اعتبارسنجی ادعای بیمه بیکاری با استفاده از روش ترکیب رده‌بندها

رحیم دهخوارقانی* [۱]، حجت امامی [۲]

[۱] دانشگاه ایشیک، استانبول، ترکیه،

[۲] دانشگاه بناب، بناب، ایران،

چکیده

بیمه بیکاری یکی از مهم‌ترین و پرطرفدارترین انواع بیمه در دنیای امروزی محسوب‌می‌شود. سازمان تأمین اجتماعی در مقابل ادعای بیکاری افراد تحت پوشش این سازمان، وظیفه بررسی صحت این موضوع را دارد. بررسی دستیِ ادعای افراد بیکار نیازمند صرف زمان و هزینه زیادی است. روش‌های داده‌کاوی و یادگیری ماشین به‌عنوان ابزارهای کارآمدِ تحلیل داده‌ها می‌تواند در خودکارسازی این فرآیند به سازمان تأمین اجتماعی کمک کنند. در این پژوهش، روشی مبتنی بر یادگیری نظارتی برای بررسی صحت ادعای بیکاری افراد متقاضی ارائه شده‌است. روش پیشنهادی، اطلاعات بیمه‌شدگان را به‌عنوان ورودی دریافت کرده و پس از تحلیل داده‌ها به هر فرد متقاضی امتیازی تخصیص می‌دهد؛ سپس بر اساس مقدار این امتیاز، مدعیان بیمه بیکاری را به دو گروه "شایسته دریافت بیمه بیکاری" و "فاقد کفایت برای دریافت بیمه بیکاری" دسته‌بندی می‌کند. روش پیشنهادی از دو ترکیب مختلف برای دسته‌بندی ادعای متقاضیان استفاده می‌کند: روش BSA-SVM و روش ترکیب ضرایب اطمینان رده‌بندها. در روش BSA-SVM برای بهبود کارایی و تخمین پارامترهای کنترلی SVM، از الگوریتم بهینه‌سازی جستجوی عقبگرد (BSA) استفاده شده‌است. در روش ترکیب ضرایب اطمینان رده‌بندها، تعدادی رده‌بند، از جمله شبکه‌های عصبی مصنوعی، درخت تصمیم و رگرسیون لجستیک داده‌ها را رده‌بندی کرده و ضرایب اطمینان این رده‌بندها با دو روش مختلف با همدیگر ترکیب می‌شوند. نتایج آزمایش‌ها نشان‌می‌دهد که روش پیشنهادی BSA-SVM با کسب ۸۷٪ و روش ترکیب رده‌بندها با ضرایب اطمینان با کسب دقت ۸۶٪، کارایی بهتری در قیاس با سایر روش‌های موجود کسب کرده‌اند.

واژه‌های کلیدی: بیمه بیکاری، داده‌کاوی، یادگیری ماشین، یادگیری نظارتی، BSA-SVM، ترکیب رده‌بندها.

Verification of unemployment benefits claims using classifier combination method

Rahim Dehkharghani[1], Hojjat Emami[2]

[1] Computer Engineering Department, Işık University, Istanbul, Turkey
[2] Computer Engineering Department, University of Bonab, Bonab, Iran
Email: rahim.dehkharghani@isikun.edu.tr
emami@ubonab.ac.ir

Abstract

Unemployment insurance is one of the most popular insurance types in the modern world. The Social Security Organization in many countries such as the Islamic Republic of Iran is responsible for checking the unemployment claims of people and paying unemployment benefits to individuals supported by this type of insurance. This payment will be done only if the unemployment claim gets approved by this organization. Manual evaluation of unemployment claims requires a big deal of time and effort, which makes it prohibitively expensive. An automatic or semi-automatic method –proposed in this work- for this evaluation can save time and cost for insurance companies. Data mining and machine learning as

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۸/۲/۱۳ • تاریخ پذیرش: ۱۳۹۹/۲/۲۴ • تاریخ انتشار: ۱۴۰۱/۱۲/۲۹ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.19.4.45]

two efficient tools for data analysis can assist Social Security Organization in automating this process. A number of research works can be found in the literature working on this problem; however, the existing research is far from ideal to solve the current problem both in covering a comprehensive feature set and obtaining high performance. This problem can be approached as a classification or regression problem. In the current work, a hybrid supervised learning method has been proposed to verify the eligibility of applicants for unemployment benefits. The proposed method takes as input the information of insured individuals with unemployment claims, as a feature vector, and assigns a numeric score to each application after analyzing the input data, according to a trained classification system. Then, claimants are classified into one of these two groups according to those scores: "Qualified" and "Unqualified". Furthermore, the effect of each feature such as the payment history, age, marriage status, supported people as family, payment amount, and occupation type on the validity of the unemployment claim has been analyzed for the first time in the literature. The proposed method includes two hybrid strategies: The BSA-SVM method and an ensemble learning method based on confidence values obtained by the base classifiers. In the BSA-SVM method, the backtracking search algorithm (BSA) is used to estimate the parameters of support vector machines (SVM) to improve classification accuracy. In the second approach, confidence values extracted from individual classifiers are combined to better classify the input data. We noticed a remarkable improvement in accuracy when using ensemble learning compared to the isolated usage of each classifier. Empirical evaluation shows an accuracy of 87% for the first approach: BSA-SVM, and 86% for the second approach: ensemble learning. Note that the obtained results are based on real data collected from the Social Security Organization of Iran. We have made a only subset of this dataset publicly available.

Keywords: Unemployment benefits, data mining, machine learning, supervised learning, BSA-SVM, classifier combination.

۱- مقدمه

افزایش اثربخشی استفاده از منابع و سطح کیفیت خدمات هر سازمان به‌منظور ایجاد توان رقابتی و ادامه حیات سازمان‌ها از اهمیت ویژه‌ای برخوردار است. یکی از راه‌های نیل به این هدف در سازمان‌ها، مدیریت صحیح داده‌ها است. افزایش حجم داده‌ها، پیشرفت سریع فناوری و بازار رقابتی، لزوم بهره‌گیری هر چه بیشتر از روش‌های خودکار پردازش داده‌ها و استخراج دانش نهفته در این داده‌ها در مدیریت صنایع و به‌طور خاص صنعت بیمه را ایجاب می‌کند.

بازار امروز با رقابت شدید سازمان‌ها و شرکت‌ها در تمام انواع کسب و کارها مواجه است. مدیریت ارتباط با مشتری، استراتژی پیش‌رو در بازار رقابتی کنونی است که می‌تواند مواردی همچون افزایش تعداد مشتریان، افزایش سطح وفاداری و سودآوری بیشتر مشتریان و ارائه خدمات بهتر و مناسب‌تر به مشتریان را به‌همراه داشته‌باشد.

در این راستا، استفاده از روش‌های هوشمند تحلیل داده‌ها مانند داده‌کاوی و یادگیری ماشین در مدیریت ارتباط با مشتری، می‌تواند برای کسب شناخت دقیق و مناسب از مشتریان مفید باشد.

در صنعت بیمه نیز از زمان پیدایش تاکنون، تغییرات زیادی رخ داده و فضای رقابتی شدید در این سازمان به وجود آمده است. یکی از سازمان‌های بیمه‌گر که قدمت زیادی در ایران دارد، سازمان تأمین اجتماعی است.

جمعیت تحت پوشش این سازمان بیش از ۱۴ میلیون و ۸۰۰ هزار نفر بیمه شده اصلی و بیش از ۳ میلیون و ۸۰۰ هزار نفر مستمری‌بگیر است که با درنظرگرفتن خانواده افراد بیمه شده نزدیک به ۴۲ میلیون نفر می‌رسد[1].

بیمه بیکاری نوعی بیمه است که در آن، در صورت بیکار شدن فرد بیمه‌شده، سازمان تأمین اجتماعی وظیفه پرداخت حق بیمه بیکاری به‌صورت ماهانه به فرد را بر عهده می‌گیرد. متأسفانه برخی از ادعاهای افراد بیمه‌شده مبنی بر بیکاری آن‌ها، صحت نداشته و ادعای دروغین هستند. بررسی صحت ادعای متقاضیان، زمان و هزینه زیادی را بر سازمان تأمین اجتماعی تحمیل می‌کند. به‌دست‌آوردن شناخت دقیق از کلیه‌ی مخاطبان سازمان تأمین اجتماعی، شامل بیمه‌شدگان، مستمری‌بگیران و کارفرمایان یکی از نیازهای اصلی در این سازمان است.

این پژوهش سعی دارد صحت یا عدم صحت ادعای بیمه‌شدگان برای دریافت حق بیمه بیکاری را با روش‌های مبتنی بر یادگیری ماشین تشخیص دهد. در روش پیشنهادی، ابتدا تعدادی رده‌بند[2] با داده‌های برچسب زده‌شده آموزش دیده و سپس در مرحله آزمون، کارایی سامانه آموزش داده‌شده، سنجیده می‌شود. برای انجام رده‌بندی، از نه ویژگی[3] شامل سن، جنسیت، وضعیت تأهل، میزان تحصیلات، تعداد عائله، سابقه پرداخت بیمه،

[1] [https://www.tamin.ir/News/Item/26660/73/26660.html, last access 25 July 2019]
[2] Classifier
[3] Features

[DOI: 10.61186/jsdp.19.4.45]

ماه‌های استحقاق، میزان پرداخت ماهانه و دلیل بیکاری استفاده می‌شود. آزمایش‌های صورت‌گرفته، کارایی بالای روش پیشنهادی را در قیاس با سایر روش‌های موجود تصدیق می‌کند.

نوآوری پژوهش حاضر را می‌توان در موارد زیر خلاصه کرد:

- **بررسی مسئله صحت ادعای بیکاری افراد تحت پوشش بیمه بیکاری سازمان تأمین اجتماعی با روش ترکیب رده‌بندها:** در این پژوهش، روشی ترکیبی که متکی بر چندین رده‌بند است، برای بررسی صحت ادعای بیکاری متقاضیان ارائه شده‌است. تا جایی که نویسندگان این مقاله اطلاع دارند، این پژوهش برای نخستین‌بار در ایران صورت‌می‌گیرد. روش پیشنهادی می‌تواند به سازمان تأمین اجتماعی در تخمین درستی ادعای افراد، با صرف کمینه هزینه، زمان و نیروی انسانی کمک کند.
- **ترکیب بهینه دسته‌بندها:** در این پژوهش، سه روش برای دسته‌بندی ادعای متقاضیان ارائه شده‌است که عبارتند از روش BSA-SVM، روش ترکیب رده‌بندها با ضرایب اطمینان و روش میانگین وزن‌دار با ضرایب اطمینان.
- **ارزیابی روش پیشنهادی:** در این پژوهش، یکی از چالش‌های اساسی در ارزیابی روش پیشنهادی، عدم وجود مجموعه‌داده استاندارد است. به همین منظور، یک مجموعه‌داده نمونه توسط نویسندگان این مقاله ایجاد شده‌است. برای ایجاد مجموعه‌داده، ابتدا ۳۶۰۰۰ رکورد جمع‌آوری‌شده، سپس با انجام پیش‌پردازش، ده‌هزار رکورد برای ایجاد این مجموعه‌داده انتخاب شده و سپس روش پیشنهادی بر اساس معیارهای مختلف بر روی مجموعه‌داده ایجادشده ارزیابی شده‌است. این مجموعه‌داده می‌تواند توسط پژوهش‌گران دیگر نیز برای ارزیابی روش‌های پیشنهادی در آینده استفاده شود. همچنین، تعیین اهمیت هر یک از ویژگی‌ها در رده‌بندی افراد به دو گروه لایق دریافت بیمه بیکاری و عدم کفایت برای دریافت آن، نیز می‌تواند جزو دستاوردهای کار فعلی محسوب شده و توسط سازمان تأمین اجتماعی استفاده شود.

