مجله
پژوهش های مدیریت
شماره ۸۴ - بهار ۱۳۸۹

طبقه بندی متقاضیان تسهیلات اعتباری بانک‌ها با استفاده از تکنیک ماشین بردار پشتیبان

دکتر عباس طلوعی اشلقی *
دکتر هاشم نیکومرام **
فرناز مقدوری شربیانی ***

چکیده

در صنعت بانکداری یکی از موضوعاتی که همواره بایستی مدنظر سیاستگذاران اعتباری قرار داشته باشد، مبحث مدیریت ریسک است. در بین ریسک‌های مختلفی که بانک‌ها با آن مواجهند, ریسک اعتباری از با اهمیت ترین آن‌ها است که از زیان‌های ناشی از ناتوانی یا عدم تمایل مشتری به ایفای تعهدات خویش در برابر بانک حاصل می‌گردد.

جهت مدیریت و کنترل ریسک مذکور, سیستم‌های طبقه بندی اعتباری مشتریان ضرورتی انکار ناپذیر است. چنین سیستمی، براساس سوابق و اطلاعات موجود، طبقه مشتریان را تعیین می‌کند. بدیهی است بهره گیری از چنین سیستمی بانک را در گزینش مطلوب مشتریان خود یاری نموده و ضمن کنترل و کاهش ریسک اعتباری، سطح بهره‌وری فرایند اعطای تسهیلات بانکی را ارتقا می‌دهد.

در مقاله حاضر مدل طبقه بندی مبتنی بر ماشین بردار پشتیبان با رویکرد هوش مصنوعی، به منظور پیش‌بینی عملکرد مالی مشتریان حقوقی بانک‌ها ارائه گردیده است. در واقع، در این نوشتار ماشین بردار پشتیبان به همراه دیگر مکانیزم‌ها از جمله تکنیک‌های F-score و Grid search جهت طبقه‌بندی متقاضیان تسهیلات اعتباری بانکی وافزایش کارایی مدل استفاده شده است. نتایج، حاکی از افزایش صحت طبقه بندی است و نشان می‌دهد که ماشین بردار پشتیبان در مقایسه با دیگر مدل‌های طبقه بندی دارای صحت بیشتری است.

واژگان کلیدی :

داده کاوی، طبقه بندی، ماشین بردار پشتیبان، انتخاب ویژگی

* دانشیار، عضو هیات علمی تمام وقت دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران (toloie@srbiau.ac.ir)
تهران- بزرگراه اشرفی اصفهانی- به سمت حصارک- دانشکده مدیریت و اقتصاد دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران
** دانشیار، عضو هیات علمی تمام وقت دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران
تهران- بزرگراه اشرفی اصفهانی- به سمت حصارک- دانشکده مدیریت و اقتصاد دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران
*** دانش آموخته کارشناسی ارشد مدیریت فناوری اطلاعات، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران (farnazmaghdouri@gmail.com)
تهران- بزرگراه اشرفی اصفهانی- به سمت حصارک- دانشکده مدیریت و اقتصاد دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران

نویسنده مسئول یا طرف مکاتبه : دکتر عباس طلوعی اشلقی

تاریخ دریافت مقاله : ۸۷/۸/۲۷ تاریخ اصلاحات مقاله : ۸۷/۹/۱۸ تاریخ پذیرش مقاله : ۸۷/۱۰/۲۶

مقدمه

بانک‌ها به عنوان بخش اصلی نظام مالی ، نقش مهمی را در تامین مالی بخش‌های مختلف اقتصادی بر عهده دارند. در راستای ایفای این نقش ، بانک‌ها با ریسک‌های متفاوتی روبرو هستند که یکی از عمده‌ترین آنها ریسک اعتباری است (طلوعی، مقدوری،۱۳۸۸). ریسک اعتباری عبارت است از احتمال قصوراستفاده کنندگان از تسهیلات بانکی در انجام تعهدات مالی خود. بر این اساس لازم است بانک‌ها برای کاهش و تقلیل ریسک اعتباری، قبل از هر پرداختی به متقاضیان، وضعیت اعتباری آنان را بررسی نمایند (طلوعی، مقدوری،۱۳۸۸). ارزیابی درست درخواست کنندگان و مدیریت ریسک برای بانک‌ها و موسسات اعتباری و فیلتر کردن متقاضیان مباحثی حیاتی هستند . برای اعطای تسهیلات باید درجه اعتبار و توان گیرنده تسهیلات را در بازپرداخت تسهیلات اعطایی تعیین نمود.بحران های مشاهده شده در نظام بانکی کشورها عمدتا ناشی از عدم کارایی در مدیریت ریسک اعتباری بوده است (طلوعی، مقدوری،۱۳۸۸). حجم قابل ملاحظه‌ای از تسهیلات اعطایی سوخت شده یا معوقه بانک‌ها ، گویای فقدان مدل‌های مناسب اندازه گیری ریسک اعتباری و سیستم‌های مدیریت ریسک در شبکه بانکی است . یکی از مهمترین ابزارهایی که بانک ها برای مدیریت و کنترل ریسک اعتباری بدان نیازمند هستند سیستم طبقه بندی مشتریان است (سینکی ،۱۹۹۲). بررسی عوامل موثر براین طبقه بندی به ما کمک خواهد کرد تا بتوانیم خیل عظیم مشتریان بانک را در ابتدای کار در طبقات مختلف دسته بندی و با اختصاص درجه و رتبه، تصمیم مناسبی در جهت تخصیص هرچه بهتر و موثرتر وجوه جمع آوری شده، اتخاذ نماییم و بازدهی مناسب را برای بانک، حاصل واحتمال نکول را به نحو بایسته کاهش دهیم. متاسفانه علی رغم اهمیت این موضوع، در کشور ما در زمینه اعطای تسهیلات اعتباری به مشتریان، روند منسجم و منظمی به منظور تعیین ریسک اعتباری، امتیازدهی، طبقه بندی و همچنین تعیین سقف‌های اعتباری بر اساس شاخص‌های ریسک ملاحظه نمی‌شود و شاخص‌ها بر اساس تشخیص کارشناسی و کمیته اعتباری صورت می پذیرد. برخورداری از یک مدل ریسک کارآمد نه تنها تصمیم گیری در زمینه اعتبار را تسهیل می نماید؛ بلکه افزون بر کاهش هزینه مبادله موجب خواهد شد که سیستم بانکی از الگوی کارآمدی در تخصیص سرمایه به بخش‌های مختلف اقتصادی برخوردار شود.

یکی از مهمترین روش‌های مدیریت ریسک اعتباری استفاده از سیستم‌های طبقه بندی برای کنترل ریسک انواع وام‌ها است(سینکی ،۱۹۹۲). در مطالعات گذشته اغلب از روش‌های آماری مانند مدل‌های رگرسیونی لاجیت و پروبیت و روش تحلیل ممیزی برای این منظور استفاده می‌گردید، ولی در سال‌های اخیر با گسترش فناوری های نوین اطلاعاتی و کاربردهای وسیع آن در پردازش و طبقه بندی اطلاعات ، مدل‌های مبتنی بر هوش مصنوعی و روش‌های ابتکاری توسعه یافته و مطالعات بسیاری در کاربرد این روش‌ها در مدل‌های طبقه بندی مشاهده می‌شوند(اخباری،۱۳۸۷). بر همین اساس در این تحقیق ماشین بردار پشتیبان را به عنوان یکی از جدیدترین تکنیک‌های داده کاوی جهت طبقه بندی مشتریان بانک‌ها معرفی نموده و از این تکنیک برای طبقه بندی مشتریان استفاده خواهیم نمود که اخیرا در صنعت اعتباری به عنوان ابزاری صحیح واثربخش برای تحلیل اعتباری مطرح شده است و منجر به عملکرد بالا در بسیاری عملیات از جمله طبقه بندی و رتبه بندی اعتباری می شود.

لذا در این مقاله با استفاده از ماشین بردار پشتیبان، چارچوبی برای طبقه بندی متقاضیان تسهیلات اعتباری پیشنهاد و کارایی این مدل با مدل‌های دیگر مقایسه شده است.

