نشریه مدیریت فناوری اطلاعات، دوره ۲، شماره ۴، بهار و تابستان ۱۳۸۹، از صفحه ۲۳ تا ۳۸

به‌کارگیری الگوریتم ژنتیک در بهینه‌سازی درختان تصمیم‌گیری برای اعتبارسنجی مشتریان بانک‌ها

محمود البرزی[۱]، محمد ابراهیم محمد پورزرندی[۲]، محمد خان‌بابایی[۳*]

۱. استادیار علمی دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران
۲. دانشیار دانشگاه آزاد اسلامی واحد تهران مرکز، تهران، ایران
۳. دانش‌آموخته کارشناسی ارشد دانشگاه آزاد اسلامی، واحد علوم و تحقیقات، تهران، و عضو باشگاه پژوهشگران جوان، ایران

(تاریخ دریافت: ۱۳۸۸/۹/۸، تاریخ تصویب: ۱۳۸۹/۳/۲۴)

چکیده

درختان تصمیم‌گیری به عنوان یکی از تکنیک‌های داده‌کاوی می‌توانند به اعتبارسنجی مشتریان بانکی بپردازند. مسئله‌ی اصلی ساخت درختان تصمیم‌گیری است که بتوانند به طور بهینه مشتریان را طبقه‌بندی کنند. در این مقاله یک مدل مناسب اعتبارسنجی مشتریان بانک‌ها برای اعطای تسهیلات اعتباری متناسب با هر طبقه مبتنی بر الگوریتم ژنتیک ارایه می‌شود. الگوریتم‌های ژنتیک می‌توانند با انتخاب ویژگی‌های مناسب و ساخت درختان تصمیم‌گیری بهینه به اعتبارسنجی مشتریان کمک کنند. در ساخت این مدل فرآیند توسعه در شناخت الگو و فرآیند CRISP برای اعتبارسنجی مشتریان به‌کار رفته است. مدل طبقه‌بندی پیشنهادی مبتنی بر تکنیک‌های خوشه‌بندی، انتخاب ویژگی‌ها، درختان تصمیم‌گیری و الگوریتم ژنتیک است. این مدل به انتخاب و ترکیب بهترین درختان تصمیم‌گیری مبتنی بر معیارهای بهینگی و ساخت درخت تصمیم‌گیری نهایی برای اعتبارسنجی مشتریان می‌پردازد. نتایج نشان می‌دهد که دقت طبقه‌بندی مدل طبقه‌بندی پیشنهادی به‌طورتقریبی از تمام مدل‌های درخت تصمیم‌گیری مقایسه شده در این مقاله بالاتر است. هم‌چنین تعداد برگ‌ها و اندازه‌ی درخت تصمیم‌گیری و در نتیجه‌ی پیچیدگی آن از همه کمتر است.

واژه‌های کلیدی: اعتبارسنجی، الگوریتم ژنتیک، انتخاب ویژگی‌ها، درختان تصمیم‌گیری، خوشه‌بندی

* نویسنده مسئول
Email: mohammadkhanbabaei@yahoo.com

۱. مقدمه

بانک‌ها به‌منظور آگاهی از نیازمندی‌ها و رفتار مشتریان خود در اعطای تسهیلات اعتباری باید به شناسایی ویژگی‌های آن‌ها بپردازند. این امر منجر به کاهش ریسک‌های بانکی از جمله ریسک اعتباری می‌شود. پژوهش‌ها و کاربردهای متعددی در حوزه اعتبارسنجی برای شناسایی مشتریان خوب و بد بانک‌ها صورت گرفته است. روش قضاوتی در اعتبارسنجی به دلیل خطا و زمان زیاد به تدریج جای خود را به روش‌های پارامتریک و ناپارامتریک داد[۲۴]. روش‌های پارامتریک مثل پروبیت، لوجیت، تحلیل تمایزی و رگرسیون لجستیک از ابتدای ظهور اعتبارسنجی مورد استفاده قرار گرفتند و سپس روش ناپارامتریک و داده کاوی مثل درختان تصمیم‌گیری، شبکه‌های عصبی و سیستم‌های خبره به‌کار گرفته شدند[۲۰]. درختان تصمیم‌گیری یکی از تکنیک‌های داده‌کاوی با قابلیت فهم بالا و سرعت مناسب در یادگیری الگو، می‌توانند برای طبقه‌بندی مشتریان در اعتبارسنجی مفید باشند.

امروزه مسئله‌ی اعتبارسنجی به یکی از مسایل مهم مدیران و کارشناسان بانکی تبدیل شده است. در یک بانک، می‌توان درختان تصمیم‌گیری متنوعی را برای طبقه‌بندی و اعتبارسنجی مشتریان ایجاد نمود. از ویژگی‌های مدل پیشنهادی می‌توان به موارد زیر اشاره نمود:

افزایش دقت در ساختار و محتوای درختان تصمیم‌گیری توسط الگوریتم ژنتیک، بهبود درختان تصمیم‌گیری برای فهم آسان مدل طبقه‌بندی مشتریان، جلوگیری از اتخاذ تصمیم‌های احتمالی غلط کارشناسان بانکی در اعتبارسنجی، کاهش نیاز به تحلیل‌های پرهزینه و زمان‌بر در طبقه‌بندی مشتریان، انتخاب بهینه ویژگی‌های اعتبارسنجی مشتریان و در نهایت رضایت‌مندی آن‌ها در ارایه‌ی تسهیلات اعتباری متناسب با هر طبقه. استفاده از الگوریتم‌های ژنتیک در بهینه‌سازی درختان تصمیم‌گیری و انتخاب ویژگی‌ها علاوه بر حل مسایل مطرح شده این مزیت را نیز دارد: الگوریتم ژنتیک در یک زمان با مجموعه‌ای از راه‌حل‌ها کار می‌کند ولی الگوریتم‌های استنتاج حریصانه راه حل جزیی را در هر مرحله بررسی می‌کنند[۴].

