مدیریت فناوری اطلاعات
دانشکدهٔ مدیریت دانشگاه تهران
دورهٔ ۷، شمارهٔ ۲
تابستان ۱۳۹۴
ص. ۲۸۲- ۲۵۹

تعیین خودکار حداقل دامنهٔ 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 حوزهٔ داده‌کاوی به‌شمار می‌رود. در الگوریتم‌های کلاسیک حوزهٔ قواعد همباش‌کاوی، مانند اپریوری، از حداقل دامنهٔ پشتیبانیِ قاعدهٔ واحد استفاده می‌شود؛ در حالی‌که در رویکردهای جدیدی که تلاش کرده‌اند الگوریتم‌های کلاسیک را بهبود بخشند، مانند ام. اس. اپریوری، از حداقل دامنهٔ پشتیبانیِ قاعدهٔ چندگانه استفاده می‌شود که در هر دو حالت، کاربر موظف است حداقل دامنهٔ پشتیبانی از قاعده را تعیین کند. در نظر بگیرید کاربر قصد اعمال الگوریتم اپریوری را بر پایگاه داده‌ای با میلیون‌ها تراکنش داشته باشد؛ به‌طور قطع کاربر نمی‌تواند دانش لازم را دربارهٔ تمام تراکنش‌های موجود در پایگاه داده داشته باشد، بنابراین نمی‌تواند حد آستانهٔ مناسبی را تعیین کند. در این پژوهش، برای اولین‌بار با استفاده از داده‌های فازی‌سازی‌شده و تکنیک میانگین‌گیری، روشی ارائه شده است که در آن، الگوریتم اپریوری به‌صورت کاملاً خودکار حداقل دامنهٔ پشتیبانی از قاعده را تعیین می‌کند. نتایج شبیه‌سازی‌شده روی نمونه‌ای واقعی نشان داد این رویکرد عملکرد مطلوب‌تری نسبت به الگوریتم اپریوری کلاسیک دارد.

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

۱. کارشناس‌ارشد مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات ایلام، ایلام، ایران
۲. کارشناس‌ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه پیام نور، ایران
۳. کارشناس‌ارشد مهندسی کامپیوتر، دانشگاه آزاد اسلامی ملایر، ملایر، ایران

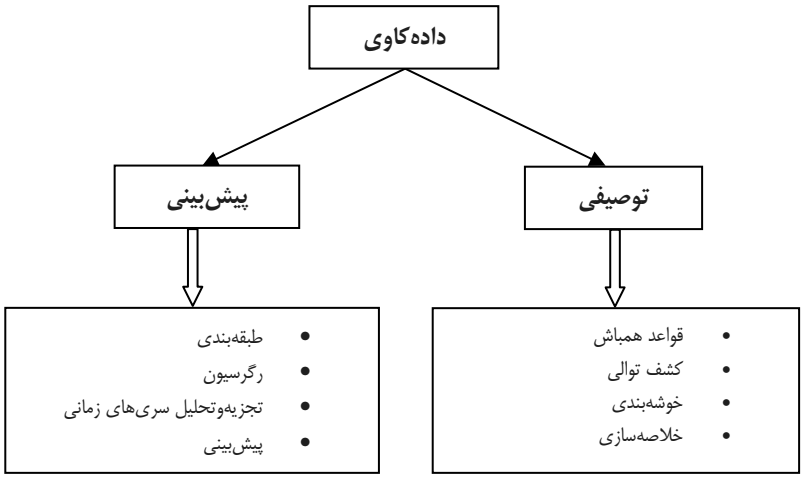
تاریخ دریافت مقاله: ۱۳۹۳/۰۶/۲۸
تاریخ پذیرش نهایی مقاله: ۱۳۹٤/۰۳/۰٦
نویسندهٔ مسئول مقاله: حیدر جعفرزاده
E-mail: heydar.jafarzadeh@gmail.com

مقدمه

داده‌کاوی فرایند منطقی‌ای است که برای جست‌وجوی داده‌ها از حجم زیادی از اطلاعات، به‌کار 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 تجزیه‌وتحلیل 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 به‌کار می‌رود.

- مدل پیش‌بینی؛ به‌منظور پیش‌بینی مقادیر آینده از نتایج به‌دست‌آمده از داده‌های مختلف، به‌کار می‌رود؛
- مدل توصیفی؛ به‌منظور ارائهٔ الگوهای کشف‌شده به شکلی فهم‌پذیر برای انسان به‌کار می‌رود.



شکل ۱. وظایف داده‌کاوی

قواعد همباش[1] یکی از وظایف بسیار مهم در حوزهٔ داده‌کاوی است که می‌تواند در دامنه‌های مختلف استفاده شود. قواعد همباش‌کاوی، روش داده‌کاوی عمومی است و برای استخراج

1. Association Rules

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 مجموعهٔ داده‌ها به جست‌وجو می‌پردازد. در واقع این روش به مطالعهٔ 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[1] و سایر الگوریتم‌هایی که در حوزهٔ استخراج قواعد همباش‌اند، اینکه کاربر مجبور است حداقل آستانه را برای دامنهٔ پشتیبانی از قاعده[2] مشخص کند. در نظر بگیرید 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 دامنهٔ پشتیبانی از قاعدهٔ مناسبی تعیین کند.