مقاله حاضر شامل بخش‌های زیر می‌باشد: پیشینه تحقیق، بیان مساله، ماشین بردار پشتیبان، ساختار اجرایی تحقیق (متدلوژی تحقیق) رویه‌ی اعتبار سنجی مبتنی بر ماشین بردار پشتیبان، چارچوب اجرائی مبتنی بر ماشین بردار پشتیبان، نتایج طبقه‌بندی با مدل ماشین بردار پشتیبان و در نهایت نتیجه گیری

پیشینه تحقیق

در سالهای اخیر ماشین بردار پشتیبان در بسیاری از مسائل طبقه بندی از جمله طبقه بندی متن، بازشناسی تصاویر و... بطور موفقیت آمیزی بکار رفته است (بلوتی، کروک، ۲۰۰۸). تحقیقات صورت گرفته در زمینه‌ی استفاده از ماشین بردار پشتیبان برای اعتبارسنجی و طبقه بندی، جدید واندک هستند، با این حال اخیرا تعدادی مقاله در این زمینه منتشر شده و به ارزیابی عملکرد ماشین بردار پشتیبان برای اعتبارسنجی پرداخته است که در ادامه به توضیح این تحقیقات پرداخته می شود. (بیزین و همکارانش، ۲۰۰۳) در تحقیقی ازماشین بردارپشتیبان همراه با دیگر طبقه بندی کننده‌ها در چندین پایگاه داده استفاده کردند.آن‌ها اعلام کردند که ماشین بردار پشتیبان در مقایسه با دیگر الگوریتم‌ها عملکرد بهتری دارد ولی همواره بهترین عملکرد را ندارد.(اسچباش و استکینگ، ۲۰۰۵) از ماشین بردار پشتیبان در وام‌های اعتباری استفاده کردند و به این نتیجه رسیدند که ماشین بردار پشتیبان نسبتا بهتر از رگرسیون لجستیک عمل می‌کند ولی تفاوت زیادی با آن ندارد. در هر دوی این مقالات از ماشین بردار پشتیبان خطی و هسته‌ی RBF استفاده شده و در هر دو مورد، اندازه‌ی پایگاه داده‌ی اعتباری بکار رفته خیلی کوچکتر از اندازه‌ی واقعی آن است.(وان گستل و همکارانش، ۲۰۰۶) از حداقل مجذورات ماشین بردار پشتیبان با هسته‌ی بیزیین برای طبقه بندی ورشکستگی بانک‌ها استفاده کردند. این محققان تفاوت مهم و معنی داری بین ماشین بردار پشتیبان، رگرسیون لجستیک و LDA پیدا نکردند. (هوانگ، چن، هسو، ۲۰۰۴) به مقایسه‌ی ماشین بردارپشتیبان با شبکه‌های عصبی در پیش بینی رتبه‌های اعتباری سازمان‌ها پرداختند ولی به تفاوت ناچیزی در عملکرد دست یافتند.(لی[1]، ۲۰۰۷) به نتایج مشابهی در مورد وام‌های سازمانی دست یافت.(هوانگ و وانگ[2]، ۲۰۰۷)به این نتیجه رسیدند که طبقه بندی اعتباری ماشین بردارپشتیبان دقیق تر ازشبکه‌های عصبی، درخت تصمیم یا الگوریتم ژنتیک نیست واهمیت نسبی استفاده از ویژگی‌های انتخاب شده توسط الگوریتم ژنتیک و ماشین بردار پشتیبان در رابطه با شبکه‌های عصبی و برنامه ریزی ژنتیک را با هم مقایسه کردند. این محققان از پایگاه داده‌ی کوچکی استفاده کردند و ویژگی‌های انتخاب شده توسط ماشین بردار پشتیبان را مقایسه نکردند و با متدهای عملی بکار رفته ازجمله رگرسیون لجستیک مقایسه نکردند.

(بلوتی، کروک، ۲۰۰۸) به مقایسه‌ی عملکرد ماشین بردار پشتیبان با چندین الگوریتم معروف پرداختند و از پایگاه داده‌ی بزرگتری نسبت به تحقیقات قبلی استفاده کردند و به این نتیجه رسیدند که ماشین بردار پشتیبان موفق‌تر از تکنیک‌های قبلی رتبه بندی است و از ماشین بردار پشتیبان به عنوان یکی از متدهای انتخاب ویژگی استفاده کردند.

همچنین (تیان-شیوگ لی و همکارانش[3]، ۲۰۰۶) در تحقیق خود به مقایسه عملکرد امتیازدهی اعتباری دو تکنیک طبقه بندی داده کاوی یعنی MARS, CART پرداخته اند . برای سنجش کارایی این دو روش از مجموعه داده های کارت اعتباری یک بانک استفاده شده است. نتایج تحقیق نشان می دهد که عملکرد دو روش مذکور نسبت به رویکردهای تحلیل ممیزی سنتی، رگرسیون لجستیک، شبکه های عصبی و ماشین بردار پشتیبان جهت امتیازدهی اعتباری بهتر می باشد و در نهایت اینکه (چنگ- لونگ، هوانگ و همکارانش[4]،۲۰۰۷) در تحقیق خود به مدلسازی تکنیکی هیبریدی برای اعتبارسنجی پرداخته‌اند. به این صورت که ترکیب الگوریتم ژنتیک و ماشین بردار پشتیبان هم برای انتخاب خصیصه ها و هم برای بهینه سازی پارامترهای مدل اعتبارسنجی بکار رفته است. مقایساتی نیز بین تکنیک ماشین بردار پشتیبان و طبقه بندهای شبکه های عصبی، برنامه ریزی ژنتیکی و درخت تصمیم صورت گرفته که نتایج نسبتا" یکسانی حاصل شده است. ولی نتایج تجربی نشان می‌دهد که ماشین بردار پشتیبان، نسبت به تکنیک‌های موجود داده کاوی، نتایج امید بخش تری را ارائه می‌دهد.

1 . Lee

2 . Huang& Wang

3 . Tian-Shyug Lee & et al.

4 . Cheng- Lung, Huang& et al.

بیان مساله

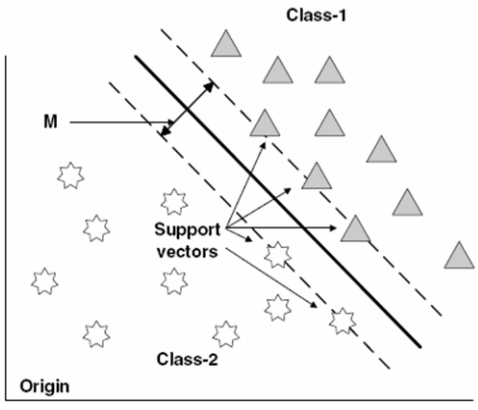
اعطای تسهیلات، بخش مهمی از عملیات هر بانک را تشکیل می‌دهد و این قسمت از فعالیت های بانکی از لحاظ اقتصادی اهمیت زیادی دارد. زیرا افزایش کمی سرمایه، باعث رشد و توسعه اقتصادی می‌شود. در اعطای تسهیلات، بانکها با خطر بزرگی که به آن ریسک اعتباری می گویند مواجه هستند. این ریسک علت مواجهه بانک ها با بحرانهای عمده مالی است. ریسک اعتباری را می توان احتمال عدم بازپرداخت وام از طرف متقاضی در نظر گرفت که بایستی مدیریت گردد (مین ،۲۰۰۶). برای مدیریت ریسک اعتباری از روشهای مختلفی می توان استفاده کرد که یکی از این روشها طراحی نظام تعیین درجه اعتباری برای دریافت کنندگان تسهیلات است(اخباری،۱۳۸۷). در واقع در این تحقیق به دنبال شناسایی الگوهای رفتاری مشتریان و در نتیجه ایجاد امکان پیش بینی رفتارهای مشتریان آتی هستیم. ارزیابی اعتبار مشتریان کار بسیار پیچیده ای است زیرا تعداد زیاد عوامل و پیچیدگی روابط مالی، اقتصادی و رفتاری، ارزیابی اعتباری را بسیار دشوار می سازد. از طرفی امر ارزیابی اغلب باید در محدوده زمانی کوتاهی صورت گیرد، زیرا طولانی شدن فرآیند ارزیابی باعث تاخیر در عملیات و در نهایت باعث افزایش هزینه‌ها خواهد شد. از طرف دیگر عدم دقت احتمالی در ارزیابی می تواند به تصمیمات اشتباه و نهایتا زیان های گزاف منجر گردد. همچنین محدودیت زمانی و ضرورت دقت در ارزیابی، پیچیدگی موضوع را دو چندان می کند(اخباری،۱۳۸۷).

همچنین یکی از مشکلات اساسی اعطای اعتبار، ضوابط اخذ وثیقه و یا آورده نقدی از طرف متقاضیان استفاده از اعتبارات و تسهیلات شبکه بانکی است.تجزیه و تحلیل اطلاعات نشان می دهد که ۷۰٪ جامعه مورد مطالعه، ضوابط اخذ وثیقه و انعطاف ناپذیر بودن معیارهای ارزیابی جهت جلوگیری از سوخت شدن اصل و سود تسهیلات را یکی از مشکلات دسترسی به تسهیلات و اعتبارات اعطایی سیستم بانکی، می دانند. همچنین حدود۹۴٪ از پاسخ دهندگان، طولانی بودن زمان ارزیابی‌ها را مشکل آفرین بیان نموده‌اند(نادعلی، ۱۳۸۷).با توجه به شرایط بازار اعتباری و با در نظر گرفتن انواع سیستم های رتبه بندی اعتباری، مساله تحقیق، پیش بینی الگوی رفتاری- بازپرداخت- مشتریان تسهیلات اعتباری در بانکهای ایران است. لذا در این مقاله به دنبال شناسایی الگوی مناسبی جهت طبقه بندی مشتریان اعتباری یک بانک با استفاده از الگوریتم ماشین بردار پشتیبان به عنوان یکی از ابزارهای داده کاوی هستیم.

