9th National Conference on Management & Humanistic Science Research in Iran
11 March 2021 - Tehran
نهمین همایش ملی پژوهش های مدیریت و علوم انسانی در ایران
۲۱ اسفند ۱۳۹۹ محل برگزاری: تهران

بکارگیری تکنیک داده کاوی در پیش بینی رفتار مشتریان بانک (مورد مطالعه: بانک توسعه تعاون)

محسن پیرمحمدی[۱]

چکیده

پیچیدگی محیطی، شدت رقابت، رواج تکنولوژی‌های نو و پیشرفته، توسعه فناوری اطلاعات و ارتباطات، و سمت گیری سازمان‌ها از دارایی‌های مشهود به نامشهود از عوامل عمده ای است که موجب شده سازمان‌ها و بانک ها در دوران حیات خود با ریسک‌های بسیار متعدد و حتی پیش بینی نشده مواجه شوند. یکی از مهمترین ریسک ها در بانک، مدیریت ریسک اعتباری می باشد. از این رو هدف از این مقاله بکارگیری تکنیک داده کاوی برای پیش بینی رفتار مشتریان بانک است که بتوان بر اساس شاخص ها و پارامترهای تاثیرگذار در انتخاب متقاضیان اعتبار، آنها را به خوبی دسته بندی کرده و از احتمال عدم بازپرداخت تسهیلات اعطایی کاست. این تحقیق، تحقیقی کاربردی است که از روش درخت تصمیم گیری ژنتیک به تجزیه و تحلیل داده های جمع آوری شده، پرداخته است. در این تحقیق یک مدل مناسب اعتبارسنجی مشتریان بانکها برای اعطای تسهیلات اعتباری متناسب با هر طبقه مبتنی بر الگوریتم ژنتیک ارایه می شود. الگوریتم های ژنتیک می توانند با انتخاب ویژگی های مناسب و ساخت درختان تصمیم گیری بهینه به اعتبارسنجی مشتریان کمک کنند. مدل طبقه بندی پیشنهادی مبتنی بر تکنیک های خوشه بندی، انتخاب ویژگی ها، درختان تصمیم گیری و الگوریتم ژنتیک است. این مدل به انتخاب و ترکیب بهترین درختان تصمیم گیری مبتنی بر معیارهای بهینگی و ساخت درخت تصمیم گیری نهایی برای اعتبارسنجی مشتریان می پردازد. نتایج نشان می دهد که دقت طبقه بندی مدل طبقه بندی پیشنهادی به طورتقریبی از تمام مدل های درخت تصمیم گیری مقایسه شده در این مقاله بالاتر است. همچنین تعداد برگ ها و اندازه درخت تصمیم گیری و در نتیجه پیچیدگی آن از همه کمتر است.

واژگان کلیدی: اعتبارسنجی، الگوریتم ژنتیک، انتخاب ویژگی ها، درختان تصمیم گیری، خوشه بندی

[۱] کارشناس ارشد فناوری اطلاعات گرایش مدیریت سیستم های اطلاعاتی mohsenpirmohamadi@gmail.com

۱- مقدمه

.به طور معمول در سازمان ها و شرکت ها ازمدیریت ریسک برای کاهش اثرات مربوط به پیدایش شرایط نامساعد ودرنتیجه کاهش تاثیررفتاری این شرایط برعملکرد وکارکردسازمان ها و شرکت ها استفاده میشود. یکی از مهمترین موضوعاتی که سازمان ها درشرایط رقابتی کنونی با آن سر وکار دارند، ارزیابی و اندازه‌گیری ریسک اعتباری در میان ریسک‌هایی که بانک در حیطه وسیع عملکرد خود با آن روبرو است از جایگاه ویژه‌ای برخوردار است. بانک ها و مؤسسات مالی زمانی با این ریسک مواجه می شوند که تسهیلات گیرنده به علت عدم توان یا تمایل،تعهدات خود را در سررسید در قبال بانک یا مؤسسات مالی ایفا نمی کند. ریسک اعتباری یکی از مهمترین ریسک هایی استکه نهادهای پولی ومالی را تحت تأثیر قرارمی دهد.هنگامی که تسهیلات گیرنده به علل مختلف با بحران مالی مواجه می شود ریسک اعتباری بانک افزایش می یابد، به عبارت دیگر کلیه ریسک های تسهیلات گیرنده ازطریق ریسک اعتباری برریسک بانک ها ومؤسسات مالی اثر می گذارد.بدین سبب ریسک اعتباری ازاهمیت بسیار زیادی برخورداربوده وکمیته بازل نیز برشناسایی وکنترل این ریسک تأکید ویژه ای دارد.(راستی،۱۳۹۰، ۲۶) یکی از تکنیک‌های مناسب در این زمینه، بکارگیری الگوها و مدل‌های تصمیم‌گیری می‌باشد. بکارگیری این تکنیک‌ها و برنامه‌ریزی در جهت کاربردی‌تر و به تبع آن مکانیزه و سیستمی‌تر نمودن آن‌ها می‌تواند به بانک‌ها کمک فراوانی در جهت مدیریت بر بحران‌ها و ریسک‌ها از جمله ریسک‌های اعتباری، نقدینگی، بازار، نرخ بهره و سودآوری می نمایند. از این رو هدف از این تحقیق بکارگیری تکنیک داده کاوی برای پیش بینی رفتار مشتریان بانک توسعه تعاون است.

۲- بیان مساله و روش تحقیق

در غالب اوقات، ریسک معرف اثر منفی بر پروژه تلقی می‌گردد؛ در صورتی که ریسک می‌تواند دریچه‌ای بر فرصت‌ها، توسعه، بهبود و یا تفکر جدید نیز باشد. (رشیدیان،۱۳۹۰، ۳۰) بدیهی است مدیریت ریسک با استفاده از نگرش هیأت مدیره و مدیران، به سازمان ها و مؤسسات کمک میکند تا با جایگزین کردن یک استراتژی جامع به جای روش های منفرد و یک جانبه در مقابل ریسک ها، نحوه اداره مؤثر و مقرون به صرفه آنها را یافته و به اجرا بگذارد.(حاجی آقایی،۱۳۸۶، ۵۷) امروزه اهمیت ارتباط با مشتریان بر کسی پوشیده نیست و تمامی سازمان ها از جمله سازمان های ارائه دهنده خدمات مالی سعی در درک بیشتر از مشتریان خود دارند. برای رسیدن به درک صحیح از مشتری، سازمان ها علاوه بر ارتباط با مشتریان نیازمند استفاده از مقیاسی برای سنجش میزان ارزش و اهمیت مشتریان مختلف هستند. این مقیاس در صورتی فراهم خواهد شد که سازمان بتواند با استفاده از ابزار مناسب به میزان ارزش مشتریان خود دست یافته و به تجزیه و تحلیل آن بپردازد. شناخت گروه های مختلف مشتریان و ایجاد ارتباط اثربخش با آنها به گونه ای که بتوان منافع اقتصادی سازمان را در آینده تضمین نمود، مساله ای مهم در کسب و کار امروز است. جذب مشتریان سودآور و همچنین حفظ و نگه داری مشتریان ارزشمند قدیمی هر دو دارای اهمیت هستند که جز با شناسایی دقیق ویژگی های آنها امکان پذیر نمی باشد.

