آمیخته‌ای از مدل‌های تحلیل عاملی با مؤلفه‌های توزیع t–چندمتغیره

مرضیه مهریار - مجتبی خزائی

دانشکده علوم ریاضی، دانشگاه شهید بهشتی

چکیده: استفاده از مدل تحلیل‌گر عاملی آمیخته در طی سال‌های اخیر در تحلیل داده‌ها رواج یافته است. این مدل که در ابتدا مبتنی بر فرض نرمال بودن مؤلفه‌ها معرفی گردید، برای حصول نتایج مطلوب در حضور مشاهدات دور افتاده به حالتی که مؤلفه‌ها دارای توزیع t–چندمتغیره هستند نیز تعمیم داده شده. در این مقاله به معرفی مدل تحلیل‌گر عاملی آمیخته با مؤلفه‌های t–چندمتغیره به تاسی از کار مک‌لاکلان و همکاران (۲۰۰۷b) پرداخته و کاربرد آن را در تحلیل داده‌های ریزآرایه مربوط به سرطان روده بررسی خواهیم کرد.

واژه‌های کلیدی: مدل تحلیل‌گر عاملی، مدل تحلیل‌گر عاملی آمیخته با مؤلفه‌های t–چندمتغیره، توزیع t–چندمتغیره، الگوریتم AECM، ریزآرایه

۱ مقدمه

تحلیل عاملی یکی از روش‌های آماری است، که چنان‌چه هدف تحقیق خلاصه کردن مقدار زیاد اطلاعات و تقلیل آن‌ها به حداقل تعداد عامل‌های مشترک باشد، کاربرد دارد. در مدل تحلیل عاملی معمول فرض بر این است که مشاهده‌ها به یک جامعه همگن تعلق دارند. حال آن که این فرض در بسیاری از کاربردها برقرار نبوده و مجموعه مشاهده‌های تحت بررسی را می‌توان تلفیقی از مشاهده‌های چند زیرجامعه یا چند گروه دانست. قهرمانی و هینتون (۱۹۹۷) مدلی تحت عنوان مدل تحلیل‌گر عاملی آمیخته ارائه کردند که در شرایط فوق به کار می‌آید، این مدل توسط مک‌لاکلان و پیل (۲۰۰۰) گسترش یافت. در این مدل‌ها فرض می‌شود توزیع زیرجامعه‌ها، نرمال چندمتغیره است. این فرض باعث گردیده در مسائل با داده‌های دور افتاده در مجموعه مشاهده‌ها، نتایج به دست آمده چندان قابل اعتماد نباشد، به‌طوری‌که به نظر می‌رسد مدل‌هایی که در آن‌ها از توزیع‌های دم‌کلفت استفاده می‌کنند بهتر قادرند رفتار مشاهده‌های تحت بررسی را تبیین کنند. برای این منظور مدل تحلیل‌گر عاملی آمیخته t توسط مک‌لاکلان و همکاران (۲۰۰۷b) معرفی شد. آن‌ها این مدل را در تحلیل داده‌های ریزآرایه سرطان

سینه به‌کار بردند. در تکمیل کار این افراد در این مقاله، کار با مجموعه کل ژن‌های موجود در ریزآرایه شروع می‌شود و با انجام آزمون‌های آماری، کاهش ژن انجام می‌شود و نهایتاً از مدل تحلیل‌گر عاملی آمیخته t جهت شناسایی خوشه‌ای از ژن‌ها که بیشترین قدرت تفکیک‌پذیری را میان بافت‌های سالم و سرطانی ایجاد می‌نماید، استفاده می‌شود. برای این منظور، الگوریتم $EMMIX-GENE$ (الگوریتمی که از مدل تحلیل‌گر عاملی آمیخته نرمال برای تحلیل‌ها استفاده کرده) مورد استفاده قرار گرفته است.

۲ توزیع tی چند متغیره

گوییم بردار تصادفی p بعدی $\boldsymbol{Y}$ دارای توزیع tی p متغیره با بردار مکانی $\boldsymbol{\mu}$، ماتریس مقیاس $\boldsymbol{\Sigma}$ و درجه آزادی v است و می‌نویسیم $\boldsymbol{Y} \sim t_p(\mu, \boldsymbol{\Sigma}, v)$، هرگاه چگالی آن به صورت زیر باشد:

$$f_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, v) = \frac{\Gamma(\frac{v+p}{2})|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi v)^{\frac{1}{2}p}\Gamma(\frac{v}{2})\{1 + \frac{1}{v}\delta(\boldsymbol{y}, \boldsymbol{\mu}; \boldsymbol{\Sigma})\}^{\frac{1}{2}(v+p)}}, \quad \boldsymbol{y} \in \boldsymbol{B}R^p$$

