مجله محاسبات نرم، جلد ۱۰، شماره ۲ (پاییز و زمستان ۱۴۰۰)، صفحه ۲ـ۱۵
https://doi.org/10.22052/scj.2022.243227.1016

مجله محاسبات نرم
SOFT COMPUTING JOURNAL
تارنمای مجله: scj.kashanu.ac.ir

دانشگاه کاشان
University of Kashan

رتبه‌بندی اعتباری مشتریان بانک به کمک روش جدید گروهی بر پایه ماشین بردار پشتیبان: مطالعه موردی بانک پاسارگاد✧

مجید ابتیاع[۱]، دانشجوی کارشناسی ارشد، سیدمحمد حسینی[۱*]، استادیار، رامین خوچیانی[۲]، استادیار
[۱] دانشکده علوم پایه، دانشگاه آیت الله العظمی بروجردی، بروجرد، ایران.
[۲] دانشکده علوم انسانی، دانشگاه آیت الله العظمی بروجردی، بروجرد، ایران.

اطلاعات مقاله

تاریخچه مقاله:
دریافت ۰۱ آبان ماه ۱۴۰۰
پذیرش ۰۶ اردیبهشت ماه ۱۴۰۱

کلمات کلیدی:
رتبه‌بندی اعتباری
مشتریان بانک
ماشین بردار پشتیبان
روش گروهی
بوت‌استرپ

چکیده

در سال‌های اخیر موضوع رتبه‌بندی اعتباری و شناسایی مشتریان خوش‌حساب و بدحساب، بسیار مورد توجه بانک‌ها قرار گرفته است. اعطای تسهیلات به مشتریان خوش‌حساب و اجتناب از اعطای تسهیلات به مشتریان بدحساب که منجر به کاهش معوقات بانکی می‌شود، همواره یکی از دغدغه‌های مهم مدیران بانک ها است که این مهم به کمک استقرار نظام رتبه بندی اعتباری کارآمد و خوب دور از دسترس نیست. در این مقاله، مدل گروهی جدیدی بر مبنای الگوریتم ماشین بردار پشتیبان برای رتبه‌بندی اعتباری مشتریان بانک ارائه می‌شود. ابتدا به روش بوت‌استرپ، مجموعه داده‌ها به چندین زیرمجموعه تقسیم می‌شود. سپس الگوریتم ماشین بردار پشتیبان بر روی هر زیرمجموعه اعمال و چندین مدل تشکیل می‌شود. در انتها بین مدل‌ها رای‌گیری انجام و مدل نهایی به دست می‌آید. به منظور نمایش دقت مدل گروهی، داده‌های ۲۲۱۸ مشتری بانک پاسارگاد شامل ۱۴ ویژگی توضیح‌دهنده به کمک روش گروهی پیشنهادی مورد ارزیابی قرار گرفتند. بر اساس معیارهای مختلف، نتایج بدست آمده بر روی داده‌های بانک پاسارگاد، برتری روش ماشین بردار پشتیبان گروهی بر روش معمولی ماشین بردار پشتیبان و روش جنگل تصادفی تایید می‌شود. خطای نوع دوم یعنی خطای شناسایی مشتریان بدحساب به عنوان خوش‌حساب در روش گروهی پیشنهادی با هسته خطی ۱۷ درصد کمتر از روش معمولی ماشین بردار پشتیبان و ۱۸ درصد کمتر از روش جنگل تصادفی است.

۱. مقدمه

در عصر حاضر، به دلیل پیشرفت تکنولوژی و سرعت بالای محاسبات و همچنین در دسترس قرار گرفتن داده‌ها و اطلاعات حوزه‌های مختلف، دیدگاه‌های علمی داده محور همچون هوش مصنوعی و یادگیری ماشین مورد استقبال بسیاری از پژوهشگران قرار گرفته است. این دیدگاه‌ها در زمینه‌های علمی و کاربردی متنوعی وارد شده و نتایج بسیار خوبی از خود بجای گذاشته است.

بانک‌ها با فعالیت‌های متنوع و گوناگون مانند تجهیز منابع، تامین نقدینگی، ابزارهای پرداخت، اعطای تسهیلات و سایر موارد بر

✧ نوع مقاله: پژوهشی

* نویسنده مسئول
پست(های) الکترونیک: m.ebtia@abru.ac.ir (ابتیاع)
sm.hoseini@abru.ac.ir (حسینی)
khochiany@abru.ac.ir (خوچیانی)

نحوه ارجاع به مقاله: ابتیاع، مجید، حسینی، سیدمحمد، خوچیانی، رامین، «رتبه‌بندی اعتباری مشتریان بانک به کمک روش جدید گروهی بر پایه ماشین بردار پشتیبان: مطالعه موردی بانک پاسارگاد»، مجله محاسبات نرم، جلد ۱۰، شماره ۲، ص۲ـ۱۵، پاییز و زمستان ۱۴۰۰.

عملکرد اقتصاد کشورها تاثیر می‌گذارد. بنابراین بانک‌ها با خطرهای متعددی روبرو هستند که مهم‌ترین آنها خطر اعتباری است، زیرا منجر به مشکل‌های زیادی در گردآوری سرمایه، اعطای وام و حاشیه سود بانک‌ها می‌شود و بانک‌ها را با کاهش ناگهانی منابع و خطر ورشکستگی مواجه می‌کند. مهم‌ترین فعالیت بانک‌ها، جمع‌آوری منابع مالی و سرمایه‌گذاری و تخصیص آن به بخش‌های مختلف اقتصادی در قالب اعطای تسهیلات است. از طرف دیگر، بخاطر محدودیت منابع مالی در اختیار بانک‌ها، مساله شناسایی و ارزیابی توان بازپرداخت متقاضیان تسهیلات به یکی از چالش‌های پیش روی بانک‌ها تبدیل شده است و بانک‌ها در صدد اعطای تسهیلات خود به مشتریانی هستند که خطر کمتری دارند. مطالبات معوق بانکی نقشی بسزا در سلامت نظام بانکی ایفا می‌کند و افزایش آن منجر به افزایش نااطمینانی و کاهش تمایل و توانایی بانک‌ها برای اعطای تسهیلات خواهد شد [۱]. بنابراین ارائه درست و بهینه تسهیلات مالی، یکی از فعالیت‌های مهم بانک‌ها به‌شمار می‌رود. این امر زمانی ممکن است که به درستی ویژگی‌های مشتریان شناسایی شود و این موضوع با اعتبارسنجی درست مشتریان بر اساس توانایی و تمایل آن‌ها نسبت به بازپرداخت کامل تسهیلات دریافت شده و در سررسید مشخص شده تحقق خواهد یافت. به احتمال پرداخت با تاخیر یا عدم بازپرداخت اصل و فرع تسهیلات اعطایی موسسه مالی از سمت مشتری، خطر اعتباری گفته می‌شود.

همچنین مدیریت نقدینگی از بزرگ‌ترین چالش‌های پیش روی سیستم بانک‌داری است، زیرا بیشتر منابع بانک‌ها از محل سپرده‌های کوتاه‌مدت تأمین مالی می‌شود. از طرف دیگر، تسهیلات اعطایی بانک ها صرف سرمایه گذاری در دارایی‌هایی می‌شود که درجه نقدشوندگی به نسبت پایینی دارند. تاکنون تحقیقات زیادی تاثیر مثبت خطر اعتباری بر خطر نقدینگی را نشان داده‌اند [۲].

پژوهش‌های بسیاری از تکنیک‌های جدید و مدرن تحلیل داده‌ها برای موضوع اعتباری‌سنجی مشتریان استفاده کرده‌اند. روش‌هایی مانند رگرسیون لجستیک، شبکه‌های عصبی، الگوریتم‌های ژنتیک، درخت تصمیم، ماشین بردار پشتیبان (SVM) و بسیاری از روش‌های دیگر که ذیل مبحث داده‌کاوی، یادگیری ماشین و هوش مصنوعی قرار می‌گیرند، برای اعتبارسنجی مشتریان و تخمین ریسک اعتباری به کار گرفته شده است.

در این مقاله، مساله رتبه‌بندی اعتباری متقاضیان حقیقی تسهیلات بانک مورد مطالعه قرار گرفته است. به منظور انجام رتبه‌بندی اعتباری مذکور، روش‌های گروهی بر مبنای SVM بکار گرفته شده است. همچنین بین چهار هسته متداول در روش SVM گروهی، مقایسه‌ای صورت گرفت. در انتها نیز بین روش مذکور و جنگل تصادفی که یکی از روش‌های گروهی متداول و پرکاربرد است، مقایسه‌ای انجام شده است.

