

مقایسه‌ی دو روش داده‌کاوی در بخش‌بندی مشتریان بیمه‌ی بدنه‌ی اتومبیل بر اساس ریسک (مورد مطالعه: شرکت بیمه‌ی ملت)

پیام حنفی‌زاده^۱

ندا رستخیز پایدار^۲

چکیده

با رشد روزافزون کامپیوتر، مقادیر زیادی از داده‌ها به‌وسیله‌ی سیستم‌های مختلف به‌وجود می‌آیند. در حال حاضر مسئله‌ی پیش روی سازمان‌ها، دیگر جمع‌آوری داده‌ها نیست، بلکه توانایی استخراج اطلاعات مفید از میان آنهاست. همانند دیگر بخش‌های اقتصادی، شناخت و جذب مشتریان کم‌ریسک و سودآور برای صنعت بیمه نیز دارای اهمیت است. بیمه‌ی اتومبیل یکی از مهم‌ترین رشته‌های بیمه‌ای در ایران است. اگر شرکت‌های بیمه به طبقه‌بندی مشتریان با توجه به ویژگی‌های قابل مشاهده بپردازند، می‌توانند نرخ پوشش‌دهی بیمه و سود خود را افزایش دهند و از سوی دیگر فشاری بر افراد با ریسک کم برای جبران خسارات وارده به‌وسیله‌ی افراد ریسک زیاد به شرکت‌های بیمه وارد نشود. در این تحقیق طبقه‌بندی ریسکی بیمه‌گذاران با استفاده از دو تکنیک شبکه‌ی خودسازمان‌ده و الگوریتم k-means انجام شد. در ابتدا عوامل تأثیرگذار بر ریسک بیمه‌گذاران شناسایی شد و سپس بخش‌بندی مشتریان با استفاده از دو روش نام‌برده به‌صورت جداگانه انجام گرفت و ویژگی‌های مشتریان در هر یک از بخش‌ها مشخص شد. در پایان مقایسه‌ای بین دو روش صورت گرفت و تفاوت‌های آنها بیان شد.

واژگان کلیدی: بخش‌بندی مشتریان، بیمه‌ی بدنه‌ی اتومبیل، شبکه‌های خودسازمان‌ده،

الگوریتم k-means

۱- استادیار، دانشکده‌ی مدیریت و حسابداری، دانشگاه علامه طباطبایی

۲- کارشناس ارشد مدیریت فناوری اطلاعات، دانشکده‌ی مدیریت و حسابداری، دانشگاه علامه طباطبایی

(نویسنده مسؤل) paydarneda@gmail.com

تاریخ پذیرش: ۹۱/۵/۱

تاریخ دریافت: ۹۱/۱/۲۸

مقدمه

با رشد روزافزون کامپیوتر، مقادیر زیادی از داده‌ها به وسیله سیستم‌های مختلف به وجود می‌آیند [۹]. با وجود اینکه با استفاده از کامپیوتر مقادیر بسیاری از داده‌ها به وسیله نهاد‌های دولتی، بنگاه‌های تجاری بزرگ و مؤسسات علمی تولید می‌شود، ولی درصد بسیار اندکی از آنها به واقع استفاده می‌شود، زیرا در بسیاری از موارد، حجم داده‌ها بزرگ‌تر از آن است که بتوان آنها را مدیریت کرد یا پیچیدگی آنها بیش از آن است که بتوان به تحلیلشان پرداخت. در حال حاضر مسئله‌ی پیش روی سازمان‌ها، دیگر جمع‌آوری داده‌ها نیست، بلکه توانایی استخراج اطلاعات مفید از میان آنهاست [۱۴]. داده‌کاوی فرایند اکتشاف و پردازش پایگاه‌های داده‌ای به منظور استخراج دانش از آنهاست [۱۷].

در کشور ما، بیمه یکی از مهم‌ترین عوامل حفظ و تضمین سرمایه به‌شمار می‌رود [۱]. بیمه‌ی اتومبیل یکی از برترین رشته‌های بیمه‌ای است که سهم عمده‌ای را در پرتفوی صنعت بیمه دارد. در ایران نرخ حق بیمه‌ی بدنه‌ی اتومبیل با توجه به تعرفه‌ی اعلام‌شده از سوی بیمه‌ی مرکزی جمهوری اسلامی ایران تعیین می‌شود. طبقه‌بندی ریسکی بیمه‌گذاران بر مبنای ویژگی‌های قابل مشاهده، می‌تواند به شرکت‌های بیمه برای کاهش زیان، افزایش نرخ پوشش بیمه و جلوگیری از وقوع انتخاب نامساعد در بازار بیمه کمک شایانی کند [۲]. نرخ حق بیمه در بسیاری از شرکت‌های بیمه‌ای در خارج از کشور با توجه به متغیرهای گوناگون جمعیت‌شناختی، مشخصات اتومبیل و سابقه‌ی خسارت بیمه‌گذار محاسبه می‌شود، زیرا عوامل بسیاری بر تصادفات خودرو تأثیر می‌گذارد. برای نمونه، طبق یکی از تحقیقات معتبر انجام شده در سال‌های اخیر، رنگ خودرو یکی از عوامل مؤثر در انواع تصادفات بوده است. احتمال رخداد تصادف در رنگ‌های تیره بیشتر از رنگ‌های روشن است [۱۶].

