تعیین مدل خوشه بندی احتمالاتی بر اساس معیار اطلاع بیزی

محمد قاسم وحیدی اصل [۱] – محسن محمدزاده [۲] – محمد قربانی [۳]

چکیده

یکی از مسائل مهم در تحلیل داده های چند متغیره، پیدا کردن ارتباط بین داده هاست. ساده ترین روش کشف رابطه موجود بین داده ها، رسم نمودار پراکنش است. در بسیاری از مفاهیم پزشکی، ژنتیکی و غیره، یکی از مسائل مهم، خوشه بندی داده ها به گروه های همگن است. روش های ابتکاری مختلفی از جمله، روش های سلسله مراتبی، با ماکسیمم کردن تشابهات درون گروهی، داده ها را به خوشه هایی افراز می کنند. بدلیل وابسته بودن این روش ها به تعریف فاصله بین دو خوشه و همچنین انتخاب یک مقدار آستانه برای تعیین تعداد خوشه ها، محققان در انتخاب بهترین معیاری که تشابهات درون گروهی را ماکسیمم نماید، با مشکل مواجه می باشند. در این مقاله روش «انتخاب مدل برای خوشه بندی احتمالاتی با استفاده از معیار اطلاع بیزی[۴] (BIC)» مورد بررسی قرار می گیرد که در آن، فرضهای مختلف برای معیار تشابه نقشی ندارند و با تجزیه طیفی ماتریس کوواریانس می توان معیارهای ساده ای برای مشخص کردن حجم، شکل و جهت خوشه ها به دست آورد. بعلاوه با استفاده از معیار BIC می توان بهترین مدل خوشه بندی را انتخاب نمود.

واژه های کلیدی: خوشه بندی کردن، مدل های آمیخته، الگوریتم EM، تجزیه طیفی، معیار BIC .

۱. مقدمه

یکی از مسائل مهم در بسیاری از مطالعات ژنتیکی و پزشکی، خوشه بندی داده ها است، که در آن N مشاهده با M ویژگی به g گروه همگن افراز می شوند بطوریکه تشابهات درون گروهها ماکسیمم گردند. روشهای مختلفی از جمله، روشهای سلسله مراتبی برای خوشه بندی کردن داده ها به کار می روند که در تعریف «فاصله بین دو خوشه» با هم تفاوت دارند. در روش «خوشه بندی با اتصال منفرد»، فاصله بین نزدیکترین عضوهای دو گروه بعنوان فاصله بین دو خوشه تعریف می شود، در روش «خوشه بندی با اتصال کامل»، فاصله بین دورترین زوجها و در روش اتصال میانگین، متوسط فاصله بین تمام زوجهای موجود در داخل دو خوشه به عنوان معیار خوشه بندی مورد استفاده واقع می شود (هارتیگان، [۶]). چون این روشها بر اساس مدل خاصی نمی باشند، استنباط از نمونه به جامعه امکان پذیر نیست و تعداد خوشه ها نیز به صورت ابتکاری با تعریف آستانه ای دلخواه تعیین می گردد. بنابر این لازم است روش خوشه بندی حتی الامکان مبتنی بر سلیقه محقق نبوده و بر اساس مدل یا توزیع احتمالی باشد تا بتوان در مورد آن استنباط آماری انجام داد. در این صورت شرایطی فراهم خواهد شد که بتوان با تجزیه طیفی ماتریس کوواریانس عناصر داخل خوشه ها، شکل، حجم و جهت خوشه ها را مشخص نمود. مجموعه مشاهدات تحت بررسی، عمدتاً همگی از جامعۀ خاص نمی باشند. برای تشخیص این که هر مشاهده از کدام یک جامعه آمده است، منطقی است فرض شود که هر مشاهده بر اساس ویژگی ها و خصوصیاتش دارای توزیع احتمال خاصی است. بنابر این جامعه ای مرکب از چند زیر جامعه، دارای توزیع احتمالی آمیخته از توزیع های احتمال زیر جامعه ها خواهد بود، که در حالت کلی بشکل

$$f(x \mid \boldsymbol{\psi}) = \lambda_1 f_1(x \mid \theta_1) + \cdots + \lambda_g f_g(x \mid \theta_g)$$

است، که در آن برای $f_j(\circ), j=1,\ldots,g$ تابع چگالی مؤلفه ها، $0 < \lambda_j \le 1$ و $\sum_{j=1}^{g} \lambda_j = 1$، $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_g)$، $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_g)$ و $\boldsymbol{\psi} = (\lambda, \theta)$. از توابع چگالی آمیخته معروف که خیلی کاربرد دارد

