مجله علوم پزشکی رازی دوره ۲۲، شماره ۱۳۶، مهر ۱۳۹۴ http://rjms.iums.ac.ir **مقاله پژوهشی**

خوشه‌بندی داده‌های بیان‌ژنی توسط عدم تشابه جنگل تصادفی

* **زهره فرهادی**: کارشناس ارشد آمار، دانشگاه شاهرود، شاهرود، ایران (*نویسنده مسئول). zohreh.farhadi87@gmail.com
داود شاهسونی: عضو هیات علمی گروه آمار، دانشگاه شاهرود، شاهرود، ایران. dshahsavani@shahroodut.ac.ir

تاریخ دریافت: ۹۴/۳/۱۸ تاریخ پذیرش: ۹۴/۶/۲۵

چکیده

زمینه و هدف: خوشه‌بندی داده‌های بیان ژنی در تشخیص و درمان سرطان، دارای اهمیت بسزایی است. مشخصه‌ی بارز این داده‌ها تعداد زیاد متغیرها (ژن‌ها) نسبت به تعداد داده‌ها (بیماران) است. بسیاری از روش‌های خوشه‌بندی بر پایه‌ی عدم تشابه داده‌ها که حاصل محاسبه‌ی یک تابع فاصله است، بنا شده‌اند و افزایش بعد، کارآیی توابع فاصله را کاهش می‌دهد. در این تحقیق معیاری جدید برای محاسبه‌ی عدم تشابه در ابعاد بالا، بر اساس یک روش رده‌بندی به نام *جنگل تصادفی* معرفی شده و کارایی آن در تحلیل داده‌های بیان ژنی، مورد ارزیابی قرار گرفته است.

روش کار: در این مقاله خوشه‌بندی مجموعه داده‌ی چاودری و همکاران توسط عدم تشابه جنگل تصادفی مد نظر قرار گرفته است. بدین منظور ابتدا مسئله‌ی خوشه‌بندی به مسئله‌ی رده‌بندی تبدیل شده و با انجام رده‌بندی جنگل تصادفی، عدم تشابه مربوطه محاسبه شده است. سرانجام داده‌ها توسط روش خوشه‌بندی افراز حول مدوید، خوشه‌بندی شده و نتیجه‌ی خوشه‌بندی توسط شاخص رند تعدیل یافته مورد ارزیابی قرار گرفته است. تمامی تحلیل‌ها با نرم‌افزار R انجام شده است.

یافته‌ها: مقدار شاخص رند تعدیل یافته (۸۱۴۹/۰)، نشان‌دهنده‌ی انطباق مطلوب خوشه‌های تخمینی با گروه‌های واقعی است. همچنین با استفاده از قابلیت تعیین اهمیت متغیرها در روش جنگل تصادفی، ژن شماره‌ی ۳۱ مؤثرترین ژن در این خوشه‌بندی شناخته شد و توانستیم خوشه‌های تخمینی را تنها به‌وسیله‌ی این ژن توصیف کنیم.

نتیجه‌گیری: عدم تشابه جنگل تصادفی، معیاری کارا برای سنجش عدم تشابه داده‌ها در خوشه‌بندی داده‌های بیان ژنی است. همچنین می‌توان با استفاده از قابلیت منحصر به‌فرد این روش، ژن‌های مؤثر در خوشه‌بندی را شناسایی نموده و خوشه‌های تخمینی را به‌وسیله‌ی آن‌ها توصیف نمود.

کلیدواژه‌ها: خوشه‌بندی، داده‌های بیان ژنی، عدم تشابه جنگل تصادفی، تعیین اهمیت متغیرها

مقدمه

بخش گسترده‌ای از تحقیقات پزشکی با اهداف شناسایی انواع سرطان، درمان و پیشگیری از آن‌ها صورت می‌گیرد. از آنجا که علل ژنتیکی نقش مهمی در ایجاد سرطان دارند، محققان به‌منظور رسیدن به اهداف ذکر شده به مطالعه و بررسی عملکرد ژن‌ها که با استخراج اطلاعات درون آن‌ها میسر می‌شود، روی آورده‌اند. در سال‌های اخیر با ظهور فناوری به نام ریزآرایه (DNA(Microarray، امکان بررسی و مطالعه فعالیت هزاران ژن به‌طور همزمان فراهم آمده است. استفاده از فناوری ریزآرایه DNA حجم انبوهی از داده‌ها را موسوم به داده‌های بیان ژنی (Gene Expression Data) تولید می‌کند که مشخصه بارز آن‌ها تعداد بسیار زیاد متغیرها (ژن‌ها) نسبت به تعداد داده‌ها (بیماران) است. تجزیه و تحلیل داده‌های حاصل از فناوری ریزآرایه DNA درهای جدیدی را در تشخیص زودهنگام سرطان، کشف انواع جدید سرطان، توسعه و بهبود روش‌های درمانی به روی محققان گشوده است و از این حیث، توجه بسیاری از آن‌ها بر روی تجزیه و تحلیل این‌گونه داده‌ها معطوف گشته است به‌طوری‌که امروزه یکی از بیشترین کاربردهای خوشه‌بندی مربوط به تجزیه و تحلیل داده‌های بیان ژنی در علوم پزشکی است.

خوشه‌بندی یکی از روش‌های پرکاربرد آماری است که به تفکیک مجموعه‌ی داده‌ها به زیرمجموعه‌های کوچک‌تری به نام خوشه می‌پردازد به‌طوری‌که اعضای هر خوشه، مشابه یکدیگر بوده و کمترین میزان تشابه را با اعضای دیگر خوشه‌ها داشته باشند. در کنار خوشه‌بندی،

نوع دیگری از گروه‌بندی به نام رده‌بندی (Classification) وجود دارد که در آن، گروه‌ها (رده‌ها) معلوم هستند و هدف، مدل‌بندی آن‌ها و یافتن قاعده‌ای برای تخصیص داده‌های جدید به گروه‌های موجود است.