۲. تاریخچه و مروری بر پیشینه پژوهش

مساله رتبه‌بندی اعتباری مشتریان بانک‌ها و موسسات مالی همواره مورد توجه مدیران و محققان قرار داشته است. اولین تحقیق و ارائه مدل برای اندازه‌گیری خطر اعتباری بر روی اوراق قرضه توسط جان موری در سال ۱۹۰۹ انجام شد [۳]. روش‌های گوناگونی از قبیل تحلیل پوششی داده‌ها [۴] و رگرسیون لاجیت [۵] برای طراحی مدلی به منظور اندازه‌گیری خطر اعتباری پیشنهاد شده است.

استفاده از روش‌های داده محور در سالیان اخیر بسیار مورد توجه قرار گرفته است که در ادامه به برخی از آنها اشاره می‌شود. مهرآرا و همکاران [۶] برای رتبه‌بندی اعتباری مشتریان بانک به مقایسه دو روش رگرسیون لجستیک و شبکه عصبی پرداخت که نتایج آن تحقیق، دقت ۸۰ درصدی برای رگرسیون و ۸۷ درصدی برای شبکه عصبی را نشان می‌دهد و حاکی از برتری عملکرد و پیش‌بینی بهتر شبکه‌های عصبی به دلیل انعطاف‌پذیری و تعمیم‌پذیری بیشتر آنها می‌باشد. توماس [۷] با مدل رگرسیون لجستیک به پیش‌بینی ریسک اعتباری وام‌گیرندگان پرداخت و دقتی نزدیک به ۷۲ درصد کسب کرد. کشاورز حداد و آیتی‌گازار [۵] به مقایسه روش درخت تصمیم و رگرسیون لجستیک در فرآیند اعتبارسنجی متقاضیان بانکی

برای دریافت تسهیلات پرداخت که هر کدام از این روش‌ها به ترتیب دقتی نزدیک به ۹۶ و ۸۲ درصد از خود ارائه کردند. از روش شبکه‌های عصبی نیز برای طبقه‌بندی متقاضیان وام استفاده شده است که دقتی نزدیک به ۶۹ تا ۸۴ درصد ارائه کردند [۸]. پیش‌بینی بحران‌های مالی همواره در کانون توجه موسسات مالی و بانک‌ها بوده است. پورزمانی و کلانتری [۹]، سه الگوریتم ژنتیک خطی و غیرخطی و شبکه‌های عصبی را در پیش‌بینی بحران مالی بکار برد و مقایسه‌ای بین نتایج آنها ارائه کرد که به ترتیب دقت‌های ۸۰، ۹۰ و ۷۰ درصدی برای این سه الگوریتم حاصل شدند. نتایج این پژوهش نشان داد که این سه روش به دلیل اینکه هیچ پیش‌فرض و پیش‌شرطی بر روی توزیع داده‌ها ندارند از دقت بالایی برخوردار و بسیار موثر هستند. همچنین الگوریتم ژنتیک در تحقیق شین و لی [۱۰] بکار گرفته شد. آنها از صورت‌های مالی حسابرسی شده ۵۲۸ شرکت صنعتی برای تخمین احتمال ورشکستی استفاده کردند و نشان دادند که الگوریتم ژنتیک می‌تواند برای این کار به دقت ۸۰ درصدی برسد. این روش به دلیل ساده بودن، بسیار سریع و موثر عمل می‌کند، با هر نوع تابع هدفی سازگار است و تمام پاسخ‌ها دقیق هستند و هیچ تقریبی وجود ندارد.

رایکرت و همکاران [۱۱] روش آنالیز تشخیص خطی را برای بررسی مسائلی که در توسعه مدل‌های امتیازدهی اعتباری بوجود می‌آیند به کار بردند و به دقتی نزدیک به ۷۰ درصد رسیدند. این روش شباهت زیادی به روش تحلیل رگرسیون دارد. هانگ و همکاران [۱۲] برای امتیازدهی اعتباری مشتریان با یک روش استخراج داده بر اساس SVM، مدل‌هایی بر روی داده‌های دو بانک آلمانی و استرالیایی ارائه دادند. دقت مدل آنها بر روی داده‌های بانک آلمانی نزدیک به ۷۶ درصد و بر روی داده‌های بانک استرالیایی نزدیک به ۸۵ درصد به دست آمد.

طلوعی‌اشلقی و همکاران در [۱۳]، به مقایسه روش‌هایی چون درخت تصمیم، رگرسیون لجستیک، شبکه بیز و SVM در مساله رتبه‌بندی اعتباری پرداخته‌اند و نتیجه گرفتند که مدل SVM نسبت به سایر مدل‌ها از دقت بیشتری برخوردار است. محمدیان‌حاجی‌کرد و همکاران [۱۴]، در پژوهشی با استفاده از مشتریان بانک تجارت با ساخت مدلی به بررسی خطر اعتباری مشتریان پرداخته‌اند. در این پژوهش ابتدا از الگوریتم ژنتیک به منظور بهینه‌سازی داده‌های ورودی استفاده شده و سپس مدلی بر اساس روش SVM شاخته شده است. لازم بذکر است که بهینه‌سازی داده‌های ورودی منجر به افزایش دقت بر روی داده‌های آزمایشی و تعمیم‌پذیری مدل می‌شود. فلاح‌شمس و مهدوی‌راد [۳]، با طراحی مدل‌هایی بر اساس روش‌های اقتصادسنجی مانند لاجیت و پروبیت، به بررسی اعتبار و پیش‌بینی خطر اعتباری مشتریان حقیقی تسهیلات لیزینگ خودرو شرکت لیزینگ ایران خودرو پرداختند و نشان دادند که کارایی مدل لاجیت بیشتر از مدل پروبیت است.

تانگ و همکاران [۱۵]، با استفاده از روش جنگل تصادفی به ارزیابی خطر اعتباری توسط کارت‌های اعتباری بر روی صنعت انرژی در چین پرداختند. هدف این پژوهش، اندازه‌گیری علمی خطر اعتباری کارت‌های اعتباری مورد استفاده در صنعت انرژی است. نتایج آنها نشان می‌دهد ویژگی‌های کارت اعتباری مانند نسبت اعتبار افزوده و میزان هزینه‌های کارت اعتباری در طی یک ماه تاثیر مهمی بر خطر اعتباری دارد. این اطلاعات با ارزش کمک شایانی به بانک‌ها برای بهبود مدیریت خطر خود خواهد کرد.

روش SVM که یکی از روش‌های پرکاربرد در مسایل مربوط به طبقه‌بندی است، علاوه بر مساله رتبه‌بندی اعتباری که موضوع بحث منابع فوق است، بر روی مسایل حوزه‌های مختلفی پیاده‌سازی شده است که از جمله تحقیقات داخلی به کمک این روش می‌توان به تحقیق آرخی و همکاران [۱۶] با هدف طبقه‌بندی اراضی و تحقیق‌های [۱۷–۱۹] در حوزه پزشکی و مقاله خسروی و همکاران [۲۰] در مساله وضعیت تحصیلی اشاره کرد.

۳. چارچوب نظری

الگوریتم‌های یادگیری ماشین در یک طبقه‌بندی به سه دسته یادگیری باناظر (نظارت‌شده) و بدون ناظر (نظارت‌نشده) و نیمه ناظر (نیمه نظارت‌شده) تقسیم می‌شوند. در یادگیری بدون ناظر،

داده‌ها بدون برچسب هستند و هدف از مطالعه این مسایل، افراز مجموعه‌ی داده‌ها به زیرمجموعه‌هایی است که داده‌های هر زیرمجموعه بیشترین شباهت به یکدیگر و داده‌های دو زیرمجموعه متمایز بیشترین عدم شباهت را به یکدیگر داشته باشند. در یادگیری باناظر، داده‌های مساله مورد نظر دارای یک ویژگی به نام ویژگی کلاس (یا برچسب) هستند. در این گونه مسایل، هدف تعیین مدلی است که به خوبی بتواند ویژگی کلاس هر داده را مشخص کند و در مواجهه با داده‌های جدید، کلاس هر یک را تعیین نماید. مدلی که بر اساس داده‌ها ساخته می‌شود را به اصطلاح یادگیرنده، طبقه‌بند یا دسته‌بند می‌نامند. در یادگیری نیمه ناظر بخشی از داده‌ها برچسب دارند و بقیه آنها برچسب ندارند و هدف انتشار برچسب به داده‌های بدون برچسب و در نهایت ارائه مدل دسته‌بند است که بتواند کلاس هر داده جدید را به درستی تشخیص دهد.

