Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

ISSN: 2783-2538

Research Article

IDCOST: A Method for Increasing Data Criterion Service by Scoring Credit Imbalanced Data Using Applied SVM

Arash Ghorbannia Delavar [a,*], **Sadaf Sadat Ziya** [a]

[a] *Department of Computer Engineering and Information Technology, Payam Noor University, Tehran, Iran*

PAPER INFO

Paper history:

Received: 2023-07-13
Revised: 2024-08-01
Accepted: 2025-01-01

Keywords:

Data criterion;
Credit imbalanced data;
Cluster head;
Scoring;
Load balancing.

ABSTRACT

Unbalanced credit data can pose significant challenges in applied data mining. To address this, we propose a method that utilizes a scoring technique and support vector machine (SVM) to enhance data criterion service. Our approach integrates index feature selection and IDCOST method, which reduces data redundancy and balances feature selection data sets with a valid index. We also use feature selection and kernel modification to improve accuracy while reducing computational complexity and execution time. Our proposed method can detect credit card fraud and credit card default data sets with higher sensitivity than other methods. It presents a promising solution for tackling credit data issues in applied SVM data mining and has the potential to improve data analysis accuracy and reduce computational complexity in various fields. The IDCOST method is presented in pre-processing, training, validation, and testing stages. We use detector threshold clustering in the pre-processing stage, sensitivity and feature validation on the models in the training stage, and score each sample in the test dataset in the testing stage. The proposed method's accuracy is optimized by selecting an appropriate cluster head in data classification and employing a scoring technique. In conclusion, our proposed method is an effective solution for tackling credit data issues in applied SVM data mining. By integrating index feature selection, IDCOST method, feature selection, and kernel modification, we can accurately detect credit card fraud and credit card default data sets while reducing data redundancy and computational complexity.

DOI: https://doi.org/10.22075/jme.2025.31252.2493

*** Corresponding author.**
E-mail address: : a_ghorbannia@pnu.ac.ir

How to cite this article:
A. Ghorbannia Delavar and S. Ziya, "IDCOST: A method for increasing data criterion service by scoring credit imbalanced data using applied SVM," Journal of Modeling in Engineering, 23 Special Issue 81 (2025): 1-18, doi: 10.22075/jme.2025.31252.2493

مقاله پژوهشی

روشی برای افزایش سرویس معیار داده‌ها با امتیازدهی داده‌های نامتعادل اعتباری با استفاده از ماشین بردار پشتیبان کاربردی

آرش قربان‌نیادلاور[1,*]، صدف السادات ضیاء[1]

اطلاعات مقاله

دریافت مقاله: ۱۴۰۲/۰۴/۲۲
بازنگری مقاله: ۱۴۰۳/۰۵/۱۱
پذیرش مقاله: ۱۴۰۳/۱۰/۱۲

واژگان کلیدی:
معیار داده‌ها،
داده‌های نامتعادل اعتباری،
سرخوشه،
امتیازدهی،
تعادل بار.

چکیده

مسائل داده‌های اعتباری یکی از مسائل مهم در داده‌کاوی ماشین بردار پشتیبان کاربردی می‌باشد که داده‌های نامتعادل نسبت به داده‌های متعادل نقش مهم‌تری دارد. لذا ما روشی برای افزایش معیارها با امتیازدهی داده‌های نامتعادل اعتباری با استفاده از ماشین بردار پشتیبان کاربردی را ارائه دادیم. به‌همین منظور تکنیک امتیازبندی و همچنین انتخاب سرخوشه مناسب در طبقه‌بندی داده‌ها می‌تواند نقش مهمی در اجرا و دقت برنامه داشته باشد که زمان اجرای برنامه بهینه شده است. با مطالعه موردی بر روی روش‌های شبکه باور عمیق و ماشین بردار پشتیبان، داده‌های نامتعادل اعتباری مورد توجه قرارنگرفته است. با ادغام نمودن انتخاب ویژگی شاخص، همچنین فاصله و اصل همسایگی تابع صلاحیتی را ارائه دادیم که نسبت به روش‌های مورد مطالعه قرار گرفته، از نظر کمی و کیفی می‌تواند معیار مجموعه داده‌ها و کشف تقلب کارت اعتباری و همچنین مجموعه داده‌های پیش‌فرض کارت اعتباری و حساسیت تشخیص روش پیشنهادی و نمونه‌های غیر پیش‌فرض با استفاده از روش IDCOST انجام شده است. که در آن با ایجاد نمونه‌برداری مناسب مجموعه داده‌ها توانستیم افزونگی داده و داده‌های تکراری را نسبت به روش‌های مورد مطالعه کاهش دهیم و همچنین تعادل مجموعه داده‌های انتخاب ویژگی را در روش IDCOST با شاخص معتبری توانستیم، انجام دهیم، تا با این روش افزایش کارایی را در بر داشته‌باشد.

DOI: https://doi.org/10.22075/jme.2025.31252.2493

۱- مقدمه

ریسک اعتباری در فرآیند وام به این احتمال اشاره دارد که بدهکاری که از حمایت اعتباری برخوردار شده است نتواند اصل و سود را به طور کامل و به موقع طبق قرارداد بازپرداخت کند. بانک‌های تجاری اولیه معمولاً ارزیابی ریسک اعتباری را از طریق بررسی کمیته اعتباری یا اتخاذ تصمیمات وام توسط کارشناسان انجام می‌دادند [۱]. فرآیند ارزیابی کاملاً مبتنی بر مشارکت انسانی است و کیفیت ارزیابی در درجه اول به تجربه حرفه ای داوران بستگی دارد. پس از آن، با افزایش مستمر مقیاس کسب و کار وام، بانک های تجاری بیشتر و بیشتری انتخاب می کنند تا از مدل امتیازدهی اعتباری برای کمی کردن ریسک اعتباری کسب و کار وام به صورت دسته ای استفاده کنند [۲]. مدل امتیازدهی اعتباری مجموعه‌ای از مدل‌های تصمیم‌گیری و فناوری‌های پشتیبان برای کمک به موسسات وام برای صدور وام است. در مقایسه با بررسی وام توسط کمیته ها یا کارشناسان، استفاده از مدل امتیازدهی اعتباری می تواند به

* پست الکترونیک نویسنده مسئول: a_ghorbannia@pnu.ac.ir
۱. گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه پیام نور، تهران، ایران

استناد به این مقاله:
آرش قربان نیادلاور و صدف السادات ضیاء، "IDCOST: روشی برای افزایش سرویس معیار داده‌ها با امتیازدهی داده‌های نامتعادل اعتباری با استفاده از SVM کاربردی،" مدل سازی در مهندسی، ۲۳ شماره ویژه ۸۱ (۱۴۰۴): ۱-۱۸، doi: 10.22075/jme.2025.31252.2493

طور موثر سطح علمی فرآیند ارزیابی و عینیت تصمیم گیری وام را افزایش دهد و همچنین از خطر اخلاقی و ریسک عملیاتی جلوگیری کند. پس از دهه‌ها توسعه، تعداد زیادی از مدل‌های امتیازدهی اعتباری معرف طراحی شده‌اند و به طور فزاینده‌ای محبوب هستند، مانند رگرسیون لجستیک، naive Bayesian، درخت تصمیم، ماشین بردار پشتیبان، مدل شبکه عصبی مصنوعی و غیره. تا به امروز، بسیاری از محققان به بررسی مدل‌های امتیازدهی اعتباری مجموعه یا ترکیبی و عملکرد بهتری نسبت به مدل‌های تک ساختاری به دست آوردند.

اگرچه در حال حاضر مدل‌های اعتباری بالغ و محبوب بسیاری وجود دارد، اما تقریباً همه انواع موسسات اعتباری، به‌ویژه بانک‌های تجاری، همچنان ترجیح می‌دهند از رگرسیون لجستیک [۳] به عنوان مدل اصلی امتیازدهی اعتباری استفاده کنند. برای چندین دهه، رگرسیون لجستیک می تواند موقعیت روش استاندارد را در سناریوهای امتیازدهی اعتباری به دلایل زیادی حفظ کند. اول از همه، توافق بازل III ایجاب می‌کند که رویکرد مبتنی بر رتبه‌بندی داخلی بانک‌های تجاری (IRB) برای ارزیابی ریسک اعتباری باید قابل تفسیر و قوی باشد و بانک‌ها را ملزم می‌کند تا با جزئیات بیشتری در مورد عملکرد خود در برآورد پارامترهای مدل توضیح دهند. اکثر مدل‌های مبتنی بر فناوری یادگیری ماشینی از نظر ساختار و پارامترها «جعبه‌های سیاه» هستند و همچنین توضیح فرآیند ارزیابی آنها برای تنظیم‌کننده‌ها آسان نیست. دوم، مدل رگرسیون لجستیک خود دارای ساختار ساده و همچنین پارامترهای قابل مشاهده و تفسیر است که می تواند الزامات تنظیم کننده ها را برآورده کند. علاوه بر این، فرآیند محاسبه مدل رگرسیون لجستیک واضح و ساده است، بدون هیچ گونه فرضی در مورد توزیع نمونه‌های مدل‌سازی، همه اینها باعث می‌شود که عملکرد قوی داشته باشد.

با ظهور عصر کلان داده، مجموعه داده‌های موجود در بانک‌های تجاری برای امتیازدهی اعتبار همچنان در ابعاد نمونه و حجم نمونه افزایش می‌یابد و به دنبال آن مسائل نوظهور عدم تعادل نمونه، پیشرفت مستمر فناوری ارزیابی ریسک اعتباری باعث می شود اکثر متقاضیان وام که پشتیبانی وام دریافت کرده‌اند در نهایت به نمونه های غیر پیش فرض تبدیل شوند، نسبت نمونه های پیش فرض در کل مجموعه داده ها همچنان کاهش می‌یابد. اگر پردازش مدل امتیازدهی اعتباری بر روی این مجموعه داده‌های نامتعادل باشد، مدل توانایی تشخیص نمونه‌های غیرپیش‌فرض به عنوان نمونه‌های کلاس اکثریت را تقویت می‌کند، در حالی که توانایی تشخیص نمونه‌های پیش‌فرض به عنوان نمونه‌های کلاس اقلیت به طور جدی کاهش می‌یابد. برای مدل رگرسیون لجستیک، مقابله موثر با مشکل عدم‌تعادل نمونه دشوار است، و به راحتی می توان نمونه‌های پرخطر بیشتر را به نمونه‌های کم خطر طبقه‌بندی کرد، که در نهایت دقت پیش بینی و ارزش کاربردی مدل را تضعیف می‌کند.

برخی برای مقابله با عدم‌تعادل از روش KTSVM [۴]، Bagging-SVM [۵]، WGMM [۶] و یا از MVQS [۷] استفاده کردند. برخی از ماشین بردار پشتیبان برای طبقه بندی صدای بیماران و تشخیص صدای غیرطبیعی بیمار استفاده میکنند [۸].

مدل‌های امتیازدهی اعتباری موجود را می‌توان به دو نوع تقسیم کرد: مدل تک ساختاری و مدل ساختار گروهی. محققان اولیه بر پیاده سازی ابزارهای آماری یا داده‌کاوی برای سناریوهای امتیازدهی اعتباری متمرکز بودند. از آن زمان، محققان بیشتری با استفاده از چندین مدل پایه برای توسعه مدل‌های مجموعه‌ای به منظور افزایش بیشتر اثر ارزیابی مورد بررسی قرار داده‌اند.

