چهارمین سمینار آمار فضایی و کاربردهای آن
4th Seminar on Spatial Statistics and Its Applications
۵–۶ آبان ماه ۱۴۰۰
27-28 Oct. 2021

خوشه‌بندی راهنماییده مجموعه داده‌های تصویر

ساجده مرادنیا و موسی گلعلی‌زاده
گروه آمار، دانشگاه تربیت مدرس

چکیده: با توجه به پیشرفت تکنولوژی در طی سالیان گذشته، حجم پایگاه‌های داده که حاوی اطلاعات بسیاری از قبیل داده‌های متن، تصویر و ویدئو هستند، به طور چشم‌گیری افزایش یافته است. از بین آن‌ها داده‌های تصویر را می‌توان در زمره داده‌های مشبکه‌ای فضایی دانست که تصویر در واقع ماتریسی دو بعدی از مقادیر عناصر تصویر (پیکسل‌ها) است که موقعیت قرارگیری پیکسل‌ها دارای همبستگی فضایی هستند. از طرفی دیگر، خوشه‌بندی یکی از روش‌های چند متغیره آماری برای واکاوی داده‌های بعد بالا است و در بسیاری از زمینه‌ها از جمله تجزیه و تحلیل تصویر نقش بارزی را ایفا می‌کند. یافتن عناصر مؤثر تصویر، نقش بسیار مهمی را در کاهش زمان و افزایش کارآمدی خوشه‌بندی ایفا می‌کند. خوشه‌بندی راهنماییده رویکردی نوین است که با استفاده از اطلاعات متغیر پاسخ، تنها متغیرهایی را که دارای اطلاعات مفیدی در خصوص تشخیص و پیش‌بینی متغیر پاسخ هستند، شناسایی می‌کند. مقاله حاضر نحوه اجرای خوشه‌بندی راهنماییده داده‌های تصویر دست نوشته الکترونیکی را تشریح نموده و عملکرد موفق آن را نشان خواهد داد.

واژه‌های کلیدی: آمار فضایی، خوشه‌بندی، داده‌های بعد بالا، ناحیه بندی تصویر، پردازش تصویر، خوشه‌بندی راهنماییده.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H30، 62H35.

۱ مقدمه

در به کارگیری اکثر روش‌های معمولی آمار، فرض بر این است که مشاهدات تحت شرایط یکسان و به صورت مستقل از هم جمع‌آوری شده‌اند. اگر چه، این فرض کمک شایانی به تسهیل مبانی نظری مدل‌بندی آماری مد نظر محقق می‌کند، اما ممکن است در عمل محقق را از کشف واقعیت‌های دیگر حول و حوش داده‌های مورد مطالعه دور کرده و موجب از بین رفتن اطلاعات زیادی شود. موارد زیادی از مثال‌های واقعی از حوزه‌های مختلف وجود دارند که در آن مشاهدات مستقل نبوده و بر حسب موقعیت قرار گرفتن خود در فضای مورد بررسی به یکدیگر وابسته هستند. اگر این وابستگی تابعی از فاصله

* نام وآدرس پست الکترونیکی ارائه‌دهنده مقاله: ساجده مرادنیا، Sajedeh.moradnia@modares.ac.ir

بین موقعیت‌های مشاهدات باشد، به گونه‌ای که مشاهدات نزدیک به هم وابسته‌تر و مشاهدات دورتر از هم وابستگی کم‌تری داشته باشند، مشاهدات مورد نظر را داده‌های فضایی و علمی که به شناسایی و تجزیه داده‌های فضایی می‌پردازد را آمار فضایی[1] می‌نامند (محمدزاده ، ۱۳۹۸).

