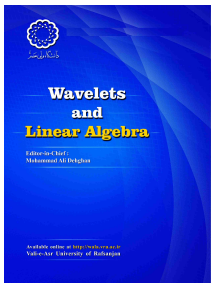
موجک‌ها و جبرخطی ۱۲(۱) (۱۴۰۴) ۱-۱۷



موجک‌ها و جبرخطی

http://wala.vru.ac.ir

دانشگاه ولیعصر (عج)
رفسنجان

خوشه‌بندی همزمان دیدگاه‌های مختلف در بررسی داده‌های مالیاتی

ام‌حکیمه اربابی[آ]، مینا جمشیدی*[آ]، محمدعلی عبدالرزاقی[ب]

[آ]گروه ریاضی کاربردی، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته، کرمان ، ایران
[ب]دانشجوی دکتری حسابداری، کارشناس ارشد اداره امورمالیاتی استان کرمان، کرمان، ایران

اطلاعات مقاله

تاریخچه مقاله:
دریافت شده: ۲۷ بهمن ۱۴۰۲
پذیرفته شده: ۳۱ فروردین ۱۴۰۴
دسترسی آنلاین:۳۱ فروردین ۱۴۰۴

کلمات کلیدی:
خوشه‌بندی، انتخاب ویژگی، داده مالیاتی.

چکیده

استفاده از روشهای داده کاوی در تحلیل داده‌های مالیاتی از جمله تقلب و یا فرار مالیاتی از راهکارهای جدید و مورد توجه می‌باشد. در روشهای ارائه شده، اغلب از الگوریتمهای بانظارت استفاده شده است که ممکن است برای تحلیل داده‌های زیاد هزینه‌بر باشد. همچنین تحلیل داده‌های مالیاتی کمتر از دیدگاه انتخاب ویژگی و یا خوشه‌بندی طیفی مورد بررسی قرار گرفته‌است. در این مقاله ابتدا مفاهیم خوشه‌بندی طیفی، خوشه‌بندی kـمیانگین و انتخاب ویژگی را مورد مطالعه قرار می‌گیرد و سپس روش جدیدی بر اساس در نظر گرفتن حالتهای متفاوتی از انتخاب ویژگی معرفی می‌گردد. در این روش ماتریس مشخصه خوشه‌بندی با در نظر گرفتن همزمان همه ماتریسهای مشخصه حاصل از انتخاب تعداد متفاوت ویژگیها، به همراه ضریب وزن تاثیر هر حالت، به دست می‌آید. سپس روش های ذکر شده روی داده‌های مالیاتی مورد بررسی قرار می‌گیرند و ویژگیهای مهم مرتبط با فرار مالیاتی با استفاده از انتخاب ویژگی به دست می‌آید. همچنین نتایج حاصل از انتخاب ویژگی و سه روش خوشه‌بندی kـمیانگین، خوشه‌بندی طیفی و روش خوشه‌بندی پیشنهادی جدید ارائه و مقایسه می‌گردد. ملاحظه می‌شود که بر اساس معیارهای ارزیابی خوشه‌بندی، این خوشه‌بندی ها نتایج نسبتا خوبی ارایه می‌نمایند و در نتیجه راهکاری مناسب برای تحلیل داده‌های مالیاتی می‌باشند و با استفاده از این روشها می‌توان فرار مالیاتی را، در داده‌هایی به صورت ارائه شده، مورد بررسی یا پیش بینی قرار داد.

*نویسنده مسئول
آدرس ایمیلها: omhakimeh75@gmail.com (ام‌حکیمه اربابی)، m.jamshidi@kgut.ac.ir (مینا جمشیدی)
http://doi.org/10.22072/wala.2025.2023159.1445

۱. مقدمه

یکی از مهم ترین منابع درآمدی دولت و کشورهای در حال توسعه مالیات محسوب می‌شود. مقدار قابل توجهی از درآمدهای مالیاتی دولت ها به دلیل استفاده کارشناسان مالیاتی از روشهای حسابرسی سنتی از دست می‌رود. از آنجا که نرخ تقلب و فرار مالیاتی رو به رشد است، نیازمند روشهایی می‌باشیم تا بتوانیم این موارد را تشخیص دهیم. با پیشرفت‌های به‌وجود آمده در جمع‌آوری داده‌ها و قابلیت‌های ذخیره‌سازی در طی دهه‌های اخیر مجموعه‌های داده‌ای با ابعاد بالا در علوم مختلف به‌سرعت در حال افزایش هستند. مسائل مربوط به امور مالیاتی نیز از این امر مستثنی نمی‌باشند و سیستم های سنتی پاسخگوی مدیریت این حجم اطلاعات نیستند. بنابراین در جهت مدیریت این حجم داده‌های بزرگ نیازمند به سیستم های هوشمند است که در این سیستم های هوشمند می‌توان از انواع الگوریتم های هوش مصنوعی و فرآیند داده کاوی در جهت کشف الگوی مورد نیاز و پنهان استفاده کرد. در واقع حجم بالای داده‌های ذخیره شده در سیستم های مالیاتی نیازمند به ابزاری است تا داده‌های ذخیره شده را پالایش و پردازش کرده و اطلاعات و دانش مورد نظر را استخراج کند و روابط موجود بین داده را بررسی و استخراج کند. بنابراین محققین به پیش بینی و بررسی عوامل مؤثر برشناسایی و تشخیص فرارها یا تقلبهای مالیاتی با استفاده از روشهای داده کاوی می‌پردازند. در این رأستا از روش های مختلفی از قبیل قوانین همبستگی، ، خوشه‌بندی، شبکه‌های عصبی، درخت های تصمیم، شبکه‌های بیزین، رگرسیون و روشهای ژنتیک استفاده شده است.

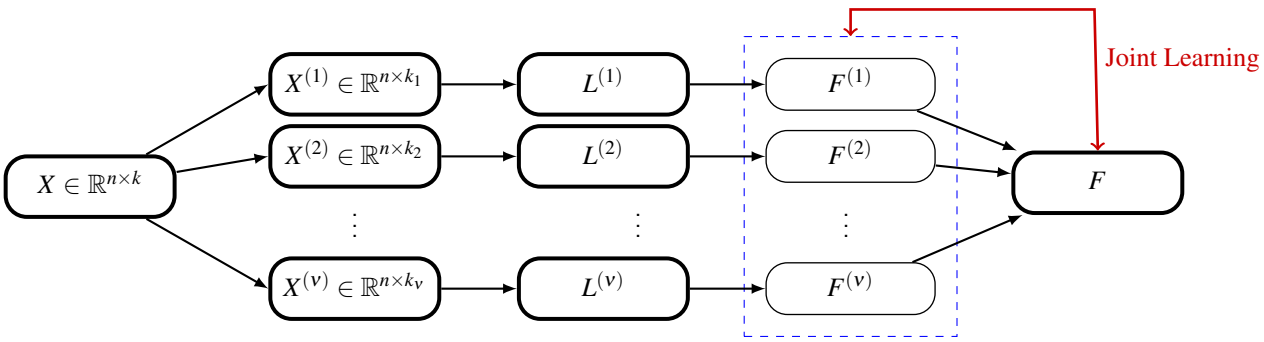
انتخاب ویژگی و خوشه‌بندی از جمله روشهایی می‌باشند که برای بررسی داده‌ها در داده‌کاوی به کار می‌رود. با اینکه این روشها می‌توانند برای داده‌های مالیاتی نیز مورد استفاده قرار گیرند، اما از روشهای انتخاب ویژگی کمتر در این امر استفاده شده‌است. در این مقاله سعی شده است که تحلیل داده‌های مالیاتی با ابزارهی خوشه‌بندی و انتخاب ویژگی به‌صورت همزمان انجام شود. برای این کار برخی از اطلاعات مؤدیان مورد بررسی قرار می‌گیرد و هر مؤدی به عنوان یك داده و اطلاعات مؤدی به عنوان ویژگیهای داده در نظر گرفته می‌شوند. انتخاب ویژگی، یکی از روشهای داده‌کاوی است که موجب صرفه‌جویی در ذخیره‌سازی و محاسبه زمان می‌شود. انتخاب ویژگی در مسائلی که تعداد نمونه‌ها نسبت به تعداد ویژگی‌ها بسیار کم می‌باشد مفید است. اما انتخاب ویژگی در برخی مسائل در تشخیص ویژگیهایی که تأثیر بیشتری در فرآیند داده‌کاوی دارند، نیز مؤثر می‌باشد و نشان می‌دهد که جمع آوری اطلاعات این ویژگیها یا تمرکز بر آنها ضروری می‌باشد. بنابراین در تشخیص برخی عوامل مؤثر در تشخیص تقلبهای مالیاتی می‌تواند مفید باشد. خوشه‌بندی نیز در رأستای تشخیص رفتار مؤدیان می‌تواند به کار گرفته شود.

الگوریتم‌های متفاوتی هم برای انتخاب ویژگی و هم برای خوشه‌بندی ارائه شده است. از متداولترین الگوریتمهای انتخاب ویژگی استفاده از امتیاز لاپلاسین می‌باشد که بر پایه گراف متناظر داده‌ها محاسبه می‌شود. همچنین روشهای دیگری در سالهای اخیر ارائه شده است که اهداف مختلفی از جمله نداشتن پارامتر، کارایی بر روی داده‌های بزرگ و ... را دنبال می‌نمایند [۱۶، ۱۸، ۲۸، ۲۹]. همچنین مشهورترین و پرکاربردترین روشهایی که برای خوشه‌بندی ارائه شده‌است روش k ـمیانگین و خوشه‌بندی طیفی می‌باشند. در بحث خوشه‌بندی در سالهای اخیر موضوع خوشه‌بندی چنددیدگاهه بسیار مورد توجه قرار گرفته است. در این روشها داده هایی وجود دارند که اطلاعات موجود برای آنها از دیدگاههای مختلف وجود دارد؛ به عنوان مثال یك خبر می‌تواند به زبانهای مختلف انتشار یابد. بنابراین برای هر مجموعه داده چندین ماتریس ویژگی وجود دارد و هدف این است که با در نظر گرفتن همزمان همه این ماتریسهای ویژگی خوشه‌بندی بهتری برای مجموعه داده ارائه گردد. در برخی روشها، که داده به صورت چند دیدگاهه نمی‌باشد، جهت دسترسی به اطلاعات بهتری از داده سعی می‌شود که با روشهای ابتکاری ماتریس ویژگیهای جدیدی برای داده به دست آورند، به عنوان مثال از توان‌های ماتریس مشابهت استفاده می‌شود [۲۳] . در این راستا تحقیقات زیادی بر اساس معیارهای مختلف انجام شده است [۶، ۸، ۱۲، ۲۲]، ولی هیچ گاه مساله از این منظر مورد بررسی قرار نگرفته است که اگر انتخاب ویژگیهایی به تعداد متفاوت انجام پذیرد، آیا با در نظر گرفتن همه این انتخاب ویژگیها به صورت همزمان به عنوان اطلاعات جدید، در خوشه‌بندی نهایی بهبودی حاصل می‌شود یا خیر؟ به بیان دیگر آیا صرف انتخاب ویژگی به نتیجه بهتری دست می‌یابیم و یا اینکه می‌توان از انتخاب ویژگی به شیوه های مختلف یا تعداد مختلف به عنوان ابزاری جهت دستیابی به اطلاعات پنهان داده ها استفاده کرد و احاطه بهتری جهت بررسی و تحلیل داده‌ها داشت؟ به دست آوردن ماتریسهای ویژگی جدید با تعداد ویژگیهای متفاوت در واقع به دیدگاهی جدید از داده منجر

