فصلنامه صنعت بیمه/سال بیست و پنجم/شماره ۴/زمستان ۱۳۸۹/شماره مسلسل ۱۰۰/صفحات ۱۳۱ تا ۱۵۲

به‌کارگیری یک درخت رگرسیونی هرس‌شده جهت انتخاب ویژگی‌های مؤثر بر پیش‌بینی رفتار خرید مشتریان در صنعت بیمه

علی‌رضا سروش[۱]
دکتر اردشیر بحرینی‌نژاد[۲]
ریحانه قاسم اصفهانی[۳]

تاریخ دریافت مقاله: ۱۳۸۹/۰۶/۱۵
تاریخ پذیرش مقاله: ۱۳۸۹/۱۲/۰۹

چکیده

از مهم‌ترین ابعاد مدیریت ارتباط با مشتری، کشف الگوی رفتاری خرید مشتری است؛ سازمان می‌تواند با تعریف استراتژی‌های بازاریابی دقیق‌تر جهت جذب مشتریان مشابه اقدام کند. لذا، انتخاب ویژگی‌های مؤثر بر روی الگوی رفتاری خرید مشتریان با اهمیت است. زیرا استفاده از ویژگی‌های نامرتبط منجر می‌شود سازمان ناخواسته هزینه‌های بسیاری را صرف افرادی کند که احتمال خرید آنها ناچیز است. این تحقیق، یک رویه جدید برای شناسایی ویژگی‌های مؤثر جهت بهبود پیش‌بینی مشتریان ارائه می‌دهد. همچنین یک سیستم تشخیص مشتری مبتنی‌بر تکنیک غیرخطی درخت تصمیم رگرسیونی جهت انتخاب ویژگی‌ها و پیش‌بینی رفتار خرید مشتریان طی دو مرحله توسعه داده می‌شود. چون شناسایی مشتریان از موضوعات حیاتی صنعت بیمه است، از مجموعه داده‌های یک شرکت بیمه هلندی برای پیش‌بینی خرید یکی از محصولات آن استفاده شده است. نتایج نشان می‌دهد که انتخاب بهترین زیرمجموعه از ویژگی‌ها با به‌کارگیری یک درخت رگرسیونی هرس‌شده علاوه‌بر کاهش معنادار پیچیدگی محاسبات می‌تواند بهبود قابل توجهی را در نتایج پیش‌بینی ایجاد کند.

واژگان کلیدی: *انتخاب ویژگی، مدیریت ارتباط با مشتری، درخت تصمیم رگرسیونی، پیش‌بینی رفتار خرید، تشخیص مشتری*

---

۱. دکتری مهندسی صنایع، دانشگاه تربیت مدرس، دانشکده فنی و مهندسی (Email: A.soroush@Modares.ac.ir)
۲. عضو هیئت علمی دانشگاه تربیت مدرس، دانشکده فنی و مهندسی (نویسنده مسئول)
(Email: Bahreininejad@Modares.ac.ir)
۳. کارشناس‌ارشد مهندسی کامپیوتر، دانشگاه علم و صنعت، دانشکده مهندسی کامپیوتر
(Email: R.Esfehani@Gmail.com)

۱. مقدمه

مسئله انتخاب ویژگی‌ها، مسئله‌ای مستقل در نظریه تشخیص الگو بوده و تاکنون حل نشده است. فرآیند انتخاب ویژگی‌ها به‌عنوان مسئله‌ای از بهینه‌سازی ترکیبی کلی در یادگیری ماشین شناخته می‌شود که تعداد ویژگی‌ها را کاهش داده و داده‌های غیرمرتبط را حذف می‌کند. هدف اصلی انتخاب ویژگی، شناسایی زیرمجموعه‌ای از ویژگی‌ها است که تأثیر بیشتری بر روی یک متغیر پاسخ معلوم دارد و بیشترین دقت را ارائه می‌کند (Kohavi & John, 1997). کشف زیرمجموعه بهینه‌ای از ویژگی‌ها معمولاً مشکل بوده و نشان داده شده است که بسیاری از مسائل مرتبط نامعین بسیار زمان‌بر[1] شناخته می‌شوند (Blum & Rivest, 1992). پیاده‌سازی مناسب انتخاب ویژگی‌ها نه تنها اطلاعات مهمی را برای پیشگویی فراهم می‌کند، بلکه تلاش‌های مورد نیاز برای تحلیل داده‌های چندبعدی را کاهش می‌دهد.

انتخاب ویژگی‌ها موفقیت‌های بسیاری را در کاربردهای دنیای واقعی داشته است، زیرا غالباً می‌تواند به‌طور قابل توجهی ابعاد را با به‌کارگیری الگوریت‌های داده‌کاوی جهت کار بر روی داده‌ها با ابعاد بزرگ کاهش دهد. به‌همین‌دلیل، در سال‌های اخیر، مدیریت ارتباط با مشتری[2] یکی از زمینه‌های تحقیقاتی بوده که در انتخاب ویژگی‌ها به‌کاررفته‌است(Ng & Liu, 2000). امروزه، توجه به CRM به‌واسطه افزایش میزان رقابت میان شرکت‌ها حیاتی‌تر شده است. CRM، وسیله عمده‌ای است که می‌تواند به کسب‌وکارها کمک کند که تقاضاهای مختلف مشتریان را شناسایی نموده و از این طریق مزیت رقابتی به‌دست‌آورد (Anderson, 2002). به‌این‌دلیل، تأثیرات انتخاب ویژگی بر روی آماده‌سازی یک سیستم CRM مطالعه شده است(Kim, 2006).

1. Non-deterministic Polymial-time hard (NP- hard)
2. Costumer Relationship Management (CRM)

به‌صورت سنتی، انتخاب بهینه مشتریان هدف از مهم‌ترین عوامل برای یک سیستم CRM درنظرگرفته‌شده‌است. ازاین‌رو، شرکت‌ها تلاش می‌کنند که مدل‌های پیش‌بینی را به‌صورت دقیق توسعه دهند که شناسایی کنند که کدام مشتریان با بیشترین احتمال خرید می‌کنند. این مدل‌ها بعنوان یک سیستم تشخیص مشتری توصیف می‌شوند. تعداد ویژگی‌های در دسترس طراح یک سیستم تشخیص مشتری معمولاً بسیار زیاد است. از جمله دلایل کاهش ویژگی‌های در دسترس می‌توان به کاهش پیچیدگی محاسباتی، حذف ویژگی‌های همبسته بدون عایدی مازاد، ارتقای قابلیت تعمیم مدل تشخیص‌دهنده اشاره کرد.

