مهندسی صنایع و مدیریت شریف
دوره‌ی ۱ - ۲۹، شماره‌ی ۱، ص. ۱۱۳-۱۲۱، (یادداشت فنی)

Research Note

ارائه‌ی مدلی برای کمینه‌سازی مجموع متوسط فواصل درون‌خوشه‌یی در طبقه‌بندی مشتریان

ابوالفضل کاظمی * **(استادیار)**
زهرا ضیایی کوچصفهانی (کارشناس ارشد)
دانشکده‌ی مهندسی صنایع و مکانیک، دانشگاه آزاد اسلامی واحد قزوین

در سال‌های اخیر اهمیت داده‌ها به‌عنوان منابع دارای پتانسیل اطلاعاتی بسیار بالا به‌نحو گسترده‌یی مورد توجه قرار گرفته شده است. داده‌کاوی با استخراج و کشف سریع و دقیق اطلاعاتِ با ارزش و پنهان از پایگاه داده‌ها به‌منظور تصمیم‌گیری و پشتیبانی تصمیم از جمله اموری است که هر کشور، سازمان و شرکتی به‌منظور توسعه علمی، فنی و اقتصادی خود به آن نیاز دارد. با توجه به ضرورت استفاده از فنون داده‌کاوی ــ خصوصاً خوشه‌بندی ــ در این نوشتار یک مدل ریاضی براساس رویکرد خوشه‌بندی ارائه‌می‌شود که در خوشه‌بندی مشتریان شرکت صنعتی پارس‌خزر کاربرد دارد. مسئله‌ی خوشه‌بندی به‌صورت مدل ریاضی با هدف کمینه‌سازی مجموع متوسط فواصل درون‌خوشه‌یی در طبقه‌بندی مشتریان فرمول‌بندی می‌شود که در تمامی موارد آزمایش شده، با بهبودبخشی شدید در فواصل درون‌خوشه‌یی همراه است. عملکرد این شیوه در یک مسئله‌ی واقعی آزموده شده و تحلیل نتایج حاکی از کارایی محاسبات این شیوه است.

واژگان کلیدی: داده‌کاوی، خوشه‌بندی، روش‌های خوشه‌بندی سلسله‌یی و غیرسلسله‌یی، مدیریت ارتباط با مشتری، الگوریتم k-means.

abkaazemi@gmail.com
shziaeem@yahoo.com

۱. مقدمه

امروزه رشد سریع فنّاوری اطلاعات در بخش‌های مختلف کسب و کار، مزایای فراوانی برای صاحبان بنگاه‌های اقتصادی به ارمغان آورده است؛ حوزه‌ی بازاریابی نیز به‌عنوان زیربخش اصلی سازمان از این قاعده مستثنی نیست. فنّاوری اطلاعات به بروز تغییراتی در روش‌های بازاریابی سازمان‌ها و ایجاد حجم وسیعی از داده‌های مشتریان انجامیده است. با توجه به رقابت فشرده در بازار و گزینه‌های متنوع محصولات و خدمات که پیش روی مشتریان قرار دارد، تحلیل رفتار مشتریان عامل بسیار مهمی برای بقای بنگاه‌ها تلقی می‌شود. با توجه به موارد یادشده، یکی از مؤثرترین ابزارهای بررسی رفتار مشتریان استفاده از تکنیک‌های داده‌کاوی است.

در سال‌های اخیر، شرکت‌ها بر شناخت نیازها و توقعات مشتریان و نیز برگروه‌بندی مشتریان فعلی و احتمالی ــ با هدف بهبودبخشی بازدهی راهبردهای بازاریابی و افزایش سهم بازار خود ــ متمرکز شده‌اند. فراوانی مجموعه‌های بزرگ داده و نیاز به استخراج دانش نهفته در آن‌ها، به ایجاد الگوریتم‌هایی برای تشخیص الگوهای ناشناخته‌ی موجود در مجموعه داده‌ها انجامیده است. آنالیز خوشه‌بندی یک تکنیک داده‌کاوی است که با هدف شناسایی گروه‌های مشابه از نقطه‌نظر معیارهای معین تشابه، توسعه یافته است. رویکردهای زیادی نسبت به مسئله‌ی خوشه‌بندی وجود دارد، شامل بهینه‌سازی براساس روش‌های دربرگیرنده‌ی مدل‌های ریاضی برای تشکیل طرح‌های خوشه‌بندی مؤثر و پرمعنا. الگوریتم‌هایی دقیق و اکتشافی برای این مدل‌ها مطرح شده است.

تحقیقات درخصوص مباحث ریاضی داده‌کاوی و خوشه‌بندی، در سال ۱۹۷۱ به دو طریق انجام شد: الف) با کمینه‌سازی مجموع مجذورهای میان‌خوشه‌یی؛ ب) با کمینه‌سازی بیشترین فاصله‌ی میان‌خوشه‌یی. در سال‌های بعد (۲۰۰۳) الگوریتم میانگین‌های k (k-means) ارائه شد؛ این الگوریتم شیوه‌یی فزاینده برای خوشه‌بندی است که به‌صورت پویا در زمانی معین از طریق یک روند جست‌وجوی عمومی درترمینانی ــ متشکل از N اجزاء از الگوریتم میانگین‌های k ــ با شروع از موقعیت‌های مناسب بر مرکز خوشه می‌افزاید. ملاحظه‌ی دیگر در خصوص خوشه‌بندی آن است که انتظار می‌رود خوشه‌های حاصله با توجه به خصوصیاتی معین، متجانس و فشرده باشند. در این نوشتار با استفاده از مدل ریاضی برای تقسیم‌بندی مجموعه داده به دسته‌های انحصاری اقدام شده است؛ در این مدل فرض شده تعداد خوشه‌های مطلوب k یک الویت است. عملکرد حقیقی مدل «کمینه‌سازی مجموع متوسط فواصل درون‌خوشه‌یی در طبقه‌بندی مشتریان با هدف کسب خوشه‌هایی به یک اندازه فشرده با کم‌ترین فاصله‌ی درون‌خوشه‌یی» است.

در این نوشتار از داده‌های حقیقیِ حاصل از شرکت صنعتی پارس‌خزر در آزمایشات محاسباتی استفاده شده است. این شرکت در بهمن ماه ۱۳۴۷ در شهر رشت، با نام شرکت صنعتی پارس‌توشیبا و به‌منظور ساخت انواع لوازم خانگی برقی شروع به فعالیت کرد. در اردیبهشت ماه سال ۱۳۵۲، نوع شرکت به سهامی خاص

---

* نویسنده مسئول
تاریخ: دریافت ۱۳۸۹/۶/۶، اصلاحیه ۱۳۹۰/۸/۱۵، پذیرش ۱۳۹۰/۱۱/۱.

تبدیل شد و در سال ۱۳۶۱ پس از واگذاری سهام متعلق به توشیبای ژاپن، نام آن به شرکت صنعتی پارس‌خزر تغییر یافت.

این شرکت شبکه‌یی گسترده، مرکب از حدود ۳۵۰ تعمیرگاه مجاز سرویس و خدمات پس از فروش در اقصی نقاط ایران دارد. در پی کسب فرصت‌هایی برای بازاریابی رابطه با مشتری، این شرکت تمایل دارد مشتریانش را براساس فاکتورهای دادوستد و موقعیت جغرافیایی طبقه‌بندی کند. تفسیری از خوشه‌بندی با استفاده از الگوریتم خوشه‌بندی روی داده‌های شرکت صنعتی پارس‌خزر در بخش ۴ ارائه خواهد شد.

