مدل ترکیبی داده کاوی الگوریتم های انتخاب ویژگی و طبقه بندی کننده های یادگیری گروهی برای امتیاز دهی اعتباری

چکیده:

تکنیک های داده کاوی در حوزه بانکداری،از کاربردهای متنوعی جهت امتیاز دهیِ اعتباری برخوردار اند. یکی از معروف ترین تکنیک های داده کاوی،روش طبقه بندی نام دارد.مطالعات پیشین نشان دادند استفاده از الگوریتم های انتخاب ویژگی ها و طبقه بندی کننده های گروهی سبب بهبود عملکرد بانک ها در مشکلات امتیاز دهیِ اعتباری می گردد.در این زمینه،موضوع اساسی،شبیه سازی و استفاده ترکیبی از این دو الگوریتم فوق جهت تعیین پارامترها و دستیابی به مدلی با عملکرد بالا،است . این مقاله به نحوه استفاده از مدل ترکیبیِ داده کاویِ الگوریتم های انتخاب ویژگی و طبقه بندی کننده های یادگیریِ گروهی در امتیاز دهیِ اعتباری در 3 مرحله می پردازد :1)جمع آوری و پردازش داده ها 2)اجراء 4 الگوریتمFS همچون آنالیز مولفه های اساسی PCA،الگوریتم ژنتیکGA، معیار ضریب بهره اطلاعات و تابع ارزیابیِ صفات مهم.در این مرحله،پارامترهای روش FS با کمک الگوریتمِ طبقه بندی کننده ماشین بردار پشتیبان،به دقت تعیین می شوند. پس از انتخاب مدلِ دقیق در هر ویژگی،باید از آن ها در الگوریتم های طبقه بندی کننده و مبناء استفاده نمود. در مرحله فوق،بهترینِ الگوریتمFS همراه با پارامترهای تعیین شده اش جهت استفاده در مرحلهِ مدلسازی،استفاده خواهند شد.3)اجراء الگوریتم های طبقه بندی بر روی مجموعه داده های آماده شده برای هر الگوریتمFS.نتایج مرحله 2،نشان می دهد که الگوریتم PCA نسبت به FS از عملکردِ مطلوب تری برخوردار است نتایج طبقه بندی مرحله 3 نیز نشان می دهد روش آدابوست در شبکه عصبی مصنوعی(ANN) از دقت طبقه بندی بسیار بالایی برخوردار است . در نهایت،مقاله حاضر به بررسیِ مدل ترکیبی به عنوان مدل عملیاتی و قدرتمندی جهت اجراء امتیاز دهیِ اعتباری اشاره خواهد داشت.

کلمات تخصصی:امتیاز دهیِ اعتباری،طبقه بندی،انتخاب ویژگی،یادگیری گروهی،داده کاوی

1- مقدمه:

اخیرا موسسات مالی و بانک ها به بررسیِ ریسک اعتباریِ مشتریان خود پرداخته اند. آن ها جهت تفکیکِ خدمات اعتباری که به مشتریانشان اعطا می کنند نیاز به سیستم های امتیاز دهیِ اعتباری،دارند. اخیرا، استفاده از روش های غیر پارامتری و داده کاوی به حوزهِ امتیاز دهیِ اعتباریِ مشتریان ،راه یافته است.روش های آماری، غیر پارامتری و هوش مصنوعیAI همگی به توسعهِ امتیاز دهیِ اعتباری مشتریان می پردازند. علاوه بر این،روش های امتیاز دهیِ اعتباریِ گروهی در بسیاری از مطالعات استفاده گردیده اند.شایان ذکر است که تعداد قابل توجهی از محققین نشان دادند روش های طبقه بندی یادگیریِ گروهی در حوزهِ امتیاز دهیِ اعتباری در مقایسه با طبقه بندی کننده های مجزا از عملکرد بسیار بهتری برخوردار اند.بر این اساس،9 رویکرد اصلی جهت مطالعه و بررسیِ معیارِ امتیاز دهیِ اعتباری ارائه می کنیم که عبارتند از:

1. مدل های امتیاز دهیِ اعتباریِ طبقه بندی کنندهِ مستقل
2. مدل های امتیاز دهیِ اعتباریِ طبقه بندی کنندهِ چندگانه
3. مدل های امتیاز دهیِ اعتباری مبتنی بر مدل های آماری
4. مدل های امتیاز دهیِ اعتباری مبتنی بر روش های AI
5. مدل های امتیاز دهیِ اعتباریِ خطی و غیر خطی
6. مدل های امتیاز دهیِ اعتباریِ پارامتری، شاملمدل احتمال خطی،؛مدل آنالیز تشخیصی،مدل های لوجیت و غیره
7. مدل های امتیاز دهیِ اعتباریِ غیر پارامتری(داده کاوی) شامل درخت تصمیم گیری،مدل نزدیک ترین همسایهknn،سیستم های خبره،ANN،منطق فازی،GA
8. مدل های امتیاز دهیِ اعتباریِ یادگیریِ گروهی
9. مدل های امتیاز دهیِ اعتباریِ ترکیبی یا هیبریدی

محققان بسیاری به بررسیِ روش های ذکر شده پرداختـه انـد. هـو و آنسـل(2007)از الگـوریتم هـایی همچـون رگرسیون لوجستیکLR،ANN و الگوریتم بهینه سازی حداقلی ترتیبیSMO در مطالعـات خـود،بهره بردنـد. مین و لی(2008) نیز به پیش بینیِ مدل امتیاز دهیِ اعتبـاری بـر اسـاس روش تحلیـل توسـعه داده هـاDEA پرداختند.در مطالعهِ دیگر،الگوریتم‌های رتبه‌دهی مبتنی بر تحلیل پیوند صفحات همراه بـا SVM جهـت امتیـاز دهیِ اعتباری،استفاده شد . ستینو و همکاران(2009) از GA جهـت بهینـه سـازی الگـوریتمKNN در زمینـهِ امتیاز دهیِ اعتباری استفاده کردند. علاوه بر این،یاه و لین(2009) به مقایسـه تکنیـک هـای داده کـاوی نظیـر KNN,IR,ANN و درخت تصمیم گیری پرداختند. ژو و همکاران(2009) از روشِ جستجوی مسـتقیم بـرای انتخاب پارامترها در الگوریتمSVM استفاده کردند. پینگ و یانگ ژن(2011) نیز از مجموعه پارامترهای مجاور و طبقه بندی کنندهSVM جهت بررسیِ امتیاز دهیِ اعتباری،اسـتفاده بـه عمـل آوردنـد. در مطالعـه دیگر،کـائو و همکاران(2012) از مدل بیزین همـراه بـا درخـت رگرسـیونِ طبقـه بنـدی کننـده اسـتفاده کردنـد. وکـوویچ و همکاران(2012) نیز از توابع اولویت همراه با مدل اسـتدلال مبتنـی بـر حالـت(CBR) در زمینـهِ امتیـاز دهـیِ اعتباری،استفاده کردند. دانِناس و گارسوا(2015) از روشِ بهینه سازی ازدحام ذراتPSO جهت انتخاب بهترین طبقه بندی کننده خطیSVM در امتیاز دهیِ اعتباری،استفاده نمودند.

همانطور که گفتیم،مدل های امتیاز دهـیِ اعتبـاریِ گروهـی در مطالعـات فراوانـی اسـتفاده شـده انـد. تِسـای و وو(2008)از شبکه عصبی پرسپترون چندلایهMLP در زمینهِ امتیاز دهیِ اعتباری و .نـانی و لـومینی(2009) از روش هایی نظیر جنگل تصادفی،فضای تصادفی و تغییر کلاس در امتیاز دهـیِ اعتبـاری، تـوالا (2010) از روش هایی همچون ANN،درخت تصمیم گیری،KNN و تحلیل تشخیص لجستیک،سیه و هانـگ(2010) از طبقـه بندی کننده های گروهیِ *Bagging* نظیر ANN,SVM و شبکه بیزین،پالئولوگو و همکاران(2010) از طبقـه بندی کننده های subagging همچون SVM,KNN،درخت تصمیم گیری و آدابوست در امتیاز دهیِ اعتباری و وانگ و ما(2012) نیز از روش یادگیریِ ترکیبی با استفاده از SVM به عنـوان یادگیرنـدهِ اصـلی در ارزیـابیِ ریسک اعتباراتِ شرکت ها،استفاده کردند.

