دانشکده علوم ریاضی

پایان‌نامه کارشناسی ارشد آمارریاضی

کاهش بعد در مساله خوشه‌بندی با استفاده از تابع جریمه لاسو گروهی

نگارنده: زینت سلیمی ثانی

استاد راهنما

دکتر داود شاهسونی

بهمن ۱۳۹۶

شماره:

تاریخ:

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای .....زینت سلیمی ثانی با شماره دانشجویی ۹۴۰۹۷۰۴ رشته آمار گرایش آمارریاضی تحت عنوان کاهش بعد در مساله خوشه بندی با استفاده از تابع جریمه لاسو گروهی که در تاریخ ۱۳۹۶/۱۱/۱۰ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

قبول (با درجه: ..خوب..) ☒ مردود ☐

نوع تحقیق: نظری ☒ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	دکتر داود شاهسونی	دانشیار	
۲- استادراهنمای دوم			
۳- استاد مشاور			
۴- نماینده تحصیلات تکمیلی	دکتر احمد نزاکتی رضازاده	دانشیار	
۵- استاد ممتحن اول	دکتر حسین باغیشنی	استادیار	
۶ـ استاد ممتحن دوم	دکتر محمد آرشی	دانشیار	

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده:

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

"تقدیم به محضر مولاو آقا امام زمان عجل الله"

تقدیم به پدرم

کوهی استوار و حامی من در طول تمام زندگی

تقدیم به مادرم

سنگ صبوری که الفبای زندگی به من آموخت

سپاس‌گزاری...

شکر شایان نثار ایزد منان که توفیق را رفیق راهم ساخت تا این پایان نامه را به پایان برسانم. از استاد فاضل و اندیشمند جناب آقای دکتر شاهسونی به عنوان استاد راهنما که همواره اینجانب را مورد لطف و محبت خود قرار داه اند، کمال تشکر را دارم.

زینت سلیمی ثانی
بهمن ۱۳۹۶

تعهد نامه

اینجانب *زینت سلیمی ثانی* دانشجوی کارشناسی ارشد رشته **آمار** دانشکده **علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **کاهش بعد در مساله خوشه‌بندی با استفاده از تابع جریمه لاسو گروهی** ، تحت راهنمایی **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 صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘ دانشگاه صنعتی شاهرود ’’ یا “ Shahrood University of Technology ” 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.

زینت سلیمی ثانی
بهمن ۱۳۹۶

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 صنعتی 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 مسائل مهم داده‌کاوی در کشف الگوهای پنهان در داده‌ها است. هنگامی که تعداد متغیرها بسیار زیاد باشد و با داده‌های با بعد بالا مواجه باشیم، مسئله کاهش بعد نیز در کنار خوشه‌بندی، مطرح می‌شود. یکی از متداول‌ترین روش‌های کاهش بعد که در هر دو موضوع یادگیری با نظارت و بدون نظارت استفاده می‌شود، تحلیل مولفه‌های اصلی است که دارای محاسن و معایب خاص خود می‌باشد. در این پایان‌نامه علاقه‌مندیم تا با معرفی روش خوشه‌بندی ممیزی بهینه (ODC) و به منظور کاهش بعد، مساله خوشه‌بندی که یک مساله یادگیری بدون نظارت است را در قالب مساله رگرسیون ریج، بیان کنیم تا بتوان همانند اندیشه‌ی مولفه‌های اصلی، نوع دیگری از ترکیب خطی متغیرهای اولیه را برای ساختن متغیرهای جدید استخراج نموده و سپس یکی از الگوریتم‌های خوشه‌بندی نظیر k ـ میانگین را برای مشاهدات تبدیل یافته‌ی جدید بکار گیریم. در این مساله، یک پارامتر تنظیم نقش بسزایی در عملکرد خوشه‌بندی خواهد داشت. همچنین وجود برخی از متغیرهای غیر ضروری در مدل، موجب عملکرد منفی و ضعیف روش خوشه‌بندی می‌شود. با اضافه کردن تابع جریمه لاسو گروهی، این ضعف را برطرف می‌کنیم و خوشه‌بندی جدیدی را معرفی می‌کنیم، که نسخه اصلاح‌شده روش خوشه‌بندی ممیزی بهینه است. نتایج حاصل از شبیه‌سازی حاکی از کارآیی این روش در مواجهه با ابعاد بالای متغیرها و همچنین برتری آن نسبت به روش مولفه‌های اصلی، مورد ارزیابی قرار گرفته است.

کلمات کلیدی: خوشه‌بندی ممیزی بهینه، خوشه‌بندی ممیزی بهینه اصلاح‌شده، داده‌های با ابعاد بالا، انتخاب متغیر، اعتبارسنجی متقابل.

پیشگفتار

امروزه، با انفجار اطلاعات در جهان، تحلیل داده‌های بزرگ نیازمند توسعه تکنیک‌های آماری است که بتواند الگوهای پنهان، همبستگی ناشناخته و سایر مفاهیم و اطلاعات داده‌ها با ابعاد بالا را کشف کند. خوشه‌بندی یکی از بهترین روش‌هایی است که برای کار با داده‌ها ارائه شده است. قابلیت آن در ورود به فضای داده و تشخیص ساختار آن‌ها، خوشه‌بندی را یکی از ایده‌آل‌ترین مکانیزم‌ها برای کار با دنیای عظیم داده‌ها کرده است. اولین بار ایده‌ی آن در دهه ۱۹۳۵ ارائه شد و امروزه با پیشرفت‌ها و جهش‌های عظیمی که در آن پدید آمده، خوشه‌بندی در کاربردها و جنبه‌های مختلفی حضور یافته است. خوشه‌بندی یکی از شاخه‌های یادگیری بدون نظارت می‌باشد و فرآیند خودکاری است که در طی آن، نمونه‌ها به دسته‌هایی که اعضای آن مشابه یکدیگر می‌باشند، تقسیم می‌شوند که به این دسته‌ها خوشه گفته می‌شود. بنابراین خوشه، مجموعه‌ای از اشیاء می‌باشد که در آن اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه‌های دیگر غیرمشابه می‌باشند. برای مشابه بودن می‌توان معیارهای مختلفی را در نظر گرفت مثلا می‌توان معیار فاصله را برای خوشه‌بندی مورد استفاده قرار داد و اشیائی را که به یکدیگر نزدیکتر هستند را بعنوان یک خوشه در نظر گرفت که به این نوع خوشه‌بندی، خوشه بندی مبتنی بر فاصله نیز گفته می‌شود. در خوشه‌بندی داده‌های با ابعاد بالا بسیاری از متغیرها نوفه هستند که با توجه به همبستگی قوی یا وابستگی قوی بین متغیرها، حاوی اطلاعات زیادی هستند. پس یافتن ساختار در فضای متغیرهای بعد بالا با زیرمجوعه کوچک از متغیرهای اصلی، بسیار مهم است. تعیین متغیرهای مهم کار دشواری است و هدف یافتن زیرمجموعه کوچکی از متغیرهای اصلی است که شامل اطلاعاتی مهم از مجموعه داده‌های اصلی هستند. این متغیرها به ما در درک ساختار چند متغیره کمک زیادی می‌کنند.

در این پایان‌نامه پس از معرفی روش خوشه‌بندی ممیزی بهینه، عملکرد آن در مقایسه با روش تحلیل مولفه‌های اصلی مورد ارزیابی قرار گرفته است. روش خوشه‌بندی ممیزی بهینه یک روش مبتنی بر رگرسیون جریمه است که همزمان خوشه‌بندی و کاهش ابعاد را انجام می‌دهد. با این مقدمه، ساختار پایان‌نامه به صورت زیر تنظیم شده است:

- در فصل اول، تعاریف و مفاهیم اولیه در ارتباط با کاهش بعد در مساله خوشه‌بندی را مطرح می‌کنیم.
- در فصل دوم، به بیان جزئی‌تر روش‌های مورد استفاده در تحقیق یعنی روش خوشه‌بندی k ـ میانگین و روش‌های انتخاب متغیر نظیر رگرسیون ریج، رگرسیون لاسو و لاسو گروهی پرداخته و معیارهایی را به منظور ارزیابی عملکرد خوشه‌بندی معرفی می‌کنیم، و روش‌های انتخاب تعداد خوشه‌ها را بیان می‌کنیم.
- در فصل سوم، ابتدا به نحوه محاسبه ماتریس امتیاز در روش تحلیل ممیزی خطی را

معرفی می‌کنیم که مبنای اندیشه معرفی ماتریس امتیاز برای یادگیری بدون نظارت است و همچنین مبنای معرفی خوشه‌بندی در قالب مدل‌های رگرسیونی انقباضی است. در این فصل با معرفی دو مدل خوشه‌بندی جدید چگونگی تبدیل مساله خوشه‌بندی به مساله رگرسیون انقباضی توضیح داده خواهد شد. همچنین به نحوه‌ی انتخاب پارامتر تنظیم بهینه و انتخاب متغیر توسط جریمه لاسو گروهی می‌پردازیم.

- در فصل چهارم، طی یک مطالعه شبیه‌سازی، ضمن برآورد تعداد خوشه‌ها، نحوه‌ی انتخاب پارامتر تنظیم بهینه نشان داده شده است. همچنین عملکرد خوشه‌بندی ممیزی بهینه در مقایسه با تحلیل مولفه‌های اصلی را مورد ارزیابی قرار می‌دهیم. با ارزیابی عملکرد خوشه‌بندی ممیزی بهینه اصلاح‌شده و نحوه‌ی انتخاب متغیر در آن، فصل را به پایان می‌بریم.
- در پیوست آ، مقادیر منفرد و روش تجزیه ماتریس بر اساس مقادیر منفرد معرفی شده و پیوست ب، حاوی کدهای نوشته‌شده در محیط R برای بازتولید مثال‌های پایان‌نامه است.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مفاهیم

۱.۱ مروری بر داده‌کاوی

امروزه با پیشرفت فن‌آوری، به ویژه فن‌آوری‌های مرتبط با اطلاعات و ارتباطات، حجم عظیمی از داده‌ها در شبکه‌های ارتباطی و اطلاعاتی در حال تولید هستند و یکی از ضرورت‌های موفقیت‌ها در سطوح مختلف علمی، پزشکی و کسب و کارها حتی در مقیاس کوچک، امکان بهره‌گیری از این اطلاعات و داده‌هاست. اگر داده‌ها فقط جمع شوند و بلا استفاده بمانند، عملا هدف اصلی از جمع‌آوری این اطلاعات، مرتفع نشده است. گام مهمی که پس از جمع آوری داده‌ها و اطلاعات می بایست برداشته شود، استخراج دانش از اطلاعات و ملموس‌تر کردن داده‌های جمع آوری‌شده، برای استنتاج قواعد و نتایج کاربردی است.

مجموعه ابزارهایی که می‌توانند شرکت‌ها، موسسات و حتی افراد را کمک کنند، تا از انبوه اطلاعات و داده‌های در دسترس، به مفاهیم کاربردی و موثر برسند، در شاخه‌ای از علوم کامپیوتر و آمار به نام داده‌کاوی[۱] مورد بررسی قرار می‌گیرد. در واقع، داده‌کاوی مجموعه‌ای از مسائل کاربردی است که در حوزه استخراج دانش از انبوه داده‌های در دسترس تعریف شده‌اند، که روش‌هایی نیز در طول زمان، توسط دانشمندان علوم کامپیوتر، ریاضیات و آمار برای آن‌ها ارائه شده است.

در داده‌کاوی نیز، انبوهی از اطلاعات و داده‌ها بررسی می‌شوند. همه این کاربردها، که هر

[۱] Data Mining

روزه ما با آن‌ها رو در رو هستیم، از طریق تحلیل داده‌های جمع‌آوری شده در گذشته انجام می‌شوند، و نمونه‌های موفقی از کاربرد داده‌کاوی در زندگی روزمره ما هستند. قطعا این موارد صرفا مثال‌هایی بودند که می‌توان به راحتی موارد دیگری را نیز به آن‌ها، افزود.

کاربردهایی که برای داده کاوی وجود دارند، بسیار بسیار گسترده اند و ما در این جا فقط امکان معرفی تعداد محدودی از آن‌ها را داریم. به عنوان مثال‌های بیشتر، می‌توان به کاربردهای داده‌کاوی در زمینه‌های زیر اشاره کرد:

- سیستم‌های مدیریتی: مدیریت ارتباط با مشتریان.
- نرم افزارهای امنیتی: نرم افزار رصد شبکه و ویروس کش‌ها.
- سیستم های بانکی: تخصیص اعتبار به مشتریان و طبقه بندی آن ها.
- مالی و اقتصادی: پیش‌بینی قیمت یک یا چند سهام یا شاخص برنامه‌ریزی و مکان یابی: چینش داخلی فروشگاه‌های بزرگ یا تخصیص امکانات شهری.
- علوم پزشکی: پیش‌بینی خطرات احتمالی ناشی از یک عمل جراحی خاص.
- علوم اجتماعی و سیاسی: پیش‌بینی یا تحلیل نتایج انتخابات.

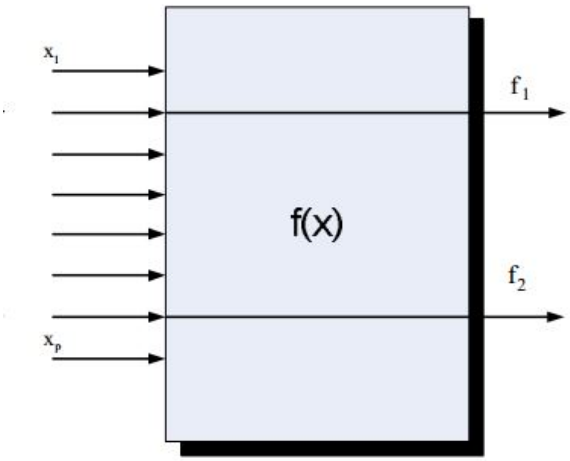
۲.۱ کاهش بعد

در این بخش به موضوع کم کردن بعد داده‌ها خواهیم پرداخت. دلیل کاهش بعد[۲] را می‌توان راحت‌تر شدن تحلیل بعد،افزایش عملکرد جداکننده [۳] بر اساس نمایش بهتر یا پایدارتر، حذف اطلاعات تکراری یا غیر مربوط یا تلاشی برای کشف ساختار اساسی با به‌دست آوردن نمایش گرافیکی از داده‌ها دانست. در ادبیات تحلیل‌های چندمتغیره[۴] [۱]. به تکنیک‌هایی که برای کاهش بعد استفاده می‌شود، روش‌های محوری[۵] یا روش‌های هندسی[۶] گفته می‌شود. روش‌هایی مانند تحلیل مولفه‌های اصلی[۷] (PCA) از این دسته‌اند. در ادبیات مربوط به تشخیص الگو و ساختار، روش‌های کاهش بعد بر اساس انتخاب متغیر[۸] و استخراج متغیر[۹] مانند روش تحلیل ممیزی خطی [۱۰] است.

---

[۲] Dimension reduction
[۳] Separator
[۴] Multivariate analysis
[۵] Ordination
[۶] Geometrical
[۷] Principal components analysis
[۸] Feature selection
[۹] Feature extraction
[۱۰] Linear discriminant analysis

برای مجموعه‌ای از داده‌ها، کاهش بعد را می‌توان از دو جنبه مختلف بررسی کرد. در دیدگاه اول، به شناسایی متغیرهایی می‌پردازیم که در کار جداسازی کمکی نمی‌کنند یا به عبارتی متغیرهای بی‌اطلاع[۱۱] هستند. در یک مساله جداسازی می‌توان از متغیرهایی که کمکی به تفکیک‌پذیری داده‌ها نمی‌کنند، صرف‌نظر کرد. پس برای کاهش بعد به دنبال پیدا کردن q متغیر مفید[۱۲] از بین p متغیر موجود هستیم، که q باید مشخص باشد. این کار انتخاب متغیر نامیده می‌شود. از این روش‌ها برای اهداف متفاوت جداسازی مانند رگرسیون لاسو [۱۳] و لاسو گروهی[۱۴] برای انتخاب زیرمجموعه‌ای از متغیرها نیز استفاده می‌شود. شکل ۱.۱ بیانگر مکانیسم ساده‌ای از انتخاب متغیر است. در این شکل x_iها مقادیر ورودی، f تابع موردنظر هستند.

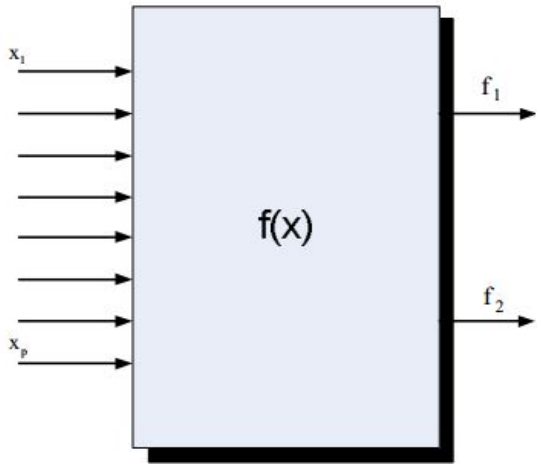


شکل ۱.۱: مکانیسم انتخاب متغیر

دیدگاه دوم پیدا کردن تبدیلی است که داده‌ها ر ااز فضای p بعدی به یک فضا با بعد کوچکتر انتقال دهد. این کار انتخاب متغیر در فضای تبدیل‌یافته یا استخراج متغیر نامیده می‌شود که مکانسیم عملکرد آن در شکل ۲.۱ نشان داده شده است. در هر دو دیدگاه، نیازمند بهینه‌سازی یک تابع معیار هستیم.

در انتخاب متغیر، بهینه‌سازی بر روی مجموعه‌ای از تمام زیرمجموعه‌های q بعدی از p متغیر موجود، انجام می‌شود. در استخراج متغیر، بهینه‌سازی بر روی تمام تبدیل‌های ممکن متغیرها، انجام می‌پذیرد[۱].

[۱۱] Non-informative features
[۱۲] Informative features
[۱۳] Lasso regression
[۱۴] Group lasso



شکل ۲.۱: مکانسیم استخراج متغیر در فضای تبدیل‌یافته

۱.۲.۱ اهداف کاهش بعد داده‌های بزرگ

یکی از موضوعات مهم در آمار و مسائل مرتبط با آن یافتن بیان مناسبی از داده‌های چند متغیره با بعد بالا است. با افزایش تعداد متغیرها اهداف زیر در مورد کاهش بعد داده‌های بزرگ مطرح می‌شود:

- سرعت الگوریتم‌ها با داده‌های با بعد کمتر بیشتر می شود.
- فضای ذخیره‌سازی کمتری نیاز است.
- برای ترسیم و به‌دست آوردن درکی از مجموعه داده‌ها گاهی بعد داده‌ها را به بعد دو یا سه تقلیل می‌دهند تا بتوانند نموداری از داده‌های با ابعاد زیاد ترسیم کنند.

به هر دلیلی که بخواهیم کاهش بعد را انجام دهیم، یکی از مرسوم‌ترین روش‌های انجام این کار روش تحلیل مولفه‌های اصلی است. در فصل دوم سعی شده تا حد ممکن روش PCA به‌صورت کامل معرفی و نحوه ی استفاده از آن در مسایل نظری و علمی بیان شود.

۳.۱ مروری بر خوشه‌بندی

ما در جهانی پر از داده زندگی می‌کنیم. هر روزه انسان‌ها با حجم وسیعی از اطلاعات رو به رو هستند که باید آن‌ها را ذخیره‌ساز ی یا نمایش دهند. یکی از روش‌های حیاتی کنترل و مدیریت این داده‌ها خوشه‌بندی[۱۵] داده‌های با خواص مشابه، درون مجموعه‌ای از دسته‌ها یا خوشه‌ها می‌باشد.

