استفاده از روش‌های داده‌کاوی برای پیش‌بینی سطح خسارت مشتریان بیمه بدنه اتومبیل

سیدمحمود ایزدپرست * | رئیس اداره نرم‌افزار و سیستم دبیرخانه مجمع تشخیص مصلحت نظام و مدرس دانشگاه تربیت معلم

احمد فراهی[1] | استادیار دانشگاه پیام نور

فرامرز فتح‌نژاد[2] | استادیار مدیریت فناوری اطلاعات بیمه مرکزی ج.ا. ایران

بابک تیمورپور[3] | استادیار دانشگاه تربیت مدرس

دریافت: ۱۳۸۹/۰۳/۲٤ | پذیرش: ۱۳۸۹/۱۲/۰٤

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۸۲۲۳-۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، SCOPUS وISC
http://jipm.irandoc.ac.ir
دوره ۲۷ | شمارهٔ ۳ | ص ص ۶۹۹- ۷۲۲
بهار ۱۳۹۱

نوع مقاله: پژوهشی

چکیده: امروزه، نقش مشتریان از حالت پیروی از تولیدکننده، به هدایت سرمایه‌گذاران، تولیدکنندگان و حتی پژوهشگران و نوآوران مبدل گشته است، به‌همین دلیل سازمان‌ها نیاز دارند مشتریان خود را بشناسند و برای آنان برنامه‌ریزی کنند. تاکنون از برخی روش‌های آماری و یادگیری ماشینی برای این منظور استفاده شده است که البته این روش‌ها به‌تنهایی دارای محدودیت‌هایی هستند که در این پژوهش سعی شده است تا با بهره‌گیری از روش‌های مختلف داده‌کاوی تا حد ممکن این محدودیت‌ها از بین برده و برطبق آن، چارچوبی برای شناسایی مشتریان بیمه بدنه اتومبیل ارائه شود. درواقع، هدف این است تا مشتریانی را که بیشتر به یکدیگر شبیه هستند دسته‌بندی و با استفاده از این دسته‌ها و ویژگی‌های آن، میزان خطرپذیری هر دسته را پیش‌بینی کرد. حال با استفاده از این معیار (میزان خطرپذیری هر دسته) و نوع بیمه‌نامه مشتری می‌توان میزان خسارت او را پیش‌بینی کرد که این معیار می‌تواند کمک شایانی برای شناسایی مشتریان و سیاستگذاری‌های تعرفه بیمه‌نامه باشد. برای این منظور، از دو روش داده‌کاوی[4]، درخت تصمیم و خوشه‌بندی برای ایجاد مدل پیش‌بینی خطرپذیری مشتریان در صنعت بیمه استفاده شده است. البته فنّ درخت تصمیم برای این منظور نتایج بهتری را به‌دست آورده است، ولی فنّ خوشه‌بندی نیز تفکیک خوبی میان مشتریان ایجاد می‌کند.

کلیدواژه‌ها: داده‌کاوی، بیمه، دسته‌بندی، درخت تصمیم، خوشه‌بندی، خسارت

* s.m.izadparast@gmail.com
1 a.farahi@pnu.ac.ir
2. fatnejad@yahoo.com
3. b.teimourpour@modares.ac.ir

4. Data mining

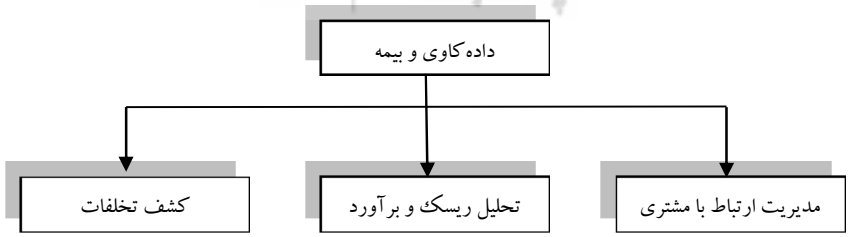
۱. مقدمه

امروزه، با گسترش روزافزون فنّاوری اطلاعات و رقابتی شدن بازار، تولید محصولات باکیفیت به‌صورت انبوه، برای بقا و موفقیت سازمان‌ها کافی نیست و با افزایش اطلاعات مشتریان، فرآیند بازاریابی هر روز پیچیده‌تر و حساس‌تر می‌شود. همچنین، توجه به مبحث مدیریت ارتباط با مشتری هر روز اهمیت بیشتری پیدا می‌کند و بودجه اختصاص‌یافته به آن در سازمان‌ها افزایش می‌یابد (رستمی ۱۳۷۸).

از جمله مسائلی که در صنعت بیمه دارای اهمیت بالا و کمک شایانی در جهت رشد آن است، شناسایی مشتریان و پیش‌بینی سطح خطرپذیری آنان است که این پژوهش در نظر دارد با کاربرد عملی داده‌کاوی در بیمه بدنه اتومبیل، بررسی نماید که چگونه می‌توان مشتریان را براساس ویژگی‌های آنها و میزان خطرپذیری‌شان خوشه‌بندی نمود. پس از آن، سطح خطرپذیری هریک از این خوشه‌ها محاسبه می‌شود. حال مشتریان جدید براساس اینکه به کدام‌یک از این خوشه‌ها شبیه‌تر هستند، در یکی از آنها قرار می‌گیرند و بدین ترتیب می‌توان سطح خطرپذیری آن مشتری را پیش‌بینی نمود. برای این منظور، از روش‌های داده‌کاوی، جهت دستیابی به قوانین تصمیم‌گیری و ایجاد مدل برای پیش‌بینی خطرپذیری مشتریان در صنعت بیمه استفاده شده است. به‌عبارت دیگر، این پژوهش با استفاده از روش‌های داده‌کاوی، مشتریان را براساس رفتار و ویژگی‌های‌شان خوشه‌بندی و سپس مشتریان آتی را با توجه به اینکه به کدام خوشه شبیه‌تر هستند، دسته‌بندی می‌نماید. در این مقاله، ابتدا روش‌های پیشین که برای این منظور استفاده می‌شده است، سپس مفهوم داده‌کاوی توضیح داده می‌شود. پس از آن، روش پیشنهادی بیان و درادامه، چگونگی پیاده‌سازی آن ارائه می‌شود. درنهایت، به ارزیابی مدل ایجادشده و تحلیل نتایج به‌دست‌آمده پرداخته می‌شود.

۲. روش‌های پیشین و فعالیت‌های مرتبط با پژوهش

پس از بررسی و مطالعه مقالات مربوط به کاربردهای داده‌کاوی، تقسیم‌بندی به‌صورت شکل ۱ از حوزه‌های داده‌کاوی در بیمه به‌دست آمد که در ادامه شرح داده می‌شود:



شکل ۱. دسته‌بندی پژوهش‌ها در زمینه داده‌کاوی

دسته اول پژوهش‌ها سعی کرده‌اند با استفاده از داده‌کاوی خطر مشتریان را برآورد نمایند. خطر مشتریان براساس مقدار خسارت و احتمال وقوع آن برای افراد تعیین می‌شود. از جمله روش‌های تعیین خسارت، روش‌هایی است که در مقالات لین اسمیت ارائه شده است. در این مقالات، از طریق روش‌های رگرسیونی و شبکه‌های عصبی میزان خسارت پیش‌بینی شده‌اند (Lin 2009; Smith 2000).

یکی از مسائل مهم در صنعت بیمه، شناخت عوامل و متغیرهایی است که برای پیش‌بینی احتمال و مقدار خسارت اهمیت دارند. هرچند برخی از این عوامل خطر واضح‌اند، روابط بین آنها و نحوه تأثیرگذاری آنها بر یکدیگر مشخص نیست. فنون داده‌کاوی مانند درخت تصمیم و شبکه عصبی می‌تواند عوامل خطر را پیش‌بینی کند. در مقاله‌ای، گوو با استفاده از درخت تصمیم عوامل مهم در مقدار فرکانس خسارت را شناسایی کرده است (Guo 2002).

دسته دوم مقالات مربوط می‌شود به کشف و کنترل تخلفات که سهم عمده‌ای در کاهش هزینه‌های شرکت‌های بیمه دارد، بنابراین لازم است از ابزارهایی برای تسهیل و تسریع و افزایش دقت این فرایند استفاده شود. داده‌کاوی یکی از ابزارهایی است که می‌تواند برای این منظور به‌کار رود (Morley 2006). به‌طور کلی، در همه روش‌های ارائه‌شده برای کشف تخلفات به‌دنبال یافتن ویژگی‌های تخلفات و متخلفان هستند. درادامه، برخی از روش‌های ارائه‌شده برای کشف تخلفات به‌خصوص تخلفات بیمه‌ای از ابعاد مختلف، انواع داده به‌کاررفته، فنون و ... مورد بررسی قرار می‌گیرند. به‌طور کلی، روش‌های استفاده‌شده برای کشف تخلفات را می‌توان به سه دسته تقسیم کرد:

- روش‌های رده‌بندی که از جمله آن می‌توان به مقالات مورلی[1] و دالکلیک[2] اشاره کرد؛
- روش‌های خوشه‌بندی که از جمله آن می‌توان به مقالات کاسترو[3] و کو[4] اشاره کرد؛ و
- ترکیب روش‌های رده‌بندی و خوشه‌بندی که از جمله آن می‌توان به مقالات سوماتی[5] و چان[6] اشاره کرد.