در این نوشتار، در بخش ۲ براساس روش‌های خوشه‌بندی، تحقیقات انجام‌شده مرور می‌شود و در بخش ۳ نیز مدل پیشنهادی ارائه می‌شود. سپس با طرح مثالی نمادین در بخش ۴، حل مدل را بر روی آن شرح داده و نتایج را با مدل ارائه‌شده‌ی قبلی مقایسه می‌کنیم. نهایتاً در بخش ۵ نتایج مطالعه‌ی حاضر، و نیز بحث و بررسی آرا برای تحقیقات آتی در زمینه‌ی بهینه‌سازی و خوشه‌بندی ارائه می‌شود.

۲. مرور مقالات و تجارب کسب شده

شاید بتوان گفت نخستین گزارش درمورد داده‌کاوی تحت عنوان «شبیه‌سازی فعالیت داده‌کاوی» در سال ۱۹۸۳ ارائه شده است. هم‌زمان، پژوهش‌گران و متخصصین علوم رایانه، آمار، هوش مصنوعی، یادگیری ماشین و... نیز به پژوهش در این زمینه و زمینه‌های مرتبط با آن پرداختند.[۱]

در سال ۲۰۰۲ نیز در مقاله‌ی «طبقه‌بندی مشتریان با استفاده از داده‌کاوی» از بعضی روش‌های داده‌کاوی برای شناسایی مشتریان (مشتریان یک بانک و یک کلوپ کتاب) استفاده شد.[۲] برای توصیف این روش در مجموعه‌ی بانک، از روش خوشه‌بندی به‌منظور یافتن ساختارها و الگوهای داخل سیستم استفاده شده است. داده‌کاوی(DM)[۱] بخشی لاینفک از مطالعات مدیریت ارتباط با مشتری (CRM)[۲] بوده، با این فرضیه که شرکت‌ها می‌توانند درصورت شناخت خصوصیات و علائق مشتریان خود، با آن‌ها رابطه‌یی موفقیت‌آمیز برقرار سازند.[۳]

برخی از محققین به مرور داده‌کاوی، مصارف آن در صنعت و تکنیک‌های مصرفی تحت این عنوان پرداخته‌اند و رابطه و تعامل بین کاربردهای داده‌کاوی و مدیریت ارتباط با مشتری (CRM) را از جوانب مختلف شرح داده‌اند.[۴] عده‌یی نیز نقش بهینه‌سازی در داده‌کاوی را مورد بررسی قرار داده، و عنوان کردند که داده‌کاوی و بهینه‌سازی می‌توانند برای توسعه‌ی روابط نوین CRM ــ روابطی مثل بیشینه‌سازی میزان طول عمر مشتری، آنالیز مشتری و تعاملات مشتریان ــ به هم کمک کنند.[۵]

با توجه به اهمیت روش‌های داده‌کاوی در مدیریت ارتباط با مشتری، مرور ادبیات و دسته‌بندی جامع باعث ایجاد برنامه‌یی جامع و کامل شده است. در این مورد ابتدا مرور ادبیات علمی درخصوص کاربرد روش‌های داده‌کاوی در CRM مورد بررسی قرارگرفته و سپس یک‌سری اطلاعات در دوره زمانی ۲۰۰۰ تا ۲۰۰۶، شامل ۲۴ مقاله تهیه شده که طی آن ۹۰۰ نویسنده ارتباط مستقیم بین روش‌های داده‌کاوی و CRM را شناسایی کرده‌اند.[۶]

«خوشه‌بندی» یکی از مهم‌ترین موارد تحقیق در حوزه‌ی داده‌کاوی، و به‌معنی ایجاد گروه‌هایی است که اجزاء داخل هر گروه بیشترین تشابه، و اجزاء در گروه‌های مجزا هیچ‌گونه شباهتی نداشته باشند. خوشه‌بندی یک روش یادگیری بدون نظارت است که مهم‌ترین مزیت آن ــ ناشی از ساختار و ویژگی‌هایش ــ امکانِ گروه‌بندی مستقیمِ داده‌های بزرگ با کم‌ترین یا بدون اطلاع قبلی است. الگوریتم خوشه‌بندی در حوزه‌های زیادی کاربرد دارد و همین نکته باعث شده تا تحقیقاتی درمورد محاسبه‌ی پیچیدگی بین الگوریتم‌های خوشه‌بندی k-means و k-medoids برای توزیع نرمال و یکنواخت داده‌های نقطه‌یی انجام شود.[۷]

روش‌های خوشه‌بندی سلسله‌یی (HC)[۳]، به‌صورت یک سری عملیات تقسیمی پیش می‌رود که شروع‌شان با یک خوشه‌ی واحد در برگیرنده‌ی تمامی موارد (مثال‌ها) و پایان‌شان، با رسیدن به معیار از پیش تعیین‌شده خاتمه‌بخشی، مشخص می‌شود. بنابراین، روش‌های HC برای ایجاد تفاسیری بر دسته داده کاربرد دارند، و حل HC را می‌توان به‌عنوان یک ورودی برای روش خوشه‌بندی غیرسلسله‌یی (NHC)[۴] و به‌منظور بهبودبخشی حل خوشه‌ی حاصله به کار برد.[۸]

روش خوشه‌بندی غیرسلسله‌یی (NHC)، موسوم به خوشه‌بندی تقسیمی (تفکیکی)، به حالتی اطلاق می‌شوند که در آن تعداد تقسیمات معلوم است. در این روش معمولاً داده ابتدا به k خوشه تقسیم می‌شود، و پس از آن برای تمامی انتقالات ممکن نقاط داده‌یی مابین خوشه‌های تشکیل شده پیش می‌رود تا این که معیار توقف تحقق یابد. بنا بر اظهار محققین، در این شیوه هر خوشه را می‌توان توسط مرکز خوشه (k-means) یا با یک مورد واقع در مرکز خوشه (میانه k) نمایش داد. شاید پرمصرف‌ترین الگوریتم خوشه‌بندی الگوریتم k-means باشد که در عمل خوب کار می‌کند، هرچند پیچیدگی زمان در بدترین حالت به‌طور تصاعدی بالا می‌رود.[۹]

در بررسی‌های اخیر یک الگوریتم جدید خوشه‌بندی سلسله‌یی برای داده‌های فازی (FHCA) مورد بررسی قرارگرفته که در آن، خوشه‌بندی فازی اعداد با استفاده از روش‌های سلسله‌مراتبی انجام می‌گیرد. در همه‌ی روش‌های مشابه قبلی، روش (FCM)[۵] بسط‌یافته (خوشه‌بندی فازی)، داده‌های فازی را پوشش می‌دهد. وجه تمایز این روش، کاربرد الگوریتم خوشه‌بندی سلسله‌یی در خوشه‌بندی داده‌های فازی است. مهم‌ترین ویژگی این الگوریتم مقاومت بالای آن در مقابل داده‌های مغشوش است.[۱۰]

در بررسی روش‌های عدد صحیح با هدف کمینه‌سازی مجموع فواصل میان هر نقطه‌ی داده و مرکز خوشه‌یِ منطبق با آن، یا به‌عبارتی بررسی اجرای الگوریتم k-means در مدل بهینه‌سازی، محققین روش‌های خوشه‌بندی را با تمرکز بر فرمولاسیون‌های ریاضی به سه دسته تقسیم کردند: روش‌های مبتنی بر مقیاس (متریک)، روش‌های مبتنی بر مدل، و روش‌های مبتنی بر تفکیک.[۱۱]