ساختار مقاله بدین صورت سازمان‌دهی شده‌است: بخش ۲ به‌مرور ادبیات موضوع می‌پردازد. در بخش ۳، به جزئیات روش پیشنهادی پرداخته شده و در بخش ۴، نحوه ارزیابی روش پیشنهادی، توضیح داده شده و سپس نتایج حاصل از آزمایش‌ها تشریح شده‌است. در بخش ۵، به جمع‌بندی مطالب و در بخش ۶ نتیجه‌گیری نهایی و همچنین ارائه پیشنهادهایی برای پژوهش‌های آینده پرداخته شده‌است.

۲- کارهای مرتبط

از دیدگاه نگای و همکاران [1] ، تقلب‌ها به چهار دسته کلی بانکی، بیمه‌ای، امنیتی-تجاری و سایر تقسیم می‌شوند. پژوهش حاضر مربوط به گروه تقلب‌های بیمه‌ای است. بررسی تقلب و ادعاهای دروغ در صنعت بیمه، به‌خصوص کشف ادعاهای دروغین برای دریافت بیمه بیکاری می‌تواند کمک شایانی به صرفه‌جویی منابع مالی و جلوگیری از هدر رفت سرمایه شرکت‌های بیمه کند؛ تا جایی که نویسندگان این مقاله اطلاع دارند، کار پیشین که به‌طور دقیق بر روی مسئله اعتبارسنجی بیمه بیکاری متمرکز شده باشد، وجود ندارد؛ اما کارهای دیگری وجود دارند که بر روی مسائل مشابه که نیاز به پیش‌بینی دارند، تمرکز کرده‌اند. کارهای مرتبط را می‌توان به دو دسته کارهای صورت‌گرفته در حوزه بیمه اشخاص و وسائط نقلیه تقسیم کرد. در ادامه به بررسی برخی از مهم‌ترین کارهای صورت‌گرفته در این زمینه‌ها می‌پردازیم.

حسینی و رضائی [2] ریسک‌های موجود در سازمان تأمین اجتماعی با استفاده از اطلاعات آماری را بررسی کردند. این اطلاعات از طریق پرسش‌نامه‌هایی که توسط بیمه‌گذاران تکمیل‌شده، جمع‌آوری‌شده است. در روش پیشنهادی، از سه الگوریتم درخت تصمیم، شبکه‌های عصبی مصنوعی و نزدیک‌ترین همسایه (KNN) برای رده‌بندی ریسک استفاده شده‌است. ویژگی‌های رده‌بندی شامل هشت ویژگی از قبیل جنسیت و میزان حق‌بیمه پرداختی است. نتایج به‌دست‌آمده از آزمایش‌ها نشان‌می‌دهد درخت تصمیم بیشترین کارایی را در دسته‌بندی ریسک دارد. در روش پیشنهادی ایشان، ویژگی‌هایی از قبیل سن، وضعیت تأهل و میزان تحصیلات نادیده گرفته شده‌است که این مسئله می‌تواند جزء نقاط ضعف این کار محسوب شود.

تقوی‌فرد و جعفری [3] روشی مبتنی بر سامانه خبره فازی برای کشف تقلب در بیمه بدنه خودرو پیشنهاد داده‌اند. در مدل ارائه‌شده، از میان ۶۱ ویژگی کیفی و کمی شناسایی‌شده، هفده ویژگی که بر اساس نظر خبرگان از اولویت بیشتری برخوردار بوده‌اند، در فرآیند کشف تقلب به کار گرفته شده‌است. در این روش، از الگوریتم ممدانی[1] برای استنتاج فازی در خصوص تعیین ادعای فرد بیمه‌شده استفاده شده‌است. مشکل اصلی این روش، وابستگی آن به افراد خبره برای طراحی قوانین

[1] Mamdani

[DOI: 10.61186/jsdp.19.4.45]

استنتاج فازی است؛ همچنین این روش وابسته به مسئله بوده و با تغییر مسئله موردنظر، قوانین استنتاجی کارایی نخواهند داشت. برای توسعه روش پیشنهادی می‌توان راهکاری خودکار برای تولید قوانین استنتاج طراحی کرد.

قربانی و فرزای [4] برای بررسی صحت ادعای افراد بیمه‌شده از روش خوشه‌بندی K-Means (KM) استفاده کرده‌اند. آن‌ها در این پژوهش، از نه ویژگی در مورد اتومبیل و فرد بیمه‌شده از جمله سن راننده، میزان جراحت و زمان تصادف استفاده کردند. ازآنجا که روش خوشه‌بندی KM، از مشکل به دام‌افتادن در بهینه محلی و نیز مقداردهی اولیه مقادیر خوشه‌ها رنج می‌برد، روش پیشنهادی در بعضی از موارد از کارایی مناسبی برخوردار نیست.

فیروزی و همکاران [5]، با استفاده از سه روش یادگیری ماشین شامل رگرسیون ترابری بیز ساده و درخت تصمیم روشی را برای شناسایی تقلب در بیمه اتومبیل ارائه داده‌اند. نتایج آزمایش‌ها بر روی یک مجموعه‌داده کوچک که شامل ۷۲ پرونده خسارت بیمه‌نامه‌های شخص ثالث و بدنه اتومبیل است، نشان داد که روش بیز ساده با کسب دقت ۹۰/۲۸٪ عملکرد بهتری در قیاس با روش‌های رگرسیون ترابری و درخت تصمیم دارد. با این وجود، استفاده از یک مجموعه‌داده کوچک، اعتبار نتایج حاصل از روش پیشنهادی را کاهش می‌دهد. با ارزیابی روش پیشنهادی بر روی مجموعه‌داده‌های بزرگ‌تر، می‌توان به نتایج به‌دست‌آمده، اعتبار بیشتری بخشید.

و این و ددن [6] از روش بیز ساده و نیز الگوریتم‌های تقویتی برای کشف تقلب در بیمه اتومبیل استفاده کردند. آن‌ها نتایج حاصل از الگوریتم بیز ساده و الگوریتم‌های تقویتی را مقایسه کرده و به این نتیجه رسیدند که الگوریتم‌های تقویتی نتایج دقیق‌تری نسبت به‌روش بیز ساده به‌دست آورده‌اند. مقایسه با تنها یک روش پایه‌ای یادگیری ماشین (بیز ساده) و عدم مقایسه با روش‌های هوشمندتر مانند شبکه‌های عصبی مصنوعی یا ماشین بردار پشتیبان یکی از نقاط ضعف این کار محسوب‌می‌شود.

حاجی حیدری و همکاران [7] روشی برای رده‌بندی بیمه‌گذاران بیمه‌ی بدنه به لحاظ ریسک دریافت یا عدم دریافت خسارت طی دوره بیمه ارائه کردند. در روش پیشنهادی، ابتدا داده‌های موردنیاز برای آموزش است طی یک دوره مشخص جمع‌آوری و سپس فرآیند پیش‌پردازش داده‌ها و شناسایی ویژگی‌های رده‌بندی انجام شده‌است؛ درنهایت، پس از آزمودن تعدادی از رده‌بندهای هوشمند، از الگوریتم رده‌بند C5 برای رده‌بندی مشتریان استفاده شده‌است که دارای دقت ۸۳٪ است. این روش تنها از پنج ویژگی برای رده‌بندی استفاده کرده است که همگی آن‌ها مربوط به خودرو است؛ درحالی‌که با درنظرگرفتن ویژگی‌های صاحب خودرو (بیمه‌گذار)، می‌توانست نتایج بهتری به‌دست‌آورد. همچنین در این روش، برای مقداردهی به ویژگی‌ها از الگوهای از پیش‌تعریف‌شده استفاده شده‌است که این کار موجب کاهش انعطاف‌پذیری آن و بروز مشکلاتی در نگهداری برنامه‌های کاربردی مبتنی بر این روش می‌شود.

در پژوهشی دیگر [8]، به دسته‌بندی مشتریان بیمه با استفاده از روش‌های داده‌کاوی پرداخته شده‌است. در این روش، بر اساس ویژگی‌های رفتاری، مشتریان در گروه‌های مختلف با ویژگی‌های مشابه قرار گرفته‌اند تا بتوان در تصمیم‌گیری‌های مختلف از آن‌ها استفاده‌کرد. این دسته‌بندی می‌تواند به تعیین اعتبار مشتریان در سازمان‌هایی از قبیل تأمین اجتماعی کمک کند.

گروه دیگری از پژوهش‌گران به بررسی ریسک و تقلب در نظام بانکی پرداختند. رضایی و آقا بیگی [9] به موضوع اعتبارسنجی کلیه مشتریان حقوقی تسهیلات اعتباری بانک ملی طی سال‌های ۸۲ لغایت ۸۵ که فعالیت تولیدی داشتند، پرداخته‌اند. در این روش، از میان ۶۱ متغیر پس از استخراج کلیه اطلاعات موجود از بانک‌های اطلاعاتی و پالایش آن‌ها، ۱۶ متغیر جهت مدل‌سازی انتخاب شده‌است که عبارت‌اند از: نوع طرح، داشتن هم‌گروه، داشتن تعهدات قبلی، نوع شرکت، مدت‌زمان تنفس، میزان تسهیلات، میزان سرمایه شرکت، سابقه فعالیت شرکت، میزان سهم متقاضی در سرمایه‌گذاری، نسبت دارایی جاری، دوره گردش موجودی برحسب روز، دوره وصول مطالبات برحسب روز، بازده فروش شرکت، نسبت مالکانه و نسبت بدهی. نتایج نشان‌می‌دهد که فرآیند اعتباردهی در بانک ملی، یک فرآیند قضاوت است که علاوه‌بر ویژگی‌های یادشده، ویژگی‌های تعهدات قبلی، نسبت دارایی جاری، نسبت مالکانه و نسبت بدهی بر میزان بدحسابی و خوش‌حسابی متقاضیان تسهیلات، متناسب با ضرایب برآورد شده نیز در آن تأثیرگذار هستند.