می‌شود و می‌توان تحلیل و خوشه‌بندی داده را با در نظر گرفتن آن به صورت داده ای چنددیدگاهه انجام داد.
در این راستا، در این مقاله روش جدیدی برای خوشه‌بندی بر اساس در نظر گرفتن انتخاب ویژگی های متفاوت به صورت همزمان ارائه می‌شود که کارایی این روش با استفاده از داده‌های مالیاتی و در مقایسه با برخی روشهای دیگر، مورد ارزیابی قرار می‌گیرد و طبق معیارهای ارزیابی می‌توان نتیجه گرفت که استفاده از این روش می‌توان راهکارهایی برای بررسی یا پیش‌بینی مؤدیانی که فرار مالیاتی انجام می‌دهند داشت. به صورت کلی در این مقاله موارد ذیل ارائه می‌گردد:

- راهکاری جدید برای به کارگیری روش انتخاب ویژگی در تحلیل داده های مالیاتی ارائه می‌گردد که در آن از انتخاب تعداد ویژگیهای مختلف برای به دست آوردن دیدگاههای جدید برای داده اصلی استفاده می‌گردد.
- ماتریسهای جدید داده با تعداد ویژگیهای مختلف، به عنوان اطلاعاتی جدید برای آن در نظر گرفته می‌شود و با استفاده از روش خوشه‌بندی طیفی ماتریسهای مشخصه خوشه‌بندی برای هر منظر و نیز ماتریس کلی مشخصه خوشه‌بندی به صورت همزمان مورد یادگیری قرار می‌گیرند.
- با استفاده از ماتریس مشخصه کلی، خوشه‌بندی بر روی داده های مالیاتی انجام می‌شود و با برخی روشهای متداول و اخیر مورد مقایسه قرار می‌گیرد.

می‌توان نمودار ۱ را به عنوان شکل روش پیشنهادی ارائه نمود.



در ادامه در بخش دوم مروری بر تحقیقات گذشته در مورد حل مسائل مالیاتی از طریق داده‌کاوی داریم. سپس در بخش سوم به توضیحاتی در مورد انتخاب ویژگی، خوشه‌بندی k ـ میانگین ، خوشه‌بندی طیفی می‌پردازیم. در بخش چهارم روش جدیدی بر پایه در نظر گرفتن حالتهای مختلف از انتخای ویژگی داده‌ها ارائه می‌نماییم. بخش پنجم به بررسی روش روی داده‌های مالیاتی اختصاص دارد و نتایج را ارزیابی و ارائه می‌نماییم. در بخش پنجم نتیجه گیری این تحقیق را ارائه می‌نماییم.

۲. تاریخچه حل مسائل مالیاتی از طریق داده‌کاوی

در این بخش مروری کوتاه ازکاربردهای علم داده‌کاوی در مسائل مالی و امور مالیاتی ارائه می‌گردد. در سالهای اخیر در زمینه کشف فرارها و تقلبهای مالی و مالیاتی، روشهایی با استفاده از الگوریتمهای مختلف داده‌کاوی ارائه شده است. شبکه‌های عصبی چند لایه، ماشین بردار پشتیبان، الگوریتمهای ژنتیک و رگرسیون لجستیک از آن دسته اند. کرکوس ،اسپاتیس و مانولوپس در سال ۲۰۰۷ از روش درخت تصمیم گیری، شبکه‌های عصبی مصنوعی و شبکه های بیزینی برای پژوهش های روشهای داده کاوی برای تشخیص صورتهای مالی جعلی به این نتیجه رسیدند که روشهای بیزینی نسبت به بقیه کارایی بهتری دارند [۲۰]. ژو و کاپور در سال ۲۰۱۱ از تکنیک های شامل مجموعه مدل های رگرسیونی، درخت تصمیم گیری و شبکه بیزینی برای تشخیص تقلب در صورتهای مالی استفاده کردند و به این نتیجه رسیدند که ترکیب این مدل ها نتایج بهتری خواهد

داشت [۳۱]. ووا و همکارانش درسال ۲۰۱۲ منظور افزایش عملکرد تشخیص فرار پرداخت از مالیات روشهای مبتنی بر داده کاوی استفاده کردند که استفاده از سیستم های مبتنی بر داده کاوی موجب بهبود چشمگیر در تشخیص شیوه‌های فرار در حوزه مالیات شد [۲۷].

باقرپور و همکاران از جمله محققان داخلی بودند که در سال ۲۰۱۲ با استفاده از تکنیک های داده کاوی مانند درخت تصمیم و شبکه‌های عصبی مصنوعی به بررسی عوامل مالی و غیر مالی برگریز مالیاتی پرداختند. [۱]. در همین سال برای پیش بینی گزارش حسابرس مستقل در ایران از روش های درخت تصمیم، شبکه‌های عصبی مصنوعی و رگرسیون لجستیک استفاده شد و این نتیجه حاصل شد که میانگین دقت مدل حاصل از روش درخت تصمیم قابل قبول تر می‌باشد [۲]. سهرابی و همکاران درسال ۲۰۱۵ برای ارزیابی مالیات عملکرد شرکت ها و تحلیل روندهای مالیاتی با استفاده از الگوریتم های داده کاوی، که شامل روش های خوشه‌بندی و طبقه بندی بود، استفاده کردند که برترین این روشها خوشه‌بندی مبتنی برچگالی است [۴]. وانهولودا و همکارانش نیز در سال ۲۰۲۰ تحقیقی درباره شیوه کشف تقلب مالیات بر ارزش افزوده با تکنیکهای خاص مانند قابل تشخیص ناهنجاری بدون نظارت انجام داده‌اند و یک روش ارزیابی جدید معرفی نموده‌اند که با استفاده از آن نشانه‌های رفتار قابل اطمینان را می‌توان بررسی نمود [۲۴]. یان و همکارانش در سال ۲۰۲۰ در یك کار پژوهشی از الگوریتم ژنتیک تطبیقی بهبودیافته برای ارائه مدلی جهت شناسایی تقلب در بیمه خودرو بر اساس شبکه عصبی استفاده کردند. نتایج تجربی نشان داد که الگوریتم ژنتیک بهبودیافته دارای دقت بهتر و مؤثر تری نسبت به الگوریتمهای ژنتیک متداول و سنتی می‌باشد [۳۰]. در سال ۲۰۲۰، دونگ و همکارانش تشخیص تقلب از طریق جنگل تصمیم گیری که ویژگی خود رمزگذار عصبی را دارد، مورد بررسی قرار دادند. آنها یك روش و الگوریتم قابل آموزش را پیشنهاد دادند و کارایی این روش را با بررسی روی داده‌ها نشان دادند [۱۱]. شا و کومار در سال ۲۰۲۱ از الگوریتمهای متعدد یادگیری ماشین مانند ماشین بردار پشتیبان، k ـنزدیکترین همسایه و شبکه عصبی مصنوعی برای پیش بینی وقوع تقلب استفاده کرده اند [۵]. در جدول ۱ مزایا و معایب برخی روشهای ذکر شده و نیز روش پیشنهادی جدید ارائه می‌گردد.

۳. مروری بر انتخاب ویژگی، خوشه‌بندی k ـمیانگین ، خوشه‌بندی طیفی

در این بخش ابتدا یك روش متداول انتخاب ویژگی بیان می‌شود و سپس روشهای خوشه‌بندی k ـمیانگین و خوشه‌بندی طیفی به صورت خلاصه ارائه می‌گردد. اما قبل از آن لازم است گراف متناظر با داده‌ها توضیح داده شود. فرض کنیم که مجموعه داده $X = \{x_1, \dots, x_n\}$ را داریم. در گراف متناظر با داده‌ها رأس i ام متناظر داده x_i می‌باشد. این گراف با یکی از سه روش ϵ - همسایگی، k - نزدیکترین همسایگی و گراف کامل به دست می‌آید چنانچه رئوس متناظر داده‌های x_i و x_j به هم متصل باشند، اغلب از تابع کرنل گاوسی برای تعیین میزان شباهت آنها یا همان وزن یالها استفاده می‌شود. با داشتن گراف متناظر داده‌ها می‌توان ماتریس مجاورت گراف، W، که ماتریسی متقارن است و نیز ماتریس لاپلاسین را به صورت $L = D - W$ که D ماتریس قطری متناظر با گراف است و درایه‌های قطر درجه وزن دار رئوس می‌باشند. مرجع [۲۵] نگاهی اجمالی به تعدادی از خواص ماتریس لاپلاسین دارند.