در مجموعه‌های داده‌ای این پژوهش، مشتریانی که از بانک مورد نظر تسهیلات دریافت کرده‌اند و آنها را به بانک بازگشت داده یا نداده‌اند به عنوان جامعه آماری در نظر گرفته شده‌اند. مشتریان مذکور دارای دو برچسب زیر هستند:

- مشتریان خوش‌حساب: مشتریانی هستند که تسهیلات اعتباری از بانک دریافت و تمام اقساط آن را در سررسید بازپرداخت کرده‌اند.
- مشتریان بدحساب: مشتریانی هستند که تسهیلات اعتباری از بانک دریافت کرده‌اند ولی قسط یا اقساطی از آن را بعد از سررسید پرداخت کرده‌اند یا نکرده‌اند.

با توجه به اینکه بانک‌ها تمایل دارند تسهیلات خود را به مشتریان خوش‌حساب دهند و از اعطای تسهیلات به مشتریان بدحساب اجتناب کنند، لذا شناسایی هر دو کلاس برای بانک اهمیت دارد و هر دو کلاس باید مورد توجه قرار گیرند. این نکته از دیدگاه یادگیری ماشین به این معناست که این دو برچسب دارای ارزش یکسان هستند.

روش SVM یکی از الگوریتم‌های یادگیری ماشین باناظر است که قادر به انجام دسته‌بندی، رگرسیون و حتی شناسایی داده‌های پرت می‌باشد. این الگوریتم یکی از پرطرفدارترین الگوریتم‌های یادگیری ماشین می‌باشد و جداسازی بهتری روی داده‌ها نسبت به سایر روش‌ها در مساله‌های دسته‌بندی دارد. الگوریتم SVM روشی برای دسته‌بندی داده‌های خطی و غیرخطی می‌باشد.

الگوریتم SVM در سال ۱۹۶۳ توسط ولادیمیر وپنیک ابداع شد و در سال ۱۹۹۵ توسط وپنیک و کورتس برای حالت غیرخطی تعمیم داده شد [۱۶]. از جمله مزایای SVM می‌توان به آموزش ساده، نظریه‌ی قوی، عملکرد خوب برای مدل‌های ساده یا پیچیده و همچنین عملکرد خوب با تعداد داده‌های آموزشی کم اشاره کرد. این روش برخلاف شبکه‌های عصبی در کمینه محلی گیر نمی‌افتد. همچنین حاشیه جداسازی آن برای دسته‌های مختلف کاملا واضح است و این ویژگی تاثیر بسیار مهمی در تعمیم‌پذیری دسته‌بند ساخته شده دارد. علاوه بر این، این روش در فضاهای با ابعاد بالاتر کارایی بیشتری دارد. هدف الگوریتم SVM پیدا کردن مرز تصمیم‌گیری بهینه است، مرزی که به بهترین شکل دسته‌های داده را از یکدیگر جدا می‌کند. در واقع الگوریتم SVM، مرزی که بیشترین حاشیه را با نقاط داده کلاس‌ها دارد، را معرفی می‌کند.

فرض کنید داده‌ها دو کلاسی (دو برچسبی) هستند و به صورت خطی تفکیک‌پذیرند. هر داده دارای d ویژگی (متغیر) است که در بردار X قرار گرفته‌اند و یک ویژگی کلاس یا برچسب که با y نمایش داده شده است. فرض کنید مجموعه‌ای شامل n داده به صورت زیر در اختیار است:

$$X = \{(x_i, y_i)|\ \ i = 1, \dots, n\}, \qquad X \subseteq \mathbb{R}^{d+1} \qquad (۱)$$

که در آن ویژگی کلاس به صورت زیر معرفی شده است:

$$y_i = \begin{cases} +1 & x_i \in C_1 \\ -1 & x_i \in C_2 \end{cases} \qquad (۲)$$

تابع تصمیم یا مرز جداکننده به صورت $\mathrm{D(x)} = \mathrm{w^T x_i} + \mathrm{b}$ تعریف می‌شود که در آن w شیب (وزن‌ها) و b عرض از مبدا را نشان می‌دهند. تابع تصمیم می‌تواند یک خط (در مورد داده‌های دو متغیری)، صفحه (در مورد داده‌های سه متغیری) یا ابرصفحه (در مورد داده‌های بیش از سه متغیر) باشد. هدف، پیدا کردن مرزی خطی است به گونه‌ای که جداکننده دو کلاس از یکدیگر با بیشترین حاشیه باشد. کمترین فاصله‌ی نقاط دو دسته

تا مرز جداکننده مورد نظر را حاشیه می‌نامند. مرزی که حاشیه زیاد دارد قابلیت تعمیم‌پذیری بیشتر بر روی داده‌ها خواهد داشت. مساله پیدا کردن مرز بهینه به فرم کلی زیر بیان می‌شود:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 \\ s.t. \quad & y_i(w^T x_i + b) \geq 1 \end{aligned} \qquad (۳)$$

می‌توان با کمک ضرایب لاگرانژ مثبت، فرم دوگان مساله را به دست آورد و با الگوریتم‌های بهینه‌سازی جواب‌ها را محاسبه کرد، لذا داریم:

$$\begin{aligned} \max \; L_d &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j\,y_i\,y_j\;\; {x_i}^T x_j + \sum_{i=1}^{n}\alpha_i \\ s.t \quad & \sum_{i=1}^{n}\alpha_i\,y_i = 0 \\ & \alpha_i \geq 0 \quad \forall i \end{aligned} \qquad (۴)$$

مساله بهینه‌سازی فوق منجر به تفکیک کامل داده‌ها از یکدیگر می‌شود و هیچ یک از داده‌های آموزشی فاصله تا مرز کمتر از مقدار حاشیه ندارند. به این حالت به اصطلاح حاشیه سخت گفته می‌شود. در مقابل حاشیه سخت، حاشیه نرم را می‌توان در نظر گرفت. به طور عمومی زمانی که داده‌ها به صورت خطی تفکیک‌پذیر نیستند، از این روش استفاده می‌شود. در مورد حاشیه نرم، مساله بهینه‌سازی اندکی تغییر می‌کند و به فرم زیر تبدیل می‌شود:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\varepsilon^i \\ s.t \quad & y_i(w^T x_i + b) \geq 1 - \varepsilon^i \\ & \varepsilon^i \geq 0 \end{aligned} \qquad (۵)$$

مقدار C یک متغیر تنظیم‌کننده است، یعنی مقدار C بزرگ منجر به خطای کم دسته‌بند روی داده‌های آموزشی می‌شود و مقدار C کوچک منجر به حاشیه بزرگ همراه با خطای بیشتر می‌شود. همانند مساله حاشیه سخت، می‌توان با استفاده از ضرایب لاگرانژ مثبت، فرم دوگان مساله را به دست آورد و با استفاده از الگوریتم‌های بهینه‌سازی جواب‌ها را محاسبه نمود.

الگوریتم SVM علاوه بر مسایلی که خطی تفکیک‌پذیر هستند بر روی مسایل غیرخطی نیز قابل اعمال است. در واقع در حالت غیرخطی، نگاشتی استفاده می‌شود که مساله غیرخطی را به یک فضای ویژگی با ابعاد بیشتر اما خطی تفکیک‌پذیر می‌نگارد. به طور عمومی این کار با استفاده از تبدیلات غیرخطی و استفاده از یک مدل خطی در فضای جدید به منظور دسته‌بندی داده‌ها انجام می‌شود. مدل خطی در فضای جدید متناظر با یک مدل غیرخطی در فضای اصلی است. بطور متداول از یک تابع هسته استفاده می‌شود. برای وجود و یکتایی جواب مساله بهینه‌سازی، تابع هسته باید متقارن و مثبت نیمه معین باشد. مساله بهینه‌سازی دوگان، با تابع هسته K به صورت زیر بیان می‌شود:

$$\begin{aligned} \max \; L_d &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_j y_j K(x_i.x_j) + \sum_{i=1}^{n}\alpha_i \\ s.t \quad & \sum_{i=1}^{n}\alpha_i\,y_i = 0 \\ & \alpha_i \geq 0 \quad \forall i \end{aligned} \qquad (۶)$$

توابع هسته معروف عبارتند از:

- تابع هسته خطی

$$K(x_i, x_j) = < x_i, x_j > \qquad (۷)$$

- تابع هسته چندجمله‌ای

$$K(x_i, x_j) = \left(\gamma < x_i, x_j > + r\right)^d \quad r > 0, d \in N \qquad (۸)$$

- تابع هسته‌ی گاوسی

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2) \qquad (۹)$$

- تابع هسته‌ی سیگموید

$$K(x_i, x_j) = \tanh(\gamma {x_i}^T x_j + r) \qquad (۱۰)$$

در روش SVM انتخاب ابر متغیرهایی مثل C و تعیین متغیرهایی که در تابع هسته وجود دارند یک چالش می‌باشد که در این پژوهش از روش اعتبارسنجی متقابل روی یک شبکه از مقادیر محتمل برای تخمین مقادیر بهینه استفاده شده است.