مطالعات مختلفی به بررسی استفاده از روش FS در مدل های امتیاز دهیِ اعتباری پرداخته اند. وانگ و هانگ(2009) از روش های پویا جهت جمع آوری داده هایی در زمینهِ امتیاز دهیِ اعتباری استفاده کردند.تسای(2009) به مقایسهِ 5 روش FSدر پیش بینیِ ورشکستگیِ شرکت ها (نظیر آزمون تی،ماتریس همبستگی،رگرسیون گام به گام،PCA و تحلیل عامل)و بررسیِ عملکرد آن ها با کمک شبکه های عصبیِMLP پرداختند. چن و لی(2010)،استراتژی هایِ ترکیبی با روش های FS همچون آنالیز تشخیص خطیLDA، نظریه مجموعه مبنا،درخت تصمیم گیری و مدل طبقه بندی SVM در زمینهِ امتیاز دهیِ اعتباری را مطرح کردند. چن(2012) ،بر اساس نظریه مجموعه مبنا، به ترکیب FS با توزیع احتمال تجمعی در حوزهِ امتیاز دهیِ اعتباری پرداخت.هاجک و میشالاک(2013)،روشی را جهت ترکیب روش های فردی و گروهیِ FS همراه با مدل های یادگیری ماشینی نظیر MLP,RBS,SVM,LDA و طبقه بندی کنندهِ نزدیک ترین میانه در پیش بینیِ امتیاز دهیِ اعتباریِ شرکت ها ارائه نمودند. اورسکی (2014) نیز به ترکیبِ GA با ANN را جهت تشخیص و شناسایی بهترین ویژگی ها و افزایش دقت طبقه بندی در ارزیابی ریسک اعتبارات،پرداخت. لیانگ و همکاران(2015) از 3 فیلتر با عنوان ,LDA،آزمون تی و رگرسیون خطی همراه با روش های FS مبتنی بر GA,PSO در ترکیب با 6 مدل پیش بینی کننده متفاوت با عناوین SVM خطی،RBF,SVM,KNN,CART و بیز جهت گردآوری داده های مربوط به ورشکستگی و امتیاز دهیِ اعتباریِ شرکت ها،استفاده کردند.

مطالعات فوق همگی در سه دیدگاه اصلیِ انجام شده اند:1)ارزیابیِ کلیِ امتیاز دهیِ اعتباری 2)ارزیابیِ امتیاز دهیِ اعتباریِ گروهی و 3)ارزیابیِ امتیاز دهیِ اعتباری مبتنی بر FS.این مقاله سعی کرده از این سه دیدگاه فاصله بگیرد.البته باید توجه داشت که مطالعات پیشین تنها یکم یا دو دیدگاه از سه دیدگاه فوق را در اختیارداشته اند. علاوه بر این،هدف اصلیِ مقاله حاضر،استفاده از الگوریتمFS و طبقه بندی کننده های مبنا و گروهی و توجه به عواملی نظیر دقت طبقه بندی،منحنی ویژگی های عملیاتی AUC و تعیین پارامترهای در مدل های امتیاز دهیِ اعتباری، می باشد . همچنین بسیاری از مطالعات،تاثیر روش های مختلفFS و پارامترهای تعیین شدهِ طبقه بندی کننده ها بر مشکلاتِ امتیاز دهیِ اعتباری را نادیده گرفته اند. بنابراین به منظور ایجاد مدل ترکیبیِ جدیدFS و مدلِ گروهیِ امتیاز دهیِ اعتباری،نه رویکرد مطرح شده اند. مدل پیشنهادیِ ما ،ترکیبی از تکنیک

های FS همراه با طبقه بندی کننده های مبنا(مستقل) و گروهی در روش های پارامتری و غیر پارامتریِ امتیـاز دهیِ اعتباری،می باشد . در تعیین پارامترها از 4 الگوریتم FS و دو الگوریتم طبقه بندی کننده،استفاده به عمـل آمده است . عملکـرد هرالگـوریتمFS، بـر اسـاس شـاخص دقـتِ طبقـه بنـدی SVM سـنجیده شـده اسـت. SVM،روش یادگیریِ موثری بوده که در مسائل طبقه بندی،استفاده می شود. SVM یکی از تکنیک های رایج استفاده شده در ادبیات محسوب می شود.بنابراین می توان از آن جهت ارزیابیِ عملکرد الگوریتم های fs استفاده نمود. همچنین،مطابق با دقت طبقه بندی و شاخص AUC،الگوریتم های طبقه بندی با یکـدیگر مقایسـه شـده اند.در این مقاله،مجموعه داده هایی از بانک توسعه صادراتِ ایـران بـه دسـت آمـده انـد. در مـدل ترکیبـی، از 4 الگوریتم FS استفاده شده که عبارتند از:1)PCA 2) GA 3) معیار ضریب بهره اطلاعات و 4)الگـوریتم Relief .علاوه بر این،دو الگوریتم طبقه بندیِ رایج در مطالعات پیشین با عنوان 1) طبقه بندی کننده های مستقل نظیر بیزین،درخـت تصـمیم گیـری CART،SVM و ANN و 2)الگـوریتم هـای گروهـی نظیـر bagging، آدابوست،درخت تصادفی و staking ،نیز مورد استفاده قرار گرفته اند. نتایج نشان می دهند مدل ترکیبیِ امتیـاز دهیِ اعتباری در مقایسه با سایر الگوریتم های دیگر،از عملکرد مطلوب تری برخوردار است .

ویژگی هایِ مهم این مقاله که از مدلِ پیشنهادی به دست آمده عبارتند از:

1. انجام مطالعهِ جامع و کامل با مقایسهِ طبقه بندی کننده ها و روش هایFS متناسـب بـا مسـائلِ امتیـاز دهیِ اعتباری
2. استفاده ترکیبی و همزمان از 3 روشِ امتیاز دهیِ اعتباریِ گروهی،کلی وFS
3. استفاده از الگوریتم های FS و مقایسه عملکرد آن ها با هدف تشخیصِ کارآییِ اگوریتمSVM و سنجش دقت AUC
4. تعیین پارامترهای الگوریتم های طبقه بندی وFS جهت بهبود عملکردِ امتیاز دهیِ اعتباری
5. استفاده همزمان و مقایسهِ الگوریتم های یادگیریِ مستقل و گروهی در مدلِ امتیاز دهیِ اعتباری
6. استفاده و مقایسهِ 9 روشِ مدلبندیِ امتیاز دهیِ اعتباری بر اساس چهارچوبِ ترکیبی

7. اگرچه امتیاز دهیِ اعتباری بر روی مشتریان واقعی بررسی شده اما مدل امتیـاز دهـیِ اعتبـاریِ مـا بـر اساس مشتریان حقیقی، ساخته شده است.

