده‌بندی در بخش ۳ می‌آید و مطالعه شبیه‌سازی تشکیل‌دهنده بخش ۴ است.

۲ مدل بلوکی پنهان

در خوشه‌بندی همزمان یا بلوکی، هدف یافتن افراز همزمان سطرها و ستون‌ها (z, w) است. در این روش z یک افرازی است که موجب خوشه‌بندی سطرها (I) به g خوشه می‌شود و w نیز یک افراز روی مجموعه ستون‌ها (J) به m خوشه است. در این حالت توزیع داده‌ها را می‌توان با تابع چگالی احتمال $f(x,\theta) = \sum_{u\in U} p(u) f(x\,|u,\theta)$ نشان داد که در آن U نشان‌دهنده کلیه افرازهای ممکن روی $I \times J$ است. اگر مجموعه افرازها روی I و J مستقل از هم فرض شوند، در این حالت $p(u) = p(z)p(w)$ و لذا

$$f(x;\theta) = \sum_{(z,w)\in Z\times W} p(z)p(w) f(x\,|z,w;\alpha)\,, \qquad (۱.۲)$$

که در آن Z و W به‌عنوان متغیرهای پنهان، مجموعه همه افرازهای ممکن I به g خوشه و J به m خوشه را تشکیل می‌دهند. با توجه به فرض استقلال مکانی[۴] مربوط به مدل کلاس پنهان[۵] تابع چگالی با شرطی شدن روی z و w به‌صورت

$$f(x\,|z,w;\theta) = \prod_{i,j,k,l} f(x_{ij};\alpha_{kl})^{z_{ik}w_{jl}}.$$

بدست می‌آید. فرض می‌شود $\pi = (\pi_1,...,\pi_g)$ و $\rho = (\rho_1,...,\rho_m)$ بردارهای احتمال‌های π_k و ρ_l باشند و این احتمال پیشین را بیان کنند که سطر و ستون به‌ترتیب متعلق به k امین مولفه و l امین مولفه باشد. باتوجه به (۱.۲) تابع درستنمایی برای مدل بلوکی به‌صورت

$$f(x;\theta) = \sum_{(z,w)\in Z\times W} \prod\nolimits_{i,k} \pi_k^{z_{ik}} \prod\nolimits_{j,l} \rho_k^{w_{jl}} \prod_{i,j,k,l} f(x_{ij};\alpha_{kl})^{z_{ik}w_{jl}} \qquad (۲.۲)$$

بدست می‌آید. از آنجایی‌که مدل شامل متغیرهایی پنهان برای تبیین افرازها هستند؛ روند الگوریتم EM که شامل فرآیندی متناوب بین گام امیدگیری (گام E) و گام ماکسیمم‌سازی (گام M) است؛ به ذهن می‌رسد. لگاریتم درستنمایی رده‌بندی با استفاده از داده کامل می

[۳] Latent Block Model (LBM)
[۴] Local independence
[۵] Latent class model

تواند به‌صورت زیر نوشته شود:

$$L_C(x,z,w,\theta)=\sum_{i,k} z_{ik}\log\pi_k+\sum_{j,l} w_{jl}\log\rho_l+\sum_{i,j,k,l} z_{ik}w_{jl}\log f(x_{ij};\alpha_{kl}).$$

اما روند الگوریتم EM برای LBM به‌دلیل ضرورت محاسبه $P(Z_{ik}W_{jl}=1|\theta^{(c)},X=x)$ غیرقابل اجراست. برای رفع این مشکل گووارت و ندیف (۲۰۰۸) یک تقریب به‌نام الگوریتم EM تغییرگرا [۶] را پیشنهاد دادند. بنا به این الگوریتم در گام E در این الگوریتم توزیع توأم $p(z,w|X,\theta)$ برابر با حاصلضرب $p(z|X,\theta)p(w|X,\theta)$ خواهد شد و ضابطه

$$F(P,\theta)=E_P(L_C(z,w,\theta))+H(P) \qquad (۳.۲)$$

مورد ماکسیمم سازی قرار می‌گیرد. به پیشنهاد کریبین و همکاران (۲۰۱۲) بعد از همگرایی، مقدار F که یک کران پایین لگاریتم درستنمایی است؛ می‌تواند به‌عنوان تقریبی از آن درنظر گرفته شود. لازم به‌ذکر است پیاده‌سازی این الگوریتم مبتنی بر LBM مستلزم اجرای گام‌هایی متناوب برای دو الگوریتم خوشه‌بندی روی سطرها ستون‌ها مبتنی بر مدل‌های آمیخته متناهی است.

