نشریه علمی‌ترویجی محاسبات نرم
سال پنجم/ شمارهٔ اول/ بهار و تابستان ۹۵/ صفحهٔ ۱۴ـ۲۵

دریافت مقاله: ۱۳۹۴/۲/۱۵
پذیرش مقاله: ۱۳۹۴/۱۰/۲۵

ارائهٔ یک الگوریتم خوشه‌بندی برای داده‌های دسته‌ای با ترکیب معیارها

مریم نبی‌لو،[۱] نگین دانشپور[۲*]

[۱] دانشجوی کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی
mnabiloo13@yahoo.com
[۲] استادیار دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی
ndaneshpour@srttu.edu

چکیده

خوشه‌بندی یکی از تکنیک‌های اصلی داده‌کاوی است. خوشه‌بندی فرایندی است که مجموعه داده‌ها را داخل گروه‌هایی طبقه‌بندی می‌کند. در خوشه‌بندی، داده‌های موجود در یک خوشه بیشترین شباهت را به‌هم دارند و داده‌های موجود در دو خوشهٔ متفاوت، بیشترین تفاوت را با هم دارند. الگوریتم‌های خوشه‌بندی با توجه به نوع داده‌ها به دو دسته تقسیم می‌شوند: الگوریتم‌های خوشه‌بندی داده‌های عددی و الگوریتم‌های خوشه‌بندی داده‌های دسته‌ای. الگوریتم‌های خوشه‌بندی داده‌های دسته‌ای به‌دلیل ماهیت و کاربرد این داده‌ها نسبت‌به الگوریتم‌های خوشه‌بندی داده‌های عددی، اهمیت بیشتری دارند. هریک از این الگوریتم‌ها با توجه به نوع داده (عددی یا دسته‌ای) از معیارهای شباهت متفاوتی در خوشه‌بندی استفاده می‌کنند. در این مقاله، ابتدا به بررسی ماهیت این نوع داده‌ها پرداخته شده و سپس معیارهای شباهت و الگوریتم‌های خوشه‌بندی مطرح‌شده در این حوزه را بررسی کرده و درنهایت، یک روش جدید در خوشه‌بندی با استفاده از ترکیب معیارهای شباهت برای داده‌های دسته‌ای ارائه می‌کنیم. آزمایش‌ها نشان می‌دهد که روش ارائه‌شده در این مقاله، توانسته است که نتایج حاصل از خوشه‌بندی را بهبود ببخشد.

واژه‌های کلیدی: داده‌کاوی، خوشه‌بندی، داده‌های دسته‌ای، معیار فاصله، معیار چگالی.

۱. مقدمه

امروزه با توجه به حجم بالای داده‌ها در علوم مختلف، می‌باید اطلاعات مفید از داخل مجموعه داده‌ها استخراج شود [۱]. زمانی که حجم داده‌ها افزایش می‌یابد، کاربران هرچند زبردست و باتجربه باشند، نمی‌توانند الگوهای مفید را در میان حجم انبوه داده‌ها تشخیص دهند [۲]. یکی از تکنیک‌هایی که طی سالیان اخیر برای انجام این امر مطرح شده، تکنیک داده‌کاوی است [۳].

داده‌کاوی به‌عنوان یک ابزار قوی برای تولید اطلاعات و دانش از داده‌های خام، شناخته شده است. برای داده‌کاوی تعاریف مختلفی وجود دارد که معمول‌ترین آن عبارت از: استخراج دانش و اطلاعات از یک پایگاه دادهٔ بسیار بزرگ و پیچیده (تبدیل داده‌ها به دانش) و همچنین کشف الگوهای پنهان بین داده‌ها [۴ـ۶]. بنابراین، ابزارهای داده کاوی الگوهای پنهانی را کشف و پیش‌بینی می‌کنند که متخصصان ممکن است به‌دلیل اینکه این اطلاعات و الگوها خارج از انتظار آن‌ها باشد، آن‌ها را مدنظر قرار ندهند و به آن‌ها دست نیابند [۷].

تکنیک‌های مختلفی نظیر دسته‌بندی[1] و رگرسیون[2] و خوشه‌بندی[3] برای داده‌کاوی ارائه شده است. خوشه‌بندی داده‌ها [۸ـ۱۰] به‌معنی گروه‌بندی نمونه‌ها در گروه‌های شبیه‌به‌هم است؛ به‌طوری‌که نمونه‌های هر گروه (خوشه) حداکثر تشابه را با یکدیگر و حداکثر فاصله را با نمونه‌های گروه‌های دیگر داشته باشند [۱۱].

روش‌های بسیاری برای خوشه‌بندی داده‌ها پیشنهاد شده است. اکثر این روش‌ها روی جنبه‌های خاصی از داده‌ها تأکید کرده و بنابراین، روی مجموعه داده‌های خاصی کارایی خوبی دارند. الگوریتم‌های زیادی در این زمینه معرفی شده است. دسته‌بندی متفاوتی برای الگوریتم‌های خوشه‌بندی وجود دارد که رویکرد زیر جامع‌ترین دسته‌بندی این روش‌هاست [۱۱]:

ـ الگوریتم‌های تفکیکی[4]

ـ الگوریتم‌های سلسله‌مراتبی[5]

ـ الگوریتم‌های مبتنی بر چگالی[6]

ـ الگوریتم‌های مبتنی بر گرید[7]

ـ الگوریتم‌های مبتنی بر مدل[8]

همچنین، الگوریتم‌های خوشه‌بندی با توجه به نوع داده‌ها تقسیم‌بندی دیگری نیز دارند: الگوریتم‌های خوشه‌بندی داده‌های عددی و الگوریتم‌های خوشه‌بندی داده‌های دسته‌ای.

الگوریتم‌های خوشه‌بندی داده‌های دسته‌ای به‌دلیل ماهیت و کاربرد این داده‌ها نسبت‌به الگوریتم‌های خوشه‌بندی داده‌های عددی، اهمیت بیشتری دارند. داده‌های دسته‌ای شامل حداقل دو مقدار متمایزند که هیچ‌گونه تقدم یا تأخر در آن‌ها وجود ندارد. به‌عبارت دیگر، میان مقادیر داده‌های دسته‌ای نمی‌توان ترتیب خاصی در نظر گرفت. به همین سبب، چالش‌هایی در زمینهٔ خوشه‌بندی داده‌های دسته‌ای مطرح است که شامل موارد زیر است [۱۲ و ۱۳].

