

تعیین خودکار حداقل دامنه پشتیبانی از قاعده در محیط فازی برای بهبود استخراج قواعد همباش با استفاده از الگوریتم اپریوری

حیدر جعفرزاده^۱، چمران عسگری^۲، امیر امیری^۳

چکیده: قواعد همباش (انجمنی) یکی از محبوب‌ترین مدل‌های حوزه داده‌کاوی به‌شمار می‌رود. در الگوریتم‌های کلاسیک حوزه قواعد همباش کاوی، مانند اپریوری، از حداقل دامنه پشتیبانی قاعده واحد استفاده می‌شود؛ در حالی که در رویکردهای جدیدی که تلاش کرده‌اند الگوریتم‌های کلاسیک را بهبود بخشند، مانند ام. اس. اپریوری، از حداقل دامنه پشتیبانی قاعده چندگانه استفاده می‌شود که در هر دو حالت، کاربر موظف است حداقل دامنه پشتیبانی از قاعده را تعیین کند. در نظر بگیرد کاربر قصد اعمال الگوریتم اپریوری را بر پایگاه داده‌ای با میلیون‌ها تراکنش داشته باشد؛ به‌طور قطع کاربر نمی‌تواند دانش لازم را درباره تمام تراکنش‌های موجود در پایگاه داده داشته باشد، بنابراین نمی‌تواند حد آستانه مناسبی را تعیین کند. در این پژوهش، برای اولین بار با استفاده از داده‌های فازی‌سازی شده و تکنیک میانگین‌گیری، روشی ارائه شده است که در آن، الگوریتم اپریوری به‌صورت کاملاً خودکار حداقل دامنه پشتیبانی از قاعده را تعیین می‌کند. نتایج شبیه‌سازی شده روی نمونه‌ای واقعی نشان داد این رویکرد عملکرد مطلوب‌تری نسبت به الگوریتم اپریوری کلاسیک دارد.

واژه‌های کلیدی: الگوریتم اپریوری، الگوهای پرتکرار، خوشه‌بندی فازی، دامنه پشتیبانی از قاعده، قواعد همباش.

۱. کارشناس ارشد مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات ایلام، ایلام، ایران

۲. کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه پیام نور، ایران

۳. کارشناس ارشد مهندسی کامپیوتر، دانشگاه آزاد اسلامی ملایر، ملایر، ایران

تاریخ دریافت مقاله: ۱۳۹۳/۰۶/۲۸

تاریخ پذیرش نهایی مقاله: ۱۳۹۴/۰۳/۰۶

نویسنده مسئول مقاله: حیدر جعفرزاده

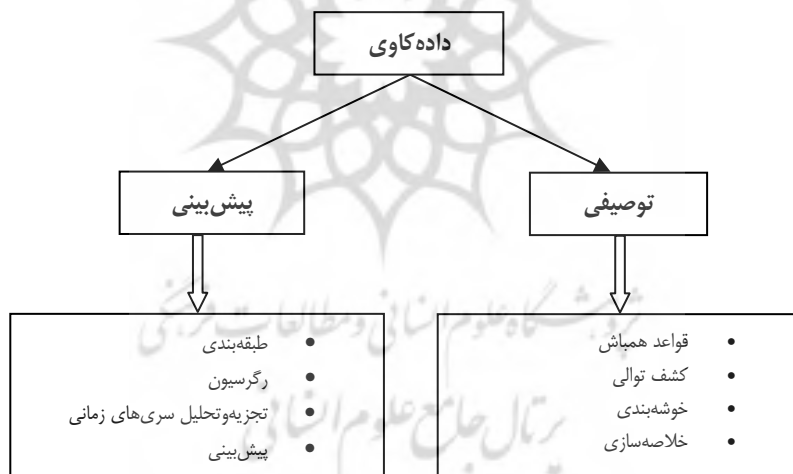
E-mail: heydar.jafarzadeh@gmail.com

مقدمه

داده‌کاوی فرایند منطقی‌ای است که برای جست‌وجوی داده‌ها از حجم زیادی از اطلاعات، به کار برده می‌شود (دونهام، ۲۰۰۲: ۵۱-۵۰). داده‌کاوی به فرایند شناخت الگوها و کشف روابط معنادار بین داده‌های انبوه اشاره دارد (عزیزی، حسین‌آبادی و بلاغی اینانلو، ۱۳۹۳). با توجه به اینکه همیشه می‌توان رابطه‌های جالبی بین داده‌ها پیدا کرد، ابزارهای خودکار و کارآمدی برای یافتن و نظم‌دهی به این رابطه‌ها لازم است. از این رو ابزارهای داده‌کاوی بسیاری با روش‌های تجزیه و تحلیل مختلف ارائه شده است.

داده‌کاوی به مدل‌های پیش‌بینی‌کننده و توصیفی تقسیم می‌شود (دونهام، ۲۰۰۲: ۷۴-۷۳) و همان‌طور که در شکل ۱ آمده است در حوزه‌های گوناگون به کار می‌رود.

- مدل پیش‌بینی؛ به منظور پیش‌بینی مقادیر آینده از نتایج به دست آمده از داده‌های مختلف، به کار می‌رود؛
- مدل توصیفی؛ به منظور ارائه الگوهای کشف شده به شکلی فهم‌پذیر برای انسان به کار می‌رود.



شکل ۱. وظایف داده‌کاوی

قواعد همبش^۱ یکی از وظایف بسیار مهم در حوزه داده‌کاوی است که می‌تواند در دامنه‌های مختلف استفاده شود. قواعد همبش‌کاوی، روش داده‌کاوی عمومی است و برای استخراج

الگوهای مفید از پایگاه داده‌های عظیم استفاده می‌شود (کاکر و آراس، ۲۰۱۲). قوانین انجمنی، یکی از روش‌های توصیفی و غیرنظارتی داده‌کاوی است که برای یافتن ارتباط بین ویژگی‌ها، در مجموعه داده‌ها به جست‌وجو می‌پردازد. در واقع این روش به مطالعه ویژگی‌هایی می‌پردازد که همراه یکدیگرند، ضمن آنکه ارتباط بین این ویژگی‌ها را کمی می‌کند (آخوندزاده نوقابی، البدوی و اقدسی، ۱۳۹۳).

یکی از مشکلات الگوریتم اپریوری^۱ و سایر الگوریتم‌هایی که در حوزه استخراج قواعد همبش‌اند، اینکه کاربر مجبور است حداقل آستانه را برای دامنه پشتیبانی از قاعده^۲ مشخص کند. در نظر بگیرد کاربر می‌خواهد الگوریتم اپریوری را بر پایگاه داده‌ای با میلیون‌ها تراکنش اعمال کند؛ به‌طور قطع کاربر نمی‌تواند دانش مد نظر برای تراکنش‌های موجود در پایگاه داده را داشته باشد و نمی‌تواند حداقل دامنه پشتیبانی از قاعده مناسبی تعیین کند.

هدف از این پژوهش بهبود الگوریتم اپریوری است. برای این کار ابتدا تراکنش‌هایی که در پایگاه داده به‌صورت قطعی ذخیره شده‌اند، به کمک یکی از الگوریتم‌های خوشه‌بندی فازی، وارد محیط فازی (لطفی‌زاده، ۱۹۶۵) خواهند شد و قبل از اعمال الگوریتم اپریوری، داده‌ها در محدوده‌های مختلفی قرار خواهند گرفت. در این فرایند تلاش شده است مناسب‌ترین حد آستانه به‌صورت خودکار به کمک یکی از روش‌های حوزه آمار، پس از تعیین، به کاربر پیشنهاد شود. امیدواریم این کار موجب شود هیچ قاعده جالبی به دلیل حد آستانه نامناسبی که کاربر مشخص کرده است، از دست نرود و هیچ قاعده بی‌استفاده‌ای استخراج نشود.

این مقاله ابتدا پیشینه استخراج قواعد همبش را مرور می‌کند و پس از آن روش پیشنهادی را معرفی می‌کند. برای درک بهتر مسئله، این رویکرد در قالب جدول‌ها و شکل‌های متعددی ارائه شده است. بخش پایانی مقاله به نتیجه‌گیری و پیشنهادها اختصاص دارد.

پیشینه پژوهش

همان‌طور که بسیاری از پایگاه داده‌ها انبوهی از نمونه‌ها و صفت‌ها را دربرگرفته‌اند، باید ابزارهای کارآمدی پیاده‌سازی شود تا بتوان اطلاعات مفید پنهان شده در این پایگاه‌ها را استخراج کرد؛ بر این اساس ابزارهای داده‌کاوی بسیاری که از روش‌های تجزیه و تحلیل گوناگونی استفاده می‌کنند و به‌طور عمده برگرفته از آمار کلاسیک‌اند، توسعه داده شده است. در این بین، روش‌های قواعد همبش کاوی علاقه‌مندی‌های بسیاری را ایجاد کرده‌اند (لی و رنهو، ۲۰۰۷)؛ زیرا محققان

1. Apriori
2. Minsup

بسیاری موضوع پژوهش در این زمینه انتخاب کرده‌اند و الگوریتم‌های متنوعی را در این حوزه توسعه داده‌اند.

