پژوهش‌های نوین در تصمیم‌گیری
دوره ۳، شماره ۲، تابستان ۱۳۹۷

خوشه‌بندی کارگاه‌های صنعتی با استفاده از رویکرد ترکیبی داده‌کاوی و تصمیم‌گیری چندمعیاره

آمنه خدیور[۱*]، فاطمه مجیبیان[۲]

۱- دانشیار، گروه مدیریت، دانشکده علوم اجتماعی و اقتصادی، دانشگاه الزهرا (س)، تهران، ایران
۲- استادیار، گروه مدیریت، دانشکده مهندسی صنایع و مدیریت، دانشگاه غیاث‌الدین جمشید کاشانی، قزوین، ایران

دریافت: ۱۳۹۶/۳/۲۹ پذیرش: ۱۳۹۶/۱۲/۲۷

چکیده

در دهه اخیر، توانایی بشر برای تولید و ذخیره داده‌ها به‌سرعت افزایش یافته است. با افزایش حجم داده‌های ذخیره‌شده، نیاز به روشی که بتوان با استفاده از آن به تحلیل اطلاعات و دانش موجود در داده‌ها پرداخت بیشتر از پیش احساس می‌شود. فنون داده‌کاوی و روش‌های تصمیم‌گیری چند شاخصه در دهه‌های اخیر هرکدام به شکلی کمک‌رسان مدیران در عرصه تصمیم‌گیری بوده‌اند. در پژوهش حاضر، با تلفیق فرآیند داده‌کاوی و روش‌های تصمیم‌گیری چندشاخصه، روشی برای خوشه‌بندی کارگاه‌های صنعتی ارائه شده است. در روش پیشنهادی، ابتدا فرآیند داده‌کاوی بر اساس روش‌های تجزیه‌وتحلیل سلسله مراتبی، K-means و شبکه عصبی کوهونن صورت گرفته و سپس عملکرد مدل طراحی‌شده جهت تعیین تعداد خوشه بهینه با شاخص‌های اعتبارسنجی مجموع خطای مربعی و واریانس بین خوشه‌ای سنجیده شده است. بخش صنایع غذایی به‌عنوان مورد مطالعاتی پژوهش موردبررسی قرار گرفته و بر اساس یافته‌های به‌دست‌آمده، چهار خوشه به‌عنوان تعداد خوشه بهینه کارگاه‌های صنعتی این بخش معرفی شده است. خوشه‌های به‌دست‌آمده بر اساس متغیرهای توزیع جمعیت، سطح

---

* نویسنده مسئول مقاله: E-mail: a.khadivar@alzahra.ac.ir

درآمد و ارزش‌افزوده فعالیت‌های صنعتی در خوشه‌ها نام‌گذاری شده‌اند و در پایان، پیشنهادهایی در دو بخش کاربردی و پژوهشی برای تصمیم‌گیرندگان و سیاست‌گذاران این صنعت و سایر محققان این حوزه ارائه شده است.

واژگان کلیدی: داده‌کاوی؛ خوشه‌بندی؛ تجزیه‌وتحلیل سلسله مراتبی؛ روش K-means؛ روش شبکه عصبی کوهونن.

۱- مقدمه

در گذشته، عموماً استخراج اطلاعات مفید از داده‌های ثبت‌شده، به‌صورت دستی و بر عهده تحلیل‌گران بود. با توجه به اینکه تجزیه‌وتحلیل دستی داده‌ها بسیار کُند و گران بود و هر روز بر پیچیدگی و حجم داده‌ها افزوده می‌شد، تحلیل‌های دستی به سمت تحلیل‌های غیرمستقیم خودکار و استفاده از روش‌های رایانه‌ای حرکت کرد و نیاز مبرمی به استفاده از فناوری‌های جدید و ابزارهای خودکار به وجود آمد تا بتوان از طریق آن‌ها حجم زیاد داده را به‌صورت هوشمند به اطلاعات و دانش تبدیل کرد. در شرایط حاضر، ضروری است از فناوری اطلاعات برای استفاده از این دانش بهره گرفت و داده‌کاوی[1] پاسخی مناسب برای استخراج این دارایی است [۱]. از طرفی خوشه‌بندی[2] از مفیدترین کارکردهای داده‌کاوی برای کشف گروه‌ها و تعیین توزیع‌های موردعلاقه و الگوها در داده‌هاست. مسئله خوشه‌بندی در مورد جداسازی یک مجموعه داده به گروه‌ها یا خوشه‌ها به نحوی است که داده‌های موجود در یک خوشه، نسبت به نقاط موجود در خوشه‌های دیگر شباهت بیشتری به یکدیگر داشته باشند [۲].

زمانی که یک مدیر بدون دید دقیق، ارزیابی واقع‌بینانه و پیش‌بینی مبتنی بر روند واقعی تغییرات رفتاری در عوامل تحت مدیریت خود، دست به اتخاذ تصمیماتی بزند، انتظار منطقی از موفقیت این تصمیمات چندان بالا نیست. یکی از مصادیق این مورد، کارگاه‌های صنعتی هستند. یک مدیر صنعتی که مدیریت یک کارگاه را بر عهده دارد، نیازمند یک دید کلی، تجمیع‌یافته و طبقه‌بندی‌شده از فعالان اقتصادی است تا بتواند در سطح کلان مدیریتی تصمیماتی را اتخاذ نمایند. یکی از موارد طلاقی علم رایانه و به‌ویژه

1. Data Mining
2. Clustering

شاخه هوش مصنوعی با علم مدیریت و فناوری اطلاعات، درست در همین نوع تصمیم‌گیری است؛ لذا در پژوهش حاضر، سعی بر آن است که در مورد کارگاه‌های صنعتی نیز این کاربرد اجرایی شود و این نوع کارگاه‌ها به کمک روش‌های خوشه‌بندی که معیارهای آن با توجه به روش‌های تصمیم‌گیری چندهدفه بر مبنای خوشه‌بندی اولویت‌دهی شده‌اند، در خوشه‌های مختلفی دسته‌بندی شوند تا با استفاده از خوشه‌های شکل‌گرفته به تصمیمات کارآمدتری دست یابیم. مطالعه موردی این پژوهش با تمرکز بر خوشه‌بندی و تعیین تعداد خوشه بهینه کارگاه‌های صنعتی فعال در بخش صنایع غذایی کشور صورت گرفته است. کاربرد نتایج این پژوهش علاوه بر خود کارگاه‌های صنعتی، تسهیل فرآیند تصمیم‌گیری سیاست‌گذاران بخش دولتی و صنعتی کشور را موجب می‌شود و این تصمیمات نیز می‌تواند توسط افراد مختلفی ازجمله مدیران کارگاه‌های صنعتی، مدیران کلان دولتی، قانون‌گذاران، بانک‌ها و مؤسسات مالی و اعتباری، سازمان بورس و سهام‌داران، مشتریان، ارائه‌کنندگان مواد اولیه، سرمایه‌گذاران خصوصی، شرکای اقتصادی و غیره اتخاذ شود.