مدل کارت امتیازی قدیمی ترین مدل تک ساختاری است و همواره یکی از فناوری‌های پیشرو در زمینه تحلیل ریسک اعتباری بوده است. این تجربه کارشناسان اعتباری را با هم ترکیب می‌کند، به هر شاخص ریسک نمره خاصی می‌دهد و با جمع کردن همه زیر امتیازها، امتیاز نهایی را به دست می‌آورد. تلاش برای مدل کارت امتیازی بسیاری از محققین را برانگیخت تا از ابزارهای آماری پیشرفته‌تری برای کاهش تداخل عوامل انسانی استفاده کنند و مدل لجستیک، مدل پروبیت، مدل Naive Bayesian و غیره را توسعه دهند. در بین همه این مدل‌ها، مدل لجستیک به دلیل قابلیت تفسیر و استحکام، بیشترین استفاده را در عمل امتیازدهی اعتباری دارد [۳].

با پیشرفت تکنولوژی یادگیری ماشینی، محققان بیشتر و بیشتری سعی در معرفی روش‌های بیشتری برای امتیازدهی اعتباری داشته‌اند. دستاوردهای نماینده شامل مدل درخت تصمیم، مدل k-نزدیکترین همسایه، مدل ماشین بردار پشتیبان، مدل شبکه عصبی مصنوعی و غیره است [۹].

تعداد زیادی از مطالعات تجربی نشان می‌دهد که مدل‌های یادگیری ماشین توانایی تعمیم قوی‌تری نسبت به مدل‌های آماری سنتی دارند و توانایی بیشتری در شناسایی رابطه غیرخطی بین متغیرها به طور مؤثرتری دارند. با این وجود، عملکرد آنها نسبت به تنظیم پارامترهای فوق حساس است و ساختار پیچیده مدل نیز باعث می‌شود فرآیند محاسباتی فاقد قابلیت تفسیر باشد. انواع مختلف مدل‌های تک سازه مزایا و معایب خاص خود را دارند، مدل سازه مجموعه چندین مدل تک سازه را به عنوان مدل فرعی در ساختار خود می‌گیرد. با بازی کامل به مزیت‌های نسبی هر مدل فرعی، بهبود عملکرد کلی مدل مجموعه قابل تحقق است. از دیدگاه ساختار، مدل‌های مجموعه را می‌توان به مدل‌های ساختار سریال، مدل‌های ساختار موازی و مدل‌های ساختار ترکیبی تقسیم کرد.

مدل ساختار سریال به مدل‌ها یا ماژول‌های مختلف اجازه می‌دهد تا عملیات و پیش‌بینی‌ها را به ترتیب خاصی انجام دهند، خروجی مدل فرعی قبلی ورودی مدل فرعی دومی است. در این فرآیند، دقت نتایج پیش‌بینی همچنان بهبود می‌یابد. به عنوان مثال بعضی از محققان یک مدل مجموعه جدید طراحی کرد، ابتدا نمونه‌های نویز را بر اساس روش یکپارچه‌سازی انباشته شناسایی کرد و سپس مدل‌های فرعی ساخت. در مرحله بعد از روش‌های بیزی برای بهینه‌سازی مدل‌های فرعی استفاده می‌شود و در نهایت وزن‌ها تخصیص داده می‌شوند و از روش رأی‌گیری برای به دست آوردن نتایج پیش‌بینی نهایی استفاده می‌شود. محققی دیگر مدل LSTM-RNN را توسعه داد که ابتدا از شبکه های عصبی مکرر برای جا دادن نمونه‌ها استفاده می‌کند و سپس از فناوری حافظه کوتاه مدت (LSTM) برای به خاطر سپردن الگوهای بلند مدت استفاده می‌کند. نتایج تجربی نشان می دهد که این مدل می‌تواند عملکرد بهتری نسبت به بسیاری از مدل های تک ساختاری داشته باشد.

مدل ساختار موازی به مدل‌های فرعی مختلف اجازه می‌دهد تا به طور جداگانه پیش‌بینی کنند و نتایج پیش‌بینی خود را از طریق مکانیسم وزن‌دهی خاصی برای به دست آوردن نتایج نهایی یکپارچه می‌کند. مدل‌های معرف زیادی از این نوع در زمینه یادگیری ماشین متولد شده‌اند که پس از اعمال در سناریوهای امتیازدهی اعتبار نتایج خوبی از خود نشان می‌دهند، مانند مدل جنگل تصادفی، مدل تقویت تطبیقی با درخت‌های تصمیم به عنوان طبقه‌بندی‌کننده اصلی، درخت تصمیم تقویت‌کننده گرادیان، مدل تقویت شیب شدید و غیره. علاوه بر این، بسیاری از محققان چندین مدل ساختار ساده بالغ را برای ساخت مدل‌های مجموعه ترکیب کرده‌اند. به عنوان مثال، محققانی با ادغام مدل خطی تعمیم یافته، مدل درخت تصمیم، ماشین بردار پشتیبان و مدل Naive Bayesian، تعدادی مدل فرعی ساختند و باعث شدند مدل مجموعه از طریق پیوندهای موازی عملکرد برتری داشته باشد. محققی دیگر یک مدل ساختار مجموعه شامل قانون فازی، شبکه عصبی تکراری و سیستم‌های اطلاعات فازی عصبی تطبیقی را طراحی کرد، نتایج تجربی نشان می‌دهد که از استحکام خوبی برخوردار است.

مدل ساختار هیبریدی شامل هر دو ویژگی ساختار سریال و موازی است، بنابراین مکانیسم آن پیچیده‌تر است. بطور مثال :

۱- مدل MLCCE که ابتدا پیش پردازش نمونه و کاهش ابعاد نمونه را در کل مجموعه داده انجام می‌دهد، سپس کل مجموعه داده را به زیر مجموعه ها تقسیم می‌کند و در نهایت مدل پیش بینی مجموعه را می‌سازد.

۲- مدل مجموعه سه مرحله‌ای، ابتدا مجموعه داده‌های اصلی به چند زیر مجموعه تقسیم می‌شود، سپس بر روی هر زیر مجموعه زیر مدل‌های LSTM یا GRU ساخته می‌شود و در نهایت مدل مجموعه شکل می‌گیرد.

۳-مدل سه مرحله‌ای دیگر، ابتدا نرمال‌سازی مجموعه داده‌های اصلی، تشخیص پرت و انتخاب شاخص‌ها را انجام می‌دهد و سپس از مدل XGBoost برای تبدیل ویژگی‌های انتخاب شده استفاده می‌شود [۱۰]. در نهایت، یک مدل امتیازدهی اعتباری بر اساس شبکه عصبی عمیق ساخته شده است.

یک مدل امتیازدهی اعتباری مجموعه‌ای ساخته شد که شامل فرآیندهای یادگیری ماشینی تحت نظارت و بدون نظارت است. ابتدا، ابزارهای مختلف یادگیری ماشینی تحت نظارت را برای ساخت چندین مدل فرعی اعمال می‌کند، سپس یک روش یادگیری ماشین بدون نظارت مانند خوشه‌بندی برای تقسیم کل مجموعه داده‌ها به زیر مجموعه‌های داده اعمال می‌شود، و یادگیری نظارت‌شده دوباره برای ایجاد برخی مدل‌های فرعی دیگر انجام می‌شود. در نهایت، تمام نتایج مدل‌های فرعی برای به دست آوردن

نتایج نهایی ترکیب می‌شوند. به طور کلی، اکثر مطالعات موجود نشان می‌دهند که در مقایسه با مدل‌های تک ساختاری، مدل‌های مجموعه می‌توانند توانایی تعمیم و استحکام بهتری به دست آورند. به همین ترتیب، پیچیدگی زمان محاسبات مدل‌های مجموعه نیز بالا است.

اگرچه بسیاری از مدل‌های تک ساختاری و مدل‌های مجموعه مبتنی بر روش‌های یادگیری ماشین عملکرد بهتری نسبت به مدل لجستیک در مجموعه داده‌های مختلف نشان داده‌اند، مدل لجستیک همچنان محبوب‌ترین مدل در عمل وام دنیای واقعی است، به دلیل الزامات محدودکننده تنظیم‌کننده‌ها در قابلیت تفسیر و استحکام مدل های ارزیابی ریسک در این راستا، بسیاری از مطالعات تلاش کرده اند تا مدل رگرسیون لجستیک پایه را به منظور بهبود بیشتر عملکرد ارزیابی بر اساس حفظ قابلیت تفسیر آن بهبود بخشند.

مشکل عدم تعادل نمونه به عدم تعادل در تعداد نمونه‌های کلاس مختلف در مشکلات طبقه‌بندی مانند امتیازدهی اعتبار و شناسایی تقلب اشاره دارد [۱۱]. در یک مفهوم کلی، اگر تعداد یک کلاس خاص از نمونه‌ها در مجموعه داده‌ها به طور مطلق در بین تمام کلاس‌های نمونه غالب باشد، در نظر گرفته می‌شود که مشکل عدم تعادل نمونه در مجموعه داده‌ها وجود دارد. به بیان دقیق، تا زمانی که تعداد نمونه‌های کلاس مختلف در مجموعه داده ناسازگار باشد، می‌توان در نظر گرفت که مشکل عدم تعادل نمونه در مجموعه داده وجود دارد.

در زمینه امتیازدهی اعتباری، پژوهشگران معمولاً به نمونه‌های پیش‌فرض علاقه دارند، اما تعداد نمونه‌های غیرپیش‌فرض همیشه بسیار بیشتر از تعداد نمونه‌های پیش‌فرض است و نسبت نمونه‌های پیش‌فرض به نمونه‌های غیرپیش‌فرض ممکن است گاهی ۱ به ۱۰۰ برسد، یا حتی بیشتر. اگر مجموعه داده با نسبت عدم تعادل نمونه بالا قبل از مدل سازی از قبل پردازش نشده باشد، بسیار آسان است که توانایی تشخیص مدل به شدت نسبت به نمونه های کلاس اکثریت، که نمونه‌های غیر پیش فرض هستند، سوگیری پیدا کند، در حالی که آن توانایی تشخیص نمونه‌های کلاس اقلیت، که نمونه های پیش فرض هستند، بسیار ضعیف است. به طور کلی، این مدل نمونه‌های پیش‌فرض واقعی‌تری را به عنوان نمونه‌های غیرپیش‌فرض شناسایی می‌کند که منجر به کاهش ارزش عملی آن می‌شود.

الگوریتم های موجود برای مقابله با مشکل عدم تعادل نمونه را می‌توان به طور تقریبی بر اساس اصل اساسی خود به سه دسته تقسیم کرد: زیر نمونه‌گیری، بیش از نمونه‌گیری و نمونه‌گیری ترکیبی.

زیر نمونه‌گیری به نمونه برداری از نمونه‌های اکثریت برای تولید یک زیرمجموعه حاوی تعداد کمی از نمونه‌های اکثریت اشاره دارد، به طوری که تعداد نمونه‌های اکثریت در زیر مجموعه با نمونه‌های اقلیت متعادل شود. یکی از راه‌های رایج این است که مرزهای نمونه‌های کلاس‌های مختلف را دورتر و واضح‌تر کنیم، از طریق حذف نمونه‌های کلاس اکثریت در جفت‌های نمونه ناهمگن مشابه‌تر، مانند قوانین فشرده نزدیک‌ترین همسایه، پیوند تامک، و غیره. یکی دیگر از راه‌های محبوب تمرکز بر روی قوانین طبقه‌بندی بر اساس نمونه‌های مرزی، مانند قانون تمیز کردن محله، قانون نزدیک به اشتباه و قانون آستانه سختی نمونه.

