پردیس بین‌المللی ارس

گروه علوم کامپیوتر

پایان‌نامه
برای دریافت درجه کارشناسی ارشد در رشته علوم کامپیوتر

عنوان

کشف قوانین انجمنی به صورت موازی بر اساس یک روش خوشه‌بندی جدید و با استفاده از یک الگوریتم تکاملی

استاد راهنما

دکتر شهریار لطفی

استاد مشاور

دکتر جابر کریم‌پور ینگجه

پژوهشگر

لیلا امیری

تابستان ۱۳۹۳

با تشکر و قدردانی از

استاد بزرگوار جناب آقای دکتر شهریار لطفی که با زحمات بی‌دریغ و دلسوزانه خویش راهنمایی پایان‌نامه را بر عهده داشتند و جناب آقای دکتر جابر کریم‌پور ینگجه که مشاوره‌ی این پایان‌نامه را تقبل فرمودند. هم‌چنین پدر و مادر دلسوزم که مرا در پیشبرد پایان نامه یاری نمودند، تشکر و قدردانی می‌نمایم.

لیلا امیری

شهریور ۱۳۹۳

نام خانوادگی: امیری **نام:** لیلا

عنوان پایان‌نامه: کشف قوانین انجمنی به صورت موازی بر اساس یک روش خوشه‌بندی جدید و با استفاده از یک الگوریتم تکاملی

استاد راهنما: دکتر شهریار لطفی

استاد مشاور: دکتر جابر کریم‌پور ینگجه

مقطع تحصیلی: کارشناسی ارشد **رشته:** علوم کامپیوتر **گرایش:** سیستم‌های کامپیوتری

دانشگاه: تبریز **دانشکده:** پردیس بین المللی ارس **تاریخ فارغ التحصیلی:** ۱۳۹۳**تعداد صفحات:** ۱۳۷

کلید واژه‌ها: کشف قوانین انجمنی، خوشه‌بندی، الگوریتم تکاملی و الگوریتم رقابت استعماری

چکیده:

کشف قوانین انجمنی، یکی از روش‌های مهم و پرکاربرد داده‌کاوی جهت کشف دانش نهفته در داده‌ها است که با استفاده از آن می‌توان روابط و وابستگی‌های مفیدی که در مجموعه‌های بزرگی از اقلام داده موجود می‌باشند را کشف نمود. این روابط و وابستگی‌ها در تصمیم‌گیری‌ها، نقش مهمی دارند. یافتن چنین روابطی، داخل مجموعه‌ی داده‌ها به دلیل ماهیت نمایی آن و حجم بالای داده‌ها بسیار زمان‌بر است. در این پایان‌نامه برای کشف قوانین انجمنی، ابتدا با ارائه‌ی یک روش خوشه‌بندی مبتنی بر الگوریتم رقابت استعماری، تراکنش‌ها به خوشه‌های مناسب تقسیم می‌شوند. سپس یک روش تکاملی بر پایه الگوریتم رقابت استعماری برای کشف قوانین انجمنی ارائه می‌گردد که این الگوریتم به طور جداگانه و مستقل بر روی هر یک از خوشه ها اجرا می‌شود. در نهایت، قوانین به دست آمده از همه خوشه‌ها در یک‌جا جمع‌آوری شده و قوانین نهایی تولید می‌گردند. نتایج حاصل از آزمایشات مختلف بر روی چند مجموعه داده شناخته شده، کارآیی راه‌کار پیشنهادی را در کشف قوانین با دقت مناسب و کاهش هزینه‌ها تایید می‌کند.

