**ارزیابی الگوریتم تصمیم گیری هرس درخت برای دقت پیچیدگی و طبقه بندی(مطالعه موردی)**

**علی سلدوزی,[[1]](#footnote-1)1، سید نوید مقدسی2، عبدالرضا اسماعیلی3**

**1- دکتری تخصصی، مهندسی برق قدرت، شرکت خدمات انفورماتیک**

**2- دکتری تخصصی، مهندسی برق قدرت، شرکت کارت اعتباری ایران کیش**

**3- دانشکده تحقیقات شکافت هسته‌ای و پلاسما، مؤسسه تحقیقات فناوری و علوم هسته‌ای، تهران**

**خلاصه**

طبقه بندی یک مشکل مهم در داده کاوی است. با توجه به پایگاه داده ای از پرونده ها، هر کدام با یک برچسب کلاس، یک طبقه بندی کننده یک توضیح مختصر و معنی دار برای هر کلاس ایجاد می کند که می تواند برای طبقه بندی پرونده های بعدی استفاده شود. تعدادی از طبقه بندی های محبوب ایجاد درخت تصمیم برای ایجاد مدل های کلاس. این طبقه بندی ها ابتدا یک درخت تصمیم گیری را ایجاد می کنند و سپس زیر شاخه ها را از درخت تصمیم گیری در یک مرحله پس از برداشتن به منظور بهبود دقت و جلوگیری از"overfiting" بریده می شود.

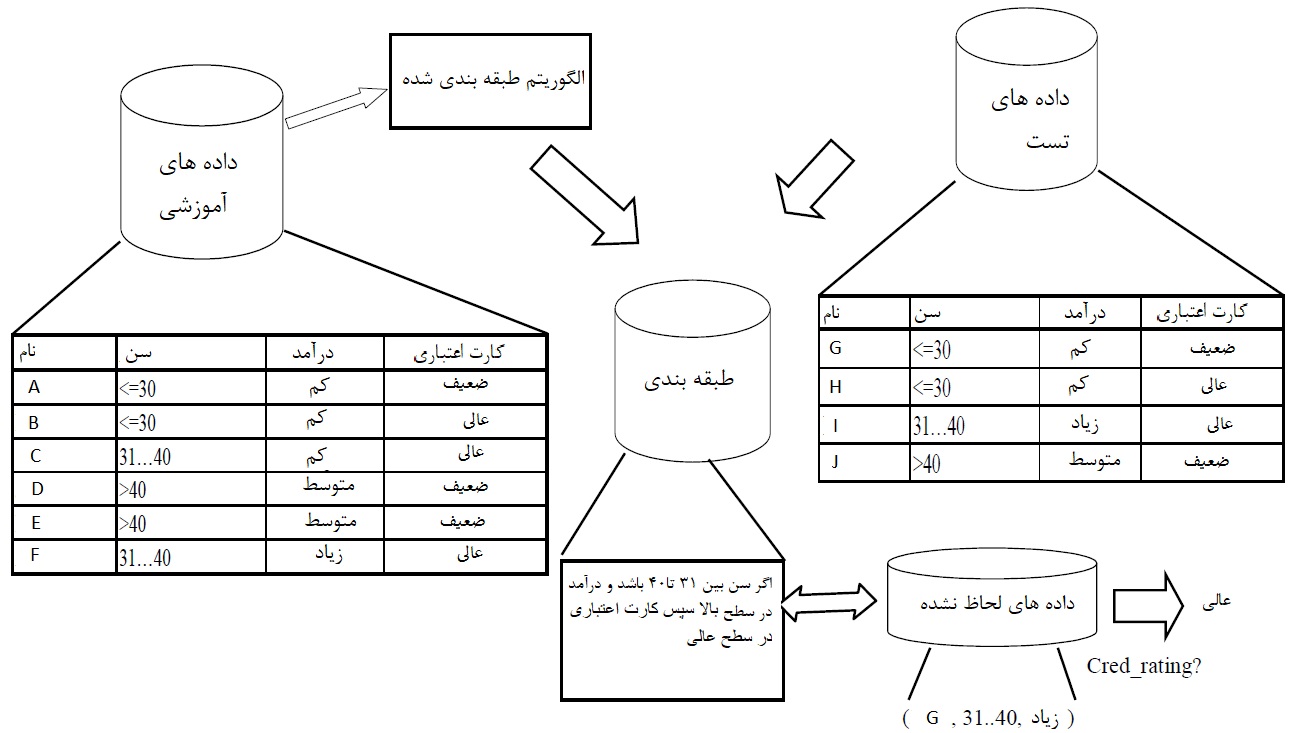
در این مقاله، روش های مختلف هرس در دسترس و ویژگی های مختلف آنها مورد بحث قرار می گیرد. همچنین اثربخشی هرس با توجه به پیچیدگی و دقت طبقه بندی با استفاده از الگوریتم طبقه بندی4.5C در پایگاه داده کارت اعتباری با هرس و بدون هرس ارزیابی می شود. به جای طبقه بندی معاملات یا تقلب یا عدم تقلب، معاملات در چهار سطح ریسک قرار می گیرند که یک مفهوم ابتکاری است.