سایر قسمت های دیگر این مقاله عبارتند از:بخش2: شرح مختصری از روش ها .بخش3: ارائه طرح آزمایش نظیـر شرح مجموعه داده ها و پردازش آن ها،ارزیابی عملکرد و توسعهِ مدل امتیاز دهیِ اعتبـاری.بخش 4:بیـان نتـایج و بخش 5: نتیجه گیری و ارائه پیشنهادات

2- پیش زمینه

2.1 امتیاز دهیِ اعتباری

توماس،امتیاز دهیِ اعتباری را به عنوان فرآیند شتاساییِ مشتریان بانک ها جهت اعطای تسهیلات و اعتبـارات بـه آن ها بر اساس برخی معیارهای از پیش تعیین شده، تعریف نمود. در واقع،مطابق با گفته اونـگ و همکاران،مـدل هایِ امتیاز دهیِ اعتباری از مزیت های متعددی برخوردار اند ازجملـه:1)کـاهش هزینـه هـای تحلیـل اعتبـارات 2)تخصیص اعتبارات بر اساس تصمیم گیری های سریع و به موقع 3)سوددهی بالا در زمان بازپرداخت تسهیلات و 4) کاهش خطرات احتمالی.در سال 1936 میلادی،فیشر با کمـکِ مـدل امتیـاز دهـیِ اعتبـاری،مفهوم آنـالیز تشخیص آماری را بیان نمود. پس از وی،دیوید دوراند در سال 1941،از روش های متعـددی جهـت تفکیـک وام های بد از خوب استفاده کرد . سپس در سال 1960،موسسات و بانک ها از کارت های اعتباری جهت انجام امتیاز دهیِ اعتباری استفاده کردند. در سال 1980،کارشناسان بانکداری،استفاده از روشِ امتیاز دهیِ اعتباری را نقطه عطفی جهت استفاده از سایر روش های دیگر،عنوان کردند. بنابراین امتیاز دهیِ اعتباری از سوی این افراد جهـت بررسیِ تصمیم گیری نهـایی در بانـک هـا اسـتفاده شـد.محققان در ارزیـابیِ امتیـاز دهـیِ اعتبـاریِ مشـتریان از 3C,4C,5C (همچـون سرمایه،شخصـیت،توانایی و وضـعیت مشـتری) بهـره بردنـد.یو و همکـاران(2009) در مطالعـات خـود بـه بررسـیِ روش هـای بهینـه سـازی و آمـاریِ امتیـاز دهـیِ اعتبـاری نظیـر LDA، تحلیـل لجستیک،تحلیل پروبیک،برنامه ریزی خطی،برنامه ریزی ترکیبی،KNN و درخت طبقـه بندی،پرداختنـد. اخیـرا برخی مطالعات به بررسیِ تکنیک های AL همانند ANN,EC,GA,SVM در حوزهِ امتیاز دهیِ اعتباری،توجه

داشته اند.علاوه بر این،مدل هایِ امتیاز دهیِ اعتباریِ مستقل و ترکیبی نیز از سوی برخی محققان استفاده شده است.

2.2 الگوریتم های انتخاب ویژگی:

الگوریتم های انتخاب ویژگی ها که به روش های پیش پردازشِ مدل نیز معروف اند،جهت افزایش عملکرد طبقه بندی،استفاده می شوند.مزیت های این روش ها عبارتند از:1- کاهش نویز در مجموعه داده ها2-کاهش هزینه های محاسباتی جهت آنایی با مدل ها 3- کمک به درک بهتری از مدل های نهایی در الگوریتم های طبقه بندی 4- استفاده آسان و 5- به روز رسانیِ مدل.در روش FS،سه شاخص عمده وجود دارد:1) شاخص ارزشیابی 2) جستجوی رفتار و 3) قانون توقف.نوع نوع معیار ارزشیابی نیز در این مدل وجود دارد که عبارتند از:1)اطلاعات 2) وابستگی 3)فاصله 4) ثبات و 5) دقت طبقه بندی.همچنین 3 نوع روش تحقیق نیز در FS موجود بوده که عبارت است از:1- جستجوی تصادفی2- اکتشافی و 3- جستجویِ کامل.همچنین قوانین توقف که از طرف وانگ و لی ارائه شده دارای ویژگی های ذیل می باند: 1) تعیین ماکزیمم تعداد تکرار 2- عدم تغییر عملکرد بدون اضافه کردن یا حذف یک ویژگی 3- تعیین مجموعه ویژگی های ایده آل.

در این مقاله،3 نوع الگوریتمFS استفاده شده اند:1-GA 2- معیار ضریب بهره اطلاعات و 3- روش relief .در ادامه ،شرحی از الگوریتم های موجود ذکر خواهند شد:

1. انتخاب ویژگیِ الگوریتم ژنتیک:در این روش،ابتدا کروموزومی به عنوان مجموعه ای از ویژگی مشتریان بانک،انتخاب می شود. ژن نیز به عنوان یکی از ویژگی های مشتریان انتخاب شده است.جهت شناساییِ بهترین مجموعه ایده آلی از متغیرها،از استراتژیِ گلدبرگ استفاده گردیده است . درادامه به منظور ارزیابی متغیرهای ورودی،از تابع ارزیاب با اعتبارn استفاده شد.علاوه بر این،زیرمجموعه ای از ویژگی ها نیز مطابق با معیار دقت طبقه بندیِ الگوریتمSVM انتخاب گردید.در نهایت،ماکزیمم تعداد نسل ها،جمعیت اولیه،تعداد جهش ها ،احتمال مقطعی ،اعتبار سنجی مقطعی و تعداد بذرهای تصادفی به ترتیب 20، 01،20/،9/،10 و 1 به دست آمدند.

2. انتخاب ویژگیِ روش Relief: این روش با کمک نرم افزار Rapid Miner می تواند مقدار هر ویژگی و نیز مقدار نزدیک ترین همسایه به آن را ارزیابی نماید.
3. انتخاب معیار ضریب بهره اطلاعات:این روش مبتنی بر مفهوم آنتروپیِ اطلاعات است. در این روش،از ابزار یادگیری ماشینیِWEKAجهت تعیین معیار ضریب بهره اطلاعات متناسب با هرگروه،استفاده شده و مقدار معیار ضریب بهره اطلاعات معادل با H منهای H می باشد .
4. انتخاب ویژگیِ تحلیل مولفه اساسی: PCA ،روشی انتقالی جهت کاهش تعداد ویژگی ها با استخراج ویژگی های جدید بوده و در آن،ویژگی های مربوطه به عنوان مولفه های اساسی انتخاب شده اند. مطابق با یافته ایِ سوسترسیک و همکاران(2009)،مولفه های اساسیِ متعدد جود داردکه به عنوان متغیرهای اصلیِ مدل انتخاب شده اند.

2.3 الگوریتم های طبقه بندیِ مبنا(مستقل):

1- درخت تصمیم گیری:این الگوریتم را می توان یکی از مشهورترین الگوریتم های استفاده شده در زمینهِ امتیاز دهیِ اعتباری نامید درخت تصمیم گیری،مدلی با ساختار درختی بوده که از نقاط،شاخه ها و برگ های فراوانی تشکیل شده است .هر نقطه به یک متغیر یا صفت تعلق دارد به بیان بهتر،شاخه ها ،داده ها را به مجموعه داده های کوچک تری تقسیم نموده و برگ ها نیز از مقدار مشخصی بهره مند اند. در این مقاله از طبقه بندی کنندهِ درخت تصمیم گیریِ CART که از دو شاخه برای هر نقطه تصمیم گیری تشکیل شده،استفاده شده است . به منظور تقسیم داده های آموزشی،CART ،تمام تعاملات را به زیرمجموعه ای از تعاملات با مقادیر مشابهی برای متغیرهای هدف، طبقه بندی می کند.CART ،یک جستجویِ جامع نیز بر روی تمام صفات انجام داده و از تمام مقادیر به دست آمده جهت دستیابی به بهترین شاخصِ تقسیم بندی شده، استفاده می کند.

2- شبکه عصبی مصنوعی:این الگوریتم طبقه بندی بر اساس روش های غیر پارامتری ارائه شده و به وفور در مسائلِ امتیاز دهیِ اعتباری،مورد استفاده قرار می گیرد .ANN می تواند در مسائل غیر خطی نیز بکار برده شود. در مطالعه حاضر،از ANN دارای 3 لایه استفاده شده که لایه درونی شامل نورون های مربوط به متغیرهای ورودی و لایه خارجی دارای یک نورون می باشد . ارتباط بین نورون های موجود درهر لایه به وزن آن ها بستگی

دارد . هر نورون در لایه های مخفی و خروجی با کمک تابعِ فعالسازی،تحریک می شود. در طی مرحلـه آمـوزش شبکه،وزن لایه ها کاهش یافته اگرچه دقت طبقه بندی افزایش می یابـد.رویکردِ آموزشـی،نوعی فرآینـد تکـراری بوده که بر اساس یادگیریِ گرادیان نزولی و با کمک شاخص ضریب یادگیری،محاسبه می شود.

