یک مدل پیش بینی برای شناسایی مشتریان اعتباری بد حساب در بانک سامان

دکتر سیامک نوری

دانشیار گروه مدیریت سیستم و بهره وری، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران

snoori@iust.ac.ir

دکتر مسعود یقینی (نویسنده مسئول)

استادیار گروه حمل و نقل ریلی، دانشکده مهندسی راه آهن، دانشگاه علم و صنعت ایران

yaghini@iust.ac.ir

تکتم ژیان

کارشناسی ارشد مهندسی صنایع، مدیریت سیستم و بهره وری، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران

zhian_toktam@yahoo.com

چکیده

یکی از تصمیمات مهم در بانک ها و مؤسسات مالی و اعتباری، تصمیم گیری در مورد اعطای وام به مشتریان و کاهش ریسک اعتباری است. هدف این مقاله، ارائه مدلی مبتنی بر شبکه های عصبی پیش خور برای شناسایی مشتریان اعتباری بد حساب در بانک سامان است. جهت یافتن ساختاری مناسب برای شبکه عصبی در مدل پیشنهادی، از سه استراتژی سریع، پویا و چندگانه استفاده شده است. سه طرح یادگیری از نسبت های مختلف داده های آموزشی، آزمایشی و اعتبارسنجی تشکیل شده و شبکه های عصبی مورد استفاده، در پیکربندی و تعداد لایه های پنهان با یکدیگر اختلاف دارند. در این پژوهش از متدولوژی داده کاوی CRISP استفاده شده است. داده های مورد استفاده در این پژوهش، داده های مربوط به مشتریان اعتباری بانک سامان طی سالهای 1379 الی 1387 است. برای آماده سازی داده ها، پیش پردازش کاملی روی داده ها صورت گرفته است. جهت پیش گیری از بیش برازش مدل با مشخصات داده های آموزشی، بر اساس روش اعتبارسنجی تقاطعی، داده ها به سه قسمت داده های آموزشی، آزمایشی و اعتبارسنجی تقسیم گردیدند. برای ارزیابی مدل پیشنهادی، نتایج حاصل از استراتژی ها و طرح های مختلف در شبکه ها با یکدیگر و با برخی از روش های رایج پیش بینی نظیر درخت تصمیم و رگرسیون لجستیک مقایسه گردیده است. نتایج حاصل نشان می دهد که شبکه عصبی سه لایه تحت الگوریتم یادگیری پس انتشار و با استراتژی سریع و طرح یادگیری اول از دقت بالاتری برخوردار است.

کلید واژه:

پیش بینی، شبکه عصبی، ریسک اعتباری، وضعیت بازپرداخت، بانک سامان.

مقدمه

با توجه به اینکه حیات صنعت بانکداری در گرو پذیرا شدن ریسک است، پرهیز از آن ممکن نبوده و تنها بایستی آنرا مدیریت کرد. مدیریت ریسک، نظام و فرآیندی حرفه ای است که هدف اصلی آن بهبود کیفیت تصمیم ها در کلیه سطوح بنگاه های اقتصادی از جمله بانک ها به منظور افزایش ثروت سهامداران است. ریسک در یک مؤسسه مالی همان عدم اطمینان نسبت به بازده مورد انتظار دارایی هاست و از جمله مهمترین اقلام تشکیل دهنده یک بانک، اعتباراتی است که به مشتریان حقیقی و حقوقی پرداخت می شود. لذا بانک ها برای کاهش ریسک و در راستای جذب مشتریان کم ریسک بایستی قبل از اعطای وام، احتمال هر نوع انتخاب نامناسب را به حداقل برسانند.

ریسک اعتباری**1** نتیجه احتمال نکول یا احتمال عدم بازپرداخت وام توسط وام گیرنده است که این ریسک همان زیان انتظاری است که قابل پیش بینی می

باشد (Saunders & Anthony, 1999). ارزیابی ریسک اعتباری عنوان مهمی در مدیریت ریسک مالی است و صنعت مالی و بانکداری تمرکز زیادی روی این عنوان دارد. روش های داده کاوی، بخصوص الگوهای دسته بندی با استفاده از داده های مشتریان گذشته، اهمیت زیادی در ساخت مدل های پیش بینی دارند (Lu,Wang & Lai, 2008).

این مدل های پیش بینی به دو گروه متمایز تقسیم می شوند. گروه اول این مدل ها برای دسته بندی متقاضیان اعتباری جدید بر اساس ریسک اعتباری آنها، به کار می روند. داده های مورد استفاده در این مدل ها شامل اطلاعات مالی و اطلاعات شخصی متقاضی وام است. گروه دوم مدل ها به پیش بینی رفتار مشتریان موجود بر اساس داده های حاصل از پرداخت گذشته آنان می پردازد (Laha, 2007; Li et al., 2006; Vellido et al., 2008).

در این پژوهش از مدل پیش بینی برای دسته بندی متقاضیان اعتباری جدید و پیش بینی وضعیت بازپرداخت آنها بر اساس تکنیک شبکه های عصبی استفاده شده است. جهت کسب دقت بالا در شبکه های عصبی از انواع مختلف شبکه، استراتژی ها و طرح های یادگیری مختلف استفاده گردید. در نهایت نتایج حاصل از شبکه های عصبی با برخی از تکنیک های دیگر نظیر درخت تصمیم و رگرسیون لجستیک مقایسه شده است.

ساختار مقاله به این صورت سازمان دهی شده است. در بخش 1 به مرور ادبیات موضوع می پردازیم. در بخش 2 به تشریح داده های جمع آوری شده مشتریان اعتباری بانک سامان پرداخته خواهد شد. در بخش 3 نخست توضیح مختصری در مورد ساختار شبکه عصبی مورد استفاده و نحوه بروز رسانی وزن ها در آن داده می شود، سپس به تشریح استراتژی های مختلف برای یافتن ساختاری مناسب برای مدل پیشنهادی پرداخته می شود. در بخش 4 ابتدا به بررسی طرح های یادگیری بر اساس نسبت های مختلف داده های آموزشی، آزمایشی و اعتبارسنجی می پردازیم، سپس نتایج حاصل از مدلسازی تشریح خواهد شد و در بخش 5 به جمع بندی مطالب و نتیجه گیری نهایی و همچنین ارائه پیشنهادهایی برای تحقیقات آینده پرداخته می شود.