مسئله‌ی اصلی ساخت درختان تصمیم‌گیری است که بتوانند به‌طور بهینه به طبقه‌بندی مشتریان خوب و بد بانک‌ها بپردازند. به‌نظر می‌رسد استفاده از تکنیک‌های بهینه‌سازی مثل

الگوریتم‌های ژنتیک در انتخاب ویژگی‌ها و ایجاد درختان تصمیم بهینه برای اعتبارسنجی مشتریان بانک‌ها مفید است. الگوریتم‌های انتخاب ویژگی ممکن است در بهینه محلی قرار گیرند و از طرف دیگر تعامل بین ویژگی‌ها را درنظر نمی‌گیرند و فرض می‌کنند، روابط بین ویژگی‌ها خطی بوده و ویژگی‌ها مستقل از هم می‌باشند. هم‌چنین این الگوریتم‌ها تنها از برخی معیارها برای انتخاب ویژگی استفاده می‌کنند. الگوریتم‌های درختان تصمیم‌گیری در فرایند رشد درخت حریص هستند[۸]. در برخی از الگوریتم‌ها حساسیت بیشتری نسبت به برخی از ویژگی‌ها برای تفکیک شدن وجود دارد. آن‌ها درختانی با اندازه‌ی بزرگ و پیچیده تولید می‌کنند که در نتیجه دارای دقت کم در طبقه‌بندی در مجموعه‌ی داده تست و ارزیابی و تناسب بیش از حد است. درختان تصمیم‌گیری نسبت به داده‌های زیاد در ویژگی‌ها تمایل نشان می‌دهند و در قسمت‌های پایینی درخت تمایلی به افزایش دقت و کاهش تناسب بیش از حد ندارند. این الگوریتم‌ها در فرایند خود تنها از برخی معیارها و توابع در ساخت درخت تصمیم‌گیری استفاده می‌کنند.

در این مقاله مدلی مناسب برای اعتبارسنجی مشتریان بانک‌ها مانند بانک ملت برای اعطای تسهیلات اعتباری متناسب با هر طبقه مبتنی بر الگوریتم ژنتیک ارایه می‌شود.

۲. پیشینه‌ی پژوهشی

پژوهش‌های متنوعی روی کاربرد روش‌های پارامتریک و ناپارامتریک در اعتبارسنجی صورت گرفته است. نگاره‌ی ۱ به چند نمونه از این پژوهش‌ها در خارج کشور اشاره دارد.

نگاره‌ی ۱. برخی پژوهش‌های روش‌های اعتبارسنجی در خارج از کشور

ردیف	مدل اعتبارسنجی	پژوهشگر، تاریخ
1	رگرسیون لجستیک، شبکه عصبی، درخت تصمیم‌گیری	(Susac, Sarlija, & Bensic, n.d.)[22]
2	ترکیب تحلیل تمایزی و الگوریتم پس انتشار در شبکه عصبی	(Lee, Chiu, Lu, & Chen, 2002)[14]
3	الگوهای طبقه‌بندی غلط	(Kim & Sohn, 2004)[12]
4	ترکیب مدل‌های شبکه‌های عصبی مصنوعی و روش MARS	(Lee & Chen, 2005)[13]
5	رتبه‌بندی تحلیل لینک با استفاده از ماشین بردار پشتیبان	(Xu, Zhou, & Wang, 2008)[26]
6	شبکه‌های عصبی و تکنیک‌های عمومی	(Abdou & Pointon, 2008)[3]
7	طبقه‌کننده‌های ترکیبی به جای یک طبقه‌کننده	(Nanni & Lumini, 2009)[17]

نگاره‌های ۲ و ۳ به ترتیب به به‌کارگیری الگوریتم‌های ژنتیک در انتخاب ویژگی‌ها و ساخت درختان تصمیم‌گیری اشاره دارد. این پژوهش‌ها مربوط به سایر حوزه‌های علوم و کسب و کار است. از نتایج آن‌ها می‌توان در اعتبارسنجی مشتریان بانک‌ها استفاده نمود.

نگاره ی ۲. برخی پژوهش‌ها در کاربرد الگوریتم ژنتیک در انتخاب ویژگی‌ها

ردیف	رویکرد به‌کارگیری الگوریتم ژنتیک در انتخاب ویژگی‌ها	پژوهشگر، تاریخ
۱	ترکیب طبقه‌بندی چندگانه مبتنی بر الگوریتم ژنتیک	[2,11](۱۳۸۷،نادعلی & خان بابایی ،Kim &Kim ,Lee,2002 cited in)
۲	استفاده از الگوریتم ژنتیک برای انتخاب متغیرهای ورودی	(D'heygere, Goethals, & Pauw, 2003)[5]
۳	الگوریتم ژنتیک در انتخاب متغیرها به کمک خوشه‌بندی مشتریان	(Liu & Ong, 2008)[15]
۴	استفاده از الگوریتم ژنتیک در ترکیب روش‌های انتخاب ویژگی	(Tan, Fu, Zhang, & Bourgeois, 2008)[23]

نگاره‌ی ۳. برخی پژوهش‌ها در کاربرد الگوریتم ژنتیک در ساخت درختان تصمیم‌گیری

ردیف	رویکرد به‌کارگیری الگوریتم ژنتیک در ساخت درختان تصمیم‌گیری	پژوهشگر، تاریخ
1	ساخت درختان تصمیم‌گیری دودویی توسط تکنیک‌های تکاملی الگوریتم ژنتیک	(Papagelis & Kalles, n.d.)[19]
۲	استفاده از الگوریتم‌های ژنتیک برای یادگیری سلسله مراتب طبقه کننده‌ها	(Martinez-Otzeta, Sierra, Lazkano, & Astigarraga, 2006)[16]
۳	بهنیه‌سازی پیشگویی مدل‌ها بر مبنای درختان تصمیم و شبکه‌های عصبی	(D'heygere, Goethals, & Pauw, 2006)[6]
۴	بررسی سودمندی تکنیک درخت تصمیم‌گیری بر مبنای الگوریتم ژنتیک	(Huang, Gong, Shi, Liu, & Zhang, 2007)[10]
۵	تعریف برازندگی درخت توسط الگوریتم ژنتیک	[2,21] (۱۳۸۷،نادعلی & خان بابایی Janssens, & Sorensen (2003 cited in
۶	تحلیل درخت طبقه‌بندی توسط الگوریتم TARGET	(Gray & Fan, 2008)[8]
۷	الگوریتم ژنتیک چند هدفه Elitist به کشف قوانین طبقه‌بندی مجموعه‌ی داده‌های بزرگ	(Dehuri, Patnaik, Ghosh, & Mall, 2008)[7]