بنابرین یکی از موضوعات دارای اهمیت، بررسی و ارزیابی ریسک اعتباری (یعنی احتمال قصور در بازپرداخت تسهیلات اعطایی از سوی مشتریان) می باشد. اندازه‌گیری این ریسک در میان ریسک هایی که بانک در حیطه وسیع عملکرد خود با آن روبرو است، از جایگاه ویژه ای برخوردار است. کاهش و کنترل ریسک به عنوان یکی از عوامل مهم و مؤثر بر بهبود فرآیند اعطای اعتبار و در نتیجه بر عملکرد بانک ها مطرح است و نقش اساسی در تداوم ارائه تسهیلات و بقای بانک ها و مؤسسات مالی دارد. استفاده از تکنیک های داده کاوی در کسب و کار امروزی به طور روز افزونی گسترش پیدا کرده است (چن، ۲۰۰۷). در واقع تکنیک های داده کاوی از ابزارهایی می باشند که که می توان از طریق شناسایی الگوهای رفتاری مشتریان و نیازهای هر گروه از مشتریان و ارائه سرویس های خاص هر گروه این مشکلات را حل نماید (جواهری، ۱۳۸۸). در واقع هدف اصلی در این

تحقیق بکارگیری تکنیک داده کاوی برای پیش بینی رفتار مشتریان بانک می باشد. تحقیق حاضر یک تحقیق کمی است که اطلاعات مورد نیاز از پایگاه داده بانک گردآوری گردیده است. بدین صورت که در ابتدا با توجه به موضوع تحقیق برای بکارگیری تکنیک داده کاوی برای پیش بینی رفتار مشتریان بانک از ادبیات تحقیق کمک گرفته شده، سپس بر اساس اطلاعات بدست آمده شاخص ها جداسازی و مرتب گردیدند.بر اساس شاخص های مورد نظر، دسته بندی و جمع بندی گردید. سپس داده های مورد نظر استخراج، طبقه بندی شده و مورد تجزیه و تحلیل قرار گرفته اند.

۳- تجزیه و تحلیل داده ها

۳-۱ توصیف داده ها

داده هایی که برای آموزش، تست و ساخت درختان تصمیم گیری استفاده می شوند، مجموعه داده های اعتباری بانک توسعه تعاون استان اردبیل است. این مجموعه داده دارای مقادیر مفقود و اختلال نیست. بر روی این مجموعه عملیات آماده سازی و تمیز کردن و پیش پردازش داده ها صورت گرفته است. به منظور انجام پیش پردازش داده ها از تکنیک های خوشه بندی و انتخاب ویژگی ها استفاده شده است. این مجموعه داده دارای ۱۰۰۰ تراکنش و ۲۱ ویژگی است. از این تعداد ویژگی ۷ ویژگی عددی و ۱۳ تای آن اسمی هستند. یک ویژگی هدف در این ویژگی ها به بررسی خوب یا بد بودن مشتری می پردازد. ویژگی های مجموعه داده های اعتباری بانک توسعه تعاون استان اردبیل به همراه نوع آنها در زیر آمده است.

جدول ۱-ویژگی های مجموعه داده

ردیف	ویژگی	نوع ویژگی
۱	وضعیت چک	اسمی
۲	مدت زمان	عددی
۳	سابقه اعتبار	اسمی
۴	هدف	اسمی
۵	مقدار اعتبار	عددی
۶	وضعیت پس انداز	اسمی
۷	سابقه کار	اسمی
۸	تعداد اقساط	عددی
۹	وضعیت شخصی و جنسیت	اسمی
۱۰	طرف های دیگر	اسمی
۱۱	محل اقامت فعلی	عددی
۱۲	اموال و دارایی ها	اسمی
۱۳	سن	عددی
۱۴	برنامه های پرداختی دیگر	اسمی
۱۵	مسکن	اسمی
۱۶	وضعیت اعتباری موجود	عددی

۱۷	شغل	اسمی
۱۸	تعداد عائله مندی	عددی
۱۹	تعداد اقساط	عددی
۲۰	آخرین مانده بدهی	عددی
۲۱	ارزش وثیقه	عددی

روش هایی که به منظور آماده سازی و تمیز کردن داده ها بر روی مجموعه داده های اعتباری اعمال بانک توسعه تعاون استان اردبیل استفاده شده است، بدین صورت است:

۱.حذف مقادیر پراکنده

۲.نرمال سازی که فقط بر روی ویژگی «سن» اعمال شد و مقادیر ویژگی هدف در محاسبات این روش لحاظ شده است.

۳.گسسته سازی مقادیر ویژگی های عددی که در این روش مقادیر ویژگی هدف در محاسبات لحاظ شده است.

۴.ادغام مقادیر داده در ویژگی های اسمی.

۵.تبدیل ویژگی های عددی به اسمی.

۳-۲ طبقه‌بندی و درخت تصمیم‌گیری

طبقه‌بندی یکی از وظایف داده کاوی است و دارای تکنیک‌های متنوعی می باشد که می‌توان از آن ها برای طبقه‌بندی استفاده کرد. در این مقاله با استفاده از درخت تصمیم گیری به طبقه‌بندی پرداخته شده است.

در این مقاله از الگوریتم C۴.۵ برای ساخت درختان تصمیم گیری به منظور اعتبار سنجی مشتریان بانک استفاده می شود. الگوریتم C۴.۵ در سال ۱۹۹۳ توسط کوئینلن تهیه شده است. در این الگوریتم متغیرهای پیوسته و گسسته در محاسبات لحاظ شده و مقادیر مفقود در الگوریتم در نظر گرفته شده است. این الگوریتم لزوماً دودویی نیست. برای انتخاب یک جداکننده بهینه در طول مسیر درخت تصمیم گیری از شاخص کسب اطلاعات یا کاهش آنتروپی استفاده می کند.

۳-۳ خوشه بندی

خوشه بندی به عنوان یکی از فعالیت های داده کاوی می باشد و به گروه بندی کردن تراکنش ها، مشاهدات یا حالت ها در کلاس های مشابه می پردازد. همچنین یک خوشه مجموعه ای از رکورد ها است که به هم شبیه می باشند و از رکوردهای بیرون خوشه تفاوت دارند. در خوشه بندی متغییر هدف وجود ندارد و به طبقه بندی، تخمین و پیشگویی مقدار متغییر هدف نمی پردازد. در این مقاله از الگوریتم خوشه بندی Simple K Means استفاده می شود. مراحل الگوریتم Simple K Means در به شرح زیر است.

۱.انتخاب تعداد مورد تمایل خوشه ها به اندازه K.

۲.انتخاب تعداد K مشاهده اولیه به عنوان seed.

۳.محاسبه متوسط مقادیر خوشه برای هر ویژگی یا متغیر.

۴.تخصیص مشاهدات آموزشی دیگر به نزدیک ترین خوشه توسط محاسبه مقیاس فاصله مورد نظر. ۵.محاسبه مجدد متوسط های خوشه بر اساس تخصیص ها در مرحله قبل.

۵.تکرار بین مراحل ۴ و ۵ می توان از تکنیک خوشه بندی به عنوان پیش پردازش داده ها استفاده کرد، که در این پژوهش بر روی مجموعه داده های اعتبار سنجی بکار می رود.

۴-۳ انتخاب ویژگی ها

به دلیل اینکه مدل های ناپارامتریک در طبقه بندی مبتنی بر داده هستند، نیاز به صرف زمان و هزینه زیاد برای کسب داده های مدل است. پس بهتر است ویژگی ها و داده هایی جمع آوری شود که از اهمیت بیشتری در ساخت مدل طبقه بندی برخوردار هستند. حذف اطلاعات غیر مرتبط و استخراج متغییرهای کلیدی در شناخت الگو، پیش پردازش نامیده می شود. در ساخت مدل مناسب طبقه بندی نیاز به داده های آموزشی با کیفیت مناسب است. انتخاب ویژگی ها به عنوان یکی از روش های پیش پردازش داده ها می تواند منجر به افزایش کیفیت مجموعه داده آموزشی در آزمون و ساخت مدل درخت تصمیم گیری شود.

تعاریف مختلفی از انتخاب ویژگی ها مطرح شده است. انتخاب ویژگی به شناسایی و انتخاب ویژگی های متمایز برای ساخت مدل ها و تفسیر بهتر داده ها می پردازد. انتخاب ویژگی ها مطرح شده است. انتخاب ویژگی به شناسایی و انتخاب ویژگی های متمایز برای ساخت مدل ها و تفسیر بهتر داده ها می پردازد. بکارگیری انتخاب ویژگی ها در ساخت مدل درخت تصمیم گیری در طبقه بندی باعث می شود تا اعتبار سنجی مشتریان اعتباری بانک به شیوه بهتری صورت گیرد. الگوریتم انتخاب ویژگی ها شامل سه قسمت است:

۱.معیار ارزیابی ویژگی.