۳ انتخاب مدل

موضوع انتخاب مدل در LBM به چند دلیل مسئله‌ساز است. اولین دلیل این است که یک زوج از تعداد خوشه‌ها یعنی (g,m) به‌جای یک عدد تنها باید انتخاب شود. دومین دلیل این است که ضابطه‌های درستنمایی جریمه‌ای موجود نظیر AIC و BIC که برای انتخاب مدل در خوشه‌بندی مدل مبنا به‌کار گرفته می‌شوند؛ نمی‌توانند به‌طور مستقیم مورد استفاده قرار گیرند. سوم اینکه انتخاب مدل معمولاً به تعداد واحدهای آماری مدل وابسته است، لذا گزینش این موضوع در LBM سؤال برانگیز است. به‌زبانی دیگر انتخاب تعداد واحد آماری مدل باید از تعداد مشاهدات n یا تعداد متغیرها d یا تعداد سلول‌ها صورت گیرد. درادامه با درنظر گرفتن این مسئله، انتخاب مدل بررسی می‌شود. به‌طور کلی، هدفی که در بحث انتخاب مدل دنبال می‌شود؛ رسیدن به بهترین مدل است. در چارچوب بیزی بهترین مدل، محتمل‌ترین آنهاست. درحقیقت در میان مجموعه متناهی مدل، مدلی انتخاب می‌شود که بیشترین احتمال پسین را دارا باشد. به‌زبانی ساده

$$M_h=\arg\max_{M\in M} p(\mathrm{M}|X).$$

اگر فرض شود که احتمال پیشین $p(\mathrm{M})$ برای هر مدل، ثابت باشد؛ آنگاه ماکسیمم نمودن $p(\mathrm{M}|X)$ معادل با ماکسیمم نمودن $p(X|M)$ است و سپس هدف یافتن مدلی می‌شود که انتگرال درستنمایی را ماکسیمم کند. لذا

$$M_h=\arg\max_{M} p(X|M),$$

طوری‌که $p(X|M)=\int_\Theta p(X|M;\theta)p(\theta|M)d\theta$. در اینجا Θ فضای پارامتر و $p(\theta|M)$ احتمال پیشین پارامتر θ برای مدل M است. در حالاتی که مشاهدات مستقل و هم‌توزیع باشند؛ ضابطه BIC یک راهی برای تقریب مجانبی از لگاریتم انتگرال درستنمایی است. فرم استاندارد این ضابطه به صورت

$$BIC(\mathrm{M})=\log L(x;\hat\theta_{ML})-\frac{1}{2}D_{\mathrm{M}}\log(n)$$

[۶] Variational EM (VEM)

است که در آن D_{M} بعد مدل، n تعداد مشاهده و $L(y;\hat{\theta}_{ML})$ ماکسیمم درستنمایی است. اما این تقریب به‌دلیل ساختار وابستگی مشاهدات بین سطرها و ستون‌ها نمی‌تواند در مدل بلوکی، به‌کار گرفته شود. برای حل این مشکل، تعریف دیگری از انتخاب مدل، به‌صورت $\mathrm{M}_c = \arg\max_{\mathrm{M}} P(X,z,w|\mathrm{M})$ ارائه شد که در آن

$$P(X,z,w|\mathrm{M}) = \int_{\Theta} p(X,z,w|\theta;\mathrm{M})p(\theta|\mathrm{M})d\theta.$$