از سوی دیگر، در برخی از الگوریتم‌های NHC ــ مثل روش اکتشافی جهانی k-means ــ خوشه‌ها به‌صورت صعودی تشکیل می‌شوند. طراحان این الگوریتم بر این باورند که حل دسته‌ی k بهینه را می‌توان از حل بهینه‌ی مسئله‌ی دسته‌ی (k-۱) با استفاده از یک جست‌وجوی محلی روی تمام مجموعه داده‌ها به دست آورد، طوری که الگوریتم k-means مکرراً انجام شود تا بهترین مرکز دسته‌ی kام ممکنه حاصل شود. اگرچه ممکن است این الگوریتم به راه‌حل‌های امیدبخشی بینجامد، متضمن رسیدن به بهترین حل کلی نیست.[۱۲]

در سری دیگری از تحقیقاتِ انجام‌شده، معیار کمینه‌سازی مجموع مجذورهای میان گروه‌ها در نظرگرفته شده است. همچنین با بهره‌گیری و بسط برخی نتایج، یک فرمول برنامه‌نویسی خطی عدد صحیح برای مسئله‌یی که در آن اجزاء، نقاطی از یک فضای تک‌بعدی یا p بعدی هندسه‌ی اقلیدسی هستند، ارائه شد. الگوریتم پویا و مؤثر برای موردی عرضه می‌شود که در آن بتوان اجزاء را نقاطی روی خط حقیقی در نظرگرفت؛ این الگوریتم کارآمدتر از الگوریتم‌های فعلی برنامه‌نویسی عدد صحیح است. به‌علاوه در این الگوریتم، از پیش مشخص شدن تعداد خوشه‌بندی‌ها ضرورت ندارد، بلکه راه حلی برای هر تعداد خوشه‌بندی ارائه می‌شود. زمانی که اجزاء نقاطی از یک فضای اقلیدسی p بعدی باشند، فرمول برنامه‌نویسی خطی عدد صحیح برای مسئله ارائه می‌شود. این فرمول اگرچه تحت شرایط معینی معتبر است، این مزیت را

دارد که علاوه برکاستن از تعداد قیود در حدی قابل توجه، مسئله را به تفکیک‌بندی مجموعه با تعدادی تقسیمات یا مجموعه‌های از پیش مشخص شده کاهش می‌دهد. پژوهش‌گرانی که این مسئله را بررسی کرده‌اند، اظهار داشته‌اند که حل مسئله‌های این‌چنینی نیازمند زمان محاسباتی مطلوبی است.[۱۳] در بخش بعد مدل ریاضی پیشنهادی برای مسئله‌ی خوشه‌بندی ارائه خواهد شد.

۳. مدل ریاضی پیشنهادی برای مسئله‌ی خوشه‌بندی

تحلیل‌های خوشه‌بندی زیادی (نظیر کمینه‌سازی مجموع مجذورهای میان‌خوشه‌یی و بررسی اجزای متعلق به فضای اقلیدسی p بُعدی) توسط رائو انجام شده است. اولین مورد در نظرگرفته شده در بررسی اجزای متعلق به فضای اقلیدسی p بُعدی این است که وقتی اجزاء به یک فضای اقلیدسی p بعدی تعلق داشته باشند مسئله پیچیده‌تر می‌شود.

معیار مصرفی در این مدل، کمینه‌سازی مجموع متوسط مجذور فواصل میان‌خوشه‌یی است. به‌طور عام، فرمول یک مسئله‌ی برنامه نویسی ۰ و ۱ غیرخطیِ کسری است که حل آن در حالت عمومی خیلی سخت است.

در مدل پیشنهادی دیگر توسط ادواردز و کاوالی، تعداد خوشه‌ها برابر ۲ است، و برای تحلیل خوشه‌بندی آن اجزاء به دو خوشه با بیشترین فشردگی تقسیم می‌شود، و این فرایند به‌صورت متوالی تکرار می‌شود. این امر حاکی از آن است که در هر مرحله از فرایند M=۲ است. مسئله‌ی برنامه‌نویسی ۰ و ۱ غیرخطی کسری فاقد قیدهای ضمنی است، و یک مسئله‌ی برنامه‌نویسی boolean است.

مدل‌های دیگر، کمینه‌سازی کل فاصله‌ی میان‌خوشه‌یی و کمینه‌سازی بیشترین فاصله‌ی میان‌خوشه‌یی است که هر دو مدل مسئله برنامه‌نویسی خطی عدد صحیح است، اما تعداد قیود سریعاً برحسب N و M افزایش می‌یابد و لذا این فرمول از حیث محاسباتی فقط برای مقادیر کوچک N و M مفید است. باید خاطرنشان ساخت که اگر این مسئله را به‌صورت یک مسئله‌ی برنامه‌نویسی خطی حل کنیم، بدون آن که متغیرها به اعداد صحیحی محدود شوند، آنگاه یک راه‌حل $x_{ik} = ۱/M$ برای هر i و k است. لذا هیچ‌وقت آن‌قدر خوش‌شانس نیستیم که به یک راه حل عدد صحیح مثل حل برنامه‌نویسی خطی، دست یابیم. البته وقتی تعداد خوشه‌ها M برابر ۲ باشد، این مشکل را می‌توان به‌صورت مؤثری حل کرد.

در تمامی تحقیقات گذشته فرمول‌های مختلفی درخصوص فواصل درون‌خوشه‌یی ارائه شده است. با توجه به توضیحات ارائه‌شده، در این مقاله مدل پیشنهادی «کمینه‌سازی مجموع متوسط فواصل درون‌خوشه‌یی در طبقه‌بندی مشتریان» با مدل مشابه «کمینه‌سازی مجموع متوسط مجذور فواصل میان‌خوشه‌یی» ارائه‌شده توسط رائو، مورد بررسی قرار گرفته است. در مدل رائو معیار مصرفی، کمینه‌سازی مجموع متوسط مجذور فواصل میان‌خوشه‌یی طبق فرمول ۴ است. این یک مسئله‌ی برنامه‌نویسی ۰ و ۱ غیرخطیِ کسری است که حل آن در حالت عمومی خیلی سخت است. البته اگر تعداد اجزاء در هر خوشه از قبل مشخص باشد می‌توان تابع حقیقی آن را تصحیح کرد، با این حال هنوز یک مسئله‌ی برنامه‌نویسی ۰ و ۱ غیرخطی با مجموعه شرط‌های جدید است. در این حالت دست‌کم دو روش ممکن برای حل این مسئله وجود دارد؛ روش اول تلقی آن به‌منزله‌ی یک مسئله‌ی برنامه‌نویسی غیرخطی boolean است، و روش دیگر خطی‌کردن تابع حقیقی است. در خطی‌کردن تابع حقیقی تعداد قیود بیشتر می‌شود، و مسئله‌ی حاصله را باید با یکی از شیوه‌های شناخته‌شده برای برنامه‌نویسی عدد صحیح خطی حل کنیم.

مراحل انجام کار در این نوشتار بدین‌ترتیب است که ابتدا فرمول پیشنهادی ارائه شده و سپس برای صحه‌گذاری، مراحل مقایسه‌ی خروجی آن با خروجی فرمول مشابه استفاده‌شده در تحلیل خوشه‌بندی در نرم‌افزار MATLAB (R۲۰۰۷b) ۷.۵.۰ انجام‌می‌شود. نتایج این مقایسه در جداول ۱ تا ۴ و نمودارهای ۱ تا ۴

جدول ۱. مقایسه‌ی نتایج فرمول الف و ب براساس کم‌ترین فاصله برای M=۲.

