



## بکارگیری تکنیک‌های خوشه‌بندی و الگوریتم ژنتیک در بهینه‌سازی درختان تصمیم‌گیری برای اعتبارسنجی مشتریان بانک‌ها

محمود البرزی

استادیار و عضو هیئت علمی دانشگاه آزاد اسلامی واحد علوم و تحقیقات  
mahmood\_alborzi@yahoo.com

محمد خان بابایی

عضو باشگاه پژوهشگران جوان، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات (مسئول مکاتبات)  
mohammadxanbababaei@yahoo.com

محمد ابراهیم محمدپور زرنندی

دانشیار دانشگاه آزاد اسلامی واحد تهران مرکزی  
pourzarandi@yahoo.com

تاریخ پذیرش: ۹۱/۸/۲۱

تاریخ دریافت: ۹۱/۵/۱۸

### چکیده

درختان تصمیم‌گیری به عنوان یکی از تکنیک‌های داده‌کاوی کاربرد زیادی در اعتبارسنجی مشتریان بانک و شناسایی آن‌ها برای اعطای تسهیلات اعتباری دارد. مسئله اصلی در پیچیدگی درختان تصمیم‌گیری، اندازه بیش از حد، عدم انعطاف‌پذیری و دقت کم در طبقه‌بندی است. هدف از این مقاله ارائه مدل ترکیبی در بهینه‌سازی درختان تصمیم‌گیری توسط تکنیک الگوریتم ژنتیک به منظور حل مسائل ذکر شده در فوق برای اعتبارسنجی مشتریان بانک است. به نظر می‌رسد بتوان با انتخاب ویژگی‌های مناسب و ساخت درختان تصمیم‌گیری توسط الگوریتم ژنتیک به کاهش پیچیدگی و افزایش انعطاف‌پذیری درختان تصمیم‌گیری پرداخت. در مدل ترکیبی پیشنهادی ابتدا داده‌های اعتباری توسط تکنیک خوشه‌بندی SimpleKmeans به دو خوشه تقسیم می‌شوند. سپس با استفاده از الگوریتم ژنتیک، پنج الگوریتم انتخاب ویژگی مبتنی بر سه رویکرد فیلتر، Wrapper و طرح‌جاسازی شده بر پایه درخت تصمیم‌گیری ژنتیکی، به انتخاب ویژگی‌های اعتبارسنجی مهم در مجموعه داده می‌پردازند. در ادامه پنج درخت تصمیم‌گیری مبتنی بر الگوریتم C4.5 در هر خوشه با مجموعه ویژگی‌های منتخب ساخته می‌شود. بهترین درختان تصمیم‌گیری در هر خوشه مبتنی بر معیارهای بهینگی مورد نظر در این مقاله انتخاب شده و با هم ترکیب می‌شوند تا درخت تصمیم‌گیری نهایی برای اعتبارسنجی مشتریان بانک ایجاد شود. ابزار یادگیری ماشین وکا و نرم‌افزار GATree برای رسیدن به نتایج بکار گرفته شده است. نتایج پژوهش نشان می‌دهد که استفاده از مدل ترکیبی پیشنهادی در ساخت درخت تصمیم‌گیری منجر به افزایش دقت طبقه‌بندی نسبت به بسیاری از الگوریتم‌های مقایسه‌شده در این مقاله می‌شود؛ ولی پیچیدگی الگوریتم مدل ترکیبی پیشنهادی از برخی الگوریتم‌های طبقه‌بندی مقایسه‌شده در این مقاله بیشتر است.

**واژه‌های کلیدی:** اعتبارسنجی، طبقه‌بندی، الگوریتم ژنتیک، درختان تصمیم‌گیری، انتخاب ویژگی، خوشه‌بندی.

## ۱- مقدمه

امروزه بانک‌ها برای شناخت مشتریان، ارضای نیازمندی‌ها و ارائه خدمات مالی مناسب نیازمند شناسایی دقیق ویژگی‌های اعتباری آن‌ها هستند. یکی از خدمات مالی در بانک، اعطای تسهیلات مالی از جمله وام به متقاضیان اعتباری است. بانک‌ها همچون کسب و کارهای دیگر در طول حیات خود با ریسک‌هایی مواجه می‌شوند. یکی از مهم‌ترین آن‌ها ریسک اعتباری است که باید با آن مقابله کنند. اعتبارسنجی<sup>۱</sup> به عنوان یک تکنیک موثر به شناخت مشتریان خوب و بد پرداخته و با این کار می‌تواند ریسک اعتباری آن‌ها را تعیین کند. تحقیقات زیادی بر روی مدل‌های اعتبارسنجی در بانک‌ها صورت گرفته است. در ابتدا مدل‌های اعتبارسنجی به صورت قضاوتی بودند. سپس روش‌های پارامتریک در اعتبارسنجی مطرح شدند. اخیراً از روش‌های ناپارامتریک در اعتبارسنجی مشتریان بانک‌ها استفاده می‌شود. فیشر<sup>۲</sup> در سال ۱۹۶۳ برای اولین بار ایده متمایز کردن گروه‌ها را مطرح کرد. سپس دیوید دراند<sup>۳</sup> در سال ۱۹۴۱ با تفکیک مشتریان به دو گروه خوب و بد به اعطای وام به آن‌ها پرداخت. کارت‌های اعتباری در سال ۱۹۶۰ وارد بانک‌ها شدند. در سال ۱۹۸۰ برای اولین بار از اعتبارسنجی در بانک‌ها استفاده شد. همچنین اعتبارسنجی در بازاریابی مستقیم در سال ۱۹۹۰ بکار رفت. موفقیت‌های اخیر در بکارگیری اعتبارسنجی منجر به استفاده از این تکنیک در ارزیابی اعتبار مشتریان شد (Thomas, 2000). در ادامه در بخش ۲ به بیان مسئله و ضرورت و اهداف پژوهش، بخش ۳ مروری بر تحقیقات صورت گرفته، بخش ۴ مواد و روش‌ها، بخش ۵ آموزش و تست مدل، بخش ۶ به مقایسه درخت تصمیم‌گیری حاصل از مدل ترکیبی پیشنهادی با سایر درختان تصمیم‌گیری و در نهایت در بخش ۷ به نتیجه‌گیری پرداخته می‌شود.

## ۲- بیان مسئله، ضرورت و اهداف تحقیق

تکنیک‌های داده‌کاوی همچون طبقه‌بندی می‌توانند با ارائه یک الگو یا مدل به کشف دانش پنهان در حجم زیادی از داده‌های تراکنش‌های اعتباری مشتریان

بانک‌ها کمک کنند. درختان تصمیم‌گیری به عنوان یکی از تکنیک‌های طبقه‌بندی از روش‌های ناپارامتریک در اعتبارسنجی هستند. درختان تصمیم‌گیری می‌توانند با شناسایی ویژگی‌های مشتریان و تفکیک آن‌ها به گروه‌های خوب و بد، به اعتبارسنجی آن‌ها بپردازند. این ویژگی‌ها با توصیف مشخصات مشتریان بانک‌ها، طبقه آن‌ها را در اعتبارسنجی مشخص می‌کنند. از طرف دیگر برای شناخت الگوی مناسب در طبقه‌بندی مشتریان بانک در حجم زیادی از داده‌ها نیاز به پیش‌پردازش داده‌ها است. روش‌های مختلفی برای پیش‌پردازش داده‌ها وجود دارد که دو مورد از آن‌ها که در این مقاله بکار می‌روند، خوشه‌بندی و انتخاب ویژگی‌ها است. با پیش‌پردازش داده‌ها الگوی بهتری برای شناسایی و اعتبارسنجی مشتریان بانک‌ها ایجاد می‌شود.

مسئله اصلی در تحقیق موضوع این مقاله ساخت درختان تصمیم‌گیری است که بتوانند به طور بهینه به طبقه‌بندی مشتریان خوب و بد بانک‌ها برای اعتبارسنجی بپردازند. به نظر می‌رسد استفاده از الگوریتم‌های ژنتیک در انتخاب ویژگی‌ها و ساخت درختان تصمیم‌گیری بتواند منجر به طبقه‌بندی و اعتبارسنجی بهتری از مشتریان بانک‌ها شود. ممکن است الگوریتم‌های انتخاب ویژگی در بهینه‌سازی محلی گیر کنند و همچنین تعامل بین ویژگی‌ها را در نظر نگیرند. در برخی از الگوریتم‌های انتخاب ویژگی فرض بر این است که روابط بین ویژگی‌ها خطی بوده و مستقل از هم می‌باشند. الگوریتم‌های انتخاب ویژگی تنها از برخی معیارها برای انتخاب ویژگی‌ها استفاده می‌کنند.

از طرف دیگر با توجه به اینکه الگوریتم درخت تصمیم‌گیری C4.5 یک الگوریتم بازگشتی<sup>۴</sup> و حریم‌ناهن<sup>۵</sup> است، منجر به ایجاد درخت تصمیم‌گیری پیچیده و بزرگ می‌شود. به همین دلیل قسمت‌های پایینی درخت تعداد کمی تراکنش را پوشش می‌دهند. تعداد کم تراکنش در قسمت‌های پایینی درخت اگرچه منجر به افزایش دقت درخت در طبقه‌بندی مشتریان می‌شود ولی از طرف دیگر انعطاف‌پذیری را در طبقه‌بندی مشتریان کاهش داده و در قسمت‌های پایینی درخت تنها با تغییر کوچک و تقریباً نامحسوس کلاس یا طبقه

۲- درصد مشاهدات درست طبقه بندی شده<sup>۹</sup> که حاصل تقسیم تعداد مشتریان درست طبقه بندی شده به تعداد کل مشتریان است.

۳- تعداد برگ ها در درخت تصمیم گیری. برگ در یک درخت تصمیم گیری گره انتهایی آن است که در این گره به تعیین خوب یا بد بودن مشتری اعتبارسنجی پرداخته می شود. افزایش تعداد برگ های درخت تصمیم گیری باعث افزایش پهنای درخت تصمیم گیری شده و پیچیدگی مدل اعتبارسنجی را افزایش می دهد.

۴- اندازه درخت تصمیم گیری که به تعداد شاخه ها و قوانین اگر آنگاه در درخت تصمیم گیری ارتباط دارد. اندازه درخت تصمیم گیری برابر با مجموع تعداد برگ ها و گره ها در آن است. با افزایش اندازه درخت تصمیم گیری پیچیدگی مدل اعتبارسنجی نیز افزایش می یابد.

با توجه به بیان مسئله و ضرورت تحقیق موضوع این مقاله، برخی از مزایایی که الگوریتم های ژنتیک می تواند در انتخاب ویژگی ها داشته باشند، شامل موارد زیر است (Carvalho & Freitas, 2004):

۱) الگوریتم های ژنتیک بر خلاف الگوریتم های حریصانه در یک لحظه با مجموعه ای از راه حل هل کار می کنند.

۲) الگوریتم های استنتاج حریصانه به بررسی تنها راه حل های جزئی در هر مرحله می پردازند.

۳) الگوریتم های ژنتیک با توجه به استفاده از قانون احتمال در بهینه محلی کمتر گیر می کنند. الگوریتم های ژنتیک به ایجاد یک رابطه مناسب بین اندازه و پیچیدگی درختان تصمیم گیری می پردازند (Aitkenhead, 2008). همچنین مزیت های بکارگیری الگوریتم ژنتیک در ساخت درختان تصمیم گیری می تواند شامل موارد زیر باشد:

۱) روش های تکاملی برای ایجاد درخت باعث می شوند که تغییرات به طور اتوماتیک در درخت ایجاد شوند. می توان توسط ترکیب روش های تکاملی و درخت تصمیم بین اندازه درخت و پیچیدگی آن رابطه مناسبی ایجاد نمود (Aitkenhead, 2008)

۲) انعطاف پذیری در نمایش تابع طبقه بندی.