درنهایت ضابطه مدنظر، با لگاریتم گرفتن به‌صورت $ICL(\mathrm{M}) = \log P(X,z,w|\mathrm{M})$ تعریف شده و به اختصار انتگرال درستنمایی کامل (یا رده‌بندی) [۷] نامیده می‌شود (لومت و همکاران (۲۰۱۲)). همچنین تحت فرض استقلال برچسب‌ها داریم:

$$ICL(\mathrm{M}) = \log p(X|z,w,\mathrm{M}) + \log p(z|\mathrm{M}) + \log p(w|\mathrm{M}). \qquad (۱.۳)$$

در ادامه یک تقریب مجانبی از ICL معرفی می‌شود. مشاهده می‌شود که این تقریب صریحاً به توزیع پیشین پارامترها بستگی ندارد. همچنین با الهام از آن ضابطه مشابه BIC با نماد BIC_{like} برای مدل بلوکی ارائه می‌شود.

۱.۳ تقریب ICL و بدست آوردن ضابطه BIC در مدل بلوکی

تحت فرض استقلال پیشین پارامترها ، یعنی $p(\theta) = p(\alpha)p(\pi)p(\rho)$ ، بخش اول ICL به‌صورت

$$\log p(X|z,w,\mathrm{M}) = \int_A p(X|z,w,\mathrm{M},\alpha)p(\alpha|\mathrm{M})d\alpha) \underbrace{\int_{P\times R} p(\pi|\mathrm{M})p(\rho|\mathrm{M})d\pi d\rho}_{=۱}$$

محاسبه می‌شود. این بخش می‌تواند با استفاده از BIC_{like} تقریب زده شود. یعنی

$$\log p(X|z,w,\mathrm{M}) = \max \log p(X|z,w,\mathrm{M},\alpha) - \frac{\lambda}{۲}\log(nd)$$

که در آن λ بعد بردار پارامتر α در فضای A است. طبق $p(z|\pi) = \prod_{i,k}\pi_k^{z_{ik}}$ و درنظر گرفتن توزیع ناآگاهی بخش دیریکله $D(a,...,a)$ به‌عنوان توزیع پیشین برای π ، و بااستفاده از قانون بیز، $p(z)$ به‌صورت

$$p(z) = \frac{p(z|\pi)p(\pi)}{p(\pi|z)} = \frac{\Gamma(ga)}{\Gamma(a)^g}\frac{\prod_k \Gamma(z_{.k}+a)}{\Gamma(n+ga)}$$

بدست می‌آید و با همین روند، $p(w)$ محاسبه می‌شود. با قرار دادن $a = \frac{۱}{۲}$ برای یک توزیع ناآگاهی‌بخش جفریز (رابرت (۲۰۰۱))، لگاریتم توزیع‌های پیشین به‌صورت

$$\log p(z|\mathrm{M}) = \log\Gamma(\tfrac{g}{۲}) + \sum_{k=۱}^{g}\log\Gamma(n_k+\tfrac{۱}{۲}) - g\log\Gamma(\tfrac{۱}{۲}) - \log\Gamma(n+\tfrac{g}{۲}),$$
$$\log p(w|\mathrm{M}) = \log\Gamma(\tfrac{m}{۲}) + \sum_{l=۱}^{m}\log\Gamma(d_l+\tfrac{۱}{۲}) - m\log\Gamma(\tfrac{۱}{۲}) - \log\Gamma(d+\tfrac{m}{۲}).$$

خواهند بود. اگر n_k و d_l به اندازه کافی بزرگ باشند؛ با به‌کارگیری تقریب استرلینگ $\Gamma(t+۱) \approx t^{t+۱/۲}\exp(-t)(۲\pi)^{۱/۲}$ برای تابع گاما و نادیده گرفتن بخش‌هایی که از $O(۱)$ هستند؛ لگاریتم توزیع پیشین به‌صورت زیر حاصل خواهند شد:

$$\log p(z|\mathrm{M}) \approx \sum_{k=۱}^{g} n_k\log n_k - n\log n - \frac{۱}{۲}(g-۱)\log n$$

[۷] Integrated Complete(Classification) Likelihood (ICL)

$$\log p(w|\mathrm{M}) \approx \sum_{l=1}^{m} d_l \log d_l - d \log d - \frac{1}{2}(m-1)\log m.$$