۱- گروه آمار، دانشگاه شهید بهشتی
۲- گروه آمار، دانشگاه تربیت مدرس
۳- گروه آمار، دانشگاه تبریز E- mail : m.ghorborbani@tabrizu.ac.ir
۴- Baysian Information Criterion

می توان با

$$f(x\mid\psi)=\lambda\phi(x\mid\mu_1,\sigma_1^2)+(1-\lambda)\phi(x\mid\mu_2,\sigma_2^2)$$
$$\phi(x\mid\mu_j,\sigma_j^2)=N(\mu_j,\sigma_j^2)\quad j=1,2$$

اشاره نمود. بنابر این خوشه بندی را می توان معادل با تفکیک توزیع آمیخته به مؤلفه های ساده دانست. قبل از بیان خوشه بندی بر اساس مدل می بایست یک سری فرضیات اساسی روی توزیع مؤلفه ها داشته باشیم. به عنوان مثال، یکی از این فرضها این است که توزیع هر مؤلفه تک مدی است. بنابر این هدف از خوشه بندی، تجزیه مؤلفه های چند بعدی مبهم و آمیخته به مؤلفه های ساده تک مدی است. به همین منظور در این مقاله برای خوشه بندی مشاهدات از مدل های آمیخته استفاده می شود تا بتوان تمام خصوصیات مشاهدات را در خوشه بندی بکار گرفت.

۱.۱ تعیین معیارهای خوشه بندی بر اساس مدل

فرض کنید $X_1,\ldots,X_n$ یک نمونه تصادفی از جامعه Π با زیر جامعه های $\Pi_1,\ldots,\Pi_g$، باشند و $\phi(x\mid\mu_j,\Sigma_j)$ تابع چگالی متغیر تصادفی X_i در زیر جامعه j ام، باشد. در این صورت اگر λ_j، احتمال تعلق X_i به جامعه Π_j باشد، توزیع X_i عبارت است از،

$$f(\mathbf{x}\mid\psi)=\sum_{j=1}^{g}\lambda_j\phi(x\mid\mu_j,\Sigma_j)\qquad(1)$$

که در آن $\boldsymbol{\theta}_j=(\mu_j,\Sigma_j)$، $\boldsymbol{\psi}=(\lambda_1,\ldots,\lambda_g,\theta_1,\ldots,\theta_g)$، $\Sigma_{j=1}^{g}\lambda_j=1$ و $\lambda_j>0$ است و تابع درستنمایی نمونه تصادفی به صورت

$$L(\boldsymbol{\psi}\mid\mathbf{x})=\prod_{i=1}^{n}\sum_{j=1}^{g}\lambda_j f_j(x\mid\theta_j).\qquad(2)$$

خواهد بود. برای $C_j=\{i,X_i\in\Pi_j\}$ معیار خوشه بندی حاصل از ماکسیمم کردن تابع درستنمایی

$$L(\mathbf{x}\mid C)=\prod_{i=1}^{g}\lambda_j^{nj}\prod_{X_i\in C_j}f\ (x_i\mid\theta_j).\qquad(3)$$

معادل با معیار حاصل از ماکسیمم کردن (۲) خواهد بود ([۵] و [۱۰]). معمولاً مؤلفه اصلی X_i ها معلوم نیستند و برای مشخص کردن مؤلفه اصلی X_i، متغیرهای گروه بندی Z_ij به صورت زیر تعریف می شود

$$Z_{ij}=\begin{cases}1 & X_i\in\Pi_j\\ 0 & X_i\notin\Pi_j\end{cases}$$

براساس مشخصه های گروه بندی تابع درستنمایی را می توان به صورت

$$L(\boldsymbol{\psi}\mid\mathbf{Z},\mathbf{X})=\prod_{j=1}^{g}\lambda_j^{z_{ij}}\prod_{i=1}^{n}f(x_i\mid\theta_j)^{z_{ij}}$$
$$=\prod_{j=1}^{g}\prod_{i=1}^{n}\{\lambda_i\ f(x_i\mid\theta_j)\}^{z_{ij}}$$

نوشت. که در این صورت لگاریتم تابع درستنمایی عبارت است از

(۴)

$$\log L(\boldsymbol{\psi}\mid\mathbf{x},\mathbf{Z})=\sum_{j=1}^{g}\sum_{i=1}^{n}z_{ij}\ \log\{\lambda_j\ f(x_i\mid\theta_j)\}.$$

اگر z_{ij} ها مشخص باشند، در این صورت عناصر خوشه jام به صورت $C_j=\{j\ ;z_{ij}>z_{ij'}\ j\neq j'\}$ خواهد بود. ولی اگر z_{ij} ها معلوم نباشند برای انجام تحلیل خوشه ای از برآورد مقدار مورد انتظار آنها استفاده می شود. پس تحلیل خوشه ای را می توان به عنوان برآورد z_{ij} تلقی نمود. چون

$$E(Z_{ij})=P(X_i\in\Pi_j)=\frac{\lambda_j f(x_i\mid\theta_j)}{\Sigma_{j=1}^{g}\lambda_j f(x_i\mid\theta_j)},$$

لازم است برای هر j پارامترهای نامعلوم λ_j و θ_j برآورد شوند. برآورد این پارامترها بروش ماکسیمم درستنمایی مستلزم استفاده از روشهای عددی است.

