مجله مهندسی مالی و مدیریت اوراق بهادار
شماره هفدهم / زمستان ۱۳۹۲

رویکرد حداقل مربعات ماشین بردار پشتیبان مبتنی بر الگوریتم ژنتیک جهت تخمین رتبه اعتباری مشتریان بانک‌ها

احمد پویان‌فر[۱]
سعید فلاح پور[۲]
محمدرضا عزیزی[۳]

تاریخ دریافت: ۹۲/۴/۱۸ تاریخ پذیرش: ۹۲/۹/۲۵

چکیده

یکی از مهم‌ترین مسائلی که همواره بانک‌ها و مؤسسات مالی با آن مواجه هستند، مسئله ریسک اعتباری یا احتمال عدم ایفای تعهدات از سوی متقاضیان دریافت کننده تسهیلات اعتباری می‌باشد. رقم قابل توجه مطالبات معوق بانک‌ها در سراسر جهان نشان دهنده اهمیت این موضوع و لزوم توجه به آن می‌باشد. از این رو تاکنون تلاش‌های بسیاری به منظور ارائه مدلی کارا جهت ارزیابی و طبقه بندی هرچه دقیق‌تر متقاضیان تسهیلات اعتباری صورت گرفته است. هدف اصلی این پژوهش بکار گیری روش حداقل مربعات ماشین بردار پشتیبان مبتنی بر الگوریتم ژنتیک (Ga-LSSVM) در ارزیابی ریسک اعتباری متقاضیان تسهیلات اعتباری می‌باشد. بدین منظور از مجموعه داده‌های بانک آلمان در پایگاه داده یادگیری ماشین UCI جهت نمایش اثربخشی و دقت طبقه بندی کننده Ga-LSSVM استفاده شده است. نتایج مدل ارائه شده با مدل آماری لاجیت و رویکردهای بهینه سازی پارامترهای ماشین بردار پشتیبان مقایسه شده است. یافته‌های پژوهش حاکی از آن است که در ارزیابی ریسک اعتباری متقاضیان تسهیلات اعتباری، مدل Ga-LSSVM نسبت به مدل‌های بررسی شده از عملکرد مطلوبی برخوردار می‌باشد.

واژه‌های کلیدی: ریسک اعتباری، حداقل مربعات ماشین بردار پشتیبان، الگوریتم ژنتیک، انتخاب ویژگی.

۱- دکترای مدیریت مالی دانشگاه تهران، ایران apouyanfar@gmail.com
۲- دکترای مدیریت مالی دانشگاه تهران، ایران sfallahpour@gmail.com
۳- کارشناسی ارشدمهندسی صنایع- گرایش مالی، دانشگاه علوم اقتصادی azizi.mr87@gmail.com

۱- مقدمه

امروزه نهادهای مالی نقشی کلیدی را در توسعه و پیشرفت اقتصادی هر کشور ایفا می‌نمایند. بانک‌ها به عنوان اصلی‌ترین تأمین کننده منابع مالی، از مهم‌ترین ارکان نهادهای مالی محسوب می‌شوند. اهمیت و نقش این نهاد مالی به اندازه‌ای می‌باشد که عملکرد کارا و موثر صنعت بانکداری هر کشور به عنوان شاخصی از ثبات مالی آن محسوب می‌شود (مصیبی فر، فرزین وش، و کمیجانی، ۱۳۸۹).

یکی از اصلی‌ترین حوزه‌های فعالیت بانک‌ها، تجهیز منابع و تخصیص آن در قالب تسهیلات بانکی می‌باشد. بانک‌ها به عنوان وکیل و امین مردم، منابع خود، که قسمت عمده‌ی آن متعلق به سپرده‌گذاران می‌باشد را در قالب تسهیلات اعتباری در اختیار متقاضیان قرار می‌دهند. در پی افزایش سرمایه‌گذاری بانک‌ها در این حوزه‌ی درآمدی، احتمال بروز ریسک اعتباری ناشی از نکول مشتریان نیز افزایش خواهد یافت.

کمیته نظارت بانکی بازل[۱] (۲۰۰۱)، ریسک اعتباری را «احتمال عدم ایفای تعهدات متقاضی در نتیجه شرایط خاص که منجر به نکول وی می‌شود»، معرفی کرده است. این ریسک بانک‌ها را متحمل زیان‌های ناشی از عدم بازپرداخت یا بازپرداخت با تأخیر اصل یا فرع وام از سوی مشتری می‌سازد. هرچه اعتبارات اعطایی از سوی بانک‌ها افزایش یابد، بانک‌ها بیشتر در معرض ریسک اعتباری قرار گرفته و احتمال تجربه بحران مالی برای آنان افزایش می‌یابد. از این رو تصمیمات این نهاد مالی در خصوص اعطای اعتبار به متقاضیان خود، از اهمیت شایان توجهی برخوردار می‌باشد.

تاکنون مطالعات و پژوهش‌های گوناگونی برای ارائه مدلی کارا، جهت کمک به اتخاذ تصمیمات صحیح اعطای اعتبار از سوی مؤسسات مالی ارائه شده است، که افزایش دقت پیش بینی احتمال نکول متقاضیان تسهیلات، از مهم‌ترین اهداف آن‌ها بوده است. در همین راستا روش‌ها و رویکردهای پیشرفته‌ای جهت بهبود عملکرد مدل‌های اعتباری ارائه شده است. رویکردهای داده کاوی، یکی از شاخص‌ترین این روش‌ها می‌باشد که تلاش می‌کند دانش نهفته در داده‌های تاریخی را استخراج کرده و از آن در جهت پیش بینی احتمال نکول مشتری، بهره گیرد. از جمله تکنیک‌های داده کاوی روش ماشین بردار پشتیبان[۲] می‌باشد که در سال‌های اخیر مطالعات متعددی جهت ارزیابی ریسک اعتباری متقاضیان با استفاده از آن صورت گرفته است، که از آن میان می‌توان به مطالعات هوانگ و همکاران (۲۰۰۷)، یو و همکاران (۲۰۱۱) و ژو و همکاران (۲۰۱۰) اشاره کرد. نتایج تحقیقات مذکور بیانگر عملکرد مطلوب این روش نسبت به دیگر روش-های ارزیابی بوده است.

در این پژوهش تلاش می‌گردد با استفاده از رویکرد داده کاوی ماشین بردار پشتیبان، مدلی کارا و موثر جهت ارزیابی ریسک اعتباری متقاضیان تسهیلات بانکی ارائه شود. بدین منظور پس از مقدمه حاضر، مروری جامع بر مبانی نظری و پیشینه پژوهش خواهیم داشت. سپس در بخش سوم، به معرفی مدل پژوهش، ویژگی‌ها و شناخت آن پرداخته می‌شود و در بخش چهارم نتایج پژوهش بیان شده و نهایتاً در بخش پایانی که نتیجه‌گیری و بحث است به تفسیر نتایج حاصل پرداخته خواهد شد.

۲- مبانی نظری و مروری بر پیشینه پژوهش

اندازه‌گیری و درجه‌بندی ریسک اعتباری برای نخستین بار در سال ۱۹۰۹ میلادی توسط جان موری بر روی اوراق قرضه انجام شد (Glantz, 2003). یکی از قدیمی‌ترین مؤسساتی که اقدام به رتبه بندی اوراق قرضه نمود، موسسه مودیز[۳] است که در سال ۱۹۰۹ تاسیس شد. به دلیل شباهت زیاد اوراق قرضه و تسهیلات اعطایی، درجه بندی اعتباری به معنای اندازه گیری ریسک عدم پرداخت اصل و بهره (سود) تسهیلات، مورد بررسی محققین قرار گرفت (فلاح شمس، ۱۳۸۷).

نظام امتیازدهی اعتباری نخستین بار در دهه ۱۹۵۰ تدوین شد، اما استفاده فراگیر از آن حدود دو دهه به درازا انجامید. در حقیقت، پایه های تاریخ ۷۰ ساله امتیازدهی اعتباری بر مقاله فیشر[۴] (۱۹۳۶) که در آن تمایز پذیری گروه‌های مشتریان بر اساس مشخصه‌های مختلف مورد بررسی قرار گرفت، بنا شده است (Fisher, 1936). پس از وی دورند[۵] (۱۹۴۱) با تکیه بر نتایج فیشر (۱۹۳۶)، اولین فردی بود که از تحلیل ممیزی برای ایجاد سیستم امتیازدهی استفاده کرد. وی نشان داد که این روش در پیش بینی بازپرداخت‌های اعتباری، از دقت مطلوبی برخوردار می‌باشد (Hand & Henley, 1997). از این جهت می‌توان گفت دورند (۱۹۴۱) ابداع کننده‌ی سیستم‌های امتیازدهی اعتباری امروزی است.

بی ور[۶] در سال ۱۹۶۷، در مقاله‌ای که در زمینه برآورد موفقیت و شکست شرکت‌ها با استفاده از برخی شاخص‌های مالی بود، نخستین مدل به کار رفته برای تعیین ورشکستگی شرکت‌ها، مدل رگرسیون چند متغیره[۷]، را ارائه نمود (Beaver, 1967).

