دسته‌بندی مشتریان بیمه با استفاده از داده‌کاوی

نویسنده: سید محمود ایزدپرست

– کارشناس‌ارشد مدیریت فناوری اطلاعات، دانشگاه پیام نور تهران

چکیده

امروزه نقش مشتریان از حالت پیروی از تولیدکننده، به هدایت تولیدکنندگان مبدل گشته است، به‌همین‌دلیل دسته‌بندی مشتریان، در هدفمند ساختن سازمان‌ها در سفارشی‌سازی خدمات‌شان و نیز الویت‌بندی محصولات براساس میزان سودآوری آن محصول کمک شایانی می‌کند. روش داده‌کاوی برای دستیابی به قوانین تصمیم‌گیری و مدل پیش‌بینی رفتار مشتریان آتی در یکی از شرکت‌های بیمه استفاده شده است. در اجرای روش دسته‌بندی در داده‌کاوی، دو تکنیک درخت تصمیم و شبکه‌های عصبی در یکی از شرکت‌های بیمهٔ خصوصی به‌کاررفته‌است. هدف این پژوهش استفاده از روش شبکه‌های عصبی و تکنیک درخت تصمیم به‌منظور دسته‌بندی مشتریان بیمه و درنهایت ارزیابی نتایج به‌دست‌آمده‌است. به‌منظور ارزیابی مدل‌ها، نتایج حاصل از دو مدل را مقایسه کردیم که تطابق آنها نشان‌دهندهٔ صحت عملکرد مدل‌هاست. البته بررسی‌ها نشان داده‌اند که روش درخت تصمیم نتایج بهتری را دربرداشته‌است و این بدان معنی است که روش درخت تصمیم، روش مناسب‌تری برای دسته‌بندی مشتریان بیمه ایجاد می‌کند. همچنین نتایج به‌دست‌آمده از این تحقیق، توسط خبرگان صنعت بیمه تأیید و نظرات آنها به شکل توضیحی از مشخصه‌ها یا متغیرهای ورودی تحقیق ارائه شده است.

واژگان کلیدی: داده‌کاوی، بیمه، فناوری اطلاعات، دسته‌بندی، درخت تصمیم، شبکه‌های عصبی

مقدمه

در جهان کنونی که امـکان تولید انبوهٔ کالا و خدمات، زمینهٔ لازم افزایش عرضه، نسـبت‌به تقاضا را فراهم آورده است، برای تولیدکنندگان راهی جز جلب رضایت مشتری باقی نمانده و دیگر نمی‌توان حیطهٔ بازار و عرضه را با ابزارهای محدود گذشته تعریف کرد (Berson et al , 2001). تجربه نشـان داده است، سـازمان‌هایی که ازنظر سنتی به مفاهیم مشـتری، کالا، بازار، فـروش، خرید، رقابـت، تبلیغات و کیفیـت نـگاه می‌کنند، علاوه‌بر عدم کسـب موفقیت، سـرمایه‌های خود را هم از دست‌می‌دهند. با ظهور اقتصاد رقابتی، مفاهیمی چون مشـتری‌مداری و کسـب رضایت مشـتری، پایه و اساس کسـب‌وکار تلقی شده و سازمانی که بدان بی‌توجه باشـد و این مطلـب را درنظر نگیرد از صحنه بازار حذف خواهد شد.

در این تحقیق ابتدا داده‌های جمع‌آوری‌شده را توسط روش‌های پیش‌پردازش غربـال می‌کنیم تا ناخالصی‌ها و داده‌های ناقص آن حذف گردد. سـپس سعی می‌کنیم با اسـتفاده از روش‌های داده‌کاوی و به‌خصوص روش‌های دسته‌بندی[1]، مشـتریان بیمه را براسـاس ویژگی‌های‌شان دسـته‌بندی کنیـم و بدین‌ترتیب مشـتریان آتـی بیمه را می‌توان در یکی از دسته‌ها قرار داد و باتوجه به اطلاعاتی که از هرکدام از این دسـته‌ها داریم می‌توان ویژگی‌های آنها را پیش‌بینی کرد. در این مقاله ابتدا مقدمه‌ای در زمینه داده‌کاوی ارائه می‌کنیم و سپس روش‌های دسته‌بندی که به‌کارگرفته‌شـده (درخت تصمیم[2] و شبکه‌های عصبی[3]) و اینکه چگونه از این روش‌ها برای دسـته‌بندی مشتریان اسـتفاده کرده‌ایم را توضیح می‌دهیم. پس از آن به‌منظور ارزیابـی نتایـج به‌دسـت‌آمده، این دو روش را مقایسـه می‌کنیـم تا ببینیم کـدام روش، نتایـج دقیق‌تری حاصل می‌نماید. در نهایت پیشـنهادهایی برای تحقیقات آتی در این زمینه ارائه می‌کنیم.

1. Classification

۱. مفهوم داده‌کاوی[4]

از زمانـی که علـم آمار به‌وجودآمد دانشـمندان نیاز به کشف خصوصیات داده‌ها را احساس کرده بودند. بااستفاده از آمار در آن زمان، خصوصیات داده‌ها از قبیل پراکندگی و تمرکز آنها بررسـی می‌شـد. با افزایش نیاز به استفاده از داده‌هـا و درک ارزش اطلاعـات، داده‌هـا به سـرعت در حال گردآوری و ذخیره‌شـدن می‌باشـند که این سـرعت همه روزه در حال افزایش اسـت (Chen, & Su, 2006). به مـوازات سـرعت زیاد ذخیره‌شـدن داده‌هـا، ابزارهای محاسـبهٔ مکانیکی، الکتریکی و درنهایت کامپیوتری نیز به‌وجودآمده‌انـد که نرخ افزایش سـرعت محاسـبهٔ این ابزارها نیز نمایی اسـت. حجم زیاد داده‌های ذخیره‌شده و نیز گستردگی ابعاد آن که معمولاً از منابع گوناگون تهیه شـده بودند و بعضاً دارای قالب‌هـای متفاوتی نیز بودند،