۲.روش جستجو.

۳.قانون توقف. به طور معمول معیارهای ارزیابی بدین صورت است:

۱.اطلاعات

۲.وابستگی

۳.فاصله

۴.سازگاری

۵.دقت طبقه بندی.

الگوریتم های انتخاب ویژگی مبتنی بر ۴ رویکرد اول در بالا از روش فیلتر استفاده می کنند. در اینجا، الگوریتم انتخاب ویژگی مستقل از الگوریتم طبقه بندی دارد. الگوریتم انتخاب ویژگی که از معیار دقت طبقه بندی استفاده می کند، از رویکرد Wrapper بهره می برد. در این رویکرد از الگوریتم یادگیری مثل الگوریتم طبقه بندی برای انتخاب ویژگی ها استفاده می شود. سه روش جستجو در انتخاب ویژگی ها وجود دارد که عبارتند از: ۱.کامل ۲.هیوریستیک ۳.تصادفی.

دو روش کامل و هیوریستیک در فضاهای کوچک کاربرد دارد که نیاز به کارایی بالا در فرآیند جستجو است. روش تصادفی مثل الگوریتم ژنتیک برای فضاهای بزرگ و پیچیده کاربرد دارد. قوانین مختلفی برای توقف الگوریتم انتخاب ویژگی ها موجود است: ماکزیمم تعداد تکرار الگوریتم، کسب نتیجه بهتر توسط اضافه یا کم کردن یک ویژگی از مجموعه ویژگی ها، رسیدن به یک زیر مجموعه بهینه از ویژگی ها. یکی دیگر از روش های انتخاب ویژگی، طرح های جاسازی شده است. در این روش الگوریتم انتخاب ویژگی به عنوان بخشی از الگوریتم طبقه بندی لحاظ می شود. در این پژوهش از رویکرد فیلتر، Wrapper و طرح جاسازی شده برای انتخاب ویژگی ها استفاده می شود. روش جستجو در انتخاب ویژگی ها به صورت تصادفی و مبتنی بر الگوریتم ژنتیک است و قانون توقف برای الگوریتم انتخاب ویژگی ها، ماکزیمم تعداد تکرار در الگوریتم انتخاب ویژگی می باشد.

۴- مدل تلفیقی پیشنهادی

مدل تلفیقی پیشنهادی از الگوریتم انتخاب ویژگی ها مبتنی بر الگوریتم ژنتیک و درخت تصمیم گیری ژنتیکی، درختان تصمیم گیری C 4.5 و همچنین الگوریتم Simple K Means برای خوشه بندی داده ها استفاده می کند. برای هر خوشه، الگوریتم های طبقه بندی متا و الگوریتم انتخاب ویژگی مبتنی بر درخت تصمیم گیری ژنتیکی برای ایجاد درختان تصمیم گیری C 4.5 بکار می رود. از یک استراتژی مناسب مبتنی بر معیارهای بهینگی مطرح در این پژوهش برای انتخاب بهترین درختان تصمیم گیری در هر خوشه استفاده می شود. الگوریتم متا به ترکیب الگوریتم انتخاب ویژگی و الگوریتم درخت تصمیم گیری C 4.5 می پردازد. روش های متا روش هایی جدید می باشند که برای ترکیب چند طبقه کننده بکار می روند.

در این مقاله در روش متا الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک با الگوریتم درخت طبقه بندی C 4.5 ترکیب می شود. ابتدا داده های اعتبار سنجی مشتریان پس از آماده سازی و تمیز شدن، به دو مجموعه داده آموزش و تست تقسیم می شوند. سپس توسط تکنیک خوشه بندی، این مجموعه داده به دو خوشه تقسیم می شوند. در هر خوشه با استفاده از پنج روش انتخاب ویژگی به انتخاب ویژگی های مهم پرداخته می شود. البته ۴ الگوریتم انتخاب ویژگی مبتنی بر رویکرد فیلتر و Wrapper به کمک الگوریتم متا با الگوریتم درخت تصمیم گیری C 4.5 ترکیب شده اند. پس از انتخاب ویژگی ها نوبت به ساخت درختان تصمیم گیری C 4.5 می رسد. تا این مرحله در هر خوشه ۵ درخت تصمیم گیری وجود دارد. مبتنی بر معیارهای بهینگی، بهترین درختان تصمیم گیری در هر خوشه انتخاب شده و با هم ترکیب می شوند تا درخت تصمیم گیری نهایی برای اعتبار سنجی مشتریان بانک ساخته شود.

در مدل تلفیقی پیشنهادی از خوشه بندی به عنوان یکی از روش های پیش پردازش داده ها استفاده می شود. تعداد خوشه ها در این مدل عدد ۲ در نظر گرفته شد. انتخاب بهینه تعداد خوشه ها از مسائل پیچیده می باشد. می توان در ابتدا تعداد خوشه ها را ۲ گرفت و مرتباً این مقدار را اضافه کرد تا جایی که دیگر هیچ بهبودی در مدل طبقه بندی حاصل نشود. البته در این مقاله از این روش استفاده نمی شود. به نظر می رسد بین درصد مشاهدات درست طبقه بندی شده و سایر معیارهای بهینگی درختان تصمیم گیری در برخی مواقع تضاد بوجود آید. به عبارت دیگر افزایش درصد مشاهدات درست طبقه بندی شده ممکن است باعث افزایش تعداد ویژگی های پیشگو منتخب، تعداد برگ ها و اندازه درخت تصمیم گیری شود. این موضوع در خود الگوریتم درخت تصمیم گیری C 4.5 و همچنین با هرس درخت تصمیم گیری و اعمال محدودیت هایی مثل مینیمم تعداد تراکنش در هر برگ در نظر گرفته می شود. ولی برای مقایسه بین درختان تصمیم گیری C 4.5 نیز باید یک تعاملی بین ۴ معار بهینگی درخت تصمیم گیری بوجود آید. ممکن است درخت تصمیم با دقت کمتر، دارای اندازه و تعداد برگ های کمتری در درخت تصمیم گیری نیز باشد. در صورتی که کاهش دقت نامحسوس باشد، با توجه به نظر کاربر درخت تصمیم گیری با دقت کمتر برای طبقه بندی مشتریان بانک ها انتخاب می شود. زیرا این درخت تصمیم گیری C 4.5 بهتر در هر خوشه به نظر کاربر یا کارشناس اعتبار سنجی بستگی دارد.

در الگوریتم طبقه بندی متا به ترکیب الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک و الگوریتم درخت تصمیم گیری C 4.5 پرداخته می شود. بدین ترتیب که ویژگی های مناسب توسط الگوریتم ژنتیک انتخاب می شوند و سپس این ویژگی ها به عنوان ورودی برای ایجاد درخت تصمیم گیری C 4.5 بکار می روند. از روش گلدبرگ برای نمایش ژنتیکی کروموزوم ها استفاده شده است. هر کروموزوم نشان دهنده زیر مجموعه ویژگی ها است. هر ژن نماد یک ویژگی است. مقدار آن ژن برابر یک و صفر است که به ترتیب نشان دهنده وجود و عدم وجود ویژگی مورد نظر در زیر مجموعه ویژگی ها است.

در این مقاله از الگوریتم ژنتیک باینری استفاده شده است. در این الگوریتم چون متغیرها دارای تغییرات پیوسته نیستند و نمی توانند هر مقداری به خود بگیرند، الگوریتم ژنتیک گسسته است. مجموعه متغیرهای مساله که باید مقدار بهینه برای آنها پیدا

شود در قالب رشته های باینری کد شده و به همدیگر الحاق شده اند. به این ترتیب یک کروموزوم از متغیرهای مساله به دست آمده است که هر کروموزوم یک جواب منحصر به فرد برای مساله مورد بررسی است. کلیه متغیرهای تصمیم به صورت یک رشته کد باینری به یکدیگر الحاق شده اند که به این ترتیب یک کروکوزوم از متغیرهای مساله به دست می آید.