3- ماشین بردار پشتیبان:SVM ابتدا توسط واپنیک،توسعه یافت می توان آن را نوعی الگوریتم یادگیریِ نظارتی و غیر پارامتری در نظر گرفت . اخیرا،SVM در مطالعاتِ مختلفی در حوزهِ امتیاز دهـیِ اعتبـاری اسـتفاده شـده است . مطابق با مطالعات پینگ و یانچنگ(2011)،رویکردِ اصلی SVM ،مینیمم سازیِ ریسک ساختاری با کمک معادلهw.x+b=0 است . در این مطالعه،کرنل های خطی،چند جمله ای وRBF جهت بهینه سـازی ابرصـفحات استفاده شده اند.

4- روش بیـز:این الگـوریتم مبتنـی بـر روش هـای پـارامتری و یـادگیری احتمـالات اسـت . مطـابق بـا گفتـه توالا(2010)،این طبقه بندی کننده از قوانین بیز جهت محاسبه احتمال دارا بودن تمام صفاتAj توسـط Ci و پیش بینیِ بیشترین احتمال یک کلاس،استفاده می کند.احتمال مقدار کلاس Ci در n مشاهدات با کمک معادله ذیل به دست می آید:

$$p(C_i|X) = \prod_{j=1}^{n} p(A_j|C_i) . \quad p(C_i) \qquad (1)$$

2.4: الگوریتم های طبقه بندی گروهی:

الگوریتم های طبقه بندی گروهی در حوزهِ امتیاز دهیِ اعتباری از کاربردهای گسترده ای برخوردار اند. یـادگیری گروهی بر اساس روش های یادگیری ماشینی بنا شده و اکثرالگوریتم های یادگیری جهت حل مسائلِامتیاز دهیِ اعتباری بکار برده می شوند. این الگوریتم ها متضاد الگوریتم های مستقل هستند.به تازگی،الگوریتم های گروهـی در مطالعاتِ امتیاز دهیِ اعتباری،به وفور استفاده می شوند. در بخش 1،برخی مـدل هـای امتیـاز دهـیِ اعتبـاریِ گروهی مطرح شد . در ادامه،4 مدل اصلیِ یادگیری گروهی بیان خواهند شد:

1. Bagging: این الگوریتم از سوی بریمن توسعه یافته و مبتنی بر مفاهیم مختلفی است . که در آن،مجموعه داده های آموزشیِ متنوع به طور تصادفی جهت آموزش به فرد یادگیرنده استفاده می شود.
2. آادبوست:این الگوریتم توسط فرند و و شافیر توسعه یافته و به رویکرد بوستینگ تعلق دارد .این الگوریتم به سنجشِ تمام تعاملات و تکرارهای آن ها می پردازد .
3. Stacking :این الگوریتم با ترکیب در الگوریتم های یادگیریِ مختلف جهت دستیابی به دقت پیش بینیِ بالا،استفاده شده و مطابق با مطالعات وانگ و همکاران(Stacking(2011 ،یا کمک یادگیرندگان در سطح متا،به پیش بینیِ فعالیت یادگیرندگان متعدد می پردازد .
4. جنگل تصادفی:این روش، مجموعه ای از درخت های تصمیم گیریِ منسجم بوده که مطابق با آزمون های سنجشِ داده های آموزشی،مورد استفاده قرار می گیرد .

3- طرح آزمایش:

3.1: تشریح مجموعه داده ها و پیش پردازشِ آن ها:

مجموعه داده های استفاده شده جهت ارزیابی عملکردِ مدل پیشنهادی به مشتریان حقیقیِ بانک توسعه صادراتِ ایران طی دو سال،وابسته اند.در ادامه،متغیرهای بکار برده شده در این مدل ارائه خواهند شد. نوع متغیرها یا پیوسته (C) بوده یا ناپیوسته (D) ،بنابراین داریم:

برگشت فروش(C) ،مجموع نمرات کیفیت(C) ، چرخه مطالبات(C) ،عملیات سیکل(C) ،موجودی دوره(C) ،ریسک بازارهای هدف(D)،تجارب بانک(D)،سابقه ورشکستگی(D)،آمار مدیران ارشد(D)، شخصیت حقوقی(D)،عوامل فصلی(D)،فعالیت در بازار داخلی(D) ،قلمرو بازار خارجی(C) ،سرمایه در گردش(C) ، جریان دارایی های جاری(C) ، دارایی های غیر جاری(C) ،نرخ بهره با سرمایه ثابت(C) ،نرخ بهره سرمایه گذاری(C) ، نسبت بدهی ها به دارایی های غیر جاری(C) ،میزان پوشش هزینه های مالی(C) ،نسبت جریان(C) ، نسبت دارایی های خالص به دارایی های غیر جاری(C) ،نسبت آنی(C) ،نسبت بدهی(C) و نسبت دارایی(متغیر هدف) (C)

در این مجموعه داده ها،متغیر هدف،یک مساله دو کلاسی بوده که بدین صورت تعریف شده است: مشتریان بد و خوب،کسانی اند که بازپرداخت های آن ها قبل و بعد از دو ماه صورت گرفته باشد .این مجموعه داده ها متشکل از 1100 مشتری حقوقی با تعداد 59 ویژگی،است .

جهت استفاده از این مجموعه،ابتدا باید عمل پردازش داده ها صورت گیرد.در این مقاله،جهت آماده سازی داده ها،روش های پیش پردازش مختلفی به صورت زیر استفاده شده است:1) حذف برخی ویژگی ها و آمار ثبت شدهِ نامعتبر 2) ترکیب داده ها 3)انتقال و تبدیل داده ها 4) حذف برخی از ویژگی های اضافی 5) نرمال سازی 6) تعیین همبستگی بین دو متغیر با کمک آزمون پیرسون 7) تجسم داده ها 8) ایجاد ویژگی جدید و 9)تعیین طرح اصلی.

بعد از انجام پیش پردازش، تعداد 30 ویژگی انتخاب شدند. علاوه بر این،تعداد رکوردهای ثبت شده بعد از آماده سازی داده ها نیز به 777 مورد کاهش یافته است .

3.2 ارزیابی عملکرد:

جهت ارزیابی آزمایشات،برخی شاخص ها از جمله دقت طبقه بندی و AUC مورد استفاده قرار گرفتند. شرح هر یک از این شاخص ها باید بر حسبِ ماتریس درهم ریختگی و مطابق با شکل 1 ،انجام شود. موارد اختصار در این ماتریس عبارتند از:TP=مثبت حقیقی،TN=منفی حقیقی FP= مثبت غیر حقیقی وFN= منفی کاذب و غیر حقیقی. در شکل 1،تعریف هر یک از این اختصارات نیز ذکر شده است.قت طبقه بندی را می توان با کمک معادله ذیل محاسبه نمود.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \qquad (2)$$

دومین شاخص ارزیابی،AUC است براون و مائوس(2012) بیان کردند که منحنی ویژگی های عملیاتیِ گیرندهROC یک نمودار دو بعدی بوده و می تواند تعاملات بین مثبتِ حقیقی و مثبت کاذب را نشان دهد. به منظور مقایسه طبقه بندی کننده های مختلف با یکدیگر،می توان از AUC استفاده نمود.