2. Decision Tree
3. Neural Network
4. Data Mining

سبب شد که در بسیاری موارد روش‌های آماری به تنهایی قادر به کشف خصوصیات داده‌ها نباشند. زمانی که می‌خواهیم به بررسی تأثیر تعداد کمی از عوامل بر روی هدف بپردازیم معمولاً روش‌های آماری مناسب است. ولی زمانی که تعداد این عوامل زیاد می‌شود دیگر این روش‌ها کارایی مناسبی ندارند. مثلاً در تحلیل داده‌های سبک زندگی افراد[1] به‌دلیل اینکه این داده‌ها دارای ابعاد بسیار زیادی می‌باشند کمتر از روش‌های آماری استفاده می‌شود. دانشمندان برای رفع این مشکل تصمیم گرفتند که از سرعت بالای کامپیوترها استفاده کنند، همین امر سبب شد که روش‌های ابتکاری دیگری علاوه‌بر روش‌های آماری مثل شبکه‌های عصبی و الگوریتم ژنتیک برای کمک به‌این‌منظور ایجاد شود (SY, 2001).

سه موضوع ذخیره‌سازی داده‌ها، افزایش سرعت کامپیوترها و پیدایش الگوریتم‌های جدید کار با داده‌ها، عامل ایجاد علمی به نام داده‌کاوی شده است. داده‌کاوی عبارت است از «استخراج اطلاعات و دانش و کشف الگوهای پنهان از یک پایگاه داده‌های بسیار بزرگ». که این الگوها و دانش‌ها معمولاً مستتر در داده می‌باشند (Chan & Lewis, 2002).

از داده‌کاوی می‌توان برای انجام کارهایی مثل دسته‌بندی، پیش‌بینی[2]، تخمین[3] و خوشه‌بندی[4] داده‌ها استفاده کرد. برای انجام این کارها تکنیک‌هایی توسعه یافته‌اند که باتوجه به پیشرفت کامپیوترها و این علم همه روزه بر تعداد و کیفیت این تکنیک‌ها افزوده می‌شود. تعدادی از معروف‌ترین این تکنیک‌ها عبارت‌اند از: الگوریتم‌های خوشه‌بندی[5]، شبکه‌های عصبی، الگوریتم ژنتیک[6]، نزدیک‌ترین همسایگی[7] و درخت تصمیم‌گیری (Clifton & Thuraisingham, 2001).

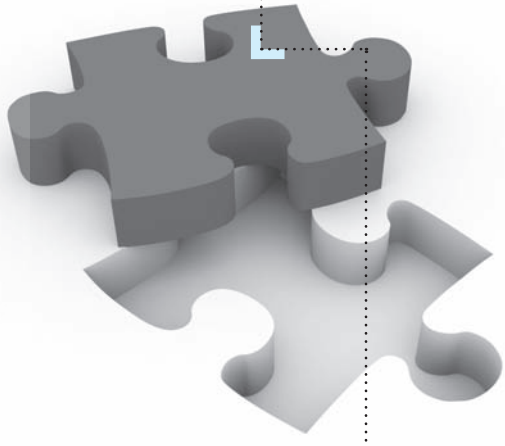
۲. تشریح تکنیک دسته‌بندی

دسته‌بندی از مسائل متعارفی است که به‌طور گسترده توسط متخصصان آمار و محققان فراگیری ماشینی مطالعه شده است. ارائه یک تعریف دقیق از روش دسته‌بندی، دشوار است اما مطابق با تعریف کلی، تکنیک دسته‌بندی، جداسازی یا قراردادن اجزا یا اشیا در تعدای از کلاس‌هاست (Tan, & Yu, 2006).

اگر کلاس‌ها بدون آزمون داده‌ها (به‌طور غیرتجربی) ساخته شده باشند، دسته‌بندی استقرایی[8] نامیده می‌شود. در مقابل، اگر کلاس‌ها به‌طور تجربی ساخته شده باشند (با آزمایش داده‌ها)، دسته‌بندی مؤخر[9] نامیده می‌شود. در

1. Demographic

2. Prediction
3. Estimation
4. Clustering
5. Cluster Detection Algorithm
6. Genetic Algorithm
7. Nearest Neighboring
8. Apriori Classification
9. Posteriori Classification



بیشتر ادبیات حاکم بر دسته‌بندی فرض می‌شود که کلاس‌ها استقرایی باشند و دسته‌بندی که شامل آموزش سیستمی شده و در زمان معرفی یک شیء جدید به سیستم آزموده شده است، قابلیت اختصاص آن شیء به یکی از کلاس‌های موجود را دارد. این روش همچنین یادگیری هدایت‌شده نیز نامیده می‌شود، برخی دیگر از این شیوه‌ها برای دسته‌بندی مؤخر یا هدایت‌نشده در دسترس هستند (Gupta, 2006) که در آن کلاس‌ها براساس داده‌های موجود تعیین شده‌اند. داده‌کاوی کاربرد جدیدی را در دسته‌بندی ایجاد کرده است، از زمانی که مجموعه‌های داده در داده‌کاوی افزایش یافته‌اند، روش‌های جدید دسته‌بندی به‌منظور سروکار داشتن با میلیون‌ها شیء توسعه یافته‌اند که شاید دارای ده‌ها یا صدها مشخصه باشند.