نقاط قوت مدل ارایه شده، در مقایسه با برخی مدل‌های مشابه داخلی و خارجی و به طور موردی شامل این موارد است. ۱. به‌کارگیری روش‌های آماده‌سازی و پیش پردازش داده‌ها. ۲. استفاده از خوشه‌بندی در پیش پردازش داده‌ها به منظور افزایش احتمالی دقت و کاهش پیچیدگی طبقه‌بندی مشتریان. ۳. تلفیق چندین الگوریتم انتخاب ویژگی مبتنی بر

رویکردهای فیلتر، Wrapper و طرح جاسازی شده برای افزایش انعطاف پذیری و دقت طبقه‌بندی در ساخت درختان تصمیم‌گیری به جای استفاده از یک طبقه کننده. ۴. تولید و مقایسه درختان تصمیم‌گیری متنوع و استفاده از آن‌ها در شرایط مختلف مطالعه‌ی موردی بر خلاف پژوهش‌های دیگر که تنها از یک نوع الگوریتم یا مدل در مطالعه‌ای خاص استفاده می‌شد. ۵. بهینه‌سازی درختان تصمیم‌گیری توسط الگوریتم ژنتیک بر خلاف تحقیقات قبل که تنها مدل طبقه‌بندی را ساخته و سعی در بهبود آن نداشتند. ۶. استفاده از معیار پیچیدگی در بهینگی درختان تصمیم‌گیری علاوه بر معیار دقت طبقه‌بندی. ۷. به‌کارگیری روش‌های هوش مصنوعی و شناخت الگو در اعتبارسنجی مشتریان بانک برای مقابله با شرایط پیچیده و لحاظ کردن روابط غیرخطی در طبقه‌بندی مشتریان و انتخاب ویژگی‌های اعتبارسنجی. ۸. وجود نگرش فرآیندی شناخت الگو و داده‌کاوی در اعتبارسنجی مشتریان بانک بر خلاف برخی پژوهش‌های دیگر درباره‌ی این حوزه.

۳. ویژگی‌های اعتباری مشتریان در مدل

مجموع ویژگی‌های اعتباری مشتریان دریافت‌کننده تسهیلات شامل موارد زیر است. ویژگی "کد نوع رکورد" ویژگی هدف است.

تاریخ تنظیم قرارداد(اسمی)، کد سرپرستی(اسمی)، شماره‌ی شعبه(اسمی)، شماره‌ی درخواست(اسمی)، نوع متقاضی(اسمی)، کد محلی(اسمی)، نام مشتری(اسمی)، شماره‌ی شناسنامه(اسمی)، تاریخ تولد(اسمی)، کد محل صدور شناسنامه(اسمی)، کد حوزه(اسمی)، کد نوع درخواست(اسمی)، تاریخ اولین سررسید(اسمی)، تاریخ آخرین سررسید(اسمی)، تعداد اقساط(عددی)، درصد سهم بانک(عددی)، مبلغ قرارداد(عددی)، مبلغ استفاده شده(عددی)، چگونگی استفاده از تسهیلات(اسمی)، مبلغ استفاده شده(عددی)، کد نوع وثیقه(اسمی)، ارزش وثیقه(عددی)، آخرین مانده‌ی بدهی(عددی)، کد نوع بخش(اسمی)، رشته‌ی فعالیت(اسمی)، کد نوع قرارداد با بانک(اسمی)، هدف از دریافت تسهیلات(اسمی)، محل مصرف تسهیلات(اسمی)، مورد مصرف تسهیلات(اسمی)، نوع رکورد تکلیفی یا غیر تکلیفی(اسمی)، تاریخ ایجاد رکورد(اسمی)، تاریخ(ماه و سال) (اسمی)، نام و نام خانوادگی برعکس‌شده(اسمی)، کد اعتباری(اسمی)، کد نوع رکورد(اسمی)، کد سرفصل حساب(اسمی) و تاریخ آخرین گردش(اسمی).

۴. تعریف عملیاتی مفاهیم مدل

درختان تصمیم‌گیری: درختان تصمیم شامل اجزایی است: ۱. گره‌ها که با نام‌های مشخصات یا ویژگی‌های اشیا برچسب گذاری شده اند. ۲. برگ‌ها که معادل طبقات مختلف هستند[۲۱]. یک درخت تصمیم شامل چند گره درونی و چند برگ است. همه گره‌های درونی شامل دو یا چند گره فرزند هستند[۲۱]. هر تصمیم در یک گره قرار دارد. گره آخر خروجی نهایی را نشان می‌دهد که در درخت تصمیم‌گیری دارای مقدار گسسته است[۹]. بهینگی درختان تصمیم‌گیری: در بررسی بهینگی درختان تصمیم‌گیری ۳ عامل در نظر گرفته می‌شود: ۱. نرخ نمونه‌های طبقه‌بندی شده‌ی صحیح(درصد مشاهده‌های درست طبقه‌بندی شده)

(۱) نرخ نمونه‌های طبقه‌بندی شده‌ی صحیح[۶] $CCI = \frac{(TP + TN)}{(TP + FP + FN + TN)} \times 100$.

۲. پیچیدگی درخت تصمیم‌گیری که شامل تعداد برگ‌ها و اندازه درخت است.