بسیاری از روش‌های خوشه‌بندی بر پایه‌ی عدم تشابه داده‌ها که حاصل محاسبه‌ی یک تابع فاصله است، بنا شده‌اند. فاصله‌ی اقلیدسی یکی از مرسوم‌ترین توابع فاصله‌ای است که در روش‌های خوشه‌بندی مورد استفاده قرار می‌گیرد. در برخی از کاربردها از جمله تجزیه و تحلیل داده‌های بیان ژنی، محققان با داده‌هایی مواجه هستند که تعداد متغیرهای توضیحی در آن‌ها بسیار زیاد است. از آنجا که افزایش تعداد ابعاد، کارآیی توابع فاصله را کاهش می‌دهد، خوشه‌بندی در ابعاد بالا، به یکی از مهم‌ترین چالش‌ها در این زمینه تبدیل گشته است (۱).

تعریف معیاری متفاوت برای سنجش عدم تشابه که وابسته به توابع فاصله نباشد، می‌تواند به‌عنوان راه‌حلی برای چالش مذکور مد نظر قرار گیرد و در بهبود عملکرد خوشه‌بندی مؤثر واقع شود. بریمن و کاتلر (Breiman and Catler) در سال ۲۰۰۳ عدم تشابه جدیدی را بر مبنای روش رده‌بندی *جنگل تصادفی*(Random Forest) معرفی کردند که با دیدگاهی کاملاً متفاوت تعریف می‌شود. با تبدیل مسئله‌ی خوشه‌بندی به مسئله‌ی رده‌بندی، می‌توان از این نوع عدم تشابه در خوشه‌بندی بهره گرفت. جنگل تصادفی روشی مناسب و کارا در مورد مجموعه داده‌هایی است که در آن‌ها تعداد متغیرها بسیار زیاد بوده و حتی بیشتر از تعداد داده‌ها است. همچنین گزینه مناسبی در مواجهه با داده‌های دورافتاده و نوفه‌ها می‌باشد و عدم تشابه حاصل از آن می‌تواند توانایی روش‌های خوشه‌بندی را در برخورد با این‌گونه چالش‌ها افزایش دهد. عدم تشابه جنگل تصادفی را می‌توان در مورد انواع متغیرها بکار برد و در صورت وجود داده‌هایی با مقادیر گمشده محاسبه نمود. این عدم تشابه وابسته به مقادیر متغیرها نیست و بکارگیری آن در روش‌های خوشه‌بندی ما را از انجام هرگونه پیش‌پردازشی روی مجموعه داده‌ها بی‌نیاز می‌کند.

همچنین از آنجا که روش جنگل تصادفی قابلیت منحصربه‌فرد تعیین متغیرهای مهم در رده‌بندی را دارا است، لذا می‌توان متغیرهای مؤثر در خوشه‌بندی را شناسایی نموده و خوشه‌های تخمین‌شده بوسیله‌ی عدم تشابه مذکور را توسط قاعده‌ای ساده توصیف کرد.

عدم تشابه جنگل تصادفی، تاکنون در بسیاری از پژوهش‌های کاربردی مورد استفاده قرار گرفته است. از جمله پژوهش‌های انجام شده در این زمینه می‌توان به تحقیقات بریمن و کاتلر، کی (Qi) و همکاران، شی (Shi) و همکاران، شی و هرود (Horvath) و همچنین چن و اسفرن (Chen and Ishveran) اشاره کرد که در آن‌ها از عدم تشابه جنگل تصادفی به‌منظور خوشه‌بندی داده‌های ژنومی، ریزآرایه‌ها و داده‌های نشانگر تومور (Tumor Marker) استفاده شده است (۲-۶).

در این مقاله قصد داریم ضمن معرفی عدم تشابه جنگل تصادفی، کارآیی آن را در تحلیل مجموعه داده‌ای بیان ژنی مورد ارزیابی قرار دهیم. در داده‌های بیان ژنی، می‌توان خوشه‌بندی را صرفاً برای متغیرها یا صرفاً برای داده‌ها انجام داد که در این تحقیق، فقط خوشه‌بندی داده‌ها مدنظر قرار گرفته است. فرض کنید تعداد p ژن از n داده(بیمار) مورد بررسی قرار گرفته است. ماتریس حاوی داده‌های بیان ژنی را به‌صورت $X=[x_{ij}]$ نمایش می‌دهیم که $1\le i\le n$ و $1\le j\le p$ و x_{ij} بیانگر سطح بیان ژن j از داده‌ی i است. همچنین عدم تشابه دو به دوی داده‌ها را که در قالب یک ماتریس متقارن نمایش داده می‌شود به‌صورت $D=[d_{ij}]$ در نظر می‌گیریم که در آیه‌ی ijام آن عدم تشابه بین دو داده‌ی i و j را نشان می‌دهد.

روش کار

داده‌های تحقیق: در این تحقیق از مجموعه داده‌ی بیان ژنی به نام چاودری(Chowdary) استفاده شده است که از طریق درگاه زیر دسترس عموم قرار دارد.

http://bioinformatics.rutgers.edu/Static/Supplements/CompCancer/datasets.htm

چاودری و همکارانش در سال ۲۰۰۶، تعداد ۱۰۴ بیمار سرطانی را مورد مطالعه قرار دادند که از این تعداد ۶۲ نفر به سرطان سینه و ۴۲ نفر به سرطان روده‌ی بزرگ مبتلا بودند. در این مطالعه، تعداد ۱۸۲ ژن از بیماران مذکور با هدف شناخت ساختار ژن‌های مسئول این دو سرطان، مورد بررسی قرار گرفت (۷).

الگوریتم خوشه‌بندی افراز حول مدوید: روش خوشه‌بندی افراز حول مدوید (Partition around medoid یاPAM) در زمره‌ی روش‌های خوشه‌بندی افرازی قرار دارد. در الگوریتم PAM برای ایجاد k خوشه، ابتدا k داده به‌عنوان نماینده‌ی خوشه‌ها انتخاب می‌شوند. هر نماینده به گونه‌ای انتخاب می‌شود که دارای بیشترین تشابه با سایر اعضای خوشه باشد. این داده که مدوید نامیده می‌شود، در مرکزی‌ترین نقطه‌ی خوشه واقع است. پس از مشخص شدن مدویدها، سایر داده‌ها به خوشه‌ای نسبت داده می‌شوند که تشابه بیشتری با مدوید آن دارند و بدین ترتیب خوشه‌های اولیه شکل می‌گیرند. سپس در یک فرآیند تکراری به‌منظور یافتن مجموعه مدویدی که بهترین خوشه‌بندی را در پی داشته باشد، تمامی داده‌ها را برای جانشینی با مدویدها آزمایش می‌کند (۸، ۹).

