Fraud Detection in Banking Electronic Payment Using Data Mining

Marzie Hosseini

IT Expert at Tourism Bank

Ma.hosseini@tourismbank.ir

Abstract

Fraud continuing variation is one of the most important Fraud Detection challenge in Electronic Payment Systems. Electronic markets need high Performance Fraud Detection methods with top accuracy. In this project we have used logistic regression, BP neural network and GMDH neural network to make a useful method detecting Fraud in a bank financial transactions. We have implemented these methods on real dataset and have measured their results. The results are GMDH neural network ۹۱,۷۳٪, logistic regression ۸۹,۲٦٪ and BP neural network ۹۰,٥٤٪. Therefore our recommended method has the best accuracy.

Keywords: Fraud, Electronic Payment Systems, Financial transactions, Data Mining, GMDH Neural Network

Archive of SID

www.SID.ir

تشخیص تقلب درسیستم‌های پرداخت الکترونیکی بانکها با استفاده از داده‌کاوی

مرضیه حسینی

کارشناس فناوری اطلاعات بانک گردشگری

Ma.hosseini@tourismbank.ir

Archive of SID

چکیده

یکی از چالش‌های تشخیص تقلب در حوزه سیستم های پرداخت الکترونیکی، تنوع و تغییر مداوم شیوه‌های تقلب است لذا نیاز به روش های تشخیص تقلب با کارایی و دقت بالا به روشنی قابل درک است. در این پژوهش روش داده کاوی رگرسیون لجستیک، شبکه عصبی BP و شبکه عصبی GMDH برای ساخت مدل‌هایی جهت شناسایی تقلب در تراکنش‌های مالی دستگاه خودپرداز یک بانک پیاده سازی شدند. در ادامه، این روش‌ها برروی داده‌های واقعی آزمایش و کارایی هر روش سنجیده شد. روش شبکه عصبی GMDH با دقت ۹۱,۷۳ درصد در شناسایی تقلب یا غیرتقلب بودن تراکنش‌های مالی بهترین کارایی را در مقایسه با دو روش رگرسیون لجستیک با دقت کلی ۸۹,۲۶ و شبکه عصبی BP با دقت کلی ۹۰,۵۴داشت. باتوجه به نتایج بدست آمده روش پیشنهادی در تشخیص تقلب نسبت به دو روش دیگر با دقت بیشتری عمل کرده است.

واژه های کلیدی: تقلب، سیستم های پرداخت الکترونیکی، تراکنش های مالی، داده کاوی، شبکه عصبی GMDH

مقدمه

www.SID.ir

طی سال‌های اخیر تجارت الکترونیک رشد قابل توجهی داشته، بطوریکه امروزه به عنوان یکی از کانال‌های رایج برای کسب و کارها مطرح است. علت این امر را باید در گسترش روزافزون اینترنت و سهولت استفاده از این کانال در بسیاری از موارد جستجو کرد. اما همان‌طور که فرایند پرداخت جز لاینفک هر کسب و کاری است، پرداخت الکترونیکی نیز از اجزای لاینفک تجارت الکترونیکی است، همان‌گونه که بازار قابل توجهی برای فروشندگان است، منبع درآمد خوبی برای سوءاستفاده متقلبین نیز بوده است.

امروزه روش‌های متفاوتی برای کشف تقلب، بویژه در حوزه بانکداری الکترونیک صورت گرفته است که از رایج ترین آنها روش داده کاوی می‌باشد. اگرچه روشها و الگوریتم‌های روش داده کاوی طی سالیان اخیر گسترش و پیشرفت چشم‌گیری داشته‌اند، اما مساله دقت و سرعت در کشف تقلب هنوز به عنوان مساله ای چالش برانگیز مطرح است. با توجه به اهمیتی که در مورد تقلب کارت‌های اعتباری و تراکنش‌های الکترونیکی مطرح شد، در این پژوهش سعی داریم تا با پیاده‌سازی روش‌هایی از داده کاوی، شیوه ای موثر برای افزایش دقت تشخیص تقلب بیابیم. در این پژوهش به سه روش داده‌کاوی می‌پردازیم. روش‌های رگرسیون و شبکه عصبی پس انتشار خطا به عنوان روش‌های موجود و روش شبکه عصبی مبتنی بر الگوریتم ژنتیک[1] به عنوان راه حلی جدید به تفسیر بیان خواهند شد. پژوهش جاری از داده‌های تراکنش‌های واقعی یک بانک بهره برده شده است، لذا نتایج حاصله نسبت به پژوهش‌های دیگر واقعی‌تر بوده و این از مزایای مهم پژوهش محسوب می‌شود. با توجه به معیارهای به دست آمده، تشخیص مدل شبکه عصبی مصنوعی GMDH از بیشترین معیار صحت برخوردار بوده و بنابراین می‌توان آن را به عنوان بهترین مدل جهت تشخیص تقلب تراکنش‌های مالی (در بین مدل‌های بکار برده شده در این پژوهش) معرفی نمود.

مفاهیم پایه

پرداخت الکترونیکی عبارت است از فراهم آوردن وسیله پرداخت برای محصولات و سرویس‌ها از طریق ارتباط عمومی به نحوی‌که کلیه اطلاعات مربوط به پرداخت به‌وسیله یک شبکه ارسال شده و نیاز به هیچ ارتباط خارجی (فکس-تلفن یا پست) نباشد[محمودی، فروزنده، ۲۰۰۹]. به عبارت دیگر پرداخت الکترونیکی شامل هرگونه پرداختی بین افراد، کسب و کارها و یا بانک است که از بستر ارتباطات از راه دور و یا شبکه‌های الکترونیکی با استفاده از فناوری‌های جدید صورت می‌پذیرد(Sumanjeet, 2009). انواع سیستم‌های پرداخت الکترونیکی عبارتند از پرداخت از طریق یک واسط (سرویس‌های تسویه حساب)، پرداخت براساس مبادله الکترونیکی وجوه (EFT)- مبادله پول نمادین، پرداخت براساس پول الکترونیکی، کارت اعتباری، چک الکترونیکی، پول دیجیتالی.

از میان روش های ذکر شده کارت اعتباری مقبول‌ترین روش پرداخت خصوصا در کشورهای توسعه یافته و در حال توسعه است(محمودی، فروزنده، ۲۰۰۹) به گونه‌ای که بیشترین خسارت در اثر تقلب در کارت‌های اعتباری متوجه فروشندگان و

شرکت‌های صادرکننده کارت است. در حالیکه در انواع دیگر پرداخت‌های الکترونیکی بیشترین خطر متوجه خریدار است. به طور کلی اجزای تشکیل دهنده یک سیستم عملیاتی کارت اعتباری شامل صاحب کارت/مشتری، فروشنده/بازرگان، بانک پرداخت کننده، بانک صادر کننده، واسط امور بانکی می باشد. صادرکننده در این سیستم‌ها، همواره حساب خریدار را زیر نظر دارد و هنگام خرید، اعتبار وی را تائید یا رد می‌کند. در واقع طرف حساب فروشنده، پایانه پرداخت و طرف حساب پایانه پرداخت، صادر کننده کارت است. ممکن است که بانک اعتباری کسی که در حسابش موجودی کافی ندارد را تائید کند، در

[1] GMDH

Archive of SID
www.SID.ir

این صورت آن بانک موظف به تسویه حساب در مهلت مقرر است و سپس صادرکننده با قوانین داخلی خود با مشتری طرف حساب خواهد بود. به طبع کارهایی که بر روی حساب‌ها انجام می‌گیرد نیز تراکنش‌هایی مشخص، استاندارد و از پیش تعریف شده است که عبارتند از تایید هویت، ابطال، تسویه، استرداد، پرداخت، کسر از حساب، مطالبه مقدار.

