Journal of Accounting, Auditing and Finance in islamic enviroments
Imam Reza International University
Vol. 2, No. 5 Spring 2025

فصلنامه حسابداری و تأمین مالی در محیط‌های اسلامی

e-issn: 2821-2258

DOI: 10.22034/aafie.2024.413225.1033

Cite the article: Malekian, E., Eghbal Sefat Ronaghi, E., Ghamari Moghaddam, A., & Malekian, M. (2025). Providing a method to detect fraud in the financial statements of companies active in the Tehran Stock Exchange using machine learning algorithms and optimized decision trees. *Journal of Accounting, Auditing and Finance in islamic enviroments*, 2(5), 60-95.

Providing a method to detect fraud in the financial statements of companies active in the Tehran Stock Exchange using machine learning algorithms and optimized decision trees

esfandiyar malekian [1], esmael eghbal sefat ronaghi[2], amin ghamari moghaddam[3], mostafa malekian [4]

1. Professor of Accounting Department, Faculty of Economic and Administrative Sciences, Mazandaran University, Babolsar, Iran (Corresponding Author) (malekian@umz.ac.ir)
2. PhD student in accounting, Faculty of Economic and Administrative Sciences, Mazandaran University, Babolsar, Iran (esmaeleghbal@yahoo.com)
3. PhD student in accounting, Department of Accounting, Central Tehran Branch, Islamic Azad University, Tehran, Iran (amin.auditor2667@gmail.com)
4. Graduated with a doctorate in accounting, Faculty of Accounting and Financial Sciences, University of Tehran, Iran (mostafa.malekian@ut.ac.ir)

ABSTRACT

Received: 24/08/2023 - Accepted: 02/06/2024

Today, knowledge is considered as a valuable and strategic resource and an asset for evaluation and prediction. It leads to providing solutions in the field of detecting frauds in the financial statements of companies, which increases accuracy and reduces ineffective labor to investigate and detect fraudulent companies. By using resolutions such as the full-time proposed solution, it is possible to identify and investigate the companies accused of fraud. This does not require human labor, but the system itself can smartly identify and inform. In the past, various solutions for detecting fraud were presented, each of which had problems. Therefore, the present research presents a method to detect fraud in the financial statements of companies with the help of artificial intelligence methods including machine learning algorithms. For this purpose, at first, after data preprocessing and data transfer, features (independent variables) were selected using the combined algorithms of the Ruff set and hierarchical analysis, and by training, calculating and testing the weights of these features through the algorithm Machine learning models of these algorithms were presented to predict the fraud of financial statements. Finally, the prediction accuracy of the proposed method was checked with some of the previous methods. The results demonstrated that the proposed method performs better than the traditional ways.

Keywords: fraud detection, support vector machine, Bayesian network, rough set, hierarchical analysis, optimized decision tree.

فصلنامه حسابداری، حسابرسی و تأمین مالی در محیط‌های اسلامی
دانشگاه بین المللی امام رضا (ع)
جلد ۲، شماره ۵، بهار ۱۴۰۴
e-issn: 2821-2258
DOI: 10.22034/aafie.2024.413225.1033
استناد به مقاله: ملکیان، اسفندیار، اقبال صفت رونقی، اسماعیل، قمری مقدم، امین و ملکیان، مصطفی. (۱۴۰۴). ارائه روشی برای کشف تقلب در صورت‌های مالی شرکت‌های فعال در بورس اوراق بهادار تهران به کمک الگوریتم های یادگیری ماشین و درخت تصمیم بهینه شده. *حسابداری، حسابرسی و تأمین مالی در محیط های اسلامی*، ۲ (۵) ، ۹۵-۶۰.

ارائه روشی برای کشف تقلب در صورت‌های مالی شرکت‌های فعال در بورس اوراق بهادار تهران به کمک الگوریتم های یادگیری ماشین و درخت تصمیم بهینه شده

اسفندیار ملکیان [۱]، اسماعیل اقبال صفت رونقی [۲]، امین قمری مقدم [۳]، مصطفی ملکیان [۴]

۱. استاد گروه حسابداری دانشکده علوم اقتصادی و اداری، دانشگاه مازندران، بابلسر، ایران. (نویسنده مسئول) (malekian@umz.ac.ir)
۲. دانشجوی دکترای حسابداری، دانشکده علوم اقتصادی و اداری، دانشگاه مازندران، بابلسر، ایران. (esmaeleghbal@yahoo.com)
۳. دانشجوی دکترای حسابداری، گروه حسابداری، واحد تهران مرکز ، دانشگاه آزاد اسلامی، تهران، ایران. (amin.auditor2667@gmail.com)
۴. دانش آموخته دکترای حسابداری، دانشکده حسابداری و علوم مالی، دانشگاه تهران، ایران. (mostafa.malekian@ut.ac.ir)

چکیده

تاریخ دریافت مقاله: ۱۴۰۲/۰۴/۲۲ تاریخ پذیرش مقاله: ۱۴۰۳/۰۳/۱۳

امروزه دانش به‌عنوان یک منبع ارزشمند و راهبردی و یک دارایی برای ارزیابی و پیش‌بینی مطرح است. ارائه راهکارها در زمینهٔ کشف شرکت‌های متقلب، سبب افزایش دقت و کاهش نیروی کار غیرمؤثر برای بررسی و تشخیص شرکت‌های متقلب می‌شود. با استفاده از راهکاری مانند راهکار پیشنهادی به‌صورت تمام‌وقت، می‌توان شرکت‌های متقلب را شناسایی و کشف کرد. این مهم نیازمند نیروی کار انسانی نیست، بلکه خود سیستم می‌تواند به‌صورت هوشمندانه تشخیص دهد و اطلاع‌رسانی کند. در گذشته، راهکارهای مختلفی برای تشخیص تقلب ارائه شد که هر یک دارای مشکلاتی بود. از این‌رو، پژوهش حاضر روشی را برای کشف تقلب در صورت‌های مالی شرکت‌ها به‌کمک روش‌های هوش مصنوعی ـ شامل الگوریتم‌های یادگیری ماشین ـ ارایه کرده است. به این منظور، پس از پیش‌پردازش داده‌ها و انتقال داده ها، با استفاده از الگوریتم‌های ترکیبی، مجموعه راف و تحلیل سلسله‌مراتبی ویژگی‌ها (متغیرهای مستقل) انتخاب شدند و با آموزش و محاسبه و آزمون اوزان این ویژگی‌های از طریق الگوریتم‌های یادگیری ماشین، مدل‌هایی از این الگوریتم‌ها برای پیش‌بینی تقلب صورت‌های مالی ارائه شد. در نهایت، میزان صحت پیش‌بینی روش پیشنهادی با چند مورد از روش‌های پیشین بررسی شد. نتایج حاکی از عملکرد بهتر روش پیشنهادی نسبت به آن‌هاست.

کلیدواژها: تشخیص تقلب، ماشین بردار پشتیبان، شبکهٔ بیزین، مجموعهٔ راف، تحلیل سلسله‌مراتبی، درخت تصمیم بهینه‌شده.

مقدمه

در سال‌های اخیر، بازارهای مالی ایالات متحده با افشای متعدد اعمال متقلبانهٔ برخی شرکت‌ها، به‌طور جدی متضرر شده‌اند. ورلدکام، انرون، آدلفیا، گلوبال کروسینگ و تیکو فقط تعداد اندکی از رسوایی‌های صورت‌های مالی هستند که بازار سهام را دچار نوسان واعتماد عمومی را سلب کرده‌اند (پائول و اسکارد[1]، ۲۰۲۰).

تقلب صورت‌های مالی به‌طور فزاینده‌ای به یک مشکل جدی برای کسب‌وکار، دولت و سرمایه‌گذاران تبدیل شده است. در واقع، این مسئله قابلیت اطمینان بازارهای سرمایه، رؤسای شرکت‌ها و حتی حرفهٔ حسابرسی را تهدید می‌کند. حسابرسان به‌طور خاص با ناتوانیِ ظاهری خود در کشف تقلب در مقیاس بزرگ، مواجه هستند. قضاوت‌های پولی در مقیاس صدها میلیون دلار برضدّ شرکت‌های دارای خدمات حسابداریِ عمومیِ رسمی، به‌طور معمول وجود دارد. از دیدگاه حسابرسی، تقلب مسئله‌ای بسیار جدی است زیرا اغلب با تلاش برای مخفی‌سازی، تحریف و گمراه‌ساختنِ استفاده‌کنندگان از سوابق و گزارش‌های واحد تحت حسابرسی همراه است. تلاش برای ارائهٔ غلطِ اطلاعات می‌تواند در سطح مدیریت نیز اتفاق بیفتد. در پیِ فروپاشی شرکت‌های بزرگ، این دیدگاه به‌صورت گسترده‌ای مورد توجه قرار گرفته است. کاربرد داده‌کاوی، تنها به تعامل‌های اجتماعی، علوم و مهندسی محدود نیست، بلکه علاوه بر آن‌ها در سامانه‌های پیشنهاددهنده، سامانه‌های مالی، ضدّجاسوسی و... نیز مورد استفاده قرار می‌گیرند. با وجود کاربردهای گستردهٔ روش‌های داده‌کاوی در علوم مختلف، تاکنون نرخ اتخاذ این روش‌ها در میان دانشگاهیان و سازمان‌های کنترلی به‌منظور کشف تقلب و مخاطره، چندان چشمگیر نبوده است. در پژوهش‌های قبل (اقدم[2] و همکاران، ۲۰۲۳؛ چی[3] و همکاران، ۲۰۱۹) با استفاده از الگوریتم‌های یادگیری ماشین تقلّب در صورت‌های مالی را پیش‌بینی کردند. تمایز این مطالعه با پژوهش‌های قبلی، تعیین ویژگی‌های تأثیرگذار در این پیش‌بینی با استفاده از مجموعه راف و تحلیل سلسله‌مراتبی و پس از آن،

1. Paul & scard
2. Aghdam
3. Chi

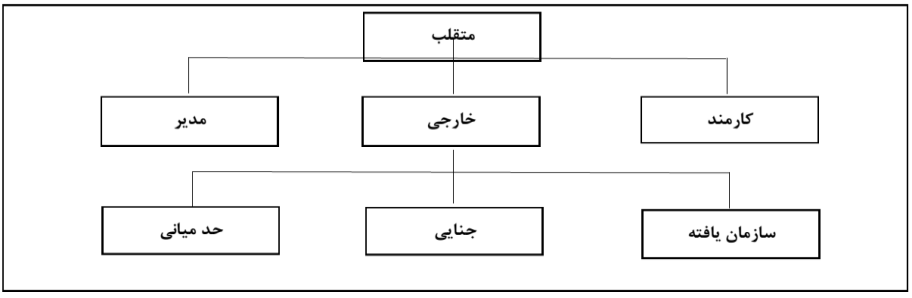
ارائه روشی با استفاده از الگوریتم‌های یادگیری ماشین برای ارزیابی و پیش‌بینی تقلب‌های مالی شرکت‌ها بود. روش ارائه‌شده دارای عملکرد مناسبی بود و بهبود نسبتاً بالایی را نسبت به الگوریتم‌های پایه خود، یعنی الگوریتم درخت تصمیم[1] و ماشین بردار پشتیبان،[2] نشان می‌دهد. پژوهش‌های قبلی بیشتر به وقوع یا عدم وقوع تقلب با استفاده از الگوریتم‌های یادگیری پرداختند و روش‌ها در پیش‌بینی در فضای دوکلاسه یا چندکلاسه (صورت‌های مالی سالم و متقلب) کاربرد داشت. این مطالعه می‌تواند خلأ پژوهش در این زمینه را پر کند تا روشی ارائه شود که با استفاده از آنتروپی و ارائه راهکاری مفید، تقلب را پیش‌بینی کند.