دسته سوم پژوهش‌ها، از فنون داده‌کاوی برای گروه‌بندی مشتریان و تحلیل الگوهای رفتاری مشتریان در حوزه مدیریت ارتباط با مشتری استفاده کرده است. در این مقالات که بیشتر از روش‌های خوشه‌بندی و رده‌بندی استفاده می‌گردد، سعی شده است تا با استفاده از روش‌های

1. Morley
2. Dalkilic
3. Castro
4. Kuo
5. Sumathi
6. Chann

خوشه‌بندی ابتدا گروه‌های لازم ایجاد گردد، سپس با استفاده از روش‌های رده‌بندی هرکدام از مشتریان براساس ویژگی‌های‌شان در هریک از این گروه‌ها قرار گیرند. از جمله مقالاتی که در این زمینه منتشر شده است می‌توان به مقالات ساها[1] و داس[2] اشاره کرد.

۳. مفهوم داده‌کاوی

از زمانی که علم آمار به‌وجود آمد، دانشمندان نیاز به کشف خصوصیات داده‌ها را احساس کرده بودند. با استفاده از آمار و روش‌های آن در آن زمان، خصوصیات داده‌ها از قبیل پراکندگی و تمرکز آنها بررسی می‌شد (Chen 2006). زمانی‌که قصد بررسی تأثیر تعداد کمی از عوامل بر روی هدف است، به‌طور معمول روش‌های آماری مناسب هستند، ولی زمانی‌که تعداد این عوامل زیاد می‌شود، دیگر این روش‌ها کارایی مناسبی ندارند و حتی در مواردی هم ناتوان هستند. به‌عنوان مثال، در تحلیل داده‌های سبک زندگی افراد[3] به‌دلیل اینکه این داده‌ها دارای ابعاد بسیار زیادی هستند، کمتر از روش‌های آماری استفاده می‌شود. دانشمندان برای رفع این مشکل تصمیم گرفتند که از سرعت بالای کامپیوترها استفاده نمایند. همین امر سبب شد که روش‌های ابتکاری دیگری علاوه بر روش‌های آماری مثل شبکه‌های عصبی و الگوریتم ژنتیک ایجاد شود (Lee 2006).

داده‌کاوی عبارت است از "استخراج اطلاعات و دانش و کشف الگوهای پنهان از یک پایگاه داده‌های بسیار بزرگ". این الگوها و دانش‌ها به‌طور معمول مستتر در داده هستند (Chan and Lewis 2002).

از داده‌کاوی می‌توان برای انجام کارهایی مثل دسته‌بندی، پیش‌بینی[4]، تخمین[5]، و خوشه‌بندی[6] داده‌ها استفاده نمود. برای انجام این کارها فنونی توسعه یافته‌اند که با توجه به پیشرفت پیشرفت کامپیوترها و این علم همه‌روزه بر تعداد و کیفیت این فنون افزوده می‌شود. تعدادی از معروف‌ترین این فنون عبارتند از: الگوریتم‌های خوشه‌بندی[7]، شبکه‌های عصبی، الگوریتم ژنتیک[8]، نزدیک‌ترین همسایگی[9]، و درخت تصمیم‌گیری.

---

1. Saha
2. Das
3. Demographic
4. Prediction
5. Estimation
6. Clustering
7. Cluster Detection Algorithm
8. Genetic Algorithm
9. Nearest Neighboring

٤. فنون دسته‌بندی و خوشه‌بندی

دسته‌بندی و خوشه‌بندی از مسائل متعارفی است که به‌طور گسترده توسط متخصصان آمار و پژوهشگران فراگیری ماشینی مطالعه شده است. ارائه یک تعریف دقیق از این دو روش، دشوار است، اما مطابق با تعریف کلی، فنّ دسته‌بندی و خوشه‌بندی، جداسازی یا قرار دادن اجزا یا اشیا در تعدادی از رده‌هاست که در خوشه‌بندی این رده‌ها از قبل وجود ندارند و طی فرآیند و با توجه به ویژگی‌های اشیاء ایجاد می‌شوند، ولی در دسته‌بندی این رده‌ها وجود دارند و اشیاء براساس ویژگی‌های‌شان در این رده‌ها قرار می‌گیرند (Tan and Steinbach 2006).

٥. درخت تصمیم

درخت تصمیم روشی معروف برای دسته‌بندی است که نتایج آن در یک فلوچارت شبیه ساختار درخت ارائه شده است که هر گره[1] نشانگر یک آزمون بر روی ارزش مشخصه و هر شاخه، خروجی هر آزمون را نمایش می‌دهد؛ برگ‌های درخت نیز نمایانگر رده‌ها هستند. به‌طور عادی، پیچیدگی یک درخت تصمیم با افزایش تعداد مشخصه‌ها افزایش می‌یابد. اگرچه در بعضی از شرایط دیده شده است که فقط تعداد کمی از مشخصه‌ها می‌توانند رده‌ای را که هر شی به آن تعلق دارد، تعیین کنند و بقیه مشخصه‌ها کم و یا بی‌تأثیرند (Tan and Steinbach 2006).

در ساخت درخت‌های تصمیم، به‌طور معمول داده‌ها را به دو دسته تقسیم می‌کنند: ۱) داده‌های آموزشی[2] که برای ساخت مدل مورد استفاده قرار می‌گیرند، و ۲) داده‌های آزمون[3] که برای آزمون و ارزیابی مدل ساخته‌شده کاربرد دارند. کیفیت داده‌های آموزشی، بیشتر نقش مهمی را در تعیین کیفیت درخت تصمیم بازی می‌کند. درصورتی‌که آموزش نظام زیاد شود، یعنی داده‌هایی که برای آموزش و ساخت مدل به‌کار می‌روند درصد زیادی از داده‌ها باشند، دچار حالتی به‌نام "آموزش بیش از حد مدل"[4] خواهیم شد که به‌دلیل وجود موارد غیرعادی در داده‌های داده‌های آموزشی، خطا تولید می‌کند (Chan and Lewis 2002).

٦. خوشه و خوشه‌بندی

خوشه‌بندی، درواقع یک عملیات غیرنظارتی است؛ این عملیات هنگامی استفاده می‌شود که به‌دنبال یافتن گروه‌هایی از داده‌های مشابه باشیم، بدون اینکه از قبل پیش‌بینی در مورد شباهت‌های موجود داشته باشیم. خوشه‌بندی، به‌طور معمول هنگامی استفاده می‌شود که به‌دنبال یافتن

1. Node
2. Train Data
3. Test Data
4. Over Fitting

گروه‌هایی از مشتریان هستیم که پیشتر شناخته نشده‌اند، یعنی گروه‌ها از قبل وجود ندارند و در فرآیند خوشه‌بندی ایجاد می‌شوند. برای مثال، می‌توان شباهت‌های مشتریان در استفاده از تلفن همراه را به‌منظور گروه‌بندی مشتریان و تشخیص خدمتی جدید جستجو نمود (Borgelt 2008).

۷. روش K-means

الگوریتم k-means هر نقطه را به خوشه‌ای که مرکز[1] آن نزدیکتر باشد، اختصاص می‌دهد. مرکز، میانگین تمام نقاط یک خوشه است، یعنی مختصات آن میانگین ریاضی داده‌های خوشه در هر بعد به‌صورت جداگانه است.

گام‌های الگوریتم عبارتند از:

۱) تعداد خوشه‌ها را انتخاب کن k؛
۲) به‌صورت تصادفی K خوشه ایجاد کن و مرکزهای آنها را تعیین کن؛
۳) هر نقطه را به نزدیک‌ترین مرکز اختصاص بده؛
۴) مرکز خوشه‌ها را باز محاسبه کن؛ و
۵) دو گام پیشین را آنقدر تکرار کن تا شرط همگرایی ارضا شود.[2]

مزیت اصلی این الگوریتم سادگی و سرعت آن است که به آن اجازه می‌دهد تا بر روی دسته داده‌های بزرگ اجرا شود. عیب آن این است که در هر اجرای الگوریتم، همان نتایج را نمی‌دهد، زیرا خوشه‌های حاصل به دسته‌بندی تصادفی اولیه وابسته هستند. این الگوریتم واریانس درون خوشه‌ای را کمینه می‌کند، ولی تضمینی برای واریانس کل نمی‌دهد. عیب دیگر آن است که الگوریتم نیاز دارد تا مفهوم میانگین برای داده‌ها تعریف‌پذیر باشد که همیشه امکان‌پذیر نیست. برای چنین داده‌هایی استفاده از الگوریتم k-medoids مناسب است. انشعابات معروف دیگر الگوریتم k-means شامل الگوریتم k-means سریع ژنتیک[3] (Lu 2004) و الگوریتم k-means افزایشی ژنتیک[4] (Lu 2004) است. البته با وجود این ایرادات این الگوریتم یکی از بهترین روش‌ها برای کار با داده‌ها با حجم بالاست.

