
بسم الله الرحمن الرحيم

دانشکده ریاضی

گروه ریاضی کاربردی

پایان‌نامه جهت اخذ درجه کارشناسی ارشد آمار ریاضی

مدیریت ریزش مشتری با استفاده از ابزارهای پیشرفته آماری و داده‌کاوی

دانشجو: **سیدعلیرضا مهدوی‌تالارپشتی**

استاد راهنما:

دکتر داود شاهسونی

مهر ماه ۱۳۹۱

تقدیم به

پدر زحمتکش

مادر مهربان

و همسر فداکارم

تشکر و قدردانی

از استاد بزرگوارم جناب آقای دکتر شاهسونی که با راهنمایی های ارزنده ی خود مرا در به ثمر رسیدن این پایان نامه یار و همراه بوده اند صمیمانه تشکر و قدردانی می کنم.

سید علیرضا مهدوی تالارپشتی

مهر ۹۱

تعهدنامه

اینجانب **سیدعلیرضا مهدوی‌تالارپشتی** دانشجوی دوره کارشناسی ارشد رشته آمار ریاضی دانشکده ریاضی دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **"مدیریت ارتباط با مشتری با استفاده از ابزارهای پیشرفته آماری و داده‌کاوی"** تحت راهنمائی **دکتر داود شاهسونی** متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است صل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

چکیده

در سال‌های اخیر، تعدد شرکت‌هایی که خدمات نسبتاً مشابه‌ای را به جامعه ارائه می‌دهند موجب شده است تا مشتریان بر اساس نیاز خود بتوانند شرکت مورد نظر را انتخاب و از خدمات آن بهره‌مند گردند. این گوناگونی به وضوح در صنایعی از قبیل بیمه، مخابرات، موسسات ارائه‌دهنده خدمات اینترنتی و شبکه‌های تلویزیون کابلی دیده می‌شود و سبب شده است تا موضوع نگه‌داشتن مشتریان فعلی و عدم جذب آن‌ها توسط غیر، از جمله مهم‌ترین استراتژی‌های مدیریتی تلقی شود. موضوع قطع ارتباط مشتری با خدمات‌دهنده فعلی، موسوم به *ریزش مشتری* است. از دست دادن یک مشتری سودآور به معنی جذب او توسط خدمات‌دهنده رقیب بوده و هزینه جذب مشتری جدید به مراتب بیشتر از هزینه‌ی نگهداری آن است. از دیدگاه مدیریت ریسک و اقتصاد، تشخیص مشتریانی که ریزش آن‌ها مخاطره زیادی دارد، بسیار مهم و قابل توجه است. اطلاعات موجود در سوابق مشتریان، اعم از مشتریان وفادار و ریزش‌شده، مبنایی برای پیش‌گویی رفتار آینده مشتریان است. اگر بتوان بر اساس اطلاعات مربوط به ویژگی‌های مشتریان، احتمال ریزش یا عدم آن را پیش‌بینی کرد، می‌توان با انجام فعالیت‌های بازدارنده، ریزش آن‌ها را به حداقل رسانید. به منظور پیش‌بینی ریزش مشتری، روش‌های گوناگون آماری و داده‌کاوی برای رده‌بندی وجود دارند که از جمله مهم‌ترین آن‌ها می‌توان به رگرسیون لجستیک، شبکه عصبی مصنوعی، درخت تصمیم و نظریه مجموعه مبهم اشاره کرد. در این پایان‌نامه، مدل‌های مذکور را معرفی نموده و دقت و کارایی آن‌ها را بر روی یک مطالعه موردی با استفاده از معیارهای ارزیابی مختلف مورد بررسی قرار دادیم. نتایج نشان داد، معیارهای ارزیابی مختلف، مدل‌های مختلفی را برای پیش‌بینی ریزش مشتری پیشنهاد می‌کنند.

کلمات کلیدی: *مدیریت ریزش مشتری، نظریه مجموعه مبهم، شبکه عصبی مصنوعی، درخت رگرسیون و رده‌بندی، رگرسیون لجستیک*

مقالات مستخرج از پایان‌نامه:

۱- مهدوی‌تالارپشتی، س.ع.، شاهسونی، د.، (۱۳۹۱)، " کاربرد روش‌های رده‌بندی در پیش‌بینی ریزش مشتری "، یازدهمین کنفرانس آمار ایران، ص **۴۰۳-۴۱۲**، دانشگاه علم و صنعت تهران.

فهرست

فهرست اشکال

فهرست جداول

فصل اول

مقدمه

مقدمه

مشتری، عنصر و سرمایه اصلی شرکت‌ها و سازمان‌ها می‌باشد. در سال‌های اخیر تعدد شرکت‌هایی که خدمات نسبتاً مشابهی را به جامعه ارائه می‌دهند، موجب شده است تا مشتریان بر اساس نیاز خود بتوانند خدمات‌دهنده مورد نظرشان را انتخاب و از خدمات آن بهره‌مند گردند. این گوناگونی به وضوح در صنایعی مانند بیمه، مخابرات، موسسات ارائه‌دهنده خدمات اینترنتی و شبکه‌های تلویزیون کابلی دیده می‌شود. افزایش قدرت انتخاب مشتری سبب خواهد شد تا مشتری به محض نارضایتی از نحوه خدمات‌رسانی، به راحتی خدمات‌دهنده خود را رها کرده و به خدمات‌دهنده دیگری رجوع کند. به این رخداد در اصطلاح *ریزش* مشتری[1] گفته می‌شود. از طرفی، هزینه جذب یک مشتری جدید به مراتب بیشتر از هزینه نگهداری مشتریان فعلی است. از این‌رو مدیران شرکت‌ها به دنبال راه‌کارهایی برای حفظ مشتریان کنونی خود هستند. برای حفظ مشتریان، ابتدا نیاز به شناسایی مشتریانی است که ریزش آن‌ها محتمل است. بدین منظور مدیران و کارشناسان شرکت‌ها نیاز به روش‌هایی برای شناسایی مشتریان دارای خاصیت ریزش خواهند داشت تا بتوانند با شناسایی آن‌ها و اعمال استراتژی-های مدیریتی مناسب، از ریزش آن‌ها جلوگیری کنند. فرآیند شناسایی مشتریانی که ریزش آن‌ها محتمل است و نحوه جلوگیری ریزش آن‌ها، *مدیریت ریزش مشتری*[2] نامیده می‌شود.

۱-۱ ضرورت و اهداف پایان‌نامه

روش‌های آماری و داده‌کاوی[3] از جمله ابزارهای قدرتمند در تحلیل اطلاعات و استخراج دانش محسوب می‌شوند. از طرفی، استخراج دانش از داده‌های خام، همیشه موضوعی چالش برانگیز در میان متخصصین علوم مختلف بوده است. از این‌رو، همکاری متخصصان آماری و داده‌کاوی با متخصصان سایر علوم قطعاً نتایج بسیار مفیدی را به همراه خواهد داشت و کارشناسان علوم مختلف می‌توانند با بهره گرفتن از دیدگاه‌ها و روش‌های نوین آماری و داده‌کاوی، دقت تجزیه و تحلیل اطلاعات خود را

[1] Customer Churn
[2] Customer Churn Management
[3] Data Mining

افزایش دهند. این مسئله، انگیزه‌ای شد تا در این پایان‌نامه بر گوشه‌ای از قابلیت‌های روش‌های آماری و داده‌کاوی تمرکز کرده و میزان قدرت آن‌ها را نمایش دهیم.

یکی از زمینه‌هایی که آمار و داده‌کاوی در آن به ایفای نقش می‌پردازند، مدیریت ریزش مشتری است که امروزه از آن به عنوان یکی از دغدغه‌های اصلی مدیران شرکت‌ها و سازمان‌ها یاد می‌شود. از جمله این سازمان‌ها می‌توان به مخابرات، بانک، خدمات‌دهنده‌های تلویزیون‌های کابلی و بیمه‌ها اشاره کرد. مشتریان، چرخ‌های گرداننده این شرکت‌ها و سازمان‌ها هستند و حفظ آن‌ها قطعاً موجب افزایش سرعت پیشرفت و ترقی آن‌ها خواهد شد. در مقابل، از دست دادن آن‌ها بسیار ضررآفرین بوده و ممکن است در شرایط بسیار وخیم، موجب ورشکستگی این فعالان اقتصادی شود. از این‌رو در مطالعات متفاوت، از مدل‌های آماری و داده‌کاوی مختلفی به منظور شناسایی مشتریانی که ریزش آن‌ها محتمل است، استفاده شده است. از جمله مهم‌ترین این مدل‌ها می‌توان به رگرسیون لجستیک[1]، شبکه عصبی مصنوعی[2]، درخت تصمیم[3] و نظریه مجموعه مبهم[4] اشاره کرد. در این پایان‌نامه، قصد داریم مدل‌های مذکور را معرفی کرده و دقت هر یک را در یک مطالعه موردی مورد ارزیابی و مقایسه قرار دهیم.

۱-۲ ساختار پایان‌نامه

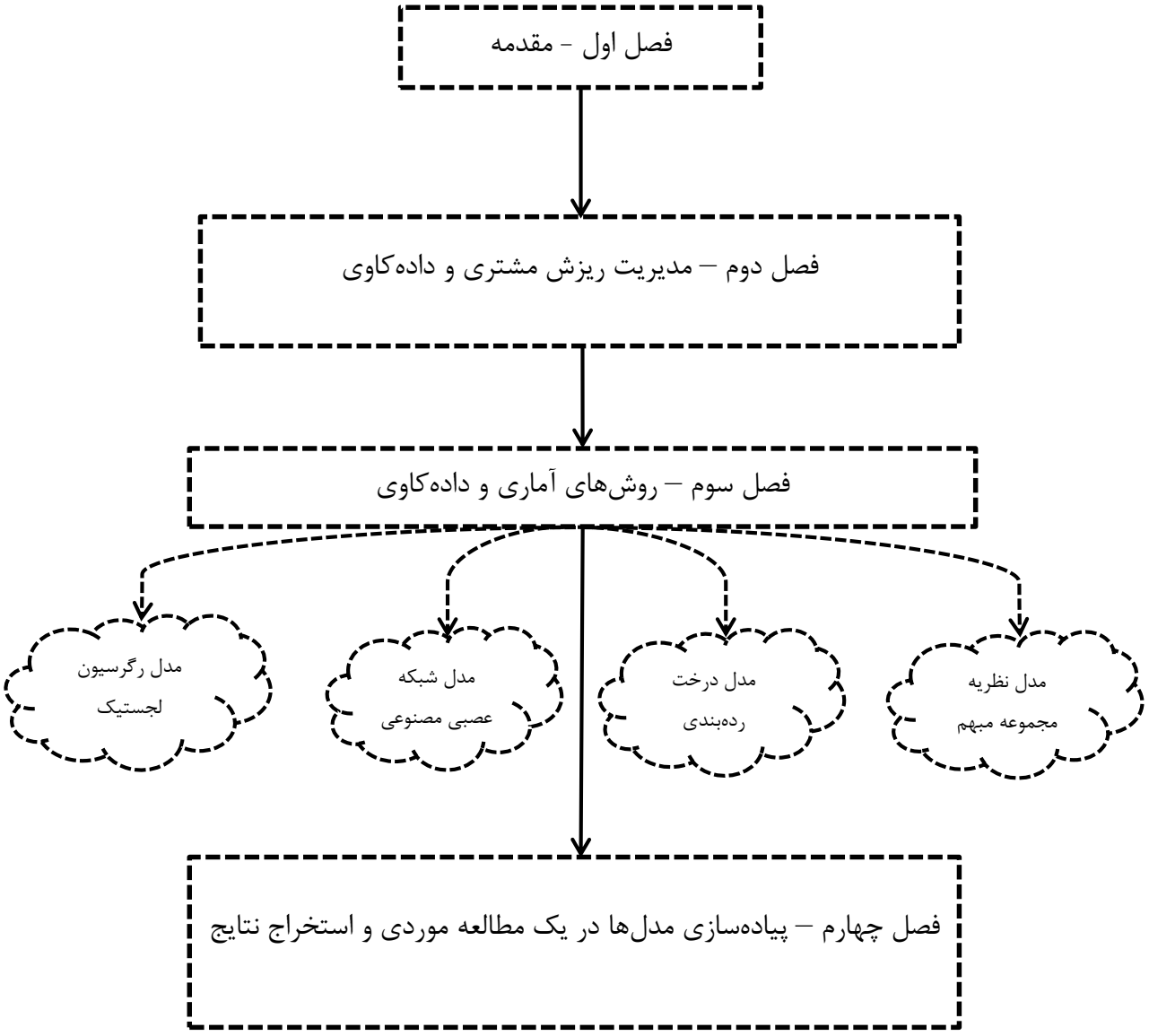
بخش‌های مختلف این پایان‌نامه مطابق ذیل تنظیم شده است:

فصل دوم اختصاص به داده‌کاوی و نحوه کاربرد آن در پیش‌بینی ریزش مشتری دارد. در فصل سوم مدل‌های نظریه مجموعه مبهم، درخت تصمیم، شبکه عصبی مصنوعی و رگرسیون لجستیک که از جمله مهم‌ترین ابزارها در پیش‌بینی ریزش مشتری هستند معرفی شده و در پایان فصل، معیارهای ارزیابی مدل، ذکر می‌شوند. در فصل چهارم ابتدا به توضیح داده‌های مورد استفاده در این پایان‌نامه که مربوط به اطلاعات مشترکان تلفن همراه است، پرداخته و سپس الگوریتم اسموت[5] به منظور انجام

---

1 Logistic regression
2 Artificial Neural network
3 Decision tree
4 Rough set theory
5 Smote Algorithm

عملیات پیش‌پردازش معرفی می‌شود و در ادامه، مدل‌های ذکر شده بر روی داده‌ها برازش داده شده و نتایج مورد ارزیابی قرار می‌گیرند. در شکل ۱-۱ ساختار شماتیکی پایان‌نامه جهت سهولت، نشان داده شده است.



شکل ۱-۱: ساختار کلی پایان‌نامه

فصل دوم

مدیریت ریزش مشتری و داده‌کاوی

مدیریت ریزش مشتری، زیرشاخه‌ای از *مدیریت ارتباط با مشتری*[1] (*CRM*) است. از این‌رو به صورت جامع به معرفی مدیریت ارتباط با مشتری می‌پردازیم.

۲-۱ مدیریت ارتباط با مشتری

برای معرفی مدیریت ارتباط با مشتری، ابتدا لازم است بدانیم مشتری کیست و چه افرادی را می‌توان مشتری نامید. مشتری، فردی است که نیاز و خواست خود را، خودش تعیین و تعریف می‌کند و بین کالاها و خدمات موجود و متنوع، حق انتخاب دارد و بابت کالاها و خدماتی که از سازمان یا شرکتی دریافت می‌کند، هزینه‌ای را پرداخت می‌کند. مدیریت ارتباط با مشتری یکی از مهم‌ترین روش‌های تجارت است که تعاریف مختلفی برای آن وجود دارد. لینگ[2] و ین[3] (۲۰۰۱)، مدیریت ارتباط با مشتری را مجموعه‌ای از فرآیندها و سیستم‌های پویا نامیدند که از یک استراتژی به منظور ایجاد و حفظ رابطه‌ای طولانی با مشتریان خاص حمایت می‌کند. کینکاید[4] (۲۰۰۳)، مدیریت ارتباط با مشتری را استفاده استراتژیک از اطلاعات، فرآیندها، تکنولوژی و مردم، برای مدیریت رابطه مشتریان با موسسه تعریف می‌کند. پارویتیار[5] و شت[6] (۲۰۰۱)، مدیریت ریزش مشتری را یک استراتژی جامع و یک فرآیند متشکل از کشف، نگهداری و همکاری با مشتریان انتخابی به منظور افزایش سود برای طرفین می‌دانند. بنابر تعاریف ذکر شده می‌توان مدیریت ارتباط با مشتری را یک راهبرد تجاری نامید که هدفش شناسایی نیازها و خواسته‌های مشتریان و پاسخ‌گویی به این نیازها و خواسته‌ها در جهت اهداف اساسی سازمان است.

از لحاظ ساختاری، چارچوب مدیریت ارتباط با مشتری را می‌توان به دو شاخه عملیاتی و تحلیلی

---

1 Customer Relationship Management
2 Ling
3 Yen
4 Kincaid
5 Parvatiyar
6 Sheth

طبقه‌بندی کرد (برسون[1] و همکاران، ۲۰۰۰). شاخه عملیاتی مدیریت ارتباط با مشتری، شامل خودکارسازی بخش‌های فروش، بازاریابی، پشتیبانی و خدمات مشتریان می‌باشد. شاخه تحلیلی مدیریت ارتباط با مشتری به تحلیل خصوصیات و رفتار مشتریان برای محافظت از استراتژی‌های مدیریتی اشاره دارد. این تجزیه و تحلیل بر روی اطلاعات حاصله از مشتریان در شاخه عملیاتی مدیریت ارتباط با مشتری انجام می‌شود. به عبارت دیگر شاخه تحلیلی مدیریت ارتباط با مشتری، بر انبار داده[2] سازمان‌ها و کشف دانش از آن‌ها تکیه دارد.

به طور کلی مدیریت ارتباط با مشتری را به چهار مرحله کلی تقسیم می‌کنند که در اصطلاح چرخه سیستم مدیریت مشتری نامیده می‌شود (ای‌یو[3] و همکاران، ۲۰۰۳). این چهار مرحله عبارتند از:

۱. شناسایی مشتریان

چرخه مدیریت ارتباط با مشتری، با شناسایی مشتریان آغار می‌شود. در این مرحله، افرادی که احتمال می‌رود که مشتری شرکت یا سازمان شده و برای آن‌ها سودآور باشند، شناسایی می‌شوند. این مرحله، شامل تحلیل مشتریانی که شرکت را ترک کرده و به سمت شرکت‌های رقیب سوق پیدا کرده‌اند نیز می‌باشد. دو عنصر مهم در شناسایی مشتری، تحلیل مشتری هدف[4] و تقسیم‌بندی مشتری[5] است. تحلیل مشتری، با تکیه بر خصوصیات مشتریان، به دنبال دسته‌های سودآوری از آن‌ها می‌گردد، در حالی که در تقسیم‌بندی مشتریان، مجموعه کامل مشتریان به زیرمجموعه‌های مجزا از هم به گونه‌ای تقسیم‌بندی می‌شوند که مشتریان درون یک مجموعه دارای بیشترین شباهت و هم‌گونی باشند (وو[6] و همکاران، ۲۰۰۵).

---

1 Berson
2 Database
3 Au
4 Target customer analysis
5 Customer segmention
6 Woo

۲. جذب مشتری

مرحله دوم مدیریت ارتباط با مشتری، دنباله‌رو مرحله اول است. بعد از شناسایی دسته مشتریان بالقوه، سازمان‌ها می‌توانند برای جذب آن‌ها تلاش کرده و آن‌ها را مورد هدف قرار دهند. یک عنصر مهم در جذب مشتری، بازاریابی مستقیم[۱] است که طی آن، مشتریان به سفارش دادن محصول تحریک می‌شوند (چئونگ[۲] و همکاران، ۲۰۰۳). از جمله فعالیت‌هایی که به منظور بازاریابی مستقیم انجام می‌شوند، می‌توان به توزیع سهمیه و ارسال مستقیم محصول به مشتریان اشاره کرد.

۳. نگهداری مشتری

برای شناسایی و جذب مشتری، قطعاً هزینه‌هایی صورت گرفته که با ترک مشتری این هزینه به هدر می‌رود. به همین دلیل، نگهداری مشتریان جذب‌شده یکی از مهم‌ترین مسائل نگران کننده برای شرکت‌ها است. رضایت‌مندی مشتری، مهم‌ترین شرط در نگهداری مشتری است که از طریق برآورده شدن انتظاراتش صورت می‌گیرد (کراکلائور[۳] و همکاران، ۲۰۰۴). از جمله مهم‌ترین عناصر نگهداری مشتری، بازاریابی یک به یک[۴]، برنامه‌های وفاداری[۵] و مدیریت شکایات[۶] هستند. بازاریابی یک به یک، به شخصی‌سازی بازاریابی می‌پردازد که از طریق تجزیه و تحلیل، کشف و پیش‌بینی رفتار مشتریان صورت می‌پذیرد (چن[۷] و همکاران، ۲۰۰۵؛ جیانگ[۸] و توژیلین[۹]، ۲۰۰۶). برنامه‌های وفاداری شامل فعالیت‌های پشتیبانی است که به واسطه آن یک رابطه طولانی با مشتری شکل می‌گیرد. از جمله فعالیت‌هایی ویژه‌ای که در زمینه برنامه‌های وفاداری صورت می‌گیرد، *تحلیل ریزش*، امتیازبندی اعتبار و کیفیت خدمات برای جلب رضایت است.

---

1 Direct Marketing
2 Cheung
3 Kracklauer
4 One to One Marketing
5 Loyalty Programs
6 Complaints Management
7 Chen
8 Jiang
9 Tuzhilin

۴. توسعه روابط با مشتری

بعد از مرحله حفظ مشتری، به منظور افزایش میزان سودآوری، مدیران شرکت‌ها در پی قوی‌تر شدن و بیشتر شدن رابطه خود با مشتریان، به‌ویژه مشتریان سودآور هستند که این مهم از طریق افزایش تعداد معامله‌ها و یا مقدار معامله‌ها صورت می‌پذیرد. تحلیل چرخه ارزش مشتری[۱]، باهم فروشی[۲] و تحلیل سبد خرید[۳]، ازجمله مهم‌ترین عناصر در توسعه روابط با مشتری هستند. تحلیل چرخه ارزش مشتری، عبارت است از پیش‌بینی حداکثر درآمدی که یک موسسه می‌تواند از یک مشتری دریافت کند (درو[۴] و همکاران، ۲۰۰۱؛ اتزیون[۵] و همکاران، ۲۰۰۵). باهم فروشی، عبارت است از مجموعه فعالیت‌های تبلیغاتی که یک مشتری را تشویق می‌کند تا به همراه خرید یک محصول، محصولات دیگری که با آن محصول در ارتباط هستند را نیز خریداری کند (پرینزی[۶] و همکاران، ۲۰۰۶). تحلیل سبد خرید نیز به مجموعه فعالیت‌هایی از قبیل نحوه چینش محصولات در قفسه‌های فروش اطلاق می‌شود که از طریق اطلاعات مربوط به رفتار خرید مشتریان پیش‌بینی می‌شود. هدف از تحلیل سبد خرید، بیشینه‌سازی مقدار و تعداد معامله‌های مشتری است.

موضوع این پایان‌نامه مرتبط با مرحله حفظ مشتری است که تحت عنوان "مدیریت ریزش مشتری" در زیر به طور خلاصه توضیح داده می‌شود.

۲-۱-۱ ریزش مشتری

ریزش مشتری که ترک مشتری یا رویگردانی مشتری نیز نامیده می‌شود، اصطلاحی تجاری است که در مواقع ترک مشتری از یک موسسه تجاری، مورد استفاده قرار می‌گیرد. به طور کلی مشتریان ریزش شده را بر اساس علت ریزش به دو دسته تقسیم می‌کنند که عبارتند از:

---

1 Customer Lifetime Value
2 Up/Cross Selling
3 Market Basket Analysis
4 Drew
5 Etzion
6 Prinzie

۱. داوطلبانه:

الف) ریزش تصادفی: زمانی رخ می‌دهد که به دلایل غیرعمدی از جمله تغییر محل زندگی و بروز مشکلات مالی، مشتری دیگر قادر به خدمات‌گیری از خدمات‌دهنده مورد نظر نباشد.

ب) ریزش عمدی: زمانی رخ می‌دهد که مشتری به علت خدمات بهتر، قیمت پایین‌تر و کیفیت بهتر محصولات رقیب، خدمات‌گیری از خدمات‌دهنده خود را ترک کرده و به خدمات‌دهنده مناسب‌تر رجوع می‌کند.

۲. غیرداوطلبانه: در این نوع ریزش، خدمات‌دهنده به دلایل بدحسابی مشتری، سوء استفاده مشتری از خدمات شرکت و از این قبیل مسائل، مشتری مورد نظر را شناسایی کرده و خدمات‌دهی به او را قطع می‌کند.

در سال‌های اخیر، رشد سریع تکنولوژی و صنعت ارتباطات، موجب افزایش تعداد شرکت‌های ارائه دهنده خدمات مشابه شده و افزایش آگاهی مشتریان باعث شده است تا مشتریان قدرت انتخاب بیشتری داشته و به راحتی خدمات‌گیرنده مورد نظر خود را انتخاب کنند. یان[1] و همکاران (۲۰۰۴) نرخ سالیانه ریزش را در صنعت مخابرات ۲۵ تا ۳۰ درصد بیان کردند. ای‌یو و همکاران (۲۰۰۳) نیز میزان مشتریان ریزش شده در خدمات‌دهنده‌های اینترنتی را ۵۰ درصد تخمین زدند. از طرف دیگر طبق مطالعات انجام شده، هزینه جذب یک مشتری جدید از هزینه نگهداری مشتریان فعلی بسیار بیشتر خواهد بود. مینگائل[2] (۲۰۰۱)، هزینه ریزش مشتریان را برای شرکت‌های مخابراتی اروپایی و آمریکایی ۴ میلیارد دلار در سال برآورد کردند. چو[3] و همکاران (۲۰۰۷)، هزینه جذب یک مشتری جدید را حدود ۵ تا ۱۰ برابر بیشتر از نگهداری مشتریان موجود برآورد کردند. بنابراین، از میان حالت-های ذکر شده برای ریزش مشتری، ترک عمدی برای مدیران شرکت‌ها بسیار حائز اهمیت است. اگر

[1] Yan
[2] Minguel
[3] Chu

مدیران شرکت‌ها قادر به شناسایی مشتریان دارای احتمال ریزش قبل از ترک شرکت باشند، ممکن است بتوانند با اعمال بعضی استراتژی‌های بازدارنده، از قبیل بسته‌های تشویقی از ریزش آن‌ها جلوگیری کنند. مدیریت ریزش مشتری لزوماً بر روی تمامی مشتریان با ریزش محتمل اعمال نمی-شود و برای موسسه‌ها، جلوگیری از ریزش مشتریان سودآور در اولویت بالاتری قرار دارد. علاوه بر این، سازمان‌ها و شرکت‌هایی از قبیل بانک‌ها، شرکت‌های مخابراتی، شرکت‌های تلویزیون کابلی، شرکت‌های بیمه و غیره اغلب از نرخ ریزش مشتریان به عنوان یکی از معیارهای کلیدی سنجش در کسب و کار استفاده می‌کنند. جلوگیری از ریزش مشتریانی که ریزش آن‌ها محتمل است، باعث کوچکتر شدن نرخ ریزش مشتری شده و موجب افزایش اعتبار شرکت خواهد شد. به عنوان مثال، نرخ ریزش ماهانه مشتریان به صورت زیر محاسبه می‌شود:

$$\text{نرخ ریزش ماهانه} = \frac{C_0+A-C_1}{C_0}$$

که در آن C_0 تعداد مشتریان در اول ماه، C_1 تعداد مشتریان در آخر ماه و A تعداد مشتریان جدید در طول ماه است. با کاهش تعداد افراد ریزش‌کننده، مقدار C_1 افزایش یافته و در نهایت نرخ ریزش ماهانه کوچکتر می‌شود.

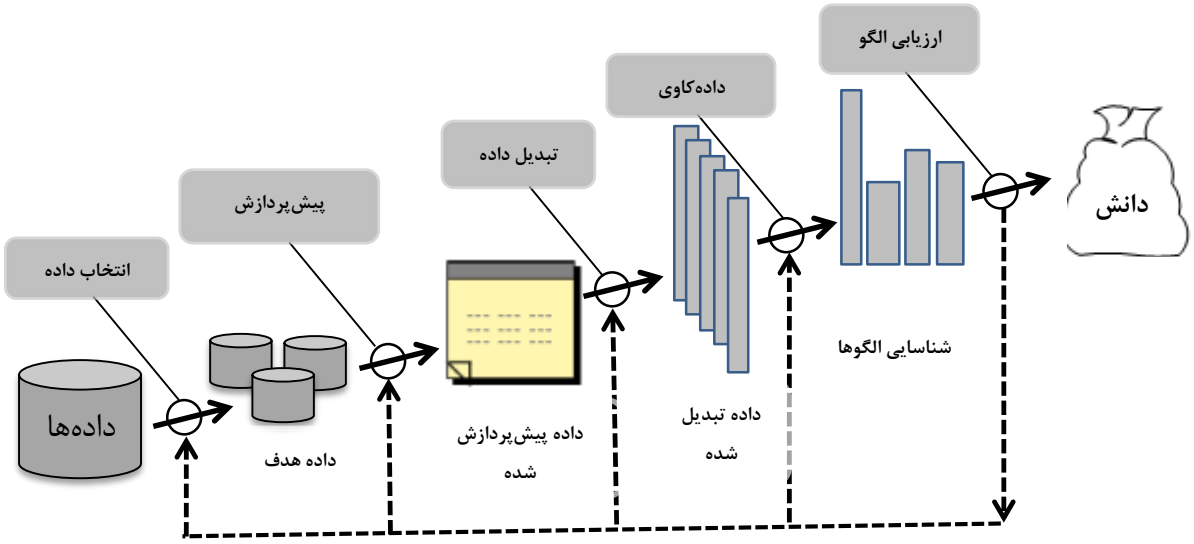
مروری بر مطالعات قبلی طی سال‌های ۲۰۰۰ تا ۲۰۰۶، نشان داده است که حدود ۶۲ درصد از مقاله‌های مرتبط با موضوع مدیریت ارتباط با مشتری، در زمینه نگهداری مشتری بوده است. بنابراین نگهداری مشتری در سال‌های اخیر بیشتر مورد توجه موسسه‌ها قرار گرفته و محققان را بر آن داشته تا راه‌کارهایی را برای حفظ مشتریان ارائه دهند.

همان‌طور که بیان شد، در شاخه تحلیلی مدیریت ارتباط با مشتری، هدف، کسب دانش از انبار داده مشتریان است. در مدیریت ریزش مشتری، دانش مورد نظر برای شناسایی مشتریانی به کار می-رود که خصوصیات ریزش در آن‌ها وجود دارد. مطالعات نشان داده‌اند که ابزارهای آماری و داده‌کاوی به منظور کشف دانش از پایگاه داده‌ی مشتریان، بسیار مفید عمل کرده‌اند و مقاله‌ها و کتاب‌های

متعددی در این زمینه به چاپ رسیده‌اند. در بخش بعدی به معرفی داده‌کاوی و نحوه کاربرد آن در استخراج دانش از پایگاه داده مشتریان می‌پردازیم.

۲-۲ کشف دانش و نقش داده‌کاوی در آن

کشف دانش، فرآیندی از شناسایی الگوهای معتبر، جدید، مفید و قابل فهم از پایگاه داده است. برای کشف دانش از پایگاه داده، چند مرحله به صورت شکل ۲-۱ باید دنبال شوند.



شکل ۲-۱: فرآیند کشف دانش

هر یک از این مراحل به صورت زیر تعریف می‌شوند.

۱. **انتخاب داده‌ها:** در این مرحله، داده‌های مرتبط با فرآیند داده‌کاوی از سایر داده‌ها جدا می‌شود. این مرحله را می‌توان بخشی از فرآیند کاهش داده‌ها نیز دانست.

۲. **پیش‌پردازش داده‌ها:** این مرحله شامل پاک‌سازی و یکپارچه‌سازی داده‌ها می‌باشد.

پاک‌سازی داده‌ها: در این مرحله داده‌های نامعتبر از مجموعه داده‌ها حذف می‌شوند. داده‌های دارای اغتشاش، داده‌های ناقص و غیره، نمونه‌هایی از داده‌هایی هستند که باید پاک‌سازی در مورد آن‌ها انجام گردد.

یکپارچه‌سازی داده‌ها: در این مرحله، منابع چندگانه داده‌ای با هم ترکیب می‌شوند.

۵. تبدیل داده‌ها: داده‌ها به قالبی قابل استفاده برای داده‌کاوی در می‌آیند. از اعمالی که در این مرحله صورت می‌گیرند می‌توان به خلاصه‌سازی یا تجمیع داده‌ها اشاره کرد.

۶. داده‌کاوی: مهم‌ترین مرحله از مراحل کشف دانش، مرحله‌ی داده‌کاوی است. در این مرحله پس از آماده‌سازی داده‌ها، بسته به نوع دانش مورد نظر، روش‌های مختلف داده‌کاوی بر روی داده‌ها پیاده شده و الگوی مورد نظر استخراج می‌شود.

۷. ارزیابی الگوها: تشخیص الگوهای صحیح مورد نظر، از سایر الگوها در این مرحله انجام می‌شود. درستی الگوها بر اساس معیارهای ارزیابی[۱] سنجیده می‌شود.

۲-۲-۱ داده‌کاوی و نحوه کاربرد آن در مدیریت ریزش مشتری

داده‌کاوی پل ارتباطی میان علم آمار، کامپیوتر، هوش مصنوعی[۲]، الگوشناسی[۳]، یادگیری ماشینی[۴] و مجسم‌سازی[۵] داده‌ها است که فرآیندهایی را جهت شناسایی الگوها و مدل‌های مفید، درحجم وسیعی از داده‌ها به کار می‌گیرد. به عبارت دیگر داده‌کاوی فرآیندی است که با استفاده از روش‌های هوشمند، دانش را از مجموعه‌ای از داده‌ها استخراج می‌کند. این روش، امروزه کاربردهای بسیار وسیعی در حوزه‌های مختلف از جمله صنعت ارتباطات، بیمه، بانک‌داری، علوم زیستی و پزشکی پیدا کرده است. سیستم‌های داده‌کاوی تقریباً از اوایل دهه ۹۰ میلادی، مورد توجه قرار گرفتند. علت این امر نیز آن بود که تا آن زمان، سازمان‌ها بیشتر در پی ایجاد سیستم‌های عملیاتی کامپیوتری بودند که به وسیله آن‌ها بتوانند داده‌های موجود را سازماندهی کنند. پس از ایجاد این سیستم‌ها، روزانه حجم زیادی از اطلاعات جمع‌آوری می‌شد که تحلیل آن‌ها از عهده انسان خارج بود. بررسی محققان نشان داده است

---

1 Evaluation measure
2 Artificial intelligence
3 Pattern recognization
4 Machine learning
5 Visualization

که حجم داده‌ها در جهان، در هر ۲۰ ماه حدوداً دو برابر می‌شود. در یک تحقیق که بر روی گروه‌های تجاری بسیار بزرگ، صورت گرفت مشخص گردید که ۱۹ درصد از این گروه‌ها دارای پایگاه داده‌هایی با حجم بیش از ۵۰ گیگا بایت می‌باشند و ۵۹ درصد از آن‌ها نیز انتظار می‌رود در آینده‌ای نزدیک در چنین سطحی از اطلاعات قرار گیرند. به همین دلیل، نیاز به روش‌هایی بود که از میان انبوه داده‌ها، اطلاعات مفید یا دانش استخراج شود. بدین ترتیب داده‌کاوی در مسیر ایجاد و رشد قرار گرفت. درسال ۱۹۸۹ و ۱۹۹۱، کارگاه‌های کشف دانش از پایگاه‌های داده توسط پیاتتسکی[1] و همکاران برگزار گردید. اصطلاح داده‌کاوی برای اولین بار توسط فیاد[2] و همکاران در اولین کنفرانس بین المللی " کشف دانش و داده‌کاوی " درسال ۱۹۹۵ مطرح شد. از سال ۱۹۹۵، داده‌کاوی به صورت جدی وارد مباحث آمار شد و در سال ۱۹۹۶ اولین مجله " کشف دانش از پایگاه داده‌ها" منتشر شد. در حال حاضر، داده‌کاوی مهم‌ترین فن‌آوری جهت بهره‌برداری موثر از داده‌های حجیم بوده و اهمیت آن رو به فزونی است.