از طرفی دیگر، کاربردهای متنوع پردازش تصویر در زمینه‌های مختلف باعث شده که تحلیل داده‌های مبتنی بر تصویر به یکی از موضوعات مورد علاقه محققین در دنیای تحلیل داده‌ها تبدیل شود (لی ، ۱۹۸۰). در این بین، پردازش تصویر[2] روشی است که در آن یک تصویر به عنوان ورودی دریافت شده و سپس با انجام عملیات‌های جبری بر روی آن، خروجی مشخص و معینی از تصویر ارائه می‌شود. این خروجی‌ها می‌تواند نمایش، چاپ، ویرایش و بهبود تصویر، کشف و تشخیص یک ویژگی خاص در تصویر و یا فشرده‌سازی تصویر باشد (پیتاس و ونتسانوپولوس ، ۱۹۹۲). نکته حائز اهمیت این است که هر تصویر را می‌توان به صورت یک ماتریس دو بعدی از مقادیر عناصر تصویر (پیکسل‌ها) در نظر گرفت که در سطرها و ستون‌ها مرتب شده‌اند و هر عنصر تصویر شامل مقادیر ترکیب‌شده سه رنگ معروف به " قرمز–سبز–آبی " که به اختصار به صورت RGB نوشته می‌شود، است. به علاوه، بسته به منظم یا نامنظم بودن مکان پیکسل‌ها می‌توان تصویر را به صورت مشبکه منظم (کرسی ، ۱۹۹۳) در R^2 در نظر گرفت.

بنا به آیزنمن (۲۰۰۸)، خوشه‌بندی، روشی ناراهنماییده[3] است که به شناسایی گروه‌های مشابه از داده‌ها، بدون هیچ اطلاع قبلی از گروه‌های مربوطه می‌پردازد. از نگاهی دیگر، خوشه‌بندی را می‌توان هم‌ارز ناحیه‌بندی تصویر[4] دانست که در آن، طی فرآیندی مشخص یک تصویر به چند بخش (مجموعه از پیکسل‌ها) قطعه‌بندی می‌شود. هدف از ناحیه‌بندی تصویر، ساده‌سازی و تغییر در نمایش تصویر است به گونه‌ای که تحلیل آن آسان‌تر شود. خروجی فرایند ناحیه‌بندی تصویر، مجموعه ای از بخش‌هاست که اجتماع آن‌ها، تشکیل‌دهنده کل تصویر است (کرمرز و همکاران ، ۲۰۰۳). اگرچه تحلیل خوشه‌بندی متوجه گروه‌بندی مشاهدات است، اما موارد بیشماری وجود دارد که محققین سعی در خوشه‌بندی متغیرها هم دارند. بنا به فریمن و همکاران (۲۰۰۸)، خوشه‌بندی متغیرها روشی برای مرتب‌سازی آن‌ها در خوشه‌های همگن است به طوری که متغیرهای موجود در هر خوشه، به شدت با یکدیگر همبسته بوده و دارای اطلاعات مشابهی باشند. موضوع تحقیق جذاب آن است که محقق قصد داشته باشد خوشه‌بندی را با نگاه بین متغیرهای تبیینی و متغیر پاسخ انجام دهد. در این راستا و بر اساس تحقیقات دتلینگ و بوهلمن (۲۰۰۲)، رویکرد نوین خوشه‌بندی راهنماییده وارد دنیای آمار شد. بنا به این رویکرد، از یک طرف متغیر پاسخ (Y) به‌طور مستقیم در فرآیند خوشه‌بندی مشارکت داده می‌شود و از طرفی دیگر تنها متغیرهایی که دارای اطلاعات مفیدی در ارتباط با متغیر پاسخ هستند، شناسایی می‌شوند. نظر به این که داده‌های تصویر مثالی از داده‌های بعد بالا هستند، اگر عناصر تشکیل‌دهنده هر تصویر (پیکسل‌ها) را به عنوان متغیرهای آن مجموعه داده در نظر بگیریم، پیاده‌سازی خوشه‌بندی راهنماییده بر روی چنین مجموعه داده‌هایی روشی نوین است که می‌تواند در مقایسه با سایر روش‌ها دارای جذابیت علمی خاصی باشد. در این مقاله، رویکرد مورد اشاره در تحلیل دست‌نوشته‌های الکترونیکی[5] که از ویلکینسون (۲۰۱۸) اخذ شده است، به کار گرفته می‌شود. نحوه عملکرد آن با روش‌های مرسوم مقایسه می‌شود و در مورد محاسن و معایب آن نیز بحث می‌شود.

