
بسم الله الرحمن الرحيم

دانشکده علوم ریاضی
گروه ریاضی کاربردی

پایان‌نامه

برای دریافت درجه کارشناسی ارشد در رشته
آمار، گرایش آمار ریاضی

عنوان

خوشه‌بندی توسط روش رده‌بندی جنگل‌های تصادفی

استاد راهنما

دکتر داود شاهسونی

دانشجو

زهره فرهادی

آبان ماه ۱۳۹۳

تقدیم به آنان‌که

عطش و اشتیاق آموختن و به‌کارگیری آموزه‌ها،
وجودشان را فروزان
و این سرزمین را پر فروغ می‌نماید.

و

تقدیم به فرزند دلبندم که تمامی اوقات مصروف برای انجام این پایان‌نامه به او تعلق داشت.

سپاس گزاری...

سپاس بی حد بر آستان بی همتای احدیت که زندگی‌ام را مملو از الطاف بی منتهایش ساخته، خدایی که همیشه هست و یادش...؟

بسیار خرسندم که افتخار شاگردی استاد فرهیخته و عالی‌قدر جناب آقای دکتر داود شاهسونی را داشته‌ام. جا دارد بدین وسیله مراتب امتنان و قدردانی خود را به پاس تمام دلسوزی‌ها، راهگشایی‌ها و راهنمایی‌های ارزنده ایشان تقدیم حضورشان نمایم. همچنین از اساتید محترم جناب آقای دکتر محمد آرشی و دکتر محمدرضا ربیعی بواسطه‌ی زحمتی که در داوری این پایان‌نامه متقبل شده‌اند، تشکر و قدردانی می‌کنم.

فرصت را غنیمت شمرده و سپاس می‌گویم خداوندگاران مهر و مهربانی پدر و مادر عزیزم را که همواره پشتیبانم بودند و در سایه‌ی وجود پر برکت ایشان امکان تحصیل برایم میسر شد و صمیمانه‌ترین سپاس‌ها را به همسر عزیزم نثار می‌کنم که بدون حمایت و همراهی او انجام این پایان نامه ممکن نبود.

زهره فرهادی
آبان ماه ۱۳۹۳

تعهدنامه

اینجانب **زهره فرهادی** دانشجوی کارشناسی ارشد رشته **آمار** **دانشکده علوم ریاضی** دانشگاه شاهرود، نویسنده پایان‌نامه با عنوان **خوشه‌بندی توسط روش رده‌بندی جنگل‌های تصادفی** ، تحت راهنمایی **دکتر داود شاهسونی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه شاهرود متعلق دارد، و مقالات مستخرج با نام " دانشگاه شاهرود " یا " Shahrood University " به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده) شده است، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

زهره فرهادی
آبان ماه ۱۳۹۳

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

خوشه‌بندی یکی از مهم‌ترین ابزارهای داده‌کاوی است و در حال حاضر به طورگسترده در حوزه‌های مختلف علوم، مورد استفاده قرار می‌گیرد. بسیاری از روش‌های خوشه‌بندی بر پایه‌ی عدم تشابه مشاهدات که حاصل محاسبه‌ی یک تابع فاصله است، بنا شده‌اند. از آنجا که با افزایش بعد داده‌ها، کارآیی توابع فاصله کاهش می‌یابد، لذا خوشه‌بندی در ابعاد بالا با این چالش، مواجه است. در این پایان‌نامه معیاری جدید برای محاسبه‌ی عدم تشابه مشاهدات در ابعاد بالا، بر اساس روش رده‌بندی جنگل تصادفی معرفی شده‌است. طی مطالعه شبیه‌سازی و کار با یک مثال واقعی، عدم تشابه جدید در ترکیبی از روش مقیاس‌گذاری چندبعدی و روش خوشه‌بندی افراز حول مدوید، اعمال گردیده و کارآیی آن نسبت به یکی از مرسوم‌ترین توابع فاصله، مورد ارزیابی قرار گرفته‌است.

کلمات کلیدی: خوشه‌بندی، روش جنگل تصادفی، اهمیت متغیرها، روش افراز حول مدوید، مقیاس‌گذاری چندبعدی

پیشگفتار

خوشه‌بندی، فرآیند تقسیم مجموعه داده‌ها به زیر مجموعه‌های کوچک‌تری به نام خوشه است به طوری‌که اعضای هر خوشه، مشابه یکدیگر بوده و کمترین میزان تشابه را با اعضای سایر خوشه‌ها داشته باشند. این فرآیند که یکی از مباحث بنیادین داده‌کاوی و تحلیل‌های چند متغیره به شمار می‌آید، در حال حاضر به طور گسترده در حوزه‌های مختلف علوم مورد استفاده قرار می‌گیرد. بسیاری از روش‌های خوشه‌بندی بر پایه‌ی عدم تشابه مشاهدات 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. از آنجا که افزایش بعد داده‌ها، به شدت توابع فاصله را تحت تاثیر قرار می‌دهد، لذا پیدا کردن خوشه‌ها در ابعاد بالا کاری دشوار بوده و به عنوان یکی از مهم‌ترین چالش‌های خوشه‌بندی مطرح می‌گردد. در این پایان‌نامه با تبدیل مسئله خوشه‌بندی به رده‌بندی، معیار متفاوتی برای سنجش عدم تشابه مشاهدات معرفی شده که معیاری مناسب و کارا در مورد داده‌هایی است که در آن‌ها تعداد متغیرها بسیار زیاد بوده و حتی بیشتر از تعداد مشاهدات است. روش جنگل تصادفی گزینه مناسبی در مواجهه با داده‌های دورافتاده و نوفه‌ها است و عدم تشابه حاصل از آن می‌تواند توانایی روش‌های خوشه‌بندی را در برخورد با این‌گونه چالش‌ها افزایش دهد. عدم تشابه جنگل تصادفی وابسته به مقادیر متغیرها نیست و به‌کارگیری آن در روش‌های خوشه‌بندی ما را از انجام هر گونه پیش‌پردازشی روی داده‌ها بی‌نیاز می‌کند. این عدم تشابه را می‌توان در مورد انواع متغیرها و حتی با وجود مقادیر گمشده محاسبه نمود. همچنین با استفاده از قابلیت تعیین اهمیت متغیرها در روش جنگل تصادفی، می‌توان متغیرهای موثر در خوشه‌بندی را شناسایی نموده و خوشه‌های تخمین‌شده را توسط قاعده‌ای ساده بیان نمود.

در این پایان‌نامه پس از معرفی عدم تشابه جنگل تصادفی، عملکرد آن در روش‌های مقیاس‌گذاری چندبعدی و خوشه‌بندی افراز حول مدوید مورد ارزیابی قرار گرفته‌است. روش مقیاس‌گذاری چند بعدی به عنوان یک روش کاهش بعد و تحلیل اکتشافی داده‌ها مد نظر قرار گرفته و روش خوشه‌بندی افراز حول مدوید به عنوان الگوریتم اصلی خوشه‌بندی استفاده شده‌است. با این مقدمه، ساختار پایان‌نامه به صورت زیر تنظیم شده است:

- در فصل اول، تعاریف و مفاهیم اولیه در ارتباط با خوشه‌بندی را مطرح می‌کنیم.

- در فصل دوم، به بیان جزئی‌تر روش‌های مورد استفاده در تحقیق یعنی روش خوشه‌بندی افراز حول مدوید و روش مقیاس‌گذاری چند بعدی پرداخته و معیارهایی را به منظور ارزیابی عملکرد خوشه‌بندی معرفی می‌کنیم.

- در فصل سوم، ابتدا روش درخت تصمیم را که مبنای اندیشه جنگل تصادفی است، مرور نموده و سپس روش جنگل تصادفی را معرفی می‌کنیم. همچنین در این فصل چگونگی تبدیل مساله خوشه‌بندی به مساله رده‌بندی توضیح داده خواهد شد.

- در فصل چهارم، طی یک مطالعه‌ی شبیه‌سازی، ضمن معرفی خواص عدم تشابه جنگل تصادفی، عملکرد آن را نسبت به فاصله‌ی اقلیدسی در روش خوشه‌بندی افراز حول مدوید مورد ارزیابی قرار می‌دهیم.
- در فصل پنجم، به منظور خوشه‌بندی یک مجموعه داده‌ی واقعی، عدم تشابه جدید را در ترکیبی از روش مقیاس‌گذاری چندبعدی و روش افراز حول مدوید، اعمال نموده و نتایج حاصل از آن را با نتایج اخذ شده از اعمال فاصله‌ی اقلیدسی مورد مقایسه قرار می‌دهیم.
- در پیوست آ، رگرسیون یکنوا و روش عددی شیب نزولی معرفی شده و پیوست ب، حاوی کدهای نوشته شده در محیط R برای بازتولید مثال‌های پایان‌نامه است.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فهرست نمادها و اختصارات

تقریباً مساوی . . . . . $\cong$
آرگومان مینیمم کننده . . . . . argmin
آرگومان ماکزیمم کننده . . . . . $\arg\max$
مجموعه‌ی داده‌ها . . . . . $\mathbf{X}$
متغیر iام . . . . . X_i
مشاهده‌ی iام . . . . . $\mathbf{x}_i$
تعداد متغیرها . . . . . p
تعداد مشاهدات . . . . . n
تعداد خوشه‌ها . . . . . k
تشابه دو مشاهده‌ی i و j . . . . . $s(\mathbf{x}_i, \mathbf{x}_j)$
عدم تشابه دو مشاهده‌ی i و j . . . . . $d(\mathbf{x}_i, \mathbf{x}_j)$
فاصله‌ی اقلیدسی بین دو مشاهده‌ی i و j . . . . . $\delta(\mathbf{x}_i, \mathbf{x}_j)$
ماتریس عدم تشابه . . . . . D
ماتریس فاصله‌ی اقلیدسی . . . . . Δ
ماتریس عدم تشابه جنگل تصادفی مبتنی بر نمونه‌ی خودگردان . . . . . RF_{boot}
ماتریس عدم تشابه جنگل تصادفی مبتنی بر توزیع یکنواخت . . . . . RF_{uni}
ماتریس کوورایانس . . . . . $\mathbf{S}$
مجموعه‌ی مقادیر میانگین اعضای خوشه‌ها . . . . . $\mathcal{M}$
مجموعه‌ی مدویدها . . . . . M
مجموعه‌ی مشاهدات غیر مدوید . . . . . U
طول گام الگوریتم شیب نزولی . . . . . α
تعداد متغیرهای مورد جستجو در هر گره از درختان جنگل تصادفی . . . . . $mtry$
تعداد درختان جنگل تصادفی . . . . . $ntree$

فصل ۱

مروری بر خوشه‌بندی

۱.۱ مقدمه

بشر برای درک بهتر جهان پیرامون، به تجزیه و تحلیل داده‌هایی می‌پردازد که خصوصیات یک پدیده‌ی طبیعی را تبیین می‌کنند. با پیشرفت تکنولوژی و ظهور فناوری‌های نو، انقلابی در تولید و ذخیره‌سازی داده‌ها رخ داده و این امر موجب پیدایش پایگاه‌های داده‌ی بزرگ و متنوع شده‌است. بدیهی است که این حجم فزاینده از داده‌ها به سادگی قابل بهره‌برداری نیست، بنابراین نیاز به کشف دانشی است که در بطن داده‌ها قرار دارد. این مهم تحت عنوان داده کاوی [۱] در صدد است تا با به‌کارگیری علم آمار، هوش مصنوعی [۲] و یادگیری ماشین [۳] ، روابط پنهان در داده‌ها را تا حد ممکن آشکار نماید.

خوشه‌بندی[۴] یکی از مباحث پرکاربرد داده‌کاوی است که به تقسیم مجموعه‌ی داده‌ها به زیرمجموعه‌های کوچک‌تری به نام خوشه می‌پردازد به طوری‌که اعضای هر خوشه، مشابه یکدیگر بوده و کمترین میزان تشابه را با اعضای دیگر خوشه‌ها داشته باشند. در واقع، خوشه‌بندی همان تقسیم داده‌ها به گروه‌هایی است که عناصر آن‌ها دارای همگنی درون گروهی و ناهمگنی میان گروهی باشند. در خوشه‌بندی، گروه‌ها مجهول هستند و هیچ اطلاعی از ساختارهای موجود در داده‌ها در دست نیست. حتی در اغلب موارد، اطلاعی از تعداد خوشه‌ها و اینکه داده‌ها به چندین گروه قابل تقسیم هستند نیز در اختیار نداریم؛ به همین علت خوشه‌بندی، یک گروه‌بندی بدون نظارت [۵] محسوب می‌شود.

---

[۱] Data mining

[۲] Artifitial inteligence

[۳] Machine learning

[۴] Clustering

[۵] Unsupervised

در حال حاضر، خوشه‌بندی در حیطه‌ی گسترده‌ای از زمینه‌های کاربردی مورد استفاده قرار می‌گیرد. در علوم پزشکی با گروه‌بندی بیماران بر اساس علائم بالینی، درمان موثرتری برای آن‌ها صورت می‌گیرد. همچنین یافتن بیماران مشابه بر اساس اطلاعات ژنومی از کاربردهای مهم خوشه‌بندی در علوم پزشکی به شمار می‌رود که در تشخیص و درمان بیماری‌ها حائز اهمیت است. شرکت‌های تجاری و سازمان‌های بیمه به کمک خوشه‌بندی، مشتریان خود را در گروه‌هایی کوچک‌تر سازماندهی نموده و از این گروه‌بندی به منظور توسعه‌ی برنامه‌های مدیریتی و راهکارهای افزایش ارتباط با مشتری استفاده می‌کنند. دسته‌بندی اسناد اینترنتی و صفحات وب، بخشی از کاربردهای خوشه‌بندی در علوم کامپیوتر به شمار می‌رود که دسترسی راحت‌تر کاربران به آن‌ها را فراهم می‌کند. از دیگر کاربردهای مهم خوشه‌بندی می‌توان به شناسایی مناطق زلزله‌خیز، تهیه‌ی نقشه‌های زمین‌شناسی و یافتن مناطق خاص جغرافیایی اشاره کرد که در علوم زمین‌شناسی به آن پرداخته می‌شود.

در کنار خوشه‌بندی، نوع دیگری از گروه‌بندی به نام رده‌بندی [۱] وجود دارد. تفاوت اصلی رده‌بندی با خوشه‌بندی، وجود متغیری به نام متغیر پاسخ است که مشخص می‌کند هر مشاهده به کدام گروه (رده) تعلق دارد، در واقع در این نوع گروه‌بندی، گروه‌ها معلوم هستند و هدف، مدل‌بندی آن‌ها و یافتن قاعده‌ای برای تخصیص مشاهدات جدید به گروه‌های موجود است. از این‌رو رده‌بندی به عنوان گروه‌بندی با نظارت [۲] شناخته می‌شود.

بسیاری از روش‌های خوشه‌بندی بر پایه‌ی عدم تشابه مشاهدات 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 مانند خوشه‌بندی داده‌های ژنومی، مستندات متنی یا پردازش تصاویر، محققان با داده‌هایی مواجه هستند که تعداد متغیرهای توضیحی در آن‌ها بسیار زیاد است. از آنجا که افزایش بعد داده‌ها، کارآیی توابع فاصله را کاهش می‌دهد، خوشه‌بندی در ابعاد بالا، به یکی از مهم‌ترین چالش‌ها در این زمینه تبدیل گشته‌است.

واژه‌ی **مصیبت بعد بالا** [۳] اولین بار توسط بلمن [۴] مطرح شد [۶]. بیر [۵] و همکاران نشان دادند که افزایش ابعاد، توابع فاصله را به شدت تحت تاثیر قرار می‌دهد و در ابعاد به قدر کافی بزرگ، فاصله‌ی دو نقطه‌ی نزدیک به هم با فاصله‌ی دیگر نقاط تفاوتی ندارد [۸]. پس از آن در بسیاری از تحقیقات، تاثیر این موضوع بر روش‌های خوشه‌بندی مورد بررسی قرار گرفت و پیدا کردن خوشه‌ها در ابعاد بالا به

---

[۱] Classification

[۲] Supervised

[۳] Curse of dimensionality

[۴] Bellman

[۵] Beyer

عنوان مسئله‌ای چالش‌برانگیز مطرح شد. در این زمینه می‌توان به مطالعات اشتاینباخ [۱] و همکاران و رادووانوویچ [۲] و همکاران اشاره کرد [۳۷، ۴۵]. خوشه‌بندی در ابعاد بالا یکی از زمینه‌های بسیار مستعد برای تحقیق است و تاکنون روش‌ها و راه‌حل‌های متعددی برای حل این چالش ارائه شده‌اند. استفاده از روش‌های کاهش بعد، یکی از راهکارهایی است که به منظور مرتفع نمودن مشکلات ناشی از ابعاد بالای داده‌ها، مورد توجه عده‌ی زیادی از محققان قرار گرفته است. فرن و برادلی [۳] از جمله پژوهشگرانی بودند که این راهکار را در خوشه‌بندی با ابعاد بالا به‌کار گرفتند و یک روش کاهش بعد به نام تصویرسازی تصادفی [۴] را با خوشه‌بندی ترکیب نمودند [۱۹]. همچنین آنها با ارائه‌ی مقاله‌ای دیگر، استفاده از چندین روش کاهش بعد را در قالب خوشه‌بندی اجماعی، برای کار در ابعاد بالا پیشنهاد دادند [۲۰].

تعریف معیاری متفاوت برای سنجش عدم تشابه که وابسته به توابع فاصله نباشد، می‌تواند به عنوان راه حلی برای چالش مذکور مد نظر قرار گیرد و در بهبود عملکرد خوشه‌بندی موثر باشد. ساختار برخی از روش‌های رده‌بندی به گونه‌ای است که همزمان با رده‌بندی مشاهدات، عدم تشابه میان آن‌ها نیز با دیدگاهی کاملاً متفاوت تعریف می‌شود. لذا می‌توان با تبدیل مسئله‌ی خوشه‌بندی به مسئله‌ی رده‌بندی، از این نوع عدم تشابه در خوشه‌بندی بهره گرفت. بریمن و کاتلر [۵] عدم تشابهی کارا را در مواجهه با تعداد زیاد متغیرهای توضیحی بر مبنای روش رده‌بندی **جنگل تصادفی** معرفی کردند [۱۱]. این روش، قادر به تعریف معیاری مناسب و کارا در مورد داده‌هایی است که در آن‌ها تعداد متغیرها بسیار زیاد بوده و حتی بیشتر از تعداد مشاهدات است. همچنین گزینه مناسبی در مواجهه با داده‌های دورافتاده و نوفه‌ها است و عدم تشابه حاصل از آن می‌تواند توانایی روش‌های خوشه‌بندی را در برخورد با این‌گونه چالش‌ها افزایش دهد. این عدم تشابه را می‌توان در مورد انواع متغیرها به‌کار برد و در صورت وجود داده‌هایی با مقادیر گمشده محاسبه نمود. عدم تشابه جنگل تصادفی وابسته به مقادیر متغیرها نیست و به‌کارگیری آن در روش‌های خوشه‌بندی ما را از انجام هر گونه پیش‌پردازشی روی داده‌ها بی‌نیاز می‌کند. همچنین از آنجا که روش جنگل تصادفی قابلیت منحصر به‌فرد تعیین متغیرهای مهم در رده‌بندی را داراست، لذا می‌توان متغیرهای موثر در خوشه‌بندی را شناسایی نموده و خوشه‌های تخمین‌شده را توسط قاعده‌ای ساده توصیف کرد.

عدم تشابه جنگل تصادفی، تاکنون در بسیاری از پژوهش‌های کاربردی مورد استفاده قرار گرفته‌است.

---

[۱] Steinbach

[۲] Radovanovic

[۳] Fern and Brodley

[۴] Random projection

[۵] Breiman and Catler

از جمله پژوهش‌های انجام شده در این زمینه می‌توان به تحقیقات بریمن و کاتلر، کی [۱] و همکاران و همچنین چن و اسفرن [۲] اشاره کرد که در آن‌ها از عدم تشابه جنگل تصادفی به منظور خوشه‌بندی داده‌های ژنومی، ریزآرایه‌ها و داده‌های نشانگر تومور استفاده شده‌است [۱۱، ۱۲، ۳۶]. همچنین پربت [۳] و همکاران و ایشیوکو [۴] عملکرد عدم تشابه جنگل تصادفی را در روش‌های بدون نظارت مورد ارزیابی قرار دادند [۲۷، ۳۵]. شی [۵] و همکاران عدم تشابه جنگل تصادفی را در ترکیبی از روش مقیاس‌گذاری چند بعدی و روش خوشه‌بندی افراز حول مدوید به منظور خوشه‌بندی یک مجموعه داده بیان ژنی اعمال کردند [۴۳]. نتایج موفقیت آمیز این تحقیق سبب شد، شی و هرود [۶] با ارائه‌ی مقاله‌ای دیگر، خوشه‌بندی بر مبنای رده‌بندی جنگل تصادفی را به طور کامل‌تر مورد پژوهش و تحقیق قرار دهند. آن‌ها در این تحقیق عدم تشابه جنگل تصادفی را جایگزین مناسبی برای فاصله اقلیدسی معرفی نموده و با استفاده از روش مقیاس‌گذاری چند بعدی، خواص این نوع عدم تشابه را مورد مطالعه قرار دادند [۴۲]. همچنین رهیافت جدیدی برای محاسبه‌ی عدم تشابه از طریق روش جنگل تصادفی توسط انگلاند و وریکاس [۷] معرفی شد [۱۶].

دراین پایان‌نامه قصد داریم ضمن معرفی عدم تشابه جنگل تصادفی، به مطالعه خواص این نوع عدم تشابه پرداخته و کارآیی آن را نسبت به فاصله‌ی اقلیدسی در ترکیبی از روش مقیاس‌گذاری چندبعدی و روش خوشه‌بندی افراز حول مدوید توسط داده‌های شبیه‌سازی شده و واقعی مورد ارزیابی قرار دهیم . شایان ذکر است که از روش مقیاس‌گذاری چند بعدی به عنوان یک روش کاهش بعد و تحلیل اکتشافی داده‌ها و از روش خوشه‌بندی افراز حول مدوید به عنوان الگوریتم اصلی خوشه‌بندی استفاده شده‌است.

۲.۱ ساختار فرآیند خوشه‌بندی

خوشه‌بندی را می‌توان در چهارگام اصلی، شامل انتخاب یا استخراج متغیرها، انتخاب یا طراحی الگوریتم خوشه‌بندی، اعتبارسنجی خوشه‌ها و تفسیر نتایج بیان نمود [۵۰]. شکل ۱.۱ ساختار فرآیند خوشه‌بندی را نشان می‌دهد. در این بخش، به توضیح و معرفی هر یک از این مراحل می‌پردازیم:

---

[۱] Qi

[۲] Chen and Ishveran

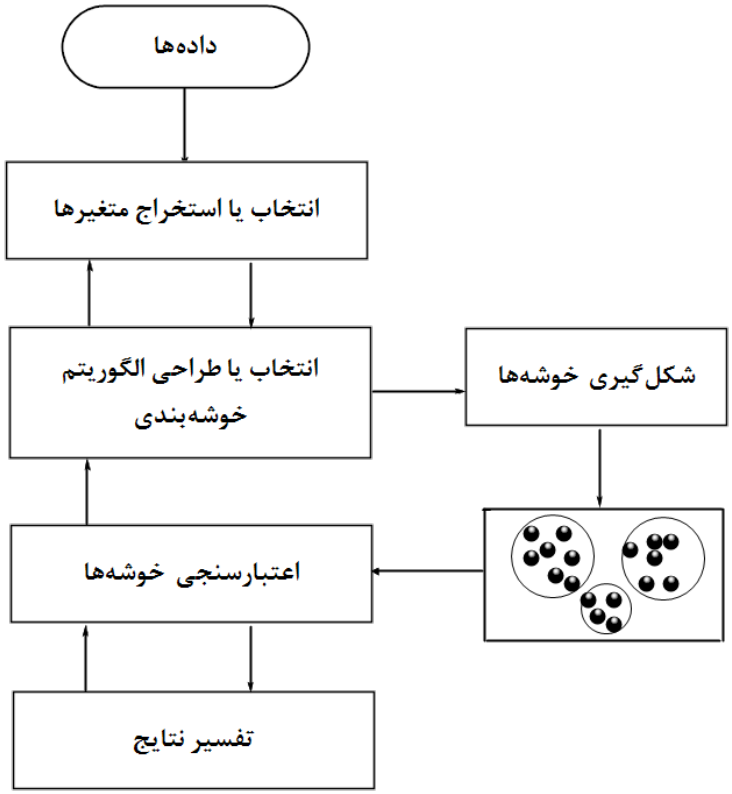
[۳] Prebet

[۴] Ishioka

[۵] Shi

[۶] Horvath

[۷] Englund and Verikas

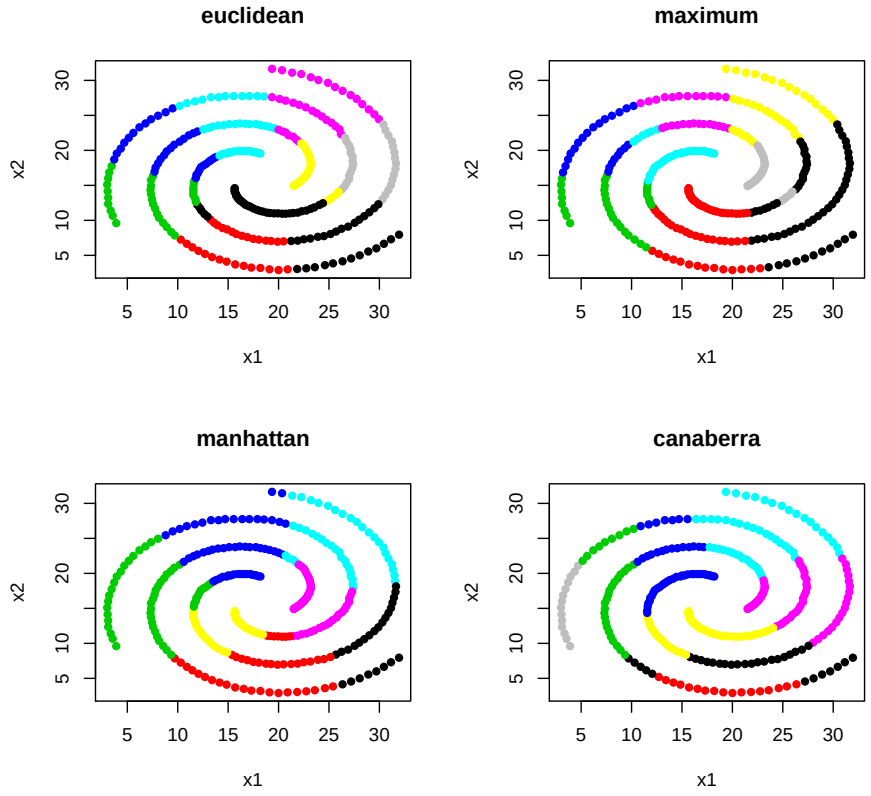


شکل ۱.۱: ساختار فرآیند خوشه‌بندی

- **انتخاب یا استخراج متغیرها** : انتخاب زیرمجموعه مناسبی از متغیرها حجم کار را تا حد زیادی کاهش می‌دهد و فرآیند خوشه‌بندی را تسهیل می‌بخشد. استخراج متغیرها به معنی استخراج متغیرهای جدید از متغیرهای اصلی است که از طریق انجام برخی تبدیلات روی متغیرهای اصلی صورت می‌گیرد. در مورد استخراج متغیرها در بخش‌های آتی با عنوان تجسم داده‌ها و کاهش بعد بحث خواهیم نمود و برای کسب اطلاعات بیشتر در مورد انتخاب متغیرها می‌توانید به مراجع [۵، ۹، ۴۴] مراجعه کنید. هر دو رهیافت انتخاب و استخراج متغیرها در اثربخشی روش‌های خوشه‌بندی و تفسیر بهتر نتایج نقش به‌سزایی را ایفا می‌کنند.

- **انتخاب یا طراحی الگوریتم خوشه‌بندی** : هر روش خوشه‌بندی بر اساس برخی ضوابط و معیارها عمل خوشه‌بندی را انجام می‌دهد و می‌تواند نتایج متفاوتی با دیگر روش‌ها ارائه دهد. حتی اگر در یک الگوریتم خوشه‌بندی از پارامترهای ورودی متفاوت استفاده شود، احتمال پیدا شدن خوشه‌های کاملاً متفاوت نیز وجود دارد. با توسعه کاربردهای خوشه‌بندی در زمینه‌های مختلف، دنیای روش‌های خوشه‌بندی رو به گسترش است و روش‌های جدید با قابلیت‌های بیشتر ابداع می‌شوند. اما با این حال هیچ الگوریتم خوشه‌بندی وجود ندارد که به طور فراگیر نسبت به

دیگر روش‌ها ترجیح داده شود. از طرفی تقریباً تمامی الگوریتم‌های خوشه‌بندی در ساز و کار خود به طور مستقیم یا غیر مستقیم با مفهوم **مجاورت** [۱] درگیر هستند و معیارهای متفاوت در سنجش مجاورت داده‌ها نیز می‌توانند منجر به نتایج کاملاً متمایز گردند. شکل ۲.۱ اجرای یک الگوریتم خوشه‌بندی با چهار معیار متفاوت سنجش مجاورت را نمایش می‌دهد که منجر به نتایج کاملاً متفاوتی شده‌اند (برای آگاهی در خصوص معیارهای سنجش مجاورت به‌کار رفته، جدول ۲.۱ را ببینید). طراحی الگوریتم خوشه‌بندی به معنای انتخاب ترکیبی مناسب از الگوریتم خوشه‌بندی، پارامترهای مربوط به الگوریتم و معیار مجاورت است.



شکل ۲.۱: نتایج اجرای یک الگوریتم خوشه‌بندی با چهار نوع معیار سنجش مجاورت به نامهای فاصله‌ی ماکزیمم، اقلیدسی، کانابرا و منهتن. خوشه‌های حاصل از الگوریتم، توسط رنگ‌ها از یکدیگر متمایز شده‌اند.

- **اعتبارسنجی خوشه‌ها** : اگر چه خوشه‌بندی یک روش بدون‌نظارت است و به طور قطعی نمی‌توان اعتبار خوشه‌های به‌دست آمده را تصدیق نمود، اما با این حال معیارهایی وجود دارند که می‌توانند در مسائلی همچون ارزیابی عملکرد خوشه‌بندی، بررسی معنی‌داری خوشه‌های به‌دست آمده، میزان

[۱] Proximity

تطابق خوشه‌ها با ساختارهای واقعی و مقایسه عملکرد روش‌های مختلف راهگشا باشند. به طور کلی معیارهای ارزیابی در خوشه‌بندی به دو دسته شاخص‌های بیرونی [۱] و درونی [۲] تقسیم می‌شوند. زمانی که حقیقت ساختارها (خوشه‌ها) موجود باشد و در واقع اطلاعات بیرونی از خوشه‌ها در دست باشد، می‌توان با مقایسه نمودن آن‌ها با نتایج خوشه‌بندی، مدل ایجاد شده را ارزیابی کرد. از شاخص‌های بیرونی برای اعتبار بخشی به الگوریتم‌های خوشه‌بندی استفاده می‌شود. شاخص‌های درونی، کیفیت خوشه‌ها را بر مبنای تراکم درون خوشه‌ای و جدایی خوشه‌ها از هم ارزیابی می‌کنند. از این رو شاخص‌های درونی به طور مستقیم بر روی داده‌های اصلی کار می‌کنند و نیازی به اطلاع از حقیقت ساختارها ندارند [۲۲، ۵۰].

- **تفسیر نتایج** : هدف نهایی خوشه‌بندی این است که اطلاعات مفیدی از ساختارهای موجود در داده‌ها را در اختیار قرار دهد تا بتوان بر پایه‌ی آن به حل مساله‌ای پرداخت. با اینکه معیارهای اعتبارسنجی تا حدی در مورد کیفیت خوشه‌بندی، آگاهی دهنده هستند اما تفسیر و تجزیه و تحلیل نهایی و بررسی اینکه نتایج تا چه اندازه معنی‌دار و قابل قبول هستند، برعهده کارشناسان در هر زمینه است.

در فرآیند خوشه‌بندی، یک معیار جامع و کارا برای انتخاب متغیرها و روش خوشه‌بندی وجود ندارد حتی انتخاب معیاری مناسب برای اعتبارسنجی هم یک چالش است. از این‌رو فرآیند خوشه‌بندی ممکن است نیاز به آزمایش و تکرار داشته باشد، به همین علت در شکل ۱.۱ برگشت‌پذیری لحاظ شده‌است.

۳.۱ معیارهای مجاورت

همانطورکه پیش از این نیز اشاره شد، تقریباً تمامی الگوریتم‌های خوشه‌بندی به طور مستقیم یا غیرمستقیم با معیارهای مجاورت سرو کار دارند. مجاورت، یک تعریف کلی از دو مفهوم تشابه و عدم تشابه است. داده‌ها توسط مجموعه‌ای از متغیرها توضیح داده می‌شوند و بسته به نوع متغیرهای توضیحی که می‌توانند اسمی [۳] ، دودویی [۴] ، عددی [۵] ، ترتیبی [۶] یا آمیخته‌ای از انواع متغیرها باشند، مجاورت به طرق مختلفی اندازه‌گیری می‌شود. فرض کنید $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_j, \cdots, \mathbf{x}_n\}$ مجموعه‌ای متشکل از n

[۱] External

[۲] Internal

[۳] Nominal

[۴] Binary

[۵] Numeric

[۶] Ordinal

داده (مشاهده) و p متغیر باشد که در آن $\mathbf{x}_j = (x_{j1}, x_{j2}, \cdots, x_{jp})^T$ مشاهده jام را نشان می‌دهد. در این صورت ماتریس مجاورت، یک ماتریس $n \times n$ خواهد بود که درایه ijام آن بیانگر تشابه یا عدم تشابه دو داده i و j است. تابع d تابع **فاصله** یا **عدم تشابه** نامیده می‌شود اگر در خواص زیر صدق کند [۳۹، ۴۶]:

- $\forall i, j \quad d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$
- $\forall i, j \quad d(\mathbf{x}_i, \mathbf{x}_j) \geq ۰$
- $\forall i \quad d(\mathbf{x}_i, \mathbf{x}_i) = ۰$

به همین ترتیب تابع s یک تابع تشابه است، اگر:

- $\forall i, j \quad s(\mathbf{x}_i, \mathbf{x}_j) = s(\mathbf{x}_j, \mathbf{x}_i)$
- $\forall i, j \quad ۰ \leq s(\mathbf{x}_i, \mathbf{x}_j) \leq ۱$
- $\forall i \quad s(\mathbf{x}_i, \mathbf{x}_i) = ۱$

در این بخش قصد داریم با در نظر گرفتن نوع متغیرهای توضیحی، معیارهایی را به منظور سنجش عدم تشابه مشاهدات معرفی کنیم. مطالب این بخش عمدتاً از مراجع [۲۱، ۲۲، ۵۰] استخراج شده‌اند.

۱.۳.۱ متغیرهای اسمی

مقادیر این‌گونه متغیرها، کیفی هستند و حالت‌هایی از مشاهدات را توصیف می‌کنند که برای آن‌ها هیچ‌گونه ترتیب و اولویتی در نظر گرفته نمی‌شود. رنگ و شغل را می‌توان به عنوان نمونه‌هایی از متغیرهای اسمی نام برد. اگر در این‌گونه متغیرها، تطابق دو مشاهده‌ی $\mathbf{x}_i$ و $\mathbf{x}_j$ را تعداد متغیرهایی در نظر بگیریم که مقدار آن‌ها برای دو مشاهده‌ی مذکور، یکسان است و آن را با m نمایش دهیم، آنگاه عدم تشابه دو مشاهده‌ی مذکور به صورت نسبت عدم تطابق آن دو، تعریف می‌شود که از طریق رابطه‌ی زیر قابل محاسبه است:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{p - m}{p} \qquad (۱.۱)$$

۲.۳.۱ متغیرهای دودویی

متغیر دودویی تنها بیانگر دو حالت برای مشاهدات است که با دو مقدار صفر و ۱ از هم تفکیک می‌شوند. اگر هر دو حالت، دارای اهمیت یکسانی باشند آن را متغیر دودویی متقارن نامیده و در غیر این صورت آن را متغیر دو دویی نامتقارن می‌نامند. به عنوان مثال پاسخ یک آزمایش برای تشخیص یک بیماری می‌تواند مثبت (۱) یا منفی (صفر) باشد، اما پاسخ مثبت آن اهمیت بیشتری نسبت به پاسخ منفی دارد، بنابراین پاسخ آزمایش مذکور، متغیر دودویی نامتقارن است. همچنین جنسیت را می‌توان به عنوان مثالی از یک متغیر دودویی متقارن نام برد. در صورتی که دو مشاهده‌ی $\mathbf{x}_i$ و $\mathbf{x}_j$ توسط متغیرهایی دودویی توصیف شده‌باشند، جدول فراوانی ۱.۱ را می‌توان برای آن‌ها تشکیل داد.

جدول ۱.۱: جدول فراوانی متغیرهای دودویی

		مشاهده‌ی iام		
		۱	۰	مجموع
مشاهده‌ی jام	۱	u	r	$u+r$
	۰	w	t	$w+t$
	مجموع	$u+w$	$r+t$	p

در جدول ۱.۱، u، r، w و t تعداد متغیرهایی هستند که مشاهده‌ی iام و jام آنها به ترتیب برابر با مقادیر متناظر سطرها و ستون‌ها هستند. به عنوان مثال u معرف تعداد متغیرهایی است که مشاهده‌ی iام و jام آنها برابر با مقدار ۱ است.

با استفاده از جدول ۱.۱، عدم تشابه مشاهداتی که توسط متغیرهای دودویی متقارن توصیف می‌شوند، را می‌توان به صورت زیر محاسبه نمود:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{r+w}{u+r+w+t} \qquad (۲.۱)$$

همچنین می‌توان از رابطه‌ی زیر به منظور سنجش عدم تشابه مشاهداتی که دارای متغیرهای دودویی نامتقارن هستند، استفاده کرد:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{r+w}{u+r+w} \qquad (۳.۱)$$

۳.۳.۱ متغیرهای عددی

معیارهای متعددی برای سنجش عدم تشابه مشاهداتی که توسط متغیرهای عددی توضیح داده می‌شوند، وجود دارند. چندین معیار به طور خلاصه در جدول ۲.۱ معرفی شده‌اند که در مورد برخی از آنها توضیحاتی در زیر آورده شده‌است:

اولین فاصله‌ی معرفی شده، فاصله‌ی مینکوفسکی است. این فاصله، تحت تاثیر متغیرهایی واقع می‌شود که واریانس بیشتری دارند و در محاسبه‌ی آن، متغیرهای مذکور بر دیگر متغیرها غالب هستند؛ از این رو در بسیاری از کاربردها قبل از محاسبه‌ی این نوع فاصله، متغیرها استاندارد می‌شوند.

مرسوم‌ترین معیار در سنجش عدم تشابه متغیرهای عددی، فاصله‌ی اقلیدسی است که حالت خاصی از فاصله‌ی مینکوفسکی به شمار می‌آید. این فاصله نسبت به انتقال و دوران پایا است، اما با انجام تبدیلات یکنوا تغییر خواهد کرد. استفاده از این فاصله در خوشه‌بندی منجر به پیدایش خوشه‌های کروی شکل خواهد شد. برای درک بهتر این موضوع، فرض کنید مشاهدات، دو بعدی هستند ($p = ۲$). خوشه‌ای مانند C_l را در نظر بگیرید. فرض کنید $\mathbf{x}_i = (x_{i۱}, x_{i۲})$ مرکز این خوشه، $\mathbf{x}_j = (x_{j۱}, x_{j۲})$ دورترین نقطه از مرکز و α فاصله‌ی بین این دو نقطه است. خوشه‌ی C_l را می‌توان به صورت زیر تعریف نمود:

$$C_l = \{\mathbf{x}_j \quad ; \quad d(\mathbf{x}_i, \mathbf{x}_j) \leq \alpha\} \tag{۴.۱}$$

$$= \{\mathbf{x}_j \quad ; \quad \sqrt{(x_{i۱} - x_{j۱})^۲ + (x_{i۲} - x_{j۲})^۲} \leq \alpha\} \tag{۵.۱}$$

ناحیه‌ای که عبارت (۵.۱) تعیین می‌کند، دایره‌ای به مرکز $\mathbf{x}_i$ و شعاع α است، لذا خوشه‌ی حاصل از اعمال فاصله‌ی اقلیدسی، یک دایره خواهد بود. با استدلالی مشابه می‌توان نشان داد که استفاده از فاصله‌ی منهتن و مجذور فاصله‌ی ماهالانوبیس به ترتیب منجر به پیدایش خوشه‌های مستطیل شکل و بیضی شکل خواهند شد.

در محاسبه‌ی مجذور فاصله‌ی ماهالانوبیس، $\mathbf{S}$ ماتریس کوواریانس تمامی داده‌ها است. در محاسبه‌ی این فاصله به متغیرها بر حسب واریانس و همبستگی خطی دو به دوی آن‌ها وزن داده می‌شود. همچنین لازم به ذکر است که اگر متغیرها ناهمبسته باشند، آنگاه ماتریس کوواریانس نمونه، یک ماتریس همانی است و در این صورت مجذور فاصله ماهالانوبیس با مجذور فاصله اقلیدسی برابر خواهد شد.

جدول ۲.۱: چندین معیار سنجش عدم تشابه مشاهدات با متغیرهای عددی .

نام	طریقه‌ی محاسبه
فاصله‌ی مینکوفسکی [آ]	$d(\mathbf{x}_i,\mathbf{x}_j)=\left[\sum_{k=۱}^{p}\lvert x_{ik}-x_{jk}\rvert^{r}\right]^{۱/r}\qquad r\geq ۱$
فاصله‌ی اقلیدسی [ب]	$d(\mathbf{x}_i,\mathbf{x}_j)=\left[\sum_{k=۱}^{p}(x_{ik}-x_{jk})^{۲}\right]^{۱/۲}$
فاصله‌ی منهتن [ج]	$d(\mathbf{x}_i,\mathbf{x}_j)=\sum_{k=۱}^{p}\lvert x_{ik}-x_{jk}\rvert$
مجذور فاصله‌ی ماهالانوبیس [د]	$d(\mathbf{x}_i,\mathbf{x}_j)=(\mathbf{x}_i-\mathbf{x}_j)^{T}\mathbf{S}^{-۱}(\mathbf{x}_i-\mathbf{x}_j)$
فاصله‌ی ماکزیمم [ه]	$d(\mathbf{x}_i,\mathbf{x}_j)=\max_{۱\leq k\leq p}\lvert x_{ik}-x_{jk}\rvert$
فاصله‌ی میانگین [و]	$d(\mathbf{x}_i,\mathbf{x}_j)=\left(\frac{۱}{p}\sum_{k=۱}^{p}(x_{ik}-x_{jk})^{۲}\right)^{۱/۲}$
فاصله‌ی کانابرا [ز]	$d(\mathbf{x}_i,\mathbf{x}_j)=\begin{cases}۰ & x_{ik}=x_{jk}=۰\\ \frac{\sum_{k=۱}^{p}\lvert x_{ik}-x_{jk}\rvert}{(\lvert x_{ik}\rvert+\lvert x_{jk}\rvert)} & x_{ik}\neq ۰ \text{یا} x_{jk}\neq ۰\end{cases}$
فاصله‌ی چرد [ح]	$d(\mathbf{x}_i,\mathbf{x}_j)=\left(۲-۲\left(\frac{\sum_{k=۱}^{p}x_{ik}x_{jk}}{\lVert\mathbf{x}_i\rVert_{۲}\lVert\mathbf{x}_j\rVert_{۲}}\right)\right)^{۱/۲}$ $\lVert\mathbf{x}_i\rVert_{۲}=\sqrt{\sum_{k=۱}^{p}x_{ik}^{۲}}$
ضریب همبستگی پیرسون [ط]	$d(\mathbf{x}_i,\mathbf{x}_j)=\frac{۱-r(\mathbf{x}_i,\mathbf{x}_j)}{۲}$ $r(\mathbf{x}_i,\mathbf{x}_j)=\frac{\sum_{k=۱}^{p}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sqrt{\sum_{k=۱}^{p}(x_{ik}-\bar{x}_i)^{۲}\sum_{k=۱}^{p}(x_{jk}-\bar{x}_j)^{۲}}}$

[آ] Minkowski
[ب] Euclidean
[ج] Manhattan
[د] Mahalanobis
[ه] Maximum
[و] Average
[ز] Canaberra
[ح] Chord
[ط] Pearson correlation cofficient

۴.۳.۱ متغیرهای ترتیبی

در متغیرهای ترتیبی، رتبه‌بندی مقادیر حائز اهمیت است. به عنوان مثال متغیر اندازه که می‌تواند شامل سه رتبه‌ی کوچک، متوسط و بزرگ باشد، یک متغیر ترتیبی است. فرض کنید مشاهدات توسط متغیرهای ترتیبی توضیح داده شده‌باشند و M_F را تعداد رتبه‌های متغیر X_f در نظر بگیرید. عدم تشابه مشاهدات را می‌توان به صورت زیر محاسبه نمود:

- مقادیر متغیرها با رتبه‌ی متناظرشان جابجا می‌شوند، یعنی به ازای هر $i = ۱, ۲ \cdots, n$ و $j = ۱, ۲, \cdots, p$ به جای مقدار x_{ij} مقدار رتبه‌ی متناظر آن یعنی r_{ij} قرار می‌گیرد که $r_{ij} \in \{۱, ۲, \cdots, M_j\}$.
- بازه‌ی متغیرهای جدید به بازه‌ی $[۰, ۱]$ نگاشت می‌شود. این کار از طریق انجام تبدیل زیر صورت می‌پذیرد:

$$r'_{ij} = \frac{r_{ij} - ۱}{M_j - ۱} \tag{۶.۱}$$

- اکنون می‌توان با متغیرهای جدید مانند متغیرهای عددی رفتار نمود و عدم تشابه مشاهدات را توسط یکی از معیارهای معرفی شده در بخش ۳.۳.۱ محاسبه کرد.