۱.۱.۱ برآورد پارامترها با استفاده از الگوریتم EM

الگوریتم امید ریاضی و ماکسیمم سازی [۱] (EM) روشی کاربردی در محاسبات تکراری، برای به دست آوردن برآورد ماکسیمم درستنمایی پارامترها در توزیع آمیخته است که در مسائل گوناگونی مانند داده های ناکامل، داده های گم شده و غیره کاربرد دارد. این الگوریتم در شرایطی به ویژه در هنگام مواجه شدن با داده های ناکامل، از الگوریتم نیوتن- افسون کار می کند[۱۱]. فرض کنید $\boldsymbol{\psi}^{(0)}$ مقدار اولیه $\boldsymbol{\psi}$ باشد، در این صورت مراحل الگوریتم EM به صورت زیر خواهد بود:

مرحله E: محاسبه امید ریاضی لگاریتم تابع درستنمایی در نقطه $\boldsymbol{\psi}^{(0)}$ به شرط مشاهده داده های کامل،

$$Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(0)})=E_{\psi^{(0)}}\{\log\ L_c\ (\boldsymbol{\psi}\mid\mathbf{x})\}$$

مرحلهٔ M: بدست آوردن مقداری مانند $\boldsymbol{\psi}^{(*)}$ برای $\boldsymbol{\psi}$ به طوری که

$$Q(\boldsymbol{\psi}^*,\boldsymbol{\psi}^{(0)})=\underset{\psi\in\Omega}{Max}(Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}))$$

۱- Expectation- Maximization Algorithm

مراحل E و M تا زمانی تکرار می شوند که شرط همگرایی $\forall \in >0: |L(\psi^{(K+1)}) - L(\psi^{(k)})| < \in$ برقرار شود.[۷].

فرض کنید $X_1,\dots,X_n$ داده های ناکامل و $\mathbf{Y}_i = (X_i, Z_i)$ داده های کامل در الگوریتم EM باشند. در این صورت برآورد پارامترهای (۴) با استفاده از این الگوریتم به صورت زیر خواهند بود.

$$\lambda_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^{(k)}$$

$$\mu_j^{(k+1)} = \sum_{i=1}^{n} \frac{z_{ij}^{(k)} X_i}{\sum_{i=1}^{n} z_{ij}^{(k)}}$$

$$(\Sigma)^{(k+1)} = \frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij}^{(k)} (X_i - \mu_j^{(k+1)})(X_i - \mu_j^{(k+1)})'$$

$$z_{ij}^{(k)} = \frac{\lambda_j^{(k)} f(x_i \mid \theta_j^{(k)})}{\Sigma_{j=1}^{g} \lambda_j^{(k)} f(x_i \mid \theta_j^{(k)})}$$

با z_{ij} برآورد شده به وسیله الگوریتم EM می توان تحلیل خوشه ای را متناسب با بیشترین مقدار $z_{ij}^{(k)}$ انجام داد[۱۱]. اما همان طوری که بدان اشاره شد در علومی چون پزشکی و فتوگرافی یکی از اهداف خوشه بندی تعیین حجم، شکل و جهت خوشه ها است. تعیین حجم خوشه ها با استفاده از تجزیه طیفی ماتریس کوواریانس نخستین بار توسط بانفیلد و رافتری [۱] بیان شد و سرانجام گیلز سیلوکس و گوورت [۳] آنرا تعمیم دادند. معیارهای تحلیل خوشه ای توسط فرالی و رفتری[۵] در نرم افزار MCLUST گنجانده شد.