۱.۳. *انتخاب ویژگی به روش امتیاز لاپلاسین*

در روشهای انتخاب ویژگی، هدف انتخاب ویژگی‌های مرتبط است که غالباً بیشترین خاصیت جداکنندگی داده‌های کلاس های مختلف را دارند و در نتیجه بیشترین تأثیر را در نتیجه مطلوب دارند[۱۹]. امتیاز لاپلاسین یکی از روشهای انتخاب ویژگی می‌باشد که اهمیت ویژگیها با توجه به قدرت آنها برای حفظ همسایگی داده‌ها مورد توجه قرار می‌گیرد. بنابراین ویژگیهای مرتبط به گونه‌ای انتخاب می‌شوند که این خاصیت حفظ شود [۱۴]. فرض کنید l_r نمایانگر امتیاز لاپلاسین ویژگی r ام باشد و نیز x_{ri} نشان دهنده میزان ویژگی r در داده i ام باشد. برای ویژگی r ام بردار $f_r = [x_{r1}, \dots, x_{rm}]^T$ معرفی می‌شود. $\bar{f}_r$ به صورت

$$\bar{f}_r = f_r - \frac{f_r^T D \mathbf{1}}{\mathbf{1}^T D \mathbf{1}} \mathbf{1}$$

روش	نوع یادگیری	مزایا	محدودیتها
kـمیانگین	بدون نظارت	کم هزینه بودن قابلیت اجرا روی داده بزرگ	حساس به شرایط اولیه عدم توانایی تشخیص شکل پیچیده
رگرسیون لجستیک	با نظارت	تعداد پارامتر کم قدرت تشخیص ویژگیهای مهم نتایج خوب روی تعداد داده کم	عدم توانایی تشخیص شکل پیچیده بیش برازش هنگام استفاده تعداد داده زیاد عدم تشخیص روابط بین ویژگیها
درخت تصمیم گیری	با نظارت	کار با هر نوع ویژگی بدون نیاز به پارامتر	احتمال تولید روابط نادرست
شبکه بیزینی	با نظارت	حساسیت کم به نقاط گم شده عملکرد خوب روی داده کم	حساس به شکل داده ورودی
شبکه عصبی	با نظارت	دقت بالا قدرت تشخیص شکلهای پیچیده	نیاز به تعداد زیاد داده تست تعداد پارامتر زیاد
ماشین بردار پشتیبان	با نظارت	تعداد پارامتر کم قدرت کار با داده های با بعد بالا	حساسیت به نویز
روش پیشنهادی	بدون نظارت	قدرت تشخیص شکل پیچیده تحلیل داده با دیدگاههای متفاوت تعداد پارامتر اولیه کم قدرت تشخیص ویژگیهای مهم	نیاز به ساخت گراف داده ها عدم تشخیص روابط بین ویژگیها

جدول ۱: مقایسه برخی روشهای استفاده‌شده برای تحلیل داده‌های مالیاتی

تعریف می‌شود و امتیاز لاپلاسین r ام ویژگی به صورت

$$l_r = \frac{\bar{f}_r^T L \bar{f}_r}{\bar{f}_r^T D \bar{f}_r}$$

محاسبه می‌شود.

۲.۳. *خوشه‌بندی k ـمیانگین، خوشه‌بندی طیفی*

روش خوشه‌بندی k ـمیانگین یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر محسوب می‌شود. روال کار در این خوشه‌بندی به این صورت است که ابتدا به تعداد خوشه‌های مورد نیاز نقاطی به صورت تصادفی انتخاب می‌شود. سپس داده‌ها با توجه به میزان نزدیکی(شباهت)، به یکی از خوشه‌ها نسبت داده می‌شوند و بدین ترتیب خوشه‌های جدیدی حاصل می‌شود. با تکرار همین روال می‌توان در هر تکرار با میانگین گیری از داده‌ها مراکز جدیدی برای آن‌ها محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. این روند تا زمانی ادامه پیدا می‌کند که شرط توقف برقرار گردد.

یکی از مشکلات اصلی خوشه‌بندی k ـمیانگین این است که خوشه‌هایی با شکلهای پیچیده را نمی‌تواند تشخیص دهد و همین موضوع باعث شده است که روش خوشه‌بندی طیفی برای تعیین خوشه‌هایی با شکلهای پیچیده تر به کار رود. در حل مسأله مربوط به خوشه‌بندی طیفی در گراف متناظر داده‌ها به حذف یالهای کم ارزش می‌پردازیم به طوریکه گراف تقریبا به چند مولفه همبندی که همان خوشه‌ها می‌باشند، برسد. ثابت شده است، [۲۵] ، که این راهکار معادل است با حل مسأله مینم

سازی زیر:

$$\begin{aligned} \min_{F} \quad & Tr(F^T LF), \\ \text{s.t.} \quad & F^T F = I_k. \end{aligned} \tag{۱.۳}$$

که k تعداد خوشه‌ها می‌باشد. با استفاده از قضیه ریلی ریتز [۲۵] جواب این مسأله به یافتن مقادیر ویژه ماتریس لاپلاسین می‌رسد و در واقع F ماتریسی است که ستونهای آن بردارهای ویژه k کوچکترین مقدار ویژه ماتریس L می‌باشد. سطرهای F به عنوان بازنمایشی از داده‌ها در فضای با بعد کمتر در نظر گرفته می‌شوند. حال خوشه‌ها با استفاده از خوشه‌بندی kـمیانگین روی سطرهای F مشخص می‌شوند.

قبل از ورود به بخش اصلی مقاله لازم است که تعریف و نکات زیر را نیز ارائه نماییم..

تعریف ۱.۳. مجموعه داده $\mathcal{X} = \{x_1, \ldots, x_n\}$ که دارای k خوشه $C_1, \ldots, C_k$ است، را در نظر بگیرید. ماتریس $F \in \{0,1\}^{n\times k}$ ماتریس مشخصه این خوشه‌بندی نامیده می‌شود چنانچه درایه (i,j) برابر یك باشد اگر $x_i \in C_j$ و در غیر اینصورت صفر باشد.

در حل مساله خوشه‌بندی طیفی، چنانچه هر مولفه گراف دقیقا مشخص کننده یك خوشه باشد،ماتریس حاصل از حل مساله مینیمم سازی ۱.۳ در شرایط تعریف فوق صدق می‌کند. اما چنانچه هر مولفه بیانگر یه خوشه نباشد ماتریس F حاصل از مساله مینیمم سازی این ویژگی را نخواهد داشت با این حال از آنجا که در روند خوشه بندی طیفی یالهای کم اهمیت حذف می شوند، گراف ایده آل داده ها(منظور گرافی که هر مولفه یك خوشه است) با گراف اصلی داده تنها در تعدادی یال کم اهمیت تفاوت دارد. بنابراین ماتریس مشابهت و به تبع آن ماتریس لاپلاسین گراف ایده آل داده ها، L_1، و ماتریس اصلی و ماتریس لاپلاسین متناظر با داده ها، L، اختلاف کمی دارند. به عبارت دیگر ماتریس E با درایه های کوچك وجود دارد به طوریکه $L = L_1 + E$. حال طبق قضیه دیویس‌ـکاهان [۲۵] بردارهای ویژه این ماتریسها به هم نزدیك می باشند. بنابراین می توان نتیجه گرفت که ماتریس حاصل از مساله بهینه سازی ، فارغ از ترتیب قرار گرفتن ستونها، تا حد زیادی به ماتریس حاصل از یك گراف ایده آل شبیه می باشد، بنابراین با اغماض می توان آن را به صورت یك ماتریس مشخصه در نظر گرفت و به این ماتریس نیز ماتریس مشخصه گوییم.

۴. خوشه‌بندی طیفی با در نظر گرفتن حالتهای مختلف انتخاب ویژگی به صورت همزمان

یکی از مباحثی که در سالهای اخیر مورد توجه قرار گرفته است یادگیری چند دیدگاهه می باشد. در یادگیری چند دیدگاهه برای یک داده چند بردار ویژگی وجود دارد. به عنوان مثال اگر داده یک انسان باشد، تصویر شخص، امضای شخص و ویژگیهای اخلاقی شخص می توانند سه بردار ویژگی برای شخص ارایه دهند. بنابراین اگر مجموعه ای از داده ها(انسانها) داشته باشیم می توان سه نوع ماتریس ویژگی(مربوط به تصویر اشخاص، امضا و ویژگیهای اخلاقی) در نظر گرفت و یا یک خبر ممکن است به زبانهای مختلف در وبسایتها موجود باشد و برای خوشه‌بندی یا دسته بندی اخبار می‌توان همه این زبانها را به صورت همزمان در نظر گرفت. تهیه اطلاعات برای یک مجموعه داده از منظرهای مختلف می‌تواند توسط تحلیل گر نیز انجام شود. به طور کلی می‌توان داده چند دیدگاهه را به صورت زیر تعریف نمود:

تعریف ۱.۴. فرض کنید مجموعه پایه $\mathcal{X} = \{x_1, \ldots, x_n\}$ را داریم. این مجموعه داده را داده چند دیدگاهه گوییم چنانچه ماتریسهای ویژگی متفاوت برای این داده به صورت $X_1, \ldots, X_v$ موجود باشد که $X_r \in \mathbb{R}^{n\times d_r}$.

در مسائل یادگیری چند دیدگاهه هدف ارایه راهکاری است که بتوان دیدگاههای مختلف را به صورت همزمان در نظر گرفت و داده پایه که همان اشخاص در این مثال می باشند را خوشه بندی یا دسته بندی کرد. یکی از رویکردهای دیگری که در راستای رویکرد یادگیری چند دیدگاهه می‌توان در نظر گرفت، این است که اگر داده ای به صورت چند دیدگاهه در دسترس

نیست به منظور در نظر گرفتن روابط پنهان بین داده ها دیدگاههای مختلفی برای داده به وجود آورد و یا ماتریسهای شباهت متفاوتی، با توجه به اینکه چه موضوعی در تحلیل داده مربوطه اهمیت دارد، به وجود آورد. به عنوان مثال چنانچه برای یک مجموعه داده ماتریس شباهت A وجود دارد، می توان دیگر دیدگاهها را ماتریس های شباهت $A^2, A^3, \ldots, A^p$ در نظر گرفت که به نوعی روابط بین رئوس برای در نظر گرفتن ویژگیهای پنهان و به دست آوردن دیدگاههای جدید از داده به کار می آید [۲۳]. در این بخش به معرفی روشی جدید برای خوشه‌بندی می‌پردازیم که در آن انتخاب ویژگیهایی به تعداد متفاوت انجام می‌شود و سپس با در نظر گرفتن همه این انتخاب ویژگیها خوشه‌بندی را انجام می‌دهیم. در واقع بررسی داده‌ها با در نظر گرفتن منظرها یا دیدگاههای مختلف راهکاری در جهت بهبود نتایج یادگیری داده ها می‌باشد. در این بخش قصد داریم که با داشتن مجموعه داده X و انجام انتخاب ویژگی های متفاوت، از دیدگاه‌های مختلف به این مجموعه داده بنگریم و با داشتن همه این دیدگاه‌ها خوشه‌بندی را انجام دهیم. در ادامه ابتدا مدل پیشنهادی را ارائه می‌نماییم. سپس با استفاده از روش تکراری به حل این مسأله بهینه سازی این روش می‌پردازیم.

۱.۴. *روش پیشنهادی*

همانطور که در مقدمه این بخش بیان گردید، راهکار پیشنهادی بر این مبنا است که برای داده های مالیاتی با توجه به اینکه چگونه روابط یا اطلاعاتی برای تحلیل داده ها مهم تر است، دیدگاههای جدید به دست آورده شود. از آنجا که تشخیص ویژگیهای مهم داده ها در تحلیل این نوع داده ها مهم است، در این روش دیدگاههای مختلف با استفاده از در نظر گرفتن ماتریس های ویژگی که با فرآیند انتخاب ویژگی به دست می‌آید، تعریف می‌شوند. در اینجا می‌توان فرآیند انتخاب ویژگی را انجام داد؛ بدین گونه که می توان انتخاب ویژگی r ویژگی را به عنوان ماتریس ویژگی جدید در نظر گرفت، X_r. یعنی هر X_r به عنوان دیدگاهی جدید برای داده پایه در نظر گرفته می‌شود. سپس همه این ماتریسهای انتخاب ویژگیهای مختلف به صورت همزمان مورد بررسی قرار می گیرد. در ادامه به بیان مساله اصلی می پردازیم.

