کاربردهای داده‌کاوی در صنعت بیمه

نویسندگان: محسن قره‌خانی- مریم ابوالقاسمی

- دانشجوی دکترای مهندسی صنایع، دانشگاه علم و صنعت
- دانشجوی کارشناسی ارشد مدیریت اجرایی، دانشگاه علم و صنعت

چکیده

باتوجه به پیشرفت سریع فناوری اطلاعات، حجم اطلاعات ذخیره‌شده در پایگاه‌های داده شرکت‌های بیمه به سرعت در حال افزایش بوده و این پایگاه‌های داده بزرگ، حاوی حجم زیادی از داده‌ها و فرصت‌های قابل استفاده و بالقوه از اطلاعات تجاری با ارزش هستند. از طرفی، یافتن اطلاعات ارزشمند پنهان در این پایگاه داده و نیز شناسایی مدل‌های مناسب کاری دشوار است. در این مقاله ضمن بررسی اجمالی داده‌کاوی، به نقش آن در کشف دانش موجود در پایگاه‌های داده و بهبود امور مرتبط با صنعت بیمه پرداخته شده است.

واژگان کلیدی: داده‌کاوی، صنعت بیمه، کشف دانش، پایگاه داده

مقدمه

صنعت بیمه در کشور به علت رویارویی با مسائلی همچون خصوصی‌سازی، تغییر و تحولات در قوانین و آزادسازی تعرفه‌ها ازیک‌سو و پیشرفت فناوری و کاهش هزینه محاسبات ازسوی‌دیگر، همواره در حال تغییرات بنیادین است تا خود را با شرایط و نیازهای جدید محیط وفق دهد. در واقع صنعت بیمه از صنعتی تولیدمحور به خدمتی مشتری‌محور تبدیل شده است. علاوه‌براین جهانی‌شدن باعث سرعت‌دادن به تغییرات بازار بیمه می‌شود. ازاین‌رو شرکت‌های بیمه در رقابت سخت جهانی هستند و بیشتر این شرکت‌ها به طور جدی به دنبال یافتن بازارهای جدید، جذب بیمه‌گذاران کم‌ریسک و ارائه خدمات هستند. فناوری مخصوصاً در حیطه فناوری‌های رایانه‌ای و ارتباطات باعث تغییرات لحظه‌ای در صنعت بیمه شده است. با وجود فناوری جدید رایانه‌ای و به‌کارگیری شبکه جهانی اینترنت، سرعت فرآیندهای تجاری و ارائه خدمات روز به روز بیشتر شده و تأثیر بسزایی در ارائه خدمات بیمه‌ای گذاشته است.

پیشرفت‌های رایانه‌ای اجازه تقسیم‌بندی و تمرکز بهتر به روی رشته‌های مختلف بیمه‌ای را فراهم آورده و واحدهای تجاری گوناگون را بیش‌ازپیش قادر به توسعه مدل‌های تجاری، طراحی تولیدات جدید و تقسیم‌بندی بیشتر مشتریان گوناگون، ساخته است. ازاین‌رو بیشتر شرکت‌های بیمه با به‌کارگیری پیشرفت‌های

به‌دست‌آمده با تأکیدی روز افزون از حالت معمولی و سنتی ارائه خدمات بیمه‌ای به خدمات جدید با تقسیم‌بندی هدفمند مشتریان روی آورده‌اند. اطلاعات مربوط‌به بیمه‌گذاران، قراردادها و داده‌های تجاری در بانک‌های اطلاعاتی ذخیره و بایگانی شده و در سیستم داده‌کاوی[1] مورد بررسی قرار می‌گیرند.

داده‌کاوی، فرآیندی تحلیلی است که برای کاوش داده‌ها (معمولاً حجم عظیمی از داده‌ها در زمینه‌های کسب‌وکار و بازار) صورت می‌گیرد و اعتبار یافته‌ها با به‌کارگیری الگوهایی احراز می‌شود.

هدف اصلی داده‌کاوی پیش‌بینی است و به‌صورت دقیق‌تر می‌توان گفت: «داده‌کاوی شناسایی الگوهای صحیح، بدیع، سودمند و قابل درک از داده‌های موجود در یک پایگاه داده است که با استفاده از پردازش‌های معمول قابل دستیابی نیستند».

ازاین‌رو به‌نظرمی‌رسد که داده‌کاوی با یافتن متغیرهای مهم، اثرات متقابل و ارتباط غیرخطی آنها می‌تواند به اخذ تصمیمات مهم تجاری شرکت‌های بیمه و تبدیل یافته‌ها به نتایج کاربردی و عملی در تجارت از قبیل توسعه محصولات، بازاریابی، تحلیل توزیع خسارت، مدیریت تعهدات و تحلیل توانگری مالی شرکت بیمه کمک شایانی نماید. در این مقاله سعی شده است که ضمن معرفی اجمالی علم داده‌کاوی و روش‌های آن به‌عنوان یکی از دانش‌های روز دنیا، به بررسی نقش آن در کشف دانش پنهان در پایگاه‌های داده موجود در صنعت بیمه و بهبود این حوزه پرداخته شود.

هدف اصلی داده‌کاوی پیش‌بینی است

۱. داده‌کاوی

در این بخش، مرور کلی از فرآیند داده‌کاوی، عملیات داده‌کاوی، روش‌ها و الگوریتم‌های داده‌کاوی و کاربردهای بالقوه آن در صنعت بیمه ارائه خواهد شد.

۱-۱. فرآیند داده‌کاوی

داده‌کاوی ترکیبی از تکنیک‌های یادگیری ماشین[2]، تشخیص الگو، آمار، تئوری پایگاه داده و خلاصه‌کردن و ارتباط بین مفاهیم و الگوهای جالب به‌صورت خودکار از پایگاه‌های داده شرکت‌های بزرگ است. هدف اصلی داده‌کاوی کمک به فرآیند تصمیم‌گیری از طریق استخراج دانش از داده‌هاست (Alpaydin, 2004).

دو کاربرد اصلی داده‌کاوی عبارت‌اند از:

- پیش‌بینی، که شامل پیداکردن روابط و الگوهای ناشناخته از مقادیر شناخته‌شده است؛
- توصیفات، که شرح یک پایگاه داده بزرگ را ارائه می‌کند (Hand et al, 2001).

فرآیند داده‌کاوی به‌طورکلی شامل این مراحل است (Michalski et al, 1998):

- پاک‌سازی داده‌ها: در این مرحله داده‌های نامعتبر از مجموعه داده‌های آموزشی خارج می‌شوند. داده‌های دارای نویز[3]، اطلاعات ناقص و مواردی از این دست، نمونه‌هایی از داده‌هایی هستند که باید پاکسازی شوند؛
- یکپارچه‌سازی داده‌ها: در این مرحله، منابع چندگانه داده‌ای با هم ترکیب می‌شوند؛
- انتخاب داده‌ها: داده‌های مرتبط به فرآیند داده‌کاوی از سایر داده‌ها جدا می‌شوند. این مبحث را می‌توان بخشی از فرآیند کاهش اطلاعات[4] نیز دانست؛
- تبدیل داده‌ها: داده‌ها به ساختاری قابل استفاده برای داده‌کاوی در می‌آیند. از اعمالی که در این مرحله صورت می‌گیرد می‌توان به خلاصه‌سازی یا محاسبه مقادیر تجمعی اشاره کرد؛
- داده‌کاوی: بخش اصلی فرآیند که در آن با استفاده از روش‌ها و تکنیک‌های خاص، استخراج الگوهای دانش صورت می‌گیرد؛
- ارزیابی الگوها: تشخیص الگوهای صحیح مورد نظر،

1. Datamining

2. Machine Learning
3. Noise
4. Data Reduction

از سایر الگوها در این مرحله انجام می‌شود. صحت الگوها براساس یک سری معیارهای جذابیت سنجیده می‌شود؛

– بازنمایی دانش: در این بخش به‌منظور ارائه دانش استخراج‌شده به کاربر، از یک سری ابزارهای بصری‌سازی به‌منظور افزایش درک استفاده می‌گردد.

۲-۱. عملیات داده‌کاوی

فرض کنید شما مجموعه داده‌ها را برای داده‌کاوی آماده کرده‌اید، بعد از آن شما نیاز به تعریف دامنه خود از مطالعات صورت گرفته و انتخاب موضوع دارید. این امر به‌عنوان انتخاب عملیات داده‌کاوی نامیده می‌شود (Haiy, 2005).