جنگل تصادفی: روش رده‌بندی جنگل تصادفی، مجموعه‌ای از درختان تصمیم است. در رده‌بندی درخت تصمیم، فضای p-بعدی متغیرها، به‌طور سلسله مراتبی به زیرفضاهای‌های کوچکتر - p بعدی تقسیم می‌شود به‌طوری‌که داده‌های واقع در هر ناحیه دارای حداکثر تجانس و همگونی باشند. این الگوی رده‌بندی، توسط ساختاری موسوم به درخت تصمیم انجام می‌گیرد. در این ساختار درختی، نقطه‌ی انشعاب به دو زیر شاخه، گره نامیده می‌شود. اولین گره‌ی درخت را ریشه نامیده و آخرین گره‌ها را با عنوان برگ می‌شناسند. همزمان با تقسیم فضای متغیرهای توضیحی به دو زیرفضای کوچکتر، یکی از گره‌ها به دو گره داخلی تقسیم می‌شود و بدین ترتیب درخت تصمیم رشد می‌کند. در هر مرحله از افراز فضا به‌منظور یافتن بهترین نوع تقسیم، همه‌ی متغیرها و همه مقادیر ممکن آن‌ها، مورد جستجو قرار می‌گیرند تا فضای متغیرها به دو قسمت مناسب تقسیم شود (۱۰).

تشکیل درختان تصمیم در جنگل تصادفی با درخت کلاسیک تصمیم دارای تفاوت‌هایی است. در جنگل تصادفی هر درخت با یک نمونه خودگردان از داده‌های اصلی رشد می‌کند و به‌منظور انجام بهترین تقسیم فضا، تعداد m متغیر که به تصادف از بین متغیرها انتخاب شده‌اند، مورد جستجو قرار می‌گیرند. تعداد درختان و مقدار m که آنها را به ترتیب با $mtry$ و $ntree$ نشان می‌دهیم، می‌بایست توسط کاربر تعیین و بهینه شوند. هرچه تعداد درختان جنگل تصادفی بیشتر باشد، پیش‌بینی از دقت بالاتری برخوردار است (۱۱)، بنابراین پارامتر $ntree$ باید به قدر کافی بزرگ انتخاب شود. در مورد پارامتر $mtry$ معمولاً مقدار $\sqrt{p}$ پیشنهاد می‌شود که p تعداد متغیرها است (۱۰، ۱۲، ۱۳).

در روش جنگل تصادفی عدم تشابه داده‌ها به طریقی کاملاً متفاوت از توابع فاصله‌ی متداول تعیین می‌گردد. تشابه داده‌ها در این روش بر مبنای قرار گرفتن آن‌ها در برگ‌های یکسان (زیرفضاهای نهایی) اندازه‌گیری می‌شود. در جنگل تصادفی، تشابه بین دو داده‌ی i و j $(s(i,j))$، نسبت تعداد دفعاتی تعریف می‌شود که دو داده‌ی مذکوردر یک برگ قرار گرفته‌اند. ماتریس تشابه جنگل تصادفی، متقارن و معین مثبت است و هر درآیه آن در بازه‌ی [۰،۱] قرار دارد. این ماتریس با انجام تبدیل زیر به یک ماتریس عدم تشابه تبدیل می‌شود:

$$d(i,j)=\sqrt{1-s(i,j)}$$

تشکیل درخت رده‌بندی وابسته به مقادیر متغیرها نیست و از این رو عدم تشابه جنگل تصادفی را می‌توان در مورد انواع متغیرها(پیوسته، دودویی، رسته‌ای، اسمی، ترتیبی و...) بکار برد (۵).

یکی از قابلیت‌های ویژه و مفید روش جنگل تصادفی تعیین اهمیت متغیرها است، بدین معنی که این روش قادر است متغیرهایی با بیشترین اهمیت در رده‌بندی را شناسایی نماید. از آنجا که

متغیرهای با اهمیت نقش عمده‌ای در رده‌بندی جنگل تصادفی دارند، در نتیجه در تعیین میزان تشابه داده‌ها نیز دارای تاثیر بیشتری هستند.

خوشه‌بندی جنگل تصادفی: خوشه‌بندی جنگل تصادفی یک نام کلی برای روش‌های خوشه‌بندی مبتنی بر ماتریس عدم تشابه است که این ماتریس برگرفته از اعمال روش جنگل تصادفی است. بدست آوردن عدم تشابه جنگل تصادفی، جز با اجرای رده‌بندی به این روش میسر نیست. به‌منظور اجرای رده‌بندی جنگل تصادفی روی مسئله خوشه‌بندی که در آن رده‌ای برای داده‌ها تعریف نشده است، می‌بایست مسئله‌ی خوشه‌بندی را به یک مسئله‌ی رده‌بندی تبدیل کرد. جهت انجام این تبدیل می‌بایست تمامی داده‌ها را متعلق به یک رده در نظر گرفت و رده‌ای دیگر را هم منتسب به داده‌هایی مصنوعی دانست که با حجمی برابر داده‌های اصلی شبیه‌سازی می‌شوند. پس از تولید داده‌های مصنوعی و الحاق آن‌ها به مجموعه داده، داده‌های اصلی با رده‌ی صفر و داده‌های مصنوعی با رده‌ی ۱برچسب‌گذاری می‌شوند. بدین ترتیب مسئله‌ی "خوشه‌بندی" به مسئله‌ی "رده‌بندی" تبدیل خواهد شد. سپس می‌توان با اجرای رده‌بندی جنگل تصادفی عدم تشابه داده‌های اصلی را مطابق آنچه در بخش قبل‌توضیح داده شد، محاسبه کرد.