در این پژوهش، نخست مفاهیم لازم برای درک روش پیشنهادی بیان می‌شود. سپس، مروری بر کارهای پیشین صورت می‌گیرد. در ادامه، روش پیشنهادی بیان‌شده ارزیابی و مقایسه می‌شود. در بخش انتهایی مقاله، جمع‌بندیِ کلّی ارائه و پیشنهادهایی برای پژوهش‌های آتی بیان می‌شود.

مبانی نظری پژوهش

تقلب

سوءاستفاده از منافع یک سازمان، بدون آنکه به پیامدها و عواقب قانونیِ مستقیم منجر شود، تقلب نامیده می‌شود (لوکمن و سلمین[3]، ۲۰۱۹).



شکل ۱: نمودار سلسله‌مراتبی از ارتکاب جرایم یقه‌سفید از هر ۲ دیدگاه: سطح ـ شرکت و سطح ـ جامعه (لوکمن و سلمین، ۲۰۱۹)

1. Iterative Dichotomiser
2. SVM
3. Lookman & Selmin

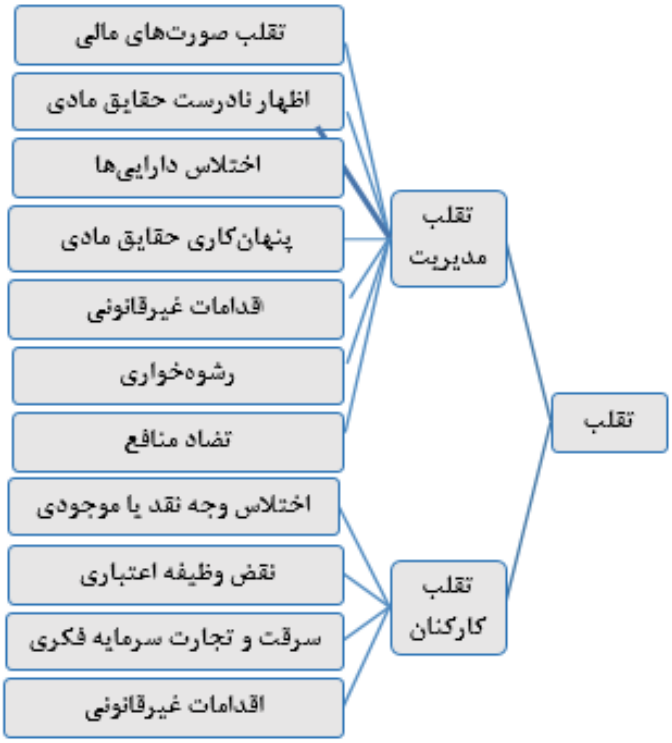
چنان‌که در شکل ۱ قابل مشاهده است، به‌منظور بررسی موشکافانهٔ چالش مورد بررسی و ارائه راهکاری مؤثر با هدف کشف تقلب، فرد یا افراد متقلب را به‌طور کلی به دو گروهِ «داخلی»(شغلی) و «خارجی» تقسیم و تعریف دقیق‌تری برای هریک از گروه‌های تقلب ارائه می‌کنند.

- تقلب شغلی یا داخلی: به‌صورت ضمنی، سوءِاستفاده یا استفادهٔ نادرست از منابع و دارایی‌های یک سازمان، به‌منظور منافع شخصی، که از طریق یکی از عنوان‌های شغلی صورت می‌گیرد، «تقلب شغلی» یا «تقلب داخلی» نامیده می‌شود.
- تقلب خارجی: به تقلبی گفته می‌شود که فرد یا افرادی خارج از سازمان، مرتکب آن می‌شوند.

انواع تقلب

رضایی و رحمانی (۱۴۰۱) تقلب را به دو گروهِ «تقلب مدیریت» و «تقلب کارکنان» تقسیم و در ذیل آن، طبقه‌بندی بیشتری از این دو نوع تقلب ارائه می‌کنند که در شکل ۲ نشان داده شده است. تقلب می‌تواند به چندین نوع تقسیم شود که معمول‌ترین آن، مصادرهٔ دارایی‌ها و اشتباهات مالی است. مصادرهٔ دارایی‌ها اغلب به تقلب کارکنان ـ شامل اختلاس، سرقت وجه نقد یا موجودی و تقلب حقوق و دستمزد ـ مربوط می‌شود. اشتباهات مالی، تقلب در صورت‌های مالی شناخته می‌شود که اغلب مسئولیت آن با مدیریت است.

وزارت دادگستری آمریکا، تقلب شرکت را در سه حوزهٔ گسترده تعریف می‌کند: «تقلب حسابداری» یا تقلب مالی، «تخطّی کارکنان» و «رفتار انحرافی». تقلب حسابداری شامل تحریف اطلاعات مالی از طریق حساب‌سازی یا گمراه‌کردن سرمایه‌گذاران است. رایج‌ترین طرح‌های حسابداری، شامل فروش موجودی‌ها، معاملات جانبی، معاملات مبادله‌ای، هزینه‌های سرمایه‌گذاری، کسب سریع درآمد و هزینه‌های معوّق است (رضایی و همکاران، ۱۴۰۱).



شکل۲: انواع تقلب (رضایی و همکاران،۱۴۰۱)

یادگیری ماشین

یادگیری ماشین یعنی چگونه می‌توان برنامه‌ای نوشت که از طریق تجربه، یادگیری و عملکرد خود را بهتر کند. یادگیری ممکن است سبب تغییر در ساختار برنامه و یا داده‌ها شود. یادگیری ماشین زمینهٔ نسبتاً جدیدی در علوم رایانه است که در حال حاضر دوران رشد و تکامل خود را می‌گذراند. یادگیری ماشین یک زمینهٔ تحقیقاتیِ بسیار فعال در علوم رایانه است (خانجانی، ۱۴۰۰). علوم مختلفی با یادگیری ماشین ارتباط دارند؛ از جمله، هوش مصنوعی، روان‌شناسی، فلسفه، تئوری اطلاعات، آمار و احتمالات، تئوری کنترل و... .

از دلایل استفاده از یادگیری ماشین برای حلّ مسائل می‌توان به موارد زیر اشاره کرد:

- ممکن است در حجم زیادی از داده‌ها، اطلاعات مهمی نهفته باشد که بشر قادر به تشخیص آن‌ها نباشد (داده‌کاوی).

- ممکن است هنگام طراحی یک سیستم، تمامی ویژگی‌های آن شناخته‌شده نباشد؛ در حالی‌که ماشین می‌تواند حین کار، آن‌ها را یاد بگیرد.
- ممکن است محیط در طول زمان تغییر کند. ماشین می‌تواند با یادگیریِ این تغییرات، خود را با آن‌ها وفق دهد.

برای برخی از کاربردهای یادگیری ماشین، می‌توان به مواردی چون کنترل روبات‌ها، داده‌کاوی، تشخیص گفتار، شناسایی متن، پردازش داده‌های اینترنتی، بیوانفورماتیک، بازی‌های رایانه‌ای و هزاران نمونهٔ دیگر اشاره کرد. مبانی ارزیابی الگوریتم‌های یادگیری ماشین شامل «دقت دسته‌بندی»، «صحّت راه‌حل و کیفیت آن» و «سرعت عملکرد» است. یادگیری ماشین به دو دستهٔ کلیِ «یادگیری با ناظر» و «یادگیری بدون ناظر» تقسیم می‌شود (لوکمن و سلمین، ۲۰۱۹).

درخت تصمیم

ساختار درخت تصمیم در یادگیری ماشین، یک مدل پیش‌بینی‌کننده است که حقایقِ مشاهده‌شده دربارهٔ یک پدیده را به استنتاج‌هایی دربارهٔ مقدار هدف آن پدیده تبدیل می‌کند. تکنیک یادگیری ماشین برای استنتاج یک درخت تصمیم از داده‌ها، «یادگیری درخت تصمیم» نامیده می‌شود که یکی از روش‌های رایجِ داده‌کاوی است.

هر گره داخلیِ متناظر یک متغیر و هر کمان به یک فرزند، نمایانگر یک مقدار ممکن برای آن متغیر است. یک گره برگ، با داشتن مقادیر متغیرها ـ که با مسیری از ریشهٔ درخت تا آن گره برگ بازنمایی می‌شود ـ مقدار پیش‌بینی‌شدهٔ متغیر هدف را نشان می‌دهد. یک درخت، تصمیم ساختاری را نشان می‌دهد که برگ‌ها نشان‌دهندهٔ دسته‌بندی و شاخه‌ها ترکیب‌های فصلیِ صفاتی هستند که نتایج این دسته‌بندی‌ها را بازنمایی می‌کنند (وانگ[1] و همکاران، ۲۰۱۹). یادگیری یک درخت می‌تواند با تفکیک‌کردن یک مجموعه منبع به زیرمجموعه‌هایی بر اساس یک تست مقدار صفت، انجام شود. این فرایند به شکل بازگشتی

1. Wang

در هر زیرمجموعهٔ حاصل از تفکیک، تکرار می‌شود. عمل بازگشت زمانی کامل می‌شود که تفکیکِ بیشتر سودمند نباشد یا بتوان یک دسته‌بندی را به همهٔ نمونه‌های موجود در زیرمجموعهٔ به‌دست‌آمده، اعمال کرد.

درختان تصمیم قادر به تولیدِ توصیف‌های قابل درک برای انسان از روابط موجود در یک مجموعه داده‌ای هستند و می‌توانند برای وظایف دسته‌بندی و پیش‌بینی به‌کار روند. این تکنیک به شکل گسترده‌ای در زمینه‌های مختلف همچون تشخیص بیماری، دسته‌بندی گیاهان و راهبردهای بازاریابی مشتری به‌کار رفته است (رضایی و همکاران، ۱۴۰۰).

این ساختار تصمیم‌گیری می‌تواند به شکل تکنیک‌های ریاضی و محاسباتی که به توصیف، دسته‌بندی و عام‌سازی یک مجموعه از داده‌ها کمک می‌کنند نیز معرفی شوند.

انواع صفات در درخت تصمیم به دو نوع «صفات دسته‌ای» و «صفات حقیقی» تقسیم می‌شود. صفات دسته‌ای، صفاتی هستند که دو یا چند مقدار گسسته می‌پذیرند (یا صفات سمبلیک)؛ در حالی که صفات حقیقی مقادیر خود را از مجموعه اعداد حقیقی می‌گیرند.

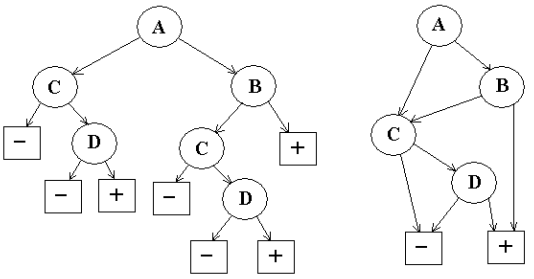
توسعه درختان تصمیم با گراف های تصمیم

گراف‌های تصمیم، تعمیمی از درخت‌های تصمیم است که دارای برگ و گرهِ تصمیم هستند. یک ویژگی که گراف‌های تصمیم را از درختان تصمیم متمایز می‌کند، آن است که گراف‌های تصمیم می‌توانند دارای پیوند باشند. پیوند، حالتی است که دو گره، یک فرزند مشترک داشته باشند و این وضعیت، بیانگر دو زیرمجموعه است که ویژگی های مشترک دارند. از این‌رو، یک مجموعه در نظر گرفته می‌شوند. در درخت تصمیم، تمام مسیرها از گرهِ ریشه به گرهِ برگ با ترکیب عطفی[1] «یا» پیش می‌رود. در یک گراف تصمیم، ممکن است از ترکیب‌های فصلی برای پیوند دو یا چند مسیر با یکدیگر استفاده کرد.