تهرانی و فلاح شمس [10] با استفاده از داده‌های اعتباری ۳۱۶ مشتری حقوقی بانک‌های کشور و با استفاده از مدل‌های احتمال خطی و شبکه‌های عصبی مصنوعی به موضوع طراحی و آزمون کارایی مدل ریسک اعتباری پرداختند. نتایج حاکی از این است که ارتباط بین متغیرها در مدل پیش‌بینی ریسک اعتباری به‌صورت خطی نبوده و توابع نمایی و سیگموئید، مناسب‌ترین مدل‌های پیش‌بینی ریسک اعتباری بوده و بیشترین کارایی برای پیش‌بینی ریسک اعتباری به‌ترتیب مربوط به شبکه‌های عصبی مصنوعی و مدل رگرسیون ترابری است.

[DOI: 10.61186/jsdp.19.4.45]

عرب مازار و روئین‌تن [4]، با مطالعه موردی بر روی داده‌های بانک کشاورزی، به بررسی عوامل مؤثر بر سنجش ریسک اعتباری مشتریان حقوقی بانک کشاورزی ایران و تدوین مدلی برای آن پرداخته‌اند؛ درنهایت، هفده متغیر که اثر قابل‌توجهی بر ریسک اعتباری و تفکیک بین دو گروه از مشتریان خوش و بدحساب داشتند، انتخاب شده و مدل نهایی به‌وسیله آن‌ها برازش شده‌است.

محمدخان و همکاران [11] با ارائه مدلی برای ارزیابی ریسک اعتباری مشتریان بانک، به بررسی ریسک اعتباری مشتریان حقوقی بانک کارآفرین پرداخته‌اند. نتایج مطالعات بر روی ۹۹ پرونده، میزان کارایی مدل را با استفاده از آزمون هاسمرلمشو، ۸۷٪ برآورد کردند که نشان‌می‌دهد مدل رگرسیون در پیش‌بینی احتمال قصور مشتریان متقاضی وام، از کارایی مطلوبی برخوردار است.

با توجه به اینکه قسمتی از روش پیشنهادی این مقاله مربوط به ترکیب رده‌بندها است، به برخی از مهم‌ترین روش‌های ترکیب رده‌بندها که در همین اواخر ارائه‌شده‌اند، در زیر اشاره شده‌است.

در [12]، به نحوه انجام رده‌بندی دودویی با وجود چند رده (روش‌های "یک در مقابل یک" و "یک در مقابل همه") به‌همراه کارایی استفاده از این روش‌ها پرداخته شده و به سوالاتی از قبیل "آیا همیشه استفاده از روش 'یک در مقابل همه' کارایی را افزایش می‌دهد؟ "، جواب داده‌شده‌است. در [13]، لیو و همکاران با استفاده از روش استدلال شهودی[1] و میانگین وزن‌دار، روشی برای ترکیب رده‌بندها ارائه دادند. این وزن‌ها با کمک قانون دمپستر[2] محاسبه‌شده‌اند. در [14]، از روش‌های خوشه‌بندی برای افزایش میزان دقت در ترکیب رده‌بندها استفاده شده‌است. در این روش، علاوه‌بر انجام رده‌بندی روی داده‌های برچسب دار، داده‌ها بدون توجه به برچسب، بر اساس عدم شباهت ویژگی‌هایشان خوشه‌بندی شده و نتایج این دو فرآیند با هم ترکیب می‌شوند.

در پژوهشی دیگر، کراوزیک و ووزنیاک [15]، با در نظر گرفتن کارایی رده‌بندها و هرس‌کردن رده‌بندهای ناکارآمد، از تابع گاوسی برای وزن‌دهی به رده‌بندها در میانگین وزنی استفاده کرده‌است. این روش نیازی به آموزش برای یادگیری وزن‌ها نداشته و یک روش بدون نظارت محسوب‌می‌شود.

پس از این بررسی کوتاه، مشخص می‌شود که روش‌های ارائه‌شده اگرچه از کارایی مناسبی برخوردار هستند، اما هنوز تلاش بیشتری برای رسیدن به کارایی ایده‌آل موردنیاز است. به‌عنوان پژوهشی در این زمینه و نیز پاسخ‌گویی به نیاز برای ایجاد استای به‌منظور بررسی صحت ادعای متقاضیان بیمه بیکاری، روشی برای رده‌بندی متقاضیان دریافت بیمه ارائه کردیم که در ادامه به تشریح جزئیات آن می‌پردازیم.

۳- روش پیشنهادی

در این پژوهش، مسئله تشخیص صحت ادعای بیکاری متقاضیان به‌صورت یک مسئله دسته‌بندی مدل‌سازی شده‌است. با فرض اینکه $I=\{i_1, i_2, ..., i_n\}$ نشان‌دهنده مجموعه افراد متقاضی بوده و $C=\{C_T, C_F\}$ دو خوشه هستند به‌نحوی‌که $C_T \cap C_F = \varnothing$ و $\cup_{i=1}^{2} C_i = I$. C_T نشان‌دهنده افرادی است که ادعای آن‌ها درست بوده و شایسته دریافت بیمه بیکاری هستند و C_F بیانگر افرادی است که ادعای آن‌ها نادرست بوده و صلاحیت دریافت بیمه بیکاری را ندارند. هدف روش پیشنهادی، دسته‌بندی متقاضیان به دو رده C_T و C_F با بیشترین دقت ممکن است. شکل ۱ روندنمای روش پیشنهادی را نشان‌می‌دهد. روش پیشنهادی شامل مراحل زیر است:

- پیش‌پردازش داده‌ها

در این مرحله، داده‌های ورودی به قالب مناسب برای رده‌بندی تبدیل می‌شوند.

- استخراج ویژگی‌ها

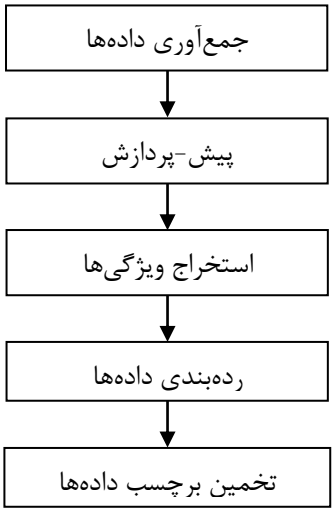
در این مرحله ویژگی‌های موردنیاز و وزن هر کدام از آن‌ها برای دسته‌بندی افراد متقاضی استخراج می‌شوند. در بررسی صحت ادعای بیکاری افراد، برخی از ویژگی‌ها تأثیر بیشتری نسبت به سایر ویژگی‌ها دارند. به‌عنوان مثال، ویژگی "تعداد عائله" تأثیر بیشتری در مقایسه با "وضعیت تأهل" دارد. به این منظور، در مرحله استخراج ویژگی، وزن متفاوتی به هر کدام از ویژگی‌های دسته‌بندی اختصاص داده‌می‌شود تا دقت رده‌بندی افزایش یابد.

- رده‌بندی داده‌ها
- در این مرحله، از دو مدل ترکیب رده‌بندها با ضرایب اطمینان و روش BSA-SVM برای رده‌بندی دادها استفاده شده‌است. در روش استفاده از ضرایب اطمینان، علاوه‌بر روش ترکیب رده‌بندها توسط یک رده‌بند دیگر، از روش میانگین وزن‌دار نیز استفاده شده‌است. به‌عبارت دیگر، در این مقاله، سه روش مختلف برای رده‌بندی رکوردها ارائه شده‌است: روش BSA-SVM[3]، روش ترکیب رده‌بندها با ضرایب اطمینان (CCCV[4]) و روش میانگین وزن‌دار با ضرایب اطمینان (WACV[5]).

[1] Evidential reasoning
[2] Dempster's rule
[3] Backtracking Search Algorithm-Support Vector Machine
[4] Classifier Combination with Confidence Values
[5] Weighted Average with Confidence Values

[DOI: 10.61186/jsdp.19.4.45]

روش‌های یاده‌شده به‌تفصیل در بخش‌های زیر توضیح داده‌شده‌اند.



شکل (۱): فلوچارت روش پیشنهادی
Figure (1): Flowchart of the proposed method

۳-۱- جمع‌آوری و پیش‌پردازش داده‌ها

برای دسته‌بندی داده‌ها، نیاز به تعداد زیادی داده با ویژگی‌های معنادار و دارای برچسب است. وجود برچسب داده‌ها برای انجام عمل دسته‌بندی ضروری است. برای این منظور، مجموعه‌داده ای شامل اطلاعات متقاضیان دریافت بیمه بیکاری جمع‌آوری‌شده است. این مجموعه‌داده‌ها متعلق به ۳۶۰۰۰ نفر تحت پوشش بیمه بیکاری سازمان تأمین اجتماعی است که ادعای بیکاری خود را طی پنج سال گذشته (از سال ۹۲ تا ۹۷) به سازمان تأمین اجتماعی اعلام کرده‌اند. هر رکورد در مجموعه‌داده ها شامل ۱۵ ویژگی است که مربوط به یک فرد متقاضی دارای بیمه بیکاری بوده و ادعای بیکاری کرده‌است. توجه شود که این داده‌ها دارای برچسب می‌باشند، بدین معنی که سازمان تأمین اجتماعی در قبل صحت یا عدم صحت ادعای بیکار شدن این افراد را توسط بازرسین خود بررسی کرده و وضعیت هر فرد را در پایگاه داده مشخص کرده است. به همین دلیل، نیازی به برچسب‌گذاری روی داده‌ها که کاری وقت‌گیر و طاقت‌فرسا است، نیست.

زیرمجموعه‌ای از مجموعه‌داده اولیه، در نشانی:
http://myweb.sabanciuniv.edu/rdehkharghani/files/2019/07/Dataset-Insurance-subset.xlsx
در دسترس همگان قرار دارد. جدول ۱ مشخصات مجموعه‌داده مورداستفاده در این پژوهش را به‌طور خلاصه نشان‌می‌دهد.

برای آماده‌سازی داده‌ها برای رده‌بندی، در مرحله پیش-پردازش اعمال زیر صورت‌می‌گیرد:

- حذف ویژگی‌هایی که دانش مفیدی را ارائه نمی‌دهند؛ در این مرحله از پیش‌پردازش، ویژگی‌هایی که اطلاعات مفیدی را برای دسته‌بندی در بر ندارند حذف می‌شوند. به‌عنوان‌مثال، نام پدرِ متقاضیان که تأثیری در دسته‌بندی ندارد از فهرست ویژگی‌ها حذف می‌شوند.
- تغییر قالب مقادیر برخی از ویژگی‌ها؛
مقادیر برخی ویژگی‌ها نظیر "سن" و "میزان درآمد" به قالب مناسب تبدیل می‌شوند. به‌عنوان‌مثال، سن افراد که برحسب سال، ماه و روز به‌عنوان ورودی داده‌شده‌است، به تعداد روز تبدیل می‌شود. همچنین مقادیر ویژگی‌هایی همانند "میزان تحصیلات" که به‌صورت نوع داده رشته‌ای داده‌شده‌است، به مقادیر عددی تبدیل می‌شود.
- حذف نویسه‌های غیرمجاز؛
در این مرحله، کاراکترهای غیرمتعارف که در پردازش داده‌ها مشکل ایجاد می‌کنند، حذف می‌شوند. به‌عنوان مثال، کاراکترهای "؟ " و "*" از مقادیر ویژگی‌ها حذف می‌شوند تا در پردازش‌های آینده مشکلی ایجاد نشود.
- جایگزینی نمونه‌هایی که دارای مقادیر گم‌شده هستند.
- فقدانِ داده‌ها یکی از مشکلات اصلی در مجموعه‌داده جمع‌آوری‌شده است. در داده‌های خام ممکن است، یک یا چند ویژگی فاقد مقدار باشند. این موضوع کیفیت دسته‌بندی را تحت‌تأثیر قرار می‌دهد. راه‌حل‌هایی از ساده تا پیچیده برای غلبه بر این مشکل و جانهی مقادیر گم‌شده وجود دارد. یکی از راه‌حل‌ها دسترسی به افراد متقاضی و کسب اطلاعات برای مقادیر گم‌شده است، اما این روش، کاری زمان‌بر و گاهی غیرممکن است.