پیشینه پژوهش

روش‌های بکار رفته در سامانه‌های کشف تقلب برای بازنمایی دانش، تحلیل اطلاعات و مدل کردن رفتار عادی، شامل روش‌هایی مانند سیستم‌های مبتنی بر قانون[۲]، روش آماری، تحلیل‌های گذار حالت[۳] و روش‌های داده‌کاوی مانند شبکه عصبی، درخت تصمیم‌گیری، الگوریتم ژنتیک، شبکه بیزین و ماشین بردار پشتیبان، خوشه‌بندی داده‌ها یا ترکیبی از روش‌های فوق است. داده‌کاوی نام خود را از شباهت بین جستجو برای یافتن اطلاعات با ارزش تجاری در یک پایگاه اطلاعات بزرگ (مثل یافتن محصولات مرتبط در یک پایگاه اطلاعات بزرگ) و کاویدن کوه برای یافتن رگه‌های اصلی سنگ معدن گرفته است(Stands, 2010) هر دو فرایند شامل حرکت و یا جستجوی هوشمندانه در یک حجم بسیار بزرگ به منظور یافتن ارزش می‌باشد. امروزه داده‌کاوی با حمایت سه تکنولوژی که به مرحله بلوغ رسیده‌اند کاملا آماده استفاده می‌باشد. این سه تکنولوژی عبارتند از جمع‌آوری و ذخیره‌سازی داده‌ها، کامپیوترهای با قدرت پردازش چندگانه، الگوریتم‌های پیشرفته داده‌کاوی. برای فهم بهتر اینکه داده‌کاوی چه کاری انجام می‌دهد، بهتر است ابتدا مفاهیم، کاربردها، عملیات و روش‌های داده‌کاوی به تفکیک بیان شوند(Yamanishi, 2011). این اطلاعات عبارتند از:

- عملیات، شامل دسته‌بندی، پیش‌بینی، خوشه‌بندی، تخمین، تعیین ارتباط بین متغیرها.
- کاربردها، شامل دسته‌بندی مشتریان، نگهداری مشتریان، شناسایی تقلب، تحلیل سایت‌های اینترنتی، تحلیل داده‌ها در بازاریابی، تحلیل و پیش‌بینی روندهای بازار بورس.
- تکنیک‌ها، شامل شبکه‌های عصبی، درخت تصمیم‌گیری، الگوریتم‌های نزدیکترین همسایه، تحلیل خوشه، استنتاج قانون، الگوریتم ژنتیک.

روش پیشنهادی

یکی از روش‌هایی که به عنوان یک روش امیدوارکننده در مسأله تشخیص تقلب معرفی گردیده است، روش شبکه عصبی مصنوعی می‌باشد که در برخی از موارد توانایی بالقوه‌ی خوبی جهت تشخیص تقلب‌های مالی از خود نشان داده‌اند. اما نکته اساسی این است که در مدل‌های شبکه عصبی مصنوعی، پیش‌داوری محقق در مورد ساختار مدل با توجه به فروض اولیه، به شبکه تحمیل می‌شود. حال آنکه در این بین مدل‌های شبکه عصبی مبتنی بر الگوریتم GMDH، نوعی شبکه‌ی خودسامانده و یک‌سویه بوده و دیگر نیازی به تحمیل شرایط اولیه توسط محقق ندارد، که استفاده از این نوع شبکه عصبی به دلیل بیان شده از جمله نوآوری‌های این پژوهش به‌شمار می‌آید.

[۲] Rule-Based Expert System

[۳] State transition analysis

Archive of SID
www.SID.ir

شبکه عصبی مصنوعی

شبکه عصبی روشی است که با استفاده از مجموعه‌ای از گره‌های به‌هم‌مرتبط، از کارکرد مغز انسان تقلید می‌کند. این روش مبتنی است بر مدلهای رایانه‌ای از نورون‌های زیستی. یک شبکه عصبی چندلایه دربرگیرنده تعداد زیادی واحد (نورون) به‌هم‌مرتبط در الگویی از ارتباطات است (Phua et al., ۲۰۰۵). این روش به شکل گسترده‌ای در رده بندی و خوشه‌بندی استفاده شده است و پس از رگرسیون پرکاربردترین روش داده‌کاوی مورد استفاده در کشف تقلبهای مالی است (Yue et al., ۲۰۰۷). نخست، شبکه با استفاده از مجموعه‌ای از داده‌های زوجی برای ترسیم ورودیها و خروجیها آموزش داده می‌شود. سپس وزن ارتباطات بین نورون‌ها تثبیت می‌شود و شبکه برای تعیین رده بندی‌های مجموعه‌ای جدید از داده‌ها مورد استفاده قرار می‌گیرد (Phua et al., ۲۰۰۵). مزایای این روش از این قرارند؛ نخست اینکه این روش انطباق‌پذیر است. دوم اینکه این روش مدلهای دارای پایایی۲۶ ایجاد می‌کند و سوم اینکه اگر وزنهای آموزشی تغییر کنند، فرایند رده بندی را نیز می‌توان اصلاح کرد. آنان با استفاده از داده‌های مالی و غیرمالی، یک مدل بحران مالی تدوین کردند. نتایج مطالعه آنان نشان می‌دهد که شبکه‌های عصبی مصنوعی بهتر از روشهای سنتی آماری، بحران مالی را پیش‌بینی می‌کنند.

الگوریتم ژنتیک

الگوریتم ژنتیک یکی از الگوریتم‌های جست‌وجو بوده که بر اساس ژنتیک ارگانیزم‌های زنده پایه‌گذاری شده است. این الگوریتم اصل "انتخاب یا بقای اصلح[٤]" داروین[٥] را با یک سری اطلاعات تصادفی ساخت، ادغام نموده و یک الگوریتم جست-وجو با خصوصیت روش‌های تکاملی طبیعی ایجاد می‌نماید (Shu-Heng, 2002). در واقع اساس کار این‌گونه الگوریتم‌ها، حل نمودن مسائل دنیای واقعی از طریق شبیه سازی فرآیندهای ژنتیک طبیعی می‌باشد. همچنین، با توجه به آنکه بسیاری از مسائل دنیای صنعتی بویژه سیستم‌های تولیدی بسیار پیچیده بوده و با روش‌های بهینه سازی سنتی و متداول قابل حل نمی-باشند، نقش روش‌های بهینه‌سازی احتمالی نظیر الگوریتم ژنتیک در این زمینه بسیار حساس و مؤثر خواهد بود.

مدل‌های شبکه عصبی مبتنی بر الگوریتم GMDH

اساساً روش‌های تکاملی[٦] مانند الگوریتم ژنتیک، کاربرد وسیعی در مراحل مختلف طراحی شبکه‌های عصبی دارند (مهرآرا، بهزادمهر، اهراری، محقق، ۲۰۱۱) چنانکه دارای قابلیت‌های منحصربه فردی در یافتن مقادیر بهینه و امکان جست‌وجو در فضاهای غیر قابل پیش‌بینی هستند (جعالی، نریمان زاده، آتش کاری، ۲۰۰۶). از این‌رو در این پژوهش جهت طراحی شکل شبکه عصبی و تعیین ضرایب آن، از الگوریتم ژنتیک استفاده شده است. زیرا شبکه عصبی GMDH مبتنی بر الگوریتم ژنتیک، ابزاری با قابلیت بالا در مدل‌سازی سیستم‌های غیرخطی پویای پیچیده به شمار می‌رود.