Predicted classification	Actual classification: Positive (Good customers)	Actual classification: Negative (Bad customers)
Positive (Good customers)	True Positive (TP)	False Positive (FP)
Negative (Bad customers)	False Negative (FN)	True Negative (TN)

شکل1:ماتریس درهم ریختگی

3.3 مدل پیشنهادی:

در این بخش،از مدل داده کاویِ ترکیبی در امتیاز دهیِ اعتباری استفاده شده است.شکل 2،نمودار مدل پیشنهادی را نشان می دهد.این مدل از سه مرحله تشکیل شده است:1)جمع آوری داده ها و پیش پردازش آن ها 2)انتخاب ویژگی ها و 3)مدلسازی یا همان طبقه بندی.درمرحله نخست،بعد از جمع آوری داده ها،از روش های پیش پردازش نیز استفاده شده است . در مرحله دوم،جهت دستیابی به بیشترین عملکرد الگوریتم های طبقه بندی در مرحله بعد،از 4 الگوریتمFS استفاده شده است.این4 الگوریتم عبارتند از:1)GA 2)روش Relief 3) معیار ضریب بهره اطلاعات و 4)PCA.در این مرحله، پارامترهای تمام تکنیک های FS ،تعیین شده است . پارامترهای بهینه نیز توسط شاخص دقت طبقه بندیِ SVM برای هر الگوریتم فوق تعیین شده است.بر این اساس،70 و 30 درصد از داده ها به ترتیب به عنوان داده هایآموزشی و آزمایشی انتخاب شدند. شاخص دقت نیز جهت مقایسه الگوریتم هایFS استفاده شدند. در مرحله 3،از 4 طبقه بندی کنندهِ مستقل استفاده شده که عبارتند از:CART,ANN,SVM و بیز و تعدادی طبقه بندی گروهی نظیر بیز-آدابوست،CART-آدابوست،ANN – آدابوست،بیز-bagging، نظیر بیز-bagging ، bagging -CART، bagging – ANN .

مجموعه داده های آموزشی(90 درصد) و آزمایشی(10درصد) نیز جهت ساخت مدل هایِ امتیاز دهیِ اعتباریِ گروهی و مستقل استفاده شده و جهت ارزیابی الگوریتم های طبقه بندیِ فوق،از دو شاخصِ 1)دقت طبقه بندی و 2)AUC استفاده به عمل آمد. با کمک این شاخص ها،الگوریتم های طبقه بندی در امتیاز دهیِ اعتباریِ مشتریان بانک استفاده شد.در مدل های ضعیف،تعیین پارامترها در مرحله مدلسازی باید به دقت انجام شود.

همانطور که در بخش مقدمه بیان شد،مطالعات متعددی به استفاده از استراتژی های ترکیبی و گروهی در مسائلِ امتیاز دهیِ اعتباری پرداخته اند. آن ها از طبقه بندی کنند هـای مختلفـی جهـت حـل مشـکلاتِ امتیـاز دهـیِ اعتباری بهره می برند. از الگوریتم های FS جهت دستیابی به عملکرد بالایِ مدل های طبقه بندیِ امتیـاز دهـیِ اعتباری استفاده شده با این حال،این مطالعات نیاز به رویکردیِ جامع برای شـبیه سـازی الگـوریتم هـایFS در چهارچوب داده کاویِ ترکیبی ،دارند.در این رابطه،واضح است که یکی از ویژگی های بارز ایـن چهـارچوب،تعیین پارامترهای FS و مدلسازیِ مدل پیشنهادی می باشد . علاوه بر این،بـه خـاطر اسـتفاده از مجموعـه داده هـای مختلف در مسائلِ امتیاز دهیِ اعتباری، تنها برخی از الگوریتم های FS با مجموعـه داده هـای موجود،همخـوانی دارند. علاوه بر این،می توان از الگوریتم های طبقه بندیِ متعددی جهت بررسی مسائلِ امتیـاز دهـیِ اعتباری،بـه شکل صحیح استفاده نمود. مدل پیشنهادی ما،می تواند مشکلات فوق را با درنظر گـرفتن نقـاط قـوت مطالعـات پیشین،به حداقل برساند.

4- نتایج و بحث و گفتگو:

در این بخش،مجموعه داده های بانک توسعه صادرات ایران، جهت ارزیابی مدل استفاده شده اسـت. از نـرم افـزار داده کاویِ Rapid Miner نیز جهت بررسی مدل استفاده شده است.در این نرم افزار،مقادیر مربوط به هرپارامتر در الگوریتم های FS به صورت پیش فرض انتخاب شده است .

در مدل پیشنهادی،بعد از گردآوری داده ها و پیش پردازش آن ها در مرحله نخست،از الگوریتم های FS جهـت تعیین پارامترها و دقت طبقه بندی استفاده شده در ادامه،نتایج این مرحله شرح داده خواهند شد.

4.1 الگوریتم ژنتیکی و تعیین پارامترها:

در الگوریتم ژنتیکِ FS،پارامتر"اندازه جمعیت" تغییر یافته است.ماکزیمم تعـداد مقـادیر نسـل هـا،جهش هـا و تبادلات به ترتیب 30،01 و 5/ می باشند. از خروجی FS نیز در مدل طبقه بندی SVM استفاده می شـود. در جدول 1 نتایجِ فوق ارائه شده اند.

با توجه به این جدول،ویژگی های انتخاب شده از پارامتر اندازه جمعیت در طبقه بنـدی SVM ،مـورد اسـتفاده قرار گرفته براین اساس،مقدار پارامتر فوق دارای دقت عملکرد مطلوبی بوده است . در جدول 1،دقت طبقه بندی و مقادیرAUC در مدل 1 به ترتیب 88.41 و 90.45درصد بوده اند از این رو،بهتر است از ویژگی های انتخابی فوق در طبقه بندی مستقل و گروهی استفاده نمود.ویژگی های انتخابیِ به دست آمـده توسـط GA عبارتنـد از: نسبت دارایی های خالص به بدهی های غیـر جـاری،نرخ بهـره سـرمایه گـذاری،تخفیف سهام،ریسـک بازارهـای هدف،عوامل فصلی،تاریخچه شرکت/ورشکستگی، آمار ثبت شدهِ مدیران ارشد و نمره کیفیِ کل.

4.2 تعیین پارامترها و روش Relief:

در روش Relief،تعدادی ازویژگی های انتخابی و نزدیک ترین همسایه ها تغییر یافته اند. از این ویژگی هـا مـی توان جهت طبقه بندی SVM استفاده نمود. نتایج طبقه بندی در جدول 2 ذکر شده است.

شکل2:نمودار کلیِ مدل پیشنهادی

جدول1: تعیین پارامترها در الگوریتم FS مبتنی بر روش GA

Model	Population size	Accuracy (%)	AUC (%)
1	5	88.41	90.45
2	10	87.70	90.17
3	15	88.41	90.13

جدول2: تعیین پارامترها در FS مبتنی بر روش Relief:

مدل	تعداد ویژگی های انتخابی	تعداد نزدیک ترین همسایه	دقت(٪)	AUC (%)
1	10	5	21.46	50.00
2	10	10	78.54	50.00
3	10	15	78.54	50.00
4	10	20	78.54	50.00
5	10	25	78.54	50.00
6	15	5	78.54	50.00
7	15	10	78.54	50.00
8	15	15	78.54	50.00
9	15	20	78.54	52.00
10	15	25	78.54	52.00
11	20	5	78.54	50.00
12	20	10	78.54	50.00
13	20	15	78.54	50.00
14	20	20	78.54	50.00
15	20	25	78.54	50.00

جدول3: تعیین تعیین پارامترها در FS مبتنی بر روش معیار ضریب بهره اطلاعات:

مدل	تعداد ویژگی های انتخابی	دقت(٪)	AUC (%)
1	10	90.99	86.27
2	13	77.25	55.81
3	16	76.39	50.00
4	19	78.97	58.71
5	22	80.69	52.22
6	25	78.11	50.00

با توجه به جدول 2،مدل های 9 و 10 از عملکرد بالاتری در طبقه بندی ,SVM برخوردار اند.مقادیر دقت و شاخص AUC نیز به ترتیب 78.54 و 52درصد بوده اند. بنابراین،ویژگیِ مدل های 9 و 10 در ساخت مدل های طبقه بندی کننده می توانند مورد استفاده قرار گیرند. جدول 2 نشان می دهد که مقدارAUC در هر مدل (به استثناء 9و10) مساوی با یکدیگر بوده است . بعد از اجراء روش Relief،ویژگیِ های مهمِ انتخابی بدین صورت به دست آمده اند:شخصیت حقوقی،نمره کیفیت کل، عقاید کارشناسان اعتباری،نسبت دارایی ها،نسبت بدهی ها،نسبت دارایی های خالص به بدهی ها،نسبت دارایی خالص به دارایی غیرجاری،ضریب بهره سرمایه گذاری،سرمایه عملیاتی/سرمایه در گردش،عوامل فصلی،ریسک بازارهای هدف و تجربه کارکنان بانک.