جدول (۱): مجموعه‌داده مورد استفاده در این پژوهش قبل از پیش‌پردازش

Table (1): The data set used in this research before pre-processing

تعداد نمونه‌ها	تعداد ویژگی‌ها	تعداد رده‌ها	توزیع داده‌های رده
۳۶۰۰۰	۱۵	۲	(۲۷۶۳۲ ادعای درست و ۶۸۷۰ ادعای کذب)

در مقابل، روش‌های داده‌کاوی برای تجزیه و تحلیل داده‌ها و جای‌گذاری مقادیر گم‌شده از سرعت بیشتری برخوردار هستند. برخی از الگوریتم‌ها همانند درخت

[DOI: 10.61186/jsdp.19.4.45]

تصمیم، می‌توانند با مقادیر گم‌شده نیز مانند سایر مقادیر رفتار کنند و آن‌ها را در قوانین بیاورند. برخی دیگر از الگوریتم‌ها مانند شبکه‌های عصبی مصنوعی روی مجموعه‌داده هایی که مقادیر گم‌شده دارند، عملکرد مطلوبی از خود نشان نمی‌دهند. حذف رکوردهایی با مقادیر گم‌شده یکی از راه‌حل‌های ممکن برای مدیریت داده‌های گم‌شده است. این کار باعث می‌شود معادلاتِ آماری در مجموعه‌داده هایی با نمونه‌های زیاد حفظ شوند. به‌عبارت دیگر، حذفِ چند رکورد از مجموعه‌داده‌ها، به‌طور عمومی در این معادلات خللی وارد نمی‌کند. راه‌حلِ دیگر، حذف ویژگی‌هایی با مقادیر گم‌شده است، اما این روش در مسئله مورد بحث در این مقاله چندان منطقی نیست؛ زیرا ممکن است، ویژگی‌های تأثیرگذار درعمل رده‌بندی نیز حذف شوند. جایگزینی مقادیر گم‌شده با مقادیری مانند میانه[1] یا میانگین[2] مقادیرِ سایر رکوردها نیز امری متداول است. به‌عنوان مثال، می‌توان مقادیر گم‌شده برای ویژگی "سن" را، از طریق محاسبه میانگین مقادیر سن در سایر رکوردها جای‌گذاری کرد. البته این کار ممکن است موجب ایجاد داده‌های ناسازگار شود. برای مثال، مقدار ویژگی کد پستی را نمی‌توان با محاسبه میانگین مقادیر محاسبه کرد و میانگین یا میانه آن بی‌معنا خواهد بود. همچنین می‌توان به‌جای مقادیر گم‌شده، یک مقدار ثابت قرار داد. این کار نیز موجب تولید داده‌های غیرواقعی خواهد شد.

برای حل مشکل یادشده، استفاده از روش‌های داده‌کاوی می‌تواند بسیار مطلوب باشد. در این پژوهش، به‌منظور حل مشکل مقادیر گم‌شده و جانهی آن‌ها، از الگوریتم نزدیک‌ترین همسایه K (KNN)[3] استفاده می‌کنیم [16]، [17]. در این روش، ابتدا نزدیک‌ترین رکوردهای همسایه را به رکوردی که دارای مقادیر گم‌شده است، پیدا می‌کنیم. این کار با محاسبه فاصله اقلیدسی بین رکورد موردنظر X^i و هر کدام از k رکورد همسایه $N = \{X^1, X^2, \ldots, X^k\}$ صورت‌می‌گیرد.

$$d(X^i, X^j) = \left\| \frac{\sum_{l=1}^{n} r_l^i . r_l^j . \|x_l^i, x_l^j\|}{\sum_{l=1}^{n} r_l^i . r_l^j} \right\| ; \begin{cases} r_l^i = 0 & x_l^i = \varnothing \\ r_l^i = 1 & x_l^i \neq \varnothing \\ r_l^j = 0 & x_l^j = \varnothing \\ r_l^j = 1 & x_l^j \neq \varnothing \end{cases} \quad (۱)$$

در رابطه بالا، X^i بیان‌گر رکوردی با مقادیر گم‌شده، X^j رکورد همسایه i، r_l^i و r_l^j به‌ترتیب بیانگر ضرایب ویژگی‌های x_l^i و x_l^j است؛ n بیان‌گر تعداد ویژگی‌ها / فیلدهای هر رکورد است. پس از پیدا کردن نزدیک‌ترین همسایه‌ها، و مرتب‌کردن آن‌ها برحسب میزان تشابه، مقدار گم‌شده به‌صورت زیر محاسبه می‌شود:

(۲)

$$x_l^{mis} = \sum_{v=1}^{k} w_v^j . x_l^v$$

در رابطه بالا، x_l^{mis} بیانگر مقدار تخمین‌زده‌شده برای مقدار گم‌شده ویژگی l است و w_v^j بیان‌گر وزن نرمال‌شده نمونه v است. w_v^j به‌صورت زیر محاسبه می‌شود:

(۳)

$$w_v^j = \frac{\frac{1}{d(x^i, x^v)}}{\sum_{v=1}^{k} \frac{1}{d(x^i, x^v)}}$$

گفتنی است که در آزمایش‌های صورت‌گرفته، مقدار k برابر ۵ در نظر گرفته شده‌است. این مقدار به‌صورت تجربی و بر اساس نتایج آزمایش‌ها انتخاب شده‌است. برای این منظور، ابتدا مقدار k برابر یک در نظر گرفته شده و هر بار، یک واحد به آن اضافه می‌شود. برای هر مقدار k، کارایی الگوریتم محاسبه می‌شود و در صورتی که کارایی است پس از افزایش مقدار k کاهش پیدا کند، مقدار قبلی آن به‌عنوان مقدار بهینه انتخاب می‌شود.

۳-۲- استخراج ویژگی‌ها

مجموعه‌داده اصلی شامل پانزده ویژگی است. برخی از این ویژگی‌ها تأثیر زیادی در رده‌بندی داده‌ها ندارند. به‌منظور تعیین ویژگی‌های مهم در رده‌بندی، از الگوریتم انتخاب ویژگی پیشرو[4] [18] استفاده کرده‌ایم.

در روش انتخاب پیشرو، هر یک از ویژگی‌ها تک‌تک به مجموعه‌ویژگی‌ها (با شروع از مجموعه تهی) اضافه شده و کارایی است رده‌بندی با افزودن این ویژگی اندازه گرفته‌می‌شود؛ در صورتی که افزودن ویژگی یادشده باعث افزایش بیشتر کارایی است شود، این ویژگی به مجموعه ویژگی‌ها افزوده شده و در غیر این‌صورت، نادیده گرفته‌می‌شود. این کار برای تمام ویژگی‌ها تکرار شده و درنهایت، زیرمجموعه‌ای از ویژگی‌ها به‌عنوان مهم‌ترین ویژگی‌ها انتخاب می‌گردد.

پس از اعمال روش انتخاب پیش‌رو بر روی مجموعه‌داده اولیه، ۹ ویژگی از بین ۱۵ ویژگی، به‌عنوان ویژگی‌های تأثیرگذار انتخاب شدند. جدول ۲ این ویژگی‌ها را به‌همراه مشخصات آن‌ها نشان‌می‌دهد. توضیح هر ویژگی به‌اختصار در زیر داده‌شده‌است:

[1] Median
[2] Mean
[3] K-Nearest Neighbors (KNN)
[4] Forward feature selection

[DOI: 10.61186/jsdp.19.4.45]

کلید اصلی: این ویژگی به‌عنوان کلید هر رکورد عمل کرده و تأثیری در رده‌بندی ندارد.

سن: این ویژگی سن یک فرد را برحسب روز نشان‌می‌دهد. به‌عنوان مثال، فردی که ۲۸ سال و ۲ ماه و ۱۵ روز از عمر خود سپری کرده است، سن این فرد ۱۰۲۹۵ روز در نظر گرفته‌می‌شود.

جنسیت: این ویژگی بیان‌گر جنسیت متقاضی با مقدار "مؤنث" یا "مذکر" است. در هنگام پیش‌پردازش، مقادیر ویژگی "جنسیت" به فرم دودویی تبدیل می‌شود به‌گونه‌ای که جنسیت "مذکر" با مقدار بیتی "۱" و جنسیت "مؤنث" با مقدار بیتی "۰" جایگزین شده‌است.

وضعیت تأهل: این ویژگی بیتی دارای مقادیر "مجرد" و "متأهل" است. در هنگام پیش‌پردازش به‌ترتیب برای مجرد و متأهل، مقادیر عددی صفر و یک در نظر گرفته شده‌است.

تحصیلات: این ویژگی میزان تحصیلات یک فرد را نشان‌می‌دهد که یک داده دودویی محسوب‌می‌شود و دارای مقادیر ۱ (معادل باسواد") و ۰ (معادل بی‌سواد) است.

تعداد عائله: این ویژگی تعداد افراد تحت تکفل فرد بیمه‌شده را که شامل فرزندان، همسر، پدر و مادر هستند، با یک عدد صحیح نشان‌می‌دهد.

سابقه پرداخت بیمه: این ویژگی سابقه پرداخت حق بیمه را برحسب ماه تعیین می‌کند. برای مثال اگر فردی تاکنون سه سال و پنج ماه حق بیمه پرداخت کرده است، عدد واردشده در این ویژگی برای او ۴۱ خواهد بود (۳×۱۲+۵).

مدت استحقاق: این ویژگی تعداد ماه‌هایی که سازمان تأمین‌اجتماعی تعهد پرداخت حق بیمه به فرد بیکار را دارد، نشان‌می‌دهد. مقدار این ویژگی توسط سازمان، با توجه به مقدار پرداختی و تعداد ماه‌های پرداخت بیمه توسط فرد بیمه‌شده در زمان اشتغال، تعیین می‌شود.

میزان پرداخت: این ویژگی میزان پرداخت حق‌بیمه به سازمان تأمین‌اجتماعی در هر ماه توسط بیمه‌شده یا کارفرمای او را نشان‌می‌دهد.