هدف این تحقیق ارائه یک رویه مفید برای طراحی یک سیستم تشخیص مشتری است. به‌طوری‌که، در این مقاله، امکان بهبود دقت پیش‌بینی مشتری از طریق انتخاب ویژگی‌های ورودی مرتبط تحلیل می‌شود. این مقاله، از یک درخت تصمیم هرس‌شده به‌عنوان ابزاری غیرخطی جهت انتخاب ویژگی‌های ورودی به سیستم استفاده می‌شود. سپس، از ویژگی‌های منتخب جهت پیش‌بینی مشتریان آتی شرکت استفاده می‌شود. به‌طور خاص، این مقاله از یک درخت تصمیم در حالت بهینه برای انتخاب ویژگی‌ها استفاده کرده و چگونگی تأثیرگذاری انتخاب ویژگی بر روی عملکرد مدل پیش‌بینی را تحلیل می‌کند.

ادامه مقاله به‌این‌صورت سازماندهی می‌شود؛ در بخش بعد، ادبیات موضوع در خصوص به‌کارگیری تکنیک‌های انتخاب ویژگی پیش از پیش‌بینی مشتریان تحلیل می‌شود. در بخش سوم، مدل درخت رگرسیونی هرس‌شده توصیف می‌شود. در بخش چهارم، موردکاوی بر روی داده‌های یک شرکت بیمه هلندی انجام شده و در بخش پنجم، نتیجه‌گیری ارائه می‌شود.

۲. مروری بر ادبیات مرتبط

در سال‌های اخیر، چندین الگوریتم برای انتخاب ویژگی‌ها در حوزه CRM به‌کاررفته و تعدادی مطالعه مقایسه‌ای نیز انجام شده است. به‌طوری‌که، آندرسون[1] از یک رویه آماری ساده یعنی انتخاب ویژگی رو به جلو مبتنی‌بر امتیاز مربع خی دو استفاده نمود که با مجموعه خالی از متغیرها شروع نموده و به تدریج متغیرهایی را اضافه می‌کند که بیشترین بهبود عملکرد را دارند. ان‌جی و لیو[2]، انتخاب ویژگی را با اجرای یک الگوریتم استقرایی بر روی مجموعه داده‌ها انجام می‌دهند. همچنین، آنها رویکرد درخت تصمیمی را برای انتخاب ویژگی پیشنهاد دادند. یان و همکارانش[3] یک رویه منحنی مشخصه عملیاتی دریافت‌کننده[4] را برای انتخاب ویژگی‌ها پیشنهاد دادند و بیان کردند که منطقه زیر منحنی را می‌توان برای انتخاب ویژگی استفاده کرد. کیم و استریت[5]، رویکرد جدیدی را برای هدف‌گذاری مشتریان در بازاریابی پایگاه داده ارائه نمودند. یک الگوریتم ژنتیک[6] استاندارد برای جستجو در میان ترکیبات ممکن ویژگی‌ها استفاده می‌شود. ویژگی‌های ورودی انتخابی توسط GA جهت آموزش شبکه‌های عصبی استفاده می‌شود. همچنین، یو و چو[7]، یک روش ایجاد تجمعی مبتنی‌بر GA برمبنای مکانیزم انتخاب زیرمجموعه‌ای از ویژگی‌های پوشه‌ای[8] پیشنهاد کردند که براساس اطلاعات گذشته خرید، مقداری را که هر مشتری خرج خواهد کرد، پیش‌بینی کنند. آن و همکارانش[9]، یک سیستم استدلال مبتنی‌بر مورد را همراه با تکنیک کاهش دو بعدی

1. Anderson, 2002
2. Ng & Liu, 2000
3. Yan et al, 2004
4. Receiver Operating Characteristic (ROC)
5. Kim & Street, 2004
6. Genetic Algoritm (GA)
7. Yu & Cho, 2006
8. Wrapper
9. Ahn et al, 2007

برای رضایت مشتریان پیشنهاد دادند که رفتار خرید مشتریان را برای یک محصول خاص با استفاده از مشخصه‌های آماری آنها پیشگویی می‌کند. باکینکز و همکارانش[۱]، یک مدل رگرسیونی خطی چندگانه را برای جلوگیری از تطبیق بیش از حد به‌عنوان یک رویه انتخاب ویژگی به‌کارگرفتند و وفاداری رفتاری مشتری را با استفاده از پایگاه داده تراکنش‌ها پیشگویی کردند. یان و چانگروی[۲] ، الگوریتم افراز تودرتو[۳] و روش ذوب شبیه‌سازی‌شده[۴] را جهت انتخاب ویژگی‌ها برای تشخیص مشتری ترکیب کردند. تی‌سنگ و هوآنگ[۵] ، نظریه مجموعه زبر[۶] را برای انتخاب ویژگی‌ها در CRM به‌کاربردند و از داده‌های گذشته خرید یک سیستم بازی تصویری برای پیشگویی رفتار خرید مشتری استفاده کردند. همچنین، هانگ[۷]، از نظریه مجموعه زبر برای کاهش ویژگی‌های پایگاه داده خانواده بسته‌بندی آی‌سی[۸] (مدار مجتمع)، جهت بهبود دقت تعمیم طبقه‌بندی استفاده کرد و از شبکه‌های عصبی مصنوعی برای ایجاد یک مدل طبقه‌بندی‌کننده کارا در خصوص نوع بسته‌بندی IC استفاده کرد. لمسن و ووف[۹]، یک مدل مرجع سلسله مراتبی را برای ماشین بردار پشتیبان[۱۰] مبتنی‌بر طبقه‌بندی در کاربردهای CRM دنیای واقعی پیشنهاد کردند که در آن حذف ویژگی بازگشتی به‌عنوان یک رویه حذف رو به عقب برای رتبه‌بندی ویژگی‌ها مبتنی‌بر SVM پیشنهاد شده است. مالدونادو و وبر[۱۱]، یک الگوریتم برای انتخاب ویژگی با

1. Buckinx et al, 2007
2. Yan & Changrui, 2007
3. Nested Partition
4. Simulated Annealing (SA)
5. Tseng & Huang, 2007
6. Rough Set
7. Hung, 2009
8. Integrated Circuit (IC)
9. Lessmann & VoB, 2009
10. Support Vector Machine (SVM)
11. Maldonado & Weber, 2009

استفاده از SVM مبتنی‌بر یک انتخاب روبه‌عقب متوالی معرفی نموده و رویکرد خود را با سایر الگوریتم‌ها همچون روش فیلتر یا SVM حذف ویژگی بازگشتی مقایسه کرده‌اند.