علاوه براین، به‌کمک تساوی $p(z|\pi) = \prod_k \pi_k^{n_k}$ و شرط $\sum_k \pi_k = 1$ رابطه $\sum_{k=1}^{g} n_k \log \frac{n_k}{n} = \max_\pi \log p(z|\pi, \mathrm{M})$ و به‌طور مشابه $\sum_{l=1}^{m} d_l \log \frac{d_l}{d} = \max_\rho \log p(w|\rho, \mathrm{M})$ بدست خواهند آمد. با جای‌گذاری پارامتر θ و بردار برچسب‌های (z, w) از طریق الگوریتم VEM سرانجام ضابطه ICL به‌صورت زیر حاصل خواهد شد:

$$ICL_{BIC}(\mathrm{M}) = L_c(X, \hat{z}, \hat{w}|\mathrm{M}, \hat{\theta}) - \frac{\lambda}{2}\log(nd) - \frac{g-1}{2}\log n - \frac{m-1}{2}\log d.$$

این ضابطه یک درستنمایی رده‌بندی جریمه‌ای محسوب می‌شود که در آن بخش جریمه جمع پارامترهای مرتبط با توزیع‌ها و همچنین پارامترهای مربوط به نسبت‌های سطری و نسبت‌های ستونی هستند. با الهام از رابطه $ICL = BIC + H$ در مدل آمیخته کلاسیک، که H تنها یک بخش آنتروپی مربوط به احتمال‌های شرطی برچسب‌هاست؛ ضابطه BIC_{like} به‌صورت

$$BIC_{like} = \log p(X|\mathrm{M}, \theta^*) - \frac{\lambda}{2}\log(nd) - \frac{g-1}{2}\log n - \frac{m-1}{2}\log d \qquad (۲.۳)$$

حاصل می‌شود.

۲.۳ محاسبه ضابطه دقیق ICL

در این بخش ضابطه دقیقی برای تعریف (۱.۳) بدست می‌آید. این ضابطه نیازی به هیچ تعریف مجانبی ندارد؛ اما توزیع پیشین برای پارامترها باید به‌طور مناسبی درنظر گرفته شوند. این موضوع برای دو حالت دنبال می‌شوند:

الف) ضابطه دقیق ICL برای داده دودویی

توزیع‌های پیشین برای پارامترهای مربوط به متغیر برنولی به‌صورت $\pi \sim D(a, ..., a)$ ، $\rho \sim D(a, ..., a)$ و $\alpha_{kl} \sim \mathrm{B}(b, b)$ درنظر گرفته می‌شود. لذا،

$$\begin{aligned} ICL_{(a,b)}(g, m) &= \log p(x, z, w) \\ &= \log\Gamma(ga) - (m+g)\log\Gamma(a) + mg(\log\Gamma(2b) - 2\log\Gamma(b)) \\ &\quad - \log\Gamma(n+ga) - \log\Gamma(d+ma) + \sum_k \log\Gamma(z_{.k}+a) + \sum_l \log\Gamma(w_{.l}+a) \\ &\quad + \sum_{k,l}\left[(\sum_{h=1}^{2}\log\Gamma(N_{kl}^h + b)) - \log\Gamma(z_{.k}w_{.l} + rb)\right]. \end{aligned}$$

طبق کریبین و همکاران (۲۰۱۴) درنظرگرفتن ابرپارامترهای $a = 4$ برای تعداد خوشه‌های کمتر از ۸ و نیز مقدار $a = 16$ برای ابعاد بزرگتر و همچنین درنظر گرفتن ابرپارامتر $b = 1$ نتایج بهتری را حاصل خواهد کرد.

