فصلنامه مدیریت نوآوری در سازمان‌های دفاعی
سال اول، شماره دو
زمستان ۹۷
صص ۱۲۹-۱۵۲

ارائه الگوریتم پویا با استفاده از مدل‌های فرااکتشافی در جریان داده‌ها بر اساس نمونه‌گیری مجدد به منظور ارتقاء پیش‌بینی پاسخ مشتریان

مهدی زکی‌پور[۱]، سینا نعمتی‌زاده[۲]، محمدعلی افشار کاظمی[۳]

چکیده

زمینه و هدف: هدف از پژوهش حاضر مواجهه با مشکل عدم توازن داده‌ها به عنوان یک مسئله نهادینه شده در حوزه تحقیقات علوم انسانی است. عدم توازن در بین داده‌های متعلق به کلاس اقلیت و اکثریت باعث تمایل الگوریتم‌های طبقه‌بند به سمت کلاس اکثریت می‌شود.

روش‌شناسی: این پژوهش از لحاظ هدف، کاربردی و از لحاظ نحوهٔ گردآوری داده‌ها پژوهشی توصیفی، علی – مقایسه‌ای می‌باشد. در این پژوهش از روش‌های مختلف متوازن‌سازی داده‌ها به منظور ارتقاء قدرت پیش‌بینی الگوریتم‌های شبکهٔ عصبی، درخت تصمیم و رگرسیون لجستیک بر اساس مدل‌های فرااکتشافی استفاده شده است.

یافته‌ها: با استفاده از داده‌های اولیه به هیچ وجه نمی‌توان به یک پیش‌بینی قابل اتکا و قابل استفاده دست یافت. به کارگیری روش‌های نمونه‌گیری مجدد با استفاده از خوشه‌بندی مشتریان و ترکیب کلاس‌های اقلیت و اکثریت به روش‌های مختلف و مطابق با الگوریتم ارائه شده می‌تواند توان پیش‌بینی طبقه‌بندها را به طرز شگفت‌انگیزی افزایش داده و در موقعیت‌های مختلف مورد استفاده قرار گیرد.

نتیجه‌گیری: این پژوهش با استفاده از نتایج حاصل از کدهای XML استخراج شده در هر مرحله به شناسایی هر چه دقیق‌تر مشتریان بالقوه پرداخته و نیز با ترکیب معیارهای مختلف ارزیابی مدل به روشی ابتکاری در جهت تغییر خروجی مدل‌های پیش‌بینی از حالت باینری به فازی از یافته‌ها و نتایج پژوهش‌های پیشین گامی فراتر برداشته شده است.

کلیدواژه‌ها: *عدم توازن داده‌ها، خوشه‌بندی، داده‌کاوی، پیش‌بینی.*

---

۱. دانشجوی دکتری، دانشکدهٔ مدیریت، دانشگاه آزاد اسلامی، واحد تهران مرکزی، تهران، ایران.
۲. دانشیار، گروه مدیریت بازرگانی، دانشکدهٔ مدیریت، دانشگاه آزاد اسلامی، واحد تهران مرکزی، تهران، ایران.
۳. دانشیار، گروه مدیریت صنعتی، دانشکدهٔ مدیریت، دانشگاه آزاد اسلامی، واحد تهران مرکزی، تهران، ایران.

تاریخ دریافت مقاله: ۹۷/۱۰/۱۲
تاریخ پذیرش نهایی مقاله: ۹۷/۱۱/۰۱
نویسنده مسئول مقاله: سینا نعمتی‌زاده
E-mail: nematizadeh51@yahoo.com

مقدمه

مجموعه داده نامتوازن بر اساس تعریف عبارت از مجموعه داده‌ای است که تعداد نمونه‌های متعلق به یک کلاس در آن با تعداد نمونه‌های کلاس دیگر به طور مساوی توزیع نشده باشد (چاولا، ۲۰۰۹). کلاس با تعداد داده‌های بیشتر را کلاس اکثریت و کلاس با داده‌های کمتر را کلاس اقلیت می‌نامند. در الگوریتم‌های طبقه‌بند استاندارد، توزیع کلاس‌ها متوازن در نظر گرفته می‌شود و این دسته از الگوریتم‌ها در مواجهه با مجموعه داده‌های نامتوازن عملکرد مناسبی را از خود ارائه نمی‌دهند، چرا که الگوریتم‌های معمول طبقه‌بند به سمت نمونه‌های آموزشی کلاس بزرگ‌تر متمایل می‌شوند که این موضوع باعث افزایش خطا در شناسایی نمونه‌های اقلیت می‌شود (سان، وانگ و کمل، ۲۰۰۹). این مسئله یکی از چالش‌های پیش رو برای طبقه‌بندی داده‌های نامتوازن محسوب می‌شود و امروزه نظر بسیاری از متخصصان و پژوهشگران حوزه تحلیل داده را به خود جلب کرده است (یانگ و وو، ۲۰۰۶). به عنوان مثال در مسائل مرتبط با حوزه مدیریت منابع انسانی: به منظور پیش‌بینی و شناسایی کارکنانی که اقدام به جابه‌جایی شغلی خواهند نمود، در مسائل مرتبط با حوزه دانشگاهی: به منظور پیش‌بینی و شناسایی دانشجویانی که اقدام به انصراف از تحصیل خواهند نمود، در مسائل مرتبط با حوزه مدیریت بازاریابی: به منظور پیش‌بینی و شناسایی مشتریانی که اقدام به خرید خواهند کرد (مشتریان بالقوه) و یا در حتی حوزه علوم پزشکی: به منظور شناسایی بیمارانی درگیر یک بیماری نادر می‌باشند، عدم توازن کلاس داده‌ها (تعداد معدود کارکنانی که اقدام به ترک یا جابه‌جایی شغلی در یک سازمان نموده‌اند، تعداد کم دانشجویانی که از یک رشته تحصیلی یا دانشگاه خواصی انصراف داده‌اند و تعداد نادر بیمارانی که درگیر با یک بیماری خاص هستند) باعث بروز سوگیری و انحراف بالایی در نتایج پیش‌بینی و طبقه‌بندی مدل‌های مختلف خواهد شد.

مدل‌های پیش‌بینی داده‌محور به طور گسترده‌ای برای پیش‌بینی پاسخ مشتری به کمپین‌های بازاریابی مورد استفاده قرار گرفته است (چن، فن و سان، ۲۰۱۵) ولی آنچه که باعث ضعف این مدل‌ها و وخیم‌تر شدن مسئله می‌شود آن است که نرخ پاسخ در شرایط کلی بازاریابی مستقیم اغلب بسیار پایین است. این امر باعث ایجاد یک پایگاه داده نامتعادل و در نتیجه اختلال در پیش‌بینی مشتریان می‌شود. اگر راه حل مناسبی برای برطرف کردن عدم تعادل کلاس به‌کار برده نشود، الگوریتم‌های طبقه‌بندی که توسط مدل پاسخ استفاده می‌شوند، احتمالاً اکثر مشتریان را به عنوان کسانی که پاسخی

ارائه نمی‌کنند در نظر می‌گیرند و این امر به ایجاد هزینه فرصت از دست رفته منجر خواهد شد. به همین دلیل، مدیریت عدم تعادل کلاس اطلاعات مشتری، به عنوان یک عامل مهم برای موفقیت بازاریابی مستقیم شناخته شده است (هیل، پرووست و وولین اسکای، ۲۰۰۶؛ لای، وانگ، لینگ، شی و ژنگ، ۲۰۰۶؛ لینگ و لی، ۱۹۹۸). هدف از پژوهش حاضر طراحی الگوریتمی پویا به منظور ارتقاء توان پیش‌بینی طبقه‌بندها با استفاده از ترکیب مدل‌ها و روش‌های ابتکاری نمونه‌گیری مجدد است. بدین منظور این مسئله در حوزه بازاریابی مستقیم که یکی از ابزارها و فنون ترفیع در استراتژی‌های بازاریابی است استفاده شده است چرا که عدم توازن کلاس مشتریان پاسخ‌گو یکی از مسائل و مشکلات به وضوح مشاهده شده در پایگاه داده مشتریان است.