[٤] Survival of the Fittest

[٥] Darvin

[٦] -Evolutionary

Archive of SID
www.SID.ir

شبکه عصبی GMDH در برگیرنده مجموعه‌ای از نرون‌ها بوده که از پیوند جفت‌های مختلف از طریق یک چندجمله‌ای درجه دوم به وجود می‌آید. فرض کنید مجموعه ای از m متغیر شامل $x_1, x_2, \ldots x_m$ و یک متغیر y وجود دارد، داده های مربوط به هر کدام از x_i ها و متغیر هدف y متغیر خروجی نیز برای یک دوره زمانی وجود دارد به عبارتی هر یک از متغیرها به صورت یک بردار که شامل اعداد سری زمانی مربوط به آن متغیر است، می‌باشد (زمانی، ۲۰۱۰). اطلاعات اولیه‌ای که جهت ساخت الگوریتم GMDH باید جمع‌آوری گردد. مجموعه‌ای از n مشاهده است که در ماتریس زیر نشان داده شده است (زمانی، ۲۰۱۰).

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_{\eta t} \\ \cdots \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} \vdots x_{12} \ldots \vdots x_{1m} \\ x_{21} \vdots x_{22} \ldots \vdots x_{2m} \\ \cdot \vdots \\ \cdot \vdots \\ x_{\eta} \vdots x_{\eta 2} \ldots \vdots x_{\eta m} \\ \cdots \vdots \cdots\cdots\cdots \\ \cdot \vdots \\ \cdot \vdots \\ x_n \vdots x_{n2} \ldots \vdots x_{nm} \end{bmatrix}$$

مجموعه آموزش

مجموعه آزمون

برای شروع به کار الگوریتم با دو مسئله مواجه هستیم تشخیص رابطه‌ای که متغیر خروجی را بر اساس متغیرهای ورودی x_i ها تولید می کند. پیش بینی y به ازای مقادیر معلوم x_i ها به عبارتی نیاز به تشخیص مدل و رابطه بین متغیرها می‌باشد (مدل سازی) که سپس بتوان از روی آن مدل مقادیر آتی متغیر هدف پیش بینی کرد [Ivakhnenko, 1995].

مبنای الگوریتم GMDH عبارت از فرآیندی جهت ساخت یک چند جمله‌ای با مراتب بالا است که به سری تابع ولترا[۷] معروف است و به شکل زیر ارائه می‌گردد: (این چند جمله-ای را ایواخننکو نیز می‌نامند.)

$$\hat{y} = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk} x_i x_j x_k + \ldots \qquad (۱)$$

برای این منظور در الگوریتم GMDH ابتدا به تجزیه سری توابع ولترا به چند جمله‌ای های دو متغیره درجه دوم می پردازیم.

$$G(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i^2 + a_1 x_j^2 + a_5 x_i x_j \qquad (۲)$$

در این تجزیه، سری ولترا به مجموعه‌ای از معادلات بازگشتی زنجیره‌ای تبدیل می گردد به گونه‌ای که مجددا با جایگذاری جبری هر یک از روابط بازگشتی در این رابطه سری ولترا برقرار می‌گردد.

$$y_i = f(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{im}) \qquad i = 1,2,3,\ldots,m \qquad (۳)$$

[۷] Volterra Functional Series

Archive of SID
www.SID.ir

تقریب زده می‌شود: f توسط تابع y_i

$$\hat{y}_i = \hat{f}(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{im}) \qquad i = 1,2,3,\ldots,m \tag{۴}$$

و در صورتی که تابع f به صورت زیر بیان شود:

$$\hat{y} = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk} x_i x_j x_k + \ldots \tag{۵}$$

در این صورت رابطه F را می توان به فرم زیر تجزیه نمود:

$$\begin{aligned}
&\hat{y}_k = G(u_i, u_j) && i, j = 1,2 (i \neq j) && && k = 1\\
&\hat{u}_k = G(s_i, s_j) && i, j = 1,2, k, F_1 (i \neq j) && F_1 \leq C_{F_2}^2 && k = 2\\
&\hat{s}_k = G(p_i, p_j) && i, j = 1,2, k, F_2 (i \neq j) && F_2 \leq C_{F_3}^2 && k = 3\\
&\vdots\\
&\hat{z}_k = G(w_i, w_j) && i, j = 1,2, k, F_l (i \neq j) && F_l \leq C_m^2 && k = F_{l+1}\\
&\hat{w}_k = G(x_i, x_j) && i, j = 1,2, k, F_m (i \neq j) && && k = F_m
\end{aligned} \tag{۶}$$

همان گونه که در معادلات بالا مشاهده می شود ترتیب روابط فوق از بالا به پایین نمایی از پروسه تجزیه رابطه (۷) به چند جمله‌ای‌های درجه دوم بوده و از طرفی نیز ترتیب این روابط از پایین به بالا بیانگر تکمیل رابطه (۷) توسط معادلات بازگشتی می باشد در واقع هدف این الگوریتم یافتن ضرایب مجهول α در سری توابع ولترا می باشد. لازم به ذکر است که تمامی مدل-های جزیی به وجود آمده از یک ساختار متشابه همانند رابطه زیر برخوردار هستند:

$$\hat{f}(x_i, x_j) = v_0 + v_1 x_i + v_2 x_j + v_3 x_i^2 + v_1 x_j^2 + v_5 x_i x_j \tag{۷}$$

با توجه به اینکه هدفی را که ما در این الگوریتم دنبال می کنیم چیزی جز مدل‌سازی سیستم اولیه نیست لذا با ترکیب مدل سیستم‌های جزیی و تکرار این عمل می‌توان به مدل اصلی سیستم دست یافت.

$$\hat{y} = v_0 + \sum_{i=1}^{m} v_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} v_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} v_{ijk} x_i x_j x_k + \ldots \tag{۸}$$

بعد از تجزیه سیستم اصلی به تعداد C_m^2 سیستم جزیی، مدلی با دو متغیر ورودی برای هر یک از آن ها محاسبه می‌شود. سپس مدل های جزیی ایجاد شده را دو به دو با هم ترکیب کرده که حاصل این ترکیب، تعداد $\frac{C_m^2(C_m^2-1)}{2}$ سیستم و یا مدل جزیی جدید با حداقل سه و چهار متغیر ورودی می باشد، البته تعداد متغیرهای وابسته به مدل و یا به عبارتی تعداد ورودی های سیستم مهم نبوده و تنها دقت تخمین واقعی سیستم اصلی توسط مدل های ایجاد شده اهمیت دارد لذا با در نظر گرفتن این قاعده برای کاهش محاسبات مضاعف و زائد و افزایش راندمان و دقت مدل سازی تعدادی از مدل های جزیی تشکیل شده را که از دقت و تخمین بالایی نسبت به مدل های دیگر برخوردار هستند انتخاب نموده و مابقی را حذف می کنیم (ابریشمی، مهرآرا، اهراری، محقق، ۲۰۱۰).

در دومین مرحله ترکیب مدل های جزیی انتخاب شده و یا به عبارتی سیستم های ایده آل شکل گرفته در مرحله قبلی مجددا دو به دو همانند مرحله قبلی ترکیب شده و سیستم های جزیی جدیدی با حداقل پنج و حداکثر شش متغیر ورودی

Archive of SID
www.SID.ir

تشکیل می شود به همین طریق در مراحل بعدی نیز با انتخاب و حذف تعدادی از مدل های جزیی ایجاد شده عمل ترکیب آنها را ادامه می دهیم تا در نهایت به مدلی نسبتا ایده‌آل دست یابیم.

هدفی که همواره در پروسه ترکیب مدل های به وجود آمده مورد نظر است دست یافتن به مدلی می‌باشد که تقریبا تمامی متغیرهای سیستم در آن نمایان و نقش داشته باشند و هدف دیگری که در انجام ترکیب‌های مکرر مورد نظر است رسیدن به مدلی می‌باشد که میزان خطای خروجی آن نسبت به سایر مدل‌های محاسبه شده در مراحل قبل کمتر باشد (شارزهی، اهراری، فخرایی، ۲۰۰۹).