هدف از این پژوهش بهبود الگوریتم اپریوری است. برای این کار ابتدا تراکنش‌هایی که در پایگاه داده به‌صورت قطعی ذخیره شده‌اند، به‌کمک 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 به‌کمک یکی از روش‌های حوزهٔ آمار، پس از تعیین، به کاربر پیشنهاد شود. امیدواریم این کار موجب شود هیچ قاعدهٔ جالبی به‌دلیل حد آستانهٔ نامناسبی که کاربر مشخص کرده است، از دست نرود و هیچ قاعدهٔ بی‌استفاده‌ای استخراج نشود.

این مقاله ابتدا پیشینهٔ استخراج قواعد همباش 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.

پیشینهٔ پژوهش

همان‌طور که بسیاری از پایگاه داده‌ها انبوهی از نمونه‌ها و صفت‌ها را دربرگرفته‌اند، باید ابزارهای کارآمدی پیاده‌سازی شود تا بتوان اطلاعات مفید پنهان‌شده در این پایگاه‌ها را استخراج کرد؛ بر این اساس ابزارهای داده‌کاوی بسیاری که از روش‌های تجزیه‌وتحلیل گوناگونی استفاده می‌کنند و به‌طور عمده برگرفته از آمار کلاسیک‌اند، توسعه داده شده است. در این بین، روش‌های قواعد همباش‌کاوی 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.

پیشینهٔ نظری

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

اصول قواعد همباش

همباش یعنی با هم‌بودن دو یا چند قلم داده که مدام با هم تکرار می‌شوند و جلو می‌روند. کشف قواعد همباش یکی از وظایف بسیار مهم در حوزهٔ داده‌کاوی است. قاعدهٔ همباش ساده به این صورت بیان شود: Bread→Cheese [support=۰/۱, confidence=۰/۸]، به‌گفتهٔ ساده‌تر این قاعده بیان می‌کند رابطه‌ای بین خرید نان و پنیر وجود دارد، ماژور دامنهٔ پشتیبانی از قاعده، بیان‌کنندهٔ ثبت نان و پنیر با هم در ۱۰ درصد تراکنش‌ها است و ماژور 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. به‌کمک 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 مسئلهٔ قواعد همباش به این صورت بیان می‌شود: $I=\{i_1,i_2,\ldots,i_m\}$ مجموعه‌ای از اقلام داده است و $T=\{t_1,t_2,\ldots,t_n\}$ مجموعه‌ای از تراکنش‌هاست که هریک شامل اقلام داده‌ای از مجموعهٔ اقلام دادهٔ I است. بنابراین هر تراکنش t_i از مجموعه‌ای از اقلام داده تشکیل شده است؛ به‌گونه‌ای که $t_i \subseteq I$. اگر X و Y اقلام داده فرض شود، قاعدهٔ همباش مفهومی به شکل $X \rightarrow Y$ است که در آن $X \subset I$، $Y \subset I$ و $X \cap Y = \phi$ (لیو، ۲۰۰۷: ۵۷-۶۰).

در قاعدهٔ همباشی 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. دامنهٔ پشتیبانی از قاعده و میزان اعتماد به قاعده، مهم‌ترین معیارهای کیفی برای ارزیابی جالب‌بودن قاعده در نظر گرفته شده‌اند.

- دامنهٔ پشتیبانی از قاعده (Support)

نحوهٔ محاسبهٔ دامنهٔ پشتیبانی از قاعده به‌صورت زیر است.

رابطهٔ ۱) $$\sup(A \Rightarrow B) = \frac{A \cap B}{X}$$

که در آن، A و B دو قلم دادهٔ 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 دادهٔ A و B وجود دارد، استخراج می‌کند و با مقدار حداقل دامنهٔ پشتیبانی از قاعده که توسط کاربر مشخص می‌شود، مقایسه می‌کند و پس از آن فقط تراکنش‌هایی را انتخاب می‌کند که دامنهٔ پشتیبانی از قاعده‌شان بزرگ‌تر یا مساوی حداقل دامنهٔ 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 به‌صورت رابطهٔ ۲ اعمال می‌شود.

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

ابتدا تراکنش‌های حاوی قلم دادهٔ A محاسبه می‌شود و از بین آنها، تراکنش‌های حاوی قلم دادهٔ B استخراج می‌شود، سپس به مقایسهٔ 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).

پیشینهٔ تجربی

الگوریتم اپریوری 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 مجموعهٔ اقلام دادهٔ پرتکرار در پایگاه داده‌های تراکنشی با استفاده از دو معیار حد آستانهٔ از قبل تعریف‌شده به نام‌های دامنهٔ پشتیبانی از قاعده و میزان اعتماد به قاعده، توسعه دادند.

از زمانی که ایدهٔ تولید قواعد همباش مطرح شده است تا کنون، محققان الگوریتم‌های متعددی برای انجام این مهم معرفی کردند که همهٔ آنها در تلاش برای تولید قواعد همباش مفیدتر بودند (هان، چنگ، شین و یان، ۲۰۰۷). الگوریتم‌های قواعد همباش‌کاوی 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 دادهٔ 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 دامنهٔ 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 به‌منظور ذخیرهٔ اطلاعات پیچیده برای الگوهای پرتکرار طراحی کردند و بر اساس این درخت، الگوریتم CFP-Growth را به‌منظور کشف مجموعهٔ اقلام دادهٔ پرتکرار ارائه دادند. سپس با بیان اینکه به‌کارگیری حداقل دامنهٔ پشتیبانی از قاعدهٔ یکسان برای اقلام داده‌ای که هریک فرکانس خاصی دارند، مشکل خواهد بود، راهکاری جدید برای تنظیم دامنهٔ پشتیبانی از قاعدهٔ 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 به‌کمک حداقل دامنهٔ پشتیبانی از قاعدهٔ چندگانه معرفی کردند. در این الگوریتم از یک بستار بالا به پایین برای دستیابی به اقلام دادهٔ پرتکرار بهره بردند. پیاده‌سازی این الگوریتم در مطالعهٔ موردی، به استخراج قواعد همباش در سطوح مختلف تحت دامنهٔ پشتیبانی از قاعدهٔ چندگانه انجامید (لی، هونگ و وانگ، ۲۰۰۸).