۵.۳.۱ آمیخته‌ای از انواع متغیرها

در بخش‌های قبل، به چگونگی سنجش عدم تشابه برای مشاهداتی پرداخته‌شد که متغیرهای توضیحی آن‌ها از نوعی یکسان هستند، اما یک مجموعه داده می‌تواند شامل انواع مختلفی از متغیرها باشد و در بسیاری از داده‌های واقعی، مشاهدات بوسیله‌ی آمیخته‌ای از انواع متغیرها توضیح داده می‌شوند. در مواجهه با چنین مواردی عدم تشابه مشاهدات از طریق رابطه‌ی زیر قابل محاسبه است:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{f=۱}^{p} \varphi^f(\mathbf{x}_i, \mathbf{x}_j) d^f(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{f=۱}^{p} \varphi^f(\mathbf{x}_i, \mathbf{x}_j)} \tag{۷.۱}$$

که در آن، اندیس بالایی f نشان دهنده‌ی نوع متغیر می‌باشد. در رابطه‌ی (۷.۱):

- در صورتی که f متغیری دودویی یا اسمی باشد؛

  اگر $x_{if} = x_{jf}$ آنگاه $d^f(\mathbf{x}_i, \mathbf{x}_j) = ۰$ و در غیر این صورت $d^f(\mathbf{x}_i, \mathbf{x}_j) = ۱$.

- در صورتی که f متغیری عددی باشد؛

  متغیرها استاندارد می‌شوند و برای محاسبه‌ی $d^f(\mathbf{x}_i, \mathbf{x}_j) = ۰$ از یکی از معیارهای معرفی شده در بخش ۳.۳.۱ استفاده می‌شود.

- اگر f ترتیبی باشد؛ عدم تشابه مشاهدات به طریقی که در بخش ۴.۳.۱ توضیح داده شد، محاسبه می‌شود.

همچنین در رابطه‌ی (۷.۱)، $\varphi^f(\mathbf{x}_i, \mathbf{x}_j) = ۰$ اگر:

- مقدار x_{if} یا x_{jf} مقداری گمشده باشد.
- $x_{if} = x_{jf} = ۰$ و متغیر f متغیر دودویی نامتقارن باشد.

و در غیر اینصورت : $\varphi^f(\mathbf{x}_i, \mathbf{x}_j) = ۱$

۴.۱ معرفی اجمالی متداول‌ترین رهیافت‌های خوشه‌بندی

تاکنون رهیافت‌های متنوعی در زمینه خوشه‌بندی ابداع شده‌اند و آن‌ها را می‌توان از منظرهای مختلفی طبقه‌بندی نمود. اما به طورکلی متداول‌ترین رهیافت‌های خوشه‌بندی عبارتند از افرازی [۱] ، سلسله مراتبی [۲] ، مبتنی بر شبکه [۳] و مبتنی بر مدل [۴] که در مراجع [۷، ۲۱، ۲۲] به تفصیل به هر یک از آن‌ها پرداخته شده است. در این بخش در مورد رایج‌ترین رهیافت‌های خوشه‌بندی یعنی سلسله مراتبی و افرازی توضیح کوتاهی ایراد می‌شود. لازم به ذکر است که ما در این پایان‌نامه توجه خود را بر روی یکی از روش‌های رهیافت افرازی به نام افراز حول مدوید [۵] (PAM) معطوف نموده‌ایم.

۱.۴.۱ رهیافت سلسله مراتبی

در اینگونه روش‌ها، خوشه‌بندی داده‌ها طی یک فرآیند سلسله مراتبی صورت می‌گیرد. با توجه به نحوه‌ی اجرای این فرآیند، روش‌های سلسله مراتبی به دو دسته تجمعی [۶] یا تقسیمی [۷] تقسیم می‌شوند. در روش‌های سلسله مراتبی تجمعی ابتدا هر مشاهده به عنوان خوشه‌ای مجزا در نظر گرفته می‌شود، سپس در هر گام خوشه‌هایی که تشابه بیشتری با هم دارند در هم ادغام شده و خوشه‌ی بزرگ‌تری را ایجاد می‌کنند تا در نهایت تمام مشاهدات درون یک خوشه قرار گیرند. معکوس این روند در روش‌های سلسله مراتبی

[۱] Partitioning

[۲] Hierachical

[۳] Grid-based

[۴] Model-based

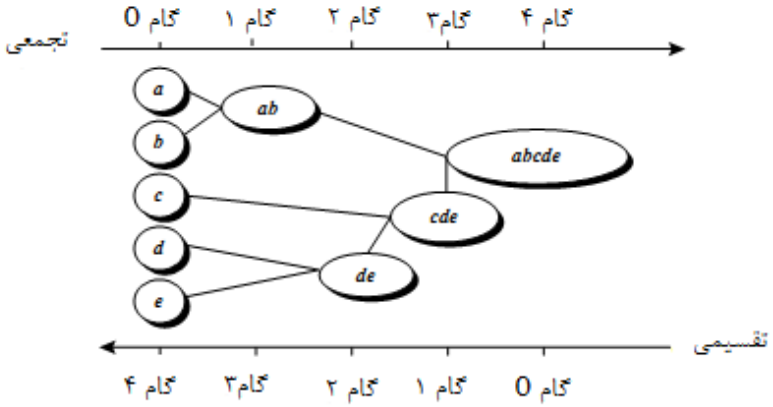
[۵] Partition around Medoid

[۶] Agglomerative

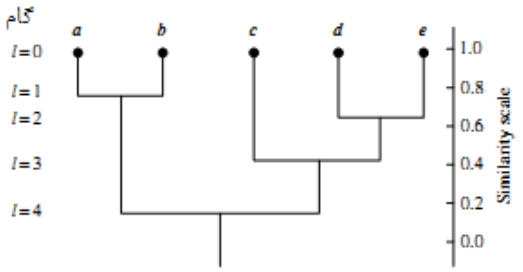
[۷] Divisive

تقسیمی رخ می‌دهد به طوری‌که ابتدا تمامی مشاهدات به عنوان یک خوشه در نظر گرفته می‌شوند، در گام بعدی این خوشه به دو خوشه‌ی کوچک‌تر تقسیم می‌شود و این روند ادامه می‌یابد تا زمانی‌که در هر خوشه، تنها یک مشاهده قرار گیرد.

شکل ۳.۱(آ) گام‌های الگوریتم سلسله مراتبی تجمعی و تقسیمی را نشان می‌دهد که بر روی مجموعه‌ی $\{a, b, c, d, e\}$ انجام شده‌است. نتایج خوشه‌بندی سلسله مراتبی را می‌توان در نموداری با ساختار درختی که دندروگرام [۱] نامیده می‌شود، نمایش داد. دندروگرام روش سلسله مراتبی تجمعی مجموعه‌ی $\{a, b, c, d, e\}$ در شکل ۳.۱(ب) نشان داده شده است. محور عمودی در این دندروگرام مقدار تشابه خوشه‌ها را نمایش می‌دهد. به عنوان مثال این دندروگرام نشان می‌دهد که دو خوشه‌ی $\{a, b\}$ و $\{c, d, e\}$ در حالی‌که تشابه آن‌ها حدود ۰/۲ بوده، در هم ادغام شده و خوشه‌ی بزرگ‌تری را ایجاد کرده‌اند.



(آ)



(ب)

شکل ۳.۱: (آ) : گام‌های الگوریتم سلسله مراتبی تجمعی و تقسیمی و (ب) : دندروگرام متناظر با روش سلسله مراتبی تجمعی [۲۲] .

[۱] Dendrogram

در روش‌های سلسله مراتبی، کاربر می‌تواند پس از پایان الگوریتم و مشاهده‌ی دندروگرام در مورد تعداد خوشه‌ها تصمیم‌گیری نموده و خوشه‌های مطلوب را استخراج کند. یکی از مشکلات الگوریتم‌های سلسله مراتبی این است که غیر قابل بازگشت هستند به این معنی که اگر در یک مرحله، داده‌ای به اشتباه به یک خوشه اختصاص داده شود، در مراحل بعدی نمی‌تواند به خوشه‌ی دیگری منتقل شود.

۲.۴.۱ رهیافت افرازی

روش‌های افرازی نیاز به دریافت تعداد خوشه‌ها برای انجام خوشه‌بندی دارند. فرض کنید قصد داریم n داده را به k خوشه تقسیم کنیم. در تقسیم‌بندی اولیه، k داده به عنوان مراکز اولیه به صورت تصادفی انتخاب می‌شوند که هر یک نمایانگر یک خوشه هستند، سپس سایر داده‌ها با توجه به میزان تشابه به یکی از مراکز نسبت داده می‌شوند و بدین ترتیب خوشه‌های اولیه شکل می‌گیرند. پس از تخصیص تمام داده‌ها به خوشه‌ها، مرکز هر خوشه دوباره محاسبه می‌شود و تخصیص نقاط بر مبنای تشابه مجدداً صورت می‌گیرد. این روند تا زمانی ادامه پیدا می‌کند که با محاسبه مراکز جدید در خوشه‌ها تغییر چندانی حاصل نشود. یک مزیت روش‌های افرازی نسبت به روش‌های سلسله مراتبی این است که در هر مرحله از روش‌های افرازی به منظور بهبود کیفیت خوشه‌بندی به داده‌ها اجازه داده می‌شود تا از خوشه‌ای به خوشه‌ی دیگر منتقل شوند.

۵.۱ چالش‌های خوشه‌بندی

گسترش کاربردهای خوشه‌بندی در حوزه‌های مختلف، سبب به‌روز چالش‌ها و پیچیدگی‌هایی در این زمینه گشته است که برخی از آنها عبارتند از [۲۲، ۲۴، ۳۸] :

- **توانایی خوشه‌بندی در ابعاد بالا:** بنابر آنچه در بخش ۱.۱ گفته شد، توابع فاصله، ممکن است در ابعاد بالا اثربخش نباشند و قطعاً این موضوع بر کارایی روش‌های خوشه‌بندی مبتنی بر عدم تشابه بی تاثیر نخواهد بود. از طرفی در مجموعه داده‌هایی با تعداد متغیرهای زیاد، احتمال وجود متغیرهای نوفه بیشتر بوده و پیدا کردن ساختارهای موجود در این مواقع، کار دشواری است.

  همچنین ذکر این نکته ضروری است که هنگام کار در ابعاد بالا تنها پیدا کردن خوشه‌ها کافی نیست بلکه باید بتوان خوشه‌های تخمینی را توسط مجموعه‌ای از متغیرها نشان داد چرا که در یک فضای بزرگ، همه‌ی متغیرها در ایجاد خوشه‌ها نقش ندارند و قسمتی از فضا برای مشخص کردن آن‌ها کافی است. بنابراین پیدا کردن خوشه‌ها در ابعاد بالا از یک سو و توصیف خوشه‌های تخمینی از سوی دیگر یکی از مهم‌ترین چالش‌هایی است که مسئله خوشه‌بندی با آن‌ها روبرو است.

در این پایان‌نامه سعی شده‌است که با استفاده از توانمندی روش جنگل تصادفی، راه‌حلی برای رویارویی با چالش مذکور ارائه گردد.

- **مقیاس پذیری:** بسیاری از روش‌های خوشه‌بندی روی مجموعه داده‌هایی با حجم کم عملکرد خوبی دارند، اما ممکن است در برخی کاربردها با پایگاه داده‌های بزرگ متشکل از میلیون‌ها یا میلیاردها داده سر و کار داشته‌باشیم؛ از این رو نیاز به الگوریتم‌های خوشه‌بندی مقیاس پذیر احساس می‌شود.

- **توانایی مواجهه با انواع مختلف متغیرها:** بسیاری از الگوریتم‌ها برای داده‌های عددی ایجاد شده‌اند و در برخی موارد نیاز به خوشه‌بندی انواع دیگر داده مانند رسته‌ای، رتبه‌ای، دودویی و یا ترکیبی از آن‌ها وجود دارد. بنابراین توسعه روش‌های خوشه‌بندی به روش‌هایی که بتوانند انواع مختلف متغیرها و خصوصیات را تحلیل کنند، یکی دیگر از نیازمندی‌ها در زمینه‌ی خوشه‌بندی است.

- **پیدا کردن خوشه‌هایی با اشکال غیر کروی:** اعمال برخی معیارهای سنجش عدم تشابه مانند فاصله اقلیدسی در الگوریتم‌های خوشه‌بندی منجر به ایجاد خوشه‌های کروی با چگالی و اشکال تقریباً یکسان می‌شوند. این در حالی است که یک خوشه می‌تواند هر شکلی داشته باشد. بنابراین توانایی تشخیص خوشه‌هایی با اشکال متفاوت توسط الگوریتم‌های خوشه‌بندی حائز اهمیت است.

- **توانایی بررسی نوفه‌ها:** بیشتر پایگاه داده‌ها در دنیای واقعی شامل داده‌های اشتباه، داده‌های پرت یا از دست رفته می‌باشند. الگوریتم خوشه‌بندی باید نسبت به این نوع داده‌ها مقاوم باشد ودر مواجهه با آن‌ها عملکرد مناسبی از خود نشان دهد.

- **نیاز به تعیین پارامترهای ورودی:** بسیاری از الگوریتم‌های خوشه‌بندی نیاز به پارامترهایی از پیش تعیین شده دارند. تعداد خوشه‌ها یکی از این پارامترها به شمار می‌آید. تعیین این پارامترها به خصوص در مجموعه‌هایی با ابعاد بزرگ و مسائلی که کاربر درک عمیقی از آن‌ها ندارد، بسیار دشوار است. این نیازمندی نه تنها بار اضافی بر کاربر تحمیل می‌کند، بلکه کنترل کیفیت خوشه‌بندی را نیز دشوار خواهد کرد.

- **قابلیت تفسیر و کارایی:** نتایج خوشه‌بندی باید برای کاربران، قابل درک و توصیف و همچنین کارا باشد؛ از این رو توسعه روش‌های خوشه‌بندی به روش‌هایی که دارای این خواص باشند، بسیار حائز اهمیت است. همچنین از آنجا که هر روش خوشه‌بندی ویژگی‌ها و خصوصیات مخصوص به خود را داراست بهتر است با مطالعه و تحقیق دقیق، از مناسب بودن آن برای هدف مورد نظر اطمینان پیدا نموده و بدین وسیله خوشه‌بندی کارآمدتری صورت پذیرد.

به منظور حل چالش‌های ذکر شده، تلاش‌های زیادی توسط پژوهشگران ومتخصصان داده‌کاوی برای رسیدن به روش‌هایی با قابلیت‌ها و بازدهی بیشتر صورت گرفته است. از آنجا که عدم تشابه داده‌ها تاثیری بنیادین در نتیجه خوشه‌بندی دارد، تعریف معیاری مناسب برای سنجش عدم تشابه هم می‌تواند توانایی روش‌های خوشه‌بندی را در برخورد با این چالش‌ها افزایش دهد. همچنین اجرای روش‌های مختلف خوشه‌بندی روی یک مجموعه داده ممکن است منجر به نتایج کاملاً متفاوتی شوند، لذا در اهداف کاربردی، با آزمودن روش‌های متفاوت روی مجموعه داده مورد نظر و یا اعمال چندین معیار سنجش عدم تشابه در روشی خاص و سپس مقایسه نتایج آن‌ها، می‌توان به یافتن خوشه‌بندی بهینه امیدوار بود.

۶.۱ تجسم اکتشافی داده‌ها از طریق کاهش بعد

همانطورکه در بخش قبل اشاره شد، بالا بودن بعد داده‌ها یکی از مهم‌ترین چالش‌های خوشه‌بندی به شمار می‌آید و اغلب روش‌های خوشه‌بندی توانایی مواجهه با آن را ندارند. از این رو کاهش بعد در خوشه‌بندی رهیافتی اثربخش است و می‌تواند در بسیاری از جنبه‌های تجزیه و تحلیل داده‌های چندمتغیری مفید باشد. کاهش بعد داده‌ها علاوه بر کاهش هزینه محاسباتی و تسهیل فرآیند خوشه‌بندی، امکان ترسیم داده‌ها و بررسی بصری آن‌ها را فراهم می‌کند.

تعیین تعداد خوشه‌ها در خوشه‌بندی از دیگر چالش‌های پیش رو است. از آنجا که با ترسیم داده‌ها در یک نمودار دو یا سه بعدی، اغلب می‌توان گروه‌هایی قابل تمییز را به صورت بصری تشخیص داد، کاهش ابعاد را می‌توان به عنوان راه حلی برای کشف تعداد خوشه‌ها به‌کار برد.

آنالیز مولفه‌های اصلی [۱] (PCA)، آنالیز مولفه‌های مستقل [۲] (ICA)، جستجوی تصویر [۳] و نقشه خود سازماندهی [۴] (SOM)، از روش‌های تجسم اکتشافی داده‌ها هستند که در مراجع [۱۷، ۲۱، ۲۸، ۵۰] به آن‌ها پرداخته شده است. مقیاس‌گذاری چند بعدی [۵] (MDS)، از دیگر روش‌های تجسم اکتشافی داده‌ها از طریق کاهش بعد است که به طور گسترده در خوشه‌بندی مورد استفاده قرار می‌گیرد [۵۰]. این روش با دریافت ماتریس مجاورت داده‌ها، به ساخت پیکره‌ای از آن‌ها در ابعاد پایین‌تر از بعد اصلی می‌پردازد بطوری‌که مجاورت بین داده‌ها حفظ شود.

---

[۱] Principle component analysis

[۲] Independent component analysis

[۳] Projection pursuit

[۴] Self-organizing map

[۵] Multidimensional scaling

شایان ذکر است که کاهش بعد داده‌ها همیشه منجر به حصول نتایج بهتر نخواهد شد. گاهی ممکن است کاهش ابعاد، سبب از دست رفتن برخی اطلاعات شود که این امر می‌تواند سبب تحریف خوشه‌ها گردد و نتایج گمراه‌کننده‌ای به همراه داشته باشد. در فصل‌های آتی در مورد کاهش بعد توسط روش مقیاس‌گذاری چند بعدی و چگونگی حفظ حداکثری اطلاعات داده‌ها، توضیحات بیشتری ارائه خواهد شد.

فصل ۲

روش‌ها

در فصل اول، مفاهیم اولیه در ارتباط با موضوع خوشه‌بندی را مرور کردیم. اکنون می‌توان به بیان جزئی‌تر روش‌هایی پرداخت که در این پایان‌نامه از آن‌ها استفاده شده است. ابتدا روش خوشه‌بندی k-میانگین [۱] را معرفی نموده و سپس با ذکر معایب آن، روش جایگزینی به نام k-مدوید [۲] را مورد بحث قرار خواهیم داد. سپس شاخص‌هایی را به منظور ارزیابی خوشه‌بندی معرفی می‌کنیم. در ادامه، روش مقیاس‌گذاری چند بعدی را به عنوان یک روش تجسم اکتشافی داده‌ها بررسی نموده و این فصل را با مبحث تعیین تعداد خوشه‌ها خاتمه می‌دهیم.

۱.۲ روش خوشه‌بندی k-میانگین

روش k-میانگین یکی از روش‌های افرازی است. همانطور که در بخش ۲.۴.۱ توضیح داده شد، تعداد خوشه‌ها در روش‌های افرازی، پارامتری از پیش تعیین شده است و هر خوشه توسط یک مرکز نشان داده می‌شود. با پیدا کردن نقاط مشابه با مرکز هر خوشه، اعضای آن خوشه تعیین خواهند شد. در الگوریتم k-میانگین که نخستین بار توسط لوید [۳] معرفی شد [۳۳]، میانگین مقادیر اعضای هر خوشه به عنوان مرکز خوشه در نظر گرفته می‌شود. مجموعه داده $\mathbf{X}$ را با n مشاهده در نظر بگیرید. فرض کنید تعداد خوشه‌ها برابر k است. مراحل این الگوریتم را می‌توان به صورت زیر بیان نمود [۷، ۵۰]:

۱. تعداد k نقطه به صورت تصادفی از مجموعه‌ی $\mathbf{x}$، به عنوان مراکز اولیه خوشه‌ها انتخاب می‌شوند. این مجموعه نقاط را با $\mathcal{M} = \{\mathbf{x}_{m_1}, \mathbf{x}_{m_2}, \cdots, \mathbf{x}_{m_k}\}$ نشان می‌دهیم.

---

[۱] k-means

[۲] k-medoid

[۳] Lloyd

۲. در این مرحله، عدم تشابه $n-k$ مشاهده باقیمانده با هر یک از اعضای مجموعه‌ی $\mathcal{M}$ محاسبه می‌شود و هر مشاهده به خوشه‌ای اختصاص داده می‌شود که تشابه بیشتری با مرکز آن دارد. به عبارتی اگر:

$$l = \arg\min_{1 \le i \le k} d(\mathbf{x}_{m_i}, \mathbf{x}_j) \qquad \mathbf{x}_j \in \mathbf{x} - \mathcal{M} \tag{۱.۲}$$

آنگاه $\mathbf{x}_j$ به lامین خوشه تعلق می‌گیرد یعنی $\mathbf{x}_j \in C_l$.

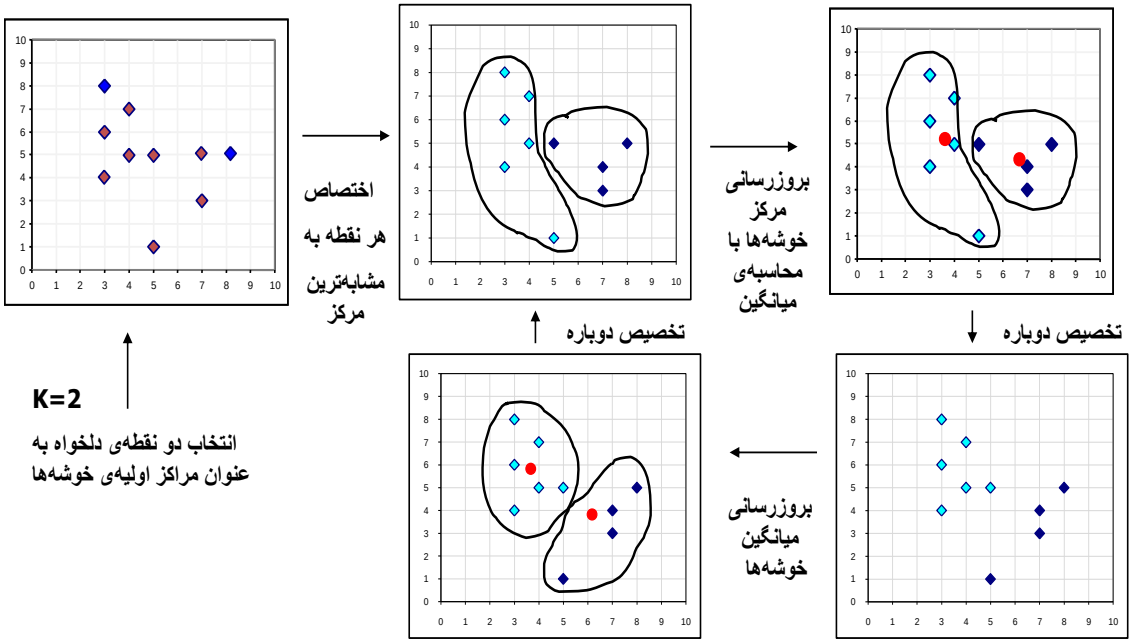
۳. پس از تخصیص تمامی مشاهدات به خوشه‌ها، میانگین مقادیر اعضای خوشه‌ها به عنوان مراکز جدید در نظر گرفته می‌شوند و مجموعه‌ی $\mathcal{M}$ بدین نحو، به‌روزرسانی می‌شود. یعنی:

$$\forall 1 \le i \le k \qquad \mathbf{x}_{m_i} = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j \tag{۲.۲}$$

که در آن n_i تعداد اعضای خوشه C_i است.

۴. مراحل ۲ و ۳ تا زمانی تکرار می‌شوند که با به‌روز شدن مجموعه‌ی $\mathcal{M}$، اعضای خوشه‌ها تغییر نکنند.

در شکل ۱.۲ مراحل اجرای الگوریتم k-میانگین در قالب یک مثال نمایش داده شده‌است.



شکل ۱.۲: مراحل اجرای الگوریتم k-میانگین

الگوریتم k-میانگین، بسیار ساده است و به آسانی می‌توان آن را در بسیاری از موارد کاربردی اجرا نمود. این الگوریتم به ویژه زمانی که خوشه‌های واقعی، کروی شکل هستند، عملکرد مناسبی از خود نشان می‌دهد و برای مجموعه داده‌های بزرگ نیز مناسب است. اما الگوریتم مذکور، دارای ضعف‌ها و اشکالاتی نیز هست که در اینجا به مهم‌ترین آن‌ها اشاره می‌کنیم [۲۲، ۵۰]:

- ضرورت تعیین تعداد خوشه‌ها قبل از انجام خوشه‌بندی، اولین ضعف الگوریتم k-میانگین و به طور کلی روش‌های افرازی است.
- این الگوریتم، قادر به پیدا کردن خوشه‌های بهینه کلی نیست بدین معنی که در هر بار اجرای الگوریتم k-میانگین، خوشه‌های نهایی فقط به ازای نقاطی که به عنوان مراکز اولیه انتخاب شده‌اند، بهینه هستند و ممکن است اگر نقاط دیگری به عنوان مراکز اولیه در گام ۱ انتخاب شوند، خوشه‌بندی باکیفیت‌تری حاصل گردد. به همین علت بهتر است الگوریتم k-میانگین به دفعات متعدد و با انتخاب مراکز اولیه متفاوت اجرا شود. اگر مقدار خطا را برای هر مشاهده، معادل عدم تشابه آن مشاهده با مرکز خوشه‌اش تعریف کنیم، مجموع توان‌های دوم خطا، معیار سنجش کیفیت خوشه‌بندی k-میانگین است که به صورت زیر تعریف می‌شود:

$$SSE = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} d^2(\mathbf{x}_{m_i}, \mathbf{x}_j) \qquad (۳.۲)$$

در واقع در دو اجرای متفاوت از این الگوریتم، اجرایی ترجیح داده می‌شود که عبارت SSE را مینیمم کند.

- از الگوریتم k-میانگین تنها زمانی می‌توان استفاده کرد که میانگین مجموعه داده‌ها قابل تعریف باشد، به همین علت نمی‌توان از این الگوریتم در حالتی که مجموعه داده‌ها شامل مقادیر اسمی است، استفاده نمود.
- این روش به دلیل استفاده از میانگین مقادیر اعضای خوشه به عنوان مرکز، به شدت نسبت به نقاط دورافتاده حساس است. در صورت وجود چنین داده‌هایی و اختصاص آن‌ها به خوشه، میانگین به شدت تغییر خواهد کرد و به این ترتیب ورود دیگر داده‌ها به خوشه نیز دستخوش تغییراتی خواهد شد.

۲.۲ روش خوشه‌بندی k-مدوید

ایده‌ی استفاده از مرکزی‌ترین عضو خوشه به جای میانگین، نوع دیگری از روش‌های افرازی به نام k-مدوید را پدید می‌آورد. در روش‌های k-مدوید، به جای میانگین مقادیر اعضای خوشه، یکی از اعضای

همان خوشه که مرکزی‌ترین عضو خوشه است، به عنوان نماینده انتخاب می‌شود. این عضو که مدوید نامیده می‌شود، دارای بیشترین تشابه با سایر اعضای خوشه است. برخلاف میانگین، مدوید در مقابل داده‌های دورافتاده مقاوم است، از این رو الگوریتم k-مدوید، مشکل الگوریتم k-میانگین در مواجهه با نقاط دورافتاده را نخواهد داشت. از طرفی مفهوم مدوید در مورد داده‌های اسمی نیز قابل تعریف است، بنابراین از این منظر هم الگوریتم k-مدوید نسبت به الگوریتم k-میانگین برتری دارد. اما ایرادی که در مورد ضرورت دریافت تعداد خوشه‌ها به الگوریتم k-میانگین وارد است، در مورد الگوریتم k-مدوید نیز صدق می‌کند و در این روش نیز تعداد خوشه‌ها می‌بایست از پیش تعیین شوند. یکی از مرسوم‌ترین روش‌های k-مدوید، روش افراز حول مدوید است که در ادامه به توضیح آن می‌پردازیم.

۱.۲.۲ روش افراز حول مدوید

روش افراز حول مدوید یکی از اولین الگوریتم‌های k-مدوید است که توسط کافمن [۱] و روسیوو[۲] معرفی شد [۲۹]. اگر خطای هر مشاهده را عدم تشابه آن با مدوید خوشه‌اش (نزدیک‌ترین مدوید) تعریف کنیم، هدف الگوریتم PAM مینیمم سازی مجموع خطای مطلق [۳] است که از آن به عنوان **تابع هزینه**[۴] یاد می‌شود و از طریق رابطه‌ی زیر قابل محاسبه است:

$$SAE = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} d(\mathbf{x}_{m_i}, \mathbf{x}_j) \qquad (۴.۲)$$

که در آن k تعداد خوشه‌ها و $\mathbf{x}_{m_i}$ مدوید خوشه‌ی C_i است. در این الگوریتم، داده‌ها بر مبنای ماتریس عدم تشابه خوشه‌بندی می‌شوند و همانطور که ذکر شد، تعیین تعداد خوشه‌ها، نقطه شروع خوشه‌بندی است. فرض کنید تعداد خوشه‌ها، k در نظر گرفته شده است و $D = [d(\mathbf{x}_i, \mathbf{x}_j)]$ ماتریس عدم تشابه مشاهدات است که درایه‌ی ijام آن، میزان عدم تشابه $\mathbf{x}_i$ و $\mathbf{x}_j$ را نشان می‌دهد. الگوریتم PAM داده‌ها را طی دو گام اصلی مرسوم به ساخت و جایگزینی به k خوشه تقسیم می‌کند. هریک از دو گام مذکور شامل مراحلی هستند که در ادامه به توضیح آن‌ها خواهیم پرداخت.

M را مجموعه مدویدها و $U = \mathbf{X} - M$ را مجموعه مشاهدات غیرمدوید در نظر می‌گیریم. همچنین برای هر مشاهده $\mathbf{x}_j$ پارامتر $d_1(\mathbf{x}_j)$ عدم تشابه بین $\mathbf{x}_j$ و نزدیک‌ترین مدوید و همچنین $d_2(\mathbf{x}_j)$ را عدم تشابه بین $\mathbf{x}_j$ و دومین مدوید نزدیک به آن در نظر می‌گیریم. با تغییر مجموعه M و U دو پارامتر مذکور به‌روزرسانی می‌شوند، همچنین $d_1(\mathbf{x}_j) = 0$ اگر و فقط اگر $\mathbf{x}_j \in M$. طبق تعریف قبل از خطای هر

[۱] Kaufman

[۲] Rousseeuw

[۳] Sum of absolute errors

[۴] Cost function

مشاهده، در نمادگذاری جدید، $d_۱(\mathbf{x}_j)$ بیانگر خطای مشاهده‌ی $\mathbf{x}_j$ است و بر این اساس، تابع هزینه‌ی الگوریتم PAM را می‌توان به صورت زیر بازنویسی کرد:

$$SAE = \sum_{j=۱}^{n} d_۱(\mathbf{x}_j) \qquad (۵.۲)$$

گام ساخت

در این گام، k مشاهده به عنوان مدویدهای اولیه انتخاب می‌شوند و با تخصیص سایر نقاط به آن‌ها k خوشه‌ی اولیه شکل می‌گیرد. در شروع این گام، مجموعه‌ی M تهی و مجموعه‌ی U شامل n مشاهده است. در اولین مرحله از این گام، مشاهده‌ای از مجموعه‌ی U که مجموع فواصلش با دیگر مشاهدات مینیمم است، به عنوان اولین مدوید انتخاب شده و به مجموعه‌ی M منتقل می‌شود. مدوید بعدی به گونه‌ای انتخاب می‌شود که منجر به مینیمم شدن تابع هزینه گردد و به همین ترتیب مشاهداتی از مجموعه‌ی U به عنوان مدوید به مجموعه‌ی M ملحق می‌شوند تا زمانی‌که k مدوید در مجموعه‌ی M قرار گیرند.

با این مقدمه، گام ساخت را می‌توان در مراحل زیر بیان نمود:

۱. برای هر مشاهده $\mathbf{x}_j$، پارامتر sd_j به صورت زیر محاسبه می‌شود:

$$sd_j = \sum_{i=۱}^{n} d(\mathbf{x}_i, \mathbf{x}_j) \qquad (۶.۲)$$

مشاهده‌ای با کمترین مقدار sd به عنوان اولین مدوید انتخاب می‌شود و به مجموعه M می‌پیوندد.

۲. مشاهده‌ی $\mathbf{x}_i \in U$ در نظر گرفته می‌شود.

۳. به منظور تصمیم‌گیری در مورد انتخاب مشاهده‌ی $\mathbf{x}_i$ به عنوان مدوید، وضعیت سایر مشاهدات مجموعه‌ی U مانند $\mathbf{x}_j$ نسبت به آن مورد بررسی قرار می‌گیرد و در صورتی‌که این انتخاب، منجر به کاهش مقدار تابع هزینه گردد، مشاهده‌ی $\mathbf{x}_i$ به مجموعه‌ی M منتقل می‌شود. بدین منظور $d_۱(\mathbf{x}_j)$ (عدم تشابه مشاهده‌ی $\mathbf{x}_j$ با نزدیکترین مدویدش) با $d(\mathbf{x}_i, \mathbf{x}_j)$ مقایسه می‌شود. برای دو عدم تشابه مذکور یکی از حالت‌های زیر صدق می‌کند:

- اگر $d(\mathbf{x}_i, \mathbf{x}_j) \leq d_۱(\mathbf{x}_j)$، آنگاه با ورود $\mathbf{x}_i$ به مجموعه‌ی M، نزدیک‌ترین مدوید به $\mathbf{x}_j$ محسوب خواهد شد. با تغییر مدوید $\mathbf{x}_j$، مقدار خطای آن به اندازه‌ی $d_۱(\mathbf{x}_j) - d(\mathbf{x}_i, \mathbf{x}_j)$ کاهش خواهد یافت.
- اگر $d(\mathbf{x}_i, \mathbf{x}_j) > d_۱(\mathbf{x}_j)$، آنگاه با انتخاب $\mathbf{x}_i$ به عنوان مدوید، نزدیک‌ترین مدوید به $\mathbf{x}_j$ تغییر نمی‌کند، لذا مقدار خطای مشاهده‌ی $\mathbf{x}_j$ نیز بدون تغییر باقی می‌ماند.

اگر پارامتر c_{ij} را میزان کاهش در خطای مشاهده‌ی $\mathbf{x}_j$ پس از ورود $\mathbf{x}_i$ به مجموعه‌ی M تعریف کنیم آنگاه بنا به دو حالت مذکور، مقدار این پارامتر به ازای هر $\mathbf{x}_j \in U$، به صورت زیر محاسبه می‌شود:

$$c_{ij} = \max\{d_1(\mathbf{x}_j) - d(\mathbf{x}_i, \mathbf{x}_j), 0\} \quad (۷.۲)$$

۴. سود حاصل از افزودن $\mathbf{x}_i$ به مجموعه‌ی M که با g_i نمایش داده می‌شود، معادل با میزان کاهش در مقدار تابع هزینه است که از طریق رابطه‌ی زیر به‌دست می‌آید:

$$g_i = \sum_{\mathbf{x}_j \in U} c_{ij} \quad (۸.۲)$$

۵. مناسب‌ترین مشاهده‌ای که می‌تواند از مجموعه‌ی U انتخاب شود و سپس به مجموعه‌ی M، به عنوان مدوید جدید منتقل شود، مشاهده‌ای است که مقدار تابع هزینه را مینیمم کند یا به طور معادل موجب ماکزیمم شدن تابع g گردد، لذا مراحل ۲ تا ۵ برای تمامی مشاهدات متعلق به U انجام می‌شود، سپس مشاهده‌ای که دارای بیشترین مقدار g است، از مجموعه‌ی U حذف شده و به مجموعه‌ی M ملحق می‌گردد. به عبارت دیگر اگر:

$$l = \arg\max_i g_i$$

آنگاه $M = M \cup \{\mathbf{x}_l\}$ و $U = U - \{\mathbf{x}_l\}$.

مراحل فوق تا زمانی تکرار می‌شوند که k مشاهده در مجموعه‌ی M قرار گیرند و بدین صورت مدویدهای اولیه که هریک نماینده‌ی یک خوشه هستند، انتخاب می‌شوند. پس از مشخص شدن مدویدها، سایر مشاهدات به خوشه‌ای نسبت داده می‌شوند که تشابه بیشتری با مدوید آن دارند و بدین ترتیب خوشه‌های اولیه شکل می‌گیرند.

گام جایگزینی

در دومین گام به منظور بهبود مجموعه M، تمامی زوج مشاهدات $(\mathbf{x}_i, \mathbf{x}_h) \in M \times U$ مورد بررسی قرار می‌گیرند. با انتقال $\mathbf{x}_i$ از مجموعه‌ی M به U و همینطور انتقال $\mathbf{x}_h$ از مجموعه‌ی U به M، ممکن است تخصیص نقاط به مدویدها و شکل‌گیری خوشه‌ها به گونه‌ای دیگر صورت گیرد، در نتیجه میزان خطای هر مشاهده و به دنبال آن مقدار تابع هزینه تغییر خواهد کرد. در صورتی که با این جایگزینی، مقدار تابع هزینه کاهش یابد و در نتیجه در کیفیت خوشه‌بندی بهبودی حاصل شود، جایگزینی زوج مشاهده‌ی مذکور اعمال شده و مجموعه‌های M و U به‌روز می‌شوند. در صورتی‌که مقدار تابع هزینه افزایش یابد یا تغییر نکند، مجموعه‌ی M و U به حالت قبل برمی‌گردند و جایگزینی صورت نمی‌گیرد.

برای محاسبه‌ی میزان تغییر در خطای هر مشاهده می‌بایست وضعیت هر مشاهده در مجموعه‌ی U مانند $\mathbf{x}_j$ پس از جایگزینی مورد بررسی قرار گیرد. با حذف $\mathbf{x}_i$ از مجموعه‌ی M و اضافه شدن مشاهده‌ی $\mathbf{x}_h$ به آن، مشاهده‌ی $\mathbf{x}_j$ ممکن است در خوشه‌ی قبلی خودش باقی بماند یا به خوشه‌ی دیگری منتقل شود. اگر عدم تشابه مشاهده‌ی $\mathbf{x}_j$ با نزدیکترین مدوید را پس از جایگزینی با $d_1^{new}(\mathbf{x}_j)$ نمایش دهیم، در این صورت، میزان تغییر در خطای مشاهده‌ی $\mathbf{x}_j$ پس از تخصیص مجدد به صورت زیر محاسبه می‌شود:

$$\Delta d_1(\mathbf{x}_j) = d_1^{new}(\mathbf{x}_j) - d_1(\mathbf{x}_j) \qquad (۹.۲)$$

بر اساس میزان تغییر خطای هر مشاهده، میزان تغییرات تابع هزینه پس از جایگزینی را می‌توان به صورت زیر محاسبه نمود:

$$T_{ih} = \sum_{\mathbf{x}_j \in U} \Delta d_1(\mathbf{x}_j) \qquad (۱۰.۲)$$

منفی بودن مقدار T_{ih} حاکی از کاهش مقدار تابع هزینه و بهبود کیفیت خوشه‌بندی است. روند کار الگوریتم در گام جایگزینی را می‌توان به صورت زیر شرح داد:

۱. زوج مشاهده‌ی $(\mathbf{x}_i, \mathbf{x}_h) \in M \times U$ برای جایگزینی در نظر گرفته می‌شود.