مرور ادبیات موضوع

بررسی های انجام شده بر روی ادبیات موضوع نشان می دهد که از تکنیک های مختلفی برای پیش بینی ریسک اعتباری استفاده شده است. تحلیل تک متغیره روشی است که اولین بار توسط بیور مطرح شد (Beaver, 1966). تحلیل نسبت ها در این روش ناقص و بالقوه گیج کننده است و پیش بینی وضعیت مالی به عوامل چند بعدی بستگی دارد. به دلیل این معایب، تحلیل تک متغیره به وسیله تحلیل چند متغیره جایگزین شد (Altman,1968). تحلیل چند متغیره بر اساس این فرض عمل می کند که ماتریس های واریانس و کوواریانس برای هر نوع بنگاه یکسان هستند و تأکید اصلی این روش بیشتر به تعیین مرز دو گروه متمرکز است. بسیاری از مطالعات نشان دادند که اغلب این فرض ها، به وسیله داده های مورد بررسی نقض می شوند (Karles, 1968). اولسون برای طبقه بندی مشتریان خوب و بد از رگرسیون لجستیک استفاده کرد. به تدریج در سال 1980 رگرسیون لجستیک و برنامه ریزی خطی به عنوان دو ابزار قوی در اعتبارسنجی معرفی شدند (Ahlson, 1980).

در جستجو برای مدل پیش بینی ریسک اعتباری با کمترین فرض محدود کننده، محققان استفاده از مدل های احتمال شرطی مانند توزیع احتمال خطی**2**، مدل لاجیت**3** و مدل پرابیت**4** را پیشنهاد کردند. موریس (Morris, 1998) در تحقیقات خود به این نتیجه رسید که نتایج حاصل از این مدل ها، هنگامی که تعداد متغیرهای توضیح دهنده خیلی زیاد و متغیرها پیوسته اند، تحت تأثیر واقع می شود. به هر حال مدل های لاجیت و پرابیت از نظر محاسباتی از مدل های آنالیز تمایزی، مشکل ترند. مشکل اصلی این مدل ها، به کار بردن یک زمان طولانی و معقولی از سری های زمانی است. این مدل ها در معرض محدودیت های اقتصادسنجی مانند دوره کوتاهتر دسترسی به سری زمانی داده های نکولی قرار می گیرند. به تدریج

سیستم های خبره و هوش مصنوعی وارد این حوزه شدند. شبکه های عصبی، الگوریتم ژنتیک و درخت های تصمیم از روش های موجود در این زمینه هستند.

استفاده از شبکه های عصبی در کاربردهای کسب و کار، اخیراً گسترش زیادی پیدا کرده است. نتایج تحقیقات مرتبط نشان می دهد که شبکه های عصبی در میان سایر تکنیک ها، ابزاری دقیق برای تحلیل ریسک اعتباری است (Min & Lee, 2008). لیم و سون (Lim & sohn, 2007) مدل رتبه بندی رفتاری مبتنی بر شبکه عصبی پیشنهاد کردند که مدل پیشنهادی آنها تغییرات مشخصه های متقاضیان را بعد از دریافت وام بصورت پویا وارد شبکه می کند و می تواند جهت کاهش زیان ناشی از متقاضیان بد، جایگزین مدل استاتیک گردد. در سال 2007 مروری بر تکنیک های استخراج قانون از ماشین های بردار پشتیبان**5**، جهت ارزیابی ریسک اعتباری انجام گرفت (Mortens et al., 2007). نتایج این تحقیق دو تکنیک استخراج قانون با استفاده از شبکه های عصبی مصنوعی را پیشنهاد کرد.

هوانگ و همکاران (Huang et al., 2007) مدل های ترکیبی رتبه بندی اعتباری مبتنی بر ماشین های بردار پشتیبان برای ارزیابی متقاضیان اعتباری با استفاده از مشخصه های آنان پیشنهاد کردند. در این پژوهش از مجموعه داده اعتباری در آلمان و استرالیا استفاده شده است. ابدو و همکاران (Abdou et al., 2008) توانایی شبکه های عصبی احتمالی و شبکه های عصبی پیش خور چندلایه را با روش های سنتی نظیر تحلیل تمایز، تحلیل پرابیت و رگرسیون لجستیک در ارزیابی ریسک اعتباری بانک های مصری با بکارگیری مدل های رتبه بندی اعتباری مقایسه کردند. نتایج حاصل از این پژوهش نشان داد که میانگین دقت شبکه های عصبی از سایر روش ها بهتر است.

در سال 2008 کاربرد شبکه های عصبی برای ارزیابی ریسک اعتباری کسب و کارهای کوچک در ایتالیا مورد بررسی قرار گرفت. این مطالعه دو ساختار از شبکه های عصبی شامل شبکه عصبی پیش خور و ساختار پیشنهادی شبکه عصبی را در زمینه ارزیابی ریسک اعتباری پیشنهاد کرد. هر دو شبکه دقت بالایی در پیش بینی احتمال نکول مشتریان دارند (Angelini, Ditollo & Roli, 2008). یو و همکاران (Yu et al., 2008) مدل یادگیری چند مرحله ای شبکه های عصبی را برای ارزیابی ریسک اعتباری پیشنهاد کردند. مدل پیشنهادی شامل شش مرحله است: 1) تولید زیرمجموعه های مختلفی از داده های آموزشی، 2) ایجاد مدل های مختلفی از شبکه های عصبی بر اساس زیرمجموعه های بدست آمده در مرحله قبل، 3) آموزش مدل های شبکه عصبی توسط داده های آموزشی و کسب نمره رتبه بندی، 4) انتخاب مجموعه اعضاء مناسب، 5) انتخاب مقادیر قابلیت اطمینان از شبکه عصبی انتخاب شده، 6) ترکیب مجموعه اعضاء شبکه عصبی انتخاب شده و نتایج حاصل از دسته بندی نهایی توسط اندازه گیری قابلیت اطمینان.

تسی و وو (Tsai & Wu, 2008) عملکرد دسته بندی کننده واحد را به عنوان دسته بندی کننده پایه با دسته بندی کننده های چندگانه توسط شبکه های عصبی و با استفاده از سه مجموعه داده مقایسه کردند. نتایج نشان داد که دسته بندی کننده جمعی تنها در یک مجموعه داده برتر از دسته بندی کننده واحد عمل می کند. ستینو و همکاران (Setiono et al., 2008) از الگوریتم بازگشتی برای استخراج قوانین دسته بندی از شبکه های عصبی پیش خور که با مجموعه داده اعتباری آموزش داده شده بودند، استفاده کردند.