تعداد نمونه	کم‌ترین فاصله	
	الف	ب
۳	۲	۰/۵
۴	۲	۰/۶۶
۵	۴	۱
۶	۱۱/۱۴	۱/۸۶
۷	۱۱	۱/۸۷
۸	۲۱/۷۷	۲/۷۷
۹	۴۵/۰۹	۴/۴۵
۱۰	۴۲	۴/۱۰

جدول ۲. مقایسه‌ی نتایج فرمول الف و ب براساس کم‌ترین فاصله برای M=۳.

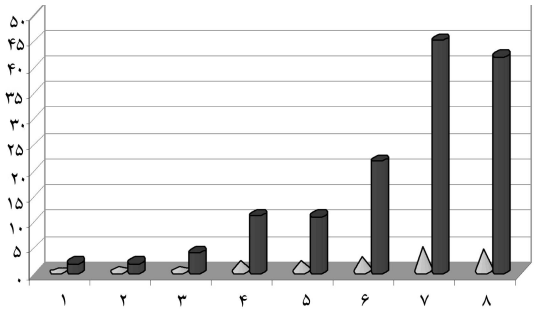
تعداد نمونه	کم‌ترین فاصله	
	الف	ب
۴	۱/۶۲	۰/۳۷
۵	۴	۰/۶۶
۶	۹/۱۱	۱/۰۴
۷	۹/۷۰	۱/۱۰
۸	۱۸/۳۳	۱/۶۱
۹	۴۵	۲/۹۲
۱۰	۴۲	۲/۷۳

جدول ۳. مقایسه‌ی نتایج فرمول الف و ب براساس کم‌ترین فاصله برای M=۴.

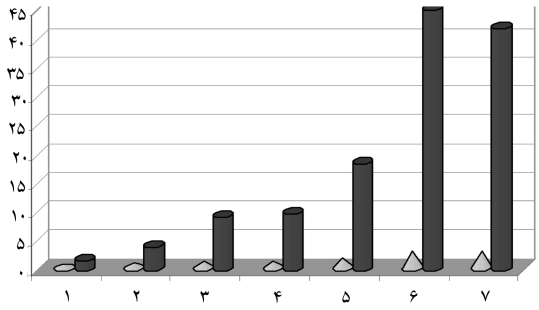
تعداد نمونه	کم‌ترین فاصله	
	الف	ب
۵	۳/۸۰	۰/۴۵
۶	۹/۱۱	۰/۷۸
۷	۹/۷۰	۰/۸۳
۸	۱۶	۱/۰۷
۹	۳۹/۱۶	۱/۹۵
۱۰	۴۲	۲/۰۵

جدول ۴. مقایسه‌ی نتایج فرمول الف و ب براساس کم‌ترین فاصله برای M=۵.

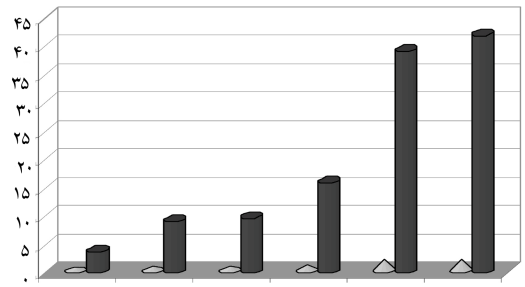
تعداد نمونه	کم‌ترین فاصله	
	الف	ب
۶	۹/۱۱	۰/۶۱
۷	۹/۷۰	۰/۶۶
۸	۱۶	۰/۸۶
۹	۳۸/۷۲	۱/۵۶
۱۰	۴۲	۱/۶۴



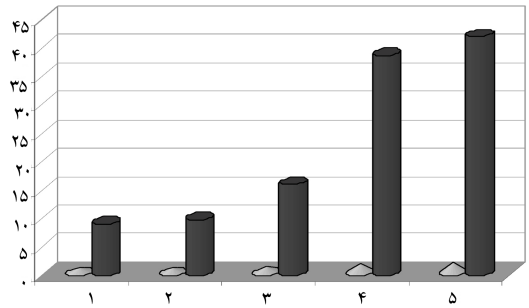
نمودار ۱. مقایسه‌ی نتایج فرمول الف و ب براساس کم‌ترین فاصله برای M=۲.



نمودار ۲. مقایسه‌ی نتایج فرمول الف و ب براساس کم‌ترین فاصله برای M=۳.



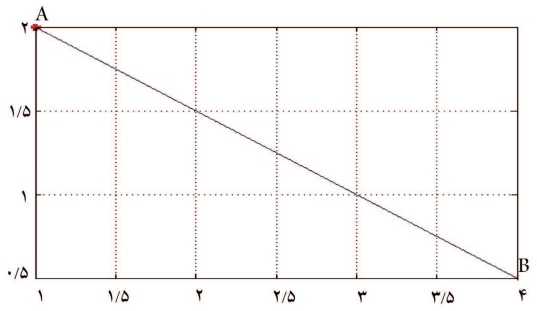
نمودار ۳. مقایسه‌ی نتایج فرمول الف و ب براساس کم‌ترین فاصله برای M=۴.



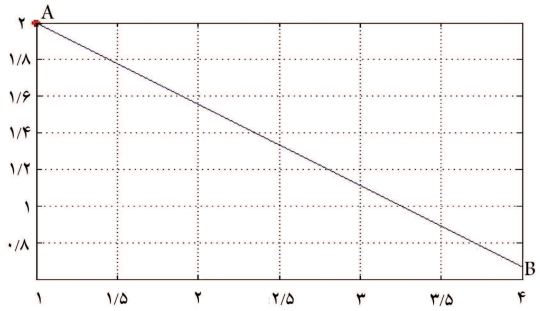
نمودار ۴. مقایسه‌ی نتایج فرمول الف و ب براساس کم‌ترین فاصله برای M=۵.

و شکل‌های ۱ تا ۲۶ آمده است. در نمودارهای ۱ تا ۴ شکل مثلث نشان‌دهنده‌ی نتایج فرمول پیشنهادی، و شکل استوانه نشان‌دهنده‌ی نتایج فرمول مشابه رائو است. ضمناً منبع داده‌های آزمون، ماتریس‌های تصادفی از عناصر ۰ و ۱ بوده‌اند که سطر و ستون آن براساس N انتخاب شده است.

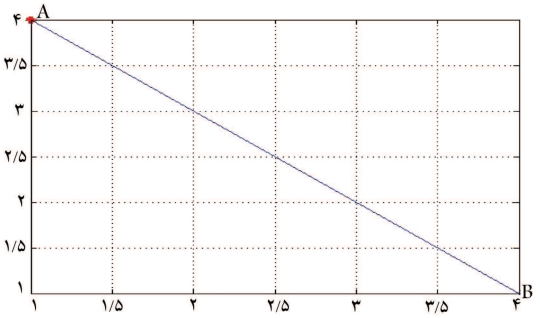
فرمول پیشنهادی که در بقیه‌ی مراحل با علامت اختصاری (ب) نشان داده خواهد شد طبق فرمول ۱ است. در اینجا تحلیل، به «تحلیل خوشه‌بندی مبتنی بر



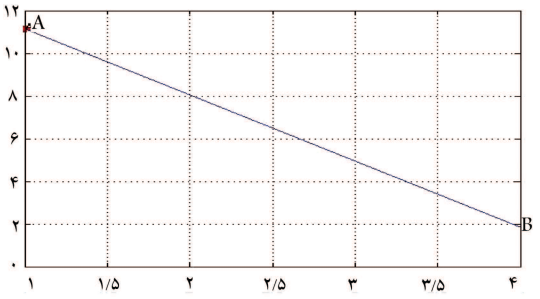
شکل ۱. برای M=۲ و N=۳.