یک فرآیند دسته‌بندی که در آن کلاس‌ها از پیش تعریف شده‌اند، به روشی نیاز دارد که سیستم دسته‌بندی را به‌منظور تخصیص اشیا به کلاس‌ها آموزش دهد؛ آموزشی که بر پایهٔ یک آموزش نمونه است، در واقع به‌صورت مجموعه‌ای از داده‌ها در جایی که برای هر نمونه، یک کلاس از قبل تعریف شده است. فرض می‌شود کلاس هر شیء با داشتن برخی مشخصه‌ها و براساس هرکدام از آنها به ما گفته می‌شود. این مشخصه برای داده‌های آموزشی شناخته شده است، اما برای داده‌های دیگر به‌جز داده‌های آموزشی (ما این داده‌های دیگر را داده‌های آزمایشی می‌نامیم) فرض می‌شود که ارزش هر مشخصه شناخته نشده و باید با روش‌های دسته‌بندی مشخص گردد. این مشخصه ممکن است به‌عنوان خروجی تمام مشخصه‌های دیگر درنظرگرفته‌شود و معمولاً به‌عنوان مشخصه خروجی یا مشخصه وابسته بیان شده است. مشخصه‌های دیگر به جز مشخصه خروجی، مشخصه‌های داخلی یا مشخصه‌های مستقل نامیده می‌شوند (Wu & Su, 2005).

۳. 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 تست بر روی ارزش مشخصه و هر شاخه، خروجی هر تست را نمایش می‌دهد، برگ‌های درخت نیز نمایانگر کلاس‌هاست. به‌طورعادی، پیچیدگی یک درخت تصمیم با افزایش تعداد مشخصه‌ها افزایش می‌یابد. اگرچه در بعضی از شرایط دیده شده است که تنها تعداد کمی از مشخصه‌ها می‌توانند کلاسی را که هر شیء به آن تعلق دارد ،تعیین کنند و بقیه مشخصه‌ها کم یا بی‌تأثیرند (Tan & Yu, 2006).

1. Node

در ساخت درخت‌های تصمیم معمولاً داده‌ها را به دو دسته تقسیم می‌کنند:

– داده‌های آموزشی[1] که برای ساخت مدل مورد استفاده قرار می‌گیرند؛

– داده‌های تست[2] که برای تست و ارزیابی مدل ساخته‌شده کاربرد دارند.

کیفیت داده‌های آموزشی اغلب نقش مهمی در تعیین کیفیت درخت تصمیم دارد. درصورتی‌که آموزش سیستم زیاد شود یعنی داده‌هایی که برای آموزش و ساخت مدل به‌کارمی‌رود درصد زیادی از داده‌ها باشد، دچار حالتی به نام «آموزش بیش از حد مدل» خواهیم شد که به‌دلیل وجود موارد غیرعادی در داده‌های آموزشی، خطا تولید می‌کند (Chan & Lewis, 2002).

اندازه‌گیری کیفیت یک درخت تصمیم مسئلهٔ مهمی است. دقت تعیین‌شدهٔ دسته‌بندی با استفاده از داده‌های آزمایشی به‌طور آشکار یک شاخص مطلوب است. اما ممکن است شاخص‌های دیگری نیز برای اندازه‌گیری به‌کاررروند. این شاخص‌ها شامل هزینهٔ متوسط و هزینهٔ بدترین مورد در دسته‌بندی یک شیء است. یک درخت تصمیم ممکن است قادر به دسته‌بندی داده‌های آموزشی با دقت ۱۰۰ درصد باشد، اما بدین معنا نیست که درخت، روی داده‌های آزمایشی که بخشی از مجموعه آموزشی نبوده، بسیار دقیق باشد. در واقع ممکن است درصورتی‌که داده‌های آموزشی نمونه خوبی از گروه داده‌ها باشند، عملکرد درخت بر روی داده‌های آزمایشی نیز تا ۱۰۰ درصد بالا بیاید. در نتیجه ورودی‌های آموزشی اهمیت بالایی دارند. به‌همین‌دلیل معیارهایی را برای این ورودی‌ها درنظرمی‌گیریم تا با رعایت آن بتوانیم نتایج بهتری را به‌دست‌آوریم. در ادامه چند مورد از این معیارها را بیان می‌کنیم:

– **معیار تعدیل دسته‌های یک پیش‌بینی‌کنندهٔ دسته‌ای**: زمانی که درخت رشد می‌کند، گره‌های جزئی (فرعی) با جداکردن گره‌های اصلی ساخته می‌شوند. پیش‌بینی‌کننده‌ای که برای این جداکردن استفاده شده، معیار تعدیل دسته نامیده می‌شود.

– **معیار حداقل اندازهٔ گره**[3]: مشخص‌کننده این امر است که هر گره چند زیرگره می‌تواند داشته باشد. هرچه این مقدار بیشتر باشد، درخت کوچک‌تر خواهد شد.

– **معیار حدکثر خلوص**[4]: می‌توان این معیار را انتخاب نکرد، اما با این انتخاب، مقدار آن باید بین صفر تا ۱۰۰٪ باشد. هرچه این مقدار بیشتر باشد، درخت بزرگ‌تر خواهد شد.

– **معیار حداکثر عمق**[5]: مقدار آن باید عدد صحیحی بین صفر تا ۲۰ باشد. هرچه این مقدار بیشتر باشد، درخت بزرگ‌تر خواهد شد.

– **قطع درخت**[6]: با این گزینه می‌توان تأثیر قطع درخت را بررسی کرد.

– **آموزش و آزمودن داده‌ها:** مشخص‌کنندهٔ درصد داده‌هایی است که برای آموزش یا تست مدل استفاده می‌شود. اگر مجموعهٔ آزمایشی به‌طور دستی انتخاب شود، ایدهٔ خوبی است برای بررسی اینکه تمام دسته‌های کلاس که در داده‌های آموزشی وجود دارد، مشابه داده‌های آزمایشی ایجاد شده باشند (Adda & Missaoui, 2007).

