فصلنامه علمی- پژوهشی مدیریت دارایی و تأمین مالی
سال پنجم، شماره چهارم، شماره پیاپی (۱۹) زمستان ۱۳۹۶
تاریخ دریافت: ۹۵/۱/۱۸ تاریخ پذیرش: ۹۵/۸/۲۴
صص: ۳۲ – ۱۷

Credit Risk Management of Banking Customers Using Support Vector Machine Optimized by Genetic Algorithm with Data Mining Approach

Meisam Jafari Eskandari[1]*, Milad Roohi[2]
1- Associate Prof., Faculty of Industrial Engineering, Industrial Engineering Department, Payame Noor University, Tehran, Iran
meisam_jafari@pnu.ac.ir
2- MSc. Industrial Engineering Department, Faculty of Industrial Engineering Payame Noor University, Tehran, Iran
milad.roohi@gmail.com

Abstract
Credit risk management, credit scoring and risk assessment of customers is an important issue in banking industry. Credit scoring is important because if the banks fail to earn resource allocation and create a balance between the processes of mobilizing and allocating resources, they are typically faced with many problems in the future. According to official figures released by the Central Bank of Iran in recent years, the rate of bad loans increased, since the systemic strict validation to evaluate and measure the credit risk of customers do not exist. In this paper, we try to predict customer's recovery rate index with data mining techniques. Markedly, in recent years in the world new method for measuring customer risk rather than default probability measure has been considered, but due to low precision of forecasting models widely different approaches in research and modeling is investigated. The method used in this paper is support vector regression model whose parameters selection is optimized with genetic algorithm.
Keywords: Credit risk, Recovery rates, Rialclaims, Support vector machine

مدیریت ریسک اعتباری مشتریان بانکی با استفاده از روش ماشین بردار تصمیم بهبودیافته با الگوریتم ژنتیک با رویکرد داده‌کاوی

میثم جعفری اسکندری[۱*]، میلاد روحی[۲]
۱- استادیار گروه مهندسی صنایع، دانشکده مهندسی صنایع، دانشگاه پیام نور تهران، ایران
meisam_jafari@pnu.ac.ir
۲- کارشناسی ارشد گروه مهندسی صنایع، دانشکده مهندسی صنایع، دانشگاه پیام نور تهران ،ایران
milad.roohi@gmail.com

چکیده
مدیریت ریسک اعتباری، رتبه‌بندی اعتباری و ارزیابی میزان ریسک مشتریان، در کنار جذب منابع از اهمیت بالایی برای بانک‌ها برخوردار است؛ زیرا اگر بانک‌ها با تخصیص بهینهٔ منابع و کسب درآمد بین فرایند تجهیز و تخصیص منابع خود نتوانند توازن ایجاد کنند، در آینده با مشکلات زیادی روبه‌رو می‌شوند. براساس آمارهای رسمی منتشرشده از سوی بانک مرکزی ج.ا.ا در سال‌های اخیر، میزان مطالبات معوق بانک‌ها بسیار افزایش یافته است؛ زیرا سیستم اعتبارسنجی دقیقی برای ارزیابی اعتبار و اندازه‌گیری میزان ریسک مشتریان وجود ندارد. در این پژوهش، الگویی با استفاده از روش‌های داده‌کاوی برای پیش‌بینی شاخص نرخ وصول مشتریان ارائه می‌شود. رویکردی که در سال‌های اخیر در دنیا به‌عنوان روشی جدید برای اندازه‌گیری ریسک مشتریان به‌جای اندازه‌گیری احتمال نکول مدّنظر قرارگرفته است. نتایج نشان می‌دهد الگوی پیشنهادی این پژوهش، دقت بیشتری دارد. به‌طور کلی، هدف پیش‌بینی درصد وصول مطالبات قراردادهای با احتمال ریسک مطالباتی بالا قبل از اعطای تسهیلات است.
واژه‌های کلیدی: ریسک اعتباری، ماشین بردار تصمیم، مطالبات ریالی، نرخ وصول

* نویسنده مسؤول

مقدمه

در سال‌های اخیر، توانایی تولید، ضبط و ذخیرهٔ داده‌ها بسیار افزایش یافته است. اطلاعاتی که در این داده‌ها می‌تواند نهفته باشد، بسیار مهم است. در دسترس‌بودن حجم بالای داده و نیاز به تبدیل آنها به دانش، صنعت فناوری اطلاعات را برای استفاده از داده‌کاوی تشویق کرده و به سمت آن سوق داده است. صنعت بانکداری در مسیر کسب‌وکار خود در سراسر جهان، دستخوش تغییر فوق‌العاده‌ای شده است. همچنین این صنعت شروع به شناخت تکنیک‌ها و مهارت‌های استفاده از داده‌کاوی برای استفاده در رقابت در بازار بانکی کرده است. بانک‌ها با استفاده از ابزار داده‌کاوی به بخش‌بندی مشتریان، مطالعه سوددهی، رتبه‌بندی اعتباری و پیش‌بینی پرداخت تسهیلات و وصول مطالبات، بازاریابی، شناسایی و کشف تقلب در تراکنش‌ها و غیره توجه کرده‌اند [۱۰]. مؤسسات اعتباری برای پیش‌بینی وضعیت افرادی که در آینده از عهدهٔ انجام تعهدات خود بر نخواهند آمد، مایل به ارزیابی مشتریان هستند. در هر دو حالت (بررسی درخواست‌های وام جدید و کنترل وام‌گیرندگان قبلی) احتمال بازپرداخت در دورهٔ وام‌دهی تخمین زده می‌شود و درنتیجه، مشتریان براساس تخمین حاصل دربارهٔ ناتوانی در بازپرداخت به سطوح متفاوتی از ریسک رتبه‌بندی خواهند شد. این روش به‌عنوان تعیین ریسک و یا طبقه‌بندی اعتبار شناخته می‌شود [۱۴]. در سال‌های اخیر در مباحث مربوط به ریسک اعتباری، مفهوم تازه‌تری برای پیش‌بینی مدّنظر قرار گرفته است و آن شاخصی با نام نرخ وصول[1] است؛ به زبان ساده‌تر، صرفاً مشتریان براساس میزان ریسک یا احتمال بازپرداخت‌نکردن وام طبقه‌بندی و رتبه‌بندی نمی‌شوند؛ بلکه از این منظر بررسی می‌شوند که اگر فرضاً احتمال نکول یا بازپرداخت‌نکردن به وجود بیاید، نرخ وصول مطالبات چه میزان خواهد بود؟ و یا به عبارتی، چند درصد از بدهی مشتری وصول می‌شود، در این پژوهش به پیش‌بینی این شاخص توجه شده است. براساس نتایج و الگوی به‌دست‌آمده و اجرای آن بر تسهیلات جاری و مطالبات‌نشده، مشتریان را می‌توان دسته‌بندی کرد و به مشتریان خوش‌حساب یا غیرمطالباتی تسهیلات با نرخ سود کمتر اعطا و مشتریان بدحساب را به‌تدریج، ابتدا از اعطای تسهیلات مجدد جلوگیری و با تدوین برنامهٔ مدونی براساس میزان وثایق دریافتی برای وصول مطالبات آنها اقدام کرد. از این نتایج برای راه‌اندازی سامانهٔ اعطای تسهیلات به‌صورت هوشمند و تسری آن به‌تمامی واحدهای بانک هنگام اعطای تسهیلات و الگوسازی رفتار مشتری جدید براساس ویژگی‌های اعتباری و مقایسه با مشتریان گذشته با ویژگی مشابه می‌توان استفاده کرد.