ماشین بردار پشتیبان

ماشین بردار پشتیبان یک طبقه بندی کننده‌ی دوتایی است که با استفاده از نگاشت داده‌ها از فضای ورودی اصلی به فضایی با بعد بالاتر برای جداسازی آن‌ها عمل می‌کند(طلوعی، مقدوری،۱۳۸۸). این مدل ابرصفحه‌ای را جستجو می‌کند که فاصله‌اش با داده‌های دو کلاس ماکزیمم است. در این روش سعی بر آنست تا جهت بدست آوردن مرز کلاس‌ها، سیستمی با ظرفیت کمینه و یا به بیان بهتر سیستمی با حداقل پیچیدگی پیاده‌سازی شود. در نتیجه ماشین بردار پشتیبان می‌تواند با استفاده از داده‌های آموزشی کمتر نسبت به روش‌های رقیب، مرزهای سیستم را با دقت مناسبی تخمین بزند، بدون آنکه تعمیم پذیری سیستم را مخدوش نماید(منیری،۱۳۸۵). در واقع هدف اصلی طبقه بندی کننده‌ی ماشین بردار پشتیبان دستیابی به یک تابع $f(x)$ است که این تابع تعیین کننده‌ی ابرصفحه است. این ابرصفحه دو کلاس را به طور بهینه از هم جدا می‌کند که در شکل۱نشان داده شده است . در این شکل ،Mحاشیه است یعنی فاصله‌ی ابرصفحه از نزدیکترین نقاط هردوکلاس (چیا مینگ،۲۰۰۹).

شکل ۱: ماشین بردار پشتیبان در حالت خطی (چیا مینگ،۲۰۰۹)



وپ نیک برای محاسبه‌ی مرز جداکننده‌ی ۲ کلاس کاملا جدا از هم، روش حاشیه‌ی بهینه را پیشنهاد کرد. در واقع باید یک مرز خطی را بین ۲ کلاس به گونه‌ای محاسبه کنیم که:

تمام داده‌هایی که به کلاس ۱+ تعلق دارند در یک طرف مرز قرار گیرند و تمام داده‌هایی که متعلق به کلاس ۱- هستند در طرف دیگر مرز واقع شوند.

مرز تصمیم گیری بایستی به گونه‌ای باشد که فاصله‌ی نزدیک ترین نمونه‌های آموزشی هر دو کلاس از یکدیگر در راستای عمود بر مرز تصمیم گیری تا جایی که ممکن است حداکثر شود (منیری،۱۳۸۵).

در ماشین بردار پشتیبان به دو طریق می توان مجموعه نقاط را از هم جداکرد: ۱- خطی ۲- غیرخطی (اوچی ،۲۰۰۹).

در حالت خطی، ماشین بردار پشتیبان با در نظر گرفتن مجموعه‌ی آموزشی$(x_k\ y_k)$،$k=1,2,\ldots,m$ و $x_k \in R^n$ و $y_k \in \{-1$ و $+1\}$ ، ابرصفحه بهینه با حداکثر حاشیه را ازطریق حل مساله بهینه سازی زیر پیدا می‌کند:

$$y_k((w.x_k)+b)\geq 1 \quad k=1,2,\ldots m; \quad \text{Min } 1/2\|w\|^2 \qquad (1)$$

با توجه به انتخاب مقادیر w, b ابرصفحه‌های جداکننده‌ی زیادی وجود دارد. در روش ماشین بردار پشتیبان تلاش می شود تا بهترین صفحه‌ی جداکننده‌ی دو کلاس بدست آید. ثابت می‌شود که صفحه‌ی بهینه دارای بیشترین فاصله‌ی بین مرزها است. رابطه‌ی $w.x + b =1$ بیانگر مرز ناحیه‌ی $y= 1$ و $w.x + b =-1$ بیانگر مرز ناحیه‌ی $y= -1$ است.

$w.x + b =0$ نشان دهنده‌ی صفحه‌ی جدا کننده‌ی مرزهای دو ناحیه است (طلوعی، مقدوری،۱۳۸۸) و(منیری،۱۳۸۵).

$y=1 \quad w.x+b\geq 1$
$y=-1 \quad w.x+b\leq 1$

درنتیجه :

$$y\ \text{sign}\ (w.x+b) \geq 1 \qquad (2)$$

حد فاصل بین دو ناحیه توسط تفاضل مرزهای ۲ناحیه‌ی زیر بدست می‌آید:

$w.x + b=1$
$w.z + b=-1$

در نتیجه داریم:

$$w.(x-z) = 2,\ w/\|w\|.(x-z)=2/\|w\| \qquad (3)$$

حال، جهت بدست آوردن مرز بهینه باید فاصله‌ی بین دو ناحیه بیشینه شود؛ یعنی مقدار $\|w\|$ کمینه شود، لذا داریم:

$$\underset{w,\,b}{\text{Min}}\ 1/2\ \|w\|^2;\ y_k((w.x_k)+b)\geq 1 \quad k=1,2,\ldots m \qquad (4)$$

حل این مساله‌ی بهینه سازی کار مشکلی است، برای ساده‌تر کردن آن از روش ضرایب لاگرانژ استفاده و مقادیر بهینه‌ی w, b بدست آورده می شود.

$$\text{Max}\ (\min L(w,b,\alpha)) \qquad (5)$$

$$\alpha\geq 0 \quad w,b$$
$$L(w,b,\alpha) = 1/2\|w\|2-\sum\alpha_k(y_k((w.x_k)+b)-1$$

دوگان روابط فوق بصورت زیربیان می شود:

$$\partial/\partial b\ L(w,b,\alpha)= 0 \rightarrow \sum \alpha_k y_k = 0 \qquad (6)$$
$$\partial/\partial w\ L(w,b,\alpha)= 0 \rightarrow w= \sum \alpha_k y_k x_k$$

حال اگر مقدار w را در تابع $f(x)= \text{sign}(w.x+b)$ جاگذاری کنیم، رابطه‌ی زیر بدست می‌آید (مینگ چی ،۲۰۰۹) و (چن ما ،۲۰۰۸):

$$f(x) = \text{sign}\ (\sum\alpha_k y_k (x . x_k) + b) \qquad (7)$$

ضـرایب α_k در رابطـه‌ی فـوق، ضـرایب لاگرانـژ نامیـده می‌شوند. به ازای بسیاری از x_k هـا، ضـرایب α_k برابـر صفر است و تنها ضرایب محدودی غیرصفر است نقاطی که به ازای آنها ضرایب α_κ مخالف صـفر هـستند، بردارهـای پشتیبان نامیده می‌شود و همین نقاط، تعیین کننده‌ی مرز جداکننده هستند (هوانگ و هسین چان ،۲۰۰۶).

هرچـه مقـدارα_k کـوچکتر باشـد تـاثیر x_k در مـرز جداکننده کمتر است.درنقـاط بردارپـشتیبان،α_k مخـالف صفر است. در این صورت در این نقاط داریم:

$$y_k((w.x_k)+b)=1 \qquad (8)$$

یعنی بردارهای پشتیبان روی مرز قراردارند . الگوریتم بـالا برای یافتن مرز دو کلاس کاملا جدا از هم بود کـه حاصـل آن یک مرز خطی است.

بردارهای پشتیبان برای داده‌های جداناپذیر

در حالـت جداناپـذیر، نتـایج بخـش قبـل گـسترش داده می‌شود.

Cores در سال۱۹۹۵برای این که بتوانـد در ایـن حالـت یک ابرصفحه‌ی بهینه را برای جداسازی ۲ کـلاس بدسـت آورد، متغیرهای نامنفی$\varepsilon_\kappa \geq 0$ را بعنوان مقدار خطا بـرای هر بردار تعریف کرد. در این حالت معادلات $w.x+b\geq 1$ و $w.x+b\leq 1$ تبدیل به معادلات زیرمی‌شوند:

$$w.x_k + b \geq 1-\varepsilon_k \quad y_k = 1$$
$$w.x_k + b \leq -1+\varepsilon_k \quad y_k = -1 \qquad (9)$$

مقدار ε_κ خطای هر داده تا مرز ناحیه‌ی مربوط بـه خـود اسـت و مقـدار $\sum_{(\kappa=1,\ldots,m)} \varepsilon_\kappa$ یـک خطـای حـد آموزش است. حال بجای کمینه کردن مقـدار $\|w\|^2/2$ بایـد عبـارت زیـر کمینـه شـود. (هوانـگ ،۲۰۰۷) و (فاتیح آکای،۲۰۰۹).

$$\min \|w\|^2 + c\left(\sum \varepsilon_\kappa\right); k=1\ldots m$$
$$y_k((w.x_k)+b)+\varepsilon_\kappa - 1\geq 0 \quad \varepsilon_k \geq 0 \qquad (10)$$