دلیل بیکاری: این ویژگی دلیل بیکاری فرد متقاضی را نشان‌می‌دهد که عبارت‌اند از: اخراج موجه، عدم نیاز، اتمام پروژه، فسخ قرارداد، عدم‌توانایی، تعطیلی کارگاه، حوادث غیرمترقبه، تغییر ساختار وزارت کار، جابجایی کارگاه، فصلی‌بودن کارگاه، اتمام قرارداد، نوسازی صنایع و غیره. در مرحله پیش‌پردازش، به‌جای مقادیر رشته‌ای بالا از معادل عددی آن‌ها استفاده می‌شود.

در انتهای رکورد مربوط به هر فرد، یک برچسب که صحیح یا کذب‌بودن ادعای بیکاری فرد را نشان‌می‌دهد اضافه می‌شود.

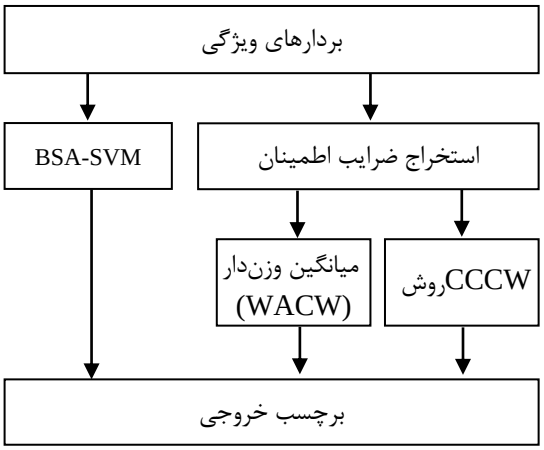
جدول (۲): ویژگی‌های افراد متقاضی بیمه بیکاری

Table (2): Features of unemployment insurance applicants

ردیف	نام ویژگی	نوع داده	واحد	پیوستگی
۱	کلید اصلی	عدد صحیح	-	گسسته
۲	سن	عدد صحیح	روز	گسسته
۳	جنسیت	بیتی	-	گسسته
۴	وضعیت تأهل	بیتی	-	گسسته
۵	تحصیلات	بیتی	-	گسسته
۶	تعداد عائله	عدد صحیح	تعداد	گسسته
۷	سابقه پرداخت بیمه	عدد صحیح	ماه	گسسته
۸	ماه‌های استحقاق	عدد صحیح	ماه	گسسته
۹	میزان پرداخت ماهانه	عدد اعشاری	ریال	پیوسته
۱۰	دلیل بیکاری	اسمی	-	گسسته

۳-۳- رده‌بندی

اغلب اطلاعات موجود در پایگاه داده‌ها توزیع‌های ناشناخته یا پیچیده‌ای دارند که به‌راحتی نمی‌توان این توزیع‌ها را شناسایی کرده و مورد استفاده قرار داد؛ بنابراین، برای تحلیل داده‌های موجود در پایگاه داده‌ها، استفاده از روش‌هایی که نیاز به دانستن توزیع متغیرها ندارد از اهمیت خاصی برخوردار است. در این مرحله، چند رده‌بند با داده‌های برچسب‌دار آموزش داده شده و سپس با هم ترکیب می‌شوند. شکل ۲ نمای کلی روش پیشنهادی برای رده‌بندی داده‌ها را نشان‌می‌دهد.



شکل (۲): شمای کلی روش پیشنهادی برای رده‌بندی داده‌ها

Figure (2): Outline of the proposed method for data classification

[DOI: 10.61186/jsdp.19.4.45]

با توجه به اینکه هر رده‌بند دارای نقاط قوت و ضعف خاص خود بوده و میزان موفقیت رده‌بندها برای دسته‌بندی انواع مختلف داده‌ها متفاوت است، ایده ترکیب رده‌بندها از سال‌ها قبل مدنظر پژوهش‌گران قرار گرفته است. رده‌بندها می‌توانند با روش‌های ساده مانند رأی بیشینه [19] با هم ترکیب شوند. انتقادی که از این روش می‌شود این است که تأثیر رأی هر رده‌بند در نتیجه نهایی با بقیه رده‌بندها برابر است، درحالی‌که میزان کارایی آن‌ها با هم متفاوت است. برای رفع این مشکل، میانگین وزن‌دار [20] به‌عنوان روش دیگر مطرح، که در این مقاله نیز استفاده شده‌است. گزینه دیگر می‌تواند استفاده از ضرایب اطمینان رده‌بندها برای تخمین برچسب یک شیء داده‌ای باشد. ویژگی این روش این است که به‌جای استفاده از مقادیر گسسته برچسب رده‌ها از مقادیر فازی احتمال تعلق شیء داده‌ای به هر رده استفاده می‌کند. با توجه به اینکه اشیای داده‌ای به‌طورمعمول با احتمال ۱۰۰٪، عضو یک رده تخمین زده نمی‌شوند، استفاده از مقادیر اطمینان هر رده‌بند در مورد برچسب بک شیء داده‌ای، منطقی به نظر می‌رسد. این روش نیز در ادامه استفاده شده است. پژوهش‌گران از روش‌های پیچیده‌تری مانند روش دمپستر-شفر [21] یا روش بوردا کانت [19] نیز برای رده‌بندی استفاده کرده‌اند.

هر یک از سه روش پیشنهادی در زیر به‌تفصیل توضیح داده‌شده‌اند.

۳-۳-۱- روش دسته‌بندی BSA-SVM

در روش BSA-SVM، از ترکیب الگوریتم بهینه‌سازی جستجوی عقب‌گرد (BSA) [22] و الگوریتم ماشین بردار پشتیبان (SVM) استفاده شده‌است. شکل ۳ روندنمای روش پیشنهادی BSA-SVM را نشان‌می‌دهد. هدف الگوریتم SVM، یافتن یک ابرصفحه بهینه برای جداسازی داده‌ها است.

اگر $D=\{x_i,d_i\}_{i=1}^{n}$، که در آن x_i رکورد ورودی، d_i برچسب داده x_i و n تعداد رکوردها باشد، تابع تقریب SVM به‌صورت زیر تعریف شده‌است:

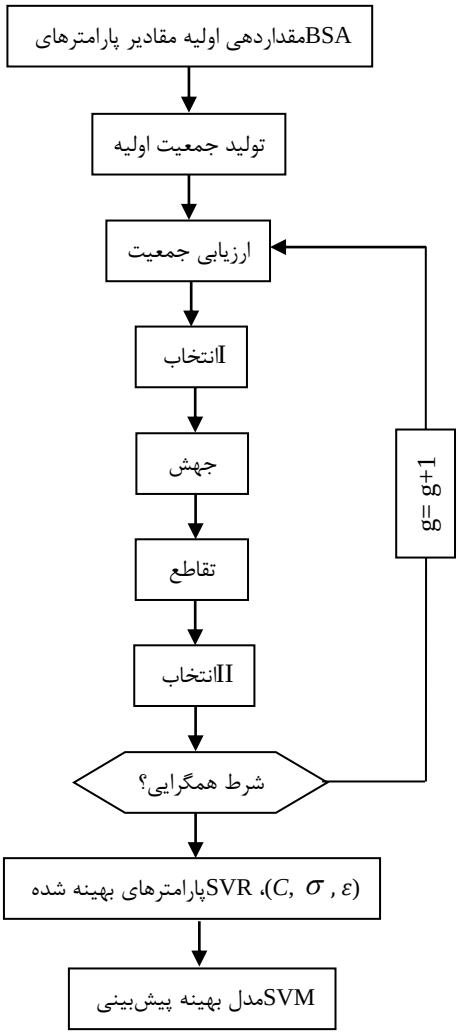
$$f(x)=w\,\varphi(x)+b \qquad (۴)$$

$$R_{SVM}(C)=\frac{1}{2}\|w\|^2+C\,\frac{1}{n}\sum_{i=1}^{n}L(x_i,d_i) \qquad (۵)$$

فضای ویژگی چندبعدی برای نگاشت بردار x_i، w بردار نرمال سازی و b مقدار ثابت است. C عددی صحیح و مثبت است که عامل تعیین جریمه در هنگام رخ دادن خطای واسنجی مدل است. φ تابع کرنل و n تعداد نمونه‌ها است. تابع $C\,\frac{1}{n}\sum_{i=1}^{n}L(x_i,d_i)$ ریسک تجربی را نشان‌می‌دهد. با کمینه‌سازی تابع ریسک، مقادیر w و b به‌صورت زیر محاسبه می‌شوند:

$$\text{minimize}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}(\zeta_i+\zeta_i^*) \qquad (۶)$$

$$\text{subject to}\begin{cases} y_i-\langle w,x_i\rangle-b & \le \varepsilon+\zeta_i \\ \langle w,x_i\rangle-b-y_i & \le \varepsilon+\zeta_i^* \\ \zeta_i,\zeta_i^* & \ge 0 \end{cases}$$



شکل (۳): روندنمای روش BSA-SVM

Figure (3): Flowchart of BSA-SVM method

$\varphi(x)$ و دو مشخصه ζ_i و ζ_i^* متغیرهای کمبود هستند که حد بالا و حد پایین خطای آموزش مرتبط با مقدار خطای مجاز ε را مشخص می‌کنند.

رابطه (۴) را می‌توان با معرفی ضرایب لاگرانژ و محدودیت‌های بهینه به‌صورت زیر حل کرد:

[DOI: 10.61186/jsdp.19.4.45]

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x,x_i)+b \qquad (۷)$$

$K(x,x_i)=\varphi(x_i)\times\varphi(x_j)$ نشان دهنده تابع کرنل است. در روش پیشنهادی، از تابع شعاعی پایه[1] (RBF) به‌عنوان تابع کرنل[2] استفاده شده‌است. دلیل استفاده از تابع کرنل RBF، کارایی و دقت بالای آن نسبت به سایر توابع کرنل است.

تابع کرنل RBF به‌صورت زیر تعریف شده‌است:

$$K(x,x_i)=\exp(-\|x_i-x\|/2\sigma^2) \qquad (۸)$$

کارایی الگوریتم SVM با تابع کرنل RBF، به تعیین دقیق پارامترهای آن، یعنی C، σ و ε بستگی دارد. برای تعیین مقادیر بهینه این پارامترها، از الگوریتم بهینه‌سازی BSA استفاده کرده‌ایم. این الگوریتم تأثیر مثبتی بر روی آموزش SVM دارد و دقت دسته‌بندی را افزایش می‌دهد.

جزئیات روش BSA-SVM به‌صورت زیر است:

تولید جمعیت اولیه

هر عضو جمعیت شامل مقادیر نامزد برای پارامترهای C، σ و ε است. این مقادیر، اعداد اعشاری هستند که به‌صورت تصادفی تولید می‌شوند.

محاسبه تابع شایستگی

شایستگی هر عضو از جمعیت با استفاده از معیار خطای میانگین مربعات (MSE) حاصل از روش اعتبارسنجی متقابل k-fold با مقدار k=5 سنجیده می‌شود. تابع هدف به‌صورت زیر تعریف شده‌است:

$$f(x)=\frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i-X_i)^2 \qquad (۹)$$

X_i نشان‌دهنده مقدار داده مشاهده‌شده، و $\hat{X}_i$ نشان‌دهنده مقدار پیش‌بینی‌شده به‌وسیله الگوریتم SVM بوده و n بیان‌گر تعداد داده‌ها در مجموعه‌داده آموزشی است. بدیهی است که اعضای جمعیت با مقدار کوچک‌تر MSE، دارای میزان شایستگی بیشتری هستند.