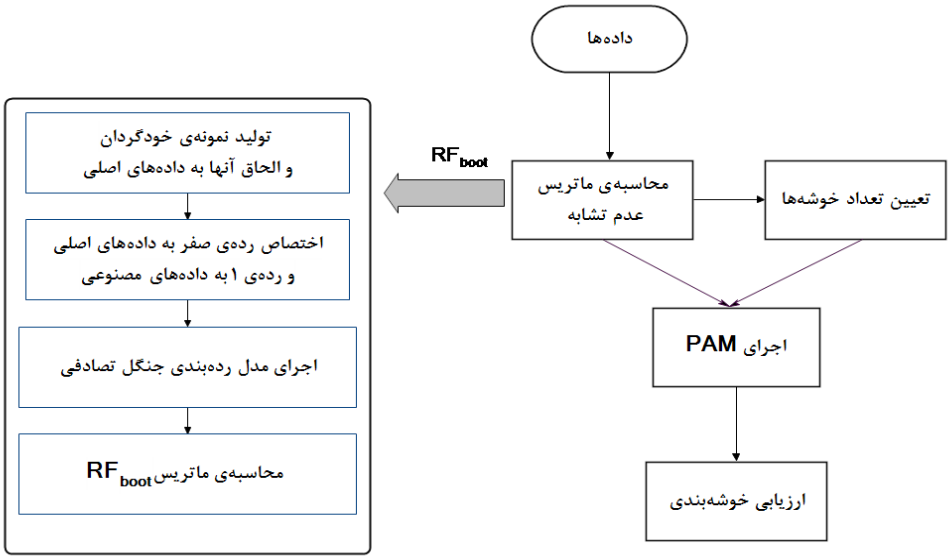
برای تولید داده‌های مصنوعی مصداقی از هر متغیر مورد نیاز است. بدین منظور می‌توان از دو روش زیر استفاده کرد (۵، ۱۴):

- تولید نمونه‌ای خودگردان از مقادیر هر متغیر در داده‌های اصلی. ماتریس عدم تشابه جنگل تصادفی حاصل از این روش را RF_{boot} می‌نامیم.
- تولید نمونه‌ای با توزیع یکنواخت در بازه‌ی مینیمم و ماکزیمم مقدار هر متغیر. ماتریس عدم تشابه جنگل تصادفی را که از این روش به‌دست می‌آید، RF_{uni} نامگذاری می‌کنیم.

در اجراهای متعدد رده‌بندی جنگل تصادفی، حتی با فرض در نظرگرفتن پارامترهای یکسان، مدل‌های رده‌بندی متفاوتی خواهیم داشت و به دنبال آن ماتریس‌های عدم تشابه مختلفی بدست می‌آید. لذا بهتر است چندین بار رده‌بندی جنگل تصادفی اجرا گردد و از عدم تشابه حاصل از آن‌ها میانگین گرفته شود. بدین ترتیب از طریق روش جنگل تصادفی برای هر مجموعه داده می‌توان دو نوع ماتریس عدم تشابه بدست آورد و هر یک از این دو ماتریس را به‌طور مجزا در روش‌های خوشه‌بندی مبتنی بر عدم تشابه بکار برد و کارایی آن‌ها را مورد بررسی قرار داد.

تعیین تعداد خوشه‌ها: تعداد خوشه‌های مناسب را می‌توان توسط میانگین شاخص نیمرخ (Silhouette index) برآورد کرد که این شاخص در اصل یکی از معیارهای ارزیابی خوشه‌بندی است. میانگین شاخص نیمرخ می‌تواند مقداری در بازه‌ی [۱،۱-] را اختیار کند. اگر میانگین شاخص نیمرخ نزدیک به عدد ۱ باشد آنگاه مدل خوشه‌بندی رضایت بخش تلقی می‌شود. مقادیر منفی و نزدیک به صفر این شاخص حاکی از نامناسب بودن مدل و عملکرد ضعیف الگوریتم خوشه‌بندی در ایجاد خوشه‌ها است. برای برآورد تعداد خوشه‌ها توسط این شاخص، خوشه‌بندی را به ازای تعداد خوشه‌های مختلف اجرا نموده و میانگین شاخص نیمرخ را برای هر حالت محاسبه می‌کنیم. تعداد مناسب خوشه‌ها برای انجام خوشه‌بندی، مقداری است که منجر به بیشترین میانگین شاخص نیمرخ گردد (۹، ۱۵، ۱۶).

ارزیابی عملکرد روش خوشه‌بندی: شاخص رند تعدیل‌یافته (Adjusted rand index) یا ARI معیاری برای ارزیابی عملکرد روش خوشه‌بندی است که در صورت اطلاع از گروه‌بندی واقعی داده‌ها می‌توان آن استفاده کرد. این شاخص، خوشه‌های حاصل از اجرای روش خوشه‌بندی را با گروه‌های واقعی مقایسه نموده و میزان تطابق آن‌ها را می سنجد. شاخص رند تعدیل یافته همواره مقداری در بازه‌ی [۱،۱-] را کسب می‌کند. زمانی‌که خوشه‌های تخمینی به‌طور کامل بر گروه‌های واقعی منطبق باشند، مقدار این شاخص برابر با ۱ است و مقادیر صفر و منفی مربوط به حالتی است که داده‌ها به‌طور تصادفی به خوشه‌ها تعلق گرفته باشند (۱۷).



شکل ۱- فلوچارت الگوریتم تحقیق مبتنی بر عدم تشابه RF_{boot}. برای اجرای خوشه‌بندی بر اساس RF_{uni} در اولین مرحله از محاسبه‌ی ماتریس عدم تشابه، نمونه‌ی یکنواخت جایگزین نمونه‌ی خودگردان می‌شود.