**کلمات کليدي:** طبقه بندی درخت تصمیم، هرس درختان، overfiting

1. معرفی

طبقه بندی یک مشکل مهم در داده کاوی است. این موضوع به طور گسترده توسط جامعه یادگیری ماشین به عنوان یک راه حل ممکن برای برداشت دانش و یا استخراج دانش شناخته شده است. ورودی به یک طبقه بندی کننده مجموعه ای از سوابق آموزشی است که هرکدام با یک برچسب کلاس برچسب گذاری می شوند. مجموعه ای از مقادیر ویژگی هر رکورد را تعیین می کند. مشخصه هایی با حوزه های گسسته به عنوان دسته بندی نامیده می شوند، در حالی که افراد دارای دامنه های مرتب به عنوان عددی نامگذاری می شوند. هدف این است که یک مدل یا توصیف برای هر کلاس را از لحاظ صفات ارائه دهیم. سپس مدل برای طبقه بندی سوابق آینده که کلاس های آنها نامشخص است استفاده می شود. طبقه بندی داده ها [1] یک فرایند دو مرحله ای است که در شکل 1 نشان داده شده است. در مرحله اول، یک مدل ساخته شده است که توصیف یک مجموعه از پیش تعیین شده از کلاسها یا مفاهیم داده است. این مدل با تجزیه و تحلیل تابع پایگاه داده توصیف شده توسط ویژگی ساخته شده است. هر تابع فرض می شود متعلق به یک کلاس از پیش تعریف شده، به عنوان تعیین شده توسط یکی از ویژگی ها، به نام ویژگی کلاس برچسب. داده ها برای تجزیه و تحلیل داده ها برای ایجاد مدل جمع آوری داده های آموزشی را تشکیل می دهند. این فرایند به عنوان یادگیری ماشین نامیده می شود [2، 3]. از آنجا که برچسب کلاس برای هر نمونه آموزشی ارائه شده است، این مرحله نیز به عنوان یادگیری نظارت شناخته می شود. این تضاد با یادگیری بی نظیر (یا خوشه بندی) است که در آن برچسب کلاس هر یک از نمونه های آموزشی شناخته شده نیست و تعداد یا مجموعه ای از کلاس های یادگیری ممکن است پیش از آن شناخته شوند. به طور معمول، یادگیری مدل در قالب قوانین طبقه بندی، قوانین تصمیم گیری، و یا فرمول های ریاضی ارائه شده است. این قوانین می تواند برای دسته بندی نمونه های داده های آینده مورد استفاده قرار گیرد و همچنین درک بهتر مطالب محتوی پایگاه داده ها را فراهم کند. در مرحله دوم، مدل برای طبقه بندی استفاده می شود. ابتدا دقت پیش بینی مدل تخمین زده می شود. اگر دقت مدل قابل قبول در نظر گرفته شود، مدل میتواند برای طبقه بندی تپهای داده آینده یا اشیائی که برچسب کلاس آن شناخته نشده است، استفاده شود. به عنوان مثال در شکل 1 داده شده با یک پایگاه داده از اطلاعات اعتبار مشتریان، قوانین طبقه بندی می تواند یاد بگیرد که مشتریان را به عنوان داشتن رتبه های اعتباری عالی یا عادلانه شناسایی می کنند. با این تجزیه و تحلیل داده ها از مشتریان موجود می توان برای پیش بینی رتبه اعتباری مشتری جدید یا آینده استفاده کرد. هنگامی که درخت تصمیم گیری با الگوریتم طبقه بندی درخت تصمیم گیری شکل می گیرد، گاهی اوقات این اتفاق می افتد که بعضی از قوانین ناخواسته و بی معنی را به صورت عمیق تر تولید می کند، به نام"overfiting" نامیده می شود [4]. این تنها با توجه به ویژگی هایی که در تشکیل حکومت خاص نقش مهمی دارند، می توان از آن اجتناب کرد. این کار با متوقف کردن رشد درخت تصمیم گیری در سطح خاص انجام می شود، به طوری که قوانین شکل گرفته به طبقه بندی بهتر می رسند. دو نوع روش هرس کردن وجود دارد: ابتدا قبل از برش [4، 5]، یعنی در حالی که ساخت درخت تصمیم گیری برای بررسی اینکه آیا درخت بیش از اندازه بر اساس اقدامات مختلف مانند خطای لاپلاس [4]،[6]LMDطول، هزینه و غیره و روش دوم، پسور هرس، که در آن ابتدا ساخته شده درخت و سپس کاهش شاخه ها و سطوح درخت تصمیم‌گیری انجام می شود. در این مقاله ما در مورد روش‌های مختلف هرس درخت تصمیم‌گیری مورد بحث قرار گرفته‌ایم.

همچنین اثربخشی هرس با استفاده از الگوریتم[7]4.5C با و بدون هرس کردن بر روی بانک اطلاعات کارت اعتباری ارزیابی می شود. به عنوان رفتار انسان غیر قابل پیش بینی است، هر گونه معامله یا به عنوان تقلب یا غیر تقلب قابل قبول نیست. در تمام مطالعات قبلی [8]، [9]، [10] معاملات در دو سطح تقسیم یا مشروعیت طبقه بندی شدند. رویکرد ما معاملات معاملاتی کارت اعتباری را در مقادیر مختلف تقلب طبقه بندی می کند که بستگی به شرایط تقلبی مختلفی از رفتار تاریخی مشتریان دارد. با در نظر گرفتن خطر درگیر در معاملات، بانک ها می توانند اقدامات پیشگیرانه لازم را انجام دهند و به مشتریان خدمات ارائه دهند. بنابراین طبقه بندی شده توسط ما طبقه بندی چند طبقه ای از معاملات را به جای طبقه بندی باینری تنها طبقه بندی می کند. بنابراین در اینجا الگوریتم4.5C با هرس به عنوان الگوریتم طبقه بندی چند سطحی(MLPC) نامیده می شود. بخش 2 در مورد تکنیک های مختلف هرس بحث می کند. بخش 3 جزئیات اجرای 4.5CوMLPC را ارائه می دهد. بخش 4 نتایج را با هر دو الگوریتم ارزیابی می کند. بخش 5 بیانگر نتیجه گیری است.

****

شکل 1- فرایند یادگیری و طبقه بندی درخت تصمیم گیری

2. تکنیک های هرس

اگر چه درخت تصمیم گیری توسط روش هایی مانند3DIو4.5C دقیق و کارآمد هستند آنها اغلب از عواقب ارائه درختان بسیار بزرگ که غیر قابل درک است برای کارشناسان رنج می برند [11]. برای حل این مشکل محققان در این زمینه علاقه مند شدید به هرس درخت هستند. روش های هرس کردن درخت یک درخت بزرگ را به یک درخت کوچک تبدیل کنید و فهمیدن آن را آسان تر کنید. چنین روش هایی معمولا از روش های آماری استفاده می کنند شاخه های کمترین قابل اطمینان را حذف کنید، به طور کلی به سرعت سریعتری می شود طبقه بندی و بهبود توانایی درخت به درستی داده های آزمایشی را طبقه بندی کنید. "لازم است بدانید مزیت و ضرر هر هرس درخت تصمیم گیری روش قبل از اینکه تصمیم بگیرد که کدام روش هرس باشد انتخاب شد". بعضی از روش های اصلی زیر برای ساده سازی درختان تصمیم گیری است

2-1. کاهش خطای هرس

این روش توسط Quinalnپیشنهاد شد [11]. این ساده ترینو قابل فهم ترین روش در هرس درخت تصمیم گیری است. . برای هرS-subtree غیر برگ درخت تصمیم اصلی، تغییر در طبقه بندی غلط بر روی مجموعه آزمون مورد بررسی قرار گرفته است. اگر این زیرشاخه توسط این جایگزین شده باشد، اشتباه طبقه بندی می شود بهترین برگ ممکن است که اکثریت برگ باشد. اگر میزان خطا از درخت جدید برابر یا کوچکتر از اصل است درخت و آنS-subtreeشامل هیچ subtree با همان ویژگیs، توسط برگ جایگزین شده است. در غیر این صورت، روند را متوقف کنید. این محدودیتی که زیر درختs حاوی هیچ زیر درخت با همان اموال تضمین کاهش هشدار خطا در القاء از بالا به بالا[12]. از آنجا که هر گره تنها یک بار برای ارزیابی آن بازدید می شود امکان هرس کردن آن، مزیت این روش خطی آن است پیچیدگی محاسباتی [12]. با این حال، این روش نیاز به یک تست جدا از موارد موجود در مجموعه آموزشی که از آن درخت ساخته شد [11]. هنگامی که مجموعه آزمون بسیار کوچکتر از مجموعه آموزش، این روش می تواند منجر به بیش از هرس شود. بسیاری محققان دریافتند که کاهش هدر خطا نیز انجام شده است به عنوان بسیاری از روش های دیگر هرس کردن از نظر دقت و بهتر از اندازه ترین درخت است [12].