در این رابطه c مقدار ارزش خطا را مشخص می‌کند. اگر c عدد بزرگی انتخـاب شـود، توجـه بیـشتری بـه خطـا داده می شود. با تعمیم معادلات لاگرانژ به حالـت جداناپـذیر و حل آن مشاهده می شود(منیری،۱۳۸۵).

$$\sum \alpha_k y_k = 0 \quad \kappa=1,\ldots,m \qquad (11)$$
$$w=\sum \alpha_k y_k x_k;\ 0\leq \alpha_k \leq c$$

در حـالتی کـه داده‌هـا جداناپذیرنـد و کـلاس‌هـا بـاهم همپوشانی دارند، جدا کردن کلاس‌هـا توسـط مـرز خطـی همواره با خطا همراه اسـت(منیـری،۱۳۸۵). بـرای حـل ایـن مشکل مـی‌تـوان ابتـدا داده‌هـا را از فـضای اولیـه Rn بـا استفاده از یک تبـدیل غیرخطـی φ بـه فـضایی ماننـد H منتقل کرد که در فضای جدید کلاس‌هـا تـداخل کمتـری باهم داشته باشند(منیری،۱۳۸۵).سپس در فـضای جدیـد بـا استفاده ازمعادلات قبلی و جایگزینی x_k با $\varphi(x_k)$، مرز بهینه را محاسبه نمـود . درواقـع در ایـن حالـت داده‌هـای آموزشی در فضای اولیه (ورودی) توسط یک تابع هـسته‌ی φ به فضای ویژگی که دارای بعد بیـشتری اسـت، منتقـل می‌شوند (بلوتی ،۲۰۰۸) و (هوانگ ،۲۰۰۷). لذا در این حالـت یافتن مرز بهینه به حل مساله‌ی بهینه سـازی زیـر تبـدیل می‌شود (بلوتی ،۲۰۰۸).

$$\text{Max}\left[-1/2\sum\sum \alpha_k y_k \alpha_j y_j K(x_k, x_j)\right] \qquad (12)$$
$$0\leq \alpha_k \leq c\ ;\ \sum \alpha_k y_k = 0$$

کـه در آن $K(X_\kappa, X_j) = (\varphi(X_\kappa).\varphi(X_j))$. لـذا ابرصفحه‌ی بهینه‌ی $f(x)$ درحالت غیرخطـی بـصورت زیـر خواهد بود:

$$F(x) = \sum \alpha_k y_k K(x, x_k) + b = \sum_{i \in sv} \alpha_k y_k K(x, x_k) \qquad (13)$$

توابع هسته ای زیادی وجود دارند که در زیر به آنها اشـاره می‌کنیم (چیا مینگ ،۲۰۰۹).

تابع هسته خطی

$$K(x_k, x_j) = (x_k, x_j)$$

تابع هسته‌ی چندجمله‌ای

$$K(x_k, x_j) = (x_k, x_j+1)^d$$

تابع هسته‌ی RBF

$$K(x_k, x_j) = \exp(-\gamma \| x_k - x_j \|^2)$$

پارامترهای توابع هسته باید به گونه‌ای انتخاب شوند که صحت طبقه بندی ماشین بردار پشتیبان را فزایش دهند (هسین ،۲۰۰۴). دراین تحقیق از تابع هسته‌ای RBF استفاده شده است.

شرح روش تحقیق(متدلوژی تحقیق) ساختار اجرایی تحقیق

تحقیق حاضر از این حیث که نوعی الگو برای مدیران و تصمیم گیرندگان ارائه می‌کند تحقیق کاربردی به حساب می آید و از جمع آوری داده شروع شده و سعی در کشف یک سری قوانین دارد. لذا در این تحقیق از روش کشف دانش استفاده شده است و مراحل اجرائی آن به شرح زیر می‌باشد:

۱- جمع آوری داده از پایگاه داده‌های موجود وپیش پردازش آن‌ها

۲-انتخاب ویژگی‌های تاثیر گذار در رفتار اعتباری مشتریان

۳- انتخاب تابع هسته و تنظیم پارامترهای مربوط

۴- تقسیم داده‌های نمونه به دو مجموعه داده‌های آموزشی[1] و داده‌های تست[2]؛

۵- استفاده از الگوریتم ماشین بردار پشتیبان روی مجموعه آموزش

۶- آزمون مدل با مجموعه داده‌های تست؛

۷- اعتبارسنجی مدل و مقایسه با دیگر مدل‌های طبقه بندی.

همچنین با توجه به اینکه ماهیت تحقیق داده محور[3] بوده و پایه اصلی تحقیق حاضر بر کشف دانش از پایگاه داده های بانک مورد مطالعه نهاده شده است، لذا از استاندارد جهانی CRISP-DM جهت انجام فرایند تحقیق استفاده شده است؛ که در این بخش ساختار اجرایی تحقیق بر اساس مراحل این استاندارد تشریح شده است.

مراحل انجام داده کاوی طبق استاندارد CRISP

مدل فرایندی داده کاوی CRISP ، یک متدولوژی استاندارد داده کاوی می باشد که چرخه عمر یک پروژه داده کاوی را نشان می‌دهد. این مدل شامل مراحل مرتبط یک پروژه و وظایف مربوط و ارتباط بین این وظایف است. چرخه عمر یک پروژه داده کاوی طبق این استاندارد ، شامل ۶ مرحله است که در ادامه بیان شده است(نادعلی،۱۳۸۷).

طبق استاندارد CRISP، فرایند داده کاوی شامل موارد زیر است:

- درک مساله کسب و کار
- درک داده‌ها
- آماده سازی داده‌ها
- مدل سازی
- ارزیابی نتایج
- بکارگیری مدل

درک مساله کسب و کار

در این فاز از فرایند تحقیق، ابتدا اهداف اصلی کسب و کار تعیین گردید، که اصلی ترین هدف کسب و کار(بانک مورد مطالعه) در این تحقیق، پیش بینی وضعیت اعتباری مشتریان حقوقی بانک مورد مطالعه می باشد و این کار بر اساس کشف الگوهای پنهان موجود بین خصیصه‌های مشتریانی است که سابقا" تسهیلات اعتباری به آنها اعطا شده است (نادعلی،۱۳۸۷). بر همین اساس موقعیت موجود محیط که بخش اعتبارات بانک می باشد مورد ارزیابی قرار گرفت؛ تا فرایند اعطای اعتبار به مشتریان و بررسیهای کارشناسی خبرگان اعتباری بررسی شد. همچنین نحوه اخذ اطلاعات از مشتریان مورد بررسی دقیق قرار گرفت تا شناخت مناسبی از وضعیت کلی طبقه بندی به مشتریان در حالت فعلی کسب شود (نادعلی،۱۳۸۷).

درک داده‌ها

در این مرحله گام های زیر دنبال شده است که جزئیات آنها در ادامه بیان شده است (نادعلی،۱۳۸۷):

- جمع آوری داده‌های اولیه

1 . Training Data Set
2 . Test Data Set
3 . Data Oriented

- توصیف داده ها
- جستجو در داده‌ها و بررسی کانال های دسترسی
- تصدیق کیفیت داده‌ها و شناسایی داده‌های هدف

در مجموع، مهمترین سوالاتی که در این بخش به آنها پاسخ داده شود، ۲ سوال می باشد: چه داده هایی مورد نیاز است؟ داده ها کجا هستند؟

در این مرحله برای ارزیابی اولیه از داده های موجود، ابتدا، پایگاه داده های موجود در بانک که داده های اعتباری مشتریان حقوقی را در خود ذخیره نموده اند؛ جستجو شد و طبق اظهارات مدیران و کارشناسان مربوطه، نحوه دسترسی به آنها مورد بررسی قرار گرفت.

در این مرحله محقق با نمونه هایی از داده های موجود در پایگاه های فوق آشنا شده و آموزش لازم جهت اخذ داده ها و اطلاعات مورد نیاز از سیستم های عملیاتی کامپیوتری (ماژول های تسهیلات اعتباری و ماژول مشتریان) را فرا گرفته تا با نحوه ذخیره سازی داده ها و کسب اطلاعات از آنها آشنایی کامل پیدا کند(نادعلی،۱۳۸۷).

داده های مورد نظر تحقیق عبارتند از : مشتریان حقوقی و خصیصه های هر یک از آنها؛ که در ادبیات اعتباری به عنوان متغیرهای تاثیرگذار در رفتار اعتباری شناخته شده اند. بنابراین فقط این نوع ویژگی ها از پایگاه داده ها مورد نظر بوده و جمع آوری خواهند شد. هر چند پایگاه های داده مملو از داده های زائد می باشند که از ثبت آنها صرف نظر می گردد.

بنابراین نهایتا" در این مرحله، محقق، داده هایی را که باید جمع آوری و پالایش نماید، شناسایی نمود.