۱.۲ تعیین حجم، شکل و جهت خوشه ها

تابع درستنمایی (۳) را می توان به صورت

$$L(\boldsymbol{\psi} \mid \mathbf{Z}, \mathbf{X}) = \prod_{j=1}^{g} \lambda_j^{n_j} |\Sigma_j|^{-\frac{n_j}{2}} \exp\{-\frac{1}{2}\sum_{X_i \in C_i} (\mathbf{X}_i - \boldsymbol{\mu}_j)' \Sigma_j^{-1} (\mathbf{X}_i - \boldsymbol{\mu}_j)\} \qquad (5)$$

نوشت، که در آن C_j تعداد مشاهدات X_{ij} ای است توسط Z_i به زیر جامعه Π_j تخصیص یافته است و n_j تعداد مشاهدات در C_j است. فرض کنید Σ_j، $j=1,\dots,g$، ماتریس کوواریانس مؤلفه jام باشد. در این صورت تجزیه طیفی ماتریس متقارن Σ_j بصورت

$$\Sigma_j = D_j B_j D_j' = \sum_{k=1}^{j} \xi_j \mathbf{e}_j \mathbf{e}_j'$$

است، که در آن $D_j=(\mathbf{e}_1,\dots,\mathbf{e}_j)$ ماتریس متعامد از بردارهای ویژه یکه و B_j ماتریس قطری از مقادیر ویژه می باشد. تجزیه طیفی فوق را می توان به صورت $\Sigma_j=\xi_j D_j A_j D_j'$ نوشت که در آن $\xi_j = |\Sigma_j|^{\frac{1}{d}}$ و A_j ماتریس قطری از مقادیر ویژه نرمال شده می باشد. پارامتر ξ_j، مشخص کننده حجم مؤلفه jام، D_j جهت آن و A_j شکل آن می باشد. فرض کنید برخی از این کمیت ها (حجم، شکل و جهت) بین خوشه ها متفاوت باشند. در این صورت می توان معیارهای ساده ای برای خوشه بندی به دست آورد که در اکثر علوم کاربرد دارند[۲]. در مدل های زیر علامت اختصاری E، بیان کننده تساوی، F نشان دهنده ثابت بودن کمیت و V (Vary) نشان دهنده متغیره بودن کمیت است.

در مدل EEE که با $\xi DAD'$ نیز نشان داده می شود حجم، شکل و جهت کلیه خوشه ها یکسان است و به راحتی ثابت می شود تخصیص بهینه $\hat{Z}$، معادل با مینیمم کردن |W| است.

در مدل VEE $(\xi DAD')$ شکل و جهت خوشه ها ثابت ولی حجم خوشه ها متغیر است. فرض کنید $\Sigma_j=\xi_j C$ که $C=DAD'$، با جایگذاری $\Sigma_j=\xi_j C$ در (۵) داریم،

$$L(\boldsymbol{\psi} \mid X) = \prod_{j=1}^{g} (\lambda_j^{n_j} (\xi_j)^{-\frac{dn_j}{2}} |C|^{-\frac{n_j}{2}}) \exp\{-\frac{1}{2}\sum_{j=1}^{g}\frac{1}{\xi_j} tr(W_j C^{-1})\}$$

بنابر این

$$\log L(\boldsymbol{\psi} \mid \mathbf{X}) \propto -\frac{d}{2}\sum_{j=1}^{g} n_j \log(\xi_j) - \frac{1}{2}\sum_{j=1}^{g}\frac{1}{\xi_j} tr(W_j C^{-1})\}$$

که در آن $\hat{\mu}_j, \hat{\lambda} = \frac{n_j}{n}$ و $\hat{\Sigma} = \frac{1}{2}\sum_{j=1}^{g} W_j$، $W_j = \sum_i (X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_j)'$، به ترتیب برآوردهای ماکسیمم درستنمایی λ_j، μ_j و Σ هستند. ماکسیمم کردن تابع درستنمایی (۵) با در نظر گرفتن مدل "VEE"هم ارز مینیمم کردن

$$F(\xi_1,\dots,\xi_g, C) = \sum_{j=1}^{g} (tr(W_j C^{-1}) + d\sum_{j=1}^{g} n_j \log(\xi_j)$$

است و نشان داده می شود که مینیمم کردن آن باید به روش تکراری (الگوریتم EM) صورت گیرد. با توجه با فرض ثابت بودن ماتریس C

مقداری از ξ_j که $F(\xi_1,\ldots,\xi_g,C)$ را مینیمم می کند عبارت است از،

$$\xi_j = \frac{tr(W_j C^{-1})}{dn_j} \qquad (۶)$$

فرض کنید ξ_j ها ثابت باشند در این صورت مینیمم کردن $F(\xi_1,\ldots,\xi_g,C)$، معادل مینیمم کردن $tr(\sum_{j=1}^{g}(\frac{1}{\xi_j}W_j)C^{-1})$ است. سیلوکس و گورت [۳] نشان دادند ماتریس متقارن d بعدی $M=\frac{Q}{|Q|^{\frac{1}{3}}}$، که در آن Q ماتریس متقارن و معین مثبت است و $|M|=1$ وجود دارد که $tr(QM^{-1})$ مینیمم می کند. با توجه به عبارت فوق ماتریسی که $tr(\sum_{j=1}^{g}(\frac{1}{\xi_j}W_j)C^{-1})$ را مینیمم کند بصورت

$$C=\frac{\sum_{j=1}^{g}(\frac{1}{\xi_j}W_j)}{|\sum_{j=1}^{g}(\frac{1}{\xi_j}W_j)|^{\frac{1}{3}}} \qquad (۷)$$

خواهد بود باتوجه به معادلات (۶) و (۷) نمی توان فرم بسته ای از یک معیار خوشه بندی ارائه نمود و بایستی از الگوریتم EM استفاده شود.