عمل‌گرهای الگوریتم BSA

تا زمانی که شرایط خاتمه الگوریتم ارضا نشده باشد، چهار عمل‌گر انتخاب-I، جهش، تقاطع و انتخاب-II به‌صورت تکراری برای به‌هنگام‌سازی اعضای جمعیت اعمال می‌شوند. شرط خاتمه، تعداد دفعات محاسبه تابع هدف است که در اینجا برابر با هزار در نظر گرفته شده‌است.

۳-۳-۲- روش ترکیب رده‌بندها با ضرایب اطمینان (CCCW)

این روش دارای دو مرحله است: ۱) استخراج ضرایب اطمینان هر رده‌بند و ۲) ترکیب رده‌بندها با این ضرایب. در مرحله اول، پنج رده‌بند مختلف به‌صورت جداگانه یک رکورد (شیء داده‌ای) را رده‌بندی می‌کنند. این رده‌بندها عبارتند از جنگل تصادفی، درخت تصمیم، ماشین بردار پشتیبان، شبکه‌های عصبی مصنوعی (پرسپترون چندلایه) و رگرسیون ترابری؛ سپس ضرایب اطمینان[3] هر رده‌بند برای تخمین برچسب این رکورد با کمک ابزار وکا[4] استخراج می‌شوند. در مرحله دوم، مجموعه ضرایب اطمینان به‌دست‌آمده از پنج رده‌بند به‌عنوان ویژگی به یک رده‌بند دیگر (SVM) داده‌می‌شود تا یک رده‌بندی دودویی دیگر انجام شود. درنهایت می‌توان خروجی این رده‌بند را به‌عنوان برچسب نهایی رکورد داده در نظر گرفت.. با توجه به اینکه دو رده در مسئله فعلی وجود دارد (ادعای درست بیکاری، و ادعای دروغین بیکاری)، خروجی مرحله یک برای هر رده‌بند، دو مقدار بین صفر و یک که مجموع آن‌ها برابر یک است خواهد بود. هر کدام از این ضرایب اطمینان، میزان اطمینان یک رده‌بند را در مورد برچسب تخمینی یک رکورد نشان‌می‌دهد. به‌عنوان مثال، رده‌بند پرسپترون چندلایه برای یک رکورد، مقادیر اطمینان (C0,C1)=(۰.۳۴, ۰.۶۶) را تولید کرده است. این بدین معنی است که از نظر این رده‌بند، برچسب شیء داده‌ای یادشده، به‌احتمال ۰.۶۶، ۱ و به‌احتمال ۰.۳۴، ۰ است. با توجه به استفاده از پنج رده‌بند، و دو ضریب اطمینان برای هر رده‌بند، ترکیب رده‌بندها دارای ده ویژگی (فیچر) خواهد بود. به‌عبارت دیگر، در مرحله دوم، یک رده‌بندی با ده ویژگی عددی با مقادیر بین صفر. یک و یک خروجی دودویی انجام‌می‌شود.

۳-۳-۴ - روش میانگین وزن‌دار با ضرایب اطمینان (WACW)

با توجه به اینکه تأثیر هر رده‌بند بر روی کارایی است نهایی متفاوت است، روش دیگری به‌نام میانگین وزن‌دار نیز استفاده شده و میزان کارایی آن در بخش ۴ با روش CCCW مقایسه شده‌است. در این روش، یک وزن

[1] Radial Basis Function
[2] Kernel
[3] Confidence values
[4] Weka

[DOI: 10.61186/jsdp.19.4.45]

برای هر رده‌بند در نظر گرفته شده و میانگین وزنی رده‌بندها بر طبق فرمول (۴) محاسبه‌شده است.

$$W = \frac{\sum_{i=1}^{n} w_i c_i}{\sum_{i=1}^{n} w_i}, \quad w_i = R_i Acc_i \qquad (۱۰)$$

در این روش نیز مانند روش CCCW، از ضرایب اطمینان (c_i ها) رده‌بندها برای تخمین برچسب یک شیء داده‌ای استفاده شده است. این برچسب با یک الگوی از پیش‌تعریف‌شده که در رابطه (۱۰) مشاهده می‌شود، تخمین زده می‌شود. خروجی این رابطه به‌ازای هر شیء داده‌ای یک‌بار برای رده صفر و یک‌بار برای رده یک محاسبه، سپس این شیء داده‌ای با شماره رده‌ای که مقدار خروجی بزرگ‌تری در این رابطه داشته‌باشد، برچسب‌گذاری می‌شود. این خروجی عددی بین صفر و یک بوده و به معنی میزان اطمینان سامانه رده‌بندی در مورد برچسب شیء داده‌ای ورودی است. مقادیر w_i وزن‌های رده‌بندهای مختلف هستند. در این وزن‌ها، Acc_i دقت رده‌بند i در رده‌بندی داده‌های اعتبارسنجی[1] با روش 5-fold cross validation بوده و مقدار R_i رتبه این رده‌بند بین بقیه رده‌بندها با توجه به میزان موفقیت آن برای تخمین برچسب داده‌ها است. این رتبه‌ها به‌صورت زیر در رابطه (۱۰) استفاده شده‌است:

R (SVM, RF, MP, LR, DT) =(۵،۴،۳،۲،۱)

این ضرایب نشان‌دهنده میزان موفقیت هر رده‌بند با توجه به معیار دقت در تخمین برچسب صحیح یک شیء داده‌ای به‌تنهایی است. رده‌بند ماشین بردار پشتیبان با ضریب ۵، بیشترین و درخت تصمیم با ضریب ۱، کمترین موفقیت را در بین رده‌بندهای دیگر کسب کرده‌اند. بدیهی است که تأثیر رأی رده‌بند کاراتر باید بیشتر از بقیه رده‌بندها باشد. این تأثیر هم به‌وسیله رتبه رده‌بند (R_i) و هم به‌وسیله دقت آن (Acc_i) روی رأی نهایی سامانه پیشنهادی اعمال می‌شود.

۴- ارزیابی روش پیشنهادی

۴-۱- معیارهای کارایی

کارایی روش پیشنهادی با استفاده از سه معیار صحت[2]، بازخوانی[3] و دقت[4] ارزیابی شده‌است. با درنظرداشتن جدول ۳ به‌عنوان ماتریس درهم‌ریختگی[5]، معیار کارایی صحت (P) به‌صورت زیر تعریف می‌شود [23]:

$$P = \frac{TP}{TP + FP} \qquad (۱۱)$$

در این رابطه، مقدار TP بیانگر تعداد بیمه‌شدگانی است که مشمول بیمه بیکاری هستند و نتیجه پیش‌بینی نیز این موضوع را تأیید می‌کند. مقدار FP نیز بیان‌گر تعداد بیمه‌شدگانی است که مشمول بیکاری نیستند، ولی مدل پیش‌بینی آن‌ها را مشمول بیکاری نشان‌می‌دهد.

معیار بازخوانی (R) به‌صورت زیر تعریف شده‌است [23]:

$$R = \frac{TP}{TP + FN} \qquad (۱۲)$$

مقدار FN بیانگر تعداد بیمه‌شدگانی است که مشمول بیمه بیکاری هستند، ولی در پیش‌بینی است، غیرمشمول شناخته‌شده‌اند.

معیار دقت (A) نیز به‌صورت زیر تعریف شده‌است [23]:

$$A = \frac{TP + TN}{TP + FP + TN + FN} \qquad (۱۳)$$

در این رابطه، TN بیانگر تعداد بیمه‌شدگانی است که مشمول بیمه بیکاری نمی‌شوند و تخمین سامانه نیز همین موضوع را نشان‌می‌دهد. توجه شود که معیارهای صحت و بازخوانی برای یک رده خاص محاسبه‌شده، ولی معیار دقت برای کل سامانه محاسبه می‌شود.

جدول (۳): فهرست نقش‌های معنایی

Table (3): Confusion matrix

		رده پیش‌بینی‌شده	
		مثبت	منفی
رده واقعی	منفی	FP	TN
	مثبت	TP	FN

۴-۲-مجموعه‌داده

به‌منظور ارزیابی مناسب روش پیشنهادی، زیرمجموعه‌ای از مجموعه‌داده اولیه ایجاد شده‌است. مجموعه ایجادشده شامل ۱۰۰۰۰ رکورد است که در آن ۵۰۰۰ رکورد به‌صورت تصادفی از موارد مثبت (ادعای صحیح) و ۵۰۰۰ رکورد به‌صورت تصادفی از موارد منفی (ادعای نادرست) انتخاب شده‌است. جدول ۴ مشخصات این مجموعه‌داده را به‌اختصار نشان‌می‌دهد. این مجموعه‌داده توازن مناسبی

[1] Validation data
[2] Precision
[3] Recall
[4] Accuracy

[DOI: 10.61186/jsdp.19.4.45]

بین نمونه‌های مثبت و منفی برقرار می‌کند که این توازن می‌تواند اعتبار نتایج به‌دست‌آمده را افزایش دهد.

جدول (۴): مجموعه‌داده مورد استفاده در آزمایش‌ها

Table (4): Data set used in experiments

تعداد نمونه‌ها	تعداد ویژگی‌ها	تعداد رده‌ها	توزیع داده‌ها
۱۰۰۰۰	۹	۲	(۵۰۰۰ ادعای درست و ۵۰۰۰ ادعای نادرست)

۵- نتایج

جدول ۵ نتایج به‌دست‌آمده توسط روش‌های رده‌بندی را بر حسب معیارهای کارایی نشان‌می‌دهد. از نظر معیارهای ارزیابی مذکور، ترتیب کارایی الگوریتم‌ها به‌صورت:

$$BSA-SVM > CCCW > WACW > SVM > RF > MP > DT > LR$$

است. نتایج گویای آن است که روش پیشنهادی BSA-SVM کارایی بیشتری را در مقایسه با سایر روش‌ها دارد. همچنین، ترکیب رده‌بندها، کارایی رده‌بندی را افزایش داده‌است. دلیل استفاده از هر کدام از رده‌بندها در مدل ترکیبی به شرح زیر است:

- رگرسیون ترابری کارایی بالا در تعمیم دقت سامانه دارد.
- ماشین بردار پشتیبان به‌دلیل قدرت بالای آن در رده‌بندی‌های دودویی استفاده شده‌است.
- جنگل تصادفی از کارایی (دقت) بیشتری نسبت به بیشتر رده‌بندهای دیگر برخوردار است.
- شبکه‌های عصبی ابزاری تحلیلی و آموزش‌پذیر هستند و به‌دلیل ماهیت پردازش موازی از سرعت و دقت مطلوبی برخوردار هستند.
- درخت تصمیم قابلیت بالایی در تفسیر نتایج داشته و همچنین قابل‌فهم برای انسان است.