آلتمن[۸] (۱۹۶۸) از دیگر پیشگامان در زمینه اندازه‌گیری ریسک اعتباری اوراق قرضه شناخته شده است. آلتمن (۱۹۶۸) جهت یافتن یک رابطه معنادار بین متغیرهای حسابداری یک شرکت و احتمال عدم توانایی در پرداخت دیون این شرکت در آینده، تلاش بسیاری انجام داد که منجر به ارائه مدل نمره Z[۹] توسط وی شد.

رهنمای رودپشتی و همکاران (۱۳۸۸) در پژوهش خود، از مدل‌های آلتمن و فالمر جهت پیش بینی ورشکستگی شرکت‌ها استفاده کردند که نتایج حاصله حاکی از آن است که در پیش بینی یک شرکت، تفاوت معنی داری بین نتایج دو مدل وجود دارد. همچنین مدل آلتمن در پیش بینی ورشکستگی محافظه کارانه از مدل فالمر عمل می‌کند.

اواخر دهه ۷۰ مدل‌های دو گزینه‌ای و چند گزینه‌ای لاجیت و پروبیت یا احتمال خطی برای پیش بینی احتمال ناکامی (شکست) بنگاه معرفی شدند. همچنین در دهه ۸۰ مطالعات الگوریتم-های تقسیم بندی بازگشتی[۱۰] بر پایه‌ی درخت طبقه بندی به کار گرفته شد.

در خلال دهه ۸۰، ویگینتن (۱۹۸۰) اولین مطالعه‌ی مقایسه لجستیک با مدل تحلیل ممیزی را ارائه کرد. بر اساس مطالعه‌ی وی، روش لجستیک در زمینه طبقه بندی نتیجه عملکرد بهتری دارد.

عرب مازار و روئین تن (۱۳۸۵) به بررسی اطلاعات کیفی و مالی یک نمونه تصادفی ۲۰۰ تایی از ۱۳۸۳ از شعب بانک کشاورزی استان تهران پرداختند. نتایج آن‌ها نشان می‌دهد مدل لاجیت در برآورد عوامل مؤثر بر ریسک اعتباری از توان بالایی برخوردار است. فلاح شمس و مهدوی راد (۱۳۸۹) در پژوهش خود با استفاده از مدل‌های اقتصادسنجی لاجیت و پروبیت، مدل پیش بینی ریسک اعتباری مشتریان حقیقی تسهیلات لیزینگ طراحی نمودند. نتایج نشان داد که کارایی مدل لاجیت با استفاده از داده های تخمین بیش از مدل پروبیت می‌باشد.

در اواخر دهه‌ی ۹۰ روش تحلیل ناپارامتری پوششی داده‌ها[۱۱] برای تحلیل رتبه بندی اعتباری معرفی شد. یه[۱۲] (۱۹۹۶) یکی از پیشگامان ترکیب تحلیل پوششی داده‌ها با آنالیز نسبت‌های مالی است. صفری و همکاران (۱۳۸۹) به منظور رتبه بندی اعتباری مشتریان بانک از رویکرد تحلیل ممیزی استفاده کردند. با مقایسه رتبه‌های حاصل از به‌کارگیری معادله رگرسیونی با رتبه‌های به دست آمده از روش تحلیل پوششی داده‌ها، ملاحظه شد که تفاوت معناداری میان مقادیر محاسبه شده و واقعی وجود ندارد و این مسئله دلالت بر تایید فرضیه کارایی مدل تحلیل پوششی داده‌ها در رتبه بندی اعتباری مشتریان حقوقی بانک تجارت می‌کند.

فلاح‌پور و راعی (۱۳۸۳) با استفاده از شبکه های عصبی مصنوعی، به پیش بینی درماندگی مالی شرکت‌های تولیدی پرداختند. نتایج حاصله از مدل‌ها، بر اساس اطلاعات ۸۰ شرکت، نشان داد که مدل شبکه عصبی مصنوعی در پیش بینی درماندگی مالی، به طور معنی داری نسبت به مدل تحلیل ممیز چندگانه از دقت پیش بینی بیشتری برخوردار است.

در سال ۲۰۰۳، هانگ و همکاران[۱۳] (۲۰۰۳) از اولین محققانی بودند که به منظور تجزیه و تحلیل رتبه بندی ریسک اعتباری از ماشین بردار پشتیبان که یک روش یادگیری ماشین است، استفاده کردند. آن‌ها به منظور پیش بینی وضعیت اعتباری مشتریان بازارهای ایالات متحده، از

ماشین بردار پشتیبان استفاده کردند و نتایج آن را با شبکه عصبی پیش خور مقایسه نمودند. طبق مطالعات آن‌ها، ماشین بردار پشتیبان همانند شبکه عصبی به دقت پیش بینی نزدیک به ۸۰ درصد دست یافت.

راعی و فلاح‌پور (۱۳۸۷) در پژوهش خود با هدف پیش بینی درماندگی مالی شرکت‌ها، به بررسی نتایج مدل ماشین بردار پشتیبان در مقایسه با مدل آماری رگرسیون لجستیک پرداختند. یافته های آن‌ها حاکی از آن است که مدل ماشین بردار پشتیبان نسبت به مدل رگرسیون لجستیک، نه تنها از دقت کلی بهتری برخوردار است، بلکه توانایی بالاتری نیز در تعمیم پذیری دارد.

از آن جهت که تخمین دقیق پارامترهای SVM بسیار حائز اهمیت است، هوانگ و همکاران (۲۰۰۷) از الگوریتم ژنتیک جهت تخمین هر چه دقیق‌تر پارامترهای SVM استفاده کردند و عملکرد آن را با روش‌های SVM کلاسیک، برنامه ریزی ژنتیک[۱۴] (GP) و شبکه عصبی پیش خور[۱۵] (BPN) مقایسه نمودند. در این پژوهش مجموعه داده بانک آلمان در نظر گرفته شده است. به منظور مقایسه دقت روش‌های فوق از آزمون فریدمن استفاده شده، که بر اساس آن، دقت طبقه بندی سه روش مشابه یکدیگر می‌باشد، درحالی‌که دقت روش GA-SVM اندکی از دو روش دیگر بهتر است.

ژو و همکاران (۲۰۱۰) به منظور ارزیابی ریسک اعتباری متقاضیان تسهیلات اعتباری، از روش LS-SVM استفاده کرده و نتایج آن را با روش‌های شبکه عصبی، تحلیل ممیزی و SVM مقایسه نمودند. بر اساس یافته‌های آن‌ها، روش LSSVM نسبت به دیگر روش‌های، نتایج بهتری را نشان می‌دهد. ون گستل و همکارانش (۲۰۰۳) از روش LS-SVM جهت پیش بینی ورشکستگی بهره بردند. پژوهش آن‌ها نشان داد که روش LS-SVM نسبت به روش‌های تحلیل ممیزی و لاجیت در طبقه بندی شرکت‌ها، دقیق‌تر می‌باشد.

در این پژوهش جهت ارزیابی ریسک اعتباری مشتریان، از روش حداقل مربعات ماشین بردار پشتیبان مبتنی بر الگوریتم ژنتیک همراه با انتخاب ویژگی بهره گرفته‌شده و سپس عملکرد آن با روش پارامتری لاجیت و همچنین دیگر روش‌های بهینه سازی حداقل مربعات ماشین بردار پشتیبان مقایسه گردیده است.

ریسک اعتباری

با توجه به تعاریف متعددی که تا کنون از ریسک ارائه شده است، می‌توان دریافت که هر یک از محققان به فراخور حال، تعریف مورد نظر خود را با اقامه دلایل و مباحث گسترده مطرح کرده‌اند

(فلاح شمس و رشنو، ۱۳۸۷). به طور کلی ریسک از نظر لغوی به معنای «احتمال وقوع چیزی بد یا نامطلوب و یا احتمال وقوع خطر» می‌باشد. ریسک‌هایی که نهادها و مؤسسات مالی را مورد تهدید قرار می‌دهند، بر اساس هدف دسته بندی، در دسته‌های گوناگونی طبقه بندی می‌شوند. در دسته بندی که به تائید کمیته نظارتی بازل (BCBS) رسیده است، انواع ریسک در چهار دسته ریسک اعتباری، ریسک بازار، ریسک نقدشوندگی و ریسک عملیاتی قرار می‌گیرند.

ریسک اعتباری، احتمال عدم بازپرداخت یا پرداخت همراه با تأخیر اصل و فرع تسهیلات اعطایی بانک‌ها و سایر ابزار بدهی از سوی مشتری می‌باشد. اهمیت این ریسک به حدی است که از آن به عنوان بزرگ‌ترین ریسک مرتبط با فعالیت‌های مالی و بانکی یاد می‌شود. در سراسر فعالیت‌های بانک‌ها، منابع مختلفی در بروز وقوع این ریسک دخیل می‌باشند، اما با این حال، تسهیلات بزرگ‌ترین و بدیهی‌ترین منشأ ایجاد ریسک اعتباری برای اغلب بانک‌ها می‌باشد. بنابراین مؤسسات اعتباری برای در اختیار قرار دادن انواع تسهیلات اعطایی به مشتریان خود، نیازمند شناسایی خصوصیات متقاضیان تسهیلات اعتباری از ابعاد کیفی و کمی می‌باشند، تا از این طریق ارزیابی کاملی از سنجش توان بازپرداخت و محاسبه احتمال عدم بازپرداخت تسهیلات و خدمات تأمین مالی از سوی آنان، به عمل آید.