لین (Lin, 2009) از مدل ترکیبی شبکه های عصبی و رگرسیون لجستیک در ارزیابی ریسک مالی بانک های تایوان استفاده کرد. وانگ و هوانگ (Wang & Huang, 2009) از شبکه های عصبی پس انتشار برای دسته بندی متقاضیان اعتباری استفاده کردند. ژو و همکاران (Xu, et al., 2009) از الگوریتم رتبه بندی اعتباری بر پایه ماشین های بردار پشتیبان برای تصمیم گیری در مورد واگذاری وام به متقاضیان استفاده کردند. چوانگ و لین (Chuang & Lin, 2004) از مدل رتبه بندی اعتبار با تخصیص مجدد استفاده نمودند. این مدل پیشنهادی شامل دو مرحله است. در مرحله اول، مدل رتبه بندی اعتباری مبتنی بر شبکه های عصبی برای دسته بندی متقاضیان به دسته های قبول شده (خوب) و رد شده (بد) ایجاد می گردد. مرحله دوم، مرحله تخصیص مجدد است که سعی در کاهش خطای نوع اول با تخصیص مجدد مشتریان خوب رد شده به

دسته مشتریان خوب شرطی دارد. کروک و همکاران (Crook et al., 2007) مقایسه ای بین روش های مختلف دسته بندی متقاضیان اعتباری انجام داده اند. در مقایسه انجام شده، شبکه های عصبی از دقت بالای در پیش بینی برخوردار بوده اند.
در این پژوهش برای پیش بینی وضعیت مشتریان از شبکه های عصبی پرسپترون دو لایه، سه لایه و چهار لایه تحت الگوریتم یادگیری پس انتشار خطا استفاده شده است. در این شبکه ها از سه استراتژی سریع، پویا و چندگانه برای پیدا کردن پیکربندی مناسب شبکه ها استفاده می شود. سپس دقت این شبکه را با روش های درخت تصمیم و رگرسیون لجستیک مورد مقایسه قرار می دهیم. پس از مقایسه نتایج بدست آمده از مدل های مختلف، بهترین مدل به عنوان مدل پیشنهادی انتخاب می گردد.

شناخت داده ها

مجموعه داده مورد استفاده در این تحقیق مربوط به مشتریان حقیقی بانک سامان است که طی سال های 1379 الی 1387 تسهیلات دریافت نموده اند. این مجموعه داده شامل 20 مشخصه است که عنوان، نوع و شرح هر مشخصه در جدول 2-1 نشان داده شده است.

جدول (2-1): مشخصه های مجموعه داده مشتریان اعتباری بانک سامان

ردیف	نام مشخصه	نوع	شرح	محدوده مقادیر مجاز
1	**Customer No.**	**Numeric**	شماره منحصر بفرد هر مشتری در بانک	[۱.۹۹۹۹۹۹۹۹]
2	**Birth Date**	**Numeric**	تاریخ تولد مشتری	[۱۳۵۰.۱۳۶۹]
3	**Entrance Date**	**Numeric**	تاریخ ورود و اولین افتتاح حساب مشتری	[۱۳۷۹.۱۳۸۷]
4	**Gender**	**Boolean**	جنسیت مشتری	مذکر و مؤنث
5	**Occupation**	**Categorical**	شغل مشتری	...
6	**Education**	**Categorical**	سطح تحصیلا ت مشتری	بیسواد، ابتدایی، سیکل، زیر دیپلم، دیپلم، دانشجو، لیسانس، فوق لیسانس، دکترا و فوق دکترا
7	**Marital Status**	**Boolean**	وضعیت تأهل مشتری	متأهل و مجرد
8	**Loan No.**	**Numeric**	شماره تسهیلات از 4 بخش تشکیل شده است. در شماره تسهیلات a-b-c-d ، a بیانگر کد شعبه ای است که مشتری از آن شعبه وام دریافت کرده است، b بیانگر کد نوع تسهیلات، c شماره مشتری و d تعداد تسهیلات دریافتی از یک شعبه را نمایش می دهد.	...
9	**Situation Of Loan**	**Categorical**	وضعیت فعلی تسهیلات در سیستم	تسویه شده، سررسید گذشته، معوق، مشکوک الوصول، فعال، آماده برای آزادسازی و آماده برای فعال سازی
10	**Way Of Repayment**	**Categorical**	نحوه بازپرداخت وام توسط مشتری	یکجا، قسطی، ارزش آتی، پلکانی، تدریجی و رأس المدتی گردان
11	**Beginning Date**	**Numeric**	تاریخ شروع قرارداد تسهیلاتی	[۱۳۷۹.۱۳۸۷]
12	**Final Date**	**Numeric**	تاریخ پایان قرارداد تسهیلاتی	[۱۳۷۹.۱۳۸۷]
13	**Interest**	**Numeric**	سود تسهیلات دریافت شده توسط مشتریان در موقع عقد قرارداد	[۵.۳۰]
14	**Wage**	**Numeric**	کارمزد تسهیلات دریافت شده توسط کارکنان	[۱.۱۰]
15	**ISICCODE**	**Categorical**	کد مربوط به تقسیم بندی بین المللی انواع تسهیلات	طبق استاندارد
16	**ISICDESC**	**Categorical**	شرح کد **ISICCODE**	طبق استاندارد
17	**Type Of Assurance**	**Categorical**	انواع وثیقه دریافت شده توسط بانک	25 نوع وثیقه
18	**Time Of Contract. D**	**Numeric**	مدت مجاز برای بازپرداخت وام (به روز)	[۱۰.۱۰۸۰۰...]
19	**Time Of Contract. M**	**Numeric**	مدت مجاز برای بازپرداخت وام (به ماه)	[۱.۳۶۰]
20	**Amount Of Loan**	**Numeric**	مبلغ مصوب تسهیلات مشتری	500000الی 150میلیارد ریال

لازم به ذکر است که انواع وثایق دریافت شده توسط بانک برای دادن اعتبار به مشتری شامل، اسناد عادی، اقرارنامه، اموال منقول در رهن بانک، اوراق مشارکت، بیمه نامه اموال در رهن بانک، تعهدنامه، چک بابت حق بیمه، چک تضمینی، سپرده سرمایه گذاری مدتدار، سفته تضمینی، سند اجاره به شرط تملیک، سند ملکی ترهینی، سند ملکی مبایعه نامه، سند ملکی مشارکت مدنی، سهام، ضمانت نامه بانکی، ورقه مبایعه نامه، وکالتنامه، بیمه نامه اعتباری می باشد.

این پایگاه داده شامل 82093 نمونه از وام های دریافتی مشتریان حقیقی است. به دلیل اینکه تمامی مشخصه ها در مدلسازی قابل استفاده نیستند، و با توجه به احساس نیاز به تغییراتی در برخی از مشخصه ها، در این مرحله ابتدا به گام های پاکسازی داده ها و ساخت مشخصه های جدید می پردازیم تا پیش از ورود به مرحله مدلسازی، داده ها آماده گردند.

با بررسی اولیه مشخصه ها، با توجه به اینکه مشخصه های سود و کارمزد ماهیت یکسانی دارند و هر یک به تنهایی دارای مقادیر مفقوده زیادی می باشند، این دو مشخصه با هم ترکیب و ادغام می شوند. مشخصه شماره مشتری به دلیل منحصر بفرد بودن از مجموعه داده حذف گردید. نتایج حاصل از تبدیل و ترکیب سایر مشخصه ها در جدول 2-2 نشان داده شده است.