1. Center or Centroid

۲. به‌طور معمول، شرط همگرایی تغییر نکردن مرکز خوشه‌هاست.

3. Fast Genetic K-means Algorithm (FGKA)
4. Incremental Genetic K-means Algorithm (IGKA)

۸. روش پیشنهادی

در این پژوهش، از دو فنّ خوشه‌بندی و درخت تصمیم استفاده شده است. در روش خوشه‌بندی، مشتریان براساس ویژگی‌هایشان در خوشه‌هایی تفکیک می‌شوند، سپس میانگین سطح خسارت در هریک از این خوشه‌ها محاسبه می‌شود. حال مشتریان آتی با توجه به اینکه به کدام‌یک از این خوشه‌ها شبیه‌تر هستند، در یکی از آنها قرار می‌گیرند تا سطح خسارت‌شان براساس خوشه‌ای که در آن قرار گرفته‌اند، مشخص شود. در فنّ درخت تصمیم با استفاده از داده‌های مشتریان درختی که براساس قوانین "اگر-آنگاه" است، ایجاد می‌شود، سپس مشتریان جدید از گره ریشه وارد می‌شوند تا به گره برگ برسند. در این قسمت، می‌توان با توجه به ویژگی‌های آن گره سطح خسارت مشتری را پیش‌بینی نمود. درنهایت، هر دو این مدل‌ها مورد ارزیابی قرار گرفته‌اند. روش‌های ارزیابی به‌کاررفته در این پژوهش به دو دسته ارزیابی درونی و بیرونی تقسیم می‌شوند. در ارزیابی درونی، صحت مدل ایجادشده بررسی می‌شود که برای این منظور، از روش پهنه سایه روشن[1] و استفاده از داده‌های آموزش و آزمایش بهره گرفته شده است و در ارزیابی بیرونی، به مقایسه نتایج به‌دست‌آمده از مدل با نظرات کارشناسان خبره بیمه بدنه و تحلیل آن پرداخته شده است.

۹. جامعه آماری و نمونه آماری

جامعه آماری مورد استفاده در این پژوهش، داده‌های مربوط به یک میلیون مشتری بیمه بدنه اتومبیل، از ۱۵ شرکت[2] ارائه‌کننده خدمات بیمه بدنه است که طی پنج سال اخیر جمع‌آوری شده و بالغ بر پنج میلیون رکورد است. این داده‌ها در پایگاه داده بیمه مرکزی ذخیره شده و شامل اطلاعات فردی مشتری و اطلاعات مربوط به خسارات مشتریان است. تعداد کل مشتریانی که دارای سابقه خسارت هستند، حدود ۷۰ هزار مشتری است که در جدولی مجزا قرار گرفته است. همچنین، برای برخی عملیات مانند تحلیل اکتشافی، نمونه‌ای از داده‌ها به‌روش نمونه‌گیری کاهشی ایجاد شده است.

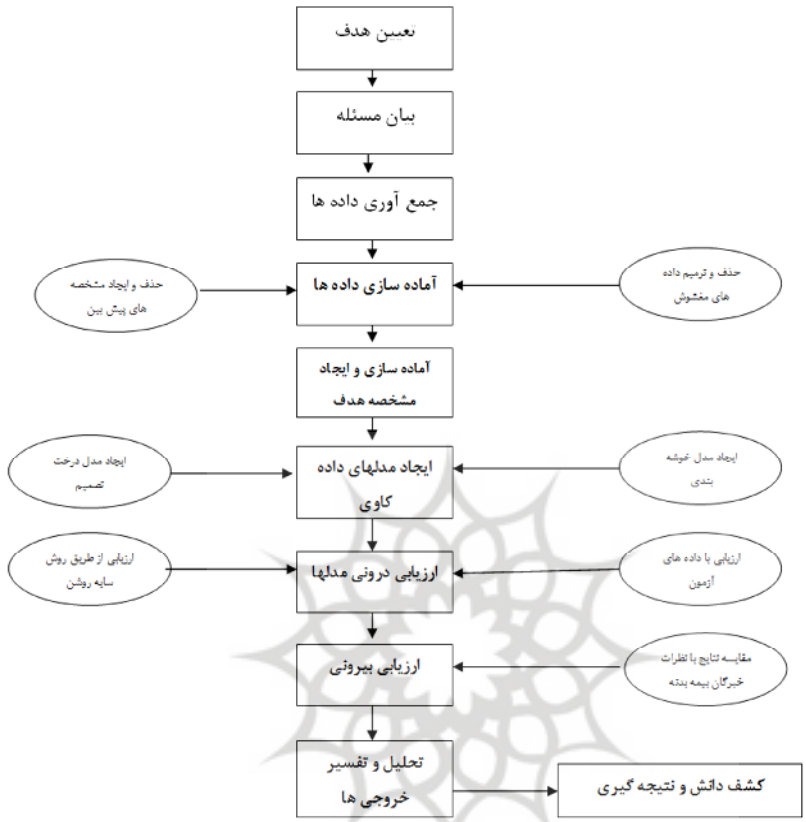
1. Silhouette

۲. شرکت‌های ایران، آسیا، البرز، پارسیان، کارآفرین، پاسارگاد، دانا، دی، رازی، سامان، سینا، معلم، ملت، نوین، و توسعه.

۱۰. مراحل پژوهش

هدف اصلی این پژوهش، کاربرد داده‌کاوی در صنعت بیمه و به‌طور خاص، در بیمه بدنه اتومبیل، به‌منظور شناسایی مشتریان و درنهایت شناسایی سطح خطرپذیری و خطر آنان براساس ویژگی‌های شخصیتی آنهاست. برای این منظور، از پایگاه داده مشتریان بیمه بدنه اتومبیل در بیمه مرکزی که بالغ بر پنج میلیون رکورد است، استفاده شده است. برای این منظور، ابتدا به آماده‌سازی داده‌ها که شامل: ۱) جمع‌آوری داده‌های بیمه‌گذاران و اتومبیل و خسارات، ۲) تجمیع داده‌های سوابق هر مشتری در یک رکورد، ۳) حذف فیلدها با داده‌های ناقص، و ۴) حذف رکوردهای مغشوش است و سپس به ساخت مدل‌های تحلیل با استفاده از این داده‌ها پرداخته شد. در این پژوهش، از دو مدل خوشه‌بندی k میانگین و دسته‌بندی درخت تصمیم استفاده شده است که پیشتر در مورد آن صحبت شد. درنهایت، از خوشه‌های به‌دست‌آمده از خوشه‌بندی، برای جداسازی مشتریان و نیز از قوانین تصمیم‌گیری به‌دست‌آمده از درخت تصمیم برای دسته‌بندی مشتریان و قرار دادن آنها در یکی از دسته‌های به‌دست‌آمده استفاده شد تا بدان وسیله بتوان به سطح خطرپذیری و خطر مشتریان آتی دست یافت.

با توضیحات ارائه‌شده، روال کار بدین صورت تعریف می‌شود که درابتدا باید هدف اصلی از این عملیات را مشخص کرد، اینکه چه نتایجی به‌دست خواهد آمد. درنتیجه براساس این هدف، مشخصه‌هایی را برای این منظور استفاده می‌شود مشخص کرد، سپس برای اینکه این داده‌ها برای انجام عملیات داده‌کاوی مناسب شوند، مجموعه عملیات پیش‌پردازش را روی آن انجام داد. پس از آن، با استفاده از مشخصه‌های به‌دست‌آمده به مدل‌سازی پرداخت و درنهایت، به تحلیل و ارزیابی مدل‌های ایجادشده پرداخت (شکل ۲).



شکل ۲. مراحل پژوهش

۱۱. پیش‌پردازش[1] داده‌ها

داده‌های مورد استفاده در این پژوهش، داده‌های مشتریان بیمه است. این داده‌ها نیز دارای اغتشاش بسیار زیادی است و شاید علت اصلی آن این است که این داده‌ها فقط برای ذخیره سوابق مشتریان کاربرد دارد و دراصل، برای کاربردهای داده‌کاوی جمع‌آوری نشده است. داده‌های بیمه که برای این پژوهش مورد استفاده قرار گرفته است نیز دارای مشکلات بسیاری از جمله مقادیر گمشده، داده‌های متناقض، و پراکنده است. با این وجود، برای اینکه این داده‌ها به‌گونه‌ای تبدیل شود که الگوریتم‌ها و فنون داده‌کاوی قادر به اعمال روی آنها باشند، از روش‌های مختلفی استفاده شده است که عبارتند از:

1. Preprocess

۱) پاک‌سازی داده‌ها، ۲) تبدیل و یکپارچه کردن داده‌ها، ۳) کاهش بعد داده که در (Han and Kamber 2006) بیان شده است.

