





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 پایان‌نامه یاری نمودند، تشکر و
قدردانی می‌نمایم.

لیلا امیری
شهریور ۱۳۹۳

نام خانوادگی: امیری	نام: لیلا
عنوان پایان نامه: 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 بین‌المللی 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 خوشه‌ها 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

صفحه	عنوان
۱	فصل ۱ - مقدمه
۵	فصل ۲ - شرح مسئله
۶	۱-۲ بیان مسئله
۸	۲-۲ ورودی و خروجی
۸	۳-۲ فرض‌های مسئله
۸	۴-۲ هدف
۸	۵-۲ نوآوری
۹	۶-۲ سوالات تحقیق
۹	۷-۲ فرضیات تحقیق
۹	۸-۲ خلاصه فصل
۱۱	فصل ۳ - مفاهیم پایه‌ای
۱۱	۱-۳ داده‌کاوی
۱۳	۲-۳ طبقه‌بندی
۱۳	۳-۳ خوشه‌بندی
۱۴	۱-۳-۳ اندازه‌گیری شباهت
۱۷	۲-۳-۳ طبقه بندی الگوریتم‌های خوشه‌بندی
۲۱	۳-۳-۳ الگوریتم K-Means
۲۲	۴-۳-۳ الگوریتم K-Medoids
۲۲	۵-۳-۳ الگوریتم K-Modes
۲۳	۶-۳-۳ اعتبار خوشه
۲۵	۴-۳ کشف قوانین انجمنی
۲۵	۱-۴-۳ اصطلاحات و تعاریف

۲-۴-۳ قوانین انجمنی	۲۶
۶-۴-۳ انواع الگوریتم‌های کشف قوانین انجمنی	۲۹
۶-۳ بهینه‌سازی	۳۱
۱-۶-۳ بهینه‌سازی چند هدفه	۳۲
۲-۶-۳ بررسی روش‌های بهینه‌سازی	۳۳
۴-۶-۳ نسخه گسسته الگوریتم رقابت استعماری	۳۹
۷-۳ خلاصه فصل	۴۰
فصل ۴ - راه‌کارهای گذشته	۴۱
۱-۴ روش‌های سنتی	۴۲
۱-۱-۴ الگوریتم AIS	۴۲
۲-۱-۴ الگوریتم Apriori	۴۳
۲-۴ روش‌های بهبود فرآیند کشف قوانین انجمنی	۴۴
۳-۴ شیوه‌ای مبتنی بر الگوریتم ICA جهت استخراج قوانین انجمنی	۵۰
۴-۴ روش‌های کشف قوانین انجمنی مبتنی بر بخش‌بندی با استفاده از خوشه‌بندی	۵۳
۱-۴-۴ کشف قوانین انجمنی نادر با استفاده از خوشه‌بندی بر اساس Seed ها ..	۵۳
۲-۴-۴ کشف قوانین با استفاده از الگوریتم ژنتیک بر اساس خوشه‌بندی	۵۵
۳-۴-۴ کشف قوانین انجمنی فازی مبتنی بر خوشه‌بندی	۵۹
۵-۴ خلاصه	۶۸
فصل ۵ - راه‌کار پیشنهادی	۶۸
۱-۵ مراحل راه‌کار پیشنهادی	۶۹
۱-۱-۵ الگوریتم پیشنهادی برای خوشه‌بندی تراکنش‌ها	۶۹
۲-۱-۵ الگوریتم پیشنهادی برای کشف قوانین انجمنی	۸۱