در ادامه مقاله و در بخش ۲ جزئیات فنی رویکرد خوشه‌بندی راهنماییده می‌آید. تحلیل مثال واقعی موضوع بخش ۳ است. جمع‌بندی و نتیجه‌گیری پایان‌بخش این مقاله است.

[1] Spatial statistics
[2] Image processing
[3] Un-supervised clustering
[4] Image segmentation
[5] Handwritten digit images

۲ خوشه‌بندی راهنماییده

خوشه‌بندی راهنماییده یکی از روش‌های خوشه‌بندی است که با استفاده از اطلاعات متغیر پاسخ، به گروه‌بندی متغیرها می‌پردازد. دتلینگ و بوهلمن (۲۰۰۲) این نوع خوشه‌بندی را در مطالعه داده‌های ژنی برای تعیین نوع بافت سرطانی معرفی کردند. پیش از این، در مواجهه با داده‌های بعد بالا و برای گروه‌بندی آن‌ها، به طور گسترده‌ای از خوشه‌بندی ناراهنماییده از جمله خوشه‌بندی سلسله مراتبی استفاده می‌شد، که واضح است نمایش هزاران خوشه در ساختار درختی سبب دشواری‌های عدیده‌ای در تحلیل خوشه‌ها می‌شد (وینستین و همکاران ، ۱۹۹۷). پیشینه روش خوشه‌بندی راهنماییده به فعالیت هستی و همکاران (۲۰۰۱) باز می‌گردد که در آن، طی روشی دو مرحله‌ای، ابتدا خوشه‌های نامزد بسیاری با استفاده از خوشه‌بندی ناراهنماییده سلسله مراتبی تولید می‌شدند.

از آن جایی که داده‌های تصویر از جمله داده‌های بعد بالایی هستند که براساس میزان همبستگی عناصر تصویر ناحیه‌بندی می‌شوند، انتظار می‌رود اعمال خوشه‌بندی راهنماییده بر روی آن‌ها بتواند عملکرد خوبی در مقایسه با روش‌های رقیب داشته باشد. برای فهم رویکرد خوشه‌بندی راهنماییده در ادامه، روشی که پیاده‌سازی آن را ممکن می‌کند تشریح می‌شود.

مدل تصادفی پایه برای داده‌های بعد بالا با متغیر پاسخ رسته‌ای، به صورت زوج‌های تصادفی $(\mathbf{X}, Y)$ با مقادیری از فضای ضربی $\mathbb{R}^p \times \mathbb{Y}$ فرض می‌شوند به طوری که $\mathbf{X} \in \mathbb{R}^p$ ماتریس داده‌ای است که با میانگین صفر و واریانس یک استاندارد شده‌اند. مجموعه $\mathbb{Y}$ برای متغیر پاسخ، مقادیر عددی خود را از مجموعه $\{۰, ۱, \dots K-۱\}$ می‌گیرد که K نمایانگر تعداد حالات متغیر پاسخ است. برای سادگی بحث و فهم آسان مطالب پیش رو، K برابر ۲ در نظر گرفته می‌شود. فرض کنید، تنها تعدادی از متغیرهای تبیینی تعیین‌کننده ارتباط با متغیر پاسخ هستند. آنگاه می‌توان احتمال شرطی

$$Pr\left(Y = ۱ \middle| \mathbf{X}\right) = f(\tilde{\mathbf{X}}) = f(\tilde{X}_{C_۱}, \tilde{X}_{C_۲}, \dots, \tilde{X}_{C_q}) \qquad (۱.۲)$$

را تعریف کرد که در آن $f(.)$ تابعی غیرخطی از $\mathbb{R}^q$ به $[۰, ۱]$ است. مجموعه $\{C_۱, \dots C_q\}$ جایی‌که $q \ll p$ خوشه‌هایی از متغیرها هستند به قسمی که $\{\cup_{i=۱}^{q} C_i\} \subset \{۱, \dots, p\}$ و به ازای $i \neq j$، $C_i \cap C_j = ۰$. به علاوه فرض می‌شود $\tilde{X}_{C_i} \in \mathbb{R}$ نشان‌دهنده نماینده هر خوشه C_i است. از آن جایی که محقق به دنبال خوشه‌هایی با متغیرهای مشابه است، ترکیب خطی ساده