یک مشتری جدید تغییر می کند. به همین دلیل به نظر می رسد اگر بتوان تا جایی که لطمه ای به دقت طبقه بندی در درخت تصمیم گیری نخورد، اندازه درخت، تعداد برگ ها و در نتیجه پیچیدگی درخت تصمیم گیری را کاهش داد، علاوه بر افزایش انعطاف پذیری در طبقه بندی، کاهش اندازه درخت و پیچیدگی آن می توان طبقه بندی بهتری را برای مشتریان بانک انجام داد.

در این مقاله فرض بر این است که بکارگیری الگوریتم ژنتیک در ساخت درختان تصمیم گیری مبتنی بر الگوریتم C4.5 منجر به بهینه سازی آن ها در دقت و پیچیدگی طبقه بندی مشتریان اعتبارسنجی بانک ها می شود. C4.5 یکی از الگوریتم های ساخت مدل طبقه بندی درخت تصمیم گیری است. این الگوریتم یک درخت تصمیم گیری به صورت نمودار گرافیکی ایجاد می کند که فهم آن برای کاربران آسان است. الگوریتم استنتاج C4.5 در سال ۱۹۹۳ توسط کوئینلن<sup>۶</sup> تهیه شد که از مفهوم شاخص کسب اطلاعات<sup>۷</sup> در ساخت درخت تصمیم گیری استفاده می کند (Larose, 2005). هدف از تحقیق موضوع این مقاله ارائه یک مدل مناسب برای اعتبارسنجی مشتریان بانک ها است. سوال تحقیق به صورت زیر است: چگونه می توان توسط الگوریتم ژنتیک به انتخاب ویژگی ها و ساخت درختان تصمیم گیری بهینه در اعتبارسنجی مشتریان بانک ها پرداخت؟

منظور از بهینگی درختان تصمیم گیری مدل پیشنهادی و سایر مدل های درخت تصمیم گیری مقایسه شده در این مقاله شامل موارد زیر است:

۱- تعداد ویژگی های پیشگویی کننده انتخابی<sup>۸</sup>؛ در یک درخت تصمیم گیری هر چه تعداد ویژگی های پیشگویی کننده کمتر باشد، پیچیدگی مدل کاهش یافته و انعطاف پذیری آن زیاد می شود. همچنین در بررسی وضعیت اعتباری متقاضیان اعتبار، زمان و هزینه کمتری صرف بررسی ویژگی های مشتریان می شود و از طرف دیگر سرعت پاسخگویی به مشتری به منظور رد یا تأیید تقاضا بیشتر می شود. از طرف دیگر نیاز به ذخیره سازی داده های کمتری در سیستم پردازش مدل طبقه بندی و اعتبارسنجی مشتریان در بانک ها است.

۳) می‌توان توابع برازندگی متنوعی در طبقه بندی بوجود آورد (Zhang & Bhattacharyya, 2004)

درختان تصمیم‌گیری در جدول ۳ اشاره می‌شود. این تحقیقات در حوزه‌های دیگر علوم و کسب و کارها می‌باشند و می‌توان از نتایج آن‌ها در اعتبارسنجی مشتریان بانک‌ها استفاده نمود.

### ۳- مروری بر تحقیقات صورت گرفته

در مورد کاربرد مدل‌های پارامتریک و ناپارامتریک اعتبارسنجی در طبقه بندی مشتریان بانک‌ها و موسسات مالی تحقیقات متنوعی در داخل و خارج از کشور صورت گرفته است. جدول ۱ به چند نمونه از این تحقیقات در خارج از کشور اشاره می‌کند.

بکارگیری الگوریتم ژنتیک در انتخاب ویژگی‌ها در سال ۱۹۸۹ و اولین بار به وسیله Siedlecki و Sklansky در مقاله‌ای تحت عنوان A note on genetic algorithms for large-scale feature selection (Tsang, Kwong, & Wang, 2007) مطرح شد.

در ادامه به تحقیقاتی راجع به بکارگیری الگوریتم‌های ژنتیک در انتخاب ویژگی‌ها در جدول ۲ و ساخت

### جدول ۱- برخی تحقیقات روش‌های اعتبارسنجی در خارج از کشور

ردیف	مدل اعتبارسنجی	پژوهشگر
۱	رگرسیون لجستیک، شبکه عصبی، درخت تصمیم‌گیری	(Susac, Sarlija, & Bensic, n.d.)
۲	ترکیب تحلیل تمایزی و الگوریتم پس انتشار در شبکه عصبی	(Lee, Chiu, Lu, & Chen, 2002)
۳	الگوهای طبقه بندی غلط	(Kim & Sohn, 2004)
۴	ترکیب مدل‌های شبکه‌های عصبی مصنوعی و روش MARS	(Lee & Chen, 2005)
۵	رتبه بندی تحلیل لینک با استفاده از ماشین بردار پشتیبان	(Xu, Zhou, & Wang, 2008)
۶	شبکه‌های عصبی و تکنیک‌های عمومی	(Abdou & Pointon, 2008)
۷	طبقه‌کننده‌های ترکیبی به جای یک طبقه‌کننده	(Nanni & Lumini, 2009)

### جدول ۲- برخی تحقیقات در کاربرد الگوریتم ژنتیک در انتخاب ویژگی‌ها

ردیف	رویکرد بکارگیری الگوریتم ژنتیک در انتخاب ویژگی‌ها	پژوهشگر
۱	الگوریتم REDUCE	(Lavrac, Gamberger, & Turney, 1995)
۲	استفاده از عملگر خودتقاطع برای انتخاب ویژگی‌ها	(PAL, NAND, & Kundu, 1998)
۳	ترکیب طبقه بندی چندگانه مبتنی بر الگوریتم ژنتیک	(Lee, 2002 cited in & Kim, Kim, نادعلی & خان بابایی, ۱۳۸۷)
۴	استفاده از الگوریتم ژنتیک برای انتخاب متغیرهای ورودی	(D'heygere, Goethals, & Pauw, 2003)
۵	الگوریتم ژنتیک در انتخاب متغیرها به کمک خوشه بندی مشتریان	(Liu & Ong, 2008)
۶	استفاده از الگوریتم ژنتیک در ترکیب روش‌های انتخاب ویژگی	(Tan, Fu, Zhang, & Bourgeois, 2008)

### جدول ۳- برخی تحقیقات در کاربرد الگوریتم ژنتیک در ساخت درختان تصمیم‌گیری

ردیف	رویکرد بکارگیری الگوریتم ژنتیک در ساخت درختان تصمیم‌گیری	پژوهشگر
۱	ساخت درختان تصمیم‌گیری دودویی توسط تکنیک‌های تکاملی الگوریتم ژنتیک	(Papagelis & Kalles, n.d.)
۲	ترکیب الگوریتم‌های ژنتیک و درختان تصمیم ID3 در طبقه بندی الگو	(Bala, Huang, Vafaie, DeJong, & Wechsler, 1995)
۳	استفاده از الگوریتم ژنتیک برای کشف قوانین در شقوق کوچک درختان تصمیم‌گیری	(Freitas & Carvalho, ۲۰۰۲, ۲۰۰۴)
۴	یکپارچه کردن الگوریتم قوانین وابستگی و الگوریتم ژنتیک در کشف درخت طبقه بندی	Hsu, & Lai, Chiu, Hsu, نادعلی & خان بابایی (۲۰۰۳ cited in)
۵	کشف قوانین تصمیم‌گیری در ورشکستگی به کمک تصمیمات کیفی خیرگان	(Kim & Han, 2003)
۶	استفاده از الگوریتم‌های ژنتیک برای یادگیری سلسله مراتب طبقه‌کننده‌ها	(Martinez-Otzeta, Sierra, Lazkano, & Astigarraga, 2006)
۷	بهینه‌سازی پیشگویی مدل‌ها بر مبنای درختان تصمیم و شبکه‌های عصبی	(D'heygere, Goethals, & Pauw, 2006)

ردیف	رویکرد بکارگیری الگوریتم ژنتیک در ساخت درختان تصمیم گیری	پژوهشگر
۸	بررسی سودمندی تکنیک درخت تصمیم گیری بر مبنای الگوریتم ژنتیک	(Huang, Gong, Shi, Liu, & Zhang, 2007)
۹	تعریف برانزنگی درخت توسط الگوریتم ژنتیک	Janssens, & Sorensen (۱۳۸۷، نادعلی & خان بابایی cited in 2003)
۱۰	تحلیل درخت طبقه بندی توسط الگوریتم TARGET	(Gray & Fan, 2008)
۱۱	الگوریتم ژنتیک چند هدفه Elitist به کشف قوانین طبقه بندی مجموعه داده های بزرگ	(Dehuri, Patnaik, Ghosh, & Mall, 2008)

#### ۴- مواد و روش ها

##### ۴-۱- مجموعه داده

داده هایی که در این مقاله برای ساخت و آزمون درختان تصمیم گیری C4.5 مورد استفاده قرار می گیرند، مجموعه داده های اعتباری آلمان<sup>۱۰</sup> است (<http://archive.ics.uci.edu/ml/datasets.html>) که در مقالات مشابه برای بررسی اثربخشی و امکان سنجی مدل مورد استفاده قرار گرفته است. این مجموعه داده که در سال ۱۹۹۲ تهیه شد، فاقد مقادیر مفقود و اختلال است. روی این مجموعه عملیات آماده سازی و تمیز کردن و پیش پردازش داده ها صورت می گیرد. 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 آلمان به همراه مقادیر و نوع آن ها در پیوست ۱ موجود است.

با انجام فرایند آماده سازی و تمیز کردن داده ها در چند مرحله و با اعمال روش های مختلف توسط نرم افزار یادگیری ماشین و کاکا<sup>۱۱</sup> نسخه ۳.۵.۸، تعداد تراکنش ها از ۱۰۰۰ به ۶۹۰ کاهش یافت. همچنین همه ویژگی ها از نوع عددی به نوع اسمی تبدیل شدند. توسط نمودارهای قابلیت تجسم سازی<sup>۱۲</sup> موجود در این نرم افزار، وضعیت کلی داده ها در هر ویژگی، مشاهده و ارتباط بین مقادیر هر یک از ویژگی های پیشگویی کننده با مقادیر ویژگی هدف بررسی شد. روش های زیر در آماده سازی داده ها برای آموزش و تست مدل ترکیبی پیشنهادی برای طبقه بندی مشتریان اعتباری از اولویت بندی خاصی پیروی نمی کند. برخی روش ها، چندین بار بر روی مجموعه داده در مراحل مختلف آماده سازی اعمال شده اند. روش های آماده سازی داده ها در این مقاله بدین صورت است:

۱. حذف مقادیر پراکنده. ۲. نرمال سازی که فقط بر روی ویژگی "سن" اعمال شد و همچنین مقادیر ویژگی هدف در محاسبات این روش در نظر گرفته شد. ۳. گسسته سازی<sup>۱۳</sup> مقادیر ویژگی های عددی که در این روش مقادیر ویژگی هدف در محاسبات لحاظ شد. ۴. ادغام مقادیر<sup>۱۴</sup> داده در ویژگی های اسمی. ۵. تبدیل<sup>۱۵</sup> ویژگی های عددی به اسمی.

##### ۴-۲- ریسک اعتباری و اعتبارسنجی

بر اساس دیدگاه کمیته بال<sup>۱۶</sup> مهم ترین ریسک هایی که بانک ها با آن مواجه هستند شامل موارد زیر است: ریسک اعتباری، ریسک کشوری و ریسک انتقال وجوه، ریسک بازار، ریسک نرخ بهره، ریسک نقدینگی، ریسک عملیاتی، ریسک حقوقی و ریسک شهرت (اداره مطالعات و کنترل ریسک بانک تجارت، ۱۳۸۶). بانک ها برای مقابله با ریسک های پیش رو باید بتوانند به مدیریت ریسک بپردازند. یکی از ریسک های بانک ها ریسک اعتباری است. ریسک اعتباری به معنی احتمال عدم بازپرداخت اصل و سود تسهیلات اعطایی توسط گیرنده اعتبار به علت عدم تمایل و یا ناتوانی مالی است (اداره مطالعات و کنترل ریسک بانک تجارت، ۱۳۸۶). برای هر موسسه ارائه دهنده خدمات مالی مثل بانک های تجاری، توانایی تفکیک مشتریان خوب و مشتریان بد، امری حیاتی و مهم می باشد. در نتیجه نیاز به مدل های معتبری است که توسط آن ها بتوان به پیش بینی قصور در بازپرداخت وام توسط مشتریان پرداخت تا ذینفعان بتوانند در زمان مناسب اقدامات پیشگیرانه و اصلاحی صحیحی انجام دهند (Yu, Wang, & Lai, 2007).