رتبه‌بندی اعتباری

رتبه‌بندی اعتباری به عنوان یکی از ابزارهای کارآمد در مدیریت ریسک اعتباری مطرح می‌باشد. این روش با استفاده از داده‌های گذشته و روش‌های آماری، سعی بر مجزا کردن آثار مشخصه‌های متفاوت متقاضیان بر نکول تسهیلات دارد. رتبه بندی اعتباری، این امکان را برای بانک یا موسسه‌های مالی به وجود می‌آورد که با کمک آن بتوانند متقاضیان یا مشتریان را بر اساس ریسک مربوط به هر یک از آن‌ها رتبه بندی بنمایند (طالبی و شیرزادی، ۱۳۹۰).

رتبه بندی‌های اعتباری، چه توسط مؤسسات رتبه بندی انجام شود و چه توسط بانک‌ها، یک هدف مشخص را دنبال می‌کنند و آن تعیین توانایی اعتبار گیرنده در بازپرداخت بدهی می‌باشد. به عبارت دیگر رتبه بندی، فرآیند کمی نمودن احتمال قصور مشتری در بازپرداخت تعهدات خود در آینده است. بنابراین رتبه اعتباری، کیفیت وام گیرنده و دورنمای بازپرداخت آن را نشان می‌دهد.

تا کنون به منظور ارزیابی ریسک اعتباری، مدل‌های رتبه بندی گوناگونی ارائه شده است که می‌توان آن‌ها را در دو دسته پارامتری[۱۶] و ناپارامتری[۱۷] طبقه بندی نمود. مدل‌های پارامتری به مدل‌هایی اطلاق می‌شود که هدف آن‌ها به دست آوردن و تخمین پارامترها به منظور طبقه بندی مشتریان است و در مقابل مدل‌هایی که در آن‌ها پارامتری محاسبه نشده و هدف آن‌ها طبقه بندی

مشتریان با توجه به روابط بین متغیرهاست، مدل‌های ناپارامتری نامیده می‌شوند (Bridges & Disney, 2001).

بر مبنای این طبقه بندی، مدل‌های رتبه بندی اعتباری پارامتری عبارتند از:

- مدل احتمال خطی[۱۸]
- مدل لاجیت[۱۹]
- مدل پروبیت[۲۰]
- مدل‌های تحلیل ممیزی یا تحلیل تفاوت‌ها[۲۱]

مدل‌های تحلیل ممیزی و رگرسیون خطی در میان طیف وسیع از مدل‌های پارامتری بیشترین کاربرد را دارند. همچنین مدل‌های رتبه بندی اعتباری ناپارامتری عبارتند از:

- روش برنامه ریزی ریاضی[۲۲]
- مدل طبقه بندی درختی یا درخت تصمیم گیری[۲۳]
- الگوهای نزدیک‌ترین همسایگان[۲۴]
- فرآیند سلسله مراتب تحلیل[۲۵]
- سیستم‌های هوش مصنوعی[۲۶]

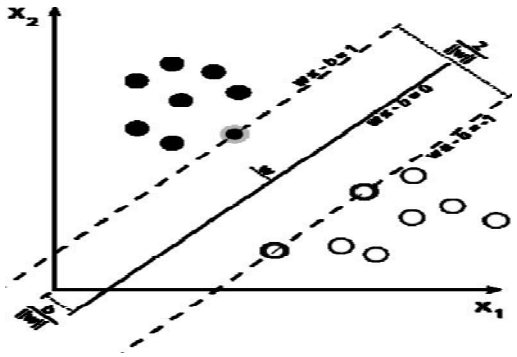
هر یک از مدل‌های عنوان شده دارای معایب و مزایای مختص به خود می‌باشند که همین امر بیان این که کدام یک توانایی بهتری در پیش بینی نکول دارد را دشوار می‌سازد.

ماشین بردار پشتیبان

ماشین بردار پشتیبان در واقع یک طبقه بندی کننده دودویی است که دو کلاس را با استفاده از یک مرز خطی از هم جدا می‌کند. در تقسیم خطی داده‌ها هدف دست‌یابی به تابعی است که تعیین کننده ابر صفحه‌ای[۲۷] با بیشترین حاشیه می‌باشد. با حداکثر شدن حاشیه این ابر صفحه، تفکیک بین طبقات حداکثر می‌گردد. فرض کنید که $S = \{x_i, y_i\}$ یک نمونه‌ی آموزشی است که از دو کلاسِ $y_i = \pm 1$ و هر کلاس از $i = 1, \ldots, m$ ویژگی تشکیل شده است.

همان‌گونه که در شکل ۱ نشان داده شده است، خط $(< w.x_i > +b) = 0$ داده‌های موجود را در دو کلاس ± 1 طبقه‌بندی می‌کند. به این خط ابر صفحه جدا کننده گفته می‌شود. دو خط $(< w.x_i > +b) = +1$ و $(< w.x_i > +b) = -1$ به ترتیب بیانگر مرز ناحیه‌ی دسته‌های $y = +1$ و $y = -1$ می‌باشند. به نزدیک‌ترین داده‌های آموزشی به ابر صفحه‌های جدا کننده، بردار پشتیبان[۲۸] نامیده می‌شوند.

در ماشین بردار پشتیبان به دو طریق خطی و غیرخطی می‌توان مجموعه نقاط را از یکدیگر جدا نمود (Avci. Engin, 2009) . در ادامه به شرح هر یک از روش‌های فوق پرداخته شده است.



شکل ۱ .ابر صفحه‌ای جدا کننده دو گروه +۱ و -۱

- **تفکیک خطی**

در حالتی که داده‌ها را بتوان به صورت خطی از هم جدا کرد، ماشین بردار پشتیبان با در نظر گرفتن مجموعه داده‌های آموزشی، با استفاده از حل مسئله بهینه سازی زیر ابر صفحه بهینه با حاشیه حداکثر را پیدا می‌نماید:

$$\begin{aligned} &\text{Min } 1/2\,\|w\|^2 \\ &S.t \quad y_i(< w.x_i > +b) \geq 1 \quad i = 1,2,...,\ m \end{aligned} \qquad (۱)$$

که در آن $\|w\|$ نرم اقلیدسی[۲۹] است. بر اساس مدل فوق، کمینه شدن مقدار $\|w\|$ با توجه به محدودیت آن، منجر به حداکثر شدن پهنای حاشیه صفحه می‌گردد.

حل این مسئله بهینه سازی دشوار است، لذا به منظور ساده‌تر نمودن حل آن، از ضرایب لاگرانژ استفاده می‌گردد. سپس دوگان آن با توجه به شرط کروش کان-تاکر[۳۰]، به صورت زیر در خواهد آمد:

$$\begin{aligned} &\max \quad W(\alpha) = \sum \alpha_i - \frac{1}{2}\sum \sum \alpha_i \alpha_j y_i y_j < x_i.x_j > \\ &Subject \quad to: \sum \alpha_i y_i = 0\ ,\ \alpha_i \geq 0\ ,\ i = 1,\dots,N \end{aligned} \qquad (۲)$$

با توجه به عدم وجود b در معادله فوق، از محدودیت‌های ابتدایی برای بدست آوردن آن استفاده می‌کنیم. پس از آنکه مقادیر α_i و b بدست آمد، می‌توان SVM را برای دسته بندی نمونه‌های جدید بکار برد. اگر x یک نمونه جدید باشد، دسته بندی آن به صورت زیر مشخص می‌گردد.

$$F(x) = sign\ [f(x, \alpha, b)] \qquad (۳)$$

که در آن $f(x,\alpha,b)$ به صورت زیر می‌باشد:

$$f(x,\alpha,b) = w.x + b = \sum \alpha_i y_i x_i x + b = \sum_{i \in SV} \alpha_i y_i x_i x + b \qquad (۴)$$

همان‌گونه که قبلاً عنوان شد، SV بردارهای پشتیبان می‌باشند.

- **تفکیک غیر خطی**

یک فرض بسیار مهم در SVM این است که داده‌ها به صورت خطی جدا پذیر باشند. در حالی که در عمل در بیشتر مواقع این فرض صحیح نیست. برای اولین بار کورتس و وپنیک در سال ۱۹۹۵ به منظور اینکه بتواند در این حالت یک ابر صفحه‌ی بهینه را برای جداسازی ۲ کلاس بدست بیاوردند، متغیرهای نا منفی $\varepsilon_i \geq 0$ را به عنوان مقادیر خطا برای هر بردار تعریف نمودند (Cortes, 1995). مطابق روش ارائه شده صورت مسئله بهینه سازی تبدیل می‌شود به یافتن wبه نحوی که معادله زیر حداقل گردد:

$$\begin{aligned} &\text{Min } 1/2 \|w\|^2 + C\sum_i \varepsilon_i \\ &S.t \quad y_i(< w.x_i > + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0 \quad \forall i \end{aligned} \qquad (۵)$$

که در آن C پارامتر تنظیم کننده حاشیه می‌باشد که وظیفه‌ی آن برقراری تعادل بین حداکثر کردن حاشیه و حداقل کردن خطای دسته بندی را بوده و همواره بزرگ‌تر از صفر است (Cristianini & Shawe-Taylor, 2000). اگر C عددی بزرگ انتخاب شود، توجه بیشتری به خطا معطوف می‌گردد. ماشین بردار پشتیبانی که به این صورت تعریف شده باشد را ماشین بردار پشتیبان حاشیه- نرم[۳۱] می‌نامند (Abe, 2005).