نبود سنجش‌های تعیین ریسک افراد در بیمه‌ی اتومبیل علاوه‌بر ناکاراسازی قراردادهای بیمه، منتج به تعیین نرخ‌های غیرعادلانه نیز می‌شود، زیرا به‌جای فرد، اتومبیل بیمه می‌شود و این امر موجب شده تا بیشتر شرکت‌های بیمه در زمینه‌ی بیمه‌ی اتومبیل، متحمل زیان شوند [۳]. داده‌کاوی ابزار ارزشمندی است که در سال‌های گذشته از آن به شکل گسترده‌ای برای استخراج اطلاعات، جست‌وجوی روابط و الگوها در بین حجم گسترده‌ی داده‌های جمع‌آوری‌شده استفاده شده است [۶]. شناسایی مشتریان مستلزم تحلیل مشتریان هدف و دسته‌بندی کردن مشتریان است که منجر به یافتن گروه‌هایی از مشتریان سودآور براساس ویژگی‌های آنها می‌شود [۸]. از نظر مفهومی خوشه‌بندی یعنی گروه‌بندی یک دسته موجودیت در گروه‌های مختلف، به‌طوری‌که داده‌های متعلق به یک خوشه به یکدیگر شبیه بوده و با دیگر خوشه‌ها متفاوت باشند [۲۱]. طبقه‌بندی ریسک در حقیقت به‌معنای گروه‌بندی مشتریان با خصوصیات ریسک مشابه است که احتمال بروز خسارت‌های مشابهی دارند [۵]. در این تحقیق پس از شناسایی عامل‌های اثرگذار بر ریسک مشتریان در بیمه‌ی بدنه‌ی اتومبیل، به بخش‌بندی مشتریان و تعیین ویژگی‌های هریک از آنها در گروه‌های مختلف ریسکی با استفاده از دو روش پرکاربرد خوشه‌بندی شامل شبکه‌ی عصبی خودسازمان‌ده^۱ و الگوریتم k-means خواهیم پرداخت.

ادبیات تحقیق

۱-۲. مروری بر شبکه‌های عصبی خودسازمان‌ده

این مدل از شبکه‌های عصبی نخستین‌بار به‌وسیله‌ی کوهنن^۲ در سال ۱۹۸۱ و با الگوبرداری از عصب‌های شبکیه‌ی چشم، معرفی شد [۱۳]. شبکه‌های عصبی

1. Self Organizing Map.

2. Kohonen.

خودسازمان ده شبکه‌های عصبی بدون نظارتی هستند که قابلیت ارائه‌ی خروجی شبکه در قالب نقشه‌های گرافیکی گویا و قابل فهم برای مدیران سازمان‌ها را دارند، از این رو سرعت درک و تفسیر نتایج برای مدیران و کارشناسان راحت‌تر خواهد بود [۱۰]. نبود حساسیت شبکه‌ی خودسازمان ده به تعداد دادگان تعلیم و حساسیت کم این نوع شبکه‌ها به وجود نویز در دادگان تعلیم، توانایی نمایش روابط خطی و غیرخطی بین متغیرها، قدرت بالا در دسته‌بندی دادگان از دیگر برتری‌های این شبکه‌ها است [۱۸]. شبکه‌های خودسازمان ده از دو لایه‌ی مجزا تشکیل می‌شوند؛ یک لایه‌ی ورودی و لایه‌ی دیگر لایه‌ی نقشه نام دارد. هر نرون در لایه‌ی نقشه مربوط به یک بردار اطلاعات با ابعادی برابر ابعاد فضای مورد تحلیل است. شکل ۱ توپولوژی شبکه‌های خودسازمان ده را نشان می‌دهد. طبق یکی از تحقیقات انجام شده از سوی «چای» و همکارانش در سال ۲۰۰۹ مشخص شده است که از بین ۳۴ تکنیک داده‌کاوی، شبکه‌های عصبی بیشترین کاربرد را داشته است. شبکه‌های عصبی که با استفاده از مغز انسان شبیه‌سازی شده، کاربردهای زیادی در زمینه‌ی دسته‌بندی، خوشه‌بندی و پیش‌بینی داشته است [۷].

۲-۱- آموزش شبکه‌های خودسازمان ده

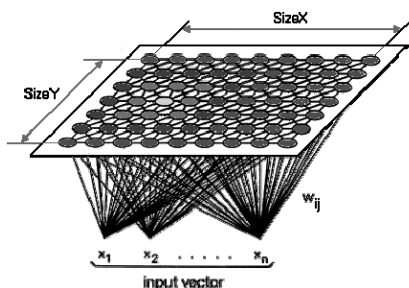
آموزش شبکه‌های خودسازمان ده بر مبنای الگوریتم یادگیری رقابتی و بدون ناظر (بدون استفاده از بردار هدف) است. در ابتدا بردار وزنی متناظر با هر نرون به‌طور تصادفی تولید شده و ساختار اولیه‌ی شبکه شکل می‌گیرد و سپس در طول فرایند آموزش شبکه، بردار وزنی متناظر با هر نرون به‌گونه‌ای تنظیم می‌شود که بتواند قسمتی از اطلاعات فضای مورد تحلیل را پوشش دهد. شکل ۱ توپولوژی شبکه‌ی خودسازمان ده را نشان می‌دهد. الگوریتم آموزش شبکه‌های خودسازمان ده دارای چهار مرحله است [۱۲]:

۱. انتخاب شناسه‌های نقشه مانند ابعاد و بردار وزن ابتدایی متناظر با هر نرون؛

۲. ارائه‌ی داده‌های مورد تحلیل به شبکه و یافتن بهترین نرون نظیر برای هر بردار داده‌ی ورودی؛

۳. به‌هنگام‌کردن بردار وزنی متناظر با هر نرون؛

۴. بررسی شرط خاتمه‌ی الگوریتم.



شکل ۱. توپولوژی شبکه‌ی خودسازمانده

اگر شرط برقرار نباشد، الگوریتم از قدم دوم ادامه می‌یابد. از آنجا که الگوریتم آموزش شبکه‌های خودسازمانده بر مبنای فاصله‌ی اقلیدسی بنا شده است، بایستی داده‌های هر بعد فضای مورد بررسی را مستقلاً نرمال استاندارد کرد.