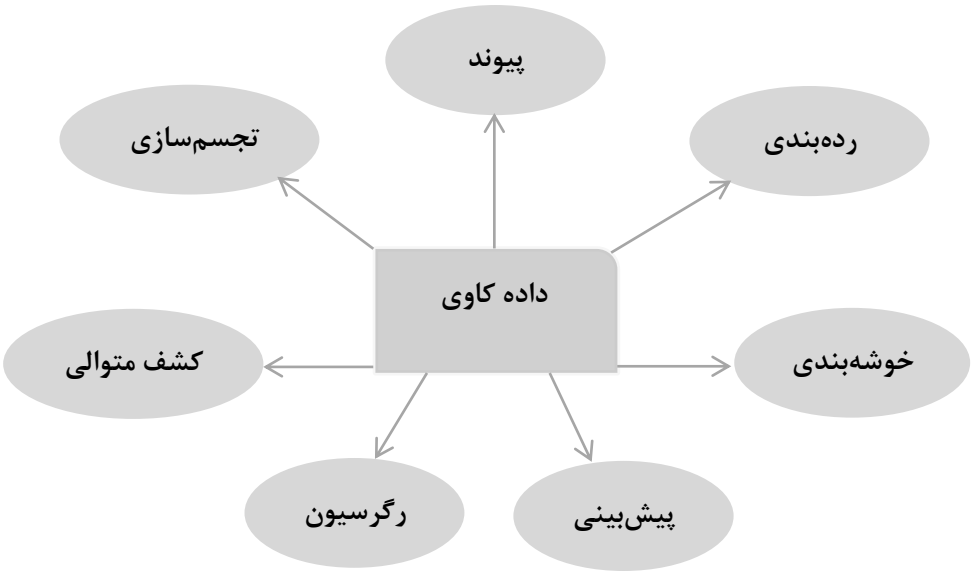
از منظرهای مختلفی می‌توان به موضوع الگویابی در داده‌ها نظاره کرد که هر یک از آن‌ها دارای مبانی و روش‌های خاصی است. شکل ۲-۲ شامل فهرستی از عناوین کلی مربوطه است که در زیر به اختصار شرح داده شده‌اند.

پیوند[3]:

هدف، یافتن ارتباط میان مجموعه‌ای از متغیرها است. این ارتباط معمولاً به صورت قوانین "اگر...آنگاه" از مجموعه داده‌ها استخراج می‌شود. به عنوان مثال می‌توان به شناسایی ارتباط میان اقلام یک فروشگاه بر اساس اطلاعات موجود در فاکتورهای فروش اشاره کرد. برنامه‌های تحلیل سبد

[1] Piatetsky
[2] Fayyad
[3] Association

بازار و باهم فروشی از جمله مثال‌هایی هستند که مدل‌های پیوند برای آن‌ها استفاده می‌شود (احمد[1]، ۲۰۰۴؛ جیائو[2] و همکاران، ۲۰۰۶).



شکل ۲-۲: دسته‌بندی تکنیک‌های داده‌کاوی

رده‌بندی[3]:

روش‌های رده‌بندی، با استفاده از اطلاعات موجود در مشاهدات، هر یک از آن‌ها را در رده‌های ازپیش تعیین‌شده‌ای قرار می‌دهند (احمد، ۲۰۰۴). تحلیل ریزش مشتری از جمله مواردی است که تکنیک-های رده‌بندی در آن کاربرد دارد.

خوشه‌بندی[4]:

روش‌های خوشه‌بندی، مشاهداتی که دارای بیشترین شباهت هستند شناسایی کرده و درون یک خوشه قرار می‌دهند (احمد، ۲۰۰۴؛ کرییر[5] و همکاران، ۲۰۰۳). تفاوت رده‌بندی و خوشه‌بندی در این است که در رده‌بندی، فرض می‌شود متغیر پاسخ دارای چند سطح بوده و هدف آن است که سطح

[1] Ahmed
[2] Jiao
[3] Classification
[4] Clustring
[5] Carrier

مربوط به مشاهدات، بر اساس متغیرهای توضیحی، پیش‌بینی شود، اما در خوشه‌بندی وجود متغیر پاسخ موضوعیت نداشته و هدف، تشخیص مشاهداتی است که به گونه‌ای، مشابه هستند.

پیش‌بینی[1]:

در پیش‌بینی، مقدار یک متغیر بر اساس یک الگوی از پیش تعریف‌شده‌ای، پیش‌بینی می‌شود. به عنوان مثال می‌توان به سری‌های زمانی اشاره کرد که در آن رفتار یک متغیر در طول زمان بررسی شده و رفتار آینده آن پیش‌بینی می‌شود.

رگرسیون[2]:

یک روش آماری برای شناسایی رابطه‌ی بین مجموعه متغیرهای توضیحی و متغیر پاسخ است. در این روش، برای شناسایی این رابطه، از تصویر متغیر پاسخ در فضای ایجادشده توسط مجموعه متغیرهای توضیحی استفاده می‌شود.

کشف متوالی[3]:

برسون[4] و همکاران (۲۰۰۰)، کشف متوالی را شناسایی روابط و الگوها در طول زمان تعریف کرده است. هدف از کشف متوالی، مدل‌سازی یک فرآیند در طول زمان به منظور شناسایی انحراف و روند در آن است.

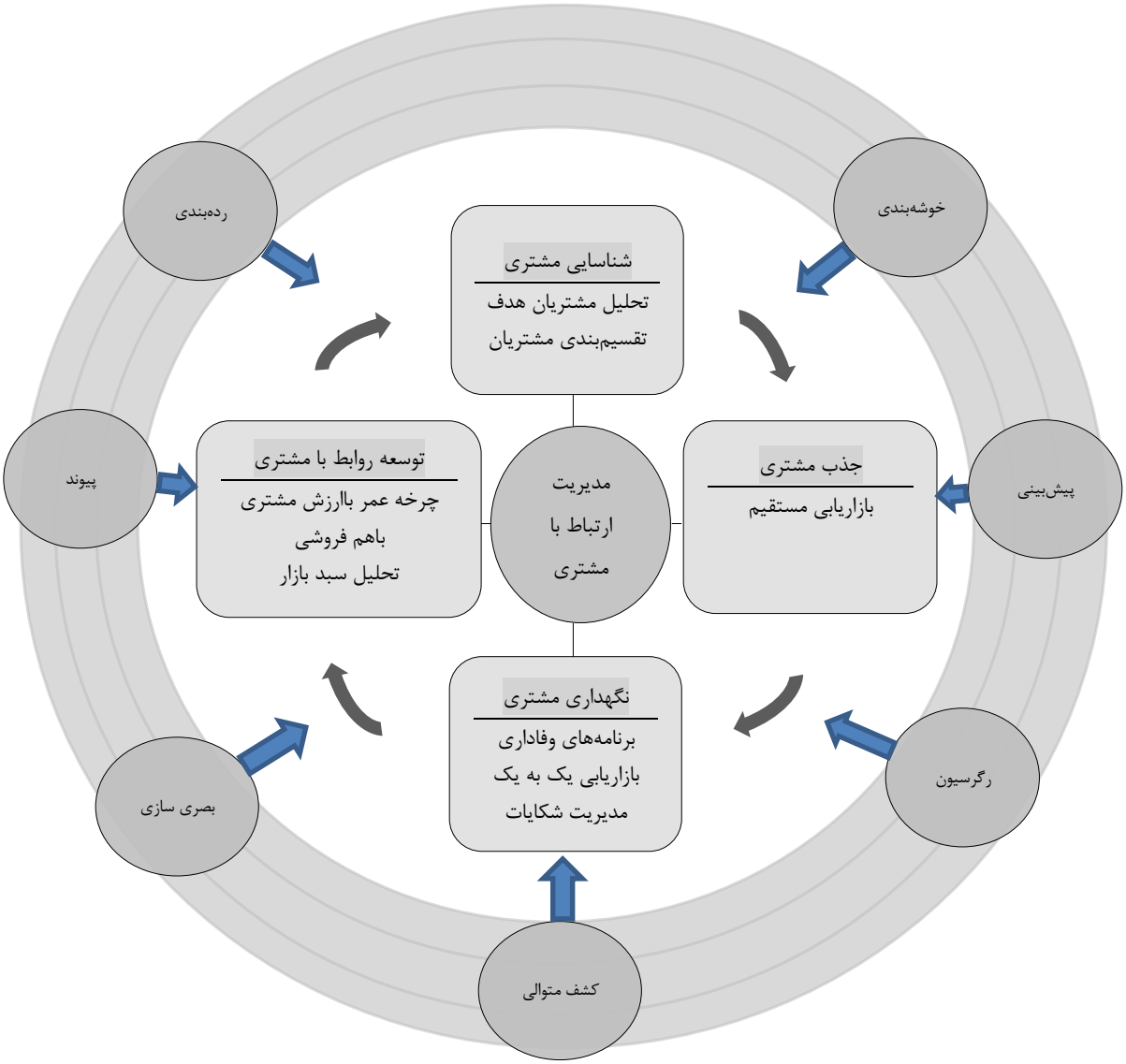
تجسم‌سازی[5]:

تجسم‌سازی، نمایش داده‌ها به نحوی است که کاربر می‌تواند الگوها و روابط پیچیده در داده‌ها را مشاهده کند (شاو[1] و همکاران، ۲۰۰۱). روش‌های بصری‌سازی به موازات روش‌های دیگر داده‌کاوی برای درک الگوها و روابط پیچیده درون داده‌ها مورد استفاده قرار می‌گیرد.

---

1 Forecasting
2 Regression
3 Sequence discovery
4 Berson
5 Visualization

هر یک از موارد فوق به نحوی در مدیریت ارتباط با مشتری مورد استفاده قرار می‌گیرند. این ارتباط را پارویتیار و همکاران (۲۰۰۱)، به صورت آنچه که در شکل ۲-۳ آمده است در نظر گرفتند.



شکل ۲-۳: نحوه کاربرد داده‌کاوی در مدیریت ریزش مشتری

به عنوان مثال، منظور از استفاده روش‌های خوشه‌بندی در مرحله شناسایی مشتری این است که کارشناسان امر با استفاده از این روش‌ها، مشتریان را بر اساس اطلاعاتشان خوشه‌بندی کرده و آن خوشه‌هایی که از نظر آنها بهترین است شناسایی می‌کنند و عملیات جذب مشتری را روی افراد درون

[1] Shaw

آن‌ها انجام می‌دهند.

برای پیش‌بینی ریزش مشتری، روش‌های رده‌بندی داده‌کاوی به منظور استخراج دانش مورد نظر، مناسب می‌باشند. در این پایان‌نامه قصد داریم با استفاده از چند ابزار رده‌بندی، مشتریان با ریزش محتمل را شناسایی کرده، دقت مدل‌ها را با هم مقایسه کنیم و در نهایت بهترین مدل را معرفی کنیم. بدین منظور ابتدا مروری بر مطالعات قبلی خواهیم داشت.

۲-۳ مروری بر مطالعات گذشته

انگای و همکاران (۲۰۰۹)، ۸۷ مقاله در زمینه مدیریت ارتباط با مشتری و کاربرد داده‌کاوی در آن را که طی سال‌های ۲۰۰۰ تا ۲۰۰۶ به چاپ رسیده بودند را مورد مطالعه قرار داده و آن‌ها را بر اساس چهار مرحله اصلی مدیریت ارتباط با مشتری طبقه‌بندی کردند. نتایج به دست آمده در جدول ۲-۱ خلاصه شده است.

جدول ۲-۱: نتایج حاصل از طبقه‌بندی مقاله‌های منتشر شده در زمینه مدیریت ارتباط با مشتری و کاربرد داده‌کاوی در آن، بر اساس چهار مرحله مدیریت ارتباط با مشتری

مراحل مدیریت ارتباط با مشتری	عنصر یا راه‌کار مورد استفاده	تعداد مقالات	تعداد کل مقالات
شناسایی مشتری	تحلیل مشتریان هدف	۵	۱۳
	تقسیم‌بندی مشتریان	۸	
جذب مشتری	بازاریابی مستقیم	۷	۷
نگهداری مشتری	برنامه‌های وفاداری	۲۴	۵۴
	بازاریابی یک به یک	۲۸	
	مدیریت شکایات	۲	
توسعه روابط با مشتری	چرخه عمر باارزش مشتری	۵	۱۳
	باهم فروشی	۲	
	تحلیل سبد بازار	۶	

با توجه به جدول ۲-۱، طی سال‌های ۲۰۰۰ تا ۲۰۰۶ بیشترین مقاله‌ها در زمینه " نگهداری مشتری" به چاپ رسیده است. همچنین در بین زمینه‌های مختلف نگهداری مشتری، عناصر "برنامه-های وفاداری" و " بازاریابی یک به یک" بیشترین مقالات را به خود اختصاص داده‌اند. قابل ذکر است که "مدیریت ریزش مشتری" جزء برنامه‌های وفاداری بوده و به همین دلیل موضوع مورد توجه در این پایان‌نامه است.

حال که اهمیت مطالعه در زمینه مدیریت ریزش مشتری مورد تأیید قرار گرفت، در ادامه به مرور مقاله‌هایی که در این زمینه به چاپ رسیده‌اند می‌پردازیم تا ببینیم برای مدیریت ریزش مشتریان از چه ابزارهای آماری و داده‌کاوی استفاده شده است. بدین منظور خاکابی و همکاران در سال ۲۰۱۰، ۷۷ مقاله را که طی سال‌های ۲۰۰۳ تا ۲۰۰۹ در کنفرانس‌های بین‌المللی در زمینه مدیریت ریزش مشتری به چاپ رسیده بودند، مورد مطالعه قرار داده و نتایج آن را به تفکیک نوع روش‌های مورد استفاده، و سال انتشار مقاله استخراج کردند که خلاصه آن در جدول ۲-۲ نمایش داده شده است. با توجه به سه سطر اول این جدول مشاهده می‌شود که شبکه عصبی مصنوعی، درخت تصمیم و رگرسیون لجستیک به‌ترتیب سه مدلی هستند که در مقالات مختلف از آن‌ها بیشترین استفاده شده است. از طرفی با توجه به مطالعات انجام شده نیز دریافتیم که اخیراً لین[1] و همکاران (۲۰۱۱)، با استفاده از مدل نظریه مجموعه مبهم توانستند با دقت بیش از ۹۰ درصد پیش‌بینی ریزش مشتری را در کارت‌های اعتباری انجام دهند.

در این پایان‌نامه، سه مدل شبکه‌ی عصبی مصنوعی، درخت تصمیم و رگرسیون لجستیک که به دفعات در پیش‌بینی ریزش مشتری مورد استفاده قرار گرفته‌اند را به همراه نظریه مجموعه مبهم که مدلی مناسب برای پیش‌بینی ریزش مشتری است، مورد مطالعه قرار داده و میزان دقت این مدل‌ها را مورد بررسی قرار دهیم.

[1] Lin

جدول ۲-۲: دفعات استفاده از مدل‌های آماری و داده‌کاوی در مقاله‌ها

		سال چاپ							تعداد کل
		۲۰۰۳	۲۰۰۴	۲۰۰۵	۲۰۰۶	۲۰۰۷	۲۰۰۸	۲۰۰۹	
مدل‌های آماری و داده‌کاوی	*Neural Networks*		۲	۳	۳	۲	۲	۳	۱۵
	Decision Tree			۲	۱	۲	۳	۵	۱۳
	Logistic Regression			۲	۲	۱	۳	۵	۱۳
	Random Forests			۲			۲	۳	۷
	Support Vector Machine						۳	۴	۷
	Survival Analysis		۱			۱	۱		۳
	Bayesian Network			۱			۲		۳
	Self Organizing Maps		۱				۱		۲
	AdaCost							۱	۱
	Gradient Boosting Machine							۱	۱
	Linear Discriminant Analysis						۱		۱
	AdaBoost						۱		۱
	Rough Set Theory						۱		۱
	K-Nearest Neighbor					۱			۱
	K-Means				۱				۱
	Tailor-Butina				۱				۱
	Time Series				۱				۱
	ROCK			۱					۱
	Regression Forests			۱					۱
	Linear Regression			۱					۱
	Association Rules		۱						۱
	Sequence Discovery	۱							۱

فصل سوم

مدل‌های آماری و داده‌کاوی

در این فصل به معرفی نظریه مجموعه مبهم، درخت تصمیم، شبکه عصبی مصنوعی و رگرسیون لجستیک که ازجمله متداول‌ترین روش‌های آماری و داده‌کاوی در پیش‌بینی ریزش مشتری هستند، می‌پردازیم. برای تمامی مدل‌ها، مجموعه متغیرهای توضیحی را با $\{X_1, X_2, \dots, X_N\}$، متغیر پاسخ را با Y، مقدار مشاهده شده متغیرهای توضیحی را با مجموعه $\{x_1, x_2, \dots, x_N\}$ و مقدار مشاهده متغیر پاسخ را با y نشان می‌دهیم. لازم به ذکر است که مجموعه متغیرهای توضیحی می‌توانند از نوع متغیرهای طبقه‌ای، فاصله‌ای و پیوسته باشند و متغیر پاسخ دارای K رده است.

۳-۱ نظریه مجموعه مبهم

نظریه مجموعه مبهم در اوایل سال ۱۹۸۰ میلادی توسط پروفسور زدیسلاو پاولاک[1] معرفی شد. این روش، الگوها و قوانین نهفته در مجموعه داده‌ها را به صورت قوانین اگر...آنگاه استخراج کرده و با استفاده از این قوانین، عمل رده‌بندی را برای یک مشاهده جدید انجام می‌دهد. همچنین این نظریه یک ابزار قدرتمند برای استدلال در موارد ابهام و نایقینی است و روش‌هایی را برای زدودن و کاستن اطلاعات و دانش نامربوط یا مازاد بر نیاز از پایگاه‌های داده ارائه می‌دهد (پاولاک، ۱۹۸۲). در نهایت، نظریه مجموعه مبهم با کاهش اطلاعات زائد، مجموعه‌ای از قوائد تلخیص شده پرمعنا را از داده‌ها استخراج می‌کند که کار تصمیم‌گیرنده را بسیار ساده می‌کنند. لذا با توجه به رشد سریع حجم اطلاعات، نظریه مجموعه مبهم نقش بسیار مؤثری را در سیستم‌های پشتیبان تصمیم‌گیری ایفا می‌کند (زیارکو[2]، ۱۹۹۳).

سرفصل مطالبی که برای معرفی نظریه مجموعه مبهم باید به آن‌ها پرداخته شود، به قرار زیر است:

۱. سیستم‌های اطلاعاتی و سیستم‌های تصمیم
۲. الگوریتم‌های تصمیم‌گیری

---

[1] Zdzislaw Pawlak
[2] Ziarko

۳. دقت و کیفیت رده‌بندی الگوریتم تصمیم‌گیری
   - ❖ روابط شباهت
   - ❖ تقریب مجموعه‌ها
   - ❖ دقت و کیفیت تقریب
   - ❖ دقت و کیفیت رده‌بندی
۴. کاهش متغیرها و هسته
۵. رده‌بندی یک مشاهده جدید
   - ❖ معیارهای محافظ تصمیم
   - ❖ عمل تصمیم‌گیری
۶. گسسته‌سازی

۳-۱-۱ سیستم‌های اطلاعاتی و سیستم‌های تصمیم

در نظریه مجموعه مبهم، سیستم‌های اطلاعاتی به جداولی اطلاق می‌شود که در سطرهای آن افراد یا اشیاء، و در ستون‌های آن متغیرها قرار دارند و به صورت رابطه (۳-۱) نمایش داده می‌شوند.

$$S = (U , C) \qquad (۱-۳)$$

که در آن U یک مجموعه ناتهی از اشیاء یا افراد و C یک مجموعه ناتهی از متغیرهای شرطی (توضیحی) است که از نوع رده می‌باشند.

سیستم‌های اطلاعاتی به صورت $S = (U , C \cup D)$ را یک سیستم *تصمیم* گویند که در آن D $(D \neq C)$، مجموعه متغیرهای تصمیم (پاسخ) از نوع رده هستند (سورج[1]، ۲۰۰۴). در مسئله مورد بررسی این پایان‌نامه، مجموعه متغیر پاسخ دارای یک عضو به صورت $D = \{d\}$ است که دارای دو رده می‌باشد.

[1] Suraj

به ازای هر $a \in \{C \cup D\}$، مجموعه مقادیری که متغیر a اختیار می‌کند را دامنه a نامیده و با V_a نشان داده می‌دهند. همچنین مقدار متغیر a برای فرد x را با $a(x)$ نشان می‌دهند.

مثال ۳-۱

فرض کنید جدول تصمیمی به صورت جدول ۳-۱ داریم که حاوی اطلاعات مربوط به ۶ فرد است.

جدول ۳-۱: جدول تصمیم

U	Patient	Headache (H)	Muscle-pain (M)	Temperature (T)	Flu (F)
	x_1	no	yes	High	Yes
	x_2	yes	no	High	Yes
	x_3	yes	yes	very high	Yes
	x_4	no	yes	normal	No
	x_5	yes	no	High	No
	x_6	no	yes	very high	Yes

مجموعه‌های C، D و U به ترتیب مجموعه متغیرهای شرطی، متغیر پاسخ و کل افراد جامعه هستند. دامنه متغیرها عبارتند از:

$V_H = \{yes, no\}$
$V_M = \{yes, no\}$
$V_T = \{normal, high, very\ high\}$
$V_F = \{yes, no\}$

به عنوان مثال، مقدار متغیر H برای فرد x_3 برابر با yes است که آن را به صورت $H(x_3) = yes$ نشان می‌دهند.

۳-۱-۲ الگوریتم‌های تصمیم‌گیری

فرض کنید در سیستم تصمیم $S = (U, C \cup D)$، مجموعه‌های C و D به ترتیب دارای n و m عضو به صورت $C = \{c_1, \dots, c_n\}$ و $D = \{d_1, \dots, d_m\}$ باشند، آنگاه برای هر $x \in U$، مجموعه مقادیر

متغیرهای شرطی و تصمیم به صورت $d_1(x), \dots, d_m(x), c_1(x), \dots, c_n(x)$ وجود خواهند داشت به طوری که بر اساس آن‌ها قوانین تصمیم به صورت رابطه (۳-۲) حاصل می‌شود.

$$C \rightarrow_x D \qquad (۳-۲)$$

که به آن قانون تصمیم ایجاد شده توسط فرد x در S گویند (پاولاک، ۲۰۰۲). در خصوص مثال ۳-۱ با توجه به اینکه $n = 3$ و $m = 1$ است، داریم:

$C = \{c_1, c_2, c_3\},\ c_1 = H,\ c_2 = M,\ c_3 = T$

$D = \{d_1\},\quad d_1 = F$

برای فرد x_1، مقادیر متغیرهای شرطی و تصمیم عبارتند از:

$$c_1(x_1) = no,\quad c_2(x_1) = yes,\quad c_3(x_1) = high,\quad d_1(x_1) = yes$$

بنابراین قانون ایجاد شده توسط فرد x_1 عبارت است از:

$$c_1(x_1) = no\ \&\ c_2(x_1) = yes\ \&\ c_3(x_1) = high \rightarrow d_1(x_1) = yes$$

یا به عبارت دیگر:

$$(1)\ H = no\ \&\ M = yes\ \&\ T = high \rightarrow_{x_1} F = yes$$

به همین صورت برای سایر افراد جامعه (U) نیز می‌توان قوانین تصمیم را به صورت زیر استخراج کرد:

$$(2)\ H = yes\ \&\ M = no\ \&\ T = high \rightarrow_{x_2} F = yes$$

$$(3)\ H = yes\ \&\ M = yes\ \&\ T = veryhigh \rightarrow_{x_3} F = yes$$

$$(4)\ H = no\ \&\ M = yes\ \&\ T = normal \rightarrow_{x_4} F = no$$

$$(5)\ H = yes\ \&\ M = no\ \&\ T = high \rightarrow_{x_5} F = no$$

$$(6)\ H = no\ \&\ M = yes\ \&\ T = veryhigh \rightarrow_{x_6} F = yes$$

به مجموعه کلیه قوانین تصمیم استخراج‌شده از یک جدول تصمیم، در اصطلاح الگوریتم تصمیم-گیری گویند. گاهی اوقات ممکن است دو قانون به ازای زیرمجموعه‌ای از متغیرهای شرطی دقیقاً مقادیر مشابهی را دارا باشند اما مقدار متغیر پاسخ آن‌ها متفاوت باشد. مثلاً قانون‌های (۲) و (۵) دارای مقادیر متغیر شرطی یکسان، اما مقدار متغیر پاسخ متفاوت هستند.

$$(2)\ H = yes\ \&\ M = no\ \&\ T = high \rightarrow_{x_2} Flu = yes$$
$$(5)\ H = yes\ \&\ M = no\ \&\ T = high \rightarrow_{x_5} Flu = no$$

در این صورت حالت *عدم‌قطعیت*[1] در مورد این قانون‌ها پیش می‌آید. یعنی برای مشاهده جدیدی که از این قانون‌ها پیروی می‌کند، نمی‌توان بطور حتمی تصمیم گرفت که مقدار متغیر پاسخ آن چیست. به الگوریتم تصمیم‌گیری که تمامی قانون‌های آن قطعی باشد، الگوریتم تصمیم‌گیری سازگار[2]، و در غیر این صورت، ناسازگار[3] گویند.

حال در اینجا سوال‌های زیر مطرح می‌شوند:

۱. آیا می‌توان الگوریتم تصمیم‌گیری استخراج‌شده را به نحوی که دقت رده‌بندی آن کاهش نیابد، ساده‌تر و خلاصه‌تر کرد؟

۲. آیا با استفاده از الگوریتم تصمیم‌گیری استخراج‌شده، می‌توان عمل رده‌بندی را برای هر مشاهده جدیدی انجام داد؟

۳. زمانی که متغیرهای شرطی از نوع پیوسته (غیر رده) باشند، چگونه می‌توان الگوریتم تصمیم‌گیری را استخراج نمود؟

نظریه مجموعه مبهم برای هر یک از سوال‌های مطرح شده راه حلی ارائه می‌کند که در زیربخش‌های بعد به آن‌ها می‌پردازیم.

---

[1] Uncertainty
[2] Consistent
[3] Inconsistent

۳-۱-۳ دقت و کیفیت رده‌بندی الگوریتم تصمیم‌گیری

در نظریه مجموعه مبهم به منظور ساده‌سازی جدول تصمیم و یا الگوریتم تصمیم‌گیری، ابتدا میزان قدرت رده‌بندی الگوریتم تصمیم‌گیری اولیه را با استفاده از معیارهایی محاسبه می‌کنند. سپس با استفاده از روشی که در زیربخش بعدی معرفی می‌شود، الگوریتم تصمیم‌گیری را به گونه‌ای ساده می‌کنند که دقت رده‌بندی الگوریتم جدید با دقت رده‌بندی الگوریتم اولیه برابر باشد. در این زیربخش معیارهایی را جهت سنجش دقت رده‌بندی یک الگوریتم تصمیم‌گیری معرفی خواهیم کرد. بدین منظور ابتدا نیاز به ذکر بعضی تعاریف پایه است که در زیر به آن می‌پردازیم.

روابط شباهت یا هم‌ارزی

در جدول اطلاعات ممکن است خصوصیات بعضی از افراد به ازای زیرمجموعه‌ای از متغیرهای شرطی یکسان باشند که در اصطلاح به آن‌ها افراد شبیه[1] گویند. به عنوان مثال در جدول ۳-۱، افراد x_2 و x_5 به ازای تمامی متغیرهای شرطی مقادیر یکسانی را دارا هستند. روابط شباهت، روابطی هستند که بر اساس آن‌ها افراد شبیه از لحاظ زیرمجموعه‌ای از متغیرها، درون یک مجموعه قرار می‌گیرند. در سیستم اطلاعاتی $S = (U, C)$، به ازای هر $B \subseteq C$ یک رابطه شباهت به صورت رابطه (۳-۳) تعریف می‌شود.

$$IND_s(B) = \{(x, x') \in u \times u | \forall a \in B, a(x) = a(x')\} \quad (۳-۳)$$

که به آن در اصطلاح رابطه B- شبیه گویند و دارای خواص زیر است:

- خاصیت بازتابی: به ازای هر $x \in U$، $(x, x) \in IND_s(B)$.
- خاصیت تقارنی: اگر $(x, x') \in IND_s(B)$ باشد، آنگاه $(x', x) \in IND_s(B)$.
- خاصیت تعدی: اگر $(x, y) \in IND_s(B)$ و $(y, z) \in IND_s(B)$، آنگاه $(x, z) \in IND_s(B)$.

[1] Indiscernible

بنابراین، اگر $(x,x') \in IND_s(B)$، آنگاه x و x' به ازای تمام متغیرهای درون مجموعه B، مقادیر یکسانی را دارا می‌باشند. بدیهی است $IND_s(B)$ یک رابطه هم‌ارزی است. خانواده‌ای از همه کلاس-های هم‌ارزی $IND_s(B)$، یک افراز ایجادشده توسط B است که آن را با نماد $U/IND_s(B)$ یا به اختصار با U/B نشان می‌دهند. همچنین بلوکی از افراز U/B که شامل x است را با $[x]_B$ نشان می-دهند. زیرنویس S در رابطه هم‌ارزی، زمانی که سیستم اطلاعاتی برای تمام روابط هم‌ارزی مشترک باشد، حذف می‌شود (سورج، ۲۰۰۴).

تعریف ۳-۱: به بلوک‌های افراز U/B در اصطلاح یک مجموعه‌ی ابتدایی[1] گویند.

مثال ۳-۲

در مثال ۳-۱، هفت زیرمجموعه ناتهی از متغیرهای شرطی وجود دارند که عبارتند از:

$$B_1 = \{\mathrm{H}\} \quad , \quad B_2 = \{\mathrm{M}\} \quad , \quad B_3 = \{\mathrm{T}\} \quad , \quad B_4 = \{\mathrm{H},\mathrm{M}\}$$
$$B_5 = \{\mathrm{H},\mathrm{T}\} \quad , \quad B_6 = \{\mathrm{M},\mathrm{T}\} \quad , \quad B_7 = \{\mathrm{H},\mathrm{M},\mathrm{T}\}$$

به عنوان مثال، برای مجموعه‌های B_1 و B_7، روابط و کلاس‌های هم‌ارزی عبارتند از:

$$IND(B_1) = \begin{Bmatrix}(x_1,x_1),(x_2,x_2),(x_3,x_3),(x_4,x_4),(x_5,x_5),(x_6,x_6),\\(x_1,x_4),(x_1,x_6),(x_4,x_6),(x_2,x_3),(x_2,x_5),(x_3,x_5)\end{Bmatrix}$$
$$U/B_1 = \{\{x_1,x_4,x_6\},\{x_2,x_3,x_5\}\}$$

$$IND(B_7) = \{(x_1,x_1),(x_2,x_2),(x_3,x_3),(x_4,x_4),(x_5,x_5),(x_6,x_6),(x_2,x_5)\}$$
$$U/B_7 = \{\{x_1\},\{x_3\},\{x_4\},\{x_2,x_5\},\{x_6\}\}$$

به همین صورت کلاس‌های هم‌ارزی را برای سایر زیرمجموعه‌ها می‌توان به دست آورد. همچنین کلاسی از افرازهای U/B_1 و U/B_7 که شامل فرد x_2 است عبارتند از:

$$[x_2]_{B_1} = \{x_2,x_3,x_5\} \quad , \quad [x_2]_{B_7} = \{x_2,x_5\}$$

[1] Elementary

تقریب مجموعه‌ها

فرض کنید $S = (U, C)$ یک سیستم اطلاعات و فرض کنید $B \subseteq C$ و $X \subseteq U$. آنگاه تقریب پایین مجموعه X بر اساس مجموعه متغیر B، به صورت رابطه (۳-۴) تعریف می‌شود و شامل تمام افرادی است که به طور قطع بر اساس مجموعه متغیرهای B، درون مجموعه‌ی X قرار می‌گیرند.

$$\underline{B}(X) = \{x | [x]_B \subseteq X\} \quad (۳-۴)$$

تقریب بالای مجموعه X بر اساس مجموعه متغیر B، به صورت رابطه (۳-۵) تعریف می‌شود و شامل تمام افرادی است که ممکن است، بر اساس مجموعه متغیرهای B، عضو مجموعه X باشند.

$$\overline{B}(X) = \{x | [x]_B \cap X \neq \emptyset\} \quad (۳-۵)$$

به اختلاف میان تقریب بالا و تقریب پایین مجموعه X بر اساس مجموعه متغیر B، ناحیه مرزی گویند که به صورت رابطه (۳-۶) تعریف می‌شود و شامل مجموعه افرادی است که، بر اساس مجموعه متغیرهای B، نمی‌توان به طور قطع در مورد تعلق آن‌ها به مجموعه X تصمیم‌گیری کرد.

$$BN_s(X) = \overline{B}(X) - \underline{B}(X) \quad (۳-۶)$$

مجموعه $U - \underline{B}(X)$ ، شامل تمام افرادی است که به طور قطع، بر اساس مجموعه متغیر B، درون مجموعه X قرار نمی‌گیرند. اگر $BN_s(X)$ تهی باشد، مجموعه X، بر اساس مجموعه متغیر B دقیق[1] است و در غیر این صورت یک مجموعه مبهم است (سورج، ۲۰۰۴).

مثال ۳-۳

در مثال ۳-۱ اگر $X = \{x: F = yes\}$ باشد، آنگاه تقریب بالا و پایین مجموعه X بر اساس مجموعه متغیر شرطی $B = \{H, M, T\}$ به صورت زیر به‌دست می‌آید:

$$X = \{x: F = yes\} = \{x_1, x_2, x_3, x_6\}$$

[1] crisp

$$U/B = \{\{x_1\},\{x_3\},\{x_4\},\{x_2,x_5\},\{x_6\}\}$$

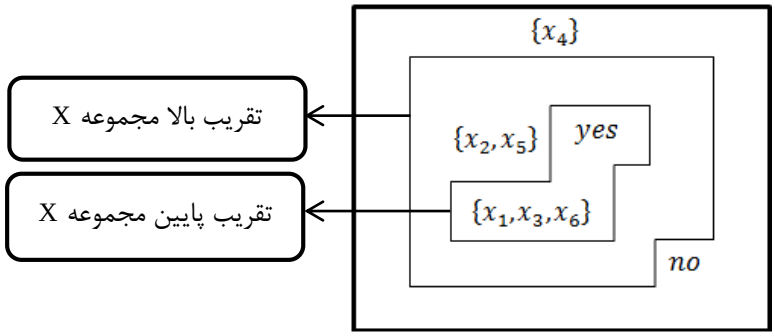
$$[x_1]_B = \{x_1\} \quad , \quad [x_2]_B = \{x_2,x_5\} \quad , \quad [x_3]_B = \{x_3\} \quad , \quad [x_4]_B = \{x_4\}$$

$$[x_5]_B = \{x_2,x_5\} \quad , \quad [x_6]_B = \{x_6\}$$

$$\Longrightarrow \quad \underline{B}(X) = \{x|[x]_B \subseteq X\} = \{x_1,x_3,x_6\} \quad ,$$

$$\overline{B}(X) = \{x|[x]_B \cap X \neq \emptyset\} = \{x_1,x_2,x_3,x_5,x_6\} \quad , \quad BN_s(X) = \{x_2,x_5\} \neq \emptyset$$

از آنجایی‌که ناحیه مرزی X ناتهی است، می‌توان نتیجه گرفت که X یک مجموعه مبهم براساس مجموعه متغیرهای B است. شمایی از مجموعه X به همراه تقریب‌های بالا و پایین آن بر اساس مجموعه متغیر B، در شکل ۳-۱ نشان داده شده است.



شکل ۳-۱: مجموعه X و دو تقریب بالا و پایین آن بر حسب مجموعه متغیرهای B

دقت و کیفیت تقریب

دقیق یا مبهم بودن مجموعه $X \subseteq U$ بر اساس مجموعه متغیر $B \subseteq C$، را می‌توان با استفاده از معیارهای دقت و کیفیت تقریب که به ترتیب در روابط (۳-۷) و (۳-۸) آمده‌اند، بررسی کرد.

$$\alpha_\beta(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \qquad (۳-۷)$$

$$\rho_B(X) = \frac{|\underline{B}(X)|}{|X|} \qquad (۳-۸)$$

که در آن‌ها $|\bullet|$، تعداد اعضا مجموعه مربوطه است. با توجه به اینکه $\underline{B}(X) \subseteq X \subseteq \overline{B}(X)$، آنگاه $0 \leq \alpha_B(X) \leq \rho_B(X) \leq 1$.

اگر $\alpha_B(X) = 1$ آنگاه X بر اساس B دقیق و در غیر این صورت مبهم می‌باشد. همچنین $\rho_B(X) = 0$، اگر و تنها اگر $\alpha_\beta(X) = 0$، و $\rho_B(X) = 1$ اگر و تنها اگر $\alpha_\beta(X) = 1$ (سورج، ۲۰۰۴).

مثال ۳-۴

در مثال ۳-۳ دقت و کیفیت تقریب را برای دو زیرمجموعه دلخواه C در زیر محاسبه می‌کنیم.

اگر $B = \{H, M, T\}$ باشد، آنگاه:

$$\underline{B}(X) = \{x_1, x_3, x_6\} \quad , \quad \overline{B}(X) = \{x_1, x_2, x_3, x_5, x_6\}$$

و بنابراین دقت و کیفیت تقریب برابرند با:

$$\alpha_\beta(X) = \frac{3}{5} \quad , \quad \rho_B(X) = \frac{3}{6}$$

حال اگر $B = \{H\}$ باشد، آنگاه:

$$\underline{B}(X) = \emptyset \quad , \quad \overline{B}(X) = U$$

بنابراین:

$$\rho_B(X) = 0 \quad , \quad \alpha_\beta(X) = 0$$

با استفاده از دقت و کیفیت تقریب برای هر زیرمجموعه دلخواه U، می‌توان دقت و کیفت تقریب را برای یک جدول تصمیم و به دنبال آن یک الگوریتم تصمیم‌گیری محاسبه کنیم که در ادامه به آن می‌پردازیم.