همچنین در حالتی که داده‌ها جداناپذیر بوده و همچنین کلاس‌ها دارای همپوشانی هستند، جدا کردن کلاس‌ها توسط مرز خطی همواره با بروز خطا همراه می‌باشد. به منظور حل مشکل

مزبور می‌توان ابتدا داده‌ها را با استفاده از یک تبدیل غیر خطی φ، از فضای اولیه به فضایی با بعد بالاتر منتقل کرد با این هدف که در فضای جدید، کلاس‌ها تداخل کمتری با یکدیگر داشته باشند (Burges, 1998). پس از نگاشت داده‌ها به فضای بالاتر، با استفاده از معادلات قبل و جایگزین کردن x_i با $\varphi(x_i)$، ابر صفحه بهینه بدست خواهد آمد (Bellotti & Crook, 2008; Chen, Ma, & Ma, 2008).

انتقال از بعد پایین به بعدی بالاتر را می‌توان توسط توابع کرنل[۳۲] انجام داد. تعدادی از توابع کرنل موجود به شرح ذیل می‌باشند:

کرنل خطی	(۶)	$K(x_i, x_j) = x_i^T x_j + c$
کرنل چند جمله‌ای	(۷)	$K(x_i, x_j) = (\alpha\, x_i^T x_j + c)^d$
کرنل تابع پایه شعاعی	(۸)	$K(x_i, x_j) = \exp(-\left\|x_i - x_j\right\|^2 \Big/ \sigma^2)$

در معادلات فوق، c عبارت ثابت، d درجه چند جمله‌ای و سیگما (σ) پارامتر قابل تنظیم می‌باشد.

کرنل تابع پایه شعاعی در دو مقوله نسبت به کرنل‌ها دارای نقطه قوت بوده و در نتیجه عملکرد مطلوب‌تری در مسائل طبقه بندی اعتباری از خود نشان داده است. این کرنل به صورت غیر خطی نمونه‌ها را به ابعاد فضایی بالاتری نگاشت می‌نماید، بنابراین بر خلاف کرنل خطی، در حالت‌هایی که روابط بین کلاس‌ها و ویژگی‌ها به صورت غیر خطی است، کاربرد خواهد داشت. دومین نقطه قوت، تعداد پارامترهایی است که بر پیچیدگی انتخاب مدل تأثیر گذار می‌باشند. کرنل چند جمله‌ای تعداد پارامترهای بیشتری نسبت به RBF دارا می‌باشد (Lean, Xiao, Shouyang, & K.K., 2011). از این رو در این پژوهش از کرنل تابع پایه شعاعی استفاده شده است. بنابراین مسئله بهینه سازی به صورت زیر تعریف می‌شود:

$$\begin{aligned} \text{Max} \quad & \sum \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ S.t \quad & \sum_i \alpha_i y_i = 0, \alpha_i \geq 0 \\ & 0 \leq \alpha_i \leq C, \forall i \end{aligned} \qquad (۹)$$

برای دسته بندی داده جدیدی همانند x، می‌توان تصمیم گیری را بر اساس تابع زیر انجام داد:

$$f(x) = sign\sum \alpha_i y_i k(x_i, x) + b \qquad (۱۰)$$

حداقل مربعات ماشین بردار پشتیبان

همان‌طور که قبلاً نیز گفته شد، برای حل مدل ماشین بردار پشتیبان از برنامه ریزی درجه دو استفاده می‌شود. زمانی که بخواهیم مسئله‌ای در ابعاد بزرگ را محاسبه نماییم، ممکن است با محاسباتی پرهزینه مواجه شویم. در همین خصوص، روش حداقل مربعات[۳۳] توسط سویکنز و وندوال (۱۹۹۹) برای ماشین‌های بردار پشتیبان طراحی شد که مسئله بهینه سازی آن به صورت زیر می‌باشد.

$$\begin{aligned} &\text{Min } 1/2\,\|\mathrm{w}\|^2 + \frac{C}{2}\sum_i \varepsilon_i{}^2 \\ &S.t \quad y_i(< w.x_i > +b) \geq 1-\varepsilon_i, \varepsilon_i \geq 0 \quad \forall i \end{aligned} \qquad (۱۱)$$

بر اساس این روش، مسائل برنامه ریزی درجه دو فوق را می‌توان به سادگی توسط مجموعه‌ای از معادلات خطی حل کرد. بر اساس معادله (۱۱) تابع لاگرانژ آن به صورت زیر می‌باشد:

$$L(w,b,\varepsilon_i,\alpha_i) = \frac{1}{2}w^T w + \frac{C}{2}\sum \varepsilon_i^2 - \sum \alpha_i \left[y_i(w^T\varphi(x_i)+b) - 1 + \varepsilon_i\right] \qquad (۱۲)$$

که در آن α_i، i امین ضریب لاگرانژ می‌باشد.
به منظور ساده‌تر کردن حل معادله فوق، فرضیات زیر را در نظر می‌گیریم.

$$\begin{aligned} &Y = (y_1, y_2, ..., y_N)^T, \quad \alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T, \quad 1 = (1,1,\dots,1)^T, \\ &\Omega_{ij} = y_i y_j^T \varphi(x_i)^T \varphi(x_j) + ((1)/C)I \\ &\quad = y_i y_j^T K(x_i, x_j) + (1/C)I, i = 1,2,\dots,N \end{aligned} \qquad (۱۳)$$

در معادله فوق $K(x_i, x_j)$ تابع کرنل و I ماتریس واحد می‌باشد. با توجه به معادله (۱۳) خواهیم داشت:

$$\begin{bmatrix} \Omega & Y \\ Y^T & 0 \end{bmatrix}\begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad (۱۴)$$

با توجه به معادله (۱۳)، پارامتر Ω مثبت می‌باشد، ضریب لاگرانژ را می‌توان با استفاده از معادله (۱۴) بدست آورد، که فرمول آن به صورت زیر می‌باشد:

$$\alpha = \Omega^{-1}(1 - bY) \qquad (۱۵)$$

با جایگذاری معادله (۱۵) با ماتریس دوم معادله (۱۴)، معادله زیر حاصل می‌گردد:

$$b = \frac{Y^T \Omega^{-1} 1}{Y^T \Omega^{-1} Y} \qquad (۱۶)$$

با توجه به این که Ω معین مثبت می‌باشد، Ω^{-1} نیز معین مثبت است. بعلاوه چون Y برداری غیر صفر است، عبارت $Y^T\Omega^{-1}Y$ بزرگ‌تر از صفر خواهد بود. بنابراین b همواره بدست خواهد آمد. با استفاده از w و b، دسته بندی کننده حداقل مربعات ماشین بردار پشتیبان در فضای ورودی به صورت زیر خواهد بود:

$$F(x) = sign[f(x)] = sign\left[\sum \alpha_i y_i K(x, x_i) + b\right] \qquad (۱۷)$$

الگوریتم ژنتیک

الگوریتم ژنتیک یکی از الگوریتم‌های جستجوی تصادفی است که ایده آن برگرفته از طبیعت و بر اساس بقای برترین‌ها یا انتخاب طبیعی استوار می‌باشد. این الگوریتم که اولین بار توسط جان هولند[۳۴] (۱۹۷۵) ارائه شد، یک روش جستجوی موثر در فضاهای بسیار بزرگ ایجاد می‌کند که در نهایت منجر به جهت گیری به سمت یافتن جواب بهینه می‌گردد (Michael, 1999).

اساس الگوریتم ژنتیک تبدیل هر مجموعه جواب به یک کدگذاری دودویی (رشته‌های بیتی) است، که اصطلاحاً آن را کروموزوم[۳۵] می‌نامند. در الگوریتم‌های ژنتیکی، هر کروموزوم نشان دهنده یک نقطه در فضای جستجو و یک راه‌حل ممکن برای مسئله مورد نظر است. اجرای الگوریتم با ایجاد یک مجموعه ابتدایی از جواب‌های تصادفی که جمعیت اولیه[۳۶] گفته می‌شود، شروع می‌گردد. در هر تکرار این الگوریتم، مجموعه جدید از کروموزوم‌ها تولید می‌شود که به آن‌ها نسل[۳۷] گفته می‌شود. طی هر نسل میزان برازش کروموزوم‌ها با تابع برازندگی[۳۸] تعیین می‌گردد. سپس طی فرآیند باز تولید[۳۹]، عملگرهای ژنتیک که شامل، آمیزش[۴۰] و جهش[۴۱] می‌باشند، بر روی کروموزوم‌ها اعمال می‌شوند. به هر یک از کروموزوم‌های تولید شده در این مرحله نوزاد[۴۲] گفته می‌شود. سپس هر یک توسط تابع برازندگی ارزیابی شده و به وسیله عملگر انتخاب[۴۳] کروموزوم‌های برتر انتخاب شده و به نسل بعد انتقال می‌یابند.در نهایت الگوریتم به بهترین کروموزوم همگرا می‌شود که نمایانگر جواب بهینه یا زیر بهینه مسئله است.