۲- مبانی نظری پژوهش

در این بخش به بررسی مفاهیم و تعاریف داده‌کاوی پرداخته و سپس، فرآیند خوشه‌بندی و دو روش خوشه‌بندی *K-means* و شبکه عصبی کوهونن مورد بحث و بررسی قرار می‌گیرند.

۲-۱- فرآیند داده‌کاوی

اصطلاح کشف دانش برای نخستین بار در دهه ۱۹۹۰ مطرح شد و توجه پژوهشگران را به سمت الگوریتم‌های داده‌کاوی معطوف کرد. هدف داده‌کاوی کشف دانش جدید، معتبر و قابل‌پیگیری با استفاده از ابزارهای هوش مصنوعی[1] و آماری در حجم بالایی از داده‌ها است [۳]. داده‌کاوی، فرآیند مرتب‌سازی و طبقه‌بندی داده‌های حجیم و آشکارسازی اطلاعات مرتبط باهم است. امروزه، داده‌کاوی به‌عنوان یکی از ابزارهای بسیار مهم مدیران جهت شناخت وضعیت

1. Artificial Intelligence

دقیق‌تر سازمان و همچنین کمک در اتخاذ تصمیمات مناسب، کاربرد دارد. با استفاده از این روش، داده‌های موجود در سازمان با به‌کارگیری ابزارهای نرم‌افزاری، موردبررسی و تحلیل دقیق قرار می‌گیرد تا الگوهای پنهان و پیچیده‌ای که در آن‌ها وجود دارد، کشف و استخراج شود [۴]. هدف اصلی داده‌کاوی پیش‌بینی است و می‌توان گفت داده‌کاوی شناسایی الگوهای صحیح، بدیع، سودمند و قابل‌درک از داده‌های موجود در یک پایگاه داده است که با استفاده از پردازش‌های معمول قابل‌دستیابی نیستند [۵]. داده‌کاوی فرآیند به خدمت گرفتن یک روش‌شناسی رایانه‌ای است که با استفاده از روش‌ها و الگوریتم‌های مختلف، در جستجوی دانش نهفته در داده‌هاست [۶]. داده‌کاوی، پایگاه‌های داده‌ای بزرگ را به‌عنوان منبع دانش در نظر می‌گیرد [۷].

۲-۲- فرآیند خوشه‌بندی

روش خوشه‌بندی یکی از زیرمجموعه‌های علم داده‌کاوی است که هدف آن، اکتشاف و پردازش پایگاه‌های داده‌ای به‌منظور استخراج دانش از آن‌هاست. خوشه‌بندی یک روش یادگیری غیرنظارتی[1] برای دسته‌بندی داده‌ها بر اساس مشابهت‌های آن‌هاست. این روش به‌عنوان ابزاری توانمند جهت استخراج ساختار اصلی نهفته در مجموعه داده‌ها معرفی شده است [۸]. در این پژوهش، روش‌های خوشه‌بندی K-means و شبکه عصبی کوهونن استفاده شده‌اند.

- **روش خوشه‌بندی K-means**

در میان الگوریتم‌های خوشه‌بندی سلسله مراتبی، الگوریتم K-means یکی از الگوریتم‌های متداول است. در این الگوریتم، موجودیت‌ها به نزدیک‌ترین مرکز خوشه تعلق می‌گیرند و این کار تا زمانی که هر موجودیت به نزدیک‌ترین خوشه تخصیص یابد، ادامه پیدا می‌کند [۹].

مجموعه داده‌هایی با N داده n بُعدی x^n مفروض است که هدف، تعیین تقسیم‌بندی طبیعی مجموعه داده‌ای به K خوشه است. K-means تلاش می‌کند تا مجموع مربعات تابع خوشه‌بندی را از رابطه ۱ به حداقل برساند.

---

1. Unsupervised learning task

$$J = \sum_{j=1}^{k} \sum_{n \in N_j} \left\| x^n - \mu_j \right\|^2 \qquad (1)$$

که در این رابطه، μ میانگین نقاط داده‌ای در خوشه S_j است و از رابطه ۲ به دست می‌آید.

$$\mu_j = \frac{1}{N_j} \sum_{n \in S} x^n \qquad (2)$$

این دنباله با تخصیص تصادفی نقاط به K خوشه انجام می‌شود. سپس، بردارهای میانگین μ از N_j نقطه را در هر خوشه محاسبه می‌کند. برای هر نقطه مجدداً خوشه جدیدی تعیین می‌شود که بر اساس آن بردار، نزدیک‌ترین میانگین به دست می‌آید. پس‌ازآن، بردارهای میانگین مجدداً محاسبه می‌شوند. فرآیند خوشه‌بندی K-means بدین‌صورت قابل تعریف است که ابتدا تعداد خوشه‌ها مشخص و دسته اولیه انتخاب می‌شود و مواردی که به عضو j از دسته j که j=1,...,k نزدیک‌ترند تعیین می‌شوند. سپس، میانگین نمونه‌ها در هر خوشه محاسبه شده و مرکز خوشه‌ها[1] به میانگین خوشه‌هایشان نزدیک می‌شود. نزدیک‌ترین موارد به مرکز خوشه جدید j متعلق به خوشه j مجدداً تخصیص داده شده و میانگین نمونه‌ها در هر خوشه به‌عنوان یک مرکز خوشه جدید در نظر گرفته می‌شود. این روش آن‌قدر تکرار می‌شود تا در خوشه‌بندی تغییر بیشتری دیده نشود [۷].

- **شبکه‌های عصبی کوهونن**

با توجه به محدودیت‌های روش‌های خوشه‌بندی کلاسیک، نیاز به یک روش تحلیلی ایجاد می‌شود که قابلیت تعلیم و مدل‌سازی سیستم‌ها با پیچیدگی دلخواه را داشته باشد. روش‌های نوین و مبتنی بر فراگیری ماشین بر مشکلات روش‌های آماری در تحلیل داده‌ها فائق آمده‌اند. ازجمله روش‌های تأمین‌کننده انتظارات فوق، شبکه‌های عصبی هستند [۱۰]. از معروف‌ترین و پرکاربردترین شبکه‌های عصبی

---

1. Seed

می‌توان به شبکه عصبی کنترل نشده کوهونن موسوم به «نقشه‌های خودسازمان‌دهنده» اشاره کرد که اولین بار توسط کوهونن و با الگوبرداری از عصب‌های شبکیه‌ی چشم معرفی شد. نقشه‌های خودسازمان‌ده از انواع شبکه‌های عصبی با قابلیت یادگیری بدون ناظر هستند که در تحلیل فضاهای پیچیده و خوشه‌بندی داده‌ها در گروه‌های همگن، توانایی زیادی دارند و یک ابزار مؤثر برای تجزیه‌وتحلیل داده‌های چندبعدی هستند. این شبکه را می‌توان برای تجزیه‌وتحلیل خوشه‌ای استفاده نمود؛ به‌طوری‌که ورودی‌های مشابه در لایه خروجی در کنار هم باقی می‌مانند [۱۱].

۲-۳- پیشینه پژوهش

تاکنون مطالعات بسیاری در خصوص فرآیند داده‌کاوی و روش‌های خوشه‌بندی ارائه شده است که در این بخش به بررسی تعدادی از آن‌ها به تفکیک موضوعات و روش‌های مختلف مورداستفاده، مطابق جدول زیر پرداخته شده است.