فهرست مطالب

عنوان **صفحه**

عنوان **صفحه**

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

عنوان **صفحه**

عنوان **صفحه**

فصل ۱

مقدمه

در سال‌های گذشته، پیشرفت‌های قابل توجهی در روش‌های جمع‌آوری، ذخیره‌سازی و انتقال حجم عظیمی از داده‌ها در زمینه‌های مختلف، صورت گرفته است. با توجه به افزایش چشم‌گیر حجم داده‌ها، نیاز به روش‌های بهتر، سریع‌تر و ارزان‌تر جهت استخراج اطلاعات و دانش نهفته‌ی موجود در این حجم از داده‌ها می‌باشد، و گرنه داده‌های موجود، فاقد ارزش خواهند بود. داده‌کاوی بخشی از فرآیند کشف دانش از داده‌ها می‌باشد. داده‌کاوی به معنای یافتن نیمه‌خودکار الگوهای پنهان موجود در مجموعه‌ی داده‌های موجود می‌باشد [۱]. از طرف دیگر صاحبان داده، درک کمی از داده‌ها دارند و دانش آن‌ها نسبت به این داده‌ها کم است. یکی از مسائل اساسی در داده‌کاوی این است که با وجود حجم بالای داده‌ها، یک مدل کوچک و نه زیاد پیچیده از داده‌ها ایجاد شود که در عین حال که داده‌ها را به خوبی توصیف نماید، هم‌چنین ساده و قابل فهم باشد. داده‌کاوی در حوزه‌های تصمیم‌گیری، پیش‌بینی و تخمین مورد استفاده قرار می‌گیرد. کشف قوانین انجمنی یکی از روش‌های هدایت نشده و بسیار مهم در داده کاوی می‌باشد. با توجه به افزایش روز افزون حجم داده‌ها و مقیاس‌پذیری الگوریتم‌های کشف قوانین انجمنی، این روش یکی از پرکاربردترین شیوه‌های داده‌کاوی است. قوانین انجمنی در زمینه‌های مختلفی از جمله تجارت (مانند بازاریابی و تعیین راهبرد برای قیمت‌گذاری)، پیش‌پردازش داده‌ها، خوشه‌بندی، سیستم‌های شخصی‌سازی و پیشنهاد برای جستجوی صفحات وب (مانند پیشنهادهایی که سایت آمازون برای خرید کتاب‌های مرتبط ارائه می‌دهد) کاربرد دارد [۲].

هدف اصلی از کشف قوانین انجمنی، یافتن روابط بین داده‌ها با استفاده از تحلیل داده‌هایی است که در بیشتر موارد به طور هدایت نشده جمع‌آوری شده‌اند. این روابط و وابستگی‌ها در تصمیم‌گیری‌ها نقش مهمی دارند. در روش‌های سنتی، کشف قوانین انجمنی از دو مرحله اصلی

تشکیل شده است. ابتدا تمامی مجموعه‌های اقلام پرتکرار، تولید شده و در گام بعدی قوانین انجمنی قوی که محدودیت‌های معینی را ارضا می‌کنند، بر اساس این مجموعه‌ها کشف می‌شوند. مرحله دوم دارای هزینه‌ی محاسباتی کمی می‌باشد اما مرحله اول دارای هزینه‌ی محاسباتی نمایی است. از طرفی در تولید قوانین انجمنی، داده‌ها ممکن است چندین بار پیمایش شوند. حجم داده‌ها در مسائل مربوط به داده‌کاوی بالا می‌باشد و در بیشتر موارد تمامی داده‌ها، قابل بارگذاری در حافظه‌ی اصلی نمی‌باشند بنابراین سربار ناشی از عملیات ورودی و خروجی در این روش‌ها بسیار بالا است [۳] [۴].

روش‌های مختلفی برای کاهش هزینه‌های مذکور و بهبود فرآیند کشف قوانین انجمنی ارائه شده است. در سال‌های گذشته، استفاده از الگوریتم‌های تکاملی در کشف قوانین انجمنی به دلیل تولید قوانین، بدون نیاز به تولید مجموعه‌های اقلام پرتکرار، بسیار مورد توجه قرار گرفته‌اند. هم‌چنین استفاده از روش‌های مبتنی بر خوشه‌بندی تراکنش‌ها، به دلیل کاهش تعداد پیمایش داده‌ها و امکان اجرای موازی الگوریتم‌های کشف قوانین انجمنی در هر خوشه، در کاهش هزینه‌ها و افزایش کارآیی الگوریتم‌های کشف قوانین انجمنی بسیار موثر می‌باشند.