پیشینه نظری

استخراج قواعد همبش از پایگاه داده‌ها، اصول و مراحل دارد و تمام محققانی که در این حوزه کار کرده‌اند با توجه به این اصول، تلاش در بهبود روش‌ها و الگوریتم‌ها داشته‌اند.

اصول قواعد همبش

همبش یعنی با هم‌بودن دو یا چند قلم داده که مدام با هم تکرار می‌شوند و جلو می‌روند. کشف قواعد همبش یکی از وظایف بسیار مهم در حوزه داده‌کاوی است. قاعده همبش ساده به این صورت بیان شود: $Bread \rightarrow Cheese$ [support=۰/۱, confidence=۰/۸]. به گفته ساده‌تر این قاعده بیان می‌کند رابطه‌ای بین خرید نان و پنیر وجود دارد، مازور دامنه پشتیبانی از قاعده، بیان‌کننده ثبت نان و پنیر با هم در ۱۰ درصد تراکنش‌ها است و مازور میزان اعتماد به قاعده بیان می‌کند پنیر در تراکنش‌هایی اتفاق افتاده است که نان هم در آن تراکنش‌ها وجود دارد. در این مورد ۸۰ درصد تراکنش‌ها شامل پنیر است که نان هم در همان تراکنش‌ها حضور دارد. به کمک این قاعده می‌توان فرض کرد، در آینده افرادی که نان خریداری می‌کنند به احتمال زیاد در همان تراکنش‌ها پنیر نیز خواهند خرید. این‌گونه اطلاعات می‌تواند به فروشندگان برای کشف فرصت‌های متقابل خرید کمک کند (گوتوالد، ۲۰۰۶).

در این مطالعه مسئله قواعد همبش به این صورت بیان می‌شود: $I = \{i_1, i_2, \dots, i_m\}$ مجموعه‌ای از اقلام داده است و $T = \{t_1, t_2, \dots, t_n\}$ مجموعه‌ای از تراکنش‌هاست که هر یک شامل اقلام داده‌ای از مجموعه اقلام داده I است. بنابراین هر تراکنش t_i از مجموعه‌ای از اقلام داده تشکیل شده است؛ به گونه‌ای که $t_i \subseteq I$. اگر X و Y اقلام داده فرض شود، قاعده همبش مفهومی به شکل $X \rightarrow Y$ است که در آن $X \subseteq I, Y \subseteq I$ و $X \cap Y = \emptyset$ (لیو، ۲۰۰۷: ۵۷-۶۰).

در قاعده همبشی که به شکل $X \rightarrow Y$ باشد؛ X را «مقدم» و Y را «نتیجه (تالی)» می‌نامند. واضح است که مقدار مقدم، مقدار نتیجه را دربرمی‌گیرد. دامنه پشتیبانی از قاعده و میزان اعتماد به قاعده، مهم‌ترین معیارهای کیفی برای ارزیابی جالب بودن قاعده در نظر گرفته شده‌اند.

• دامنه پشتیبانی از قاعده (Support)

نحوه محاسبه دامنه پشتیبانی از قاعده به صورت زیر است.

$$\text{sup}(A \Rightarrow B) = \frac{A \cap B}{X} \quad \text{رابطه (۱)}$$

که در آن، A و B دو قلم داده متفاوت در پایگاه داده و X کل اقلام‌هایی است که در پایگاه داده وجود دارد. این قاعده تمام تراکنش‌هایی را که در آنها دو قلم داده A و B وجود دارد، استخراج می‌کند و با مقدار حداقل دامنه پشتیبانی از قاعده که توسط کاربر مشخص می‌شود، مقایسه می‌کند و پس از آن فقط تراکنش‌هایی را انتخاب می‌کند که دامنه پشتیبانی از قاعده‌شان بزرگ‌تر یا مساوی حداقل دامنه پشتیبانی از قاعده باشد و باقی تراکنش‌ها قواعد به‌دردنخوری (غیر جالبی) هستند که حذف می‌شوند.

• میزان اعتماد به قاعده (Confidence)

معیار میزان اعتماد به قاعده در فهرست جدید به صورت رابطه ۲ اعمال می‌شود.

$$\text{conf}(A \Rightarrow B) = P(B|A) = \frac{\text{sup}(A \rightarrow B)}{\text{sup}(A)}; \quad \text{sup}(A) = \frac{A}{X} \quad \text{رابطه (۲)}$$

ابتدا تراکنش‌های حاوی قلم داده A محاسبه می‌شود و از بین آنها، تراکنش‌های حاوی قلم داده B استخراج می‌شود، سپس به مقایسه خروجی این رابطه با حداقل میزان اعتماد به قاعده پرداخته خواهد شد. قواعدی که از این فیلتر عبور کنند، قواعد همباش نامیده می‌شوند (آگراوال و شافر، ۱۹۹۶).

پیشینه تجربی

الگوریتم ابریوری کلاسیک‌ترین و محبوب‌ترین الگوریتم برای استخراج قواعد همباش از پایگاه داده است که آگراوال و سریکانت آن را در سال ۱۹۹۴ برای استخراج قواعد همباش از مجموعه اقلام داده پرتکرار در پایگاه داده‌های تراکنشی با استفاده از دو معیار حد آستانه از قبل تعریف شده به نام‌های دامنه پشتیبانی از قاعده و میزان اعتماد به قاعده، توسعه دادند.

از زمانی که ایده تولید قواعد همباش مطرح شده است تا کنون، محققان الگوریتم‌های متعددی برای انجام این مهم معرفی کردند که همه آنها در تلاش برای تولید قواعد همباش مفیدتر بودند (هان، چنگ، شین و یان، ۲۰۰۷). الگوریتم‌های قواعد همباش کاوی در زمینه‌های گوناگونی در کانون توجه قرار گرفته است. در ادامه به برخی از این مطالعات پرداخته می‌شود.

هیو، وو و لایو (۲۰۱۳) ابتدا یک ساختار داده‌ای فشرده با عنوان PLMS-Tree برای ذخیره و فشرده‌سازی پایگاه داده رابطه‌ای، در قالب درختی معرفی کردند و پس از آن، الگوریتم

MSCP-Growth را بر اساس آن طراحی و ارائه کردند که به کاربران اجازه می‌دهد حداقل دامنه پشتیبانی از قاعده‌های متفاوت را برای اقلام داده بر اساس فرکانس طبیعی‌شان تعیین کند. در یک ساختار مشابه، هیو و چن FP-Tree را با عنوان MIS-Tree به منظور ذخیره اطلاعات پیچیده برای الگوهای پرتکرار طراحی کردند و بر اساس این درخت، الگوریتم CFP-Growth را به منظور کشف مجموعه اقلام داده پرتکرار ارائه دادند. سپس با بیان اینکه به کارگیری حداقل دامنه پشتیبانی از قاعده یکسان برای اقلام داده‌ای که هر یک فرکانس خاصی دارند، مشکل خواهد بود، راهکاری جدید برای تنظیم دامنه پشتیبانی از قاعده اقلام داده معرفی کردند و با اجرای مکرر الگوریتم به پایانی رضایت‌بخش دست یافتند (هیو و چن، ۲۰۰۶).

لی و همکارانش نوعی الگوریتم کاوش چندسطحی فازی، برای استخراج دانش از پایگاه داده‌های تراکنشی به کمک حداقل دامنه پشتیبانی از قاعده چندگانه معرفی کردند. در این الگوریتم از یک بستر بالا به پایین برای دستیابی به اقلام داده پرتکرار بهره بردند. پیاده‌سازی این الگوریتم در مطالعه موردی، به استخراج قواعد همبستگی در سطوح مختلف تحت دامنه پشتیبانی از قاعده چندگانه انجامید (لی، هونگ و وانگ، ۲۰۰۸).

در مطالعه تسنگ و لین (۲۰۰۷) بیان شده است تعیین حداقل دامنه پشتیبانی از قاعده یکسان برای تمام اقلام داده‌ای که در سطح مشابهی طبقه‌بندی شده‌اند، به از دست رفتن برخی اطلاعات جالب در بین اقلام داده منجر خواهد شد. آنها با مطرح کردن مشکلات استفاده از الگوریتم اپریوری برای کشف الگوهای پرتکرار، الگوریتم‌های MMS-Cumulate و MMS-Stratify را ارائه کردند. این الگوریتم‌ها با استفاده از حداقل دامنه پشتیبانی از قاعده‌های متعدد، الگوهای پرتکرار را از مجموعه داده استخراج می‌کنند.

در مطالعه هوانگ (۲۰۱۳) تلاش شد الگوریتم FQSP بهبود داده شود. محقق معتقد است اولاً از آنجاکه FQSP از حداقل دامنه پشتیبانی از قاعده واحد برای کشف الگوهای پرتکرار استفاده می‌کند؛ اگر حداقل دامنه پشتیبانی از قاعده، عددی بزرگ در نظر گرفته شود، اقلام داده غیرجالب بسیاری تولید خواهد شد و اگر عددی پایین در نظر گرفته شود، به از دست رفتن اقلام داده‌ای می‌انجامد که می‌تواند جالب باشد. از این رو در رویکرد جدید، روش حداقل دامنه پشتیبانی از قاعده چندگانه را به کار برد. دوم، با توجه به اینکه FQSP، فقط از تابع عضویت واحد استفاده می‌کند و عملکردی آشفته‌ای دارد، در رویکرد جدید برای پوشش این مشکل، ایده تابع عضویت قابل تنظیم را ارائه کرد.

در پژوهشی تلاش شده است الگوهای جزئی دوره‌ای با استفاده از حداقل دامنه پشتیبانی از قاعده چندگانه استخراج شود. در این پژوهش ابتدا به مشکل کارهای قبلی برای استخراج الگوهای جزئی دوره‌ای مبنی بر مشکل ناکارآمدی ناشی از به کارگیری حداقل دامنه پشتیبانی از

قاعده واحد، اشاره شده است. سپس برای غلبه بر این مشکل، الگوریتمی ارائه شد که نیاز به تولید الگوهای جزئی دوره‌ای کاندید را برطرف می‌کند و تعداد بازدیدهای متوالی از پایگاه داده را کاهش می‌دهد. سپس برای هر رویداد بر اساس فرکانس رخ دادنش، حداقل دامنه پشتیبانی از قاعده، تعیین شده است (پی، کولیت و تیل، ۲۰۰۷).

چن، هونگ و تسنگ (۲۰۰۹) تلاش کردند کار قبلی خود را مبنی بر الگوریتم داده‌کاوی ژنتیک فازی برای استخراج حداقل دامنه پشتیبانی از قاعده و توابع عضویت برای اقلام داده در پایگاه داده کمی، بهبود بخشند. آنها الگوریتم بهبودیافته را که بر اساس تقسیم و حل کار می‌کند، DGFMMMS نامیدند. در این روش، از الگوریتم ژنتیک فازی برای استخراج قواعد همباش فازی استفاده شده است.

با توجه به اینکه بررسی کامل الگوریتم‌ها و رویکردهای حوزه استخراج قواعد همباش خارج از حوصله این پژوهش است، به همین مقدار بسنده می‌شود. با این حال در مقاله هان، چنگ، شین و یان (۲۰۰۷)، نگاه ویژه و مفصلی به رویکردهای ارائه‌شده در حوزه استخراج قواعد همباش و کشف الگوهای پرتکرار تا سال ۱۳۸۶ شده است.

نکته حائز اهمیتی که انگیزه ارائه مقاله حاضر بوده است، اینکه در هر دو حالت به‌کارگیری حداقل دامنه پشتیبانی از قاعده واحد و چندگانه، بدون استثنا از کاربر درخواست می‌شود حداقل دامنه پشتیبانی از قاعده را تعیین کند. فرض کنید کاربر قصد اعمال الگوریتم اپریوری را بر پایگاه داده‌ای با میلیون‌ها تراکنش داشته باشد؛ به‌طور مسلم کاربر نمی‌تواند دانش لازم را درباره تمام تراکنش‌های موجود در پایگاه داده داشته باشد و بنابراین نمی‌تواند حداقل دامنه پشتیبانی از قاعده مناسبی را تعیین کند. در این مطالعه با استفاده از روش میانگین‌گیری، روشی معرفی شده است که در آن الگوریتم اپریوری به‌صورت کاملاً خودکار این حد آستانه را تعیین می‌کند.

فازی‌سازی

لطفی‌زاده، اولین فردی بود که نظریه مجموعه‌های فازی را مطرح کرد. این نظریه چارچوب ریاضی محکمی برای مطالعه پدیده‌ها و مفاهیم مبهم، مهیا می‌کند؛ زبان مدل‌سازی محسوب می‌شود و هنگامی که روابط و محدودیت‌ها فازی می‌شوند، بسیار مناسب است (آذر، سنگی، ایزدخواه و انوری، ۱۳۹۴). در این پژوهش از روش خوشه‌بندی فازی استفاده می‌شود. با بهره‌مندی از این روش می‌توان داده‌های قطعی را وارد مجموعه‌های فازی کرد. از میان الگوریتم‌هایی که برای خوشه‌بندی فازی وجود دارد، در این مطالعه از الگوریتم محبوب خوشه‌بندی فازی C_means (بزدک، ۱۹۸۱: ۱۵۰-۱۶۲ و شیهاب و بورگر، ۱۹۸۸) بهره برده شده است.

قواعد همبانش فازی

پس از انتقال داده‌ها به محیط فازی، به استخراج قواعد همبانش فازی از بین این داده‌ها اقدام خواهد شد. این کار به کمک رویکرد اپریوری صورت خواهد گرفت؛ همان‌طور که پیش از این بیان شد، دو معیار اندازه‌گیری به نام‌های «دامنه پشتیبانی از قاعده» و «میزان اعتماد به قاعده» در استخراج قواعد همبانش استفاده می‌شود که در محیط فازی این معیارها به صورت زیر مطرح می‌شوند:

- دامنه پشتیبانی از قاعده فازی

در اولین گام برای تعیین تعداد تکرار هر قلم داده در پایگاه داده فازی‌سازی شده، مجموع درجه عضویت‌ها برای هر داده محاسبه می‌شود که این رقم میزان تکرار هر داده در نظر گرفته خواهد شد. رابطه ۳ گویای مطلب فوق است.

$$\text{fuzzysum} = \sum_{(x) \in D} \mu(x) \quad \text{رابطه ۳}$$

سپس برای تولید مجموعه ارقام دو - داده‌ای، باید پس از مقایسه تمام درجه عضویت‌های هر دو قلم داده با هم، حداقل^۱ را از بین آنها انتخاب کرد. در پایان، مجموع حداقل‌ها میزان تکرار دو قلم داده مد نظر معرفی می‌شود. برای تولید مجموعه ارقام سه - داده‌ای باید پس از مقایسه تمام درجه عضویت‌های هر سه داده با هم، از بین آنها حداقل را انتخاب کرد که مجموع حداقل‌ها، میزان تکرار سه قلم داده مد نظر معرفی می‌شود. این کار تا زمانی ادامه پیدا خواهد کرد که داده پرتکرار جدیدی تولید نشود. معمول‌ترین انتخاب برای عملگر t-norm (جانگ، سون و میزوتانی، ۱۹۷۷: ۱۲۱-۱۱۷)، یافتن حداقل در بین مجموعه‌ها است که در پژوهش پیش رو از این عملگر استفاده شده است. رابطه ۴ گویای مطلب فوق است.

$$\text{fuzzysup}(A \rightarrow B) = \sum_{(x) \in D} T[A(x), B(y)] \quad \text{رابطه ۴}$$

$T[A(x), B(y)]$ بیانگر t-norm است و حداقل مقدار $A(x)$ و $B(y)$ را محاسبه می‌کند.

- میزان اعتماد به قاعده فازی

پس از استخراج الگوهای پرتکرار فازی، برای تولید قواعد همبانش تلاش خواهد شد که این کار مستلزم استفاده از میزان اعتماد به قاعده فازی است. قواعدی که از این فیلتر عبور کنند، قواعد

همباش فازی معرفی خواهند شد. در این مطالعه از عملگر t-norm برای محاسبه میزان اعتماد به قاعده فازی استفاده شده است. رابطه ۵ مطلب فوق را به صورت رسمی بیان می کند.

$$\text{fuzzyconf}(A \Rightarrow B) = \frac{\sum_{(x,y) \in D} T[A(x), B(y)]}{\sum_{(x) \in D} \min[A(x)]} \quad (\text{رابطه ۵})$$

نحوه محاسبه حداقل دامنه پشتیبانی از قاعده با روش پیشنهادی

این پژوهش تلاش می کند الگوریتم، به صورت خودکار حداقل آستانه مناسبی را برای معیار اندازه گیری حداقل دامنه پشتیبانی از قاعده به کاربر معرفی کند. هنگامی که متخصصان بخواهند مجموعه ای از داده ها را با شاخصی مقایسه کنند و چنین شاخصی وجود نداشته باشد که داده ها نسبت به آن سنجیده شود و داده هایی که از این شاخص بزرگ تر (بهتر، قوی تر و ...) یا کوچک ترند (بدتر، ضعیف تر و ...) را دسته بندی کند، از روش های آماری کمک می گیرند و با توجه به معادله هایی مانند میانگین گیری، انحراف معیار، میانه، مد، واریانس و ... شاخصی را به منظور مقایسه داده ها تعریف می کنند و با این کار اطلاعات را به دست می آورند.

از آنجا که هیچ شاخص مناسبی وجود ندارد که بتوان حداقل آستانه مناسبی را برای معیار اندازه گیری دامنه پشتیبانی از قاعده با توجه به آن تعریف کرد، در این مطالعه نیز از روش های آماری استفاده شده است. از بین روش هایی که در این حوزه وجود دارد، روش میانگین گیری انتخاب شده است تا حداقل آستانه مناسب برای استخراج الگوهای پرتکرار تعریف شود؛ به طوری که این حداقل آستانه، نه خیلی پایین باشد که به تولید حجم زیادی از الگوهای به دردنخور منجر شود و نه خیلی بالا در نظر گرفته شود که به از دست دادن الگوهای جالب بینجامد. به کارگیری این ایده می تواند سبب بهبود روند کشف الگوهای پرتکرار و استخراج قواعد همباش شود. پس از انتقال داده ها به محیط فازی توسط الگوریتم خوشه بندی FCM، از طریق رابطه ۶ مجموع تمام درجه های عضویت برای هر فیلد محاسبه می شود و این رقم بر تعداد کل اقلام داده (T) تقسیم خواهد شد. عدد به دست آمده حداقل دامنه پشتیبانی از قاعده معرفی می شود.

$$\text{fuzzy min sup}(A, B) = \frac{\sum_{(x,y) \in D} [\text{sum}(f_A(x), f_B(y))]}{|T|} \quad (\text{رابطه ۶})$$

در واقع، در این پژوهش دخالت کاربر در روند اجرای الگوریتم اپریوری حذف شده است و ادعا می شود که با ارائه این روش، الگوریتم اپریوری به الگوریتمی خودکار تبدیل شده است.

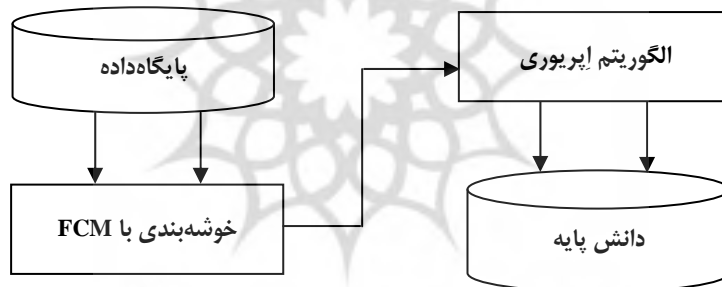
روش‌شناسی پژوهش

به‌طور کلی رویکرد پیشنهادی به‌منظور بهبود الگوریتم اپریوری در دو مرحله اجرا می‌شود:

- تبدیل داده‌های قطعی به فازی که از الگوریتم خوشه‌بندی FCM استفاده خواهد شد؛
- رویکرد اپریوری با بهبود در نحوه تعیین حداقل دامنه پشتیبانی از قاعده خودکار که از مجموعه داده‌های فازی برای کشف الگوهای پرتکرار و استخراج قواعد همباش، استفاده خواهد شد.

شکل ۲ مراحل اجرای کار را در قالب مدل نشان می‌دهد.

۱. دریافت مجموعه داده مد نظر از پایگاه داده که به‌منظور هماهنگی و حذف داده‌های پرت بر اساس قوانین پیش‌پردازش، تجزیه و تحلیل شده است؛
۲. انتقال مجموعه داده‌های قطعی به محیط فازی با استفاده از الگوریتم FCM؛
۳. استفاده از الگوریتم اپریوری با تکیه بر روش پیشنهادی برای تعیین حداقل دامنه پشتیبانی از قاعده به‌منظور کشف الگوهای پرتکرار و استخراج قواعد همباش و درنهایت، ذخیره قواعد با نام KB.



شکل ۲. مدل رویکرد پیشنهادی

در ادامه به تعریف مفاهیمی که در رویکرد پیشنهادی به کار می‌رود، پرداخته شده است:

- فیلد: عبارت است از صفت، قلم داده یا ستون جدول؛
- رکورد: عبارت است از سطر با همه فیلدها؛
- $\mu(X)$: مقدار مجموعه فازی که عددی خواهد بود در بازه $[0, 1]$ ؛
- C_k : معرف مجموعه اقلام داده کاندید است که $1 \leq k \leq n$ ؛
- L_k : معرف مجموعه اقلام داده پرتکرار است که $1 \leq k \leq n$ ؛
- Fuzzy minsup: معیار اندازه‌گیری حداقل دامنه پشتیبانی از قاعده فازی؛
- Fuzzy minconf: معیار اندازه‌گیری حداقل میزان اعتماد به قاعده فازی.

مراحل ۱ تا ۱۲ شبه کد رویکرد پیشنهادی ارائه شده در شکل ۳ را با جزئیات توضیح می دهند:

۱. FCM پس از خوشه بندی داده های قطعی، مرکز هر مجموعه فازی را تعیین می کند و مقادیر حداکثر و حداقل را برای هر فیلد از مجموعه داده ورودی پیدا می کند؛
۲. با استفاده از trimf از ارقام داده میان خوشه ها توزیع می شود؛
۳. حداقل دامنه پشتیبانی از قاعده با استفاده از رابطه ۶ به صورت خودکار محاسبه می شود؛
۴. دامنه پشتیبانی از قاعده هر مجموعه قلم داده برای هر رکورد با استفاده از رابطه ۳ به دست می آید، سپس در فهرست مجموعه ارقام داده کاندید C_1 مرتب و ذخیره می شود؛
۵. از بین مجموعه ارقام داده کاندید C_1 ، آنهایی که بتوانند شرط حداقل دامنه پشتیبانی از قاعده را برآورده کنند، مجموعه ارقام داده پرتکرارند که در L_1 مرتب و ذخیره می شوند؛
۶. ارقام داده موجود در L_1 به صورت زیر با هم الحاق می شوند:

$$(L_1 \text{ Join } L_1) = \{ \{C[1], C[i], \{C[1], C[i+1]\} \dots \{C[1], C[n]\} \}$$
 که $C[1]$ اولین مجموعه فازی، $C[i]$ دومین مجموعه فازی و $C[n]$ آخرین مجموعه فازی است و $(C[1] \cap C[i]=\emptyset, C[1] \cap C[i+1]=\emptyset, \dots, C[1] \cap C[n]=\emptyset)$. نتیجه در فهرست مجموعه ارقام داده کاندید C_2 ذخیره می شود. دامنه پشتیبانی از قاعده برای هر مجموعه قلم داده کاندید با استفاده از رابطه ۴ محاسبه می شود؛
۷. از بین مجموعه ارقام داده کاندید C_2 ، آنهایی که بتوانند شرط حداقل دامنه پشتیبانی از قاعده را برآورده کنند، مجموعه ارقام داده پرتکرارند که در L_2 مرتب و ذخیره می شوند؛
۸. ارقام داده هایی که در L_2 هستند با هم الحاق می شوند. باید تمام زیرمجموعه های C_2 به مثابه مجموعه ارقام داده پرتکرار در L_1 حضور داشته باشند، یعنی تمام زیرمجموعه های C_k به مثابه مجموعه ارقام داده پرتکرار در L_{k-1} حضور داشته باشند؛
۹. نتیجه عملیات الحاق مرحله قبل در فهرست مجموعه ارقام داده کاندید C_3 ذخیره می شود. دامنه پشتیبانی از قاعده برای هر مجموعه قلم داده کاندید محاسبه می شود؛
۱۰. ارقام داده کاندید C_3 که شرط حداقل دامنه پشتیبانی از قاعده را برآورده کنند به L_3 انتقال داده می شوند؛
۱۱. این فرایند تا زمانی که L_k خالی شود، تکرار می شود؛

۱۲. الگوهای پرتکرار کشف شده در قالب (IF-Then) مرتب می شوند و میزان اعتماد به قاعده با استفاده از رابطه ۵ به دست می آید و نتیجه در KB ذخیره می شود.

```

Begin
Find the fuzzy set of the quantitative data set, based on FCM.
Calculate the fuzzy minsup value using equation 6.
Calculate the summation of the membership value for each fuzzy set with all records using Equation 3.
    IF fuzzysum  $\geq$  fuzzy minsup Then
        Insert the fuzzy set into  $L_1$ ,  $L_1 = \{\text{frequent itemset}\}$ 
For ( $k = 2$ ;  $L_{k-1} \neq \emptyset$ ;  $k++$ ) do
     $C_k =$  generate candidate from  $L_{k-1}$  (join  $L_{k-1}$  call (p) with  $L_{k-1}$  call (q));
    {
    Insert into  $C_k$ 
    Select itemset;  $p.item_1, p.item_2 \dots p.item_{k-1}, q.item_{k-1}$  From p, q
    Where  $p.item_1 = q.item_1, p.item_2 = q.item_2 \dots p.item_{k-1} = q.item_{k-1}$ 
    }
        For each  $c \in C_k$  do
    Check all subsets of all itemsets in  $C_k$ , should be frequent itemsets in  $L_{k-1}$ 
        For each (k-1) subset s of c do
            IF  $s \notin L_{k-1}$  Then Delete c from  $C_k$ 
            ENDIF
        EndFor
    EndFor
        For each itemset candidate in  $C_k$  do
            Calculate the fuzzy support value using Equation 4.
            Insert the fuzzy set into  $L_k$ ,  $L_k = \{\text{frequent itemset}\}$ 
        EndFor
    EndFor
    Select the frequent itemsets that exist in  $L_2$  to  $L_k$  under form "IF-Then".
    For each frequent itemset
        Calculate the fuzzy confidence for frequent itemset using Equation 5.
    Endfor
EndBegin

```

شکل ۳. شبه کد رویکرد پیشنهادی

یافته‌های پژوهش

برای تجزیه و تحلیل و اعتبارسنجی اهداف، رویکردی که در بخش قبل مطرح شد، بر مجموعه داده انتخاب شده از یک کتابخانه الکترونیک، اعمال شده است.

پیاده‌سازی چگونگی کارکرد رویکرد پیشنهادی

با توجه به سطح اهمیت کتابخانه‌های الکترونیک میان مؤسسه‌های تحقیقاتی و دانشگاهی، در این پژوهش چگونگی استخراج قواعد همبانش فازی با استفاده از رویکرد پیشنهادی و با کمک گرفتن از داده‌های یکی از کتابخانه‌های الکترونیک توصیف می‌شود. در این مطالعه اطلاعات با توجه به اطلاعات مقاله جلیل منش و همکارش با عنوان «نگاشت دانش سازمانی بر اساس سیستم اطلاعاتی کتابخانه» به دست آمده است (جلیل منش و همایون والا، ۱۳۹۰). در کل هدف از ارائه این مثال، یافتن ارتباطات جالب بین کتاب‌های یکی از کتابخانه‌های الکترونیک با استفاده از رویکرد پیشنهادی بوده است. با تجزیه و تحلیل کتاب‌هایی که محتمل است با هم به امانت برده شوند، اطلاعات مفیدی برای محققان و کاربران به دست می‌آید و می‌توانند در کمترین زمان جست‌وجو، بهترین نتایج را استخراج کنند و این امکان در اختیار آنها قرار خواهد گرفت که بدانند موضوع مد نظرشان با چه موضوعات دیگری در ارتباط است. در این مطالعه برای تشکیل انبار داده، اطلاعات از این پایگاه داده دریافت شده است. جدول ۱ نمونه‌ای از این انبار داده را نشان می‌دهد.

جدول ۱. نمونه‌ای از انبار داده کتابخانه الکترونیک

| ردیف | نام | ژانویه | اقلام داده | ... |
|------|-------|--------|--|-----|
| ۱ | ناصری | ۱ | تکنولوژی نانو، زیست‌شناسی، فیزیک، ... | ... |
| ۲ | شفیعی | ۲ | کامپیوتر، مکانیک، ... | ... |
| ۳ | علوی | ۳ | فیزیک، بیولوژی، الکترونیک، کامپیوتر، ... | ... |
| ۴ | اصغری | ۴ | الکترونیک، مکانیک، تکنولوژی نانو، ... | ... |
| ۵ | کبیری | ۵ | فیزیک، شیمی، تکنولوژی نانو، ... | ... |
| ۶ | اکبری | ۶ | الکترونیک، ... | ... |
| ۷ | حیدری | ۷ | فیزیک، بیولوژی، کامپیوتر، شیمی، ... | ... |
| ۸ | محسنی | ۸ | شیمی، معماری، ... | ... |
| ۹ | ناصری | ۹ | فیزیک، زیست‌شناسی، تکنولوژی نانو، ... | ... |
| ۱۰ | شفیعی | ۱۰ | فیزیک، زیست‌شناسی، کامپیوتر، مکانیک | ... |
| ... | ... | ... | ... | ... |

وظایف آماده‌سازی داده‌ها در چند دوره انجام می‌گیرد و هیچ تعریف از پیش تعیین شده‌ای ندارد. این وظایف شامل انتخاب جدول، رکوردها، خصیصه‌ها و همچنین انتقال و پاک‌سازی داده برای مدل‌سازی است (رادفر، نضافتی و یوسفی اصلی، ۱۳۹۳). در مرحله پیش‌پردازش، فقط اطلاعات لازم برای استخراج قواعد همبش فازی، از این انبار داده استخراج خواهد شد؛ به این صورت که در هر تراکنش چه کتاب‌هایی حضور داشته‌اند. جدول ۲ اطلاعات قطعی پیش‌پردازش شده‌ای را که آماده فازی‌سازی‌اند، نشان می‌دهد. سپس قسمتی از این مجموعه داده برای استخراج قواعد همبش انتخاب می‌شود. جدول ۳ یکسری اطلاعات آماری از مقادیر حداقل، حداکثر و مراکز داده‌های ورودی را نشان می‌دهد.

جدول ۲. نمونه‌ای از مجموعه داده قطعی کتابخانه الکترونیک

| ردیف | ماه | فیزیک | زیست‌شناسی | الکترونیک | کامپیوتر | شیمی | ... |
|------|---------|-------|------------|-----------|----------|------|-----|
| ۱ | ژانویه | ۱ | ۱۲ | ۲۰ | ۵ | ۰ | ... |
| ۲ | فوریه | ۵ | ۲ | ۱۲ | ۱۳ | ۱۶ | ... |
| ۳ | مارس | ۱۱ | ۴ | ۱۱ | ۱۸ | ۱۸ | ... |
| ۴ | آوریل | ۲ | ۰ | ۱ | ۲ | ۵ | ... |
| ۵ | مه | ۳ | ۱ | ۸ | ۶ | ۱۲ | ... |
| ۶ | ژوئن | ۰ | ۶ | ۳ | ۷ | ۱۸ | ... |
| ۷ | جولای | ۱ | ۱۵ | ۷ | ۱۹ | ۱۴ | ... |
| ۸ | اگوست | ۸ | ۰ | ۶ | ۲۱ | ۰ | ... |
| ۹ | سپتامبر | ۴ | ۸ | ۴ | ۱ | ۱۸ | ... |
| ۱۰ | اکتبر | ۷ | ۷ | ۰ | ۱۶ | ۱۴ | ... |
| ... | ... | ... | ... | ... | ... | ... | ... |

جدول ۳. اطلاعات آماری از مجموعه داده قطعی کتابخانه الکترونیک

| | کامپیوتر | الکترونیک | زیست‌شناسی | فیزیک | حداقل |
|--------|----------|-----------|------------|---------|-------|
| حداکثر | ۲۱ | ۲۰ | ۱۵ | ۱۱ | ۰ |
| مرکز ۱ | ۰/۹۸۸۲ | ۱/۷۲۷۵ | ۰/۲۵۴۵ | ۰/۷۱۱۵ | ۰ |
| مرکز ۲ | ۶/۰۱۶۵ | ۷/۱۵۷۴ | ۳/۱۷۴۶ | ۳/۵۹۱۷ | ۰ |
| مرکز ۳ | ۱۴/۲۲۹۴ | ۱۱/۴۶۸۲ | ۷/۰۷۸۷ | ۷/۰۴۶۹ | ۰ |
| مرکز ۴ | ۱۹/۲۹۱۱ | ۱۹/۹۹۴۹ | ۱۳/۶۵۵۳ | ۱۰/۴۶۲۲ | ۰ |

حال مجموعه داده‌های قطعی به محیط فازی منتقل می‌شود که برای این کار از الگوریتم خوشه‌بندی FCM استفاده شده است. در این مطالعه موردی، برای هر فیلد چهار خوشه با عنوان‌های «ضعیف»، «متوسط»، «خوب» و «عالی»^۴ تعریف شده است و تمام ارقام داده در این چهار خوشه توزیع شده‌اند. برای خلاصه‌کردن نام خوشه‌ها از کلمه اول نام هر خوشه استفاده شده است که به ترتیب عبارت‌اند از: (W, M, G, T). جدول ۴، هر فیلد را با خوشه‌های متناظر و تابع عضویتش نشان می‌دهد.

جدول ۴. فیلدها و مجموعه‌های فازی متناظر با آن‌ها

| حوزه دانشی | مجموعه فازی | تابع عضویت |
|------------|---|------------|
| فیزیک | (W _۱)، متوسط (M _۲)، خوب (G _۳)، عالی (T _۴) | Trimf |
| زیست‌شناسی | (W _۵)، متوسط (M _۶)، خوب (G _۷)، عالی (T _۸) | Trimf |
| الکترونیک | (W _۹)، متوسط (M _{۱۰})، خوب (G _{۱۱})، عالی (T _{۱۲}) | Trimf |
| کامپیوتر | (W _{۱۳})، متوسط (M _{۱۴})، خوب (G _{۱۵})، عالی (T _{۱۶}) | Trimf |

درجه عضویت برای تمام قسمت‌هایی که مجموعه‌ها با هم تداخلی نداشته باشند، برابر با ۱ است، اما در مناطقی که مجموعه‌ها با هم تداخل دارند، تعیین درجه عضویت به این بستگی دارد که در مرز بالای مجموعه قرار دارد یا مرز پایین. به‌طور مثال برای بیان درجه عضویت فازی فیلد فیزیک (فرض کنید p) از رابطه‌های ۷ تا ۱۰ استفاده شده است.

$$\mu_{\text{Physics.Middle}}(p) = \begin{cases} 1 & p \leq 0.7 \\ \frac{3/6 - p}{2/9} & 0.7 \leq p \leq 3/6 \end{cases} \quad \text{رابطه ۷}$$

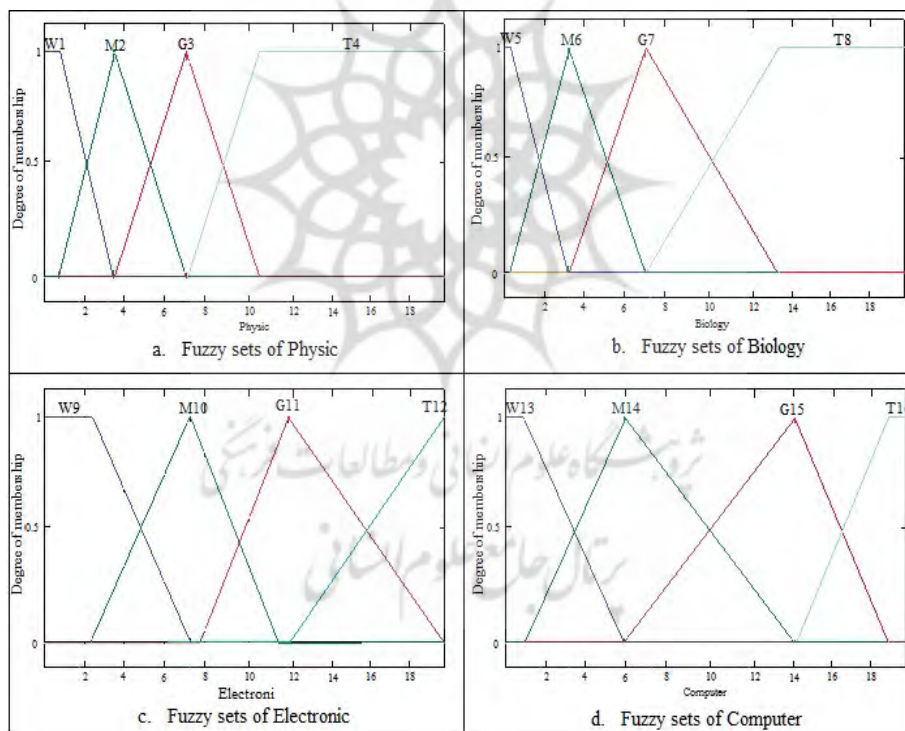
$$\mu_{\text{Physics.Middle}}(p) = \begin{cases} \frac{p - 0.7}{2/9} & 0.7 \leq p \leq 3/6 \\ \frac{7/1 - p}{3/5} & 3/6 \leq p \leq 7/1 \end{cases} \quad \text{رابطه ۸}$$

1. Weak
2. Middle
3. Good
4. Top

$$\mu_{\text{Physics.Good}}(p) = \begin{cases} \frac{p - 3/6}{3/5} & 3/6 \leq p \leq 7/1 \\ \frac{10/5 - p}{3/4} & 7/1 \leq p \leq 10/5 \end{cases} \quad \text{رابطه ۹}$$

$$\mu_{\text{Physics.Top}}(p) = \begin{cases} \frac{p - 7/1}{3/4} & 7/1 \leq p \leq 10/5 \\ 1 & p \geq 10/5 \end{cases} \quad \text{رابطه ۱۰}$$

شکل ۴ خوشه‌بندی فازی با استفاده از الگوریتم FCM و محدوده‌های خوشه‌ها را نشان می‌دهد. همچنین جدول ۵ توزیع برخی داده‌ها را میان خوشه‌ها نشان می‌دهد.



شکل ۴. خوشه‌بندی فازی کتابخانه الکترونیک. الف) مجموعه‌های فازی فیلد فیزیک؛ ب) مجموعه‌های فازی فیلد زیست‌شناسی؛ ج) مجموعه‌های فازی فیلد الکترونیک؛ د) مجموعه‌های فازی فیلد کامپیوتر

جدول ۵. توزیع برخی داده‌ها میان خوشه‌ها (کتابخانه الکترونیک)

| الکترونیک | زیست‌شناسی | | | | | فیزیک | | | |
|-----------|------------|-------|-------|-------|-------|-------|-------|-------|-------|
| | W{۹} | T{۸} | G{۷} | M{۶} | W{۵} | T{۴} | G{۳} | M{۲} | W{۱} |
| ... | ۰ | ۰/۸۵۶ | ۰/۰۹۷ | ۰/۰۳ | ۰/۰۱۷ | ۰/۰۰۱ | ۰/۰۰۳ | ۰/۰۱۲ | ۰/۹۸۵ |
| ... | ۰/۰۰۳ | ۰/۰۰۷ | ۰/۰۳۵ | ۰/۶۵۹ | ۰/۲۹۹ | ۰/۰۴ | ۰/۲۸۷ | ۰/۶۰۷ | ۰/۰۶۶ |
| ... | ۰/۰۰۳ | ۰/۰۰۷ | ۰/۰۶۳ | ۰/۸۸۷ | ۰/۰۴۳ | ۰/۹۷۴ | ۰/۰۱۸ | ۰/۰۰۵ | ۰/۰۰۳ |
| ... | ۰/۹۸ | ۰ | ۰/۰۰۱ | ۰/۰۰۷ | ۰/۹۹۲ | ۰/۰۱۳ | ۰/۰۳۷ | ۰/۳۷۶ | ۰/۵۷۴ |
| ... | ۰/۰۱۷ | ۰/۰۰۴ | ۰/۰۱۳ | ۰/۱۰۳ | ۰/۸۸ | ۰/۰۰۵ | ۰/۰۲ | ۰/۹۱۴ | ۰/۰۶۱ |
| ... | ۰/۸۹۱ | ۰/۰۱۷ | ۰/۸۳۳ | ۰/۱۲۱ | ۰/۰۲۹ | ۰/۰۰۴ | ۰/۰۱ | ۰/۰۳۷ | ۰/۹۴۹ |
| ... | ۰/۰۰۱ | ۰/۹۵۲ | ۰/۰۲۷ | ۰/۰۱۳ | ۰/۰۰۸ | ۰/۰۰۱ | ۰/۰۰۲ | ۰/۰۱۲ | ۰/۹۸۵ |
| ... | ۰/۰۶۵ | ۰ | ۰/۰۰۱ | ۰/۰۰۷ | ۰/۹۹۲ | ۰/۱۲۴ | ۰/۸۲۴ | ۰/۰۳۸ | ۰/۰۱۴ |
| ... | ۰/۶۱۳ | ۰/۰۲۵ | ۰/۹۲۸ | ۰/۰۳۴ | ۰/۰۱۳ | ۰/۰۰۴ | ۰/۰۱۷ | ۰/۹۶۴ | ۰/۰۱۵ |
| ... | ۰/۹۱۹ | ۰ | ۱ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ |
| ... | ۰/۰۱۷ | ۰ | ۰/۰۰۲ | ۰/۹۹۴ | ۰/۰۰۴ | ۰ | ۰ | ۰/۰۱ | ۰/۹۹ |
| ... | ۰/۹۹۶ | ۰ | ۰/۰۰۱ | ۰/۰۰۷ | ۰/۹۹۲ | ۰/۰۰۳ | ۰/۰۰۱ | ۰/۰۳۷ | ۰/۹۴۹ |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

از این مرحله به بعد است که از الگوریتم اپریوری برای استخراج الگوهای پرتکرار و تولید قواعد همبش استفاده می‌شود. با توجه به رابطه ϵ معیار اندازه‌گیری حداقل دامنه پشتیبانی از قاعده در این مطالعه موردی برابر با $\text{Fuzzyminsup} = ۰/۲۶۳$ به دست آمده است. جدول‌های ۶ تا ۱۲ کشف اقلام داده پرتکرار را با استفاده از رویکرد پیشنهادی به تصویر می‌کشند.

| جدول ۶. قسمتی از C_1 | | جدول ۷. قسمتی از L_1 |
|-------------------------|----------|------------------------|
| دامنه پشتیبانی از قاعده | قلم داده | مجموعه داده |
| ۳/۶۵۲ | W{۱} | W{۱} |
| ۲/۹۶۵ | M{۲} | M{۲} |
| ۲/۲۱۸ | G{۳} | G{۳} |
| ۱/۱۶۶ | T{۴} | T{۴} |
| ۲/۲۸۱ | W{۵} | W{۵} |
| ... | ... | ... |

جدول ۸. نحوه محاسبه دامنه پشتیبانی از قاعده برای دو قلم داده

| دامنه پشتیبانی از قاعده | T{۸} | W{۱} |
|-------------------------|-------|-------|
| ۰/۸۵۶ | ۰/۸۵۶ | ۰/۹۸۵ |
| ۰/۰۰۷ | ۰/۰۰۷ | ۰/۰۶۶ |
| ۰/۰۰۳ | ۰/۰۰۷ | ۰/۰۰۳ |
| . | . | ۰/۵۷۴ |
| ۰/۰۰۴ | ۰/۰۰۴ | ۰/۰۶۱ |
| ... | ... | ... |
| $\Sigma = ۱/۸۵۴$ | ... | ... |

جدول ۱۰. قسمتی از L_2

| مجموعه داده | مجموعه داده | دامنه پشتیبانی از قاعده |
|-------------|-------------|-------------------------|
| W{۱}, M{۲} | W{۱}, M{۲} | ۰/۵۹۶ |
| W{۱}, M{۶} | W{۱}, G{۳} | ۰/۱۷ |
| W{۱}, G{۷} | W{۱}, T{۴} | ۰/۰۸۵ |
| W{۱}, T{۸} | W{۱}, W{۵} | ۰/۲۱۱ |
| W{۱}, W{۹} | W{۱}, M{۶} | ۰/۳۲۳ |
| W{۱}, M{10} | W{۱}, G{۷} | ۱/۰۲۵ |
| W{۱}, G{۱۱} | W{۱}, T{۸} | ۱/۸۵۴ |
| M{۲}, G{۳} | W{۱}, W{۹} | ۱/۵۱۸ |
| ... | ... | ... |

جدول ۱۲. قسمتی از L_3

| مجموعه داده | مجموعه داده | دامنه پشتیبانی از قاعده |
|-------------------|-------------------|-------------------------|
| W{۱}, M{۲}, W{۹} | W{۱}, M{۲}, M{۶} | ۰/۲۲ |
| M{۲}, G{۳}, W{۵} | W{۱}, M{۲}, G{۷} | ۰/۱۲۹ |
| M{۲}, G{۳}, M{۶} | W{۱}, M{۲}, W{۹} | ۰/۴۶۶ |
| M{۲}, G{۳}, M{۱۰} | W{۱}, M{۲}, M{۱۰} | ۰/۱۹۵ |
| M{۲}, W{۵}, M{۶} | M{۲}, G{۳}, W{۵} | ۰/۳۷۸ |
| M{۲}, W{۵}, M{۱۰} | M{۲}, G{۳}, M{۶} | ۰/۳۵۸ |
| M{۲}, W{۵}, G{۱۱} | M{۱}, G{۳}, G{۷} | ۰/۰۸۳ |
| W{۱}, M{۶}, M{۱۰} | M{۲}, G{۳}, W{۹} | ۰/۱ |
| ... | ... | ... |

تا این مرحله الگوهای پرتکرار با دامنه پشتیبانی از قاعده‌شان کشف شده‌اند. از این مرحله به بعد استخراج قواعد همبش آغاز خواهد شد. ابتدا تمام مجموعه ارقام داده پرتکرار موجود در L_2 تا L_n در قالب‌های IF-Then قرار می‌گیرند. سپس برای هر یک میزان اعتماد به قاعده براساس رابطه ۵ محاسبه می‌شود. جدول ۱۳ محاسبه این معیار را نشان می‌دهد.

جدول ۱۳. محاسبه معیار اندازه‌گیری میزان اعتماد به قاعده

| الگوهای پرتکرار | Form (x→y) | Support (x,y) | Support (x) | Conf = Support (x,y)/ Support (x) |
|------------------|--------------|---------------|-------------|--------------------------------------|
| {۱, ۸} | (۱→۸) | ۱/۸۵۴ | ۳/۶۵۲ | ۰/۵۰۸ |
| {۱۰, ۴} | (۱۰→۴) | ۱/۰۷۳ | ۲/۳۳۸ | ۰/۴۵۹ |
| {۱, ۲, ۹} | (۱→۲,۹) | ۰/۴۶۶ | ۳/۶۵۲ | ۰/۱۲۸ |
| {۳, ۹, ۷} | (۳,۹→۷) | ۰/۹۶۸ | ۱/۰۷۲ | ۰/۹۰۳ |
| {۶, ۱۰, ۳, ۵} | (۶,۱۰→۳,۵) | ۰/۳۵۶ | ۱/۷۶۸ | ۰/۲۰۱ |
| {۱, ۹, ۱۱, ۷} | (۱,۹,۱۱→۷) | ۰/۸۷ | ۰/۹۰۷ | ۰/۹۵۹ |
| {۵, ۲, ۳, ۶, ۱۰} | (۵→۲,۳,۶,۱۰) | ۰/۳۴۳ | ۲/۲۸۱ | ۰/۱۵ |
| {۵, ۶, ۱۰, ۲, ۳} | (۵,۶,۱۰→۲,۳) | ۰/۳۴۳ | ۰/۴۴۶ | ۰/۷۶۹ |

در نهایت جدول ۱۴ برخی از قواعد همبش تولیدشده با استفاده از روش پیشنهادی را نشان می‌دهد که در آن $Fuzzyminsup = ۰/۲۶۳$ و $Fuzzyminfer = ۰/۷$ ، در نظر گرفته شده‌اند.

جدول ۱۴. برخی از قواعد همبش تولید شده با استفاده از رویکرد پیشنهادی

| قواعد همبش | Support (x,y) | Support (x) | Confidence (x→y) |
|-------------|---------------|-------------|------------------|
| (۷→۹) | ۲/۳۸۷ | ۲/۹۹۸ | ۰/۷۹۶ |
| (۷,۱۱→۹) | ۱/۲۲۶ | ۱/۲۷۲ | ۰/۹۶۴ |
| (۷,۱۰→۹,۱۱) | ۰/۳ | ۰/۳۹۳ | ۰/۷۶۵ |
| (۱۰,۱۱→۹) | ۰/۳۴۷ | ۰/۴۰۴ | ۰/۸۵۹ |
| (۲,۳→۱۰) | ۰/۳۸۳ | ۰/۴۱۹ | ۰/۹۱۴ |
| (۱,۷,۹→۱۱) | ۰/۸۷ | ۰/۸۷ | ۱ |
| (۹,۱۰→۱۱) | ۰/۳۴۷ | ۰/۳۴۷ | ۱ |
| (۷,۹,۱۰→۱۱) | ۰/۳ | ۰/۳ | ۱ |

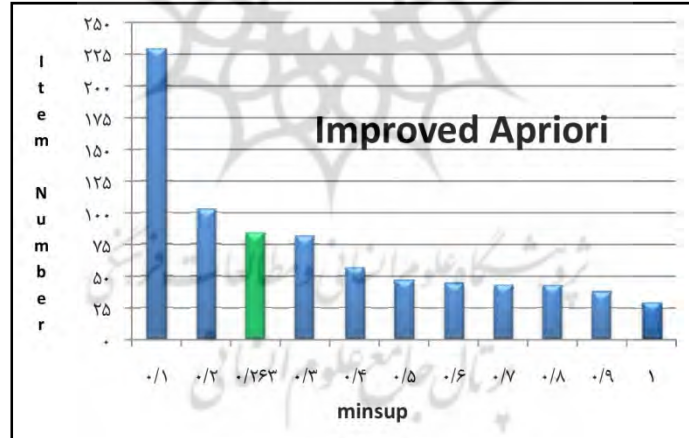
حال می‌توان به جای اعداد به‌کاررفته در این پژوهش، اقلام داده متناظر با آنها را قرار داد. این کار درک بهتری از خروجی‌ها را به همراه می‌آورد. شکل ۵ برخی از قواعد همبش را در قالب IF-Then نشان می‌دهد که در آنها از عنوان خود اقلام داده استفاده شده است.

IF Biology = Good Then Electronic = Weak, Confidence=0/796
 IF Physics=Top and Electronic= Middle Then Biology= Middle, Confidence= 0/8
 IF Physics= Good and Electronic= Middle Then Physics= Middle and Biology= Middle
 Confidence= 0/9

شکل ۵. برخی از قواعد همبش استخراج‌شده در قالب IF-Then

تجزیه و تحلیل یافته‌ها

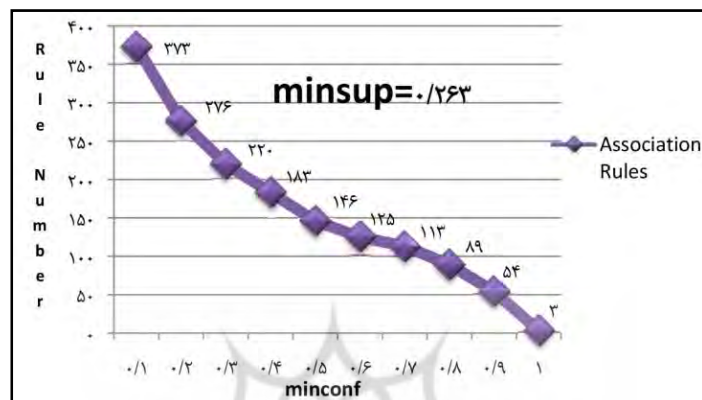
از آنجا که در این پژوهش حداقل آستانه برای حداقل دامنه پشتیبانی از قاعده به صورت خودکار تعریف شد، می‌توان ادعا کرد که تمام الگوهای پرتکرار استخراج خواهند شد. شکل ۶ مقایسه‌ای از تعداد اقلام داده پرتکرار تولیدشده روش پیشنهادی و حداقل آستانه‌های مختلف را نشان می‌دهد.



شکل ۶. تعداد الگوهای پرتکرار تولیدشده با استفاده از روش پیشنهادی و حداقل دامنه پشتیبانی از قاعده‌های مختلف

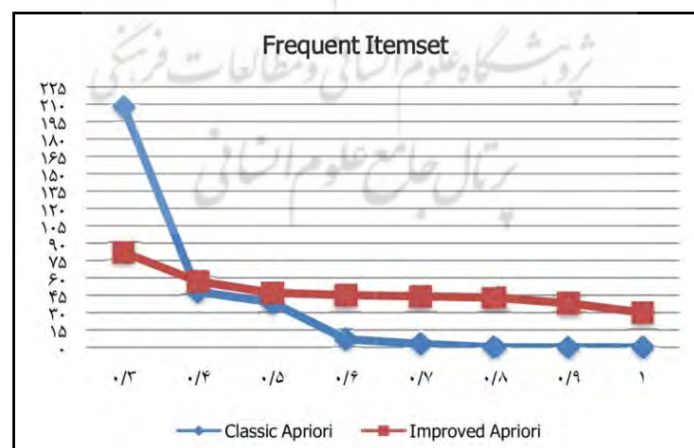
حداقل دامنه پشتیبانی از قاعده به دست آمده برای مطالعه موردی این پژوهش با استفاده از رویکرد پیشنهادی برابر با ۰/۲۶۳ است و بر اساس آن، قواعد همبش با توجه به معیار

اندازه‌گیری میزان اعتماد به قاعده‌شان استخراج شده‌اند. شکل ۷ تعداد قواعد همباش تولید شده با حداقل دامنه پشتیبانی از قاعده = 0.263 و حداقل میزان اعتماد به قاعده‌های مختلف را نشان می‌دهد.



شکل ۷. تعداد قواعد همباشی تولید شده با استفاده از حداقل میزان اعتماد به قاعده‌های مختلف

در این مطالعه، الگوریتم اپریوری کلاسیک با استفاده از نرم‌افزار و داده‌های پژوهش، اعمال شد. مقایسه نتایج نشان داد الگوهای پرتکرار استخراج شده با رویکرد پیشنهادی از انسجام بیشتری برخوردارند و الگوهای پرتکراری که در الگوریتم اپریوری کلاسیک از دست رفته‌اند، در رویکرد پیشنهادی به راحتی کشف خواهند شد.



شکل ۸. مقایسه رویکرد پیشنهادی و الگوریتم اپریوری کلاسیک

نتیجه‌گیری و پیشنهادها

در این پژوهش، روشی کارآمد برای بهبود الگوریتم اپریوری مبنی بر تعیین خودکار حداقل دامنه پشتیبانی از قاعده به‌منظور تولید الگوهای پرتکرار و استخراج قواعد همبش از پایگاه داده‌ها ارائه شد. از ویژگی‌های شایان توجه رویکرد پیشنهادی، انتقال داده‌های قطعی به محیط فازی است؛ چرا که بر اساس دلایل بیان شده در پیشینه نظری، استخراج قواعد همبش از مجموعه داده‌های قطعی، به تولید قواعدی با دقت کم و ضعیف منجر خواهد شد. تعیین حداقل دامنه پشتیبانی از قاعده این پژوهش به کمک روش‌های آماری و روش میانگین‌گیری انجام گرفت و پرتکرار بودن اقلام داده در پایگاه داده‌ها با آن سنجیده شد. اقلام داده‌ای که دامنه پشتیبانی از قاعده بزرگ‌تر یا مساوی حداقل دامنه پشتیبانی از قاعده داشتند، به‌مثابه مجموعه اقلام داده پرتکرار به مراحل بعدی راه یافتند و بقیه آنها حذف شدند.

همان‌طور که گفته شد، مهم‌ترین مزیت رویکرد پیشنهادی این است که کاربران از تعیین حداقل دامنه پشتیبانی از قاعده معاف‌اند. برای تعیین حداقل دامنه پشتیبانی از قاعده، از روش میانگین‌گیری از حوزه دانشی آمار استفاده شده است؛ با این حال سایر روش‌هایی که در این حوزه وجود دارد (مانند واریانس، میانه، انحراف معیار و...) را نیز می‌توان برای انجام این مهم به کار برد و نتایج را با روش این پژوهش مقایسه کرد.

از آنجاکه تمام الگوریتم‌های حوزه استخراج قواعد همبش از کاربران درخواست تعیین حداقل دامنه پشتیبانی از قاعده را دارند، پیشنهاد می‌شود رویکرد پیشنهادی این پژوهش برای الگوریتم‌های دیگر این حوزه به کار رود و نتایج، تجزیه و تحلیل شود.

در نهایت، الگوریتم اپریوری معیار اندازه‌گیری دیگری دارد که حداقل میزان اعتماد به قاعده نامیده می‌شود. کاربران هنگام کار، باید این معیار را نیز تعیین کنند؛ پیشنهاد می‌شود برای تعیین حداقل میزان اعتماد به قاعده خودکار نیز مطالعاتی صورت گیرد، با این کار انتظار می‌رود قواعد جالب‌تری استخراج شود.

References

- Agrawal, R. & Shafer, J.C. (1996). Parallel mining of association rules, *IEEE Transactions on Knowledge and Data Engineering*, 8(6): 962-969.
- Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases [A]. *Proc. of the 20th Int'l Conf on Very Large Data Bases [C]*. Santiago: Morgan Kaufmann, 478-499.
- Akhondzadeh-Noughabi, L. & Albadvi, A. & Aghdasi, M. (2014). Mining customer dynamics in designing customer segmentation using data mining techniques.

- Quarterly Journal of Information technology management*, 6(1): 1-30. (in Persian)
- Azar, A., Sangi, M., Izadkhah, M-M. & Anvari, A. (2015). Synergy management model of the holding by fuzzy approach, the role of information technology in its implementation. *Quarterly Journal of Information technology management*, 7(1): 1-22. (in Persian)
- Azizi, SH., Abadi, V.H. & Balaghi Inanlou, M. (2014). Segmentation of Internet Banking Users Based on Expectations: A Data Mining Approach. *Quarterly Journal of Information technology management*, 6(3): 419-434. (in Persian)
- Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers Norwell, MA, USA.
- Cakir, O. & Aras, M.E. (2012). A recommendation engine by using association rules. *Procedia – social and Behavioral Sciences*, 62(24): 452 – 456.
- Chen, C., Hong, T. & Tseng, V. (2009). An improved approach to find membership functions and multiple minimum supports in fuzzy data mining. *Expert Systems with Applications*, 36(6): 10016–10024.
- Dunham, M.H. (2002). *Data Mining: Introductory and Advanced Topics*. Prentice Hall PTR Upper Saddle River, NJ, USA.
- Gottwald, S. (2006). Universes of Fuzzy Sets and Axiomatizations of Fuzzy Set Theory. *Studia Logica*, 82(2): 211-244.
- Han, J., Cheng, H., Xin, D. & Yan, X. (2007). Frequent pattern mining: current status and future Directions. *Data Mining and Knowledge Discovery*, 15(1): 55-88.
- Hu, Y. & Chen, Y. (2006). Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. *Decision Support Systems*, 42(1): 1 – 24.
- Hu, Y., Wu, F. & Liao, Y. (2013). An efficient tree-based algorithm for mining sequential patterns with multiple minimum supports. *The Journal of Systems and Software*, 86(5): 1224- 1238.
- Huang, T. (2013). Discovery of fuzzy quantitative sequential patterns with multiple minimum supports and adjustable membership functions. *Information Sciences*, 222(10): 126-146.
- Jalilmanesh, A. & Homaionvala, A. (2011). Organizational Knowledge Mapping Based on Library Information System. *IADIS Collaborative Technologies, Rome (Italy)*, 20-26.
- Jang, J., Sun, C. & Mizutani, E. (1997). *Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence*. Paperback.
- Lee, Y., Hong, T. & Wang, T. (2008). Multi-level fuzzy mining with multiple minimum supports. *Expert Systems with Applications*, 34(1): 459–468.

- Lei, Z. & Ren-Hou, L. (2007). An Algorithm for Mining Fuzzy Association Rules Based on Immune Principles. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE*. Boston, MA.
- Liu, B. (2007). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer.
- Pea, J., Qualite, L. & Tille, Y. (2007). Systematic smpeling is a minimum support design. *Computational statistics & Data Analysis*, 51(12): 5591-5602.
- Radfar, R., Nezafati, N. & Yousefi Asli, S. (2014). Classification of Internet banking customers using data mining algorithms. *Quarterly Journal of Information technology management*, 6(1): 71-90. (in Persian)
- Shihab, A.I. & Burger, P. (1998). The Analysis of Cardiac Velocity MR Images Using Fuzzy Clustering. *Proceeding of SPIE Medical Imaging Physiology and Function from Multidimensional Images*, 3337(14): 176-183.
- Tseng, M. & Lin, W. (2007). Efficient mining of generalized association rules with non-uniform minimum support. *Data & Knowledge Engineering*, 62(1): 41-64.
- Lotfizadeh, A. (1965). Fuzzy sets. *Information and Control*, 8 (3): 338-353.