۵- عملگرهای الگوریتم ژنتیک

۵-۱ انتخاب

پس از اینکه برازندگی تمام افراد یک نسل مشخص شد، طبق اصول طبیعی، فرزندانی که از زوج های برازنده تر به وجود می آیند، برازندگی بیشتری دارند و همان طور که در طبیعت، افرادی که برتری هایی نسبت به دیگران دارند، به زوج های برتری دست می یابند، الگوریتم ژنتیک این فرآیند را شبیه سازی می کند و به افراد برازنده تر شانس تولید مثل بیشتری می دهد. فرآیند انتخاب تعیین تعداد دفعاتی است که یک فرد می تواند در مرحله تکثیر شرکت کند. برای انتخاب در این تحقیق از تکنیک چرخ گردان استفاده شده است. در این روش احتمال انتخاب کروموزوم ها با برازندگی بیشتر بالاتر است. به عبارت دیگر، به هر کروموزوم به نسبت برازندگی آن یک احتمال انتخاب داده می شود. در نتیجه ممکن است بعضی از کروموزوم ها چند بار انتخاب شوند یا اصلا انتخاب نشوند. احتمال انتخاب متناظر p_k با هر کروموزوم بر مبنای برازندگی آن محاسبه می شود، طوری که اگر f_k مقدار برازندگی کروموزوم kام باشد، آنگاه مساحت هر بخش p_k متناسب با مقدار برازندگی f_k در کروموزوم kام است. همچنین بر مبنای رابطه زیر داریم:

$$\sum_{i=1}^{k} p_k = 1 \qquad (۱)$$

از آنجا که مجموع p_k برابر با ۱ است. در نهایت عددی بین صفر و یک به طورتصادفی انتخاب خواهد شد. اگر r بر مبنای توزیع یکنواخت در بازه $[0,1]$ به طور تصادفی انتخاب شود، آنگاه r به عنوان نشانگر (انتخاب عدد تصادفی) در فرآیند مدل سازی ریاضی چرخه گردان استفاده می شود و اعضا بر مبنای رابطه زیر انتخاب می شوند:

$$\begin{array}{lll} if & 0 \le r \le p_1 & select\ p_1 \\ if & p_1 \le r \le p_1 + p_2 & select\ p_2 \end{array}$$

$$if \quad p_1 + \cdots + p_{k-1} \le r \le p_1 + \cdots + p_k \qquad select\ p_k$$

۵-۲ پیوند

پیوند مهم ترین عملگر الگوریتم ژنتیک و کلید موفقیت آن است. عملگر انتخاب برای کشف نواحی جدید فضای جست وجو ابزاری ندارد و اگر تنها، به نسخه برداری ساختارهای قدیمی، بدون تغییر آن اکتفا شود، نمی توان به بررسی موارد جدید پرداخت. عملگری است که اطلاعات بین رشته ای را به طور اتفاقی مبادله می کند در این مقاله از عملگر پیوند یک نقطه برش استفاده شده است و از دو عضو والد دو عضو فرزند ایجاد شده است. در این نوع از پیوند بخشی از ژن های ذرات والد با هم جابجا می شوند و ذرات فرزند را به وجود می آورند. در این مرحله جواب های اولیه به عنوان والدین در نظر گرفته شده بر اساس عملگر یک نقطه برش تقاطع اعضای والد ادغام و اعضای ثانویه که در اصطلاح جمعیت فرزند نامیده می شود، ایجاد می شود.

اگر $x_1(t)$ و $x_2(t)$ به عنوان عضوهای والد تعریف شوند. $m(t)$ یک عملگر صفر و یک باشد که به تعداد ژن های هر عضو n_x تعریف شده و مقدار آن صفر و یک است. اگر مقدار آن صفر باشد عمل پیوند صورت نمی گیرد ولی اگر مقدار آن یک باشد عمل پیوند انجام می شود و فرزند $\widetilde{x_1}(t)$ و $\widetilde{x_2}(t)$ را به وجود می آورد.

شبه کد پیوند برای رشته ذرات

$\widetilde{x_1}(t)$ را فرزند ذره $x_1(t)$ و $\widetilde{x_2}(t)$ را فرزند ذره $x_2(t)$ قرار دهید.

اگر $u(0,1) \leq p_c$ باشد آنگاه:

مقدار $m(t)$ باینری را برای هر ژن تعیین کنید.

برای هر ژن $j = 1, \dots, n_x$ انجام دهید.

اگر m_j برابر با ۱ باشد آنگاه:

ژن ها را به صورت زیر جابجا کنید:

$$\widetilde{x_{1j}}(t) = x_{2j}(t)$$

$$\widetilde{x_{2j}}(t) = x_{1j}(t)$$

احتمال پیوند برابر با ۰/۸ و جمعیت فرزندان برابر با ۰/۰۵ کل جمعیت اولیه در نظر گرفته شده است. از عملگر تقاطع تک نقطه ای برای تقاطع استفاده شده است.

۳-۵جهش

سومین عملگر در الگوریتم ژنتیک جهش نام دارد. گرچه عملگرهای انتخاب و پیوند، جست وجوی موثری درفضای جستجو طراحی می کنند، گاهی باعث می شوند از بین خصوصیات مفید رشته ها بروند. عملگر جهش امکان دستیابی مجدد به این ویژگی های مثبتی را که در جمعیت نیست، فراهم می کند. عملگر جهش بدین صورت است که اگر مقدار یک ژن صفر باشد، آن را تبدیل به یک می کند و اگر مقدار آن ژن یک باشد، آن را به صفر تبدیل می نماید. این عملگر در کروموزوم های متفاوت تغییرات تصادفی برنامه ریزی نشده ایجاد می کند و ژن هایی را که در جمعیت اولیه وجود نداشته اند را وارد جمعیت می کند. عملگر جهش با احتمال معین p_m برای هر ژن از عضو $x_i(t)$ اعمال می شود و عضو فرزند جهش یافته $\widetilde{x_i}(t)$را به وجود می آورد. احتمال جهش به عنوان نرخ جهش شناخته می شود. در این تحقیق از عملگر جهش یکنواخت (تصادفی) استفاده شده است. در این روش تعدادی ژن به طورتصادفی انتخاب و مقدار آن تعویض می شود. یک رشته به طور تصادفی انتخاب و مقدار یکی از سلول هایش از صفر به یک تغییر داده می شود. نرخ جهش نیز برابر با ۰/۰۱ در نظر گرفته شده است.

شبه کد الگوریتم یکنواخت (تصادفی) جهش

برای هر ژن j=1,….,n_x انجام دهید.