۴. طراحی معماری شبکهٔ عصبی

در ساخت معماری شبکهٔ عصبی، موارد زیر تعیین می‌شوند:

– تعداد ورودی‌ها؛

– تعداد لایه‌های پنهان شبکه؛

– پارامترهای یادگیری؛

– تعیین نوع انتخاب داده‌های آموزشی و آزمایشی؛

– سایز اولیهٔ تابع وزنی؛

– تعداد سیکل‌های آموزشی داده‌ها؛

– افزودن تکانه[7]: می‌توان قانون تغییر وزن‌ها را طوری درنظرگرفت که تغییر وزن در تکرار حدی، به اندازه تغییر وزن در تکرار قبلی بستگی داشته باشد. افزودن تکانه

1. Train Data
2. Test Data
3. Minimum Node Size Criterion
4. Maximum Purity Criterion
5. Maximum Depth Criterion
6. Pruning Option
7. Momentum

باعث می‌شود تا با حرکت در مسیر قبلی در سطح خطا از گیرافتادن در مینیمم محلی و قرارگرفتن در سطوح صاف پرهیز شود و با افزایش تدریجی مقدار پله تغییرات، سرعت جستجو افزایش یابد (Belhadjali & Whaley, 2004).

۵. دریافت نتایج مدل‌سازی

خروجی اول، یک مدل شبکه عصبی است که عمدتاً مجموعه‌ای از وزن‌ها بین لایه‌های شبکه است. بعد از اتمام اجرا، باید مجموعهٔ نهایی وزن‌ها نیز در صفحهٔ محاسبه ذخیره شود. خروجی دیگر مدل، محاسبه مقادیر خطای نسبی مطلق[1] و میانگین مربعات خطا[2] بر روی مجموعه داده‌های آموزشی می‌باشد. نهایتاً اعتبارسنجی صورت گرفته روی مدل ارائه خواهد شد. میانگین مربعات خطا معیار مناسبی برای تعیین این امر است که مدل ایجادشده تا چه حدی می‌تواند واقعیات را پیش‌بینی کند. در واقع خروجی نهایی مدل‌سازی، وزن‌های تعیین‌شده برای یال‌ها و نیز میزان خطایی است که این وزن‌ها در پیش‌بینی دارند (Sivanandam & Sumathi, 2006).

۶. تجزیه‌وتحلیل نتایج درخت تصمیم

درخت تصمیم، روشی شناخته‌شده به‌منظور دسته‌بندی است که نتایج آن در یک فلوچارت شبیه ساختار درخت ارائه می‌شود (بدین‌صورت که هر گره نشانگر یک آزمون بر روی ارزش مشخصه و هر شاخه، خروجی آن را نمایش می‌دهد، برگ‌های درخت نیز نمایانگر کلاس‌هاست).

برای تحلیل داده‌ها از نرم‌افزار قدرتمند Ctree استفاده می‌کنیم، که ابزار و امکانات زیادی برای ایجاد درخت تصمیم و بررسی نتایج آن و تحلیل در اختیار ما قرار می‌دهد. همان‌طور که بیان شد، نیاز است بین تعداد نمونه‌های آزمایشی و تعداد مشخصه‌های مستقل توازن باشد. به‌طورکلی اگر تعداد مشخصه‌های مستقل کم باشد، بهتر است تعداد نمونه‌های آموزشی مورد نیاز کم باشد و به‌همین‌ترتیب زمانی که تعداد مشخصه‌ها زیاد است، بهتر است تعداد نمونه‌های آموزشی مورد نیاز نیز زیاد باشد. باتوجه به توضیحات روش ارائه‌شده جهت اجرای نرم‌افزار دسته‌بندی بااستفاده از درخت تصمیم در فصل دوم، خلاصه کار به‌این صورت است:

- ورود داده‌ها؛
- ساخت مدل؛
- دریافت نتایج مدل‌سازی؛
- تولید قوانین.

خروجی اولیه درخت تصمیم شامل گره‌های به‌دست‌آمده به‌صورت نمودار ۱ است. نتایج درخت تصمیم به‌صورت جزیی و تفکیک‌شده در مدل دسته‌بندی در نمودار ۲ آورده شده است.

براساس این مدل نهایی، خروجی نرم‌افزار CTree، تعداد کل داده‌های آزمایشی و داده‌های آموزشی ارائه شده است. تعداد پیش‌بینی‌کننده‌ها به تعداد ورودی‌های نرم‌افزار هستند که برابر ۶ است. تحقیق شامل سه کلاس A، B و C به‌ترتیب بیمه عمر و پس‌انداز، بیمه حوادث خانواده و بیمه حوادث فردی است. کلاس حداکثر، A بوده و از این کلاس به‌عنوان پیش‌بینی‌کننده استفاده شده است که دسته‌بندی ۳۲ درصد خطای دسته‌بندی را دارد. به این معنا که مدل به میزان ۳۲ درصد کلاس را به اشتباه پیش‌بینی کرده است. باتوجه به اینکه درصد اشتباه پایین بوده است، رکوردها تا حدود ۷۰ درصد، کلاس‌های خروجی پیش‌بینی‌شده را پشتیبانی می‌کنند. معیار کلی پایین یا بالا بودن به‌طورمعمول ۵۰درصد محسوب می‌شود که در تحقیق‌های کاربردی درخت تصمیم نیز از همین ملاک استفاده شده است.

ماتریس نهایی شامل اطلاعات داده‌های آزمایشی و آموزشی است. بدین‌صورت که تعداد موجود در هریک از کلاس‌ها را مشخص نموده و در خانه‌های ماتریس قرار می‌دهد. محور افقی ماتریس، نشانگر کلاس پیش‌بینی‌کننده و محور عمودی، معرف کلاس صحیح هریک از رکوردهاست. مثلاً کلاس A به‌عنوان پیش‌بینی‌کننده دارای ۱۵۸۳ داده آموزشی است، کلاس B شامل ۵۷۵ رکورد صحیح در B و کلاس C