طی بررسی صورت گرفته در پژوهش‌های داخلی، خوشه‌بندی شرکت‌ها یا کارگاه‌ها با رویکرد داده‌کاوی یافت نشد و در اکثر خوشه‌بندی‌های انجام‌شده سطح تحلیل افراد، مشتریان یا محصولات هستند. از طرفی، هیچ پژوهش مشابهی در صنایع غذایی نیز صورت نگرفته است.

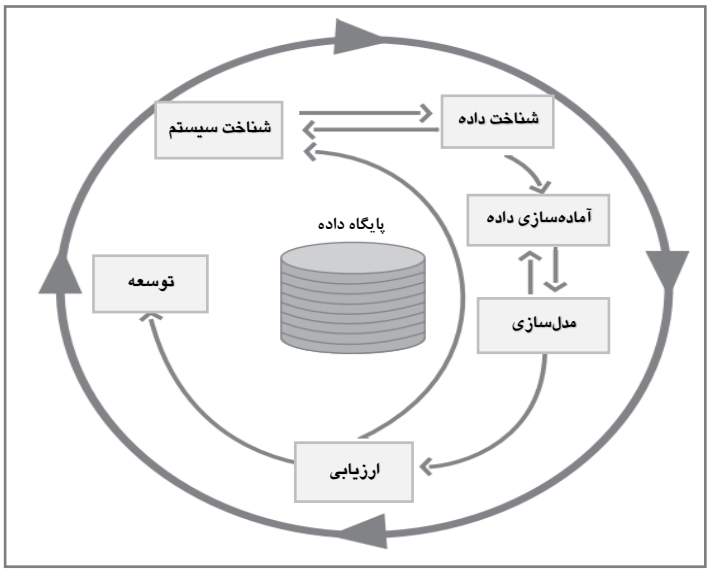
۳- روش پیشنهادی پژوهش

در این پژوهش به‌منظور پیاده‌سازی فرآیند داده‌کاوی، از روش استاندارد داده‌کاوی میان‌صنعتی[1] استفاده شده است. این روش در سال ۱۹۹۶ طرح‌ریزی شده و امروزه بیشتر طرح‌های داده‌کاوی نیز از این روش استفاده می‌کنند [۳]. داده‌کاوی CRISP دارای ۶ مرحله است که هرکدام دارای اهداف جداگانه‌ای شامل شناخت سیستم، شناخت داده، آماده‌سازی داده، مدل‌سازی، ارزیابی و توسعه است [۲۲]. مراحل استاندارد فرآیند CRISP-DM در شکل ۱ نمایش داده شده است [۲۳].

1. CRISP Date Mining

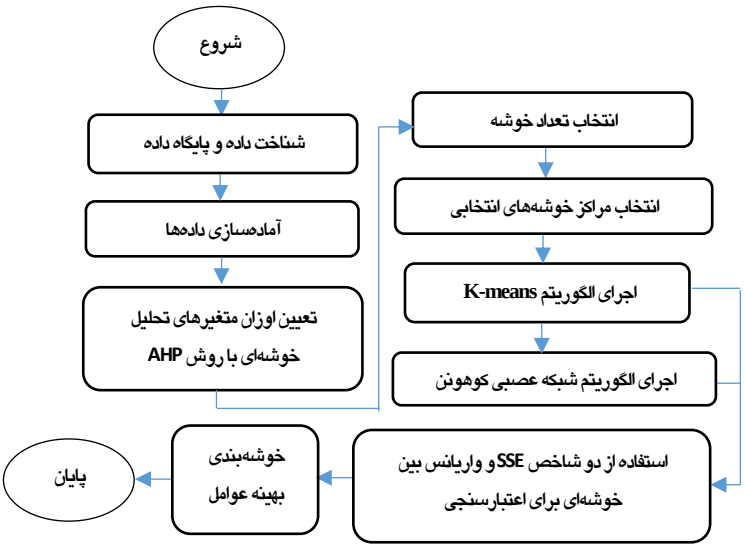
جدول ۱ خلاصه‌ای از پژوهش‌های انجام‌شده

محقق	موضوع	روش	توضیحات
شین و سوهن [۱۲]	بخش‌بندی مشتریان بازار بورس کشور کره و تبیین سیاست‌های بازاریابی برای حداکثر کردن سود بلندمدت	K-means SOM fuzzy K-means	روش خوشه‌بندی بر روی داده‌های تراکنشی مشتریان بازار بورس کشور کره جنوبی صورت گرفته است
راد و همکاران [۱۳]	خوشه‌بندی و رتبه‌بندی رشته‌های دانشگاهی	K-means AHP	تعداد ۱۷۷ رشته دانشگاهی را در ایران با توجه به شباهت‌ها و تفاوت‌های آن‌ها در ۱۰ خوشه، خوشه‌بندی نمودند
حنفی‌زاده و رستخیز [۱۴]	مقایسه‌ی دو روش داده‌کاوی در بخش‌بندی مشتریان بیمه بدنه خودرو بر اساس ریسک	K-means SOM	عوامل تأثیرگذار بر ریسک بیمه‌گذاران شناسایی شده و سپس بخش‌بندی مشتریان انجام گرفته و ویژگی‌های مشتریان در هر یک از بخش‌ها مشخص شده است
بای و همکاران [۱۵]	ارزیابی عملکرد سازمان‌ها برمبنای عملکردهای عملیاتی و استراتژیک	Fuzzy C-Means TOPSIS	این روش با استفاده از داده‌های واقعی شرکت‌ها و ابعاد عملکردی کارت امتیاز متوازن پیاده‌سازی و ارزیابی شده است
لولی و همکاران [۱۶]	دسته‌بندی موجودی‌ها بر اساس شاخص‌های متفاوت	AHP K-Means	یک رویکرد ترکیبی برای تسهیل فرآیند مدیریت و کنترل موجودی استفاده شده است
هوانگ و همکاران [۱۷]	ارائه الگوریتم سری زمانی بر اساس روش K-Means برای خوشه‌بندی داده‌های سری زمانی	Time Series K-Means	الگوریتم پیشنهادی می‌تواند از اطلاعات سری‌های زمانی به‌منظور افزایش عملکرد خوشه بهره‌برداری نماید
بشیری موسوی و همکاران [۱۸]	تحلیل ارزش مشتری در بانک با استفاده از روش داده‌کاوی و تحلیل سلسله مراتبی فازی	FAHP K-Means	این تحقیق با هدف تعیین سازوکاری برای ارزش‌گذاری مشتریان در شعب بانک تجارت صورت گرفته است
بروفر و همکاران [۱۹]	شناسایی الگوی رفتاری مشتریان در بیمه عمر و تشکیل سرمایه با استفاده از داده‌کاوی	RFM Two-step K-Means	بخش‌بندی مشتریان بر اساس مهم‌ترین ویژگی‌های مالی و جمعیت شناختی در قالب عوامل مؤثر بر شاخص‌های RFM صورت گرفته است
فضلی و جماعتی تفتی [۲۰]	پیش‌پردازش تصمیم‌گیری چندشاخصه با استفاده از داده‌کاوی	SWARA VIKOR	استفاده از هم‌افزایی بین داده‌کاوی و تصمیم‌گیری چندشاخصه برای خوشه‌بندی شرکت‌های وارانتی شخص‌ثالث
هیلرمان و همکاران [۲۱]	به‌کارگیری روش‌های خوشه‌بندی و AHP برای ارزیابی شواهد بهداشتی مشکوک	Cluster analysis AHP method	الگوریتم تلفیقی پیشنهادی برای خوشه‌بندی و رتبه‌بندی موجودیت‌های مشکوک



شکل ۱ مراحل انجام فرآیند CRISP-DM

بدین‌ترتیب، الگوریتم خوشه‌بندی پیشنهادی با بهره‌گیری از دو روش کلاسیک K-means و شبکه عصبی کوهونن به‌صورت زیر است.