آماده سازی داده ها برای مدل

در این قسمت ، همانطور که از نام آن پیداست باید داده‌های خام را برای ورود به الگوریتم آماده کرد. در این تحقیق، آماده سازی داده‌ها به عنوان بخشی از فرایند داده‌کاوی، دارای مراحل زیر می‌باشد که در ادامه به تشریح فرایند بکار رفته می‌پردازیم:

جمع‌آوری و انتخاب داده‌ها

یکپارچه‌سازی داده‌ها

پاکسازی داده‌ها

کاهش ابعاد داده‌ها

تعریف کلاسها

قالب بندی داده‌ها (برای مدلسازی در نرم‌افزار ا

از آنجا که حجم داده‌های خام بسیار بالا است و فرمت آنها نیز فرمت استاندارد الگوریتم بکار رفته نیست، لذا در این مرحله بایستی اقدامات پاکسازی داده‌ها و کاهش حجم آنها صورت گیرد. در این مرحله همچنین بررسی‌هایی در مورد ویژگی‌های کلاس‌های خوش حساب و بد حساب صورت گرفت .

- **جمع‌آوری و انتخاب داده‌ها:** در این مرحله، داده‌های مرتبط با مساله(که در مرحله اول فرایند داده کاوی مشخص شد)از منابع موجود که شامل پرونده‌های متقاضیان تسهیلات اعتباری و پایگاه داده‌های مشتریان بود، جمع‌آوری شد.
- **یکپاچه‌سازی و پاکسازی سازی داده‌ها:** در صورتیکه داده‌ها از منابع مختلف بدست آمده و دارای فرمت یکسانی نباشند، در این مرحله بایستی یک شکل و متجانس شوند و پس از آن کار پاکسازی و ادغام صورت گیرد که عملا کار سخت و زمانبری است.حذف برخی رکوردها و فیلدها از کارهای صورت گرفته در این مرحله می‌باشد. بسیاری از رکوردها غیرضروری بوده و کمکی به ساخت یک طبقه‌بندی خوب و شناسایی مشتریان بدحساب نمی‌کردند و لذا حجم داده‌ها را کاهش داده و این رکوردها را حذف نمودیم. گاها در مجموعه داده‌های اولیه، مقادیر گم‌شده در فیلدهای برخی رکوردها وجود داشت و یا در بعضی موارد، مقادیر تخصیص یافته به متغیرها اشتباه بودند که تصحیح و یا حذف شدند.
- **کاهش داده‌ها:** در این مرحله، علی‌رغم پاکسازی و پیش پردازش داده‌ها در مرحله قبل، مواردی از مشتریان با اطلاعات ناکافی وجود داشت که از مجموعه داده‌ها حذف شدند.
- **کاهش ابعاد داده‌ها:** افزایش ابعاد داده‌ها باعث پیچیدگی فرایند و افزایش زمان آموزش و همچنین پائین آمدن کیفیت جواب می‌شود. لذا تعدادی از ویژگی‌ها از جمله نام متقاضی و ... که بی تاثیر بودند با استفاده از روش F-score شناسایی و حذف شدند که نتیجه آن افزایش استقلال بین متغیرها می‌باشد.

- **تعریف کلاسها:** در تعداد اندکی از پرونده‌های مورد بررسی، نوع کلاسها مشخص نشده و یا به اشتباه مشخص شده بود که اصلاحات لازم صورت گرفت.
- **قالب بندی داده‌ها (برای مدلسازی در نرم‌افزار):** همچنین با توجه به ماهیت داده‌ها درالگوریتم ماشین بردار پشتیبان، کلیه متغیرهای اسمی، عددی شده و پس از آن بی‌مقیاس شده و با استفاده از فرمول زیر به بازه $[0,1]$ منتقل شده‌اند.

$$x' = \frac{x - \min_a}{\max_a - \min_a}$$

تحلیل داده‌ها و ساخت مدل

با داشتن مجموعه داده مناسب با فرمت استاندارد، می‌توان مرحله بعد را شروع کرد. هدف این مرحله ساخت مدلی جهت اعتبارسنجی متقاضیان تسهیلات اعتباری با استفاده از داده‌های در دسترس است. از تکنیک‌های داده کاوی متعددی می‌توان در این مرحله استفاده نمود که در این تحقیق از ماشین بردار پشتیبان استفاده شده است.

آموزش مجموعه داده‌های آماده سازی شده

در این مرحله، داده‌های آماده توسط الگوریتم ماشین بردار پشتیبان دو کلاسه غیرخطی آموزش داده شدند و بعد نتایج آن با دیگر الگوریتم ها مقایسه شدند که در بخش‌های آتی به توضیح آن خواهیم پرداخت.

- **زیر فرایند مرتبط با داده**

در این قسمت از تکنیک f-score برای پیش پردازش داده‌ها جهت بهبود کیفیت داده‌های ورودی استفاده شده است.

- **زیر فرایند مرتبط با مدل**

در این بخش داده‌ها را با استفاده از تکنیک ماشین بردار پشتیبان مدل سازی نموده و پس از آن در مرحله بعدی تحلیل های لازم در خصوص اعتبارسنجی مدل و کارایی آن ارائه می گردد.

ارزیابی نتایج

در این مرحله مدل موردنظر با دیگر مدل‌ها مقایسه و ارزیابی می‌شود. معیارهای ارزیابی شامل دقت پیش بینی، سرعت و... می‌باشد. روش‌های ارزیابی شامل ماتریس اغتشاش[1]، منحنی ROC، تابع هزینه و غیره است که در این تحقیق از ماترس اغتشاش استفاده شده است و نتایج آن در ادامه آمده است. هدف ارزیابی، انتخاب مدل مناسب و یا در صورت امکان، مدل بهینه در بین تعدادی از مدل‌هاست.

مرحله اعتبارسنجی وارزیابی مدل :در این بخش شیوه ارزیابی مدل و همچنین سنجش کارایی مدل به لحاظ قدرت تفکیک مشتریان در مقایسه با الگوریتم های مختلف و کارشناسان بانک مورد مطالعه، تشریح شده است.

اعتبارسنجی عرضی n لایه: برای دستیابی به یک تخمین پایا و کمینه کردن تاثیر و وابستگی به داده‌های خاص در مدل ارزیابی اعتبار جدید، برای ساخت بخش‌های تصادفی از مجموعه داده‌ها، از n-fold cross validation ، استفاده شد. در این رویه، مجموعه داده‌ها به k گروه مستقل تقسیم می‌شوند. الگوریتم با استفاده از k-1 گروه آموزش داده می‌شود و تنها با استفاده از گروه k ام تست می‌شود. این رویه تا زمانی تکرار می‌شود که هر گروه به عنوان گروه آزمایش نیز بکار رفته باشد. مزیت رویه اعتبارسنجی k لایه، استفاده از تمامی لایه‌ها برای آموزش و همچنین تست الگوریتم‌ها است. برآورد دقت الگوریتمها در اعتبارسنجی چندلایه بسیار بیشتر از روش‌های معمولی قابل اعتماد است.در این تحقیق مقدار k برابر ۵ است.

در انتها برای ارزیابی قابلیت اطمینان روش ارائه شده، نتایج الگوریتم مورد استفاده را با الگوریتم های دیگر مانند الگوریتم درخت های تصمیم ، شبکه های عصبی، شبکه های بیزین، رگرسیون لجستیک و ... نیز استفاده شده است. برای اعمال این تکنیک ها از نرم افزار های مربوطه استفاده شده است. نتایج حاصله از انجام این کار در جداول ۱-۵ تا ۱-۱۰ ارائه شده‌است. نتایج حاکی از آن است که مدل پیشنهادی نه تنها دارای بهترین نرخ

1 . Confusion matrix

طبقه بندی است بلکه دارای کمترین خطای نوع دوم نیز هست.

بکارگیری مدل

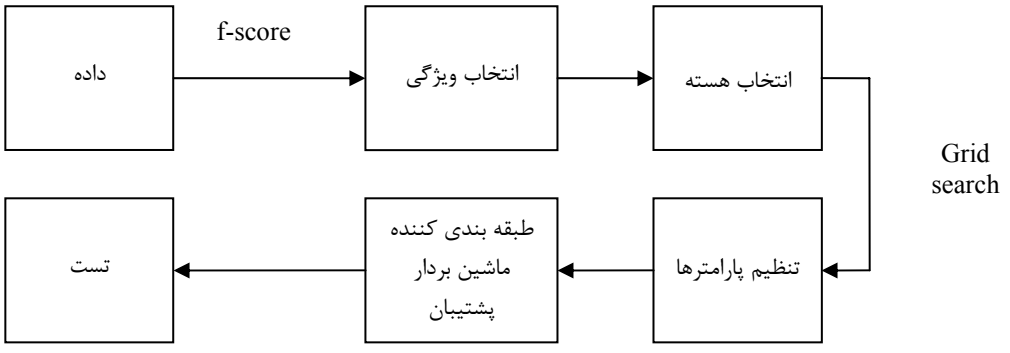
کاراصلی تحلیل گر داده[1] یا داده کاو[2] در مرحله قبل تقریبا به اتمام رسیده و در این تحقیق نیز نتایج، بصورت یک گزارش از کل کار در اختیار بانک مورد مطالعه قرار می گیرد تا بانک نیز بر اساس استراتژی های خود راجع به استفاده از نتایج تصمیم گیری نماید.