در این پایان‌نامه برای کشف قوانین انجمنی، ابتدا با ارائه‌ی یک الگوریتم تکاملی جدید برای خوشه‌بندی، تراکنش‌ها به خوشه‌های مناسب تقسیم می‌شوند. سپس یک الگوریتم تکاملی جدید برای کشف قوانین انجمنی ارائه می‌گردد که این الگوریتم به طور جداگانه و مستقل بر روی هر یک از خوشه ها اجرا می‌شود. در نهایت، قوانین به دست آمده از همه خوشه‌ها در یک‌جا جمع‌آوری شده و قوانین نهایی تولید می‌گردند. راه‌کار پیشنهادی به اختصار ICA-ARMC[۱] نام‌گذاری شده است.

در ادامه در فصل دوم، به بیان مسئله‌ی کشف قوانین انجمنی، ورودی و خروجی، و فرض‌های مسئله، اهداف خواهیم پرداخت. در فصل سوم، مفاهیم و اصطلاحات پایه‌ای، تعاریف و روش‌های مختلف مورد استفاده در کشف قوانین انجمنی شرح داده می‌شود. در فصل چهارم، روش‌های سنتی و روش‌های ارائه شده برای بهبود فرآیند کشف قوانین انجمنی را مورد بررسی قرار خواهیم داد. در

[۱] ICA Association Rules Mining Using Clustering

ادامه در فصل پنجم، به تشریح راه‌کار پیشنهادی و مراحل مختلف آن خواهیم پرداخت. فصل ششم، شامل آزمایش‌ها و بررسی نتایج راه‌کار پیشنهادی و مقایسه با روش‌های موجود می‌باشد و در آخر در فصل هفتم، نتیجه‌گیری و پیشنهادات برای راه‌کارهای آتی را خواهیم داشت.

فصل ۲

شرح مسئله

کشف قوانین انجمنی یکی از مهمترین روش‌های داده‌کاوی می‌باشد. با توجه به افزایش چشم‌گیر داده‌ها، کشف قوانین انجمنی به دلیل درک راحت‌تر و مقیاس‌پذیری الگوریتم‌های آن یکی از پرکاربردترین روش‌های کشف قوانین انجمنی است. در این فصل، به بیان مسئله‌ی کشف قوانین انجمنی، ورودی و خروجی، و فرض‌های مسئله و اهداف خواهیم پرداخت.

۲-۱ بیان مسئله

کشف قوانین انجمنی[۱]، یکی از روش‌های بسیار مهم داده‌کاوی[۲] جهت کشف دانش نهفته در داده‌ها می‌باشد. با استفاده از این شیوه، می‌توان روابط و وابستگی‌های مفیدی که در مجموعه‌های بزرگ از اقلام داده[۳] موجود می‌باشد را کشف نمود. اقلام داده، در قالب تراکنش‌ها[۴] ذخیره می‌شوند. تراکنش‌ها را می‌توان به کمک یک فرآیند خارجی تولید نمود، یا از پایگاه‌های داده یا انباره‌های داده استخراج نمود. یکی از کاربردهای بارز قوانین انجمنی، تحلیل سبد خرید[۵] می‌باشد که سعی می‌شود با یافتن وابستگی و روابط موجود بین اجناس خریداری شده به وسیله مشتری‌ها، الگوهای خرید، شناسایی و تحلیل شوند. به عنوان مثال، مشتری‌ها، شیر و نان را با هم می‌خرند. این نوع قوانین در افزایش سود و رقابت‌های بازاریابی استفاده می‌شود.

