INTERNATIONAL CONFERENCE ON
SCIENCE AND ENGINEERING
1st December 2015 DUBAI - UAE
کنفرانس بین المللی علوم و مهندسی
۱۰ آذرماه ۱۳۹٤ دبی - امارات

استفاده از یک مدل داده کاوی برای پیش بینی رفتار مشتریان مشاغل خرد در بنگاه های اقتصادی و تعیین ریسک اعتباری در اعطای تسهیلات بانکی

سمیه مهاجری

کارشناس ارشد مهندسی کامپیوتر گرایش نرم افزار - Somayyeh_Mohajeri@yahoo.com

خلاصه

هدف بالقوه هر سازمان مالی حفظ مشتریان موجود و رسیدن به مشتریان جدید در بلند مدت است. رفتار اقتصادی مشتری و ماهیت سازمان با یک توصیفی با نام "شناخت مشتریان خود (KYC) [۱] " در نظام بانکداری کنترل می شود. زمانی که بانک ها قصد دارند به مشتریان وام یا تسهیلاتی پرداخت کنند، وضع اعتبار دارایی ها و خوش حسابی یا بدحسابی مشتری مورد نظر را می سنجند .مشتریان سپرده گذار در برخی از بخش ها با ریسک بالا هستند؛ در حالی که در برخی از بخش ها خطر متوسط و ریسک بسیار پائینی دارند. در حال حاضر ریسک اعتباری را می توان بطور گسترده تحت عوامل کمی و کیفی طبقه بندی کرد. اگر چه سیستم های موجود بسیاری در حفظ مشتری در بانک نهادینه شده است اما روش های سختگیرانه و بدون رویکرد روشن و تعریف شده ای برای پرداخت تسهیلات در بخش بنگاه اقتصادی های وجود دارد . در این مقاله از اطلاعات مشتریان مشاغل خرد بنگاه های اقتصادی در یکی از شعب بانک های ایران برای تجزیه و تحلیل و عوامل عمده تراکنش های مشتریان و پیش بینی رفتار ایشان در پرداخت تسهیلات بانکی استفاده شده است و بهره گیری از روش های داده کاوی در شعب بانک برای موضوع فوق و ساده تر شدن فعالیت های سنجش و گزینش درخواست تسهیلات مشتریان ، از میان تعداد زیاد درخواست تسهیلات به عنوان راه حل مناسب اعطای تسهیلات به مشتریان بنگاه های اقتصادی پیشنهاد شده است و از درخت تصمیم گیری و الگوریتم C۴,۵ در مدل استفاده شده است و در نهایت عملکرد خود را با نتایج MATLAB مورد آزمایش قرار داده ایم .

کلمات کلیدی: داده کاوی، درخت تصمیم، مشتری، بانک، تسهیلات

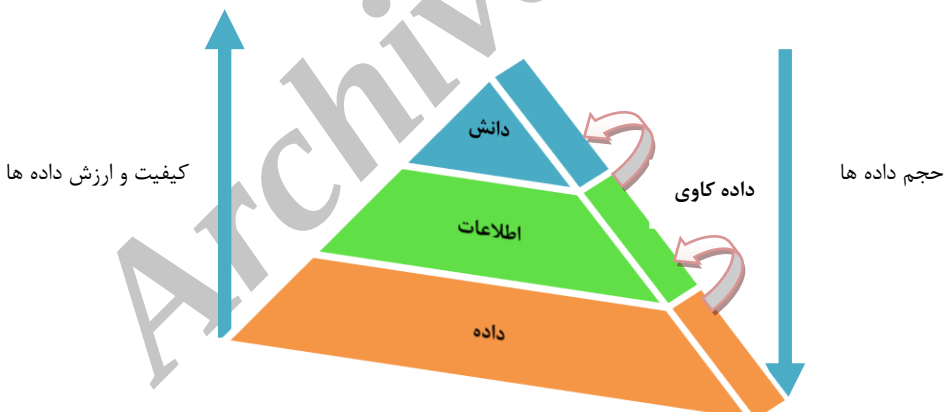
[۱] Know Your Customer

۱.مقدمه

وقوع رویدادهایی نظیر توسعه رقابت جهانی، پیشرفت فناوری اطلاعات و دسترسی به سیستمهای اطلاعاتی ارزان در سالهای اخیر و نیز تلاش واحدهای اقتصادی در جهت احراز رتبه جهانی و ورود به بازارهای بین‌المللی، داشتن نگرشهایی همچون رضایت مشتریان، مدیریت اطلاعات را اجتناب‌ناپذیر کرده است. آنچه مسلم است موفقیت و تداوم فعالیت در محیط رقابتی جدید، مستلزم استفاده از روشه ای نوینی است که کسب‌وکار را در رده جهانی قرار می دهد که در این میان بانک ها نیز به عنوان یک بنگاه اقتصادی از این امر مستثنی نبوده و در سال های اخیرتعداد بانک ها و موسسات مالی که خدمات بانکی و محصولات بانکداری الکترونیکی را معرفی و گسترش می دهند افزایش یافته است .