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

در مطالعهٔ هوانگ (۲۰۱۳) تلاش شد الگوریتم FQSP، بهبود داده شود. محقق معتقد است اولاً از آنجاکه FQSP از حداقل دامنهٔ پشتیبانی از قاعدهٔ واحد برای کشف الگوهای پرتکرار استفاده می‌کند؛ اگر حداقل دامنهٔ پشتیبانی از قاعده، عددی بزرگ در نظر گرفته شود، اقلام دادهٔ غیرجالب بسیاری تولید خواهد شد و اگر عددی پایین در نظر گرفته شود، به از دست‌رفتن اقلام داده‌ای می‌انجامد که می‌تواند جالب باشد. از این رو در رویکرد جدید، روش حداقل دامنهٔ پشتیبانی از قاعدهٔ چندگانه را به‌کار 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، ایدهٔ تابع عضویت قابل تنظیم را ارائه کرد.

در پژوهشی تلاش شده است الگوهای جزئی دوره‌ای با استفاده از حداقل دامنهٔ پشتیبانی از قاعدهٔ چندگانه استخراج شود. در این پژوهش ابتدا به مشکل کارهای قبلی برای استخراج الگوهای جزئی دوره‌ای مبنی بر مشکل ناکارآمدیِ ناشی از به‌کارگیری حداقل دامنهٔ پشتیبانی از

قاعدهٔ 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 دامنهٔ 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 دامنهٔ پشتیبانی از قاعده و توابع عضویت برای اقلام داده در پایگاه دادهٔ کمی، بهبود بخشند. آنها الگوریتم بهبودیافته را که بر اساس تقسیم و حل کار می‌کند، DGFMMS 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 حوزهٔ استخراج قواعد همباش خارج از حوصلهٔ این پژوهش است، به همین مقدار بسنده می‌شود. با این حال در مقالهٔ هان، چنگ، شین و یان (۲۰۰۷)، نگاه ویژه و مفصلی به رویکردهای ارائه‌شده در حوزهٔ استخراج قواعد همباش و کشف الگوهای پرتکرار تا سال ۱۳۸۶ شده است.

نکتهٔ حائز اهمیتی که انگیزهٔ ارائهٔ مقالهٔ حاضر بوده است، اینکه در هر دو حالت به‌کارگیری حداقل دامنهٔ پشتیبانی از قاعدهٔ واحد و چندگانه، بدون استثنا از کاربر درخواست می‌شود حداقل دامنهٔ 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 دربارهٔ تمام تراکنش‌های موجود در پایگاه داده داشته باشد و بنابراین نمی‌تواند حداقل دامنهٔ پشتیبانی از قاعدهٔ 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 نظریهٔ مجموعه‌های فازی را مطرح کرد. این نظریه چارچوب ریاضی محکمی برای مطالعهٔ 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 همباش فازی

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

- دامنهٔ پشتیبانی از قاعدهٔ فازی

در اولین گام برای تعیین تعداد تکرار هر قلم داده در پایگاه دادهٔ فازی‌سازی‌شده، مجموع درجهٔ عضویت‌ها برای هر داده محاسبه می‌شود که این رقم میزان تکرار هر داده در نظر گرفته خواهد شد. رابطهٔ ۳ گویای مطلب فوق است.

$$\text{fuzzysum} = \sum_{(x)\in D} \mu(x) \qquad \text{رابطهٔ ۳)}$$

سپس برای تولید مجموعهٔ اقلام دو ـ داده‌ای، باید پس از مقایسهٔ تمام درجهٔ عضویت‌های هر دو قلم داده با هم، حداقل[1] را از بین آنها انتخاب کرد. در پایان، مجموع حداقل‌ها میزان تکرار دو قلم دادهٔ مد نظر معرفی می‌شود. برای تولید مجموعهٔ اقلام سه ـ داده‌ای باید پس از مقایسهٔ تمام درجهٔ عضویت‌های هر سه داده با هم، از بین آنها حداقل را انتخاب کرد که مجموع حداقل‌ها، میزان تکرار سه قلم دادهٔ مد نظر معرفی می‌شود. این کار تا زمانی ادامه پیدا خواهد کرد که دادهٔ 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. رابطهٔ ۴ گویای مطلب فوق است.

$$\text{fuzzy sup}(A \to B) = \sum_{(x)\in D} T[A(x), B(y)] \qquad \text{رابطهٔ ۴)}$$

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

- میزان اعتماد به قاعدهٔ فازی

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

1. Minimum

همباش فازی معرفی خواهند شد. در این مطالعه از عملگر t-norm برای محاسبهٔ میـزان اعتمـاد به قاعدهٔ فازی استفاده شده است. رابطهٔ ۵ مطلب فوق را به‌صورت رسمی بیان می‌کند.