ب) محاسبه ضابطه دقیق ICL برای داده پیوسته

با فرض متفاوت بودن واریانس‌های هر بلوک، یک توزیع پیشین نرمال با پارامترهای $(\mu_0, \frac{\sigma_{kl}^2}{\kappa_0})$ برای میانگین شرطی هر بلوک یعنی $\mu_{kl}|\sigma_{kl}^2$ در نظر گرفته می‌شود. همچنین یک توزیع پیشین معکوس خی‌دوی مقیاس شده برای پارامتر واریانس هر بلوک (σ_{kl}^2) ، لحاظ می‌شود. به‌زبانی دقیق‌تر، $\sigma_{kl}^2 \sim Scale - Inv - \chi^2(\upsilon_0, \sigma_0^2)$. آنگاه بخش اول معادله (۱.۳) به‌صورت

$$\log p(X|z, w, \mathrm{M}) = -\frac{nd}{2}\log\pi + \frac{gm\upsilon_0}{2}\log(\upsilon_0\sigma_0^2) - gm\log\Gamma(\frac{\upsilon_0}{2}) + \frac{gm}{2}\log\kappa_0$$

$$-\sum_{k,l}\frac{\upsilon_\circ+n_kd_l}{۲}\log(\upsilon_\circ\sigma_\circ^{۲}+(n_kd_l-۱)s_{kl}^{۲*}+\frac{\kappa_\circ n_kd_l}{\kappa_\circ+n_kd_l}(\bar{x}_{kl}-\mu_\circ)^{۲})$$
$$+\sum_{k,l}\log\Gamma\left(\frac{\upsilon_\circ+n_kd_l}{۲}\right)-\frac{۱}{۲}\log(\kappa_\circ+n_kd_l).$$

خواهد شد که در آن n_k و d_l به‌ترتیب تعداد سطرها و ستون‌ها در بلوک (k,l) $\bar{x}_{kl}=۱/(n_kd_l)\sum_{i,j}z_{ik}w_{jl}x_{ij}$ ، میانگین نمونه‌ای، $s_{kl}^{۲*}=۱/(n_kd_l-۱)\sum_{i,j}z_{ik}w_{jl}(x_{ij}-\bar{x}_{kl})^{۲}$ ، واریانس نمونه‌ای ناریب هستند. برای پارامترهای نسبت (π,ρ) ، با فرض اینکه درایه‌های آن از همدیگر متفاوت هستند، یک توزیع پیشین دیریکله با پارامترهای $(a,...,a)$ درنظر گرفته می‌شود. بنابراین دوبخش آخر معادله (۱.۳) به‌ترتیب به‌صورت زیر تغییر می‌یابند:

$$\log p(z|\mathrm{M})=\log\Gamma(ga)+\sum_k\log\Gamma(n_k+a)-g\log\Gamma(a)-\log\Gamma(n+ga)$$

$$\log p(w|\mathrm{M})=\log\Gamma(ma)+\sum_l\log\Gamma(d_l+a)-m\log\Gamma(a)-\log\Gamma(d+ma).$$

برخلاف ضابطه‌های تقریبی BIC و ICL ، ضابطه دقیق ICL به توزیع پیشین پارامترها بستگی دارد. اگرچه درنظر گرفتن این اطلاعات پیشین می‌تواند در استنباط $\hat{z}$ و $\hat{w}$ و بنابراین تخمین پارامترهای مدل اثر بگذارد؛ با این حال در شبیه‌سازی‌ها ترجیح داده شده که از برآورد ماکسیمم درستنمایی استفاده می‌شود. گرچه این انتخاب غیر متعارف است و ممکن است برای ضابطه دقیق ICL مناسب نباشد؛ ولی بنا به لومت و همکاران (۲۰۱۴). اجازه می‌دهد تا با ضابطه‌های انتخاب مدل با برآورد پارامترهای یکسان مقایسه شود.