دقت و کیفیت رده‌بندی

فرض کنید $S = (U, C \cup D)$ یک سیستم تصمیم و $G = \{X_1, X_2, \dots, X_n\}$، مجموعه‌ای از زیرمجموعه‌های جامعه U باشد که در آن $X_i \cap X_j = \emptyset$ و $U = \cup_{i=1}^{n} X_i$ هستند، آنگاه G یک افراز جامعه U است و هر کدام از X_i ها یک رده از این افراز محسوب می‌شوند.

به ازای هر $B \subseteq C$، تقریب پایین و بالای مجموعه G به ترتیب به صورت زیر تعریف می‌شوند.

$$\underline{B}(G) = \{\underline{B}(X_1), \underline{B}(X_2), \dots, \underline{B}(X_n)\}$$

$$\overline{B}(G) = \{\overline{B}(X_1), \overline{B}(X_2), \dots, \overline{B}(X_n)\}$$

آنگاه دقت و کیفیت رده‌بندی به ترتیب به صورت روابط (۳-۹) و (۳-۱۰) تعریف می‌شوند (دیمیتراس[1] و همکاران، ۱۹۹۹).

$$\mu_B(G) = \frac{\sum_{i=1}^{n}|\underline{B}(X_i)|}{\sum_{i=1}^{n}|\overline{B}(X_i)|} \quad (۳-۹)$$

$$\partial_B(G) = \frac{\sum_{i=1}^{n}|\underline{B}(X_i)|}{|U|} \quad (۳-۱۰)$$

مثال ۳-۵

در جدول ۳-۱، اگر $X_1 = \{x|F(x) = yes\}$، $X_2 = \{x|F(x) = no\}$، $U = \cup_{i=1}^{2} X_i$ و $B =$ $\{H, M, T\}$، آنگاه:

$$\underline{B}(X_1) = \{x_1, x_3, x_6\} \quad , \quad \overline{B}(X_1) = \{x_1, x_2, x_3, x_5, x_6\}$$

$$\underline{B}(X_2) = \{x_4\} \quad , \quad \overline{B}(X_2) = \{x_2, x_4, x_6\}$$

بنابراین دقت و کیفیت رده‌بندی الگوریتم تصمیم‌گیری موجود برای G عبارتند از:

$$\mu_B(G) = \frac{3+1}{5+3} = \frac{4}{8} = \frac{1}{2} \quad , \quad \partial_B(G) = \frac{3+1}{6} = \frac{4}{6} = \frac{2}{3}$$

بنابراین الگوریتم تصمیم‌گیری استخراج شده از جدول ۳-۱ به ترتیب دارای دقت و کیفیت رده‌بندی $\frac{1}{2}$ و $\frac{2}{3}$ است.

[1] Dimitras

۳-۱-۴ کاهش‌ها و هسته متغیرها

در یک سیستم تصمیم، ممکن است وجود بعضی متغیرها غیرضروری بوده و باعث بزرگی بیهوده جدول و به دنبال آن پیچیدگی الگوریتم تصمیم‌گیری استخراج شده از آن شود. بنابراین یافتن متغیرهای غیرضروری[1] و حذف آن‌ها از جداول تصمیم بسیار مهم و حائز اهمیت است. از این‌رو در این بخش، روش کاهش متغیرها را با استفاده از نظریه مجموعه مبهم، معرفی می‌کنیم.

تعریف ۳-۲: فرض کنید $S = (U, C)$ یک سیستم تصمیم بوده و $B \subseteq C$ و $a \in B$. گوییم a یک متغیر غیرضروری در B است اگر $IND_s(B) = IND_s(B - \{a\})$ ، در غیر این صورت متغیر a در B ضروری است.

تعریف ۳-۳: مجموعه B مستقل است اگر همه متغیرهای آن ضروری باشند.

تعریف ۳-۴: هر زیرمجموعه B' از B را یک کاهش[2] مجموعه B نامند، اگر B' مستقل و $IND_s(B') = IND_s(B)$. به عبارت دیگر کاهش‌ها، به زیرمجموعه‌هایی از متغیرها اطلاق می‌شوند که قادر هستند افراز یکسانی همانند مجموعه اصلی متغیرها، روی افراد جامعه ایجاد کنند. کاهش مجموعه متغیر B را با $Red(B)$ نشان می‌دهند. متغیرهایی که متعلق به کاهش‌ها نیستند، غیرضروری محسوب می‌شوند. بدیهی است یک مجموعه متغیر می‌تواند کاهش‌های بسیاری داشته باشد.

تعریف ۳-۵: به مجموعه تمام متغیرهای ضروری در B که $B \subseteq C$، هسته B گویند و آن را با نماد $Core(B)$ به صورت زیر نشان می‌دهند.

$$Core(B) = \cap Red(B)$$

[1] Dispensable
[2] Reduct

که در آن $Red(B)$، مجموعه همه کاهش‌های B است. از آنجایی‌که هسته، اشتراک همه کاهش‌ها است، از این‌رو مهمترین زیرمجموعه از متغیرها است و حذف هر یک از اعضاء آن قطعاً در توان رده-بندی الگوریتم تصمیم‌گیری مؤثر خواهد بود (سورج، ۲۰۰۴).

مثال ۳-۶

در جدول ۳-۱، برای مجموعه متغیر $B = \{H, M, T\}$ کاهش‌ها و هسته به صورت زیر به دست می-آیند.

$$IND(B) = \{(x_1, x_1), (x_2, x_2), (x_3, x_3), (x_4, x_4), (x_5, x_5), (x_6, x_6), (x_2, x_5)\}$$

$$B_1' = \{H\} \quad , \quad B_2' = \{M\} \quad , \quad B_3' = \{T\} \quad , \quad B_4' = \{H, M\}$$

$$B_5' = \{H, T\} \quad , \quad B_6' = \{M, T\} \quad \& \quad B_1', B_2', B_3', B_4', B_5', B_6' \subset \mathrm{B}$$

حال روابط هم‌ارزی را برای هر یک از زیرمجموعه‌ها به دست می‌آوریم:

$$IND(B_1') = \begin{Bmatrix} (x_1, x_1), (x_2, x_2), (x_3, x_3), (x_4, x_4), (x_5, x_5), (x_6, x_6) \\ , (x_1, x_4), (x_1, x_6), (x_4, x_6), (x_2, x_3), (x_2, x_5), (x_3, x_5) \end{Bmatrix}$$

$$IND(B_2') = \begin{Bmatrix} (x_1, x_1), (x_2, x_2), (x_3, x_3), (x_4, x_4), (x_5, x_5), (x_6, x_6) \\ , (x_1, x_3), (x_1, x_4), (x_1, x_6), (x_3, x_4), (x_3, x_6), (x_4, x_6), (x_2, x_5) \end{Bmatrix}$$

$$IND(B_3') = \begin{Bmatrix} (x_1, x_1), (x_2, x_2), (x_3, x_3), (x_4, x_4), (x_5, x_5), (x_6, x_6) \\ , (x_1, x_2), (x_1, x_5), (x_2, x_5), (x_3, x_6) \end{Bmatrix}$$

$$IND(B_4') = \begin{Bmatrix} (x_1, x_1), (x_2, x_2), (x_3, x_3), (x_4, x_4), (x_5, x_5), (x_6, x_6) \\ , (x_1, x_4), (x_1, x_6), (x_4, x_6), (x_2, x_5) \end{Bmatrix}$$

$$IND(B_5') = \{(x_1, x_1), (x_2, x_2), (x_3, x_3), (x_4, x_4), (x_5, x_5), (x_6, x_6), (x_2, x_5)\}$$

$$IND(B_6') = \{(x_1, x_1), (x_2, x_2), (x_3, x_3), (x_4, x_4), (x_5, x_5), (x_6, x_6), (x_3, x_6), (x_2, x_5)\}$$

با توجه به روابط هم‌ارزی ایجاد شده توسط هر یک از زیرمجموعه‌ها داریم:

$$IND(B) = IND(B_5')$$

از طرفی B_5' مستقل است، بنابراین تنها کاهش ممکن برای مجموعه متغیر B، مجموعه B_5' است و در نتیجه هسته B نیز $\acute{B}_5$ می‌باشد.

$$Red(B) = \{H, T\}$$

$Core(B) = \{H, T\}$

بنابراین با استفاده از روشی که در این زیربخش معرفی شد می‌توانیم یک الگوریتم تصمیم‌گیری را با حفظ قدرت رده‌بندی ساده کنیم.

۳-۱-۵ نحوه رده‌بندی یک مشاهده

هدف اصلی از معرفی نظریه مجموعه مبهم، انجام عمل رده‌بندی برای یک مشاهده جدید است. پس از استخراج الگوریتم تصمیم‌گیری و ساده کردن آن، نوبت به استفاده از آن برای رده‌بندی یک مشاهده جدید می‌رسد. با توجه به این‌که ممکن است مشخصات یک مشاهده جدید با چند قانون مختلف مطابقت داشته باشد، از این‌رو برای انتخاب بهترین قانون برای این مشاهده و انجام عمل رده‌بندی، نیاز به معرفی معیارهایی به منظور سنجش میزان اهمیت هر قانون خواهیم داشت. بنابراین در ادامه، ابتدا معیارهای محافظ تصمیم‌گیری را معرفی کرده و سپس با استفاده از آن‌ها عمل رده‌بندی را برای یک مشاهده جدید انجام می‌دهیم.

معیارهای محافظ تصمیم

در سیستم‌های تصمیم‌گیری، معیارهایی تعریف می‌شوند که بر اساس آن‌ها می‌توان به میزان قطعیت و اهمیت هر قانون پی برد. در ادامه برخی از این معیارهای مهم را معرفی می‌کنیم.

تعریف ۳-۶: معیار پشتیبان[1] قانون، عبارت است از تعداد دفعات تکرار یک قانون که به صورت رابطه زیر تعریف می‌شود.

$Support_x(C, D) = |[x]_C \cap [x]_D|$

تعریف ۳-۷: معیار توان[2] قانون، به نسبت تعداد دفعات تکرار یک قانون به تعداد کل قانون‌ها گفته می‌شود. که عبارت است از:

$\sigma_x(C, D) = \frac{Support_x(C,D)}{|U|}$

[1] Rule Support Factor
[2] Rule Strength Factor

تعریف ۳-۸: معیار قطعیت[1] قانون، عبارت است از:

$$cer_x(C,D) = \frac{Support_x(C,D)}{|[x]_C|} = \frac{\sigma_x(C,D)}{\pi([x]_C)} \quad , \quad \pi([x]_C) = \frac{|[x]_C|}{|U|}$$

اگر $cer_x(C,D) = 1$ باشد، آنگاه تصمیم مربوطه یک تصمیم قطعی و در غیر این صورت یک تصمیم غیرقطعی می‌باشد (پاولاک، ۲۰۰۲).

تعریف ۳-۹: معیار پوشش[2] قانون تصمیم، عبارت است از:

$$cov_x(C,D) = \frac{Support_x(C,D)}{|[x]_D|} = \frac{\sigma_x(C,D)}{\pi([x]_D)} \quad , \quad \pi([x]_D) = \frac{|[x]_D|}{|U|}$$

قانونی که معیارهای محافظ تصمیم بزرگتری داشته باشد، دارای اعتبار بیشتری است.

مثال ۳-۷

برای جدول ۳-۱، الگوریتم تصمیم‌گیری ساده شده را به همراه معیارهای محافظ تصمیم به دست می‌آوریم. با توجه به مثال ۳-۶، دیدیم که $C' = \{H,T\}$ تنها کاهش مجموعه متغیر $C = \{H,M,T\}$ است. بنابراین قانون تصمیم استخراج شده از مشخصات فرد x_1 عبارت است از:

$$1.\ (H = no)\&(T = high) \rightarrow (F = yes)$$

معیارهای محافظ تصمیم برای قانون (۱) به صورت زیر به دست می‌آیند.

$$\mathrm{U}/C' = \{\{x_1\},\{x_2,x_5\},\{x_3\},\{x_4\},\{x_6\}\} \quad , \quad \mathrm{U}/D = \{\{x_1,x_2,x_3,x_6\},\{x_4,x_5\}\}$$

$$[x_1]_{C'} = \{x_1\} \quad , \quad [x_1]_D = \{x_1,x_2,x_3,x_6\}$$

$$\Rightarrow Support_{x_1}(C',D) = |[x_1]_{C'} \cap [x_1]_D| = |\{x_1\} \cap \{x_1,x_2,x_3,x_6\}| = |\{x_1\}| = 1$$

$$\sigma_{x_1}(C',D) = \frac{Support_{x_1}(C',D)}{|U|} = \frac{1}{6} \quad , \quad cer_{x_1}(\acute{C},D) = \frac{Support_{x_1}(C',D)}{|[x_1]_{C'}|} = \frac{1}{1} = 1$$

$$cov_{x_1}(C',D) = \frac{Support_{x_1}(C',D)}{|[x_1]_D|} = \frac{1}{4}$$

به همین صورت سایر قانون‌ها را استخراج کرده و معیارهای محافظ تصمیم برای آن‌ها محاسبه می‌شوند که نتایج به قرار زیر است:

---

[1] Rule Certainty Factor
[2] Rule Coverage Factor

2. $(H = yes)\&(T = high) \rightarrow (F = yes)$,

$$support = 1\ ,\ \sigma = \frac{1}{6}\ ,\ cer = \frac{1}{2}\ ,\ cov = \frac{1}{4}$$

3. $(H = yes)\&(T = very\ high) \rightarrow (F = yes)$,

$$support = 1\ ,\ \sigma = \frac{1}{6}\ ,\ cer = 1\ ,\ cov = \frac{1}{4}$$

4. $(H = no)\&(T = normal) \rightarrow (Fl = no)$,

$$support = 1\ ,\ \sigma = \frac{1}{6}\ ,\ cer = 1\ ,\ cov = \frac{1}{2}$$

5. $(H = yes)\&(T = high) \rightarrow (F = no)$,

$$support = 1\ ,\ \sigma = \frac{1}{6}\ ,\ cer = \frac{1}{2},\ cov = \frac{1}{2}$$

6. $(H = no)\&(T = very\ high) \rightarrow (F = yes)$,

$$support = 1\ ,\ \sigma = \frac{1}{6}\ ,\ cer = 1,\ cov = \frac{1}{4}$$

بنابراین شاهد یک الگوریتم تصمیم‌گیری ساده‌شده هستیم که شامل مجموعه‌ای از قانون‌های قطعی و غیرقطعی است.

رده‌بندی مشاهده جدید

حال اگر یک مشاهده جدید داشته باشیم و بخواهیم رده متغیر پاسخ مربوط به آن را پیش‌بینی کنیم، ممکن است یکی از حالت‌های زیر برای آن اتفاق بیافتد:

۱. مشخصات متغیرهای شرطی مشاهده جدید، دقیقاً با یکی از قانون‌های قطعی موجود در الگوریتم تصمیم‌گیری مطابقت داشته باشد.

۲. مشخصات متغیرهای شرطی مشاهده جدید، با یکی از قانون‌های غیرقطعی موجود در الگوریتم تصمیم‌گیری مطابقت داشته باشد.

۳. مشخصات متغیرهای شرطی مشاهده جدید، با هیچ‌یک از قانون‌های موجود در الگوریتم تصمیم‌گیری مطابقت نداشته باشد.

۴. مشخصات متغیرهای شرطی مشاهده جدید، با بیش از یک قانون موجود در الگوریتم تصمیم-گیری مطابقت داشته باشد.

برای حالت اول، پیش‌بینی به آسانی انجام می‌شود. برای حالت دوم قانون موجود غیرقطعی است، بنابراین این قانون به همراه چند قانون دیگر دارای مقادیر متغیرهای شرطی یکسان و مقدار متغیر تصمیم متفاوت هستند. از این‌رو تصمیم‌گیرنده از میان این قانون‌ها، قانونی را که از لحاظ معیارهای توان و پوشش دارای اعتبار بیشتری باشد به عنوان قانون پیش‌بینی برای مشاهده جدید انتخاب می-کند. برای حالت چهارم همه قانون‌هایی که با مشخصات مشاهده جدید مطابقت دارند می‌توانند به تصمیم‌گیرنده پیشنهاد شوند. اگر همه قانون‌ها رده مشابهی را انتخاب کنند آنگاه هیچ ابهامی وجود نخواهد داشت، در غیر این صورت معیارهای محافظ تصمیم، تعیین‌کننده بهترین قانون خواهند بود و مشابه حالت دوم، عمل پیش‌بینی صورت می‌پذیرد. تصمیم‌گیری برای حالت سوم نسبت به حالت‌های دیگر مشکل‌تر است. برای این حالت مجموعه‌ای از قوانین تصمیم که به مشخصات مشاهده جدید شبیه‌تر و نزدیک‌تر باشند به تصمیم‌گیرنده پیشنهاد می‌شوند. بدین منظور اسلوینسکی و همکاران[1] (۱۹۹۴)، روشی را تحت عنوان رابطه نزدیکی ارزشمند[2] پیشنهاد کردند که با استفاده از آن مجموعه-ای از قوانین که بیشترین شباهت را به مشخصات فرد مورد نظر دارند به تصمیم‌گیرنده پیشنهاد می-شود.

رابطه نزدیکی ارزشمند

فرض کنید x یک مشاهده و r یک قانون باشد. رابطه نزدیکی ارزشمند، به میزان نزدیکی یا شباهت مشاهده x به قانون r گفته می‌شود که با $(x\ S\ r)$ نشان داده می‌شود. از آنجایی‌که ممکن است قانون r نسبت به یک مشاهده دارای شرط‌های (تعداد متغیرهای شرطی) کمتری باشد، بنابراین میزان نزدیکی قانون r به مشاهده x بر اساس متغیرهای موجود در قانون r مورد بررسی قرار می‌گیرد. اگر

[1] Slowinski
[2] Valued Closeness Relation

$A = \{a_1, a_2, \dots, a_m\}$ مجموعه متغیرهای شرطی موجود در قانون r باشد، و a_1^x، a_2^x،...، a_m^x مقادیر این متغیرها برای مشاهده x باشند، آنگاه رابطه نزدیکی ارزشمند برای متغیر i ام به صورت رابطه (۳-۱۱) محاسبه می‌شود.

$$a_i^x \, S_i a_i^r \Leftrightarrow g_i(a_i^x) \geq g_i(a_i^r) \qquad (۳-۱۱)$$

که در آن، g_i یک تابع حقیقی‌مقدار است. با تعریف رابطه (۳-۱۱)، مجموعه متغیرهای شرطی به دو زیرمجموعه جدا از هم با نام‌های اتحاد موافق[1] و اتحاد ناسازگار[2] تقسیم می‌شوند که به ترتیب به صورت روابط زیر تعریف می‌شوند.

$$C(x\, S\, r) = \{a_i \in A;\ g_i(a_i^x) \geq g_i(a_i^r)\}$$

$$D(x\, S\, r) = \{a_i \in A; g_i(a_i^x) < g_i(a_i^r)\}$$

به عبارت دیگر، اتحاد موافق شامل متغیرهایی است که نزدیکی مشاهده x و قانون r را تأیید می‌کنند و در مقابل، اتحاد ناسازگار شامل متغیرهایی است که نزدیکی مشاهده x و قانون r را تأیید نمی‌کنند. برای تأیید یا رد گزاره‌ی $(x\, S\, r)$، دو شرط زیر باید برقرار باشند:

۱. شرط توافق: $C(x\, S\, r)$ به قدر کافی با اهمیت باشد.

برای ارزیابی شرط توافق از شاخص توافق[3] که به صورت رابطه زیر تعریف شده استفاده می‌شود.

$$c(x,r) = \frac{\sum_{a_i \in C(x\, S\, r)} k_i}{\sum_{a_i \in A} k_i} \quad , \quad c(x,r) \in [0,1]$$

که در آن k_i، یک وزن مثبت برای متغیر a_i است. به عبارت دیگر $c(x,r)$ بیانگر اهمیت نسبی اتحاد $C(x\, S\, r)$ در مجموعه A می‌باشد. با در نظر گرفتن یک مقدار آستانه مانند $S \in [0,1]$ برای شاخص توافق، میزان اهمیت مورد نظر (تعیین شده توسط کارشناس) در شرط توافق مورد بررسی قرار می‌گیرد.

[1] Concordance Coalition
[2] Discordance Coalition
[3] Concordance Index

۲. شرط ناسازگاری: برای هر $a_i \in D(x\ S\ r)$، میزان ناسازگاری مشاهده x با قانون r به اندازه کافی بزرگ باشد.

برای اعمال شرط دوم، نیاز به معرفی یک مقدار آستانه‌ی v_i برای هر متغیر a_i داریم. به عبارت دیگر باید رابطه زیر برقرار باشد.

$$g_i(a_i^x) - g_i(a_i^r) < v_i \quad , \quad a_i \in D(x\ S\ r)$$

بنابراین به طور کلی برای تأیید وجود رابطه نزدیکی ارزشمند بین مشاهده x و قانون r داریم:

$$x\ S\ r \Leftrightarrow \begin{cases} c(x,r) \geq s \\ \text{و} \\ g_i(a_i^x) - g_i(a_i^r) < v_i \quad , \quad a_i \in D(x\ S\ r) \end{cases}$$

مقادیر مجهول k_i، S و v_i باید توسط تصمیم‌گیرنده (کارشناس) تعیین شوند. به همین صورت برای تمامی قانون‌های موجود، وجود رابطه نزدیکی ارزشمند را با مشاهده x بررسی می‌کنیم. در نهایت، شاهد مجموعه‌ای از قوانین خواهیم بود که وجود شباهت بین آن‌ها و مشاهده x مورد تأیید قرار گرفته و تصمیم‌گیرنده می‌تواند بر اساس معیارهای محافظ تصمیم، یکی از این قانون‌ها را برای پیش‌بینی رده مشاهده جدید انتخاب کند (بویسو[1]، ۱۹۹۶).

۳-۱-۶ گسسته‌سازی

نظریه مجموعه مبهم، تنها قابلیت کار کردن با متغیرهای گسسته را دارد. اما در جداول و سیستم‌های تصمیم، ممکن است بعضی از متغیرها از نوع پیوسته باشند. لذا برای کار کردن با نظریه مجموعه مبهم، لازم است این متغیرها به متغیر گسسته تبدیل شوند. به عمل تبدیل یک متغیر پیوسته به گسسته، در اصطلاح عمل گسسته‌سازی گویند. گسسته‌سازی یک تکنیک پیش‌پردازش است که طی آن، حدود تغییرات یک متغیر به زیرفاصله‌هایی تقسیم می‌شود که بتوانند بهترین رده‌بندی را برای اشیاء یا افراد انجام دهند. فاصله‌های بلند موجب از دست دادن اطلاعات و به وجود آمدن اغتشاش در

[1] Bouyssou

داده‌ها خواهند شد، اما برای رده‌بندی یک مشاهده جدید مناسب‌اند. فاصله‌های کوتاه اغتشاش موجود در داده‌ها را کم می‌کنند، اما قدرت رده‌بندی یک مشاهده جدید را کاهش می‌دهند. بنابراین انتخاب یک فاصله مناسب امری مهم و حائز اهمیت است. روش‌های مختلفی برای انجام عمل گسسته‌سازی وجود دارد که یکی از آن‌ها، روشی مبتنی بر الگوریتم استدلال بولی[1] است (انگوین[2] و همکاران، ۱۹۹۸).

روش مبتنی بر الگوریتم استدلال بولی (*BRA*)

فرض کنید $S = (U, C \cup D, V, f)$ صورت دیگری از نمایش یک سیستم تصمیم باشد، که در آن $C = \{a_1, \dots, a_k\}$، $U = \{x_1, \dots, x_n\}$ و $D = \{d\}$. مجموعه $V = \bigcup_{a \in C} V_a$ مجموعه مقادیر متغیرها است که در آن $V_a = [l_a, r_a) \subset R$ حدود تغییرات متغیر a است. برای هر $a \in C$، تابع $f: U \to V_a$ را خواهیم داشت که بیانگر مقدار متغیر a برای فرد x است.

تعریف ۳-۱۰: هر جفت (a, b) به عنوان عضوی از یک افراز روی V_a تعریف می‌شود که در آن $a \in C$ و $b \in R$ است. این جفت در اصطلاح یک برش روی V_a نامیده می‌شود.

فرض کنید $B_a = \{(a, b_1^a), (a, b_2^a), \dots, (a, b_{k_a}^a)\}$، مجموعه برش‌های ممکن برای $a \in C$ باشد که در آن $k_a \in N$ و $l_a = b_0^a < b_1^a < b_2^a < \cdots < b_{k_a}^a < b_{k_a+1}^a = r_a$ هستند. این مجموعه برش‌ها، افرازی روی مجموعه V_a به صورت زیر ایجاد می‌کنند:

$$B_a = [b_0^a, b_1^a) \cup [b_1^a, b_2^a) \cup \dots \cup \left[b_{k_a}^a, b_{k_a+1}^a\right]$$

بنابراین مجموعه کل افرازهای ایجاد شده توسط متغیرهای شرطی عبارت است از:

$$B = \bigcup_{a \in C} B_a$$

[1] Boolean Reasoning Algorithm
[2] Nguyen

مجموعه B، سیستم تصمیم اصلی S را به سیستم تصمیم گسسته $S^B = (U, C \cup D, V^B, f^B)$ تبدیل می‌کند، به طوری که:

$$\forall x \in U, a \in C, i \in \{0,1,\dots,k_a\} \;\; ; \;\; f^B(x_a) = i \Leftrightarrow f(x_a) \in [b_i^a, b_{i+1}^a]$$

مجموعه برش‌های متفاوت، سیستم‌های تصمیم گسسته متفاوت تولید می‌کنند. هدف اصلی در گسسته‌سازی، انتخاب کوچکترین زیرمجموعه از برش‌ها است که بتواند افراد جامعه U را مانند مجموعه اصلی برش‌ها از هم ممیزی کند.

حال باید نقاط ممکن برای انجام برش مشخص شوند. فرض کنید برای هر $a \in C$، دنباله $\mathrm{V}_1^{\mathrm{a}} < \mathrm{V}_2^{\mathrm{a}} < \cdots < \mathrm{V}_{\mathrm{n_a}}^{\mathrm{a}}$ مجموعه مقادیر مرتب‌شده‌ای باشد که متغیر a می‌تواند اختیار کند، آنگاه متغیرهای بولی به صورت رابطه زیر تعریف می‌شوند.

$$\forall a \in C \;,\; 1 \le l \le n_{a-1} : \; p_l^a = [V_l^a, V_{l+1}^a)$$

و مجموعه همه متغیرهای بولی[1] جامعه‌ی U را به صورت زیر نشان می‌دهند.

$$BV(U) = \{p_1^{a_1}, \dots, p_{n_{a-1}}^{a_1}, \dots, p_1^{a_k}, \dots, p_{n_{a-1}}^{a_k}\}$$

مجموعه همه برش‌های ممکن روی a را می‌توان به صورت رابطه (۳-۱۲) تعریف نمود.

$$B_a = \left\{\left(a, \frac{V_1^a + V_2^a}{2}\right), \left(a, \frac{V_2^a + V_3^a}{2}\right), \dots, \left(a, \frac{V_{n_a-1}^a + V_{n_a}^a}{2}\right)\right\} \qquad (۳-۱۲)$$

در گام بعدی باید زیرمجموعه‌ای از این برش‌ها به گونه‌ای انتخاب شود که بتواند تمامی افراد جامعه را از هم ممیز کند. برای پیدا کردن این برش‌ها از روشی مبتنی بر استدلال بولی استفاده می‌کنیم. این روش، جدول S^* را از روی جدول S طی مراحل زیر می‌سازد.

[1] Boolean Variable

۱. ستون‌های جدول S^* را مقادیر p_l^a ها، و سطرهای آن را جفت متغیرهایی با رده تصمیم متفاوت تشکیل می‌دهند. اگر جفت متغیر i ام توسط برش j ام به درستی از هم ممیز شوند، آنگاه درایه S_{ij}^* مقدار یک و در غیر این صورت مقدار صفر را می‌گیرد.

۲. تابع تمایز را برای تمام جفت افرادی که در سطرهای جدول قرار دارند طبق رابطه (۳-۱۳) به دست می‌آوریم.

$$\forall\ k = 1, \dots, n_a \quad , \quad \psi(x_i, x_j) = \vee_{a \in C} \{p_k^a | p_k^a = 1\} \qquad (۱۳-۳)$$

که در آن، ∨ عملگر انفصال[1] نامیده می‌شود. رابطه (۳-۱۳) به این معنی است که برای ممیز کردن جفت فرد (x_i, x_j)، حداقل به یک برش در فاصله‌هایی که متغیر بولی آن‌ها مقدار یک را گرفته، نیاز است. مثلاً اگر $\psi(x_i, x_j) = (p_1^{a_1} \vee p_2^{a_3})$، یعنی برای ممیز کردن x_i و x_j از هم، نیاز به حداقل یک برش در $[V_1^{a_1}, V_2^{a_1})$ یا $[V_2^{a_3}, V_3^{a_3})$ خواهد بود.

۳. در پایان، برای یافتن کوچکترین زیرمجموعه از متغیرهای بولی، تابع بولی را به صورت رابطه (۳-۱۴) تعریف می‌کنند.

$$\Phi^U = \wedge \{\psi(x_i, x_j) : d(x_i) \neq d(x_j)\} \qquad (۱۴-۳)$$

که در آن منظور از $d(x_i)$، مقدار متغیر پاسخ مربوط به متغیر x_i و ∧ عملگر اتصال[2] است. با استفاده از قوانین جبر بولی (معرفی‌شده در قسمت ضمیمه)، مجموعه‌های موجود در Φ^U تا حد امکان ساده شده و از میان مجموعه‌های باقیمانده، کوچکترین مجموعه به عنوان مجموعه برش‌های مناسب، انتخاب می‌شود.

برای فهم بهتر موضوع، به بیان یک مثال ساده می‌پردازیم. فرض کنید یک سیستم تصمیم به شکل جدول ۳-۲ داشته باشیم.

---

[1] Disjunction
[2] Conjunction

جدول ۳-۲: جدول تصمیم (مثال)

S	a_1	a_2	d
x_1	0.8	2	1
x_2	1	0.5	0
x_3	1.3	3	0
x_4	1.4	1	1
x_5	1.4	2	0
x_6	1.6	3	1
x_7	1.3	1	1

در این جدول $C = \{a_1, a_2\}$ مجموعه دو متغیر شرطی و $D = \{d\}$ مجموعه متغیر پاسخ است. متغیرهای شرطی هر دو از نوع پیوسته هستند و نیاز به گسسته‌سازی دارند. مقادیر ممکنی این دو متغیر عبارتند از:

$V_{a_1} = [0,2)$

$V_{a_2} = [0,4)$

همچنین مجموعه مقادیر a_1 و a_2 برای مشاهدات U عبارتند از:

$V_1^{a_1} = 0.8\,,\quad V_2^{a_1} = 1\,,\quad V_3^{a_1} = 1.3\,,\quad V_4^{a_1} = 1.4\,,\quad V_5^{a_1} = 1.6$

$V_1^{a_2} = 0.5\,,\quad V_2^{a_2} = 1\,,\quad V_3^{a_2} = 2\,,\qquad V_4^{a_2} = 3$

$p_1^{a_1} = [0.8,1)\ ,\ p_2^{a_1} = [1,1.3)\ ,\ p_3^{a_1} = [1.3,1.4)\ \ ,\ \ p_4^{a_1} = [1.4,1.6)$

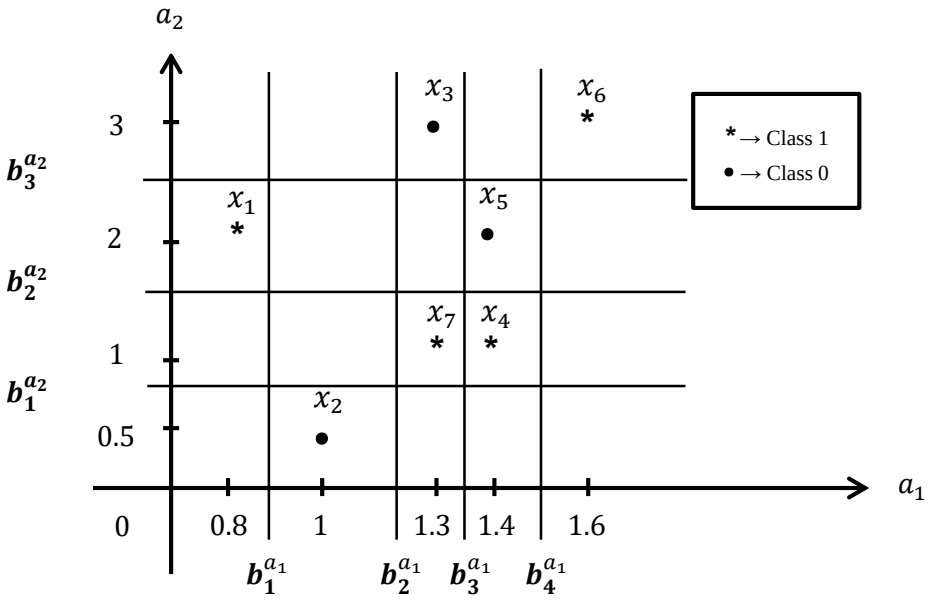
$p_1^{a_2} = [0.5,1)\ ,\ p_2^{a_2} = [1,2)\ \ \ ,\ p_3^{a_2} = [2,3)$

در نتیجه خواهیم داشت:

$$B_{a_1} = \left\{\left(a_1, \frac{V_1^{a_1} + V_2^{a_1}}{2}\right), \left(a_1, \frac{V_2^{a_1} + V_3^{a_1}}{2}\right), \left(a_1, \frac{V_3^{a_1} + V_4^{a_1}}{2}\right), \left(a_1, \frac{V_4^{a_1} + V_5^{a_1}}{2}\right)\right\}$$
$$= \{(a_1, 0.9), (a_1, 1.15), (a_1, 1.35), (a_1, 1.5)\}$$

$$B_{a_2} = \left\{\left(a_2, \frac{V_1^{a_2} + V_2^{a_2}}{2}\right), \left(a_2, \frac{V_2^{a_2} + V_3^{a_2}}{2}\right), \left(a_2, \frac{V_3^{a_2} + V_4^{a_2}}{2}\right)\right\}$$
$$= \{(a_2, 0.75), (a_2, 1.5), (a_2, 2.5)\}$$

بنابراین فضای متغیرهای شرطی توسط مجموعه برش‌های موجود به صورت شکل ۳-۲ افراز می‌شود.



شکل ۳-۲: افراز فضای مشاهدات توسط تمامی برش‌ها

حال جدول S^* را تشکیل می‌دهیم.