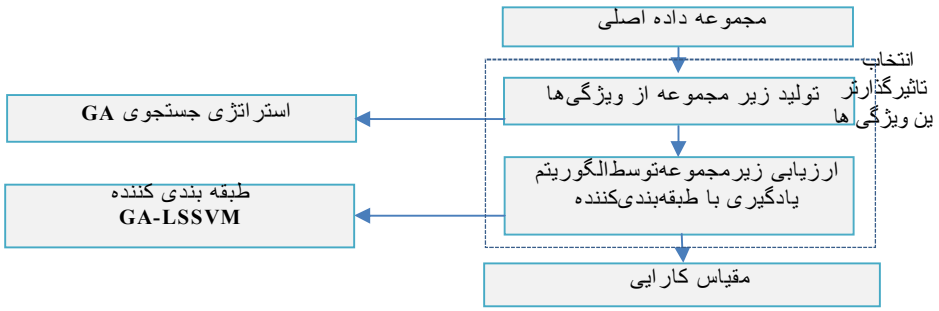
۳- مدل پژوهش، ویژگی‌ها و شناخت آن

اعتبارسنجی شامل طبقه بندی و رتبه بندی متقاضیان تسهیلات اعتباری در دو گروه خوب و بد بر اساس شرایط متقاضی از قبیل موقعیت کاری، اطلاعات شخصی، سن و غیره که به عنوان متغیرهای طبقه بندی مطرح می‌باشند، در نظر گرفته می‌شود. در این بخش به شرح مدل ارائه شده پژوهش، به منظور دسته بندی جامعه آماری متقاضیان تسهیلات اعتباری بانک‌ها و تعیین وضعیت اعتباری آن‌ها پرداخته می‌شود.

نمونه آماری این پژوهش از پایگاه داده یادگیری ماشین UCI جمع آوری شده است. این نمونه آماری شامل متقاضیان اعتباری بانک آلمان می‌باشد. مجموعه داده جهانی بانک آلمان شامل ۱۰۰۰ متقاضی بوده که ۷۰۰ نمونه آن در دسته خوب و ۳۰۰ نمونه دیگر در دسته بد قرار گرفته‌اند و برای هر نمونه ۲۰ ویژگی در نظر گرفته شده است. پایگاه داده یادگیری ماشین UCI در آدرس (http://archive.ics.uci.edu/ml) قابل دسترس می‌باشد. به منظور غنی‌تر شدن مدل‌های ایجاد شده، از ۴ دسته نمونه فرعی استفاده شده است. در هر نمونه فرعی، ۷۵۰ نمونه برای آموزش مدل و ۲۵۰ نمونه برای آزمایش قدرت مدل استفاده گردیده است. به عبارت دیگر، در هر نمونه فرعی جای ۲۵۰ نمونه از مجموعه آموزشی با ۲۵۰ نمونه آزمایشی تعویض می‌گردد. بنابراین، با توجه به این که برای هر ترکیب یک مدل ارائه می‌شود، در مجموع برای هر یک از رویکردهای ارائه شده چهار مدل مختلف ارائه گردیده است.

انتخاب ویژگی

روش‌های انتخاب ویژگی به دو دسته رَپر (Wrapper) و فیلتر طبقه بندی می‌شود. در روش Wrapperکه به جعبه سیاه[۴۴] معروف است، از یک تابع دسته بندی برای ارزیابی شایستگی زیرمجموعه های ویژگی استفاده می‌گردد. تفاوت اصلی دو روش فیلتر و Wrapper در دو مقوله می‌باشد. اول اینکه در روش فیلتر انتخاب بهترین ویژگی‌ها بر اساس معیاری مستقل از معیار برازندگی اصل مسئله می‌باشد، اما در روش Wrapper انتخاب بهترین ویژگی‌ها بر اساس معیار نهایی بوده و معیار برازندگی برای هر یک از زیرمجموعه‌های انتخابی مسئله اصلی می‌باشد.



شکل ۲. فلوچارت روش Wrapper در انتخاب ویژگی

از آنجایی که روش Wrapper می‌تواند خودش را با الگوریتم یادگیری ماشین استفاده شده تطبیق دهد، معمولاً نسبت به روش فیلتر نتایج بهتری را ارائه می‌دهد. در روش فیلتر، انتخاب ویژگی به عنوان یک مرحله پیش پردازشی انجام می‌شود. از معایب روش فیلتر این است که اثرات زیرمجموعه ویژگی انتخاب شده را بر روی کارایی الگوریتم استقرایی در نظر نمی‌گیرد (Liu & Motoda, 1998). فلوچارت کلی استفاده از روش انتخاب ویژگی در شکل زیر نشان داده شده است.

با توجه به نکات ذکر شده، در این مقاله از روش Wrapper برای ایجاد مدل ارائه شده استفاده شده است.

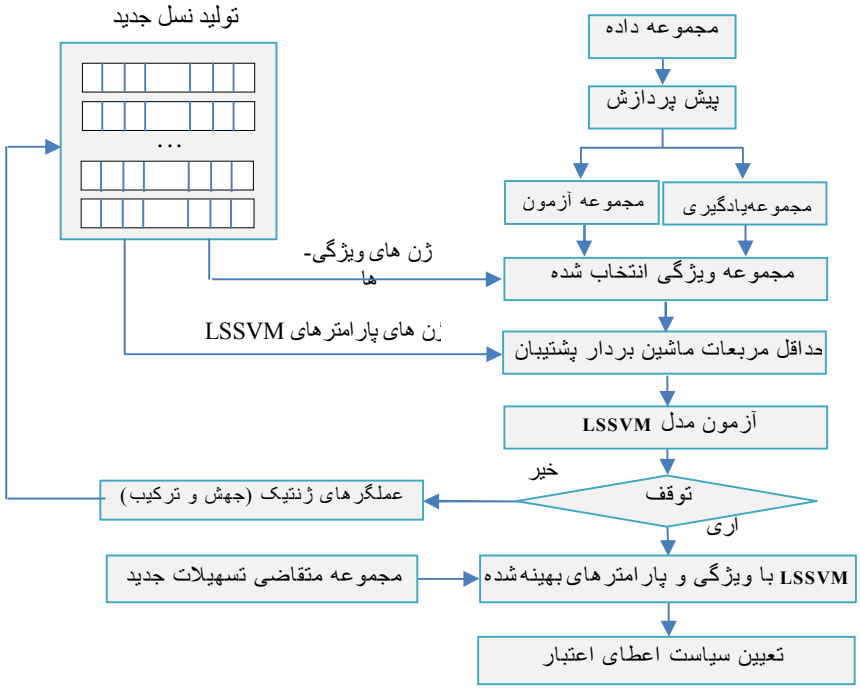
مدل پژوهش

رویه رتبه‌بندی توسط مدل ارائه شده به این صورت است که در جهت افزایش قابلیت تفسیر و تعمیم‌پذیری مدل، از الگوریتم ژنتیک برای انتخاب زیر مجموعه بهینه از ویژگی‌های ورودی و همچنین بهینه کردن پارامترهای LSSVM بهره گرفته می‌شود. در ادامه به شرح گام‌هایی که به منظور پیاده سازی این رویکرد می‌باید پیموده شوند پرداخته شده است.

در گام اول، بر روی مجموعه داده‌ها پیش پردازش صورت گرفته و آن‌ها نرمال می‌شوند. به طور کلی هر ویژگی را می‌توان به صورت خطی در محدوده‌ی [۱,۱-] یا [۰,۱] توسط معادله (۱۸) مقیاس بندی کرد(Liu & Motoda, 1998) .

$$v' = \frac{v - \min_a}{\max_a - \min_a} \qquad (۱۸)$$

که در آن v مقدار اصلی، v' مقدار مقیاس بندی شده، $\max_a$ حد بالای ویژگی و $\min_a$ حد پایین مقدار ویژگی است. در گام دوم، مجموعه داده‌ها در دو دسته یادگیری و آزمون طبقه بندی می‌گردند. در گام سوم الگوریتم ژنتیک اقدام به تولید نسل اولیه از کروموزوم‌ها می‌نماید. هر کروموزوم شامل ژن‌هایی که معرف ویژگی‌های انتخاب شده، پارامترهای C و γ می‌باشد. در گام چهارم بر اساس مقادیر تعیین شده در هر کروموزوم، دقت عملکرد مدل LSSVM آزمون می‌گردد. در گام پنجم شرایط توقف مدل بررسی می‌گردد. در صورتی که این شرایط محقق نشده باشد، کروموزوم‌ها بر اساس دقت حاصل شده در گام قبل مرتب شده، سپس الگوریتم ژنتیک با استفاده از عملگرهای خود مقادیر کروموزوم‌ها را به سمت بهترین جواب سوق می‌دهد. گام‌های سوم تا پنجم تا جایی که شرایط توقف به وقوع بپیوندد ادامه می‌یابند. پس از آن که کروموزومی با مقادیر بهینه تعداد ویژگی، C و γ انتخاب شد، آن‌گاه مدل LSSVM بر اساس مقادیر مذکور شکل می‌گیرد. در انتها برای ارزیابی ریسک اعتباری متقاضیان جدید و تعیین سیاست اعطای اعتبار آن‌ها، کافی است اطلاعات مربوط به ویژگی‌های آن‌ها را به مدل بهینه شده وارد نماییم. شکل ۳ مراحل کلی این الگوریتم را نشان می‌دهد.