۱۲. معرفی مشخصه‌ها

همان‌طور که مطرح شد، داده‌های این پژوهش، داده‌های بیمه‌گذاران بیمه بدنه اتومبیل است. این داده‌ها در دو جدول قرار گرفته‌اند که یک دسته شامل داده‌های مربوط به فرد بیمه‌گذار و اتومبیل وی است و جدول دوم شامل داده‌های خسارت است که اطلاعات فرد خسارت‌زننده را شامل می‌گردد. تعداد این فیلدها بسیار زیاد بود (حدود ۶۰ مشخصه) که البته بخش بزرگی از این فیلدها دارای داده‌های بسیار مغشوش و یا بدون مقدار بودند، بنابراین سعی شد تا براساس مصاحبه و جلسات با مدیران فنّی بیمه بدنه اتومبیل[1] و بررسی پژوهش‌های مشابه قبلی، مشخصه‌هایی را که برای این پژوهش مناسب‌ترند انتخاب گردد. درنتیجه، با توجه به اینکه هدف این پژوهش دسته‌بندی بیمه‌گذاران و یافتن روابطی پنهان در میان داده‌ها برای پیش‌بینی‌های آتی است، مشخصه‌های مندرج در جدول ۱، به‌عنوان مشخصه‌های اصلی که بر میزان خسارات بیمه‌گذاران در بیمه بدنه اتومبیل تأثیرگذارند، انتخاب گردیده است.

جدول ۱. فهرست مشخصه‌های اصلی برای تحلیل

داده‌های جمعیت‌شناختی بیمه‌گذاران	داده‌های مربوط به ویژگی‌های اتومبیل	داده‌های خسارات
سن مشتری	نوع استفاده اتومبیل (آموزش، مسافربری، بارکش، آتش‌نشانی، و شخصی)	سطح خسارت
تاریخ صدور گواهینامه مشتری	سال ساخت اتومبیل	میزان خسارت پرداختی
شغل مشتری	تعداد سال که مشتری خسارت نداشته است.	
جنسیت مشتری	مقدار پژوهش‌های مشتری	
وضعیت تأهل مشتری	نوع اتومبیل (اتوبوس، بارکش، کشاورزی، سواری، و موتور)	
شهر محل زندگی مشتری	ظرفیت اتومبیل	
وضعیت بیمه‌نامه (انفرادی، گروهی)	تعداد سیلندر اتومبیل	
سطح تحصیلی مشتری	کد شهر پلاک اتومبیل	
نوع پلاک اتومبیل (دولتی، شخصی)	نوع تیپ اتومبیل	
شهر محل صدور شناسنامه مشتری	نوع مالکیت (حقوقی، خصوصی، و دولتی)	

۱. این مصاحبه به‌منظور پالایش داده‌ها و نیز انتخاب مشخصه‌هایی که در خسارت مشتریان تأثیرگذارند با تعداد هفت نفر از مدیران فنّی بیمه مرکزی ج.ا.ا و بیمه ایران که سابقه زیادی در این زمینه دارند، صورت گرفته است.

این مشخصه‌ها در سه گروه قرار گرفته‌اند که از مشخصه‌های جمعیت‌شناختی و اتومبیل برای پیش‌بینی مشخص خسارت استفاده خواهد شد.

۱۳. مشخصه هدف

پژوهش حاضر به‌دنبال پیش‌بینی سطح خطر و خسارت مشتریان است و مشخصه مبلغ کل خسارت، به‌خوبی نشان‌دهنده این مسأله است. ولی باید این نکته را در نظر گرفت که مبلغ خسارت همبستگی بالایی با نوع بیمه‌نامه و مبلغ تعهد بیمه‌نامه دارد. این مشخصه از ویژگی‌های جمعیت‌شناختی شخص نیست و مربوط به نوع وسیله نقلیه و مبلغ وسیله نقلیه اوست. به‌همین دلیل، برای حذف آن و نرمال‌سازی سطوح خطرپذیری، مبلغ خسارت بر مبلغ تعهد بیمه‌نامه تقسیم شده است تا نتایج نرمال شود و بتوان مشخصه سطح خطرپذیری مشتری را گسسته ساخت. این مشخصه، سطح خسارت به‌معنای سطح خطرپذیری مشتری نامیده می‌شود. در جدول ۲، حدود این سطوح مشخص شده است.

جدول ۲. سطوح خسارت تعریف‌شده برپایه مبلغ خسارت پرداختی بیمه به مبلغ تعرفه بیمه‌نامه

سطح خسارت	حد پایین	حد بالا
۱	کمتر از یک برابر	یک برابر
۲	یک برابر	دو برابر
۳	دو برابر	پنج برابر
۴	پنج برابر	ده برابر
۵	ده برابر	بیشتر از ده برابر

۱۴. انتخاب مشخصه

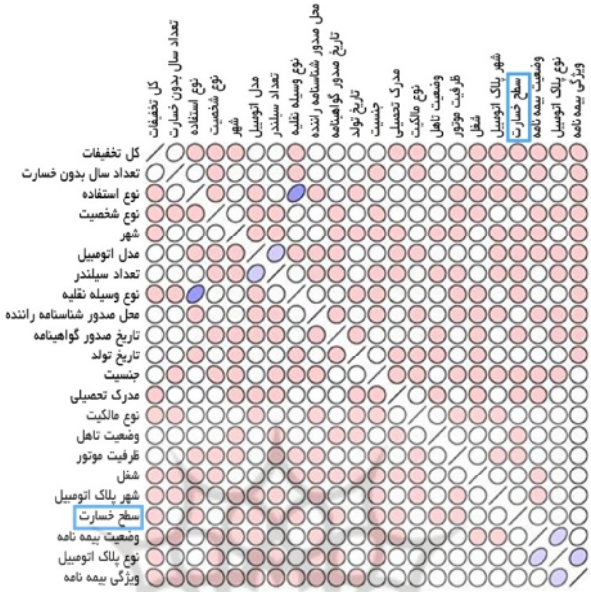
همان‌طور که پیشتر اشاره شد، تعداد مشخصه‌ها حدود ۶۰ مشخصه بود که شامل مشخصه‌هایی در مورد شخص بیمه‌کننده، اتومبیل، و داده‌های مربوط به خسارت می‌شد. البته این داده‌ها دارای مشکلات زیادی از جمله ناقص بودن، مغشوش بودن و ... بود که درنهایت طی فرایند انتخاب مشخصه، از بین آنها ۲۰ مشخصه اصلی استخراج شد. برای انتخاب مشخصه‌های[1] مناسب در این پژوهش، ابتدا هرکدام از مشخصه‌های موجود بررسی شد. در این مرحله، تعدادی از مشخصه‌ها مانند (نام، نام خانوادگی، شماره شناسنامه و ...) که اطلاعات شخصی افراد بود مربوط

1. Future Selection

به ویژگی‌های رفتاری آنها نمی‌شد؛ مشخصه‌هایی مانند (شماره شاسی، شماره پلاک و ...) که مشخصه‌های یکتای اتومبیل بود و نیز برخی مشخصه‌های مربوط به جدول خسارت مثل (شماره پرونده، وضعیت پرونده و ...) که یکتا بود و نسبت به متغیر هدف تأثیرگذار نبود با توجه به مصاحبه‌ها و مشاوره‌هایی که با کارشناسان بیمه صورت گرفت، حذف شد. درواقع، این متغیرها دارای تعداد حالت‌های بسیار زیادی هستند و علاوه بر آنکه نمی‌توانند عامل مناسبی برای تفکیک اشیاء باشند، ممکن است روی مدل تأثیر منفی بگذارند و نتایج را مغشوش نمایند. مشخصه‌های باقیمانده بررسی شد تا مشخص شود تغییرات هرکدام از آنها نسبت به متغیر هدف چگونه است. درواقع، بررسی شد که آیا آن مشخصه نسبت به متغیر هدف تغییر معنی‌داری دارد یا خیر و اینکه این تغییر به چه صورتی است (مستقیم یا معکوس). برای این منظور، از ماتریس همبستگی[1]، نمودار هیستوگرام[2]، و نمودار جعبه‌ای[3] استفاده شد.

نمودار هیستوگرام نشان‌دهنده توزیع متغیر و اینکه مقادیر به چه شکلی پراکنده شده‌اند است و نمودار جعبه‌ای نشان‌دهنده این مطلب است که داده‌ها در چه بازه‌ای بیشترین تجمع را دارند و به کمک آن می‌توان داده‌های پرت را شناسایی نمود (Han and Kamber 2006). برای این کار، ابتدا با استفاده از نرم‌افزار R یک نمونه از داده‌ها وارد یک صفحه گسترده شد، سپس این داده‌ها در نرم‌افزار بارگذاری شد. پس از آن، با استفاده از دستورات برنامه‌نویسی این نرم‌افزار، نمودارهای تمام مشخصه‌ها رسم شد تا توزیع هرکدام از آنها مشاهده شود. در پایان، ماتریس همبستگی بین این مشخصه‌ها ایجاد گردید (شکل ۳). ماتریس همبستگی، درواقع نشانگر رابطه بین دو متغیر است بدین معنی که تا چه اندازه‌ای این دو متغیر به یکدیگر وابسته هستند یا به‌عبارت دقیق‌تر اینکه تا چه اندازه می‌توان مقدار یک متغیر را از روی دیگری به‌دست آورد که این معیار با ضریب همبستگی شناخته می‌شود (Tan and Steinbach 2006).