جدول4: تعیین پارامترها در FS مبتنی بر روش PCA:

Model	Number of factor	Accuracy (%)	AUC (%)
1	10	87.12	83.73
2	20	87.12	84.26
3	30	87.12	84.14
4	40	87.12	84.17
5	50	87.12	84.10

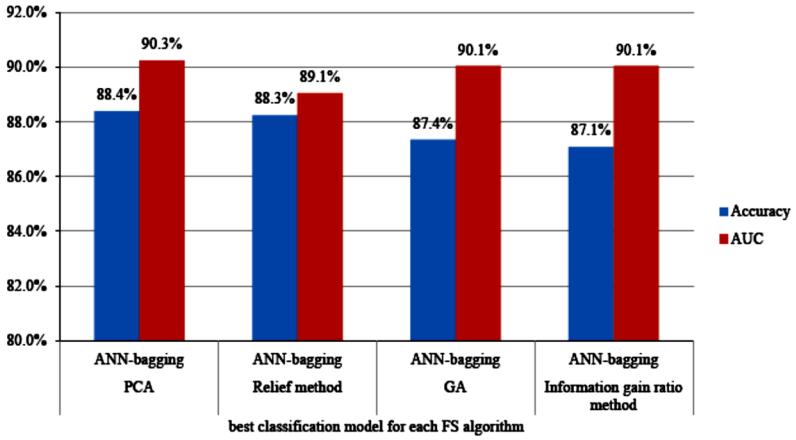
4.3 روش معیار ضریب بهره اطلاعات و تعیین پارامترها:

در روش معیار ضریب بهره اطلاعاتFS،تنها تعدادی از ویژگی ها تغییر می کنند.از این ویژگی ها نیـز در طبقـه بندی SVM استفاده می شود. نتایج در جدول 3 موجوداند.

مطابق با این جدول،مدل 1 از عملکرد بالایی در طبقه بندی SVM برخوردار است .مقـادیر مربـوط بـه دقـت و شاخص AUC به ترتیب 90.99 و 86.27درصد بوده است . در ادامه،ویژگی های مدل 1 نیز جهت ساخت مدل طبقه بندی استفاده می شوند.با استفاده از روش معیار ضریب بهره اطلاعات،ویژگی هـای ذیـل بـه دسـت آمـده اند:نسبت بدهی ها،نسبت دارایی ها، نمره کیفیت کل ،شخصـیت حقـوقی، ، عقایـد کارشناسـان اعتباری،نسـبت دارایی ها،نسبت دارایی های خالص به بدهی ها،نسبت بدهی ها،نسبت دارایی خالص به دارایـی غیرجاری،ضـریب بهره سرمایه گذاری وسرمایه عملیاتی/سرمایه در گردش.

4.4 تحلیل مولفه اساسی و تعیین پارامترها:

در روش کاهش ویژگی های PCA،تعداد عوامل تغییر یافته اند. ویژگی های انتخابیِ به دست آمده نیز در طبقه بندی SVM استفاده می شوند. جدول 4،نتایج مربوطه را نشان می دهد.



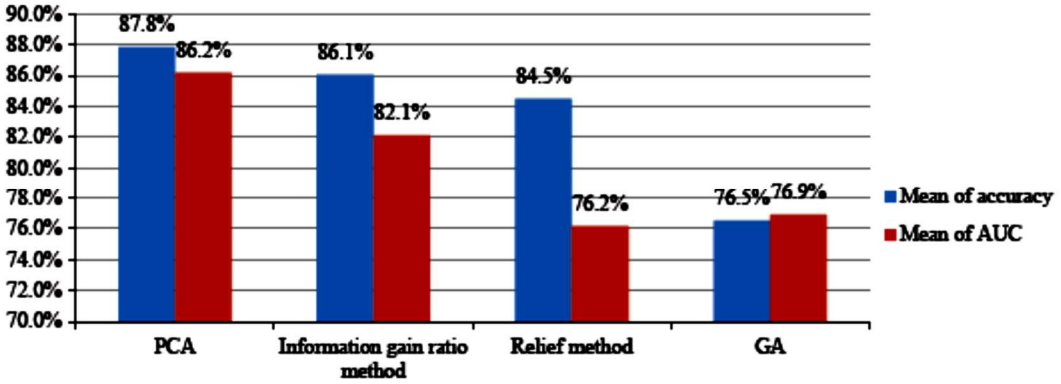
شکل3: نتایج الگوریتمFS

مطابق با جدول 4،مدل 2 دارای عملکرد بالایی در طبقه بندی SVM است . مقادیر مربوط به دقت و شاخص AUC به ترتیب 87.12 و 84.26درصد بوده بنابراین ویژگی های انتخابیِ مدل 2 در مدل طبقه بندی مورد استفاده قرار می گیرند. جدول 4 نشان می دد که مقدار دقت در هر مدل با دیگری یکی بوده زیرا تغییر پارامتر "تعداد عوامل" باعث تغییری در مقدار دقت مدل ها نمی کند.انتخاب ویژگیِPCA تنها با کمک 19 متغیر پیوسته در هنگام محاسبات و بعد از اجراء الگوریتم،امکان پذیر خواهد بود در نتیجه تعداد 13 متغیر جدید به دست خواهند آمد.این متغیرهای جدید با 19 متغیر دیگر ترکیب شده اند و میانگین اهمیت نسبی متغیرهای فوق باید به دقت محاسبه شده هرچند که وزن آن ها بیشتر از 15/. بوده است.این متغیرهای عبارتند از:چرخه مطالبات،دارایی های جاری،نسبت بدهی به دارایی های غیر جاری،نمره کل کیفیت،نسبت بهره به سرمایه ثابت،بازده فروش، سرمایه در گردش یا سرمایه عملیاتی،نسبت بدهی های جاری به دارایی خالص و سود سهام تقسیم شده.

بعد از اجراء FS و تعیین پارامترها،جهت یافتن بهترین ویژگی ها،باید تمام الگوریتم های گروهی و مستقل در هر یک از مدل های FS(جداول 1-4) اجرا شوند. شکل 3،بهترین نتایج طبقه بندی برای هرالگوریتم FS را نشان می دهد به عنوان مثال،با استفاده از ویژگی های انتخابی GA در تمام الگوریتم ها،می توان گفت الگوریتم ANN-bagging از بهترین عملکرد با مقادیر دقت و AUC به ترتیب 87.4 و 90.1 درصد برخوردار است .

مطابق با شکل 3، در تمام الگوریتم های FS،مدل طبقه بندی ANN-bagging از عملکرد طبقه بندی بسیار مطلوبی بهره مند بوده علاوه بر این،مدل فوق که از الگوریتم PCA به دست آمده نسبت به سایر الگوریتم های دیگر،نتایج بسیار خوبی را به دنبال داشته است.مقادیر مربوط به دقت و AUC در بهترین الگوریتم FS به ترتیب 88.4 و 90.3 درصد بوده اند .

شکل 4،میانگین مقادیر دقت و AUC در این 4 الگوریتمFS و در تمام الگوریتم های طبقه بندی،را نشان می دهدواضح است که PCA از عملکرد بهتری نسبت به سایر الگوریتم ها بر اساس دو شاخص دقت و مقدار AUC به ترتیب 87.8 و 86.2درصد برخوردار بوده بنابراین با توجه به اشکال 3 و 4،ما از این الگوریتم جهت ساخت مدل های طبقه بندیِ گروهی و مستقل استفاده خواهیم کرد .



