مجله علمی «نوآوری‌های فناوری اطلاعات و ارتباطات کاربردی»
سال اول، شماره ۴، فصل زمستان ۱۴۰۰؛ ص۱۵تا۳۲

ارائه راهکاری برای پیش‌بینی شرکت‌های متقلب مالیاتی مبتنی بر الگوریتم‌های داده‌کاوی درخت تصمیم بهینه‌شده، ماشین بردار پشتیبان و شبکه بیزین

امین قمری مقدم — دانشجو دکتری،رشته حسابداری،دانشگاه آزاد اسلامی واحد قائنات

حبیب اله نخعی — استادیار گروه حسابداری دانشگاه آزاد اسلامی واحد بیرجند

(دریافت: ۱۴۰۰/۱۰/۲۹، پذیرش: ۱۴۰۱/۰۱/۲۸)

چکیده

بسیاری از شرکت‌ها با تخلف در صورت‌های مالی موجب به‌هم‌ریختگی نظام اقتصادی می‌شوند و این امر موجب یک بحران مهم نظام اقتصادی شده است. راهکارهای مختلفی برای تشخیص وجود دارد که اغلب انسانی می‌باشند. این راهکارها دارای هزینه‌های بالایی برای محاسبه و بررسی صورت‌های مالی تمامی شرکت‌ها دارند ازاین‌جهت باید به دنبال راهکاری بود که بتواند با استفاده از داده‌کاوی و خودکار این فرایند تشخیص را انجام دهد. البته روش‌های داده‌کاوی نیز برای این قضیه ارائه‌شده‌اند که هر یک دارای مزایا و معایبی می‌باشند. روش‌های داده‌کاوی که تا به اینجا برای این کار ارائه شدند دارای سربار بالای محاسباتی و یا دقت پایین بودند. حال‌آنکه در روش پیشنهادشده در این تحقیق از درخت تصمیم‌گیری ID3 بهبودیافته به همراه شبکه بیزین و ماشین بردار پشتیبان به‌عنوان یک روش ترکیبی استفاده‌شده است. در این روش پیشنهادی برای بهبود عملکرد و دقت از الگوریتم مجموعه راف و تحلیل سلسله‌مراتبی در جهت انتخاب ویژگی‌های مؤثر کمک گرفته‌شده است. درختی که در روش پیشنهادی ایجاد می‌شود دارای کمترین عمق ممکن است و ازاین‌رو دارای سرعت بالایی است. سربار محاسباتی روش پیشنهادی به دلیل استفاده از الگوریتمی بهینه پایین می‌باشد. داده‌های استفاده‌شده در ارزیابی داده‌های مربوط به ۳ سال از ۶۰ شرکت است. در ارزیابی روش پیشنهادی نشان داده شده است که روش پیشنهادی دارای دقت ۸۰ درصد می‌باشد که نسبت به روش‌های مشابه به خود دقت بالایی به‌حساب می‌آید. سربار زمانی در روش پیشنهادی O(m.n) و سربار حافظه O(n) است که m، نشان‌دهنده اندازه مجموعه آموزش و n نشان‌دهنده مجموعه ویژگی مورد استفاده در آموزش می‌باشد.

کلیدواژه‌ها: شبکه بیزین، ماشین بردار پشتیبان، درخت تصمیم، فازی راف، تحلیل سلسله‌مراتبی، تقلب.

Providing a solution for predicting tax fraud companies based on optimized decision tree, support vector machine and Bayesian network

A.Ghamari Moghadm, H.Nakhaei

PhD Student Accounting Cours.Islamic Azad university.Ghaenat Branch.Ghaenat, Iran

Assistant professor.Department of accounting Birjand Branch Islamic Azad university, Birjand, Iran

(Received: 2022/January/19; Accepted: 2022/April/28)

Abstract

Many companies disrupt the economic system by violating their financial statements, which has led to a major economic crisis. There are various diagnostic strategies that are mostly human. These solutions have high costs for calculating and reviewing the financial statements of all companies, so we must look for a solution that can use data mining and automation to perform this diagnostic process. Of course, data mining methods have also been proposed for this case, each of which has advantages and disadvantages. The data mining methods presented so far have high computational overhead or low accuracy. However, in the method proposed in this research, the improved ID3 decision tree with Bayesian network and support vector machine has been used as a combined method. In this proposed method, to improve performance and accuracy, the rough set algorithm and hierarchical analysis are used to select effective features. The tree created in the proposed method has the lowest possible depth and therefore has a high velocity. The computational overhead of the proposed method is low due to the use of an optimal algorithm. The data used in the evaluation of 3-year data from 60 companies. In evaluating the proposed method, it is shown that the proposed method has an accuracy of 80%, which is considered to be high accuracy compared to similar methods. The time overhead in the proposed method is O(m.n) and the memory overhead is O(n) where m represents the size of the training set and n represents the feature set used in the training.

Keywords: Bayesian network, Support vector machine, decision tree, fuzzy rough, hierarchical analysis, fraud.

۱. مقدمه

کاربرد داده‌کاوی، تنها محدود به تعاملات اجتماعی، علـوم و مهندسـی نمـی‌شـود، بلکـه عـلاوه بـر آنهـا در سـامانه‌هـای پیشنهاددهنده، سامانه‌های مالی، ضدجاسوسی و غیره، نیـز مـورد استفاده قرار می‌گیرند. بااین‌وجود که روش‌هـای داده‌کـاوی دارای کاربردهای بسیاری در علوم مختلف هستند. اما تاکنون نرخ اتخاذ این روش‌ها در میان دانشگاهیان و سازمان‌های کنترلی به‌منظـور کشف تقلب و مخاطره، چندان چشمگیر نبـوده اسـت. در فصـول ابتدایی این پژوهش، بیان مسـئله و مفـاهیم مقـدماتی در دامنـه موردنظر بررسی مـی‌شـود. بـر اسـاس وقـایع و مشـاهدات اخیـر، سرقت‌های اطلاعاتی و رسوایی‌های مالی می‌توان استدلال کرد که احتمال وقوع تقلب‌های درون‌سازمانی در هر شرکت یا نهادی، اعم از تجاری و غیرتجاری وجود دارد و مختص به سطح یا رده خاصی از آن مجموعه‌ها نمی‌شود. به‌عنوان نمونه طبق گزارش ارائـه‌شـده توسط سازمان جهانی بررسـی جـرائم اقتصـادی در سـال ۲۰۲۰، "تقلب‌های شغلی" صـورت‌گرفتـه در کشـور فرانسـه حـدود ۵۶٪ جرائم مالی و اقتصادی این کشور را به خود اختصاص داده‌اند [۱] ، [۲] این مستندات و وقایع سبب می‌شوند تا شناسایی این گـروه از تخلفات اهمیـت ویـژه‌ای بیابـد و بـرای کشـف و شناسـایی آن روش‌های بسیاری ارائـه شـود. از جملـه نـوآوری‌هـا در شناسـایی تقلب، استفاده از روش‌هـای داده‌کـاوی اسـت [۳] ، [۲]. یکـی از دلایل مهمـی کـه موجـب شـده اسـت روش‌هـای داده‌کـاوی در تشخیص تقلب مورداستفاده قـرار گیرنـد، پویـا بـودن ذات تقلـب است. همین پویایی و تغییر دائمی سبب شده است، به‌روز بودن را به‌عنوان یک معیار مهـم، در اعتبارسـنجی ابـزار تشـخیص تقلـب قلمداد کنند. ازآنجایی‌که روش‌های داده‌کاوی بـه‌صـورت پویـا بـا محیط‌های در حال تغییر، سازگار مـی‌شـوند، مـدت‌زمـانی را کـه صرف کشف الگو می‌شود، تا حد قابل‌توجهی نسبت به روش‌هـای غیرخودکار (دستی) کاهش می‌دهنـد [۵]، [۴]. مـی‌تـوان کشـف تقلب و مدیریت مخاطره را به‌عنوان یکی از زمینه‌های تحقیقـاتی مهم و کاربردی در علم داده‌کاوی در نظر گرفت.

در کشور ما هیچ نهادی بـه طـور مسـتقیم بـرای تحقیـق و کشف موارد تقلب احتمالی و نیز هیچ پایگـاه اطـلاع‌رسـانی بـرای گزارش این قبیل موارد وجود نـدارد. نهادهـایی از قبیـل سـازمان بورس اوراق بهادار اطلاعات احتمالی مربوط به هرگونه تحریـف و به طور خاص تقلب در صـورت‌هـای مـالی را در اختیـار عمـوم و تحلیلگران قرار نمی‌دهند. تنها مواردی که در سازمان بورس اوراق بهادار پیگیری می‌شود احتمال تقلب توسـط دارنـدگان اطلاعـات نهانی (به‌ویژه مدیران) در این شرکت‌ها و در صورت صدور رأی از طریق محـاکم قضـایی؛ مـوارد بـه طـور خصوصـی اطـلاع‌رسـانی می‌شود. کشف تقلب با استفاده از روش‌های معمـولی حسابرسـی، کار بسیار مشکلی است. دلیل آن این است که نخست دانش کمی در ارتباط با ویژگی‌هـای تقلـب مـدیریت وجـود دارد. دوم اینکـه بخشی از حسابرسان فاقد تجربه مورد نیـاز در کشـف تحریـف‌هـا به‌ویژه موارد تقلب هستند. نهایتاً اینکه برخی مدیران عمداً سـعی می‌کنند حسابرسان را فریب دهند. برای چنـین مـدیرانی کـه بـه محـدودیت‌هـای حسابرسـی آگـاه هسـتند؛ روش‌هـای مرسـوم حسابرسی ممکن است کافی نباشد. ایـن محـدودیت‌هـا نیـاز بـه روش‌هـای تحلیلـی اضـافی بـرای کشـف مـؤثر تقلـب را تـداعی می‌سازند [۶].

در این تحقیق راهکاری در جهت تشخیص تقلب ارائـه‌شـده است که مبتنی بر درخت تصمیم، ماشین بردار پشتیبان و شـبکه بیزین می‌باشد. در این راهکار پیشنهادی برای اینکه بتوان نویز در داده‌ها را کاهش داد و در کنار آن سربار پردازشـی و حافظـه‌ای را کاهش داد از راهکاری مبتنی بر راهکاری ترکیبی مجموعه راف و تحلیل سلسله‌مراتبی برای انتخاب ویژگی استفاده‌شـده اسـت کـه بدین شکل تنها ویژگی‌های مؤثر مورد استفاده قرار بگیرند.

در ادامه ابتدا مفاهیم لازم جهت درک روش پیشنهادی بیان می‌شود و سپس در بخش بعدی روش پیشنهادی با جزئیات بیان مـی‌شـود. در ادامـه آن ارزیـابی روی روش پیشـنهادی صـورت می‌گیرد و در انتهای مقاله جمع‌بندی کلی و سپس پیشنهادهایی برای کارهای آتی بیان می‌شود.

۲. مفاهیم پایه

در این بخش مفاهیم لازم جهت درک روش پیشنهادی بیان می‌شود.

۲-۱. تقلب

تقلب سوءاستفاده از منافع یک سازمان، بدون آن کـه الزامـاً به‌صورت مستقیم منجر (منتها) به پیامدها و عواقب قانونی شود را تقلب گویند [۷].

شکل (۱): ساختار شبکه نمودار سلسله مراتبی از ارتکاب جرائم یقه سفید از هر ۲ دیدگاه: سطح - شرکت و سطح - جامعه [۷]

محققین، همان گونه که در شکل (۱) قابل‌مشاهده است، به‌منظور بررسی موشکافانه چالش مورد بررسی و ارائه راهکاری مؤثر به‌منظور کشف تقلب، فرد یا افراد متقلب را به‌طورکلی به ۲ گروه "داخلی (شغلی) " و "خارجی" تقسیم می‌کنند و تعریف دقیق‌تری برای هریک از گروه‌های تقلب ارائه می‌دهند.

- تقلب شغلی یا داخلی: به‌صورت ضمنی سوءاستفاده با استفاده نادرست از منابع و دارایی‌های یک سازمان، به‌منظور منافع شخصی که از طریق یکی از عناوین شغلی صورت می‌گیرد را "تقلب شغلی یا تقلب داخلی" گویند. در این گروه از تقلب‌ها سود و انگیزه فرد متقلب متأثر از تعاملات کسب‌وکار است.
- تقلب خارجی: به تقلبی گفته می‌شود که فردی یا افرادی خارج از سازمان، مرتکب آن می‌شود. این تخلفات شامل جرائم رایانه‌ای، نفوذ به شبکه، کارت‌های اعتباری، بیمه، مخابرات، تقلب در نرم‌افزار و غیره می‌شود [۷][۱۰] [۹] [۸]. این بخش یا بخش‌های خارجی متقلب می‌تواند در قالب مشتری (مصرف‌کننده) فعلی یا آتی (آینده) و یا در قالب عرضه‌کننده ارائه‌دهنده فعلی یا آتی (آینده) مرتکب تقلب شود. متقلبین خارجی به ۳ گروه اصلی تقسیم می‌شوند:
  - مجرم حد میانی (میانگین)
  - مجرم جنایی
  - مجرم جنایی سازمان‌یافته

مجرم حد میانی (میانگین)، بیانگر رفتاری گروهی از متقلبین است که فرد متقلب به‌صورت تصادفی یا گاه‌به‌گاهی که فرصت و شرایط مناسب تخلف است یا وسوسه ناگهانی مرتکب آن می‌شود و نیاز مالی و مشکلات اقتصادی افراد است که موجب بروز آن می‌شود [۷].