پنج نوع از عملیات داده‌کاوی وجود دارد : طبقه‌بندی[1]، رگرسیون[2]، تحلیل لینک[3]، بخش‌بندی[4] و تشخیص انحراف[5]. طبقه‌بندی و رگرسیون برای پیش‌بینی مفید هستند، درحالی‌که بخش‌بندی، و تشخیص انحراف برای توصیف الگوهای موجود در داده‌ها به‌کارمی‌روند (Larose, 2004). جدول ۱ عملیات داده‌کاوی و تکنیک‌های مربوطه را به‌طور مختصر بیان می‌کند

۱-۲-۱. طبقه‌بندی

طبقه‌بندی در واقع ارزشیابی ویژگی‌های مجموعه‌ای از داده‌ها و سپس اختصاص‌دادن آنها به مجموعه‌ای از گروه‌های از پیش تعریف شده بوده و می‌توان گفت که متداول‌ترین قابلیت داده‌کاوی است. داده‌کاوی را می‌توان با استفاده از داده‌های تاریخی برای تولید یک مدل یا نمایی از یک گروه براساس ویژگی‌های داده‌ها به‌کاربرد. سپس می‌توان از این مدل تعریف‌شده برای طبقه‌بندی مجموعه داده‌های جدید استفاده کرد. همچنین می‌توان با تعیین نمایی که با آن سازگار است برای پیش‌بینی‌های آتی از آن بهره گرفت (غضنفری و همکاران، ۱۳۸۷).

۲-۲-۱. رگرسیون

این عملیات مدلی را می‌سازد که آیتم‌های داده را به یک متغیر پیش‌بینی‌شده با ارزش واقعی نگاشت می‌کند. مدل به طور سنتی با استفاده از روش‌های آماری مانند لجستیک، خطی و رگرسیون توسعه می‌یابد. هم طبقه‌بندی و هم رگرسیون برای پیش‌بینی استفاده می‌شود.

1. Classification
2. Regression
3. Link Analysis
4. Segmentation
5. Deviation Detection

جدول ۱. تکنیک‌های داده‌کاوی برای انجام عملیات داده‌کاوی

تکنیک داده‌کاوی / عملیات داده‌کاوی	استنتاج	الگوریتم ژنتیک	خوشه‌بندی	رگرسیون لجستیک	کشف انجمنی (وابستگی)	کشف توالی	مصور ساختن
دسته‌بندی	✓	✓					
رگرسیون	✓	✓		✓			
تقسیم‌بندی		✓	✓				
انحراف				✓			✓

(Alpaydin, 2004)

۳-۲-۱. شبکه‌های عصبی

تکنیک شبکه‌های عصبی[1] به‌طور گسترده در داده‌کاوی استفاده می‌شود. شبکه‌های عصبی را می‌توان با اغماض زیاد، مدل‌های الکترونیکی از ساختار عصبی مغز انسان نامید.

مکانیسم فراگیری و آموزش مغز اساساً بر تجربه استوار است. مدل‌های الکترونیکی شبکه‌های عصبی طبیعی نیز براساس همین الگو بنا شده‌اند و روش برخورد چنین مدل‌هایی با مسائل، با روش‌های محاسباتی که به‌طورمعمول توسط سیستم‌های کامپیوتری در پیش گرفته شده‌اند، تفاوت دارد. روش شبکه‌های عصبی می‌تواند برای توسعه طبقه‌بندی، رگرسیون، تحلیل لینک، بخش‌بندی و مدل‌سازی استفاده شود. دو دسته کلی از الگوریتم‌های عصبی خالص وجود دارد:

- تحت نظارت: که به ارزش‌های خروجی به‌منظور توسعه طبقه‌بندی مدل نیاز دارد.
- بدون سرپرست که به ارزش‌های خروجی نیاز نداشته و یکی از بهترین روش‌های شناخته‌شده برای آن کوهنن[2] است.

برای سازمان‌های با حجم زیادی از اطلاعات آماری، شبکه عصبی بسیار ایده‌ال هستند زیرا آنها می‌توانند شناسایی و تجزیه‌وتحلیل تغییر در الگوها، شرایط یا تاکتیک‌ها را به مراتب بیشتر از رگرسیون عددی انجام دهند (شهرابی، ۱۳۸۸).

۴-۲-۱. بخش‌بندی

هدف قطعه‌بندی، شناسایی خوشه‌ها با استفاده از سوابق و رفتار مشابه یا ویژگی‌های پنهان در داده‌هاست. خوشه ممکن است منحصربه‌فرد و جامع یا شامل مقوله‌های سلسله مراتبی و متداخل باشد (Hair, 2005).

۲. به‌کارگیری داده‌کاوی در صنعت بیمه

متدولوژی داده کاوی اغلب می‌تواند مدل‌های اکچوئری[3] موجود را از طریق پیداکردن متغیرهای مهم، تعیین روابط بین آنها و کشف روابط غیرخطی آنها ارتقا دهد. داده‌کاوی می‌تواند در تصمیم‌گیری‌های حیاتی کسب‌وکار به شرکت‌های بیمه کمک کند و دانش تازه به‌دست‌آمده را به نتایج قابل اقدام در کسب‌وکار شامل توسعه محصول، بازاریابی، تحلیل توزیع خسارت، مدیریت دارایی- بدهی و تحلیل توانایی بازپرداخت دیون تبدیل کند. مثالی از نحوه کارکرد داده‌کاوی در بیمه درمان را می‌توان در تحقیق بوروک ۱۹۹۷ مشاهده کرد. به طور خاص داده‌کاوی موارد زیر را می‌تواند انجام دهد.

1. Neural Networks
2. Kohonen
3. Actuarial

۲-۱. شناسایی عوامل ریسک که سود، خسارت و زیان را پیش‌بینی می‌کند

از مهم‌ترین سؤالات در اکچوئری این است که کدام عوامل و متغیرهای ریسک در پیش‌بینی توزیع خسارت و اندازه آن مهم هستند؟

هرچند که بسیاری از عوامل ریسک که بر نرخ اثر می‌گذارند بدیهی‌اند، ممکن است بین متغیرها روابط دقیق و غیرشهودی برقرار باشد که کشف آنها بدون استفاده از تکنیک‌های پیچیده‌تر اگر محال نباشد، کار بسیار سختی است. مدل‌های جدید داده‌کاوی از قبیل درخت تصمیم‌گیری و شبکه‌های عصبی، ریسک را با دقت بیشتری نسبت‌به مدل‌های اکچوئری موجود پیش‌بینی می‌کنند. بنابراین یک شرکت بیمه می‌تواند نرخ‌های دقیق‌تری بدهد که به نوبه خود منجربه قیمت‌گذاری دقیق‌تر و نیز موقعیت رقابتی بهتر می‌شود.

۲-۲. تحلیل در سطح مشتری

حفظ موفق مشتری نیازمند تحلیل داده‌ها در مناسب‌ترین سطح ممکن یعنی سطح فردی مشتری است (به جای درنظرگرفتن مجموعه‌ای از مشتریان به‌صورت کلی). با استفاده از تکنیک داده‌کاوی کشف مرتبط[1]، شرکت‌های بیمه می‌توانند دقیق‌تر انتخاب کنند که چه خدمات و قراردادهایی را به مشتری ارائه کنند. با این تکنیک شرکت‌های بیمه می‌توانند:

– پایگاه داده خود را برای ایجاد پروفایل مشتری بخش‌بندی[2] کنند.

– روی بخش خاصی از مشتریان و برای یک محصول، تحلیل خسارت و نرخ انجام دهند. برای مثال شرکت‌ها می‌توانند تحلیل عمیقی از محصولات جدید بالقوه را روی بخش خاصی از مشتریان انجام دهند.

– برای چند محصول با استفاده از پردازش گروهی[3] و متغیرهای چند هدفی[4] تحلیل بخش انجام دهند، برای مثال اینکه قراردادهای ترکیبی (اتومبیل، منزل مسکونی و زندگی) در مورد بخش خاصی از مشتریان چقدر سودده بوده است؟

– انجام تحلیل‌های بازار متوالی (در طول زمان) روی بخش‌های مختلف مشتریان. برای مثال چند درصد از بیمه‌گذاران اتومبیل در طول ۵ سال، بیمه عمر هم خریده‌اند؟

بخش‌بندی پایگاه داده و تکنیک‌های پیشرفته‌تر مدل‌سازی، تحلیلگران را قادر می‌سازد که تعیین کنند برای برنامه‌های حفظ مشتری روی کدام بخش هدف‌گذاری کنند. بیمه‌گذاران فعلی را– که احتمال دارد شرکت بیمه خود را عوض کنند– می‌توان با استفاده از مدل‌سازی پیشگویانه تشخیص داد.

مدل رگرسیون لجستیک رویکردی سنتی برای پیشگویی آن دسته از بیمه‌گذارانی است که احتمال بیشتری دارد که بیمه خود را عوض کنند[5]. شناسایی گروه هدف برای برنامه‌های حفظ مشتری را می‌توان با مدل‌سازی رفتار بیمه‌گذاران ارتقا داد.

