مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۵، شماره‌ی ۲، پاییز و زمستان ۱۳۹۳، صص ۱۱۵ – ۱۴۸

درخت تصمیم داده‌های نامطمئن (مطالعه‌ی موردی داده‌های نامطمئن طرح اطلاعات اقتصادی خانوار)

مهسا قائمی[†,*]، میرمحسن پدرام[‡] و عادل آذر[★]

[†] دانشگاه آزاد اسلامی، واحد علوم و تحقیقات

[‡] دانشگاه خوارزمی تهران

[★] دانشگاه تربیت مدرس و مرکز آمار ایران

چکیده: درخت تصمیم یکی از تکنیک‌های بسیار رایج در طبقه‌بندی داده‌ها است. در این مقاله درخت تصمیم داده‌های نامطمئن مورد بررسی قرار گرفته است. از عواملی که سبب عدم اطمینان در داده‌ها می‌شوند می‌توان به محدودیت در دقت اندازه‌گیری، منابع قدیمی، اظهار نشدن اطلاعات و مسائلی که در انتقال داده‌ها بوجود می‌آید اشاره نمود. در داده‌های نامطمئن، مقدار داده با یک مقدار مشخص، نشان داده نمی‌شود و با چند مقدار به شکل توزیع احتمالی نشان داده می‌شود. داده‌های طرح اطلاعات اقتصادی خانوار نیز به‌دلیل کم‌گویی یا نبود برخی از داده‌ها، در دسته‌ی داده‌های نامطمئن قرار می‌گیرند، بنابراین لازم است که از الگوریتمی استفاده شود که بتواند با داده‌های نامطمئن کار کرده و با دقت قابل قبولی طبقه‌بندی داده‌ها را انجام دهد. در این مقاله، الگوریتم درخت تصمیم نامطمئن پیشین تعمیم داده شده است. این الگوریتم از روش‌های پیش‌بینی مثل نرخ بهره و آنتروپی و همچنین داده‌های نامطمئن بازه‌ای استفاده می‌کند و توانسته است با استفاده از توابع چگالی احتمال متفاوت سبب کاهش اثر داده‌های نامتوازن در خروجی الگوریتم شود. این الگوریتم برای هر دو مجموعه داده‌های مطمئن و نامطمئن کار می‌کند و نتایج این مقاله نشان می‌دهد که الگوریتم پیشنهادی، دقت پیش‌بینی رضایت بخشی دارد. ساخت درخت تصمیم داده‌های نامطمئن، حجم پردازش بیش‌تری را در پردازنده نسبت به ساخت درخت روی داده‌های مطمئن اشغال می‌کند، بنابراین در الگوریتم پیشنهادی از تکنیک ماکسیمم

---

* نویسنده‌ی عهده‌دار مکاتبات

دریافت: ۱۳۹۲/۱۲/۶، پذیرش: ۱۳۹۴/۶/۲۴.

سطح استفاده می‌شود که مصرف پردازنده را بهینه خواهد کرد.

واژگان کلیدی: داده‌ی نامطمئن؛ درخت تصمیم؛ طبقه‌بندی؛ داده‌کاوی.

۱- مقدمه

تعداد مقالاتی که در زمینه‌ی داده‌های نامطمئن و پایگاه داده‌های احتمالی در دهه‌های اخیر نوشته شده است نشان می‌دهد که این حوزه مورد توجه بسیاری از محققین قرار گرفته و کار بر روی زمینه‌های متفاوت آن تازه شروع شده است. داده‌های نامطمئن زمینه‌ای است که جامعه‌ی محققان هوش مصنوعی در سال‌های اخیر بسیار بر روی آن کار کرده‌اند.

الگوریتم‌های داده‌کاوی به افراد در زمینه‌ی تحلیل داده‌های خام، استخراج الگوهای پنهان، تصمیم‌گیری و یافتن دانش[۱] کمک می‌کند [۱۰]. در داده‌کاوی دو نوع جهت‌گیری متفاوت وجود دارد. اولین نوع، یادگیری نظارت‌شده[۲] است، که اصلی‌ترین کار آن یادگیری کشف ارتباط میان ویژگی‌ها و کلاس مقصد است. نوع دوم یادگیری نظارت‌نشده[۳] است که داده در این حالت کلاس هدف مستقلی ندارد و روش یادگیری، روابط قطعی میان داده‌ها و ویژگی‌ها را کشف می‌کند.

نظر به کاربرد بسیار زیاد طبقه‌بندی در پدیده‌های واقعی، این روش یکی از رویکردهای اصلی داده‌کاوی در بیش‌تر دهه‌ها بوده است. در این مقاله، طبقه‌بندی داده‌ها مورد مطالعه قرار گرفته و درخت تصمیم به‌عنوان یکی از محبوب‌ترین روش‌های طبقه‌بندی مورد توجه قرار گرفته است. لازم به یادآوری است که، روش‌های طبقه‌بندی زیادی توسط دانشمندان ارایه شده است، که می‌توان به روش k- نزدیک‌ترین همسایگی[۴]، شبکه‌ی عصبی مصنوعی[۵]، درخت تصمیم[۶] و ماشین بردار پشتیبان[۷] اشاره کرد.

درخت تصمیم یکی از روش‌های بسیار محبوب و پرکاربرد برای طبقه‌بندی داده‌ها است. زیرا، به‌دلیل دارا بودن ساختاری کاربردی و قابل فهم، به راحتی می‌توان قوانین لازم را از آن استخراج کرد. الگوریتم‌های زیادی مثل ID۳ و C۴/۵ برای ساخت درخت تصمیم ایجاد شده‌اند. این الگوریتم‌ها در بسیاری از برنامه‌ها و نرم‌افزارها مانند پردازش تصویر، تشخیص بیماری، شناخت نظریه‌ی فروید و بازاریابی مورد استفاده قرار گرفته‌اند [۱۹].

از آن‌جا که عدم اطمینان داده‌ها در بسیاری از شرایط وجود دارد، یافتن روشی به‌منظور طبقه‌بندی این داده‌ها ضروری به نظر می‌رسد. در این مقاله درخت تصمیم برای طبقه‌بندی

داده‌های نامطمئن طرح اطلاعات اقتصادی خانوار ارایه شده است. در بسیاری از داده‌ها عدم اطمینان به‌صورت ساختاری وجود دارد. برخی از عواملی که سبب عدم اطمینان می‌شوند عبارت‌اند از: اندازه‌گیری‌های غلط، داده‌های غیر قابل اعتماد و از بین رفتن داده‌ها [۱۰]. به‌دلیل وجود عدم اطمینان در داده‌ها، ارایه‌ی مدلی جهت طبقه‌بندی ضروری به نظر می‌رسد. در این مقاله استفاده از درخت تصمیم برای طبقه‌بندی داده‌های نامطمئن بررسی شده است. استفاده از این روش با استناد بر نقاط قوت درخت تصمیم (به [۱۷] مراجعه شود) انجام شده است.

درخت تصمیم ساختار ساده و قابل درکی دارد و نیاز زیادی به آماده‌سازی داده‌ها ندارد در صورتی که سایر روش‌ها معمولاً نیاز به نرمال‌سازی داده‌ها دارند. درخت تصمیم را می‌توان بر روی هر دو نوع داده‌ی عددی و رسته‌ای اعمال کرد، در صورتی که برخی روش‌ها فقط با نوع داده‌ی خاصی کار می‌کنند. درخت تصمیم از مدل جعبه‌ی سفید استفاده می‌کند که شفافیت داشته و مدلی بسیار قوی است که قابلیت مقیاس‌بندی دارد و همچنن داده‌های بزرگ را در زمان کمی پردازش می‌کند [۱۶].

در این مقاله، درخت تصمیم جدیدی برای طبقه‌بندی داده‌های مطمئن و نامطمئن ارایه شده است. این درخت به‌منظور طبقه‌بندی داده‌های طرح اطلاعات اقتصادی خانوار استفاده شده است. نظر به این که در طرح اطلاعات اقتصادی خانوار برخی افراد به هنگام تکمیل پرسشنامه اطلاعات خود را به‌صورت کامل ارایه نداده‌اند و با مقایسه‌ی این داده‌ها با برخی داده‌های مطمئن که از طریق نهادهای کشوری در اختیار مرکز آمار ایران قرار داده شده است، عدم اطمینان داده‌ها مشخص شده است. بنابراین به‌منظور جداسازی افراد و تشخیص تمکن مالی آن‌ها الگوریتمی دقیق و قابل قبول به‌منظور طبقه‌بندی داده‌ها نیاز است. یکی از الگوریتم‌های کارا و مناسب به‌منظور طبقه‌بندی داده‌ها الگوریتم درخت تصمیم است که برای داده‌های نامطمئن این الگوریتم با نام درخت تصمیم داده‌های نامطمئن ارایه خواهد شد. به‌منظور ساخت این درخت، الگوریتمی مشابه درخت تصمیم سنتی وجود دارد، با این تفاوت که در این الگوریتم همه چیز به‌صورت احتمالی ارایه خواهد شد.

داده‌های طرح اطلاعات اقتصادی خانوار به دو دسته‌ی داده‌های خوداظهاری خانوار مربوط به سال ۸۸ و داده‌های دریافتی از نهادهای دولتی مربوط به سال ۹۱ تقسیم می‌شوند که از طریق مرکز آمار ایران به‌منظور انجام این طرح در اختیار قرار گرفته‌اند.

داده‌های طرح اطلاعات اقتصادی خانوار داده‌هایی با عدم اطمینان زیاد و همچنین نامتوازن هستند، بنابراین به‌منظور کار با این داده، دو عامل فوق باید مورد توجه ویژه قرار گیرند. درخت تصمیم داده‌های نامطمئن پیش‌نهاد شده در این مقاله، الگوریتم عمومی برای حل مسئله‌ی طبقه‌بندی داده‌های نامطمئن است که توانسته است نتایج مطلوبی در طبقه‌بندی داده‌های طرح اطلاعات اقتصادی خانوار داشته باشد.

ساختار کلی مقاله به‌صورت زیر است: در بخش ۲، مرور نوشتگان ارایه خواهد شد. در بخش ۳، تعاریف و مفاهیم لازم جهت ورود به مبحث اصلی ارایه خواهد شد. در بخش ۴، روش پیشنهادی این مقاله مورد بررسی قرار خواهد گرفت. در بخش ۵، مجموعه داده‌ها و موارد و مشکلات مربوط به آن توضیح داده خواهد شد. در بخش ۶ نتایج آزمایش‌های انجام‌شده مطرح می‌شود و در بخش ۷ نیز، نتیجه‌گیری کلی ارایه خواهد شد.

۲- مرور نوشتگان

در سال‌های اخیر، روش‌های غیرمستقیم جمع‌آوری داده منجر به ایجاد داده‌های نامطمئن شده است. برای مثال، شبکه‌های سنسوری، داده‌های نامطمئن زیادی را تولید می‌کنند. به‌صورت مشابه، داده‌های طرح‌های آماری و سرشماری‌ها نیز به‌صورت ذاتی مقادیر نامطمئن دارند [۲]. مطالعات بسیار گسترده‌ای در حوزه‌ی داده‌های نامطمئن انجام شده است که از مهم‌ترین حوزه‌های این مبحث می‌توان به سه مورد زیر اشاره کرد [۳، ۹]:

۱- مدل‌سازی داده‌های نامطمئن: فرایند مدل‌سازی داده‌های نامطمئن مسئله‌ی اصلی سیستم‌های نامطمئن است، بنابراین در مدل‌سازی، باید پیچیدگی‌های داده را به‌منظور مدیریت پایگاه داده در نظر گرفت.

۲- مدیریت داده‌های نامطمئن: در این حالت، هدف این است که از روش‌های مدیریت سنتی پایگاه داده استفاده شود، این روش‌ها می‌توانند پردازش اتصال[۸]، پردازش پرس‌وجو[۹]، شاخص‌گذاری[۱۰] و مجتمع سازی[۱۱] پایگاه داده باشد.

۳- کاوش در داده‌های نامطمئن: در این حالت، روش‌هایی به‌منظور استخراج گزارش‌ها و نتایج وجود دارد. در کاوش داده‌های نامطمئن نتایج حاصل از برنامه‌های داده‌کاوی تحت تأثیر عدم اطمینان در داده‌ها قرار می‌گیرند. بنابراین، ضروری است که روش‌های داده‌کاوی برای رفتار با داده‌های نامطمئن طراحی شود. از مهم‌ترین

روش‌های کاوش در داده‌ها می‌توان به روش طبقه‌بندی و خوشه‌بندی اطلاعات اشاره کرد.

مطالعات انجام‌شده در حوزه‌ی داده‌های نامطمئن تمرکز بیش‌تری بر روی خوشه‌بندی این داده‌ها داشته‌اند [۸، ۱۳، ۱۵]. در مقاله‌ی [۵]، الگوریتم کاربردی خوشه‌بندی k – میانگین به UK – میانگین توسعه داده شده است و از آن برای خوشه‌بندی داده‌های نامطمئن استفاده می‌شود. علاوه بر این، ابعاد دیگر مورد مطالعه در این زمینه شامل خوشه‌بندی بر مبنای چگالی[۱۲][۱۳]، کاوش داده‌های تکراری[۱۳] [۷] و طبقه‌بندی بر مبنای چگالی[۱۴] [۱] است. در طبقه‌بندی بر مبنای چگالی نیاز است که توزیع احتمالی اتصالات ویژگی‌های داده، وجود داشته باشد.

طبقه‌بندی یکی از مسائل اصلی داده‌کاوی است. به هنگام اجرای الگوریتم طبقه‌بندی داده‌ها می‌توان داده‌ها را به دو گروه داده‌ی آموزشی و آزمونی تقسیم کرد و از این داده‌ها به‌منظور آموزش و آزمون استفاده نمود. در داده‌ی آموزشی، هر داده کلاس مشخصی دارد و باید با الگوریتم مشخص، مدلی برای پیش‌بینی برچسب کلاس‌های مربوط به داده‌های جدید یافت. هدف طبقه‌بندی، متعلق نمودن نمونه‌ی آزمون به یک برچسب از مجموعه‌ی برچسب‌ها است. الگوریتم‌های طبقه‌بندی زیادی در مقالات مختلف ارایه شده است که می‌توان به الگوریتم‌های درخت تصمیم [۱۲]، طبقه‌بندی بیزی [۱۴]، ماشین بردار پشتیبان (SVM) [۲۰] و شبکه‌ی عصبی مصنوعی [۲] اشاره کرد.

با وجود الگوریتم‌های زیاد طبقه‌بندی، طبقه‌بندی بر پایه‌ی داده‌های نامطمئن یک چالش اساسی است. در مطالعات اخیر، توسعه‌ی درخت تصمیم برای داده‌هایی با مقادیر گمشده یا دارای نوفه انجام شده [۱۱، ۱۸] و همچنین رویکردهای زیادی برای پیش‌بینی یا جانهی مقادیر گمشده ارایه شده است. تفاوت مطالعه‌ی انجام‌شده در این مقاله در این است که همه‌ی داده‌های موجود نامطمئن هستند و با مطالعات پیشین که بخشی از داده‌ها دارای مقادیر گمشده یا دارای نوفه هستند متفاوت خواهد بود. عدم اطمینان داده‌های طرح اطلاعات اقتصادی خانوار به‌صورت مقادیر گمشده یا نادرست نیست و به‌صورت فواصل نامطمئن و توابع توزیع احتمالی وجود دارد. طبقه‌بندی داده‌های نامطمئن در مقاله [۴] نیز مطالعه شده است که در این مقاله داده‌های آزمونی انتخاب شده و سعی در حل کردن مسائل طبقه‌بندی خاصی داشته است. بنابراین الگوریتم عمومی برای طبقه‌بندی داده‌های نامطمئن تا کنون ارایه نشده است.

۳- تعاریف و مفاهیم

۳-۱- عدم اطمینان در داده‌ها

در این بخش، عدم اطمینان در داده‌ها برای ویژگی‌های عددی و رسته‌ای بررسی خواهد شد. داده‌های مورد مطالعه در این مقاله داده‌هایی با ویژگی‌های نامطمئن و کلاس خروجی مطمئن بودند.

اگر در مجموعه‌ی داده‌های عددی[۱۵] نسبت به دقت و صحت یک یا چند ویژگی اطمینان وجود نداشته باشد، آن ویژگی، «ویژگی عددی نامطمئن» (UNA)[۱۶] گفته می‌شود که با $A_i^{u_n}$ نشان داده می‌شود [۱۶]. از مهم‌ترین ویژگی‌های داده‌های عددی نامطمئن، مقادیر پویا آن‌ها است. پویایی در مقادیر به این معنی است که هر پارامتر به جای مقدار ثابت و مشخص در یک بازه ذخیره خواهد شد [۶]. بنابراین از $A_{ij}^{u_n}$ برای مشخص کردن jامین نمونه $A_i^{u_n}$ استفاده می‌شود. در این مقاله با روش‌های مختلف آماری و استفاده از توابع چگالی احتمال متفاوت که سبب کاهش اثر داده‌های نامتعادل[۱۷] مورد مطالعه می‌شود، بازه‌ای مناسب حول عدد مربوطه ساخته خواهد شد و مقدار $A_i^{u_n}$ به‌صورت بازه‌ای ذخیره خواهد شد. نمونه‌ای که بازه‌ی $A_i^{u_n}$ دارد با $A_{ij}^{u_n}.U$ مشخص می‌شود که در بازه‌ی $[A_{ij}^{u_n}.l\,, A_{ij}^{u_n}.r]$ است و $A_{ij}^{u_n}.l\,, A_{ij}^{u_n}.r \in R$ و $A_{ij}^{u_n}.r \geq A_{ij}^{u_n}.l$ است. L و r حدود چپ و راست بازه را مشخص می‌کنند. تابع چگالی احتمال نامطمئن $A_{ij}^{u_n}$ با $A_{ij}^{u_n}.f(x)$ مشخص خواهد شد و اگر $x \in A_{ij}^{u_n}.U$ است حاصل انتگرال برابر $\int_{A_{ij}^{u_n}.l}^{A_{ij}^{u_n}.r} A_{ij}^{u_n}.f(x)dx = ۱$ و در صورتی که $x \notin A_{ij}^{u_n}.U$ است حاصل انتگرال برابر خواهد بود با $\int_{A_{ij}^{u_n}.l}^{A_{ij}^{u_n}.r} A_{ij}^{u_n}.f(x)dx = ۰$.

اگر در مجموعه‌ی داده‌ی رسته‌ای[۱۸] نسبت به دقت و صحت یک یا چند ویژگی اطمینان وجود نداشته باشد، آن ویژگی، «ویژگی‌های رسته‌ای نامطمئن» (UCA)[۱۹] گفته می‌شود، که با $A_i^{u_c}$ مشخص می‌شود. از $A_{ij}^{u_c}$ برای مشخص کردن jامین نمونه‌ی $A_i^{u_c}$ استفاده خواهد شد. به‌منظور اعمال عدم اطمینان در داده‌های رسته‌ای نامطمئن، برای نمونه $A_{ij}^{u_c}$ مقادیر از دامنه رسته‌ای Dom[۲۰] با کاردینالیتی |Dom|=n گرفته خواهد شد. به‌عنوان مثال در مجموعه‌ی داده‌های مطمئن، اگر داده‌ای با مشخصه‌ی «جنسیت»، دارای مقدار «مرد» باشد، مرد $= A_{ij}^{u_c}$ و ۲ =|Dom| با مقادیر {مرد و زن} خواهد بود. بنابراین مقدار

ویژگی، احتمالی برابر با عدد یک در دامنه دارد و $P_۲(A = d_k) = ۱$ است، یعنی به احتمال ۱ این فرد «مرد» و با احتمال صفر «زن» خواهد بود، زیرا در مجموعه‌ی داده‌های مطمئن احتمالات به‌صورت ۰ و ۱ خواهد بود. در رابطه با مجموعه‌ی داده‌های نامطمئن، اطلاعات با توزیع احتمال آن‌ها در Dom ثبت می‌شود. در صورتی که دامنه‌ی رسته‌ای $\text{Dom} = \{d_۱, \dots, d_n\}$ باشد، ویژگی رسته‌ای نامطمئن (UCA)، A^{u_c} با توزیع احتمال آن در آن دامنه مشخص می‌شود. این احتمالات با بردار احتمالات $P = \{P_۱, \dots, P_n\}$ مشخص می‌شود که مجموع آن‌ها برابر یک خواهد بود.

$$P\left(A_{ij}^{u_c} = u_k\right) = P_{jk} \quad \text{و} \quad \sum_{k=۱}^{n} P_{jk} = ۱ \; (۱ \le K \le n).$$

در رابطه با مثال فوق می‌توان گفت داده‌ی مورد نظر با احتمال $P_۱$، «مرد» و با احتمال $P_۲$ «زن» خواهد بود به‌صورتی که $P_۱ + P_۲ = ۱$. به‌عنوان مثال، به‌دلیل وجود داده‌های نامطمئن اگر داده‌ی مربوطه در پایگاه داده مقدار «مرد» داشته باشد با احتمال $P_۱ = ۰/۷$ «مرد» خواهد بود و با احتمال $P_۲ = ۰/۳$ «زن».

۲-۳- شرایط انتخاب ویژگی به‌عنوان یافتن بهترین نقطه‌ی تقسیم جهت ساخت درخت

نکته‌ی بسیار مهم در الگوریتم درخت تصمیم، یافتن راهی برای تقسیم کردن رکوردها است. هر مرحله که درخت بزرگ می‌شود، نیاز به انتخاب ویژگی برای تقسیم رکوردها به زیرمجموعه‌های کوچک‌تر وجود دارد. روش‌های معمول جداسازی داده مثل ضریب جینی برای داده‌های نامطمئن مناسب نیستند. در این مقاله، روشی برای طبقه‌بندی داده‌های نامطمئن عددی و داده‌های رسته‌ای‌شده نامطمئن ارایه شده است.

۱-۲-۳- داده‌های عددی نامطمئن

همان‌طور که پیش از این ذکر شد، مقادیر ویژگی‌های عددی نامطمئن با استفاده از تابع چگالی احتمال رسته‌ای می‌شوند. جدول ۱ مثالی از داده‌های نامطمئن عددی را نشان می‌دهد [۱۶] که داده‌های این جدول برای پیش‌بینی افراد داوطلب گرفتن وام استفاده

می‌شود. در تمام این ویژگی‌ها، درآمد سالیانه عددی نامطمئن است که دقت مقادیر آن معلوم نیست.

بررسی‌هایی که تا کنون انجام شده است فقط از تابع چگالی احتمال $f(x)$ یکنواخت در بازه‌ی مربوطه استفاده کرده‌اند. در این مقاله از توابع چگالی احتمال متفاوت استفاده شده است و نتایج آن در ادامه آمده است. مزیت اصلی استفاده از توابع چگالی متفاوت این است که می‌توان عدم توازن داده‌ها[۲۱] را نیز با استفاده از این توابع پوشش داد. زیرا تابع چگالی احتمال یکنواخت، توزیع داده‌ها را به‌صورت متوازن در نظر می‌گیرد. در مثال زیر که داده‌های جدول ۱ را مورد بررسی قرار داده است از تابع چگالی احتمال یکنواخت استفاده شده است. نکته‌ای که در استفاده از تابع چگالی احتمال یکنواخت وجود دارد این است که، این تابع به‌دلیل توزیع یکنواخت آن معمولاً نقطه‌ای در حوالی نقاط میانی را به‌عنوان نقطه‌ی تقسیم قرار خواهد داد. مطالعات انجام‌شده در داده‌کاوی داده‌های نامطمئن طرح اطلاعات اقتصادی خانوار، این موضوع را ثابت می‌کند که به‌دلیل وجود عدم توازن بسیار زیاد این داده‌ها توزیع‌های آماری دیگر نتایج بسیار بهتری را در پیش‌بینی داده‌های جدید تولید خواهند نمود.

هر بازه‌ی عددی نامطمئن، مقادیر ماکسیمم و مینیمم دارد که به آن‌ها نقاط بحرانی گفته می‌شود. برای هر مقدار عددی نامطمئن، باید نقاط بحرانی را به‌صورت صعودی مرتب و موارد تکراری را حذف کرد. پس از این کار می‌توان داده‌های عددی نامطمئن را تقسیم کرد. هر بخش ممکن است با داده‌های عددی نامطمئن زیادی همپوشانی داشته باشد. هنگامی که یک نمونه از داده‌های عددی نامطمئن با بخش $[a,b)$ همپوشانی داشته باشد، احتمال این که داده‌ی عددی نامطمئن در این بازه قرار بگیرد برابر است با $\int_a^b f(x)dx$. بر مبنای احتمال هر نمونه که در بخش $[a,b)$ قرار بگیرد، می‌توان احتمال‌ها را محاسبه کرد که به این احتمال‌ها، احتمال کاردینالیتی گفته می‌شود.

تعریف ۱. احتمال کاردینالیتی یک مجموعه‌ی داده در بازه‌ی $Pa = [a,b)$ مجموع احتمالات هر نمونه (به ازای $j = 1$ تا $j = n$) است که n تعداد کل نمونه‌ها است. داده‌های عددی نامطمئن با بازه‌ی $[a,b)$ همپوشانی داشته باشد به‌صورتی که:

$$PC(Pa) = \sum_{j=1}^{n} P\left(A_{ij}^{u_n} \in [a,b)\right) = \sum_{j=1}^{n} \int_a^b A_{ij}^{u_n}.f(x)dx$$

جدول ۱- داده‌های عددی نامطمئن

ID	مسکن دارد؟	وضعیت تاهل	درآمد سالیانه	آیا وام می‌گیرند؟
۱	بله	مجرد	۱۲۰-۱۱۰	خیر
۲	خیر	متاهل	۱۲۰-۱۰۰	خیر
۳	خیر	مجرد	۸۵-۶۰	خیر
۴	بله	متاهل	۱۴۵-۱۱۰	خیر
۵	خیر	مطلقه	۱۲۰-۱۱۰	بله
۶	خیر	متاهل	۸۰-۵۰	خیر
۷	بله	مطلقه	۲۵۰-۱۷۰	خیر
۸	خیر	مجرد	۱۰۰-۸۵	بله
۹	خیر	متاهل	۱۰۰-۸۰	خیر
۱۰	خیر	مجرد	۱۴۵-۱۲۰	بله
۱۱	خیر	مطلقه	۱۲۵-۱۰۵	بله
۱۲	خیر	مطلقه	۹۵-۸۵	خیر

احتمال کاردینالیتی برای کلاس c_j (کلاس هدف مربوط به هر یک از نمونه‌های داده‌ی نامطمئن) برای بازه‌ی $Pa = [a, b)$ برابر است با مجموع احتمالات هر نمونه‌ی T_j در c_j (هر نمونه از کل داده‌ها که با کلاس هدف بازه‌ی مورد مطالعه همپوشانی داشته باشد) به شرطی که این نمونه با بازه‌ی $[a, b)$ همپوشانی داشته باشد که در آن مجموع همه‌ی داده‌هایی که همپوشانی دارند با $PC(Pa, C)$ مشخص خواهد شد به‌صورتی که:
$PC(Pa, C) = \sum_{j=1}^{n} P(A_{ij}^{u_n} \in [a, b) \wedge C_{T_j} = C_j)$ برچسب $C_{T_j} = C_j$ کلاس نمونه T_j را مشخص می‌کند. با توجه به مجموعه داده موجود در جدول ۱، احتمال کاردینالیتی برای بازه‌ی [۱۱۰,۱۲۰) روی مشخصه درامد سالیانه مجموع احتمالات نمونه‌هایی است که درامد سالیانه‌ی آن‌ها با بازه‌ی [۱۱۰,۱۲۰) همپوشانی دارد. اگر درامد سالیانه برای هر نمونه به‌صورت یکسان در رسته‌های نامطمئن توزیع شده باشد، نمونه‌های ۱،۲، ۴، ۵ و ۱۱ با بازه‌های [۱۱۰,۱۲۰) همپوشانی دارند و احتمال برای نمونه‌ی ۱ با درامد سالیانه [۱۱۰,۱۲۰) برابر است با

$$P\left(I_1 \in [۱۱۰, ۱۲۰)\right) = \frac{۱۲۰-۱۱۰}{۱۲۰-۱۱۰} = ۱$$

$$P\left(I_{۲} \in \left[۱۱۰, ۱۲۰\right)\right) = ۰/۵$$
$$P\left(I_{۴} \in \left[۱۱۰, ۱۲۰\right)\right) = ۰/۲۹$$
$$P\left(I_{۵} \in \left[۱۱۰, ۱۲۰\right)\right) = ۱$$
$$P\left(I_{۱۱} \in \left[۱۱۰, ۱۲۰\right)\right) = ۰/۵$$

بنابراین احتمال کاردینالیتی $PC(Pa, C)$ برای این مجموعه‌ی داده در بازه‌ی [۱۱۰,۱۲۰) برابر مجموعه‌ی احتمالات و ۳/۲۹ است. احتمال کاردینالیتی برای کلاس داوطلب گرفتن وام = خیر در بازه‌ی [۱۱۰,۱۲۰) برای مشخصه‌ی درامد سالیانه، مجموع احتمالات نمونه‌هایی است که داوطلب گرفتن وام نیستند و درامد آن‌ها در بازه‌ی [۱۱۰,۱۲۰) است. در نمونه‌های ۱،۲،۴،۵،۱۱ که با بازه‌ی [۱۱۰,۱۲۰) همپوشانی دارند، فقط نمونه‌های ۱،۲،۴ در کلاس خیر هستند، پس احتمال کاردینالیتی برای داوطلب گرفتن وام = خیر در بخش [۱۱۰,۱۲۰) برابر ۱/۷۹ و برای کلاس بله در این بازه ۱/۵ است.

با دو تعریف فوق، می‌توان آنتروپی احتمالی را برای یافتن بهترین ویژگی تقسیم در مجموعه‌ی داده‌های نامطمئن به‌صورت زیر محاسبه کرد. در اینجا منظور از آنتروپی میزان خلوص (بی‌نظمی یا عدم خالص بودن) مجموعه‌ای از نمونه‌ها است. برای تعیین خلوص خواص ورودی به یک فرمول ریاضی احتیاج است که این خواص را به‌صورت پیشگویانه بسنجد. فرمول ریاضی مربوط به آنتروپی احتمالی مجموع احتمال‌های کاردینالیتی هر برچسب در لگاریتم احتمال‌های کاردینالیتی مربوطه است. $PC(D, i)/PC(D)$ احتمال تعلق هر یک از داده‌های موجود در مجموعه‌ی داده‌ی D به کلاس c_i است. تابع لگاریتم در مبنای ۲ به این دلیل استفاده شده است که اطلاعات به‌صورت بیت به بیت رمزگشایی می‌شوند. $ProbInfo(D)$ میانگین اطلاعات مورد نیاز برای یافتن برچسب کلاس برای هر یک از داده‌های موجود در D است.

تعریف ۲. آنتروپی احتمالی برای مجموعه‌ی داده‌ی D برابر است با:

$$ProbInfo(D) = -\sum_{i=1}^{m} \frac{PC(D,i)}{PC(D)} \times \log_2((PC(D,i)/PC(D))$$

اگر ویژگی A به‌عنوان ویژگی تقسیم‌شونده انتخاب شود و مجموعه داده‌ی D را به K زیرمجموعه تقسیم کند (در حالت عددی به دو زیرمجموعه و در حالت رسته‌ای به k زیر مجموعه) $\{D_1, D_2, \dots, D_k\}$، $ProbInfo_A(D)$ اطلاعات مورد نیاز برای طبقه‌بندی داده‌های موجود در D بر مبنای قسمت‌بندی A است. این مقدار به‌صورت اطلاعات مورد انتظار بر مبنای تقسیم‌بندی به طریق زیر قابل اندازه‌گیری است:

$$ProbInfo_A(D) = \sum_{i=1}^{k} \frac{PC(D_j)}{PC(D)} \times ProbInfo(D_j)$$

عبارت $PC(D_j)$، وزن jامین بخش است. هر چه این مقدار کوچک‌تر باشد، زیرمجموعه‌ی تقسیم‌شده خالص‌تر است. $ProbGain(A)$ یا همان نرخ بهره‌ی احتمالی میزان تفاوت میان آنتروپی احتمالی داده‌ها را پیش از تقسیم و بعد از آن محاسبه می‌کند. به بیان دیگر نرخ بهره‌ی اطلاعاتی میزان دقت را در هنگام انتخاب ویژگی A به‌عنوان ویژگی تقسیم مشخص می‌کند. این نرخ به‌صورت زیر محاسبه می‌شود:

$$ProbGain(A) = ProbInfo(D) - ProbInfo_A(D)$$

ویژگی که بزرگ‌ترین نرخ بهره‌ی احتمالی را داشته باشد به‌عنوان ویژگی تقسیم انتخاب خواهد شد. اطلاعاتی که از آزمون به‌دست آمده است در صورتی ماکسیمم می‌شود که فقط یک حالت در هر زیرمجموعه D_j وجود داشته باشد. این مورد به‌صورت زیر مشخص می‌شود:

$$ProbGain\_ration(A) = ProbGain(A)/ProbSplitInfo_A(D)$$

که در این‌جا:

$$ProbSplitInfo_A(D) = -\sum_{j=1}^{k} \frac{PC(D_j)}{PC(D)} \times \log_2((PC(D_j)/PC(D))$$

در بسیاری از حالات، محاسبه‌ی نرخ بهره به روش فوق برای داده‌هایی با مقادیر مختلف و زیاد مناسب نیست، مثل شناسه‌ی خانوار که شامل اعداد یکتا است و مقادیر متفاوتی

دارد. در این حالات نرخ بهره عدد بزرگی خواهد شد و بنابراین تقسیم روی چنین مشخصه‌ای از داده سبب ایجاد نقاط تقسیم به تعداد داده‌ها خواهد شد و هر شاخه حاوی یک عدد متفاوت خواهد بود. مشخص است که چنین تقسیم‌بندی قابل قبول نخواهد بود. روش فوق در الگوریتم درخت تصمیم ID۳ استفاده می‌شود و برای الگوریتم C۴٫۵ که توسعه‌یافته‌ی الگوریتم فوق است، از ضریب بهره برای یافتن ویژگی تقسیم استفاده می‌شود. از این ضریب به‌منظور نرمال‌سازی نرخ بهره استفاده خواهد شد. ضریب بهره به‌صورت زیر محاسبه می‌شود:

$$ProbGain\_ratio(A) = ProbGain(A)/ProbSplitInfo_A(D)$$

که در این‌جا:

$$ProbSplitInfo_A(D) = -\sum_{j=1}^{k} \frac{PC(D_j)}{PC(D)} \times \log_2((PC(D_j)/PC(D))$$

این مقدار نشان‌دهنده‌ی اطلاعات بالقوه‌ی تولیدشده توسط تقسیم مجموعه‌ی داده‌های آموزشی D، به k زیربخش است. بنابراین ویژگی با بزرگ‌ترین ضریب بهره به‌عنوان ویژگی تقسیم انتخاب خواهد شد.

۲-۲-۳- داده‌های رسته‌ای نامطمئن

ویژگی‌های رسته‌ای نامطمئن، $A_i^{u_c}$ با توزیع احتمالی در دامنه مشخص می‌شوند. همان‌طور که پیش از این ذکر شد، می‌توان این ویژگی‌ها را با بردار احتمال $\{P_1, \dots, P_n\}$ نشان داد به‌صورتی که

$$P(A_{ij}^{u_c} = d_j) = P_j(1 \le j \le n).$$

(جدول۲) مثالی از داده‌های رسته‌ای نامطمئن را نشان می‌دهد. این مجموعه‌ی داده شامل اطلاعاتی از مشکلات وسایل نقلیه است. مشکل ماشین می‌تواند از تایر، ترمز، اگزوز یا سایر بخش‌ها باشد.

جدول ۲- داده‌های رسته‌ای نامطمئن

ID	شرکت سازنده	تاریخ	مشکل	محل	کلاس
۱	Explorer	۲۰۰۸/۰۵/۰۴	(ترمز: ۰٫۵; تایر: ۰٫۵)	کانادا	۰
۲	Camry	۲۰۰۲/۰۸/۰۳	(ترانس: ۰٫۲ ; تایر: ۰٫۸)	هندوستان	۱
۳	Civic	۱۹۹۹/۰۹/۱۲	(اگزوز: ۰٫۴ ; ترمز: ۰٫۶)	تگزاس	۰
۴	Pantiac	۲۰۰۱/۰۴/۰۲	(تایر: ۱)	ایرلند	۱
۵	Caravan	۲۰۰۴/۰۱/۲۳	(ترانس: ۰٫۳ ; ترمز: ۰٫۵)	استرالیا	۱

مشابه داده‌های نامطمئن عددی، احتمال کاردینالیتی برای مجموعه‌ی داده‌های رسته‌ای در d_j (نمونه‌ی مورد مطالعه، مثلا اگر مشکل از ترمز ماشین باشد) مجموع احتمالات هر نمونه است که داده‌های رسته‌ای نامطمئن مربوط به آن برابر d_j است. d_j در این‌جا نشان دهنده‌ی احتمال مربوط به هر یک از موارد موجود در ستون یک داده‌ی رسته‌ای است، پس $PC(d_j) = \sum_{j=1}^{n} P(A_{ij}^{u_c} = d_j)$. بنابراین احتمال کاردینالیتی برابر است با مجموع احتمالات مربوط به هر نمونه در داده. این احتمال برای کلاس C از مجموعه‌ی داده‌ها روی d_j مجموع احتمالات هر نمونه در c_j ($j = 1$ تا $j = n$ که شامل همه‌ی داده‌ها خواهد بود) است که داده‌های رسته‌ای نامطمئن مربوطه برابر d_j است. پس:

$$PC(d_j, C) = \sum_{j=1} P(A_{ij}^{u_c} = d_j \wedge C_j = C)$$

با توجه به مجموعه‌ی داده در جدول ۲، احتمال کاردینالیتی هنگامی که مشکل ماشین از ترمز باشد برابر است با مجموع احتمالات هر نمونه که مشکل آن از ترمز است که برابر است با ۱٫۸=۰٫۵+۰٫۶+۰٫۵. احتمال کاردینالیتی برای کلاس صفر هنگامی که مشکل ماشین از ترمز آن است برابر مجموع احتمالات نمونه‌ها در کلاس صفر است که یکی از ویژگی‌هایی که در ستون مشکل وجود دارد، ترمز باشد، که این مجموع آن برابر ۱٫۱ خواهد بود. بر مبنای احتمال کاردینالیتی برای هر کلاس C، می‌توان احتمال آنتروپی و احتمال نرخ بهره[۲۲] را در حالتی که مشخصه‌ی «مشکل ماشین» به‌صورت رسته‌ای باشد، با همان فرایندی که در حالت عددی انجام می‌شود محاسبه کرد. اگر مشخصه‌ی «مشکل ماشین» یا هر مشخصه‌ی دیگری بالاترین نرخ بهره را داشته باشد، به‌عنوان ویژگی تقسیم‌شونده بعدی انتخاب خواهد شد.

۴- روش پیشنهادی

در این طرح به‌منظور جداسازی داده‌ها از درخت تصمیم نامطمئن استفاده شده است. داده‌های این طرح به چهار روش متفاوت، نامطمئن می‌شوند و به درخت تصمیم ارسال خواهند شد. پس از این که درخت تصمیم به خوبی آموزش دید از آن می‌توان برای پیش‌بینی و ارزیابی سایر داده‌ها و داده‌های جدید استفاده کرد.

درخت تصمیم نامطمئن ساختاری فلوچارتی شبیه به درخت دارد که هر گره داخلی (گره‌هایی غیر از برگ) نشان‌دهنده‌ی ارزیابی روی ویژگی‌ها است، هر شاخه نشان‌دهنده‌ی خروجی ارزیابی است و هر برگ (گره‌های لایه‌ی آخر که منجر به تصمیم‌گیری می‌شوند) برچسب یک کلاس است و بالاترین گره آن گره ریشه است. در این الگوریتم داده‌های آموزشی با D نشان داده می‌شود که شامل مجموعه داده‌های آموزشی و برچسب کلاس‌های آن است. به‌منظور انتخاب ویژگی برای تقسیم طبق توضیحات فوق عمل خواهد شد. بنابراین در صورتی که داده عددی است تقسیم دودویی داده‌ها صورت می‌پذیرد یعنی داده از نقطه‌ای شکسته خواهد شد و بخشی از داده‌ها در شاخه‌ی سمت راست و بخشی در شاخه‌ی سمت چپ قرار خواهد گرفت. حال اگر داده رسته‌ای به‌عنوان ویژگی تقسیم انتخاب شود، به تعداد موارد موجود در هر رسته، درخت شاخه خواهد داشت. به‌عنوان مثال برای داده‌ی سطح سواد با ۵ مقدار متفاوت ۵ شاخه تولید خواهد شد. هنگامی که ویژگی تقسیم یافته می‌شود، همه‌ی نمونه‌های آموزشی به‌عنوان ورودی به درخت تصمیم نامطمئن وارد می‌شوند و در یکی از رسته‌های موجود قرار می‌گیرند و وزن‌دهی می‌شوند. وزن‌دهی به این صورت خواهد بود که در صورتی که داده ورودی در یکی از شاخه‌ها قرار بگیرد با وزن ۱ در آن شاخه و با وزن صفر در شاخه‌های دیگر قرار خواهد گرفت. در صورتی که داده‌ی ورودی با دو یا چند شاخه همپوشانی داشته باشد، وزن‌ها به‌صورت احتمالی در شاخه‌ها تقسیم خواهند شد. در نهایت خروجی این الگوریتم درخت تصمیم نامطمئن خواهد بود.

۱-۴- الگوریتم درخت تصمیم نامطمئن

الگوریتم درخت تصمیم نامطمئن[۲۳]

ورودی: مجموعه‌ی داده‌ی آموزشی D، مجموعه‌ی ویژگی‌های کاندید موجود در فهرست ویژگی
خروجی: درخت تصمیم نامطمئن
شروع
۱. گره N را بساز
۲. اگر[۲۴] (تمام موارد موجود در D متعلق به یک کلاس هستند) آن‌گاه
۳. N به‌عنوان گره برگ با برچسب کلاس C برگردانده می‌شود.
۴. اگر (فهرست ویژگی خالی است) آن‌گاه
۵. N به‌عنوان گره با برچسب بزرگ‌ترین وزن کلاس موجود در D، برگردانده می‌شود.
۶. در غیر این صورت اگر (سطح درخت به ماکسیمم سطح مورد نظر برسد) آن‌گاه
۷. N به‌عنوان گره با برچسب بزرگ‌ترین وزن کلاس موجود در D، برگردانده می‌شود.
۸. اتمام اگر[۲۵]
۹. ویژگی آزمون با بالاترین نرخ بهره احتمالی اطلاعات[۲۶] برای برچسب گره N انتخاب می‌شود.
۱۰. اگر (ویژگی آزمون، عددی یا عددی نامطمئن است) آن‌گاه
۱۱. از نقطه‌ی تقسیم y داده به‌صورت دودویی تقسیم می‌شود.
۱۲. برای[۲۷] (هر نمونه‌ی R_j) تکرار کن
۱۳. اگر (ویژگی آزمون $\geq$ y است) آن‌گاه
۱۴. y در مجموعه‌ی D_l با وزن $R_j.w$ قرار می‌گیرد
۱۵. در غیر این صورت اگر (ویژگی آزمون $>$ y است) آن‌گاه
۱۶. y در مجموعه D_r با وزن $R_j.w$ قرار می‌گیرد
۱۷. در غیر این صورت

۱۸. y در مجموعـه D_l بـا وزن $R_j.w \times \int_{x_1}^{y} f(x)dx$ قـرار می‌گیرد
۱۹. y در مجموعـه D_r بـا وزن $R_j.w \times \int_{y}^{x_2} f(x)dx$ قـرار می‌گیرد
۲۰. اتمام اگر
۲۱. اتمام حلقه
۲۲. در غیر این صورت
۲۳. برای (هر مقدار $a_i(i = 1, \dots, n)$ هر ویژگی) تکرار می‌شود
۲۴. یک شاخه‌ی D_i ساخته می‌شود
۲۵. اتمام حلقه
۲۶. برای (هر نمونه‌ی R_j) تکرار می‌شود
۲۷. اگر (ویژگی آزمون نامطمئن است) آنگاه
۲۸. در D_i با وزن $R_j.a_i.w \times R_j.w$ قرار می‌گیرد
۲۹. در غیر این صورت
۳۰. D_i مطمئن با وزن $R_j.w$ قرار می‌گیرد
۳۱. اتمام اگر
۳۲. اتمام حلقه
۳۳. اتمام اگر
۳۴. برای (هر D_i) تکرار می‌شود
۳۵. گره بازگشتی از $DTU(D_i, att - list)$ به درخت اضافه می‌شود
۳۶. اتمام حلقه
۳۷. پایان

توضیحات مرحله به مرحله این الگوریتم بازگشتی به‌صورت زیر است:

۱- درخت از یک گره تنها (همان گره ریشه) شروع خواهد شد. این گره به‌عنوان نمونه‌ای از داده‌های آموزشی است.

۲- اگر تمام نمونه‌ها متعلق به یک کلاس است، گره‌ی مربوطه، تبدیل به برگ خواهد شد و با نام آن کلاس برچسب می‌خورد. تبدیل شدن یک گره به برگ یکی از شروط پایان الگوریتم درخت تصمیم نامطمئن است (مراحل ۲ و ۳).

۳- اگر فهرست ویژگی‌ها خالی است و ویژگی دیگری برای اضافه شدن به درخت وجود نداشته است، الگوریتم پایان می‌پذیرد (مراحل ۴ و ۵).

۴- دیگر شرط خاتمه این است که سطح درخت به ماکسیمم سطح مورد نظر برسد (مراحل ۶ و ۷).

۵- در غیر این صورت، الگوریتم از احتمال بر مبنای آنتروپی که به آن ضریب بهره‌ی اطلاعات نیز گفته می‌شود استفاده می‌کند تا بتواند ویژگی‌ای را انتخاب کند که به بهترین حالت، نمونه‌ها را تقسیم کند (مرحله‌ی ۹). این ویژگی، ویژگی «آزمون» در گره است.

۶- اگر ویژگی آزمون عددی یا عددی نامطمئن است، در محل y داده‌ها به‌صورت دودویی تقسیم و دو شاخه تولید خواهد شد (مراحل ۱۰ و ۱۱).

نکته‌ی قابل توجهی که در این مرحله قابل ذکر است این است که پس از بررسی‌ها و مشاهده‌ی خروجی‌های برنامه می‌توان به این نتیجه رسید که نقطه‌ی تقسیم هر مرحله به‌صورت حدودی در ناحیه‌ی میانگین داده‌ها است. بنابراین به‌منظور افزایش سرعت الگوریتم به هنگام یافتن نقطه‌ی تقسیم، ۰٫۳ داده‌ها از بالا و ۰٫۳ از پایین را حذف می‌کنیم تا سرعت یافتن احتمالات بیش‌تر شود. زیرا با بررسی‌های زیادی که در این‌جا انجام شد می‌توان گفت که نقطه‌ی تقسیم همواره در ۰٫۴ باقی داده‌ها وجود دارد.

۷- شاخه‌ای برای ویژگی $y \leq$ ویژگی آزمون یا $y >$ ویژگی آزمون ساخته می‌شود. اگر ویژگی نمونه‌ی آزمون $[x_1, x_2]$ کم‌تر یا برابر با $y(x_2 \leq y)$ است، آن را در شاخه‌ی سمت چپ با وزن $R_j \times w$ می‌گذاریم. اگر نمونه‌ی ویژگی آزمون $[x_1, x_2]$ بزرگ‌تر از $y(x_1 > y)$ است، آن را در شاخه‌ی سمت راست با وزن $R_j \times w$ می‌گذاریم. اگر مقدار $[x_1, x_2]$ نقاط تقسیم را بپوشاند یعنی $y(x_1 \leq y \leq x_2)$ آن را در شاخه‌ی سمت چپ با وزن $R_j \times w \times \int_{x_1}^{y} f(x)dx$ و شاخه‌ی سمت راست با وزن $R_j \times w \times$

$\int_{y}^{x_2} f(x)dx$ می‌گذاریم. پس مجموعه‌ی داده به D_L و D_R تقسیم می‌شود (مراحل ۱۲ تا ۲۱).

۸- اگر ویژگی آزمون، رسته‌ای یا رسته‌ای نامطمئن است، داده‌ها به چند مسیر تقسیم می‌شوند (مراحل ۲۳ تا ۳۲). برای هر مقدار ویژگی آزمون، شاخه‌ای ساخته می‌شود و نمونه‌های آن نیز به همین صورت تقسیم می‌شود. هنگامی که یک ویژگی مطمئن است، برای هر مقدار a_i، یک نمونه را در D_i با وزن $R_j \times w$ می‌گذاریم. اگر ویژگی نامطمئن است، احتمال مقادیر ویژگی a_i برابر $R_j \times a_i \times P$ است و نمونه را در شاخه‌ی a_i با وزن $R_j \times a_i \times P \times R_j \times w$ می‌گذاریم.

۹- الگوریتم به‌صورت بازگشتی این فرایند را برای تمام نمونه‌ها، به‌منظور ساخت درخت تصمیم انجام می‌دهد.

۱۰- فرایند بخش‌بندی بازگشتی هنگامی که یکی از سه شرط زیر اتفاق بیفتد متوقف می‌شود:

۱. تمام نمونه‌های گره مربوطه برای یک کلاس است (مراحل ۲ و ۳) یا

۲. ویژگی باقی نمانده است که بخواهد بخش‌بندی شود (مرحله‌ی ۴). در این حالت، کلاس با وزن بیش‌تر انتخاب می‌شود. (مرحله‌ی ۵). در این حالت گره داده‌شده به برگ تبدیل خواهد شد و برچسب این برگ، کلاسی با بالاترین وزن از نمونه‌ها است.

۳. سطح درخت به ماکسیمم سطح مورد نظر برسد. (مرحله‌ی ۶) در این حالت، کلاس با وزن بیش‌تر انتخاب می‌شود (مرحله‌ی ۷). در این حالت گره داده شده به برگ تبدیل خواهد شد و برچسب این برگ، کلاسی با بالاترین وزن از نمونه‌ها است.

۲-۴- پیش‌بینی با درخت تصمیم نامطمئن

هنگامی که درخت تصمیم نامطمئن ساخته می‌شود، می‌توان از آن برای پیش‌بینی کلاس‌ها استفاده کرد. پیش‌بینی در این‌جا به این معنا است که پس از آن‌که الگوریتم درخت تصمیم نامطمئن با داده‌های آموزشی، ساخته شد. هر داده‌ی جدیدی را می‌توان به‌صورت ورودی به درخت داد و کلاس خروجی را دریافت کرد. به‌عنوان مثال اگر بخواهیم بدانیم فرد جدیدی که به یارانه اضافه می‌شود از تمکن مالی خوبی برخوردار است یا خیر، داده را به

درخت تصمیم نامطمئن ایجادشده خواهیم داد و نتیجه به ما بردار احتمالی از میزان تعلق نمونه به کلاس بلی یا خیر خواهد داد. مثلاً با احتمال ۰/۷ این فرد متمکن است و به احتمال ۰/۳ متمکن نیست. فرایند پیش‌بینی از گره ریشه آغاز می‌شود و شرایط آزمون در هر گره درخت تصمیم نامطمئن اعمال خواهد شد و شاخه‌ی مناسب ادامه می‌یابد.

هنگامی که نمونه‌ی آزمون R مطمئن است، فرایند راحت خواهد بود و نتیجه‌ی آزمون و پیش‌بینی به یک شاخه‌ی مشخص و بدون ابهام می‌رسد، یعنی با احتمال ۱ به یک شاخه ختم خواهد شد. هنگامی که ویژگی آزمون نامطمئن است، الگوریتم پیش‌بینی به‌صورت زیر است:

۱. اگر شرایط آزمون روی ویژگی A در درخت تصمیم نامطمئن اعمال شود و نقطه‌ی تقسیم a باشد. R×a را در بازه $[x_1, x_2]$ در نظر گرفته می‌شود و تابع چگالی احتمال مربوط به آن برابر است با$(x)f \times A \times R$.

اگر $a < x_1$ است به این معنی است که مینیمم مقدار ممکن برای $R \times A$ بزرگ‌تر از a است، بنابراین $P(R \times A > a) = R \times W$. در این‌جا این اطمینان وجود دارد که $R \times A > a$ است و R شاخه‌ی راست را ادامه می‌دهد.

اگر $a \geq x_2$ است به این معنی است که مقدار ماکسیمم ممکن برای $R \times A$ کوچک‌تر از a است، بنابراین $P(\mathrm{R} \times \mathrm{A} < \mathrm{a}) = R \times W$ و بدون شک $R \times A < a$ است و R شاخه‌ی چپ را ادامه می‌دهد.

اگر $x_1 < a < x_2$ است، یعنی نمونه‌ی آزمون با دو طرف بازه‌ی نقطه‌ی شکست همپوشانی دارد و احتمال $R \times A < a$ برابر است با :

$$P(R \times A < a) = R \times W \times \int_{x_1}^{a} f(x)dx$$

و احتمال $R \times A > a$ برابر است با:

$$P(\mathrm{R} \times \mathrm{A} > \mathrm{a}) = R \times W \times \int_{a}^{x_2} f(x)dx$$

همچنین R می‌تواند با احتمال $R \times W \times \int_{x_1}^{a} f(x)dx$ در شاخه‌ی سمت چپ است و با احتمال $R \times W \times \int_{a}^{x_2} f(x)dx$ در شاخه‌ی سمت راست است.

۲. اگر شرایط آزمون روی ویژگی A در درخت تصمیم نامطمئن است و $a_1, a_2, \dots, a_k$ مقادیر ویژگی‌های رسته‌ای است، $R \times A$ را درخت تصمیم نامطمئنی در نظر بگیرید که در آن $R \times A = \{P_1, P_2, \dots \dots, P_k\}$ با $p_i (i = 1, 2, \dots, k)$ و احتمال $R \times A = a_i$. پس R در شاخه‌ی iام با احتمال p_i است.

برای گره‌ی برگ درخت تصمیم نامطمئن، هر کلاس C_i احتمال $PL(C_i)$ را دارد که $PL(C_i)$ احتمال این است که هر نمونه در کلاس C_i (کلاس هدف) است. $PL(C_i)$ از تقسیم کاردینالیتی یک نمونه در کلاس C_i بر مجموع احتمالات کاردینالیتی نمونه‌ها در کلاس C_i است. فرض کنید مسیر L از ریشه به گره‌ی برگ، t مقدار آزمون داشته باشد و داده به یک کلاس C_i ختم شود. $p(T_i)$ احتمال این است که نمونه‌ای مسیری را در iامین آزمون طی کند. پس احتمال برای هر نمونه که در کلاس C_i است و مسیر مشخص L را داشته باشد برابر است با:

$$P_{C_i}^{L} = PL(C_i) \times \prod_{i=1}^{T} P(T_i)$$

هنگام پیش‌بینی نوع کلاس برای نمونه‌ی T با مقادیر نامطمئن، ممکن است که این فرایند چند مسیر را بیابد. فرض کنید m مسیر مجموعاً به‌عنوان خروجی داده شده است، احتمال T در کلاس C_i برابر است با: $P_{C_i} = \sum_{i=1}^{m} P_{C_i}^{i}$. در نهایت، پیش‌بینی می‌شود نمونه‌ای در کلاس C_i است که بالاترین P_{C_i} را در تمام P_{C_i}ها دارد انتخاب می‌شود، $i = 1, \dots, n$ یعنی آن مسیری انتخاب خواهد شد که در نهایت به بالاترین P_{C_i} برسد و با بیش‌ترین احتمال کلاس هدف آن مشخص شود.

۵- مجموعه‌ی داده

از جمله مسائلی که هم اکنون در کشورمان مورد توجه خاص است طرح اطلاعات اقتصادی خانوار است که از سال ۱۳۸۷ در کشور اجرایی شده است. مرکز آمار ایران نیز به‌عنوان تنها مرجع رسمی آماری کشور در این طرح همکاری‌های گسترده‌ای را داشته است. این همکاری‌ها در سال ۱۳۸۷ با خوشه‌بندی خانوارها و ارایه‌ی دهک‌های مختلف برای جامعه آغاز شد و در سال ۱۳۹۰ با حذف خانوارهای پردرآمد ادامه یافت و به همین دلیل است که یکی از چالش‌های جدی در مرکز آمار ایران مسئله جداسازی داده‌ها با ضریب اطمینان مناسب است.

در سال ۱۳۹۰ به‌منظور حذف خانوارهای پردرآمد و برای افزایش ضریب اطمینان، داده‌هایی از نهادهای کشوری به مرکز آمار ایران ارایه شد که از بسیاری از این داده‌ها به‌منظور راستی‌آزمایی و حذف دقیق‌تر افراد پردرآمد استفاده شد. به‌عنوان مثال این داده‌ها شامل اطلاعاتی مثل تعداد دفعات خروج از کشور، درآمد خانوار و ... است. داده‌های مورد مطالعه در این مقاله تلفیقی از داده‌های خوداظهاری سال ۱۳۸۷ و برخی داده‌های نهادهای کشوری مربوط به سال ۱۳۹۰ در استان ایلام است که از طریق مرکز آمار ایران به‌منظور انجام طرح تحقیقاتی در اختیار اینجانب قرار گرفته است.

در این مقاله سعی بر این است که در طبقه‌بندی اطلاعات، صرفاً از داده‌های خود اظهاری استفاده شود که هم بتوان تا حد قابل قبولی دقت این داده‌ها را تعیین کرد و همچنین به دلیل این که داده‌های نهادهای کشوری، داده‌های امنیتی و شخصی هستند و همیشه قابل دسترس نیستند، بتوان با میزان دقت قابل قبولی از داده‌های خود اظهاری استفاده کرد. در مرحله‌ی دوم به‌منظور مقایسه‌ی نتایج و افزایش دقت جداسازی مشخصه‌های خوداظهاری، یکی از مشخصه‌های دریافت‌شده از نهادهای کشوری به اطلاعات مردمی اضافه شده است، تا بتوان مقایسه‌ای میان نتایج خروجی انجام داد.

داده‌های طرح اطلاعات اقتصادی خانوار به دو نوع عددی و رسته‌ای تقسیم می‌شوند. برای داده‌های عددی، یک عدد ثابت در پایگاه داده ذخیره شده است ولی این عدد ثابت عدد مطمئنی نیست. بنابراین باید آن‌ها را تبدیل به بازه‌ی عددی حول آن عدد نموده تا عدم اطمینان موجود در نظر گرفته شود و اگر داده‌های نامطمئن از نوع رسته‌ای باشند، باید به آن‌ها احتمالی نسبت داد تا بتوان عدم اطمینان را در آن‌ها در نظر گرفت.

داده‌های فعلی طرح اطلاعات اقتصادی خانوار طبقه‌بندی قطعی دارند و در خروجی دو دسته مشخص و مطمئن داریم که مشخص می‌کند خانوار تمکن مالی دارد یا خیر. تقسیم داده‌ها به این دو دسته با استفاده از داده‌های مطمئن نهادهای کشوری انجام شده است و طبقه‌بندی داده‌ها با استفاده از داده‌های نامطمئن خوداظهاری در این مقاله مطالعه شده است.

۶- نتایج آزمایش‌های انجام‌شده

۱-۶- پیاده‌سازی

داده‌هایی که مقادیر نامطمئن دارند با تابع توزیع احتمال و به‌صورت بازه‌ای حول مقدار مربوطه مشخص می‌شوند. بنابراین مقادیر به جای این‌که مقدار عددی مشخص داشته باشند با توزیع احتمال مشخص می‌شوند. در بسیاری حالات، مثل داده‌های سرشماری یا طرح‌های آماری، به‌دلیل مسائل محرمانگی فقط داده‌های تجمعی در اختیار قرار می‌گیرد، که هر یک از رکوردهای تجمعی با تابع احتمال نشان داده می‌شوند. در برخی از طرح‌های آماری، افراد اطلاعات صحیح و درست نداده و داده‌ها مشکل کم‌گویی خواهند داشت. در طرح اطلاعات اقتصادی خانوار نیز مسئله کم‌گویی یا در برخی موارد حتی نبود برخی از اطلاعات وجود دارد. این مورد در داده‌های درآمدی یا داده‌هایی که قیمت ملک یا ماشین را مشخص می‌کنند بیش‌تر مشاهده می‌شود. بنابراین این داده‌ها با تابع توزیع احتمال مشخص خواهند شد.

پیاده‌سازی الگوریتم درخت تصمیم نامطمئن با زبان برنامه‌نویسی متلب انجام شده است. به‌منظور شروع الگوریتم در ابتدا باید داده‌ها از حالت خام، به شکلی که درخت تصمیم نامطمئن بتواند با آن کار کند تبدیل شود. داده‌های موجود در طرح اطلاعات اقتصادی خانوار به دو گروه عددی و رسته‌ای تقسیم می‌شوند.

داده‌های عددی موجود در این طرح برای خانوار و افراد آن عبارت‌اند از: میزان درآمد دریافتی از نهادهای کشوری، مجموع میزان حقوق خانوار و سایر درآمدها، مجموع قیمت منزل(در صورت مالکیت) و سایر مستغلات، قیمت ماشین، مقدار وامی که در هر ماه پرداخت می‌کند، مقدار کل وامی که دریافت کرده است، مبلغ اجاره‌ی ماهیانه منزل (در صورت مستاجر بودن)، مبلغ رهن منزل (در صورت مستاجر بودن).

داده‌های رسته‌ای موجود در این طرح برای خانوار و افراد آن عبارت‌اند از: تعداد افراد، تحت حمایت نهادهای حمایتی، نوع حمایت (اگر تحت حمایت است)، نوع مالکیت منزل، وضع سواد، وضع فعالیت، کد طبقه‌ی شغلی، تحت پوشش بودن یکی از انواع بیمه‌ها، وضعیت خوردو سواری، وضع واحد مسکونی و وضع اقساط وام بانکی.

به‌منظور تبدیل داده‌های فوق به شکلی که درخت تصمیم نامطمئن بتواند با آن کار کند از چهار روش زیر استفاده شده است:

۱- تبدیل بازه‌ای دوطرفه: در این روش برای داده‌های عددی بازه‌ای حول داده‌ی خام ساخته می‌شود. اگر بازه‌ی نامطمئن ۰/۱ داده‌ی خام است، به مجموعه‌ی داده نامطمئن، U۱۰ گفته می‌شود. مثلا اگر فرد درآمد خود را ۵۰۰۰۰۰۰ ریال اعلام کرده است و داده‌ها با توزیع احتمالی U۲۰ نامطمئن شده است این داده به بازه‌ی [۴۰۰۰۰۰۰,۶۰۰۰۰۰۰] تبدیل می‌شود.

داده‌های رسته‌ای به‌صورت بردار احتمال نشان داده می‌شوند. بنابراین ویژگی رسته‌ای A_i ممکن است k مقدار ممکن v_j را داشته باشد. $1 \le j \le k$. برای نمونهI_j، مقدار A_{ij} به بردار$P = (p_{j1}, p_{j2}, \dots\dots, p_{ji}, \dots\dots p_{jk})$ تبدیل می‌شود، که p_{jl} احتمال $A_{ij}^{u_c}$ است که برابر است با v_l که برابر است با $P(A_{ij}^{u_c} = v_l) = p_{jl}$. مثلاً، هنگامی که ۰/۱ عدم اطمینان داریم، این ویژگی مقدار اصلی خود را با احتمال ۰/۹ می‌گیرد و ۰/۱ مابقی به‌صورت برابر به سایر مقادیر می‌رسد. در نظر بگیرید که مقدار اصلی یک داده برابر است با $A_{ij} = v_1$، احتمال این مقدار برابر است با $P_{j1} = 0/9$ و احتمال سایر مقادیر $P_{jl}(2 \le l \le k)$ برابر است با $\sum_{i=2}^{k} p_{jl} = 0/1$. اگر فرد اعلام کرده است که ماشین دارد و داده‌ها با توزیع احتمالی U۲۰ نامطمئن شوند با احتمال ۰/۸ در نظر گرفته می‌شود که این فرد ماشین دارد و ۰/۲ نیز برای حالتی که ماشین ندارد.

۲- تبدیل بازه‌ای یک‌طرفه: در این روش مانند روش فوق عمل می‌شود با این تفاوت که فقط احتمال کم‌گویی اطلاعات داده می‌شود، بنابراین در این‌جا بازه‌ی نامطمئن از خود داده تا تابع چگالی احتمال داده شده، خواهد بود و داده‌های رسته‌ای بدون هیچ تفاوتی مانند مورد فوق خواهند بود.

۳- روش ترکیبی با اعمال داده‌های دریافت‌شده از نهادهای کشوری: این روش، به جهت افزایش دقت از داده‌های دریافت‌شده از نهادهای کشوری استفاده می‌شود. در دو روش اول از این داده‌ها هیچ استفاده‌ای نشد و دلیل آن، احتمال عدم وجود این داده‌ها در برخی موارد بوده است. در این‌جا برای نامطمئن‌سازی داده‌ها به‌صورت ترکیبی عمل خواهد شد. در داده‌های دریافت‌شده از نهادهای کشوری فقط یک داده وجود دارد که می‌توان آن را با داده‌های خود اظهاری مرتبط کرد، این داده درآمد خانوار است که از نهادهای کشوری دریافت شده است. مقایسه‌ی این دو مشخصه این گونه انجام می‌شود که مشخصه‌ی

مجموع درآمدهای خانوار را ابتدای بازه‌ی عددی نامطمئن و مشخصه‌ی درآمد دریافت‌شده از نهادها به‌عنوان انتهای بازه در نظر گرفته شد.

۴- استفاده‌ی مستقیم از داده‌های نهادها در داده‌های ورودی: تا قبل از این روش صرفاً از داده‌های خود اظهاری استفاده می‌شد و در روش قبل نیز مقایسه‌ای بین داده‌های خود اظهاری و داده‌های نهادها انجام شد. اما در این روش این مشخصه، مستقیماً در داده‌های ورودی اعمال می‌شود و به‌عنوان یکی از مشخصه‌ها با آن برخورد می‌شود و توزیع‌ها نیز روی آن اعمال می‌شود.

۲-۶- تابع توزیع احتمال

همان‌طور که پیش از این نیز ذکر شد، مقادیر ویژگی‌های عددی نامطمئن با استفاده از تابع چگالی احتمال رسته‌ای می‌شوند. می‌توان از توابع توزیع احتمال متفاوتی به‌منظور نامطمئن‌سازی داده‌های عددی استفاده کرد. در مطالعاتی که پیش از این انجام شده است فقط از تابع توزیع احتمال یکنواخت استفاده شده است، که مزیت اصلی در این مقاله استفاده از توابع چگالی متفاوت در این است که، می‌توان عدم توازن داده‌ها را نیز با استفاده از این توابع پوشش داد. توابع چگالی احتمالی که در این برنامه استفاده شده است عبارت‌اند از:

۱- **توزیع یکنواخت:** تابع چگالی توزیع یکنواخت به‌صورت $\boldsymbol{f(x) = \frac{1}{\beta-\alpha}}$ مشخص می‌شود به‌صورتی که $\boldsymbol{\alpha < x < \beta}$ است.

۲- **توزیع مثلثی:** تابع چگالی احتمال این توزیع برابر است با:

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a \le x \le c \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c \le x \le b \end{cases}$$

۳- **توزیع بتا:** این تابع بیان‌گر توزیع بتا با پارامترهای $\boldsymbol{\beta}$ و $\boldsymbol{\alpha}$ است که در آن $\boldsymbol{\alpha, \beta > 0}$ و $\boldsymbol{0 < x < 1}$ است. تابع چگالی احتمال این توزیع برابر است با:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

۴- **توزیع نرمال:** تابع چگالی احتمال توزیع نرمال با پارامترهای μ و $\boldsymbol{\sigma^2}$ به‌صورت زیر است:

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)},\qquad x\in\mathbb{R}.$$

۵- **توزیع گاما:** تابع چگالی توزیع گاما به‌صورت $\boldsymbol{f(x;\alpha,\beta)=\frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}}$ مشخص می‌شود به‌صورتی که $\boldsymbol{\alpha}$ پارامتر شکل[۲۸] توزیع است و $\boldsymbol{\beta}$ پارامتر نرخ[۲۹] توزیع را مشخص می‌کنند. تغییر در $\boldsymbol{\beta}$ شکل نمودار را به‌صورت عمودی تغییر می‌دهد و در $\boldsymbol{\alpha}$ این تغییر به‌صورت افقی خواهد بود. در این توزیع $\boldsymbol{\alpha,\beta,x>0}$ است.

۳-۶- نتایج

همان‌طور که ذکر شد، مراحلی که در راستای اجرای برنامه طی می‌شود به‌صورت زیر است:

۱- تبدیل داده‌ها به شکلی که درخت تصمیم نامطمئن بتواند با آن کار کند. این عمل، به یکی از چهار طریق زیر انجام می‌شود که در هر یک از این مراحل می‌توان از توزیع‌های مختلف آماری استفاده کرد.
- تبدیل بازه‌ای دوطرفه
- تبدیل بازه‌ای یک‌طرفه
- روش ترکیبی با اعمال داده‌های دریافت‌شده از نهادها
- استفاده‌ی مستقیم از داده‌های نهادها در داده‌های ورودی

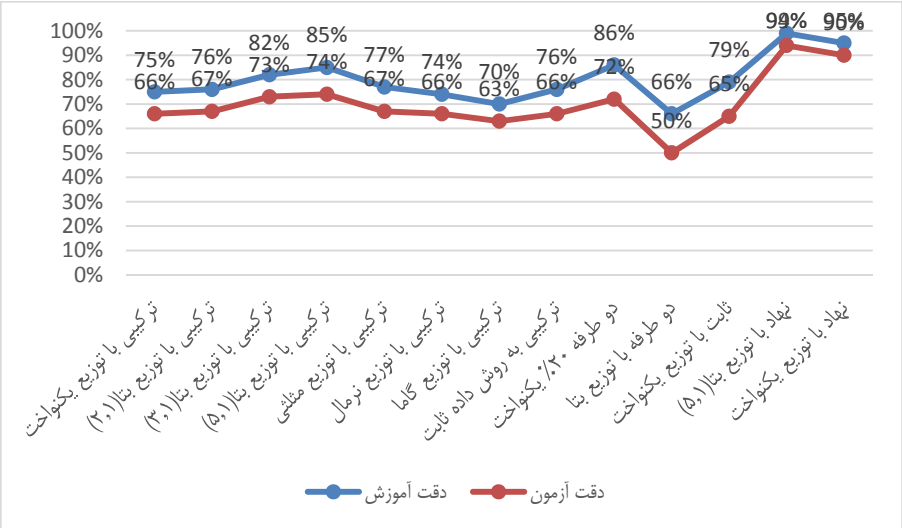
۲- ساخت و یادگیری درخت تصمیم با ۰/۷ داده‌ها

۳- آزمون و پیش‌بینی ۰/۳ باقیمانده‌ی اطلاعات

۴- اعتبارسنجی اطلاعات.

داده‌های طرح اطلاعات اقتصادی خانوار را با تمام حالات و توابع توزیع احتمال فوق نامطمئن کرده و پس از ساخت درخت با ۰/۷ داده‌ها نتایج آزمون و اعتبارسنجی آن‌ها را با هم مقایسه کرده و بهترین نتایج در نمودارهای زیر آورده شده است. مقایسه‌ی میان دقت آموزش در الگوریتم درخت تصمیم نامطئمن با دقت آزمون در شکل ۱ قابل مشاهده است.

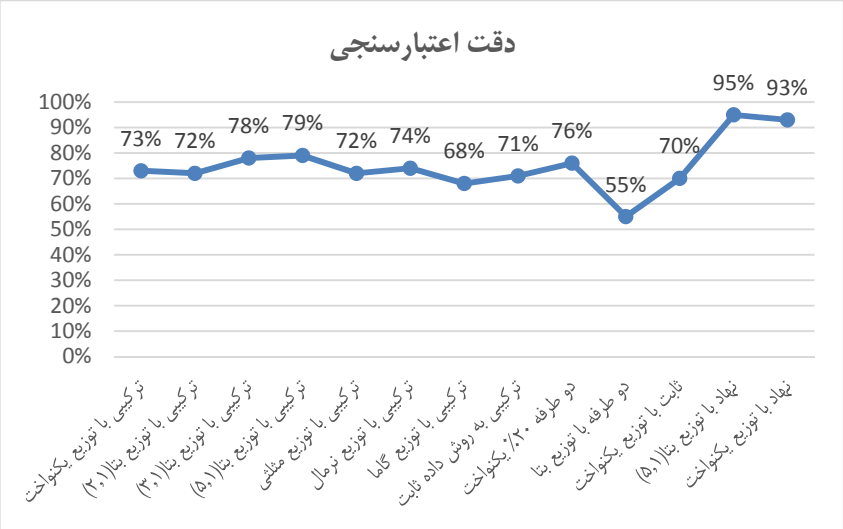
در این‌جا منظور از دقت آموزش این است که، الگوریتم درخت تصمیم نامطئمن تا چند درصد توانسته است به درستی ۰/۷ داده‌ی آموزشی را یاد بگیرد. این کار به این صورت انجام خواهد شد که درخت تصمیم با داده‌های آموزشی ساخته خواهد شد و داده‌ها مجدداً به‌صورت داده‌ی جدید به درخت داده می‌شود تا بتوان دقت درخت را در آموزش سنجید. دقت آزمون نیز به این مفهوم خواهد بود که ۰/۳ داده‌ی باقیمانده و جدید به‌صورت ورودی به درخت تصمیم داده خواهد شد و میزان دقت درخت در درست پیش‌بینی کردن کلاس خروجی ارزیابی شده و به‌صورت درصد بیان خواهد شد. در این نمودار «U۲۰ با توزیع یکنواخت» و «U۲۰ با توزیع بتا(۲،۱)» در گروه تبدیل بازه‌ای دوطرفه، «U+ با توزیع یکنواخت» در گروه تبدیل بازه‌ی یک‌طرفه، «ترکیبی با توزیع یکنواخت»، «ترکیبی با توزیع بتا (۲،۱)»، «ترکیبی با توزیع بتا (۳،۱)»، «ترکیبی با توزیع بتا (۵،۱)»،«ترکیبی با توزیع مثلثی» و «ترکیبی به روش داده‌ی ثابت» در گروه روش ترکیبی با اعمال داده‌های دریافت‌شده از نهادها و «نهاد با توزیع یکنواخت» و «نهاد با توزیع بتا (۵۲،۱)» در گروه استفاده‌ی مستقیم از داده‌های نهادها در داده‌های ورودی هستند که بهترین نتایج را داده‌اند. نمودار آزمون نتیجه‌ی عملکرد آموزش درخت در هر مرحله است.



شکل ۱- مقایسه‌ی دقت آموزش و آزمون درخت تصمیم نامطمئن

۴-۶- اعتبارسنجی[۳۰] اطلاعات

نتایج اصلی و قابل مقایسه از اعتبارسنجی اطلاعات به وجود می‌آید. اعتبارسنجی اطلاعات به این طریق انجام می‌شود که کل مجموعه‌ی داده‌ها (D) به تعداد K بار مورد آزمایش قرار خواهد گرفت و میانگین نتایج K بار آزمایش، به‌عنوان نتیجه‌ی نهایی خواهد بود. بنابراین به تعداد K بار داده‌های آموزشی به درخت تصمیم داده می‌شوند و نتیجه‌ی آموزش با داده‌های آزمونی مقایسه خواهند شد. بنابراین، $\frac{D}{K}$ عدد از داده‌ها به‌عنوان داده‌ی آزمون و $\frac{D}{K} - ۱$ عدد از داده‌ها داده آموزشی خواهند بود. به‌عنوان مثال اگر k=۱۰ است و تعداد کل داده‌های ما ۱۰۰ است، ۱۰ دور آموزش و آزمون انجام خواهد شد و میانگین این ۱۰ بار آزمون به‌عنوان نتیجه‌ی اعتبارسنجی خواهد بود. این ۱۰ بار به این صورت طی می‌شوند که در دور اول $\frac{۱}{۱۰}$ اول داده‌ها به‌عنوان داده‌ی آزمون و $\frac{۹}{۱۰}$ باقیمانده به‌عنوان داده‌ی آموزشی هستند. در گذر دوم، $\frac{۱}{۱۰}$ دوم داده‌ها را به‌عنوان آزمون و $\frac{۹}{۱۰}$ باقیمانده به‌عنوان داده‌ی آموزشی در نظر گرفته می‌شود و همین‌طور الی آخر. نتیجه‌ی اعتبارسنجی را برای حالات فوق در شکل ۲ می‌توان دید.