۲. وضعیت سایر مشاهدات مجموعه‌ی U مانند $\mathbf{x}_j$ پس از جایگزینی $\mathbf{x}_i$ و $\mathbf{x}_h$ مورد بررسی قرار می‌گیرد و میزان تغییر در خطای آن محاسبه می‌گردد. این مشاهده ممکن است در یکی از وضعیت‌های زیر قرار گیرد:

(آ) اگر $\mathbf{x}_i$ نزدیک‌ترین مدوید به $\mathbf{x}_j$ باشد یعنی $\mathbf{x}_i$ مدوید خوشه‌ی $\mathbf{x}_j$ باشد، با حذف $\mathbf{x}_i$ از مجموعه‌ی مدویدها، می‌بایست مشاهده‌ی $\mathbf{x}_j$ به دومین مدوید نزدیک به آن اختصاص یابد. با اضافه شدن $\mathbf{x}_h$ به مجموعه‌ی M، ممکن است دومین مدوید نزدیک به $\mathbf{x}_j$ تغییر کند، لذا لازم است که به منظور تخصیص مجدد، وضعیت مشاهده‌ی $\mathbf{x}_j$ نسبت به دومین مدوید نزدیک به آن و همینطور نسبت به مشاهده‌ی $\mathbf{x}_h$ مورد بررسی قرار گیرد. در این وضعیت یکی از دو حالت زیر رخ می‌دهد:

آ-۱) $d(\mathbf{x}_j, \mathbf{x}_h) < d_2(\mathbf{x}_j)$: در این حالت با اضافه شدن مشاهده‌ی $\mathbf{x}_h$ به مجموعه‌ی M، این مشاهده دومین مدوید نزدیک به $\mathbf{x}_j$ محسوب شده و با حذف $\mathbf{x}_i$ از مجموعه‌ی M، نزدیک‌ترین مدوید به آن خواهد بود، لذا میزان تغییر در خطای مشاهده‌ی $\mathbf{x}_j$ را می‌توان به صورت زیر محاسبه کرد:

$$\Delta d_1(\mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_h) - d_1(\mathbf{x}_j)$$

آ-۲) $d(\mathbf{x}_j, \mathbf{x}_h) \geq d_2(\mathbf{x}_j)$: در این حالت با اضافه شدن مشاهده‌ی $\mathbf{x}_h$ به مجموعه‌ی M دومین مدوید نزدیک به $\mathbf{x}_j$ تغییر نمی‌کند و پس از حذف $\mathbf{x}_i$، مشاهده‌ی $\mathbf{x}_j$ به آن اختصاص می‌یابد، بنابراین میزان تغییری که در مقدار خطای مشاهده‌ی $\mathbf{x}_j$ رخ می‌دهد برابر است با:

$$\Delta d_1(\mathbf{x}_j) = d_2(\mathbf{x}_j) - d_1(\mathbf{x}_j)$$

(ب) اگر $\mathbf{x}_i$ نزدیک‌ترین مدوید به مشاهده‌ی $\mathbf{x}_j$ نباشد یعنی مدوید خوشه‌ای که $\mathbf{x}_j$ به آن تعلق دارد، مشاهده‌ای غیر از $\mathbf{x}_i$ باشد، در این صورت حذف $\mathbf{x}_i$ وضعیت $\mathbf{x}_j$ را تحت تاثیر قرار نمی‌دهد، اما با اضافه شدن مشاهده‌ی $\mathbf{x}_h$ به مجموعه‌ی M، عدم تشابه بین $\mathbf{x}_j$ و $\mathbf{x}_h$ در یکی از دو حالت زیر صدق می‌کند:

ب-۱) $d(\mathbf{x}_j, \mathbf{x}_h) < d_1(\mathbf{x}_j)$: در این حالت پس از جایگزینی، مشاهده‌ی $\mathbf{x}_h$ نزدیک‌ترین مدوید مشاهده‌ی $\mathbf{x}_j$ محسوب می‌شود و $\mathbf{x}_j$ به $\mathbf{x}_h$ تعلق می‌گیرد، بنابراین میزان تغییر در خطای مشاهده‌ی $\mathbf{x}_j$ را می‌توان به صورت زیر محاسبه نمود:

$$\Delta d_1(\mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_h) - d_1(\mathbf{x}_j)$$

ب-۲) $d(\mathbf{x}_j, \mathbf{x}_h) > d_1(\mathbf{x}_j)$: در این حالت با ورود $\mathbf{x}_h$ به مجموعه‌ی M، نزدیک‌ترین مدوید $\mathbf{x}_j$ تغییر نمی‌کند و پس از جایگزینی هم $\mathbf{x}_j$ در همان خوشه‌ی قبلی باقی می‌ماند، لذا در این حالت:

$$\Delta d_1(\mathbf{x}_j) = 0$$

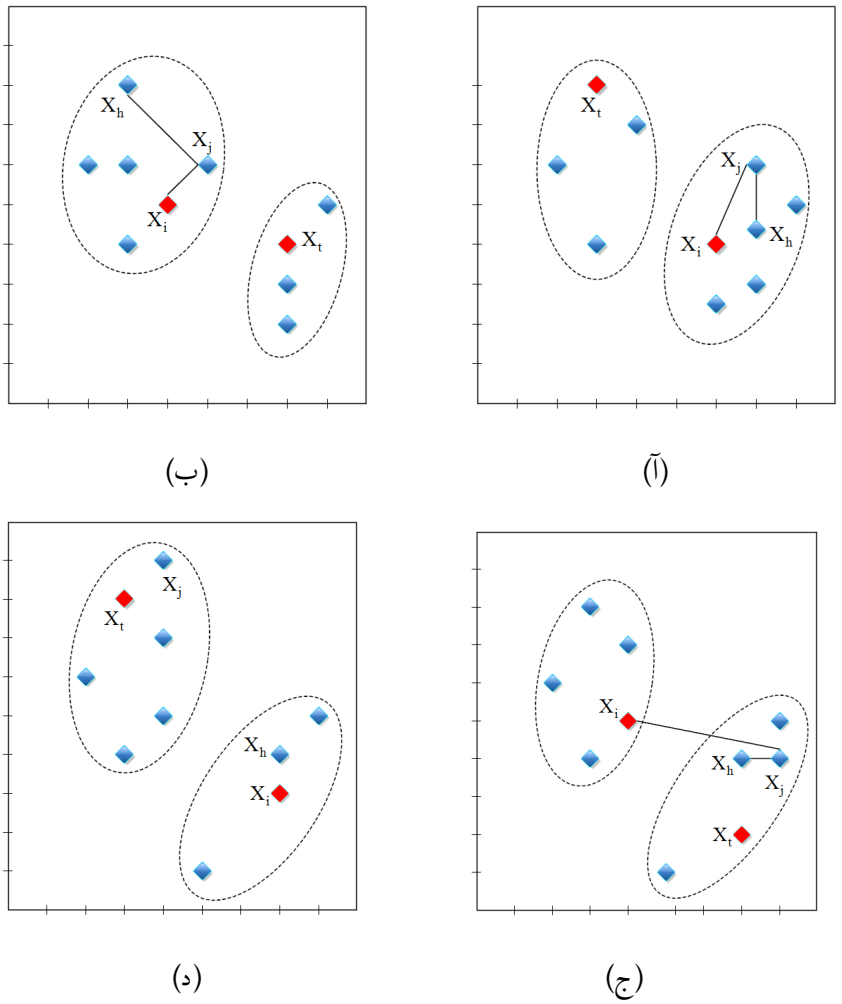
در شکل ۲.۲ مثالی برای هر یک از چهار حالت آ-۱، آ-۲، ب-۱ و ب-۲ نمایش داده شده‌است.

۳. بر اساس میزان تغییر خطای مشاهدات که در مرحله‌ی قبل محاسبه شد، میزان تغییرات تابع هزینه طبق رابطه‌ی (۱۰.۲) محاسبه می‌شود.

۴. مراحل ۲ و ۳ برای تمامی زوج مشاهدات $(\mathbf{x}_i, \mathbf{x}_h) \in M \times U$ انجام می‌شود و از بین آن‌ها زوجی که دارای کمترین مقدار T است، انتخاب می‌گردد. به عبارتی زوج مشاهده‌ی $(\mathbf{x}_p, \mathbf{x}_q) \in M \times U$ انتخاب می‌شود اگر:

$$(\mathbf{x}_p, \mathbf{x}_q) = \underset{i,h}{\operatorname{argmin}}\, T_{ih}$$

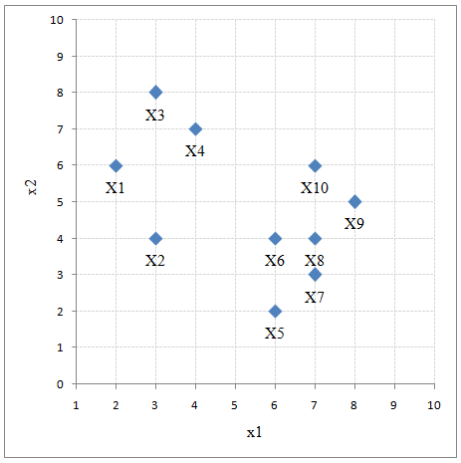
۵. اگر تغییرات تابع هزینه زوج $(\mathbf{x}_p, \mathbf{x}_q)$ منفی باشد، پس از جایگزینی آن‌ها مقدار تابع هزینه کمتر شده و در نتیجه خوشه‌بندی بهتری صورت خواهد گرفت، لذا اگر $T_{pq} < 0$ مدوید $\mathbf{x}_p$ با مشاهده‌ی



شکل ۲.۲: (آ) : نمایش مثالی از موقعیت (آ-۱)، (ب) : (آ-۲)، (ج) : (ب-۱) و (د) : (ب-۲) در گام جایگزینی. $\mathbf{x}_i$ و $\mathbf{x}_t$ مدویدهای اولیه هستند و $\mathbf{x}_h$ مشاهده‌ای است که جایگزینی آن با $\mathbf{x}_i$ تحت بررسی است.

$\mathbf{x}_q$ جابجا شده و پس از به‌روز رسانی پارامترهای d_1 و d_2 برای هر مشاهده مجموعه‌ی U به مرحله‌ی ۱ بازمی‌گردیم. اما در صورتی که $T_{pq} > 0$ امکان بهبود کیفیت خوشه‌بندی وجود ندارد و الگوریتم متوقف می‌شود [۲۹].

الگوریتم PAM در یک فرآیند تکراری در صدد یافتن مجموعه مدویدی است که کمترین مقدار مجموع خطای مطلق و به دنبال آن بهترین خوشه‌بندی را در پی داشته باشد، از این رو تمامی داده‌ها را برای جانشینی با مدویدهای مجموعه M آزمایش می‌کند. بنابراین حتی اگر مدویدهای اولیه انتخاب‌های مناسبی نباشند، این الگوریتم منجر به خوشه‌های بهینه کلی خواهد شد. این در حالی است که کیفیت خوشه‌های حاصل از الگوریتم k-میانگین کاملاً به انتخاب مراکز اولیه وابسته است و اصطلاحاً الگوریتم k-میانگین به یافتن خوشه‌های بهینه محلی منجر می‌شود. هر چند ممکن است در نگاه اول این مساله حاکی از برتری روش PAM نسبت به الگوریتم k-میانگین باشد، اما آزمودن تمامی داده‌ها به عنوان



شکل ۳.۲: نمودار پراکنش نقاط مثال ۱.۲

مدوید در این الگوریتم، هزینه محاسباتی بیشتری را به‌ویژه در مواجهه با مقادیر بزرگ n و k، در پی خواهد داشت، به همین دلیل استفاده از الگوریتم PAM در مورد داده‌هایی با حجم بالا توصیه نمی‌شود. در این مواقع می‌توان از روشی به نام خوشه‌بندی کاربردهای بزرگ [۱] (CLARA) که تعمیمی از الگوریتم PAM است، استفاده نمود [۲۲].

مثال ۱.۲. فرض کنید قصد داریم توسط الگوریتم خوشه‌بندی PAM، ۱۰ نقطه دو بعدی را که در شکل ۳.۲ نمایش داده شده‌اند، به ۲ خوشه تفکیک کنیم. برای اجرای الگوریتم PAM به ماتریس عدم تشابه دو به دوی نقاط نیاز داریم. در این مثال، استفاده از فاصله منهتن، ماتریس عدم تشابه زیر را نتیجه داده است:

$$D = \begin{bmatrix} ۰ & ۳ & ۳ & ۳ & ۸ & ۶ & ۸ & ۷ & ۷ & ۵ \\ ۳ & ۰ & ۴ & ۴ & ۵ & ۳ & ۵ & ۴ & ۶ & ۶ \\ ۳ & ۴ & ۰ & ۲ & ۹ & ۷ & ۹ & ۸ & ۸ & ۶ \\ ۳ & ۴ & ۲ & ۰ & ۷ & ۵ & ۷ & ۶ & ۶ & ۴ \\ ۸ & ۵ & ۹ & ۷ & ۰ & ۲ & ۲ & ۳ & ۵ & ۵ \\ ۶ & ۳ & ۷ & ۵ & ۲ & ۰ & ۲ & ۱ & ۳ & ۳ \\ ۸ & ۵ & ۹ & ۷ & ۲ & ۲ & ۰ & ۱ & ۳ & ۳ \\ ۷ & ۴ & ۸ & ۶ & ۳ & ۱ & ۱ & ۰ & ۲ & ۲ \\ ۷ & ۶ & ۸ & ۶ & ۵ & ۳ & ۳ & ۲ & ۰ & ۲ \\ ۵ & ۶ & ۶ & ۴ & ۵ & ۳ & ۳ & ۲ & ۲ & ۰ \end{bmatrix}$$

[۱] Clustering large applications

گام اول، تعیین مدویدهای اولیه است. اولین مدوید، مشاهده‌ای است که مجموع فواصلش با سایر مشاهدات مینیمم باشد. برای پیدا کردن این مشاهده پارامتر sd برای همه مشاهدات محاسبه شده و در جدول ۱.۲ آمده است.

جدول ۱.۲: مقدار پارامتر sd برای مشاهدات مثال۱.۲

j	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
sd_j	۵۰	۴۰	۵۶	۴۴	۴۶	۳۲	۴۰	۳۴	۴۲	۳۶

طبق مقادیر جدول ۱.۲ اولین مدوید، مشاهده‌ی $\mathbf{x}_۶$ است پس $M = \{\mathbf{x}_۶\}$ و $U = \mathbf{X} - \{\mathbf{x}_۶\}$. در مرحله بعد، باید عضو دیگری از مجموعه U به مجموعه M بپیوندد. با محاسبه‌ی مقدار g_i برای هر $\mathbf{x}_i \in U$ دومین مدوید تعیین می‌شود. مشاهده‌ی $\mathbf{x}_۱ \in U$ را در نظر بگیرید. برای محاسبه مقدار تابع $g_۱$ نیاز است مقادیر $c_{۱j}$ برای هر $\mathbf{x}_j \in U - \{\mathbf{x}_۱\}$ محاسبه شود. محاسبه این مقادیر در جدول ۲.۲ صورت گرفته‌است.

جدول ۲.۲: مقدار پارامتر $c_{۱j}$ برای مشاهدات مثال۱.۲

j	$d_۱(\mathbf{x}_j)$	$d(\mathbf{x}_۱, \mathbf{x}_j)$	$d_۱(\mathbf{x}_j) - d(\mathbf{x}_۱, \mathbf{x}_j)$	$c_{۱j}$
۲	۳	۳	۰	۰
۳	۷	۳	۴	۴
۴	۵	۳	۲	۲
۵	۲	۸	-۶	۰
۷	۲	۸	-۶	۰
۸	۱	۷	-۶	۰
۹	۳	۷	-۴	۰
۱۰	۳	۵	-۲	۰

طبق مقادیر جدول ۲.۲، مقدار تابع $g_۱$ به صورت زیر محاسبه می‌شود:

$$g_۱ = \sum_{\mathbf{x}_j \in U} c_{۱j} = ۶$$

به همین ترتیب برای سایر اعضای مجموعه U محاسبه می‌شود. مقدار تابع g مربوط به اعضای U در جدول ۳.۲ نمایش داده است.

جدول ۳.۲: مقدار پارامتر g برای مشاهدات مثال۱.۲.

j	۱	۲	۳	۴	۵	۷	۸	۹	۱۰
g_j	۶	۷	۳	۸	۰	۰	۳	۱	۴

عضوی از U با بیشترین مقدار g به مجموعه M می‌پیوندد، لذا مشاهده‌ی $\mathbf{x}_۴$ از مجموعه U به مجموعه M منتقل می‌شود و مجموعه‌های M و U جدید عبارتند از: $M = \{\mathbf{x}_۴, \mathbf{x}_۶\}$ و $U = \mathbf{X} - \{\mathbf{x}_۴, \mathbf{x}_۶\}$. بدین ترتیب گام اول به پایان می‌رسد.

در گام دوم، تمامی زوج مشاهدات $(\mathbf{x}_i, \mathbf{x}_j) \in M \times U$ برای جابجایی، مورد بررسی قرار می‌گیرند. به منظور تصمیم‌گیری در مورد جایگزینی این دو مشاهده لازم است ابتدا مقدار پارامتر $\Delta d_۱(\mathbf{x}_j)$ به ازای هر $\mathbf{x}_j \in U - \{\mathbf{x}_h\}$ محاسبه شود. به عنوان مثال زوج مشاهده‌ی $(\mathbf{x}_۴, \mathbf{x}_۷)$ را در نظر بگیرید. مقادیر $\Delta d_۱(\mathbf{x}_j)$ برای هر $\mathbf{x}_j \in U - \{\mathbf{x}_۷\}$ محاسبه شده و در جدول ۴.۲ داده شده است.

جدول ۴.۲: مقدار پارامتر $\Delta d_۱(\mathbf{x}_j)$ برای مشاهدات مثال۱.۲

j	$d(\mathbf{x}_j, \mathbf{x}_۶)$	$d(\mathbf{x}_j, \mathbf{x}_۴)$	$d(\mathbf{x}_j, \mathbf{x}_۷)$	$d_۱(\mathbf{x}_j)$	$d_۲(\mathbf{x}_j)$	$\Delta d_۱(\mathbf{x}_j)$
۱	۶	۳	۸	۳	۶	۳
۲	۳	۴	۵	۳	۴	۰
۳	۷	۲	۹	۲	۷	۵
۵	۷	۲	۲	۷	۲	۵
۸	۶	۱	۱	۶	۱	۸
۹	۶	۱	۱	۶	۳	۹
۱۰	۶	۳	۳	۴	۳	۱۰

بر اساس مقادیر $\Delta d_۱(\mathbf{x}_j)$ محاسبه شده، مقدار تابع $T_{۴۷}$ به صورت زیر خواهد بود:

$$T_{۴۷} = \sum_{\mathbf{x}_j \in U} \Delta d_۱(\mathbf{x}_j) = ۸$$

به همین ترتیب مقدار تابع T_{ih} برای تمامی زوج مشاهدات $(\mathbf{x}_i, \mathbf{x}_h) \in M \times U$ محاسبه شده و در نهایت زوج مشاهده‌ای با مینیمم مقدار T انتخاب می‌گردد. اگر مقدار مینیمم T منفی بود، دو مشاهده‌ی انتخاب شده جابجا شده، مجموعه‌های M و U به‌روز می‌شوند. پس از به‌روزرسانی مجموعه‌های مذکور، با تکرار گام دو، برای بهبود آن‌ها تلاش می‌شود. در صورتی‌که مقدار مینیمم T مثبت باشد ، الگوریتم متوقف خواهد شد (مختصات مشاهدات در این مثال از درگاه http://en.wikipedia.org/wiki/K-medoids استخراج شده‌است).

۳.۲ ارزیابی خوشه‌بندی

در بخش ۲.۱ به طور مختصر در مورد معیارهای ارزیابی خوشه‌بندی توضیح داده شد. لازم به یادآوری است که معیارهای ارزیابی خوشه‌بندی به دو دسته معیارهای درونی و بیرونی تقسیم می‌شوند. در این بخش، ابتدا یکی از شاخص‌های بیرونی ارزیابی خوشه‌بندی به نام شاخص رند [۱] را معرفی می‌کنیم و سپس با ذکر مشکل آن، حالت تعدیل‌شده‌ی آن را توضیح می‌دهیم. همچنین شاخصی درونی تحت عنوان شاخص نیمرخ [۲] را معرفی خواهیم نمود.

۱.۳.۲ شاخص رند

شاخص رند یکی از معیارهای بیرونی خوشه‌بندی است که توسط رند معرفی شد [۲۸]. این شاخص، میزان تطابق خوشه‌های به‌دست آمده از الگوریتم خوشه‌بندی (خوشه‌های تخمینی) و خوشه‌های واقعی را می‌سنجد. فرض کنید $\mathbf{X}$ مجموعه داده معرفی شده در بخش ۳.۱ باشد و افرازهای C و B به ترتیب شامل خوشه‌های حاصل از انجام الگوریتم خوشه‌بندی و خوشه‌های واقعی (رده‌ها) باشند. با فرض اینکه $C = \{C_۱, C_۲, \cdots, C_s\}$ و $B = \{B_۱, B_۲, \cdots, B_r\}$ که s و r بیانگر تعداد خوشه‌ها در هر مجموعه هستند، داریم:

$$\cup_{i=۱}^{r} B_i = \mathbf{X} = \cup_{j=۱}^{s} C_j$$

و

$$C_i \cap C_{i'} = \emptyset = B_j \cap B_{j'} \quad ; \quad ۱ \leq i \neq i' \leq s \quad , \quad ۱ \leq j \neq j' \leq r$$

فرض کنید فراوانی مشاهدات در خوشه‌های تخمینی و واقعی به صورت جدول ۵.۲ که در آن n_{ij} بیانگر تعداد اعضای مشترک بین رده‌ی B_i و خوشه‌ی تخمینی C_j می‌باشد. مقدار $n_{i.}$ برابر مجموع مقادیر سطر iام یا تعداد اعضای خوشه‌ی واقعی B_i و همچنین مقدار $n_{.j}$ مجموع مقادیر ستون jام یا تعداد اعضای خوشه‌ی C_j در نظر گرفته شده‌است.

اکنون کمیت‌های a، b، c و d را بدین صورت تعریف می‌کنیم:

---

[۱] Rand

[۲] Silhouette index

جدول ۵.۲: فراوانی مشاهدات در خوشه‌های تخمینی و رده‌های واقعی

		افراز C				
		$C_۱$	$C_۲$	$\cdots$	C_s	مجموع
افراز B	$B_۱$	$n_{۱۱}$	$n_{۱۲}$	$\cdots$	$n_{۱s}$	$n_{۱.}$
	$B_۲$	$n_{۲۱}$	$n_{۲۲}$	$\cdots$	$n_{۲s}$	$n_{۲.}$
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	B_r	$n_{r۱}$	$n_{r۲}$	$\cdots$	n_{rs}	$n_{r.}$
	مجموع	$n_{.۱}$	$n_{.۲}$	$\cdots$	$n_{.s}$	$n_{..} = n$

a : تعداد جفت مشاهداتی که در مجموعه B متعلق به یک رده هستند و طی خوشه‌بندی هم درون یک خوشه قرار گرفته‌اند. کمیت a را توسط نمادهای جدول ۵.۲ می‌توان به صورت زیر تعریف نمود:

$$a = \sum_{i=۱}^{r} \sum_{j=۱}^{s} \binom{n_{ij}}{۲}$$

b : تعداد جفت مشاهداتی که دورن یک رده قرار دارند، اما در خوشه‌های مختلف واقع شده‌اند و به صورت زیر قابل محاسبه است:

$$b = \sum_{i=۱}^{r} \binom{n_{i.}}{۲} - \sum_{j=۱}^{r} \sum_{j=۱}^{s} \binom{n_{ij}}{۲}$$

c : تعداد جفت مشاهداتی که طی خوشه‌بندی به یک خوشه نسبت داده شده‌اند، اما متعلق به یک رده نیستند. کمیت c بنا به جدول ۵.۲ به صورت زیر تعریف می‌شود:

$$c = \sum_{j=۱}^{s} \binom{n_{.j}}{۲} - \sum_{j=۱}^{r} \sum_{j=۱}^{s} \binom{n_{ij}}{۲}$$

d : تعداد جفت مشاهداتی که در رده‌های متفاوت قرار دارند و در خوشه‌های متفاوت نیز واقع شده‌اند. کمیت d به صورت زیر به‌دست می‌آید:

$$d = \binom{n}{۲} - a - b - c$$

اگر $m_۱$ را تعداد جفت مشاهداتی در نظر بگیریم که متعلق به رده یکسانی هستند، آنگاه با توجه به تعریف کمیت‌های فوق، $m_۱$ برابر است با:

$$m_۱ = a + b$$

همچنین اگر تعداد جفت مشاهداتی که طی الگوریتم خوشه‌بندی درون خوشه‌ی یکسان واقع شده‌اند را $m_۲$ بنامیم آنگاه $m_۲$ به صورت زیر قابل محاسبه است:

$$m_۲ = a + c$$

تعداد کل جفت مشاهدات را M در نظر می‌گیریم که از طریق زیر محاسبه می‌شود:

$$M = \binom{n}{۲} = a + b + c + d$$

اکنون می‌توان به کمک کمیت‌های تعریف شده، جدول ۵.۲ را در جدول ۶.۲ خلاصه نمود.

جدول ۶.۲: فراوانی جفت مشاهدات

	واقع شده در خوشه‌های متفاوت	واقع شده در خوشه‌ی یکسان	
واقع شده در رده‌ی یکسان	b	a	$m_۱$
واقع شده در رده‌های متفاوت	d	c	$M - m_۱$
	$M - m_۲$	$m_۲$	M

بنابر جدول ۶.۲ تعداد جفت مشاهداتی که موقعیت قرار گرفتن آن‌ها در خوشه‌های تخمینی منطبق بر واقعیت است برابر با $a + d$ است و همچنین $b + c$ حاکی از تعداد جفت مشاهداتی است که خوشه‌ی آن‌ها به درستی تشخیص داده نشده است. مقدار $a + d$ به عنوان میزان تطابق خوشه‌های واقعی و تخمینی و مقدار $b + c$ به عنوان میزان عدم تطابق آن‌ها تفسیر می‌شود. شاخص رند که میزان تطابق خوشه‌های تخمینی و رده‌های واقعی را می‌سنجد به صورت نسبت میزان تطابق، تعریف شده و از طریق رابطه‌ی زیر محاسبه می‌شود:

$$RI = \frac{a + d}{M} \qquad (۱۱.۲)$$

مقدار این شاخص در بازه‌ی $[۰, ۱]$ خواهد بود. اگر خوشه‌های تشخیص داده شده توسط الگوریتم خوشه‌بندی، کاملاً منطبق بر رده‌های واقعی باشد، مقدار این شاخص برابر با ۱ است مقادیر نزدیک به صفر عدم تطابق خوشه‌های تخمینی و واقعی را نشان می‌دهد. اگرچه حداقل مقدار شاخص رند عدد صفر است اما این مقدار در مورد داده‌های واقعی به ندرت صفر می‌شود [۳۴].

مثال ۲.۲. مجموعه داده‌ی $\mathbf{X} = \{\mathbf{x}_۱, \mathbf{x}_۲, \cdots, \mathbf{x}_{۱۰}\}$ را در نظر بگیرید. فرض کنید این مجموعه داده شامل دو خوشه‌ی زیر باشد:

$$B_۱ = \{\mathbf{x}_۱, \mathbf{x}_۲, \mathbf{x}_۳, \mathbf{x}_۴, \mathbf{x}_۵, \mathbf{x}_۶\} \quad و \quad B_۲ = \{\mathbf{x}_۷, \mathbf{x}_۸, \mathbf{x}_۹, \mathbf{x}_{۱۰}\}$$

اگر مشاهدات مذکور به طور تصادفی با اعداد ۱ و ۲ و مطابق جدول ۷.۲ برچسب‌گذاری شوند آنگاه از این طریق، دو خوشه‌ی تصادفی به صورت زیر ایجاد می‌شود:

$$C_۱ = \{\mathbf{x}_۳, \mathbf{x}_۵, \mathbf{x}_۶, \mathbf{x}_۸, \mathbf{x}_۹\} \quad و \quad C_۲ = \{\mathbf{x}_۱, \mathbf{x}_۲, \mathbf{x}_۴, \mathbf{x}_۷, \mathbf{x}_{۱۰}\}$$

جدول ۷.۲: خوشه‌ی تعیین شده برای هر مشاهده طی برچسب‌گذاری تصادفی

مشاهده	$\mathbf{x}_۱$	$\mathbf{x}_۲$	$\mathbf{x}_۳$	$\mathbf{x}_۴$	$\mathbf{x}_۵$	$\mathbf{x}_۶$	$\mathbf{x}_۷$	$\mathbf{x}_۸$	$\mathbf{x}_۹$	$\mathbf{x}_{۱۰}$
خوشه	۲	۲	۱	۲	۱	۱	۲	۱	۱	۲

اکنون قصد داریم با محاسبه‌ی شاخص رند، خوشه‌بندی انجام شده را مورد ارزیابی قرار دهیم. فراوانی مشاهدات در خوشه‌های واقعی و تخمینی در جدول ۸.۲ درج شده‌است.

جدول ۸.۲: مربوط به مثال ۲.۲. فراوانی مشاهدات در خوشه‌های تخمینی و رده‌های واقعی

		افراز C		
		$C_۱$	$C_۲$	مجموع
افراز B	$B_۱$	۳	۳	۶
	$B_۲$	۲	۲	۴
	مجموع	۵	۵	۱۰

طبق مقادیر جدول ۸.۲ کمیت‌های a، b، c و d به صورت زیر محاسبه می‌شود:

$$a = \sum_{i=۱}^{۲}\sum_{j=۱}^{۲}\binom{n_{ij}}{۲} = \binom{۳}{۲} + \binom{۳}{۲} + \binom{۲}{۲} + \binom{۲}{۲} = ۸$$

$$b = \sum_{i=۱}^{۲}\binom{n_{i.}}{۲} - \sum_{j=۱}^{۲}\sum_{j=۱}^{۲}\binom{n_{ij}}{۲} = \binom{۶}{۲} + \binom{۴}{۲} - ۸ = ۱۳$$

$$c = \sum_{j=۱}^{۲} \binom{n_{.j}}{۲} - \sum_{j=۱}^{۲}\sum_{j=۱}^{۲} \binom{n_{ij}}{۲} = \binom{۵}{۲} + \binom{۵}{۲} - ۸ = ۱۲$$
$$d = \binom{۱۰}{۲} - a - b - c = ۴۵ - ۸ - ۱۳ - ۱۲ = ۱۲$$

اکنون طبق رابطه‌ی (۱۱.۲)، شاخص رند به صورت زیر محاسبه می‌شود:

$$RI = \frac{۱۰}{۱۵} = ۰/۴۴$$

همان‌گونه که در مثال فوق ملاحظه می‌شود، حتی زمانی که مشاهدات به طور تصادفی برچسب‌گذاری شوند، مقدار شاخص رند صفر یا نزدیک به صفر نخواهد شد زیرا در محاسبه‌ی این شاخص، حالتی که خوشه‌های تخمینی به طور اتفاقی بر رده‌های واقعی منطبق شوند، در نظر گرفته نشده است. به اصطلاح گفته می‌شود که شاخص رند برای وقوع تصادفی تطابق، تصحیح نشده است.

۲.۳.۲ شاخص رند تعدیل یافته

شاخص رند تعدیل یافته [۱] (ARI) حالت تصحیح شده‌ی شاخص رند است که توسط هوبرت [۲] و ارابی [۳] معرفی شد [۲۵]. به طور کلی می‌توان آماره‌ی T را با در نظر گرفتن شانس به صورت زیر تصحیح نمود:

$$T' = \frac{T - E(T)}{Max(T) - E(T)} \qquad (۱۲.۲)$$

که در آن $Max(T)$ حداکثر مقداری است که آماره T می‌تواند اختیار کند و $E(T)$ مقدار مورد انتظار T است.

می‌توان نشان داد که [۲۵]:

$$E(a) = \frac{m_۱ m_۲}{M}$$
$$E(d) = \frac{(M - m_۱)(M - m_۲)}{M}$$

[۱] Adjusted rand index

[۲] Hubert

[۳] Arabie

بنابراین مقدار مورد انتظار شاخص رند برابر است با:

$$\begin{aligned} E(Rand) &= E(\frac{a+d}{M}) \\ &= \frac{m_۱ m_۲/M + (M-m_۱)(M-m_۲)/M}{M} \\ &= \frac{m_۱ m_۲ + (M-m_۱)(M-m_۲)}{M^۲} \end{aligned}$$

در نتیجه شاخص رند تعدیل یافته به صورت زیر تعریف می‌شود:

$$\begin{aligned} ARI &= \frac{(a+d)/M - E((a+d)/M)}{۱ - E((a+d)/M)} \\ &= \frac{a - (m_۱ m_۲)/M}{(m_۱ + m_۲)/۲ - (m_۱ m_۲)/M} \end{aligned}$$

اگر مقدار شاخص رند برابر با مقدار مورد انتظار باشد، مقدار شاخص رند تعدیل‌یافته برابر با صفر است و اگر شاخص رند مقدار ۱ را اختیار کند مقدار شاخص رند تعدیل‌یافته هم برابر ۱ خواهد بود. شاخص رند تعدیل یافته، مقداری در بازه‌ی $[-۱, ۱]$ را اختیار می‌کند. زمانی‌که خوشه‌های تخمینی به طورکامل بر رده‌های واقعی منطبق باشند مقدار شاخص رند تعدیل یافته برابر ۱ است و مقادیر صفر و منفی مربوط به حالتی است که تطابق آن‌ها به صورت تصادفی رخ داده باشد.

مثال ۳.۲. بار دیگر خوشه‌های تخمینی و واقعی را در مثال ۲.۲ در نظر بگیرید، شاخص رند تعدیل‌یافته برای این مثال به صورت زیر محاسبه می‌شود:

$$m_۱ = a + b = ۲۱$$

$$m_۲ = a + c = ۲۰$$

$$ARI = \frac{۸ - ۲۰(۲۱)/۴۵}{۲۰ + ۲۱/۲ - ۲۰(۲۱)/۴۵} = -۰/۱۱$$

مقدار شاخص رند تعدیل یافته نشان می‌دهد که خوشه‌بندی، ضعیف بوده و تطابق خوشه‌های تخمینی و واقعی به صورت تصادفی رخ داده‌است.

۳.۳.۲ شاخص نیمرخ

شاخص نیمرخ یکی از معیارهای درونی خوشه‌بندی است که برای ارزیابی نتایج خوشه‌بندی، نیازی به اطلاع از خوشه‌های واقعی ندارد. این شاخص برای اولین بار توسط روسیو [۱] معرفی شد [۴۰]. شاخص نیمرخ را می‌توان ابتدا برای هر مشاهده، سپس برای هر خوشه و در نهایت به صورت جامع برای یک مسئله‌ی خوشه‌بندی تعریف نمود. طبق نمادگذاری قبلی، مجموعه‌ی $C = \{C_۱, C_۲, \cdots, C_k\}$ را که حاوی خوشه‌های تشخیص داده شده توسط الگوریتم خوشه‌بندی است، در نظر بگیرید. فرض کنید مشاهده‌ی $\mathbf{x}_i$ به خوشه‌ی C_l نسبت داده شده باشد، $a(\mathbf{x}_i)$ را میانگین عدم تشابه مشاهده‌ی $\mathbf{x}_i$ با مشاهدات هم‌خوشه‌اش تعریف می‌کنیم که به صورت زیر قابل محاسبه است:

$$a(\mathbf{x}_i) = \frac{۱}{n_l - ۱} \sum_{\substack{\mathbf{x}_j \in C_l \\ j \neq i}} d(\mathbf{x}_i, \mathbf{x}_j) \qquad (۱۳.۲)$$

همچنین فرض کنید میانگین عدم تشابه مشاهده‌ی $\mathbf{x}_i$ با مشاهدات خوشه‌ی دیگری مانند $C_{l'}$ به صورت زیر تعریف شود:

$$\eth(\mathbf{x}_i, C_{l'}) = \frac{۱}{n_{l'}} \sum_{X_j \in C_{l'}} d(\mathbf{x}_i, \mathbf{x}_j) \qquad (۱۴.۲)$$

اگر کمیت $b(\mathbf{x}_i)$ را به صورت زیر در نظر بگیریم:

$$b(\mathbf{x}_i) = \min_{l' \neq l} \eth(\mathbf{x}_i, C_{l'}) \qquad (۱۵.۲)$$

شاخص نیمرخ مشاهده‌ی $\mathbf{x}_i$ عبارت است از:

$$S(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}} \qquad (۱۶.۲)$$

شاخص نیمرخ می‌تواند مقداری در بازه‌ی $[-۱, ۱]$ را اختیار کند. هر چه $b(\mathbf{x}_i)$ بزرگ‌تر باشد، میزان عدم تشابه مشاهده‌ی $\mathbf{x}_i$ با مشاهدات سایر خوشه‌ها بیشتر است و هر چه $a(\mathbf{x}_i)$ کوچک‌تر باشد، مشاهده‌ی $\mathbf{x}_i$ تشابه بیشتری با مشاهدات هم‌خوشه‌اش دارد، لذا هر چه اختلاف بین $b(\mathbf{x}_i)$ و $a(\mathbf{x}_i)$ بیشتر و در نتیجه $S(\mathbf{x}_i)$ به عدد ۱ نزدیکتر باشد، خوشه‌ی حاوی مشاهده‌ی $\mathbf{x}_i$ متراکم‌تر و از سایر خوشه‌ها مجزاتر است. لذا نزدیک بودن مقدار $S(\mathbf{x}_i)$ به عدد ۱ مطلوب بودن خوشه‌ای را نشان می‌دهد که $\mathbf{x}_i$ به آن تعلق دارد. میانگین شاخص نیمرخ خوشه‌ی C_l، میانگین شاخص نیمرخ مشاهدات آن خوشه است که به صورت زیر تعریف می‌شود:

$$S(C_l) = \frac{۱}{n_l} \sum_{X_i \in C_l} S(\mathbf{x}_i) \qquad (۱۷.۲)$$

[۱] Rousseeuw

نزدیک بودن مقدار $S(C_l)$ به عدد ۱ نشان می‌دهد که شاخص نیمرخ اکثر مشاهدات خوشه‌ی C_l به عدد ۱ نزدیک بوده و مطلوب بودن این خوشه توسط اکثریت اعضای آن تصدیق شده است. اکنون اگر شاخص نیمرخ سایر خوشه‌ها نیز به عدد ۱ نزدیک باشد، می‌توان نتیجه گرفت که مدل خوشه‌بندی ایجاد شده مناسب است. از این رو به منظور سنجش کیفیت خوشه‌بندی از میانگین شاخص نیمرخ تمامی خوشه‌ها استفاده می‌شود که به صورت زیر قابل محاسبه است:

$$S_k = \frac{۱}{k} \sum_{l=۱}^{k} S(C_l) \qquad (۱۸.۲)$$

بنابر آنچه گفته شد، نزدیک بودن مقدار S_k به عدد ۱ حاکی از رضایت‌بخش بودن مدل خوشه‌بندی است و منفی بودن این شاخص، عملکرد ضعیف الگوریتم خوشه‌بندی در ایجاد خوشه‌ها را نشان می‌دهد.

۴.۲ مقیاس‌گذاری چند بعدی

مقیاس‌گذاری چند بعدی (MDS) یک عنوان کلی برای دسته‌ای از روش‌های آماری است که به منظور تجسم اکتشافی داده‌ها به‌کار می‌روند. این روش، داده‌ها را در بعد کوچک‌تر از بعد واقعی آن‌ها پیکربندی می‌کند به گونه‌ای که مجاورت بین آن‌ها حفظ شود. MDS قادر است نمایشی از داده‌ها را روی یک نقشه فراهم کند به طوری‌که حاوی اطلاعات مفیدی از روابط بین داده‌ها باشد و امکان تعبیر بهتر و کشف ساختارهای نهفته داده‌ها را تحقق بخشد. این روش می‌تواند به عنوان یک روش مستقل در خوشه‌بندی مورد استفاده قرار گیرد. همچنین بنا به دلایلی که در بخش ۶.۱ اشاره شد، می‌توان آن‌را پیش از سایر روش‌های خوشه‌بندی اجرا نمود. بر اساس نحوه‌ی تجزیه و تحلیل داده‌های مجاورت، می‌توان برای روش‌های مقیاس‌گذاری چند بعدی تقسیم‌بندی‌های متفاوتی را انجام داد [۱۸]. برای کسب اطلاعات بیشتر در این زمینه می‌توانید به مرجع [۱۴] مراجعه کنید. اگر کمی یا کیفی بودن داده‌های مجاورت را معیار این طبقه‌بندی قرار دهیم مقیاس‌گذاری چند بعدی به دو دسته کلاسیک و غیرمتری [۱] تقسیم می‌شود. در روش کلاسیک، پیکره مشاهدات را به صورت مستقیم از اندازه‌های ماتریس عدم تشابه یعنی درایه‌های ماتریس D به‌دست می‌آورند. در حالی‌که در روش‌های غیرمتری، رتبه‌بندی عدم تشابه مشاهدات، در ساخت پیکره مد نظر قرار می‌گیرد.

هدف MDS پیدا کردن مختصات مشاهدات (پیکره) در فضای اقلیدسی q بعدی ($q \leq p$) است به طوری‌که فاصله مشاهدات در این پیکره تطبیق خوبی برای عدم تشابه داده شده، ارائه دهد. علاوه بر مختصات مشاهدات، بعد q نیز مجهول است و می‌تواند ۱، ۲، ۳ و یا حتی بیشتر باشد. در ادامه به توضیح روش‌های مقیاس‌گذاری چندبعدی کلاسیک و غیرمتری می‌پردازیم.

[۱] nonmetric

۱.۴.۲ مقیاس‌گذاری چند بعدی کلاسیک

فرض کنید مختصات چندین شهر روی نقشه را در اختیار داریم. فاصله بین آن‌ها به راحتی بوسیله یکی از معیارهای سنجش عدم تشابه که در بخش ۳.۱ توضیح داده شد قابل محاسبه است. عکس این مساله را در نظر بگیرید، اگر تنها ماتریس فاصله دو به دوی شهرها را در اختیار داشته باشیم، آیا می‌توان مختصات هر یک از آن‌ها را پیدا کرد؟

مقیاس گذاری چند بعدی کلاسیک که اولین بار توسط تورگرسون [۱] مطرح شد [۴۷]، به حل این مساله می‌پردازد. ماتریس عدم تشابه $D = [d(\mathbf{x}_i, \mathbf{x}_j)]$ را در نظر بگیرید. هدف از مقیاس‌گذاری چند بعدی کلاسیک، پیدا کردن مختصات n مشاهده در بعد q است به طوری‌که ماتریس فواصل اقلیدسی آن‌ها که با $\Delta = [\delta(\mathbf{x}_i, \mathbf{x}_j)]$ نمایش می‌دهیم، تا حد ممکن به ماتریس D نزدیک باشد. به عبارتی:

$$\forall i, j \qquad d(\mathbf{x}_i, \mathbf{x}_j) \cong \delta(\mathbf{x}_i, \mathbf{x}_j) \tag{۱۹.۲}$$

که در آن $\delta(\mathbf{x}_i, \mathbf{x}_j)$ بیانگر فاصله اقلیدسی دو مشاهده $\mathbf{x}_i$ و $\mathbf{x}_j$ در فضای تقلیل یافته q بعدی است.

الگوریتم مقیاس‌گذاری کلاسیک، پیکره‌ای از نقاط را در بعد q به صورت زیر به‌دست می‌آورد [۳۹، ۴۶]:

۱. ماتریس A بر اساس ماتریس عدم تشابه D و به صورت زیر ساخته می‌شود:
$$A = (a_{ij}) = d^{۲}(\mathbf{x}_i, \mathbf{x}_j)$$

۲. ماتریس B طبق رابطه‌ی زیر تشکیل می‌شود:
$$B = -\frac{۱}{۲} JAJ$$
که در آن $J = I - \frac{۱}{n}۱۱'$.

۳. مقادیر ویژه ماتریس B و بردارهای ویژه متناظر با آن‌ها محاسبه می‌شود.

۴. از بین مقادیر ویژه ماتریس B و بردارهای متناظر آن‌ها، q مقدار ویژه بزرگ‌تر یعنی $\lambda_۱ > \lambda_۲ > \cdots > \lambda_q$ و بردارهای ویژه متناظر با آن‌ها در نظر گرفته می‌شود. ممکن است تعدادی از مقادیر ویژه منفی باشند که در این صورت از آن‌ها صرف نظر می‌کنیم.

[۱] Torgerson

۵. سر انجام، مختصات نقاط در پیکره‌ی q بعدی به صورت زیر محاسبه خواهد شد:

$$\begin{aligned} Z &= V_q \Lambda_q^{۱/۲} \\ &= (\sqrt{\lambda_۱} v_۱, \sqrt{\lambda_۲} v_۲, \cdots, \sqrt{\lambda_q} v_q) \\ &= \begin{bmatrix} z'_۱ \\ z'_۲ \\ \vdots \\ z'_n \end{bmatrix} \end{aligned}$$

که $V_q = (v_۱, v_۲, \cdots, v_q)$ ماتریسی است که هر یک از ستون‌های آن یعنی v_i که $i = ۱, ۲, \cdots, q$، به ترتیب بردارهای ویژه متناظر با مقادیر ویژه λ_i هستند و همچنین Λ_q ماتریس قطری شامل مقادیر ویژه آن است. بردارهای سطری $z'_۱$، $z'_۲$ و..... z'_n مختصات نقاط در بعد q هستند که فاصله اقلیدسی بین آن‌ها یعنی $\delta_{ij} = (z_i - z_j)'(z_i - z_j)$ به عدم تشابه اولیه نزدیک است.