پرسش اساسی این پژوهش، یافتن پارامترهای بهینهٔ الگوهای پیش‌بینی مجموعهٔ SVR برای پیش‌بینی میزان مطالبات قراردادهای با ریسک زیاد قبل از اعطای تسهیلات است. هدف پژوهش، ارائهٔ الگوی جامعی برای پیش‌بینی درصد میزان وصول مطالبات قراردادهای با ریسک زیاد است و نوآوری پژوهش در مقایسه با پژوهش‌های داخلی در درجهٔ اول پیش‌بینی نرخ وصول مشتریان اعتباری در صورت مطالبات شدن برای اولین بار با استفاده از داده‌کاوی در ایران و همچنین پیش‌بینی مبلغ زیان ناشی از وصول‌نشدن قراردادهای با نرخ وصول کمتر از ۱ به‌جای صرفاً طبقه‌بندی مشتریان به بخش‌های مختلف است که به مورد دوم در پژوهش‌های قبلی در داخل کشور

1 RR (Recovery Rate)

به‌دفعات توجه شده است؛ زیرا در حال حاضر، بانک‌ها در شرایطی مجبور به پرداخت تسهیلات به برخی مشتریان هستند و صرفاً قرارگرفتن مشتری در طبقهٔ پرریسک، ملاک پرداخت نکردن تسهیلات نیست و باید با پیش‌بینی دقیق مشخص شود در صورتی‌که تسهیلات به آنها اعطا شود، در آینده چه میزان از مبلغ بدهی مشتری با گذراندن مراحل قانونی قابل وصول خواهد بود.

مبانی نظری

در مطالعات لوترمن،[1] به الگوریتم‌های مختلف رگرسیون در الگوسازی شاخص درصد مطالبات قابل وصول توجه شده است. ۲۴ روش الگوسازی براساس الگوریتم‌های رگرسیون خطی، لجستیک، کمترین مربعات و ... بررسی و همچنین با بررسی روش‌های دیگر نظیر شبکهٔ عصبی و ماشین بردار تصمیم مشخص شد که این دو روش نسبت به روش‌های سنتی خطی، کارایی بیشتری دارند [۸].با بررسی کاربردهای تحلیل تکنیک‌های ابقا زمانی در الگوسازی شاخص مذکور و روش‌هایی نظیر رگرسیون کاکس، خطی و لجستیک برای پیش‌بینی این شاخص استفاده شده است [۱۵]. با مقایسهٔ الگوهای رگرسیون برای تخمین شاخص نرخ وصول با مقایسهٔ الگوهای مختلف با تأکید بر الگوهای رگرسیونی نظیر رگرسیون ساده، لجستیک، درخت تصمیم و... می‌توان نتیجه گرفت با توجه به نوع داده‌ها هیچ‌کدام از الگوها لزوماً به‌صورت در خور توجهی بهتر از دیگر الگوها نیستند [۲]. اطلاعات حسابداری برای وام‌دهندگان در اقسام قراردادهای بدهی، آنها را در تخصیص مناسب شاخص مذکور به مشتریان می‌تواند مجهز کند. با داشتن اطلاعات حسابداری قراردادهای بدهی مشتریان ۴۷ ماه قبل از نکول، درصد مطالبات غیر قابل وصول را به‌صورت طبیعی قبل از وقوع می‌توان به دست آورد. همچنین مشخص می‌شود افزایش نرخ بهرهٔ وام، رابطهٔ مستقیمی با این شاخص و میزان بدهی نیز ارتباط زیادی دارد [۱]. یااو[2] و همکاران (۲۰۱۵) با استفاده از روش رگرسیون بردار تصمیم به الگوسازی شاخص درصد مطالبات وصول‌ناپذیر گرفتند. در این پژوهش، تکنیک مذکور و ۱۳ الگوریتم دیگر بررسی و درانتها، مشخص شد این تکنیک بسیار اعتمادپذیرتر از سایر روش‌های الگوسازی شاخص مذکور است [۱۶]. در پژوهش گرتلر[3] و هیبلن[4] (۲۰۱۳) بهبود پیش‌بینی شاخص درصد مطالبات غیر قابل وصول بانک بررسی شد. در این مطالعه با استفاده از روش‌هایی نظیر نمونه‌گیری مغرضانه، مشخصه‌های متفاوت وامی با توجه به نوع پایان دورهٔ مطالبات و تنظیمات اطلاعاتی متفاوت براساس وضعیت نکول بررسی شد [۵]. باستوس[5] (۲۰۱۰) برای پیش‌بینی درصد مطالبات غیر قابل وصول وام‌ها در بانک، شاخص مذکور را با استفاده از پیش‌بینی نرخ وصول بررسی و از روش رگرسیون در بخش پارامتریک و درخت تصمیم در بخش نا پارامتریک استفاده کرده است. در نتایج پژوهش نشان داده شده است درخت‌های تصمیم، جایگزین مناسبی برای روش‌های پارامتریک در الگوسازی شاخص درصد مطالبات وصول‌ناپذیر هستند [۳]. حاجی کرد و همکاران (۱۳۹۵) با استفاده از الگوی ماشین بردار تصمیم و الگوی هیبریدی الگوریتم ژنتیک برای پیش‌بینی ریسک اعتباری و تقسیم‌بندی مشتریان به دو دستهٔ خوش‌حساب و بدحساب استفاده کرده‌اند که