۲-۱-۲. نمایش فضای مورد تحلیل با استفاده از شبکه‌های خودسازمانده

پس از آموزش شبکه‌های خودسازمانده، به تعداد نرون‌های انتخاب‌شده برای شبکه، بردارهای وزنی n بعدی به‌دست می‌آیند که هر یک نمایانگر بخشی از فضای مورد تحلیل هستند. متناظر با مقدار هر مشخصه در بردار وزنی یک بردار RGB^1 و در نتیجه یک رنگ در نظر گرفته می‌شود؛ به‌گونه‌ای که کلیه‌ی مقادیر با استفاده از طیف رنگی، از آبی تیره (برای کمترین مقادیر) تا قرمز تیره (برای بیشترین مقادیر)، قابل نمایش باشند. بدین ترتیب به‌ازای هر مشخصه، رنگ هر نرون تعیین می‌شود و

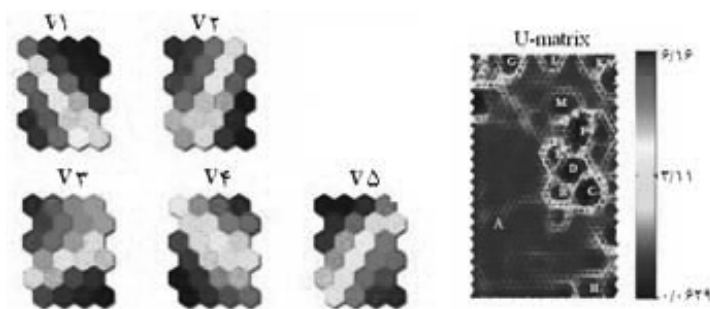
۱. RGB (Red-Green-Blue) از فرمت‌های استاندارد تعریف رنگ‌هاست که هر رنگی را با توجه به میزان شدت رنگ‌های اصلی و از ترکیب آنها قابل دستیابی است.

نقشه‌ی متناظر با آن مشخصه به دست می‌آید. با به دست آمدن نقشه‌های مشخصات، بررسی ارتباط متقابل میان آنها (آزمون همبستگی) ممکن می‌شود [۱۱]. شکل ۲، نشان‌دهنده‌ی نمونه‌ای از کاربرد شبکه‌های خودسازمان‌ده در تحلیل الگوهای پیچیده و نمایش هم‌زمان اثرات متغیرهای مختلف بر یکدیگر است.

• متغیرهای V_2 و V_5 و همچنین V_1 و V_4 ، در تمام دامنه‌ی تغییرات خود دارای همبستگی معکوس هستند و هر جا V_2 دارای رنگ قرمز است (مقادیر بالا به خود گرفته)، V_5 رنگ آبی دارد (مقادیر پایین به خود گرفته است و برعکس). گرچه شدت همبستگی V_2 و V_5 در تمام نقاط فضا تقریباً ثابت است، ولی این مطلب در خصوص متغیرهای V_1 و V_4 صادق نیست.

• ماتریس U^1

از جمله خروجی‌های دیگر شبکه‌های خودسازمان‌ده، ماتریس دسته‌بندی و متناظر با آن نقشه‌ی دسته‌بندی است. درایه‌های این ماتریس، فاصله‌ی جبری نرون‌های همسایه را از یکدیگر نشان می‌دهند. شکل ۲ یک ماتریس U را با تعدادی خوشه و زیرخوشه از یک فضای ۶۲ بعدی نشان می‌دهد [۱۹].



شکل ۲. نمایشی از یک U_Matrix و کاربرد شبکه‌های خودسازمان‌ده در تحلیل هم‌زمان روابط

1. U_Matrix .

۲-۲. الگوریتم K-means

برای بخش‌بندی داده‌ها به‌طور کلی دو رویکرد اصلی وجود دارد: رویکرد سلسله‌مراتبی و رویکرد تفکیکی [۲۰]. الگوریتم k-means یکی از تکنیک‌های پُر استفاده در رویکرد تفکیکی به‌شمار می‌رود [۷]. ایده‌ی اصلی آن تعریف کردن k مرکز است که هر مرکز برای یک خوشه می‌باشد [۲۲]. گام‌های این الگوریتم به‌صورت زیر است:

- ۱- در ابتدا K نقطه به‌عنوان مراکز خوشه‌ها انتخاب می‌شوند.
- ۲- هر نمونه داده به خوشه‌ای که مرکز آن خوشه کمترین فاصله تا آن داده را دارد، نسبت داده می‌شود.
- ۳- پس از تعلق تمام داده‌ها به یکی از خوشه‌ها برای هر خوشه یک نقطه‌ی جدید به‌عنوان مرکز محاسبه می‌شود (میانگین نقاط متعلق به هر خوشه).
- ۴- مراحل ۲ و ۳ تکرار می‌شوند تا زمانی که دیگر هیچ تغییری در مراکز خوشه‌ها به‌وجود نیاید [۱۵].

مدل پیشنهادی

در این تحقیق مشتریان بیمه‌ی بدنه‌ی اتومبیل در شرکت بیمه‌ی ملت براساس ریسک آنها بخش‌بندی می‌شوند. به‌طور کلی گام‌های مدل پیشنهادی شامل تعریف متغیرها، جمع‌آوری داده‌ها، پاکسازی داده‌ها، طراحی مدل و تجزیه و تحلیل داده‌ها است.

۳-۱. تعریف متغیرها

یکی از عوامل مهم در ساخت مدلی با صحت زیاد، انتخاب درست مشخصه‌ها است. برای بخش‌بندی مشتریان براساس ریسک، نخستین و مهم‌ترین گام اصلی شناسایی عوامل ریسک است. فاکتورهای مهم و اثرگذار بر ریسک در این تحقیق طبق

فاکتورهای شناسایی شده در یکی از تحقیقات معتبر انجام شده در سال ۹۰ در کشور تعیین شد [۴]. طبق تحقیق صورت گرفته فاکتورها در دو فاز مطالعه و بررسی بر روی مقالات علمی معتبر منتشر شده در بازه‌ی زمانی ۲۰۰۹-۲۰۰۰ و سپس نظرسنجی از خبرگان انتخاب شدند. عوامل (فاکتورهای) انتخاب شده مطابق جدول ۱ است.