نمونه‌برداری بیش از حد به نمونه‌برداری از نمونه‌های اقلیت برای تولید زیرمجموعه بزرگتری از نمونه‌های اقلیت اشاره دارد، به طوری که تعداد نمونه‌های اقلیت در زیر مجموعه کم و بیش با نمونه‌های اکثریت متعادل باشد. بعضی از محققین روش SMOTE را پیشنهاد کردند و پس از آن، مدل‌های مشتق و مدل‌های متنوعی مانند ADASYN، Borderline-SMOTE، KMeans-SMOTE، SVM-SMOTE، LR-SOMTE و غیره طراحی شدند. اکثر این روش‌ها از منظر انتخاب نمونه اولیه، شناسایی نمونه‌های مرزی، الگوریتم و قوانین جدید تولید نمونه اقلیت و غیره، بررسی می‌شوند.

نمونه‌گیری مختلط مفهوم زیر نمونه‌گیری و نمونه برداری بیش از حد را ترکیب می‌کند و با حذف برخی از نمونه‌های اکثریت و گسترش نمونه‌های اقلیت به طور همزمان تعادل دو نمونه کلاس را محقق می‌کند. به عنوان مثال، SMOTE برای تولید نمونه‌های اقلیت جدید، سپس یک بخش پیوند تامک را اضافه کنید تا نمونه‌های نویز را دور بیندازید و مدل SMOTE-Tomek یا روش SMOTE-ENN را بسازید، که از قانون نزدیک‌ترین همسایه ویرایش‌شده وزنی برای شناسایی و حذف کلاس اکثریت و اقلیت نویز استفاده می‌کند. نمونه ها پس از استفاده از روش SMOTE، علاوه بر این، برخی از مقالات بر روی "نمونه‌های مرزی" (نزدیک

نمونه های کلاس های مختلف) تمرکز می‌کنند و راه‌های بهتری برای تشخیص آنها پیدا می‌کنند.

۲- مطالعات انجام شده

برخی نویسندگان روش پیشنهادی را با سه گروه از الگوریتم‌ها مقایسه کردند، به شرح زیر [۱۲]:

۱- الگوریتم‌هایی مانند رگرسیون لجستیک[۲]، شبکه عصبی[۳]، ماشین بردار پشتیبان[۴]، شبکه بیزی[۵] و درخت تصمیم[۶].

۲- الگوریتم‌های تقویت‌کننده مانند Logit Boost، AdaBoost، RUSBoost، RBBoost، SMOTEBoost، Bagging، Voting و BalanceCascade.

۳- الگوریتم‌های حساس به درآمد که شامل AdaCost و MetaCost می‌شود.

نتایج نشان داد که CSDE در سه معیار ارزیابی عملکرد بهتری از روش‌های دیگر دارد. میانگین رتبه‌بندی نشان می‌دهد که CSDE و CSDNN به ترتیب رتبه‌های اول و سوم را در بین تمام روش‌ها دارند. روش‌های پیشنهادی همچنین به شکاف‌های تعمیم متوسط بسیار پایینی دست یافتند که نشان می‌دهد روش‌ها بعید است مشکل بیش‌برازش داشته باشند. علاوه بر این، CSDE نتایج امیدوارکننده‌ای را در اکثر مجموعه‌های داده در طیف وسیعی از نسبت‌های عدم‌تعادل ایجاد کرد. در این مقاله از تکنیک امتیازبندی و میانگین گیری استفاده نشده است و همچنین زمان اجرا محاسبه نشده است.

برخی نویسندگان مدل پیشنهادی، پیش‌بینی را به‌عنوان یک مجموع وزنی مدل‌سازی می‌کند و درک چگونگی ایجاد پیش‌بینی را آسان می‌کند [۱۳]. در این مقاله از روش میانگین گیری استفاده نشده است و همینطور زمان اجرا محاسبه نشده است. اما روشی برای بدست آوردن امتیاز پیش بینی ریسک مالی ارائه داده اند.

برخی الگوریتم استراتژی امتیازدهی خودکار اعتبار[۷] ارائه می‌دهند. روش پیشنهادی به طور مداوم با توجه به دقت و معیار HM[۸] رتبه‌بندی می‌شود. آزمایش‌های بیشتر نشان داد که معیارهای خودکارسازی و دقت را می‌توان در مقایسه با سایر روش‌های امتیازدهی اعتباری متعادل کرد [۱۴]. در این مقاله از روش میانگین گیری استفاده نشده است و لازم بذکر است که زمان اجرا نیز محاسبه نشده است.

برخی نویسندگان از الگوریتم‌های خوشه‌بندی k-means و توصیف دامنه بردار پشتیبان[۹] استفاده می‌کنند که نتیجه آن نشان داد که تعداد بهینه خوشه‌های پیش‌بینی شده در یک سال آینده ۱۲ است، در حالی که تابع هسته بهینه آن است. به ترتیب، در کل، استفاده از هسته RBF و Poly در SVDD ترجیح داده می‌شود [۱۵]. در این مقاله از تکنیک امتیازبندی و میانگین گیری استفاده نشده است. اینجا از خوشه بندی استفاده شده است.

برخی دیگر مدلی پیشنهاد کردند که عملکرد عالی را برای مجموعه‌داده با نسبت عدم‌تعادل کم ارائه می‌دهد، در اکثر موارد از ۱۴ روش مرتبط برتر از لحاظ F-measure، G-mean، AUC- area، MMD-score، Silhouette-score بهتر عمل می‌کند [۱۶]. در این مقاله تکنیک امتیازبندی و میانگین گیری استفاده نشده است.

برخی دیگر از نویسندگان از الگوریتم‌های گروهی تقویت‌کننده، AugBoost-RFS و AugBoost-RFU استفاده می‌کنند که روش‌های پیشنهادی از مهارت جاسازی مبتنی بر درخت برای تحقق بخشیدن به افزایش گام به گام ویژگی برای چارچوب تقویتی استفاده کردند. استراتژی گروه تقویتی، فراگیران پایه را در AugBoost-RFS/AugBoost-RFU قادر می‌سازد تا به طور مکرر «اشتباه‌ها» را تصحیح کنند. تعبیه مبتنی بر درخت به عملکرد تعبیه ویژگی دست یافت در حالی که پیچیدگی افزایش گام‌به‌گام ویژگی را کاهش داد [۱۷]. در این مقاله از میانگین گیری استفاده نشده است هرچند که یک تکنیک امتیازدهی ارائه داده است.

برخی دیگر از الگوریتم‌های KNN[۱۰]، SVM، RF[۱۱]، NB[۱۲]، همسایه‌های وزنی تطبیقی[۱۳] استفاده می‌کنند که با ترکیب میانگین وزنی افزونگی با عدم‌قطعیت متقارن بین ویژگی‌ها و کلاس‌های تصمیم، الگوریتم کاهش ویژگی با استفاده از

[2] LR: Logistic Regression
[3] NN: Neural Network
[4] SVM: Support Vector Machine
[5] BN: Bayesian Network
[6] DT: Decision Tree
[7] ACSS: Automatic Credit Scoring Strategy

[۸] معیار HM برای اندازه‌گیری عملکرد مدل استفاده می‌شود که خودکارسازی و دقت را متعادل می‌کند.
[9] SVDD: support vector domain description
[10] K-nearest neighbor
[11] Random Forest
[12] NaiveBayes
[13] AWKNN: adaptive weighted k-nearest neighbors

مبتنی بر عدم‌قطعیت متقارن طراحی شده‌است تا ویژگی‌های نمونه را به عنوان زیرمجموعه ویژگی بهینه انتخاب‌کند که ارتباط و افزونگی را در بین ویژگی‌ها در نظر می‌گیرد [۱۸]. در این روش هم تکنیک امتیازبندی رعایت نشده است.

برخی دیگر روش credit default forecasting called (CDFS) ماژول چندجانبه پردازش داده، انتخاب ویژگی، تعادل داده، پیش‌بینی، ارزیابی و تفسیر. ارائه شده است. مجموعه داده عملکرد پیش‌بینی‌کننده عالی و قابلیت تفسیر رضایت‌بخشی را نشان داد. این مطالعه به بهبود دقت پیش‌بینی داده های اعتباری و کاهش ریسک اعتباری در مؤسسات مالی مورد استفاده قرار گرفته شد [۱۹]. روش CDFS اتکاپذیری چارچوبی که با امتیازدهی داده‌های نامتعادل اعتباری انجام نشده است همچنین ماژول چندجانبه مورد ارزیابی مستقیم تابع صلاحیت قرار گرفته نشد.

برخی دیگر با روشی پیشنهاد کردند که می‌توان ابعاد نمونه‌ها را کاهش داد، همچنین برای مدیریت داده‌های مالی با توزیع غیرخطی مناسب‌تر باشد و پیچیدگی محاسباتی را از دست بدهد. بهبود SMOTE، که می‌تواند نمونه‌های مصنوعی جدید را غیرمتمرکزتر کند و از برازش بیش از حد جلوگیری کند [۲۰]. در این روش از خوشه‌بندی استفاده نشده است.

برخی دیگر طبقه‌بندی ریسک اعتباری با پراکندگی رفتار با معرفی ماژول Behavior2-Shapelets استفاده شده است [۲۱]. برای افزایش استحکام مدل، از یک استراتژی پویا برای تعیین آستانه استفاده شده است که بر اساس مدل آماری کولموگروف-اسمیرنوف است. دراین روش نسبت به مورد پایه از نظر دقت طبقه‌بندی و استحکام بهتر عمل می‌کند. درماژول Behavior2-Shapelets در صورتیکه طبقه بندی با دسته بندی ورودی ها و ایجاد سرخوشه مناسب استفاده شده بود دقت طبقه‌بندی و استحکام بهینه تر بوده است.

برخی دیگر با استفاده از KNN ماژول‌های موجود در مدل پیشنهادی می‌توانند به طور موثر عملکرد تشخیص سیستم را بهبود بخشند. برخی دیگر مدعی هستند که با این الگوریتم می‌تواند به طور قابل توجهی اثربخشی مجموعه‌داده را با نرخ عدم‌تعادل بالا و بعد ویژگی بالا بهبود بخشد، به‌ویژه مجموعه‌داده‌هایی که طبقه‌بندی آن دشوار است [۲۲]. در این روش تکنیک امتیازبندی و میانگین وزنی رعایت نشده است.

برخی دیگر برای مجموعه داده¬های بزرگ با استفاده از شبکه عصبی مصنوعی مقیاس پذیر کارایی یک مدل شبکه عصبی مصنوعی (ANN) را با استفاده از استراتژی اعتبارسنجی متقاطع K fold برای مقابله با یک مجموعه داده نامتعادل انجام شده است ارائه یک مدل ANN کم عمق مقیاس پذیر است با در نظر گرفتن قابلیت عملکرد بالا و سربار کم ANN، هدف ما این است که با ارائه یک مدل ANN بهینه شده است [۲۳]، اما مشکل شبکه عصبی مصنوعی (ANN) در تحلیل محاسبه پیچیده عددی و همچنین استفاده از سرخوشه مناسب تا تعادلی برای متعادل نمودن داده تصادفی نامنظم می باشد.