جدول (2-2): ساخت مشخصه های جدید از مشخصه های موجود

ردیف	مشخصه قبلی	مشخصه جدید
1	***Birth date***	***Age***
2	***Entrance Date***	***Background***
3	***Entrance Date***	***Entrance Year***
4	***Loan No.***	***City***
5	***Loan No.***	***Type Of Loan***
6	***Beginning Date***	***Beginning Year***
7	***Final Date***	***Final Year***
8	***Beginning & final Date***	***Time Of contract***

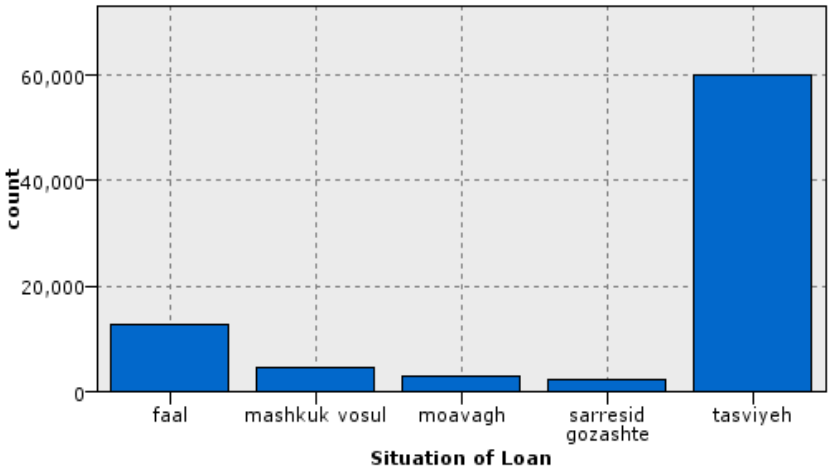
بطور کلی تغییرات داده شده به شرح ذیل است:

- مشخصه سن مشتری از تاریخ تولد مشتری استخراج گردید.
- مشخصه های سابقه و سال ورود از تاریخ ورود مشتری به بانک استخراج گردید.
- با توجه به توضیح ذکر شده در مورد شماره تسهیلات، مشخصه های شهر و نوع اصلی وام از شماره تسهیلات استخراج گردید.
- مشخصه سال شروع قرارداد تسهیلاتی از تاریخ شروع و مشخصه سال پایان قرارداد تسهیلاتی از تاریخ پایان استخراج گردید.
- مشخصه مدت قرارداد تسهیلات از مشخصه های تاریخ شروع و پایان استخراج گردید.

در مرحله بعد داده ها از نظر کیفیت و اعتبار مورد بررسی قرار گرفته و از روش های مختلفی به منظور پاکسازی داده های مغشوش استفاده گردیده است.

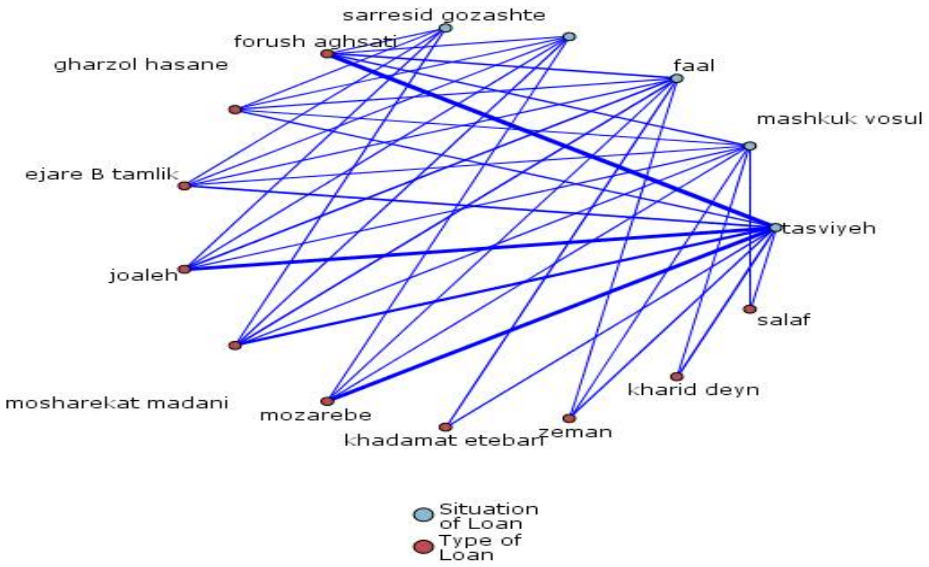
در مشخصه های سن، سابقه و مدت قرارداد با توجه به محدود بودن مقادیر مفقوده و مغشوش از روش های جایگزینی دستی مقادیر، استفاده شده است. مشخصه شغل به دلیل تعداد بسیار زیاد مقادیر مفقوده و عدم دسترسی به این مقادیر، بطور کلی از پایگاه داده حذف گردید. مقادیر مفقوده سایر مشخصه ها، با استفاده از الگوریتم درخت تصمیم C&RT جایگزین گردیده اند.

با توجه به اینکه هدف این پژوهش پیش بینی وضعیت بازپرداخت مشتریان است، شکل 2-1 نحوه توزیع مقادیر این متغیر را نشان می دهد. همان طور که از نمودار بر می آید اکثر وام ها در وضعیت تسویه شده و فعال قرار دارند، تعداد محدودی از این وام ها در حالت های ریسکی مشکوک الوصول، معوق و سررسید گذشته قرار دارند.