شکل ۳. مراحل کلی مدل ارائه شده

برای تشریح بیشتر روش ارائه شده، در ادامه برخی از مراحل را به تفصیل شرح داده شده‌اند.

تابع مطلوبیت

مقدار مطلوبیت مدل ارائه شده از دید عملکرد تعمیم‌پذیری به وسیله نسبت برخورد[۴۵] و از دید پیچیدگی مدل به وسیله تعداد ویژگی‌های انتخاب شده ارزیابی می‌گردد. پس از اینکه Ga-LSSVM مجموعه متقاضیان را دسته بندی نمود، آنگاه دقت عملکرد آن با مجموعه آزمون ارزیابی می‌گردد. تابع هدف در این قسمت حداکثر نمودن نسبت برخورد و حداقل سازی تعداد ویژگی‌های انتخاب شده می‌باشد.

$$f = E_{accuracy} - \alpha E_{complexity} \qquad (۱۹)$$

که در آن $E_{accuracy}$ دقت دسته بندی یا نسبت برخورد و $E_{complexity}$ پیچیدگی مدل است. پارامتر α با هدف تعدیل تعداد متغیرهای استفاده شده توسط LSSVM در نظر گرفته شده است. در معادله (۱۹) هدف قسمت اول انتخاب زیرمجموعه متغیرها با بیشترین قدرت تمیز دهی است در حالی که قسمت دیگر با هدف یافتن راه حلی با صرفه توسط حداقل کردن تعداد متغیرهای انتخاب شده به شکل زیر، در نظر گرفته شده است (Huang, Chen, & Wang, 2007).

$$\mathrm{E_{accuracy}} = \frac{\text{تعداد دسته بندی صحیح}}{\text{تعداد کل نمونه ها}} \qquad (۲۰)$$

$$E_{complexity} = \frac{n_v}{N} \qquad (۲۱)$$

که در آن n_v تعداد متغیرهای استفاده شده در مدل LSSVM و N تعداد کل متغیرها می‌باشد.

معیارهای ارزیابی

به منظور آزمون عملکرد روش‌های ارائه شده، از سه معیار بهره گرفته شده است. معیار یک، نسبتی از تعداد مشتریان بد حساب است که به درستی طبقه بندی شده‌اند. این معیار، که حساسیت[۴۶] نامیده می‌شود، بسیار حائز اهمیت می‌باشد چرا که هزینه‌ای که از عدم تشخیص صحیح مشتریان بدحساب مؤسسات اعتباری را تهدید می‌نماید، بسیار بیشتر از هزینه‌ی عدم

تشخیص صحیح مشتری خوش‌حساب می‌باشد. این هزینه شامل از دست دادن اصل و فرع تسهیلات به همراه هزینه پیگیری مطالبات معوق می‌باشد. (Yu, Wang, & Cao, 2009)

$$\text{معیار دقت یک} = \frac{\text{مشتریان بدحسابی که صحیح طبقه بندی شده اند}}{\text{تعداد کل مشتریان بد حساب}} \qquad (۲۲)$$

با توجه به مفهوم معیار فوق، اختلاف آن از عدد یک را می‌توان به عنوان ریسک اعتباری روش مورد استفاده در نظر گرفت. منظور از ریسک اعتباری، احتمال اعطای اعتبار به مشتریان بدحساب می‌باشد که مدل توانایی تشخیص درست آن‌ها را نداشته است. معیار دوم، با عنوان تشخیص[۴۷]، بر دقت مدل در شناسایی صحیح مشتریان خوش‌حساب دلالت دارد. با استفاده از این معیار دقت روش مورد استفاده در پیش بینی صحیح مشتریان خوش حساب نمایان می‌گردد. پیش بینی نادرست مشتریان خوش حساب، کاهش حاشیه سود بانکی را در پی خواهد داشت.

$$\text{معیار دقت دو} = \frac{\text{مشتریان خوش‌حسابی که صحیح طبقه بندی شده اند}}{\text{تعداد کل مشتریان خوش حساب}} \qquad (۲۳)$$

اختلاف معیار فوق از عدد یک را می‌توان به عنوان ریسک تجاری در نظر گرفت که به معنای احتمال عدم اعطای اعتبار به مشتریان خوش حسابی است که مدل نتوانسته آن‌ها را به درستی تشخیص دهد. معیار سوم (دقت کلی)، نشان دهنده توانایی مدل در دسته بندی کلیه مشتریان، چه خوب و چه بد، می‌باشد.

$$\text{معیارسوم (دقت کل)} = \frac{\text{مشاهده هایی که صحیح طبقه بندی شده اند}}{\text{تعداد کل مشاهدات}} \qquad (۲۴)$$

از آن‌جا که هر یک از معیارهای فوق از درجه اهمیتی مختص به خود برخوردار می‌باشند، لذا در سنجش عملکرد مدل لازم است هر سه در کنار یکدیگر مورد بررسی قرار گیرند. توجه بیش از حد به یک یا دو معیار، شیوه‌ی درستی برای بررسی عملکرد یک مدل نخواهد بود. از این رو در این پژوهش به بررسی هر سه معیار پرداخته شده است.

۵- نتایج پژوهش

در این قسمت به داده‌ها و نتایج تجربی حاصل از پیاده سازی روش‌های ذکر شده پرداخته می-شود. همان‌گونه که قبلاً نیز عنوان شد، مجموعه داده پژوهش به چهار نمونه فرعی تقسیم بندی شده‌اند. نتایج بکار گیری رویکرد LSSVM، Ga-LSSVM و لاجیت بر اساس هر یک از نمونه های فرعی و به تفکیک معیارهای پژوهش در جدول ۱ ارائه شده است.

جدول ۱.مقایسه نتایج روش‌ها

نمونه فرعی	روش	معیار اول	معیار دوم	معیار سوم
یک	LSSVM	٪۵۱.۹۵ (۲)	٪۸۶.۱۳ (۳)	٪۷۵.۶۰ (۲)
	Ga-LSSVM	٪۶۱.۵۴ (۱)	٪۹۴.۲۲ (۱)	٪۸۴.۰۰ (۱)
	LOGIT	٪۳۸.۹۶ (۳)	٪۸۹.۰۲ (۲)	٪۷۳.۶۰ (۳)
دوم	LSSVM	٪۵۱.۴۳ (۲)	٪۸۴.۴۴ (۳)	٪۷۵.۲۰ (۲)
	Ga-LSSVM	٪۶۰.۰۰ (۱)	٪۹۱.۶۷ (۱)	٪۸۲.۸۰ (۱)
	LOGIT	٪۳۴.۲۹ (۳)	٪۸۸.۸۹ (۲)	٪۷۳.۶۰ (۳)
سوم	LSSVM	٪۵۱.۹۵ (۲)	٪۸۶.۱۳ (۳)	٪۷۵.۶۰ (۲)
	Ga-LSSVM	٪۶۱.۰۴ (۱)	٪۹۴.۲۲ (۱)	٪۸۴.۰۰ (۱)
	LOGIT	٪۳۸.۹۶ (۳)	٪۸۹.۰۲ (۲)	٪۷۳.۶۰ (۳)
چهارم	LSSVM	٪۵۱.۴۳ (۲)	٪۸۴.۴۴ (۳)	٪۷۵.۲۰ (۲)
	Ga-LSSVM	٪۶۰.۰۰ (۱)	٪۹۱.۶۷ (۱)	٪۸۲.۸۰ (۱)
	LOGIT	٪۳۴.۲۹ (۳)	٪۸۸.۸۹ (۲)	٪۷۳.۶۰ (۳)

در هر ستون عدد اول نشان دهنده دقت روش مورد نظر بر اساس معیارهای پژوهش می‌باشد و عدد دوم که در پرانتز قرار داده شده است نشانگر رتبه روش مزبور در میان دیگر روش‌ها می‌باشد.

به عنوان مثال بر اساس نتایج بدست آمده از نمونه فرعی اول، عملکرد روش Ga-LSSVM در طبقه بندی صحیح مشتریان بد حساب ۶۱/۵۴ درصد بوده است. از این رو با توجه مفهوم این معیار، ریسک اعتباری این رویکرد ۳۸/۴۶ درصد خواهد بود. این در حالی است که روش لاجیت با دقت ۳۸/۹۶ درصد و ریسک اعتباری ۶۱/۰۴ درصد، ضعیف‌ترین عملکرد را به خود اختصاص داده است.

از سویی دیگر رویکرد Ga-LSSVM ۹۴/۲۲ درصد از متقاضیان خوش حساب را به درستی شناسایی کرده است که به نوعی نشان دهنده ریسک تجاری ۵/۷۸ درصد خواهد بود. همچنین

عملکرد روش‌های لاجیت و LSSVM به ترتیب با اختلاف ۵/۲۰ و ۸/۰۹ از روش Ga-LSSVM در جایگاه دوم و سوم جای دارند.