مثال ۴.۲. فرض کنید ماتریس عدم تشابه ۴ مشاهده به صورت زیر داده شده است:

$$D = \begin{bmatrix} ۰ & ۹۳ & ۸۲ & ۱۳۳ \\ ۹۳ & ۰ & ۵۲ & ۶۰ \\ ۸۲ & ۵۲ & ۰ & ۱۱۱ \\ ۱۳۳ & ۶۰ & ۱۱۱ & ۰ \end{bmatrix}$$

قصد داریم پیکره‌ای دوبعدی از این ۴ مشاهده به‌دست آوریم. بدین منظور الگوریتم مقیاس‌گذاری چندبعدی کلاسیک را اجرا می‌کنیم. اولین مرحله ساخت ماتریس A و سپس تشکیل ماتریس B است.

$$A = D^{۲} = \begin{bmatrix} ۰ & ۸۶۹۴ & ۶۷۲۴ & ۱۷۶۸۹ \\ ۸۶۴۹ & ۰ & ۲۷۰۴ & ۳۶۰۰ \\ ۶۷۲۴ & ۲۷۰۴ & ۰ & ۱۲۳۲۱ \\ ۱۷۶۸۹ & ۳۶۰۰ & ۱۲۳۲۱ & ۰ \end{bmatrix}$$

$$J = \begin{bmatrix} ۱ & ۰ & ۰ & ۰ \\ ۰ & ۱ & ۰ & ۰ \\ ۰ & ۰ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۱ \end{bmatrix} - \frac{۱}{۴} \begin{bmatrix} ۱ & ۱ & ۱ & ۱ \\ ۱ & ۱ & ۱ & ۱ \\ ۱ & ۱ & ۱ & ۱ \\ ۱ & ۱ & ۱ & ۱ \end{bmatrix}$$

$$= \begin{bmatrix} ۰/۷۵ & -۰/۲۵ & -۰/۲۵ & -۰/۲۵ \\ -۰/۲۵ & ۰/۷۵ & -۰/۲۵ & -۰/۲۵ \\ -۰/۲۵ & -۰/۲۵ & ۰/۷۵ & -۰/۲۵ \\ -۰/۲۵ & -۰/۲۵ & -۰/۲۵ & ۰/۷۵ \end{bmatrix}$$

$$B = -\frac{۱}{۲} JAJ$$

$$= \begin{bmatrix} ۵۰۳/۰۶۲۵ & -۱۵۵۳/۰۶۲۵ & ۲۵۸/۹۳۷۵ & -۳۷۴۰/۹۳۸ \\ -۱۵۵۳/۰۶۲۵ & ۵۰۷/۸۱۲۵ & ۵/۳۱۲۵ & ۱۰۳۹/۹۳۸ \\ ۲۵۸/۹۳۷۵ & ۵/۳۱۲۵ & ۲۲۰۶/۸۱۲۵ & -۲۴۷۱/۰۶۲ \\ -۳۷۴۰/۹۳۷۵ & ۱۰۳۹/۹۳۷۵ & -۲۴۷۱/۰۶۲۵ & ۵۱۷۲/۰۶۲ \end{bmatrix}$$

مقادیر ویژه ماتریس B عبارتند از :

$$\lambda_۱ = ۹۷۲۴/۱۶۸, \quad \lambda_۲ = ۳۱۶۰/۹۸۶, \quad \lambda_۳ = ۳۶/۵۹۶۴۱, \quad \lambda_۴ = ۱/۵۱۰۰۸۹e - ۱۳$$

جهت به‌دست آوردن پیکره‌ی دو بعدی می‌بایست از دو مقدار ویژه بزرگ‌تر B یعنی $\lambda_۱$ و $\lambda_۲$ و بردارهای ویژه متناظر با آن‌ها استفاده کرد. بردارهای ویژه متناظر با این دو مقدار عبارتند از:

$$e_۱ = \begin{pmatrix} -۰/۶۳۷ \\ ۰/۱۸۷ \\ -۰/۲۵۳ \\ ۰/۷۰۴ \end{pmatrix} \qquad e_۲ = \begin{pmatrix} -۰/۵۸۶ \\ ۰/۲۱۴ \\ ۰/۷۰۶ \\ -۰/۳۳۴ \end{pmatrix}$$

در نهایت مختصات دو بعدی نقاط به صورت زیر محاسبه می‌شود:

$$z = \begin{bmatrix} -۰/۶۳۷ & -۰/۵۸۶ \\ ۰/۱۸۷ & ۰/۲۱۴ \\ -۰/۲۵۳ & ۰/۷۰۶ \\ ۰/۷۰۴ & -۰/۳۳۴ \end{bmatrix} \begin{bmatrix} \sqrt{۹۷۲۴/۱۶۸} & ۰ \\ ۰ & \sqrt{۳۱۶۰/۹۸۶} \end{bmatrix} = \begin{bmatrix} -۶۲/۸۳۱ & -۳۲/۹۷۴۴۸ \\ ۱۸/۴۰۳ & ۱۲/۰۲۶۹۷ \\ -۲۴/۹۶۰ & ۳۹/۷۱۰۹۱ \\ ۶۹/۳۸۸ & -۱۸/۷۶۳۴۰ \end{bmatrix}$$

با محاسبه‌ی فاصله‌ی اقلیدسی مشاهدات در پیکره‌ی حاصل، ماتریس Δ به صورت زیر به‌دست می‌آید:

$$\Delta = \begin{bmatrix} ۰ & ۹۲/۸۶۶۰۰ & ۸۱/۹۵۹۶۲ & ۱۳۲/۹۸۰۵۲ \\ ۹۲/۸۶۶۰۰ & ۰ & ۵۱/۴۴۶۵۸ & ۵۹/۵۶۱۰۴ \\ ۸۱/۹۵۹۶۲ & ۵۱/۴۴۶۵۸ & ۰ & ۱۱۰/۹۹۹۰۵ \\ ۱۳۲/۹۸۰۵۲ & ۵۹/۵۶۱۰۴ & ۱۱۰/۹۹۹۰۵ & ۰ \end{bmatrix}$$

برای مقایسه‌ی ماتریس فوق با عدم تشابه اولیه (ماتریس D) می‌توان از یکی از روش‌های مقایسه‌ی ماتریس‌ها استفاده کرد. این مثال برگرفته از مرجع [۴۹] می‌باشد.

نیکویی برازش در مقیاس‌گذاری کلاسیک

همانطور که در بخش ۶.۱ اشاره شد، انتخاب نامناسب q منجر به از دست رفتن اطلاعات و نتایج گمراه‌کننده خواهد شد، از این رو سوال مهمی که در الگوریتم کلاسیک با آن مواجه هستیم این است که آیا تعداد بعد درنظر گرفته شده در ساخت پیکره، برای حفظ اطلاعات داده‌ها و نمایش مناسب عدم تشابه آن‌ها کفایت می‌کند؟ برای پاسخ به این سوال می‌توان از دو معیار زیر استفاده کرد:

$$GOF_۱ = \frac{\sum_{i=۱}^{q} |\lambda_i|}{\sum_{i=۱}^{n} |\lambda_i|} \qquad (۲۰.۲)$$

$$GOF_۲ = \frac{\sum_{i=۱}^{q} \lambda_i^۲}{\sum_{i=۱}^{n} \lambda_i^۲} \qquad (۲۱.۲)$$

مقادیر GOF_i $(i = ۱, ۲)$، بزرگ‌تر از ۰/۸ بر مناسب بودن تعداد بعد انتخاب شده دلالت می‌کنند [۱۸، ۴۶].

۲.۴.۲ مقیاس‌گذاری چند بعدی غیرمتری

درایه‌های ماتریس عدم تشابه همیشه نتیجه‌ی محاسبه‌ی معیارهای عدم تشابه نیستند به طوری‌که گاهی ماتریس عدم تشابه از قضاوت‌های انسانی ناشی می‌شود. برای مثال فرض کنید از مشتریان یک کارخانه خواسته شده است محصولات آن کارخانه را دو به دو مقایسه و رتبه‌بندی نمایند. در این صورت ماتریس عدم تشابه، ترتیبی نامیده می‌شود. وجود ماتریس‌های عدم تشابه ترتیبی موجب شد تا افرادی چون شپرد [۱] (۱۹۶۲) و کروسکال [۲] (۱۹۶۴) روش مقیاس‌گذاری چند بعدی را به الگوریتمی که مختصات نقاط را بر مبنای رتبه‌بندی عدم تشابه می‌سازد، گسترش دهند [۳۱، ۴۱]. این روش را تحت عنوان مقیاس‌گذاری چندبعدی غیرمتری می‌شناسند.

فرض کنید اعضای بالا مثلثی ماتریس D یعنی $\{d(\mathbf{x}_i, \mathbf{x}_j) \quad : \quad ۱ \leq i \leq j \leq n\}$ به صورت زیر مرتب شوند:

$$d(\mathbf{x}_{r_۱}, \mathbf{x}_{s_۱}) \leq d(\mathbf{x}_{r_۲}, \mathbf{x}_{s_۲}) \leq \cdots \leq d(\mathbf{x}_{r_m}, \mathbf{x}_{s_m}) \quad m = \frac{n(n-۱)}{۲} \qquad (۲۲.۲)$$

هدف مقیاس‌گذاری غیرمتری ساخت پیکره‌ای از مشاهدات در بعد کوچک‌تر از بعد واقعی است به قسمی که رتبه‌بندی فواصل آن‌ها در این پیکره تقریباً با ترتیب درایه‌های ماتریس عدم تشابه که در (۲۲.۲) داده شده است، منطبق باشد. به عبارت دیگر

$$\forall i, j, r, s \qquad d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_r, \mathbf{x}_s) \Rightarrow \delta(\mathbf{x}_i, \mathbf{x}_j) \leq \delta(\mathbf{x}_r, \mathbf{x}_s) \qquad (۲۳.۲)$$

که در آن طبق نمادگذاری قبلی، $\delta(\mathbf{x}_i, \mathbf{x}_j)$ فاصله‌ی اقلیدسی دو مشاهده‌ی $\mathbf{x}_i$ و $\mathbf{x}_j$ در فضای تقلیل یافته q بعدی است. عبارت (۲۳.۲) قید یکنوایی نامیده می‌شود.

فرض کنید از طریق مقیاس‌گذاری کلاسیک، مشاهدات در بعد q پیکربندی و ماتریس فاصله اقلیدسی آن‌ها (Δ) محاسبه شده باشد. با مرتب کردن درایه‌های ماتریس Δ، فواصل اقلیدسی مشاهدات در فضای جدید رتبه‌بندی می‌شوند. به ندرت اتفاق می‌افتد که این رتبه‌بندی با ترتیب موجود بین درایه‌های ماتریس D منطبق بوده و قید یکنوایی برقرار باشد. توسط رگرسیون یکنوا می‌توان اعدادی را برآورد نمود که رتبه‌بندی آن‌ها با ترتیب درایه‌های ماتریس D منطبق باشد که آنها را با $\hat{\delta}$ نمایش می‌دهیم. مقادیر $\hat{\delta}$ فاصله نیستند، بلکه اعدادی هستند که از آن‌ها برای برقرارسازی قید یکنوایی کمک گرفته می‌شود. طی الگوریتم مقیاس‌گذاری غیرمتری، مختصات نقاط در فضای جدید به گونه‌ای تغییر داده می‌شود که فواصل اقلیدسی بین آن‌ها تا حد ممکن به $\hat{\delta}$ها نزدیک شوند یعنی اختلاف فواصل اقلیدسی نقاط با مقادیر $\hat{\delta}$ مینیمم گردد. این کار در الگوریتم مقیاس‌گذاری غیرمتری از طریق مینیمم‌سازی تابعی به نام تابع استرس

[۱] Shepard

[۲] Kruskal

[۱] صورت می‌گیرد که به صورت زیر تعریف می‌شود:

$$STRESS = \left\{ \frac{\sum_{r<s} [\delta(\mathbf{x}_r, \mathbf{x}_s) - \hat{\delta}(\mathbf{x}_r, \mathbf{x}_s)]^2}{\sum_{r<s} \delta^2(\mathbf{x}_r, \mathbf{x}_s)} \right\}^{1/2} \qquad (۲۴.۲)$$

که در آن $r = ۱, ۲, \cdots, n-۱$ و $s = ۲, ۳, \cdots, n$. مینیمم‌سازی تابع فوق به صورت تحلیلی امکان‌پذیر نیست و می‌بایست از روش‌های عددی استفاده کرد. کروسکال روش عددی شیب نزولی را بدین منظور پیشنهاد داده‌است [۳۱]. در این روش طی یک فرآیند تکراری، یک نقطه‌ی اولیه در جهت شیب نزولی تابع حرکت می‌کند تا به اولین مینیمم محلی برسد. بدین منظور در هر تکرار، مختصات نقطه‌ی اولیه طبق یک رابطه‌ی تعریف شده تغییر می‌کند.

اکنون مختصات نقاط در فضای تقلیل یافته را نقاط اولیه و تابع هدف مینیمم‌سازی را تابع استرس فرض کنید. با پیاده‌سازی روش عددی شیب نزولی می‌توان مختصات نقاط اولیه را به گونه‌ای تغییر داد که تابع استرس مینیمم گردد.

با داشتن ماتریس عدم تشابه D مربوط به n مشاهده، الگوریتم یافتن پیکره‌ای بهینه در بعد q توسط مقیاس‌گذاری چندبعدی غیرمتری به شرح زیر است [۲۶، ۳۹، ۴۶]:

۱. درایه‌های ماتریس عدم تشابه مطابق رابطه‌ی (۲۲.۲) رتبه‌بندی می‌شوند.

۲. پیکره‌ای از مشاهدات در بعد q تشکیل می‌شود. این پیکربندی اولیه را می‌توان از روش مقیاس‌گذاری کلاسیک به‌دست آورد.

۳. فواصل اقلیدسی مشاهدات در پیکربندی فعلی محاسبه می‌شود.

۴. رگرسیون یکنوای $d(\mathbf{x}_i, \mathbf{x}_j)$ روی $\delta(\mathbf{x}_i, \mathbf{x}_j)$ اجرا می‌شود و $\hat{\delta}(\mathbf{x}_i, \mathbf{x}_j)$ها محاسبه می‌گردند.

۵. تابع استرس مربوط به فواصل اقلیدسی و $\hat{\delta}(\mathbf{x}_i, \mathbf{x}_j)$هایی که به ترتیب حاصل گام‌های ۳ و ۴ هستند، مطابق فرمول(۲۴.۲) محاسبه می‌شود.

۶. توسط روش عددی شیب نزولی، مختصات مشاهدات در پیکره فعلی تغییر داده می‌شوند تا تابع استرس مینیمم گردد. مقادیر جدید مشاهده $\mathbf{x}_i = (x_{i۱}, x_{i۲}, \cdots, x_{iq})$، به صورت زیر محاسبه می‌شود:

$$x_{il}^{new} = x_{il} + \frac{\alpha}{n-۱} \sum_{\substack{j=۱ \\ j \neq i}}^{n} \left(۱ - \frac{\hat{\delta}(\mathbf{x}_i, \mathbf{x}_j)}{\delta(\mathbf{x}_i, \mathbf{x}_j)} \right) (x_{jl} - x_{il}) \qquad l = ۱, ۲, \cdots, q \quad (۲۵.۲)$$

[۱] Standardized residual sum of square (STRESS)

که در رابطه‌ی ۲۵.۲، α طول گام الگوریتم شیب نزولی است و در مورد نحوه‌ی انتخاب آن در پیوست، بخش آ.۱.۲ توضیح داده شده‌است [۳۱].

۷. مراحل ۳ تا ۶ تا زمانی‌که تابع استرس به مینیمم مقدار همگرا شود، تکرار می‌شوند. یعنی تا زمانی‌که با تکرار بیشتر، کاهش چندانی در میزان استرس رخ ندهد.

مثال ۵.۲. فرض کنید ماتریس عدم تشابه ۴ مشاهده به صورت زیر داده شده‌است.

$$D = \begin{bmatrix} ۰ & ۳ & ۲ & ۵ \\ ۳ & ۰ & ۱ & ۴ \\ ۲ & ۱ & ۰ & ۶ \\ ۵ & ۴ & ۶ & ۰ \end{bmatrix}$$

برای اجرای الگوریتم مقیاس گذاری چند بعدی غیرمتری برای تشکیل پیکره‌ای بهینه از این مشاهدات در بعد ۲، مراحل زیر را انجام می‌دهیم:

در اولین مرحله می‌بایست درایه‌های ماتریس D رتبه‌بندی شوند. درایه‌های ماتریس D به صورت زیر مرتب می‌شوند:

$$d(\mathbf{x}_۲, \mathbf{x}_۳) \leq d(\mathbf{x}_۱, \mathbf{x}_۳) \leq d(\mathbf{x}_۱, \mathbf{x}_۲) \leq d(\mathbf{x}_۲, \mathbf{x}_۴) \leq d(\mathbf{x}_۱, \mathbf{x}_۴) \leq d(\mathbf{x}_۳, \mathbf{x}_۴) \qquad (۲۶.۲)$$

در مرحله دوم به یک پیکره‌ی اولیه از مشاهدات در بعد ۲ نیاز داریم. مختصات مشاهدات در این پیکره‌ی اولیه مطابق مقادیر جدول ۹.۲ فرض شده که برگرفته از مرجع [۲۶] می‌باشد.

جدول ۹.۲: مختصات اولیه مشاهدات مثال ۵.۲

i	$x_{i۱}$	$x_{i۲}$
۱	۳	۲
۲	۲	۷
۳	۱	۳
۴	۱۰	۴

محاسبه فاصله‌ی اقلیدسی مشاهدات در پیکره‌ی فعلی، سومین مرحله از الگوریتم مقیاس‌گذاری غیرمتری است.

ماتریس فاصله‌ی اقلیدسی نقاط جدول ۹.۲ عبارت است از:

$$\Delta = \begin{bmatrix} ۰ & ۵/۱ & ۲/۲ & ۷/۳ \\ ۵/۱ & ۰ & ۴/۱ & ۸/۵ \\ ۲/۲ & ۴/۱ & ۰ & ۹/۱ \\ ۷/۳ & ۸/۵ & ۹/۱ & ۰ \end{bmatrix}$$

رتبه‌بندی درایه‌های ماتریس Δ به صورت زیر می‌باشد:

$$\delta(\mathbf{x}_۱, \mathbf{x}_۳) \leq \delta(\mathbf{x}_۲, \mathbf{x}_۳) \leq \delta(\mathbf{x}_۱, \mathbf{x}_۲) \leq \delta(\mathbf{x}_۱, \mathbf{x}_۴) \leq \delta(\mathbf{x}_۲, \mathbf{x}_۴) \leq \delta(\mathbf{x}_۳, \mathbf{x}_۴) \qquad (۲۷.۲)$$

همانطورکه ملاحظه می‌کنید رتبه‌بندی برخی از درایه‌های ماتریس فاصله‌ی اقلیدسی از ترتیب درایه‌های ماتریس عدم تشابه پیروی نمی‌کنند. توسط رگرسیون یکنوا می‌توان اعداد $\hat{\delta}$ را که رتبه‌بندی درایه‌های ماتریس D را حفظ می‌کنند، برآورد نمود و به کمک این اعداد، مختصات نقاط را برای رسیدن به پیکره‌ی ایده‌آل تغییر داد، در ادامه روند کار رگرسیون یکنوا در این مثال، شرح داده شده‌است.

درایه‌های ماتریس‌های Δ و D به ترتیب از سمت چپ به صورت دو به دو با هم مقایسه می‌شوند. در اولین مقایسه داریم: $d(\mathbf{x}_۲, \mathbf{x}_۳) \leq d(\mathbf{x}_۱, \mathbf{x}_۳)$ در حالی‌که $\delta(\mathbf{x}_۱, \mathbf{x}_۳) \leq \delta(\mathbf{x}_۲, \mathbf{x}_۳)$. طبق این مقایسه، رتبه‌بندی دو درایه از ماتریس عدم تشابه توسط $\delta(\mathbf{x}_۱, \mathbf{x}_۳)$ و $\delta(\mathbf{x}_۲, \mathbf{x}_۳)$ نقض شده‌است، لذا بر اساس آنچه در پیوست آ بخش آ.۱ توضیح داده شده‌است، میانگین آن‌ها محاسبه شده و جایگزین مقادیر فعلی می‌شود، یعنی:

$$\begin{aligned} \hat{\delta}(\mathbf{x}_۱, \mathbf{x}_۳) &= \hat{\delta}(\mathbf{x}_۲, \mathbf{x}_۳) \\ &= \frac{\delta(\mathbf{x}_۱, \mathbf{x}_۳) + \delta(\mathbf{x}_۲, \mathbf{x}_۳)}{۲} \\ &= \frac{۲/۲ + ۴/۱}{۲} \\ &= ۳/۱۵ \end{aligned}$$

پس از اعمال تغییر فوق در عبارت (۲۷.۲)، زوج بعدی از درایه‌ها مورد مقایسه قرار می‌گیرند. در این مقایسه داریم: $d(\mathbf{x}_۱, \mathbf{x}_۳) \leq d(\mathbf{x}_۱, \mathbf{x}_۲)$ و $\hat{\delta}(\mathbf{x}_۱, \mathbf{x}_۳) \leq \delta(\mathbf{x}_۱, \mathbf{x}_۲)$. از آنجا که ترتیب این زوج از درایه‌های ماتریس D در عبارت (۲۷.۲) حفظ شده است، مقادیر $\hat{\delta}(\mathbf{x}_۱, \mathbf{x}_۳)$ و $\delta(\mathbf{x}_۱, \mathbf{x}_۲)$ بدون تغییر باقی می‌مانند. اکنون این دو عدد با درایه قبل یعنی $\hat{\delta}(\mathbf{x}_۲, \mathbf{x}_۳)$ مقایسه می‌شوند. اگر محاسبه‌ی مقادیر جدید، صعودی بودن عبارت (۲۷.۲) را بر هم نزند، سه مقدار مذکور بدون تغییر باقی می‌مانند، در غیر این صورت میانگین آن‌ها جایگزین مقادیر فعلی خواهد شد.

سایر مقادیر $\hat{\delta}$ نیز به طریق مشابه محاسبه می‌شوند. این مقادیر در جدول ۱۰.۲ آمده است.

جدول ۱۰.۲: مقادیر عدم تشابه، فاصله اقلیدسی و فاصله برآوردشده جفت مشاهدات

(i,j)	$d(\mathbf{x}_i,\mathbf{x}_j)$	$\delta(\mathbf{x}_i,\mathbf{x}_j)$	$\hat{\delta}(\mathbf{x}_i,\mathbf{x}_j)$	$[\delta(\mathbf{x}_i,\mathbf{x}_j)-\hat{\delta}(\mathbf{x}_i,\mathbf{x}_j)]^2$	$\delta^2(\mathbf{x}_i,\mathbf{x}_j)$
(۲,۳)	۱	۴/۱	۳/۱۵	۰/۹	۱۶/۸
(۱,۳)	۲	۲/۲	۳/۱۵	۰/۹	۴/۸
(۱,۲)	۳	۵/۱	۵/۱	۰	۲۶
(۲,۴)	۴	۸/۵	۷/۹	۰/۴	۷۲/۳
(۱,۴)	۵	۷/۳	۷/۹	۰/۴	۵۳/۳
(۳,۴)	۶	۹/۱	۹/۱	۰	۸۲/۸

در مرحله‌ی بعد تابع استرس مطابق رابطه‌ی (۲۴.۲) و با استفاده از مقادیر جدول ۱۰.۲ به صورت زیر محاسبه می‌شود:

$$STRESS = \sqrt{\frac{۰/۹ + ۰/۹ + ۰/۴ + ۰/۴}{۱۶/۸ + ۴/۸ + ۲۶ + ۷۲/۳ + ۵۳/۳ + ۸۲/۸}}$$
$$= \sqrt{\frac{۲/۶}{۲۵۶}} = ۰/۱$$

اکنون می‌بایست به منظور کاهش مقدار استرس، مختصات اولیه نقاط مندرج در جدول ۹.۲ را مطابق رابطه (۲۵.۲) کمی تغییر داد. طبق بخش ۲.۱.آ در اولین تکرار می‌بایست α را ۰/۲ در نظر گرفت. در این صورت، مختصات جدید نقاط مطابق جدول ۱۱.۲ به‌دست می‌آیند.

جدول ۱۱.۲: مختصات جدید نقاط مثال ۵.۲ پس از یک تکرار الگوریتم مقیاس‌گذاری غیرمتری

i	x_{i1}^{new}	x_{i2}^{new}
۱	۳/۰۱۹۲	۲/۰۲۴۵
۲	۲/۰۲۲۱	۷/۰۷۵۸
۳	۰/۹۵۷۹	۳/۰۳۲۹
۴	۹/۵۱۰۲	۳/۰۹۳۲۶

به عنوان مثال مقدار $x_{۱۱}^{new}$ به صورت زیر محاسبه می‌شود:

$$
\begin{aligned}
x_{۱۱}^{new} &= x_{۱۱} + \frac{۰/۲}{۴-۱} \sum_{\substack{j=۱ \\ j\neq i}}^{۴} \left(۱ - \frac{\hat{\delta}\mathbf{x}_i, \mathbf{x}_j)}{\delta(\mathbf{x}_i, \mathbf{x}_j)}\right)(x_{j۱} - x_{۱۱}) \\
&= ۳ + \frac{۰/۲}{۳}\left[\left(۱ - \frac{\hat{\delta}(\mathbf{x}_۱, \mathbf{x}_۲)}{\delta(\mathbf{x}_۱, \mathbf{x}_۲)}\right)(x_{۲۱} - x_{۱۱}) + \left(۱ - \frac{\hat{\delta}(\mathbf{x}_۱, \mathbf{x}_۳)}{\delta(\mathbf{x}_۱, \mathbf{x}_۳)}\right)(x_{۳۱} - x_{۱۱})\right. \\
&\quad \left. + \left(۱ - \frac{\hat{\delta}(\mathbf{x}_۱, \mathbf{x}_۴)}{\delta(\mathbf{x}_۱, \mathbf{x}_۴)}\right)(x_{۴۱} - x_{۱۱})\right] \\
&= ۳ + ۰/۰۶۶۶\left[(۱ - \frac{۵/۱}{۵/۱})(۲-۳) + (۱ - \frac{۳/۱۵}{۲/۲})(۱-۳) + (۱ - \frac{۷/۹}{۷/۳})(۱۰-۳)\right] \\
&= ۳/۰۱۹۲
\end{aligned}
$$

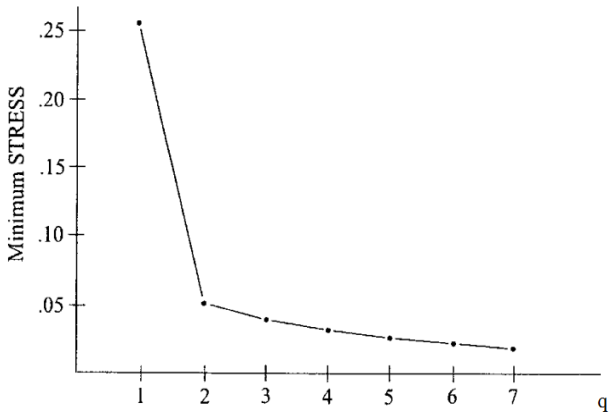
سپس مراحل ۳ تا ۶ الگوریتم برای نقاط جدید اجرا می‌گردد تا زمانی‌که با تکرار بیشتر کاهش چندانی در مقدار استرس رخ ندهد [۲۶].

نیکویی برازش در مقیاس‌گذاری چندبعدی غیرمتری

با افزایش بعد q، تابع استرس کاهش پیدا می‌کند. مشابه مقیاس‌گذاری کلاسیک، در مقیاس گذاری غیرمتری نیز با این سوال مواجه هستیم که حداقل بعد لازم برای ساخت بهترین پیکره را چگونه می‌توان پیدا کرد؟ بدین منظور الگوریتم مقیاس‌گذاری غیرمتری را به ازای ابعاد مختلف ($q = ۱, ۲, ۳, ...$) اجرا می‌کنیم و برای هر q، تابع استرس مربوط به پیکربندی نهایی را محاسبه می‌کنیم. سپس نمودار این مقادیر را در مقابل ابعاد در نظر گرفته شده، رسم می‌کنیم. بهترین بعد برای ساخت پیکربندی، بعدی است که بیشترین کاهش استرس در آن رخ داده و پس از آن شاهد کاهش چشمگیری در میزان استرس نباشیم [۳۹، ۴۶]. نمونه‌ای از این نمودار در شکل ۴.۲ قابل مشاهده است که در آن حداقل بعد مناسب برای نمایش مشاهدات توسط مقیاس‌گذاری غیرمتری برابر با ۲ است.

۵.۲ تعیین تعداد خوشه‌ها

تعیین تعداد خوشه‌ها کار بسیار دشواری است و در اغلب موارد، تعداد صحیح خوشه‌ها مبهم است. تاکنون روش‌های متعددی برای برآورد تعداد خوشه‌ها پیشنهاد شده‌است. یک روش ساده تعیین تعداد خوشه‌ها این است که تعداد خوشه‌ها را برابر با $\sqrt{n/۲}$ در نظر بگیریم که n تعداد مشاهدات است [۲۲].



شکل ۴.۲: نمودار بعد در مقابل استرس

در این بخش، دو روش را به منظور برآورد تعداد مناسب خوشه‌ها معرفی می‌کنیم.

۱.۵.۲ استفاده از شاخص نیمرخ

از میانگین شاخص نیمرخ علاوه بر ارزیابی کیفیت خوشه‌بندی به منظور برآورد تعداد مناسب خوشه‌ها نیز استفاده می‌شود. بدین منظور خوشه‌بندی را به ازای تعداد خوشه‌های مختلف اجرا نموده و برای هر حالت، میانگین شاخص نیمرخ را محاسبه می‌کنند. تعداد مناسب خوشه‌ها برای انجام خوشه‌بندی، تعدادی است که منجر به بیشترین مقدار میانگین شاخص نیمرخ گردد. به عبارتی اگر m پارامتری تعیین شده باشد، خوشه‌بندی به ازای $k \in \{۲, ۳, ..., m\}$ اجرا می‌گردد و تعداد مناسب خوشه‌ها برابر است با:

$$k^* = \underset{۲\leq l\leq m}{\arg\max}\, S_l \qquad (۲۸.۲)$$

که در آن S_l بیانگر میانگین شاخص نیمرخ به ازای $k = l$ است [۲۹].

۲.۵.۲ استفاده از نمودار مقیاس‌گذاری چند بعدی

یکی از راه‌های برآورد تعداد خوشه‌ها ترسیم داده‌ها است. با استفاده از روش‌های کاهش بعد می‌توان بعد داده‌ها را به دو یا سه کاهش داده و امکان ترسیم آن‌ها را فراهم آورد. روش مقیاس‌گذاری چند بعدی، گزینه‌ی مناسبی جهت نیل به این هدف است چرا که این روش با اخذ ماتریس عدم تشابه، داده‌ها را در فضای اقلیدسی با بعدی کوچک‌تر پیکربندی می‌کند به گونه‌ای که فاصله اقلیدسی مشاهدات در این

پیکره تا حد ممکن به عدم تشابه اخذ شده نزدیک باشد. بنابراین مشاهداتی که تشابه بیشتری دارند در فضای تقلیل‌یافته به فاصله کمتری از یکدیگر قرار می‌گیرند. با ترسیم نمودار پیکره‌ی حاصل اغلب می‌توان گروه‌های قابل تمییزی را روی نمودار تشخیص داده و تعداد خوشه‌ها را حدس زد [۱۷، ۴۳].

فصل ۳

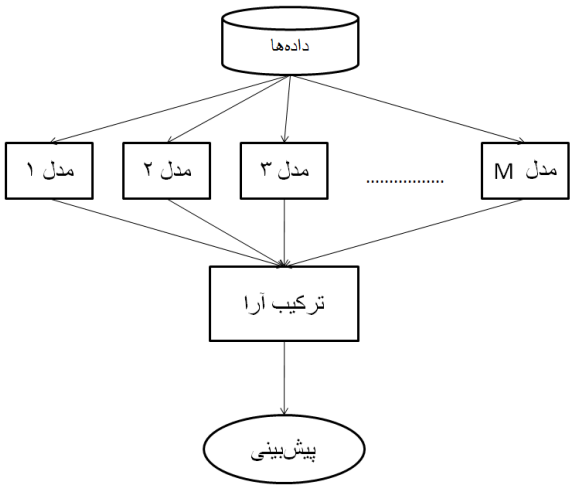
عدم تشابه جنگل تصادفی

بنا بر توضیحات قبلی، بسیاری از روش‌های خوشه‌بندی، مبتنی بر عدم تشابه مشاهدات هستند. در بخش ۳.۱ با توجه به نوع متغیرهای توضیحی، معیارهایی را برای سنجش عدم تشابه مشاهدات معرفی کردیم. در این فصل به معرفی معیاری دیگر خواهیم پرداخت که حاصل یک روش رده‌بندی است و با دیدگاهی کاملاً متفاوت از عدم تشابهات پیشین محاسبه می‌شود. روش جنگل‌های تصادفی یک روش رده‌بندی کارا به ویژه در مواجهه با تعداد زیادی از متغیرهای توضیحی است. ساختار این روش به گونه‌ای است که در آن علاوه بر رده‌بندی داده‌ها، عدم تشابه مشاهدات به طریقی متفاوت تعریف می‌شود. خوشه‌بندی مبتنی بر این عدم تشابه تحت عنوان خوشه‌بندی جنگل‌های تصادفی شناخته می‌شود که هدف اصلی این فصل، معرفی این نوع خوشه‌بندی است. از آنجا که روش جنگل تصادفی یک روش رده‌بندی اجماعی و ترکیبی از چندین درخت رده‌بندی و رگرسیونی است، لذا در این فصل ابتدا توضیحی کلی در مورد روش‌های اجماعی ارائه می‌دهیم، سپس روش جنگل تصادفی را مورد مطالعه قرار داده و عدم تشابه حاصل از آن را معرفی خواهیم نمود و در نهایت خوشه‌بندی جنگل تصادفی را توضیح خواهیم داد.

۱.۳ روش‌های رده‌بندی اجماعی

در روش‌های رده‌بندی اجماعی، مجموعه‌ای از مدل‌های رده‌بندی با هدف افزایش دقت رده‌بندی مورد استفاده قرار می‌گیرند. برای روشن شدن مفهوم روش‌های اجماعی مثال ساده‌ای را در نظر بگیرید؛ فرض کنید بیماری برای تشخیص بیماری خود به جای مراجعه به یک پزشک، با تعداد بیشتری از پزشکان مشورت کند. به احتمال زیاد، فرد بیمار، تشخیصی را به عنوان بیماری خود قبول می‌کند که نظر اکثر پزشکان بوده‌است. اگر فرد بیمار را یک مشاهده و نظر هریک از پزشکان را خروجی یک مدل رده‌بندی فرض کنیم و برآیند نظرات را به عنوان پیش‌بینی نوع بیماری (رده‌ی مشاهده) در نظر بگیریم، در

این صورت می‌توان مسئله‌ی پیش‌بینی بیماری را به عنوان یک مسئله‌ی رده‌بندی اجماعی تعبیر نمود. روش‌های اجماع‌سازی خودگردانی [۱] و افزایشی [۲] نمونه‌هایی از روش‌های رده‌بندی اجماعی هستند که در مراجع [۳۰، ۲۲] در مورد آن‌ها توضیح داده شده است. در این بخش روش اجماع‌سازی خودگردانی مد نظر قرار گرفته و ساختار آن در شکل ۱.۳ نمایش داده شده‌است.



شکل ۱.۳: ساختار روش اجماع‌سازی خودگردانی

طبق شکل ۱.۳، روش مذکور از M مدل مبنا ساخته شده است. هر یک از مدل‌های مبنا، از یک مجموعه داده‌ی آموزشی که از داده‌های اصلی استخراج می‌شود، استفاده می‌کند. در رده‌بندی یک مشاهده جدید، تمامی مدل‌های مبنای رده‌بندی اجماعی شرکت خواهند داشت. به این ترتیب که هریک از مدل‌ها، رده‌ای را برای آن مشاهده تعیین می‌کند و در نهایت، مشاهده به رده‌ای تعلق خواهد گرفت که دارای بیشترین فراوانی در بین تمامی رده‌های تعیین‌شده، باشد. به عبارت دیگر مد رده‌های تخمین‌شده توسط مدل‌های مبنا، رده‌ی نهایی آن مشاهده را تعیین می‌کند. دقت رده‌بندی اجماعی از یک مدل تکی برای رده‌بندی بیشتر است چرا که استفاده از فقط یک مدل رده‌بندی ممکن است در پیش‌بینی رده‌ی یک مشاهده دچار خطا شود، اما یک مدل رده‌بندی اجماعی زمانی رده‌ی یک مشاهده را اشتباه پیش‌بینی می‌کند که بیش از نیمی از مدل‌های مبنا، اشتباه پیش‌بینی کنند [۲۲].

روش جنگل تصادفی [۳] نیز یک روش رده‌بندی اجماعی است که دارای ساختاری مشابه روش اجماع‌سازی خودگردانی است و هر مدل در آن توسط روش درخت رده‌بندی و رگرسیونی تولید می‌شود.

---

[۱] Bootstrap aggregating (Bagging)

[۲] Boosting

[۳] Random Forest

در بخش بعدی روش درخت رده‌بندی و رگرسیونی را که مبنای روش جنگل‌های تصادفی است مرور خواهیم نمود.

۲.۳ درخت رده‌بندی و رگرسیونی

روش درخت رده‌بندی و رگرسیونی [۱] (CART) روشی است که در هر دو موضوع رده‌بندی و رگرسیون به‌کار می‌رود. در این بخش ما تنها به توضیح مبحث رده‌بندی می‌پردازیم. طی این روش با پاسخ بلی-خیر به مجموعه‌ای از سوالات، فضای p-بعدی متغیرهای توضیحی، به طور سلسله مراتبی به ابرمکعب‌های کوچک‌تر p-بعدی تقسیم می‌شود به طوری‌که مشاهدات واقع در هر ناحیه دارای حداکثر تجانس و همگونی باشند. مشاهدات واقع در هر یک از این ابرمکعب‌ها دارای رده‌ی مختص به خود هستند و به آسانی می‌توان در هر ابرمکعب، رده‌ای با بیشترین فراوانی (مد) را تعیین نمود. کار رده‌بندی یک مشاهده‌ی جدید به این صورت است که این مشاهده بر حسب مختصاتش، در یکی از این ابرمکعب‌ها قرار خواهد گرفت و رده‌ی تخمین‌شده برای این مشاهده همان مد رده‌ها در آن ابرمکعب است.

این الگوی رده‌بندی توسط ساختاری موسوم به درخت تصمیم انجام می‌گیرد. در این ساختار درختی، نقطه‌ی انشعاب به دو زیر شاخه، گره [۲] نامیده می‌شود. اولین گره‌ی درخت را ریشه نامیده و آخرین گره‌ها را با عنوان برگ می‌شناسند. همزمان با تقسیم فضای متغیرهای توضیحی به دو ناحیه کوچک‌تر، یکی از گره‌ها به دو گره داخلی تقسیم می‌شود و بدین ترتیب درخت تصمیم رشد می‌کند.

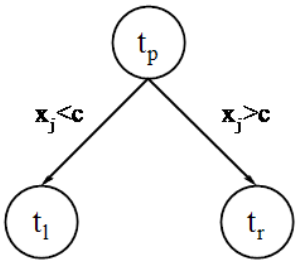
در هر مرحله از افراز فضا به منظور یافتن بهترین نوع تقسیم، همه‌ی متغیرهای توضیحی و همه مقادیر ممکن آن متغیرها، مورد جستجو قرار می‌گیرند تا فضای متغیرها به دو قسمت مناسب تقسیم شود. به عبارت دیگر، در هر مرحله سوال این است که جهت نیل به حداکثر تجانس داده‌ها، فضای متغیرها بایستی در راستای کدام متغیر توضیحی و از کدام مقدار مشاهده شده‌ی آن متغیر افراز گردد. بدین منظور میزان تغییرات تابعی به نام تابع ناخالصی [۳] برای تمام افرازهای ممکن محاسبه می‌شود. فرض کنید فضای متغیرها در راستای متغیر X_j و به ازای مقدار $x_{ij} = c$، به صورت شکل ۲.۳ افراز شده باشد که در آن t_p فضای اصلی و t_l و t_r به ترتیب دو زیرفضای چپ و راست حاصل از این افراز هستند. اگر تابع ناخالصی را با $i(t)$ نشان دهیم، در این صورت تغییرات تابع ناخالصی برای افراز انجام‌شده به صورت زیر تعریف می‌شود:

$$\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r) \qquad (۱.۳)$$

[۱] Classification and regression tree

[۲] Node

[۳] impurity function



شکل ۲.۳: نمایی از افراز فضا در درخت تصمیم

که در آن P_l و P_r نسبتی از کل مشاهدات در ناحیه t_p هستند که به ترتیب در نواحی t_l و t_r واقع شده‌اند. با تکرار فرآیند فوق برای تمامی متغیرها و به ازای مقادیر مختلف آن‌ها، مقدار $\Delta i(t)$ برای تمامی افرازهای ممکن محاسبه شده و در نهایت افرازی که بیشترین تغییرات تابع را نتیجه دهد، به عنوان افراز بهینه در نظر گرفته خواهد شد. به عبارت دیگر تعیین افراز بهینه ناحیه t_p به دو ناحیه کوچک‌تر t_l و t_r معادل با جواب مسئله‌ی ماکزیمم‌سازی زیر است:

$$\underset{\substack{x_j<x_{ij}\\ ۱\le i\le n, ۱\le j\le p}}{\arg\max} \quad \Delta i(t) \tag{۲.۳}$$

تابع ناخالصی می‌تواند به یکی از صورت‌های زیر تعریف شود:

- شاخص جینی [۱] : مرسوم‌ترین تابع ناخالصی، شاخص جینی است. با فرض داشتن k رده، شاخص جینی ناحیه‌ی t به صورت زیر محاسبه می‌شود:

$$i(t) = ۱ - \sum_{j=۱}^{k} P_{j|t}^{۲} \tag{۳.۳}$$

که در آن $P_{j|t}^{۲}$ احتمال شرطی رده‌ی j در ناحیه‌ی t است که توسط فراوانی نسبی رده‌ی j در ناحیه‌ی t برآورد می‌شود.

- شاخص خطای رده‌بندی [۲] : طبق نمادگذاری قبل، این شاخص به صورت زیر تعریف می‌شود:

$$i(t) = ۱ - P_{j|t} \tag{۴.۳}$$

- شاخص آنتروپی متقاطع [۳] : این شاخص نیز یکی دیگر از توابع ناخالصی معرفی شده‌است که از

[۱] Gini index

[۲] Misclassification errore

[۳] Cross entropy

طریق فرمول زیر محاسبه می‌شود:

$$i(t) = -\sum_{j=۱}^{k} P_{j|t} \log_۲(P_{j|t}) \qquad (۵.۳)$$

فرآیند سلسله مراتبی افراز فضا تا زمانی اجرا می‌شود که تعداد مشاهدات واقع در هر ناحیه، برابر با حداقل تعداد از قبل تعیین شده(مثلاً r مشاهده)، باشد یا مشاهدات مذکور دارای رده‌ی مشابهی باشند. این درخت پیچیده که شامل حداکثر تعداد شاخه است، اگرچه برای داده‌های مدلساز مناسب است اما کارآیی آن برای داده‌های جدید که مشارکتی در ساخت مدل نداشته‌اند (داده‌های آزمون) ضعیف بوده و لذا بایستی با هرس نمودن برخی از زیرشاخه‌ها بهینه گردد [۲۲، ۲۳].

۱.۲.۳ هرس درخت

در روش CART پس از شکل‌گیری درخت، از الگوریتمی به نام هزینه‌ی پیچیدگی هرس [۱] به منظور هرس آن استفاده می‌شود. هزینه‌ی پیچیدگی هرس به صورت تابعی از تعداد برگ‌ها و نرخ خطای رده‌بندی در نظر گرفته می‌شود. روند کار این الگوریتم به صورت زیر است [۲۲]:

- گره‌ی N از آخرین گره‌های داخلی انتخاب می‌شود.
- هزینه‌ی پیچیدگی هرس برای دو حالت زیر محاسبه می‌گردد:

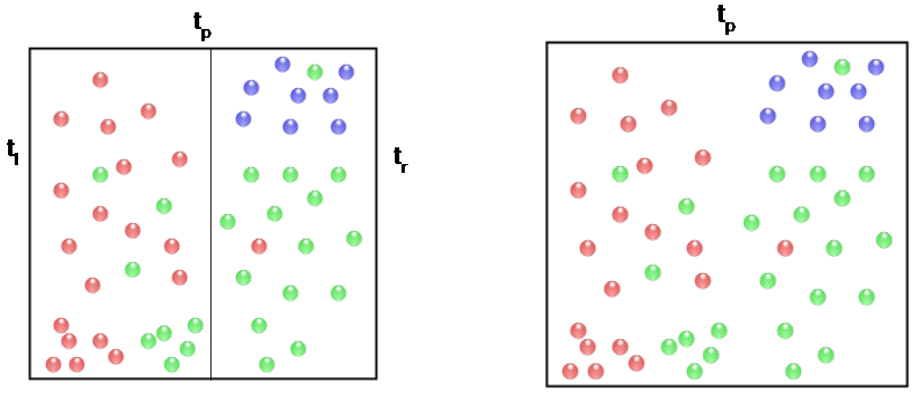
  الف- حالتی که درخت شامل گره‌ی N باشد.