فرض می‌کنیم که برای مجموعه داده $X = \{x_1, \ldots, x_n\}$ را در نظر بگیرید. راهکار برای به دست آوردن دیدگاههای جدید این است که برای این مجموعه انتخاب ویژگیها به تعداد متفاوت انجام دهیم. برای این کار روند انتخاب ویژگی با یك یا با چند روش انتخاب ویژگی به تعداد $n_1, \ldots, n_\nu$ ویژگی به صورت مستقل انجام می شود. منظور با چند روش انتخاب ویژگی اینست که به عنوان مثال می توان n_1 ویژگی را با روش شماره یك و n_2 ویژگی را با روش شماره دو انتخاب کرد. حال ماتریسهای حاصل از این انتخاب ویژگیها، یعنی $X_1, \ldots, X_\nu$، را داریم. بنابراین برای مجموعه داده ذکر شده یک مجموعه از ماتریسها، که می توان آنها را دیدگاههای مختلف مجموعه داده اصلی در نظر بگیریم، داریم. حال برای دیدگاه t ، با استفاده از حل مسأله

$$\begin{aligned} \min_{F_t} \quad & Tr(F_t^T L_t F_t), \\ \text{s.t.} \quad & F_t^T F_t = I_k. \end{aligned} \tag{۱.۴}$$

می‌توان یک ماتریس مشخصه برای خوشه‌بندی طیفی آن دیدگاه پیدا کرد. در اینجا به دنبال ماتریس مشخصه مشترکی هستیم که تا حد امکان برای خوشه‌بندی همه این دیدگاه‌های مختلف مناسب باشد. در نگاه اول شاید به نظر برسد که بهینه کردن فاصله همه ماتریسهای مشخصه دیدگاههای مختلف از ماتریس مشخصه مشترك راهکار مناسبی باشد، به عبارت دیگر از عبارت $\min_F \|F - F_t\|$ برای به دست آوردن یك ماتریس مشخصه F که تا حد امکان به همه این ماتریسهای مشخصه نزدیك باشد، استفاده نماییم. اما نکته حائز اهمیت این است که در به دست آوردن ماتریسهای مشخصه دیدگاههای مختلف ترتیب ستونها که همان بردارهای مشخصه خوشه هستند لزوما رعایت نمی شود. بنابراین برای حل این مشکل از این خاصیت استفاده می‌نماییم که در ماتریس $F_t F_t^T$ در حالت ایده ال درایه (i, j) برابر ۱ است اگر داده‌های i داده j ام داخل یك خوشه قرار بگیرند و در غیر اینصورت صفر است. حال برای به دست آوردن ماتریس مشخصه حاصل از در نظرگرفتن انتخاب ویژگیهای با تعداد متفاوت به صورت همزمان، ماتریسی در نظر می‌گیریم که تا حد امکان این خاصیت را حفظ نماید، یعنی

عبارت

$$\|FF^T - F_tF_t^T\|_F^2 \tag{۲.۴}$$

را برای همه دیدگاه‌ها باید به حداقل برسانیم [۲۱] . با توجه به روابط ۱.۴ و ۲.۴ مسأله بهینه سازی زیر را در نظر می‌گیریم:

$$\begin{array}{ll}\min\limits_{F_t,F} & \sum_{t=1}^{v}\beta Tr(F_t^TL_tF_t)+\alpha_t\|FF^T-F_tF_t^T\|_F^2,\\ \text{s.t.} & F^TF=I_k, F_t^TF_t=I_k.\end{array} \tag{۳.۴}$$

که در اینجا α_t وزنی است که برای میزان تأثیر هر جمله در نظر گرفته می‌شود که مشابه آنچه که در [۲۱] انجام شده است؛ می‌توان در نظر گرفت $\alpha_t = \frac{1}{2\|FF^T-F_tF_t^T\|_F}$. در واقع هر چه F_t به F نزدیکتر باشد، مقدار α_t بیشتر است و تأثیر بیشتری دارد.

۲.۴. *حل مسأله*

برای به دست آوردن هر یك از مجهولات مسأله ۳.۴ دیگر متغیرها را ثابت در نظر می‌گیریم و مسأله را نسبت به آن مجهول خاص حل می‌نماییم و سپس با استفاده از روش بازگشتی الگوریتمی‌جهت به دست آوردن همه مجهولات ارائه می‌نماییم. بنابراین برای به دست آورن F مقادیر F_t را ثابت می‌گیریم و از رابطه زیر استفاده می‌نماییم:

$$\sum_{t=1}^{v}\alpha_t\|FF^T-F_tF_t^T\|_F^2=\sum_{t=1}^{v}\alpha_tTr[(FF^T-F_tF_t^T)^T(FF^T-F_tF_t^T)].$$

بنابراین باید مسأله زیر را حل نماییم:

$$\begin{array}{ll}\min\limits_{F} & Tr(F^TMF),\\ \text{s.t.} & F^TF=I_k.\end{array} \tag{۴.۴}$$

که در آن

$$M=\sum_{t=1}^{v}\alpha_t(-2F_tF_t^T+I)$$

که این مسأله مشابه آنچه که در روش خوشه‌بندی طیفی بیان شد با استفاده از قضیه ریلی ریتز قابل حل می‌باشد و F ماتریسی است که ستونهای آن بردارهای ویژه متناظر با k کوچکترین مقادیر ویژه ماتریس M می‌باشد.
برای به دست آوردن هر F_t نیز دیگر مقادیر را ثابت می‌گیریم؛ بنابراین باید مسأله زیر را حل نماییم:

$$\begin{array}{ll}\min\limits_{F_t} & \beta Tr(F_t^TL_tF_t)+\alpha_t\|FF^T-F_tF_t^T\|_F^2,\\ \text{s.t.} & F_t^TF_t=I_k.\end{array} \tag{۵.۴}$$

که معادل است با حل مسأله

$$\begin{array}{ll}\min\limits_{F_t} & Tr(F_t^TNF_t),\\ \text{s.t.} & F_t^TF_t=I_k.\end{array} \tag{۶.۴}$$

که در آن

$$N = \beta L_t - ۲\alpha_t FF^T + \alpha_t I.$$

از آنجا که α_t، F_t و F به یکدیگر وابسته اند، این مقادیر را نمی‌توان به صورت مستقل به دست آورد و با یك مقدار دهی اولیه مقادیر ذکر شده به صورت تکراری به روزرسانی می‌شوند. الگوریتم این روش را می‌توان به صورت الگوریتم ۱ ارائه نمود.

الگوریتم ۱ روش پیشنهادی

ورودی: مجموعه داده X، مقادیر اولیه α_t، پارامتر β، تعداد خوشه‌ها k و تعداد تکرار r
خروجی: F

۱: دیدگاه‌های مختلف $\{X_۱, \ldots, X_v\}$ را با استفاده از انتخاب ویژگی به تعداد مختلف به دست آور.
۲: مجموعه ماتریس های مشخصه $\{F_۱, \ldots, F_v\}$، را با حل رابطه ۱.۴ محاسبه کن.
۳: مراحل زیر را به تعداد تکرار r انجام بده.
۴: F را باتوجه به رابطه ۴.۴ به روز رسانی کن.
۵: F_t را با توجه به رابطه ۶.۴ به روز رسانی کن.
۶: α_t را با استفاده از $\alpha_t = \frac{۱}{۲\|FF^T - F_tF_t^T\|_F}$ به روز رسانی کن.

۵. ارزیابی روش پیشنهادی برای مجموعه داده‌های مالیاتی

در این بخش به ارزیابی روش ارائه شده برای دو مجموعه داده مالیاتی می‌پردازیم. ابتدا این داده ها را توضیح می‌دهیم و سپس روشهای ارزیابی را به صورت مختصر معرفی می‌نماییم. بعد از آن نتایج روشها را روی داده ها ارائه می‌دهیم.

۱.۵. *معرفی مجموعه داده مالیاتی*

مجموعه $X = \{x_۱, \ldots, x_n\}$ دارای n داده که در اینجا داده‌ها مؤدیان مالیاتی می‌باشند را در نظر میگیریم. اطلاعات هر مؤدی نمایانگر ویژگیهای هر داده است که در ارتباط با انجام یا عدم انجام تکالیف مالیاتی مالیات دهنده می‌باشد. این ویژگیها شامل موارد زیر می‌باشد: اعلامیه متمم، تسلیم اظهارنامه(صورت مجلس)، تسلیم دفاتر، عدم تسلیم دفاتر یا اظهارنامه، صدور برگ تشخیص اولیه، صدور برگ متمم، نوع مؤدی(کوچک، متوسط، بزرگ)، تمکین مؤدی، اعتراض و توافق، اعتراض و عدم توافق، برگ قطعی و واریز، برگ قطعی و عدم واریز.

بنابراین جدولی از مؤدیان داریم که در هرسطر جدول اطلاعات هر مؤدی بر اساس بله یا خیر نمایش داده شده است. قسمتی از این جدول برای ده مؤدی شهر کرمان در شکل زیر آمده است. حال این اطلاعات را به ماتریس داده‌ها تبدیل

مودی	تسلیم اظهارنامه	صورت مجلس	عدم صورتمجلس یا اظهارنامه	صدور برگ تشخیص اولیه	صدور برگ متمم	مودی کوچک	مودی متوسط	مودی بزرگ	تمکین مودی	اعتراض و توافق ماده ۲۳۸	اعتراض و عدم توافق	برگ قطعی و واریز	برگ قطعی و عدم واریز
1	بلي	خير	بلي	خير	خير	بلي	خير	خير	خير	خير	خير	خير	بلي
2	بلي	خير	بلي	خير	خير	بلي	خير	خير	خير	خير	خير	خير	بلي
3	بلي	بلي	خير	بلي	خير	بلي	خير	خير	بلي	بلي	خير	خير	بلي
4	بلي	خير	بلي	خير	خير	بلي	خير	خير	خير	خير	خير	خير	بلي
5	بلي	خير	بلي	خير	خير	بلي	خير	خير	خير	خير	خير	خير	بلي
6	بلي	بلي	خير	بلي	خير	بلي	خير	خير	خير	خير	خير	خير	بلي
7	بلي	خير	بلي	خير	خير	بلي	خير	خير	خير	خير	خير	خير	بلي
8	بلي	بلي	خير	بلي	خير	بلي	خير	خير	خير	خير	خير	خير	بلي
9	بلي	خير	بلي	خير	خير	بلي	خير	خير	خير	خير	خير	خير	بلي
10	بلي	خير	بلي	خير	خير	بلي	خير	خير	خير	خير	خير	خير	بلي

شکل ۱: نمونه داده های مالیاتی

می‌نماییم. بدین گونه که در ماتریس چنانچه این اطلاعات در مورد مؤدی برقرار باشد عدد ۱ و در غیر اینصورت عدد صفر قرار دارد. بنابراین ماتریس ویژگیها را به این صورت می‌سازیم.