- عدم امکان تفکیک‌پذیری روی دامنه برای داده‌های دسته‌ای: برای داده‌های عددی نمایندهٔ هر خوشه اغلب شامل میانگین داده‌ها در دامنهٔ صفات هر خوشه است. محاسبهٔ میانگین برای داده‌های دسته‌ای، غیرعملی است [۱۴].
- عدم امکان اندازه‌گیری شباهت برای داده‌های دسته‌ای به روش معمول داده‌های عددی: توابع فاصله نظیر فاصلهٔ منهتن و فاصلهٔ اقلیدسی برای داده‌های دسته‌ای به‌دلیل اینکه هیچ ترتیبی بین مقادیر داده وجود ندارد، کارایی ندارد [۱۵].

تابه‌حال، الگوریتم‌های بسیاری برای خوشه‌بندی داده‌های دسته‌ای ارائه شده است که هریک از آن‌ها معیار شباهت جدیدی را برای این نوع داده‌ها معرفی می‌کند. این معیارهای شباهت ارائه‌شده، هریک مزایا و معایب خاص خود را دارد. به همین منظور، یعنی برای کاهش معایب معیارهای شباهت در این مقاله، دو تا از معروف‌ترین این معیارها را با هم ترکیب کرده‌ایم.

در این مقاله، ابتدا به تعریف و تبیین مفهوم خوشه‌بندی،

1. Classification
2. Regression
3. Clustering
4. Partitioning Algorithms
5. Hierarchical Algorithms
6. Density –Based Algorithms
7. Grid-Based Algorithms
8. Model - Based Algorithms

به‌خصوص خوشه‌بندی داده‌های دسته‌ای پرداخته شده است. در ادامه، روش ترکیبی جدیدی بر پایهٔ ترکیب دو الگوریتم خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی تفکیکی برای خوشه‌بندی دادهٔ دسته‌ای ارائه شده است. آزمایش‌ها نشان می‌دهد که روش یادشده موجب بهبود در نتایج خوشه‌بندی داده‌ها می‌شود. تمام الگوریتم‌های خوشه‌بندی ترکیبی در حوزهٔ داده‌های عددی قرار دارند و از روش‌های ترکیبی در زمینهٔ خوشه‌بندی داده‌های دسته‌ای استفاده نشده است. بنابراین، در این مقاله یک رویکرد جدید بر مبنای ترکیب الگوریتم‌های موجود برای خوشه‌بندی داده‌های دسته‌ای ارائه می‌شود.

۲. پیشینهٔ پژوهش

خوشه‌بندی فرایندی است که مجموعه داده‌ها را داخل گروه‌هایی طبقه‌بندی می‌کند. الگوریتم‌های متنوعی برای خوشه‌بندی داده‌ها ارائه شده است. هریک از این الگوریتم‌ها معیار شباهت خاص خود را دارد. این الگوریتم‌ها بسته به معیار شباهتی که استفاده می‌کنند و روشی که در خوشه‌بندی داده‌ها دارند، توانایی متفاوتی در مدیریت داده‌های پرت و دارای نویز دارند. الگوریتم k-means یکی از اولین الگوریتم‌های خوشه‌بندی تفکیکی بود که به‌صورت وسیع و مؤثر، در خوشه‌بندی مجموعه داده‌های مورد استفاده قرار می‌گرفت [۱۶]. هرچند استفاده از انواع الگوریتم‌های k-means فقط محدود به داده‌های عددی است، بیشتر داده‌های جهان واقعی از نوع داده‌های دسته‌ای هستند. خوشه‌بندی داده‌های دسته‌ای به‌دلیل ماهیت و کاربرد این داده‌ها نسبت‌به خوشه‌بندی داده‌های عددی، اهمیت بیشتری دارند. برای خوشه‌بندی داده‌های دسته‌ای الگوریتم k-means به الگوریتم k-modes [۱۷] گسترش یافت. در این الگوریتم، محدودیت درخصوص داده‌های عددی از بین رفت و این الگوریتم قادر بود که به‌طور مؤثر روی داده‌های دسته‌ای اجرا شود. در این الگوریتم، معیار شباهت جدیدی برای داده‌های دسته‌ای معرفی شد. معیار شباهت بین دو دادهٔ دسته‌ای برابر تعداد مقادیر صفات غیرمشابه میان آن دو داده است. الگوریتم خوشه‌بندی k-modes میانگین خوشه‌ها را با مفهوم mode جایگزین کرد و از روش‌های مبتنی بر فرکانس برای به‌روزرسانی این modeها در فرایند خوشه‌بندی استفاده کرد تا به این وسیله، تابع هدف خوشه‌بندی را مینیمم کند. الگوریتم خوشه‌بندی k-modes چندین بار با مقادیر متفاوت modeهای اولیه اجرا می‌شود تا پایداری راه حل خوشه‌بندی به‌دست آید. اما این الگوریتم همانند الگوریتم k-means توانایی مدیریت داده‌های پرت و نویز را در فرایند خوشه‌بندی ندارد و همچنین جواب نهایی به انتخاب مراکز اولیه خوشه‌ها وابسته است. علاوه بر این، الگوریتم fuzzy k-modes نیز ارائه شد که در آن، داده‌ها می‌توانند با یک درجه عضویت، عضوی از تمام خوشه‌ها باشند [۱۸]. این امر مهم‌ترین مزیت این الگوریتم است و موجب شده این الگوریتم برای داده‌های دارای نویز مناسب باشد. اما این الگوریتم نیز مانند الگوریتم k-modes توانایی مدیریت داده‌های پرت را ندارد. در ادامه، تعدادی از الگوریتم‌های خوشه‌بندی مختص داده‌های دسته‌ای را بررسی می‌کنیم:

الگوریتم خوشه‌بندی weighting k-modes [۳]، در واقع همان الگوریتم k-modes است با این تفاوت که به هر صفت، وزنی اختصاص داده است. دلیل وزن دادن به صفات در این الگوریتم، این است که ممکن است بعضی از صفات در مجموعه داده‌ها اهمیت بیشتری داشته باشند. این‌گونه صفات در فرایند خوشه‌بندی، نقش بیشتری را ایفا می‌کنند. وزن یک صفت معمولاً عددی بین ۰ تا ۱ است و صفتی که اهمیت بیشتری داشته باشد، وزن بیشتری دارد. مکانیزم وزن‌دهی به صفات در این الگوریتم، موجب بهبود کیفیت خوشه‌بندی می‌گردد. اما این الگوریتم توانایی مدیریت داده‌های پرت و دارای نویز را ندارد.

الگوریتم خوشه‌بندی weighting fuzzy k-modes [۱۹]، مدل فازی الگوریتم weighting k-modes است. این الگوریتم در محاسبهٔ شباهت به هر صفت، بسته به اهمیت آن یک وزن در بازه ۰ و ۱ می‌دهد. در این الگوریتم، هر داده می‌تواند با یک درجه عضویت، عضوی از تمام خوشه‌ها باشد. مزیت این الگوریتم نسبت‌به الگوریتم weighting k-modes این است که این الگوریتم توانایی مدیریت داده‌های دارای نویز را دارد. اما همچنان توانایی مدیریت داده‌های پرت را ندارد و مشکل انتخاب مراکز اولیهٔ خوشه‌ها در آن، به قوّت خود باقی است.