روشی که اشیا در گراف‌های تصمیم دسته‌بندی می‌شوند، همان روشِ به‌کار رفته در درختان تصمیم است. هر درخت تصمیم و گراف تصمیم، یک دسته‌بندی را تعریف می‌کنند

1. AND

(یک افراز از فضای شیء به دسته‌های مجزا). مجموعه توابع قابل نمایش توسط گراف، دقیقاً همانند مجموعهٔ قابل نمایش توسط درخت است؛ هرچند مجموعه‌دسته‌هایی که در تعریف یک تابع تصمیم وارد می‌شوند، متفاوت است. برای مثال، دسته‌بندی برای تابع $(A \wedge B) \vee (C \wedge D)$ متفاوت است. گراف و درخت تصمیم متناظر این تابع، در شکل ۳ نشان داده شده است. درخت تصمیم فضای شیء را به هفت دسته تقسیم می‌کند، در حالی‌که گراف تصمیم این فضا را به دو دسته افراز می‌کند (وانگ و همکاران، ۲۰۱۹).



شکل ۳: نمونه‌ای از گراف تصمیم (وانگ و همکاران، ۲۰۱۹)

مروری بر کارهای گذشته

اقدم و همکاران (۲۰۲۳) در مطالعه‌ای نشان دادند کشف تقلب مالی به چهار دلیل اصلی، یک موضوع چالش‌برانگیز است: «کلاهبرداری همواره در حال تغییر رفتار»، «نبودِ سازوکاری برای ردیابی داده‌های تقلب»، «محدودیت‌های خاصّ تکنیک‌های شناسایی موجود» (به‌عنوان الگوریتم‌های یادگیری ماشین) و «مجموعه داده‌های کلاهبرداریِ مالیِ بسیار پراکنده». بنابراین، می‌توان گفت که الگوریتم‌های آموزشی پیچیده هستند. افزون بر این، نگران با ارزش حساسیت مدل‌ها، مدل درختی رگرسیون تقویت‌شده بالاترین حساسیت را به درستیِ نبودِ تقلب مالی دارد. بی آووکس[1] و همکاران (۲۰۲۲) در پژوهشی یک رویکرد جدید برای تشخیص تقلب بر اساس تقویت گرادیان و شبکه‌های عصبی را پیشنهاد

1- Biao X

کردند. به‌منظور ترکیب مزایای هر دو روش مرسوم و یادگیری عمیق، ابتدا درخت تصمیم‌گیریِ نرمِ یک مدلِ ساختاریافتهٔ درخت تصمیم با شبکه‌های عصبی، به‌عنوان گره‌های آن، را به کار بردند و نشان دادند که این الگوریتم‌ها می‌تواند عملکرد را به طور قابل توجهی بهبود ببخشد و در عین حال، قابلیتِ تفسیرپذیریِ خوبی را حفظ کند. وانگ و همکاران (۲۰۱۹) در مطالعه‌ای با عنوان «نقش حسابرسان در جلوگیری، کشف و گزارشگری تقلب در کشور امارات» فرایندهایی را که حسابرسان داخلی و خارجی برای شناسایی تقلب در طول حسابرسی دنبال می‌کنند، شناسایی کردند. نتایج این تحقیق نشان‌دهندهٔ این امر بود که مسئولیتِ شناساییِ تقلب و گزارش آن با حسابرسان داخلی بوده است. همچنین، حسابرسان خارجی نیز باید جستجوی خود را برای شناسایی و کشف تقلب در گزارش‌های مالی افزایش دهند (چی و همکاران، ۲۰۱۹). راوندا[1] و همکاران (۲۰۱۸) پروکسی‌های جدید مدیریت معاملات را برای کشف شواهد تجربی از مدیریت راهبردیِ معاملات حسابداری، با هدف انجام عملیات پول‌شویی، در نمونه‌ای از ۳۵۵ شرکت تحت کنترل مافیای ایتالیا بررسی کردند. در این مقاله نشان داده شد، با استفاده از تجزیه‌وتحلیل خوشه‌ای، شرکت‌های تحت کنترل مافیا را می‌توان به دو گروهِ مختلفِ 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 مدیریت معاملات[2] این شرکت‌ها را ارائه می‌دهد که ممکن است با فعالیت‌های خاصّ پول‌شویی ارتباط داشته باشد. این مطالعه، پیشنهاد جدید پروکسی‌های را بر اساس ماهیت معاملهٔ هزینه انجام می‌دهد که می‌تواند توسط مقامات، به‌عنوان پرچم‌های قرمز برای فعالیت‌های پول‌شویی مورد استفاده قرار گیرد. افزون بر این، این مطالعه ممکن است از استدلال‌های مهم در برابر دیدگاه ارتدوکس از نقش پول نقد حسابداری و مناسب بودن پروکسی‌های سنّتی TRM برای نشان‌دادن شیوه‌های درونِ شرکت‌هایی که ویژگی‌های مشترکی با شرکت‌های تحت کنترل مافیا دارند، پشتیبانی کند

1. Ravenda
2. TRM

(چن[1]، ۲۰۱۶). همچنین، چن در پژوهشی دیگر با هدف ایجاد یک مدل معتبر کشف تقلب در صورت‌های مالیِ شرکت‌های فعال در بورس تایوان در بین سال‌های ۲۰۰۲ تا ۲۰۱۳، هر دو صورت‌های مالیِ جعلی و غیرجعلی را به آزمون گذاشت. در مرحله اول، در انتخاب متغیرهای اصلی، دو الگوریتم درخت تصمیم‌گیری، استفاده شد. مرحلهٔ دوم شامل ترکیب درخت‌های تصمیم با شبکهٔ اعتقادی بیزی، دستگاه بردار پشتیبان و شبکهٔ عصبیِ مصنوعی به‌منظور ساخت مدل‌های تشخیص تقلب بود. نتایج نشان داد عملکرد تشخیص مدل درخت تصمیم با دقت کلی ۸۷.۹۷٪ مؤثرترین مدل است.

روش‌شناسی پژوهش

برای تشخیص و طبقه‌بندی واحدهای اقتصادی، به شرکت‌های متقلب یا سالم در گزارشگری مالی چارچوب نظری خاصی وجود ندارد. با توجه به استانداردهای شماره ۲۴۰ و ۴۵۰ حسابرسی ایران، می‌توان برمبنای مقدار و محتوای ویژگی‌های داده‌های صورت‌های مالی حسابرسی‌شده توسط حسابداران رسمی یا سازمان حسابرسی، معیارهایی را برای محسوب نمودن تحریف ارائه کرد. از این‌رو، با توجه به تحقیقات پیشین، از قبیل تحقیق داغمه‌چی فیروزجایی (۱۳۸۹)، فرقاندوست حقیقی و بروازی (۱۳۸۸) برای طبقه‌بندی شرکت‌های نمونه به متقلب و سالم، باید معیارهای زیر طیّ حداقل سه سال (۱۳۹۳ تا ۱۳۹۵) در صورت‌های مالیِ شرکت‌های متقلب، صادق بوده و وجود سه شرط زیر به‌عنوان شرایط طبقه‌بندی در گروه شرکت‌های دارای احتمال تقلب است. این شروط عبارتند از:

۱- اظهارنظر غیر مقبول حسابرسی ۲- وجود اختلاف‌های مالیاتی با حوزهٔ مالیاتی، طبق یادداشتِ ذخیرهٔ مالیات بردرآمد و پروندهٔ مالیاتی و بند شرط گزارش حسابرسی۳- وجود تعدیلات سنواتی با اهمیت و صورت‌های مالیِ تجدید ارائه‌شده.

دلایل انتخاب این معیارها این است که دربارهٔ معیار اول، وجود تقلب با اهمیت، می‌تواند زمینه‌ساز اظهارنظر غیر مقبول باشد. دربارهٔ معیار دوم، اختلاف‌های مالیاتی به‌طور عمده ناشی

1. Chen

از تفسیرِ نادرستِ قوانینِ مالیاتی و اشتباه در به‌کارگیریِ بندهای قوانین ذی‌ربط و در برخی موارد، تأخیر در شناسایی مالیات و حفظ نقدینگی و سایر موارد احتمالیِ تخلف است. دربارهٔ معیار سوم، موارد اشتباه و دستکاری اقلام، به‌ویژه اقلام سود و زیانی در سنوات قبل، زمینه‌ساز بروز ارائهٔ دوبارهٔ صورت‌های مالی و دلیلی بر احتمال تقلب در صورت‌های مالی است.

بدین‌ترتیب، به‌وسیلهٔ این معیارها نخست فهرستی از شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران که بین سال‌های ۱۳۹۳ تا ۱۳۹۵ مرتکب تقلب در صورت‌های مالی شده‌اند، تهیه و با توجه به در دسترس‌بودن اطلاعات شرکت‌ها، تعداد شرکت‌های متقلّب تعیین می‌شود. سپس، تعداد شرکت‌های سالم را در همین دامنهٔ زمانی تعیین و بر اساس روش نمونه‌گیریِ تصادفیِ ساده، نمونه‌های کافی انتخاب می‌شوند.

امکانِ تطبیق دادنِ شرکت‌های دو گروه از نظر صنعت وجود ندارد، زیرا صنعت یا صنایع مشابهی که به‌اندازهٔ کافی هم دارای شرکت‌های متقلب و هم شرکت‌های سالم باشد، موجود نیست. بنابراین، در نمونه‌گیری از کلّ شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران، استفاده خواهد شد. البته، حسنِ متنوع‌بودنِ صنایع این است که تعمیم‌پذیری مدل افزایش می‌یابد. تنها شرکت‌ها از نظر سال مالی با هم تطبیق داده می‌شوند.

متغیرهای مستقل در این پژوهش، نسبت‌های مالی هستند؛ به‌طوری‌که با مطالعه و بررسیِ پژوهش‌های صورت‌گرفته، نسبت‌های مالیِ مورد نیاز در این پژوهش، ابتدا ۱۲۵ نسبت مالی انتخاب شدند. اما به‌منظور پرهیز از وجود همبستگی بالا میان برخی از نسبت‌ها و ارائه‌نکردن اطلاعات مشابه، نسبت‌های دارای همبستگی بالا به‌وسیلهٔ آزمون تی شناسایی و حذف گردید. این ترکیب، از تحلیل همبستگی و آزمون تی، به انتخاب نهاییِ تعداد ۵۴ متغیر مستقل منجر شد که اطلاعاتِ معنادار و غیرهم‌پوشی را ارائه می‌کنند.