شکل ۲ الگوریتم خوشه‌بندی پیشنهادی

۴- مطالعه موردی: خوشه‌بندی کارگاه‌های صنعتی فعال در بخش صنایع غذایی

در این بخش، کارگاه‌های صنعتی ثبت‌شده در کشور در گروه صنایع غذایی به‌عنوان مورد مطالعاتی بررسی شده و نتایج حاصل از خوشه‌بندی این کارگاه‌ها بر اساس پیاده‌سازی روش پیشنهادی در بخش بعد مورد تجزیه‌وتحلیل و بحث قرار خواهد گرفت. بدین ترتیب با توجه به اطلاعات مربوط به مورد مطالعاتی، الگوریتم خوشه‌بندی پیشنهادی پژوهش طی مراحل زیر پیاده‌سازی شده است:

- شناخت سیستم و پایگاه داده

بر اساس اطلاعات دریافت شده از وزارت صنایع در خصوص کارگاه‌های صنعتی فعال در صنایع غذایی، تعداد ۲۷۱۵ کارگاه برای کار خوشه‌بندی انتخاب شدند. فهرست اطلاعات مربوط به این کارگاه‌ها مطابق جدول زیر ارائه شده است:

جدول ۲ اطلاعات مربوط به حوزهای مختلف مورداستفاده برای خوشه‌بندی (بر اساس گزارش وزارت صنایع در سال ۱۳۹۴)

ارزش مواد اولیه مصرفی	تشکیل سرمایه ثابت	ارزش تغییرات موجودی انبار	مالیات غیرمستقیم و عوارض	جبران خدمات	پرداختی غیرصنعتی	ارزش‌افزوده فعالیت‌های صنعتی	تعداد کارکنان
۰	۰	-۱۴۰/۷۷۱۵۶	۰	۶۸۰/۶۷۰۰۷	۴۰۲/۱۸۵۸۱	۲۵۴۰/۱۲۶۵۴۴	۲۴
۰	۰	۰	۰	۵۸۴/۲۸۲۴۲۸	۱۱۲/۶۰۰۶۴۳	۱۷۶۳/۰۹۱۲۹۹	۳۱
۴۳/۶۳۲۸۴۸	۰	۰	۰	۶۱۱/۳۵۸۱۹۵	۴۰۱/۱۷۶۹۳۴	۱۶۷۶/۵۳۴۳۸۴	۱۲۹۹
۵	۵۷/۸	-۱۵۴/۹۰۳۸۱۲	۰	۹۹۸/۸۲۲۴	۲۵۷/۳۸۰۱۲	۲۲۵۲/۲۷۱۹۵	۳۸
۱۸	۲۰۳۲	۱۴۲	۲۵	۳۹۴/۷۴۶۲۴	۴۶/۷	۸۸۲/۹۷۶	۱۶
۰	۶۰/۴۸۴۵	۱۷۳۸/۲۶۲	۰	۹۷۳/۷۴۴۸۹	۴۵/۷۹۷۹۶	۳۳۹۳/۲۲۰۲۰۷	۵۱
۰	۸۵/۴۰۵۹۳۶	۲۶/۰۳۵۶۳	۰	۱۲۹/۱۱۱۸۳۶	۳۴/۳۱۳۴۹۴	۲۳۲۶/۶۶۷۵۸۴	۶۰
۰	۱۰۴/۶۸۳	۱۴۱۹/۰۵۱۸۴۵	۴۷/۴۰۳۰۶۸	۱۳۹/۱۵۹۸۷۴	۱۱۳/۳۰۹۳۶۹	۱۴۵۶/۷۱۳۵۶۱	۷۰
۰	۵۸۷	۰	۰	۷۹۷/۱۳۷۸۴۸	۸۹/۰۶۲۴	۱۴۶۸۹/۸۱۷۷۱	۲۷
۰	۲/۰۸	-۴۶۵/۵۷۹۱۷۴	۰	۶۰۵/۵۵۶۳۲۴	۵۰/۳۲۱۸	۹۶۷/۰۱۸۷۹۴	۳۶
۵/۸۰۲۶۶۶	۱۴۶	۰	۶/۲۷۲	۱۲۹/۵۶۳۳۷۹	۵۱/۳۲۵۸۶۵	۵۳۴/۰۲۳۵۸۳	۱۰
۰	۴۷۳۰	۰	۰	۵۲۳/۹	۸۳۳/۷	۱۳۸۴/۸۵	۲۹
۰	۳۵۱/۱۲۸	۰	۱۲۰/۶۳۱۸۴	۲۱۸/۰۸۴۰۴۹	۴۲۳/۳۵۷۶۲۸	۴۸۴۳/۳۷۳۳۰۴	۹۸
۰	۲۴۹/۷۷۰۰۰۳	۳۳۱۶/۰۸۱۰۲۹	۰	۳۵۷/۷۴۹۹۶۵	۱۵۴/۵۸۹۲۹۲	۷۶۰۱/۸۴۴۲۳	۱۵۵
۰	۰	۰	۰/۲۴۷۲۶	۵۵۷/۰۴۶۸۲۳	۷۴/۶۵۶۹۰۵	۶۳۷۷/۴۴۹۸۵۹	۲۰
۰	۲۵۰۵/۳۵۱۳۵	۲۷۵۵/۸۹۹۴۴	۱۲/۵۹۶۹۳۲	۷۱۶/۹۳۴۰۳۹	۳۷۹/۲۲۷۰۴۴	۳۱۶۵/۳۴۷۴۲۹	۳۹