  ب- حالتی که گره‌ی N از درخت حذف و تبدیل به برگ شود.

- در صورتی که هزینه‌ی محاسبه شده در حالت ب کمتر از هزینه‌ی محاسبه شده در حالت الف باشد، هرس انجام شده و گره‌ی N از درخت حذف می‌شود. در غیر اینصورت زیردرخت مذکور در درخت باقی می‌ماند.

مثال ۱.۳. نمودار پراکنش ۳۸ مشاهده با دو متغیر توضیحی را که در شکل ۳.۳(آ) نمایش داده شده است را در نظر بگیرید. با فرض اینکه مقادیر متغیر پاسخ، سه رده‌ی ''آبی''، ''سبز'' و ''قرمز'' باشند. رده‌ی مربوط به هر مشاهده با رنگ متناظرش در این شکل مشخص شده‌است، که تعداد مشاهدات این سه رنگ به ترتیب ۹، ۱۳ و ۱۶ است. یکی از حالت‌های افراز این فضا(t_p) به دو فضای کوچک‌تر t_l و t_r که در نمودار ۳.۳(ب) نشان داده شده است، را در نظر بگیرید که نسبت مشاهدات در این فضا بترتیب $\frac{۲۷}{۵۱}$ و $\frac{۲۴}{۵۱}$ است.

[۱] Cost complexity pruning



(ب) (آ)

شکل ۳.۳: نمودارهای پراکنش داده‌ها در فضای اصلی(قاب (آ)) و فضای افراز شده(قاب (ب))

اگر از شاخص جینی به عنوان تابع ناخالصی استفاده کنیم، تابع ناخالصی مربوط به فضای t_p، t_l و tr به صورت زیر محاسبه می‌شود:

$$i(t_p) = ۱ - \sum_{j=۱}^{k} P^۲_{j|t_p} = ۱ - \left[(\frac{۸}{۵۱})^۲ + (\frac{۲۳}{۵۱})^۲ + (\frac{۲۰}{۵۱})^۲\right] = ۰/۶۲۵$$

$$i(t_r) = ۱ - \sum_{j=۱}^{k} P^۲_{j|t_l} = ۱ - \left[(\frac{۸}{۲۴})^۲ + (\frac{۱۵}{۲۴})^۲ + (\frac{۱}{۲۴})^۲\right] = ۰/۵۱۹$$

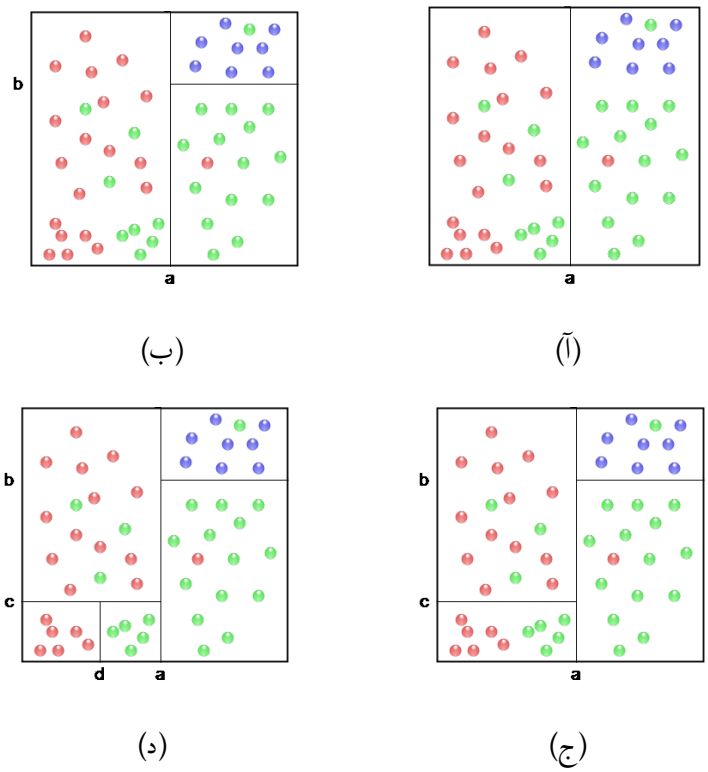
$$i(t_l) = ۱ - \sum_{j=۱}^{k} P^۲_{j|t_r} = ۱ - \left[۰ + (\frac{۸}{۲۷})^۲ + (\frac{۱۹}{۲۷})^۲\right] = ۰/۴۲۵$$

تغییرات تابع ناخالصی نیز به ازای افراز صورت گرفته ، از طریق زیر به‌دست می‌آید:

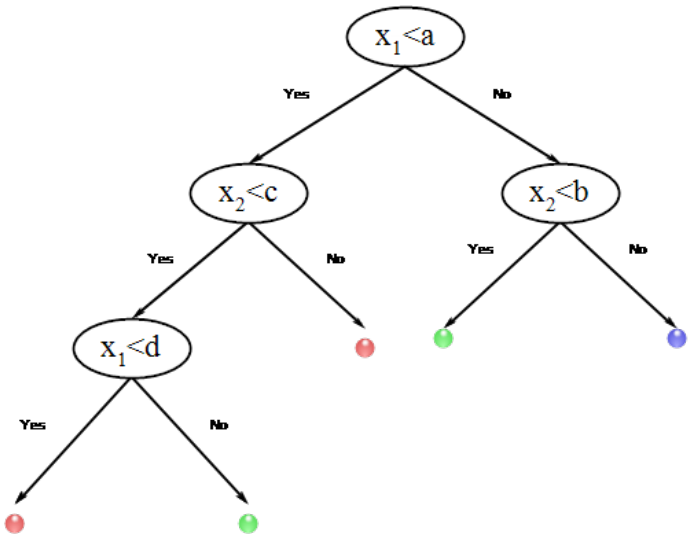
$$\begin{aligned}\Delta i(t) &= i(t_p) - P_l i(t_l) - P_r i(t_r)\\ &= ۰/۶۵۲۱ - \frac{۲۷}{۵۱}(۰/۵۱۹) - \frac{۲۴}{۵۱}(۰/۴۲۵)\\ &= ۰/۶۲۵ - ۰/۲۶۹ - ۰/۱۹۸\\ &= ۰/۱۵۸\end{aligned}$$

در رده‌بندی درخت تصمیم تمامی جهت‌های ممکن برای افراز فضا در نظر گرفته می‌شود و تغییرات تابع ناخالصی به ازای هر حالت محاسبه خواهد شد. بهترین افراز ناحیه‌ی t_p حالتی است که در آن بیشترین مقدار تغییرات تابع ناخالصی رخ دهد. بدین ترتیب هر ناحیه به دو ناحیه کوچک‌تر تقسیم می‌شود تا نقاط هر ناحیه دارای حداکثر همگونی باشند. شکل ۴.۳ مراحل افراز فضای t_p و تشکیل درخت تصمیم

را نشان می‌دهد که به حداکثر همگونی مشاهدات واقع در هر ناحیه منجر شده است. همچنین ساختار درختی این افراز در شکل ۵.۳ نشان داده شده‌است.



شکل ۴.۳: مراحل افراز فضای t_p در رده‌بندی درخت تصمیم طی چهار مرحله‌ی (آ)، (ب)، (ج) و (د)



شکل ۵.۳: ساختار درختی افراز فضای t_p

۳.۳ روش جنگل تصادفی

روش جنگل تصادفی که ترکیبی از روش اجماع‌سازی خودگردانی و انتخاب تصادفی متغیرهاست، توسط بریمن [۱] پیشنهاد شد [۱۰]. در این رده‌بندی اجماعی هریک از مدل‌های مبنا، یک درخت تصمیم است که با مجموعه آموزشی (داده‌های مدل‌ساز) مختص به خود که حاصل یک نمونه خودگردان [۲] از داده‌های اصلی است، رشد می‌کند. از آنجا که این نمونه‌گیری به شیوه‌ی باجایگذاری صورت می‌گیرد، در هر یک از این نمونه‌ها تعدادی از مشاهدات تکرار می‌شوند و تعدادی دیگر حضور ندارند. مشاهداتی را که در نمونه‌های خودگردان حضور ندارند، داده‌های بیرون افتاده [۳] (OOB) نامیده و نقش مجموعه‌ی آزمون را در هر مدل مبنا برعهده خواهند داشت. تفاوت اصلی تشکیل درخت‌های جنگل تصادفی با درخت کلاسیک تصمیم در تعداد گزینه‌های مورد جست و جو جهت انتخاب بهترین افراز است، به طوری‌که در هر مرحله برای انجام انشعاب و تشکیل شاخه، جست و جو برای انجام بهترین افراز فضای متغیرها تنها در بین m متغیر ($m < p$) به تصادف انتخاب شده از p متغیر صورت می‌گیرد. در جنگل تصادفی، هر درخت به طور کامل رشد نموده ولی هرس نخواهد شد.

با افزایش تعداد درختان می‌توان از طریق داده‌های OOB به برآوردی نااریب از خطای رده‌بندی دست یافت[۱۰] . مقدار m و تعداد درختان را به ترتیب با $mtry$ و $ntree$ نشان می‌دهیم. در مورد پارامتر $mtry$ معمولاً مقدار $\sqrt{p}$ پیشنهاد می‌شود که p تعداد متغیرها است [۲۳، ۴۸]. در روش جنگل تصادفی هرچه تعداد درختان بیشتر باشد، پیش‌بینی از دقت بالاتری برخوردار است[۱۰]، بنابراین پارامتر $ntree$ باید به قدر کافی، بزرگ انتخاب شود [۴۲]. اما به طور کلی نمی‌توان مرزی را برای انتخاب پارامتر $ntree$ تعیین کرد چرا که مقدار این پارامتر بستگی به نوع داده‌ها دارد. برای یافتن مقادیر بهینه دو پارامتر $ntree$ و $mtry$ می‌توان مقادیر متفاوت از این دو پارامتر را مورد بررسی قرار داد.

الگوریتم ساخت یک جنگل تصادفی با پارمترهای $mtry = m$ و $ntree = B$ به شرح زیر است [۴۸]:

۱. نمونه‌ای خودگردان از مشاهدات به حجم n انتخاب می‌شود.

۲. درخت T_l روی نمونه خودگردان حاصل از مرحله‌ی قبل رشد می‌کند به طوری‌که در هر گره‌ی آن:

- از میان متغیرهای توضیحی، تعداد m متغیر به تصادف انتخاب می‌شود.

---

[۱] Beriman

[۲] Bootstrap

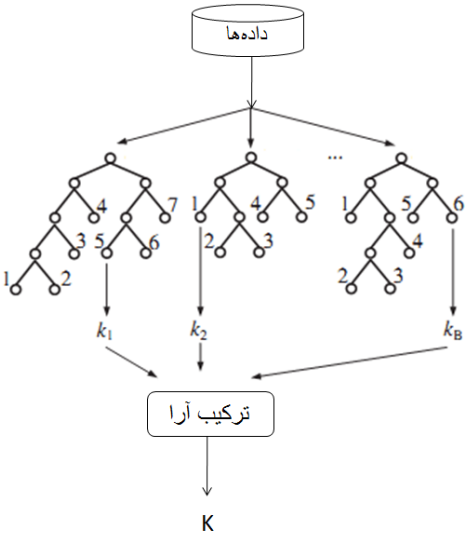
[۳] Out of Bag

- جستجو برای بهترین افراز و تقسیم به دو زیرگره‌ی داخلی، از بین m متغیر منتخب صورت می‌گیرد و پس از یافتن متغیر و مقدار بهینه، افراز صورت می‌گیرد.

۳. درخت T_l به طور کامل رشد می‌کند و هرس نمی‌شود.

۴. گام‌های ۱ تا ۳، B بار تکرار می‌شوند و بدین ترتیب B درخت تصمیم شکل می‌گیرد.

به منظور رده‌بندی یک مشاهده‌ی جدید، هر درخت به یکی از رده‌ها رأی می‌دهد و آن مشاهده در نهایت به رده‌ای نسبت داده می‌شود که مد رده‌های تشخیص داده‌شده (رده‌ی با اکثریت آرا) توسط تمامی درخت‌ها است. در شکل ۶.۳ ساختار یک جنگل تصادفی با B درخت تصمیم نمایش داده شده‌است. در این ساختار، k_i بیانگر رده‌ای است که iامین درخت برای مشاهده جدید تخمین زده‌است، در نهایت نتایج حاصل از تمام درختان با هم ترکیب شده و مشاهده‌ی جدید به رده‌ی k که اکثریت درختان برای آن مشاهده پیش‌بینی کرده‌اند، نسبت داده می‌شود.



شکل ۶.۳: ساختار کلی جنگل تصادفی [۱۶]

۱.۳.۳ تعیین اهمیت متغیرها در جنگل تصادفی

مسلم است که همه‌ی متغیرها در رده‌بندی به یک میزان، دارای اهمیت نیستند و برخی از متغیرها نقش بیشتری را ایفا می‌کنند. روش جنگل تصادفی قادر است متغیرها را بر اساس تاثیری که در رده‌بندی داشته‌اند، اولویت‌بندی نموده و بدین‌ترتیب متغیرهای با بیشترین اهمیت را شناسایی نماید. تعیین اهمیت متغیرها یکی از قابلیت‌های ویژه و مفید این روش است که به‌وسیله داده‌های OOB در هر درخت منفرد فراهم می‌شود. اهمیت هر متغیر به صورت میانگین تغییرات حاصل در خطای رده‌بندی داده‌های OOB،

پس از جابجایی تصادفی(جایگشت) مقادیر آن متغیر محاسبه می‌شود. اگر خطای رده‌بندی مربوط به داده‌های OOB درخت tام را با $OOBerror_t$ و خطای این مجموعه را پس از اعمال یک جایگشت تصادفی بر روی مقادیر jامین متغیر با $\widetilde{OOBerror_t^j}$ نمایش دهیم، در این صورت اهمیت متغیر jام به صورت میانگین تغییرات خطای ناشی از جابجایی مقادیر در کل درخت‌ها تعریف شده و به صورت زیر قابل محاسبه است [۴۸]:

$$VI(x_j) = \frac{1}{ntree} \sum_{t=1}^{ntree} (\widetilde{OOBerror_t^j} - OOBerror_t) \qquad (۶.۳)$$

رابطه‌ی (۶.۳) نشان می‌دهد که هرچه پس از جابجایی مقادیر یک متغیر در داده‌های OOB، تغییر بیشتری در میزان خطای رده‌بندی رخ دهد، اهمیت آن متغیر بیشتر است.

۲.۳.۳ عدم تشابه جنگل تصادفی

در روش جنگل تصادفی عدم تشابه مشاهدات به طریقی کاملاً متفاوت از توابع فاصله‌ی متداول و بر اساس ساختار این مدل تعیین می‌گردد. تشابه مشاهدات در این روش بر مبنای قرار گرفتن آن‌ها در برگ‌های یکسان اندازه‌گیری می‌شود. در جنگل تصادفی تشابه بین دو مشاهده i و j ($s(\mathbf{x}_i, \mathbf{x}_j)$) به صورت نسبت تعداد درخت‌هایی تعریف می‌شود که در آن‌ها دو مشاهده i و j در یک برگ قرار گرفته‌اند. ماتریس تشابه جنگل تصادفی، متقارن و معین مثبت است و هر درایه آن در بازه‌ی $[0, 1]$ قرار دارد. درایه‌های ماتریس عدم تشابه جنگل تصادفی به صورت زیر محاسبه می‌شوند [۳۲] :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{1 - s(\mathbf{x}_i, \mathbf{x}_j)} \qquad (۷.۳)$$

از آنجا که متغیرهای با اهمیت نقش عمده‌ای در تشکیل مدل رده‌بندی جنگل تصادفی و در نتیجه در تعیین میزان تشابه مشاهدات دارند، در تعریف این نوع عدم تشابه، متغیرهای با اهمیت در رده‌بندی تاثیر بیشتری خواهند داشت. همچنین عدم تشابه جنگل تصادفی را می‌توان در مورد انواع متغیرها(پیوسته، دودویی، رسته‌ای، اسمی،ترتیبی و...)به‌کار برد، چرا که تشکیل درخت رده‌بندی، وابسته به نسبت مشاهدات است و با مقادیر متغیرهای توضیحی مرتبط نیست [۴۲].

۴.۳ 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. به‌دست آوردن عدم تشابه جنگل تصادفی، جز با اجرای رده‌بندی به این روش میسر نیست. اما چگونه می‌توان رده‌بندی جنگل تصادفی را در مورد مساله خوشه‌بندی که در آن رده‌ای برای مشاهدات تعریف نشده است، اجرا نمود؟

بدین منظور می‌بایست مسئله‌ی خوشه‌بندی را به یک مسئله‌ی رده‌بندی تبدیل کرد. جهت انجام این تبدیل می‌بایست تمامی مشاهدات 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 افزودن آن‌ها به داده‌های اصلی، مشاهدات اصلی با رده‌ی صفر و داده‌های مصنوعی با رده‌ی ۱ برچسب‌گذاری می‌شوند. بدین ترتیب مسئله‌ی **خوشه‌بندی** به مسئله‌ی **رده‌بندی** تبدیل خواهد شد (شکل ۷.۳). سپس می‌توان توسط جنگل تصادفی دو رده‌ی مذکور را مدل‌بندی نمود و بر اساس ساختار مدل تعیین‌شده عدم تشابه مشاهدات اصلی را مطابق آنچه در بخش ۲.۳.۳ توضیح داده شد، محاسبه کرد.

		X_1	.	.	.	X_p	Y
داده‌های اصلی	1						0
	2						0
	.						0
	.						.
	.						.
	n						0
داده‌های مصنوعی	n+1						1
	n+2						1
	.						1
	.						.
	.						
	2n						1

شکل ۷.۳: چگونگی تبدیل مسئله خوشه‌بندی به رده‌بندی

از آنجا که عدم تشابه جنگل تصادفی به شدت به داده‌های مصنوعی وابسته است [۴۲]، چگونگی تولید این داده‌ها از اهمیت زیادی برخوردار است. داده‌های مصنوعی می‌بایست از یک توزیع مناسب و با حجمی برابر با داده‌های اصلی تولید شوند. برای تولید داده‌های مصنوعی مصداقی از هر متغیر مورد نیاز است. بدین منظور می‌توان از دو روش زیر استفاده کرد [۳۲، ۴۲]:

۱. نمونه‌گیری تصادفی از توزیع حاشیه‌ای تجربی داده‌ها یا به عبارتی تولید نمونه‌ای خودگردان از مقادیر هر متغیر در داده‌های اصلی. ماتریس عدم تشابه جنگل تصادفی حاصل از این روش را RF_{boot} می‌نامیم.

۲. نمونه‌گیری تصادفی ساده از ابرمکعبی که فضای متغیرهای توضیحی را تشکیل می‌دهد. بدین منظور می‌توان داده‌های مصنوعی را از توزیع یکنواخت در بازه‌ی $[m_i, M_i]$ تولید نمود که در آن m_i و M_i به ترتیب مینیمم و ماکزیمم مقدار متغیر iام هستند. ماتریس عدم تشابه جنگل تصادفی که از این روش به‌دست می‌آید را 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 به‌دست 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 به‌دست آورد و هر یک از این دو ماتریس را به طور مجزا در روش‌های خوشه‌بندی مبتنی بر عدم تشابه به‌کار برد و کارایی آن‌ها را مورد بررسی قرار داد.

فصل ۴

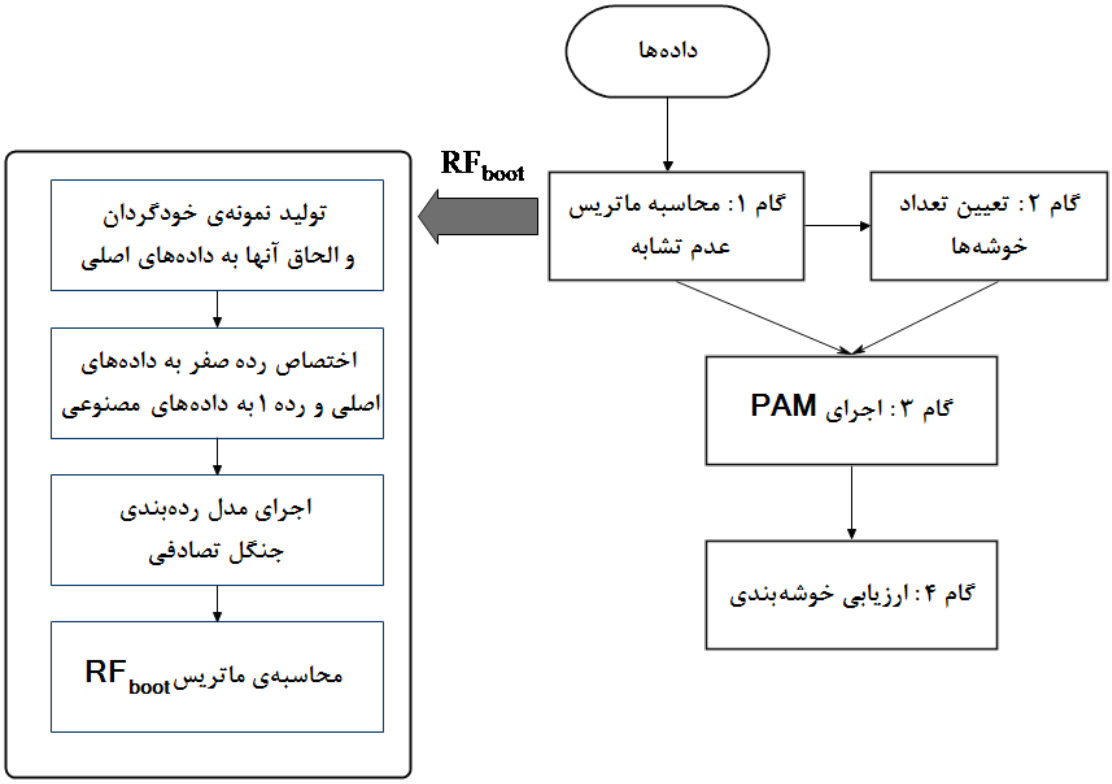
مطالعه شبیه‌سازی

در فصل سوم، عدم تشابه جدیدی مبتنی بر روش رده‌بندی جنگل تصادفی معرفی شد و به چگونگی تبدیل مسئله‌ی خوشه‌بندی به مسئله‌ی رده‌بندی و محاسبه این نوع عدم تشابه پرداختیم. لازم به یادآوری است که با توجه به ساختار تولید داده‌های مصنوعی در مسئله‌ی رده‌بندی جنگل تصادفی، می‌توان دو نوع ماتریس عدم تشابه مختلف (RF_{boot} و RF_{uni}) به‌دست آورد و آن‌ها را در روش‌های خوشه‌بندی مبتنی بر عدم تشابه به‌کار گرفت. در این فصل ابتدا الگوریتم مورد استفاده برای مطالعه‌ی شبیه‌سازی را شرح داده و سپس آن را بر روی مجموعه داده‌های شبیه‌سازی شده اجرا نموده و ضمن مطالعه خواص این نوع عدم تشابه، نتایج حاصل از آن را با نتایج اخذ شده از اعمال عدم تشابه مبتنی بر فاصله‌ی اقلیدسی مقایسه می‌کنیم.

۱.۴ الگوریتم تحقیق در مطالعه شبیه‌سازی

در این فصل برای انجام خوشه‌بندی، رهیافتی مبتنی بر عدم تشابه جنگل تصادفی و فاصله‌ی اقلیدسی را معرفی نموده و نتایج آن‌ها را با هم مقایسه می‌کنیم.

ابتدا بر اجرای رهیافت توسط عدم تشابه RF_{boot} تمرکز می‌کنیم. در یک نگاه اجمالی می‌توان فلوچارت رهیافت مذکور را به صورتی که در شکل ۱.۴ نمایش داده شده، در نظر گرفت.



شکل ۱.۴: فلوچارت الگوریتم تحقیق مبتنی بر عدم تشابه RF_{boot} در مطالعه‌ی شبیه‌سازی

گام‌های این الگوریتم به صورت جزئی‌تر در زیر شرح داده شده‌است:

۱. در اولین گام، ماتریس عدم تشابه مشاهدات محاسبه می‌شود. به منظور محاسبه‌ی ماتریس RF_{boot}، لازم است در این گام، مراحل زیر اجرا شوند:

(آ) نمونه‌ای خودگردان به روش با جایگذاری از مقادیر هر متغیر تولید شده و به عنوان داده‌های مصنوعی، به مشاهدات اصلی الحاق می‌شود.

(ب) رده‌های صفر و یک برای متغیر پاسخ به ترتیب برای مشاهدات اصلی و داده‌های مصنوعی تعیین می‌گردند، بدین ترتیب مسئله‌ی خوشه‌بندی به مسئله‌ی رده‌بندی تبدیل می‌شود.

(ج) برای اجرای رده‌بندی جنگل تصادفی، پارامترهای $mtry$ و $ntree$ تعیین می‌شوند. در تمامی مثال‌های این فصل، پارامتر $mtry$ برابر با $\sqrt{p}$ و پارامتر $ntree$ برابر با ۵۰۰ در نظر گرفته شده‌است (به بخش ۳.۳ رجوع شود).

(د) برای رسیدن به پایداری نسبی، رده‌بندی جنگل تصادفی با پارامترهای تعیین شده در گام قبل، به تعداد ۵۰ بار اجرا گردیده و سپس میانگین مقادیر عدم تشابه حاصل شده در هر اجرا محاسبه می‌شود.

۲. تعداد خوشه‌های مناسب برای خوشه‌بندی از طریق روش‌های پیشنهادی در بخش ۵.۲ تعیین می‌گردد.

۳. با داشتن ماتریس عدم تشابه و همچنین تعداد خوشه‌های بهینه متناظر، روش خوشه‌بندی PAM اجرا می‌شود.

۴. خوشه‌های حاصل از مرحله قبل، توسط یکی از معیارهای ارزیابی خوشه‌بندی معرفی شده در بخش ۳.۲ به نام شاخص رند تعدیل‌یافته ارزیابی می‌گردند.

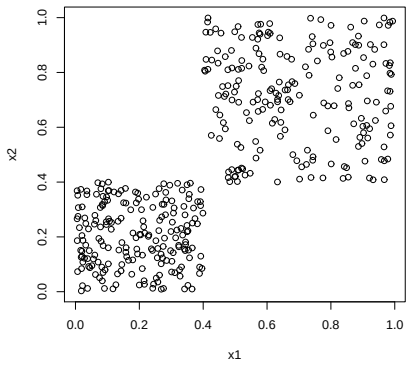
خوشه‌بندی مبتنی بر عدم تشابه RF_{uni} نیز کاملاً مشابه رهیافت مبتنی بر RF_{boot} صورت می‌گیرد، با این تفاوت که در مرحله‌ی آ در گام ۱ داده‌های مصنوعی از طریق نمونه‌ی یکنواخت در بازه‌ی مینیمم و ماکزیمم مقدار هر متغیر تولید می‌شوند. همچنین به منظور اجرای رهیافت فوق توسط فاصله‌ی اقلیدسی، ماتریس فاصله‌ی اقلیدسی مشاهدات (Δ) در گام ۱ محاسبه می‌گردد.

۲.۴ تاثیر ماتریس عدم تشابه در کارآیی خوشه‌بندی

با توجه به نقش بنیادین معیار سنجش عدم تشابه در مسئله‌ی خوشه‌بندی، انتظار می‌رود که استفاده از معیار عدم تشابه جنگل تصادفی نتایج متفاوتی نسبت به فاصله‌ی اقلیدسی ارائه کند، در این بخش علاوه بر نشان دادن این تفاوت، کارآیی دو معیار مذکور را مورد ارزیابی قرار می‌دهیم.

مثال ۱.۴. مجموعه داده‌ای با نام RFcl1، متشکل از ۱۰ متغیر، ۴۰۰ مشاهده و ۲ خوشه به صورت زیر شبیه‌سازی شده است:

خوشه‌ها توسط متغیرهای X_1 و X_2 ایجاد شده و هشت متغیر به عنوان نوفه در نظر گرفته شده‌اند. ۲۰۰ مشاهده‌ی اول متغیرهای X_1 و X_2 از توزیع یکنواخت در بازه‌ی $[0, 0/4]$ و ۲۰۰ مشاهده‌ی دوم در بازه‌ی $[0/4, 1]$ تولید شده‌اند. ۲۰۰ مشاهده‌ی اول را متعلق به خوشه‌ی ۱ و بقیه مشاهدات را متعلق خوشه‌ی ۲ در نظر می‌گیریم. متغیر X_3 دارای توزیع برنولی با احتمال موفقیت ۰/۵ است و متغیرهای X_4، X_5 و... X_{10} حاصل جایگشت تصادفی مقادیر متغیر X_1 هستند. بدین ترتیب متغیرهای مذکور، مستقل از متغیر X_1 و هم‌توزیع با آن می‌باشند. نمودار پراکنش متغیرهای X_1 و X_2 و خوشه‌های تشکیل شده در شکل ۲.۴ نمایش داده شده است.

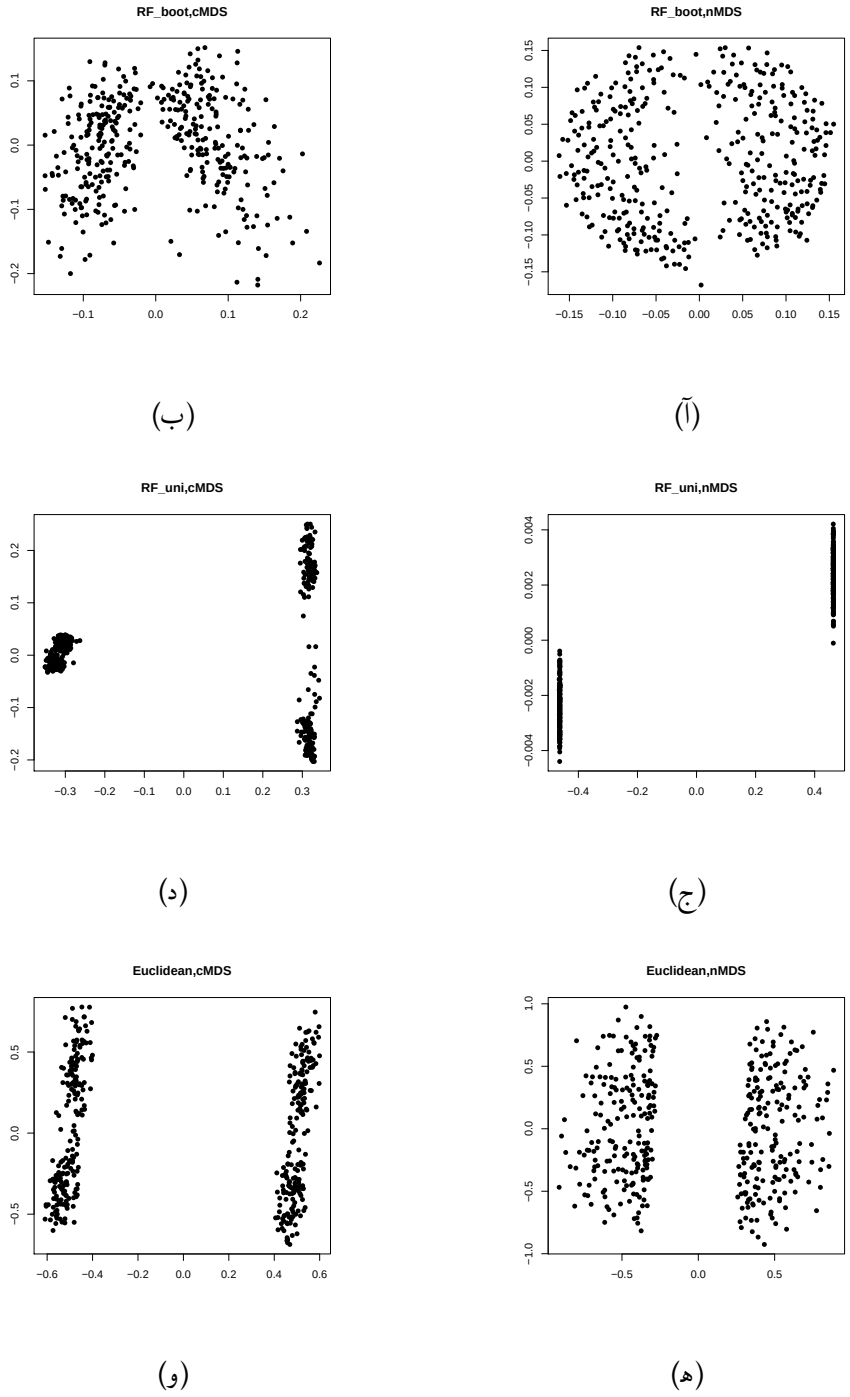


شکل ۲.۴: نمودار پراکنش متغیرهای X_1 و X_2 در مجموعه داده RFcl1 و خوشه‌های تشکیل شده

پس از محاسبه‌ی سه ماتریس عدم تشابه RF_{boot}، RF_{uni} و Δ برای مشاهدات مجموعه داده‌ی RFcl1، لازم است تعداد خوشه‌های متناظر با هر یک از آن‌ها به طور مجزا تعیین گردد. علیرغم اینکه در این مثال تعداد واقعی خوشه‌ها مشخص است، قصد داریم عملکرد سه ماتریس مذکور را در روش‌های برآورد تعداد خوشه‌ها یعنی استفاده از شاخص نیمرخ و نمودار مقیاس‌گذاری چند بعدی (معرفی شده در بخش ۵.۲) مورد ارزیابی قرار دهیم که جزئیات آن به ترتیب در بندهای الف و ب آمده‌است.

الف- به منظور برآورد تعداد خوشه‌ها توسط شاخص نیمرخ، خوشه‌بندی PAM به ازای تعداد خوشه‌های مختلف ($k = ۱, ۲, \cdots, ۲۰$) اجرا گردیده و مقدار شاخص نیمرخ برای هریک از خوشه‌بندی‌های صورت‌گرفته، محاسبه شده‌است. سپس تعداد خوشه‌ای که منجر به ماکزیمم شدن مقدار این شاخص گردیده، به عنوان تعداد بهینه‌ی خوشه‌ها در نظر گرفته شده‌است. در این مثال تعداد خوشه‌ها برای هر یک از سه ماتریس عدم تشابه RF_{boot}، RF_{uni} و Δ، برابر با عدد ۲ تعیین شده و لذا برآوردی بدون خطا صورت گرفته‌است.

ب- به منظور تعیین تعداد خوشه‌ها می‌توان با اعمال هر یک از ماتریس‌های عدم تشابه RF_{boot}، RF_{uni} و Δ در روش‌های مقیاس‌گذاری کلاسیک و غیرمتری، بعد داده‌ها را به ۲ تقلیل داده و آن‌ها را روی یک نمودار دو بعدی ترسیم نمود. بدین ترتیب برای هر ماتریس عدم تشابه، دو ترکیب مختلف خواهیم داشت که هر یک می‌توانند ساختار داده‌ها را به گونه‌ای متفاوت از دیگری نمایش دهند. در این قسمت قصد داریم کارایی روش مقیاس‌گذاری چند بعدی را در برآورد تعداد خوشه‌های مجموعه‌ی RFcl1 مورد بررسی قرار دهیم. همچنین در صدد پاسخگویی به این سوال هستیم که در مورد هر ماتریس عدم تشابه، کدام روش مقیاس‌گذاری چند بعدی برای تشخیص تعداد خوشه‌ها می‌تواند موثرتر واقع شود. نمودارهای مذکور مربوط به داده‌ی RFcl1 در شکل ۳.۴ قابل مشاهده است.



شکل ۳.۴: نمایش داده RFcl1 در نمودارهای حاصل از مقیاس‌گذاری چندبعدی مبتنی بر سه ماتریس عدم تشابه. ستون سمت چپ مربوط به مقیاس‌گذاری کلاسیک و ستون سمت راست مربوط به مقیاس‌گذاری غیرمتری است. سطرهای اول و دوم و سوم به ترتیب نمایانگر نمودارهای مقیاس‌گذاری چند بعدی مبتنی بر عدم تشابه RF_{boot}، RF_{uni} و فاصله اقلیدسی هستند.

نمودارهای (آ) و (ب) شکل ۳.۴، نمودارهای مقیاس‌گذاری چندبعدی مبتنی بر عدم تشابه RF_{boot} را نمایش می‌دهند. هرچند در هر دو نمودار مذکور می‌توان دو گروه را برای مشاهدات تشخیص داد اما در

نمودار (ب) تفکیک بهتری از داده‌ها صورت گرفته‌است، بنابراین در خوشه‌بندی مبتنی بر RF_{boot} بهتر است از نمودار مقیاس‌گذاری کلاسیک جهت تعیین تعداد خوشه‌ها استفاده شود.

در مورد عدم تشابه RF_{uni} نمودار مقیاس‌گذاری غیرمتری (نمودار ۳.۴-ج) نمایانگر دو گروه برای داده‌ها است درحالی‌که مقیاس‌گذاری کلاسیک (نمودار ۳.۴-د) سه گروه را برای داده‌ها نمایش می‌دهد، لذا در مورد عدم تشابه RF_{uni} بهتر است تعیین تعداد خوشه‌ها برمبنای نمودار مقیاس‌گذاری غیرمتری صورت گیرد.

نمودارهای مقیاس‌گذاری چندبعدی مبتنی بر فاصله اقلیدسی که در سطر سوم شکل ۳.۴ نمایش داده شده‌اند، نمایانگر دو گروه برای داده‌ها هستند، لذا در مورد ماتریس Δ هر دو روش مقیاس‌گذاری کلاسیک و غیرمتری در برآورد تعداد خوشه‌ها مفید هستند.

پس از تعیین تعداد خوشه‌ها،می‌توان خوشه‌بندی PAM را انجام داد که در آن با استفاده از ماتریس عدم تشابه و تعداد خوشه‌های بهینه، داده‌ها خوشه‌بندی می‌شوند. در این مثال، الگوریتم PAM سه بار و هر بار با عدم تشابه متفاوتی اجرا شده‌است. نتایج خوشه‌بندی توسط شاخص رند تعدیل‌یافته (ARI) مورد ارزیابی قرار گرفته و مقادیر آن در جدول ۱.۴ درج شده‌است. طبق این جدول بیشترین مقدار شاخص ARI مربوط به RF_{boot} و برابر با ۰/۶۷۱۷ است لذا بهترین خوشه‌بندی بر اساس عدم تشابه جنگل تصادفی و با استفاده از نمونه خودگردان یعنی RF_{boot} حاصل گردیده‌است. همچنین این جدول نشان می‌دهد که عدم تشابه RF_{uni} و Δ عملکرد ضعیفی را از خود نشان داده‌اند.

جدول ۱.۴: مقادیر شاخص ARI به منظور ارزیابی عملکرد الگوریتم PAM مبتنی بر سه ماتریس عدم تشابه مربوط به داده‌ی RFcl1

عدم تشابه	شاخص رند تعدیل یافته
RF_{boot}	۰/۶۷۱۷
RF_{uni}	−۰/۰۰۲۱
Δ	۰/۰۰۷۵

ترسیم نمودن داده‌ها علاوه بر اینکه در مرحله تعیین تعداد خوشه‌ها مفید است، در مرحله ارزیابی (گام ۴) هم می‌تواند موثر واقع شود. اگر توسط روش مقیاس‌گذاری چند بعدی نمایش مناسبی از داده‌ها را روی یک نمودار دوبعدی در اختیار داشته باشیم، با منعکس کردن خوشه‌های تخمینی و واقعی روی این نمودار، می‌توان میزان تطابق آن‌ها را مورد بررسی قرار داده و بدین وسیله امکان ارزیابی بصری نتایج خوشه‌بندی را فراهم آورد. منظور از مناسب بودن این تجسم، ارائه نمایشی از داده‌ها است به گونه‌ای که در آن نقاط کمترین هم‌پوشانی را داشته باشند.

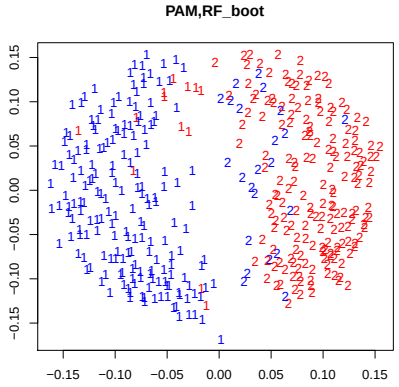
در مورد این مثال، نمودار (آ) شکل ۳.۴ را به عنوان یک تجسم مناسب از داده‌ها انتخاب می‌کنیم. رهیافت مبتنی بر RF_{boot} را در نظر بگیرید. برای نمایش نتایج این رهیافت روی نمودار مذکور، مشاهداتی که طی خوشه‌بندی به خوشه‌ی ۱ تعلق گرفته‌اند را با رنگ "آبی" و مشاهدات نسبت داده‌شده به خوشه‌ی ۲ را با رنگ "قرمز" نشان می‌دهیم. همچنین به منظور منعکس کردن خوشه‌های واقعی روی این نمودار، اعضای خوشه‌ی ۱ را با برچسب 1 و اعضای خوشه‌ی ۲ را با برچسب 2 نمایش می‌دهیم، بدین ترتیب می‌توان خوشه‌های واقعی و تخمینی را به طور همزمان روی یک نمودار مشاهده کرد. در واقع این نمودار طبق جدول ۲.۴ موقعیت مشاهدات در خوشه‌های تخمینی و واقعی را نمایش می‌دهد.

جدول ۲.۴: نحوه‌ی تشخیص موقعیت مشاهدات در خوشه‌های تخمینی و واقعی از طریق نمودار

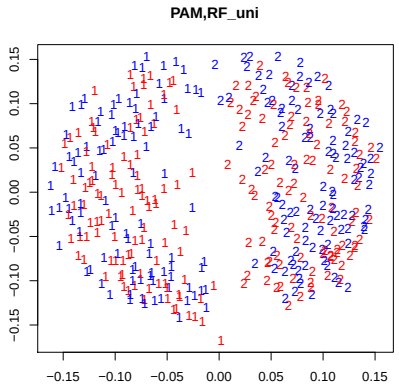
		تخمینی ۱	تخمینی ۲
واقعی	۱	1	1
	۲	2	2

همین رویه را برای ارزیابی رهیافت مبتنی بر RF_{uni} و Δ تکرار می‌کنیم. نمودارهای حاصل در شکل ۴.۴ نمایش داده شده‌اند. اگر نقاطی که دارای برچسب یکسان هستند، هم‌رنگ شوند، آنگاه می‌توان نتیجه گرفت که مشاهداتی که در واقعیت هم‌خوشه بوده‌اند، طی فرآیند خوشه‌بندی نیز درون یک خوشه قرار گرفته‌اند و لذا تطابق خوشه‌های تخمینی و واقعی رخ داده‌است.