در اینجا از دو جدول استفاده می‌شود یکی اطلاعاتی که در مقاله [۳] برای ۴۰ مؤدی ارائه شده است و دیگری اطلاعات مربوط به ۱۵۰ مؤدی سازمان مالیاتی کرمان در سال ۱۴۰۰.

۲.۵. *معیارهای ارزیابی*

بعد از انتخاب ویژگی و خوشه‌بندی داده‌ها، با معیارهای ارزیابی روند کار را می‌سنجند تا کارایی آن مورد تحلیل قرار گیرد. روشهای متعددی برای ارزیابی روشهای خوشه‌بندی وجود دارد که از بین آنها دو روش بسیار متداول را انتخاب نموده ایم که توضیحات آنها در زیر آمده است. این روشها در مراجع [۷، ۱۵] به تفصیل بیشتری توضیح داده شده اند.

- **شاخص دقت**[1]: ACC به بررسی نقطه به نقطه داده‌ها می‌پردازد، میزان قرارگرفتن داده‌ها در خوشه صحیح را محاسبه می‌نماید و به صورت زیر تعریف می‌شود:

$$ACC = \frac{N_{cor}}{N}$$

که N تعداد داده‌ها و N_{cor} تعداد داده‌های قرارگرفته در خوشه درست می‌باشد.

- **شاخص خلوص**[2]:

یکی از متداولترین شاخص‌های ارزیابی در خوشه‌بندی، شاخص خلوص است که درصد مطابقت بین برچسب‌های خوشه‌بندی و برچسب‌های واقعی را می‌سنجد. در این حالت برچسب هر خوشه با برچسب واقعی دسته‌ای که بیشترین اشتراک را دارد مطابقت پیدا کرده و تعداد نقاطی از خوشه که در دسته صحیح طبقه‌بندی شده‌اند شمارش می‌شوند. نسبت این تعداد به تعداد کل نقاط، شاخص خلوص را می‌سازد و شکل محاسباتی آن به صورت زیر است:

$$Purity = \frac{\sum_{i=1}^{k} \max_j(|S_i \cap C_j|)}{N} = \frac{\sum_{i=1}^{k} \max_j(m_{ij})}{n}$$

که در اینجا S_i ها دسته های واقعی می‌باشند و C_i ها خوشه هایی می‌باشند که با استفاده از یك روش خوشه‌بندی به دست آمده اند.

- **اطلاعات متقابل نرمال شده**[3]: اطلاعات متقابل نرمال شده یکی از روشهای آماری ارزیابی نتایج خوشه‌بندی داده ها می‌باشد که در آن از میزان همبستگی بین متغیرها استفاده می‌شود. در واقع اطلاعات متقابل میزان اطلاعاتی که بین دو متغیر مشترک است را با استفاده از آنتروپی متغیرهای تصادفی به دست آورد. این معیار به صورت زیر فرمول بندی شده است:

$$\text{Nmi} = ۲\frac{M(X_i, X_j)}{E(X_i) + E(X_j)}.$$

که $M(X_i, X_j)$ اطلاعات متقابل بین دو متغیر را برآورد می‌نماید و $E(X_i)$ آنتروپی هر متغیر می‌باشد.

---

[1] Accuracy
[2] Purity Index
[3] Normalized Mutual Information

- **شاخص رند اصلاح شده**[4]: شاخص رند اصلاح شده به بررسی قرار گرفتن زوج داده های هم دسته یا جدا، داخل خوشه ها می‌پردازد و به صورت زیر تعریف شده است:

$$ARI(S,C) = \frac{\sum_{i=1}^{l}\sum_{j=1}^{k}\binom{m_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}$$

که $t_1 = \sum_{i=1}^{k}\binom{|C_i|}{2}$، $t_2 = \sum_{j=1}^{l}\binom{|S_j|}{2}$ و $t_3 = \frac{2t_1 t_2}{n(n-1)}$.

۳.۵. *بررسی و تحلیل نتایج برای داده های مالیاتی*

در این قسمت داده‌های معرفی شده در زیربخش ۱.۵ را مورد بررسی قرار می‌دهیم و این داده‌ها را با روشهای داده‌کاوی با دقت بیشتری نگاه می‌کنیم. هدف این است که بررسی کنیم، که آیا با روش خوشه‌بندی، داده‌های دارای واریز و عدم واریز می‌توان تفکیک نمود. ابتدا با استفاده از روش انتخاب ویژگی، ویژگیهای برتر داده‌ها را انتخاب می‌نماییم. سپس روشهای خوشه‌بندی را برای معین کردن خوشه‌ها به کار می‌بریم.

برای انتخاب ویژگی از روش روش امتیاز لاپلاسین استفاده می‌نماییم. بعد از آن از روش خوشه‌بندی طیفی استفاده می‌نماییم و این داده‌ها را خوشه‌بندی می‌نماییم. بعد از انجام روش امتیاز لاپلاسین ویژگیها با ترتیبی که در ادامه آمده مرتب می‌شوند: اعلامیه متمم، صدور برگ متمم، تمکین مؤدی، اعتراض و توافق ماده ۲۳۸، تسلیم اظهارنامه، تسلیم دفاتر، عدم تسلیم دفاتر یا اظهارنامه، صدور برگ تشخیص اولیه، اعتراض و عدم توافق، مؤدی کوچک، مؤدی بزرگ، مؤدی متوسط.

ملاحظه می‌شود ارتباط خوبی بین ویژگیهای به دست آمده در این روش و ویژگیهای به دست آمده در [۳] که عبارت بودند از واصل شدن اعلامیه، تسلیم اظهارنامه، تسلیم دفاتر و یا عدم تسلیم یکی از آنها وجود دارد. لازم به توضیح است برخی از ویژگیهای ارائه شده در ماتریس داده به یکدیگر وابستگی خاصی دارند به عنوان مثال بردار ویژگی مربوط به اعلامیه متمم و صدور متمم کاملا یکسان می‌باشد.

در ادامه با نگه داشتن تعدادی از ویژگیهای مهم، با استفاده از روشهای k- میانگین و خوشه‌بندی طیفی داده‌ها را خوشه‌بندی می‌نماییم. همچنین از دو روش اخیر انتخاب ویژگی با ایده های بر پایه کاهش مختصات [۲۸] و بر اساس واریانس کواریانس فاصله زیرفضایی [۱۸] نیز برای مقایسه روش استفاده می‌نماییم. در جدولهایی که در ادامه می‌آید به ترتیب سه تا نه ویژگی را مورد استفاده قرار داده‌ایم و نتایج ارزیابی را با روش جدید ارائه شده مقایسه کرده‌ایم. برای روش پیشنهادی، تعداد دیدگاههای انتخاب شده در جدول متفاوت می‌باشد. به عنوان مثال در ستون اول دیدگاه‌هایی انتخاب شده‌اند که در آنها یك، دو و سه ویژگی انتخاب شده‌اند و این دیدگاه‌ها را با استفاده از روش پیشنهادی به صورت همزمان در نظر گرفته ایم. در جدولها بهترین نتایج را با هایلایت مشخص نموده‌ایم. با توجه به نتایج ارائه شده، مشخص می‌شود که همه روش های مطرح شده قابلیت تشخیص نسبتا خوبی برای بررسی این نوع داده‌های مالیاتی دارند. اما از بین روشها، روش خوشه‌بندی چند دیدگاهه در مجموع می‌تواند بیشتر مورد توجه قرار گیرد.

[4] Adjust Rand Index

معیار	روش	$n=۳$	$n=۴$	$n=۵$	$n=۶$	$n=۷$	$n=۸$	$n=۹$
دقت	k-میانگین	۸۲/۵۰	۸۲/۴۶	۸۲/۴۳	۸۳/۳۲	۸۴/۱۱	۸۱/۱۵	۸۱/۰۹
	طیفی	۷۲/۵۰	۷۲/۵۰	۷۲/۵۰	۷۲/۴۱	۷۲/۵	۷۲/۵۳	۷۲/۲۳
	کاهش مختصات	۸۲/۵	۸۰	۸۱/۷۵	۸۲/۵	۷۹/۲۵	۸۲/۵	۷۹/۶۲
	فاصله زیرفضایی	۶۵	۸۰	۸۰	۸۲/۳۷	۸۰	۶۵/۵	۷۸/۲۵
	چنددیدگاهه	۸۵/۸۵	۸۵/۸۵	۸۵/۳۵	۸۶	۸۵/۱	۸۵/۷۵	۸۵/۵
خلوص	k-میانگین	۹۰/۰۰	۹۰/۰۰	۹۰/۰۰	۹۰/۶۵	۹۱/۰۵	۹۰/۱۰	۹۰/۱۰
	طیفی	۹۰	۹۰/۰۲	۹۰	۹۰	۹۰	۹۰/۱۲	۹۰/۰۳
	کاهش مختصات	۹۰	۹۰	۹۰/۷۵	۹۰	۹۰	۹۰	۹۰
	فاصله زیرفضایی	۹۰	۹۰	۹۰	۹۰	۹۰	۹۰	۹۰
	چنددیدگاهه	۹۰	۹۰	۹۰	۹۰	۹۰	۹۰	۹۰
اطلاعات متقابل نرمال شده	k-میانگین	۲۴/۶۳	۲۴/۵۸	۲۴/۵۵	۲۹/۱۹	۳۲/۲۹	۲۲/۴۶	۲۲/۸۶
	طیفی	۴/۳۹	۴/۳۹	۴/۳۹	۴/۳۹	۴/۳۹	۴/۳۹	۴/۳۹
	کاهش مختصات	۲۴/۶۲	۲۱/۹۵	۲۶/۷۳	۲۴/۶۲	۲۰/۲۰	۲۴/۶۲	۲۱/۰۸
	فاصله زیرفضایی	۰/۱۱	۲۱/۹۵	۲۱/۹۶	۲۴/۴۹	۲۱/۹۶	۰/۱۱	۱۷/۰۵
	چنددیدگاهه	۲۷/۳۲	۲۷/۳۲	۲۵/۹۱	۲۷/۷۵	۲۵/۲۱	۲۷/۰۴	۲۶/۳۳
شاخص رند اصلاح شده	k-میانگین	۷۰/۳۸	۷۰/۳۸	۷۰/۳۸	۷۰/۳۸	۷۰/۳۸	۷۰/۳۸	۷۰/۳۸
	طیفی	۵۹/۱۰	۵۹/۱۰	۵۹/۱۰	۵۹/۱۰	۵۹/۱۰	۵۹/۱۰	۵۹/۱۰
	کاهش مختصات	۷۰/۳۸	۶۷/۱۸	۶۷/۱۸	۷۰/۳۸	۶۷/۱۹	۷۰/۳۸	۶۷/۱۸
	فاصله زیرفضایی	۵۳/۳۳	۵۳/۳۳	۶۷/۱۸	۷۰/۳۸	۵۳/۳۳	۵۳/۳۳	۶۷/۱۸
	چنددیدگاهه	۷۷/۵۶	۷۷/۵۶	۷۷/۵۶	۷۷/۵۶	۷۷/۵۶	۷۷/۵۶	۷۷/۵۶