2-2. خطای هرس پیشبینی نشده

این روش همچنین توسطQuinaln پیشنهاد شده بود [11]. توسعه یافته در زمینه 3DI. کوینلن متوجه شد که این نیز به طور خوشبختانه برای ما است که استفاده از یک مجموعه آموزشی برای آزمون میزان خطایa درخت تصمیم گیری استفاده می کنیم، زیرا درختان تصمیم گرفته شده اند مجموعه آموزشی در این مورد، میزان خطا می تواند 0. . اما برخی از داده ها غیر از مجموعه آموزشی استفاده شده است نرخ خطا به طور چشمگیری افزایش خواهد یافت. برای حل این مشکل،Quinaln از تداوم استفاده کرد اصلاح برای توزیع binomial برای دریافت میزان خطا که واقعی تر است در آمار، اصلاح تداوم مفید است روش در استفاده از توزیع نرمال به محاسبه احتمال دوگانه. هنگامی که توزیع عادی است (توزیع مداوم) برای یافتن تقریبی استفاده می شود پاسخ به مشکلات ناشی از توزیع دوجمله ای (توزیع گسسته)، سازگاری برای عدم تطابق ساخته شده است انواع توزیع این اصلاح تداوم نامیده می شود. Quinaln از معادلات زیر برای به دست آوردن تعداد اشتباهات از معادلات استفاده می کند.

n’(t) = e(t) + (1/2) (1)

n’(Tt) = e(Tt ) + (NT/2) (2)

معادله (1) تعداد اشتباهات اشتباه برای گرهt و معادله (2) تعداد اشتباهات طبقه بندی نشده برای زیر تیرهT است.

از انجایی که:

تعداد برگ برای زیر تیرهT است NT

)t(e تعداد اشتباهات در گرهt است

تعداد اشتباهات طبقه بندي براي زيربخشT است. e(Tt)

2/1 در معادله (1) و (2) ثابت است که نشان دهنده سهم یک برگ در پیچیدگی درخت است

این روش هرس تنها درخت زیر را نگه می دارد اگر شکل تصحیح شده آن (از معادله 2) بیش از یک خطای استاندارد بهتر از شکل گره (از معادله 1) باشد. این روش بسیار سریعتر از کاهش هدر خطا است و همچنین تضمین های بالاتر را فراهم می کند. معایب آن این است که، در بدترین حالت، هنگامی که درخت نیازی به هرس کردن ندارد، هر گره یک بار به آنجا خواهد رفت [12]

2-3. هزینه هرس پیچیدگی

این روش توسطBreiman و همکاران پیشنهاد شده است[13]. این در نظر گرفتن هر دو تعداد خطاها و پیچیدگی درخت است. اندازه درخت برای نشان دادن پیچیدگی درخت استفاده می شود. همچنین به عنوان روش برشTRAC شناخته می شود و فلورینا اسپوزیتو و همکارانش آن را در دو مرحله توصیف می کند:

1. انتخاب یک خانواده پارامتری از زیر شاخه های{LT.......,1T,0T} با توجه به برخی اکتشافات 0T درخت تصمیم اصلی است و هر1+iT با جایگزینی یک یا چندزیر شاخه ها از iT با برگ توسط هرس کردن شاخه هایی که کمترین افزایش نرخ خطا ظاهری در برگ برگ را نشان می دهد تاLT درخت نهایی فقط یک برگ است به دست می آید.

2. انتخاب بهترین درخت با توجه به برآورد میزان خطای واقعی درختان در خانواده پارامتری

برای مثال، زیر تیرهT مورد استفاده برای طبقه بندی هر یک از مواردN در مجموعه آموزش وE از نمونهN به طور غیرمجاز طبقه بندی شده است اگر زیر درختT با بهترین برگ جایگزین شود. اجازه دهیدTN تعداد برگ در زیر تیرهT باشد، معادله زیر برای تعریف کل هزینه پیچیدگی زیر درختT استفاده می شود:

هزینه پیچیدگی= (E/N) + α \* NT (3)

که در آنα هزینه یک برگ اضافی در درخت است و باعث کاهش خطا در هر برگ می شود.

اگر زیر درخت چیده شود، درخت جدیدM را بیشتر از موارد موجود در مجموعه آموزشی تعریف می کند، اما شامل1-TN برگ کمتر است. همان هزینه پیچیدگی زمانی به دست می آید

α = (M / (N\* (NT – 1)) (4)

از معادله فوق،α برای هر زیر درخت محاسبه می شود و زیر درخت (ها) با کوچکترین مقدارα برای هرس انتخاب می شود. این روند را ادامه دهید تا برگ حاصل شود. کار بعدی این است که یکی از درختان را انتخاب کنید. خطای استانداردES نرخ اشتباه طبقه بندی است.

SE = (R \* (100 – R)) / N (5)

از انجایی که

نرخ اشتباه طبقه بندی درخت بری(هرس کردن) =R

عداد نمونه ها در داده های آزمون=N

کوچکترین درخت که تعدادی از اشتباهات مشاهده شده در مجموعه آزمون ازES+R فراتر نمی رود انتخاب شده است.

این روش نیاز به یک مجموعه هرس متمایز از مجموعه آموزش اصلی دارد.

معایب آن اینست که تنها می توان یک درخت را در مجموعه{LT.......,1T,0T} که در مرحله اول بدست آمده است، به جای مجموعه ای از همه زیر شاخه های ممکن [12]. همچنین به نظر نمی رسد که مدل پیچیدگی هزینه استفاده شده برای ایجاد توالی زیرشاخه ها هنگام انتخاب بهترین درخت رها شود[11].

2-4. حداقل خطای هرس

این روش توسط Niblett و Brotko توسعه داده شد.[14] این روش رو به پایین است که یک درختی را دنبال می کند که میزان خطای مورد انتظار را در یک مجموعه داده مستقل به حداقل می رساند. فرض کنید کلاس هایk برای یک مجموعه ای از داده ها وجود دارد که تعداد آنn و ncکلاس c با بیشترین تعداد داده است. اگر پیش بینی می شود که تمام نمونه های آینده در کلاسC باشد، برای پیش بینی میزان خطای انتظار برای معادله زیر استفاده می شود:

Ek = (n- nc + k -1) / (n+k) (6)

از انجایی که:

تعداد کلاس ها برای همه داده ها=k

میزان خطای انتظار می رود اگر ما پیش بینی کنیم که تمام نمونه های آینده در کلاسc قرار خواهند گرفت=kE

روش متشکل از سه مرحله [15] است

1. در هر گره غیر برگ درخت تصمیم گیری، از معادله (6) برای محاسبه میزان خطای انتظار می رود اگر این زیر درختی بریده شود.

2. اگر مقدار گره نشتی نباشد، با توجه به نسبت مشاهدات در طول هر شاخه، با توجه به وزن، ترکیب با نرخ خطای مورد انتظار را محاسبه کنید. [12] [14]

3. اگر هرس گره منجر به یک میزان خطای پیش بینی شده بیشتر شود، سپس زیر درخت را نگه دارید؛ در غیر این صورت، آن را بنوشید.

J.Mingers اشاره می کند که در این روش چندین معایب وجود دارد. اولا، در عمل تقریبا درست است که همه کلاس ها به همان اندازه احتمال دارد. دوم، این روش فقط یک درخت تنها تولید می کند.