یکی از تکنیک های مهم در ارزیابی ریسک اعتباری، اعتبارسنجی است. توماس<sup>۱۷</sup> در سال ۲۰۰۲ اعتبارسنجی را تکنیکی تعریف کرد که به بانک ها و شرکت های

اعتباری در زمینه اعطای اعتبار به مشتریان بر مبنای معیارهای از قبل تعیین شده، کمک می‌کند (Yu, Wang, & Lai, 2007). اعتبارسنجی دارای مزیت‌هایی مثل موارد زیر می‌باشد: ۱. کاهش هزینه تحلیل اعتبار. ۲. تصمیم‌گیری سریع در اعتبارسنجی مشتریان. ۳. تضمین اعتبارات و حذف ریسک‌های احتمالی (Nanni & Lumini, 2009) (Ong, Huang, & Tzeng, 2005) بریل<sup>۱۸</sup> در سال ۱۹۹۸ علاوه بر دو مورد بالا به موارد زیر نیز اشاره کرد (نادعلی & خان بابایی, 1387). ۱. نظارت نزدیک به حساب‌های موجود. ۲. تعیین اولویت در مجموعه اعتبارات.

روش‌های اعتبارسنجی در ابتدا قضاوتی بود. کارشناس اعتبارسنجی با بررسی فرم تقاضانامه مبتنی بر تحلیل پنج ضابطه معروف به تحلیل (پنج C) به تصمیم‌گیری در مورد اعطای وام یا رد تقاضانامه می‌پرداخت. 5C برگرفته از حروف اول ۵ کلمه است. این ۵ کلمه عبارتند از: ویژگی<sup>۱۹</sup> شخص گیرنده وام، سرمایه<sup>۲۰</sup> وی، ضمانت<sup>۲۱</sup>، توانایی<sup>۲۲</sup> بازپرداخت و شرایط<sup>۲۳</sup> (Thomas, 2000). امروزه اکثر تحقیقات و کاربردها حول محور اعتبارسنجی توسط دو روش صورت می‌گیرند: ۱. روش‌های سنتی مثل رگرسیون لجستیک و مدل‌های لوجیت و پروبیت. ۲. روش‌های داده‌کاوی (Sabzevari, Soleymani, & Noorbakhsh, n.d.) در تقسیم‌بندی دیگری مدل‌های اعتبارسنجی به مدل‌های پارامتریک و ناپارامتریک تقسیم می‌شوند (Sabzevari, Soleymani, & Noorbakhsh, n.d.). مدل‌های پارامتریک مثل تحلیل تمایزی، رگرسیون خطی، پروبیت و لوجیت و مدل‌های ناپارامتریک مثل درختان طبقه‌بندی، شبکه‌های عصبی، سیستم‌های خبره و ...

بسیاری از تحقیقات اشاره شده در بخش ۳ این مقاله از روش‌های پارامتریک و ناپارامتریک برای اعتبارسنجی مشتریان استفاده کردند. نکته قابل توجه در این تحقیقات این است که هر یک از مدل‌های موجود در اعتبارسنجی به شیوه‌ای خاص به متمایز کردن مشتریان خوب و بد پرداختند. در این تحقیقات هر یک از مدل‌ها بر دیگری برتری داشتند و در سال‌های اخیر نیز محققین به این

قضیه توجه خاصی کرده‌اند. آن‌ها سعی دارند به شیوه‌ای بهتر به طبقه‌بندی مشتریان اعتبارسنجی بپردازند. دقت، پیچیدگی و انعطاف‌پذیری در طبقه‌بندی مورد توجه آن‌ها در مقایسه مدل‌های اعتبارسنجی است. در حال حاضر روش‌های جدیدی از هوش مصنوعی نیز در اعتبارسنجی بکار رفته‌اند که از آن‌ها می‌توان به موارد زیر اشاره کرد: شبکه‌های عصبی مصنوعی، محاسبات تکاملی، الگوریتم‌های ژنتیک و ماشین بردار پشتیبان. همچنین اخیراً از مدل‌های ترکیبی در اعتبارسنجی استفاده می‌شود؛ مثل مدل عصبی فازی، مدل فازی ماشین بردار پشتیبان و مدل ترکیبی شبکه‌های عصبی (Yu, Wang, & Lai, 2007). در این مقاله از ترکیب روش‌های خوشه‌بندی، الگوریتم انتخاب ویژگی مبتنی بر الگوریتم‌های ژنتیک، درختان تصمیم‌گیری مبتنی بر الگوریتم ژنتیک و درختان تصمیم‌گیری C4.5 برای ارائه مدل ترکیبی پیشنهادی اعتبارسنجی مشتریان بانک‌ها استفاده می‌شود. این روش‌ها از نوع روش‌های داده‌کاوی، ناپارامتریک، هوش مصنوعی و مدل‌های ترکیبی هستند.

#### ۴-۳- طبقه‌بندی و درختان تصمیم‌گیری

یکی از وظایف داده‌کاوی طبقه‌بندی است. طبقه‌بندی دارای تکنیک‌های متنوعی است که در تحقیقات مختلف از آن‌ها استفاده می‌شود. از تکنیک‌های رایج در طبقه‌بندی می‌توان به موارد زیر اشاره کرد: K نزدیک‌ترین همسایه، درختان تصمیم‌گیری، شبکه‌های عصبی، ماشین بردار پشتیبان، طبقه‌بندی بی‌زین، رگرسیون، تئوری‌های مجموعه دانه درشت، منطق یابی مبتنی بر حالت، سیستم‌های خبره، منطق فازی، الگوریتم‌های ژنتیک. درختان تصمیم‌گیری به علت سادگی و قابلیت فهم بالا از محبوبیت بالایی در کاربرد برخوردار هستند. این تکنیک در زمره درختان طبقه‌بندی قرار می‌گیرد. درختان طبقه‌بندی به پیشگویی مقادیر ویژگی‌ها یا متغیرهای وابسته و گسسته می‌پردازد. درختان تصمیم‌گیری تنها مقادیر ویژگی‌های گسسته را پیشگویی می‌کنند. این پیشگویی توسط متغیر کلاس که ویژگی هدف یا ویژگی وابسته نیز نامیده می‌شود، صورت می‌گیرد.

#### ۴-۴- 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. یک خوشه مجموعه ای از رکوردها است که به هم شبیه می باشند و از رکوردهای بیرون خوشه تفاوت دارند. در خوشه بندی متغیر هدف وجود ندارد و به طبقه بندی، تخمین و پیشگویی مقدار متغیر هدف نمی پردازد (Larose, 2005). در این مقاله از الگوریتم خوشه بندی SimpleKmeans استفاده می شود. معیار نزدیکی در پیدا کردن نزدیک ترین مرکز خوشه برای هر رکورد، معمولاً فاصله اقلیدسی است. معیار توقف می تواند به طور مثال مجموع مربعات خطا باشد. الگوریتم SimpleKmeans در (Olson & Shi, 2007, p.75) آمده 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<sup>۲۶</sup> دیگر به نزدیک ترین خوشه توسط محاسبه مقیاس فاصله مورد نظر. ۵. محاسبه مجدد متوسط های خوشه بر اساس تخصیص ها در مرحله ۴. ۶. تکرار بین مراحل ۴ و ۵. می توان از تکنیک خوشه بندی به عنوان پیش پردازش داده ها استفاده کرد (Olson & Shi, 2007) که در این مقاله این تکنیک بر روی مجموعه داده های اعتباری آلمان اعمال می شود.

#### ۴-۵- انتخاب ویژگی ها

در مدل ناپارامتریک طبقه بندی هزینه و زمان زیادی باید صرف کسب داده های مدل شود؛ زیرا مدل های ناپارامتریک مبتنی بر داده هستند. پس باید به جمع آوری ویژگی ها و داده هایی پرداخت که از اهمیت بیشتری برای ساخت مدل طبقه بندی برخوردارند. حذف اطلاعات غیر مرتبط و استخراج متغیرهای کلیدی در شناخت الگو، پیش پردازش نامیده می شود. (Kennedy, Lee, Roy, Reed, & Lippmann, 1998) برای ساخت یک مدل طبقه بندی مناسب نیاز به داده های آموزشی با کیفیت است. این کیفیت با تعداد داده ها و ویژگی ها در مجموعه آموزش ارتباط دارد. انتخاب ویژگی ها به عنوان

مقادیر ویژگی هدف، وابسته به مقادیر متغیرهای (ویژگی های) مستقل (توصیف کننده) و وجود آن ها در ساختار درخت تصمیم گیری است (D'hegyere, Goethals, & Pauw, 2003). درختان تصمیم گیری دارای الگوریتم های مختلفی هستند که برخی از آن ها شامل موارد زیر است: ID3, C4.5, C4, C5, CART, CHAID, QUEST. در این مقاله از الگوریتم درخت تصمیم گیری C4.5 برای ساخت درختان تصمیم گیری به منظور طبقه بندی و اعتبارسنجی مشتریان بانک ها استفاده می شود.

الگوریتم استنتاج C4.5 در سال ۱۹۹۳ توسط کوئینلن تهیه شد. این الگوریتم متغیرهای پیوسته و گسسته را در محاسبات خود لحاظ کرده و مقادیر مفقود را در الگوریتم خود در نظر می گیرد (Aitkenhead, 2008). این الگوریتم لزوماً دودویی نیست. برای انتخاب یک جداکننده بهینه در طول مسیر درخت تصمیم گیری از شاخص کسب اطلاعات یا کاهش آنتروپی<sup>۲۴</sup> استفاده می کند (Larose, 2005). برای فهم شاخص کسب اطلاعات، آنتروپی و شاخص کسب می توان به (Tsang, Kwong, & Wang, 2007) رجوع کرد و الگوریتم استنتاج C4.5 در (Larose, 2005) آمده است. این الگوریتم نسخه جدید الگوریتم ID3 است. در این مقاله به چند دلیل از الگوریتم درخت تصمیم گیری C4.5 به جای الگوریتم ID3 استفاده می شود:

۱. الگوریتم ID3 تنها ویژگی های اسمی را در ساخت درخت تصمیم گیری در نظر می گیرد، ولی الگوریتم C4.5 هر دو نوع ویژگی اسمی و عددی را لحاظ می کند.
۲. در الگوریتم ID3 ابتدا باید داده های مفقود را از بین برد، ولی C4.5 در الگوریتم خود با داده های مفقود مقابله می کند.
۳. الگوریتم C4.5 بر خلاف ID3 به هرس کردن درخت می پردازد (Hall, 1999). هرس درخت باعث کاهش اندازه درخت و پیچیدگی آن می شود. الگوریتم درخت تصمیم گیری C4.5 برای هر مقدار ویژگی اسمی به طور پیش فرض به تفکیک شاخه می پردازد که موجب پریشتر شدن<sup>۲۵</sup> درخت تصمیم گیری می شود (Larose, 2005).

یکی از روش‌های پیش‌پردازش داده می‌تواند باعث افزایش کیفیت مجموعه داده آموزش برای ساخت مدل طبقه‌بندی گردد. درختان تصمیم‌گیری به عنوان یکی از تکنیک‌های طبقه‌بندی نیز از این قاعده مستثنی نیستند (SALAPPA, DOUMPOS, & ZOPOUNIDIS, 2007). در این مقاله از الگوریتم‌های انتخاب ویژگی به عنوان یکی از روش‌ها برای پیش‌پردازش داده‌ها استفاده می‌شود.