یافته‌ها

در یک نگاه اجمالی می‌توان فلوچارت الگوریتم تحقیق را به‌صورتی که در شکل ۱ نمایش داده شده است، در نظر گرفت. به‌منظور بدست آورن عدم تشابه جنگل تصادفی برای داده‌های (بیماران) مجموعه‌ی چاودری، داده‌های مصنوعی به دو روش پیشنهادی (نمونه خودگردان و توزیع یکنواخت) تولید گردید و به داده‌های اصلی الحاق شدند. سپس با اختصاص رده‌ی صفر به داده‌های اصلی و رده‌ی ۱ به داده‌های مصنوعی، مسئله‌ی خوشه‌بندی به مسئله‌ی رده‌بندی تبدیل شد. به‌منظور اجرای رده‌بندی جنگل تصادفی پارامترهای *mtry* و *ntree*، به ترتیب برابر با $p=$ ۱۳ $\sqrt{}$ و ۵۰۰ در نظر گرفته شد. برای رسیدن به پایداری نسبی، رده‌بندی جنگل تصادفی به تعداد ۵۰ بار اجرا گردید و سپس میانگین مقادیر عدم تشابه حاصل شده در هر اجرا محاسبه شد. بدین ترتیب، دو ماتریس عدم تشابه RF_{boot} و RF_{uni} مبتنی بر تولید داده‌های مصنوعی توسط نمونه‌ی خودگردان و توزیع یکنواخت حاصل شدند.

به‌منظور برآورد تعداد خوشه‌ها توسط شاخص نیمرخ، خوشه‌بندی PAM به ازای تعداد خوشه‌های مختلف $(k=2,3,\ldots,10)$ اجرا گردید و مقدار شاخص نیمرخ برای هریک ازخوشه‌بندی‌های صورت‌گرفته، محاسبه شد. سپس تعداد خوشه‌ای که منجر به بیشینه شدن مقدار این شاخص گردید، به‌عنوان تعداد بهینه‌ی خوشه‌ها در نظر گرفته شد. در مورد این مجموعه داده، تعداد خوشه‌ها برای هر یک از دو ماتریس عدم تشابه مذکور، برابر با عدد ۲ برآورد شده که برآوردی بدون خطا است.

پس از محاسبه‌ی ماتریس‌های عدم تشابه و برآورد تعداد خوشه‌ها، الگوریتم PAM به‌طور مجزا برای هر یک از دو ماتریس اجرا گردید و در نهایت خوشه‌های حاصل از الگوریتم PAM توسط شاخص رند تعدیل یافته مورد ارزیابی قرار گرفت. مقادیر این شاخص در جدول ۱ حاکی از تفاوت معنی دار بین دو نوع تولید نمونه های مصنوعی برای انجام عمل رده بندی جنگل تصادفی است.

توصیف خوشه‌های تخمینی موضوع دیگری است که در این تحقیق مد نظر قرار گرفته است، بدین معنا که قاعده‌ای متشکل از ژن‌ها را برای توصیف خوشه‌ها بیابیم. واقعیت آن است که در ایجاد خوشه‌ها، تمامی ژن‌ها نقش ندارند بلکه زیرمجموعه‌ای از آن‌ها در خوشه‌بندی مؤثر هستند

جدول ۱- مقادیر شاخص رند تعدیل یافته

عدم تشابه	شاخص ARI
RF_{boot}	۰/۸۱۴۹
RF_{uni}	۰/۰۲۹۶

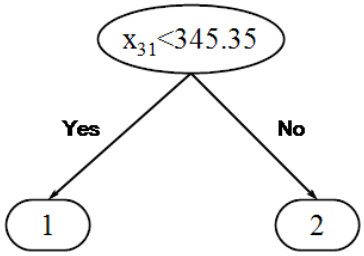
جدول ۲- ژن‌های با اهمیت در رده‌بندی جنگل تصادفی بر اساس تولید داده‌ی مصنوعی توسط نمونه‌ی خودگردان

اولویت	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
متغیر	x_6	x_{143}	x_7	x_{31}	x_{116}	x_{109}	x_{169}	x_{74}	x_{171}	x_{55}

که تعیین آن‌ها، خوشه‌ها را تفسیر پذیر نموده بطوری که می‌توان عوامل اصلی بیماری را از بین تعداد انبوه ژن‌ها تمییز داد.

همانطور که پیش از این نیز اشاره شد، در روش جنگل تصادفی، متغیرهای با اهمیت در رده‌بندی نقش موثرتری در عدم تشابه حاصل از روش مذکور خواهند داشت. با توجه به تاثیر غیر قابل انکار معیارهای عدم تشابه در نتیجه‌ی خوشه‌بندی، انتظار می‌رود که متغیرهای مؤثر در محاسبه‌ی عدم تشابه RF_{boot} نیز در مسئله‌ی خوشه‌بندی تاثیر بیشتری نسبت به سایر متغیرها داشته باشند و خوشه‌های تعیین‌شده به آن‌ها وابسته شوند. اکنون میزان صحت این ادعا را در خوشه‌بندی انجام‌شده، مورد بررسی قرار دهیم. بدین منظور ابتدا به شناسایی ژن‌های با اهمیت در رده‌بندی جنگل تصادفی می‌پردازیم. از بین ۱۸۲ ژن، تعداد ده ژن که دارای بیشترین میزان اهمیت هستند، به ترتیب اولویت در جدول ۲ درج شده‌اند.

به منظور توصیف خوشه‌های تخمینی بوسیله‌ی ژن‌های با اهمیت مذکور، مجموعه داده‌ی چاودری را تنها با این ژن‌ها در نظر گرفته و از سایرین صرف نظر می‌کنیم. اکنون اگر خوشه‌ی تخمین‌شده برای هر داده را به‌عنوان مقدار متغیر پاسخ برای آن در نظر بگیریم، یافتن قاعده‌ای برای توصیف خوشه‌های تخمینی معادل با یک مسئله‌ی رده‌بندی است. هرچه نرخ خطا در این رده‌بندی کمتر باشد توصیف بهتری از خوشه‌ها توسط



شکل ۲- ساختار درختی رده‌بندی مجموعه داده‌ی چاودری با در نظر گرفتن ده ژن با اهمیت به عنوان متغیرهای توضیحی و خوشه‌های تخمینی به عنوان متغیر پاسخ

ژن‌های با اهمیت صورت گرفته است. در اینجا از روش رده‌بندی درخت تصمیم برای یافتن مدل رده‌بندی داده‌های مذکور استفاده شده‌است. ساختار درختی رده‌بندی صورت‌گرفته در شکل ۲ نمایش داده شده‌است.