در الگوریتم‌های باناظر، داده‌ها به دو بخش مجموعه آموزشی و مجموعه آزمایشی تقسیم می‌شوند. مجموعه آزمایشی پس از ساخت مدل نهایی برای ارزیابی و تعمیم‌پذیری آن به کار گرفته می‌شوند. مجموعه آموزشی نیز اغلب به دو مجموعه آموزشی و اعتبارسنجی تقسیم می‌شود. با توجه به این که اکثر روش‌ها نیاز

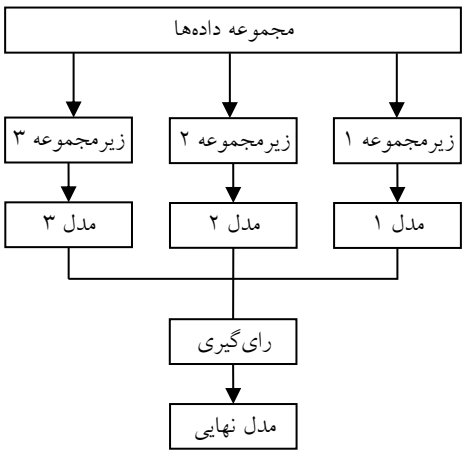
به تعیین متغیرهایی توسط کاربر دارند مشخص کردن آنها توسط روش‌های مختلفی انجام می‌شود که یکی از متداول‌ترین آنها، روش‌های جستجوی شبکه‌ای است. در ابتدا مقدار متغیرهای دسته‌بند بر اساس نقاط شبکه مقداردهی می‌شوند. سپس با استفاده از مجموعه آموزشی مدل تعیین شده و با بهره گیری از مجموعه اعتبارسنجی ارزیابی می‌شود. در نهایت بر اساس ارزیابی‌های به دست آمده متغیرهای بهینه ایجاد می‌شوند و بر پایه آنها مدل نهایی ساخته می‌شود. بعد از ساخت دسته‌بند نهایی، تعمیم‌پذیری آن به کمک مجموعه آزمایشی تخمین زده می‌شود.

اعتبارسنجی متقابل، یکی از روش‌های اعتبارسنجی است که می‌توان با آن متغیرهای مدل را به طور بهینه مشخص کرد. در یک فرآیند تکراری که آن را اعتبارسنجی متقابل می‌نامند، در هر بار تکرار این فرآیند داده‌های آموزشی برای یادگیری و داده‌های اعتبارسنجی برای آزمایش دسته‌بند مورد استفاده قرار می‌گیرند. پس در هر مرحله از فرآیند، داده‌های آزمایشی برای پیش‌بینی خطا یا دقت حاصل از برازش مدل بر روی داده‌های آموزشی به کار گرفته می‌شوند، در نهایت میانگین این خطاها را به عنوان خطای مدل در نظر می‌گیرند. پس مدلی انتخاب می‌شود که دارای کمترین خطا باشد. روش‌های مختلف اعتبارسنجی وجود دارد که در این مقاله، روش k بخشی برای اعتبارسنجی بر روی داده‌ها استفاده خواهد شد.

یکی از دیدگاه‌های مهم و کاربردی که اخیرا بسیار مورد توجه محققین حوزه یادگیری ماشین و هوش مصنوعی قرار گرفته است، ادغام و ترکیب چند دسته‌بند برای ساخت یک دسته‌بند قوی و بهتر است. ترکیب چند دسته‌بند برای ساخت یک دسته‌بند قوی را دسته‌بند گروهی می‌نامند. نمای کلی روش‌های گروهی در شکل (۱) به ازای سه دسته‌بند نمایش داده شده است. در دسته‌بند گروهی دو دیدگاه وجود دارد:

- رای‌گیری بوت‌استرپ: در این دیدگاه، داده‌ها اصلی به صورت تصادفی با جایگذاری به زیرمجموعه‌هایی از اندازه برابر تقسیم می‌شود و هر زیرمجموعه به یکی از دسته‌بندها نشان داده می‌شود. یعنی هر دسته‌بند یک بخش از مجموعه داده‌ها را مشاهده می‌کند و باید مدل خود را بر اساس همان بخش از داده‌ها که در اختیارش قرار گرفته است، بسازد. بعد از ساخت مدل‌ها، بین آنها رای‌گیری می‌شود و دسته‌بند نهایی به دست می‌آید.
- ارتقا یا تقویت: این دیدگاه شبیه دیدگاه اول، داده‌ها را تقسیم می‌کند، اما بعد از هر بار آموزش، بر روی داده‌هایی تمرکز می‌کند که به درستی دسته‌بندی نشده‌اند. بدین شکل که در ابتدا همه نمونه‌ها وزن یکسانی دارند ولی پس از آموزش، وزن داده‌هایی که به درستی دسته‌بندی نشده‌اند افزایش می‌یابد و از وزن داده‌هایی که به درستی دسته‌بندی شده‌اند کم می‌شود.

در الگوریتم‌های گروهی، به طور معمول بعد از ساخت دسته‌بندها، برای دسته‌بندی نهایی، بین دسته‌بندها رای‌گیری انجام می‌شود و آن دسته‌بندی که بیشترین رای را بیاورد، به عنوان دسته‌بند نهایی انتخاب می‌شود.



شکل (۱): نمودار الگوریتم‌های گروهی

روش جنگل تصادفی یک دسته‌بند باناظر است و بر اساس ایده روش‌های گروهی شکل گرفته است. این روش، یک رای‌گیری بوت‌استرپ روی چندین درخت تصمیم ساخته شده از داده‌ها می‌باشد. در روش جنگل تصادفی، ابتدا داده‌ها به شیوه بوت‌استرپ به چندین زیرمجموعه تقسیم می‌شود و روی هر زیرمجموعه یک درخت تصمیم که ویژگی‌ها به طور تصادفی شاخه‌ها را تشکیل می‌دهند، ساخته می‌شود. تعداد ویژگی‌های مورد استفاده در هر درخت تصمیم را عمق درخت می‌گویند.

بنابراین هر درخت تصمیم تنها بخشی از داده‌ها را مشاهده می‌کند و همچنین برخی از ویژگی‌ها به طور تصادفی و نه همه آنها مبنای شاخه‌های درخت تصمیم خواهد بود. در انتها، بین تمام درخت‌های تصمیم ساخته شده، رای‌گیری می‌شود و مدل نهایی بر اساس آن تشکیل می‌شود. بعلاوه در مواجه با داده‌های جدید نیز هر داده به همه درخت‌ها ارائه می‌شود و بین نتایج همه آنها رای‌گیری و دسته‌ی داده جدید مشخص می‌شود. در روش جنگل تصادفی، رای‌گیری از دسته‌بندهای ضعیف انجام می‌شود و از طرفی نتایج مطلوبی در مسایل مختلف از خود نشان داده است به گونه‌ای که این روش جزو روش‌های محبوب، پرکاربرد و به‌صرفه در بین روش‌های مختلف یادگیری ماشین تبدیل شده است.

۴. روش پیشنهادی

در این بخش، روش SVM گروهی معرفی و نحوه استفاده از آن برای دسته‌بندی مساله اعتبارسنجی مشتریان تسهیلات بانک توضیح داده می‌شود.

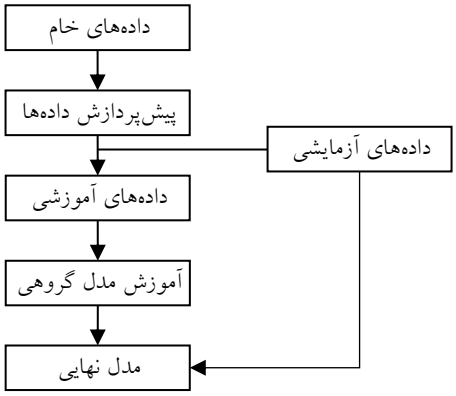
مجموعه داده‌ای را به صورت زیر در نظر بگیرید.

$$X = \{(x_i, y_i) |\ x_i \in \mathbb{R}^d, y_i = \pm 1,\ \ i = 1, \dots, n\} \quad (۱۱)$$

که در آن $y_i = 1$ نشان‌دهنده تعلق x_i به کلاس C_1 و $y_i = -1$ نشان‌دهنده تعلق x_i به کلاس C_2 است. ابتدا مجموعه X به دو بخش داده‌های آموزشی و آزمایشی تقسیم می‌شود. در این پژوهش ۲۰ درصد داده‌ها به صورت تصادفی به عنوان داده‌های آزمایشی در نظر گرفته می‌شود و بنابراین ۸۰ درصد داده‌ها به آموزش مدل اختصاص داده می‌شود. داده‌های آموزشی به N زیرمجموعه A_1، A_2، ... و A_N با اندازه یکسان تقسیم می‌شوند. برای تشکیل این زیرمجموعه‌ها از روش بوت‌استرپ استفاده می‌شود، یعنی شیوه انتخاب به روش با جایگذاری، اعضای هر زیرمجموعه را تعیین می‌کند.