که $\delta(\boldsymbol{y}, \boldsymbol{\mu}; \boldsymbol{\Sigma}) = (\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ مربع فاصله ماهالانوبیس میان $\boldsymbol{y}$ و $\boldsymbol{\mu}$ است. توزیع tی چند متغیره را می‌توان به صورت آمیخته‌ای (نامتناهی) از توزیع نرمال و گاما نوشت. اثبات می‌شود اگر $\boldsymbol{Y}|W = w \sim N_p(\boldsymbol{\mu}, \frac{1}{w}\boldsymbol{\Sigma})$ و $W \sim \gamma(\frac{v}{2}, \frac{v}{2})$، آن‌گاه $\boldsymbol{Y} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, v)$، که در آن، Y مشاهده می‌شود ولی W قابل مشاهده نیست. از این رو، در برآورد ماکسیمم درستنمایی پارامترهای توزیع tی چندمتغیره در حضور مقادیر گمشده w_jها از الگوریتم EM استفاده می‌شود. در این حالت به سادگی می‌توان نشان داد که

$$w_j^{(k)} = E_{\Psi(k)}(W_j|\boldsymbol{y}_j) = \frac{v + p}{v + \delta(\boldsymbol{y}_j, \boldsymbol{\mu}^{(k)}; \boldsymbol{\Sigma}^{(k)})}, \qquad (۱)$$

رابطه (۱) نشان می‌دهد $w_j^{(k)}$ها یا وزن‌های کوچک متناظر با $\delta(\boldsymbol{y}_j, \boldsymbol{\mu}^{(k)}; \boldsymbol{\Sigma}^{(k)})$ بزرگ بوده و بیان‌گر این است که مشاهده $\boldsymbol{y}_j$ از $\boldsymbol{\mu}^{(k)}$ دور است. به عبارت دیگر نشان از دور افتاده بودن مشاهده $\boldsymbol{y}_j$ دارد.

۳ مدل تحلیل‌گر عاملی آمیخته t

فرض کنید که $y_1, \ldots, y_n$ بردارهای تصادفی p بعدی از جامعه Π باشند که $\prod$ نیز ترکیبی از $g \geq 2$ زیر جامعه $\Pi_1, \ldots, \Pi_g$ است و $f_1, \ldots, f_g$ تابع چگالی توزیع t−چند

متغیره بردار تصادفی y، در زیرجامعه‌های $\Pi_1, \ldots, \Pi_g$ است. مدل تحلیل‌گرهای عاملی آمیخته با مؤلفه‌های t–چندمتغیره به شکل زیر بیان می‌شود:

$$f(y_j;\psi) = \sum_{i=1}^{g} \pi_i f(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\nu_i), \qquad (2)$$

که در آن

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i\boldsymbol{B}_i' + \boldsymbol{D}_i,$$

و بردار پارامتری $\boldsymbol{\Psi}$ متشکل از درجه‌های آزادی مؤلفه‌ها، ν_iها، نسبت‌های آمیختگی، π_iها، عناصر $\boldsymbol{\mu}_i$، $\boldsymbol{B}_i$ و $\boldsymbol{D}_i$ می‌باشد.

مدل تحلیل‌گر عاملی آمیخته، آمیخته‌ی متناهی از مدل‌های خطی است و در کل مدلی غیرخطی در نظر گرفته می‌شود. این مدل علاوه بر این‌که یک شیوه کاهش بعد موضعی خطی، و در کل ناخطی است، در خوشه‌بندی داده‌ها شامل ساختارهای گروه‌بندی نیز مفید است.

مشابه با مدل تحلیل عاملی معمول، مدل آمیخته (۲) را به شکل (۳) می‌توان تعریف کرد. برای این منظور فرض می‌کنیم $\boldsymbol{Z}_j$ یک بردار تصادفی g بعدی باشد، که iامین مؤلفه آن Z_{ij}، به صورت زیر تعریف می‌شود:

$$Z_{ij} = \begin{cases} 1 & y_j \in \Pi_i \\ 0 & y_j \notin \Pi_i \end{cases}$$

و احتمال اینکه Z_{ij} مقدار ۱ را اختیار کند برابر است با π_i. بر این اساس توزیع $\boldsymbol{Z}_j$، توزیع چندجمله‌ای زیر است:

$$\boldsymbol{Z}_j \sim MB_g(1, (\pi_1, \ldots, \pi_g)').$$

تحت شرایط فوق، به طور معادل می‌توان گفت که مشاهده y_j با احتمال π_i به صورت زیر است:

$$\boldsymbol{y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i\boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad i = 1, \ldots, g \qquad (3)$$