اساس و پایه استراتژی این موسسات مالی توانایی در حفظ مشتریان موجود و رسیدن به مشتریان جدید و آینده نگر است.صنعت بانکداری در ایران به سرعت در حال رشد است واین درحالی است که همگام با رشد صنعت بانکداری از طریق پیشرفت های تکنولوژیکی ،ارائه ایده های جدید برای بازار بین بانکی برای حفظ آن به ضروریات تبدیل شده است و لاجرم پرتفوی محصولات ارائه شده توسط بانک ها متنوع شده است.با توجه به این که موضوع تجهیز منابع و تخصیص آن به عنوان خط اصلی فعالیت بانکها شناخته شده به طوریکه حتی در تعریف بانکها این مهم گنجانده شده است ؛ لذا به نظر می رسد هر چه از عمر یک بانک می گذرد می بایست منابع بیشتری جذب و به همان نسبت تسهیلات بیشتری پرداخت نموده و مقداری از منابع را نیز مصرف دارائی های غیر تسهیلاتی کرده باشد. در طول سال های اخیر، جذب مشتریان بیشتر از بیش در رسیدن به این هدف مهم شده است ؛ تجمع داده های عملیاتی به ناچار از این رشد در این صنعت منتج می شود.

نیاز فزاینده ای برای تبدیل داده ها به اطلاعات مفید و باارزش یکسان سازی شده به منظور حرکت رو به جلو و به دست آوردن یک مزیت رقابتی وجود دارد . داده کاوی[۱] نقش مهمی در این تلاش ها دارد .



شکل۱: نقش داده کاوی در تبدیل داده ها

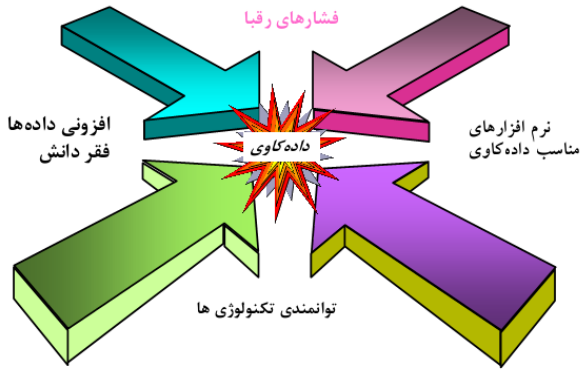
در این مقاله با استفاده از مدل درخت های تصمیم[۲]به عنوان یکی مدل های داده کاوی ، سعی شده است اطلاعات ارزشمندی از مشتریان تسهیلات بانکی با توجه به اینکه پیدا کردن مشتریان خوب برای پرداخت تسهیلات واقعا یک مسئله چالش برانگیز در عصر بانکی است ، استخراج شود .در مدل درخت های تصمیم از الگوریتم C۴,۵ برای استخراج اطلاعات مشتریان دارای ریسک اعتباری کمتر و سودمندتر برای پرداخت تسهیلات بانکی استفاده شده است.

[۱] Data Mining
[۲] Decision Trees

۲. داده کاوی و نقش آن در بانک ها

در بسیاری از رشته های علمی و فنی، در نهایت ما با مجموعه ای از داده ها روبرو هستیم که حجم کم یا زیادی را دارند؛ اما مهم ترین کار، به دست آوردن چنین پایگاه داده ای نیست. بلکه باید بتوانیم، سطح بالاتری از دانش را با توجه به پایگاه داده مذکور به دست بیاوریم؛ این یعنی نتیجه گیری و جمع بندی تمام تلاش هایی که برای جمع آوری آن داده ها صرف شده است.

داده کاوی فرآیند تکرار شونده است که ترکیبی از دانش بنگاه اقتصادی، روش های یادگیری ماشین و ابزار و مقادیر زیادی از اطلاعات دقیق و مربوط به توانایی کشف بینش غیر بصری پنهان دراطلاعات حقوقی سازمان است. این اطلاعات می تواند اصلاح فرآیندهای موجود و کشف روندها بوده و به تدوین سیاست های مربوط به رابطه این شرکت با مشتریان و کارکنانش کمک کند. در بانک ها و موسسات مالی، داده کاوی به طور موفقیت آمیزی در تعیین کاندیدهای واجد شرایط احتمالی برای پرداخت تسهیلات بانکی،پیدا کردن مشتریان سودآور، محصولات بانکی، مشخص کردن عوامل ارائه خدمات مختلف استفاده شده است(Hu, ۲۰۰۵).تمامی این عوامل،بهبود روش های قدیمی انجام امور و وادار کردن بانک ها به در نظر گرفتن بازنگری فعالیت های خود را برای افزایش سهم از بازار در بر دارد.



شکل۲: نقش داده کاوی در رقابت بین بانکی

تحقیقات متعددی در زمینه تجزیه و تحلیل برخورد وحفظ مشتری در بخش های بانکی ساخته شده است. برخی مطالعات نشان می دهد که مهم ترین متغیرهای موثر بر انتخاب مشتری خدمات ارائه شده موثر و کارآمد، سرعت و کیفیت خدمات، انواع خدمات عرضه شده، هزینه های خدمات الکترونیکی پایین، امکانات بانکداری آنلاین، ایمنی وجوه و صندوق امانات و در دسترس بودن خدمات مبتنی بر فن آوری ، نرخ بهره پایین تسهیلات، محل مناسب شعبه ، منظربانک، مدیریت خوب، محیط کلی بانک می باشد(Aregbeyen,۲۰۱۱- Rehman and Ahmed,۲۰۰۸- Siddiqi,۲۰۱۱).