برای هر $j = 1,2, \dots, N$ بر روی زیرمجموعه A_j مراحل زیر انجام می‌شود. زیرمجموعه A_j به الگوریتم SVM ارائه می‌شود. برای ساخت مدل ابتدا به روش اعتبارسنجی، متغیرهای مدل SVM تنظیم و بهترین آنها انتخاب می‌شود. متغیرهای بهینه دسته‌بندی به صورت زیر محاسبه می‌شوند. ابتدا یک شبکه از مقادیر ابرمتغیرها انتخاب می‌شود و به ازای هر نقطه از این شبکه، ابرمتغیرهای روش، مقداردهی می‌شوند. سپس از روش اعتبارسنجی متقابل برای ارزیابی عملکرد ابرمتغیر در نقطه مذکور بهره برده خواهد شد. در روش اعتبارسنجی متقابل، داده‌های آموزشی به طور تصادفی و با جایگذاری، به k بخش مساوی تقسیم می‌شوند. از زیرنمونه‌ها، یکی به عنوان داده‌های اعتبارسنجی برای آزمایش مدل حفظ می‌شود و بقیه زیرنمونه‌ها مدل را می‌سازند. بعد از ساخت مدل، زیرنمونه آزمایشی برای میزان اعتبار مدل ساخته شده به کار گرفته می‌شود. به این ترتیب به ازای هر نقطه شبکه، فرآیند اعتبارسنجی متقابل، k بار تکرار می‌شود و در هر بار معیار عملکرد آن نقطه سنجیده می‌شود. در انتها از میانگین معیارهای بدست آمده برای سنجش میزان اعتبار آن نقطه بهره برده می‌شود. در میان نقاط مختلف شبکه مذکور، نقطه‌ای که بهترین اعتبار را تولید کرده است به عنوان ابرمتغیرهای مدل اصلی استفاده می‌شود. بعد از انتخاب بهترین متغیرها، الگوریتم SVM روی مجموعه A_j اعمال و مدل نهایی دسته‌بند ساخته می‌شود.

پس از ساخت مدل‌ها، در انتها N مدل دسته‌بند به دست می‌آید که هر کدام در واقع بخشی از داده‌های آموزشی را مشاهده کرده‌اند. اکنون برای ارزیابی روش، بین این مدل‌ها روی داده‌های آزمایشی رای‌گیری می‌شود تا کلاس هر کدام از داده‌های آزمایشی پیش‌بینی شود. به این معنا که هر یک از داده‌های آزمایشی به N مدل ارائه می‌شود و کلاس آن توسط هر کدام از دسته‌بندها پیش‌بینی می‌شود و در نهایت بین این پیش‌بینی‌ها رای‌گیری می‌شود. با انجام این کار می‌توان بین پیش‌بینی کلاس و کلاس واقعی داده‌های آزمایشی مقایسه انجام داد و عملکرد روش را سنجید. نمودار روش پیشنهادی در شکل(۲) نمایش داده شده است. این روش مزایای زیادی دارد که مهم‌ترین آن سادگی و سرعت بالای اجرای آن است. بعلاوه تمام ویژگی‌های روش SVM از جمله تعمیم‌پذیری بالا و عملکرد مطلوب روی داده‌های چندکلاسی را هم خواهد داشت.



شکل (۲): نمودار روش پیشنهادی

در این تحقیق، فرآیند رتبه‌بندی مشتریان موسسات مالی نسبت به خطر اعتباری آنها به صورت زیر انجام می‌شود. ابتدا مجموعه داده‌ای مربوط به مشاهدات جمع‌آوری و پیش‌پردازش می‌شوند. پس از جمع‌آوری داده‌ها، ابتدا داده‌ها نرمال و سپس با کمک روش مولفه‌های داده‌های پرت محلی، داده‌های پرت کنار گذاشته می‌شوند. سپس داده‌ها به دو بخش آموزشی و آزمایشی تقسیم می‌شود. اکنون روش SVM گروهی اعمال می‌شود. در این پژوهش از چندین هسته مختلف از جمله چندجمله‌ای، خطی، گاوسی و سیگموید استفاده خواهد شد. هر کدام از دسته‌بندها به صورت ترکیبی ساخته می‌شوند، یعنی چندین SVM بر اساس یک هسته تشکیل می‌شوند. در مرحله آخر بر اساس داده‌های آزمایشی، مدل نهایی مورد ارزیابی قرار می‌گیرد و میزان اعتماد و خطای مدل به دست می‌آید. در آخر با مقایسه بین انواع ماشین‌های بردار پشتیبان گروهی، بهترین دسته‌بند اصلی معرفی می‌شود و همچنین مقایسه‌ای بین SVM با روش جنگل تصادفی انجام می‌گیرد.

برای ارزیابی و مقایسه نتایج حاصل شده و آزمون عملکرد روش‌های ارائه شده از معیارهای زیر بهره گرفته شده است. این معیارها در تحقیقات زیادی مورد استفاده و استناد قرار گرفته‌اند [۲۱]. در ادامه برچسب خوش‌حساب را کلاس مثبت و برچسب بدحساب را کلاس منفی در نظر می‌گیریم. بر اساس ماتریس درهم‌ریختگی، معیارهای زیر برای بیان دقت و خطای دسته‌بند استفاده می‌شود.

$$P = \frac{TP}{TP + FP} \quad (۱۲)$$

$$R = \frac{TP}{TP + FN} \quad (۱۳)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (۱۴)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (۱۵)$$

$$EI = \frac{FP}{FP + TN} \quad (۱۶)$$

$$EII = \frac{FN}{TP + FN} \quad (۱۷)$$

که در آنها TP تعداد نمونه‌های کلاس مثبت (مشتریان خوش‌حساب) است که به درستی مثبت تشخیص داده شده‌اند، TN تعداد نمونه‌های کلاس منفی (بدحساب) است که به درستی منفی تشخیص داده شده‌اند، FP تعداد نمونه‌های کلاس منفی که به اشتباه مثبت تشخیص داده شده‌اند و FN تعداد نمونه‌های کلاس مثبت که به اشتباه منفی تشخیص داده شده‌اند. شش معیار فوق را به صورت زیر ارجاع می‌دهند: P صحت مدل، R پوشش مدل، ACC دقت مدل، MCC ضریب همبستگی متیوز مدل و مقادیر EI و EII به ترتیب خطای نوع اول و نوع دوم مدل نامیده می‌شوند. از آنجا که تشخیص اشتباه مشتری بدحساب به عنوان خوش‌حساب منجر به هزینه برای بانک می‌شود لذا کوچک بودن خطای نوع اول اهمیت زیادی دارد و به عنوان یک معیار مهم در این تحقیق لحاظ خواهد شد.

لازم به ذکر است که الگوریتم پیشنهادی، بر روی یک رایانه شخصی با مشخصات فنی Intel Core i7 و ۶ گیگابایت RAM و به کمک نرم‌افزار پایتون پیاده‌سازی شده است. همچنین از آنجا که روش پیشنهادی به صورت گروهی است و هر مدل به طور مستقل از سایر مدل‌ها ساخته می‌شود، بنابراین می‌توان به کمک فناوری پردازش موازی، هزینه محاسباتی روش را به شدت کاهش داد.