در مثال تحلیل سبد خرید، هر یک از محصولات موجود در فروشگاه، متناظر با یک قلم داده می‌باشد. سبد خرید متعلق به هر یک از مشتری‌ها، نشان دهنده‌ی اقلام خریداری شده به وسیله‌ی آن مشتری است. هر سبد خرید مربوط به یک مشتری به عنوان یک تراکنش در مجموعه‌ی داده

---

۱ Association Rules Mining (ARM)
۲ Data Mining (DM)
۳ itemset
۴ transactions
۵ market basket analyse

می‌باشد و شامل یک شناسه‌ی یکتا است. که در جدول ۲-۱ نمونه‌ای از یک سبد خرید مشتری یا به عبارت دیگر یک تراکنش نشان داده شده است:

جدول ۲-۱ نمونه‌ای از یک تراکنش [۲]

تراکنش (سبد خرید مربوط به هر مشتری)	TID (شناسه‌ی یکتا)
شیر، نان، تخم‌مرغ	۱۰۰۰

فرض کنید $I = \{i_1, i_2, \dots, i_m\}$، مجموعه‌ی اقلام داده و m، تعداد اقلام داده باشد. D، مجموعه‌ی داده‌های تراکنشی می‌باشد که در آن برای هر تراکنش T، یک شناسه‌ی یکتا TID وجود دارد و $T \subseteq I$ است. تراکنش T شامل مجموعه‌ی اقلام A می‌باشد اگر و فقط اگر $T \subseteq A$ باشد. یک قانون انجمنی به شکل $A \to B$ می‌باشد که در آن $A \subset I$، $B \subset I$ و $A \cap B = \emptyset$ است. با داشتن مجموعه‌ی تراکنش‌ها، مسئله‌ی کشف قوانین انجمنی شامل تولید قوانین انجمنی می‌باشد که پشتیبانی و اطمینان بالاتر از کمینه‌ی پشتیبان و اطمینان تعیین شده را داشته باشند.

معیارهای مختلفی برای ارزیابی کیفیت یک قانون وجود دارد که بر اساس آن قوانین قوی، از میان مجموعه‌ی وسیعی از قوانین ممکن انتخاب می‌شوند. از معروف‌ترین و پرکاربردترین این معیارها، پشتیبانی[۱] و اطمینان[۲] می‌باشد که به صورت درصد بیان می‌شود. پشتیبانی قانون $A \to B$ عبارت است از نسبت تعداد تراکنش‌هایی که شامل A و B هستند به تعداد کل تراکنش‌ها. به عبارت دیگر احتمال آن که A و B با هم در کل تراکنش‌ها رخ دهند. پشتیبانی، فراوانی کل قانون را با توجه به کل تراکنش‌ها بیان می‌کند:

$$\text{Support}(A \to B) = P(A \cup B) = \frac{A \cup B}{D} \qquad (۱-۲)$$

اطمینان قانون $A \to B$ عبارت است از نسبت تعداد تراکنش‌هایی که شامل A و B هستند به تراکنش‌هایی که شامل A هستند. اطمینان، قدرت دلالت قانون را بیان می‌کند:

[۱] support
[۲] confidence

$$\text{Confidence}(A \rightarrow B) = P(A|B) = \frac{A \cup B}{A} \qquad (۲-۲)$$

۲-۲ ورودی و خروجی‌های مسئله

داده‌های ورودی برای یک الگوریتم کشف قوانین انجمنی، مجموعه‌ای از تراکنش‌ها می‌باشد که هر تراکنش زیرمجموعه‌ای از اقلام می‌باشد و شامل یک شناسه‌ی یکتا است. با داشتن مجموعه‌ی تراکنش‌ها، خروجی مسئله‌ی کشف قوانین انجمنی شامل قوانین انجمنی می‌باشد که پشتیبانی و اطمینان بالاتر از کمینه‌ی پشتیبان و کمینه‌ی اطمینان تعیین شده را داشته باشند.