رویه‌ی اعتبار سنجی مبتنی بر ماشین بردار پشتیبان

اعتبار سنجی در واقع شامل طبقه بندی و رتبه‌های (خوب و بد)تخصیص یافته به متقاضیان و شرایط متقاضی از قبیل مقدار وام ،موقعیت کاری، اطلاعات شخصی ، سن و... به عنوان متغیرهای طبقه بندی است . رویه‌ی اعتبار سنجی مبتنی بر ماشین بردار پشتیبان شامل : جمع آوری داده و پیش پردازش آن، انتخاب ویژگی‌ها ، انتخاب هسته و پارامترهای مربوطه ، آموزش، استفاده از طبقه بندی کننده‌ی ماشین بردار پشتیبان ،تست و در نهایت اعتبار سنجی برای نمونه‌های جدید می‌باشد (چن ،۲۰۰۸). همانطور که می‌دانیم تنظیم درست پارامترهای مدل ، می‌توانند صحت مدل طبقه بندی کننده‌ی بهینه‌ی ماشین بردار پشتیبان را بالا ببرند (چن ،۲۰۰۸). پارامترهایی که بایستی درصورت استفاده از هسته‌ی RBF بهینه شوند ، عبارتند از: گاما . در این تحقیق برای پیدا کردن ویژگی‌های مدل از تکنیک F-Score و پارامترها از Grid Search استفاده شده است. شکل ۲ رویه‌ی مدل مبتنی بر ماشین بردار پشتیبان را نشان می‌دهد .

برای بهبود صحت اعتبارسنجی از F-Score برای انتخاب ویژگی ها والگوریتم Grid Search برای بدست آوردن پارامترهای بهینه استفاده شده است (چن ،۲۰۰۸). درالگوریتم Grid Search برای هرجفت پارامتر، 5-fold cross validation روی مجموعه آموزش بکاررفته است واز نرم افزار Weka برای بکارگیری واجرای ماشین بردارپشتیبان استفاده شده است.

شکل ۲: رویه مبتنی بر ماشین بردار پشتیبان



1 . Data Analyst

2 . Data Miner

انتخاب ویژگی

سه مساله‌ی اساسی هنگام استفاده ازماشین بردار پشتیبان درطبقه بندی وجود دارد که عبارتنداز:۱) انتخاب ویژگی‌های بهینه ۲) انتخاب هسته ۳) تعیین پارامترهای هسته (چن ،۲۰۰۸).

انتخاب ویژگی یکی ازمباحث مهم درساخت مدل‌های طبقه بندی است . هرچه تعدادویژگی‌های ورودی کمترباشد پیش بینی بهترصورت گرفته ومیزان پیچیدگی محاسباتی مدل‌ها کاهش می‌یابد(هوانگ ،۲۰۰۷) .درواقع انتخاب ویژگی‌ها یکی ازمسائل اساسی بهینه‌سازی است که شامل جستجو در فضا برای یافتن ویژگی‌هایی است که یا بهینه‌اند و یا نزدیک به بهینه‌اند واین امر با در نظر گرفتن یک معیارعملکردی معین مانند صحت ، صورت می‌گیرد (مین،۲۰۰۶).

روشF-Score

روش F-Score ساده‌ترین تکنیک برای انتخاب ویژگی‌هاست که تفاوت دو مجموعه از اعداد حقیقی را اندازه گیری می‌کند. با درنظر گرفتن بردارهای آموزشxk ; k=1,…,m و n+ و- n به عنوان تعداد نمونه‌های مثبت و منفی، F-Score مربوط به i امین ویژگی بصورت زیر تعریف می‌شود(طلوعی، مقدوری،۱۳۸۸):

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2},$$

که $\bar{x}_i, \bar{x}_i^{(+)}, \bar{x}_i^{(-)}$ درآن به ترتیب متوسط کل، مثبت و منفی iامین ویژگی مجموعه‌ی داده‌ها هستند و $x_{k,i}^{(+)}$ iامین ویژگی مربوط به k امین نمونه‌ی مثبت و $x_{k,i}^{(-)}$ iامین ویژگی مربوط به k امین نمونه‌ی منفی است. صورت کسر نشان دهنده‌ی تفاوت بین مجموعه‌های مثبت و منفی، و مخرج بیانگر یکی از دو مجموعه است و هرچه مقدار F- Score بیشتر باشد با احتمال بیشتری، آن ویژگی متمایز است (مین ،۲۰۰۶). رویه‌ی به‌کار رفته به شرح زیر می‌باشد:

برای هر زیر مجموعه‌ی k تایی از مجموعه داده‌ی D، مجموعه‌ی آموزشی T=D-K را درنظر می‌گیریم، سپس فرایند اعتبارسنجی متقابل را انجام داده و میانگین صحت کلی را برای تمامی k بخش حساب می شود.

این رویه به ترتیب زیر می‌باشد:

گام ۱) محاسبه‌ی F-Score برای هریک از ویژگی‌ها

گام۲) دسته بندی F-Score و محاسبه‌ی مقدار ممکن ویژگی‌ها با استفاده از فرمول $f = [n / 2^i]$ که $i \in \{0,\ldots,m\}$ و m عدد صحیحی است که: $n / 2^m \geq 1$

گام۳) انجام مراحل زیر برای هر یک از fها:

- حفظ ویژگی‌های f اول بدست آمده ازF-Score
- تقسیم داده‌های آموزشی بطور تصادفی به ۲ قسمت $D_{training}$ و $D_{validation}$ با استفاده از 5-fold cross validation و تکرار مراحل زیر برای هر یک از fold ها.
- در نظرگرفتن $D_{training}$ به عنوان داده‌ی آموزشی جدید و استفاده از رویه‌ی ماشین بردار پشتیبان جهت رسیدن به یک پیشگویی کننده و پس از آن، استفاده از پیشگویی کننده جهت پیش بینی $D_{validation}$
- محاسبه‌ی متوسط خطای صحت 5-fold cross validation

گام۴) انتخاب f دارای حداقل متوسط خطای صحت

گام۵) حذف ویژگی‌هایی که دارای F-Score کمتر از f هستند .

استخراج ویژگی‌های مهم

جدول زیر ویژگی‌های مهم انتخاب شده توسط تکنیک f-score و ویژگی‌های انتخابی طبق نظر خبرگان را نشان می‌دهد.محاسبه مقادیر Fi در f-score توسط نرم‌افزارExcel و بقیه مراحل آن توسط نرم‌افزار weka انجام شده است. جدول ۱، مقادیرf-score مربوط به هر ویژگی را نشان می‌دهد.با توجه به محاسبات، مقدار f = ۱۹ می‌باشد . همچنین تعداد ویژگی‌های اولیه n=26 است، لذا داریم :

جدول ۱: مقادیر f-score مربوط به ویژگی‌ها

ردیف	ویژگی	f-score	نرخ صحت متوسط(٪)
۱	شغل (job)	14.89	57.6
۲	تحصیلات (education)	13.74	59.3
۳	نوع وثیقه (colla-kind)	11.31	61.7
٤	سابقه اعتباری (bank-ref)	9.52	65.2
۵	سن شرکت (company-age)	4.73	67.5
۶	میزان سرمایه (capital)	2.90	71.8
۷	مقدار وام (money)	2.11	77.4
۸	نوع تسهیلات (contract)	1.07	79.1
۹	نوع مصرف وام (use-kind)	0.83	81.3
۱۰	نوع شرکت (company-kind)	0.57	82.3
۱۱	نسبت جاری (current-ratio)	0.41	83.6
۱۲	نسبت بدهی (debt-ratio)	0.32	85.8
۱۳	نرخ بهره (,benefit ratio)	0.20	86.3

جدول ۲: ویژگی های بدست آمده از تکنیک F-score

ردیف	عنوان متغیر	مقادیر	نوع متغیر
۱	شغل (job)	بازرگانی- خدماتی- تولیدی	کیفی
۲	تحصیلات (education)	۱-زیردیپلم۲- دیپلم ۳- فوق دیپلم ٤- لیسانس ۵- فوق لیسانس و بالاتر	کیفی
۳	نوع وثیقه (colla-kind)	۱-چک عادی۲- چک تضمینی۳- سفته مدیران ٤-سفته اعتباری ۵-سند ملکی	کیفی
٤	سابقه اعتباری (bank-ref)	۱- :سابقه منفی ۱:سابقه مثبت	کمی
۵	سن شرکت (company-age)	به صورت عدد است.	کمی
۶	میزان سرمایه (capital)	به صورت عدد است.	کمی
۷	مقدار وام (money)	به صورت عدد است.	کمی
۸	نوع تسهیلات (contract)	۱-اقساطی۲- مضاربه ۳- مشارکت مدنی ٤- جعاله	کیفی
۹	نوع مصرف وام (use-kind)	تولید و صنعت- خدمات- واردات- بازرگانی داخلی	کیفی
۱۰	نوع شرکت (company-kind)	تعاونی- محدود- عام-خاص	کیفی
۱۱	نسبت جاری (current-ratio)	به صورت درصد است.	کمی
۱۲	نسبت بدهی (debt-ratio)	به صورت درصد است.	کمی
۱۳	نرخ بهره (,benefit ratio)	به صورت درصد است.	کمی