جدول ۲: نتایج خوشه‌بندی برای ۴۰ داده

معیار	روش	$n=۳$	$n=۴$	$n=۵$	$n=۶$	$n=۷$	$n=۸$	$n=۹$
دقت	k-میانگین	۵۸/۷۸	۵۸/۶۶	۵۸/۸۷	۵۸/۶۶	۵۸/۶۶	۵۸/۹۲	۵۸/۶۸
	طیفی	۷۹/۴۱	۷۹/۴۱	۷۸/۹۸	۷۹/۲۰	۷۹/۹۴	۷۹/۹۴	۷۹/۴۱
	کاهش مختصات	۵۸/۶۸	۶۰/۶۶	۵۹/۳۳	۶۰	۵۹/۳۳	۵۸/۶۶	۵۹/۳۷
	فاصله زیرفضایی	۵۸/۶۶	۵۸/۶۶	۵۹/۳۳	۵۸/۶۶	۵۸/۶۶	۵۸/۶۶	۵۸/۶۶
	چنددیدگاهه	۸۷/۱۷	۸۸/۳۳	۹۳/۳۳	۸۷/۱۷	۸۶/۵	۸۶/۱۷	۸۴/۳۳
خلوص	k-میانگین	۷۲/۰۵	۷۲/۱۱	۷۲/۱۴	۷۲/۱۲	۷۲/۱۴	۷۲/۱۴	۷۲/۱۲
	طیفی	۸۰/۴۹	۸۰/۴۹	۸۰/۳۰	۸۰/۴۰	۸۰/۷۲	۸۰/۷۲	۸۰/۴۹
	کاهش مختصات	۷۲	۷۲	۷۲	۷۲	۷۲	۷۲	۷۲
	فاصله زیرفضایی	۷۲	۷۲	۷۲	۷۲	۷۲	۷۲	۷۲
	چنددیدگاهه	۸۹/۳۹	۹۰/۱۳	۹۳/۳۳	۸۹/۳۹	۸۸/۹۶	۸۸/۷۵	۸۷/۵۷
اطلاعات متقابل نرمال شده	k-میانگین	۰/۶۵	۲/۴۵	۰/۹۵	۰/۴۵	۰/۴۵	۰/۴۳	۰/۴۴
	طیفی	۳۶/۷۳	۳۶/۷۳	۳۵/۹۲	۳۶/۳۳	۳۷/۷۴	۳۷/۷۴	۳۶/۷۳
	کاهش مختصات	۰/۴۵	۰/۰۱	۰/۰۸	۰/۰۴	۰/۰۸	۰/۴۵	۰/۰۸
	فاصله زیرفضایی	۰/۴۵	۰/۴۵	۰/۴۵	۰/۴۵	۰/۴۵	۰/۴۵	۰/۴۵
	چنددیدگاهه	۵۰/۱۳	۵۲/۲۸	۶۱/۵۰	۵۰/۱۳	۴۸/۹۰	۴۸/۲۸	۴۴/۹۰
شاخص رند اصلاح شده	k-میانگین	۵۱/۱۷	۵۱/۱۸	۵۱/۱۸	۵۱/۱۸	۵۱/۱۸	۵۱/۱۸	۵۱/۶۷
	طیفی	۶۹/۴۳	۶۹/۴۳	۶۹/۴۳	۵۱/۶۷	۶۹/۴۳	۶۹/۴۳	۶۹/۴۳
	کاهش مختصات	۵۱/۱۷	۵۱/۹۵	۵۱/۴۲	۵۱/۶۷	۵۱/۴۲	۵۱/۱۸	۵۱/۴۱
	فاصله زیرفضایی	۵۱/۱۸	۵۱/۱۸	۵۱/۱۸	۵۱/۱۸	۵۱/۱۸	۵۱/۱۸	۵۱/۱۸
	چنددیدگاهه	۵۱/۶۸	۸۷/۴۷	۸۷/۴۷	۸۷/۴۷	۵۱/۶۸	۸۷/۴۷	۸۷/۴۷

جدول ۳: نتایج خوشه‌بندی برای ۱۵۰ داده

داده	تعداد نمونه ها	تعداد ویژگی ها	تعداد دسته ها
Wine	۱۷۸	۱۳	۳
Glass	۲۱۴	۹	۶
Control	۶۰۰	۶۰	۶
۵Isolet	۱۵۶۰	۶۱۷	۲۶

جدول ۴: معرفی داده های مورد آزمایش

۶. ارزیابی روش پیشنهادی روی داده های غیر مالیاتی

در این بخش به ارزیابی عملکرد روش پیشنهادی روی چهار مدل داده حقیقی متداول می پردازیم. مشخصات این داده ها در جدول ۴ آورده شده است. معیارهای ارزیابی و روشهای پیشنهادی مطابق با بخش قبل انتخاب شده اند.

۱.۶. *تحلیل نتایج*

نتایج آزمایشات روی این داده های ذکر شده در جدولهای ۵، ۶، ۷ و ۸ آورده شده است. در سطر اول تعداد ویژگیهای انتخاب شده مشخص شده است و میانگین نتایج برای هر مورد در ستون آخر آورده شده است. همچنین بهترین نتایج را هایلایت نموده ایم.

با بررسی میانگین نتایج می توان به این نتیجه رسید که غیر از داده Wine که در آن روش کاهش مختصات بهترین نتایج را دارد در بقیه داده ها روش پیشنهادی چند دیدگاهه به طور کلی عملکرد بهتری دارد. به طور دقیق تر در داده Wine روش پیشنهادی برای معیارهای دقت و خلوص با اختلاف خیلی کم در جایگاه دوم قرار دارد. همچنین برای معیارهای اطلاعات متقابل نرمال شده و شاخص رند اصلاح شده، نتایج به دست آمده کاملا قابل قبول و قابل رقابت با روشهای جایگاه اول و دوم می باشند. همچنین در مورد داده Glass روش چند دیدگاهه در شاخص های دقت، خلوص و اطلاعات متقابل نرمال شده بهترین نتایج را دارد و در شاخص رند اصلاح شده با اختلاف کم بعد از روش طیفی در جایگاه دوم قرار دارد. در داده Control برای همه معیارها روش پیشنهادی به طور میانگین بهترین نتایج را دارد و در داده ۵Isolet غیر از معیار شاخص رند اصلاح شده باز هم روش پیشنهادی در سایر معیارهای بهترین نتیجه را به طور میانگین به دست آورده است. همچنین در مورد شاخص رند اصلاح شده نتیجه به دست آمده نزدیک به بهترین نتیجه و قابل قبول می باشد.

۲.۶. *بررسی اثر پارامتر روی داده های غیرمالیاتی*

در روش پیشنهادی یك پارامتر در بیان مساله بهینه سازی مطرح شده است که پارامتر β می‌باشد. با توجه به نمودارهای ارائه شده در ذیل تغییر این پارامتر می‌تواند در نتیجه نهایی موثر باشد. در این زیربخش اثر این پارامتر را برای داده های غیرمالیاتی در مورد شاخص دقت بررسی می نماییم. در شکل ۲ مشخص است که شاخص دقت با تغییر این پارامتر تغییر می کند و این تغییرات برای داده های Control و Glass واضح تر می باشد.

۷. نتیجه گیری و پیشنهادات

در این مقاله روش جدیدی با توجه به در نظر گرفتن دیدگاه‌های مختلف حاصل از انتخاب ویژگیهای متفاوت ارائه گردید. این روش به همراه دو روش خوشه‌بندی طیفی و خوشه‌بندی k ـ میانگین پس از انتخاب ویژگی بر روی داده‌های مالیاتی و نیز دو روش انتخاب ویژگی دیگر مورد ارزیابی قرار گرفتند. نتایج حاصل، نشان می‌دهد که روش ارائه شده نسبت به سایر روشها روش کارایی بهتری برای تحلیل داده‌های مالیاتی دارد و نتایج بهتری را ارائه می‌دهد. بررسی روشهای ریاضی و داده‌کاوی بر روی داده‌های مالیاتی می‌تواند کمکی به بهبود تشخیص رفتارهای مالیاتی داشته باشد و می‌توان برای انواع دیگری از داده‌های

معیار	روش	$n=۶$	$n=۷$	$n=۸$	$n=۹$	$n=۱۰$	$n=۱۱$	میانگین
دقت	kـمیانگین	۶۹/۵۸	۶۹/۶۱	۶۹/۵۵	۶۹/۵۵	۶۹/۶۶	۶۹/۶۶	۶۹/۶۰
	طیفی	۶۷/۵۴	۶۷/۵۷	۶۷/۴۷	۶۷/۵۰	۶۷/۴۶	۶۷/۶۱	۶۷/۵۲
	کاهش مختصات	۶۴/۶۱	۷۴/۷۵	۶۹/۶۶	۷۰/۶۷	۶۹/۶۶	۶۹/۶۶	۶۹/۸۴
	فاصله زیرفضایی	۷۰/۲۲	۶۹/۶۶	۶۶/۸۵	۶۹/۶۶	۶۹/۱۰	۶۹/۱۱	۶۹/۱
	چنددیدگاهه	۶۹/۶۹	۶۹/۶۳	۶۹/۶۳	۶۹/۶۴	۶۹/۶۹	۶۹/۶۹	۶۹/۶۶
خلوص	kـمیانگین	۳۷/۴۴	۳۷/۴۹	۳۷/۳۷	۳۷/۳۷	۳۷/۵۶	۳۷/۵۶	۳۷/۴۷
	طیفی	۴۰/۴۱	۴۰/۴۱	۴۰/۳۶	۴۰/۳۶	۴۰/۴۸	۴۰/۲۸	۴۰/۳۸
	کاهش مختصات	۶۴/۶۱	۷۴/۷۵	۶۹/۶۶	۷۰/۶۷	۶۹/۶۶	۶۹/۶۶	۶۹/۸۳
	فاصله زیرفضایی	۴۰/۳	۳۷/۶۲	۳۱/۹۳	۳۷/۶۲	۳۵/۷۴	۳۵/۷۴	۳۶/۴۹
	چنددیدگاهه	۴۳/۰۳	۴۲/۸۷	۴۲/۹۵	۴۲/۹۵	۴۳/۰۳	۴۲/۸۷	۴۲/۹۵
اطلاعات متقابل نرمال شده	kـمیانگین	۶۹/۵۸	۶۹/۶۰	۶۹/۵۵	۶۹/۵۵	۶۹/۶۶	۶۹/۶۶	۶۹/۶۰
	طیفی	۶۷/۵۴	۶۷/۵۶	۶۷/۴۷	۶۷/۵۰	۶۷/۴۵	۶۷/۶۰	۶۷/۵۲
	کاهش مختصات	۶۶/۲۹	۷۴/۷۵	۶۹/۶۶	۷۰/۶۷	۶۹/۶۶	۶۹/۶۶	۷۰/۱۱
	فاصله زیرفضایی	۷۰/۲۲	۶۹/۶۶	۶۶/۸۵	۶۹/۶۶	۶۹/۱	۶۹/۱	۶۹/۱
	چنددیدگاهه	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴
شاخص رند اصلاح شده	kـمیانگین	۷۰/۳۰	۷۰/۳۱	۷۰/۳۰	۷۰/۳۰	۷۰/۳۱	۷۰/۳	۷۰/۳
	طیفی	۶۸/۸۲	۶۸/۱۶	۶۸/۱۶	۶۸/۳۷	۶۸/۸۱	۶۸/۸۲	۶۸/۵۳
	کاهش مختصات	۶۹/۵۴	۷۱/۷	۷۰/۰۹	۷۲/۰۶	۷۰/۳	۷۱/۴۲	۷۰/۸۵
	فاصله زیرفضایی	۷۲/۱۳	۷۰/۳	۶۷/۹۷	۷۰/۳	۶۹/۶	۶۹/۶۵	۶۹/۹۸
	چنددیدگاهه	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴	۶۹/۵۴