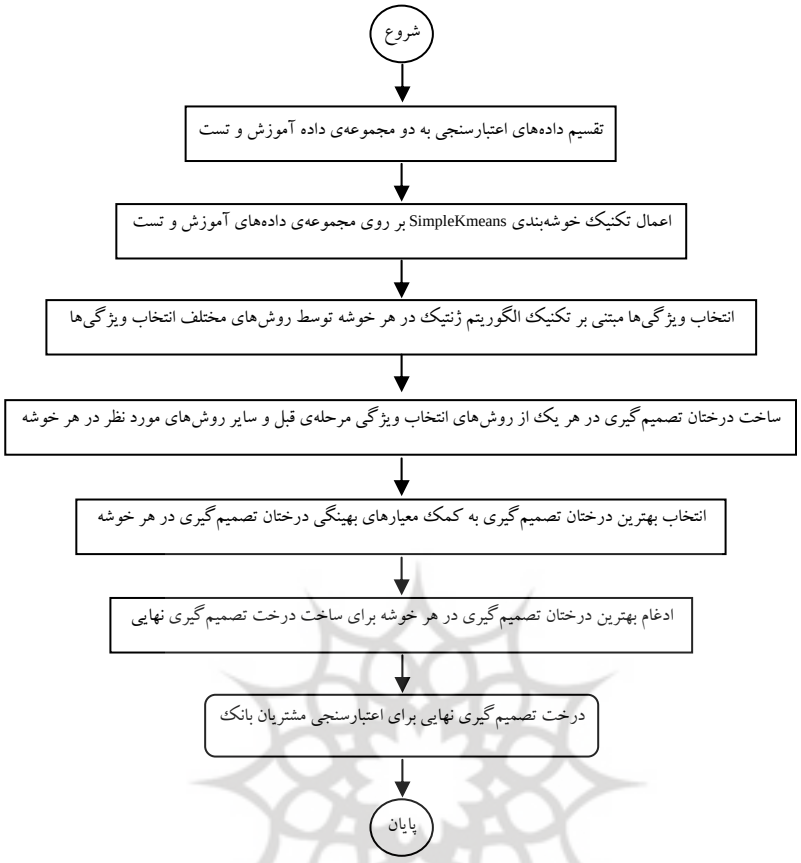
۳. تعداد ویژگی‌های پیشگو موجود در درخت تصمیم‌گیری.

اندازه‌ی درخت: به تعداد شاخه‌ها تا رسیدن به برگ در درخت تصمیم‌گیری ارتباط دارد و برابر با مجموع تعداد برگ‌ها و گره‌ها در یک درخت تصمیم‌گیری است[۲۵].

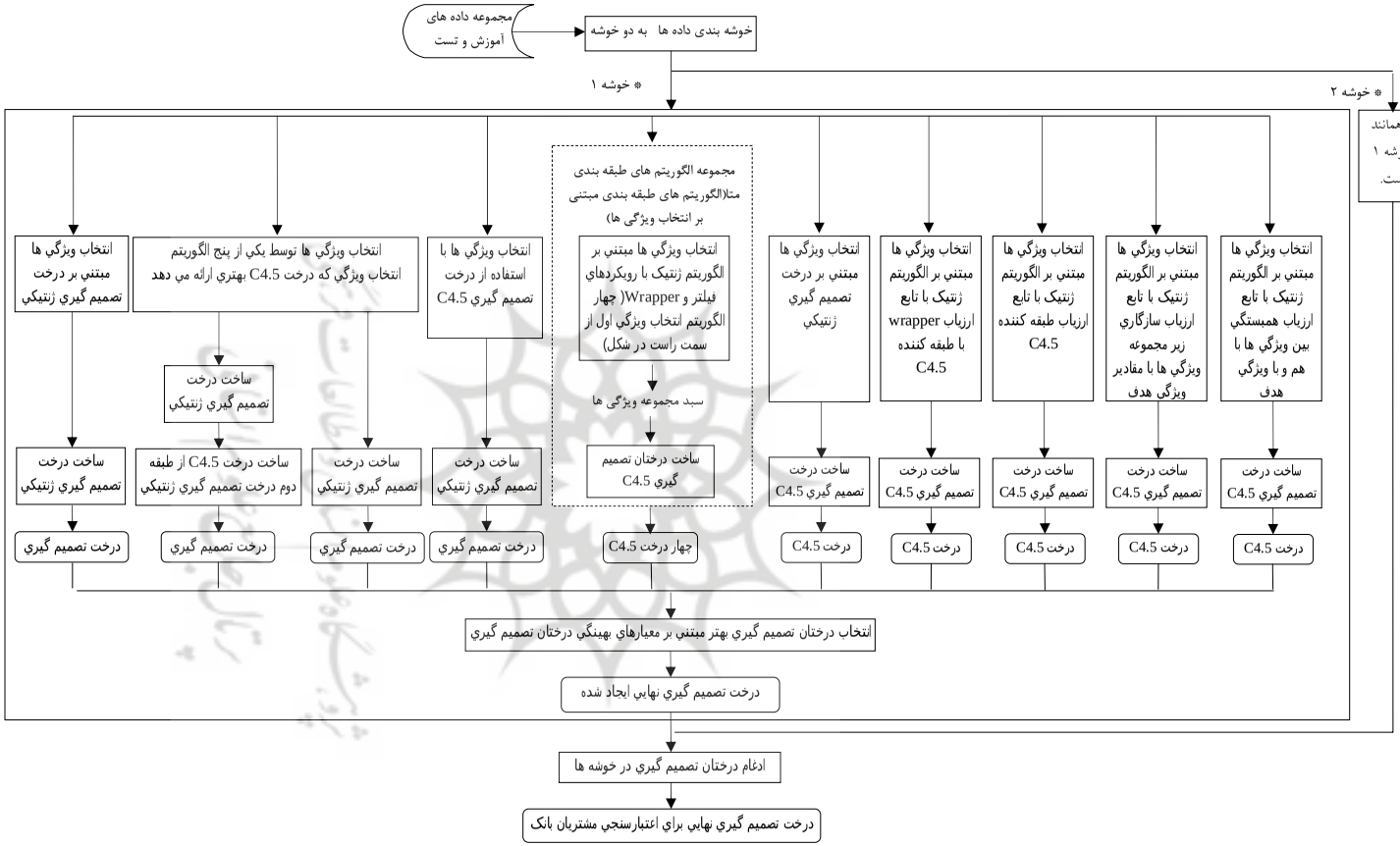
طبقه‌ی مشتریان در ویژگی هدف: شامل سه طبقه مشتریان جاری، سررسید گذشته و مطالبات معوقه است. مشتریان جاری، مشتریانی هستند که دچار رفتار عدم بازپرداخت تسهیلات دریافتی نشده اند. مشتریان سررسید گذشته مشتریانی هستند از مدت بازپرداخت تسهیلات اعطایی آن‌ها سه ماه گذشته است. مشتریان مطالبات معوقه مشتریانی هستند که از مدت بازپرداخت آن‌ها بیش از شش ماه گذشته باشد[۱].