این مضر در زمینه سیستم های متخصص است، در صورتی که چندین درخت، که به درجه های مختلف در دسترس است، بیشتر مفید خواهد بود. سوم، تعداد طبقات به شدت بر درجه هرس تاثیر می گذارد، منجر به نتایج ناپایدار می شود. حداقل هرس خطا توسط Cestnik و Brotko بهبود یافته [14] و آخرین نسخه از حداقل هشدار خطا دو مشکل از روش اصلی را برطرف کرده

عصب خوش بینانه و وابستگی میزان خطای انتظار برای تعداد کلاس ها

2-5. ارزش بحرانی هرس

این روش توسط Mingers [17] پیشنهاد شده است. در این روش، آستانه، به نام ارزش بحرانی، برای ارزیابی اهمیت یا قدرت یک گره تعیین می شود. هنگامی که گره به مقدار بحرانی نمی رسد، آن را قطع می کند.

اما هنگامی که یک گره با شرایط هرس مطابقت دارد، اما فرزندانش همه شرایط هورمون را برآورده نمی کنند، این شاخه باید نگه داشته شود زیرا حاوی گره های مربوطه است. اگر یک مقدار بحرانی بزرگتر انتخاب شده باشد، یک درخت کوچکتر حاصل می شود به دلیل هرس شدن شدید تر.

Mingers هرس کردن ارزش بحرانی را به عنوان دو مرحله اصلی توصیف می کند [17]

1. زیرسطحی برای افزایش مقادیر بحرانی

2. اهمیت درختان زراعتی را به طور کلی و توانایی پیش بینی آنها را اندازه گیری کنید و بهترین درخت را در میان آنها انتخاب کنید.

ضعف این روش تمایل شدیدی به زیر گله دارد و این روش درختان را با دقت پیش بینی نسبتا کم انتخاب می کند [18].

2-6. هرس بهینه

Breiman همکاران، یک اصطلاحات مناسب را که برای تعیین و تأیید خواص ریاضی هرس مطلوب استفاده می شود معرفی می کنند [13]. آنها همچنین یک الگوریتم را برای انتخاب یک زیر درخت به طور مطلوب برش از میان کاندیدان معرفی می کنند [13]

Bratkoو]Bohanec 19] Almuallim [20[ به مسئله یافتن هرس مطلوب به روش دیگری اشاره می کنند. Bohanec یک الگوریتم تضمین هرس کردن بهینه را ارائه داد و(OPT) را از لحاظ پیچیدگی محاسباتی بهبود داد .انگیزه آنها برای ساده سازی درخت های تصمیم گیری متفاوت است از انگیزه های معمول برای برداشتن تصمیم گیری درختان هنگام یادگیری از داده های پر سر و صدا. هر دو [19] و [20] فرض می کنند که درخت تصمیم اولیه و غیرقابل قبول کاملا درست است. با این حال، در یادگیری از داده های پر سر و صدا، که مورد ما است، تصور می شود که درخت تصمیم گیری اولیه،هرس نشده ی نادرست است و هرس مناسب، دقت آن را بهبود می بخشد.

2-7. هرس درخت تصمیم گیری حساس به هزینه

یک مشکل اصلی برای بسیاری از روش های برش درخت تصمیم گیری این است که وقتی یک درخت تصمیم گیری بریده شود، همیشه فرض می شود که همه کلاس ها به همان اندازه احتمال و به همان اندازه مهم هستند. با این حال، در مشکلات طبقه بندی دنیای واقعی، هزینه های مربوط به نمونه های اشتباه طبقه بندی از هر کلاس نیز وجود دارد. در حال حاضر شایعترین روش برای روش هرس کردن حساس به استفاده از تکنیک های آمار برای مقابله با مشکل است. استفاده از مدل های احتمالی و روش های آماری برای تجزیه و تحلیل داده ها تقریبا در تمام رشته های علمی به کار گرفته شده است. برای مثال، محمد اردن از یک روش آماری برای ساخت یک مدل درخت تصمیم استفاده کرد [21] یک پارامتر را می توان از داده های نمونه بر اساس یک شماره واحد (برآورد نقطه) و یا یک فاصله کامل از ارزش های قابل اعتماد (فاصله اطمینان) تخمین زده می شود. با این حال، اغلب هدف تحقیق این نیست که یک پارامتر را برآورد کنیم، بلکه باید تصمیم بگیریم که کدام یک از دو ادعای متناقض در مورد پارامتر صحیح است (برخی از روش های هرس کردن هزینه با استفاده از این [22]). روشهای انجام این کار بخشی از استنتاج آماری به نام آزمون فرضیه است. فرضیه صفر، توسط0H نشان داده شده است ادعا در مورد یک یا چند ویژگی های جمعیت است که در ابتدا فرض می شود درست است. فرضیه جایگزین، که توسط aH مشخص می شود، ادعایی است که متناقض با0H است. . فرضیه صفر به نفع فرضیه های جایگزین رد می شود، فقط اگر شواهد نمونه نشان می دهد که0H نادرست است. اگر نمونه به شدت با0H مخالف باشد، همچنان به حقیقت فرضیه صفر اعتقاد دارد.

3. پیاده سازی

این بخش در مورد الگوریتم القاء درخت تصمیم گیری4.5C و MLPC است که در پایگاه داده کارت اعتباری استفاده می شود.

3-1. بانک اطلاعات اعتباری

بانک اطلاعات کارت اعتباری که برای آموزش و طبقه بندی مورد استفاده قرار می گیرد براساس عکس فوری بانک اطلاعات کارت اعتباری که توسط بانک ساخته شده است، توسعه یافته است. به منظور اهداف امنیتی، بانك ما اجازه نداد كه داده های واقعی را فاش كند، به همین دلیل پایگاه داده به صورت دستی از اطلاعات داده شده و بررسی كلی از دنیای كارت اعتباری پیش پردازش شده است. جدول تراکنش کارت اعتباری که برای یادگیری ساخته شده شامل 101580 رکورد است.

جدول معامله براساس اطلاعات معامله فعلی مانند مقدار، زمان معامله، محل مبادله، تاریخ انقضا وارد شده، محدودیت کارت، علاوه بر آن، برخی از اطلاعات تاریخی نیز با این زمینه ها مانند خرید متوسط ​​سه ماهه گذشته، میانگین خرید از دوازده ماه گذشته، مکان و زمان ترمینال ترمینال مشتری، محدودیت تعداد معاملات در یک روز برای ردیابی رفتار نرمال مشتری. رکورد معامله شامل شماره حساب حساب مشتری نیست، زیرا به جای یادگیری، مدل رفتار حسابهای انفرادی مشتری، مدل های کلی که تلاش می کنند معاملات قانونی را از تقلبات جدا کنند، ساخته شده است. بنابراین مدل مستقل از مشتری است.

3-2. انواع تقلب

در عوض طبقه بندی معاملات داده شده تنها در دو نوع است که تقلب یا عدم تقلب است، در معامله پیاده سازی سیستم می تواند به چهار نوع مختلف کلاس های L1، L2، L3، L4 طبقه بندی شود که بر اساس شرایط تقلبی مختلف تصمیم گیری می شود از تصویر فوری داده شده از پایگاه داده توسط بانک و بررسی انجام شده در دنیای کارت اعتباری ردیابی شده است. وضعیت تقلبی بر اساس آن سطح کلاس به معاملات اختصاص داده شده است.