۳-۱-۵ جمع‌آوری قوانین تولید شده از کل خوشه‌ها و تولید قوانین نهایی	۹۲
۲-۵ فلوجارت راه‌کار پیشنهادی	۹۲
۳-۵ خلاصه فصل	۹۳
فصل ۶ - ارزیابی و نتایج عملی	۹۵
۱-۶ آزمایشات	۹۶
۱-۱-۶ قابلیت اطمینان	۹۶
۲-۱-۶ بررسی هم‌گرایی	۹۹
۳-۱-۶ بررسی پایداری	۱۰۲
۲-۶ بررسی نتایج حاصل از راه‌کار پیشنهادی و مقایسه آن با روش‌های موجود	۱۰۴
۱-۲-۶ بررسی نتایج الگوریتم پیشنهادی و مقایسه با الگوریتم Apriori	۱۰۴
۲-۲-۶ مقایسه الگوریتم پیشنهادی و الگوریتم AICluster	۱۱۳
۳-۲-۶ مقایسه الگوریتم پیشنهادی و الگوریتم MINICA	۱۱۵
۳-۶ بحث	۱۱۶
۴-۶ خلاصه فصل	۱۱۷
فصل ۷ - نتیجه‌گیری و کارهای آتی	۱۱۸
۱-۷ نتیجه‌گیری	۱۱۹
۲-۷ راه‌کارهای آتی	۱۲۱
مراجع	۱۲۳

فهرست جدول‌ها

عنوان	صفحه
جدول ۱-۲ نمونه‌ای از یک تراکنش [۲].....	۷
جدول ۱-۳ توابع فاصله میان داده‌های x و y [۲].....	۱۴
جدول ۲-۳ ترکیب‌های ممکن مقادیر x و y برای اندازه‌گیری شباهت [۲].....	۱۶
جدول ۱-۴ محاسبه‌ی سرعت (بر اساس ثانیه) بر اساس تعداد رکوردها [۱۹].....	۵۸
جدول ۲-۴ نتایج خوشه‌بندی پس از اجرای الگوریتم خوشه‌بندی [۶].....	۶۴
جدول ۱-۶ پارامترهای راه‌کار پیشنهادی برای اجرای بر روی TestData.....	۹۸
جدول ۲-۶ پارامترهای الگوریتم پیشنهادی برای خوشه‌بندی بر روی داده TestData.....	۹۹
جدول ۳-۶ مقدار پارامترهای الگوریتم در مجموعه داده Zoo.....	۱۰۷
جدول ۴-۶ مقایسه نتایج اجرا الگوریتم‌های مختلف بر روی مجموعه داده Zoo.....	۱۰۷
جدول ۵-۶ مقدار پارامترها در مجموعه داده Data5000.....	۱۰۸
جدول ۶-۶ مقایسه نتایج حاصل از اجرای الگوریتم در Data5000.....	۱۰۹
جدول ۷-۶ پارامترهای در نظر گرفته شده برای مجموعه داده Mushroom.....	۱۱۱
جدول ۸-۶ پارامترهای در نظر گرفته شده برای مجموعه داده ChessData.....	۱۱۱
جدول ۹-۶ مقایسه نتایج حاصل از اجرای الگوریتم بر روی مجموعه داده Mushroom.....	۱۱۲
جدول ۱۰-۶ مقایسه نتایج حاصل از اجرای الگوریتم بر روی مجموعه داده ChessData.....	۱۱۲
جدول ۱۱-۶ مقدار پارامترهای الگوریتم در مجموعه داده Zoo.....	۱۱۳
جدول ۱۲-۶ نتایج مقایسه الگوریتم پیشنهادی و الگوریتم Apriori Inverse و AICluster.....	۱۱۴
جدول ۱۳-۶ مقدار پارامترهای الگوریتم در مجموعه داده Car Evaluation.....	۱۱۵
جدول ۱۴-۶ نتایج مقایسه الگوریتم پیشنهادی و الگوریتم MINICA.....	۱۱۵

