
بسم الله الرحمن الرحيم

دانشکده: ریاضی

گروه : آمار

کشف تقلب با استفاده از روش‌های داده‌کاوی

دانشجو:

سجاد رضایی

استاد راهنما:

دکتر داود شاهسونی

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

شهریور ۹۱

تقدیم به

پدر و مادرم

که از نگاهشان صلابت

از رفتارشان محبت

و از صبرشان ایستادگی را آموختم

و همسرم

که سایه مهربانیش سایه سار زندگیم می‌باشد

تشکر و قدردانی

نمی توانم معنایی بالاتر از تقدیر و تشکر بر زبانم جاری سازم و سپاس خود را در وصف استاد خویش آشکار نمایم، که هر چه گویم ، کم گفته ام.

همچنین بر خود واجب می‌دانم، تشکر صمیمانه‌ای را از تمام کسانی که مرا در انجام این تحقیق یاری نموده‌اند داشته باشم.

از خداوند منان آرزوی سلامتی و توفیق روز افزون برای آنان خواهانم.

تعهد نامه

اینجانب **سجاد رضایی** دانشجوی دوره کارشناسی ارشد رشته آمار گرایش ریاضی دانشکده ریاضی دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **کشف تقلب با استفاده از روش‌های داده‌کاوی** تحت راهنمائی آقای دکتر**داود شاهسونی** متعهد می-شوم:

- 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 شده‌است، ضوابط و اصول اخلاقی رعایت شده‌است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده‌است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده‌است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

هدف این پژوهش، بررسی و کاربرد داده‌کاوی در کشف تقلب و ارائه سه روش طبقه‌بندی در داده-کاوی می‌باشد.

در این تحقیق، چارچوب طبقه‌بندی تقلبات مالی را بیان می‌کنیم و جایگاه داده‌کاوی را در این چارچوب مشخص خواهیم کرد. از آنجائیکه موضوع کشف تقلب به روش‌های طبقه‌بندی در داده‌کاوی مربوط می‌شود، سه روش طبقه‌بندی در داده‌کاوی را که، رگرسیون لجستیک، بیزی خام و شبکه بیزی نام دارد مطالعه می‌کنیم و ارتباط رگرسیون لجستیک و بیزی خام را مورد بررسی قرار می‌دهیم. برای ارزیابی مدل‌های بیان‌شده، معیارهای ارزیابی مدل را بیان کرده و با توجه به یک مجموعه داده، بهترین مدل را انتخاب خواهیم کرد. با توجه به تجزیه و تحلیل نتایج بدست آمده، نتیجه خواهیم گرفت که مدل‌های رگرسیون لجستیک و شبکه بیزی عملکرد یکسانی داشته و بهتر از مدل بیزی خام عمل می‌کنند.

کلمات کلیدی : کشف تقلب، داده‌کاوی، رگرسیون لجستیک، بیزی خام، شبکه بیزی

مقالات استخراج شده

۱. رضایی س، مهدوی ع، الیاسی ا، (۱۳۹۱) " معرفی داده‌کاوی و کاربرد آن در کشف تقلب" اولین همایش بین المللی مدیریت، دانشگاه آزاد، نور

۲. رضایی س، (۱۳۹۱) "مقایسه روش طبقه بندی کننده نیو بیز گوسی با روش رگرسیون لجستیک" یازدهمین کنفرانس آمار ایران، دانشگاه علم و صنعت، تهران

فهرست :

فهرست شکل‌ها

فهرست جدول‌ها

فصل اول

مقدمه

مقدمه

امروزه با توجه به تنوع زیاد مخاطبین، مشتریان، بازارها، تنوع و پیچیدگی خدمات و محیط‌های کسب و کار، دسترسی به اطلاعات مناسب برای تصمیم‌گیری صحیح ضروری است. از این‌رو استفاده از راه-کارهای مناسب برای طبقه‌بندی و تولید اطلاعات از میان انبوهی از داده‌ها برای سازمان‌ها امری ضروری و حیاتی است. پیچیدگی سازمان‌ها و تراکنش‌ها باعث افزایش فرصت برای استفاده‌های سود-جویانه و تقلب‌ها شده است. اثرات سازمانی این تقلب‌ها علاوه بر ضررهای اقتصادی، اثرات روانی مختلفی را نیز بر روی خود سازمان و مشتریان آن خواهد داشت. این اثرات ممکن است باعث تاثیر بر روی اعتبار سازمان و سطح رضایت مشتریان شده و درجه اعتماد مشتریان نسبت به سرویس‌های جدید ارائه شده را کاهش دهد. معضلی که امروزه مدیران سازمان‌ها به آن دچار شده‌اند، تقلبات مالی است و به دنبال راه‌حل‌هایی برای کشف و جلوگیری آن هستند. روش‌های مختلف **داده‌کاوی**[1] با ارائه الگوهای مناسب و هم‌چنین طبقه‌بندی مناسب اطلاعات ورودی، این مهم را فراهم می‌کنند. علم داده-کاوی، مبتنی بر روش‌های آماری، پاسخی به این نیاز به منظور کشف دانش پنهان داده‌ها و تامین اطلاعات مورد نیاز مدیران می‌باشد. روش‌های طبقه‌بندی مختلفی در داده‌کاوی در زمینه کشف تقلب پیشنهاد شده است که در هر یک، میزان دقت طبقه‌بندی اطلاعات ورودی، توسط مدل‌های موجود بیان می‌گردد. مهم‌ترین چالشی که محققین را با آن مواجه می‌کند، فقدان اطلاعات است. در بسیاری از مقالات و کتب ارائه شده، نبود اطلاعات کافی در رابطه با اطلاعات مشتریان و هم‌چنین تقلبات صورت گرفته در سازمان‌ها، به مساله‌ای جدی برای پژوهشگران تبدیل شده است. تاکنون روش‌های متعددی توسط محققین برای رسیدن به مدلی مناسب، با دقت طبقه‌بندی بالا ارائه شده است. در این پایان‌نامه نیز به دلیل در دسترس نداشتن اطلاعات و داده‌های واقعی، روش‌هایی سودمند در زمینه کشف تقلبات مالی را بیان کرده و با استفاده از یک مجموعه داده فرضی به ارزیابی مدل‌های طبقه-بندی بیان‌شده می‌پردازیم. از طرف دیگر با فرض در اختیار داشتن داده، به دلیل حجم عظیم آن

[1] Data mining

پردازش اطلاعات بسیار سخت و زمان‌بر می‌باشد که در فصل پنجم با ارائه روشی مناسب، به حل این چالش خواهیم پرداخت . روش‌های مورد مطالعه، رگرسیون لجستیک[1]، طبقه‌بندی کننده بیزی خام[2] و روش شبکه بیزی[3] است که در آن، ساختار روش‌های طبقه‌بندی‌کننده بیزی خام و شبکه بیزی، همان‌طور که از اسمشان مشخص است بر پایه قانون بیز می‌باشد. در این تحقیق قصد داریم با ارائه چارچوبی از تقلبات مالی، جایگاه داده‌کاوی را در ارتباط با این موضوع بررسی کنیم، سپس سه روش موثر در داده‌کاوی را بطور کامل مطالعه نموده و ارتباط روش طبقه‌بندی‌کننده بیزی خام با روش رگرسیون لجستیک را بیان کنیم، در نهایت با استفاده از مجموعه داده مورد نظر مدل‌های بیان شده را مورد ارزیابی قرار دهیم .

۱-۱ ضرورت و اهداف پایان‌نامه

با توجه به اخبار و اطلاعاتی که امروزه منتشر می‌شود، افراد متعددی دچار ضررهایی هنگفت بر اثر تقلبات مالی شده‌اند. هر ساله با افزایش تقلبات مالی، این مساله به عنوان راهی برای کسب درآمد تبدیل شده است. در سال ۲۰۰۷، ارزش مطالبات جعلی بیمه‌گرهای انگلستان در مجموع ۱/۶ میلیارد پوند در سال بوده است و یا در مطالعه‌ای که بروی ۹۰۰ شرکت در سراسر جهان در سه سال ۲۰۰۶ تا ۲۰۰۸ صورت گرفت، بطور متوسط ۸/۲ میلیارد دلار در هر سال شرکت‌ها دچار زیان مالی شدند که افزایش ٪۲۲ نسبت به بررسی منتشر شده را نشان می‌دهد. در سال ۲۰۰۸ شرکت‌هایی که حداقل یک بار تقلب در آنها صورت گرفته است به ٪۸۵ رسید که افزایش ٪۸۰ نسبت به سال قبل را داشته است. با توجه به اهمیت این موضوع، ارتباط بین شاخه آمار با سایر علوم که موجب شده است تا پژوهشگران از دست‌آوردها و روش‌های آماری به عنوان ابزاری قوی برای تجزیه و تحلیل اطلاعات استفاده کنند، امری اجتناب‌ناپذیر است. تکنیک‌های داده‌کاوی، با بهره‌گیری از روش‌های آماری، می-

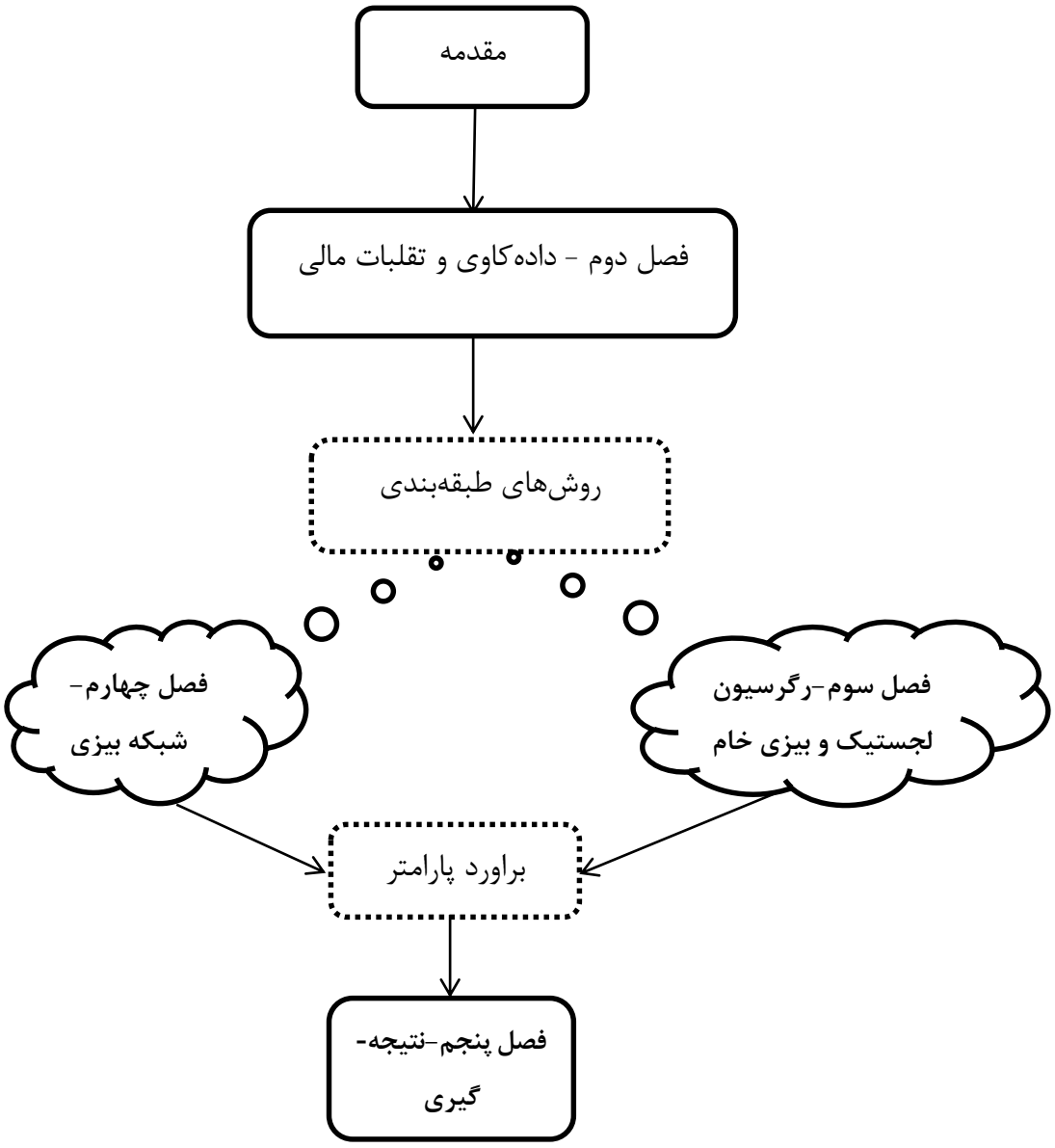
[1] Logistic regression
[2] Naïve bayes classifier
[3] Bayesian network

توانند چالش‌های موجود در تحلیل اطلاعات را شناسایی نموده و راه‌حل‌هایی برای مرتفع کردن آن ارائه نمایند. از سوی دیگر، مدیران سازمان‌ها با بهره‌گیری از دیدگاه‌ها و روش‌های نوین آماری، می‌توانند برای بالا بردن سطح رضایت‌مندی مشتریان و هم‌چنین ارتقاء سطح کیفی سازمانشان، از آن‌ها به عنوان ابزاری پرکاربرد استفاده نمایند. با توجه به ضررهای هنگفتی که افراد و سازمان‌ها هر ساله دچار آن می‌شوند و هم‌چنین اهمیتی که امروزه کشف تقلب در سراسر دنیا پیدا کرده است، انگیزه‌ای شد تا در این پایان‌نامه رویکرد کاربردی آمار مورد توجه قرار گیرد.

۱-۲ ساختار پایان‌نامه

بخش‌های مختلف این پایان‌نامه مطابق ذیل تنظیم شده است :

فصل دوم شامل معرفی داده‌کاوی و تعاریفی از تقلبات مالی است، و با ارائه چارچوبی از تقلبات مالی، جایگاه و اهمیت روش‌های داده‌کاوی در ارتباط با کشف تقلبات مالی را نشان داده‌ایم. در فصل سوم و چهارم که اساس کار این پایان‌نامه می‌باشد، به روش‌های طبقه‌بندی شامل رگرسیون لجستیک، بیزی خام و شبکه بیزی پرداخته شده است که در فصل سوم ابتدا روش‌های رگرسیون لجستیک و بیزی خام را مطالعه نموده، سپس ارتباط میان این دو روش را مورد بررسی قرار می‌دهیم و روش شبکه بیزی را در فصل چهارم بیان می‌کنیم. در فصل پنجم با استفاده از معیارهای ارزیابی مدل، به جمع‌بندی مطالب و بیان پیشنهادات در رابطه با ادامه این موضوع با اهمیت پرداخته‌ایم. نمودار زیر ساختار کلی این تحقیق را نشان می‌دهد.



شکل(۱-۱) ساختار کلی پایان‌نامه

فصل دوم

داده‌کاوی و تقلبات مالی

۲-۱ تعریف تقلب

فرهنگ انگلیسی آکسفورد تقلب را فریب غیرقانونی یا جزایی به منظور رسیدن به سود مالی یا شخصی تعریف کرده است. همچنین می‌توان تقلب را این‌گونه بیان کرد "بدست آوردن سود مالی و یا از دست دادن آن، بصورت آشکار یا پنهان، که در آن متقلب به سود یا ضرر غیر قانونی می رسد". اگرچه تعریف جهانی پذیرفته شده‌ای از تقلب مالی وجود ندارد ولی دانشمندان هر یک تعریفی را بیان کرده اند، فوآ[1] و همکاران تقلب را این‌گونه تعریف می کنند:"یک اقدام که منجر به سواستفاده از نظام سود سازمان، بدون آنکه لزوما منجربه عواقب قانونی مستقیم شود" همچنین وانگ[2] و همکارانش آن-را" یک اقدام عمدی که در تضاد با قانون، حکومت یا سیاست است به منظور بدست آوردن سود غیر مجاز مالی" تعریف کنند[1].

۲-۲ تقلب مالی

تقلب به یک مساله مهم و اصلی برای بسیاری از صنایع تبدیل شده است. در تمام سازمان‌ها مانند بیمه، ارتباطات از راه‌دور و صنایع مرتبط به امور مالی این مساله وجود دارد. در امور دولتی نیز سو استفاده‌هایی از عموم مردم صورت می‌گیرد، تقلب در پرداخت مالیات و تهیه و انجام خدمات دولتی نمونه‌هایی از آن می باشد[1]. هر ساله با افزایش تقلبات مالی، این مساله به عنوان راهی برای کسب درامد تبدیل شده است. ٪۴۵ از شرکت‌ها و سازمان‌ها در سراسر جهان در سال ۲۰۰۴ و ۲۰۰۵ قربانی جرایم اقتصادی شدند. بطور متوسط زیانی که شرکت‌ها از تقلب دچار شدند نزدیک به ۱/۷ میلیون دلار بوده است. با توجه به تیتر مجلات و موضوعات خبری رفتار تقلب روندی افزایشی را به خود گرفته است[2]. کیگان[3] [3] در مطالعه‌ای که بروی ۹۰۰ شرکت در سراسر جهان در سه سال ۲۰۰۶ تا ۲۰۰۸ صورت داده بود بیان کرد که، بطور متوسط ۸/۲ میلیارد دلار در هر سال شرکت‌ها دچار زیان

[1] Phua
[2] Wang
[3] Kaygan

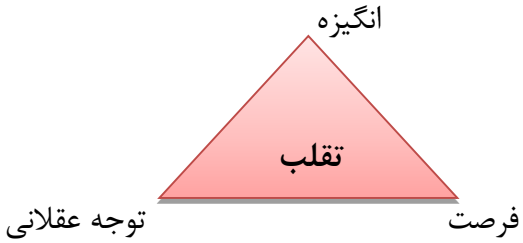
مالی شده‌اند که افزایش ٪۲۲ نسبت به بررسی منتشر شده را نشان می‌دهد. در سال ۲۰۰۸ شرکت-هایی که حداقل یک‌بار تقلب در آنها صورت گرفته بود به ٪۸۵ رسیده که افزایش ۸۰ درصدی نسبت به سال قبل را داشته است. با توجه به اخبار و اطلاعاتی که امروزه منتشر می‌شود، افراد متعددی دچار ضررهای هنگفت، بر اثر تقلب‌های مالی شده‌اند. رئیس سابق نزدک برنارد مادوف[1] که باعث خسارت قریب به ۵۰ میلیارد دلار در سراسر جهان شده، همچنین با توجه به گزارش اخبار بی‌بی‌سی در سال ۲۰۰۷، ارزش مطالبات جعلی بیمه‌گرهای انگلستان در مجموع ۱/۶ میلیارد پوند در سال است. مثالی دیگرجوزف هریکو[2] مدیر عامل سابق شرکت انرون برود بند سرویس (EBS[3]) است که بعد از پذیرفتن اتهام تقلب، زیان نزدیک به ۸/۷ میلیون دلار را اعتراف کرد[1]. مجموع تقلبات گزارش شده در سال ۲۰۰۵ در آمریکا، ۲,۷ میلیارد دلار بوده که در سال ۲۰۰۷ به ۳/۲ میلیارد دلار رسیده است [4]. در سال ۲۰۰۵ دو شرکت ویزا[4] و مسترکارت[5] که از بزرگترین شرکت‌های کارت اعتباری در دنیا به شمار می‌آیند،۱/۱۴ میلیارد دلار بر اثر تقلب، متضرر شده‌اند که نسبت به سال ۱۹۹۵، ٪۶۲/۹ افزایش داشته است. در انگلستان، تقلب در کارت‌های اعتباری یکی از جرایمی به حساب می‌آید که به شدت در حال رشد است. در سال ۲۰۰۸ ، ۶۰۹ میلیون پوند بر اثر تقلب در کارت‌های اعتباری این کشور دچار زیان شده است که ۵۲/۵ میلیون آن بطور خاص تقلب آنلاین بوده است (انجمن پرداخت پاک[6]۲۰۰۹)[3]. مسلما مقادیر بالا را می‌توان تا حدی بوسیله افزایش تراکنش‌ها و یا افزایش چشمگیر کارت‌های اعتباری در سال‌های گذشته دانست ولی باید توجه داشت تقلب مالی یک شکل جدی در حال افزایش است. تقلب می‌تواند طیف نامحدودی از اشکال مختلف را در برگیرد، با این حال در سال-های اخیر با توسعه فن‌آوری های جدید، اشکال سنتی تقلب‌های جعلی از قبیل پول‌شویی به اشکال راحت و جدیدی از تقلب، مانند ارتباطات از راه‌دور تلفن همراه و نفوذ کامپیوتری تبدیل شده، این

[1] Bernard madoff
[2] Joseph Hirko
[3] Enron Broadband Services
[4] visa
[5] Mastercart
[6] Association for Payment Clearing Services

موضوع راه‌های بیشتری را برای تقلب‌کنندگان فراهم ساخته تا بتوانند تقلب صورت دهند، اثرات سازمانی این تقلب‌ها علاوه بر ضررهای اقتصادی اثرات روانی مختلفی را نیز بر روی خود سازمان و مشتریان آن خواهد داشت. با وجود خسارت‌های جبران‌ناپذیری که سازمان‌ها و افراد دچار آن شده‌اند، کشف تقلب به یک موضوع مهم و ارزشمند تبدیل شده است.

۲-۳ مثلث تقلب[1]

دونالد کریسی[2] مدلی را از تقلب تعریف کرد که اولین دیدگاه‌های تعریف تقلب به شمار می‌رود. او علل تقلب را در مدل شناخته شده‌ای به عنوان مثلث تقلب که در شکل زیر آمده است خلاصه کرد[5].



شکل(۲-۱) مثلث تقلب

سه عنصر مثلث تقلب عبارتند از انگیزه یا فشار، فرصت، وتوجیه عقلانی. با توجه به این مثلث، تقلب زمانی رخ می‌دهد که کسی با توجه به نیازهای مالی و یا فشار، در خود انگیزه‌ای می‌بیند تا با استفاده از فرصتی، از طریق راه نادرستی از منابع مالی اقدام کرده و توجیه عقلانی از عمل خود داشته باشد. هنگامی‌که ثبات یا سودآوری مالی بوسیله اقتصاد یا صنعت وجود نداشته باشد و یا فشار بیش از حدی برای پرداخت بدهی‌های موجود بروی مدیران حاصل گردد، انگیزه یا فشاری را می‌بینند که به فعالیت‌های تقلبی روی آورند. هنگامی‌که حسابرسی‌های دقیقی وجود داشته باشد یا نظارت بر گزارشات مالی صورت گیرد، و یا گردش مالی شرکت بالا و بررسی‌های داخلی بی‌اثر باشد، تحقیق و بررسی کار بسیار دشواری خواهد بود، اینجاست که فرصت برای تقلب وجود دارد. هنگامی‌که ارتباطات

[1] Fraud Triangle
[2] Donald Cressey

و حمایت از موسسات مستقل بطور مشهود ناکارامد و نامناسب باشد و یا سابقه تخلف از قانون مشهود باشد آنگاه نگرش نسبت به گزارش جعلی انجام شده توجیه عقلانی به خود می‌گیرد. در ارتباط با تعریف مثلث تقلب می‌توان به این نتیجه رسید که حداقل سه راه برای جلوگیری از تقلب وجود دارد که عبارتند از :

۱. تغییر در انگیزه‌های افراد

۲. محدود کردن فرصت‌ها برای دسترسی مخفیانه به منابع مالی

۳. تضعیف توجیهات عقلانی مشترک و معمول.