1. Absolute Relative Error
2. Mean Squared Error

شـامل تمام ۵۷۲ رکورد کلاس خـود و ۲ رکورد متعلق به کلاس B اسـت. در داده‌های آزمایشـی نیز کلاس A شـامل ۵۰۵ رکورد صحیح، کلاس B شامل ۱۷۰ رکورد صحیح در B و ۳ رکورد در کلاس C و کلاس C شـامل تمام ۱۹۶ رکورد صحیح اسـت. درنتیجـه تنها ۵ رکورد به‌طورکلی به اشتباه در کلاس اولیه خود حضور داشته‌اند و دیگـر رکوردها در خروجی صحیح خـود قرار دارند. به‌عبارت‌دیگـر خطـای دسـته‌بندی اشـتباه در داده‌های آموزشـی۰/۰۷ درصـد و در داده‌های آزمایشـی۰/۳۴ درصد اسـت کـه قابل توجـه نیسـتند. همچنین درخت تصمیم نیز شـامل ۳۳ گره، ۱۹ برگ و ۱۰ سـطح است. جـدول خلاصه قوانیـن، کیفیـت قوانین واحـد را ارائه داده اسـت که این کیفیت توسـط سه روش متفاوت زیر اندازه‌گیری می‌شود:

– **پشتیبانی**[1]: درصدی از داده‌های آموزشی که سمت چپ قانون برای آنها صدق می‌کند. اگر در یک مشاهده، قاعده سمت چپ[2] قانون صدق کند، می‌گوییم که قانون برای آن مشـاهده به‌کارمی‌رود. این شاخص بیان می‌کند که قانون مورد نظر تا چه حد قابل کاربرد است.

اطمینـان[3]: درصد رکوردهای آزمایشـی اسـت که قاعده سمت چپ در آنها صادق است و علاوه بر آن قاعده سمت راسـت[4] نیز در آنها صدق می‌کند. به‌عبارت‌دیگر، این شـاخص نشـان می‌دهد قانون برای چـه درصدی از مشاهداتی که قانون در آنها به‌کارمی‌رود، صدق می‌کند. این شاخص صحت قوانین را بررسی می‌کند.

– **دریافـت**[5]: چه درصـدی از رکوردهای مورد نظر به‌طور صحیح توسـط یک قانون دریافت می‌شـوند، در واقع نوعی انعکاس از ساختار مسئله است. اگر یک قانون با دریافت نزدیک به ۱۰۰ درصد وجود داشته باشد، به این معنی اسـت که در فضای پیش‌بینی‌کننده، تمام مشاهدات بـا این کلاس به‌طـور نزدیکی کنار هم قـرار گرفته‌اند و قانون توانسـته است آن بخش از فضای پیش‌بینی‌کننده را به‌خوبی دریافت کند.

برای قضاوت در رابطه با کیفیت قانون، هر سـه معیار بایـد درنظرگرفته‌شـوند. از طرفی نمی‌تـوان گفت یک قانـون به‌عنوان‌مثال بـا ۱۰۰ درصد اطمینان و پشـتیبانی۵ درصد، لزوماً ضعیف اسـت، چرا که ممکن اسـت قانون پیش‌بینی‌کننـده در واقـع تنهـا ۵ درصـد از زمان‌ها در داده‌های آموزشی رخ دهد و همان مقدار را قانون به‌طور مطلوبی دریافت کرده باشـد و کیفیت خوبی داشته باشد. کیفیت این قوانین به‌صورت بصری در نمودار ۳ ارائه‌شده که محور افقی آن، تعداد رکوردهای مشاهدات و محور عمودی، قوانین تولیدشده است.

هر نقطه در این نمودار نشانگر یک مشاهده و رنگ آن نشان‌دهنده کلاس صحیح آن است. رنگ شماره ۱ کلاس A، رنگ شماره ۳ کلاس B و رنگ شماره ۲ کلاس C اسـت. اولین خط از پایین قانون ۱ را نشان می‌دهد که این ردیف باتوجه به کلاس پیش‌بینی‌شـده توسـط این قانون رنگ خاصی می‌گیرد.

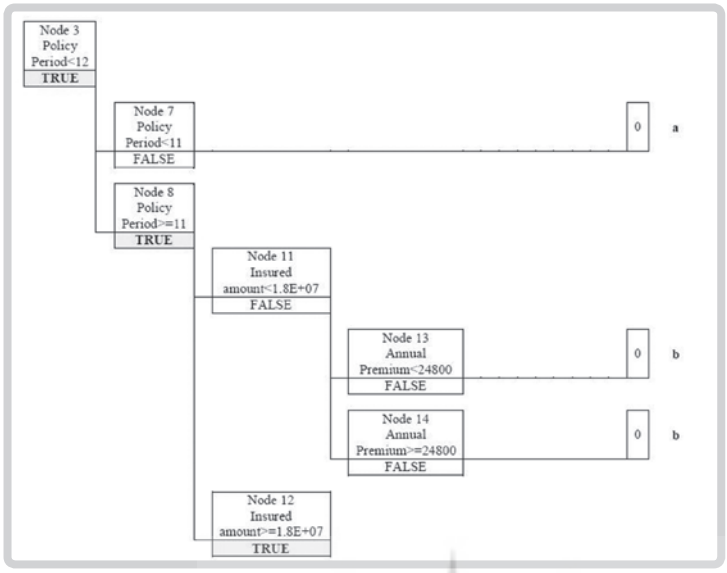
تعداد نقطه‌ها در یک ردیف، میزان پشتیبانی یک قانون را می‌رسـاند کـه در اینجا تنها قوانین ۷ تا ۱۰، پشـتیبانی کمی داشته یا به عبارتی کمتر رخ می‌دهند.

اگـر هریـک از نقـاط را در ردیـف قوانین بـا نقاط پایین‌ترین ردیف مقایسـه کنیم و رنگ یکسانی را داشته باشیم، قانون، کلاس را به‌درسـتی پیش‌بینی کرده است. در اینجا ۸ قانون باتوجه به شـکل، کلاس‌ها را به‌درسـتی پیش‌بینی کرده‌اند که قابل توجه است. درعین‌حال قانون ۱۴ و ۱۵ پیش‌بینی درستی نداشته و درصد اطمینان پایینی را کسب کرده‌اند.