با بررسی کلیه روش‌ها و تکنیک‌های انتخاب ویژگی می‌توان متوجه شد که همه آنها چند هدف مشترک شامل حداکثرسازی دقت هم‌زمان با حداقل‌سازی تعداد ویژگی‌ها، بهبود دقت با حذف ویژگی‌های غیرمرتبط، کاهش پیچیدگی داده‌ها و حذف محاسبات، کاهش مقدار داده‌ها برای مرحله یادگیری را دنبال می‌کنند. به‌علاوه، معمولاً یک الگوریتم انتخاب ویژگی به عناصری شامل یک معیار ارزیابی ویژگی، یک رویه جستجو، یک معیار توقف یا استراتژی انتخاب مدل نیاز دارد. در میان روش‌های ارائه‌شده، نقطه ضعف روش ROC آن است که روابط بین ویژگی‌ها را به‌صورت خطی تحلیل کرده و همچون SVM تنها آنها را رتبه‌بندی می‌کند. همچنین، GA نیز به زمان بسیار زیادی جهت تعیین ویژگی‌های بهینه نیاز دارد و SA در نقاط موضعی می‌افتد. ازآنجایی‌که، نویسندگان به دنبال ابزاری بوده‌اند که قادر به شناسایی روابط غیرخطی باشد و با سرعت مناسبی ویژگی‌های مؤثر بر رفتار مشتریان را شناسایی کند، از درخت رگرسیونی استفاده می‌شود که هر دو این قابلیت‌ها را دارد.

۳. درخت رگرسیونی هرس‌شده

دسته بزرگی از طبقه‌بندی کننده‌های غیرخطی به درخت‌های تصمیم معروف است. یک درخت تصمیم یا درخت رگرسیونی، نمونه‌ای از یک فرآیند تصمیم چند مرحله‌ای است که دسته‌ها به‌صورت متوالی رد می‌شوند تا درنهایت به یک دسته قابل قبول برسیم. سرانجام، فضای ویژگی به روشی متوالی به دو ناحیه منحصربه‌فرد برطبق دسته‌ها تقسیم می‌شود. با ورود یک بردار ویژگی، جستجو برای ناحیه‌ای که بردار ویژگی از طریق تصمیمات متوالی تخصیص داده خواهد شد همراه با مسیری از گره‌های یک درخت ساخته‌شده مناسب انجام می‌شود. چنین طرح‌هایی زمانی برتری

دارند که تعداد زیادی دسته دربرگرفته‌می‌شود. عمومی‌ترین شیوه میان درخت‌های تصمیم آنهایی هستند که فضا را به مافوق چهارگوشه‌هایی[1] با کناره‌های موازی با محورها تقسیم می‌کند. توالی تصمیمات برای ویژگی‌های منفرد به‌کارمی‌رود و پرسش‌هایی که پاسخ داده می‌شوند به شکل «آیا ویژگی $x_i \leq \alpha$? است؟» که α یک مقدار آستانه است. چنین درخت‌هایی به‌عنوان «درخت‌های رگرسیونی دوتایی معمولی[2](OBCTs)» مشهور هستند. همچنین انواع دیگر درخت‌ها هستند که امکان تقسیم فضا را به سلول‌های چندوجهی محدب یا قسمت‌های کروی[3] فراهم می‌کنند (Webb, 2002).

۱-۳. عناصر درخت تصمیم دوتایی

به‌طورکلی، به‌منظور توسعه یک درخت تصمیم دوتایی، طراح باید این عناصر طراحی را در مرحله آموزش درنظر بگیرد:

– در هر گره، در خصوص مجموعه سؤالاتی که پرسیده می‌شود، باید تصمیم‌گیری انجام شود. هر سؤال به یک دودویی خاص مربوط می‌شود که به دو گره همزاد تبدیل می‌شود. هر گره، t، با یک زیرمجموعه خاص X_t از مجموعه آموزشی X همبسته می‌شود. دو نیم‌سازی یک گره معادل با دو نیم‌سازی زیرمجموعه X_t به دو زیرمجموعه همزاد گسسته X_{tY} و X_{tN}است. اولین جزء بردار در X_t به پاسخ «بله» و دومی به «خیر» مربوط می‌شود. اولین گره (ریشه) درخت با مجموعه آموزشی X همبسته می‌شود. برای هر دو نیمه، این رابطه صحیح است:

$$X_{tY} \cap X_{tN} = \varphi \qquad (۱)$$

$$X_{tY} \cup X_{tN} = X_t \qquad (۲)$$

1. Hyperrectangles
2. Convex Polyhedral Cells
3. Pieces of Spheres

– یک معیار دو نیم‌سازی باید جهت تعیین بهترین دو نیم از مجموعه کاندیداهای انتخابی، اتخاذ شود.

– یک قاعده توقف دو نیم‌سازی لازم است که رشد درخت را کنترل کند و یک گره به‌عنوان یک گره پایانی شناسایی شود.

– یک قاعده مورد نیاز است که هر گره پایانی را به یک دسته خاص تخصیص دهد (Theodoridis & Koutroumbas, 2006).

به‌طورکلی، به‌جای استفاده از کلیه ویژگی‌ها به‌منظور تصمیم‌گیری از زیرمجموعه‌های مختلف ویژگی‌ها در سطوح مختلف درخت استفاده می‌شود.

۲-۳. مجموعه سؤالات

برای درخت‌های از نوع OBCT سؤالات به این شکل است: « $x_i \le \alpha$? است؟». برای هر ویژگی، هر مقدار ممکن آستانه دو نیمه خاصی از زیرمجموعه X_t را تعریف می‌کند. ازاین‌رو، ازنظر تئوری مجموعه‌ای نامحدود از سؤالات باید پرسیده شود، اگر α در یک فاصله $Y_\alpha \subset R$ تغییر می‌کند. از نظر عملی، تنها مجموعه محدودی از سؤالات را می‌توان درنظرگرفت؛ برای مثال، ازآنجایی‌که تعداد نقاط آموزشی، N، در X محدود است، هر ویژگی x_k، $k = ۱,...,l$، حداکثر می‌تواند $N_t \le N$ مقدار متفاوت بگیرد که N_t عدد اصلی زیرمجموعه $X_t \subseteq X$ است. ازاین‌رو، برای ویژگی x_k، فرد می‌تواند از α_{kn}، $(N_{tk} \le N_t)$ $n = ۱,۲,...,N_{tk}$ استفاده کند که α_{kn} میانه گرفته‌شده بین مقادیر مجزای پیاپی x_k در زیرمجموعه آموزشی X_t است. بطور مشابه باید برای کلیه ویژگی‌ها تکرار شود. ازاین‌رو، در چنین موردی کل تعداد سؤالات کاندید برابر $\sum_{k=۱}^{l} N_{tk}$ است. هرچند، تنها یکی از آنها باید جهت ایجاد دو نیمه در گره فعلی، t، درخت انتخاب شود. موردی انتخاب خواهد شد که به بهترین دو نیمه زیرمجموعه همبسته X_t منجر شود. درباره بهترین دو نیمه براساس معیار دو نیم‌سازی تصمیم‌گیری می‌شود (Theodoridis & Koutroumbas, 2006).

۳-۳. معیار دو نیم‌سازی[1]

یک قاعده دو نیم‌سازی، یک دستورالعمل است برای تصمیم‌گیری در خصوص اینکه کدام متغیر یا ترکیبی از متغیرها باید در یک گره جهت تقسیم نمونه‌ها به زیرگروه‌ها استفاده شود و اینکه چه آستانه‌ای برای آن متغیر درنظرگرفته‌شود. یک دو نیم‌سازی از یک شرط بر روی مختصات یک بردار تشکیل شده است. سؤال این است که چگونه داده‌هایی را که در زیرفضای $u(t)$ در گره t قرار دارند، چند دسته کنیم (Webb, 2002).