1 Loterman, G

2 Yao, X
3 Gürtler, M
4 Hibbeln, M
3 Bastos, J A

نتیجه نشان داد الگوی بهینه‌سازی‌شدهٔ ماشین بردار تصمیم با الگوریتم ژنتیک، تأثیر بهتری در پیش‌بینی ریسک اعتباری و دسته‌بندی مشتریان به خوش‌حساب و بدحساب دارد [۹].کرانی و آقایی‌پور (۱۳۹۳) نظریهٔ تحلیل بقا در مدیریت ریسک اعتباری دریافت‌کنندگان تسهیلات (مطالعهٔ موردی: بانک مسکن) را بررسی و احتمال‌های نکول آن را براساس الگوی خطرهای متناسب کاکس و برآوردگر حد حاصل‌ضربی تعمیم‌یافته برآورد کرده‌اند [۷]. نظرپور و رضایی (۱۳۹۲) عقود اسلامی و الگوی پرداخت تسهیلات را بررسی کردند و دریافتند عقود اسلامی غیرمشارکتی (مبادله‌ای) در مقایسه با عقود مشارکتی، ریسک کمتری در بردارد؛ اما این عقود نیز ریسک است و بانک‌های اسلامی را در معرض ریسک اعتباری قرار می‌دهد و بانک‌ها در صورت استقرارنداشتن یک نظام مدیریت ریسک اعتباری متناسب، میزان ریسک اعتبارات و تعیین زیان‌های احتمالی بازپرداخت‌نکردن وام‌ها را نمی‌توانند تشخیص بدهند و درنتیجه، سرمایهٔ خود را نخواهند توانست تخصیص بهینه کنند [۱۱].

روش پژوهش

برای پیش‌بینی و الگوسازی شاخص نرخ وصول، علاوه بر مشخصات اعتباری مشتریان، اطلاعات متغیرهای اقتصادی نیز در الگوسازی استفاده و برای محاسبهٔ این شاخص، از مشتریانی در الگوسازی استفاده می‌شود که تسهیلات دریافتی آنها به طبقهٔ مطالبات غیرجاری (سررسید گذشته/ معوق/ مشکوک‌الوصول) منتقل شده است.

$$RR = \frac{DNR}{EAD} = \frac{FR}{EAD} \times \frac{FR - AC}{FR} \times (1+r)^{-T} \qquad (۱)$$

RR نرخ وصول تعهدات مطالبات‌شده، DNR ارزش تنزیل خالص وصول که به خالص تمام هزینه‌های وصول مطالبات گفته می‌شود. EAD میزان بدهی در زمان ایجاد مطالبات. FR ارزش اسمی میزان وصول در دوره (در این پژوهش ارزش وثیقه‌های دریافتی به جای میزان وصولی در دورهٔ واقعی در نظر گرفته شده است. AC هزینه‌های اداری مرتبط با فرایند وصول تعهدات مطالبات‌شده. r نرخ تنزیل و T زمان فرایند وصول مطالبات [۱۲]. با توجه به اطلاعات اعتباری موجود و مشخص‌نبودن هزینه‌های اداری، نرخ تنزیل و زمان فرایند وصول برای تک‌تک مشتریان و همچنین فرایند وصول مطالبات که بعضاً به‌دلیل بوروکراسی اداری، مکاتبات واحدهای حقوقی، تشکیل دادگاه در کشور ایران که کاملاً متغیر و بعضاً غیر قابل اندازه‌گیری است، فرمول مذکور با توجه به شرایط بانک‌های کشور و اطلاعات موجود بومی‌سازی می‌شود. در این پژوهش، ارزش اسمی میزان وصول به دست آمده است که از اندازه‌گیری ارزش مبلغی وثیقه‌های دریافتی با ضرایب اعلامی از سوی بانک مرکزی ج.ا.ا براساس (دستورالعمل نحوهٔ محاسبهٔ ذخیرهٔ مطالبات، ۱۳۹۰) اعلام شده است و به‌عنوان DNR در نظر گرفته می‌شود. در حالت کلی برای محاسبهٔ نرخ وصول واقعی باید میزان وصول واقعی اتفاق‌افتاده در دوره را جایگزین کرد؛ بنابراین فرمول محاسبهٔ نرخ واقعی وصول تعهدات مطالبات شده به‌صورت رابطهٔ (۲) و (۳) است:

$$RR = \frac{DNR}{EAD} = \frac{FR}{EAD} = \frac{\text{ارزش اسمی مطلوب}}{\text{میزان بدهی حال مشتری}} \qquad (۲)$$

$$\text{ارزش وثایق مطلوب} = \sum_{i=1}^{2} (W_i \times FR_i) \qquad (۳)$$

که در آن W_i ضریب ارزش وثیقه i و FR_i ارزش اسمی مطلوب است. در پژوهش حاضر در بانک مدّنظر، دو نوع وثیقهٔ i=1,2 با ارزش متفاوت با مشتری گرفته می‌شود در جدول (۱) لیست وثیقه‌ها به همراه ارزش وزنی آنها آمده است. ارزش وزنی سایر وثیقه‌ها (سفته و برات، قرارداد، چک، ضامن معتبر و امضای مدیران) صفر است (توجه شود در حالت کلی، صورت کسر، میزان وصولی در دوره است که در این پژوهش، ارزش وثیقه‌های دریافتی در نظر گرفته شده است)

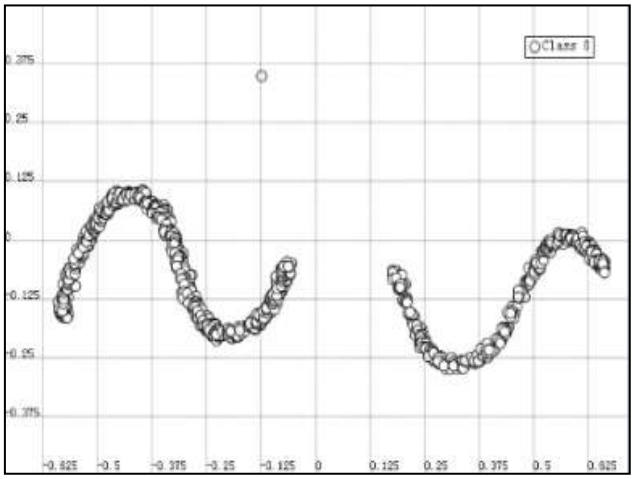
(۴) میزان بدهی حال مشتری = سایر بدهی‌ها+ (سود آینده- اصل مبلغ وام)

(۵) سایر بدهی‌ها = ∑(معوق + سررسید گذشته + سود دریافتنی + مشکوک الوصول + سود دریافتنی مطالبات + وجه التزام دریافتنی)