3-3. الگوریتم درخت تصمیم گیری

الگوریتمMLPC با پیش برش زدن اجرا می شود که در آن هنگام ساخت رشد درخت در سطح بریده شده متوقف می شود. الگوریتم برخی موارد پایه را که در زیر آورده شده است را در نظر می گیرد.

**موارد پایه:**

●تمام نمونه ها در لیست متعلق به یک کلاس هستند. وقتی این اتفاق می افتد، به سادگی یک گره برگ را برای درخت تصمیم گیری برای انتخاب این کلاس ایجاد می کند.

●هیچ یک از ویژگی های ارائه هیچ اطلاعاتی را دریافت نمی کند. در این مورد، الگوریتم یک گره تصمیمی را بالاتر از درخت با استفاده از مقدار پیش بینی شده از کلاس ایجاد می کند. هیچ یک از این ویژگی ها هیچ گونه افزایش اطلاعات را فراهم نمی کند. در این مورد، الگوریتم گره تصمیمی را بالاتر از درخت با استفاده از مقدار پیش بینی شده کلاس ایجاد می کند.

●نمونه کلاس قبلا غیرقابل مشاهده مواجه شد. باز هم الگوریتم یک گره تصمیم را به بالا درخت با استفاده از مقدار مورد انتظار ایجاد می کند.

الگوریتمMLPC:

a)درخت را در یک تقسیم بازگشتی از بالا به پایین بسازید به نحوی به دست آوردن

b)در ابتدا، تمام نمونه های آموزشی را در ریشه در ابتدا حفظ کنید، تمام نمونه های آموزش را در ریشه نگه دارید

c)نمونه های پارتیشن به طور بازگشتی بر اساس ویژگی های انتخاب شده

d) ویژگی تقسیم بر اساس اندازه گیری آنتروپی را انتخاب کنید

e)تمام مراحل را تا زمانی که یکی از سه شرایط راضی باشد، تکرار کنید

1) تمام نمونه ها برای یک گره داده شده متعلق به یک کلاس هستند.

2)هیچ ویژگی باقیمانده برای پارتیشن بندی بیشتر وجود ندارد

3)هیچ نمونه ای وجود ندارد

4) سطح خمیر را تنظیم کنید

**اندازه گیری آنتروپی**

اندازه گیری آنتروپی با استفاده از معادله زیر ارائه می شود. برای مجموعه ای از رکوردs

Entropy E(S) = -∑ pj log pj (7)

در کجا، j=1,2,….,m فرکانس Pj نسبی کلاس j در S آنتروپی است که S را با n پرونده در دو مجموعه S1 با سوابق n1 و S2 با سوابق n2 تقسیم می کند.

E(S1,S2)= [E(S1)]+ [E(S2)] (8)

در محدوده درخت های تصمیم گیری، اگر نتیجه یک گره، طبقه بندی سوابق به دو کلاس C1 و C2 باشد، نتیجه می تواند به عنوان پیام ای که تولید می شود مشاهده شود و آنتروپی اندازه گیری اطلاعات را برای یک پیام می دهد C1 یا C2 اگر مجموعه ای از سوابق T به مجموعه ای از کلاس های جامع غیر متشکل از C1، C2، ....،Cn بر اساس یک مقدار از ویژگی کلاس طبقه بندی شود، سپس اطلاعات مورد نیاز برای شناسایی کلاس یک عنصر ازT

Info (T) = Entropy (P) (9)

Info به معنی اطلاعات

Entropy همان آنتروپی

جایی که P توزیع احتمالی پارتیشن C1، C2، ...، Cn است. P بر اساس فرکانس نسبی خود محاسبه می شود، یعنی

P = ((|C1|/|T|, |C2|/|T| , …|Cn|/|T|) (10)

هدف این است که آنتروپی را پایین بیاوریم. 3.4

3-4. طبقه بندی الگوریتم

دو طبقه در دسته طبقه بندی درخت تصمیم گیری وجود دارد: ابتدا نسل تصمیم گیری از داده های آموزش داده شده تولید می شود و دوم، طبقه بندی واقعی است که در آن قوانین تصمیم گیری درخت تصمیم گیری شکل برای معامله با برچسب کلاس نامشخص برای طبقه بندی آن در یکی از کلاس ها اعمال می شود . الگوریتم برای این طبقه بندی در زیر آمده است:

1. برای هر معامله طبقه بندی شده، یک قاعده تصمیم گیری از جدول تصمیم گیری را بخوانید.

2. زمینه ها را از معامله با هر حکم تصمیم گیری مطابقت دهید. (زمینه هایی که دارای نوشته های خالی در جدول تصمیم گیری هستند، نشان دهنده ی وضعیت مراقبت نیستند).

3. ابتدا سعی کنید برای پیدا کردن یک بازی کامل و فیلد Class از معامله را با کلاس قوانین همسان پر کنید.

4. اگر هماهنگی کامل بین قواعد هماهنگ وجود نداشته باشد، قاعده ای که دارای بالاترین سطح ریسک است انتخاب می شود و فیلد کلاس از معامله با آن دسته از قانون هماهنگ پر می شود.

4. نتایج

4-1. مجموعه داده

داده های مورد استفاده در این مقاله داده های واقعی جهان است که توسط یک بانک ملی متشکل شده است. همانطور که داده ها حاوی اطلاعات حساس هستند، پایگاه داده نمی تواند به عنوان توافق با بانک مشخص شود. حدود 1 کارت اعتباری LAC بر اساس موارد تقلبی مختلف استفاده می شود. سپس تراکنش ها به مجموعه های مختلف آزمون تقسیم می شوند.

طبقه بندی شده با مجموعه های معامله گوناگون آموزش دیده و برای طبقه بندی هر یک از این مجموعه ها مورد استفاده قرار می گیرد. برای مقایسه، الگوریتم اولیه C4.5 و الگوریتم MLPC برای آموزش استفاده می شود. همانگونه که کلاس های این معاملات قبلا شناخته شده است، دقت طبقه بندی با مقایسه معاملات تراکنش با ارزش کلاس اصلی معاملات ارزیابی می شود. معیارهای طبقه بندی مورد استفاده برای ارزیابی نتایج عبارتند از: نرخ واقعی مثبت (TPR)، نرخ مثبت کاذب (FPR)، نرخ منفی اشتباه (FNR)و دقت [2],[15]. C4.5 یک الگوریتم طبقه بندی اساسی درخت تصمیم است. سیستم تشخیص تقلب کارت اعتباری یکی از کاربردهای آن است که توسعه یافته است. این نرم افزار برای بین بانکی مفید است که بانک ها می توانند قوانین مربوط به تقلب خود را با یکدیگر تعریف کنند تا از تهدید تقلب که به طور وسیع در جهان کارت های اعتباری گسترش می یابد، غلبه کند.