مروری بر مبانی نظری

مواجهه با عدم توازن داده‌ها

از روش‌های متنوعی برای حل مسئله عدم توازن در علم یادگیری ماشین استفاده می‌شود. یکی از این روش‌ها، روش‌های بازبینی در سطح الگوریتم[1] است که با تغییر در الگوریتم طبقه‌بندی، به نوعی مسئله عدم توازن مرتفع می‌شود (گالار، فرناندز، بارنچی، بوستینس و هررا، ۲۰۱۲؛ باراندلا، سنچزب و گارسیا، ۲۰۰۳). از دیگر روش‌های حل نامتوازن بودن داده‌ها روش‌های مبتنی بر ترکیب طبقه‌بندها است. هدف اصلی روش ترکیب[2] تلاش برای بهبود عملکرد طبقه‌بندی داده‌ها از طریق ترکیب چندین طبقه‌بند است. به‌طوری‌که ترکیب چند طبقه‌بند عملکرد بهتری نسبت به یکی از همان طبقه‌بندها خواهد داشت. یان و همکاران با استفاده از این روش و ماشین بردار پشتیبان[3] توانستند مسئله پیش‌بینی کلاس اقلیت را بهبود دهند (یان، لی، چین و هاپمن، ۲۰۰۳). روش سوم روش‌های سطح داده[4] است. در این دسته از روش‌ها، توزیع کلاس نامتوازن با نمونه‌گیری مجدد[5] در فضای داده‌ها متوازن می‌شود

1. Algorithm level
2. Ensemble Methodology
3. Support Vector Machine
4. Data Level
5. Re-sampling

(نافیرالا، استفانوفسکی و ولیک، ۲۰۱۰) و نهایتاً روش‌های حساس به هزینه[1] دسته دیگری از روش‌های ارائه شده برای حل عدم توازن در داده‌ها محسوب می‌شود. این دسته از روش‌ها به نوعی از ترکیب روش‌های تغییر در الگوریتم طبقه‌بند و روش‌های سطح داده حاصل می‌شوند (ژنگ، لی، ژو و ژنگ، ۲۰۰۸). در این بین، روش‌های سطح داده با دو رویکرد کم‌نمونه‌برداری[2] و بیش‌نمونه‌برداری[3] از روش‌های مؤثر در متوازن نمودن داده‌ها محسوب می‌شوند (لی، لئو و هو، ۲۰۱۰).

بازاریابی مستقیم

بازاریابی مستقیم یکی از انواع بازاریابی است که در آن، فرایندهای مربوط به بازاریابی و ارتباط با مشتریان، به جای اینکه به صورت انبوه انجام شود، صرفاً برای مشتریان بالقوه یا بالفعل که هویت و تعداد و مشخصات آن‌ها تا حد زیادی مشخص است، طراحی و اجرا می‌شود (لنکستر و مسینگهام[4]، ۲۰۰۱). علاوه بر اجزای قدیمی ترویج در آمیخته بازاریابی، بازاریابی مستقیم با انجام بازاریابی تعاملی، اسپانسرینگ، ارتباطات شفاهی، تبلیغات پاسخ مستقیم، پست مستقیم مصرف کننده، پست مستقیم کسب و کار، سفارش پستی، استفاده از ابزارهای آنلاین و شبکه اینترنت، بازاریابی تلفنی، فروش شخصی مستقیم و در نهایت بازاریابی چند سطحی سعی در برقراری ارتباط اثربخش‌تر با مشتری دارد (بون و کورتز[5]، ۲۰۱۳). بازاریابی مستقیم نوعی از بازاریابی است که در آن مستقیماً با مشتری مشکوک مشکوک ارتباط برقرار می‌شود. مشتری مشکوک به کسی گفته می‌شود که امکان فروش خدمات و یا محصول مورد نظر به وی وجود دارد (لینوف و بری، ۲۰۱۱). بازاریابی مستقیم به بازاریاب اجازه می‌دهد که پاسخ‌های مستقیم بیشتری را از مشتری دریافت کرده، بازار هدف را به گونهٔ بهتری نشانه‌گیری کند و محصول را بدون قرار گرفتن در فرایند عریض و طویل و پرهزینه کانال‌های سنتی توزیع، به فروش رساند (ویلکینسون، مک آلیستر و ودمیر[6]، ۲۰۰۷).

---

1. Cost-sensitive learning
2. Under-Sampling
3. Over-Sampling
4. Geoff Lancaster and Lester Massingham
5. Boone and Cortez
6 Wilkinson, McAlister, & Widmier

مدل‌سازی پاسخ

مدل‌سازی پاسخ یکی از مؤثرترین ابزارها برای شرکت‌هایی است که به دنبال برقراری روابط درازمدت با مشتریان خود هستند (سان، لی و ژو، ۲۰۰۶). هدف از مدل‌سازی پاسخ برای بازاریابی مستقیم این است که بر اساس تاریخچه رفتار مشتری و سایر اطلاعات در دسترس، مشتریانی را که احتمالاً یک محصول رقابتی را خریداری می‌کنند شناسایی شوند. مدل‌سازی پاسخ یکی از مؤثرترین ابزارها برای شرکت‌هایی است که به دنبال برقراری روابط درازمدت با مشتریان خود هستند (بری و لینوف، ۱۹۹۷؛ گونول و هافستد، ۲۰۰۶؛ سان و همکاران، ۲۰۰۶). هدف از مدل‌سازی واکنش این است که بر اساس تاریخچه خرید مشتری و سایر اطلاعات در دسترس، مشتریانی را شناسایی کند که به احتمال زیاد یک محصول را خریداری می‌کنند. بر اساس پیش‌بینی‌های مدل، شرکت‌ها تلاش می‌کنند تا خریداران بالقوه بیشتری را برای خرید محصول کمپین شده با استفاده از کانال‌های ارتباطی خودشان مانند تلفن، کاتالوگ ارسال شده یا ایمیل ترغیب و تهییج کنند.

پیشینه پژوهش

رضائی نوایی و همکاران (۲۰۱۷) در مقاله‌ای با عنوان «به‌کارگیری و ارزیابی تکنیک‌های داده‌کاوی جهت پیش‌بینی رویگردانی مشتری در صنعت بیمه» به بررسی علل و پیش‌بینی رویگردانی مشتریان در صنعت بیمه پرداخته است. در این مقاله نخست از الگوریتم ژنتیک برای فرایند انتخاب مشخصه‌های تأثیرگذار استفاده شده است. پس از مدل‌سازی مسئله، پارامترهای مدل ماشین بردار پشتیبان با استفاده از دو روش جستجوی شبکه و اعتبارسنجی متقابل K لایه، بهینه می‌شوند. عملکرد پیش‌بینی روش SVM با روش‌های درخت تصمیم، شبکه‌های عصبی، رگرسیون لجستیک، جنگل تصادفی، دسته‌بندی‌کننده بیزی، K نزدیک‌ترین همسایگی، مقایسه و بهینه‌سازی پارامترهای هر روش با استفاده از جستجوی شبکه انجام شده است. یافته‌های تحقیق نشان می‌دهد که روش ماشین بردار پشتیبان از عملکرد بالاتری نسبت به سایر روش‌ها برخوردار است. در مدل پیشنهادی مبتنی بر این روش، مشخصه‌های سابقه خرید، نحوه آشنایی با سازمان و تمایل به خرید، به عنوان مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی مشتری شناسایی شدند.

بصیری (۱۳۸۶) در پژوهشی با عنوان «کاربرد تکنیک داده‌کاوی در مدیریت روابط مشتری»

مدیریت روابط مشتری را به‌عنوان حوزه جدیدی که شرکت در آن می‌تواند مزیت رقابتی کسب نماید معرفی نموده است. همچنین در این مقاله چنین آمده است که ابزارهای داده‌کاوی به سؤالاتی از کسب و کار پاسخ می‌دهد که در گذشته پیگیری آن‌ها بسیار وقت‌گیر بوده است. با این حال توانمندی‌های موجود در داده‌کاوی، مدیریت روابط مشتری را به نحو بهتری ممکن می‌سازد. از مزایای دیگر استفاده از ابزارهای داده‌کاوی در مدیریت روابط مشتریان، در نظر گرفتن حریم خصوصی مشتریان است. تلفیق CRM و داده کاوی باعث پاسخ‌گویی مؤثر به نیاز مشتری، بهینه نمودن بازدهی سرمایه و بهره‌وری نیروی انسانی، ارتقاء کیفیت در محصولات و در نهایت موجب پاسخ‌گویی سریع به تغییرات محیطی می‌گردد.

مورو و همکاران[1] (۲۰۱۴) در تحقیقی به روش داده‌کاوی به پیش‌بینی موفقیت تماس تلفنی برای فروش سپرده‌های بلندمدت بانکی پرداخته‌اند. در این تحقیق ویژگی‌های اثرگذار در مدل‌های پیش‌بینی به‌صورت نیمه خودکار انتخاب شده‌اند و از چهار مدل داده‌کاوی از قبیل رگرسیون لجستیک، درخت تصمیم، شبکه عصبی و ماشین بردار استفاده شده است. با استفاده از دو معیار، مساحت زیر منحنی‌های راک و منحنی تجمعی لیفت هر چهار مدل در یک مجموعه ارزیابی شدند. نتایج بدست آمده حاکی از این است که شبکهٔ عصبی بهترین پیش‌بینی را بعمل آورده است (سرجیو مورو، کورتز و ریتا، ۲۰۱۴).