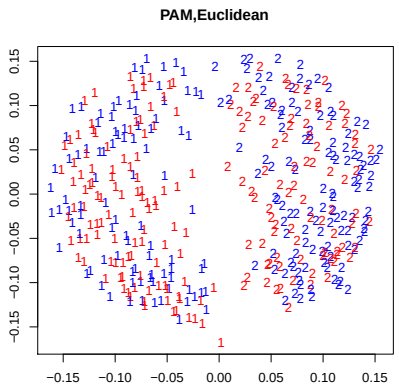
نمودار ۴.۴(آ) که مربوط به ارزیابی نتایج خوشه‌بندی PAM مبتنی بر عدم تشابه RF_{boot} است، نشان می‌دهد که از بین مشاهداتی که دارای برچسب 1 هستند، تعداد ۱۳ مشاهده با رنگ قرمز و بقیه با رنگ آبی مشخص شده‌اند، یعنی از بین ۲۰۰ مشاهده‌ای که در واقعیت به خوشه‌ی ۱ تعلق دارند، ۱۳ مشاهده طی خوشه‌بندی به خوشه‌ی ۲ و ۱۸۷ مشاهده به خوشه‌ی ۱ نسبت داده‌شده‌اند. همچنین از بین ۲۰۰ مشاهده‌ای که دارای برچسب 2 هستند به جز تعداد ۲۳ مشاهده، بقیه به رنگ قرمز درآمده‌اند، لذا اکثر مشاهدات متعلق به خوشه‌ی ۲، طی خوشه‌بندی هم به خوشه‌ی ۲ اختصاص داده شده‌اند. با توجه به این ارزیابی بصری، می‌توان تطابق خوشه‌های واقعی و تخمینی و همچنین عملکرد الگوریتم PAM مبتنی بر عدم تشابه RF_{boot} را نسبتاً مطلوب ارزیابی کرد.



(آ)



(ب)



(ج)

شکل ۴.۴: نمایش مجموعه داده RFcl1 و منعکس کردن خوشه‌های واقعی و خوشه‌های حاصل از الگوریتم PAM عدم تشابه با اخذ (آ):RF_{boot}، (ب):RF_{uni} و (ج):Δ. خوشه‌های تخمینی ۱ و ۲ به ترتیب با دو رنگ آبی و قرمز و خوشه‌های واقعی به ترتیب با دو برچسب 1 و 2 از هم تفکیک شده‌اند.

به همین ترتیب نمودارهای (ب) و (ج) شکل ۴.۴ را مورد بررسی قرار می‌دهیم. همانگونه که در این نمودارها ملاحظه می‌شود از بین مشاهداتی که دارای برچسب 1 هستند، تعداد زیادی به رنگ قرمز نمایش داده شده‌اند. همچنین تعداد زیادی از مشاهدات با برچسب 2 هم به رنگ آبی درآمده‌اند، لذا مشاهداتی که در واقعیت به خوشه‌ای یکسان تعلق داشته‌اند، طی خوشه‌بندی، بین دو خوشه‌ی ۱ و ۲ پراکنده شده‌اند. بنابراین نمودارهای مذکور عدم تطابق خوشه‌های واقعی و تخمینی و عملکرد ضعیف دو عدم تشابه مذکور در خوشه‌بندی PAM را نشان می‌دهند.

مثال ۲.۴. در شبیه‌سازی دوم، متغیر X_1 با ۱۰۰ مشاهده از توزیع برنولی با احتمال موفقیت ۰/۵ تولید شده است. مشاهداتی که مقدار آن‌ها صفر است را خوشه‌ی ۱ و سایر مشاهدات را خوشه‌ی ۲ در نظر بگیرید. بدین ترتیب دو خوشه توسط متغیر X_1 تشکیل شده است. اکنون ۱۹ متغیر ناهمبسته که از توزیع یکنواخت در بازه‌ی $[0, 1]$ تولید شده‌اند را به عنوان نوفه در نظر می‌گیریم و مجموعه داده حاصل را که شامل ۱۰۰ مشاهده، ۲۰ متغیر و ۲ خوشه است، RFcl2 می‌نامیم.

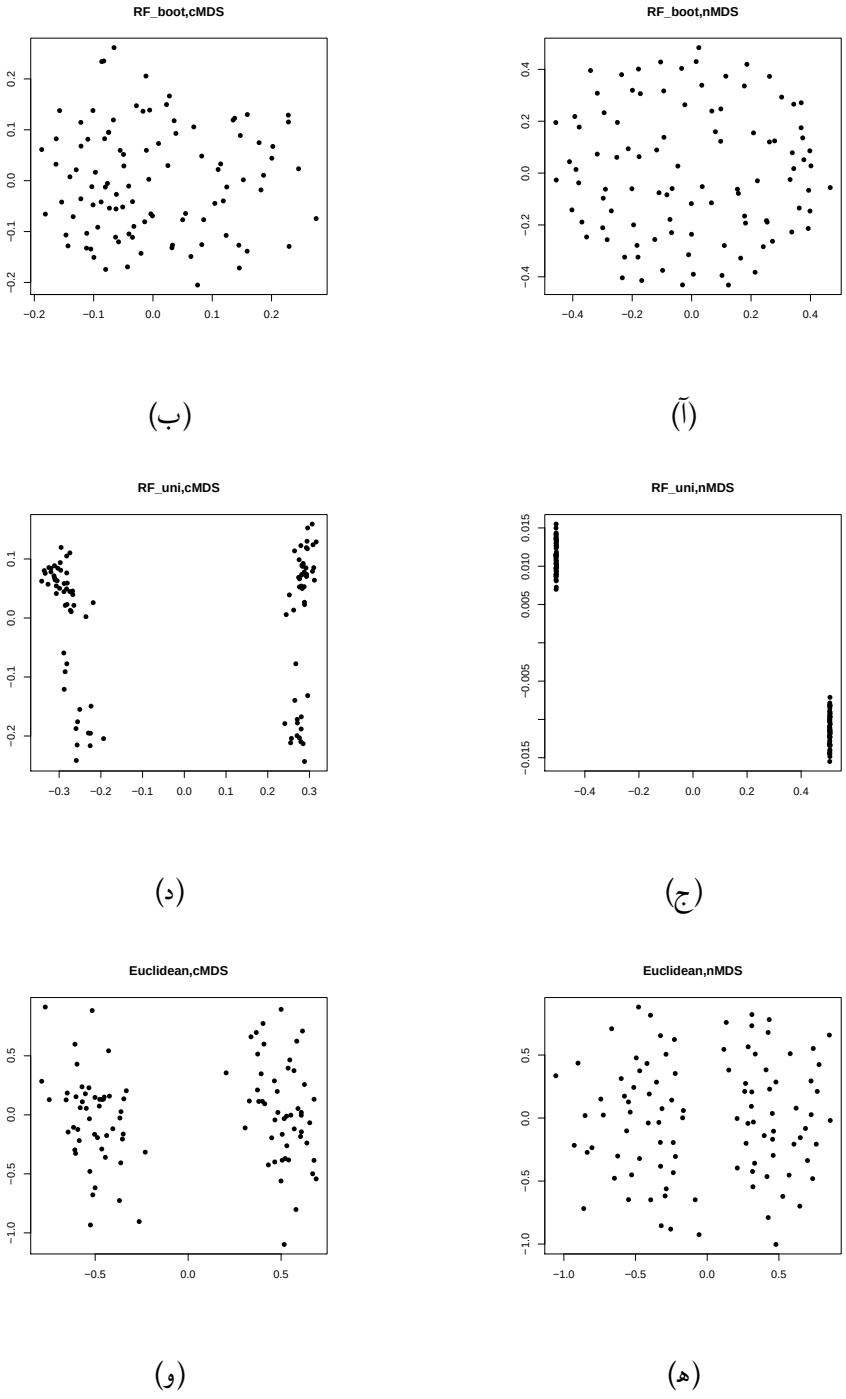
پس از محاسبه‌ی ماتریس‌های عدم تشابه RF_{boot}، RF_{uni} و Δ، تعداد خوشه‌ها با استفاده از شاخص نیمرخ و ترسیم نمودار مقیاس‌گذاری چند بعدی برآورد شده است که جزئیات آن به ترتیب در بند الف و ب آمده است.

الف- تعداد خوشه‌های بهینه متناظر با هر یک از سه ماتریس عدم تشابه، توسط شاخص نیمرخ از مجموعه‌ی $\{2, 3, \cdots, 10\}$ انتخاب شده است. در این مثال تعداد بهینه‌ی خوشه‌ها متناظر با سه ماتریس عدم تشابه RF_{boot}، RF_{uni} و فاصله اقلیدسی برابر با ۲ تعیین شده و در نتیجه برآوردی صحیح، صورت گرفته است.

ب- به منظور تعیین تعداد خوشه‌ها از طریق روش‌های مقیاس‌گذاری کلاسیک و غیرمتری، این نمودارها برای سه ماتریس عدم تشابه RF_{boot}، RF_{uni} و Δ ترسیم شده و به ترتیب در سطرهای اول، دوم و سوم شکل ۵.۴ نمایش داده شده‌اند. در نمودارهای (آ) و (ب) شکل ۵.۴ نمی‌توان گروه‌های قابل تمییز را مشاهده کرد، لذا نمودارهای مقیاس‌گذاری چندبعدی در تعیین تعداد خوشه‌های بهینه متناظر با عدم تشابه RF_{boot} کارآیی ندارند.

در نمودارهای مقیاس‌گذاری چند بعدی مبتنی بر عدم تشابه RF_{uni}، نمودار (ج) دو گروه را برای داده‌ها نمایش می‌دهد، در حالی‌که نمودار (د) منجر به پیدایش خوشه‌های کاذب گشته و نمایانگر چهار گروه برای مشاهدات است. لذا این مثال هم نشان می‌دهد که تعیین تعداد خوشه‌ها در خوشه‌بندی مبتنی بر RF_{uni} بهتر ست بر مبنای نمودار مقیاس‌گذاری غیرمتری صورت پذیرد.

همچنین در نمودارهای مقیاس‌گذاری چند بعدی مبتنی بر ماتریس Δ (سطر سوم شکل ۵.۴) داده‌ها به دو گروه کاملاً مجزا تفکیک شده‌اند و لذا می‌توان تعداد خوشه‌های بهینه متناظر با دو



شکل ۵.۴: نمایش داده RFcl2 در نمودارهای حاصل از مقیاس‌گذاری چندبعدی مبتنی بر سه ماتریس عدم تشابه. ستون سمت چپ مربوط به مقیاس‌گذاری کلاسیک و ستون سمت راست مربوط به مقیاس‌گذاری غیرمتری است. سطرهای اول و دوم و سوم به ترتیب نمایانگر نمودارهای مقیاس‌گذاری چند بعدی مبتنی بر عدم تشابه RF_{boot}، RF_{uni} و فاصله اقلیدسی هستند.

ماتریس عدم تشابه مذکور را برابر با ۲ در نظر گرفت و هردو روش مقیاس‌گذاری چند بعدی را در برآورد تعداد خوشه‌های متناظر با این ماتریس، مفید دانست.

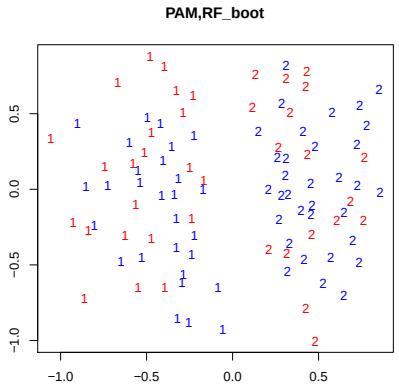
بنابر توضیحات فوق، در مورد این مجموعه داده روش‌های مقیاس‌گذاری چندبعدی در برآورد تعداد خوشه‌های بهینه‌ی متناظر با عدم تشابه RF_{boot} موفق نبوده‌اند اما بنابر برآورد صورت گرفته توسط شاخص نیمرخ، تعداد خوشه‌های بهینه‌ی متناظر با این عدم تشابه را برابر ۲ در نظر می‌گیریم. در گام بعدی الگوریتم PAM با سه ماتریس عدم تشابه RF_{boot}، RF_{uni} و Δ اجرا شده و خوشه‌های حاصل توسط شاخص رند تعدیل یافته مورد ارزیابی قرار گرفته‌اند. جدول ۳.۴ که حاوی نتایج این ارزیابی است، نشان می‌دهد که خوشه‌های حاصل از الگوریتم PAM مبتنی بر عدم تشابه RF_{uni} کاملاً بر خوشه‌های واقعی منطبق شده‌اند. اعمال فاصله اقلیدسی هم در الگوریتم خوشه‌بندی مذکور نتیجه نسبتاً مطلوبی را در پی داشته است در حالی‌که استفاده از عدم تشابه RF_{boot} عملکرد بسیار ضعیفی را از خود نشان داده است.

جدول ۳.۴: مقادیر شاخص ARI به منظور ارزیابی عملکرد الگوریتم PAM مبتنی بر سه ماتریس عدم تشابه مربوط به داده‌ی RFcl2

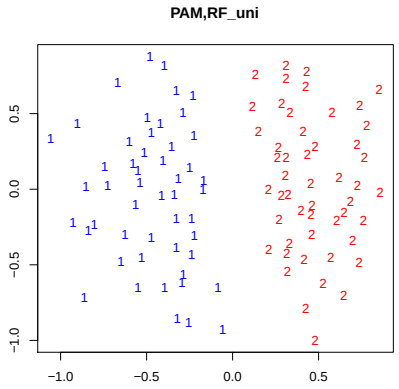
عدم تشابه	شاخص رند تعدیل یافته
RF_{boot}	۰/۰۰۰۴
RF_{uni}	۱
Δ	۰/۷۳۷۰

همچنین به منظور فراهم آوردن ارزیابی بصری، از نمودار (ه) شکل ۵.۴ به منظور نمایش همزمان خوشه‌های تخمینی و واقعی استفاده می‌کنیم. مشابه مثال قبل، در مورد هر عدم تشابه، خوشه‌های تخمینی همزمان با خوشه‌های واقعی روی این نمودار منعکس شده و در شکل ۶.۴ نمایش داده شده است. لازم به ذکر است که در هر نمودار خوشه‌های تخمینی ۱ و ۲ به ترتیب با دو رنگ آبی و قرمز و خوشه‌های واقعی به ترتیب با دو برچسب 1 و 2 از هم تفکیک شده‌اند. باز هم طبق جدول ۲.۴ می‌توان موقعیت مشاهدات در خوشه‌های تخمینی و واقعی را تشخیص داده و نتیجه‌ی خوشه‌بندی را مورد ارزیابی قرار داد.

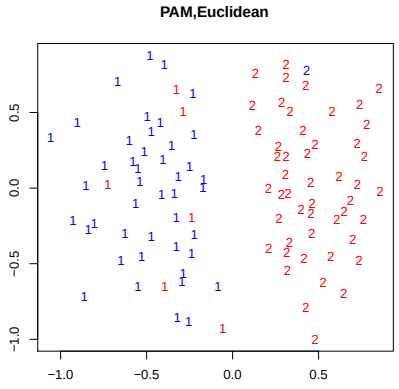
در نمودار ۶.۴(آ) که نمایانگر نتایج خوشه‌بندی PAM مبتنی بر عدم تشابه RF_{boot} است، مشاهداتی که دارای برچسب یکسان هستند، با دو رنگ مختلف نمایش داده شده‌اند و نمی‌توان اکثریت آن‌ها را به یک رنگ نسبت داد لذا مشاهداتی که در واقعیت، درون خوشه‌ای یکسان قرار دارند، طی خوشه‌بندی



(آ)



(ب)



(ج)

شکل ۶.۴: نمایش مجموعه داده RFcl2 و منعکس کردن خوشه‌های واقعی و خوشه‌های حاصل از الگوریتم PAM عدم تشابه با اخذ (آ):RF_{boot}، (ب):RF_{uni} و (ج):Δ. خوشه‌های تخمینی ۱ و ۲ به ترتیب با دو رنگ آبی و قرمز و خوشه‌های واقعی به ترتیب با دو برچسب 1 و 2 از هم تفکیک شده‌اند.

PAM ، بین دو خوشه‌ی مختلف پراکنده شده‌اند. بنابراین از این نمودار می‌توان عدم تطابق خوشه‌های تخمینی و واقعی و عملکرد ضعیف خوشه‌بندی مذکور را نتیجه گرفت.

در نمودار ۶.۴(ب) که نمایانگر نتایج خوشه‌بندی مبتنی بر عدم تشابه RF_{uni} است، تمامی مشاهدات که دارای برچسب ۱ هستند، با رنگ آبی و تمامی مشاهدات با برچسب ۲ با رنگ قرمز ظاهر شده‌اند، لذا تطابق کامل خوشه‌های واقعی و تخمینی در این خوشه‌بندی رخ داده‌است.

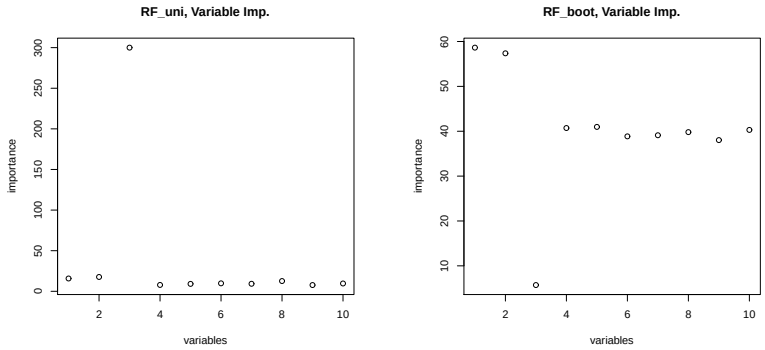
نمودار (ج) شکل ۶.۴ که مربوط به خوشه‌بندی PAM مبتنی بر ماتریس Δ است، نشان می‌دهد که از بین مشاهداتی که در واقعیت به خوشه‌ی ۱ تعلق دارند، ۶ مشاهده به رنگ قرمز و بقیه به رنگ آبی ظاهر شده‌اند، یعنی از بین ۱۰۰ مشاهده‌ی متعلق به خوشه‌ی ۱، تعداد ۹۶ مشاهده طی خوشه‌بندی هم درون خوشه‌ی ۱ قرار گرفته‌اند. همچنین به جز یک مشاهده تمامی مشاهدات متعلق به خوشه‌ی ۲، طی خوشه‌بندی هم به خوشه‌ی ۲ نسبت داده شده‌اند. لذا این نمودار تطابق مطلوب خوشه‌های واقعی و تخمینی و عملکرد نسبتاً مطلوب خوشه‌بندی مذکور را نشان می‌دهد.

۳.۴ متغیرهای موثر در خوشه‌بندی

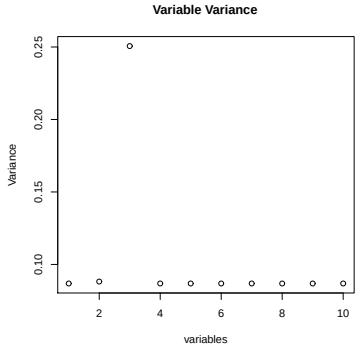
همانگونه که پیش از این اشاره شد، فاصله‌ی اقلیدسی بیشتر تحت تاثیر متغیرهایی قرار می‌گیرد که واریانس بیشتری دارند. همچنین در روش جنگل تصادفی، متغیرهای با اهمیت در رده‌بندی نقش بیشتری در تعریف عدم تشابه حاصل از روش مذکور خواهند داشت. با توجه به تاثیر غیر قابل انکار معیارهای عدم تشابه در روش‌های خوشه‌بندی، انتظار می‌رود متغیرهای موثر در محاسبه‌ی عدم تشابه، در مسئله‌ی خوشه‌بندی نیز تاثیر بیشتری نسبت به سایر متغیرها داشته باشند و خوشه‌های تخمینی به آن‌ها وابسته شوند. در این بخش قصد داریم در قالب یک مثال، نقش متغیرهای موثر در تعریف سه ماتریس عدم تشابه RF_{boot}، RF_{uni} و Δ را در خوشه‌بندی مبتنی بر آن‌ها، مورد بررسی قرار دهیم.

مثال ۳.۴. بار دیگر مجموعه داده RFcl1 را در نظر بگیرید. به منظور شناسایی متغیرهای موثر در محاسبه سه ماتریس عدم تشابه، نمودار میزان اهمیت متغیرها در رده‌بندی جنگل تصادفی مربوط به تولید دو نوع داده مصنوعی و همچنین نمودار واریانس متغیرها در شکل ۷.۴ ترسیم شده است.

همانطور که در نمودار ۷.۴(آ) ملاحظه می‌شود، در تعریف عدم تشابه RF_{boot} متغیرهای X_1 و X_2 دارای بیشترین اهمیت هستند. برای تولید عدم تشابه RF_{boot} نمونه‌ای خودگردان از مشاهدات اصلی به عنوان داده‌های مصنوعی در نظر گرفته می‌شوند. از آنجا که متغیرهای X_1 و X_2 دارای بیشترین همبستگی هستند، با تولید نمونه‌ی خودگردان، متغیرهای مذکور همبستگی خود را در داده‌های مصنوعی از دست می‌دهند، لذا تفاوت میان مشاهدات اصلی و داده‌های مصنوعی به دو متغیر X_1 و X_2 برمی‌گردد. بنابراین تشخیص این متغیرها به عنوان متغیرهای با اهمیت توسط جنگل تصادفی قابل توجیه است.



(ب) (آ)



(ج)

شکل ۷.۴: نمودار اهمیت متغیرها در رده‌بندی جنگل تصادفی مربوط به عدم تشابه RF_{boot} (آ)، عدم تشابه RF_{uni} (ب) و نمودار واریانس متغیرها (ج) در مجموعه داده RFcl1

همچنین در تعریف عدم تشابه RF_{uni} اهمیت متغیر $X_۳$ بیشتر از سایر متغیرها تشخیص داده شده است (نمودار ۷.۴-ب). در تعریف عدم تشابه مذکور داده‌های مصنوعی از توزیع یکنواخت در بازه‌ی مینیمم و ماکزیمم مقدار متغیرها تولید می‌شوند. متغیر $X_۳$ که در مشاهدات اصلی دارای مقادیر ۰ و ۱ است، در داده‌های مصنوعی به متغیری پیوسته تبدیل می‌شود که مقادیر آن در بازه‌ی $[۰, ۱]$ قرار دارند، لذا متغیر مذکور عامل تمایز مشاهدات اصلی از داده‌های مصنوعی خواهد بود و به عنوان با اهمیت‌ترین متغیر شناخته می‌شود. در تعریف عدم تشابه RF_{uni} در صورت عدم وجود متغیرهای برنولی، متغیرهای همبسته به عنوان با اهمیت‌ترین متغیرها شناخته می‌شوند زیرا در این نوع از تولید داده‌ی مصنوعی نیز، متغیرهای همبسته، همبستگی خود را در داده‌های مصنوعی از دست خواهند داد و متغیرهای مذکور بین مشاهدات اصلی و داده‌های مصنوعی تفاوت ایجاد خواهند کرد. همچنین نمودار واریانس متغیرها (نمودار ۷.۴-ج) نشان می‌دهد که متغیر $X_۳$ دارای بیشترین واریانس بین متغیرها است.

پس از شناسایی متغیرهای موثر در محاسبه ماتریس‌های عدم تشابه، نقش متغیرهای مذکور را در خوشه‌بندی مورد بررسی قرار می‌دهیم. بدین منظور بار دیگر، از نمایش دو بعدی داده‌ها توسط روش مقیاس‌گذاری چند بعدی استفاده می‌کنیم. لازم به یادآوری است که روش مقیاس‌گذاری چند بعدی با اخذ ماتریس عدم تشابه، داده‌ها را به گونه‌ای در فضای دو بعدی پیکربندی می‌کند که عدم تشابه مشاهدات طی این تقلیل فضا حفظ شود. بنابراین گروه‌هایی که روی این نمودار نمایان می‌شوند را می‌توان خوشه‌هایی در نظر گرفت که توسط این روش تخمین شده‌اند. در واقع با رسم نمودار مقیاس‌گذاری چند بعدی نوعی خوشه‌بندی صورت گرفته‌است که در آن خوشه‌ی هر مشاهده به طور دقیق مشخص نیست و محقق بنا به تحلیل خود آن را تشخیص می‌دهد. از این رو، تنها به ترسیم نمودار مقیاس‌گذاری چند بعدی اکتفا نموده و وابستگی گروه‌ها (خوشه‌ها)ی نمایان شده را به متغیرهای موثر در محاسبه ماتریس‌های عدم تشابه، مورد بررسی و تحقیق قرار می‌دهیم.

شکل ۸.۴ نمودارهای مقیاس‌گذاری چند بعدی مبتنی بر سه ماتریس عدم تشابه را نشان می‌دهد. هر مشاهده روی این نمودارها با توجه مقادیر متغیرهای $X_۱$ و $X_۳$ و به صورت زیر رنگ‌آمیزی شده‌است:

اگر $X_۱ > ۰/۴$ و $X_۳ = ۱$: مشکی

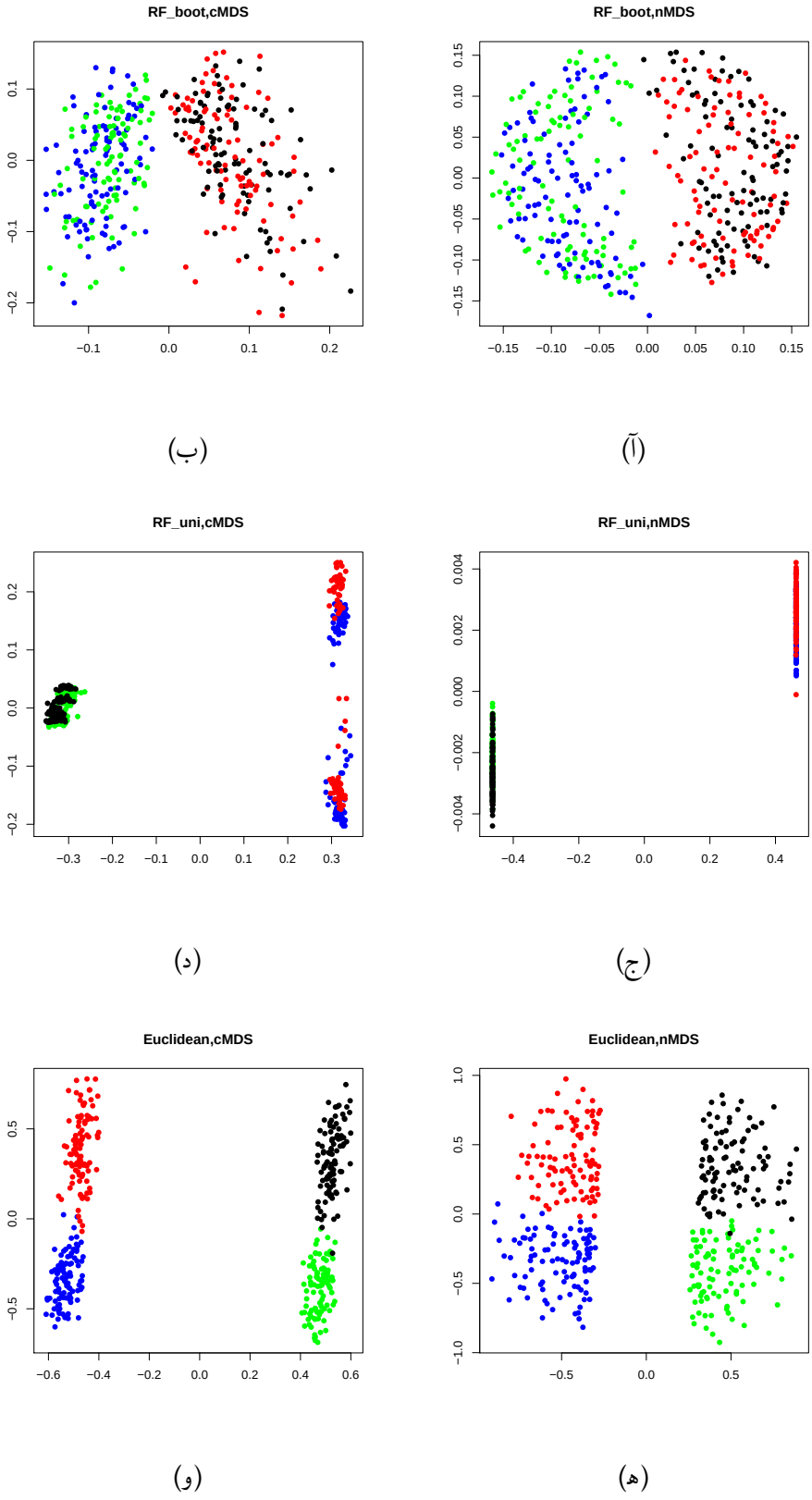
اگر $X_۱ > ۰/۴$ و $X_۳ = ۰$: قرمز

اگر $X_۱ \leq ۰/۴$ و $X_۳ = ۰$: آبی

اگر $X_۱ \leq ۰/۴$ و $X_۳ = ۱$: سبز

همانگونه که در نمودارهای (آ) و (ب) مشاهده می‌شود، نقاط قرمز و مشکی در یک گروه و نقاط آبی و سبز در گروه دیگر قرار گرفته‌اند. نقاط قرمز و مشکی از نظر متغیر $X_۱$ مشابه هستند و مقدار متغیر $X_۳$ در آن‌ها متفاوت است. از طرفی تفاوت دو گروه (قرمز و مشکی) و (سبز و آبی) به متغیر $X_۱$ بر می‌گردد و دو گروه مذکور دارای مقداری یکسان از متغیر $X_۳$ هستند. بنابراین تفکیک رنگ‌ها روی این نمودار نشان می‌دهد که خوشه‌های حاصل از عدم تشابه RF_{boot} به متغیر $X_۱$(با اهمیت‌ترین متغیر در محاسبه‌ی RF_{boot}) وابسته هستند و متغیر $X_۳$ در این خوشه‌بندی نادیده گرفته شده است.

در نمودار (ج) و (د) شکل ۸.۴ یعنی نمودارهای مقیاس‌گذاری چند بعدی مربوط به RF_{uni}، نقاط سبز و مشکی هم‌گروه هستند و نقاط قرمز و آبی در یک گروه قرار دارند لذا در نمودارهای مذکور نقاط هم گروه از نظر متغیر $X_۳$ مشابه هستند و در متغیر $X_۱$ با هم تفاوت دارند. همچنین تفاوت دو گروه (سبز و مشکی) و (آبی و قرمز) به مقدار متغیر $X_۳$ برمی‌گردد و این دو گروه از لحاظ متغیر $X_۱$ مشابه هستند. توزیع رنگ‌ها در این گروه‌ها نشان می‌دهد که تفکیک گروه‌های ایجاد شده روی نمودار مذکور بر اساس متغیر با اهمیت $X_۳$ صورت گرفته‌است و متغیر $X_۱$ در این خوشه‌بندی بی تاثیر بوده است. همانطور که قبلاً ذکر شد، متغیر $X_۳$ با اهمیت‌ترین متغیر در محاسبه عدم تشابه RF_{uni} می‌باشد.



شکل ۸.۴: نمایش داده RFcl1 در نمودارهای حاصل از مقیاس‌گذاری چندبعدی مبتنی بر سه ماتریس عدم تشابه. ستون سمت چپ مربوط به مقیاس‌گذاری کلاسیک و ستون سمت راست مربوط به مقیاس‌گذاری غیرمتری است. سطرهای اول و دوم و سوم به ترتیب نمایانگر نمودارهای مقیاس‌گذاری چند بعدی مبتنی بر عدم تشابه RF_{boot}، RF_{uni} و فاصله اقلیدسی هستند. نحوه‌ی رنگ‌آمیزی نقاط در متن توضیح داده شده‌است.

همچنین در نمودارهای مربوط به فاصله‌ی اقلیدسی یعنی ۸.۴(ه) و ۸.۴(و) مشاهده می‌شود که رنگ‌ها به طور مجزا از یکدیگر قرار گرفته‌اند و در هم ادغام نشده‌اند. اگر نقاط همرنگ را یک خوشه در نظر بگیریم، توزیع نقاط روی این نمودار نشان می‌دهد که خوشه‌بندی صورت‌گرفته در این سه نمودار تحت تاثیر هر دو متغیر $X_۳$ و $X_۱$ بوده‌است.

بررسی نمودارهای شکل ۸.۴ که در مجموعه داده RFcl1، خوشه‌های حاصل از خوشه‌بندی مبتنی بر عدم تشابه جنگل تصادفی به متغیرهای با اهمیت وابسته هستند و این متغیرها در ایجاد خوشه‌ها سهم به‌سزایی دارند. از این وابستگی می‌توان در توصیف خوشه‌های تخمینی استفاده کرد و به جای توصیف خوشه‌ها با تمامی متغیرها، آن‌ها را توسط متغیرهای موثر بیان نمود. اما خوشه‌های حاصل از اعمال فاصله‌ی اقلیدسی را نمی‌توان تنها به متغیر $X_۳$ که دارای بیشترین واریانس است، وابسته دانست.

بنابر نتیجه‌ی فوق، در این مثال خوشه‌های تخمینی حاصل از اعمال عدم تشابه RF_{boot}، به متغیرهای با اهمیت $X_۱$ و $X_۲$ وابسته‌اند. از طرفی در این مجموعه داده، خوشه‌های واقعی نیز توسط متغیرهای $X_۱$ و $X_۲$ تشکیل شده‌اند، بنابراین می‌توان عملکرد مناسب عدم تشابه RF_{boot} نسبت به عدم تشابه RF_{uni} و فاصله اقلیدسی را توجیه نمود و عملکرد نامطلوب عدم تشابه RF_{uni} و فاصله اقلیدسی را ناشی از این دانست که خوشه‌های حاصل از اعمال این دو عدم تشابه تحت تاثیر متغیر $X_۳$ که یک نوفه است، قرار گرفته‌اند.

۴.۴ تاثیر پارامتر رده‌بندی جنگل تصادفی در خوشه‌بندی

همانطور که اشاره شد به منظور اجرای رده‌بندی جنگل تصادفی می‌بایست پارامتر $mtry$ تعیین شود. برای بررسی تاثیر پارامتر مذکور بر خوشه‌بندی، در مجموعه داده RFcl1، با فرض ثابت بودن پارامتر $ntree$، به ازای مقادیر مختلف پارامتر $mtry$، دو ماتریس عدم تشابه RF_{boot} و RF_{uni} را به‌دست آوردیم و الگوریتم PAM را اجرا نمودیم. جدول ۴.۴ نرخ خطای رده‌بندی (OOBerror) و همچنین مقدار شاخص رند تعدیل‌یافته (ARI) متناظر با هر مقدار $mtry$ را نشان می‌دهد. همچنین نرخ خطا و شاخص رند تعدیل‌یافته به ازای مقادیر مختلف $mtry$ در مجموعه داده RFcl2 نیز محاسبه شده و در جدول ۵.۴ آمده است.

جدول ۴.۴: مقادیر شاخص رند تعدیل یافته و نرخ خطای رده‌بندی جنگل تصادفی به ازای مقایر مختلف پارامتر $mtry$ مربوط به مجموعه داده RFcl1

	RF_{boot}		RF_{uni}	
$mtry$	$OOBerror$	ARI	$OOBerror$	ARI
۲	۰/۴۰۵۵	۰/۶۱۵۳	۰/۰۰۸۱	−۰/۰۰۲۱
۴	۰/۳۷۳۸	۰/۷۱۳۳	۰/۰۰۷۴	−۰/۰۰۲۱
۶	۰/۳۵۹۱	۰/۵۹۹۶	۰/۰۰۷۱	−۰/۰۰۲۱
۸	۰/۳۵۲۵	۰/۶۳۹۱	۰/۰۰۷۰	−۰/۰۰۲۱

جدول ۵.۴: مقادیر شاخص رند تعدیل یافته و نرخ خطای رده‌بندی جنگل تصادفی به ازای مقایر مختلف پارامتر $mtry$ مربوط به مجموعه داده RFcl2

	RF_{boot}		RF_{uni}	
$mtry$	$OOBerror$	ARI	$OOBerror$	ARI
۵	۰/۶۹۵۹	۰/۰۰۳۹	۰/۰۱۳۹	۱
۱۰	۰/۶۹۵۴	۰/۰۱۶۴	۰/۰۱۳۹	۱
۱۵	۰/۶۹۷	۰/۰۱۰۵	۰/۰۱۳۹	۱
۲۰	۰/۶۹۱۱	−۰/۰۰۶۵	۰/۰۱۳۸	۱

نتایج مربوط به داده RFcl1 در جدول ۴.۴ نشان می‌دهد که با تولید داده‌های مصنوعی توسط نمونه خودگردان، با افزایش $mtry$ دقت رده‌بندی (کاهش نرخ خطا) افزایش پیدا کرده اما لزوماً خوشه‌بندی بهتری صورت نگرفته‌است. به عنوان مثال همانگونه که ملاحظه می‌شود با تغییر $mtry$ از ۴ به ۶ مقدار شاخص ARI، ۰/۱۲ کاهش یافته و در نتیجه کیفیت خوشه‌بندی پایین آمده‌است. مقادیر نرخ خطا و شاخص رند مربوط به RF_{uni} نیز نشان می‌دهد که با افزایش مقدار $mtry$ نرخ خطای رده‌بندی کاهش چندانی نداشته و مقدار شاخص رند ثابت باقی مانده‌است. همچنین طبق مقادیر جدول ۵.۴ (مربوط به داده‌ی RFcl2) نیز در مورد هر دو ماتریس عدم تشابه، با تغییر $mtry$، تغییرات نرخ خطای رده‌بندی و مقدار شاخص رند، ناچیز بوده و لذا در مورد این داده، نتیجه‌ی رده‌بندی و همچنین خوشه‌بندی نسبت به پارامتر $mtry$ پایدار است.

بنابراین می‌توان نتیجه گرفت که گاهی انتخاب مناسب پارامتر $mtry$ می‌تواند مقدار شاخص رند و در نتیجه کیفیت خوشه‌بندی را به طور قابل ملاحظه‌ای افزایش دهد، اما مقدار بهینه‌ی پارامتر $mtry$ برای انجام خوشه‌بندی لزوماً مقداری نیست که دقت رده‌بندی را افزایش می‌دهد.

فصل ۵

خوشه‌بندی داده‌های بیان ژنی

در فصل چهارم عملکرد خوشه‌بندی مبتنی بر عدم تشابه جنگل تصادفی توسط داده‌های شبیه‌سازی‌شده، مورد مطالعه قرار گرفت. در این فصل قصد داریم کارآیی عدم تشابه جنگل تصادفی را در خوشه‌بندی مجموعه داده‌ای واقعی مورد ارزیابی قرار دهیم. یکی از بارزترین خصوصیات این مجموعه داده، تعداد بسیار زیاد متغیرها در مقایسه با تعداد مشاهدات است که عملکرد روش‌های خوشه‌بندی را به شدت تحت تاثیر خود قرار می‌دهد. کاهش بعد فضای متغیرهای توضیحی با حفظ حداکثری خواص آن‌ها، یکی از راههایی است که می‌توان بر این موضوع فائق آمد، به همین دلیل کاهش بعد داده‌ها قبل از اجرای خوشه‌بندی به عنوان راهکاری دیگر در مواجهه با ابعاد بالا مورد توجه قرار گرفته‌است. در اولین بخش از این فصل، مجموعه داده مورد استفاده، معرفی شده‌است و سپس در بخش دوم، روند انجام کار توسط یک فلوچارت بیان شده که در آن بر ضرورت ترکیب نمودن خوشه‌بندی با مقیاس‌گذاری چندبعدی تاکید شده‌است. نتایج اجرای رهیافت معرفی شده بر روی داده‌ها در بخش سوم ارائه خواهد شد. همچنین در بخش چهارم و پنجم به ترتیب توصیف خوشه‌های تخمینی و حساسیت خوشه‌بندی نسبت به پارامترهای جنگل تصادفی مورد بررسی قرار گرفته‌است.

۱.۵ معرفی مجموعه داده

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، محققین به منظور رسیدن به اهداف ذکر شده به مطالعه و بررسی عملکرد ژن‌های سلول‌های سرطانی روی آورده‌اند که این مهم منوط به استخراج اطلاعات درون ژن‌هاست. بیان ژن [۱] فرآیندی است که طی آن، اطلاعات درون ژن

[۱] Gene expression

استخراج می‌شود که به داده‌های حاصل از این فرآیند، داده‌های بیان ژنی می‌گویند. در گذشته، محققین تنها قادر به مطالعه بخش کوچکی از ژنهای یک سلول بودند اما در سالهای اخیر با ظهور فناوری به نام ریزآرایه DNA [۱] که تلفیقی از دستگاه‌های پیچیده مهندسی و ابزارهای محاسباتی است، امکان بررسی و مطالعه فعالیت هزاران ژن به طور همزمان فراهم آمده است. استفاده از فناوری ریزآرایه DNA حجم انبوهی از داده‌ها را تولید می‌کند که مشخصه بارز آن‌ها بالا بودن تعداد متغیرها (ژن‌ها) و کم بودن تعداد مشاهدات (بیماران) است .تجزیه و تحلیل داده‌های حاصل از فناوری ریزآرایه DNA درهای جدیدی را در تشخیص زود هنگام سرطان، کشف انواع جدید سرطان و توسعه و بهبود روش‌های درمانی به روی پژوهشگران و محققین گشوده است و از این حیث توجه بسیاری از پژوهشگران روی تجزیه و تحلیل این داده‌ها معطوف گشته است. از سال ۱۹۹۸، خوشه‌بندی به عنوان یکی از راه‌های تجزیه و تحلیل داده‌های بیان ژنی مطرح شد و پس از آن مطالعات زیادی در این زمینه صورت گرفت به طوری‌که امروزه یکی از بیشترین کاربردهای خوشه‌بندی مربوط به تجزیه و تحلیل داده‌های بیان ژنی در علوم پزشکی است. در داده‌های بیان ژنی، می‌توان خوشه‌بندی را صرفاً برای متغیرها یا صرفاً برای مشاهدات انجام داد که در این پایان‌نامه، فقط خوشه‌بندی مشاهدات مدنظر قرار گرفته‌است [۱، ۲، ۳، ۴، ۱۵].

با وجود مشخص بودن گروه‌بندی واقعی در اینگونه داده‌ها، پژوهش‌های زیادی در ارتباط با خوشه‌بندی آن‌ها صورت گرفته‌است و هنوز هم ادامه دارد. در صورتی که خوشه‌های تخمین زده شده توسط یک روش، قابل انطباق با گروه‌بندی واقعی باشند، می‌توان از روش خوشه‌بندی مذکور در مواردی که اطلاع دقیقی از رده‌های موجود در دست نیست، استفاده کرد. به علاوه نتایج خوشه‌بندی ممکن است زیرگروه‌هایی از مشاهدات را به نحوی متمایز کند که برای انطباق آن با یافته‌های بالینی، پژوهش‌های آزمایشگاهی یا بالینی جدیدی لازم باشد [۲].

در این فصل از مجموعه داده‌ی بیان ژنی به نام چاودری [۲] استفاده شده است که از طریق درگاه http://bioinformatics.rutgers.edu/Static/Supplements/CompCancer/datasets.htm در دسترس عموم قرار دارد. چاودری و همکارانش در سال ۲۰۰۶ تعداد ۱۰۴ بیمار مبتلا به دو نوع سرطان (سرطان‌های سینه و روده بزرگ) را مورد مطالعه قرار دادند. در این تحقیق تعداد ۱۸۲ ژن از بیماران مذکور با هدف شناخت ساختار ژن‌های مسئول این دو سرطان، مورد بررسی قرار گرفته‌است [۱۳]. هدف از خوشه‌بندی این مجموعه داده، بررسی توانایی رهیافت پیشنهادی در تفکیک این دو گروه از بیماران است.

[۱] DNA microarray

[۲] Chowdary

۲.۵ الگوریتم مطالعه

روند اجرای خوشه‌بندی به این صورت طراحی شده‌است که ابتدا ماتریس عدم تشابه محاسبه می‌شود. در مورد عدم تشابه جنگل تصادفی پس از تولید داده‌های مصنوعی و اختصاص متغیر پاسخ ساختگی، مسئله موجود را به مسئله رده‌بندی تبدیل نموده و با انجام رده‌بندی جنگل تصادفی، عدم تشابه مربوطه را محاسبه می‌کنیم. سپس در خصوص کاهش بعد داده‌ها، ابتدا بعد مناسب را تشخیص داده و پس از آن توسط روش MDS، مختصات مشاهدات در فضای تقلیل‌یافته،محاسبه خواهد شد. سرانجام بر مبنای فاصله اقلیدسی و استفاده از روش PAM عمل خوشه‌بندی صورت می‌پذیرد. در واقع عدم تشابه جنگل تصادفی به طور غیر مستقیم در خوشه‌بندی دخیل بوده و خوشه‌بندی توسط ترکیبی از روش مقیاس‌گذاری چند بعدی و الگوریتم خوشه‌بندی PAM صورت می‌پذیرد. فلوچارت رهیافت مذکور مبتنی بر عدم تشابه RF_{boot} در شکل ۱.۵ نمایش داده شده است.