شکل4:نتایج الگوریتم های FS با توجه به میانگین دقت و AUC در تمام الگوریتم های طبقه بندی

Accuracy (%)	AUC (%)
82.05	79.81

جدول 5: بهترین مدل بیز

Confidence for pruning	Minimal size for split	Accuracy (%)	AUC (%)
0.125	16	85.90	89.31

جدول6: بهترین مدل CART

Learning rate	Epochs	Accuracy (%)	AUC (%)
0.6	400	87.18	84.54

جدول7: بهترین مدل ANN

Kernel type	Coefficient	Accuracy (%)	AUC (%)
RBF	0	85.90	81.85

جدول 8: بهترین مدل SVM

بعد از انتخاب بهترین الگوریتم FS و تعیین پارامترهای آن،حال باید از آن در ساخت الگوریتم های طبقه بندی استفاده نمود. در ادامه،نتایج طبقه بندی همراه با پارامترهای تعیین شده برای هر الگوریتم ذکر خواهد شد.در برنامه Rapid Miner،تمام مقادیر پارامترها(به جزء پارامترهای تعیین شده) در هر دو طبقه بندی کننده مستقل و گروهی به صورت پیش فرض در نظر گرفته شده اند.در درخت تصمیم گیری CART،دو پارامتر" کمترین اندازه شکاف و آماده برای هرس" دارای مقدار 0-5/. و 4-40 بوده اند.علاه بر این،در مدل ANN،پارامترهای"تعداد دورها و میزان یادگیری" به ترتیب 4/.-6/. و 400-600 تعیین شده اند. تعداد لایه های مخفی و سرعت حرکت شتاب نیز 1 و 2/. انتخاب شده اند. شایان ذکر است که نوع لایه های ورودی و خروجی به ترتیب خطی و حلقوی بوده اند. در SVM،پارامتر "نوع کرنل" به صورت زیر بیان شده است: 1- خطی 2- چند جمله ای و 3- RBF.همچنین پارامتر ضریب ثابت نیز بین 100-100 و سایر پارامترها به صورت پیش فرض انتخاب شده اند.در مدل آدابوست،پارامتر تکرار دارای مقادیر 8،5 و 10 بوده در مدل bagging،پارامتر نسبت نمونه ها نیز بین 7/. تا 1 انتخاب شده در درخت تصادفی،ارامتر تعداد درخت ها بین 8-10 بوده و کمترین اندازه شکاف در دو مرحله بین 4-15 در نظر گرفته شده است.پارامتر آماده برای هرس در مرحله بین 2/. تا 5/. انتخاب شده،دو پارامتر مینیمم بهره و ماکزیمم مق نیز 1/. و 20 انتخاب شده اند. در مدل stacking،بهترین یادگیرنده ها عبارتند از:CART-SVM,ANN .،بهترین مدل در هر الگوریتم بعد از تعیین پارامترها در جداول 5-12 نشان داده شده اند.

Row	Model	Iterations	Accuracy (%)	AUC (%)
1	ANN–AdaBoost	8	91.03	91.20
2	CART–AdaBoost	8	83.33	84.72
3	Naïve Bayes–AdaBoost	8	82.05	83.84
4	SVM–AdaBoost	8	85.90	75.28

جدول9: بهترین مدل آدابوست

Row	Model	Sample ratio	Accuracy (%)	AUC (%)
1	ANN-bagging	1	88.46	90.28
2	CART-bagging	1	84.62	89.31
3	Naïve Bayes-bagging	0.7	83.33	83.70
4	SVM-bagging	0.8	85.90	81.39

جدول10: بهترین مدل bagging

در ادامه،شکل 5،به مقایسه الگوریتم های مستقل و گروهی با توجه به دو شاخص دقت AUC خواهد پرداخت. الگوریتم آدابوست در زمینهِ امتیاز دهیِ اعتباری بهترین الگوریتم طبقه بندی،شناخته شده است و مقادیر مربوط به دقت و شاخصAUC در آن به ترتیب 91 و 91.2درصد بوده اند. الگـوریتم ANN-bagging نیـز جایگـاه بعدی را به خود اختصاص داده است. الگوریتم های بیز و SVM- آدابوست به خاطر مقـادیر پـایین دو شـاخص دقت و شاخصAUC،جزء بدترین الگوریتم ها شناخته شده اند.

دو شکل 6 و7 به مقایسه نتایج دقت طبقه بندی و شاخص AUC در الگوریتم های طبقه بندی اشـاره دارند.بـه منظور مقایسه الگوریتم های طبقه بندی ،دو شاخصِ میانگین دقت و میانگین AUC با یکدیگر مقایسه شده که نتایج آن درشکل 8 قابل مشاهده است .

جدول11: بهترین مدل جنگل تصادفی

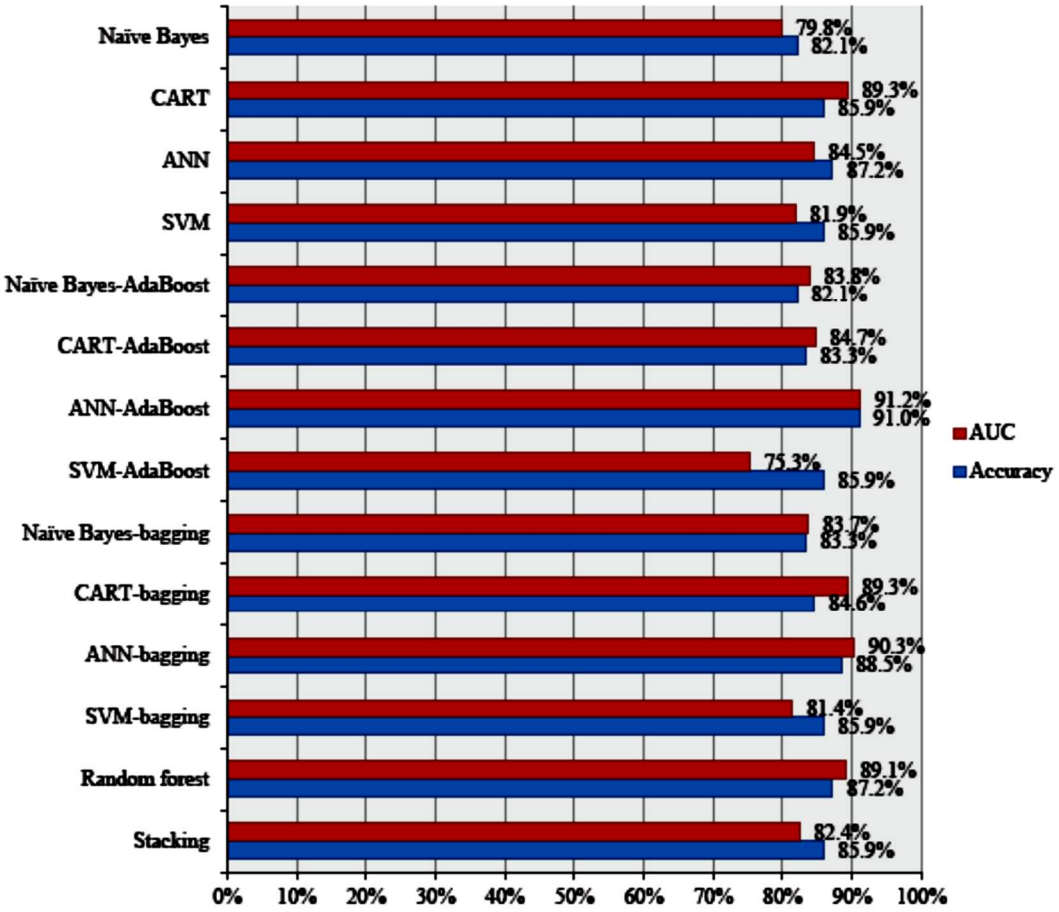
Confidence for pruning	Minimal size for split	Number of trees	Accuracy (%)	AUC (%)
0.2	4	10	87.18	89.12

جدول12: بهترین مدل stacking

Accuracy (%)	AUC (%)
85.90	82.41

بنابراین می توان چنین نتیجه گرفت که:

1. در الگوریتم های FS،الگوریتم PCA بهترین عملکرد را نسبت به سایر الگوریتم ها داشته بنابراین ویژگی های انتخابی از این الگوریتم توانسته دقت طبقه بندی و شاخص AUC را بهبود بخشد (دو شکل 3و4)