روش پیشنهادی، شامل استفاده از شبکه عصبی مبتنی بر الگوریتم ژنتیک (GMDH) است. این شبکه‌ها مبتنی بر شبیه‌سازی‌های کامپیوتری از نرون‌های عصبی هستند و با استفاده از یک پایه ریاضیاتی اعمال خود را انجام می‌دهند. ایده اصلی این روش عبارتست از طراحی یک مدل بهینه‌ی پیچیده، که فقط مدل را بر پایه داده‌ها و اطلاعات طراحی کند و هیچ‌گونه پیش‌زمینه نظری از نحوه عملکرد داده‌ها از سوی محقق انجام نگیرد و این کار تنها بر اساس کشف ارتباط ساده و پیچیده میان داده‌های ورودی و خروجی سیستم انجام گیرد، بنابراین یک مدل خودتنظیم کننده خواهد ساخت که قابلیت حل مسائلی نظیر تشخیص، پیش‌بینی و سایر مسائل سیستمی به کار برده شده را دارد. در ادامه، ضمن پیاده‌سازی روش مذکور، نخست روش‌های گوناگون بر روی داده‌ها بکار گرفته خواهد شد و با استفاده از معیارهای ارزیابی، دقت و کارایی هر روش محاسبه و سپس با مقایسه کارکرد روش ها با یکدیگر، توانمندی‌های هر روش ارزیابی خواهد شد.

تحلیل و ارزیابی نتایج

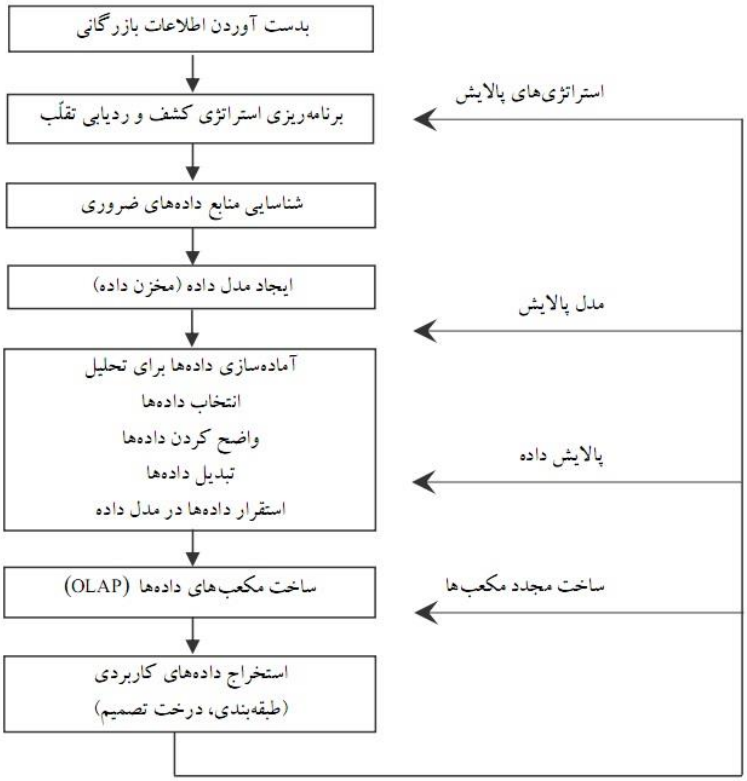
در این پژوهش شبکه عصبی مبتنی بر الگوریتم ژنتیک (GMDH) پیشنهاد شد. مزیت روش پیشنهادی نسبت به شبکه‌های عصبی مصنوعی، کشف ارتباط ساده و پیچیده میان داده‌های ورودی و خروجی سیستم و به تبع آن سازمان‌دهی خودکار ساختار شبکه می‌باشد. در ادامه به جزئیات استفاده از این نوع از شبکه‌های عصبی به همراه ارزیابی و تحلیل نتایج پرداخته خواهد شد.

یک ساختار هوشمند تشخیص فعالیت‌ها و تراکنش‌های مشکوک به تقلب به طور عام از چند لایه متشکل شده است که در سطوح مختلف تحلیل می‌شوند که اصلی‌ترین سطح، سطح تراکنش‌های بانکی می‌باشد. تحلیل تراکنش یکی از مهمترین و اصلی‌ترین کارهای یک سیستم مبارزه با تقلب می‌باشد. تراکنش‌ها و حساب‌ها را نمی‌توان به طور مجزا مورد بررسی قرار داد. بلکه مجموعه تراکنش و حساب بانکی به همراه هم باید یک دید کلی از رفتار مشتریان ارائه کنند. تحلیل‌ها بر پایه دو ویژگی مهم می‌باشند؛ فرکانس تراکنش‌ها و مقدار هر تراکنش. زیرسیستم پیشنهادی از ترکیب این دو ویژگی به صورت یک مجموعه قوانین برای تشخیص موارد مشکوک استفاده می‌کند.

اغلب روش‌ها و نرم افزارهای شناسایی تقلب موجود براساس فرکانس و میزان یک نوع تراکنش مشخص و براساس میانگین و انحراف معیار آن عمل می‌کنند. در واقع چنین سیستم‌هایی به نوعی، فقط مقایسه یک طرفه را در نظر می‌گیرند؛ یعنی رفتار یک مشتری خاص در برابر همین مشتری در گذشته. از نظر منطقی این رویکرد برای ردیابی تراکنش‌های یک مشتری کافی است اما این کار به تنهایی کافی نمی‌باشد، چراکه فاکتورهای بسیاری وجود دارند که میزان فرکانس تراکنش‌ها را در بانکداری تحت تاثیر قرار می‌دهند، مثل محیط سیاسی، اقلیم و نوع بازار سرمایه، ارزش سرمایه و غیره. یکی از راه‌حل‌های موجود برای تعیین پارامترهای تاثیرگذار و همچنین میزان تاثیر هر پارامتر در جهت افزایش دقت و کارایی یک سیستم ضدتقلب استفاده از روش‌های داده‌کاوی است. در واقع سیستم پیشنهادی به عنوان یک سامانه مبارزه با تقلب مبتنی بر روش‌های داده‌کاوی در زمینه تحلیل تراکنش‌های بانک و بهبود دقت و سرعت کشف موارد مشکوک به تقلب، عملکرد بسیار خوبی را خواهد داشت. در

Archive of SID
www.SID.ir

این سیستم از شبکه عصبی GMDH جهت شناسایی موارد رفتاری و تراکنش‌های غیرمعمول و مشکوک و اعلان هشدار به سیستم بانکی استفاده می‌شود. از ویژگی‌های بارز این سیستم، استفاده از عملیات اکتشاف متوالی جهت تعیین الگوهای توالی داده‌ها و تراکنش‌ها و همچنین استفاده از تحلیل داده‌ها جهت شناسایی تراکنش‌ها و شبکه جرایم مالی و استخراج داده‌های محرمانه در اسناد بانک است. اجزای سیستم پیشنهادی به منظور پیاده‌سازی این وظیفه در شکل ۱ نشان داده شده است.



شکل ۱- اجزای سیستم پیشنهادی جهت پیاده‌سازی وظیفه کشف تقلب

برای ساختن یک مدل ریاضی، نیاز به داده‌هایی از هردو دسته تراکنش‌های تقلب و غیر تقلب داریم. در این بخش با استفاده از اطلاعات تراکنش‌های دستگاه POS مربوط به یک بانک که در طی آذرماه سال ۱۳۹۱ تا پایان اردیبهشت ماه ۱۳۹۲ ثبت شده‌اند و همچنین با به‌کارگیری روش‌های ذکرشده در بخش قبل، مدل‌هایی جهت شناسایی تقلب در تراکنش‌های مالی ارائه شده است. این داده‌ها شامل حدود یکصد و نه هزار تراکنش است که پس از انتخاب حدود ده هزار داده از این میان، تعداد ۶۸ تراکنش توسط کارشناسان به عنوان تقلب تشخیص داده شده است.