تعاریف مختلفی از انتخاب ویژگی‌ها مطرح شده است. انتخاب ویژگی به شناسایی و انتخاب ویژگی‌های متمایز برای ساخت مدل‌ها و تفسیر بهتر داده‌ها می‌پردازد (Tan, Fu, Zhang, & Bourgeois, 2008). انتخاب ویژگی‌ها دارای مزایای متعددی است: ۱. باعث فهم آسان داده‌ها می‌شود. ۲. زمان یادگیری را در مدل کاهش می‌دهد. ۳. با انتخاب ویژگی‌ها نیاز کمتری به اندازه‌گیری و ذخیره‌سازی مقادیر ویژگی‌ها است (Guyon & Elisseeff, 2003).

با توجه به این موضوع می‌توان گفت که انتخاب ویژگی‌ها باعث می‌شود یک مدل اعتبارسنجی بهتری برای طبقه‌بندی مشتریان بانک‌ها تولید شود و از طرف دیگر در حجم زیاد داده‌های اعتبارسنجی، هزینه و زمان جمع‌آوری و بررسی ویژگی‌های مشتریان جدید کاهش یابد. در نتیجه کارشناسان اعتبارسنجی می‌توانند سریعتر به تصمیم‌گیری در مورد قبول یا رد تقاضای اعتبارسنجی مشتریان بانک‌ها بپردازند. همچنین با آسان شدن فهم داده‌ها، تفسیر رد یا قبول اعتبار برای کارشناسان اعتبارسنجی و متقاضیان اعتبار راحت‌تر می‌شود.

الگوریتم انتخاب ویژگی‌ها از سه قسمت تشکیل می‌شود (WANG & LI, 2008):

۱. معیار ارزیابی ویژگی.
۲. روش جستجو.
۳. قانون توقف.

به طور معمول معیارهای ارزیابی شامل موارد زیر است: ۱. اطلاعات. ۲. وابستگی. ۳. فاصله. ۴. سازگاری. ۵. دقت طبقه‌بندی. الگوریتم‌های انتخاب ویژگی که از ۴ معیار اول اشاره شده در بالا استفاده می‌کنند مبتنی بر رویکرد

فیلتر هستند. در این رویکرد، الگوریتم انتخاب ویژگی مستقل از الگوریتم طبقه‌بندی است. الگوریتم انتخاب ویژگی که از معیار دقت طبقه‌بندی برای انتخاب ویژگی‌ها استفاده می‌کند، از رویکرد Wrapper بهره می‌برد. در این رویکرد الگوریتم انتخاب ویژگی از الگوریتم یادگیری مثل الگوریتم طبقه‌بندی برای انتخاب ویژگی‌ها استفاده می‌کند. روش‌های جستجو در انتخاب ویژگی‌ها شامل ۳ روش است که عبارتند از: ۱. کامل. ۲. هیوریستک. ۳. تصادفی. دو روش کامل و هیوریستیک مربوط به فضاهای کوچک است و در مواردی مناسب می‌باشد که نیاز به کارایی بالا در فرایند جستجو می‌باشد. روش تصادفی مثل الگوریتم ژنتیک برای فضاهای بزرگ و پیچیده مناسب‌تر است. قوانین مختلفی برای توقف الگوریتم انتخاب ویژگی‌ها موجود است: ماکزیمم تعداد تکرار الگوریتم، کسب نتیجه بهتر توسط اضافه یا کم کردن یک ویژگی از مجموعه ویژگی‌ها، رسیدن به یک زیرمجموعه بهینه از ویژگی‌ها و ... (WANG & LI, 2008). یکی دیگر از روش‌های انتخاب ویژگی، طرح‌های جاسازی شده است. در این روش الگوریتم انتخاب ویژگی به عنوان بخشی از الگوریتم طبقه‌بندی لحاظ می‌شود (SALAPPA, DOUMPOS, & ZOPOUNIDIS, 2007).

در این مقاله از رویکردهای فیلتر، Wrapper و طرح‌جاسازی شده برای انتخاب ویژگی‌ها استفاده می‌شود. روش جستجو در انتخاب ویژگی‌ها به صورت تصادفی و مبتنی بر الگوریتم ژنتیک است و قانون توقف برای الگوریتم انتخاب ویژگی‌ها، ماکزیمم تعداد تکرار در الگوریتم انتخاب ویژگی می‌باشد. هر یک از الگوریتم‌های انتخاب ویژگی دارای یک تابع ارزیابی<sup>۲۷</sup> برای ارزیابی ویژگی‌ها هستند. سه الگوریتم اول از رویکرد فیلتر، الگوریتم چهارم از رویکرد Wrapper و الگوریتم آخر از رویکرد طرح‌جاسازی شده برای انتخاب ویژگی‌ها استفاده می‌کنند. در ادامه به توضیحی مختصر راجع به روش‌های انتخاب ویژگی در این مقاله پرداخته می‌شود. ۱) الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک توسط تابع ارزیابی همبستگی بین ویژگی‌ها با هم و با ویژگی هدف<sup>۲۸</sup>: تابع ارزیابی در این الگوریتم به بررسی همبستگی تک تک ویژگی‌ها با هم و با



(۴) الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک توسط تابع ارزیاب wrapper با طبقه کننده C4.5: توسط الگوریتم های یادگیری مثل الگوریتم C4.5، ویژگی ها را ارزیابی می کند. به عبارت دیگر الگوریتم انتخاب ویژگی، برای انتخاب ویژگی های مناسب از یک الگوریتم یادگیری بهره می برد. برای بکارگیری الگوریتم یادگیری در انتخاب ویژگی از داده های تست و آموزش در تکنیک اعتبارسنجی متقاطع استفاده می شود.

(۵) الگوریتم انتخاب ویژگی مبتنی بر درخت تصمیم گیری ژنتیکی: این الگوریتم مبتنی بر طرح جاسازی شده در انتخاب ویژگی ها است. زیرا با اجرای این الگوریتم درخت تصمیم گیری ایجاد می شود که از مجموعه ویژگی های این درخت تصمیم گیری به عنوان ویژگی های منتخب در ساخت درخت تصمیم گیری C4.5 استفاده می شود. الگوریتم درخت تصمیم گیری ژنتیکی مورد استفاده در انتخاب ویژگی ها برگرفته از (Papagelis & Kalles, n.d.) است. در این الگوریتم برای نمایش راه حل ها از روش درختی استفاده می شود. همچنین در ساخت درختان تصمیم گیری به تغییرات اساسی در الگوریتم ژنتیک می پردازد. هر ویژگی دارای یک مقدار تصادفی می باشد. اگر ویژگی اسمی باشد، یکی از مقادیر آن در هر تکرار به طور تصادفی انتخاب می شود و اگر ویژگی مورد نظر عددی باشد، در بازه تعریف شده آن مقدار آن تغییر می کند. عملگر جهش ویژگی را به طور تصادفی انتخاب و مقدار آن را به طور تصادفی تغییر می دهد و عملگر تقاطع با انتخاب ویژگی ها به صورت تصادفی زیر درخت های آن ها را جابجا می کند. تابع برازندگی در اینجا به بهینه کردن اندازه درخت و دقت طبقه بندی می پردازد.

#### ۴-۶- الگوریتم ژنتیک

واژه الگوریتم ژنتیک به توصیف یک مجموعه ای از رویه های جستجوی تصادفی می پردازد که از اصول ژنتیک طبیعی و اصل بقای برترین ها نشات گرفته شده

ویژگی هدف می پردازد. تابع ارزیاب در این الگوریتم، الگوریتمی است که با جستجوی هیوریستیک، به بررسی همبستگی بین ویژگی ها می پردازد. این تابع ارزیاب فرض می کند که بهترین ویژگی ها نسبت به هم همبستگی کمتر و نسبت به ویژگی هدف دارای همبستگی بیشتری می باشند. البته همبستگی در این تابع ارزیاب به وابستگی ویژگی ها با هم دلالت دارد و منظور همبستگی خطی کلاسیک نیست (Hall, 1999).

(۲) الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک توسط تابع ارزیابی سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف: به جستجوی کامل و جامع در فضای زیر مجموعه ویژگی ها می پردازد، تا اینکه کمینه ترین ترکیب از ویژگی ها را پیدا کند. سپس این مجموعه ویژگی ها به تقسیم مجموعه آموزشی در کلاس ها می پردازند. این الگوریتم توسط Liu و Setiono ابداع شد و ویژگی بارز آن این است که با اختلال در داده ها به خوبی برخورد می کند. این الگوریتم ابتدا به طور تصادفی یک زیر مجموعه با نام S از کل ویژگی ها انتخاب می کند. سپس در مرحله بعد یک زیر مجموعه دیگر از ویژگی ها تولید می شود. سطح سازگاری مقادیر ویژگی هدف توسط قرار دادن نمونه های مجموعه داده در این مجموعه ویژگی سنجیده می شود. اگر این سطح سازگاری کمتر باشد، این مجموعه جایگزین مجموعه قبلی می شود. نرخ ناسازگاری در مجموعه ویژگی ها در هر مرحله محاسبه می شود. این روند مرتباً ادامه می یابد تا مناسب ترین مجموعه ویژگی ها انتخاب شوند (Hall, 1999). برای فهم بیشتر می توان به (Liu&Setiono,n.d.) مراجعه کرد.

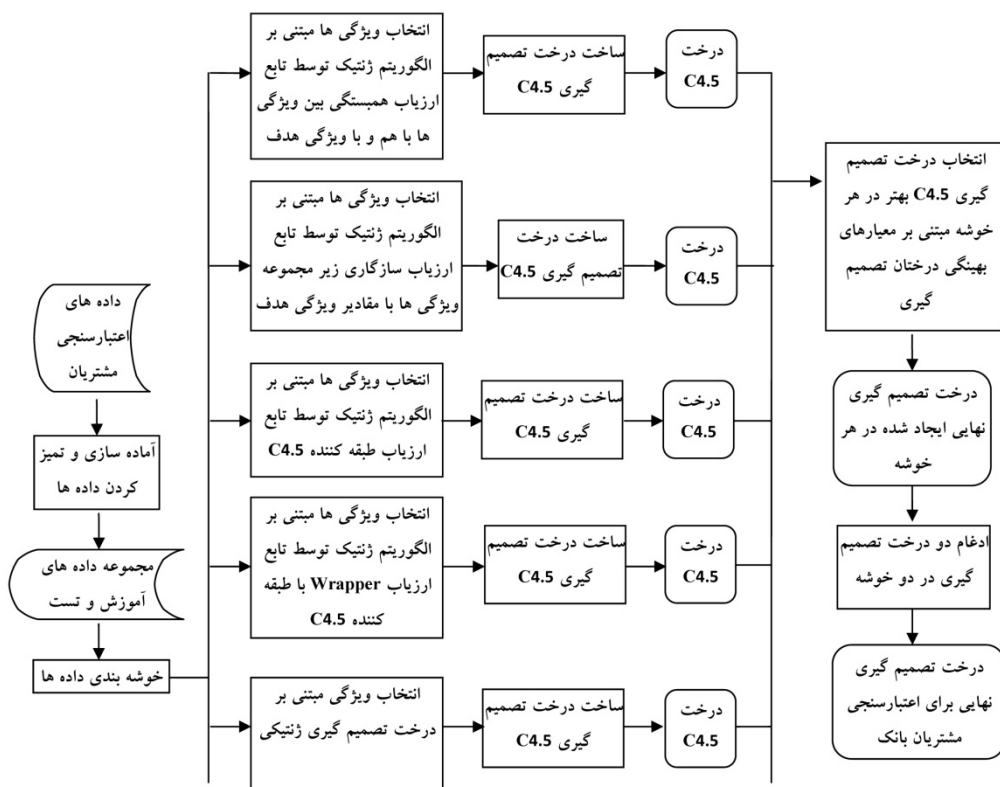
(۳) الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک توسط تابع ارزیاب طبقه کننده C4.5: مجموعه ویژگی ها را توسط تابع طبقه کننده C4.5 در مجموعه داده آموزشی ارزیابی می کند. میزان شایستگی زیر مجموعه ویژگی منتخب توسط شاخص دقت طبقه بندی الگوریتم C4.5 تعیین می شود.