مجرم جنایی و مجرم جنایی سازمان‌یافته در مقابل گروه مجرمین حد میانی، متقلبین خطرناک‌تری هم وجود دارند. مجرمین جنایی فردی سازمان‌یافته گروهی این گروه خطرناک را تشکیل می‌دهند. آنان کلاهبرداران حرفه‌ای هستند به این سبب که آن‌ها بارهاوبارها هویت واقعی خود را تغییر داده و شیوهٔ عملکرد و روال کاری خود را در طی زمان ارتقا و بهبود می‌بخشند تا در قالبی در ظاهر قانونی قرار بگیرند و توسط روش‌ها و ابزار مقابله با تقلب، کشف و شناسایی نشوند [۷].

رضایی و ریلی در کتاب خود تقلب را به دو گروه تقلب مدیریت و تقلب کارکنان تقسیم نموده و در ذیل آن طبقه¬بندی بیشتری از این دو نوع تقلب را ارائه می‌کنند که در شکل (۲) نشان‌داده‌شده است. تقلب می‌تواند به چندین نوع تقسیم شود که معمول‌ترین آن مصادره دارایی‌ها و اشتباهات مالی است. مصادره دارایی¬ها اغلب مربوط به تقلب کارکنان شامل اختلاس، سرقت وجه نقد یا موجودی و تقلب حقوق و دستمزد، می‌شود؛ اشتباهات مالی به‌عنوان تقلب صورت‌های مالی در نظر گرفته می¬شود که اغلب مدیریت مسئول آن است. وزارت دادگستری ایالات متحده امریکا تقلب شرکت را در سه حوزه گسترده تعریف می‌کند: تقلب حسابداری یا تقلب مالی، تخطی کارکنان، و رفتار انحرافی. تقلب حسابداری شامل تحریف اطلاعات مالی از طریق حساب¬سازی کردن یا گمراه‌کردن سرمایه‌گذاران می‌شود. رایج‌ترین طرح‌های حسابداری شامل فروش موجودی¬ها، معاملات جانبی، معاملات مبادله‌ای، هزینه‌های سرمایه‌گذاری، کسب سریع درآمد و هزینه‌های معوق است [۱۱].

شکل (۲): انواع تقلب [۱۱]

۲-۲. درخت تصمیم

ساختار درخت تصمیم[1] در یادگیری ماشین، یک مدل پیش‌بینی‌کننده می‌باشد که حقایق مشاهده‌شده در مورد یک پدیده را به استنتاج‌هایی در مورد مقدار هدف آن پدیده نقش می‌کند. تکنیک یادگیری ماشین برای استنتاج یک درخت تصمیم از داده‌ها، یادگیری درخت تصمیم نامیده می‌شود که یکی از رایج‌ترین روش‌های داده‌کاوی است [۱۰].

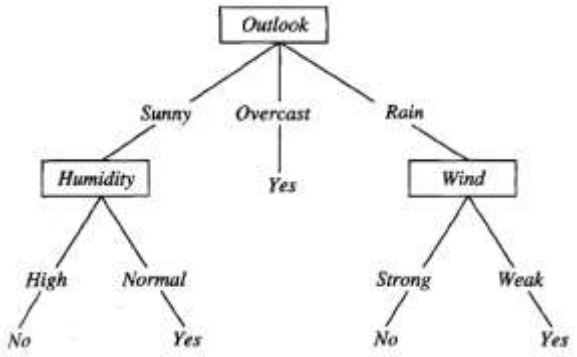
[1] Decision Tree

هر گره داخلی متناظر یک متغیر و هر کمان به یک فرزند، نمایانگر یک مقدار ممکن برای آن متغیر است. یک گره برگ، با داشتن مقادیر متغیرها که با مسیری از ریشه درخت تا آن گره برگ بازنمایی می‌شود، مقدار پیش‌بینی‌شده متغیر هدف را نشان می‌دهد. یک درخت تصمیم ساختاری را نشان می‌دهد که‌برگ‌ها نشان‌دهنده دسته‌بندی و شاخه‌ها ترکیبات فصلی صفاتی که منتج به این دسته‌بندی‌ها را بازنمایی می‌کنند [۱۲].

درختان تصمیم، نمونه‌ها را با مرتب‌کردن آن¬ها در درخت از گره ریشه به سمت گره‌های برگ دسته‌بندی می‌کنند. هر گره داخلی در درخت، صفتی از نمونه را آزمایش می‌کند و هر شاخه‌ای که از آن گره خارج می‌شود متناظر یک مقدار ممکن برای آن صفت است. همچنین به هر گره برگ، یک دسته‌بندی منتسب می‌شود. هر نمونه، با شروع از گره ریشه درخت و آزمایش صفت مشخص‌شده توسط این گره و حرکت در شاخه متناظر با مقدار صفت دیده‌شده در نمونه، دسته‌بندی می‌شود. این فرایند برای هر زیر درختی که گره جدید ریشه آن است، تکرار می‌شود.

در حالت کلی، درختان تصمیم یک ترکیب فصلی از ترکیبات عطفی قیود روی مقادیر صفات نمونه‌ها را بازنمایی می‌کنند. هر مسیر از ریشه درخت به یک، برگ متناظر با یک ترکیب عطفی صفات تست موجود در آن مسیر بوده و خود درخت نیز متناظر با ترکیب فصلی همه این ترکیبات عطفی است. برای مثال درخت تصمیم شکل (۳) متناظر با عبارت زیر می‌باشد [۸].

((Outlook=Sunny) ∧ (Humidity=Normal)) ∨ ((Outlook=Overcast)∨ (Outlook=Rain∧Wind=Weak))



شکل (۳): انواع نمونه‌ای از درخت تصمیم[۸]

۲-۳. شبکه بیزین

شبکه‌های بیزین[1] که با نام شبکه‌های اعتقاد (باور) هم شناخته می‌شوند، متعلق به خانواده مدل‌های گرافیکی احتمالاتی هستند. این ساختارهای گرافیکی برای نشان‌دادن اطلاعات در یک حوزه دارای عدم‌قطعیت به کار می‌روند. به طور خاص هر گره در گراف نشان‌دهنده یک متغیر تصادفی است و شاخه‌ها (کمان) وابستگی‌های احتمالاتی بین متغیرها را نشان می‌دهند. این وابستگی‌های شرطی غالباً به‌وسیله روش‌های آماری و احتمالاتی مشخص ارزیابی می‌شوند. شبکه‌های بیزین اصولی از نظریه گراف، نظریه احتمالات، علوم کامپیوتر و آمار را باهم ترکیب می‌کنند.

به‌طورکلی مدل‌های گرافیکی با شاخه‌های بدون جهت، میدان‌های تصادفی مارکوف یا شبکه‌های مارکوف نامیده می‌شوند. این شبکه‌ها یک تعریف ساده برای استقلال بین متغیرها بر مبنای مفهوم لایه مارکوف فراهم می‌کنند. شبکه‌های مارکوف در زمینه‌هایی نظیر فیزیک آماری و بینایی کامپیوتر بسیار مشهور هستند.

شبکه‌های بیزین متعلق به ساختار دیگری از مدل‌های گرافیکی به نام گراف‌های غیرمقدور جهت‌دار هستند که در زمینه‌های آماری، یادگیری ماشینی و هوش مصنوعی بسیار مشهور هستند. شبکه‌های بیزین نمایش و محاسبات مؤثری از توزیع احتمالاتی مشترک به روی یک سری متغیر تصادفی را فراهم می‌آورند. به‌علاوه شبکه‌های بیزین شدت ارتباط بین متغیرها را به‌صورت کمی مدل می‌کنند که اجازه می‌دهند با دسترسی به اطلاعات جدید، اعتقاد شرطی در مورد آن‌ها به‌صورت خودکار به‌روزرسانی شود.

۲-۴. ماشین بردار پشتیبان

ماشین بردار پشتیبانی[2] یکی از روش‌های یادگیری با نظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند.

این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی نشان داده است. مبنای کاری دسته‌بندی‌کننده SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد. حل معادله پیداکردن خط بهینه برای داده‌ها به‌وسیله روش‌های QP که روش‌های شناخته شده‌ای در حل مسائل محدودیت دار هستند صورت می‌گیرد. قبل از تقسیمِ خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به‌وسیله تابعِ phi به فضای با ابعاد خیلی بالاتر می‌بریم. برای اینکه بتوانیم مسئله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کنیم از قضیه دوگانی لاگرانژ برای تبدیلِ مسئله مینیمم‌سازی موردنظر به فرم دوگانی آن که در آن به‌جای تابع پیچیده phi که ما را به فضایی با ابعاد بالا می‌برد،

[1] Baysian Network

[2] Support Vector Machine(SVM)

تابعِ ساده‌تری به نامِ تابع هسته که ضرب برداری تابع phi است ظاهر می‌شود استفاده می‌کنیم. از توابع هسته مختلفی از جمله هسته‌های نمایی، چندجمله‌ای و سیگمید می‌توان استفاده نمود [۱۴]، [۱۳].

الگوریتم SVM، جز الگوریتم‌های تشخیص الگو دسته‌بندی می‌شود. از الگوریتم SVM، در هر جایی که نیاز به تشخیص الگو یا دسته‌بندی اشیا در کلاس‌های خاص باشد می‌توان استفاده کرد.

۳. مروری بر کارهای گذشته

کلادون و رموندی (۲۰۲۰) فرصت‌های کاربردی تکنیک‌های تحلیل شبکه را برای جلوگیری از پول‌شویی بررسی کردند. در این مقاله با تجزیه‌وتحلیل پایگاه‌داده مرکزی شرکت فاکتور که عمدتاً در ایتالیا کار می‌کند، طی یک دوره ۱۹ماهه، داده‌های واقعی جهان جمع‌آوری‌شده است. این پایگاه‌داده شامل عملیات مالی مرتبط با تجارت فاکتور، همراه با سایر اطلاعات مفید در مورد مشتریان شرکت است. در این مقاله یک رویکرد جدید برای مرتب‌کردن و ارزیابی داده‌های ارتباطی ارائه‌شده است و مدل‌های پیش‌بینی را بر اساس معیارهای شبکه ارائه می‌دهد تا پروفایل‌های ریسک مشتریان درگیر در تجارت فاکتور را ارزیابی کند. می‌توان مشاهده نمود که با استفاده از معیارهای شبکه اجتماعی می‌توان پروفایل‌های خطر را پیش‌بینی کرد [۱۵].

چی، چو و چن در سال ۲۰۱۹ راهکاری را مبتنی بر درخت تصمیم C5.0 ماشین بردار پشتیبان و الگوریتم Chaid ارائه نمودند. در این تحقیق اطلاعات تحقیقاتی مربوط به سال¬های ۲۰۰۷ تا ۲۰۱۶ از مجله اقتصادی تایوان (TEJ) می‌باشد. نمونه شامل ۲۸ شرکت درگیر در کلاهبرداری در صورت¬های مالی و ۸۴ شرکت درگیر در چنین کلاهبرداری‌ها (با نسبت ۱ به ۳) است که در بورس کالا تایوان و بورس تایپه در طول دوره تحقیق ذکرشده است. این مقاله قبل از ایجاد مدل با CHAID و SVM متغیر¬های کلیدی را با SVM و C5.0 انتخاب می‌کند. هر دو متغیر مالی و غیرمالی برای افزایش دقت مدل تشخیص داده می - شوند و برای کلاهبرداری‌های گزارشگری مالی استفاده می‌شوند [۱۶].

وانگ و همکارانش (۲۰۱۹)، در مطالعه‌ای با عنوان " نقش حسابرسان در جلوگیری، کشف و گزارشگری تقلب در کشور امارات"، به شناسایی فرآیندهایی که حسابرسان داخلی و خارجی جهت شناسایی تقلب در طول حسابرسی دنبال می‌کنند، می‌پردازند. نمونه‌ی این تحقیق تعداد ۵۳ حسابرس در امارات متحده عربی می‌باشند. نتایج نشان می‌دهد که مسئولیت شناسایی وقایع متقلبانه با حسابرسان داخلی می‌باشد و فرآیندهای دنبال شده توسط حسابرسان خارجی تا حدودی سخت‌گیرانه‌تر نسبت به حسابرسان داخلی می‌باشد. به‌طورکلی نتایج این تحقیق نشان‌دهنده این امر است که مسئولیت شناسایی تقلب و گزارش آن با حسابرسان داخلی بوده و همچنین حسابرسان خارجی نیز باید جست‌وجوی خود را جهت شناسایی و کشف تقلب در گزارش‌های مالی افزایش دهند [۱۷].

تجزیه‌وتحلیل پول‌شویی یک کار بسیار پیچیده است که نیاز به پردازش مقادیر زیادی داده‌ها از منابع مختلف از قبیل صورت‌حساب‌ها و معاملات حساب بانکی برای دستیابی به دانش مفید برای یک محقق است. به‌منظور پشتیبانی از توانایی‌های تحلیلی انسان، ابزارهای نرم‌افزاری اختصاصی موردنیاز است [۱۸].