[۱۵] Clustering

مروزه، خوشه‌بندی نقش حیاتی در روش‌های بازیابی اطلاعات برای سازمان‌بندی مجموعه‌های بزرگ داده‌ها درون تعداد کمی خوشه معنادار دارد. معمولا در خوشه‌بندی، با ابعاد بسیار بالای فضای داده مواجه هستیم که انجام خوشه‌بندی به این شکل، مشکل به نظر می‌رسد. اساسا، سیستم‌های خوشه‌بندی همراه با نظارت[۱۶] یا بدون نظارت[۱۷] هستند. برخلاف رده‌بندی[۱۸] در خوشه‌بندی ، گروه‌ها از قبل مشخص نمی‌باشند و همچنین معلوم نیست که بر حسب کدام خصوصیات گروه‌بندی صورت می‌گیرد. در نتیجه پس از انجام خوشه‌بندی باید یک فرد خبره خوشه‌های ایجادشده را تفسیر کند و در بعضی مواقع لازم است که پس از بررسی خوشه‌ها بعضی از پارامترهایی که در خوشه‌بندی در نظر گرفته شده‌اند، ولی بی‌ربط بوده یا اهمیت چندانی ندارند، حذف شده و جریان خوشه‌بندی از اول صورت گیرد. هدف نهایی خوشه‌بندی این است که داده‌های موجود را به چند گروه تقسیم کنند و در این تقسیم‌بندی داده‌های گروه‌های مختلف باید حداکثر تفاوت ممکن را به هم داشته باشند و داده‌های موجود در یک گروه باید بسیار به هم شبیه باشند. البته کیفیت نتایج خوشه‌بندی، به روش اندازه‌گیری شباهت و توانایی و قدرت الگوریتم در کشف الگوهای مخفی میان داده‌ها بستگی دارد.
همچنین می‌توان بیان کرد که تحلیل خوشه‌ای[۱۹]، ابزاری برای اکتشاف ساختار داده‌هاست بدون فرضیاتی که در روش‌های آماری در نظر می‌گیریم. الگوریتم‌های خوشه‌بندی ابزارهای هستند که در جریان پیدا کردن ساختار داده‌ها مورد استفاده قرار می‌گیرند. تحلیل خوشه‌ای روش پیدا کردن خوشه‌ها در داده‌ها نیز نامیده می‌شود.

۴.۱ کاربردهای خوشه‌بندی

از آنجا که خوشه‌بندی یک روش یادگیری بدون نظارت محسوب می‌گردد، در موارد بسیاری می‌تواند کاربرد داشته باشد:

- بازاریابی[۲۰]: دسته‌بندی مشتری‌ها به دسته‌هایی بر حسب رفتارها و نیازهای آن‌ها از طریق مجموعه زیادی از ویژگی‌ها و آخرین خریدهای آن‌ها.
- زیست‌شناسی[۲۱]: دسته‌بندی حیوانات و گیاهان از روی ویژگی‌های آن‌ها.
- کتابداری: دسته‌بندی کتاب‌ها.

[۱۶] Supervised
[۱۷] Unsupervised
[۱۸] Classification
[۱۹] Cluster analysis
[۲۰] Marketing
[۲۱] Biology

- نقشه‌برداری شهری[۲۲]: دسته‌بندی خانه‌ها بر اساس نوع و موقعیت جغرافیایی آن‌ها.
- مطالعات زلزله‌نگاری[۲۳]: تشخیص مناطق حادثه‌خیز بر اساس مشاهدات قبلی.
- وب [۲۴]: دسته‌بندی اسناد یا دسته‌بندی مشتریان به سایت‌ها.
- داده‌کاوی: کشف اطلاعات و ساختار جدید از داده‌های موجود.
- تشخیص گفتار[۲۵]: ساخت کتاب کد از بردارهای ویژگی در تقسیم کردن گفتار بر حسب گویندگان آن یا فشرده‌سازي گفتار.
- تقسیم‌بندی تصاویر[۲۶]: تقسیم‌بندی تصاویر پزشکی یا ماهواره‌ای.

۵.۱ چالش‌های الگوریتم‌های خوشه‌بندی

به طور کلی الگوریتم‌های خوشه‌بندی داده‌ها با چالش‌های زیر روبرو هستند:

۱. مقیاس‌پذیری : چگونه می‌توان الگوریتم‌های خوشه‌بندی را تنظیم نمود تا برای پایگاه داده‌های با حجم بالا کارایی مناسب داشته باشند.

۲. توانایی مواجهه با انواع مختلف صفات و داده‌ها : الگوریتم‌های خوشه‌بندی باید برای داده‌های کمی و کیفی و هم برای داده‌های نوعی (اسمی) قابل اجرا باشند.

۳. حداقل نیازمندی به دانش اولیه که با پارامترهای ورودی مشخص می‌شود : بسیاری از انواع الگوریتم‌های خوشه‌بندی نیاز دارند تا کاربر پارامترهای ورودی خاصی را (مثل تعداد خوشه‌های مورد نظر) به عنوان ورودی تحلیل خوشه‌ها مشخص کند. مشخص نمودن بسیاری از این پارامترها مسئله دشواری خواهد بود.

۴. کشف خوشه‌ها با اشکال مختلف: اغلب الگوریتم‌های خوشه‌بندی بر پایه فاصله اقلیدسی کار می‌کنند . پس خوشه‌های کروی شکل با اندازه یا چگالی مشابه را پیدا می‌کنند . پس مهم است که الگوریتم خوشه‌بندی بتواند خوشه‌هایی متناسب با توزیع داده‌ها بیابد.

۵. توانایی مقابله با داده‌های نوفه‌دار : بیشتر پایگاه‌های داده شامل داده‌های پرت، جاافتاده و نادرست می‌باشند . چنانچه الگوریتم به این نوع داده‌ها حساس باشد ، خوشه‌های با کیفیت پایین تولید خواهند نمود.

---

[۲۲] City-planning
[۲۳] Earthquake studies
[۲۴] WWW
[۲۵] Speech recognition
[۲۶] Image segmentation

۶. عدم حساسیت به ترتیب داده‌های ورودی: نباید الگوریتم خوشه‌بندی با ترتیب متفاوت ورود داده‌ها، خروجی‌های مختلفی ایجاد نماید.

۷. ابعاد بالای داده های ورودی: یک پایگاه داده یا انبار داده ممکن است شامل صدها صفت یا بعد باشد. مطلوب است که الگوریتم مستقل از تعداد ابعاد، کارایی مناسبی داشته باشد.

۸. خوشه‌بندی همراه با اعمال محدودیت‌های کاربر: گاهی نیاز داریم تا برخی از محدودیت‌ها مثل تعداد خوشه‌ها، را برای الگوریتم تعریف نماییم.

۹. قابلیت تفسیر و استفاده: نتایج خوشه‌بندی باید برای کاربر قابل تفسیر، جامع و مفید باشد.

۱.۵.۱ کاهش بعد در مساله خوشه‌بندی

در تحلیل خوشه‌ای برای تجسم داده‌ها، نمودار پراکنش را رسم می‌کنند، اما این کار برای داده‌های با بیش از سه متغیر کار سختی است. پس برای تجسم اینگونه مجموعه داده‌ها می‌توان از ابزار کاهش بعد مانند مولفه‌های اصلی بهره برد که نمودارهای پراکنش بر اساس چند مولفه اصلی اول رسم می‌شوند. اگرچه در محبوبیت روش PCA برای کاهش بعد در بسیاری از زمینه‌ها نمی‌توان شکی کرد اما باید این موضوع را در نظر گرفت که دیدن ساختار خوشه‌ها موضوع مهمی است که روش PCA در این زمینه ضعف عمده‌ای دارد. به دنبال این موضوع می‌توان روش‌هایی ارائه کرد که به موازات کاهش بعد، خوشه‌بندی را انجام دهند و همچنین بتوانند ساختار خوشه‌ای را به طور واضح نشان دهند. لذا می‌توان با تبدیل مسائل خوشه‌بندی به مسائل رگرسیون انقباضی، مدل‌هایی را ارائه کرد که این کار را به خوبی انجام می‌دهند.

۶.۱ پیشینه تحقیق

تحلیل ممیزی خطی[۲۷]، (LDA) روشی کلاسیک و کاربردی است که به موازات طبقه‌بندی، کاهش بعد را نیز انجام می‌دهد. در این راستا یک فضای ممیزی کاهش بعد یافته را با استفاده از ترکیبات خطی و ماکزیمم‌سازی نسبت پراکندگی [۲۸] بین گروه‌ها به پراکندگی داخل گروه‌ها، را تخمین می‌زند. با توجه به این موضوع که در مسائل طبقه‌بندی دودویی[۲۹]، LDA معادل با روش کمترین میانگین توان‌های دوم خطا[۳۰] است [۵]، یافتن ارتباط مشابه‌ای بین LDA

[۲۷] Linear discriminant analysis
[۲۸] Scatter
[۲۹] Binary classification
[۳۰] least mean squared error

و مسئله طبقه‌بندی چند_کلاسه [۳۱] جالب به‌نظر می‌رسد. از این رو یافتن ارتباط مشابه در طبقه‌بندی چندکلاسه و مسائل رگرسیونی، موضوع مهمی است که راه‌های زیادی در این زمینه مطرح شده‌اند که رویکرد دیگری برای انجام LDA در قالب رگرسیونی فراهم می‌کند تا بتوان با تحمیل برخی معیار های جریمه، به برخی خواص آماری دست پیدا کند. از جمله این روش‌ها تحلیل ممیزی اصلاح‌شده، [۳۲] [۴] و روش تحلیل ممیزی جریمه‌شده [۳۳] هستند که توسط تیبشیرانی و هستی معرفی شده‌اند. [۶] اخیرا زو و همکارن مدل اصلاح‌شده PCA را با عنوان تحلیل مولفه‌های اصلی اصلاح شده [۳۴] [۲۴] معرفی کردند که تحلیل مولفه‌های اصلی را با اعمال جریمه لاسو یا شبکه منعطف[۳۵] روی یک بردار رگرسیونی را به‌عنوان یک مسئله رگرسیونی در نظر می‌گیرد. در این پایان‌نامه با استفاده از ماتریس امتیاز و جریمه ریج یک قالب بدون نظارت را برای انجام همزمان کاهش بعد و خوشه‌بندی، بازگو می‌کنیم که ایده اصلی آن از انجام روش LDA در فرم رگرسیونی، مطرح شده است [۸]. این ایده زمینه‌ساز معرفی الگوریتم‌های کارآمد بدون نظارت، یا اعمال جرایم ریج، لاسو یا شبکه منعطف می‌شود.

[۳۱] Multi-class classification
[۳۲] Sparse discriminant analysis
[۳۳] Penalized discriminant analysis
[۳۴] Sparse principal component analysis
[۳۵] Elastic net

فصل ۲

روش‌های مورد استفاده

در فصل اول، مفاهیم اولیه در ارتباط با خوشه‌بندی را مرور کردیم. اکنون می‌توان به بیان جزئی‌تر روش‌هایی پرداخت که در این پایان‌نامه از آن‌ها استفاده شده است. ابتدا روش خوشه‌بندی k ـ میانگین[۱] را معرفی نموده و مزایا و معایب این روش را بیان می‌کنیم. سپس به معرفی روش خوشه‌بندی سلسله‌مراتبی می‌پردازیم. همچنین در ادامه ابزارهای متداول در کاهش بعد و انتخاب متغیر از جمله تحلیل مولفه‌های اصلی، رگرسیون ریج، لاسو و لاسو گروهی را بیان می‌کنیم. سپس شاخص‌هایی به منظور ارزیابی خوشه‌بندی معرفی می‌کنیم و این فصل را با معرفی روش‌های تعیین تعداد خوشه‌ها خاتمه می‌دهیم.

۱.۲ روش خوشه‌بندی k ـ میانگین

روش k ـ میانگین یکی از محبوب‌ترین روش‌های خوشه‌بندی داده‌ها در داده‌کاوی است. این روش علی‌رغم سادگی، یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر محسوب می‌شود. الگوریتم‌های مختلفی برای این روش بیان شده‌اند ولی همه آن‌ها دارای مکانیسمی مشابه و تکراری هستند. از جمله مشکلات این روش آن است که وابسته به انتخاب اولیه مراکز بوده و بنابراین بهینه نیست. مشکلات دیگر آن تعیین تعداد خوشه‌ها و پوچ شدن خوشه‌ها است.

[۱] k-means

در الگوریتم k ـ میانگین که نخستین بار توسط لوید[۲] [۱۲] معرفی شد، میانگین مقادیر اعضای هر خوشه به عنوان مرکز خوشه در نظر گرفته می‌شود. برای مجموعه داده X با n مشاهده‌ی $\{x_۱, ..., x_n\}$، تعداد k خوشه‌ی $c_۱, c_۲, ..., c_k$ را در نظر می‌گیریم. مراحل انجام الگوریتم k ـ میانگین به شرح زیر است:

۱. تعداد k نقطه به صورت تصادفی از مجموعه‌ی X، به عنوان مراکز اولیه خوشه‌ها انتخاب می‌شوند. این مجموعه نقاط را به عنوان مراکز اولیه و با مجموعه $M = \{x_۱^*, x_۲^*, \cdots, x_k^*\}$ نشان می‌دهیم.

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

$$l = \arg\min_{۱\leq i\leq k} d(x_i^*, x_j) \qquad x_j \in X - M \tag{۱.۲}$$

آن‌گاه x_j به خوشه lام یعنی C_l تعلق می‌گیرد که d تابع فاصله است.

۳. پس از تخصیص تمامی مشاهدات به خوشه‌ها، مراکز جدید هر خوشه با میانگین گرفتن از مقادیر اعضای هر خوشه، دوباره محاسبه شده و مجموعه M به‌صورت زیر به‌روزرسانی می‌شود.

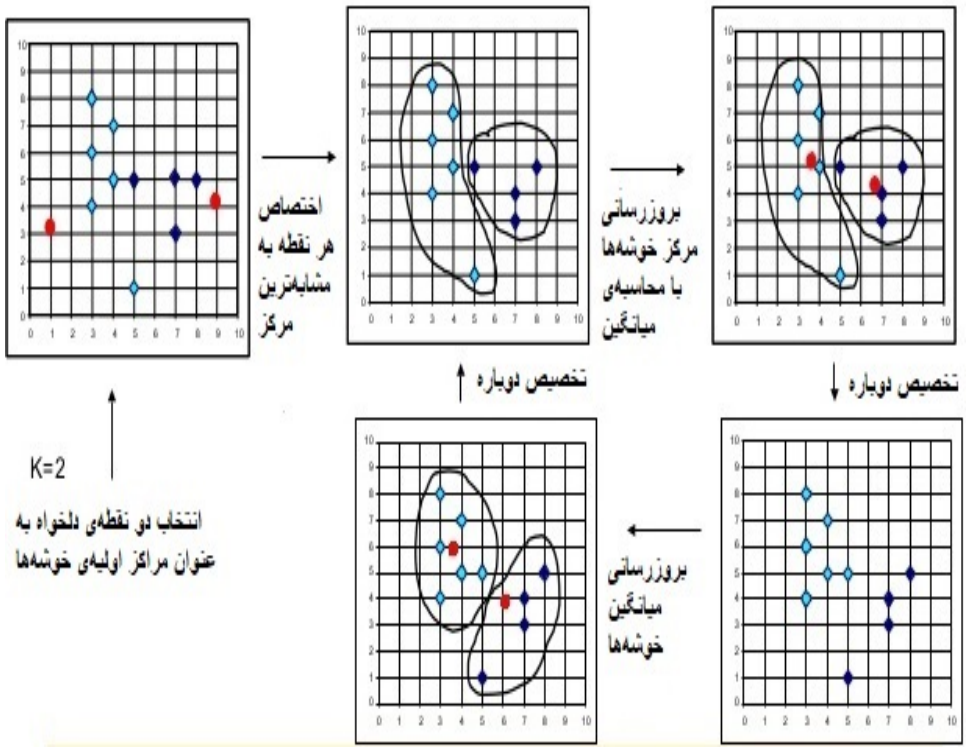
$$x_i^* = \frac{۱}{n_i} \sum_{x_j \in C_i} x_j \qquad ۱ \leq i \leq k \tag{۲.۲}$$

که در آن n_i تعداد اعضای خوشه C_i است.

۴. روند به روز رسانی مجموعه M تا زمانی ادامه می‌یابد که مراکز خوشه‌ها یا به عبارت دیگر اعضای خوشه‌ها تغییر نکنند.

در شکل ۱.۲ مراحل اجرای الگوریتم k ـ میانگین در قالب یک مثال نمایش داده شده است.

[۲] Lloyd



شکل ۱.۲: مراحل اجرای الگوریتم k ـ میانگین

معمولا مرکز خوشه‌های اولیه به صورت تصادفی از میان مشاهدات انتخاب می‌شوند. بنابراین خوشه‌های به‌دست آمده منحصر به فرد نیستند چرا که مرکز خوشه‌های اولیه در دو خوشه‌بندی مستقل k ـ میانگین می‌توانند متفاوت باشند. در الگوریتم k ـ میانگین می‌توان از معیارهای فاصله‌ی گوناگون بهره گرفت و مزایا و معایب به‌کارگیری آن معیار به نوع داده‌هایی که قرار است خوشه بندی شوند، بستگی دارد [۲].

۱.۱.۲ مزایا و معایب

الف ـ مزایا

- الگوریتم k ـ میانگین در مجموعه داده‌های بزرگ، کارامد و موثر عمل می‌کند.
- در صورت زیاد بودن تعداد متغیر ها، این روش نسبت به روش سلسله‌مراتبی دارای سرعت بالاتری می‌باشد.
- الگوریتم k ـ میانگین نسبت به روش سلسله‌مراتبی، خوشه‌های متراکم‌تری تولید می‌کند به خصوص هنگامی که خوشه‌ها به صورت کروی باشند.

ب_ **معایب**

- از جمله مشکلات این روش این است که بهینه بودن آن به انتخاب اولیه مراکز وابسته است. یعنی با انتخاب مراکز اولیه متفاوت ممکن است عملکرد خوشه‌بندی تغییر کند. اگر مقدار خطا را برای هر مشاهده، معادل عدم تشابه آن مشاهده با مرکز خوشه‌ای که در آن قرارگرفته، تعریف کنیم، مجموع توان‌های دوم خطا، معیار سنجش کیفیت خوشه‌بندی k_میانگین است. در واقع در این الگوریتم برای حل مشکل بهینه‌سازی باید عبارت زیر را به حداقل رساند:

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in C_i} d^2(x_i^*, x_j) \qquad (۳.۲)$$

که در آن $d(x_i^*, x_j)$ فاصله بین x_i^* و x_j است. در واقع در دو بار اجرای متفاوت این الگوریتم، اجرایی ارجح‌تر است که در آن عبارت SSE حداقل باشد.

- مشکلات دیگر آن تعیین تعداد خوشه‌ها و پوچ شدن خوشه‌ها می‌باشد. یعنی اگر در تکراری از الگوریتم، تعداد داده‌های متعلق به خوشه‌ای صفر شد راهی برای تغییر و بهبود ادامه روش وجود ندارد.
- از الگوریتم k_میانگین تنها زمانی که مجموعه داده‌ها مقادیر قابل تعریف داشته باشند و بتوان میانگین آن‌ها را محاسبه کرد، استفاده می‌شود.
- در این روش فرض شده است که تعداد خوشه‌ها از ابتدا مشخص است. اما معمولا در کاربردهای زیادی تعداد خوشه‌ها معلوم نیست.
- الگوریتم k_میانگین به نقاط دورافتاده حساس است و به شدت نتایج خوشه‌بندی را تحت تاثیر قرار می‌دهد.
- این الگوریتم عملکرد ضعیفی در خوشه‌بندی داده‌های غیر محدب یا U_شکل دارد.