مدل EVV (مدل $\xi D_j A_j D_j'$): با قرار دادن $C_j = D_j A_j D_j'$، ماکسیمم کردن (۵) معادل با مینیمم کردن

$$F(\xi_1,C_1,\ldots,C_g)=\sum_{j=1}^{g}\frac{1}{\xi}(tr(W_j C_j^{-1})+nd\log(\xi)$$

است. بنابر این با توجه به سیلوکس و گورت [۳]، $C_j=\frac{W_j}{|W_j|^{\frac{1}{3}}}$. از طرفی ξای که تابع F را مینیمم می کند عبارت است از

$$\xi=\frac{\sum_{j=1}^{g}tr(W_J c_J^{-1})}{ND}$$

با جایگذاری C_j، در معادله فوق داریم $\xi=\frac{\sum_{j=1}^{g}(|W_j|^{\frac{1}{d}})}{nd}$. بنابر این مدل "$EVV$" ، افرازی که $\sum_{j=1}^{g}(|W_j|^{\frac{1}{d}})$ را مینیمم می کند، معادل با ماکسیمم تابع درستنمایی است.

در مدل های قبلی شکل خوشه ها چه ثابت و چه متغیر بیضوی هستند. اکنون دو مدل که شکل خوشه ها آنها کروی است، معرفی می شوند. در مدل $\Sigma_j=\xi I$ شکل خوشه ها کروی و حجم آنها با هم مساوی است. با فرض این مدل، ماکسیمم کردن (۵) هم ارز مینیمم کردن

$$F(\xi)=\frac{1}{\xi}tr(W)+nd\log\xi$$

می باشد و $\xi=\frac{tr(W)}{nd}$ تابع $F(\xi)$ را مینیمم می کند. با جایگذاری ξ در F داریم، $F(\xi)=nd+nd\log\left(\frac{tr(W)}{nd}\right)$. پس به جای مینیمم کردن F می توان $nd\log(W)$ و یا در حقیقت $tr(W)$ را مینیمم کرد، که کاربردی ترین و قدیمی ترین معیار خوشه بندی است.

در مدل $\Sigma_j=\xi_j I$ شکل خوشه ها کروی ولی حجم آنها با هم متفاوت است. در این صورت ماکسیمم کردن (۵) هم ارز مینیمم کردن

$$F(\xi_1,,\ldots,\xi_g)=\sum_{j=1}^{g}\frac{1}{\xi_j}tr(W_j)+d\sum_{j=1}^{g}n_j\log\xi_j$$

جدول ۱: معیار های خوشه بندی بر اساس مدلهای آمیخته با مؤلفه های نرمال چد متغیره است. با مشتق گیری از F نسبت به ξ_j و مساوی صفر قرار دادن آن داریم

$$\xi_j=\frac{tr(W_j)}{dn_j}$$

با جایگذاری ξ_j در F خواهیم داشت،

$$F(\xi_1,,\ldots,\xi_g)=\sum_{j=1}^{g}dn_j+d\sum_{j=1}^{g}n_j\log(\frac{tr(W_j)}{dn_j}).$$

بنابر این مینیمم کردن F هم ارز مینیمم کردن $\sum_{j=1}^{g}n_j\log\left(\frac{tr(W_j)}{n_j}\right)$ می باشد. معیارهای خوشه بندی بر اساس مدل را می توان به طور خلاصه در جدول ۲ ملاحظه نمود.