برخی دیگر زمانی که طبقه‌بندی‌کننده پایه ارزیابی اعتبار چند طبقه‌بندی را انجام می‌دهد، عملکرد هر الگوریتم طبقه‌بندی متفاوت است و در نتیجه سطوح مختلفی از خطاهای دقت ارزیابی در دسته‌های مختلف ایجاد می‌شود، و مکملی بین هر طبقه‌بندی پایه وجود دارد. مدل MIFCA ساخته شده در این مقاله از مکمل بودن منابع اطلاعاتی متعدد استفاده می‌کند، اطلاعات اضافی هر طبقه‌بندی‌کننده را ترکیب می‌کند و عدم‌قطعیت کلی مدل را کاهش می‌دهد و در نتیجه دقت طبقه‌بندی را بهبود می‌بخشد [۲۴]. در این روش تکنیک امتیازبندی رعایت نشده است.

برخی دیگر روش‌های نمونه‌گیری تصادفی و SMOTE در امتیازدهی اعتبار کارآمد هستند. نمرات رتبه این روش‌ها همگی کمتر از ۱۰ است، به این معنی که عملکرد آنها در رسیدگی به مشکل امتیازدهی اعتباری کاملاً پایدار است [۲۵]. مشکل نمونه گیری مجدد، زمانی است داده های تکراری و افزونگی در روش مورد مطالعه در نظر گرفته شود تا با استفاده نمونه‌های غیر پیش‌فرض بتوانیم زمان اجرای واقعی را بهینه نماییم.

برای امتیازدهی اعتباری BSAC یک مدل یادگیری است که به طور همزمان از قدرت برتر رمزگذارهای خودکار نظارت شده و یادگیری بازنمایی در طبقه بندی و همچنین مکانیسم Bagging برای رسیدگی به بی نظمی ها در فضای ویژگی استفاده شده است [۲۶]. استفاده داده‌های نامتعادل و بی‌نظمی ویژگی‌ها در فضای پنهان زمان اجرای و تاخیر آن عملکرد بهینه نداشته است در صورتیکه با دسته

بندی ورودی ها وزمان اجرای واقعی نسبت به معیار، برتری و استحکام راندمان را افزایش دهیم.

برخی دیگر خوشه‌ای مجدد مقیاس‌شده و سپس پیش‌بینی برای امتیازدهی اعتباری طبقه‌بندی‌کننده XGBoost مورد استفاده واقع شده است افزایش تفسیرپذیری و عملکرد پیش‌بینی مدل‌های امتیازدهی اعتباری انجام شده است [۲۷]. خوشه‌ای مجدد مقیاس‌شده و سپس پیش‌بینی برای ابتدا دسته بندی ورودی انجام سپس با شکل دهی با درجه حساسیت مثبت و منفی وزن دقت و اصل همسایگی اتکاپذیری تفسیرپذیری و عملکرد پیش‌بینی را بهینه نمود. گروهی دیگر نمونه‌برداری فعال تأثیر زیادی در بهبود دقت مدل TLC دارد. علاوه‌براین، برای رسیدگی به مشکل استفاده کم از داده‌های بلادرنگ، یک سیستم تحلیل سرتاسر طراحی می‌کنیم و مدل پیشنهادی می‌تواند مستقیماً در ماژول تحلیل سیستم اعمال شود [۲۸]. در این روش تکنیک امتیازبندی و میانگین وزنی رعایت نشده است. برخی دیگر معتقدند که مدل رگرسیون لجستیک به دلیل تفسیرپذیری قوی از نتایج، به طور گسترده در عمل امتیازدهی اعتباری استفاده می‌شود، اما عملکرد تشخیص آن برای نمونه‌های پیش‌فرض که در مجموعه داده‌های نامتعادل دنیای واقعی اقلیت هستند، باید بهبود یابد. پس یک مدل مجموعه جدید بر اساس رگرسیون لجستیک به عنوان مدل لجستیک-BWE پیشنهاد می‌کنند [۳]. به منظور مقایسه عملکرد مدل پیشنهادی با مدل‌های نماینده موجود به طور جامع‌تر، ۱۰ مدل نماینده به عنوان مدل‌های معیار انتخاب شده‌اند. به طور مشخص، دو مدل آماری به عنوان مدل لجستیک (لجستیک) و GaussianNB، چهار مدل یادگیری ماشین به عنوان مدل درخت تصمیم (DT)، مدل k نزدیک‌ترین همسایه (k-NN)، مدل ماشین برداری پشتیبان (SVM) و مدل شبکه عصبی مصنوعی (BP) و چهار مدل مجموعه به عنوان مدل جنگل تصادفی (RF)، روش تقویت تطبیقی با درخت‌های تصمیم به عنوان طبقه‌بندی‌کننده اصلی (AdaBoost)، مدل درخت تصمیم تقویت کننده گرادیان (GBDT) و مدل تقویت شیب شدید (XGBoost). در مدل لجستیک-BWE برای داده‌های از دست رفته و یا گمشده از معادله ۱ استفاده می‌شود:

$$x' = \frac{x - min(X)}{max(X) - min(X)} \qquad (۱)$$

در اینجا X نشان دهنده بردار است، x مقدار ویژگی یک نمونه خاص است، و x' مقدار مقیاس شده همان ویژگی در نمونه است [۳]. برای پیش‌بینی نهایی از ماتریس درهم‌ریختگی استفاده شده است. جدول ۱ ماتریس درهم‌ریختگی می‌باشد [۲۹] :

جدول ۱- ماتریس درهم‌ریختگی

	پیش‌بینی مقدار صفر	پیش‌بینی مقدار یک
مقدار صفر واقعی	TN	FP
مقدار یک واقعی	FN	TP

اگر مقادیر پیش‌فرض با ۱ یا مثبت، و مقادیر غیرپیش‌فرض با صفر یا منفی ارائه شوند، TN (منفی واقعی) تعداد نمونه‌های غیرپیش‌فرض را نشان می‌دهد که به درستی به‌عنوان غیرپیش‌فرض، پیش‌بینی شده‌اند. FP (مثبت غلط) تعداد نمونه‌های غیرپیش‌فرض را نشان می‌دهد که به اشتباه به‌عنوان پیش‌فرض، پیش‌بینی شده‌اند. FN (منفی نادرست) تعداد نمونه‌های پیش‌فرض را نشان می‌دهد که به اشتباه به‌عنوان غیر پیش‌فرض، پیش‌بینی شده‌اند و TP (مثبت واقعی) تعداد نمونه‌های پیش‌فرض را نشان می‌دهد که به درستی به‌عنوان پیش‌فرض، پیش‌بینی شده‌اند. معادله‌های ۲ و ۳ با استفاده از ماتریس درهم‌ریختگی شکل گرفته اند و در لجستیک-BWE از آنها استفاده شده [۳] :

$$Sensitivity = \frac{TP}{TP + FN} \qquad (۲)$$

$$Specificity = \frac{TN}{TN + FP} \qquad (۳)$$

در جایی که TP، TN به ترتیب تعداد نمونه‌های پیش‌فرض طبقه‌بندی‌شده درست و نمونه‌های غیرپیش‌فرض را نشان می‌دهد، FP نشان‌دهنده تعداد نمونه‌های غیرپیش‌فرض است که اشتباهاً به‌عنوان نمونه‌های پیش‌فرض طبقه‌بندی شده‌اند، FN نشان‌دهنده تعداد نمونه‌های پیش‌فرض است که اشتباهاً به‌عنوان نمونه‌های غیرپیش‌فرض طبقه‌بندی شده‌اند. بدیهی است که Sensitivity توانایی تشخیص نمونه‌های پیش‌فرض مدل را اندازه‌گیری می‌کند، در حالی که Specificity توانایی تشخیص مدل از نمونه‌های غیرپیش‌فرض را اندازه‌گیری می‌کند. معادله ۴، پیش‌بینی نهایی را حاصل می‌شود [۳]:

$$P(X) = \frac{\sum_{i=1}^{n} P_i \times W_i}{\sum_{i=1}^{n} W_i}, W_i \quad (۴)$$
$$= \begin{cases} Sensitivity_i, if\ P_i \geq threshold \\ 1 - Specificity_i, if\ P_i < threshold \end{cases}$$

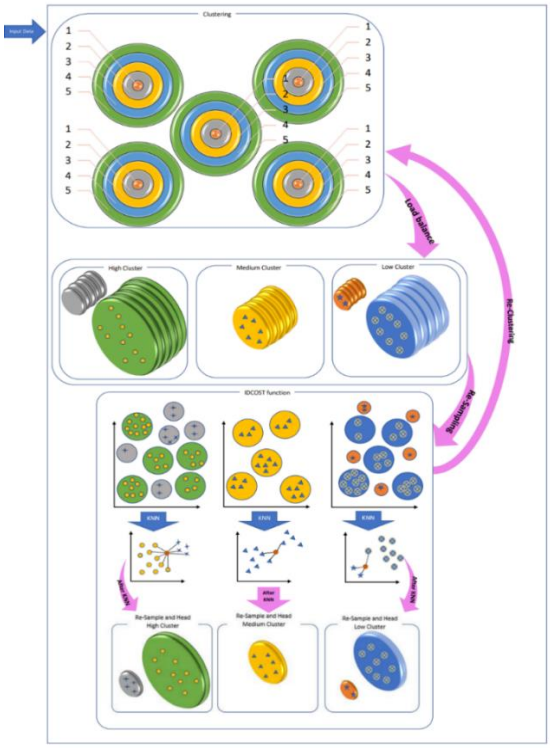
در معادله ۴، P(X) نتیجه پیش‌بینی نهایی را به‌عنوان احتمال پیش‌فرض بودن نمونه X نشان می‌دهد، n تعداد مدل‌های فرعی است و Pi نشان‌دهنده نتیجه پیش‌بینی فرعی تولید شده توسط مدل فرعی i است. Wi وزن ناهمگن برای نتایج زیر پیش‌بینی‌های مختلف است [۳].

یکی از معایب deep learning مشتریان باید هزینه‌ای بپردازند، شناسه کاربری و رمز عبور اختصاصی تنظیم کنند تا بیشتر مطالب ارائه شده توسط این سایت‌ها را دریافت کنند. فقط کسانی که مایل و قادر به پرداخت هزینه‌های این سایت‌ها هستند می‌توانند به محتوای آنها دسترسی داشته باشند. در مواردی که داده‌ها دارای ویژگی پیچیده‌تر و امتیاز بندی درست برحسب اولویت پارامترهای شاخص مورد استفاده واقع نشده است . در مواجهه با داده‌های کم یا ویژگی‌های ساده، حتی دچار پیچیدگی نامنظم باشد و انجام محاسبه آنها ، زمان اجرای واقعی و دقت را نتواند بهینه نماید.

بهبود همگرایی و جلوگیری از پیش‌برازش کمک مستقیم ارائه داده نشده است . که درمطالعه موردی هم محدودیت داشته است نمی تواند ساختاری داده نامتعادل و معیارها با امتیازدهی داده‌های نامتعادل اعتباری را با دقت و عملکرد بالا افزایش دهد. به همین دلیل داده‌کاوی SVM کاربردی تر ومقایسه آن با روش مورد مطالعه دارای کارآمدی بیشتری است. اما لازم بذکر است که SVM به همراه شبکه باور عمیق مطالعه و بررسی شد که بخوبی نتوانسته اند داده های نامتعادل را پوشش دهند. همچنین ادغام شبکه باور عمیق و IDCOST به عنوان یکی از کارهای مورد بحث آینده است.