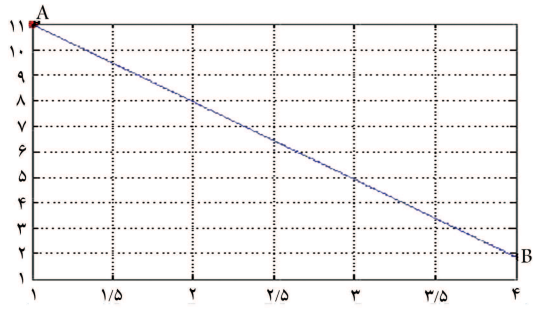
شکل ۲. برای M=۲ و N=۴.



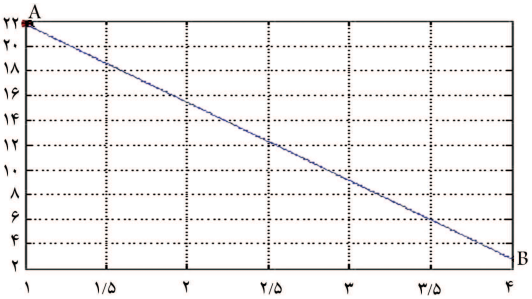
شکل ۳. برای M=۲ و N=۵.



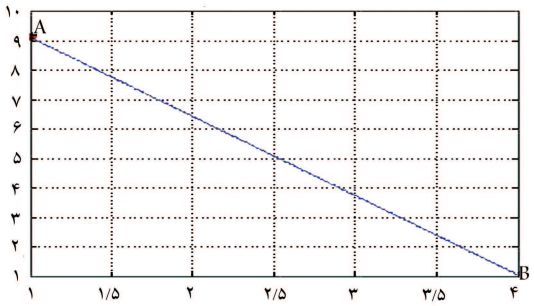
شکل ۴. برای M=۲ و N=۶.



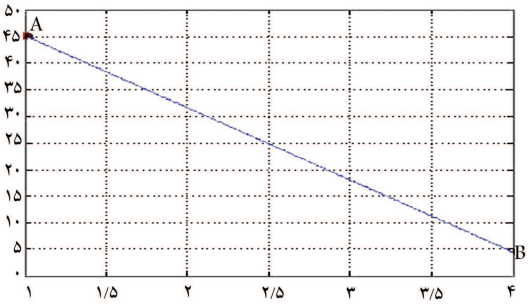
شکل ۵. برای M=۲ و N=۷.



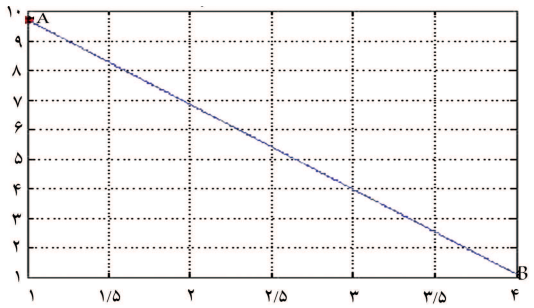
شکل ۶. برای M=۲ و N=۸.



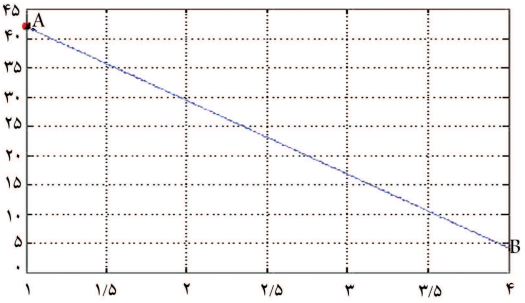
شکل ۱۱. برای M=۳ و N=۶.



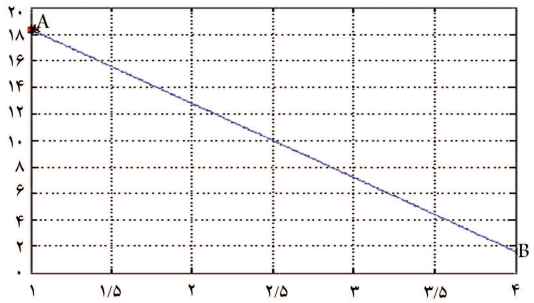
شکل ۷. برای M=۲ و N=۹.



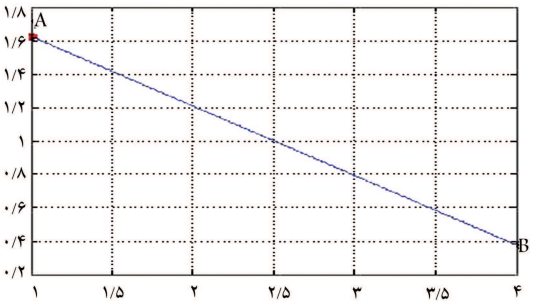
شکل ۱۲. برای M=۳ و N=۷.



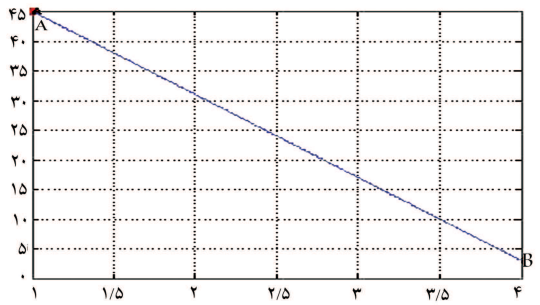
شکل ۸. برای M=۲ و N=۱۰.



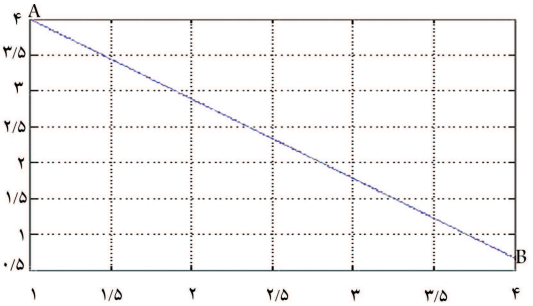
شکل ۱۳. برای M=۳ و N=۸.



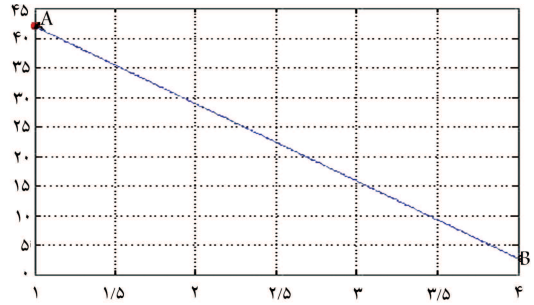
شکل ۹. برای M=۳ و N=۴.



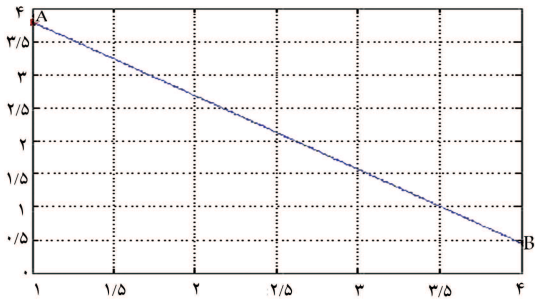
شکل ۱۴. برای M=۳ و N=۹.



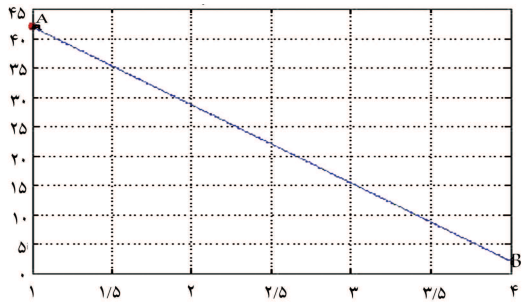
شکل ۱۰. برای M=۳ و N=۵.