۲. انتخاب بهترین مدل با استفاده از معیار *BIC*

فرض کنید $X_1,\ldots,X_n$ یک نمونه تصادفی از توزیع آمیخته با تابع چگالی احتمال(۱) باشد. در استنباط بیزی برای مدل های آمیخته گاوسی، یک روش ساده برای انتخاب بهترین مدل و تعیین تعداد مؤلفه ها محاسبه عامل بیزی[۱] است. عامل بیزی B_{10} برای مدل M_1 در مدل M_0 به صورت

۱- Bayese Factor

$$B_{۱۰}=\frac{P(\mathbf{X}\mid M_{۱})}{P(\mathbf{X}\mid M_{۰})}$$

تعریف می شود، که در آن

(۸)

$$P(\mathbf{X}\mid M_j)=\int P(\mathbf{X}\mid \psi_j, M_j)\,P(\psi_j \mid M_j)\,d\psi_j$$

و ψ_j بردار پارامتری (λ_j,θ_j) در مدل M_j و $P(\psi_j \mid M_j)$ تابع چگالی پیشین است. بنابر این عامل بیزی شانس پسین یک مدل در برابر دیگر می باشد [۸]. هنگامی که از الگوریتم EM برای یک روش کلاسیک برای تقریب (۸) استفاده از معیار اطلاع بیزی *(BIC)* است. این تقریب عبارت است از

$$۲\log P(\mathbf{X}\mid M_j)=۲L_M(\mathbf{X}\mid\hat{\psi})-m_M\ \ln(n)=BIC(M,g).$$

که در آن $L_M(\mathbf{X}\mid\hat{\psi})$، لگاریتم تابع درستنمایی و m_M تعداد پارامترهایی است که با در نظر گرفتن مدل M می بایست برآورد شوند. بر این اساس مدلی که دارای بیشترین مقدار BIC باشد، بهترین خواهد بود. در حالیکه براساس عامل بیزی اگر $B_{۱۰}>۱$ باشد آنگاه مدل $M_{۱}$ بهترین از $M_{۰}$ است.

مثال: خوشه بندی کردن داده های گل زنبق: برای تشریح روشهای خوشه بندی بر اساس مدل از داده های گل زنبق فیشر استفاده می شود. این داده ها در ماردیا [۹] داده شده اند که شامل طول و عرض کاسبرگ ها و گلبرگ برای ۱۵۰ گل زنبق از سه نوع نوع خاردار، رنگانگ و ورجینیایی می باشند.

۲ . ۱ کاربرد الگوریتم *EM* در خوشه بندی بر اساس مدل

همانطوری که بیان شد در مرحله E الگوریتم EM ماتریس مشخصه گروهبندی، $\{Z_{ij}\}$، محاسبه می شود که در آن $\hat{Z}_{ij}$، برآورد احتمال شرطی تعلق مشاهده i ام به گروه k با انتخاب مقادیر اولیه برای پارامترها یا برآورد آنها است. سپس در «مرحله M » برآورد ماکسیمم درستنمایی پارامترها مشخص می شود.

برای این منظور از نرم افزار MCLUST استفاده شده است. براساس یک سری محاسبات با $MCLUST$ داریم $z_{۱,۱}=۱$، $z_{۱,۲}=۵.۸e^{-۰۲۲}$ و $z_{۱,۳}=۳.۱۷e^{-۰۴۲}$ همچنین $z_{۱۱۹,۳}=۱$ $z_{۱۱۹,۱}=۲.۱۷e^{-۰۵۸}$ و $z_{۱۱۹,۲}=۱.۹۶e^{-۰۰۹}$. چون $z_{۱,۱}$ از $z_{۱,۲}$ و $z_{۱,۳}$ بزرگتر است مشاهده ۱ام در خوشه اول قرار می گیرد، در حالی که مشاهده ۱۱۹ام به خوشه سوم تعلق می یابد. به همین ترتیب بر اساس z برآورد شده، در مورد خوشه اصلی سایر مشاهدات نیز می توان تصمیم گرفت. بر این اساس مشاهدات به گروهی تعلق می گیرند که دارای مقدار $\hat{Z}$ بیشتری باشند. بعد از خوشه بندی کردن داده ها، می توان برآورد ماکسیمم درستنمایی پارامترها را به دست آورد که به صورت زیر می باشند.

$$\hat{\mu}=\begin{pmatrix} ۵/۰۰۶ & ۵/۹۴۲۳۳۱ & ۶/۵۷۴۶۲۳ \\ ۳/۴۲۸ & ۲/۷۶۰۷۵۷ & ۲/۹۸۰۷۹۲ \\ ۱/۴۶۲ & ۴/۲۵۸۷۱۸ & ۵/۵۳۹۰۱۶ \\ ۰/۲۶۴ & ۱/۳۱۹۲۰۳ & ۲/۰۲۴۹۳۳ \end{pmatrix}$$

$$\hat{\Sigma}=\begin{pmatrix} ۰/۲۶۳۹۳۴۸۰ & ۰/۰۸۹۸۴۹۸۲ & ۰/۱۶۹۶۵۷۷۸ & ۰/.۰۳۹۳۳۸۰۸ \\ ۰/۰۸۹۸۴۹۸۲ & ۰/۱۱۱۹۴۷۷۶ & ۰/۰۵۱۱۲۰۷۶ & ۰/۰۲۹۹۷۸۳۴ \\ ۰/۱۶۹۶۵۷۷۸ & ۰/۰۵۱۱۲۰۷۶ & ۰/۱۸۶۵۳۴۷۶ & ۰/۰۴۱۹۷۳۳۵ \\ ۰/۰۳۹۳۳۸۰۸ & ۰/۰۲۹۹۷۸۳۴ & ۰/۰۴۱۹۷۳۳۵ & ۰/۰۳۹۷۱۱۹۵ \end{pmatrix}$$