الگـوریتم خوشـه‌بنـدی k-modes based on entropy [۲۰]، مشابه الگوریتم خوشه‌بندی k-modes است، با این تفاوت که ایـن الگوریتم به‌جای استفاده از معیار شـباهت overlap بـرای محاسـبهٔ شباهت میان اشیاء داده و مراکـز خوشـه، از معیـار آنتروپـی بـرای محاسبهٔ شباهت استفاده می‌کند و بـا مینـیمم‌سـازی آنتروپـی کلـی موردانتظار، خوشه‌بندی را انجـام مـی‌دهـد. ایـن الگـوریتم کـه از جدیدترین نسـخه‌هـای ارائـه‌شـده از الگـوریتم k-modes اسـت، قابلیت فراوانی در فرایند خوشه‌بندی دارد. این الگـوریتم همچنـین توانسته است مشکل مدیریت داده‌های پـرت و دارای نـویز را کـه الگـوریتم k-modes در آن نـاتوان بـود، بـه‌خـوبی حـل کنـد. امـا همچنان مشکل انتخاب مراکز اولیهٔ خوشه‌ها در این الگـوریتم نیـز به قوّت خود باقی است.

در الگـوریتم خوشـه‌بنـدی fuzzy k-prototype [۲۱]، k عـدد prototype از میان اشیاء داده به‌وسـیلهٔ مینـیمم‌سـازی تـابع فاصـله انتخاب می‌شود. این الگوریتم مانند الگوریتم خوشـه‌بنـدی fuzzy k- modes عمل می‌کند، با این تفاوت که مفهوم mode بـا مفهـوم prototype جایگزین شـده اسـت. انگیـزهٔ ایـن الگـوریتم بررسـی اهمیت صفات مختلف در فرایند خوشه‌بندی است. ایـن الگـوریتم هم صفات عددی و هم صفات دسـته‌ای را بـا وزنـی متفـاوت در فرایند خوشه‌بندی سهیم می‌کند. این امر موجب می‌شـود کـه ایـن الگوریتم برای داده‌ها با صفات ترکیبی (صفات عددی و دسـته‌ای) مناسب عمل کند. اما همچنان توانایی مـدیریت داده‌هـای پـرت را ندارد و همچنان مشکل انتخـاب مراکـز اولیـهٔ خوشـه‌هـا در ایـن الگوریتم، به قوّت خود باقی است.

الگوریتم خوشه‌بندی Squeezer [۲۲] شـباهت بـین داده‌هـا را با یک معیار شباهت ویژه بررسی مـی‌کنـد تـا دادهٔ مـوردنظر را در یکی از خوشه‌های موجود یا در یک خوشهٔ جدید قرار دهـد؛ کـه برای این کار از یک مقـدار آسـتانه اسـتفاده مـی‌کنـد. اگـر میـزان شباهت یک شیء به یک خوشه، بیشـتر از مقـدار آسـتانه بـود، آن شیء در آن خوشه قرار می‌گیرد و در غیـر ایـن صـورت، در یـک خوشهٔ جدید مستقر مـی‌شـود. الگـوریتم خوشـه‌بنـدی Squeezer برای داده‌های دسته‌ای استفاده مـی‌شـود و بـا دادن وزن بیشـتر بـه صفات غیرمشابه، خوشه‌بندی را انجام می‌دهد. این الگـوریتم تنهـا با یک‌بار بررسی مجموعه داده‌ها می‌تواند بـه نتیجـهٔ خوشـه‌بنـدی خوبی دست یابد و به‌منظور افزایش کیفیت خوشه‌ها می‌تـوان ایـن عمل را چندین بار تکـرار کـرد. ایـن الگـوریتم داده‌هـای پـرت و نویزدار را نیز به‌خوبی مدیریت می‌کند. عیب این الگـوریتم تعیـین مقدار آستانه برای مقایسهٔ شباهت است که این مقـدار بـه‌صـورت تجربی به‌دست می‌آید.

الگوریتم خوشـه‌بنـدی COOLCAT [۲۳] از مفهـوم آنتروپـی برای گروه‌بندی رکوردهـا اسـتفاده مـی‌کنـد. ایـن الگـوریتم، یـک الگوریتم افزایشی با هدف مینیمم‌سازی آنتروپی موردانتظـار بـرای خوشـه‌هاسـت. بـا داشـتن یـک مجموعـه از خوشـه، الگـوریتم COOLCAT نقطهٔ بعدی در مجموعه نقاط داده را با مینیمم‌سـازی آنتروپی کلی موردانتظار، خوشه‌بندی می‌کند. الگوریتم خوشه‌بندی COOLCAT بدون هـیچ پـیش‌پردازشـی روی مجموعـه داده‌هـا، خوشه‌بندی را انجام می‌دهد. بنابراین الگوریتم COOLCAT بـرای داده‌های جریانی مناسب است، اما توانایی مدیریت داده‌های پـرت را ندارد.

الگوریتم خوشه‌بندی CLOPE [۲۴] بـرای هـر خوشـه، یـک نمودار ستونی ترسیم می‌کند و بر مبنای مینیمم‌سازی نسـبت طـول به عرض مربوط به آن خوشـه، خوشـه‌بنـدی را انجـام مـی‌دهـد. الگوریتم خوشه‌بندی CLOPE سریع و مقیاس‌پذیر اسـت و بـرای جداسازی تراکنش‌های با ابعاد بالا مناسب است. این الگـوریتم بـه ترتیـب داده‌هـای ورودی حسـاس نیسـت، امـا توانـایی مـدیریت داده‌های پرت را ندارد.

در این بخش، الگوریتم‌های مختلفـی کـه بـرای خوشـه‌بنـدی داده‌های دسته‌ای معرفی شـده بـود، بررسـی شـد. هریـک از ایـن الگوریتم‌ها از یک معیار شباهت برای خوشه‌بندی داده‌هـا اسـتفاده می‌کند و برای موقعیت خاصی ارائه شده است. همان‌طور که بیـان شد، هریک از الگوریتم ها توانایی متفـاوتی در مـدیریت داده‌هـای پرت دارد. در این مقاله، روشی ترکیبـی بـرای خوشـه‌بنـدی بهتـر داده‌های دسته‌ای بر مبنای الگوریتم سلسله‌مراتبی تجمعی و بهبـود این روش ارائه می‌شود. در روش پیشنهادی، مشکل داده‌های پرت و دارای نویز در فرایند خوشه‌بندی، تا حد زیادی مرتفع شد.