۲-۳. ایجاد رشته‌های محصول جدید

شرکت‌های بیمه می‌توانند قابلیت سوددهی خود را با شناسایی سودمندترین بخش از مشتریان و اولویت‌دهی به برنامه‌های بازاریابی برهمین‌اساس افزایش دهند. مشکلات مربوط‌به سوددهی شرکت زمانی رخ می‌دهد که شرکت قادر نباشد قرارداد مناسب[6] را با نرخ مناسب و به مشتری مناسب در زمان مناسب ارائه کند؛

به‌عنوان‌مثال، برای یک بیمه‌گر استفاده از توزیع لگ‌نرمال به‌منظور نرخ‌گذاری، زمانی که توزیع پارتو توزیع صحیح باشد، اشتباه بزرگی است و هزینه‌های زیادی درپی‌دارد. این مسئله لزوم وجود ابزار مناسبی برای شناسایی و تخمین توزیع زیان را روشن می‌کند. امروزه با عملیات داده‌کاوی از قبیل بخش‌بندی[7] یا تحلیل وابستگی[8]

1. Associated Discovery
2. Segment
3. Group Processing
4. Multi Target Variables

5. Switching
6. Right
7. Segmentation
8. Association Analysis

شـرکت‌های بیمه می‌توانند از همه اطلاعات موجود خود اسـتفاده کنند تا محصولات و برنامه‌های بازاریابی بهتری طراحی کنند.

۲-۴. بیمه اتکایی

از داده‌کاوی می‌تـوان بـرای سـاماندهی مؤثرتـر بیمه اتکایی نسبت‌به روش‌های سـنتی استفاده کرد. تکنولوژی داده کاوی معمـولاً بـرای وضـوح بخش‌بندی اسـتفاده می‌شـود. در مورد بیمه اتکایی، گروهی از خسـارت‌های پرداختی برای مدل‌سـازی خسـارت‌های انتظـاری گروه دیگری از بیمه‌نامه اسـتفاده می‌شـود. بـا بخش‌بندی‌های دقیق‌تر، تحلیلگران می‌توانند اطمینان بیشـتری به خروجی مدل داشـته باشند. انتخاب قراردادها برای بیمه اتکایی باید بر مبنای مدل ریسـک تجربه شده باشد و تنها بر پایه تعمیم نباشد. چرا که آن مجموعه‌ای از کسب‌وکاری با توزیع دم سنگین[1] است.

۲-۵. تخمین ذخایر برای خسارت‌های معوق

تسـویه حساب خسـارت‌ها اغلب با تأخیر همراه است. بنابراین تا زمانی که میزان واقعی ارزش خسـارت مشخص نشده از تخمین شدت خسارت استفاده می‌شود. این تخمین به موارد زیر بستگی دارد:

- شدت خسارت؛
- مدت زمان تا تسویه حساب؛
- اثرات متغیرهای مالی نظیر نرخ تورم و بهره؛
- اثرات تغییر در آداب و رسـوم اجتماعی؛ مثلاً صنعت تنباکو به‌شدت از تغییر رویکرد نسبت‌به استعمال دخانیات متأثر است.

برای بهبود تخمین خسـارت می‌تـوان از تکنیک‌های داده‌کاوی نظیر تحلیل لینک[2] و کشـف انحراف[3] استفاده کرد.

تخمین میزان خسـارت با اسـتفاده از مدل پیشگویانه بر این فرض اسـتوار است که آینده شـبیه به گذشته خواهد بود. اگر مدل در طول زمان به‌روزرسـانی نشود و داده‌های بیشـتری در دسترس نباشد این فرض تبدیل به این می‌شود کـه آینده مثل گذشـته دور خواهد بود. مـدل داده‌کاوی پیشگویانه را می‌توان به‌روز کرد و فرض تبدیل می‌شود به اینکه آینده مثل گذشـته نزدیک رفتار می‌کند. تکنولوژی داده‌کاوی، تحلیل‌گران را قادر می‌سازد تا مدل‌های جدید و قدیمی را با هم مقایسه کنند و آنها را براساس عملکردشان ارزیابی کنند. اگر مدلی که به تازگی به‌روزرسانی‌شـده از مـدل قدیمی بهتر کار کند، زمان این فرارسـیده که آن را با مدل جدیـد جایگزین کنیم. باتوجه بـه تکنولوژی‌های جدید، تحلیل‌گران امروزه قادرند مدل‌های پیشـگویانه را کنترل کرده و در صورت نیاز به‌روز کنند.

تفـاوت اصلی بیـن تکنیک‌های اکچوئـری موجود و داده‌کاوی در آن است که داده‌کاوی بیشتر به کاربرد تمایل دارد تـا توصیف ماهیت پدیده. برای مثال، آشـکارکردن ماهیت توزیع خسـارت فردی یا رابطه خاصی بین سـن و نوع اتومبیل راننده جزو اهداف اصلی داده‌کاوی نیستند.

در عـوض تمرکـز بر ایجاد راه‌حلی اسـت کـه بتواند پیش‌بینی‌های حق‌بیمه آینـده را بهبود دهد. داده‌کاوی در تعیین رابطه بین حق‌بیمه و فاکتورهای چند بعدی ریسـک نظیر سـن و اتومبیل راننده بسـیار مؤثر اسـت. دو مثال از به‌کارگیـری تکنیک‌هـای داده‌کاوی در صنعـت بیمه و اکچوئری در دو بخش بعدی ارائه می‌شود.

۳. خوشه‌بندی و داده‌کاوی توصیفی

خوشـه‌بندی[4] از مفیدترین کارکردهـای داده‌کاوی برای کشـف گروه‌هـا و تعیین توزیع‌هـای مورد علاقـه و الگوها در داده‌هاسـت. مسئله خوشـه‌بندی در مورد جداسازی یک مجموعه داده به گروه‌ها (خوشه) به نحوی است که داده‌های موجود در یک خوشـه، نسبت به نقاط موجود در خوشه‌های دیگر شباهت بیشـتری به یکدیگر داشته باشـند . برای مثال بخش‌بنـدی بیمه‌گـذاران فعلی بـه گروه‌هایی مشـخص و مرتبط کـردن یـک پروفایل با هـر گـروه می‌تواند در استراتژی‌های نرخ‌گذاری آینده مفید باشد.

در روش‌های خوشـه‌بندی، تحلیل خوشـه گسسـته را

1. Heavy Tailed
2. Link Analysis
3. Deviation Detection
4. Clustering

از داده‌کاوی می‌توان برای ساماندهی مؤثرتر بیمه اتکایی نسبت‌به روش‌های سنتی استفاده کرد

بـر پایه فاصله اقلیدسـی انجـام می‌دهند. ایـن فاصله‌های اقلیدسـی از یـک یا تعدادی متغیر هسـته محاسـبه شـده اسـت کـه توسـط الگوریتـم تولیـد و به‌روز می‌شـوند (Tan, 2004).

می‌تـوان معیـار خوشـه‌بندی را تعیین کرد کـه برای اندازه‌گیری فاصله بین مشـاهدات و هسته‌ها به‌کارمی‌رود. مشـاهدات به خوشه‌هایی تقسیم می‌شوند به‌گونه‌ای که هر مشاهده حداکثر به یک خوشه تعلق داشته باشد.

همچنیـن از مطالعـات خوشـه‌بندی بـه یادگیـری یـا بخش‌بنـدی کنترل‌نشـده تعبیـر می‌شـود. یادگیـری کنترل‌نشـده، فرآیند خوشه‌بندی با هدف نامشخص است. یعنی کلاس هر مورد نامشـخص است. هدف تقسیم کردن موارد به کلاس‌های گسسـته اسـت که نسـبت به ورودی همگن باشند.

مطالعات خوشه‌بندی هیچ متغیر وابسته‌ای ندارند و مانند مطالعات طبقه‌بندی یک ویژگی خاص را پروفایل نمی‌کنیم.

یک پایـگاه داده را با اسـتفاده از مـوارد زیر می‌توان تقسیم‌بندی کرد:

- روش‌های سنتی تشخیص الگو؛
- شـبکه‌های عصبی کنترل‌نشـده نظیر ART و نقشـه کوهنن؛
- تکنیک‌هـای خوشـه‌بندی مفهومـی از قبیـل COBWEB،UNIMEN؛
- رویکرد بیزی مانند AutoClass .

الگوریتم‌های خوشـه‌بندی مفهومی، کلیه ویژگی‌های توصیف‌کننده هر رکورد را درنظرگرفته و زیرمجموعه‌ای از ویژگی‌هایی که هر خوشه ایجادشده را توصیف می‌کنند، برای ایجاد مفهوم تعیین می‌کنند. مفاهیم در الگوریتم‌های خوشـه‌بندی به‌صورت اتصال ویژگی‌ها با مقدار یا مقادیر درنظرگرفته‌می‌شـوند. الگوریتم‌هـای خوشـه‌بندی بیزی به صـورت خودکار خوشـه‌بندی را کشـف می‌کنند که باتوجه به داده‌ها محتمل‌ترین است. الگوریتم‌های مختلف خوشه‌بندی را می‌توان با نوع مقدار ویژگی قابل قبول آنها نظیر پیوسته، گسسته، کیفی یا با روش نمایش هر خوشه یا با روش‌های سازماندهی مجموعه خوشه‌ها (چه به صورت سلسله مراتبی یا به‌صورت فایل‌های مسطح) مشخص کرد. در بخـش بعدی روش خوشـه‌بندی K-means که یک روش پایه‌ای خوشه‌بندی است، ارائه می‌شود.