در هریک از مدل‌هایی که در این بخش برای شناسایی تقلب استفاده خواهند شد، تقلب یا غیرتقلب بودن یک تراکنش، به عنوان متغیر وابسته در نظر گرفته می‌شود. مقدار ۱ برای متغیر وابسته به معنای تقلب بودن تراکنش و مقدار ۰ به معنای غیرتقلب بودن آن تراکنش است. در این مطالعه، از ۳۷ متغیر مستقل ثبت شده، ۹ تای آن برای فرآیند شناسایی تقلب انتخاب

www.SID.ir

شده است. برای انتخاب متغیرهای مستقل از کارشناسان بانکی در حوزه شناسایی تقلب بهره گرفته شد. در جدول ۱ متغیرهای مستقل مورد استفاده جهت فرآیند شناسایی تقلب آورده شده است.

جدول ۱- متغیرهای انتخاب شده در فرآیند تشخیص تقلب

متغیر مستقل	توصیف متغیر
CARD_NUMBER	شماره کارت مبدأ
PROCESSING_CODE	کد پردازش
TRANSACTION_AMOUNT	مقدار تراکنش
CARD_ACCEPTOR_ACTIVITY	شناسه پذیرنده کارت
TRX_TIME	زمان تراکنش
TRX_DATE	تاریخ تراکنش
CARD_MAGSAD	شماره کارت مقصد
PAZIRANDEH	کد پذیرنده
POS_DATA	اطلاعات پیگیری درگاه

Archive

روش رگرسیون لجستیک

زمانی که متغیر وابسته، متغیری کیفی با دو سطح باشد، مدل‌های رگرسیون معمولی قابل استفاده نیستند. در این‌گونه موارد معمولا از رگرسیون لجستیک استفاده می‌شود. مدل رگرسیون لجستیک به این صورت تعریف می‌شود:

$$logit\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d \qquad (۹)$$

در این مدل $X_1, \ldots, X_d$ متغیرهای مستقل و p احتمال مشاهده مقدار ۱ برای متغیر وابسته به شرط مشاهده مقادیر $X_1, \ldots, X_d$ است. ضرایب رگرسیونی در این حالت با فرض دوجمله‌ای بودن توزیع متغیر وابسته از روش برآورد حداکثر درست نمایی[۸] (MLE) محاسبه می‌شوند. با توجه به اینکه در این تحقیق متغیر وابسته (وضعیت تقلب بودن تراکنش) یک متغیر

[۸] Maximum Likelihood Estimation

www.SID.ir

دوسطحی است، از رگرسیون لجستیک برای تشخیص تقلب یا غیرتقلب بودن تراکنش‌های مالی استفاده شده است. برای این منظور و شبیه‌سازی مدل رگرسیون لجستیک از نرم افزار MATLAB بهره گرفته شد.

پس از انجام شبیه‌سازی و ارائه خروجی، برای به‌دست آوردن دقت مدل، رابطه ۱۰ محاسبه می‌گردد. در این رابطه منظور از $|a|$ همان قدرمطلق a است و F_i برابر با خروجی واقعی و F'_i برابر با خروجی محاسبه شده توسط مدل است.

$$Accuracy = \left(\sum_{j=1}^{m} \left(1 - \left(|F_i - F'_{i,j}| \,/\, |F_i|\right)\right)\right)/m \qquad (۱۰)$$

به‌طوری که m تعداد داده‌های آزمون باشد. مقدار صحت الگوریتم بین ۰ تا ۱ تغییر می‌کند و هرچه بیشتر باشد بیانگر دقت بهتر در انجام عملیات است. با به‌کارگیری این مدل بر روی داده‌های اولیه، دقت مدل سنجیده شده است. همان‌طور که در جدول۲ مشاهده می‌شود، دقت مدل در شناسایی تراکنش‌های تقلب، ۸۶,۵۷ درصد و در شناسایی تراکنش‌های غیرتقلب ۸۹,۳۵ بوده است. ضمن اینکه دقت مدل در شناسایی صحیح تقلب یا غیرتقلب هر تراکنش برابر ۸۹,۲۶ درصد است.

جدول ۲- نتایج برآورد دقت تشخیص تقلب مدل رگرسیون لجستیک

		برآورده شده		درصد صحیح
		تقلب	غیر تقلب	
مشاهده شده	تقلب	۵۸	۹	۸۶,۵۷
	غیر تقلب	۲۱۷	۱۸۲۰	۸۹,۳۵
کل				۸۹,۲۶

روش شبکه عصبی پس انتشار خطا

شبکه‌های عصبی مصنوعی و شبکه‌های عصبی پرسپترون چندلایه[۹] (MLP) با الگوریتم آموزشی پس انتشار خطا[۱۰] (BP) برای حل هر مسئله‌ای، سه مرحله را طی می‌کنند که این مراحل به ترتیب عبارتند از آموزش، تعمیم و اعتبارسنجی (امینیان، محقق، عامری، ۱۹۹۵). الگوریتم یادگیری BP، بر اساس الگوریتم تقریبی بیشترین تنزل[۱۱] (S.D) است. تنظیم پارامترهای شبکه، مطابق با سیگنال‌های خطا که بر اساس ارائه هر الگو به شبکه محاسبه می شود، صورت می‌گیرد.

در نهایت الگوریتم بیشترین تنزل با معادله ۱۱ توصیف می‌شود:

[۹] MultiLayer Perceptron

[۱۰] Back-Propagation Algorithm

[۱۱] Steepest Descent

www.SID.ir

$$S^L j(k) = \frac{\delta F(K)}{\delta nj^L(K)} = \left[\sum_{m=1}^{S^{L+1}} S_m{}^{l+1}(k) W_{mj}^{l+1}(k)\right] f^L(N_j^L(k)) \quad (۱۱)$$

به طوری‌که $S^L_j(k)$، حساسیت رفتار شبکه در لایه Lام است.

در قدم اول برای آموزش شبکه، مجموعه دادگان آماده شده در بخش قبل را که در قالب فایل اکسل تنظیم شده بود، به عنوان مجموعه آموزشی یا به عبارت دیگر بردارهای ورودی انتخاب کرده و ۹ ستون مربوط به متغیرهای انتخاب شده برای تراکنش‌های مالی را از نرم افزار اکسل به شبکه عصبی اعمال می‌کنیم (در نرم افزار MATLAB). به‌طور مشابه در مورد مقادیر هدف (تقلب یا غیرتقلب بودن تراکنش‌ها) نیز که به صورت کد ۰ و ۱ در نظر گرفته شده‌اند، کدهای مربوط به همان تراکنش‌ها در قسمت قبل را به عنوان خروجی مطلوب به شبکه وارد می‌کنیم. قدم بعدی، ایجاد ساختار اولیه شبکه و پس از آن آموزش شبکه می‌باشد. با استفاده از بردارهای ورودی و هدف و نیز با تعیین تعداد نرون‌های لایه میانی می‌توان یک شبکه با دولایه به صورت پیش‌فرض ایجاد کرد. سپس شبکه مورد نظر در مرحله آموزش، روابط بین داده‌های ورودی و هدف را با استفاده از الگوریتم‌های آموزشی (BP) یاد می‌گیرد. بدین منظور و در خلال فرایند آموزش، بردارهای ورودی و هدف به‌طور تصادفی به سه دسته تقسیم می‌شوند. دسته اول که شامل ۶۰ درصد از بردارهای ورودی است، به عنوان مجموعه آموزشی وارد مرحله آموزش می شود. دسته دوم که شامل ۲۰ درصد دیگر از بردارها می‌باشد، به منظور اعتبارسنجی انتخاب می‌شوند تا توسط آنها قدرت تعمیم شبکه سنجیده شود. این داده‌ها برخلاف داده‌های آموزشی که در تعیین وزن‌های شبکه مورد استفاده قرار می‌گیرند، در تعیین وزن‌ها نقش مستقیم ندارند. در حقیقت پس از ارائه داده‌های آموزشی به شبکه، داده‌های مربوط به اعتبارسنجی با استفاده از وزن‌های حاصله، وارد شبکه شده و مقادیر خروجی شبکه، مانند قبل در قالب کدهای ۰ و ۱ برای نشان دادن مقادیر غیرتقلب و تقلب محاسبه می‌شوند. مرحله آموزش تا زمانی ادامه می‌یابد که سبب کاهش خطا در بردارهای مربوط به اعتبارسنجی شود. اما در حالتی که خطای مربوطه بیش از حد زیاد شود، شبکه به جایی می‌رسد که قدرت تعمیم خود را از دست می‌دهد. در این حالت شبکه، قابلیت دریافت ارتباط صحیح بین بردارهای ورودی و خروجی را از دست می‌دهد. اما هر زمان که این مشکل رخ دهد، شبکه با تقسیم‌بندی ذکرشده که به صورت خودکار انجام می‌گیرد، آموزش را متوقف می‌کند تا از مشکل فوق‌الذکر جلوگیری شود. ۲۰ درصد بردارهای باقیمانده نیز که تاکنون در شبکه مورد استفاده قرار نگرفته‌اند، برای سنجش نهایی قابلیت تعمیم شبکه استفاده می‌شوند.