۳. روش پیشنهادی

در اینجا از یک الگوریتم سلسله‌مراتبی با ترکیب خطی دو معیار شباهت، یعنی معیار همگرایی Jaccard و معیار چگالی استفاده می‌کنیم. به‌منظور تست الگوریتم پیشنهادی، دو مجموعه داده soybean data و zoo data انتخاب شده و الگوریتم پیشنهادی را روی آن‌ها به اجرا رسانده‌ایم. در ادامه، به شرح روش خوشه‌بندی سلسله‌مراتبی و هریک از معیارهای شباهت همگرایی Jaccard و چگالی پرداخته شده و سپس روش ترکیبی پیشنهادی بیان می‌شود.

۱.۳. خوشه‌بندی سلسله‌مراتبی

روش خوشه‌بندی سلسله‌مراتبی یک روش تحلیل خوشه‌ای است که برای داده‌های کم‌حجم و داده‌های دسته‌ای به‌کار می‌رود. روش خوشه‌بندی سلسله‌مراتبی با یک سری از تقسیم‌بندی‌های متوالی انجام می‌شود. روش‌های خوشه‌بندی سلسله‌مراتبی بر دو دسته‌اند: سلسله‌مراتبی تجمعی[1] و سلسله‌مراتبی تقسیمی[2] [۲۵ و ۲۶]. روشی که در این مقاله از آن استفاده شده، خوشه‌بندی سلسله‌مراتبی تجمعی است. در مکانیزم اجرایی این روش، ابتدا با استفاده از یک معیار، فواصل دو خوشه تعریف می‌شود و سپس روش مناسب برای تشکیل خوشه‌ها و پیوند آن‌ها با یکدیگر انتخاب می‌گردد. د نهایت نیز تعداد خوشه‌های مناسب برای داده‌ها تعیین شده و خوشه‌بندی انجام می‌گیرد. خوشه‌بندی سلسله‌مراتبی با جداسازی هر مورد در یک خوشهٔ جداگانه شروع می‌شود. در هر مرحله از تحلیل، جداسازی موارد تا جایی انجام می‌گیرد که شبیه‌ترین دو خوشه در هم ادغام شوند و درنهایت نیز تمامی موارد در یک درخت طبقه‌بندی کامل ادغام گردند. معیاری که خوشه‌بندی براساس آن انجام می‌پذیرد، فاصله است. مواردی که نزدیک یکدیگرند، در یک خوشه ادغام شده و مواردی که نسبت‌به یکدیگر فاصلهٔ بیشتری دارند، در خوشه‌های متفاوت قرار می‌گیرند [۲۷]. شکل (۱) الگوریتم خوشه‌بندی سلسله‌مراتبی تجمعی را بیان می‌کند.

1. Agglomerative
2. Divisive

The Agglomerative Hierarchical Clustering
1: Initial Clustering, Put each data in a cluster
2: for each cluster do
3: Compute similarity measure for all pairs
4: Merge clusters C_i and C_j, in which Similarity(C_i,C_j) is max
5: end for
6: Determine # of clusters

شکل (۱): الگوریتم خوشه‌بندی سلسله‌مراتبی تجمعی

۲.۳. معیار فاصله و چگالی

معیار شباهت یا فاصله بین دو داده، یک مسئلهٔ چالش‌برانگیز در داده‌کاوی و تکنولوژی اطلاعات است. اندازه‌گیری شباهت می‌تواند به دو بخش تقسیم شود: اندازه‌گیری شباهت برای داده‌های دسته‌ای و داده‌های عددی. اندازه‌گیری شباهت برای داده‌های دسته‌ای به‌راحتی داده‌های عددی نیست. انواع فاصله نظیر فاصلهٔ منهتن[3] و فاصلهٔ اقلیدسی[4] دو نوع از مهم‌ترین معیارهای فاصله برای داده‌های عددی هستند. شباهت یا فاصله برای داده‌های دسته‌ای، به‌راحتی داده‌های عددی محاسبه نمی‌شود و این مسئله یک موضوع چالش‌برانگیز است. این حقیقت که مقادیر مختلف در داده‌های دسته‌ای هیچ ترتیب خاصی ندارند، موجب می‌شود که مقایسهٔ بین داده‌های دسته‌ای ممکن نباشد. به‌علاوه معیار شباهت مورداستفاده ممکن است به دامنهٔ ویژه‌ای وابسته باشد. معیارهای مختلفی برای داده‌های دسته‌ای به‌کار می‌رود، اما بیشتر این معیارها در به‌کارگیری سایر ویژگی‌ها در مجموعه داده با شکست روبه‌رو می‌شوند. یک راه حل برای این مشکل ساخت یک منبع مشترک برای اندازه‌گیری شباهت برای همهٔ مفاهیم مشترک است. اندازه‌گیری شباهت برای یک دامنه نسبت‌به دیگری متفاوت است. بنابراین برای تعیین شباهت داده‌ها باید درک صحیحی از دامنهٔ داده‌ها به‌دست آورد. بنابراین اندازه‌گیری شباهت توسط یک خبرهٔ دامنه که به‌خوبی مفهوم دامنه را درک کرده، تعریف می‌شود. در بیشتر کاربردها کارشناس دامنه موجود نیست و کاربر ارتباط درونی بین داده‌ها را به‌خوبی درک نمی‌کند تا براین‌اساس شباهت و فاصلهٔ بین داده‌ها را محاسبه کند [۲۸ و ۲۹].

3. Manhattan distance
4. Euclidean distance

ساده‌ترین راه برای پیدا کردن شباهت بین دو صفت دسته‌ای بر این اساس است کـه دو داده اگـر یکسـان باشـند، شباهتشـان ۱ است و در غیر این صورت، شباهتشان صفر است. بـرای داده‌هـای دسته‌ای چندمتغیره، شباهت بین آن‌ها برابر تعـداد صـفات یکسـان میان آن‌هاست. این روش overlap نام دارد. میزان تشابه هر صفت برای اندازه‌گیری overlap بین ۰ و ۱ است؛ مقـدار صـفر بـه ایـن معناست که دو داده مشابه نیستند و مقدار ۱ به این معناست که دو داده کاملاً یکسان‌اند. عیب بـزرگ روش overlap ایـن اسـت کـه فرقی بین مقادیر مختلف یک صـفت قائـل نیسـت. درواقـع همـهٔ صفات یکسان و غیریکسان میان دو داده، رفتار برابـر دارنـد [۲۹]. رابطهٔ (۱) این روش را نشان می‌دهد:

$$S_k(X_k,Y_k)=\begin{cases}1 & if \quad X_k=Y_k\\ 0 & otherwise\end{cases} \qquad (۱)$$

کـه X_k و Y_k اشـیاء داده در مجموعـه عناصـر داده هسـتند و $S_k(X_k,Y_k)$ تابع شباهت بین این دو شیء را محاسبه می‌کند.