۵. تجزیه و تحلیل داده‌ها

این پژوهش به ارائه روش نوین SVM گروهی و مقایسه آن با الگوریتم هم‌طراز آن، یعنی جنگل تصادفی برای دسته‌بندی مشتریان خوش‌حساب و بدحساب بانک می‌پردازد. روش‌های

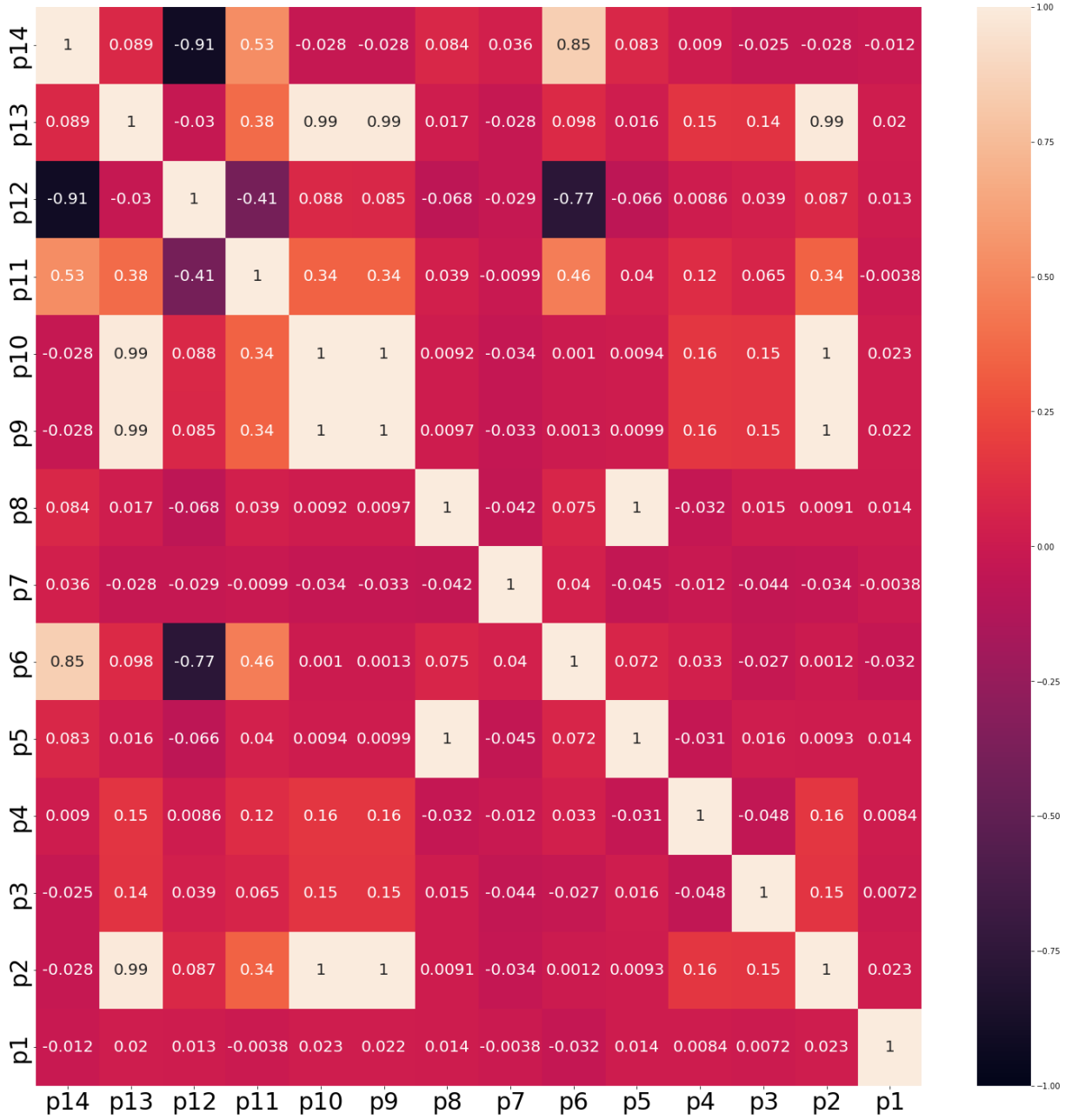
مذکور بر روی مجموعه داده‌های مشتریان اعتباری یکی از شعب بانک پاسارگاد اعمال خواهد شد. ابتدا داده‌ها معرفی و آماره‌های مختلف داده‌های بانکی به طور خلاصه بررسی می‌شوند.

داده‌های بانک پاسارگاد شامل داده‌های مربوط به ۲۲۱۸ مشتری است که سالیان قبل از بانک تسهیلات دریافت کرده‌اند و در زمان انجام این تحقیق حداقل یک سال از موعد آخرین سررسید آن گذشته است. هر مشتری برمبنای تسهیلات دریافتی و طبق نظر کارشناسان و مدیران بانک یکی از دو وضعیت خوش‌حساب و بدحساب را دریافت کرده است. از هر مشتری ۱۴ متغیر (ویژگی) در مجموعه داده‌ای ثبت شده است که عبارتند از: (۱) جنسیت، (۲) مبلغ تسهیلات دریافتی، (۳) مبلغ وجه التزام دریافتی، (۴) نوع وثیقه، (۵) سن، (۶) نحوه بازپرداخت، (۷) میزان تحصیلات، (۸) میزان سابقه کار، (۹)حدود میزان درآمد ماهانه، (۱۰) اموال و دارایی‌های فعلی، (۱۱) وضعیت مسکن، (۱۲) سابقه چک برگشتی، (۱۳) متوسط سپرده قبل از تسهیلات و (۱۴) وضعیت مشتری (خوش‌حساب یا بدحساب). از ۲۲۱۸ مشتری موجود در این پایگاه داده، تعداد ۱۷۹۵ نفر به عنوان خوش‌حساب و ۴۲۳ نفر به عنوان بدحساب شناخته شده‌اند. همچنین مبلغ کل تسهیلات پرداختی به تمام مشتریان مذکور ۱۷۱۸ میلیارد ریال می‌باشد که بیش از ۲۰ درصد این مبلغ یعنی ۳۴۵ میلیارد ریال به مشتریان بدحساب تخصیص یافته است.

در ابتدا بر روی داده‌ها پیش‌پردازش انجام می‌شود. در شکل(۳)، همبستگی بین متغیرهای مختلف در مجموعه داده‌ای بانک پاسارگاد را مشاهده می‌کنید. همان گونه که مشخص است بین متغیر مبلغ تسهیلات دریافتی و متغیرهای درآمد ماهانه، اموال فعلی و متوسط سپرده و همچنین بین متغیر سن و متغیر سابقه کار همبستگی مثبت بالایی (یک یا بسیار نزدیک به یک) برقرار است. بنابراین می‌توان از میان این متغیرها که همبستگی بالایی دارند برخی را حذف کرد. در این پژوهش درآمد ماهانه، اموال فعلی، متوسط سپرده و سابقه کار را حذف می‌کنیم. از طرف دیگر، بین متغیر چک برگشتی و متغیر وضعیت مشتری همبستگی منفی بالایی (نزدیک به ۱-) وجود دارد. به بیان دیگر، با توجه به متغیر چک برگشتی هر مشتری می‌توان با دقت بسیار بالایی (نزدیک به ۱۰۰ درصد) خوش‌حساب بودن یا بدحساب بودن وی را پیش‌بینی کرد. همچنین متغیر نحوه بازپرداخت و متغیر وضعیت مشتری دارای همبستگی ۰/۸۵ هستند.

با اعمال روش بردار پشتیبان گروهی با هسته‌های مختلف از قبیل خطی، گاوسی و چندجمله‌ای بر روی این مجموعه‌ی داده‌ای که تنها شامل دو متغیر نحوه بازپرداخت و چک برگشتی همراه با متغیر وضعیت مشتری است، تمام معیارهای دقت، صحت، پوشش و ضریب متیوز بالای ۹۹ درصد و خطای نوع اول و دوم کمتر از دو درصد نتیجه شد. لذا نتیجه می‌شود که در این مجموعه داده‌ای، ویژگی چک برگشتی و نحوه بازپرداخت از اهمیت بسیار بالایی برای پیش‌بینی وضعیت مشتری برخوردار است و مدیران آن شعبه از بانک می‌توانند تنها با دانستن این دو متغیر، وضعیت مشتری را به خوبی پیش‌بینی کنند.

برای اعمال و مقایسه روش پیشنهادی با روش‌های دیگر در ادامه، از ستون‌های مذکور در بالا، ویژگی چک برگشتی و نحوه بازپرداخت هم از مجموعه داده‌ای حذف شد و با توجه به ستون‌های زیر، روش پیشنهادی اعمال شد: (۱)جنسیت، (۲)مبلغ تسهیلات دریافتی، (۳) مبلغ وجه التزام دریافتی، (۴) نوع وثیقه، (۵) سن، (۶) میزان تحصیلات، (۷) وضعیت مسکن و (۸)وضعیت مشتری (خوش‌حساب یا بدحساب). به منظور استفاده از روش SVM گروهی داده‌های موجود بایستی به صورت عددی بیان شود. همچنین بعد از کمی کردن ویژگی‌های کیفی، داده‌ها نرمال‌سازی شدند. روش پیشنهادی و سایر روش‌ها در محیط نرم‌افزار Python پیاده‌سازی و بر روی داده‌های بانک پاسارگاد اجرا شدند. مقادیر بهینه متغیرهای هر یک از این مدل‌ها در Python در جدول (۱) گزارش شده است. همچنین در مورد آموزش مدل، برای اعتبارسنجی از روش اعتبارسنجی متقابل ۱۰ بخشی استفاده شده است.