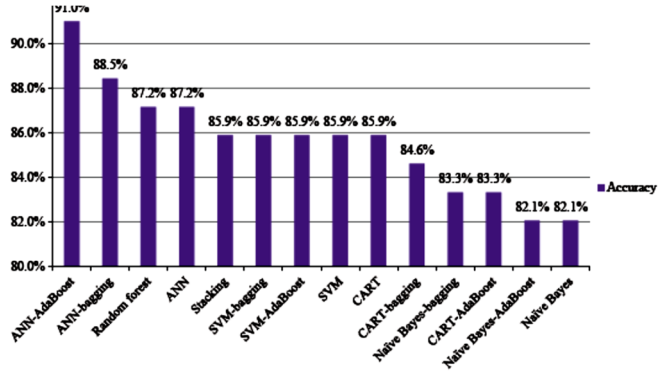
شکل5: نتایج الگوریتم های طبقه بندی بر حسب دقت و شاخص AUC

2. مطابق با شکل 8، الگوریتم های یادگیری گروهی از عملکرد مطلوب ترین نسبت به الگوریتم های مستقل بهره مند اند .علاوه بر این، در دو شکل 6و7،دقت و شاخص AUC در سطح مطلوبی قراردارد

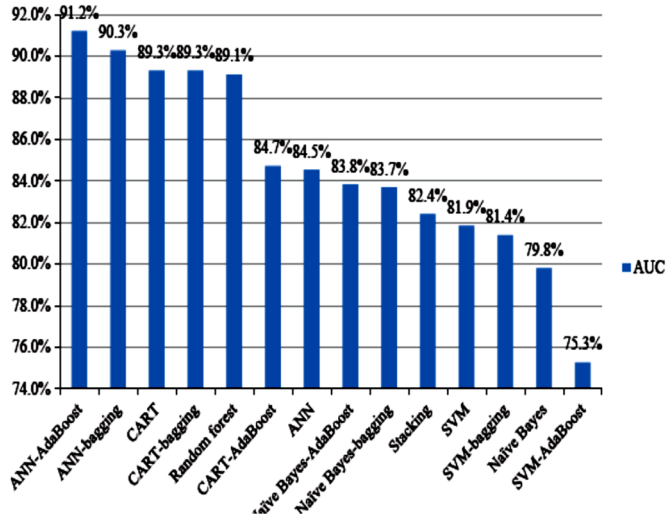
.اما باید توجه داشت که الگوریتم مستقل قادر است به خوبی با الگوریتم گروهی ترکیب شود به عنوان مثال،SVM قادر به ترکیب با الگوریتم گروهی نخواهد بود.(اشکال 6-8)

3. مطابق با شکل 7 و 8،ANN نسبت به سایر روش های یادگیریِ دیگر همانند بیز وSVM(به استثناء CART)از نتایج مطلوب تری برخوردار بوده است.مطابق با شکل 6،ANN از دقت طبقه بندی بهتری نسبت به سایرین بهره مند است . همچنین ANN- آدابوست نیز نسبت به سایر الگوریتم مستقل و گروهیِ دیگر،بهترین عملکرد را به خود اختصاص داده است(اشکال 6-8). به عبارت بهتر،بهترین طبقه بندی کنندهِ مستقل در الگوریتم های یادگیری،ANN است.

4. مطابق با شکل 8، الگوریتم های بیز،بیز-آدابوست و بیز-bagging ،از بدترین عملکرد برخوردار بوده همچنین الگوریتم بیز در استراتژی یادگیریی مستقل،هیچ گونه نقش قابل توجهی را ایفا نمی کند.



شکل6: نتایج الگوریتم های طبقه بندی بر حسب دقت



شکل7: نتایج الگوریتم های طبقه بندی بر حسب شاخص AUC

5- نتیجه گیری و پیشنهادات:

در این مقاله،مدلِ امتیاز دهیِ اعتباری ترکیبیِ جدیدی از الگوریتم های FS و طبقه بندی کننده های یادگیریِ گروهی و مستقل در بانک توسعه صادرات ایران مورد استفاده قرارگرفتند. در مدل پیشنهادی، در وهله اول،تعداد 4 الگوریتم FS جهت دستیابی به مجموعه داده مناسب با دقت طبقه بندی مطلوب، انتخاب شدند. الگوریتم FS که بهترین دقت طبقه بندی را برای SVM به ارمغان می آورد نیز در این مدل انتخاب گردید.علاوه بر این،از مرحله تعیین پارامترها جهت دستیابی به بهترین نتایج در الگوریتم FS استفاده شد.در این بین،روش هایی همچون GA,PCA,FS, و معیار بهره اطلاعات به عنوان بهترین گزینه ها انتخاب شدند. پس از انتخاب بهترین الگوریتمFS،از الگوریتم های طبقه بندی در مدل استفاده شده و در مدل های ضعیف تر نیز،از روش FSجهت مدلسازی استفاده به عمل آمد. اشکال 6-8 نشان می دهند که الگوریتم ANN- آدابوست بهترین عملکرد را در حوزهِ امتیاز دهیِ اعتباری از آن خود کرده اند همچنین الگوریتم های SVM- آدابوست، بیز و بیز-آدابوست بدترین عملکرد را بر حسب شاخص های ارزیابی،در اختیار داشته اند.

مدل ترکیبیِ ما جهت نشان دادن شاخص های مالی به مشتریان شایسته،از روشِ امتیاز دهیِ اعتباری استفاده می کند.علاوه بر این،جهت برطرف نمودن مشکلات موجود،رویکردی جامع همراه با تعداد زیادی از الگوریتم های FS و طبقه بندی و نیز چهارچوب داده کاویِ ترکیبی جهت افزایش سطحِ اطمینانِ مدل های امتیاز دهیِ اعتباری،ارائه شده اند.

بر این اساس،می توان چنین نتیجه گرفت که الگوریتم های طبقه بندیِ گروهی نسبت به الگوریتم های مستقل در بررسیِ امتیاز دهیِ اعتباریِ داده کاوی از عملکرد مطلوبتری برخوردار اند.همچنین مطالعات آتی باید به بررسی روش های ترکیبی در زمینه امتیاز دهیِ اعتباری نیز توجه داشته باشند. این مقاله از یک مدل ترکیبی جهت شبیه سازی الگوریتم های FS استفاده نموده و دو الگوریتم گروهی و مستقل را با یکدیگر مقایسه نموده است .نتایج نشان دادند مدل های ترکیبی نسبت به مستقل ،عملکرد قابل توجهی دارند بنابراین جهت دستیابی به عملکرد بالا،باید تعیین پارامترها برای FS و الگوریتم های طبقه بندی به دقت انجام شود.در واقع،نتایج بیان

کرده اند روش های غیر پارامتریِ امتیاز دهیِ اعتباری نسبت به روش های پارامتری،نمایش بهتری از خود نشـان داده اند.

در مطالعات آینده،سایر الگوریتم های FS نظیر الگوریتم تبرید شبیه‌سازی شدهSA،PSO،نمرهF،LDA ،کلنـی مورچه ها و تئوری مجموعه های راف نیـز بایـد بررسـی شـوند. همچنین،سـایر الگـوریتم هـای طبقـه بنـدی از جملهLR-CHAID-KNN، نیز باید بررسی شده و نتایج هر یک با دیگری مقایسه شوند. شایان ذکر است کـه مرحلهِ تعیین پارامترها در این مقاله جهت دستیابی به نتایج بهتر می توانـد توسـعه یابـد.بنابراین الگـوریتم هـای یادگیری گروهی باید با سایر الگوریتم های دیگر در آینده مقایسه شوند.پیشنهاد می کنیم در میان انواع مختلـف مدل های ترکیبیِ امتیاز دهیِ اعتباری،از دو نوع مدل استفده کنید که عبارتند از:1) تکنیک های طبقه بندیِ هر کلاستر با کمک الگوریتم های اجحرا شده در مجموعه داده های موجود و 2) رویکردهای ترکیبـیِ 2 مرحلـه ای همراه با دو الگوریتم طبقه بندی.