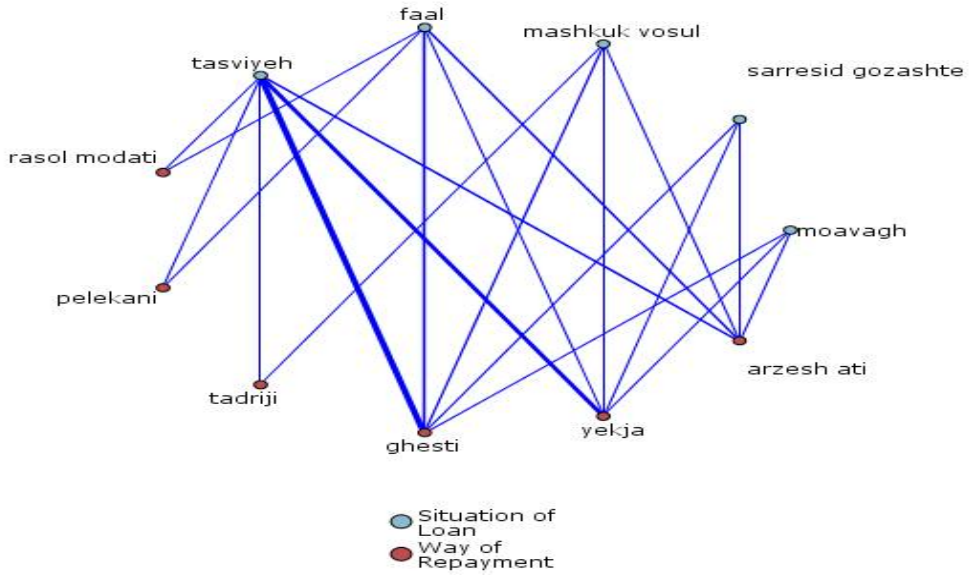


شکل (1): نمودار هیستوگرام وضعیت تسهیلات

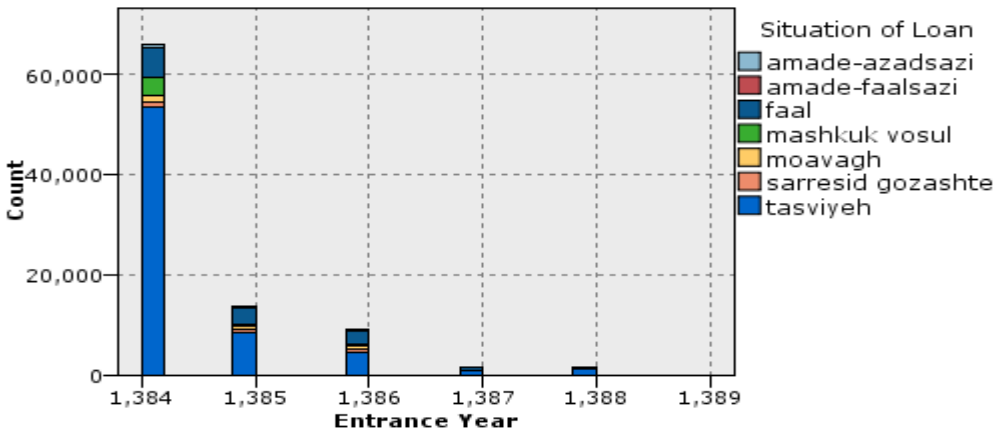
شکل 2-2 و 2-3 به ترتیب ارتباط بین مشخصه وضعیت تسهیلات را با نوع اصلی وام و روش بازپرداخت نشان می دهد. این دو شکل بیانگر این موضوع است که اکثر وام های تسویه شده مربوط به تسهیلاتی از نوع فروش اقساطی، جعاله و مضاربه و با نحوه بازپرداخت قسطی و یکجا می باشند. بیشتر تسهیلات مشکوک الوصول از نوع فروش اقساطی، تسهیلات معوق و سررسید گذشته از نوع مشارکت مدنی هستند. شکل 2-4 نشان می دهد که اکثر وام های تسویه شده، فعال، معوق، سررسید گذشته و مشکوک الوصول مربوط به مشتریانی است که در سال 1384 وارد بانک شده اند.



شکل (2): رابطه بین مشخصه وضعیت تسهیلات و نوع اصلی وام



شکل (3): رابطه بین مشخصه وضعیت تسهیلات و روش بازپرداخت



شکل (4): نمودار هیستوگرام متغیر سال ورود

مدل شبکه عصبی

3-1- پیکربندی شبکه عصبی

شبکه عصبی مصنوعی بعنوان یکی از تکنیک‌های مهم در یادگیری نظارتی و یادگیری غیر نظارتی در داده کاوی بشمار می آید. شبکه عصبی مورد استفاده در این مقاله، شبکه عصبی پیشخور**6** است که به آن پرسپترون چند لایه**7** نیز می گویند. برای آموزش شبکه از الگوریتم پس انتشار**8** خطا بهمراه اندازه حرکت**9**، استفاده شده است. تابع فعال سازی مورد استفاده برای این شبکه، تابع سیگموید**10** بصورت رابطه (1) است.

$$\sigma(x) = 1/1+e^{-x} \qquad (1)$$

در ابتدای یادگیری تمام وزن‌ها در شبکه بصورت تصادفی در فاصله $-0.5 \le w_{ij} \le 0.5$ انتخاب می‌شوند. نحوه بروز رسانی وزن‌ها بعد از ورود هر نمونه از مجموعه داده‌ها بصورت رابطه (2) است:

$$\Delta w_{ij}(t+1) = \eta \delta_{pj} o_{pi} + \alpha \Delta w_{ij}(t) \qquad (2)$$

که در آن پارامتر η نرخ یادگیری، δ_{pj} خطای انتشار یافته، o_{pi} خروجی نرون i برای نمونه p، پارامتر α اندازه حرکت و $\Delta w_{ij}(t)$ مقدار تغییر w_{ij} در تکرار قبلی است. مقدار α در حین فرآیند یادگیری ثابت است ولی اندازه η در طول تکرارهای فرآیند یادگیری تغییر می‌کند. در ابتدا η بصورت یک مقدار اولیه داده می‌شود و سپس بصورت لگاریتمی به مقدار η_{low} کاهش می‌یابد، و هنگامی که مقدار η کمتر از مقدار η_{low} شد، آنگاه مقدار η_{high} را برای آن در نظر می‌گیریم، یعنی اگر $\eta(t-1) < \eta_{low}$ آنگاه $\eta(t) = \eta_{high}$. تابع کاهش لگاریتمی برای پارامتر نرخ یادگیری بصورت رابطه (3) است.

$$\eta(t) = \eta(t-1).\exp(\log(\eta_{low}/\eta_{high})/d) \qquad (3)$$

که در آن d مقداری است که توسط کاربر تنظیم می‌شود و به آن زوال η می‌گوئیم. این فرآیند تا پایان زمان یادگیری ادامه می‌یابد. مقدار خطا δ_{pj} برای لایه خروجی بصورت رابطه (4) محاسبه می شود.

$$\delta_{pj} = (t_{pj} - o_{pj}) o_{pj} (1 - o_{pj}) \qquad (4)$$

برای دیگر لایه ها بصورت رابطه (5) می باشد.

$$\delta_{pj} = o_{pj} (1 - o_{pj}) \sum_k \delta_{pk} w_{kj} \qquad (5)$$

که t_{pj} مقدار خروجی مورد انتظار برای پیش بینی است. وزن‌های شبکه بعد از پیش بینی خروجی هر نمونه به روز می شوند.

3-2- ساختار شبکه عصبی

یکی از مباحث چالش بر انگیز در بحث شبکه‌های عصبی یافتن ساختار بهینه شبکه براساس داده‌های موجود است. برای غلبه بر این مشکل در این تحقیق چندین روش متفاوت استفاده شده است که در نهایت روشی با بیشترین دقت بعنوان روش برتر انتخاب شده است.

3-2-1- استراتژی سریع

در روش سریع فقط یک شبکه عصبی از نوع پرسپترون سه لایه به صورت مجزا آموزش داده می‌شود. این شبکه بطور پیش فرض دارای یک لایه مخفی با $\max\left(\left(\Theta_i+\Theta_o\right)/20,3\right)$ نرون است که Θ_i تعداد نرون‌ها در لایه ورودی و Θ_o تعداد نرون‌ها در لایه خروجی هستند. وزن های شبکه با استفاده از الگوریتم آموزشی پس انتشار بروز رسانی می شوند.