$$\tilde{X}_{C_i} = \frac{۱}{|C_i|} \sum_{g \in C_i} \alpha_g X_g \qquad (۲.۲)$$

که در آن $|C_i|$ تعداد متغیرهای موجود در خوشه C_i است و $\alpha_g \in \{-۱, ۱\}$ بهترین انتخاب برای $\tilde{X}_{C_i}$ است. از آن جا که علامت نقش تعیین‌کننده‌ای در تحلیل بازی نمی‌کند، در محاسبه (۲.۲) امکان مشارکت متغیر خاص gـام توسط X_g و همچنین $-X_g$ فراهم شده است. در نتیجه، با هر متغیر به طور متقارن رفتار می‌شود. با این حال، یافتن زیرمجموعه‌ای از p متغیر و تشکیل خوشه‌های $\{C_۱, \dots, C_q\}$ با ساختار احتمالاتی تا حدی دشوار است. بنابراین، ارائه روشی محاسباتی که به صورت تقریبی مدل‌بندی ارائه‌شده در (۱.۲) را تسهیل کرده و از نظر تجربی نتایج خوبی را به همراه داشته باشد، ضروری بنظر می‌رسد.

در الگوریتم خوشه‌بندی راهنماییده، متغیرها یکی پس از دیگری به خوشه اضافه شده و در هر مرحله، متغیرهایی که به غلط به خوشه اضافه شده‌اند، حذف می‌شوند. این عمل تا زمان تثبیت خوشه ادامه می‌یابد و سپس تشکیل خوشه جدید آغاز می‌شود. برای فهم آسان‌تر الگوریتم خوشه‌بندی راهنماییده، فرض کنید $(\mathbf{x}_۱, y_۱), \dots, (\mathbf{x}_n, y_n)$ طوری که $\mathbf{x}_j \in \mathbb{R}^p$ و $y_j \in \{۰, ۱\}$، n تحقق مستقل و هم‌توزیع از بردار تصادفی $(\mathbf{X}, Y)$ باشند و مقادیر $\mathbf{x}_j$ با میانگین صفر و واریانس یک استاندارد شده‌اند. طبیعی است که برای خوشه‌بندی درست و دقیق، لازم است تابع هدف خاصی مد نظر قرار گیرد. تابع

هدف مورد اشاره باید کمی، کارآمد و دارای مقیاسی توانمند در جداسازی وضعیت متغیر پاسخ باشد. از آن جایی که هدف یافتن زیرمجموعه‌هایی از متغیرها با قدرت تفکیک دقیق در مسائل دوحالتی است، دتلینگ و بوهلمن (۲۰۰۲) آماره آزمون ویلکاکسون را که به عنوان تابع امتیاز رتبه محور ناپارامتری به کار برده می‌شود، برای انجام این امر پیشنهاد دادند و این معیار را ویلما[6] نامیدند. در نتیجه، امتیاز متغیر $i-$ام از بردار n بعدی مقادیر مشاهده‌شده متغیرها یعنی $\xi_i = (x_{i1}, \ldots, x_{in})$ به صورت

$$Score(\xi_i) = s(\xi_i) = \sum_{j \in N_0} \sum_{l \in N_1} \mathbb{1}_{[x_{ij} \geq x_{il}]} \tag{۳.۲}$$