از سوی دیگر، مشتری هسته عملیات بانک ها می باشد، به طوری که پرورش و حفظ آنها برای موفقیت بانکها مهم است.با رشد چشمگیر سیستمهای اطلاعاتی در سامانه‌های فروش و جمع آوری داده‌های عملکردی مشتریان نیاز به ابزارهای مناسب جهت تحلیل داده‌های مشتریان و رفتارسنجی آنان جهت ارتباط موثر با مشتریان دوچندان شده است. بانک ها و موسسات مالی جهت ارتباط موثر با مشتریان خود ضمن ثبت مشخصات مشتریان، داده‌های تراکنش های مالی آنها مانند زمان انجام تراکنش، نوع کالا یا خدمات خریداری شده را در سامانه‌های اینترنتی یا کامپیوتری ثبت می‌کنند. ثبت این داده‌ها بدون پردازش آنها دانش موجود در رفتار آنها جهت ارائه خدمات من جمله تسهیلات بانکی را ندارد. داده‌کاوی ابزار

کاوش در داده‌های زیاد به دنبال کشف دانش از آنها است. لذا استفاده از تکنیکهای داده‌کاوی بر روی داده‌های مشتری در حل مشکل یاد شده و تحقق سودآوری و کاهش ریسک اعتباری مشتریان کمک خواهد کرد.
بسیاری از مسائل داده کاوی را می توان به صورت یک مسأله کلاسه بندی[۱] بیان نمود، که در آن در نهایت یک عامل پیش بین تربیت می شود که می تواند با در دست داشتن دانش موجود برای طبقه بندی یک مجموعه از موارد، آن دانش را به کلاسه بندی سایر موارد تعمیم دهد. در واقع مسأله کلاسه بندی، یک مسأله یادگیری نظارت شده است.

۳. کلاسه بندی با درخت‌های تصمیم در داده کاوی

درخت تصمیم شیوه منحصر به فردی از ارائه یک سیستم است ، که تصمیم گیری های آتی را تسهیل و سیستم را به نحو مناسبی تعریف می کند . با توجه به اینکه اکثر سیستم های مهندسی ، اجرایی و محاسباتی را می توان در قالب یک سری داده (ویژگی یا ویژگی ها و خروجی منطبق با آنها) تعریف کرد ، می توان با استفاده از یک الگوریتم (ایجاد درخت) ویژگی ها و خروجی ها را آنالیز کرد و سیستم را بر اساس این داده ها در قالب یک درخت تصمیم ارائه نمود.
استفاده از درخت های تصمیم به عنوان کلاسه بندی کننده در حال حاضر کاملا کاربردی شده و انواع مختلف پیاده سازی آن من جمله ID۳ وC۴,۵ عملا در کاربردهای متنوع استفاده می شود. درخت تصمیم بر اساس آنالیز داده های ورودی و به منظور پیدا کردن یک ویژگی[۲] به عنوان مبنای تصمیم گیری هر گره بکار می رود . در هر گره ویژگی های مختلف داده ها بررسی شده و یک ویژگی که در صورت انتخاب باعث کاهش بی نظمی (آنتروپی) می شود، گزینش می شود(Quinlan,۱۹۹٦).
بعد از ایجاد یک درخت می توان از ساختار ایجاد شده برای کلاسه بندی داده های تست استفاده کرد ؛ در اینصورت توسط این ساختار سریع ، داده های تست به خوبی برای جداسازی در کلاس های مختلف تصمیم گیری می شوند. عموما داده های تعلیم که برای ایجاد درخت استفاده می شوند ،از داده های تست که برای ارزیابی درخت ایجاد شده استفاده می شوند متفاوت بوده و تعداد خطا در تشخیص کلاس های داده تست ، معیار مناسب بودن الگوریتم می باشد.
الگویتم ID۳ یک الگوریتم برای ساختن درخت تصمیم می باشد. در این الگوریتم از مفهوم بی نظمی[۳] برای دسته بندی داده ها استفاده شده است و الگوریتم در صدد آن است که میزان بی نظمی در گره های بالایی درخت حداقل باشد تا بتوان درختی با حداقل ارتفاع داشت . پس ابتدا برای تمامی ویژگی های داده های اولیه بی نظمی را محاسبه کرده و سپس آن ویژگی که بیشترین سودمندی را خواهد داشت به عنوان ریشه انتخاب می کند .این الگوریتم فقط قادر به دسته بندی داده ها با دامنه ویژگی های گسسته و محدود می باشد و در مورد داده های نویزی و مخدوش کارایی ندارد.

۳,۱. الگوریتم C۴,۵

الگوریتم C۴,۵ تکمیل شده الگوریتم ID۳ می باشد . این الگوریتم قادر به دسته بندی داده های پیوسته و نویزی نیز می باشد. برای این منظور ابتدا داده ها را مرتب کرده سپس مقادیر سودمندی را برای تمامی حالت هایی که امکان جداشدن این داده های مرتب شده از هم وجود دارد را بدست آورده و جداساز متناظر با بزرگترین مقدار سودمندی را به عنوان جداکننده انتخاب می کنیم. به طور کلی کلاسه‌بندی داده‌ها یک فرآیند دو مرحله‌ای است :

[۱] Classification
[۲] Feature
[۳] Entropy

۱. در مرحله اول یک مدل ساخته‌می‌شود که مجموعه‌ای از کلاسهای داده‌ای یا مفاهیم را مشخص می‌کند . این مرحله را مرحله یادگیری[۱] گوییم که در آن یک الگوریتم کلاسه‌بندی یک مدل را با تحلیل یک مجموعه آموزشی[۲] که مجموعه‌ای از تاپل‌های پایگاه است می سازد و برچسب کلاس‌های مربوط به این تاپل‌ها را مشخص می‌کند . یک تاپل X با یک بردار صفت $X=(x_1,x_2,\ldots,x_n)$ نمایش داده می‌شود . فرض می شود که هر تاپل به یک کلاس از پیش تعریف شده متعلق است و کلاس با یک صفت که به آن صفت برچسب کلاس می‌گوییم مشخص می‌شود . مجموعه آموزشی به صورت تصادفی از پایگاه انتخاب می شود .