ارزش وثیقه‌های مندرج در بانک اطلاعات مشتریان اعتباری با توجه به ضرایب اعلامی بانک مرکزی ج.ا.ا براساس (دستورالعمل نحوهٔ محاسبهٔ ذخیرهٔ مطالبات، ۱۳۹۰) غیرمنقول ۷۰ درصد، سپرده (قرض‌الحسنه پس‌انداز، سرمایه‌گذاری مدت‌دار و ...) ۱۰۰، ضمانت صادرات، سهام بورس، غیرمنقول از محل اجرای طرح/ خارج از محل، غیرمنقول ملکی، غیرمنقول کارخانه همگی ۷۰ و غیرمنقول ماشین‌آلات و تجهیزات و کالا ۵۰ درصد است. الگوریتم استفاده‌شده، روش ماشین بردار تصمیم است. روش ماشین بردار تصمیم به دو صورت مبتنی بر طبقه‌بندی و مبتنی بر پیش‌بینی است. در این پژوهش، پیش‌بینی با روش‌های مبتنی بر رگرسیون انجام می‌گیرد. علت این موضوع، پیوستگی متغیر هدف (نرخ وصول) است و شامل زیرمجموعه‌ای از الگوهای پیش‌بینی است که شامل الگوهای مختلف - εSVR،

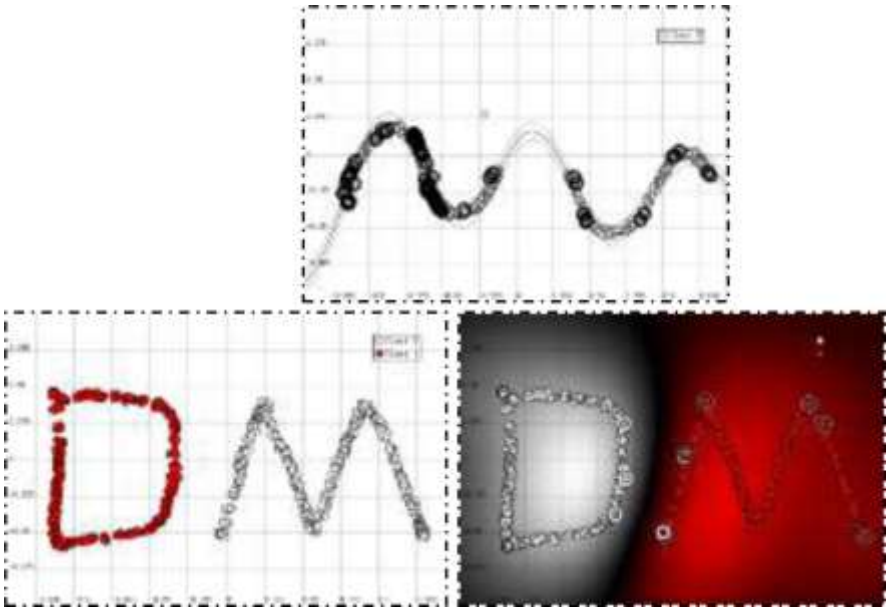
υ-SVR است؛ با این تفاوت که کلیّهٔ این الگوها با روش الگوریتم ژنتیک برای انتخاب متغیرهای الگو، بهینه‌سازی و درنهایت، میزان خطا و کارایی آنها مقایسه می‌شود. مبنای الگوسازی در این پژوهش [۴] و [۱۳] است.

روش رگرسیون بردار تصمیم (مبتنی بر پیش‌بینی یا تخمین عددی) برای اهداف پیوسته در داده‌کاوی است.



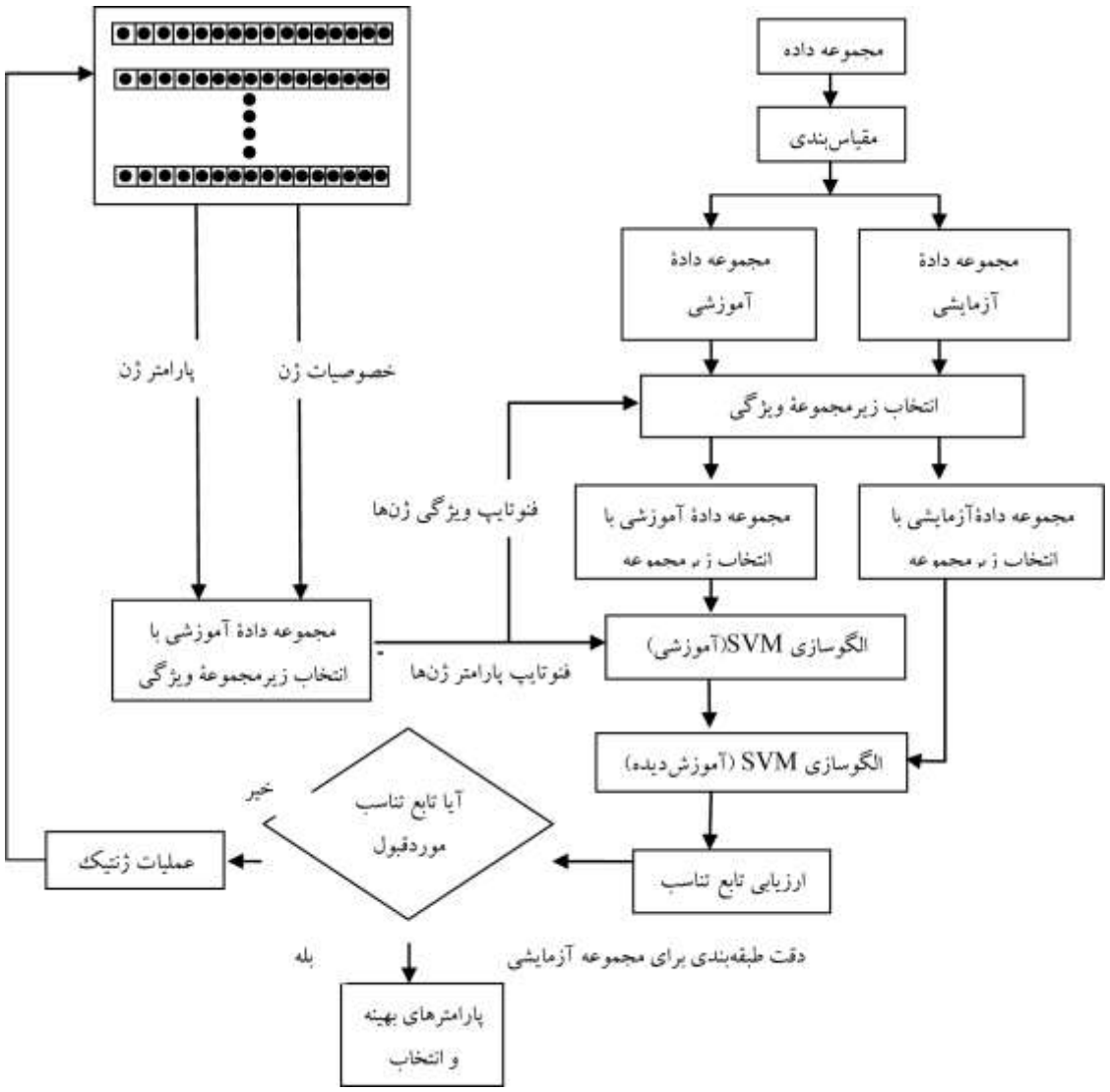
شکل (۱) الگوسازی به روش SVR

روش الگوسازی طبقه‌بندی بردار تصمیم (مبتنی بر طبقه‌بندی) برای پیش‌بینی اهداف گسسته و یا عدد صحیح به کار گرفته می‌شود. برای پیش‌بینی مشتریان بد و خوب با تفکیک هدف از این الگو برای پیش‌بینی می‌توان استفاده کرد.