سپس، این متغیرها را از صورت‌های مالیِ شرکت‌های متقلب و سالم در سال‌های مورد مطالعه استخراج و به‌وسیلهٔ مدل رگرسیون خطی، متغیرهای دارای همبستگی معنادار با صورت‌های مالی متقلبانه، انتخاب و از آن‌ها به‌عنوان متغیرهای ورودی استفاده می‌شود. البته، در ادامه بیان می‌شود که این داده‌ها، داده‌های ورودیِ روش پیشنهادی می‌باشند و جزءِ روش

پیشنهادی نیستند. در روش پیشنهادی، یک مرحله انتخابِ ویژگی‌های مؤثر نیز وجود دارد که موجب می‌شود مؤثرترین ویژگی‌ها از طریق الگوریتم ترکیبی فازی-راف و تحلیل سلسله‌مراتبی، انتخاب شوند.

شکل۴، میانگین، انحراف استاندارد و آزمون آنالیز واریانس[1] را برای نسبت‌های شرکت‌های متقلب و غیرمتقلب، گزارش می‌کند. آزمون‌های تک‌متغیّره به چندین متغیّر اشاره دارند که ممکن است در کشف شرکت‌های متقلب مفید باشند. از بین ۵۴ متغیر مورد آزمون، ۲۳ متغیر که در سطوح ۱ الی۵ درصد معنادار هستند، در شکل ۴ خلاصه شده‌اند. سایر متغیرها فاقد معناداری مناسب بوده‌اند.

ردیف	متغیر	شرکتهای متقلب		شرکتهای غیرمتقلب		آزمون ANOVA	
		میانگین	انحراف استاندارد	میانگین	انحراف استاندارد	آماره F	Prob
X3	بدهی ها به دارایی ها *	۰.۸۵۸۳۳۲	۰.۴۴۸۹۳	۰.۶۵۹۲۵۲	۰.۳۸۴۱۰۱	۲.۵۶۳۹۱۹	۰.۰۴۲۴
X5	نسبت جاری *	۱.۲۱۸۹۲۳	۰.۶۱۰۱۶۱	۱.۳۸۸۹۰۵	۰.۶۳۲۵۳۶	۲۸.۴۶۷۵۸	۰
X6	نسبت سریع *	۰.۶۰۱۶۰۸	۰.۹۸۰۰۳۶	۰.۳۳۷۹۱۷	۰.۴۶۲۸۶۵	۱۷.۵۰۵۷۹	۰
X17	لگاریتم طبیعی بهای تمام شده کالای فروش رفته *	۶.۰۴۴۳۴۹	۰.۷۳۶۵۰۶	۵.۹۷۱۵۴۲	۰.۵۷۳۶۵۲	۳.۷۲۵۲۲۵	۰.۰۱۱۴
X18	سود خالص به مجموع دارایی ها *	۰.۰۶۴۹۰۴	۰.۱۵۷۳۲۴	-۰.۰۴۲۳۲	۰.۲۱۹۸۸۳	۳.۵۶۱۹۹۹	۰.۰۰۰۵
X20	سود خالص به بهای تمام شده کالای فروش رفته *	۰.۰۱۶۷۲۵	۰.۳۸۰۵۴۹	-۰.۲۷۳۲۲	۰.۷۹۴۸۸۳	۴.۷۳۹۹۱۷	۰.۰۰۱۸
X21	بهای تمام شده کالای فروش رفته به مجموع دارایی ها *	۰.۹۸۶۳۰۲	۰.۷۷۰۶۵۲	۰.۷۴۰۷۴۷	۰.۴۷۴۹۶۳	۴.۰۴۶۹۳۱	۰.۰۴۵۸
X24	سود عملیاتی به فروش *	-۰.۰۹۴۹۹	۰.۵۰۹۱۰۱	۰.۰۸۰۸۵	۰.۲۵۴۷۰۲	۹.۰۸۳۳۳۶	۰.۰۰۳
X25	سود قبل از بهره و مالیات به فروش *	-۰.۲۶۸۴	۰.۷۹۸۵۲	۰.۰۳۰۱۱۷	۰.۳۸۸۷۶۴	۱۰.۸۸۶۷۴	۰.۰۰۱۲
X27	سود ناخالص به کل داراییها *	۰.۰۷۹۸۳۶	۰.۱۳۴۵۷۱	۰.۱۵۴۱۱	۰.۱۳۰۱۴۵	۱۰.۸۰۷۶۶	۰.۰۰۱۲
X28	سود قبل از بهره و مالیات به کل داراییها *	-۰.۰۳۹۲۲	۰.۲۳۳۲۵۲	۰.۰۷۷۴۶۵	۰.۱۷۰۲۱۶	۱۳.۴۶۲۸۶	۰.۰۰۰۳
X32	سود قبل از بهره و مالیات به بدهی های جاری *	۰.۰۸۷۴۵۵	۰.۴۸۴۹۶۹	۰.۲۵۰۵۲	۰.۴۶۸۷۸۶	۴.۰۱۳۹۴۴	۰.۰۴۶۶
X33	(دارایی جاری- موجودی کالا) به بدهی جاری *	۰.۷۱۱۹۲	۱.۱۶۴۹۳۳	۰.۳۹۵۱۱۳	۰.۴۷۶۷۹	۶۶۸۶۶۵۱	۰.۰۱۰۵
X34	موجودی کالا به بدهی جاری *	۰.۶۷۶۹۸۵	۰.۶۱۷۸۸۳	۰.۸۲۳۸۱۱	۰.۵۵۲۱۲۷	۴.۰۶۳۹۰۱	۰.۰۲۴۴
X35	وجه نقد به جمع بدهی ها *	۰.۰۸۹۷۵۷	۰.۲۰۶۸۰۸	۰.۰۸۹۶۶۱	۰.۲۰۵۵۴۴	۳.۷۱۴۹۴۹	۰.۰۱۸۸
X38	بدهی های جاری به جمع دارایی ها *	۰.۷۳۱۳۲۸	۰.۴۲۷۴۵	۰.۵۷۹۰۷۳	۰.۳۲۲۴۲۴	۶.۳۳۸۸۷۳	۰.۰۱۲۷
X39	سرمایه به جمع داراییها *	۰.۱۴۱۶۶۸	۰.۴۴۸۹۳	۰.۳۴۰۷۴۸	۰.۳۸۴۱۰۱	۸.۳۶۰۵۶۳	۰.۰۰۴۳
X42	موجودی کالا به فروش *	۰.۸۵۸۲۸۷	۰.۹۱۷۳۷۷	۰.۵۷۳۶۵۲	۰.۵۲۳۸۹۳	۶.۴۹۳۸۶۴	۰.۰۱۱۷
X43	حسابهای دریافتنی به فروش *	۰.۶۸۴۰۳۱	۰.۶۴۲۴۹۳	۰.۴۳۵۷۷۸	۰.۶۳۳۹۳۶	۵.۱۰۱۴۴۲	۰.۰۲۵۱
X44	فروش به دارایی ثابت *	۳.۹۴۵۹۸۵	۳.۶۶۵۴۱۸	۵.۴۴۶۸۹۲	۶.۰۵۷۳۱۹	۵.۷۸۷۱۰۸	۰.۰۰۲۱
X47	بهای تمام شده کالای فروش رفته به فروش *	۰.۸۹۰۱۴۸	۰.۱۸۴۶۵۴	۰.۸۲۴۶۶۳	۰.۱۵۳۰۹۷	۵.۴۹۶۰۵۷	۰.۰۲۰۲
X48	هزینه های عملیاتی به فروش *	۰.۰۸۰۸۴۶	۰.۳۴۵۱۷۷	۰.۰۱۲۳۴۷	۰.۰۴۳۰۵۲	۴.۹۸۷۰۰۸	۰.۰۲۶۸
X53	موجودی کالا به دارایی جاری *	۰.۵۴۷۱۲۴	۰.۲۷۷۷۳۱	۰.۶۹۲۹۹	۰.۳۹۵۷۴۹	۵.۲۴۲۷۴	۰.۰۲۳۲

شکل٤: میانگین، انحراف استاندارد و آزمون ANOVA را برای نسبت‌های شرکت‌های متقلب و غیرمتقلب.

1. ANOVA

تفاوت‌های عمده در میانگین ارزش‌ها بین شرکت‌های متقلّب و غیرمتقلب و معناداریِ آماریِ بالا ($p<0.000$) نشان می‌دهد که این نسبت‌ها با تقلب صورت‌های مالی مرتبط هستند.

مدل پیشنهادی دارای مراحل مختلفی به ترتیب زیر است:

۱. پیش‌پردازش داده‌ها؛

۲. انتقال داده‌ها؛

۳. انتخاب ویژگی‌های مؤثر با استفاده از الگوریتم ترکیبیِ مجموعهٔ راف و تحلیل سلسله‌مراتبی؛

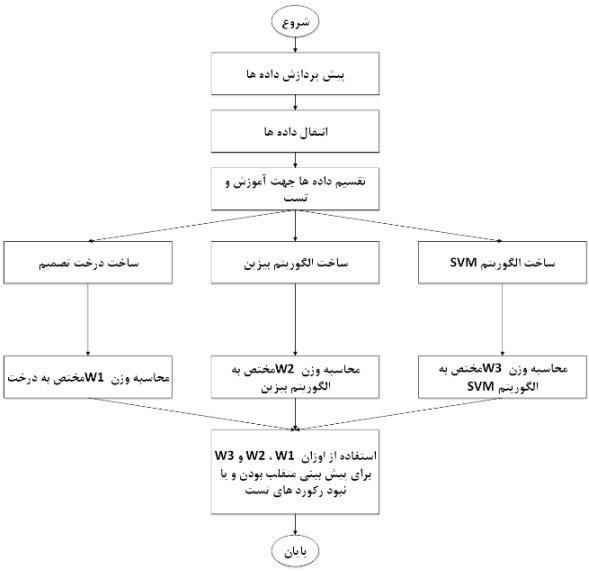
٤. آموزش و محاسبهٔ اوزان الگوریتم‌های درخت تصمیم، شبکه بیزین و ماشین بردار پشتیبان؛

٥. ساخت درخت تصمیم‌گیری؛

٦. تبدیل درخت تصمیم‌گیری و بهینه‌سازی آن؛

۷. آموزش و محاسبهٔ اوزان الگوریتم‌های درخت تصمیم، ماشین بردار پشتیبان و الگوریتم بیزین.

نمودار کلی روش تحقیق را می‌توان در شکل۵ مشاهده کرد.



شکل۵: نمودارِ کلیِ روش پیشنهادی

پیش‌پردازش داده‌ها

نخست، داده‌ها جمع‌آوری و سپس آماده‌سازی و پیش‌پردازش می‌شوند. در آماده‌سازی و پیش‌پردازش داده‌ها از روش‌های مختلفی استفاده می‌شود. نخست، 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. برای نمونه، می‌توان به ویژگیِ نام و نام خانوادگی اشاره کرد. همچنین، ممکن است برخی تراکنش‌ها دارای مقادیرِ مفقودِ فراوانی باشند. از این‌رو، این تراکنش‌ها نیز باید از مجموعهٔ داده‌ها حذف شوند. از طرفی، ممکن است مقادیر برخی ویژگی‌ها دارای مقادیر نویز و مفقود باشند. از این‌رو، این مقادیر نیز باید در مجموعهٔ داده اصلاح شوند. مرحلهٔ بعدی، استفاده از ابزار کشف آنومالی است. داده‌هایی که در نقاط خارج از قانون مجموعه داده قرار دارند، شناسایی و حذف می‌شوند. برای اینکه بتوان روی داده‌ها به‌عنوان ورودی کار کرد، باید ویژگی‌هایی را از آن‌ها استخراج کرد. به‌طور معمول، پیش از انتخاب و استخراج ویژگی‌ها، برخی عملیات پیش‌پردازش بر روی داده‌ها انجام می‌شود.