۲. در مرحله دوم ، یادگیری از طریق یک تابع $y=f(X)$ انجام می‌شود که می‌تواند برچسب کلاس هر تاپل X از پایگاه را پیش بینی کند . این تابع به صورت قواعد کلاسه‌بندی ، درخت‌های تصمیم گیری یا فرمول‌های ریاضی است . آنچه ما در اینجا برای کلاسه بندی از درختان تصمیم‌استفاده می کنیم .

عملکرد الگوریتم فوق به شرح ذیل است :

- الگوریتم با پارامترهای:مجموعه آموزشی و برچسب کلاس‌های متناظر با آنها (D) ، لیستی از صفات موجود در تاپل‌ها روال انتخاب ویژگی فراخوانی می‌شود . D در واقع یک بخش[۳] داده‌ای است وروال انتخاب ویژگی یک روال ابتکاری[۴] است که بهترین صفت را برای جداکردن تاپل‌ها براساس کلاس ها می‌دهد . این متد از یک معیار انتخاب صفت مانند بهره اطلاعاتی[۵] یا شاخص جینی[۶] استفاده می کند .
- درخت در گام اول با یک گره تنها (N) که مجموعه آموزشی را نشان می دهد ایجاد می شود .
- اگر تاپل‌های D همه از یک کلاس باشند گره N یک برگ خواهد بود و با آن کلاس برچسب می خورد.
- در غیر این صورت روال انتخاب ویژگی فراخوانی می‌شود تا معیار شکاف[۷] را مشخص کند . معیار شکاف مشخص می‌کند که کدام صفت باید در گره N مورد آزمون قرار گیرد . معیار شکاف همچنین بیان می‌کند که چه شاخه‌هایی باید از گره N با توجه به آزمون مربوطه ، خارج شوند . به عبارت دیگر معیار شکاف ، صفت یا نقطه شکاف را تعیین می‌کند . نقطه شکاف ، D را به یکسری بخش تبدیل می‌کند . این بخش ها باید تا حد ممکن خالص[۸] باشند به این معنی که همه تاپل های موجود در یک بخش باید مربوط به یک کلاس باشند .
- گره N با معیار شکاف برچسب می‌خورد. یک شاخه از گره N به هر یک از خروجی‌های معیار شکاف می رود . تاپل‌های D متناظر بخش بندی می‌شوند.
- الگوریتم فرایند مشابهی را به صورت بازگشتی در هر یک از بخش های حاصل شده دنبال می‌کند .
- بخش بندی بازگشتی در صورتی که یکی از شرایط زیر بوجود آید متوقف می‌شود :
  ۱. اگر تمام تاپل ها در بخش D متعلق به یک کلاس باشند .
  ۲. صفتی برای بخش بندی بیشتر وجود نداشته باشد.در این حالت گره N به یک برگ تبدیل می شود و برچسب کلاس آن کلاس متداول در D خواهد بود .
  ۳. تاپلی برای یک شاخه وجود نداشته باشد ، در واقع اگر یکی از بخش های D مانندD_j تهی باشد در این موارد یک برگ با برچسب کلاس متداول در D ایجاد می شود (Han and Kamber,۲۰۰۱)

---

[۱] Learning
[۲] training set
[۳] Partition
[۴] Heuristic
[۵] Information Gain
[۶] Gini Index
[۷] Splitting criterion
[۸] pure

معیارهای مختلفی برای انتخاب صفتی که شکاف باید بر اساس آن انجام شود وجود دارد که ما از بهره اطلاعاتی استفاده می کنیم.اطلاعات مورد نیاز برای کلاسه بندی یک تاپل در D برابر رابطه (۱) است :

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (۱)$$

که در آن P_i احتمال آن است که یک تاپل دلخواه در D متعلق به کلاس C_iباشد که این احتمال به صورت $|C_i,D|/|D|$ تخمین زده می‌شود ($|D|$ و $|C_i,D|$ تعداد تاپل‌ها در D و C_i,D را نشان می‌دهد) . تعداد کلاسهای موجود m است . در واقع رابطه (۱) نشان دهندهبی نظمی می‌باشد .

فرض می‌کنیم صفت A دارای v مقدار متمایز بصورت $\{a_۱,a_۲,\ldots,a_v\}$ باشد یا بعبارت دیگر A یک صفت گسسته است . اگر بخواهیم D را بر حسب صفت A بشکافیم v بخش یا زیرمجموعه مانند $\{D_۱,D_۲,\ldots,D_v\}$ حاصل می شود که در آن D_j شامل تاپل‌هایی از D است که مقدار صفت A در آنها برابر a_j است . اگر فرض کنیم که D در گره‌ای چون N قرار داشته باشد آنگاه این بخش‌ها متناظر با شاخه‌هایی هستند که از N خارج می‌شوند . اطلاعات مورد نیاز برای کلاسه‌بندی یک تاپل از D بر حسب صفت A برابر با :

$$Info_A(D) = \sum_{j=1}^{v} |D_j| / |D| \times Info(D_j) \qquad (۲)$$

است . عبارت $|Dj|/|D|$ در واقع وزن بخش j را نشان می‌دهد .