فهرست شکل‌ها

عنوان	صفحه
شکل ۱-۳ داده‌های دودویی x و y [۲].....	۱۵
شکل ۲-۳ مثالی از یک سلسله مراتب برای مواد غذایی [۲].....	۲۷
شکل ۳-۳ شمای کلی فرآیند سیاست جذب.....	۳۶
شکل ۴-۳ شمای کلی فرآیند جابه‌جایی کشور استعمارگر و مستعمره.....	۳۷
شکل ۵-۳ شمای کلی فرآیند رقابت استعماری میان امپراطوری‌ها.....	۳۸
شکل ۶-۳ شمای کلی فرآیند سقوط امپراطوری‌های ضعیف.....	۳۹
شکل ۱-۴ نحوه‌ی کدگذاری قانون در یک کشور [۸].....	۵۱
شکل ۲-۴ نحوه‌ی نمایش قانون در یک کشور [۸].....	۵۱
شکل ۳-۴ ساختار کلی روش پیشنهادی [۱۹].....	۵۷
شکل ۴-۴ کدگذاری کروموزوم برای قانون $ACF \rightarrow BE$ [۱۹].....	۵۸
شکل ۱-۵ نحوه‌ی نمایش یک کشور.....	۷۱
شکل ۲-۵ نحوه‌ی نمایش یک کشور.....	۷۱
شکل ۳-۵ A، کشور استعمارگر و B، کشور مستعمره.....	۷۴
شکل ۴-۵ نتیجه محاسبه A-B.....	۷۴
شکل ۵-۵ نتیجه محاسبه B-A.....	۷۴
شکل ۶-۵ Add و Del پس حذف برخی از عناصر.....	۷۵
شکل ۷-۵ position جدید برای B پس از اعمال سیاست جذب.....	۷۵
شکل ۸-۵ عناصر انتخاب شده از Position کشور برای تغییر.....	۷۶
شکل ۹-۵ Position جدید کشور پس از اجرای انقلاب با استفاده از روش اول.....	۷۶
شکل ۱۰-۵ Position جدید کشور پس از اجرای انقلاب با استفاده از روش دوم.....	۷۷
شکل ۱۱-۵ نمایش قانون $AD \rightarrow BCE$	۸۲

- شکل ۵-۱۲ نحوه نمایش قانون $AB \rightarrow CD$ ۸۴
- شکل ۵-۱۳ نحوه نمایش قانون $E \rightarrow AC$ ۸۴
- شکل ۵-۱۴ مجموعه‌های Addl و Remove ۸۵
- شکل ۵-۱۵ Remove و Addl پس حذف برخی از عناصر ۸۵
- شکل ۵-۱۶ مجموعه‌های AddR و RemoveR ۸۵
- شکل ۵-۱۷ مجموعه AddR و RemoveR پس از حذف برخی از عناصر ۸۶
- شکل ۵-۱۸ کشور مستعمره پس از اعمال سیاست جذب ۸۶
- شکل ۵-۱۹ مقداردهی اولیه آرایه Chrom ۸۸
- شکل ۵-۲۰ قرار گرفتن اقلام موجود در قانون $AB \rightarrow C$ به طور تصادفی در آرایه chrom ۸۸
- شکل ۵-۲۱ عناصر انتخاب شده از آرایه chrom برای تغییر ۸۸
- شکل ۵-۲۲ تغییر مکان‌های انتخابی با محتوای ۰ در آرایه chrom ۸۹
- شکل ۵-۲۳ تغییر مکان‌های انتخابی با محتوای غیر صفر در آرایه chrom ۸۹
- شکل ۵-۲۴ نحوه نمایش قانون تولید شده پس از اجرای عملگر انقلاب بر روی قانون $AB \rightarrow C$ ۹۰
- شکل ۵-۲۵ فلوجارت راه کار پیشنهادی ۹۳
- شکل ۶-۱ نتایج اجرای الگوریتم Apriori بر روی TestData ۹۷
- شکل ۶-۲ نتایج اجرای الگوریتم پیشنهادی بر روی TestData ۹۸
- شکل ۶-۳ نمودار هم‌گرایی الگوریتم پیشنهادی برای خوشه‌بندی بر روی داده TestData ۱۰۰
- شکل ۶-۴ نمودار هم‌گرایی الگوریتم کشف قوانین انجمنی بر روی داده TestData ۱۰۲
- شکل ۶-۵ نمودار پایداری الگوریتم پیشنهادی برای خوشه‌بندی بر روی داده TestData ۱۰۳
- شکل ۶-۶ نتایج آزمون پایداری الگوریتم پیشنهادی برای خوشه‌بندی تراکنش‌ها ۱۰۳
- شکل ۶-۷ نمودار پایداری الگوریتم پیشنهادی کشف قوانین انجمنی بر روی داده TestData ۱۰۴