معیار فاصلهٔ دیگر در مورد داده‌های دسته‌ای، معیـار همگرایـی Jaccard است که بـا الهـام از معیـار overlap و بهبـود آن توسـعه یافته است. معیار همگرایی Jaccard تشابه بین داده‌ها را به‌صـورت رابطهٔ (۲) بررسی می‌کند:

$$Sim(X_k,Y_k)=\frac{|X_k\bigcap Y_k|}{|X_k\bigcup Y_k|} \qquad (۲)$$

که $|X_k|$ تعداد عناصر X_k است و X_k یک شیء داده در مجموعـه عناصـر داده هسـتند. رابطـهٔ (۲) نزدیکـی بـین عناصـر دو داده را محاسبه می‌کند و عناصری که در هر دو داده X_k,Y_k مشترک‌انـد، به‌صورت $|X_k\bigcap Y_k|$ محاسبه می‌شوند و بعـد از نرمـال‌سـازی بـا پارامتر θ مقایسه می‌شوند [۳۰ و ۳۱].

معیار دیگری که برای تعیین تشـابه داده‌هـای دسـته‌ای بـه‌کـار می‌رود، معیار چگالی است. معیار چگالی فرکانس تکرار صفات را برای هر داده محاسبه می‌کند. برای محاسـبهٔ چگـالی از رابطـهٔ (۳) استفاده می‌کنیم.

$$Dens(x)=-\frac{1}{|U|}\sum_{y\in U}d(x,y) \qquad (۳)$$

کـه $d(x,y)$ فرکـانس تکـرار دو داده را محاسـبه مـی‌کنـد و درنهایت، برای نرمال‌سازی چگالی به‌دست‌آمده، آن را به $|U|$ کـه برابر تعداد داده‌ها در مجموعه داده است، تقسیم می‌کند [۳۲].

در این مقالـه، ترکیبـی از دو معیـار چگـالی و فاصـله جهـت بهره‌وری از محاسن هر دو معیار استفاده شده اسـت. در ادامـه، دو معیار فاصلهٔ Jaccard و چگالی مبتنی بر فرکـانس تکـرار صـفات، با یکدیگر ترکیب شده‌اند.

۳.۳. روش ترکیبی

هریک از دو معیار فاصله و چگالی، دارای مزایا و معایـب مربـوط به خود هستند. برای محاسبهٔ شباهت دو داده، عـلاوه بـر چگـالی، معیار فاصله را هم در نظر می‌گیریم. اگر تنهـا معیـار فاصـله را در نظر بگیریم، ممکن اسـت کـه داده‌هـای پـرت[1] در تصـمیم‌گیـری کیفیت خوشه‌بندی ما دخالت کند و اگر تنهـا معیـار چگـالی را در نظر بگیریم، بیشتر خوشـه‌هـا در یـک ناحیـه از مجموعـه داده‌هـا انتخاب می‌شوند. بنابراین برای اجتناب از بروز این‌گونه مشـکلات در اینجا معیار چگالی با معیار فاصله برای پیداکردن خوشـه‌هـا بـا هم ادغام می‌شوند [۳۲]. برای ترکیب معیارهای فاصله و چگـالی، دو ترکیب بیان شده است که در هر دو، از یـک فـاکتور وزن‌دهـی به این دو معیار به نام η استفاده می‌شود:

- ترکیب خطی $\eta\ Dens + (1-\eta)d$ که در آن $\eta\in[0,1]$ است.
- ترکیب غیرخطی $Dens^{\eta}\times d^{(1-\eta)}$ که در آن $\eta\in[0,1]$ است.

در اینجا به سه دلیل، از ترکیب خطی استفاده می‌کنیم:

- ترکیب غیرخطی ارائه شده با Logگیری، به رابطهٔ خطی تبـدیل می‌شود.
- تر کیب غیرخطی به مقـدار فـاکتور وزن‌دهـی بسـیار حسـاس است.
- در مطالعات انجام‌شده در این زمینـه، اغلـب از ترکیـب خطـی استفاده می‌شود.

شکل (۲) الگوریتم خوشه‌بندی سلسله‌مراتبی تجمیعی را بیـان می‌کند.

این ترکیب در تحقیقات [۵ و ۳۲] برای یافتن مراکز اولیه خوشه‌ها استفاده شده است و در این مقالـه، بـرای ترکیـب دو معیار خوشه‌بندی استفاده مـی‌شـود. در بخـش بعـد، ملاحظـه می‌شود که ترکیب دو معیار چگالی و فاصـله بـه‌عنـوان معیـار

1. Outliers

شباهت در خوشه‌بندی سلسله‌مراتبی، به نتایج بهتری نسبت‌به زمانی که تنها از یک معیار استفاده می‌شد، منجر می‌شود.

The Agglomerative Hierarchical Algorithm.
Input: Data set D, Number of Clusters k, Dimensions d:
Output: results.
Begin
Initial Clustering, Put each data in a cluster
for each cluster do
Find two cluster that sim(C_i , C_j)=
$\max_{1\le i,j\le d}(\eta * \text{Jaccard\_sim}(x_i, x_j) +$
$(1-\eta) * \text{Overlay\_sim}(x_i, x_j))$
Merge clusters C_i and C_j, in which Similarity(C_i,C_j) is max
Compute Center new cluster
end for
Determine # of clusters
End

شکل (۲): الگوریتم خوشه‌بندی سلسله‌مراتبی تجمیعی

۴.۳. نتایج پژوهش

در این بخش، ابتدا خوشه‌بندی انجام‌شده برای الگوریتم سلسله‌مراتبی با ترکیب معیارها (فاصله و چگالی) بررسی و سپس کارایی این الگوریتم با حالت بدون ترکیب هریک از معیارهای فاصله و چگالی با هم مقایسه می‌شود. این پیاده‌سازی روی دو مجموعه دادهٔ soybean data و zoo data انجام شده که شرح آن‌ها به‌صورت زیر است [۳۳]:

soybean data: این مجموعه داده شامل ۴۷ رکورد است که هرکدام توسط ۳۶ صفت توصیف می‌شوند. و هر رکورد با چهار ویژگی برچسب زده شدهٔ Diaporthe Stem Canker, Phytophthora Rot Charcoal Rot, Rhizoctonia Root Rot, به غیر از Phytophthora Rot شامل ۱۷ رکورد است.

zoo data: این مجموعه داده شامل ۱۰۱ نمونه است که هرکدام توسط ۱۷ صفت باینری توصیف می‌شوند. داده‌ها در ۷ خوشه طبقه‌بندی می‌شوند.