۱-۳. خوشه‌بندی K-means

شرح مسـئله: مجموعه داده‌هایی با N داده n بعدی x^n را در نظر بگیرید، هدف تعیین تقسیم‌بندی طبیعی مجموعه داده‌ای به k خوشـه و نویز اسـت. می‌دانیم که k خوشـه ناپیوسته شامل N_j نقطه داده‌ای با بردار نشانگر μ_i وجود دارد کـه $j=1,....,k$ الگوریتـم k-means تـلاش می‌کند تا مجموع مربعات تابع خوشـه‌بندی را از رابطه (۱) به حداقل برساند.

$$J=\sum_{j=1}^{k}\sum_{n\in S_j}\left\|x^n-\mu_j\right\|^2 \qquad (۱)$$

کـه در این رابطه μ_j میانگین نقاط داده‌ای در خوشـه S_j است و از رابطه (۲) به‌دست می‌آید.

$$\mu_i=\frac{1}{N_j}\sum_{n\in S_j}x^n \qquad (۲)$$

این دنباله با تخصیص تصادفی نقاط به k خوشـه انجام می‌شـود. سپس بردارهای میانگین μ_j از N_j نقطه را در هر خوشـه محاسـبه می‌کند. برای هر نقطه مجدداً خوشـه جدیدی تعیین می‌شـود که براساس آن بردار نزدیک‌ترین میانگین به‌دست می‌آید. سـپس بردارهای میانگین مجدداً محاسبه می‌شوند.

خوشه‌بندی k-means به شرح زیر پیش می‌رود:

- تعداد خوشه‌ها را که با K نشان داده می‌شود، مشخص کنید؛
- k دسته اولیه را انتخاب کنید؛
- مـواردی را تعییـن کنید که به عضو j از دسـته j که $j=1,...,k$ نزدیک‌ترند؛
- میانگین نمونه‌ها در هر خوشه را محاسبه کنید و مرکز

جدول ۲. تکنیک های داده‌کاوی برای انجام عملیات داده‌کاوی

متغیر	نوع متغیر	سطوح اندازه‌گیری	شرح
سن	پیوسته	فاصله	سن راننده در سال
سن ماشین	پیوسته	فاصله	سن خودرو در سال
نوع ماشین	رده‌ای	اسمی	انواع خودرو
جنس	رده‌ای	باینری	زن ، مرد
سطح پوشش	رده‌ای	اسمی	بیمه‌نامه
آموزش	رده‌ای	اسمی	سطح آموزش رانندگی
محل	رده‌ای	اسمی	محل اقامت
اقلیم	رده‌ای	اسمی	کد آب‌وهوا برای اقامت
رتبه‌بندی اعتباری	پیوسته	فاصله	اعتبار (امتیاز) راننده
شناسایی	ورودی	اسمی	شماره شناسایی راننده

خوشه‌های[1] k را به میانگین خوشه‌هایشان نزدیک کنید؛

– نزدیک‌ترین موارد به مرکز خوشه جدید j متعلق به خوشه j را مجدداً تخصیص دهید؛

– میانگین نمونه‌ها را در هر خوشه به‌عنوان یک مرکز خوشه جدید در نظر بگیرید.

این روش آنقدر تکرار می‌شود تا در خوشه‌بندی تغییر بیشتری دیده نشود (Larose, 2004).

خوشه‌بندی k-means یک روش طبقه‌بندی بدون نظارت است. ازنظر محاسباتی روشی کارآمد است که دانه‌های خوشه اولیه را فراهم می‌کند که به طرز هوشمندانه‌ای قرار دارند. روش‌های خوشه‌بندی به معیار فاصله یا تشابه بین نقاط بستگی دارد. استانداردهای فاصله‌ای متفاوتی در خوشه‌بندی به روش k-means استفاده می‌شود که می‌تواند خوشه‌های متفاوتی به‌وجود بیاورند.

۲-۳. مثال: خوشه‌بندی رانندگان اتومبیل

شرکت بیمه ABC به‌طور متناوب فهرستی از رانندگان را از منابع خارجی خریداری می‌کند. بخش اکچوئری ABC می‌خواهد توالی شکایت برای ادعای خسارت را ارزیابی کند. آنها برمبنای تجربیات می‌دانند توالی شکایت راننده به فاکتورهای جمعیت‌شناسی و جغرافیایی بستگی دارد. در نتیجه آنها می‌خواهند رانندگان را به گروه‌هایی تقسیم کنند که نسبت به این ویژگی‌ها مشابه باشند. پس از اینکه رانندگان تقسیم شدند نمونه‌ای اتفاقی از انتظارات در هر قسمت مورد استفاده قرار می‌گیرد تا توالی را تخمین بزنند. نتایج آزمون به اکچوئری‌ها اجازه خواهد داد تا سود بالقوه را هم به‌طورکلی و هم برای هر دسته خاص ارزیابی می‌کند. داده‌های شبیه‌سازی شده به‌دست‌آمده از فروشنده در جدول ۲ ارائه شده است.

پس از پردازش داده‌ها که باید انتخاب نمونه‌ای اتفاقی از داده‌ها برای تحلیل اولیه را در برگیرد، فیلترکردن مشاهدات و استانداردکردن متغیرها از دسته‌بندی k-means استفاده می‌کنیم تا خوشه‌ها شکل بگیرند.

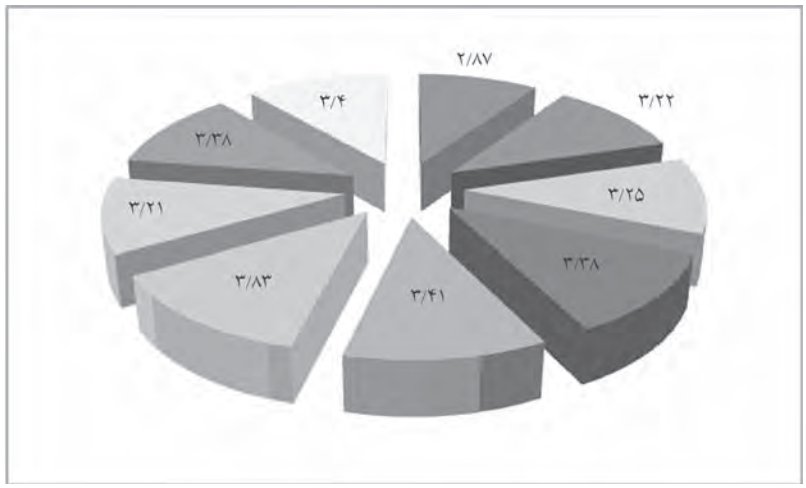
نمودار ۱ نمایش گرافیکی از ویژگی‌های کلی خوشه‌ها فراهم می‌کند.

در نمودار ۱، پهنای هر تکه فاصله میانگین حسابی (انحراف استاندارد) بین نمونه‌ها در دسته ارتفاع به معنای فرکانس است و رنگ، فاصله از دورترین عضو خوشه را نشان می‌دهد.

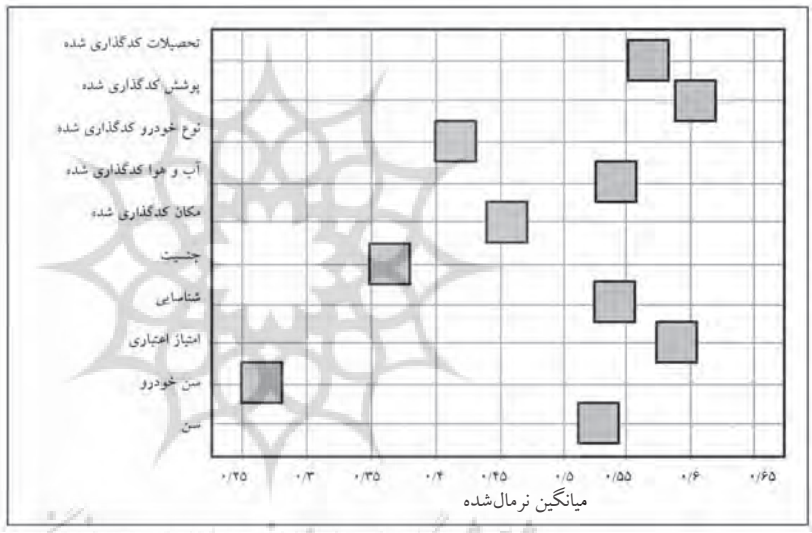
خوشه ۵ بیشترین نمونه را دارد درحالی‌که دسته ۹

1. Seed

نمودار ۱. نمودار خوشه ها برای نمایندگان EMDATA



نمودار ۲. میانگین ورودی کلی



کم‌ترین نمونه را دارد.