آموزش عمومی و یا بازجویی و مجازات افراد نیز می‌تواند به عنوان روشی دیگر برای جلوگیری از تقلب بیان گردد. باید توجه داشت که انگیزه و فرصت یا وجود دارند یا ندارند، درحالی‌که توجیه عقلانی وابسته به افراد و شرایط است و با توجه به آن است‌که اتفاق می‌افتد. فهم و درک مناسب مثلث تقلب در ارزیابی تقلب مالی بسیار ضروری است. زمانی قادر به درک بهتر تقلب مالی خواهیم شد که قادر به درک مفاهیم پایه‌ای مثلث تقلب باشیم و اینکه، چگونه آن‌ها رخ می‌دهند و چه باید کرد تا آن‌ها را متوقف کرد.

۲-۴ چارچوب طبقه‌بندی تقلب مالی

بر اساس چارچوب جرائم مالی دفتر تحقیقات فدرال ایالت متحده[1](اف.بی.آی) ، تقلبات مالی به ۴ دسته تقسیم شده است [1] که عبارتند از ۱. تقلب بانک ۲. تقلب بیمه ۳. تقلب اوراق بهادار و اجناس ۴. سایر کلاه‌برداری مالی از قبیل تقلب صنفی و تقلب بازاریابی و فروش.

[1] Federal Bureau of Investigation

۲-۴-۱ تقلب بانک

تقلب بانک به عقیده CULS[1] اینگونه تعریف شده است :

هر کسی که بطور آگاهانه طرح یا تزویری را برای کلاه‌برداری از یک نهاد مالی یا برای بدست آوردن پول، بودجه، اعتبارات، دارایی، اوراق بهادار، یا سایر اموال متعلق یا تحت حفاظت یا کنترل یک نهاد مالی به وسیله ادعاهای جعلی یا نادرست، تلاش برای اجرای آن، برای خود یا دیگران کند را تقلب بانک گویند. تقلب بانک شامل تقلب در کارت اعتباری، پول شویی و تقلب در رهن و فروش املاک می‌باشد.

۲-۴-۲ تقلب بیمه

تقلب بیمه معمولا شامل تقلب در مراقبت‌های بهداشتی و بیمه خودرو می‌باشد. اف.بی.آی بیان می-کند که تقلب مراقبت‌های بهداشتی توسط بسیاری از بخش‌های نظام بهداشت و درمان به روش‌های مختلفی انجام می‌شود که برخی از انواع بسیار رایج آن عبارتند از: صورت‌حساب برای خدمات ارائه نشده، مطالبات و ادعاهای تکراری، خدمات پزشکی غیر ضروری و غیره . تقلب بیمه خودرو که یکی از بیشترین تقلبات صورت گرفته می‌باشد، یک سری فعالیت‌های جعلی شامل صحنه تصادف، تعمیرات غیر ضروری و آسیب دیدگی‌های شخصی جعلی را در بردارد.

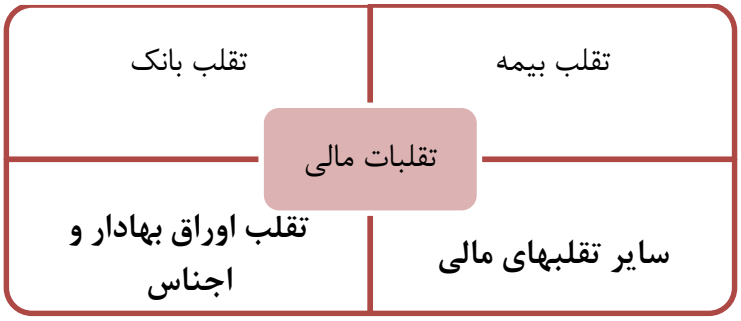
۲-۴-۳ تقلب اوراق بهادار و اجناس

اف.بی.آی شرح مختصری از شایع‌ترین تقلب اوراق بهادار و اجناسی که امروزه با آن مواجه‌ایم را ارائه داده است. برای مثال تقلب در بازار، تقلب سرمایه‌گذاری های پربازده، برنامه‌های هرمی، تقلب کالاها، تقلب ارز خارجی، اختلاس دلالان، نمونه‌هایی از آن می‌باشد.

[1] Connell University Law School

۲-۴-۴ سایر تقلبهای مالی

دسته نهایی از انواع تقلب‌های مالی، غیر از دسته های مذکور را شامل می‌شود. تقلب شرکت‌های بزرگ و تقلب بازاریابی نمونه‌هایی از آن می‌باشند. این چارچوب در شکل زیر آمده است :



شکل(۲-۲) چارچوب طبقه‌بندی تقلبات مالی

۲-۵ داده‌کاوی و روش‌های آماری

تا چند سال پیش کلمه داده‌کاوی فقط برای تعداد کمی از افراد شناخته شده بود. امروزه در بسیاری از مقالات فن‌آوری اطلاعات و مجلات بازاریابی از داده‌کاوی استفاده می شود. داده‌کاوی در سال ۱۹۹۰ معرفی شد اگرچه داده دارای تاریخچه تکامل طولانی می‌باشد[6]. دانش داده‌کاوی یکی از دانش‌های در حال توسعه‌ای است که دهه آینده را با انقلاب تکنولوژیک مواجه خواهد ساخت و بدین رو در سال‌های اخیر در دنیا گسترش فوق العاده سریعی داشته است. امروزه در دنیای توسعه یافته مکان و موضوعی بدون بهره از دانش داده‌کاوی یافت نمی‌شود و در تمامی حوزه‌ها نقش دارد. داده-کاوی اغلب به عنوان یک حوزه چند رشته‌ای نامیده می‌شود، اولین دلیل آن این است‌که، منشا داده-کاوی در سه موضوع مختلف است. آمار، هوش مصنوعی[1]، و یادگیری ماشین داده‌کاوی را تشکیل می-دهند که در آن روش‌های آماری پایه و اساس روش‌های مورد استفاده در داده‌کاوی می باشند. هوش مصنوعی نیز در داده‌کاوی تلاش می‌کند تا با استفاده از روش‌ها و فن‌آوری‌های هوشمند مسائل پیش-رو را حل کند. سومین جز داده‌کاوی که منجر به تکامل آن شده، یادگیری ماشین می‌باشد که به

[1] Artificial intelligence

عنوان پیونددهنده دو جز دیگر داده‌کاوی یعنی آمار و هوش مصنوعی بکار می‌رود. یادگیری ماشین شامل آمار و هم‌چنین استفاده از فن‌آوری هوشمند پیشرفته و الگوریتم‌هایی از هوش مصنوعی می-باشد[6]. می‌توان گفت که هر روندی در داده‌کاوی به نحوی می‌تواند یک یا ترکیبی از این سه جز را در برگیرد. دومین دلیل را پائولو گودیچی[1] این‌گونه بیان می‌کند : داده‌کاوی در یک فرایند معتبر، در یک شرکت، شامل حداقل سه فرد متفاوت با مهارت قوی می‌باشد: متخصصین بازاریابی و کسب و کار، برای مجموعه‌ای از اهداف و تفسیر نتایج داده‌کاوی. متخصصین فناوری اطلاعات که در مورد داده‌ها و فن آوری‌های مورد نیاز دانش کافی دارند و متخصصین روش‌های آماری برای تجزیه و تحلیل داده‌ها و روش‌های مورد بررسی[7]. هدف هر پروژه داده‌کاوی آشکار ساختن الگوها و وابستگی در داده‌هایی که قبلا شناخته نشده‌اند به منظور تصمیم‌گیری مفید بر اساس این یافته‌ها می‌باشد. امروزه با توجه به تنوع زیاد مخاطبین، مشتریان، بازارها، تنوع و پیچیدگی خدمات و محیط‌های کسب و کار، دسترسی به اطلاعات مناسب برای تصمیم‌گیری صحیح ضروری می‌باشد. از این‌رو استفاده از راه‌کارهای مناسب برای طبقه‌بندی و تولید اطلاعات از میان انبوهی از داده‌ها برای سازمان‌ها امری ضروری و حیاتی است. علم داده‌کاوی پاسخی به این نیاز به منظور کشف دانش پنهان داده‌ها و تامین اطلاعات مورد نیاز مدیران می‌باشد. دانش داده‌کاوی ابزاری است که مدیران را قادر می‌سازد تا سریع‌تر نسبت به آینده عمل نمایند، فعال باشند بجای آنکه واکنش‌پذیر باشند و بدانند و مطمئن باشند بجای آنکه حدس بزنند. واضح است که با داشتن این قابلیت، سازمان‌ها قادر می‌باشند دانش ارزشمندی را از داده‌هایشان بدست آورند. دانش داده‌کاوی سازمان‌ها را قادر می‌سازد تا از سرمایه داده‌هایشان بهره-برداری نمایند. این ابزار برای پشتیبان فرایند تصمیم‌گیری استفاده می‌گردد. داده‌کاوی با پردازش جامع داده و انجام فرایند تصمیم‌سازی از طریق استخراج دانش با ارزش از داده، تصمیم‌گیری را برای مدیران سازمان تسهیل می‌نماید.

---

[1] Paolo Giudici

۲-۶ داده کاوی از نظر محققین

تعاریف متعددی از داده‌کاوی وجود دارد، برخی از محققین داده‌کاوی را این‌گونه تعریف کرده‌اند[1]: باس و ماهاپاترا[1] [8] داده‌کاوی را "فرآیند تشخیص الگوهای جالب توجه در پایگاه داده، که می‌تواند در تصمیم‌گیری مورد استفاده قرار گیرد" تعریف می‌کند. توربان[2] و همکاران [9] داده‌کاوی را "فرآیندی که با استفاده از آمار، ریاضی، هوش مصنوعی و تکنیک‌های یادگیری ماشین، اطلاعات مفید را از یک پایگاه داده بزرگ استخراج و شناسایی می‌کند و سپس به نتیجه دلخواه می‌رسد" تعریف می‌کند. فراولی[3] و همکاران [10] بیان می‌کنند که هدف از داده‌کاوی بدست آوردن اطلاعات مفید و غیر صریح از داده‌های ذخیره شده در مخازن بزرگ است. کو[4] و همکاران [11] نیز تایید کرد که مزیت مهم داده‌کاوی این است که می تواند برای توسعه یک دسته از مدل‌های جدید پیش از آنکه توسط کارشناسان انسانی تشخیص داده شوند مورد استفاده قرار می‌گیرد. فوآ و همکاران [12] اشاره می‌کنند تشخیص تقلب به یکی از معروف‌ترین کاربردهای داده‌کاوی هم در صنعت و هم در حکومت تبدیل شده است. با این وجود، داده کاوی را می توان به عنوان تکنیکی که با استفاده از روش‌های یادگیری ماشین و تجزیه و تحلیل بر پایه روش‌های آماری، بر حجم عظیمی از داده‌ها، تعریف کرد.

۲-۷ اهمیت استفاده از روش های داده‌کاوی

زیان کلی ناشی از تقلب مالی بی‌شمار است، تشخیص تقلب مالی برای پیشگیری از اغلب عواقب ویران‌گر تقلب مالی بسیار حیاتی است. تشخیص تقلب مالی شامل تشخیص اطلاعات مالی جعلی از اطلاعات مالی معتبر می‌باشد، در نتیجه رفتارها یا فعالیت‌های جعلی را آشکار می‌کند و تصمیم گیرندگان را قادر به توسعه استراتژی‌های مناسب برای کاهش اثر تقلب می‌سازد. داده‌کاوی نقش مهمی در تشخیص تقلب مالی دارد که اغلب شامل استخراج و کشف حقایق پنهان در مقادیر بسیار

[1] Bos and Mahapatra
[2] Turban
[3] Frawley
[4] Kou

زیاد داده است. همچنین با رشد جمعیت و حجم عظیمی از اطلاعات و تراکنش‌ها از مشتریان، پردازش این اطلاعات کار دشوار و زمان‌بری خواهد بود. در طول دهه گذشته حجم زیادی از داده‌ها در پایگاه داده‌ها انباشته و ذخیره شده‌اند و نتیجه این انباشتگی این است که سازمان‌ها در داده غنی ولی در کسب دانش بسیار ضعیف می‌باشند. امروزه میزان داده‌های در دسترس هر ۵ سال دو برابر می‌شود و سازمانی توانا است که قادر باشد حداقل ۷ درصد از اطلاعاتش را مدیریت نماید. تحقیقات انجام یافته نشان از آن دارد که سازمان‌ها امروزه کمتر از یک درصد از داده‌هایشان را برای تحلیل استفاده می‌نمایند. به‌عبارت دیگر امروزه سازمان‌ها در اطلاعات غرق شده‌اند در حالی‌که گرسنه دانش هستند چرا که سازمان‌ها داده‌های زیادی را در تصرف خود دارند درحالی‌که هنوز با فقدان دانش پنهان درون داده‌ها مواجه هستند. روش‌های آماری مبتنی بر یادگیری ماشین با استفاده از ابزارهای پیشرفته این مهم را فراهم می‌کنند. همچنین تکنیک‌های داده‌کاوی از ابزارهای مهمی است که در پردازش حجم عظیمی از داده‌ها مورد استفاده قرار می‌گیرد. با وجود انجام پردازش توسط ابزارهای داده‌کاوی باید ارتباط معناداری از اعداد و رقم‌های به وجود آمده پیدا کرد، اینجاست که ما نیاز به علم آمار داریم. این علم آمار است که به تمام اعداد و پردازش‌های بدست آمده معنی و مفهوم می‌بخشد و الگوهای مناسبی را برای پیشگیری تقلب به کاربران ارائه می‌دهد.

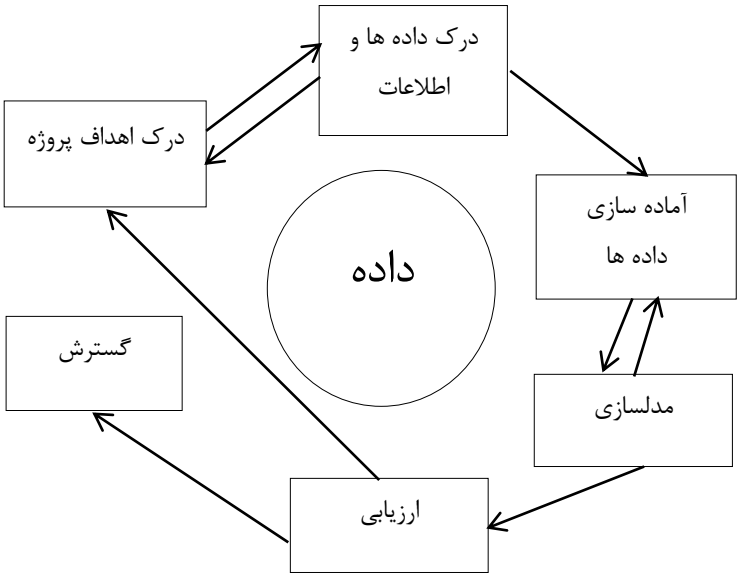
همان طور که بیان شد تقلب هر ساله در حال افزایش است و بر اثر این تقلبات هزینه های جانبی هنگفتی نیز در پی خواهد داشت که می‌توان چند نمونه آن را به این صورت نام برد:

۱. از دست دادن مشتریان بر اثر نداشتن امنیت خاطر نسبت به حساب خود
۲. اثر بازدارنده در گسترش تجارت الکترونیک
۳. هزینه ای که برای تشخیص و پیشگیری باید صرف کرد
۴. هزینه راه‌اندازی و نصب سیستم های تشخیص و پیشگیری تقلب
۵. هزینه تجارت از دست رفته برای مثال هزینه جایگزین کردن کارتهای به سرقت رفته

با وجود ضررهای مالی هنگفتی که اشاره شد، اهمیت این موضوع برای سازمان‌ها دو چندان شده است. امروزه با پیشرفت تکنولوژی و با استفاده از روش‌های پیشرفته آماری، کشف تقلب مالی موثر گشته است، براین اساس ضرورت دانستن روشهای داده‌کاوی که اساس آن یادگیری آماری است بیشتر از قبل شده است.

۲-۸ مراحل داده کاوی

همان طور که در شکل زیر مشاهده می شود فرایند داده کاوی ۶ مرحله دارد[6].



شکل(۲-۳) مراحل داده کاوی

مرحله اول در داده‌کاوی درک اهداف پروژه می‌باشد. در این مرحله به دنبال این هستیم که چگونه می‌توان سوالات موجود را حل کرد و اینکه چگونه می توان آن را در مسائل داده‌کاوی آورد. در زمینه کشف تقلب سوال این است‌که چگونه موارد غیرجعلی را از موارد جعلی تفکیک کنیم؟ مرحله دوم، درک داده‌ها و اطلاعات می‌باشد که شامل مجموعه‌ای از داده‌ها و همچنین آشنایی با مفهوم آن می-باشد. درک معانی و مفاهیم هر یک از متغیرها برای تمام فرایند بسیار ضروری است. در مرحله سوم که به عنوان آماده‌سازی داده‌ها نام دارد، داده‌های جمع‌آوری شده خام (اولیه)، با توجه به اهداف خاص پروژه، ساختارسازی، فیلتر، تمیز و تبدیل شده و در آخر از روش‌های داده‌کاوی مورد نیاز استفاده می-

شود. انتخاب متغیرها و هم‌چنین تغییرات دیگر بر روی متغیرها نیز ممکن است بسته به روش‌های استفاده شده برای یادگیری مورد استفاده قرار گیرند. مرحله چهارم، مدل‌سازی می‌باشد که داده‌کاوی بصورت واقعی در این مرحله صورت می‌پذیرد. بر اساس داده‌های آماده‌شده، پیش‌بینی یا توصیف مدل داده‌ها با توجه به الگوریتم انتخاب شده صورت می‌گیرد. این انتخاب بسیار مهم است و بر عملکرد مدل ارائه شده تاثیرگذار خواهد بود. برای مثال برخی از روش‌های طبقه‌بندی قابلیت اجرا با داده‌هایی که دارای واریانس بالایی می‌باشند را دارند و بعضی از آن‌ها به داده‌های پرت و یا به توزیع کلاس‌های چوله بسیار حساس می‌باشند. مرحله پنجم، ارزیابی است. با توجه به مرحله قبل مدلی برای داده‌ها ارائه و تفسیر شد، حال باید دید آیا آن مدل پاسخگوی سوالات پروژه می‌باشد یا فقط الگویی شناخته-شده را ارائه داده است ؟ آیا توانایی پیش‌بینی و کمک به اتخاذ تصمیمی درست در کسب‌وکار را دارد؟ در این مرحله به ارزیابی آنچه مدنظر ما بوده است و نتایجی که در مرحله قبل بدست آمده است می-پردازیم. مرحله آخر در فرایند داده‌کاوی گسترش[1] (صف آرایی) می‌باشد. گسترش مرحله ایستکه در آن مدلی جامع و کامل ارائه می‌شود. در حوزه کشف تقلب ، این به معنی اجرای قوانین ایجاد شده توسط روش‌های طبقه‌بندی در فرایند بازاریابی یا کسب‌وکار خواهد بود. داده‌کاوی فرایندی پویاست، بنابراین رفتن به مراحل قبلی در آن امکان‌پذیر است. اگر مدل دارای عملکرد مناسبی نباشد این فرایند باید تکرار شود. الگوریتم‌های داده‌کاوی مختلف نیاز به داده‌هایی با شکل‌های متفاوت هستند بنابراین آماده‌سازی داده‌ها ممکن است چندین بار تکرار شده، یا ممکن است شما تشخیص دهید که این مرحله هرگز نمی‌تواند پاسخگویی سوالات تعریف شده در مراحل قبل باشد. در این صورت با توجه به اهداف مورد نظر ممکن است مجموعه داده‌ها تعدیل شده یا تعداد متغیرهای آن افزایش داده شود. برای ارزیابی نتایج در روش‌های داده‌کاوی، ابتدا داده‌ها را بطور تصادفی به دو دسته بعنوان داده‌های مدل‌ساز و داده‌های آزمون تقسیم می‌کنیم. سپس برای تجزیه و تحلیل داده‌ها، ابتدا مدل‌های موجود

[1] Deployment

را بر روی داده‌های مدلساز برازش داده و سپس با استفاده از مدل ارائه شده با توجه به داده‌های مدل-ساز، عملکرد مدل‌ها را توسط داده‌های آزمون مورد ارزیابی قرار می‌دهیم.