کیم و دیگران (۲۰۲۰)، در پژوهش خود تحت عنوان " کشف اظهارات اشتباه مالی با مقاصد متقلبانه با استفاده از آموزش چند سطحی حساس به هزینه" با استفاده از رگرسیون لجستیک چندجمله‌ای، ماشین بردار حمایتی و شبکه‌های بیزین، به‌عنوان ابزار پیش‌بینی جهت کشف و طبقه‌بندی اظهارات اشتباه بر اساس وجود مقاصد متقلبانه، به بسط سه سطحی طبقه‌بندی پرداختند. آن‌ها به ارزیابی جنبه‌هایی از پژوهش‌های قبلی در کشف مقاصد متقلبانه و ملزومات اظهارات اشتباه پرداختند. جنبه‌هایی مانند نسبت سود کوتاه‌مدت و مقیاس کارایی بنگاه، نشان‌دهنده پتانسیل بالقوه‌ی تبعیض‌آمیز می‌باشد [۱۹].

لاری دشتبان و همکارانش (۲۰۱۹)، در تحقیقی تحت عنوان " فرایند جستجو و کشف اطلاعات برای تقلب در صورت‌های مالی " به یک مرور کلی از فرایندهای داده‌کاوی مورداستفاده برای شناسایی تقلب مالی، به‌ویژه تقلب در صورت‌های مالی شرکت‌ها، پرداختند. آن‌ها در نتایج پژوهش خود اظهار داشتند که مهم‌ترین روش‌های مورداستفاده برای شناسایی تقلب مالی شامل رگرسیون لجستیک، شبکه‌های عصبی، شبکه‌های استنتاج بیزی و درخت تصمیم‌گیری است که راه‌حل‌های مهمی برای مشکلات ذاتی شناسایی و طبقه‌بندی داده‌ها هستند [۲۰].

محمد یوسف و همکارانش (۲۰۱۵)، در مطالعه‌ای تحت عنوان "کاربرد مدل‌های تقلب در شرکت‌های بورسی مالزی" به بررسی احتمال تقلب صورت‌های مالی در شرکت‌های بورسی مالزی با استفاده از مدل مثلث تقلب، مدل لوزی تقلب و مدل

پنج‌ضلعی تقلب پرداختند. در این تحقیق به‌وسیله فاکتورهای ریسک تقلب به‌دست‌آمده از این مدل‌های تقلب، چشم‌انداز جدیدی در کشف احتمال تقلب در صورت‌های مالی شرکت‌های مالزی به دست آمد. همچنین نتایج این تحقیق معیارهای قابل‌اندازه‌گیری و فاکتورهای ریسک تقلب جدیدی مانند طمع و جهل را معرفی می‌کند [۲۱].

راوندا و همکارانش (۲۰۱۸) پروکسی¬های جدید مدیریت معاملات را برای کشف شواهد تجربی از مدیریت استراتژیک معاملات حسابداری، با هدف انجام عملیات پول‌شویی، در نمونه‌ای از ۳۵۵ شرکت تحت کنترل مافیای ایتالیا، توسعه داده شده است. در این مقاله نشان‌داده‌شده است که با استفاده از تجزیه‌وتحلیل خوشه‌ای، شرکت‌های تحت کنترل مافیا را می‌توان به دو گروه مختلف متصل به شرکت‌های واقعی و شرکت‌های پوسته تقسیم‌بندی کرد، بر اساس فرضیه‌های خاص بر ویژگی‌های متمایز آن‌ها. مهم‌تر از همه، برآورد رگرسیون شواهدی از شیوه‌های مختلف TRM این شرکت‌ها را ارائه می‌دهد که ممکن است با فعالیت‌های خاص پول‌شویی ارتباط داشته باشد. این مطالعه پیشنهاد جدید پروکسی¬های TRM را بر اساس ماهیت معامله هزینه انجام می‌دهد که می‌تواند توسط مقامات به‌عنوان پرچم‌های قرمز برای فعالیت‌های پول‌شویی مورداستفاده قرار گیرد. علاوه بر این، این مطالعه ممکن است از استدلال‌های مهم در برابر دیدگاه ارتدوکس از نقش پول نقد حسابداری و مناسب‌بودن پروکسی‌های سنتی TRM برای نشان‌دادن شیوه‌های درون شرکت‌هایی که ویژگی‌های مشترکی با شرکت‌های تحت کنترل مافیا دارند، پشتیبانی کند [۲۲].

۴. روش پیشنهادی

در این تحقیق برای جمع‌آوری داده‌ها از روش کتابخانه‌ای و میدانی استفاده‌شده است. اطلاعات مربوط به مبانی نظری و تئوریک پژوهش از کتب و مقالات فارسی و لاتین جمع‌آوری گردیده است. منبع جمع‌آوری داده‌های مورد نیاز این پژوهش، صورت‌های مالی، گزارش‌های حسابرس مستقل و بازرس قانونی شرکت‌ها بوده است. داده‌های مالی مورد نیاز این پژوهش از طریق بانک اطلاعات رایانه‌ای ره‌آورد نوین و مراجعه به سامانه جامع اطلاع‌رسانی ناشران http://www.codal.ir/ و http://www.rdis.ir جمع‌آوری‌شده است.

داده‌های جمع‌آوری‌شده پس از طبقه‌بندی لازم بر اساس متغیرهای مورد بررسی، با استفاده از نرم‌افزار اکسل وارد رایانه شده است.

از جمله مهم‌ترین مواردی که لازم است هنگام جمع‌آوری داده‌ها به آن توجه کرد، روایی ابزارهای گردآوری داده‌ها است. منظور از روایی ابزارهای گردآوری داده‌ها این است که ابزارها بتوانند واقعیت‌ها را به‌خوبی نشان دهند.

در اینجا، قصد رسیدن به روشی می‌باشد که به‌وسیله آن بتوان تا با درصد بالایی تقلب را در صورت‌های مالی تشخیص داد. با استفاده از روشی که در اینجا بیان می‌شود قصد داریم تا راهکاری ارائه کنیم که مبتنی بر درخت تصمیم، SVM و الگوریتم بیزین است. در اینجا هدف اصلی ارائه راهکاری می‌باشد که به‌وسیله آن بتوان تصمیمات مهمی را در زمینه تشخیص تقلب در صورت‌های مالی و همچنین جلوگیری از آن‌ها ارائه نمود. در این تحقیق برای اینکه بتوان به‌دقت بالاتری در روش پیشنهادی دست‌یافت باید ابتدا با استفاده از داده‌کاوی سیستم مورد آموزش قرار گیرد.

روشی که در این تحقیق استفاده‌شده است یک روش ترکیبی می‌باشد که بتوان به این شکل دقت روش را بسیار افزایش داد. در اینجا از الگوریتم بیزین، SVM و درخت تصمیم ID3 استفاده‌شده است. با استفاده از این راهکار پیشنهادی برای هر یک از این الگوریتم‌ها وزنی در نظر گرفته می‌شود و برای محاسبه و پیش‌بینی مقدار محاسبه‌شده توسط هر یک از این الگوریتم‌ها در وزن آن الگوریتم ضرب می‌شود و در انتها نتیجه واقعی به دست می‌آید که دارای دقت بالاتری خواهد بود؛ زیرا از فواید هر دو این الگوریتم‌ها استفاده نموده است.

روش پیشنهادی دارای مراحل مختلفی می‌باشد که به ترتیب عبارت‌اند از:

۱- پیش‌پردازش داده‌ها

۲- انتقال داده‌ها

۳- انتخاب ویژگی‌های مؤثر با استفاده از مجموعه راف و AHP

۴- ساخت درخت تصمیم‌گیری

۵- آموزش و محاسبه اوزان الگوریتم‌های درخت تصمیم، ماشین بردار پشتیبان و الگوریتم بیزین

فلوچارت کلی روش پیشنهادی را می‌توان در شکل (۴) مشاهده نمود.

شکل (۴): فلوچارت کلی روش پیشنهادی

۴-۱. عملیات پیش‌پردازش داده‌ها

در ابتدا مجموعه داده جمع‌آوری‌شده و به آماده‌سازی و پیش‌پردازش داده‌ها پرداخته می‌شود. در آماده‌سازی و پیش‌پردازش داده‌ها از روش‌های مختلفی استفاده می‌شود. اول این که برخی ویژگی‌ها دارای مقادیر منحصربه‌فرد هستند. این ویژگی‌ها نمی‌توانند دانش مفیدی را در مجموعه داده ایجاد کنند. لذا این مجموعه ویژگی‌ها باید از داده‌ها حذف شوند. به طور نمونه می‌توان به ویژگی نام و نام خانوادگی اشاره نمود. همچنین ممکن است برخی تراکنش‌ها دارای مقادیر مفقود فراوان باشند. لذا این تراکنش‌ها نیز باید از مجموعه داده‌ها حذف شوند. از طرفی ممکن است، مقادیر برخی ویژگی‌ها دارای مقادیر نویز و مفقود باشند؛ لذا این مقادیر نیز باید در مجموعه داده اصلاح شوند. مرحله بعدی نوبت به استفاده از ابزار کشف آنومالی پرداخته می‌شود. داده‌هایی که در نقاط خارج از قانون مجموعه داده قرار دارند شناسایی‌شده و حذف می‌شوند. برای اینکه بتوان روی داده‌ها به‌عنوان ورودی کارکرد باید ویژگی‌هایی را از آن‌ها استخراج نمود. به طور معمول پیش از انتخاب و استخراج ویژگی‌ها، برخی عملیات پیش‌پردازش بر روی داده‌ها انجام می‌شود.

۴-۲. انتقال داده‌ها

در این قسمت داده‌ها در دامنه‌های درست قرار می‌گیرند. بدین معنا که داده‌ها باید به رنج‌هایی که در سیستم مشخص‌شده است منتقل شوند و داده‌های خارج از رنج، داده‌های مشکل‌دار بوده و باید حذف شوند. داده‌ها باید در رنج درست قرار بگیرند بدین معنا که برای مثال اگر فیلد سن وجود داشته باشد فردی که محدوده سنی بین ۵۵ تا ۷۰ دارد باید در سیستم به‌صورت خیلی پیر شود که این قسمت به‌صورت خودکار از روی دیتاست تکمیل می‌شود

۴-۳. انتخاب ویژگی‌های مؤثر با استفاده از مجموعه راف

بسیاری از مفاهیم و تئوری‌های عدم‌قطعیت نظیر مجموعه‌های فازی، سیستم‌های خاکستری و مجموعه‌های راف، درگذشته معرفی‌شده و در سال‌های اخیر ابزارهای ریاضی مبتنی بر آن‌ها با سرعت بالایی توسعه‌یافته‌اند. هریک از این رویکردها، مفاهیم خاص خود را داشته و دارای ویژگی‌های منحصر به خود می‌باشد. به‌عنوان مثال، تئوری کلاسیک به دنبال تحلیل داده‌های احتمالی یا قطعی بوده و تئوری فازی، محاسبات نرم را اساس کار خود قرار داده است. تئوری خاکستری به کنترل سیستم‌ها در شرایط کمبود داده‌ها و اطلاعات ناکامل پرداخته و تئوری راف، تقریب و استدلال درباره داده‌ها را به دنبال دارد. داده‌هایی که از دنیای واقعی اخذ می‌گردند معمولاً شامل تمامی انواع نویزها بوده و عدم‌قطعیت بسیار و اطلاعات غیرکامل فراوانی به همراه دارند. روش‌های سنتی برخورد با این عدم‌قطعیت نظیر تئوری فازی، تئوری گواه، تئوری احتمالات و نظایر آن، به اطلاعات اضافی

مانند توزیع احتمال و تابع عضویت نیازمند هستند. به بیان دیگر، کار با این سیستم‌ها به دلیل حجم بالایی از داده‌ها مشکل است، ازاین‌رو به‌کارگیری سایر تئوری‌ها نظیر تئوری مجموعه‌های راف می‌تواند در این راه کمک‌کننده باشد.

مجموعه راف ابزاری قابل‌استفاده از شرایط ابهام و عدم‌قطعیت است که اولین‌بار توسط پاولاک (۱۹۸۲) ارائه شد. این تئوری راف در زمینه‌های مختلفی مورد استفاده قرار می‌گیرد از جمله تجزیه‌وتحلیل تصمیم‌گیری، سیستم‌های پشتیبان تصمیم. بعد از آقای پاولاک، سه محقق دیگر به نام‌های ژای، خو و ژانگ در سال ۲۰۰۸ اعداد راف را ارائه کردند. یک عدد راف دارای حد پایین (L)، حد بالا (U) و حد میانی که به فاصله مرزی راف مشهور است تشکیل شده است. اعداد راف در مسائلی استفاده می‌شود که نظرات خبرگان در آن دخیل هستند و به‌نوعی باعث ایجاد عدم‌قطعیت و ابهام بشود.