۲.۲ روش خوشه‌بندی سلسله‌مراتبی

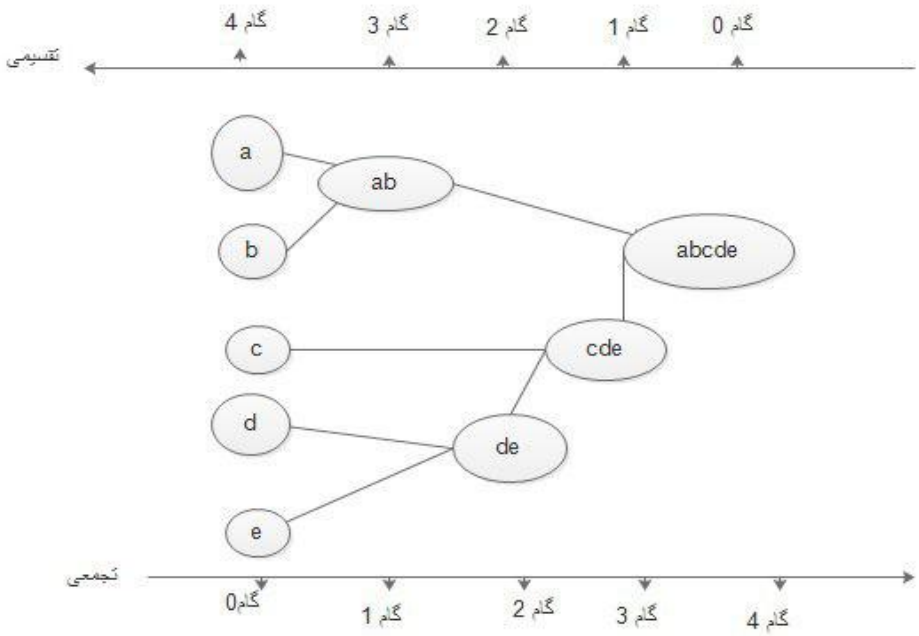
در این گونه روش‌ها، خوشه‌بندی داده‌ها طی یک فرآیند سلسله مراتبی صورت می‌گیرد. با توجه به نحوه‌ی اجرای این فرآیند، روش‌های سلسله‌مراتبی به دو دسته تجمعی[۳] یا تقسیمی[۴] تقسیم می‌شوند. در روش‌های سلسله‌مراتبی تجمعی ابتدا هر مشاهده به عنوان خوشه‌ای مجزا در نظر گرفته می‌شود. سپس در هر گام خوشه‌هایی که تشابه بیشتری با هم دارند در هم ادغام شده و خوشه‌ی بزرگتری را ایجاد می‌کنند تا در نهایت تمام مشاهدات درون یک

[۳] Agglomerative

[۴] Divisive

خوشه قرار گیرند. معکوس این روند در روش‌های سلسله‌مراتبی تقسیمی رخ می‌دهد به طوری که ابتدا تمامی مشاهدات به عنوان یک خوشه در نظر گرفته می‌شوند. در گام بعدی این خوشه به دو خوشه کوچکتر تقسیم می‌شوند و این روند تا زمانی ادامه می‌یابد که در هر خوشه، تنها یک مشاهده قرار گیرد.

شکل ۲.۲ گام‌های الگوریتم سلسله‌مراتبی تجمعی و تقسیمی را نشان می‌دهد که بر روی مجموعه $\{a, b, c, d, e\}$ انجام شده است. نتایج خوشه‌بندی سلسله‌مراتبی را می‌توان در نموداری با ساختار درختی که دندروگرام [۵] نامیده می‌شود، نمایش داد. در روش‌های سلسله‌مراتبی، 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 سلسله‌مراتبی 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].



شکل ۲.۲: گام‌های الگوریتم سلسله‌مراتبی تجمعی و تقسیمی

[۵] Dandrugram

۳.۲ ابزارهای معروف کاهش بعد در یادگیری بدون راهنما و با راهنما

۱.۳.۲ تحلیل مولفه‌های اصلی

در مسئله یادگیری بدون راهنما[۶] تحلیل مولفه‌های اصلی یکی از روش‌های کلاسیک چندمتغیره و شاید قدیمی‌ترین و معروف‌ترین آن‌ها باشد. این روش ابتدا توسط پیرسون [۱۳] به عنوان وسیله‌ای برای برازاندن صفحات از طریق کمترین توان‌های دوم متعامد معرفی شد و مستقلاً به وسیله هتلینگ [۹] به منظور تحلیل ساختارهای ماتریس‌های واریانس – کواریانس و ضریب همبستگی توسعه داده شد. مانند بسیاری از روش‌های چندمتغیره تا قبل از اختراع رایانه‌ها، به دلیل پیچیدگی در محاسبات، به طور گسترده‌ای مورد استفاده واقع نشد. بعد از آن از دیدگاه نظری و کاربردی به طور وسیعی توسعه پیدا کرده و بکار برده شده است. این نوع تجزیه را می‌توان از دیدگاه‌های مختلف مورد توجه قرار داد:

- تبدیل متغیرهای وابسته به متغیرهای ناهمبسته.
- یافتن ترکیبات خطی با تغییرپذیری نسبی بزرگ یا کوچک.
- کاهش حجم داده‌ها.
- تفسیر داده‌ها.

این نوع تجزیه معمولاً یک تجزیه نهایی تلقی نمی‌شود بلکه به عنوان وسیله‌ای میانی برای مطالعات و بررسی‌های بیشتر مورد استفاده قرار می‌گیرد. جنبه‌های ریاضی مورد استفاده در این روش شامل مقادیر ویژه و بردارهای ویژه ماتریس‌های همیشه مثبت متقارن است.
کاهش حجم داده‌ها، هدف اصلی این تجزیه را تشکیل می‌دهد که این داده‌ها شامل تعداد زیادی متغیرهای با همبستگی‌های درونی می‌باشند به طریقی که حداکثر ممکن اطلاعات موجود در داده‌ها محفوظ بماند. این امر از طریق تبدیل داده ها (متغیرها) به متغیرهای جدیدی است که مولفه‌های اصلی نامیده شده و ناهمبسته هستند و به ترتیبی اولویت‌بندی می‌شوند که تعداد اندکی از آن‌ها اغلب تغییرات موجود در متغیرهای اولیه را با خود به همراه دارند.

[۶] Unsupervised learning

تعریف مولفه‌های اصلی

فرض کنید بردار $\mathbf{X}$ شامل p متغیر بوده و بررسی و مطالعه واریانس متغیرها و کواریانس بین دو به دوی این متغیرها مورد توجه است. در شرایطی که p کوچک باشد یا ماتریس واریانس – کواریانس ساده باشد این مطالعه به سادگی و با ملاحظه p واریانس و $\frac{1}{2}p(p-1)$ کواریانس قابل انجام است. در غیر این صورت یک روش مطلوب آن است که تعداد اندکی متغیر جدید (بسیار کمتر از p) پیدا کنیم به طوری که شامل اکثر اطلاعات مربوط به واریانس‌ها و کواریانس‌ها باشد. در تجزیه مولفه‌های اصلی گرچه ظاهراً توجه اصلی روی واریانس متغیرها است اما با توجه به روابط بین واریانس ها و کواریانس‌ها این روش به طور ضمنی کواریانس‌ها یا ضرایب همبستگی را نیز مورد توجه قرار می‌دهد.

از لحاظ جبری، مولفه‌های اصلی، ترکیبات خطی به خصوصی از عوامل بردار $\mathbf{X}$ می‌باشند. از لحاظ هندسی ترکیبات خطی به منزله انتخاب یک سیستم محورهای مختصات جدید است که از دوران محورهای اولیه یعنی $X = x_1, x_2, ..., x_p$ به‌دست می‌آید به طوری که محورهای جدید جهات با بیشترین تغییرات را نشان می‌دهند و توصیفی ساده‌تر از ساختار ماتریس واریانس – کواریانس ارائه می‌دهد. شکل ۳.۲ در واقع تحلیل مولفه‌های اصلی PCA را نمایش می‌دهد. در این شکل، فرض بر این است که مولفه‌ها تحت تاثیر متغیرهای مشاهده شده‌اند. در اینجا متغیرهای مشاهده‌شده، همان شاخص‌ها هستند که بر مولفه‌ها تاثیرگذار هستند. این نوع تحلیل، زمانی استفاده خواهد شد که قصد داریم بعد داده‌ها را کاهش دهیم.

فرض کنید متغیرهای $X_1, ..., X_p$ متغیرهای جدید $c_1, ..., c_p$ را تولید کنند به طوری که

$$\begin{aligned} c_1 &= l_1'X = l_{11}X_1 + l_{21}X_2 + ... + l_{p1}X_p \\ c_2 &= l_2'X = l_{12}X_1 + l_{22}X_2 + ... + l_{p2}X_p \\ &\vdots \\ c_p &= l_p'X = l_{1p}X_1 + l_{2p}X_2 + ... + l_{pp}X_p \end{aligned}$$

و

$$\text{var}(c_i) = l_i'\Sigma l_i, \quad \text{cov}(c_i, c_k) = l_i'\Sigma l_k, \quad i = 1, 2, ..., p$$

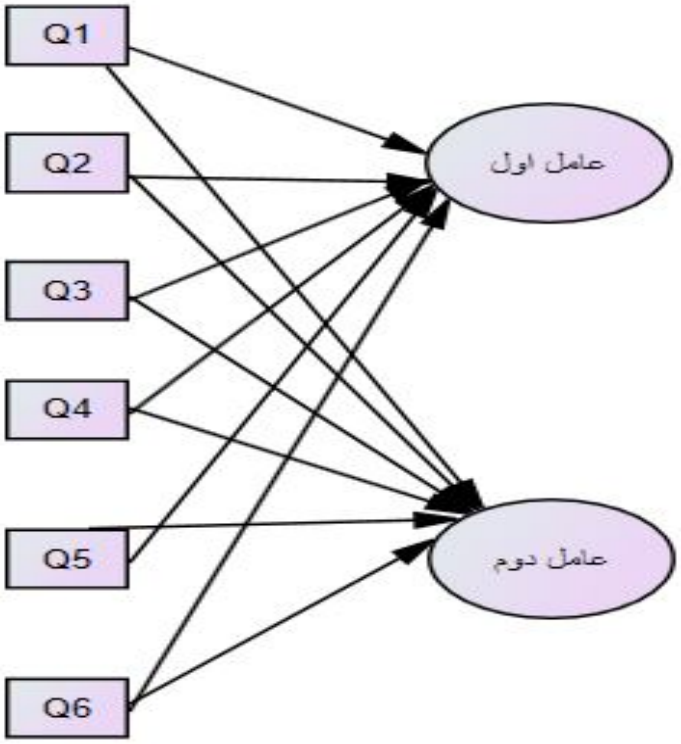
مولفه‌های اصلی ترکیبات خطی از $c_1, ..., c_p$ هستند که دارای بیشترین واریانس هستند. فرض می‌کنیم Σ ماتریس کواریانس برای متغیرهای $X = [X_1, ..., X_p]$ با بردارهای ویژه $\lambda_1, ..., \lambda_p$ است که

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$$

و همچنین دارای بردارهای ویژه $e_1, ..., e_p$ است. مولفه‌ی اصلی jام را c_j نایده و به صورت زیر نشان داده می‌شود.

$$c_j = e_jX = e_{1j}X_1 + e_{2j}X_2 + ... + e_{pj}X_p$$

c_1 اولین مولفه اصلی است و دارای بیشترین واریانس ممکن است. همچنین c_2 دومین مولفه‌ی اصلی است. یکی از کاربردهای PCA کشف ساختارهای نهفته در فضای متغیرها در داده‌های با



شکل ۳.۲: نمایی از عملکرد تحلیل مولفه‌های اصلی

ابعاد بالاست اما همیشه این ساختارها خطی نیستند. بنابراین روش‌های کاهش بعد و انتخاب متغیر همانند PCA قادر به کشف ساختارهای غیرخطی نیستند.

۲.۳.۲ تابع جریمه ریج

در مسائل یادگیری با راهنما[۷]، هنگام ساخت مدل‌های رگرسیون برای داده‌های با ابعاد بزرگ که شامل بسیاری از متغیرها هستند، مشکل همخطی اغلب مشاهده می‌شود. یکی از علائم همخطی این است که برآورد ضرایب رگرسیون بسیار بزرگ بوده و خطاهای معیار مرتبط نیز بسیار بزرگ هستند. این بدان معنی است که ضرایب به خوبی تعریف نشده است. رگرسیون ریج [۱۰] برای غلبه بر مشکل همخطی طراحی شده است که ضریب برآوردشده را به صفر کاهش می‌دهد. به طور خاص ضرایب مدل رگرسیون ریج به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned}\hat{\beta}_{Ridge} &= \arg\min_{\beta\in R^p} \sum_{i=1}^{n} (y_i - x'_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\ &= \arg\min_{\beta\in R^p} \sum_{i=1}^{n} ||y_i - x'_i\beta||_2^2 + \lambda \sum_{j=1}^{p} ||\beta||_2^2\end{aligned} \qquad (۴.۲)$$

[۷] Supervised learning

که در آن $||.||_۲$ تابع نرم $L_۲$ و تابع جریمه ریج یک تابع درجه دوم است و $\lambda > ۰$ پارامتر تنظیم است. در مدل رگرسیون جریمه ریج زمانی که $\lambda = ۰$ است، برآورد OLS [۸] یا همان روش حداقل مربعات معمولی نتیجه می‌شود. با بزرگ شدن λ، مقادیر $\hat{\beta}_{Ridge}$ به صفر نزدیک می‌شوند. پارامتر تنظیم برای ایجاد تعادل دو طرفه و کاهش خطای برآورد انتخاب می‌شود. رگرسیون ریج وقتی ضرایب مقادیر کوچکی هستند عملکرد بهتری دارد اما وقتی ضرایب نسبتا بزرگ هستند و دامنه پارامتر λ کوچک است، عملکرد مناسبی ندارد. رگرسیون ریج ضرایب را دقیقا صفر نمی‌کند مگر در حالت $\lambda = \infty$. پس می‌توان گفت رگرسیون ریج به عنوان یک روش انقباضی پیوسته پیش بینی بهتری را به‌دست می‌آورد. با این حال، رگرسیون ریج همیشه تمام ضرایب را در مدل نگه می‌دارد. از این رو، نمی‌تواند انتخاب متغیر را انجام دهد بنابراین از نظر ارائه یک تفسیر واضح ضعیف است، هرچند که از لحاظ دقت پیش‌بینی خوب است. همانطور که قبلا اشاره شد در مجموعه داده‌ها با ابعاد بالا برخی از متغیرها غیر ضروری هستند و نباید برای انجام پیش بینی در مدل باشند. پس نیاز است ضرایب آن‌ها دقیقا صفر شوند. این کار برای تفسیر مدل و دقت پیش‌بینی بسیار مهم است. جریمه لاسو در بخش بعدی می‌تواند در این وضعیت نتیجه بهتری داشته باشد.

۳.۳.۲ تابع جریمه لاسو

در روش لاسو[۹] [۱۹]، به طور همزمان انقباض ضرایب و انتخاب خودکار متغیر انجام می‌شود، برآورد ضرایب لاسو به صورت زیر تعریف می‌شوند:

$$\hat{\beta}_{lasso} = \underset{\beta \in R^p}{\arg\min} \sum_{i=۱}^{n} ||y_i - x'_i\beta||_۲^۲ + \lambda||\beta||_۱ \qquad (۵.۲)$$

لاسو از تابع نرم $L_۱$ استفاده می‌کند در حالی که ریج از تابع نرم $L_۲$ استفاده می‌کند. تابع نرم $L_۱$ مجموع قدر مطلق ضرایب و $\lambda > ۰$، پارامتر تنظیم است و قدرت تابع جریمه را کنترل می‌کند. اگر $\lambda = ۰$، رگرسیون خطی نتیجه می‌شود و اگر $\lambda = \infty$ آن‌گاه $\hat{\beta}_{lasso} = ۰$ است. بر خلاف جریمه ریج که با افزایش λ دقت برآورد افزایش پیدا می‌کند و واریانس کاهش می‌یابد، در جریمه لاسو انتخاب λ مناسب یک مساله مهم و دشوار است. همچنین بر خلاف رگرسیون ریج که نمی‌تواند ضرایب را صفر کند، رگرسیون لاسو بسیاری از ضرایب متغیرهای زائد را در مدل صفر می‌کند. با افزایش λ بیشتر ضرایب صفر می‌شوند، و ضرایب غیر صفر نیز بیشترین انقباض را دارند. یکی از مزیت‌های خاص جریمه لاسو در مقابل جریمه ریج بحث انتخاب متغیر است. در مقایسه با روش‌های کلاسیک انتخاب متغیر مانند روش انتخاب زیرمجموعه‌ای از متغیرها، لاسو دو مزیت ویژه دارد. اولین مزیت این است که فرآیند انتخاب متغیرها در لاسو به طور مداوم انجام می‌شود و در مقایسه با روش‌های قبلی مانند انتخاب زیرمجموعه و روش

[۸] Ordinary least squares

[۹] LASSO

گام به گام، پایدارتر است. دومین مزیت لاسو، انجام محاسبات برای داده‌های با ابعاد بالا را امکان‌پذیر می‌کند، و بر خلاف روش انتخاب زیرمجموعه‌ای، که محاسبات را به صورت ترکیبی انجام می‌دهد و زمانی که p بزرگ باشد امکان انجام محاسبات را ندارد، عملکرد بهتری دارد. با این حال لاسو نیز محدودیت‌هایی دارد. از جمله حالتی که تعداد مشاهدات بیشتر از تعداد متغیرها باشند یعنی $n > p$ باشد و همبستگی بالایی بین متغیرها وجود داشته باشد، دقت پیش‌بینی رگرسیون ریج در مقایسه با لاسو بهتر است [۱۹]. علاوه بر این لاسو توانایی پیشگویی گروهی را ندارد [۲۵]. یعنی اگر گروهی از متغیرها همبستگی بالایی داشته باشند، لاسو به دلخواه فقط چند تا از متغیرها را از این گروه انتخاب می‌کند. به‌همین دلیل روش لاسو در برخی موارد برای انتخاب متغیر نامناسب است. جریمه لاسو به همه‌ی ضرایب به یک میزان جریمه اعمال می‌کند اما در مواردی که نیاز است تا به برخی از ضرایب جریمه‌های متفاوتی اعمال شود نامناسب است. در همین راستا ژو [۲۴] مدل جریمه لاسو منطبق با این موضوع را پیشنهاد کرد که در این مدل برای اعمال جریمه به ضرایب مختلف وزن‌های متفاوتی (λ_j) در نظر می‌گیرد.

۴.۳.۲ تابع جریمه لاسو گروهی

در بسیاری از مسائل رگرسیونی، متغیرهای توضیحی ساختار گروهی دارند و تمام ضرایب متعلق به یک گروه، همزمان صفر یا غیر صفر می‌شوند. برآوردگر لاسو به صورت زیر تعریف می‌شود:

$$\hat{\beta}_{Lasso} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} (y_i - {x'}_i\beta)^2 + \sum_{j=1}^{p} \lambda_j |\beta_j|. \qquad (۶.۲)$$

صورت‌های مختلفی از جریمه لاسو گروهی برای مواجهه با چنین حالتی طراحی شده‌اند [۲۲].