- شکل ۶-۸ نتایج آزمون پایداری الگوریتم کشف قوانین انجمنی بر روی داده TestData ۱۰۴
- شکل ۶-۹ نتایج الگوریتم Apriori بر روی مجموعه داده Zoo ۱۰۵
- شکل ۶-۱۰ نتایج الگوریتم پیشنهادی بر روی مجموعه داده Zoo ۱۰۶
- شکل ۶-۱۱ نتایج الگوریتم پیشنهادی بدون مرحله پس پردازش بر روی مجموعه داده Zoo ۱۰۶
- شکل ۶-۱۲ نتایج اجرای الگوریتم پیشنهادی غیرخوشه‌بندی بر روی مجموعه داده Zoo ۱۰۶
- شکل ۶-۱۳ نتایج الگوریتم بدون مرحله خوشه‌بندی بر روی مجموعه داده Data5000 ۱۰۸
- شکل ۶-۱۴ نتایج الگوریتم پیشنهادی بر روی مجموعه داده Data5000 ۱۰۹
- شکل ۶-۱۵ نتایج الگوریتم پیشنهادی بر روی مجموعه داده Zoo ۱۱۳
- شکل ۶-۱۶ نمودار هم‌گرایی الگوریتم پیشنهادی و الگوریتم MINICA ۱۱۶

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 داده‌کاوی 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 کارایی 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 خوشه‌ها اجرا می‌شود. در نهایت، قوانین به دست آمده از همه خوشه‌ها در یک جا 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 راه کار پیشنهادی و مراحل مختلف آن خواهیم پرداخت. فصل ششم، شامل آزمایش‌ها و بررسی نتایج راه کار پیشنهادی و مقایسه با روش‌های موجود می‌باشد و در آخر در فصل هفتم، نتیجه‌گیری و پیشنهادات برای راه کارهای آتی را خواهیم داشت.

فصل ۲

شرح مسأله

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 انبار‌های 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 ۱-۲ نمونه‌ای از یک سبد خرید مشتری یا به عبارت دیگر یک تراکنش نشان داده شده است:

جدول ۱-۲ نمونه‌ای از یک تراکنش [۲]

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

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 $A \rightarrow B$ 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 $A \rightarrow B$ 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:

$$\text{Support}(A \rightarrow B) = P(A \cup B) = \frac{A \cup B}{D} \quad (1 - 2)$$

اطمینان قانون $A \rightarrow B$ 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

$$\text{Confidence}(A \rightarrow B) = P(A|B) = \frac{A \cup B}{A} \quad (2-2)$$

۲-۲ ورودی و خروجی های مسئله

داده های ورودی برای یک الگوریتم کشف قوانین انجمنی، مجموعه ای از تراکنش ها می باشد که هر تراکنش زیرمجموعه ای از اقلام می باشد و شامل یک شناسه ی یکتا است. با داشتن مجموعه ی تراکنش ها، خروجی مسئله ی کشف قوانین انجمنی شامل قوانین انجمنی می باشد که پشتیبانی و اطمینان بالاتر از کمینه ی پشتیبان و کمینه ی اطمینان تعیین شده را داشته باشند.

۳-۲ فرض های مسئله

مجموعه ی داده های تراکنشی، به شکل دودویی می باشند. 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 می باشد.

۴-۲ هدف

از مشکلات مهم روش های سنتی برای کشف قوانین انجمنی تولید بسیار زیاد مجموعه های پرتکرار و پیمایش زیاد داده ها می باشد. باتوجه به حجم بالای داده ها در مسائل مربوط به داده کاوی، در بیشتر موارد تمامی داده ها قابل بارگذاری در حافظه ی اصلی نمی باشند. از طرفی تولید مجموعه های پرتکرار دارای هزینه محاسباتی نمایی می باشد. بنابراین ارائه ی روش های مناسب برای کاهش این هزینه ها، حائز اهمیت می باشد.

۵-۲ نوآوری

در الگوریتم پیشنهادی برای کشف قوانین انجمنی یک روش کدگذاری سریع و ساده و هم چنین دو روش برای عمل گر انقلاب ارائه شده است. یکی از محدودیت های بیشتر الگوریتم های تکاملی ارائه شده برای کشف قوانین انجمنی، نحوه ی 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 و موثر خواهد بود.