شکل (۲) الگوسازی به روش SVC

در روش بردار تصمیم، پارامترهای الگو نظیر C,υ,ε و γ قابل تنظیم است که قبل از الگوسازی با انتخاب یک عدد خاص، الگو بر داده‌های آموزشی فرا گرفته شود که با استفاده از روش الگوریتم ژنتیک، انتخاب این پارامترها را با رویکرد کم‌کردن خطای الگو در مقایسه با داده‌های تستی به‌طور بهینه می‌توان انتخاب کرد؛ به عبارتی، طرح کروموزوم، تابع تناسب و نوع معماری سیستم برای انتخاب ویژگی مبتنی بر الگوریتم ژنتیک و بهینه‌سازی پارامترها است. الگوریتم کلی که در تمامی قسمت‌ها تقریباً با اندکی تغییر برای الگوسازی استفاده‌شده است، به‌صورت شکل (۳) نشان داده شده است.



شکل (۳) معماری سیستم پیشنهادی برای انتخاب ویژگی و پارامترهای بهینه با استفاده از الگوریتم ژنتیک

فرایند داده‌کاوی و الگوسازی بر داده‌های یکی از بانک‌های دولتی در فاصلهٔ مهر تا شهریورماه سال‌های ۱۳۸۹-۱۳۹۴ به‌صورت ماهانه به تعداد ۶۰ مقطع انجام شده است. حجم جامعهٔ آماری مدّنظر، تعداد ۲۵،۰۱۷،۰۳۶ قرارداد است. نمونهٔ آماری برای آموزش الگو، نمونه‌ای به حجم ۵۰۰ قرارداد از جامعهٔ مدّنظر براساس روش نمونه‌گیری تصادفی ساده بدون جایگزینی بر کلیّهٔ قراردادهای مذکور انجام شده است و برای آزمایش دقت داده‌ها در مرحلهٔ بررسی دقت الگو از نمونه‌ای با حجم ۲۰۰۰ و با روش نمونه‌گیری تصادفی ساده بدون جایگزینی استفاده شده است. برای هر یک از الگوها نیز نمونه‌گیری جدید انجام شده است.

جدول (۱) متغیرهای استفاده‌شده در الگوسازی نرخ وصول

ردیف	شرح متغیر	نام متغیر
۱	کد منطقه	TR_ADMIN
۲	کد شعبه	TR_BR
۳	کد نوع قرارداد	C_noegharardad
۴	سال قرارداد	Year_gharardad
۵	کد رشته فعالیت	C_reshtefaaliat
۶	مبلغ پرداختنی	Mablagh_pardakhtani
۷	مبلغ بازگشتی	Mablagh_bazgashti
۸	سود سهم مشتری	Soodsahmmoshtari
۹	سود سهم دولت	Soodsahmdolat
۱۰	نرخ سود سهم مشتری	Nerkhsoodsahmmoshtari
۱۱	نرخ سود سهم دولت	Nerkhsood_sahmdolat
۱۲	درصد مشتری	Darsadmoshtari
۱۳	تعداد اقساط	Tedadaghsat
۱۴	کد نوع رکورد	C_noerecord
۱۵	کد بودجه	C_budget
۱۶	کد هدف	C_hadaf
۱۷	مبلغ مصوب	Mablagh_mosavab
۱۸	کد ادارهٔ ابلاغ کننده	C_edareeblaghconandeh
۱۹	کد مصوبه	C_mosavabe
۲۰	کد تکلیف	C_taklif
۲۱	کد مدّنظر	C_moredmasraf
۲۲	روش تقسیط	Ravesh_taghseet
۲۳	کد شهر مدّنظر	C_shahrmoredmasraf
۲۴	جنسیت	Gender
۲۵	نرخ تورم	Inflation
۲۶	تولید ناخالص داخلی به قیمت پایه	GDP
۲۷	ارزش اسمی وثیقه‌ها	Arzeshvasaiegh

با توجه به اینکه برای الگوسازی در داده‌کاوی، روش‌های مختلفی وجود دارد؛ نظیر داده‌کاوی نظارت‌شده و غیرنظارت‌شده، برای الگوسازی، سناریوهای مختلفی براساس الگوهای مذکور در بخش قبل در نظر گرفته می‌شود.

جدول (۲) الگوهای زیرمجموعهٔ ماشین بردار تصمیم به تفکیک سناریوهای استفاده شده

روش داده‌کاوی	یک مرحله‌ای	دومرحله‌ای	سه مرحله‌ای
نظارت‌شده	✓	✓	-
غیرنظارت‌شده	-	✓	-
نظارت‌شده + غیرنظارت‌شده	-	-	✓

در روش‌های الگوسازی، روش یک مرحله‌ای مستقیماً نرخ وصول مشتریان را پیش‌بینی می‌کند؛ اما روش‌های دومرحله‌ای و سه مرحله‌ای، ابتدا طبقهٔ ریسک را با توجه به طبقه‌بندی با استفاده از درخت تصمیم مشخص و سپس طبقه با ریسک بالا و کم از الگوسازی حذف و طبقه با ریسک متوسط پیش‌بینی استفاده می‌شود. در روش داده‌کاوی نظارت‌شدهٔ یک مرحله‌ای به‌صورت مستقیم به الگوسازی نرخ وصول مشتریان با استفاده از روش‌های بخش قبل توجه می‌شود. الگوهای استفاده‌شده در این سناریو GA-υ-SVR و GA-ε-SVR است. در روش داده‌کاوی نظارت‌شدهٔ دومرحله‌ای، حالت اول بخش الگوسازی به‌صورت دومرحله‌ای انجام می‌گیرد؛ یعنی ابتدا، متغیر هدف (نرخ وصول) به سه‌طبقه تقسیم می‌شود. نرخ وصول با مقدار ۱ (احتمال وصول ۱۰۰ درصد)، نرخ وصول با مقدار بین ۰ و ۱، نرخ وصول با مقدار ۰ (احتمال وصول صفر). طبقات بالا به‌ترتیب، با ستون وضعیت (State) در فایل مشتریان با کدهای ۲،۱ و ۳ طبقه‌بندی می‌شوند.