۵.۳.۲ تابع جریمه شبکه منعطف

شبکه منعطف[۱۰] یک ساختار کلی بر اساس رگرسیون ریج و لاسو دارد و تابع جریمه آن هم به صورت خطی و درجه دوم است که به صورت زیر معرفی می‌شود:

$$\begin{aligned} \hat{\beta}_{\text{ElasticNet}} &= \arg\min_{\beta \in R^p} \sum_{i=1}^{n} (y_i - {x'}_i\beta)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \\ &= \arg\min_{\beta \in R^p} \sum_{i=1}^{n} ||y_i - {x'}_i\beta||_2^2 + \lambda_2 \sum_{j=1}^{p} ||\beta||_2^2 + \lambda_1 \sum_{j=1}^{p} ||\beta||_1. \end{aligned} \qquad (۷.۲)$$

مشابه لاسو، شبکه منعطف به طور همزمان انقباض ضرایب و انتخاب خودکار متغیر را انجام می‌دهد و گروهی از متغیرهای وابسته را انتخاب می‌کند. دقت پیش‌بینی در شبکه منعطف نسبت به جریمه لاسو بهتر است. شبکه منعطف بخصوص وقتی که تعداد ضرایب خیلی بیش‌تر از تعداد مشاهدات است، مفید است.

[۱۰] Elastic net

۴.۲ ابزارهای بررسی عملکرد خوشه‌بندی

برای پاسخ به این سوال که نتایج تولیدشده توسط یک روش خوشه‌بندی، تا چه اندازه مناسب است و چگونه می‌توان نتایج حاصل از روش‌های خوشه‌بندی مختلف را با یکدیگر مقایسه نمود، روش‌های محدودی برای ارزیابی کیفیت خوشه‌بندی معرفی شده‌اند. بطور کلی می‌توان این روش‌ها را بر اساس اطلاعاتی که از خوشه‌بندی ایده‌آل وجود دارد به دو گروه تقسیم نمود.
دسته اول روش‌هایی با مبدا خارجی[۱۱] هستند که در صورت وجود اطلاعات مربوط به خوشه‌بندی ایده‌آل از آن‌ها استفاده می‌شود. در این روش‌ها نتایج حاصل از خوشه‌بندی با این اطلاعات مقایسه می‌شوند. در دسته دوم که با نام روش‌های ذاتی[۱۲] شناخته می‌شوند، به دلیل آنکه اطلاعات ایده‌آل بودن خوشه‌ها در دسترس نیست، مناسب بودن روش با توجه به درجه‌ی تکفیک‌پذیری خوشه‌ها از یکدیگر تعیین می‌شود. از آنجا که اطلاعات مربوط به خوشه‌بندی ایده‌آل را می‌توان نظارتی در شکل‌دهی برچسپ‌های خوشه‌ها دانست، روش‌هایی با مبدا خارجی با نام روش‌های بانظارت و روش‌های ذاتی با نام روش‌های بدون نظارت شناخته می‌شوند. در این پایان‌نامه برای ارزیابی عملکرد خوشه‌بندی از دسته دوم بدون نظارت استفاده می‌کنیم. که در آن تفکیک خوشه‌ها و فشردگی داخل آن‌ها مدنظر است. در بسیاری از این روش‌ها از معیار تشابه بین اشیاء موجود در مجموعه داده‌ها استفاده می‌کنند.

۱.۴.۲ شاخص رند

شاخص رند[۱۳] که توسط ویلیام رند در سال معرفی شده، یک ابزار ارزشیابی برای مساله خوشه‌بندی بر اساس میزان شباهت بین دو خوشه است [۱۴]. این شاخص میزان تطابق خوشه‌های به‌دست آمده و خوشه‌های واقعی را اندازه‌گیری می‌کند.

جدول احتمالی

فرض کنید مجموعه مشاهدات S، دارای n مشاهده $S = (s_1, s_2, ..., s_n)$ باشد و $V = \{v_1, ..., v_C\}$ خوشه‌های واقعی داده‌ها و $U = \{u_1, ..., u_R\}$ خوشه‌های پیش‌بینی شده مجموعه مشاهدات S، باشند با شرایط زیر باشند:

$$\cup_{i=1}^{R} u_i = S = \cup_{j=1}^{C} v_j$$
$$u_i \cap u_i' = \emptyset = v_j \cap v_j'$$

فرض کنید n_{ij} برابر با فراوانی مشاهدات مشترک خوشه‌های u_i و v_j باشد که در جدول ۱.۲ نشان داده شده است. همچنین فرض کنید

[۱۱] Extrinsic methods
[۱۲] Intrinsic methods
[۱۳] Rand Index

جدول ۱.۲: فراوانی مشاهدات مشترک در دو گروه U و V

U/V	$v_{1.}$	$v_{2.}$	$\dots$	$v_{C.}$	مجموع
u_1	n_{11}	n_{12}	$\dots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	$\dots$	n_{2C}	$n_{2.}$
u_R	n_{R1}	n_{R2}	$\dots$	n_{RC}	$n_{R.}$
	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.C}$	$n_{..} = n$

الف‌ـ تعداد جفت مشاهداتی که در گروه V متعلق به یک خوشه هستند و طی فرآیند خوشه‌بندی نیز درون یک خوشه قرار می‌گیرند را با نماد a نشان می‌دهیم:

$$a = \sum_{i=1}^{R} \sum_{j=1}^{C} \binom{n_{ij}}{2}$$

ب‌ـ تعداد جفت مشاهداتی را که درون یک رده قرار داشته‌اند و طی فرآیند خوشه‌بندی در خوشه‌های مختلف قرار گرفته‌اند را با نماد b نشان می‌دهیم:

$$b = \sum_{i=1}^{R} \binom{n_{i.}}{2} - \sum_{i=1}^{R} \sum_{j=1}^{C} \binom{n_{ij}}{2}.$$

ج‌ـ تعداد جفت مشاهداتی را که طی فرآیند خوشه‌بندی در یک خوشه قرار گرفتند اما متعلق به یک رده نیستند را با نماد c نشان می‌دهیم:

$$c = \sum_{j=1}^{C} \binom{n_{.j}}{2} - \sum_{i=1}^{R} \sum_{j=1}^{C} \binom{n_{ij}}{2}$$

د‌ـ تعداد جفت مشاهداتی که در رده‌های متفاوتی قرار دارند و طی فرآیند خوشه‌بندی نیز در خوشه‌های متفاوت قرار گرفته‌اند را با کمیت d نشان می‌دهیم:

$$d = \binom{n}{2} - a - b - c.$$

تعداد کل جفت مشاهدات را M در نظر می‌گیریم و به صورت زیر نشان می‌دهیم:

$$M = \binom{n}{2} = a + b + c + d.$$

در این صورت شاخص رند به صورت زیر تعریف می‌شود:

$$RI = \frac{a+d}{M}.$$

که مقدار آن در بازه‌ی $[۰,۱]$ قرار می‌گیرد. حال می‌توانیم شاخص رند تعدیل‌یافته را به صورت زیر تعریف کنیم.

۲.۴.۲ شاخص رند تعدیل‌یافته

شاخص رند تعدیل‌یافته (ARI) [۱۴] یکی از محبوب‌ترین ابزار ارزیابی سازگاری بین دو تقسیم‌بندی داده‌ها در زمینه تشخیص الگو است. شاخص رند تعدیل‌یافته، نسخه‌ی اصلاح‌شده شاخص رند است. مقدار این شاخص در بازه $[-۱,۱]$ قرار می‌گیرد. مقدار $ARI = ۱$ نشان‌دهنده همپوشانی کامل بین دو گروه است، در حالیکه مقدار $ARI = -۱$ نشان‌دهنده‌ی ناهماهنگی کامل دو گروه است. صورت کلی این شاخص به صورت زیر تعریف می‌شود:

$$ARI = \frac{index - Expectedindex}{Maximu\min dex - Expectedindex}.$$

همچنین می‌توان شاخص رند تعدیل‌یافته را به صورت زیر نوشت:

$$ARI = \frac{\binom{n}{۲}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{۲}^{۲} - [(a+b)(a+c)+(c+d)(b+d)]}. \qquad (۸.۲)$$

۵.۲ انتخاب تعداد خوشه‌ها

یکی از مسائل مهم در تحلیل خوشه‌ای، مسئله انتخاب تعداد خوشه‌هاست. تعیین درست خوشه‌ها در یک مجموعه داده نه تنها برای بعضی از الگوریتم‌های خوشه‌بندی مانند K-میانگین به عنوان پارامتر لازم است، بلکه گروه‌بندی صحیح در تحلیل خوشه‌ای را نیز کنترل می‌کند. تعیین تعداد خوشه‌ها می‌تواند توازن مناسبی میان قابلیت فشردگی و درستی خوشه‌بندی برقرار نماید.

اغلب به دلیل آنکه تعداد خوشه‌ها مبهم است، تعیین تعداد درست خوشه‌ها کار ساده‌ای نیست. یافتن تعداد مناسب خوشه‌ها به شکل توزیع و مقیاس مجموعه داده‌ها و همچنین دقتی که از فرآیند خوشه‌بندی انتظار داریم، بستگی دارد. روش‌های متعددی جهت برآورد تعداد خوشه‌ها وجود دارند. از جمله این روش‌ها را کالینسکی و هاراباس [۳]، هارتیگان [۷] و

[۱۴] Adjusted rand index

کارزانفسکی [۱۱] معرفی کرده‌اند که همه این روش‌ها بر اساس مجموع توان دوم فاصله درون خوشه‌ای و میان خوشه‌ای است، و به تازگی نیز روش‌هایی دیگر از جمله شاخص Gap [۱۸] و شاخص پایائی معرفی شده‌اند [۲۱]. در اینجا به صورت خلاصه برخی از روش‌های رایج و کارآمد را معرفی می‌کنیم.
یکی از این روش‌ها جهت تعیین تعداد درست خوشه‌ها این است که منحنی مجموع واریانس‌های درون خوشه‌ای نسبت به تعداد خوشه‌ها رسم شود و نقطه‌ای از منحنی که پس از آن تغییر محسوسی وجود ندارد را برای تعداد خوشه‌های مناسب در نظر بگیریم. از نظر تکنیکی با داشتن یک مقدار مانند k که بزرگتر از صفر است و اعمال یک الگوریتم خوشه‌بندی مانند k ـ میانگین و سلسله‌مراتبی می‌توان مجموع واریانس‌های درون خوشه‌ای را محاسبه نمود. سپس منحنی این مجموع واریانس‌ها را نسبت به مقدار k رسم می‌شود. اولین و مهم‌ترین نقطه‌ی تغیر و تحول در منحنی، تعداد درست خوشه‌ها را نشان می‌دهد[۱۶].

۱.۵.۲ شاخص پایائی

برای برآورد تعداد خوشه‌ها از شاخص پایای[۱۵] استفاده می‌شود [۱۶]. منظور از شاخص پایایی این است که روش‌های مختلف خوشه‌بندی در تکرارهای مختلف، با شرایط یکسان دارای نتایج تقریبا مشابه باشند. برای این منظور در تکرارهای مختلف با تعداد خوشه‌های مختلف چندین روش خوشه‌بندی را روی نمونه‌های مستقل اجرا می‌کنیم و در حالتی که اجرای این روش‌ها رفتار یکسانی داشته باشند را به عنوان تعداد خوشه انتخابی برای انجام خوشه‌بندی برآورد می‌کنیم. بدین صورت که برای هر k داده‌شده، شاخص پایائی برای تابع $\phi(.;k,z^n)$ به صورت زیر محاسبه می‌شود:

$$stab(\phi,k)=E\left\{corr(\phi(.;k,z^n),\phi(.;k,z'^n))\right\}$$

که $\phi(.;k,z^n)$ و $\phi(.;k,z'^n)$ به ترتیب خوشه‌بندی به‌دست آمده از اجرای ϕ روی z و z' است که z و z' دو نمونه مستقل با احتمال $P(x)$ هستند و ϕ یکی از روش‌های خوشه‌بندی است که بر روی مشاهدات اعمال می‌کنیم. همبستگی بین خوشه‌بندی $\phi_1(x)$ و $\phi_2(x)$ اجرا شده روی داده‌های z^n به صورت زیر قابل محاسبه است:

$$corr(\phi_1,\phi_2)=\frac{P(I_1=I_2=I)-P(I_1=1)P(I_2=1)}{\sqrt{P(I_1=1)(1-P(I_1=1))P(I_2=1)(1-P(I_2=1))}}$$

که در آن $I_1=I\left\{\phi_1(X),\phi_1(Y)\right\}$ و $I_2=I\left\{\phi_2(X),\phi_2(Y)\right\}$ و X و Y دو نمونه مستقل و هم توزیع با احتمال $P(x)$ هستند.

[۱۵] Stability

اعتبارسنجی متقابل برای برآورد تعداد خوشه‌ها با استفاده از شاخص پایائی

این روش شاخص پایائی را برای تقسیم‌بندی ایجادشده روی داده‌ها در هر بار انجام اعتبارسنجی متقابل، محاسبه می‌کند و تعداد خوشه‌ها را برآورد می‌کند [۲۱] که برای تعداد تکرار $c = ۱, \ldots, C$ گام اول تا گام سوم را تکرار می‌کنیم.

گام اول: تکرار cام مشاهدات $(x_۱, \ldots x_n)$ را به‌صورت $(x_۱^{*c}, \ldots x_n^{*c})$ بازنویسی می‌کنیم.

گام دوم: مشاهدات $(x_۱^{*c}, \ldots x_n^{*c})$ را به صورت تصادفی در سه دسته با طول m تقسیم می‌کنیم و گروه‌بندی‌های جدید را به صورت زیر می‌نویسیم:

$$z_۱^{*c} = (x_۱^{*c}, \ldots x_m^{*c}); \qquad z_۲^{*c} = (x_m^{*c}, \ldots x_{۲m}^{*c}); \qquad z_۳^{*c} = (x_{۲m+۱}^{*c}, \ldots x_n^{*c})$$

گام سوم: دو گروه اول را به‌عنوان مجموعه آموزشی در نظر می‌گیریم و برای $i = ۱, ۲$ مقدار $z_i^{*c}; i$ زیر را محاسبه می‌کنیم:

$$\begin{aligned} V(x_i^{*c}, x_j^{*c}, \phi, k, z_۱^{*c}, z_۲^{*c}) =& I(I(\phi(x_i^{*c}; k, z_۱^{*c}) = \phi(x_j^{*c}; k, z_۱^{*c})) \\ &+ I(\phi(x_i^{*c}; k, z_۲^{*c}) = \phi(x_j^{*c}; k, z_۱^{*c}))) = ۱) \end{aligned} \tag{۹.۲}$$

با استفاده از مجموعه آزمون $z_۳^{*c}$، مقدار شاخص پایائی را به صورت زیر بدست می‌آوریم:

$$\hat{stab}^{*c}(\phi, k) = \sum_{۲m+۱ \le i \le j \le n} V(x_i^{*c}, x_j^{*c}, \phi, k, z_۱^{*c}, z_۲^{*c}). \tag{۱۰.۲}$$

گام چهارم: پس از C بار تکرار گام اول تا گام سوم مقدار شاخص پایائی را به صورت زیر می‌نویسیم:

$$\hat{stab}(\phi, k) = \frac{۱}{C} \sum_{c=۱}^{C} stab^{*c}(\phi, k). \tag{۱۱.۲}$$

گام پنجم: بیشترین مقدار برآوردشده در گام چهارم را به‌عنوان تعداد خوشه‌های مورد نیاز برای انجام خوشه‌بندی انتخاب می‌کنیم:

$$\hat{k} = \arg \max_{۲ \le k \le K} \hat{stab}(\phi, k). \tag{۱۲.۲}$$

۲.۵.۲ شاخص Gap

شاخص Gap، یک روش استاندارد برای تعیین تعداد خوشه‌ها در مجموعه‌ای از داده‌ها است [۱۸]. فرض کنید $\{x_i \quad ; i = ۱, \ldots, n\}$ مجموعه مشاهدات باشند که $x_i = (x_{i۱}, \ldots, x_{ip})$ بردار مشاهده iام برای p متغیر است که در k خوشه مجزا، $c_۱, c_۲, \ldots, c_r$ گروه‌بندی شده باشند همچنین D_r را مجموع فواصل نقاط داخل خوشه rام به صورت زیر تعریف می‌کنیم:

$$D_r = \sum_{i, i' \in C_r} d_{i,i'}$$

که در آن $d_{i,i'}$ فاصله مشاهده i و مشاهده i' داخل خوشه rام هستند. در این صورت با فرض اینکه

$$n_r = |D_r|$$

و

$$W_k = \sum_{r=۱}^{k} \frac{۱}{۲n_r} D_r$$

شاخص Gap به صورت زیر معرفی می‌شود:

$$Gap_n(k) = E'_n(\log(W_k)) - \log(W_k)$$

که در آن E'_n مقدار مورد انتظار از نمونه‌ای با حجم n از توزیع مرجع است. برای برآورد تعداد خوشه‌ها، k، کمترین مقدار به‌دست آمده از شاخص Gap را در نظر می‌گیرند [۱۸].

فصل ۳

روش‌های بهینه خوشه‌بندی و خوشه‌بندی اصلاح‌شده

اگر مساله خوشه‌بندی را در یک چارچوب رگرسیونی تعریف کنیم، می‌توانیم به بسیاری از ویژگی‌های آماری با تحمیل معیارهای جریمه دست پیدا کنیم. در این فصل به امتیازبندی بهینه که فیشر برای انجام تحلیل ممیزی خطی استفاده کرد، اشاره می‌کنیم. سپس از این ایده استفاده کرده و با معرفی روش خوشه‌بندی ممیزی بهینه[۱] (ODC) و به منظور کاهش بعد، مساله خوشه‌بندی که یک مساله یادگیری بدون نظارت است در قالب مساله رگرسیون ریج بیان شده است تا بتوان همانند اندیشه‌ی مولفه‌های اصلی، نوع دیگری از ترکیب خطی متغیرهای اولیه را برای ساختن متغیرهای جدید استخراج نمود. سپس یکی از الگوریتم‌های خوشه‌بندی نظیر k ـ میانگین را برای مشاهدات تبدیل‌یافته‌ی جدید به‌کار می‌گیریم.