- آماده‌سازی داده‌ها

ازآنجایی‌که یکی از فنون اصلی استفاده‌شده در این پژوهش برای خوشه‌بندی کارگاه‌ها K-means بوده و این الگوریتم نسبت به داده‌های پرت، داده‌های خالی و داده‌های غیرنرمال حساس است، سعی بر آن شده که تا حد امکان با استفاده از روش‌های پاک‌سازی داده‌ها، از میزان داده‌های مشکل‌دار کاسته شود. برای این منظور از نرم‌افزار اکسل برای پیدا کردن داده‌های خالی استفاده شده و مقادیر داده‌های گمشده با میانگین خانه پر شده است. برای عملیات نرمال‌سازی نیز از نرم‌افزار اکسل استفاده شده و همه ستون‌ها طبق فرمول زیر نرمال‌سازی شده‌اند:

$$Z = \frac{X - @Global\_Mean(x)}{@Global\_SDEV(x)} \quad (۳)$$

در این فرمول، X مقدار مربوط به هر خانه است. @GLOBAL_MEAN میانگین مربوط به یک متغیر (ستون) خاص از جدول و @GLOBAL_SDEV مقدار انحراف از معیار یک متغیر خاص را برمی‌گرداند. لازم به توضیح است در مدل پژوهش حاضر توابع گلوبال در نرم‌افزار Clementine از قبل تعریف شده‌اند. با اجرای توابع گلوبال اطلاعات لازم ازجمله میانگین مربوط به یک متغیر، انحراف از معیار یک متغیر، بیشینه و کمینه یک متغیر به دست می‌آید. پس از آماده‌سازی داده‌های بی مقیاس‌شده، داده‌ها وارد نرم‌افزار Clementine شده و جهت تشخیص داده‌های پرت از آزمون آنومالی استفاده شده است. از بین ۲۷۱۵ رکورد فقط ۲۷ رکورد دارای آنومالی هستند. استراتژی مواجهه با این داده‌های پرت، استفاده از میانگین خانه به‌جای داده پرت بوده است.

سپس برای تعیین اولویت و ضرایب متغیرهای خوشه‌بندی از روش مقایسه زوجی یا تحلیل سلسله مراتبی[1] (AHP) استفاده خواهد شد. روش کار بدین‌صورت بوده که اهمیت یا ارجحیت شاخص‌های مختلف نسبت به یکدیگر، در قالب ماتریسی که در اکسل ساخته شده بود، به‌صورت زوجی از ۲۶ تن از خبرگان وزارت صنایع و مرکز آمار ایران مورد پرسش قرار گرفته است. ماتریس حاصل،

1. Analytical Hierarchy Process

نرمال شده تا برای هر شاخص وزنی به دست آید که نمایانگر مقدار ارجحیت این شاخص برای خوشه‌بندی باشد.

به‌منظور نرمال کردن، ابتدا حاصل جمع هر ستون به دست می‌آید. سپس، هر عنصر در ماتریس زوجی را به جمع ستون خودش تقسیم کرده تا ماتریس زوجی، نرمالایز شود. مقدار میانگین هر سطر در ماتریس نرمال‌شده محاسبه می‌شود. اعداد حاصل، ضریب وزنی ارجحیت هر شاخص‌اند. برای محاسبه نهایی وزن‌ها از نرم‌افزار Expert Choice11 استفاده شده است. این نرم‌افزار بر پایه روش تصمیم‌گیری چندمعیاره است. نتایج نهایی به‌دست‌آمده پس از رندسازی به شرح جدول زیر است:

جدول ۳ وزن‌های نهایی به‌دست‌آمده برای متغیرهای تأثیرگذار بر خوشه‌بندی

توضیح ستون‌ها	وزن نهایی ستون‌ها
تعداد کل کارکنان	۲۰
ارزش‌افزوده فعالیت‌های صنعتی	۱۵
پرداختی غیر صنعتی	۱۰
جبران خدمات	۸
مالیات غیرمستقیم و عوارض	۱۰
ارزش تغییرات موجودی انبار	۱۲
تشکیل سرمایه ثابت	۱۵
ارزش مواد اولیه مصرفی	۱۰

بعد از به دست آوردن وزن‌ها، وزن هر متغیر وارد نرم‌افزار Clementine شده است؛ سپس از الگوریتم K-means برای فرآیند خوشه‌بندی داده‌های وزن‌دار استفاده می‌شود. یکی از مسائل مهم در خوشه‌بندی، تعیین تعداد بهینه خوشه-هاست که در اکثر الگوریتم مانند K میانگین باید توسط خود کاربر معین شود و راه خاصی برای تعیین آن مشخص نشده است. بدین ترتیب، پس اجرای الگوریتم K-means برای تعداد خوشه‌های مختلف، الگوریتم کوهونن پیاده‌سازی می‌شود.

در شبکه کوهونن، هر گره دارای موقعیت مکانی مشخص بوده (یک جدول مشخصات x و y) و دارای برداری با همان ابعاد ورودی است. با ورود هر بردار ورودی جدید، همه گره‌ها بررسی می‌شود تا گره‌ای که مشابه‌ترین اوزان را به بردار ورودی دارد، یافت شود. بدین ترتیب، حالات زیر برای تعداد خوشه‌ها و ابعاد شبکه در نظر گرفته شده است:

تعداد خوشه‌ها: ۳، ۴، ۵ و ۶ خوشه

ابعاد شبکه بسته به تعداد خوشه:

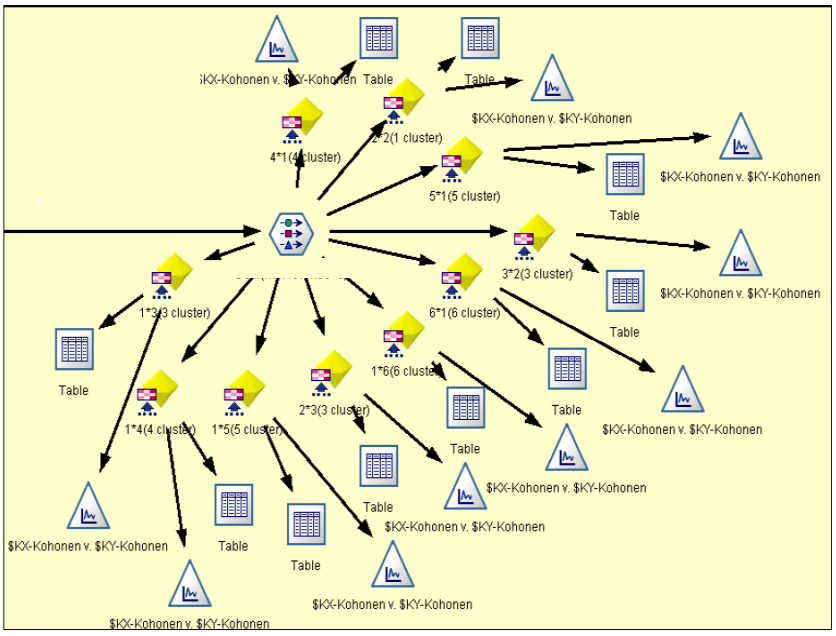
۳: (۱و۳)، (۳و۱)، (۲و۳) و (۳و۲)

۴: (۱و۴) و (۴و۱)

۵: (۱و۵) و (۵و۱)

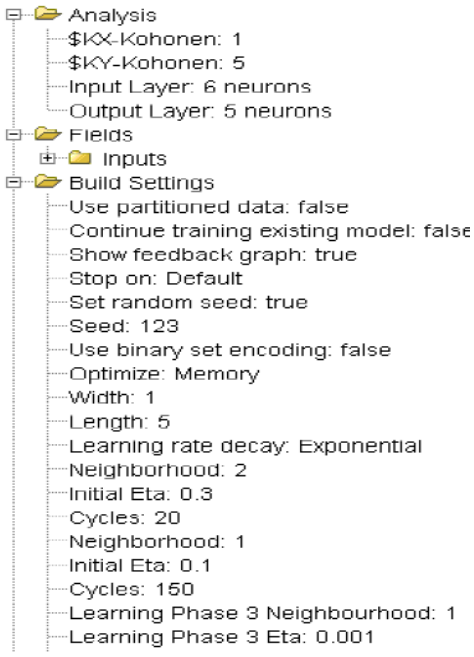
۶: (۱و۶) و (۶و۱)