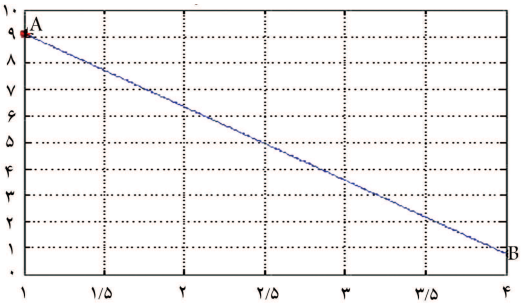
شکل ۱۵. برای M=۳ و N=۱۰.



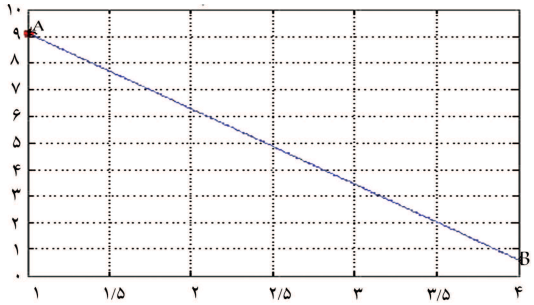
شکل ۱۶. برای M=۴ و N=۵.



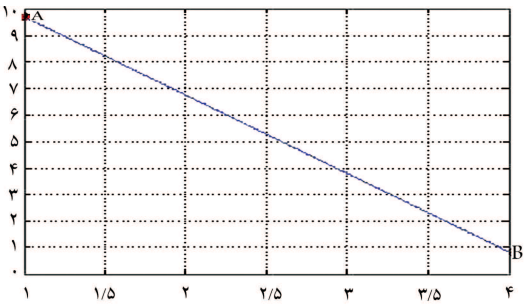
شکل ۲۱. برای M=۴ و N=۱۰.



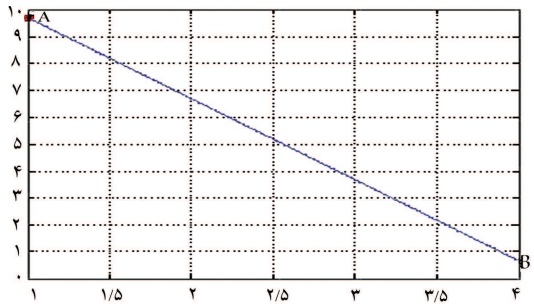
شکل ۱۷. برای M=۴ و N=۶.



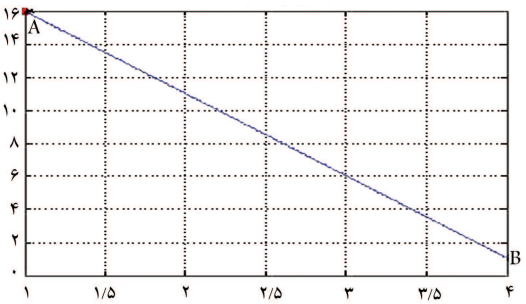
شکل ۲۲. برای M=۵ و N=۶.



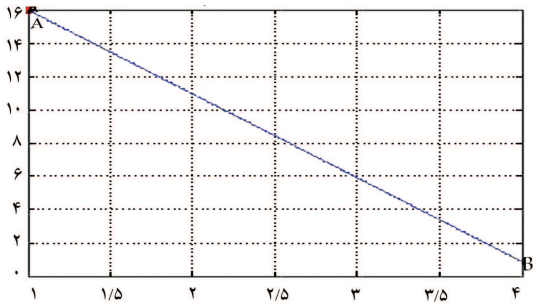
شکل ۱۸. برای M=۴ و N=۷.



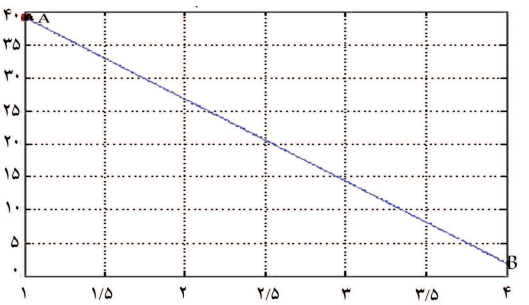
شکل ۲۳. برای M=۵ و N=۷.



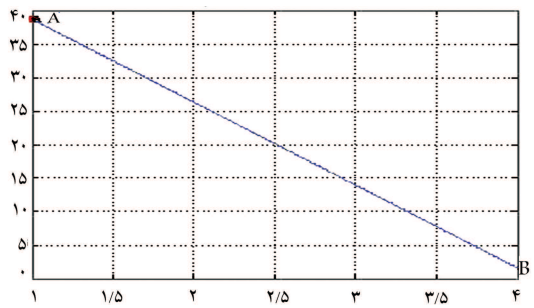
شکل ۱۹. برای M=۴ و N=۸.



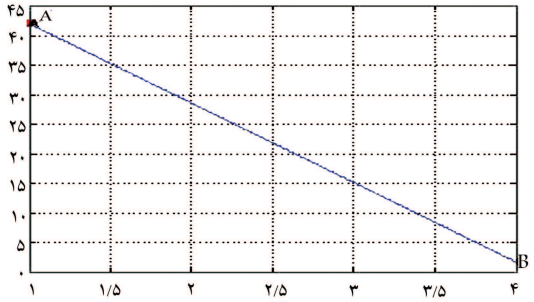
شکل ۲۴. برای M=۵ و N=۸.



شکل ۲۰. برای M=۴ و N=۹.



شکل ۲۵. برای M=۵ و N=۹.



شکل ۲۶. برای M=۵ و N=۱۰.

فاصله» محدود می‌شود که در آن معیار سنجش فاصله بین موجودیت‌های مختلف فراهم است. اگرچه معیار فاصله بین اجزای مختلف شناخته شده، اما هنوز هم معیار تقسیم بهینه به کاربرد مورد نظر بستگی دارد و همیشه، N بیان‌گر تعداد اجزاء و M بیان‌گر تعداد خوشه‌هاست.

$$\min \sum_{k=1}^{M} \left[\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |d_{ij}| x_{ik} x_{jk} \right) / \sum_{i=1}^{N} x_{ik} \right] / M \quad (1)$$

$$\sum_{k=1}^{M} x_{ik} = 1$$

$$i = 1, 2, \ldots, N \quad (2)$$

$$x_{ik} \geq 0 \qquad k = 1, 2, \ldots, M, \qquad i = 1, 2, \ldots, N \quad (3)$$

این فرمول پیشنهادی با فرمول مشابه «کمینه‌سازی مجموع متوسط مجذور فواصل میان خوشه‌یی» طبق فرمول شماره ۴ ــ که در بقیه مراحل با علامت اختصاری (الف) نشان داده خواهد شد ــ مقایسه شده است؛ در مدل مذکور نیز از نرم‌افزار MATLAB استفاده شده است:

$$\min \sum_{k=1}^{M} \left[\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^{2} x_{ik} x_{jk} \right) / \sum_{i=1}^{N} x_{ik} \right] \quad (4)$$

$$\sum_{k=1}^{M} x_{ik} = 1$$

$$i = 1, 2, \ldots, N \quad (5)$$

$$x_{ik} \geq 0 \qquad k = 1, 2, \ldots, M,$$

$$i = 1, 2, \ldots, N \quad (6)$$

در این رابطه‌ها d_{ij} فاصله‌ی بین جزء i و j است. برحسب این که آیا جزء iاُم به گروه k نسبت داده شده یا خیر، x_{ik} برابر ۱ یا صفر خواهد بود.