بر اساس ساختار فوق، خوشه‌های تخمینی را تنها می‌توان بوسیله‌ی ژن شماره‌ی ۳۱ بیان نمود، بدین ترتیب که اگر سطح بیان ژن داده‌ای بزرگتر از مقدار ۳۵/۳۴۵ باشد، به خوشه‌ی ۱ تعلق دارد و در غیر اینصورت متعلق به خوشه‌ی ۲ است. نرخ خطای این رده‌بندی برابر با ۰۱۹۲/۰ و تعداد خطاها برابر با ۲ گزارش شد؛ یعنی فقط تعداد اندکی از داده‌ها (دو داده) وجود دارند که خوشه‌ی تخمینی آن‌ها با این قاعده قابل توصیف نیست که این تعداد خطا قابل چشم‌پوشی است. بنابراین می‌توان نتیجه گرفت که قاعده‌ی تعیین شده توسط رده‌بندی درخت تصمیم به خوبی توانسته است خوشه‌ی تخمینی داده‌ها را بر اساس ژن‌های با اهمیت توصیف کند.

بحث و نتیجه‌گیری

خوشه‌بندی داده‌های بیان ژنی به دلیل اهمیت در تشخیص و درمان سرطان، از زمینه‌های مستعد تحقیق در حوزه‌ی علوم پزشکی است. تاکنون رهیافت‌های متنوعی توسط پژوهشگران برای خوشه‌بندی این نوع داده‌ها، مورد ارزیابی قرار گرفته است (۱۸-۲۰). در میان انواع مختلف این رهیافت‌ها، دو رهیافت سلسله مراتبی و افرازی به دلیل بازدهی مطلوبی که نسبت به سادگی ساز و کار و سهولت اجرا دارند، بیش از سایرین مورد استفاده قرار می‌گیرند. هر دو رهیافت مذکور بر پایه‌ی عدم تشابه داده‌ها که اغلب، حاصل محاسبه‌ی یک تابع فاصله است، عمل می‌کنند. در روش‌های سلسله مراتبی، داده‌ها بر اساس میزان تشابهی که با یکدیگر دارند، طی یک فرآیند سلسله مراتبی، در هم ادغام شده یا از هم جدا

می‌شوند. مشکل اصلی روش‌های مذکور این است که اگر در یک مرحله، داده‌ای به اشتباه به یک خوشه اختصاص داده شود، در مراحل بعدی نمی‌تواند به خوشه‌ی دیگری منتقل شود. یک راهکار حل این مشکل استفاده از روش‌های افرازی است که در آن به داده‌ها اجازه داده می‌شود تا به‌منظور بهبود کیفیت خوشه‌بندی، در هر مرحله از اجرای الگوریتم مربوطه، از خوشه‌ای به خوشه‌ی دیگر منتقل شوند؛ اما روش‌های افرازی دارای ضعف‌ها و اشکالاتی نیز هستند که از جمله مهم‌ترین آنها می‌توان به ضرورت تعیین تعداد خوشه‌ها قبل از شروع خوشه‌بندی اشاره کرد.

روش خوشه‌بندی k- میانگین یکی از روش‌های مرسوم رهیافت افرازی است که در آن میانگین مقادیر اعضای خوشه به‌عنوان نماینده خوشه انتخاب می‌شود. از آنجا که میانگین به شدت به نقاط دورافتاده حساس است، در صورت وجود چنین داده‌هایی و اختصاص آنها به یک خوشه، میانگین به شدت تغییر می‌کند و خوشه‌ها نیز دستخوش تغییراتی خواهند شد. الگوریتم PAM که در این تحقیق از آن به‌عنوان الگوریتم اصلی خوشه‌بندی استفاده شده است، یک روش جایگزین برای الگوریتم k-میانگین است که به جای میانگین، مرکزی‌ترین عضو خوشه را به‌عنوان نماینده انتخاب می‌کند و نسبت به داده‌های دور افتاده مقاوم است (۲۱). با وجود توضیحات ذکر شده باید به این نکته توجه داشت که هر روش خوشه‌بندی ممکن است در کاربردهای مختلف، عملکرد کاملاً متفاوتی را از خود بروز دهد و هیچ الگوریتمی وجود ندارد که به‌طور فراگیر بر سایر الگوریتم‌ها ترجیح داده شود. در کنار تنوعی که در مورد روش‌های خوشه‌بندی وجود دارد، گزینه‌های متفاوتی هم برای سنجش عدم تشابه داده‌ها وجود دارند. از آنجا که عدم تشابه داده‌ها تاثیری بنیادین در نتیجه‌ی خوشه‌بندی دارد، لازم است در مورد نحوه‌ی سنجش عدم تشابه داده‌ها دقت کافی به عمل آید.

در خصوص داده‌های بیان ژنی مورد استفاده در این تحقیق (چاودری) می‌توان اذعان داشت که تحقیقات متعددی در زمینه‌ی خوشه‌بندی آن صورت گرفته و کماکان ادامه دارد. به‌عنوان نمونه می‌توان به پژوهش سوتو (Souto) و همکاران اشاره کرد که آنها الگوریتم k- میانگین را به همراه پنج روش سلسله مراتبی بر روی سی و پنج مجموعه داده‌ی بیان ژنی از جمله مجموعه‌ی چاودری اجرا کردند. نتایج این پژوهش حاکی از عملکرد ضعیف روشهای سلسه‌مراتبی نسبت به روش k- میانگین بود (۲۲). همچنین جاسکوویک (Jaskowiak) و همکاران، پانزده معیار عدم تشابه متفاوت را برای خوشه‌بندی سی و پنج مجموعه داده مذکور، مورد استفاده قرار دادند. نتایج نشان داد که نمی‌توان به‌طور کلی در خوشه‌بندی این داده‌ها، یک معیار سنجش عدم تشابه را بر سایرین ترجیح داد و عملکرد معیار سنجش عدم تشابه، به شدت به داده‌ها و الگوریتم خوشه‌بندی حساس است (۲۳). گاسپارووییکا (Gasparovica) و همکاران نیز از روش خوشه‌بندی فازی برای مجموعه داده چاودری استفاده کردند (۲۴).