هر دو نیمه از یک گره t، دو گره اولاد تولید می‌کند که آنها را باتوجه به پاسخ «بلی» یا «خیر» به سؤال درنظرگرفته‌شده برای گره t نامیده‌شده به‌عنوان گره جد[2] توسط t_Y و t_N نشان می‌دهیم. همان‌طورکه پیش از این ذکر شد، گره‌های اولاد با دو زیرمجموعه جدید یعنی X_{tY} و X_{tN} همبسته هستند. در خصوص متدولوژی رشد درخت، از گره ریشه‌ای تا گره‌های پایانی، هر دو نیمه باید زیرمجموعه‌هایی را تولید کند که در مقایسه با زیرمجموعه جد X_t مشابه‌تر هستند. بدین معنی که بردار ویژگی آموزشی در هریک از زیرمجموعه‌های جدید اولویت بالاتری را برای دسته(های) خاصی نشان می‌دهد، درحالی‌که، داده‌ها در X_t به‌صورت یکنواخت‌تری میان دسته‌ها توزیع شده‌اند. بنابراین، هدف، تعریف سنجه‌ای است که ناخالصی گره را کمی نموده و گره را دو نیمه می‌کند، به‌طوری‌که، ناخالصی کل گره‌های اولاد به‌صورت بهینه نسبت‌به ناخالصی گره جد کاهش پیدا می‌کند.

$P(\omega_i|t)$ را درنظر بگیرید که احتمال اینکه یک بردار در زیرمجموعه X_t همبسته با یک گره t بوده و متعلق به دسته ω_i، $i=1,2,...,M$ باشد را نشان می‌دهد. یک تعریف متداول مورد استفاده از ناخالصی گره که با $I(t)$ نشان داده شده، به‌این‌صورت تعیین می‌شود:

1. Splitting Rules
2. Ancestor Node

$$I(t) = -\sum_{i=1}^{M} P(\omega_i | t) \log_2 P(\omega_i | t) \qquad (۳)$$

که $\log_2$، لگاریتمی با پایه ۲ است که هیچ‌چیز دیگری مگر آنتروپی همبسته با زیرمجموعه X_t معروف به نظریه اطلاعاتی شانون نیست. نمایش این موضوع مشکل نیست که $I(t)$ مقدار حداکثر خود را می‌گیرد، اگر کلیه احتمالات برابر با $\frac{1}{M}$(بیشترین ناخالصی) هستند و برابر صفر می‌شود، اگر کلیه داده‌ها به یک دسته تعلق دارد و تنها یکی $P(\omega_i | t) = 1$ و بقیه برابر صفر (کمترین ناخالصی) هستند. در عمل، احتمالات توسط درصدهای نسبی $\frac{N_t^i}{N_t}$ برآورد می‌شوند که تعداد نقاط در N_t^i است که به دسته ω_i تعلق دارد. اکنون یک عمل دو نیم‌سازی را فرض کنید که N_{tY} عدد از نقاط به‌سوی گره «بلی» (X_{tY}) و N_{tN} عدد به سوی گره «خیر» (X_{tN}) ارسال می‌شوند. کاهش در ناخالصی گره به‌این‌صورت تعریف می‌شود:

$$\Delta I(t) = I(t) - \frac{N_{tY}}{N_t} I(t_Y) - \frac{N_{tN}}{N_t} I(t_N) \qquad (۴)$$

که $I(t_Y)$ و $I(t_N)$ به‌ترتیب ناخالصی‌های گره‌های t_Y و t_N هستند. اکنون هدف اتخاذ یکی از مجموعه سؤالات کاندید می‌شود که دو نیم‌سازی را اجرا می‌کند که به بیشترین کاهش ناخالصی منجر می‌شود.

۴-۳. قاعده توقف دو نیم‌سازی

سؤالی که اکنون مطرح می‌شود، آن است که چه زمانی فرد باید دو نیم‌سازی یک گره را متوقف کند و آن را به‌عنوان یک گره پایانی درخت لحاظ کند. یک راه ممکن اتخاذ یک آستانه T و توقف دو نیم‌سازی در صورتی است که حداکثر مقدار $\Delta I(t)$ در مورد کلیه دو نیم‌های ممکن کمتر از T است. راه‌های دیگر، توقف دو نیم‌سازی در صورتی است که عدد اصلی زیرمجموعه X_t به اندازه کافی کوچک بوده یا X_t مطلق است بدین معنی که کلیه نقاط در آن به یک دسته تعلق دارد.

۵-۳. قاعده تخصیص دسته

زمانی که یک گره به‌عنوان یک گره پایانی شناسایی می‌شود، سپس باید برای آن یک برچسب دسته تعیین شود. یک قاعده متداول مورد استفاده، قاعده اکثریت است یعنی گره پایانی بعنوان ω_i برچسب زده می‌شود که:

$$j=\arg\max_i P(\omega_i|t) \qquad (۵)$$

به‌عبارتی، یک گره پایانی t به آن دسته‌ای تخصیص داده می‌شود که اکثریت بردارها در X_t به آن تعلق دارد.

با داشتن عناصر اصلی مورد نیاز بحث‌شده برای رشد یک درخت تصمیم، اکنون می‌توانیم مراحل الگوریتم پایه‌ای را برای ساخت یک درخت تصمیم دودویی به‌صورت خلاصه بیان کنیم:

- ابتدا با گره ریشه‌ای آغاز کنید، یعنی $X_t = X$؛
- برای هر گره جدید t؛
  - برای هر ویژگی x_k، $k = ۱,۲,...,l$؛
  - برای هر مقدار α_{kn}، $n = ۱,۲,...,N_{tk}$؛
- X_{tY} و X_{tN} را باتوجه به پاسخ سؤال تولید کنید: $x_k(i) \le \alpha_{kn}$، $i = ۱,۲,...,N_t$ است.
- کاهش ناخالصی را محاسبه کنید؛
  - پایان؛
  - $\alpha_{kn_\circ}$ را که منجر به حداکثر کاهش w.r. به x_k می‌شود را انتخاب کنید.
  - پایان؛
  - $x_{k_\circ}$ و $\alpha_{k_\circ n_\circ}$ همبسته‌ای را که منجربه حداکثر کاهش کلی ناخالصی می‌شود را انتخاب کنید؛

- اگر قاعده توقف دو نیم‌سازی حاصل شده است، گره t را به‌عنوان یک گره پایانی درنظرگرفته و به آن یک برچسب دسته تخصیص دهید.
- درغیراین‌صورت، دو گره اولاد t_Y و t_N را با زیرمجموعه‌های همبسته X_{tY} و X_{tN} باتوجه به پاسخ به سؤال $x_{k_\circ} \leq \alpha_{k_\circ n_\circ}$، تولید کنید؛