قابل محاسبه است که در آن، x_{ij} مقدار اطلاعات متغیر $i-$ام در مورد پاسخ $j-$ام و N_k به ازای $k \in \{0, 1\}$ نشان‌دهنده زیر مجموعه‌ای از $\{1, \ldots, n\}$ است. با توجه به رابطه (۳.۲)، تابع امتیاز به طور غیرمستقیم از اطلاعات موجود در متغیر پاسخ استفاده می‌کند و بنابراین می‌تواند به عنوان معیاری برای خوشه‌بندی راهنماییده تلقی شود. محاسبه امتیاز برای خوشه C_i نیز از طریق مقادیر نمایندگان مشاهده‌شده $\xi_{C_i} = (x_{C_{i,1}}, \ldots, x_{Ci,n})$ قابل محاسبه است. با در نظر گرفتن تابع امتیاز به عنوان آماره آزمون ویلکاکسون، امکان ترتیب‌بندی متغیرها و خوشه‌ها براساس میزان معنی‌دار بودن‌شان برای جداسازی متغیر پاسخ وجود دارد. در حالتی که جداسازی دقیق اتفاق بیفتد، مقدار تابع امتیاز کمترین مقدار $(s_{\min} = 0)$ و در غیر این صورت، ماکزیمم مقدار $(s_{\max} = n_0 n_1)$ خود را خواهد گرفت. همان‌طور که اشاره شد، هدف آن است که تمام متغیرها با بار اطلاعاتی کم، به کلاس صفر اختصاص یابند و این هدف تنها با استفاده از تغییر علامت مقادیر $\tilde{\xi}_i$ برای تمام متغیرهای $i \in \{1, \ldots, p\}$ محقق می‌شود. با پیروی از دتلینگ و بوهلمن (۲۰۰۲)، با معرفی $\alpha_g \in \{-1, 1\}$ می‌توان اطلاعات متغیرهای تبیینی درباره متغیر پاسخ را به صورت رابطه شرطی

$$\tilde{\xi}_i = \alpha_i \xi_i = \begin{cases} \xi_i & s(\xi_i) \leq \frac{s_{\max}}{2} \\ -\xi_i & s(\xi_i) > \frac{s_{\max}}{2} \end{cases}$$

نوشت. پس از اعمال عملگر تغییر علامت برای متغیرهای واجد شرایط، امتیازات تمام متغیرها در ماتریس داده‌ها از طریق رابطه $s(\tilde{\xi}_i) = \min\{s(\xi_i), s_{\max} - s(\xi_i)\}$ قابل محاسبه خواهد بود و تمام متغیرها قطبیت مشابهی خواهند داشت. حال می‌توان با خیال آسوده، با میانگین‌گیری، مقادیر اطلاعاتی موجود در متغیرهای تبیینی درباره متغیر پاسخ را محاسبه کرد. اگرچه تابع امتیاز تغییریافته دارای قابلیت‌های فراوانی است، اما در برخی مواقع عملکرد خوبی ندارد. به عنوان مثال، مواردی وجود دارد که به دلیل گسسته بودن دامنه تابع امتیاز، امتیاز تعدادی از متغیرهای پیشگو در طول فرآیند خوشه‌بندی تقریبا برابر (صفر) می‌شود. لذا، ضروری است تابع امتیاز مد نظر به طریقی اصلاح شود. برای این منظور، دتلینگ و بوهلمن (۲۰۰۲) تابع حاشیه

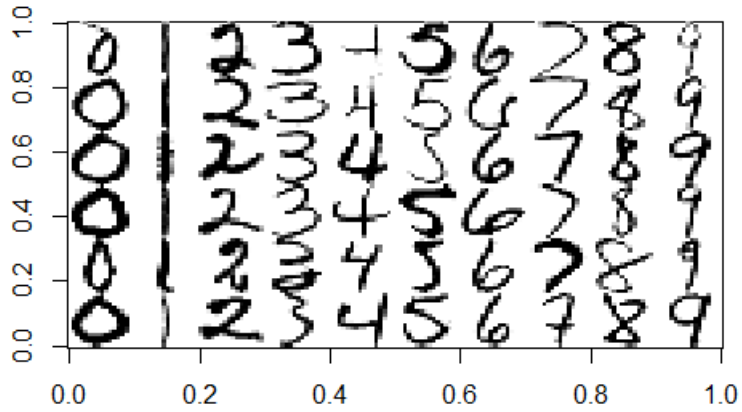
$$Margin(\xi_i) = m(\xi_i) = \min_{l \in N_1}(x_{il}) - \max_{j \in N_0}(x_{ij}) \tag{۴.۲}$$

را معرفی کردند که مقیاسی پیوسته برای تفکیک متغیر پاسخ است، که در آن N_0، N_1 و x_{ij} مشابه آن چیزی است که در رابطه (۳.۲) تعریف شدند. می‌توان ملاحظه کرد که مقدار تابع حاشیه مثبت خواهد بود اگر و تنها اگر تابع امتیاز صفر باشد و $\tilde{\xi}_i$ به طور کامل و به بهترین نحو، متغیر پاسخ را تفکیک نماید. در غیر این صورت، مقدار تابع حاشیه منفی خواهد بود. در ادامه نحوه پیاده‌سازی خوشه‌بندی راهنماییده تشریح‌شده در این بخش برای تحلیل دست‌نوشته‌های الکترونیکی افراد تشریح می‌شود.