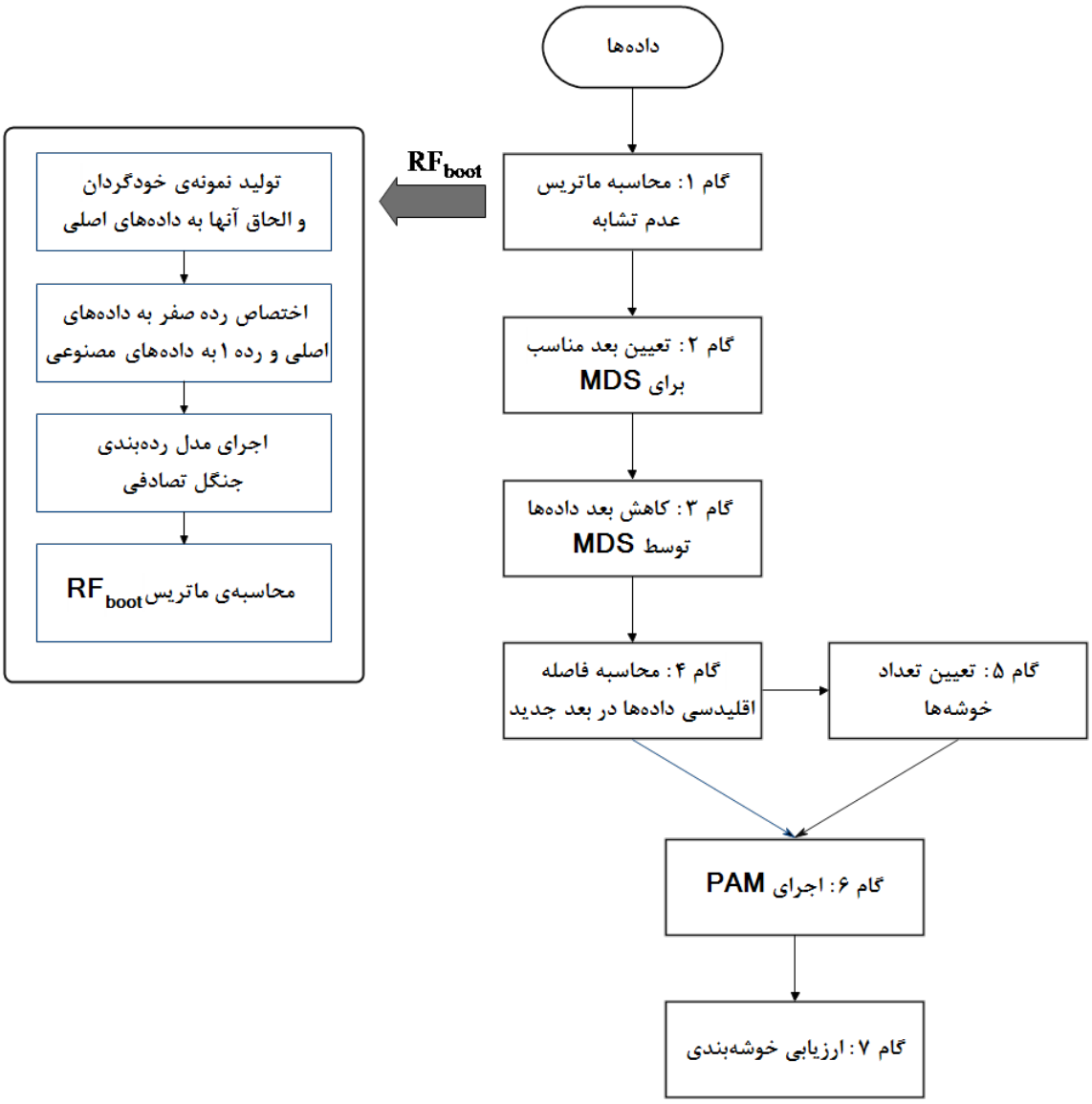
در بخش ۳.۵ الگوریتم مذکور برای داده‌های واقعی به صورت جزئی‌تر بیان شده‌است. همچنین لازم به ذکر است که برای انجام خوشه‌بندی مبتنی بر عدم تشابه RF_{uni} می‌بایست در اولین مرحله از گام ۱، داده‌های مصنوعی را توسط نمونه‌ی یکنواخت در بازه‌ی مینیمم و ماکزیمم مقدار هر متغیر تولید نمود. همچنین می‌توان در این گام، ماتریس فاصله‌ی اقلیدسی(Δ) را محاسبه کرد و نتایج حاصل از آن را با نتایج به‌دست‌آمده از اعمال عدم تشابه جنگل تصادفی مورد مقایسه قرار داد.

۳.۵ اجرای الگوریتم بر روی مجموعه داده چاودری و ارائه نتایج

- **گام اول:** در اولین گام از رهیافت معرفی شده، به منظور محاسبه‌ی عدم تشابه جنگل تصادفی می‌بایست مراحل زیر اجرا شوند:

  - در اولین مرحله از گام ۱، داده‌های مصنوعی به دو روش پیشنهادی(نمونه خودگردان و توزیع یکنواخت) تولید گردیده و به مجموعه داده‌ی اصلی الحاق می‌شوند.

  - در مرحله‌ی دوم، مشاهدات اصلی با رده‌ی صفر و داده‌های مصنوعی با رده‌ی یک برچسب‌گذاری می‌شوند و بدین ترتیب مسئله‌ی خوشه‌بندی به مسئله‌ی رده‌بندی تبدیل می‌شود.



شکل ۱.۵: فلوچارت الگوریتم مطالعه مبتنی بر عدم تشابه RF_{boot}

- در سومین مرحله، رده‌بندی جنگل تصادفی اجرا می‌شود. به منظور اجرای رده‌بندی توسط این روش، نیاز به تعیین دو پارامتر $mtry$ و $ntree$ داریم که در اینجا مقادیر این دو پارامتر به ترتیب برابر با $\sqrt{p} = ۱۳$ و ۵۰۰ در نظر گرفته شده است. برای رسیدن به پایداری نسبی، رده‌بندی جنگل تصادفی به تعداد ۵۰ بار اجرا گردیده و سپس میانگین مقادیر عدم تشابه حاصل شده در هر اجرا محاسبه شده‌است. با اجرای این مرحله، دو ماتریس عدم تشابه RF_{boot} و RF_{uni} متناظر با داده‌های مصنوعی مبتنی بر نمونه‌ی خودگردان و توزیع یکنواخت به‌دست می‌آیند.

همچنین در این گام ماتریس فاصله‌ی اقلیدسی مشاهدات، محاسبه می‌گردد.

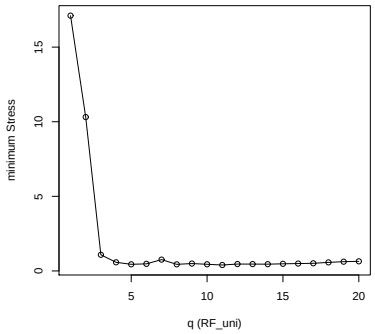
- **گام دوم:** در مورد این مجموعه داده، با ۱۸۲ متغیر توضیحی مواجه هستیم. پیش از این در بخش ۶.۱ اشاره شد که در مواجهه با تعداد بالای متغیرهای توضیحی، کاهش بعد مشاهدات قبل از خوشه‌بندی می‌تواند تا حدودی مشکلات ناشی از ابعاد بالا را رفع نموده و به خوشه‌بندی کاراتر منجر شود. به همین دلیل قصد داریم قبل از اجرای خوشه‌بندی، بعد داده‌ها را توسط روش‌های مقیاس‌گذاری کلاسیک و غیرمتری کاهش دهیم. روش مقیاس‌گذاری چند بعدی، با استفاده از ماتریس عدم تشابه مشاهدات، آن‌ها را در فضایی کوچک‌تر پیکربندی می‌کند به گونه‌ای که فاصله اقلیدسی مشاهدات در فضای جدید به عدم تشابه اولیه نزدیک باشد. باید در نظر داشت که این کاهش بعد نباید موجب از دست رفتن اطلاعات مفید شود چرا که در اینصورت ساختارهای اصلی موجود در داده‌ها دچار تحریف گشته و نمی‌توان به نتیجه خوشه‌بندی اعتماد کرد.

به منظور پیدا کردن بعد مناسب برای روش مقیاس‌گذاری چند بعدی معیارهایی پیشنهاد شده است. همانگونه که در بخش ۴.۲ اشاره شد، در مقیاس‌گذاری کلاسیک با استفاده از یکی از معیارهای $GOF_۱$ و $GOF_۲$ در مورد تعداد بعد مناسب تصمیم‌گیری می‌شود. بدین منظور به ازای هر بعد، مقدار تابع GOF محاسبه می‌شود. مقادیر بزرگ‌تر از ۰/۸ نشان می‌دهند که بعد مورد نظر برای حفظ اطلاعات داده‌ها کفایت می‌کند. در جدول ۱.۵ حداقل بعد مناسب برای اجرای مقیاس‌گذاری چند بعدی کلاسیک، بر اساس معیار $GOF_۱$ و متناظر با سه ماتریس عدم تشابه نشان داده شده‌است.

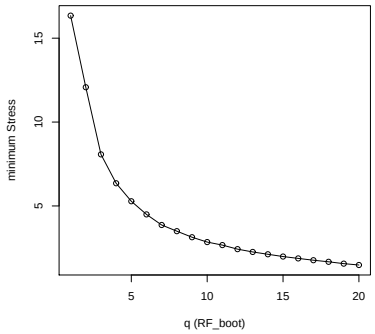
جدول ۱.۵: بعد مناسب برای اجرای مقیاس‌گذاری کلاسیک مربوط به داده‌ی چاودری

عدم تشابه	بعد مناسب بر اساس معیار $GOF_۱$
Δ	۳
RF_{boot}	۵۵
RF_{uni}	۱۵

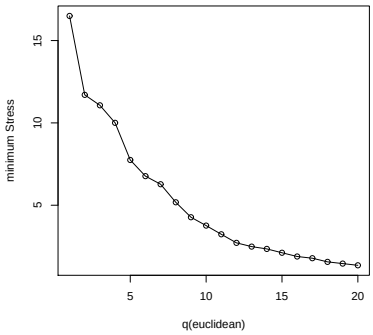
در مقیاس‌گذاری غیرمتری، بعد مناسب به طریقی متفاوت نسبت به روش کلاسیک تعیین می‌شود. در این روش با رسم نمودار مقادیر تابع استرس در مقابل تعداد ابعاد، در این مورد تصمیم‌گیری می‌شود و بهترین بعد، کوچک‌ترین بعدی است که پس از آن شاهد کاهش چندانی در میزان تابع استرس نباشیم. نمودارهای متناظر با سه ماتریس عدم تشابه در مورد داده‌ی چاودری در شکل ۲.۵ نمایش داده شده است و بعد مناسب بر اساس نمودارهای مذکور، در جدول ۲.۵ درج شده است.



(ب)



(آ)



(ج)

شکل ۲.۵: نمودار بعد در مقابل مینیمم استرس جهت تعیین بعد مناسب برای اجرای مقیاس‌گذاری چند بعدی غیرمتری مربوط به مجموعه داده‌ی چاودری و متناظر با سه ماتریس عدم تشابه (آ): RF_{boot}، (ب): RF_{uni} و (ج): Δ

جدول ۲.۵: بعد مناسب برای اجرای مقیاس‌گذاری غیرمتری مربوط به داده‌ی چاودری

بعد مناسب	ماتریس عدم تشابه
۱۰	Δ
۹	RF_{boot}
۳	RF_{uni}

- **گام سوم:** در این گام توسط روش‌های مقیاس‌گذاری چند بعدی کلاسیک و غیرمتری، مشاهدات در بعد تعیین شده(گام دوم) پیکربندی می‌شوند.

- **گام چهارم:** با توجه به هدفی که در مقیاس‌گذاری چند بعدی دنبال می‌شود، می‌توان اطمینان داشت که فاصله اقلیدسی مشاهدات در فضای تقلیل یافته نزدیک به عدم تشابه اولیه است لذا برای انجام خوشه‌بندی کافی است فاصله اقلیدسی مشاهدات را در فضای تقلیل یافته محاسبه کرد تا بتوان از ماتریس حاصل برای اجرای PAM استفاده نمود.

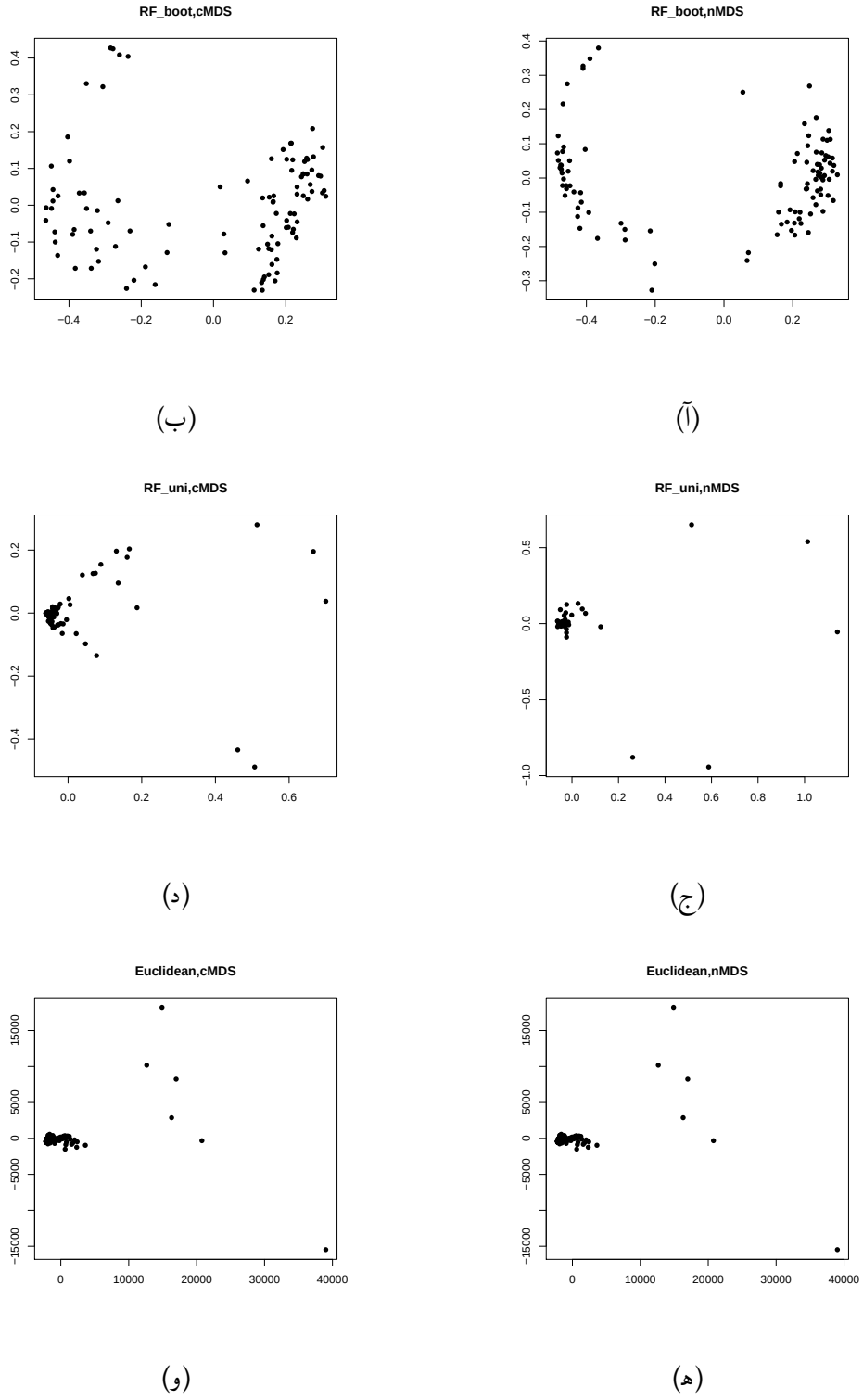
- **گام پنجم:** اکنون می‌بایست تعداد خوشه‌های بهینه برای خوشه‌بندی را تعیین کرد. بدین منظور از دو روش پیشنهادی در بخش ۵.۲ که توضیحات آن در دو بند زیر آمده، استفاده شده است.

الف- با استفاده از شاخص نیمرخ، تعداد خوشه‌ها در مجموعه‌ی $\{۲, ۳, \cdots, ۱۰\}$مورد بررسی قرارگرفته و تعداد بهینه‌ی خوشه‌ها، برابر با ۲ تعیین شده است که معادل با تعداد خوشه‌های واقعی است.

ب- همچنین از نمودارهای مقیاس‌گذاری چند بعدی نیز به منظور حدس تعداد خوشه‌ها استفاده شده است (شکل ۳.۵). همانگونه که ملاحظه می‌شود، در نمودارهای ۳.۵(آ) و ۳.۵(ب) (مربوط به ماتریس عدم تشابه RF_{boot}) می‌توان دو گروه را برای مشاهدات در نظر گرفت. البته در نمودارهای مذکور تعداد اندکی از مشاهدات نیز وجود دارند که نمی‌توان آن‌ها را جزو هیچ یک ازاین دو گروه به حساب آورد.

همچنین در نمودارهای (ج)، (د)(مربوط به ماتریس عدم تشابه RF_{uni})، (ه) و (و)(مربوط به ماتریس Δ) شکل ۳.۵ تمامی مشاهدات به جز تعداد اندکی از آن‌ها در یک گروه قرار گرفته‌اند، لذا نمودارهای مذکور نمایش مناسبی از مشاهدات را ارائه نداده‌اند و نمی‌توانند در تعیین تعداد خوشه‌ها راهگشا باشند.

هر چند در مورد این مجموعه داده، خوشه‌های واقعی مشخص هستند اما نمودارهای مقیاس‌گذاری چند بعدی می‌توانند محققان را به سوی کشف خوشه‌های جدید سوق دهند. به عنوان مثال همانگونه که در نمودارهای شکل ۳.۵ بویژه نمودارهای مربوط به عدم تشابه RF_{uni} و فاصله اقلیدسی ملاحظه نمودید تعدادی از مشاهدات به طور مجزا از سایرین قرار گرفته‌اند که شناسایی این مشاهدات و تحقیق بیشتر در مورد آن‌ها ممکن است زیر مجموعه جدیدی از سرطان را آشکار سازد.



(آ) (ب)

(ج) (د)

(ه) (و)

شکل ۳.۵: نمایش داده چاودری در نمودارهای حاصل از مقیاس‌گذاری چندبعدی مبتنی بر سه ماتریس عدم تشابه. ستون سمت چپ مربوط به مقیاس‌گذاری کلاسیک و ستون سمت راست مربوط به مقیاس‌گذاری غیرمتری است. سطرهای اول و دوم و سوم به ترتیب نمایانگر نمودارهای مقیاس‌گذاری چند بعدی مبتنی بر عدم تشابه RF_{boot}، RF_{uni} و فاصله اقلیدسی هستند.

- **گام ششم:** اکنون مشاهدات در فضای تقلیل‌یافته خوشه‌بندی می‌شوند. بدین منظور الگوریتم خوشه‌بندی PAM با اخذ تعداد خوشه‌ها و ماتریس فاصله اقلیدسی که در گام چهارم محاسبه شده است، به خوشه‌بندی مشاهدات می‌پردازد. در واقع این خوشه‌بندی حاصل ترکیبی از الگوریتم PAM با روش‌های مقیاس‌گذاری کلاسیک و غیرمتری است. از این پس به منظور سهولت، ترکیب الگوریتمPAM با روش‌های مقیاس‌گذاری چند بعدی کلاسیک و غیرمتری را به ترتیب با CPAM و NPAM نامگذاری می‌کنیم.
- **گام هفتم:** در این گام نتایج خوشه‌بندی با خوشه‌های واقعی مقایسه می‌شوند. این مقایسه توسط شاخص رند تعدیل‌یافته (ARI) انجام شده و به صورت خلاصه در جدول ۳.۵ درج شده است.

جدول ۳.۵: ارزیابی نتایج خوشه‌بندی داده‌ی چاودری بر اساس شاخص ARI

عدم تشابه	$CPAM$	$NPAM$
Δ	۰/۰۶۵۷	۰/۶۴۷۶
RF_{boot}	۰/۸۱۴۸	۰/۸۱۴۸
RF_{uni}	۰/۰۲۹۶	۰/۱۰۸۳

جدول فوق نشان می‌دهد که در مورد داده‌ی چاودری، بهترین خوشه‌بندی از اجرای رهیافت‌های CPAM و NPAM با اخذ عدم تشابه RF_{boot} حاصل شده است و پس از آن اعمال فاصله اقلیدسی در رهیافت NPAM در رده‌ی دوم جای می‌گیرد. همچنین مقادیر ARI مربوط به عدم تشابه RF_{uni} حاکی از عملکرد ضعیف این عدم تشابه در هر دو رهیافت CPAM و NPAM است. به عبارت دیگر تولید داده‌های مصنوعی بر اساس توزیع یکنواخت، نتوانسته‌است در خصوص این مجموعه داده منجر به خوشه‌بندی مناسبی شود.

همچنین با مقایسه مقادیر ARI دو رهیافت CPAM و NPAM مربوط به هر ماتریس عدم تشابه در جدول ۳.۵ می‌توان نتیجه گرفت که در مورد دو ماتریس عدم تشابه RF_{boot} و RF_{uni}، نتیجه‌ی خوشه‌بندی نسبت به نوع روش مقیاس‌گذاری چند بعدی حساس نیست، در واقع، اینکه کدامیک از روش‌های مقیاس‌گذاری جهت کاهش بعد انتخاب گردد، تفاوت چندانی در نتیجه خوشه‌بندی ایجاد نخواهد کرد در حالیکه در مورد فاصله اقلیدسی، کاهش بعد توسط مقیاس‌گذاری غیرمتری نسبت به مقیاس‌گذاری کلاسیک نتیجه بهتری را در پی داشته است.

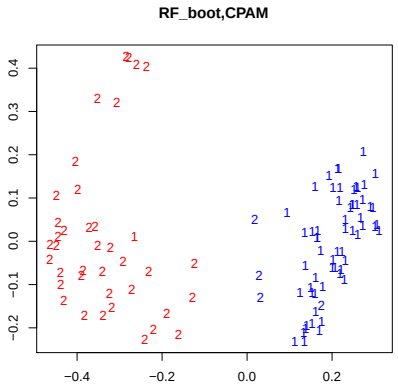
همچنین به منظور فراهم آوردن یک ارزیابی بصری، از نمودار (ب) شکل ۳.۵ به عنوان یک تجسم مناسب از مجموعه داده‌ی چاودری استفاده می‌کنیم. بدین منظور خوشه‌های تخمینی حاصل از اجرای هر

یک از رهیافت‌ها همزمان با خوشه‌های واقعی در نمودار مذکور لحاظ گشته و در شکل ۴.۵ نمایش داده شده‌اند. خوشه‌های تخمینی با دو رنگ قرمز و آبی از هم تفکیک شده‌اند و برچسب‌گذاری نقاط با اعداد 1 و 2 نمایانگر خوشه‌ی واقعی مشاهدات است. بدین ترتیب می‌توان با مشاهده‌ی این نمودارها، طبق جدول ۲.۴ موقعیت مشاهدات در خوشه‌های تخمینی و واقعی را تشخیص داده و خوشه‌بندی را ارزیابی نمود.

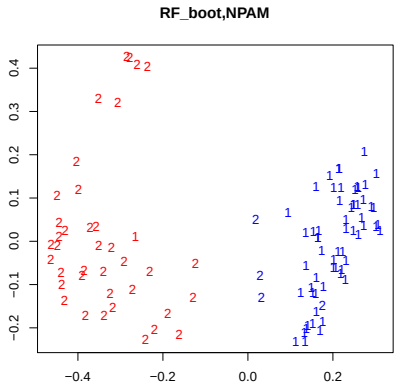
نمودار (آ) شکل ۴.۵ مربوط به رهیافت NPAM با اخذ عدم تشابه RF_{boot} است. همان‌گونه که در این نمودار ملاحظه می‌شود، از بین مشاهداتی که که با برچسب 1 روی نمودار مشخص‌شده‌اند، تنها ۱ مشاهده به رنگ قرمز درآمده و بقیه مشاهدات با رنگ آبی نشان داده شده‌اند، یعنی طی رهیافت مذکور از بین مشاهداتی که در واقعیت به خوشه‌ی ۱ تعلق دارند، تنها خوشه‌ی یک مشاهده، به اشتباه تشخیص داده شده‌است. همچنین از بین مشاهدات دارای برچسب 2 تنها سه مشاهده به رنگ آبی و بقیه با رنگ قرمز ظاهر شده‌اند یعنی مشاهدات متعلق به خوشه‌ی ۲ طی این رهیافت نیز به خوشه‌ی ۲ تعلق گرفته‌اند و در این میان تنها سه مشاهده به اشتباه به خوشه‌ی ۱ اختصاص داده‌شده‌اند. نقاط روی نمودار (ب) این شکل نیز کاملاً مشابه نمودار (آ) ظاهر شده‌اند و این نشان می‌دهد که دو رهیافت NPAM و CPAM با اخذ عدم تشابه RF_{boot}، خوشه‌بندی یکسانی را انجام داده‌اند. بنابراین نمودارهای مذکور مبین تطابق خوشه‌های تخمینی و واقعی و عملکرد مناسب رهیافت مبتنی بر عدم تشابه RF_{boot} واقعی بوده و نتایج ارزیابی مبتنی بر شاخص رند تعدیل یافته را تصدیق می‌کنند.

با توجه به اینکه در نمودارهای ۴.۵(ج)، ۴.۵(د) تعداد اندکی از مشاهدات به خوشه‌ی۲ نسبت داده شده‌اند، قطعاً عدم تشابه RF_{uni} در تشخیص ساختارهای موجود در داده‌ها ناتوان بوده است. به طور مشابه این اتفاق در نمودار ۴.۵(و) نیز رخ داده است، لذا خوشه‌بندی CPAM با اخذ فاصله اقلیدسی نیز یک خوشه‌بندی ضعیف تلقی می‌شود.

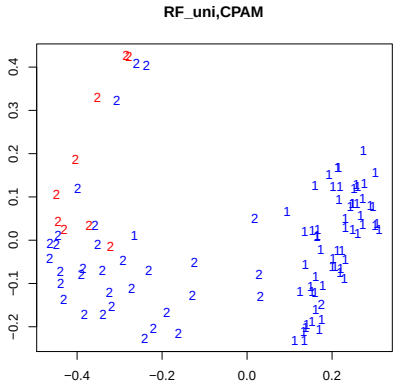
نتایج رهیافت NPAM با اعمال فاصله اقلیدسی روی نمودار (ه) منعکس شده‌است. بررسی میزان انطباق خوشه‌های واقعی با خوشه‌های تخمینی روی این نمودار نشان می‌دهد که عدم تطابق خوشه‌های تخمینی و واقعی تنها در مورد یازده مشاهده اتفاق افتاده‌است که از این تعداد یک مشاهده متعلق به خوشه‌ی ۱ و ده مشاهده متعلق به خوشه‌ی ۲ بوده‌اند. با توجه به اینکه خوشه‌ی اکثر مشاهدات به درستی تشخیص داده شده، لذا می‌توان نتیجه گرفت که رهیافت مذکور عملکرد مطلوبی را از خود نشان داده است.



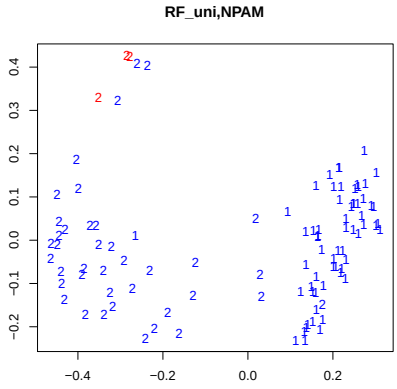
(ب)



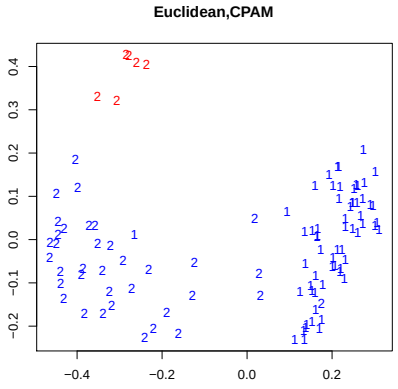
(آ)



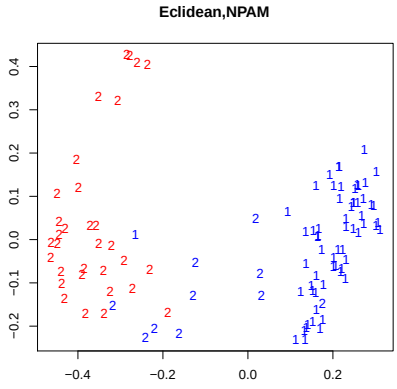
(د)



(ج)



(و)



(ه)

شکل ۴.۵: نمایش نتایج خوشه‌بندی داده چاودری در نمودارهای مقیاس‌گذاری چندبعدی مبتنی بر سه ماتریس عدم تشابه. ستون سمت راست مربوط به مقیاس‌گذاری کلاسیک و ستون سمت چپ مربوط به مقیاس‌گذاری غیرمتری است. سطرهای اول و دوم و سوم به ترتیب نمایانگر نمودارهای مقیاس‌گذاری چند بعدی مبتنی بر عدم تشابه RF_{boot}، RF_{uni} و فاصله اقلیدسی هستند. خوشه‌های تخمینی ۱ و ۲ به ترتیب با دو رنگ آبی و قرمز و خوشه‌های واقعی به ترتیب با دو برچسب 1 و 2 از هم تفکیک شده‌اند.

۴.۵ توصیف خوشه‌های تخمینی

منظور از توصیف خوشه‌های تخمینی، پیدا کردن قاعده‌ای متشکل از متغیرهای توضیحی برای بیان آن‌هاست. واقعیت این است که در ایجاد خوشه‌ها، تمامی متغیرهای توضیحی نقش ندارند بلکه زیرمجموعه‌ای از آن‌ها در خوشه‌بندی موثر هستند، بنابراین استفاده از تمامی متغیرها برای ساخت این قاعده، کاری بیهوده است. شناسایی متغیرهای موثر در خوشه‌بندی می‌تواند به ساخت قاعده‌ای ساده‌تر برای توصیف خوشه‌ها منجر شود و خوشه‌ها را تفسیرپذیر نماید. این موضوع بویژه در مواجهه با تعداد زیادی از متغیرهای توضیحی اهمیت پیدا می‌کند.

ارزیابی نتایج خوشه‌بندی چاودری در بخش قبل نشان داد که بهترین خوشه‌بندی توسط دو رهیافت NPAM و CPAM با اخذ عدم تشابه RF_{boot} صورت گرفته‌است و هر دو رهیافت، خوشه‌های کاملاً یکسانی را نتیجه داده‌اند. پس از آن رهیافت NPAM با اخذ فاصله‌ی اقلیدسی توانسته است خوشه‌بندی نسبتاً مطلوبی را نتیجه دهد. در این بخش در صدد یافتن قاعده‌ای ساده برای توصیف خوشه‌های تخمینی مبتنی بر دو عدم تشابه RF_{boot} و Δ هستیم.

۱.۴.۵ توصیف خوشه‌های تخمینی مبتنی بر عدم تشابه RF_{boot}

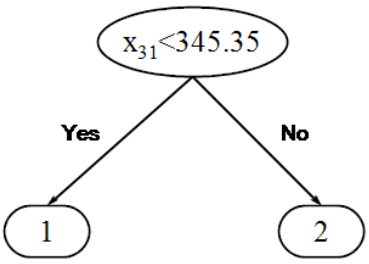
بر اساس آنچه در بخش ۳.۴ توضیح داده‌شد، ادعا می‌کنیم که خوشه‌های تخمینی مبتنی بر عدم تشابه RF_{boot} به متغیرهای موثر در تعریف این عدم تشابه(متغیرهای با اهمیت در رده‌بندی جنگل تصادفی) وابسته هستند و می‌توان خوشه‌های تخمینی را تنها بوسیله متغیرهای مذکور توصیف نمود. در این قسمت قصد داریم میزان صحت این ادعا را مورد بررسی قرار دهیم. بدین منظور ابتدا به شناسایی متغیرهای با اهمیت در رده‌بندی جنگل تصادفی می‌پردازیم. از بین متغیرهای توضیحی، ۱۰ متغیر با بیشترین میزان اهمیت را انتخاب می‌کنیم. متغیرهای انتخاب شده به ترتیب اولویت در جدول ۴.۵ درج شده‌اند.

جدول ۴.۵: متغیرهای با اهمیت در رده‌بندی جنگل تصادفی مربوط به عدم تشابه RF_{boot}

اولویت	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
متغیر	$X_{۶}$	$X_{۱۴۳}$	$X_{۷}$	$X_{۳۱}$	$X_{۱۱۶}$	$X_{۱۰۹}$	$X_{۱۶۹}$	$X_{۷۴}$	$X_{۱۷۱}$	$X_{۵۵}$

به منظور توصیف خوشه‌های تخمینی بوسیله‌ی متغیرهای با اهمیت فوق، مجموعه داده‌ی چاودری را تنها با این متغیرها در نظر گرفته و از سایر متغیرها صرف نظر می‌کنیم. اکنون اگر خوشه‌ی تخمین‌شده برای هر مشاهده را به عنوان مقدار متغیر پاسخ برای آن مشاهده در نظر بگیریم، یافتن قاعده‌ای برای توصیف

خوشه‌های تخمینی معادل با یک مسئله‌ی رده‌بندی است و با پیاده‌سازی یک روش رده‌بندی می‌توان به مدلی برای توصیف مقادیر متغیر پاسخ(خوشه‌های تخمینی) توسط متغیرهای با اهمیت دست یافت. در اینجا از روش رده‌بندی CART برای برای یافتن مدل رده‌بندی داده‌های مذکور استفاده شده‌است. ساختار درختی رده‌بندی صورت‌گرفته در شکل ۵.۵ نمایش داده شده‌است.



شکل ۵.۵: ساختار درختی رده‌بندی مجموعه داده‌ی چاودری با در نظر گرفتن ۱۰ متغیر با اهمیت به عنوان متغیرهای توضیحی و خوشه‌های تخمینی به عنوان متغیر پاسخ

بر اساس ساختار فوق، خوشه‌های تخمینی را تنها می‌توان بوسیله‌ی متغیر $X_{۳۱}$ بیان نمود بدین ترتیب که اگر مقدار متغیر $X_{۳۱}$ مشاهده‌ای بزرگ‌تر از مقدار ۳۴۵/۳۵ باشد، به خوشه‌ی ۱ تعلق دارد و در غیر اینصورت متعلق به خوشه‌ی ۲ است.

اکنون میزان درستی قاعده‌ی تعیین شده را به صورت بصری مورد بررسی قرار می‌دهیم. بدین منظور بار دیگر نمودار (ب) شکل ۳.۵ را به عنوان تجسم مناسبی از داده‌ها انتخاب می‌کنیم. مشاهدات متعلق به خوشه‌های تخمینی ۱ و ۲ را به ترتیب با دو علامت دایره و مثلث نمایش داده و آن‌ها را به صورت زیر رنگ‌آمیزی می‌کنیم:

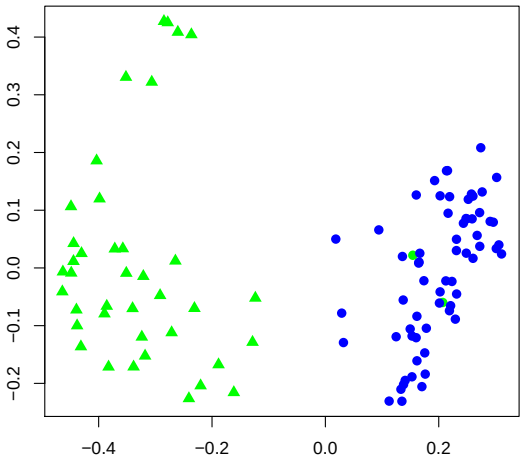
اگر $X_{۳۱} < ۳۴۵/۳۵$: آبی

اگر $X_{۳۱} > ۳۴۵/۳۵$: سبز

پس از اعمال تغییرات فوق نمودار حاصل در شکل ۶.۵ نمایش داده شده است. همان‌گونه که ملاحظه می‌شود، در نمودار مذکور تعداد سه دایره با رنگ سبز ظاهر شده و بقیه به رنگ آبی درآمده‌اند. آبی شدن دایره‌ها روی نمودار به این معنی است که مشاهداتی که خوشه‌ی ۱ برای آن‌ها تخمین زده شده است، در قاعده‌ی $X_{۳۱} < ۳۴۵/۳۵$ صدق می‌کنند.

همچنین ظاهر شدن مثلث‌ها با رنگ سبز نشان می‌دهد که مقدار متغیر $X_{۳۱}$ مشاهداتی که به خوشه‌ی ۲ نسبت داده شده‌اند، بزرگ‌تر از ۳۴۵/۳۵ است.

لذا طبق نمودار ۶.۵ فقط تعداد اندکی از مشاهدات (دو مشاهده) 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 روش CART به خوبی توانسته است خوشه‌ی تخمینی مشاهدات



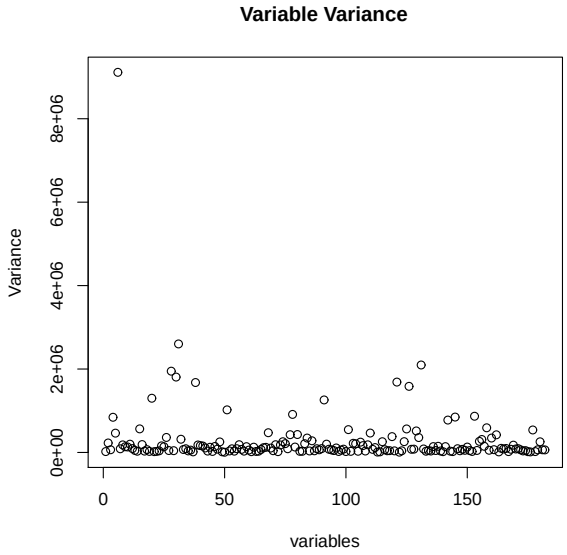
شکل ۶.۵: نمایش همزمان خوشه‌های تخمینی مبتنی بر عدم تشابه RF_{boot} و خوشه‌های تعیین شده توسط روش CART

را بر اساس متغیرهای با اهمیت توصیف کند و علیرغم اینکه عدم تشابه RF_{boot} به طور غیر مستقیم در خوشه‌بندی دخیل بوده‌است، توانستیم خوشه‌های حاصل از دو رهیافت NPAM و CPAM را توسط قاعده‌ای ساده که تنها از متغیر با اهمیت $X_{۳۱}$ ساخته شده‌است، توصیف کنیم.

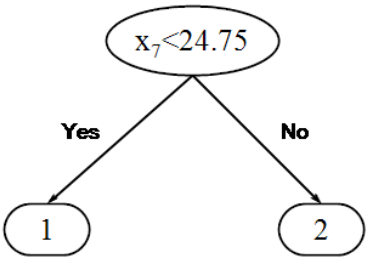
۲.۴.۵ توصیف خوشه‌های تخمینی مبتنی بر فاصله‌ی اقلیدسی

در این قسمت، قصد داریم توسط متغیرهای موثر در محاسبه‌ی فاصله‌ی اقلیدسی، قاعده‌ای را برای توصیف خوشه‌های حاصل از اعمال این نوع عدم تشابه تعیین نموده و سپس میزان اعتبار آن را مورد سنجش قرار دهیم.

به منظور شناسایی متغیرهای موثر در محاسبه‌ی فاصله‌ی اقلیدسی، نمودار واریانس متغیرها در شکل ۷.۵ ترسیم شده‌است. در این شکل، متغیر با بیشترین واریانس، متغیر $X_{۷}$ است. اکنون مشابه بخش قبل، با در نظر گرفتن متغیر $X_{۷}$ به عنوان متغیر توضیحی و خوشه‌های تخمینی به عنوان متغیر پاسخ، از رده‌بندی CART برای توصیف خوشه‌های تخمینی توسط متغیر $X_{۷}$ استفاده می‌کنیم. ساختار درختی این رده‌بندی در شکل ۸.۵ داده شده است. با توجه به این ساختار درختی، قاعده‌ی توصیف خوشه‌های تخمینی به این صورت تعیین می‌شود: اگر مقدار متغیر $X_{۷}$ مشاهده‌ای، کوچک‌تر از مقدار ۲۴/۷۵ باشد، مشاهده‌ی مذکور به خوشه‌ی ۱ و در غیراین‌صورت به خوشه‌ی ۲ تعلق گرفته‌است.



شکل ۷.۵: نمودار واریانس متغیرها در داده‌ی چاودری



شکل ۸.۵: ساختار ردختی رده‌بندی داده‌ی چاودری با در نظر گرفتن متغیر با بیشترین واریانس به عنوان متغیر توضیحی و خوشه‌های تخمینی به عنوان متغیر پاسخ

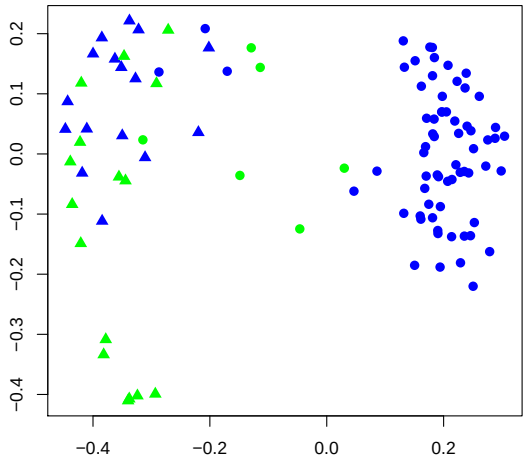
اکنون برای سنجش میزان اعتبار این قاعده، خوشه‌های تخمینی را روی تجسمی دو بعدی از داده‌ها منعکس می‌کنیم. خوشه‌های تخمینی ۱ و ۲ را به ترتیب با دو علامت دایره و مثلث روی این نمودار نمایش داده شده‌اند. همچنین قاعده‌ی تعیین‌شده را توسط دو رنگ متمایز و به صورت زیر روی نمودار مذکور اعمال می‌کنیم:

اگر $X_۷ < ۲۴/۷۵$: آبی

اگر $X_۷ > ۲۴/۷۵$: سبز

نمودار حاصل در شکل ۹.۵ قابل ملاحظه است.در صورتی‌که قاعده‌ی تعیین‌شده برای توصیف خوشه‌های تخمینی مناسب باشد، می‌بایست دایره‌ها با رنگ آبی و مثلث‌ها با رنگ سبز روی نمودار ۹.۵ ظاهر شوند. اما همانگونه که ملاحظه می‌کنید روی این نمودار، تعداد دایره‌های سبز و مثلث‌های

آبی(تعداد خطاها) قابل چشم‌پوشی نیست. بنابراین توصیف خوشه‌های تخمینی توسط این قاعده با خطاهای زیادی همراه بوده و قاعده‌ی تعیین‌شده قادر به توصیف خوشه‌ها نیست. در نتیجه در خوشه‌بندی



شکل ۹.۵: نمایش همزمان خوشه‌های تخمینی مبتنی بر عدم تشابه Δ و خوشه‌های تعیین شده توسط روش CART

مبتنی بر فاصله‌ی اقلیدسی، نمی‌توان به قاعده‌ای ساده برای توصیف خوشه‌ها دست یافت.