جدول ۳-۳: جدول S^* حاصل از جدول ۳-۲

S^*	$p_1^{a_1}$	$p_2^{a_1}$	$p_3^{a_1}$	$p_4^{a_1}$	$p_1^{a_2}$	$p_2^{a_2}$	$p_3^{a_2}$
x_1, x_2	1	0	0	0	1	1	0
x_1, x_3	1	1	0	0	0	0	1
x_1, x_5	1	1	1	0	0	0	0
x_2, x_4	0	1	1	0	1	0	0
x_2, x_6	0	1	1	1	1	1	1
x_2, x_7	0	1	0	0	1	0	0
x_3, x_4	0	1	1	0	0	1	1
x_3, x_6	0	0	1	1	0	0	0
x_3, x_7	0	0	0	0	0	1	0
x_5, x_4	0	0	0	0	0	1	0
x_5, x_6	0	0	0	1	0	0	1
x_5, x_7	0	0	1	0	0	1	0

و توابع تمایز را از روی جدول ۳-۳ به صورت زیر می‌نویسیم.

$$\psi(x_1, x_2) = \left(p_1^{a_1} \vee p_1^{a_2} \vee p_2^{a_2}\right) \quad , \quad \psi(x_1, x_3) = \left\{p_1^{a_1} \vee p_2^{a_1} \vee p_3^{a_2}\right\}$$

$$\psi(x_1, x_5) = \left\{p_1^{a_1} \vee p_2^{a_1} \vee p_3^{a_1}\right\} \quad , \quad \psi(x_2, x_4) = \left\{p_2^{a_1} \vee p_3^{a_1} \vee p_1^{a_2}\right\}$$

$$\psi(x_2, x_6) = \left\{p_2^{a_1} \vee p_3^{a_1} \vee p_4^{a_1} \vee p_1^{a_2} \vee p_2^{a_2} \vee p_3^{a_2}\right\} \quad , \quad \psi(x_2, x_7) = \left\{p_2^{a_1} \vee p_1^{a_2}\right\}$$

$$\psi(x_3, x_4) = \left\{p_2^{a_1} \vee p_2^{a_2} \vee p_3^{a_2}\right\} \quad , \quad \psi(x_3, x_6) = \left\{p_3^{a_1} \vee p_4^{a_1}\right\}$$

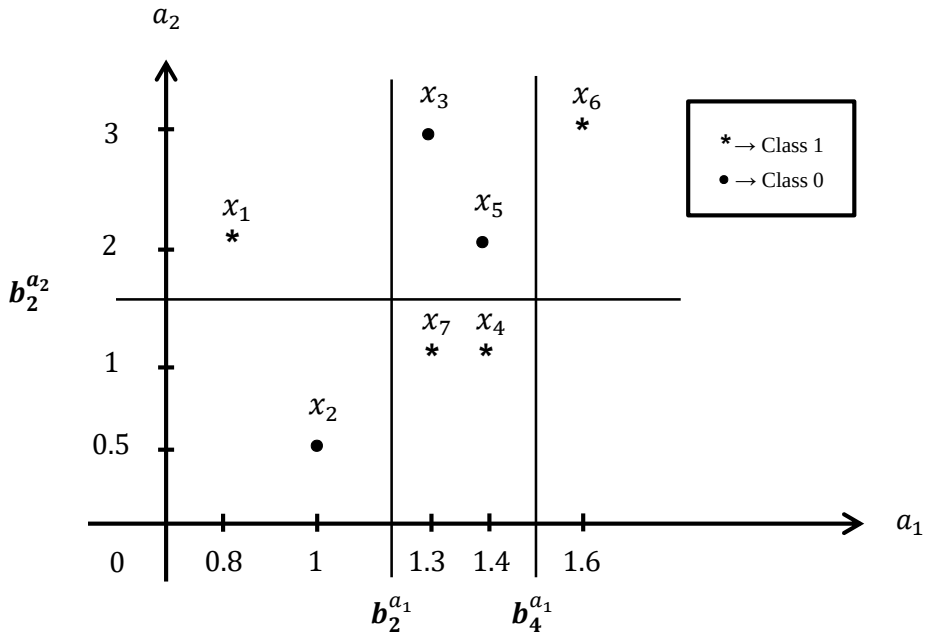
$$\psi(x_3, x_7) = \left\{p_2^{a_2} \vee p_3^{a_2}\right\} \quad , \quad \psi(x_5, x_4) = \left\{p_2^{a_2}\right\}$$

$\psi(x_5, x_6) = \{p_4^{a_1} \vee p_3^{\boldsymbol{a_2}}\}$, $\psi(x_5, x_7) = \{p_3^{a_1} \vee p_2^{\boldsymbol{a_2}}\}$

حال تابع بولی را برای به دست آوردن کوچکترین زیرمجموعه از متغیرهای بولی محاسبه می‌کنیم.

$$\Phi^U = (p_1^{a_1} \vee p_1^{\boldsymbol{a_2}} \vee p_2^{\boldsymbol{a_2}}) \wedge \{p_1^{a_1} \vee p_2^{a_1} \vee p_3^{\boldsymbol{a_2}}\} \wedge \{p_1^{a_1} \vee p_2^{a_1} \vee p_3^{a_1}\} \wedge \{p_2^{a_1} \vee p_3^{a_1} \vee p_1^{\boldsymbol{a_2}}\} \wedge \{p_2^{a_1} \vee p_3^{a_1} \vee p_4^{a_1} \vee p_1^{\boldsymbol{a_2}} \vee p_2^{\boldsymbol{a_2}} \vee p_3^{\boldsymbol{a_2}}\} \wedge \{p_2^{a_1} \vee p_1^{\boldsymbol{a_2}}\} \wedge \{p_2^{a_1} \vee p_2^{\boldsymbol{a_2}} \vee p_3^{\boldsymbol{a_2}}\} \wedge \{p_3^{a_1} \vee p_4^{a_1}\} \wedge \{p_2^{\boldsymbol{a_2}} \vee p_3^{\boldsymbol{a_2}}\} \wedge \{p_2^{\boldsymbol{a_2}}\} \wedge \{p_4^{a_1} \vee p_3^{\boldsymbol{a_2}}\} \wedge \{p_3^{a_1} \vee p_2^{\boldsymbol{a_2}}\} = \{p_2^{a_1} \vee p_4^{a_1} \vee p_2^{\boldsymbol{a_2}}\} \wedge \{p_2^{a_1} \vee p_3^{a_1} \vee p_2^{\boldsymbol{a_2}} \vee p_3^{\boldsymbol{a_2}}\} \wedge \{p_3^{a_1} \vee p_1^{\boldsymbol{a_2}} \vee p_2^{\boldsymbol{a_2}} \vee p_3^{\boldsymbol{a_2}}\} \wedge \{p_1^{a_1} \vee p_4^{a_1} \vee p_1^{\boldsymbol{a_2}} \vee p_2^{\boldsymbol{a_2}}\}$$

در نتیجه کوچکترین مجموعه از متغیرهای بولی مجموعه $\{p_2^{a_1} \vee p_4^{a_1} \vee p_2^{\boldsymbol{a_2}}\}$ است. بنابراین فضای مشاهدات توسط مجموعه برش‌های انتخابی به صورت شکل ۳-۳ افراز می‌شود.



شکل ۳-۳: افراز فضای مشاهدات توسط برش‌های انتخابی

با توجه به شکل ۳-۳ می‌بینیم که مجموعه افراد جامعه با استفاده از مجموعه برش‌های انتخابی، با همان دقت مجموعه برش‌های اصلی از هم ممیز شده‌اند. بنابراین جدول ۳-۳ بر اساس برش‌های ذکر شده، طی روند زیر به جدول ۳-۴ تبدیل می‌شود.

۱. اگر $a_1(x_i) < C_2^{\boldsymbol{a_1}}$ آنگاه $a_1(x_i) = 0$، اگر $C_2^{a_1} \leq a_1(x_i) < C_4^{a_1}$ آنگاه $a_1(x_i) = 1$، و اگر $a_1(x_i) \geq C_4^{a_1}$ آنگاه $a_1(x_i) = 2$.

۲. اگر $a_2(x_i) < C_2^{a_2}$ آنگاه $a_2(x_i) = 0$، و اگر $a_2(x_i) \geq C_2^{a_2}$ آنگاه $a_2(x_i) = 1$.

جدول ۳-۴: جدول تصمیم گسسته‌سازی شده

S^B	a_1	a_2	d
x_1	0	1	1
x_2	0	0	0
x_3	1	1	0
x_4	1	0	1
x_5	1	1	0
x_6	2	1	1
x_7	1	0	1

۳-۲ درخت رگرسیون ورده‌بندی

درخت رگرسیون و رده‌بندی یکی از روش‌های ناپارامتری در داده‌کاوی است که توسط بریمن[1] و همکارانش (۱۹۸۴) ارائه شد. این روش هر دو قابلیت رگرسیون و رده‌بندی را داراست به این معنی که وقتی متغیر پاسخ از نوع کیفی باشد، از قابلیت رده‌بندی و درصورتی‌که متغیر پاسخ از نوع کمی یا پیوسته باشد از قابلیت رگرسیونی برخوردار است. از آنجایی‌که در پیش‌بینی ریزش مشتری، متغیر پاسخ از نوع کیفی و دارای دو رده ریزش و عدم‌ریزش است، لذا فقط به معرفی قابلیت رده‌بندی آن می‌پردازیم. در اینجا نیز مجموعه متغیرهای توضیحی و متغیر پاسخ (با K رده) را به ترتیب با $\{X_1, X_2, \dots, X_N\}$ و Y نشان می‌دهیم.

درخت رده‌بندی

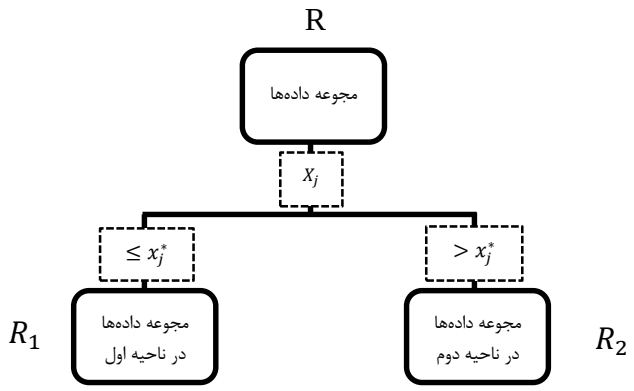
درخت رده‌بندی که آن را درخت تصمیم نیز می‌نامیم، قادر به تولید قانون‌هایی بر اساس مقادیر متغیرهای توضیحی، برای پیش‌بینی رده متغیر پاسخ می‌باشد. این قوانین را می‌توان برای سهولت درک آن، به شکل ساختاری درختی نمایش داد. فرآیند ساخت درخت رده‌بندی، طی دو مرحله‌ی ساخت و هرس درخت انجام می‌شود که در ادامه به معرفی آن‌ها می‌پردازیم.

[1] Breiman

۳-۲-۱ ساخت درخت

فضای به وجود آمده توسط متغیرهای توضیحی، یک فضای N بعدی است و هدف در درخت تصمیم، افراز این فضا به ابرمکعب مستطیل‌های مجاور هم است به گونه‌ای که مشاهدات واقع در هر ابرمکعب دارای بیشترین هم‌گونی از لحاظ رده خاصی از متغیر پاسخ باشند. نحوه افراز فضای متغیرهای توضیحی، سلسله مراتبی و به صورت دودویی است به طوری که در هر مرحله، این فضا (مجموعه داده‌ها) به دو قسمت تقسیم می‌شود. این افراز در امتداد یکی از متغیرهای توضیحی صورت گرفته و بسته به نوع متغیر توضیحی به یکی از دو صورت زیر انجام می‌شود:

۱. اگر متغیر توضیحی از نوع کمی و یا کیفی ترتیبی باشد، آنگاه این افراز به ازای مقدار خاصی در دامنه آن متغیر که بیشترین هم‌گونی را در نواحی جدید ایجاد می‌کند صورت می‌گیرد. فرض کنید فضای متغیرها در راستای متغیر X_j و به ازای مقدار x_j^* (مقداری در دامنه متغیر X_j) افراز شده باشد، آنگاه این افراز را می‌توان به صورت شکل ۳-۴ نشان داد.



شکل ۳-۴: نحوه افراز در درخت تصمیم برای متغیر کمی و یا کیفی ترتیبی

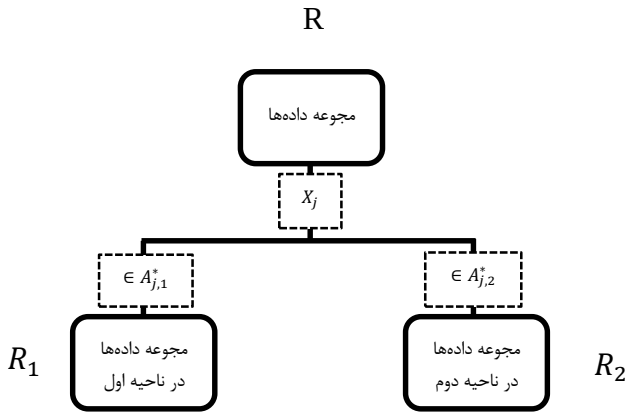
۲. اگر متغیر توضیحی از نوع اسمی باشد، آنگاه مجموعه حالات متغیر توضیحی به دو زیرمجموعه نا-تهی به‌گونه‌ای افراز می‌شود که بیشترین هم‌گونی را در نواحی جدید ایجاد کند. برای مثال، فرض کنید متغیر X_j مجموعه مقادیر $\{a, b, c\}$ را اختیار کند، آنگاه زیرمجموعه‌های ناتهی ممکن برای مجموعه X_j عبارتند از:

$$A_{j,1} = \{a\}, \qquad A_{j,2} = \{b, c\}$$

$A'_{j,1} = \{a, b\}, \quad A'_{j,2} = \{c\}$

$A^*_{j,1} = \{a, c\}, \quad A^*_{j,2} = \{b\}$

حال فرض کنید فضای متغیرها در راستای متغیر X_j و به ازای زیرمجموعه‌های $A^*_{j,1}$ و $A^*_{j,2}$ افراز شده باشد، آنگاه می‌توان این افراز را به صورت شکل ۳-۵ نشان داد.



شکل ۳-۵: نحوه افراز در درخت تصمیم برای متغیر اسمی

در هر دو حالت افراز بیان شده، R ناحیه اصلی شامل کل داده‌ها و R_1 و R_2 به ترتیب نواحی چپ و راست ایجاد شده هستند. هر یک از نواحی موجود در روند تشکیل درخت تصمیم در اصطلاح یک گره[1] نامیده می‌شوند. ناحیه اولیه (R) که شامل کل داده‌ها است را گره مادر و سایر گره‌ها را گره‌های فرزند می‌نامند. به گره‌های فرزندی که خود نیز به دو گره جدید تقسیم شده‌اند گره داخلی و به سایر گره‌های فرزند، گره پایانی گویند. فرض کنید فراوانی نسبی رده k واقع در ناحیه t را با $p(k|t)$ نشان دهیم که:

$$p(k|t) = \frac{1}{N_t} \sum_{x_i \in t} I(y_i = k) \qquad (۱۵-۳)$$

در این رابطه N_t، تعداد مشاهدات درون ناحیه t و I به صورت زیر تعریف می‌شود:

$$I = \begin{cases} 1 & , y_i = k \\ 0 & , y_i \neq k \end{cases} \qquad (۱۶-۳)$$

[1] Node

آنگاه *شاخص جینی*[1] که معرف میزان ناخالصی ناحیه t می‌باشد، عبارت است از:

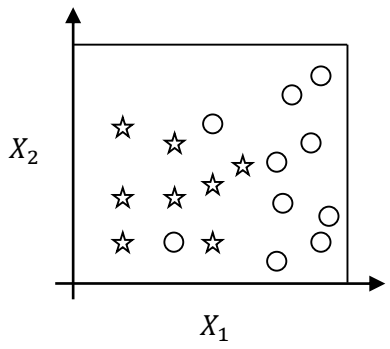
$$i(t) = \sum_{k \neq l} p(k|t)p(l|t) \quad (۱۷-۳)$$

شاخص جینی را می‌توان برای هر یک از سه ناحیه R، R_1 و R_2 محاسبه نمود. تغییرات تابع ناخالصی[2]، ترکیبی از این سه مقدار است که با $\Delta i(t)$ نشان داده شده و به صورت رابطه (۳-۱۸) تعریف می‌شود.

$$\Delta i(t) = i(t_R) - E[i(t_c)] = i(t_R) - P_{R_1} i(t_{R_1}) - P_{R_2} i(t_{R_2}) \quad (۱۸-۳)$$

که در آن t_c نواحی R_1 و R_2 ایجاد شده و P_{R_1} و P_{R_2} نسبتی از کل مشاهدات ناحیه R است که به‌ترتیب در R_1 و R_2 واقع شده‌اند. با تکرار فرآیند فوق برای تمامی متغیرها و به ازای مقادیر مختلف آن‌ها، مقدار $\Delta i(t)$ برای همه افرازهای ممکن محاسبه شده و در نهایت افرازی که مقدار بیشینه $\Delta i(t)$ را ایجاد کند، به عنوان افراز بهینه در نظر گرفته خواهد شد. نواحی جدید ایجاد شده نیز به همین صورت به نواحی کوچکتر تقسیم می‌شوند.

برای فهم بهتر موضوع، مثالی از نحوه افراز فضای متغیرها را برای حالتی که با دو متغیر توضیحی پیوسته سروکار داریم و متغیر پاسخ دو مقدار ۰ و ۱ را اختیار می‌کند، ارائه می‌کنیم. فرض کنید مجموعه داده‌ها به‌صورت شکل ۳-۶ توزیع شده باشند که در آن منظور از دایره و ستاره، به ترتیب مشاهداتی هستند که متغیر پاسخ آن‌ها مقدار صفر و یک را اختیار می‌کنند.

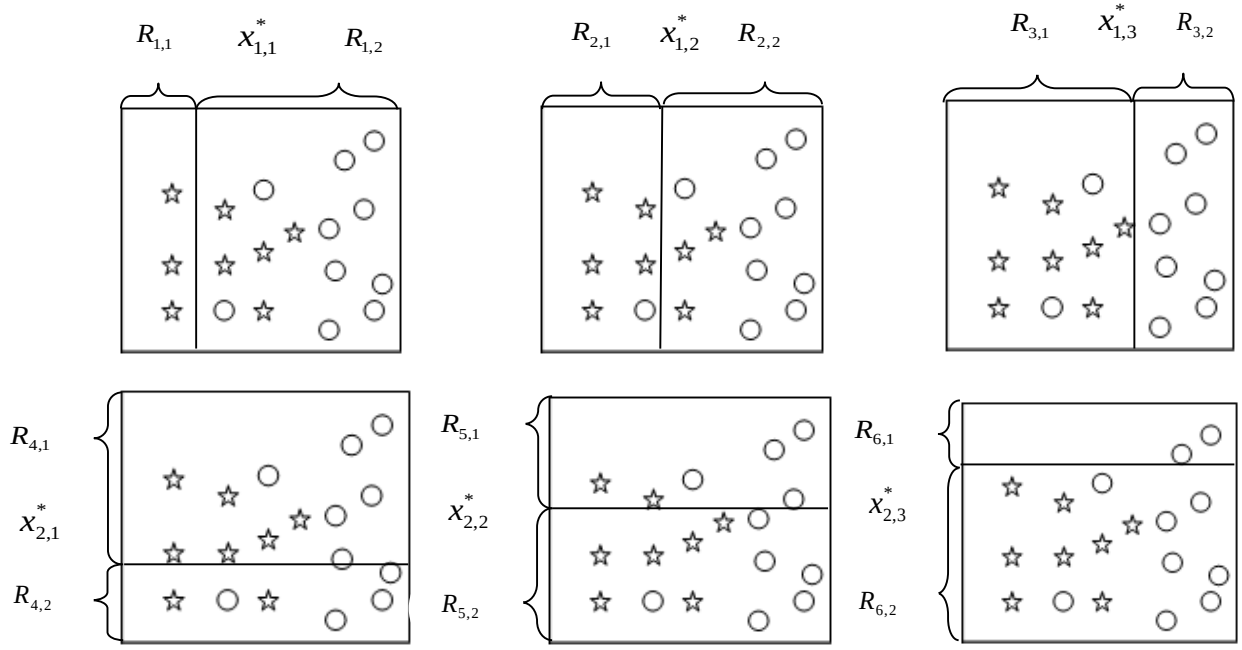


شکل ۳-۶: فضای ایجاد شده توسط دو متغیر توضیحی

[1] Gini Index
[2] Impurity function

همان‌طور که بیان شد، هدف در روش درخت تصمیم، افراز این فضای دوبعدی به مستطیل‌های مجاور هم به‌گونه‌ای است که مشاهدات موجود در هر مستطیل بیشترین هم‌گونی را داشته باشند. از آنجایی‌که برای هر یک از متغیرهای X_1 و X_2، تعداد نقاط افراز زیادی می‌توان در نظر گرفت، لذا در اینجا جهت سهولت، فقط چند نقطه‌ای که به ظاهر برای انجام افراز در راستای هر یک از متغیرها محتمل‌تر هستند در نظر می‌گیریم (شکل ۳-۷).

حال باید معیار ناخالصی (۳-۱۸) را برای تک‌تک حالات فوق محاسبه کرده و از میان آن‌ها، راستا و نقطه‌ای که این معیار به‌ازای آن ماکزیمم می‌شود را انتخاب کنیم. در اینجا معیار ناخالصی مربوط به



شکل ۳-۷: نقاط افراز محتمل در راستای هر یک از متغیرها، به منظور یافتن بهترین افراز

افراز ایجاد شده توسط $x^*_{j,k}$ را با $\Delta_{j,k} i(t)$ نشان می‌دهیم. از آنجایی‌که متغیر پاسخ دارای دو سطح است، روابط (۳-۱۶) و (۳-۱۷) را به صورت زیر بازنویسی می‌کنیم.

$$I = \begin{cases} 1 & , y_k = 1 \\ 0 & , y_k = 0 \end{cases} \qquad i(t) = p(1|t)p(0|t)$$

به عنوان مثال، تابع ناخالصی مربوط به افراز ایجاد شده در راستای متغیر X_1 و مقدار $x^*_{1,1}$ (قاب بالا سمت چپ شکل ۳-۲) به صورت زیر محاسبه می‌شود:

$$p(1|R) = \frac{8}{18}, \quad p(0|R) = \frac{10}{18}, \quad p(1|R_{1,1}) = 1, \quad p(1|R_{1,2}) = \frac{5}{15}$$

$$p(0|R_{1,1}) = 0, \quad p(0|R_{1,2}) = \frac{10}{15}, \quad p(R_{1,1}) = \frac{3}{18}, \quad p(R_{1,2}) = \frac{15}{18}$$

$$i(R_{1,1}) = p(1|R_{1,1})p(0|R_{1,1}) = 1 \times 0 = 0$$

$$i(R_{1,2}) = p(1|R_{1,2})p(0|R_{1,2}) = \frac{5}{15} \times \frac{10}{15} = \frac{2}{9}$$

$$i(R) = p(1|R)p(0|R) = \frac{8}{18} \times \frac{10}{18} = \frac{20}{81}$$

بنابراین:

$$\Delta_{1,1} i(t) = \frac{20}{81} - \left(\frac{3}{18} \times 0\right) - \left(\frac{15}{18} \times \frac{2}{9}\right) = \frac{20}{81} - \frac{15}{81} = 0.061$$

تابع ناخالصی برای بقیه نقاط نیز به همین صورت محاسبه می‌شود که نتایج به قرار زیر است.

$$\Delta_{1,2} i(t) = 0.075, \quad \Delta_{1,3} i(t) = \mathbf{0.158}$$

$$\Delta_{2,1} i(t) = 0.006, \quad \Delta_{2,2} i(t) = 0.006, \quad \Delta_{2,3} i(t) = 0.02$$

با توجه به نتایج به دست آمده، داریم:

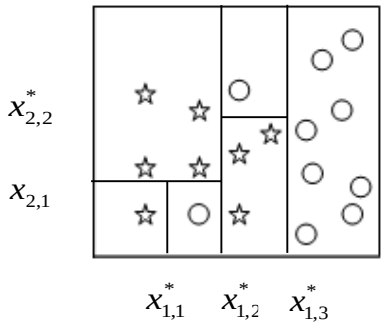
$$max_{\ j=1,2\ ,\ k=1,2,3}\ \Delta_{j,k} \boldsymbol{i(t)} = \Delta_{1,3} i(t)$$

بنابراین اولین افراز در راستای متغیر X_1 و به ازای مقدار $x_{1.3}^*$ ایجاد می‌گردد. نواحی $R_{3,1}$ و $R_{3,2}$ نیز به همین صورت به نواحی کوچکتر افراز می‌شوند.

فرآیند رشد درخت زمانی متوقف می‌شود که حداقل یکی از حالات زیر اتفاق بیافتد:

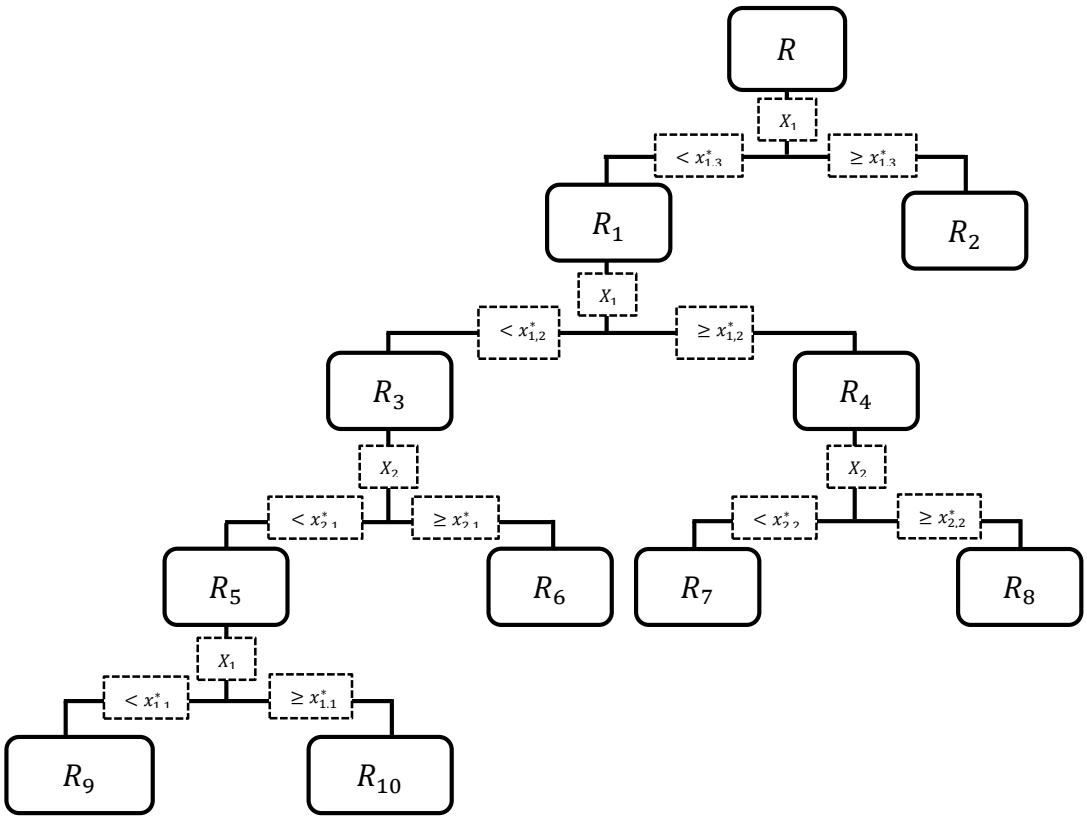
۱. در هر یک از نواحی جدید ایجاد شده، تنها یک مشاهده وجود داشته باشد.
۲. همه مشاهدات داخل یک ناحیه، دارای رده مشابهی از متغیر پاسخ باشند.
۳. یک پیش فرض توسط کاربر در نظر گرفته شود. مثلاً کاربر علاقمند باشد که در هر گره پایانی حداقل ۵ مشاهده وجود داشته باشد.

در نهایت پس از رشد کامل درخت، فضای متغیرها به صورت شکل ۳-۸ افراز می‌شود.



شکل ۳-۸- افراز کامل فضای ایجاد شده توسط دو متغیر توضیحی، در مدل درخت تصمیم

این افراز را می‌توان به صورت یک درخت (شکل ۳-۹) نیز نشان داد که علت نام‌گذاری این روش به نام درخت تصمیم، همین ساختار درخت‌گونه آن است.

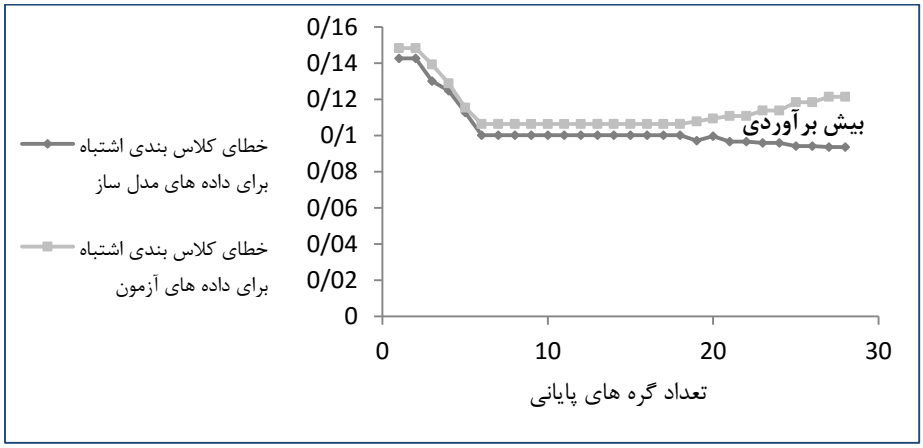


شکل ۳-۹: ساختار درختی مدل درخت تصمیم

رشد بیش از حد درخت ممکن است موجب بروز پدیده بیش‌برآوردی[1] شود. اصطلاح بیش-برآوردی زمانی برای یک مدل به کار می‌رود که اختلاف خطای رده‌بندی مدل برای داده‌های مدل-

[1] Overfitting

ساز[1]، و داده‌های آزمون[2]، بسیار زیاد باشد. لازم به ذکر است که داده‌های مدل‌ساز، داده‌هایی هستند که برای برازش مدل به کار می‌روند و مدل برازش شده، توسط داده‌های آزمون مورد ارزیابی قرار می‌گیرد. شکل ۳-۱۰ نمونه‌ای از پدیده بیش‌برآوردی را نمایش می‌دهد.



شکل ۳-۱۰: شمایی از پدیده بیش‌برآوردی

رخداد این پدیده زمانی که تعداد داده‌های مدل‌ساز کم باشند و یا داده‌های مدل‌ساز حاوی اغتشاش ذاتی زیادی باشند، محتمل‌تر است. برای حل این مشکل دو راه‌حل وجود دارد.

۱. اندازه مناسب درخت را تخمین زده و قبل از اینکه داده‌های مدل‌ساز به طور کامل رده‌بندی شوند از رشد درخت جلوگیری کنیم.

۲. اجازه دهیم درخت به طور کامل رشد کند، سپس آن را هرس کنیم.

از آنجایی‌که تخمین اندازه صحیح درخت کار دشواری است، در عمل استفاده از روش دوم مرسوم‌تر است که در ادامه به آن می‌پردازیم.

۳-۲-۲ هرس‌کردن درخت

پس از این‌که درخت به طور کامل رشد کرد، نوبت به هرس‌کردن آن می‌رسد. فرآیند هرس کردن درخت به این صورت است که در هر گام یکی از شاخه‌های درخت که نبود آن باعث کمترین افزایش

[1] Train data
[2] Test data

در خطای رده‌بندی می‌شود، حذف می‌گردد. این گام تا جایی که حداقل دو گره در درخت باقی بماند ادامه می‌یابد. در نهایت دنباله‌ای از زیردرختان تولید می‌شوند که به ترتیب تعداد گره‌های پایانی کمتری نسبت به زیردرخت قبلی خود دارند. حال باید زیردرخت بهینه از لحاظ میزان خطای رده-بندی و میزان پیچیدگی درخت انتخاب شود. بدین منظور از روشی تحت عنوان مینیمم‌سازی تابع هزینه- پیچیدگی[1] را که به صورت رابطه (۳-۱۹) تعریف شده استفاده می‌شود.

$$R_\alpha(T) = R(T) + \alpha(\tilde{T}) \qquad (۱۹-۳)$$

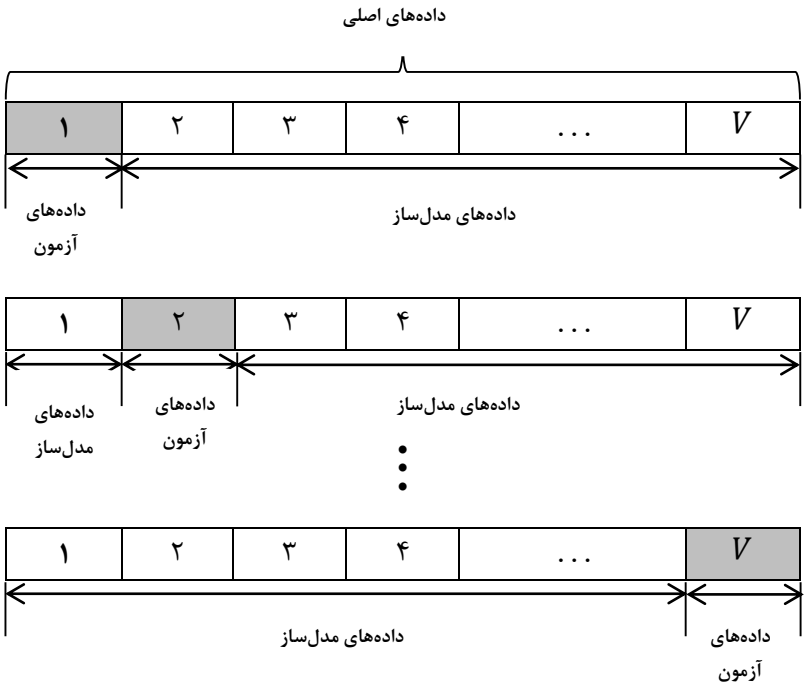
که در آن $R(T)$ خطای رده‌بندی اشتباه مربوط به درخت T، و $\tilde{T}$، مجموع کل گره‌های پایانی در درخت T است. α معیار پیچیدگی می‌باشد که به $\tilde{T}$ وابسته است. تابع هزینه-پیچیدگی به دنبال یک نسبت بهینه بین خطای رده‌بندی و میزان پیچیدگی درخت می‌گردد. هر یک از زیردرختان به ازای مقدار خاصی از α بهینه می‌شوند اما سوال اساسی اینجا مطرح می‌شود که α بهینه، چه α ای است؟ در جواب باید گفت α بهینه، α ای می‌باشد که زیردرخت متناظر با آن موجب بروز پدیده بیش-برآوردی نشود. برای یافتن α بهینه، از روشی به نام اعتبارسنجی متقابل[2] استفاده می‌شود که در زیربخش بعدی آن را معرفی می‌کنیم.

اعتبارسنجی متقابل

اعتبارسنجی متقابل، روشی برای سنجش اعتبار یک مدل است. در این روش، مجموعه داده‌ها به طور تصادفی به V بخش به گونه‌ای تقسیم می‌شوند که داده‌های موجود در همه بخش‌ها، دارای توزیع مشابهی از لحاظ مقادیر متغیر پاسخ باشند. روند کار در روش اعتبارسنجی متقابل به این صورت است که در هر مرحله یکی از زیرمجموعه‌ها به عنوان داده آزمون و $V-1$ زیرمجموعه باقی‌مانده به عنوان داده‌های مدل‌ساز در نظر گرفته می‌شوند. این روند را به طور واضح می‌توان در شکل ۳-۱۱ دید. بنابراین شاهد V مدل مختلف خواهیم بود که هر کدام از آن‌ها توسط مجموعه‌های مستقلی از داده‌ها

[1] Cost-Complexity Function
[2] Cross Validation

مورد آزمون قرار گرفته‌اند. میانگین خطای رده‌بندی اشتباه این V مدل، برآورد بسیار خوبی برای کارایی مدل اصلی (مدلی که توسط مجموعه داده‌ی کامل به دست آید) خواهد بود.



شکل۳-۱۱: اعتبارسنجی متقابل

استفاده از اعتبارسنجی متقابل برای یافتن α بهینه

زمانی که از روش اعتبارسنجی متقابل برای مدل درخت تصمیم استفاده می‌کنیم، شاهد V بار ساخت و هرس درخت خواهیم بود. در هر بار، درخت با استفاده از $V-1$ زیرمجموعه از داده‌های اصلی ساخته و هرس می‌شود و با یک مجموعه باقی‌مانده، عمل آزمون مدل برای هر یک از زیردرختان انجام می‌شود و خطای رده‌بندی محاسبه می‌شود. بنابراین در هر بار فرآیند اعتبارسنجی متقابل، دنباله‌ای از زیردرختان تولید می‌شوند. هر یک از زیردرختان موجود در هر دنباله، به ازای مقدار خاصی از α بهینه می‌باشند. بنابراین به ازای یک α خاص، در هر یک از دنباله‌ها، می‌توان زیردرخت بهینه را یافت. با محاسبه خطای رده‌بندی، زیردرختان انتخابی برای داده‌های آزمون مربوطه و میانگین گرفتن از همه آن‌ها، مقدار خطای رده‌بندی اشتباه برای آن α خاص برآورد می‌شود. با تکرار این عمل برای α های مختلف، خطای رده‌بندی اشتباه به ازای هر مقدار α برآورد می‌شود. سرانجام، α ای که خطای

رده‌بندی متناظر با آن کوچکتر از همه است به عنوان α بهینه انتخاب می‌شود. با مشخص شدن α بهینه و قرار دادن آن در رابطه (۳-۱۹)، بهترین زیردرخت از درخت اصلی انتخاب می‌شود (روگر[1] و همکاران، ۲۰۰۰).

در نهایت با داشتن یک مشاهده جدید و قرار دادن آن در درخت، این مشاهده درون یکی از گره‌های پایانی قرار می‌گیرد. عمل رده‌بندی برای این مشاهده به این صورت است که رده‌ای که بیشترین فراوانی را در گره پایانی مربوطه داراست، به عنوان رده مشاهده مربوطه انتخاب می‌شود.

۳-۲-۳ مزایا و معایب درخت رده‌بندی

درخت تصمیم دارای مزایای بسیار مهم و مفید است که بعضی از آن‌ها را در زیر بیان می‌کنیم.

- برای کار کردن با درخت تصمیم به هیچ فرضیه‌ای در مورد توزیع داده‌ها نیاز نیست.
- متغیرهای توضیحی می‌توانند مخلوطی از متغیرهای گسسته و پیوسته باشند.
- زمانی که بُعد فضای متغیرهای توضیحی بالا است و حجم داده‌ها بسیار زیاد است، درخت تصمیم بسیار خوب عمل می‌کند.
- یک روش قدرتمند در مواقع برخورد با مقادیر گمشده است.
- داده‌های پرت، همبستگی بین متغیرها و ناهمگنی واریانس داده‌ها روی این روش تأثیر اندکی دارد.
- نسبت به تبدیل‌های یکنواخت روی متغیرهای توضیحی، پایا است.
- وجود متغیرهای غیرضروری، موجب کاهش دقت مدل نمی‌شود.

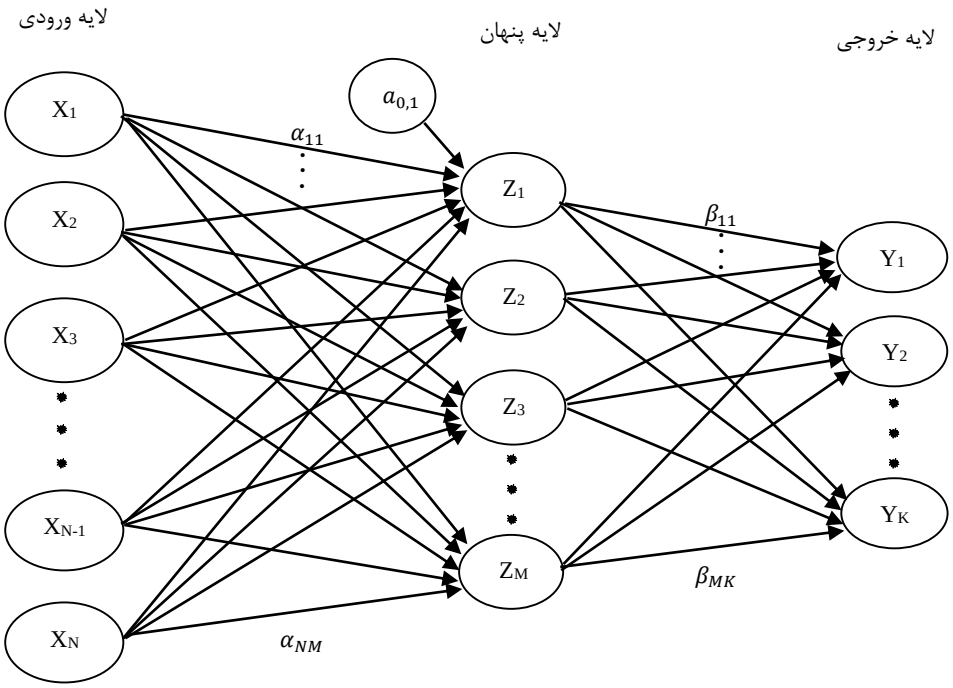
از جمله معایب روش درخت رده‌بندی عبارتند از:

- تغییرات ناچیز در داده‌های مدل‌ساز، موجب تغییر اساسی در درخت تصمیم قبلی خواهد شد.
- با تعداد متغیر اندک، ممکن است نتایج غیرقابل قبولی ارائه دهد.

[1] Roger

۳-۳ شبکه‌های عصبی مصنوعی

شبکه‌های عصبی مصنوعی که از این پس به اختصار آن را شبکه‌های عصبی می‌نامیم، رده‌ای از روش-های یادگیری ماشین است که در زمینه‌های آمار و هوش مصنوعی رشد و توسعه یافته است و دارای قابلیت رگرسیونی و رده‌بندی می‌باشد. اساس این روش، از شبکه‌های عصبی پیچیده موجود در مغز انسان، الگوبرداری شده است. در مدل شبکه‌های عصبی، متغیرهای توضیحی و متغیر پاسخ می‌توانند از نوع کمی یا کیفی باشند و همچنین متغیر پاسخ، رابطه‌ای غیرخطی و غیرمستقیم با متغیرهای توضیحی دارد (هستی[1] و همکاران، ۲۰۰۹). چارچوب شبکه‌های عصبی از چند لایه به نام‌های لایه ورودی، لایه(های) پنهان و لایه خروجی تشکیل شده است. شکل ۳-۱۲، ساختار شماتیکی یک شبکه عصبی را با یک لایه پنهان، N متغیر توضیحی و یک متغیر پاسخ با K سطح را نشان می‌دهد.



شکل ۳-۱۲: شمایی از ساختار شبکه عصبی با یک لایه پنهان

که در آن Z_m ها و Y_k ها با استفاده از توابع خطی یا غیرخطی از گره‌های لایه‌های قبل، محاسبه می-شوند.

[1] Hastie

هر یک از لایه‌ها شامل اجزائی هستند که در اصطلاح نورون یا گره[1] نامیده می‌شوند. مجموعه‌های $\{X_1, X_2, \dots, X_N\}$، $\{Z_1, Z_2, \dots, Z_M\}$، $\{Y_1, Y_2, \dots, Y_K\}$ به ترتیب مجموعه گره‌های ورودی، لایه پنهان، و لایه خروجی هستند. هر یک از اتصالات موجود بین دو گره با وزن (ضریب) خاصی صورت می‌پذیرد. به عنوان مثال در شکل ۳-۱۲، گره X_1 در لایه ورودی با وزن $\alpha_{1,1}$ به گره Z_1 متصل می‌شود. همچنین ممکن است در یک شبکه عصبی علاوه بر مقادیر گره‌های موجود در لایه‌های قبل که به یک گره خاص در لایه‌های بعد وارد می‌شوند، یک مقدار نیز به عنوان مقدار اریبی به آن گره خاص وارد شود. مثلاً در شکل ۳-۱۲، علاوه بر گره‌های $X_1, X_2, \dots, X_N$ که به گره Z_1 متصل شده‌اند، یک مقدار ثابتی با نام α_{01} نیز به Z_1 وارد شده است. بنابراین به طور کلی می‌توان گفت α_{0m} و β_{0k} به ترتیب مقادیر اریبی هستند که به گره‌های Z_m و Y_k وارد می‌شوند و α_{im} و β_{mk} به ترتیب وزن اتصال-دهنده‌هایی است که گره X_i را به Z_m و گره Z_m را به گره Y_k متصل می‌کنند.

تعداد گره‌ها در لایه ورودی و خروجی، بسته به نوع متغیرهای توضیحی و پاسخ تعیین می‌شوند. در لایه ورودی، به ازای هر متغیر توضیحی از نوع کمی یک گره، و به ازای هر متغیر توضیحی از نوع کیفی دارای l کلاس، l گره در نظر گرفته می‌شود. در لایه خروجی برای حالت رگرسیون، زمانی که K متغیر پاسخ داشته باشیم، K گره و برای حالت رده‌بندی با K رده، K گره در نظر گرفته می‌شود. تعداد گره‌ها در لایه‌های پنهان به طور دلخواه انتخاب می‌شود. از آنجایی‌که متغیر پاسخ در پیش‌بینی ریزش مشتری از نوع کیفی بوده و دارای ۲ رده است، در اینجا به معرفی قابلیت رده‌بندی شبکه عصبی پرداخته و از آن استفاده می‌کنیم.

۳-۳-۱ انواع شبکه‌های عصبی

شبکه عصبی از لحاظ نحوه‌ی اتصال گره‌ها به دو نوع شبکه‌های عصبی *پیش‌خورانی*[2] و *پس‌خورانی*[1] تقسیم می‌شوند.

---

1 Node
2 Feed-Forward

۱. شبکه‌های عصبی پیش‌خورانی: در این شبکه، گره‌های هر لایه تنها می‌توانند به گره‌های لایه‌های بعد متصل شوند. به عبارت دیگر، از خروجی هیچ گره‌ای نمی‌توان به خودش یا به گره‌های موجود در همان لایه و گره‌های لایه‌های قبل اتصالی برقرار کرد (شکل ۳-۱۲).

۲. شبکه‌های عصبی پس‌خورانی: در این شبکه، گره‌های هر لایه می‌توانند علاوه بر اتصال به گره‌های لایه‌های بعد، به گره‌های همان لایه یا به گره‌های لایه قبل و به خودشان متصل شوند.

در ادامه با شبکه‌های عصبی پیش‌خورانی کار خواهیم کرد.

۳-۳-۱ توابع ترکیب[۲] و فعال‌سازی[۳]

هر کدام از گره‌های موجود در یک ساختار شبکه عصبی مقداری را اختیار می‌کنند. مقادیر گره‌های لایه‌ی ورودی، همان مقادیر متغیرهای توضیحی می‌باشند. اما مقادیر گره‌های لایه پنهان و لایه خروجی، طی دو مرحله محاسبه می‌شوند. در مرحله اول مقادیر ورودی به هر گره، توسط توابع خاصی با هم ترکیب شده و به یک عدد تبدیل می‌شوند. به این توابع، در اصطلاح توابع ترکیب گویند. دو نوع عمده از این توابع که بیشتر مورد استفاده قرار می‌گیرند، عبارتند از:

۱. توابع ترکیب خطی: ترکیب خطی از مقادیر و وزن‌های ورودی به یک گره را به دست آورده و سپس یک مقدار ثابت (اریبی) به آن اضافه می‌کند.

۲. توابع ترکیب رادیال[۴]: توان دوم فاصله اقلیدسی بین بردار مقادیر و بردار وزن‌های ورودی به یک گره را محاسبه کرده و سپس توان دوم مقدار اریبی را در آن ضرب می‌کند.

از جمله این توابع ترکیب، می‌توان به توابع موجود در جدول ۳-۵ اشاره کرد که در اینجا به منظور محاسبه مقدار Z_m تعریف شده‌اند.

---

[1] Feedback
[2] Combination Function
[3] Activation Function
[4] Radial

جدول ۳-۵: توابع ترکیب

	توابع ترکیب	Z_m
توابع خطی	Linear	$\alpha_{0m} + \sum_i \alpha_{im} X_i$
	EQSlopes	$\alpha_{0m} + \sum_i \alpha_i X_i$
توابع رادیال	EHRadial	$-\alpha_{0m}^2 \sum_i (\alpha_{im} - X_i)^2$
	EVRadial	$s\ log(\alpha_{0m}) - \alpha_{0m}^2 \sum_i (\alpha_{im} - X_i)^2$
	EQRadial	$-\alpha_0^2 \sum_i (\alpha_{im} - X_i)^2$

در اینجا $\alpha_0 = \sum_{m=1}^{M} \alpha_{0m}$ مجموع اریبی‌هایی است که به گره‌های لایه‌ی پنهان وارد شده، s تعداد ورودی‌های ممکن به یک گره در لایه پنهان و $\alpha_i = \sum_{m=1}^{M} \alpha_{im}$، مجموع وزن‌هایی که i امین ورودی را به گره‌های لایه‌ی پنهان متصل می‌کند. به همین صورت تابع ترکیب را برای گره‌های موجود در لایه خروجی نیز می‌توان محاسبه نمود.

در مرحله بعد مقادیر به دست آمده توسط توابع ترکیب، با استفاده از توابعی دیگر به مقادیر جدید، تبدیل می‌شوند. این توابع، در اصطلاح توابع فعال‌سازی نامیده می‌شوند. توابع فعال‌سازی انواع مختلفی دارند که فهرست آنها در جدول ۳-۶ آمده است.

جدول ۳-۶: توابع فعال‌سازی

تابع	نام تابع فعال‌سازی	تابع	نام تابع فعال‌سازی
x	همانی	$\frac{1}{1+exp(-x)}$	لجستیک
$1/x$	معکوس	$Tanh(x) = 1 - \frac{1}{1+exp(2x)}$	تانژانت هایپربولیک
$Sin(x)$	سینوس	$\frac{exp(x)}{\sum_{all\ neouon\ in\ same\ layer} exp(x)}$	سافتمکس
$Cos(x)$	کسینوس	$exp(-x^2)$	گوسین
$\frac{2}{\pi}\arctan(x)$	آرک تانژانت	$exp(x)$	نمائی

که در آن x، مقدار حاصل از تابع تبدیل در گره مربوطه است. لازم به ذکر است که توابع S- شکل مانند لجستیک، آرک تانژانت و تانژانت هایپربولیک که به توابع سیگموئید[1] موسومند و همچنین تابع سافتمکس[2] که آن را تابع لجستیک چندگانه می‌نامند، از پرکاربردترین توابع فعال‌سازی می‌باشند. مطالعات نشان داده است که تابع سافتمکس برای مسائل رده‌بندی، و تابع همانی برای مسائل رگرسیون در لایه‌ی خروجی مناسب‌تر از سایر توابع فعال‌سازی هستند.

در حالت کلی، اگر توابع ترکیب استفاده شده در لایه‌های پنهان را با g_H و g_O و توابع فعال‌سازی استفاده شده در لایه خروجی را با σ و ρ_k نشان دهیم، آنگاه مقادیر Z_m و Y_k به صورت روابط (۳-۲۰) تا (۳-۲۲) محاسبه می‌شوند.

$$Z_m = \sigma(\mathrm{g}_H(\alpha_{0m}, \dots, \alpha_{Nm}, X_1, \dots, X_N)); \qquad m = 1,2, \dots, M \qquad (۳-۲۰)$$

$$T_k = \mathrm{g}_O(\beta_{0k}, \dots, \beta_{Mk}, Z_1, \dots, Z_M); \qquad k = 1,2, \dots, K \qquad (۳-۲۱)$$

$$f_k(\boldsymbol{X}) = Y_k = \rho_k(T); \qquad k = 1,2, \dots, K \qquad (۳-۲۲)$$

که در آن $T = \{T_1, T_2, \dots, T_K\}$ است.

تعداد متغیرهای ورودی، نوع تابع ترکیب و تابع فعال‌سازی و نحوه اتصال گره‌ها، ساختار کلی یک شبکه عصبی را مشخص می‌کنند. یکی از مهم‌ترین ساختارها، ساختار پرسپترون چندلایه[3] نام دارد که دارای خصوصیات زیر است:

- ❖ از نوع شبکه عصبی پیش‌خورانی است.
- ❖ به تعداد دلخواه می‌تواند متغیرهای ورودی داشته باشد.
- ❖ دارای یک یا چند لایه پنهان با هر تعداد گره دلخواه است.
- ❖ در لایه‌های پنهان و خروجی از تابع ترکیب خطی استفاده می‌کند.

---

[1] Sigmoid
[2] Softmax
[3] Multilayer Perceptron

- ❖ از تابع فعال‌سازی S- شکل در لایه پنهان استفاده می‌کند.
- ❖ دارای هر تعداد خروجی با هر تابع فعال‌سازی دلخواه است..

به عنوان مثال، اگر شبکه عصبی پیش‌خورانی شامل یک لایه پنهان، توابع ترکیب در لایه‌های پنهان و خروجی آن خطی و توابع فعال‌سازی در لایه‌های پنهان و خروجی به ترتیب تانژانت هایپربولیک و سافتمکس در نظر گرفته شوند، آنگاه شبکه عصبی ایجاد شده تحت این خصوصیات، یک شبکه عصبی با ساختار پرسپترون چندلایه است که با فرض:

$$\boldsymbol{\alpha}_m = \begin{pmatrix} \alpha_{1m} \\ \alpha_{2m} \\ \vdots \\ \alpha_{Nm} \end{pmatrix} \quad , \quad \boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} \quad , \quad \boldsymbol{\beta}_k = \begin{pmatrix} \beta_{1k} \\ \beta_{2k} \\ \vdots \\ \beta_{mk} \end{pmatrix} \quad , \quad \boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}$$

روابط (۳-۲۰) تا (۳-۲۲) به فرم زیر خواهند بود.

$$Z_m = Tanh(\alpha_{0m} + \boldsymbol{\alpha_m^T X}); \qquad m = 1,2,\dots,M$$

$$T_k = \beta_{0k} + \boldsymbol{\beta_k^T Z}; \qquad k = 1,2,\dots,K$$

$$f_k(\boldsymbol{X}) = Y_k = \frac{exp(T_k)}{\sum_{l=1}^{K} exp(T_l)}; \qquad k = 1,2,\dots,K$$

۳-۳-۳ برآورد ضرایب و عمل رده‌بندی

برای محاسبه $f_k(\boldsymbol{X})$ نیاز به برآورد ضرایب مجهول مدل می‌باشد. برای برآورد ضرایب از مینیمم-سازی تابع خطا استفاده می‌شود. توابع خطای مختلفی بدین منظور مورد استفاده قرار می‌گیرند که یکی از مرسوم‌ترین آن‌ها مجموع مربعات خطا[1] است. فرض کنید داده‌های مدل‌ساز شامل مجموعه n تائی از مشاهدات به صورت $(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ باشند که در آن $\boldsymbol{x_i^T} = (x_{i1}, x_{i2}, \dots, x_{iN})$ و y_i کلاس متغیر پاسخ مربوط به مشاهده i ام است، آنگاه تابع مجموع مربعات خطا به صورت رابطه (۳-۲۳) تعریف می‌شود.

[1] Sum of square error

$$R(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\left(y_{ik} - f_k(\boldsymbol{x_i})\right)^2 = \sum_{i=1}^{n} R_i \qquad (۲۳-۳)$$

که در آن θ، ضرایب مجهول مدل و y_{ik} به صورت زیر تعریف می‌شود.

$$y_{ik} = \begin{cases} 1 & ; \quad y_i = k \\ 0 & ; \quad y_i \neq k \end{cases}$$

روشی که به منظور مینیمم‌سازی تابع خطا مورد استفاده قرار می‌گیرد، روش پس‌انتشار[1] است که الگوریتم آن از روش گرادیان کاهشی[2] جهت برآورد ضرایب مجهول استفاده می‌کند. با استفاده از این الگوریتم، برداری از ضرایب مجهول که تابع خطا را مینیمم می‌کنند، شناسایی می‌شوند. فرض کنید $z_{mi} = \sigma(\alpha_{0m} + \alpha_m^T x_i)$ و $\boldsymbol{z_i^T} = (z_{1i}, z_{2i}, \dots, z_{Mi})$ باشد، آنگاه الگوریتم پس‌انتشار برای شبکه عصبی با یک لایه پنهان به صورت زیر عمل می‌کند.

ابتدا از تابع خطا نسبت به ضرایب مشتق گرفته می‌شود.

$$\frac{\partial R_i}{\partial \beta_{mk}} = -2\left(y_{ik} - f_k(\boldsymbol{x}_i)\right)\rho_k'\left(\boldsymbol{\beta_k^T z_i}\right)z_{mi} = \delta_{ki} z_{mi} \qquad (۲۴-۳)$$

$$\frac{\partial R_i}{\partial \alpha_{lm}} = -\sum_{k=1}^{K} 2\left(y_{ik} - f_k(x_i)\right)^2 \rho_k'\left(\boldsymbol{\beta_k^T z_i}\right)z_{mi}\beta_{mk}\sigma'(\boldsymbol{\alpha_m^T x_i})x_{il} = s_{mi}x_{il}$$

$$\delta_{ki} = -2\left(y_{ik} - f_k(x_i)\right)\rho_k'\left(\boldsymbol{\beta_k^T z_i}\right) \qquad (۲۵-۳)$$

$$s_{mi} = -\sum_{k=1}^{K} 2\left(y_{ik} - f_k(x_i)\right)^2 \rho_k'\left(\boldsymbol{\beta_k^T z_i}\right)z_{mi}\beta_{mk}\sigma'(\boldsymbol{\alpha_m^T x_i}) \qquad (۲۶-۳)$$

که در آن‌ها δ_{ki}، خطای مدل فعلی در گره‌های لایه خروجی و s_{mi}، خطای مدل فعلی در گره‌های لایه پنهان می‌باشند. با توجه به روابط (۳-۲۵) و (۳-۲۶) می‌توان به رابطه مفیدی به صورت رابطه (۳-۲۷) دست یافت که در اصطلاح برابری پس‌انتشار نامیده می‌شود.

$$s_{mi} = \sigma'(\boldsymbol{\alpha_m^T x_i})\sum_{k=1}^{K}\beta_{mk}\delta_{ki} \qquad (۲۷-۳)$$

---

[1] Back-propagation
[2] Gradient descent

اگر مشتقات (۳-۲۴) وجود داشته باشند، آنگاه مقادیر ضرایب با کمک روش نیوتن-رافسون به صورت روابط (۳-۲۸) بروزرسانی می‌شوند.

$$\beta_{mk}^{(r+1)} = \beta_{mk}^{(r)} - \gamma_r \sum_{i=1}^{n} \frac{\partial R_i}{\partial \beta_{mk}^{(r)}}$$

$$\alpha_{lm}^{(r+1)} = \alpha_{lm}^{(r)} - \gamma_r \sum_{i=1}^{n} \frac{\partial R_i}{\partial \alpha_{lm}^{(r)}} \quad (۲۸-۳)$$

که در آن‌ها γ_r نرخ یادگیری است. مقادیر اولیه ضرایب معمولاً به طور تصادفی نزدیک به صفر انتخاب می‌شوند. بروزرسانی در (۳-۲۸) یک نوع یادگیری دسته‌ای[1] می‌باشد. یادگیری دسته‌ای به یادگیری گفته می‌شود که در آن بروزرسانی ضرایب، توسط همه داده‌های مدل‌ساز انجام می‌شود. در یادگیری دسته‌ای، نرخ یادگیری معمولاً مقدار ثابتی است که می‌تواند طی یک روند جستجو به منظور مینیمم-سازی تابع خطا، بهینه شود.

با استفاده از برابری (۳-۲۷)، بروزرسانی ضرائب در روابط (۳-۲۸) بر اساس یک الگوریتم دو مسیری محاسبه می‌شوند. در مسیر پیش‌رو، با استفاده از وزن‌های موجود مقادیر پیش‌بینی $\hat{f}_k(\boldsymbol{x_i})$ طبق رابطه (۳-۲۲) محاسبه می‌شوند. با داشتن $\hat{f}_k(\boldsymbol{x_i})$، در مسیر پس‌رو مقادیر خطای δ_{ki} با استفاده از رابطه (۳-۲۵) محاسبه شده و سپس با استفاده از برابری (۳-۲۷) این خطاها برای محاسبه s_{mi} به عقب انتشار داده می‌شوند. هر دو مجموعه از خطاها برای محاسبه گرادیان‌ها در روابط (۳-۲۴) مورد استفاده قرار می‌گیرند. با محاسبه‌ی گرادیان‌ها، عمل بروزرسانی در روابط (۳-۲۸) انجام می‌شود. این روش دو مسیره را در اصطلاح روش پس‌انتشار می‌نامند.

بروزرسانی ضرایب زمانی که تغییرات در خطای رده‌بندی اشتباه مدل بسیار کم باشد، متوقف می‌شود. در نهایت پس از برآورد ضرایب مجهول، مقادیر $\hat{f}_k(\boldsymbol{x})$ برای $k = 1,2,\dots,K$ برآورد خواهند شد. لازم به ذکر است در حالت بیش از یک لایه پنهان، برآورد ضرایب نیز تقریباً به همین منوال با کمی محاسبات پیچیده‌تر انجام می‌شود. پیش‌بینی رده متغیر پاسخ مربوط به مشاهده x به این

[1] Batch learning

صورت است که آن رده‌ای از متغیر پاسخ که مقدار $\hat{f}_k(\boldsymbol{x})$ مربوط به آن بزرگتر باشد به عنوان رده پیش‌بینی در نظر گرفته می‌شود. مثلاً در پیش‌بینی ریزش مشتری با توجه به اینکه تعداد رده‌ها برابر با ۲ ($K=۲$) است و با فرض اینکه "$k=۱$"بیانگر متغیر پاسخ با رده یک و "$k=۰$" بیانگر متغیر پاسخ با رده صفر باشد، آنگاه اگر $\hat{f}_1(\boldsymbol{x}) > \hat{f}_2(\boldsymbol{x})$، رده متغیر پاسخ مربوط به مشاهده $\boldsymbol{x}$، یک پیش-بینی شده و در غیر این صورت صفر پیش‌بینی می‌شود.

۳-۳-۴ بیش‌برآوردی

در شبکه عصبی نیز همانند سایر مدل‌های داده‌کاوی ممکن است پدیده بیش‌برآوردی اتفاق بیافتد. از جمله دلائل بروز این پدیده می‌توان به وجود تعداد زیاد ضرایب در مدل اشاره کرد. بدین منظور نیاز به یک قاعده به منظور توقف بروزرسانی ضرایب داریم. از این‌رو روشی تحت عنوان تباهی وزن‌ها[1] وجود دارد که مشابه معیار هزینه-پیچیدگی در درخت رده‌بندی عمل کرده و با اضافه کردن یک فاکتور تاوان[2] به تابع خطا، برآورد ضرایب را تنظیم می‌کند. به عبارت دیگر عمل زیر انجام می‌شود.

$$R_\lambda(\theta) = R(\theta) + \lambda J(\theta)$$

که در آن $\lambda > 0$ پارامتر میزان‌سازی[3] بوده و $J(\theta)$ عبارت است از:

$$J(\theta) = \sum_{l,m} \alpha_{lm}^2 \sum_{m,k} \beta_{mk}^2$$

هرچه λ بزرگتر باشد، وزن‌ها کوچکتر شده و به سمت صفر میل می‌کنند. برای انتخاب λ از روش اعتبارسنجی متقابل استفاده می‌شود. سرانجام برآوردهایی که تابع $R_\lambda(\theta)$ را مینیمم کنند به عنوان بهترین برآورد انتخاب می‌شوند.

[1] Weight Decay
[2] Penalty
[3] Tuning Parameter

۳-۳-۵ استانداردسازی متغیرهای توضیحی

از آنجایی‌که مقیاس متغیرهای توضیحی، اثر وزن‌ها را در یک لایه مشترک مشخص می‌کند، از این‌رو تأثیر زیادی در کیفیت نتایج نهایی خواهد داشت. لذا به منظور پیش‌گیری از بروز چنین مشکلی، تمامی متغیرهای توضیحی را استاندارد می‌کنند. با انجام این عمل تمامی متغیرها در فرآیند تنظیم ضرایب، نقش یکسانی را بازی می‌کنند و به کاربر اجازه می‌دهد دامنه معنی‌داری را برای انتخاب تصادفی مقادیر اولیه انتخاب کند.

۳-۳-۶ تعداد لایه‌ها و گره‌های پنهان

با داشتن تعداد گره‌های کم، ممکن است مدل انعطاف‌پذیری کافی به منظور نمایش خواص غیرخطی در داده‌ها را نداشته باشد. از طرفی تعداد گره‌های زیاد زمانی که از روش تباهی وزن‌ها به منظور تنظیم وزن‌ها استفاده می‌شود، ممکن است موجب نزدیک شدن زیاد وزن‌ها به صفر شود. بنابراین انتخاب تعداد گره‌ها و لایه‌های پنهان می‌تواند تأثیر زیادی بر میزان دقت یک مدل داشته باشد. بهترین روشی که به منظور انتخاب بهترین تعداد گره و لایه پنهان وجود دارد، استفاده از روش سعی و خطا است.

۳-۴ رگرسیون لجستیک

رگرسیون لجستیک، عضو خانواده‌ای از مدل‌های آماری به نام مدل‌های خطی تعمیم‌یافته[1] است. این مدل‌ها، تعمیمی از مدل‌های خطی هستند که در آن‌ها شرط نرمال بودن توزیع خطا وجود ندارد و برای مدل‌بندی داده‌هایی که توزیع احتمال آن‌ها متعلق به خانواده توزیع‌های نمایی مانند پواسن، دوجمله‌ای و چندجمله‌ای است مدل‌های مناسبی محسوب می‌شوند. این مدل‌ها، از پذیره ثابت بودن واریانس که برای انجام آزمون‌های فرضیه در مدل‌های خطی لازم است، مبرا هستند. رگرسیون خطی

[1] Generalized Linear Models

ساده، مدل آنالیز واریانس[1]، مدل آنالیز کواریانس[2] و رگرسیون لگ خطی[3] از جمله مدل‌هایی هستند که به خانواده مدل‌های خطی تعمیم‌یافته تعلق دارند (مک‌کولاق[4] و همکاران، ۱۹۸۹).

رگرسیون لجستیک تنها دارای قابلیت رده‌بندی بوده و متغیرهای توضیحی در آن می‌توانند از نوع طبقه‌ای، پیوسته و یا مخلوطی از آن‌ها باشند. متغیر پاسخ در رگرسیون لجستیک باید از نوع کیفی (اسمی و ترتیبی) باشد و بسته به نوع متغیر کیفی و تعداد رده‌ها به سه دسته تقسیم می شود:

۱) رگرسیون لجستیک دودویی[5]: وقتی متغیر پاسخ دارای دو رده (دو سطح) باشد، از قبیل بله/خیر، سلامتی/ بیماری و غیره مورد استفاده قرار می‌گیرد.
۲) رگرسیون لجستیک ترتیبی[6]: وقتی مورد استفاده قرار می‌گیرد که متغیر پاسخ از نوع ترتیبی باشد.
۳) رگرسیون لجستیک اسمی[7]: وقتی مورد استفاده قرار می‌گیرد که متغیر پاسخ از نوع اسمی باشد.

از آنجایی‌که متغیر پاسخ در پیش‌بینی ریزش مشتری به صورت دودوئی (ریزش (۱) و عدم‌ریزش (۰)) است، در ادامه به معرفی رگرسیون لجستیک دودویی می‌پردازیم.

۳-۴-۱ رگرسیون لجستیک دودویی

در رگرسیون لجستیک دودویی، متغیر پاسخ مقدار ۱ را با احتمال p و مقدار صفر را با احتمال $q = 1 - p$ اختیار می‌کند. در این مدل زمانی که N متغیر توضیحی وجود داشته باشد، رابطه میان

[1] ANOVA
[2] ANCOVA
[3] Loglinear
[4] McCullagh
[5] Binary Logistic Regression
[6] Ordinal Logistic Regression
[7] Nominal Logistic Regression

متغیرهای توضیحی و متغیر پاسخ توسط تابع لجستیک بیان شده و به صورت رابطه (۳-۲۹) نشان داده می‌شود (هستی و همکاران).

$$P(Y=1|\boldsymbol{X}=\boldsymbol{x})=\frac{\mathrm{e}^{(\boldsymbol{x}'\,.\,\boldsymbol{\gamma}^{\boldsymbol{T}})}}{1+\mathrm{e}^{(\boldsymbol{x}'\,.\,\boldsymbol{\gamma}^{\boldsymbol{T}})}}=p(\boldsymbol{x}',\boldsymbol{\gamma}) \qquad (۳\text{-}۲۹)$$

که در آن $\boldsymbol{x}'=(1,x_1,x_2,\dots,x_N)$ و $\boldsymbol{\gamma}=(\beta_0,\beta_1,\beta_2,\dots,\beta_N)$.

بنابراین $q(\boldsymbol{x}',\boldsymbol{\gamma})$ نیز به صورت زیر محاسبه می‌شود:

$$q(\boldsymbol{x}',\boldsymbol{\gamma})=P(Y=0|\boldsymbol{X}=\boldsymbol{x})=1-p(\boldsymbol{x}',\boldsymbol{\gamma})=\frac{1}{1+\mathrm{e}^{(\boldsymbol{x}'\,.\,\boldsymbol{\gamma}^{\boldsymbol{T}})}}$$

هدف در رگرسیون لجستیک، پیش‌بینی رده متغیر پاسخ یک مشاهده $\boldsymbol{x}$ بر اساس احتمال $p(\boldsymbol{x}',\boldsymbol{\gamma})$ است، به طوری که اگر $p(\boldsymbol{x}',\boldsymbol{\gamma})>\frac{1}{2}$ باشد، رده متغیر پاسخ مربوط به مشاهده $\boldsymbol{x}$، یک و در غیر این صورت صفر پیش‌بینی می‌شود. به منظور انجام عمل تصمیم‌گیری، از تابع تبدیل لجیت[1] نیز می‌توان استفاده کرد:

$$\mathrm{Logit}\big(p(\boldsymbol{x}',\boldsymbol{\gamma})\big)=\ln\left[\frac{p(\boldsymbol{x}',\boldsymbol{\gamma})}{q(\boldsymbol{x}',\boldsymbol{\gamma})}\right]=\boldsymbol{x}'\,.\,\boldsymbol{\gamma}^{\boldsymbol{T}}$$

که در آن به $\left[\frac{p(\boldsymbol{x}',\boldsymbol{\gamma})}{q(\boldsymbol{x}',\boldsymbol{\gamma})}\right]$ نسبت بخت گفته می‌شود. اگر نسبت بخت بیشتر از ۱ باشد، رده متغیر پاسخ یک و در غیر این صورت صفر پیش‌بینی می‌شود.

برای محاسبه $p(\boldsymbol{x}',\boldsymbol{\gamma})$ نیاز به برآورد پارامترهای مجهول مدل خواهیم داشت. با توجه به اینکه $(Y|\boldsymbol{x_i})$ دارای توزیع برنولی با پارامتر $p(\boldsymbol{x_i'},\boldsymbol{\gamma})$ است، لذا از روش درست‌نمایی ماکزیمم می‌توان برای برآورد پارمترهای مجهول مدل استفاده نمود. به عبارت دیگر:

$$P(Y|\boldsymbol{x_i})=p(\boldsymbol{x_i'},\boldsymbol{\gamma})^{Y_i}\big(1-p(\boldsymbol{x_i'},\boldsymbol{\gamma})\big)^{(1-Y_i)}$$

بنابراین تابع درست‌نمایی عبارت است از:

$$L(\boldsymbol{\gamma})=\prod_{i=1}^{n}p(\boldsymbol{x_i},\boldsymbol{\gamma})^{Y_i}\big(1-p(\boldsymbol{x_i},\boldsymbol{\gamma})\big)^{(1-Y_i)}$$

[1] Logit

با لگاریتم گرفتن از طرفین تابع درست‌نمایی خواهیم داشت:

$$\begin{aligned} ln\, L &= \sum_{i=1}^{n}\{y_i ln\big(p(\boldsymbol{x_i'}, \boldsymbol{\gamma})\big) + (1-y_i) ln\big(1-p(\boldsymbol{x_i'}, \boldsymbol{\gamma})\big)\} \\ &= \sum_{i=1}^{n} ln\big(1-p(\boldsymbol{x_i'}, \boldsymbol{\gamma})\big) + \sum_{i=1}^{n} y_i\, ln\left(\frac{p(\boldsymbol{x_i'}, \boldsymbol{\gamma})}{1-p(\boldsymbol{x_i'}, \boldsymbol{\gamma})}\right) \\ &= \sum_{i=1}^{n} ln\big(1-p(\boldsymbol{x_i'}, \boldsymbol{\gamma})\big) + \sum_{i=1}^{n} y_i(\boldsymbol{x_i'}\,.\,\boldsymbol{\gamma^T}) \\ &= \sum_{i=1}^{n} - ln\Big(1+\mathrm{e}^{(\boldsymbol{x_i'}\,.\,\boldsymbol{\gamma^T})}\Big) + \sum_{i=1}^{n} y_i(\boldsymbol{x_i'}\,.\,\boldsymbol{\gamma^T}) \\ &= \sum_{i=1}^{n}\Big\{y_i(\boldsymbol{x_i'}\,.\,\boldsymbol{\gamma^T}) - ln\Big(1+e^{(\boldsymbol{x_i'}\,.\,\boldsymbol{\gamma^T})}\Big)\Big\} \end{aligned} \qquad (۳۰-۳)$$

برای به دست آوردن برآوردهای درست‌نمائی ماکزیمم، باید از $ln\, L$ نسبت به پارامترها مشتق گرفته و آن‌ها را برابر صفر قرار داد. مشتق $ln\, L$ نسبت به بردار مؤلفه‌های $\boldsymbol{\gamma}$ به صورت رابطه (۳-۳۱) خواهد بود.

$$\frac{\partial\, ln\, L}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \boldsymbol{x_i'}\big(y_i - p(\boldsymbol{x_i'}, \boldsymbol{\gamma})\big) = 0 \qquad (۳۱-۳)$$

که شامل $N+1$ برابری غیرخطی است. از آنجایی‌که اولین مؤلفه $\boldsymbol{x_i'}$، عدد ۱ است، بنابراین برای اولین برابری خواهیم داشت:

$$\sum_{i=1}^{n}\big(y_i - p(\boldsymbol{x_i'}, \boldsymbol{\gamma})\big) = 0 \implies \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p(\boldsymbol{x_i'}, \boldsymbol{\gamma})$$

برای حل سایر برابری‌های (۳-۳۱) و به دست آوردن برآوردها باید از روش‌های عددی استفاده شود. برای این منظور، روش‌های عددی مختلفی وجود دارند که یکی از بهترین آن‌ها روش نیوتن-رافسون[1] است. برای استفاده از روش نیوتن-رافسون به مشتق مرتبه دوم تابع $ln\, L$ که به آن ماتریس هسین[2] گویند نیز نیاز داریم که به صورت رابطه (۳-۳۲) محاسبه می‌شود.

$$\frac{\partial^2\, ln\, L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} = -\sum_{i=1}^{n} \boldsymbol{x_i'} \boldsymbol{x_i'}^T p(\boldsymbol{x_i'}, \boldsymbol{\gamma})\big(1-p(\boldsymbol{x_i'}, \boldsymbol{\gamma})\big) \qquad (۳۲-۳)$$

در نهایت با استفاده از روش نیوتن-رافسون داریم:

---

[1] Newton Raphson Method
[2] Hessian

$$\boldsymbol{\gamma}^{r+1} = \boldsymbol{\gamma}^{r} - \left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\right)^{-1} \frac{\partial \ln L}{\partial \boldsymbol{\gamma}} \qquad (۳۳-۳)$$

که در آن مشتقات اول و دوم مربوط به $\boldsymbol{\gamma}^r$ می‌باشند.

روابط (۳-۳۱) تا (۳-۳۳) را با تعریف بعضی نمادهای جدید می‌توان به ترتیب به صورت روابط زیر بازنویسی کرد.

$$\frac{\partial \ln L}{\partial \boldsymbol{\gamma}} = X^T(y - p)$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} = -X^T W X$$

$$\begin{aligned} \boldsymbol{\gamma}^{r+1} &= \boldsymbol{\gamma}^{r} + (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} X^T (\boldsymbol{y} - \boldsymbol{p}) \\ &= (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \underbrace{(\boldsymbol{X}\boldsymbol{\gamma}^r + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{p}))}_{\longrightarrow\ z} \\ &= (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z} \end{aligned}$$

که در آن $\boldsymbol{y}$، برداری متشکل از مقادیر y_i، $\boldsymbol{X}$ یک ماتریس $n \times (N+1)$ بعدی متشکل از مقادیر $\boldsymbol{x_i'}$، $\boldsymbol{p}$ برداری متشکل از مقادیر $p(\boldsymbol{x_i'}, \boldsymbol{\gamma}^r)$ و $\boldsymbol{W}$ ماتریس قطری است که مقادیر روی قطر آن را $p(\boldsymbol{x_i'}, \boldsymbol{\gamma}^r)(1 - p(\boldsymbol{x_i'}, \boldsymbol{\gamma}^r))$ تشکیل می‌دهند.

در هر تکرار از عمل بروزرسانی ضرایب، روش نیوتن-رافسون به دنبال $\boldsymbol{\gamma}$ ای می‌گردد که در رابطه زیر صدق کند:

$$\boldsymbol{\gamma}^{r+1} = \arg\min_{\boldsymbol{\gamma}} (z - \boldsymbol{X}\boldsymbol{\gamma})^T W (z - \boldsymbol{X}\boldsymbol{\gamma})$$

در نهایت ضرایب مجهول مدل طی چندین مرحله مینیمم‌سازی و بروزرسانی، برآورد می‌شوند. در نتیجه خواهیم داشت:

$$\hat{p}(\boldsymbol{x}', \boldsymbol{\gamma}) = \frac{\mathrm{e}^{(x' \cdot \hat{\gamma}^T)}}{1 + \mathrm{e}^{(x' \cdot \hat{\gamma}^T)}}$$

۳-۴-۲ معنی‌داری مدل

در رگرسیون لجستیک، همانند رگرسیون خطی ساده، معنی‌داری تابع رگرسیون مورد آزمون قرار می-گیرد. فرضیه مورد آزمون عبارت است از:

$$\begin{cases} H_0: & \beta = 0 \\ H_1: & \beta \neq 0 \end{cases}$$

که در آن $\beta = \{\beta_1, \beta_2, \dots, \beta_N\}$. بدیهی است فرضیه H_0، عدم وجود متغیر در مدل و فرضیه H_1، وجود همه متغیرها در مدل را بیان می‌کنند.

یکی از آزمون‌هایی که برای این منظور می‌توان مورد استفاده قرار داد، آزمون درست‌نمایی ماکزیمم می‌باشد که در آن آماره آزمون عبارت است از:

$$G = -2ln\left[\frac{\text{تابع درست‌نمائی بدون وجود متغیرها}}{\text{تابع درست‌نمائی مدل اشباع شده}}\right]$$

که در آن منظور از مدل اشباع شده[1]، مدلی است که تمام متغیرها در آن وجود دارند. آماره G، دارای توزیع کی-دو با n درجه آزادی است که به صورت زیر محاسبه می‌شود:

(n = تعداد پارامترهای مدل بدون وجود متغیر- تعداد پارامترهای مدل با وجود متغیرها)

اگر $G > \chi^2_{1-\frac{\alpha}{2}}(n)$ یا $G \leq \chi^2_{\frac{\alpha}{2}}(n)$، فرضیه معنی‌داری رگرسیون در سطح α رد شده، و در غیر این-صورت رد نمی‌شود.

۳-۴-۳ معنی‌داری وجود متغیرها

در مدل رگرسیون لجستیک، ممکن است وجود بعضی از متغیرها در مدل غیرضروری باشد و حذف آن‌ها موجب افزایش دقت پیش‌بینی و ساده‌سازی مدل گردد. از این‌رو، معنی‌داری وجود تک‌تک متغیرها در مدل، مورد آزمون قرار می‌گیرد. برای متغیر i ام فرضیه مورد آزمون به صورت زیر است:

[1] Saturated Model

$$\begin{cases} H_0: & \beta_i = 0 \\ H_1: & \beta_i \neq 0 \end{cases}$$

از روش درست‌نمایی ماکزیمم نیز می‌توان برای آزمون معنی‌داری یک متغیر در مدل استفاده کرد، اما از آماره آزمون دیگری به نام آماره والد[1] برای این منظور استفاده می‌شود که به صورت رابطه زیر تعریف می‌شود:

$$W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

که در آن $SE(\hat{\beta}_i)$، خطای معیار $\hat{\beta}_i$ می‌باشد. این آماره تحت فرضیه H_0 دارای توزیع نرمال استاندارد بوده و توان دوم آن دارای توزیع کی-دو با یک درجه آزادی است. اگر $W^2 > \chi^2_{1-\frac{\alpha}{2}}(1)$ یا $W^2 \leq \chi^2_{\frac{\alpha}{2}}(1)$، آنگاه فرضیه معنی‌داری وجود متغیر i ام در مدل در سطح α رد شده، و در غیر این صورت رد نمی‌شود.

۳-۵ معیارهای ارزیابی مدل

برای سنجش دقت و کارایی مدل‌ها، از معیارهای مختلفی استفاده می‌شود که در اینجا بعضی از این معیارها را معرفی می‌کنیم. فرض کنید مشاهدات دارای متغیر پاسخ با رده ۱ را به عنوان مشاهده مثبت (P) و مشاهدات دارای متغیر پاسخ با رده صفر را به‌عنوان مشاهده منفی (N) در نظر بگیریم. در جدول ۳-۷ تعداد این مشاهدات بر اساس رده پیش‌بینی آن‌ها نمادگذاری شده‌اند.

جدول ۳-۷: ماتریس حالات پیش‌بینی برای مسائل رده‌بندی با دو رده

	مشاهده مثبت	مشاهده منفی
پیش‌بینی مثبت	TP	FP
پیش‌بینی منفی	FN	TN

[1] Wald Statistic

که در آن TP تعداد مشاهدات مثبتی است که مثبت پیش‌بینی شده‌اند، FP تعداد مشاهدات منفی است که مثبت پیش‌بینی شده‌اند، FN تعداد مشاهدات مثبتی است که منفی پیش‌بینی شده‌اند و TN تعداد مشاهدات منفی است که منفی پیش‌بینی شده‌اند.

با استفاده از ماتریس حالات پیش‌بینی، معیارهای مختلف برای سنجش دقت و کارائی مدل به‌صورت روابط زیر تعریف می‌شوند.

$$Recall = TPrate = \frac{TP}{TP+FN}$$
$$Precision = \frac{TP}{TP+FP}$$
$$FPrate = \frac{FP}{TN+FP}$$
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

که در آن *Recall*، نسبتی از کل مشاهدات مثبت است که درست رده‌بندی شده‌اند، *Precision*، نسبتی از پیش‌بینی‌های مثبت است که واقعاً مثبت بوده‌اند، $FPrate$، تعداد مشاهدات منفی است که به اشتباه مثبت پیش‌بینی شده‌اند و میزان دقت یا *Accuracy*، معرف نسبت مشاهداتی است که درست پیش‌بینی شده‌اند. مقادیر بزرگ معیارهای *Recall*، *Precision* و *Accuracy* و مقدار کوچک $FPrate$ مبین مناسب‌تر بودن مدل و عملکرد بهتر آن است (گارسیا[1] و همکاران، ۲۰۰۷). بدیهی است با در نظر گرفتن رده صفر به عنوان رده مثبت می‌توان تمامی معیارهای معرفی شده را برای این رده نیز به دست آورد. انتخاب بهترین مدل با استفاده از معیارهای ارزیابی به دو صورت زیر انجام می‌شود:

۱. اگر هزینه رده‌بندی اشتباه برای همه رده‌های متغیر پاسخ یکسان باشد، یا به عبارت دیگر پیش‌بینی اشتباه رده متغیر پاسخ برای تمام رده‌ها ضرر یکسانی را در پی داشته باشد، آنگاه معیارهای

[1] Garcia

Accuracy و *Precision* تعیین کننده بهترین مدل می‌باشند. علت این امر این است که در این دو معیار، هر دو رده به میزان یکسانی تأثیرگذار می‌باشند.

۲. اگر هزینه رده‌بندی اشتباه برای یکی از رده‌های متغیر پاسخ بیشتر از رده دیگر باشد، آنگاه اگر رده مثبت دارای اهمیت بیشتری باشد بهترین مدل، مدلی خواهد بود که از لحاظ معیار *Recall* دارای بیشترین دقت است. اما اگر رده منفی دارای اهمیت بیشتری باشد، آنگاه بهترین مدل، مدلی خواهد بود که از لحاظ معیار $FPrate$ دارای بیشترین دقت باشد.

فصل چهارم

پیاده‌سازی مدل‌ها و استخراج نتایج

در این فصل، ابتدا داده‌های مورد استفاده در پایان‌نامه معرفی شده و الگوریتم اسموت به منظور عملیات پیش‌پردازش معرفی می‌شود. سپس مدل‌های معرفی شده در فصل ۳ را بر روی مجموعه داده‌ها برازش داده و معیارهای سنجش دقت مدل را برای هر یک از این مدل‌ها محاسبه می‌کنیم.

۴-۱ معرفی مورد مطالعاتی و انجام پیش‌پردازش بر روی داده‌ها

داده‌های مورد استفاده در این پایان‌نامه مربوط به مجموعه داده‌های ریزش مشتری موجود در پایگاه داده دانشگاه ایروین[1] کالیفرنیا (بلیک[2]، ۱۹۹۸) می‌باشد. این مجموعه داده که مربوط به مشترکان شبکه تلفن همراه است، شامل ۳۳۳۳ مشاهده و ۱۴ متغیر است که ۱۳ متغیر آن، توضیحی و یک متغیر به عنوان متغیر پاسخ در نظر گرفته شده است. مشخصات این متغیرها در جدول ۴-۱ بیان شده‌اند.

جدول ۴-۱: متغیرهای موجود در مورد مطالعاتی و مشخصات آن‌ها

نام متغیر	نقش متغیر	نوع متغیر	نام متغیر	نقش متغیر	نوع متغیر
کل زمان مکالمه در روز "Day-Mins"	توضیحی	عددی	تعداد تماس‌های بین-المللی "Intl-Calls"	توضیحی	عددی
تعداد تماس‌ها در روز "Day-Calls"	توضیحی	عددی	مدت حساب "Account-Length"	توضیحی	عددی
کل زمان مکالمه در غروب "Eve-Mins"	توضیحی	عددی	طرح بین‌المللی "Intl-Plan"	توضیحی	اسمی
تعداد تماس‌ها در غروب "Eve-Calls"	توضیحی	عددی	طرح پیام صوتی "VMail-Plan"	توضیحی	اسمی
کل زمان مکالمه در شب "Night-Mins"	توضیحی	عددی	تعداد پیام‌های صوتی "VMail-Message"	توضیحی	عددی
تعداد تماس‌ها در شب "Night-Calls"	توضیحی	عددی	تعداد تماس‌های مشتری با خدمات‌دهنده "CustServ-Calls"	توضیحی	عددی
کل زمان مکالمه بین-المللی "Intl-Mins"	توضیحی	عددی	ریزش کردن مشتری "Churn"	پاسخ	اسمی

[1] Irvine
[2] Blake

مشاهدات به طور کامل ثبت شده‌اند و هیچ یک از متغیرها دارای مقادیر گمشده نیستند. متغیر پاسخ، مقادیر ۱ و صفر را متناسب با ریزش و عدم‌ریزش مشتریان اختیار می‌کند. در مطالعات آماری و داده-کاوی به منظور سنجش دقت مدل‌ها، معمولاً مجموعه اصلی داده‌ها را به دو زیرمجموعه مجزا (مثلاً به نسبت ۷۰٪ و ۳۰٪) تقسیم نموده و یکی از آن‌ها را به عنوان مجموعه داده مدل‌ساز و دیگری را به عنوان مجموعه داده آزمون در نظر می‌گیرند. در این مطالعه ۷۰٪ از داده‌های اصلی به عنوان داده‌های مدل‌ساز، و ۳۰٪ مابقی به عنوان داده‌های آزمون در نظر گرفته شد. طبق این تقسیم‌بندی، مجموعه داده‌های مدل‌ساز شامل ۲۳۳۳ مشاهده است که در ۸۵٫۲ درصد آن‌ها، متغیر پاسخ مقدار صفر و در ۱۴٫۸ درصد مابقی مقدار ۱ را اختیار کرده است و می‌توان گفت نسبت ریزش به عدم‌ریزش مشتری، ۱ به ۶ می‌باشد. هنگامی که یکی از سطوح متغیر پاسخ نسبت به سطوح دیگر دارای فراوانی بسیار بیشتری باشد، حالت عدم‌تعادل[1] یا عدم‌بالانس پیش می‌آید. در بروز چنین مواردی، الگوریتم‌های یادگیری و روش‌های آماری، عملکرد خوبی نداشته و نتایج قابل اعتمادی ارائه نمی‌دهند (گارسیا، ۲۰۰۷). برای حل این مشکل، روش‌های مختلفی پیشنهاد شده که با افزایش یا کاهش حجم داده‌ها، نسبت مشاهدات با رده‌های مختلف را به میزان قابل قبولی تعدیل می‌کنند. یکی از این روش‌ها، روش بیش‌نمونه‌گیری[2] است. در این روش با به‌کارگیری الگوریتم‌های مختلف، فراوانی رده‌ای از متغیر پاسخ که در اقلیت است، افزایش می‌یابد. یکی از این الگوریتم‌ها، الگوریتم اسموت می‌باشد که در ادامه به شرح آن می‌پردازیم.

الگوریتم اسموت

الگوریتم اسموت توسط چائولا[3] و همکارانش (۲۰۰۲) معرفی شد. این الگوریتم با تولید مشاهدات ساختگی[4] از رده اقلیت (رده‌ای که فراوانی بسیار کمی دارد)، عمل بیش‌نمونه‌گیری انجام می‌دهد. فرض کنید در اینجا نیز مشابه فصل سوم $\{X_1, X_2, \dots, X_N\}$ مجموعه متغیرهای توضیحی، Y متغیر

[1] Unbalanced situation
[2] Over sampling
[3] Chawla
[4] Synthetic

پاسخ، $\{x_1, x_2, \dots, x_N\}$ مقدار مشاهده شده متغیرهای توضیحی، y مقدار مشاهده متغیر پاسخ، $\vec{o}_i =$ $\{\vec{o}_1, \vec{o}_2, \dots, \vec{o}_n\}$ مجموعه مقادیر متغیرهای توضیحی برای فرد i ام و $\{x_1(o_i), x_2(o_i), \dots, x_N(o_i)\}$ مجموعه کل مشاهدات باشند. مجموعه متغیرهای توضیحی می‌توانند از نوع متغیرهای گسسته و پیوسته باشند و متغیر پاسخ به صورت گسسته می‌باشد که دارای K رده است. اگر فرض کنیم رده l در اقلیت باشد و تعداد T مشاهده از n مشاهده دارای رده l باشند، آنگاه نحوه تولید مشاهدات مصنوعی به این صورت است که ابتدا باید برای هر یک از این T مشاهده، k نزدیک‌ترین همسایگی از همان رده پیدا کنیم. سپس اگر بخواهیم M% بیش‌نمونه‌گیری انجام دهیم (یعنی تعداد مشاهدات با رده اقلیت را به اندازه M% افزایش دهیم)، آنگاه باید برای هر مشاهده، از میان k نزدیکترین همسایگی، M همسایگی را به طور تصادفی انتخاب کنیم. به عنوان مثال، اگر ۶۰ مشاهده از رده اقلیت داشته باشیم و بخواهیم تعداد مشاهدات این رده را دو برابر کنیم (یعنی بخواهیم بیش‌نونه‌گیری ٪۲۰۰ انجام دهیم)، آنگاه باید ۲ همسایگی از میان k نزدیک‌ترین همسایگی برای هر یک از این مشاهدات به طور تصادفی انتخاب کنیم. حال اگر مشاهده ساختگی تولیدشده توسط مشاهده i ام و نزدیک‌ترین همسایگی j ام را با $\vec{o}^*_{i,j}$ نشان دهیم، آنگاه این مشاهده با استفاده از رابطه زیر تولید می‌شود.

$$\vec{o}^*_{i,j} = \vec{o}_i + \lambda(\vec{o}_j - \vec{o}_i) \quad , \quad i = 1,2, \dots, T \quad , \quad j \in \{1,2, \dots, k\}$$

یا به عبارت دیگر:

$$\begin{pmatrix} x_1(o^*_{i,j}) \\ x_2(o^*_{i,j}) \\ . \\ . \\ . \\ x_N(o^*_{i,j}) \end{pmatrix} = \begin{pmatrix} x_1(o_i) \\ x_2(o_i) \\ . \\ . \\ . \\ x_N(o_i) \end{pmatrix} + \lambda \begin{pmatrix} x_1(o_j) - x_1(o_i) \\ x_2(o_j) - x_2(o_i) \\ . \\ . \\ . \\ x_N(o_j) - x_N(o_i) \end{pmatrix}$$

که در آن λ مقداری بین صفر و یک می‌باشد. به عنوان مثال اگر دو متغیر توضیحی داشته باشیم و $\vec{o}_i = (6,4)$ مقادیر متغیرهای توضیحی مربوط به مشاهده‌ای از رده اقلیت باشد و فرض کنیم $=$ $\vec{o}_j$

(4,3) یکی از k نزدیکترین همسایگی برای مشاهده $\vec{o}_i$ باشد، آنگاه با فرض $\lambda = 0.2$ مشاهده ساختگی $\vec{o}^*_{i,j}$ به صورت زیر ساخته می‌شود:

$$\vec{o}^*_{i,j} = \begin{pmatrix}6\\4\end{pmatrix} + 0.2 * \begin{pmatrix}4-6\\3-4\end{pmatrix} = \begin{pmatrix}6\\4\end{pmatrix} + \begin{pmatrix}-0.4\\-0.2\end{pmatrix} = \begin{pmatrix}5.6\\3.8\end{pmatrix}$$

مقادیر λ و k باید توسط کاربر مشخص شوند و k حداقل باید برابر با M باشد. همچنین $M \geq 100$ می‌باشد.

حال با استفاده از الگوریتم اسموت، قصد داریم عمل بیش‌نمونه‌گیری را به ازای مقادیر ۱۰۰٪، ۲۰۰٪، ۳۰۰٪، ۴۰۰٪ و ۴۹۰٪، روی داده‌های مدل‌ساز انجام دهیم. برای انجام این کار از نرم‌افزار "وکا[1]" که یک نرم‌افزار داده‌کاوی است، استفاده کردیم که نتایج آن به قرار جدول ۲-۴ است. از این پس به منظور سهولت در بیان، مجموعه داده‌های مدل‌سازی که عمل بیش‌نمونه‌گیری ۱۰۰٪، ۲۰۰٪، ۳۰۰٪، ۴۰۰٪ و ۴۹۰٪ روی آن‌ها انجام گرفته است را به ترتیب با نام‌های اسموت ۱۰۰٪، اسموت ۲۰۰٪، اسموت ۳۰۰٪، اسموت ۴۰۰ و اسموت ۴۹۰٪ مورد استفاده قرار می‌دهیم.

جدول ۴-۲: نتایج بیش‌نمونه‌گیری

تعداد مشاهدات / داده مدل‌ساز	رده صفر (اکثریت)	رده یک (اقلیت)
مجموعه داده اصلی مدل‌ساز (*Original Data*)	۱۹۹۶	۳۳۷
بیش‌نمونه‌گیری با اسموت ۱۰۰٪ (*Smote 100%*)	۱۹۹۶	۶۷۴
بیش‌نمونه‌گیری با اسموت ۲۰۰٪ (*Smote 200%*)	۱۹۹۶	۱۰۱۱
بیش‌نمونه‌گیری با اسموت ۳۰۰٪ (*Smote 300%*)	۱۹۹۶	۱۳۴۸
بیش‌نمونه‌گیری با اسموت ۴۰۰٪ (*Smote 400%*)	۱۹۹۶	۱۶۸۵
بیش‌نمونه‌گیری با اسموت ۴۹۰٪ (*Smote 490%*)	۱۹۹۶	۱۹۸۸

[1] Weka

علت اینکه عمل بیش‌نمونه‌گیری را تا اسموت ۴۹۰٪ ادامه دادیم این است که با انجام اسموت ۴۹۰٪ نسبت مشاهدات با رده صفر و یک، تقریباً برابر با یک می‌شود و افزایش بیش از این مقدار الگوریتم اسموت موجب بیشتر شدن تعداد مشاهدات با رده یک از تعداد مشاهدات با رده صفر می‌شود.

در زیربخش‌های بعدی مدل‌های معرفی شده را روی هر یک از این مجموعه داده‌ها برازش داده، سپس مقادیر معیارهای ارزیابی مدل را برای مجموعه داده آزمون به ازای تمامی مدل‌ها به دست آورده و نتایج را مورد بررسی قرار می‌دهیم. لازم به ذکر است، برای برازش مدل‌های رگرسیون لجستیک، شبکه عصبی و درخت تصمیم، از نرم‌افزار "*SAS*" و برای برازش مدل نظریه مجموعه مبهم، از نرم‌افزار "روزتا[1]" استفاده کردیم. نتایج به دست آمده به طور مفصل در ادامه بیان می‌شوند.

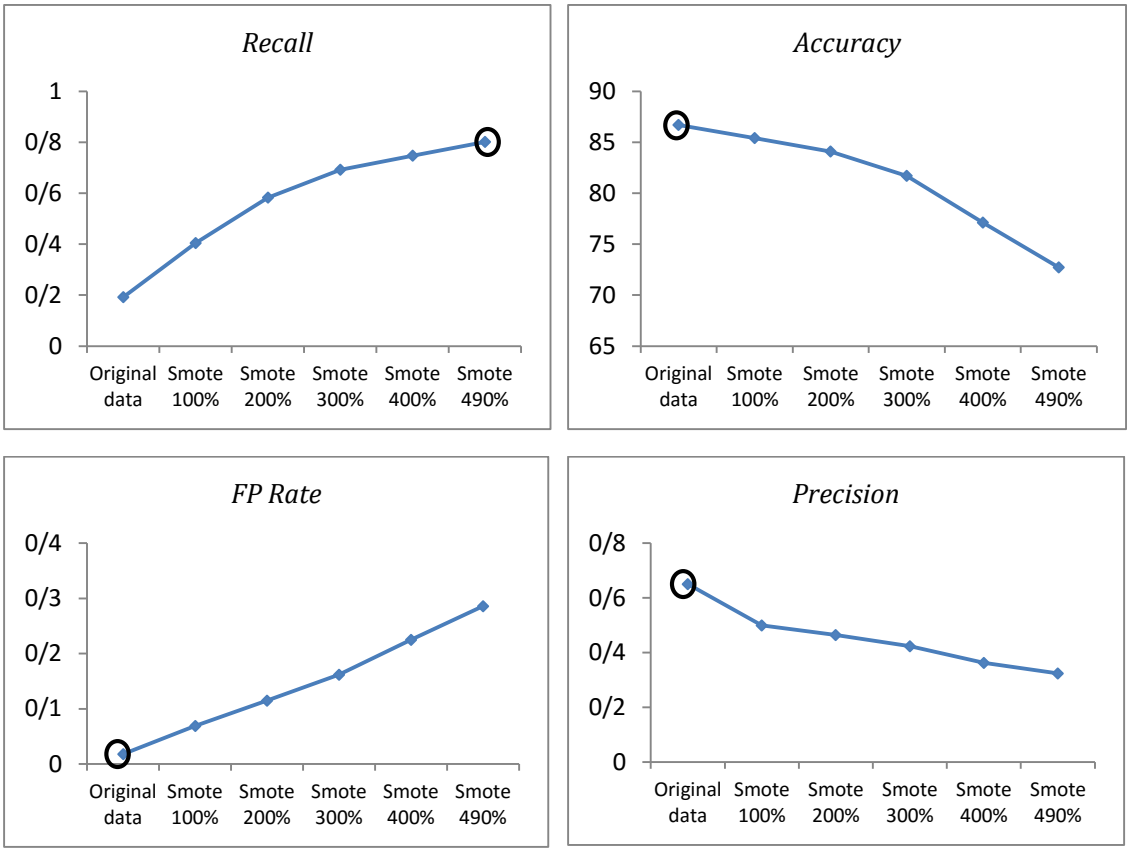
۴-۲ نتایج برازش مدل رگرسیون لجستیک

شکل ۴-۱، نمودارهای معیارهای ارزیابی مدل رگرسیون لجستیک را نشان می‌دهد. هر کدام از نقاط واقع بر نمودار، مربوط به نتایج اعمال این مدل بر روی یکی از مجموعه داده‌های مدل‌ساز است.

با توجه به شکل ۴-۱، از لحاظ معیارهای *Accuracy*، *Precision* و *FPrate*، مدل رگرسیون لجستیک برازش داده شده بر روی مجموعه داده اصلی دارای دقت بالاتری می‌باشد. این به معنی عدم نیاز به بیش‌نمونه‌گیری بوده و مجموعه داده اصلی به عنوان بهترین مجموعه داده انتخاب می‌شود. اما از لحاظ معیار *Recall*، میزان بیش‌نمونه‌گیری توسط اسموت ۴۹۰٪ بهترین مقدار بیش‌نمونه‌گیری است. بنابراین دو مورد پیشنهادی برای انتخاب بهترین مقدار بیش‌نمونه‌گیری عبارتند از:

۱. اگر هزینه رده‌بندی اشتباه رده ۱ (ریزش مشتری) برابر با هزینه رده‌بندی اشتباه رده صفر (عدم ریزش مشتری) باشد، آنگاه مجموعه داده اصلی به دلیل اینکه مدل رگرسیون لجستیک متناظر با آن

[1] Rosseta



شکل ۴-۱: نمودارهای معیارهای ارزیابی مدل رگرسیون لجستیک

دارای بیشترین دقت را از لحاظ معیارهای *Accuracy* و *Precision* است، به عنوان بهترین مجموعه داده پیشنهاد می‌شود. این مدل دارای *Accuracy* برابر با ۸۶٫۷ ٪ می‌باشد.

۲. اگر هزینه رده‌بندی اشتباه رده ۱ (ریزش مشتری) بیشتر از هزینه رده‌بندی اشتباه رده صفر (عدم ریزش مشتری) باشد، آنگاه مقدار بیش‌نمونه‌گیری ۴۹۰٪، به دلیل اینکه مدل رگرسیون لجستیک متناظر با آن از لحاظ معیار *Recall* بیشترین دقت را دارا بود، به عنوان بهترین میزان بیش‌نمونه‌گیری انتخاب می‌شود. این مدل دارای *Accuracy* برابر با ۷۲٫۷٪ می‌باشد.

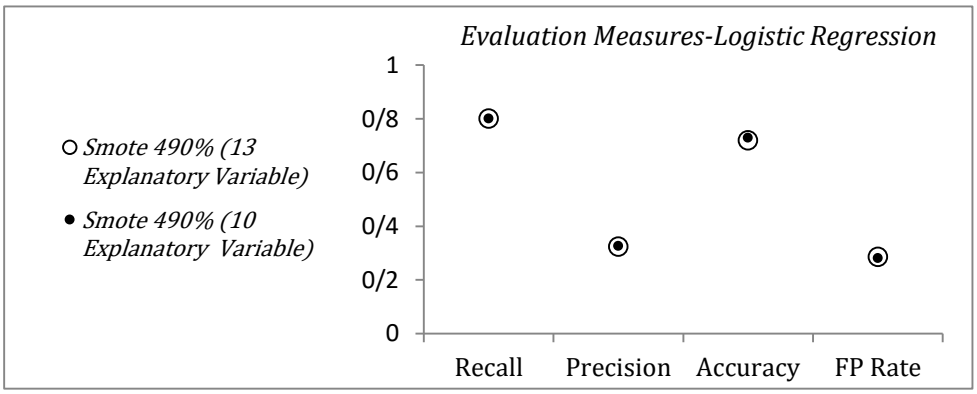
از آنجایی‌که انتظار می‌رود حالت دوم پیشنهاد شده، بیشتر مد نظر کارشناسان در مدیریت ریزش مشتری باشد، لذا ما به بیان نتایج مربوط به مدل رگرسیون لجستیک برازش داده شده روی مجموعه داده اسموت ۴۹۰٪ می‌پردازیم. با برازش مدل رگرسیون لجستیک روی مجموعه داده اسموت ۴۹۰٪، معنی‌داری رگرسیون با استفاده از آزمون درست‌نمایی ماکزیمم، مورد تأیید قرار گرفت. سپس معنی-

داری وجود متغیرها با استفاده از آماره‌های آزمون والد مورد بررسی شد که نتایج آن به شرح جدول ۴-۳ است.

جدول ۴-۳: نتایج آزمون معنی‌داری وجود متغیرها در مدل رگرسیون لجستیک

متغیر	پارامتر	درجه آزادی	مقدار آماره	مقدار-p
Account_Length	α_1	1	8.1784	0.0042
CustServ_Calls	α_2	1	387.0766	<.0001
Day_Calls	α_3	1	0.6378	0.4245
Day_Mins	α_4	1	303.3432	<.0001
Eve_Calls	α_5	1	2.0235	0.1549
Eve_Mins	α_6	1	56.9649	<.0001
Intl_Calls	α_7	1	37.6913	<.0001
Intl_Mins	α_8	1	62.4115	<.0001
Intl_Plan	α_9	1	25.0854	<.0001
Night_Calls	α_{10}	1	0.0430	0.8357
Night_Mins	α_{11}	1	16.3104	<.0001
VMail_Message	α_{12}	1	5.6854	0.0171
VMail_Plan	α_{13}	1	59.6719	<.0001

با توجه به جدول ۴-۳، متغیرهای *Day_Calls*، *Eve_Calls* و *Night_Calls* تأثیر معنی‌داری در مدل ندارند. به عبارت دیگر حذف این متغیرها تأثیری منفی در قدرت رده‌بندی مدل رگرسیون لجستیک نخواهد داشت. با حذف این متغیرها از مجموعه داده اسموت ۴۹۰٪ و اجرای دوباره مدل بر مجموعه داده جدید، شاهد این بودیم که مقادیر معیارهای ارزیابی تقریباً معادل با مدل اصلی بوده و حتی به ازای بعضی از آن‌ها به میزان اندکی افزایش نیز داشته است. نتایج را می‌توان در شکل ۴-۲ مشاهده نمود.



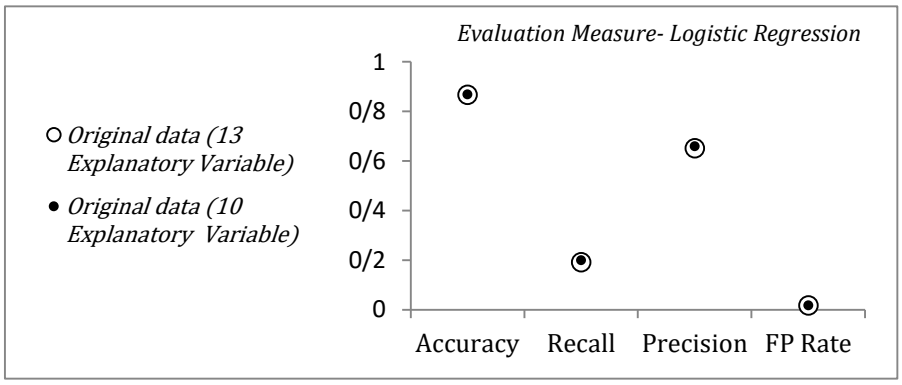
شکل ۴-۲: نمودار معیارهای ارزیابی مربوط به مدل رگرسیون لجستیک برازش داده شده روی مجموعه داده اسموت ۴۹۰٪، قبل و بعد از حذف متغیرهای غیرضروری

بنابراین برای مورد پیشنهادی دوم، مجموعه داده اسموت ۴۹۰٪ با ۱۰ متغیر توضیحی، مناسب می‌باشد. پارامترهای مجهول مدل رگرسیون لجستیک برازش داده شده بر این مجموعه داده، پس از ۴ مرحله بروزرسانی برآورد شدند که به شرح جدول ۴-۴ می‌باشند.

جدول ۴-۴: برآورد پارامترهای مدل رگرسیون لجستیک

پارامتر	α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}
برآورد	-6.330	0.002	0.559	0.011	0.005	-0.100	0.113	-0.335	0.003	0.015	0.764

با توجه به غیرضروری بودن سه متغیر نامبرده برای مجموعه داده اسموت ۴۹۰٪، این سوال مطرح می‌شود که آیا این سه متغیر برای سایر مجموعه داده‌های مدل‌ساز نیز غیرضروری هستند؟ با حذف این سه متغیر از سایر مجموعه داده‌های مدل‌ساز و برازش مدل، درستی این امر تأیید شد و متغیرهای فوق‌الذکر برای تمامی مجموعه داده‌های مدل‌ساز غیرضروری شناخته شدند. از این‌رو به دلیل اینکه در مورد پیشنهادی اول، مجموعه داده‌ی اصلی به عنوان بهترین مجموعه داده‌ی مدل‌ساز در نظر گرفته شده بود، معیارهای ارزیابی را قبل و بعد از حذف سه متغیر با هم مقایسه کردیم که نتایج آن در شکل ۴-۳ نشان داده شده است.