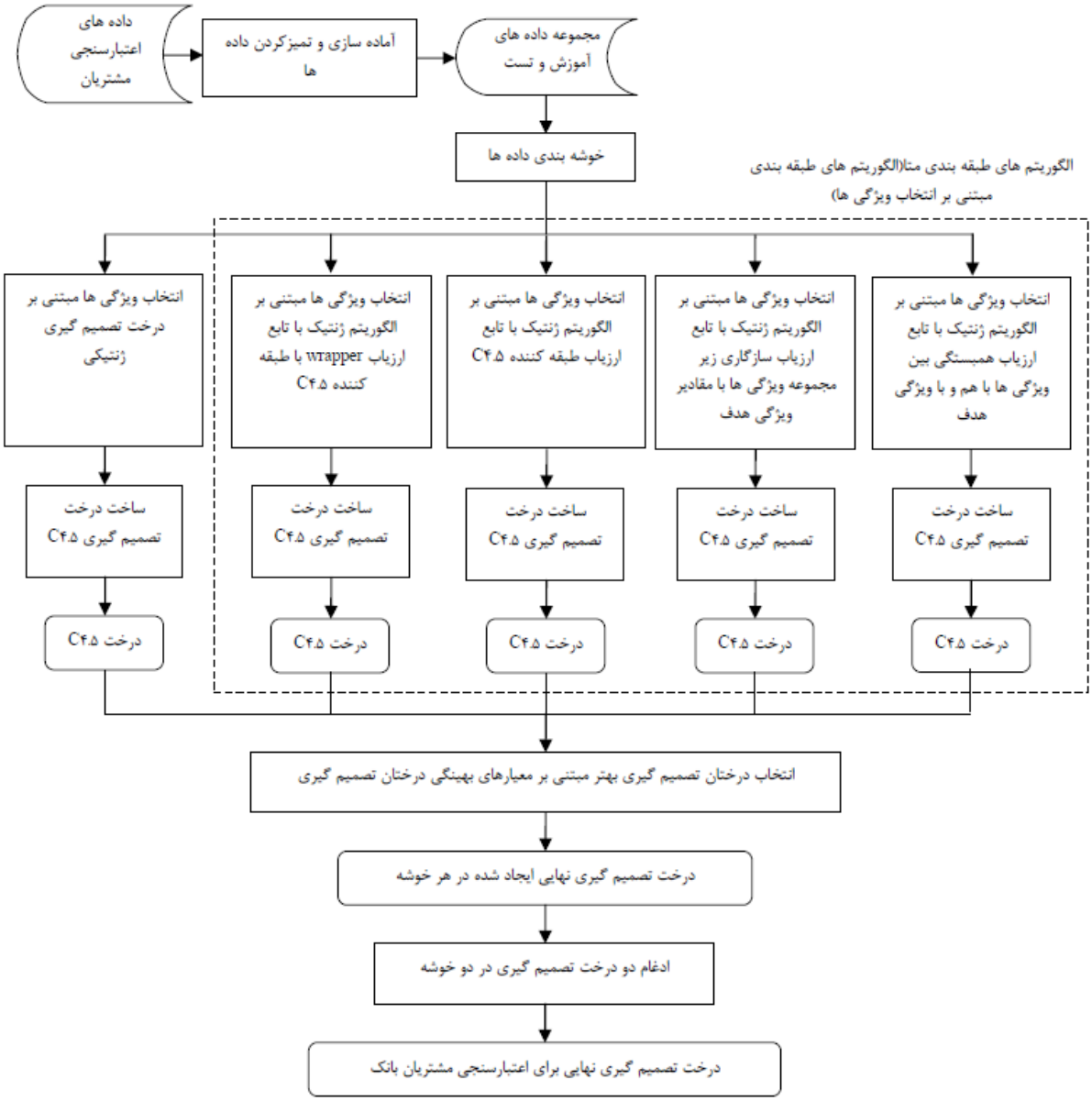
اگر ⟦ u(0,1)≤p⟧ _m آنگاه

x_ij (t)=~(x_ij)˜(t))˙)

۶- بررسی شرایط خاتمه

بعد از اجرای عملگر پیوند و جهش مساله دارای سه جمعیت شامل جمعیت فرزندان، جمعیت اصلی و جمعیت جهش یافتگان است. به همین منظور کروموزوم های این سه جمعیت بر مبنای کیفیت جواب هایشان و تابع هدف مساله با یکدیگر ادغام شده، جمعیت ادغام شده را به وجود می آورند. به عبارت دیگر مقدار متناظر هر کروموزوم در تابع هدف مساله قرار می گیرد و کروموزوم هایی با بهترین جواب انتخاب می شوند. پس از ادغام جمعیت ها و ایجاد نسل های جدید و انجام تکرارهای مختلف، لازم است شرایط خاتمه در الگوریتم بررسی شود. در این تحقیق شرایط خاتمه شرط پایانی، تغییر نکردن بهترین رشته، در هر نسل، بر ای پنجاه نسل متوالی در نظر گرفته شده است؛ بدین معنی که اگر بهترین رشته، در پنجاه نسل متوالی بدون تغییر

باقی ماند، الگوریتم پایان می پذیرد. شرط توقف الگوریتم ژنتیک در اینجا تعداد نسل ها در نظر گرفته شد. عملگر جایگزینی کروموزوم با کروموزوم های دیگر مبتنی بر پایه شایستگی است.



شکل ۱: فرآیند ساخت و آزمون مدل تلفیقی پیشنهادی در اعتبار سنجی مشتریان بانک

۷- آموزش، تست مدل

بعد آماده سازی و تمیز کردن داده ها و به منظور آموزش و تست مدل از تعداد ۶۹۰ تراکنش استفاده شده است. تعداد خوشه ها در الگوریتم خوشه بندی ۲ و عدد Seed برابر یک در نظر گرفته شده است. ویژگی هدف در محاسبات الگوریتم خوشه بندی

لحاظ نشده اند. از تکنیک اعتبارسنجی متقاطع در آموزش و تست مدل تلفیقی پیشنهادی این پژوهش استفاده شده است. پارامترهای الگوریتم ژنتیک در انتخاب ویژگی ها مبتنی بر رویکرد های فیلتر و Wrapper به قرار زیر است:

جدول ۲- پارامترهای الگوریتم ژنتیک در انتخاب ویژگی ها مبتنی بر رویکرد های فیلتر و WRAPPER

ردیف	پارامتر	میزان/ نوع
۱	اندازه جمعیت اولیه	۲۰
۲	نوع ایجاد جمعیت اولیه	عدد تصادفیSeed برابر ۱
۳	اندازه کروموزوم	۱۶ (رشته باینری)
۴	نوع پیوند	یک نقطه برش
۵	احتمال پیوند	۰/۹
۶	نوع جهش	یکنواخت تصادفی
۷	نرخ جهش	۰/۰۱
۸	روش انتخاب والد	چرخ گردان
۹	انتخاب کروموزوم بهینه	رتبه ای

۱۰

از عدد اعتبارسنجی متقاطع ۱۰ برای آموزش و تست مدل استفاده شده است. بدین ترتیب که ابتدا یک دهم اول داده ها برای تست استفاده می شود و بقیه برای آموزش الگوریتم انتخاب ویژگی یا درخت تصمیم گیریC4.5 بکار می رود. سپس یک دهم بعدی و به همین ترتیب ۱۰ بار این عمل صورت می گیرد و از نتایج این مراحل میانگین گرفته می‌شود. تعداد دسته ها و عددSeed و حد آستانه در الگوریتم انتخاب ویژگی با تابع ارزیاب Wrapper با طبقه کننده C4.5 به ترتیب برابر ۱۰ و ۱و ۰/۰۱ است. همچنین برای ایجاد درخت تصمیم گیری مبتنی برالگوریتم ژنتیک نیز از عدد اعتبارسنجی متقاطع ۱۰ استفاده شده است.

۸- مقایسه نتایج درخت تصمیم گیری مدل تلفیقی پیشنهادی با سایر درخت تصمیم گیری

تا این بخش به ارائه مدل تلفیقی پیشنهادی برای ساخت درخت تصمیم گیری نهایی به منظور اعتبار سنجی مشتریان بانک پرداخته شد. این مدل به طور مختصر راحل زیر را برای ساخت درخت تصمیم گیری نهایی در اعتبار سنجی مشتریان بانک انجام می دهد:

۱. خوشه بندی داده ها

۲. انتخاب ویژگی ها توسط الگوریتم های متا و درخت تصمیم گیری ژنتیکی در هر خوشه

۳.ساخت درختان تصمیم گیری C4.5 از مجموعه ویژگی های منتخب از هر الگوریتم انتخاب ویژگی در هر خوشه

۴.بکارگیری استراتژی مناسبی به منظور انتخاب درخت تصمیم گیری C4.5 بهتر مبتنی بر معیارهای بهینگی در هر خوشه

۵.ترکیب دو درخت تصمیم گیری منتخب در دو خوشه و ساخت درخت تصمیم گیری نهایی در اعتبار سنجی مشتریان بانک.

از مجموعه داده های اعتباری بانک توسعه تعاون استان اردبیل بعد از اعمال آماده سازی و تمیز کردن داده ها برای ساخت در خت تصمیم گیری مدل تلفیقی پیشنهادی و سایر درختان تصمیم گیری مقایسه شده مطرح در این پژوهش استفاده شده است. جداول(۲)الی ۱۲ به تبیین نتایج حاصل از اجرای الگوریتم طبقه بندی این درختان تصمیم گیری می پردازد. جدول(۲)نتایج حاصل از اجرای الگوریتم های ساخت درخت تصمیم گیری C4.5 مبتنی بر مدل تلفیقی پیشنهادی در خوشه اول را نشان می دهد.