1. Correlation
2. Histogram chart
3. BoxPlot



شکل ۳. نمایش گرافیکی ماتریس همبستگی

همان‌طور که در شکل ۳ مشاهده می‌شود، بیشتر مشخصه‌ها همبستگی دو به دو با یکدیگر ندارند و نیز مشخصه سطح خسارت که به‌عنوان متغیر هدف است با مشخصه‌های دیگر همبستگی مشخصی ندارد. در صورتی که متغیر هدف، با مشخصه‌ای همبستگی بالایی داشته باشد باید آن مشخصه را حذف کرد، زیرا در صورت وجود آن مشخصه پیش‌بینی به‌دست‌آمده فقط ناشی از آن مشخصه خواهد شد و تأثیر سایر مشخصه‌ها بر مشخصه هدف نشان داده نخواهد شد.

۱۵. نرم‌افزارهای داده‌کاوی به‌کارگرفته‌شده

برای انجام عملیات داده‌کاوی، نرم‌افزارهای زیادی موجود است که می‌توان برحسب نیاز از آنها استفاده نمود. از جمله این نرم‌افزارها:

۱. Microsoft SQLServer که قابلیت کار با حجم بالای داده را دارد، ولی امکانات آن هنوز محدود است.

۲. Rapid Miner که نرم‌افزار قدرتمندی برای ساخت مدل‌های زیادی از داده‌کاوی است.

۳. R درواقع، نرم‌افزار آماری است که برای انجام عملیات تحلیل اکتشافی بسیار مناسب است.

در این پژوهش، با توجه به اینکه حجم داده‌ها بسیار بالاست، برای مدل‌سازی از نرم‌افزار

Microsoft SQLServer استفاده شده است. درواقع، مجموعه عملیات پیش‌پردازش توسط خود این نرم‌افزار انجام می‌شود و برای انجام عملیات داده‌کاوی و مدل‌سازی از قسمتی از این نرم‌افزار با نام SQLServer Analysis Services استفاده می‌شود. درواقع، این نرم‌افزار برای پشتیبانی از OLAP[1] است که هر دو هدف داده‌کاوی[2] و انباره داده[3] را پشتیبانی می‌کند. بدین صورت که می‌توان یک انباره داده ایجاد کرد و سپس روش‌ها و مدل‌های داده‌کاوی را روی این انباره اعمال نمود. البته برای عملیات تحلیل اکتشافی با توجه به محدودیت‌های این نرم‌افزار از نرم‌افزارهای R و Rapid Miner استفاده شده است.

۱۶. پیاده‌سازی مدل‌های داده‌کاوی

همان‌طور که پیشتر بیان شد، در این پژوهش از دو روش درخت تصمیم و خوشه‌بندی استفاده شده است که درادامه به بیان چگونگی مدل‌سازی آن پرداخته می‌شود.

۱۶-۱. خوشه‌بندی

در این پژوهش، از الگوریتم خوشه‌بندی K-means که توسط شرکت مایکروسافت[4] ارائه شده و در نرم‌افزار SQLServer نیز موجود است، استفاده می‌شود. همان‌طور که پیشتر اشاره شد، این الگوریتم با وجود مزایای زیادی که دارد معایبی را نیز شامل می‌شود که در این اینجا برای حل هریک از معایب این الگوریتم، راهکارهایی ارائه شده است که درادامه بیان می‌شود. این الگوریتم برای حل مشکل تعداد خوشه بهینه از مجموعه‌ای از روش‌های هیوریستیک[5] استفاده می‌نماید به‌گونه‌ای که داده‌ها را با مقادیر مختلف k خوشه‌بندی می‌کند و سعی می‌نماید تا با استفاده از روش‌های بهینه‌سازی مثل الگوریتم ژنتیک[6] و ... مقدار بهینه k را به‌دست آورد. البته برای اینکه محاسبه شود که خوشه‌های به‌دست‌آمده دارای کیفیت مناسب هستند یا خیر و اینکه تفکیک خوبی صورت گرفته است یا خیر، می‌توان از روش‌هایی مثل پهنه سایه روشن[7] استفاده نمود. این روش‌ها براساس محاسبه تفاوت بین خوشه‌ها و نیز همسایگی هر عنصر می‌تواند تعداد خوشه‌های مناسب را تشخیص دهد. در مورد ایراد دومی که بر الگوریتم K-means وارد می‌شود نیز تمهیداتی در نظر گرفته شده است. در این پژوهش، برای جلوگیری از مواجهه با این مشکل سعی شده است

1. Online Analysis Services
2. Data Mining
3. Data Warehouse
4. Microsoft
5. heuristics
6. Genetic Algorithm
7. Silhouette

تا با استفاده از روش‌های دیگر مثل خوشه‌بندی سلسله مراتبی، نقاط مناسب مرکز خوشه تشخیص داده شود. البته با توجه به محدودیت‌های این روش و هزینه بالای آن از نظر زمان و حافظه، از نمونه‌ای از داده‌ها برای این الگوریتم استفاده شده است. همچنین، در روش خوشه‌بندی Microsoft نیز نقاط ابتدایی خوشه به‌صورت کاملا تصادفی نیست بلکه از مجموعه‌ای روش‌های هیوریستیک در مورد پراکندگی داده‌ها، برای تعیین مراکز خوشه‌ها استفاده می‌شود (Robert 2001).

پس از خوشه‌بندی داده‌ها دسته‌هایی ایجاد می‌گردند که هر کدام ویژگی‌هایی دارند و می‌توان با داشتن این دسته مشتریان جدید را براساس صفات‌شان در یکی از این خوشه‌ها قرار داد که به این قسمت از کار یعنی قرار دادن هر کدام از مشتریان در یکی از این دسته‌ها، دسته‌بندی گفته می‌شود.

۱۶-۲. درخت تصمیم

درخت تصمیم ابزاری برای پیش‌بینی است که در اینجا برای پیش‌بینی سطح خسارت از آن استفاده می‌شود. برای ایجاد درخت تصمیم نیز از الگوریتم درخت تصمیم Microsoft استفاده می‌شود. در این الگوریتم مانند سایر روش‌های درخت تصمیم مجموعه‌ای از مشخصه‌ها به‌عنوان ورودی تعریف می‌شوند و یک متغیر به‌عنوان مشخصه هدف. یعنی از مشخصه‌های ورودی برای پیش‌بینی مشخصه هدف استفاده می‌شود. روش کار این الگوریتم به این صورت است که مشخصه‌های ورودی را براساس میزان تأثیرگذاری‌شان بر متغیر هدف درنظر می‌گیرد و به آنها الویت می‌دهد و درنهایت، بر طبق این الویت‌بندی، درخت پیش‌بینی را ایجاد می‌کند (Microsoft 2010).

۱۷. تحلیل مدل‌های ایجادشده

۱۷-۱. تحلیل درخت تصمیم

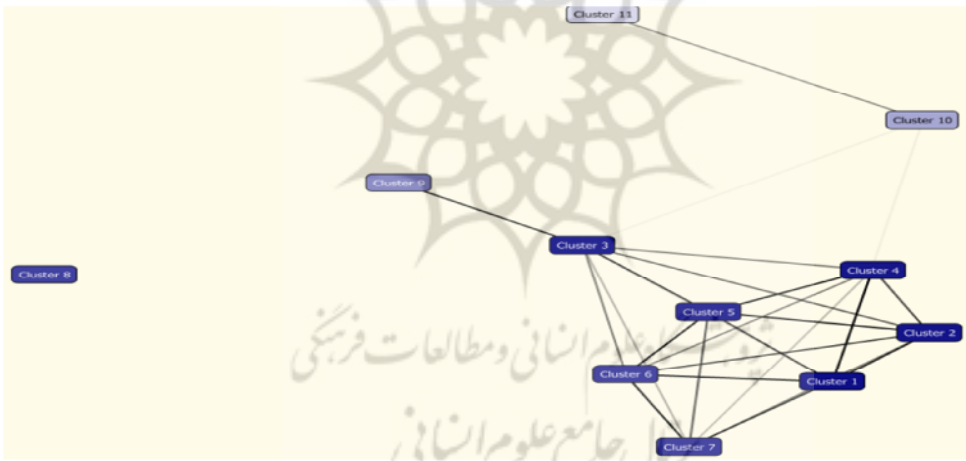
برای ساخت درخت از ۲۰ مشخصه که به‌عنوان مشخصه ورودی تعریف شده بود، استفاده گردید. همچنین، پارامترهای لازم در الگوریتم درخت تصمیم با توجه به مطالبی که در فصل سوم بیان شد، تنظیم گردید تا بتوان درخت مورد نظر را ایجاد کرد. برای ساخت درخت ابتدا با استفاده از ۷۰٪ داده‌ها ایجاد شد و سپس مدل ایجادشده با ۳۰٪ باقیمانده داده‌ها مورد ارزیابی قرار گرفت.

پس از ایجاد، مدل مورد ارزیابی قرار گرفت؛ صحت مدل روی داده‌های آزمون ۶۰٪ به‌دست آمد. برای بررسی صحت نتیجه، با درصد رده غالب سنجیده می‌شود. در این مدل، رده غالب

دارای ۳۵٪ داده‌هاست، یعنی بدون هیچ مدل و به‌صورت تصادفی می‌توان تا دقت ۳۵٪ پیش‌بینی کرد. در اینجا دقت ۶۰٪ دقت بسیار خوبی است.