مطالعات گذشته نشان داده است در حالی که افزایش میزان پاسخ کار آسانی نیست، تأثیر آن کاملاً خارق‌العاده است. به عنوان مثال کونن، سوئینن، وانهوف و وتس (۲۰۰۰) اشاره کردند که حتی بهبود کمی از میزان پاسخ می‌تواند نتیجه کلی یک کمپین پستی مستقیم از شکست به موفقیت تغییر دهد. باسنس، ویائن، وان دل‌پوئل، وانتیئنن و ددن (۲۰۱۲) نشان دادند که چگونه بهبود کوچکی در نرخ پاسخ ممکن است سود عظیمی را به همراه داشته باشد. در مثال آن‌ها، تنها ۱ درصد افزایش در نرخ پاسخگویی برای یک شرکت سفارش پستی، ۵۰۰۰۰۰ یورو اضافه درآمد به‌همراه داشت. نوت، هیز و اسلین (۲۰۰۲) گزارش کردند که برای یک بانک خرده فروشی تنها ۷ درصد افزایش پاسخ کل، درآمد را به میزان سه برابر و درآمد هر پاسخ‌گو را به میزان ۲۰ درصد افزایش داد. سان و همکاران

1. Moro et al.

(۲۰۰۶) اشاره کردند که بهبود میزان پاسخ نه تنها می‌تواند سود را افزایش دهد، بلکه موجب تقویت وفاداری مشتری نیز می‌شود، زیرا مشتریانی که به طور صحیحی هدف قرار گرفته‌اند به احتمال بیشتری راضی می‌شوند و مدت زمان طولانی‌تری همراه شرکت می‌ماند.

بر اساس تحقیقات انجام شده بر روی دو مجموعه داده‌های واقعی مشتری اثر خوشه‌بندی، کم نمونه‌برداری و ترکیب[۱] به شرح زیر گزارش شده است: اول، به وضوح باعث بهبود ضریب تصحیح متعادل[۲] شده است. این موفقیت با افزایش نرخ پاسخ واقعی[۳]، با بالا نگه‌داشتن نرخ عدم پاسخ واقعی[۴]، حاصل شده است. دوم، مدل پاسخ را پایدارتر از روش‌های متعادل کننده دیگر ساخته است. سوم، انحراف استاندارد ضریب تصحیح متعادل را در اکثر موارد حداقل ۵۰٪ کاهش داده است. نسبت به چهار شیوهٔ طبقه بندی متفاوت، خوشه‌بندی، کم نمونه‌برداری و ترکیب عملکرد فوق العاده‌ای را از لحاظ ضریب تصحیح متعادل، تغییرات آن و ارتقاء معیارهای ارزیابی مدل به دست آورده است. چهارم، خوشه‌بندی، کم نمونه‌برداری و ترکیب به افزایش سود بیش از هر روش دیگر کمک کرده است (کان، چو و مک لخلان، ۲۰۱۲).

روش‌شناسی پژوهش

این پژوهش از لحاظ هدف، کاربردی و از لحاظ نحوهٔ گردآوری داده‌ها پژوهشی توصیفی، علی – مقایسه‌ای[۵] با روش ایجاد گروه‌های همگن می‌باشد. جامعهٔ آماری این پژوهش شامل ۱۰۰۰۰ نفر از مشتریان یک مؤسسه بانکی در کشور پرتغال[۶] است. در کارهای پژوهشی (مورو و همکاران، ۲۰۱۴) و (سرجیو مورو، لورئانو و کورتز، ۲۰۱۱) نیز از این دیتاست استفاده شده است و فایل داده‌های مربوطه به جهت استفادهٔ سایر محققین در دسترس است[۷]. روش نمونه‌گیری در این پژوهش به صورت تصادفی ساده است. نمونهٔ آماری پژوهش شامل ۱۹۸۳ نفر از مشتریان بانک است که با استفاده از فرمول

1. Clustering, Undersampling, And Ensemble (CUE)
2. Balanced Correction Rate (BCR)
3. True Respondents (TR)
4 True Non-Respondents (TN)
5 causal Comparative
6 .Portuguese banking institution
7. https://archive.ics.uci.edu

کوکران با جامعهٔ محدود بدست آمده است (حافظنیا، ۱۳۸۹: ۱۴۴):

$$n = \frac{NZ^2pq}{Nd^2 + Z^2pq} = \frac{10000 \times (1.96)^2 \times 0.05 \times 0.05}{10000 \times (0.01)^2 + (1.96)^2 \times 0.05 \times 0.05} = 1983$$

که در آن N حجم کل جامعهٔ آماری، Z ضریب اطمینان (در سطح معنی‌داری ۰/۰۵ برابر با ۱/۹۶ است)، p احتمال وجود صفت در جامعهٔ آماری (در اینجا پاسخ مشتری)، q احتمال عدم وجود صفت در جامعهٔ آماری و d خطای نمونه‌گیری (در اینجا ۱ درصد) است. شیوه بازاریابی مستقیم بدین‌صورت بوده است که با این مشتریان جهت گشایش یک حساب سپرده مدت‌دار تماس برقرار شده است. در نهایت پس از کسر ۶۶ داده گم‌شده، تعداد ۹۷ نفر به کمپین دریافت شده پاسخ داده و ۱۸۲۰ نفر هیچ پاسخی ارائه نکرده‌اند. در این پژوهش با استفاده از داده‌های فوق اقدام به ارائه الگوریتمی برای پیش‌بینی و در نهایت استخراج الگویی برای پیش‌بینی پاسخ سایر مشتریان در جامعهٔ آماری (۸۰۸۳) نموده‌ایم.

همان‌گونه که در ابتدای مقاله عنوان شد، استفاده از مدل‌های پیش‌بینی مختلف با داده‌های اولیه به ارائه یک نتیجه واحد منجر شده است و آن اینکه هیچکدام از مشتریان به عنوان یک مشتری بالقوه شناسایی نشده‌اند. این در واقع مشکلی است که نشأت گرفته از عدم توازن کلاس مشتریان است. به منظور فائق آمدن بر این مشکل، در این مرحله اقدام به توسعه و معرفی الگوریتمی اثربخش شده است، به نحوی که پس از گروه‌بندی بازار هدف به دو گروه از مشتریانی که به فعالیت‌های بازاریابی مستقیم شرکت پاسخ داده و نداده‌اند، اقدام به ایجاد دسته‌هایی مرکب از خوشه‌های مختلف مشتریان غیر پاسخ‌گو و پاسخ‌گو نموده و سپس پاسخ مشتریان را به‌وسیله مدل‌های مختلف طبقه‌بند، پیش‌بینی و در نهایت عملکرد آن‌ها را مورد بررسی و تحلیل قرار خواهد گرفت.

در این مرحله مشتریان غیر پاسخ‌گو را با مشتریانی که پاسخ داده‌اند (۹۷ نفر) به چند شیوه ترکیب می‌شود. یادآوری می‌شود که خوشه‌بندی مشتریان غیر پاسخ‌گو به منظور مد نظر قرار دادن مشتریان (نمایندگانی) از تمامی خوشه‌های متفاوت مشتریان است. ولی سوالی که در اینجا مطرح می‌شود این است که از هر خوشه چه تعداد مشتری (و به چه شیوه‌ای) انتخاب شده و چگونه با مشتریان پاسخ‌گو ترکیب شوند؟

بدین‌منظور ضمن مد نظر قرار دادن روش‌های بیش‌نمونه‌گیری و کم‌نمونه‌گیری تصادفی که در پژوهش‌های پیشین به کرات استفاده شده است سه راه حل ابتکاری به شرح زیر معرفی می‌گردد:

- به منظور مورد توجه قرار دادن حجم خوشه‌ها، تعداد اعضای انتخاب شده از هر خوشه برابر با نسبت حجم خوشه به کل مشتریان غیر پاسخ‌گو باشد. بدین ترتیب اعضاء نمونه انتخاب شده از مشتریان غیر پاسخ‌گو متناسب با حجم خوشهٔ عضویت آن‌ها خواهد بود و نیز عدم تعادل کلاس هم از بین خواهد رفت. این روش را (C-RUS-01) نامگذاری کرده و در تحلیل‌های مقاله مورد استفاده قرار خواهد گرفت.
- از ترکیب روش‌های خوشه‌بندی، کم نمونه‌گیری تصادفی کلاس اکثریت[1] و بیش نمونه‌گیری تصادفی کلاس اقلیت[2] استفاده نماییم. در واقع هدف از معرفی این روش، پوشش ضعف هر یک از روش‌های اشاره شدهٔ پیشین است. روش پیشنهادی بدین صورت است که (در راستای استفاده از مزیت بیش نمونه‌گیری تصادفی) ابتدا مشتریان پاسخ‌گو را ضربدر تعداد خوشه‌های شناسایی شده از مشتریان غیر پاسخ‌گو نموده (در اینجا ۳۸۸=۴×۹۷) و به همان تعداد از مشتریان غیر پاسخ‌گو اقدام به نمونه‌گیری شود. نمونه‌گیری از مشتریان غیر پاسخ‌گو متشکل از کلیه خوشه‌ها و به نسبت حجم آن‌ها خواهد بود. این روش را (C-ROS-RUS-02) نامگذاری کرده و در تحلیل‌های مقاله مورد استفاده قرار خواهد گرفت.
- از ترکیب روش‌های خوشه‌بندی کم نمونه‌گیری تصادفی و بیش نمونه‌گیری تصادفی استفاده می‌شود، بدین‌نحو که کوچکترین خوشهٔ شناسایی شده ملاک قرار داده شود. از سایر خوشه‌ها به اندازهٔ کوچک‌ترین خوشهٔ شناسایی شده نمونه‌گیری کرده و اندازه کلاس اقلیت را به اندازه جمع نمونه‌های اخذ شده بسط داده شود. این روش را (C-ROS-RUS-03) نامگذاری کرده و در تحلیل‌های مقاله مورد استفاده قرار خواهد گرفت.