– پایان؛

یک عامل حیاتی در طراحی یک درخت تصمیم، اندازه آن است. همچون پرسپترون‌های چندلایه، اندازه یک درخت باید به اندازه کفایت بزرگ باشد اما نه بسیار بزرگ؛ درغیراین‌صورت درخت، گرایش به یادگیری جزئیات خاص مجموعه آموزشی داشته و عملکرد تعمیم ضعیفی را نشان خواهد داد. تجربه نشان داده است که استفاده از یک مقدار آستانه برای کاهش ناخالصی به‌عنوان قاعده توقف دو نیم‌سازی به درخت‌هایی با اندازه درست منجر نمی‌شود. به‌طوری‌که، بسیار دیده شده است که رشد درخت بسیار زود یا بسیار دیر متوقف می‌شود. متداول‌ترین رویکرد مورد استفاده، ابتدا رشد یک درخت تا یک اندازه بزرگ و سپس هرس‌کردن گره‌ها براساس یک معیار هرس است. این رویه مشابه با هرس‌کردن پرسپترون‌های چندلایه است. یک معیار متداول هرس، ترکیب برآوردی از احتمال خطا همراه با یک عبارت سنجش پیچیدگی (به‌طورمثال، تعداد گره‌های پایانی) است (Theodoridis & Koutroumbas, 2006).

۶-۳. الگوریتم هرس[1]

هرس‌کردن فرآیند کاهش یک درخت از طریق تبدیل برخی گره‌های شاخه‌ای به سوی گره‌های پایانی و حذف گره‌های پایانی زیرشاخه اصلی است. الگوریتم هرس به‌طورکلی برای درخت‌هایی به‌کارمی‌رود که لزوماً طبقه‌بندی نبوده بلکه درخت‌های رگرسیونی هستند.

1. Pruning Algorithm

فرض کنید که $R(t)$ اعداد حقیقی همبسته با هر گره t از یک درخت معلوم T باشد. اگر t یک گره پایانی است، یعنی $t \in \tilde{T}$ آنگاه، $R(t)$ می‌تواند نسبت نمونه‌های طبقه‌بندی نشده (تعداد نمونه‌ها در $u(t)$ که به دسته همبسته با گره پایانی تعلق ندارد، تعریف شده به‌نام $M(t)$ که به کل تعداد نقاط داده‌ای n تقسیم می‌شود) را نمایش دهد:

$$R(t) = \frac{M(t)}{n} \quad t \in \tilde{T} \qquad (۶)$$

فرض کنید برای یک عدد حقیقی α، $R_\alpha(t) = R(t)+\alpha$ باشد. قرار دهید:

$$R(t) = \sum_{t \in \tilde{T}} R(t) \qquad (۷)$$

$$R_\alpha(t) = \sum_{t \in \tilde{T}} R_\alpha(t) = R(T) + \alpha\left|\tilde{T}\right| \qquad (۸)$$

در یک مسئله طبقه‌بندی، $R(t)$، نرخ عدم‌طبقه‌بندی برآورد شده است؛ $\left|\tilde{T}\right|$ عدد اصلی مجموعه $\tilde{T}$ است؛ $R_\alpha(T)$ نرخ عدم‌طبقه‌بندی-پیچیدگی برآوردشده از یک درخت طبقه‌بندی است و α یک مقدار ثابت است که می‌تواند به‌عنوان هزینه پیچیدگی هر گره پایانی درنظرگرفته‌شود. اگر α کوچک است، آنگاه جریمه کمی برای داشتن تعداد زیادی گره وجود دارد. همچنان که α افزایش می‌یابد، با کوچک‌سازی زیردرخت $T' \leq T$ که $R_\alpha(T')$ را حداقل می‌کند) گره‌های پایانی کمتری دارد (Webb, 2002).

۷-۳. وارسی اعتبار[1]

وارسی اعتبار، روشی برای برآورد میزان خطاست که ایده‌ای ساده دارد. مجموعه داده‌ها به اندازه نمونه به دو قسمت تفکیک می‌شوند. پارامترهای مدل با استفاده از یک مجموعه (با حداقل‌سازی چند معیار بهینه‌سازی) برآورد شده و معیار خوبی

1. Cross-Validation (CV)

برازش[1] بر روی مجموعه دوم ارزیابی می‌شود. نسخه معمول وارسی اعتبار یک روش صرف‌نظر از یکی[2] است که در آن مجموعه دوم متشکل از تنها یک نمونه است. آنگاه، برآورد وارسی اعتبار معیار خوبی برازش، متوسط کلیه مجموعه‌های آموزشی ممکن به اندازه n-۱ است. خطای وارسی اعتبار به‌عنوان وسیله‌ای برای تعیین یک مدل مناسب، برای هر عضو از خانواده مدل‌های کاندید محاسبه می‌شود، $\{M_k, k = ۱,...,K\}$، و مدل $M_{\hat{k}}$ انتخاب می‌شود که:

$$\hat{k} = \arg\min CV(k) \qquad (۹)$$

وارسی اعتبار هنگام انتخاب یک مدل صحیح، گرایش به برازش بیش از حد دارد، به‌طوری‌که، برای مجموعه داده‌ها یک مدل بسیار پیچیده انتخاب می‌کند. شواهدی وجود دارد که وارسی اعتبار چندتایی، زمانی‌که $d > ۱$ نمونه از مجموعه آموزشی حذف می‌شود، انتخاب مدل بهتر از وارسی اعتبار صرف‌نظر از یکی انجام می‌شود. برای n بزرگ، میزان محاسبات زیاد بوده و به طراحی n طبقه‌بندی‌کننده نیاز دارد. اگرچه، با هزینه افزایش در واریانس برآوردکننده، به‌صورت تخمینی نااریب است. یکی از معایب رویکرد وارسی اعتبار این است که امکان دارد به مقدار قابل توجهی محاسبات نیاز داشته باشد (Webb, 2002).

در ادامه مقاله، به پیاده سازی الگوریتم درخت رگرسیونی هرس‌شده جهت انتخاب ویژگی‌ها بر روی یک موردکاوی پرداخته و سپس نشان داده می‌شود که ویژگی‌های انتخاب‌شده به چه میزان منجربه بهبود نتایج پیش‌بینی رفتار خرید مشتریان خواهند شد.

۴. موردکاوی یک شرکت بیمه

از پرکاربردترین حوزه‌هایی که می‌توان از ابزارهای داده‌کاوی برای شناسایی و تحلیل مشتریان آن استفاده کرد، صنعت بیمه است. به‌ویژه اینکه، شناسایی ویژگی‌های

1. Goodness-of-Fit
2. Leave-one-Out

مشتریان یک محصول شرکت بیمه از طریق رفتاری که در قبال سایر محصولات بیمه‌ای شرکت از خود نشان داده‌اند، می‌تواند جالب توجه باشد. بدین‌طریق، شرکت می‌تواند براساس انواع مختلف مشتریان، استراتژی‌های بازاریابی خود را به‌منظور کاهش هزینه‌های بازاریابی تدوین کند. در این تحقیق، مجموعه داده‌ها از طریق یک شرکت بیمه هلندی تهیه شده و مربوطبه یک کسب‌وکار در دنیای واقعی می‌شود. این شرکت بیمه می‌خواهد مشتریان بالقوه برای یک محصول معین را شناسایی کند.