۱۷-۲. تحلیل خوشه‌بندی

با استفاده از مشخصه‌ها و تنظیم پارامترهای الگوریتم خوشه‌بندی، این فن نیز روی داده‌ها اعمال شد. شکل ۴ نشان‌دهنده خوشه‌های ایجادشده است. خطوطی که بین خوشه‌ها قرار دارد و نیز شدت رنگ آن نشان‌دهنده نزدیکی مراکز این خوشه‌ها به یکدیگر است. همچنین، محل قرار گرفتن آنها براساس الگوریتم MDS مشخص شده است. این الگوریتم در (Tan and Steinbach 2006) بیان شده و مشخص‌کننده فاصله بین خوشه‌هاست. همان‌طور که مشاهده می‌شود خوشه هشتم نسبت به سایر خوشه‌ها دورتر قرار گرفته است و نیز با هیچ‌کدام ارتباط نزدیکی ندارد که نشان‌دهنده استقلال و دور بودن مرکز آن از سایر خوشه‌هاست. به‌عبارت دیگر، در صورت تلفیق آن با سایر خوشه‌ها کیفیت خوشه‌بندی بسیار خراب خواهد شد.



شکل ٤. شمای گرافیکی خوشه‌های ایجادشده

این خوشه‌ها براساس مشخصه‌هایی که پیشتر در مورد آنها صحبت شد و مقادیر آنها به‌وجود آمده‌اند. هر کدام از این خوشه‌ها براساس ۲۰ مشخصه تعیین‌شده بررسی شدند تا مشخص گردد پراکندگی این مشخصه‌ها در هرکدام از این خوشه‌ها به چه صورتی است. به‌عنوان مثال، از نظر جنسیت خوشه ۱ تا ۸ بیشتر مرد بوده و در خوشه‌های ۹ تا ۱۱ بیشتر زنان هستند. همه خوشه‌ها با توجه به مشخصه‌ها مورد بررسی قرار گرفتند. سرانجام با تحلیل این خوشه‌ها از روی ویژگی‌ها و

فراوانی هر کدام از آنها دریافت شد خوشه‌های (۲، ۴، ۱، ۷) دارای سطح خسارت کمتر هستند و خوشه‌های (۵، ۶، ۳، ۹) دارای سطح خسارت متوسط و درنهایت، خوشه ۸ دارای سطح خسارت بالاست. همان‌طور که از شکل ۴ مشخص است، خوشه ۸ دارای فاصله زیاد از سایر خوشه‌هاست که این مطلب نشان‌دهنده مغایرت زیاد اعضای این خوشه با سایر خوشه‌هاست.

۱۸. ارزیابی نتایج

برای ارزیابی نتایج به‌دست‌آمده از دو روش ارزیابی درونی و بیرونی استفاده می‌شود. در ارزیابی درونی، مدل ایجادشده توسط روش‌های داده‌کاوی مورد آزمون قرار می‌گیرد تا صحت آن را سنجیده شود و در ارزیابی بیرونی، نتایج به‌دست‌آمده از مدل، با ذهنیات کارشناسان خبره بیمه بدنه اتومبیل مقایسه می‌شود.

۱۸-۱. ارزیابی درونی

ارزیابی درونی برای تأیید اینکه آیا پارامترهای مدل به‌صورت مناسب، برای هدف در نظر گرفته شده، تنظیم شده است. برای ارزیابی درونی از دو روش استفاده می‌شود:

۱۸-۱-۱. ارزیابی با داده‌های آموزش و آزمایش

در روش اول برای بررسی اینکه مدل ایجادشده تا چه حدی قابل اعتماد است و می‌تواند متغیر هدف را پیش‌بینی نماید داده‌ها به دو دسته آموزش و آزمون تقسیم می‌شود. روش کار بدین صورت است که ابتدا، مدل با استفاده از داده‌های آموزشی ایجاد می‌شود و سپس مدل ساخته‌شده با داده‌های آزمون، آزمون می‌شود تا صحت عملکرد آن به‌دست آید. این روش برای آزمون مدل‌های پیش‌بین بسیار مناسب است (Han and Kamber 2006).

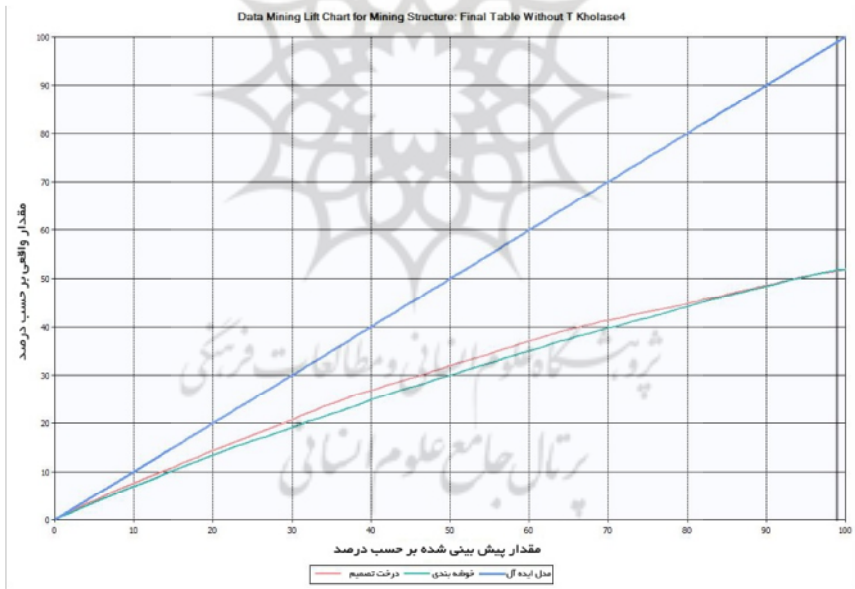
در این پژوهش، داده‌ها به دو دسته ۷۰٪ برای آموزش و ۳۰٪ برای آزمون تقسیم شدند. سپس، با استفاده از داده‌های آموزش مدل درخت تصمیم و خوشه‌بندی ایجاد شدند. پس از آن، ۳۰٪ داده‌های آزمون وارد مدل شدند. از مدل برای پیش‌بینی مشخصه هدف یعنی سطح خطرپذیری مشتری استفاده گردید.

برای نمایش نحوه عملکرد مدل و پیش‌بینی آن از نمودار ترفیع[1] استفاده می‌شود. این نمودار بدین صورت است که در یک نمودار به‌عنوان حالت ایده‌آل (شکل نمودار در صورتی که همه حالت‌ها صحیح پیش‌بینی شده باشد) رسم می‌شود و همچنین، نمودار دیگری که بیان‌کننده حالت مدل این پژوهش است را در همان گراف رسم می‌کند (Han and Kamber 2006). درنتیجه، با

1. Lift Chart

مقایسه این دو نمودار می‌توان دریافت مدل در چه قسمت‌هایی توانسته پیش‌بینی خوبی داشته باشد و یا اینکه در کدام قسمت‌ها ضعیف عمل کرده است. همچنین، پارامتری به‌نام امتیاز[1] در این نمودار وجود دارد که نشان‌دهنده درصدی از مواردی است که به‌صورت صحیح پیش‌بینی شده است. این پارامتر برای درخت تصمیم برابر ۶۰ و برای خوشه‌بندی ۵۹ به‌دست آمد.

در شکل ۵ نمودار ترفیع مدل ایجاد شده است. همان‌طور که در شکل مشاهده می‌شود سه نمودار رسم شده است که نمودار آبی رنگ نشان‌دهنده نمودار ایده‌آل، نمودار قرمز رنگ درخت تصمیم و نمودار سبز برای خوشه‌بندی است. در اینجا، از هر دو روش خوشه‌بندی و درخت تصمیم برای پیش‌بینی سطح خطرپذیری مشتریان مقادیر استفاده شده است. همان‌طور که در نمودار ملاحظه می‌شود این دو فن به‌طور تقریبی به یک اندازه توانسته‌اند پیش‌بینی صورت دهند (درخت تصمیم ۶۰٪ و خوشه‌بندی ۵۹٪) که با توجه به آن می‌توان دریافت فنّ درخت تصمیم برای پیش‌بینی سطح خطرپذیری مشتریان دارای دقت بیشتری است و برای این منظور مناسب‌تر است.