جدول ۳: نتایج حاصل از اجرای الگوریتم‌های ساخت درخت تصمیم گیری C4.5 مبتنی برمدل تلفیقی پیشنهادی در خوشه اول

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی های پیشگو منتخب	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	wrapperبا طبقه کننده C۴.۵	۴۴۵	۱۱	۳۶۵	٪۸۲.۰۲۲۵	۲۸	۴۲	۰.۸۲	۰.۸۲۱
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۴۴۵	۵	۳۵۴	٪۷۹.۵۵۰۶	۱۸	۲۶	۰.۸۰۹	۰.۷۶۶
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۴۴۵	۱۴	۳۰۲	٪۶۷.۸۶۵۲	۴۳	۶۳	۰.۷۰۵	۰.۶۰۲
۴	طبقه کننده C۴.۵	۴۴۵	۱۴	۳۵۸	٪۸۰.۴۴۹۴	۳۶	۵۵	۰.۸۰۶	۰.۸۰۲
۵	مبتنی بر درخت تصمیم گیری ژنتیکی	۴۴۵	۶	۳۴۳	٪ ۷۷.۰۷۸۷	۵	۸	۰.۷۳۵	۰.۹۷۱

بیشترین درصد مشاهدات درست طبقه بندی شده مربوط به درخت تصمیم گیری است که در انتخاب ویژگی ها از تابع ارزیاب Wrapper با طبقه کننده C4.5 استفاده می کند. کمترین پیچیدگی مربوط به درخت تصمیم گیری است که در انتخاب ویژگی ها از درخت تصمیم گیری ژنتیکی استفاده کرده است. زیرا درخت تصمیم گیری آن دارای کمترین تعداد برگ ها و اندازه درخت است. دقت طبقه بندی این درخت تصمیم گیری نسبت به بسیاری از درختان تصمیم گیری جدول (۳)کمترین مقدار است. به نظر میرسد بهترین درخت در این خوشه درخت تصمیم گیری باشد که در انتخاب ویژگی ها از تابع ارزیاب Wrapper با طبقه کننده C4.5 استفاده می کند. زیرا بالاترین دقت را در طبقه بندی داشته و زیاد بودن پیچیدگی آن در مقابل دقت بالای آن قابل چشم پوشی است. در نظر گرفتن تعادل بین معیارهای بهینگی در انتخاب بهترین درختان تصمیم گیری می تواند به عهده کارشناس اعتبار سنجی باشد که بر طبق نظر و تصمیم خود بهترین درخت را انتخاب کند. برای مجموعه داده های اعتباری در خوشه دوم نیز همانند خوشه اول به ساخت درخت تصمیم گیری C4.5 مبتنی بر مدل تلفیقی پیشنهادی این مقاله در خوشه دوم در جدول (۴) آورده شده است.

جدول ۴ : نتایج حاصل از اجرای الگوریتم های ساخت درخت تصمیم گیری C4.5 مبتنی برمدل تلفیقی پیشنهادی درخوشه دوم

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی های پیشگو منتخب	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	wrapperبا طبقه کننده C۴.۵	۲۴۵	۳	۲۰۳	٪۸۲.۸۵۷۱	۶	۹	۰.۸۴۹	۰.۷۹۱
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۲۴۵	۳	۲۰۳	٪۸۲.۸۵۷۱	۶	۹	۰.۸۴۹	۰.۷۹۱
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۲۴۵	۸	۱۸۴	٪۷۵.۱۰۲	۱۹	۲۷	۰.۷۵۳	۰.۷۴۶
۴	طبقه کننده C۴.۵	۲۴۵	۹	۱۹۵	٪۷۹.۵۹۱۸	۱۹	۲۹	۰.۸۲۸	۰.۷۳۹
۵	مبتنی بر درخت تصمیم گیری ژنتیکی	۲۴۵	۳	۱۹۰	٪۷۷.۵۵۱	۷	۱۰	۰.۷۴۵	۰.۹۱۱

جدول (۴) نتایج حاصل از اجرای الگوریتم های تصمیم گیری C4.5 را در خوشه دوم مدل تلفیقی پیشنهادی نشان می دهد . درختان تصمیم گیری با توابع ارزیاب Wrapper با طبقه کننده C4.5 و همبستگی بین ویژگی ها باهم و با ویژگی هدف دارای بالاترین دقت طبقه بندی و کمترین پیچیدگی هستند . بعد از انتخاب بهترین درختان تصمیم گیری در هر خوشه نوبت به ادغام این دو درخت و در نهایت ساخت درخت تصمیم گیری نهایی برای اعتبار سنجی مشتریان بانک مبتنی بر مدل تلفیقی پیشنهادی میرسد. جدول(۴)نتایج حاصل از اجرای الگوریتم مدل تلفیقی پیشنهادی در ساخت درخت تصمیم گیری نهایی برای اعتبار سنجی مشتریان بانک نشان می دهد. کل مشاهدات در این جدول برابر با مجموع کل مشاهدات در خوشه های اول و دوم است.

تعداد ویژگی هالی پیشگو منتخب در الگوریتم طبقه بندی مدل پیشنهادی برابر اجتماع ویژگی های پیشگو منتخب در الگوریتم درختان تصمیم گیری انتخاب شده در هر خوشه به علاوه عدد یک است. ویژگی های پیشگو منتخب در دو درخت تصمیم گیری بهتر در خوشه دوم (ردیف ۱ و ۲ جدول ۴) با هم یکی است. از طرفی این ۳ ویژگی پیشگو درون ویژگی های پیشگو منتخب درخت تصمیم گیری بهینه خوشه اول قرار دارند. پس تعداد ویژگی های پیشگو منتخب در درخت تصمیم گیری مدل تلفیقی پیشنهادی برابر ۱۱ به علاوه ۱ یعنی ۱۲ است. عدد یک در اینجا به ویژگی «نوع خوشه» اشاره دارد. زیرا در مدل تلفیقی پیشنهادی به منظور طبقه بندی مشتریان اعتبار سنجی نوع خوشه آن ها در ابتدا تعیین می شود.

جدول ۵: نتایج حاصل از اجرای الگوریتم های ساخت درخت تصمیم گیری C4.5 مبتنی برمدل تلفیقی پیشنهادی

ردیف	الگوریتم درخت تصمیم گیری	کل مشاهدات	تعداد ویژگی های پیشگو منتخب	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	مدل تلفیقی پیشنهادی طبقه بندی مشتریان بانک ها با داده های اعتباری	۶۹۰	۱۲	۵۶۸	٪ ۸۲.۳۱۸۸	۳۴	۵۲	۰.۸۲۹۸	۰.۸۰۹۱

تعداد مشاهدات درست طبقه بندی شده در مدل تلفیقی پیشنهادی برابر مجموع تعداد مشاهدات درست طبقه بندی شده درختان تصمیم گیری منتخب در خوشه های اول و دوم است. برای تعیین دقت کلاس مشتریان خوب درخت تصمیم گیری مدل تلفیقی پیشنهادی به صورت زیر عمل شد:

تعداد مشتریان درست طبقه بندی شده در کلاس مشتریان خوب هر دو خوشه با هم جمع شد؛ مقدار این عدد بر مجموع مشتریان طبقه بندی شده در کلاس مشتریان خوب هر دو خوشه تقسیم شد. به همین ترتیب دقت کلاس مشتریان بد درخت تصمیم گیری مدل تلفیقی پیشنهادی به صورت زیر بدست آمد: مجموع تعداد مشتریان درست طبقه بندی شده در کلاس مشتریان بد هر دو خوشه، تقسیم بر مجموع تعداد مشتریان طبقه بندی شده در کلاس مشتریان بد هر دو خوشه. تعداد برگ ها در درخت تصمیم گیری نهایی مدل تلفیقی پیشنهادی توسط مجموع تعداد برگ های درختان تصمیم گیری منتخب در هر خوشه بدست می آید. اندازه درخت تصمیم گیری مدل تلفیقی پیشنهادی به صورت زیر تعیین می شود: مجموع اندازه درختان تصمیم گیری منتخب در هر دو خوشه به علاوه ۱ . عدد یک در اینجا بر گره تصمیم گیرنده «نوع خوشه» دلالت دارد. این ویژگی در ابتدای درخت تصمیم گیری مدل تلفیقی پیشنهادی به تعیین نوع خوشه به منظور طبقه بندی توسط درخت تصمیم گیری منتخب در هر خوشه می پردازد. ویژگی «نوع خوشه» یک ویژگی اسمی می باشد و در مدل تلفیقی پیشنهادی دارای مقادیر «خوشه اول» و «خوشه دوم» است. درخت تصمیم گیری نهایی مدل تلفیقی پیشنهادی تحقیق پیش رو در ابتدا دارای گرهی می باشد که این گره مربوط به ویژگی «نوع خوشه» است. به عبارت دیگر با استفاده از این درخت در ابتدا نوع خوشه مشتری اعتبار سنجی جدید تعیین می شود. سپس با در نظر گرفتن نوع خوشه، بهترین درخت تصمیم گیری در هر خوشه برای طبقه بندی و اعتبار سنجی مشتریان بانک ها مورد استفاده قرار می گیرد. در ادامه به مقایسه مدل تلفیقی پیشنهادی با سایر روش های ساخت درخت تصمیم گیری C4.5 پرداخته می شود. مقادیر پارامترهای این روش ها همانند مقادیر پارامترهای مدل تلفیقی پیشنهادی است. جدول ۹ نتایج درخت تصمیم گیری C4.5 را نشان می دهد که در آن از الگوریتم های انتخاب ویژگی ها و خوشه بندی به عنوان روش های پیش پردازش داده ها استفاده نمی شود.

جدول ۶ نتایج حاصل از اجرای درخت تصمیم گیری C4.5 بدون اعمال الگوریتم های انتخاب ویژگی ها و خوشه بندی

ردیف	الگوریتم درخت تصمیم گیری	کل مشاهدات	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	درخت تصمیم گیری C۴.۵ بدون اعمال الگوریتم های انتخاب ویژگی و خوشه بندی	۶۹۰	۵۷۳	٪۸۳.۰۴۳۵	۴۰	۶۰	۰.۸۲۶	۰.۸۴۱

همانطور که در جدول(۶)مشاهده می شود، دقت طبقه بندی این درخت به میزان کمی از درخت تصمیم گیری مدل تلفیقی پیشنهادی بیشتر است؛ ولی از طرف دیگر پیچیدگی درخت تصمیم گیری مدل تلفیقی پیشنهادی از درخت تصمیم گیری جدول ۹ کمتر است. در صورتی که در ساخت درختان تصمیم گیری تنها از الگوریتم های انتخاب ویژگی مبتنی بر الگوریتم ژنتیک استفاده شود و الگوریتم خوشه بندی به کار نرود، نتایج به صورت جدول (۷)قابل مشاهده است.

جدول ۷: نتایج حاصل از اجرای درخت تصمیم C4.5 با بکارگیری انتخاب ویژگی ها و بدون اعمال خوشه بندی

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	wrapperبا طبقه کننده C۴.۵	۶۹۰	۵۷۰	٪۸۲.۶۰۸۷	۱۴	۲۱	۰.۸۲۶	۰.۸۲۵
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۶۹۰	۵۵۳	٪۸۰.۱۴۴۹	۲۲	۳۴	۰.۸۰۹	۰.۷۸۴
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۶۹۰	۴۸۹	٪۷۰.۸۶۹۶	۶۸	۱۰۱	۰.۷۲۳	۰.۶۶۷
۴	طبقه کننده C۴.۵	۶۹۰	۵۶۷	٪۸۲.۱۷۳۹	۴۹	۷۳	۰.۸۱۵	۰.۸۱۵

طبقه بندی و پیچیدگی درخت تصمیم گیری C4.5 ردیف ۱ که از تابع ارزیابWrapper با طبقه کننده C4.5 در انتخاب ویژگی ها استفاده می کند، از درخت تصمیم گیری مدل تلفیقی پیشنهادی بهتر است. همچنین دقت طبقه بندی درخت تصمیم گیری با تابع ارزیاب طبقه کننده C4.5 از دقت طبقه بندی درخت تصمیم گیری مدل مدل تلفیقی پیشنهادی بیشتر است؛ ولی درخت تصمیم گیری مدل تلفیقی پیشنهادی دارای پیچیدگی کمتری نسبت به این درخت تصمیم گیری است. پیچیدگی درخت تصمیم گیری با تابع ارزیاب ردیف ۲ از جدول ۱۰ از پیچیدگی درخت تصمیم گیری مدل تلفیقی پیشنهادی کمتر است؛ ولی دقت طبقه بندی درخت تصمیم گیری مدل تلفیقی پیشنهادی از این درخت بیشتر می باشد. اگر در انتخاب ویژگی ها از الگوریتم جستجوی اول بهترین به جای الگوریتم ژنتیک استفاده شود، نتایج به صورت جدول (۸)است. روش جستجوی اول بهترین، برای انتخاب ویژگی ها با مجموعه تهی از ویژگی ها شروع می کند و به کشف همه زیر مجموعه های ممکن می پردازد و این کار توسط اضافه کردن تک تک ویژگی ها انجام می دهد. زیر مجموعه با بالاترین دقت انتخاب می شود و این روند تا جایی ادامه می یابد که دیگر بهبودی حاصل نشود. روش جستجوی اول بهترین از الگوریتم انتخاب رو به جلو برای انتخاب ویژگی ها استفاده کرده و ضریب تعداد ویژگی ها در مجموعه داده برابر یک و مقدار بازگشت به عقب برابر عدد ۵ است.

جدول ۸: نتایج حاصل از اجرای درخت تصمیم گیری C4.5 با انتخاب ویژگی ها مبتنی بر جستجوی اول بهترین و بدون اعمال خوشه بندی

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	Wrapperبا طبقه کننده C۴.۵	۶۹۰	۵۷۰	٪ ۸۲.۶۰۸۷	۱۴	۲۱	۰.۸۲۶	۰.۸۲۵
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۶۹۰	۵۵۶	٪ ۸۰.۵۷۹۷	۱۱	۱۸	۰.۸۱۷	۰.۷۸۲
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۶۹۰	۵۶۷	٪ ۸۲.۱۷۳۹	۳۸	۵۷	۰.۸۱۵	۰.۸۳۹
۴	طبقه کننده C۴.۵	۶۹۰	۵۷۳	٪ ۸۳.۰۴۳۵	۴۲	۶۴	۰.۸۲۱	۰.۸۵۴

دقت طبقه بندی درختان تصمیم گیری با توابع ارزیاب ردیف های ۱ و ۴ از دقت طبقه بندی درخت تصمیم گیری مدل تلفیقی پیشنهادی بیشتر است. ولی پیچیدگی درخت تصمیم گیری مدل تلفیقی پیشنهادی از پیچیدگی درخت تصمیم گیری با تابع ارزیاب ردیف ۴کمتر است. می توان ویژگی «نوع خوشه» را در ساخت درخت تصمیم گیری در نظر گرفت. برای این کار به مجموعه داده اعتباری آلمان یک ویژگی پیشگو با نام ویژگی «نوع خوشه» اضافه می شود. جدول (۹) نتایج حاصل از اجرای این الگوریتم را نشان می دهد. در اینجا ابتدا توسط الگوریتم های انتخاب ویژگی با توابع ارزیاب مطرح در جدول(۹)، ویژگی های مناسب انتخاب شده و سپس درختان تصمیم گیری C4.5 ساخته می شود.