۵.۵ تحلیل حساسیت خوشه‌بندی نسبت به پارامترهای جنگل تصادفی

همانطورکه پیش از این نیز اشاره شد، در اجرای رده‌بندی جنگل تصادفی و به دنبال آن یافتن عدم تشابه حاصل از این روش، ملزم به تعیین پارامترهای $mtry$ و $ntree$ هستیم. در بخش ۴.۴ تاثیر پارامتر $mtry$ در رده‌بندی و خوشه‌بندی جنگل تصادفی در مورد مثالهای شبیه‌سازی مورد بررسی قرار گرفت.در این بخش قصد داریم با تمرکز بر روی هر دو پارامتر $mtry$ و $ntree$، حساسیت نتیجه‌ی خوشه‌بندی مبتنی بر عدم تشابه جنگل تصادفی نسبت به پارامترهای مذکور را به طور کامل‌تر مورد مطالعه قرار دهیم.

بدین منظور پارامتر $mtry$ در مجموعه‌ی $M = \{۱, ۴, ۷, \cdots, ۲۸\}$ و پارامتر $ntree$ در مجموعه‌ی $N = \{۱۰۰, ۲۰۰, \cdots, ۱۰۰۰\}$ در نظر گرفته شده است. به ازای هر زوج پارامتر $(mtry, ntree) \in M \times N$، دو ماتریس عدم تشابه RF_{boot} و RF_{uni} محاسبه شده و هر یک از آن‌ها به

طور مجزا در رهیافت CPAM اعمال شده‌اند. پس از ارزیابی نتایج در هر حالت، به این نتیجه رسیدیم که به ازای مقادیر مختلف از پارامترهای مذکور تغییر چندانی در مقدار شاخص رند تعدیل‌یافته رخ نمی‌دهد لذا می‌توان نتیجه گرفت که در مورد مجموعه‌ی داده‌ی چاودری، خوشه‌بندی نسبت به پارامترهای جنگل تصادفی تقریباً پایدار است. قطعاً این نتیجه را نمی‌توان همیشه برقرار دانست و ممکن است در مورد برخی داده‌ها، خوشه‌بندی حساسیت بیشتری نسبت به پارامترهای جنگل تصادفی داشته باشد.

انتظار می‌رود که با افزایش دقت رده‌بندی(کاهش نرخ خطا)، عدم تشابه دقیق‌تری حاصل شود و تخمین بهتری از خوشه‌ها (افزایش شاخص رند تعدیل‌یافته) صورت گیرد. اکنون با محاسبه‌ی مقادیر نرخ خطای رده‌بندی به ازای مقادیر متفاوت $mtry$ و $ntree$ چگونگی برآورده‌شدن این انتظار را مورد بررسی قرار می‌دهیم. بدین منظور نرخ خطای رده‌بندی در محاسبه‌ی هر یک از ماتریس‌های عدم تشابه RF_{boot} و RF_{uni} را به ازای هر زوج $(mtry, ntree) \in M \times N$، به‌دست آوردیم. این بررسی نشان داد که در محاسبه‌ی هر دو ماتریس عدم تشابه به ازای مقادیر مختلف از پارامترهای جنگل تصادفی نرخ خطای رده‌بندی پایین و تقریباً ثابت بوده اما کمترین نرخ خطا به ازای $mtry > ۷$ و $ntree > ۵۰۰$ رخ داده است. بررسی مقادیر شاخص رند تعدیل یافته با همین تنظیمات نشان می‌دهد که به ازای این تنظیمات نه تنها افزایشی در مقدار شاخص رند تعدیل یافته رخ نداده بلکه در مورد عدم تشابه RF_{uni} کاهش نیز داشته است لذا نمی‌توان رابطه‌ی معنی داری را بین شاخص رند تعدیل یافته و نرخ خطای رده‌بندی تعریف نمود.

بنا بر بررسی انجام شده در این بخش و نیز بخش ۴.۴ به این نتیجه می‌رسیم که لزوماً مقادیری از پارامترهای $ntree$ و $mtry$ که دقت رده‌بندی را افزایش می‌دهند، منجر به خوشه‌بندی بهینه نخواهند شد. بنابراین انتخاب مقادیر دو پارامتر $mtry$ و $ntree$ بر اساس کاهش نرخ خطای رده‌بندی قابل اعتماد نیست و این انتخاب می‌بایست با هدف بهبود کیفیت خوشه‌بندی صورت گیرد. برای این بهینه‌سازی می‌توان از معیارهای داخلی خوشه‌بندی بهره گرفت. به عنوان مثال می‌توان برای یافتن مقادیر بهینه پارامترهای $ntree$ و $mtry$، رویه‌ای مشابه بهینه‌سازی تعداد خوشه‌ها از طریق میانگین شاخص نیمرخ را استفاده نمود، یعنی به ازای مقادیر مختلف پارامتر $mtry$ و $ntree$، رده‌بندی جنگل تصادفی اجرا شده و عدم تشابه حاصل از آن را در خوشه‌بندی اعمال کرد و سپس میانگین شاخص نیمرخ متناظر با هر خوشه‌بندی محاسبه شود . مقادیر بهینه $mtry$ و $ntree$ مقادیری هستند که به ازای آن‌ها میانگین شاخص نیمرخ ماکزیمم شده باشد.

۶.۵ نتیجه‌گیری

در این پایان‌نامه معیاری جدید برای سنجش عدم تشابه مشاهدات بر مبنای روش رده‌بندی جنگل تصادفی معرفی شد. در یک مطالعه‌ی شبیه‌سازی کارآیی این نوع عدم تشابه نسبت به فاصله اقلیدسی در روش خوشه‌بندی افراز حول مدوید مورد ارزیابی قرار گرفت و خواص آن معرفی گردید. همچنین عدم تشابه مذکور در ترکیبی از روش مقیاس‌گذاری چند بعدی و افراز حول مدوید به منظور خوشه‌بندی مجموعه داده‌ای با اهمیت در حوزه‌ی پزشکی اعمال گردید و نتایج حاصل از آن با نتایج اخذ شده از اعمال فاصله اقلدسی مورد مقایسه قرار گرفت. با توجه به نتایج به‌دست آمده از مطالعه‌ی شبیه‌سازی و خوشه‌بندی مجموعه داده‌ی واقعی می‌توان نتیجه گرفت که عدم تشابه جنگل تصادفی انتخاب مناسبی در مواجهه با داده‌هایی است که در آن‌ها تعداد متغیرهای توضیحی بسیار زیاد بوده و حتی بیشتر از تعداد مشاهدات است. همچنین توسط قابلیت منحصر به‌فرد تعیین اهمیت متغیرها در رده‌بندی جنگل تصادفی می‌توان متغیرهای موثر در خوشه‌بندی را شناسایی نموده و خوشه‌های تخمینی را توسط قاعده‌ای ساده توصیف نمود.

۷.۵ پیشنهادات

به منظور انجام تحقیقات و پژوهش‌های آتی در زمینه‌ی خوشه‌بندی جنگل تصادفی، می‌توان از راهکارهای ذیل بهره گرفت:

- از آنجا که روش‌های خوشه‌بندی مختلف ممکن است به نتایج بسیار متفاوتی منجر شوند، لذا می‌توان کارآیی عدم تشابه جنگل تصادفی را در سایر روش‌های مبتنی بر عدم تشابه بررسی نمود. بدین منظور می‌توان از انواع روش‌های سلسله مراتبی بهره گرفت.
- می‌توان به منظور دست یافتن به بهترین خوشه‌بندی، پارامترهای $mtry$ و $ntree$ را نسبت به معیارهای درونی ارزیابی خوشه‌بندی مانند شاخص نیمرخ بهینه نمود.
- با استفاده از خوشه‌بندی اجماعی می‌توان از چندین پارامتر $mtry$ و $ntree$ به طور همزمان در محاسبه‌ی عدم تشابه جنگل تصادفی استفاده کرد و یا عدم تشابه حاصل از این روش را به طور همزمان در چندین روش خوشه‌بندی اعمال نمود.
- همچنین استفاده از دیگر روش‌های کاهش بعد قبل از اجرای خوشه‌بندی جنگل تصادفی در خور توجه می‌باشد.

پیوست آ

روش‌های عددی

آ.۱ رگرسیون یکنوا

فرض کنید $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=۱}^{n}$ مجموعه داده‌ای دو بعدی است. داده‌ها را به صورت زیر مرتب می‌کنیم:

$$\mathbf{x}_{(۱)} \leq \mathbf{x}_{(۲)} \leq \mathbf{x}_{(۳)} \leq \cdots \leq \mathbf{x}_{(n)}$$

$\mathbf{y}_{(i)}$ را $\mathbf{y}$ متناظر با $\mathbf{x}_{(i)}$ قرار می‌دهیم. مسئله رگرسیون یکنوا پیدا کردن مجموعه $\{\hat{\mathbf{y}}_{(i)}\}_{i=۱}^{n}$ است به طوریکه عبارت ^{۲} $\frac{۱}{n}\sum_{i=۱}^{n}(\mathbf{y}_{(i)} - \hat{\mathbf{y}}_{(i)})$ را مینیمم کند و در قید یکنوایی صدق کند یعنی:

$$\hat{\mathbf{y}}(۱) \leq \hat{\mathbf{y}}_{(۲)} \leq \hat{\mathbf{y}}_{(۳)} \leq \cdots \leq \hat{\mathbf{y}}_{(n)}. \qquad (آ.۱)$$

یکی از مرسوم‌ترین الگوریتم‌ها در انجام رگرسیون یکنوا، الگوریتم ادغام مجاورهای ناقض ^{۱} (PAV) است که در ادامه آن را شرح می‌دهیم.

بررسی برقرار بودن یا نبودن قید یکنوایی را به ترتیب از مقدار $\mathbf{y}$ متناظر با $\mathbf{x}_{(۱)}$ یعنی $\mathbf{y}_{(۱)}$ آغاز می‌کنیم. اگر جفت $(\mathbf{y}_{(i)}, \mathbf{y}_{(i+۱)})$ قید یکنوایی را نقض کردند یعنی داشته باشیم: $\mathbf{y}_{(i)} \geq \mathbf{y}_{(i+۱)}$ ، در اینصورت جهت برقراری یکنوایی می‌بایست مقادیر جدیدی که حاصل ادغام دو مجاور مذکور هستند، جایگزین آن‌ها گردند. بدین منظور میانگین $\mathbf{y}_{(i)}$ و $\mathbf{y}_{(i+۱)}$ را محاسبه می‌کنیم. بنابراین داریم:

$$\mathbf{y}_{(i)} = \mathbf{y}_{(i+۱)} = \frac{\mathbf{y}_{(i)} + \mathbf{y}_{(i+۱)}}{۲}$$

در مرحله بعد $\mathbf{y}_{(i-۱)}$ را با $\mathbf{y}_{(i)}$ مقایسه می‌کنیم. در صورت برقراری قید یکنوایی به سراغ دو مجاور بعد می‌رویم ، در غیر اینصورت میانگین $\mathbf{y}_{(i-۱)}$، $\mathbf{y}_{(i)}$ و $\mathbf{y}_{(i+۱)}$، جایگزین هر یک از آن‌ها خواهد شد.

---

[۱] Pool Adjacent Violators

روند فوق ادامه می‌یابد تا زمانیکه به $\mathbf{y}_{(n)}$ برسیم و یکنوایی به ازای همه $\mathbf{y}_{(i)}$ها برقرار شده باشد. $\mathbf{y}_{(i)}$های نهایی، $\hat{\mathbf{y}}_{(i)}$هایی هستند که به دنبال پیدا کردن آن‌ها بودیم [۵۱].

۲.آ روش شیب نزولی

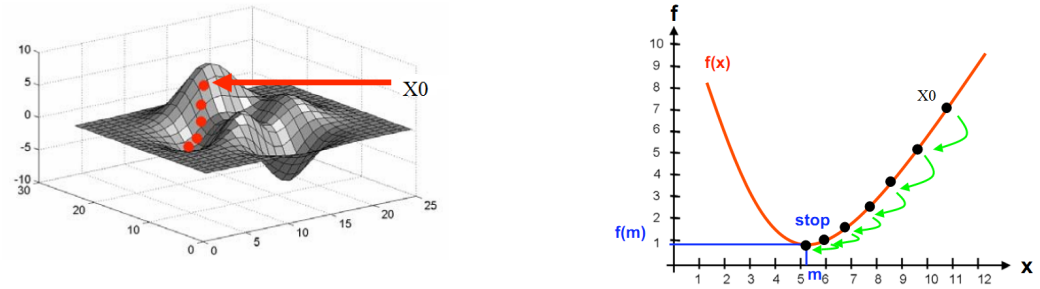
روش شیب نزولی یک روش عددی است که برای پیداکردن نزدیک‌ترین مینیمم محلی تابع $\mathbf{y} = f(\mathbf{x})$ به‌کار می‌رود. این روش برای تعیین نقطه‌ی مینیمم، روی ابر رویه‌ی تابع، گام به گام به سمت پایین حرکت می‌کند تا به نقطه‌ی مورد نظر برسد. $-\bigtriangledown f(\mathbf{x})$ جهت مسیر حرکت به سمت مینیمم و کاهش مقدار تابع f را نشان می‌دهد. طول گام حرکت مقداری دلخواه است که انتخاب مناسب آن در همگرایی به مینیمم دارای اهمیت است. حرکت به سمت مینیمم از یک نقطه‌ی اولیه مانند $\mathbf{x}_۰$ آغاز می‌شود، نقطه‌ی $\mathbf{x}_۱ = \mathbf{x}_۰ - \alpha \bigtriangledown f(\mathbf{x}_۰)$ نقطه‌ی بعدی است که به مینیمم f نزدیک‌تر است. با ادامه پیشروی به سمت پایین و در امتداد دنباله‌ای از نقاط انتظار می‌رود که به نزدیکترین نقطه‌ی مینیمم f برسیم. دنباله نقاط مذکور از طریق فرمول زیر محاسبه می‌شوند:

$$\mathbf{x}_{i+۱} = \mathbf{x}_i - \alpha \bigtriangledown f(\mathbf{x}_i) \qquad (۲.آ)$$

این روش زمانی متوقف می‌شود که اختلاف دو مقدار $\mathbf{x}_i$ و $\mathbf{x}_{i+۱}$ از مقدار تعیین شده ϵ کوچک‌تر شود به عبارتی:

$$\triangle \mathbf{x} < \epsilon \qquad (۳.آ)$$

شکل آ.۱ نحوه‌ی یافتن مینیمم توسط روش شیب نزولی را در مورد یک تابع دو متغیره و چند متغیره نمایش می‌دهد [۵۱].



(آ) تعیین مینیمم یک تابع یک متغیره توسط روش شیب نزولی

(ب) تعیین مینیمم یک تابع چند متغیره توسط روش شیب نزولی

شکل آ.۱: تعیین مینیمم تابع توسط روش شیب نزولی

مثال آ.۱. تابع $f(\mathbf{x}) = \mathbf{x}^۲ - ۲\mathbf{x} + ۳$ را در نظر بگیرید. به راحتی می‌توان مینیمم این تابع را به صورت زیر محاسبه نمود:

$$f'(\mathbf{x}) = ۲\mathbf{x} - ۲ = ۰$$

نقطه‌ی $\mathbf{x} = ۱$ در معادله‌ی بالا صدق می‌کند و چون $f''(۱) = ۲ > ۰$ ، در نتیجه $\mathbf{x} = ۱$ مینیمم تابع است. اکنون قصد داریم با استفاده از روش عددی شیب نزولی و به ازای $\epsilon = ۰/۰۵$ مینیمم این تابع را پیدا کنیم. با استفاده از معادله‌ی تکرار $\mathbf{x}_{i+۱} = \mathbf{x}_i - \alpha(۲\mathbf{x} - ۲)$ و در نظر گرفتن $\alpha = ۰/۴$ و $\mathbf{x}_۱ = ۷$، نتیجه‌ی محاسبه در تکرارهای مختلف در جدول آ.۱ درج شده‌است.

جدول آ.۱: نتایج محاسبه‌ی مینیمم تابع $f(\mathbf{x})$ در تکرارهای مختلف

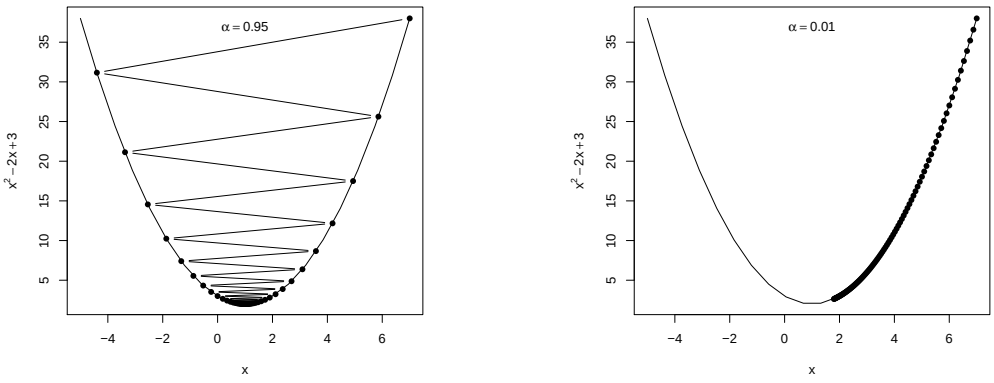
i	$\mathbf{x}_i$	$\Delta\mathbf{x}_i$
۱	۷	–
۲	۲/۲۰	۴/۸
۳	۱/۲۴	۰/۹۶
۴	۱/۰۴۸	۰/۱۹۲
۵	۱/۰۰۹۶	۰/۰۳۴۸

طبق جدول فوق، در تکرار پنجم با کوچک‌تر شدن مقدار $\Delta\mathbf{x}_i$ از عدد ۰/۰۵، الگوریتم متوقف گشته است. در این تکرار مقدار مینیمم تابع f برابر با عدد ۱/۰۰۹۶ به‌دست آمده‌است که بسیار نزدیک به مینیمم واقعی تابع می‌باشد. بنابراین پس از پنج تکرار توانستیم به تقریب خوبی از مینیمم تابع برسیم. تقریب مینیمم تابع مذکور به ازای دو مقدار دیگر از پارامتر α در شکل آ.۲ داده شده‌است.

همانطور که در نمودار آ.۲(آ) ملاحظه می‌شود، به ازای $\alpha = ۰/۰۱$ تکرارهای زیادی صورت گرفته و مینیمم تابع به درستی تشخیص داده نشده‌است. نمودار آ.۲(ب) نیز نشان می‌دهد که با در نظر گرفتن $\alpha = ۰/۹۵$ نیز تعداد تکرارهای زیادی برای رسیدن به مینیمم تابع لازم است. بنابراین پارامتر α به شدت در سرعت همگرایی و تقریب صحیح مینیمم تاثیرگذار است.

۱.۲.آ نحوه‌ی انتخاب α در الگوریتم مقیاس‌گذاری غیرمتری

در الگوریتم مقیاس‌گذاری غیرمتری از روش عددی شیب نزولی استفاده می‌شود. طول گام مورد نیاز برای اجرای روش عددی مذکور(پارامتر α) در الگوریتم مقیاس‌گذاری غیرمتری به پیشنهاد کروسکال به صورت زیر تعیین می‌شود[۳۱]:



(ب) (آ)

شکل آ.۲: تاثیر پارامتر α در همگرایی روش شیب نزولی

در اولین تکرار مقدار این پارامتر ۰/۲ در نظرگرفته می‌شود و در تکرارهای بعد از رابطه‌ی زیر محاسبه می‌گردد:

$$\alpha_{new} = \alpha_{old} \times \beta \times \gamma \times \eta \qquad (آ.۴)$$

که در آن:

$$\beta = ۴\cos^{۳}\theta \qquad (آ.۵)$$

$$\gamma = \frac{۱/۳}{۱ + \gamma_{۰}^{۵}} \qquad (آ.۶)$$

و

$$\eta = \min\{۱, \frac{\text{مقدار استرس فعلی}}{\text{مقدار استرس قبلی}}\} \qquad (آ.۷)$$

در رابطه‌ی (آ.۵)، θ زاویه‌ی بین گرادیان فعلی و گرادیان قبلی است. اگر گرادیان فعلی را g و گرادیان قبلی را g' فرض کنیم آنگاه:

$$\cos\theta = \frac{\sum gg'}{\sqrt{\sum g^{۲}}\sqrt{\sum g'^{۲}}}$$

همچنین در رابطه‌ی (آ.۶):

$$\gamma_{۰} = \min\{۱, \frac{\text{مقدار استرس فعلی}}{\text{مقدار استرس ۵ تکرار قبل}}\}$$

اگر تعداد تکرارها کمتر از ۵ باشد برای محاسبه‌ی $\gamma_{۰}$ مقدار استرس تکرار اول به جای مقدار استرس ۵ تکرار قبل قرار می‌گیرد.

پیوست ب

دستورات نرم‌افزار R

در این پیوست دستورات مربوط به بدست آوردن عدم تشابه جنگل تصادفی، اجرای مقیاس‌گذاری چند بعدی و همین‌طور الگوریتم خوشه‌بندی PAM ارائه شده است.

دستورات مربوط به خواندن داده‌ها در R، تولید داده‌های مصنوعی، افزودن آنها به داده‌های اصلی و برچسب زدن مجموعه داده‌ی جدید:

```
data=read.csv("e:/data.csv",header=FALSE,sep=",",dec=".")
obs=dim(data)[[1]]
var=dim(data)[[2]]
trueclusters=data[,var]
x=data[1:obs,1:var-1]

n.row=dim(x)[[1]]
n.col=dim(x)[[2]]

##### method 1 : bootstrape sample ##########
synboot=matrix(0,nrow=n.row,ncol=n.col)
for(j in 1:n.col){
 synboot[,j]=sample(x[,j],replace=TRUE)
}
colnames(synboot)=colnames(x)
```

```
dataframe=rbind(x,synboot)
y<-rep(c(1,2),c(n.row,n.row))
addsyn_boot=data.frame(cbind(y,dataframe))

##### method 2 : uniform distrobution#########
synuni=matrix(0,nrow=n.row,ncol=n.col)
for(i in 1:n.row){
for(j in 1:n.col){
 synuni[,j]=runif(length(x[,j]), min=min(x[,j]), max =max(x[,j]))
}
}
colnames(synuni)=colnames(x)
dataframe=rbind(x,synuni)
y<-rep(c(1,2),c(n.row,n.row))
addsyn_uni=data.frame(cbind(y,dataframe))
```

دستورات مربوط به اجرای رده‌بندی جنگل تصادفی و بدست آوردن عدم تشابه مربوط به آن:

```
############RF_boot############
library(randomForest)
prox_boot=matrix(0,nrow=n.row,ncol=n.row)
importance_boot=matrix(0,nrow=n.col,ncol=4)
err_boot=0
for(i in 1:nforest){
y_boot=addsyn_boot[,1]
rf_boot<-randomForest(factor(y_boot)~.,data=addsyn_boot[,-1], ntree=ntree,
oob.prox=FALSE, proximity=TRUE,do.trace=F,mtry=mtry,importance=TRUE)
rfprox_boot=rf_boot$proximity
prox_boot=prox_boot+(rfprox_boot[c(1:n.row),c(1:n.row)])
importance_boot= importance_boot+rf_boot$importance
err_boot=err_boot+ rf_boot$err.rate[ntree]
```

```
}
importance_boot=importance_boot/nforest
err_boot=err_boot/nforest
disrf_boot=(sqrt(1-prox_boot[1:n.row,1:n.row]/nforest))



################RF_uni#####################
prox_uni=matrix(0,nrow=n.row,ncol=n.row)
importance_uni=matrix(0,nrow=n.col,ncol=4)
err_uni=0
for(i in 1:nforest){
y_uni=addsyn_uni[,1]
rf_uni<-randomForest(factor(y_uni)~.,data=addsyn_uni[,-1], ntree=ntree,
oob.prox=FALSE, proximity=TRUE,do.trace=F,mtry=mtry,importance=TRUE)
rfprox=rf_uni$proximity
prox_uni=prox_uni+(rfprox[c(1:n.row),c(1:n.row)])
importance_uni= importance_uni+rf_uni$importance
err_uni=err_uni+ rf_uni$err.rate[ntree]
}
importance_uni=importance_uni/nforest
err_uni=err_boot/nforest
disrf_uni=(sqrt(1-prox_uni[1:n.row,1:n.row]/nforest))
```

دستور محاسبه‌ی فاصله‌ی اقلیدسی:

```
dist_euclid=dist(x,method="euclidean")
```

دستورات مربوط به پیدا کردن بعد مناسب برای مقیاس‌گذاری چند بعدی کلاسیک:

```
cmd_eclid=cmdscale(as.dist(dist_euclid),n.row-1,eig=TRUE)
cumsum(abs(cmd_eclid$eig))/sum(abs(cmd_eclid$eig))
cumsum((cmd_eclid$eig)^2)/sum((cmd_eclid$eig)^2)
```

```
cmd_boot=cmdscale(as.dist(disrf_boot),n.row-1,eig=TRUE)
cumsum(abs(cmd_boot$eig))/sum(abs(cmd_boot$eig))
cumsum((cmd_boot$eig)^2)/sum((cmd_boot$eig)^2)

cmd_uni=cmdscale(as.dist(disrf_uni),n.row-1,eig=TRUE)
cumsum(abs(cmd_uni$eig))/sum(abs(cmd_uni$eig))
cumsum((cmd_uni$eig)^2)/sum((cmd_uni$eig)^2)
```

دستورات مربوط به پیدا کردن بعد مناسب برای مقیاس‌گذاری چند بعدی غیرمتری:

```
library(MASS)

maxq=20
iso=matrix(0,nrow=maxq,ncol=2)
for(k in 1:maxq){
iso[k,1]=k
iso[k,2]=isoMDS(as.dist(dist_euclid),k=k)$stress
}
plot(iso[,1],iso[,2],xlab="q(euclidean)", ylab="minimum Stress",type="o")

maxq=20
iso=matrix(0,nrow=maxq,ncol=2)
for(k in 1:maxq){
iso[k,1]=k
iso[k,2]=isoMDS(as.dist(disrf_boot),k=k)$stress
}
plot(iso[,1],iso[,2],xlab="q (RF_boot)", ylab="minimum Stress",type="o")

maxq=20
```

```
iso=matrix(0,nrow=maxq,ncol=2)
d=disrf_uni
for(k in 1:maxq){
iso[k,1]=k
iso[k,2]=isoMDS(as.dist(d),k=k)$stress
}
plot(iso[,1],iso[,2],xlab="q (RF_uni)", ylab="minimum Stress",type="o")
```

دستورات مربوط به کاهش بعد داده‌ها توسط مقیاس‌گذاری چند بعدی :

```
library(MASS)
cmd_euclid=cmdscale(as.dist(dist_euclid),2)
cmdrf_boot=cmdscale(as.dist(disrf_boot),2)
cmdrf_uni=cmdscale(as.dist(disrf_uni),2)
iso_euclid=isoMDS(as.dist(dist_euclid),k=2)$points
isorf_boot=isoMDS(as.dist(disrf_boot),k=2)$points
isorf_uni=isoMDS(as.dist(disrf_uni),k=2)$points
```

دستورات مربوط به تعیین تعداد مناسب خوشه‌ها توسط شاخص نیمرخ:

```
library(fpc)
n1=pamk(dist(cmd_euclid))$nc
n2=pamk(dist(cmdrf_boot))$nc
n3=pamk(dist(cmdrf_uni))$nc
n4=pamk(dist(iso_euclid))$nc
n5=pamk(dist(isorf_boot))$nc
n6=pamk(dist(isorf_uni))$nc
```

دستورات مربوط به خوشه‌بندی PAM:

```
pamcmd_eclid=pam(dist(cmd_eclid),k=n1)$cluster
pamcmdrf_boot=pam(dist(cmdrf_boot),k=n2)$cluster
```

```
pamcmdrf_uni=pam(dist(cmdrf_uni),k=n3)$cluster
pamiso_eclid=pam(dist(iso_eclid),k=n4)$cluster
pamisorf_boot=pam(dist(isorf_boot),k=n5)$cluster
pamisorf_uni=pam(dist(isorf_uni),k=n6)$cluster
```

دستورات مربوط به محاسبه‌ی شاخص رند تعدیل یافته:

```
library(mclust)

randpamcmd_eclid=adjustedRandIndex(trueclusters,pamcmd_eclid)
randpamcmdrf_boot=adjustedRandIndex(trueclusters,pamcmdrf_boot)
randpamcmdrf_uni=adjustedRandIndex(trueclusters,pamcmdrf_uni)
randpamiso_eclid=adjustedRandIndex(trueclusters,pamiso_eclid)
randpamisorf_boot=adjustedRandIndex(trueclusters,pamisorf_boot)
randpamisorf_uni=adjustedRandIndex(trueclusters,pamisorf_uni)

rbind(randpamcmd_eclid,randpamcmdrf_boot,randpamcmdrf_uni,
randpamiso_eclid,randpamisorf_boot,randpamisorf_uni)
```

دستورات مربوط به رسم نمودارهای مقیاس‌گذاری چند بعدی و منعکس کردن نتایج روی نمودارها:

```
plot(cmd_boot[,1],cmd_boot[,2],xlab=NA,ylab=NA,main="RF_boot,cMDS",
col=pam_boot,pch=as.character(trueclusters))
plot(iso_boot[,1],cmd_boot[,2],xlab=NA,ylab=NA,main="RF_boot,nMDS",
col=pam_boot,pch=as.character(trueclusters))
plot(cmd_uni[,1],cmd_uni[,2],xlab=NA,ylab=NA,main="RF_uni,cMDS",
col=pam_uni,pch=as.character(trueclusters))
plot(iso_uni[,1],iso_uni[,2],xlab=NA,ylab=NA,main="RF_uni,nMDS",
col=pam_uni,pch=as.character(trueclusters))
plot(cmd_eclid[,1],cmd_eclid[,2],xlab=NA,ylab=NA,main="Euclidean,cMDS",
col=pam_ecli,pch=as.character(trueclusters))
```

```
plot(iso_eclid[,1],iso_eclid[,2],xlab=NA,ylab=NA,main="Euclidean,nMDS",
col=pam_ecli,pch=as.character(trueclusters))
```

دستورات مربوط به رسم نمودار اهمیت متغیرها در رده‌بندی جنگل تصادفی و رسم نمودار واریانس متغیرها:

```
plot(importance_boot[,4],xlab="variables",ylab="importance",
main="RF_boot, Variable Imp.")
plot(importance_uni[,4],xlab="variables",ylab="importance",
 main="RF_uni, Variable Imp.")
plot(apply(x,2,var) ,xlab="variables",ylab="Variance",
main="Variable Variance")
```

دستورات مربوط به توصیف خوشه‌های تخمینی مبتنی بر عدم تشابه RF_{boot} و فاصله‌ی اقلیدسی با استفاده از رده‌بندی CART:

```
x=cbind(x,pamcmd_boot)
fit=rpart(factor(x[,183])~x[,6]+x[,143]+x[,7]+x[,31]+x[,116]+x[,109]+
x[,169]+x[,74]+x[,171]+x[,55],method="class")

par(mfrow = c(1,2), xpd = NA)
plot(fit)
text(fit, use.n=TRUE)

col1=ifelse(x[,178]<27.65,"blue","green")
pch=ifelse(pamcmd_boot==1,16,17)
plot(cmdrf_boot,pch=pch,xlab=NA,ylab=NA,col=col1,cex=1.2)

x=cbind(x,pamiso_eclid)
fit=rpart(factor(x[,183])~x[,7],method="class")

par(mfrow = c(1,2), xpd = NA)
```

```
plot(fit)
text(fit, use.n=TRUE)

col1=ifelse(x[,7]<24.75,"blue","green")
pch=ifelse(pamiso_eclid==1,16,17)
plot(cmdrf_boot,pch=pch,xlab=NA,ylab=NA,col=col1,cex=1.2)
```

دستورات مربوط به بررسی نرخ خطای رده‌بندی و شاخص رند تعدیل‌یافته به ازای مقادیر مختلف $mtry$ و $ntree$:

```
trees=seq(100,1000,100)
mtrys=seq(1,30,3)
ntrees=as.vector(trees)
randboot=matrix(0,nrow=10,ncol=10)
randuni=matrix(0,nrow=10,ncol=10)
errboot=matrix(0,nrow=10,ncol=10)
erruni=matrix(0,nrow=10,ncol=10)
for(i in 1:10){
for(j in 1:10){
err_boot=0
prox_boot=matrix(0,nrow=n.row,ncol=n.row)
for(m in 1:nforest){
mtry=mtrys[j]
ntree=ntrees[i]
y_boot=addsyn_boot[,1]
rf_boot<-randomForest(factor(y_boot)~.,data=addsyn_boot[,-1],ntree=ntree,
 oob.prox=FALSE, proximity=TRUE,do.trace=F,mtry=mtry,importance=FALSE)
rfprox_boot=rf_boot$proximity
prox_boot=prox_boot+(rfprox_boot[c(1:n.row),c(1:n.row)])
err_boot=err_boot+ rf_boot$err.rate[ntree]
```

```
}
err_boot=err_boot/nforest
disrf_boot=(sqrt(1-prox_boot[1:n.row,1:n.row]/nforest))

err_uni=0
prox_uni=matrix(0,nrow=n.row,ncol=n.row)
for(t in 1:nforest){
y_uni=addsyn_uni[,1]
rf_uni<-randomForest(factor(y_uni)~.,data=addsyn_uni[,-1], ntree=ntree,
oob.prox=FALSE, proximity=TRUE,do.trace=F,mtry=mtry,importance=FALSE)
rfprox_uni=rf_uni$proximity
prox_uni=prox_uni+(rfprox_uni[c(1:n.row),c(1:n.row)])
err_uni=err_uni+ rf_uni$err.rate[ntree]
}
err_uni=err_boot/nforest
disrf_uni=(sqrt(1-prox_uni[1:n.row,1:n.row]/nforest))



for(k in 1:103){
cmd_boot=cmdscale(as.dist(disrf_boot),k,eig=TRUE)
if(cmd_boot$GOF[1]>=0.8){
optimk=k
break()
}}
cmd_boot=cmdscale(as.dist(disrf_boot),optimk)
n1=pamk(dist(cmd_boot))$nc
pam_boot=pam(dist(cmd_boot),n1)$cluster
randboot[i,j]=round(adjustedRandIndex(trueclusters,pam_boot),4)
```

```
for(q in 1:103){
cmd_uni=cmdscale(as.dist(disrf_uni),k,eig=TRUE)
if(cmd_uni$GOF[1]>=0.8){
optimq=q
break()
}}
cmd_uni=cmdscale(as.dist(disrf_uni),optimq)
n2=pamk(dist(cmd_uni))$nc
pam_uni=pam(dist(cmd_uni),n2)$cluster
randuni[i,j]=round(adjustedRandIndex(trueclusters,pam_uni),4)
errboot[i,j]=round(err_boot,4)
erruni[i,j]=round(err_uni,4)
}}
```

مراجع

[۱] زالی، ح.، امینی، ر. و شیری هریس، ر. (۱۳۹۲)، آنالیز داده‌های مکرواری بیماری لوکمیا با برنامه DAVID. مجله علمی دانشگاه علوم پزشکی ایلام، ۲۱، ۹۲-۱۰۲.

[۲] زرنگارنیا، ی.، علوی مجد، ح.، رضایی طاویرانی، ن. و خادم معبودی، ع. (۱۳۸۹)، کاربرد خوشه‌بندی فازی در تحلیل پروتئین‌های مرتبط با سرطان مری، معده و کلون بر اساس تشابهات تفسیر هستی شناسی ژنی، مجله علمی دانشگاه علوم پزشکی سمنان، ۱۲، ۱۴-۲۲.

[۳] علوی مجد، ح.، واحدی، م.، محرابی، ی. و نقوی، ب. (۱۳۸۶)، بکارگیری روش‌های خوشه‌بندی در ریزآرایه DNA. مجله پژوهش در پزشکی، ۳۱، ۱۹-۲۵.

[۴] واحدی، م.، علوی مجد، ح.، محرابی، ی. و نقوی، ب. (۱۳۸۶)، خوشه‌بندی داده‌های بیان ژنی و کاربرد آن در تحلیل افتراق انواع سرطان خون، مجله علمی دانشگاه علوم پزشکی سمنان، ۹، ۱۶۳-۱۶۹.

[5] Babu, G. P and Murty, M. N. (1993), A near-optimal initial seed value selection in K-means algorithm using a genetic algorithm, *Pattern Recognition Letters*, 14, 763-769.

[6] Bellman, R. E. (1961), *Adaptive control processes - A guided tour* , Princeton, New Jersey, U.S.A. : Princeton University Press.

[7] Berkhin, P. (2002), Survey of clustering data mining techniques ,Tec. Rep., Accrue Software, San Jose, CA.

[8] Beyer, K. S., Goldstein, J., Ramakrishnan, R. and Shaft, U. (1999), When is nearest neighbor meaningful?, *Lecture Nottes in Computer Science*, 1540, 217-235.

[9] Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford: Oxford Universiy Press.

[10] Breiman, L. (2001), Random Forests, *Machine Learning*, 45, 5-32.

[11] Breiman, L. and Cutler, A. (2003), manual v4, Tec. Rep., UC Berkel.

[12] Chen, X. and Ishwaran, H. (2012), Random forests for genomic data analysis, *Genomics*, 99, 323-329.

[13] Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y. and Mazumder, A. (2006), Prognostic gene expression signatures can be measured in tissues in RNA later preservative, *J Mol Diagn*, 8, 31-39.

[14] Cox, T. F. and Cox, M. (2000), *Multidimensional Scaling Second Edition*, Chapman and Hall/CRC, 2nd ed.

[15] de Souto, M. C. P., Costa, I. G., de Araujo, D. S. A., Ludermir, T. B. and Schlieq, A. (2008), Clustering cancer gene expression data: acomparative study, *BMC Bioinformatics*, 9, 101-114.

[16] Englund, C. and Verikas, A. (2012), A novel approach to estimate proximity in a random forest: An exploratory study, *Expert Syst. Appl*, 39, 13046-13050.

[17] Everitt, B. (2011), *Cluster analysis*, London : Wiley, 5nd ed.

[18] Everitt, B. and Hothorn, T. (2011), *Introduction to applied multivariate analysis with r*, Berlin: Springer.

[19] Fern, X. Z. and Brodley, C. E. (2003), Random projection for high dimensional data clustering: Acluster ensemble approach, *ICML*, 3, 186-193.

[20] Fern, X. Z. and Brodley, C. E. (2006), Cluster ensemble for high dimensional clustering: an empirical study, Tec. Rep, Corvallis, OR: Oregon State University, Dept. of Computer Science.

[21] Gan, G., Ma, C. and Wu, J. (2007), *Data clustering: theory, algorithms and applications*, Philadelphia: SIAM, society for Industrial and Applied Mathematics.

[22] Han, J., Kamber, M. and Pei, J.(2012), *Data mining concepts and techniques, third edition*, Waltham, Mass: Morgan Kaufmann Publishers.

[23] Hatie, T., Tibshirani, R. and Friedman, J. (2009), *The elements of statistical learning: data mining inference and prediction*, California: Springer, 2nd ed.

[24] Hosseininezhad, F. and Salajegheh, A. (2012), Study and comparison of partitioning clustering algorithms, *Iranian Journal of Medical Informatics* , 2, 38-43.

[25] Hubert, L. and Arabie, P. (1985), Comparing partitions, *Journal of Classification*, 2, 193-218.

[26] Härdle, W. and Simar, L. (2003), *Applied multivariate statistical analysis*, Berlin [u.a.]: Springer.

[27] Ishioka, T. (2013), Imputation of missing values for semi-supervised data using the proximity in random forests *IJBIDM*, 8, 155-166.

[28] Jain, A. K. and Dubes, R. (1988), *Algorithms for Clustering Data*, New Jersey: Prentice Hall.

[29] Kaufman, L. and Rousseeuw, P. J. (1990), *Finding groups in data: An introduction to cluster analysis*, Canada: John Wiley.

[30] Kotsiantis, S. B. (2014), Bagging and boosting variants for handling classifications problems: a survey, *The Knowledge Engineering Review*, 29, 78-100.

[31] Kruskal, J. (1964), Nonmetric multidimensional scaling: A numerical method, *Psychometrika*, 29, 115-129.

[32] Liaw, A. and Wiener, M. (2002), Classification and Regression by randomForest, *R News*, 2, 18-22.

[33] Lloyd, S. P. (1982), Least squares quantization in PCM, *IEEE Transactions on Information Theory*, 28, 129-136.

[34] Milligan, G. and Cooper, M. (1986), A study of the comparability of external criteria for hierarchical cluster analysis *Multivariate Behavioral Research*, 21, 441-458.

[35] Perbet, F., Stenger, B. and Maki, A. (2009). Random Forest Clustering and Application to Video Segmentation, *Proceedings of the British Machine Vision Conference*, 101-110, London .

[36] Qi, Y., Klein-seetharaman, J. and Bar-joseph, Z. (2005), Random Forest Similarity for Protein-Protein Interaction Prediction, *Pac Symp Biocomput*, 10, 531-542.

[37] Radovanovic, M., Nanopoulos, A. and Ivanovic, M. (2010), Hubs in space: Popular nearest neighbors in high-dimensional data, *J. Mach. Learn. Res*, 11, 2487-2531.

[38] Rai, P. and Singh, S. (2010), A Survey of Clustering Techniques, *International Journal of Computer Applications*, 7, 645-678.

[39] Rencher, A. C. and Christensen, W. F. (2012), *Methods of multivariate analysis*, Hoboken, New Jersey: Wiley.

[40] Rousseeuw, P. (1987), Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, 20, 53-65.

[41] Shepard, R. N. (1962), The analysis of proximities: multidimensional scaling with an unknown distance function I., *Psychometrika*, 27, 125-140.

[42] Shi, T. and Horvath, S. (2006), Unsupervised Learning With Random Forest Predictors, *Journal of Computational and Graphical Statistics*, 15, 118-138.

[43] Shi, T., Seligson, D., Belldegrun, A. S., Palotie, A. and Horvath, S. (2004), Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma, *Modern Pathology*, 18, 547-557.

[44] Siedlecki, W. and Sklansky, J. (1989), A note on genetic algorithms for large-scale feature selection, *Pattern Recognition Letters*, 10, 335-347.

[45] Steinbach, M., Ertöz, L. and Kumar, V. (2003), The challenges of clustering high-dimensional data, *New Directions in Statistical Physics*, Springer-Verlag, 273-309.

[46] Timm, N. (2002), *Applied multivariate analysis*, Springer texts in statistics, New York: Springer.

[47] Torgerson, W. S. (1952), Multidimensional scaling: I. Theory and method, *Psychometrika*, 17, 401-419.

[48] Verikas, A., Gelzinis, A. and Bacauskiene, M. (2011), Mining data with random forests: A survey and results of new tests, *Pattern Recognition*, 44, 330-349.

[49] Wickelmaier, F. (2003), An introduction to MDS, Tec. Rep., Aalborg University, Denmark.

[50] Xu, R. and Wunsch, I. (2005), Survey of clustering algorithms, *Neural Networks, IEEE Transactions on*, 16, 645-678.

[51] Zhou, W. (2004), A review and implementation of some approaches to metric clustering, Tec. Rep., Aalborg University, Denmark.

Aabstract

Clustering, that is widely used in various fields of science, is one of the important tools in data mining. Many clustering methods have been built based on dissimilarity of observations which is calculated by a distance function. The progress of science and technology in the world today has resulted in large data sets with very large number of underlying variables. As the performance of distance functions are affected by increasing the dimension, identifying the clusters in the high dimensional space can be viewed as a challenge for researchers. In this thesis, a new kind of dissimilarity measure based on the classification method of Random Forests is investigated to provide a new basis for any clustering method. Thereafter, the Multidimensional scaling method is combined with Partition around medoid clustering algorithm to handle some simulation and real examples. The obtained results confirm the superiority of this approach in comparison with common methods.

Key Words: Clustering, Random forest, Variables importance, Partition around medoid, Multidimensional scaling

Shahrood University
Faculty Of Mathematical Sciences

Dissertation Submitted in Partial
Fulfillment of The Requirements For The
Degree of Master of Science in
Statistic

Clustering with the classification method of Random Forests

Supervisor

Dr. D. Shahsavani

by

Zohreh Farhadi

November 2014