۱.۳ تعاریف

فرض کنید $X_{n\times p} = [x_1, ..., x_n]'$ ماتریسی حاوی n مشاهده p بعدی باشد که در آن $x_i = (x_{i1}, .., x_{ip})$ معرف مقادیر p متغیر توضیحی برای i امین مشاهده است. همچنین فرض کنید

[۱] Optimal discriminant clustering

C رده مجزا داشته باشیم که هر x_i متعلق به یک رده مجزا است به‌طوری که

$$\forall j = ۱, ..., C \Rightarrow x_j \in C_j, C_j \underset{j \neq j}{\cap} C'_j = \emptyset$$

همچنین مجموعه شاخص و اندیس‌گذار $V = \{۱, ..., n\}$ را به C زیرمجموعه جدا از هم با شرایط زیر افراز می‌کنیم:

$$\begin{aligned} \forall i \neq j \quad &; V_i \cap V_j = \emptyset \\ &\bigcup_{j=۱}^{C} V_j = V \end{aligned} \tag{۱.۳}$$

که در آن $|V_j| = card(V_j) = n(V_j) = n_j$ و $\sum_{j=۱}^{C} n_j = n$ است. همچنین ماتریس زیر

$$E = [e_{ij}]_{n \times c} = \begin{bmatrix} e_{۱۱} & \cdots & e_{۱c} \\ \vdots & \vdots & \vdots \\ e_{n۱} & \cdots & e_{nc} \end{bmatrix}$$

ماتریس نشانگر است، و به صورت زیر تعریف می‌شود:

$$e_{ij} = \begin{cases} ۱ & ; x_i \in c_j \\ ۰ & ; o.w \end{cases} \tag{۲.۳}$$

که در آن c_j رده jام است، همچنین ماتریس قطری Π را به صورتی که قطر اصلی شامل تعداد عناصر هر رده باشد، تعریف می‌کنیم. یعنی $\Pi = diag(n_۱, ..., n_c)$. در ادامه چند نماد مرتبط با Π را به صورت زیر تعریف می‌کنیم:

$$\pi = (n_۱, ..., n_c)'$$
$$\pi^{\frac{۱}{۲}} = (\sqrt{n_۱}, ..., \sqrt{n_c})'$$

و همچنین ۱_n بردار nـبعدی از یک‌هاست، بنا به تعاریف بالا روابط زیر را در نظر می‌گیریم:

لم ۱.۱.۳.

$$۱'_n E = ۱'_c \Pi = \pi' \tag{۳.۳}$$

برهان. با توجه به تعاریف فوق به سادگی می‌توان نتیجه گرفت

$$۱'_n E = \begin{bmatrix} ۱, & \cdots & ,۱ \end{bmatrix} \begin{bmatrix} e_{۱۱} & \cdots & e_{۱c} \\ \vdots & \vdots & \vdots \\ e_{n۱} & \cdots & e_{nc} \end{bmatrix}_{n \times c} = \left[\sum_{i=۱}^{n} e_{i۱}, \cdots, \sum_{i=۱}^{n} e_{ic} \right]_{۱ \times c}$$

با توجه به رابطه (۲.۳) داریم،

$$\forall i \neq j; \quad e_{ij} = ۰$$

$$۱'_n E = [n_۱, \cdots, n_c]_{n\times ۱} = \pi'.$$

از طرف دیگر رابطه نیز داریم

$$۱'_n \Pi = [۱, \cdots, ۱]_{۱\times n} \begin{bmatrix} n_۱ & \cdots & ۰ \\ \vdots & \ddots & \vdots \\ ۰ & \cdots & n_c \end{bmatrix}_{n\times c} = [n_۱, \cdots, n_c]_{۱\times c} = \pi'$$

پس تساوی برقرار است. □

لم ۳.۱.۲.

$$E۱_c = ۱_n \tag{۴.۳}$$

برهان. از تعاریف فوق داریم

$$E۱_c = \begin{bmatrix} e_{۱۱} & \cdots & e_{۱c} \\ \vdots & \ddots & \vdots \\ e_{n۱} & \cdots & e_{nc} \end{bmatrix}_{n\times c} \begin{bmatrix} ۱ \\ \vdots \\ ۱ \end{bmatrix}_{c\times ۱} = \begin{bmatrix} \sum_{i=۱}^{n} e_{i۱} \\ \vdots \\ \sum_{i=۱}^{n} e_{ic} \end{bmatrix}_{n\times ۱}$$

با توجه به رابطه (۲.۳) داریم،

$$\forall i \neq j; \quad e_{ij} = ۰$$

$$E۱_c = \begin{bmatrix} ۱ \\ \vdots \\ ۱ \end{bmatrix}_{n\times ۱} = ۱_n.$$

□

لم ۳.۱.۳.

$$۱'_c \pi = n \tag{۵.۳}$$

برهان. اثبات بدیهی است. □

لم ۴.۱.۳.

$$E'E = \Pi \tag{۶.۳}$$

برهان. با توجه به تعاریف قبلی داریم،

$$E'E = \begin{bmatrix} e_{11} & \cdots & e_{n1} \\ \vdots & \vdots & \vdots \\ e_{1c} & \cdots & e_{nc} \end{bmatrix}_{c\times n} \begin{bmatrix} e_{11} & \cdots & e_{1c} \\ \vdots & \vdots & \vdots \\ e_{n1} & \cdots & e_{nc} \end{bmatrix}_{n\times c}$$
$$= \begin{bmatrix} \sum_{i=1}^{n} e_{i1}^2 & \cdots & \sum_{i=1}^{n} < e_{i1}.e_{ic} > \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^{n} < e_{in}.e_{i1} > & \cdots & \sum_{i=1}^{n} e_{ic}^2 \end{bmatrix}_{n\times c}$$

که منظور از $< e_{ij}.e_{ij'} >$ ضرب داخلی بین e_{ij} و $e_{ij'}$ است، با توجه به رابطه (۲.۳)

$$\forall j \neq j' < e_{ij}.e_{ij'} >= 0$$

پس داریم

$$= \begin{bmatrix} \sum_{i=1}^{n} e_{i1}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sum_{i=1}^{n} e_{ic}^2 \end{bmatrix}_{n\times c} = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & n_c \end{bmatrix}_{n\times c} = \Pi$$

□

لم ۵.۱.۳.

$$\Pi^{-1}\pi = 1'_c \qquad (۷.۳)$$

اثبات بدیهی است.

۲.۳ ماتریس امتیازدهی

برای مسائل طبقه‌بندی cـرده، هستی [۸]، ماتریس امتیازدهی $\Theta \in R^{c\times(c-1)}$ را با شرایط زیر تعریف کرد:

$$\Theta'(E'E)\Theta = \Theta'\Pi\Theta = I_{c-1}$$

که در آن jامین سطر ماتریس Θ را یک امتیاز یا وزن برای رده jام در نظر می‌گیرند. اکنون با توجه به کار انجام شده می‌توانیم تعریف فوق را اصلاح کنیم.

تعریف ۱.۲.۳. در مساله طبقه‌بندی cـرده، که تعداد عناصر رده jام برابر n_j است ماتریس $\Theta \in R^{c\times(c-1)}$ را به‌عنوان ماتریس امتیاز با شرایط زیر در نظر می‌گیریم:

$$\Theta'\Pi\Theta = I_{c-1} \quad \pi'\Theta = 0$$

که می‌توان نتیجه گرفت رابطه زیر برقرار است:

$$\Theta'\Theta = \Pi^{-1} - \frac{1}{n} I_{c-1} I'_{c-1}. \qquad (۸.۳)$$

اثبات در [۸] است.

۱.۲.۳ امتیازبندی بهینه برای LDA

در مساله طبقه‌بندی چند_رده، مدل جریمه‌ای را بر اساس امتیازبندی بهینه تعریف می‌کنند. در واقع داریم

$$\min_{\Theta,W} \left\{ f(\Theta, W) \triangleq \frac{1}{2} \|E\Theta - \mathrm{H}_n XW\| + \frac{\sigma^2}{2} tr(W'W) \right\} \qquad (۹.۳)$$

به طوری که

$$\Theta'\Pi\Theta = I_{c-1} \qquad ; \pi'\Theta = 0$$

که در آن ماتریس امتیاز $\Theta \in R^{c\times(c-1)}$ و $W \in R^{p\times(c-1)}$ ماتریس ضرایب بدست آمده است و در مقایسه با مدل جریمه معرفی‌شده تیبشیرانی و هستی [۸]، شرط $\pi'\Theta = 0$ به رابطه (۹.۳) اضافه شده است.

۲.۲.۳ ماتریس امتیازدهی در یادگیری بدون نظارت

معرفی ماتریس امتیازدهی بهینه در مسائل یادگیری بدون ناظر، از مهمترین مسائلی است که در این فصل به‌دنبال آن هستیم تا بتوانیم بر اساس این ماتریس، امتیاز بهینه یک مدل جریمه برای انجام کاهش بعد همزمان با خوشه‌بندی را براساس مدل جریمه (۹.۳) معرفی کنیم.
اگر در رابطه (۹.۳) عبارت $E\Theta = Y$ قرار دهیم، مدل جریمه جدید به صورت زیر تعریف می‌شود:

$$\min_{Y,W} \left\{ f(Y, W) \triangleq \frac{1}{2} \|Y - H_n XW\|_F^2 + \frac{\sigma^2}{2} tr(W'W) \right\} \qquad (۱۰.۳)$$

به طوری که

$$1'_n Y = 0 \qquad Y'Y = I_{c-1}.$$

۳.۳ خوشه‌بندی در قالب مدل‌های رگرسیونی انقباضی

در مدل‌های رگرسیونی چندگانه $Y = X\beta + \varepsilon$، که در آن Y, β, ε به ترتیب بردارهای p_بعدی، خطا، پارامترهای رگرسیونی و مقادیر پاسخ است، اگر متغیرهای توضیحی دارای همخطی باشند، آنگاه مقادیر برآورد پارامترها بسیار بزرگ بوده و ناپایدار هستند. برای حل این مشکل،

موضوعی به نام رگرسیون انقباضی پدید آمده است که مینیمم‌سازی مجموع توان‌های دوم خطا را با شرط محدود کردن مقادیر β در نظر می‌گیرد. این مسئله مینیمم‌سازی، که در حالت خاص موسوم به رگرسیون Ridge است، به صورت زیر بیان می‌شود:

$$\hat{\beta} = \arg\min_{\beta \in R^p} ||Y - X\beta|| + \lambda ||\beta||_F^2$$

که در آن $\|\beta\|_F = \sqrt{tr(\beta'\beta)}$ و λ پارامتر تنظیم نامیده شده و میتوان با استفاده از اعتبارسنجی متقابل مقدار بهینه این پارامتر را تعیین نمود.

۴.۳ خوشه‌بندی ممیزی بهینه (ODC)

در روش جدید خوشه‌بندی موسوم به خوشه‌بندی ممیزی بهینه یا ODC با در نظر گرفتن اندیشه‌ی روش مولفه‌های اصلی (PCA) ، ترکیبی خطی از متغیرهای X به صورت

$$S = H_n X (X' H_n X + \sigma^2 I_r)^{-1} X' H_n \qquad (۱۱.۳)$$

بنا به رابطه (۱۱.۳)، مولفه‌های Z به شکل زیر می‌شوند:

$$Z = H_n X (X' H_n X + \sigma^2 I_p)^{-1} X' H_n Y = SY. \qquad (۱۲.۳)$$

بنابراین با انتخاب $W = (X' H_n X + \sigma^2 I_p)^{-1} X' H_n Y$، داریم

$$Z = H_n X W = (z^{(1)}, ..., z^{(k-1)})$$

که در آن $H_n = I_n - \frac{1}{n} 1_n 1_n'$ ماتریس مرکزی و I_n ماتریس همانی $n \times n$ است. همچنین

$$W = (w_{ij})_{p \times (k-1)} = (w^{(1)}, ..., w^{(k-1)}) = (w_1, ..., w_p)'$$

ماتریسی است که همانند ضرایب رگرسیونی β در روند حل یک مساله مینیمم‌سازی به‌دست خواهد آمد که k معرف تعداد خوشه‌های از قبل تعیین‌شده است. بردارهای $z^{(1)}, z^{(2)}, ...$ بترتیب اولین مولفه ODC ، دومین مولفه‌ی ODC و... نامیده می‌شوند که می‌توان عملکرد آن را با مولفه‌های اصلی مقایسه کرد. اگر چه خوشه‌بندی یک مسئله یادگیری بدون نظارت است و در آن متغیر پاسخ وجود ندارد، اما ژانگ و دی [۲۳] با رویکرد فوق و با تعریف ماتریس امتیازدهی در نقش متغیر پاسخ Y مساله خوشه‌بندی را در قالب رگرسیون Ridge عنوان کردند.

تعریف ۱.۴.۳. برای مسائل خوشه‌بندی ماتریس امتیاز نمونه‌ای $Y_{n \times (K-1)}$ را با شرط

$$Y'Y = I_{k-1} \qquad 1_n' Y = 0$$

تعریف می‌کنیم، که j امین سطر ماتریس امتیازدهی Y، یک امتیاز یا نمره برای خوشه jام محسوب می‌شود. در اینجا k لزوما تعداد خوشه‌ها نیست. مثلا می‌تواند نشانگر فضای کاهش بعد یافته باشد.

در روش خوشه بندی ODC، با الگویی برگرفته از رگرسیون Ridge ، برآورد از حل مساله مینیمم‌سازی زیر به‌دست می‌آیند:

$$(\hat{W}, \hat{Y}) = \underset{W,Y}{\arg\min} ||Y - H_n X W||_F^2 + \lambda ||W||_F^2 \qquad (۱۳.۳)$$

که

$$\|W\|_F = \sqrt{tr(W'W)}$$

نرم اقلیدسی است. و λ پارامتر تنظیم است که مقدار بهینه آن بر اساس روش پیشنهادی [۲۰] که به آن در بخش ۲.۴.۳ اشاره خواهد شد، محاسبه می‌شود.

قضیه ۱.۴.۳. فرض کنید $\hat{Y}$ و $\hat{W}$ مینیمم کننده (۱۳.۳) هستند. در این صورت $\hat{W} = (X'H_nX + \lambda_2 I_p)^{-1}X'H_n\hat{Y}$ و $\hat{Y}$ ماتریس متعامد مقادیر ویژه بالا برای $H_nX(X'H_nX + \lambda_2 I_p)^{-1}X'H_n$ هستند.

برهان. اثبات در [۲۳] بیان شده است. □

.

۱.۴.۳ الگوریتم ODC

گام اول : برای λ داده شده، با استفاده از رابطه (۱۳.۳) داریم:

$$\hat{W} = (X'H_nX + \lambda_2 I_P)^{-1}X'H_n\hat{Y}$$

گام دوم : مولفه‌های ODC را به صورت زیر محاسبه می‌کنیم:

$$Z = (z_1, ..., z_p)' = H_nX\hat{W}$$

گام سوم : خوشه‌بندی k ـ میانگین را روی z_i ها برای $i = 1, ..., n$ انجام می‌دهیم.
گام چهارم : گروه‌بندی z_i ها را جایگزین گروه‌بندی x_i ها در نظر می‌گیریم. در گام سوم بجای روش خوشه‌بندی k ـ میانگین می‌توان از روش‌های خوشه‌بندی متداول دیگر استفاده کرد.

۲.۴.۳ الگوریتم انتخاب بهینه پارامتر تنظیم

پارامتر تنظیم تاثیر معنی‌داری بر عملکرد روش ODC دارد. در اینجا و همانند روش‌های آماری، برآورد این پارامتر را توسط روش اعتبارسنجی متقابل انجام می‌دهیم.
گام اول : ابتدا داده‌ها را به گروه‌هایی با اندازه برابر C (به طور مثال $C = 5$) که با نماد

$X_{(1)}, ..., X_{(C)}$ نشان داده می‌شوند، تقسیم می‌کنیم.

گام دوم : بر اساس روش اعتبارسنجی 5-fold یکی از گروه‌ها $c \in \{1, \ldots, C\}$ را به عنوان مجموعه آزمون انتخاب کرده و بقیه گروه‌ها را به عنوان مجموعه آموزشی در نظر می‌گیریم. با استفاده از داده‌های آموزشی خارج از گروه c، $\hat{W}^{\lambda_2}_{(-c)}$ را با توجه به λ های داده‌شده و با توجه به رابطه (۱۳.۳) محاسبه می‌کنیم.

گام سوم : با توجه به (۱۳.۳) و مقدار λ_2 های داده‌شده $\hat{Y}_c^{\lambda_2}$ را برای داده‌های گروه c به صورت زیر محاسبه می‌کنیم:

$$\hat{Y}_c^{\lambda_2} = \arg\min_{W,Y} ||Y - H_{n_c} X_{(c)} \hat{W}^{\lambda_2}_{(-c)}||^2_F$$

$$s.t. \;\; Y^{'} Y = I_{k-1} \; and \; 1^{'}_{n_c} Y = 0 \quad (2)$$

گام چهارم با توجه به رابطه (۱۴.۳) می‌توانیم مقدار $\hat{\lambda}_2$ را بهینه انتخاب کنیم:

$$\hat{\lambda}_2 = \arg\min \left\{ \frac{1}{c} \sum_{c=1}^{C} \left\| \hat{Y}^{\lambda_2}_{(c)} - H_n X_{(c)} \hat{W}^{\lambda_2}_{(-c)} \right\|^2_F \right\} \qquad (۱۴.۳)$$

۳.۴.۳ نحوه‌ی انتخاب تعداد خوشه‌ها در ODC

به‌طور خلاصه به نحوه‌ی انتخاب تعداد خوشه‌ها در الگوریتم ODC اشاره می‌کنیم. فرض کنید K متعلق به مجموعه $\{1, ..., k_{\max}\}$ باشد و k_{max} بیش‌ترین تعداد خوشه مد نظر است. بعضی از روش‌ها با پیش‌فرض $K = 2$ خوشه‌بندی را انجام می‌دهند. با اجرای خوشه‌بندی ممیزی بهینه ODC، با انتخاب λ از روش اعتبارسنجی متقابل معرفی‌شده، مقداری را با عنوان شاخص تعداد خوشه تخمین می‌زند (مثلا شاخص Gap) و بسته به شاخص مورد نظر ماکزیمم یا مینیمم این مقادیر را تحت عنوان $\hat{k}$ معرفی می‌کند.

۵.۳ خوشه‌بندی ممیزی بهینه اصلاح‌شده (SODC)

در مواجهه با داده‌های دارای ابعاد بالا با p متغیر که p خیلی بزرگ است، برای انجام خوشه‌بندی و کاهش ابعاد این داده‌ها در بخش ۴.۳ خوشه‌بندی ممیزی بهینه را در قالب مدل‌های رگرسیون انقباضی معرفی کردیم و مولفه‌های $Z = H_n X \hat{W}$ که ترکیب خطی از متغیرهای $x_{i1}, \cdots, x_{ip}$، $i = 1, \cdots, n$، هستند، را بیان کردیم. تفسیر مولفه‌های کاهش بعد یافته $z_i = \sum_{j=1}^{p} \hat{w}_j x_{ij}$ ها که شامل p متغیر هستند، خیلی سخت است. مشارکت متغیرهای زائد در خوشه‌بندی بر عملکرد آن تاثیر منفی می‌گذارد و خوشه‌بندی را تحت تاثیر قرار می‌دهد. برای حل این مشکل در خوشه‌بندی ODC، به رابطه (۱۳.۳) جریمه لاسو گروهی را اضافه می‌کنیم و برآورد اصلاح‌شده از $\hat{W}$ را به‌دست می‌آوریم.

برای حذف تاثیر متغیر jام از تمام $k-۱$ مولفه‌ی ODC باید $\hat{w}_j$ را صفر کنیم که این کار مرسوم نیست. پس برای رسیدن به این هدف با اضافه کردن جریمه لاسو گروهی به الگوریتم ODC، مدل جریمه جدیدی را با نام خوشه‌بندی ممیزی بهینه اصلاح‌شده[۲] (SODC) را معرفی می‌کنیم:

$$(\hat{W},\hat{Y}) = \arg\min_{W,Y} \|Y - H_n X W\|_F^۲ + \lambda_۲ \|W\|_F^۲ + \lambda_۱ \sum_{j=۱}^{p} \|w_j\|_۲ \qquad (۱۵.۳)$$

به طوری که

$$Y'Y = I_{k-۱} \quad ۱_n' Y = ۰$$

که در آن $\|w_j\|$ نرم اقلیدسی w_j است. مراحل انجام خوشه‌بندی SODC همانند الگوریتم خوشه‌بندی ODC انجام می‌شود به این ترتیب که پس از برآورد $\hat{W}$ اصلاح‌شده، خوشه‌بندی بر اساس $Z = H_n X \hat{W}$ که یک ماتریس $n \times (k-۱)$ بعدی از مولفه‌های SODC است، انجام می‌شود. همچنین ماتریس $Z_{n\times(k-۱))}$ به صورت زیر می‌توان نوشت:

$$Z = (z^{(۱)}, ..., z^{(j)})$$

که $z^{(۱)}$ مولفه‌ی اول، $z^{(۲)}$ مولفه‌ی دوم و $z^{(j)}$ مولفه‌ی jام SODC هستند.