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

نحوهٔ محاسبهٔ حداقل دامنهٔ پشتیبانی از قاعده با روش پیشنهادی

این پژوهش تلاش می‌کند الگوریتم، به‌صورت خودکـار حـداقل آسـتانهٔ مناسـبی را بـرای معیـار اندازه‌گیری حداقل دامنهٔ پشتیبانی از قاعده به کاربر معرفی کند. هنگامی‌که متخصصان بخواهنـد مجموعه‌ای از داده‌ها را با شاخصی مقایسه کنند و چنین شاخصی وجود نداشته باشد کـه داده‌هـا نسبت به آن سنجیده شود و داده‌هایی کـه از ایـن شـاخص بـزرگ‌تـر (بهتـر، قـوی‌تـر و ...) یـا کوچک‌ترند (بدتر، ضعیف‌تر و...) را دسته‌بندی کند، از روش‌های آماری کمک می‌گیرند و با توجه به معادله‌هایی مانند میانگین‌گیری، انحراف معیار، میانه، مد، واریانس و... شاخصـی را بـه‌منظـور مقایسهٔ داده‌ها تعریف می‌کنند و با این کار اطلاعات را به‌دست می‌آورند.

از آنجاکه هیچ شاخص مناسبی وجود ندارد که بتوان حداقل آسـتانهٔ مناسـبی را بـرای معیـار اندازه‌گیری دامنهٔ پشتیبانی از قاعده با توجه به آن تعریف کرد، در این مطالعـه نیـز از روش‌هـای آماری استفاده شده است. از بین روش‌هایی که در این حـوزه وجـود دارد، روش میـانگین‌گیـری انتخاب شده است تا حداقل آستانهٔ مناسب برای استخراج الگوهای پرتکرار تعریف شود؛ به‌طوری که این حداقل آستانه، نه خیلی پایین باشد که به تولید حجم زیادی از الگوهای به‌دردنخور منجر شود و نه خیلی بالا در نظر گرفته شود که به از دست‌دادن الگوهای جالب بینجامد. بـه‌کـارگیری این ایده می‌تواند سبب بهبود روند کشف الگوهای پرتکرار و استخراج قواعد همباش شود. پس از انتقال داده‌ها به محیط فازی توسط الگوریتم خوشه‌بندی FCM، از طریق رابطهٔ ۶ مجموع تمـام درجه‌های عضویت برای هر فیلد محاسبه می‌شود و این رقم بر تعداد کل اقلام داده (T) تقسـیم خواهد شد. عدد به‌دست‌آمده حداقل دامنهٔ پشتیبانی از قاعده معرفی می‌شود.

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

در واقع، در این پژوهش دخالت کاربر در روند اجرای الگوریتم اپریوری حـذف شـده اسـت و ادعا می‌شود که با ارائهٔ این روش، الگوریتم اپریوری به الگوریتمی خودکار تبدیل شده است.

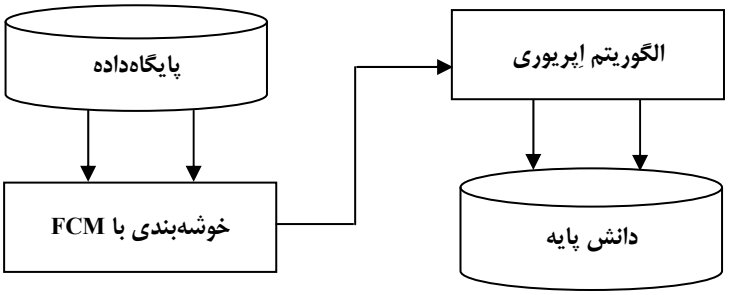
روش‌شناسی پژوهش

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 نحوهٔ تعیین حداقل دامنهٔ پشتیبانی از قاعدهٔ خودکار که از مجموعهٔ 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 مجموعهٔ دادهٔ مد نظر از پایگاه داده که به‌منظور هماهنگی و حذف داده‌های پرت بر اساس قوانین پیش‌پردازش، تجزیه‌وتحلیل شده است؛
۲. انتقال مجموعهٔ داده‌های قطعی به محیط فازی با استفاده از الگوریتم FCM؛
۳. استفاده از الگوریتم اپریوری با تکیه بر روش پیشنهادی برای تعیین حداقل دامنهٔ پشتیبانی از قاعده به‌منظور کشف الگوهای پرتکرار و استخراج قواعد همباش و درنهایت، ذخیرهٔ قواعد با نام KB.



شکل ۲. مدل رویکرد پیشنهادی

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

- فیلد: عبارت است از صفت، قلم داده یا ستون جدول؛
- رکورد: عبارت است از سطر با همهٔ فیلدها؛
- $\mu(X)$: مقدار مجموعهٔ فازی که عددی خواهد بود در بازهٔ [۰,۱]؛
- C_k: معرف مجموعهٔ اقلام دادهٔ کاندید است که $1 \leq k \leq n$؛
- L_k: معرف مجموعهٔ اقلام دادهٔ پرتکرار است که $1 \leq k \leq n$؛
- Fuzzy minsup: معیار اندازه‌گیری حداقل دامنهٔ پشتیبانی از قاعدهٔ فازی؛
- Fuzzy minconf: معیار اندازه‌گیری حداقل میزان اعتماد به قاعدهٔ فازی.