نتایج حاصل از نمونه‌برداری به روش‌های بدون نمونه‌گیری مجدد[3]، کم نمونه‌گیری تصادفی و بیش نمونه‌گیری تصادفی و راه حل‌های ابتکاری ارائه شده در جدول (۱) نشان داده شده است.

1. Random Under-sampling (RUS)
2. Random Over-sampling (ROS)
3. No Sampling (NS)

جدول ۱. نمونه‌ها و مدل‌های مورد استفاده جهت پیش‌بینی پاسخ مشتریان

ردیف	عنوان روش	اختصار استفاده شده	ترکیب مشتریان					حجم نمونه
			C4	C3	C2	C1	R	
۱	بدون نمونه‌گیری مجدد	NS	۱۸۲۰				۹۷	۱۹۱۷
۲	کم نمونه‌گیری تصادفی کلاس اکثریت	RUS	۹۷				۹۷	۱۹۴
۳	بیش نمونه‌گیری تصادفی کلاس اقلیت	ROS	۱۸۲۰				۱۸۲۰	۳۶۴۰
۴	روش ابتکاری اول	C-RUS-01	۲۴	۲۵	۲۲	۲۶	۹۷	۱۹۴
۵	روش ابتکاری دوم	C-ROS-RUS-02	۹۷	۱۰۱	۸۸	۱۰۲	۳۸۸	۷۷۶
۶	روش ابتکاری سوم	C-ROS-RUS-03	۴۱۲	۴۱۲	۴۱۲	۴۱۲	۱۶۴۸	۳۲۹۶

معیارهای ارزیابی نقشی اساسی در مسائل طبقه‌بندی ایفا می‌کنند. برای ارزیابی کارایی طبقه‌بندها، ساده‌ترین روش تحلیل بر اساس ماتریس اغتشاش[1] (ماتریس درهم ریختگی) است. این ماتریس از ابزارهای مناسب برای بررسی میزان موفقیت و کارایی سیستم‌های طبقه‌بندی محسوب می‌شود و در واقع نمونه‌هایی را که برای هر طبقه به درستی و به اشتباه تشخیص داده شده‌اند گزارش می‌کند که در آن مثبت حقیقی[2]: تعداد نمونه‌های طبقهٔ مثبت که به درستی طبقه‌بندی شده‌اند، منفی حقیقی[3]: تعداد نمونه‌های طبقهٔ منفی که به درستی طبقه‌بندی شده‌اند، مثبت کاذب[4]: تعداد نمونه‌های طبقه منفی است که به اشتباه مثبت پیش‌بینی شده‌اند و منفی کاذب[5]: تعداد نمونه‌های طبقه مثبت است که به اشتباه منفی پیش‌بینی شده‌اند، است. رایج‌ترین معیار مستخرج از این ماتریس، معیار دقت[6] است. معیار دقت، بیان می‌کند دو مقدار مثبت حقیقی و منفی حقیقی مهم‌ترین مقادیری هستند که در یک مسئله دو دسته‌ای باید بیشینه شوند. رابطه سایر معیارهای استفاده شده در پژوهش حاضر به شرح زیر است:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad Sensivity = \frac{TP}{TP+FN}$$

1. Confusion matrix
2. True Positive (TP)
3. True Negative (TN)
4. False Positive (FP)
5. False Negative
6. Accuracy

$$Specificity = \frac{TN}{TN+FP} \qquad Precision = \frac{TP}{TP+FP}$$
$$G-Mean = \sqrt{Precision.Recall} \qquad F-Measure = \frac{2.Recall.Precision}{Recall+Precision}$$

از معیار سطح زیر منحنی[1] نیز که نشان دهنده سطح زیر منحنی مشخصهٔ عملکرد سیستم[2] است استفاده شده است. هر چه مقدار این عدد مربوط به یک دسته‌بند بزرگتر (نزدیک به یک) باشد کارایی نهایی دسته‌بند مطلوب‌تر ارزیابی می‌شود.

یافته‌ها

اعداد ارائه شده در جدول (۲) مرتبط با معیارهای ارزیابی هر مدل و نمونهٔ مرتبط با آن است. هدف اصلی این تحقیق تهیهٔ نمونه‌ها و طراحی الگوریتمی کارا به منظور استفاده بهینه از مدل‌های پیش‌بینی و نیل به نرخ بالایی از پاسخ مشتری است. همان‌گونه که از جدول (۲) می‌توان فهمید، نمونهٔ اولیه به هیچ وجه توانایی پیش‌بینی مشتریان بالقوه را نخواهد داشت.

شایان ذکر است سطح زیر منحنی‌های راک هرچقدر به عدد ۱ نزدیک‌تر باشد بیان‌گر توان و معیار بهینه‌ای از پیش‌بینی الگوریتم طبقه‌بند است که با اختصار سطح زیر منحنی در جدول (۲) نشان داده شده است. با توجه به معیارهای ارائه شده و سطح زیر منحنی نمودارهای راک مشخص است که نمونهٔ C-ROS-RUS-03 بهترین پیش‌بینی را با طبقه‌بند شبکه عصبی مصنوعی بعمل آورده است.

جدول ۲. نتایج حاصل معیارهای ارزیابی طبقه‌بندی مشتریان با توجه به روش‌های مختلف نمونه‌گیری

مدل	معیارهای ارزیابی مدل		NS	RUS	ROS	C-RUS-01	C-ROS-RUS-02	C-ROS-RUS-03
رگرسیون لجستیک	Accuracy		0.949	0.613	0.578	0.695	0.592	0.588
	Sensitivity		0	0.649	0.596	0.752	0.618	0.618
	Specificity		1	0.577	0.559	0.639	0.567	0.558
	Precision		0	0.605	0.575	0.675	0.588	0.583
	G-mean		0	0.612	0.577	0.693	0.592	0.587
	F-measure		0	0.626	0.585	0.712	0.602	0.600
	ROC	AUC	0.500	0.613	0.578	0.696	0.593	0.589
		Std. Error	0.030	0.040	0.009	.0038	0.020	0.010
		Sig	1.000	0.006	0.000	0.000	0.000	0.000

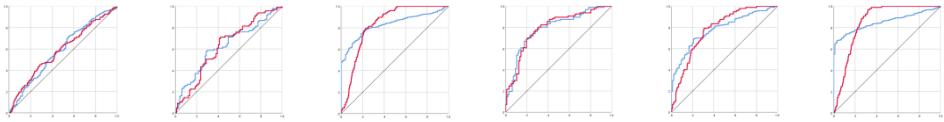
1. Area Under Curve (AUC)
2. Receiver Operating Characteristic (ROC)

	Chi-square		9.764	7.156	89.415	19.380	25.222	87.497
	Sig		0.202	0.209	0.000	0.000	0.000	0.000
	R Square		0.015	0.048	0.032	0.127	0.043	0.035
شبکه عصبی مصنوعی	Accuracy		0.936	0.561	0.759	0.706	0.687	0.790
	Sensitivity		0	0.655	0.807	0.774	0.823	0.837
	Specificity		0.936	0.464	0.709	0.629	0.556	0.742
	Precision		0	0.558	0.738	0.705	0.640	0.766
	G-mean		0	0.551	0.757	0.698	0.676	0.788
	F-measure		0	0.603	0.771	0.738	0.720	0.800
	ROC	AUC	0.500	0.536	0.768	0.742	0.722	0.794
		Std. Error	0.037	0.041	0.008	0.036	0.019	0.008
		Sig	1.000	0.385	0.000	0.000	0.000	0.000
درخت تصمیم	Estimate	Risk	0.051	0.464	0.371	0.376	0.393	0.316
		Std. Error	0.005	0.036	0.008	0.035	0.018	0.008
	Accuracy		0.944	0. 536	0. 628	0.623	0.606	0.684
	Sensitivity		0	0. 896	0. 782	0.560	0.701	0.762
	Specificity		1	0. 175	0. 474	0.680	0.512	0.605
	Precision		0	0. 520	0. 598	0.639	0.590	0.659
	G-mean		0	0. 157	0. 609	0.621	0.599	0.679
	F-measure		0	0. 659	0. 678	0.601	0.640	0.707
	ROC	AUC	0.500	0.536	0.629	0.624	0.607	0.684
		Std. Error	0.030	0.041	0.009	0.040	0.020	0.009
		Sig	1.000	0.385	0.000	0.003	0.000	0.000