به‌همین‌ترتیـب آن میزان از رنگ قانـون که به رنگ کلاس صحیح می‌باشد، در واقع میزان درصدی از کلاس صحیح اسـت که توسـط قانون موردنظر پیش‌بینی شده است. باتوجه به شـکل، قانون‌های اول تا چهارم دریافت بالایی داشته‌اند.

1. Support
2. Left Hand Side
3. Confidence
4. Right Hand Side
5. Capture

نمودار ۱. درخت تصمیم حاصل از تحلیل داده توسط نرم‌افزار CTree



نمودار ۲. خلاصه اطلاعات حاصل‌شده از تحلیل داده توسط نرم‌افزار CTree

Classification Tree Model

Number of Training observations	2732
Number of Test observations	874
Number of Predictors	6
Class Variable	Class
Number of Classes	3
Majority Class	a
% MissClassified if Majority Class is used as Predicted Class	32%

Tree Information

Total Number of Nodes	33
Number of Leaf Nodes	19
Number of Levels	10

% Missclasssified

On Training Data	0.07%
On Test Data	0.34%

Time Taken

Data Processing	56 Sec
Tree Growing	3 Min : 36 Sec
Tree Pruning	0 Sec
Tree Drawing	1 Sec
Classification using final tree	1 Min : 5 Sec
Rule Generation	24 Sec
Total	6 Min : 3 Sec

Confusion Matrix

Training Data

	Predicted Class			
True Class	a	b	c	
a	1583			1583
b		575		575
c		2	572	574
	1583	577	572	2732

Test Data

	Predicted Class			
True Class	a	b	c	
a	505			505
b		170	3	173
c			196	196
	505	170	199	874

۷. فرضیهٔ شبکه مصنوعی

بااستفاده از شبکه‌های عصبی مصنوعی، مشتریان بالقوهٔ محصولات جدید بیمه‌ای و خدمات متناسب آنها شناسایی می‌شوند. بااستفاده از روش شبکه‌های عصبی مصنوعی، مدل دسته‌بندی مشتریان ارائه شده و پس از اطمینان از صحت نتایج می‌توان براساس آن، مشخصه‌های مشتریان جدید را وارد مدل کرده و درنهایت دسته مورد نظر مشتری را پیش‌بینی کرد. آزمون شبکه‌های عصبی ابتدا با درنظرگرفتن داده‌های آموزشی و آزمایشی و بررسی انطباق آنها در طول دورهٔ آموزشی شبکه عصبی مصنوعی

نمودار ۳. نمایش بصری کیفیت داده‌ها توسط نرم‌افزار CTree

نمودار ۴. نمایش اطلاعات شبکه عصبی حاصل‌شده توسط نرم‌افزار R

Neural Network Model for Classification

% MissClass.(Training)	14.51%	% MissClass.(Validation)	13.47%

Number of Hidden Layers	2			
Layer Sizes	11	2	2	3
True Output (if available)				
Model Output	A			

نمودار ۵. نمایش داده‌های تبدیل‌شده برای شبکه عصبی با استفاده از نرم‌افزار R

Transformed Input		Bias	Income level.low	Income level.average	Income level.very low	Income level.high	Income level.very high	Gender.male
	Hdn1_bias	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Hdn1_Nrn1	-15.5528	-3.1816	-2.0158	-7.7086	-1.2370	-1.7267	-6.6269
	Hdn1_Nrn2	-3.8321	8.3451	8.3627	-35.8282	7.3738	8.1817	-2.6685
		1.0000	#N/A	#N/A				
	Hdn2_bias	0.0000	0.0000	0.0000	0.0000			
	Hdn2_Nrn1	-4.1095	-1.4429	10.4773	#N/A			
	Hdn2_Nrn2	2.6959	-5.1714	7.6796	#N/A			

و سپس مقایسه نتایج آن با نتایج روش درخت تصمیم انجام می‌شود.

۸. تجزیه‌وتحلیل نتایج شبکه‌های عصبی

همان‌طور که بیان شد، اجزای یک شبکه عصبی عبارت‌اند از:

– ورودی‌ها: ورودی‌ها می‌توانند خروجی سایر لایه‌ها بوده یا این‌که به حالت خام در اولین لایه و به‌صورت‌های داده‌های عددی و رقمی، متون ادبی، فنی و تصویر یا شکل باشند.

– وزن‌ها: میزان تأثیر ورودی بر خروجی توسط وزن اندازه‌گیری می‌شود.

– تابع جمع: خروجی مسئله را تا حدودی مشخص می‌کند.

– تابع تبدیل: این تابع توسط طراح مسئله انتخاب می‌گردد و براساس انتخاب الگوریتم یادگیری[1]، پارامترها تنظیم می‌گردند.

– خروجی: منظور از خروجی، پاسخ مسئله است.

برای تحلیل شبکه‌های عصبی، ابزارهای گوناگونی وجود دارد که ازنظر قدرت و امکانات متفاوت‌اند. در اینجا ما از نرم‌افزار R استفاده می‌کنیم که گزارش‌های بسیار کارایی از تحلیل داده‌ها در اختیار ما قرار می‌دهد و نیز قابلیت‌های فراوانی دارد. براساس ورودی‌های تعیین‌شده و معیارهای انتخاب‌شده در معماری شبکه عصبی و اجرای نرم‌افزار، مدل دسته‌بندی ساخته‌شده بااستفاده از شبکه‌های عصبی به‌صورت نمودار ۴ است.

همان‌طور که مشاهده می‌شود درصد خطای دسته‌بندی نادرست ۱۴ و ۱۳ درصد در داده‌های آموزشی و آزمایشی است که مطابق با ۱۴ درصد خطای مشابه در خروجی نهایی درخت تصمیم است.