پس از اعمال بردارهای ورودی و خروجی مطلوب به شبکه و انجام فرایند آموزش، نوبت به ارزیابی میزان دقت پاسخ‌های شبکه و یا به عبارت دیگر خروجی‌های شبکه می‌رسد. پس از محاسبه دقت الگوریتم نتایج موجود در جدول ۳ حاصل شد.

جدول ۳- نتایج برآورد دقت تشخیص تقلب مدل شبکه عصبی BP

		برآورده شده		درصد صحیح
		تقلب	غیر تقلب	
مشاهده شده	تقلب	۶۹	۱۱	۸۶,۲۵
	غیر تقلب	۱۸۸	۱۸۳۶	۹۰,۷۱
کل				۹۰,۵۴

Archive of SID
www.SID.ir

همان‌طور که در جدول ۳ مشاهده می‌شود، دقت مدل در شناسایی تراکنش‌های تقلب، ۸۶,۲۵ درصد و در شناسایی تراکنش‌های غیرتقلب ۹۰,۷۱ بوده است. ضمن اینکه دقت مدل در شناسایی صحیح تقلب یا غیرتقلب هر تراکنش برابر ۹۰,۵۴ درصد است.

روش شبکه عصبی مبتنی بر الگوریتم GMDH

موضوع طراحی شبکهٔ عصبی GMDH با مسائل عنوان شده در بخش قبل متمایز است. در این نوع از طراحی، هدف، جلوگیری از رشد واگرایی شبکه و نیز مرتبط کردن شکل و ساختار شبکه به یک یا چند پارامتر عددی است، به گونه‌ای که با تغییر این پارامترها، ساختار شبکه‌ها نیز تغییر کند. روش‌های تکاملی مانند الگوریتم ژنتیک، کاربرد وسیعی در مراحل مختلف طراحی شبکه های عصبی به دلیل قابلیت‌های منحصر به فرد خود در پیدا کردن مقادیر بهینه و امکان جستجو در فضاهای غیرقابل پیش‌بینی، دارند (نریمان زاره، ۸۰-۸۷). در تحقیق حاضر، برای طراحی شکل شبکهٔ عصبی و تعیین ضرایب آن، از الگوریتم ژنتیک استفاده شده است. به‌منظور تدوین الگو و تشخیص تقلب تراکنش‌های مالی در این مدل، از نرم افزار GEvoM[۱۲] استفاده می‌شود (برای توسعه این نرم‌افزار Net Framework. بکار رفته است). آدرس وب‌گاه برنامه نصب‌کننده نرم‌افزار از آدرس زیر، قابل دانلود است:

http://research.guilan.ac.ir/gevom/docs/GEvoM۱۰۰-Demo.exe

نرم افزار محاسباتی مبتنی بر بهینه سازی چند منظوره (آتش کاری، نریمان زاده، ۲۰۰۷)، با هدف کمینه کردن خطای الگوسازی و تشخیص، طراحی شده که با استفاده از نمودار پرتو (امانی فرد، نریمان زاده، برجی، خلخالی، حبیب دوست، ۲۰۰۸) دقت تشخیص و پایداری فرآیند را افزایش می‌دهد. منظور از خطای الگوسازی، خطایی است که به واسطهٔ اختلاف مقادیر برآورد شده با مقادیر واقعی در دوره آموزش شبکه حاصل شده است و منظور از خطای تشخیص، میزان خطایی است که به واسطهٔ اختلاف مقادیر برآورد شده و مقادیر واقعی در مشاهدات دورهٔ پیش بینی محاسبه شده است.

این نرم‌افزار مبتنی بر شبکه‌های عصبی مبتنی بر الگوریتم GMDH بوده و امکان ایجاد شبکه‌های عصبی چندجمله‌ای را فراهم می‌کند. این نرم‌افزار امکان مدل‌سازی داده‌های آزمایشی و شبیه‌سازی را داراست. در این سیستم از ژنتیک و تکنیک SVD برای بهینه‌سازی پیکربندی شبکه و ضرایب چندجمله‌ای حاصله استفاده می‌شود. این نرم‌افزار از چهار متد مختلف ISP[۱۳]، PSD[۱۴]، ED[۱۵] و روش تکاملی برای طراحی شبکه‌های عصبی مبتنی بر GMDH استفاده می‌کند. روش تکاملی از الگوریتم ژنتیک برای پیکربندی بهینه اتصالات ساختاری شبکه عصبی GMDH استفاده می‌کند. در کلیه روش‌های فوق از SVD[۱۶] ویا SNE[۱۷] برای به دست آوردن ضرایب چندجمله‌ای استفاده می‌شود.