برآورد ماکسیمم درستنمایی پارامترهای مدل آمیخته استفاده شود یک روش کلاسیک برای تقریب (۸) استفاده از معیار اطلاع بیزی *(BIC)* است. این تقریب عبارت است از

$$\hat{\lambda}=\begin{pmatrix} ۰/۳۳۳۳۳۳۳ & ۰/۳۲۹۶۱۹۱ & ۰/۳۳۷۰۴۷۵ \end{pmatrix}$$

پس از خوشه بندی کردن داده ها توسط الگوریتم EM و برآورد پارامترها، به انتخاب بهترین مدل با استفاده از معیار BIC می پردازیم. با توجه به شکل ۱ و جدول ۲ مقدار BIC برای مدل خوشه بندی *"VEV"* با تعداد دو خوشه بیشتر از سایر حالت هاست. لذا این مدل با فرض دو خوشه بهترین مدل براساس معیار BIC در بین سه مدل فوق با کلیه حالات در نظر گرفته شده برای تعداد خوشه ها می باشد.

۳. بحث و نتیجه‌گیری

در این مقاله روش خوشه بندی بر اساس مدل با مؤلفه های نرمال چند متغیره مورد بررسی قرار گرفت و نشان داده شد که این روش معادل با برآورد پارامترهای مؤلفه ها می باشد. سپس از الگوریتم EM برای برآورد پارامترها استفاده شد. همچنین بیان شد که در این روش فرضیات دلخواه اشخاص در مورد معیار تشابه نقشی ندارد و با تجزیه طیفی ماتریس کوواریانس می توان حجم ، شکل و جهت خوشه ها را به دست آورد. بعلاوه می توان با استفاده از معیار BIC بهترین مدل را تعیین نمود که هم ارز بهترین مقدار ممکن برای تعداد خوشه ها نیز می باشد.

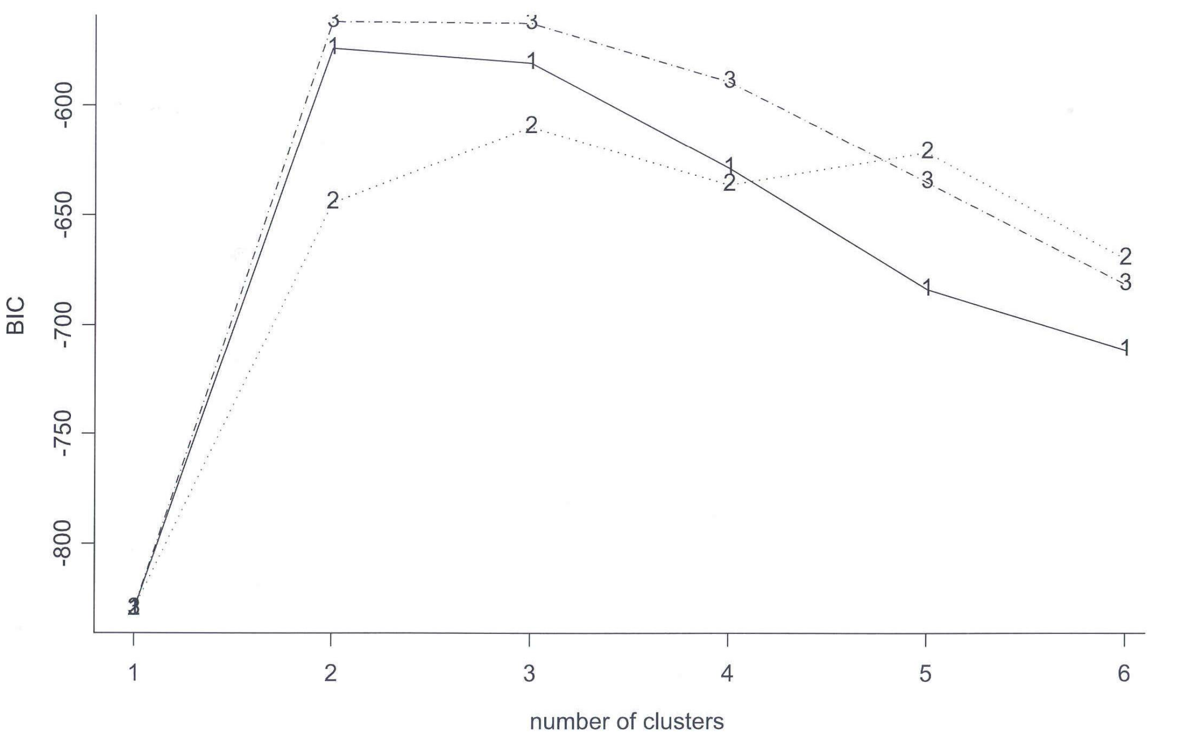
جدول شمار ه ۱: معیار خوشه بندی بر اساس مدلهای آمیخته با مولفه های نرمال چند متغیره