اطلاعات حاصل از انشعاب بر حسب صفت A را به صورت زیر تعریف می‌کنیم :

$$Gain(A) = Info(D) - Info_A(D) \qquad (۳)$$

هر چه مقدار بهره صفت A بیشتر باشد یا به عبارت دیگر هر چه $Info_A(D)$ کمتر باشد صفت A به عنوان صفت شکاف انتخاب می شود .

۴. سیستم پیشنهادی

سیستم پیشنهادی در شکل ۳ نشان داده شده است. اطلاعات مربوط به تراکنش های مشتری از پایگاه داده گرفته شده و تکنیک پیش پردازشی داده کاوی روی این داده ها استفاده می شود. بعد از چندتجزیه و تحلیل آماری، یک کلاس سلسله مراتبی از پیش تعریف شده (مقادیر هدف و یا ویژگی های اختیاری) در هر اطلاعات حساب اختصاص داده می شود. سپس روی این داده هایک روش طبقه بندی مناسب انجام می گیرد. روش کلاسه بندی درخت تصمیم گفته شده برای استخراج قواعد برای پیش بینی رفتار مشتریان استفاده شده است.

شکل ۳ : نمایی از سیستم پیشنهادی

۵. پیاده سازی و طراحی سیستم

۵،۱. شناسایی دوره و گردش داده

برای اطلاعات اولیه از پایگاه داده اطلاعات مشتریان شعبه مرکزی یکی از بانکهای خصوصی ایران استفاده کردیم . با توجه به موقعیت شعبه که در مرکز شهر واقع بوده و اندازه شعبه که بزرگترین شعبه بانک فوق است و درجه ممتاز را در درجه بندی شعب استان ها از سال ۱۳۹۱ اخذ نموده است؛ تنوع و تعدد مشتریان اطمینان از همه جانبه بودن و صحت نتایج خروجی را به ما می دهد.

با توجه به در دسترس بودن داده ها و دوره تناوب زمان خود، تصمیم به شروع انتخاب داده بااستخراج یک زیر مجموعه از تراکنش های حساب ها که در وضعیت "فعال" برای مدت یک سال قبل از پرداخت تسهیلات بود ، نمودیم .دلیل این انتخاب ، وضعیت حساب مشتری قبل از پرداخت تسهیلات می باشد که در آن معدل موجودی شش ماهه و یک ساله حساب مشتری در زمان پرداخت تسهیلات مشخص می شود.

با توجه به اینکه درخواست هایی که از سوی سازمان ها و ادارات جهت پرداخت تسهیلات معرفی می شوند و تحت عنوان تسهیلات تکلیفی در بانک ها شناخته می شود و بانک موظف به پرداخت این تسهیلات می باشد اطلاعات این مشتریان در پایگاه داده استفاده نشد از این جمله به پرداخت تسهیلات قرض الحسنه ازدواج می توان اشاره نمود و تسهیلات پرداختی شش ماهه سال از ابتدای سال ۱۳۹۴ یعنی ۱۳۹۴/۰۱/۰۵ تا ۱۳۹۴/۰۶/۳۱ انتخاب شده و با توجه به وضعیت انتقال تسهیلات به سرفصل مطالبات سررسید گذشته که پس از دو ماه از سررسید اقساط پرداخت نشده و به طبقه معوق پس از گذشت شش ماه از اولین قسط پرداخت نشده بازه زمانی را حداکثر ۱۳۹۴/۰۸/۱۵ در نظر گرفتیم."از تاریخ" و"تا تاریخ" زیر پرس و جوی مورد استفاده در پرس و جو اصلی برای بازیابی اطلاعات از پایگاه داده هستند.

برای انواع تسهیلات، از انواع سرفصل و معین هایی که در سیستم بانکی استفاده می شود ، بهره گرفتیم هر سرفصل تسهیلاتی یک عدد نهایت سه رقمی بوده که نوع تسهیلات را نشان می دهد . معین طبقه ریز تری از سرفصل بوده و اطلاعات کامل تری از نوع تسهیلات به ما می دهد برای مثال سرفصل ۴۳ مربوط به تسهیلات مشارکتی بوده و معین ۱ بخش صنعت و معدن و معین ۵ مربوط به بخش کشاورزی می باشد که ما ترکیب سرفصل و معین را در تحقیق خود استفاده نمودیم (همانند ۴۳۱و ۴۳۵ در نمونه گفته شده). همچنین برای بدست آوردن ریسک اعتباری هر شخص از فرمول زیر استفاده نمودیم:

$$Risk=m*۱۰۰ /M \quad (۴)$$

که در آن Risk ریسک اعتباری مشتری تسهیلاتی و m میزان اقساط سررسید شده پرداخت نشده و Mمیزان تسهیلات می باشد. نهایتا اطلاعات به دست آمده در یک جدول موقت برای مورد استفاده قرار گرفتن با پیوستن با تمام جداول مربوطه واکشی شده در بخش های بعدی ذخیره می شوند.

۵،۲. انتخاب داده ها

اطلاعات استخراج شده از پایگاه داده بانک ، یک ردیف از هر حساب برای داده های اساسی بود مانند تعداد حساب ، عنوان حساب، شغل ، توانایی پرداخت و غیره . نتیجه داده های اولیه در جدول ۱ نشان داده شده است مقادیر این متغیر های ثابت در طول زمان تغییر نمی کند.مقادیر میانگین داده های تراکنشی متغیرهای حساس به زمان روزانه تغییر پیدا می کند و حفظ تمام این مقادیر مختلف برای شش ماه در جهت پیدا کردن رفتار فصلی مرتبط به فعالیت های حساب امری ضروری است. برای همین منظور معدل موجودی سه ماهه و شش ماهه و یکساله حساب استخراج شده است.