قضیه ۱.۵.۳. در الگوریتم بهینه‌سازی (۱۵.۳) برآورد Y از تجزیه مقادیر منفرد U و V به صورت زیر محاسبه می‌شود:

$$\hat{Y} = UV'$$

و تجزیه svd [۳] ماتریس مولفه‌ای Z به صورت زیر است:

$$H_n X W = UDV'$$

برهان. با استفاده از روش ضرایب لانگراژ و با توجه به Y ، W داده شده، فرض کنید $X^* = H_n X W$ و $k^* = k-۱$ تابع هدف را به صورت زیر می‌نویسیم:

$$L(Y, X^*, \Lambda, b) = \frac{۱}{۲} tr(Y'Y) - tr(Y'X^*) + \frac{۱}{۲} tr(X^{*'}X^*) - \frac{۱}{۲} tr(\Lambda(Y'Y - I_{k^*})) - tr(b'Y'۱_n)$$

که در آن Λ، ماتریس متقارن $k^* \times k^*$ بعدی از ضرایب لانگراژ و b بردار $۱ \times k^*$ بعدی از ضرایب لانگراژ باشد، داریم:

$$\frac{\partial L}{\partial Y} = Y - X^* - Y\Lambda - ۱_n b' = ۰$$

طرفین رابطه را در $۱_n'$ ضرب می‌کنیم و $b = ۰$ می‌شود پس می‌توان نوشت $Y - X^* - Y\Lambda = ۰$، و با ضرب Y' در طرفین رابطه داریم:

$$I_{k^*} - Y'X^* = \Lambda. \qquad (۱۶.۳)$$

[۲] Sparse optimal discriminant clustering

[۳] Singular value decomposition

الف‌ـ اگر $n \geq k^*$ باشد تجزیه SVD برای X^* به صورت $X^* = UDV'$ است که در آن $U \subset R^{n \times k^*}$، $D \subset R^{k^* \times k^*}$ و $V' \subset R^{k^* \times k^*}$ هستند. داریم:

$$1'_n X^* V D^{-1} = 1'_n U = 0$$

از آنجایی که

$$\hat{Y} = UV'$$

با ضرب طرفین مساوی در $1'_n$ داریم:

$$1'_n \hat{Y} = 1'_n UV' = 0$$

و همچنین با ضرب $\hat{Y}'$ در طرفین رابطه

$$\hat{Y} = UV'$$

داریم:

$$\hat{Y}'\hat{Y} = \hat{Y}'UV' = VU'UV' = I_{k^*}$$
$$I_{k^*} - \hat{Y}'X^* = I_{k^*} - VU'UDV' = I_{k^*} - VDV' = \Lambda$$

پس ثابت می‌شود که $\hat{Y}$ مینیم‌کننده رابطه (۱۵.۳) است.

ب‌ـ اگر $n < k^*$ باشد و تجزیه SVD، $X^* = UDV'$ باشد که $U \subset R^{n \times n}$، $D \subset R^{n \times n}$ و $V' \subset R^{n \times k^*}$ هستند، با استفاده از تجزیه SVD، به طور مشابه اثبات می‌شود.

□

برای به‌دست آوردن مینیمم رابطه (۱۵.۳) با توجه به W و Y داده‌شده، مطابق روش گفته شده در [۲۲] عمل می‌کنیم. فرض می‌کنیم ستون‌های Y بردارهای $n(k-1)$ بعدی y باشند و ماتریس X به صورت زیر باشد. $X = (X_1, ..., X_j)$ و X_jها ماتریس $n(k-1) \times (k-1)$ بعدی $X_j = diag(x^{(j)}, ..., x^{(j)})$ که هر $x^{(j)}$ ستون jام ماتریس X است و $j = 1, \dots p$ است.

$$\hat{W} = \arg\min_w \left\| Y - \sum_{j=1}^{p} X_j w_j \right\|_F^2 + \lambda_2 \|w_j\|_2^2 + \lambda_1 \sum_{j=1}^{p} \|w_j\|_2 \qquad (۱۷.۳)$$
$$s.t. Y'Y = I_{k-1} and 1'_n Y = 0$$

قضیه ۲.۵.۳. $\hat{W} = (w_1, ..., w_p)'$ از حل مساله مینیمم‌سازی رابطه (۱۷.۳) به‌دست می‌آید و هر $\hat{w}_j$ به صورت زیر محاسبه می‌شود:

$$\hat{w}_j = \frac{(\|v_j\|_2 - \frac{\lambda_1}{2})_+}{(1 + \lambda_2)\|v_j\|_2} v_j$$

که در آن $v_j = X_j(Y - \sum_{l \neq j} X_l \hat{w}_l)$ و $(a)_+ = \begin{cases} a & ; a > 0 \\ 0 & ; o.w \end{cases}$

برهان. با توجه به رابطه (۱۷.۳)، نسبت به w_j مشتق می‌گیریم و برابر صفر قرار می‌دهیم:

$$-۲X_j'\left(Y-\sum_{j=۱}^{p}X_jw_j\right)+۲\lambda_۲w_j+\lambda_۱\frac{w_j}{\|w_j\|_۲}=۰ \qquad j=۱,...,p$$

با توجه به اینکه

$$\hat{W}=(w_۱,...,w_p)'$$

اگر نامساوی

$$\left\|X'_j(Y-\sum_{l\neq j}X_l\hat{w}_l)\right\|_۲<\left\|\frac{\lambda_۱}{۲}\right\|$$

برقرار باشد $\hat{w}_j=۰$ است بنابراین برای هر j، $\hat{w}_j=۰$ اگر

$$\lambda>\lambda_{\max}=\max\left\|X'_jY\right\|$$

و در غیر اینصورت $\hat{w}_j$ به صورت زیر است:

$$\hat{w}_j=(X'_jX_j+\lambda_۲+\frac{\lambda_۱}{۲\|w_j\|_۲})^{-۱}V_j \qquad (۱۸.۳)$$

که در آن

$$V_j=X_j(Y-\sum_{l\neq j}X_l\hat{w}_l)$$

و با توجه به اینکه $X'_jX_j=I_{k-۱}$ ماتریس قطری است، پس معکوس پذیر است، می‌توان رابطه بالا را به صورت زیر نوشت:

$$\hat{w}_j=\frac{۲\|w_j\|_۲}{\lambda_۱+۲(۱+\lambda_۲)\|w_j\|_۲}V_j$$

رابطه بالا را به صورت زیر نیز می‌توان ساده کرد:

$$\|w_j\|_۲=\frac{۲\|v_j\|_۲-\lambda_۱}{۲(۱+\lambda_۲)}$$

با جایگذاری عبارت بالا در رابطه [؟]، قضیه اثبات می‌شود. □

با توجه به قضیه ۲.۵.۳ اگر $\lambda_۱\geq۲\|v_j\|_۲$ باشد $\hat{w}_j=۰$ می‌شود. بنابراین برای به‌دست آوردن برآورد اصلاح‌شده $\hat{W}$ پارامتر تنظیم $\lambda_۱$ باید به اندازه کافی بزرگ باشد. بر اساس قضایای ۱.۵.۳ و ۲.۵.۳ برای شروع ابتدا مقادیر اولیه $\hat{W}_۰$ و $\hat{Y}_۰$ را برآورد می‌کنیم و در هر تکرار برآوردهای جدیدی از $\hat{W}$ و $\hat{Y}$ را به‌دست می‌آوریم. این کار را تا زمانی که مقادیر برآوردشده همگرا شوند ادامه می‌دهیم. بنا به رابطه (۱۷.۳) چون تابع هدف محدب است پس دارای مقادیر یکتاست. همچنین در محاسبات عددی الگوریتم به سرعت پس از چند تکرار، همگرا می‌شود.
برای انجام محاسبات یک فهرست از $\lambda_۱$ را بر اساس زیر تهیه می‌کنیم:

$$\lambda_۱\in\left\{۱۰^{-\tau+\log_{۱۰}(\lambda_{\max})\times\frac{l}{l}}\right\};l=\{۱,...,L\}$$

که $\lambda_{\max} \leq \lambda_1 \leq \lambda_1^\circ = 10^{-\tau}$، و λ_{max} به صورت زیر است

$$\lambda_{\max} = \max_j \left\| X'_j Y \right\|.$$

اگر $\lambda < \lambda_{\max}$ باشد، برآورد به‌دست آمده از w_j یعنی $\hat{w}_j = 0$ می‌شود. به این ترتیب با در نظر گرفتن $\lambda_1 = \lambda_1^\circ$ مقادیر اولیه $\hat{W}_0$ و $\hat{Y}_0$ را از الگوریتم ODC محاسبه می‌کنیم و هر بار بر اساس λ_1 انتخاب‌شده از فهرست بالا، مقادیر برآوردشده برای $\hat{W}$ و $\hat{Y}$ را به روز می‌کنیم.

۱.۵.۳ الگوریتم انتخاب پارامتر تنظیم λ_1 در SODC

با توجه به اینکه در داده‌های با ابعاد بالا بسیاری از متغیرها بر عملکرد خوشه‌بندی تاثیر منفی دارند، برای بهبود عملکرد خوشه‌بندی این متغیرها را از مدل حذف می‌کنیم. در الگوریتم خوشه‌بندی SODC، با استفاده از جریمه لاسو گروهی انتخاب متغیر را انجام می‌دهیم. در یادگیری با نظارت بسیاری از معیارهای معرفی‌شده برای سنجش عملکرد انتخاب متغیر مانند معیار AIC و BIC معرفی شده‌اند. اما در یادگیری بدون نظارت چون تعریف دقیقی از تابع هدف وجود ندارد کار انتخاب متغیر بسیار سخت است. در این روش بجای انتخاب همزمان λ_1 و λ_2 می‌توانیم از الگوریتم زیر استفاده کنیم و λ_2 را با استفاده از روش اعتبارسنجی متقابل گفته شده در بخش ۲.۴.۳ محاسبه کنیم و آن را به عنوان یک مقدار ثابت در نظر بگیریم.

۲.۵.۳ مراحل اجرای روش کاپا برای انتخاب λ_1

گام اول: فهرستی از λ_1ها را انتخاب می‌کنیم.

گام دوم: به طور تصادفی داده‌ها را داخل دو گروه X_1^{b*} و X_2^{b*} تقسیم می‌کنیم که $b = 1, \ldots, B$.

گام سوم: به ازای هر کدام از λ_1ها خوشه‌بندی SODC را روی مجموعه داده‌های گروه X_1^{b*} و X_2^{b*} انجام می‌دهیم $\hat{A}^{b*}_{1\lambda_1}$ و $\hat{A}^{b*}_{2\lambda_1}$ را محاسبه می‌کنیم.

گام چهارم: ضریب کاپا بین $\hat{A}^{b*}_{1\lambda_1}$ و $\hat{A}^{b*}_{2\lambda_1}$ را به صورت زیر محاسبه می‌کنیم:

$$k(\hat{A}^{b*}_{1\lambda_1}, \hat{A}^{b*}_{2\lambda_1}) = \frac{pr(a) - pr(e)}{1 - pr(e)} \qquad b = 1, ..., B$$

که در آن

$$pr(a) = \frac{(|\hat{A}_{1\lambda_1} \cap \hat{A}_{2\lambda_1}| + |\hat{A}^c_{1\lambda_1} \cap \hat{A}^c_{2\lambda_1}|)}{p}$$

و

$$pr(e) = \frac{(|\hat{A}_{1\lambda_1}||\hat{A}_{2\lambda_1}| + |\hat{A}^c_{1\lambda_1}||\hat{A}^c_{2\lambda_1}|)}{p^2}.$$

گام پنجم: در هر مرحله مقدار

$$k(\lambda_۱) = \frac{۱}{B}\sum_{b=۱}^{B} k(\hat{A}^{b*}_{۱\lambda_۱}, \hat{A}^{b*}_{۲\lambda_۱})$$

را محاسبه می‌کنیم.

گام ششم: بیشترین مقدار به‌دست آمده در مرحله قبل در B تکرار پنج مرحله قبل را به عنوان $\lambda_۱$ انتخاب می‌کنیم:

$$\lambda_۱ = \arg\max k(\lambda_۱).$$

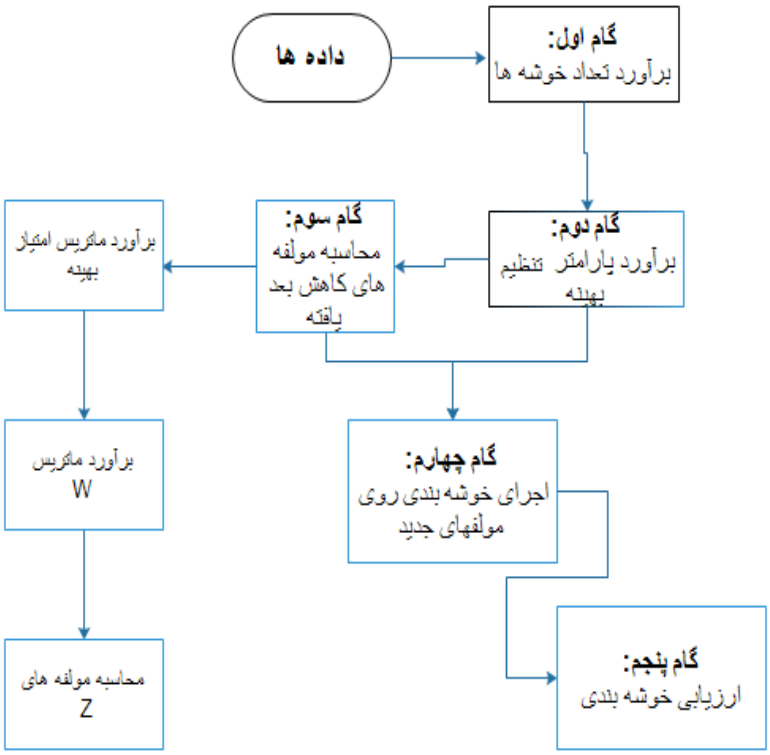
فصل ۴

مطالعه شبیه‌سازی

در فصل سوم، روش خوشه‌بندی مبتنی بر تابع جریمه ریج و لاسو معرفی شد و به چگونگی تبدیل مسئله ی خوشه‌بندی که یک روش بدون ناظر است به مسئله رگرسیون جریمه پرداختیم. در این فصل ابتدا الگوریتم مورد استفاده برای مطالعه ی شبیه‌سازی را شرح داده و سپس آن را بر روی مجموعه داده‌های شبیه‌سازی شده اجرا نموده و ضمن مطالعه خواص روش پیشنهادی، نتایج حاصل از آن را با نتایج اخذشده از اعمال الگوریتم PCA مقایسه می‌کنیم.

۱.۴ الگوریتم ODC در مطالعه شبیه‌سازی

در این فصل برای انجام همزمان کاهش بعد و خوشه‌بندی، روشی مبتنی بر رگرسیون ریج را معرفی نموده و هچنین برای کاهش تاثیر عملکرد متغیرهای زاید به مدل معرفی‌شده، جریمه لاسو گروهی را اعمال می‌کنیم تا بتوانیم متغیرهای زاید را از مدل حذف کنیم. مدل خوشه‌بندی جدید، نسخه اصلاح شده خوشه‌بندی ODC است. نتایج حاصل از انجام این دو مدل خوشه‌بندی را با روش تحلیل مولفه‌های اصلی مقایسه می‌کنیم. ابتدا بر اجرای الگوریتم ODC تمرکز می‌کنیم. در یک نگاه اجمالی، می‌توان فلوچارت الگوریتم ODC را به صورتی که در شکل ۱.۴ نمایش داده شده، در نظر گرفت.



شکل ۱.۴: فلوچارت الگوریتم خوشه‌بندی ODC

گام‌های این الگوریتم را به صورت جزئی‌تر در زیر شرح داده شده است:

۱. در اولین گام، تعداد خوشه‌های مناسب برای خوشه بندی از طریق روش های پیشنهادی در بخش ۵.۲ تعین می‌گردد.

۲. با استفاده از الگوریتم معرفی‌شده در بخش ۲.۴.۳، پارامتر تنظیم بهینه λ_2 را به‌دست می‌آوریم.

۳. برای محاسبه مولفه‌های کاهش بعد یافته در این گام، لازم است مراحل زیر اجرا شوند:

(آ) ماتریس امتیاز بهینه را برآورد می‌کنیم.

(ب) ماتریس ضرایب W را برآورد می‌کنیم.

(ج) مولفه‌های جدید کاهش بعد یافته را محاسبه می‌کنیم.

۴. یکی از روش‌های خوشه‌بندی مانند kـمیانگین یا روش‌های دیگر خوشه‌بندی روی مولفه‌های کاهش بعد یافته جدید اجرا می‌شود.

۵. خوشه‌های حاصل از مرحله قبل، توسط یکی از معیارهای ارزیابی خوشه‌بندی معرفی‌شده در بخش ۴.۲ به نام شاخص رند تعدیل‌یافته ارزیابی می‌گردند.