۲-۳ فرض‌های مسئله

مجموعه‌ی داده‌های تراکنشی، به شکل دودویی می‌باشند. به این ترتیب که اگر قلم موردنظر در تراکنش موجود باشد، مقدار آن قلم در تراکنش مربوطه ۱ و در غیر این صورت ۰ خواهد بود. هدف کشف قوانین انجمنی مثبت می‌باشد.

۲-۴ هدف

از مشکلات مهم روش های سنتی برای کشف قوانین انجمنی تولید بسیار زیاد مجموعه‌های پرتکرار و پیمایش زیاد داده‌ها می‌باشد. باتوجه به حجم بالای داده‌ها در مسائل مربوط به داده‌کاوی، در بیشتر موارد تمامی داده‌ها قابل بارگذاری در حافظه‌ی اصلی نمی‌باشند. از طرفی تولید مجموعه‌های پرتکرار دارای هزینه محاسباتی نمایی می‌باشد. بنابراین ارائه‌ی روش‌های مناسب برای کاهش این هزینه‌ها، حائز اهمیت می‌باشد.

۲-۵ نوآوری

در الگوریتم پیشنهادی برای کشف قوانین انجمنی یک روش کدگذاری سریع و ساده و همچنین دو روش برای عمل‌گر انقلاب ارائه شده است. یکی از محدودیت‌های بیشتر الگوریتم‌های تکاملی ارائه شده برای کشف قوانین انجمنی، نحوه‌ی کد کردن قوانین است که برای هر قلم موجود در

مجموعه‌ی اقلام یک یا چند مکان در نظر گرفته می‌شود و نیز ترتیب اقلام مهم می‌باشد. در این روش کدگذاری، تنها برای اقلام موجود در قانون مکان در نظر گرفته شده و هم‌چنین ترتیب اقلام مهم نیست. هم‌چنین با استفاده از دو روش ارائه شده برای عمل‌گر انقلاب، با جستجوی سریع نقاط مختلف فضای بزرگ مسئله، تعداد بیشتری قوانین در زمان کم تولید می‌گردد. راه‌کار پیشنهادی علاوه بر قوانین انجمنی، توانایی بالایی در کشف قوانین انجمنی نادر دارد.

۲-۶ سوالات تحقیق

- آیا روش خوشه‌بندی جدید تراکنش‌ها، می‌تواند تعداد پیمایش داده‌ها را کاهش دهد؟
- آیا الگوریتم تکاملی می‌تواند قوانین انجمنی مناسب را در زمان مناسب تولید نماید؟
- آیا استفاده از الگوریتم تکاملی به طور جداگانه و موازی در هر یک از خوشه‌ها می‌تواند قوانین انجمنی را در زمان مناسب تولید نماید؟

۲-۷ فرضیات تحقیق

- روش خوشه‌بندی جدید تراکنش‌ها، تعداد پیمایش داده‌ها را کاهش خواهد داد.
- الگوریتم تکاملی، قوانین انجمنی مناسب را در زمان مناسب تولید خواهد کرد.
- استفاده از الگوریتم تکاملی به طور جداگانه و موازی در هر یک از خوشه‌ها، قوانین انجمنی را در زمان مناسب تولید خواهد نمود.

۲-۸ خلاصه فصل

کشف قوانین انجمنی به دلیل درک راحت‌تر، یکی از مهم‌ترین شیوه‌های داده‌کاوی می‌باشد که با استفاده از این شیوه می‌توان روابط و وابستگی‌های مفیدی که در مجموعه‌های بزرگ از اقلام داده موجود می‌باشد را کشف نمود. فرآیند کشف قوانین انجمنی در روش‌های سنتی بسیار هزینه‌بر می‌باشد از طرفی حجم داده‌ها در مسائل مربوط به داده‌کاوی بالا می‌باشد، بنابراین ارائه‌ی روش‌هایی کاهش این هزینه‌ها در فرآیند کشف قوانین انجمنی بسیار مفید وموثر خواهد بود.