داده های تراکنشی یعنی حداقل مقدار اعتبار یک تک تراکنش ازهر حساب با استفاده از پرس و جوی زیر از پایگاه داده که در آن "از تاریخ" و"تا تاریخ" در بخش قبلی توضیح داده شد واکشی شده است . برخی اطلاعات دیگر از جمله هزینه استعلام از بانک مرکزی، کارمزد ارائه خدمات ، سود متعلقه ،هزینه پست،۱/۵در هزار هزینه کارشناسی ،حق ماموریت، حق بیمه، هزینه ارزیابی وثایق ملکی و سایر هزینه های دیگر در پرس و جوی تراکنش های تولید شده نادیده گرفته شد. خروجی در بررسی دیگر برای مقادیر صفر مورد آزمون قرار گرفت.

جدول ۱: برخی از داده های اولیه پایگاه داده مشتریان تسهیلاتی شعبه بانک

شماره قرارداد	سرفصل	تاریخ ایجاد	مبلغ تسهیلات	نوع وثیقه	ارزش وثیقه	شغل
۹۴۸۸۸۴۰۱۶۹۸	۴۷۱	۹۴۰۵۲۹	۱۰۰	۶۰	۶	فروشنده لوازم یدکی اتومبیل
۹۴۸۲۹۰۲۰۶۸۱	۴۳۱	۹۴۰۶۱۶	۲۰۰	۶۰	۱۱	شرکت خدماتی
۹۴۳۹۶۶۴۴۳۴۰	۴۳۱	۹۴۰۵۱۱	۱۰۰	۶۰	۱۲	راننده تریلر
۹۴۳۱۲۳۶۹۱۱۷	۱۹۳۸	۹۴۰۵۰۸	۶۳	۶۰	۱۳	کارگر شرکت
۹۴۸۸۶۸۶۴۲۸۸	۴۳۱	۹۴۰۵۱۸	۱۱۰	۶۰	۱۵	مهندس معدن
۹۴۸۱۹۴۷۱۸۳۱	۴۳۱	۹۴۰۵۲۸	۱۱۵	۶۰	۱۷	پیمانکار آب و فاضلاب روستایی
۹۴۳۴۵۴۹۹۱۳۷	۴۳۱	۹۴۰۵۰۸	۱۰۰	۶۰	۲۰	کارگر کارخانه
۹۴۸۸۶۳۴۵۳۷۲	۴۳۱	۹۴۰۶۰۳	۱۰۰	۶۰	۲۰	کارمند دولتی
۹۴۳۹۴۴۶۴۲۱۲	۴۳۱	۹۴۰۵۱۱	۲۰	۸۰	۲۳	منشی
۹۴۸۱۱۴۰۹۱۱۵	۴۳۱	۹۴۰۵۲۵	۱۳۰	۶۰	۲۴	راننده تریلر
۹۴۸۱۹۱۲۳۷۲۱	۱۹۳۸	۹۴۰۵۱۷	۷۵	۶۰	۲۵	فرشباف
۹۴۸۱۵۰۵۲۹۱۹	۴۳۱	۹۴۰۵۱۳	۱۲۵	۶۰	۲۵	نمایشگاه اتومبیل
۹۴۸۲۷۴۳۱۳۹۱	۴۳۱	۹۴۰۶۳۱	۲۴	۸۰	۲۹	کارمند دولتی
۹۴۰۷۲۷۳۴۹۳۷	۴۳۱	۹۴۰۳۲۰	۲۶	۸۰	۳۱	دامدار
۹۴۸۱۹۸۳۳۳۳۸	۱۹۳۸	۹۴۰۶۰۹	۱۰۰	۶۰	۳۳	راننده تریلر
۹۴۰۵۹۲۸۲۵۷۴	۴۳۱	۹۴۰۲۲۷	۴۰	۸۰	۵۰	کارمند دولتی

۶. تحلیل آماری

بخش قابل توجهی از مجموعه داده ها فبلد های تهی هستند. فیلتر کردن این فیلدها در مراحل اولیه به طور قابل توجهی می تواند به کاهش زمان پردازش و بهبود مدل دقت کمک کند(Hu,۲۰۰۵)تعدادی از رکوردها که یک ویژگی تک مقداری داشتند و برخی از رکوردها نیز مقادیر صفر داشتند، آماری معنی دار تلقی نشدند و ارزش آماری نداشتند. این فیلدها از جدول داده های هدف به منظور کاهش زمان محاسبات برای حصول اطمینان از فرایند مدل سازی سریع تر حذف شدند.

بسیاری از داده ها در جدول نهایی عددی بودند به عنوان مثال نوع وثیقه از جمله داده های با اهمیت یافته اصلی از داده های تراکنشی از مشتریان تسهیلاتی شعبه بانک بود . در تحقیق ازنرمالسازی مقیاس دهدهی[۱] استفاده شد همچنین ما برای یکسان سازی ارقام مبالغ را در واحد میلیون ریال در نظر گرفتیم .