[۱۲] GMDH-type Neural Network Design by Evolutionary Method for Modeling

[۱۳] Increasing Selection Pressure Method

[۱۴] Pre-specified-Design Method

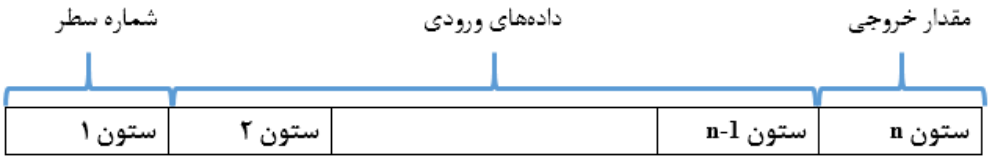
[۱۵] Error-Driven Method

[۱۶] Singular Value Decomposition

[۱۷] Solving Normal Equations

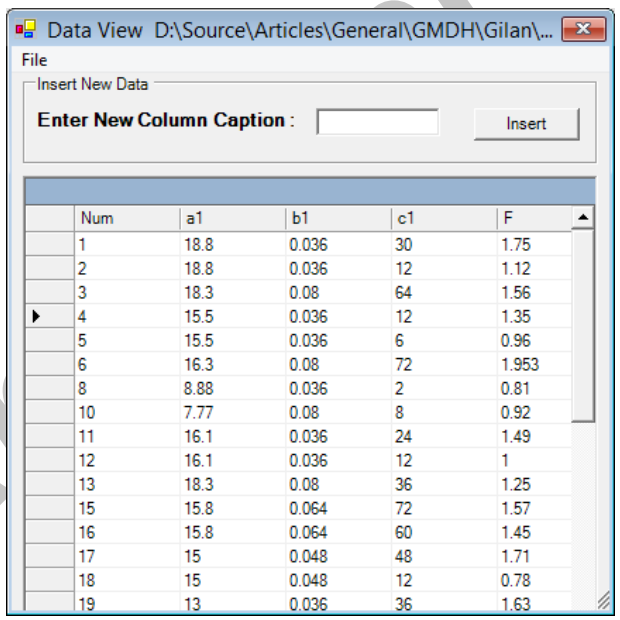
Archive of SID
www.SID.ir

برای کار با این ابزار می‌بایست داده‌های ورودی در قالب یک فایل CSV[۱۸] -که اعداد در هر سطر با یک فضای خالی از هم جدا شده باشند- به نرم‌افزار داده شود. لازم به ذکر است که اگر داده‌ها با کاما ویا سایر کاراکترها از هم جدا شده باشند نرم‌افزار آن‌ها را بارگذاری نخواهد کرد. امکان بازبینی و ویرایش داده‌ها با استفاده از DataView نیز وجود دارد. داده‌ها باید دارای تعداد ستون‌های مساوی در هر یک از سطرها باشند و اولین ستون به صورت یک عدد ترتیبی است که از ۱ شروع شده و در سطرهای بعدی یکی افزایش می‌یابد تا نشان دهنده شماره سطر باشد. بنابراین اگر داده‌های ورودی دارای n ستون باشند، ستون اول بیانگر شماره سطر خواهد بود، ستون ۲ تا n-۱ بعنوان متغیرهای ورودی بوده و دارای نام‌های مشخص هستند و در هر سطر مقدار مشخصی را می‌پذیرند. ستون آخر داده‌ها، حاوی مقدار خروجی است بنابراین قالب داده‌ها مطابق شکل ۲ خواهد بود.



شکل ۲- قالب داده‌های ورودی

شکل۳، نحوه ویرایش داده‌ها با قالب فوق بعد از ورود به سیستم را نشان می‌دهد.



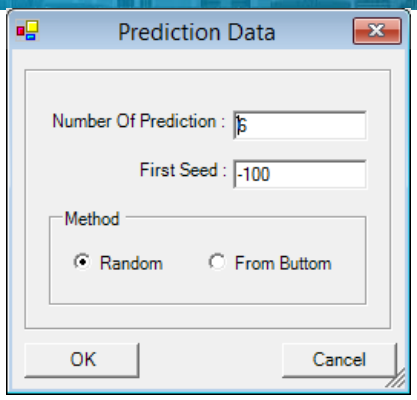
Num	a1	b1	c1	F
1	18.8	0.036	30	1.75
2	18.8	0.036	12	1.12
3	18.3	0.08	64	1.56
4	15.5	0.036	12	1.35
5	15.5	0.036	6	0.96
6	16.3	0.08	72	1.953
8	8.88	0.036	2	0.81
10	7.77	0.08	8	0.92
11	16.1	0.036	24	1.49
12	16.1	0.036	12	1
13	18.3	0.08	36	1.25
15	15.8	0.064	72	1.57
16	15.8	0.064	60	1.45
17	15	0.048	48	1.71
18	15	0.048	12	0.78
19	13	0.036	36	1.63

شکل ۳- صفحه ویرایشگر داده‌های ورودی

تنظیمات مربوط به پیش‌بینی براساس داده‌های ورودی، مطابق شکل ۴، وارد می‌شود.

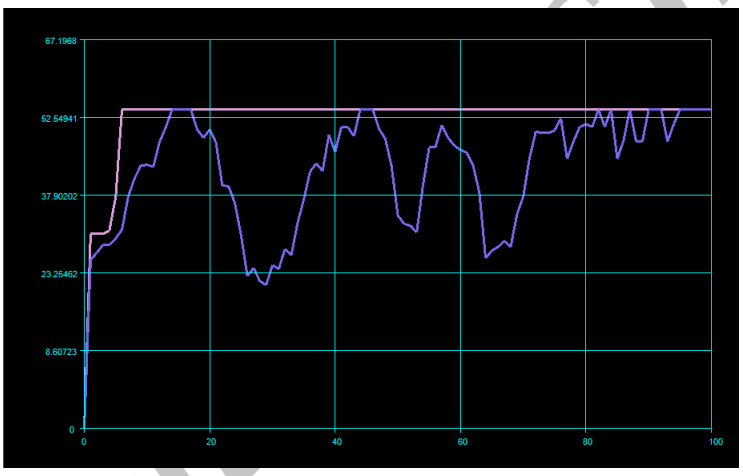
[18] Comma Separated Values

www.SID.ir



شکل ۴- نحوه مقداردهی پارامترهای پیش بینی

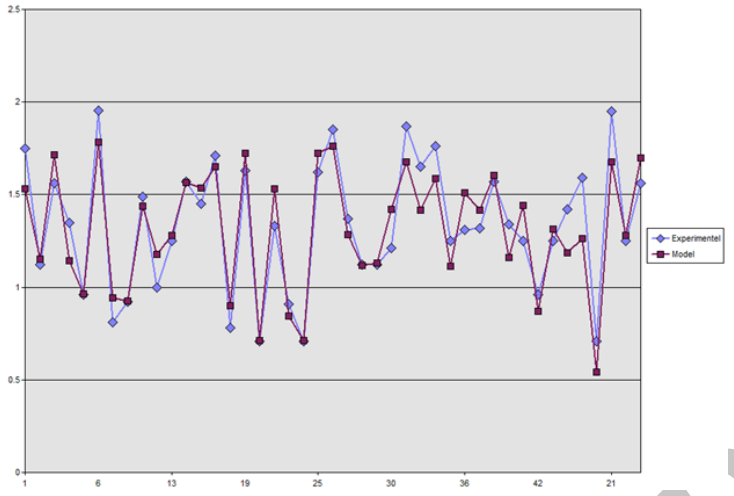
پس از مقداردهی پارامترهای مربوط به پیش بینی، نحوه پیشرفت اجرای سیستم توسط شکل ۵ نمایش داده می‌شود.



شکل ۵- نحوه پیشرفت اجرای سیستم

نمودار شکل ۶، امکان مقایسه دقیق بین داده‌های ورودی و مقادیر محاسبه شده توسط الگوریتم GMDH را نشان می‌دهد. محور افقی در این نمودار نمایش‌دهنده شماره سطر داده ورودی و محور عمودی نمایش‌دهنده مقدار خروجی شبکه‌عصبی GMDH است. همان‌طور که در این شکل مشخص است مقادیر واقعی با مقادیر محاسبه شده توسط سیستم، نظیر به نظیر در مجاور هم هستند و با دارا بودن دو رنگ متفاوت، امکان مقایسه دقیق و به دست آوردن دقت جواب حاصله توسط شبکه‌عصبی GMDH مقدور است.

www.SID.ir



شکل ۶- مقایسه مقادیر مطلوب داده‌های ورودی و مقادیر به دست آمده از مدل GMDH

پس از محاسبه دقت الگوریتم (با توجه به رابطه ۱۰) نتایج موجود در جدول ۴ حاصل شد.

جدول ۴- نتایج برآورد دقت تشخیص تقلب مدل شبکه عصبی GMDH

		برآورده شده		درصد صحیح
		تقلب	غیر تقلب	
مشاهده شده	تقلب	۶۷	۹	۸۸,۱۶
	غیر تقلب	۱۶۵	۱۸۶۳	۹۱,۸۶
کل				۹۱,۷۳

همان‌طور که در جدول۴ مشاهده می‌شود، دقت مدل در شناسایی تراکنش‌های تقلب، ۸۸,۱۶ درصد و در شناسایی تراکنش‌های غیرتقلب ۹۱,۸۶ بوده است. ضمن اینکه دقت مدل در شناسایی صحیح تقلب یا غیرتقلب هر تراکنش برابر ۹۱,۷۳ درصد است.