ضابطه ICL به ۵ ابرپارامتر $(a,\mu_\circ,\kappa_\circ,\nu_\circ,\sigma_\circ^{۲})$ نیاز دارد. برای توزیع دیریکله مربوط به پارامترهای نسبت‌ها، یک پیشین ناآگاهی بخش جفریز یعنی $a=\frac{۱}{۲}$ انتخاب خواهد شد. برای پارامتر مربوط به میانگین‌های هر بلوک یعنی $\mu_\circ$ ، میانگین کلی ماتریس داده درنظر گرفته می‌شود. ابرپارامتر $\kappa_\circ$ نیز برابر $\sqrt{nd/gm}$ لحاظ می‌شود. واضح است که این مقدار با بالارفتن تعداد درایه‌ها در هر خوشه، افزایش می‌یابد که با این انتخاب، پیچیدگی مسئله لحاظ شده است و درواقع گویای این نکته نیز هست که با افزایش بعد داده نسبت خطا کاهش می‌یابد. طبق لومت و همکاران (۲۰۱۴) به‌صورت تجربی نشان داده شده گرفتن ریشه دوم از این نسبت، نتیجه بهتری را حاصل خواهد کرد. نسبتی که برای واریانس پیشین میانگین هر بلوک انتخاب می‌شود؛ برای پیشین مربوط به واریانس هر بلوک، ابرپارامتر $\nu_\circ$ و $\sigma_\circ^{۲}$ را طوری انتخاب می‌کند که مد توزیع معکوس خی‌دو مقیاس شده، یعنی $(\frac{\nu_\circ\sigma_\circ^{۲}}{\nu_\circ+۲})$ برابر واریانس نمونه‌ای ماتریس داده باشد. درنتیجه با درنظر گرفتن مقدار $۱\circ^{-۱}$ به‌منظور ایجاد واگرایی بیشتر، مقدار $\sigma_\circ^{۲}$ به‌صورت $۲\circ s^{۲*}$ درنظر گرفته می‌شود که در اینجا $s^{۲*}$ واریانس نمونه‌ای ناریب است.

۴ مطالعه شبیه‌سازی

برای g و m های مختلف، الگوریتم خوشه‌بندی بلوکی روی ۲۰ داده شبیه‌سازی شده با ابعاد بلوک صحیح ۲×۳ از نوع دودویی اجرا شدند تا عملکرد ضابطه‌های ICL_{BIC} ، BIC_{like} و ICL واقعی مورد مقایسه قرار بگیرند. در جدول ۱ ، اندازه ماتریس برای سه بعد مختلف و دو دسته از تداخل خوشه‌ها روی داده‌های دودویی درنظر گرفته شده است. بدیهی است تعیین پارامترها در میزان تداخل خوشه‌ها موثر خواهد بود و این تداخل براساس نسبت خطا بین دو افراز برآورد شده تعیین می‌شود. نسبت خطای حاصل از الگوریتم خوشه‌بندی، یعنی $u=(z,w)$ و افراز صحیح $u'=(z',w')$ به‌صورت $z_{ik}w_{jl}z'_{ik}w'_{jl}$ و به‌عنوان $\delta(u,u')=۱-\frac{۱}{nd}\sum_{i,j,k,l}$

میانگین تعداد خطای رده‌بندی‌های اشتباه [۸] تعریف می‌شود. بازه $\bar{\delta}$ برای این دو دسته از داده با تداخل نسبتاً تفکیک‌شده $(++)$ و تفکیک‌شده ضعیف $(+++)$ به‌ترتیب بین [۰/۰۱, ۰/۱۰] و [۰/۱۲, ۰/۲۲] قرار گرفته‌اند. اعداد داخل جدول تعداد تشخیص ابعاد بلوک در حالت‌های مختلف و تحت ضابطه‌های مربوطه را نشان می‌دهند.

جدول ۱: نتایج حاصل از مطالعه شبیه‌سازی

		++			+++		
		۳۰ × ۷۰	۱۲۰ × ۲۰۰	۲۰۰ × ۵۰۰	۳۰ × ۷۰	۱۲۰ × ۲۰۰	۲۰۰ × ۵۰۰
ICL_{BIC}	۲	۲	۱		۱۱		
	۳	۱ ۴ ۹	۱۹	۲۰	۹	۲ ۱۳	۱۶ ۴
	۴	۱ ۵				۱ ۱ ۳	
BIC_{like}	۲	۲	۱۶		۱۸		
	۳	۱ ۴ ۹	۱۴	۲۰	۲	۳ ۱۳	۶
	۴	۴				۱ ۳	
ICLدقیق	۲	۲	۲		۹		۱
	۳	۱ ۴ ۹	۲۰	۲۰	۱ ۱۰	۱ ۲ ۱۳	۱۶
	۴	۱ ۵				۱ ۳	۱ ۲