مثال ۱.۱.۴. در این بخش برای بررسی روش ODC از یک مجموعه داده با نام data.all شبیه سازی شده که در نرم افزار R از پکیج SODC فراخوانی شده‌اند، استفاده نموده، داده‌های شبیه سازی شده متشکل از ۱۰ متغیر و ۱۵۰ مشاهده هستند که خوشه‌ی هر یک از مشاهدات، $C_i,\ i = ۱, ..., ۱۵۰$ ، یکی از مقادیر ۱،۲ یا ۳ است که به تصادف و از توزیع یکنواخت گسسته (DU) انتخاب شده‌اند.

$$C_i \sim DU\{۱, ۲, ۳\}$$

در اینجا فرض شده است که فقط $q = ۲$ متغیر در خوشه‌بندی موثر هستند و نوع تاثیر آنها به این صورت تعریف شده است که برای هر i مقادیر این دو متغیر $X_۱, X_۲$ از توزیع نرمال $N_۲(m(C_i), \Sigma)$ تولید شده‌اند که میانگین توزیع نرمال برای هر مشاهده عبارت است از

$$m(C_i) = \mu(-۱/۲, ۱/۲)I(C_i = ۱) + \mu ۱_q I(C_i = ۲) + \mu(۱/۲, -۱/۲)I(C_i = ۳) \qquad (۱.۴)$$

$$X = (X_۱, X_۲) \sim N_۲(m(C_i), \Sigma) \Rightarrow$$
$$f(X_۱, X_۲) = \frac{۱}{(۲\pi)^{۱/۲}|\Sigma|^{۱/۲}} e^{-(-\frac{۱}{۲})(X-\mu)'\Sigma^{-۱}(X-\mu)}$$

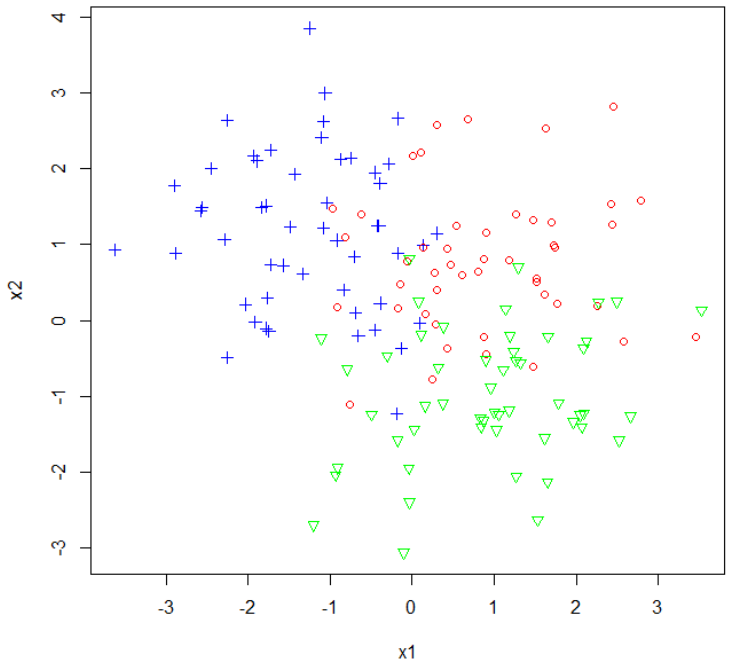
در اینجا q نمایانگر تعداد متغیرهای موثر در خوشه‌بندی است، مقادیر زوج را می‌پذیرد و $\Sigma = (\sigma_{jk} = r^{I(j\neq k)})$ مشخص‌کننده‌ی ماتریس کوواریانس $q \times q$ بین آن‌ها است که در آن $r = ۰, ۰/۵$ و به‌صورت زیر محاسبه می‌شود:

$$\Sigma = (\sigma_{jk} = r^{I(j\neq k)}) = \begin{bmatrix} \sigma_{۱۱} & \sigma_{۱۲} \\ \sigma_{۲۱} & \sigma_{۲۲} \end{bmatrix}_{۲\times ۲}$$
$$= \begin{bmatrix} r & r^{۰} \\ r^{۰} & r \end{bmatrix} = \begin{bmatrix} ۰/۵ & ۱ \\ ۱ & ۰/۵ \end{bmatrix}$$

از رابطه (۱.۴) مقادیر $m(C_i)$ را محاسبه می‌کنیم

$$m(C_۱) = \mu(-۱/۲, ۱/۲)I(C_۱ = ۱) + \mu ۱_۲ I(C_۲ = ۲) + \mu(۱/۲, -۱/۲)I(C_۱ = ۳)$$
$$= \mu(-۱, ۱) = \begin{bmatrix} -\mu \\ \mu \end{bmatrix}$$
$$m(C_۲) = \mu(-۱/۲, ۱/۲)I(C_۲ = ۱) + \mu ۱_۲ I(C_۲ = ۲) + \mu(۱/۲, -۱/۲)I(C_۲ = ۳)$$
$$= \mu(۱, ۱) = \begin{bmatrix} \mu \\ \mu \end{bmatrix}$$
$$m(C_۳) = \mu(-۱/۲, ۱/۲)I(C_۳ = ۱) + \mu ۱_۲ I(C_۳ = ۲) + \mu(۱/۲, -۱/۲)I(C_۳ = ۳)$$
$$= \mu(۱, -۱) = \begin{bmatrix} \mu \\ -\mu \end{bmatrix}$$

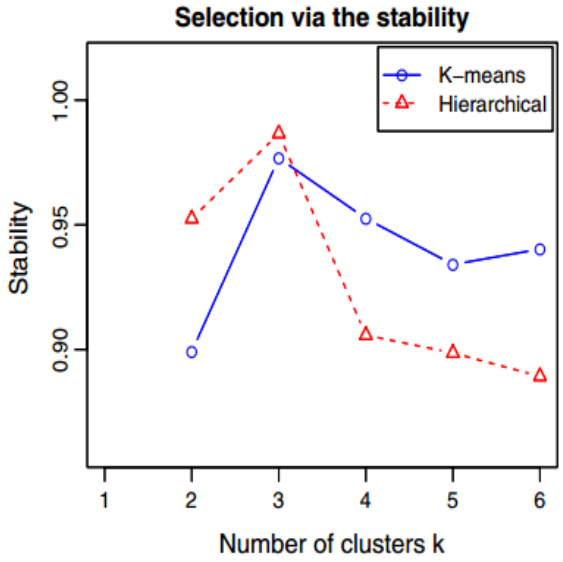
μ میزان همپوشانی خوشه‌ها در نظر گرفته شده‌است بطوری که با افزایش μ همپوشانی خوشه‌ها کمتر و با کاهش آن، همپوشانی خوشه‌ها بیشتر می‌شود [۲۰]. سایر متغیرهای غیر موثر در خوشه‌بندی، یعنی $p-q=۸$ متغیر دیگر از توزیع $N_۸(۰, I_{۸\times ۸})$ تولید شده و به عنوان زاید به مجموعه داده‌ها اضافه شده‌اند. نمودار پراکنش متغیرهای موثر $X_۱$ و $X_۲$ و خوشه‌های تشکیل شده، در شکل ۲.۴ نشان داده شده‌اند.



شکل ۲.۴: نمودار پراکنش متغیرهای $X_۱$ و $X_۲$ در مثال شبیه‌سازی

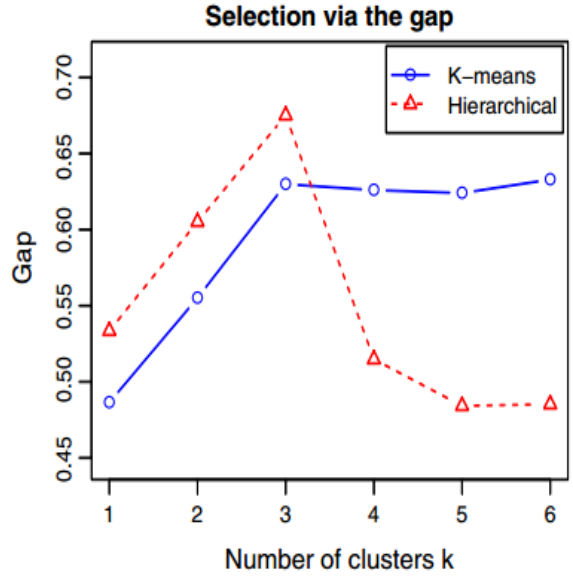
علیرغم اینکه در این مثال تعداد واقعی خوشه‌ها مشخص است، قصد داریم تعداد خوشه‌ها را با استفاده از شاخص پایا و شاخص Gap که در بخش ۵.۲ اشاره شده است، برآورد کنیم که جزئیات آن به ترتیب در بندهای الف و ب آمده است.

الف ــ به منظور برآورد تعداد خوشه‌ها توسط شاخص پایایی، خوشه‌بندی k ــ میانگین و خوشه‌بندی سلسله‌مراتبی را به ازای تعداد خوشه‌های مختلف $(k=۲,۳\ldots,۶)$ اجرا گردیده و مقدار شاخص پایایی برای هر یک از خوشه‌بندی‌های صورت‌گرفته، محاسبه شده است. سپس تعداد خوشه‌هایی که منجر به پایا بودن نتایج در هر دو روش اجراشده می‌شود به عنوان تعداد بهینه‌ی خوشه‌ها در نظر گرفته شده است. در این مثال تعداد خوشه‌ها برای هر یک از روش‌های خوشه‌بندی اجراشده برابر با عدد ۳ تعیین شده است. لذا برآوردی بدون خطا صورت گرفته است. شکل ۳.۴ نحوه برآورد تعداد خوشه‌ها را با استفاده از شاخص پایایی نشان می‌دهد.



شکل ۳.۴: روند انتخاب براورد تعداد خوشه‌ها با شاخص پایایی

ب‌ـ به منظور تعیین خوشه‌ها با استفاده از شاخص Gap، ابتدا با تعداد تکرار $k = ۱,...,۶$، مقدار شاخص را محاسبه می‌کنیم و کمترین مقدار به‌دست آمده را برای تعداد خوشه‌ها به عنوان k برآورد می‌کنیم. در شکل ۴.۴ بعد از انجام خوشه‌بندی k‌ـمیانگین و خوشه‌بندی سلسله‌مراتبی روی مجموعه داده‌های data.all، در هر بار تکرار $k = ۱,...,۶$، مقدار شاخص Gap را محاسبه می‌کنیم.



شکل ۴.۴: روند انتخاب تعداد خوشه‌ها با استفاده از آماره Gap

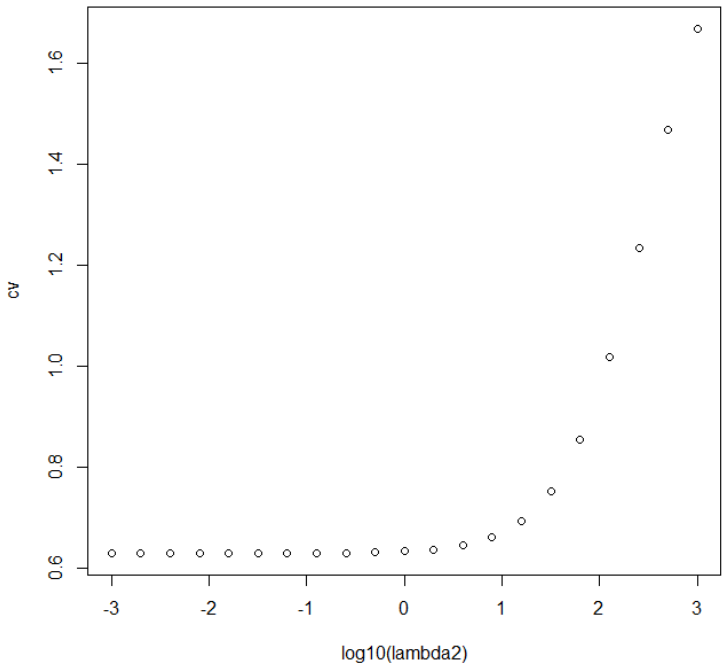
در اینجا برای برآورد تعداد خوشه‌ها با استفاده از شاخص Gap، ۵۰ بار شبیه‌سازی انجام دادیم که در جدول ۱.۴ نتایج را مشاهده می‌کنید.

جدول ۱.۴: انتخاب تعداد خوشه‌ها با استفاده از شاخص فاصله با ۵۰ بار شبیه‌سازی

p	μ	فراوانی k					
		۱	۲	۳	۴	۵	۶
۱۰	۲.۰	۰	۰	۴۹	۰	۱	۰
	۲.۲	۰	۰	۴۵	۱	۰	۴
	۲.۴	۰	۰	۴۸	۱	۱	۰
۵۰	۲.۰	۰	۰	۵۰	۰	۰	۰
	۲.۲	۰	۰	۵۰	۰	۰	۰
	۲.۴	۰	۰	۵۰	۰	۰	۰
۱۰۰	۲.۰	۰	۱۲	۳۸	۰	۰	۰
	۲.۲	۰	۰	۵۰	۰	۰	۰
	۲.۴	۰	۰	۵۰	۰	۰	۰
۲۰۰	۲.۰	۲۷	۲۳	۰	۰	۰	۰
	۲.۲	۷	۳۴	۹	۰	۰	۰
	۲.۴	۰	۷	۴۳	۰	۰	۰

با توجه به نتایج مشاهده‌شده در جدول ۱.۴، در حالتی که تعداد متغیرها $p = ۱۰,۵۰,۱۰۰$ است، شاخص Gap خیلی خوب عمل می‌کند. اما در حالت $p = ۲۰۰$ و $\mu = ۲/۰, \mu = ۲/۲$ عملکرد نامناسبی در برآورد تعداد خوشه‌ها دارد که ضرورت انتخاب متغیر در خوشه‌بندی را نشان می‌دهد. در ادامه تعداد خوشه‌ها را برابر ۳ در نظر می‌گیریم.

برای انتخاب پارامتر تنظیم $\lambda_۲$ مشبکه‌ای شامل مقادیر $(۱۰^{-۳+۳\times(l/۱۹)})$ که $l = ۱,\ldots,۱۹$ در نظر گرفته شد، و سرانجام مقدار بهینه آن بر اساس اعتبارسنجی متقابل 5-fold برابر با $\hat{\lambda} = ۳۵/۴۸۱۳۴$ به‌دست آمد. شکل ۵.۴ روند خطای به‌دست آمده را با تغییر مقدار پارامتر تنظیم نشان می‌دهد.



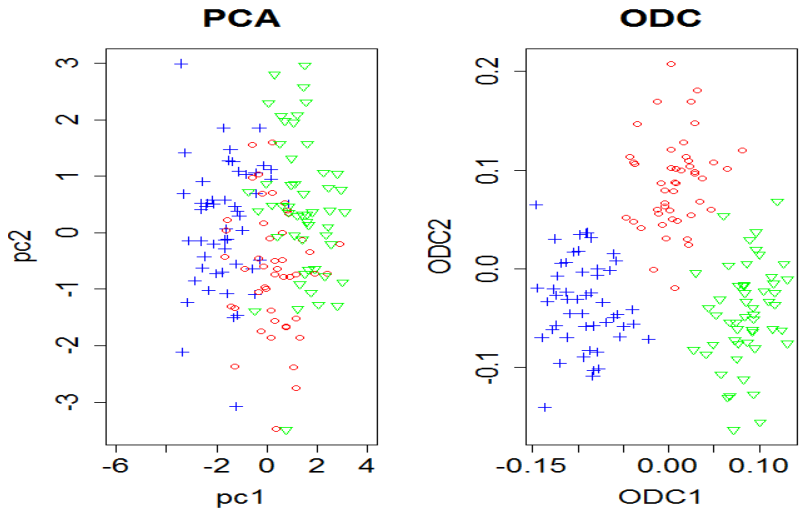
شکل ۵.۴: روند انتخاب λ در ODC

با استفاده از λ بهینه، خوشه‌بندی این مجموعه داده‌ها توسط روش ODC انجام شد و نتایج حاصل از آن با اعمال روش‌های خوشه‌بندی k ـ میانگین روی مولفه‌های اصلی (PCA)، اعمال خوشه‌بندی فقط روی متغیرهای اصلی (ORACLE) و خوشه‌بندی روی تمام متغیرها (ALL) توسط شاخص رند تعدیل‌یافته مقایسه گردید. جدول ۲.۴ عملکرد این چهار خوشه‌بندی را در مقایسه با یکدیگر نشان می‌دهد.

جدول ۲.۴: مقادیر شاخص رند تعدیل یافته

الگوریتم خوشه‌بندی	شاخص رند تعدیل یافته (ARI)
ORACLE	۰/۴۴۰۴۰۴۱
ALL	۰/۳۰۴۸۶۰۵
PCA	۰/۳۹۶۲۹۱۷
ODC	۰/۹۲۲۶۳۵۶

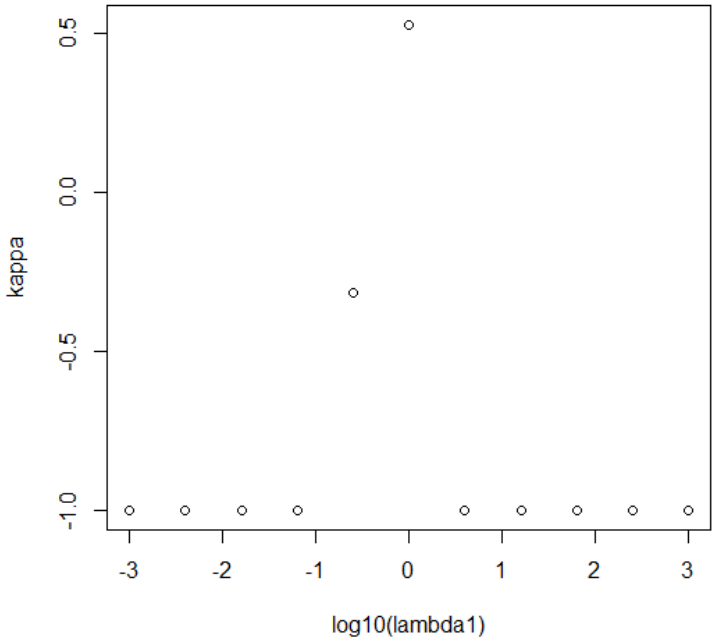
با توجه به اینکه مقادیر ARI نزدیک به عدد ۱ نشان از کارآیی حداکثری خوشه‌بندی است لذا میتوان نتیجه گرفت که با افزایش بعد و همپوشانی خوشه‌ها، PCA کارآیی نداشته و ابزار کاهش بعد ODC عملکرد بسیار بهتری نسبت به سایر حالت‌ها دارد. در شکل ۶.۴ نمودار اولین دو مولفه اصلی و همچنین اولین دو مولفه ODC نشان داده شده‌اند. واضح است که روش پیشنهادی ODC به خوبی قادر به جداسازی خوشه‌ها بوده در حالیکه روش مولفه‌های اصلی چنین قابلیتی را از خود نشان نمی‌دهد.



شکل ۶.۴: نمودار پراکنش حاصل از مولفه‌های ODC (راست) و نمودار پراکنش حاصل از انجام k ـ میانگین روی مولفه‌های PCA (چپ)

۲.۴ الگوریتم SODC در مطالعه شبیه‌سازی

خوشه‌بندی SODC نیز کاملا مشابه خوشه‌بندی ممیزی بهینه ODC انجام می‌شود با این تفاوت که علاوه بر اعمال جریمه ریج، جریمه لاسو گروهی را اعمال می‌کنیم تا بتوانیم متغیرهایی که عملکرد خوشه‌بندی را ضعیف می‌کنند شناسایی و از مدل حذف کنیم. برای انتخاب متغیر در الگوریتم ODC روش انتخاب بر اساس ضریب کاپا در بخش ۱.۵.۳ معرفی شده است. با توجه به مثال بخش ۱.۱.۴ برای انتخاب پارامتر تنظیم λ_1، مشبکه‌ای شامل مقادیر $(۱۰^{-۳+\lambda_1^{\max}\times(l/۱۹)})$ که $l = ۱,\ldots,۱۹$ را در نظر می‌گیریم و با استفاده از ضریب کاپای معرفی‌شده در بخش ۲.۵.۳ نحوه‌ی انتخاب پارامتر تنظیم بهینه λ_1 را شبیه‌سازی کردیم و در شکل ۷.۴ نشان می‌دهیم.
با به‌دست آوردن پارامتر تنظیم λ_1 خوشه‌بندی SODC را روی مجموعه داده‌های مثال قبل انجام دادیم و نتایج حاصل از این کار را با خوشه‌بندی ODC توسط شاخص رند تعدیل‌یافته مقایسه می‌کنیم. جدول ۳.۴ عملکرد این خوشه‌بندی را در مقایسه با خوشه‌بندی ODC نشان



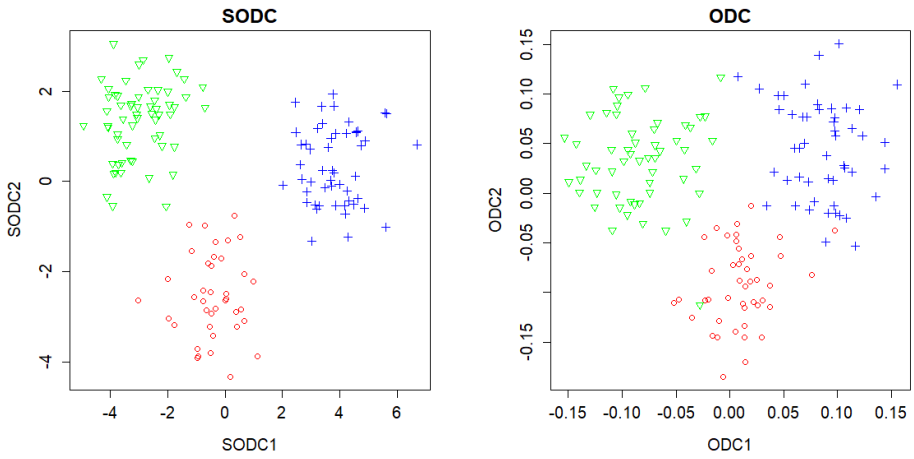
شکل ۷.۴: روند انتخاب λ_1 با ضریب کاپا

می‌دهد. با توجه به مقادیر ARI خوشه‌بندی SODC کاراتر از روش ODC است.