فرض کنید که در یک مجموعه تصمیم‌گیری مجموعه U شامل تمام اعضای مجموع باشد. Y یک عضو دلخواه از مجموعه U و R یک مجموعه از t کلاس است که تمام اعضای U را پوشش می‌دهد اگر این کلاس‌ها به‌صورت ترتیبی همانند $G1<G2<\ldots<Gt$ باشند آنگاه حدهای پایین، بالا و ناحیه مرزی از کلاس G به‌صورت رابطه (۱) تعریف می‌شود.

$$\underline{Apr}(G_q) = \bigcup[Y \in U | R(Y) \le G_q] \quad (1)$$

$$\overline{Apr}(G_q) = \bigcup[Y \in U | R(Y) \ge G_q]$$

$$Bnd(G_q) = \bigcup[Y \in U | R(Y) \ne G_q] = [Y \in U | R(Y) > G_q] \cup [Y \in U | R(Y) < G_q]$$

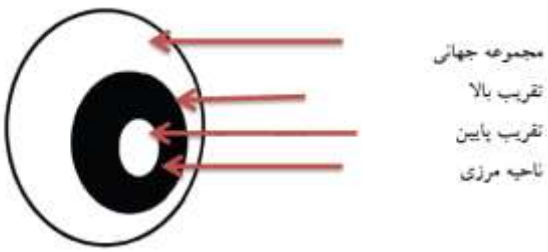
سپس این کلاس G می‌تواند به‌صورت یک عدد راف در حدهای پایین و بالا به‌صورت رابطه (۲) ارائه شود

$$\underline{Lim}(G_q) = \frac{1}{M_L}\sum R(Y) | Y \in \underline{Apr}(G_q) \quad (2)$$

$$\overline{Lim}(G_q) = \frac{1}{M_U}\sum R(Y) | Y \in \overline{Apr}(G_q)$$

$$RN(G_q) = [\underline{Lim}(G_q), \overline{Lim}(G_q)]$$

همچنین فاصله مرزی راف به‌صورت شکل (۵) محاسبه می‌شود این فاصله مرزی ابهام را بیان می‌کند به‌طوری‌که هر چقدر این عدد بزرگ‌تر باشد نشان‌دهنده ابهام بیشتر است و اگر عدد کوچک‌تر باشد نشان‌دهنده دقت بیشتر است.



شکل (۴): محاسبه محدوده در مجموعه راف

چون اعداد راف مشابه اعداد فاصله‌ای هستند؛ بنابراین قوانین محاسباتی اعداد فاصله‌ای برای اعداد راف نیز یکسان است که در ادامه به آن می‌پردازیم.

- الف) ضرب یک عدد صحیح در یک عدد راف
- ب) جمع دو عدد راف
- ج) ضرب دو عدد راف

این سه عملگر اصلی در روابط (۳) نشان‌داده‌شده است که در حالت الف، آن عدد صحیح در درایه‌های عدد راف ضرب می‌شود. در حالت ب، درایه‌های نظیر دو عدد راف باهم جمع می‌شود و در حالت ج نیز، درایه‌های متناظر دو عدد راف در هم ضرب می‌شوند.

$$RN(\alpha) \times \mu = [\underline{Lim}(\alpha), \overline{Lim}(\alpha)] \times \mu = [\mu \times \underline{Lim}(\alpha), \mu \times \overline{Lim}(\alpha)] \quad (3)$$

$$RN(\alpha) + RN(\beta) = [\underline{Lim}(\alpha), \overline{Lim}(\alpha)] + [\underline{Lim}(\beta), \overline{Lim}(\beta)] = [\underline{Lim}(\alpha) + \underline{Lim}(\beta), \overline{Lim}(\alpha) + \overline{Lim}(\beta)]$$

$$RN(\alpha) \times RN(\beta) = [\underline{Lim}(\alpha), \overline{Lim}(\alpha)] \times [\underline{Lim}(\beta), \overline{Lim}(\beta)] = [\underline{Lim}(\alpha) \times \underline{Lim}(\alpha), \overline{Lim}(\alpha) \times \overline{Lim}(\beta)]$$

در این تحقیق از روش مجموعه راف ادغام‌شده با روش AHP که یک روش تصمیم‌گیری می‌باشد وزن‌دهی پارامترها صورت‌گرفته و پارامترهای مؤثر شناسایی می‌شوند.

روش AHP (فرایند تحلیل سلسله‌مراتبی) از روش‌های پر کاربرد در تصمیم‌گیری چندمعیاره است که هدف آن محاسبه وزن معیارها و گزینه‌های پژوهش تحت یک مدل سلسله‌مراتبی است. در این مدل ابتدا مقایسات زوجی تشکیل شده و در اختیار خبرگان قرار داده می‌شود تا بر اساس طیف ۱ تا ۹ نظرات خود را نسبت به مقایسه دو به دوی معیارها بیان کنند. برای استفاده از اعداد راف در روش AHP (rough AHP) به طریق زیر عمل می‌کنیم.

۱. ابتدا مقایسات زوجی خبره‌ها را از نظر نرخ ناسازگاری بررسی کرده و چنانچه نرخ ناسازگاری کمتر از ۰٫۱ باشد یعنی مقایسه زوجی سازگار است و در صورتیکه بزرگ‌تر از ۰٫۱ باشد باید اعداد مقایسه زوجی اصلاح شود.
۲. ایجاد اعداد راف از اعداد خبره‌ها با استفاده از روابطی که در تئوری گفته شد.
۳. محاسبه وزن فاصله‌ای معیارها با استفاده از روش میانگین هندسی

به‌عنوان مثال فرض کنید ۷ معیار داریم می‌خواهیم با استفاده از تئوری راف وزن این ۷ معیار را محاسبه کنیم. تعداد خبرگان نیز ۵ نفر می‌باشد. ابتدا مقایسات زوجی ۷ معیار در اختیار خبرگان قرار داده می‌شود که به طریق زیر پاسخ داده‌اند و همچنین نرخ ناسازگاری هر ماتریس مقایسه زوجی از ۰٫۱ نیز کمتر است (شکل (۶)).

$$B_1=\begin{bmatrix}1&5&4&2&3&5&9\\1/5&1&1/3&1/5&1/4&1/2&5\\1/4&3&1&1/5&1/3&2&5\\1/2&5&5&1&3&5&9\\1/3&4&3&1/3&1&3&9\\1/5&2&1/2&1/5&1/3&1&5\\1/9&1/5&1/5&1/9&1/9&1/5&1\end{bmatrix},\quad CR_1=0.0585<0.1$$

$$B_3=\begin{bmatrix}1&7&3&1&3&5&9\\1/7&1&1/2&1/7&1/3&1/3&3\\1/3&2&1&1/5&1/2&3&5\\1&7&5&1&2&4&9\\1/3&3&2&1/2&1&3&5\\1/5&3&1/3&1/4&1/3&1&3\\1/9&1/3&1/5&1/9&1/5&1/3&1\end{bmatrix},\quad CR_3=0.0388<0.1$$

$$B_2=\begin{bmatrix}1&5&3&1&1&4&9\\1/5&1&1/2&1/5&1/5&1/3&5\\1/3&2&1&1/5&1/2&3&7\\1&5&5&1&2&4&9\\1&5&2&1/2&1&3&7\\1/4&3&1/3&1/4&1/3&1&3\\1/9&1/5&1/7&1/9&1/7&1/3&1\end{bmatrix},\quad CR_2=0.0506<0.1$$

$$B_4=\begin{bmatrix}1&7&5&2&3&7&9\\1/7&1&1/3&1/6&1/5&1/2&3\\1/5&3&1&1/3&1/3&2&5\\1/2&6&3&1&3&5&7\\1/3&5&3&1/3&1&4&7\\1/7&2&1/2&1/5&1/4&1&4\\1/9&1/3&1/5&1/7&1/7&1/4&1\end{bmatrix},\quad CR_4=0.0458<0.1$$

$$B_5=\begin{bmatrix}1&7&5&1&2&5&7\\1/7&1&1/3&1/5&1/4&1/2&3\\1/5&3&1&1/3&1/3&2&5\\1&5&3&1&2&5&7\\1/2&4&3&1/2&1&3&5\\1/5&2&1/2&1/5&1/3&1&2\\1/7&1/3&1/5&1/7&1/5&1/2&1\end{bmatrix},\quad CR_5=0.0329<0.1$$

شکل (۶): ماتریس نمونه

بعد از پیاده‌سازی گام‌های تئوری راف وزن معیارها به‌صورت زیر محاسبه می‌شود.

$$w=\{w_1,w_2,w_3,w_4,w_5,w_6,w_7\}= \quad (۴)$$
$$\{\lfloor 2.902,3.556\rfloor,\lfloor 0.428,0.506,0.898,1.073\rfloor,$$
$$\lfloor 2.663,3.123\rfloor,$$
$$\lfloor 1.563,1.935\rfloor,\lfloor 0.595,0.703\rfloor,\lfloor 0.216,0.254\rfloor\}$$

۴-۴. ساخت درخت تصمیم

در این قسمت از درخت تصمیم‌گیری استفاده می‌شود. ویژگی‌هایی که برای شناسایی تقلب در صورت¬های مالی در نظر گرفته می‌شوند و به‌عنوان پارامترهای تأثیرگذار شناسایی شدند در این مرحله باید به‌عنوان گره‌های درخت در نظر گرفته شوند؛ ولی اینکه هر یک از این ویژگی‌ها در کدام گره قرار گیرند و یا در چه سطحی از درخت قرار گیرند بسیار مهم می‌باشد. از طرفی چون در این قسمت در راهکار پیشنهادی تفاوتی ندارد که روی چه دیتاستی کار شود و صرفاً یک مدل بهینه برای پیش‌بینی ارائه می‌شود در این قسمت ویژگی‌ها به‌صورت A، B و... در نظر گرفته می‌شود چرا که می‌تواند هر ویژگی ای باشد و برای فهم بهتر کار این قسمت با مثال توضیح داده می‌شود. درخت تصمیم‌گیری، درختی است که هر شاخه از آن به‌عنوان یک انتخاب می‌باشد. بدین معنی برای رفتن از گره ریشه به گره پایین‌تر می‌توان از شاخه‌هایی که به آن گره متصل هستند یکی انتخاب شود. در انتها هر یک از گره‌های انتهایی یا اصطلاحاً گره برگ تصمیمی را بازگو می‌کند. هر کدام از شاخه‌ها تا رسیدن به برگ دارای سناریو ای است که موجب تصمیمی می‌شود. در این روش پیشنهادی از درخت تصمیم ID3 بهبودیافته توسط ما استفاده‌شده است که بهبود ناشی از ما موجب سرعت عمل بالای آن شده است. درخت ID3 یک درخت تصمیم‌گیری است که دارای یادگیری نیز می‌باشد و اولین‌بار توسط راس کوینلن مطرح شد. ایده الگوریتم ID3، ساخت درخت تصمیم‌گیری بالابه‌پایین است که انتخاب گره در آن به‌وسیله جستجوی حریصانه از میان مجموعه‌ای از صفت‌ها می‌باشد. در اینجا ما برای اینکه قادر باشیم تا مفیدترین صفت را از میان صفات بیابیم که در کلاسه‌بندی مفیدتر باشد از الگویی بخصوص استفاده نمودیم. برای اینکه بتوان کلاسه‌بندی مفیدی را برای مجموعه یادگیری انجام داد، باید تعداد سؤالات را کاهش داد یا می‌توان گفت باید عمق درخت تصمیم‌گیری را کاهش داد. ازاین‌رو در این قسمت نیاز به تابعی است که قادر باشد تا متعادل‌ترین تقسیم را انجام دهد که در این صورت عمق درخت بسیار کاهش می‌یابد و گره‌ها به‌صورت متعادل در درخت تقسیم می‌شوند.

جدولی را در نظر بگیرید که دارای صفات و کلاسی از صفات نیز می‌باشد. در صورتی به این جدول همگن گفته می‌شود که تنها شامل یک کلاس باشد. اگر یک جدول دارای چندین کلاس باشد، در این حالت به آن ناهمگن گویند. توابع زیادی همچون آنتروپی، gini index و classification error برای سنجش میزان همگن پذیری وجود دارند. در این میان در اینجا از آنتروپی[1] استفاده‌شده است

$$\text{Entropy}=\sum_j -p_j\log_2 p_j \quad (۵)$$

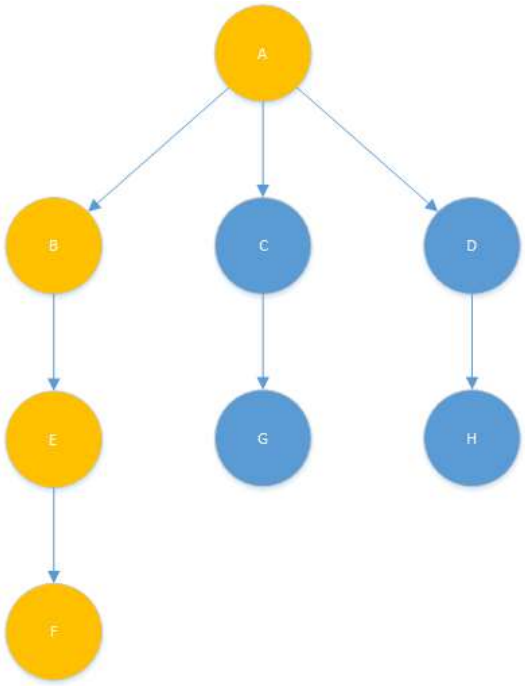
[1] Entropy

آنتروپی یک جدول صفر است؛ زیرا احتمال آن مقداری برابر یک است (تنها دارای یک کلاس باشد). آنتروپی زمانی به بیشترین مقدار خود می‌رسد که تمامی کلاس‌های موجود در جدول دارای احتمالی برابر باشند. آنتروپی را می‌توان به‌نوعی معیاری برای سنجش بی‌نظمی در نظر گرفت هر چه مجموعه منظم‌تر باشد و دارای گوناگونی کمتری باشد آنگاه آنتروپی آن کمتر است و به‌نوعی بی‌نظمی آن نیز کمتر است و برعکس. البته در اینجا چون ما در مرحله قبل یک کلاسه‌بندی ابتدایی را انجام دادیم تقریباً بی‌نظمی نیز پایین می‌باشد و این خود باعث سرعت عمل بالاتر روش پیشنهادی ما می‌شود؛ زیرا این قضیه باعث می‌شود که عمق درخت تصمیم‌گیری کم شود و هر چه عمق این درخت کمتر باشد، سرعت تصمیم‌گیری نیز بیشتر می‌شود.