جدول ۹: نتایج حاصل از اجرای درخت تصمیم گیری C4.5 با انتخاب ویژگی ها مبتنی بر جستجوی الگوریتم ژنتیک با در نظر گرفتن ویژگی نوع خوشه

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	Wrapperبا طبقه کننده C۴.۵	۶۹۰	۵۷۴	٪ ۸۳.۱۸۸۴	۲۴	۳۷	۰.۸۱۶	۰.۸۷۴
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۶۹۰	۵۵۸	٪ ۸۰.۸۶۹۶	۱۱	۱۸	۰.۸۱۹	۰.۷۸۶
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۶۹۰	۵۱۵	٪ ۷۴.۶۳۷۷	۳۹	۵۶	۰.۷۷۱	۰.۶۹۳
۴	طبقه کننده C۴.۵	۶۹۰	۵۵۷	٪ ۸۰.۷۲۴۶	۵۱	۷۷	۰.۸۰۶	۰.۸۱۱

دقت طبقه بندی درخت تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 از دقت درخت تصمیم گیری مدل تلفیقی پیشنهادی بیشتر و پیچیدگی آن کمتر است.

۹-نتیجه گیری

این پژوهش با هدف تعیین مدلی بهینه برای درجه بندی ریسک اعتباری مشتریان بانک توسعه تعاون استان اردبیل در بخش مشتریان حقیقی صورت گرفته است. دقت طبقه بندی درخت تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 از دقت درخت تصمیم گیری مدل تلفیقی پیشنهادی بیشتر و پیچیدگی آن کمتر است . در این بخش از پژوهش به ارائه نتایج حاصل از اجرای مدل تلفیقی پیشنهادی در آزمون و ساخت درخت تصمیم گیری برای اعتبار سنجی مشتریان بانک پرداخته شده است. این نتایج به نتایج حاصل از اجرای سایر روش های ساخت درختان تصمیم گیری مقایسه شده است. می توان در انتخاب درختان تصمیم گیری بهتر مبتنی بر معیارهای بهینگی در هر خوشه از نظرات کارشناسان اعتبار سنجی استفاده کرده و تعادل بین دقت طبقه بندی و پیچیدگی درختان تصمیم گیری را در انتخاب بهترین درخت تصمیم گیری لحاظ کرد.

درخت تصمیم گیری مدل تلفیقی پیشنهادی با تعداد ۱۳ درخت تصمیم گیری دیگر در مجموعه داده اعتباری بانک توسعه تعاون استان اردبیل مقایسه شده است. دقت طبقه بندی درخت تصمیم گیری این مدل از دقت طبقه بندی ۸ درخت تصمیم گیری دیگر در مجموعه داده اعتباری آلمان مقایسه شد . دقت طبقه بندی درخت تصمیم گیری این مدل از دقت طبقه بندی ۸ درخت تصمیم گیری مقایسه شده بیشتر بوده است. همچنین پیچیدگی درخت تصمیم گیری مدل تلفیقی پیشنهادی از پیچیدگی ۷ درخت تصمیم گیری مقایسه شده کمتر بود. تنها ۳ درخت تصمیم گیری که در انتخاب ویژگی ها از تابع ارزیاب Wrapper با طبقه کننده C4.5 استفاده کردند، دارای دقت طبقه بندی و پیچیدگی بهتری نسبت به درخت تصمیم گیری مدل تلفیقی پیشنهادی بوده اند. هر چه میزان مالکیت متقاضی بر دارایی‌ها بیشتر و مبلغ بدهی به ویژه بدهی های آنی و کوتاه‌مدت کمتر باشد، حصول اطمینان نسبت به برگشت منابع بانک بیشتر خواهد بود. به طوری که بررسی‌ها نشان داد، در بسیاری از مطالعات تجربی بانک‌های موفق خارجی بیش از ده‌ها هزار مشاهده مورد استفاده قرار گرفته است. حجم اطلاعاتی که ما در اختیار داشتیم بسیار اندک بود و به نظر می‌رسید که مدل پیشنهادی تا حدودی می‌توان مشکل کم بودن اطلاعات را مرتفع سازد.

منابع

- جواهری کامل، م؛ ع. اسعدی. و م. کوثرنشان(۱۳۸۸)مدیریت دانش در تحقیقات پلیس. توسعهٔ انسانی پلیس، ۱۰۷-۱۲۴.
-راستی،م،ر.اختیاری،م(۱۳۹۰)." تصمیم گیری گروهی برای رتبه بندی اعتباری مشتریان"،بانک سپه، شماره ۱۲۲ ،خرداد ۱۳۹۰،ص۲۶_۳۲.
- رشیدیان،س.(۱۳۹۰)."طبقه بندی مشتریان شبکه بانکی بر اساس ریسک اعتباری با استفاده از مدل های پیش بینی و تصمیم گیری چند معیاره"،پایان نامه کارشناسی ارشد،دانشکده مدیریت و حسابداری،دانشگاه آزاد اسلامی_واحد سنندج.
-حاجی آقایی،ب.(۱۳۸۶)." انواع ریسک و پوشش های آن"،تازه های اقتصاد، ۱۳۸۷ سال ششم،شماره ۱۲۲ ، ص ۵۸ _ ۶۵.
-خالصی، نرگس و شکوهی، امیرحسین،۱۳۸۹،ارائه روشی جدید برای اعتبار سنجی مشتریان بانکی با استفاده ازتکنیک های داده کاوی،چهارمین کنفرانس داده کاوی ایران،تهران،، https://civilica.com/doc/109117

-سلیمانی، پروانه و فکری، مهدی و ایمانی، مهدی و مظهری، نیلوفر،۱۳۸۹،ارائه روشی جدید برای سیست مهای تشخیص چهره و بازیابی تصاویر با استفاده از ویژگیهای معنایی در تصویرکاوی،چهارمین کنفرانس داده کاوی ایران،تهران،،https://civilica.com/doc/109061

-فیلی، حمیدرضا و ظهوری، سپیده و بیجاری، ریحانه،۱۳۸۹،ارائه روشی مبتنی بر قوانین فازی جهت شناسایی رفتار مشتری در حوزه بازاریابی،چهارمین کنفرانس داده کاوی ایران،تهران،،https://civilica.com/doc/109077

-خیابانی، ایدا و شهرابی، جمال و علیان نژاد، رسول و صالحی، مهسا،۱۳۸۹،ارائه مدل داده کاوی جهت تخمین میزان تحمل بار ورزشی در جانبازان شیمیایی براساس مقادیر پارامترهای تست اسپیرومتری،چهارمین کنفرانس داده کاوی ایران،تهران،،https://civilica.com/doc/109137

-فرقانی، محمدعلی و حیدری، خسرو و نوجوان، صمد،۱۳۸۹،ارائه مدلی برای ا صلاح الگوی تولید با استاندارد سازی محصولات، قطعات و فرایندها رویکرد داده کاوی،چهارمین کنفرانس داده کاوی ایران،تهران،،https://civilica.com/doc/109127

-Chen, M.S., Han, J., Yu, P.S. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8, 866-883. (2007)

-Azgomi, H., & Sohrabi, M. K. (2019). Finding similar documents using frequent pattern mining methods. International Journal of Uncertainty, Fuzziness and KnowledgeBased Systems.

-Azgomi, H., & Sohrabi, M. K. (2018). RTLTDS Dataset. 10.13140/RG.2.2.35186.81608.

-Azgomi, H. (2018). RTLTDS Data Warehouse. 10.13140/RG.2.2.35513.88166.

-Sohrabi, M. K., & Azgomi, H. (2018). Parallel set similarity join on big data based on Locality-Sensitive Hashing. Science of Computer Programming, 145, 1-12

-Hassan, Mohammed Mehedi, Md Zia Uddin, Amr Mohamed, and Ahmad Almogren. A robust human activity recognition system using smartphone sensors and deep learning. Future Generation Computer Systems, Vol. 81, pp. 307-313, 2018

-Basu, Sayantani, Marimuthu Karuppiah, K. Selvakumar, Kuan-Ching Li, SK Hafizul Islam, Mohammad Mehedi Hassan, and Md Zakirul Alam Bhuiyan. An intelligent/cognitive model of task scheduling for IoT applications in cloud computing environment. Future Generation Computer Systems, Vol. 88, pp. 254-261, 2018