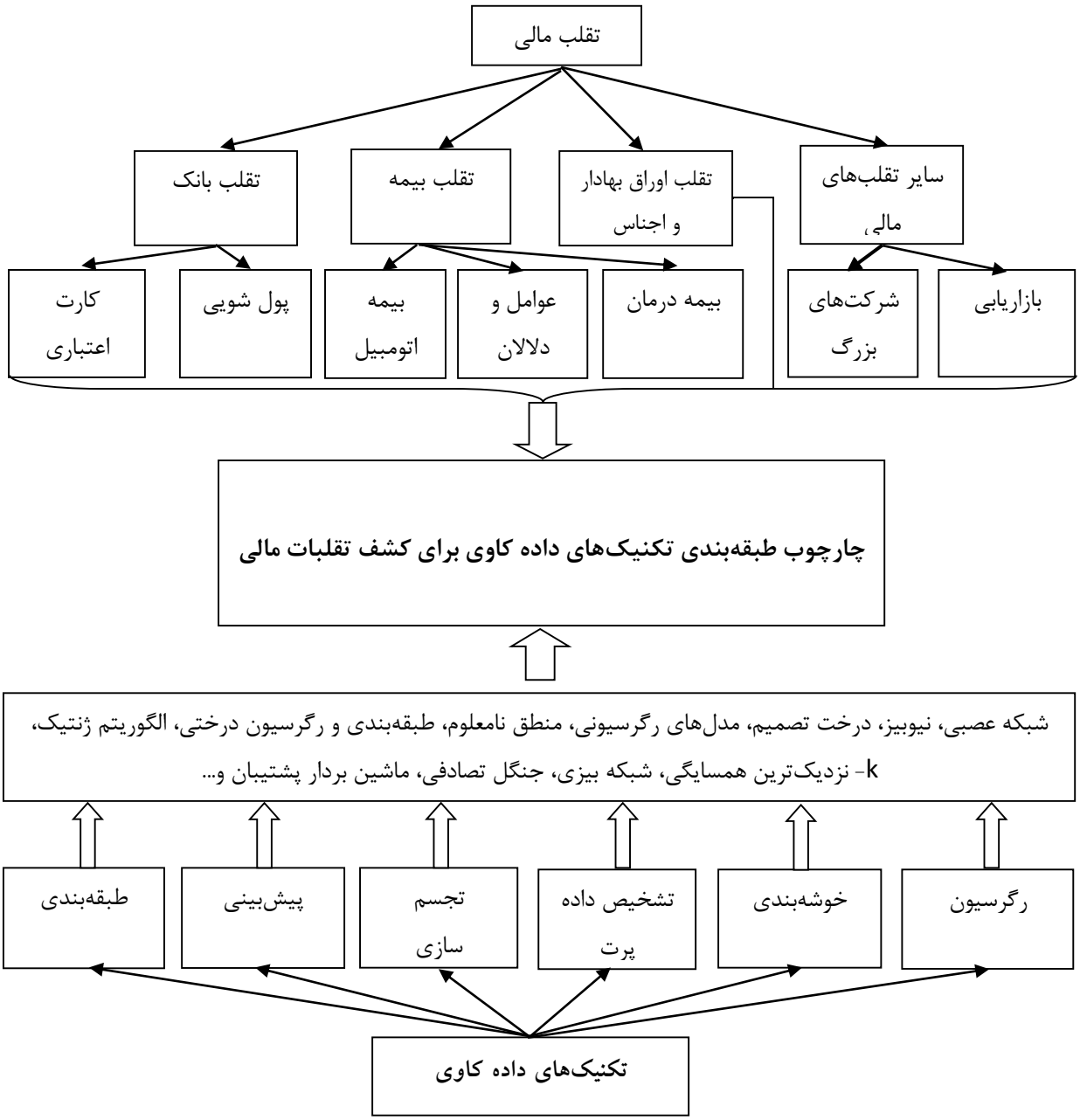
۲-۹ روش‌های کشف تقلب

تکنیک‌های داده‌کاوی کشف تقلب به دو دسته روش‌های باراهنما[1] و روش‌های بدون‌راهنما[2] تقسیم می‌شوند. در روش‌های باراهنما ، بر اساس معاملات جعلی و غیرجعلی مدلی فراهم می‌شود و با استفاده از آن مدل، حالت‌های جدید را پیش‌بینی می‌کنیم ولی در روش بدون‌راهنما ، معاملات جعلی و غیر جعلی را در پایگاه داده نداریم و با توجه به روش‌های موجود، الگویی مناسب برای کشف تقلب ارائه خواهد شد. در شکل زیر، چارچوبی از تکنیک‌های داده‌کاوی در کشف تقلب مالی را آورده‌ایم. این شکل بر پایه شکل ۲-۲ می‌باشد، که در طبقه‌بندی اول، همان‌طور که در شکل۲-۲ آورده شد تقلبات مالی به ۴ دسته تقسیم شده‌اند، و در سطح دیگر هر یک از انواع تقلبات مالی به فعالیت‌های جعلی که در آنها صورت می‌گیرد دسته‌بندی شده است. باید توجه داشت این دسته‌بندی بر اساس بیشترین فعالیت‌های جعلی صورت گرفته در هر یک از طبقات تقلبات مالی می‌باشد. در طبقه‌بندی دوم، ابتدا روش‌های داده کاوی طبقه‌بندی شده و سپس الگوریتم‌های موجود در داده‌کاوی برای کشف تقلب آورده شده است که در ادامه به ۳ روش از آن خواهیم پرداخت.

حال این سوال مطرح می‌شود که اگر داده‌کاوی روشی بسیار پرکاربرد است چرا هنوز تقلب وجود دارد؟. اول این‌که بهترین نتایجی که در کشف تقلب بدست می‌آید بوسیله بکار بردن روش‌های باراهنما در داده‌کاوی می‌باشد که در آن تقلب‌های صورت گرفته مشخص است و داده‌کاوی به ارائه مدلی می‌پردازد که دارای بیشترین دقت در کشف تقلب است. داشتن داده‌هایی که برچسب دارند دیدگاه واقع گرایانه‌ای در بسیاری از مشکلات شرکت‌ها نیست و سازمان‌ها و شرکت‌ها مشکل عمده‌ای که سروکار دارند بوجود آمدن تقلب و جلوگیری از آن است. دلیل دوم در دسترس نبودن اطلاعات

[1] supervised methods
[2] unsupervised methods

می‌باشد که با وجود اهمیت این موضوع ، مدیران سازمان‌ها به‌دلیل از دست ندادن اعتبار سازمانشان، حاضر به ارائه اطلاعات در این زمینه نیستند، به‌همین علت نمی‌توان تقلب را بطورکامل بررسی کرد.



شکل(۲-۴) چارچوب طبقه‌بندی تکنیک‌های داده‌کاوی برای کشف تقلبات مالی

در این پایان‌نامه نیز به‌دلیل نداشتن مجموعه داده واقعی، تنها به ارائه روش‌هایی موثر در داده‌کاوی پرداخته و با مجموعه داده‌ای غیرواقعی به ارزیابی مدل‌های بیان شده خواهیم پرداخت.

۲-۱۰ مروری بر تحقیقات گذشته

در سال‌های اخیر مقالات مختلفی در زمینه کشف تقلب منتشر شده است و افراد متعددی در مقالات خود از تکنیک‌های داده‌کاوی در تشخیص تقلب مالی استفاده کرده‌اند، برای مثال، راویسانکار[۱] و همکاران با استفاده از شش روش شبکه عصبی چندلایه مقدم[۲]، ماشین بردار پشتیبان[۳]، برنامه‌نویسی ژنتیکی[۴]، روش‌های گروهی داده‌گردانی[۵]، رگرسیون لجستیک و شبکه عصبی احتمالاتی[۶]، به بررسی تشخیص تقلب مالی پرداخته‌اند [13]، بولتون و هند[۷] روش‌های آماری تشخیص تقلب در کارت اعتباری، پول‌شویی، تقلب از راه دور و... را بررسی کرده‌اند [14]، باتاچاریا[۸] و همکاران با استفاده از روش‌های آماری و مقایسه سه تکنیک رگرسیون لجستیک، ماشین بردار پشتیبان و جنگل تصادفی[۹]، به تجزیه و تحلیل در کشف تقلب در کارت اعتباری پرداخته‌اند [15]. ژو و ژانگ[۱۰] کاربردهای مالی داده‌کاوی شامل بازار سهام، پیش بینی ورشکستگی و تشخیص تقلب را مورد مطالعه قرار داده-اند[16]. همچنین فوآ و همکاران یک تحقیق از تشخیص تقلب با استفاده از داده‌کاوی، شامل تقلب معاملات در کارت‌های اعتباری، تقلب در بیمه خودرو را بررسی کرده‌اند [12]. نیل آدامز[۱۱] تحلیل جامع بر روی موضوع کشف تقلب در کارت‌های اعتباری با استفاده از ۷ روش از تکنیک‌های داده‌کاوی صورت داده است[17]. این روش‌ها عبارت بودند از، رگرسیون لجستیک، ماشین بردار پشتیبان، تحلیل ممیزی، جنگل تصادفی، طبقه‌بندی و رگرسیون درختی، بیزی خام و k-نزدیکترین همسایگی. همچنین سانچز[۱۲]و همکارانش با استفاده از قوانین پیوند[۱۳] به این موضوع پرداخته‌اند[18]. پانیگراهی[۱]

[1] Ravisankar
[2] Multilayer feed forward neural network
[3] Support vector machines
[4] Genetic programming
[5] Group method of data handling
[6] Probabilistic neural network
[7] Bolton and Hand
[8] Bhattacharyya
[9] Random farest
[10] zhang and zhou
[11] Niall Adams
[12] Sanchez
[13] Association rules

پانیگراهی[1] و همکارانش نیز با استفاده از قضیه دمستر- شفر[2] موضوع تقلب در کارت‌های اعتباری را بررسی کردند[4]. با استفاده از روش سیستم مقاوم مصنوعی[3] نیز تحقیقاتی در این زمینه انجام شده است‌که می‌توان به گادی[4] و همکارانش[19] و وانگ[5] و همکارانش اشاره نمود[20]. یکی از بهترین تحقیقات مروری انجام شده، توسط ان‌گای[6] و همکارانش [1] بر روی ۴۹ مقاله ژورنالی در بین سال-های ۱۹۹۷ تا ۲۰۰۸ صورت گرفته، که تمام آنها در زمینه کشف تقلبهای مالی با استفاده از روش‌های داده‌کاوی صورت گرفته است. از آنجائیکه در این پایان‌نامه به مطالعه روش‌های رگرسیون لجستیک(LR)، بیزی خام(NB) و شبکه بیزی(BN) پرداخته شده، لذا فهرستی از مقالاتی که در آن‌ها از این سه روش برای کشف تقلب استفاده گردیده‌اند در جدول ۲-۱ آمده است.

جدول(۲-۱) تحقیقات انجام شده در کشف تقلب با استفاده از روش‌های LR,NB,BN

مقالات مورد بررسی	روش مورد استفاده برای کشف تقلب
[4],[21],[22],[23],[15]	LR
[24]	NB
[25],[26],[27],[28]	BN
[3],[6],[17]	LR,NB
[14]	LR,BN
[29],[30],[31]	BN,NB
[32],[33],[34]	LR,NB,BN

[1] Panigrahi
[2] Dempster-Shafer
[3] Artificial immune systems
[4] Gadi
[5] Wong
[6] Ngai

فصل سوم

رگرسیون لجستیک و بیزی خام

۳-۱ مقدمه

در فصل‌های قبل تعاریفی از داده‌کاوی و تقلبات مالی را بیان کردیم و به اهمیت روش‌های داده‌کاوی در کشف تقلبات مالی پی بردیم. در فصل ۳ و ۴ روش‌هایی موثر در ارتباط با این موضوع را مطالعه می‌کنیم. روش اول، رگرسیون لجستیک می‌باشد که با وجود متداول بودن آن دارای اهمیت فراوانی است. روش دوم، طبقه‌بندی‌کننده بیزی خام است‌که تحت شرایط خاصی فرم پارامتری آن با رگرسیون لجستیک یکسان خواهد شد. سرانجام شبکه بیزی که روشی برای بهبود طبقه‌بندی‌کننده بیزی خام است را در فصل چهارم بیان خواهیم کرد.

۳-۲ رگرسیون لجستیک

در روش‌های مدل‌سازی، هنگامی‌که مقادیر متغیر پاسخ، گسسته است دیگر نمی‌توان از روش‌های رگرسیون خطی استفاده کرد. روش رگرسیون لجستیک توانایی مدل‌سازی این‌گونه مسائل را برای متغیرهای توضیحی پیوسته و گسسته داراست. در این‌جا ابتدا مدل رگرسیون خطی را به صورت مختصر توضیح می دهیم و سپس به ارائه مدل رگرسیون لجستیک خواهیم پرداخت. مدل رگرسیون خطی تابعی پارامتریک است و به صورت زیر بیان می شود :

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \rightarrow \quad E(Y = y|X = x) = \beta_0 + \beta_1 x \quad \rightarrow \quad Y = E(Y = y|x) + \varepsilon$$

که در آن β_0, β_1 پارامترهایی هستند که به روش مینیمم‌سازی مجموع مربعات خطا برآورد شده و X متغیر توضیحی مدل (متغیر ورودی) با مقدار مشاهده شده x و ε متغیر تصادفی دارای توزیع نرمال با میانگین صفر و واریانس σ^2 می‌باشد. رگرسیون لجستیک روشی است برای یادگیری تابعی به فرم $f: X \to Y$ یا $P(Y = y \mid X = x)$ که در آن Y متغیر پاسخ با مقادیر گسسته و $X =< X_1, \dots, X_n >$ برداری از متغیرهای پیوسته یا گسسته می‌باشد. فرض می‌کنیم Y مقادیر ۰ و ۱ را شامل می‌شود که در بحث کشف تقلب، تقلب را به ۱، و مشاهداتی که در آن‌ها تقلبی صورت نپذیرفته است را به ۰ اختصاص می‌دهیم.

توزیع $P(Y = y \mid X = x)$ ، مانند رگرسیون خطی، فرمی پارامتری بصورت زیر را دارد[16] :

$$\pi(x) = E(Y|X = x) = P(y = 1|X = x) = \frac{1}{1 + exp(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)} \qquad (1-3)$$

و

$$1 - \pi(x) = P(y = 0|X = x) = \frac{\exp(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)} \qquad (2-3)$$

بنابراین $(Y|x)$ دارای توزیع برنولی با پارامتر $\pi(x)$ می‌باشد.

۳-۲-۱ برآورد پارامتر

همانند روش رگرسیون خطی می‌توان مدل رگرسیون لجستیک را به صورت زیر بیان کرد:

$$Y = E(Y|x) + \varepsilon \rightarrow Y = \pi(x) + \varepsilon$$

با توجه به توزیع $Y \mid X$ ، از روش درستنمایی ماکزیمم برای برآورد پارامترها استفاده می‌کنیم. از آنجائی‌که $Y \mid x \sim B(1, \pi(x))$ است، لذا داریم [35]:

$$P(Y = y_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

در این صورت تابع درستنمایی بر اساس یک نمونه تصادفی n تایی مستقل عبارتست از:

$$L(\boldsymbol{\beta} \mid x) = \prod_{i=1}^{n} P(Y = y_i)$$

که در آن $\boldsymbol{\beta} = (\beta_0, \beta_1)$ می‌باشد.

با لگاریتم گرفتن از طرفین داریم:

$$\ln L(\boldsymbol{\beta} \mid x) = \sum_{i=1}^{n} \{y_i \ln(\pi(x_i)) + (1 - y_i) \ln(1 - \pi(x_i))\} \qquad (3-3)$$

با جایگذاری رابطه ۳-۱ داریم:

$$lnL = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{1}{1+\exp(\beta_0 + \sum_{i=1}^{n}\beta_i X_i)}\right) + (1-y_i)\ln\left(\frac{\exp(\beta_0 + \sum_{i=1}^{n}\beta_i X_i)}{1+\exp(\beta_0 + \sum_{i=1}^{n}\beta_i X_i)}\right)\right\}$$

$$= \sum_{i=1}^{n} \left\{ -y_i \ln\left(1+\exp\left(\beta_0 + \sum_{i=1}^{n}\beta_i X_i\right)\right) + \ln\left(\frac{\exp(\beta_0 + \sum_{i=1}^{n}\beta_i X_i)}{1+\exp(\beta_0 + \sum_{i=1}^{n}\beta_i X_i)}\right) - y_i \ln\left(\frac{\exp(\beta_0 + \sum_{i=1}^{n}\beta_i X_i)}{1+\exp(\beta_0 + \sum_{i=1}^{n}\beta_i X_i)}\right)\right\}$$

$$= \sum_{i=1}^{n} \left\{ \ln\left(\frac{\exp(\beta_0 + \sum_{i=1}^{n}\beta_i X_i)}{1+\exp(\beta_0 + \sum_{i=1}^{n}\beta_i X_i)}\right) - y_i \ln\left(\exp\left(\beta_0 + \sum_{i=1}^{n}\beta_i X_i\right)\right)\right\}$$

$$= \sum_{i=1}^{n} \left\{ \ln\left(\exp\left(\beta_0 + \sum_{i=1}^{n}\beta_i X_i\right)\right) - y_i \ln\left(\exp\left(\beta_0 + \sum_{i=1}^{n}\beta_i X_i\right)\right) - ln\left(1+\exp\left(\beta_0 + \sum_{i=1}^{n}\beta_i X_i\right)\right)\right\}$$

$$= \sum_{i=1}^{n} \left\{ (1-y_i)\left(\beta_0 + \sum_{i=1}^{n}\beta_i X_i\right) - \ln\left(1+\exp\left(\beta_0 + \sum_{i=1}^{n}\beta_i X_i\right)\right)\right\} \qquad (4-3)$$

حال با مشتق‌گیری از ۳-۴ نسبت به پارامترها داریم :

$$\frac{\partial lnL}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \qquad (5-3)$$

$$\frac{\partial lnL}{\partial \beta_i} = \sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0 \qquad (6-3)$$

باید به این نکته توجه داشت که در مدل رگرسیون لجستیک ارتباط Y و X غیرخطی است و حل معادلات فوق مانند روش رگرسیون خطی نمی‌باشد و لذا باید برای حل معادلات فوق از روش‌های حل

عددی استفاده نمود. پس از برآورد پارامترهای مورد نظر، حال می‌توان برای اختصاص دادن هر مشاهده به یکی از مقادیر ۱ یا ۰، بیشترین احتمال $P(Y = y_i \mid X = x)$ را در نظرگرفت که در آن y_i دارای مقادیر ۰ یا ۱ می‌باشند. بنابراین y را به ۰ نسبت می‌دهیم هرگاه :

$$P(Y = 0|x) > P(Y = 1|x) \rightarrow \frac{P(Y = 0|x)}{P(Y = 1|x)} > 1 \quad \Rightarrow \quad \exp\left(\beta_0 + \sum_{i=1}^{n} \beta_i X_i\right) > 1$$

که اگر از طرفین نامساوی فوق لگاریتم بگیریم خواهد شد:

$$\beta_0 + \sum_{i=1}^{n} \beta_i X_i > 0 \Rightarrow \ln \frac{1-\pi}{\pi} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i > 0$$

بنابراین قاعده تصمیم برای پیش‌بینی کلاس مورد نظر بر اساس روش رگرسیون لجستیک به صورت زیر می‌باشد:

$$y = 1 \qquad if \quad \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_i X_i < 0$$

$$y = 0 \qquad if \quad \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_i X_i > 0 \qquad (7-3)$$

در بخش بعد، ابتدا طبقه‌بندی کننده بیزی را یادآوری نموده و در بخش ۳-۴ به بیان روش بیزی خام خواهیم پرداخت.

۳-۳ طبقه بندی بر اساس قانون بیز

در روش‌های طبقه‌بندی، به دنبال طبقه‌بندی متغیر پاسخ بر اساس متغیرهای توضیحی می‌باشیم، یعنی پیش‌بینی می‌کنیم مشاهدات متغیرهای ورودی به کدام کلاس از متغیر پاسخ تعلق دارند. بنابراین فرض کنید متغیر پاسخ Y دارای j حالت از مقادیر گسسته و X نیز برداری شامل n متغیر توضیحی باشد. برای پیش‌بینی مشاهدات به کلاس مناسب، باید $P(Y = y \mid X = x)$ را محاسبه کرده و بر اساس آن تصمیم‌گیری کنیم.

با استفاده از قانون بیز، $P(Y = y_j \mid X = x)$ به صورت زیر می‌باشد :

$$P(Y = y_j|X = x_i) = \frac{P(X = x_i|Y = y_j)P(Y = y_j)}{\sum_l P(X = x_i|Y = y_l)P(Y = y_l)} \qquad (8-3)$$

که در آن y_j، j - امین حالت از مقادیر متغیر پاسخ و x_i، i - امین حالت از بردار مقادیر متغیرهای توضیحی می‌باشد که در۳-۳-۱ منظور از حالت‌های ممکن بیان خواهد شد. همچنین مجموع مخرج، تمام حالت‌های ممکن از مقادیر متغیر پاسخ را در بر می‌گیرد. با فرض اینکه Y دارای مقادیر ۰ یا ۱ است، مشاهدات X را به کلاس ۱ اختصاص می دهیم اگر:

$$P(Y = 1|X) > P(Y = 0|X) \qquad (9-3)$$

در حقیقت اگرچه مخرج ثابت است، تاثیرش در این است‌که برای کار طبقه‌بندی باید صورت‌ها را با هم مقایسه کرد بنابراین داریم :

$$P(Y|X) \approx P(X|Y)P(Y) \qquad (10-3)$$

حال با توجه به مطالب بیان شده در ۲-۸، با استفاده از داده‌های آموزشی یا همان داده‌های مدل‌ساز، احتمال‌های سمت راست را بر اساس فراوانی نسبی مشاهدات محاسبه می‌کنیم.

هنگامیکه تعداد متغیرها زیاد باشد، برای تخصیص مشاهدات به کلاس مورد نظر، تعداد احتمالات مورد نیاز زیاد خواهد شد که محاسبه آن‌ها زمان‌بر است[36]. مشکل دیگر، تزاحم ابعادی[1] است‌که در محاسبه $P(X_1 = x_1, \dots, X_n = x_n|y)$ برای تمام ترکیب‌های ممکن از متغیرهای توضیحی بوجود می‌آید. هنگامی‌که تعداد متغیرها زیاد باشد، حالت‌های متعددی از ترکیب‌های ممکن خواهیم داشت که این امر باعث تزاحم ابعادی می‌شود. در این حالت، روش طبقه‌بندی بر اساس قانون بیز عملکرد مناسبی نخواهد داشت. طبقه‌بندی‌کننده بیزی خام روشی مناسب برای رفع این مشکل

[1] Curse of dimensionality

است. در ادامه، با محاسبه تعداد احتمال‌های مورد نیاز به روش بیزی، خواهیم دید هنگامی‌که تعداد متغیرهای مورد بررسی زیاد باشد طبقه‌بندی توسط قانون بیز بسیار مشکل و زمان‌بر خواهد شد.

۳-۳-۱ محاسبه تعداد پارامترها به روش طبقه‌بندی بیزی

ابتدا حالتی ساده را برای محاسبه تعداد پارامترهای $P(Y = y \mid X = x)$ بیان می‌کنیم. در این بخش احتمالات مورد نیاز برای طبقه‌بندی مشاهدات را به‌عنوان *پارامتر* تعریف می‌کنیم. فرض می‌کنیم Y و هر یک از متغیرهای توضیحی دارای مقادیر بولی هستند. با توجه به معادله ۳-۸، می‌خواهیم $P(Y = y \mid X = x)$ را بدست آوریم. همان طور که بیان شد، با استفاده از داده‌های آموزشی $P(X = x \mid Y = y)$ و $P(Y = y)$ بر اساس فراوانی نسبی مشاهدات در کلاس مورد نظر قابل محاسبه می‌باشند. اما مساله‌ای که اینجا مطرح می‌شود این است‌که احتمالات مورد نیاز برای محاسبه $P(Y = y \mid X = x)$ چه تعداد خواهد بود.

برآورد $P(Y = y)$ به راحتی محاسبه می‌شود، به عبارتی باید بررسی کنیم چه تعداد پارامتر برای برآورد $P(X = x \mid Y = y)$ نیاز است. بنابراین به دنبال برآورد مجموعه پارامتر زیر خواهیم بود[37] :

$$\theta_{ij} \equiv P(X = x_i | Y = y_j) \qquad (11-3)$$

از آنجائی‌که X برداری شامل n متغیر بولی است، تعداد حالتهای ممکن در i، برابر 2^n، و در j برابر با 2 حالت خواهد بود، بنابراین نیاز به محاسبه 2^{n+1} پارامتر داریم. اما باید به این مطلب توجه داشت که، برای هر j ثابت، چون $\sum_i \theta_{ij} = 1$، تنها نیازمند محاسبه $(2^n - 1)$ پارامتر مستقل هستیم و با توجه به دو حالت ممکن Y، مجموع کل پارامترهایی که باید براورد شوند برابر با $2(2^n - 1)$ خواهد بود.