های دیگر می‌پردازد. ۵. ایجاد جمعیت جدید توسط عملیات تقاطع و جهش: برخی راه حل‌ها ممکن است در یک جمعیت از راه حل‌های دیگر برتر باشند. به همین دلیل با انتخاب این مجموعه راه حل‌ها و توسط عملیات تقاطع و جهش جمعیت دیگری شکل می‌گیرد. مجدداً مراحل ۴ و ۵ انجام می‌شود تا شرط توقف الگوریتم پذیرفته شود.

#### ۴-۷- مدل ترکیبی پیشنهادی

در مدل ترکیبی پیشنهادی تحقیق از الگوریتم‌های انتخاب ویژگی مبتنی بر الگوریتم ژنتیک، درخت تصمیم‌گیری ژنتیکی، درختان تصمیم‌گیری C4.5 و همچنین الگوریتم SimpleKmeans برای خوشه‌بندی داده‌ها استفاده می‌شود. شکل ۱ فرایند ساخت و آزمون مدل ترکیبی پیشنهادی موضوع این مقاله را برای اعتبارسنجی مشتریان بانک نشان می‌دهد.

است. جان هلند<sup>۲۹</sup> اولین شخصی بود که در سال ۱۹۷۰ به طراحی الگوریتم ژنتیک پرداخت (1387, نادعلی & خان بابایی (Grune & Jooste, 2004 cited in). از آن موقع تا حال تحقیقات فراوانی پیرامون این تکنیک و کاربردهای آن انجام شده است. مراحل اجرای الگوریتم ژنتیک می‌تواند شامل موارد زیر باشد (1387, نادعلی & خان بابایی (Grune & Jooste, 2004 cited in): ۱. شناخت ژن‌ها: در الگوریتم ژنتیک برای نمایش یک ژن از یک بیت استفاده می‌شود. یک ژن نشان‌دهنده رفتار یک بخشی از راه حل مسئله است. ۲. سازمان‌دهی ژن‌ها در کروموزوم‌ها: به مجموعه ژن‌ها کروموزوم می‌گویند. هر کروموزوم شامل مقادیری است که جمعاً رفتار یک کروموزوم یا یک راه حل مسئله تبیین می‌کنند. ۳. ایجاد یک جمعیتی از راه حل‌های ممکن: به مجموعه‌ای از راه حل‌ها یا کروموزوم‌ها یک جمعیت می‌گویند. ۴. ارزیابی تک تک کروموزوم‌ها: در این مرحله الگوریتم ژنتیک به ارزیابی اثربخشی جمعیت اولیه برای مقایسه با جمعیت



شکل ۱ - فرایند ساخت و آزمون مدل ترکیبی پیشنهادی موضوع این مقاله در اعتبارسنجی مشتریان بانک

معیار 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 دارای اندازه و تعداد برگ های کمتری نسبت به درخت تصمیم دیگر است. در نهایت اینکه تصمیم گیری نهایی در انتخاب درخت تصمیم گیری C4.5 بهتر در هر خوشه به نظر کاربر یا کارشناس اعتبارسنجی بستگی دارد.

در چهار الگوریتم اول انتخاب ویژگی مبتنی بر الگوریتم ژنتیک مطرح در بخش ۵.۴ این مقاله، از روش گلدبرگ برای نمایش ژنتیکی کروموزوم ها استفاده می شود. 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. عملگر انتخاب از چرخ گردان برای انتخاب کروموزوم های برتر استفاده می کند. از عملگر تقاطع تک نقطه ای برای انجام عمل تقاطع استفاده می شود. بدین ترتیب که به طور تصادفی برشی بر روی دو کروموزوم ایجاد می شود و قسمت های برش خورده به صورت ضربدری یا اریب با هم جابجا می شوند. با انجام این عمل دو کروموزوم جدید بوجود می آید. عملگر جهش بدین صورت است که اگر مقدار یک ژن که به صورت تصادفی انتخاب می شود، صفر باشد، آن را تبدیل به یک می کند و اگر مقدار آن ژن یک باشد، آن را به صفر تبدیل می نماید. عملگر جایگزینی کروموزوم ها با کروموزوم های قبلی بر پایه شایستگی است. شرط توقف الگوریتم ژنتیک در اینجا تعداد نسل ها در نظر گرفته شد.

##### ۵- آموزش، تست مدل

به منظور آموزش و تست مدل ترکیبی پیشنهادی پس از آماده سازی داده های اعتباری آلمان، تعداد ۶۹۰ تراکنش از ۱۰۰۰ تراکنش مورد استفاده قرار گرفت. تعداد خوشه ها در الگوریتم خوشه بندی

همان طور که مشاهده می شود، این مجموعه داده، بعد از عملیات آماده سازی و تمیز کردن به دو قسمت داده های آموزش و تست تقسیم می شود و به وسیله تکنیک های خوشه بندی و الگوریتم های انتخاب ویژگی، پیش پردازش داده ها صورت می گیرد. توسط تکنیک خوشه بندی SimpleKmeans این داده ها به دو دسته خوشه بندی می شوند. در ادامه در هر خوشه توسط پنج الگوریتم انتخاب ویژگی، ویژگی های مهم انتخاب شده و از این ویژگی ها تعداد پنج درخت تصمیم گیری C4.5 (J48) در نرم افزار وکا ساخته می شود. سپس به انتخاب بهترین درختان تصمیم گیری در هر خوشه مبتنی بر معیارهای بهینگی درختان تصمیم گیری مطرح شده در بخش ۲ این مقاله پرداخته می شود. بعد از اینکه درختان تصمیم گیری C4.5 برتر در هر دو خوشه تعیین شد، این دو درخت با هم ترکیب شده و در نهایت درخت تصمیم گیری نهایی برای اعتبارسنجی مشتریان بانک ساخته می شود.

با توجه به اینکه انتخاب تعداد بهینه خوشه ها از مسائل پیچیده می باشد، در مدل ترکیبی پیشنهادی می توان ابتدا تعداد خوشه را عدد ۲ در نظر گرفت و با توجه به نیاز و نظر کاربر، مرتباً تعداد آن ها را افزایش داد تا به بهترین تعداد خوشه ای دست یافت که درختان تصمیم گیری در آن خوشه ها بهترین درختان تصمیم گیری مبتنی بر معیارهای بهینگی باشند. البته در این مقاله تعداد خوشه عدد ۲ در نظر گرفته شده و از این روش برای انتخاب تعداد بهینه خوشه ها 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 پیشگویی کننده انتخابی، تعداد برگ ها و اندازه درخت تصمیم گیری شود. می توان با روش هایی مثل هرس درخت تصمیم گیری و اعمال محدودیت هایی مثل مینیموم تعداد تراکنش در هر برگ به برقراری تعادل بین معیارهای بهینگی درختان تصمیم گیری پرداخت. ولی برای مقایسه درختان تصمیم گیری C4.5 نیز باید یک تعاملی بین ۴

هرس درخت، استفاده از فاکتور اطمینان<sup>۳۲</sup> 0.25 در فرایند ساخت درخت تصمیم‌گیری و تعیین مینیموم تعداد تراکنش با عدد ۲ در هر برگ درخت تصمیم‌گیری برای هرس درخت و تعیین اندازه و پیچیدگی درخت، عدم استفاده از جداکننده‌های دودویی، تعداد دسته‌ها عدد ۳ (یعنی تعداد دو و یک دسته به ترتیب برای رشد و هرس درخت)، عدم استفاده از روش هرس خطای کاهش یافته، عدد تصادفی seed برابر ۱ برای تصادفی کردن داده در زمان استفاده از روش هرس خطای کاهش یافته، زیاد کردن زیر درخت، رویکرد اعتبارسنجی متقاطع در آموزش و تست درخت تصمیم‌گیری C4.5 با عدد ۱۰. بهتر است در ساخت درختان تصمیم‌گیری و مخصوصاً اعتبارسنجی مشتریان بانک‌ها همواره بین میزان پیچیدگی و دقت طبقه‌بندی مدل‌های طبقه‌بندی یک تعادل برقرار باشد. کاهش فاکتور اطمینان منجر به افزایش هرس درخت می‌شود. افزایش مینیموم تعداد تراکنش در هر برگ باعث می‌شود تعداد بیشتری تراکنش در یک برگ قرار گیرد و در نتیجه تعداد برگ‌ها، تعداد ویژگی‌های پیشگو منتخب و اندازه درخت کاهش یافته، ولی احتمال کاهش دقت طبقه‌بندی وجود دارد. پس برقراری تعادل بین معیارهای بهینگی درختان تصمیم‌گیری برای اعتبارسنجی مشتریان بانک‌ها ضروری است که در این مقاله این موضوع لحاظ شده است.

#### ۶- مقایسه درخت تصمیم‌گیری حاصل از مدل

##### ترکیبی پیشنهادی با سایر درختان تصمیم‌گیری

تا این قسمت از مقاله به ارائه مدل ترکیبی پیشنهادی برای اعتبارسنجی مشتریان بانک پرداخته شد. این مدل به طور مختصر با توجه به شکل ۱ مراحل زیر را برای اعتبارسنجی مشتریان بانک‌ها انجام می‌دهد: ۱. خوشه‌بندی داده‌ها. ۲. انتخاب ویژگی‌ها توسط پنج الگوریتم انتخاب ویژگی. ۳. ساخت درختان تصمیم‌گیری C4.5 از هر یک از مجموعه ویژگی‌های انتخابی در هر خوشه. ۴. انتخاب بهترین درختان تصمیم‌گیری در هر خوشه مبتنی بر معیارهای بهینگی. ۵. ترکیب دو درخت

SimpleKmeans و عدد seed به ترتیب برابر دو و یک در نظر گرفته شد. پارامترهای الگوریتم ژنتیک در چهار الگوریتم انتخاب ویژگی ابتدایی در بخش ۵.۴ به قرار زیر است: نرخ تقاطع ۰.۹، نرخ جهش ۰.۰۱، تعداد نسل و جمعیت اولیه ۲۰ و عدد تصادفی seed برابر ۱ در نظر گرفته شد. از عدد اعتبارسنجی متقاطع ۱۰ برای آموزش و تست مدل استفاده شد. بدین ترتیب که ابتدا یک دهم اول داده‌ها برای تست استفاده می‌شود و بقیه برای آموزش الگوریتم انتخاب ویژگی یا درخت تصمیم‌گیری C4.5 بکار می‌رود. سپس یک دهم بعدی و به همین ترتیب ۱۰ بار این عمل صورت می‌گیرد و از نتایج این مراحل میانگین گرفته می‌شود. تعداد دسته‌ها و عدد seed و حد آستانه<sup>۳۰</sup> در الگوریتم انتخاب ویژگی با تابع ارزیاب Wrapper با طبقه‌بندی کننده 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 یعنی درصد تعداد درختان بد که در نسل‌ها جایگزین می‌شوند، برابر ۰.۲۵، نرخ خطا برابر ۰.۹۵ (زمانی که خطای طبقه‌بندی یک درخت از حد معین تعریف شده توسط این پارامتر بیشتر شود، از طبقه‌بندی تراکنش‌ها در مجموعه تست جلوگیری می‌شود تا منابع حفظ شود. با افزایش مقدار این پارامتر، سرعت تکامل<sup>۳۱</sup> در الگوریتم درخت تصمیم‌گیری ژنتیکی افزایش می‌یابد.)، ترجیح قابلیت درخت تصمیم‌گیری با دقت بالاتر بر درخت تصمیم‌گیری کوچکتر، عدم تغییر پویا در ترجیح درختان تصمیم‌گیری با دقت بالاتر بر درخت تصمیم‌گیری کوچکتر در ابتدا و انتهای فرایند تکامل، نرخ تقاطع عدد ۰.۹۹، نرخ جهش عدد ۰.۰۱، تعداد نسل‌ها عدد 100، جمعیت اولیه عدد 100، عدد تصادفی seed برابر 100. ۱۲۳۴۵۶۷۸۹.