در این قسمت ما از آنتروپی استفاده کردیم تا مقدار بی‌نظمی را برای هر یک از صفت‌های دیتاست به دست آوریم. برای اینکه بتوانیم صفتی را در درخت تصمیم‌گیری انتخاب کنیم که در رتبه بالاتری از بقیه صفت‌ها باشد به‌نوعی دارای اهمیت بالاتری از بقیه صفت‌ها باشد از رابطه (۶) استفاده کردیم. با توجه به این فرمول آنتروپی همه صفات در مجموعه S محاسبه می‌شود و مقدار صفت مجموعه A از آن کاسته می‌شود. مجموعه A، مجموعه همه صفات انتخاب‌شده از پدر تا به اینجا در یک مسیر خاص می‌باشد.

$$G(S,A) = Entropy(S) - \sum_{v \in Values(A)} Entropy(v) \qquad (۶)$$

برای درک بهتر رابطه (۶) می‌توان شکل (۷) را مشاهده نمود.



شکل (۷): مثالی برای انتخاب صفت‌ها

همان‌طور که در شکل (۷) قابل‌مشاهده است، باید آنتروپی تمامی صفات از آنتروپی صفات انتخاب‌شده تا به اینجای مسیر کاسته شود (یعنی باید G(S,F)=E(S)-(E(A)+E(B)+E(E)+E(F)) را به دست آورد). البته باید توجه شود که ما باید در مجموعه A صفاتی که تا به اینجای کار استفاده‌شده است به‌علاوه صفتی که می‌خواهیم قرار دهیم را محاسبه نماییم. بعدازاین کار از بین این مجموعه صفات باقیمانده که برای هر کدام رابطه (۶) را محاسبه نمودیم، صفتی را که دارای G بیشتری است را انتخاب نماییم. در این حالت اگر دو صفت دارای G برابر بودند که احتمال این پیشامد نیز کم نیست، باید به گره دو یا هر تعداد صفت که دارای بیشترین مقدار G هستند و باهم برابر نیز می‌باشند به گره مربوطه بیفزاییم؛ یعنی اگر برای مثال در گره‌ای دو صفت دارای G برابر بودند، آنگاه این دو به گره مربوطه دو فرزند می‌افزاییم و هر کدام از این صفت‌ها به‌عنوان یک فرزند این گره در نظر گرفته می‌شوند و بعدازاین کار روند الگوریتم را برای هر یک از این گره‌ها ادامه می‌دهیم. برای مثال در شکل (۷) می‌توان مشاهده نمود که مقدار G برای صفت‌های B، C و D برابر است؛ ازاین‌رو همه این صفت‌ها در یک سطح قرار گرفته‌اند.

این کار باعث می‌شود تا صفت‌هایی که دارای آنتروپی بیشتری هستند را بیابیم؛ چراکه این صفت‌ها تأثیر بیشتری را در تصمیم نهایی ما می‌گذارند. این روند جلو رفتن در درخت تصمیم‌گیری تا جایی ادامه می‌یابد که در هر مسیر دیگر صفتی باقی نمانده باشد. در این حالت درخت تصمیم‌گیری کاملاً ساخته‌شده است و به پایان رسیده است.

۴-۵. ماشین بردار پشتیبان در روش پیشنهادی

در این تحقیق سعی داریم تا از این الگوریتم استفاده نماییم تا پیوستگی بین ویژگی‌ها حفظ شود و در نتیجه با توجه به پیوستگی بین نسبت‌های مالی بتوان پیش‌بینی‌های دقیق‌تری را انجام داد. با توجه به اینکه در این تحقیق نسبت‌های مالی و متخلف بودن و یا نبودن مشخص است؛ بنابراین از نوع نظارت شده می‌باشد. در این تحقیق از الگوریتم SVM و شبکه بیزین استفاده‌شده است چرا که این روش‌ها دارای قدرت تحلیل بسیار بالایی می‌باشند و همچنین با توجه به‌دقت این دو روش از این دو روش در کنار روش پیشنهادی کمک گرفته‌شده است تا اینکه بتوان راهکاری ارائه داد که دارای دقت بسیار بالاتری از حالت عادی باشد و بتوان به نتایج مطلوب‌تری دست‌یافت.

۴-۶. شبکه بیزین در روش پیشنهادی

شبکه بیزین با استفاده از قدرتی که دریافت احتمالات و پیوستگی بین ویژگی‌ها دارد برای این تحقیق انتخاب‌شده است.

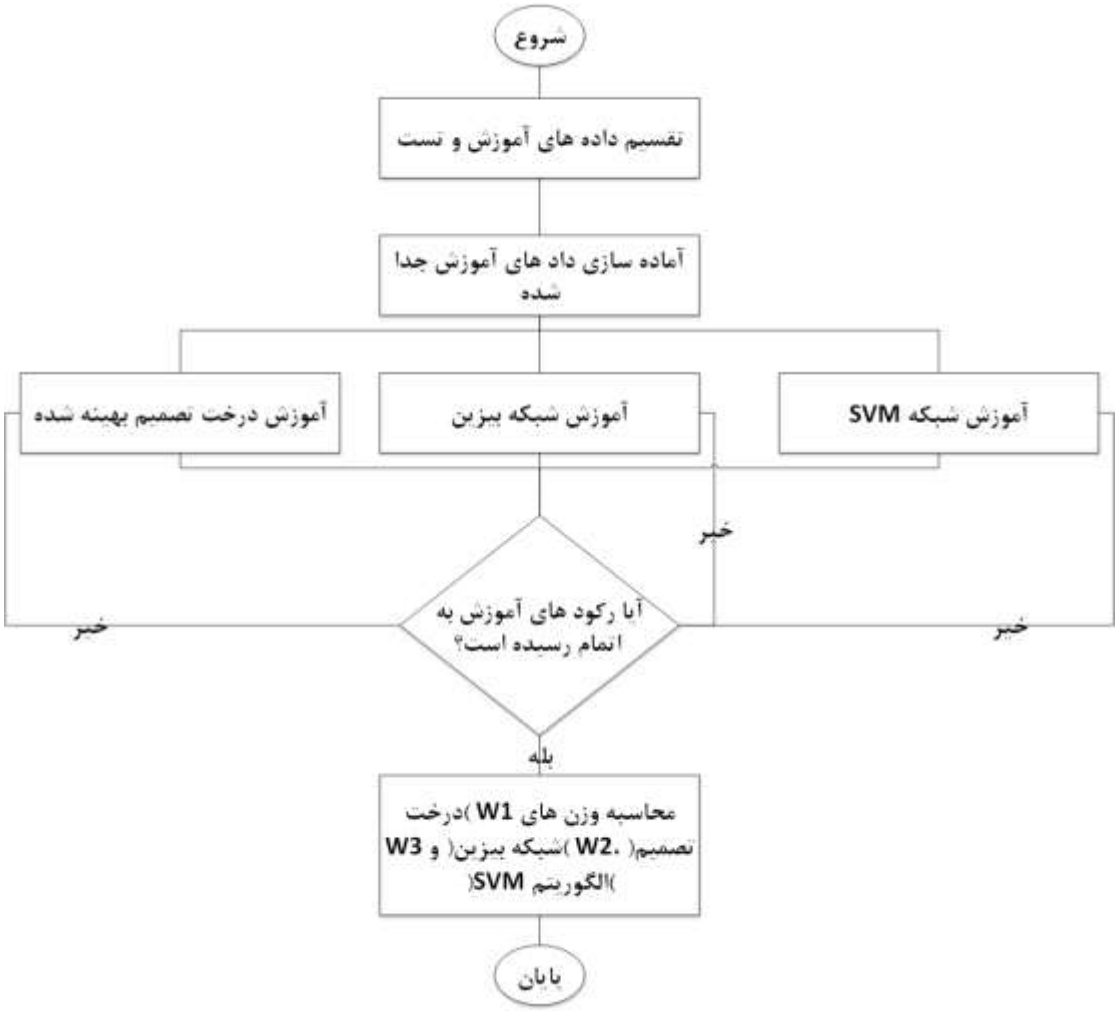
این الگوریتم در کنار دو روش دیگر یعنی درخت تصمیم و ماشین بردار پشتیبان در این روش پیشنهادی استفاده‌شده است تا اینکه بتواند باقدرت بالاتری پیش‌بینی‌ها را انجام دهد.

در این مرحله در ابتدای کار نرمال‌سازی انجام می‌شود و سپس نتایج استخراج‌شده از نرمال‌سازی، قسمتی از این داده‌ها به‌عنوان داده‌های آموزش استفاده‌شده و مدل شبکه بیزین ایجاد می‌شود و سپس با استفاده از داده‌های آزمون وزن این الگوریتم محاسبه می‌شود تا اینکه بتوان در ادامه میزان تأثیرگذاری این قسمت از الگوریتم را محاسبه نمود.

۴-۷. محاسبه اوزان ماشین بردار پشتیبان، شبکه بیزین و درخت تصمیم

برای اینکه بتوان میزان تأثیرگذاری هر یک از این سه الگوریتم را در خروجی و یا پیش‌بینی نهایی در نظر گرفت باید برای هر یک از الگوریتم‌های ماشین بردار پشتیبان، شبکه بیزین و درخت تصمیم وزنی را در نظر گرفت و متناسب با این وزن‌های در نظر گرفته‌شده میزان تأثیرگذاری در پیش‌بینی جواب نهایی را محاسبه نمود.

در ابتدای کار درصدی از مجموعه دیتاست مورد استفاده برای محاسبه آموزش و محاسبه وزن مورد استفاده قرار می‌گیرد. در این قسمت قصد استفاده از یک الگوریتم ترکیبی می‌باشد که از دو الگوریتم درخت تصمیم و الگوریتم شبکه بیزین استفاده می‌کند. این دو الگوریتم هر کدام سهمی از جواب نهایی را خواهند داشت که بدین شکل دقت سیستم افزایش می‌یابد. می‌توان نمایی از این مرحله را در شکل (۸) مشاهده نمود.



شکل (۸): فلوچارت روش استفاده‌شده در محاسبه وزن

همان‌طور که در شکل (۸) مشاهده می‌شود الگوریتم پیشنهادی با استفاده از مجموعه داده‌ای در ابتدا به محاسبه وزن می‌پردازد و این محاسبه بدین شکل می‌باشد که هر الگوریتم با استفاده از ۷۰ درصد از مجموعه کل دیتاست آموزش می‌بینند. در نهایت با توجه به تعداد جواب‌های صحیح امتیاز و یا وزنی به آن تعلق می‌گیرد تا اینکه با استفاده از وزن تعلق‌گرفته بتوان در مرحله بعدی وزنی را برای خروجی هر الگوریتم در نظر گرفت همان‌طور که در معماری نیز می‌توان مشاهده نمود بعد از

محاسبه وزن که به‌صورت تقسیم تعداد جواب‌های درست به تعداد کل جواب‌های حدس زده شده می‌باشد، می‌توان میزان تأثیرگذاری هر کدام از این الگوریتم را در خروجی نهایی بهتر تشخیص داد. در این روش بعد از محاسبه وزن‌ها، به ازای هر رکورد، پیش‌بینی‌ای توسط الگوریتم شبکه بیزین و توسط درخت تصمیم ID3 صورت می‌گیرد که مقدار پیش‌بینی‌شده باید در وزن آن الگوریتم ضرب شود و خروجی نهایی پیش‌بینی الگوریتم برابر است با جمع نتایج هر یک از الگوریتم ضرب در وزن آن الگوریتم که بدین صورت نتیجه نهایی به دست می‌آید و دسته‌بندی درست صورت می‌گیرد.

۵. ارزیابی روش پیشنهادی

در این بخش روش پیشنهادی که در بخش قبل بیان شد مورد بررسی قرار می‌گیرد و روش ارائه‌شده با الگوریتم‌های معروف به نام ID3، الگوریتم SVM بهبود داده شده [۱۷] که مقاله پایه راهکار پیشنهادی است و شبکه بیزین [۲۳] که اوئویا و همکارانش در سال ۲۰۱۹ در جهت کشف تقلب در صورت‌های مالی مبتنی بر شبکه بیزین ارائه دادند، مورد مقایسه قرار می‌گیرد. راهکار ارائه‌شده با استفاده از مجموعه داده‌هایی که مربوط به سه سال از شرکت‌ها می‌باشد و در این داده‌ها متقلب بودن و یا نبودن شرکت مشخص‌شده است، ارزیابی صورت می‌گیرد.

روشی که برای انتخاب داده‌های آزمون و آزمون مورد استفاده قرار گرفته است، روش k-fold cross validation است. در این گزارش مقدار k برابر مقدار متعارف ۱۰ قرار داده خواهد شد. زیرا این مقدار اثبات شده است که بهترین نسبت برای ارزیابی روش‌های داده‌کاوی می‌باشد؛ بنابراین در این تحقیق نیز از همین تعداد استفاده‌شده است.

داده‌ها در ابتدا توسط برنامه به فرمت مناسب برای تحلیل قرار می‌گیرد یا به عبارتی پیش‌پردازش ابتدایی صورت می‌گیرد فایلی با فرمت ARFF ایجاد است که ساختاری مناسب و استاندارد برای تحلیل می‌باشد.