مدل ساخته‌شده در نرم‌افزار برای اجرای الگوریتم، در شکل زیر نمایش داده شده است:



شکل ۳ مدل ساخته‌شده برای اجرای الگوریتم کوهونن

خلاصه عملیات خوشه‌بندی شبکه عصبی کوهونن با ابعاد (۵و۱)، ۵ نـرون در لایـه خروجی و ۵ خوشه، برای نمونه در شکل زیر نشان داده شده است:



شکل ۴ خلاصه نتایج اجرای الگوریتم کوهونن

- محاسبه شاخص‌های اعتبارسنجی

شاخص‌های اعتبارسنجی برای سنجش میزان صحت نتایج خوشه‌بنـدی بـه‌منظـور مقایسه بین روش‌های خوشه‌بندی مختلف یا مقایسـه نتـایج حاصـل از یـک روش بـا مولفـه‌هـای مختلـف مورداسـتفاده قـرار مـی‌گیرنـد. در ایـن پـژوهش دو شـاخص اعتبارسنجی مجموع خطای مربعی و واریانس بین خوشه‌ای به‌منظور انتخاب بهتـرین شکل خوشه‌بندی از بین تمام شکل‌های موجود استفاده شده است.

- شاخص مجموع خطای مربعی:

در این معیار، خطا برابرِ فاصـله هـر نقطـه بـا نزدیـک‌تـرین خوشـه اسـت. پـس از مشخص کردن خطای تمامی نقاط، معیار SSE از رابطه زیر محاسبه می‌شود:

$$SSE = \sum_{i=1}^{c} \sum_{x \in D_i} dist^2(m_i, x) \qquad (۴)$$

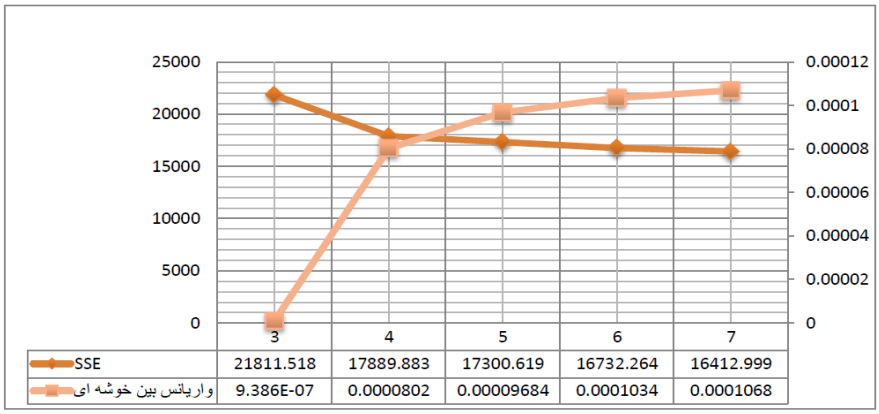
در این رابطه، c بیانگر تعداد خوشه‌ها، m_i بیانگر مرکز خوشه، x معرف نقطه‌ای متعلق به خوشه و D مجموعه داده است.

- واریانس بین خوشه‌ای:

این معیار متداول‌ترین معیار مقایسه نتایج خوشه‌بندی بوده و بر اساس منطق حداکثر تفاوت بین خوشه‌ها مورداستفاده قرار می‌گیرد. هر چه این شاخص بزرگ‌تر باشد، گویای پراکندگی خوشه‌ها نسبت به هم و معمولاً نشانه خوشه‌بندی بهتر است.

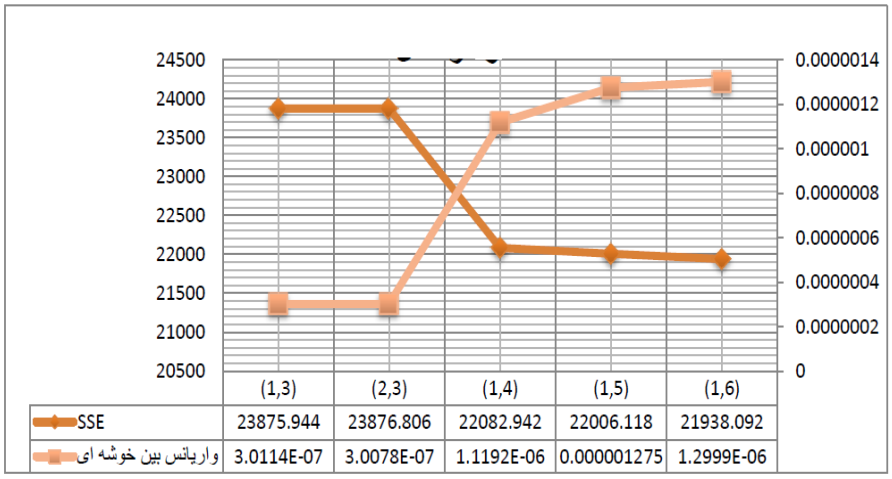
$$\sigma^2 = \frac{\sum_i (x - \mu_x)^2}{n} \qquad (۵)$$

که در آن μ_x میانگین مراکز خوشه‌ها و x مرکز هر یک از خوشه‌هاست. بعد از محاسبه این شاخص‌ها، مقدار آن‌ها برای انواع حالت‌های ممکن (تعداد خوشه‌های مختلف) مقایسه شده که نتایج آن برای الگوریتم‌های K-means و کوهونن در نمودارهای ۱ و ۲ نشان داده شده است.



نمودار ۱ نمودار مقایسه شاخص SSE و واریانس بین خوشه‌ای برای الگوریتم K-means

همان‌طور که نمودار ۱ نشان می‌دهد، پس از خوشه‌بندی با تعداد ۳ خوشه (که دارای شاخص SEE بالاتر و شاخص واریانس بین خوشه‌ای پایین‌تری نسبت به نقطه بعدی است)، شیب منحنی تا نقطه K=4 برای شاخص SEE با سرعت زیاد، کاهش و برای شاخص واریانس بین خوشه‌ای، افزایش یافته است و شرایط در نقطه چهار نسبت به نقطه سه بهبود یافته است. لذا این نقطه بیانگر تعداد بهینه خوشه‌هاست؛ چراکه در ادامه نمودار، با افزایش تعداد خوشه‌ها تغییر چندانی (با شیب زیاد) در منحنی‌ها و مقادیر شاخص‌ها به وجود نمی‌آید؛ بنابراین، چهار خوشه به‌عنوان تعداد خوشه بهینه کارگاه‌ها انتخاب می‌شود.