۳- روش پیشنهادی

در این بخش رویکرد پیشنهادی ما برای امتیازدهی، افزایش دقت و همچنین کاهش افزونگی داده ارائه می‌شود. قصد داریم در روش پیشنهادی خود که آن را IDCOST می‌نامیم، در عین حال که افزونگی داده را کاهش داده و برای مشخصه ها وزن تعیین می‌کنیم دقت را بهبود دهیم و همچنین امتیازدهی مناسب لحاظ شود. IDCOST برای داده‌ها با ابعاد بزرگ نیز کارایی دارد. در شکل (۱) می‌توانید طرح این الگوریتم را مشاهده کنید. طبق شکل (۱) داده های ورودی بعد از خوشه‌بندی به قسمت تعادل بار وارد می‌شوند و به سه دسته تعادل بار کوچک، متوسط و بزرگ تقسیم‌بندی می‌شوند.



شکل ۱- طرح روش پیشنهادی IDCOST

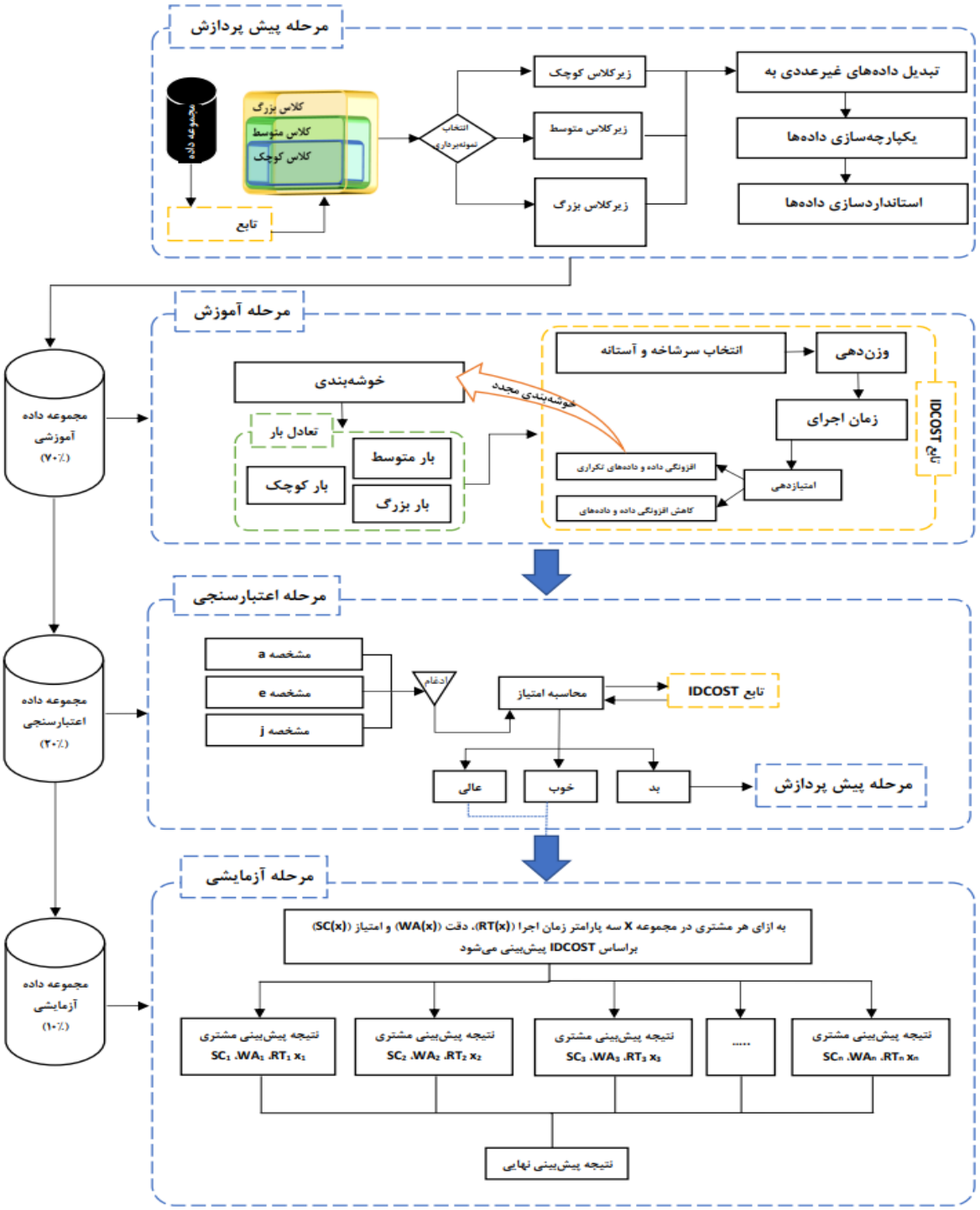
بعد از آن وارد تابع IDCOST شده که از بین شاخه‌های موجود سرشاخه انتخاب می‌شود. بطور مثال از ۲۵ خوشه ایجاد شده بعد از تقسیم‌بندی در قسمت تعادل بار در تابع IDCOST از بین خوشه‌های مرتبط یکی به عنوان سرخوشه یا سرشاخه انتخاب می‌شود که دراینجا ۵ سرشاخه انتخاب شدند.

درصورتی که نتوان داده را در تابع IDCOST در هیچ دسته ای قرارداد، خوشه‌بندی مجدد صورت میگیرد. در شکل (۲) چارچوب مدل پیشنهادی ما IDCOST را مشاهده می‌کنید.

از منظر جریان داده، جریان داده را می توان به چهار مرحله تقسیم نمود. ابتدا در مرحله اولیه، پیش پردازش با دسته بتدی ورودی ها در مرحله دوم آموزش روش با استفاده از داده های اعتباری ارائه شده انجام می شود. متعاقباً، در مرحله سوم ، تعیین آستانه آشکارساز طبقه‌بندی بر روی مجموعه داده اعتبارسنجی انجام شد. روش پیشنهادی آموزش‌دیده برای پیش‌بینی احتمالات پیش‌فرض برای

مجموعه داده‌های اعتبارسنجی اعمال می‌شود. و شاخص $SC(X)$ برای شناسایی آستانه طبقه بندی مناسب استفاده می شود. در نهایت، در مرحله چهارم مجموعه داده آزمایشی برای ارزیابی عملکرد روش استفاده می شود. با وارد کردن مجموعه داده های اعتباری آزمایشی در مدل آموزش‌دیده، احتمالات پیش‌فرض پیش‌بینی‌شده برای هر نمونه آزمایشی به‌دست می‌آیند و متعاقباً بر اساس آستانه آشکارساز اولیه تعیین‌شده به عنوان پیش‌فرض یا غیرپیش‌فرض طبقه‌بندی می‌شوند. عملکرد روش پیشنهادی نسبت به مطالعه های موردی بهینه شده است.

همانطور که مشاهده می‌کنید شامل چهار مرحله است: مرحله اول پیش پردازش داده و مرحله دوم آموزش و سپس اعتبارسنجی و در آخر مرحله آزمایش می‌باشد. در ادامه درباره هر مرحله از شکل (۲) توضیح خواهیم داد.



شکل ۲- چارچوب روش پیشنهادی IDCOST

۳-۱- مجموعه داده‌ها

این مطالعه از شش مجموعه داده اعتبار عمومی با اندازه‌های نمونه مختلف، اندازه ویژگی‌ها و نسبت‌های عدم تعادل استفاده می‌کند تا به طور جامع توانایی پیش‌بینی IDCOST را نشان دهد.

این مجموعه‌داده‌ها شامل ۳ مجموعه داده عمومی پرکاربرد از مخزن داده UCI به عنوان مجموعه داده های استرالیا، مجموعه داده‌های داده‌های آلمانی و مجموعه‌داده مشتری کارت اعتباری پیش فرض، ۲ مجموعه داده عمومی دیگر از مجموعه آتا و مجموعه‌داده‌های تشخیص تقلب کارت اعتباری از سایت کاگل است. علاوه بر این، یک مجموعه داده منبع باز عمومی و در مقیاس بزرگ دیگر به عنوان مجموعه‌داده‌های وام شخصی چینی انتخاب شده است. شرح این مجموعه داده‌ها در جدول ۲ ارائه شده است. این مجموعه داده‌ها به طور گسترده در مطالعات فعلی استفاده می‌شود. جدول ۲ نشان می‌دهد که این مجموعه داده ها شامل نمونه‌هایی از ۶۹۰ تا ۲۸۴۸۰۷، با تعداد ویژگی های مختلف از ۱۰ تا ۳۰ و نسبت عدم تعادل از تقریباً متعادل ۱/۲۵ به عنوان مجموعه‌داده‌های استرالیا تا سناریوهای بسیار نامتعادل ۵۷۷/۸۸ به عنوان تقلب کارت اعتباری است.

جدول ۲- مجموعه داده‌های استفاده شده در مقاله

مجموعه داده	تعداد نمونه	تعداد ویژگی	نمونه‌های غیر پیش‌فرض	نمونه‌های پیش‌فرض	نسبت عدم‌تعادل
استرالیا	690	14	383	307	1.25
آلمان	1000	24	700	300	2.33
وام شخصی چینی	10,000	30	8317	1683	4.94
پیش‌فرض مشتری کارت اعتباری	30,000	23	23,364	6636	3.52
به من اعتبار بده	150,000	10	139,974	10,026	13.96
کشف تقلب کارت اعتباری	284,807	28	284,315	492	577.88

۳-۲- مرحله پیش پردازش

در این مرحله مجموعه داده اصلی به تابع IDCOST وارد می‌شود و آستانه آشکارساز را بدست می‌آورد، براساس این پارامتر سه کلاس ایجاد می‌شود. براساس پیش بینی نمونه انتخابی نسبت به آستانه آشکارساز، نمونه برای پردازش به یکی از سه زیرکلاس منتقل می‌شود. سپس داده‌ها وارد فرآیند تبدیل داده‌های غیرعددی به عددی، یکپارچه‌سازی و استانداردسازی می‌شوند. داده‌های از دست رفته و یا گمشده را نیز از معادله ۵ بدست‌آورده و جایگزین می‌کنیم:

$$x' = \frac{2(x - Avg(min(X)))}{Avg(max(X) - min(X))} \qquad (۵)$$

در اینجا X نشان دهنده بردار است، x مقدار ویژگی یک نمونه خاص است، و x' مقدار مقیاس شده همان ویژگی در نمونه است.

۳-۳- مرحله آموزش

مرحله بعد از پیش پردازش، مرحله آموزش است. همانطور که می‌دانیم بیشتر مجموعه داده‌های دنیای واقعی نامتعادل هستند و مجموعه داده‌های آموزشی نامتعادل، مدل را به سمت داشتن توانایی تشخیص قوی برای نمونه‌های کلاس اکثریت سوق می‌دهد درحالی‌که برای نمونه‌های کلاس اقلیت توانایی تشخیص ضعیفی دارد. در مدل پیشنهادی برای اصلاح عدم‌تعادل مجموعه داده‌های آموزشی از خوشه بندی و الگوریتم نزدیک ترین همسایه استفاده شده است.

۳-۴- مرحله اعتبارسنجی

مرحله بعد از آموزش، مرحله اعتبارسنجی است. هدف از این مرحله قابلیت تعمیم مدل‌های فرعی مختلف آموزش داده شده در مرحله قبل است. در این مرحله عمدتا بر روی حساسیت و ویژگی مدل‌های فرعی تمرکز دارد. با ادغام ویژگی‌ها و محاسبه امتیاز از تابع IDCOST ، رتبه‌بندی صورت می‌گیرد و درصورتی که رتبه "بد" باشد به مرحله پیش پردازش برمی‌گردد در غیر اینصورت به مرحله بعد آزمایشی وارد می‌شود.