چنان که در جداول و نمودارهای ۱ تا ۴ مشاهده می‌شود هدف خوشه‌بندی داده‌های پیشنهادی مختلف درگروه‌های متفاوت از M=۲ الی M=۵ خوشه است.

۱.۳. مراحل انجام کار

مراحل انجام کار و نتایج حاصل از آن به شرح زیر است:

مرحله‌ی اول. در جدول ۱ داده‌های ورودی از N=۳ الی N=۱۰ در دو خوشه قرار داده شده است و نشان داده شده که کم‌ترین فاصله‌ی درون‌خوشه‌یی در مدل پیشنهادی در مقایسه با مدل ارائه‌شده‌ی قبلی (طبق فرمول ۴) با توجه به افزایش داده‌ها دارای روند بهبود مناسب‌تری بوده و خوشه‌های ایجادشده از لحاظ کم‌ترین فاصله نیز دارای وضعیت بهتری هستند. همچنین خروجی داده‌های اعمال‌شده و روند بهبود آن‌ها در شکل‌های ۱ تا ۸ ــ که گویای ایجاد خوشه‌هایی با کم‌ترین فاصله‌ی درون‌خوشه‌یی است ــ نشان داده شده است. ضمناً چنان که در جدول ۱ و نمودار ۱ نشان داده شده فواصل درون خوشه‌یی مدل ارائه‌شده‌ی قبلی به‌مراتب بیشتر از مدل پیشنهادی است.

مرحله‌ی دوم. در جدول ۲ داده‌های ورودی از N=۴ الی N=۱۰ در سه خوشه‌ی متفاوت قرار داده شده است. چنان که در جدول ۲ و نمودار ۲ نشان داده شده باز هم کم‌ترین فاصله‌ی درون‌خوشه‌یی مدل پیشنهادی به‌مراتب بهتر از مدل ارائه‌شده‌ی قبلی است؛ همچنین در مدل ارائه‌شده‌ی قبلی کم‌ترین فاصله‌ی درون‌خوشه‌یی براساس مجموع متوسط مجذور فواصل نیز دارای اعداد بزرگ‌تری است. بازه تغییرات در مدل پیشنهادی کم‌تر از مدل ارائه‌شده‌ی قبلی است.

در شکل‌های ۹ تا ۱۵، مدل پیشنهادی با مدل ارائه‌شده‌ی قبلی برای خوشه‌بندی داده‌ها در سه گروه مقایسه شده، که در همه‌ی شکل‌های ارائه‌شده روند بهبود مدل پیشنهادی به‌وضوح دیده شده است.

مرحله‌ی سوم. در جدول ۳ داده‌های ورودی از N=۵ الی N=۱۰ در چهار خوشه‌ی متفاوت قرار داده شده است. چنان که در جدول ۳ و نمودار ۳ مشاهده می‌شود کم‌ترین فواصل درون‌خوشه‌یی براساس مدل پیشنهادی بهتر از مدل ارائه‌شده‌ی قبلی بوده که روند بهبود به‌طور کامل در همه‌ی شکل‌های ۱۶ تا ۲۱ به‌صورت روند نزولی بهبود نشان داده شده است. در جدول ۳ نشان داده شده که با افزایش مقدار خوشه‌ها (M) فواصل درون‌خوشه‌یی تغییرات کم‌تری دارد؛ این نکته گویای آن است که خوشه‌بندی با مدل پیشنهادی مناسب‌تر و بهتر از خوشه‌بندی از طریق مدل ارائه‌شده‌ی قبلی است.

مرحله‌ی چهارم. در جدول ۴ و نمودار ۴ داده‌های ورودی از N=۶ تا N=۱۰ در پنج خوشه‌ی متفاوت قرار داده شده است. همانند سه مرحله‌ی قبلی (طبق شکل‌های ۲۲ تا ۲۶) روند بهبود و نیز مقایسه‌ی کم‌ترین فاصله‌ی درون‌خوشه‌یی مدل پیشنهادی با مدل ارائه‌شده‌ی قبلی نشان داده شده، که همچنان شاهد همان روند بهبود هستیم.

۴. ارزیابی مدل پیشنهادی روی یک مجموعه داده‌ی حقیقی

در این بخش عملکرد و صحت مدل ریاضی و مدل خوشه‌بندی پیشنهادی، بر روی یک مجموعه داده‌ی حقیقی شرکت صنعتی پارس‌خزر آزمایش می‌شود. در بازار فرارقابتی[۶] امروز، پارس‌خزر زمانی قادر به بقاست که نیاز مشتری و تغییرات بازار را به‌درستی شناسایی کند و منابع در دسترس خود را برای خلق فاکتورهای حیاتی موفقیت و مزیت‌های رقابتی خود به کارگیرد. لذا شناسایی ویژگی‌های مشتریان و انجام خوشه‌بندی مناسب، طوری که مشتریان با بیشترین شباهت در یک خوشه قرار بگیرند، از اهمیت به‌سزایی برخوردار است. از طرفی با توجه به ۳۵۰ مرکز خدمات پس از فروش شرکت صنعتی پارس‌خزر برای ارائه‌ی خدمات مناسب‌تر به مشتریان، از مدل پیشنهادی برای خوشه‌بندی مناسب مراکز استفاده شده، و نتیجه‌ی آن با مدل ارائه‌شده‌ی قبلی براساس فرمول مشابه «کمینه‌سازی مجموع متوسط مجذور فواصل میان‌خوشه‌یی» (رابطه‌ی ۴) مقایسه شده است. مراحل انجام‌شده عبارت است از:

۱. با توجه به پراکندگی مشتریان شرکت در سراسر ایران، ابتدا نقشه‌ی جغرافیای ایران به چهار ناحیه‌ی «شمال شرقی، شمال غربی، جنوب شرقی و جنوب غربی» تفکیک شد.

۲. سپس تعیین شد که استان مورد نظر درکدام چهار ناحیه‌ی تعیین‌شده قرار دارد.

۳. براساس چهار ناحیه‌ی تعیین‌شده، اعداد ۱ تا ۴ به آن استان اختصاص داده شد.

۴. با توجه به محل استقرار استان در ناحیه‌ی تعیین‌شده، اعداد مختصات جغرافیایی همان ناحیه به استان مربوطه اختصاص یافته است.

با توجه به توضیحات داده‌شده، استان‌هایی که در چهار ناحیه قرار گرفته‌اند عبارت‌اند از:

- ناحیه‌ی شمال شرقی با اعداد مختصات (۱ ۱):
  ۱. خراسان رضوی (۱ ۱ ۱)
  ۲. خراسان جنوبی (۱ ۱ ۲)
  ۳. خراسان شمالی (۱ ۱ ۳)
  ۴. سمنان (۱ ۱ ۴)
  ۵. یزد (۱ ۱ ۵)
  ۶. گلستان (۱ ۱ ۶)
- ناحیه‌ی شمال‌غربی با اعداد مختصات (۲ ۲):
  ۱. آذربایجان شرقی (۲ ۲ ۱)
  ۲. اردبیل (۲ ۲ ۲)
  ۳. آذربایجان غربی (۲ ۲ ۳)
  ۴. گیلان (۲ ۲ ۴)
  ۵. زنجان (۲ ۲ ۵)
  ۶. کردستان (۲ ۲ ۶)
  ۷. قزوین (۲ ۲ ۷)
  ۸. مازندران (۲ ۲ ۸)
  ۹. همدان (۲ ۲ ۹)
  ۱۰. کرمانشاه (۲ ۲ ۱۰)
  ۱۱. تهران (۲ ۲ ۱۱)
  ۱۲. قم (۲ ۲ ۱۲)
  ۱۳. مرکزی (۲ ۲ ۱۳)
  ۱۴. لرستان (۲ ۲ ۱۴)
  ۱۵. ایلام (۲ ۲ ۱۵)
  ۱۶. اصفهان (۲ ۲ ۱۶)
- ناحیه‌ی جنوب شرقی با اعداد مختصات (۳ ۳):
  ۱. سیستان و بلوچستان (۳ ۳ ۱)
  ۲. کرمان (۳ ۳ ۲)
  ۳. هرمزگان (۳ ۳ ۳)
- ناحیه‌ی جنوب غربی با اعداد مختصات (۴ ۴):
  ۱. خوزستان (۴ ۴ ۱)
  ۲. چهارمحال و بختیاری (۴ ۴ ۲)

جدول ۵. مقایسه‌ی نتایج فرمول الف و ب براساس کم‌ترین فاصله.