۳-۲. جمع آوری داده

در این تحقیق داده‌های مربوط به تصادفات در پایگاه داده‌ی بیمه‌ی بدنه‌ی اتومبیل در بیمه‌ی ملت در سال ۸۸ مورد استفاده قرار گرفت. پنج فاکتور از فاکتورهای نهایی شامل سن، نوع گواهینامه و سال دریافت گواهینامه‌ی بیمه‌گذار و همچنین ظرفیت موتور و سرعت راننده به دلیل وجود نداشتن اطلاعات در پایگاه داده در ابتدای کار حذف شد.

جدول ۱. فاکتورهای اثرگذار بر ریسک

عوامل برگزیده	گروه عوامل
سن	مشخصات جمعیت شناختی
جنسیت	
محل زندگی	
نوع گواهینامه رانندگی	
سال اخذ گواهینامه (سابقه رانندگی)	مشخصات خودرو
سیستم خودرو	
کاربری خودرو	
نوع خودرو	
رنگ خودرو	
ظرفیت موتور	
مدل خودرو	
تجهیزات ایمنی (ABS)	سابقه رانندگی
میزان کارکرد بر حسب کیلومتر	
تعداد ادعای خسارت در سال قبل	
سرعت رانندگی	مشخصات بیمه نامه
میزان پوشش‌های بیمه‌ای	

با توجه به اینکه فراوانی سه نوع خودرووی پراید، زانتیا و وانت مزدا با دیگر انواع خودروهای بیمه‌شده در این شرکت تفاوت چشمگیری دارند، از این رو تنها سه نوع خودرووی نام‌برده در این تحقیق تجزیه و تحلیل شد.

۳-۳. پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها یکی از گام‌های مهم در فرایند داده‌کاوی است که میزان دقت نتایج به‌دست‌آمده رابطه‌ای قوی با نحوه‌ی انجام آن دارد. در این تحقیق حذف دادگان نامناسب، تبدیل کلیه‌ی مقادیر به مقادیر عددی و نرمال‌سازی مقادیر برای پیش‌پردازش داده‌ها صورت گرفت. همچنین به منظور نرمال‌سازی دادگان تعلیم شبکه‌ی خودسازمان‌ده، از روش نرمال‌سازی Z که کاربردهای فراوانی در آمار دارد، استفاده شد.

۳-۴. طراحی و ساخت مدل

همان‌گونه که بیان شد در این تحقیق از دو روش پرکاربرد در تکنیک خوشه‌بندی در داده‌کاوی استفاده شده است. در ابتدا داده‌ها با استفاده از تکنیک شبکه‌های خودسازمان‌ده آموزش داده شد و نتایج به‌دست‌آمده از خروجی شبکه در قالب نقشه‌های گرافیکی مورد تحلیل قرار گرفت. سپس برای مقایسه با روش به‌کارگرفته‌شده در این تحقیق از الگوریتم k -means استفاده شده و نتایج به‌دست‌آمده تحلیل می‌شود.

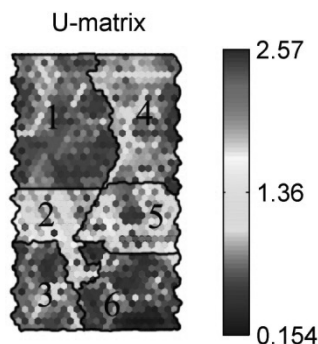
۳-۴-۱. شبکه‌ی خودسازمان‌ده

داده‌های تعلیم شبکه از ۳۴۷۲۶ بردار یازده‌بعدی تشکیل شده است که هر بردار نماینده‌ی یک رکورد از رکوردهای پایگاه داده‌ی بیمه‌ی بدنه‌ی اتومبیل بیمه‌ی ملت است. ابعاد این بردار برابر تعداد فاکتورها است. رابطه‌ی ۱ مجموعه دادگان تعلیم شبکه را تعریف می‌کند:

$$X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6}, X_{i7}, X_{i8}, X_{i9}, X_{i10}, X_{i11}) \quad \text{رابطه ی (۱)}$$

i: ۱ to ۳۴۷۲۶

در خصوص تعداد نرون های لایه ی نقشه تحقیقات گسترده ای انجام گرفته است. کوهنن مبدع شبکه های خودسازمان ده در کتابی با همین عنوان، فرمول $5\sqrt{n}$ را برای تعداد نرون های لایه ی نقشه توصیه می کند که در آن n تعداد داده های آموزش است [۱۳]. در اینجا نیز از همین رابطه استفاده شده و با توجه به اینکه تعداد داده های آموزش ۳۴۷۲۶ نمونه بوده است، تعداد نرون های لایه ی نقشه ۹۳۲ نرون انتخاب شده است. شکل ۳ ماتریس U حاصل از شبکه ی خودسازمان ده آموزش دیده به وسیله ی داده های بیمه ی بدنهی اتومبیل را نشان می دهد.