۳-۵- مرحله آزمایشی

قسمت آخر مرحله تست یا مرحله آزمایش است. هدف از این مرحله بدست آوردن امتیاز هر نمونه در مجموعه‌داده‌های آزمایشی است. برای هر نمونه X در مجموعه داده‌های آزمایش، ابتدا هر مدل فرعی i یک نتیجه پیش‌بینی SC_i را ارائه می‌دهد که نشان‌دهنده احتمال پیش‌فرض پیش‌بینی‌شده X است. سپس وزن هر مدل فرعی برای پیش‌بینی‌های آن‌ها تعیین می‌شود. وزن هر مدل فرعی از طریق آستانه آشکارساز، زمان اجرای واقعی (T_r) و سرخوشه تعیین می‌شود. آستانه آشکارساز اولیه،t_d مقدار ثابت ۰/۶۶ می‌باشد و در مراحل بعدی طبق معادله ۶

حاصل می‌شود:

$$I_{td}(X) = (\frac{t_d + t'_d}{2}) \times t_e \quad (۶)$$

طبق معادله ۶ آستانه آشکارساز را بصورت پویا بدست می‌آوریم. آستانه آشکارساز در هر مرحله از آستانه آشکارساز مرحله قبلی و زمان اجرا برنامه حاصل می‌شود.

زمان اجرای واقعی طبق معادله ۷ با محاسبه بر روی زمان اجرا (t_e) و زمان رفت (t_{rq}) و برگشت (t_{rs}) حاصل می‌شود.

$$T_r(X) = t_e - (t_{rq} + t_{rs}) \quad (۷)$$

سرخوشه طبق معادله ۱۰ از طریق فاصله اصل همسایگی (معادله ۸) و سایز نمونه (معادله ۹) بدست می‌آید:

$$D_n = \{d_1, d_2, d_3, \dots, d_n\} \quad (۸)$$

$$s = \sum_{i=1}^{n} a_i + \sum_{k=1}^{n} e_k + \sum_{z=1}^{n} j_z \quad (۹)$$

$$F_s(X) = D_n \times s \quad (۱۰)$$

وزن دقت از نتیجه معادله ۱۱ که از طریق سرخوشه شاخص، آستانه آشکارساز و زمان اجرای واقعی که به ترتیب معادله‌های ۱۰، ۶ و ۷ می‌باشد، استفاده می‌کند، حاصل می‌شود:

$$wa_i = \frac{2}{3} I_{td(i)} \times F_{s(i)} + T_{r(i)} => i \in \{1,2,\dots,n\} \quad (۱۱)$$

به این ترتیب معادله ۱۲ وزن دقت را حاصل می‌شود:

$$W_a(X) = \begin{cases} 1 - Specificity = s_i < wa_i \\ Sensitivity = s_i \geq wa_i \\ l_i = m_i \geq wa_i \\ if\ l_i \geq wa_i\ then\ ReClustering \end{cases} \quad (۱۲)$$

برای محاسبه امتیاز نهایی با وزن دادن به تمام SC_i های پیش‌بینی‌شده توسط مدل‌های فرعی مختلف با یک وزن خاص به‌دست می‌آید. معادله ۱۳ پیش‌بینی نهایی مدل پیشنهادی را حاصل می‌شود:

$$SC(X) = \frac{\sum_{i=1}^{n} SC_i \times W_{a(i)}}{\sum_{i=1}^{n} W_{a(i)}} \quad (۱۳)$$

شکل (۳) سودوکد روش IDCOST می‌باشد:

۴- پیاده‌سازی و ارزیابی نتایج

۴-۱- مدل‌های معیار

به منظور مقایسه عملکرد روش IDCOST با مدل‌های نماینده موجود به طور جامع‌تر، ۱۰ مدل نماینده و همچنین مدل logistic-BWE به عنوان مدل‌های معیار انتخاب شده‌اند.

IDCOST Sub Algorithm

Input: a set of training samples: $S=\{(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)\}$,
intelligence threshold: I_{td},
wight of accuracy: W_a,
size of sample: s,
cluster head of feature selection: F_S,
distance principle of neighborhood: D_n,
execution time: t_e,
request time: t_{rq},
response time: t_{rs},
real execution time: T_r.

1. **Process:**
2. **Initialize the** $t_d = 0.66$
3. **for** each sample of S **do**
4. calculate the threshold detector:
5. $I_{td}(X) = \left(\frac{t_d + t'_d}{2}\right) \times t_e$
6. calculate the real execution time:
7. $T_r(X) = t_e - (t_{rq} + t_{rs})$
8. calculate D_n , s then F_s:
9. $D_n = \{d_1, d_2, d_3, \dots, d_n\}$
10. $s = \sum_{i=1}^{n} a_i + \sum_{k=1}^{n} e_k + \sum_{z=1}^{n} j_z$
11. $F_s(X) = D_n \times s$
12. calculate w_a:
13. $wa = \frac{2}{3} I_{td} \times F_s + T_r$
14. **end for**
15. **return** W_a = wa

Output: score is: $SC(X) = \frac{\sum_{i=1}^{n} SC_i \times W_{a(i)}}{\sum_{i=1}^{n} W_{a(i)}}$

شکل ۳- زیرالگوریتم IDCOST

۴-۲- معیارهای ارزیابی

تمام معیارهای ارزیابی از ماتریس درهم‌ریختگی، همانطور که در جدول ۱ تعریف شده است، تولید می‌شوند.

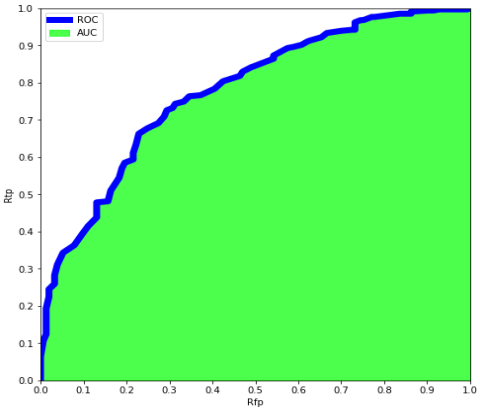
۱- AUC، به عنوان ناحیه ای تعریف می شود که توسط منحنی مشخصه عملکرد گیرنده و دو محور مختصات محصور شده است، که توانایی تمایز بین نمونه های کلاس مختلف یک مدل را اندازه گیری می‌کند. هر چه منحنی ROC به گوشه سمت چپ بالا نزدیکتر باشد. محدوده مقادیر آن در [۰,۱] است و هر چه مقدار AUC بزرگتر باشد، عملکرد طبقه بندی بهتر است.

برای محاسبه AUC لازم است ابتدا منحنی ROC، با محور افقی به عنوان R_{fp} (نرخ مثبت کاذب) و محور عمودی به عنوان R_{tp} (نرخ مثبت واقعی) ترسیم شود که به صورت زیر تعریف می‌شوند:

$$R_{fp} = \frac{FP}{FP + TN} \quad (۱۴)$$

$$R_{tp} = \frac{TP}{TP + FN} \quad (۱۵)$$

FP تعداد نمونه‌های اکثریتی را نشان می‌دهد که اشتباهاً به عنوان نمونه‌های اقلیت طبقه‌بندی شده‌اند، TN نشان‌دهنده تعداد نمونه‌های اکثریت به‌درستی طبقه‌بندی‌شده، TP نشان‌دهنده تعداد نمونه‌های اقلیت به‌درستی طبقه‌بندی‌شده، FN نشان‌دهنده تعداد نمونه‌های اقلیت است که اشتباهاً به عنوان نمونه‌های اکثریت طبقه‌بندی شده‌اند. شکل (۴) نمایش بصری AUC می‌باشد.



شکل ۴- تصویر بصری از AUC

۲- Sensitivity همچنین R_{tp} نامیده می شود، که تشخیص مدل را به نمونه‌های پیش فرض و اقلیت اندازه‌گیری می‌کند.

۳- Specificity حساسیت تشخیص مدل را نسبت به نمونه‌های غیر پیش فرض و اکثریت اندازه‌گیری می‌کند.

۴- G-mean، روش دیگری برای اندازه‌گیری توانایی تشخیص کلی عملکرد مدل در مورد عدم تعادل نمونه است. هر چه به ۱ نزدیکتر باشد، عملکرد بهتری دارد. به صورت زیر محاسبه می‌شود:

$$G - mean = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (۱۶)$$

۵- معیار F-score، میانگین هم‌آهنگ درستی (p) و یادآوری(r) است که در آن به بیشینه مقدار خود امتیاز ۱ و کمینه مقدار، امتیاز صفر می‌رسد. F-score، دقت و یادآوری به صورت زیر تعریف می‌شوند:

$$r = \frac{TP}{TP + FN} \quad (۱۷)$$

$$p = \frac{TP}{TP + FP} \quad (۱۸)$$

$$F - scocre = \frac{2 \times r \times p}{r + p} \quad (۱۹)$$

۶- معیار MCC که مخفف ضریب همبستگی متیو است. در اصل یک مقدار ضریب همبستگی بین [۱،-۱] است. ضریب ۱ ، یک پیش‌بینی کامل، صفر یک پیش‌بینی تصادفی متوسط و ۱- یک پیش‌بینی معکوس را نشان می‌دهد. می‌توان آن را به صورت زیر محاسبه کرد:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (۲۰)$$

۴-۳- شاخص‌های IDCOST

در جدول ۳ شاخص های استفاده شده در روش IDCOST را مشاهده می‌کنید.

جدول ۳- شاخص‌های IDCOST

عنوان شاخص	شرح شاخص
F_S	سرخوشه
t_d	آستانه آشکارساز اولیه
I_{td}	آستانه آشکارساز
T_r	زمان اجرای واقعی
D_n	فاصله اصل همسایگی
SC	امتیاز
W_a	وزن دقت
s	سایز نمونه
t_e	زمان اجرا
t_{rq}	زمان رفت
t_{rs}	زمان برگشت
r	یادآوری
p	درستی
F-score	میانگین درستی و یادآوری
MCC	ضریب هم‌بستگی متیو
G-mean	اندازه توانایی تشخیص عملکرد مدل در مورد عدم تعادل نمونه
Sensitivity	نرخ مثبت واقعی
Specificity	نرخ منفی واقعی
R_{fp}	نرخ مثبت کاذب
R_{tp}	نرخ مثبت واقعی

۴-۴- آزمایش‌های انجام شده

معیارهایی که عنوان کردیم را برتمام مدل‌های ذکر شده به ازای هر مجموعه داده تست و بررسی کردیم و نتایج آنها را در جداول قرار داده‌ایم. جداول ۴ تا ۹ نتیجه معیارها را برای مجموعه داده‌ها را نشان می‌دهد و همانطور که مشاهده می‌کنید مدل IDCOST پاسخ بهینه‌تری را نسبت به دیگر مدل‌ها ارائه داده است. طبق جدول ۴، مدل پیشنهادی ما در مجموعه داده استرالیا به نتیجه بهینه‌تری در هر معیار دست یافته‌است.