انتقال داده‌ها

در این قسمت، داده‌ها در دامنه‌های درست قرار می‌گیرند؛ بدین‌معنا که داده‌ها باید به رنج‌هایی که در سیستم مشخص شده است، منتقل شوند و داده‌های خارج از رنج، داده‌های مشکل‌دار بوده و باید حذف شوند. داده‌ها باید در محدودهٔ درست قرار بگیرند؛ برای مثال، اگر فیلدِ «سن» وجود داشته باشد، فردی که محدودهٔ سنّی بین ۵۵ تا ۷۰ دارد، باید در سیستم به‌صورت «خیلی پیر» ثبت شود که این قسمت به‌صورت خودکار از روی مجموعه داده‌ها تکمیل می‌شود.

انتخاب ویژگی‌های مؤثر با استفاده از مجموعه راف و تحلیل سلسله‌مراتبی

بسیاری از مفاهیم و تئوری‌های عدم قطعیت نظیر مجموعه‌های فازی، سیستم‌های

خاکستری و مجموعه‌های راف، در گذشته معرفی شده و در سال‌های اخیر، ابزارهای ریاضیِ مبتنی بر آن‌ها با سرعت بالایی توسعه یافته‌اند. هریک از این رویکردها، مفاهیم خاصّ خود را دارد و از ویژگی‌های منحصر به خود برخوردارند. به‌عنوان مثال، «نظریهٔ کلاسیک» به‌دنبال تحلیل داده‌های احتمالی یا قطعی است و «نظریهٔ فازی» محاسبات نرم را اساس کار خود قرار داده است. «نظریهٔ خاکستری» به کنترل سیستم‌ها در شرایط کمبود داده‌ها و اطلاعات ناکامل پرداخته و «نظریهٔ راف» تقریب و استدلال دربارهٔ داده‌ها را به‌دنبال دارد. داده‌هایی که از دنیای واقعی گرفته می‌شوند، معمولاً شامل تمامی انواع نویز بوده و عدم قطعیت بسیار و اطلاعات غیرکاملِ فراوانی را به‌همراه دارند. روش‌های سنتیِ برخورد با این عدم قطعیت ـ نظیر نظریهٔ فازی ، نظریهٔ گواه، نظریهٔ احتمالات و نظایر آن ـ به اطلاعات اضافی مانند توزیع احتمال و تابع عضویت نیازمندند. به بیان دیگر، کار با این سیستم‌ها به‌دلیل حجم بالای داده‌ها دشوار است. از این‌رو، به‌کارگیری سایر نظریه‌ها، نظیر نظریهٔ مجموعه‌های راف، می‌تواند در این راه کمک‌کننده باشد.

مجموعهٔ راف ابزاری قابل استفاده از شرایط ابهام و عدم قطعیت است که نخستین‌بار توسط پاولاک (۱۹۸۲) ارائه شد. «نظریهٔ راف» در زمینه‌های مختلفی مورد استفاده قرار می‌گیرد؛ از جمله، تجزیه‌وتحلیل تصمیم‌گیری، سیستم‌های پشتیبان تصمیم. بعد از آقای پاولاک، سه پژوهشگر دیگر به نام‌های ژای، خو و ژانگ در سال ۲۰۰۸ اعداد راف را ارائه کردند. یک عدد راف دارای حدّ پایین (L)، حدّ بالا (U) و حدّ میانی تشکیل شده است که به «فاصلهٔ مرزی راف» مشهور است. اعداد راف در مسائلی استفاده می‌شوند که نظرهای خبرگان در آن دخیل است و به‌نوعی سبب ایجاد عدم قطعیت و ابهام بشود.

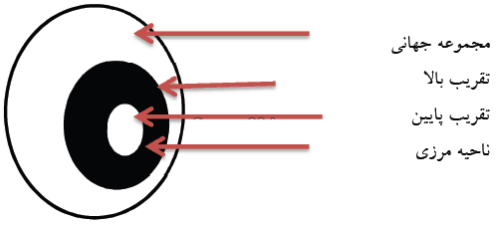
فرض کنید در یک مجموعه تصمیم‌گیری، مجموعه U شامل تمام اعضای مجموعه باشد. Y یک عضو دلخواه از مجموعه U و R یک مجموعه از t کلاس است که تمام اعضای U را پوشش می‌دهد. اگر این کلاس‌ها به‌صورت ترتیبی همانند $G1<G2<\ldots<Gt$ باشند، آنگاه حدهای پایین، بالا و ناحیهٔ مرزی، از کلاس G به‌صورت رابطه (۱) تعریف می‌شود.

$$\begin{aligned}
&\underline{Apr}(G_q) = \bigcup\{Y \in U \mid R(Y) \le G_q\} \\
&\overline{Apr}(G_{q)=} \bigcup\{Y \in U \mid R(Y) \ge G_q\} \\
&Bnd(G_q) = \bigcup\{Y \in U \mid R(Y) \ne G_q\} = \\
&\{Y \in U \mid R(Y) \succ G_q\} \cup \{Y \in U \mid R(Y) \prec G_q\}
\end{aligned} \qquad (۱)$$

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

$$\begin{aligned}
&\underline{Lim}(G_q) = \frac{1}{M_L}\sum R(Y) \mid Y \in \underline{Apr}(G_q) \qquad (۲) \\
&\overline{Lim}(G_q) = \frac{1}{M_L}\sum R(Y) \mid Y \in \overline{Apr}(G_q) \\
&RN(G_q) = \left[\underline{Lim}(G_q), \overline{Lim}(G_q)\right]
\end{aligned}$$

همچنین فاصله مرزی راف به‌صورت شکل۶ محاسبه می‌شود. این فاصلهٔ مرزی، ابهام را بیان می‌کند، به‌طوری‌که هر چقدر این عدد بزرگ‌تر باشد، نشان‌دهندهٔ ابهام بیشتر و اگر کوچک‌تر باشد، نشان‌دهندهٔ دقت بیشتر است.



شکل۶: محاسبه محدده در مجموعه راف

چون اعداد راف مشابه اعداد فاصله‌ای هستند، قوانین محاسباتی اعداد فاصله‌ای برای اعداد راف نیز یکسان است، که عبارتند از:

الف) ضرب یک عدد صحیح در یک عدد راف؛

ب) جمع دو عدد راف؛

ج) ضرب دو عدد راف.

در این پژوهش، از روش مجموعل راف ادغام‌شده با روش تحلیل سلسله‌مراتبی[1] که یک

1- AHP

روش تصمیم‌گیری است، وزن‌دهی پارامترها صورت گرفته و پارامترهای مؤثر شناسایی می‌شوند.

روش فرایند تحلیل سلسله‌مراتبی، از روش‌های پرکاربرد در تصمیم‌گیریِ چندمعیاره است که هدف آن، محاسبهٔ وزن معیارها و گزینه‌های پژوهش تحت یک مدل سلسله‌مراتبی است. در این مدل، ابتدا مقایسه‌های زوجی تشکیل و در اختیار خبرگان قرار داده می‌شود تا بر اساس طیف ۱ تا ۹، نظرهای خود را نسبت به مقایسهٔ دوبه‌دوی معیارها، بیان کنند. برای استفاده از اعداد راف در روش AHP (rough AHP) به طریق زیر عمل می‌شود:

ابتدا مقایسه‌های زوجی خبره‌ها را از نظر نرخ ناسازگاری بررسی می‌کنیم. چنانچه نرخ ناسازگاری کمتر از ۰.۱ باشد، یعنی مقایسهٔ زوجی سازگار است و در صورتی‌که بزرگ‌تر از ۰.۱ باشد، باید اعداد مقایسهٔ زوجی اصلاح شود.

ایجاد اعداد راف از اعداد خبره‌ها با استفاده از روابطی که در تئوری گفته شد.

محاسبهٔ وزن فاصله‌ایِ معیارها با استفاده از روش میانگین هندسی.

در این تحقیق نیز از خبرگانی استفاده شده است که برای انتخاب ویژگی‌های مؤثر کمک کردند و در این حالت، بهترین نسبت‌های مالی انتخاب شدند.

در نهایت، در این قسمت تعداد ۲۳ نسبت مالی بررسی می‌شود و در صورت امکان، تعداد آن‌ها کاهش پیدا می‌کند که در این راستا، سربار محاسباتی و همچنین نویز نیز در صورت ممکن کاهش پیدا می‌کند. این فرایند در شکل۸ نشان داده شده است.

به‌منظور سهولت در انجام عملیات، هر نسبت با استفاده از Att و یک عدد مشخص شده که در شکل۷ نشان داده شده است.

مجموع بدهی‌ها / مجموع دارایی‌ها	Att1
نسبت جاری = دارایی‌های جاری/ بدهی‌های جاری	Att2
نسبت آنی = دارایی‌های جاری ـ (موجودی کالا+پیش پرداخت) / بدهی های جاری	Att3
لگاریتم طبیعی (بهای تمام‌شده کالای فروش‌رفته)	Att4
سود خالص/ مجموع دارایی‌ها	Att5
سود خالص/ فروش	Att6
فروش/ مجموع دارایی‌ها	Att7

سود عملیاتی/ فروش	Att8
سود قبل از بهره و مالیات/ فروش	Att9
سود ناخالص/ کلّ دارایی‌ها	Att10
سود قبل از بهره و مالیات/ کلّ دارایی‌ها	Att11
سود قبل از بهره و مالیات / بدهی‌های جاری	Att12
بدهی جاری/ (دارایی جاری- موجودی کالا)	Att13
موجودی کالا / بدهی جاری	Att14
وجه نقد / جمع بدهی‌ها	Att15
بدهی‌های جاری / جمع دارایی‌ها	Att16
سرمایه / جمع دارایی‌ها	Att17
موجودی کالا / فروش	Att18
حساب‌های دریافتنی/ فروش	Att19
فروش/ دارایی ثابت	Att20
بهای تمام شده کالای فروش رفته / فروش	Att21
هزینه‌های عملیاتی/ فروش	Att22
موجودی کالا/ دارایی جاری	Att23

شکل (۷): معماری استفاده‌شده در محاسبهٔ وزن

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

شکل۸: ترتیب تأثیر ویژگی‌های نسبت‌های مالی استفاده شده

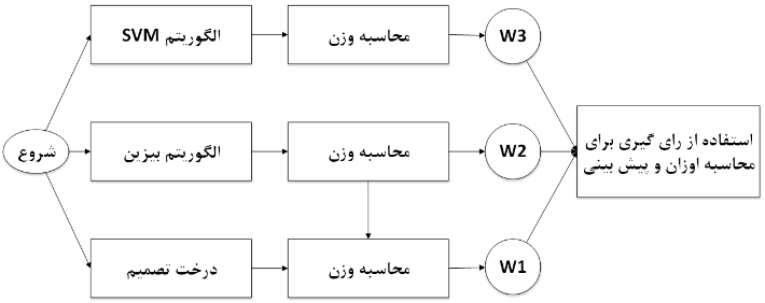
در روش پیشنهادی در ابتدا، کار انتخاب ویژگی صورت می‌گیرد و ویژگی‌های مؤثرتر در روش پیشنهادی انتخاب می‌شوند که برای مجموعه داده‌های ورودی می‌توان در شکل(۸) تأثیر ویژگی‌ها را مشاهده کرد که مرتب‌سازی شده است.