در این پژوهش به‌صورتی ابتکاری اقدام به محاسبه معیار جدیدی شده است که از حاصل‌ضرب هفت معیار فوق به‌دست می‌آید، بدین‌معنی که اگر تمامی معیارهای پیش‌گفته مقدار مناسب و مطلوبی را ارائه دهند حاصل‌ضرب آن‌ها نیز مقداری بالاتری خواهد شد. این معیار را تحت عنوان «حاصل‌ضرب[۱]» به شرح زیر معرفی شده است؛

Multiple = (Accuracy × Sensitivity × Specificity × Precision × G-mean × F-measure × AUC)

مقدار این معیار برای ۱۸ نمونه–روش مختلف پیش‌بینی استفاده شده در جدول (۳) نشان داده شده است. در نهایت از معیار «حاصل‌ضرب» به عنوان نشان‌گری برای مطلوبیت پیش‌بینی طبقه‌بندها با هر نمونهٔ تولید شده استفاده خواهد شد. منحنی‌های راک با توجه به شش نمونهٔ مورد بررسی برای شبکه‌های عصبی مصنوعی به‌صورت زیر ترسیم شده است:



شکل ۱. منحنی‌های راک مربوط به نمونه‌های مورد بررسی

جدول ۳. حاصل‌ضرب به‌دست آمده از هفت معیار ارزیابی مدل‌های طبقه‌بندی و مجموع آن‌ها

مدل	NS	RUS	ROS	C-RUS-01	C-ROS-RUS-02	C-ROS-RUS-03	جمع
رگرسیون لجستیک	0	0.032	0.021	0.077	0.025	0.024	0.218
شبکه عصبی	0	0.016	0.143	0.092	0.070	0.188	0.566
درخت تصمیم	0	0.002	0.036	0.035	0.029	0.068	0.171
جمع ضرایب	0	0.051	0.201	0.205	0.126	0.280	0.956

با توجه به نمودار (۲) و نیز توضیحات ارائه شده، مشخص است که شبکه عصبی و درخت تصمیم بهترین پیش‌بینی خود را با نمونهٔ C-ROS-RUS-03 و رگرسیون لجستیک بهترین پیش‌بینی خود را با نمونهٔ C-RUS-01 انجام داده است. حال سوالی که مطرح می‌شود این است؟ بالاخره کدام نوع نمونه‌برداری از هر حیث بهترین شیوه نمونه‌برداری و ترکیب مشتریان برای آموزش بهینهٔ مدل‌های پیش‌بینی است؟ کدام مدل پیش‌بینی قوی‌ترین مدل برای پیش‌بینی مشتریان پاسخ‌گو است؟ آیا یک بهترین مدل و یا یک بهترین روش نمونه‌برداری برای این منظور وجود دارد؟ این سوالاتی است که ما را برای طراحی الگوریتم بهینه‌سازی که هدف پژوهش جاری است رهنمون می‌سازد. الگوریتم موصوف به شکل زیر ارائه شده است و در نهایت فلوچارت الگوریتم مربوطه در بخش پیشنهادات کاربردی به عنوان یافته اصلی پژوهش طراحی و معرفی شده است.

الگوریتم پویای پیشنهادی

شروع.

انتخاب زیر مجموعهٔ ویژگی‌ها[1] (FSS)، حذف ویژگی‌های غیرمرتبط با استفاده از روش Embeded (رگرسیون لجستیک).

اگر دقت مجموعهٔ انتخاب شده حداکثر نباشد برو به مرحله ۲.

تقسیم مجموعهٔ داده‌ها به داده‌های آموزشی و آزمایشی.

ورود داده‌های آموزشی به مدل‌های شبکه عصبی، درخت تصمیم و رگرسیون لجستیک.

آموزش مدل‌ها از طریق داده‌های آموزشی.

تشکیل ماتریس آشفتگی برای هر مدل.

محاسبه معیارهای ارزیابی برای هر یک از مدل‌ها.

اعتبارسنجی نتایج هر مدل.

اگر مقدار خطای مدل‌ها بیشتر از یک مقدار قابل قبول باشد برو به مرحله ۶.

محاسبه مضروب (حاصل‌ضرب) دقت روش‌های طبقه‌بندی با استفاده از معیار جدید.

انتخاب مدل طبقه بندی باقدرت پیش‌بینی نزدیک به بهینه

پایان.

الگوریتم پیشنهادی در مرحله اول از یک گام پیش پردازشی به منظور آماده‌سازی داده‌ها و از بین بردن ویژگی‌های نامرتبط با برچسب داده‌ها استفاده می‌کند. در مرحله بعدی داده‌ها با استفاده از سه طبقه بند شبکه عصبی (جود، ۱۹۹۰)، درخت تصمیم (باهرمن و ولف، ۲۰۰۲) و رگرسیون لجستیک (بیندر، ۱۹۸۳) استفاده شده است. در واقع الگوریتم پیشنهادی از این سه روش به صورت موازی برای پیش‌بینی پاسخ مشتریان استفاده می‌کند. از آنجایی که این صحت روش‌های طبقه‌بندی فوق‌الذکر اثبات شده است، بر این اساس صحت الگوریتم پیشنهادی نیز به صورت پیش فرض به اثبات می‌رسد.

همچنین در مورد پیچیدگی روش پیشنهادی می‌توان به قضیهٔ زیر اشاره کرد.

1. Feature Subset Selection

پیچیدگی الگوریتم پویای پیشنهادی از مرتبه $O(n^2)$ است.

اثبات: با توجه به این که روش پیشنهادی در گام پیش‌پردازشی ویژگی‌های مربوط به n داده را بررسی نموده و در هر گام از تعداد ویژگی‌های این مجموعه داده کاسته می‌شود، پیچیدگی زمانی گام پیش پردازشی را می‌توان از مرتبه O(nlogn) یاد کرد. از سوی دیگر روش پیشنهادی از ترکیب شبکه‌های عصبی با پیچیدگی $O(n^2)$ و درخت تصمیم با پیچیدگی O(n) و رگرسیون لجستیک خطی با پیچیدگی O(n) در مرحله آموزش تشکیل شده است. از آنجایی در مرحله آموزش سه روش طبقه‌بندی به صورت موازی استفاده شده است، پیچیدگی فاز آموزش را می‌توان از مرتبه $O(n^2)$ عنوان کرد. در نهایت پیچیدگی کل روش پیشنهادی به صورت زیر محاسبه می‌شود.

$$T(n)=\begin{cases} O(\text{nlogn}) <= O(n2) \\ O(n^2) <= O(n^2) \end{cases} <= O(n^2)$$

بحث و نتیجه‌گیری

در پژوهش حاضر خوشه‌بندی مشتریان غیرپاسخ‌گو و ترکیب آن‌ها با روش‌های مختلف با مشتریان پاسخ‌گو، به عنوان ابزاری در جهت متعادل‌سازی پایگاه داده مشتریان معرفی شد. بدیهی است استفاده از شیوه‌های مختلف خوشه‌بندی، متغیرهای مختلف برای انجام این کار، تعداد خوشه‌های شناسایی یا تعیین شده و روش‌های مختلف ترکیب اعضاء خوشه‌ها به طرق مختلف و بسیار متنوعی قابل اجراست که به جهت جلوگیری از تفصیل مقاله از ارایهٔ آن‌ها خودداری شد. هدف پژوهش حاضر دست‌یابی به بهینه‌ترین پیش‌بینی برای یک جامعهٔ خاص یا مورد بررسی در این مقاله نمی‌باشد، بلکه هدف اصلی ارایهٔ یک الگو و الگوریتمی ابتکاری است که به‌صورتی پویا برنامه‌ریزی شده و بهترین روش ترکیب و نمونه‌برداری را جهت پیش‌بینی مشتریان در هر جامعه یا بازاری انجام دهد. به‌طور خلاصه، لزوم ارزیابی مدل‌های مختلف خوشه‌بندی و تحلیل خوشه‌های به‌دست آمده از هر روش، انتخاب و ارزیابی روش‌های مختلف ترکیب خوشه‌ها با مشتریان پاسخ‌دهنده و در نهایت استفاده از ابزارهای مختلف پیش‌بینی از قبیل رگرسیون لجستیک، شبکه عصبی مصنوعی، درخت تصمیم و ... از جملهٔ گام‌های مهم الگوریتم پیشنهاد شده توسط محقق است که در بخش پیشنهادات کاربردی به‌طور خلاصه نشان داده شده است. جدول (۴) مقایسه‌ای بین یافتهٔ روش پیشنهادی این مقاله با روش اجرا شده در پژوهش (Kang et al. 2012) ارائه کرده است:

جدول ۴. مقایسه یافته‌های مقاله با روش اجرا شده در پژوهش (Moro et al. 2014)

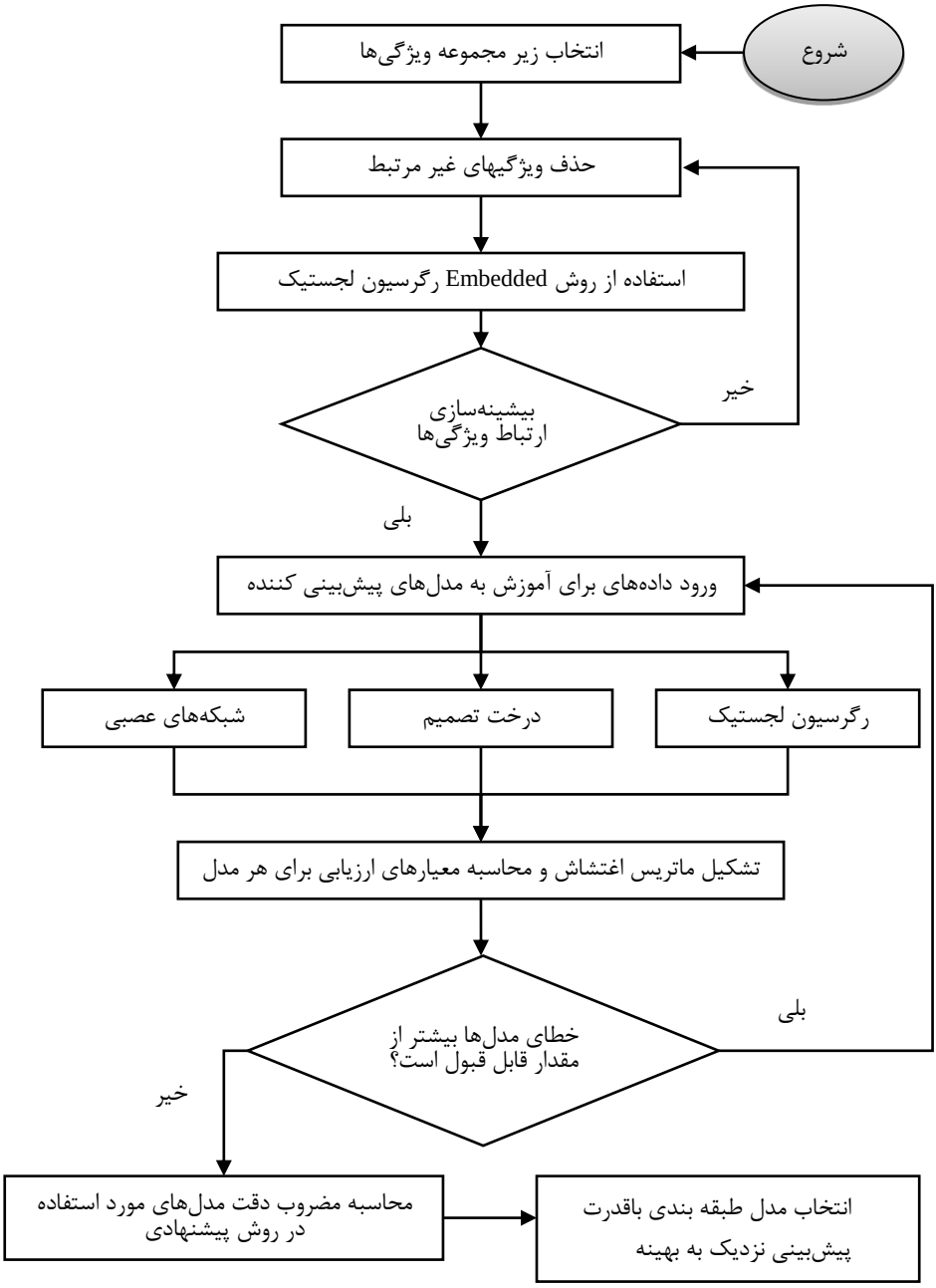
پژوهش	معیار ارزیابی	رگرسیون لجستیک				شبکه عصبی			
		NS	RUS	ROS	CUE	NS	RUS	ROS	CUE
کانگ و همکاران ۲۰۱۲	ACC	۰/۹۳۹	۰/۶۴۱	۰/۶۹۳	۰/۶۹۹	۰/۹۳۸	۰/۵۹۸	۰/۷۴۸	۰/۶۷۳
	G-Mean	۰/۱۲۹	۰/۶۴۸	۰/۶۵۴	۰/۶۷۳	۰/۰۷۱	۰/۶۱۴	۰/۵۸۵	۰/۶۶۲
پژوهش حاضر	ACC	۰/۹۴۹	۰/۶۱۳	۰/۵۷۸	۰/۶۹۵	۰/۹۳۶	۰/۵۶۱	۰/۷۵۹	۰/۷۹۰
	G-Mean	۰	۰/۶۱۲	۰/۵۷۷	۰/۶۹۳	۰	۰/۵۵۱	۰/۷۵۷	۰/۷۸۸

نتایج مقایسهٔ نتایج پژوهش حاضر با پژوهش (Kang et al. 2012) حاکی از آن است که الگوریتم ارائه شده در این پژوهش با چند نمونه‌گیری مجدد و یک جستجو توانایی ارتقاء معیارهای ACC و G-Mean را با استفاده از مدل شبکه عصبی و ارتقاء معیار G-Mean را با استفاده از مدل رگرسیون لجستیک دارد و همان‌گونه که در بخش ارائه الگوریتم اشاره شد، پیچیدگی سیستم برابر با $O(n^2)$ است.

مهم‌ترین نکته‌ای که به عنوان پیشنهادات کاربردی این پژوهش قابل ذکر است آن است که یک بهترین روش برای پیش‌بینی مشتریان بالقوه وجود ندارد، بلکه استفاده از ترکیبی از روش‌ها و به منظور ایجاد نمونه‌های مختلف از مشتریان جهت کاهش عدم توازن کلاس مشتریان پاسخ‌گو می‌تواند بهترین نتایج را برای بازاریابان به ارمغان بیاورد. فلوچارت الگوریتم توضیح داده در فرایند پژوهش حاضر در شکل (۳) ترسیم شده است.

روش کار الگوریتم معرفی شده بدین نحو است که ابتدا کل مشتریان موجود در پایگاه داده به دو دستهٔ آموزش و آزمایش تفکیک می‌شود. بدون نمونه‌گیری مجدد دستهٔ آزمایش را تحت تأثیر یک کمپین بازاریابی قرار داده و با توجه به نتایج مشاهده شده به پیش بینی پاسخ دستهٔ مشتریان آزمایش پرداخته می‌شود. اگر نتایج پیش‌بینی شده مطلوب بود (که طی مطالعات و تجربیات به‌دست آمده بعید است) فرایند کار به اتمام رسیده است، در غیر این‌صورت مشتریان دستهٔ آموزش به دو گروه پاسخ‌گویان و غیر پاسخ‌گویان طبقه‌بندی می‌شود. در این مرحله اقدام به نمونه‌برداری مجدد به روش کم نمونه‌برداری کلاس اکثریت (مشتریان غیر پاسخ‌گو) و بیش نمونه‌برداری کلاس اقلیت (مشتریان پاسخ‌گو) نموده و مجدداً اقدام به محاسبه و ارزیابی مدل‌های مختلف پیش‌بینی (در این پژوهش مدل‌های رگرسیون لجستیک، شبکه عصبی مصنوعی و درخت تصمیم) می‌شود. در مرحله بعد اقدام به خوشه‌بندی مشتریان غیرپاسخ‌گو نموده و با سه روش پیشنهادی مشتریان غیرپاسخ‌گو در خوشه‌های