جدول ۳.۴: مقادیر شاخص رند تعدیل یافته

الگوریتم خوشه بندی	شاخص رند تعدیل یافتهARI
SODC	۰/۹۸۰۵۳۳
ODC	۰/۹۰۲۱۴۴

در شکل ۸.۴ نمودار اولین دو مولفه ODC و اولین دو مولفه‌های SODC نشان داده شده است. واضح است که روش پیشنهادی SODC به خوبی قادر به جداسازی خوشه‌ها می‌باشد.



شکل ۸.۴: نمودار پراکنش مولفه‌های اصلی حاصل از دو روش SODC (سمت چپ) و ODC (سمت راست).

۳.۴ بحث و نتیجه‌گیری

در این پایان‌نامه الگوریتم جدیدی برای کاهش ابعاد در خوشه‌بندی معرفی شد و برای عملکرد بهتر این الگوریتم، از روش اعتبارسنجی متقابل برای انتخاب پارامتر تنظیم بهینه بهره گرفتیم. همچنین برای برآورد تعداد خوشه‌ها از آماره gap و شاخص پایایی استفاده کردیم. در ادامه نشان دادیم که ODC به عنوان یکی از ابزار کاهش ابعاد در خوشه‌بندی بهتر از بسیاری از روش‌های موجود مانند PCA، که ساختار خوشه‌ها را در نظر نمی‌گیرد، عمل می‌کند. علاوه بر این برای انجام خوشه‌بندی همزمان با انتخاب متغیر به الگوریتم ODC، جریمه لاسو گروهی اضافه شده و الگوریتم جدید SODC معرفی شده است.

الگوریتم SODC همسو با روش‌هایی مانند، روش PCA اصلاح‌شده [۲۴]، روش تحلیل همبستگی کانونی اصلاح‌شده [۱۷] و تحلیل ممیزی اصلاح‌شده (SDA) [۴] پیشنهاد شده است. اندیشه ODC برگرفته از روش FDA که طبقه‌بندی را بر اساس امتیازبندی بهینه انجام می‌دهد، گرفته شده است، و سپس نسخه اصلاح‌شده آن SDA می‌نامند، با اضافه کردن جریمه لاسو به روش FDA معرفی شده‌است. از این رو ایده انجام خوشه‌بندی با امتیازبندی بهینه منجر به معرفی الگوریتم ODC شده‌است و نسخه اصلاح‌شده آن SODC با اضافه کردن جریمه لاسو گروهی به روش ODC معرفی شده است. موضوع انتخاب پارامتر تنظیم در SODC خیلی متفاوت با روش SDA است زیرا در مسائل طبقه‌بندی انتخاب پارمتر تنظیم با طبقه‌بندی نادرست، منجر به رخداد خطا می‌شود. همچنین در مسائل خوشه‌بندی انتخاب پارامتر تنظیم خیلی سخت است. بر همین اساس در الگوریتم SODC با معرفی ضریب کاپا انتخاب پارامتر تنظیم را انجام دادیم. ضریب کاپا در مسائل رگرسیونی سازگار است اما سازگاری ان در مسائل خوشه‌بندی

ثابت نشده است.

۴.۴ پیشنهادات

به‌منظور انجام تحقیقات و پژوهش‌های آتی در زمینه کاهش ابعاد در مساله خوشه‌بندی می‌توان از راهکارهای ذیل بهره گرفت:
علاوه بر بسیاری از ابزارهای موجود، از روش SODC نیز می‌توان در زمینه کاهش ابعاد برای رسم نمودارهای پراکنش استفاده کرد. رسم نمودارهای پراکنش قبل از تحلیل بسیار مهم است چون با بررسی این نمودارها می‌توانیم به ایده‌های بزرگی در مورد موضوعات مهم دست یابیم. موضوعات مهمی که در خوشه بندی مورد توجه هستند مشخص نبودن خوشه‌های واقعی، بزرگتر بودن برخی از خوشه‌ها از دیگر خوشه‌ها، یا خوشه‌های محدب می‌باشند. برای حل این مشکلات باید چندین روش خوشه‌بندی مختلف را اجرا کنیم تا به نتایج پایدار و سازگار برسیم.

پیوست آ

روش‌های عددی

آ.۱ مقادیر منفرد

برای ماتریس $A_{m\times n}$ ماتریس $A'A$ و AA' یک ماتریس هرمیتی [۱] و مثبت معین است:

۱. اگر $m < n$ باشد جذر مقادیر ویژه AA' را مقادیر منفرد [۲] ماتریس A می‌نامند.

۲. اگر $m > n$ باشد جذر مقادیر ویژه $A'A$ را مقادیر منفرد ماتریس A می‌نامند.

مثال آ.۱.۱. مقادیر منفرد ماتریس A را بیابید.

$$A_{۲\times ۳} = \begin{bmatrix} ۳ & ۱ & ۱ \\ -۱ & ۳ & ۱ \end{bmatrix}$$

از آنجائیکه ماتریس حقیقی است، لذا مقادیر منفرد بصورت جذر مقادیر ویژه ماتریس AA^T تعریف می‌شود. بنابراین

$$AA^T = \begin{bmatrix} ۳ & ۱ & ۱ \\ -۱ & ۳ & ۱ \end{bmatrix} \begin{bmatrix} ۳ & -۱ \\ ۱ & ۳ \\ ۱ & ۱ \end{bmatrix} = \begin{bmatrix} ۱۱ & ۱ \\ ۱ & ۱۱ \end{bmatrix}$$

[۱] self-adjoint

[۲] singular value

حال مقادیر ویژه ماتریس AA^T را بدست می‌آوریم،

$$|\lambda I - AA^T| = \begin{vmatrix} \lambda - ۱۱ & -۱ \\ -۱ & \lambda - ۱۱ \end{vmatrix} = (\lambda - ۱۰)(\lambda - ۱۲)$$

پس مقادیر ویژه برابر هستند با

$$\lambda_۱ = ۱۲, \qquad \lambda_۲ = ۱۰.$$

از این‌رو مقادیر منفرد برای ماتریس A بصورت زیر بدست می‌آیند:

$$\sigma_۱ = \sqrt{۱۲}, \qquad \sigma_۲ = \sqrt{۱۰}.$$

۲.آ تجزیه ماتریس‌ها بر اساس مقادیر منفرد

یکی از مهم‌ترین روش‌های تجزیه ماتریس‌ها تجزیه بر اساس مقادیر منفرد است. در این روش یک ماتریس مانند $A_{m\times n}$ با رتبه K را می‌توان به‌صورت زیر تجزیه کرد:

$$A = U\Sigma V^T \tag{۱.آ}$$

که در آن $U_{\mathrm{m}\times m} = \begin{bmatrix} u_{۱۱} & \dots & u_{۱m} \\ \vdots & \vdots & \vdots \\ u_{m۱} & \cdots & u_{mm} \end{bmatrix}$ و $V_{n\times n} = \begin{bmatrix} v_{۱۱} & \dots & v_{۱n} \\ \vdots & \vdots & \vdots \\ u_{n۱} & \cdots & u_{nn} \end{bmatrix}$ ماتریس‌های متعامد هستند. ستون‌های ماتریس $U_{m\times m}$ از بردارهای ویژه یکامتعامد [۳] ماتریس AA^T و ستون‌های ماتریس $V_{n\times n}$ از بردارهای ویژه‌ی یکامتعامد ماتریس A^TA تشکیل می‌شوند و $\Sigma_{m\times n}$ یک ماتریس قطری است که عناصر روی قطر آن مقادیر منفرد غیر صفر ماتریس AA^T و A^TA می‌باشند. یعنی

$$\begin{aligned} &\Sigma_{m\times n} = diag(\sigma_۱, ..., \sigma_p) \qquad p = \min\{m, n\} \\ &\sigma_۱ \geq \sigma_۲ \geq ... \geq \sigma_k > ۰ \qquad \sigma_{k+۱} = ... = \sigma_p = ۰. \end{aligned} \tag{۲.آ}$$

در اینجا $\sigma_۱$ و σ_k به ترتیب بزرگترین و کوچکترین مقادیر منفرد غیر صفر ماتریس A هستند.

مثال آ.۱.۲. ماتریس A داده‌شده را توسط مقادیر منفرد تجزیه می‌کنیم.

$$A_{۲\times ۳} = \begin{bmatrix} ۳ & ۱ & ۱ \\ -۱ & ۳ & ۱ \end{bmatrix}$$

[۳] Orthonormal

باید ماتریس $A_{۲\times۳}$ را به‌صورت $A = U\Sigma V^T$ تجزیه کنیم. برای بدست آوردن ماتریس $U_{۲\times۲}$ باید بردارهای ویژه یکامتعامد ماتریس AA^T را بیابیم.

$$A^T = \begin{bmatrix} ۳ & -۱ \\ ۱ & ۳ \\ ۱ & ۱ \end{bmatrix} \rightarrow AA^T = \begin{bmatrix} ۳ & ۱ & ۱ \\ -۱ & ۳ & ۱ \end{bmatrix} \begin{bmatrix} ۳ & -۱ \\ ۱ & ۳ \\ ۱ & ۱ \end{bmatrix} = \begin{bmatrix} ۱۱ & ۱ \\ ۱ & ۱۱ \end{bmatrix}$$

ابتدا مقادیر ویژه ماتریس AA^T را بدست می‌آوریم. داریم

$$|\lambda I_۲ - AA^T| = \begin{vmatrix} \lambda - ۱۱ & -۱ \\ -۱ & \lambda - ۱۱ \end{vmatrix} = (\lambda - ۱۲)(\lambda - ۱۰)$$

لذا مقادیر ویژه ماتریس AA^T عبارتند از

$$\lambda_۱ = ۱۲ \qquad \lambda_۲ = ۱۰.$$

ماتریس AA^T دو مقدار ویژه حقیقی و متمایز دارد. حال بردارهای ویژه متناظر با هر یک از آن‌ها را بدست می‌آوریم:

$$(AA^T - \lambda_۱ I_۲)u_۱ = ۰ \rightarrow \begin{bmatrix} -۱ & ۱ \\ ۱ & -۱ \end{bmatrix} \begin{bmatrix} u_{۱۱} \\ u_{۲۱} \end{bmatrix} = ۰ \rightarrow u_۱ = \begin{bmatrix} ۱ \\ ۱ \end{bmatrix}$$

$$(AA^T - \lambda_۲ I_۲)u_۲ = ۰ \rightarrow \begin{bmatrix} ۱ & ۱ \\ ۱ & ۱ \end{bmatrix} \begin{bmatrix} u_{۱۲} \\ u_{۲۲} \end{bmatrix} = ۰ \rightarrow u_۲ = \begin{bmatrix} ۱ \\ -۱ \end{bmatrix}$$

$$\begin{bmatrix} u_۱ & u_۲ \end{bmatrix} = \begin{bmatrix} ۱ & ۱ \\ ۱ & -۱ \end{bmatrix}$$

می‌توان با اعمال فرآیند گرام‌ـاشمیت [۴] این دو بردار را بصورت یکامتعامد تبدیل کرد. لذا ماتریس $U_{۲\times۲}$ به‌شکل زیر به‌دست می‌آید:

$$U = \begin{bmatrix} \frac{۱}{\sqrt{۲}} & \frac{۱}{\sqrt{۲}} \\ \frac{۱}{\sqrt{۲}} & \frac{-۱}{\sqrt{۲}} \end{bmatrix}.$$

ماتریس $V_{۳\times۳}$ نیز باید از بردارهای ویژه یکامتعامد ماتریس A^TA به‌دست آید. مقادیر ویژه ماتریس A^TA عبارتند از

$$\lambda_۱ = ۱۲, \qquad \lambda_۲ = ۱۰, \qquad \lambda_۳ = ۰.$$

[۴] Gram–Schmidt process

ماتریس A^TA سه مقدار ویژه حقیقی و متمایز دارد. حال بردارهای ویژه متناظر با هر یک از آن‌ها را به‌دست می‌آوریم:

$$(A^TA-\lambda_۱I_۳)v_۱=۰\rightarrow\begin{bmatrix}-۲ & ۰ & ۲\\ ۰ & -۲ & ۴\\ ۲ & ۴ & -۱۰\end{bmatrix}\begin{bmatrix}v_{۱۱}\\ v_{۲۱}\\ v_{۳۱}\end{bmatrix}=۰\rightarrow v_۱=\begin{bmatrix}۱\\ ۲\\ ۱\end{bmatrix}$$

$$(A^TA-\lambda_۲I_۳)v_۲=۰\rightarrow\begin{bmatrix}۰ & ۰ & ۲\\ ۰ & ۰ & ۴\\ ۲ & ۴ & -۸\end{bmatrix}\begin{bmatrix}v_{۱۲}\\ v_{۲۲}\\ v_{۳۲}\end{bmatrix}=۰\rightarrow v_۲=\begin{bmatrix}۱\\ ۲\\ ۱\end{bmatrix}$$

$$(A^TA-\lambda_۳I_۳)v_۳=۰\rightarrow\begin{bmatrix}۱۰ & ۰ & ۲\\ ۰ & ۱۰ & ۴\\ ۲ & ۴ & ۲\end{bmatrix}\begin{bmatrix}v_{۱۳}\\ v_{۲۳}\\ v_{۳۳}\end{bmatrix}=۰\rightarrow v_۳=\begin{bmatrix}۱\\ ۲\\ -۵\end{bmatrix}$$

$$\begin{bmatrix}v_۱ & v_۲ & v_۳\end{bmatrix}=\begin{bmatrix}۱ & ۲ & ۱\\ ۲ & -۱ & ۲\\ ۱ & ۰ & -۵\end{bmatrix}$$

با اعمال فرآیند یکامتعامدسازی گرام‌ـاشمیت بردارهای ویژه یکامتعامد را می‌توان به‌دست آورد. لذا ماتریس $V_{۳\times۳}$ به‌صورت زیر به‌دست می‌آید:

$$V=\begin{bmatrix}\frac{۱}{\sqrt{۶}} & \frac{۲}{\sqrt{۵}} & \frac{۱}{\sqrt{۳۰}}\\ \frac{۲}{\sqrt{۶}} & \frac{-۱}{\sqrt{۵}} & \frac{۲}{\sqrt{۳۰}}\\ \frac{۱}{\sqrt{۶}} & ۰ & \frac{-۵}{\sqrt{۳۰}}\end{bmatrix}.$$

نهایتا ماتریس $\Sigma_{۲\times۳}$ با استفاده از مقادیر منفرد محاسبه شده به‌صورت زیر به‌دست می‌آید:

$$\Sigma_{۲\times۳}=\begin{bmatrix}\sigma_۱ & ۰ & ۰\\ ۰ & \sigma_۲ & ۰\end{bmatrix}=\begin{bmatrix}\sqrt{۱۲} & ۰ & ۰\\ ۰ & \sqrt{۱۰} & ۰\end{bmatrix}$$

بنابراین، تجزیه به مقادیر منفرد ماتریس A به‌شکل زیر به‌دست می‌آید:

$$A=U\Sigma V^T=\begin{bmatrix}\frac{۱}{\sqrt{۲}} & \frac{۱}{\sqrt{۲}}\\ \frac{۱}{\sqrt{۲}} & \frac{-۱}{\sqrt{۲}}\end{bmatrix}\begin{bmatrix}\sqrt{۱۲} & ۰ & ۰\\ ۰ & \sqrt{۱۰} & ۰\end{bmatrix}\begin{bmatrix}\frac{۱}{\sqrt{۶}} & \frac{۲}{\sqrt{۶}} & \frac{۱}{\sqrt{۶}}\\ \frac{۲}{\sqrt{۵}} & \frac{-۱}{\sqrt{۵}} & ۰\\ \frac{۱}{\sqrt{۳۰}} & \frac{۲}{\sqrt{۳۰}} & \frac{-۵}{\sqrt{۳۰}}\end{bmatrix}.$$

مراجع

[۱] محمدی تاکامی م، (۱۳۸۴)، **"روش‌های پیش‌پردازش داده‌ها و تشخیص الگو"** جلد اول، چاپ دوم، انتشارات دانشگاه خواجه نصرالدین طوسی.

[۲] فرهادی ز، (۱۳۹۳)، **"رده‌بندی با استفاده از جنگل‌های تصادفی"** پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود.

[3] Caliński, T., Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, **3(1)**, 1-27.

[4] Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, **53(4)**, 406-413.

[5] Duda, R. O., Hart, P. E., Stork, D. G. (2012). Pattern classification. *John Wiley and Sons*, New York, Second edition.

[6] Hastie, T., Buja, A., Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, **23(1)**, 73-102.

[7] Hartigan, J. A., Hartigan, J. A. (1975). Clustering Algorithms. *Wiley:New York*,Vol.209.

[8] Hastie, T., Tibshirani, R., Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, **89(428)**, 1255-1270.

[9] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24(6)**, 417.

[10] Hoerl, A., Kennard, R. (1988). Ridge Regression. *in Encyclopedia of Statistical Sciences*. Vol.8.

[11] Krzanowski, W. J., Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*,**44(1)** 23-34.

[12] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, **28(2)**, 129-137.

[13] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2(11)**, 559-572.

[14] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, **66(336)**, 846-850.

[15] Rocci, R. A. (2011). A new dimension reduction method: factor discriminant K-means. *Journal of Classification*,**28(2)** 210–226.

[16] Reizer, G. V. (2011). **Stability Selection of the Number of Clusters**.Thesis Georgia State University.

[17] Sun, L. a. (2008). A least squares formulation for canonical correlation analysis.*Proceedings of the 25th international conference on Machine learning*, 1024–1031

[18] Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63(2)**,411-423.

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), 267-288.

[20] Wang, Y. a. (2016). Sparse optimal discriminant clustering.*Statistics and Computing*, **26(3)**, 629–639.

[21] Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, **97(4)**, 893-904.

[22] Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68(1)**, 49-67.

[23] Zhang, Z. a. (2009). Optimal scoring for unsupervised learning. *Advances in neural information processing systems*. pp, 2241-2249.

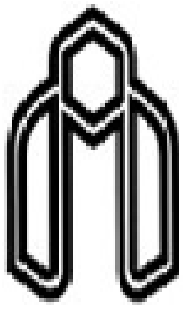
[24] Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, **15(2)**, 265-286.

[25] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology), **67(2)**, 301-320.

Aabstract

Clustering is one of the important data mining issues for discovering hidden patterns in data. If the number of variables are too much and we encounter with high dimensional data , the problem of the dimension reduction is paralled to clustering or in the same direction of it. One of the most commonly methods of dimension reduction, which is used in both monitoring and uncontrolled learning topics, is the principal component method that has its own merits and disadvantages. In this thesis we trying by introducing optimal discriminant clustering (ODC) method, that is used to reduce the dimension, describe the clustering problem, which is an uncontrolled learning problem, as a problem with ridge regression, so that, like the thought of the principal components, one can extrapolate another kind of linear combination of initial variables to construct new variables, and then use one of the algorithms k-means clustering for new converted observations. Since the ridge regression problem plays prominent a role in the Tuning parameter, In this case, an tuning parameter will play a significant role in clustering performance. Also, the existence of some non-essential variables in the model leads to negativ and weak performing of clustering method, so by adding the group's fine fractional function, we will eliminate this weakness and introduce a new clustering, that is the revised version of the optimal discriminant clustering. The results of the simulation indicate the effectiveness of this method in dealing with the high dimensions of the variables as well as its superiority to the principal components analysis method.

key words: Optimal discriminant clustering، Sparse optimal discriminant clustering، High dimension data، Selection variable، Cross validation.



Shahrood University of Technology

Faculty of Mathematical Sciences

M.Sc. Thesis in Statistics

Dimension Reduction in Clustering by Gorup-Lasso Penalty Function

By: zinat salimi sani

Supervisor

Dr. Davood Shahsavani

January 2018