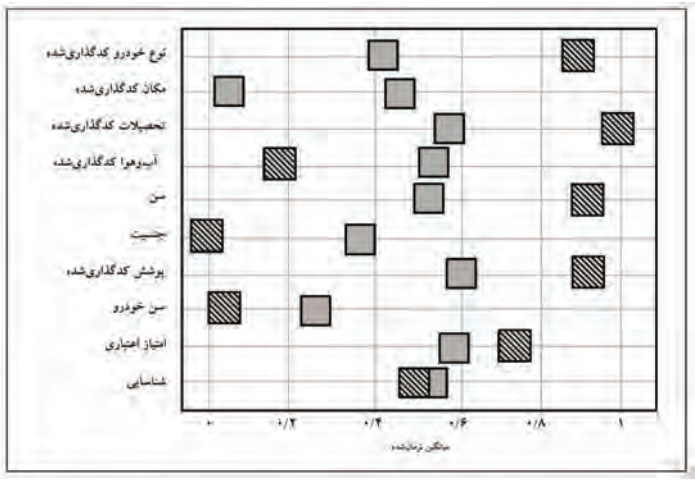
نمودار۲ میانگین ورودی‌ها برای کل مجموعه داده‌های کلی دسته‌ها را نشان می‌دهد. میانگین ورودی‌ها با استفاده از یک مقیاس تبدیل، نرمالیزه می‌شوند.

$$y = \frac{x - \min(x)}{\max(x) - \min(x)}$$

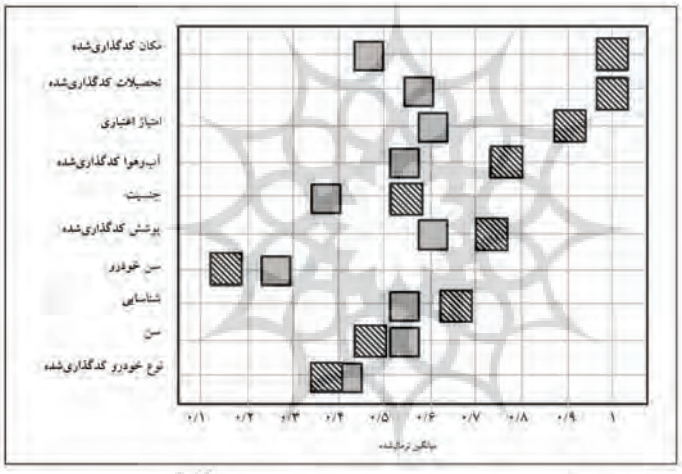
نمودار میانگین نرمال‌شده می‌تواند برای مقایسه میانگین‌های نرمالیزه در کل با میانگین‌های نرمال در هر خوشه مورد استفاده قرار گیرد. نمودار ۳ میانگین ورودی‌ها از دسته ۱ (بلوک‌های هاشورخورده) را با میانگین‌های ورودی کلی (بلوک‌های خاکستری) مقایسه می‌کند. باید میانگین ورودی‌ها برای خوشه‌هایی که با میانگین‌های ورودی کلی تفاوت دارند، شناسایی شوند.

نمودار، ورودی‌ها را برمبنای اینکه چگونه پراکندگی میانگین ورودی‌ها برای خوشه منتخب نسبت‌به میانگین کلی ورودی‌ها است، دسته‌بندی می‌کند. ورودی که بیشترین پراکندگی را دارد، در بالاترین و ورودی که کوچک‌ترین پراکندی را دارد در پایین‌ترین بخش فهرست می‌شود. ورودی با بیشترین پراکندگی معمولاً به بهترین شکل دسته انتخاب‌شده را مشخص می‌سازد (خوشه۱ در نمودار ۳). نمودار ۳ نشان می‌دهد که متغیر «نوع ماشین» و «مکان»، ورودی‌هایی کلیدی هستند که

نمودار ۳. مقایسه میانگین‌های ورودی برای خوشه ۱ و میانگین‌های کلی



نمودار ۴. مقایسه میانگین‌های ورودی برای خوشه ۱ و میانگین‌های کلی



کمک می‌کنند رانندگان را در دسته (۱) از تمام رانندگان مجموعه داده‌ای متمایز کنند.

رانندگان در دسته اول تمایل دارند سطح آموزش بالاتری از متوسط داشته باشند تا رانندگان متوسط در مجموعه داده‌ای دسته ۵، همان‌طور که در نمودار ۴ نشان داده شده تحصیلات بالاتر از متوسط دارند و امتیاز اعتباری بیشتری از متوسط دارند. بیشتر رانندگان در دسته ۵ در منطقه ۴ زندگی می‌کنند و آنها نسبت به رانندگان متوسط ماشین جدیدتری دارند (جدول ۳).

جدول ۳ اطلاعاتی درباره هر دسته را نشان می‌دهد. انحراف استاندارد خطای میانگین را در عرض متغیرهایی از انحراف استاندارد میانگین می‌گیرد که با فاصله جذر مربع میانگین بین نمونه‌ها در دسته برابر است.

در طی فرآیند خوشه‌بندی یک مقدار در بازه ۰ و ۱ برای هر متغیر محاسبه می‌شود.

این مقدار بیانگر اهمیت آن متغیر در تعیین خوشه‌هاست. همان‌طور که در جدول۴ نشان داده شده متغیر «جنسیت» درجه اهمیت صفر دارد که یعنی متغیر به‌عنوان متغیر ایجاد شکاف در تعیین خوشه‌ها مورد استفاده قرار نمی‌گیرد. معیار اهمیت نشان می‌دهد که متغیرها چگونه داده‌ها را به خوشه‌ها تقسیم می‌کنند. متغیرهایی با درجه اهمیت صفر نباید الزاما حذف شود.

جدول ۳. اطلاعات خوشه‌ها

آموزش	پوشش	نوع خودرو	آب و هوا	مکان	جنسیت	سن	سن خودرو	اعتبار امتیاز	میزان فاصله از شبیه‌ترین خوشه	شبیه‌ترین خوشه	حداکثر فاصله از مرکز خوشه	فراوانی خوشه	خوشه
۱/۸۶	۲/۴۳	۳/۵۷	۱/۲۹	۳/۴۳	۱/۰۰	۳۵/۵۷	۳/۲۹	۰/۸۶	۲/۸۲	۵/۰۰	۲/۸۷	۷	۹
۱/۸۵	۲/۸۵	۲/۲۵	۲/۵۵	۲/۸۰	۰/۶۵	۴۶/۶۵	۲/۱۵	۰/۶۲	۲/۴۰	۷/۰۰	۳/۲۲	۲۰	۸
۲/۲۷	۲/۳۶	۱/۴۵	۲/۰۹	۱/۹۵	۰/۲۷	۲۴/۵۹	۲/۷۳	۰/۶۵	۲/۲۵	۲/۰۰	۳/۲۵	۲۲	۷
۱/۷۶	۱/۱۹	۱/۶۷	۱/۴۸	۲/۰۰	۰/۴۳	۳۵/۱۹	۶/۵۲	۰/۸۱	۲/۴۱	۴/۰۰	۳/۳۸	۲۱	۶
۳/۰۳	۲/۳۹	۲/۰۳	۲/۳۳	۳/۸۲	۰/۵۸	۳۲/۷۹	۳/۰۰	۰/۸۲	۲/۳۷	۴/۰۰	۳/۴۱	۳۳	۵
۲/۵۶	۱/۴۴	۲/۷۲	۱/۸۳	۳/۵۰	۰/۳۹	۳۴/۴۴	۵/۱۷	۰/۵۹	۲/۳۷	۵/۰۰	۳/۸۳	۱۸	۴
۲/۰۰	۱/۴۳	۱/۱۴	۲/۴۳	۳/۵۷	۰/۴۳	۲۰/۵۷	۸/۰۰	۰/۴۶	۳/۱۴	۷/۰۰	۳/۲۱	۷	۳
۱/۳۹	۲/۲۸	۱/۲۸	۲/۶۷	۲/۸۹	۰/۲۸	۲۶/۰۰	۳/۵۶	۰/۵۶	۲/۲۵	۷/۰۰	۳/۳۸	۱۸	۲
۳/۰۰	۲/۷۰	۳/۳۰	۱/۵۲	۲/۰۴	۰/۰۷	۴۴/۱۵	۲/۳۷	۰/۷۵	۲/۵۵	۵/۰۰	۳/۴۰	۲۷	۱

جدول ۴. ارزش متغیرها

نام	ارزش
جنسیت	۰
شناسایی	۰
مکان	۰
آب‌وهوا	۰
نوع خودرو	۰/۵۲۹۹۹۳
پوشش	۰/۳۶۳۹۷۲
امتیاز اعتباری	۰/۳۴۳۴۷۸
سن خودرو	۰/۹۴۱۹۵۲
سن	۰
تحصیلات	۰/۷۵۱۲۰۳

تحلیل خوشه‌بندی می‌تواند به‌وسیله ویژگی‌های بیمه مالکیت / تصادفات استفاده شود تا دقت پیشگویانه را با تقسیم‌بندی پایگاه داده‌ها به گروه‌های همگن‌تر تقویت کند. سپس داده‌های هر گروه می‌تواند کشف، تحلیل و مدل‌سازی شود. تقسیم‌بندی‌ها برمبنای انواع متغیرهایی که با ریسک فاکتورها، منافع یا رفتارها همراه می‌شوند اغلب کنتراست‌های شدیدی به‌وجودمی‌آورند که می‌تواند آسان‌تر تفسیر شود.