برای ارزیابی کارایی یک الگوریتم خوشه‌بندی، فاکتورهایی بدین صورت ارائه شده است:

Accuracy: با فرض اینکه $C=\{C_1,...,C_k\}$ مجموعه خوشه‌ها و a_i برابر تعداد داده‌هایی است که در خوشه صحیح خود (C_i) قرارگرفته‌اند و مجموعه داده دارای n شیء و K تعداد خوشه‌ها باشد، میزان Accuracy با رابطهٔ (۴) اندازه‌گیری می‌شود [۳۴]:

$$AC=\frac{\sum_{i=1}^{K} a_i}{n} \qquad (۴)$$

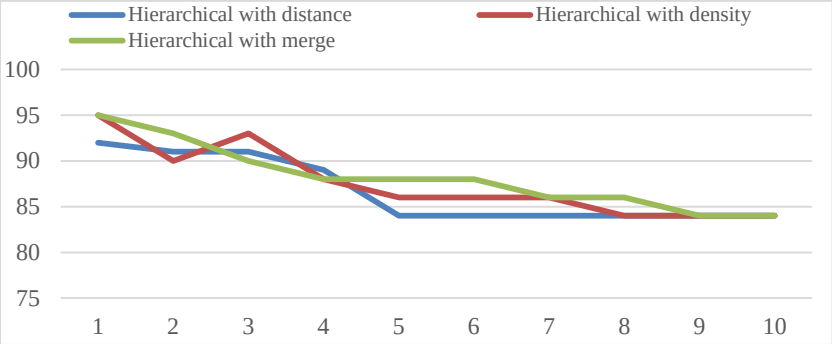
Precision: با فرض اینکه $C=\{C_1,...,C_k\}$ مجموعه خوشه‌ها و a_i برابر تعداد داده‌هایی است که در خوشهٔ صحیح خود (C_i) قرار گرفته و b_i برابر تعداد داده‌هایی است که در خوشهٔ نادرست C_i قرارگرفته‌اند و مجموعه داده دارای n شیء و K تعداد خوشه‌هاست، میزان Precision با رابطهٔ (۵) اندازه‌گیری می‌شود [۳۴]:

$$PR=\frac{\sum_{i=1}^{K}\left(\frac{a_i}{a_i+b_i}\right)}{K} \qquad (۵)$$

جدول (۱) مقایسهٔ بین الگوریتم‌های سلسله‌مراتبی با سه معیار مختلف روی مجموعه داده soybean data را با تعداد داده ورودی متفاوت نشان می‌دهد. جدول (۲) مقایسهٔ بین الگوریتم‌های سلسله‌مراتبی با سه معیار مختلف روی مجموعه دادهٔ zoo data را با تعداد داده ورودی متفاوت نشان می‌دهد. اعداد جدول نتایج حاصل از خوشه‌بندی الگوریتم‌ها با استفاده از معیار ارزیابی Accuracy را نشان می‌دهد و عدد بزرگ‌تر نشان‌دهندهٔ خوشه‌بندی بهتر است. همان‌طور که در جدول (۱) و (۲) و نمودارهای (۱) و (۲) نشان داده شده است، با افزایش تعداد داده‌ها، کارایی خوشه‌بندی افت پیدا می‌کند؛ این امر نشان‌دهندهٔ تعداد بهینهٔ خوشه‌ها برای این نوع داده‌هاست. مثلاً برای دادهٔ soybean تعداد بهینهٔ خوشه‌ها ۱ یا ۲ است. نتایج نشان می‌دهد که روش پیشنهادی در مقایسه با دو روش دیگر رشد چشمگیری دارد.

جدول (۱): نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار Accuracy برحسب درصد

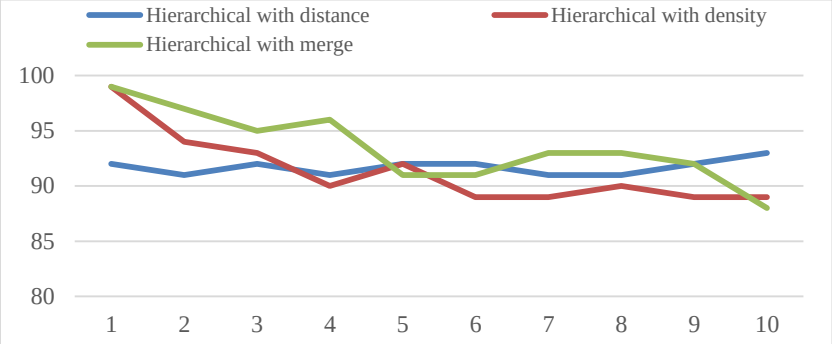
	تعداد خوشه									
	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical with distance	۹۲	۹۱	۹۱	۸۹	۸۴	۸۴	۸۴	۸۴	۸۴	۸۴
Hierarchical with density	۹۵	۹۰	۹۳	۸۸	۸۶	۸۶	۸۶	۸۴	۸۴	۸۴
Hierarchical with merge	**۹۵**	**۹۳**	**۹۰**	**۸۸**	**۸۸**	**۸۸**	**۸۶**	**۸۶**	**۸۴**	**۸۴**



نمودار (۱): نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار Accuracy برحسب درصد

جدول (۲): نتایج خوشه‌بندی داده‌های zoo با استفاده از معیار Accuracy برحسب درصد

تعداد خوشه										
	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical with distance	۹۲	۹۱	۹۲	۹۱	۹۲	۹۲	۹۱	۹۱	۹۲	۹۳
Hierarchical with density	۹۹	۹۴	۹۳	۹۰	۹۲	۸۹	۸۹	۹۰	۸۹	۸۹
Hierarchical with merge	**۹۹**	**۹۷**	**۹۵**	**۹۶**	**۹۱**	**۹۱**	**۹۳**	**۹۳**	**۹۲**	**۸۸**