مشروط بر این‌که

$$\begin{aligned} \boldsymbol{U}_{ij}|z_{ij} = 1 &\sim t_q(0, \boldsymbol{I}_q, \nu_i), \\ \boldsymbol{e}_{ij}|z_{ij} = 1 &\sim t_q(0, \boldsymbol{D}_q, \nu_i), \\ cov[(\boldsymbol{e}_{ij}, \boldsymbol{U}ij)|z_{ij} = 1] &= 0. \end{aligned}$$

به عبارت دیگر، توزیع خطاها و عامل‌ها، توزیع t–چندمتغیره‌ای با درجه‌های آزادی یکسان و ناهمبسته‌اند.

۴ برآورد پارامترهای مدل تحلیل‌گر عاملی

در این بخش با چگونگی برآورد ماکسیمم درستنمایی پارامترهای مدل تحلیل‌گر عاملی آمیخته t آشنا می‌شویم، با توجه به نامعلوم بودن $\mathbf{Z}_j$ها و $\mathbf{U}_{ij}$ها(و البته W_jها) می‌توان از الگوریتم EM استفاده کرد، که الگوریتم EM یکی از ابزارهای متداول برای یافتن برآورد ماکسیمم درستنمایی، در حضور مقادیر گمشده است. در استفاده از الگوریتم EM، ایده‌آل این است که نقطه ماکسیمم در گام M، به سادگی و توسط روابط بسته‌ای قابل حصول باشند. اما متاسفانه در بعضی از موقعیت‌ها این‌گونه نیست و این نقاط را باید به روش‌های عددی محاسبه کرد که در جای خود می‌توانند پیچیده بوده و به‌خصوص سرعت همگرایی را پایین آورند. در مواردی که اطلاع از مقدار بعضی از پارامترها کار ماکسیمم‌سازی پارامترهای دیگر ر بسیارساده‌ترمی‌کند، استفاده از الگوریتم ECM توسط منگ و روبین (۱۹۹۳) پیشنهاد شد. در مواردی مثل این‌جا که علاوه بر این، مقادیر گمشده ازانواع مختلفی است، روش $AECM$[۱] که توسط منگ و ون‌دیک (۱۹۹۷) ارائه شد، برای محاسبه برآورد پارامترهای مدل تحلیل‌گر عاملی t استفاده می‌شود.

در ادامه با جزئیات الگوریتم $AECM$ برای برآورد بردار پارامترهای نامعلوم، $\boldsymbol{\Psi}$، آشنا می‌شویم. در ابتدا بردار $\boldsymbol{\Psi}$ به دو بخش $(\boldsymbol{\Psi}_1', \boldsymbol{\Psi}_2')'$ افراز می‌شود که در آن $\boldsymbol{\Psi}_1$ شامل عناصر $\boldsymbol{\mu}_i$ها، نسبت‌های آمیختگی π_iها و درجه‌های آزادی ν_iها و $\boldsymbol{\Psi}_2$ شامل عناصر $\boldsymbol{B}_i$ها و $\boldsymbol{D}_i$هااست. هر تکرار الگوریتم $AECM$ از دو چرخه تشکیل شده است، که هر چرخه شامل دو مرحله E و M است:

چرخه اول:

مرحله E: امید ریاضی لگاریتم درستنمایی داده‌های کامل، به شرط داده‌های مشاهده شده، به ازای برآورد پارامترها، $\boldsymbol{\Psi}^{(k)}$، در تکرار $(k+1)$ام به صورت زیر تعریف می‌شود:

$$Q_1(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \sum_{j=1}^{n} \sum_{i=1}^{g} ln[\pi_i f_p(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i)]_{\boldsymbol{\Psi}^{(k)}}(Z_{ij}|\boldsymbol{y}_j) \qquad (۴)$$

برای محاسبه امید شرطی (۴) کافیست $E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij}|\boldsymbol{y}_j)$ محاسبه شود.

$$E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij}|\boldsymbol{y}) = \tau_{ij}^{(k)} = \frac{\pi_i^{(k)} f(\boldsymbol{y}_j; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}, \nu_i^{(k)})}{f(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)})}$$

[۱] Alternating Expectation Conditional Maximization

مرحله M: شامل ماکسیمم کردن تابع $Q_1(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ نسبت به پارامترهای $\boldsymbol{\pi}_i$، $\boldsymbol{\mu}_i$ و ν_i است. برآورد $\boldsymbol{\pi}_i$ و $\boldsymbol{\mu}_i$ به صورت زیر برآورد خواهد شد.

$$\pi_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)}}{n} \quad , \quad \boldsymbol{\mu}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)} \boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)}}$$

برای تعیین برآورد ν_i نیاز به حل معادله زیر است.

$$\Big\{\psi(\frac{\nu_i}{2}) + log\frac{\nu_i}{2} + 1 + \frac{1}{n_i^{(k)}} \sum_{j=1}^{n} \tau_{ij}^{(k)} (log\, w_{ij}^{(k)} - w_{ij}^{(k)}) + \psi(\frac{\nu_i^{(k)} + p}{2}) + log(\frac{\nu_i^{(k)} + p}{2})\Big\} = 0 \qquad (5)$$

که در آن $n_i^{(k)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}$. معادله (۵) جواب صریحی نداشته و باید با به‌کارگیری روش‌های عددی مانند روش نیوتن-رافسون برآورد ν_i را به‌دست آورد.

چرخه دوم:

در چرخه دوم الگوریتم $AECM$ داده‌های کامل، y_c به صورت زیر تعریف می‌شود.

$$\boldsymbol{y}_c = (\boldsymbol{y}', z_1', \ldots, z_n', \boldsymbol{u}_1', \ldots, \boldsymbol{u}_n', w_1, \ldots, w_n)'$$

مرحله E: امید ریاضی لگاریتم درستنمایی داده‌های کامل به شرط مشاهده‌ها به ازای برآورد جاری پارامترها، $Q_2(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k+\frac{1}{2})})$، محاسبه می‌گردد.

$$Q_2(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k+\frac{1}{2})}) = \sum_{j=1}^{n}\sum_{i=1}^{g} ln[\pi_i f_p(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\nu_i)] E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij}|\boldsymbol{y}_j)$$

مرحله M: شامل ماکسیمم کردن تابع $Q_2(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k+\frac{1}{2})})$ به ازای $\boldsymbol{B}_i$ و $\boldsymbol{D}_i$ است. برآورد $\boldsymbol{B}_i$ و $\boldsymbol{D}_i$ در تکرار $(K+1)$ به‌ترتیب به‌صورت زیر برآورد خواهد شد.

$$\boldsymbol{B}_i^{(k+1)} = \boldsymbol{V}_i^{(k+\frac{1}{2})} \gamma_i^{(k)} \Big(\gamma_i^{(k)\prime} \boldsymbol{V}_i^{(k+\frac{1}{2})} \gamma_i^{(k)} + \boldsymbol{\Omega}_i^{(k)}\Big)^{-1},$$
$$\boldsymbol{D}_i^{(k+1)} = diag\Big\{\boldsymbol{V}_i^{(k+\frac{1}{2})} - \boldsymbol{V}_i^{(k+\frac{1}{2})} \gamma_i^{(k)} \boldsymbol{B}_i^{(k+1)\prime}\Big\}$$

که

$$\boldsymbol{V}_i^{(k+\frac{1}{2})} = \frac{\sum_{j=1}^{n} w_{ij}^{(k+\frac{1}{2})} \tau_{ij}^{(k+\frac{1}{2})} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})'}{\sum_{j=1}^{n} \tau_{ij}^{(k+\frac{1}{2})}},$$
$$\gamma_i = (\boldsymbol{B}_i \boldsymbol{B}_i' + \boldsymbol{D}_i)^{-1} \boldsymbol{B}_i \quad , \quad \boldsymbol{\Omega}_i = \boldsymbol{I}_q - \gamma_i' \boldsymbol{B}_i$$

چرخه‌ی اول و دوم تا همگرایی الگوریتم تکرار می‌شوند.
بر اساس آنچه به آن پرداخته شد در مدل تحلیل‌گر عاملی t فرض کردیم که تعداد مؤلفه‌ها و تعداد عامل‌ها معلوم هستند، اما متاسفانه در عمل تعداد مؤلفه‌ها و تعداد عامل‌ها مشخص نیستند. لذا در ادامه، در انتخاب یک مدل مناسب از میان یک مجموعه متناهی از مدل‌ها، از ملاک BIC استفاده می‌شود. ملاک اطلاع بیزی BIC که توسط شوارتز (۱۹۷۸) ارائه شد، عبارت است از:

$$BIC = -۲log\ l(y,\widehat{\boldsymbol{\theta}}) + klog\ n$$

که در آن $log\ l(y,\widehat{\boldsymbol{\theta}})$ مقدار لگاریتم تابع درستنمایی در نقطه ماکسیمم و k تعداد پارامتر های مدل است. مطابق تعریف فوق، مدلی که BIC کمتری داشته باشد مدل بهتری خواهد بود.
روش خوشه‌بندی، روشی برای شناسایی گروه‌های مشابه از اشیا در بین داده‌هاست که با استفاده از آن، حجم وسیعی از داده‌ها، به خوبی سازماندهی و خلاصه می‌شوند، به طوری که هر کدام از خوشه‌ها شامل مشاهده‌های مشابه بوده و همچنین هر خوشه تا جای ممکن متمایز از دیگر خوشه‌ها باشد. برای خوشه‌بندی داده‌ها روش‌های زیادی وجود دارد، از این‌رو جهت مقایسه‌ی روش‌های مختلف خوشه‌بندی دو شاخص رند و رند تعدیل‌شده در این مقاله معرفی خواهند شد.
یک روش عمومی برای ارزیابی نتایج خوشه‌بندی یک روش معین، مقایسه خوشه‌بندی به‌دست آمده از این روش با یک خوشه‌بندی درست و از پیش تعریف شده است. یک معیار معمول برای چنین مقایسه‌ای، شاخص رند است که اولین بار توسط رند (۱۹۷۱) معرفی شد. فرض کنید $C = \{C_۱,\ldots,C_s\}$ ساختار خوشه‌بندی حاصل از یک روش خوشه‌بندی معین و $P = \{P_۱,\ldots,P_t\}$ خوشه‌بندی درست و از پیش تعریف‌شده باشد. شاخص رند تطابق بین دو ساختار خوشه‌بندی P و C را به‌صورت زیر اندازه می‌گیرد.

$$RI(P,C) = \frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{n}{۲}}$$

که در آن a تعداد جفت مشاهده‌هایی است که متعلق به خوشه‌های مشابهی در P و همچنین به خوشه‌های مشابهی در C تخصیص داده شده‌اند، d جفت مشاهده‌هایی که متعلق به خوشه‌های متفاوت در P و همچنین خوشه‌های متفاوت در C هستند، b جفت مشاهده‌هایی که متعلق به خوشه‌های متفاوتی در P اما در خوشه‌های مشابهی در C قرار گرفته‌اند، c تعداد جفت مشاهده‌هایی که متعلق به خوشه‌های مشابهی در P هستند ولی در خوشه‌های متفاوتی در C قرار گرفته‌اند و n تعداد کل مشاهدات است.

این شاخص مقادیر بین صفر و یک می‌پذیرد و هرچه مقدار آن به یک نزدیک‌تر باشد، تطابق بین دو ساختار خوشه‌بندی بیشتر است.
موردی که به عنوان نقطه ضعف شاخص رند تلقی می‌شود این است که، مقدار متوسط این شاخص نامشخص است و با تغییر ساختار خوشه‌بندی، مقدار متوسط این شاخص تغییر می‌کند. برای اجتناب از این مشکل هوبرت و عربی (۱۹۸۵) شاخص رند تعدیل شده، ARI[۲]، را ارائه دادند و آن را به شکل زیر تعریف کردند.

$$ARI = \frac{\frac{a+d}{\binom{n}{2}} - \frac{(a+b)(a+c)+(c+d)(b+d)}{\binom{n}{2}^2}}{1 - \frac{(a+b)(a+c)+(c+d)(b+d)}{\binom{n}{2}^2}}$$

این شاخص مقادیر کمتر از یک را با مقدار متوسط ثابت صفر می‌پذیرد و رفتاری مشابه رفتار شاخص رند دارد به این معنا که مقادیر بزرگ شاخص رند تعدیل شده بیان‌گر تطابق بیشتر بین دو ساختار خوشه‌بندی است. برای جزئیات بیشتر می‌توانید به یونگ و روزو (۲۰۰۰) مراجعه نمایید.

۵ تحلیل داده‌های ریزآرایه

فناوری استفاده از ریزآرایه درنگرش بشر، در مسائل زیستی، تحول عظیمی ایجاد نموده است. این فناوری پژوهشگران را قادر می‌سازد به‌جای بررسی تک به تک ژن‌ها، به طور موازی چندین هزار ژن را مورد بررسی قرار دهند تا عملکرد و نوع تعامل آن‌ها با یکدیگر بهتر مشخص شود. در اکثر داده‌های ریزآرایه، تعداد ژن‌ها از مرتبه چند هزار و چند ده هزار است، اما تعداد نمونه‌ی بافت‌ها به ندرت از ۱۰۰ مورد تجاوز می‌کند. به همین دلیل استفاده از تکنیک‌ها و روش‌های انتخاب ژن و کاهش بعد امری اجتناب ناپذیر می‌باشد. هدف خوشه‌بندی، گروه‌بندی ژن‌ها و نمونه‌ها با ویژگی‌های شبیه هم است. این مسأله می‌تواند به عنوان کاهش ابعاد سیستم نیز در نظر گرفته شود.
در این مقاله به معرفی روش $EMMIX-GENE$ بر روی مجموعه داده‌های ریزآرایه سرطان روده بزرگ پرداخته خواهد شد. برای کسب اطلاعات به مک‌لاکلان و همکاران (۲۰۰۲) و مک‌لاکلان و همکاران (۲۰۰۷a) مراجعه نمایید. روش $EMMIX-GENE$ شامل سه مرحله زیر است:

مرحله ۱ : تشخیص و حذف ژن‌های غیر مرتبط

برای کشف ژن‌های غیر مرتبط، هر ژن را به صورت جداگانه بررسی، و این که آیا ژن مربوطه قادر به تشخیص دو مؤلفه متناظر با نمونه‌های بافتی سالم و سرطانی هست یا

[۲] Adjusted Rand Index

نه را با آزمون $H_0 : G = 1$ در مقابل $H_1 : G = 2$ بررسی می‌کنیم. در این روش جهت کاستن حساسیت آزمون نسبت به نقاط دورافتاده، توزیع مؤلفه‌های آمیخته t، در نظر گرفته می‌شود. در انتخاب یک ژن، شرط‌های زیر باید برقرار باشد:

$$-2Ln\lambda > b_1 \qquad (۶)$$

و

$$min\{N_1, N_2\} \geq b_2, \qquad (۷)$$

که N_1 و N_2 تعداد بافت‌هایی است که به مؤلفه‌ها تخصیص داده شده‌اند و نقاط آستانه b_1 و b_2، از قبل توسط محقق تعیین می‌شود. اگر شرط (۷) برقرار ولی شرط (۸) برقرار نباشد، یکی از خوشه‌ها شامل تعداد کمی مشاهده و در واقع دورافتاده‌ها است. در این حالت ممکن است مؤلفه با تعداد مشاهده‌های بیشتر خود شامل دو زیرمؤلفه باشد. برای این‌که شانس مجددی برای شناسایی چنین ژن‌هایی وجود داشته باشد، آزمون $H_0 : G = 2$ در مقابل $H_1 : G = 3$ انجام می‌شود. شرط (۷) به همراه این شرط که حداقل حجم دو مؤلفه از سه مؤلفه، حداقل b_2 است، در نظر گرفته می‌شود. اگر هر دو شرط برقرار باشد، ژن انتخاب و در غیر این صورت کنار گذاشته می‌شود.

مرحله ۲ : خوشه‌بندی ژن‌ها

منظور از یک خوشه‌ی ژنی، گروهی از ژن‌ها هستند که در میان همه‌ی ژن‌ها بردار بیان شبیه و نزدیک به هم دارند. ژن‌هایی که دارای بردار بیان شبیه و نزدیک به هم هستند، با یکدیگر همبستگی داشته و این همبستگی باعث می‌شود که قدرت تمییز یکسانی در دسته‌بندی نمونه‌ها داشته باشند. با این انگیزه، ژن‌های باقی‌مانده از مرحله اول با استفاده از الگوریتم K–میانگین خوشه‌بندی می‌شوند. تعداد خوشه‌های مدنظر توسط محقق تعیین می‌گردد. این تعداد باید طوری انتخاب شود که علاوه بر کم بودن مجموع تغییرپذیری‌های درون خوشه‌ها، متوسط تعداد ژن‌ها در هر خوشه خیلی بزرگ نباشد (از مشاهده‌ها کمتر باشد).

مرحله ۳ : خوشه‌بندی بافت‌ها

در این مقاله، بخش آخر الگوریتم را به مدل تحلیل‌گر عاملی آمیخته t، تعمیم داده و در هر خوشه‌ی ژنی مرحله قبل سعی می‌کنیم بعد فضای ویژگی را کوچک‌تر و افراد را خوشه‌بندی نماییم. به منظور تعیین تعداد عامل‌ها در هر خوشه ژنی، معیار BIC مورد استفاده قرار می‌گیرد. در نهایت از میان خوشه‌های ژنی، خوشه‌ای که رند و رند تعدیل شده بزرگ‌تری داشته باشد انتخاب می‌شود.

۶ $EMMIX-GENE$ در تحلیل ریزآرایه سرطان روده بزرگ

در این بخش به تحلیل یک مجموعه داده بیان ژنی ریزآرایه مربوط به سرطان روده بزرگ حاوی ۲۰۰۰ ژن از ۶۲ بافت نمونه خواهیم پرداخت. که ۴۰ نمونه از بافت‌های سرطانی بیماران و ۲۲ نمونه از بافت‌های سالم بیماران هستند. این داده‌ها، از داده‌های موجود در سایت «$http://www.maths.uq.edu.au/gim/emmix-gene$» است.

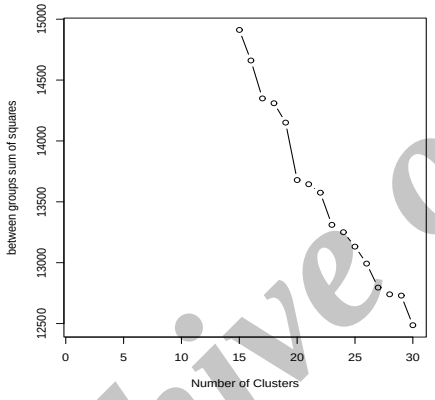
جدول ۱: مقدار معیار RI و ARI در هر خوشه‌ی ژنی

شماره خوشه	q	ARI	RI	شماره خوشه	q	ARI	RI
۱	۵	۰/۰۷	۰/۵۶	۱۱	۶	۰/۰۷	۰/۵۴
۲	۳	۰/۰۶	۰/۵۶	۱۲	۷	۰/۰۳	۰/۵۲
۳	۳	۰/۱۶	۰/۵۸	۱۳	۵	۰/۰۷	۰/۵۶
۴	۷	۰/۰۳	۰/۵۵	۱۴	۴	۰/۰۹	۰/۵۷
۵	۳	−۰/۰۳	۰/۴۹	۱۵	۳	۰/۰۹	۰/۵۷
۶	۶	۰/۱۱	۰/۵۶	۱۶	۴	۰/۰۸	۰/۵۵
۷	۴	۰/۰۸	۰/۵۵	۱۷	۵	۰/۱۱	۰/۵۶
۸	۶	۰/۰۲	۰/۵۲	۱۸	۹	۰/۰۱	۰/۵۰
۹	۸	۰/۰۴	۰/۵۴	۱۹	۴	−۰/۰۲	۰/۴۹
۱۰	۸	۰/۰۴	۰/۵۴	۲۰	۹	۰/۰۸	۰/۵۵

در اجرای روش $EMMIX-GENE$ در مرحله اول، $b_1 = b_2 = ۸$ در نظر می‌گیریم، ۱۵۴۷ ژن به عنوان ژن‌های نامرتبط حذف می‌شود و ۴۵۳ ژن باقی می‌ماند. پس از این مرحله، ژن‌های ریزآرایه را با استفاده از الگوریتم k–میانگین در k خوشه قرار خواهیم داد. جهت تعیین تعداد خوشه‌ها، مجموع تغییرپذیری‌های درون خوشه‌ای را به ازای $k = ۱۵,\ldots,۳۰$ محاسبه کرده و در شکل ۱ رسم کرده‌ایم. همان‌طور که ملاحظه می‌شود با افزایش k مجموعه تغییرپذیری‌های درون خوشه‌ای همواره کاهش می‌یابد، اما چند کاهش زیاد به‌خصوص از ۱۹ به ۲۰ خوشه اشاره‌ای به این است که تعداد ۲۰ خوشه می‌تواند انتخاب مناسبی باشد. در ادامه با اجرای روش تحلیل‌گر عاملی آمیخته t با $G = ۲$ و $q = ۳,\ldots,۹$ در هر خوشه، بافت‌ها را به دو گروه تقسیم می‌کنیم. برای ارزیابی نتایج خوشه‌بندی در هر خوشه از دو ملاک رند و رند تعدیل شده استفاده می‌کنیم که نتایج آن در جدول ۲ آمده است. مطابق جدول ۲، خوشه سوم بهترین نتیجه را در میان خوشه‌ها داراست. این خوشه در میان بافت‌های سرطانی ۱۴ نمونه و در میان بافت‌های سالم ۴ نمونه به اشتباه تخصیص داده شده‌اند. هرچند هنوز نرخ خطا نسبتاً بالا است، ولی مقایسه روش‌های فوق با روش‌های معمول خوشه‌بندی سسلسله مراتبی، k–میانگین و مدل–پایه که در مک‌نیکلاس و مورفی (۲۰۱۰) ارائه شده است، نشان از برتری این روش دارد (جدول ۲).

۷ بحث و نتیجه‌گیری

مسأله‌ای که محققین در کاربردهای علمی با آن مواجه می‌شوند، بعد بالای مشاهده‌ها است. از این‌رو، مدل‌های تحلیل‌گرعاملی آمیخته دارای کاربردهای وسیعی درزمینه‌های مختلف علوم کامپیوتر، پزشکی و ژنتیک می‌باشد. یکی از حوزه‌هایی که در آن کمتر از این تکنیک استفاده شده، تحلیل داده‌های ریزآرایه است، که در این مقاله این روش را در تحلیل ریزآرایه‌های سرطان روده بزرگ استفاده کردیم. روش‌های آماری تحلیل داده‌های ریزآرایه، بدلیل تعداد متغیرهای بسیار زیاد و مشاهده‌های کم، هنوز به عملکرد خیلی رضایت‌بخشی منجر نشده‌اند. مک‌نیکلاس و مورفی (۲۰۱۰) تکنیک‌های دیگری را در تحلیل این داده‌ها به کار گرفتند و روش‌های مختلف خوشه‌بندی را بر اساس شاخص‌های رند و رند تعدیل شده مقایسه نمودند. نتایج به دست آمده توسط آن‌ها نشان می‌دهد، روش‌های معمول خوشه‌بندی سلسله‌مراتبی، k–میانگین و مدل–پایه نتوانسته‌اند نتایج بهتری از مدل تحلیل‌گر عاملی آمیخته t با $ARI = ۰/۱۶$ و $RI = ۰/۵۸$ کسب کنند.



شکل ۱.: مجموع تغییرپذیری‌های درون خوشه‌ای به ازای تعداد خوشه‌های متفاوت

جدول ۲: شاخص RI و ARI حاصل از روش‌های معمول خوشه‌بندی داده‌های ریزآرایه سرطان روده بزرگ

روش	ARI	RI
سلسله‌مراتبی کامل	-۰/۰۲	۰/۵۰
سلسله‌مراتبی میانگین	-۰/۰۱	۰/۵۳
سلسله‌مراتبی منفرد	-۰/۰۱	۰/۵۳
k–میانگین	-۰/۰۲	۰/۵۰
مدل–پایه با ماتریس واریانس–کوواریانس VII	-۰/۰۱	۰/۵۰

مراجع

Ghahramani, Z. and Hinton, G. E. (1997), The EM algorithm for factor analyzers, *Technical Report NO. CRG-TR-96-1.Toronto: The University of Toronto*.

Hubert, L. and Arabie, P. (1985), Comparing partitions, *Journal of classification*, **2**, 193-218.

McLachlan, G. J. and Bean, R. and Wen, S. (2007a), Application of gene shaving and mixture models to cluster microarray gene expression data, *Cancer Inform*, **2**, 25-43.

McLachlan, G. J. and Bean, R. W. and Ben-Tovim Jones, L. (2007b), Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution, *Computational Statistics and Data Analysis*, **51**, 5327-5338.

McLachlan, G. J. and peel, D. (2000), Mixtures of factor analyzers, *Proceedings of the Seventeenth International Conference on Machine Learning*.

McLachlan, G. J. and Peel, D. and Bean, R. W. (2002), A mixture model-based approach to the clustering of microarray expression data, *Bioinformatics*, **18**, 413-422.

McNicholas, P. D. and Murphy, T. B. (2010), model-based clustering of microarray expression data via latent Gaussian mixture models, *Bioinformatics*, **21**, 2705-2712.

Meng, X. L. and Rubin, D. B. (1993), Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika*, **80**, 267-278.

Meng, X. L. and van Dyk, D. (1997), The EM algorithm-an old folk song sung to a fast new tune (with discussion), *J. Roy. Soc. B*, **59**, 511-567.

Rand, W. M. (1971), Objective criteria for the evaluation of clustering methods, *J. Amer. Statist. Assoc*, **66**, 846-850.

Schwartz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, **6**, 31-38.

Yeung, K. Y. and Ruzzo, W. L. (2000), An empirical study on principal component analysis for clustering gene expression data, *Tech. Report, University of Washington*, **17**, 763-774.