جدول شماره ۲: مقادیر *BIC* برای سه مدل مختلف

معیار خوشه بندی	جهت	شکل	حجم	شکل خوشه ها	مدل	علامت اختصاری
$Tr(w)$	موجود نیست	مساوی	مساوی	کروی	ξI	EI
$\sum_{j=1}^{g} n_j \log(\frac{tr(w_j)}{dn_j})$	موجود نیست	مساوی	متغیر	کروی	$\xi_j I$	VI
$\lvert W \rvert$	مساوی	مساوی	مساوی	بیضوی	$\xi DAD'$	EEE
$\sum_{j=1}^{g} n_j \log(\frac{\lvert W_j \rvert}{n_j})$	متغیر	متغیر	متغیر	بیضوی	$\xi_j D_j A_j D'_j$	VVV
از الگوریتم EM	مساوی	مساوی	متغیر	بیضوی	$\xi_j DAD'$	VEE
از الگوریتم EM	متغیر	ثابت	متغیر	بیضوی	$\xi_j D_j AD'_j$	VFV
$\sum_{j=1}^{g} \lvert W \rvert^{\frac{1}{d}}$	متغیر	متغیر	مساوی	بیضوی	$\xi D_j A_j D'_j$	EVV

مدل	۱	۲	۳	٤	٥	٦
VVV	-۸۲۹/۹۷۸۲	۵۷۴/۰۱۷۸	-۵۸۰/۸۳۸۹	-۶۲۸/۹۵۶۴	-۶۸۳/۸۱۱۴	-۷۱۱/۵۶۵۷
EEV	-۸۲۹/۹۷۸۲	۶۴۴/۵۹۹۷	-۶۱۰/۰۸۳۶	-۶۴۵/۹۹۵۰	-۶۲۱/۶۹۰۱	-۶۶۹/۷۰۶۹
VEV	-۸۹۲/۹۷۸۲	-۵۶۱/۷۲۸۵	-۵۶۲/۵۵۰۷	-۵۸۹/۳۵۱۰	-۶۳۵/۲۰۵۱	-۶۸۱/۲۹۷۶

شکل ۱: نمودار *BIC* برای سه مدل (۱): *VVV*، (۲): *EEV* و (۳): VEV



مراجع

[1] Banfield, J.D and Raftey, A. E., 1993, Model-Based Gaussian and Non-Gaussin Clustering. *Biometrics,* 49, 803-821.

[2] Bensmail, H. and Celeux, G., 1996, Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition. *J. Amer. Stat. Assoc*., 91, 1743-1748.

[3] Celeux, G. and Govert, G., 1995, Gaussian Parsimonious Clustering Models. *Pattern Recognition*, 28, 781-793.

[4] Fraly, C. and Raftery, A. E., 1998, How Many Clusters? Which Clustering Mehod? Answer via Model-Based Cluster Analysis. *Techincal Report*, No. 329. Seattle: Department of Statistics, University of Washington.

[5] Fraley, C. and Raftery, A. E., 1999, MCLUST: Software for Model-Based Cluster Analysis, *J. Classification,* 16, 297-306.

[6] Hartigan, J. A., 1975, *Clustering Algorithms*. Wiley, New York.

[7] Jeff C. F., 1983, On the Convergence of the EM Algorithm, *Annals of Statistcs*, 11, 95-103.

[8] Kass, R. E. and Raftery, A. E., 1995, Bayes Factors. J. *Amer. Stat. Assoc*., 90, 773-775.

[9] Mardia, K. V. and Kent, J. M., 1979, *Multivariate Analysis*, Bibby, Academic Press.

[10] McLachlan, G. J., 1982, The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis. In Krishnan, P. R. and Kanal, L. N. (eds), *Handbook of Statistics*, 2, 199-208. North-Holland, Amesterdam.

[11] McLachlan, G. J. and Krishnan, T., 1997, *The EM Algorithm and Extensions*. Wiley, New York.