مراحل ۱ تا ۱۲ شبه کد رویکرد پیشنهادی ارائه‌شده در شکل ۳ را با جزئیات توضیح می‌دهند:

۱. FCM پس از خوشه‌بندی داده‌های قطعی، مرکز هر مجموعهٔ فازی را تعیین می‌کند و مقادیر حداکثر و حداقل را برای هر فیلد از مجموعهٔ دادهٔ ورودی پیدا می‌کند؛

۲. با استفاده از trimf اقلام داده میان خوشه‌ها توزیع می‌شود؛

۳. حداقل دامنهٔ پشتیبانی از قاعده با استفاده از رابطهٔ ۶ به‌صورت خودکار محاسبه می‌شود؛

۴. دامنهٔ پشتیبانی از قاعدهٔ هر مجموعهٔ قلم داده برای هر رکورد با استفاده از رابطهٔ ۳ به‌دست می‌آید، سپس در فهرست مجموعهٔ اقلام دادهٔ کاندید C_1 مرتب و ذخیره می‌شود؛

۵. از بین مجموعهٔ اقلام دادهٔ کاندید C_1، آنهایی‌که بتوانند شرط حداقل دامنهٔ پشتیبانی از قاعده را برآورده کنند، مجموعهٔ اقلام دادهٔ پرتکرارند که در L_1 مرتب و ذخیره می‌شوند؛

۶. اقلام دادهٔ موجود در L_1 به‌صورت زیر با هم الحاق می‌شوند:

$$(L_1 \text{ Join } L_1) = \{ \{C[1], C[i], \{C[1], C[i+1]\} \ldots . \{C[1], C[n]\} \}$$

که C [۱] اولین مجموعهٔ فازی، C[i] دومین مجموعهٔ فازی و C[n] آخرین مجموعهٔ فازی است و ($C[1] \cap C[i]=\emptyset$, $C[1] \cap C[i+1]=\emptyset$, … , $C[1] \cap C[n]=\emptyset$). نتیجه در فهرست مجموعهٔ اقلام دادهٔ کاندید C_2 ذخیره می‌شود. دامنهٔ پشتیبانی از قاعده برای هر مجموعهٔ قلم دادهٔ کاندید با استفاده از رابطهٔ ۴ محاسبه می‌شود؛

۷. از بین مجموعهٔ اقلام دادهٔ کاندید C_2، آنهایی‌که بتوانند شرط حداقل دامنهٔ پشتیبانی از قاعده را برآورده کنند، مجموعهٔ اقلام دادهٔ پرتکرارند که در L_2 مرتب و ذخیره می‌شوند؛

۸. اقلام داده‌هایی که در L_2 هستند با هم الحاق می‌شوند. باید تمام زیرمجموعه‌های C_2 به‌مثابهٔ مجموعهٔ اقلام دادهٔ پرتکرار در L_1 حضور داشته باشند، یعنی تمام زیرمجموعه‌های C_K به‌مثابهٔ مجموعهٔ اقلام دادهٔ پرتکرار در L_{K-1} حضور داشته باشند؛

۹. نتیجهٔ عملیات الحاق مرحلهٔ قبل در فهرست مجموعهٔ اقلام دادهٔ کاندید C_3 ذخیره می‌شود. دامنهٔ پشتیبانی از قاعده برای هر مجموعهٔ قلم دادهٔ کاندید محاسبه می‌شود؛

۱۰. اقلام دادهٔ کاندید C_3 که شرط حداقل دامنهٔ پشتیبانی از قاعده را برآورده کنند به L_3 انتقال داده می‌شوند؛

۱۱. این فرایند تا زمانی که L_K خالی شود، تکرار می‌شود؛

۱۲. الگوهای پرتکرار کشف‌شده در قالب (IF-Then) مرتب می‌شوند و میـزان اعتمـاد بـه قاعده با استفاده از رابطهٔ ۵ به‌دست می‌آید و نتیجه در KB ذخیره می‌شود.

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* ≥ *fuzzy minsup* Then
Insert the fuzzy set into L_1, $L_1 =$ {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.$item_2$ … p.$item_{k-1}$, q.$item_{k-1}$ From p, q
Where p.$item_1$ = q.$item_1$, p.$item_2$ = q.$item_2$ … 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, $L_k =$ {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 تجزیه‌وتحلیل و اعتبارسنجی اهداف، رویکردی که در بخش قبل مطرح شد، بر مجموعهٔ دادهٔ انتخاب‌شده از یک کتابخانهٔ 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 همباش فازی با استفاده از رویکرد پیشنهادی و با کمک‌گرفتن از داده‌های یکی از کتابخانه‌های الکترونیک توصیف می‌شود. در این مطالعه اطلاعات با توجه به اطلاعات مقالهٔ جلیل‌منش و همکارش با عنوان «نگاشت دانش سازمانی بر اساس سیستم اطلاعاتی کتابخانه» به‌دست آمده است (جلیل‌منش و همایون والا، ۱۳۹۰).

در کل هدف از ارائهٔ این مثال، یافتن ارتباطات جالب بین کتاب‌های یکی از کتابخانه‌های الکترونیک با استفاده از رویکرد پیشنهادی بوده است. با تجزیه‌وتحلیل کتاب‌هایی که محتمل است با هم به امانت برده شوند، اطلاعات مفیدی برای محققان و کاربران به‌دست 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 دادهٔ کتابخانهٔ الکترونیک

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

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

جدول ۲. نمونه‌ای از مجموعهٔ دادهٔ قطعی کتابخانهٔ 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	آگوست	۸	۰	۶	۲۱	۰	...
۹	سپتامبر	۴	۸	۴	۱	۱۸	...
۱۰	اکتبر	۷	۷	۰	۱۶	۱۴	...
...	...	...	...	...	...	...	...