اقدامات طبقه بندی برای ارزیابی دقت

دقت طبقه بندی در صورت شناسایی تقلب کارت اعتباری براساس معیارهای زیر ارزیابی می شود:

نرخ واقعی مثبت (TPR)

نرخ مثبت کاذب (FPR)

نرخ منفی نادرست (FNR)

دقت (Accuracy)

4-2. نتایج طبقه بندی شده در سطوح مختلف هرس

از کل مجموعه معامله، برخی از تراکنش ها برای آموزش گرفته می شود و بخشی از آن برای آزمایش تلقی می شود و سپس این روش برای کل پایگاه داده تراکنش تکرار می شود. نتایج با هر دو الگوریتم آموزش C4.5 و MLPC در تمام ترکیب مجموعه ها ارزیابی می شوند، اما با توجه به محدودیت فضا، تنها برخی از نتایج ذکر شده و مقایسه می شوند. مشخصات مجموعه معامله در جدول 1 آمده است. روش قبل از بریدن در سیستم توسعه یافته، که در آن درخت در حال رشد در سطح خاص متوقف شده است برای جلوگیری از تشکیل حکم بی معنی و یا قوانین خاص تر. نتایج توسط ضریب اصلاح درخت در حال رشد در سطوح مختلف و دقت در هر سطح برش مقایسه شده است. جدول 2 مقايسه دقت طبقه بندي با سطوح مختلف لوبيا را براي 4 سطح خطر L1، L2، L3، L4 به عنوان ارزش طبيعي با Set1 به عنوان تست file و Main Set as file آموزش نشان مي دهد.

جدول 2، ارزیابی دقت را در هر سطح خاك نشان می دهد.

نتایج نشان می دهد که زمانی که سطح خراش کم است، به معنی داشتن تعداد کمی از صفات برای طبقه بندی، دقت نوع کلاس L1 بسیار بالا است

زیرا بیشتر معاملات به عنوان غیر فریبنده طبقه بندی می شوند که دقت کمتری از نوع دیگر کلاس را نشان می دهد که سطح تقلبی را نشان می دهد.

در حالی که با افزایش درجه لکه با در نظر گرفتن ویژگی های بیشتر برای طبقه بندی، دقت نوع کلاس L1 کمی کاهش می یابد، اما دقت انواع دیگر کلاس ها بهبود می یابد. شکل 2 همان نتیجه را به صورت گرافیکی نشان می دهد.

جدول 1- مشخصات هر مجموعه معامله

|  |  |
| --- | --- |
| No. Of Transactions | Test Set Name |
| 10000 | Test1 |
| 20000 | Test2 |
| 30000 | Test3 |
| 40000 | Test4 |
| 10000 | Test5 |
| 20000 | Test6 |
| 30000 | Test7 |
| 40000 | Test8 |
| 50780 | Set1 |
| 50780 | Set2 |
| 101560 | Main set |

جدول 2- ارزیابی دقت کلاس Class vise در هر سطح Prune

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| L4 | L3 | L2 | L1 | Prune Level |
| 75.79 | 79.12 | 70.51 | 88.84 | 1 |
| 75.79 | 96.03 | 70.51 | 87.65 | 2 |
| 75.79 | 96.03 | 75.32 | 87.51 | 3 |
| 75.79 | 96.03 | 80.6 | 85.38 | 4 |
| 75.79 | 82.35 | 72.53 | 88.69 | 5 |
| 75.79 | 82.35 | 81.2 | 88.43 | 6 |
| 75.79 | 82.35 | 82.1 | 88.43 | 7 |
| 75.79 | 80.73 | 82.32 | 86.6 | 8 |
| 75.79 | 79.12 | 82.33 | 87.57 | 9 |
| 75.79 | 79.12 | 82.33 | 86.74 | 10 |
| 75.79 | 79.12 | 82.32 | 86.74 | 11 |

جدول 3- مقايسه دقت، TPR، FPR و FNR براي سطوح مختلف

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy | TPR | FNR | FPR | Prune Level |
| 79.24 | 80.02 | 19.98 | 11.16 | 1 |
| 81.72 | 84.55 | 15.45 | 12.35 | 2 |
| 82.9 | 86.49 | 13.51 | 12.49 | 3 |
| 83.55 | 88.47 | 11.53 | 14.62 | 4 |
| 80.25 | 81.59 | 18.41 | 11.31 | 5 |
| 82.38 | 85.22 | 14.78 | 11.57 | 6 |
| 82.61 | 85.62 | 14.38 | 11.57 | 7 |
| 81.79 | 85.77 | 14.23 | 13.4 | 8 |
| 81.84 | 85.77 | 14.23 | 12.43 | 9 |
| 81.57 | 85.77 | 14.23 | 13.26 | 10 |
| 81.57 | 85.77 | 14.23 | 13.26 | 11 |

نتایج نشان می دهد که FPR ارائه شده توسط سطوح متوسط خوب است، هرچند که پایین ترین سطح به پایین ترین FPR می دهد، آن را به طور کلی دقت نمی دهد. همچنین سطوح قاعدگی متوسط، TPR بهتر از سایر سطوح را فراهم می کند.

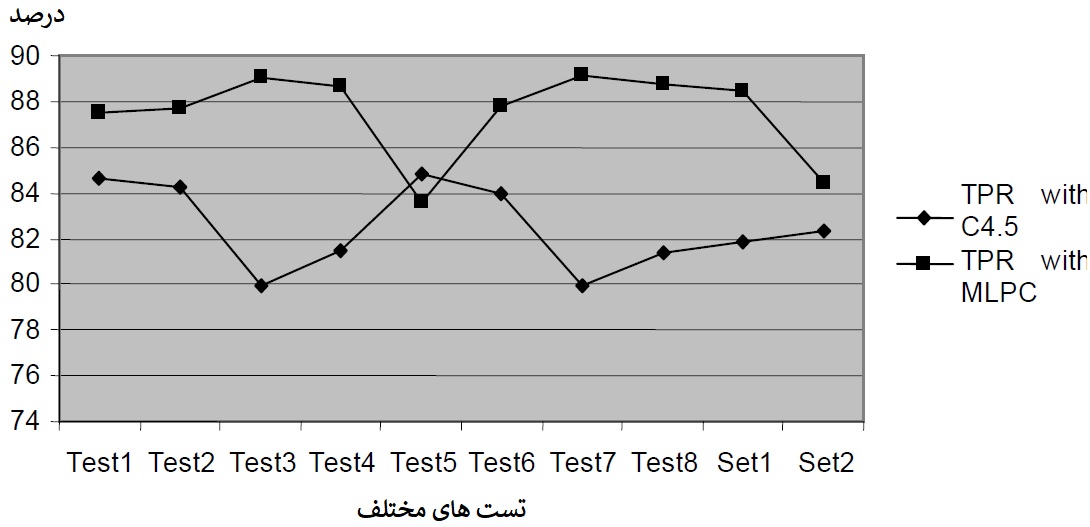
4-3. مقايسه اندازه های جدول تصميم

الگوریتم C4.5 درخت تصمیم گیری برای تمام ترکیبات هر ویژگی رشد می کند، سطح درخت تصمیم گیری بالا است که در شرایط تولید تعداد زیادی از قوانین تصمیم گیری در حالی که MLPC تنها ویژگی هایی را که بالاترین سهم در طبقه بندی را در نظر می گیرند در نظر می گیرد بنابراین تعداد از قوانین تولید شده کمتر است. این کمک می کند در داده کاوی توزیع که در آن داده های یک شرکت در سایت های مختلف توزیع شده و به جای گذشتن از داده ها آنها قوانین تولید شده از پایگاه داده به اشتراک گذاری. در اینجا، اگر اندازه جدول تصمیم گیری بالا باشد، پهنای باند شبکه بیشتر خواهد شد و زمان انتقال نیز بالا خواهد بود. نتایج در جدول 4 نشان می دهد که تعدادی از قوانین تولید شده توسط الگوریتم MLPC 10 برابر کمتر از الگوریتم C4.5 است.