جدول ۴- نتیجه معبارهای مجموعه‌داده استرالیا

مدل	AUC	Sensitivity	Specificity	G-mean	F-score	MCC
IDCOST	0.912	0.922	0.986	0.905	0.899	0.776
Logistic-BWE	0.884	0.861	0.869	0.865	0.85	0.73
XGBoost	0.863	0.857	0.868	0.863	0.847	0.727
AdaBoost	0.86	0.846	0.834	0.84	0.849	0.718
GaussianNB	0.857	0.853	0.841	0.847	0.841	0.712
GBDT	0.855	0.851	0.859	0.855	0.838	0.708
RF	0.851	0.841	0.861	0.851	0.835	0.701
logistic	0.851	0.853	0.848	0.851	0.835	0.699
k-NN	0.831	0.853	0.809	0.831	0.819	0.662
DT	0.657	0.531	0.783	0.645	0.588	0.33
SVM	0.578	0.186	0.969	0.425	0.304	0.259
BP	0.596	0.28	0.913	0.505	0.33	0.22

طبق جدول ۵، مدل پیشنهادی ما در مجموعه داده آلمان نیز به نتیجه بهینه‌تری در هر معیار نسبت به دیگر مدل‌ها دست یافته‌است.

جدول۵- نتیجه معیارهای مجموعه‌داده آلمانی

مدل	AUC	Sensitivity	Specificity	G-mean	F-score	MCC
IDCOST	0.779	0.837	0.944	0.756	0.654	0.459
logistic-BWE	0.763	0.789	0.682	0.734	0.624	0.438
GaussianNB	0.696	0.634	0.759	0.694	0.577	0.377
SVM	0.695	0.658	0.733	0.694	0.575	0.368
logistic	0.67	0.441	0.9	0.63	0.527	0.391
XGBoost	0.666	0.478	0.854	0.639	0.524	0.354
GBDT	0.662	0.49	0.834	0.639	0.521	0.337
AdaBoost	0.652	0.417	0.887	0.608	0.486	0.346
DT	0.615	0.42	0.81	0.583	0.442	0.239
RF	0.609	0.327	0.892	0.54	0.414	0.273
k-NN	0.608	0.352	0.865	0.552	0.419	0.25
BP	0.591	0.259	0.923	0.489	0.298	0.205

طبق جدول ۶، مدل پیشنهادی ما در مجموعه داده کشف تقلب کارت اعتباری نیز به نتیجه بهینه‌تری در هر معیار نسبت به دیگر مدل‌ها دست یافته‌است. به نحوی که در معیار AUC از مدل logistic-BWE افزایش یافته‌است.

جدول۶- نتیجه معیارهای مجموعه‌داده کشف تقلب کارت اعتباری

مدل	AUC	Sensitivity	Specificity	G-mean	F-score	MCC
IDCOST	0.886	0.855	0.93	0.891	0.857	0.852
AdaBoost	0.858	0.79	0.926	0.855	0.837	0.836
XGBoost	0.855	0.784	0.926	0.852	0.837	0.836
k-NN	0.857	0.788	0.926	0.854	0.837	0.835
SVM-rbf	0.85	0.773	0.926	0.846	0.831	0.831
RF	0.84	0.755	0.906	0.827	0.825	0.826
logistic-BWE	0.88	0.842	0.928	0.884	0.841	0.825
DT	0.853	0.781	0.925	0.85	0.824	0.822
BP	0.854	0.788	0.92	0.851	0.789	0.789
GBDT	0.81	0.695	0.925	0.802	0.757	0.761
logistic	0.844	0.773	0.916	0.841	0.731	0.735
GaussianNB	0.839	0.769	0.91	0.836	0.668	0.664

طبق جدول ۷، مدل پیشنهادی ما در مجموعه داده وام شخصی چینی نیز به نتیجه بهینه‌تری در هر معیار نسبت به دیگر مدل‌ها دست یافته‌است. به نحوی که در معیار AUC وSensitivity با مدل logistic-BWE در رقابت نزدیکی است اما در دیگر معیارها به طور قابل ملاحظه‌ای این تفاوت مشهود است.

جدول۷- نتیجه معیارهای مجموعه‌داده وام شخصی چینی

مدل	AUC	Sensitivity	Specificity	G-mean	F-score	MCC
IDCOST	0.829	0.979	0.998	0.824	0.597	0.502
logistic-BWE	0.807	0.911	0.683	0.789	0.565	0.475
GBDT	0.721	0.582	0.861	0.708	0.503	0.41
DT	0.687	0.466	0.908	0.65	0.479	0.388
SVM	0.756	0.874	0.637	0.746	0.477	0.386
GaussianNB	0.752	0.874	0.629	0.742	0.475	0.382
AdaBoost	0.653	0.376	0.931	0.591	0.436	0.352
XGBoost	0.654	0.393	0.914	0.6	0.431	0.334
RF	0.621	0.314	0.928	0.54	0.371	0.29
logistic	0.592	0.216	0.968	0.457	0.313	0.284
k-NN	0.602	0.329	0.875	0.537	0.337	0.21
BP	0.563	0.165	0.961	0.399	0.213	0.177

طبق جدول ۸، مدل پیشنهادی ما در مجموعه داده پیش‌فرض مشتری کارت اعتباری نیز به نتیجه بهینه‌تری در هر معیار نسبت به دیگر مدل‌ها دست یافته‌است.

جدول۸- نتیجه معیارهای مجموعه‌داده پیش‌فرض مشتری کارت اعتباری

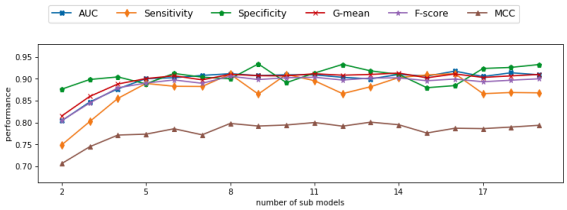
مدل	AUC	Sensitivity	Specificity	G-mean	F-score	MCC
IDCOST	0.672	0.741	1	0.68	0.697	0.418
AdaBoost	0.648	0.346	0.949	0.573	0.456	0.384
DT	0.658	0.396	0.92	0.604	0.476	0.371
RF	0.64	0.331	0.948	0.56	0.439	0.367
logistic-BWE	0.653	0.679	0.58	0.627	0.639	0.36
XGBoost	0.643	0.379	0.908	0.587	0.448	0.331
GBDT	0.641	0.373	0.909	0.583	0.444	0.328
GaussianNB	0.6	0.647	0.553	0.598	0.408	0.17
SVM	0.597	0.599	0.595	0.597	0.403	0.165
k-NN	0.55	0.262	0.838	0.469	0.286	0.108
BP	0.52	0.131	0.91	0.346	0.115	0.047
logistic	0.5	0.001	0.999	0.035	0.002	0.014

در جدول ۹، نتایج مدل پیشنهادی ما در مجموعه داده به من اعتبار بده را نشان می‌دهد. به نتایج بهینه‌تری در هر معیار نسبت به دیگر مدل‌ها دست یافته‌است.

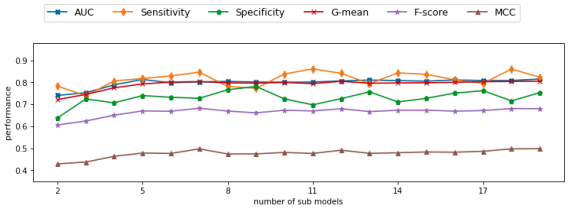
جدول۹- نتیجه معیارهای مجموعه‌داده به من اعتبار بده

مدل	AUC	Sensitivity	Specificity	G-mean	F-score	MCC
IDCOST	0.824	0.801	1	0.838	0.855	0.579
logistic-BWE	0.797	0.734	0.804	0.768	0.784	0.531
GBDT	0.69	0.441	0.939	0.643	0.377	0.333
AdaBoost	0.613	0.244	0.981	0.49	0.324	0.316
XGBoost	0.595	0.208	0.981	0.452	0.282	0.272
DT	0.594	0.209	0.978	0.452	0.274	0.259
RF	0.552	0.112	0.993	0.333	0.175	0.194
logistic	0.52	0.041	0.998	0.203	0.072	0.116
BP	0.51	0.021	0.999	0.145	0.04	0.099
SVM	0.575	0.719	0.43	0.556	0.152	0.078
k-NN	0.5	0.002	0.999	0.045	0.004	0.006
GaussianNB	0.5	0	1	0	0	0

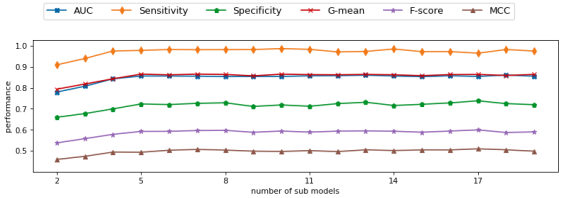
جداول ۴ تا ۹ برای زمانی است که ما ۷ زیرمدل داشته باشیم در ادامه نتیجه نمودارهای IDCOST را برای هر مجموعه داده و برای تعداد زیرمدل‌های متغیر بطور مثال ۲ تا ۱۹ زیرمدل نشان می‌دهیم.



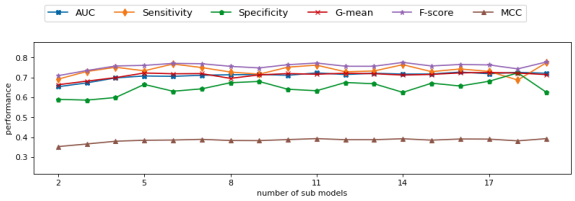
نمودار ۱- نتیجه معیارها برای مجموعه داده استرالیا با استفاده از IDCOST



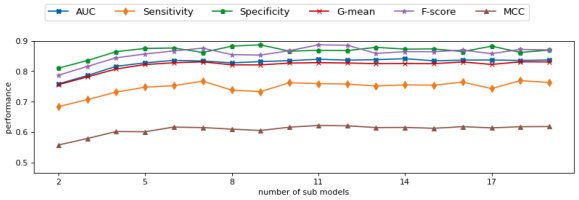
نمودار ۲- نتیجه معیارها برای مجموعه داده آلمان با استفاده از IDCOST



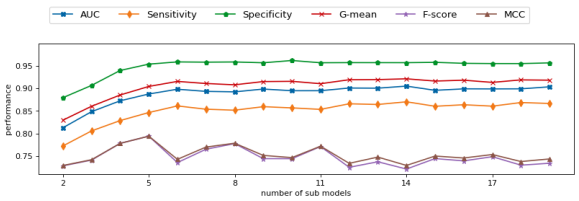
نمودار ۳- نتیجه معیارها برای مجموعه داده وام شخصی چینی با استفاده از IDCOST



نمودار ۴- نتیجه معیارها برای مجموعه داده پیش‌فرض مشتری کارت اعتباری با استفاده از IDCOST

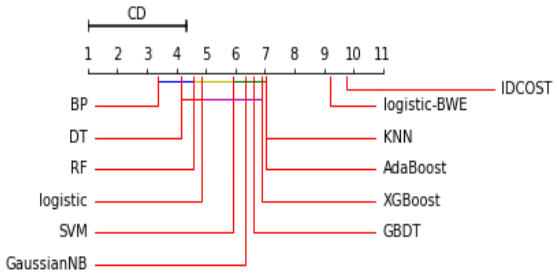


نمودار ۵- نتیجه معیارها برای مجموعه داده به من اعتبار بده با استفاده از IDCOST



نمودار ۶- نتیجه معیارها برای مجموعه داده کشف تقلب کارت اعتباری با استفاده از IDCOST

نمودارهای حاصل از IDCOST نتایج بهینه تری را نسبت به دیگر مدل‌ها نشان می‌دهد همچنین نسبت به مدل مقایسه‌ای ما logistic-BWE نتایج بهینه‌تری برای معیارها ارائه داد. طبق آزمایش فردمن کارایی روش IDCOST از دیگر مدل‌های مورد مطالعه بیشینه است و نمودار ۷ نشان دهنده این ادعا می‌باشد:



نمودار ۷- مقایسه کارایی مدل‌ها

۵- نتیجه‌گیری و کارهای آینده

همانطور که در جداول و نمودارها نشان داده شده است، توانستیم با تابع IDCOST افزونگی داده و داده‌های تکراری را کاهش دهیم و همچنین با انتخاب سرخوشه و نمونه‌برداری، داده‌ها متعادل شده که باعث افزایش وزن دقت و امتیازدهی بهینه شد.