[6] Wilma

۳ تحلیل مثال واقعی

مجموعه داده تصاویر دست‌نوشته‌های الکترونیکی شامل تصاویری از ارقام دست‌نویس افرادی است که از آن‌ها خواسته شده تا اعداد ۰ تا ۹ را در یک صفحه الکترونیکی به صورت لاتین ثبت کنند (ویلکینسون ، ۲۰۱۸). تعداد زیادی از نمونه‌های طبقه‌بندی‌شده دستی در مجموعه داده zip.train قابل دسترسی در بسته ElemStatLearn موجود است. نمونه‌ای از تصاویر در شکل ۱ آمده است. این مجموعه داده به صورت ماتریس ۷۲۹۱ × ۲۵۷ بعدی ذخیره شده است. نظر به این که هر سطر



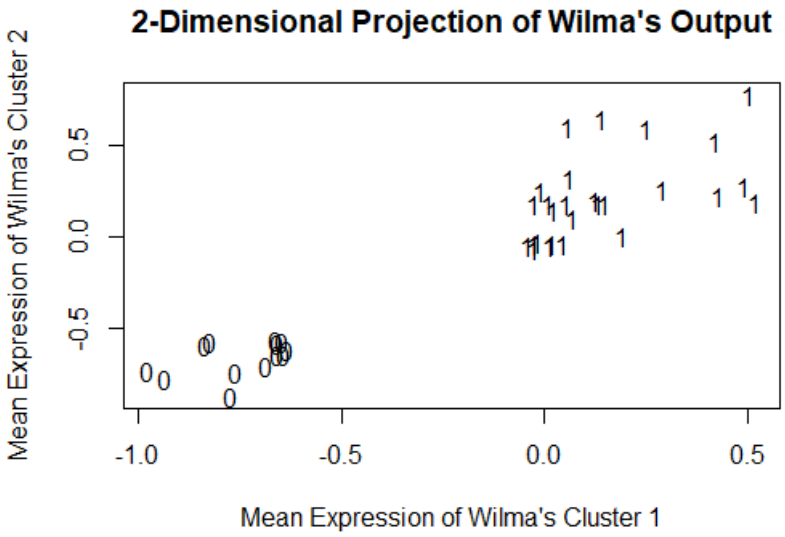
شکل ۱: نمونه‌ای از تصاویر مجموعه داده zip.train.

این ماتریس، بیانگر یک تصویر است، بنابراین ملاحظه می‌شود که $n = ۷۲۹۱$ و $p = ۲۵۷$. در حقیقت، هر تصویر، ماتریسی ۱۶ × ۱۶ از پیکسل‌ها با طیف رنگی خاکستری است. بنابراین هر تصویر دارای ۲۵۶ بُعد (پیکسل) است که می‌توان به صورت بردار ۲۵۶ بعدی در فضای $\mathbb{R}^{۲۵۶}$ در نظر گرفت. اولین عدد هر سطر نشان‌دهنده رقم هر تصویر یا به عبارتی متغیر پاسخ است. بنابراین، هر سطر این مجموعه داده از فضای $\mathbb{R}^{۲۵۶} \times \{۰, ۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸, ۹\}$ تشکیل شده است.