۵. مدل پیشنهادی

نمودار ۲ مدل پیشنهادی را نشان می‌دهد که به فرآیند ساخت و آزمون درخت تصمیم-گیری در اعتبارسنجی مشتریان بانک اشاره دارد. پس از این‌که سه مرحله‌ی اول از فرآیند توسعه در شناخت الگو انجام شد، پیش پردازش داده‌ها صورت گرفت. می‌توان روش خوشه‌بندی را برای پیش پردازش داده‌ها به‌کار برد[۱۸]. با توجه به نمودار ۲ فلوچارت مراحل کلی کار در ساخت مدل طبقه‌بندی پیشنهادی به صورت نمودار ۱ است.



نمودار ۱. فلوچارت مراحل کلی کار در ساخت مدل طبقه‌بندی پیشنهادی



نمودار ۲. فرایند ساخت و آزمون مدل پیشنهادی در اعتبارسنجی مشتریان بانک

در انتخاب ویژگی‌ها، روش جستجو، تصادفی و بر مبنای الگوریتم ژنتیک است. الگوریتم‌های انتخاب ویژگی مبتنی بر رویکرد فیلتر به ارزیابی موارد زیر در مجموعه‌ی ویژگی‌ها می‌پردازند: همبستگی بین ویژگی‌ها با هم و با ویژگی هدف، سازگاری زیر مجموعه‌ی ویژگی‌ها با مقادیر ویژگی هدف، دقت طبقه کننده C4.5. الگوریتم انتخاب ویژگی با تابع ارزیاب Wrapper با طبقه‌کننده C4.5، مبتنی بر رویکرد Wrapper است. انتخاب ویژگی مبتنی بر درخت تصمیم‌گیری ژنتیکی: به علت این‌که این الگوریتم از یک الگوریتم طبقه‌بندی هم‌چون درخت تصمیم‌گیری ژنتیکی برای انتخاب ویژگی‌های مناسب استفاده می‌کند، مبتنی بر رویکرد طرح جاسازی شده در انتخاب ویژگی‌ها است. الگوریتم درخت تصمیم‌گیری ژنتیکی برگرفته از [۱۹] است.

در انتخاب ویژگی مبتنی بر رویکردهای فیلتر و Wrapper توسط الگوریتم ژنتیک از شیوه‌ی کدگذاری صفر و یک برای کدگذاری کروموزوم‌ها(مجموعه‌ی ویژگی‌ها) استفاده می‌شود. عدد یک و صفر به ترتیب نشان دهنده‌ی حضور و عدم حضور یک ویژگی در مجموعه‌ی ویژگی‌ها است. یک مجموعه‌ای از کروموزوم‌ها به‌صورت تصادفی تولید می‌شوند. بعد از این مرحله نوبت به ارزیابی تک تک کروموزوم‌ها توسط توابع ارزیاب می‌رسد. کروموزوم‌های برتر مبتنی بر روش چرخ گردان برگرفته از گلدبرگ انتخاب می‌شوند و برای تولید مجدد، عملیات تقاطع و جهش به‌طور تصادفی بر روی آن‌ها صورت می‌گیرد. ابتدا یک عدد احتمالی تعیین می‌شود. سپس الگوریتم، یک عددی تصادفی را به هر دو کروموزوم تخصیص می‌دهد. در صورتی که این عدد از عدد احتمالی از قبل تعیین شده بیشتر باشد، عمل تقاطع تک نقطه‌ای صفر و یک، برگرفته از گلدبرگ صورت می‌گیرد. سپس عمل جهش تک نقطه ای صفر و یک بر روی کروموزوم‌های جدید اعمال می‌شود. در ادامه کروموزوم‌های جدید ایجاد شده، دوباره به‌وسیله‌ی توابع ارزیاب، ارزیابی‌شده و برترین آن‌ها جایگزین کروموزوم‌های ضعیف‌تر از نسل قبل می‌شوند. نوع عملگر جایگزینی بر پایه شایستگی است. شرط خاتمه‌ی این الگوریتم رسیدن به تعداد معینی تکرار الگوریتم یا ماکزیمم تعداد نسل‌ها است.

در روش‌های انتخاب ویژگی مقادیر پارامترها به صورت زیر است. عدد اعتبارسنجی متقاطع برابر ۱۰، نرخ تقاطع ۰.۹، نرخ جهش ۰.۰۱، تعداد نسل و جمعیت اولیه ۲۰ و عدد تصادفی seed برابر ۱ در نظر گرفته شد. تعداد دسته‌ها و عدد seed و حد آستانه در

الگوریتم انتخاب ویژگی با تابع ارزیاب Wrapper با طبقه‌کننده C4.5 به ترتیب برابر ۱۰ و ۱ و ۰.۰۱ است. مقادیر پارامترهای الگوریتم انتخاب ویژگی مبتنی بر درخت تصمیم‌گیری ژنتیکی به صورت زیر است. استفاده از رویکرد اعتبارسنجی متقاطع با عدد ۱۰ در آموزش و تست درخت تصمیم‌گیری ژنتیکی، عملگر تقاطع: تصادفی استاندارد، عمل جهش: تصادفی استاندارد، درصد جایگزینی ژنوم برابر ۰.۲۵، نرخ خطا برابر ۰.۹۵، ترجیح قابلیت درخت تصمیم با دقت بالاتر بر درخت تصمیم کوچک‌تر، عدم تغییر پویا در ترجیح درختان تصمیم‌گیری با دقت بالاتر بر درخت تصمیم‌گیری کوچک‌تر در ابتدا و انتهای فرآیند تکامل، نرخ تقاطع عدد ۰.۹۹، نرخ جهش عدد ۰.۰۱، تعداد نسل‌ها عدد ۱۰۰، جمعیت اولیه عدد ۱۰۰، عدد تصادفی seed برابر ۱۲۳۴۵۶۷۸۹.