جدول ۳. اطلاعات آماری از مجموعهٔ دادهٔ قطعی کتابخانهٔ الکترونیک

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

حال مجموعهٔ داده‌های قطعی به محیط فازی منتقل می‌شود که برای این کار از الگوریتم خوشه‌بندی FCM استفاده شده است. در این مطالعهٔ موردی، برای هر فیلد چهار خوشه با عنوان‌های «ضعیف»[1]، «متوسط»[2]، «خوب»[3] و «عالی»[4] تعریف شده است و تمام اقلام داده در این چهار خوشه توزیع شده‌اند. برای خلاصه‌کردن نام خوشه‌ها از کلمهٔ اول نام هر خوشه استفاده شده است که به‌ترتیب عبارت‌اند از: (T، G، M، W). جدول ۴، هر فیلد را با خوشه‌های متناظر و تابع عضویتش نشان می‌دهد.

جدول ٤. فیلدها و مجموعه‌های فازی متناظر با آن‌ها

تابع عضویت	مجموعهٔ فازی	حوزه دانشی
Trimf	ضعیف (W۱)، متوسط (M۲)، خوب (G۳)، عالی (T۴)	فیزیک
Trimf	ضعیف (W۵)، متوسط (M۶)، خوب (G۷)، عالی (T۸)	زیست‌شناسی
Trimf	ضعیف (W۹)، متوسط (M۱۰)، خوب (G۱۱)، عالی (T۱۲)	الکترونیک
Trimf	ضعیف (W۱۳)، متوسط (M۱۴)، خوب (G۱۵)، عالی (T۱۶)	کامپیوتر

درجهٔ 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 درجهٔ عضویت به این بستگی دارد که در مرز بالای مجموعه قرار دارد یا مرز پایین. به‌طور مثال برای بیان درجهٔ عضویت فازی فیلد فیزیک (فرض کنید p) از رابطه‌های ۷ تا ۱۰ استفاده شده است.

رابطهٔ ۷)
$$\mu_{\text{Physics.Middle}}(p)=\begin{cases}۱ & p\le ۰/۷\\ \dfrac{۳/۶-p}{۲/۹} & ۰/۷\le p\le ۳/۶\end{cases}$$

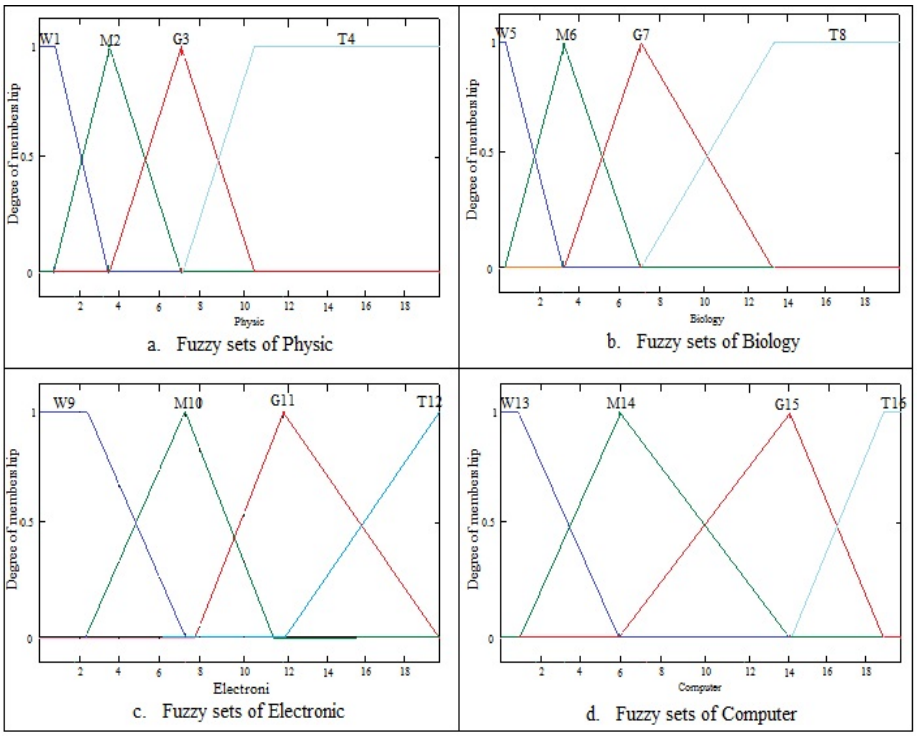
رابطهٔ ۸)
$$\mu_{\text{Physics.Middle}}(p)=\begin{cases}\dfrac{p-۰/۷}{۲/۹} & ۰/۷\le p\le ۳/۶\\ \dfrac{۷/۱-p}{۳/۵} & ۳/۶\le p\le ۷/۱\end{cases}$$

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

$$\mu_{\text{Physics .Good}}(p)=\begin{cases}\dfrac{p-۳/۶}{۳/۵} & ۳/۶\le p\le ۷/۱\\ \dfrac{۱۰/۵-p}{۳/۴} & ۷/۱\le p\le ۱۰/۵\end{cases}$$ رابطهٔ ۹)

$$\mu_{\text{Physics .Top}}(p)=\begin{cases}\dfrac{p-۷/۱}{۳/۴} & ۷/۱\le p\le ۱۰/۵\\ ۱ & p\ge ۱۰/۵\end{cases}$$ رابطهٔ 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 ٤. خوشه‌بندی فازی کتابخانهٔ 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 (کتابخانهٔ الکترونیک)

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

از این مرحله به بعد است که از الگوریتم اپریوری برای استخراج الگوهای پرتکرار و تولید قواعد همباش استفاده می‌شود. با توجه به رابطهٔ ۶، معیار اندازه‌گیری حداقل دامنهٔ پشتیبانی از قاعده در این مطالعهٔ موردی برابر با Fuzzyminsup = ۰/۲۶۳ به‌دست آمده است. جدول‌های ۶ تا ۱۲ کشف اقلام دادهٔ پرتکرار را با استفاده از رویکرد پیشنهادی به تصویر می‌کشند.

جدول ٦. قسمتی از C_1

دامنهٔ پشتیبانی از قاعده	قلم داده
۳/۶۵۲	W{۱}
۲/۹۶۵	M{۲}
۲/۲۱۸	G{۳}
۱/۱۶۶	T{۴}
۲/۲۸۱	W{۵}
...	...

جدول ۷. قسمتی از L_1

مجموعه داده
W{۱}
M{۲}
G{۳}
T{۴}
W{۵}
...

جدول ۸. نحوهٔ محاسبه دامنهٔ پشتیبانی از قاعده برای دو قلم داده

W{۱}	T{۸}	دامنهٔ پشتیبانی از قاعده
۰/۹۸۵	۰/۸۵۶	۰/۸۵۶
۰/۰۶۶	۰/۰۰۷	۰/۰۰۷
۰/۰۰۳	۰/۰۰۷	۰/۰۰۳
۰/۵۷۴	۰	۰
۰/۰۶۱	۰/۰۰۴	۰/۰۰۴
...	...	...
...	...	∑ = ۱ / ۸۵۴

جدول ۹. قسمتی از C_2

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

جدول ۱۰. قسمتی از L_2

مجموعه داده
W{۱} , M{۲}
W{۱} , M{۶}
W{۱} , G{۷}
W{۱} , T{۸}
W{۱} , W{۹}
W{۱} , M{10}
W{۱} , G{۱۱}
M(۲) , G(۳)
...

جدول ۱۱. قسمتی از C_3

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

جدول ۱۲. قسمتی از L_3

مجموعه داده
W{۱} , M{۲} , W{۹}
M{۲} , G{۳} , W{۵}
M{۲} , G{۳} , M{۶}
M{۲} , G{۳} , M{۱۰}
M{۲} , W{۵} , M{۶}
M{۲} , W{۵} , M{۱۰}
M{۲} , W{۵} , G{۱۱}
W{۱} , M{۶} , M{۱۰}
...

تا این مرحله الگوهای پرتکرار با دامنهٔ پشتیبانی از قاعده‌شان کشف شده‌اند. از این مرحله بـه بعد استخراج قواعد همباش آغاز خواهد شد. ابتدا تمام مجموعهٔ اقلام دادهٔ پرتکرار موجـود در L_2 تا L_n در قالب‌های IF-Then قرار می‌گیرند. سپس برای هر یک میزان اعتماد به قاعده براساس رابطهٔ ۵ محاسبه می‌شود. جدول ۱۳ محاسبهٔ این معیار را نشان می‌دهد.

جدول ۱۳. محاسبهٔ 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 همباش تولیدشده با استفاده از روش پیشنهادی را نشـان می‌دهد که در آن Fuzzyminsup =۰/۲۶۳ و Fuzzyminconf =۰/۷، در نظر گرفته شده‌اند.

جدول ١٤. برخی از قواعد همباش تولید شده با استفاده از رویکرد پیشنهادی

قواعد همباش	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	۰/۳۹۲	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

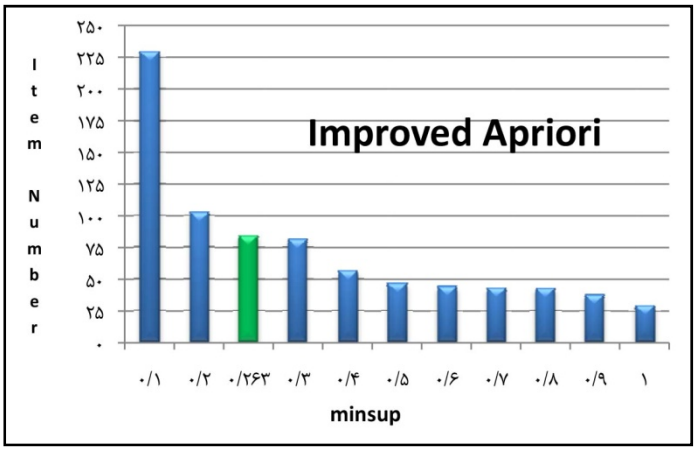
حال می‌توان به جای اعداد به‌کاررفته در این پژوهش، اقلام دادهٔ متناظر بـا آنهـا را قـرار داد. این کار درک بهتری از خروجی‌ها را به همراه می‌آورد. شـکل ۵ برخـی از قواعـد همبـاش را در قالب IF-Then نشان می‌دهد که در آنها از عنوان خود اقلام داده استفاده شده است.