نمودار ۲ نمودار مقایسه شاخص SSE و واریانس بین خوشه‌ای برای الگوریتم کوهونن

با مقایسه منحنی‌های نمودار ۲، شیب منحنی شاخص SSE تا نقطه (۴و۱) که تعداد خوشه ۴ است، با شدت بیشتری نسبت به سایر نقاط، کاهش یافته و شیب منحنی شاخص واریانس بین خوشه‌ای نیز افزایش یافته است. بنابراین همان‌گونه که نشان داده شده است، بعدازاین نقطه، یا شیب‌ها با این شکل تغییر نکرده‌اند یا فقط مقدار یکی از شاخص‌ها بهبود داشته است؛ لذا برای الگوریتم کوهونن نیز تعداد ۴ خوشه مناسب به نظر می‌رسد.

۵- تجزیه‌وتحلیل یافته‌های پژوهش

نتایج به‌دست‌آمده از مقایسه دو الگوریتم *K-means* و شبکه عصبی کوهونن در جدول زیر خلاصه شده است:

جدول ۴ مقایسه نتایج دو الگوریتم استفاده‌شده برای خوشه‌بندی

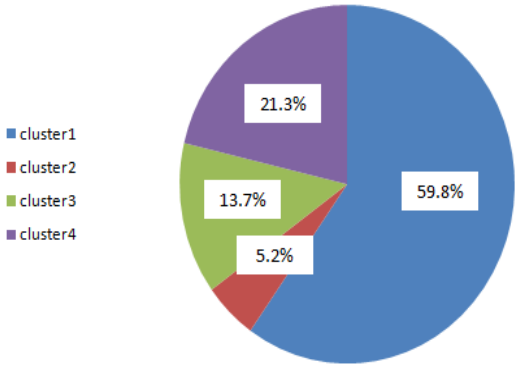
روش	تعداد خوشه‌ها	SSE	واریانس بین‌خوشه‌ای
K میانگین	۴	۲۰۹۵۹/۸۲۳	۰/۰۰۰۰۷۵۸۵۵۴
شبکه عصبی کوهونن	۴	۲۴۶۳۲/۰۳۳	۰/۰۰۰۰۰۰۳۸۹۸۹۹

همان‌طور که در جدول بالا نشان داده شده است، نتایج به‌دست‌آمده از دو روش نشان می‌دهد که تعداد ۴ خوشه می‌تواند ترکیب بهینه‌ای برای داده‌ها باشد. نظر خبرگان وزارت صنایع و مرکز آمار ایران نیز این موضوع را تائید می‌کند که این چهار خوشه می‌تواند تقسیم‌بندی منطقی از کارگاه‌های بخش صنایع غذایی فعال در کشور باشد. توزیع تعداد داده‌ها در درون چهار خوشه تائید شده به‌صورت جدول زیر است:

جدول ۵ توزیع داده‌ها در چهار خوشه تعیین‌شده

خوشه	تعداد رکوردهای هر خوشه	درصد
۱	۵۷۸۲۹۵	۲۱/۳
۲	۳۷۱۹۵۵	۱۳/۷
۳	۱۴۱۱۸	۵/۲
۴	۱۶۲۳۵۷	۵۹/۸

نمودار دایره‌ای مربوط به توزیع خوشه‌ها نیز در شکل ۶ نمایش داده شده است.



نمودار ۳ نمودار دایره‌ای توزیع داده‌ها در خوشه‌ها

۵-۱- نام‌گذاری و تفسیر ماهیت خوشه‌ها

در این مرحله با توجه به درصد توزیع و مقادیر بیشینه و کمینه خانه‌های تأثیرگذار در خوشه‌بندی، خوشه‌های به‌دست‌آمده نام‌گذاری می‌شوند. تحلیل ماهیت خوشه‌ها نشان می‌دهد که توزیع جمعیت، سطح درآمد و ارزش‌افزوده فعالیت‌های صنعتی در خوشه‌ها به‌طور معنی‌داری متفاوت است و لذا از این متغیرها برای نام‌گذاری خوشه‌ها استفاده می‌شود، بنابراین با توجه به نتایج، خوشه‌ها به‌صورت زیر نام‌گذاری می‌شوند:

I. کارگاه‌های کم‌جمعیت با پایین‌ترین سطح ارزش‌افزوده فعالیت‌های صنعتی و درآمد کم؛

II. کارگاه‌های با پایین‌ترین سطح جمعیت و ارزش‌افزوده فعالیت‌های صنعتی متوسط و درآمد متوسط؛

III. کارگاه‌های پرجمعیت با بالاترین سطح ارزش‌افزوده فعالیت‌های صنعتی و درآمد بالا؛

IV. کارگاه‌های با جمعیت متوسط و ارزش‌افزوده فعالیت‌های صنعتی و درآمد متوسط.

۶- نتیجه‌گیری و پیشنهادها

در این پژوهش با در نظر گرفتن لزوم استفاده از تحلیل خوشه‌ای با رویکرد داده‌کاوی و استفاده از تصمیم‌گیری چندشاخصه، الگوریتمی جهت خوشه‌بندی

عوامل ارائه شد. در الگوریتم تحلیل خوشه‌ای پیشنهادی برای اطمینان از حصول نتایج معتبر، از ترکیب الگوریتم‌های *K-means* و کوهونن برای خوشه‌بندی استفاده شده است. روش پیشنهادی پژوهش در بخش صنایع غذایی به‌منظور خوشه‌بندی کارگاه‌های صنعتی این بخش پیاده‌سازی شد که بنا به نتایج به‌دست‌آمده، تعداد خوشه‌های بهینه این صنعت تعداد چهار خوشه معرفی شدند. در ادامه پیشنهادهایی در دو بخش کاربردی و پژوهشی ارائه می‌شود.

۶-۱- پیشنهادهای کاربردی

با توجه به نتایج حاصل از بررسی مورد مطالعاتی این پژوهش و خوشه‌بندی صورت گرفته بر روی کارگاه‌های صنعتی صنایع غذایی کشور، پیشنهادهایی کاربردی در جهت بهبود عملکرد کارگاه‌ها در مورد هر خوشه شناسایی‌شده ارائه می‌شود.

خوشه I: خوشه‌ای است که بیشترین تعداد کارگاه در آن عضو هستند. این کارگاه‌ها عمدتاً از تعداد نیروی انسانی کمی برخوردار بوده و درعین‌حال طبق اطلاعات به‌دست‌آمده ارزش‌افزوده کمی هم ایجاد می‌کنند. وضعیت این کارگاه‌ها اصولاً مطلوب نیست و اصولاً در معرض خطر تعطیلی قرار دارند. طبق نظرات خبرگان این حوزه، سیاست‌گذاران باید برای حفظ این کارگاه‌ها تلاش کنند. ممکن است استراتژی ادغام این کارگاه‌ها ازجمله استراتژی‌های موفق برای حفظ این کارگاه‌ها باشد.

خوشه II: این خوشه که جمعیت کمی هم دارد، شامل کارگاه‌هایی است که تعداد نیروی انسانی کمی دارند و جزو شرکت‌های کوچک و متوسط به‌حساب می‌آیند؛ اما توانسته‌اند که ارزش‌افزوده متوسط و درآمد متوسطی ایجاد کنند. لذا عملکرد این خوشه مثبت بوده و احتمالاً نشان از بهره‌وری بالا در این واحدها یا وجود سرمایه اولیه خوب است. پیشنهاد سیاست‌گذاری برای این خوشه، محک‌زنی خوشه I یا خرید کارگاه‌های خوشه I است.