روش‌های ساخت انواع درختان تصمیم‌گیری در مدل طبقه‌بندی پیشنهادی به‌این صورت است. ۱. ساخت پنج درخت تصمیم‌گیری C4.5 توسط پنج روش انتخاب ویژگی. ۲. ساخت چهار درخت تصمیم‌گیری C4.5 توسط روش متا(ترکیب الگوریتم‌های انتخاب ویژگی و درخت تصمیم‌گیری C4.5) با استفاده از چهار الگوریتم انتخاب ویژگی مبتنی بر رویکردهای فیلتر و Wrapper. ۳. استفاده از درخت تصمیم‌گیری C4.5 در انتخاب ویژگی‌ها و ساخت درخت تصمیم‌گیری ژنتیکی با ویژگی‌های منتخب(رویکرد مبتنی بر طرح جاسازی شده). ۴. درخت تصمیم‌گیری ژنتیکی. ۶. استفاده از درخت تصمیم‌گیری ژنتیکی در انتخاب ویژگی‌ها و ساخت درخت تصمیم‌گیری ژنتیکی با ویژگی‌های منتخب(رویکرد مبتنی بر طرح جاسازی شده). ۵. استفاده از سلسله مراتب درختان تصمیم-گیری: ساخت درخت تصمیم‌گیری ژنتیکی و سپس ساخت درخت تصمیم‌گیری C4.5 در دو شاخه‌ی حاصل از طبقه‌ی دوم بالای درخت تصمیم‌گیری ژنتیکی.

۶. مطالعه موردی مبتنی بر مدل پیشنهادی

مدل طبقه‌بندی پیشنهادی در بانک ملت مورد بررسی قرار گرفت. مجموعه‌ی داده‌های اعتبارسنجی مورد نیاز در این مدل به صورت یک نسخه‌ی الکترونیکی از مرکز تحقیقات و برنامه‌ریزی بانک ملت دریافت شد. به‌منظور تجزیه و تحلیل داده‌ها و اطلاعات اعتبارسنجی بانک ملت و رسیدن به نتایج لازم از ابزارهای آمار توصیفی، یادگیری ماشین و داده کاوی استفاده شد. از نرم‌افزارهای یادگیری ماشین WEKA و نرم‌افزار GATree و نرم‌افزار

Excel به‌منظور تحلیل اطلاعات و ثبت نتایج استفاده شد. برای ساخت و اعتبارسنجی مدل از مجموعه‌ی داده‌های اعتباری مشتریان حقیقی بانک ملت استفاده شده که در سه ماهه‌ی اول سال ۱۳۸۲ تسهیلات در قالب قرارداد دریافت کرده اند.

نتایج درخت تصمیم‌گیری مدل طبقه‌بندی پیشنهادی به صورت نگاره‌ی ۴ است.

نگاره‌ی۴. نتایج مدل طبقه‌بندی پیشنهادی اعتبارسنجی مشتریان در مجموعه‌ی داده‌های اعتباری بانک ملت

ردیف	الگوریتم طبقه‌بندی	کل مشاهدات	تعداد ویژگی‌های پیشگو منتخب	تعداد مشاهدات درست طبقه‌بندی شده	درصد مشاهدات درست طبقه‌بندی شده	تعداد برگ‌ها	اندازه‌ی درخت	دقت کلاس مشتریان ۱	دقت کلاس مشتریان ۲	دقت کلاس مشتریان ۳
ـ	مدل طبقه‌بندی پیشنهادی	۵۱۷۳	۱۷	۴۹۹۲	٪۹۶.۵۰۱	۲۱۳	۲۹۰	۰.۹۸۳۹	۰.۹۵۹۴	۰.۸۷۹۴

در ادامه به مقایسه‌ی درخت تصمیم‌گیری مدل طبقه‌بندی پیشنهادی با سایر درختان تصمیم‌گیری C4.5 در مجموعه‌ی داده‌های اعتباری بانک ملت پرداخته می‌شود.

نگاره‌ی ۵. نتایج اجرای C4.5 بدون انتخاب ویژگی‌ها و خوشه‌بندی مجموعه‌ی داده‌های اعتباری بانک ملت

ردیف	کل مشاهده‌ها	تعداد مشاهده‌های درست طبقه‌بندی شده	درصد مشاهده‌های درست طبقه‌بندی شده	تعداد برگ‌ها	اندازه‌ی درخت	دقت کلاس مشتریان ۱	دقت کلاس مشتریان ۲	دقت کلاس مشتریان ۳
۱	۵۱۷۳	۴۹۶۴	٪۹۵.۹۶	۳۱۶	۳۹۸	۰.۹۷۹	۰.۹۴۵	۰.۸۸۷

نگاره‌ی ۶. نتایج اجرای C4.5 با انتخاب ویژگی‌ها مبتنی بر الگوریتم ژنتیک و بدون خوشه‌بندی مجموعه‌ی داده‌های اعتباری بانک ملت

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد مشاهدات درست طبقه‌بندی شده	درصد مشاهدات درست طبقه‌بندی شده	تعداد برگ‌ها	اندازه‌ی درخت	دقت کلاس مشتریان ۱	دقت کلاس مشتریان ۲	دقت کلاس مشتریان ۳
۱	Wrapper با طبقه کننده C4.5	۵۱۷۳	۴۹۹۱	٪۹۶.۴۸۱۷	۲۹۷	۳۸۰	۰.۹۸۲	۰.۹۵	۰.۹
۲	همبستگی بین ویژگی‌ها با هم و با ویژگی هدف	۵۱۷۳	۴۹۳۸	٪۹۵.۴۵۷۲	۲۲۶	۳۲۳	۰.۹۷۴	۰.۹۴	۰.۸۸۱
۳	سازگاری زیرمجموعه ویژگی‌ها با مقادیر ویژگی هدف	۵۱۷۳	۴۹۲۶	٪۹۵.۲۲۵۲	۲۳۸	۳۲۶	۰.۹۷۶	۰.۹۲۳	۰.۸۸۵
۴	طبقه‌کننده C4.5	۵۱۷۳	۴۹۵۱	٪۹۵.۷۰۸۵	۲۷۷	۳۷۲	۰.۹۸	۰.۹۳۲	۰.۸۸۸