شکل ۲- اعتبارسنجی اطلاعات

همان‌طور که مشاهده می‌شود نتایج قابل قبولی در اعتبارسنجی اطلاعات به‌دست آمد و می‌توان طبقه‌بندی را با استفاده از داده‌های خود اظهاری به تنهایی و با حداکثر ۰٫۸ صحت انجام داد و در مورد ۰٫۲ باقیمانده نیز از بخش کوچکی از داده‌های نهادهای دولتی کمک گرفت و با ۰٫۹۵ اطمینان طبقه‌بندی را انجام داد.

۷- بحث و نتیجه‌گیری

در این مقاله، درخت تصمیم نامطمئن، با تغییرات در الگوریتم به‌منظور بهینه‌سازی الگوریتم، بهبود سرعت اجرای آن و کاربردی سازی این درخت برای داده‌های حقیقی، پیاده‌سازی شده است. نتایج نهایی الگوریتم درخت تصمیم نامطمئن به‌صورت زیر خواهد بود:

درخت تصمیم نامطمئن ارایه‌شده برای نخستین بار روی پایگاه‌داده‌ی نامطمئن واقعی آزمون و اعمال شده است.

مطالعاتی که تا این‌جا در رابطه با درخت تصمیم نامطمئن انجام شده است فقط با تابع توزیع احتمال یکنواخت کار می‌کرده است، اما در این مقاله الگوریتم با توابع توزیع احتمال متفاوت کار می‌کند و نتایج نشان می‌دهد که سایر توابع توزیع در داده‌های نامطمئن بسیار بهتر از توزیع یکنواخت عمل می‌کند، زیرا در این توابع عدم توازن و عدم اطمینان داده‌ها بیش‌تر مورد توجه قرار می‌گیرد.

تجارب این طرح نشان می‌دهد که این درخت به‌صورت هم‌زمان برای داده‌های عددی و رسته‌ای کار می‌کند و می‌تواند طبقه‌بندی و پیش‌بینی را با دقت قابل قبولی، حتی برای داده‌هایی با عدم اطمینان بالا انجام دهد.

در سال‌های اخیر یکی از مسائل اساسی و چالشی کشور بحث تشخیص خانوارهای متمکن و پردرآمد در طرح اطلاعات اقتصادی خانوار بوده است. به‌منظور تشخیص و یافتن این خانوارها در سال‌های گذشته از داده‌های مطمئن سایر نهادهای دولتی کمک گرفته شد و به‌دلیل عدم اطمینان داده‌های خود اظهاری امکان استفاده از این داده‌ها هرگز فراهم نگردید. با توجه به نیاز مرکز آمار ایران در جداسازی خانوارها بر مبنای داده‌های خود اظهاری پژوهش فوق انجام گردیده است و نتایج این مورد را اثبات می‌کند که می‌توان با دقت ۰٫۸ از داده‌های خود اظهاری به‌منظور تشخیص این خانوارها استفاده نمود و در صورتی که بخواهیم این دقت به ۰٫۹۹ برسد می‌توان از یکی از مشخصه‌های

دریافتی از نهادهای کشوری به جای ۱۷ مشخصه که در گذشته دریافت می‌شد، استفاده کرد. نکته‌ی بسیار مهم در این جا این است که با استفاده از داده‌های خوداظهاری و مردمی می‌توان پاسخ بهتر و محکم‌تری به افراد معترض در جامعه ارایه نمود زیرا مبنای اصلی این جداسازی داده‌هایی است که افراد اعلام کرده‌اند.

۸- پیش‌نهادهای آتی

در ادامه‌ی این کار می‌توان طرحی مطالعاتی به‌منظور بهبود بیش‌تر کیفیت طراحی پرسشنامه‌ها از خروجی‌های این طرح تعریف کرد. در این راستا می‌توان مدلی از پرسشنامه را برای افرادی که از پاسخ خود مطمئن نیستند ارایه کرد. همچنین می‌توان طرحی مطالعاتی به‌منظور تعیین شاخصی دقیق‌تر و مناسب‌تر از میزان کم‌گویی افراد تعریف کرد، تا از این عدد در پیش‌بینی‌ها و طرح‌های بعدی استفاده شود.

سپاس‌گزاری

در اینجا لازم است از همکاری بی‌دریغ مرکز آمار ایران برای پیشبرد این مقاله تشکر نمایم.

توضیحات

1. Knowledge
2. Supervised Learning
3. Unsupervised Learning
4. K-nearest Neighbor Method (KDD)
5. Artificial Neural Network (NN)
6. Decision Tree (DT)
7. Support Vector Machine
8. Join
9. Query
10. Indexing
11. Integration
12. Density-Based Clustering

13. Frequent Item Set Mining
14. Density-Based Classification
۱۵. داده‌های عددی مانند فیلد سن، درآمد، ارزش ملک و ...
16. Uncertain Numerical Attributes
17. Imbalance Data
۱۸. داده‌های رسته‌ای مانند گروه خونی، جنسیت، سطح سواد و ...
19. Uncertain Categorical Attributes
20. Instead of Domain
21. Imbalance Data
22. Information Gain
۲۳. در بخش پیوست‌ها الگوریتم به زبان انگلیسی نیز آورده شده است.
24. If
25. End if
26. Probabilistic information gain ration
27. for
28. Shape
29. Rate
30. Cross Validation

مرجع‌ها

[1] Aggarwal, C.C. (2007). On Density Based Transforms for Uncertain Data Mining, in ICDE. Istanbul, Turkey: IEEE, pp. 866–875.

[2] Andrews, R., Diederich, J. and Tickle, A. (1995). A Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks, *Knowledge Based Systems*, **8**, 373-389.

[3] Andritsos, P., Fuxman, A. and Miller, R.J. (2006). Clean Answers over Dirty Databases: A Probabilistic Approach, Proc. 22nd IEEE Int'l Conf. Data Eng. (ICDE).

[4] Bi, J. and Zhang, T. (2004). Support Vector Classification with Input Data Uncertainty, *Advances in Neural Information Processing Systems*, **17**, 161-168.

[5] Chau, M., Cheng, R., Kao, B. and Ng, J. (2006). Uncertain Data Mining: An Example in Clustering Location Data, in PAKDD, ser. Lecture Notes in Computer Science, vol. 3918. Singapore: Springer, pp. 199–204.

[6] Cheng, R., Kalashnikov, D. and Prabhakar, S. (2003). Evaluating Probabilistic Queries over Imprecise Data, In: Proceedings of the ACM SIGMOD.

[7] Chui, C.K., Kao, B. and Hung, E. (2007). Mining Frequent Item Sets from Uncertain Data, in PAKDD, ser. Lecture Notes in Computer Science, vol. 4426. Nanjing, China: Springer, pp. 47–58.

[8] Cormode, G. and McGregor, A. (2008). Approximation Algorithms for Clustering Uncertain Data, *PODS 2008*, 191-199.

[9] Das Sarma, A., Benjelloun, O., Halevy, A. and Widom, J. (2006). Working Models for Uncertain Data, Proc. 22nd IEEE Int'l Conf. Data Eng. (ICDE).

[10] Ding, Z. (2011). Diversified Ensemble Classifiers for Highly Imbalanced Data Learning and their Application in Bioinformatics, Computer Science Dissertations, Computer Science Department, Georgia State University.

[11] Hawarah, L., Simonet, A. and Simonet, M. (2006). Dealing with Missing Values in a Probabilistic Decision Tree During Classification, The Second International Workshop on Mining Complex Data, pp. 325-329.

[12] Kaufmann, M. (1993). C4.5: Programs for Machine Learning, ISBN 1-55860-238-0.

[13] Kriegel, H. and Pfeifle, M. (2005). Density-Based Clustering of Uncertain Data, In: Proceedings of the KDD'05, pp. 672-677.

[14] Langley, P., Iba, W. and Thompson, K. (1992). An Analysis of Bayesian Classifiers, In: Proceedings of the tenth National Conference on Artificial Intelligence, pp. 223–228.

[15] Ngai, W.K., Kao, B., Chui, C.K., Cheng, R., Chau, M. and Yip, K.Y. (2006). Efficient Clustering of Uncertain Data, In: Proceedings of ICDM'06, pp. 436–445.

[16] Qin, B., Xia, Y. and Li, D. (2009). DTU: A Decision Tree for Uncertain Data, Advances in Knowledge Discovery and Data Mining, Springer.

[17] Quinlan, J.R. (1986). Induction of Decision trees. *Machine Learning*, **1**, 81–106.

[18] Quinlan, J.R. (1990). Probabilistic Decision Trees. In Machine Learning: an Artificial Intelligence Approach, **3**, 140–152

[19] Tsang, S., Kao, B., Yip, K.Y., Ho, W.S. and Lee, S.D. (2011). Decision Trees for Uncertain Data, *IEEE Transactions on Knowledge and Data Engineering*, **23**.

[20] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer Verlag.

پیوست‌ها

Uncertain Decision Tree Algorithm

```
Begin
1: create a node N;
2: if (D are all of the same class, C) then
3:       return N as a leaf node labeled with the class C;
4: else if (attribute-list is empty) then
5:       return N as a leaf node labeled with the highest weight class in D;
6: else if (Level = Max Level) then
7:       return N as a leaf node labeled with the highest weight class in D
8: end if;
```

9: select a test-attribute with the highest probabilistic information gain ratio to label node N;
10: if (test-attribute is numeric or uncertain numeric) then
11: binary split the data from the selected position y;
12: for (each instanceR_j) do
13: if (test-attribute ≤ y) then
14: put it into D_l with weight$R_j.w$;
15: else if (test-attribute ≻ y) then
16: put it into D_r with weight$R_j.w$;
17: else
18: put it into D_l with weight$R_j.w \times \int_{x_1}^{y} f(x)dx$;
19: put it into D_r with weight$R_j.w \times \int_{y}^{x_2} f(x)dx$;
20: end if;
21: end for;
22: else
23: for (each value $a_i(i = 1, \dots, n)$ of the attribute) do
24: grow a branch D_i for it;
25: end for;
26: for (each instanceR_j) do
27: if (test-attribute is uncertain) then
28: put it into D_i with weight$R_j.a_i.w \times R_j.w$;
29: else
30: put it into a certain D_i with weight$R_j.w$;
31: end if
32: end for;
33: end if;
34: for each D_i do
33: attach the node returned by$DTU(D_i, att - list)$;
35: end for;
End

مهسا قائمی
فوق لیسانس مهندسی کامپیوتر- هوش مصنوعی
تهران، انتهای بزرگراه شهید ستاری، میدان دانشگاه، بلوار شهدای حصارک، دانشگاه آزاد اسلامی واحد علوم و تحقیقات.
رایانشانی: ghaemi.mahsa@gmail.com

میرمحسن پدرام
دکتری برق
تهران، میدان هفت تیر، خیابان مفتح جنوبی، دانشگاه خوارزمی، گروه مهندسی برق و کامپیوتر.
رایانشانی: mmpedram@gmail.com

عادل آذر
دکتری مدیریت صنعتی
تهران، خیابان فاطمی، خیابان رهی معیری، پلاک ۱، مرکز آمار ایران.
رایانشانی: azara@modares.ac.ir