مشخصات درخت تصمیم‌گیری C4.5 به قرار زیر است: ماتریس هزینه‌های طبقه‌بندی غلط در ساخت درخت تصمیم‌گیری C4.5 در نظر گرفته نشد، استفاده از

در طبقه بندی بوده و به میزان کمی دارای پیچیدگی بیشتر نسبت به برخی از درختان تصمیم گیری است. کمترین تعداد ویژگی پیشگویی کننده انتخابی، تعداد برگ ها و اندازه درخت، مربوط به درخت تصمیم گیری است که از الگوریتم درخت تصمیم گیری ژنتیکی برای انتخاب ویژگی ها بهره برده است. شاید دلیل این امر تعداد ویژگی های انتخابی کم برای ساخت درخت تصمیم گیری C4.5 است که از الگوریتم درخت تصمیم گیری ژنتیکی برای انتخاب ویژگی ها استفاده کرده است. اما این درخت دارای دقت کمتری در طبقه بندی مشتریان نسبت به اکثر درختان تصمیم گیری در جدول ۴ است. کمترین میزان دقت طبقه بندی، مربوط به درخت تصمیم گیری C4.5 با تابع ارزیاب "سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف" در انتخاب ویژگی است. به نظر می رسد بدترین درخت به لحاظ معیارهای بهینگی، این درخت باشد. زیرا دارای کمترین دقت در طبقه بندی و بیشترین تعداد برگ ها و اندازه درخت را در بین درختان تصمیم گیری جدول ۴ دارا است. جدول ۵ به ارائه نتایج حاصل از اجرای مدل ترکیبی پیشنهادی و ساخت درختان تصمیم گیری C4.5 در خوشه دوم می پردازد. تعداد ویژگی های حاصل از اجرای پنج الگوریتم انتخاب ویژگی مطرح در بخش ۵.۴ به ترتیب برابر است با: ۶، ۱۲، ۱۵، ۶ و ۹.

تصمیم گیری بهتر در هر خوشه و ساخت درخت تصمیم گیری نهایی برای اعتبارسنجی مشتریان بانک.

در این بخش به ارائه یافته ها و تحلیل نتایج حاصل از اجرای مدل ترکیبی پیشنهادی این مقاله پرداخته شده و این نتایج با نتایج سایر روش های ساخت درختان تصمیم گیری مطرح در این مقاله مقایسه می شود. بدین منظور از مجموعه داده های اعتباری آلمان که عملیات آماده سازی روی آن ها اعمال شد، استفاده می شود. نتایج جداول ۴ الی ۹ مبتنی بر خروجی نرم افزار وکا در ساخت درختان تصمیم گیری C4.5 است. تنها برای انتخاب ویژگی ها مبتنی بر درخت تصمیم گیری ژنتیکی از نرم افزار GATree استفاده شده است. جدول ۴ به ارائه نتایج حاصل از اجرای مدل ترکیبی پیشنهادی و ساخت درختان تصمیم گیری C4.5 در خوشه اول می پردازد.

تعداد ویژگی های حاصل از اجرای پنج الگوریتم انتخاب ویژگی مطرح در بخش ۵.۴ به ترتیب برابر است با: ۷، ۱۵، ۱۶، ۱۴ و ۷. همان طور که در جدول ۴ مشاهده می شود، درخت تصمیم گیری C4.5 با تابع ارزیاب "طبقه کننده C4.5" در انتخاب ویژگی، دارای بیشترین درصد مشاهدات درست طبقه بندی شده است. از طرفی این درخت بیشترین ویژگی پیشگویی کننده انتخابی را دارا می باشد. به نظر می رسد این درخت در خوشه اول بهترین درخت باشد؛ زیرا نسبت به بقیه درختان تصمیم گیری در خوشه اول دارای بالاترین دقت

#### جدول ۴ - نتایج حاصل از اجرای الگوریتم های ساخت درخت تصمیم گیری C4.5

##### مبتنی بر مدل ترکیبی پیشنهادی در خوشه اول

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی های پیشگویی کننده انتخابی	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	wrapper با طبقه کننده C4.5	۴۴۵	۱۱	۳۷۳	۸۳.۸۲۰۲٪	۲۸	۴۲	۰.۸۲۹	۰.۸۶۲
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۴۴۵	۵	۳۵۳	۷۹.۳۲۵۸٪	۱۸	۲۶	۰.۸۰۹	۰.۷۶۱
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۴۴۵	۱۳	۳۱۶	۷۱.۰۱۱۲٪	۴۳	۶۳	۰.۷۲۲	۰.۶۷۳
۴	طبقه کننده C4.5	۴۴۵	۱۴	۳۷۶	۸۴.۴۹۴۴٪	۳۶	۵۵	۰.۸۳	۰.۸۸۲
۵	مبتنی بر درخت تصمیم گیری ژنتیکی	۴۴۵	۶	۳۴۳	۷۷.۰۷۸۷٪	۵	۸	۰.۷۳۵	۰.۹۷۱

## جدول ۵ - نتایج حاصل از اجرای الگوریتم‌های ساخت درخت تصمیم‌گیری C4.5

مبتنی بر مدل ترکیبی پیشنهادی در خوشه دوم

ردیف	تابع ارزیابی انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی‌های پیشگویی‌کننده انتخابی	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ‌ها	اندازه درخت	دقت مشتریان خوب	دقت مشتریان بد
۱	wrapper با طبقه‌کننده C4.5	۲۴۵	۳	۲۶	۸۴.۰۸۱۶٪	۶	۹	۰.۸۵۶	۰.۸۱۲
۲	همبستگی بین ویژگی‌ها با هم و با ویژگی هدف	۲۴۵	۳	۲۰۶	۸۴.۰۸۱۶٪	۶	۹	۰.۸۵۶	۰.۸۱۲
۳	سازگاری زیر مجموعه ویژگی‌ها با مقادیر ویژگی هدف	۲۴۵	۹	۱۸۰	۷۳.۴۶۹۴٪	۱۹	۲۷	۰.۷۲	۰.۸
۴	طبقه‌کننده C4.5	۲۴۵	۹	۲۰۳	۸۲.۸۵۷۱٪	۱۹	۲۹	۰.۸۴۵	۰.۷۹۸
۵	مبتنی بر درخت تصمیم‌گیری ژنتیکی	۲۴۵	۳	۱۹۰	۷۷.۵۵۱٪	۷	۱۰	۰.۷۴۵	۰.۹۱۱

## جدول ۶ - نتایج حاصل از اجرای الگوریتم‌های ساخت درخت تصمیم‌گیری C4.5 مبتنی بر مدل ترکیبی پیشنهادی

ردیف	الگوریتم درخت تصمیم‌گیری	کل مشاهدات	تعداد ویژگی‌های پیشگویی‌کننده انتخابی	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ‌ها	اندازه درخت	دقت مشتریان خوب	دقت مشتریان بد
۱	درخت تصمیم‌گیری حاصل از مدل ترکیبی پیشنهادی	۶۹۰	۱۵	۵۸۲	۸۴.۳۴۷۸٪	۴۲	۶۵	۰.۸۳۸۹	۰.۸۵۳۷

طبقه بندی، برتری نسبت به دو درخت تصمیم‌گیری با دقت بالا در جدول ۵ ندارد. بهترین درخت در خوشه دوم، درخت تصمیم‌گیری است که تابع ارزیابی آن در انتخاب ویژگی‌ها برابر همبستگی بین ویژگی‌ها با هم و با ویژگی هدف یا Wrapper با طبقه‌کننده C4.5 است؛ زیرا هر دو درخت نتایج یکسانی را ارائه می‌دهند. این درختان دارای بالاترین دقت طبقه بندی در بین درختان تصمیم‌گیری خوشه دوم بوده و به لحاظ پیچیدگی نسبت به بسیاری از درختان تصمیم‌گیری بهتر هستند.

بر طبق مدل ترکیبی پیشنهادی مقاله در اعتبارسنجی مشتریان بانک، بعد از اینکه در هر خوشه درختان تصمیم‌گیری ساخته شدند، نوبت به انتخاب بهترین درخت تصمیم‌گیری مبتنی بر معیارهای بهینگی می‌رسد. این کار در بالا انجام شد. بعد از انتخاب بهترین درختان تصمیم‌گیری در هر خوشه نوبت به ترکیب دو درخت تصمیم‌گیری و ساخت درخت تصمیم‌گیری نهایی برای اعتبارسنجی مشتریان بانک می‌رسد. نتایج حاصل از ساخت درخت تصمیم‌گیری توسط مدل

تعداد ویژگی‌های انتخابی در الگوریتم‌های انتخاب ویژگی با توابع ارزیابی همبستگی بین ویژگی‌ها با هم و با ویژگی هدف و "Wrapper با طبقه‌کننده C4.5" کم است. از طرف دیگر پیچیدگی درختان تصمیم‌گیری C4.5 که از این دو روش برای انتخاب ویژگی‌ها استفاده کرده‌اند، نسبت به سایر درختان تصمیم‌گیری در جدول ۵ کمتر شده است. زیرا دارای تعداد ویژگی‌های پیشگویی‌کننده انتخابی، تعداد برگ‌ها و اندازه درخت کمتری هستند. بیشترین میزان "درصد مشاهدات درست طبقه بندی شده" نیز مربوط به این دو درخت است. درخت تصمیم‌گیری با تابع ارزیابی "سازگاری زیر مجموعه ویژگی‌ها با مقادیر ویژگی هدف" دارای کمترین میزان دقت در طبقه بندی مشتریان بوده و پیچیدگی آن نیز نسبتاً بالا است. درخت تصمیم‌گیری حاصل از الگوریتم انتخاب ویژگی مبتنی بر درخت تصمیم‌گیری ژنتیکی دارای دقت طبقه بندی کمتری نسبت به ۳ درخت تصمیم‌گیری جدول ۵ است. پیچیدگی آن نیز در حد قابل قبولی پایین است. ولی به دلیل کم بودن میزان دقت

خوشه اول و دوم استفاده نکرد. جدول ۷ نتایج حاصل از اجرای مدل ترکیبی پیشنهادی بدون خوشه بندی را نشان می دهد.

همان طور که در جدول ۷ مشاهده می شود، درختان تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 و "مبتنی بر درخت تصمیم گیری ژنتیکی در انتخاب ویژگی" بهترین درختان به ترتیب در معیارهای دقت و پیچیدگی هستند. بهترین درخت تصمیم گیری در جدول ۷، درخت تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 است. زیرا بالاترین دقت را در طبقه بندی و اعتبارسنجی مشتریان دارد. همچنین پیچیدگی آن نسبت به بقیه درختان تصمیم گیری به جز درخت تصمیم گیری با انتخاب ویژگی ها مبتنی بر درخت تصمیم گیری ژنتیکی کمتر است. با مقایسه نتایج جداول ۶ و ۷ مشاهده می شود که خوشه بندی مشتریان قبل از ساخت درختان تصمیم گیری در مدل ترکیبی پیشنهادی نتایج بهتری را در معیار درصد مشاهدات درست طبقه بندی شده در مجموعه داده اعتباری آلمان ارائه داد. در صورتی که نیاز به تصمیم گیری منعطف، ساده و بدون پیچیدگی در اعتبارسنجی است، از درخت تصمیم گیری استفاده شود که در انتخاب ویژگی ها از تابع ارزیاب Wrapper با طبقه کننده C4.5 استفاده کرده است و از قبل عمل خوشه بندی برای ساخت آن صورت نگرفته است.

ترکیبی پیشنهادی این مقاله در مجموعه داده های اعتباری آلمان به صورت جدول ۶ است.