نگاره‌ی ۷. نتایج اجرای C4.5 با انتخاب ویژگی‌ها مبتنی بر جستجوی اول بهترین و بدون خوشه‌بندی مجموعه‌ی داده‌های اعتباری بانک ملت

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد مشاهدات درست طبقه‌بندی شده	درصد مشاهدات درست طبقه‌بندی شده	تعداد برگ‌ها	اندازه‌ی درخت	دقت کلاس مشتریان ۱	دقت کلاس مشتریان ۲	دقت کلاس مشتریان ۳
۱	Wrapper با طبقه کننده C4.5	۵۱۷۳	۴۹۹۸	٪۹۶.۶۱۷۱	۳۱۸	۳۸۹	۰.۹۸۴	۰.۹۶۵	۰.۸۷۹
۲	همبستگی بین ویژگی‌ها با هم و با ویژگی هدف	۵۱۷۳	۴۹۴۳	٪۹۵.۵۵۳۸	۲۱۵	۳۱۵	۰.۹۷۴	۰.۹۴۲	۰.۸۸۴
۳	سازگاری زیرمجموعه ویژگی‌ها با مقادیر ویژگی هدف	۵۱۷۳	۴۹۸۰	٪۹۶.۲۶۹۱	۲۷۶	۳۵۲	۰.۹۷۹	۰.۹۵۸	۰.۸۸۶
۴	طبقه‌کننده C4.5	۵۱۷۳	۴۹۶۸	٪۹۶.۰۳۷۱	۳۱۴	۴۰۴	۰.۹۸	۰.۹۳۸	۰.۹۰۱

نگاره‌ی ۸. نتایج اجرای C4.5 با انتخاب ویژگی‌ها مبتنی بر الگوریتم ژنتیک با در نظر گرفتن ویژگی نوع خوشه در مجموعه‌ی داده‌های بانک ملت

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد مشاهدات درست طبقه‌بندی شده	درصد مشاهدات درست طبقه‌بندی شده	تعداد برگ‌ها	اندازه‌ی درخت	دقت کلاس مشتریان ۱	دقت کلاس مشتریان ۲	دقت کلاس مشتریان ۳
۱	Wrapper با طبقه کننده C4.5	۵۱۷۳	۴۹۸۶	٪۹۶.۳۸۵۱	۳۴۶	۴۳۷	۰.۹۸۳	۰.۹۵۳	۰.۸۸۸
۲	همبستگی بین ویژگی‌ها با هم و با ویژگی هدف	۵۱۷۳	۴۹۴۳	٪۹۵.۵۵۳۸	۲۱۵	۳۱۵	۰.۹۷۴	۰.۹۴۲	۰.۸۸۴
۳	سازگاری زیرمجموعه ویژگی‌ها با مقادیر ویژگی هدف	۵۱۷۳	۴۹۶۲	٪۹۵.۹۲۱۱	۲۴۰	۳۱۰	۰.۹۷۴	۰.۹۵۷	۰.۸۸۸
۴	طبقه کننده C4.5	۵۱۷۳	۴۹۷۰	٪۹۶.۰۷۵۸	۳۰۱	۴۰۰	۰.۹۷۹	۰.۹۴۹	۰.۸۸۹

درخت تصمیم‌گیری مدل طبقه‌بندی پیشنهادی دارای دقت طبقه‌بندی بالا و پیچیدگی کمتر نسبت به درختان تصمیم‌گیری مقایسه شده بود. تنها درخت تصمیم‌گیری C4.5 حاصل از انتخاب ویژگی‌ها با جستجوی اول بهترین و تابع ارزیاب Wrapper با طبقه‌کننده C4.5 دارای دقت طبقه‌بندی بالاتر فقط به میزان ۰.۱٪ بود. اما تعداد برگ‌ها و اندازه‌ی درخت این درخت تصمیم‌گیری خیلی بیشتر از تعداد برگ‌ها و اندازه‌ی درخت تصمیم-گیری مدل طبقه‌بندی پیشنهادی بود.

مبتنی بر مدل طبقه‌بندی پیشنهادی، از الگوریتم ژنتیک در ساخت درختان تصمیم‌گیری مانند C4.5 استفاده شده است. هم‌چنین دقت طبقه‌بندی و پیچیدگی درخت تصمیم‌گیری مدل طبقه‌بندی پیشنهادی برای اعتبارسنجی مشتریان بانک (بانک ملت) نسبت به درختان تصمیم‌گیری مقایسه شده، بهتر شده است.

۷. نتیجه‌گیری و پیشنهادها

بانک‌ها در اعطای تسهیلات اعتباری به مشتریان خود نیازمند اعتبارسنجی آن‌ها هستند. درختان تصمیم‌گیری می‌توانند در این زمینه به طبقه‌بندی مشتریان بپردازند. مسئله‌ی اصلی