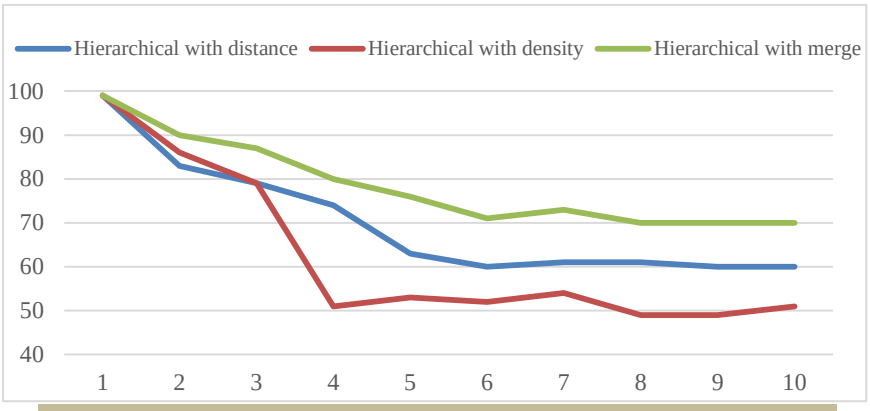


نمودار (۲): نتایج خوشه‌بندی داده‌های zoo با استفاده از معیار Accuracy برحسب درصد

جدول (۳) و نمودار (۳) مقایسهٔ بین الگوریتم‌های سلسله‌مراتبی با سه معیار مختلف با معیار Precision روی مجموعه داده soybean data با تعداد خوشه‌های متفاوت را نشان می‌دهد. جدول (۴) و نمودار (۴) همین مقایسه را روی مجموعه دادهٔ zoo data با تعداد خوشه‌های متفاوت نشان می‌دهد. اعداد جدول نتایج حاصل از خوشه‌بندی الگوریتم‌ها با استفاده از معیار ارزیابی Precision را نشان می‌دهد و عدد بزرگ‌تر نشان‌دهندهٔ خوشه‌بندی بهتر است. همان‌طور که در جدول (۳) و (۴) و نمودار (۳) و (۴) نشان داده شده است، با افزایش تعداد خوشه‌ها، کارایی خوشه‌بندی افت پیدا می‌کند. همان‌طور که در جدول (۳) و (۴) دیده می‌شود، میزان Precision خوشه‌بندی الگوریتم پیشنهادی نسبت‌به روش دیگر رشد چشمگیری داشته است.

جدول (۳): نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار Precision برحسب درصد

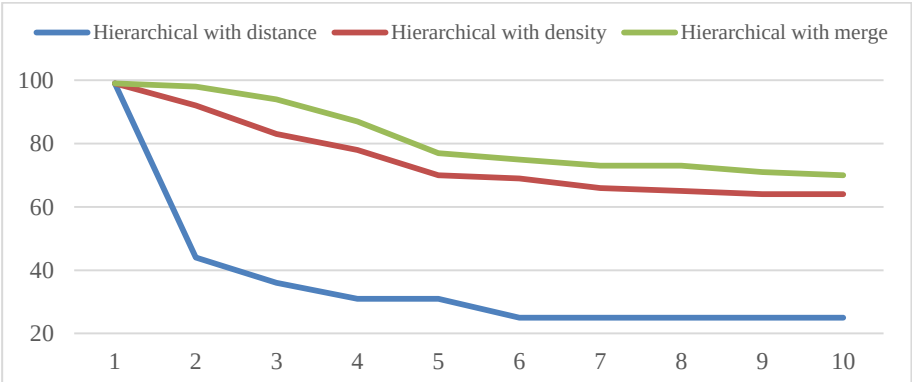
تعداد خوشه										
	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical with distance	۹۹	۸۳	۷۹	۷۴	۶۳	۶۰	۶۱	۶۱	۶۰	۶۰
Hierarchical with density	۹۹	۸۶	۷۹	۵۱	۵۳	۵۲	۵۴	۴۹	۴۹	۵۱
Hierarchical with merge	**۹۹**	**۹۰**	**۸۷**	**۸۰**	**۷۶**	**۷۱**	**۷۳**	**۷۰**	**۷۰**	**۷۰**



نمودار (۳): نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار Precision برحسب درصد

جدول (۴): نتایج خوشه‌بندی داده‌های zoo با استفاده از معیار Precision برحسب درصد

تعداد خوشه										
	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical with distance	۹۹	۴۴	۳۶	۳۱	۳۱	۲۵	۲۵	۲۵	۲۵	۲۵
Hierarchical with density	۹۹	۹۲	۸۳	۷۸	۷۰	۶۹	۶۶	۶۵	۶۴	۶۴
Hierarchical with merge	**۹۹**	**۹۸**	**۹۴**	**۸۷**	**۷۷**	**۷۵**	**۷۳**	**۷۳**	**۷۱**	**۷۰**



نمودار (۴): نتایج خوشه‌بندی داده‌های zoo با استفاده از معیار Precision برحسب درصد

در نتایج به‌دست‌آمده در نتایج خوشه‌بندی، ضریب وزنی معیارهای شباهت ثابت در نظر گرفته شده بود. جدول (۵) ارزیابی کیفیت خوشه‌بندی ترکیبی را با ضریب وزنی متفاوت در فرایند خوشه‌بندی نشان می‌دهد. همان‌طور که در جدول (۵) نشان داده شده است، نتایج حاصل نشان می دهد که هرچه ضریب وزنی برای معیار Jaccard بیشتر باشد، نتایج بهتری به‌دست می‌آید.

جدول (۵): نتایج خوشه‌بندی با ضرایب متفاوت برای ترکیب معیارهای شباهت

ضرایب متفاوت برای معیار Jaccard								
معیار ارزیابی	۰,۹	۰,۸	۰,۷	۰,۶	۰,۵	۰,۴	۰,۳	۰,۲
Accuracy	۹۲٪	٪۹۰	۹۰٪	۸۹٪	۸۸٪	۸۶٪	۸۴٪	۸۴٪

جدول (۶) مقایسه‌ای بین الگوریتم پیشنهادی در این مقاله با دو الگوریتم بیان‌شده در پیشینهٔ پژوهش (الگوریتم Squeezer و الگوریتم fuzzy k-modes) را روی سه مجموعه دادهٔ ذکرشده، با دو معیار Accuracy و Precision با تعداد خوشهٔ ثابت ۴ بیان می‌کند. همان‌طور که در جدول (۶) نشان داده شده است، میزان دقت الگوریتم پیشنهادی در مقایسه با

دو الگوریتم Squeezer و fuzzy k-modes رشد قابل‌مقایسه‌ای را به همراه دارد. این امر به‌دلیل ترکیب دو معیار شباهت است؛ زیرا ترکیب این دو معیار موجب از بین بردن نقاط ضعف هریک از آن‌ها و نیز موجب افزایش درصد دقت خوشه‌بندی می‌شود. به‌کارگیری همزمان دو معیار شباهت سبب می‌شود که مشکل تجمع خوشه‌ها در یک نقطه از مجموعهٔ داده از بین برود و تجمع خوشه‌ها حالت متوازن‌تری به خود بگیرد.