برای برداشت بهتر موضوع مثالی ساده را در این رابطه بیان می‌کنیم . فرض کنید متغیر پاسخ Y و متغیرهای توضیحی X_1, X_2, X_3 دارای مقادیر ۰ و ۱ می‌باشند. بنابراین ۸ حالت برای $Y = 0$ و بطور مشابه ۸ حالت نیز برای $Y = 1$ محاسبه می‌شود که تعداد کل آن برابر $8 + 8 = 2^{3+1} = 16$ حالت است. پارامترهای مورد نظر برای $Y = 0$ عبارتند از :

$$\Theta_{10} = P(X_1 = 0, X_2 = 1, X_3 = 0 | Y = 0) \quad \& \quad \Theta_{20} = P(X_1 = 0, X_2 = 1, X_3 = 1 | Y = 0)$$

$$\Theta_{30} = P(X_1 = 0, X_2 = 0, X_3 = 0 | Y = 0) \quad \& \quad \Theta_{40} = P(X_1 = 0, X_2 = 0, X_3 = 1 | Y = 0)$$

$$\Theta_{50} = P(X_1 = 1, X_2 = 1, X_3 = 0 | Y = 0) \quad \& \quad \Theta_{60} = P(X_1 = 1, X_2 = 1, X_3 = 1 | Y = 0)$$

$$\Theta_{70} = P(X_1 = 1, X_2 = 0, X_3 = 0 | Y = 0) \quad \& \quad \Theta_{80} = P(X_1 = 1, X_2 = 0, X_3 = 1 | Y = 0)$$

همچنین بر اساس اینکه یکی از احتمال‌ها را می‌توان با استفاده از احتمال‌های دیگر بدست آورد، تعداد کل پارامترهای مستقل برابر $7 + 7 = 2(2^3 - 1) = 14$ خواهد شد. لازم به ذکر است‌که اگر بردار X شامل۵۰ متغیر و متغیرها غیربولی باشند نیاز به برآورد میلیاردها پارامتر داریم و روند محاسبه آن بسیار زمان‌بر و دشوار خواهد بود. با وجود این پیچیدگی در روش طبقه‌بندی بیزی، روش طبقه‌بندی کننده بیزی خام راه‌حلی مناسب برای رفع این مشکل بوده و باعث کاهش تعداد پارامترهای مورد نیاز می‌شود.

۳-۴ طبقه بندی‌کننده بیزی خام[1]

الگوریتم‌های بسیاری به منظور طبقه‌بندی داده‌ها در روش‌های یادگیری باراهنما ارائه شده‌اند که می-توان روش طبقه‌بندی کننده بیزی خام (NBC) را یکی از ساده‌ترین آن‌ها دانست. این روش، از فرایند طبقه‌بندی سریعی برخوردار است و توانایی اجرا برای متغیرهای پیوسته و گسسته را داراست. همچنین از نقاط قوت این روش می‌توان به پیش‌بینی و طبقه‌بندی مجموعه داده‌های ناقص اشاره

[1] Naïve Bayes Classifier

نمود[37]. در مجموع می‌توان در روش‌های طبقه‌بندی باراهنما، از روش بیزی خام به عنوان اولین انتخاب برای این‌گونه مسائل یاد کرد.

۳-۴-۱ تعداد پارامترهای مورد نیاز در روش طبقه‌بندی‌کننده بیزی خام

طبقه‌بندی کننده بیزی خام با توجه به فرضیه‌ای که در آن برقرار است، باعث کاهش تعداد پارامترهای مورد نیاز برای برآورد $P(Y = y \mid X = x)$ خواهد شد. این مدل دارای فرضیه استقلال شرطی بین متغیرهای توضیحی و متغیر پاسخ داده شده می‌باشد. قبل از بیان مدل طبقه بندی کننده بیزی خام ابتدا تعریفی از استقلال شرطی را ذکر نموده و سپس به بیان این روش می‌پردازیم.

۳-۴-۲ استقلال شرطی

تعریف: متغیرهای تصادفیX_1, X_2, Y را در نظر بگیرید، با فرض معلوم بودن مقدار Y ، متغیرهای X_1 و X_2 را مستقل شرطی گویند اگر توزیع احتمال X_1 مستقل از X_2 ، نسبت به Y داده شده باشد[16].

$$P(X_1 = x_i | X_2 = x_j \, , \mathrm{Y} = \mathrm{Y}_k) = P(X_1 = x_i | \mathrm{Y} = \mathrm{Y}_k) \qquad \forall\, i, j, k$$

باید توجه داشت زمانیکه متغیرهای توضیحی پیوسته هستند، احتمالات فوق بصورت تابع چگالی احتمال بیان می‌شود. در حالت n متغیر توضیحی، فقط نیاز به محاسبه $2n$ پارامتر است و پیچیدگی روش بیز را نخواهد داشت. بنابراین مدل ۳-۱۰ بصورت زیر بیان می‌شود:

$$P(Y = y_j | X_1 = x_1, \dots, X_n = x_n) = \underbrace{\prod_i P(X = x_i | Y = y_j)}_{n} \underbrace{P(Y = y_j)}_{2} \qquad (12-3)$$

تعداد پارامترهای مورد نیاز در این روش به طور چشمگیری از $2(2^n - 1)$ حالت به $2n$ حالت کاهش داده می‌شود و باعث ساده‌شدن محاسبات برای برآورد $P(Y = y \mid X = x)$ می‌گردد. باید

توجه داشت، تعداد پارامترها در حالت‌هایی که متغیرهای توضیحی گسسته یا پیوسته باشند متفاوت است، که در ادامه به آن پرداخته می‌شود.

۳-۵ طبقه‌بندی کننده بیزی خام با متغیر توضیحی گسسته

در بخش‌های قبل به بیان مدلی پرداختیم که X و Y، مقادیر بولی را شامل می‌شدند. همچنین بر اساس این مقادیر، تعداد پارامترهای لازم برای برآورد احتمال $Y|X$ را در روش بیزی و بیزی خام محاسبه کردیم. حال به بیان کلی روش طبقه‌بندی‌کننده بیزی خام می‌پردازیم.

فرض کنید Y دارای k حالت گسسته و X نیز برداری از n متغیر توضیحی که هر کدام دارای j مقدار گسسته می‌باشد. همان طور که بیان کردیم، به دنبال بدست آوردن $P(Y=y|X=x)$ هستیم. بنابراین برای پیش‌بینی کلاس مربوط به مشاهدات مورد نظر، احتمال $Y|X$ را با توجه به آن مشاهدات محاسبه کرده و سپس بیشترین احتمال را در نظر گرفته و مشاهدات مربوط به متغیرهای ورودی را به آن کلاس اختصاص می‌دهیم.

با توجه به فرضیه استقلال شرطی در این مدل، معادله ۳-۸ بصورت زیر در می‌آید:

$$P(Y=y_k|X_1 \dots X_n)=\frac{P(Y=y_k)\prod_i P(X_i|Y=y_k)}{\sum_j P(Y=y_j)\prod_i P(X_i|Y=y_j)} \quad i=1,\dots,n \qquad (13-3)$$

تساوی ۳-۱۳را تساوی بنیادی طبقه بندی کننده بیزی خام گویند[16]. اگرچه این فرضیه در حالت کلی درست نیست و در طبیعت امکان وابستگی میان متغیرها وجود دارد، اما این فرضیه بطور چشمگیری پیش بینی مشاهدات به کلاس مورد نظر را ساده می‌کند. مشاهده x را به کلاس L نسبت می‌دهیم اگر :

$$L \leftarrow \underset{y_k}{argmax}\frac{P(Y=y_k)\prod_i P(X_i|Y=y_k)}{\sum_j P(Y=y_j)\prod_i P(X_i|Y=y_j)} \qquad (14-3)$$

که در آن $argmax$ بیان‌گر کلاسی از متغیر پاسخ است‌که دارای بیشترین مقدار عبارت فوق باشد. با توجه به معادله ۳-۱۰ ، از آنجائی‌که مخرج معادله ۳-۱۴ به y_k وابسته نیست داریم:

$$Y \leftarrow \underset{y_k}{argmax}\, P(Y = y_k) \prod_i P(X_i|Y = y_k) \qquad (15-3)$$

۳-۵-۱ محاسبه تعداد پارامتر

برای بررسی تعداد پارامترهای مورد نیاز باید دو مجموعه از پارامترها را برآورد کنیم. مجموعه اول برابر خواهد بود با:

$$\Theta_{ijk} \equiv P\big(X_i = x_{ij}\big|Y = y_k\big) \qquad i = 1, \dots, n \qquad (16-3)$$

که در آن برای هر متغیر ورودی X_i ، حالت‌های ممکن از مقادیر x_{ij} و y_k را باید در نظر گرفت. با توجه به توضیحاتی که برای محاسبه تعداد پارامترها در بخش‌های قبل داشتیم، کاملا مشخص است که تعداد پارامترها برابر njk حالت می‌باشد. همچنین باید توجه داشت برای هر جفت مقادیر i و k، از آنجائیکه $\sum_j \Theta_{ijk} = 1$، بنابراین تعداد پارامترهای مجموعه اول $n(j-1)k$ حالت می‌باشد.

مجموعه دوم، پارامترهای احتمال پیشین yاست که باید تعداد آن‌را محاسبه نمود. به عبارتی پارامترهای مجموعه زیر:

$$\pi_k \equiv P(Y = y_k) \qquad k = 1, \dots, K \qquad (17-3)$$

تعداد پارامترهای $17-3$ برابر k حالت است، اما از آنجائی‌که یکی از احتمالات بر اساس احتمالات دیگر محاسبه می‌شود بنابراین تعداد پارامترهای مجموعه دوم $k-1$ حالت می‌باشد. حال با استفاده از روش درستنمایی ماکزیمم به برآورد پارامترهای فوق می‌پردازیم. روش MLE بر پایه محاسبه نسبت فراوانی مقادیر متفاوت متغیرها در داده می‌باشد و با استفاده از نمونه آموزشی مانند D، برآورد درستنمایی Θ_{ijk} به صورت زیر خواهد شد :

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij}\,\Lambda\, Y = y_k\}}{\#D\{Y = y_k\}} \qquad (18-3)$$

که در آن $\#D\{X_i = x_{ij}\,\Lambda\, Y = y_k\}$ تعداد نمونه‌ای با خاصیت $(X_i = x_i\,\Lambda\, Y = y_k)$ و $\#D\{Y =$ $y_k\}$ تعداد نمونه‌ای با خاصیت $Y = y_k$ می‌باشند. براورد درستنمایی π_k نیز بصورت زیر محاسبه می‌شود:

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|} \qquad (19-3)$$

که در آن $|D|$ نشان دهنده تعداد عناصر مجموعه آموزشی می‌باشد[16].

۳-۶ نقاط ضعف طبقه بندی کننده بیزی خام

مدل بیزی خام دارای نقاط ضعفی نیز می‌باشد. این مدل در مجموعه داده‌هایی با حجم و تعداد متغیرهای ورودی بسیار زیاد ، عملکرد مناسبی ندارد و دو محدودیت برای این مدل ایجاد می‌کند. این دو محدودیت بیش برآوردی[1] و پاریز[2] است[38]. پاریز به شرایطی گفته می‌شود که کامپیوتر سعی می‌کند، یک عدد بسیار کوچک نزدیک به صفر را نمایش دهد. برنامه‌ها در صورت مواجه شدن با این وضعیت به چند صورت عمل می‌کنند. بعضی‌ها یک خطا را گزارش می‌کنند در حالی‌که بعضی دیگر به بهترین وجه ممکن این عدد را گرد کرده و به پردازش ادامه می‌دهند. برای مثال اگر کامپیوتر شما تا ۸ رقم اعشار پشتیبانی کند و محاسبات برنامه ما عددی با ۹ رقم اعشار تولید کند وضعیت پاریز رخ می‌دهد. هنگامی‌که تعداد زیادی از احتمالات در هم ضرب شوند و حاصل آن عددی بسیار کوچک باشد پاریز اتفاق می‌افتد و به دلیل بسیار کوچک بودن، قابل بیان نمی‌باشد. از آنجائیکه احتمال همیشه بین صفر و یک است بنابراین:

$$\prod_{i=1}^{n} P(X_i|Y = y_k) \to 0 \quad as \quad n \to \infty$$

[1] Over fitting
[2] Underflow

در اینصورت احتمال پسین هر کلاس برابر صفر خواهد شد، بنابراین در اینگونه موارد NBC در طبقه‌بندی داده‌های آزمون دچار خطا می‌شود.

محدودیت دیگر آن این است‌که احتمال درستنمایی با توجه به هر کلاس برابر صفر شود. این محدودیت را بیش‌برآوردی گویند. بیش برآوردی زمانی رخ می‌دهدکه مقدار متغیر X_i هرگز در نمونه آموزشی برای کلاس y_k مشاهده نشود. بنابراین احتمال درستنمایی $P(X = x \mid Y = y)$ صفر برآورد خواهد شد و در نتیجه ضرب دنباله $P(Y = y_k) \prod_i P(X_i|Y = y_k)$ برابر صفر می‌شود. برای رفع این مشکل از روش‌های هموارسازی برای برآورد پارامتر θ استفاده می‌شود.

۳-۷ برآورد هموارسازی[1]

با توجه به مشکلی که در بخش قبل بیان شد با اضافه کردن عامل هموارساز به معادله ۳-۱۸داریم[16]:

$$\hat{\theta}_{ijk} = \hat{P}\big(X_i = x_{ij} \big| Y = y_k\big) = \frac{\#D\{X_i = x_{ij}\, \Lambda\, Y = y_k\} + L}{\#D\{Y = y_k\} + LJ} \qquad (20-3)$$

که در آن $\#D\{X_i = x_{ij}\, \Lambda\, Y = y_k\}$ تعداد نمونه‌ای با خاصیت $(X_i = x_i\, \Lambda\, Y = y_j)$ می‌باشند، $\#D\{Y = y_k\}$ تعداد نمونه‌ای با خاصیت $Y = y_k$ و J تعداد حالت‌های X_i و L نیز مقداری قراردادی است‌که برای تمام مشاهدات یکسان بوده و توانایی هموارسازی نام دارد. اگر $L = 1$ در نظر گیریم برآورد هموارسازی لاپلاس گویند که در بخش بعد بیان می‌کنیم. همچنین برآورد هموار شده $\hat{\pi}_k$ نیز بصورت زیر می‌باشد :

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + L}{|D| + LK} \qquad (21-3)$$

[1] Smoothing Estimate

که در آن $\#D\{Y = y_k\}$ تعداد نمونه‌ای با خاصیت $Y = y_j$ هستند، K تعداد کلاس‌ها و L نیز توانایی هموارسازی می‌باشد.

۳-۷-۱ برآورد لاپلاس[1] و برآورد M [2]

در سال ۱۹۹۰ کسنیک[3] دو روش برای هموارسازی ارائه داد. این دو روش، با استفاده از عامل هموار-سازی، معمولا برای افزایش مقیاس مقادیر بکار می‌رود. کسنیک این دو روش را که، برآورد لاپلاس و برآورد M می‌باشند، برای محاسبه احتمالات معادلات ۳-۲۰ و ۳-۲۱ ارائه داد. برآورد لاپلاس برای محاسبه $\hat{P}(Y = y_k)$ بصورت زیر می‌باشد[37]:

$$P_{laplace} = \frac{\#D\{Y = y_k\} + 1}{|D| + K} \qquad (22-3)$$

که در آن $\#D\{Y = y_k\}$ تعداد نمونه‌ای با خاصیت $Y = y_j$ هستند، $|D|$ تعداد نمونه‌های آموزشی، K تعداد کلاس‌ها می‌باشد. برای بدست آوردن $P(X_i = x_i|Y = y_j)$ نیز داریم:

$$\hat{P}_{P_{laplace}}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij}\ \Lambda\ Y = y_k\} + 1}{\#D\{Y = y_k\} + J} \qquad (23-3)$$

$\#D\{X_i = x_{ij}\ \Lambda\ Y = y_k\}$ تعداد نمونه‌ای با خاصیت $(X_i = x_i\ \Lambda\ Y = y_j)$ می‌باشند، $\#D\{Y = y_k\}$ تعداد نمونه‌ای با خاصیت $Y = y_j$ و J تعداد حالت‌های X_i است.

$P(X_i = x_i|Y = y_j)$ را می‌توان از روش دیگری که روش برآورد M نام دارد محاسبه کرد :

$$P_m = \frac{\#D\{X_i = x_{ij}\ \Lambda\ Y = y_k\} + m * p}{\#D\{Y = y_k\} + m} \qquad (24-3)$$

که در آن p احتمال پیشین $P(Y = y_k)$ (که بوسیله براورد لاپلاس بدست می‌آید) و m نیز مقداری قراردادی است که برای تمام متغیرها برابر است. در حالتی خاص، هنگامی‌که $m = n$ و $p = 1/n$

[1] Laplace-Estimate
[2] M-Estimate
[3] Cestnik

باشد، برآورد m معادل براورد لاپلاس خواهد شد.

۳-۸ بیزی خام با متغیر توضیحی پیوسته

رفتار مدل بیزی خام برای متغیرهای پیوسته وگسسته متفاوت است. در این بخش به محاسبه $P(X = x \mid Y = y)$ هنگامی‌که متغیر توضیحی پیوسته است می‌پردازیم. یک فرضیه رایج در این روش این است‌که، توزیع توام متغیرهای پیوسته در هر کلاس، نرمال است. باید توجه داشت که این فرضیه در ذات مدل بیزی خام وجود ندارد اما اغلب از آن استفاده می‌کنند. هنگامی‌که Y دارای k مقدار گسسته و X برداری از n متغیر توضیحی پیوسته نرمال باشد، با استفاده از برآورد میانگین و انحراف معیار مشاهدات، $P(X = x \mid Y = y)$ را بدست می‌آوریم. بنابراین :

$$\mu_{ik} = E[X_i|Y = y_k] \quad \& \quad \sigma_{ik}^2 = E((X_i - \mu_{ik})^2 \mid Y = y_k)) \qquad (25-3)$$

با توجه به اینکه متغیرهای توضیحی پیوسته می‌باشند، $P(X_i = x_{ij}|Y = y_k)$ را باید بصورت تابع چگالی احتمال در نظر گرفت، یعنی $g(x; \mu_{ik}, \sigma_{ik})$ جایگزین $P(X_i = x_{ij}|Y = y_k)$ می‌شود که $g(x; \mu, \sigma)$ تابع چگالی احتمال توزیع نرمال با پارامترهای μ و σ^2 می باشد :

$$g(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (26-3)$$

حال باید تعداد پارامترهای مورد نیاز برای محاسبه معادله ۳-۲۶ را بدست آوریم.

همان‌طور که از معادله ۳-۲۶ مشخص است، پارامترهای این معادله، μ_{ik} و σ_{ik} می‌باشند. در نتیجه برای محاسبه تعداد پارامترهای مورد نیاز معادله ۳-۲۶ ، کافیست تعداد پارامترهای μ_{ik} و σ_{ik} را محاسبه کنیم. با توجه به اینکه هر یک از μ_{ik}ها و σ_{ik}ها، برابر با nk است لذا تعداد کل پارامترهای مورد نیاز در این مدل برابر $2nk$ خواهد شد. همچنین برآورد احتمال پیشین Y نیز برابر $k-1$ پارامتر می‌باشد. در این روش نیز با استفاده از روش درستنمایی ماکزیمم به برآورد پارامترهای μ_{ik} و σ_{ik} می‌پردازیم. برآوردگر درستنمایی ماکزیمم μ_{ik} و σ_{ik}^2 ، مانند $\bar{X}$ و S^2 می-

باشد ولی از آنجائی‌که هر یک از پارامترها وابسته به کلاس مورد نظرشان می‌باشند، تنها مقادیر اختصاص داده به آن کلاس در نظر گرفته می‌شود، بنابراین داریم [16] :

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta\left(Y^j = y_k\right) \qquad (27-3)$$

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta\left(Y^j = y_k\right) \qquad (28-3)$$

که در آن X_i^j، j- امین مشاهده متغیر X_i ، و $\delta(Y = y_k)$ بصورت زیر می‌باشد:

$$\delta(Y = y_k) = \begin{cases} 1 & if \quad Y = y_k \\ 0 & if \quad Y \neq y_k \end{cases}$$

۳-۹ طبقه‌بندی‌کننده بیزی خام گوسی[1]

طبقه‌بندی کننده بیزی خام هنگامی‌که بر پایه فرضیات زیر باشد به آن طبقه‌بندی‌کننده بیزی خام گوسی گویند که به اختصار آن را با GNB نشان می‌دهیم. این فرضیات به‌صورت زیر می‌باشد[16]:

۱. Y مقادیر ۰ یا ۱ را شامل شود و دارای توزیع برنولی با پارامتر $\pi = P(Y = 1)$ باشد

۲. $X =< X_1, \dots, X_n >$ متغیرهای تصادفی پیوسته باشد

۳. $\left(X_i = x_{ij} | Y = y_k\right) \sim N(\mu_{ik}, \sigma_i)$

۴. برای تمام i و $j \neq i$ ، X_i , X_j با توجه به y داده شده‌شان مستقل شرطی‌اند .

۳-۱۰ ارتباط طبقه‌بندی‌کننده بیزی خام گوسی با رگرسیون لجستیک

حال به ارتباط رگرسیون لجستیک با روشGNB می‌پردازیم. باید توجه داشت که در اینجا فرض بر این است‌که σ_i برای هر متغیر، وابسته به کلاس متغیر پاسخ نمی‌باشد.