مختلف شناسایی شده را با مشتریان پاسخ‌گو ترکیب نموده و مجدداً مدل‌های پیش‌بینی را در مورد آن‌ها به اجرا در آورده می‌شود. همان‌گونه که پیش‌تر یادآوری شده است این امر به‌خاطر طبقه‌بندی و خوشه‌بندی مشتریان غیرپاسخ‌گو در خوشه‌های مختلف و متمایز است؛ زیرا همانگونه که از تحقیقات پیشین مطرح شده است قرار دادن مشتریان غیر پاسخ‌گو در یک دسته به صرف پاسخ ندادن‌شان به کمپین بازاریابی بسیار ساده لوحانه است. به عبارتی مشتریانی که از ما خرید نکرده‌اند صرف خرید نکردن‌شان نباید به عنوان یک طیف مستقل و یکسان مدنظر قرار گیرند بلکه آن‌ها نیز دارای وجوه تشابه و تمایز زیادی هستند که با ابزار خوشه‌بندی می‌توان بدآن‌ها پی برد و در نهایت ترکیب‌های بهینه‌ای از آن‌ها را با مشتریان پاسخ‌گو به‌وجود آورد. در این مرحله نیز ترکیب‌ها یا نمونه‌های به‌دست آمده به‌وسیله مدل‌های مختلف مورد پیش‌بینی قرار گرفته و مشتریان محتمل شناسایی می‌شوند. در نهایت به شیوه‌های مختلف اقدام به آموزش طبقه‌بندهای پیش‌بینی نموده و نتایج در هر مرحله در فایل‌های XML خروجی ذخیره می‌شود. نمونه‌ای از خروجی کدهای XML در مراحل مختلف پژوهش جهت پیش‌بینی پاسخ مشتریان نمونهٔ آزمایش (۸۰۸۳ نفر) در پیوست (۱) ارائه شده است. اصلی‌ترین خروجی پژوهش پس از استخراج کدهای XML مربوط به هر نمونه و در هر مرحله شناسایی مشتریان محتمل خواهد بود. ولی به این امر اکتفا نشده است و با استفاده از معیار ضریب به‌دست آمده برای هر مرحله خروجی نرم‌افزار از حالت باینری به فازی تبدیل می‌شود، بدین شرح که هر مشتری (از بین ۸۰۸۳ مشتری پژوهش حاضر) که به عنوان مشتری پاسخ‌گو در هر مرحله شناسایی می‌شود یک امتیاز گرفته و امتیاز مربوطه ضرب‌در معیار «حاصل‌ضرب» مربوط به آن مرحله از پیش‌بینی می‌شود. در پژوهش حاضر در بهترین حالت، اگر یک مشتری بتواند در تمامی نمونه‌های هفت‌گانه به عنوان مشتری بالقوه شناسایی شود (که بسیار به ندرت اتفاق می‌افتد) امتیازی برابر با ۰/۱۷ کسب خواهد کرد: (۰/۱۷= ۰/۰۶۸ + ۰/۰۲۹ + ۰/۰۳۵ + ۰/۰۳۶ + ۰/۰۰۲+۰/۰۰۰). اگر این مشتری توسط هر سه طبقه‌بند و در هر شش نمونه به‌دست آمده (در تست ۱۸ گانهٔ پژوهش) به عنوان مشتری بالقوه شناسایی شود امتیازی برابر با ۰/۵۱ کسب خواهد کرد؛ بنابراین کلیه مشتریان مورد آزمایش (۸۰۸۳ نفر) امتیازی مابین صفر تا ۰/۵۱ به‌دست خواهند آورد.



شکل ۳. فلوچارت الگوریتم پیشنهادی پژوهش

در نهایت پیشنهاد می‌شود تمامی مشتریان موجود در پایگاه داده مشتریان با توجه به امتیازات کسب شده رتبه‌بندی شده و به ترتیب اولویت مورد هدف بازاریابان شیوهٔ مستقیم قرار گیرند؛ بنابراین نتایج این پژوهش ما را به سمتی رهنمون ساخته است تا علاوه بر ارتقاء توانایی پیش‌بینی طبقه‌بندهای مختلف با متوازن‌سازی داده‌ها و استفاده از مدل‌های مختلف پیش‌بینی یکی دیگر از نواقص این طبقه‌بندها را برطرف نماییم و آن تبدیل خروجی پیش‌بینی از حالت باینری به فازی است. به عنوان پیشنهاد کاربردی دیگر چنین مطرح نمود که مدیران بازاریابی پس از اجرا و حصول نتایج این پژوهش می‌توانند مشتریان را در طیف‌های گوناگونی دسته‌بندی کنند.

در زمینهٔ تحقیقات آتی به محققان پیشنهاد می‌شود، ضمن تلاش در جهت تشخیص مشتریان محتمل خرید و یا به عبارتی شناخت مشتریان وفادار اقدام به شناسایی مشتریان ارزشمند نمایند که معمولاً ترکیبی از وفاداری و سودمندی را در کارنامه خودشان دارند.

پیشنهاد می‌شود در تحقیقات آتی که به روش مشابهی صورت می‌پذیرد، عامل هزینه خطای عدم طبقه بندی صحیح (هزینه تماس با مشتریانی که پاسخ نخواهند داد و هزینه عدم موفقیت در برقراری تماس با مشتریانی که پاسخ خواهند داد یعنی درصورت برقراری تماس خرید خواهند کرد) را نیز وارد مدل شود.

در پژوهش حاضر تأیید شد که نباید مشتریان غیر پاسخ‌گو را به عنوان یک گروه همگن در نظر گرفت، بلکه خوشه‌بندی و تفکیک آن‌ها و ترکیب آن‌ها با مشتریان پاسخ‌گو می‌تواند به طرز عجیبی توان پیش‌بینی و تشخیص مشتریان بالقوه را افزایش دهد. پیشنهاد می‌شود برای مطالعات آتی، اقدام مشابهی برای مشتریان پاسخ‌گو نیز انجام شود، زیرا تصور بر این است که مشتریان پاسخ‌گو نیز از یک گروه همگن تشکیل نشده‌اند، بنابراین تنوع‌بخشی و استفاده از ترکیب آن‌ها برای ایجاد نمونه‌های جدید چه بسا موجب تقویت بیشتر معیارهای ارزیابی الگوریتم‌های پیش‌بینی شود.

توصیه می‌شود در پژوهش‌های آتی مدیران و پژوهشگران اطلاعات و استنباط‌های کیفی و انتزاعی مبتی بر مضمون کسب کار و سایر شرایط و متغیرهای اثرگذار بر پیش‌بینی رفتار مصرف‌کننده را در راستای پیش‌بینی پاسخ آن‌ها مد نظر قرار دهند. به عنوان مثال دخالت عواملی چون شخصیت و خصوصیات و ویژگی‌های شخصی مصرف کننده و چگونگی تأثیر آن‌ها در فرایند پاسخ وی می‌تواند راه‌گشای مطالعات آتی باشد.

با توجه به الگوریتم معرفی شده در بخش پیشنهادات کاربردی، مشخص است کـه ادامـهٔ چرخـهٔ طراحی شده در هر مرحله به پیش‌بینی مشتریان و در نهایت مشاهدهٔ عملکرد واقعی آنان منجر خواهـد شد. اطلاعات ثبت شده در هر مرحله به تولید شاخص‌های جدید منجر خواهد شد که به عنوان خریدها و پیش‌بینی خریدهای مشتریان ثبت می‌گردد. فلذا پیشنهاد می‌شود پژوهشگران آتـی بـا تهیـهٔ چنـین بانک اطلاعاتی اقدام به تحلیل RFM مشتریان نمایند.

با توجه به اینکه شبکه‌های عصبی مصنوعی ابزاری قدرتمند در طبقه‌بندی و شناسایی مشتریان و نیز شناسایی متغیرهای مهم در راستای طبقه‌بندی آن‌ها می‌باشند، پیشنهاد می‌شود از نتـایج حاصـل از شبکه‌های عصبی در زمینهٔ پیش‌بینی مشتریان و شناسایی متغیرهای اثـربخش بـه عنـوان بـازخوردی جهت تهیه نمونه‌های ورودی استفاده شود. بدین نحو که متغیرهای تأثیرگذار در مدل‌های پـیش‌بینـی شبکهٔ عصبی به عنوان متغیرهای مهم در زمینهٔ خوشه‌بندی مشتریان غیرپاسخ‌گو در نظر گرفته شده و با توجه به این متغیرها، یک‌بار دیگر اقدام به خوشه‌بندی و ترکیب مشتریان غیرپاسخ‌گـو بـا مشـتریان پاسخ‌گو (برای هر روش نمونه‌گیری مجدد) به‌صورت مجزا نمایند.

منابع

بصیری، مهدی، «کاربرد تکنیک داده‌کاوی در مدیریت روابط مشتری»، همایش ملی تجـارت الکترونیکـی، وزارت صـنعت، معدن و تجارت، انجمن علمی تجارت الکترونیکی ایران، ۱۳۸۶، دوره ۴.

بون و کورتز (۲۰۱۳)، مدیریت بازاریابی نوین، ترجمه نوروزی، حسین، و مهدی مهذبی (۱۳۹۵)، تهران، انتشارات فوژان.

حافظ‌نیا، محمدرضا (۱۳۸۹)، مقدمه‌ای بر روش تحقیق در علوم انسانی، تهران، انتشارات سمت.

رضائی نوائی، سمیرا، کوشا، حمیدرضا، «به‌کارگیری و ارزیابی تکنیک‌های داده‌کاوی جهت پیش‌بینی رویگردانـی مشـتری در صنعت بیمه»، نشریهٔ بین المللی مهندسی صنایع و مـدیریت تولیـد، دوره (۲۷)، شـماره (۴)، سـال (۱-۲۰۱۷)، صفحات (۶۳۵-۶۵۳).