نمودار ۶. نمایش شیب و عرض از مبدأ برای داده‌های آموزشی و آزمایشی بااستفاده از نرم‌افزار R

1. Learning Algorithm

در این مسئله، داده‌های رکوردهای مختلف را به شبکه داده و نام گروه هر رکورد را به‌عنوان خروجی مشخص می‌کنیم، پس از آموزش، شبکه توانسته است با دریافت داده‌های مربوط‌به نمونه‌های جدید مشخص کند که این نمونه به کدام دسته متعلق است. حال بخشی از داده‌های تبدیل‌شده در نرم‌افزار با دو لایه پنهان در شبکه به‌صورت نمودار ۵ است:

وجود Bias در تابع تبدیل که به‌صورت تابع مقابل است:

$$y=f(n)=f(wx+b)$$

w باید به نوعی کارایی داشته باشد که با ضرب در x مقدار را مشخص کند، اما در واقعیت نمی‌تواند، به‌همین‌دلیل مقداری چولکی دارد که باید به آن اضافه شود. مانند ترازویی که در یک مقدار ثابت مشکل دارد و باید به یک طرف آن وزنه‌ایی را افزود. در این خروجی میزان b صفر است و وزن‌ها تا حد زیادی نزدیک به صحیح انتخاب شده‌اند. اعداد منفی در این جدول نشان‌دهنده حرکت در جهت کاهش شیب است که در نمودار ۵ قابل مشاهده است.

شیب و عرض از مبدأ تابع تبدیل به میزان تمام داده‌های آموزشی و آزمایشی در نمودار ۶ فراهم آمده است:

باتوجه با اینکه دورهٔ آموزشی در معماری شبکه عصبی این تحقیق ۵۰ انتخاب شده است، محور افقی، شامل ۵۰ آموزش شبکه است و محور عمودی، میزان خطای دسته‌بندی نادرست را در هرس کل نشان می‌دهد. همان‌طور که در نمودار مشاهده می‌شود اختلاف بسیار ناچیزی میان دوره‌های آموزش داده‌های آزمایشی و داده‌های آموزشی وجود داشته که تا حد زیادی براساس گفته‌های قبلی در روش تحقیق، اعتبار تحقیق را تأیید می‌کند. برای بررسی دقیق‌تر می‌توان درصدها را از روی جدول با مقادیر عددی مقایسه کرد. همچنین هرچه شبکه بیشتر آموزش می‌بیند، میزان خطای دسته‌بندی اشتباه کاهش می‌یابد، به‌طوری‌که در دوره‌های پایانی به ۱۳ تا ۱۴درصد رسیده و شبکه موفق شده است تا تابع هدف را بیاموزد.

۹. نتیجه‌گیری از خروجی‌های درخت تصمیم و شبکه‌های عصبی

همان‌طورکه خروجی‌های نرم‌افزار درخت تصمیم و شبکه‌های عصبی یکدیگر را تأیید کردند و در راستای یکدیگر قرار گرفتند، نتایجی در رابطه با مشخصه‌های این پژوهش (متغیرهای ورودی) و بررسی قوانین تصمیم‌گیری تولیدشده به‌دست‌آمده که توسط خبرگان این صنعت نیز تأیید شده است. نتایج به‌دست‌آمده از هر دو روش به‌کارگرفته شده، طی جلسه‌ای به مدیر بیمه درمان و کارشناس‌ارشد بیمه عمر و درمان شرکت بیمه موردنظر ارائه و نظرات بیان‌شده مستند گردید.

۱۰. نتیجه‌گیری

نتایج به‌دست‌آمده از این تحقیق بیان می‌کنند که فضای خالی بین تئوری و اجرا هنوز باید کمتر شود، به‌ویژه در ترغیب شرکت‌های بیمه به‌منظور تغییردادن مفاهیم استفاده از داده‌کاوی در جایی که اطلاعات امنیتی نیز درگیر شده‌اند. این تحقیق به‌این‌دلیل که شواهد کاربردی موجب جلب توجه به نقش مهم داده‌کاوی به‌کاررفته در صنعت بیمه می‌شود، بسیار با ارزش است. همچنین ممکن است مسائل دیگری هم از قبیل درنظرگرفتن دیگر مشخصه‌های مشتریان (مانند شغل، نوع زندگی، تحصیلات و ...) و طراحی‌کردن تمام مراحل برای یک توسعه پیچیده (مانند ایجاد ماجول[1] و به‌روزکردن نتایج کشف‌شده در هر زمان) مانند توسعه نرم‌افزاری مهم باشند.

با اهمیت یافتن بازاریابی فردی جهت پاسخ به مشتری، داده‌کاوی به‌صورت مؤثر در دسته‌بندی مشتریان هدف استفاده شده است. بازاریابی فردی بر این مطلب تأکید می‌کند که استراتژی بازاریابی باید بر مشخصات فردی مشتری متمرکز شود. مطابق با همین موضوع، مدل تجربه‌گرا که بیشتر شرکت‌های بیمه از آن استفاده می‌کنند، به‌دلیل اینکه باید بسیاری از مشتریان و همچنین بسیاری از خصوصیات آنها درنظرگرفته‌شود، دیگر برای اجرای امور پیچیده در دنیای امروز کارآمد نخواهد بود.

1. Module

با رشد سریع در حجم داده‌هایی که سیستم‌های اطلاعاتی جمع‌آوری می‌کنند، بازاریاب‌های بیشتری تمایل به استفاده از ابزارهای پشتیبانی تصمیم‌گیری مبتنی‌بر داده برای بالابردن و بهبود کارایی و مؤثربودن تصمیمات بازاریابی خود خواهند داشت.