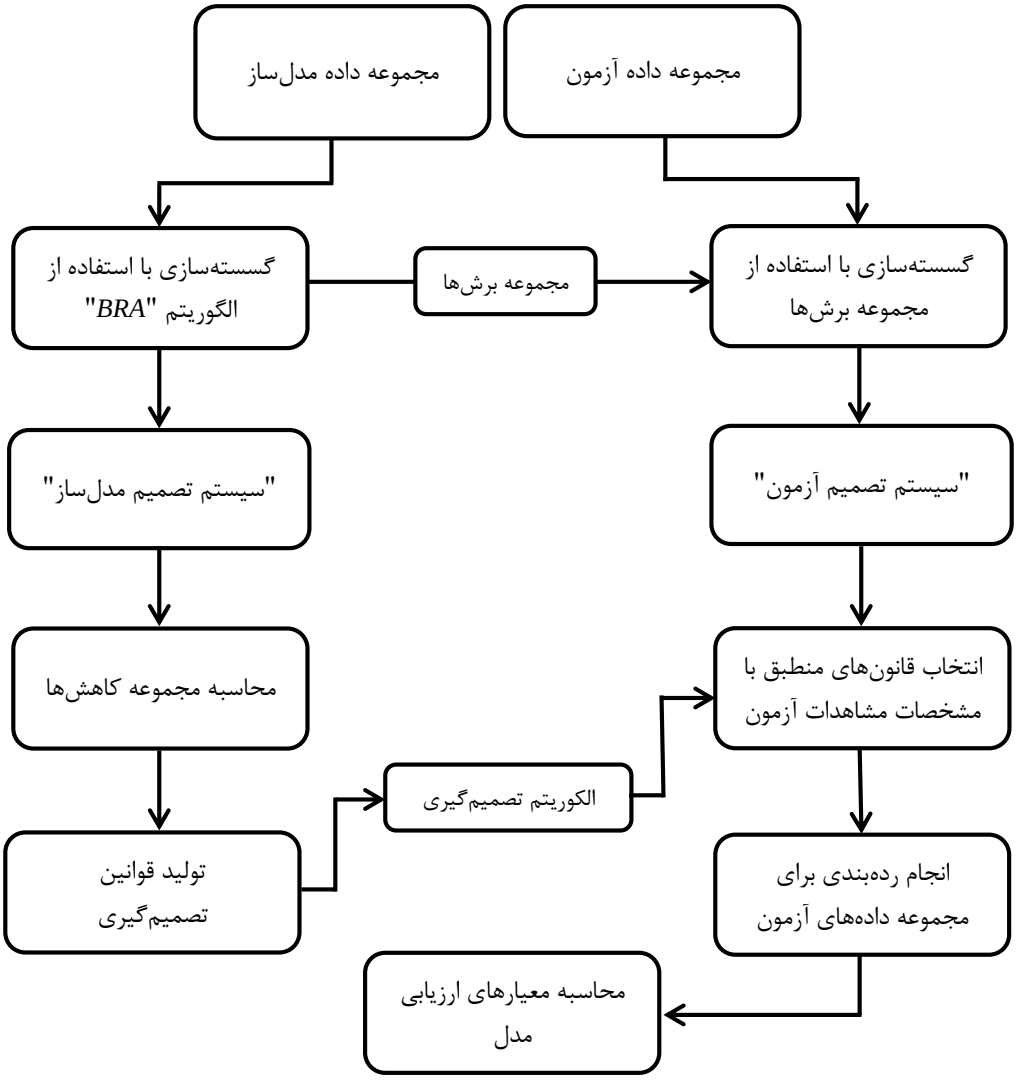
شکل ۴-۳: نمودار معیارهای ارزیابی مربوط به مدل رگرسیون لجستیک برازش داده شده روی مجموعه داده اصلی، قبل و بعد از حذف متغیرهای غیرضروری

از این نمودار نیز به وضوح مشخص است که مدل رگرسیون لجستیک برازش داده شده بر مجموعه داده اصلی، قبل و بعد از حذف متغیرهای غیرضروری، از لحاظ معیارهای مختلف ارزیابی

مدل دارای دقتی تقریباً برابر می‌باشند. بنابراین برای انتخاب بهترین میزان بیش‌نمونه‌گیری برای مدل رگرسیون لجستیک، همان دو مورد پیشنهادی را با حذف ۳ متغیر غیرضروری پیشنهاد می‌کنیم.

۴-۳ نتایج برازش مدل نظریه مجموعه مبهم

برای پیاده‌سازی مدل نظریه مجموعه مبهم بر مجموعه داده مدل‌ساز و استخراج نتایج، باید دنباله‌ای از مراحل طی شوند که در شکل ۴-۴ نمایش داده شده‌اند.



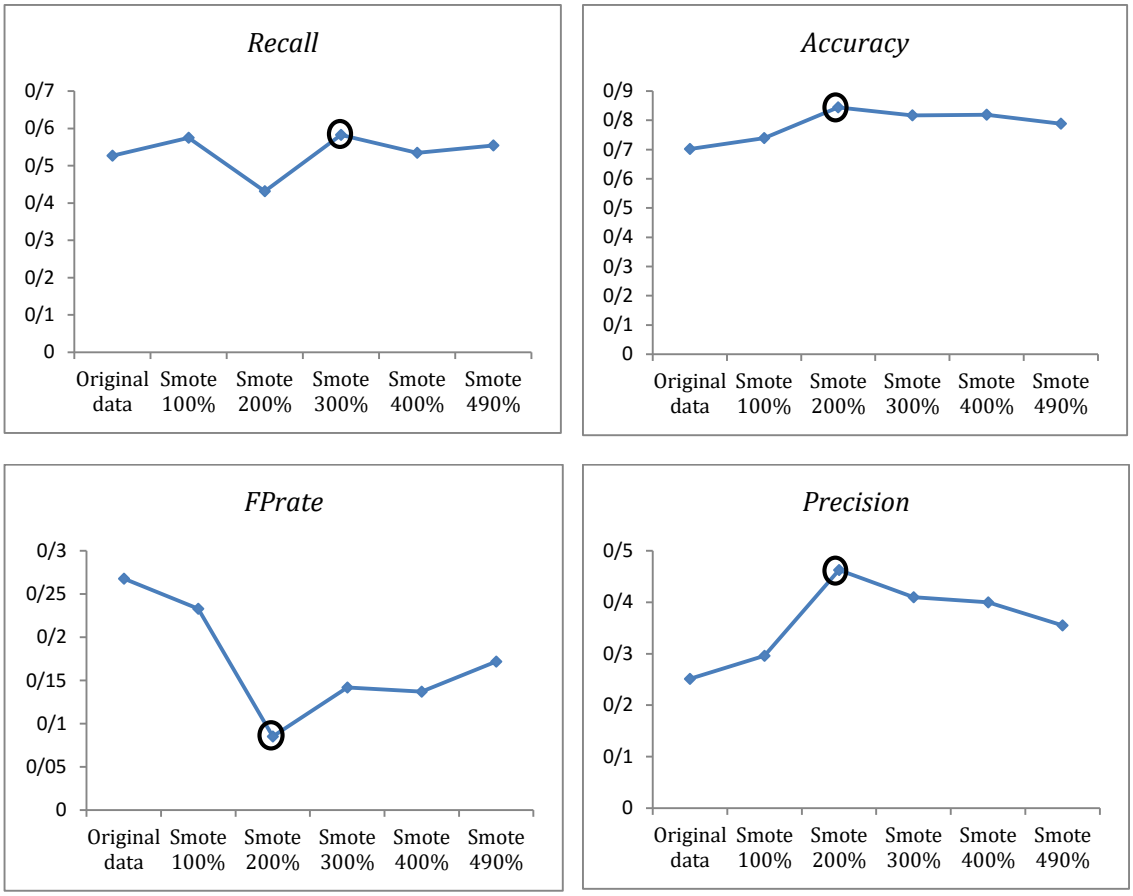
شکل ۴-۴: فرآیند پیاده‌سازی مدل نظریه مجموعه مبهم

این نمودار بیانگر این است که برای پیاده‌سازی یک مدل نظریه مجموعه مبهم، ابتدا باید مجموعه داده مدل‌ساز و آزمون به شکل یک جدول یا سیستم تصمیم‌گیری در آیند. از آنجایی‌که بعضی از

متغیرهای مورد استفاده در پیش‌بینی ریزش مشتری از نوع کمی هستند، باید گسسته‌سازی را برای مجموعه داده مدل‌ساز انجام دهیم. سپس با استفاده از برش‌های ایجاد شده طی گسسته‌سازی روی مجموعه داده‌های مدل‌ساز، عمل گسسته‌سازی را برای مجموعه داده آزمون انجام می‌دهیم. با انجام این دو عمل، سیستم تصمیم مدل‌ساز و سیستم تصمیم آزمون تولید می‌شوند. در مرحله بعد مجموعه کاهش را برای جدول تصمیم مدل‌ساز پیدا می‌کنیم. با استفاده از این کاهش‌ها، قانون‌های تصمیم-گیری و به دنبال آن الگوریتم تصمیم‌گیری استخراج می‌شود. سپس از میان قانون‌های تصمیم‌گیری موجود، آن دسته از قانون‌هایی که با مشخصات مشاهدات موجود در مجموعه داده‌ی آزمون مطابقت دارند جدا شده و سایر قانون‌ها حذف می‌شوند. در نهایت عمل رده‌بندی، انجام شده و معیارهای ارزیابی مدل محاسبه می‌شوند. بدیهی است برای پیاده‌سازی مدل بر هر کدام از مجموعه داده‌های مدل‌ساز (مجموعه داده اصلی، اسموت ۱۰۰٪، اسموت ۲۰۰٪، اسموت ۳۰۰٪، اسموت ۴۰۰٪، اسموت ۴۹۰٪)، باید فرآیند پیاده‌سازی مدل نظریه مجموعه مبهم را یک بار تکرار کنیم. در نهایت پس از برازش مدل روی تمام مجموعه داده‌های مدل‌ساز، نتایج به صورت نمودارهای شکل ۴-۵ خلاصه شده‌اند. در هر یک از این نمودارها، مقادیر معیارهای ارزیابی مدل به ازای مجموعه داده‌های مختلف قرار دارند و بهترین مدل با دایره روی نقطه متناظر با آن مشخص شده است.

با توجه به نمودارهای شکل ۴-۵، مقادیر بهینه معیارهای *Accuracy*، *Recall*، *Precision* و *FPrate*، به ترتیب به ازای میزان بیش‌نمونه‌گیری ۲۰۰٪، ۳۰۰٪، ۲۰۰٪ و۲۰۰٪ به دست آمده است. حال بسته به اینکه برای کارشناسان مربوطه کدام رده از متغیر پاسخ مهم‌تر باشد می‌توان به صورت زیر میزان بیش‌نمونه‌گیری مناسب را پیشنهاد داد:

۱. اگر هزینه رده‌بندی اشتباه رده ۱ (ریزش مشتری) و رده صفر (عدم ریزش مشتری) یکسان باشد، آنگاه مجموعه داده اسموت ۲۰۰٪ به دلیل دارا بودن بیشترین مقادیر معیارهای *Accuracy* و *Precision*، به عنوان بهترین مجموعه پیشنهاد می‌شود. مدل برازش داده شده روی این مجموعه داده دارای *Accuracy* برابر با ۸۴٫۴٪ می‌باشد.

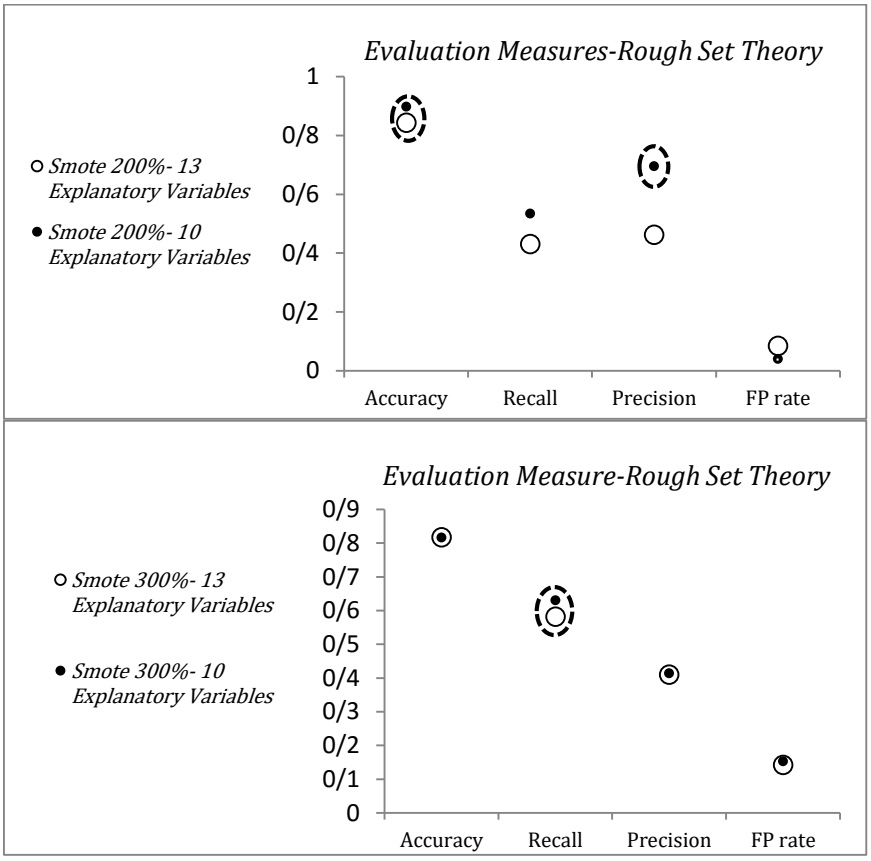


شکل ۴-۵: نمودار معیارهای ارزیابی مدل نظریه مجموعه مبهم برای مجموعه داده‌های مدل‌ساز اصلی، اسموت ۱۰۰٪، ۲۰۰٪، ۳۰۰٪، ۴۰۰٪ و ۴۹۰٪

۲. اگر هزینه رده‌بندی اشتباه رده ۱ (ریزش مشتری) بیشتر از هزینه رده‌بندی اشتباه رده صفر (عدم ریزش مشتری) باشد، آنگاه مجموعه داده اسموت ۳۰۰٪ به دلیل دارا بودن بیشترین مقدار *Recall*، به عنوان بهترین مجموعه داده پیشنهاد می‌شود. مدل برازش داده شده روی این مجموعه داده دارای *Accuracy* برابر با ۸۱٫۶٪ می‌باشد.

با توجه به اینکه در مدل رگرسیون لجستیک، سه متغیر *Day_Calls*، *Eve_Calls* و *Night_Calls* غیرضروری بودند، قصد داریم تأثیرگذاری این سه متغیر را بر قدرت رده‌بندی مدل نظریه مجموعه مبهم نیز مورد بررسی قرار دهیم. بدین منظور، با توجه به اینکه دو مجموعه داده اسموت ۲۰۰٪ و اسموت ۳۰۰٪ از لحاظ معیارهای ارزیابی به عنوان مقدار بیش‌نمونه‌گیری مناسب‌تر انتخاب شدند، متغیرهای غیرضروری را از این دو مجموعه داده حذف کرده و مدل نظریه مجموعه

مبهم را مجدداً بر مجموعه داده‌های جدید، برازش دادیم. نتایج حاصل از حذف متغیرهای غیرضروری برای هر یک از معیارهای ارزیابی به همراه نتایج قبل از حذف، برای حالت‌های اسموت ۲۰۰٪ و ۳۰۰٪ در شکل ۴-۶ آمده است.



شکل ۴-۶: نمودار معیارهای ارزیابی مربوط به مدل نظریه مجموعه مبهم برازش داده شده بر مجموعه داده‌های اسموت ۲۰۰٪ (قاب بالا) و اسموت ۳۰۰٪ (قاب پایین)، قبل و بعد از حذف متغیرهای غیرضروری

با توجه به شکل ۴-۶ می‌بینیم که حذف متغیرهای غیرضروری موجب افزایش دقت مدل نظریه مجموعه مبهم در هر دو مورد پیشنهادی (اسموت ۲۰۰٪، اسموت ۳۰۰٪) شده است. از این‌رو، می‌توان نتیجه‌گیری کرد که مدل نظریه مجموعه مبهم نسبت به متغیرهای غیرضروری حساس بوده و حذف این متغیرها باعث افزایش دقت آن می‌شود. با توجه به این نتیجه، در نهایت مجموعه داده اسموت ۲۰۰٪ و مجموعه داده اسموت ۳۰۰٪ (پس از حذف سه متغیر غیرضروری)، به ترتیب به

عنوان دو مورد پیشنهادی اول و دوم به کارشناسان ارائه می‌شود. مدل برازش داده شده بر این دو مجموعه داده به ترتیب دارای معیار *Accuracy*، ۸۹٫۸٪ و ۸۱٫۷٪ می‌باشند.

از آنجایی‌که در پیش‌بینی ریزش مشتری حالت دوم پیشنهادی (اسموت ۳۰۰٪ پس از حذف سه متغیر غیرضروری) برای انتخاب کارشناسان محتمل‌تر است، لذا نتایج مربوط به برازش مدل نظریه مجموعه راف را روی این مجموعه داده مورد بررسی قرار می‌دهیم. با توجه به اینکه بعضی از متغیرهای شرطی از نوع پیوسته بود، ابتدا با استفاده از روش "*BRA*" عمل گسسته‌سازی را روی مجموعه داده اسموت ۳۰۰٪ (پس از حذف سه متغیر غیرضروری) انجام داده و سپس با استفاده از برش‌های ایجاد شده عمل گسسته‌سازی را روی مجموعه داده آزمون انجام دادیم. با این عمل، سیستم تصمیم مدل‌ساز و سیستم تصمیم آزمون ساخته شد. سپس کاهش‌های ممکن را برای مجموعه متغیرهای شرطی (توضیحی)، از روی سیستم تصمیم مدل‌ساز به دست آوردیم که تنها یک کاهش به صورت زیر به دست آمد.

اگر مجموعه متغیرهای شرطی را با C نشان دهیم، آنگاه کاهش و هسته به دست آمده عبارتند از:

$$Core(C) = Red(C) = \{Account\_Length, Intl\_Plan, VMail\_Message, Day\_Mins, Eve\_Mins, Night\_Mins, Intl\_Mins, Intl\_Calls, CustServ\_Calls\}$$

بر اساس مجموعه متغیرهای درون کاهش، ۳۱۷۴ قانون تولید شد که در زیر ۲ مورد از قانون‌های تولید شده ذکر شده‌اند.

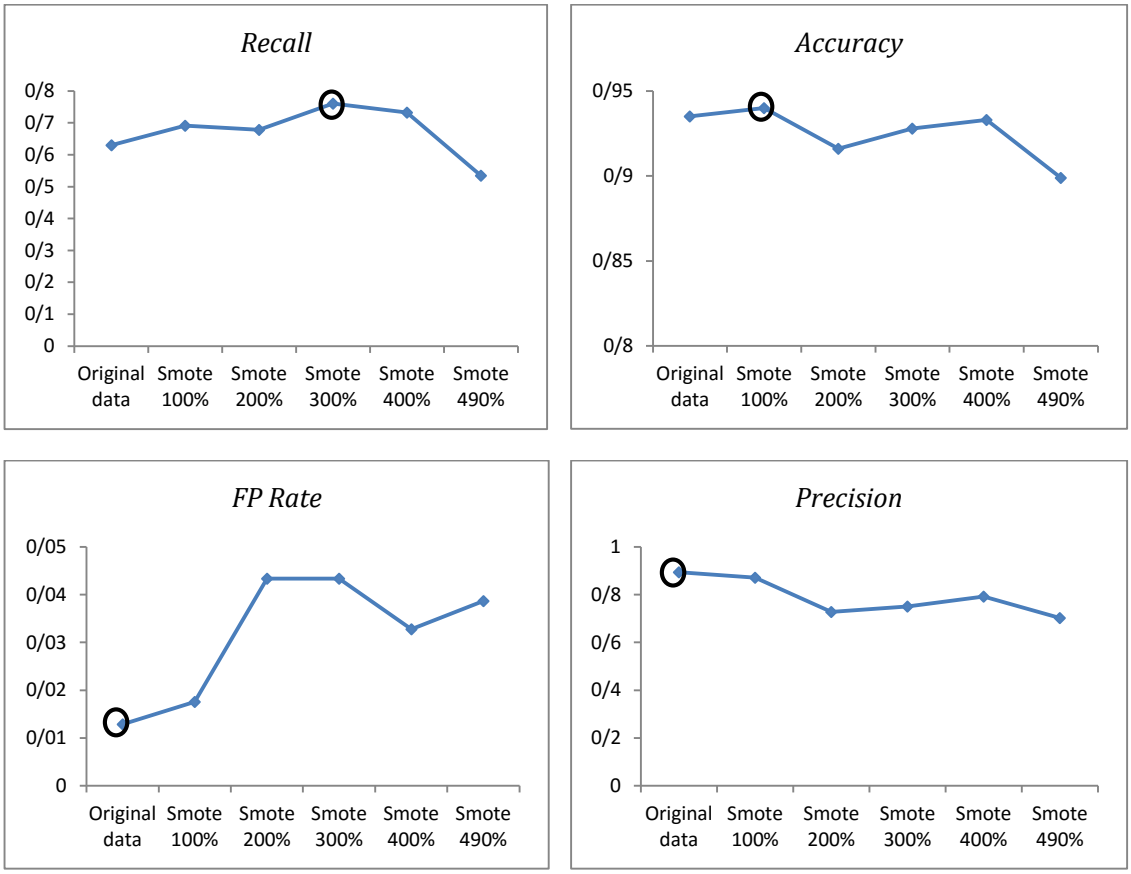
1. *Account_Length([103, 126)) AND Intl_Plan([*, 1)) AND VMail_Message([*, 1)) AND Day_Mins([197.3, 243.5)) AND Eve_Mins([236.7, 293.2)) AND Night_Mins([201.8, 227.4)) AND Intl_Mins([10.4, 12.1)) AND Intl_Calls([*, 3)) AND CustServ_Calls([3, 4)) => Churn_(1)*

2. *Account_Length([103, 126)) AND Intl_Plan([*, 1)) AND VMail_Message([*, 1)) AND Day_Mins([149.3, 197.3)) AND Eve_Mins([223.3, 236.7)) AND Night_Mins([*, 168.9)) AND Intl_Mins([8.8, 10.4)) AND Intl_Calls([3, 4)) AND CustServ_Calls([4, *)) => Churn_(0)*

با استفاده از قانون‌های تولید شده، عمل رده‌بندی را برای مجموعه داده آزمون انجام دادیم که در نتیجه آن به مدلی با دقت ۶۳٪ از لحاظ معیار *Recall* رسیدیم.

۴-۴ نتایج برازش مدل درخت تصمیم

در شکل ۴-۷، نمودارهای معیارهای ارزیابی مدل درخت تصمیم به ازای مجموعه داده‌های آزمون مختلف نشان داده شده است و بهترین مدل در هر نمودار با دایره مشخص شده است.



شکل ۴-۷: نمودار معیارهای ارزیابی مدل درخت تصمیم برای مجموعه داده‌های مدل‌ساز اصلی، اسموت ۱۰۰٪، ۲۰۰٪، ۳۰۰٪، ۴۰۰٪ و ۴۹۰٪

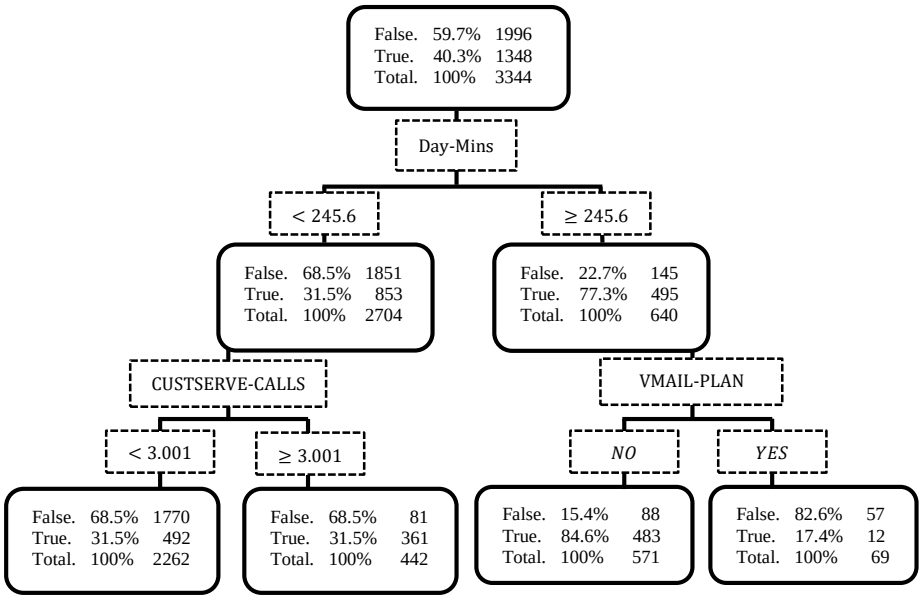
با توجه به نمودارهای مربوط به معیارهای *Accuracy* و *Recall*، به ترتیب میزان بیش‌نمونه‌گیری با الگوریتم اسموت ۲۰۰٪ و ۳۰۰٪ به عنوان بهترین میزان بیش‌نمونه‌گیری انتخاب می‌شوند و از لحاظ معیارهای *Precision* و *FPrate*، مجموعه داده اصلی به عنوان بهترین مجموعه داده انتخاب می‌شود.

حال بسته به اینکه برای کارشناسان مربوطه کدام رده از متغیر پاسخ مهم‌تر باشد می‌توان به صورت زیر میزان بیش‌نمونه‌گیری مناسب را پیشنهاد داد:

۱. اگر هزینه رده‌بندی اشتباه رده ۱ (ریزش مشتری) برابر با هزینه رده‌بندی اشتباه رده صفر (عدم ریزش مشتری) باشد، آنگاه مجموعه داده اسموت ۱۰۰٪، به دلیل دارا بودن بیشترین دقت از لحاظ معیارهای *Accuracy* و دقت بالا از لحاظ معیار *Precision* ، به عنوان بهترین مجموعه داده انتخاب می‌شود. مدل برازش داده شده روی این مجموعه داده دارای *Accuracy* برابر با ۹۴٪ می‌باشد.

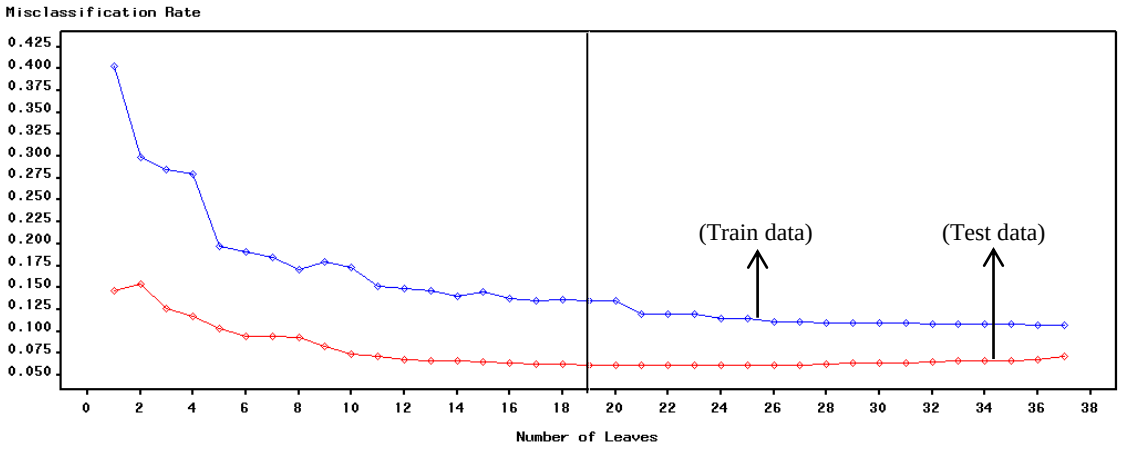
۲. اگر هزینه رده‌بندی اشتباه رده ۱ (ریزش مشتری) بیشتر از هزینه رده‌بندی اشتباه رده صفر (عدم ریزش مشتری) باشد، آنگاه مجموعه داده اسموت ۳۰۰٪، به دلیل دارا بودن بیشترین دقت از لحاظ معیار *Recall*، به عنوان بهترین داده انتخاب می‌شود. مدل برازش داده شده روی این مجموعه داده دارای *Accuracy* برابر با ۹۲٫۸٪ می‌باشد.

حال به ذکر جزئیات مدل پیشنهادی دوم یعنی مدل درخت تصمیم برازش داده شده بر مجموعه داده اسموت ۳۰۰٪ می‌پردازیم. بخشی از درخت ایجاد شده توسط این مدل در شکل ۴-۸ آمده است.



شکل ۴-۸: بخشی از درخت ایجاد شده توسط مدل درخت تصمیم

در این شکل فرآیند افراز درخت تصمیم تا ۴ گره پایانی نشان داده شده است و به دلیل وجود محدودیت، از نشان دادن درخت تصمیم کامل صرفه نظر شده است. نتایج نشان داده است که بهترین تعداد گره پایانی برای درخت تصمیم، ۱۹ گره می‌باشد و از این تعداد گره به بعد، تغییرات خطای رده‌بندی مدل، برای مجموعه داده آزمون بسیار اندک می‌شود. در شکل ۴-۹ این امر به وضوح دیده می‌شود.



شکل ۴-۹: خطای رده‌بندی به ازای تعداد گره‌های پایانی در فرآیند افراز درخت تصمیم برای مجموعه داده مدل‌ساز اسموت ۳۰۰٪ و مجموعه داده آزمون

نتیجه جالب به دست آمده این بود که دقیقاً همان سه متغیری که در مدل رگرسیون لجستیک به عنوان متغیرهای غیرضروری شناخته شدند، در فرآیند افراز درخت تصمیم نیز هیچ نقشی نداشتند. لذا غیرضروری بودن این متغیرها توسط مدل درخت تصمیم نیز تأیید می‌گردد. برای اطمینان بیشتر، این سه متغیر را از مجموعه داده‌های مدل‌ساز اسموت ۲۰۰٪ و اسموت ۳۰۰٪ حذف کردیم و مدل درخت تصمیم را مجدداً بر مجموعه داده جدید برازش دادیم. نتایج نشان داد، دقت مدل درخت تصمیم قبل و بعد از حذف متغیرها دقیقاً باهم یکسان می‌باشند و این امر غیرضروری بودن این سه متغیر را برای مدل درخت تصمیم تأیید می‌کند.

۴-۵ نتایج برازش مدل شبکه عصبی مصنوعی

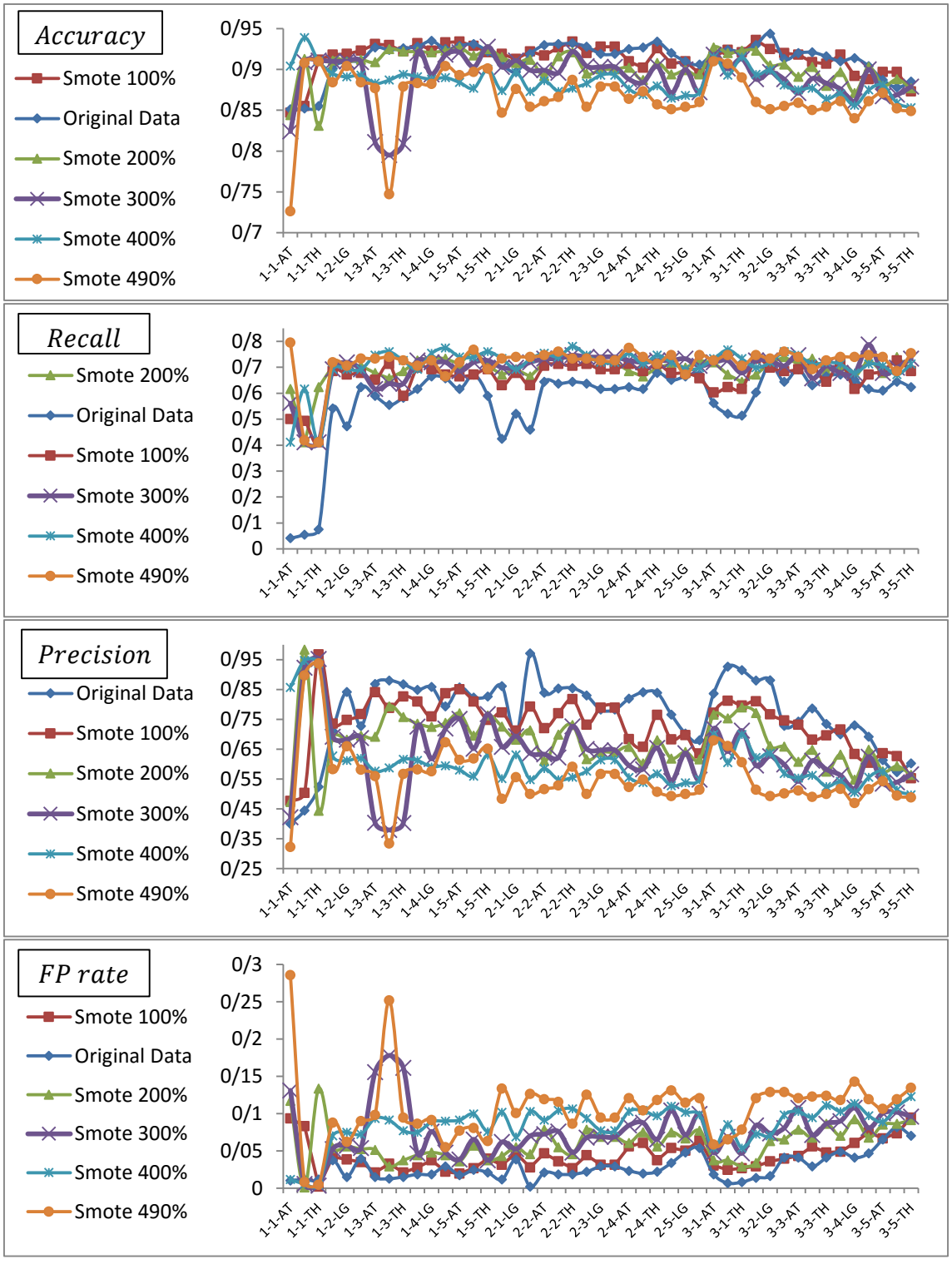
با توجه به تعریف شبکه عصبی با ساختار پرسپترون چندلایه، بدیهی است می‌توان با تغییر تعداد لایه

پنهان، تعداد گره در هر لایه پنهان و نوع تابع فعال‌سازی مورد استفاده در لایه پنهان و لایه خروجی، شبکه‌های عصبی مختلفی ایجاد کرد که همگی دارای ساختار پرسپترون چندلایه هستند. در اینجا تنها شبکه عصبی با ساختار پرسپترون چندلایه را به ازای تعداد ۱ تا ۳ لایه پنهان، در هر لایه ۱ تا ۵ گره، توابع فعال‌سازی تانژانت هایپربولیک (*TH*)، لجستیک (*LG*) و آرک تانژانت (*AT*) در لایه پنهان و تابع سافتمکس در لایه خروجی را بر روی مجموعه داده‌های مدل‌ساز مختلف برازش دادیم و دقت هر یک از مدل‌ها را با استفاده از یک مجموعه داده آزمون بر اساس معیارهای ارزیابی مدل، مورد بررسی قرار دادیم. علت استفاده از تابع سافتمکس در لایه پنهان این بود که با توجه به مطالعات انجام شده، این تابع در مسائل رده‌بندی بهتر از سایر توابع عمل می‌کند.

در شکل ۴-۱۰، نمودار معیارهای ارزیابی به ازای مدل‌های مختلف شبکه عصبی و مجموعه داده-های مدل‌ساز مختلف نمایش داده شده است. در این شکل منظور از 2-1-AT در محور افقی این است که شبکه عصبی متناظر، دارای ۲ لایه پنهان می‌باشد که در هر لایه آن ۱ گره وجود دارد و تابع فعال‌سازی مورد استفاده در لایه‌های پنهان آرک تانژانت می‌باشد.

با توجه به شکل ۴-۱۰، از لحاظ معیار *Accuracy* (قاب اول)، مقدار بیش‌نمونه‌گیری ۱۰۰٪ تقریباً به ازای همه ساختارها، بهترین میزان بیش‌نمونه‌گیری است. از لحاظ معیار *Precision* و *FPrate* (قاب سوم و چهارم)، مجموعه داده اصلی، تقریباً عملکرد بهتری در همه ساختارها از خود نشان داده است. همچنین برای معیار *Recall* (قاب دوم)، مقدار بیش‌نمونه‌گیری ۴۹۰٪ تقریباً به ازای همه ساختارها، بهترین میزان بیش‌نمونه‌گیری می‌باشد. در اینجا نیز دو مورد پیشنهادی برای انتخاب بهترین میزان بیش‌نمونه‌گیری به صورت زیر خواهد بود:

۱. اگر هزینه رده‌بندی اشتباه رده ۱ (ریزش مشتری) برابر با هزینه رده‌بندی اشتباه رده صفر (عدم ریزش مشتری) باشد، آنگاه مدل‌های برازش داده شده بر مجموعه داده اصلی به دلیل دارا بودن بیشترین دقت از لحاظ معیارهای *Accuracy*، *Precision* و *FPrate* بهترین مدل‌ها می‌باشد. بنابراین



شکل ۴-۱۰: نمودار معیارهای ارزیابی مربوط به مدل شبکه عصبی با ساختارهای مختلف

بهترین مجموعه داده، همان مجموعه داده اصلی می‌باشد. در میان این مدل‌ها نیز ساختار 3-2-LG با *Accuracy* برابر با ۹۴٫۴٪ بهترین ساختار است.