در نتیجه اکچوئرها می‌توانند به‌طور دقیق‌تری احتمال شکایت و میزان شکایت را پیش‌بینی کنند.

مثلاً یک شرکت بیمه متوجه شد که دسته رانندگان مردی که مدت ۱۸ تا ۲۰ سال رانندگی می‌کنند، نرخ تصادف پائین‌تری نسبت به کل مردان دارند. این زیرگروه در چه متغیری اشتراک دارند که این اختلاف را می‌تواند توجیه کند؟

بررسی داده‌ها نشان می‌دهد که اعضای زیرگروهی که کمترین ریسک را دارند، ماشین‌هایی می‌رانند که به طرز معنی‌داری از سن متوسط پیرتر است و رانندگان ماشین‌های قدیمی‌تر وقت بیشتری صرف ماشین‌های قدیمی‌شان می‌کنند در نتیجه اعضای زیرگروه احتمالاً نسبت‌به سایرین در گروه همسالان، اتومبیل‌هایشان را با احتیاط بیشتری می‌رانند. در نهایت، شناساگر هر مشاهده می‌تواند به سایر گروه‌ها برای استفاده به عنوان یک ورودی، کد شناسایی[1] گروه، یا متغیر هدف انتقال داده شود. مثلاً می‌توانید

1. Identity

خوشه‌هایی برمبنای گروه‌های سنی مختلفی تشکیل دهید، سپس شما می‌توانید الگوهای پیش‌بینی‌کننده‌ای برای هر گروه سنی با انتقال متغیر دسته به‌عنوان متغیر گروه به یک گروه مدل‌سازی بسازید.

۴. داده‌کاوی و پیش‌بینی

در این بخش الگوهای داده‌کاوی برای پیش‌بینی معرفی می‌گردد. بخش ۱-۴ مقدمه‌ای بر الگوریتم درخت تصمیم‌گیری داده‌کاوی ارائه می‌دهد و بخش ۲-۴ الگوی توالی شکایت با درخت‌های تصمیم‌گیری و رگرسیون منطقی ارائه می‌دهد.

۱-۴. درخت‌های تصمیم‌گیری

درخت‌های تصمیم‌گیری بخشی از دسته القایی تکنیک‌های داده‌کاوی هستند. یک درخت تجربی تقسیم‌بندی داده‌ها را ارائه می‌دهد که با به‌کارگیری مجموعه‌ای از قوانین ساده ایجاد می‌شود. هر قانون نگرشی نسبت‌به یک تقسیم‌بندی برمبنای مقدار یک ورودی تعیین می‌کند. یک قانون کلی پس از دیگری به‌کار گرفته‌می‌شود و سلسله مراتبی از تقسیم‌بندی‌ها درون تقسیمات دیگر به‌وجودمی‌آورد. این سلسله مراتب درخت نامیده می‌شود و هر بخش یک گره نامیده می‌شود. یک دسته اصلی کل مجموعه داده‌ای را دربرمی‌گیرد و گره ریشه‌ای یک درخت نامیده می‌شود. یک گره با تمام اجزای بعد از آن تشکیل یک شاخه[1] می‌دهد.

گره‌های نهایی برگ نامیده می‌شوند. برای هر برگ تصمیمی گرفته می‌شود و برای مدل‌سازی به‌کارگرفته‌می‌شود تصمیم خیلی ساده مقدار پیش‌بینی‌شده است.

تکنیک درخت تصمیم‌گیری داده‌کاوی شما را قادر می‌سازد درخت‌های تصمیم‌گیری بسازید که:

- مشاهدات را برمبنای مقدار اسمی، دو تایی یا اهداف ترتیبی طبقه‌بندی کنید؛
- خروجی‌هایی برای اهداف فاصله‌ای پیش‌بینی کنید؛
- هنگامی که جایگزین‌های تصمیم‌گیری را مشخص می‌کنید، تصمیم‌گیری مناسب را پیش‌بینی کنید.

روش‌های درخت تصمیم‌گیری خاص[2]، درخت رگرسیون و طبقه‌بندی و الگوریتم تصمیم‌گیری[3] را شامل می‌شود (Tan, 2004).

CART و CHAID تکنیک‌هایی درخت تصمیم‌گیری هستند که برای طبقه‌بندی مجموعه داده‌ها می‌توانند مورد استفاده قرار بگیرند. این تصمیم‌گیری توصیف مختصری از الگوریتم CHAID برای ساختن درخت‌های تصمیم‌گیری است.

در الگوریتم CHAID ورودی‌ها اسمی یا ترتیبی نیستند. بسیاری از بسته‌های نرم‌افزاری، ورودی‌های با مقیاس فاصله‌ای را می‌پذیرد و به‌طور اتوماتیک مقادیر را به محدوده‌هایی قبل از رشد درخت تقسیم‌بندی می‌کند. برای گره‌هایی با بسیاری از مشاهدات الگوریتم نمونه‌ای برای جستجوی شکاف از محاسبه ارزش استفاده می‌کند (معیار ارزش نشان می‌دهد که یک متغیر چقدر خوب داده‌ها را دسته‌بندی می‌کند و برای مشاهده محدودیت حداقل اندازه یک انشعاب).

نمونه‌ها در گره‌های مختلف به‌طور مستقل به‌دست‌می‌آیند. برای شکاف‌های دو تایی یا اهداف فاصله‌دار شکاف بهینه همیشه یافت می‌شود. برای سایر موقعیت‌ها داده‌ها ابتدا یکی می‌شوند و سپس یا تمام شکاف‌های ممکن ارزیابی می‌شوند یا یک جستجوی درختی مورد استفاده قرار می‌گیرد.

فاز یکسان‌سازی به دنبال گروه‌هایی از مقادیر ورودی می‌گردد که احتمالاً به‌نظرمی‌رسد همان انشعاب در بهترین شکاف را تعیین می‌کنند جستجوی شکاف به مشاهداتی در همان گروه یکسان‌سازی به‌عنوان همان مقدار ورودی مرتبط می‌شود. جستجوی شکاف سریع‌تر است چون کاندیدهای کمتری نیاز به ارزیابی دارند.

بررسی اولیه به هنگام توسعه یک درخت برای پیش‌بینی

1. Branch

2. CART
3. CHAID

درخت‌های تصمیم‌گیری بخشی از دسته القایی تکنیک‌های داده‌کاوی هستند

تصمیم‌گیری این است که درخت چقدر بزرگ باشد یا چه چیزی همان نتیجه را می‌دهد، چه گره‌هایی درخت را هرس می‌کنند.

روش CHAID برای ساخت درخت سطح اهمیتی از آزمون مربع کای[1] را مشخص می‌کند تا رشد درخت را متوقف کند، شاخص‌های شکاف برمبنای سطح معناداری[2] توزیع فیشر[3] با توزیع Chi-Square می‌باشند. برای این شاخص‌ها بهترین شکاف، شکافی با کوچک‌ترین p-value است طبق قرارداد، p-value طوری تنظیم می‌شود تا چندین آزمایش حساب آورده شود.

یک مقدار ازدست‌داده‌شده ممکن است به‌عنوان مقداری مجزا با آن رفتار شود. برای ورودی‌های اسمی یک مقدار از دست داده شده، یک طبقه جدید می‌سازد. برای ورودی‌های ترتیبی یک مقدار ازدست‌رفته آزاد از هرگونه محدودیت ترتیبی است.

جستجو برای شکاف در ورودی‌ها گام‌به‌گام پیش می‌رود. ابتدا یک انشعاب برای هر مقدار از ورودی‌ها تخصیص می‌یابد. انشعابات به‌طور جایگزینی ظهور می‌کنند و همان‌طور که به‌نظرمی‌رسد با p-value تضمین‌شده مجدداً دچار شکاف می‌شوند.

روش جایگزینی مشترک ادامه پیدا می‌کند تا یک شکاف دو تایی ظاهر شود. سپس شکاف با مطلوب‌ترین p-value در میان تمام شکاف‌هایی که الگوریتم بررسی کرده است، سازگار می‌شود.

پس از آنکه شکافی برای یک ورودی سازگار شد p-value آن تنظیم می‌شود و ورودی با بهترین p-value به‌عنوان متغیر شکافتگی انتخاب می‌شود. اگر p-value انتخاب‌شده کوچک‌تر از آستانه مشخص‌شده باشد، گره شکافته می‌شود. ساخت درخت هنگامی که تمام p-value های تنظیم‌شده متغیرهای شکافتگی در گره‌های غیرشکافته هم بالاتر از آستانه مشخص‌شده توسط کاربر باشند، پایان می‌یابد.