با توجه به معیارهای به‌دست آمده، دقت تشخیص مدل شبکه عصبی مصنوعی مبتنی بر الگوریتم ژنتیک (GMDH) از بیشترین معیار صحت برخوردار بوده و بنابراین می‌توان آن‌را به عنوان بهترین مدل جهت تشخیص تقلب تراکنش‌های مالی (در بین مدل‌های بکار برده شده در این پژوهش) معرفی نمود.

در این پژوهش روش داده‌کاوی رگرسیون لجستیک، شبکه عصبی BP و شبکه عصبی GMDH برای ساخت مدل‌هایی جهت شناسایی تقلب در تراکنش‌های مالی معرفی شدند. در ادامه، این روش‌ها برروی داده‌های واقعی آزمایش و کارایی هر روش سنجیده شد. روش شبکه عصبی GMDH با دقت ۹۱,۷۳ درصد در شناسایی تقلب یا غیرتقلب بودن تراکنش‌های مالی بهترین کارایی را در مقایسه با دو روش رگرسیون لجستیک با دقت کلی ۸۹,۲۶ و شبکه عصبی BP با دقت کلی ۹۰,۵۴

Archive of SID
www.SID.ir

داشت. البته باید به این نکته توجه داشت که در هر سه مدل معرفی شده برای تشخیص تقلب تراکنش‌های مالی، تعداد ۹ متغیر که از دید کارشناسان حوزه بانکی انتخاب گردید، استفاده شد. لذا همان‌طور که پیش از این نیز گفته شد، در این تحقیق متغیرهای دیگری نیز وجود دارند که به شرط ثبت اطلاعات آن‌ها در تراکنش‌های پیشین، می‌توانند با ورود به مدل به تشخیص بهتر تقلب در تراکنش‌های مالی کمک کنند.

نتیجه گیری

از جمله الگوریتم‌های داده کاوی مورد استفاده برای کشف تقلب می‌توان به شبکه‌های عصبی، خوشه بندی، درخت‌های تصمیم‌گیری، گراف تحلیل روابط، تحلیل آماری، تجمیع موارد مشکوک و موتور استنتاج اشاره نمود. همان‌طور که گفته شد، یکی از روش‌هایی امیدوارکننده در مسأله تشخیص تقلب، روش شبکه عصبی مصنوعی است که در برخی از موارد توانایی بالقوه‌ی خوبی جهت تشخیص تقلب‌های مالی از خود نشان داده‌اند. اما نکته اساسی این است که، در مدل‌های شبکه عصبی مصنوعی، پیش‌داوری محقق در مورد ساختار مدل با توجه به فروض اولیه، به شبکه تحمیل می‌شود. حال آنکه در این بین مدل‌های شبکه عصبی مبتنی بر الگوریتم GMDH، نوعی شبکه‌ی خودسامانده و یک‌سویه بوده و دیگر نیازی به تحمیل شرایط اولیه توسط محقق ندارد، که استفاده از این نوع شبکه عصبی به دلیل بیان شده به عنوان روش پیشنهادی در این پژوهش مطرح گردید. نتایج حاصل از این پژوهش نشان داد که مدل شبکه‌ی عصبی مصنوعی مبتنی بر الگوریتم GMDH، عملکرد قابل توجهی در شناسایی تقلب داشته و قادر به ارائه تشخیص‌های دقیق‌تری نسبت به مدل‌های پایه خود (مدل‌های به‌کار گرفته شده در این پژوهش) می‌باشد.

Archive of SID

www.SID.ir

مراجع

Mahmoudi,M. T, Forouzideh, N, Lucas, C, and Taghiyareh, F. (۲۰۰۹). "Artificial Neural Network Weight Optimization based on Imperialist Competitive Algorithm," Seventh International Conference on Computer Science and Information Technologies, Yerevan, Amenia, ۲۸ September-۲ October.

Sumanjeet, S. (۲۰۰۹). "Emergence of Payment Systems in the age of Electronic Commerce: the state of art," Global Journal of International Business Research, Vol. ۲, No. ۲, PP. ۱۷-۳٦.

Morgan, J. P. (March ۲۰۱۱). "Payment Fraud and Control Survey,".

Stands, K. (۲۰۱۰). On Knowledge Discovery-Data Mining Community Accessed from www.Kdnuggets.com.

Yamanishi, et al, K. (۲۰۱۱). "Online Unsupevised Outlier Detection Using Finite Mixtures with Discounting.

Shu-Heng, C. (۲۰۰۲). "Genetic Algorithm and Genetic Programming in Computational Finance", Springer Kluwer Academic Publications, New York; USA: ۹۰-۱۰۲.

Mehrara, M., Behzadmehr, N., Ahrari, M., Mohaghegh, M., (۲۰۱۱). Prediction of Volatility of Crude Oil Prices Using GMDH Neural Network, Qurterly Energy Economics Review, Vol. ۱(۲۵), pp. ۸۹-۱۱۲.

Jamali, A., Nariman -zadeh, N., Atashkari, K., (۲۰۰٦). "Inverse Modeling of Multi-Objective Thermodynamically Optimized and GAI " ۱٤th Annual (International) Mechanical Engineering Conference, Isfahan University of Technology, Isfahan, Iran.

Zamani, N., (۲۰۱۰), "Analysis and Forecast Crude Oil Sales Revenues Using Neural Networks and Their Comparison", MSc Thesis, University of Tehran.

Ivakhnenko.G.A (۱۹۹٥),"The Review of Problems Solvable by Algorithms of the Method of Data Handling (GMDH)", Pattern Recognition and Image Analysis, Vol.٥, No.٤, PP ٥۲۷-٥۳٥.

Abrishami, H., Mehrara, M., Ahrari, M., Mirghasemi, S., (۲۰۱۰). "Modeling and Forecasting Economic Growth of Iran GMDH Neural Network Approach", Journal of Economic Research, Vol. ۱(۸۸), PP. ۱-۲٤.

Sharzehi, G., Ahrari, M., Fakhraie, H., (۲۰۰۹). "Projection of Water Demand In Tehran Using Structural Patterns", Time Series and the GMDH-type Neural Network, Journal of Economic Research, Vol. ۱(۸٤), pp. ۱٥۱-۱۷٥.

White A.C., Molnar D., Aminian K., Mohaghegh S., Ameri S. & Esposito P., (۱۹۹٥). "The application of ANN for zone identification in a complex reservoir", SPE Paper ۳۰۹۷۷, SPE Eastern Regional Conference & Exhibition, Morgantown, West Virginia, USA.

Demuth H., Beal M. & Hagan M., (۱۹۹۲-۲۰۰۹). "Neural network toolbox ٦" (Users Guide), The MathWorksTM.

Nariman-zadeh, N.; Darvizeh, A.; Darvizeh, M.; Gharababaei, H., (۲۰۰۲), "Modelling of Explosive Cutting Process of Plates Using GMDH-type Neural Network and Singular Value Decomposition". Journal of Materials Processing Technology, ۱۲۸ (۱-۳): ۸۰-۸۷.

Atashkari, K. Nariman-Zadeh, N. Gölcü,M. Khalkhali, A , Jamali, A. (۲۰۰۷) "Modelling and Multi-objective Optimization of a Variable Valve-timing Spark-ignition Engine Using Polynomial Neural Networks and Evolutionary Algorithms", Energy Conversion and Management, ٤۸ (۳):۱۰۲۹-۱۰٤۱.

Archive of SID
www.SID.ir

Amanifard, N., Nariman-Zadeh, N., Borji, M., Khalkhali, A., Habibdoust, A. (۲۰۰۸). "Modelling and Pareto Optimization of Heat Transfer and Flow Coefficients in Microchannels Using GMDH type Neural Networks and Genetic Algorithms", Energy Conversion and Management, ٤٩(٢): ٣١١ – ٣٢٥.

Archive of SID

www.SID.ir