در انتها برای نمونه فرعی اول، دقت عملکرد رویکرد Ga-LSSVM بر اساس معیار سوم پژوهش، که طبقه بندی کلی متقاضیان می‌باشد، ۸۴ درصد می‌باشد. این در حالی است که در این نمونه، روش‌های LSSVM و لاجیت به ترتیب از دقت ۷۵/۶۰ و ۷۳/۶۰ درصد برخوردار می‌باشند.

با توجه به نتایج بدست آمده برای چهار نمونه فرعی، عملکرد روش Ga-LSSVM به تفکیک سه معیار برای هر چهار نمونه‌ی فرعی در رتبه اول قرار دارد. از سوی دیگر عملکرد روش LSSVM بر اساس معیار های اول و سوم برای کلیه نمونه‌های فرعی نسبت به روش لاجیت مطلوب‌تر بوده و در جایگاه دوم قرار دارد. این در حالی است که روش لاجیت تنها در معیار دوم نسبت به روش LSSVM برتری دارد.

در ادامه به منظور بررسی آماری برتری عملکرد مدل Ga-LSSVM نسبت به دو روش دیگر از آزمون مقایسه زوجی استفاده شده است. نتایج حاصل از این آزمون در جدول ۲ نشان داده شده است.

جدول ۲ نتایج آزمون مقایسه زوجی

معیار	عنوان	Ga-Lssvm	Logit	Ga-Lssvm	Lssvm
معیار اول	میانگین	۰/۵۹۹	۰/۳۶۲	۰/۵۹۹	۰/۵۱۶
	آماره T	۳۰/۰۹۶۱۷		۲۵/۴۱۳۳۸۱	
	p-value	۰/۰۰۰۰۸		۰/۰۰۰۱۳۳۶	
معیار دوم	میانگین	۰/۹۲۷	۰/۸۹۴	۰/۹۲۷	۰/۸۴۷
	آماره T	۴/۴۰۷۰۵۷		۱۸/۸۴۰۵۲	
	p-value	۰/۰۲۱۶۷		۰/۰۰۰۲۸	
معیار سوم	میانگین	۰/۸۲۹	۰/۷۳۵	۰/۸۲۹	۰/۷۴۷
	آماره T	۱۴/۱۷۱۰۳		۲۱/۴۱۱۵۵	
	p-value	۰/۰۰۰۷۶۱		۰/۰۰۰۲۲۹	

در روش آزمون مقایسه زوجی، از تفاضل‌های بین زوج مشاهدات استفاده شده و درباره تفاضل میانگین‌ها در جامعه استنباط می‌شود. در این آزمون فرض صفر این است که تفاضل مشاهدات بین دو روش بررسی شده برابر صفر است و فرض مقابل بیان می‌کند که این تفاضل مخالف صفر می‌باشد.

در جدول فوق فرضیه آزمون مقایسه زوجی بین عملکرد روش‌های LSSVM و لاجیت با رویکرد Ga-LSSVM برای هر یک از معیارهای مورد بررسی قرار گرفته است.

همان گونه که ملاحظه می‌کنید برای دو روش لاجیت و Ga-LSSVM سطح معناداری بدست آمده کوچک‌تر از ۰/۰۵ می‌باشد، بنابراین در سطح اطمینان ۹۵ درصد فرضیه صفر مبنی بر تساوی میانگین لاجیت با روش Ga-LSSVM بر اساس معیارهای اول تا سوم رد می‌شود. این بدان معناست که روش حداقل مربعات ماشین بردار پشتیبان بر اساس هر سه معیار عملکرد بهتری نسبت به روش لاجیت داشته است.

نظر به این که برای دو روش Ga-LSSVM و LSSVM نیز سطح معناداری بدست آمده کوچک‌تر از ۰/۰۵ می‌باشد، بنابراین این امر نشان دهنده رد فرضیه صفر و عملکرد بهتر روش Ga-LSSVM نسبت به LSSVM می‌باشد.

در ادامه نیز جهت قیاس عملکرد رویکردهای ارائه شده در پژوهش‌های انجام شده با پژوهش حاضر، نتایج حاصل از هر یک را بر اساس معیارهای اول تا سوم در جدول ذیل نشان داده شده است. روش‌های فرا ابتکاری بکار گرفته شده در این مطالعات شامل طراحی آزمایش‌ها[۴۸] (DOE)، جستجوی شبکه‌ای[۴۹] (GS) و جستجوی مستقیم[۵۰] (DS) می‌باشد. لازم به ذکر است که دقت ارائه شده برای رویکردهای این پژوهش از میانگین دقت نمونه های اول تا چهارم بدست آمده است.

جدول ۳ مقایسه نتایج پژوهش‌های مشابه

روش	منبع اطلاعات	تعداد ویژگی	معیار اول	رتبه	معیار دوم	رتبه	معیار سوم	رتبه
Ga-LSSVM	محاسبات پژوهش	۱۴	٪۶۰/۵۲	۱	٪۹۲/۹۵	۱	٪۸۳/۴۰	۱
LSSVM	محاسبات پژوهش	۲۴	٪۵۱/۶۹	۵	٪۸۵/۲۹	۸	٪۷۵/۴۰	۶
LOGIT	محاسبات پژوهش	۱۱	٪۳۶/۶۳	۷	٪۸۸/۹۶	۶	٪۷۳/۶۰	۸
SVM	هوانگ و همکاران (۲۰۰۷)	۲۴	٪۳۳/۳۳	۸	٪۹۰/۶۰	۴	٪۷۳/۹۰	۷
GA-SVM	هوانگ و همکاران (۲۰۰۷)	۲۴	٪۶۰/۴۴	۲	٪۸۸/۳۰	۷	٪۷۷/۹۲	۳
GS-LSSVM	ژو و همکاران (۲۰۰۹)	۲۴	٪۴۸/۶۳	۶	٪۹۲/۳۳	۲	٪۷۷/۶۱	۴
DS-LSSVM	یو و همکاران (۲۰۱۱)	۲۴	٪۵۲/۴۵	۴	٪۹۰/۶۷	۳	٪۷۶/۹۲	۵
DOE-LSSVM	یو و همکاران (۲۰۱۱)	۲۴	٪۵۵/۵۵	۳	٪۸۹/۲۳	۵	٪۷۸/۴۶	۲

همان گونه که در جدول ۳ ملاحظه می‌کنید، بر اساس معیار اول یا درجه حساسیت، روش Ga-LSSVM با دقت ۶۰/۵۲ درصد در میان دیگر روش‌ها بهترین عملکرد را داشته است. بعد از این روش، رویکرد Ga-SVM در رتبه‌ی دوم و رویکرد DOE-LSSVM در رتبه سوم قرار دارند.

بر اساس معیار دوم که درجه تشخیص می‌باشد، روش Ga-LSSVM با دقت ۹۲/۹۵ درصد، بالاترین دقت را به خود اختصاص داده است. این در حالی است که رویکرد GS-LSSVM با اختلاف ۰/۶۲ درصد از روش ارائه در این پژوهش، در رتبه‌ی دوم قرار گرفته و عملکرد بهتری نسبت به شش روش دیگر ارائه نموده است.

بر اساس معیار سوم که بیانگر رویکرد کلی سنجش دقت روش‌های مربوطه می‌باشد، بهترین عملکرد از آن روش ارائه شده Ga-LSSVM مبتنی بر انتخاب ویژگی است.

۶- نتیجه گیری و بحث

ریسک اعتباری، به معنای «احتمال قصور در بازپرداخت تسهیلات اعطایی از سوی مشتریان»، از مهم‌ترینِ ریسک‌هایی است که هر سازمان مالی با آن مواجه می‌باشد، به گونه‌ای که کمیته بال تاکید فراوانی بر مدیریت آن نموده و همچنین نزد بانک‌های تجاری از اهمیت قابل توجهی برخوردار می‌باشد.

در این پژوهش تلاش گردیده تا با استفاده از روش حداقل مربعات ماشین بردار پشتیبان به عنوان یکی از رویکردهای برجسته داده‌کاوی و ترکیب آن با روش‌های الگوریتم ژنتیک و انتخاب ویژگی، بتوان مدلی کارا جهت ارزیابی وضعیت اعتباری مشتریان ارائه نمود. با توجه به پیاده سازی مدل ارائه شده، می‌توان نتایج زیر را استخراج نمود:

بر اساس مزیت نسبی روش Ga-LSSVM نسبت به دیگر روش‌های مبتنی بر بهینه سازی پارامترهای ماشین بردار، می‌توان دریافت که استفاده از رویکرد فرا ابتکاری الگوریتم ژنتیک دقت عملکرد رویکرد حداقل مربعات ماشین بردار پشتیبان را بهبود می‌بخشد. که بر اساس آن می‌توان نتیجه گرفت که تعیین بهینه پارامترهای LSSVM، تأثیر بسزایی در بهبود عملکرد آن‌ها خواهد داشت.