3-2-2- استراتژی پویا

در روش پویا ساختار شبکه در طول فرآیند یادگیری تغییر می‌کند و نرون‌ها به شبکه اضافه می‌شوند تا کارایی شبکه بالاتر رفته و به دقت مطلوب برسد. این مرحله خود دارای دو زیر مرحله‌ی یافتن ساختار مناسب و سپس آموزش شبکه نهایی است. برای یافتن ساختار مناسب ابتدا شبکه‌ای با دو لایه مخفی ساخته می شود، که هرکدام دارای دو نرون هستند. نرخ یادگیری اولیه را 0,05 و α = 0,9 درنظر می گیریم و شبکه را به اندازه یک تکرار آموزش می دهیم و یک کپی از شبکه ایجاد کرده و یکی را راست و دیگری را چپ می نامیم، سپس به دومین لایه مخفی شبکه سمت راست یک نرون اضافه کرده و مجدداً هر دو شبکه را به اندازه یک تکرار آموزش می دهیم و مجموع خطا در هردو شبکه را بدست می آوریم. در صورتی که شبکه سمت چپ خطای کمتری داشته باشد آن را نگه داشته و یک نرون به اولین لایه مخفی شبکه سمت راست اضافه می کنیم. درصورتی که شبکه سمت راست خطای کمتری داشته باشد یک نسخه از شبکه سمت راست را به جای شبکه سمت چپ قرار می دهیم و مجدداً یک نرون به دومین لایه مخفی شبکه سمت راست اضافه می کنیم. هر دو شبکه را به اندازه یک تکرار آموزش می دهیم و این فرآیند را تا رسیدن به شرط خاتمه تکرار می کنیم. برای تنظیم نرخ یادگیری در هر تکرار دو بردار محاسبه می‌شود اولی بردار حرکت $M(t)$ بر مبنای تغییرات وزن‌ها در یک تکرار است و دومی بردار تغییر $C(t)$ که بر مبنای اندازه حرکت تکرار فعلی است. بردارهای $M(t)$ و $C(t)$ از روابط زیر محاسبه می‌شود:

$$M(t)=2\left[W(t)-W(t-1)\right] \quad (6)$$

$$C(t)=0.8C(t-1)-M(t) \quad (7)$$

که در آن $W(t)$ بردار وزن تکرار فعلی و $W(t-1)$ بردار وزن تکرار قبلی است. بردار نسبت بزرگی این دو بردار را به صورت زیر تعریف می کنیم:

$$m(t)=\|M(t)\|/\|C(t)\| \quad (8)$$

این مقدار شاخص شتاب یادگیری است در صورتی که این شاخص کمتر از $1+\|C(t)\|/10$ باشد یادگیری در حال کند شدن است و نرخ یادگیری در عدد 1,2 ضرب می شود و اگر شاخص بیشتر از 5 باشد یادگیری در حال شتاب گرفتن است و نرخ یادگیری در عدد $4/m(t)$ ضرب می‌شود. بعد از یافتن ساختار مناسب در روش پویا برای شبکه عصبی، نوبت به آموزش شبکه نهایی براساس روش پس انتشار خطا می رسد برای شروع آموزش نرخ یادگیری اولیه را برابر 0,02 و α = 0,9 و بر این اساس شبکه را آموزش می دهیم.

3-2-3- استراتژی چندگانه

در روش چندگانه تا رسیدن به شرط خاتمه، چندین شبکه بصورت شبه موازی باهم آموزش می بینند و شبکه با بالاترین دقت بعنوان شبکه نهایی انتخاب می شود. به این صورت که در ابتدا چندین شبکه با یک لایه مخفی ساخته می شود، که تعداد نرون‌ها در لایه مخفی از 3 تا حداکثر تعداد نرون‌های لایه ورودی تغییر می کند. سپس به ازای هر شبکه با یک لایه مخفی چندین شبکه با دو لایه مخفی ایجاد می کنیم که تعداد نرون‌های لایه مخفی اول درست به اندازه شبکه یک لایه‌ای است، ولی تعداد نرون‌های لایه مخفی دوم متغیر است و به ترتیب 2، 5، 10، 17 و به همین ترتیب حداکثر تا تعداد نرون‌های لایه مخفی اول در لایه مخفی دوم می توان نرون داشته باشیم و با ساخت این شبکه‌ها، آنها را آموزش می دهیم (SPSS Clementine, 2008).

پیش بینی وضعیت بازپرداخت مشتریان

یکی از مشکلات رایج هنگام آموزش مدل های پیش بینی، تطبیق بیش از حد مدل با خصوصیات داده های آموزشی فعلی یا به عبارتی بیش برازش است، که این مسئله منجر به پیش بینی نادرست و با دقت کمتر از حد انتظار برای مقادیر جدید می شود. برای جلوگیری از چنین حالتی با توجه به حجم عظیم داده ها، با استفاده از روش اعتبارسنجی تقاطعی **11** داده ها به سه بخش داده های آموزشی، آزمایشی و اعتبارسنجی تقسیم شده است، سپس آموزش به صورت دوره ای متوقف شده و بعد از هر دوره آموزشی شبکه با استفاده از مجموعه داده اعتبارسنجی، مورد ارزیابی قرار می گیرد. با استفاده از این شیوه قادر به مشخص کردن شروع بیش برازش خواهیم بود. در این مقاله برای مقایسه تأثیر نسبت داده های آموزشی، آزمایشی و اعتبارسنجی در پیش بینی مدل از سه طرح یادگیری با نسبت های مختلف از این سه گروه داده ها استفاده خواهیم کرد. در جدول1 جزئیات این سه طرح یادگیری ارائه شده است.

جدول -1): جزئیات مربوط به طرح های یادگیری

	طرح یادگیری 1	طرح یادگیری 2	طرح یادگیری 3
داده های آموزشی	۶۰٪	۵۰٪	۷۰٪
داده های آزمایشی	۳۰٪	۴۰٪	۲۰٪
داده های اعتبارسنجی	۱۰٪	۱۰٪	۱۰٪

در نهایت شبکه عصبی با دقیق ترین استراتژی و طرح یادگیری به عنوان مدل نهایی انتخاب می گردد.

در پایگاه داده ای موجود، وضعیت تسهیلات مشتری شامل 5 طبقه بندی تسویه شده، فعال، سررسید گذشته، معوق و مشکوک الوصول است. لایه خارجی شبکه عصبی شامل 5 نرون است و هر یک از این حالات به عنوان یک نرون در لایه خروجی در نظر گرفته شده است. مدل پیشنهادی می تواند هر یک از این حالت ها را پیش بینی کند.

متغیرهای لایه ورودی در مدل شبکه عصبی پیشنهادی عبارتند از : سن، سال ورود، جنسیت، سطح تحصیلات، وضعیت ازدواج، نوع اصلی وام، روش بازپرداخت، سود، کد ISIC، مدت قرارداد، مبلغ مصوب، وثیقه نوع 1 الی وثیقه نوع 19، مجموع وثایق، پارتیشن بندی داده ها. با توجه به تعداد حالات هر یک از این متغیرها، لایه ورودی شبکه عصبی شامل 61 نرون می شود.