جدول (۳) طبقه‌بندی وضعیت نرخ وصول مشتریان

شرح	طبقه	برچسب شرح طبقه
نرخ وصول ۱	۱	LOW
نرخ وصول بین ۰ و ۱	۲	MEDIUM
نرخ وصول ۰	۳	HIGH

الگوهای استفاده‌شده در این سناریو C5.0 +GA-ε-SVR و C5.0 +GA-υ-SVR است.

در روش داده‌کاوی غیرنظارت‌شدهٔ دومرحله‌ای حالت دوم، ابتدا داده‌ها خوشه‌بندی می‌شوند و سپس بر اساس خروجی هر خوشه، نمونه‌گیری بر هر خوشه و پیش‌بینی انجام می‌گیرد. روش یک مرحله‌ای در این سناریو بی‌معنی است؛ زیرا خوشه‌بندی به‌صورت خودکار، مرحلهٔ اول را انجام می‌دهد (خوشه‌بندی+پیش‌بینی). الگوهای استفاده‌شده در این بخش عبارت است از K-MEANS+GA-υ-SVR و

K-MEANS+ GA-ε-SVR. در روش داده‌کاوی نظارت‌شده و غیرنظارت‌شدهٔ سه مرحله‌ای، ابتدا داده‌ها برچسب‌گذاری و سپس براساس سناریوی الگوسازی براساس روش داده‌کاوی غیرنظارت‌شده خوشه‌بندی می‌شوند و سپس براساس خروجی هر خوشه، طبقه‌بندی انجام می‌شود و درنهایت، پیش‌بینی بر جامعهٔ طبقه‌بندی‌شده انجام می‌شود. الگوسازی در این سناریو به‌صورت سه مرحله‌ای است (خوشه‌بندی+طبقه‌بندی+پیش‌بینی). الگوهای استفاده‌شده در این سناریو به‌شرح K-MEANS+C5.0 +GA-ε-SVR و K-MEANS+C5.0 +GA-υ-SVR است.

یافته‌ها

درادامه، به‌صورت تفکیک‌شده، نتایج هر سناریو به تفکیک الگو آمده است.

جدول (۴) نتایج الگوی GA-υ-SVR – داده‌کاوی نظارت‌شدهٔ یک‌مرحله‌ای

شرح پارامتر	مقدار	مقدار بهینه
بازه تغییرات C	۰/۰۰۰۱–۱۰	۶/۶۰۱۷۱۵
بازه تغییرات γ	۰/۰۰۱–۲	۰/۰۲۲۶۳۷۰۲
بازه تغییرات υ	۰–۱	۰/۳۹۳۲۷۷۸
تعداد تکرار الگوریتم	۱۰	-
حجم جمعیت	۴۰	-
احتمال جهش	۰/۹	-
احتمال عبور	۰/۱	-
مقدار تابع تناسب	-۰/۰۳۵۹۲۱۵۷	-

جدول (۵) نتایج الگوی GA-ε-SVR – داده‌کاوی نظارت‌شدهٔ یک‌مرحله‌ای

شرح پارامتر	مقدار	مقدار بهینه
بازه تغییرات C	۰/۰۰۰۱–۱۰	۵/۹۹۳۲۳۹
بازه تغییرات γ	۰/۰۰۱–۲	۰/۰۲۷۶۹۸۱۷
بازه تغییرات ε	۰/۰۱–۲	۰/۱۷۷۱۲۸۸
تعداد تکرار الگوریتم	۱۰	-
حجم جمعیت	۵۰۰	-
احتمال جهش	۰/۸	-
احتمال عبور	۰/۰۵	-
مقدار تابع تناسب	-۰/۰۳۵۶۳۵۴۸	-

جدول (۶) نتایج الگوی GA-ε-SVR – داده‌کاوی نظارت‌شدهٔ دومرحله‌ای

شرح پارامتر	مقدار	مقدار بهینه
بازه تغییرات C	۰/۰۰۰۱–۱۰	۵/۱۵۵۱۲۶
بازه تغییرات γ	۰/۰۰۱–۲	۰/۰۱۵۸۰۵۸
بازه تغییرات ε	۰/۰۱–۲	۰/۵۶۱۸۸۰۴
تعداد تکرار الگوریتم	۱۰	-
حجم جمعیت	۵۰۰	-
احتمال جهش	۰/۸	-
احتمال عبور	۰/۰۵	-
مقدار تابع تناسب	-۰/۰۲۶۱۵۴۶۸	-

جدول (۷) نتایج الگوی GA-υ-SVR – داده‌کاوی نظارت‌شدهٔ دومرحله‌ای

شرح پارامتر	مقدار	مقدار بهینه
بازه تغییرات C	۰/۰۰۰۱–۱۰	۸/۸۵۸۹۷۶
بازه تغییرات γ	۰/۰۰۱–۲	۰/۰۰۲۷۲۸۷۳
بازه تغییرات υ	۰–۱	۰/۳۴۶۹۳۹۹
تعداد تکرار الگوریتم	۱۰	-
حجم جمعیت	۴۰	-
احتمال جهش	۰/۹	-
احتمال عبور	۰/۱	-
مقدار تابع تناسب	-۰/۰۲۵۵۰۶۴۱	-

جدول (۸) ستون‌های استفاده‌شده در فرایند خوشه‌بندی (نرمال‌سازی‌شده)

شرح	نام متغیر	وضعیت
ارزش اسمی وثیقه‌ها	Arzeshvasaiegh_Transformed	نرمال‌شده
کد نوع قرارداد (عقد تسهیلاتی)	C_noegharardad	نرمال‌شده
کد رشته فعالیت (بخش اقتصادی)	C_reshtefaaliat	نرمال‌شده
مبلغ تسهیلات پرداختنی	Mablagh_Pardakhtani_Transformed	نرمال‌شده
مبلغ تسهیلات بازگشتی	Mablagh_Bazgashti_Transformed	نرمال‌شده

جدول (۹) نتایج خوشه‌بندی براساس تقسیم‌بندی مجموعه دادهٔ آموزشی و آزمایشی

بخش	مجموعه دادهٔ آموزشی	مجموعه دادهٔ آزمایشی	جمع کل
بخش اول	۱۴/۱۲۱/۸۶۴	۷/۶۰۶/۶۸۴	۲۱/۷۲۸/۵۴۸
بخش دوم	۲/۱۳۶/۶۷۵	۱/۱۵۱/۸۰۷	۳/۲۸۸/۴۸۲
جمع کل	۱۶/۲۵۸/۵۳۹	۸/۷۵۸/۴۹۱	۲۵/۰۱۷/۰۳۰
سهم درصد از کل	%۶۵	%۳۵	

جدول (۱۰) نتایج خوشه‌بندی براساس تقسیم‌بندی مجموعه دادهٔ آموزشی و آزمایشی

تکرار	میزان خطا
۱	۰/۵۶۷
۲	۰/۵۷
۳	۰/۳۳۵
۴	۰/۰۹۶
۵	۰/۰
۶	۰/۰

کل مجموعهٔ هر بخش با توجه به اینکه به دو بخش تقسیم شده است، در بخش داده‌های آموزشی آموزش‌دیده و پس از کنترل بر داده‌های آزمایشی در صورت کمترین خطا انتخاب می‌شود.