شکل (۳): میزان همبستگی متغیرهای مشتریان بانک پاسارگاد. منبع: محاسبات محقق

در مورد روش جنگل تصادفی، اسمی بودن متغیرها اهمیتی ندارد ولی به منظور مقایسه نتایج، در این مقاله روش جنگل تصادفی نیز بر روی داده‌های نرمال‌سازی شده اعمال گردید. بعد از نرمال‌سازی، همانگونه که در بخش قبل بیان شد، ویژگی‌ها و متغیرهای مهمی که از ارزش اطلاعاتی بیشتری برخوردار هستند شناسایی می‌شوند. از الگوریتم تحلیل مولفه‌های اصلی برای استخراج ویژگی استفاده شد، به این ترتیب تبدیل‌های این الگوریتم روی ستون‌های داده‌ای، ستون‌های جدیدی را به وجود می‌آورند که بیشترین اطلاعات را در خود دارند. سپس به کمک الگوریتم رگرسیون لجستیک، از ستون‌های جدید ستون‌هایی که اهمیتی نزدیک به صفر دارد شناسایی و حذف می‌شوند. در انتها روش‌های یادگیری گروهی بر روی ستون‌های باقی‌مانده اعمال و ارزیابی می‌شوند.

جدول (۱): ابرمتغیرهای بهینه‌ی مدل‌های مختلف (منبع محاسبات محقق)

مدل		متغیرها	مقدار بهینه
SVM معمولی	گاوسی	Gamma	۷۷۴۲/۶۳۶۸
	خطی	C	۳/۵۹۳۸
	چندجمله‌ای	Degree	۳
	سیگموید	Gamma	۴۶/۴۱۵۹
SVM ترکیبی	گاوسی	n_estimators	۲۱
	خطی	n_estimators	۱۴
	چند جمله‌ای	n_estimators	۱۷
	سیگموید	n_estimators	۱۹
درخت تصمیم		max_depth	۲۷
جنگل تصادفی		max_depth	۷
		n_estimators	۱۹

جدول (۲)، معیارهای دقت، صحت، پوشش، ضریب همبستگی متیوز، خطای نوع اول و دوم برای روش SVM گروهی را نشان می‌دهد. همچنین به منظور مقایسه، روش معمولی SVM نیز روی مجموعه داده‌ای اعمال گردید و نتایج آن در جدول (۲) آمده است. همان طور که ملاحظه می‌شود، تمام معیارهای شش‌گانه در روش SVM گروهی بهبود یافته‌اند. به ویژه در نتایج با هسته خطی، خطای نوع اول که شناسایی اشتباه مشتریان بدحساب به عنوان خوش‌حساب است، شاهد کاهش ۱۷ درصدی است. همان گونه که پیش از این اشاره شد این کاهش ۱۷ درصدی به این معناست که روش پیشنهادی با هسته خطی مشتریان بدحساب را با دقت بیشتری شناسایی می‌کند و می‌تواند به بانک در کاهش خطر اعطای تسهیلات کمک کند.

نتایج به دست آمده با هسته گاوسی در هر دو روش عادی و گروهی معیارهای شش‌گانه یکسانی را نشان می‌دهد، اما با توجه به این که روش گروهی قابل انجام به صورت موازی است لذا روش گروهی با سرعت بیشتری می‌تواند پردازش را انجام دهد زیرا هر دسته‌بند به طور جداگانه آموزش می‌بیند و مدل خود را می‌سازد. به عبارت دیگر، در صورتی که هر دسته‌بند روی ۱۰ درصد از داده‌ها اعمال شود، هزینه محاسباتی روش گروهی تقریباً یک دهم هزینه محاسباتی روش عادی می‌شود. این نکته در داده‌های حجیم از اهمیت بسیار بالایی برخوردار است.

روش گروهی تنها با هسته سیگموید عملکرد ضعیف‌تری نسبت به روش عادی در معیارهای دقت، ضریب متیوز و خطای نوع اول دارد و در سایر معیارها بهبود اندکی حاصل شده است.

در مقایسه با روش گروهی جنگل تصادفی، روش SVM گروهی با هسته‌های گاوسی، خطی و چندجمله‌ای در تمام معیارها بهتر عمل نموده است و نشان از ارجحیت روش پیشنهادی نسبت به جنگل تصادفی است. همچنین شناسایی مشتریان بدحساب به روش گروهی با هسته خطی کاهش ۱۸ درصدی را نشان می‌دهد.

جدول (۲): مقایسه درصد معیارهای ارزیابی دو روش SVM و روش پیشنهادی بر روی داده‌های بانک پاسارگاد. منبع: محاسبات محقق

	روش	ACC	P	R	MCC	EI	EII
SVM معمولی	گاوسی	۱	۱	۱	۱	۰	۰
	خطی	۰/۸۷	۰/۸۷	۰/۹	۰/۸۳	۰/۲	۰/۱
	چندجمله‌ای	۰/۸۸	۰/۸۸	۰/۸۹	۰/۸۴	۰/۱۷	۰/۱۱
	سیگموید	۰/۷۳	۰/۸۱	۰/۷۸	۰/۴۱	۰/۴۵	۰/۲۲
SVM گروهی	گاوسی	۱	۱	۱	۱	۰	۰
	خطی	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۷	۰/۰۳	۰/۰۰۵
	چندجمله‌ای	۰/۹۹	۰/۹۹	۱	۰/۹۸	۰/۰۲	۰/۰۰۳
	سیگموید	۰/۸۳	۰/۸۹	۰/۸۹	۰/۴۹	۰/۴۰	۰/۱۱
درخت تصمیم		۰/۸۶	۰/۸۸	۰/۸۷	۰/۷۶	۰/۲۳	۰/۱۰
جنگل تصادفی		۰/۸۸	۰/۸۷	۰/۸۷	۰/۸۱	۰/۲۱	۰/۱۱

در قسمت پایانی این بخش به مقایسه یافته‌های این پژوهش با پژوهش‌های دیگر می‌پردازیم. هانگ و همکاران [۱۲] در سال ۲۰۰۷ با استفاده از روش SVM بر روی داده‌های بانک آلمانی و استرالیایی به ارزیابی خطر اعتباری می‌پردازد. به عنوان نمونه دقت پژوهش بر روی داده‌های آلمانی با روش SVM با هسته گاوسی دقت ۷۶ درصدی گزارش شده است. در پژوهش دیگری، محمدیان‌حاجی‌کرد و همکاران [۱۴] در سال ۱۳۹۵ با استفاده از روش SVM به بررسی خطر اعتباری مشتریان بانک تجارت می‌پردازد. نتایج حاصل در پژوهش مذکور ۵۷ درصد دقت بر روی داده‌های آزمایشی می‌باشد که در مقایسه با مقادیر

جدول (۲) نسبت به تمام روش‌ها از دقت کمتری برخوردار است. طلوعی و همکاران [۱۳] با مقایسه چندین روش SVM، رگرسیون لجستیک، شبکه بیز و درخت تصمیم به رتبه‌بندی مشتریان بانکی می‌پردازد. نتایج حاصل شده در مقایسه با دقت‌های روش SVM گروهی در جدول (۲) نشان از کارایی روش‌های این پژوهش می‌باشد. به طور خلاصه از مطالب بالا می‌توان نتیجه گرفت روش‌های ترکیبی از کارایی و تعمیم‌پذیری بهتری نسبت به روش‌های معمولی برخوردار هستند.

۶. نتیجه‌گیری

با توجه به اینکه افزایش دقت در رتبه‌بندی اعتباری و تشخیص درست مشتریان بدحساب می‌تواند جلوی خسارت‌هایی عظیم به بانک‌ها و موسسات مالی را بگیرد، لذا ارائه مدلی کارا و دقیق در این راستا از اهمیت بالایی برای مدیران موسسات مالی دارد. روش گروهی پیشنهاد شده در این تحقیق بر پایه روش SVM توانست مدل‌هایی با دقت بالاتر، تعمیم‌پذیری بیشتر و هزینه محاسباتی بسیار کمتر به وجود آورد. روش پیشنهادی بر روی داده‌های یکی از شعب بانک پاسارگاد اعمال شد و در شناسایی مشتریان خوش‌حساب و بدحساب نتایج بسیار خوبی از خود نشان داد. روش پیشنهادی با هسته‌های خطی، چندجمله‌ای و گاوسی بر اساس شش معیار سنجیده شد که در معیارهای دقت، صحت، پوشش، ضریب همبستگی متیوز نزدیک به ۱۰۰ درصد و در معیارهای خطای نوع اول و دوم، کمتر از ۲ درصد بودند. روش پیشنهادی با هسته گاوسی و خطی به ترتیب بهترین نتایج را به دست آوردند. نتایج این رویکرد نشان‌دهنده کارایی بیشتر دسته‌بندهای گروهی است.