"ویژگی های پیشگویی کننده انتخابی" بهترین درخت تصمیم گیری خوشه اول در "ویژگی های پیشگویی کننده انتخابی" بهترین درخت تصمیم گیری خوشه دوم موجود است. اگر ویژگی "نوع خوشه" به این مجموعه ویژگی اضافه شود، تعداد کل ویژگی های پیشگویی کننده انتخابی درخت تصمیم گیری مدل ترکیبی پیشنهادی برابر با ۱۵ است. تعداد مشاهدات درست طبقه بندی شده برابر با مجموع تعداد مشاهدات درست طبقه بندی شده در درختان تصمیم گیری منتخب در دو خوشه است. تعداد برگ ها در درخت تصمیم گیری مدل ترکیبی پیشنهادی برابر با مجموع تعداد برگ های درختان تصمیم گیری منتخب در هر دو خوشه است. اندازه درخت در این مدل برابر مجموع اندازه درخت در درختان تصمیم گیری دو خوشه به علاوه گره با ویژگی "نوع خوشه" است. این ویژگی در ابتدای درخت تصمیم گیری مدل ترکیبی 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 پیشنهادی دارای مقادیر "خوشه اول" و "خوشه دوم" است.

می توان به ساخت درختان تصمیم گیری به وسیله مدل ترکیبی پیشنهادی پرداخت. تنها با این تفاوت که در ابتدا از خوشه بندی برای تفکیک مشتریان به دو

جدول ۷ - نتایج حاصل از اجرای مدل ترکیبی پیشنهادی بدون استفاده از خوشه بندی در ساخت درخت تصمیم گیری C4.5

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی های پیشگویی کننده انتخابی	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت مشتریان خوب	دقت مشتریان بد
۱	wrapper با طبقه کننده C4.5	۶۹۰	۶	۵۷۰	٪۸۲.۶۰۸۷	۱۴	۲۱	۰.۸۲۶	۰.۸۲۵
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۶۹۰	۶	۵۵۳	٪۸۰.۱۴۴۹	۲۲	۳۴	۰.۸۰۹	۰.۷۸۴
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۶۹۰	۱۴	۴۸۹	٪۷۰.۸۶۹۶	۶۸	۱۰۱	۰.۷۲۳	۰.۶۶۷
۴	طبقه کننده C4.5	۶۹۰	۱۵	۵۶۷	٪۸۲.۱۷۳۹	۴۹	۷۳	۰.۸۱۵	۰.۸۳۹
۵	مبتنی بر درخت تصمیم گیری ژنتیکی	۶۹۰	۴	۵۶۲	٪۸۱.۴۴۹۳	۱۱	۱۸	۰.۸۲۲	۰.۷۹۸

۸ مشاهده می‌شود که درخت تصمیم‌گیری C4.5 با تابع ارزیاب Wrapper با طبقه‌بندی کننده C4.5 دارای دقت بالا و پیچیدگی کمتر نسبت به بهترین درخت تصمیم‌گیری C4.5 حاصل از مدل ترکیبی پیشنهادی است.

برای اینکه بتوان درخت تصمیم‌گیری مدل ترکیبی پیشنهادی را با درختان تصمیم‌گیری دیگری مقایسه کرد که در آن‌ها از خوشه‌بندی و انتخاب ویژگی‌ها استفاده نشده است، نتایج دو درخت تصمیم‌گیری C4.5 در جدول ۹ آمده است.

جدول ۹ به ارائه نتایج اجرای دو نوع دیگر الگوریتم درخت تصمیم‌گیری C4.5 و مقایسه آن‌ها با درخت تصمیم‌گیری حاصل از مدل ترکیبی پیشنهادی می‌پردازد. در ردیف ۱ از جدول ۹ نتایج الگوریتم درخت تصمیم‌گیری C4.5 ارائه شده است که در آن از تکنیک‌های خوشه‌بندی و انتخاب ویژگی‌ها استفاده نشده است. این درخت دارای دقت طبقه‌بندی کمتری نسبت به درخت تصمیم‌گیری حاصل از مدل ترکیبی پیشنهادی است؛ ولی تعداد برگ‌ها، اندازه درخت و تعداد ویژگی‌های پیشگویی کننده انتخابی آن کمتر است. در نتیجه پیچیدگی کمتری دارد. در ردیف ۲ از جدول ۹ نتایج درخت تصمیم‌گیری C4.5 نشان داده شده است که همانند درخت ردیف ۱ از خوشه‌بندی و انتخاب ویژگی‌ها استفاده نکرده ولی ویژگی "نوع خوشه" در مجموعه ویژگی‌ها برای ساخت این درخت وجود دارد. این درخت تصمیم‌گیری دارای دقت طبقه‌بندی، تعداد ویژگی‌های پیشگویی کننده انتخابی، تعداد برگ‌ها و اندازه درخت کمتری نسبت به درخت تصمیم‌گیری مدل ترکیبی پیشنهادی است.

می‌توان ویژگی "نوع خوشه" را در ساخت درخت تصمیم‌گیری در نظر گرفت. برای این کار به مجموعه داده اعتباری آلمان یک ویژگی پیشگویی کننده با نام ویژگی "نوع خوشه" اضافه می‌شود که توسط آن نوع خوشه هر تراکنش مشتری تعیین شده است. پارامترهای خوشه‌بندی در اینجا همانند مدل ترکیبی پیشنهادی است. جدول ۸ نتایج حاصل از اجرای این الگوریتم را نشان می‌دهد. الگوریتم ساخت درختان تصمیم‌گیری در این روش همانند الگوریتم ساخت درختان تصمیم‌گیری در مدل ترکیبی پیشنهادی است. مقادیر پارامترهای الگوریتم‌ها در این روش همانند الگوریتم‌های ساخت درختان تصمیم‌گیری در مدل ترکیبی پیشنهادی است. همان‌طور که در جدول ۸ مشاهده می‌شود، درخت تصمیم‌گیری با تابع ارزیاب Wrapper با طبقه‌بندی کننده C4.5 دارای بالاترین میزان دقت طبقه‌بندی است. کمترین پیچیدگی مربوط به درخت تصمیم‌گیری است که در انتخاب ویژگی‌ها از درخت تصمیم‌گیری ژنتیکی استفاده کرده است. به نظر می‌رسد بهترین درخت تصمیم‌گیری در جدول ۸، درخت تصمیم‌گیری با تابع ارزیاب Wrapper با طبقه‌بندی کننده C4.5 باشد. زیرا دارای دقت قابل توجه بالا در طبقه‌بندی مشتریان اعتبارسنجی نسبت به سایر درختان تصمیم‌گیری جدول ۸ است. به گونه‌ای که شاید بتوان از پیچیدگی نسبی این درخت نسبت به برخی از درختان تصمیم‌گیری جدول ۸ چشم‌پوشی کرد. نکته قابل ذکر در اینجا این است که ویژگی "نوع خوشه" در درخت تصمیم‌گیری با تابع ارزیاب Wrapper با طبقه‌بندی کننده C4.5 به عنوان یک ویژگی پیشگویی کننده وجود دارد. با مقایسه نتایج جداول ۶ و

جدول ۸ - نتایج حاصل از اجرای الگوریتم ساخت درخت تصمیم‌گیری C4.5 با در نظر گرفتن ویژگی "نوع خوشه" در

#### مجموعه ویژگی‌ها

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی‌های پیشگویی کننده انتخابی	تعداد مشاهدات درست طبقه‌بندی شده	درصد مشاهدات درست طبقه‌بندی شده	تعداد برگ‌ها	اندازه درخت	دقت کلاس مشتریان بد	دقت کلاس مشتریان خوب
۱	wrapper با طبقه‌بندی کننده C4.5	۶۹۰	۱۳	۵۸۴	٪ ۸۴.۶۳۷۷	۳۷	۵۹	۰.۸۶۹	۰.۸۳۷
۲	همبستگی بین ویژگی‌ها با هم و با ویژگی هدف	۶۹۰	۴	۵۵۶	٪ ۸۰.۵۷۹۷	۱۱	۱۸	۰.۷۸۲	۰.۸۱۷
۳	سازگاری زیر مجموعه ویژگی‌ها با مقادیر ویژگی هدف	۶۹۰	۱۳	۴۸۲	٪ ۶۹.۸۵۵۱	۶۲	۹۳	۰.۶۱۳	۰.۷۳۹



۴	طبقه کننده C4.5	۶۹۰	۱۵	۵۶۴	% ۸۱.۷۳۹۱	۴۸	۷۶	۰.۸۱۹	۰.۸۱۴
۵	مبتنی بر درخت تصمیم گیری ژنتیکی	۶۹۰	۴	۵۴۴	% ۷۸.۸۴۰۶	۹	۱۳	۰.۷۴۷	۱

جدول ۹ - نتایج حاصل از بکارگیری الگوریتم های درخت تصمیم گیری C4.5

ردیف	الگوریتم درخت تصمیم گیری	کل مشاهدات	تعداد ویژگی های پیشگویی کننده انتخابی	تعداد مشاهدات درست 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 C4.5	۶۹۰	۱۴	۵۷۳	% ۸۳.۰۴۳۵	۴۰	۶۰	۰.۸۲۶	۰.۸۴۱
۲	درخت تصمیم گیری C4.5 با ویژگی "نوع خوشه"	۶۹۰	۱۲	۵۷۲	% ۸۲.۸۹۸۶	۳۶	۵۶	۰.۸۳۱	۰.۸۲۴

ترکیب آن ها درخت تصمیم گیری حاصل از فرایند ساخت و آزمون مدل ترکیبی پیشنهادی برای اعتبارسنجی مشتریان بانک ها ایجاد شد. نتایج حاصل از اجرای الگوریتم مدل ترکیبی پیشنهادی نشان داد که دقت طبقه بندی درخت تصمیم گیری حاصل از آن نسبت به بسیاری از درختان تصمیم گیری مقایسه شده در این مقاله بیشتر بود. تنها درخت تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 و مبتنی بر الگوریتم ژنتیک در انتخاب ویژگی ها که ویژگی "نوع خوشه" را در ساخت خود لحاظ کرد، دارای بیشترین دقت طبقه بندی بوده و پیچیدگی آن نیز از درخت تصمیم گیری مدل ترکیبی پیشنهادی کمتر بود. پیچیدگی درخت تصمیم گیری مدل ترکیبی پیشنهادی نسبت به برخی از درختان تصمیم گیری بیشتر بود؛ ولی دقت بالای طبقه بندی آن باعث شد این درخت یک برتری نسبی نسبت به آن ها در اعتبارسنجی مشتریان داشته باشد. با توجه به نکات بالا می توان از مدل ترکیبی پیشنهادی برای ساخت درختان تصمیم گیری بهینه در اعتبارسنجی مشتریان بانک ها استفاده نمود. همچنین می توان هزینه طبقه بندی غلط مشتریان خوب و بد را در ساخت درختان تصمیم گیری لحاظ کرد که این شاخص در این مقاله، در نرم افزار وکا برای دو نوع مشتری یکسان لحاظ شد.

در این بخش از مقاله به ارائه نتایج حاصل از اجرای مدل ترکیبی پیشنهادی مقاله و مقایسه آن با برخی درختان تصمیم گیری C4.5 پرداخته شد. در انتخاب بهترین درختان تصمیم گیری در هر خوشه نظرات کارشناسان اعتبارسنجی را می توان لحاظ کرد. همچنین پیچیدگی طبقه بندی و هزینه طبقه بندی غلط مشتریان اعتبارسنجی در انتخاب بهترین درختان تصمیم گیری مهم هستند.