[1] Gaussian Naive Bayes Classifier

در حالت کلی با استفاده از قانون بیز داریم:

$$P(y=1|X=x)=\frac{P(y=1)g(x;\mu_{i1},\sigma_{i1})}{P(y=1)g(x;\mu_{i1},\sigma_{i1})+P(y=0)g(x;\mu_{i0},\sigma_{i0})}$$

$$=\frac{1}{1+\frac{P(y=0)g(x;\mu_{i0},\sigma_{i0})}{P(y=1)g(x;\mu_{i1},\sigma_{i1})}}$$

$$=\frac{1}{1+exp(ln\frac{P(y=0)g(x;\mu_{i0},\sigma_{i0})}{P(y=1)g(x;\mu_{i1},\sigma_{i1})})}$$

با استفاده از استقلال شرطی و فرضیه طبقه بندی کننده بیزی خام گوسی خواهیم داشت :

$$P(y=1|X=x)=\frac{1}{1+\exp(ln\frac{P(y=0)}{P(y=1)}+\sum_i ln\frac{g(x;\mu_{i0},\sigma_{i0})}{g(x;\mu_{i1},\sigma_{i1})})}$$

$$=\frac{1}{1+exp(ln\frac{1-\pi}{\pi}+\sum_i ln\frac{g(x;\mu_{i0},\sigma_{i0})}{g(x;\mu_{i1},\sigma_{i1})})} \qquad (29-3)$$

حال باید مجموع عبارت $ln\frac{g(x;\mu_{i0},\sigma_{i0})}{g(x;\mu_{i1},\sigma_{i1})}$ را بررسی کنیم. با توجه به این فرض که، $g(x;\mu_{ik},\sigma_{ik})$ نرمال است، بنابراین با بسط دادن عبارت مورد بررسی خواهیم داشت :

$$\sum_i ln\frac{g(x;\mu_{i0},\sigma_{i0})}{g(x;\mu_{i1},\sigma_{i1})}=\sum_i ln\frac{\frac{1}{\sqrt{2\pi\sigma_i^2}}exp\left(\frac{-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_i^2}}exp\left(\frac{-(X_i-\mu_{i1})^2}{2\sigma_i^2}\right)}$$

$$=\sum_i lnexp\left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)=\sum_i\left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)$$

$$=\sum_i(\frac{(X_i^2-2X_i\mu_{i1}+\mu_{i1}^2)-(X_i^2-2X_i\mu_{i0}+\mu_{i0}^2)}{2\sigma_i^2}$$

$$= \sum_i \left(\frac{2X_i(\mu_{i0} - \mu_{i1}) + \mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} \right) = \sum_i \left(\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} \right) \quad (30-3)$$

با قرار دادن ۳-۳۰ در ۳-۲۹ خواهیم داشت :

$$P(y = 1|X = x) = \frac{1}{1 + exp(ln\frac{1-\pi}{\pi} + \sum_i (\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}))}$$

با در نظر گرفتن $\beta_i = \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2}$ و $\beta_0 = ln\frac{1-\pi}{\pi} + \sum_i (\frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2})$ داریم :

$$P(y = 1|X) = \frac{1}{1 + exp(\beta_0 + \sum_i \beta_i X_i)}$$

$$P(y = 0|X) = \frac{exp(\beta_0 + \sum_i \beta_i X_i)}{1 + exp(\beta_0 + \sum_i \beta_i X_i)} \quad (31-3)$$

با توجه به ۳-۳۱، مشاهده می‌شود که معادلات پارامتری رگرسیون لجستیک با GNB یکسان شده است. این ارتباط به این معنا نیست که عملکرد روش رگرسیون لجستیک و روش بیزی خام یکسان است. در عمل دقت طبقه‌بندی این دو روش باهم متفاوتند.

۳-۱۱ تفاوت روش رگرسیون لجستیک و روش بیزی خام

ابتدا طبقه‌بندی‌کننده ممیزی[1] و طبقه‌بندی‌کننده مولد[2] را به‌طور مختصر تعریف می‌کنیم. برای محاسبه $P(Y = y \mid X = x)$ به روش مولد، از قانون بیز استفاده می‌شود که در آن نیاز به توزیع احتمال توام متغیرهای توضیحی در هر کلاس است. اما در روش ممیزی، از قانون بیز استفاده نشده و مدل صریح دیگری که شامل پارامترهای مجهول است در نظر گرفته می‌شود. رگرسیون لجستیک، گونه‌ای از طبقه‌بندی‌کننده ممیزی است، در صورتیکه روش بیزی خام به طبقه‌بندی‌کننده مولد اختصاص داده می‌شود[39]. بطور خلاصه داریم :

[1] Discriminative
[2] Generative

محاسبه $P(Y = y \mid X = x)$ به روش طبقه‌بندی‌کننده مولد

- براورد پارامترهای $P(X = x \mid Y = y)$ و $P(Y = y)$ از طریق داده‌های مدل‌ساز بر اساس فراوانی نسبی مشاهدات
- استفاده از قانون بیز برای محاسبه $P(Y = y \mid X = x)$

محاسبه $P(Y = y \mid X = x)$ به روش طبقه‌بندی‌کننده ممیزی

- محاسبه $P(Y = y \mid X = x)$ از طریق داده‌های مدل‌ساز بر اساس شکل پارامتری تابع

۳-۱۱ تفاوت روش تحلیل ممیزی خطی با طبقه بندی کننده بیزی خام گوسی

روش تحلیل ممیزی، یکی از روش‌های طبقه‌بندی برای داده‌هایی با متغیرهای ورودی پیوسته می‌باشد. در این روش طبقه‌بندی نیز، نیاز به محاسبه احتمال پسین مشاهدات ورودی در کلاس مورد نظرشان هستیم. بنابراین به دنبال بدست آوردن $P(Y = y \mid X = x)$ می‌باشیم. فرض می کنیم $f_k(x)$ تابع چگالی شرطی X در کلاس $Y = k$ و π_k احتمال پیشین کلاس k باشد که $\sum_{k=1}^{K} \pi_k =$ 1 است[35]. با توجه به قضیه بیز :

$$P(Y = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

همچنین:

$$\hat{\pi}_k = \frac{\#samples\ in\ class\ k}{total\ \#\ of\ samples}$$

همچنین فرض می کنیم توزیع توام xها، نرمال چند متغیره بصورت زیر باشد :

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k^{-1}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

در روش تحلیل ممیزی خطی، فرض بر این است‌که ماتریس واریانس-کواریانس متغیرها برای هر کلاس یکسان می‌باشند یعنی :

$$\sum_k = \sum \quad \forall k$$

تفاوت این روش با روش طبقه بندی کننده بیزی خام گوسی در این است‌که در این روش توزیع توام متغیرها را چند متغیره نرمال در نظر گرفته ولی روش طبقه‌بندی کننده بیزی خام گوسی، بر اساس تابع چگالی حاشیه‌ای متغیرهای ورودی، که نرمال یک متغیره است بیان می‌شود. همچنین در روش تحلیل ممیزی خطی فرض بر این است‌که متغیرهای ورودی دارای ماتریس واریانس کواریانس برابر هستند ولی در روش طبقه‌بندی کننده بیزی خام گوسی فرض بر این است‌که تمام متغیرهای ورودی در هر کلاس، از هم مستقل می‌باشند.

نتیجه‌گیری

با توجه به مطالبی که بیان شد، می‌توان ساختاری را برای روش بیزی ساده تعریف کرد. این ساختار شامل دو قسمت است که در زیر شرح داده شده است [36].

I) ساختاری برای طبقه بندی

1) استفاده از داده‌های آموزشی برای تعیین احتمالات $P(Y = y_j)$

2) استفاده از فراوانی نسبی داده‌ها با توجه به داده‌های آموزشی برای تعیین احتمالات شرطی کلاس-ها برای متغیرهای گسسته و میانگین و واریانس کلاس‌ها برای متغیرهای پیوسته

II) طبقه بندی کردن نمونه جدید X

1) محاسبه $P(X = x|y_j)P(y_j)$ برای هر y_j، که در آن $P(X = x|y_j) = \prod_{k=1}^{n} P(x_k|y_j)$

2) تعیین طبقه X به کلاس مناسب بر اساس بیشترین مقدار

فصل چهارم

شبکه بیزی

۴-۱ مقدمه

شبکه بیزی نیز مانند روش بیزی خام بر پایه قانون بیز بنا شده است. همان طور که در فصل قبل بیان شد روش NBC بر پایه استقلال متغیرهای توضیحی در هر کلاس بیان می‌شود. این شرط محدودیت بسیار بزرگی است زیرا در عمل، امکان وابستگی میان متغیرها وجود دارد. بنابراین روش NBC عملکرد مناسبی نخواهد داشت. رهیافتی که در این‌گونه مواقع پیشنهاد می‌شود، روش شبکه بیزی است. قبل از بیان این مدل ابتدا به ذکر تعاریفی پایه‌ای می‌پردازیم.

۴-۲ احتمالات فیزیکی و بیزی

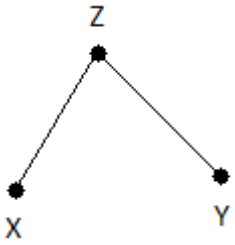
احتمالات بدست آمده توسط شبکه بیزی ممکن است بیزی یا فیزیکی باشند. هنگامی‌که شبکه بیزی تنها با دانش قبلی ساخته شود، احتمالات بیزی خواهند بود و هنگامی‌که یادگیری این شبکه توسط داده‌ها صورت پذیرد، این احتمالات را فیزیکی گویند[40]. در این فصل، روش شبکه بیزی را با استفاده از احتمالات فیزیکی مورد مطالعه قرار خواهیم داد.

۴-۳ مدل گرافیکی

برای درک مدل‌های گرافیکی، فهم استقلال شرطی برای یک مجموعه از متغیرها لازم است. همان‌طور که در فصل قبل بیان شد، با در نظر گرفتن ۳ متغیر تصادفی X,Y,Z ، متغیرهای X و Y را با شرط Z مستقل گویند اگر توزیع احتمال توام X و Y به شرط Z را بتوان بصورت حاصل ضرب دو مولفه تجزیه شده زیر نوشت:

$$f(x,y|Z=z)=f(x|Z=z)f(y|Z=z) \qquad (1-4)$$

این استقلال را با نماد $X\perp Y|Z$ نمایش داده و نمودار گرافیکی آن بصورت شکل ۴-۱است.



شکل (۴-۱) گراف مستقل شرطی $X \perp Y|Z$

حال به تعریف مدل‌های گرافیکی می‌پردازیم. مدل‌های گرافیکی مدل‌هایی هستند که می‌توانند بطور مستقیم از طریق روابط استقلال شرطی میان متغیرها در یک گراف مشخص شوند. داروچ، لاریزن و اسپید[1] در سال ۱۹۸۰ با استفاده از مفهوم گراف و مدل‌های آماری، مدل‌های گرافیکی را ارائه دادند.

گراف $G = (V, E)$ را در نظر بگیرید که ساختار آن شامل تعداد متناهی از رئوس یا گره‌ها[2] می‌باشد. مجموعه این رئوس که همان متغیرها هستند را با V نشان داده و E نیز نشان‌دهنده مجموعه‌ای از یال‌های[3] جهت‌دار بین متغیرها می‌باشد. بطور کلی، تاثیر علت و معلولی هر متغیر بروی متغیر دیگر بوسیله یال‌های جهت‌دار صورت می‌گیرد. شکل ۴-۲ نمونه‌ای از گراف غیرجهت‌دار و جهت‌دار را نشان می‌دهد که در گراف جهت‌دار، x و c علتی برای A می‌باشند.

شکل(۴-۲) نمونه‌ای از گراف غیرجهت‌دار(گراف سمت راست) و گراف جهت‌دار(گراف سمت چپ)

گره X را اولیا[4] گره Y گوئیم و آن‌را بصورت $X \to Y$ نمایش می‌دهیم اگر آن‌ها با یال جهت‌داری از X

[1]Darroch Lauritzen Speed
[2] Node
[3] Edge
[4] Parents

به Y متصل شده باشند، یعنی: $if\ (X,Y) \in E\ and\ (Y,X) \notin E$. بطور کلی، مدل گرافیکی مبین خانواده‌ای از توزیع‌های احتمال می‌باشد که قوانین استقلال شرطی آن توسط یک گراف بیان می‌شود.

۴-۴ مدل گرافیکی بازگشتی[1]

مدل‌های گرافیکی بازگشتی به عنوان ابزاری با اهمیت برای پیش‌بینی در داده‌کاوی در نظر گرفته می‌شوند. برای مشخص کردن هر مدل بازگشتی، نیاز به گراف جهت‌داری داریم که روابط علت و معلولی میان متغیرها را نشان دهد. به محض اینکه این گراف مشخص شد، توزیع احتمال آن با توجه به داشتن مجموعه‌ای از متغیرهای $\mathrm{X} = \{\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{n}\}$ محاسبه خواهد شد. با توجه به قانون زنجیری احتمال داریم:

$$p(\mathrm{X}_1 = x_1, \dots, \mathrm{X}_\mathrm{n} = x_n) = p(x_n|x_1, \dots, x_{n-1}).p(x_1, \dots, x_{n-1})$$

که در معادله بالا $p(x_1, \dots, x_{n-1})$ بصورت زیر می‌باشد:

$$p(x_1, \dots, x_{n-1}) = p(x_{n-1}|x_1, \dots, x_{n-2}).p(x_1, \dots, x_{n-2})$$

با تکرار فرایند فوق خواهیم داشت:

$$p(\mathrm{X}_1 = x_1, \dots, \mathrm{X}_\mathrm{n} = x_n) = \prod_{i=1}^{n} p(x_i|x_1, \dots, x_{i-1}) \qquad (2-4)$$

حال برای هر X_i، زیر مجموعه‌ای بصورت $\Pi_i \subseteq \{\mathrm{X}_1, \dots, \mathrm{X}_{i-1}\}$ داریم بطوریکه، X_i و $\{\mathrm{X}_1, \dots, \mathrm{X}_{i-1}\} - \Pi_i$ با توجه به Π_i داده شده، مستقل شرطی‌اند. بنابراین برای هر $\{\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{n}\}$ داریم:

$$p(x_i|x_1, \dots, x_{i-1}) = p(x_i|\pi_i) \qquad (3-4)$$

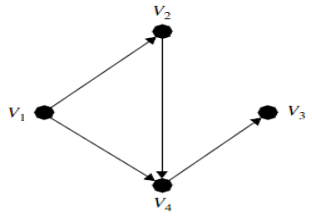
با ادغام ۴-۲ و ۴-۳ داریم :

[1] Recursive graphical models

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i) \qquad (4-4)$$

که در آن به تعداد متغیرها گره داریم (n گره) و π_i نشان‌دهنده اولیای گره‌ها می‌باشد. بنابراین توزیع توام n متغیر، بصورت ضرب دنباله‌ای از عبارات می‌باشد. به آن عبارات احتمالات موضعی[1] گویند که هر یک از آن‌ها، وابستگی متغیر X_i نسبت به متغیرهای توضیحی وابسته به آن، که اولیای آن می-باشند، را نشان می‌دهد. شبکه بیزی، گرافی بدون دور و جهت‌دار می‌باشد که در ادامه برای پرداختن به این مدل نیاز به بیان تعریف گراف بدون‌دور جهت‌دار هستیم.

گرافی را بدون‌دور[2] گویند که گره‌ها بصورت حلقه، با هم در ارتباط نباشند و نتوان راهی در گراف پیدا کرد که با شروع از یک گره به همان گره بازگشت. شکل زیر گرافی جهت‌دار بدون‌دور را به ما نشان می‌دهد که در آن یال‌ها، خط‌های ارتباط‌دهنده بین گره‌ها می‌باشد و یال‌هایی که در گراف بین متغیرها رسم نشده‌اند توسط استقلال شرطی بین آنها مشخص می‌شوند. برای مثال همان‌طور که در شکل ۴-۳ می‌بینید از آنجائیکه V_2 و V_3 یالی برای اتصال ندارند پس می‌توان نتیجه گرفت از هم مستقل‌اند و وابستگی بین آن‌ها وجود ندارد.



شکل(۴-۳) نمونه‌ای از گراف بدون‌دور جهت‌دار

با توجه به شکل ۴-۳، مدل بازگشتی زیر را داریم :

$$p(V_1, V_2, V_3, V_4) = p(V_1)p(V_2|V_1)p(V_4|V_1, V_2)p(V_3|V_4)$$

[1] Local probability
[2] Acyclic graph

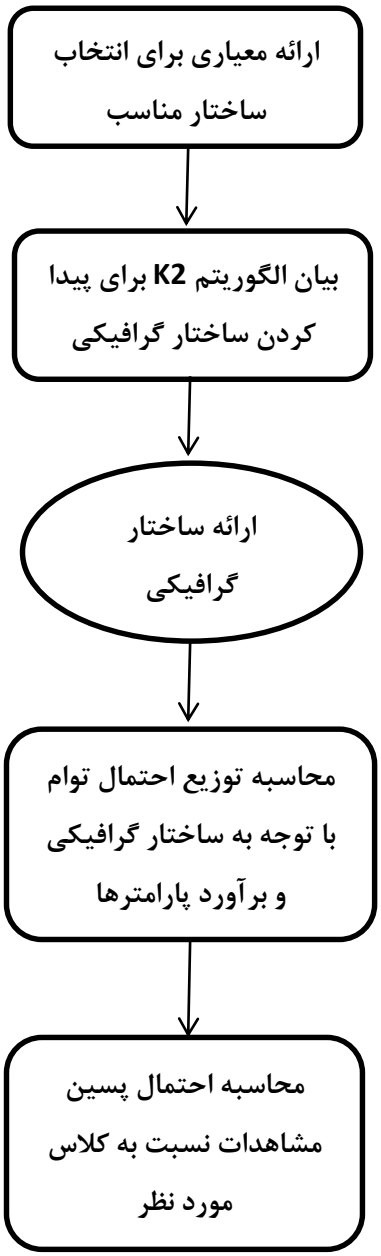
همان‌طور که از معادله بالا مشاهده می‌شود، استقلال شرطی متغیرها بر اساس ارتباطشان با اولیای آن‌ها و یا گره‌ای که به آن متصل شده‌اند بیان می‌شود و در صورتیکه گره‌ای اولیا نداشته باشد دیگر بصورت شرطی بیان نمی‌شود، مانند گره V_1 . مزیت مدل بازگشتی این است‌که به سادگی بر اساس مدل گرافیکی جهت‌دار بیان می‌شود و مدل‌های موضعی بسیاری تبدیل شده و باعث کاهش بعد می-شود. مدل‌های گرافیکی جهت‌دار در برنامه‌های کاربردی هوش مصنوعی استفاده می‌شود که شناخته-ترین آن‌ها شبکه بیزی می‌باشد. شبکه بیزی می‌تواند به عنوان روشی از طبقه‌بندی بیزی نام برده شود که در آن ارتباط وابستگی شرطی میان متغیرهای توضیحی بوسیله گراف جهت‌داری مشخص می-شود[7].

۴-۵ شبکه بیزی

شبکه بیزی را پیرل[1] در سال ۱۹۸۶ ارائه داد که به عنوان پیوندی از هوش مصنوعی و آمار می‌باشد. شبکه بیزی مدلی از وابستگی‌ها میان مجموعه‌ای از متغیرها را نشان می‌دهد که بوسیله ساختاری شبکه‌ای باهم در ارتباط هستند [41]. شبکه بیزی گراف بدون‌دور جهت‌داری[2] است (DAG) ، که نودها در آن مطابق با متغیرهای تصادفی می‌باشند. در این گراف هر نود بوسیله مجموعه‌ای از احتمالات شرطی $p(x_i|\pi_i)$ ارتباط داده می‌شوند که در آن x_i متغیری است‌که با نود خاصی ارتباط داده شده و π_i مجموعه‌ای از اولیاها در گراف می‌باشد[40]. در شبکه بیزی هر متغیر دارای مجموعه متناهی از حالت‌های دو به دو ناسازگار می‌باشند و با هر متغیر یک مجموعه از جدول احتمال شرطی ارتباط داده می‌شود. یادگیری شبکه بیزی را می‌توان به دو بخش تقسیم کرد: یادگیری ساختاری، که توپولوژی‌ای از شبکه را تعیین می‌کند، و یادگیری پارامتری که برای محاسبه پارامترها یا همان احتمالات شرطی با توجه به ساختار شبکه بدست آمده بکار می‌رود. در روش‌های طبقه‌بندی برای اختصاص دادن مشاهدات به کلاس مورد نظر، می‌بایست احتمال پسین آن را محاسبه کرد. در روش

[1] pearl
[2] Directed Acyclic Graph

شبکه بیزی نیز به‌همین صورت می‌باشد و با توجه به ساختاری که از شبکه بیزی ارائه می‌شود، احتمالات پسین را با توجه به توزیع احتمال توام در مدل شبکه بیزی محاسبه می‌کنیم. بنابراین مهم-ترین بخش در ارائه مدل شبکه بیزی، بدست آوردن ساختار گرافیکی مناسب برای متغیرهای مدل می‌باشد. شکل۴-۴ فرایند شبکه بیزی را نمایش می‌دهد.



شکل(۴-۴) ساختاری از مدل شبکه بیزی

باید توجه داشت که در این مطالعه، مدل شبکه بیزی را در حالتی‌که داده‌های آن کامل (یعنی دارای داده گمشده نباشد) و با متغیرهای گسسته بررسی خواهیم کرد و علاقه‌مندان به این موضوع می‌توانند در تحقیقات آینده، حالتی متفاوت با این مطالعه یعنی داده‌های ناقص و متغیرهای پیوسته را بررسی نمایند. ابتدا به علامت‌گذاری‌های لازم در رابطه با این بحث می‌پردازیم:

(X, Y, X_i) را به عنوان متغیر و (x, y, x_i) را به عنوان مقادیر یا حالت‌های متغیر و مجموعه‌ای از مشاهدات را این‌گونه $D = \{X_1 = x_1, \dots, X_1 = x_n\}$ نشان می‌دهیم.

شبکه بیزی برای یک مجموعه از متغیرهای $\{X_1, \dots X_n\}$ شامل دو مولفه می‌باشد :

۱) ساختار شبکه(S) : که در آن مجموعه مستقل شرطی از متغیرها نمایش داده می‌شوند.

۲) مجموعه P : که توزیع احتمال محلی ارتباط داده شده بوسیله هر متغیر می‌باشد.

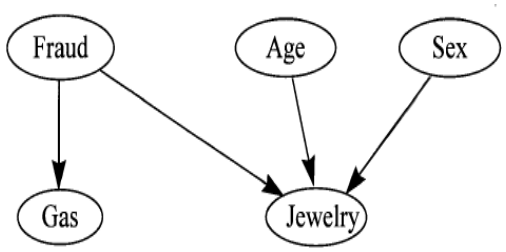
بوسیله این دو مولفه، توزیع احتمال توام تعریف می‌شود. ساختار شبکه (S) یک گراف بدون‌دور جهت‌دار می‌باشد و نودها در S تناظر یک به یک با متغیرها را دارند. از X_i هم برای نشان دادن متغیر و هم گره مطابق با آن استفاده می‌کنیم. بنابراین همان‌طور که در بخش ۴-۴ بیان شد، با توجه به ساختار S داده شده، توزیع احتمال توام $\{X_1, \dots X_n\}$ بصورت زیر خواهد شد :

$$p(X_1 = x_1, \dots, X_1 = x_n) = \prod_{i=1}^{n} p(x_i|\pi_i) \qquad (5-4)$$

در شبکه بیزی به DAG ساختار گفته می‌شود و به مقادیر توزیع‌های احتمال شرطی پارامتر گفته می‌شود که با θ_{ijk} نشان داده می‌شود. قبل از اینکه به یادگیری پارامتری شبکه بیزی بپردازیم، ابتدا یادگیری ساختاری آنرا مورد مطالعه قرار می‌دهیم.

۴-۶ یادگیری ساختار شبکه بیزی

با مثالی به فرایند ساختار شبکه بیزی می‌پردازیم که موضوع آن کشف تقلب در کارت اعتباری است[40]. ابتدا متغیرهای این مثال را بیان می‌کنیم که در شکل ۴-۵ مشخص است: جنسیت (S)، سن (A)، جواهرفروشی (J)، پمپ بنزین (G) و تقلب (F)، که بیانگر این مساله است‌که معامله انجام شده در آن تقلب صورت گرفته یا خیر. در آن G به عنوان معامله‌ای که در آن بنزین خریداری شده است یا خیر و هم‌چنین J به عنوان معامله‌ای در جواهرفروشی انجام شده است یا خیر. باید توجه داشت که در مسئله واقعی متغیرهای زیادی وجود دارند که برای راحت توصیف کردن مثال تعداد کمی از متغیرها را در ارتباط با کشف تقلب در کارت اعتباری بیان کردیم.