فرض کنید هدف پیاده‌سازی خوشه‌بندی راهنماییده بر روی مجموعه داده مدنظر است. ابتدا حالت دو کلاسه و تنها افرادی که دو عدد ۳ و ۸ را در صفحه الکترونیکی ثبت کرده بودند، مورد بررسی قرار می‌گیرد. تعداد این افراد ۱۲۰۰ نفر بود که با توجه به تعریف و ماهیت داده‌های بعد بالا، ۶۰ نفر از این مجموعه را انتخاب کرده و الگوریتم خوشه‌بندی راهنماییده ویلما بر روی آن‌ها اجرا شد. علی‌رغم این که دو عدد انتخابی لاتین، از نظر نوشتاری شباهت بسیاری با هم دارند، اما خوشه‌بندی راهنماییده با استفاده از اطلاعات متغیر پاسخ با دقت ۹۵ درصدی پیکسل‌ها را از نظر شباهت و همبستگی مکانی که با هم داشتند، در دو کلاس قرار داد که نشان از موفقیت اعمال این روش بر روی این مجموعه داده است. در شکل ۲ میانگین بار اطلاعاتی متغیرهای موجود در هر دو خوشه (برچسب‌های صفر و یک به ترتیب برای عدد ۳ و ۸) نشان داده شده است.

در این قسمت قصد داریم خوشه‌بندی راهنماییده را در حالت چند کلاسه پیاده نماییم. روش‌های متعددی برای تفکیک داده‌های چند کلاسه به منظور یافتن بهترین کلاس برای مشاهدات وجود دارد. در این مقاله، روش "یکی مقابل همه"[۷] را برای اعمال خوشه‌بندی راهنماییده بر روی داده‌های zip.train به کار گرفتیم. به جای گزارش دقت بالای (حدود ۹۶ درصد) نقش تاثیر متغیرهای موثر در خوشه‌بندی راهنماییده را برجسته می‌کنیم. برای این منظور، رقم ۹ را با در نظر گرفتن همه متغیرها (۲۵۶ بعد)، در نظر گرفتیم. سپس رقم ۹ را در مقابل سایر حالات متغیر پاسخ قرار دادیم و الگوریتم ویلما را بر روی آن اجرا کردیم. در این حالت و با در نظر گرفتن همه متغیرها، دقت خوشه‌بندی راهنماییده ۹۴ درصد به دست آمد.

[7] One-Against-All



شکل ۲: میانگین بار اطلاعاتی متغیرهای موجود در خوشه اول و دوم در روش ویلما.

همچنین متغیرهای (پیکسل‌های) ضروری را در سه خوشه قرار دادیم. این متغیرهای مهم در جدول ۱ آورده شده است.

جدول ۱: متغیرهای مهم رقم ۹ در مجموعه داده zip.train.

خوشه اول	خوشه دوم					خوشه سوم
۲۲۰	۲۵۶	۱۴۶	۲۱۰	۲۰۹	۱۶۲	۲۳۵
۱۵۸	۲۵۵	۱۴۵	۱۷۳	۱۴۳	۱۶۱	۱۱
۱۵۷	۲۵۲	۱۴۸	۲۳۷	۱۴۴	۱۶۰	۱۹
۱۸۲	۲۵۴	۱۴۹	۲۳۶	۱۴۲	۱۷	۲۱۹
۱۹۷	۲۴۴	۱۶۶	۲۰۴	۱۵۹	۹۶	۵۸
۱۰	۱	۱۶۵	۲۳۸	۱۶۴	۱۱۳	۴
۱۰۲	۲	۱۹۵	۲۳۹	۱۶۳	۱۱۲	۲۴۵
۱۵۲	۱۷۵	۱۹۳	۲۲۱	۱۷۴	۱۶	۲۱۳
۶۷	۱۷۶	۱۹۲	۲۴۰	۱۸۱	۱۸	
۹	۱۷۷	۱۹۱	۲۴۱	۱۸۰	۸۰	
۴۲	۱۷۸	۱۹۰	۲۰۷	۳	۳۳	
۶	۱۷۹	۱۸۹	۲۰۵	۵۷	۲۵۱	
۱۲۱	۲۲۸	۱۸۸	۲۰۶	۱۳۱	۲۵۳	
۲۴۹	۲۲۷	۲۱۱	۲۲۲	۱۳۰	۲۴۳	
۲۳۲	۲۲۶	۱۹۶	۲۲۳	۱۲۹	۲۴۲	
	۲۲۵	۲۱۲	۲۲۴	۱۲۸		
	۱۴۷	۱۹۴	۲۰۸	۱۲۷		