شکل ۳. ماتریس U شبکه ی خودسازمان ده پس از آموزش داده ها

همان گونه که مشاهده می شود ماتریس U به دست آمده دارای شش بخش یا شش خوشه ی اطلاعاتی است و هریک از مشتریان بیمه ی اتومبیل در یکی از این شش خوشه قرار می گیرند. مشتریان متعلق به هریک از این بخش ها از لحاظ مشخصه های بررسی شده، شباهت نزدیک به یکدیگر داشته و اختلاف زیادی با مشخصه های مشتریان موجود در دیگر بخش ها دارند. به همین سبب انتظار می رود که ریسک مشتریان قرار گرفته در یک گروه نیز مشابه یکدیگر بوده و با ریسک دیگر مشتریان

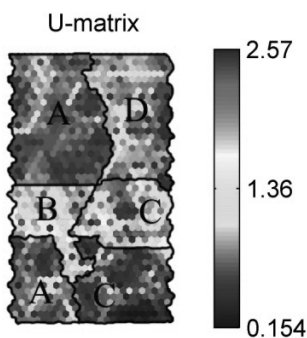
در گروه‌های دیگر متفاوت باشد. به منظور تعیین برچسب و نام هریک از خوشه‌ها، تعدادی نمونه‌ی تصادفی از هر بخش انتخاب شد و میانگین ضریب خسارت نمونه‌های انتخابی محاسبه شد. نسبت خسارت واردشده به شرکت بیمه بر میزان حق بیمه‌ی دریافتی از مشتری ضریب خسارت نام دارد. اگر میزان خسارت پرداختی شرکت بیمه بیش از حق بیمه‌ی دریافتی باشد در حقیقت این ضریب مقدار بالاتر از یک به خود می‌گیرد و به معنای آن است که شرکت بیمه متحمل زیان شده است. جدول ۲ نحوه‌ی برچسب‌زنی بخش‌های نقشه‌ی خودسازمان‌ده را نشان می‌دهد. همان‌گونه که در جدول ۲ قابل مشاهده است میانگین ضریب خسارت خوشه‌ی یک و خوشه‌ی سه نزدیک به هم هستند، از این‌رو هر دو خوشه در طبقه‌ی مشتریان با ریسک کم جای گرفتند.

جدول ۲. برچسب‌زنی گروه‌های مختلف مشتریان براساس ریسک

بخش	نمونه	ضریب	میانگین ضریب خسارت	میزان ریسک بخش
۱	۱	۰.۱۵	۰.۱۷	پایین
	۲	۰		
	۳	۰.۳۶		
۲	۱	۰.۳۸	۰.۴۵	نسبتاً پایین
	۲	۰.۴۴		
	۳	۰.۵۴		
۳	۱	۰	۰.۱۰	پایین
	۲	۰.۳۰		
	۳	۰		
۴	۱	۱.۴۰	۱.۵۰	بالا
	۲	۲.۲۰		
	۳	۰.۹۰		
۵	۱	۱	۱.۰۳	بالا
	۲	۱.۲۰		
	۳	۰.۹۰		
۶	۱	۰.۷۷	۰.۷۴	نسبتاً بالا
	۲	۰.۶۵		
	۳	۰.۸		

همچنین دو خوشه ی چهار و پنج نیز میانگین ضریب خسارت نزدیک به هم دارند و از این سبب این دو خوشه هم با توجه به آنکه میانگین ضریب خسارت آنها بیشتر از یک است، در طبقه ی مشتریان با ریسک زیاد جای گرفتند. این بدان معنی است که خسارت وارد شده به بیمه از سوی مشتریان این خوشه بیش از حق بیمه ی پرداختی است، از این رو به طور کلی می توان ماتریس U را به چهار خوشه ی مطابق شکل ۴ تقسیم بندی کرد:

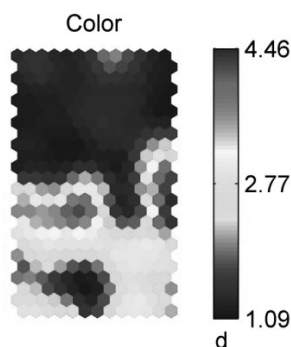
مشتریان خوشه ی A مشتریانی با ریسک کم، مشتریان خوشه ی B مشتریانی با ریسک نسبتاً کم، مشتریان خوشه ی C مشتریانی با ریسک زیاد و مشتریان خوشه ی D مشتریانی با ریسک نسبتاً زیاد هستند.



شکل ۴. تعیین برچسب هر یک از خوشه های ماتریس U

به طور کلی نقشه های رسم شده امکان سه نوع تحلیل مختلف را برای ما فراهم می آورند. ابتدا می توان با توجه به نقشه ی مربوط به هر متغیر، به تنهایی اطلاعاتی راجع به متغیرها استخراج کرد. شکل ۵ نقشه ی متغیر رنگ خودرو را نشان می دهد. در این نقشه، خودروهای با طیف رنگ روشن به رنگ آبی تیره و خودروهای با طیف رنگ های تیره (مانند مشکی، خاکستری و...) با آبی و خودروهای با رنگ هایی از خانواده ی آبی در این نقشه به رنگ سبز روشن نمایش داده شده و خودروهای با خانواده ای از رنگ های تند (مانند نارنجی، قرمز و...) در نقشه با رنگ قرمز تیره نمایش

داده شده است. نکته‌ای که به‌تنهایی از روی این نقشه می‌توان دریافت این است که فراوانی خودروهای با رنگ‌های روشن و تیره از دیگر رنگ‌ها بیشتر است و خودروهای با رنگ‌های تیره از دیگر خودروها فراوانی کمتری دارد.



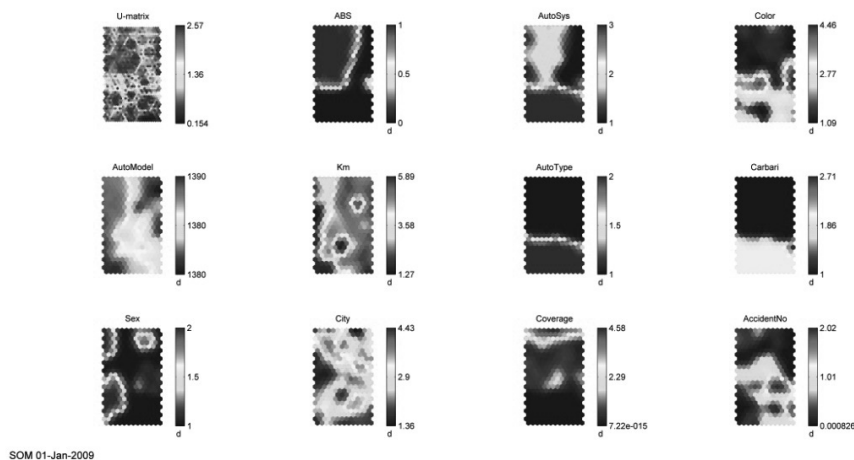
شکل ۵. نقشه متغیر رنگ خودرو

از مقایسه‌ی ماتریس U با نقشه‌ی تک‌تک متغیرها امکان تحلیل ویژگی‌های هر یک از خوشه‌ها فراهم می‌شود. در ادامه به‌عنوان نمونه‌ی خوشه با ریسک نسبتاً زیاد تحلیل می‌شود:

• تحلیل خوشه با ریسک نسبتاً زیاد

- سیستم خودرو: خودروهای این منطقه شامل خودروی پراید و زانتیا است. از آنجایی که رنگ غالب در این ناحیه آبی تیره است، از این‌رو بیشتر خودروهای این ناحیه خودروی پراید است.
- تجهیزات ایمنی (ABS): اکثر خودروهای پراید این ناحیه از این نوع سیستم ترمز بی‌بهره هستند و تعداد کمی از خودروها در این منطقه سیستم ترمز ABS دارند.
- رنگ خودرو: تعداد بیشتری از خودروهای زانتیا در این منطقه رنگ تیره دارند. تقریباً می‌توان گفت خودروهای پراید این منطقه نیز همگی طیف‌های رنگی

- سفید، تیره، آبی و رنگ‌های تند را به خود اختصاص داده‌اند. باین حال رنگ‌های تیره، آبی و رنگ‌های تند بیشتر به چشم می‌خورد.
- مدل خودرو: بیشتر خودروها در این ناحیه مدل‌های ساخت بالا و نسبتاً بالا دارند.
 - میزان کیلومتر مصرفی: خودروهای این ناحیه میزان کیلومتر مصرفی نسبتاً زیادی دارند (حدود ۳۰۰۰۰-۱۰۰۰۰۰ کیلومتر). همچنین در برخی قسمت‌ها کیلومترهای کم و نسبتاً کم نیز قابل مشاهده است.
 - نوع خودرو: نوع خودروهای این منطقه همه سواری هستند.
 - کاربری خودرو: کاربری خودروها در این ناحیه از نوع شخصی است.
 - جنسیت بیمه‌گذار: در این ناحیه، هم بیمه‌گذاران زن و هم مرد هر دو به چشم می‌خورند. فراوانی بیمه‌گذاران مرد در این ناحیه بیشتر است.
 - شهر محل زندگی: بیشتر شهرهای این ناحیه شامل شهرهای بزرگ و نسبتاً بزرگ است. طیف رنگی غالب در این منطقه طیف آبی روشن و زرد است.
 - میزان پوشش بیمه‌ای خریداری‌شده: با توجه به اینکه در برخی قسمت‌ها میزان پوشش بیمه‌ای بسیار زیاد، با وجود این اکثریت خودروها از پوشش بیمه‌ای زیاد بی‌بهره هستند.
 - تعداد خسارت در سال قبل: میزان خسارت خودروها در این ناحیه شامل خسارت‌های صفر و یک است.
- همچنین تحلیل دیگری که می‌توان با کمک نقشه‌ی متغیرها انجام داد، یافتن روابط معنی‌دار و بی‌معنی در بین متغیرها از طریق مقایسه‌ی نقشه‌ی دوبه‌دوی متغیرها است. شکل ۶ نقشه‌ی هم‌زمان همه‌ی متغیرها و ماتریس U را نشان می‌دهد.



شکل ۶. نمایش هم‌زمان ماتریس U و دیگر متغیرها

برای نمونه از روی هم قراردادن نقشه‌ی متغیرهای جنسیت و متغیر شهر محل زندگی می‌توان نتیجه گرفت که فراوانی بیمه‌گذاران زن در کلان‌شهرها و شهرهای بزرگ و نسبتاً بزرگ بیشتر از شهرهای کوچک است.

۳-۴-۲. الگوریتم K-means

در این قسمت داده‌ها با استفاده از الگوریتم k-means به‌عنوان یکی دیگر از پرکاربردترین الگوریتم‌های خوشه‌بندی تفکیکی برای مقایسه با شبکه‌ی عصبی خودسازمان‌ده آموزش داده شد و مدل‌سازی گردید.

در این تحقیق از شاخص میانگین مربعات خطا برای تعیین تعداد خوشه‌ی اولیه استفاده شد. در این مرحله مدل بر روی تعداد خوشه‌های از دو خوشه تا ۱۲ خوشه تکرار شد (با توجه به آنکه تعداد خوشه‌ها در مدل شبکه‌ی خودسازمان‌ده برابر شش تعیین شده است، از این‌رو در این قسمت بیشترین تعداد خوشه ۱۲ خوشه در نظر گرفته شد و نیازی به تکرار مدل بر روی تعداد خوشه‌های بیشتر نیست). کمترین خطا مربوط به تعداد هشت خوشه است، از این‌رو تعداد خوشه‌ی بهینه برابر هشت

خوشه است که با این تعداد خوشه مدل با ۱۶ تکرار به اتمام رسید. به منظور پیاده سازی این مدل از محیط نرم افزار Clementine 11.1 به عنوان یکی از تخصصی ترین نرم افزارهای موجود در زمینه ی داده کاوی استفاده شد. جدول ۳ نتایج به دست آمده از پیاده سازی را نشان می دهد. برای متغیرهایی که نوع آنها از نوع متغیر پیوسته است، میانگین آن متغیر در آن خوشه و برای دیگر انواع متغیرها درصد فراوانی آنها مشخص شد. در جدول ۳ بالاترین فراوانی در هر خوشه مشخص شده است. همان گونه که نشان داده شد، ویژگی های هر یک از خوشه ها پس از خوشه بندی مشخص شد. برای نمونه خوشه ی یک شامل همه ی بیمه گذاران خانمی است که دو نوع خودروی پراید و زانتیا دارند که مجهز به سیستم ترمز ABS هستند. رنگ غالب خودروها در این ناحیه متعلق به خانواده ی طیف رنگ های روشن است. شهر محل زندگی بیمه گذاران در خوشه ی یک با فراوانی ۲۹ درصد بیشتر متعلق به شهرهای نسبتاً بزرگ است. همچنین میانگین سابقه ی خسارت داده های خوشه ی یک ۲۸ درصد است.