پیاده‌سازی در برنامه Visual Studio 2017 و با زبان برنامه‌نویسی C# صورت‌گرفته است و در حین کار از کتابخانه‌هایی نیز کمک گرفته‌شده است که عبارت‌اند از:

- Weka: نرم‌افزار Weka در دانشگاه Waikato واقع در نیوزلند توسعه‌یافته است و اسم آن از عبارت "Waikato Environment for knowledge Analysis" استخراج گشته است. همچنین Weka ، نام پرنده‌ای با طبیعت جستجوگر است که پرواز نمی‌کند و در نیوزلند، یافت می‌شود. این نرم‌افزار، یک واسط همگون برای بسیاری از الگوریتم‌های یادگیری متفاوت، فراهم کرده است.
- ZedGraph: این کتابخانه برای رسم نمودار می‌باشد که با استفاده از این کتابخانه، نمودارهای مورد نیاز برنامه ایجاد شده‌اند.

سیستم مورد استفاده در اینجا دارای سیستم‌عامل Windows 10، دارای ۶ گیگ RAM و Corei7 می‌باشد.

در این برنامه در ابتدای کار که قصد داریم تا فایل ورودی برنامه را ایجاد کنیم، پیش‌پردازش و عملیات انتقال را روی داده‌ها انجام می‌دهیم. می‌توان در شکل (۹) تصویری از این برنامه را مشاهده نمود.

شکل (۹): برنامه ساخت ورودی الگوریتم پیشنهادی

با استفاده از نرم‌افزاری که برای ساخت ورودی استفاده می‌شود و تصویر آن نیز در شکل (۹) نشان داده شد، داده‌ها به‌صورت شکل (۱۰) ساخته می‌شوند.

```
@RELATION Test

@ATTRIBUTE Att1 Real
@ATTRIBUTE Att2 Real
@ATTRIBUTE Att3 Real
@ATTRIBUTE Att4 Real
@ATTRIBUTE Att5 Real
@ATTRIBUTE Att6 Real
@ATTRIBUTE Att7 Real
@ATTRIBUTE Att8 Real
@ATTRIBUTE Att9 Real
@ATTRIBUTE Att10    Real
@ATTRIBUTE Att11    Real
@ATTRIBUTE Att12    Real
@ATTRIBUTE Att13    Real
@ATTRIBUTE Att14    Real
@ATTRIBUTE Att15    Real
@ATTRIBUTE Att16    Real
@ATTRIBUTE Att17    Real
@ATTRIBUTE Att18    Real
@ATTRIBUTE Att19    Real
@ATTRIBUTE Att20    Real
@ATTRIBUTE Att21    Real
@ATTRIBUTE Att22    Real
@ATTRIBUTE Att23    Real
@ATTRIBUTE Class {0,1}

@DATA
0.682749,1.611791,0.4957,6.297421,0.030668,0.017538,1.748626,0.04183
0.772834,1.365646,0.81061,6.193666,0.001899,0.002049,0.929795,0.0657
0.82052,1.287812,0.82734,6.284927,0.001605,0.001587,1.011191,0.04950
0.948676,1.147852,-0.2017,5.815516,-0.10852,-0.14391,0.754138,-0.009
1.087529,0.875712,0.33826,6.052555,-0.12507,-0.11055,1.13142,-0.0374
0.859954,1.100309,0.21613,6.142555,0.292207,0.219297,1.33247,0.11012
1.089156,0.868582,0.3291,6.781989,-0.23289,-0.65809,0.853895,-0.1296
```

شکل (۱۰): قسمتی از ورودی آماده شده برای الگوریتم ارائه‌شده توسط روش پیشنهادی

باید خاطرنشان کرد که در ابتدا داده‌های ورودی به‌صورت جدول (۱) می‌باشند.

جدول (۱): داده‌های خام ورودی برنامه قبل از پیش‌پردازش

ردیف	مجموع بدهیها به مجموع داراییها	نسبت جاری	نسبت آنی	لگاریتم طبیعی (بهای تمام شده کالای فروش رفته)	سود خالص به مجموع دارایی ها	سود خالص به فروش	فروش به مجموع دارایی ها	سود عملیاتی به فروش	سود قبل از بهره و مالیات به فروش	سود ناخالص به کل داراییها	سود قبل از بهره و مالیات به کل داراییها	سود قبل از بهره و مالیات به بدهی های جاری	-دارایی جاری) (موجودی کالا به بدهی جاری
0	0.682749	1.611791	0.4957	6.297421	0.030668	0.017538	1.748626	0.041834	0.020268	0.139082	0.035441	0.060261	0.53
0	0.772834	1.365645	0.81061	6.133666	0.001899	0.002043	0.929795	0.065774	0.002043	0.124121	0.001899	0.002714	0.85
0	0.82052	1.287812	0.82734	6.284927	0.001605	0.001587	1.011191	0.049502	0.001587	0.106791	0.001605	0.002162	0.87
0	0.948676	1.147852	-0.2017	5.815516	-0.10852	-0.14391	0.754138	-0.00911	-0.14391	0.118806	-0.10852	-0.18445	-0.09
0	1.087529	0.875712	0.33826	6.052355	-0.12507	-0.11055	1.13142	-0.03741	-0.11055	0.100021	-0.12507	-0.15154	0.37
0	0.859954	1.100309	0.21613	6.142555	0.292207	0.219297	1.33247	0.110123	0.219297	0.288807	0.292207	0.417118	0.3
1	1.089156	0.868582	0.3291	6.781989	-0.23289	-0.65809	0.353895	-0.12968	-0.65809	0.053447	-0.23289	-0.30743	0.46
1	1.071176	1.31503	0.52773	6.657439	-0.19308	-0.86736	0.222609	-0.48255	-0.86736	0.030489	-0.19308	-0.47597	0.64
1	1.648045	1.038729	0.51893	6.856278	-0.26867	-0.58538	0.458968	-0.22249	-0.58538	0.058507	-0.26867	-0.41322	0.56
0	0.516661	1.306147	0.13886	6.277285	0.192015	0.183279	1.047661	0.183305	0.183279	0.23151	0.192015	0.407388	0.15
0	0.536026	1.304754	0.41889	6.195376	0.092011	0.149325	0.616175	0.050279	0.155448	0.091782	0.095783	0.191289	0.5
0	0.601829	1.157904	0.48435	6.30496	0.052251	0.082158	0.635991	0.056311	0.082158	0.074911	0.052251	0.092328	0.64
1	0.855973	1.112174	-0.0609	7.029139	0.104526	0.085584	1.221337	0.126966	0.105428	0.206548	0.128763	0.235167	0.07
1	0.748089	0.700876	-0.1047	7.070353	-0.12133	-0.15375	0.789119	-0.09617	-0.15375	-0.03336	-0.12133	-0.22042	-0.04
1	0.789207	0.539476	-0.0349	7.04097	-0.04807	-0.0571	0.841854	-0.00338	-0.0571	0.045284	-0.04807	-0.07833	0.08

در این پیاده‌سازی برای بررسی الگوریتم‌های دیگر و الگوریتم پیشنهادی از پارامترهای بررسی خطا بسیار معتبر و استاندارد استفاده‌شده است که عبارت‌اند است از:

- خطای میانگین مربعات (MSE): این معیار میانگین خطای مربعات را با استفاده از رابطه (۵-۱)به دست می‌آورد و با استفاده از این معیار می‌توان به طور دقیق پیش‌بینی را سنجید و میزان خطای پیش‌بینی را محاسبه نمود.

(۷)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(r_i - p_i)^2$$

- معیار صحیح مثبت (TP): درصورتی‌که جوابی که از پیش‌بینی به دست می‌آید p باشد و مقدار واقعی نیز p باشد آنگاه ۱ واحد به TP اضافه می‌شود بنابراین در مجموعه داده‌هایی که برای آزمون استفاده می‌شود این مقدار جمع‌آوری می‌شود.
- معیار صحیح منفی (TN): درصورتی‌که جوابی که از پیش‌بینی به دست می‌آید n باشد و مقدار واقعی نیز n باشد آنگاه ۱ واحد به TN اضافه می‌شود بنابراین در مجموعه داده‌هایی که برای آزمون استفاده می‌شود این مقدار جمع‌آوری می‌شود.تفاوت این قسمت با TP در این می‌باشد که در TP ما جواب مثبت و در اینجا جواب‌های منفی را شمارش می‌کنیم.
- معیار غلط مثبت (FP): درصورتی‌که جوابی که از پیش‌بینی به دست می‌آید p باشد و مقدار واقعی n باشد آنگاه ۱ واحد به FP اضافه می‌شود؛ بنابراین در مجموعه‌داده‌هایی که برای آزمون استفاده می‌شود این مقدار جمع‌آوری می‌شود. یعنی به‌طورکلی جوابی که از پیش‌بینی به دست می‌آید متفاوت از مقدار واقعی باشد.
- معیار غلط منفی (FN): درصورتی‌که جوابی که از پیش‌بینی به دست می‌آید n باشد و مقدار واقعی p باشد آنگاه ۱ واحد به Fn اضافه می‌شود بنابراین در مجموعه داده‌هایی که برای آزمون استفاده می‌شود این مقدار جمع‌آوری می‌شود. یعنی به‌طورکلی جوابی که از پیش‌بینی به دست می‌آید متفاوت از مقدار واقعی باشد. تفاوت این قسمت با قسمت قبل در این می‌باشد که در قسمت قبل ما جواب‌های منفی که به اشتباه مثبت در نظر گرفته می‌شدند را در نظر گرفتیم و در اینجا جواب‌های مثبتی که به اشتباه منفی در نظر گرفته‌شده‌اند در نظر می‌گیریم.

بعدازاین که داده‌ها به شکل ورودی برنامه ایجاد شد، الگوریتم پیشنهادی روی داده‌های ورودی اجرا می‌شود.

در این قسمت ویژگی‌های مؤثر در همان ابتدای کار الگوریتم انتخاب می‌شوند و البته باید به این نکته توجه نمود که این ویژگی‌های مؤثر تنها در روش پیشنهادی ما استفاده می‌شوند و در الگوریتم‌های دیگر مورد مقایسه کلیه این ویژگی‌ها داده شده است. بنابراین در این روش پیشنهادی قطعاً سربار محاسبات نیز

پایین‌تر خواهد بود؛ زیرا تنها ویژگی‌های مؤثرتر استفاده‌شده‌اند. به‌منظور سهولت در انجام عملیات، هر نسبت با استفاده از برچسب Att مشخص‌شده که در شکل (۱۱) نشان‌داده‌شده است. در این پژوهش از الگوریتم مجموعه راف و تحلیل سلسله‌مراتبی که در بخش قبل توضیح داده شد به‌منظور تشخیص استفاده و یا عدم استفاده از نسبت‌های مالی استفاده نمودیم که می‌توان در شکل (۱۲) نتایج به‌دست‌آمده برای پارامترهای استفاده‌شده در داده‌ها را مشاهده نمود.

Att1	مجموع بدهی‌ها / مجموع دارایی‌ها
Att2	نسبت جاری = دارایی‌های جاری / بدهی‌های جاری
Att3	نسبت آنی = دارایی‌های جاری – (موجودی کالا+پیش‌پرداخت) / بدهی‌های جاری
Att4	لگاریتم طبیعی (بهای تمام‌شده کالای فروش رفته)
Att5	سود خالص/ مجموع دارایی‌ها
Att6	سود خالص / فروش
Att7	فروش / مجموع دارایی‌ها
Att8	سود عملیاتی/ فروش
Att9	سود قبل از بهره و مالیات/ فروش
Att10	سود ناخالص / کل دارایی‌ها
Att11	سود قبل از بهره و مالیات/ کل دارایی‌ها
Att12	سود قبل از بهره و مالیات / بدهی‌های جاری
Att13	بدهی جاری/ (دارایی جاری – موجودی کالا)
Att14	موجودی کالا / بدهی جاری
Att15	وجه نقد / جمع بدهی‌ها
Att16	بدهی‌های جاری / جمع دارایی‌ها
Att17	سرمایه / جمع دارایی‌ها
Att18	موجودی کالا / فروش
Att19	حساب‌های دریافتنی / فروش
Att20	فروش / دارایی ثابت
Att21	بهای تمام‌شده کالای فروش رفته / فروش
Att22	هزینه‌های عملیاتی / فروش
Att23	موجودی کالا / دارایی جاری

شکل (۱۱): لیست نسبت‌های استفاده‌شده

Att1: 0
Att16: 0
Att15: 0
Att14: 0
Att13: 0
Att23: 0
Att19: 0
Att20: 0
Att18: 0
Att21: 0
Att7: 0
Att22: 0
Att4: 0
Att3: 0
Att2: 0
Att17: 0
Att9: 0.0905061414027294
Att6: 0.0949449035096511
Att12: 0.100474368655747
Att11: 0.104421869607332
Att5: 0.105127953476521
Att8: 0.114018997382377
Att10: 0.120272737712267

شکل (۱۲): مقدار وزن‌های اختصاص داده‌شده به پارامترها

با توجه به نتایج فوق می‌توان نتیجه گرفت که نسبت‌های مالی با ضریب صفر، تأثیری در خروجی ندارد و این در حالی است که نسبت‌های مالی همچون سود قبل از بهره و مالیات به فروش، سود خالص به فروش، سود قبل از بهره و مالیات به بدهی‌های جاری، سود قبل از بهره و مالیات به کل دارایی‌ها، سود خالص به مجموع دارایی‌ها، سود عملیاتی به فروش و سود ناخالص به کل دارایی‌ها بیشترین تأثیر را در متقلب بودن و یا نبودن شرکت دارد. در اینجا می‌توان نسبت‌های مالی بی‌تأثیر را حذف نمود تا الگوریتم قادر باشد با سرعت بیشتری داده‌کاوی را انجام دهد و همین‌طور در همین قسمت می‌توان دریافت که چه نسبت‌هایی بیشترین تأثیر را در متقلب بودن و یا نبودن شرکت دارد و چه عواملی باعث سوق‌دادن شرکت به سمت تقلب است. یعنی می‌توان پارامترهای تأثیرگذار را در اینجا شناسایی نمود و روی این پارامترها توجه بیشتری صورت گیرد تا اینکه بتوان شرکت‌های متقلب را بهتر شناسایی نمود و تصمیمات مهمی را نیز در این راستا اتخاذ نمود.