جدول 4- مقایسه مقیاس جدول تصمیم گیری

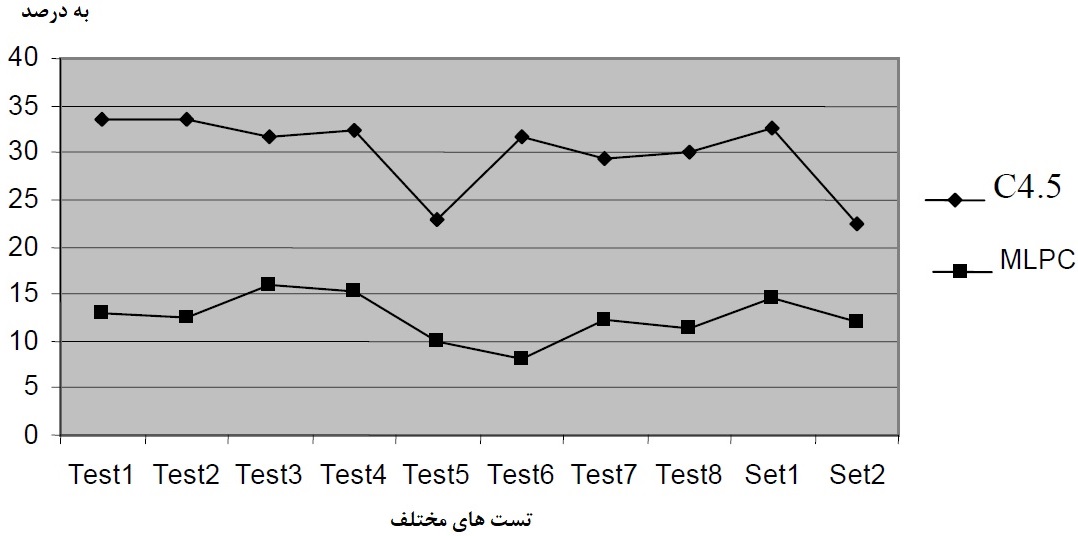
|  |  |  |
| --- | --- | --- |
| No. Of Decision rules generated | | Training File |
| MLPC | C4.5 |
| 15 | 134 | Test1 |
| 12 | 131 | Test2 |
| 16 | 237 | Test3 |
| 15 | 238 | Test4 |
| 15 | 96 | Test5 |
| 15 | 95 | Test6 |
| 18 | 156 | Test7 |
| 15 | 156 | Test8 |
| 15 | 237 | Set1 |
| 15 | 156 | Set2 |
| 16 | 207 | Main Set |

4-4. مقایسه مقیاس دقت و طبقه بندی با Main\_set به عنوان فایل آموزشی



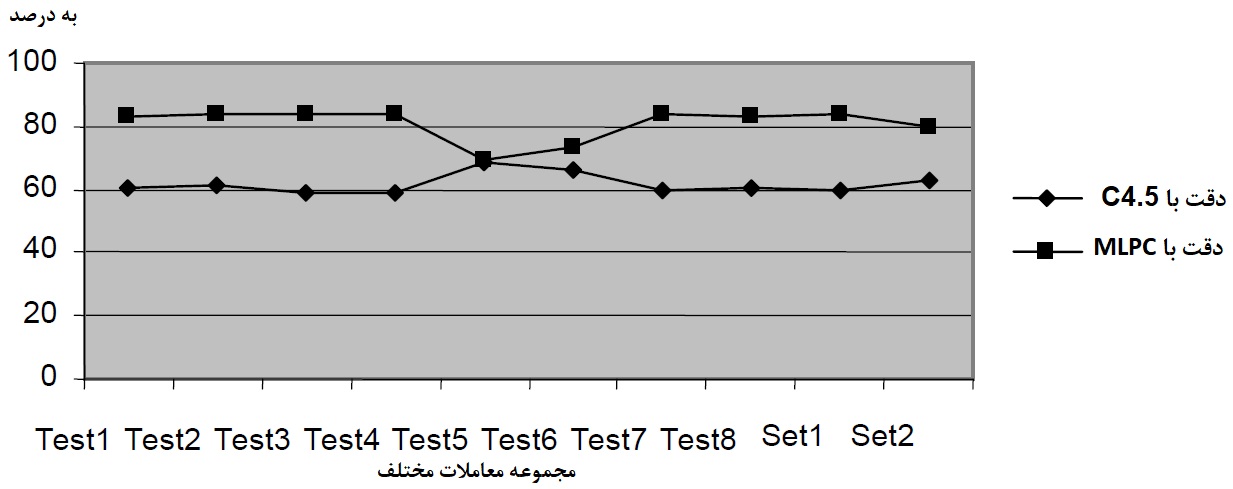
شکل 2- مقایسه نرخ مثبت واقعی (TPR)

شکل 2 نشان می دهد که MLPC به طور متوسط 87% از TPR را نشان می دهد در حالی که C4.5 به میزان 82% از TPR می دهد. شکل 3، نتایج مقایسهای از میزان امتیاز مثبت کاذب با الگوریتم C4.5 و MLPC را ارائه می دهد. نتایج نشان می دهد که MLPC باعث کاهش FPR در حدود 12% نسبت به C4.5 می شود که میانگین FPR 30% را نشان می دهد.



شکل 3- مقایسه میزان امتیاز مثبت کاذب

شکل 4 نتایج حاصل از ارزیابی کلی دقت را نشان می دهد. دقت کلی هر معامله طبقه بندی شده به سطح صحیح کلاس بسیار اهمیت دارد برای رسیدن به اهداف سیستم. با توجه به مجموعه اصلی اصلی به عنوان طبقه بندی پایه و طبقه بندی مجموعه های داده های مختلف، MLPC دارای بالاترین دقت متوسط 80% است که بسیار بهتر از C4.5 است که دقت کلی 62% را نشان می دهد.



شکل 4- مقایسه دقت کلی

5. نتیجه گیری ها

در این مقاله، الگوریتم MLPC که فاز برش را در فاز ساختمان ادغام می کند، پیشنهاد شده است. در رسیدن به تنظیم گره سطح خراش در مرحله ساخت گسترش نمی یابد. در نتیجه، گره های کمتر در مرحله ساخت گسترش می یابند و بنابراین پیچیدگی ساخت درخت تصمیم گیری کاهش می یابد.