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 همباش استخراج‌شده در قالب IF-Then

تجزیه‌وتحلیل یافته‌ها

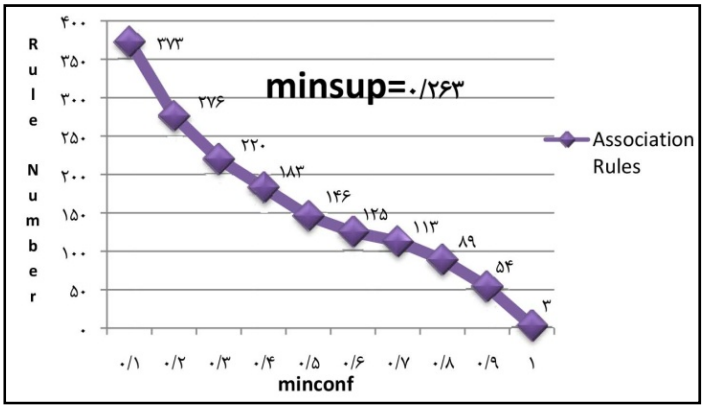
از آنجا که در این پژوهش حداقل آستانه برای حداقل دامنهٔ پشتیبانی از قاعده به‌صورت خودکـار تعریف شد، می‌توان ادعا کرد که تمام الگوهای پرتکرار استخراج خواهند شد. شکل ۶ مقایسـه‌ای از تعداد اقلام دادهٔ پرتکرار تولیدشدهٔ روش پیشـنهادی و حـداقل آسـتانه‌هـای مختلـف را نشـان می‌دهد.



شکل ٦. تعداد الگوهای پرتکرار تولیدشده با استفاده از روش پیشنهادی و حداقل دامنهٔ پشتیبانی از قاعده‌های مختلف

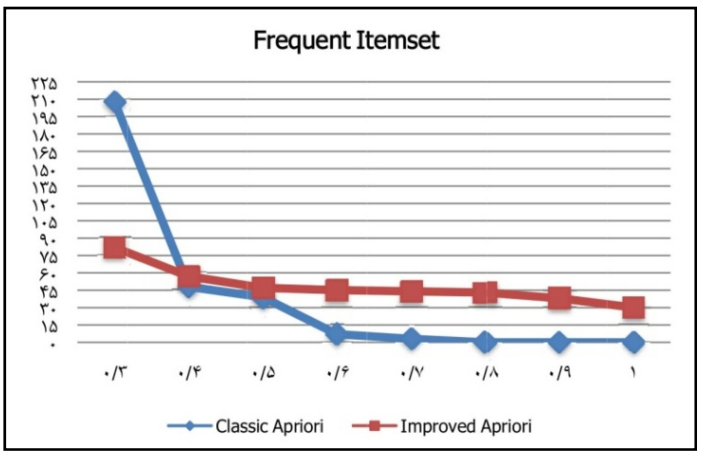
حداقل دامنهٔ پشتیبانی از قاعده به‌دست‌آمده برای مطالعهٔ موردی این پـژوهش بـا اسـتفاده از رویکرد پیشنهادی برابر بـا ۰/۲۶۳ اسـت و بـر اسـاس آن، قواعـد همبـاش بـا توجـه بـه معیـار

اندازه‌گیری میزان اعتماد به قاعده‌شان استخراج شده‌اند. شکل ۷ تعداد قواعد همباش تولید شده با حداقل دامنهٔ پشتیبانی از قاعده = ۰/۲۶۳ و حداقل میزان اعتماد به قاعده‌های مختلف را نشان می‌دهد.



شکل ۷. تعداد قواعد همباش 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. مقایسهٔ نتایج نشان داد الگوهای پرتکرار استخراج‌شده 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. مقایسهٔ رویکرد پیشنهادی و الگوریتم اپریوری کلاسیک

نتیجه‌گیری و پیشنهادها

در این پژوهش، روشی کارآمد برای بهبود الگوریتم اپریوری مبنی بر تعیین خودکار حداقل دامنهٔ پشتیبانی از قاعده به‌منظور تولید الگوهای پرتکرار و استخراج قواعد همباش از پایگاه داده‌ها ارائه شد. از ویژگی‌های شایان توجه رویکرد پیشنهادی، انتقال داده‌های قطعی به محیط فازی است؛ چرا که بر اساس دلایل بیان‌شده در پیشنیهٔ نظری، استخراج قواعد همباش از مجموعهٔ داده‌های قطعی، به تولید قواعدی با دقت کم و ضعیف منجر خواهد شد. تعیین حداقل دامنهٔ پشتیبانی از قاعده این پژوهش به‌کمک روش‌های آماری و روش میانگین‌گیری انجام گرفت و پرتکراربودن اقلام داده در پایگاه داده‌ها با آن سنجیده شد. اقلام داده‌ای که دامنهٔ پشتیبانی از قاعدهٔ بزرگ‌تر یا مساوی حداقل دامنهٔ پشتیبانی از قاعده داشتند، به‌مثابهٔ مجموعهٔ اقلام دادهٔ پرتکرار به مراحل بعدی راه یافتند و بقیهٔ آنها حذف شدند.

همان‌طور که گفته شد، مهم‌ترین مزیت رویکرد پیشنهادی این است که کاربران از تعیین حداقل دامنهٔ پشتیبانی از قاعده معاف‌اند. برای تعیین حداقل دامنهٔ پشتیبانی از قاعده ، از روش میانگین‌گیری از حوزهٔ 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 به‌کار برد و نتایج را با روش این پژوهش مقایسه کرد.

از آنجاکه تمام الگوریتم‌های حوزهٔ استخراج قواعد همباش از کاربران درخواست تعیین حداقل دامنهٔ پشتیبانی از قاعده را دارند، پیشنهاد می‌شود رویکرد پیشنهادی این پژوهش برای الگوریتم‌های دیگر این حوزه به‌کار رود و نتایج، تجزیه‌وتحلیل 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. & Homaiounvala, 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.