تکنیک‌های درخت دیدگاه‌هایی در فرآیند تصمیم‌گیری فراهم می‌کند که بیان می‌کند که چگونه نتایج به‌دست‌می‌آیند.

درخت تصمیم‌گیری، کارآمد است و بنابراین برای مجموعه‌های داده‌ای بزرگ مناسب است. درخت‌های تصمیم‌گیری شاید موفق‌ترین روش برون‌یابی برای کشف ساختار داده‌های دارای انحراف باشد. درخت‌ها به‌طور برگشت‌پذیر معنای داده‌ای ورودی را تقسیم‌بندی می‌کنند تا تست‌هایی که همگن هستند را شناسایی کنند. اگرچه درخت‌های تصمیم‌گیری می‌توانند داده‌ها را به چند بخش همگن بشکنند و قوانین ایجادشده توسط درخت می‌تواند مورد استفاده قرار بگیرد تا اثرات متقابل میان متغیرها را ردیابی کند، ولی نسبتاً ناپایدار است و ردیابی روابط خطی یا درجه دوم بین متغیر پاسخ و متغیرهای دشوار می‌باشد (Michalski et al, 1998).

۲-۴. مدل‌سازی توالی شکایت

اکنون ما فرآیند مدل‌سازی را با مطالعه رابطه بین فراوانی ادعای خسارت و فاکتورهای ریسک شامل سن، جنس، اعتبار، مکان، سطح تحصیلات، پوشش و سن ماشین شروع می‌کنیم. دوباره داده‌های شبیه‌سازی شده مورد استفاده قرار می‌گیرد.

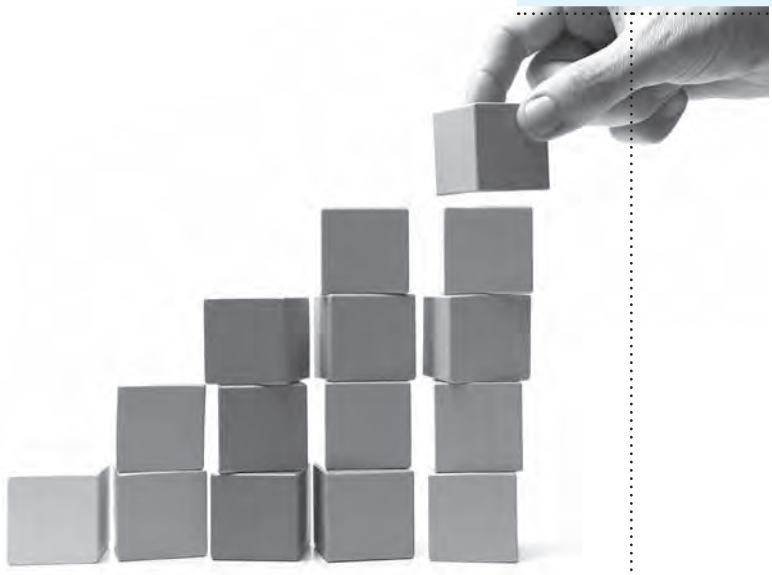
فرآیند مدل‌سازی ترکیبی از تکنیک‌های درخت تصمیم‌گیری رگرسیون منطقی است.

ابتدا الگوریتم درخت تصمیم‌گیری را استفاده می‌کنیم تا فاکتورهایی را شناسایی کنیم که توالی شکایت را تحت‌تأثیر قرار می‌دهند. پس از آنکه این فاکتورها شناسایی شدند تکنیک رگرسیون منطقی استفاده می‌شود تا توالی شکایت و اثر فاکتور ریسک کمی شود.

اکنون از الگوریتم درخت تصمیم‌گیری استفاده می‌کنیم تا تأثیرات و اهمیت فاکتورهای ریسک توالی شکایت را آنالیز کنیم. الگوریتم درخت مورد استفاده در این تحقیق SAS / Enterprise Miner است.

۱۰۰ درخت رگرسیون دوتایی و ۱۰۰ درخت شبه CHAID را برای درخت تصمیم‌گیری بهینه ساختیم.

1. Chi-Square
2. P- Value
3. Fisher



تحلیل درخت تصمیم‌گیری نشـان داد که میزان اعتبار بیشـترین تأثیر را بر توالی شـکایت دارد. توالی شکایت و اندرکنـش میان فاکتورهای مختلف که توالی شـکایت را تحت‌تأثیر قـرار می‌دهد با تغییر حالت میـزان اعتبار تغییر می‌کند.

به‌عـلاوه تأثیر معنـی‌داری در وضعیت اعتبـاری بالاتر وجود دارد.

یک نمودار درختی موارد زیر را شامل می‌شود:

- **گـره ریشـه‌ای:** بالاترین گـره در درخت که تمام مشاهدات را شامل می‌شود.
- **گره‌های داخلی:** گره‌های غیرانتهایی (شـامل گره ریشه‌ای) که قوانین شکافتگی را دربرمی‌گیرد.
- **گره‌هـای بـرگ:** گره‌های انتهایی کـه طبقه‌بندی نهایی برای مجموعه‌ای از مشاهدات را دربرمی‌گیرد.

نمـودار درختی آمار گره، نام متغیرهای مورد اسـتفاده برای شـکافتن داده‌ها به گره‌ها و مقادیر متغیرها برای چند سـطح از گره‌هـا در درخت را نشـان می‌دهـد. نمودار ۵ پروفایل جزئی نمودار درختی را برای آنالیز نشان می‌دهد.

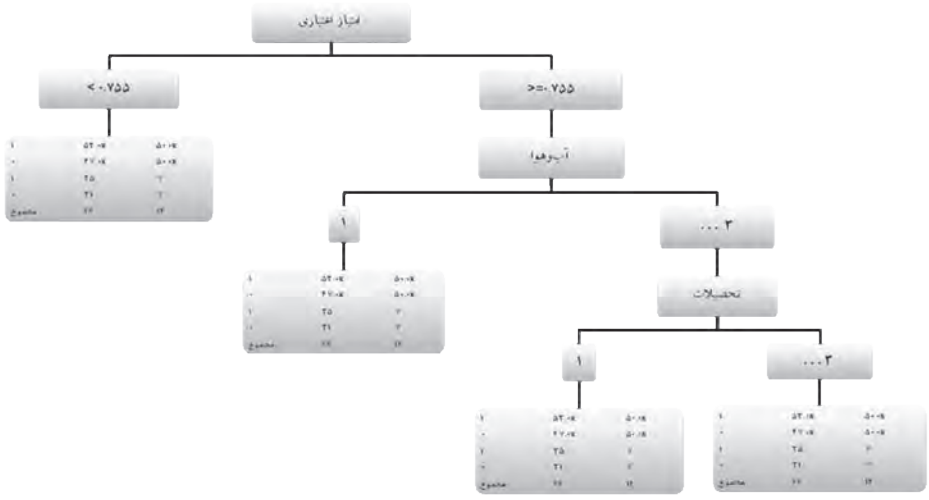
در نمـودار ۵، هر برگ درصد و تعـدادی از مقادیری را نشـان می‌دهد که برای تعیین انشـعابات مورد اسـتفاده قرار می‌گیرند. سـتون دوم اطلاعاتی از داده‌های آزمایشی شـامل درصـد برای هر سـطح هدف، شـمارش برای هر سـطح هدف و شـمارش کل را در برمی‌گیرد. ستون سوم اطلاعاتی برای اعتبارسـنجی داده‌ها شامل درصد برای هر سـطح هدف، شمارش برای هر سـطح هدف و تعداد کل را در برمی‌گیرد. مثلاً در میـان این رانندگان با مقادیر زیر ۷۵/۵٪، ۵۳٪ از آنها شـکایتی از داده‌های آزمایشـی را به ثبت رساندند.

مقادیـر ارزیابی مورد اسـتفاده قرار می‌گیـرد تا به‌طور برگشـت‌پذیر داده‌ها را در زیرگروه‌ها همگن تقسیم‌بندی کند.

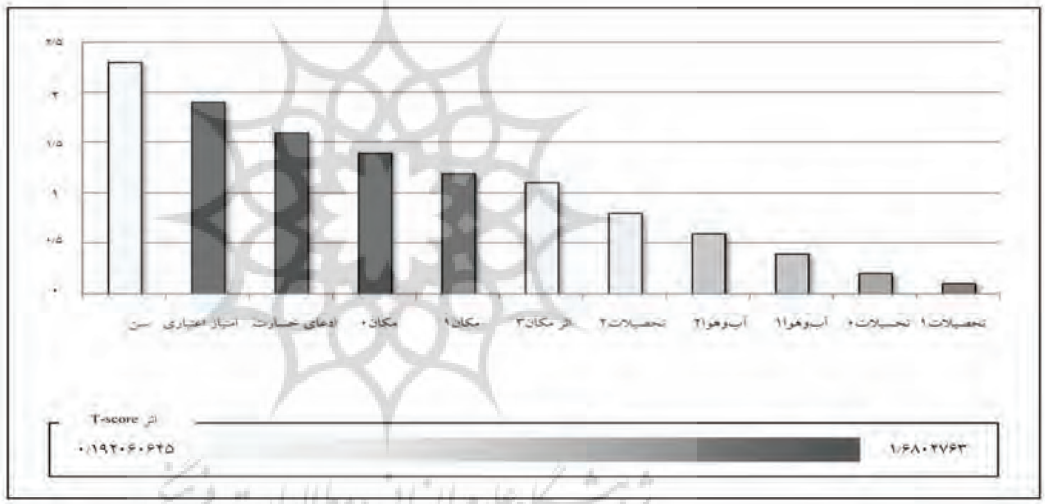
روش بازگشـتی اسـت؛ زیرا هر زیرگروه از شکافتگی یک زیرگروه از یک شکاف قبلی نتیجه می‌شود.