لینکستر و مسینگهام (۲۰۰۱)، اصول مدیریت بازاریابی، ترجمه نـوروزی، حسـین، و نیمـا سـلطانی‌نـژاد (۱۳۹۵)، تهـران، مؤسسه کتاب مهربان نشر.

Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. European Journal of Operational Research, 138(1), 191-211.

Barandelaa, R., Sanchezb, J., & Garcia, V. (2003). Strategies for learning in class imbalance problems.

Berry, M. J., & Linoff, G. (1997). Data mining techniques: for marketing, sales, and customer support: John Wiley & Sons, Inc.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review/Revue Internationale de Statistique, 279-292.

Błaszczyński, J., Dembczyński, K., Kotłowski, W., & Pawłowski, M. (2006). Mining direct marketing data by ensembles of weak learners and rough set methods. Paper presented at the International Conference on Data Warehousing and Knowledge Discovery.

Buhrman, H., & De Wolf, R. (2002). Complexity measures and decision tree complexity: a survey. Theoretical Computer Science, 288(1), 21-43.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview Data mining and knowledge discovery handbook (pp. 875-886): Springer.

Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2015). Behavior-aware user response modeling in social media: Learning from diverse heterogeneous data. European Journal of Operational Research, 241(2), 422-434.

Coenen, F., Swinnen, G., Vanhoof, K., & Wets, G. (2000). The improvement of response modeling: combining rule-induction and case-based reasoning. Expert Systems with Applications, 18(4), 307-313.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 463-484.

Gönül, F. F., & Hofstede, F. T. (2006). How to compute optimal catalog mailing decisions. Marketing Science, 25(1), 65-74.

Hill, S., Provost, F., & Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. Statistical Science, 21(2), 256-276.

Judd, J. S. (1990). Neural network design and the complexity of learning: MIT press.

Kang, P., Cho, S., & MacLachlan, D. L. (2012). Improved response modeling based on clustering, under-sampling, and ensemble. Expert Systems with Applications, 39(8), 6738-6753.

Knott, A., Hayes, A., & Neslin, S. A. (2002). Next-product-to-buy models for cross-selling applications. Journal of interactive Marketing, 16(3), 59-75.

Lai, Y.-T., Wang, K., Ling, D., Shi, H., & Zhang, J. (2006). Direct marketing when there zre voluntary buyers. Paper presented at the Data Mining, 2006. ICDM'06. Sixth International Conference on.

Li, D.-C., Liu, C.-W., & Hu, S. C. (2010). A learning method for the class imbalance problem with medical data sets. Computers in biology and medicine, 40(5), 509-518.

Ling, C. X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. Paper presented at the Kdd.

Linoff, G. S., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship management: John Wiley & Sons.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. Paper presented at the Proceedings of European Simulation and Modelling Conference-ESM'2011.

Napierała, K., Stefanowski, J., & Wilk, S. (2010). Learning from imbalanced data in presence of noisy and borderline examples. Paper presented at the International Conference on Rough Sets and Current Trends in Computing.

Sun, B., Li, S., & Zhou, C. (2006). "Adaptive" learning and "proactive" customer relationship management. Journal of Interactive Marketing, 20(3-4), 82-96.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687-719.

Wilkinson, T. J., McAlister, A., & Widmier, S. (2007). Reaching the international consumer: An assessment of the international direct marketing environment. Direct Marketing: An International Journal, 1(1), 17-37.

Yan, R., Liu, Y., Jin, R., & Hauptmann, A. (2003). On predicting rare classes with SVM ensembles in scene classification. Paper presented at the Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on.

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making, 5(04), 597-604.

Zhang, S., Liu, L., Zhu, X., & Zhang, C. (2008). A strategy for attributes selection in cost-sensitive decision trees induction. Paper presented at the Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on.

پیوست

نمونه‌ای از کدهای XML استخراج شده جهت شناسایی مشتریان خارج از کمپین بازاریابی (۸۰۸۳ نفر)


```
<?xml version="1.0"?>
<PMML xsi:schemaLocation="http://www.dmg.org/PMML-4_3 pmml-4-3.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.dmg.org/PMML-4_3"
version="4.3">
 - <Header copyright="Copyright (c) IBM Corp. 1989, 2017.">
      <Application version="25.0.0.0" name="IBM SPSS Statistics"/>
      <Timestamp/>
   </Header>
 - <DataDictionary numberOfFields="8">
      <DataField name="Children" dataType="double" isCyclic="0" optype="continuous"
         displayName=""/>
      <DataField name="Gender" dataType="double" isCyclic="0" optype="categorical" displayName=""/>
    - <DataField name="Education" dataType="double" isCyclic="0" optype="ordinal" displayName="">
         <Value property="valid" displayValue="Some high school or less" value="1"/>
         <Value property="valid" displayValue="High school" value="2"/>
         <Value property="valid" displayValue="Some college" value="3"/>
         <Value property="valid" displayValue="College" value="4"/>
         <Value property="valid" displayValue="Post-graduate" value="5"/>
      </DataField>
      <DataField name="Age" dataType="double" isCyclic="0" optype="continuous" displayName=""/>
      <DataField name="Reside" dataType="double" isCyclic="0" optype="continuous"
         displayName="Years at current residence"/>
    - <DataField name="Married" dataType="double" isCyclic="0" optype="categorical" displayName="">
         <Value property="valid" displayValue="No" value="0"/>
         <Value property="valid" displayValue="Yes" value="1"/>
      </DataField>
    - <DataField name="Income" dataType="double" isCyclic="0" optype="ordinal"
      displayName="Income category (thousands)">
         <Value property="valid" displayValue="<25" value="1"/>
         <Value property="valid" displayValue="25-49" value="2"/>
         <Value property="valid" displayValue="50-74" value="3"/>
         <Value property="valid" displayValue="75+" value="4"/>
      </DataField>
    - <DataField name="Responded" dataType="double" isCyclic="0" optype="categorical"
      displayName="Responded to test offer">
         <Value property="valid" displayValue="No" value="0"/>
         <Value property="valid" displayValue="Yes" value="1"/>
      </DataField>
   </DataDictionary>
 - <TreeModel functionName="classification" algorithmName="CRT">
    - <Extension extender="spss.com">
         <X-risk value="0.463917525773196"/>
         <X-seOfRisk value="0.0358043117150723"/>
      </Extension>
    - <MiningSchema>
         <MiningField name="Income" usageType="active"/>
         <MiningField name="Responded" usageType="predicted"/>
      </MiningSchema>
    - <ModelStats>
       - <UnivariateStats field="Income">
            <Counts invalidFreq="0" missingFreq="0" totalFreq="194"/>
         </UnivariateStats>
      </ModelStats>
    - <Node id="0" recordCount="194" score="0">
       - <Extension>
          - <X-Node>
               <X-NodeStats improvement="0.00543357728986443"/>
```

```
            </X-Node>
        </Extension>
        <True/>
    - <ScoreDistribution value="0" recordCount="97">
        <Extension name="probability" value="0.5" extender="spss.com"/>
      </ScoreDistribution>
    - <ScoreDistribution value="1" recordCount="97">
        <Extension name="probability" value="0.5" extender="spss.com"/>
      </ScoreDistribution>
    - <Node id="1" recordCount="27" score="0">
      - <CompoundPredicate booleanOperator="surrogate">
        - <CompoundPredicate booleanOperator="and">
            <True/>
            <SimplePredicate value="1" field="Income" operator="lessOrEqual"/>
          </CompoundPredicate>
          <False/>
        </CompoundPredicate>
      - <ScoreDistribution value="0" recordCount="17">
          <Extension name="probability" value="0.62962962962963" extender="spss.com"/>
        </ScoreDistribution>
      - <ScoreDistribution value="1" recordCount="10">
          <Extension name="probability" value="0.37037037037037" extender="spss.com"/>
        </ScoreDistribution>
      </Node>
    - <Node id="2" recordCount="167" score="1">
      - <CompoundPredicate booleanOperator="surrogate">
        - <CompoundPredicate booleanOperator="and">
            <SimplePredicate value="1" field="Income" operator="greaterThan"/>
            <True/>
          </CompoundPredicate>
          <True/>
        </CompoundPredicate>
      - <ScoreDistribution value="0" recordCount="80">
          <Extension name="probability" value="0.479041916167665" extender="spss.com"/>
        </ScoreDistribution>
      - <ScoreDistribution value="1" recordCount="87">
          <Extension name="probability" value="0.520958083832335" extender="spss.com"/>
        </ScoreDistribution>
      </Node>
  </Node>
- <Extension>
  - <X-TreeModel>
    - <X-PredictorImportanceList>
        <X-PredictorImportance importance="100" predictorName="Income"/>
      </X-PredictorImportanceList>
    - <X-Priors>
        <X-Prior-Value value="0.5" targetCategory="0"/>
        <X-Prior-Value value="0.5" targetCategory="1"/>
      </X-Priors>
    </X-TreeModel>
  </Extension>
</TreeModel>
</PMML>
```