مجموعه داده‌های اصلی شامل رکوردهای ۵۸۲۲ مشتری می‌شود که به‌منظور آموزش و اعتبارسنجی مدل‌های پیش‌بینی و ایجاد یک توصیف از مشتریان استفاده می‌شوند. هر رکورد از ۸۶ ویژگی شامل داده‌های آمارگیری اجتماعی (ویژگی‌های ۱-۴۳) و داده‌های مالکیت محصول و بهره‌گیری از محصولات بیمه‌ای مختلف (ویژگی‌های ۴۴-۸۶) می‌شود که ویژگی ۸۶ متغیر هدف است. قصد داریم پیشگویی کنیم که چه کسی به خرید یک قرارداد بیمه‌ای کاروان (خانه متحرک) علاقمند خواهد بود؟ از میان این مشتریان، ۳۴۸ نفر حدود ۶٪ بیمه‌نامه کاروان دارند. داده‌های مورد استفاده برای پیشگویی شامل رکوردهای ۴۰۰۰ مشتری می‌شود که جهت ارزیابی مدل استفاده می‌شوند. در میان این مشتریان، ۲۳۸ مشتری خریداران بیمه‌نامه کاروان بوده‌اند که شناسایی ویژگی‌های آنها مدنظر است. شرکت بیمه از ما خواسته است که زیرمجموعه‌ای ۸۰۰(۴۰۰۰×۲۰٪)تایی از مجموعه تست را معرفی کنیم که بیشترین تعداد خریداران بیمه‌نامه کاروان را دربرگیرد. ازاین‌رو، گام اول انتخاب ویژگی‌های مناسب به‌منظور پیش‌بینی رفتار مشتریان است.

۱-۴. انتخاب ویژگی‌های مؤثر مشتریان بر خرید بیمه‌نامه کاروان

در بخش قبل در خصوص نحوه عملکرد درخت تصمیم رگرسیونی به‌عنوان یک ابزار جهت انتخاب ویژگی‌ها صحبت شد. در این قسمت قصد داریم از این ابزار به‌منظور

شناسایی ویژگی‌های تأثیرگذار بر متغیر خرید بیمه‌نامه کاروان استفاده کنیم. مشخصات درخت رگرسیونی به‌کارگرفته شامل این موارد می‌شود:

– نوع درخت، درخت رگرسیونی است، زیرا نوع متغیر هدف به‌صورت کمی بوده و ۰ یا ۱ است.

– از وارسی اعتبار ۱۰ تکه‌ای جهت ارزیابی خوبی برازش از طریق محاسبه بردار هزینه استفاده می‌شود. بدین معنی که تابع، نمونه را به ۱۰ زیرنمونه که به‌صورت تصادفی انتخاب شده و تقریباً دارای اندازه برابر هستند، تفکیک می‌کند. هریک از زیرنمونه‌ها، یک درخت را به داده‌های باقی‌مانده برازش می‌دهد و از آن برای پیش‌بینی زیرنمونه استفاده می‌کند. سپس، اطلاعات کلیه زیرنمونه‌ها را برای محاسبه هزینه برای کل نمونه با یکدیگر ترکیب می‌کند. همچنین خطای استاندارد هر مقدار هزینه محاسبه‌شده و تعداد گره‌های پایانی برای هر زیردرخت و بهترین سطح برآوردشده جهت هرس‌زنی تعیین می‌شود. بهترین سطح، کوچک‌ترین درختی را تولید می‌کند که در محدوده یک خطای استاندارد از زیردرخت حداقل هزینه است. همچنین، درخت با حداقل هزینه انتخاب می‌شود.

– جهت هرس‌کردن درخت، معیار کمترین میزان خطای عدم‌طبقه‌بندی به‌کارمی‌رود.

– هزینه درخت، مجموع کلیه گره‌های پایانی شامل احتمال برآوردشده هر گره در هزینه گره است.

– احتمال یک گره از طریق نسبت مشاهدات از داده‌های اصلی که شرایط گره را برآورده می‌کنند، محاسبه می‌شود. برای درخت رگرسیونی، نسبت برای هر احتمال اولیه‌ای تخصیص داده‌شده به هر دسته تعدیل می‌شود.

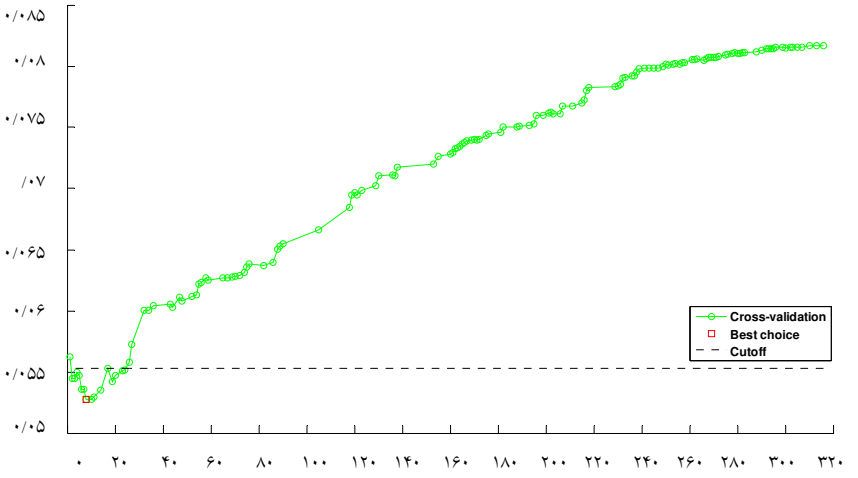
– هزینه یک گره، میانگین مربعات خطا در کلیه مشاهدات در آن گره است.

– خطا برای هر گره، واریانس مشاهدات تخصیص داده‌شده به آن گره است.

– اندازه یک گره به‌صورت تعداد مشاهدات از داده‌های مورد استفاده برای ایجاد درختی که شرایط را برای گره برآورده کند، تعریف می‌شود.

بدین‌ترتیب، با ورود داده‌های ۵۸۲۲ مشتری که شامل ۸۵ ویژگی است، درخت تصمیم خرید بیمه‌نامه کاروان ترسیم می‌شود. نمودار ۱ نقطه بهینه برای هرس‌کردن درخت به‌منظور لحاظ ویژگی‌های مؤثرتر بر متغیر هدف یعنی خرید بیمه‌نامه کاروان توسط مشتریان را نشان می‌دهد.