جدول ۳: ویژگی های بدست آمده طبق نظر خبرگان

ردیف	عنوان متغیر	مقادیر	نوع متغیر
۱	زمینه فعالیت	بازرگانی- خدماتی- تولیدی	کیفی
۲	تحصیلات	۱-زیردیپلم۲- دیپلم ۳- فوق دیپلم ٤- لیسانس ۵- فوق لیسانس و بالاتر	کیفی
۳	نوع وثیقه	۱-چک عادی۲- چک تضمینی۳- سفته مدیران٤-سفته اعتباری۵-سند ملکی	کیفی
٤	شخصیت حقوقی و اجتماعی فرد(سابقه اعتباری)	۱- :سابقه منفی ۱:سابقه مثبت	کمی
۵	نوع مصرف وام	تولید و صنعت- خدمات- واردات- بازرگانی داخلی	کیفی
۶	سال تاسیس شرکت	به صورت عدد است.	کمی
۷	نوع شرکت	تعاونی- محدود- عام-خاص	کیفی

تنظیم پارامترها

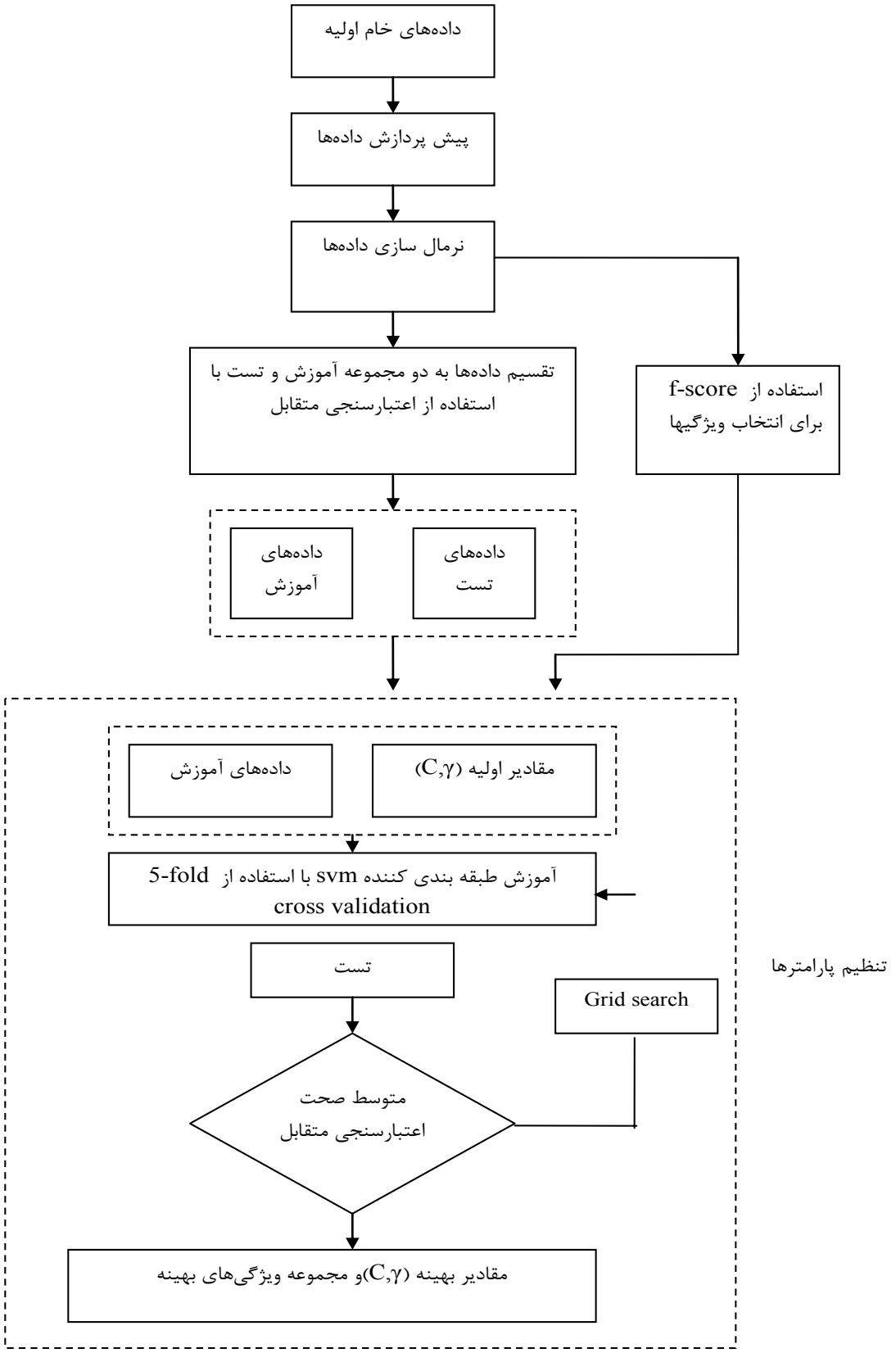
ماشین بردار پشتیبان نیازمند تنظیم یکسری پارامترها قبل از طبقه بندی است (منیری،۱۳۸۵). این پارامترها روی یک مجموعه داده ی جداگانه ی مستقل از مجموعه داده ی بکار رفته برای طبقه بندی ، تنظیم می شوند که این مجموعه داده بکار رفته برای تنظیم پارامترها شامل ۲۳۶ پرونده است . این مجموعه داده به ۲ قسمت آموزش و تست ، تقسیم شده است و از Grid-Search برای تعیین و شناسایی آن مقادیر پارامترهایی که دارای بیشترین متوسط صحت اعتبار سنجی متقابل هستند، استفاده شده است. جدول زیر مقادیر(C,γ) بدست آمده را برای مقادیر $C= 2^{-5},\ldots..2^{13}$ و $\gamma =2^{-13},\ldots,2^{5}$ و مقدار بهینه انتخابی(C,γ)=۷۹,۹۲ را نشان می‌دهد.

جدول ۴: نتایج استفاده از grid search با استفاده از 5-fold cross validation

C	γ									
	2^{-13}	2^{-11}	2^{-9}	2^{-7}	2^{-5}	2^{-3}	2^{-1}	2^{1}	2^{3}	2^{5}
2^{-5}	61.32	61.32	61.32	61.37	61.32	61.32	61.32	61.32	61.32	61.32
2^{-3}	61.32	61.32	61.32	61.37	61.32	77.48	65.3	66.4	67.39	61.32
2^{-1}	61.32	61.32	61.32	61.37	68.2	79.92	66.29	66.93	67.7	62.02
2^{1}	61.32	61.32	62.46	62.46	62.68	78.39	67.09	67.94	68.45	62.28
2^{3}	61.32	61.32	62.59	62.51	63.74	77.46	67.52	68.53	66.9	62.69
2^{5}	61.32	62.58	62.59	63.61	63.89	75.46	69.49	71.44	70.44	63.7
2^{7}	62.47	63.62	63.54	65.24	67.12	75.32	72.46	69.36	71.33	63.54
2^{9}	63.54	66.68	67.36	68.37	69.8	75.32	73.38	75.28	71.96	66.34
2^{11}	66.34	69.84	69.63	70.73	70.72	75.24	75.62	76.39	72.03	69.73
2^{13}	69.73	69.84	69.91	70.92	72.48	74.11	75.83	76.9	72.28	69.93

برای تنظیم پارامترها نیز از نرم افزار weka-libsvm استفاده شده است.

چارچوب اجرائی مبتنی بر ماشین بردار پشتیبان



مقایسه الگوریتم های مختلف طبقه بندی جهت بررسی کارایی مدل

برای سنجش کارایی مدل ارائه شده توسط ماشین بردار پشتیبان، صحت نتایج این مدل با الگوریتم های طبقه بندی دیگری مورد مقایسه قرار گرفته اند. این الگوریتم ها عبارتند از:درخت تصمیم C4.5 Tree، رگرسیون لجستیک، شبکه های بیزین، شبکه‌های عصبی چند لایه و درخت ژنتیکی که نتایج طبقه بندی هر یک از آنها در جدول های ۱-۵ تا ۱-۱۰نمایش داده شده است. همچنین کارایی مدل با عملکرد کارشناسان اعتبارسنج بانک نیز مقایسه شده است که نتایج آن در جدول ۱۲ آمده است.