خوشه III: کارگاه‌هایی که در این خوشه هستند، بهترین وضعیت دارند. این کارگاه‌ها عمدتاً کارخانه‌های بزرگ مواد غذایی هستند که با نیروی انسانی و سرمایه بالا اداره می‌شوند و دارای ارزش‌افزوده بالا و درآمد بالا هستند. این کارخانه وضع موجود مطلوبی داشته و سیاست استمرار وضع موجود برای آن‌ها توصیه می‌شود.

به‌عبارت‌دیگر، ازنظر وزارت صنایع این گروه اولویتی برای کمک‌رسانی دولتی، اخذ تسهیلات و غیره ندارد.

خوشه IV: کارگاه‌هایی که در این خوشه هستند، وضعیت متوسطی دارند؛ ازنظر نیروی انسانی در وضعیت متوسطی هستند، یعنی کوچک به شمار نمی‌آیند، ولی نتوانسته‌اند درآمد یا ارزش‌افزوده صنعتی خوبی ایجاد کنند. این کارگاه‌ها نیز مانند خوشه نخست نیاز به تسهیلات و کمک دارند؛ چراکه ممکن است آن‌ها نیز به خوشه یک تبدیل شوند، نیروی انسانی خود را تعدیل کرده و به درآمد کمتر هم راضی شوند. لذا سیاست‌گذاری برای این خوشه عمدتاً باید به نحوی باشد که آن‌ها را به خوشه ۳ نزدیک کند و از رفتن آن‌ها در خوشه یک جلوگیری کند.

۶-۲- پیشنهادهای پژوهشی

پیشنهاد می‌شود به‌منظور واقعی‌ترسازی خروجی‌های مدل پیشنهادی از نظریه فازی در فرآیند اعمال تصمیم خبرگان استفاده شود؛ چراکه تصمیم‌گیرندگان اغلب به علت طبیعت غیرقطعی مقایسه‌های زوجی، قادر نیستند به‌صراحت نظرشان را در مورد برتری‌ها اعلام کنند. ازاین‌رو برای غلبه بر این مشکلات، استفاده از روش تحلیل سلسله مراتبی فازی در تلفیق با رویکرد داده‌کاوی پیشنهاد می‌شود. از طرفی می‌توان به‌جای دو الگوریتم *K-means* و کوهونن از الگوریتم‌های خوشه‌بندی مبتنی بر روش‌های فراابتکاری نوین و تلفیقی نیز استفاده نمود و نتایج حاصل را با یافته‌های پژوهش حاضر مورد مقایسه قرار داد.

۷- منابع

[1] Taghavi Fard, M.T., Mansori, T., NaserZadeh. M.R., Ferasat. A.R., Data mining and its application in decision making, Journal of Management Knowledge 79(1), 2007, pp. 3-14. (in Persian).

[2] Jahangiri, M., Ahmadi, M.R., Naderi Dehkordi, M., Application of data mining in the insurance industry and customer categories, National

Conference on Computer Engineering and Information Technology Management. 2014. (in Persian).

[3] Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. Toward data mining engineering: A software engineering approach. Information systems, 34(1), 2009. pp. 87-107.

[4] Nori Borojerdi, P., Eskandari, V., Introduction to Quantitative Studies in Management (Case study: Data mining application in management studies), 1(3), 2010, pp. 1-13. (in Persian).

[5] Gharekhani, M., Abolghasemi, M. Data mining applications in the insurance industry, Journal of New Worlds Insurance, 158, 2011, pp.5-22.

[6] Kantardzic, M. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons. 2011.

[7] Larose, D. T. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons. 2014.

[8] Valente J & Pedrycz, W. (Eds.). Advances in fuzzy clustering and its applications. New York: Wiley. 2007.

[9] Hsu, C. H. Data mining to improve industrial standards and enhance production and marketing: An empirical study in apparel industry. Expert Systems with Applications, 36(3), 2009, pp. 4185-4191.

[10] Decker, R., Monien, K., Market basket analisis with neural gas network and self-organising maps, Journal of Targeting, Measurement and Analysis for Marketing, 11(4), 2003, pp. 373-386.

[11] Ghaseminezhad, M.H., & Karimi, A. A novel self-organizing map (SOM) neural network for discrete groups of data clustering, Applied Soft Computing, 11, 2011, pp. 3771-3787.

[12] Shin, H. W., & Sohn, S. Y. Segmentation of stock trading customers according to potential value. Expert systems with applications, 27(1), 2004, 27-33.

[13] Rad, A., Naderi, B., & Soltani, M. Clustering and ranking university majors using data mining and AHP algorithms: A case study in Iran. Expert Systems with Applications, 38(1), 2011, pp. 755-763.

[14] Hanafi Zadeh, P., Rastkhisz Paydar, N. Comparison between two data mining methods in segmentation of car insurance customers (Case Study: Mellat Insurance Company), Industrial Management Studies, 11(30), 2013, pp. 77-98. (in Persian).

[15] Bai, C., Dhavale, D., & Sarkis, J. Integrating Fuzzy C-Means and TOPSIS for performance evaluation: An application and comparative analysis. Expert Systems with Applications, 41(9), 2014, pp. 4186-4196.

[16] Lolli, F., Ishizaka, A., & Gamberini, R. New AHP-based approaches for multi-criteria inventory classification. International Journal of Production Economics, 156, 2014, pp. 62-74.

[17] Huang, X., & Ye, Y., & Xiong, L., & Lau, R.Y.K., & Jiang, N., & Wang, Sh., Time series *k*-means: A new *k*-means type smooth subspace clustering for time series data, Information Sciences, 13(1), 2016, pp. 367-368.

[18] Bashiri mousavi S.A, Afsar A., Mahjubifard A. "Customer value analysis in bank with data mining technique and fuzzy analytic hierarchy process", Management Researches in Iran, 19(1), 2015, pp. 23-43, (in Persian).

[19] Boroufar, A., Rezaeian, A., Shokohyar, S., "Identifying the customer behavior model in life insurance Sector using data mining, Management Researches in Iran, 20(4), 2016, pp. 65-94.

[20] Fazli, S., Jamaati Tafti, R. Preprocessing Multiple Criteria Decision-Making Using Data Mining (Case Study: Selection of third party logistic in outsourcing warranty services of an electronic facilities company), Modern Researches in Decision Making, 2(3), 2017, pp. 215-239. (in Persian).

[21] Hillerman, T., Carlos, H., Carla., A. Rommel, N. Applying clustering and AHP methods for evaluating suspect healthcare claims, Journal of computational science, 19, 2017, pp. 97-111.

[22] Bryson, Osei. Muata, Kweku. Towards supporting expert evaluation of clustering results using a data mining process model. Information Sciences, 180(3), 2010, pp. 414-431.

[23] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., CRISP-DM 1.0 Step-by-step data mining guide 2000. SPSS Inc. 2008.