تعداد نمونه	تعداد خوشه	حداقل فاصله	
		الف	ب
۳۰	۳	۴۶۷۵/۸۵	۱۰۶/۳۰
۳۰	۴	۴۶۷۵/۸۵	۷۹/۷۲

۳. فارس (۴ ۴ ۳)

۴. کهکیلویه و بویراحمد (۴ ۴ ۴)

۵. بوشهر (۴ ۴ ۵)

اعداد مختصاتی که به هر استان داده شده براساس ترتیب استقرار در چهار ناحیه‌ی انتخاب شده بوده که بعد از تعیین محل جغرافیایی هریک از استان‌ها، داده‌های تعیین‌شده وارد نرم‌افزار Matlab ۷.۵.۰ (R۲۰۰۷b) شد، وکلیه‌ی داده‌ها در سه و چهار خوشه قرارگرفت (جدول ۵). چنان که در جدول ۵ مشاهده می‌شود، خروجی مدل پیشنهادی برای تعداد نمونه‌های ۳۰تایی و در خوشه‌های ۳ و ۴تایی به‌مراتب بهتر از مدل ارائه‌شده‌ی قبلی است و روند بهبود آن نیز به‌وضوح مشخص است.

۵. نتیجه‌گیری

در این نوشتار نشان داده شد که برخی مسائل تحلیل خوشه‌بندی مبتنی بر فاصله، مسائل برنامه‌نویسی ریاضی‌اند. مدل ریاضی ارائه‌شده در این تحقیق مبتنی بر مدل «کمینه‌سازی مجموع متوسط فواصل درون‌خوشه‌یی» است که مزیت اصلی آن مطلوب بودن نتایج محاسباتی است. از مدل ارائه‌شده به‌صورت موردی برای داده‌های مشتریان شرکت صنعتی پارس‌خزر در اقصی‌نقاط ایران استفاده شد، و براین اساس خوشه‌های پیشنهادی تشکیل شد. صحت مدل پیشنهادی با نتایج حاصل از مدل ارائه‌شده‌ی قبلی «کمینه‌سازی مجموع متوسط مجذور فواصل میان‌خوشه‌یی» بر روی مثالی نمادین و مجموعه داده‌یی پیشنهادی مقایسه و مشاهده شد که خروجی مدل پیشنهادی نسبت به مدل ارائه‌شده‌ی قبلی بهتر است.

درچند دهه‌ی اخیر، مدل‌های ریاضی‌کاربرد موفقیت‌آمیزی درتحلیل خوشه‌بندی داشته‌اند. اگرچه نتایج زیادی حاصل شده، اما همچنان برای ادغام کامل تحلیل خوشه‌بندی در مدل‌های ریاضی باید با جدیت کوشید و در این راه به اصول بدیهی، به‌ویژه خوشه‌بندی نیازمندیم. بنابراین موضوعات پیشنهادی برای تحقیقات بعدی، ابداع الگوریتم‌های دقیق جدید عمدتاً برای خوشه‌بندی سلسله‌یی تقسیمی، خوشه‌بندی دنباله‌یی و جمعی است. انجام مقایسه‌ی تجربی روش‌های ابداعی نیز خیلی نادر است، همچنین این روش‌ها هیچ اطلاعاتی درمورد ساختار خوشه‌ها در اختیار ما قرار نمی‌دهند، مثلاً کدام داده‌های خوشه به مرکز یا متوسط آن نزدیک‌ترند یا داده‌هایی که به‌نوعی به دو یا چند خوشه شباهت دارند چگونه بیان می‌شوند. برای رفع این مشکلات بهتر است از رویکرد فازی در مدل‌های تحلیل خوشه‌بندی استفاده شود. همچنین پیشنهاد می‌شود مدل پیشنهادی در این نوشتار به‌صورت ترکیبی با روش BRICH ــ که الگوریتمی جدید برای خوشه‌بندی مجموعه داده‌های خیلی بزرگ است و مسئله‌ی خوشه‌بندی بزرگ را از طریق تمرکز بر بخش‌های متراکم و خلق خلاصه‌یی فشرده ایجاد می‌کند ــ نیز در بررسی‌های آتی استفاده شود.

پانوشت‌ها

1. Data Mininy
2. customer relationship management
3. hierarchical clustering
4. nonhierarchical clustering
5. Fuzzy Clusteriny Method
6. hyper competitive

منابع (References)

1. Saeedi, A."Concept and application of data mining in high education", **18** (In Persain)(2005).
2. Ahola, J. and Rinta-Runsala, E. "Data mining case studies in customer profiling", Research Report TTE1-2001-29, VTT Information Tecknology, pp. 1-24 (2002).
3. Nemati, H.R. and Barko, C.D. "Enhancing enterprise decisions through organizational data mining", *J. of Computer Information Systems,* **42**(4) pp. 21-28 (2002).
4. Rygielski, C., Wang, J.C. and Yen, D. "Data mining techniques for customer relationship management", *Technology in Society,* **24**, pp. 483-502 (2002).
5. Padmanabhan, B. and Tuzhilin, A. "On the use of optimization for data mining: Theoretical interactions and CRM opportunities Management", *Science,* **49**, pp. 1327-1343 (2003).
6. Ngai, E.W.T., Xiu, L. and Chau, D.C.K. "Application of data mining techniques in customer relationship management: A literature review and classification", *Expert Systems with Applications,* **36**, pp. 2592-2602 (2009).
7. Velmurugan, T. and Santhanam, T. "Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points", *J. of Computer Science,* **6**(3), pp. 363-368 (2010).
8. Sharma, S., *Applied Multivariate Techniques*, John Wiley & Sons, USA (1996).
9. Zhang, T., Ramakrishnan, R. and Livny, M. "BIRCH: A new data clustering algorithm and its applications", *Data Mining and Knowledge Discovery,* **1**, pp. 141-182 (1997).
10. GhasemiGol, M., Sadoghi Yazdi, H. and Monsefi, R. "A new hierarchical clustering algorithm on fuzzy data (FHCA)", *Int. J. of Computer and Electrical Eng.*, **2**(1), pp. 1793-8163 (2010).
11. Bradley, P.S., Fayyad, U.M. and Angasarian, O.L. "Mathematical programming for data mining: Formulations and challenges", *J. on Computing,* **11**, pp. 217-238 (1999).
12. Likas, A., Vlassis, N. and Verbeek, J.J. "The global k-means clustering algorithm", *Pattern Recognition,* **36**, pp. 451-461 (2003).
13. Rao, M.R. "Cluster analysis and mathematical programming", *J. of the American Statistical Association,* **66**, pp. 622-626 (2009).