شکل (۴-۵) شبکه بیزی برای کشف تقلب در کارت اعتباری

در تعیین ساختار بیزی ابتدا باید به طریقی متغیرها را مرتب کرد. شکل ۴-۵ نمونه‌ای از مرتب‌سازی متغیرها می‌باشد. سپس بایستی مجموعه متغیرهایی که در رابطه ۴-۵ صدق می‌کنند را تعیین نمود. در مثال بالا، با توجه به مرتب‌سازی در نظر گرفته شده روابط زیر را داریم :

$$p(a|f) = p(a) \qquad p(s|f,a) = p(s)$$

$$p(g|f,a,s) = p(g|f) \qquad p(j|f,a,s,g) = p(j|f,a,s,) \qquad (6-4)$$

بنابراین احتمال پسین مشاهدات با توجه به کلاس مورد نظر بصورت زیر خواهد شد :

$$P(f \mid a,s,g,j) = \frac{P(f,a,s,g,j)}{P(a,s,g,j)} = \frac{P(f,a,s,g,j)}{\sum_{\acute{f}} P(\acute{f},a,s,g,j)}$$

که در آن $\acute{f}$ حالت‌های مختلف متغیر f را نشان می‌دهد، یعنی حالت‌های تقلب و غیر تقلب. حال با توجه به ۴-6 داریم :

$$P(f \mid a,s,g,j) = \frac{p(f)p(a)p(s)p(g \mid f)p(j \mid f,a,s)}{\sum_{\acute{f}} p(\acute{f})p(a)p(s)p(g \mid \acute{f})p(j \mid \acute{f},a,s)}$$

با توجه به ثابت بودن مخرج خواهیم داشت:

$$P(f \mid a,s,g,j) \propto p(f)p(a)p(s)p(g \mid f)p(j \mid f,a,s) \qquad (7-4)$$

باید توجه داشت که اگر متغیرها به‌طور نامناسبی در گراف مرتب شوند، ممکن است ساختار شبکه بدست آمده ، استقلال‌های غیر واقعی را نشان دهد. یادگیری ساختار شبکه بیزی بوسیله داده‌ها مسئله‌ای بسیار سخت می‌باشد و حالت‌های متعددی را در بر می‌گیرد، یعنی بر اساس داده‌های موجود و متغیرها، ساختاری را برای مدل حدس زده و بر اساس آن مدل را ارائه دهیم. الگوریتم‌های متعددی برای یادگیری ساختار شبکه بیزی طراحی شده است. این الگوریتم‌ها هر یک دارای فرایند خاصی می-باشند. بعضی از این الگوریتم‌ها، تمام ساختارهای ممکن را در نظر گرفته و از میان آن‌ها با توجه به معیاری تعریف شده، بهترین ساختار را انتخاب می‌کنند. در صورتیکه در نوع دیگر از الگوریتم‌ها، محتمل‌ترن ساختار انتخاب می‌شود و روند آن مانند الگوریتم‌هایی که تمام ساختارهای ممکن را جستجو کرده و از میان آنها بهترین را انتخاب می‌کنند نمی‌باشد. فرایند الگوریتم K2، به گونه‌ای است‌که با توجه به مجموعه داده و متغیرهای آن، محتمل‌ترین ساختار ارائه می‌شود و جزء دسته دوم از الگوریتم‌ها می‌باشد. در ادامه به مطالعه این الگوریتم خواهیم پرداخت.

۴-۷ الگوریتم *K2*

در سال‌های اخیر الگوریتم‌های متعددی برای یادگیری شبکه بیزی بیان شده که از پرکاربردترین آن‌ها

می‌توان الگوریتم $K2$ که توسط کوپر و هرسویت[1] در سال ۱۹۹۲ ارائه شده را نام برد[42]. همان‌طور که بیان کردیم، فرض می‌کنیم متغیرها گسسته هستند. بنابراین برای هر گره i متناظر با متغیر تصادفی X_i ، r_i حالت ممکن $x^1, \dots, x^{r_i}$ در نظر می‌گیریم.

$$node1 \rightarrow X_1 \in \{x_1^1, \dots, x_1^{r_1}\}$$

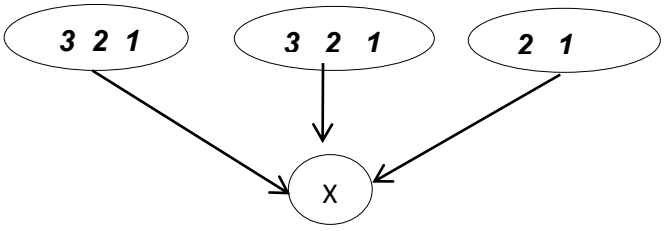
$$node2 \rightarrow X_2 \in \{x_2^1, \dots, x_2^{r_2}\}$$

$$\vdots$$

$$nodei \rightarrow X_i \in \{x_i^1, \dots, x_i^{r_i}\}$$

حال فرض کنید D مجموعه داده و N_{ijk}، تعداد مشاهداتی در D باشد که متغیر X_i در حالت k- ام و اولیاء آن حالت j- ام را شامل می‌شود یعنی، $X_i = x_i^k \quad k = 1, \dots, r_i$ و $\pi(X_i) = x_i^j$. باید توجه داشت اولیاء متغیرهای مورد نظر نشان‌دهنده ترکیب‌های π_i می‌باشد. به این معنی که، چون هر گره ممکن است ۲ یا تعداد بیشتری اولیاء داشته باشد، حالت j برابر است با ترکیب حالت‌های اولیاء یعنی $\pi_i{}^1, \dots, \pi_i{}^{q_i}$ که در آن $q_i = \prod_{X_i \in \pi_i} r_i$ ، $j = 1, \dots, q_i$.

برای درک بهتر این موضوع، گره‌ای دارای ۳ اولیاء که هر یک دارای ۳ حالت می‌باشد را در نظر بگیرید (شکل ۴-۶). بنابراین حالت‌های اولیاء عبارتند از $q_i = 3 * 3 * 3 = 27$ یعنی، ۲۷ حالت از اولیاء وجود دارد و j از ۱ تا ۲۷ حالت دارد. در جدول ۴-۱ این ۲۷ حالت نمایش داده شده است :



شکل (۴-۶) گره‌ای با ۳ اولیا

[1] *Cooper & Hersovits*

جدول (۴-۱) تعداد حالت‌های گره‌ای با ۳ اولیا

J	state	j	State	J	State
1	1 1 1	10	2 1 1	19	3 1 1
2	1 1 2	11	2 1 2	20	3 1 2
3	1 1 3	12	2 1 3	21	3 1 3
4	1 2 1	13	2 2 1	22	3 2 1
5	1 2 2	14	2 2 2	23	3 2 2
6	1 2 3	15	2 2 3	24	3 2 3
7	1 3 1	16	2 3 1	25	3 3 1
8	1 3 2	17	2 3 2	26	3 3 2
9	1 3 3	18	2 3 3	27	3 3 3

معیاری که در این الگوریتم بکار برده می‌شود تا مناسب‌ترین ساختار انتخاب شود بصورت زیر است[43] :

$$f(i,\pi_i) = \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}! \qquad (8-4)$$

این معیار در فرایند الگوریتم $K2$ برای محاسبه اولیاء گره‌های مورد نظر مورد استفاده قرار می‌گیرد که در آن $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. حال به ارائه الگوریتم $K2$ می‌پردازیم.

این الگوریتم به دنبال پیدا کردن اولیاء هر گره می‌باشد. روند این الگوریتم به این صورت است‌که ابتدا به گره‌ها یا همان متغیرها یک ترتیب اولیه می‌دهد، یعنی اگر ۳ متغیر داشته باشیم ترتیب اولیه آن به ترتیب چپ به راست بصورت X_1, X_2, X_3 فرض می‌شود. این ترتیب به این معنی است‌که ابتدا گره X_1 را در چرخه الگوریتم قرار داده سپس اولیاء آن با توجه به فرایند الگوریتم مشخص می‌شوند. در مرحله بعد X_2 را در چرخه الگوریتم قرار داده و اولیاء آن مشخص می‌شوند. این

روند تا آخرین متغیر که X_3 می‌باشد، ادامه پیدا می‌کند. هنگامی‌که اولیاء تمامی گره‌ها مشخص شد، ساختار شبکه بیزی تعیین می‌شود. این الگوریتم بصورت زیر می‌باشد[43] :

```
1.  procedure K2;
2.  {Input: A set of n nodes, an ordering on the nodes, an upper bound u on the
3.          number of parents a node may have, and a database D containing m cases.}
4.  {Output: For each node, a printout of the parents of the node.}
5.  for i:= 1 to n do
6.      π_i := ∅;
7.      P_old := f(i, π_i); {This function is computed using Equation 20.}
8.      OKToProceed := true;
9.      While OKToProceed and |π_i| < u do
10.             let z be the node in Pred(x_i) - π_i that maximizes f(i, π_i ∪ {z});
11.             P_new := f(i, π_i ∪ {z});
12.             if P_new > P_old then
13.                     P_old := P_new;
14.                     π_i := π_i ∪ {z};
15.             else OKToProceed := false;
16.     end {while};
17.     write('Node: ', x_i, ' Parent of x_i: ',π_i);
18. end {for};
19. end {K2};
```

در ادامه با ارائه مثالی ساده، فرایند این الگوریتم را مورد مطالعه قرار می‌دهیم. مجموعه داده-ای که نمونه‌ای دارای ۱۰ مشاهده است را در نظر بگیرید. این مجموعه دارای سه متغیر می‌باشد که هر یک، دو مقدار ۰ و۱ را شامل می‌شوند. این مجموعه داده در جدول ۴-۲ نشان داده شده است. باید توجه داشت در الگوریتم *K2* تمام متغیرها حتی متغیر پاسخ، در فرایند الگوریتم قرار می‌گیرند.

جدول(۴-۲) مجموعه داده‌ای شامل ۱۰ مشاهده و ۳ متغیر

case	x_1	x_2	x_3
۱	۱	۰	۰
۲	۱	۱	۱
۳	۰	۰	۱
۴	۱	۱	۱
۵	۰	۰	۰
۶	۰	۱	۱
۷	۱	۱	۱
۸	۰	۰	۰
۹	۱	۱	۱
۱۰	۰	۰	۰

در مثال فوق متغیر X_1 به عنوان متغیر پاسخ در نظر گرفته شده است. فرایند تعیین ساختار شبکه بیزی بصورت زیر می‌باشد:

ورودی:

- ترتیب اولیه بصورت x_1, x_2, x_3
- حد بالایی تعداد اولیاء برای هر متغیر u=2
- تعداد مشاهدات در مجموعه داده m=10

الگوریتم $K2$:

i=1، بنابراین برای i=1 متغیر مربوط به آن x_1 را در نظر می‌گیریم که تعداد حالت‌های آن $r_1 = 2$ می‌باشد.

i. $\pi_1 = \emptyset$ ، یعنی اولیائی برای اولین متغیر ورودی در نظر گرفته نمی‌شود.

ii. $P_{old} = f(1, \emptyset) = \prod_{j=1}^{q_1} \frac{(r_1-1)!}{(N_{1j}+r_1-1)!} \prod_{k=1}^{r_1} N_{1jk}!$

حال با توجه به مجموعه داده، مقادیر لازم برای محاسبه P_{old} را بدست می‌آوریم. از آنجائیکه $\pi_1 = \emptyset$ بنابراین $q_1 = 0$ خواهد شد. هنگامی‌که این حالت در الگوریتم رخ دهد، از اندیس j صرف‌نظر کرده و مقادیر را بر اساس اندیس i و k محاسبه می‌کنیم.

$N_{1-1} = 5$ تعداد حالت‌هایی که $x_1 = 0$

$N_{1-2} = 5$ تعداد حالت‌هایی که $x_1 = 1$

$$N_{1-} = N_{1-1} + N_{1-2} = 10$$

$$P_{old} = f(1, \emptyset) = \frac{(2-1)!}{(10+2-1)!} * 5! * 5! = \frac{1}{2772}$$

iii. ازآنجائیکه $pred(x_1) = \emptyset$، (به این معنی است‌که در مرتب‌سازی متغیرها، متغیری مقدم‌تر از x_1 قرار ندارد) بنابراین تکرار برای i=1 بوسیله $\pi_1 = \emptyset$ به اتمام می‌رسد.

---

i=2، بنابراین برای i=2 متغیر مربوط به آن x_2 را در نظر می‌گیریم که تعداد حالت‌های آن $r_2 = 2$ می‌باشد.

i. $\pi_2 = \emptyset$

ii. $P_{old} = f(2, \emptyset) = \prod_{j=1}^{q_2} \frac{(r_2-1)!}{(N_{2j}+r_2-1)!} \prod_{k=1}^{r_2} N_{2jk}!$

مانند حالت اول از آنجائیکه $\pi_2 = \emptyset$ ، بنابراین $q_2 = 0$ خواهد شد. بنابراین از اندیس j صرفنظر کرده و مقادیر را بر اساس اندیس i و k محاسبه می‌کنیم.

$N_{2-1} = 5$ تعداد حالت‌هایی که $x_2 = 0$

$N_{2-2} = 5$ تعداد حالت‌هایی که $x_2 = 1$

$$N_{2-} = N_{2-1} + N_{2-2} = 10$$

$$P_{old} = f(2, \emptyset) = \frac{(2-1)!}{(10+2-1)!} * 5! * 5! = \frac{1}{2772}$$

iii. ازآنجائیکه $pred(x_2) = \{x_1\}$، (به این معنی است‌که در مرتب‌سازی متغیرها، متغیر مقدم x_2، x_1 است) این معنی استکه در مرتب‌سازی متغیرها، متغیری مقدم‌تر از x_1 قرار ندارد بنابراین فقط تکرار برای i=2 بوسیله $z = x_1$ می‌باشد.

$$P_{new} = f(2, \pi_2 \cup \{x_1\}) = f(2, \{x_1\}) = \prod_{j=1}^{q_2} \frac{(r_2-1)!}{(N_{2j}+r_2-1)!} \prod_{k=1}^{r_2} N_{2jk}!$$

- $q_2 = 2$ که در آن $x_1 = 0$ و $x_1 = 1$ حالت‌های آن می‌باشند.
- $N_{211} = 4$ تعداد حالت‌هایی که $x_2 = 0$ و $x_1 = 0$
- $N_{212} = 1$ تعداد حالت‌هایی که$x_2 = 1$ و $x_1 = 0$
- $N_{221} = 1$ تعداد حالت‌هایی که $x_2 = 0$ و $x_1 = 1$
- $N_{222} = 4$ تعداد حالت‌هایی که $x_2 = 1$ و $x_1 = 1$
- $N_{21} = N_{211} + N_{212} = 5$
- $N_{22} = N_{221} + N_{222} = 5$

$$P_{new} = \frac{1}{6!} * 4! * 1! * \frac{1}{6!} * 1! * 4! = \frac{1}{900}$$

iv. ازآنجائیکه $P_{new} = \frac{1}{900} > P_{old} = \frac{1}{2772}$، تکرار برای i=2 بوسیله $\pi_2 = \{x_1\}$ به اتمام می‌رسد.

---

i=3، بنابراین برای i=3 متغیر مربوط به آن x_3 را در نظر می‌گیریم که تعداد حالت‌های آن $r_3 = 2$ می‌باشد.

i. $\pi_3 = \emptyset$

ii. $P_{old} = f(3, \emptyset) = \prod_{j=1}^{q_3} \frac{(r_3-1)!}{(N_{3j}+r_3-1)!} \prod_{k=1}^{r_3} N_{3jk}!$

مانند حالت‌های قبل از آنجائیکه $\pi_3 = \emptyset$ ، بنابراین $q_3 = 0$ خواهد شد و از اندیس j صرفنظر کرده و مقادیر را بر اساس اندیس i و k محاسبه می‌کنیم.

$N_{3-1} = 4$ تعداد حالت‌هایی که $x_3 = 0$

$N_{3-2} = 6$ تعداد حالت‌هایی که $x_3 = 1$

$$N_{3-} = N_{3-1} + N_{3-2} = 10$$

$$P_{old} = f(3, \emptyset) = \frac{(2-1)!}{(10+2-1)!} * 4! * 6! = \frac{1}{2310}$$

iii. در مرتب‌سازی متغیرها، متغیر مقدم x_3، x_2 و x_1 است بنابراین $pred(x_3) = \{x_1, x_2\}$، و $\pi_3 := \emptyset$، نیاز به محاسبه بیشترین مقدار تابع زیر هستیم.

$$argmax(f(3, \pi_3 \cup \{x_1\}), f(3, \pi_3 \cup \{x_2\})$$

❖ $f(3, \pi_3 \cup \{x_1\}) = f(3, \{x_1\}) = \prod_{j=1}^{q_3} \frac{(r_3-1)!}{(N_{3j}+r_3-1)!} \prod_{k=1}^{r_3} N_{3jk}!$

- $q_3 = 2$ که در آن $x_1 = 0$ و $x_1 = 1$ حالت‌های آن می‌باشند.
- $N_{311} = 3$ تعداد حالت‌هایی که $x_3 = 0$ و $x_1 = 0$
- $N_{312} = 2$ تعداد حالت‌هایی که $x_3 = 1$ و $x_1 = 0$
- $N_{321} = 1$ تعداد حالت‌هایی که $x_3 = 0$ و $x_1 = 1$
- $N_{322} = 4$ تعداد حالت‌هایی که $x_3 = 1$ و $x_1 = 1$
- $N_{31} = N_{311} + N_{312} = 5$
- $N_{32} = N_{321} + N_{322} = 5$

$$f(3, \{x_1\}) = \frac{1}{6!} * 3! * 2! * \frac{1}{6!} * 1! * 4! = \frac{1}{1800}$$

❖ $f(3,\pi_3 \cup \{x_2\}) = f(3,\{x_2\}) = \prod_{j=1}^{q_3} \frac{(r_3-1)!}{(N_{3j}+r_3-1)!} \prod_{k=1}^{r_3} N_{3jk}!$

- $q_3 = 2$ که در آن $x_2 = 0$ و $x_2 = 1$ حالت‌های آن می‌باشند.
- $N_{311} = 4$ تعداد حالت‌هایی که $x_3 = 0$ و $x_2 = 0$
- $N_{312} = 1$ تعداد حالت‌هایی که $x_3 = 1$ و $x_2 = 0$
- $N_{321} = 0$ تعداد حالت‌هایی که $x_3 = 0$ و $x_2 = 1$
- $N_{322} = 5$ تعداد حالت‌هایی که $x_3 = 1$ و $x_2 = 1$
- $N_{31} = N_{311} + N_{312} = 5$
- $N_{32} = N_{321} + N_{322} = 5$

$$f(3,\{x_2\}) = \frac{1}{6!} * 4! * 1! * \frac{1}{6!} * 0! * 5! = \frac{1}{180}$$

iv. ازآنجائیکه $f(3,\{x_2\}) = \frac{1}{180} > f(3,\{x_1\}) = \frac{1}{1800}$ بنابراین $z = x_2$. همچنین ازآنجائیکه $f(3,\{x_2\}) > P_{old}$، بنابراین $\pi_3 = \{x_2\}$ و$P_{old} = P_{new} = \frac{1}{180}$.

v. حال فرایند الگوریتم را برای i=3 با اضافه شدن x_1 ادامه می‌دهیم تا بررسی کنیم که آن به عنوان اولیائی از x_3 می‌باشد یا خیر.

$$f(3,\pi_3 \cup \{x_1\}) = f(3,\{x_1,x_2\}) = \prod\prod_{j=1}^{q_3} \frac{(r_3-1)!}{(N_{3j}+r_3-1)!} \prod\prod_{k=1}^{r_3} N_{3jk}!$$

- $q_3 = 4$. $(x_2 = 0, x_1 = 0)$ ، $(x_2 = 0, x_1 = 1)$، $(x_2 = 1, x_1 = 0)$ و $(x_2 = 1, x_1 = 1)$ حالت‌های آن می‌باشند.
- $N_{311} = 3$ تعداد حالت‌هایی که $x_3 = 0$ و $x_2 = 0$ و $x_1 = 0$
- $N_{312} = 1$ تعداد حالت‌هایی که $x_3 = 1$ و $x_2 = 0$ و $x_1 = 0$
- $N_{321} = 0$ تعداد حالت‌هایی که $x_3 = 0$ و $x_2 = 1$ و $x_1 = 0$
- $N_{322} = 1$ تعداد حالت‌هایی که $x_3 = 1$ و $x_2 = 1$ و $x_1 = 0$

- $N_{331} = 1$ تعداد حالت‌هایی که $x_3 = 0$ و $x_2 = 0$ و $x_1 = 1$
- $N_{332} = 0$ تعداد حالت‌هایی که $x_3 = 1$ و $x_2 = 0$ و $x_1 = 1$
- $N_{341} = 0$ تعداد حالت‌هایی که $x_3 = 0$ و $x_2 = 1$ و $x_1 = 1$
- $N_{342} = 4$ تعداد حالت‌هایی که $x_3 = 1$ و $x_2 = 1$ و $x_1 = 1$
- $N_{31} = N_{311} + N_{312} = 4$
- $N_{32} = N_{321} + N_{322} = 1$
- $N_{33} = N_{331} + N_{332} = 1$
- $N_{34} = N_{341} + N_{342} = 4$

$$f(3, \{x_1, x_2\}) = \frac{1}{5!} * 3! * 1! * \frac{1}{2!} * 0! * 1! * \frac{1}{2!} * 1! * 0! * \frac{1}{5!} * 0! * 4! = \frac{1}{400}$$

vi. ازآنجائیکه $P_{new} = \frac{1}{400} < P_{old} = \frac{1}{180}$، بنابراین تکرار برای i=3 بوسیله $\pi_3 = \{x_2\}$ به اتمام می‌رسد.

خروجی: برای هر گره اولیاء آن تعیین شود.

$node\ x_1 \rightarrow \pi_1 = \emptyset$

$node\ x_2 \rightarrow \pi_2 = \{x_1\}$

$node\ x_3 \rightarrow \pi_3 = \{x_2\}$

بنابراین ساختار شبکه بر اساس مجموعه داده‌های جدول ۴-۲ بصورت زیر خواهد شد:

$$x_1 \rightarrow x_2 \rightarrow x_3$$

با انتخاب ساختار گرافیکی بدست آمده و استفاده از استقلال شرطی، توزیع توام متغیرها بصورت زیر خواهد شد:

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2) \qquad (9-4)$$

با در نظر گرفتن متغیر x_1 به عنوان متغیر کلاس یا پاسخ، برای اختصاص دادن مشاهدات به این کلاس با استفاده از معادله بالا احتمال پسین آن‌را محاسبه کرده و بیشترین مقدار را به آن کلاس اختصاص می‌دهیم. حال با توجه به پیدا کردن ساختاری مناسب برای متغیرها، برای محاسبه احتمالات محلی نیاز به برآورد آن‌ها داریم که به عنوان یادگیری پارامتری شبکه بیزی نام برده می‌شود. در بخش بعد برآورد احتمالات فوق را بیان می‌کنیم.