۷. تعریف مقادیر هدف

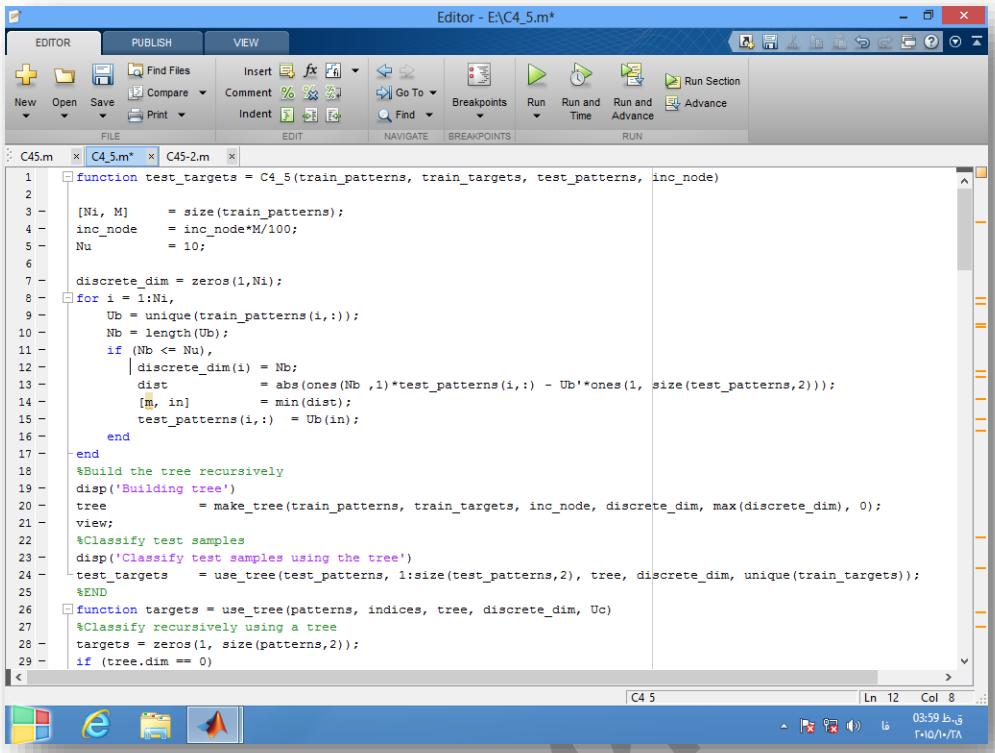
بخشی از روش داده کاوی تعریف درست هدف مناسب است که با توجه به اهداف پرداخت تسهیلات برای تجزیه و تحلیل داده کاوی با کمک از کارشناسان حوزه اعتباری که در بخش وام مشغول هستند ، مقادیر هدف از نظر داده های موجود تعریف شد و با این، مقدار متغیر هدف، یعنی متغیری که ویژگی های اختیاری را تعیین می کند ، تعریف شد . بدین وسیله به عنوان برچسب کلاس تعریف شد . شرایط استفاده از تعریف: عالی (برابرو بالاتر از ۹۰)؛ خیلی خوب (برابر و بالاتر از ۸۰ و کمتر از ۹۰)؛ خوب (برابر و بالاتر از ۷۰ولی کمتر از ۸۰)؛ مرزی (برابر و بالاتر از ۶۰ و کمتر از ۷۰)؛ و بد (کمتر از ۶۰)این است.

۸. درخت تصمیم

درخت تصمیم از یک درخت برای ساخت یک مدل پیش بینی (تخمین) استفاده می کند که مشاهدات درباره یک آیتم را به نتیجه گیری هایی درباره مقدار هدف آن آیتم نگاشت می دهد. در واقع درخت تصمیم گیری یک ساختار درختی مانند فلوچارت است که مجموعه ای از قوانین را برای استفاده مدل پیش بینی ایجاد می کند. مزیت این روش این است که قوانین برای درک بسیار آسان است و آنهاغالبا برای کشف فرآیندهای پرداخت تسهیلات اصولی مفید هستند . الگوریتم با یک مجموعه آموزش و برچسب کلاس مرتبط با خود را شروع می شود.

مجموعه آموزش به صورت بازگشتی به زیر مجموعه های کوچکتر مانند درخت که ساخته شده است تقسیم می شود.

[۱] Decimalscaling normalization.