جدول ۵: نتایج بر روی داده Wine

معیار	روش	$n=۳$	$n=۴$	$n=۵$	$n=۶$	$n=۷$	میانگین
دقت	kـمیانگین	۴۷/۳۷	۴۷/۲۰	۴۶/۵۴	۴۶/۵۹	۴۶/۵۲	۴۶/۸۵
	طیفی	۵۹/۰۹	۵۹/۲۷	۵۸/۷۴	۵۸/۷۲	۵۸/۴۱	۵۸/۸۵
	کاهش مختصات	۴۷/۰۸	۴۸/۶۹	۵۰/۸۴	۴۶/۹۶	۴۶/۰۷	۴۷/۹۳
	فاصله زیرفضایی	۴۲/۵۲	۴۷/۶۶	۴۸/۶۰	۴۵/۷۹	۴۹/۰۶	۴۶/۷۳
	چنددیدگاهه	۵۹/۲۳	۵۹/۰۳	۶۰/۰۱	۶۰/۱۳	۵۹/۲۱	۵۹/۵۲
خلوص	kـمیانگین	۵۷/۶۲	۵۷/۶۳	۵۸/۹۴	۵۹/۴۱	۵۹/۶۱	۵۸/۶۴۳
	طیفی	۶۴/۲۳	۶۴/۳۲	۶۳/۸۱	۶۳/۹۲	۶۳/۷۴	۶۴/۰۱
	کاهش مختصات	۴۷/۰۸	۴۸/۶۹	۵۰/۸۴	۴۶/۹۶	۴۶/۰۷	۳۸/۵۱
	فاصله زیرفضایی	۴۹/۱۴	۵۷/۰۶	۵۶/۹۳	۶۱/۰۶	۵۷/۰۷	۵۶/۲۵
	چنددیدگاهه	۶۴/۳۴	۶۴/۶۰	۶۵/۰۶	۶۵/۱۸	۶۴/۴۹	۶۴/۷۳
اطلاعات متقابل نرمال شده	kـمیانگین	۷۶/۷۲	۷۶/۸۰	۷۸/۲۸	۷۸/۸۷	۷۹/۱۲	۷۷/۹۶
	طیفی	۸۱/۲۹	۸۱/۷۷	۸۱/۱۲	۸۰/۸۴	۸۰/۸۱	۸۱/۱۷
	کاهش مختصات	۷۶/۵۹	۷۷/۲۲	۷۷/۵۵	۷۷/۸۵	۸۰/۳۵	۷۷/۹۱
	فاصله زیرفضایی	۶۵/۴۲	۷۶/۱۷	۷۷/۵۷	۸۱/۳۱	۷۷/۵۷	۷۵/۶۱
	چنددیدگاهه	۸۱/۷۸	۸۲/۴۰	۸۲/۵۹	۸۲/۷۸	۸۲/۲۲	۸۲/۳۵
شاخص رند اصلاح شده	kـمیانگین	۷۷	۷۶/۹۵	۷۶/۵۱	۷۷	۷۹/۱۱	۷۷/۳۱
	طیفی	۸۳/۲۱	۸۳/۱۴	۸۱/۸۴	۸۱/۷۷	۸۱/۸۴	۸۲/۳۶
	کاهش مختصات	۷۶/۹۵	۷۸/۲۲	۷۹/۳۹	۷۸/۹۹	۷۹/۲	۷۸/۵۵
	فاصله زیرفضایی	۶۹/۷۷	۷۷	۷۸/۲۷	۷۹/۲	۷۸/۳۶	۷۶/۵۲
	چنددیدگاهه	۸۱/۷۸	۸۱/۷۸	۸۲/۱۳	۸۳/۱۵	۷۵/۶۹	۸۰/۹۱

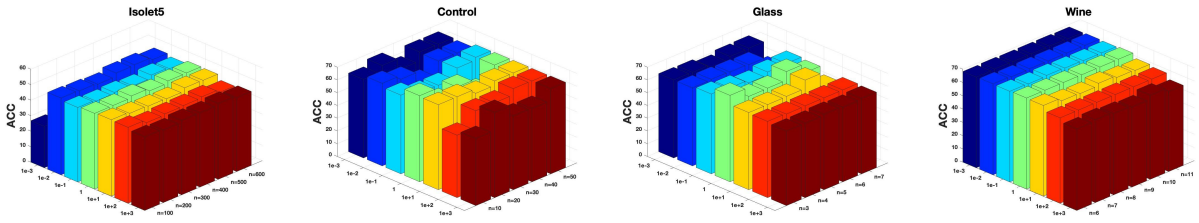
جدول ۶: نتایج بر روی داده Glass

معیار	روش	$n=۱۰$	$n=۲۰$	$n=۳۰$	$n=۴۰$	$n=۵۰$	میانگین
دقت	kـمیانگین	۳۸/۷۷	۴۲/۳۷	۴۶/۵۸	۵۱/۸۴	۵۷/۸۹	۴۷/۴۹
	طیفی	۶۰/۹۲	۶۰/۸۳	۶۱/۳۲	۶۰/۵۲	۶۰/۴۸	۶۰/۸۲
	کاهش مختصات	۶۲/۲۶	۵۸/۴۸	۵۹/۶۲	۵۶/۴۲	۵۸/۹۱	۵۹/۱۳
	فاصله زیرفضایی	۵۲/۳۳	۶۰/۱۷	۵۵/۵	۵۹/۶۷	۶۱	۵۷/۷۳
	چنددیدگاهه	۶۶/۶۷	۶۶/۶۷	۶۶/۶۷	۵۷/۳۳	۶۴/۹۲	۶۴/۴۵
خلوص	kـمیانگین	۳۴/۱۵	۳۹/۱۲	۴۱/۹۱	۴۸/۰۷	۵۳/۵۷	۴۳/۳۷
	طیفی	۷۰/۱۲	۶۹/۵۴	۷۰/۱۸	۶۹/۶۷	۶۹/۵۳	۶۹/۸۱
	کاهش مختصات	۶۲/۲۶	۵۸/۴۸	۵۹/۶۲	۵۶/۴۲	۵۸/۹۱	۵۹/۱۳
	فاصله زیرفضایی	۵۸/۵۶	۶۶/۷۴	۶۷/۹۷	۶۶/۳۹	۶۷/۰۱	۶۵/۳۳
	چنددیدگاهه	۷۶/۷۵	۷۶/۷۵	۷۶/۷۵	۶۵/۵۵	۷۱/۰۳	۷۳/۳۶
اطلاعات متقابل نرمال شده	kـمیانگین	۴۲/۹۶	۴۹/۲۸	۵۳/۸۵	۵۷/۸۴	۶۲/۴۸	۶۸/۱۲
	طیفی	۶۸/۳۰	۶۸/۰۸	۶۸/۴۱	۶۷/۹۴	۶۷/۸۷	۵۳/۲۸
	کاهش مختصات	۶۸/۲۶	۶۵/۸۸	۶۵/۸۲	۶۴/۴۵	۶۵/۲۴	۶۵/۹۳
	فاصله زیرفضایی	۶۱/۱۷	۶۴/۵	۶۵/۳۳	۶۴/۵	۶۵/۳۳	۶۴/۱۶
	چنددیدگاهه	۷۴	۷۴	۷۴	۶۵/۷۵	۷۰/۶۷	۷۱/۶۸
شاخص رند اصلاح شده	kـمیانگین	۷۵/۷۹	۷۷/۷۷	۷۹/۱۸	۸۰/۴۷	۸۲/۲۱	۷۹/۰۸
	طیفی	۸۳/۵۳	۸۷/۰۶	۸۶/۹۲	۸۶/۹۴	۷۷/۰۱	۸۴/۲۹
	کاهش مختصات	۸۶/۵۴	۸۵/۹۱	۸۴/۱۷	۸۵/۴۹	۸۵/۸۵	۸۵/۵۹
	فاصله زیرفضایی	۸۲/۸۱	۸۵/۳	۸۵/۵۹	۸۵/۱۵	۸۵/۳۱	۸۴/۸۳
	چنددیدگاهه	۸۸/۴۸	۸۸/۴۸	۸۸/۴۸	۸۸/۴۸	۸۴/۷۶	۸۷/۷۴