بعد از اجرا، خروجی‌های شکل (۱۳) تا شکل (۱۶) در برنامه ایجاد می‌شود که مربوط به هر الگوریتم می‌باشد.

```
--------Naive Baysian----------
confusionMatrix:
[0,0] = 18 [0,1]=11
[1,0] = 27 [1,1]=124
Correct Prediction Percent = 78.8888888888889%
InCorrect  Prediction Percent = 21.1111111111111%
MeanAbsoluteError(MAE) = 0.215093567412967
MeanSquaredError(MSE) = 0.450450314025597
RelativeAbsoluteError(REA) = 57.1003786873271
Correct Prediction Number = 142
InCorrect Prediction Number = 38
TP: 124
FP: 11
FN: 27
TN: 18
```

شکل (۱۳): خروجی مربوط به الگوریتم بیزین

```
--------MyAlgorithm----------
confusionMatrix:
[0,0] = 18 [0,1]=9
[1,0] = 27 [1,1]=126
Correct Prediction Percent = 80%
InCorrect  Prediction Percent = 20%
MeanAbsoluteError(MAE) = 0.251683833684513
MeanSquaredError(MSE) = 0.390570453145535
RelativeAbsoluteError(REA) = 66.813909805457
Correct Prediction Number = 144
InCorrect Prediction Number = 36
TP: 126
FP: 9
FN: 27
TN: 18
```

شکل (۱۴): خروجی مربوط به الگوریتم پیشنهادی

```
--------ID3----------
confusionMatrix:
[0,0] = 7 [0,1]=7
[1,0] = 38 [1,1]=128
Correct Prediction Percent = 75%
InCorrect  Prediction Percent = 25%
MeanAbsoluteError(MAE) = 0.343346480854847
MeanSquaredError(MSE) = 0.430065588743028
RelativeAbsoluteError(REA) = 91.147375133409
Correct Prediction Number = 135
InCorrect Prediction Number = 45
TP: 128
FP: 7
FN: 38
TN: 7
```

شکل (۱۵): خروجی مربوط به الگوریتم ID3

```
--------SVM----------
confusionMatrix:
[0,0] = 2 [0,1]=4
[1,0] = 43 [1,1]=131
Correct Prediction Percent = 73.8888888888889%
InCorrect  Prediction Percent = 26.1111111111111%
MeanAbsoluteError(MAE) = 0.261111111111111
MeanSquaredError(MSE) = 0.510990323891863
RelativeAbsoluteError(REA) = 69.3165467625899
Correct Prediction Number = 133
InCorrect Prediction Number = 47
TP: 131
FP: 4
FN: 43
TN: 2
```

شکل (۱۶): خروجی مربوط به الگوریتم ماشین بردار پشتیبان

با توجه به نتایج دریافتی می‌توان به‌وضوح مشاهده نمود که الگوریتم پیشنهادی با ۸۰ ٪ دقت دارای بالاترین دقت صحت و با ۲۰ ٪ اشتباه دارای کمترین میزان اشتباه می‌باشد. می‌توان مشاهده نمود در الگوریتم پیشنهادی که از روش ترکیبی مجموعه راف و تحلیل سلسله‌مراتبی برای انتخاب نسبت‌های مالی مؤثرتر استفاده‌شده است و در ادامه نیز با استفاده از الگوریتم‌های درخت تصمیم بهینه‌شده و شبکه بیزین به بیشترین دقت ممکن رسیده است.

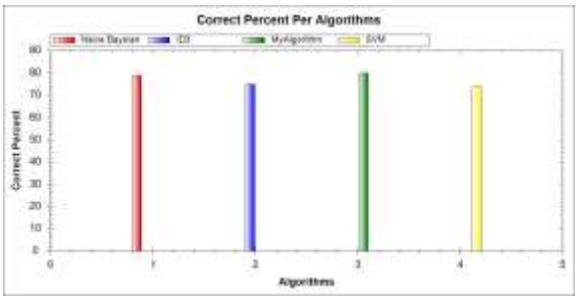
می‌توان در جدول (۲) نسبت بهتری را مشاهده نمود. کاملاً مشخص است که الگوریتم پیشنهادی از باقی بهتر و از خود الگوریتم‌های ID3 و شبکه بیزین که الگوریتم پایه روش پیشنهادی می‌باشند نیز بسیار بهتر عمل می‌کند.

جدول (۲): نسبت درصد صحت پیش‌بینی‌ها و خطای پیش‌بینی‌ها

	درصد صحت پیش‌بینی‌ها	درصد خطای پیش‌بینی‌ها
الگوریتم پیشنهادی	٪۸۰	٪۲۰
الگوریتم بیزین	٪۷۸٫۸۸	٪۲۱٫۱۱
الگوریتم ID3	٪۷۵	٪۲۵
الگوریتم SVM	٪۷۳٫۸۸	٪۲۶٫۱۱

با توجه به این دقت‌های به‌دست‌آمده کاملاً قابل‌مشاهده می‌باشد که روش پیشنهادی از دیگر روش‌ها بسیار بهتر عمل کرده است و دراین‌بین الگوریتم بیزین نیز از الگوریتم‌های ID3 و SVM نیز بهتر عمل کرده است. همچنین الگوریتم ID3 نیز از الگوریتم SVM بهتر عمل کرده است. در اینجا می‌توان کاملاً مشاهده کرد که در روش ترکیبی پیشنهادی بیان شده می‌توان با ادغام روش‌های ID3 و بیزین که هر یک دارای دقتی می‌باشند، به‌دقت بالاتری از هر یک از آن‌ها به‌تنهایی دست‌یافت و در همین قسمت می‌توان مشاهده نمود راهکار پیشنهادی دارای عملکرد بهتری است.

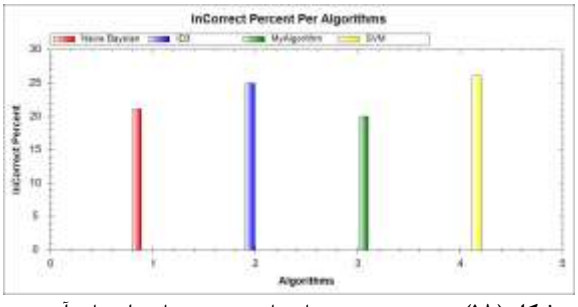
در ادامه نمودارهای به‌دست‌آمده از برنامه مورد بررسی قرار گرفته است. اولین نمودار، نمودار مربوط به درصد پیش‌بینی صحیح در میان داده‌های آزمون است که می‌توان در شکل (۱۷) مشاهده نمود



شکل (۱۷): درصد پیش‌بینی صحیح در میان داده‌های آزمون برای الگوریتم پیشنهادی و دیگر الگوریتم‌های مورد بررسی

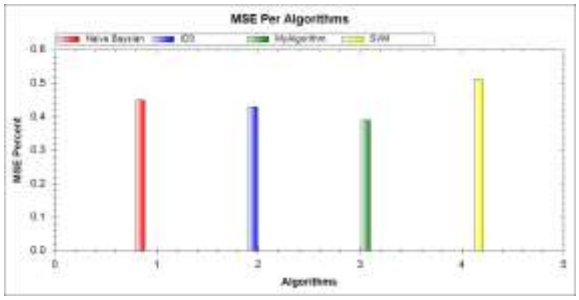
همان‌طور که می‌توان از این نمودار دریافت روش پیشنهادی ما دارای دقت پیش‌بینی صحیح بیشتری نسبت به دیگر الگوریتم‌های مورد بررسی یعنی شبکه بیزین، الگوریتم SVM و الگوریتم ID3 می‌باشد. این بدین دلیل است که ما در روش پیش‌بینی خود تنها مواردی از داده‌های آزمون را در نظر گرفتیم که تأثیر بیشتری را در نتیجه خروجی داشتند و در نتیجه داده‌هایی که در خروجی تأثیر نداشتند را استفاده نکردیم و بدین شکل زمان تحلیل را بسیار کاهش دادیم و حال‌آنکه الگوریتم‌های دیگر به دلیل استفاده از تمامی پارامترها دارای دقت کمتری هستند؛ زیرا ممکن است بعضی از پارامترها دارای مقادیر دوری باشند که ممکن است هیچ تأثیر در نتیجه خروجی نداشته باشند؛ ولی چون در الگوریتم‌های دیگر در ساخت مدل برای پیش‌بینی این پارامترها مورد استفاده قرار گرفته‌اند باعث ایجاد نویز و کاهش دقت می‌شوند و در روش پیشنهادی ما چون این پارامترها بی‌فایده وجود نداشتند در نتیجه دقت در روش پیشنهادی ما افزایش یافت و از دیگر الگوریتم‌ها بهتر عمل نموده است. همچنین در الگوریتم پیشنهادی از یک روش ترکیبی استفاده‌شده است که می‌توان مشاهده نمود به‌خوبی عمل نموده است و از دیگر روش‌ها بسیار بهتر عمل کرده است و از روش‌هایی که بر پایه آن‌ها ایجاد شده است نیز بهتر عمل کرده است.

در نمودار ارائه‌شده در شکل (۱۸) می‌توان درصد پیش‌بینی غلط را مشاهده نمود. با توجه به این نمودار می‌توان درک نمود که روش پیشنهادی به همان دلیل که پیش‌تر دررابطه‌با درصد پیش‌بینی درست گفته شد از باقی روش‌ها دارای مقدار کمتری است؛ یعنی پیش‌بینی اشتباه کمتری دارد؛ بنابراین روش پیشنهادی بهتر از باقی روش‌ها عمل کرده است.



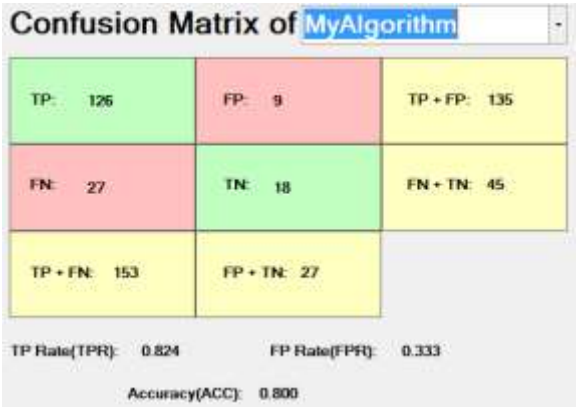
شکل (۱۸): درصد پیش‌بینی‌های ناصحیح در میان داده‌های آزمون برای الگوریتم پیشنهادی و دیگر الگوریتم‌های مورد بررسی

با توجه به نمودار شکل (۱۹) می‌توان به این قضیه پی برد که خطای پیش‌بینی تقریباً در الگوریتم‌های مورد بررسی برابر و نزدیک است؛ ولی الگوریتم پیشنهادی کمتر می‌باشد؛ زیرا هر چه نرخ صحت بیشتر باشد قاعدتاً نرخ غلط پایین‌تر خواهد بود و این نشان‌دهنده عملکرد مناسب روش پیشنهادی است. در ادامه می‌توان مشاهده نمود که نرخ خطای میانگین (MSE) در روش پیشنهادی از تمامی روش‌های دیگر کمتر می‌باشد و الگوریتم ID3 نیز از الگوریتم SVM کمتر است و درعین‌حال الگوریتم بیزین از الگوریتم SVM دارای نرخ خطای کمتری است. این نرخ خطا تنها غلط‌بودن را بررسی نمی‌کند؛ بلکه میزان دور بودن جواب پیش‌بینی‌شده نسبت به جواب واقعی را نیز محاسبه می‌کند که در این حالت مشاهده می‌شود که الگوریتم پیشنهادی بسیار بهتر از دیگر الگوریتم‌ها رفتار می‌کند. اگر در این نمودار توجه شود می‌توان به این قضیه پی برد چرا که الگوریتم ID3 دارای پیش‌بینی غلط بیشتری از الگوریتم بیزین است؛ ولی چون در آنجا میزان دور بودن در نظر گرفته نمی‌شد ID3 بدتر از شبکه بیزین بود ولی می‌توان در شکل (۱۹) مشاهده نمود که الگوریتم ID3 دارای MSE کمتری از الگوریتم بیزین است که این یعنی دارای نرخ خطای کمتری می‌باشد. به‌طورکلی در علم داده‌کاوی MSE دارای اهمیت بسیار بالایی است و دارای اعتبار بسیار زیادی است.