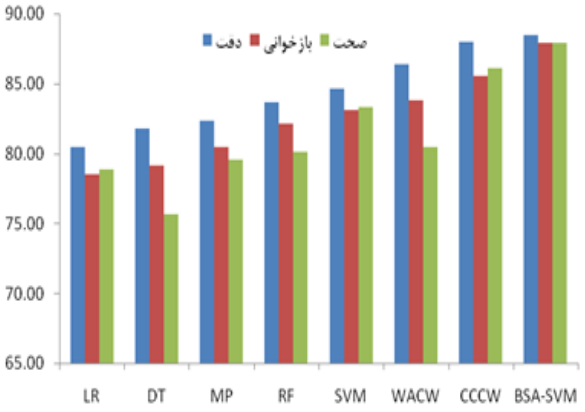
جدول (۵): نتایج به‌دست‌آمده از روش‌های پیشنهادی بر روی مجموعه‌داده آزمون

Table (5): Results obtained from the proposed methods on the test dataset

روش رده‌بندی	صحت	بازخوانی	دقت
درخت تصمیم (DT)	۰/۸۱.۷۶	۰/۷۶.۱۶	۰/۷۶.۶۶
رگرسیون (LR)	۰/۸۰.۴۳	۰/۷۸.۵۴	۰/۷۸.۸۶
شبکه‌های عصبی (MP)	۰/۸۲.۳۲	۰/۸۰.۴۴	۰/۷۹.۵۲
جنگل تصادفی (RF)	۰/۸۳.۶۵	۰/۸۳.۱۲	۰/۸۰.۰۹
ماشین بردار پشتیبان (SVM)	۰/۸۴.۶۲	۰/۸۳.۰۹	۰/۸۳.۲۹
روش میانگین وزن‌دار (WACW)	۰/۸۶.۳۹	۰/۸۳.۷۷	۰/.۸۵.۷۱
روش ترکیب رده‌بندها با ضرایب اطمینان (CCCW)	۰/۸۸.۰۱	۰/۸۵.۵۵	۰/۸۶.۱۱
روش BSA-SVM	۰/۸۸.۴۴	۰/۸۷.۹۳	۰/۸۷.۹۳

همچنین به‌جهت وضوح بیشتر، مقایسه کارایی روش‌های ارائه‌شده از نظر معیارهای دقت، بازخوانی و صحت در شکل ۴ به تصویر کشیده شده‌است.

بدیهی است که هر کدام از ویژگی‌ها تأثیر متفاوتی بر روی رده‌بندی داده‌ها دارند. به‌منظور تعیین میزان تأثیر هر ویژگی بر روی کارایی رده‌بندی، روش رده‌بندی BSA-SVM بر مبنای هر یک از ویژگی‌ها بررسی‌شده و تأثیر هر ویژگی بر حسب معیار دقت در جدول ۶ نشان داده‌شده‌است.



شکل (۴): مقایسه کارایی روش‌های رده‌بندی

Figure (4): Comparing the performance of classification methods

جدول (۶): تأثیر هر ویژگی بر رده‌بندی داده‌ها با روش BSA-SVM

Table (6): The effect of each feature on data classification with BSA-SVM

شماره	نام ویژگی	دقت به‌دست‌آمده
۱	سن	۰/۷۶.۵۴
۲	جنسیت	۰/۶۸.۲۱
۳	وضعیت تأهل	۰/۶۳.۰۹
۴	تحصیلات	۰/۷۰.۲۳
۵	تعداد عائله	۰/۶۳.۱۶
۶	سابقه پرداخت بیمه	۰/۶۹.۹۷
۷	ماه‌های استحقاق	۰/۷۲.۹۱

[DOI: 10.61186/jsdp.19.4.45]

۸	میزان پرداخت ماهانه	۰/۷۳.۸۵
۹	دلیل بیکاری	۰/۷۶.۶۹
۱۰	قطع مقرری	۰/۷۱.۰۵

همان‌گونه که در جدول ۶ آمده است، ویژگی‌های شماره ۱ و ۹ (سن و دلیل بیکاری)، در مقایسه با سایر ویژگی‌ها، بیشترین تأثیر را بر روی عمل رده‌بندی دارند. این نتیجه می‌تواند به این دلیل باشد که افرادی که سن بالاتری دارند، معمولاً کمتر از بقیه ادعای دروغ مبنی بر بیکاری به سازمان تأمین اجتماعی ارائه می‌دهند. از طرف دیگر، کسانی که به‌دلایل خارج از اختیار خود مانند تعطیل‌شدن محل کار خود بیکار شده‌اند، نسبت به بقیه افراد که به‌دلایل دیگری بیکار شده‌اند، راست‌گوتر هستند.

برای ارائه دقیق‌تر نتایج و بررسی بیشتر آن‌ها، ماتریس درهم‌ریختگی رده‌بندی داده‌ها با روش BSA-SVM به‌همراه مقادیر صحت و بازخوانی هر رده به‌طور جداگانه در جدول‌های ۷ و ۸ ارائه شده‌است. همان‌طور که در این جداول مشاهده می‌شود، کارایی سامانه در تشخیص افراد **مشمول بیمه بیکاری** نسبت به افراد غیر **مشمول بیمه بیکاری**، مقداری بیشتر است. این پدیده می‌تواند به‌دلیل کیفیت و وضوح بالاتر داده‌های مربوط به افراد مشمول بیمه بیکاری در مرحله آموزش رده‌بندها باشد.

جدول (۷): ماتریس درهم‌ریختگی رده‌بندی داده‌ها

Table (7): Confusion matrix for classification of data

	مشمول بیمه بیکاری (واقعی)	غیر مشمول بیمه بیکاری (واقعی)
مشمول بیمه بیکاری (پیش‌بینی‌شده)	۴۱۰۵	۳۱۲
غیرمشمول بیمه بیکاری (پیش‌بینی‌شده)	۸۹۵	۴۶۸۸

جدول (۸): مقادیر معیارهای دقت و بازخوانی برای رده‌ها

Table (8): Precision and readability metrics values for classes

رده	صحت	بازخوانی
مشمول بیمه بیکاری	۹۲.۹۳	۰/۸۲.۱۰
غیرمشمول بیمه بیکاری	۰/۸۳.۹۶	۰/۹۳.۷۶

جدول ۹ مقایسه‌ی نتایج روش پیشنهادی با دو روش ارائه شده توسط سایر پژوهش‌گران را نشان‌می‌دهد.

روش اول، یک مدل ترکیبی خبره فازی برای رتبه‌بندی اعتباری مشتریان مؤسسه مالی و اعتباری قوامین است که هدف آن بررسی اعطای تسهیلات اعتباری متناسب با هر درجه از ریسک اعتباری مشتریان است [24]. این روش مبتنی بر رویکرد ترکیبی انتخاب ویژگی، الگوریتم ژنتیک و سیستم خبره فازی است. در این روش، از الگوریتم ژنتیک برای انتخاب ویژگی‌های رتبه‌بندی اعتباری استفاده شده‌است. ویژگی‌های منتخب، ورودی‌های سامانه خبره فازی در نظر گرفته شده و در ساخت قوانین رتبه‌بندی اعتباری مورد استفاده قرار گرفت.

روش دوم بر اساس یک مدل ترکیبی از الگوریتم بهینه‌سازی رقابت استعماری و شبکه عصبی برای افزایش دقت دسته‌بندی در ارزیابی و سنجش ریسک اعتباری مشتریان بانکی ارائه شده‌است [25]. ما این دو روش را بر روی مسئله اعتبارسنجی بیمه بیکاری اعمال کرده و با روش پیشنهادی مقایسه کردیم.

همان‌گونه که در جدول شماره ۹ نشان داده‌شده‌است، روش پیشنهادی در مقایسه با سایر روش‌های ارائه‌شده به کارایی بالاتری دست یافته‌است. این موضوع نشان‌می‌دهد که ترکیب دسته‌بندها به‌صورت لایه‌ای، کارایی دسته‌بندی را افزایش داده‌است.

جدول (۹): مقایسه کارایی روش پیشنهادی با روش‌های ترکیبی رده‌بندی

Table (9): Comparison of the perfrormane of the proposed method with the combined classification methods

روش رده‌بندی	صحت	بازخوانی	دقت
روش ترکیبی خبره فازی [24]	۰/۸۳	۰/۸۱	۰/۷۹
روش ترکیبی الگوریتم رقابت استعماری و شبکه‌های عصبی [25]	۰/۸۱	۰/۸۳	۰/۸۰
روش پیشنهادی (BSA-SVM)	۰/۸۸	۰/۸۷	۰/۸۷

جدول شماره ۱۰، زمان اجرای الگوریتم‌ها را بر حسب ثانیه نشان‌می‌دهد. همان گونه که در این جدول نشان داده‌شده‌است، روش ترکیبی پیشنهادی، از زمان اجرای بیشتری در مقایسه با مؤلفه‌های خود یعنی رگرسیون ترابری، درخت تصمیم، شبکه‌های عصبی، جنگل تصادفی و بیز ساده برخوردار است که این مورد به‌دلیل سربار ناشی از ترکیب رده‌بندها امری بدیهی است. در مقابل، کارایی رده‌بندی ترکیبی از کارایی رده‌بندی کننده‌های منفرد بیشتر است. همچنین روش پیشنهادی با

[DOI: 10.61186/jsdp.19.4.45]

وجود اینکه زمان کمتری نسبت به روش‌های موجود دیگر در جدول ۱۰ برای اجراشدن استفاده می‌کند، از کارایی بالاتری نسبت به آن‌ها برخوردار است.

۶- نتیجه‌گیری

در این مقاله، مسئله ادعای دریافت بیمه بیکاری توسط افراد تحت پوشش سازمان تأمین اجتماعی بررسی‌شده، و روشی مبتنی بر یادگیری نظارتی و ترکیب رده‌بندها ارائه شده‌است. نتایج آزمایش‌ها نشان‌می‌دهد که روش پیشنهادی از کارایی مطلوبی در مقایسه با سایر روش‌ها برخوردار است. سامانه ارائه‌شده در این مقاله می‌تواند به‌عنوان یک ابزار کمکی برای تصمیم‌گیری در سازمان تأمین اجتماعی استفاده شود. به‌عبارت دیگر، سازمان تأمین اجتماعی می‌تواند به‌جای صرف زمان و هزینه زیاد برای بررسی دقیق و کامل تمامی افراد مدعی بیکاری، تنها تعدادی کمی از افرادی را که سامانه آن‌ها را متقلب تعیین کرده است، مورد کنکاش و بررسی قرار دهد، و برای بررسی صحت ادعای افراد دیگری که راست‌گو تخمین زده‌شده‌اند، زمان و دقت کمتری به خرج دهد. سامانه آموزش داده‌شده می‌تواند پس از استخراج ویژگی‌های یک فرد جدید (به‌عنوان داده آزمون)، درست یا نادرست‌بودن ادعای بیکاری فرد را تخمین بزند. استفاده از ویژگی‌های بیشتر، از جمله سوابق کیفری افراد یا وضعیت مالی آن‌ها به‌عنوان کار آینده در نظر گرفته شده‌است.