نمودار۱. تعیین نقطه بهینه برای هرس‌کردن درخت رگرسیونی خرید بیمه‌نامه کاروان



درخت تصمیم شامل ۳۱۶ گره پایانی است که برای زیردرخت‌های مختلف ترسیم شده است (نمودار ۱). به‌طورکلی درخت شامل ۶۳۱ گره اولیه می‌باشد که از گره ریشه‌ای درخت به‌تدریج هزینه عدم‌طبقه‌بندی (میانگین مربعات خطا) کاهش یافته و در گره پایانی ۹ به حداقل مقدار خود برابر ۰/۰۵۲۸ می‌رسد. سپس، این هزینه با اندکی نوسان به تدریج روندی صعودی را طی می‌کند. بنابراین، درخت تصمیم بهینه در گره پایانی ۹ با حداقل هزینه عدم‌طبقه‌بندی هرس می‌شود. دراین‌حالت، پیچیدگی‌های درخت نیز با کاهش تعداد گره‌های پایانی کاهش می‌یابد. جدول ۱

ویژگی‌های مؤثر بر خرید بیمه‌نامه کاروان را به‌عنوان خروجی درخت رگرسیونی هرس‌شده نمایش می‌دهد.

جدول ۱. قواعد درخت رگرسیونی برای مجموعه داده‌های آموزشی پس از هرس‌زنی

گره	قاعده	احتمال	اندازه
۱	اگر بیمه‌نامه اتومبیل مشارکتی کوچک‌تر از ۵/۵ است برو به گره ۲ درغیراین‌صورت گره ۳.	۱	۵۸۲۲
۲	fit = ۰/۰۲۴۸	۰/۵۹۴۱	۳۴۵۹
۳	اگر نوع عمده مشتری کوچک‌تر از ۲/۵ است، برو به گره ۴ درغیراین‌صورت گره ۵.	۰/۴۰۵۸	۲۳۶۳
۴	اگر بیمه‌نامه آتش‌سوزی مشارکتی کوچک‌تر از ۳/۵ است، برو به گره ۶ درغیراین‌صورت گره ۷.	۰/۰۷۹۶	۴۶۴
۵	اگر بیمه‌نامه قایق مشارکتی کوچک‌تر از ۰/۵ است، برو به گره ۸ درغیراین‌صورت گره ۹.	۰/۳۲۶۱	۱۸۹۹
۶	fit = ۰/۱۲۸۶	۰/۰۴۱۳	۲۴۱
۷	اگر کارگر بی‌تجربه کوچک‌تر از ۳/۵ است، برو به گره ۱۰ درغیراین‌صورت گره ۱۱.	۰/۰۳۸۳	۲۲۳
۸	اگر بیمه‌نامه آتش‌سوزی مشارکتی کوچک‌تر از ۲/۵ است، برو به گره ۱۲ درغیراین‌صورت گره ۱۳.	۰/۳۲۳۲	۱۸۸۲
۹	اگر بیمه شخص ثالث شخصی مشارکتی کوچک‌تر از ۱ است، برو به گره ۱۴ درغیراین‌صورت گره ۱۵.	۰/۰۰۲۹	۱۷
۱۰	fit = ۰/۲۶۱۶	۰/۰۳۶۷	۲۱۴
۱۱	fit = ۰/۷۷۷۷	۰/۰۰۱۵	۹
۱۲	fit = ۰/۰۵۰۹	۰/۱۶۵۰	۹۶۱
۱۳	fit = ۰/۱۱۹۴	۰/۱۵۸۱	۹۲۱
۱۴	fit = ۰/۸۱۸۱	۰/۰۰۱۸	۱۱
۱۵	fit = ۰	۰/۰۰۱۰	۶

بدین‌ترتیب براساس جدول ۱، شش متغیر شامل بیمه‌نامه اتومبیل مشارکتی، نوع عمده مشتری، بیمه‌نامه آتش‌سوزی مشارکتی، بیمه‌نامه قایق مشارکتی، کارگر بی‌تجربه و بیمه شخص ثالث مشارکتی به‌عنوان مؤثرترین عوامل تأثیرگذار بر خرید بیمه‌نامه کاروان با استفاده از درخت رگرسیونی پس از هرس‌کردن انتخاب شده‌اند.

۲-۴. پیش‌بینی رفتار خرید مشتریان

یک مرحله مهم در طراحی یک سیستم تشخیص مشتری، مرحله ارزیابی عملکرد است که در آن احتمال خطای پیش‌بینی سیستم طراحی‌شده برآورد می‌شود. باید تأکید شود که این مرحله بسیار حیاتی است. اگر ویژگی‌هایی با قدرت تشخیص مشتری پایین انتخاب شوند، متعاقباً سیستم عملکرد ضعیفی خواهد داشت.

بنابراین، پس از انتخاب ویژگی‌ها با استفاده از درخت رگرسیونی هرس‌شده، باید بررسی شود که آیا ویژگی‌های انتخاب‌شده منجربه بهبود پیش‌بینی رفتار خرید مشتریان این محصول شرکت شده‌اند یا خیر. بدین‌منظور، پیش‌بینی در دو حالت انجام شده است. در حالت اول، از کلیه ۸۵ ویژگی به‌منظور آموزش و اعتبارسنجی درخت رگرسیونی استفاده می‌شود. در حالت دوم، درخت رگرسیونی هم به‌منظور انتخاب ویژگی‌ها و هم به‌منظور پیش‌بینی مشتریان به‌کارمی‌رود. جدول ۲ نتایج به‌دست‌آمده از پیش‌بینی رفتار خرید مشتریان را برای خرید بیمه‌نامه کاروان نشان می‌دهد.

جدول ۲. نتایج حاصل از پیش‌بینی در دو حالت مختلف

تعداد ویژگی‌ها	روش پیش‌بینی	مشتریان بیمه‌نامه کاروان میان ۲۰٪ از مشتریان پیش‌بینی‌شده					
		مجموعه آموزشی			مجموعه اعتبارسنجی		
		تعداد	درصد از کل مشتریان	درصد از مشتریان پیش‌بینی‌شده	تعداد	درصد از کل مشتریان	درصد از مشتریان پیش‌بینی‌شده
۸۵	درخت تصمیم	۲۶۹	٪۷۷/۳۰	٪۲۳/۱۱	۹۲	٪۳۸/۶۶	٪۱۱/۵۰
۶	درخت تصمیم	۱۷۸	٪۵۱/۱۵	٪۱۵/۳۲	۱۰۰	٪۴۲/۰۲	٪۱۲/۵۰