همان‌طور که ملاحظه می‌شود؛ کیفیت عملکرد ضابطه‌های با افزایش اندازه ماتریس و کاهش میزان تفکیک‌پذیری مولفه‌ها افزایش می‌یابد. در بیشتر موارد ضابطه ICL_{BIC} قادر به انتخاب مدل درست است؛ درحالیکه BIC_{like} تمایل به کم‌برآورد تعداد خوشه‌ها دارد؛ به‌خصوص هنگامی‌که ابعاد ماتریس کوچک‌تر باشد. دلیلی که می‌توان برای این امر ارائه کرد؛ این است که بعد ماتریس نمی‌تواند برای اعتبار استفاده از تقریب مجانبی BIC_{like} کافی باشد. به‌بیانی دیگر در این ضابطه به‌دلیل غیر دسترس بودن ماکسیمم درستنمایی از کران پایین آن به‌عنوان تقریبی از ماکسیمم انرژی آزاد استفاده شد که بر تفاوت رفتار درستنمایی و کران پایین آن با افزایش تعداد مولفه‌ها تاکید می‌کند.

بحث و نتیجه‌گیری

مدل بلوکی پنهان یک مدل آماری به‌صرفه از نظر برآورد تعداد پارامترها برای سازماندهی یک ماتریس داده به بلوک‌های همگن است. مسئله انتخاب تعداد مناسب خوشه‌های سطری و ستونی یک چالشی برای این ساختارها است که مقاله حاضر بر روی تعمیم ضابطه مبتنی بر انتگرال درستنمایی رده‌بندی متمرکز شده است. اما در این میان یکی از مشکلات برای داده دودویی و به‌طور کلی‌تر داده‌های رده‌ای، امکان روبرو شدن با پدیده خالی بودن برخی خوشه‌ها به‌خصوص در ابعاد کوچک‌تر است. به‌نظر می‌رسد رهیافت بیزی می‌تواند تا حدی از این مسئله جلوگیری کند که به‌عنوان مطالعات آینده مدنظر نویسندگان مقاله است.

[۸] Misclassification

مراجع

Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7), 719-725.

Charrad, M., Lechevallier, Y., Saporta, G. and Ahmed, M. B. (2010). Détermination du nombre de classes dans les méthodes de bipartitionnement, 17ème Rencontres de la Société Francophone de Classification, 119-122.

Govaert, G., and Nadif, M. (2008). Block clustering with Bernoulli mixture models: Comparison of different approaches, *Computational Statistics and Data Analysis*, 52(6), 3233-3245.

Keribin, C., Brault, V., Celeux, G. and Govaert, G. (2012). Model selection for the binary latent block model, In COMPSTAT 2012 20th International Conference on Computational Statistics.

Keribin, C., Brault, V., Celeux, G. and Govaert, G. (2014). Estimation and selection for the latent block model on categorical data, *Statistics and Computing*, 1-16.

Lomet, A., Govaert, G. and Grandvalet, Y. (2012). Model selection in block clustering by the integrated classification likelihood, In 20th International Conference on COMPUTATIONAL STATISTICS (COMPSTAT 2012), 519-530.

Lomet, A., Govaert, G. and Grandvalet, Y. (2014). Model selection for Gaussian latent block clustering with the integrated classification likelihood, *Advances in Data Analysis and Classification*, 1-20.

Robert C (2001). *The Bayesian choice*, Springer, Berlin.

Rocci, R. and Vichi, M. (2008). Two-mode multi-partitioning, *Computational Statistics and Data Analysis*, 52(4), 1984-2003.

Schepers, J., Ceulemans, E. and Van Mechelen, I. (2008). Selecting among multi-mode partitioning models of different complexities: A comparison of four model selection criteria, *Journal of Classification*, 25(1), 67-85.

van Dijk, B., van Rosmalen, J. and Paap, R. (2009). A Bayesian approach to two-mode clustering, Technical Report, Econometric Institute Research Papers.