با توجه شکل (۸) 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 کلّ دارایی‌ها، سود خالص به مجموع دارایی‌ها، سود عملیاتی به فروش و سود ناخالص به کلّ دارایی‌ها بیشترین تأثیر را در متقلب بودن و یا نبودن شرکت دارد. در اینجا می‌توان نسبت‌های مالی بی‌تأثیر را حذف کرد تا الگوریتم قادر باشد با سرعت بیشتری داده‌کاوی را انجام دهد. همین‌طور، در همین قسمت می‌توان دریافت که چه نسبت‌هایی بیشترین تأثیر را در متقلب بودن و یا نبودن شرکت دارد و چه عواملی سبب سوق‌دادنِ شرکت به سمت تقلب می‌شود؛ یعنی می‌توان مؤلفه‌های تأثیرگذار را در اینجا شناسایی کرد و به آن‌ها توجه بیشتری داشت تا اینکه بتوان شرکت‌های متقلب را بهتر شناسایی کرد و دربارهٔ آن‌ها تصمیم‌های مهمی گرفت.

آموزش و محاسبهٔ اوزان الگوریتم‌های درخت تصمیم، ماشین بردار پشتیبان و شبکهٔ بیزین

در ابتدای کار، درصدی از مجموعه داده‌های مورد استفاده برای محاسبهٔ آموزش و وزن، استفاده می‌شود. در این قسمت، استفاده از یک الگوریتم ترکیبی مبنای نظر است که از دو الگوریتمِ درخت تصمیم و ماشین بردار پشتیبان، استفاده می‌کند. این دو الگوریتم هر کدام سهمی از جواب نهایی را خواهند داشت که بدین شکل دقت سیستم افزایش می یابد. می‌توان

نمایی از این مرحله را در شکل ۹ مشاهده کرد.



شکل۹: معماری استفاده‌شده در محاسبهٔ وزن

همان‌طور که در شکل۹ مشاهده می‌شود، الگوریتم پیشنهادی در ابتدا با استفاده از یک مجموعه داده به محاسبهٔ وزن می‌پردازد. این محاسبه بدین ترتیب است که هر الگوریتم با استفاده از ۷۰ درصد مجموعه داده‌های موجود، آموزش می‌بیند و با استفاده از ۳۰ درصد باقی‌مانده مورد آزمون قرار می‌گیرد. در نهایت، با توجه به تعداد جواب‌های صحیح، امتیاز و یا وزنی به آن تعلق می‌گیرد تا اینکه با استفاده از وزن تعلق توان در مرحلهٔ بعدی، وزنی را برای خروجی هر الگوریتم بتوان در نظر گرفت. همان‌طور که در معماری نیز قابل مشاهده است، پس از محاسبهٔ وزن که به‌صورت تقسیم تعداد جواب‌های درست به تعداد کلّ جواب‌های حدس زده‌شده است، می‌توان میزان تأثیرگذاری هر کدام از این الگوریتم‌ها را در خروجی نهایی، بهتر تشخیص داد. در این روش، پس از محاسبهٔ وزن‌ها، به ازای هر رکورد، پیش‌بینی‌ای به‌وسیلهٔ ماشین بردار پشتیبان، شبکه بیزین و درخت تصمیم بهینه صورت می‌گیرد. 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 بدین‌صورت، نتیجهٔ نهایی به‌دست می‌آید و دسته‌بندی درست صورت می‌گیرد. در واقع، در این قسمت، از روش رأی‌گیری[1] استفاده می‌شود.

1.Voting

ساخت درخت تصمیم‌گیری

در این قسمت، از درخت تصمیم‌گیری استفاده می‌شود. در درخت تصمیم‌گیری، هر شاخه یک انتخاب است؛ بدین‌معنا، برای رفتن از گرهِ ریشه به گرهِ پایین‌تر، می‌توان از شاخه هایی که به آن گره متصل هستند، یکی را انتخاب کرد. در انتها، هر یک از گره‌های انتهایی یا به اصطلاح گرهِ برگ، تصمیمی را بازگو می‌کند. هر کدام از شاخه‌ها تا رسیدن به برگ، دارای سناریویی است که سبب اتخاذ یک تصمیم می‌شود. در این پژوهش، از مدل پیشنهادی مبتنی بر درخت تصمیم ID3 بهبودیافته استفاده شده است. این بهبود سبب سرعت عمل بالای آن شده است. درخت ID3 یک درخت تصمیم‌گیری است که دارای یادگیری نیز هست و نخستین‌بار توسط «راس کوینلن» مطرح شد. ایدهٔ الگوریتم ID3، ساخت درخت تصمیم‌گیری بالا به پایین است که انتخاب گره در آن به‌وسیلهٔ جستجوی حریصانه از میان مجموعه‌ای از صفت‌هاست. در اینجا، برای اینکه قادر باشیم تا مفیدترین صفت را از میان صفات بیابیم که در کلاسه‌بندی مفیدتر باشد از الگویی بخصوص استفاده کرده‌ایم. برای اینکه بتوان کلاسه‌بندیِ مفیدی را برای مجموعهٔ یادگیری انجام داد، باید تعداد سؤال‌ها را کاهش داد. یا می‌توان گفت باید عمق درخت تصمیم‌گیری را کاهش داد. از این‌رو، در این قسمت به تابعی نیاز است که قادر باشد متعادل‌ترین تقسیم را انجام دهد. 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 است. 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 می‌گویند. توابع زیادی همچون آنتروپی، شاخص جینی[1] و خطای طبقه‌بندی[2] برای سنجش میزان همگن‌پذیری وجود دارند. در این میان، در اینجا از آنتروپی استفاده شده است.

$$\text{Entropy} = \sum_{j} - \mathrm{p}_j \log_2 p_j \qquad (۳)$$

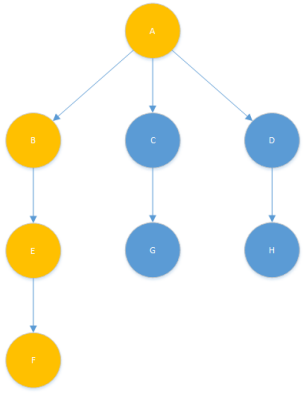
1. Gini index
2. classification error

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. هرچه مجموعه منظم‌تر و دارای گوناگونی کمتری باشد، آنتروپی و بی‌نظمی آن کمتر است و برعکس. البته، در اینجا چون ما در مرحلهٔ قبل یک کلاسه‌بندی ابتدایی را انجام دادیم، تقریبا بی‌نظمی نیز پایین است و این خود سبب سرعت عمل بالاتر روش پیشنهادی ما می‌شود، زیرا این قضیه باعث می‌شود عمق درخت تصمیم‌گیری کم شود و هر چه عمق این درخت کمتر باشد، سرعت تصمیم‌گیری نیز بیشتر می‌شود.

برای اینکه بتوانیم صفتی را در درخت تصمیم‌گیری انتخاب کنیم که در رتبهٔ بالاتری از بقیهٔ صفت‌ها و دارای اهمیت بالاتری از بقیه صفت‌ها باشد، از فرمول (۳) استفاده کردیم. با توجه به این فرمول، آنتروپیِ همهٔ صفات را در مجموعهٔ S محاسبه و مقدار صفت مجموعه A را از آن کم می‌کنیم. مجموعه A، مجموعهٔ همهٔ صفات انتخاب‌شده از پدر تا به اینجا در یک مسیر خاص است.

$$G(\mathrm{S},\mathrm{A}) = Entropy(S) - \sum_{v \in Values(A)} Entropy(\mathrm{v}) \qquad (۴)$$

برای درک بهتر فرمول (۴) می توان شکل ۱۰ را مشاهده کرد.



شکل ۱۰: مثالی برای انتخاب صفت‌ها

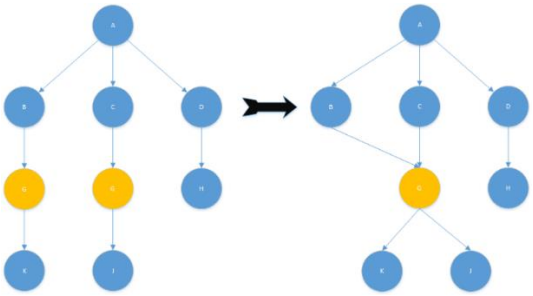
چنان‌که در شکل ۱۰ قابل مشاهده است، ما باید آنتروپیِ تمامی صفات را از آنتروپیِ صفاتِ انتخاب‌شده تا به اینجای مسیر بکاهیم (یعنی باید G(S,F)=E(S)-(E(A)+E(B)+E(E)+E(F)) را به‌دست آورد). البته، باید توجه داشت که باید در مجموعه A صفاتی را که تا به اینجای کار استفاده شده است همراه با صفتی که می‌خواهیم قرار دهیم، محاسبه کنیم. سپس، از بین این مجموعه صفات باقی‌مانده که برای هر کدام فرمول (۴) را محاسبه کردیم، صفتی را که دارای G بیشتری است، انتخاب کنیم. در این حالت، اگر دو صفت دارای G برابر بودند ـ که احتمال این پیشامد نیز کم نیست ـ باید به گره دو یا هر تعداد صفت که دارای بیشترین مقدار G و با هم برابر هستند، به گره مربوط 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. سپس، روند الگوریتم را برای هر یک از این گره‌ها ادامه می‌دهیم. برای مثال، در شکل ۱۰ می‌توان مشاهده کرد مقدار G برای صفت‌های B، C و D برابر است. از این‌رو، همهٔ این صفت‌ها در یک سطح قرار گرفته‌اند.

این کار سبب می‌شود صفت‌های دارای آنتروپی بیشتر را بیابیم، زیرا این صفت‌ها تأثیر بیشتری را در تصمیم نهایی ما می‌گذارند. این روندِ جلو رفتن در درخت تصمیم‌گیری تا جایی ادامه می‌یابد که در هر مسیر دیگر، صفتی باقی نمانده باشد. در این حالت، درخت تصمیم‌گیری کاملا ساخته شده و به پایان رسیده است.

تبدیل درخت تصمیم‌گیری

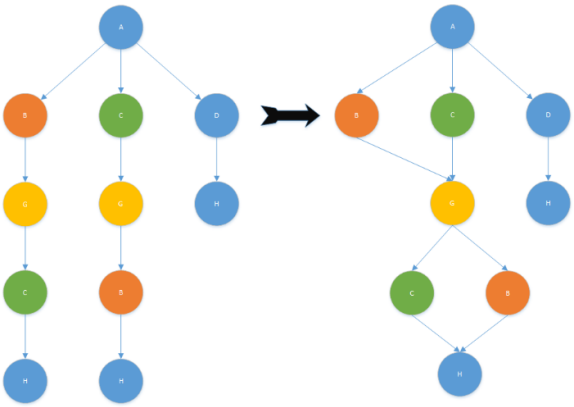
برای استخراج شرط‌هایی از این درخت، باید به‌صورت If…then…. شود که در آن بتوان از قوانین انجمنی نیز کمک گرفت. ما درخت را تبدیل کردیم که در موارد بسیاری به گراف تبدیل و از حالت درخت خارج می‌شود. البته، اگر این درخت از نظر ظاهری به گراف تبدیل شود، برای ما همچنان به‌صورت درخت در نظر گرفته می‌شود. بنابراین، می‌توان گفت چیزی مابین گراف و درخت به‌دست می‌آید.

در این قسمت، در هر سطح گره‌های همنام را با هم ادغام می‌کنیم و فرزندان آن‌ها به این گرهِ ادغام‌شده افزوده می‌شوند. در شکل۱۱ این کار را می‌توان مشاهده و بهتر آن را درک کرد.