جدول (۱۱) نتایج خوشه‌بندی حجم خوشه‌ها و نسبت بزرگ‌ترین خوشه به کوچک‌ترین خوشه

شرح	مقدار	درصد
حجم کمترین خوشه	۲/۱۳۶/۶۷۵	٪۱۳/۱
حجم بیشترین خوشه	۱۴/۱۲۱/۸۳۴	٪۸۶/۹
نسبت بیشترین خوشه به کمترین خوشه	۶/۶۱	

جدول (۱۲) نتایج اجرای الگوی GA-ε-SVR بر مبنای انتخاب پارامترها با روش الگوریتم ژنتیک براساس روش داده‌کاوی غیرنظارت‌شده و پیش‌بینی نظارت‌شده

شرح پارامتر	مقدار (خوشهٔ اول)	مقدار بهینه (خوشهٔ اول)	مقدار (خوشهٔ دوم)	مقدار بهینه (خوشهٔ دوم)
بازه تغییرات C	۰/۰۰۰۱–۱۰	۲/۸۴۰۰۴۷	پاسخی از الگوی بهینه‌سازی در هیچ حالتی دریافت نمی‌شود و الگو، جواب بهینه ندارد.	
بازه تغییرات γ	۰/۰۰۱–۲	۰/۰۷۵۱۶۹۷۱		
بازه تغییرات ε	۰/۰۱–۲	۰/۰۱۶۳۷۸۷۳		
تعداد تکرار الگوریتم	۵	-		
حجم جمعیت	۵۰	-		
احتمال جهش	۰/۸	-		
احتمال عبور	۰/۱	-		
مقدار تابع تناسب	-۰/۰۳۲۹۵۰۷۸	-		

جدول (۱۳) نتایج اجرای الگوی GA-υ-SVR بر مبنای انتخاب پارامترها با روش الگوریتم ژنتیک براساس روش داده‌کاوی غیرنظارت‌شده و پیش‌بینی نظارت‌شده

شرح پارامتر	مقدار (خوشهٔ اول)	مقدار بهینه (خوشهٔ اول)	مقدار (خوشهٔ دوم)	مقدار بهینه (خوشهٔ دوم)
بازه تغییرات C	۰/۰۰۰۱–۱۰	۳/۳۱۵۰۳۱	۰/۰۰۰۱–۱۰	۹/۰۲۹۳۶۷
بازه تغییرات γ	۰/۰۰۱–۲	۰/۰۳۹۵۱۵۳۷	۰/۰۰۱–۲	۰/۰۱۷۹۲۴۴۳
بازه تغییرات υ	۰–۱	۰/۶۸۰۵۱۵۹	۰–۱	۰/۲۱۵۴۸۷
تعداد تکرار الگوریتم	۱۰	–	۱۰	–
حجم جمعیت	۵۰	–	۵۰	–
احتمال جهش	۰/۸	–	۰/۸	–
احتمال عبور	۰/۱	–	۰/۱	–
مقدار تابع تناسب	-۰/۰۱۷۸۰۸۸۹	–	-۰/۰۶۰۹۱۸۷۱	–

جدول (۱۴) نتایج اجرای الگوی GA-ε-SVR براساس روش داده‌کاوی نظارت‌شده و غیرنظارت‌شدهٔ سه مرحله‌ای

شرح پارامتر	مقدار (خوشهٔ اول)	مقدار بهینه (خوشهٔ اول)	مقدار (خوشهٔ دوم)	مقدار بهینه (خوشهٔ دوم)
بازه تغییرات C	۰/۰۰۰۱–۱۰	۸/۶۰۲۲۱۴	۰/۰۰۰۱–۱۰	۴/۴۴۹۲۹۷
بازه تغییرات γ	۰/۰۰۱–۲	۰,۰۰۵۹۵۰۷۰۶	۰/۰۰۱–۲	۰/۰۱۰۵۶۴۴۱
بازه تغییرات ε	۰/۰۱–۲	۰,۰۸۳۵۹۷۲۵	۰/۰۱–۲	۰/۱۵۲۷۱۸
تعداد تکرار الگوریتم	۱۰	–	۱۰	–
حجم جمعیت	۵۰۰	–	۵۰۰	–
احتمال جهش	۰/۸	–	۰,۸	–
احتمال عبور	۰/۰۵	–	۰,۰۵	–
مقدار تابع تناسب	–۰/۰۳۰۴۹۵۲۸	–	–۰/۰۳۳۲۳۲۴۳	–

جدول (۱۵) نتایج اجرای الگوی GA-υ-SVR براساس روش داده‌کاوی نظارت‌شده و غیرنظارت‌شدهٔ سه مرحله‌ای

شرح پارامتر	مقدار (خوشهٔ اول)	مقدار بهینه (خوشهٔ اول)	مقدار (خوشهٔ دوم)	مقدار بهینه (خوشهٔ دوم)
بازه تغییرات C	۰/۰۰۰۱–۱۰	۲/۲۳۱۸۲۸	۰/۰۰۰۱–۱۰	۳/۴۸۷۶۷۱
بازه تغییرات γ	۰/۰۰۱–۲	۰/۱۱۶۹۸۶۱	۰/۰۰۱–۲	۰/۰۲۶۲۰۴۵۷
بازه تغییرات υ	۰–۱	۰/۲۸۸۰۸۶۹	۰–۱	۰/۸۳۷۹۷۱۲
تعداد تکرار الگوریتم	۱۰	–	۱۰	–
حجم جمعیت	۴۰	–	۴۰	–
احتمال جهش	۰/۹	–	۰/۹	–
احتمال عبور	۰/۱	–	۰/۱	–
مقدار تابع تناسب	–۰/۰۳۰۸۹۳۷۶	–	–۰/۰۳۳۴۷۲۵	–

نتایج و پیشنهادها

برای راحتی فراخوانی الگوها به‌صورت جدول (۱۶)کدبندی می‌شود.