نتایج طبقه‌بندی با مدل SVM

جدول ۵ : ماشین بردار پشتیبان

SVM(Support Vector Machine)		
نمونه های طبقه بندی شده بطور صحیح	181	%76.7
نمونه های طبقه بندی شده بطورغلط	55	%23.3
کل نمونه ها	236	%100

	طبقه بندی شده در کلاس خوش حساب	طبقه بندی شده در کلاس بد حساب	مجموع
کلاس خوش حساب	12	30	150
کلاس بد حساب	25	61	86
مجموع	145	91	236

جدول ۶ : درخت تصمیمC4.5 Tree

C4.5 Tree		
نمونه های طبقه بندی شده بطور صحیح	158	%67
نمونه های طبقه بندی شده بطورغلط	78	%33
کل نمونه ها	236	%100

	طبقه بندی شده در کلاس خوش حساب	طبقه بندی شده در کلاس بد حساب	مجموع
کلاس خوش حساب	99	51	150
کلاس بد حساب	27	59	86
مجموع	126	110	236

جدول۷ : رگرسیون لجستیکLogistic –R

Logistic –R		
نمونه های طبقه بندی شده بطور صحیح	153	%64.8
نمونه های طبقه بندی شده بطورغلط	83	%35.2
کل نمونه ها	236	%100

	طبقه بندی شده در کلاس خوش حساب	طبقه بندی شده در کلاس بد حساب	مجموع
کلاس خوش حساب	106	44	150
کلاس بد حساب	39	47	86
مجموع	145	91	236

جدول۸ : شبکه‌های بیزین

BayesNet		
نمونه های طبقه بندی شده بطور صحیح	170	%72
نمونه های طبقه بندی شده بطورغلط	66	%28
کل نمونه ها	236	%100

	طبقه بندی شده در کلاس خوش حساب	طبقه بندی شده در کلاس بد حساب	مجموع
کلاس خوش حساب	95	55	150
کلاس بد حساب	11	75	86
مجموع	106	130	236

جدول ۹ : شبکه‌های عصبی چند لایه MultilayerPerceptron

MultilayerPerceptron(MLP)		
نمونه های طبقه بندی شده بطور صحیح	146	%61.8
نمونه های طبقه بندی شده بطورغلط	90	%38.2
کل نمونه ها	236	%100

	طبقه بندی شده در کلاس خوش حساب	طبقه بندی شده در کلاس بد حساب	مجموع
کلاس خوش حساب	128	22	150
کلاس بد حساب	68	18	86
مجموع	196	40	236

جدول۱۰ : درخت ژنتیکیGATree

GATree		
نمونه های طبقه بندی شده بطور صحیح	165	%69.9
نمونه های طبقه بندی شده بطورغلط	71	%30.1
کل نمونه ها	236	%100

	طبقه بندی شده در کلاس خوش حساب	طبقه بندی شده در کلاس بدحساب	مجموع
کلاس خوش حساب	۱۰۳	۴۷	۱۵۰
کلاس بد حساب	۲۴	۶۲	۸۶
مجموع	۱۲۷	۱۰۹	۲۳۶

جدول ۱۱ : مقایسه نتایج صحت مدل‌ها

الگوریتم	SVM	C4.5 Tree	Logistic R	Bayes Net	NN(MLP)	GA Tree
صحت پیش بینی مدل	%76.7	%67	%64.8	%72	%61.8	%69.9

کارایی مدل در مقایسه با عملکرد کارشناسان اعتبارسنج بانک

جدول۱۲ :عملکرد کارشناسان اعتبارسنج بانک در خصوص پیش بینی طبقه مشتریان

کارشناسان اعتبارسنجی بانک		
نمونه های طبقه بندی شده بطور صحیح	134	%56.8
نمونه های طبقه بندی شده بطورغلط	102	%43.2
کل نمونه ها	236	%100

	طبقه بندی شده در کلاس خوش حساب	طبقه بندی شده در کلاس بد حساب	مجموع
کلاس خوش حساب	84	56	150
کلاس بد حساب	46	50	86
مجموع	130	106	236

نتیجه گیری

در این تحقیق که هدف اصلی آن بررسی کارائی استفاده از ماشین بردار پشتیبان در طبقه‌بندی متقاضیان بوده است، نتایج مدل ماشین بردار پشتیبان در مقایسه با مدل‌های رگرسیون لجستیک و... بررسی شده است. یافته‌های تحقیق حاکی از آنست که در طبقه‌بندی متقاضیان، مدل ماشین بردار پشتیبان نسبت به مدلهای دیگر بطور معناداری از دقت کلی بیشتری برخورداراست .در مقاله حاضر با استفاده از آمار و اطلاعات مربوط به مشتریان حقوقی بانک مورد مطالعه، که در طی سالهای ۸۴ تا ۸۶ از بانک تسهیلات دریافت نموده‌اند و با بهره گیری از روش ماشین بردار پشتیبان، سعی شد عوامل موثر بر عدم بازپرداخت بموقع تسهیلات اعتباری شناسایی شود تا بدین وسیله اطلاعات لازم برای درجه بندی مشتریان آتی بانک فراهم گردد.نتایج بدست آمده از این مطالعه نشان داد که متغیرهای نوع وثیقه، سابقه اعتباری و زمینه فعالیت ،از جمله مهمترین عوامل تاثیر گذار بر بازپرداخت تسهیلات هستند.

نتایج بدست آمده از برآورد مدل ماشین بردار پشتیبان نشان داد که اگر بانک در پرداخت تسهیلات به اشخاص حقوقی، به عوامل تاثیرگذاری که در این مطالعه مشخص شد توجه کند، می تواند احتمال بازپرداخت تسهیلات را افزایش دهد. همچنین نتایج حاصل از مقایسه تکنیک ها نشان داد که ماشین بردار پشتیبان در مقایسه با دیگر تکنیک‌های طبقه‌بندی از کارائی بالاتری برخوردار است.از این رو تدوین یک نظام جامع اعتبارسنجی برای بانک در جهت کاهش ریسک عدم بازپرداخت تسهیلات یک ضرورت و در عین حال یک گام موثر در این راستاست.

منابع و مآخذ :

۱. اخباری، مهدیه، « رتبه بندی اعتباری مشتریان حقوقی بانک‌ها با رویکرد هوش مصنوعی» ، پایان نامه کارشناسی ارشد، دانشگاه صنعتی اصفهان،۱۳۸۷

۲. طلوعی اشلقی، عباس، مقدوری شربیانی، فرناز، دانشگر، فرید، « امتیاز دهی اعتباری متقاضیان کارتهای اعتباری بانک ها با استفاده از تکنیک ماشین بردار پشتیبان »، دومین کنفرانس شهر الکترونیکی،۱۳۸۸

۳. منیری آرش،« استفاده از ماشین بردار پشتیبان در باز شناسی کلمات گسسته فارسی »،پایان نامه کارشناسی ارشد، دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران،۱۳۸۵

۴. نادعلی، احمد، « طبقه بندی متقاضیان تسهیلات اعتباری بانکی با استفاده از داده کاوی و منطق فازی » ، پایان نامه کارشناسی ارشد، دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران،۱۳۸۷

5. Avci, Engin, "Selecting of the optimal feature subset and kernel parameters in digital modulation classification by using hybrid genetic algorithm-support vector machines:HGASVM", Expert systems with applications, Vol.36, pp. 1391-1402, 2009.
6. Bellotti, Tony, Crook, Jonathan, "Support vector machines for credit scoring and discovery of significant features", Expert systems with applications, pp. 102-109, 2008.
7. Chia-Ming Wanmg, Yin-Fu Huang, " Evolutionary- based feature selection approaches with new criteria for data mining:" A case study of credit approval data", Expert systems with applications, Vol.36, pp. 5900-5908, 2009.
8. Chen, Weimin, Ma, Chaoqun, Ma, Lin, "Mining the customer credit using hybrid support vector machine technique", Expert systems with applications, pp.1-6, 2008.

9. Fatih Akay, Mehmet, "support vector machines combined with feature selection for breast cancer diagnosis", Expert systems with pp.3240-3247, 2009 applications, Vol.36.
10. Hsieh, N. C., "An integrated data mining and behavioral scoring model for analyzing bank customers", Expert systems with applications, Vol.4, pp.623-633, 2004.
11. Huang, Cheng-Lung, Chen, Mu-Chen, "Credit scoring with a data mining approach based on support vector machines", Expert systems with applications, Vol.3,pp.847-856,2007.
12. Huang, Zan, Chen, Hsinchun, "Credit Rating analysis with support vector machines and neural networks: a market comparative study", Decision support systems, Vol.37, pp.549-551, 2006.
13. Ming-Chi Lee, "Using support vector machine with a hybrid feature selection method to the stock trend prediction", Expert systems with applications, 2009.
14. Min, Sung-Hwan, Lee, Jumin, "Hybrid genetic algorithms and support vector machines for bankruptcy prediction", Expert systems with applications, Vol.31, pp. 652-660, 2006.
15. Sinky jr.,Joseph f., "commercial Bank Financial Management", 4th Edition; Macmillan, p.519,1992.