جدول ۳. ویژگی هر یک از متغیرها در خوشه های به دست آمده با استفاده از k-means

متغیرها شماره خوشه	مشخصات خودرو							مشخصات بیمه گذار		سابقه بیمه گذار	بیمه نامه
	مدل	ترمز ABS	میزان کارکرد	سیستم	نوع	کاربری	رنگ	جنسیت	شهر محل زندگی	سابقه خسارت	میزان پوشش بیمه ای
1	1386	1->100%	1->24%	1- >52.14%	1->100%	1->99%	1->61%	2->100%	3->29%	0.334	1.5
2	1386	0->100%	3->34%	3->100%	2->100%	2->100%	1->45%	1->75%	3->32%	0.35	0
3	1385	1->100%	3->49%	1->100%	1->100%	1->100%	1->66%	1->100%	3->37%	0.333	1.5
4	1385	1->100%	6->32%	2->100%	1->100%	1->100%	1->54%	1->100%	2->40%	0.432	1.4
5	1383	0->85%	5->44%	1->100%	1->100%	1->98%	2->56%	1->97%	2->28%	0.416	1
6	1386	1->100%	1->33%	2->71%	1->100%	1->100%	1->62%	1->100%	1->100%	0.22	2.2
7	1385	0->100%	6->28%	3->100%	2->100%	2->100%	3->87%	1->79%	2->54%	0.734	0
8	1383	0->100%	6->100%	1->100%	1->100%	1->99%	2->56%	1->84%	2->32%	0.633	0.9

به منظور برچسب‌زنی هریک از خوشه‌ها همانند مدل طراحی‌شده در شبکه‌ی خودسازمان‌ده، تعدادی نمونه‌ی تصادفی از هر خوشه انتخاب شد و با توجه به میانگین ضریب خسارت نمونه‌های انتخاب‌شده خوشه‌ها براساس ریسک آنها نام‌گذاری شد. میانگین ضریب خسارت داده‌های هر خوشه و برچسب خوشه‌ها در جدول ۴ نشان داده شده است.

جدول ۴. برچسب‌زنی خوشه‌های تشکیل شده در الگوریتم K-means

شماره خوشه	میانگین ضریب خسارت	برچسب خوشه
۱	۰/۱۵	ریسک پایین
۲	۰/۲۴	ریسک پایین
۳	۰/۴۵	ریسک نسبتاً پایین
۴	۰/۱۹	ریسک پایین
۵	۰/۷	ریسک نسبتاً بالا
۶	۰/۲۱	ریسک پایین
۷	۱/۲	ریسک بالا
۸	۰/۹	ریسک بالا

۳-۵. تحلیل نتایج

دو تکنیک شبکه‌ی خودسازمان‌ده و k-means بر روی حجم داده‌های بسیار زیاد و با متغیرهای زیاد قابل استفاده است. در تکنیک k-means تعیین تعداد خوشه‌ها به‌عنوان پارامتر ورودی بسیار بااهمیت است و انتخاب نادرست این پارامتر بر روی نتایج خروجی بسیار تأثیرگذار خواهد بود. یکی از نقاط قوت شبکه‌های خودسازمان‌ده، قابلیت ارائه‌ی خروجی شبکه در قالب نقشه‌های گرافیکی گویا و قابل فهم برای مدیران است که می‌تواند سرعت درک و تفسیر نتایج را برای مدیران و کارشناسان بیشتر و راحت‌تر کند. شبکه‌های خودسازمان‌ده امکان تجزیه و تحلیل داده‌ها را از سه بعد فراهم می‌کنند. با استفاده از نقشه‌ی هریک از متغیرها، به‌تنهایی می‌توان اطلاعاتی را درباره‌ی آن متغیر خاص به‌دست آورد. همچنین امکان یافتن

با کمک ماتریس U و رنگ‌های تشکیل‌شده، امکان یافتن خوشه‌های مختلف فراهم می‌شود. به‌علاوه با استفاده از نقشه‌ی دیگر متغیرها امکان تحلیل‌کردن ویژگی هر متغیر در خوشه‌ی تشکیل‌شده در ماتریس U فراهم می‌شود. در خصوص مسئله‌ی این تحقیق شبکه‌ی خودسازمان‌ده برتری بیشتری نسبت به تکنیک k-means دارد.