شکل ۵. نمودار ترفیع ایجادشده برای ارزیابی مدل درخت تصمیم و خوشه‌بندی

1. Score

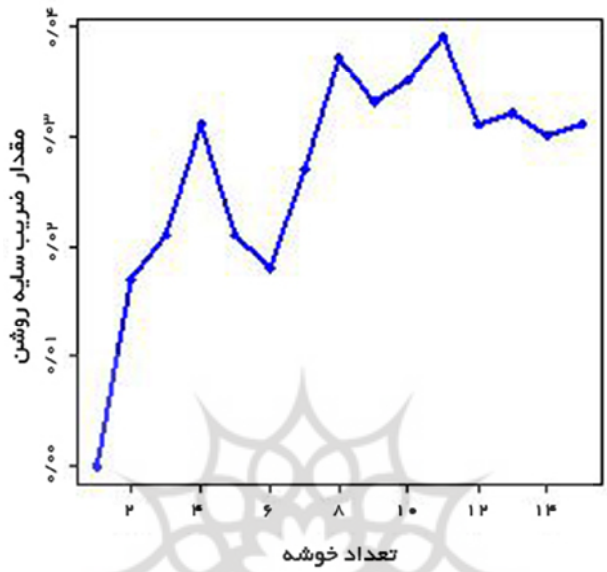
۱۸-۱-۲. ارزیابی به‌روش پهنه سایه روشن

روش پهنه سایه روشن برای ارزیابی خوشه‌ها مورد استفاده قرار می‌گیرد. یکی از ایراداتی که بر فنّ K-means وارد می‌شود این است که نمی‌تواند مقدار K بهینه را پیشنهاد دهد. برای تعیین مقدار k بهینه و یا ارزیابی اینکه آیا خوشه‌بندی خوبی صورت گرفته است می‌توان از روش سایه روشن استفاده نمود (Tan and Steinbach 2006).

برای انجام این کار، از نرم‌افزار R استفاده شد. روش کار بدین صورت است که ابتدا یک نمونه[1] از داده‌ها ایجاد می‌شود، سپس ماتریس فاصله‌ها[2] محاسبه و با استفاده از آن خوشه‌بندی سلسله مراتبی انجام می‌شود. خوشه‌بندی سلسله مراتبی دارای مزایای بسیاری از جمله استفاده از آن برای تعیین k است، ولی این فن به‌دلیل اینکه نیاز به ماتریس فاصله‌ها دارد دارای هزینه بالایی از نظر زمان و حافظه است. بنابراین برای این منظور، نمونه‌ای از داده‌ها به تعداد ۵۰۰۰ رکورد با استفاده از روش نمونه‌گیری طبقه‌ای[3] انتخاب شد. در این روش، ابتدا داده‌ها براساس رده‌شان مشخص می‌شوند، سپس از هر رده به‌میزان درصد معینی از رکوردها انتخاب می‌شود. پس از این کار، مقدار ضریب سایه روشن برای تعداد خوشه‌های متفاوت محاسبه می‌گردد. در این پژوهش، تعداد خوشه‌ها برابر ۱۱ درنظر گرفته شد. همان‌طور که مشاهده می‌شود با توجه به مقدار ضریب سایه روشن، بهترین تعداد خوشه که می‌تواند تفکیک‌کننده خوبی برای مشتریان باشد آن تعداد خوشه‌ای است که ضریب سایه روشن آن بیشینه باشد. در اینجا، این ضریب برای تعداد خوشه ۱۱ بیشینه شده و دوباره رو به کاهش است. بنابراین، تعداد خوشه ۱۱ بهترین تعداد خوشه برای تفکیک مشتریان است. درنتیجه، می‌توان نتیجه‌گیری کرد خوشه‌بندی صورت‌گرفته خوشه‌بندی مناسبی است.

در شکل ۶ نمودار توزیع مقادیر میانگین پهنه سایه روشن مشاهده می‌شود که این مقادیر به چه شکلی تغییر می‌کنند.

1. Sample
2. Distance Matrix
3. Stratified Sampling



شکل ٦. توزیع مقادیر میانگین پهنه سایه روشن

همان‌طور که در این شکل نیز مشاهده می‌شود، مقدار خوشه ۱۱ داری بیشترین مقدار ضریب سایه روشن است. یعنی این تعداد خوشه بهترین تعداد خوشه برای تفکیک این داده‌هاست. نتایج حاصل از دو فنّ درخت تصمیم و خوشه‌بندی مورد ارزیابی قرار گرفت و نتایج بسیار خوبی دریافت شد. همچنین، مشاهده شد که نتایج به‌دست‌آمده از این دو فن با یکدیگر هم‌راستا هستند، درنتیجه می‌توان نتایج تقریبا یکسانی از این دو فن به‌دست آورد که این مطلب نیز خود می‌تواند به‌عنوان صحّه‌ای برای نتایج به‌دست‌آمده باشد.

۱۸-۲. ارزیابی بیرونی

در ارزیابی بیرونی، مقایسه نتایج به‌دست‌آمده با ارزیابی‌های ذهنی کارشناسان خبره بیمه بدنه اتومبیل جستجو می‌شود. برای این منظور، از روش پرسشنامه استفاده شده است تا به‌وسیله آن ذهنیاتی را که کارشناسان خبره نسبت به پارامترهای تأثیرگذار بر خطرپذیری و خطر مشتریان دارند استخراج شود و با نتایج به‌دست‌آمده از عملیات داده‌کاوی و مدل ایجادشده مقایسه گردد. این پرسشنامه شامل مشخصه‌هایی است که برای مدل‌سازی از آنها استفاده شد. کارشناسان به هر کدام از این مشخصه‌ها امتیازی از ۱ تا ۵ می‌دهند که این امتیاز تعیین‌کننده الویت و اهمیت آن

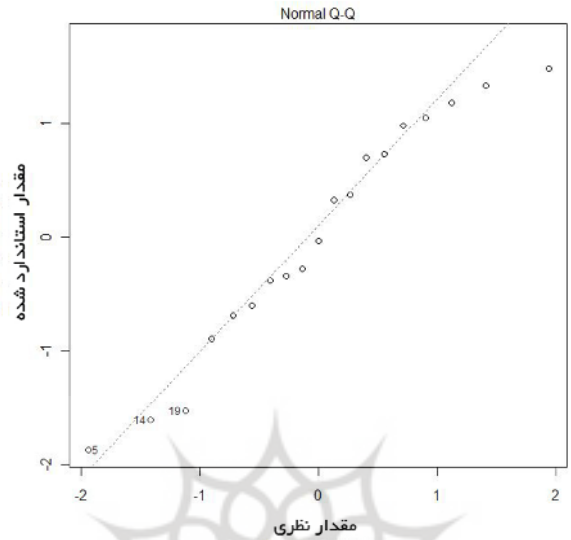
مشخصه در خطرپذیری مشتری و درنهایت، سطح خسارت اوست. پرسشنامه ایجادشده شامل ۲۰ سؤال است که برگرفته از مشخصه‌های اصلی تأثیرگذار در خطرپذیری مشتریان است و در ساخت مدل نیز از آنها استفاده شده است.

به‌دلیل اینکه شرکت‌های بیمه‌ای مختلفی وجود دارد برای پرکردن پرسشنامه‌ها از چهار شرکت بزرگ بیمه‌ای (آسیا، ایران، دانا، و البرز) استفاده شد. تعداد ۲۲ پرسشنامه توسط کارشناسان این شرکت‌ها پر گردید. سپس، با استفاده از نرم‌افزار صفحه گسترده[1] امتیازاتی را که کارشناسان به هر کدام از این مشخصه‌ها داده‌اند ذخیره و مجموع امتیاز برای هر مشخصه محاسبه گردید. درنهایت، براساس این امتیازات، مشخصه‌ها اولویت‌بندی شد، به‌گونه‌ای که مشخصه‌ای که دارای امتیاز بالاتری است، دارای اولویت بالاتری است.

۱۹. تحلیل نتایج نظرات کارشناسان

پس از اولویت‌بندی مشخصه‌ها براساس نظرات کارشناسان، نتایج حاصل از مدل‌ها با آن مقایسه می‌شود تا مشخص گردد این نتایج تا چه حدی هم‌راستا هستند. برای این منظور، نمودار اولویت‌های به‌دست‌آمده از مدل، برحسب اولویت‌های به‌دست‌آمده از نظرسنجی خبرگان رسم می‌شود. هر چه نمودار ایجادشده به خط نیمساز ناحیه اول نزدیکتر باشد، بدین معنی است که نتایج به‌دست‌آمده از دو روش به یکدیگر نزدیکتر هستند (شکل ۷). در شکل ۷، نقاط نشان‌دهنده نسبت نتایج به‌دست‌آمده از پرسشنامه به نتایج حاصل از مدل است و نیز خط رسم‌شده خط رگرسیونی است که براساس این نقاط به‌دست آمده است. خط رگرسیون خطی است که مجموع فواصل آن از نقاط ایجادشده کمتر باشد. به‌عبارت دیگر، هر چه نقاط به این خط نزدیکتر باشند به‌معنای آن است که این نتایج به‌دست‌آمده بیشتر به یکدیگر نزدیک هستند. مقدار ضریب همبستگی محاسبه‌شده برابر ۸۲٪ است که به‌معنای همبستگی بالا میان نتایج به‌دست‌آمده از مدل و پرسشنامه است.

1. Spread Sheet



شکل ۷. نمودار همبستگی و خط رگرسیون نتایج حاصل از مدل و نتایج حاصل از پرسشنامه

همان‌طور که در شکل ۷ مشاهده می‌شود نمودار به‌دست‌آمده به‌صورت خط مستقیم نیست و این به‌معنای آن است که نتایج به‌صورت دقیق یکسان نیستند، ولی این مطلب قابل توجیه است. اولویت‌هایی که در ذهن کارشناسان است، به‌صورت کامل دقیق نیست، زیرا براساس ذهنیات و تجربیات آنان است و پایه علمی قوی ندارد. به‌همین دلیل، اختلاف‌هایی بین نتایج وجود دارد ولی نتایج به‌دست‌آمده از دو مدل به‌طور تقریبی هم‌راستا هستند. ضریب همبستگی برای این دو برابر ۰.۸۲ است که این مطلب می‌تواند به‌عنوان تأییدی بر نتایج به‌دست‌آمده باشد؛ این دو نتیجه یکدیگر را تأیید می‌کنند.