همان‌طورکه در جدول ۲ مشاهده می‌شود، اگر پیش از پیش‌بینی، اقدام به تعیین ویژگی‌های مؤثر بر هدف نکنیم، در مجموعه آموزشی با انتخاب ۱۱۶۴ مشتری از ۵۸۲۲ مشتری موجود (۲۰٪)، ۲۶۹ مشتری بیمه‌نامه کاروان از کل ۳۴۸ مشتری بیمه‌نامه کاروان (۷۷/۳٪) شناسایی می‌شوند که این تعداد ۲۳/۱۱٪ از مشتریان پیش‌بینی‌شده را تشکیل می‌دهد. درحالی‌که، نتایج داده‌های اعتبارسنجی دراین‌حالت نشان می‌دهد که با انتخاب ۲۰٪ از مشتریان تنها می‌توان ۹۲ مشتری از ۲۳۸ مشتری که اقدام به خرید بیمه‌نامه کاروان (۳۸/۶۶٪) نموده‌اند را پیش‌بینی کرد که تنها ۱۱/۵٪ از کل مشتریان پیش‌بینی‌شده را شامل می‌شود. اختلاف بسیار زیادی که بین مقادیر داده‌های آموزشی و داده‌های اعتبارسنجی وجود دارد، نشان‌دهنده وجود ویژگی‌های نامرتبط در میان ورودی‌هاست که منجربه کاهش دقت در برآورد دقیق پیش‌بینی شده

است. به‌طوری‌که، اگر پیش‌بینی با ۶ ویژگی منتخب به‌عنوان ورودی به درخت رگرسیونی انجام شود، با انتخاب ۲۰٪ از کل مشتریان در مجموعه آموزشی ۱۷۸ مشتری بیمه‌نامه کاروان (۵۱/۱۵٪) شناسایی شده که ۱۵/۳۲٪ از مشتریان پیش‌بینی‌شده را تشکیل می‌دهند. به‌طوری‌که، با به‌کارگیری این ورودی‌ها جهت پیش‌بینی در مجموعه اعتبارسنجی ۱۰۰ مشتری بیمه‌نامه کاروان (۴۲/۰۲٪) پیش‌بینی می‌شوند که ۱۲/۵٪ از کل مشتریان پیش‌بینی‌شده را شامل می‌شود که مقایسه حالت ۶ ویژگی به‌عنوان ورودی با حالت ۸۵ ورودی، بهبود قابل توجه نتیجه پیش‌بینی را نشان می‌دهد.

۵. نتیجه‌گیری

در این مقاله، در خصوص اهمیت انتخاب ویژگی‌ها در حوزه پیش‌بینی رفتار خرید مشتری بحث کردیم که یکی از حوزه‌های مهم مدیریت ارتباط با مشتری است. بدین‌منظور، پس از تحلیل ادبیات موضوع و تکنیک‌های مختلف، درخت رگرسیونی را به‌عنوان ابزار مناسبی جهت انتخاب ویژگی‌ها برگزیدیم. ازآنجایی‌که این تکنیک یک روشی غیرخطی بوده و رفتار مشتریان نیز غیرخطی است، به‌نظرمی‌رسد به‌کارگیری این ابزار بسیار مناسب باشد. همچنین مزیت دیگر این تکنیک سرعت قابل توجه آن در شناسایی ویژگی‌های تأثیرگذار بر روی هدف است. به‌منظور انتخاب ویژگی‌های مؤثرتر، درخت تصمیم حاصل‌شده هرس گردید تاآنجایی‌که میزان هزینه عدم‌طبقه‌بندی مشتریان به حداقل برسد. تکنیک بیان‌شده را بر روی یک موردکاوی شرکت بیمه برای پیش‌بینی یکی از محصولات آن اجرا کردیم. به‌طوری‌که، ویژگی‌های بهینه با استفاده از درخت رگرسیونی هرس‌شده تعیین شدند. کاملاً واضح است که قوی‌ترین پیشگویی‌کننده برای خرید بیمه‌نامه کاروان، ویژگی خرید بیمه‌نامه اتومبیل مشارکتی است. به‌طوری‌که، بیش از سایر ویژگی‌ها خریداران و غیرخریداران را از هم تفکیک می‌کند. درنهایت، به‌منظور نمایش میزان بهبودی که در پیش‌بینی مشتریان با انتخاب ویژگی‌های درست حاصل می‌شود، در دو حالت اقدام به پیش‌بینی

کردیم. نتایج پیش‌بینی نشان می‌دهد که انتخاب ویژگی‌ها علاوه‌بر کاهش ویژگی‌های موجود از ۸۵ به ۶ ویژگی منجربه بهبود قابل توجهی در نتیجه پیش‌بینی شده است.

منابع

1. Ahn, H, Kim, K & Han, I 2007, 'A case-based reasoning system with the two-dimensional reduction technique for customer classification', *Expert Systems with Applications*, vol. 32, pp. 1011–19.

2. Anderson, ET 2002, 'Sharing the wealth: when should firms treat customers as partners?', *Management Science,* vol. 48, no. 8, pp. 955-71.

3. Blum, AL & Rivest, RL 1992, 'Training a 3-node neural networks is NP-complete', *Neural Networks*, vol. 5, pp. 117-27.

4. Buckinx, W, Verstraeten, G & Poel, DV 2007, 'Predicting customer loyalty using the internal transactional database', *Expert Systems with Applications*, vol. 32, pp. 125–34.

5. Hung, YH 2009, 'A neural network classifier with rough set-based feature selection to classify multiclass IC package products', *Advanced Engineering Informatics*, vol. 24, no. 4, pp. 348-57.

6. Kim, YS 2006, 'Toward a successful CRM: variable selection, sampling, and ensemble', *Decision Support Systems,* vol. 41, pp. 542–53.

7. Kim, YS & Street, WN 2004, 'An intelligent system for customer targeting: a data mining approach', *Decision Support Systems,* vol. 37, pp. 215– 28.

8. Kohavi, R & John, GH 1997, 'Wrappers for feature subset selection', *Artificial Intelligence,* no. 1–2, vol. 1-2, pp. 273–324.

9. Lessmann, S & VoB, S 2009, 'A reference model for customer-centric data mining with support vector machines', *European Journal of Operational Research*, vol. 199, pp. 520–30.

10. Maldonado, S & Weber, R 2009, 'A wrapper method for feature selection using Support Vector Machines', *Information Sciences*, vol. 179, pp. 2208–17.

11. Ng, KS & Liu, H 2000, 'Customer retention via data mining', *AI Rev,* vol. 14, no. 6, pp. 569-90.

12. Theodoridis, S & Koutroumbas, K 2006, *Pattern recognition*, Academic Press, 3rd ed.

13. Tseng, TL & Huang, CC 2007, 'Rough set-based approach to feature selection in customer relationship management', *Omega,* vol. 35, pp. 365-83.

14. Webb, AR 2002, *Statistical pattern recognition*, John Wiley & Sons, Ltd., 2nd ed.

15. Yan, L & Changrui, Y 2007, 'A new hybrid algorithm for feature selection and its application to customer recognition', *Lecture Notes in Computer Science*, vol. 4616, pp. 102-11.

16. Yan, L, Wolniewicz, R & Dodier, R 2004, 'Predicting customer behaviour in telecommunications', *IEE Intelligent Systems,* vol. 19, pp. 50–8.

17. Yu, E & Cho, S 2006, 'Constructing response model using ensemble based on feature subset selection', *Expert Systems with Applications,* vol. 30, pp. 352–60.