جدول ۷: نتایج بر روی داده Control

معیار	روش	$n=۱۰۰$	$n=۲۰۰$	$n=۳۰۰$	$n=۴۰۰$	$n=۵۰۰$	$n=۶۰۰$	میانگین
دقت	kـمیانگین	۱۵/۰۹	۱۷/۸۲	۱۸/۴۳	۱۹/۷۷	۲۳/۰۱	۲۳/۳۶	۱۹/۵۸
	طیفی	۲۷/۶۷	۲۷/۶۲	۲۷/۹۱	۲۷/۶۹	۲۷/۵۷	۲۷/۸۰	۲۷/۷۱
	کاهش مختصات	۴۸/۳۵	۴۸/۴۳	۵۱/۵۵	۵۲/۴۰	۵۰/۳۶	۵۰/۱۵	۵۰/۲۱
	فاصله زیرفضایی	۴۹/۸۴	۴۸/۲۴	۴۵/۳۵	۴۹/۵۸	۵۰/۱۶	۵۱/۸۳	۴۹/۱۷
	چنددیدگاهه	۴۹/۳۵۲	۵۱/۰۳۹	۵۰/۷۲	۴۸/۹۹	۴۹/۲۱	۴۹/۶۲	۵۰/۵۷
خلوص	kـمیانگین	۲۸/۶۱	۳۴/۷۵	۳۵/۴۳	۳۶/۰۱	۳۹/۰۶	۳۹/۶۹	۳۵/۵۹
	طیفی	۴۴/۱۵	۴۴/۲۷	۴۴/۵۵	۴۴/۲۷	۴۴/۰۶	۴۴/۲۸	۴۴/۲۶
	کاهش مختصات	۴۸/۳۵	۴۸/۴۳	۵۱/۵۵	۵۲/۴۰	۵۰/۳۶	۵۰/۱۵	۵۰/۲۱
	فاصله زیرفضایی	۶۷/۸۳	۶۸/۴۹	۶۴/۳۹	۶۷/۵۸	۶۸/۷۱	۶۹/۴۹	۶۷/۷۵
	چنددیدگاهه	۶۷/۶۲	۶۸/۶۳	۶۸/۶	۶۷/۷	۶۷/۴۵	۶۸/۱۷	۶۸/۰۳
اطلاعات متقابل نرمال شده	kـمیانگین	۱۷/۲۱	۲۱/۴۷	۲۲/۴۰	۲۳/۵۷	۲۶/۲۶	۲۶/۷۲	۲۲/۹۴
	طیفی	۳۰/۳۶	۳۰/۳۰	۳۰/۵۸	۳۰/۳۵	۳۰/۲۶	۳۰/۴۵	۳۰/۳۸
	کاهش مختصات	۵۲/۳۴	۵۲/۴۹	۵۵/۲۴	۵۶/۸۳	۵۴/۲۴	۵۴/۷۸	۵۴/۳۲
	فاصله زیرفضایی	۵۴/۸۴	۵۳/۴۳	۵۰/۷۴	۵۳/۰۵	۵۲/۷۳	۵۳/۶۹	۵۳/۰۸
	چنددیدگاهه	۵۴/۱۸	۵۴/۹	۵۴/۵۶	۵۳/۰۳	۵۳/۶۷	۵۴/۰۳	۵۴/۴۹
شاخص رند اصلاح شده	kـمیانگین	۹۰/۴۶	۹۰/۰۸	۸۹/۸۱	۹۰/۰۶	۹۲/۳۴	۹۲/۷۴	۹۰/۹۲
	طیفی	۶۶/۸۳	۷۲/۳۰	۷۱/۴۲	۷۴/۳۲	۶۰/۱۷	۷۱/۴۶	۶۹/۴۱
	کاهش مختصات	۹۵/۲۶	۹۴/۹۵	۹۵/۵۱	۹۵/۳۲	۹۵/۵۲	۹۴/۹۷	۹۵/۲۵
	فاصله زیرفضایی	۹۵/۴۴	۹۵/۳۲	۹۴/۷۶	۹۵/۳۶	۹۵/۵۱	۹۵/۵۵	۹۵/۳۲
	چنددیدگاهه	۹۴/۶۱	۹۴/۱۸	۹۵/۳	۹۴/۳۲	۹۴/۷۷	۹۲/۹۵	۹۴/۳۵

جدول ۸: نتایج بر روی داده ۵Isolet



(د) (ج) (ب) (آ)

شکل ۲: تاثیر پارامتر β روی شاخص دقت

مالیاتی نیز روشهای جدیدی ارائه نمود. اغلب روشهایی که قبلا ارائه شده است روشهای بانظارت می‌باشد،اما برای حجم زیادی داده مالیاتی این روشها ممکن است کارا نباشد و روش ارائه شده، به عنون یك روش بدون نظارت می‌تواند ارزش بیشتری داشته باشد.

مراجع

[۱] م. باقرپور ولاشانی، م. باقری، ح. خادم و ر. حسینی‌پور، بررسی عوامل مالی و غیر مالی مؤثر بر گریز مالیاتی با استفاده از تکنیک‌های داده‌کاوی: صنعت خودرو و ساخت قطعات، *مطالعات تجربی حسابداری مالی*، ۱(۳۴) (۱۳۹۱) ۱۰۳-۱۲۸.

[۲] م. باقرپور ولاشانی، م. ساعدی، ع. مشکانی و م. باقری، پیش‌بینی گزارش حسابرس مستقل در ایران: رویکرد داده‌کاوی، *دهمین همایش حسابداری ایران*، دانشگاه الزهرا، ۱۳۹۱.

[۳] م. سامعی‌راد و ا. شاه‌بهرامی، بهبود کارایی الگوریتم‌های تشخیص تقلب مالیاتی با استفاده از الگوهای پردازش موازی، *پژوهشنامه مالیات*، ۲۴(۲۹) (۱۳۹۵) ۱۱-۳۲.

[۴] ب. سهرابی، ا. رئیسی وانانی و و. قاونی شیشوان، ارزیابی مالیات عملکرد شرکت‌ها و تحلیل روندهای مالیاتی با استفاده از الگوریتم‌های داده‌کاوی، *تحقیقات مالی*، ۱۷(۴۰) (۱۳۹۴) ۲۱۹-۲۳۸.

[5] R. B. Asha, K. R. Suresh Kumar, Credit card fraud detection using artificial neural network, *Global Transitions Proceedings*, (2021). doi: https://doi.org/10.1016/j.gltp.2021.01.006.

[6] Y. Cai, C. Hangjun, P. Baicheng, L. Man-Fai, L. Cheng, W. Shiping, Projected cross-view learning for unbalanced incomplete multi-view clustering, *Information Fusion*, **105** , (2024), 102245.

[7] D. Cai, X. He, J. Han, Document clustering using locality preserving indexing, *IEEE Transactions on Knowledge and Data Engineering*, **17** (12), (2005), 1624–1637.

[8] C. Cui, R. Yazhou, P. Jingyu, L. Jiawei, P. Xiaorong, W. Tianyi, S. Yutao, H. Lifang, A novel approach for effective multi-view clustering with information-theoretic perspective, *Advances in Neural Information Processing Systems*, **36**, (2024).

[9] C. H. Chen, Unsupervised margin-based feature selection using linear transformations with neighbor preservation, *Neurocomputing*, **171**, (2016), 1354–1366.

[10] G. Chandrashekar, F. Sahin, A survey on feature selection methods, *Computers and Electrical Engineering*, **40**, (2014), 16–28.

[11] M. Dong, L. Yao, X. Wang, B. Benatallah, Ch. Huang, X. Ning, Opinion fraud detection via neural autoencoder decision forest, *Pattern Recognition Letters*, **132**, (2020), 21–29.

[12] U. Fang, L. Man, L. Jianxin, G. Longxiang, J. Tao, Z. Yanchun, A comprehensive survey on multi-view clustering, *IEEE Transactions on Knowledge and Data Engineering*, **35**(12), (2023), 12350–12368.

[13] X. Fang, Y. Xu, X. Li, Z. Fan, H. Li, Y. Chen, Locality and similarity preserving embedding for feature selection, *Neurocomputing*, **128**, (2014), 304–315.

[14] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, *Advances in Neural Information Processing Systems*, **18**, (2006), 507–514.

[15] L. Hubert, P. Arabie, Comparing partitions, *Journal of Classification*, **2** (1), (1985), 193–218.

[16] H. Htun, M. Biehl, N. Petkov, Survey of feature selection and extraction techniques for stock market prediction, *Financial Innovation*, **9**(1), (2023), 26.
[17] Z. Kang, Z. Xinjia, P. Chong, Z. Hongyuan, Z. Joey, P. Xi, C. Wenyu, X. Zenglin, Partition level multiview subspace clustering, *Neural Networks*, **122**, (2020), 279–288.
[18] S. Karami, F. Saberi-Movahed, P. Tiwari, P. Marttinen, S. Vahdati, Unsupervised feature selection based on variance–covariance subspace distance, *Neural Networks*, **166**, (2023), 188–203.
[19] N. N. Karnik, J. M. Mendel, Operations on type-2 fuzzy sets, *Fuzzy Sets and Systems*, **122**, (2001), 327–348.
[20] Y. Manolopoulos , Data mining techniques for the detection of fraudulent financial statements, *Expert Systems with Applications*, **32**(4), (2007), 995–1003.
[21] F. Nie, J. Li, X. Li, Self-weighted multiview clustering with multiple graphs, *International Joint Conferences on Artificial Intelligence* (2017), 2564–2570.
[22] F. Sadjadi, M. Jamshidi, Z. Kang, Multi-view subspace clustering using drop out technique on points, *International Journal of Machine Learning and Cybernetics*, **15**(5), (2024), 1841–1854.
[23] F. Sadjadi, V. Torra, M. Jamshidi, Preprocessed Spectral Clustering with Higher Connectivity for Robustness in Real-World Applications, *International Journal of Computational Intelligence Systems*, **17**(1), (2024), 86.
[24] J. Vanhoeyvelda, D. Martensa, B. Peetersb, Value-Added tax fraud detection with scalable anomaly detection techniques, *Applied Soft Computing*, **86**, (2020). doi:https://doi.org/10.1016/j.asoc.2019.105895.
[25] U. Von Luxburg, A tutorial on spectering, *Statistics and Computing*, **17**(4), (2007), 395–416.
[26] S. Wang, W. Pedrycz, Q. Zhu, W. Zhu, Unsupervised feature selection via maximum projection and minimum redundancy, *Knowledge-Based Systems*, **75**, (2015), 19–29.
[27] R.S. Wu, C.S. Ou, S.I. Chang, D.C. Yen, Using data mining technique to enhance tax evasion detection performance, *Expert Sestems With Applications*, **39**, (2012), 8769–8777.
[28] L. Xu, R. Wang, F. Nie, X. Li, Efficient top-k feature selection using coordinate descent method, *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, (2023). 10594–10601.
[29] C. Tang, Z. Zheng, Z. Wei, L. Xinwang, Z. Xinzhong, Z. En, Unsupervised feature selection via multiple graph fusion and feature weight learning, *Science China Information Sciences*, **66**(5), (2023), 152101.
[30] Ch. Yan, M. Li , W. Liu , M. Qi, Improved adaptive genetic algorithm for the vehicle insurance fraud identification model based on a BP neural network, *Theoretical Computer Science*, **817**, (2020), 12–23.
[31] W. Zhou, G. Kapoor, Detecting evolutionary financial statement fraud, *Decision Support Systems*, **50**(3), (2011), 570–575.