علی‌رغم حذف ویژگی‌های زائد و یا اضافی از مجموعه داده های مشتریان توسط رویکرد انتخاب ویژگی در روش Ga-LSSVM، همچنان دقت مدل در سطح قابل قبولی حفظ شده است. این رویداد بیانگر عملکرد مطلوب رویکرد انتخاب ویژگی بر مبنای الگوریتم ژنتیک در ارزیابی متقاضیان می‌باشد.

همچنین با توجه به مقایسه نتایج پژوهش حاضر با تحقیقات مشابه، عملکرد رویکرد ارائه شده Ga-LSSVM هم نسبت به رویکردهایی که از روش SVM استفاده کردند و هم نسبت به روش‌هایی که از الگوریتم فرا ابتکاری جهت بهینه کردن پارامترها استفاده نمودند، برتر می‌باشد. کاهش هزینه های محاسباتی از طریق خطی کردن روابط و به‌کارگیری الگوریتم ژنتیک با توجه به قدرت و توانایی‌های آن، از علل این برتری می‌باشند.

همان‌گونه که نشان داده شد، عملکرد رویکرد لاجیت تنها در معیار دوم نسبت به روش LSSVM بهتر بوده و در دیگر معیارها در پایین‌ترین سطح قرار گرفته است. این امر موید آن است که در صورتی که به دنبال کاهش ریسک اعتباری و افزایش دقت طبقه بندی متقاضیان تسهیلات باشیم، می‌توان رویکرد حداقل مربعات ماشین بردار پشتیبان را به عنوان جایگزینی مناسب برای روش لاجیت در نظر گرفت.

بر اساس نتایج بدست آمده و مقایسه های صورت گرفته، رویکرد Ga-LSSVM در هر سه معیار اول تا سوم عملکرد بهتری از خود نشان داده است، بنابراین در صورت بنا نهادن سیاست اعتبار دهی بر اساس این رویکرد، نه تنها این امر منجر به شناسایی بهتر متقاضیان خوش حساب شده و سودآوری را افزایش می‌دهد بلکه ضریب خطا در شناسایی متقاضیان بد حساب کمتر شده و در نتیجه از بروز زیان‌های ناشی از نکول این متقاضیان جلوگیری می‌شود. لذا به‌کارگیری این رویکرد از دو جنبه مذکور بانک یا مؤسسات اعتباری را منتفع می‌سازد.

با توجه به تائید کارایی روش‌های فرا ابتکاری جهت حل مسائل پیچیده، توصیه می‌شود که در عمل استفاده از این روش‌ها در سنجش ریسک اعتباری مشتریان مد نظر قرار گیرد، اما به دلیل این که ساختار داده‌های مشتریان تأثیر بسزایی در عملکرد مدل خواهد داشت و همچنین در روش-های فرا ابتکاری امکان قرار گرفتن جواب در بهینه محلی وجود دارد، بهتر است در مراحل اولیه طراحی، از این مدل‌ها در کنار مدل‌های مرسوم و سنتی استفاده گردد.

فهرست منابع

* راعی، ر.، و فلاح پور، س. (۱۳۸۷). کاربرد ماشین بردار پشتیبان در پیش بینی درماندگی مالی شرکت‌ها با استفاده از نسبت‌های مالی. *بررسی‌های حسابداری و حسابرسی، دوره ۱۵، شماره ۵۳*، ۱۷-۳۴.
* رهنمای رودپشتی، ف.، علی خانی، ر.، و مران‌جوری، م. (۱۳۸۸). *بررسی کاربرد مدل‌های پیش بینی ورشکستگی آلتمن و فالمر در شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران.* بررسی‌های حسابداری و حسابرسی، ۳۴-۱۹
* صفری، س.، ابراهیمی شقاقی، م.، و شیخ، م. (۱۳۸۹). مدیریت ریسک اعتباری مشتریان حقوقی در بانک‌های تجاری با رویکرد تحلیل پوششی داده‌ها. *پژوهش‌های مدیریت در ایران*، ۱۳۷-۱۶۴.
* طالبی، م.، و شیرزادی، ن. (۱۳۹۰). *اندازه گیری ریسک اعتباری.* سمت.
* عرب مازار، ع.، و روئین تن، پ. (۱۳۸۵). *عوامل موثر بر ریسک اعتباری مشتریان بانکی؛ مطالعه موردی بانک کشاورزی.* دو فصلنامه علمی - پژوهشی جستارهای اقتصادی، سال سوم، شماره ششم.
* فلاح شمس، م. (۱۳۸۷). *مدل‌های اندازه گیری ریسک اعتباری در بانک‌ها و موسسه های اعتباری.* نشریه تازه های اقتصاد، ۲۲ - ۲۸.
* فلاح شمس، م.، و رشنو، م. (۱۳۸۷). *مدیریت ریسک اعتباری در بانک‌ها و مؤسسات مالی و اعتباری.* تهران: دانشکده علوم اقتصادی.
* فلاح شمس، م.، و مهدوی راد، ح. (۱۳۸۹). *طراحی مدل اعتبارسنجی و پیش بینی ریسک اعتباری مشتریان تسهیلات لیزینگ. مهندسی مالی و مدیریت پرتفوی، ۱-۲۲.*
* *مصیبی فر، ا.، فرزین وش، ا.، و کمیجانی، ا. (۱۳۸۹). نقش نظام بانکی در توسعه اقتصادی (با توجه خاص به نظام بانکی ایران). دانشکده اقتصاد، دانشگاه تهران*
* Abe, S. (2005). Support Vector Machines for Pattern Classification. Springer.
* Altman, E. (1968). Financial ratios discriminate analysis and the prediction of corporate Bankruptcy. The Journal of finance.
* Avci. Engin. (2009). Selecting of the optimal feature subset and kernel parameters in digital modulation classification by using hybrid genetic algorithm- support vector machines. Expert systems with applications, 1391-1402.
* Beaver, W. (1967). Financial ratios as Predicators of Failure. Journal of Acounting Resarch.

* Bellotti, T., & Crook, J. (2008). Support vector machines for credit scoring and discovery of significant features. Expert systems with applications, 102-109.
* Burges, C. (1998). tutorial on support vector machines for patternrecognition. Data Mining and Knowledge Discovery, 121-167.
* Chen, W., Ma, C., & Ma, L. (2008). Mining the customer credit using hybrid support vector machine technique. Expert systems with applications, 1-6.
* Cortes, C. V. (1995). Support-vector networks. Machine Learning. 273-297.
* Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK.
* Desai, V. S., Crook, J. N., & Overstreet, J. (1996). A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research, 24-37.
* Fisher, R. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 179-188.
* Glantz, M. (2003). Managing Bank Risk. Academic Press.
* Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring. Journal of the Royal Statistical Society, 523-541.
* Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 847-856.
* Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems, 543-558.
* Lean, Y., Xiao, Y., Shouyang, W. a., & K.K., L. (2011). Credit risk evaluation using least squares SVM classifier with design of experiment for parameter selection. Expert Systems with Applications, 15392-15399.
* Liu, H., & Motoda, H. (1998). Feature selection for knowledge discovery and data mining. Norwell, MA: Kluwer Academic.
* Michael, D. (1999). The simple Genetic Algorithm: Foundation and Theory. The MIT Press.
* Shin S, K., Lee S, T., & Kim J, H. (2005). An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, 127-135.
* Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural Processing Letters, 293–300.
* Van Gestel, T., Baesens, B., Suykens, j., Espinoza, M., Baestaens, D., Vanthienen, J., et al. (2003). Bankruptcy prediction with least squares support vector machine

classifiers. Proceedings on IEEE International Conference on Computational Intelligence for Financial Engineering.

* Yu, L., Wang, S., & Cao, J. (2009). A modified least squares support vector machine classifier with application to credit risk analysis. International Journal of Information Technology and Decision Making, 697-710.
* Zhou, L., Lai, K. K., & Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. Expert System with Applications, 127-133.

یادداشت‌ها

[1] Basel Committee on Banking Supervision (BCBS)
[2] Support Vector Machine (SVM)
[3] Moody's
[4] Fisher
[5] Durand
[6] Beavor
[7] Logit Regression
[8] Altman
[9] Z-Score
[10] Recursive Partitioning Algorithm (RPA)
[11] Data Envelopment Analysis (DEA)
[12] Yeh
[13] Hung and et.al
[14] Genetic Programing
[15] Backpropagation neural network
[16] Parametric
[17] Non-Parametric
[18] Linear Probability Model (LPM)
[19] Logit Model
[20] Probit Model
[21] Discriminant Analysis
[22] Mathematical Programming
[23] Decision Trees Model
[24] Nearest Neighbors Model
[25] Analysis Hierarchy Process
[26] Artificial Intelligence
[27] Hyper plane
[28] Support Vector (SV)
[29] Euclidean norm
[30] Karush Kuhn–Tucker (KKT)
[31] Soft-Margin
[32] Kernel methods
[33] Least Square Support Vector Machine (LSSVM)
[34] John Holland

[35] Chromosome
[36] Initial Population
[37] Generation
[38] Fitness Function
[39] Reproduction
[40] Crossover
[41] Mutation
[42] Offspring
[43] Selection
[44] Black Box
[45] Hit Ratio
[46] Sensitivity
[47] Specificity
[48] Design of Experience
[49] Grid Search
[50] Direct Search