نشـانگرهای عـددی مسـتقیماً بـالای هر گره نشـان می‌دهـد که در کدام نقطه الگوریتم درخت شـکاف‌های معنی‌داری در سطح وقفه‌ها یافت می‌شود یا در شکاف‌های طبقه‌بندی‌شـده برای توزیع اسـمی یا ترتیبی نشـانگرهای کاراکتری نسـبت به هر شـکاف و نام متغیرها تقسیم‌بندی می‌شـود. شما می‌توانید مسـیرهایی از ریشه به هر برگ را دنبال کنید و نتایج را به عنوان قانون بیان کنید.

نمودار ۵. نمودار درختی



نمودار۶. نمودار میله‌ای اثر T-erocs از تحلیل رگرسیون منطقی



نمودار۷. Lift Chart

همان‌طور که در نمودار ۵ نشان داده شده توالی شکایت با مهم‌ترین ریسک فاکتور (وضعیت میزان اعتبار در این مطالعه) در میان سایر متغیرها تغییر می‌کند، برمبنای آنالیز درخت، سن ماشین، پوشش و نوع ماشین فاکتورهای غیرمرتبط هستند. آنها نباید در الگوی توالی شکایت وارد شوند.

برمبنای تحلیل درخت، اکنون رگرسیون را مورد استفاده قرار می‌دهیم تا احتمال ادعای خسارت برای هر راننده را بر مبنای فاکتورهای تحت بررسی تخمین بزنیم. همان‌طور که در بخش ۲ مورد بحث قرار گرفت، رگرسیون لجستیک تلاش می‌کند تا احتمال یک شکایت را به عنوان تابعی از یک یا چند ورودی مستقل پیش‌بینی کنید.

نمودار ۶ نمودار میله‌ای اثر T-score از تحلیل رگرسیون را نشان می‌دهد. اثر T-score با پارامتر تخمین‌زده‌شده تقسیم بر خطای استاندارد برابر است.

امتیازات به‌وسیله کاهش مقدار مطلق در نمودار مرتب می‌شود. شدت رنگ اندازه امتیاز برای یک میله را نشان می‌دهد. امتیازات حداقل و حداکثر امتیاز را در سمت چپ و راست اختصارات نشان می‌دهد. محور عمودی مقدار مطلق برای اثر را نشان می‌دهد. در این مثال، اولین متغیر، سن، بالاترین مقدار مطلق را دارد و اعتبار دومین مقدار مطلق بزرگ است. تخمین مکان و تحصیلات مثبت هستند بنابراین مقدار میله‌ها با سایه خاکستری رنگ می‌شود.

برآوردهای سن و امتیاز اعتبار مقدار منفی دارند بنابراین میله‌هایشان به رنگ متفاوتی نشان داده می‌شوند.

ارزیابی بخش نهایی فرآیند داده‌کاوی است. شاخص ارزیابی مقایسه سود واقعی و مورد انتظار یا زیان واقعی و مورد انتظار از نتایج الگو است. این شاخص مقایسه‌ها و ارزیابی‌های بین مدلی مستقل از تمام فاکتورهای دیگر را مقدور می‌سازد. نمودار ۷ درصد تجمعی ادعای خسارت برای الگوی رگرسیون را نشان می‌دهد.

Lift Chart درصد ادعای خسارت را روی محور عمودی نشان می‌دهد. در این نمودار رانندگان هدف از چپ به راست به وسیله افرادی که بیشترین احتمال تصادف را دارند طبقه‌بندی می‌شوند. گروه دسته‌بندی شده به دهک‌ها در طول محور x توزیع می‌شود. اولین دهک از سمت چپ، بیشترین احتمال تصادف را دارد. محور عمودی درصد تجمعی ادعای خسارت را پیش‌بینی می‌کند. Lift Chart درصد تجمعی را برای یک الگوی خطی تصادفی نشان می‌دهد. کیفیت کارایی یک الگو با درجه‌ای که Lift Chart به سمت بالا و چپ می‌رود، بیان می‌شود. در این مثال الگوی رگرسیون حدود ۳۰٪ از رانندگان را در دهک آخر در بر می‌گیرد. الگوی رگرسیون قدرت پیش‌بینی بیستمین تا هشتاد بین درصدها دارد. در حدود ۹۰ اطمینان مقدار درصد ادعای خسارت تجمعی برای الگوی پیش‌بینی کننده، حدود همان خط پایه مدل است.

۵. نتیجه‌گیری

این مقاله رویکرد داده‌کاوی نسبت به ریسک بیمه و چند کاربرد آن را معرفی کرد. در این مقاله مقدمه‌ای برای عملیات و تکنیک‌های داده‌کاوی فراهم شد و دو کاربرد بالقوه آن در اکچوئری اموال/ حوادث توصیف شده که در یک بخش از خوشه‌بندی k-means استفاده کردیم تا گروهی از رانندگان را با تقسیم‌بندی بهتر توصیف کنیم. در بخش دیگری چندین فاکتور ریسک برای رانندگان خود را با هدف پیش‌بینی توالی شکایت بررسی کردیم.

تأثیرات و روابط این فاکتورها بر توزیع شکایت با آنالیز داده‌ای برون‌یابی و الگوریتم درخت تصمیم‌گیری شناسایی می‌شود. رگرسیون لجستیک سپس برای مدل کردن توالی شکایت استفاده شد.

به‌دلیل کاربرد داده‌های شبیه‌سازی شده فرضی مثال‌ها مزیت محدودی از داده‌کاوی نسبت به تحلیل اکچوئری سنتی مشخص می‌شود. اهمیت اصلی داده‌کاوی را تنها در پایگاه‌داده بزرگ مشخص می‌شود. کلید به‌دست آوردن مزیت رقابتی در صنعت بیمه در شناسایی پایگاه‌داده‌های مشتریان است و اگر بدرستی مدیریت، تحلیل و بکار گرفته شوند دارایی‌های منحصر بفرد و با ارزش هستند. شرکت‌های بیمه می‌توانند بینش موجود در پایگاه‌های

داده‌ای مشـتریان را از طریق تکنولـوژی مدرن داده کاوی رمزگشایی کنند. داده کاوی از مدل‌سازی پیش‌بینی کننده، تقسـیم‌بندی پایگاه‌داده‌هـا، تحلیل سـبد بـازار و ترکیب تئوری‌ها برای بدست آوردن پاسخ‌های سریع‌تر به سؤالات برای تجارت با دقت بیشـتر اسـتفاده می‌کند. با استفاده از ابزار داده کاوی می‌توان محصولات جدیدی توسـعه داد و اسـتراتژی‌های بازاریابی مدرن بکار گرفت و نهایتاً شرکت بیمـه قادر خواهد بـود اطلاعات را از ابـزار داده کاوی به قابلیت پیش‌بینی، تفسیرپذیری و دانش تبدیل کند.

منابع

۱. شهرابی، جمال ۱۳۸۸، *داده کاوی*، انتشارات دبیرخانه دائمی کنفرانس داده کاوی ایران، تهران، چ ا.

۲. غضنفری، مهدی، علیزاده، سمیه و تیمورپور، بابک ۱۳۸۷، *داده کاوی و کشـف دانش*، انتشـارات دانشگاه علم و صنعت ایران، تهران، چ ا.

۳. صفوی، بهاره ۱۳۸۷، 'کشف دانش پنهان درون داده‌ها'، روزنامه جام‌جم، ش ۲۴۲۰، ۸۷/۸/۱۴

۴. مقدم، قربان، '*داده کاوی یک ابزار تحلیل مدیریتی*'، موسسه آموزش عالی روزبه زنجان.

5. Alpaydin, E 2004, *Introduction to machine learning*, MIT Press.

6. Hair, JF 2005, *Multivariate data analysis*, Prentice Hall.

7. Hand, D, Heikki, M & Padhraic, S 2007, *Principles of data mining*, The MIT Press.

8. Larose, DT 2004, *Discovering knowledge in data: an introduction to data mining*, Wiley.

9. Michalski, RS, Bratko, I & Kubat, M 1998, *Machine learning and data mining: methods and applications*, Wiley Edition.

10. Tan, SK 2004, *An introduction to data mining*, Wiley.