جدول (۱۰): زمان اجرای روش‌های رده‌بندی بر حسب ثانیه

Table (10): Execution time of classification methods in seconds

روش رده‌بندی	زمان اجرا (ثانیه)
رگرسیون (LR)	۱/۹۳
درخت تصمیم (DT)	۱/۲
شبکه‌های عصبی (MP)	۳/۲
جنگل تصادفی (RF)	۲/۹
ماشین بردار پشتیبان (SVM)	۱/۲
روش پیشنهادی (BSA-SVM)	۸/۹
روش ترکیبی خبره فازی [24]	۹/۴
روش ترکیبی الگوریتم رقابت استعماری و شبکه‌های عصبی [25]	۱۱/۷

۷- مراجع 7- Refrence

[1] E.W.T.Ngai, Y. Hu, Y.H.Wong, Y.Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, 2011.

[2] ع. حسینی و ع. رضائی، "کشف تقلب و راهکارهای مقابله با آن در سازمان‌های بیمه‌ای با استفاده از داده‌کاوی (مطالعه موردی: سازمان تأمین اجتماعی)، " فصلنامه تأمین اجتماعی، دوره ۱۴، شماره ۱، ص ۱۱۱-۱۳۶، ۱۳۹۷.

[2] A. Hosseini, A. Rezaei, " Fraud detection and solutions to deal with it in insurance organizations using data mining (case study: Social Security Organization) ", Social Security Quarterly, vol. 14, no. 1, pp. 111-136.

[3]س. م. تقوی فرد، ز. جعفری، "کشف تقلب در بیمه بنده خودرو با بهره مندی از سامانه خبره فازی"، مدیریت فناوری اطلاعات، دوره ۷، شماره ۲، ص ۲۳۹-۲۵۸، ۱۳۹۴.

[3] S. M. Taqwa Fard, z. Jafari, "Detecting Fraud in Car Insurance Using Fuzzy Expert System", Information Technology Management, vol. 7, no. 2, pp. 239-258, 2014.

[4]A. Ghorbani and S. Farzai, "Fraud Detection in Automobile Insurance using a Data Mining Based Approach, " *Int. J. Mechatronics, Electr. Comput. Technol.*, vol. 8, no. 27, pp. 3764–3771, 2018, doi: IJMEC/10.225163.

[5]م. فیروزی، م. شکوری، ل. کاظمی، س. زاهدی؛ "شناسایی تقلب در بیمه اتومبیل با استفاده از روش‌های داده‌کاوی"، پژوهشنامه بیمه، دوره ۲۶، شماره ۳، ص. ۱۰۳-۱۲۸، ۱۳۹۰.

[5] M. Firouzi, M. Shakuri, L. Kazemi, S. ascetic; "Identifying fraud in car insurance using data mining methods", Insurance Research Journal, vol. 26, no. 3, p. 103-128, 1390.

[6]"Viaene, S. and Dedene, G., 2004. Insurance fraud: issues and challenges. Geneva Papers on Risk and Insurance and Practice, 29, pp.313-33.".

[7]ن. حاجی حیدری، س. خالهء و ا. فراهی؛ "رده‌بندی میزان ریسک بیمه‌گذاران بیمه بنده خودرو با استفاده از الگوریتم‌های داده‌کاوی (مورد مطالعه: یک شرکت بیمه) "، پژوهشنامه بیمه، دوره ۲۶، شماره ۴، ص. ۱۰۷-۱۲۹، ۱۳۹۰.

[7] N. Haji Heydari, S. Khalaha and A. Farahi; "Classification of the risk level of car insurance policyholders using data mining algorithms (case study: an insurance company), " Insurance Research Journal, vol. 26, no. 4, p. 107-129, 1390.

[8]ل. حسین‌زاده، "دسته بندی مشتریان هدف در صنعت بیمه با استفاده از داده‌کاوی"، پایان نامه کارشناسی‌ارشد، دانشگاه تربیت‌مدرس، سال ۱۳۸۶.

[8] L. Hosseinzadeh, "Categories of target customers in the insurance industry using data mining", master's thesis, Tarbiat Modares University, 2016.

[9]ژ. آقابیگی و س. رضائی، "اعتبار سنجی مشتریان اعتباری بانک

[DOI: 10.61186/jsdp.19.4.45]

ملی بر اساس تکنیک‌های داده‌کاوی (رگرسیون لجستیک) " اولین کنفرانس داده‌کاوی ایران، ۱۳۸۶.

[9] J. Aghabeigi and S. Rezaei, "Validation of credit customers of Melli Bank based on data mining techniques (logistic regression) ", 2016.

[10] ر. تهرانی و م. ف. شمس, "طراحی و تبیین مدل ریسک اعتباری در نظام بانکی کشور, " مجله علوم اجتماعی و انسانی دانشگاه شیراز، دوره ۴۳، ص. ۴۵-۶۰، ۱۳۸۴.

[10] R. Tehrani and M. F. Shams, "Designing and explaining the credit risk model in the country's banking system, " Journal of Social Sciences and Humanities of Shiraz University

[11]م. محمدخان, م. اسماعیلی و م. یاراحمدی، "طراحی مدل ارزیابی ریسک اعتباری مشتریان بانک با استفاده از مدل رگرسیون نجستیک" ششمین کنفرانس بین المللی مهندسی صنایع، ۱۳۸۷.

[11] M. Mohammad Khan, M. Ismaili, and M. Yarahamdi, "Designing a credit risk assessment model for bank customers using a logistic regression model," 2017.

[12] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes, " *Pattern Recognit.*, vol. 44, no. 8, pp. 1761–1776, 2011.

[13] Z.-G. Liu, Q. Pan, J. Dezert, and A. Martin, "Combination of classifiers with optimal weight based on evidential reasoning," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1217–1230, 2017.

[14] M. A. Duval-Poo, J. Sosa-Garcia, A. Guerra-Gandón, S. Vega-Pons, and J. Ruiz-Shulcloper, "A new classifier combination scheme using clustering ensemble, " in *Iberoamerican Congress on Pattern Recognition*, 2012, pp. 154–161.

[15]B. Krawczyk and M. Woźniak, "Untrained weighted classifier combination with embedded ensemble pruning, " *Neurocomputing*, vol. 196, pp. 14–22, 2016.

[16]R. Pan, T. Yang, J. Cao, K. Lu, and Z. Zhang, "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information, " *Appl. Intell.*, vol. 43, no. 3, pp. 614–632, 2015, doi: 10.1007/s10489-015-0666-x.

[17]D. E. N. Frossard, I. O. Nunes, and R. A. Krohling, "An approach to dealing with missing values in heterogeneous data using k-nearest neighbors, " *arXiv Prepr. arXiv1608.04037*, 2016, [Online]. Available: http://arxiv.org/abs/1608.04037

[18]V. Kumar and S. Minz, "Feature selection: a literature review, " *Smart Comput. Rev.*, vol. 4, no. 3, pp. 211–229, 2014, doi: 10.1504/ijise.2013.052279.

[19] M. V. Erp, L. G. Vuurpijl, and L.. Schomaker, "An Overview and Comparison of Voting Methods for Pattern Recognition, " in *Proc. of the 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR-8), Niagara-onthe-Lake, Canada*, 2002, pp. 195–200.

[20] L. A. Alexandre, A. C. Campilho, and M. Kamel, "Combining independent and unbiased classifiers using weighted average, " in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, 2000, vol. 2, pp. 495–498.

[21]R. Yager and L. Liu, *Classic Works of the Dempster-Shafer Theory of Belief Functions*, vol. 219. 2008. doi: 10.1007/978-3-540-44792-4.

[22]P. Civicioglu, "Backtracking Search Optimization Algorithm for numerical optimization problems, " *Appl. Math. Comput.*, vol. 219, no. 15, pp. 8121–8144, 2013.

[23]D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation, " *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[24]م. ت. فرد، ف. س. حسینی، و م. خ. بابایی، "مدل رتبه بندی اعتباری هیبریدی با استفاده از الگوریتم‌های ژنتیک و سیستم‌های خبره فازی (مطالعه موردی: مؤسسه مالی و اعتباری قوامین)، " *مدیریت فناوری اطلاعات*, دوره ۶، شماره ۱، ص. ۳۱-۴۶، ۱۳۹۳.

[24] M. T. Fard, F. s. Hosseini, and M. Kh. Babaei, "Hybrid Credit Rating Model Using Genetic Algorithms and Fuzzy Expert Systems (Case Study: Qavamin Financial and Credit Institute), " Information Technology Management, vol. 6, no. 1, pp. 31-46, 2013.

[25]م. صالحی و ع. کتولی، "انتخاب ویژگی‌های بهینه به‌منظور تعیین ریسک اعتباری مشتریان بانکی, " فصلنامه مطالعات مدیریت کسب و کار هوشمند، دوره ۶، شماره ۲، ص. ۱۲۹-۱۵۴، ۱۳۹۶.

[25] M. Salehi and A. Katoli, "Choosing the optimal features in order to determine the credit risk of bank customers, " Smart Business Management Studies Quarterly, vol. 6, no. 2, pp. 129-154, 2016.

رحیم دهخوارقانی در سال ۸۵، مدرک کارشناسی خود را در رشته مهندسی رایانه شاخه نرم‌افزار از دانشگاه پیام‌نور و در سال ۸۷ مدرک کارشناسی‌ارشد خود را از دانشگاه شهید بهشتی در شاخه نرم‌افزار رشته مهندسی رایانه دریافت کرده است و از سال ۹۰ تا ۹۴ دوره دکترا را در دانشگاه سابانجی استانبول (ترکیه) در شاخه هوش مصنوعی گذرانید. از نقاط برجسته کارنامک ایشان، کسب مقام نخست جهانی در مسابقات SUMO Prize 2007 به‌دلیل ساخت بهترین آنتولوژی دامنه برای آنتولوژی SUMO توسط تیم دانشگاه شهید بهشتی است. ایشان از سال ۱۴۰۰ شمسی عضو

[DOI: 10.61186/jsdp.19.4.45]

هیئت علمی دانشگاه ایشیک استانبول (ترکیه) در گروه مهندسی رایانه است.
نشانی رایانامه ایشان عبارت است از:
rahim.dehkharghani@isikun.edu.tr

حجت امامی در رشته مهندسی کامپیوتر گرایش هوش مصنوعی فارغ‌التحصیل شده‌است. وی دانشیار گروه مهندسی کامپیوتر دانشگاه بناب است. زمینه‌های پژوهشی وی شامل داده‌کاوی، یادگیری ماشین، سامانه‌های چند عاملی، الگوریتم‌های اکتشافی و فرااکتشافی، هوش جمعی و کاربردهای آن است. هم‌اکنون او در زمینه استفاده از یادگیری ماشین در حوزه مدیریت ترافیک هوایی، تشخیص بیماری در حوزه پزشکی و جایابی گره‌ها در شبکه‌های نوری کار پژوهشی خود را ادامه می‌دهد.
نشانی رایانامه ایشان عبارت است از:
emami@ubonab.ac.ir

[DOI: 10.61186/jsdp.19.4.45]