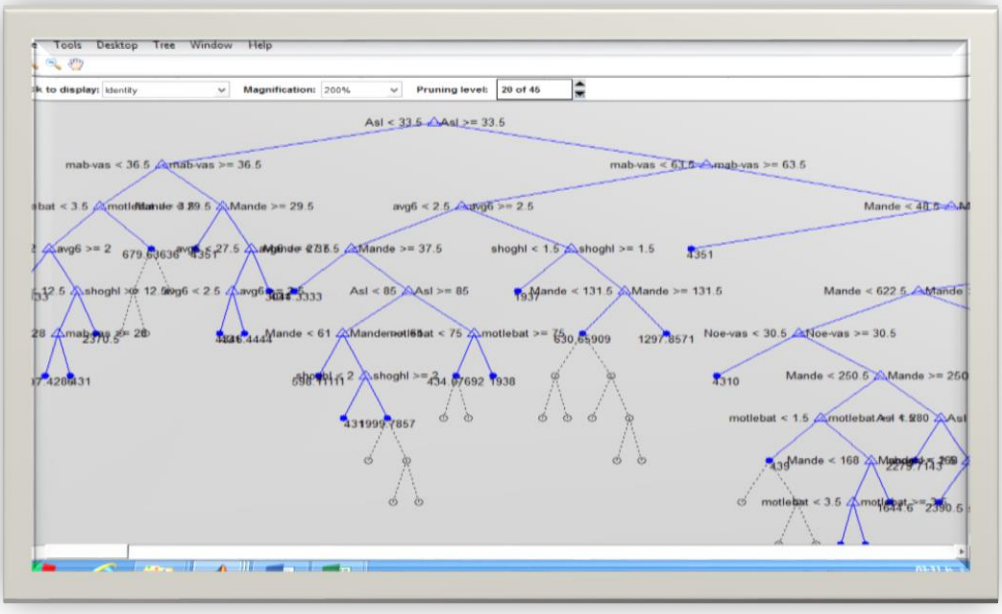


شکل ۴ : پیاده سازی الگوریتم C۴.۵ در نرم افزار Matlab

۹. نتیجه و بحث

داده های آموزشی از پایگاه داده تراکنشی بعد از۱۳۹۳/۰۱/۰۵ و قبل از۱۳۹۴/۰۸/۱۵ ، جمع آوری شد. این مدل با برخی از نمونه داده ها برای قبل از ۱۳۹۳/۰۱/۰۵ و بعد از آن ۱۳۹۴/۰۸/۱۵ از پایگاه داده مورد آزمایش قرار گرفت . این مدل همچنین با داده های جمع آوری شده از نزدیک مورد آزمایش قرار گرفت. یافته های اولیه برخی نتایج جالب را نشان می دهد.

به منظور آزمایش اثر از مدل، مدل با نرم افزار MATLABتست شده و متوجه شدیم که خروجی آنها تقریبا با اطلاعات



اولیه یکسان هستند. خروجی نتیجه MATLAB در شکل ۵ نشان داده شده است.از خروجی می توان گفت که مدل به اندازه کافی دقیق برای پیش بینی وضعیت مشتریان بنگاه اقتصادی در دریافت تسهیلات از یک شعبه بانک تجاری است.

شکل ۵ : خروجی از نرم افزار MATLAB برای داده های مشتریان تسهیلاتی

مطالعات اولیه تعدادی از روابط بین متغیرها را پرده برداری کرد که بحث و تجزیه و تحلیل بیشتری را تصدیق می کرد. صداقت و امانتداری متقاضی دریافت تسهیلات و سابقه دریافت تسهیلات ، عملکرد مثبت ۵ ساله ، وضعیت رقابتی بازار و شرایط و مشخصات آن از جمله عواملی بود که می توانست بررسی ما را در پیش بینی رفتار مشتریان خرد بهبود بخشد.

۱۰. نتیجه گیری

ما عوامل و بخش های چشم انداز اعطای تسهیلات به بنگاه های اقتصادی را در شعبه بانک با اجرای رویه ای از وظایف داده کاوی مورد بحث قرار دادیم . ما تعداد مطلوب خوشه ها را (عالی،روش های طبقه بندی خیلی خوب، خوب، مرزی و بد) از پیش تعریف کردیم واز متد طبقه بندی (درخت تصمیم) برای پیش بینی رفتار مشتریان بخش بنگاه اقتصادی برای دریافت تسهیلات بانکی بر روی داده های رفتاری تراکنشی مشتریان موجود استفاده نمودیم. یافته های اولیه برخی از نتایج جالبی را نشان می دهد. آزمون ما را راهنمایی و متوجه کرد که مدل پیش بینی داده کاوی برای چشم انداز بخش های بنگاه اقتصادی برای پرداخت وام بسیار دقیق است .

تمرکز اصلی این مقاله بر صفات کتبی از مشتری بنگاه اقتصادی بود اما مشتری ویژگی های بسیاری از رفتار غیرکتبی و شفاهی دارد . پس موضوع از برخی از ویژگی های مهم نانوشته همچون صداقت، شخصیت، شرایط بازار مستثنی است . این مدل بر اساس روش های داده کاوی طبقه بندی درخت تصمیم گیری به عنوان یک گام اولیه توسعه یافته است . ما یک طرح برای گسترش مدل برای روشهای داده کاوی دیگر در آینده داریم.

۱۱.مراجع

[۱] Xiaohua Hu, (۲۰۰۵) A Data Mining Approach for Retailing Bank Customer Attrition Analysis. Applied Intelligence. Vol. ۲۲, pp. ۴۷–۶۰.

[۲] OmoAregbeyen, Ph.D, (۲۰۱۱) The Determinants of Bank Selection Choices by Customers: Recent and Extensive Evidence from Nigeria. International Journal of Business and Social Science. Vol.۲, No. ۲۲, pp.۲۷۶-۲۸۸.

[۳] Hafeez Ur Rehman and Saima Ahmed, (۲۰۰۸) An Empirical Analysis of the determinants of bankselection in Pakistan; A customer view. Pakistan Economic and Social Review. Vol. ۴۶, no. ۲, pp. ۱۴۷-۱۶۰.

[۴] Kazi Omar Siddiqi, (۲۰۱۱) Interrelations between Service Quality Attributes, Customer Satisfaction and Customer Loyalty in the Retail Banking Sector in Bangladesh. International Journal of Businessand Management. Vol. ۶, No. ۳, pp.۱۲-۳۶.

[۵] J. R. Quinlan, (۱۹۹۶)Improved use of continuous attributes in c۴.۵. Journal of Artificial Intelligence Research, ۴:۷۷-۹۰.

[٦] Jiawei Han and MichelineKamber,(٢٠٠١) "Data Mining: Concepts andTechniques", Morgan Kaufmann, ٢٠٠١.