4-1- مقایسه نتایج پیش بینی شبکه های عصبی

نتایج حاصل از پیش بینی شبکه های عصبی دو لایه، سه لایه و چهار لایه، با سه استراتژی و سه طرح یادگیری مختلف با استفاده از نرون های ورودی و نرون های خروجی ذکر شده در بالا در جداول 2 الی 7 نشان داده شده است. کلیه نتایج مدل با استفاده از نرم افزار SPSS Clementine 12,0 به دست آمده است.

جدول (-2): نتایج مدل شبکه عصبی دو لایه بر روی داده ها

طرح یادگیری	استراتژی تعیین ساختار شبکه	دقت پیش بینی در داده های آموزشی (درصد)	دقت پیش بینی در داده های آزمایشی (درصد)	دقت پیش بینی در داده های اعتبارسنجی (درصد)	دقت کلی
اول	سریع	85,64	86,15	85,77	85,84
دوم	سریع	85,65	86,01	85,74	85,56
سوم	سریع	85,64	86,11	85,75	85,73

جدول (-3): نتایج مدل شبکه عصبی سه لایه بر روی داده ها

طرح یادگیری	استراتژی تعیین ساختار شبکه	دقت پیش بینی در داده های آموزشی (درصد)	دقت پیش بینی در داده های آزمایشی (درصد)	دقت پیش بینی در داده های اعتبارسنجی (درصد)	دقت کلی
اول	سریع	86,36	86,71	86,33	86,12
دوم	سریع	86,49	86,49	86,57	85,91
سوم	سریع	86,31	86,66	86,00	85,89

جدول (-4): نتایج مدل شبکه عصبی چهار لایه بر روی داده ها

طرح یادگیری	استراتژی تعیین ساختار شبکه	دقت پیش بینی در داده های آموزشی (درصد)	دقت پیش بینی در داده های آزمایشی (درصد)	دقت پیش بینی در داده های اعتبارسنجی (درصد)	دقت کلی
اول	سریع	85,55	85,78	84,70	84,97
اول	پویا	85,97	85,98	85,47	85,30
اول	چندگانه	86,16	86,46	86,20	85,35
دوم	سریع	86,06	86,27	85,99	85,70
دوم	پویا	85,65	85,81	85,42	85,55
دوم	چندگانه	86,38	86,54	86,03	85,94
سوم	سریع	86,45	86,82	86,58	85,80
سوم	پویا	85,73	85,90	85,54	85,60
سوم	چندگانه	86,22	86,77	86,34	85,95

با مقایسه سه جدول بالا مشاهده می شود که بالاترین دقت پیش بینی مربوط به شبکه عصبی سه لایه با استراتژی سریع و طرح یادگیری اول است. استراتژی پویا و چندگانه نیاز به حافظه بیشتر و همچنین زمان بیشتر برای اجرای الگوریتم دارد. کمترین دقت پیش بینی مربوط به شبکه عصبی چهار لایه با استراتژی سریع و طرح یادگیری اول می باشد. همان طور که نتایج بدست آمده نشان می دهد، نمی توان یک روش خاص را بطور مطلق بعنوان بهترین روش انتخاب کرد، بلکه باید تحلیلی روی دقت، زمان و میزان حافظه مورد نیاز هر یک از روش ها انجام داد.

تعداد پیش بینی های درست در هر یک از حالت های وضعیت تسهیلات (نرون های خروجی) توسط شبکه عصبی انتخابی، در جدول 5 نشان داده شده است. همچنین نتایج حاصل از پیش بینی وضعیت تسهیلات با استفاده از شبکه عصبی چهار لایه با استراتژی سریع و طرح یادگیری اول که در بین تمام شبکه ها، کمترین دقت را دارد، در جدول 5- ارائه شده است.

جدول (-5): نتایج پیش بینی وضعیت تسهیلات با استفاده از شبکه عصبی پیشنهادی

	پیش بینی وضعیت تسهیلات				
وضعیت تسهیلات	فعال	مشکوک الوصول	معوق	سررسید گذشته	تسویه شده
فعال	۷۳۰۲	۰	۰	۰	۵۲۹۱
مشکوک الوصول	۰	۳۵۱۸	۳۳۱	۵۱۳	۱
معوق	۱	۷۷۰	۷۳۲	۱۲۲۷	۱
سررسید گذشته	۱	۴۹۵	۳۵۱	۱۴۵۳	۰
تسویه شده	۲۱۳۱	۰	۰	۰	۵۷۹۷۵

جدول (-6): نتایج پیش بینی وضعیت تسهیلات با استفاده از شبکه عصبی با کمترین دقت

	پیش بینی وضعیت تسهیلات				
وضعیت تسهیلات	فعال	مشکوک الوصول	معوق	سررسید گذشته	تسویه شده
فعال	۷۹۰۳	۰	۰	۰	۴۶۹۰
مشکوک الوصول	۰	۳۰۹۹	۱۴۰	۱۱۲۴	۱
معوق	۰	۴۲۸	۱۹۱	۲۱۱۲	۱
سررسید گذشته	۰	۳۰	۸۸	۲۲۱۸۲	۰
تسویه شده	۳۲۶۵	۰	۰	۰	۵۶۸۴۱

برای ارزیابی نتایج بدست آمده از دو روش درخت تصمیم و رگرسیون لجستیک چندجمله ای**12** استفاده شده است. برای ساخت درخت تصمیم از الگوریتم CHAID**13** استفاده شده است. الگوریتم CHAID یک تکنیک کارآمد طبقه‌بندی است که از قابلیت های بالای یک تست آماری بعنوان معیاری جهت ارزیابی مقدار احتمالی یک مشخصه پیش بینی کننده، استفاده می‌کند. الگوریتم، مقادیری که با توجه به متغیر هدف، مشابه تشخیص می دهد با هم ادغام می کند و سایر مقادیر غیر مشابه را نگه می دارد. سپس الگوریتم بهترین پیش بینی کننده را انتخاب می کند تا اولین شاخه درخت را تشکیل دهد، بدین شکل که هر گره فرزند از گروهی از مقادیر همسان مشخصه انتخابی تشکیل شده است. این فرآیند تا جایی ادامه پیدا می کند که درخت رشد کامل کرده باشد. تست آماری مورد استفاده بستگی به سطح اندازه گیری مشخصه کلیدی دارد. اگر مشخصه

هدف پیوسته باشد، از تست F استفاده می‌شود و اگر گسسته باشد تست کای- مربع مورد استفاده قرار می گیرد (Witten & Frank, 2005). نتایج حاصل از پیش بینی با درخت تصمیم و رگرسیون لجستیک در جدول 5-7 ارائه شده است.