#### ۷- نتیجه گیری

بانک ها به منظور اعطای تسهیلات اعتباری به مشتریان خود نیازمند اعتبارسنجی آن ها هستند. تکنیک های متنوعی مثل درخت تصمیم گیری C4.5 می تواند به بانک ها در شناسایی مشتریان خوب و بد به لحاظ اعتباری کمک کند. انتخاب درخت تصمیم گیری بهتر که بتواند نسبت به سایر درختان تصمیم گیری به طبقه بندی مشتریان با دقت بالا و پیچیدگی کمتر پردازد، می تواند باعث شود تا کارشناسان اعتبارسنجی بتوانند با صرف زمان و هزینه کم و افزایش رضایت مشتری اعتبارسنجی به طبقه بندی مشتریان خود پردازند. در این مقاله به ارائه مدل جدیدی پرداخته شد که با استفاده از تکنیک های الگوریتم ژنتیک، انتخاب ویژگی ها و خوشه بندی به ساخت درختان تصمیم گیری C4.5 پردازد. برای آزمون مدل از مجموعه داده اعتباری آلمان استفاده شد. در نهایت مبتنی بر معیارهای بهینگی درختان تصمیم گیری مطرح در این مقاله، بهترین درختان تصمیم گیری در هر خوشه انتخاب شده و با

## فهرست منابع

- 12) Grupe, F. H., & Jooste, S.(2004), "Genetic Algorithms A Business Perspective", *Information Management & Computer Security*, Vol.12, No.3, pp.1-4.
- 13) Guyon, I., & Elisseeff, A.(2003), "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, Vol.3, pp.2,4,9.
- 14) Hall, M. A.(1999), "Corrolation Based Feature Selection for Machine Learning." Ph.D., University of Waikato.
- 15) Hsu, P. L., Lai, R., Chiu, C. C., & Hsu, C. I., (2003), "The hybrid of association rule algorithms and genetic algorithms for tree induction: an example of predicting the student course performance". *Expert Systems with Applications*, Vol.25, pp.1,2,4,5,6,7,11.
- 16) Huang, C. L., Chen, M. C., & Wang, C. J., (2007), "Credit Scoring with a data mining Approach Based on Support Vector Machines", *Expert Systems with Applications*, Vol.33, pp.1-3.
- 17) Huang, M., Gong, J., Shi, Z., Liu, C., & Zhang, L. (2007), "Genetic algorithm-based decision tree classifier for remote sensing mapping with SPOT-5 data in the HongShiMao watershed of the loess plateau, China". *Neural Comput & Applic*, Available at: [www.springer.com](http://www.springer.com), pp.1-3.
- 18) <http://archive.ics.uci.edu/ml/datasets.html>.
- 19) Kennedy, R. L., Lee, Y., Roy, B. V., Reed, C. D., & Lippmann, R. P.(1998), "Solving Data Mining Problems through Pattern Recognition", S.I., Prentice Hall.
- 20) Kim, E., Kim, W., & Lee, Y.(2002), "Combination of multiple classifiers for the customer's purchase behavior prediction", *Decision Support Systems*, Vol.34, pp.2-8.
- 21) Kim, M. J., & Han, I.(2003), "The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms". *Expert Systems with Applications*, Vol.25, pp.1-5,8.
- 22) Kim, Y. S., & Sohn, S. Y.(2004), "Managing loan customers using misclassification patterns of credit scoring model", *Expert Systems with Applications*, Vol.26, pp.1-3.
- 23) Larose, D. T.(2005), "Discovering Knowledge in Data, an Introduction to Data Mining", New Jersey: WILEY.
- 1) اداره مطالعات و کنترل ریسک بانک تجارت (۱۳۸۶). "مدیریت ریسک در بانکداری". چاپ اول. تهران: موسسه انتشارات و چاپ دانشگاه تهران.
- ۲) نادعلی، ا.، & خان بابایی، م. (۱۳۸۷)، "بکارگیری تکنیک‌های درخت تصمیم و الگوریتم ژنتیک جهت اعتبارسنجی مشتریان بانک‌ها در یک سیستم پشتیبانی تصمیم‌گیری"، دومین کنفرانس ملی داده کاوی، تهران: دانشگاه صنعتی امیرکبیر، صفحه 6، 8.
- 3) Abdou, H., & Pointon, J.( 2008), "Neural nets versus conventional techniques in credit scoring in Egyptian banking", *Expert Systems with Applications*, Vol.xxx, pp.1.
- 4) Aitkenhead, M. J.(2008), "A co-evolving decision tree classification method", *Expert Systems with Applications*, Vol.34, pp.2-3.
- 5) Bala, J., Huang, J., Vafaie, H., DeJong, K., & Wechsler, H. (1995), "Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification", *IJCAI, Montreal: IJCAI conference*, pp. 1,2,4.
- 6) Carvalho, D. R., & Freitas, A. A. (2002) "A genetic-algorithm for discovering small-disjunct rules in data mining", *Applied Soft Computing*, Vol.2, pp. 2-8,12.
- 7) Carvalho, D. R., & Freitas, A. A. (2004), "A hybrid decision tree/genetic algorithm method for data mining", *Information Sciences*, Vol.163, pp.1-18.
- 8) D'heygere, T., Goethals, P. L., & Pauw, N. D. (2006), "Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks", *Ecological Modelling*, Vol.195, pp.1-5.
- 9) D'heygere, T., Goethals, P. L., & Pauw, N. D.(2003), "Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates", *Ecological Modelling*, Vol.160, pp.1-8.
- 10) Dehuri, S., Patnaik, S., Ghosh, A., & Mall, R., (2008), "Application of elitist multi-objective genetic algorithm for classification rule generation". *Applied Soft Computing*, Vol.8, pp.1,2,3,5.
- 11) Gray, J. B., & Fan, G. (2008), "Classification tree analysis using TARGET". *Computational Statistics & Data Analysis*, Vol.52, pp.1-3.

- experimental evaluation", *Optimization Methods and Software*, Vol.22, No.1, pp.2-5.
- 36) Sorensen, K., & Janssens, G. K., (2003), "Data mining with genetic algorithms on binary trees", *European Journal of Operational Research* Vol.151, pp.2,10.
- 37) Susac, M. Z., Sarlija, N., & Bencic, M., (n.d.), "Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models", s.n, pp.1-4.
- 38) Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G., (2008), "A genetic algorithm-based method for feature sub set selection", *Soft Comput*, form [www.springer.com](http://www.springer.com), pp.1,3,4,5,6.
- 39) Thomas, L. C., (2000), "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers", *International Journal of Forecasting*, Vol.16, pp.2-4.
- 40) Tsang, C. H., Kwong, S., & Wang, H., (2007), "Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection", *Pattern Recognition*, Vol.40, pp.7,10,13.
- 41) WANG, Y. Y., & LI, J., (2008), "Feature-selection ability of the decision-tree algorithm and the impact of feature-selection/extraction on decision-tree results based on hyperspectral data", *International Journal of Remote Sensing*, Vol.22, No.10, pp.4,6,7.
- 42) Xu, X., Zhou, C., & Wang, Z., (2008), "Credit scoring algorithm based on link analysis ranking with support vector machine", *Expert Systems with Applications*, Vol.xxx, pp.6.
- 43) Zhang, Y., & Bhattacharyya, S., "Genetic programming in classifying large-scale data: an ensemble method". *Information Sciences*, Vol.163, pp.2,6.
- 24) Lavrac, N., Gamberger, D., & Turney, P., (1995), "Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm", *Journal of Artificial Intelligence Research*, Vol.2, pp.1,2,4,5.
- 25) Lee, T. S., & Chen, I. F., (2005), "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines", *Expert Systems with Applications*, Vol.28, pp.1,2,5,6,7,8.
- 26) Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F., (2002), "Credit scoring using the hybrid neural discriminant technique", *Expert Systems with Applications*, Vol.23, pp.1,5,6,8.
- 27) Liu, H. H., & Ong, C. S., (2008), "Variable selection in clustering for marketing segmentation using genetic algorithms", *Expert Systems with Applications*, Vol.34, pp.1,3,4,5,6.
- 28) Martinez-Otzeta, J. M., Sierra, B., Lazkano, E., & Astigarraga, A., (2006), "Classifier hierarchy learning by means of genetic algorithms". *Pattern Recognition Letters*, Vol.27, pp.1,2,3,5,6.
- 29) Nanni, L., & Lumini, A., (2009), "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring", *Expert Systems with Applications*, Vol.36, pp.1-4.
- 30) Olson, D., & Shi, Y., (2007), "Introduction to Business Data Mining". Singapore: McGraw Hill Education.
- 31) Ong, C. S., Huang, J. J., & Tzeng, G. H., (2005), "Building credit scoring models using genetic programming", *Expert Systems with Applications*, Vol.29, pp.1-3.
- 32) PAL, N. R., NAND, S., & Kundu, M. K., (1998), "Self-crossover-a new genetic operator and its application to feature selection", *International Journal of Systems Science*, Vol.29, No.2, pp.2,4,5,6.
- 33) Papagelis, A., & Kalles, D., (n.d.), "Breeding Decision Trees Using Evolutionary Techniques", from website [www.siteceer.com](http://www.siteceer.com), pp.1-7.
- 34) 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", s.n., pp.1-7.
- 35) SALAPPA, A., DOUMPOS, M., & ZOPOUNIDIS, C., (2007), "Feature selection algorithms in classification problems: an



## پیوست‌ها

### پیوست ۱- مقادیر ویژگی‌های داده‌های اعتبارسنجی آلمان

مقادیر ویژگی‌ها					نوع ویژگی‌ها	ویژگی‌ها
no checking		0<=X<200		>=200	<0	وضعیت چک
عددی					اسمی	مدت زمان
no credits/all paid		existing paid		delayed previously	critical/other existing credit	سابقه اعتبار
ماشین دست دوم	آموزش مجدد	تعمیرات	راديو تلویزیون	ماشین جدید	مبلمان و تجهیزات	هدف
عددی					اسمی	مقدار اعتبار
عدم پس انداز مشخص		500<=X<1000		100<=X<500	>=1000	وضعیت پس انداز
بیکار		4<=X<7		1<=X<4	>=7	سابقه کار
عددی					اسمی	تعداد اقساط
male single		male mar/wid		male div/sep	female div/dep/mar	وضعیت شخصی و جنسیت
نه				متقاضی مشترک		طرف‌های دیگر
عددی					اسمی	محل اقامت فعلی
املاک و ساختمان		اموال ناشناخته		بیمه زندگی	ماشین	اموال و دارایی‌ها
عددی					اسمی	سن
موجودی		نه		بانک		برنامه‌های پرداختی دیگر
اجاره		مالک		رایگان		مسکن
عددی					اسمی	وضعیت اعتباری موجود
مقیم غیر ماهر		ماهر		unemp/unskilled non res	high qualif/self emp/mgmt	شغل
عددی					اسمی	تعداد عائله
بلی		نه			مالکیت تلفن	
بلی					کارگر خارجی	
خوب		بد			طبقه (کلاس)	

## یادداشت‌ها

- <sup>24</sup> Entropy
- <sup>25</sup> Bushiness
- <sup>26</sup> Training Observations
- <sup>27</sup> Evaluator Function
- <sup>28</sup> Correlation Based Feature Selection
- <sup>29</sup> John Holland
- <sup>30</sup> Threshold
- <sup>31</sup> Evolution
- <sup>32</sup> Confidence Factor

- <sup>1</sup> Credit Scoring
- <sup>2</sup> Fisher
- <sup>3</sup> David Durand
- <sup>4</sup> Recursive
- <sup>5</sup> Greedy
- <sup>6</sup> Quinlan
- <sup>7</sup> Information Gain
- <sup>8</sup> Selected Predictive Attributes
- <sup>9</sup> Correctly Classified Instances
- <sup>10</sup> German Credit Data
- <sup>11</sup> WEKA
- <sup>12</sup> Visualization
- <sup>13</sup> Discretization
- <sup>14</sup> Merge Values
- <sup>15</sup> Turning
- <sup>16</sup> Ball Commitee
- <sup>17</sup> Thomas
- <sup>18</sup> Brill
- <sup>19</sup> Character
- <sup>20</sup> Capital
- <sup>21</sup> Collateral
- <sup>22</sup> Capacity
- <sup>23</sup> Condition