در این مقاله، برای استفاده از روش PAM، معیاری جدید برای سنجش عدم تشابه داده‌ها بر مبنای روش رده‌بندی جنگل تصادفی معرفی شد. انتخاب نوع داده‌های مصنوعی، تنظیم پارامترهای روش جنگل تصادفی و انتخاب روش تخمین تعداد خوشه‌ها از جمله موارد مهم اعمال شده هستند که در زیر مورد بحث واقع شده‌اند.

در اینجا پس از تولید دو نوع داده مصنوعی بر اساس روش‌های نمونه خودگردان و توزیع یکنواخت، مسئله‌ی خوشه‌بندی را به مسئله‌ی رده‌بندی تبدیل نموده و برای هر یک از این دو نوع داده مصنوعی، با اجرای روش رده‌بندی جنگل تصادفی، محاسبه میزان عدم تشابه جدید بین داده‌ها و استفاده از آنها در الگوریتم PAM، کارایی هر یک از این دو نوع داده را به‌طور مجزا در خوشه‌بندی مجموعه داده‌ی چاودری مورد ارزیابی قرار دادیم. یافته‌های مندرج درجدول ۱ نشان داد که خوشه‌بندی مبتنی برعدم تشابه RF_{boot}، تطابق مطلوبی با گروه‌بندی واقعی داده‌ها داشته و خوشه‌های مناسبی را نتیجه داده‌است در حالی‌که تولید داده‌های مصنوعی بر اساس توزیع یکنواخت نتوانسته است منجر به خوشه‌بندی مناسبی شود.

باید توجه داشت که ممکن است در مورد داده‌ای دیگر، عدم تشابه RF_{uni} عملکرد بهتری داشته باشد. همچنین ذکر این نکته ضروری است که در محاسبه‌ی عدم تشابه جنگل تصادفی، انجام هیچ‌گونه پیش پردازشی روی داده‌ها لازم نیست اما محاسبه‌ی این عدم تشابه جدید، زمان اجرای بیشتری را نسبت به سایر معیارهای عدم تشابه مانند فاصله اقلیدسی می‌طلبد.

انتخاب مقادیر پارامترهای جنگل تصادفی موضوع مهم دیگری است که به شدت بر نتیجه‌ی خوشه‌بندی تاثیرگذار است و انتخاب نامناسب این پارامترها می‌تواند کیفیت خوشه‌بندی را به‌طور قابل ملاحظه‌ای کاهش دهد. در این مقاله مقدار پارامتر *mtry* برابر با $\sqrt{p}$ در نظر گرفته شد که مقدار پیشنهادی برای این پارامتر در پژوهش‌های انجام شده در زمینه‌ی روش جنگل تصادفی است. همچنین مقدار ۵۰۰ را برای پارامتر *ntree* انتخاب کردیم که مقدار بزرگی به حساب می‌آید. شایان ذکر است که در تنظیم پارامترهای مذکور، لزوماً مقادیری از پارامترهای *ntree* و *mtry* که دقت رده‌بندی را افزایش می‌دهند، منجر به خوشه‌بندی بهینه نخواهند شد. بنابراین انتخاب مقادیر دو پارامتر بر اساس کاهش نرخ خطای رده‌بندی قابل اعتماد نیست و این انتخاب می‌بایست با هدف بهبود کیفیت خوشه‌بندی صورت گیرد. در اجراهای متعدد رده‌بندی جنگل تصادفی، حتی با فرض در نظرگرفتن پارامترهای یکسان، مدل‌های رده‌بندی متفاوتی خواهیم داشت و به دنبال آن ماتریس‌های عدم تشابه مختلفی به‌دست می‌آید، به همین علت برای رسیدن به پایداری، رده‌بندی جنگل تصادفی به تعداد ۵۰ بار اجرا گردید و از عدم تشابه حاصل از آن‌ها میانگین گرفته شد.

در اجرای روشهای افرازی مانند PAM، تعیین تعداد خوشه‌ها برای شروع خوشه‌بندی، مسئله‌ای بسیار مهم و دشوار است و روش‌های متعددی برای حل آن وجود دارد. در این مقاله برای برآورد تعداد خوشه‌ها، خوشه‌بندی را به ازای تعداد خوشه‌های مختلف اجرا نموده و مقدار میانگین شاخص نیمرخ را در هر حالت محاسبه کردیم. نتایج نشان داد که بیشترین مقدار شاخص نیمرخ یا به عبارتی بهترین خوشه‌بندی به ازای $k=2$ حاصل می‌شود که برابر با تعداد گروه‌های واقعی در مجموعه داده‌ی مورد بررسی است. رویه‌ای مشابه بهینه‌سازی تعداد خوشه‌ها از طریق میانگین شاخص نیمرخ را می‌توان برای یافتن مقادیر بهینه پارامترهای جنگل تصادفی، استفاده نمود، یعنی به ازای مقادیر مختلف از پارامترهای *mtry* و *ntree*، رده‌بندی جنگل تصادفی را اجرا نموده و عدم تشابه حاصل از آن را در خوشه‌بندی اعمال کرد. پس از محاسبه‌ی میانگین شاخص نیمرخ متناظر با هر خوشه‌بندی، مقادیر بهینه‌ی دو پارامتر مذکور، مقادیری هستند که به ازای آن‌ها میانگین شاخص نیمرخ ماکزیمم شده باشد.

همچنین توسط قابلیت منحصربه‌فرد روش جنگل تصادفی در خصوص تعیین اهمیت متغیرها توانستیم ژن‌های مؤثر در خوشه‌بندی را شناسایی نموده و خوشه‌های تخمینی را بوسیله قاعده‌ای ساده، قابل توصیف کنیم که تاکنون به این موضوع در پژوهش‌های مشابه پرداخته نشده بود. طبق یافته‌های حاصل از تحقیق ژن شماره‌ی ۳۱ موثرترین ژن در خوشه‌بندی شناخته شد و توسط این ژن خوشه‌های ۱ و ۲ به ترتیب به‌صورت $x_{31} < 345.35$ و $x_{31} > 345.35$ توصیف شدند.