جدول (7: نتایج پیش بینی با درخت تصمیم و رگرسیون لجستیک

الگوریتم	طرح یادگیری	دقت پیش بینی در داده های آموزشی (درصد)	دقت پیش بینی در داده های آزمایشی (درصد)	دقت پیش بینی در داده های اعتبارسنجی (درصد)
درخت تصمیم	اول	83,71	83,70	83,23
درخت تصمیم	دوم	83,70	83,72	83,23
درخت تصمیم	سوم	83,71	83,73	83,23
رگرسیون لجستیک	اول	84,95	85,33	84,67
رگرسیون لجستیک	دوم	84,90	85,25	84,62
رگرسیون لجستیک	سوم	85,02	85,42	84,64

با مقایسه نتایج حاصل از دو روش درخت تصمیم و رگرسیون لجستیک با شبکه های عصبی مشاهده می شود که شبکه های عصبی دقت بالاتری در پیش بینی وضعیت بازپرداخت مشتریان دارند. البته قابل ذکر است که روش های مورد استفاده در این تحقیق از نظر دقت پیش بینی اختلاف زیادی با هم ندارند ولی از نظر زمان پردازش و ساخت مدل تفاوت زیادی بین روش های مختلف وجود دارد. در نهایت مدل شبکه عصبی دو لایه با استراتژی سریع و طرح یادگیری اول به عنوان مدل انتخابی این تحقیق انتخاب می گردد.

نتیجه گیری

در این مقاله مدلی مبتنی بر شبکه های عصبی با دقت بالا برای پیش بینی وضعیت بازپرداخت مشتریان اعتباری بانک سامان ارائه شده است. برای تعیین ساختار بهینه مدل، از سه شبکه با تعداد لایه ها و نرون های مختلف، سه استراتژی و سه طرح یادگیری مختلف استفاده شده است. جهت ساخت مدل پیش بینی، از اطلاعات مشتریان اعتباری بانک طی سال های 1379 الی 1387 استفاده شده است. برای ارزیابی مدل پیشنهادی، نتایج بدست آمده با نتایج حاصل از دو روش دیگر شامل درخت تصمیم و رگرسیون لجستیک مقایسه شده است. نتایج حاصل نشان می دهد که مدل شبکه عصبی دو لایه با استراتژی سریع و طرح یادگیری اول (با نسبت □□□ داده های آموزشی، □□□ داده های آزمایشی و □□□ داده های اعتبارسنجی) بالاترین دقت را در میان سایر مدل ها دارد. با استفاده از مدل ارائه شده می توان پیش بینی دقیقی از وضعیت بازپرداخت مشتریان داشته و نسبت به برنامه ریزی مناسب جهت کاهش ریسک اعتباری بانک اقدام نمود.

زمینه های مختلفی برای ادامه این مطالعه می توان معرفی نمود. برای بهبود دقت پیش بینی می‌توان از روش‌های فرا ابتکاری نظیر الگوریتم ژنتیک یا تبرید شبیه‌سازی شده برای یافتن ساختار بهینه شبکه عصبی استفاده نمود. برای افزایش سرعت یادگیری شبکه، می‌توان از الگوریتم‌های فرا ابتکاری استفاده کرد. همچنین می توان از مدل های ترکیبی جهت پیش بینی دقیق وضعیت بازپرداخت مشتریان استفاده نمود. سیستم های فازی نیز در این زمینه می توانند مورد استفاده قرار گیرند.

منابع

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Elsevier Inc., 2006.

[2] D.T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, Inc., 2006.

[3] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Elsevier Inc., 2005.

[4] S. J. Russell and P. Norvig , Artificial Intelligence, A Modern Approach, Prentice Hall, 1995.

[5] A. Saunders and L. Allen, Credit Risk Measurement: New Approaches to Value At Risk and Other Paradigms, 2nd Edition, Wiley Finance, 1999.

[6] N.C. Hsieh, "Hybrid mining approach in the design of credit scoring models", Expert Systems with Applications 28, 2005, pp. 655-665.

[7] N.C. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customer", Expert Systems with Applications 27, 2004, pp. 623-633.

[8] Y.M. Huang, C.M. Huang, and H.C. Jiau, "Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem", Nonlinear Analysis: Real world Application 7, 2006, pp. 720-747.

[9] I.C. Yeh and C.H. Lien,"The comparison of data mining techniques for the predictive accuracy of probability of default of credit card clients", Expert Systems with Applications 36, 2009, pp. 2473-2480.

[10] A. Khashman, " Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes", Expert Systems with Application, 2010.

[11] C.F. Tsai and J.W. Wu, "Using Neural Network ensembles for bankruptcy prediction and credit scoring", Expert Systems with Applications 34, 2008, pp. 2639-2649.

[12] C.L. Chuang and R.H. Lin, "Constructing a reassigning credit scoring model", Expert Systems with Applications 36, 2009, pp. 1685-1694.

[13] E. Angelini, G.D. Tollo, and A. Roli, "A neural network approach for credit risk evaluation", the Quarterly Review of Economics and Finance 48, 2008, pp. 733-755.

[14] L. Yu, Sh. Wang, and K. Lai, " Credit risk assessment with a multistage neural network ensemble learning approach", Expert Systems with Application 34, 2008, pp. 1434-1444.

[15] M. Sustersic, D. Mramor, and J. Zupan, " Consumer credit scoring models with limited data", Expert Systems with Applications 36, 2009, pp. 4736-4744.

[16] H. Abdou, J. Pointon, and A. EI-Masry, " Neural nets versus conventional techniques in credit scoring in Egyptian banking", Expert Systems with Applications 35, 2008, pp. 1275-1292.

[17] R. Malhotra and D.K. Malhotra, " Evaluating consumer loans using neural networks", Omega 31, 2003, pp. 83-96.

[18] R. Malhotra, D.K. Malhotra, " Differentiating between good credits and bad credits using neuro-fuzzy systems", European Journal of Operational Research 136, 2002, pp. 190-211.

[19] SPSS Inc., "Clementine® 12,0 User's Guide," Integral Solutions Limited, 2007.

[20] SPSS Inc., "Clementine® 12,0 Modeling Nodes," Integral Solutions Limited, 2007.

[21] SPSS Inc., "Clementine® 12,0 Source, Process, and Output Nodes," Integral Solutions Limited, 2007.

[22] SPSS Inc., "Clementine® 12,0 Algorithms Guide," Integral Solutions Limited, 2007.

[23] م. غضنفری، س. علیزاده، و ب. تیمورپور، " داده کاوی و کشف دانش"، مرکز انتشارات دانشگاه علم و صنعت ایران، 1387.

[24] وب سایت www.defaultrisk.com

پی نوشت:

[1] *Credit risk*

[2] *Linera probability distribution*

[3] *Logit model*

[4] *Probit model*

[5] *Support vector machine*

[6] *Feed forward neural network*

[۷] *Multilayer perceptrons (MLP)*

[۸] *Back propagation algorithm*

[۹] *Momentum*

[۱۰] *Sigmoid function*

[۱۱] *Crossed validation*

[12] *Multinomial Logistic Regression*

[13] *Chi-squared Automatic Interaction Detector (CHAID)*