۸-۴ یادگیری پارامتری (احتمالات) در شبکه بیزی

معمولا احتمالات زیادی در مدل سهیم هستند و برای برآورد مشاهدات به کلاس مورد نظر، نیاز به تعیین تعداد زیادی از آنها هستیم. برای مثال، در مسئله مربوطه در رابطه با کشف تقلب، می‌خواهیم احتمالات تقلب در مشاهدات متغیرهای داده شده را بدانیم. این احتمالات بطور مستقیم در مدل ذخیره نشده‌اند و باید آن‌را محاسبه کرد. به همین منظور باید برآوردی مناسب را برای محاسبه احتمالات ارائه کرد. مانند روش بیزی خام، احتمالات مورد استفاده در مدل را به عنوان پارامتر می‌دانیم و بصورت زیر بیان می‌کنیم:

$$\theta_{ijk} = p(X_i = k | \pi_i = j)$$

θ_{ijk}، احتمال شرطی X_i در k- امین حالت، با توجه به j- امین حالت از π_i می‌باشد[44].

هکرمن[1] فرمول زیر را به‌عنوان برآوردی مناسب برای محاسبه احتمالات در مدل شبکه بیزی ارائه داده است[41]:

$$\Rightarrow \hat{\theta}_{ijk} = \frac{1 + N_{ijk}}{r_i + N_{ij}} \qquad (10-4)$$

که N_{ijk}، تعداد حالت‌هایی است‌که، متغیر X_i حالت k- ام و اولیاء آن حالت j-ام را در مجموعه داده D شامل می‌شود و $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. بنابراین با توجه به برآورد بالا می‌توان حالت‌هایی که در گره

[1] Heckerman

مشاهده نشده است را بر اساس گره‌های موجود با محاسبه احتمالات فیزیکی آن پیش‌بینی کرد. برای مثال با توجه به مشاهدات جدول ۴-۲، می‌خواهیم بدانیم مشاهدات $(x_2 = 1, x_3 = 1)$ به کدام کلاس تعلق دارد. با توجه به ۴-۷ داریم:

$$p(x_1 = 0 \mid x_2 = 1, x_3 = 1) \propto p(x_1 = 0)p(x_2 = 1|x_1 = 0)p(x_3 = 1|x_2 = 1)$$

بنابراین پارامترهای فوق را با توجه به ۴-۱۰ محاسبه می‌کنیم:

$$p(x_1 = 0) = \hat{\theta}_{1-1} = \frac{1 + N_{1-1}}{r_i + N_{1-}} = \frac{1+5}{2+10} = \frac{1}{2}$$

$$p(x_2 = 1|x_1 = 0) = \hat{\theta}_{212} = \frac{1 + N_{212}}{r_i + N_{21}} = \frac{1+1}{2+5} = \frac{2}{7}$$

$$p(x_3 = 1|x_2 = 1) = \hat{\theta}_{322} = \frac{1 + N_{322}}{r_i + N_{32}} = \frac{1+5}{2+5} = \frac{6}{7}$$

بنابراین احتمال تعلق مشاهده $(x_2 = 1, x_3 = 1)$ به کلاس ۰ برابر است با :

$$p(x_1 = 0 \mid x_2 = 1, x_3 = 1) \propto \frac{1}{2} * \frac{2}{7} * \frac{6}{7} = \frac{6}{49}$$

از طرف دیگر داریم :

$$p(x_1 = 1) = \hat{\theta}_{1-2} = \frac{1 + N_{1-2}}{r_i + N_{1-}} = \frac{1+5}{2+10} = \frac{1}{2}$$

$$p(x_2 = 1|x_1 = 1) = \hat{\theta}_{222} = \frac{1 + N_{222}}{r_i + N_{22}} = \frac{1+4}{2+5} = \frac{5}{7}$$

$$p(x_3 = 1|x_2 = 1) = \hat{\theta}_{322} = \frac{1 + N_{322}}{r_i + N_{32}} = \frac{1+5}{2+5} = \frac{6}{7}$$

$$p(x_1 = 1 \mid x_2 = 1, x_3 = 1) \propto \frac{1}{2} * \frac{5}{7} * \frac{6}{7} = \frac{15}{49}$$

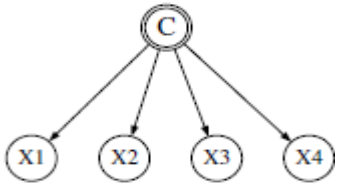
از آنجائی‌که:

$$p(x_1 = 1 \mid x_2 = 1, x_3 = 1) > p(x_1 = 0 \mid x_2 = 1, x_3 = 1)$$

بنابراین مشاهده $(x_2 = 1, x_3 = 1)$ را به کلاس ۱ اختصاص می‌دهیم.

۴-۹ ارتباط روش بیزی خام و شبکه بیزی

با توجه به تعاریفی که از این دو روش داشتیم، در این بخش ارتباط میان آن‌ها را مطالعه می‌کنیم. روش بیزی خام بر اساس فرضیه قوی استقلال شرطی میان متغیرها با توجه به متغیر پاسخ داده شده درآن بیان می‌شد در صورتیکه روش شبکه بیزی بر اساس وابستگی میان متغیرها. شبکه بیزی و بیزی خام را می‌توان گراف بدون‌دور جهت‌داری تعریف نمود که در روش بیزی خام بر اساس فرضیه موجود در آن متغیرها یا همان گره‌ها باهم در ارتباط نیستند ولی در روش شبکه بیزی به دلیل عدم وجود این فرضیه استقلال شرطی، متغیرها دارای وابستگی می‌باشند و گراف آن با گراف بیزی خام متفاوت خواهد بود. گراف زیر نمونه‌ای از ساختار بیزی خام می‌باشد که متغیرها با توجه به کلاس داده شده در آن مستقل از هم می‌باشند[45].



شکل(۴-۷)- ساختار گرافیکی مدل بیزی خام

با توجه به ساختار گرافیکی مدل بیزی خام توزیع توام مشاهدات با توجه به استقلال شرطی میان متغیرها بصورت زیر خواهد شد :

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i) = p(x_1 \mid c)p(x_2 \mid c)p(x_3 \mid c)p(x_4 \mid c)$$

با توجه به معادله بالا می‌توان نتیجه گرفت روش بیزی خام یک شبکه بیزی است‌که در آن متغیرها با توجه به فرضیه استقلال شرطی باهم مستقل می‌باشند.

نتیجه‌گیری

شبکه بیزی، گراف بدون‌دور جهت‌داری است‌که گره‌ها همان متغیرها هستند. به این دلیل جهت‌دار گویند که یال‌ها که نشان‌دهنده وابستگی میان متغیرها می‌باشند، فقط در یک جهت قرار دارند. بدون‌دور نیز به این دلیل گویند که، یک گره در یک چرخه به اولیای خود باز نمی‌گردد (گره‌ها در یک سیکل قرار ندارند). شبکه بیزی مدلی‌است گرافیکی برای بیان روابط احتمالاتی میان مجموعه‌ای از متغیرها. شبکه بیزی به دلایلی عمده و مهم به عنوان تکنیک مناسبی در روش‌های داده‌کاوی همواره پیشنهاد می‌شود. اول اینکه، توانایی اجرا برای مجموعه داده‌های ناقص (دارای داده گمشده) را داراست. دوم اینکه به یادگیری روابط علت و معلولی می‌پردازد. همچنین روش‌های بیزی در ارتباط با شبکه بیزی و انواع دیگر از مدل‌های پیشنهاد شده رویکرد کارآمد و اصولی برای جلوگیری از بیش-براوردی در داده‌ها را ممکن می‌سازد و باعث کاهش بعد نیز می‌شود. شبکه‌های بیزی بر اساس استقلال شرطی توام میان زیرمجموعه‌ای از متغیرها تعریف شده است.

شبکه بیزی با تعیین ۲ فاکتور زیر بیانگر توزیع احتمال توام می‌باشند :

۱. مجموعه‌ای از استقلال شرطی متغیرها
۲. احتمالات هر متغیر با توجه به اولیا آنها

که این اطلاعات برای هر متغیر فراهم می‌شود.

فصل پنجم

نتیجه‌گیری و پیشنهادات

۵-۱ مقدمه

با توجه به روش‌های طبقه‌بندی بیان شده، برای آن‌که بتوانیم مدلی را انتخاب کنیم که بهترین دقت پیش‌بینی را داشته باشد، نیاز به ارزیابی آن‌ها با استفاده از مجموعه داده‌ای مورد نظر خواهیم داشت. در این فصل ابتدا روشی را برای پردازش اطلاعات هنگامی‌که مجموعه داده‌های مورد بررسی دارای حجم زیادی می‌باشند ارائه داده، سپس به ارزیابی مدل‌های طبقه‌بندی خواهیم پرداخت.

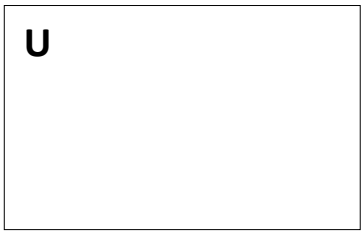
۵-۲ پردازش مجموعه داده در ارتباط با موضوع کشف تقلب

همان‌طور که در فصل۲ بیان کردیم، موضوع کشف تقلب با توجه به محدودیتی که دارد، در اختیار داشتن داده و اطلاعات بسیار مشکل است. از طرف دیگر با فرض در اختیار داشتن داده، به دلیل حجم عظیم آن پردازش اطلاعات بسیار سخت و زمان‌بر می‌باشد. چالش دیگری که در رابطه با این موضوع وجود دارد، عدم تعادل در داده‌ها می‌باشد. رخ دادن یک تقلب تنها از میان تعداد زیادی از مشاهدات امکان می‌پذیرد و این‌طور نیست که مشاهداتی که در آن‌ها تقلب صورت می‌پذیرد تعداد زیادی می-باشند. به عنوان مثال در [15] از بین ۵۰ میلیون کارت اعتباری تنها ۲۵۰۰ تراکنش جعلی یا تقلب در آن صورت پذیرفته است. بر همین اساس با فرض در اختیار داشتن داده نیز پردازش این اطلاعات فرایندی سخت و زمان‌بر می‌باشد. با توجه به این چالش، باید به فکر راه‌حلی برای کاهش حجم داده باشیم. در این بخش روشی را برای کاهش حجم داده‌ها در ارتباط با موضوع کشف تقلب ارائه می-دهیم.

داده‌های مورد بررسی در رابطه با تقلب در کارت اعتباری می‌باشد و با توجه به تراکنش‌های صورت پذیرفته توسط هر کارت، فرایند کاهش حجم انجام می‌پذیرد. تراکنش‌های مختلفی را می‌توان با یک کارت اعتباری انجام داد، در این بررسی به تراکنش‌هایی خواهیم پرداخت که حداقل یکبار در آن تقلب صورت پذیرفته است. این مطالعه بر اساس اطلاعات موجود در مجموعه داده [15] می‌باشد به دلیل در اختیار نداشتن آن، تنها به ارائه فرایند کاهش حجم در این مجموعه داده می‌پردازیم.

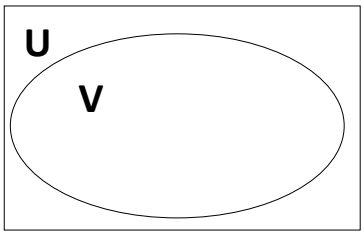
این مجموعه داده حدودا شامل ۵۰ میلیون تراکنش توسط ۱ میلیون کارت اعتباری در مدت زمان ۱۳ ماه (از ژانویه ۲۰۰۶ تا ژانویه ۲۰۰۷) می‌باشد. همان‌طور که بیان شد پردازش این حجم اطلاعات، بسیار زمان‌بر است و از آنجائیکه نسبت متعادلی میان تراکنش‌های جعلی و غیر جعلی وجود ندارد، مجموعه داده‌ها دارای عدم تعادل یا بالانس می‌باشند و مدل‌سازی با توجه به تمام داده‌ها مناسب نخواهد بود. در این مطالعه، مجموعه داده‌ها را به دسته‌های مختلفی تقسیم‌بندی کرده تا با مجموعه داده‌ای با حجم مناسب و کم فرایند پردازش صورت پذیرد. مجموعه U را به عنوان تمام تراکنش‌های انجام شده توسط کارت‌های اعتباری در نظر می‌گیریم.

۵۰ میلیون تراکنش:U



تراکنش‌های مختلفی با استفاده از کارت اعتباری انجام می‌گیرد، ما به دنبال تراکنش‌هایی هستیم که حداقل یکبار در آن تقلب صورت پذیرفته است. تراکنش‌هایی که در مجموعه داده U، با توجه به هیچ کارت اعتباری در آن تقلب صورت نگرفته است را حذف می‌کنیم تا مجموعه داده کوچکتری را در اختیار داشته باشیم. این مجموعه داده را V می‌نامیم.

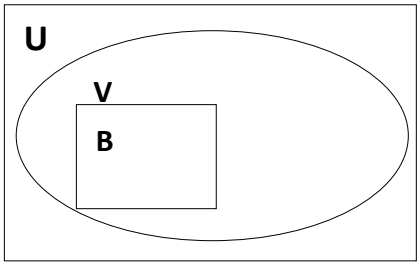
۳۰ میلیون تراکنش:V



با انجام این عمل، حجم داده از ۵۰ میلیون به ۳۰ میلیون تراکنش کاهش می‌یابد. در مجموعه داده V، تراکنش‌های جعلی وجود دارند که مشخص نیست به کدام نوع از تراکنش‌ها تعلق دارند و فقط نسبت به جعلی بودن آن با توجه به یک کارت اعتباری مطلع می‌باشیم. تمام تراکنش‌های جعلی

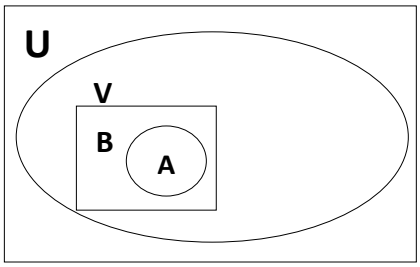
صورت گرفته در مجموعه V، شامل تراکنش‌های جعلی شناخته‌شده و تراکنش‌های جعلی شناخته-نشده، ۳۷ هزار تراکنش می‌باشد که در مجموعه B قرار دارد.

۳۷ هزار تراکنش: B



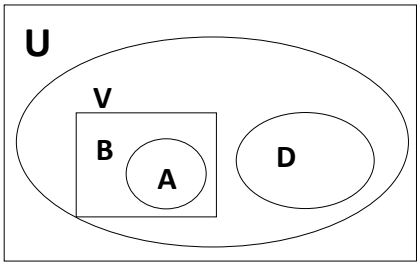
در این بررسی به دنبال تراکنش‌های جعلی شناخته شده هستیم و برای مقایسه تکنیک‌های مختلف در داده‌کاوی برای پیش‌بینی تقلب در کارت اعتباری، نیاز به مجموعه داده‌ای داریم که مشاهدات در آن کاملا معلوم و تعریف شده باشند. بر همین اساس در مجموعه B تراکنش‌های جعلی شناخته نشده را حذف می‌کنیم و مجموعه حاصل شده از تراکنش‌های جعلی شناخته‌شده را A می‌نامیم. این مجموعه شامل ۲هزار تراکنش می‌باشد.

۲ هزار تراکنش: A



مجموعه داده A را مجموعه داده تقلب می‌نامیم. باید توجه داشت تنها مجموعه داده تقلب برای پیش-بینی بر اساس تکنیک‌های داده‌کاوی کافی نمی‌باشد و نیاز به ترکیبی از تراکنش‌های جعلی و غیرجعلی داریم. تراکنش‌های غیر جعلی در مجموعه V، مجموعه‌ای شامل تمام تراکنش‌ها به‌جز تراکنش‌های مجموعه B می‌باشد. همان‌طور که از ابتدای بحث بیان کردیم، به دنبال کاهش حجم برای پردازش اطلاعات هستیم. چالش دیگر، عدم تعادل در مجموعه داده می‌باشد. به‌همین منظور

نمونه‌ای تصادفی از مجموعه V-B به عنوان داده‌های غیرجعلی انتخاب می‌کنیم و آن‌را مجموعه D می‌نامیم. شکل زیر حالت کلی تقسیم‌بندی مجموعه داده اولیه (U) می‌باشد.



D: ۳۴۰ هزار تراکنش

حال مجموعه‌ای شامل داده‌های جعلی و مجموعه‌ای شامل داده‌های غیرجعلی داریم که بر اساس این دو مجموعه فرایند پردازش اطلاعات را انجام می‌دهیم. این تقسیم‌بندی ایده‌ای مناسب برای کاهش مجموعه داده‌هایی که حجم بسیار زیادی دارند می‌باشد. در داده‌هایی مانند داده‌های تقلب در کارت اعتباری که تعادل یا بالانس بودن مشاهدات دیده نمی‌شود، این تقسیم‌بندی روشی مناسب برای تجزیه و تحلیل اطلاعات می‌باشد. در مجموعه داده‌ای که شامل ۵۰ میلیون مشاهده بوده است، بر اساس این روش، فرایند پردازش اطلاعات با مجموعه داده‌ای حدودا شامل ۳۴۲ هزار مشاهده انجام می‌پذیرد و این مجموعه داده به داده‌های آموزشی و داده‌های آزمون تقسیم شده و فرایند پیش‌بینی مشاهدات انجام می‌پذیرد.

همان‌طور که بیان شد برای ارائه مدلی مناسب میان مدل‌های مورد بررسی، نیاز به ارزیابی آن‌ها داریم. در ادامه با استفاده از مجموعه داده‌ای شبیه به مجموعه داده تقلب، مدل‌های طبقه‌بندی را با ارائه معیارهایی برای سنجش دقت آن‌ها ارائه می‌دهیم.

۵-۳ مجموعه داده

مجموعه داده‌های مورد بررسی در این بخش، اطلاعاتی شامل ۴۶۰۰ مشاهده با ۵۷ متغیر توضیحی و ۱ متغیر پاسخ است. هدف، طبقه‌بندی مناسب بر اساس اطلاعات ورودی است. مجموعه داده‌های مورد

بررسی در این بخش، اطلاعاتی شامل فراوانی حروف و کاراکترهای مختلف، در ۴۶۰۰ نامه الکترونیکی می‌باشد. در این داده‌ها ۵۴ متغیر، مقادیری بین [0,100]، ۳ متغیر مقادیر [... ,1]، و متغیر پاسخ مقادیر ۰ و ۱ را در بر می‌گیرد. برای تجزیه و تحلیل مجموعه داده مورد بررسی در رابطه با موضوع کشف تقلب، مقادیر متغیر پاسخی که برابر با (spam = 1) باشد را تقلب (Fraud = 1)، و (spam = 0) را غیر تقلب (nonFraud = 0) در نظر می‌گیریم. یعنی مشاهداتی که در آن تقلب صورت پذیرفته را با (1) و مشاهداتی که جعلی نبوده را با (0) نشان می‌دهیم. بنابراین باید مدلی را ارائه دهیم که دقت طبقه‌بندی بالایی را برخوردار باشد و با داشتن مشاهدات جدید، جعلی یا غیر جعلی بودن آنرا به درستی تشخیص دهد. بر همین اساس نیاز به ارائه معیارهایی برای ارزیابی مدل‌ها داریم. در بخش بعد با ارائه معیارهای سنجش دقت مدل، به ارزیابی مدل‌های مورد بررسی می‌پردازیم.

۵-۴ معیارهای ارزیابی[1] مدل

برای ارزیابی دقت مدل‌های طبقه‌بندی، از معیارهای مختلفی استفاده می‌شود که در این بخش آنها را معرفی و مورد استفاده قرار می‌دهیم. در جدول ۵-۱ ماتریس حالات پیش‌بینی برای مسائل طبقه-بندی با دو کلاس نشان داده شده است که در آن مشاهده‌ای که متغیر پاسخ مربوط به آن ۱ است را به عنوان مشاهده مثبت (P) و مشاهده‌ای که متغیر پاسخ مربوط به آن ۰ است را مشاهده منفی (N) می‌نامیم[15].

جدول (۵-۱) ماتریس حالات پیش بینی برای طبقه بندی با دو کلاس

	پیش بینی مثبت (1)	پیش بینی منفی (0)
مشاهده مثبت (1)	TP	FN
مشاهده منفی (0)	FP	TN

که در آن TP[۱] تعداد مشاهدات مثبتی است که مثبت پیش‌بینی شده‌اند، FN[۲] تعداد مشاهدات مثبتی است که منفی پیش‌بینی شده‌اند، FP[۳] تعداد مشاهدات منفی است که مثبت پیش بینی شده-

[1] Performance measures

اند و TN[4] تعداد مشاهدات منفی است که منفی پیش بینی شده اند. حال با توجه به ماتریس حالات پیش‌بینی، معیارهای ارزیابی مدل را بیان کرده و با توجه به مجموعه داده مورد بررسی به تجزیه و تحلیل آن می‌پردازیم.

معیارهای ارزیابی مدل را بصورت زیر تعریف می‌شوند:

$$Recall = TPrate = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$FPrate = \frac{FP}{TP + FP}$$

$$\text{F} - \text{Measure} = \sqrt{\text{Precision} * \text{Recall}}$$

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

$$Misclassification = 1 - Accuracy$$

که در آن $Recall$ نسبتی از کل مشاهدات مثبت است‌که درست طبقه‌بندی شده‌اند، $Precision$ نسبتی از پیش‌بینی‌های مثبت است که واقعاً مثبت بوده‌اند، $FPrate$ نسبتی از پیش‌بینی‌های مثبت است که اشتباه طبقه‌بندی شده‌اند، اندازه F، میانگین هندسی $Recall$ و $Precision$ است که کارایی $Recall$ را نسبت به$Precision$ کنترل می‌کند، میزان دقت یا $Accuracy$ معرف نسبت مشاهداتی است که درست پیش‌بینی شده‌اند و $Misclassification$ نسبت مشاهداتی استکه به اشتباه طبقه-بندی شده است. مقادیر بزرگ معیارهای $Recall$ ، $Precision$، F و $Accuracy$ و مقادیرکوچک معیارهای $Missclassification$ و $FPrate$ مبین مناسب‌تر بودن مدل و عملکرد بهتر آن‌ها است. حال به نتایج حاصل توسط مدل‌های بیان شده بر روی مجموعه داده تعریف شده می‌پردازیم.

---

[1] True positives
[2] False negatives
[3] False positives
[4] True negatives

۵-۵ ارزیابی مدل

ابتدا داده‌ها را بطور تصادفی به دو دسته تقسیم کرده و ۷۰٪ داده‌ها را بعنوان داده‌های مدل‌ساز و ۳۰٪ آن را بعنوان داده‌های آزمون در نظر می‌گیریم. برای تجزیه و تحلیل داده‌ها، ابتدا مدل‌های موجود را بر روی داده‌های مدل‌ساز برازش داده و سپس با استفاده از معیارهای ارزیابی مدل، عملکرد آن‌ها را توسط داده‌های آزمون مورد ارزیابی قرار می‌دهیم. در جدول ۵-۲ مقادیر حالات پیش‌بینی برای هر سه مدل با توجه به داده‌های آزمون است. اعداد جدول ۵-۲ بر اساس محاسبات در نرم افزار وکا[1] بدست آمده است.