نتیجه‌گیری

نرخ حق بیمه در بسیاری از شرکت‌های بیمه‌ای در کشورهای توسعه‌یافته با توجه به متغیرهای گوناگون جمعیت‌شناختی، مشخصات اتومبیل و سابقه‌ی خسارت بیمه‌گذار محاسبه می‌شود. در ایران، نرخ حق بیمه‌ی بدنه‌ی اتومبیل با توجه به تعرفه‌ی اعلام‌شده از سوی بیمه‌ی مرکزی جمهوری اسلامی ایران تعیین می‌شود. در این تحقیق سعی شد با کمک شبکه‌های عصبی خودسازمان‌ده و تکنیک k-means مشتریان بیمه‌ی بدنه‌ی اتومبیل در یکی از شرکت‌های خصوصی بیمه‌ای فعال براساس ریسک آنها خوشه‌بندی شده و نقاط قوت و ضعف هر یک از تکنیک‌ها نشان داده شود. پس از تجزیه و تحلیل خروجی هر دو مدل الگوهای جالبی یافت شد. برای نمونه از بین سه نوع خودروی وانت، زانتیا و پراید که تمرکز اصلی این تحقیق بر آنها بود، مشخص شد که ریسک خودروی وانت از سه خودروی دیگر بیشتر بوده و ریسک خودروی پراید از خودروی زانتیا نیز بیشتر است. شرکت‌های بیمه می‌توانند با یافتن الگوهای مناسب در میان داده‌های ذخیره شده و بهره‌مندی از سیستم طبقه‌بندی ریسک افراد، سعی در افزایش بهره‌وری و سوددهی صنعت بیمه‌ی خود داشته باشند. در این تحقیق با توجه به این موضوع که فراوانی سه نوع خودروی پراید، زانتیا و وانت از دیگر خودروها در شرکت مورد مطالعه بیشتر است، از این رو تنها بر این سه نوع خودرو تمرکز کردیم. یکی از پیشنهادات برای تحقیقات آینده تحلیل و بررسی بر روی انواع خودروهای بیشتر است که فراوانی به نسبت مشابهی با یکدیگر داشته باشند، زیرا این امر بر نتایج تحلیل، تأثیرگذار خواهد بود. همچنین تحلیل بر روی

تک‌تک انواع خودروها می‌تواند نتایج دقیق‌تر و بهتری را برای تأثیرگذاری دیگر متغیرها در خصوص آن خودروی خاص نشان دهد.

منابع و مأخذ

۱. جعفری صمیمی، احمد؛ مرادی، محمدعلی، خصوصی سازی و بیمه‌ی اتومبیل در ایران، همایش بین‌المللی صنعت بیمه، چالش‌ها و فرصت‌ها، ۱۳۸۷.
۲. چوبدار، سروناز، طراحی چهارچوبی برای پیش‌بینی مشتریان بیمه‌ی بدنه‌ی اتومبیل بر پایه‌ی داده‌کاوی، پایان‌نامه‌ی کارشناسی ارشد، دانشگاه تربیت مدرس، ۱۳۸۷.
۳. حسین‌زاده، لیلا، دسته‌بندی مشتریان هدف در صنعت بیمه با استفاده از داده‌کاوی، پایان‌نامه‌ی کارشناسی ارشد، دانشگاه تربیت مدرس، اسفند ۱۳۸۶.
۴. حنفی‌زاده، پیام؛ رستخیز پایدار، ندا، مدلی جهت دسته‌بندی ریسکی گروه‌های مشتریان بیمه‌ی بدنه‌ی اتومبیل براساس ریسک با استفاده از داده‌کاوی، فصلنامه‌ی علمی - پژوهشی بیمه، ۱۳۹۰.
۵. ماجد، وحید، راهکارهای طبقه‌بندی ریسکی بیمه‌گذاران در بازار بیمه: شواهدی از بازار بیمه‌ی بدنه‌ی اتومبیل در ایران، همایش بین‌المللی صنعت بیمه، چالش‌ها و فرصت‌ها، ۱۳۸۷.
6. E.W.T. Ngai a, Yong Hu b, Y.H. Wong a, Yijun Chen b, Xin Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", Decision Support Systems, 2011, Vol. 50, pp. 559-569.
7. E.W.T. Ngai, Li Xiu, D.C.K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications, 2009, Vol. 36, pp. 2592-2602.
8. Fahim, A. M., Saake, G., Salem, A. M., Torkey, F.A. Ramadan, M. A., "K-means for Spherical clusters with large variance in sizes", World Academy of Science, Engineering and Technology, 2008, Vol. 45, pp. 177-182.
9. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, November 1996, Vol. 39, No. 11. 27.
10. Hanafizadeh, P., Mirzazadeh, M. (2010). Visualizing market segmentation using selforganizing maps and Fuzzy Delphi method – ADSL market of a telecommunication company. Expert system with application, 38(1), 198-205
11. Hung, Ch., Tsai, Ch., "Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand", Expert Systems with Applications, 2008, Vol. 34, pp: 780-787.
12. Kohonen, T., Self-Organizing Maps, Springer series in Information Sciences, 30,

- Springer, Berlin, Heidelberg New York, 2001, 3th Ed.
13. Kohonen, T., "Automatic formation of topological maps of patterns in a self-organizing system, In Oja, E. and Simula, O. (Eds), proceedings of SCIA Scand. Conference on Pattern Recognition, Los Alamitos, CA. IEEE Computer Soc. Press, 1981, pp. 182-185.
 14. Lee, S. C., Suh, Y. H., Kim, J. K., and Lee, K. J., "A cross-national market segmentation of online game industry using SOM." *Expert Systems with Applications*, 2004, Vol.27, pp: 559-570.
 15. Leung, Y., Zhang, J., Xu, Z., "Clustering by scale-space filtering". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22 (12), pp. 1396-1410.
 16. Newstead, S., D'Elia, A., "Does vehicle colour influence crash risk", *Safety Science*, 2010. Vol. 48, pp. 1327-1338.
 17. Nong Ye, *The Hand Book of Data Mining*. New Jersey, LAWRENCE ERLBAUM ASSOCIATES, 2003
 18. Vesanto, J., Alhoniemi, E., "Clustering of the Self-Organizing Map, *IEEE Transactions on Neural Networks*", 2000, Vol. 11, No.3, pp. 586-600.
 19. Vesanto, J., *Data mining techniques based on the Self-Organizing map*, Helsinki university of Technology, 1997.
 20. Witten IH, Frank E, 2000, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems.
 21. Woo, J. Y., Bae, S. M., & Park, S. C., Visualization method for customer targeting using customer map. *Expert Systems with Applications*, 2000, Vol. 28, pp: 763-772.
 22. Yang, Y., & Padmanabhan, B., "A hierarchical pattern-based clustering algorithm for grouping web transactions". *IEEE Transaction on Knowledge and Data Engineering*, 2005, Vol.17, pp.1300-1304.