بطور کلی و با توجه به نتایج به‌دست آمده می‌توان نتیجه گرفت که عدم تشابه جنگل تصادفی می‌تواند انتخاب مناسبی برای تجزیه و تحلیل داده‌های بیان ژنی باشد .از آنجا که روش‌های خوشه‌بندی مختلف ممکن است به نتایج بسیار متفاوتی منجر شود، لذا می‌توان به‌عنوان پیشنهادی برای آینده‌ی تحقیق، کارآیی عدم تشابه جنگل تصادفی برای تحلیل داده‌های بیان ژنی را در سایر روش‌های مبتنی بر عدم تشابه بررسی نمود. بدین منظور می‌توان از انواع روش‌های سلسله مراتبی بهره گرفت.

منابع

1. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is nearest neighbor meaningful. in 7th Int.

Conf. Database Theary; 1999.

2. Catler A, Breiman L. Random forests manual v4.0. tech.rep, UC Berkeley; 2003.

3. Qi Y, Klein-seetharaman J, Bar-joseph Z. Random Forest Similarity for Protein-Protein Interaction Prediction, Pac Symp Biocomput; 2005.10:531-542.

4. Shi T, Seligson D, Belldegrun AS, Palotie A, Horvath S. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma, Modern Pathology; 2005.18:547-557.

5. Shi T, Horvath S. Unsupervised Learning with Random Forest Predictors, Journal of Computational and Graphical Statistics; 2006.15:118-138.

6. Chen X, Ishwaran H. Random forests for genomic data analysis, Genomics; 2012.99:323-329.

7. Chowdary D, Lathrop J, Skelton J, Curtin K, Briggs T, Zhang Y, et al. Prognostic gene expression signatures can be measured in tissues in RNA later preservative, J Mol Diagn; 2006.8:31-39.

8. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. New York, John Wiley and Sons; 1990.

9. Reynolds AP, Richrds G, Delaiglesia B, Rayward-Smith V. Clustering Rules: A comparison of Partitioning and Hierachical Clustering Algorithm. Journal of Mathematical Modelling and Algorithms; 2006.5:475-504.

10. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning:data mining inference and prediction, 2nd ed, California, Springer, 2009.

11. Breiman L. Random Forests, Machine Learning; 2001.45:5-32.

12. Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of new tests, Pattern Recognition; 2011.44:330-349.

13. Diaz-Uriarte R, Alvares de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics; 2006.7:3.

14. Liaw A, Wiener M. Classification and Regression by random Forest, R News; 2002. 2:18-22.

15. Dalton L, Ballarin V, Brun M. Clustering Algorithms: on Learning, Validation, Performance, and Applications to Genomics. Curr Genomics; 2009. 430-445.

16. Chunmei Y, Baikun W, Xiaofeng G. Effectivity of Internal Validation Techniques for Gene Clustering. In Biological and Medical Data Analysis; 2006. pp49-59.

17. Warrens MJ. On the Equivalence of Cohen's Kappa and the Hubert-Arabie Adjusted Rand Index. Journal of Classificatin; 2008. 25:177-183.

18. Jiang D, Tang C, Zhang A. Cluster Analysis for Gene Expression Data: A Survey.TKDE; 2004.16:1370-1386.

19. Maria I, Kurian M. A Survey on Clustering Approaches for Gene Expression Patterns. International Journal of Computer Trends and Technology; 2013.6:214-218.

20. Gan G, Ma C, Wu J. Data clustering: theory, algorithms and applications, Philadelphia: SIAM, society for Industrial and Applied Mathematics, 2007.

21. Han J, Kamber M, Pei, J. Data mining concepts and techniques, 3 nd ed, Waltham, Mass: Morgan Kaufmann Publishers, 2012.

22. de Souto MCP, Costa IG, de Araujo DSA, Ludermir TB, Schlieq A. Clustering cancer gene expression data: acomparative study, BMC Bioinformatics; 2008.9:101-114.

23.Jaskowiak AP, Campello JGBR, Costa GI. On the selection of appropriate distances for gene expression data clustering , BMC Bioinformatics; 2014.15:1-17.

24. Gasparovica M, Aleksejeva L, Nazaruks V. Using Fuzzy clustering with bioinformatics data, International Conference on Applied Information and Communication Technologies; 2013. pp. 62-70.

Original Article http://rjms.iums.ac.ir

Gene expression data clustering with random forest dissimilarity

***Zohreh Farhadi,** MSc of Statistics, Shahrood University, Shahrood, Iran.
Zohreh.farhadi87@gmail.com
Davood Shahsavani, PhD, Assistant Professor of Statistics, Shahrood University, Shahrood, Iran.
dshahsavani@shahroodut.ac.ir

Abstract

Background: The clustering of gene expression data plays an important role in the diagnosis and treatment of cancer. These kinds of data are typically involve in a large number of variables (genes), in comparison with number of samples (patients). Many clustering methods have been built based on the dissimilarity among observations that are calculated by a distance function. As increasing the dimensions reduces the performance of distance functions, most of the methods provide low accuracy. In this paper a new dissimilarity measure is introduced based on a classification method, called Random forests (RF). The performance of this new measure has been evaluated in the gene expression data.

Methods: In this article, the clustering problem of Chowdary data set, using the RF dissimilarity measure, is under consideration. At the first step, the clustering problem is converted to classification problem, thereafter; the new dissimilarity is calculated using the classification method of random forests. Finally, the data are clustered with a partition around mediod algorithm and the results are then evaluated by adjusted rand index. All the analysis is implemented with R software.

Results: The value of adjusted rand index (0.8149) represents an acceptable agreement between clusters and true groups. The most effective gene in constructing the clusters was gene no.31 which was detected by using the unique ability of RF that is identifying the importance of variables.

Conclusion: The random forest dissimilarity is an efficient criterion for measuring dissimilarity in gene expression data clustering. Detection of effective genes in clustering that is done with RF, helps the researcher in the diagnosing and treatment of the cancers.

Keywords: Clustering, Gene expression data, Random forest dissimilarity, Variables importance