جدول (۵-۲) مقادیر ماتریس حالات پیش‌بینی بر روی داده‌های آزمون

		پیش بینی مثبت	پیش بینی منفی
مشاهده مثبت	شبکه بیزی	۴۳۹	۸۷
	بیزی خام	**۵۰۸**	**۱۸**
	لجستیک	۴۷۶	۵۰
مشاهده منفی	شبکه بیزی	۴۴	۸۱۰
	بیزی خام	۲۴۰	۶۱۴
	لجستیک	۵۰	۸۰۴

برای تجزیه و تحلیل مدل‌های بیان شده، باید موضوع مورد بررسی را در نظر داشت و بسته به اهمیت کلاس‌های متغیر پاسخ، به ارزیابی مدل‌های بیان شده پرداخت. در ارتباط با موضوع کشف تقلب، اهمیت مشاهده مثبت یا تقلب بیشتر از کلاس دیگر است و باید مدلی را در نظر گرفت که مشاهداتی که در آن‌ها تقلب صورت گرفته را تقلب تشخیص داده و دارای اشتباه کمتری باشد. برای مثال اگر مشاهده‌ای که در آن تقلب صورت گرفته را به اشتباه تشخیص دهیم هزینه زیادی را از این طبقه‌بندی اشتباه از دست خواهیم داد، در صورتی‌که اهمیت اشتباه تشخیص دادن مشاهده غیرتقلب به کلاس تقلب بسیار کمتر می‌باشد. با توجه به اهمیت کلاس تقلب، مدلی را در نظر می‌گیریم که مقدار TP آن بیشتر باشد. بنابراین با در نظر داشتن مقادیر حالات پیش‌بینی که در جدول ۵-۲ آورده

[1] Weka software

شده است، مدل بیزی خام عملکرد مناسب‌تری نسبت به دو مدل دیگر دارد. اما باید به این نکته توجه داشت که مقدار بالای TP نشان‌دهنده مدلی مناسب میان این سه روش نخواهد بود زیرا، همان‌طور که در جدول ۵-۲ مشاهده می‌کنید مقدار FP یعنی تعداد مشاهدات منفی که مثبت پیش بینی شده‌اند در مدل بیزی خام ۲۴۰ می‌باشد که نشان‌دهنده عدم دقت در پیش‌بینی مشاهدات است. به-همین خاطر برای ارائه مدلی مناسب، نیاز به عملکرد کلی آن مدل در پیش‌بینی مشاهدات خواهیم داشت. حال با توجه به مقادیر ماتریس حالات پیش‌بینی، معیارهای ارزیابی مدل را محاسبه می‌کنیم و با توجه به نتایج بدست آمده، به ارزیابی ۳ مدل می‌پردازیم. نتایج معیارهای ارزیابی مدل در جدول ۵-۳ آورده شده است.

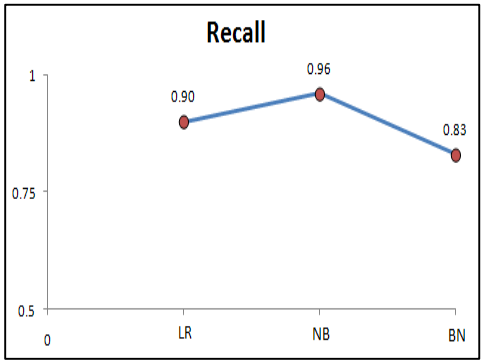
جدول (۵-۳) نتایج معیارهای ارزیابی مدل

معیارهای ارزیابی مدل	رگرسیون لجستیک	بیزی خام	شبکه بیزی
Recall	۰/۹۰	**۰/۹۶**	۰/۸۳
precision	۰/۹۰	۰/۶۷	۰/۹۰
FPrate	۰/۰۵	۰/۲۸	۰/۰۵
F-Measure	۰/۹۰	۰/۷۹	۰/۸۷
Accuracy	**۰/۹۲**	۰/۸۱	**۰/۹۱**
Misclassification	۰/۰۸	۰/۱۹	۰/۰۹

۵-۶ ارزیابی نتایج

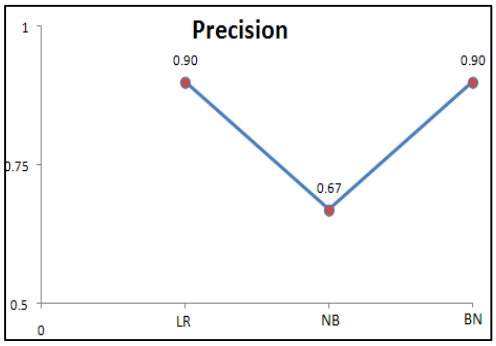
با توجه به مطالبی که برای انتخاب بهترین مدل در ارتباط با موضوع کشف تقلب بیان کردیم، معیاری را باید در نظر گرفت که مشاهدات جعلی را درست تشخیص دهد. بنابراین از میان معیارهای ارزیابی مدلی که ارائه شده است، معیار $Recall$ به دلیل اینکه بیان‌کننده نسبت کل مشاهدات مثبتی است که درست طبقه‌بندی شده‌اند، اهمیت بیشتری را داراست. با توجه به جدول ۵-۳ معیار $Recall$ برای مدل بیزی خام ۰/۹۶ می‌باشد که از دو مدل دیگر عملکرد بهتری در پیش‌بینی مشاهدات مثبت یا تقلب داشته است. بنابراین بیزی خام در ارتباط با موضوع کشف تقلب مدلی مناسب‌تر از رگرسیون

لجستیک و شبکه بیزی می‌باشد. اما باید توجه داشت بیزی خام، بهترین مدل در میان این سه مدل نخواهد بود. همان‌طور که بیان شد مقادیر بزرگ معیارهای $Recall$ ، $Precision$، F و $Accuracy$ و مقادیرکوچک معیارهای $Missclassification$ و $FPrate$ مبین مناسب‌تر بودن مدل و عملکرد بهتر آن‌ها است. بنابراین در حالت کلی با مقایسه این سه مدل مشاهده می‌شود که تنها معیار $Recall$ در مدل بیزی خام مناسب‌تر از دو مدل دیگر است و با توجه به معیارهای دیگر، دارای عملکرد ضعیف‌تری می‌باشد. شکل‌های زیر مقایسه این سه مدل را بر اساس معیارهای ارزیابی مدل نشان می-دهد که نشان‌دهنده عملکرد بهتر رگرسیون لجستیک و شبکه بیزی نسبت به بیزی خام است.



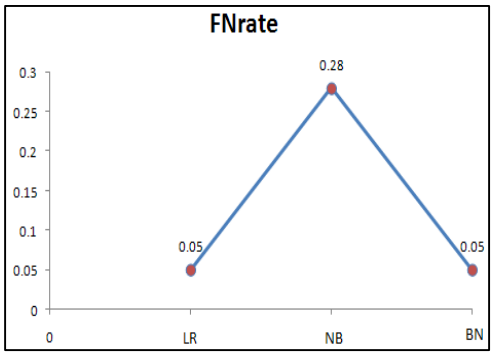
شکل (۵-1) نمودار معیار Recall برای مدل‌های طبقه‌بندی

نمودار ۵-۱، معیار $Recall$ که نسبتی از کل مشاهدات مثبت است که درست طبقه‌بندی شده‌اند، در سه روش طبقه‌بندی نشان می‌دهد. مشاهده می‌کنید که در این نمودار روش بیزی خام از دو روش دیگر بهتر بوده و با دقت ۹۶٪ مشاهداتی را که مثبت بوده‌اند را درست پیش‌بینی کرده است. مناسب بودن این معیار برای روش بیزی خام به تنهایی نمی‌تواند بیانگر مدلی مناسب میان دو مدل دیگر باشد بنابراین به تجزیه و تحلیل معیارهای دیگری برای ارزیابی مدل می‌پردازیم.



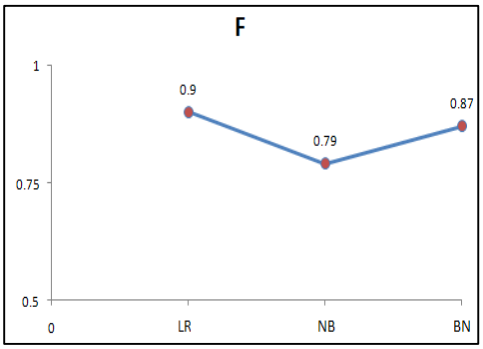
شکل (۵-۲) نمودار معیار Precision برای مدل‌های طبقه‌بندی

نمودار ۵-۲، معیار *precision* که نشان‌دهنده نسبتی از پیش‌بینی‌های مثبت است‌که واقعاً مثبت بوده را برای روش‌های لجستیک و شبکه بیزی نشان می‌دهد که در آن، رگرسیون لجستیک و شبکه بیزی عملکرد یکسانی داشته‌اند، در مقابل روش بیزی خام عملکرد نامناسبی دارد. این نمودار نشان می‌دهد که، رگرسیون لجستیک و شبکه بیزی، مشاهداتی را که واقعا مثبت بوده‌اند را با ٪۹۰ دقت درست پیش‌بینی می‌کند ، در صورتی‌که روش بیزی خام ٪۶۷ دقت را داراست.



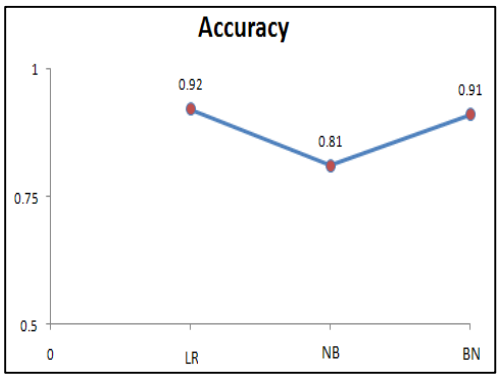
شکل (۵-۳) نمودار معیار FPrate برای مدل‌های طبقه‌بندی

در نمودار ۵-۳، نسبتی از پیش‌بینی مشاهدات مثبت که اشتباه طبقه‌بندی شده‌اند را نشان داده شده است. همان‌طور که مشاهده می‌کنید، عملکرد بیزی خام ضعیف‌تر از دو روش دیگر است و روش‌های لجستیک و شبکه بیزی عملکرد یکسانی دارند و خطای طبقه‌بندی آنها برابر ۰/۰۵ می‌باشد.



شکل(۵-۴) نمودار معیار F برای مدل‌های طبقه‌بندی

همان‌طور که در نمودار بالا مشخص است، معیار F نیز که کارایی *Recall* را نسبت به *Precision* کنترل می‌کند و هرچه مقدار آن بیشتر باشد عملکرد مناسب مدل را نشان می‌دهد، روش بیزی خام در میان سه روش ضعیف‌تر بوده و روش رگرسیون لجستیک از دو روش دیگر بهتر است.



شکل(۵-۵) نمودار معیار Accuracy برای مدل‌های طبقه‌بندی

معیار $Accuracy$ نشان‌دهنده دقت مدل در پیش‌بینی مشاهدات می‌باشد. این معیار کمک فراوانی را برای انتخاب بهترین مدل در ارزیابی مدل‌های طبقه‌بندی به ما خواهد کرد. با توجه به نمودار بالا، میزان دقت مدل در روش‌های لجستیک و شبکه بیزی تقریبا یکسان بوده است، در صورتی‌که دقت مدل بیزی خام کمتر از دو مدل دیگر است. بنابراین می‌توان نتیجه گرفت مدل بیزی خام مدلی مناسب نخواهد بود.

۵-۷ نتیجه‌گیری

موضوع کشف تقلب به دلیل آن‌که پیش‌بینی مشاهدات تقلب اهمیت بیشتری را دارد، معیار $Recall$ در میان معیارهای ارزیابی مدل دارای اهمیت بیشتری است. معیار $Recall$ در مدل بیزی خام بیشترین مقدار را نسبت به دو مدل دیگر دارد و مقدار ۰/۹۶ را نشان می‌دهد، اما باید توجه داشت نمی‌توان بر اساس تنها یک معیار بهترین مدل را انتخاب کرد. همان‌طور که در شکل‌های مقایسه مدل‌های طبقه‌بندی مشاهده کردید، شبکه بیزی و رگرسیون لجستیک عملکردی بهتری را نسبت به بیزی خام نشان می‌دهد. میزان دقت یا $Accuracy$ که معرف نسبت مشاهداتی است که درست پیش‌بینی شده‌اند در این دو مدل تقریبا برابر بوده و ۱۰٪ بیشتر از مدل بیزی خام می‌باشد که ۸۱٪ دقت را نشان می‌دهد. بنابراین با توجه به ارزیابی کلی از نتایج بدست آمده، مدل لجستیک و شبکه بیزی دارای عملکردی تقریبا برابر و بهتر از بیزی خام می‌باشند. بنابراین می‌توان نتیجه گرفت، با وجود اهمیتی که معیار $Recall$ در ارتباط با موضوع کشف تقلب خواهد داشت، تنها این معیار ملاک مناسبی برای انتخاب مدل نمی‌باشد.

۵-۸ تحقیقات پیشنهادی

- در صورت در دسترس بودن مجموعه داده، ارزیابی مدل‌های طبقه‌بندی دیگر
- تحقیق در زمینه مدل بیزی خام هنگامیکه تحت فرضیه بیزی خام گوسی نباشد
- بررسی مدل شبکه بیزی با متغیرهای پیوسته
- بررسی مدل شبکه بیزی با مجموعه داده ناقص
- استفاده از الگوریتم دیگری برای یافتن ساختاری مناسب در مدل شبکه بیزی

منابع:

مقالات مجلات علمی

*[1]. Ngai E.W.T. and Hu Y. Wong Y.H and Chen Y.J. and Sun X. (**2011**)" The application of data mining techniques in financial fraud detection :A classification framework and an academic review of literature"* ***J. Decision Support Systems****, p.p* ***559–569***

*[3]. Manuela P. and Antonello P. (**2010**) "A discrete choice approach to model credit card fraud"* ***MPRA Paper No. 20019***

*[4] Panigrahi S. and Kundu A. and Sural S. , A.K. Majumdar (**2009**)" Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning"* ***J. Information Fusion****, p.p* ***354–363***

*[8]. Bose I. and Mahapatra I.R.K. (**2001**) "Business data mining — a machine learning perspective"* ***J. Information Management****, p.p* ***211–225***

[9]. Turban E. and Aronson J.E. and Liang T.P. and Sharda R. (2007). Decision Support and Business Intelligence Systems, New Jersey, Pearson Prentice Hall.

*[10]. Frawley W.J. and Shapiro G. And Matheus C.J. (**1992**) "Knowledge discovery in databases an overview"* ***AI Magazine, 13, 3,*** *pp* ***57–70****.*

*[12]. Phua C. And Lee V. And Smith K. And Gayler R. (**2005**) " A comprehensive survey of data mining-based fraud detection"* ***research Artificial Intelligence Review,*** *p.p* ***1–14***

*[13]. Ravisankar P. And Ravi V . And Rao G. And Bose I. (**2011**) "Detection of financial statement fraud and feature selection using data mining techniques"* ***J. Decision Support Systems*** *, p.p* ***491–500***

*[14]. Bolton J . and Hand J .(**2002**) "Statistical Fraud Detection :A Review"* ***J. Statistical Science****, No. 3, p.p* ***235-255***

*[15]. Bhattacharyya S. and Jha S. and Tharakunnel K. and J.Christopher Westland J.C. (**2011**) "Data mining for credit card fraud :A comparative study"* ***J. Decision Support Systems*** *, p.p* ***602–613***

*[17] Adams N.and Whitrow C.and Hand D. and Weston D. (**2009**) "Credit Card Transaction Fraud Detection" .* ***Data Min. Knowl. Discov****, p.p.**30-55***

*[18] Sanchez D. and Vila M.A. and Cerda L. and Serrano J.M. (**2009**) "Association rules applied to credit card fraud detection"**J. Expert Systems with Applications**, p.p **3630–3640***

*[19] Gadi M.F.A. and Wang X. and do Lago A.P .(**2008**) "Credit Card Fraud Detection with Artifcial Immune System"* ***J.Computer science***

*[20] Wong N. and Ray P. and Stephens G. and Lewis L. (**2012**) "Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results"**J. Info Systems**, p.p **53–76***

*[22] Whitrow C. and Hand D. J. and Juszczak P. and Weston D. and Adams N. M. (**2008**)"Transaction aggregation as a strategy for credit card fraud detection"* ***J.Springer Science and Business Media.***

*[23] Lenard M.J. and Alam P.(**2009**) "An Historical Perspective on Fraud Detection: From Bankruptcy Models to Most Effective Indicators of Fraud in Recent Incidents"* ***J.Forensic and Investigative Accounting, vol 1***

*[25] Kirkos E. and Spathis C. and Manolopoulos Y. (**2007**) "Data Mining techniques for the detection of fraudulent financial statements"* ***J. Expert Systems with Applications****,p.p **995–1003***

*[26] Delamaire L. and Abdou H. and Pointon J. (**2009**) "Credit card fraud and detection techniques: a review"* ***J.Banks and Bank Systems****, **Vol 4***

*[27] Ezawa K. and Norton S. (**1996**) "Constructing Bayesian Networks to Predict Uncollectible Telecommunications Accounts"**J. IEEE, Expert Intelligent Systems**, p.p **45 - 51***

*[28] Maes S. and Tuyls K. and Vanschoenwinkel B. and Manderick B. (**2002**) "Credit Card Fraud Detection using Bayesian and Neural Networks" available at www.personeel.unimaas.nl/k-tuyls/publications/papers/maenf02.pdf*

*[29] Gadi M.F.A. and Wang X. and do Lago A.P.(**2008**) "A comparasion of classification methods applyed on credit card fraud detection" available at www.ime.usp.br/~mgadi/Article_1_eng.pdf*

*[31] Aral K.D. and Güvenir H.A. and Sabuncuogluc I. and Akar A.R. (**2011**) "A prescription fraud detection model" **J. computer methods and programs in biomedicine**, **No. of Pages 10***

*[33] Bose I. and Piramuthu S. and Shaw M.J.(**2011**) "Quantitative methods for Detection of Financial Fraud" **J.Decision Support Systems**, p.p **557–558***

*[34] Phua C. and Lee V. and Smith K. and Gayler R. (**2005**) "A Comprehensive Survey of Data Mining-based Fraud Detection Research" **j.Artificial Intelligence**, available at http://www.citeulike.org/user/jydelort/article/3393241*

*[36] Yager R . (**2006**)"An extension of the naive Bayesian classifier "**J. Information Sciences**, p.p **577–588***

*[37] Soria D. and Garibaldi J.M. and Ambrogi F. and Biganzoli M.E. and Ellis I.O.(**2011**)" A non-parametric version of the naive Bayes classifier" **J. Knowledge-Based Systems**, p.p **775–784***

*[38] Chandra B. Gupta M. (**2011**) "Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data" **J. Expert Systems**, p.p **1293–1298***

*[39] Lu S. and Chiang D. and Keh H. and Huang H.(**2010**)" Chinese text classification by the Naïve Bayes Classifier and the associative classifier with multiple confidence threshold values" **J.Knowledge-Based Systems**, p.p **598–604***

*[40] Heckerman D. (**1997**) " Bayesian Networks for Data Mining" **J.Data Mining and Knowledge Discovery**, p.p **179–119***

*[41] Hassen B.H. and Masmoudi A. and Rebai A.(**2008**) "Causal inference in biomolecular pathways using a Bayesian network approach and an Implicit method "**J. Theoretical Biology**, p.p **717– 724***

*[43] Gregory F. and Cooper E. H.(**1993**) "A bayesian method for the induction of probabilistic networks from data" **J.Medical Computer Science**, Stanford University*

*[44] Bouchaala L. and Masmoudi A. and Gargouri F. and Rebai A.(**2010**) "Improving algorithms for structure learning in Bayesian Networks using a new implicit score" **J.Expert Systems with Applications**, p.p **5470–5475***

*[45] Aritz Perez A. and Larranaga P. and Inza I. (**2006**) "Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes" **International Journal of Approximate Reasoning**, p.p **1–25***

کتاب

*[7] Giudici P.(**2003**)" Applied Data Mining Statistical Methods for Business and Industry" Publisher: John Wiley & Sons, ISBN : 0470871393, 9780470871393*

*[16] Mitchell T. (**1997**)" Machine Learning" McGraw Hill, Chapter 1, ISBN: 0070428077*

*[35] Hastie T. and Tibshirani R. and Friedman J. (**2009**)" the elements of statistical learning" ISBN: 978-0387848570 , Edition:2nd ed.*

*[42] Rebai A. (**2010**)" Bayesian Network" available at **www.sciyo.com***

پایان‌نامه

*[5]. Paasch. C.A.W (**2008**), **P.h.D thesis**, "credit card fraud detection using artificial neural networks tuned by genetic algoritm" information and system management. Hong Kong university.*

*[6]. Strehl C. (**2005**) **Master-thesis,** "Fraud detection using cost-sensitive classification procedures in E–Business"*

کنفرانس

*[2]. Jans M. and Lybaert N. and Vanhoof K. (**2007**)" Data Mining for Fraud Detection : Toward an Improvement on Internal Control System" European A ccounting Association-Annual congress, URL:http://hdl.handle.net/1942/7872*

*[11]. Kou Y. And Lu C. And Sirwongwattana S. And Huang Y. (**2004**) " Survey of fraud detection techniques" IEEE International Conference on Networking, p.p **749–754***

*[21] Shen A. and Tong R. and Deng Y. (**2007**) "Application of Classification Models on Credit Card Fraud Detection" International Conference, Service Systems and Service Management*

*[24] Deng Q. (**2010**) "Detection of Fraudulent Financial Statements Based on Naïve Bayes Classifier" The 5th International Conference on Computer Science & Education, Nanchang,China.*

*[30] Gadi M.F.A. and Wang X. and do Lago A.P.(**2008**) "Comparison with Parametric Optimization in Credit Card Fraud Detection" Seventh International Conference Machine Learning and Applications*

*[32] Kirkos E. and Spathis C. and Manolopoulos Y. (**2005**) "Detection of Fraudulent Financial Statements through the use of Data Mining Techniques" 2nd International Conference on Enterprise Systems and Accounting,Thessaloniki,Greece.*

Abstract

The aim of this study is to investigate data mining and the use of its techniques in fraud detection and provide three classification methods. We describe financial classification and the place of data mining in this framework. Since the subject of fraud detection is related to classification methods in data mining , we describe three of them that are logistic Regression, Naive Bayes and Bayesian networks and investigate the relation between logistic Regression and Naive Bayes. For evaluating the stated models, some criteria is explained and finally the best model is chosen. Results showed that the logistic regression and Bayesian networks have the same performance and better than the Naive Bayes.

Keywords: *Fraud detection, Data mining, Logistic regression, Naive bayes, Bayesian network*

Shahrood University of Technology

Faculty of mathematical

Fraud detection using data mining techniques

sajjad rezaei

supervisor:

Dr.D.Shahsavani

September 2012