ساخت درختان تصمیم‌گیری است که بتوانند به طور بهینه مشتریان را طبقه‌بندی کنند. هدف، ارایه‌ی یک مدل مناسب اعتبارسنجی مشتریان بانک‌ها مانند بانک ملت برای اعطای تسهیلات اعتباری متناسب با هر طبقه بود. این مدل در قالب فرآیند توسعه در شناخت الگو و فرآیند CRISP به ساخت درخت تصمیم‌گیری نهایی برای اعتبارسنجی مشتریان بانک پرداخت. تکنیک‌های خوشه‌بندی و انتخاب ویژگی‌ها مبتنی بر الگوریتم ژنتیک در ساخت درختان تصمیم‌گیری به‌کار رفتند. درخت تصمیم‌گیری حاصل از مدل طبقه‌بندی پیشنهادی دارای دقت طبقه‌بندی بالاتر، تعداد برگ‌ها و اندازه‌ی درخت تصمیم‌گیری و در نتیجه پیچیدگی کمتری نسبت به همه‌ی درختان تصمیم‌گیری مقایسه شده در این مقاله بود. با توجه به موارد گفته شده می‌توان از مدل طبقه‌بندی پیشنهادی برای ساخت و آزمون درختان تصمیم‌گیری به‌منظور اعتبارسنجی مشتریان بانک استفاده نمود. با توجه به پیشینه ی پژوهشی و مدل پیشنهادی این موارد پیشنهاد می‌شود: ۱. لحاظ کردن هزینه‌ی طبقه‌بندی غلط در الگوریتم‌های درخت تصمیم‌گیری و هزینه‌ی انتخاب ویژگی‌های غلط در الگوریتم انتخاب ویژگی در مدل پیشنهادی. ۲. توسعه‌ی مدل پیشنهادی با به‌کارگیری سایر روش‌های طبقه‌بندی درخت تصمیم‌گیری هم‌چون ID3 در ساخت مدل طبقه‌بندی پیشنهادی. هم‌چنین پیشنهادهای کاربردی برای بانک‌ها به‌این ترتیب است: ۱. استفاده از مدل طبقه‌بندی پیشنهادی در اعتبارسنجی مشتریان بانکی برای تخصیص بهینه‌ی تسهیلات اعتباری. ۲. به‌کارگیری فرآیند توسعه در شناخت الگو برای ساخت مدل‌های طبقه‌بندی برای اعتبارسنجی بهتر مشتریان بانک‌ها. ۳. طراحی و ساخت سیستم پشتیبانی تصمیم و نرم‌افزار کاربردی برای اعتبارسنجی مشتریان بانک مبتنی بر مدل پیشنهادی.

منابع

۱. شریفی ک. (۱۳۸۸/۲/۶). شیوه‌یاعتبارسنجی در بانک ملت. (م. خان بابایی، مصاحبه کننده).

۲. نادعلی ا.، خان بابایی م. به‌کارگیری تکنیک‌های درخت تصمیم و الگوریتم ژنتیک جهت اعتبارسنجی مشتریان بانک‌ها در یک سیستم پشتیبانی تصمیم‌گیری. دومین کنفرانس ملی داده‌کاوی. تهران: دانشگاه صنعتی امیرکبیر؛ ۱۳۸۷.

3. Abdou H. Pointon J. Neural Nets Versus Conventional Techniques in Credit Scoringin Egyptian Banking. Expert Systems with Applications; 2008. Available at: www.sciencedirect.com.

4. Carvalho D. R., Freitas A. A. A Hybrid Decision Tree/Genetic Algorithm Method for Data Mining. Information Sciences 2004; 163.

5. D'heygere, T., Goethals, P. L., Pauw N. D. Genetic Algorithms for Optimisation of Predictive ecosystems Models Based on Decision Trees and Neural Networks. Ecological Modelling 2006; 195

6. D'heygere T., Goethals P. L., Pauw N. D. Use of Genetic Algorithms to Select Input Variables in Decision Tree Models for the Prediction of Benthic Macroinvertebrates. Ecological Modelling 2003; 160. Available at: www.elsevier.com.

7. Dehuri S., Patnaik S., Ghosh A., Mall R., Application of Elitist Multi-objective Genetic Algorithm for Classification Rule Generation. Applied Soft Computing 2008; 8.

8. Gray J. B., Fan G. Classification Tree Analysis Using TARGET. Computational Statistics & Data Analysis 2008; 52.

9. Hsu P. L., Lai, R., Chiu, C. C., & Hsu, C. I. The hybrid of Association Rule Algorithms and Genetic algorithms for Tree Induction: An Example of Predicting the Student Course Performance. Expert Systems with Applications 2003; 25

10. Huang M., Gong J., Shi Z., Liu C., Zhang L. Genetic Algorithm-based Decision Tree Classifier for Remote Sensing Mapping with SPOT-5 Data in the Hong Shi Mao Watershed of the Loess Plateau, China. Neural Comput & Applic 2007.

11. Kim E., Kim W., Lee Y. Combination of Multiple Classifiers for the Customer's Purchase Behavior Prediction. Decision Support Systems 2002; 34.

12. Kim Y. S., Sohn S. Y. Managing Loan Customers Using Misclassification Patterns of Credit Scoring Model. Expert Systems with Applications 2004; 26.

13. Lee T. S., Chen I. F. A Two-stage Hybrid Credit Scoring Model Using Artificial Neural Networks and Multivariate Adaptive Regression Splines. Expert Systems with Applications 2005; 28.

14. Lee T. S., Chiu C. C., Lu C. J., Chen I. F. Credit Scoring Using the Hybrid Neural Discriminant Technique. Expert Systems with Applications 2002; 23.

15. Liu H. H., Ong C. S. Variable Selection in Clustering for Marketing Segmentation Using Genetic Algorithms. Expert Systems with Applications 2008; 34.

16. Martinez-Otzeta J. M., Sierra B., Lazkano E., Astigarraga A. Classifier Hierarchy Learning by Means of Genetic Algorithms. Pattern Recognition Letters 2006; 27.

17. Nanni L., Lumini A. An Experimental Comparison of Ensemble of Classifiers for Bankruptcy Prediction and Credit Scoring. Expert Systems with Applications 2009; 36.

18. Olson D., Shi Y. Introduction to Business Data Mining. Singapore: McGraw Hill Education; 2007.

19. Papagelis A., Kalles D. (n.d.) Breeding Decision Trees Using Evolutionary Techniques.

20. Sabzevari H., Soleymani M., Noorbakhsh E. (n.d.) A Comparison between Statistical and Data Mining Methods for Credit Scoring in Case of Limited Available Data.

21. Sorensen K., Janssens G. K.. Data Mining with Genetic Algorithms on Binary Trees. European Journal of Operational Research 2003; 151.

22. Susac M. Z., Sarlija N., Bensic M., (n.d.) Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models. s.n.

23. Tan F., Fu X., Zhang Y., Bourgeois A. G., A Genetic Algorithm-based Method for Feature Subset Selection. Soft Comput; 2008.

24. Thomas L. C. A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers. International Journal of Forecasting 2000; 16.

25. Weka Version 3.5.8 Software, 1999-2008.

26. Xu X., Zhou C., Wang Z. (in press). Credit Scoring Algorithm Based on Link Analysis Ranking with Support Vector Machine. Expert Systems with Applications; 2008. Available at: