مهندسی صنایع و مدیریت شریف
دوره‌ی ۱-۲۸، شماره‌ی ۲، ص. ۴۳-۵۴

Original Article

ارائه‌ی یک سامانه‌ی هوشمند با تلفیقی از درخت رگرسیونی و نقشه‌ی خودسازمانده بهینه‌شده برای تقسیم‌بندی بهینه‌ی مشتریان

علیرضا سروش * **(دانشجوی دکتری)**
اردشیر بحرینی‌نژاد (استادیار)
محمدرضا امین ناصری (دانشیار)
دانشکده‌ی فنی و مهندسی بخش مهندسی صنایع، دانشگاه تربیت مدرس

تقسیم‌بندی بهینه‌ی مشتریان برمبنای ویژگی‌های مرتبط می‌تواند به توسعه‌ی استراتژی‌های بازاریابی دقیق‌تر به‌منظور صرف کاراتر منابع کمک کند. اما ایجاد سامانه‌ی تقسیم‌بندی مشتریان که علاوه‌بر پیچیدگی کم از قابلیت تقسیم‌بندی بهینه‌یی برخوردار باشد، به‌دلیل حجم زیاد ویژگی‌ها کاری بسیار مشکل است. هدف این نوشتار، ارائه‌ی یک سامانه‌ی تلفیقی هوشمند مبتنی بر درخت رگرسیونی و نقشه‌ی خودسازمانده بهینه‌سازی شده است که از نظر محاسباتی کارا و دقیق باشد. نتایج نشان می‌دهد که درخت رگرسیونی ۹۳٪ از ویژگی‌ها را در حالت بهینه حذف می‌کند و لذا به کاهش قابل توجه هزینه‌ی محاسبات می‌انجامد. به‌علاوه، نتایج اعتبارسنجی نشان می‌دهد که این سامانه با دقت قابل توجهی خوشه‌ها را تفکیک کرده است و بدین طریق می‌توان منابع بازاریابی را برای جذب مشتریان مشابه با مشتریان بهترین خوشه‌ها صرف کرد.

واژگان کلیدی: مدیریت ارتباط با مشتری، انتخاب ویژگی، درخت رگرسیونی، تقسیم‌بندی مشتریان، نقشه‌ی خودسازمانده.

a.soroush@modares.ac.ir
bahreininejad@modares.ac.ir
amin_nas@modares.ac.ir

۱. مقدمه

با گسترش رقابت در عرصه‌های جهانی و محدودیت منابع، به‌کارگیری صحیح آن به‌عنوان یکی از چالش‌های عمده‌ی مدیریت مطرح شد. روابط بین کارکنان و مشتریان نیز به‌عنوان مهم‌ترین و کمیاب‌ترین منابع سازمان ــ که نقش قابل توجهی در سودآوری آن دارند ــ از چنان اهمیت و توجه روزافزونی برخوردار شد که می‌توان گفت مشتریان تنها مرکز سودده در شرکت‌ها هستند. بنابراین سازمان‌ها همواره سعی دارند براساس مفاهیم جدید بازاریابی نوین که به‌معنی دانش و هنر یافتن مشتری، و نگه‌داری و افزودن بر تعداد آن است، با خلق نیازها و خواسته‌های جدید و بدیع برای مشتریان و دوری جستن از اعمال قدرت و ضوابط خشک و به‌کارگیری مشارکت و تفاهم به مدیریت روابط با مشتریان پرداخته و آن‌ها را برای تضمین سودآوری و بقای خود در اختیار بگیرند.

در سال‌های اخیر با گسترش پایگاه‌های داده مشتریان و افزایش رقابت میان سازمان‌ها در جهان بیش از پیش به موضوع «مدیریت ارتباط با مشتری (CRM)[۱]» توجه کرده‌اند تا بتوانند ضمن شناسایی تقاضاهای مختلف مشتریان، به مزیتی رقابتی دست یابند.[۱] به‌طوری که سازمان‌ها به دنبال روش‌هایی برای انجام بازاریابی‌های دقیق‌تر و کسب هرچه بیشتر سهم بازار بوده‌اند تا بتوانند سودآوری خود را افزایش داده و هزینه‌ها را کمینه کنند. از آنجا که موفقیت سازمان منوط به قابلیت ایجاد و حفظ روابط ارزشمند با مشتری است، تعریف راهکارهایی پالوده برای مواجهه با مشتریان ضروری است. لازمه‌ی تعریف راهکاری‌های مناسب برای مشتریان مورد نظر سازمان، شناسایی ویژگی‌های آن‌هاست. با شناسایی این ویژگی‌ها، سازمان می‌تواند بر تعداد مشتریان کم‌تری متمرکز شود و بالتبع، هزینه‌ی کم‌تری به‌منظور انجام فعالیت‌های بازاریابی را متقبل شود. در این راستا، قانون پارتو درخصوص مشتریان چنین بیان می‌دارد که ۲۰٪ از مشتریان انجام‌دهنده‌ی ۸۰٪ معاملات شرکت هستند و ۸۰٪ سود شما را ایجاد می‌کنند.از این‌رو، مهم است که سازمان با شناسایی ویژگی‌های مشتریان مورد نظر و تقسیم‌بندی‌شان برحسب این ویژگی‌ها، محتمل‌ترین دسته تا کم‌احتمال‌ترین مشتریان را برای تأمین هدف سازمان اولویت‌بندی کند. بدین‌ترتیب، سازمان می‌تواند منابع بازاریابی خود را در راستای جذب مشتریان در هر دسته، با توجه به احتمال خرید آن‌ها به‌صورت سودآورتری هزینه کند. تقسیم‌بندی بهینه‌ی مشتریان یکی از مهم‌ترین عناصر برای ایجاد سامانه‌ی کامل CRM محسوب می‌شود. از این‌رو، مدل‌های بسیاری برای تقسیم‌بندی مشتریان سازمان‌های مختلف براساس ویژگی‌های متنوع ارائه شده است. برخی از تحقیقات انجام‌شده از سه متغیر تأخر، تکرر و کل مبلغ خرج‌شده،[۲-۴] برخی از ارزش طول زندگی مشتری،[۵و۶] و

---

* *نویسنده مسئول*

تاریخ: دریافت ۱۳۸۹/۶/۳۱، اصلاحیه ۱۳۸۹/۱۰/۲۵، پذیرش ۱۳۹۰/۳/۲۱.

برخی از ترکیبی از آن‌ها[۷] و یا پیمایش رضایت مشتری[۸] برای تقسیم‌بندی مشتریان استفاده کرده‌اند. هدف این محققین از تقسیم‌بندی مشتریان شناسایی و اولویت‌بندی مشتریان براساس سودآوری آن‌ها بوده است. برخی دیگر نیز، براساس احتیاجات مشتریان[۹و۱۰] با هدف شناسایی انواع سلیقه‌های آنان به تفکیک مشتریان پرداخته‌اند. همچنین به‌منظور بهبود دقت پیش‌بینی مصرف روزانه‌ی ایستگاه‌های برق، براساس پروفایل روزانه‌ی مشتریان ایستگاه‌های برق را تقسیم کرده‌اند.[۱۱] در مطالعه‌یی دیگر، تقسیم‌بندی به‌گونه‌یی است که برای کاهش هزینه‌ی عملیات لجستیک نسبت به خوشه‌بندی مشتریان براساس ویژگی‌های تقاضای آن‌ها قبل از اجرای مسیریابی ناوگان در عملیات لجستیک اقدام کرده‌اند.[۱۲] بررسی‌ها نشان می‌دهد که بیشتر محققین شیوه‌ی شبکه‌ی عصبی نقشه‌ی خودسازمانده (SOM)[۲] یا حالت ساده‌تر آن، تکنیک k-means را به کارگرفته‌اند. اما هیچ‌کدام از نویسندگان اقدام به محاسبه و تعیین ویژگی‌های مؤثرتر پیش از تقسیم‌بندی نکرده‌اند و ویژگی‌ها به‌صورت شهودی انتخاب شده‌اند؛ این موضوع می‌تواند به عدم کارایی و دقت پایین تقسیم‌بندی منجر شود.

مسئله‌ی انتخاب ویژگی‌ها، مسئله‌یی مستقل در نظریه‌ی تشخیص الگو بوده و تاکنون حل نشده است. فرایند انتخاب ویژگی‌ها به‌عنوان مسئله‌یی از بهینه‌سازی ترکیبی کلی در یادگیری ماشین شناخته می‌شود که تعداد ویژگی‌ها را کاهش داده و داده‌های غیرمرتبط و زائد را حذف می‌کند. هدف اصلی انتخاب ویژگی، شناسایی زیرمجموعه‌یی از ویژگی‌هاست که تأثیر بیشتری بر یک متغیر پاسخ معلوم دارند.[۱۳] کشف زیرمجموعه‌ی بهینه‌یی از ویژگی‌ها معمولاً مشکل بوده و نشان داده شده که بسیاری از مسائل مرتبط NP-hard شناخته می‌شوند.[۱۴] انتخاب ویژگی‌ها موفقیت‌های بسیاری را در کاربردهای دنیای واقعی داشته است، زیرا غالباً می‌تواند ابعاد داده‌ها را برای به‌کارگیری الگوریتم‌های داده‌کاوی به‌طور چشمگیری کاهش دهد. در سال‌های اخیر، CRM یکی از زمینه‌های تحقیقاتی بوده است که در روش‌های انتخاب ویژگی به کار گرفته شده‌اند.[۱۵و۱۶] پیاده‌سازی مناسب انتخاب ویژگی‌ها نه‌تنها اطلاعات مهم‌تر را برای تقسیم‌بندی بهینه‌ی مشتریان برمی‌گزیند، بلکه محاسبات مورد نیاز برای تحلیل داده‌های چندبعدی را کاهش داده و دقت آن را افزایش می‌دهد. به‌همین دلیل، پیاده‌سازی انتخاب ویژگی برای آماده‌سازی سامانه‌ی تقسیم‌بندی مشتریان بهینه دارای اهمیت است.

در سال‌های اخیر، چندین الگوریتم برای انتخاب مناسب ویژگی‌ها در حوزه‌ی CRM به کارگرفته شده و تعدادی مطالعه‌ی مقایسه‌یی نیز انجام شده است. به‌علاوه، از آنجا که هر ویژگی مورد استفاده به‌عنوان بخشی از یک رویه‌ی تقسیم‌بندی می‌تواند هزینه و زمان اجرای سامانه‌ی تقسیم‌بندی مشتریان را افزایش دهد، انگیزه‌ی قوی برای طراحی و پیاده‌سازی سامانه‌یی با مجموعه‌ی ویژگی‌های کم وجود دارد. هم‌زمان یک نیاز متضاد برای لحاظ مجموعه‌ی کافی از ویژگی‌ها به‌منظور دست‌یابی به نرخ‌های تشخیص بالا تحت شرایط مشکل وجود دارد. این موضوع منجر به توسعه‌ی تکنیک‌های متنوعی برای کشف یک زیرمجموعه‌ی بهینه از ویژگی‌ها شده است. ساده‌ترین روش جست‌وجوی بهینه یک روش جامع است، هرچند که با این روش تعداد زیرمجموعه‌های ممکن به سرعت رشد کرده و برای مجموعه ویژگی‌های با اندازه‌ی متوسط نیز غیرقابل استفاده است. روش‌های جست‌وجوی بهینه‌یی همچون الگوریتم شاخه و تحدید وجود دارند که مانع از رویکردی جامع می‌شود. دو روش ابتکاری مشهور نیز انتخاب پیشرو متوالی و انتخاب پسرو متوالی هستند. پویا شده‌ی این دو روش به‌نام‌های جست‌وجوی شناور پیشرو متوالی و جست‌وجوی شناور پسرو متوالی معرفی می‌شوند.

یک رویه‌ی آماری ساده، انتخاب ویژگی رو به جلو مبتنی بر امتیاز مجذور خی‌دو است، اگرچه برای دست‌یابی به حل بهینه با محدودیت‌هایی مواجه‌اند. روش دیگر، رویکرد مبتنی بر جست‌وجوی تصادفی الگوریتم ژنتیک (GA) است که از گام‌های احتمالی یا تکنیک‌های نمونه‌گیری استفاده می‌کند. این روش وزن‌ها را به ویژگی‌ها تخصیص می‌دهد. ویژگی‌های وزن داده‌شده با پیشروی تا حد آستانه‌یی که توسط کاربر تعریف شده انتخاب می‌شوند.[۱۷] به‌منظور هدف‌گذاری مشتریان[۱۸] و نیز به منظور بهبود پیش‌بینی میزان خرید مشتری[۱۹] از یک الگوریتم ژنتیک (GA) استفاده می‌شود. به‌کارگیری GA برای تعیین ویژگی‌های مؤثر به زمان بسیار زیادی برای محاسبات نیاز خواهد داشت. محققین یک سامانه‌ی استدلال مبتنی بر مورد همراه با تکنیک کاهش دوبعدی برای رضایت مشتریان پیشنهاد کرده‌اند[۲۰] که رفتار خرید مشتریان را برای یک محصول خاص با استفاده از مشخصه‌های آماری آن‌ها پیش‌گویی می‌کند. برای انتخاب ویژگی، ترکیبی از الگوریتم افراز تودرتو[۳] و روش ذوب شبیه‌سازی‌شده[۴] به‌منظور تشخیص مشتری[۲۱] و نظریه‌ی مجموعه‌ی ناهنجار[۵] به‌منظور پیشگویی رفتار خرید مشتری[۲۲] به کار برده شده‌اند. همچنین، ماشین برداری پشتیبان[۶] برای رتبه‌بندی ویژگی‌ها درکاربردهای CRM دنیای واقعی پیشنهاد شده است.[۲۳] مهم‌ترین نقطه ضعف این روش‌ها این است که خطی بوده و روابط غیرخطی بین هر ویژگی و متغیر هدف را لحاظ نمی‌کنند. برخی از محققین رویکرد درخت تصمیم را برای انتخاب ویژگی پیشنهاد، و بیان کرده‌اند که این روش بسیار دقیق بوده و سابقه‌ی عملکرد خوبی دارد.[۱۵]

مقصود این نوشتار توسعه‌ی یک سامانه‌ی هوشمند تلفیقی است که به‌لحاظ محاسباتی کارا و دقیق است. رویکرد انتخابی بدین صورت است: ابتدا یک درخت رگرسیونی هرس‌شده با سرعت محاسباتی بالا و دقتی قابل توجه برای انتخاب ویژگی‌های بهینه طراحی می‌شود. سپس یک شبکه‌ی عصبی SOM برای تقسیم‌بندی بهینه‌ی مشتریان مبتنی بر شاخص دیویس ـ بولدین[۷] ایجاد شده و بهینه‌ترین خوشه‌ها به‌ترتیب اولویت تعیین‌شده برای ویژگی‌های مشتریان در هر خوشه، به‌منظور تدوین راهکارهای بازاریابی توصیف می‌شوند. بدین‌ترتیب، سازمان می‌تواند بر خوشه‌هایی که احتمال تأمین هدف مورد نظر سازمان در آن‌ها بیشتر است متمرکز شود تا هزینه‌ی بازاریابی کم‌تری متوجه آن شود. از طریق به‌کارگیری چنین رویکردی، دست‌یابی به هر دو نیاز ــ پیچیدگی کم و عملکرد تقسیم‌بندی بهینه ــ ممکن می‌شود. پس از تعیین تعداد خوشه‌های بهینه، نتایج آن با داده‌هایی جدا از داده‌های آموزشی برای طراحی سامانه، اعتبارسنجی می‌شود. در ادامه سامانه‌ی هوشمند تلفیقی طراحی‌شده روی داده‌های یک شرکت بیمه پیاده‌سازی شده و قابلیت آن بررسی می‌شود.

در ادامه‌ی این تحقیق به توصیف موردکاوی تحقیق می‌پردازیم و سپس از طریق درخت رگرسیونی بهینه‌شده به انتخاب ویژگی خواهیم پرداخت. پس از آن، در بخش چهارم، تقسیم‌بندی بهینه‌ی مشتریان با توسعه‌ی SOM صورت گرفته و نتایج مورد ارزیابی قرار می‌گیرند.

۲. توصیف داده‌ها: موردکاوی یک شرکت بیمه

یکی از پرکاربردترین زمینه‌هایی که می‌توان برای شناسایی مشتریان از آن استفاده کرد، صنعت بیمه است. به‌ویژه آن که، شناسایی ویژگی‌های مشتریانِ یکی از محصولات شرکت بیمه از طریق رفتاری که در قبال سایر محصولات بیمه‌یی شرکت از خود نشان داده‌اند، می‌تواند جالب توجه باشد. از این‌رو، در این تحقیق مجموعه‌ی داده‌ها که مربوط به کسب‌وکاری در دنیای واقعی است، از طریق یک شرکت بیمه تهیه شده است. این شرکت بیمه قصد دارد مشتریان بالقوه برای یک محصول معین را

شناسایی کند. داده‌ها مربوط به ۹۸۲۲ مشتری این شرکت بیمه است که تعدادی از آن‌ها اقدام به خرید بیمه‌نامه‌ی خودرویی به‌نام کاروان کرده‌اند. این داده‌ها امکان ارزیابی قابلیت سامانه‌ی هوشمند در تقسیم‌بندی بهینه‌ی مشتریان را فراهم می‌سازد.

در این مقاله، دو مجموعه داده‌ی جداگانه به کار برده می‌شود: یک مجموعه‌ی آموزشی با ۵۸۲۲ مشتری و یک مجموعه‌ی ارزیابی با ۴۰۰۰ مشتری. هر رکورد متشکل از ۸۶ ویژگی شامل داده‌های آمارگیری اجتماعی (۴۳ ویژگی) و مالکیت محصول (۴۲ ویژگی) می‌شود. ویژگی آخر یعنی «تعداد بیمه‌نامه‌ی کاروان»، متغیر هدف است. از داده‌های آموزشی برای آماده‌سازی سامانه‌ی هوشمند استفاده می‌شود و نتیجه براساس مجموعه‌ی ارزیابی اعتبارسنجی می‌شود. از ۵۸۲۲ مشتری احتمالی در مجموعه داده‌ی آموزشی، ۳۴۸ نفر بیمه‌نامه‌ی کاروان را خریده‌اند که از آن نرخ هدف ۵٫۹۷٪=۳۴۸/۵۸۲۲ برای مشتریان مورد نظر حاصل می‌شود. همچنین، از ۴۰۰۰ مشتری احتمالی در مجموعه داده‌ی ارزیابی، ۲۳۸ نفر بیمه‌نامه‌ی کاروان را خریده‌اند که نرخ هدف ۵٫۹۵٪=۲۳۸/۴۰۰۰ حاصل می‌شود. چنان که مشاهده می‌شود نرخ هدف تقریباً در هر دو مجموعه برابر است. سامانه‌ی هوشمند برای تقسیم‌بندی مشتریان و شناسایی ویژگی‌های ۲۰ درصد اولِ مشتریانی طراحی می‌شود که انتظار می‌رود در مجموعه داده ارزیابی محتمل‌ترین افراد برای خرید بیمه‌نامه‌ی کاروان باشند. ذکر این نکته ضروری است که تنها اطلاعات در مجموعه داده‌ی آموزشی در توسعه‌ی سامانه‌ی هوشمند استفاده می‌شود و مجموعه داده‌ی ارزیابی منحصراً برای اعتبارسنجی استفاده می‌شود.

افزون براین، در طراحی چنین سامانه‌یی سه موضوع ضرورت می‌یابد: ویژگی‌های منتخب، الگوریتم خوشه‌بندی، ارزیابی خروجی خوشه‌بندی. این سه‌موضوع کارایی، اثربخشی و کیفیت تقسیم‌بندی را تعیین می‌کنند.

۳. انتخاب ویژگی مبتنی بر یک درخت رگرسیونی

چنان که اشاره شد، یکی از مشکلات عمده درخصوص مسائل تشخیص الگو، ابعاد زیاد و انتخاب متغیرهای غیر سودآور است. بدین‌معنا که معمولاً تعداد ویژگی‌های در اختیار طراح سامانه‌ی طبقه‌بندی یا خوشه‌بندی بسیار زیاد است و احتمال انتخاب ویژگی‌های غیرمرتبط بسیار زیاد است. انتخاب ویژگی به فرایند انتخاب مرتبط‌ترین ویژگی‌ها گفته می‌شود که بیشترین ارتباط را با متغیر(های) خروجی داشته باشند.[۲۴] این فرایند باید بسیار دقیق انجام شود؛ اگر ویژگی‌های مهم به‌درستی انتخاب نشوند، ویژگی‌های غیرمرتبط می‌تواند بر پیچیدگی مسئله بیفزاید و دقت مدل را کم کند. هدف شناسایی کم‌ترین زیرمجموعه‌از متغیرهاست که بالاترین دقت را ارائه می‌کند. این موضوع ممکن است ساده به نظر آید، که البته چنین نیست. زیرا اولاً، برای یک پایگاه داده تنها با ۲۰ ویژگی، بیش از ۱ میلیون زیرمجموعه‌ی ممکن وجود دارد. ثانیاً برای ارزیابی هر زیرمجموعه، دسترسی به یک مدل و ارزیابی آن از طریق اندازه‌گیری خطا ضروری خواهد بود.

از آنجا که اندازه‌ی فایل داده‌های مشتری در حال افزایش است، گرایش به مسئله‌ی انتخاب متغیر به‌شدت رشد داشته است. تنها حفظ مرتبط‌ترین ویژگی‌ها، اجرای کاهش ابعاد با کم‌ترین هزینه‌ی اطلاعات داده‌یی برای کاهش پیچیدگی محاسبات و افزایش دقت تقسیم‌بندی را ممکن می‌سازد. امتیازات استفاده از درخت رگرسیونی به‌عنوان یک پیش‌پردازشگر برای خوشه‌بندی مشتریان با استفاده از SOM عبارت است از: غیرخطی بودن (با توجه به غیرخطی بودن رفتار مشتری)، زمان آموزش بسیار سریع، عدم نیاز به تبدیل یا آماده‌سازی داده‌ها، به‌کارگیری خودکار پیشگویی‌کننده‌های قیاسی (اسمی)، قابلیت پشتیبانی پیشگویی‌کننده‌های بسیار زیاد (حداکثر ۸۰۰۰ عدد) و انتخاب یک به یک متغیرهای مرتبط در طول فرایند. از این‌رو، درخت رگرسیونی می‌تواند رویکردی مناسب برای شناسایی ویژگی‌های بهینه برای مسئله‌ی تقسیم‌بندی مشتریان باشد.[۲۵و۲۶]

۱.۳. درخت رگرسیونی هرس‌شده

درخت‌های تصمیم یکی از ساده‌ترین و موفق‌ترین الگوریتم‌های یادگیری در داده‌کاوی و یادگیری ماشین هستند. شهرت این درخت‌های تصمیم از آن روست که به‌سادگی قابل تفسیر و از نظر محاسباتی کم‌هزینه‌اند. در صنعت و محیط کسب‌وکار، از شیوه‌هایی برای ارزیابی اعتبار، کشف کلاه‌برداری و مدیریت ارتباط با مشتری استفاده شده‌اند.[۲۵] معمولاً هدف پیداکردن درخت تصمیم بهینه با کمینه‌سازی خطای تعمیم است. ویژگی‌های ورودی و مقادیر خروجی می‌توانند گسسته یا پیوسته باشند. درخت تصمیم دو نام دیگر نیز دارد: ۱. درخت تصمیم با محدوده‌یی از برچسب‌های دسته‌یی گسسته (قیاسی یا طبقه‌یی)، ۲. درخت طبقه‌بندی. این در حالی است که درخت تصمیم با محدوده‌یی از مقادیر خروجی پیوسته (عددی) را «درخت رگرسیونی» می‌نامند. به‌طور کلی، این درخت‌های تصمیم را درخت طبقه‌بندی و رگرسیونی (CART)[۸] می‌نامند.[۲۷]

میزان اهمیت یک ویژگی براساس مجموع بهبودها درکلیه‌ی گره‌ها تعیین می‌شود که ویژگی نقش دو نیم‌کننده (وزن‌دهی شده توسط بخشی از داده‌های آموزشی در هر شکاف گره) را دارد. جانشین‌ها نیز در محاسبات اهمیت لحاظ می‌شوند، بدین‌معنا که حتی به متغیری که هرگز یک گره را دو نیم نمی‌کند، ممکن است امتیاز اهمیت بزرگی اختصاص داده شود. این موضوع، رتبه‌بندی اهمیت متغیر برای مشخص‌سازی پوشانه‌های متغیر و همبستگی غیرخطی میان ویژگی‌ها را ممکن می‌سازد. امتیازهای اهمیت را می‌توان به‌صورت ارادی به دونیم‌کننده‌ها محدود کرد؛ مقایسه‌ی دونیم‌کننده‌ها و رتبه‌بندی‌های اهمیت کل تشخیصی مفید است.[۲۵]

از آنجا که درخت‌های تصمیم با پیچیدگیِ کم‌تر جامع‌ترند، معمولاً تصمیم‌گیرندگان آن‌ها را ترجیح می‌دهند. به‌علاوه، پیچیدگی درخت به‌دلیل یادگیری جزئیات خاص تأثیر شدیدی بر عملکرد دقت آن دارد و منجر به عملکرد تعمیمی ضعیفی خواهد شد. تکنیک درخت رگرسیونی دونیم‌کننده‌هایی را جست‌وجو می‌کند که مربعات خطای (انحراف حداقل مربعات) پیشگویی را کمینه می‌کند. پیشگویی در هر گره پایانی برمبنای میانگین وزنی برای گره تعیین می‌شود. متداول‌ترین رویکرد مورد استفاده، ابتدا رشد یک درخت تا یک اندازه‌ی بزرگ و سپس هرس‌کردن گره‌ها براساس معیار هرس است. معیار توقف مورد استفاده و روش هرس به کار برده شده پیچیدگی را کنترل می‌کند.[۲۸] استفاده از معیار ارزیابی عملکرد به‌طوری که درخور مسئله‌ی تحت بررسی باشد، بسیار مهم است. هزینه‌ی میانگین مربعات خطا باید مورد محاسبه قرار گیرد، زیرا آن‌ها در انتخاب روش تأثیرگذارند. فرض برابری هزینه‌های میانگین مربعات خطا در موارد بسیار کمی مناسب است. معمولاً ــ حتی اگر این هزینه‌ها دقیقاً معلوم نباشند ــ تا اندازه‌یی می‌توان درخصوص هزینه‌ها صحبت کرد.

۲.۳. الگوریتم هرس[۹]

کارهای اولیه در حوزه‌ی درخت‌های تصمیم امکان هرس را نمی‌دادند. در آن شرایط، درخت‌ها آن‌قدر رشد می‌کردند تا با شرط توقف برخورد کنند، و نهایتاً درخت منتج به‌عنوان درخت نهایی در نظر گرفته می‌شد. «هرس» فرایند کاهش یک درخت از طریق تبدیل برخی گره‌های شاخه‌یی به گره‌های پایانی و حذف گره‌های پایانی

زیرشاخه‌ی اصلی است. سازوکار هرس اکیداً بر داده‌های آموزشی مبتنی می‌شود و با یک سنجه هزینه‌ی پیچیدگی آغاز می‌شود.

الگوریتم هرس به‌طور کلی برای درخت‌هایی به کار گرفته می‌شود که لزوماً طبقه‌بندی نشده‌اند بلکه درخت‌های رگرسیونی‌اند. فرض کنید که $R(t)$ اعداد حقیقی همبسته با هر گره t از یک درخت معلوم T باشد. مقدار $R(t)$ مطابق رابطه‌ی ۱ تعیین شود:

$$R(t) = e(t)p(t) \quad (۱)$$

که در آن $e(t)$ میانگین مربعات خطا معلوم می‌کند که یک مورد در گره می‌افتد. اگر ما فرض کنیم که $N(t)$ تعداد نمونه‌های $L = \{(x_i, y_i), i = ۱, ..., n\}$ را نشان دهد، آنگاه $p(t)$ را می‌توان چنین تعریف کرد:

$$p(t) = \frac{N(t)}{n} \quad (۲)$$

از این‌رو، اگر t به‌عنوان یک گره پایانی درنظرگرفته شود، $R(t)$ سهم آن گره به خطای کل است. فرض کنید که T_t زیردرختی با ریشه‌ی t باشد. اگر $R_\alpha(T_t) < R_\alpha(t)$، آنگاه سهم به هزینه‌ی پیچیدگی زیردرخت کم‌تر از سهم گره t است. این موضوع برای α کوچک اتفاق می‌افتد. همچنان که α افزایش می‌یابد، تساوی حاصل می‌شود، زمانی که:

$$\alpha = \frac{R(t) - R(T_t)}{N_d(t) - ۱} \quad (۳)$$

که در آن $N_d(t)$ تعداد گره‌های پایانی در T_t است، یعنی $N_d(t) = |\widetilde{T}_t|$ و حذف درخت در t برتری می‌یابد. بنابراین:

$$g(t) = \frac{R(t) - R(T_t)}{N_d(t) - ۱} \quad (۴)$$

به‌عنوان سنجه‌یی از شدت پیوند گره t تعریف می‌شود. اولین مرحله‌ی الگوریتم، گره با کوچک‌ترین مقدار $g(t)$ را جست‌وجو می‌کند. آن گره تبدیل به یک گره پایانی می‌شود و مقدار $g(t)$ برای کلیه‌ی اجدادش محاسبه می‌شود. این فرایند تکرار می‌شود و ادامه می‌یابد تا به گره ریشه برسیم. بدین‌ترتیب، الگوریتم هرس، یک رشته از درخت‌ها را تولید می‌کند.[۲۷]

۳.۳. بررسی اعتبار[۱۰]

بررسی اعتبار (CV)، روشی برای برآورد میزان خطاست که دارای ایده‌یی ساده است. مجموعه داده‌ها به‌اندازه‌ی نمونه به دو قسمت تفکیک می‌شوند. پارامترهای مدل با استفاده از یک مجموعه (با کمینه‌سازی چند معیار بهینه‌سازی) برآورد شده و معیار خوبی برازش[۱۱] بر روی مجموعه‌ی دوم ارزیابی می‌شود. نسخه‌ی معمول بررسی اعتبار یک روش صرف‌نظر از یکی[۱۲] است که در آن مجموعه‌ی دوم متشکل از تنها یک نمونه است. آنگاه برآورد بررسی اعتبار معیار خوبی برازش (CV) متوسط کلیه‌ی مجموعه‌های آموزشی ممکن به‌اندازه‌ی $n - ۱$ است. خطای بررسی اعتبار (CV) به‌عنوان وسیله‌یی برای تعیین یک مدل مناسب، برای هر عضو از خانواده مدل‌های کاندید $\{M_k, k = ۱, ..., K\}$ و مدل $M_{\hat{k}}$ انتخاب شده محاسبه می‌شود، که:

$$\hat{k} = \arg\min CV(k) \quad (۵)$$

بررسی اعتبار هنگام انتخاب یک مدل صحیح گرایش به برازش بیش از حد دارد، به‌طوری که، برای مجموعه داده یک مدل بسیار پیچیده انتخاب می‌کند. شواهدی وجود دارد که بررسی اعتبار چندتایی، زمانی که $d > ۱$ نمونه از مجموعه آموزشی حذف می‌شود، انتخاب مدل بهتر از بررسی اعتبار صرف‌نظر از یکی انجام می‌شود. برای n بزرگ، میزان محاسبات زیاد بوده و به طراحی n طبقه‌بندی‌کننده نیاز دارد. با این که با وجود هزینه‌ی یک افزایش در واریانس برآوردکننده، به‌صورت تخمینی نااریب است.[۲۹]

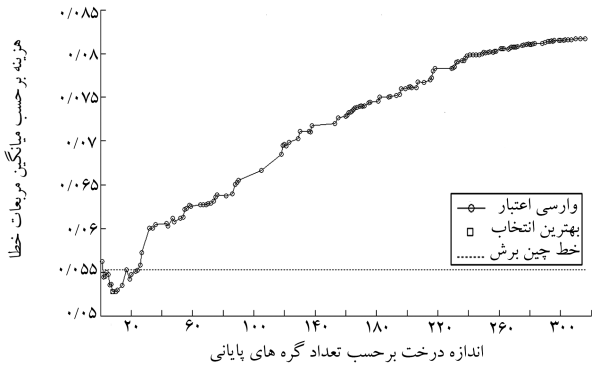
۴.۳. مشخصه‌های درخت رگرسیونی هرس‌شده

در قسمت‌های قبل نحوه‌ی عملکرد درخت رگرسیونی هرس‌شده به‌عنوان ابزاری برای انتخاب ویژگی‌ها توصیف شد. در این بخش نیز به بیان مشخصه‌های آن برای انتخاب ویژگی‌ها در موردکاوی می‌پردازیم. مشخصات درخت رگرسیونی هرس‌شده‌ی مورد استفاده عبارت است از:

- نوع درخت رگرسیونی است، زیرا نوع متغیر هدف عددی است (۰ یا ۱).
- از بررسی اعتبار ۵۰ تکه‌یی به‌منظور برآورد خطای واقعی برای درخت‌ها در اندازه‌های مختلف استفاده می‌شود؛ بدین‌معنا که تابع نمونه به ۵۰ زیرنمونه ــ که به‌صورت تصادفی انتخاب شده و تقریباً دارای اندازه‌ی برابرند ــ تفکیک می‌شود. برای هریک از زیرنمونه‌ها، یک درخت به داده‌های باقی‌مانده برازش می‌یابد و از آن برای پیش‌بینی زیرنمونه استفاده می‌شود. سپس اطلاعات تمامی زیرنمونه‌ها برای محاسبه‌ی هزینه‌ی کل نمونه با یکدیگر ترکیب می‌شود. همچنین برداری شامل خطای معیارِ هر مقدار هزینه، برداری شامل تعداد گره‌های پایانی برای هر زیردرخت، و اسکالری شامل بهترین سطح برآوردشده‌ی هرس محاسبه می‌شود. بهترین سطح، کوچک‌ترین درختی را که در محدوده‌ی خطای معیار از زیردرختِ کم‌ترین هزینه است، تولید می‌کند. متغیرهای منتخب آن‌هایی خواهند بود که بهترین عملکرد را دارند؛ یعنی درخت دارای کم‌ترین هزینه انتخاب می‌شود.
- برای هرس‌کردن درخت، معیار میانگین مربعات خطا به کار برده می‌شود.
- هزینه‌ی درخت، مجموع کلیه‌ی گره‌های پایانی شامل احتمال برآوردشده‌ی هرگره در هزینه‌ی گره است. از آنجا که درخت یک درخت رگرسیونی است، هزینه‌ی یک گره میانگین مربعات خطا در کلیه‌ی مشاهداتِ آن گره است.
- خطا برای هر گره، واریانس مشاهدات تخصیص داده شده به آن گره است.
- احتمال یک گره از طریق نسبتی از مشاهداتِ داده‌های اصلی که شرایط گره را برآورده می‌کنند، محاسبه می‌شود.
- اندازه یک گره به‌صورت تعدادی از مشاهداتِ داده‌های مورد استفاده به‌منظور ایجاد درختی که شرایط را برای گره برآورده کند، تعریف می‌شود.

۵.۳. نتایج انتخاب ویژگی‌ها

براساس مشخصه‌های یادشده، درخت رگرسیونی برای ۵۸۲۲ مشتری ترسیم می‌شود که شامل ۸۵ ویژگی است. با توجه به تعداد ویژگی‌ها و مشتریان می‌توان دریافت که اندازه و پیچیدگی درخت بسیار زیاد است؛ بنابراین مجموعه‌قواعد تولیدشده نمی‌توانند خیلی شهودی باشند. برای دست‌یابی به توصیفِ فشرده‌تر داده‌ها، برخی از شاخه‌های درخت را هرس می‌کنیم به‌طوری که درختی کوچک‌تر را خواهیم دید. بنابراین، درختی انتخاب می‌شود که کم‌ترین هزینه با تعداد ویژگی کم‌تر و تعداد گره‌های پایانی کم‌تر را دارد. در شکل ۱ نقطه‌ی بهینه برای هرس درخت با کم‌ترین هزینه نمایش داده شده است.



شکل ۱. نقطه‌ی بهینه برای هرس درخت رگرسیونی با کم‌ترین هزینه.

چنان که در تصویر مشاهده می‌شود، خط ترسیم‌شده نشان‌گر هزینه‌ی برآوردی برای هر اندازه درخت است؛ خط چین نشان‌گر یک خطای استاندارد بالاتر از کم‌ترین خطاست؛ و مربع کوچک بر روی خط نشان‌گر کم‌ترین هزینه‌ی درخت زیر خط‌چین است. درخت کامل شامل ۳۱۶ گره پایانی است که هزینه برمبنای معیار میانگین مربعات خطا برای زیردرخت‌های مختلف محاسبه شده است. به‌طورکلی، درخت ۶۳۱ گره دارد که هزینه‌ی درخت به‌تدریج و همزمان با رشد اندازه درخت و رسیدن به مقدار ۰/۰۵۲۸ با ۸ گره پایانی، کاهش می‌یابد. هزینه‌ی درخت در ادامه، در روندی صعودی شروع به نوسان می‌کند به‌طوری که، با افزایش تعداد گره‌های پایانی پس از رسیدن به نقطه‌ی بهینه، تعداد ویژگی‌های بیشتری برگزیده می‌شوند اما منجر به هزینه‌ی‌بالاتری می‌شود. بنابراین، بهترین انتخاب، بهترین سطح برآوردی هرس یا نقطه‌ی بهینه را نشان می‌دهد که شامل هشت گره پایانی می‌شود. در این حالت، می‌توان متوجه شد که پیچیدگی‌های درخت از طریق کاستن تعداد گره‌های پایانی کاهش می‌یابد. زمانی که درخت رگرسیونی هرس‌شده در مجموعه‌ی آموزشی ترسیم می‌شود، ویژگی‌هایی که در ترسیم آن به نظر می‌رسند انتخاب می‌شوند. به‌طوری که، ویژگی‌هایی که در مسیرهای منتهی به هر گره پایانی در درخت هرس شده به‌نظر می‌رسند، زیرمجموعه‌ی منتخب بهینه می‌شوند. جدول ۱، قواعد درخت، اندازه درخت و مقادیر $p(t)$، $e(t)$، $R(t)$در هر گره را بعد از مرحله‌ی هرس بهینه نشان می‌دهد.

چنان که در جدول ۱ مشاهده می‌شود، پس از مرحله‌ی هرس بهینه ۱۵ گره باقی می‌ماند. شماره‌ی گره‌های پایانی ۲، ۶، ۱۰، ۱۱، ۱۲، ۱۳، ۱۴ و ۱۵ هستند و هریک نشان می‌دهند که چه تعداد مشتری در ستون اندازه براساس چه قاعده‌یی و چه احتمالی در نظر گرفته می‌شوند. خطا برای هر گره در ستون پنجم نشان‌گر واریانس مشاهدات تخصیص داده شده به آن گره است، و ریسک برای هر گره در ستون ششم نشان‌گر خطای گره وزن‌دهی شده توسط احتمال گره است.

نهایتاً، شش ویژگی به‌نام‌های بیمه‌نامه‌ی خودرو، نوع مشتری، بیمه‌نامه‌ی آتش‌سوزی، بیمه‌نامه‌ی قایق، کارگر بی‌تجربه و بیمه‌ی شخص ثالث به‌عنوان تأثیرگذارترین ویژگی‌ها بر هدف، یعنی خرید بیمه‌نامه‌ی کاروان، انتخاب می‌شوند. این رویه‌ی انتخابِ ویژگی منجر به کاهشی قابل توجه، معادلِ ۹۳٪ در داده‌های ورودی می‌شود. در مرحله‌ی بعد پیاده‌سازی سامانه‌ی هوشمند تقسیم‌بندی مشتریان طراحی‌شده، خریداران بیمه‌نامه‌ی کاروان براساس ویژگی‌های منتخب بهینه خوشه‌بندی‌می‌شوند.

۴. تقسیم‌بندی مشتریان

در این مرحله پس از تعیین ویژگی‌های مؤثر بر رفتار خرید مشتری به تقسیم‌بندی مشتریان جهت شناسایی ویژگی‌های مشترک مشتریان هر خوشه‌برای صرف بهینه هزینه‌های بازاریابی خواهیم پرداخت. بنابراین، نیاز به ابزاری است که یاد بگیرد چگونه به ویژگی‌های مختلف وزن‌دهی کرده و همزمان یک نقشه تعمیم یافته که بیش از حد منطبق نباشد، تولید کند. بدین‌منظور از آنجا که دسته‌بندی اولیه‌ی مشخصی وجود ندارد (مسئله‌ی تقسیم‌بندی از نوع خوشه‌بندی است نه طبقه‌بندی) و با توجه به روابط غیرخطی که در میان ویژگی‌ها دیده می‌شود، از شبکه‌ی عصبی کوهونن

جدول ۱. قواعد درخت بعد از مرحله‌ی هرس‌بهینه.

گره (t)	قاعده	اندازه	احتمال $p(t)$	خطا $e(t)$	ریسک $R(t)$
۱	اگر بیمه‌نامه‌ی خودرو کوچک‌تراز ۵/۵ است برو به گره ۲ در غیر این‌صورت گره ۳.	۵۸۲۲	۱	۰/۰۵۶۲	۰/۰۵۶۲
۲	تطابق = ۰/۰۲۴۹	۳۴۵۹	۰/۵۹۴۱	۰/۰۲۴۲	۰/۰۱۴۴
۳	اگر نوع مشتری کوچک‌تراز ۲/۵ است، برو به گره ۴ در غیر این‌صورت گره ۵.	۲۳۶۳	۰/۴۰۵۸	۰/۰۹۸۶	۰/۰۴۰۰
۴	اگر بیمه‌نامه‌ی آتش‌سوزی کوچک‌تراز ۳/۵ است، برو به گره ۶ درغیر این‌صورت گره ۷.	۴۶۴	۰/۰۷۹۶	۰/۱۶۱۵	۰/۰۱۲۹
۵	اگر بیمه‌نامه‌ی قایق کوچک‌تراز ۰/۵ است، برو به گره ۸ درغیر این‌صورت گره ۹.	۱۸۹۹	۰/۳۲۶۱	۰/۰۸۰۶	۰/۰۲۶۳
۶	تطابق = ۰/۱۲۸۶	۲۴۱	۰/۰۴۱۳	۰/۱۱۲۱	۰/۰۰۴۶
۷	اگر کارگر بی‌تجربه کوچک‌تراز ۳/۵ است، برو به گره ۱۰ در غیر این‌صورت گره ۱۱.	۲۲۳	۰/۰۳۸۳	۰/۲۰۲۷	۰/۰۰۷۸
۸	اگر بیمه‌نامه‌ی آتش‌سوزی کوچک‌تراز ۲/۵ است، برو به گره ۱۲ درغیر این‌صورت گره ۱۳.	۱۸۸۲	۰/۳۲۳۲	۰/۰۷۷۳	۰/۰۲۵۰
۹	اگر بیمه‌ی شخص ثالث شخصی کوچک‌تراز ۱ است، برو به گره ۱۴ درغیر این‌صورت گره ۱۵.	۱۷	۰/۰۰۲۹	۰/۲۴۹۱	۰/۰۰۰۷
۱۰	تطابق = ۰/۲۶۱۷	۲۱۴	۰/۰۳۶۷	۰/۱۹۳۲	۰/۰۰۷۱
۱۱	تطابق = ۰/۷۷۷۸	۹	۰/۰۰۱۵	۰/۱۷۲۸	۰/۰۰۰۳
۱۲	تطابق = ۰/۰۵۱۰	۹۶۱	۰/۱۶۵۰	۰/۰۴۸۴	۰/۰۰۸۰
۱۳	تطابق = ۰/۱۱۹۴	۹۲۱	۰/۱۵۸۱	۰/۱۰۵۲	۰/۰۱۶۶
۱۴	تطابق = ۰/۸۱۸۲	۱۱	۰/۰۰۱۸	۰/۱۴۸۸	۰/۰۰۰۳
۱۵	تطابق = ۰	۶	۰/۰۰۱۰	۰	۰

یا نقشه‌ی خودسازمانده (SOM) به‌عنوان یکی از بهترین روش‌های خوشه‌بندی یا دسته‌بندی هدایت‌نشده استفاده می‌شود. محققین بسیاری SOM را الگوریتمی کارا برای خوشه‌بندی مشتریان ارزیابی کرده‌اند.

به‌علاوه، باید توجه داشت که یک مرحله‌ی مهم در طراحیِ هر سامانه، مرحله‌ی ارزیابی عملکرد است که در آن احتمال خطای تقسیم‌بندی سامانه‌ی طراحی‌شده برآورد می‌شود. اگر ویژگی‌هایی با قدرت تمایزکم انتخاب شوند، طراحی سامانه منجر به عملکردی ضعیف خواهد شد. از این‌رو، باید ویژگی‌هایی انتخاب شوند که در فضای بردار ویژگی منجر به فاصله بین دسته‌یی بزرگ و واریانس درون دسته‌یی کوچک می‌شوند.[۲۸] به‌علاوه، توجه به این نکته ضروری است که داده‌های مورد استفاده برای استخراج ویژگی‌ها باید کاملاً از داده‌های اعتبارسنجی مستقل باشند، در غیر این‌صورت خطر تطابق بیش از حد وجود خواهد داشت.

۱.۴. نقشه‌ی خودسازمانده

گاهی ممکن است متغیرها به‌روشی کاملاً غیرخطی به‌هم مرتبط باشند که در این صورت، مدل‌های تحلیلی خطی قادر به شناسایی این روابط نخواهند بود. از آنجا که SOM یک روش نمایش غیرخطی است، غالباً می‌توان چنین حالت‌ها یا خوشه‌های مشخصه‌یی را بدون تشریح مدل‌سازی سامانه در نقشه‌ی خودسازمانده مشاهده کرد. این نوع شبکه‌ی عصبی بیشترین ساختارهای مشخصه‌یی تابع چگالی ورودی را در نمایش‌گری با ابعاد کم ارائه می‌کند. از این‌رو SOM ابزاری قوی برای کشف و تصویرسازی ساختارهای کلی فضای حالت و رفتار سامانه است که عملیات تصویرسازی و خوشه‌بندی را با هم ترکیب می‌کند. عمده‌ترین مزایای این روش، نسبت به سایر روش‌ها، عبارت‌اند از:[۳۰]

- مجموعه‌یی بازاز مشاهدات چندمتغیره را توسط مجموعه‌یی نامحدود از مشاهدات مدل مانند خوشه‌بندی K-means نمایش می‌دهد.
- مشاهدات را در قالب یک شبکه‌ی منظم دوبعدی مرتب‌شده نمایش داده و سپس مشاهدات مدل با گره‌های شبکه همبسته می‌شوند.
- محاسبه‌ی مجدد کل نقشه برای هر نمونه‌ی جدید ضروری نیست، زیرا درصورت ثابت‌فرض شدن آماره‌ها، نمونه‌ی جدید می‌تواند مستقیماً در قالب نزدیک‌ترین مشاهده‌ی مدل قدیمی نقشه قرارگیرد.

خوشه‌بندی Q، به‌معنای افراز مجموعه‌یی از داده‌ها به مجموعه‌یی از خوشه‌های Q_j، $j = ۱, ..., C$ است. شبکه‌ی عصبی کوهونن توصیف‌شده چنین عمل می‌کند:[۳۰]

مرحله‌ی ۱: واحد خروجی برنده را به‌عنوان بزرگ‌ترین سنجه‌ی مشابهت (یا کوچک‌ترین سنجه‌ی عدم تجانس) بین تمامی بردارهای وزنی w_i و بردار ورودی x انتخاب کنید. اگر فاصله‌ی اقلیدسی به‌عنوان سنجه‌ی عدم مشابهت انتخاب شود، آنگاه واحد برنده (c) معادله‌ی ۶ را ارضاء می‌کند.

$$||x - w_c|| = \min ||x - w_i|| \qquad (۶)$$

که در آن شاخص c به واحد برنده نسبت داده می‌شود.

مرحله‌ی ۲: NB_c مجموعه‌یی از شاخص متناظر با یک همسایه اطراف c برنده را مشخص می‌کند. وزن‌های برنده و واحدهای همسایه‌اش از طریق رابطه‌ی ۷ بروز می‌یابد:

$$\Delta w_i = \eta\gamma(i)(x - w_i), \qquad i \in NB_c \qquad (۷)$$

که در آن η نرخ یادگیری مثبت کوچک است. به‌جای تعریف همسایگی یک واحد برنده، می‌توان از یک تابع همسایگی $\gamma(i)$ در اطراف واحد برنده‌ی c استفاده کرد.

$$\gamma(i) = \exp\left(\frac{-||p_i - p_c||^۲}{۲\sigma^۲}\right) \qquad (۸)$$

که در آن $p_i - p_c$ فاصله‌ی فیزیکی بین واحد i و واحد برنده‌ی c است.

الگوریتم با افزایش تعداد دوره‌های آموزشی و کاهش یکنواخت پارامتر آموزشی اندازه‌ی همسایگی (نشان‌دهنده‌ی فاصله‌ی تمامی نرون‌های قرارگرفته در شعاعی مشخص از واحد برنده‌ی c)، کاهش نرخ یادگیری (ضریب تطبیق) و کاهش مقدار $\sigma^۲$ نمایش‌گر پهنای تابع همسایگی تکرار می‌شود تا وزن‌ها به ثبات برسند. در طول فرایند، حوزه‌ی داده‌های مرتبط به‌تدریج در قالب یک گروه کوچک‌تر برمبنای مفهوم همسایگی محدود خواهد شد.

یک تعریف پذیرفته‌شده از خوشه‌بندی بهینه افرازی است که فواصل بین نمونه‌های داخلی را کمینه، و فواصل بین خوشه‌ها را بیشینه کند.[۳۱] در این نوشتار نیز برای تعیین بهترین تعداد خوشه از شاخص دیویس ـ بولدین استفاده می‌شود که در آن C تعداد خوشه‌ها، S_c فاصله‌ی درون‌خوشه‌یی (مجموع فواصل بین کلیه‌ی بردارهای ورودی قرارگرفته در یک خوشه از مرکز همان خوشه)، و d_{ce} فاصله‌ی بین خوشه‌یی (مجموع فواصل بین مراکز کلیه خوشه‌ها) را نمایش می‌دهد. براساس شاخص اعتبارسنجی دیویس ـ بولدین، بهترین خوشه‌بندی رابطه‌ی ۹ را کمینه می‌کند:

$$\frac{۱}{C}\sum_{j=۱}^{C}\max_{l\neq j}\left\{\frac{S_c(Q_j) + S_c(Q_l)}{d_{ce}(Q_j, Q_l)}\right\} \qquad (۹)$$

این شاخص هم فاصله‌ی درون‌خوشه‌یی و هم فاصله‌ی بین‌خوشه‌یی را هنگام ارزیابی خوشه‌بندی حاصله مورد ارزیابی قرار می‌دهد و شاخص مناسبی برای سنجش تقسیم‌بندی SOM است، زیرا هرقدر مقدار رابطه‌ی ۹ کم‌تر باشد، نشان‌دهنده‌ی خوب‌بودن نتایج خوشه‌بندی به‌لحاظ کروی‌بودن خوشه‌ها[۱۳] است.[۳۱]

۲.۴. نرمال‌سازی داده‌ها

غالباً طراح با ویژگی‌هایی مواجه می‌شود که مقادیر در محدوده‌های متفاوتی قرار می‌گیرند. بنابراین ویژگی‌های با مقادیر بزرگ ممکن است تأثیر بیشتری به نسبت ویژگی‌های با مقادیر کوچک در تابع هزینه داشته باشند؛ اگرچه لزوماً اهمیت نسبی آن‌ها در طراحی ابزار تقسیم‌بندی را منعکس نمی‌کند. موضوع غلبه بر این مشکل از طریق نرمال‌سازی ویژگی‌هاست، به‌طوری که مقادیر آن‌ها در محدوده‌های مشابهی قرارگیرد. در این تحقیق نیز با توجه به این که، شش ویژگی منتخب بهینه مقیاس‌های متفاوتی دارند، از نرمال‌سازی با میانگین صفر و واریانس ۱ استفاده می‌شود. میانگین و انحراف معیار داده‌های ورودی با استفاده از رابطه‌ی ۱۰ محاسبه می‌شوند:

$$y_{new} = \frac{y_{old} - mean}{std}, \qquad (۱۰)$$

که در آن y_{old} مقدار اصلی، y_{new} مقدار جدید، و $mean$ و std به‌ترتیب میانگین و انحراف معیار داده‌های اصلی هستند.

۳.۴. الگوهای آرایه‌یی خوشه‌ها و نرم فاصله

هدف از خوشه‌بندی، تقسیم‌بندی بهینه‌ی مشتریان محصول مورد نظر، به‌منظور شناسایی ویژگی‌های مشترک محتمل‌ترین مشتریان برای تدوین استراتژی‌های بازاریابی

آتی و صرف بهینه‌ی منابع است. بنابراین افزایش تعداد خوشه‌ها تا جایی ادامه می‌یابد که بهبودی درکاهش شاخص دیویس ـبولدین مشاهده نشود. به‌همین دلیل تقسیم‌بندی را با دو خوشه (نرون) آغاز، و تا دست‌یابی به بهینه‌ترین حالت خوشه‌بندی ـــ با توجه به هدف تعریف‌شده ـــ ادامه داده‌ایم. در این راستا از ۲ تا ۳۰ خوشه مورد آموزش قرار گرفتند که شامل ۹۵ آرایه‌ی مختلف می‌شود: (۲ ۱)، (۳ ۱)، (۴ ۱)، (۵ ۱)، (۶ ۱)، (۷ ۱)، (۸ ۱)، (۹ ۱)، (۱۰ ۱)، (۱۱ ۱)، (۱۲ ۱)، (۱۳ ۱)، (۱۴ ۱)، (۱۵ ۱)، (۱۶ ۱)، (۱۷ ۱)، (۱۸ ۱)، (۱۹ ۱)، (۲۰ ۱)، (۲۱ ۱)، (۲۲ ۱)، (۲۳ ۱)، (۲۴ ۱)، (۲۵ ۱)، (۲۶ ۱)، (۲۷ ۱)، (۲۸ ۱)، (۲۹ ۱)، (۳۰ ۱)، (۱ ۲)، (۲ ۲)، (۳ ۲)، (۴ ۲)، (۵ ۲)، (۶ ۲)، (۷ ۲)، (۸ ۲)، (۹ ۲)، (۱۰ ۲)، (۱۱ ۲)، (۱۲ ۲)، (۱۳ ۲)، (۱۴ ۲)، (۱۵ ۲)، (۱ ۳)، (۲ ۳)، (۳ ۳)، (۴ ۳)، (۵ ۳)، (۶ ۳)، (۷ ۳)، (۸ ۳)، (۹ ۳)، (۱۰ ۳)، (۱ ۴)، (۲ ۴)، (۳ ۴)، (۴ ۴)، (۵ ۴)، (۶ ۴)، (۷ ۴)، (۱ ۵)، (۲ ۵)، (۳ ۵)، (۴ ۵)، (۵ ۵)، (۶ ۵)، (۱ ۶)، (۲ ۶)، (۳ ۶)، (۴ ۶)، (۵ ۶)، (۱ ۷)، (۲ ۷)، (۳ ۷)، (۴ ۷)، (۱ ۸)، (۲ ۸)، (۳ ۸)، (۱ ۹)، (۲ ۹)، (۳ ۹)، (۱ ۱۰)، (۲ ۱۰)، (۳ ۱۰)، (۱ ۱۱)، (۲ ۱۱)، (۱ ۱۲)، (۲ ۱۲)، (۱ ۱۳)، (۲ ۱۳)، (۱ ۱۴)، (۲ ۱۴)، (۱ ۱۵)، (۲ ۱۵).

یادآور می‌شود تعداد خوشه‌ها در هر آرایه از ضرب دو مقدار آرایه در یکدیگر حاصل می‌شود. همچنین برای سنجش فواصل درون‌خوشه‌یی و بین خوشه‌یی از معیار فاصله‌ی اقلیدسی استفاده می‌شود، بدین‌ترتیب که با فرض در اختیار داشتن یک ماتریس داده‌یی $y\,(k*n)$ که شامل $k\,(n*1)$ بردار ورودی $y_1, y_2, ..., y_n$ باشد، فواصل مختلف بین بردار y_r و y_s به‌عنوان فاصله‌ی اقلیدسی و با استفاده از رابطه‌ی ۱۱ محاسبه می‌شود:

$$d_{rs} = \sqrt{((y_r - y_s).^{\wedge} 2)} \qquad (11)$$

۴.۴. آموزش SOM

نقشه‌ی خودسازمانده با استفاده از الگوریتم آموزشی دسته‌یی[14] برای تقسیم‌بندی مجموعه داده‌های ۳۴۸ مشتری محصول مورد نظر آموزش داده شده است. در این الگوریتم کل مجموعه داده‌ها یک‌باره، و پیش از این که هیچ وزنی به‌روز شود، به شبکه ارائه می‌شود. سپس الگوریتم یک نرون برنده را برای هر بردار ورودی تعیین می‌کند. در ادامه، هر بردار وزنی به موقعیت میانگین کلیه بردارهای ورودی، که یک برنده است یا در همسایگی یک برنده است حرکت می‌کند.

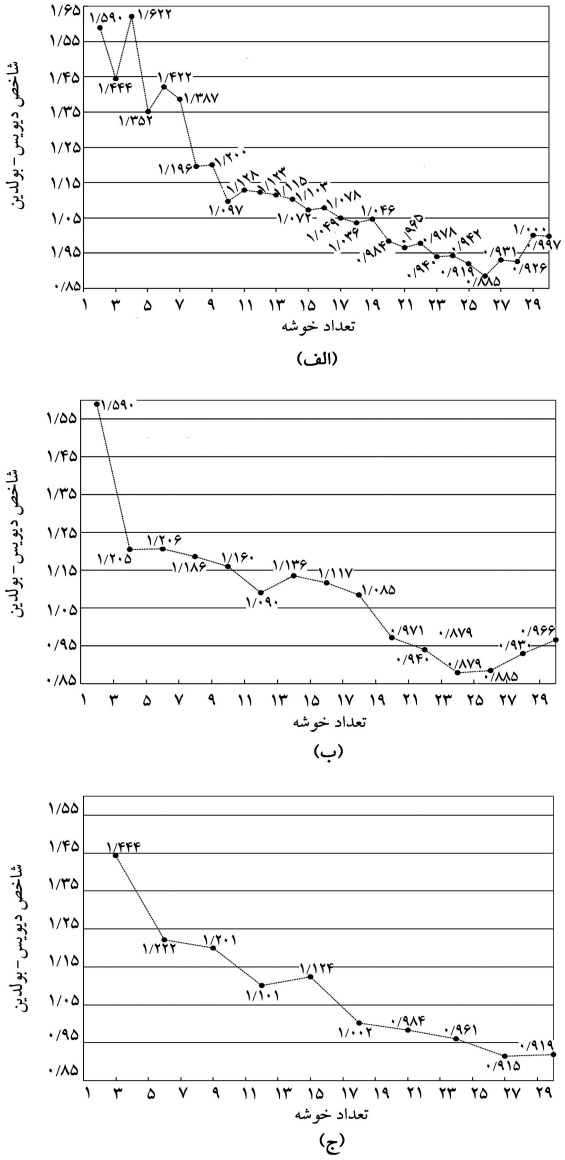
یادآور می‌شود نرخ یادگیری و فاصله‌ی همسایگی در قالب دو مرحله تغییر داده می‌شوند: ۱. ترتیب؛ ۲. تنظیم. در این تحقیق، مرحله‌ی «ترتیب» با انجام ۱۰۰۰ گام پایان می‌پذیرد. در طول این مرحله، نرخ یادگیری از ۰٫۹ آغاز و تا ۰٫۰۲ تعدیل می‌شود و فاصله‌ی همسایگی از بیشترین فاصله‌ی نرونی تا ۱ تعدیل می‌شود. در طول این مرحله انتظار می‌رود که وزن‌های نرونی خودشان را در فضای ورودی سازگار با موقعیت‌های نرونی همبسته مرتب کنند. در طول مرحله‌ی «تنظیم» نرخ یادگیری از ۰٫۰۲ به کندی کاهش می‌یابد و فاصله‌ی همسایگی همیشه برابر با ۱ است. در طول این مرحله انتظار می‌رود وزن‌ها تقریباً به‌صورت تصادفی در کل فضای ورودی پراکنده شوند در حالی که نظم توپولوژیکی‌شان در مرحله‌ی ترتیب را حفظ می‌کنند.

به این ترتیب، نقشه‌های ویژگی[15] در حین یادگیری گروه‌بندی ورودی‌شان، توپولوژی و توزیع ورودی‌شان را نیز یاد می‌گیرند.[۳۲و۳۳] همچنین، هریک از شبکه‌های عصبی SOM با ۲۰۰۰ دوره‌ی آموزشی مورد آموزش قرار می‌گیرد.

۵.۴. نتایج خوشه‌بندی

چنان که پیش‌تر نیز بیان شد، SOM برای ۹۵ آرایه‌ی مختلف با به‌کارگیری الگوریتم آموزشی دسته‌یی و در دو مرحله‌ی ترتیب و تنظیم آموزش داده شد. نتایج حاصل از آموزش هریک از آرایه‌ها، مختصات مراکز هریک از نرون‌ها و مکان قرارگیری هر بردار ورودی درون هریک از نرون‌ها را ارائه می‌کند. سپس برای تعیین بهترین تعداد خوشه از میان آرایه‌های مختلف، شاخص اعتبارسنجی دیویس ـبولدین با استفاده از نرم فاصله‌ی اقلیدسی محاسبه می‌شود. با مقایسه‌ی نتایج حاصله برای آرایه‌های مختلف بهینه‌ترین حالت خوشه‌بندی مشخص می‌شود. در شکل ۲ روند مقادیر شاخص دیویس ـبولدین برای الگوی آرایه‌های دارای عنصر اول ۱ و عنصر دوم از ۲ تا ۳۰ (شکل الف)، عنصر اول ۲ و عنصر دوم از ۱ تا ۱۵ (شکل ب)، و عنصر اول ۳ و عنصر دوم از ۱ تا ۱۰ (شکل ج) نمایش داده شده است.

چنان که در شکل ۲ مشاهده می‌شود، کم‌ترین مقدار شاخص دیویس ـبولدین از



شکل ۲. مقادیر شاخص دیویس ـبولدین برای الگوی آرایه‌های دارای عنصر اول ۱، ۲ و ۳.

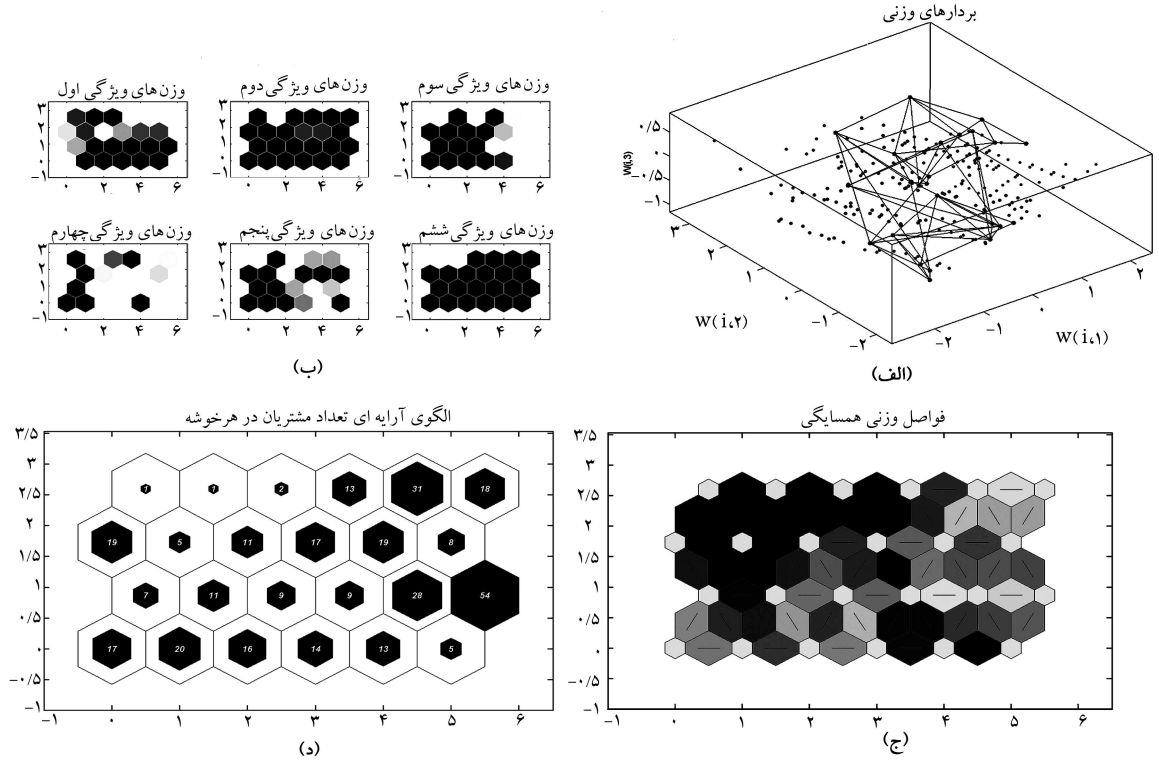
میان ۲۹ آرایه‌ی اجراشده با عنصر اول ۱ مربوط به آرایه‌ی (۲۶ ۱) با مقدار ۰٫۸۸۵، میان ۱۵ آرایه با عنصر اول ۲ مربوط به آرایه‌ی (۱۲ ۲) با مقدار ۰٫۸۷۹ و میان ۱۰ آرایه با عنصر اول ۳ مربوط به آرایه‌ی (۹ ۳) با مقدار ۰٫۹۱۵ است که میان این سه آرایه، الگوی آرایه‌یی (۱۲ ۲) مقدار کم‌تری را ارائه می‌کند. همچنین محاسبات شاخص دیویس ـبولدین برای ۴۱ آرایه‌ی دیگر نامبرده نیز انجام شد. جدول ۲ مقادیر شاخص دیویس ـ بولدین برای سایر الگوهای آرایه‌یی را نشان می‌دهد.

براساس جدول ۲، محاسبات نشان می‌دهد که میان این آرایه‌ها، الگوی آرایه‌یی (۴ ۶) با مقدار ۰٫۸۷۷ کم‌ترین مقدار را دارد که در مقایسه با آرایه‌ی (۱۲ ۲) نیز مقدارش کم‌تر است ــ اگرچه هر دو آرایه ۲۴ خوشه را به‌عنوان حالت بهینه برای خوشه‌بندی مشتریان نمایش می‌دهند. بنابراین، الگوی آرایه‌یی (۴ ۶) به‌عنوان بهینه‌ترین حالت برای تقسیم‌بندی مشتریان تعیین می‌شود. در شکل ۳، چیدمان بهینه‌ی قرارگیری خوشه‌های مشتریان مبتنی بر شش ویژگی بهینه و تعداد مشتریان هر خوشه در مجموعه‌ی آموزشی نمایش داده شده است.

در شکل ۳الف تعداد ۳۴۸ بردار ورودی (مشتری) و چگونگی خوشه‌بندی فضای ورودی با چیدمان ۲۴ بردار وزنی هر نرون (مرکز یک خوشه از بردارهای ورودی) متصل شده به نرون‌های همسایه نشان داده شده است. با توجه به وجود شش

جدول ۲. مقادیر شاخص دیویس ـ بولدین برای سایر الگوهای آرایه‌یی.

عنصر دوم / عنصر اول	۱	۲	۳	۴	۵	۶	۷
۴	۱٫۶۲۲	۱٫۱۶۴	۱٫۱۳۲	۱٫۰۵۷	۱٫۰۶۱	۰٫۹۶۲	۰٫۹۱۶
۵	۱٫۳۵۲	۱٫۰۹۲	۱٫۰۸۳	۱٫۰۲۷	۰٫۹۱۸	۰٫۹۹۹	
۶	۱٫۴۲۲	۱٫۱۳۳	۱٫۰۹۸	۰٫۸۷۷	۰٫۸۸۱		
۷	۱٫۳۸۷	۱٫۰۷۹	۱٫۰۱۵	۰٫۸۸۸			
۸	۱٫۱۹۶	۱٫۰۱۹	۰٫۹۲۷				
۹	۱٫۲۰۰	۱٫۰۶۴	۰٫۹۳۱				
۱۰	۱٫۰۹۷	۱٫۰۵۰	۰٫۹۸۰				
۱۱	۱٫۱۲۸	۰٫۹۳۷					
۱۲	۱٫۱۲۳	۰٫۹۶۲					
۱۳	۱٫۱۱۵	۰٫۸۸۸					
۱۴	۱٫۱۰۳	۰٫۹۲۴					
۱۵	۱٫۰۷۲	۰٫۸۸۲					



شکل ۳. چیدمان بهینه‌ی قرارگیری خوشه‌های مشتریان مبتنی بر شش ویژگی بهینه و تعداد مشتریان هر خوشه در مجموعه‌ی آموزشی.

جدول ۳. ویژگی‌های مشتریان هر خوشه.

شماره‌ی خوشه	ویژگی‌های مشتریان
۱	نوع مشتری(۶-۱۰)، کارگر بی‌تجربه(۰-۲)، بیمه شخص ثالث(۲،۰)، بیمه‌نامه‌ی خودرو(۰)، بیمه‌نامه‌ی آتش‌سوزی(۴،۳،۰)، بیمه‌نامه‌ی قایق(۰)
۲	نوع مشتری(۱-۳)، کارگر بی‌تجربه(۰-۳)، بیمه شخص ثالث(۲،۰)، بیمه‌نامه‌ی خودرو(۰)، بیمه‌نامه‌ی آتش‌سوزی(۳،۰)، بیمه‌نامه‌ی قایق(۰)
۳	نوع مشتری(۱-۳)، کارگر بی‌تجربه(۱،۰)، بیمه شخص ثالث(۰)، بیمه‌نامه‌ی خودرو(۶)، بیمه‌نامه‌ی آتش‌سوزی(۰)، بیمه‌نامه‌ی قایق(۰)
۴	نوع مشتری(۱-۳)، کارگر بی‌تجربه(۱،۰)، بیمه شخص ثالث(۱،۰)، بیمه‌نامه‌ی خودرو(۶،۵)، بیمه‌نامه‌ی آتش‌سوزی(۴،۳)، بیمه‌نامه‌ی قایق(۰)
۵	نوع مشتری(۱-۳)، کارگر بی‌تجربه(۰-۳)، بیمه شخص ثالث(۰-۲)، بیمه‌نامه‌ی خودرو(۰)، بیمه‌نامه‌ی آتش‌سوزی(۲-۴)، بیمه‌نامه‌ی قایق(۰)
۶	نوع مشتری(۱-۳)، کارگر بی‌تجربه(۰-۲)، بیمه شخص ثالث(۲)، بیمه‌نامه‌ی خودرو(۵،۶)، بیمه‌نامه‌ی آتش‌سوزی(۱،۰)، بیمه‌نامه‌ی قایق(۰)
۷	نوع مشتری(۷-۹)، کارگر بی‌تجربه(۴،۵)، بیمه شخص ثالث(۰)، بیمه‌نامه‌ی خودرو(۰)، بیمه‌نامه‌ی آتش‌سوزی(۰)، بیمه‌نامه‌ی قایق(۰)
۸	نوع مشتری(۱،۳،۵)، کارگر بی‌تجربه(۲-۴)، بیمه شخص ثالث(۰)، بیمه‌نامه‌ی خودرو(۶)، بیمه‌نامه‌ی آتش‌سوزی(۱،۰)، بیمه‌نامه‌ی قایق(۱،۰)
۹	نوع مشتری(۱-۳،۵)، کارگر بی‌تجربه(۲،۳،۵)، بیمه شخص ثالث(۰)، بیمه‌نامه‌ی خودرو(۶)، بیمه‌نامه‌ی آتش‌سوزی(۳،۴،۶)، بیمه‌نامه‌ی قایق(۰)
۱۰	نوع مشتری(۱-۳)، کارگر بی‌تجربه(۴،۵)، بیمه شخص ثالث(۲)، بیمه‌نامه‌ی خودرو(۶)، بیمه‌نامه‌ی آتش‌سوزی(۳-۵)، بیمه‌نامه‌ی قایق(۰)
۱۱	نوع مشتری(۱-۳،۵)، کارگر بی‌تجربه(۲،۳)، بیمه شخص ثالث(۲)، بیمه‌نامه‌ی خودرو(۶)، بیمه‌نامه‌ی آتش‌سوزی(۲-۴)، بیمه‌نامه‌ی قایق(۰)
۱۲	نوع مشتری(۱-۳،۵)، کارگر بی‌تجربه(۰،۱)، بیمه شخص ثالث(۲،۳)، بیمه‌نامه‌ی خودرو(۶)، بیمه‌نامه‌ی آتش‌سوزی(۳-۵)، بیمه‌نامه‌ی قایق(۰)
۱۳	نوع مشتری(۹،۶،۸)، کارگر بی‌تجربه(۰-۳)، بیمه شخص ثالث(۰)، بیمه‌نامه‌ی خودرو(۶،۵)، بیمه‌نامه‌ی آتش‌سوزی(۰)، بیمه‌نامه‌ی قایق(۰)
۱۴	نوع مشتری(۹،۳،۸)، کارگر بی‌تجربه(۰-۲)، بیمه شخص ثالث(۰)، بیمه‌نامه‌ی خودرو(۶،۰)، بیمه‌نامه‌ی آتش‌سوزی(۰)، بیمه‌نامه‌ی قایق(۲-۴)
۱۵	نوع مشتری(۶-۱۰)، کارگر بی‌تجربه(۰-۳)، بیمه شخص ثالث(۲،۰)، بیمه‌نامه‌ی‌خودرو (۶،۵)، بیمه‌نامه‌ی‌آتش‌سوزی(۳-۶)، بیمه‌نامه‌ی‌قایق(۰-۲)
۱۶	نوع مشتری(۷-۵،۹)، کارگر بی‌تجربه(۸،۵،۷،۴)، بیمه شخص ثالث(۰)، بیمه‌نامه‌ی خودرو(۶)، بیمه‌نامه‌ی آتش‌سوزی(۰،۳)، بیمه‌نامه‌ی قایق(۰)
۱۷	نوع مشتری(۷،۵،۳-۹)، کارگر بی‌تجربه(۴-۶)، بیمه شخص ثالث(۲،۱)، بیمه‌نامه‌ی خودرو(۶،۵)، بیمه‌نامه‌ی آتش‌سوزی(۰،۲-۴)، بیمه‌نامه‌ی قایق(۰)
۱۸	نوع مشتری(۵-۹)، کارگر بی‌تجربه(۰-۲)، بیمه شخص ثالث(۲)، بیمه‌نامه‌ی خودرو(۶،۵)، بیمه‌نامه‌ی آتش‌سوزی(۰-۲)، بیمه‌نامه‌ی قایق(۰)
۱۹	نوع مشتری(۷)، کارگر بی‌تجربه(۲)، بیمه شخص ثالث(۲)، بیمه‌نامه‌ی خودرو(۰)، بیمه‌نامه‌ی آتش‌سوزی(۴)، بیمه‌نامه‌ی قایق(۶)
۲۰	نوع مشتری(۵)، کارگر بی‌تجربه(۶)، بیمه شخص ثالث(۰)، بیمه‌نامه‌ی خودرو(۶)، بیمه‌نامه‌ی آتش‌سوزی(۳)، بیمه‌نامه‌ی قایق(۶)
۲۱	نوع مشتری(۱،۹)، کارگر بی‌تجربه(۱،۳)، بیمه شخص ثالث(۲)، بیمه‌نامه‌ی خودرو(۵)، بیمه‌نامه‌ی آتش‌سوزی(۴)، بیمه‌نامه‌ی قایق(۴)
۲۲	نوع مشتری(۷-۹)، کارگر بی‌تجربه(۴،۶،۷-۲،۰)، بیمه شخص ثالث(۰-۲)، بیمه‌نامه‌ی خودرو(۰)، بیمه‌نامه‌ی آتش‌سوزی(۳،۴)، بیمه‌نامه‌ی قایق(۰)
۲۳	نوع مشتری(۷-۹)، کارگر بی‌تجربه(۳،۲)، بیمه شخص ثالث(۱-۳)، بیمه‌نامه‌ی خودرو(۵،۶)، بیمه‌نامه‌ی آتش‌سوزی(۳-۵)، بیمه‌نامه‌ی قایق(۰)
۲۴	نوع مشتری(۸-۱۰)، کارگر بی‌تجربه(۱،۰)، بیمه شخص ثالث(۲)، بیمه‌نامه‌ی خودرو(۵،۶)، بیمه‌نامه‌ی آتش‌سوزی(۴،۶،۳)، بیمه‌نامه‌ی قایق(۰)

ویژگی (بعد)، از تحلیل مؤلفه‌ی اصلی برای نمایش موقعیت وزنی خوشه‌ها و هریک از مشتریان در فضای سه بعدی استفاده شده است. در شکل ۳ب، یک صفحه‌ی وزنی برای هریک از شش عنصر بردار ورودی (نوع مشتری، کارگر بی‌تجربه، بیمه‌ی شخص ثالث شخصی، بیمه‌نامه‌ی خودرو، بیمه‌نامه‌ی آتش‌سوزی و بیمه‌نامه‌ی قایق به‌ترتیب ورودی) به‌صورت جداگانه در یک فضای دوبعدی نشان داده شده است. هریک از عناصر تصاویری از وزن‌ها هستند که هر ورودی را به یکی از نرون‌ها متصل می‌کند. هر قدر رنگ خوشه تیره‌تر باشد، نشان‌دهنده‌ی وزن بیشتر است. اگر الگوهای اتصال دو ورودی شباهت بسیاری به یکدیگر داشته باشند، می‌توان نتیجه گرفت که آن دو ورودی همبستگی شدیدی دارند. اما در خصوص این شش ورودی اتصالات بسیار متفاوت‌اند. در شکل ۳ج نیز فاصله‌ی بردارهای وزنی (مرکز خوشه‌ها) هر نرون از نرون‌های همسایه با لحاظ هر شش ویژگی به‌صورت همزمان نشان داده شده است. میزان تیرگی رنگ فاصله‌ی بین خوشه‌ها، نشان‌دهنده‌ی فاصله بیشتر مرکز خوشه از خوشه‌های مجاور است. به‌طور کلی، فاصله‌ی همسایگی میان خوشه‌های قرارگرفته در سمت چپ بالای تصویر بسیار بیشتر از سایر مناطق شکل فاصله همسایگی است. بین دو خوشه‌ی ۱۳ و ۱۹ و دو خوشه‌ی ۱۵ و ۲۰ که فاصله‌ی همسایگی در آن‌ها تیره‌تر نشان داده شده، بیشترین فاصله بین مراکز خوشه‌ها دیده می‌شود. در شکل ۳د، تعداد بردارهای ورودیِ در برگرفته توسط هر خوشه نشان داده شده است. در میان خوشه‌ها، خوشه‌ی ۱۲ با ۵۴ بردار ورودی بیشترین تعداد مشتری را شامل می‌شود. جدول ۳ ویژگی‌های مشتریان هر خوشه را نشان می‌دهد.

براساس اطلاعات در دسترس، اعداد داخل پرانتز درمورد کارگر بی‌تجربه از ۰ تا ۹؛ درمورد بیمه‌نامه‌ی خودرو شامل ۰، ۴ تا ۸؛ درمورد بیمه‌نامه‌ی آتش‌سوزی از ۰ تا ۸؛ و درمورد بیمه‌نامه‌ی قایق از ۰ تا ۶ هستند. به‌علاوه، دو جدول ۴ و ۵ دسته‌بندی نوع مشتری و مقادیر بیمه شخص ثالث را توصیف می‌کنند.

همچنین به‌منظور صرف بهینه‌ی منابع شرکت در قالب هزینه‌های بازاریابی، باید از شاخصی برای سنجش آن بهره جست. شاخص PPC که از آن برای محاسبه‌ی

جدول ۴. نوع مشتری.

ردیف	نوع مشتری
۱	لذت‌جوهای موفق
۲	پرورش‌دهندگان گیاهان
۳	خانواده‌ی متوسط
۴	افراد تنها در کل زندگی
۵	افراد بدون نگرانی مالی
۶	افراد مهم در حال سفر
۷	بازنشسته و دیندار
۸	خانواده‌یی با افراد بزرگسال
۹	خانواده‌های محافظه‌کار
۱۰	کشاورزان

جدول ۵. مقادیر بیمه شخص ثالث.

ردیف	مقادیر بیمه
۰	۰
۱	۱-۴۹
۲	۵۰-۹۹
۳	۱۰۰-۱۹۹

جدول ۶. بهینه‌ترین خوشه‌ها برای صرف هزینه‌های بازاریابی و خوشه‌بندی مشتریان مجموعه داده‌های ارزیابی.

شماره‌ی خوشه	مجموعه آموزشی			مجموعه اعتبارسنجی		
	PC	TC	PPC (٪)	PC	TC	PPC (٪)
۱	۱۷	۵۹۷	۲٫۸	۱۲	۴۱۸	۲٫۹
۲	۲۰	۶۱۷	۳٫۲	۱۲	۴۳۴	۲٫۸
۳	۱۶	۲۵۷	۶٫۲	۸	۱۷۳	۴٫۶
۴	۱۴	۷۷	۱۸٫۲	۵	۵۶	۸٫۹
۵	۱۳	۴۴۴	۲٫۹	۲۱	۲۹۵	۷٫۱
۶	۵	۳۳	۱۵٫۲	۰	۲۴	۰٫۰
۷	۷	۵۶۴	۱٫۲	۶	۳۹۳	۱٫۵
۸	۱۱	۲۷۲	۴٫۰	۵	۱۸۴	۲٫۷
۹	۹	۶۹	۱۳٫۰	۱۰	۵۰	۲۰٫۰
۱۰	۹	۳۸	۲۳٫۷	۷	۳۲	۲۱٫۹
۱۱	۲۸	۲۰۰	۱۴٫۰	۱۱	۱۴۰	۷٫۹
۱۲	۵۴	۲۵۷	۲۱٫۰	۳۷	۱۶۸	۲۲٫۰
۱۳	۱۹	۴۷۵	۴٫۰	۱۷	۳۴۸	۴٫۹
۱۴	۵	۱۱	۴۵٫۵	۱	۲	۵۰٫۰
۱۵	۱۱	۱۶۹	۶٫۵	۸	۱۱۶	۶٫۹
۱۶	۱۷	۲۸۸	۵٫۹	۱۰	۱۶۸	۶٫۰
۱۷	۱۹	۲۵۴	۷٫۵	۱۲	۱۵۰	۸٫۰
۱۸	۸	۱۲۶	۶٫۳	۶	۷۵	۸٫۰
۱۹	۱	۴	۲۵٫۰	۰	۰	-
۲۰	۱	۱	۱۰۰	۰	۰	-
۲۱	۲	۱۰	۲۰٫۰	۲	۷	۲۸٫۶
۲۲	۱۳	۶۰۷	۲٫۱	۱۵	۴۳۸	۳٫۴
۲۳	۳۱	۲۹۸	۱۰٫۴	۱۹	۱۹۸	۹٫۶
۲۴	۱۸	۱۵۴	۱۱٫۷	۱۴	۱۳۱	۱۰٫۷
مجموع	۳۴۸	۵۸۲۲	۵٫۹۸	۲۳۸	۴۰۰۰	۵٫۹۵
۲۰٪ برتر	۱۷۷	۱۱۵۲	۱۵٫۳۶	۱۰۶	۸۰۸	۱۳٫۱۲

درصد مشتریان محصول بیمه‌نامه‌ی کاروان در هر خوشه استفاده شده، در قالب رابطه‌ی ۱۲ تعریف می‌شود:

$$PPC = \frac{PC}{TC} \qquad (۱۲)$$

که در آن PC نشان‌گر مشتریان محصول در هر خوشه، و TC نشان‌گر کل مشتریان شرکت در هر خوشه است. هرچه مقدار این شاخص بزرگ‌تر باشد، صرف هزینه‌های بازاریابی برای این قبیل از مشتریان با احتمال برگشت سرمایه‌ی بیشتری همراه است. به‌علاوه، چنان که پیش‌تر نیز بیان شد، برای اطمینان از عملکرد مناسب سامانه‌ی هوشمند طراحی‌شده، لازم است عمکرد سامانه مورد سنجش قرار گیرد. بدین‌منظور از مجموعه داده‌های ارزیابی برای اعتبارسنجی سامانه استفاده می‌شود. در جدول ۶ بهینه‌ترین خوشه‌ها برای صرف هزینه‌های بازاریابی و خوشه‌بندی مشتریان مجموعه داده‌های ارزیابی مبتنی بر بردارهای وزنی تعیین شده برای هر خوشه در مجموعه آموزشی نمایش داده شده است.

چنان که در جدول ۶ مشاهده می‌شود،براساس بردارهای وزنی بدست آمده برای ۲۴ خوشه، اقدام به خوشه‌بندی ۵۸۲۲ مشتری کلیه محصولات شرکت در مجموعه آموزشی کرده‌ایم. بدین‌ترتیب هر مشتری با توجه به میزان شباهت ویژگی‌هایش به مشتریان هریک از خوشه‌ها در آن خوشه قرار می‌گیرد. ستون سوم جدول تعداد کل مشتریان در هر خوشه را نشان می‌دهد. سپس، از تقسیم مشتریان محصول بیمه‌نامه‌ی کاروان در هر خوشه بر تعداد کل مشتریان آن خوشه، درصد مشتریان محصول بیمه‌نامه‌ی کاروان در آن خوشه حاصل می‌شود. با مشاهده‌ی ستون چهارم جدول می‌توان متوجه شد که خوشه‌های ۴، ۶، ۹، ۱۰، ۱۱، ۱۲، ۱۴، ۱۹، ۲۰، ۲۱، ۲۳ و ۲۴ بالاترین درصد مشتریان بیمه‌نامه‌ی کاروان را در بر می‌گیرند. به‌طوری که، این ۱۲ خوشه با داشتن نزدیک به ۲۰٪ (۱۱۵۲/۵۸۲۲) ازکل مشتریان، ۵۰٫۹٪ (۱۷۷/۳۴۸) از مشتریان محصول بیمه‌نامه‌ی کاروان را پوشش می‌دهند. بدین‌ترتیب، درصد مجموع مشتریان محصول بیمه‌نامه‌ی کاروان به کل مشتریان شرکت با بهبودی قابل توجه از ۵٫۹۸٪ به ۱۵٫۳۶٪ افزایش می‌یابد. درنتیجه، شرکت می‌تواند با تمرکز بر مشتریانی با ویژگی‌های مشابه با این ۱۲ خوشه منابع بازاریابی خود را به‌صورت بهینه‌تری هزینه کند. در مقابل، تمرکز بر مشتریان شش خوشه‌ی ۱، ۲، ۷، ۸، ۱۳ و ۲۲ تنها هزینه‌بر خواهد بود، زیرا مقدار شاخص در این خوشه‌ها کم‌تر از مقدار متوسط (یعنی ۵٫۹۸٪) است.

عموماً افرادی که بیمه‌نامه‌ی خودرو خریداری می‌کنند، محتمل‌ترین افراد برای خرید بیمه‌نامه‌ی کاروان هستند و افرادی که مقادیر ۴، ۷ یا ۸ به بیمه‌نامه‌ی خودرو آن‌ها تخصیص یافته، قطعاً مشتریان این محصول نیستند. به‌علاوه، احتمال خرید گروه «افراد تنها درکل زندگی» وجود نداشته است. یکی از بزرگ‌ترین شگفتی‌ها در تحلیل ما این حقیقت بود که گروه «افراد مهم در حال سفر» بعید به نظر می‌رسد که اقدام به خرید بیمه‌نامه‌ی کاروان کنند. اگر ویژگی کارگر بی‌تجربه مقدار ۹ داشته باشد، احتمال خرید وجود ندارد. اگر بیمه‌نامه‌ی آتش‌سوزی مقدار ۷ یا ۸ بگیرد، احتمال خرید برابر صفر است همچنان که اگر بیمه‌نامه‌ی قایق مقدار بیش از ۲ بگیرد، احتمال خرید بسیار زیاد است. البته این تحلیل‌ها اطلاعاتی کلی را ارائه می‌کنند و اطلاعات دقیق‌تر از داخل هر خوشه استخراج می‌شود.

در ادامه، خوشه‌بندی مشتریان مبتنی بر بردارهای وزنی تعیین شده برای هر خوشه در مجموعه آموزشی بر روی ۴۰۰۰ مشتری شرکت در مجموعه داده ارزیابی نیز انجام شد. باید توجه داشت که همچون مجموعه‌ی آموزشی خوشه‌های ۴، ۹، ۱۰، ۱۱، ۱۲، ۱۴، ۲۱، ۲۳ و ۲۴ بالاترین مقدار شاخص را به خود اختصاص می‌دهند. با توجه به کم‌بودن مراجعه‌ی مشتریانی با ویژگی‌های مشابه با مشتریان خوشه‌های ۱۹ و ۲۰، در مجموعه ارزیابی چنین مشتریانی وجود نداشته‌اند. اما نمی‌توان این دو خوشه را از میان خوشه‌های مناسب حذف کرد، زیرا به‌احتمال فراوان، این مشتریان درصورت مراجعه به شرکت، مشتری محصول بیمه‌نامه‌ی کاروان نیز خواهند بود. لیکن، تنها درمورد مشتریان خوشه‌ی ۶ باید تأمل بیشتری کرد. بهتر است در تخصیص منابع بازاریابی مشتریان این خوشه پس از ۱۱ خوشه‌ی مناسب‌تر قرار گیرد. در مجموع، ۱۲ خوشه‌ی نامبرده با پوشش نزدیک به ۲۰٪ (۸۰۸/۴۰۰۰) از کل مشتریان ۴۴٫۵٪ (۱۰۶/۲۳۸) از مشتریان محصول بیمه‌نامه‌ی کاروان را دارند. با این اوصاف، درصد مجموع مشتریان محصول بیمه‌نامه‌ی کاروان به کل مشتریان شرکت با بهبودی چشم‌گیر از ۵٫۹۵٪ به ۱۳٫۱۲٪ افزایش می‌یابد. درخصوص

خوشه‌های هزینه‌بر نیز مجموعه‌ی اعتبارسنجی بر دقت تشخیص شش خوشه‌ی نامبرده صحه می‌گذارد. بنابراین، مجموعه داده‌ی ارزیابی مؤید کارآمدی و دقت قابل توجه خوشه‌بندی حاصله است. بدین‌ترتیب شرکت می‌تواند بنابر ویژگی‌های مشتریان بهینه‌ترین خوشه‌ها اقدام به تدوین استراتژی‌های بازاریابی خود برای ترغیب و جذب مشتریان مشابه کند. درصورتی که منابع بازاریابی شرکت محدود نباشد، شش خوشه‌یی (با احتمال جذب کم‌تر) که نام آن‌ها آورده نشد می‌توانند گزینه‌های بعدی آن باشند.

۵. نتیجه‌گیری

پیشرفت‌های اخیر در انتخاب ویژگی مسئله‌ی بهبود عملکرد تقسیم‌بندی مشتریان را از نقطه‌نظر عملی مورد توجه قرار می‌دهد. موضوعی که به‌ندرت در ادبیات به آن پرداخته‌اند. در این نوشتار، تلاش کردیم یک سامانه‌ی هوشمند تلفیقی برای تقسیم‌بندی بهینه‌ی مشتریان یک محصول مشخص، به‌نام بیمه‌نامه‌ی کاروان، توسعه دهیم. رویکرد جدیدی برای تقسیم‌بندی مشتریان با طراحی یک شبکه عصبی SOM بهینه‌شده مبتنی بر شش ویژگی منتخب بهینه با به‌کارگیری درخت رگرسیونی هرس‌شده پیشنهاد شده، به‌طوری که کم‌ترین نرخ میانگین مربعات خطا را برای تعیین بهترین درخت رگرسیونی هرس شده و شاخص دیویس ـ بولدین را برای تعیین بهترین الگوی آرایه‌یی لحاظ کردیم. این سامانه‌ی هوشمند تلفیقی به یک بهینه‌سازی ترکیبی نامحدود منجر شد که در آن نرخ MSE و شاخص نامبرده معیار جست‌وجو هستند. معیار کم‌ترین میانگین مربعات خطا برای تعیین نقطه‌ی بهینه برای هرس درخت رگرسیونی به کار برده شده بود که نهایتاً منجر به انتخاب شش ویژگی شد. به‌کارگیری این رویه‌ی انتخاب ویژگی پیشنهادی منجر به بهبودی قابل توجه درکاهش ویژگی‌ها شد. این رویه همچنین سرعت و دقت محاسباتی بسیار زیادی در انتخاب ویژگی دارد. سپس تقسیم‌بندی مشتریان با توسعه‌ی SOM براساس ویژگی‌های منتخب انجام شد، به‌طوری که به‌کارگیری SOM مبتنی بر یک شاخصِ مناسب نتایج قابل توجهی در بر دارد. نتایج نشان داد امکان صرف منابع بازاریابی کم‌تر برای دست‌یابی به مشتریانی با احتمال جذب بیشتر وجود دارد. با انتخاب نیمی از خوشه‌ها شامل ۲۰٪ از کل مشتریان می‌توان حدود ۵۰٪ از مشتریان بیمه‌نامه‌ی کاروان را مورد پوشش قرار داد. همچنین به‌منظور ارزیابی اثر بخشی سامانه‌ی هوشمند از یک مجموعه داده‌ی ارزیابی استفاده شد که انتخاب صحیح بهینه‌ترین خوشه‌ها در مجموعه‌ی آموزشی برای صرف کاراتر منابع را مورد تأیید قرار داد. بدین‌ترتیب، شرکت بیمه می‌تواند با تدوین راهکارهای بازاریابی خود برای خوشه‌های بهینه، انتظار کسب بیشترین سودآوری را داشته باشد.

نگارندگانِ این مقاله مدل خود را روی داده‌های یک شرکت بیمه اجرا کردند. اما یکی دیگر از حوزه‌های مهم کاربردی این مدل داده‌های بانک است که می‌تواند کمک شایانی به بانک در شناسایی سودآورترین مشتریان کند. استفاده از شبکه‌های عصبی به‌عنوان یک تکنیک غیرخطی برای انتخاب ویژگی بسیار زمان‌بر است، اما ممکن است منجر به انتخاب دقیق‌تر ویژگی‌ها شود. همچنین، نویسندگان بر این باورند که روش خوشه‌بندی فازی $c-means$ به‌عنوان یک تکنیک خوشه‌بندی به‌تنهایی نمی‌تواند نتیجه‌ی خوبی در بر داشته باشد، اما ترکیب آن با SOM ممکن است منجر به نتایج بهتری شود که نیاز به تحقیقات بیشتری دارد.

۶. تقدیر و تشکر

بدینوسیله از پشتیبانی مالی مرکز تحقیقات مخابرات ایران (ITRC) در پیشبرد اهداف این تحقیق تقدیر و تشکر به عمل می آید.

منابع (References)

1. Anderson, E.T. "Sharing the wealth: When should firms treat customers as partners?", *Management Science*, **48**(8), pp. 955-71 (2002).
2. Ha, S.H. "Applying knowledge engineering techniques to customer analysis in the service industry", *Advanced Engineering Informatics*, **21**, pp. 293-301 (2007).
3. Hsieh, N.C. "An integrated data mining and behavioral scoring model for analyzing bank customers", *Expert Systems with Applications*, **27**, pp. 623-633 (2004).
4. Jonker, J.J.; Piersma, N. and Poel, D.V. "Joint optimization of customer segmentation and marketing policy to maximize long-term profitability", *Expert Systems with Applications*, **27**, pp. 159-168 (2004).

پانوشت‌ها

1. customer relationship management
2. self organizing map
3. nested partition
4. simulated annealing
5. rough set
6. support vector machine
7. Davies-Bouldin index
8. classification and regression tree
9. pruning algorithm
10. cross-validation
11. goodness-of-fit
12. leave-one-out
13. spherical clusters
14. batch training algorithm
15. feature maps

5. Hwang, H.; Jung, T. and Suh, E. "An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry", *Expert Systems with Applications*, **26**, pp. 181-188 (2004).
6. Kim, S.Y.; Jung, T.S.; Suh, E.H. and Hwang, H.S. "Customer segmentation and strategy development based on customer lifetime value: A case study", *Expert Systems with Applications*, **31**, pp. 101-107 (2006).
7. Chan, C.C.H. "Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer", *Expert Systems with Applications*, **34**, pp. 2754-2762 (2008).
8. Lee, J.H. and Park, S.Ch., "Intelligent profitable customers segmentation system based on business intelligence tools", *Expert Systems with Applications*, **29**, pp. 145-152 (2005).
9. Chen C.H.; Khoo, L.P. and Yan, W. "A strategy for acquiring customer requirement patterns using laddering technique and ART2 neural network", *Advanced Engineering Informatics*, **16**, pp. 229-240 (2002).
10. Shieh, M.D.; Yan, W. and Chen, C.H., "Soliciting customer requirements for product redesign based on picture sorts and ART2 neural network", *Expert Systems with Applications*, **34**, pp. 194-204 (2008).
11. Espinoza, M.; Joye, C.; Belmans, R. and Moor, B.D. "Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series", *IEEE Transactions on Power Systems*, **20**(3), pp. 1622-1630 (2005).
12. Hu, T.L. and Sheu, J.B. "A fuzzy-based customer classification method for demand-responsive logistical distribution operations", *Fuzzy Sets and Systems*, **139**, pp. 431-450 (2003).
13. Kohavi, R. and John, G.H. "Wrappers for feature subset selection", *Artificial Intelligence*, (1-2), pp. 273-324 (1997).
14. Blum, A.L. and Rivest, R.L. "Training a 3-node neural network is NP-complete", *Neural Networks*, **5**, pp. 117-127 (1992).
15. Ng, K.S. and Liu, H. "Customer retention via data mining", *Artificial Intelligence Review*, **14**(6), pp. 569-590 (2000).
16. Kim, Y.S. "Toward a successful CRM: Variable selection, sampling, and ensemble", *Decision Support Systems*, **41**, pp. 542-553 (2006).
17. Hadden, J.; Tiwari, A.; Roy, R. and Ruta, D. "Computer assisted customer churn management: State-of-the-art and future trends", *Computers & Operations Research*, **34**, pp. 2902-2917 (2005).
18. Kim, Y.S. and Street, W.N. "An intelligent system for customer targeting: A data mining approach", *Decision Support Systems*, **37**, pp. 215-228 (2004).
19. Yu, E. and Cho, S. "Constructing response model using ensemble based on feature subset selection", *Expert Systems with Applications*, **30**, pp. 352-360 (2006).
20. Ahn, H.; Kim, K. and Han, I. "A case-based reasoning system with the two-dimensional reduction technique for customer classification", *Expert Systems with Applications*, **32**, pp. 1011-1019 (2007).
21. Yan, L. and Changrui, Y. "A new hybrid algorithm for feature selection and its application to customer recognition", *LNCS*, **4616**, pp. 102-111 (2007).
22. Tseng, T.L. and Huang, C.C. "Rough set-based approach to feature selection in customer relationship management", *Omega*, **35**, pp. 365-383 (2007).
23. Lessmann, S. and Voß, S. "A reference model for customer-centric data mining with support vector machines", *European Journal of Operational Research*, **199**, pp. 520-530 (2009).
24. Han, J. and Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann, Second Edition (2006).
25. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, Ph.S.; Zhou, Zh.H.; Steinbach, M.; Hand, D.J. and Steinberg, D. "Top 10 algorithms in data mining", *Knowledge Information System*, **14**, pp. 1-37 (2008).
26. Liu, H.H. and Ong, Ch.Sh. "Variable selection in clustering for marketing segmentation using genetic algorithms", *Expert Systems with Applications*, **34**, pp. 502-510 (2008).
27. Breiman, L.; Friedman, J.; Olshen, R. and Stone, C. "Classification and regression trees", Boca Raton, FL: CRC Press (1984).
28. Theodoridis, S. and Koutroumbas, K. "Pattern recognition", *Academic Press,* Third Edition (2006).
29. Webb, A.R., *Statistical Pattern Recognition*, John Wiley & Sons, Ltd., Second Edition (2002).
30. Kohonen, T. "Self-organizing maps", *Springer Series*, Third Edition (2001).
31. Vesanto, J. and Alhoniemi, E. "Clustering of the self-organizing map", *IEEE Transactions on Neural Networks*, **11**(3), pp. 586-600 (2000).
32. Demuth, H.; Beale, M. and Hagan, M., *Neural Network Toolbox^{TM} User's Guide*, Version 6.0.3, The MathWorks, Inc. (2009).
33. Hagan, M.; Demuth, H. and Beale, M., *Neural Network Design*, USA, PWS Publishing Company, First Edition (1996).