همچنین روش گروهی پیشنهادی با روش معمولی SVM مقایسه شد و همان طور که انتظار می‌رفت روش گروهی با هسته‌های خطی و چندجمله‌ای در تمام شش معیار بهبود قابل توجهی ایجاد کرد. علاوه بر بهبود عملکرد، برای اجرای روش گروهی، زمان محاسباتی بسیار کمتری در مقایسه با روش معمولی نیاز خواهد شد و این نکته در مورد داده‌های بسیار حجیم اهمیت زیادی دارد. در این تحقیق زمان محاسباتی روش گروهی پیشنهادی تقریباً یک دهم روش معمولی SVM بود.

بر اساس تحلیل انجام شده، ویژگی چک برگشتی و نحوه بازپرداخت، در این شعبه بانک پاسارگاد از قدرت پیش‌بینی بسیار بالایی برای شناسایی مشتریان برخوردار است. اما چنانچه ویژگی چک برگشتی در اختیار نباشد، سایر ویژگی‌های مشتریان که در بخش قبل اشاره شد نیز به کمک روش‌های دقیق مانند روش گروهی پیشنهادی، می‌توانند به خوبی مشتریان خوش‌حساب و بدحساب را تفکیک کنند.

در این تحقیق بر روی داده‌های با حجم کم، روش گروهی SVM اعمال شد و چنانچه داده‌های حجیم در دسترس باشد ویژگی کاهش هزینه‌های محاسباتی به روشنی قابل مشاهده خواهد بود. به عنوان پیشنهاد برای تحقیقات آتی، پیشنهاد می‌شود روش گروهی پیشنهادی بر روی داده‌های تمامی شعب بانک پاسارگاد یا تمامی بانک‌ها استفاده شود. همچنین می‌توان بجای روش گروهی Bagging از روش‌های دیگر مانند ارتقا و یا رای‌گیری استفاده کرد.

تعارض منافع: نویسندگان اعلام می‌کنند که هیچ تعارض منافعی ندارند.

مراجع

[۱] ابوالحسنی م. ج.، صمدی س. و واعظبرزانی م.، «تعیین اثرات کوتاه‌مدت و بلندمدت متغیرهای کلان اقتصادی و بانکی بر حجم مطالبات معوق بانک‌های پذیرفته شده در بورس اوراق بهادار تهران (۱۳۹۶-۱۳۸۶)»، مطالعات اقتصادی کاربردی ایران، جلد ۱۰، شماره ۳۷، ص. ۲۰۱-۲۳۳، ۱۴۰۰.

[۲] بزرگ‌اصل م. و صمدی م. ت.، «رابطه بین ریسک نقدینگی و ریسک اعتباری و تاثیر آن بر ناپایداری مالی در صنعت بانکداری ایران»، فصلنامه پژوهش‌های پولی-بانکی، جلد ۱۰، شماره ۳۳، ص. ۵۰۹-۵۳۱، ۱۳۹۶.

[۳] فلاح‌شمس م. و مهدوی‌راد ح.، «طراحی مدل اعتبارسنجی و پیش‌بینی ریسک اعتباری مشتریان تسهیلات لیزینگ (مورد مطالعه: شرکت لیزینگ ایران خودرو)»، پژوهش‌نامه اقتصادی (رویکرد اسلامی-ایرانی)، جلد ۱۲، شماره ۴۴، ص. ۲۱۳-۲۳۴، ۱۳۹۱.

[۴] عیسی‌زاده س. و عریانی ب.، «رتبه‌بندی مشتریان حقوقی بانک‌ها بر حسب ریسک اعتباری به روش تحلیل پوششی داده‌ها: مطالعه موردی شعب بانک کشاورزی»، فصلنامه پژوهش‌ها و سیاست‌های اقتصادی، شماره ۵۵، ص. ۸۶-۵۹، ۱۳۸۹.

[۵] کشاورزحداد غ. و آیتی گازار ح.، «مقایسه کارکرد مدل لاجیت و روش درخت‌های طبقه‌بندی و رگرسیونی در فرآیند اعتبارسنجی متقاضیان حقیقی برای استفاده از تسهیلات بانکی»، پژوهش‌های رشد و توسعه پایدار (پژوهش‌های اقتصادی)، شماره ۴، ص. ۷۱-۹۷، ۱۳۸۶.

[۶] مهرآرا م.، موسایی م.، تصوری م. و حسن‌زاده ا.، «رتبه‌بندی اعتباری مشتریان حقوقی بانک پارسیان»، فصل‌نامه مدل‌سازی اقتصادی، جلد ۳، شماره ۹، ص. ۱۵۰-۱۲۱، ۱۳۸۸.

[7] Thomas L. C., "A survey of credit and behavioral scoring: Forecasting financial risks of lending to customers", International Journal of Forecasting, 16(2): 149-172, 2000.

[8] Kim Y. S. and Sohn S. Y., "Managing loan customers using misclassification patterns of credit scoring model", Expert Systems with Applications, 26(4): 567-573, 2004.

[۹] پورزمانی ز. و کلانتری ح.، «مقایسه‌ی قدرت پیش‌بینی بحران مالی توسط تکنیک‌های مختلف هوش مصنوعی»، پژوهش‌های حسابداری مالی و حسابرسی، شماره ۱۷، ص. ۶۴-۳۳، ۱۳۹۲.

[10] Shin K. S. and Lee Y. J., "A Genetic Application in Bankruptcy Prediction Modelling", Expert Systems with Applications, 23(3): 321-328, 2002.

[11] Reichert K. A., Cho C. C., and Wagner M. G., "An examination of the conceptual issues involved in developing credit-scoring models", Journal of Business and Economic Statistics, 1(2): 101-114, 1983.

[12] Huang C. l., Chen M. C., and Wang, C., "Credit scoring with a data mining approach based on support vector machines", Expert Systems with Applications, 33(4): 847-856, 2007.

[۱۳] طلوعی‌اشلقی ع.، نیکومرام ه. و مقدوری‌شربیانی ف.، «طبقه‌بندی متقاضیان تسهیلات اعتباری بانک‌ها با استفاده از تکنیک ماشین بردار پشتیبان»، مجله پژوهش های مدیریت، شماره ۸۴، ۱۳۸۹.

[۱۴] محمدیان‌حاجی‌کرد ا.، اصغرزاده‌زعفرانی م. و امام‌دوست م.، «بررسی ریسک اعتباری مشتریان حقوقی با استفاده از مدل ماشین بردار پشتیبان و مدل هیبریدی الگوریتم ژنتیک- مطالعه موردی بانک تجارت»، مهندسی مالی و اوراق بهادار، جلد ۷، شماره ۲۷، ص. ۱۷-۳۲، ۱۳۹۵.

[15] Tang L., Cai F., and Ouyang Y., "Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China", Technological Forecasting and Social Change, 144: 563-572, 2019.

[۱۶] آرخی ص. و ادیب‌نژاد م.، «ارزیابی کارایی الگوریتم‌های ماشین بردار پشتیبان جهت طبقه‌بندی کاربری اراضی با استفاده از داده‌های ماهواره‌ای ETM+ لندست (مطالعه موردی: حوزه سد ایلام)»، تحقیقات مرتع و بیابان ایران، جلد ۱۸، شماره ۳، ص. ۴۲۰-۴۴۰، ۱۳۹۰.

[۱۷] نوازی س. و رسول‌زادگان ع.، «ارائه یک روش جدید دو مرحله‌ای جهت تخمین هوشمند سن افراد»، مجله محاسبات نرم، جلد ۲، شماره ۲، ص. ۵۲-۶۱، ۱۳۹۲.

[۱۸] وثیقی‌ذاکر ا. و جلیلی س.، «پیش‌بینی ژن‌های بیماری با استفاده از دسته‌بند تک‌کلاسی ماشین بردار پشتیبان»، مجله محاسبات نرم، جلد ۴، شماره ۱، ص. ۷۴-۸۳، ۱۳۹۴.

[۱۹] ویسی ه.، قایدشرف ح. و ابراهیمی م.، « بهبود کارایی الگوریتم‌های یادگیری ماشین در تشخیص بیماری‌های قلبی با بهینه‌سازی داده‌ها و ویژگی‌ها»، مجله محاسبات نرم، جلد ۸، شماره ۱، ص. ۷۰-۸۵، ۱۳۹۸.

[۲۰] خسروی ا.، عبدالمالکی ه. و فیاضی م.، « پیش‌بینی وضعیت تحصیلی متقاضیان پذیرش‌شده دانشگاه، مبتنی بر داده‌های

آموزشی و پذیرشی با استفاده از تکنیک‌های داده کاوی»، مجله محاسبات نرم، جلد ۹، شماره ۲، ص. ۹۴-۱۱۳، ۱۳۹۹.

[21] Bao W., Lianju N., and Yue K., "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment", Expert Systems with Applications, 128: 301-315, 2019.