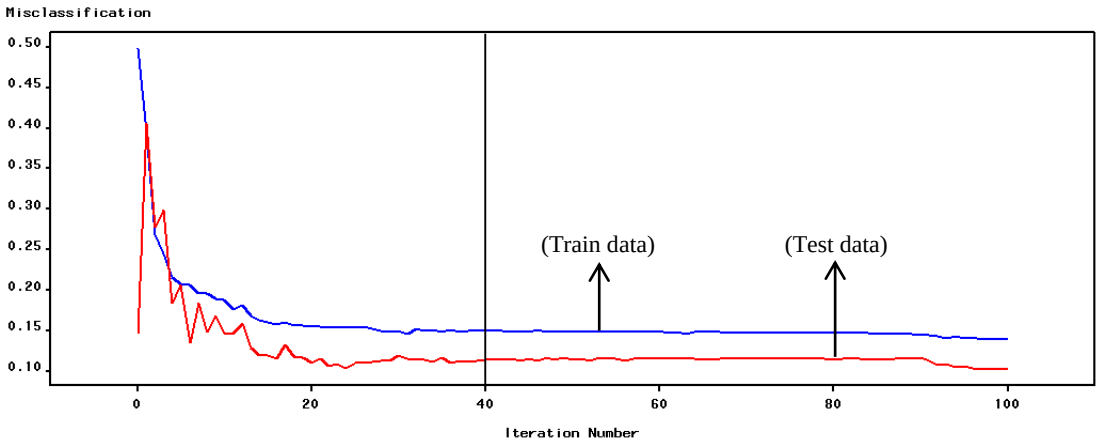
۲. اگر هزینه رده‌بندی اشتباه رده ۱ (ریزش مشتری) بیشتر از هزینه رده‌بندی اشتباه رده صفر (عدم ریزش مشتری) باشد، آنگاه مدل‌های برازش داده شده بر مجموعه داده اسموت۴۹۰٪ بهترین ساختار-ها از لحاظ معیار *Recall* بوده و بهترین میزان بیش‌نمونه‌گیری، اسموت ۴۹۰٪ است. حال باید از میان مدل‌هایی که بر روی مجموعه داده اسموت ۴۹۰٪ برازش داده شده‌اند، بهترین ساختار را انتخاب کنیم. مدل 1-1-TH، بهترین ساختار از لحاظ معیار *Recall* می‌باشد اما این مدل از لحاظ معیارهای دیگر بسیار ضعیف عمل کرده و دقت بسیار پایینی دارد. از این‌رو 1-5-LG را که بعد از 1-1-TH، یکی از بهترین ساختارها از لحاظ معیار *Recall* می‌باشد به عنوان بهترین ساختار انتخاب می‌کنیم. این مدل دارای *Accuracy* برابر با ۸۹٫۷٪ می‌باشد.

در اینجا نتایج مربوط به مدل پیشنهادی دوم یعنی مدل شبکه عصبی با ساختار 1-5-LG برازش داده شده روی مجموعه داده اسموت ۴۹۰٪ را مورد بررسی قرار می‌دهیم. با توجه به مطالب بیان شده در معرفی مدل شبکه عصبی، رابطه مدل شبکه عصبی پیش‌رو با یک لایه پنهان و ۵ گره و تابع فعال-سازی آرک تانژانت در لایه پنهان و تابع فعال‌سازی سافتمکس در لایه خروجی به صورت زیر است:

$$Z_m = \frac{1}{1 + exp\big((\alpha_{0m} + \boldsymbol{\alpha_m^T X})\big)} \quad ; \quad \boldsymbol{\alpha_m^T} = (\alpha_{1,m}, \alpha_{2,m}, \dots, \alpha_{13,m}) \quad ; \quad m = 1,2, \dots ,5$$

$$y_k = f_k(\boldsymbol{X}) = \frac{exp(\beta_{1,k} * Z_1 + \beta_{2,k} * Z_2 + \beta_{3,k} * Z_3 + \beta_{4,k} * Z_4 + \beta_{5,k} * Z_5)}{\sum_{l=1}^{2} exp(\beta_{1,l} * Z_1 + \beta_{2,l} * Z_2 + \beta_{3,l} * Z_3 + \beta_{4,l} * Z_4 + \beta_{5,l} * Z_5)}$$

بنابراین برای به دست آوردن y_k ها نیاز به برآورد ۸۲ ضریب مجهول داریم. در شکل ۴-۱۱، نمودار خطای رده‌بندی در هر مرحله از بروزرسانی برآورد ضرایب برای مجموعه داده اسموت ۴۹۰٪ و مجموعه داده آزمون رسم شده است. واضح است که مقدار خطای رده‌بندی از مرحله ۴۰ به بعد تقریباً به حالت پایداری رسیده و لذا وزن‌های به دست آمده قابل اعتماد هستند.



شکل ۴-۱۱: نمودار خطای رده‌بندی در طی بروزرسانی ضرایب

ضرایب برآورد شده به شرح جدول ۴-۵ می‌باشند:

جدول ۴-۵: برآورد ضرایب مدل شبکه عصبی با ساختار 1-5-LG پس از ۴۰ مرتبه بروزرسانی

$\alpha_{0,1}$	**-8.47537**	$\alpha_{5,2}$	**0.046857**	$\alpha_{10,3}$	**0.081974**	$\alpha_{1,5}$	**-0.14834**
$\alpha_{1,1}$	**-0.04137**	$\alpha_{6.2}$	**0.419756**	$\alpha_{11,3}$	**-0.27077**	$\alpha_{2,5}$	**0.138187**
$\alpha_{2,1}$	**-0.07125**	$\alpha_{7,2}$	**-0.267**	$\alpha_{12,3}$	**0.394944**	$\alpha_{3,5}$	**-0.14171**
$\alpha_{3,1}$	**0.003843**	$\alpha_{8,2}$	**0.333473**	$\alpha_{13,3}$	**-0.98537**	$\alpha_{4,5}$	**0.213812**
$\alpha_{4,1}$	**-0.13557**	$\alpha_{9,2}$	**0.27359**	$\alpha_{0,4}$	**10.65019**	$\alpha_{5,5}$	**-0.06877**
$\alpha_{5,1}$	**0.021564**	$\alpha_{10,2}$	**0.128834**	$\alpha_{1,4}$	**0.107427**	$\alpha_{6,5}$	**0.150184**
$\alpha_{6,1}$	**-0.11899**	$\alpha_{11,2}$	**0.079002**	$\alpha_{2,4}$	**0.147558**	$\alpha_{7,5}$	**0.105965**
$\alpha_{7,1}$	**-0.05224**	$\alpha_{12,2}$	**-0.968**	$\alpha_{3,4}$	**0.070485**	$\alpha_{8,5}$	**-0.22741**
$\alpha_{8,1}$	**-0.00429**	$\alpha_{13,2}$	**0.103339**	$\alpha_{4,4}$	**-7.13389**	$\alpha_{9,5}$	**-0.03751**
$\alpha_{9,1}$	**0.003086**	$\alpha_{0,3}$	**2.720117**	$\alpha_{5,4}$	**0.172409**	$\alpha_{10,5}$	**0.064806**
$\alpha_{10,1}$	**0.023658**	$\alpha_{1,3}$	**-0.08618**	$\alpha_{6,4}$	**-3.02181**	$\alpha_{11,5}$	**0.104344**
$\alpha_{11,1}$	**4.028**	$\alpha_{2,3}$	**0.752395**	$\alpha_{7,4}$	**-0.23159**	$\alpha_{12,5}$	**0.337551**
$\alpha_{12,1}$	**0.015788**	$\alpha_{3,3}$	**-0.48534**	$\alpha_{8,4}$	**-0.31621**	$\alpha_{13,5}$	**1.2528**
$\alpha_{13,1}$	**9.966859**	$\alpha_{4,3}$	**1.088511**	$\alpha_{9,4}$	**0.204226**	$\beta_{1,1}$	**35.09228**
$\alpha_{0,2}$	**5.469187**	$\alpha_{5,3}$	**-0.20909**	$\alpha_{10,4}$	**-1.39469**	$\beta_{2,1}$	**-7.29842**
$\alpha_{1,2}$	**-0.00201**	$\alpha_{6,3}$	**0.903053**	$\alpha_{11,4}$	**16.15131**	$\beta_{3,1}$	**11.33663**
$\alpha_{2,2}$	**-2.1129**	$\alpha_{7,3}$	**0.091907**	$\alpha_{12,4}$	**-0.4865**	$\beta_{4,1}$	**-14.0777**
$\alpha_{3,2}$	**0.140739**	$\alpha_{8,3}$	**-0.34974**	$\alpha_{13,4}$	**7.443245**	$\beta_{5,1}$	**-22.3145**
$\alpha_{4,2}$	**1.137768**	$\alpha_{9,3}$	**0.016722**	$\alpha_{0,5}$	**-0.2463**	$\beta_{0,1}$	**14.71601**

نرم‌افزارهای داده‌کاوی از جمله *SAS* برای کاهش محاسبات، زمانی که متغیر پاسخ دارای K رده است، برآورد ضرایب را برای $(K-1)$ رده اول در لایه خروجی به دست آورده و با این برآوردها مقدار

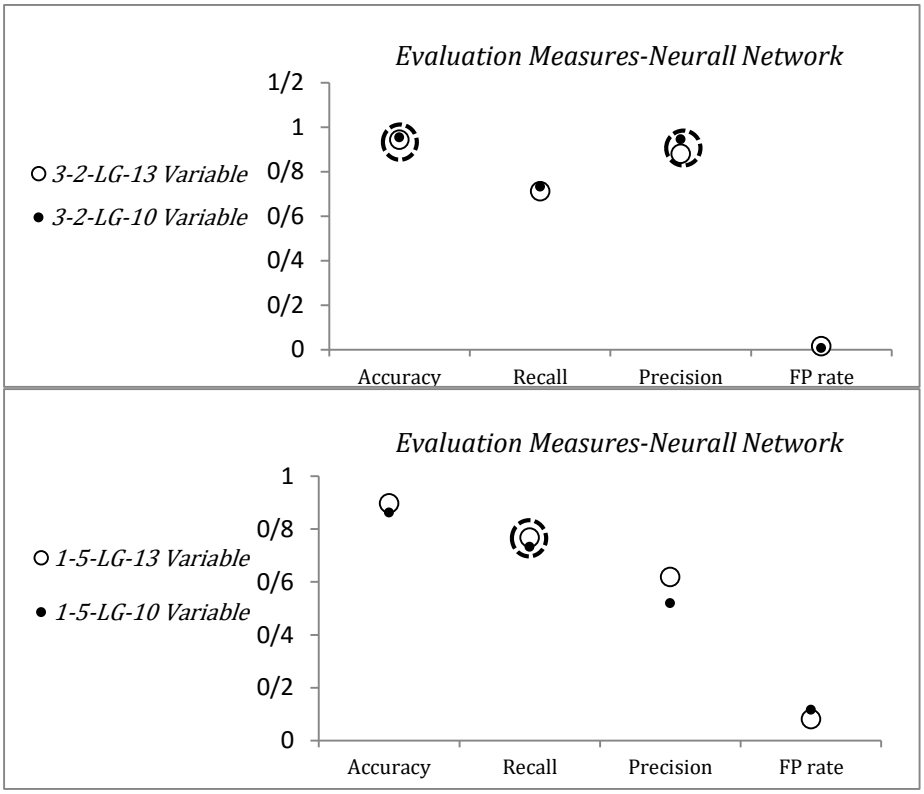
$\hat{y}_k$ را برای $(k = 1,2, \dots, K-1)$ محاسبه می‌کنند. سپس رده $\hat{y}_K$ که ضرایب مجهول آن برآورد نشده، از رابطه زیر را محاسبه می‌کند.

$$\hat{y}_K = 1 - \sum_{i=1}^{K-1} \hat{y}_i$$

علت این امر این است که مقادیر متغیر پاسخ به دست آمده توسط مدل شبکه عصبی، همان مفهوم احتمال پسین را داشته و از خاصیت احتمال پیروی می‌کنند. از این‌رو با توجه به جدول ۴-۳ می‌بینیم که برآورد ضرایب $\beta_{0,2}$، $\beta_{1,2}$،...، $\beta_{5,2}$ محاسبه نشده‌اند.

با توجه به اینکه در مدل رگرسیون لجستیک و مدل درخت تصمیم سه متغیر *Day_Calls*، *Eve_Calls* و *Night_Calls* غیرضروری شناخته شدند، و همچنین مدل نظریه مجموعه مبهم نیز این موضوع را تأیید کرد، قصد داریم تأثیرگذاری این سه متغیر را بر قدرت رده‌بندی مدل شبکه عصبی نیز مورد بررسی قرار دهیم. بدین منظور این سه متغیر غیرضروری را از مجموعه داده اصلی و مجموعه داده بیش‌نمونه‌گیری شده توسط اسموت ۴۹۰٪، حذف کردیم و بهترین مدل‌هایی که در دو مورد پیشنهادی انتخاب شدند، بر مجموعه داده‌های جدید برازش دادیم. نتایج به دست آمده در شکل ۴-۱۲ نشان داده شده است.

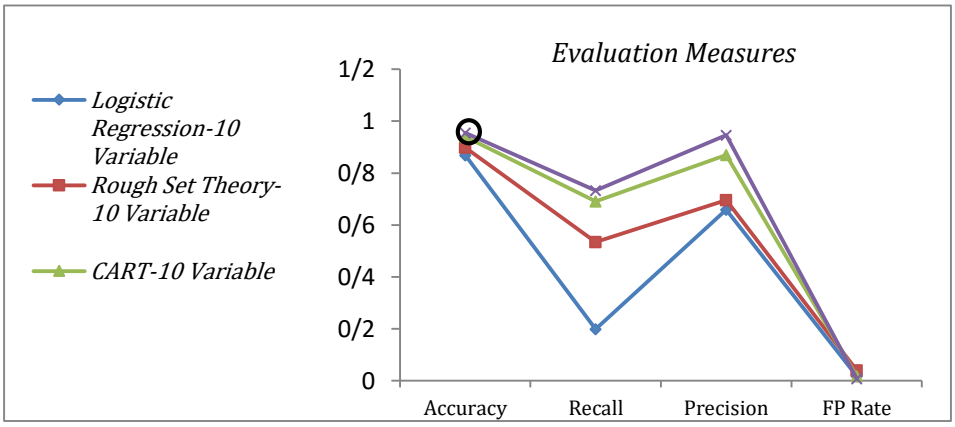
با توجه به نمودارهای شکل ۴-۱۲، می‌بینیم که با حذف سه متغیر غیرضروری از مجموعه داده مدل‌ساز اصلی و برازش مدل 3-2-LG روی آن، دقت مدل به دست آمده بیشتر از مدل قبل می‌باشد و این امر غیرضروری بودن سه متغیر را تأیید می‌کند. مدل برازش داده شده روی مجموعه داده اصلی بعد از حذف سه متغیر دارای *Accuracy* برابر با ۹۵٫۵٪ می‌باشد. اما با حذف متغیرهای غیرضروری از مجموعه داده بیش‌نمونه‌گیری شده توسط اسموت ۴۹۰٪ و برازش مدل 1-5-LG روی آن، شاهد کاهش جزئی دقت معیار *Recall* بودیم. از آنجایی‌که میزان کاهش دقت از لحاظ معیار *Recall* و *Accuracy* جزئی بود، لذا فرضیه غیرضروری بودن سه متغیر فوق‌الذکر با قدرت تأیید نمی‌شود و فقط می‌توان نتیجه گرفت که وجود آن متغیرها تأثیر اساسی در مدل ندارد.



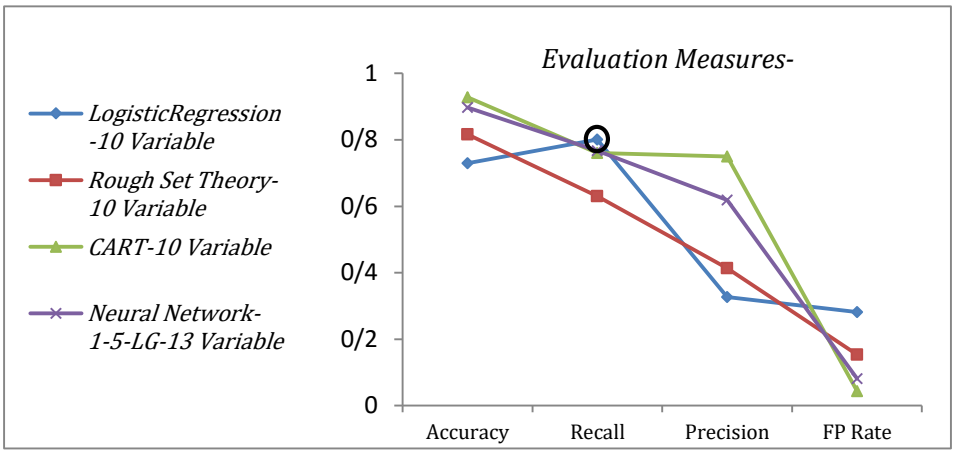
شکل ۴-۱۲: نمودار معیارهای ارزیابی مدل شبکه عصبی با ساختارهای 3-2-LG و 1-5-LG، به ترتیب قبل و بعد از حذف متغیرهای غیرضروری از مجموعه داده اصلی مدل‌ساز و اسموت ۴۹۰٪

۴-۶ انتخاب بهترین مدل

پس از انتخاب بهترین میزان بیش‌نمونه‌گیری برای هر یک از مدل‌ها، می‌توان در خصوص تعیین بهترین مدل پیش‌بینی ریزش مشتری بحث و بررسی نمود. با توجه به دو مورد پیشنهادی اول و دوم برای هر مدل، نتایج معیارهای ارزیابی، به طور جداگانه در شکل‌های ۴-۱۳ و ۴-۱۴ آمده است.



شکل ۴-۱۳: نمودار معیارهای ارزیابی مربوط به موردهای پیشنهادی اول برای چهار مدل



شکل ۴-۱۴: نمودار معیارهای ارزیابی مربوط به موردهای پیشنهادی دوم برای چهار مدل

با توجه به شکل ۴-۱۳، از لحاظ معیار *Accuracy* مدل شبکه عصبی 3-2-LG با ۱۰ متغیر توضیحی بهترین مدل است. این مدل دارای *Accuracy* برابر با ۹۵٫۵٪ می‌باشد. لازم به ذکر است مجموعه داده‌ای که این مدل روی آن برازش داده شد، مجموعه داده اصلی است. اما از لحاظ معیار *Recall* در شکل ۴-۱۴، مدل رگرسیون لجستیک با ۱۰ متغیر دارای بهترین کارایی می‌باشد. مجموعه متغیری که این مدل روی آن برازش داده شد، مجموعه داده اسموت ۴۹۰٪ می‌باشد. این مدل دارای *Accuracy* برابر با ۷۳٪ می‌باشد.

۴-۷ بحث و نتیجه‌گیری

با توجه به نتایج به دست آمده در این فصل می‌توان نتیجه‌گیری کرد که تقریباً تمام مدل‌ها، انجام عمل بیش‌نمونه‌گیری را برای افزایش معیار *Recall* امری ضروری دانسته و در مقابل در حالت کلی این عمل را باعث کاهش سایر معیارهای سنجش دقت، علی‌الخصوص *Accuracy* می‌دانند. بنابراین ماکزیمم‌سازی دو معیار *Accuracy* و *Recall* با هم غیرممکن بوده و افزایش یکی، موجب کاهش دیگری می‌شود. البته باید این نکته را ذکر کرد که این موضوع همیشه صحت نداشته و فقط در مطالعه‌ای که ما انجام دادیم مورد تأیید می‌باشد. بدیهی است با داشتن مورد مطالعاتی دیگر ممکن است این نتیجه‌گیری رد شود. در نهایت باز هم نظر کارشناسان امر تعیین‌کننده خواهد بود که آیا

برای آن‌ها دقت کلی مدل یعنی *Accuracy* مهم‌تر است یا دقت مدل از لحاظ معیار *Recall*. پاسخ این سوال منوط به میزان اهمیت رده‌بندی مشاهدات با رده‌های صفر و یک می‌باشد. اگر برای آن‌ها معیار *Accuracy* مهم‌تر باشد، مدل شبکه عصبی 3-2-LG با ۱۰ متغیر توضیحی بهترین مدل پیشنهادی می‌باشد. اما اگر معیار *Recall* مهم‌تر باشد، مدل رگرسیون لجستیک با ۱۰ متغیر به کارشناسان امر پیشنهاد می‌شود. اما اگر بخواهیم در حالت کلی مدلی را پیشنهاد کنیم که از لحاظ هر دو معیار عملکرد خوبی داشته باشد، باز هم همان مدل شبکه عصبی 3-2-LG با ۱۰ متغیر توضیحی بهترین گزینه برای پیشنهاد خواهد بود. چون این مدل علاوه بر اینکه بهترین مدل از لحاظ معیار *Accuracy* است، از لحاظ معیار *Recall* نیز عملکرد خوبی داشته و *Recall* آن برابر با ۷۳٪ می‌باشد.

نتیجه دیگری که حاصل شد این بود که متغیرهای *Day_Calls*، *Eve_Calls* و *Night_Calls* به طور قطع از لحاظ مدل‌های رگرسیون لجستیک و درخت تصمیم غیرضروری شناخته شده‌اند و از لحاظ دو مدل دیگر نیز تقریباً این فرضیه مورد تأیید واقع شد. همچنین دریافتیم، مدل‌های نظریه مجموعه مبهم و شبکه عصبی، نسبت به وجود متغیرهای غیرضروری در مدل حساس می‌باشند و ممکن است این امر موجب کاهش دقت آن‌ها گردد.

نتیجه‌گیری

مدیریت ریزش مشتری، یکی از علومی است که قابلیت‌های آمار و داده‌کاوی در آن به خوبی به چشم می‌خورد.

نتایج حاکی از آن هستند که می‌توان پیش‌بینی ریزش مشتریان را بر اساس مدل‌های مختلف رده‌بندی مورد ارزیابی قرار داد. در این راستا می‌توان از روش‌های رگرسیون لجستیک، شبکه عصبی، درخت رگرسیون و رده‌بندی و نظریه مجموعه مبهم استفاده نمود. این روش‌ها از جمله روش‌های متداولی هستند که در این زمینه مورد استفاده قرار گرفته‌اند. با توجه به نتایج می‌توان گفت، هر یک از این مدل‌ها دارای نقاط قوت و ضعفی هستند و با توجه به نظر کارشناس مربوطه، مدلی که مناسب-تر است انتخاب می‌شود. برای اعمال نظر کارشناس مربوطه، از معیارهای ارزیابی مختلف استفاده شد. همچنین مسئله عدم تعادل داده‌ها مورد بررسی قرار گرفت و نتایج نشان داد که عدم بررسی این موضوع، ممکن بود منجر به نتیجه‌گیری اشتباه در مورد انتخاب بهترین مدل گردد.

پیشنهادات

به منظور تحقیقات آینده در زمینه پیش‌بینی ریزش مشتری می‌توان از راه‌کارهای پیشنهادی ذیل بهره گرفت:

۱. استفاده از سایر الگوریتم‌های بیش‌نمونه‌گیری، به منظور بالانس داده‌ها

۲. استفاده از سایر مدل‌های آماری و یادگیری ماشین برای پیش‌بینی ریزش مشتری

۳. استفاده از سایر معیارهای سنجش دقت مدل به منظور سنجش کارایی مدل‌ها

۴. مدل‌بندی طول عمر مشتریان با روش‌های تحلیل بقا و ارتباط دادن آن به ریزش مشتری

فهرست منابع

Ahmed, S.R., (2004), "Applications of data mining in retail business", ***Information Technology: Coding and Computing,*** *vol* ***2,*** *pp* ***455–459.***

Au, W.H., Chan, K.C.C., Yao, X., (2003), "A novel evolutionary data mining algorithm with applications to churn prediction", ***IEEE Transactions on Evolutionary Computation,*** *vol* ***7,*** *pp* ***532–545.***

Berson, A., Smith, S., Thearling, K., (2000), "Building data mining applications for CRM", ***McGraw-Hill.***

Blake, C. L., Merz, C. J.,"Churn Data Set, UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.htm", ***University of California, Department of Information and Computer Science, Irvine, CA, 1998.***

Bouyssou, D., (1996), "Outranking relations: do they have special properties?", ***Journal of Multiple Criteria Decision Analysis,*** *vol* ***5,*** *pp* ***99-111.***

Breiman, L., Jerome, H.F., Richard, A.O., Charles, J.S., (1984), "Classification and Regression Trees", ***New York: Wadsworth, Inc.***

Brown, F.M., (1990), "Boolean Reasoning: The Logic of Boolean Equations", ***Kluwer Academic Publishers, Dordrecht, The Netherlands.***

Carrier, C.G., Povel, O., (2003), "Characterising data mining software", ***Intelligent Data Analysis,*** *vol* ***7,*** *pp* ***181–192.***

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.Ph., (2002), "SMOTE: Synthetic minority over-sampling technique", ***Journal of Artificial Intelligence Research.***

Chen, M.C., Chiu, A.L., Chang, H.H., (2005), "Mining changes in customer behavior in retail marketing" ***Expert Systems with Applications,*** *vol* ***28,*** *pp* ***773–781.***

Cheung, K.W., Kwok, J. T., Law, M.H., Tsui, K.C., (2003), "Mining customer product ratings for personalized marketing", ***Decision Support Systems,*** *vol* ***35,*** *pp* ***231–243.***

Chu, B.H., Tsai, M.S., Ho, C.S., (2007), "Toward a hybrid data mining model for customer retention", ***Knowledge-Based Systems****,vol.* ***20****, pp.****703–718****.*

Dimitras, A.I., Slowinski, R., Susmaga, R., Zopounidis, C., (1999), "Business failure prediction using rough sets", ***European Journal of Operational Research ,****vol* ***114,*** *pp* ***263-280.***

Drew, J.H., Mani, D.R., Betz, A.L., Datta, P., (2001), "Targeting customers with statistical and data-mining techniques", ***Journal of Service Research,*** *vol* ***3,*** *pp* ***205– 220.***

Etzion, O., Fisher, A., Wasserkrug, S., (2005), "E-CLV: A modeling approach for customer lifetime evaluation in e-commerce domains, with an application and case study for online auction", ***Information Systems Frontiers,*** *vol* ***7,*** *pp* ***421–434.***

Garcia, V., Sanchez, J.S., Mollineda, R.A., Alejo, R., Sotoca, J.M., (2007), "The class imbalance problem in pattern classification and learning", ***II CongresoEspanol de Informatica.***

Hastie, T., Tibshirani, R., Friedman, J., (2009), "The Elements of Statistical Learning", ***Springer Series in Statistics ,http://www.springer. com/series/692.***

Jiang, T., Tuzhilin, A., (2006), "Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever", ***IEEE Transactions on Knowledge and Data Engineering,*** *vol* ***18,*** *pp* ***1297–1311.***

Jiao, J.R., Zhang, Y., Helander, M., (2006), "A Kansei mining system for affective design", ***Expert Systems with Applications,*** *vol* ***30,*** *pp* ***658–673.***

*KhakAbi, S., Gholamian, M., Namvar***. *M., (2010), "Data Mining Applications in Customer Churn Management",* ***International Conference on Intelligent Systems, Modelling and Simulation****, pp* ***200-225****.*

Kincaid, J.W., (2003), "Customer relationship management: Getting it right.Upper saddle river", ***N.J.: Prentice Hall PTR.***

Kracklauer, A.H., Mills, D.Q., Seifert, D., (2004), "Customer management as the origin of collaborative customer relationship management", ***Collaborative Customer Relationship Management - taking CRM to the next level,*** *pp* ***3–6.***

Ling, R., Yen, D.C., (2001), "Customer relationship management: An analysis framework and implementation strategies", ***Journal of Computer Information Systems,*** *vol* ***41,*** *pp* ***82–97****.*

Lin, CH.S., Tzeng, G.H., Chin, Y.CH., (2011), " Combined rough set theory and flow network graph to predict customer churn in credit card accounts", ***Expert Systems with Applications****, Vol.* ***38****, pp.****8-15.***

McCullagh, P., Nelder, J.A., (1989), "Generalized Linear Models", ***Chapman and Hall: London.***

Minguel, A.P., (2001), "Measuring the impact of data mining on churn management. Internet Research", ***Electronic Network Applications and Policy, 11(5):*** *pp****. 375-387.***

Ngai, E.W.T., Xiu, L., Chau, D.C.K., (2009), "Application of data mining techniques in customer relationship management: A literature review and classification", ***Expert Systems with Applications,*** *vol* ***36,*** *pp* ***2592–2602***

Nguyen, H.S., Nguyen, S.H., (1998), "Discretization Methods in Data Mining", in L. Polkowski and A. Skowron (eds.) ***Rough Sets in Knowledge Discovery,*** *Vol.**1,** pp **451-482, Physica-Verlag.***

Parvatiyar, A., Sheth, J.N., (2001), "Customer relationship management: Emerging practice, process, and discipline", ***Journal of Economic & Social Research,*** *vol **3,** pp **1–34**.*

Pawlak, Z., (1982), "Rough sets", ***International Journal of Computer and Information Sciences,*** *vol **11,** pp **341-356.***

Pawlak, Z., (2002), "Rough set theory and its applications", ***Journal of telecommunications and information technology.***

Prinzie, A., Poel, D.V.D., (2006), "Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models", ***European Journal of Operational Research,*** *vol **170,** pp **710–734.***

Roger, J., Lewis, M.D., (2000), "An Introduction to Classification and Regression Tree (CART) Analysis", ***Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.***

Shaw, M.J., Subramaniam, C., Tan, G.W., Welge, M.E., (2001), "Knowledge management and data mining for marketing", ***Decision Support Systems,*** *vol **31,** pp **127–137.***

Slowinski, R., Stefanowski, J., (1994), "Rough Classification with Valued Closeness Relation, in: E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B. Burtschy Eds.., New Approaches in Classification and Data Analysis, ***Springer, Berlin.***

Stone, M. H., (1936), "The theory of representations for Boolean algebras", ***Trans. Amer. Math. Soc.,*** *vol **40,** pp **37–111.***

Suraj, Z., (2004), "An Introduction to Rough Set Theory and Its Applications: A tutorial", ***ICENCO'2004,*** *December **27-30,** Cairo, Egypt.*

Woo, J.Y., Bae, S.M., Park, S.C., (2005), "Visualization method for customer targeting using customer map", ***Expert Systems with Applications,*** *vol **28,** pp **763–772.***

Yan, L., Wolniewicz, R.H., Dodier, R., (2004), "Predicting Customer Behavior in Telecommunications", ***IEEE Intelligent Systems,*** *pp **50-58.***

Ziarko, W., (1993), "The Discovery ,Analysis and Representation of Data Dependencies in Databases", ***Knowledge Discovery in Databases,AAAI MIT Press ,Cammridge ,****MA* ***,pp 213-228.***

ضمیمه

استدلال بولی

جبر بولی

به پنج‌تائی $(B, +, \cdot, 0, 1)$ که دارای چهار خاصیت پایین باشند، یک جبر بولی گفته می‌شود. مجموعه B، یک مجموعه‌ی حامل نامیده می‌شود. " $+$ " و " $\cdot$ " ، عملگرهای دودویی هستند که روی B تعریف می‌شوند. 0 و 1 نیز اعضاء مجزایی از مجموعه B می‌باشند. چهار خاصیت اساسی عبارتند از:

۱. خاصیت جابه‌جایی: برای هر "$a, b \in B$" خواهیم داشت "$a + b = b + a$" و "$a.b = b.a$".

۲. خاصیت توزیع‌پذیری: برای هر "$a, b, c \in B$" خواهیم داشت "$a + (b.c) = (a + b).(a + c)$" و "$a.(b + c) = (a.b) + (a.c)$" .

۳. یکتایی: برای هر "$a \in B$"، خواهیم داشت "$0 + a = a$" و "$1.a = a$".

۴. مکمل: برای هر "$a \in B$" مجموعه‌ی "$\acute{a} \in B$" وجود خواهد داشت به طوری که "$a + \acute{a} = 1$" و "$a.\acute{a} = 0$".

تعریف ۱:

رابطه " $\leq$ " در جبر بولی به‌صورت زیر تعریف می‌شود.

$$a \leq b \Leftrightarrow a.\acute{b} = 0 \qquad (۱)$$

رابطه " $\leq$ " یک ترتیب جزیی است. ترتیب‌های جزئی را به راحتی در نموداری به نام نمودار هسه[1] نمایش می‌دهند که دارای خواص زیر است:

۱. خاصیت بازتابی: برای هر "$a \in B$" خواهیم داشت، "$a \leq a$".

۲. خاصیت پادتقارنی: برای هر "$a, b \in B$"، اگر $a \leq b$ و $b \leq a$ ، آنگاه $a = b$.

۳. خاصیت انتقال‌پذیری: برای هر "$a, b, c \in B$"، اگر $a \leq b$ و $b \leq c$، آنگاه $a \leq c$.

[1] Hasse

با توجه به نظریه معرفی شده توسط استون[1] (۱۹۳۶)، همه جبرهای بولی دارای مجموعه حامل متناهی، متناظر با یک جبر بولی از زیرمجموعه‌های $(2^A, \cup, \cap, \emptyset, A)$ هستند که در آن A یک مجموعه متناهی می‌باشد. چون که رابطه " $\leq$ " در یک جبر بولی B، متناظر با رابطه "$\subseteq$" در جبر زیرمجموعه‌های متناظر با B است، لذا " $\leq$ " را یک رابطه تضمنی می‌نامند.

تعریف ۲:

یک تابع m متغیره $f: B^m \to B$ را یک *تابع بولی* می‌نامند اگر و تنها اگر بتوان آن را به صورت یک فرمول بولی بیان کرد.

تعریف ۳:

لفظ p را یک *دلالت‌کننده* از تابع بولی f نامند، اگر $p \leq f$. به عبارت دیگر دلالت‌کننده، هر اجتماعی از الفاظ (متغیرها یا منفی آن‌ها) می‌باشد، به طوری که اگر مقداری از این الفاظ برای ارزش دلخواه v از متغیرها مثبت باشد، آنگاه مقدار f نیز تحت ارزش v مثبت باشد. هر اصطلاحی از یک فرمول مجموع حاصل‌ضرب (SOM) برای f به طور واضح یک دلالت کننده از f می‌باشد.

تعریف ۴:

یک دلالت‌کننده از f را یک عمده دلالت‌کننده از f نامند اگر زمانی‌که یکی از الفاظ آن حذف شود، دیگر یک دلالت‌کننده f نباشد. یک دلالت‌کننده p از f، یک عمده دلالت‌کننده از f است در صورتی که برای هر اصطلاح q رابطه ۲ برقرار باشد.

$$p \leq q \leq f \Longrightarrow p = q \qquad ۲$$

انفصالی[2] از عمده دلالت‌کننده‌های f، یک شکل متعارفی بلیک[1] نامیده می‌شود که با $BCF(f)$ نمایش داده می‌شود. به دست آوردن عمده دلالت‌کننده‌ها در حالت کلی بسیار سخت می‌باشد. بدین

[1] Stone
[2] disjunction

منظور الگوریتم‌های تقریبی و ابتکاری وجود دارند که عمده دلالت‌ها را جستجو می‌کنند.

طرح کلی کاربرد استدلال بولی برای حل مسائل P

برون[۲] (۱۹۹۰) طرح کلی را برای حل مسائل P به صورت زیر ارائه کرده است.

۱. کدگذاری کردن مسائل P به صورت یک دستگاه شامل چند معادله‌ی بولی به صورت رابطه (۳).

$$P = \begin{cases} g_1 = h_1 \\ \quad . \\ \quad . \\ \quad . \\ g_k = h_k \end{cases} \qquad (۳)$$

که در g_1 و h_1 توابع بولی هستند.

۲. تبدیل دستگاه معادلات به یک معادله‌ی بولی[۳] به صورت "$f_P = 0$".

$$f_P = \sum_{i=1}^{k} (g_i'.h_i + g_i.h_i') \qquad (۴)$$

۳. محاسبه $BCF(f_P)$، یا همان عمده دلالت کننده‌های f_P.

۴. پس از محاسبه عمده دلالت کننده‌های f_P، مسئله P حل می‌شود.

طرح کلی تعریف شده یک ابزار قدرتمند و مفید برای حل مسائلی از قبیل کشف قانون، خوشه‌بندی، استخراج خصیصه و گسسته‌سازی می‌باشد.

---

[1] Blake
[2] Brown
[3] Boolean equation

Abstract

In recent years, the large number of firms which offer similar services, has helped customers to choose from among several firms based on their needs and enjoy their services. This diversification is clearly seen in many industries such as insurances, telecommunications, internet service providers and cable TV networks and caused the strategy of "keeping customers and avoiding to lose them to the rivals" to become one of the most important managerial strategies. Cutting the relationship with the current service provider is called "customer churn". To lose a profitable customer means that he is being attracted by rival service provider and the cost of attracting a new customer is much more than the cost of keeping him (Moser, 2002). From the economic and risk management point of view, identifying the customers whose churn will incur great loss is crucial. The existing information about customers, including loyal and churned customers, is a base for predicting the future customer behavior. If we could predict the probability of churning customers by examining the characteristics of customers, we would be able to bring the churning rate of customers to the minimum by exerting preemptive activities. There are a number of methods for predicting the customer churn, such as logistic regression, neural networks, decision tree and the Rough Set Theory. In this thesis, we introduce the aforementioned models and investigate the accuracy and efficiency of them on the case study using different evaluation criteria. The results showed that, these criteria suggest different models for predicting customer churn.

Keywords: Customer Churn Management, Rough Set Theory, Neural Network, Classification and Regression Tree, Logistic Regression

Shahrood University of Technology

Faculty of mathematical

This thesis submitted in part fulfillment of the degree of Master Science in mathematical statistics

Customer churn management using the advanced statistical and data mining methods

By:
Seyyed Alireza Mahdavi Talarposhti

Supervisor:
Dr. D. Shahsavani

October 2012