۲۰. دستاوردهای پژوهش و پیشنهادات آتی

از جمله دستاوردهای مهم این پژوهش، نتایج حاصل از عملیات داده‌کاوی بر روی داده‌های بیمه بدنه اتومبیل است. براساس نتایج این مرحله مشاهده گردید که علاوه بر مشخصه‌های ظاهری اتومبیل مانند تعداد سیلندر، ظرفیت، سال ساخت، و نوع کاربری، مشخصه‌های رفتاری مشتری یا همان ویژگی‌های جمعیت‌شناختی آنها نیز در پیش‌بینی سطح خسارت مشتریان بیمه تأثیر عمده‌ای دارند. این نکته نشان‌دهنده آن است که پارامترهای مورد استفاده برای نرخ‌گذاری بیمه نیازمند بازنگری و توجه بیشتر به سوابق، رفتار، و خسارت مشتری است. درواقع، در تعیین تعرفه بیمه که

هم اکنون فقط براساس ویژگی‌های اتومبیل است، نقصان بسیار زیادی وجود دارد. البته این مسأله بدین معنی نیست که ویژگی‌های اتومبیل در تعیین تعرفه بیمه تأثیرگذار نیست، بلکه بدین معنی است که این ویژگی‌ها برای این امر ناقص است و به‌تنهایی نمی‌تواند نتیجه مناسب را داشته باشد، زیرا در یک تصادف، راننده اتومبیل و ویژگی‌های او نیز نقش بسیار مهمی دارد.

از جمله دستاوردهای دیگر این پژوهش، شناسایی مشخصه‌های مفید در صنعت بیمه بدنه اتومبیل و نیز تعیین مشخصه‌های تأثیرگذار بر میزان خطرپذیری رانندگان است که با توجه به آن می‌توان میزان خطر رانندگان را شناسایی نمود و بر این اساس، سطح خسارت آنها را پیش‌بینی کرد.

در ارتباط با مشخصه‌های شناسایی مشتریان بیمه بدنه اتومبیل، بهتر است مشخصه‌های جمعیت‌شناختی بیشتری نیز در نظر گرفته شود. زیرا این مشخصه‌ها می‌توانند توصیف‌کننده میزان خطرپذیری مشتریان باشند و همچنین جمع‌آوری این اطلاعات از مشتریان ساده است و منع قانونی ندارد. از جمله این مشخصه‌های جمعیت‌شناختی می‌توان به موارد زیر اشاره کرد:

۱. منطقه جغرافیایی مشتری؛
۲. منطقه محل کار مشتری؛
۳. داده‌های مربوط به جرائم رانندگی؛
۴. میزان فاصله محل زندگی فرد تا محل کارش؛
۵. میزان مسافت طی‌شده در سال؛
۶. تعداد افراد استفاده‌کننده از وسیله نقلیه؛
۷. تعداد سانحه رانندگی مشتری در سال؛
۸. میزان جرائم رانندگی مشتری در سال؛
۹. نوع جرائم رانندگی مشتری؛
۱۰. میزان نقص عضو راننده؛
۱۱. داشتن یا نداشتن پارکینگ اختصاصی؛
۱۲. منطقه محل زندگی و کار مشتری از نظر تراکم جمعیت؛ و
۱۳. نوع نگرش مشتری که اتومبیل را در اختیار افراد دیگر از جمله افراد جوان‌تر قرار می‌دهد یا نه.

درنتیجه می‌توان مشتریان را بهتر شناسایی کرد و برای آنها تعرفه بیمه‌نامه‌ای با توجه به ویژگی‌های آنها و سطح خطرپذیری آنها تعیین نمود.

۲۱. منابع

حسین‌زاده، لیلا، و شعبان الهی. ۱۳۸۶. دسته‌بندی مشتریان هدف در صنعت بیمه با استفاده از داده‌کاوی. پایان‌نامه کارشناسی ارشد رشته مدیریت فناوری اطلاعات، دانشگاه تربیت مدرس.

رستمی، حمیدرضا. ۱۳۷۸. نقش عوامل ایجادکننده خطر در قیمت‌گذاری بیمه بدنه اتومبیل. *فصلنامه صنعت بیمه* ۴۸ (۴) ۹۰-۸۵.

Borgelt, C. 2008. Accelerating fuzzy clustering. *Information Sciences* 179 (23): 3985-3997.

Castro, E. 2000. Automatic clustering via boundary extraction for mining massive point-data sets. In *Proceedings of the Fifth International Conference on Geo-computation* 33: 133-146.

Chan, C., and B. Lewis. 2002. A basic primer on data mining. *Information Systems Management*, 56-60.

Chen, S. 2006. A KanoCKM model for customer knowledge discovery. *Total Quality Management* 17 (5): 589-608.

Dalkilic, T. 2009. Neural networks approach for determining total claim amounts in insurance. *Insurance: Mathematics and Economics* 45 (2): 236-241.

Das, S. 2009. Utilization of self-organizing map and fuzzy clustering for site characterization using piezocone data. *Computers and Geotechnics* 36: 241-248.

Guo, L. 2002. Applying data mining techniques in property. *Casualty Insurance* 13 (2): 230-472.

Gupta, G. K. 2006. *Data mining with case studies*. New Delhi: Prentice Hall India.

Han, J., and M. Kamber. 2006. *Data mining: Concepts and techniques.* San Francisco: Morgan Kaufman.

Kuo, R. J. 2007. Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan. *Expert Systems with Applications* 33 (3): 794-808.

Lee, S. 2006. Decision tree approaches for zeroinflated count data. *Journal of Applied Statistics Journal of Applied Statistics* 33 (8): 853-865.

Lin, C. 2009. Using neural networks as a support tool in the decision making for insurance industry. *Expert Systems With Applications* 36 (3): 6914-6917.

Lu, F. 2004. A fast genetic k-means algorithm. In *Proceedings of the 19th ACM Symposium on Applied Computing* 5: 101-110.

Microsoft. 2010. *Microsoft SqlServer 2008*. http://msdn.microsoft.com/en-us/sqlserver/bb895906.aspx (accessed 2 May 2011).

Moon, C. 2001. Data mining approach to policy analysis in a health insurance domain. *International journal of medical informatics* 62 (2-3): 101-103.

Morley, B. 2006. How the detection of insurance fraud. *Psychology, Crime & Law* 12: 163-180.

Robert, A. 2001. SQL Server Analysis Services. *Designing SQL server 2000 databases* 41 (10): 453-498.

Saha, S. 2009. A new point symmetry based fuzzy genetic clustering technique for automatic evolution of clusters. *Information Sciences* 20 (11): 1441-1457

Smith, K. 2000. An analysis of customer retention and insurance claim patterns using data mining: a case study. *Journal of the Operational Research Society*, 532-541 51 (5): 532-541.

Sumathi, S. 2006. Data mining for insurance. Berlin: Springer.

Tan, P. N., and M. Steinbach. 2006. *Intruduction to data mining.* United State Of America: Addison Wesly.

Using Data Mining Techniques to Predict the Detriment Level of Car Insurance Customers

Seyyed Mahmood Izadparast*
Master of Management of Information Technology

Ahmad Farahi[1]
PhD of Software Engineering, Assistant Professor in Payame Noor University

Faramarz Fath Nejad[2]
PhD of Applied Mathematics

Babak Teimourpour[3]
PhD of Information Technology

Iranian Journal of Information Processing & Management
Iranian Research Institute For Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in LISA, SCOPUS & ISC
Vol.27 | No.3 | pp: 699-722
spring 2012

Abstract: Nowadays customers' role is changed from just accepting the producers, to leading investors, producers, and even researchers and inventors. Therefore, it is necessary for organizations to identify their customers well and to make plans for them. Some statistical and machine-based learning methods are used so far. However these methods alone are not without limitations. Using various methods of data mining, this research was to eliminate those restrictions as far as possible, so that a framework for identification of car insurance customers could be provided. In fact, the purpose was to categorize the most similar customers and to estimate the amount of risk in each category, according to their characteristics. Now, using this scale (i.e. amount of risk in each category) and considering the type of customer's policy, the level of recompense could be estimated. This criterion can be helpful to identify customers and for making insurance tariff policies. For this purpose, in insurance industry the two data mining methods were been used to estimate customers' detriment: the decision tree and clustering. Nevertheless, the decision tree method appears to give better results, although at the same, the clustering method generates a good categorization.

Keywords: Data mining, insurance, categorize, decision tree, clustering, detriment

* **Corresponding Author:** s.m.izadparast@gmail.com
1. a.farahi@pnu.ac.ir
2. fatnejad@yahoo.com
3. b.teimourpour@modares.ac.ir