تعدادی از قوانین تصمیم گیری شده توسط MLPC خیلی کمتر از قوانین تولید شده توسط الگوریتم C4.5 است که در جدول 4 ذکر شده است. بسیاری از قوانین تولید شده توسط C4.5 از کار بی نظیر و بی معنی است. در الگوریتم MLPC، قوانین کمتری دارند که به طور مستقیم بر اندازه جدول تصمیم تاثیر می گذارد. این به طور مستقیم پیچیدگی زمانی طبقه بندی را کاهش می دهد. به عنوان اندازه جدول تصمیم گیری تولید شده با الگوریتم MLPC کوچک است، پهنای باند شبکه مورد نیاز در هنگام انتقال جدول تصمیم نیز کاهش می یابد. بنابراین تکنیک هرس اثبات مؤثر و مقیاس پذیر در القاء درخت تصمیم گیری است.

عملکرد مبتنی بر دقت و درستی میزان مثبت بین الگوریتم اساسی C4.5 و الگوریتم جدید MLPC توسعه یافته مقایسه شده است. MLPC به طور متوسط ​​80% دقت می کند، در حالی که الگوریتم C4.5 به طور متوسط ​​به میزان 62% دقت می دهد. بنابراین الگوریتم هرس از دیدگاه دقت طبقه بندی مؤثر است.

 نرخ تسخیر تکرار (TPR) هر دو طبقه بندی به ترتیب در هر رقم 85% است. 2. میزان خطای هشدار اشتباه (FPR) MLPC 12% است و C4.5 نرخ هشدار اشتباه 30% را در شکل 3 نشان می دهد. الگوریتم MLPC درخت تصمیم گیری است الگوریتم یادگیری با قبل از هرس کردن. مشاهدات نشان می دهد که در سطح 4، بالاترین دقت را برای مجموعه های معامله مختلف برنامه می دهد. اما بعد از آن هم همیشه سطح هرس "مطلوب" برای برنامه های مختلف و الزامات وجود دارد که باید شناسایی و انتخاب کنید.

سیستم تشخیص تقلب کارت اعتباری یکی از برنامه های کاربردی MLPC است که توسعه یافته است. در مقایسه با سیستم های تشخیص تقلب در کارت اعتباری قبلا توسعه یافته بود که در آن معاملات فقط در دو سطح یا تقلب و یا عدم تقلب تقسیم بندی می شد، سیستم توسعه یافته می تواند در میان شرایط تقلبی متفاوت متمایز شود و معاملات را در چهار سطح خطر تقلب تقسیم کند.

 کار جاری ما بر روی ترکیب و آزمایش راه حل های مختلف برش مورد بحث در مقاله در الگوریتم پیشنهاد شده متمرکز شده است. هدف این است که الگوریتم هورمونی مقرون به صرفه را توسعه دهیم که تا کنون کارهای زیادی انجام نداده است.

6. مراجع

[1] Jiawei Han، Micheline Kamber، "مفاهیم و تکنیک های داده کاوی"، صفحات 279-328، 2001.

[2] Doina Caragea، "طبقه بندی های یادگیری از منابع داده توزیع شده، معنایی ناهمگن، داده های مستقل داده ها"، پایان نامه، دانشگاه ایالتی آیووا ، 2004.

[3] تام. مایکل میچل، "یادگیری ماشین"، انتشارات McGraw-Hill، 1997   
[4] ژانگ یونگ، "الگوریتم برش بر اساس تصمیم گیری درختان بر اساس کمیت مجموعه داده ها"، در مقالات ششمین کنفرانس بین المللی رایانه، برنامه های کاربردی و توزیع شده تکنولوژی، 2005.

[5] آرون پوجیری، "تکنیک های داده کاوی"، pp 150-200، 1999.

[6] Manish Mehta، Rakesh Agrawal et al: "SLIQ - طبقه بندی سریع مقیاس پذیر برای داده کاوی."، در 5th Intl . Conf. در گسترش فناوری پایگاه داده، مارس

[7] J. Quinlan برنامه C4.5 برای یادگیری ماشین، San Mateo، CA: Morgan Kaufmann، 1993-1996.

[8] Salvatore، Philip et al: "عوامل یادگیری متقابل برای تشخیص تقلب و نفوذ در سیستم های اطلاعات مالی"، مقاله پژوهشی مقالات در کنفرانس بین المللی از دانش کشف و داده کاوی، 1996.

[9] S. Stolfo و همکاران، "JAM: نمایندگان جاوا برای مطالعه فلزی از پایگاه داده های توزیع شده،" Proc. سومین اینترانت کشف دانش و داده کاوی، AAAI Press، Menlo Park، Calif، 1997، pp. 74-81.

[10] Salvatore J. Stolfo، David W. Fan، Wenke Lee و Andreas L. Prodromidis، "تشخیص تقلب کارت اعتباری با استفاده از متال آموزش: مسائل و نتایج اولیه"، DARPA، 1999.

[11] جین کینلان. ساده سازی تصمیم درختان، بین المللی J. Human Computer Studies، (1999) 51، pp. 497-491، 1999.

[12] F. Esposito، D. Malerba، و G. Semeraro. تجزیه و تحلیل تطبیقی روش های درخت تصمیم گیری هرس، معاملات IEEE در تجزیه و تحلیل الگوی و هوش ماشین، 19 (5): pp. 476-491، 1997.

[13] L. Briemann، J. Friedman، R. Olshen و C. Stone. CART: طبقه بندی و رگرسیون درختان، EBelmont، CA: Wadsworth آمار آماری، 1984.

[14] B. Cestnik، و I. Bratko. برآورد احتمال ها در هرس درخت، EWSL، pp. 138-150، 1991

[15] Esposito F.، Malerba D.، Semeraro G. تجزیه و تحلیل تطبیقی روش های درختان تصمیم گیری هرس، معاملات IEEE در تجزیه و تحلیل الگوی و هوش ماشین، VOL. 19، NO 5، 1997، ص 476-491

[16] J. Mingers. مقایسهای تجربی از روشهای انتخاب برای القاء درخت تصمیم گیری، یادگیری ماشین، 3 (3): pp. 319-342، 1989

[17] J. Mingers. یک مقایسه تجربی از روش های برش برای القاء درخت تصمیم گیری، یادگیری ماشین، 4: ص 227243، 1989

[18] L. A. Breslow و D. W. Aha. ساده سازی درخت های تصمیم گیری: یک نظرسنجی، گزارش فنی شماره AIC-96-014، مرکز نیروی دریایی تحقیقات کاربردی در هوش مصنوعی، آزمایشگاه تحقیقات دریایی واشنگتن دی سی، 1996.

[19] I. Bratko و M. Bohanec. دقت تجاری برای سادگی در درخت تصمیم گیری، ماخ. یادگیری، حجم 15، صص 223-250، 1994.

[20]H. Almuallim. الگوریتم کارآمد برای هرس کردن بهینه درخت تصمیم، Artif. Intell.، vol. 83، نه 2، pp. 347-362، 1996.

[21]M. I. Jordan. یک رویکرد آماری برای تصمیم گیری مدل سازی درخت، مقالات هفتمین کنفرانس سالانه ACM تئوری یادگیری محاسباتی، نیویورک: ACM Press، 1994.

[22] B. اندرو، L. Brain C.، حساس به هزینه حساس درخت درخت هرس استفاده از منحنی ROC، جلسات هشت کنفرانس مشترک در مورد هوش مصنوعی استرالیا، pp.18، نوامبر 1995.

1. **Corresponding author: ALI SOLDOOZY**

   **Email: soldoozy@gmail.com** [↑](#footnote-ref-1)