شکل۱۱: مثالی از تبدیل درخت تصمیم‌گیری

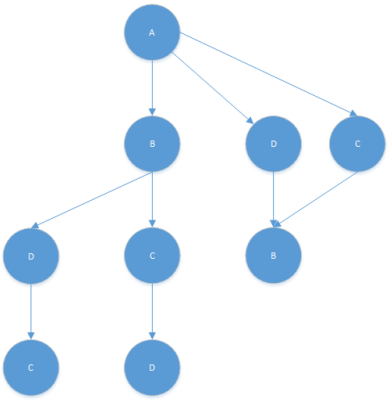
همان‌طور که در شکل۱۱ می‌توان مشاهده کرد، گره‌های B و C دارای فرزند مشترک G بودند. این دو گره G در یک سطح نیز قرار داشتند. در این حالت، این دو گره به یک گره تبدیل و فرزندان آن‌ها نیز به گرهِ جدید افزوده می‌شود. ممکن است شرایط مانند شکل۱۲ پیش آید که در آن، دو گره G با هم ادغام شدند؛ ولی حالتی است که در آن گره G اول دارای فرزند C و گره G دوم دارای فرزند B است و این گره‌ها هر کدام برای گره G مقابل در پیشینیان آن، بازدید شده است. در این حالت نیز مشکلی پیش نمی‌آید و تنها می‌توان گفت در هنگام نوشتن شرط می‌توان در نظر گرفت که با داشتن شرط B و C، همراه با And منطقی آن با G آن مسیر پیموده می‌شود؛ یعنی A And (B OR C) AND G.



شکل۱۲: مثالی از حالت خاص تبدیل

در اینجا، می‌رسیم به این قسمت که هر کدام از این صفت‌ها در هر یک از داده‌ها می‌تواند داری مقدار متفاوتی باشد. برای مثال، در شکل۱۳، G می‌تواند دارای دو مقدار true و یا false باشد. برای این قسمت، داده‌ای را انتخاب می‌کنیم که فراوانی آن در صفت مربوطه در کلّ داده‌ها بیشتر است. برای مثال، فراوانی true بیشتر است یا false. در بین صفت داده‌های از نوع G آنگاه آن را برمی‌گزینیم. در صورتی که فراوانی دو مقدار برابر باشد، در تصمیم، or را برای این دو مقدار قید می‌کنیم. برای مثال، G=true or G=false.

استخراج تصمیمات از گراف تصمیم‌گیری که در اینجا ساخته شده است. برای این کار، از گره ریشه شروع کرده و به‌سمت هر یک از برگ‌ها که برویم، یک تصمیم شکل می‌گیرد.



شکل۱۳: مثالی از گراف تصمیم روش پیشنهادی

ماشین بردار پشتبان در مدل پیشنهادی

در این مرحله، در ابتدای کار نرمال‌سازی انجام می‌شود. پس از استخراج نتایج از نرمال‌سازی، قسمتی از این داده‌ها به‌عنوان داده‌های آموزش استفاده و مدل ماشین بردار ایجاد می‌شود. سپس، با استفاده از داده‌های تست وزن، این الگوریتم محاسبه می‌شود تا بتوان در ادامه، میزان تأثیرگذاری این قسمت از الگوریتم را محاسبه کرد.

شبکهٔ بیزین در مدل پیشنهادی

در این مرحله نیز ابتدا نرمال‌سازی انجام می‌شود و پس از استخراج نتایج از آن، قسمتی از این داده‌ها به‌عنوان داده‌های آموزش استفاده و مدل شبکه بیزین ایجاد می‌شود. سپس، با استفاده از داده‌های تست وزن، این الگوریتم محاسبه می‌شود تا بتوان در ادامه، میزان تأثیرگذاری این قسمت از الگوریتم را محاسبه کرد.

ارزیابی مدل پیشنهادی

در این بخش، روش پیشنهادی که در فصل قبل بیان شد، بررسی و روش ارائه‌شده، با الگوریتم‌های معروف به نام ID3، الگوریتم بردار پشتیبان بهبود داده شده (اوییوآ[1]،۲۰۱۹) که مقالهٔ اصلیِ راهکار پیشنهادی است، و شبکهٔ بیزین که اوئویا و همکارانش در سال ۲۰۱۹ در جهت کشف تقلب در صورت‌های مالی مبتنی بر شبکه بیزین ارائه دادند، مقایسه می‌شود.

نخست، داده‌ها به‌وسیلهٔ برنامه به فرمت مناسب برای تحلیل تبدیل می‌شود و پیش‌پردازش ابتدایی صورت می‌گیرد. سپس فایلی با فرمت ARFF ایجاد می‌شود که ساختاری مناسب و استاندارد برای تحلیل است.

پیاده‌سازی در برنامهٔ Visual Studio 2017 و با زبان برنامه نویسی C# صورت گرفته است و در حین کار از کتابخانه‌های weka و zedgraph کمک گرفته شده است.

سیستم مورد استفاده در اینجا دارای سیستم عامل Windows 10 ، دارای ۶ گیگ RAM و Corei7 است.

1. Eweoya

در این پژوهش، الگوریتم پیشنهادی با الگوریتم‌های ID3، SVM و بیزین مقایسه شده است. الگوریتم‌های ID3 و SVM الگوریتم‌های پایهٔ روش پیشنهادی هستند. این روش پیشنهادی از ترکیبِ این دو روشِ ایجادشده است. همچنین، روش با یکی از الگوریتم‌های معروف به‌نام «شبکه بیزین» نیز مقایسه شده است. در ادامه می‌توان نتایج پیاده‌سازی را برای این الگوریتم‌ها مشاهده کرد.

```
--------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
```

شکل۱٤: خروجی مربوط به الگوریتم شبکه بیزین مقاله [۱٦]

```
--------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

شکل۱۵: خروجی مربوط به الگوریتم روش 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

شکل۱۶: 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
```

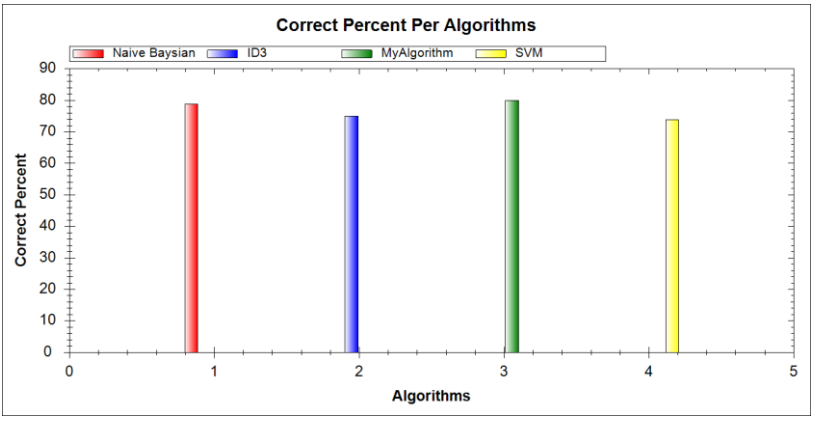
شکل۱۷: خروجی مربوط به الگوریتم SVM مقاله [۱۵]

با توجه به نتایج دریافتی، به‌وضوح قابل مشاهده است که الگوریتم پیشنهادی با ۸۰ ٪ میزان دقت، دارای بالاترین دقت صحّت و با ۲۰ ٪ اشتباه، دارای کمترین میزان اشتباه است. در جدول۱، الگوریتم پیشنهادی با سایر الگوریتم‌ها مقایسه شده است. مشخص است که الگوریتم پیشنهادی از سایر الگوریتم‌ها بهتر است.

جدول۱: نسبت درصد صحّت پیش‌بینی‌ها و خطای پیش‌بینی‌ها

الگوریتم‌ها	درصد صحت پیش‌بینی‌ها	درصد خطای پیش‌بینی‌ها
الگوریتم پیشنهادی	۸۰٪	۲۰٪
الگوریتم بیزین	۷۸.۸۸٪	۲۱.۱۱٪
الگوریتم ID3	۷۵٪	۲۵٪
الگوریتمsvm	۷۳.۸۸٪	۲۶.۱۱٪

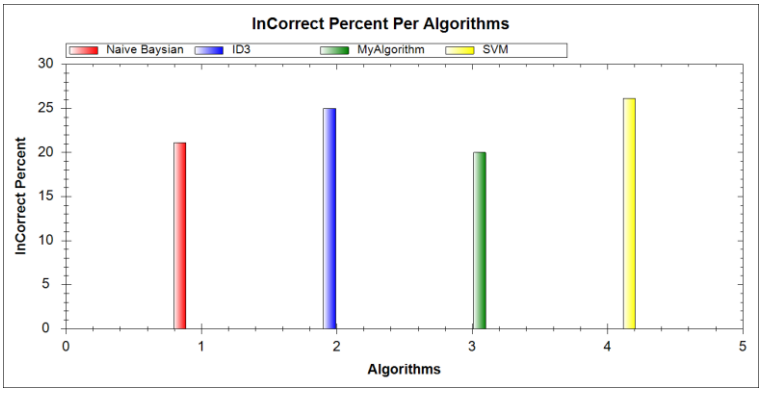
در ادامه، نمودارهای به‌دست‌آمده از برنامه، بررسی شده است. اولین نمودار، به درصد پیش‌بینی صحیح در میان داده‌های آزمون مربوط است که در شکل۱۸ قابل مشاهده است.



شکل۱۸: درصد پیش‌بینیِ صحیح در میان داده‌های آزمون برای الگوریتم پیشنهادی و دیگر الگوریتم‌های مورد بررسی

همان‌طور که می‌توان از این نمودار دریافت، روش پیشنهادی دارای دقتِ پیش‌بینیِ صحیحِ بیشتری نسبت به دیگر الگوریتم‌های مورد بررسی است؛ یعنی شبکه بیزین و الگوریتم بردار پشتیبان در مقاله (اوییوآ[1]، ۲۰۱۹) و الگوریتم درخت تصمیم. این بدین دلیل است که در روش پیش‌بینی، تنها مواردی از داده‌های آزمون در نظر گرفته شده است که تأثیر بیشتری را در نتیجهٔ خروجی داشتند. در نتیجه، داده‌هایی که در خروجی تأثیر نداشتند، استفاده نشده‌اند. بدین‌شکل، زمانِ تحلیل بسیار کاهش پیدا کرده است؛ حال آنکه الگوریتم‌های دیگر به‌دلیل استفاده از تمامی پارامترها دارای دقت کمتری هستند، زیرا ممکن است بعضی از پارامترها دارای مقادیر دوری باشند که در نتیجهٔ خروجی تأثیری نداشته باشند؛ ولی چون در الگوریتم‌های دیگر در ساخت مدل برای پیش‌بینی این پارامترها مورد استفاده قرار گرفته‌اند، سبب ایجاد نویز و کاهش دقت می‌شوند. در روش پیشنهادی، چون این پارامترهای بی‌فایده وجود نداشتند، دقت در روش پیشنهادی می‌تواند افزایش یابد.