همان‌طور که در جدول ۱ قابل مشاهده است، از بین ۲۵۶ پیکسل رقم ۹ ، تنها ۱۰۶ عدد از آن‌ها به عنوان پیکسل‌های ضروری تشخیص داده شده‌اند. بنابراین، این‌بار خوشه‌بندی راهنماییده را با استفاده از این ۱۰۶ متغیر پیاده‌سازی کردیم و مجددا رقم ۹ را در مقابل سایر کلاس‌ها قرار دادیم. با توجه به اعمال کاهش بعد و تنها در نظر گرفتن پیکسل‌های ضروری، انتظار داشتیم تا دقت مدل کاهش یابد. اما با این وجود باز هم به دقت ۹۴ درصد رسیدیم که این نتیجه حاکی از قدرتمند بودن الگوریتم ویلما است. نکته قابل توجه دیگر این است که اگر پیکسل‌های ضروری موجود در هر خوشه در جدول ۱ را به ترتیب صعودی مرتب کنیم، خواهیم دید که متغیرهای ضروری در همسایگی یکدیگر و به عبارتی دارای وابستگی مکانی با هم هستند. به عبارتی دیگر می‌توان گفت پیکسل‌های ضروری دارای برچسب یکسان، دارای ویژگی‌های مشابهی مانند شدت روشنایی،بافت و یا رنگ یکسانی هستند.

بحث و نتیجه‌گیری

روش‌های خوشه‌بندی ناراهنماییده از هیچ‌گونه اطلاعات اضافی برای خوشه‌بندی داده‌ها جز ماهیت و ساختار متغیرهای پیشگو استفاده نمی‌کنند. بنظر می‌رسد دیدگاه دخالت دادن متغیر پاسخ تا حد زیادی به خوشه‌بندی بهتر مجموعه داده کمک کند. هدف اصلی مقاله حاضر، تشریح خوشه‌بندی راهنماییده و پیاده‌سازی آن در یک مثال واقعی بود. نشان داده شد که در نظر گرفتن وجود ارتباط میان متغیرهای پیشگو و پاسخ، جایی که قرار گرفتن متغیرها در خوشه‌ها بی‌تأثیر از این ارتباط نیست، منجر به نتایج رضایت‌بخشی از تحلیل خوشه‌ای می‌شود. همچنین به بررسی خوشه‌بندی راهنماییده وقتی که تعداد خوشه‌ها بیش از دو باشد، پرداخته شد. در نظر گرفتن متغیرهای ضروری و اعمال کاهش بعد بر روی مجموعه داده، دقت پیش‌بینی برچسب کلاس مشاهدات را کاهش نداد که مسئله‌ای حائز اهمیت در دنیای داده‌های بعد بالا و به ویژه تحلیل تصاویر است.

مراجع

محمدزاده، م.، (۱۳۹۸)، *آمار فضایی و کاربردهای آن*، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،

Cremers, D., Kohlberger, T., and Schnörr, C. (2003), Shape Statistics in Kernel Space for Variational Image Segmentation, *Pattern Recognition*, **36**, 1929-1943.

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley, New York.

Dettling, M., and Bühlmann, P. (2002), Supervised Clustering of Genes, *Genome Biology*, **3**, 0069.1–0069.15.

Fraiman, R., Justel, A., and Svarc, M. (2008), Selection of Variables for Cluster Analysis and Classification Rules, *Journal of the American Statistical Association*, **103**, 1294 -1303.

Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. (2001), Supervised Harvesting of Expression Trees, *Genome Biology*, **2**, 1-12.

Izenman, A. J. (2008), *Modern Multivariate Statistical Techniques, Regression, Classification and Manifold Learning*, Springer, New York.

Lee, J. S. (1980), Digital Image Enhancement and Noise Filtering by Use of Local Statistics, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**, 165-168.

Pitas, I., and Venetsanopoulos, A. N. (1992), Order Statistics in Digital Image Processing, *Proceedings of the IEEE*, **80**, 1893-1921.

Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J., Kohn, K. W., Fojo, T., Bates, S. E. Rubinstein, L. V., Anderson, N. L., and Buolamwini, J. K. (1997) An Information-Intensive Approach to the Molecular Pharmacology of Cancer, *Science*, **275**, 343-349.

Wilkinson, D. J., (2018), *Stochastic Modelling for Systems Biology*, Chapman and Hall/CRC.