جدول (۱۶) کدبندی الگوهای استفاده‌شده

کد الگو	نام الگو
M1	ε-SVR
M2	υ-SVR
M3	C5.0 +GA-ε-SVR
M4	C5.0 +GA-υ-SVR
M5	K-MEANS+GA-ε-SVR
M6	K-MEANS+GA-υ-SVR
M7	K-MEANS+C5.0+GA-ε-SVR
M8	K-MEANS+C5.0+GA-υ-SVR

در این بخش، ابتدا جامعهٔ اصلی مشتریان دارای مطالبات غیرجاری (بدهکار به بانک) به دو بخش آموزشی و آزمایشی تقسیم می‌شود؛ سپس تمامی الگوهای به‌دست‌آمده با تنظیمات بهینه بر نمونه‌ای تصادفی با حجم ۲۰۰۰ مشاهده از جامعهٔ اصلی (مجموعه داده آزمایشی مربوط به خود) اجرا و نتایج بررسی و خطاها مقایسه می‌شود. برای الگوهای M1 و M2 مجموعه دادهٔ آزمایشی از کل جامعه انتخاب می‌شود و نیازی به تفکیک نیست. با توجه به نتایج به‌دست‌آمده، رتبه‌بندی الگوهای منتخب نهایی براساس کمترین خطاها و همچنین بیشترین پوشش سطح زیر منحنی به‌شرح جدول (۱۷) است.

جدول (۱۷) الگوهای منتخب براساس رتبه‌بندی کمترین میزان خطا

رتبه	کد الگو	MAE	MSE	RMSE
۱	M1	۰/۰۷۲۴۴۲۳۲	۰/۰۳۲۰۱۴۷۹	۰/۱۷۸۹۲۶۸
۲	M6	۰/۰۷۲۰۷۱۷	۰/۰۳۵۸۱۶۳	۰/۱۸۷۳۵۰۱
۳	M2	۰/۰۵۶۲۰۰۸۱	۰/۰۳۹۷۷۴۶۹	۰/۱۹۹۴۳۵۹

با توجه به نتایج به‌دست‌آمده براساس اجرای الگوهای بهینه‌شده با استفاده از الگوریتم ژنتیک مشخص می‌شود الگوی ε-SVR بهترین الگو برای پیش‌بینی نرخ وصول مطالبات در روش یک‌مرحله‌ای بدون پیش‌بینی طبقهٔ ریسک مشتریان است و در روش دومرحله‌ای، الگوی K-MEANS+GA-υ-SVR بهترین الگو برای پیش‌بینی نرخ وصول مطالبات است. پیشنهاد می‌شود در حوزهٔ بانکداری، بانک‌ها به راه‌اندازی سامانهٔ جامع اعطای تسهیلات بر مبنای تحلیل اطلاعات مشتریان با استفاده از الگوریتم‌های داده‌کاوی به‌ویژه الگوهای پیشنهادی در پژوهش حاضر اقدام کنند و قبل از اعطای تسهیلات شعبه با واردکردن اطلاعات لازم مشتری و مقایسه و الگوسازی با داده‌های گذشته و بانک اطلاعاتی بانک و یا حتی به‌طور کامل‌تر با جامعهٔ سیستم بانکی کشور با تأمل و بررسی بیشتری به اعطای تسهیلات اقدام کنند. بانک‌ها و مؤسسات مالی با روش مذکور در پژوهش حاضر، علاوه بر پیش‌بینی احتمال مطالبات شدن هر مشتری و در صورت مطالبات شدن پیش‌بینی میزان وصول قرارداد مطالبات شده، میزان زیان بانک ناشی از وصول‌نشدن قراردادهای اعتباری را

می‌توانند به‌پیش‌بینی و برآورد کنند. راه‌اندازی سامانهٔ هوشمند اعطای تسهیلات برای الگوسازی با روش‌های داده‌کاوی قبل از اعطای تسهیلات می‌تواند اقدامی در راستای کنترل انضباط اعتباری براساس الزامات بانک مرکزی ج.ا.ا و کمیتهٔ بال باشد.

در این پژوهش با توجه به محدودیت‌های موجود، محاسبهٔ نرخ وصول به‌صورت کامل انجام نشد و برای الگوسازی به‌دلیل مشخص‌نبودن میزان وصول واقعی از ارزش مبلغی وثیقه‌ها برای الگوسازی استفاده و صرفاً بر روش الگوسازی تأکید شد. پیشنهاد می‌شود در پژوهش‌های آینده با استفاده از اطلاعات کامل‌تر نظیر زمان فرایند وصول مطالبات، نرخ تنزیل، هزینه‌های وصول مطالبات هر قرارداد نرخ وصول محاسبه شود که طبیعتاً به عدد صفر نزدیک‌تر است. همچنین از روش‌های دیگر الگوسازی نیز استفاده شود.

منابع

[1] Amiram, D. (2011). Debt contracts and loss given default. *Job Market Paper*. University of North Carolina-Chapel Hill.

[2] Arsova, A. Haralampieva, M. & Tsvetanova, T. (2011). Comparison of regression models for LGD estimation. *Credit Scoring and Credit Control XII Edinburgh*: Experian Limited:1-23

[3] Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance*, Vol. 34 (10): 2510-2517.

[4] Chen, K.Y. & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*. (28): 215-226

[5] Gürtler, M. & Hibbeln, M. (2013). Improvements in loss given default forecasts for bank loans. *Journal of Banking & Finance*. (37): 2354-2366.

[6] Huang, C. L. & Wang, C. J. (2006). A GA-based feature selection and parameters optimization. *Expert Systems with Applications*. (31): 231–240.

[7] Karani, H. & Aghaei Pour, M. (2014). Application of the theory of survival analysis of credit risk management loan recipients. *Ravand Quarterly* (21):175-200

[8] Loterman, G. (2013). *Predicting Loss Given Default* PHD Thesis: Ghent University.

[9] Mohammadian H. K. A., Asgharzadeh Z. M. & Emam D. M. (2016). Credit risk assessment of corporate customers using support vector machine and genetic algorithm hybrid model - A case study of Tejarat Bank. *Financial Engineering & Portfolio Management*. (7): 17-32.

[10] Moin, K. & Baseer A. D. (2012). Use of data mining in banking. *International Journal of Engineering Research and Applications (IJERA*). (2): 738-742.

[11] Nazarpour, M. T. & Rezaei, A. (2013). Credit risk management in Islamic banking with approach to review contracts and loan payment pattern. *Islamic Financial Research*. (2):123-156.

[12] Resti, A. & Sironi, A. (2007). *Risk Management and Shareholders' Value in Banking*. England: John Wiley & Sons.

[13] Sermpinis, G. Stasinakis, C. & Theofilatos, K. (2015). Modeling, forecasting and trading the EUR exchange rates with hybrid rolling genetic algorithms support vector regression forecast combinations. *European Journal of Operational Research*. (247): 831-846.

[14] Shahrabi, J., Hadavandi, E. (2011). *Data Mining In Banking.*Tehran: Iranian Academic Center for Education Culture & Research.

[15] Witzany, J. Rychnovsky, M. & Charamza, P. (2012). Survival analysis in LGD modeling. *European Financial and Accounting Journal*. (7): 6-27.

[16] Yao, X. Crook, J. & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*. (240): 528-538.