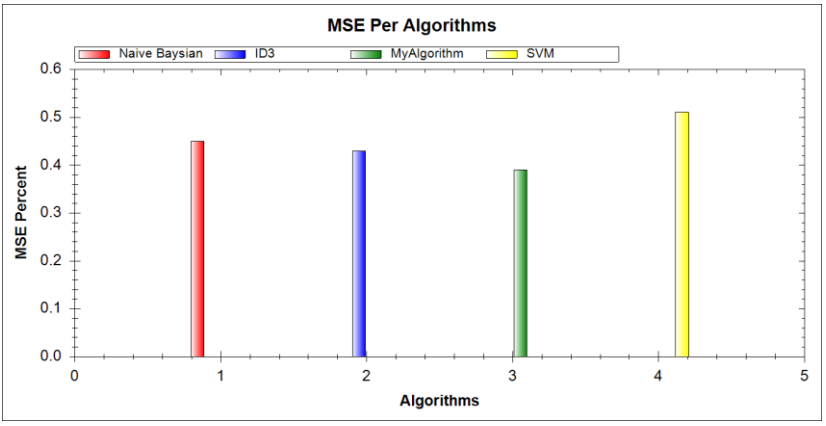
در نمودار ارائه‌شده در شکل۱۹، می‌توان درصد پیش‌بینیِ غلط را مشاهده کرد. با توجه به این نمودار، می‌توان درک کرد روش پیشنهادی به همان دلیل که پیش‌تر دربارهٔ درصد پیش‌بینیِ درست گفته شد، از روش‌های دیگر دارای مقدار کمتری است؛ یعنی پیش‌بینیِ اشتباهِ کمتری دارد. بنابراین، روش پیشنهادی بهتر از سایر روش‌ها عمل کرده است.



شکل۱۹: درصد پیش‌بینیِ ناصحیح در میان داده‌های آزمون برای الگوریتم پیشنهادی و دیگر الگوریتم‌های مورد بررسی

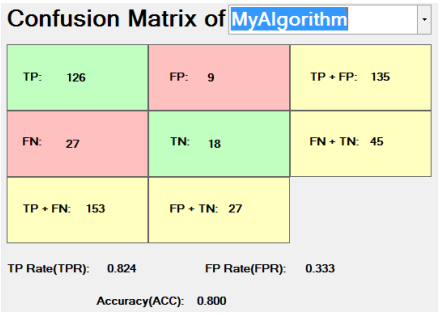
1. Eweoya

با توجه به نمودار شکل۱۹، می‌توان به این قضیه پی‌برد که خطای پیش‌بینی تقریباً در الگوریتم‌های مورد بررسی، برابر و نزدیک است ولی الگوریتم پیشنهادی کمتر است، زیرا هرچه نرخ صحت بیشتر باشد، نرخ غلط پایین‌تر خواهد بود و این نشان‌دهندهٔ عملکرد مناسب روش پیشنهادی است. در ادامه، در شکل۲۰ می‌توان مشاهده کرد که نرخ خطای میانگین(MSE) در روش پیشنهادی، از تمامی روش‌های دیگر کمتر است.

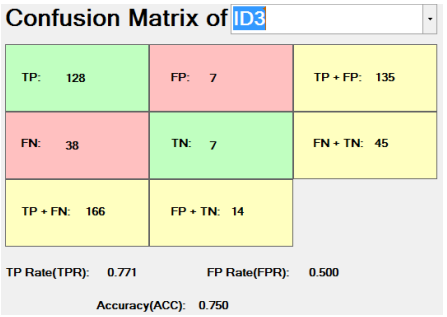


شکل۲۰: میزان معیار MSE در میان الگوریتم پیشنهادی و دیگر الگوریتم‌های مشابه مورد بررسی

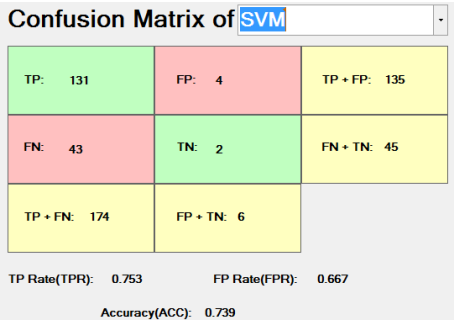
در ادامه Confusion Matrix مربوط به روش پیشنهادی و دیگر روش‌های مورد بررسی در این تحقیق، نشان داده شده است.



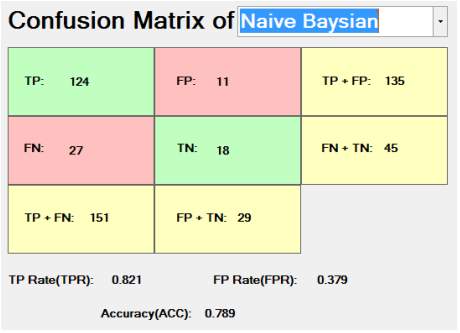
شکل۲۱: جدول Confusion matrix روش پیشنهادی



شکل۲۲: جدول Confusion matrix روش ID3



شکل۲۳: جدول Confusion matrix روش SVM مقاله [۱۵]



شکل۲٤: جدول Confusion matrix روش بیزین مقاله [۱٦]

می‌توان مشاهده کرد که روش پیشنهادی دارای دقت بالاتری است، زیرا در این جدول دارای مقدار TP و TN بیشتری از دیگر الگوریتم‌هاست و در کنار آن نیز دارای FP و FN کمتری از دیگر الگوریتم‌های مورد بررسی است. هرچه یک الگوریتم دارای TP و TN

بیشتری باشد، یعنی متقلب بودن و یا نبودن‌های در مجموعه داده‌های تست را به مقدار بیشتری درست تشخیص داده است و FP و FN نشان‌دهندهٔ عکسِ این قضیه، یعنی پیش‌بینی اشتباه است.

بحث و نتیجه‌گیری

در این پژوهش، راهکاری برای ارزیابی و پیش‌بینی تقلب‌های مالی شرکت‌ها ارائه و مشاهده شد که روش ارائه‌شده، عملکرد مناسبی داشته و بهبود نسبتاً بالایی را نسبت به الگوریتم‌های پایهٔ خود، یعنی بردار پشتیبان و درخت تصمیم، نشان داده است. روش پیشنهادی نسبت به الگوریتم درخت تصمیم، ۶.۶۶ درصد بهبود و نسبت به بردار پشتیبان دارای ۸.۲۷ درصد بهبود عملکرد داشته است. همچنین، کار در ادامه با الگوریتم شبکهٔ بیزین نیز بررسی و مشاهده شد که روش پیشنهادی بسیار بهتر از الگوریتم بیزین عمل می‌کند و دارای دقت بالاتر و نرخ خطای کمتری است. 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 نیست. داده‌های استفاده شده در این پژوهش، مربوط به ۶۰ شرکت در طی ۳ سال است؛ یعنی داده‌های مورد بررسی دارای ۱۸۰ رکورد بود. در اینجا، داده‌ها ابتدا پیش‌پردازش شد و انتقالاتی نیز روی داده‌ها صورت گرفت تا اینکه داده‌ها به داده‌های ورودیِ مورد نیاز الگوریتمِ پیشنهادی تبدیل شوند. نتایج به‌دست‌آمده کاملاً بهبود روش پیشنهادی را نشان می‌دهد. ازاین رو، با توجه به نتایج دریافتیِ مدلِ ارائه‌شده با ۸۰ ٪ دقت دارای بالاترین دقت صحت و با ۲۰ ٪ اشتباه دارای کمترین میزان اشتباه است که می‌توان آن را برای پیش‌بینی تقلب و یا به‌عنوان نمایندهٔ تقلب در بررسی‌ها و تحقیقات مختلف پیشنهاد کرد. در این روش پیشنهادی، از آنتروپی استفاده شده است، ولی می‌توان از روش‌های دیگری نیز استفاده یا آن را با روش‌های دیگری ادغام کرد. برای مثال، در صورتی که این روش را با روشی مانند Gain که تابع ارزش است، ادغام کنیم؛ به احتمال زیاد دارای عملکرد

بهتری است زیرا آنتروپی نیز دارای معایبی است، ولی سرعت آن بالاست. در این پژوهش به دنبال روشی بودیم که سرعت بالایی داشته باشد، ولی می‌توان با ادغام این روش و یا روش‌های جایگزین، دقت این روش ارائه‌شده را به‌شدت افزایش داد. می‌توان از روش پیشنهادی با بهبود در الگوریتم C4.5 نیز استفاده کرد؛ یعنی روش پیشنهادی را روی C4.5 با بهبودی مشابه بهبودی که روی ID3 در این پژوهش انجام شد، انجام داد تا عملکرد آن افزایش یابد. البته، نمی‌توان به‌طور قطع گفت که عملکرد آن بهتر می‌شود، بلکه باید این روش آزمایش و صحّت عملکرد آن بررسی شود.

منابع

خانجانی، رضا و سعید، پارسا. (۱۴۰۰). مبهم‌سازی نرم افزار با استفاده از تحلیل سلسله مراتبی و شبکه‌های پتری. *فصلنامه نوآوری‌های فناوری اطلاعات و ارتباطات کاربردی*. ۴ (۳۳)، ۱۲۲-۱۳۴.

داغمه‌چی فیروزجایی، مهدی. (۱۳۸۹). *استفاده از اطلاعات مکانی مشترک در تولید کلید رمز شبکه GSM، مدل مطالعاتی الگوریتم Geo-encryption* ، کنفرانس ملی امنیت اطلاعات و ارتباطات، جهاد دانشگاهی استان خوزستان،اهواز.

رضایی، ذبیح اله، رحمانی، علی و منتی، وحید. (۱۴۰۱). *تقلب در صورت‌های مالی، کشف و پیشگیری*. تهران: انتشارات دانشگاه الزهرا.

فراقان‌دوست حقیقی، کامبیز و برواری، فرید. (۱۳۸۸). بررسی کاربرد روش‌های تحلیلی در ارزیابی ریسک تحریف، *دانش و پژوهش حسابداری*، ۳ (۱۶)، ۷۹-۹۱.

Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. *Springer Plus*, 1(5), 89-101.

Chi, D. J., Chu, C. C., & Chen, D. (2019). *Applying support vector machine, C5. 0, and CHAID to the detection of financial statements frauds in Intelligent Computing Methodologies,* 15th International Conference, ICIC 2019, Nanchang, China.

Eweoya, I. O., Adebiyi, A., Azeta, A., Chidozie, F., Agono, F. O., & Guembe, B. (2019). A Naive Bayes approach to fraud prediction in loan default. *Journal of Physics: Conference Series*, 1299(1).

Lookman, S., Selmin, N. (2019). *A Framework for Occupational Fraud Detection by Social Network Analysis,* 27th International Conferences on Advanced Information Systems Engineering, Stockholm, Sweden.

Nahri Aghdam Ghalejoogh, J., Rezaei, N., Aghdam Mazarae, Y., & Abdi, R. (2023). Detecting financial fraud using machine learning techniques. *International Journal of Nonlinear Analysis and Applications*, 15(1), 199-214.

Paul, A., Free, C., & Scard, B. (2020). Pathways to accountant fraud: Australian evidenceand analysis. *Accounting Research Journal,* 28(1), 10-44.

Onuwa, O. B. (2014). Fuzzy Expert System for Malaria Diagnosis. *Oriental Journal of Computer Science & Technology*, 7, (2), 273-284.

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*. 60, 65-85.

Xu, B., Wang, Y., Liao, X., & Wang, K. (2023). Efficient fraud detection using deep boosting decision trees. *Decision Support Systems*, 157, 114037.

Wang, H., Mao, Ch., He, H., Zhao, M., Jaakkola, T. S., & Katabi, D. (2019). *Bidirectional Inference Networks: A Class of Deep Bayesian Networks for Health Profiling*. The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19). Hawaii, United States of America.