مشتریان و میزان ریسک آنان، عامل اصلی در سودآوری شرکت‌های بیمه است و شناخت آنها مهم‌ترین مسئله در این مورد است. استفاده از ابزارهای دسته‌بندی به‌منظور جداسازی مشتریان و شناخت آنها بسیار مفید است. داده‌کاوی، ابزار بسیار هوشمندی به‌منظور دسته‌بندی مشتریان است. بااستفاده از دسته‌بندی مشتریان می‌توان مشتریان بیمه را بهتر شناخت و براساس گروهی که مشتری باتوجه به خصیصه‌هایش در آن قرار گرفته است سیاست‌گذاری‌های مناسبی برای آنها درنظر گرفت. در این پژوهش با استفاده از روش‌های درخت تصمیم و شبکه عصبی، مشتریان بیمه را دسته‌بندی کرده‌ایم و سپس نتایج این دو روش را مقایسه کرده‌ایم. همان‌طور که پیشتر ذکر شد نتایج به‌دست‌آمده از این دو روش یکدیگر را تأیید می‌کنند که این امر بر صحت نتایج می‌افزاید. همچنین از نظرات کارشناسان برای ارزیابی نتایج استفاده کرده‌ایم. نظرات کارشناسان نیز نتایج به‌دست‌آمده از روش‌های داده‌کاوی را تأیید می‌کند. لذا با استفاده از داده‌کاوی توانسته‌ایم مشتریان بیمه را در دسته‌هایی قرار داده و با توجه به ویژگی‌های این دسته‌ها شناخت بهتری از مشتریان به‌دست‌آوریم. همچنین می‌توان مشتریان آتی را باتوجه به ویژگی‌هایشان در یکی از این دسته‌ها قرار دارد. دسته‌ای که مشتری در آن قرار گرفته، نشان‌دهنده خصایص مشتری و میزان ریسک آن برای شرکت بیمه است. بدین‌صورت می‌توان در مورد میزان ریسک مشتری برای شرکت بیمه پیش‌بینی داشته باشیم که این امر می‌تواند کمک شایانی در سودآوری شرکت‌های بیمه داشته باشد.

۱۱. پیشنهاد برای تحقیقات آتی

مباحث این تحقیق عمدتاً بر سه نکته متمرکز شده است که شامل تجربه کاربرد داده‌کاوی در دسته‌بندی مشتریان هدف برای صنعت بیمه، قابلیت استفاده از داده‌کاوی در پشتیبانی از عملیات تصمیم‌گیری و جهشی از مباحث تئوری به سمت نتایج عملی است (Adda & Missoui, 2007). فناوری داده‌کاوی عمدتاً یک نمونه از پایگاه داده است که به‌عنوان ابزار مدیریتی در پشتیبانی تصمیم‌گیری استفاده شده است، اگرچه سبک تصمیم‌گیری تجربه‌گرا در حقیقت برای یک مدت طولانی به‌کاررفته و تا نقش مهمی در حمایت از مدیریت بازی کند، ادامه پیدا خواهد کرد. همچنین با افزایش اطلاعات، سبک پشتیبانی تصمیم به‌طور تصاعدی از تجربه‌گرایی به اطلاعات‌گرایی تغییر کرده است.

هرچند که این تحقیق تنها در سطح یکی از واحدهای شرکت بوده، اما دانش کشف‌شده، توجه زیادی را در شرکت مورد نظر در جلسه با مدیر بیمه عمر و درمان جذب کرده است. این امر بدین دلیل است که نتایج نشان داده‌اند در شرکت مورد نظر، هرگز به رابطه بین محصولات و مشتریان توجه نکرده‌اند. قابلیت استفاده از دانش کشف‌شده و نمایش‌داده‌شده در این تحقیق، نشان داده است که در این مرحله، داده‌کاوی می‌تواند یک راه ممکن به سمت افزایش دسته‌بندی مشتریان هدف باشد. برای تأثیر دانش کشف‌شده در مرحله دوم، به‌منظور جمع‌آوری و مقایسه داده‌ها برای شرکت موردنظر، زمان زیادی مورد نیاز است؛ برای مثال، آیا دانش کشف‌شده به‌طور قابل ملاحظه‌ای به افزایش پاسخ‌های مشتری برای سال بعد (دو، سه یا چهار سال بعد) کمک می‌کند، این بخش می‌تواند یکی از مهم‌ترین تحقیقات آینده باشد.

منابع

1. Adda, M & Missaoui, R 2007, 'Relation rule mining', *International Journal of Parallel, Emergent and Distributed Systems*, vol.22, no.6, pp. 439- 49.

2. Belhadjali, M & Whaley, GL 2004, 'A data mining approach to neural network training', *Information Management & Computer Security* , vol. 12, no. 1, pp. 117-24.

3. Berson, A; Smith, S & Therling, K 2001, *Building data mining applications for CRM*, Mc GrawHill.

4. Brockett, P & Xiaohua, X 1999, 'Operations research in insurance, A review', *Journal of Mathematics and Economics*, vol. 19, no. 2, p. 154.

5. Chalmeta, R 2005, 'Methodology for customer relationship management', *The Journal of System and Software*, vol. 15, pp. 192-201.

6. Chan, C & Lewis, B 2002, 'A basic primer on data mining', *Information Systems Management*, vol. 19, no. 4, pp. 56-60.

7. Chen, YH & Su, CT 2006, 'A kanoCKM model for customer knowledge discovery,' *Total Quality Management & Business Excellence*, vol. 17, no.5, pp.589-608.

8. Clifton, C & Thuraisingham, B 2001, 'Emerging standards for data mining', *Computer Standards and Interfaces*, vol. 23, pp. 187-93.

9. Gupta, GK 2006, *Data mining with case studies*, New Delhi, Prentice Hall India.

10. Sivanandam, SN & Sumathi, S 2006, *Neural networks using MATLAB 6.0*, McGrawHill.

11. SY, BK 2001, 'Information statistical pattern based approach for data mining', *Journal of Statistical Computation and Simulation*, vol. 69, no. 2.

12. Tan, KC & Yu, Q 2006, 'A coevolutionary algorithm for rules discovery in data mining', *International Journal of Systems Science*, vol. 12, no. 37, pp. 835-64.

13. Wu, Ch & Su, Y 2005, 'Targeting customers via discovery knowledge for the insurance industry, '*Expert Systems with Applications*, vol. 29, pp. 291 -99.