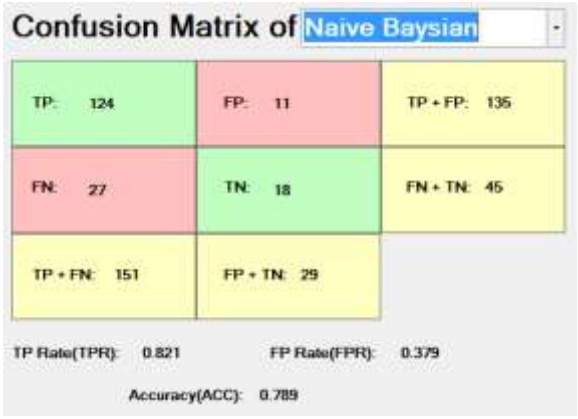


شکل (۱۹): میزان معیار MSE در میان الگوریتم پیشنهادی و دیگر الگوریتم‌های مشابه مورد بررسی

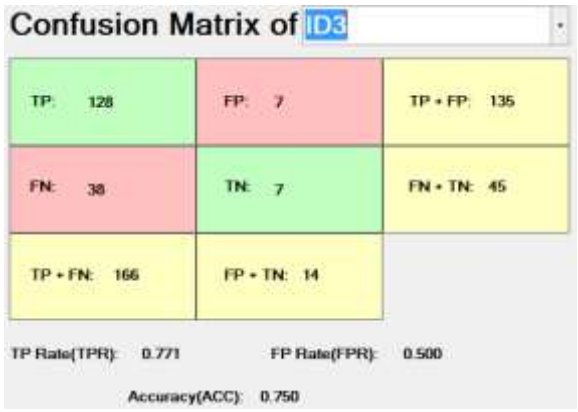
در ادامه Confusion Matrix مربوط به روش پیشنهادی و دیگر روش‌های مورد بررسی در این تحقیق نشان‌داده‌شده است



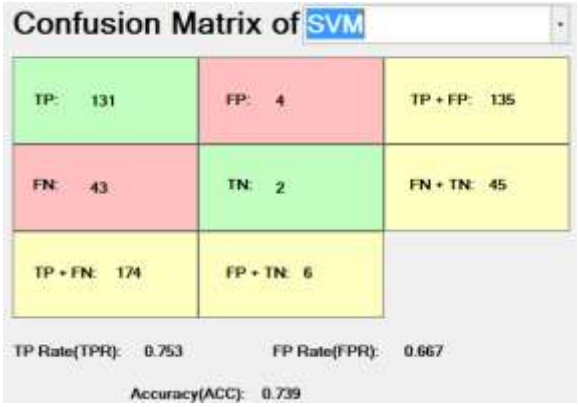
شکل (۲۰): جدول Confusion matrix روش پیشنهادی



شکل (۲۱): جدول Confusion matrix الگوریتم شبکه بیزین



شکل (۲۲): جدول Confusion matrix الگوریتم ID3



شکل (۲۳): جدول Confusion matrix الگوریتم ماشین بردار پشتیبان

در اینجا می‌توان مشاهده نمود که روش پیشنهادی دارای دقت بالاتری می‌باشد؛ زیرا در این جدول دارای مقدار TP و TN بیشتری از دیگر الگوریتم‌ها است و در کنار آن نیز دارای FP و FN کمتری از دیگر الگوریتم‌های مورد بررسی در اینجا است. زیرا هر چه یک الگوریتم دارای TP و TN بیشتری باشد؛ یعنی متقلب بودن و یا نبودن‌های در مجموعه داده‌های آزمون را به مقدار بیشتری درست تشخیص داده است و FP و FN نشان‌دهنده برعکس این قضیه یعنی پیش‌بینی اشتباه می‌باشد.

با توجه به ماتریس‌های Confusion می‌توان نرخ خطا و نرخ صحت عملکرد روش را محاسبه نمود که این روابط در (۸) و (۹) نشان‌داده‌شده‌اند.

$$FPR = \frac{FP}{FP + TP} \quad (۸)$$

$$TPR = \frac{TP}{FP + TP} \quad (۹)$$

۶. جمع‌بندی و کارهای آتی

امروزه دانش به‌عنوان یک منبع ارزشمند و استراتژیک و نیز یک دارایی برای ارزیابی و پیش‌بینی مطرح است و ارائه این راهکارها در زمینه کشف شرکت‌های متقلب باعث افزایش دقت و همچنین کاهش نیروی کار مؤثر و همیشگی برای بررسی و تشخیص شرکت‌های متقلب می‌شود یعنی می‌توان با استفاده از راهکاری مانند راهکار پیشنهادی به‌صورت تمام‌وقت به بررسی و کشف شرکت‌های متقلب پرداخت و این نیازمند به نیروی کار انسانی نمی‌باشد؛ بلکه خود سیستم می‌تواند به‌صورت هوشمندانه تشخیص را انجام داده و اطلاع‌رسانی را انجام دهد. در این پژوهش راهکاری برای ارزیابی و پیش‌بینی بینی تقلب‌های مالی شرکت‌ها ارائه شد و مشاهده شد که روش ارائه‌شده در اینجا دارای عملکرد مناسبی بود و بهبود نسبتاً بالایی را نسبت به الگوریتم‌های پایه خود یعنی ID3 و بیزین نشان داده است که روش پیشنهادی نسبت به الگوریتم ID3، ۶٫۶۶ درصد بهبود و نسبت به بیزین دارای ۸٫۲۷ درصد بهبود عملکرد داشته است. همچنین کار در ادامه با الگوریتم ماشین بردار پشتیبان نیز مورد بررسی قرار گرفت و مشاهده شد که روش پیشنهادی بسیار بهتر از الگوریتم ماشین بردار پشتیبان عمل می‌کند و دارای دقت بالاتر و نرخ خطای کمتری می‌باشد. الگوریتم بیزین از الگوریتم‌های SVM و ID3 بسیار بهتر عمل می‌کند البته مشاهده شد که در صورت بررسی نرخ خطای MSE، ID3 از بیزین دارای نرخ خطای کمتری است و این بدین دلیل است که MSE تنها وابسته به TP، TN، FP و FN نمی‌باشد. داده‌های استفاده‌شده در این تحقیق مربوط به ۶۰ شرکت در طی ۳ سال است که یعنی داده‌های مورد بررسی در اینجا دارای ۱۸۰ رکورد بود. در علم داده‌کاوی داده بسیار دارای اهمیت بالایی می‌باشد؛ زیرا این داده‌ها هستند که باعث ایجاد علم و پیش‌بینی می‌شوند بدین معنی که اگر تعداد داده‌ها در اینجا بیشتر بود قطعاً می‌توانستیم نتایج بهتری را نیز به دست آوریم؛ زیرا راهکار پیشنهادی در اینجا مبتنی بر داده‌کاوی است و برای داده‌کاوی کیفیت داده‌های ورودی و تعداد این داده‌ها بسیار پراهمیت می‌باشد. در اینجا

داده‌ها در ابتدا مورد پیش‌پردازش قرار گرفت و انتقالاتی نیز روی داده‌ها صورت گرفت تا اینکه داده‌ها به داده‌های ورودی مورد نیاز الگوریتم پیشنهادی تبدیل شوند. نتایج به‌دست‌آمده کاملاً بهبود روش پیشنهادی را نشان می‌دهد.

۶-۱. پیشنهادهای آتی

- در این روش پیشنهادی از آنتروپی استفاده‌شده است ولی می‌توان از روش‌های دیگری نیز استفاده نمود و یا این روش را با روش‌های دیگری ادغام نمود. برای مثال درصورتی‌که این روش را با روشی مانند Gain که تابع ارزش است، ادغام نماییم، به‌احتمال زیاد دارای عملکرد بهتری است؛ چراکه آنتروپی نیز دارای معایبی می‌باشد؛ ولی سرعت آن بالاست و در اینجا هم ما به دنبال روشی بودیم که دارای سرعت بالا باشد ولی می‌توان با ادغام این روش و یا روش‌های جایگزین دقت این روش ارائه‌شده را به‌شدت افزایش داد.
- می‌توان روش پیشنهادی را با بهبود در الگوریتم C4.5 نیز استفاده نمود یعنی روش پیشنهادی را روی C4.5 با بهبودی مشابه بهبود که روی ID3 در این تحقیق انجام شد، انجام داد تا عملکرد آن افزایش یابد البته نمی‌توان به‌طورقطع گفت که عملکرد آن بهتر می‌شود بلکه باید این روش مورد آزمایش قرار گیرد تا صحت عملکرد بررسی شود.

مراجع

[1] Andon, Paul, Clinton Free, and Benjamin Scard, 2019, Pathways to accountant fraud: Australian evidence and analysis, Accounting Research Journal 28, vol. 1, pp. 10-44.

[2] Lookman, Sanni, and Selmin Nurcan, 2020, A Framework for Occupational Fraud Detection by Social Network Analysis, In CAISE 2015 FORUM.

[3] Sarno, Riyanarto, Rahadian Dustrial Dewandono, Tohari Ahmad, Mohammad Farid Naufal, and Fernandes Sinaga, 2019, Hybrid Association Rule Learning and Process Mining for Fraud Detection, IAENG International Journal of Computer Science 42, p. no. 2.

[4] Van Vlasselaer, Véronique, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens, 2019, Afraid: fraud detection via active inference in time-evolving social networks, In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis, pp. 659-666.

[5] Mugarura, N., 2020, Uncoupling the relationship between corruption and money laundering crimes.Journal of Financial Regulation and Compliance, 24(1).

[6] Nikoloska, s., Simonovski, I., 2019, Role of banks as entity in the system for prevention of money laundering in the Macedonia, Procedia - Social and Behavioral Sciences 44, 453 – 459.

[7] Armideh, Javad, Asghari Vasksi. Shideh, 2013, Application of Fuzzy Logic in Presenting a Fuzzy Expert System for Diagnosis of Different Mental Illnesses, The First Conference on New Approaches in Computer Engineering and Information Retrieval in Iran, October 6, (In Persian).

[8] S. Krishna Anand, R. Kalpana and S. Vijayalakshmi, 2019, Design and Implementation of a Fuzzy Expert System for Detecting and Estimating the Level of Asthma and Chronic Obstructive Pulmonary Disease, Middle-East Journal of Scientific Research 14 (11): 1435-1444, [ISSN 1990-9233 © IDOSI Publications].

[9] Rupinder Kaur, Amrit Kaur, 2019, Hypertension Diagnosis Using Fuzzy Expert System, International Journal of Engineering Research and Application (IJERA), ISSN: 2248-9622.

[10] Kantesh Kumar Oad, Xu DeZhi & Pinial Khan Butt , 2019, A Fuzzy Rule based Approach to Predict Risk Level of Heart Disease, Global Journal of Computer Science and Technology: C Software & Data Engineering, Volume 14 Issue 3 Version 1.0, Online ISSN: 0975-4172 & Print ISSN: 0975-4350

[11] Adams, J.; and Sargin, E., 2016, Deep neural networks for youtube recommendations, In Proceedings of the 10th ACM Conference on Recommender Systems, 191–198. ACM.

[12] Ojeme Blessing Onuwa et. All, 2019, Fuzzy Expert System for Malaria Diagnosis, An International Open Free Access, Peer Reviewed Research Journal, Published By: Oriental Scientific Publishing Co., India, Vol.7,No. (2):Pgs. 273-284 [ISSN: 0974-6471]

[13] Ziming Yin, Yinhong Zhao, Xudong Lu, and Huilong Duan, 2020, Screening of Alzheimer's Disease Based on Multiple Neuropsychological Rating Scales, Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine, Volume, Article ID 258761, 13 pages.

[14] Site: http://farsithesis.ht3.ir/farsithesis/41484/html, 2018, (In Persian).

[15] Colladon, A. F., Remondi, E. 2020. Using Social Network Analysis to Prevent Money Laundering, Elsevier Science, DOI: 10.1016/j.eswa.2020.09.029.

[16] Chi DJ., Chu CC., Chen D. 2019, Applying Support Vector Machine, C5.0, and CHAID to the Detection of Financial Statements Frauds. In: Huang DS., Huang ZK., Hussain A. (eds) Intelligent Computing Methodologies. ICIC 2019. Lecture Notes in Computer Science, vol 11645. Springer, Cham.

[17] Hao Wang, Chengzhi Mao, Hao He, Mingmin Zhao, Tommi S. Jaakkola, Dina Katabi, 2019, Bidirectional Inference Networks: A Class of Deep Bayesian Networks for Health Profiling", Machine Learning (stat.ML); Artificial Intelligence (cs.AI); Computer Vision and Pattern Recognition (cs.CV); Machine Learning (cs.LG).

[18] Hariri, N. Mobasher, B. and Burke, R., 2017, Context-aware music recommendation based on latenttopic sequential patterns, In Proceedings of the sixth ACM conference on Recommender systems, 131 –138. ACM.

[19] Kim, Yeonkook J., Baik, Bok. Cho, Sungzoon, 2020, Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning, Expert Systems with Applications, No. 62, pp. 32-43.

[20] Lari. Dashtbayaz, Mahmoud, 2019, Data search and discovery process for financial statement fraud, Research Journal of Finance and Accounting, Vol.6, No.3.

[21] Mohamed Yusof. K., Ahmad Khair A.H. & Jon Simon., 2015, Fraudulent Financial Reporting: An Application of Fraud Models to Malaysian Public Listed Companies, The Macrotheme Review. 4(3), (In Persian).

[22] Ravenda D., Valencia-Silva M. M., Argiles-Bosch J. M., García-Blandón J., 2018, Money laundering through the strategic management of accounting transactions, Critical Perspectives on Accounting.

[23] I O Eweoya, A Adebiyi, A Azeta, F Chidozie, F O Agono, B Guembe, 2019, A Naive Bayes approach to fraud prediction in loan default, Journal of Physics: Conference Series, Volume 1299, Number 1.