طبق نتایج حاصل شده برای مجموعه داده استرالیا برای معیار AUC مقدار IDCOST نسبت به مدل بیشینه logistic-BWE ، ۲/۸ درصد و نسبت به مدل کمینه SVM ۳۳/۴ درصد افزایش یافته است. در مجموعه داده آلمان برای معیار AUC مقدار IDCOST نسبت به مدل بیشینه logistic-BWE ۱/۶ درصد و نسبت به مدل کمینه BP ۱۸/۸ درصد افزایش یافته است. در مجموعه داده کشف تقلب کارت اعتباری برای معیار AUC مقدار IDCOST نسبت به مدل بیشینه logistic-BWE ۰/۶ درصد و نسبت به مدل کمینه GBDT ۷/۶ درصد افزایش یافته است. در مجموعه داده وام شخصی چینی برای معیار AUC مقدار IDCOST نسبت به مدل بیشینه logistic-BWE ۲/۲ درصد و نسبت به مدل کمینه BP ۲۶/۶ درصد افزایش یافته است. در مجموعه داده به من اعتباربده برای معیار AUC مقدار IDCOST نسبت به مدل بیشینه logistic-BWE ۲/۷ درصد و نسبت به مدل کمینه GNB ۳۲/۴ درصد افزایش یافته است و در نهایت برای مجموعه داده پیش‌فرض کارت مشتری برای مدل بیشینه DT ۱/۴ درصد و برای مدل کمینه ۱۷/۲ درصد افزایش داشته‌ایم.

با استفاده از تکنیک امتیازبندی و انتخاب سرخوشه توانستیم دقت و زمان اجرا را بهبود دهیم و همچنین با استفاده از انتخاب ویژگی شاخص و محاسبه فاصله و اصل همسایگی توانستیم معیار مجموعه داده ها را کشف کنیم.

در مجموعه داده اعتباری، ویژگی ها را می توان به سه دسته تقسیم کرد: ویژگی های دسته ایستا، ویژگی های عددی ایستا، و ویژگی های رفتاری پویا. اول، روش‌های مختلف تعبیه‌سازی برای و ویژگی‌های طبقه‌بندی استاتیک اتخاذ شده‌اند. در روش پیشنهادی ویژگی های عددی وایستا مورد توجه قرار گرفته شد اما ویژگی‌های رفتاری پویا نیز می توانست مورد استفاده قرار گرفته شود.

معیار ما بر اساس مجموعه داده‌های امتیازدهی، اعتبار عمومی از مجموعه یک داده اعتباری می باشد ، عملکرد معیار انتخابی در سایر مجموعه‌های داده ها و ارتباط آن با اعتبار عمومی بین مجموعه با مقایسه سناریو واقعی برای استفاده از افراد خبره موردتوجه قرار گرفته شود.

تقدیر و تشکر

با تشکر از واحد پژوهش دانشگاه پیام نور تهران که بستر مناسبی برای انجام پژوهش مورد نظر ایجاد نمودند.

تعارض منافع

نویسندگان اعلام می‌کنند که در مورد انتشار این مقاله تعارض منافع وجود ندارد.

تاییدیه اخلاقی

نویسندگان متعهد میشوند که مطالب این مقاله را در هیچ مجله دیگری به چاپ نرسانده‌اند.

مشارکت های نویسندگان

آرش قربان نیا دلاور: روش‌شناسی، نوآوری، بازبینی پردازش داده ها و ارزیابی آن، بازبینی پیش‌نویس اصلی، اعتبارسنجی، بررسی و ویرایش، تجزیه و تحلیل داده و توابع، راهنمایی

صدف السادات ضیاء: نرم افزار، پردازش داده ها، بررسی و ویرایش، پیش‌نویس اصلی، منابع

منابع مالی

در انجام پژوهش هیچ منابع مالی مورد استفاده قرار نگرفته است.

مراجع

[1] T. Aliheidari Bioki, and Hassan Khademizare. 2015. "Improvement of DEA Approach for Clustering Credit Rating of Customer in Banks." *Journal of Modeling in Engineering* 13, no. 41 (2015): 59–74. (in Persian)

[2] X. Xie, X. Shi, J. Gu, and X. Xu. "Examining the Contagion Effect of Credit Risk in a Supply Chain under Trade Credit and Bank Loan Offering." *Omega* 115 (2023): 102751.

[3] R. Zhang, X. Liguo, and W. Qin. "An Ensemble Credit Scoring Model Based on Logistic Regression with Heterogeneous Balancing and Weighting Effects." *Expert Systems with Applications* 212 (2023): 118732.

[4] C. Jimenez-Castaño, A. Álvarez-Meza, D. Cárdenas-Peña, A. Orozco-Gutíerrez, and J. Guerrero-Erazo. "Kreĭn Twin Support Vector Machines for Imbalanced Data Classification." *Pattern Recognition Letters* 182 (2024): 39–45.

[5] W. Zhai, Xiya Xiong, Guozhao Mo, Yuzhen Xiao, Caicong Wu, Zhi Xu, and Jiawen Pan. "A Bagging-SVM Field-Road Trajectory Classification Model Based on Feature Enhancement." *Computers and Electronics in Agriculture* 217 (2024): 108635–35.

[6] C. Dou, Yan Lv, Zhen Wang, and Lan Bai. "Handling Imbalanced Classification Problems by Weighted Generalization Memorization Machine." *Applied Artificial Intelligence* 38, no. 1 (2024): 2355424.

[7] Z. Hou, J. Tang, Y. Li, S. Fu, and Y. Tian. "MVQS: Robust multi-view instance-level cost-sensitive learning method for imbalanced data classification." *Information Sciences* 675 (2024): 120467.

[8] M. Zakariah, M. Al-Razgan, and T. Alfakih. "Pathological voice classification using MEEL features and SVM-TabNet model." *Speech Communication* 162 (2024): 103100.

[9] X. Gao, Z. Meng, X. Jia, J. Liu, X. Diao, B. Xue, Z. Huang, and K. Li. "An imbalanced binary classification method based on contrastive learning using multi-label confidence comparisons within sample-neighbors pair." Neurocomputing 517 (2023): 148-164.

[10] R. Asencios, C. Asencios, and E. Ramos. "Profit Scoring for Credit Unions Using the Multilayer Perceptron, XGBoost and TabNet Algorithms: Evidence from Peru." *Expert Systems with Applications* 213 (2023): 119201.

[11] Liu, Wanan, Hong Fan, Min Xia, and Meng Xia. "A Focal-Aware Cost-Sensitive Boosted Tree for Imbalanced Credit Scoring." *Expert Systems with Applications* 208 (2022): 118158.

[12] Wong, Man Leung, Kruy Seng, and Pak Kan Wong. "Cost-Sensitive Ensemble of Stacked Denoising Autoencoders for Class Imbalance Problems in Business Domain." *Expert Systems with Applications* 141 (2020): 112918.

[13] V. Moscato, A. Picariello, and G. Sperlí. "A Benchmark of Machine Learning Approaches for Credit Score Prediction." *Expert Systems with Applications* 165 (2021): 113986.

[14] F. Yang, Y. Qiao, C. Huang, S. Wang, and X. Wang. "An Automatic Credit Scoring Strategy (ACSS) Using Memetic Evolutionary Algorithm and Neural Architecture Search." *Applied Soft Computing* 113 (2021): 107871.

[15] K. Yuan, G. Chi, Y. Zhou, and H. Yin. 2022. "A Novel Two-Stage Hybrid Default Prediction Model with K-Means Clustering and Support Vector Domain Description." *Research in International Business and Finance* 59 (2022): 101536.

[16] J. Zhai, J. Qi, and C. Shen. "Binary Imbalanced Data Classification Based on Diversity Oversampling by Generative Models." *Information Sciences* 585 (2022): 313–43.

[17] W. Liu, H. Fan, and M. Xia. "Credit Scoring Based on Tree-Enhanced Gradient Boosting Decision Trees." *Expert Systems with Applications* 189 (2022): 116034.

[18] L. Sun, J. Zhang, W. Ding, and J. Xu. "Feature Reduction for Imbalanced Data Classification Using Similarity-Based Feature Clustering with Adaptive Weighted K-Nearest Neighbors." *Information Sciences* 593 (2022): 591–613.

[19] W. Sun, X. Zhang, M. Li, and Y. Wang. "Interpretable High-Stakes Decision Support System for Credit Default Forecasting." *Technological Forecasting & Social Change* 196 (2023): 122825.

[20] L. Wang. "Imbalanced Credit Risk Prediction Based on SMOTE and Multi-Kernel FCM Improved by Particle Swarm Optimization." *Applied Soft Computing* 114 (2022): 108153.

[21] L. Yu, and C. He. "A Shapelet-Based Behavioral Pattern Extraction Method for Credit Risk Classification with Behavior Sparsity." *Advanced Engineering Informatics* 58 (2023): 102227.

[22] H. Ding, L. Chen, L. Dong, Z. Fu, and X. Cui. "Imbalanced Data Classification: A KNN and Generative Adversarial Networks-Based Hybrid Approach for Intrusion Detection." *Future Generation Computer Systems* 131 (2022): 240–54.

[23] S. Sen, K. Pratap Singh, and P. Chakraborty. "Dealing with Imbalanced Regression Problem for Large Dataset Using Scalable Artificial Neural Network." *New Astronomy* 99 (2023): 101959.

[24] T. Wang, R. Liu, and G. Qi. "Multi-Classification Assessment of Bank Personal Credit Risk Based on Multi-Source Information Fusion." *Expert Systems with Applications* 191 (2022): 116236.

[25] C. Jiang, W. Lu, Z. Wang, and Y. Ding. "Benchmarking State-of-The-Art Imbalanced Data Learning Approaches for Credit Scoring." *Expert Systems with Applications* 213 (2023): 118878.

[26] M. Abdoli, M. Akbari, and J. Shahrabi. "Bagging Supervised Autoencoder Classifier for Credit Scoring." *Expert Systems with Applications* 213 (2023): 118991.

[27] H.W. Teng, M.H. Kang, I.H. Lee, and L.C. Bai. "Bridging Accuracy and Interpretability: A Rescaled Cluster-Then-Predict Approach for Enhanced Credit Scoring." *International Review of Financial Analysis* 91 (2024): 103005–5.

[28] Y. Liu, G. Yang, S. Qiao, M. Liu, L. Qu, N. Han, T. Wu, G. Yuan, T. Wu, and Y. Peng. "Imbalanced Data Classification: Using Transfer Learning and Active Sampling." *Engineering Applications of Artificial Intelligence* 117 (2023): 105621.

[29] Y. Wang, Y. Jia, Y. Tian, and J. Xiao. "Deep Reinforcement Learning with the Confusion-Matrix-Based Dynamic Reward Function for Customer Credit Scoring." *Expert Systems with Applications* 200 (2022): 117013.