جدول (۶): نتایج خوشه‌بندی سه الگوریتم روی سه مجموعه داده

معیار ارزیابی	تعداد خوشه‌ها	Squeezer algorithm	Weighting K-modes	Hierarchical with merge
Accuracy	Zoo data	۹۰٪	۸۲٪	**۹۱٪**
	Soy data	۸۶٪	۹۰٪	**۸۹٪**
	Mushroom date	۷۸٪	۷۹٪	**۸۳٪**
Precision	Zoo data	۸۰٪	۸۵٪	**۸۸٪**
	Soy data	۸۶٪	۸۲٪	**۹۰٪**
	Mushroom date	۷۳٪	۷۸٪	**۸۳٪**

۴. نتیجه‌گیری

هدف از آنالیز خوشه‌بندی، تقسیم داده‌ها داخل خوشه‌های بامعنی است. در مجموعه داده‌های دسته‌ای، به‌دلیل ماهیت درونی آن‌ها، خوشه‌بندی آن‌ها با خوشه‌بندی داده‌های عددی متفاوت است. چالش‌هایی در زمینهٔ خوشه‌بندی داده‌های دسته‌ای مطرح است که شامل عدم امکان تفکیک‌پذیری روی دامنه و عدم امکان اندازه‌گیری شباهت برای داده‌های دسته‌ای است. مهم‌ترین ویژگی داده‌های دسته‌ای این است که معیار فاصلهٔ طبیعی دربارهٔ آن‌ها کاربردی ندارد. تغییر شکل داده‌های دسته‌ای به داده‌های عددی، معنای داده را از بین می‌برد. در این مقاله، یک الگوریتم خوشه‌بندی سلسله‌مراتبی با مفهوم ترکیب معیارهای شباهت برای داده‌های دسته‌ای ارائه شد. در اینجا دو معیار فاصله Jaccard و چگالی مبتنی بر فرکانس تکرار صفات با یکدیگر ترکیب شدند. هریک از این معیارها در الگوریتم‌های جداگانه‌ای برای خوشه‌بندی داده‌های دسته‌ای استفاده شده‌اند. نتایج حاصل از خوشه‌بندی این الگوریتم نشان می‌دهد که خوشه‌های ایجادشده توسط ترکیب معیارها نتایج بهتری را نسبت‌به استفاده از هر معیار به‌طور جداگانه برای خوشه‌بندی به‌دست می‌دهد؛ این امر نشان‌دهندهٔ تایر کمتر داده‌های پرت و دارای نویز در فرایند خوشه‌بندی است.

مراجع

[1] Hosseininezhad, F., Salajegheh, A., *Study and Comparison of Partitioning Clustering Algorithms*, Iranian Journal of Medical Informatics, Vol. 2, No. 1, pp. 32-44, 2012.

[2] Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J., Models and issues in data stream systems, *Proceedings of Principles of DataBase Systems*, pp. 112-118, 2002.

[3] Cao, F., Liang, J., Li, D., Zhao, X., A weighting k-modes algorithm for subspace clustering of categorical data, *International Journal of Neurocomputing*. Vol. 108, No. 13, pp. 23–30, 2013.

[4] Berry, M., Linoff, G., Mastering Data Mining: The Art and Science of Customer Relationship Management, New York, NY, USA: John Wiley and Sons, pp. 987-998, 1999.

[5] Bai, L., Liang, J., Dang, C., An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. *Knowledge-Based Systems*, Vol. 24, No. 6, pp. 785–795, 2011.

[6] Wiederhold, G., Advances in Knowledge Discovery in Databases, California: *Proceedings of National Conference on Artificial Intelligence*, Vol. 32, No. 2, pp. 322-330, 1996.

[7] Pendharkar, P., Managing Data Mining Technologies in Organizations: Techniques and

Applications. Idea Group Publishing. pp. 123-154, 2003.

[8] Dubes, R., Jain, K., Algorithms for Clustering Data. Englewood Cliffs: NJ: Prentice- Hall. pp. 432-453, 1988.

[9] Parvin, H., Aliizadeh, H., Minaei-Bidgoli, B., Analoui, M., CCHR: Combination of Classifiers using Heuristic Retraining. *Proceedings of International Conference on Networked Computing and advanced Information Management*. pp. 231-240, 2008.

[10] Yun, C., Chuang, K., Chen, M., Adherence clustering: an efficient method for mining market-basket clusters, *International Journal of Information Systems*, Vol. 31, No. 2, pp. 170-186, 2006.

[11] Mesakar , S., Chaudhari, S., Review Paper On Data Clustering Of Categorical Data, *International Journal of Engineering Research & Technology (IJERT)*, Vol. 1, No. 10, pp. 1-18, 2012.

[12] Chen, H., Chuang, K., Chen, M., On Data Labeling for Clustering Categorical Data, *IEEE Transaction on Knowledge and Data Engineering*, Vol. 20, No. 11, pp. 1458-1472, 2008.

[13] Li, Y., Li, D., Wang, S., Zhai, Y., Incremental entropy-based clustering on categorical data streams with concept drift, *International Journal of Knowledge-Based Systems*, Vol. 59, No. 12, pp. 33–47, 2014.

[14] Han, J., Kamber, M., Data Mining Concepts And Techniques, Harcourt India Private Limited, pp. 398-450, 2010.

[15] Kim, M., Ramakrishna, S., Projected clustering for categorical datasets, *Pattern Recognition Letters*, Vol. 27, No. 12, pp. 1405–1417, 2006.

[16] Zhexue, H., (ReviewPaper) Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and Knowledge Discovery*, Vol. 2, No. 3, pp. 283-304, 1998.

[17] Huang, Z., A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining, *Proceedings of ACM Sigmod on Research Issues on data Mining and knowledge Discovery*, pp. 56-70, 1997.

[18] Bai, L., Liang, J., Dang, C., Cao, F., A novel fuzzy clustering algorithm with between-cluster information for categorical data, *International Journal of Fuzzy Sets and Systems*, Vol. 215, No. 1, pp. 55–73, 2013.

[19] Saha, A. Das, S., Categorical fuzzy k-modes clustering with automated feature weight learning. *Neurocomputing*, Vol. 166, No. 1, pp. 422–435, 2015.

[20] Sankar Sangam, R., Om, H., The k-modes algorithm with entropy based similarity coefficient, *International Journal of Procedia Computer Science*, Vol. 50, No. 1, pp. 93-98, 2015.

[21] Ji, J., Pang, W., Zhou, C., Han, X., Wang, Z., A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data, *International Journal of Knowledge-Based Systems*, Vol. 30, No. 12, pp. 129–135, 2012.

[22] Zengyou, H., Xiaofei, X., Shengchun, D., Squeezer: An Efficient Algorithm for Clustering Categorical Data, *International Journal of Computer Science and Technology*, Vol. 17, No. 5, pp. 611-624, 2002.

[23] Barbar'a, D., Couto, J., Li, Y., COOLCAT: An entropy-based algorithm for categorical clustering, *Proceedings of International Conference ACM CIKM on Information and Knowledge Management*, pp. 590–599, 2002.

[24] Yang, Y., Guan, X., You, J., CLOPE: a fast and effective clustering algorithm for transactional data, *Proceedings of international conference ACM Sigkdd on Knowledge discovery and data mining*, pp. 682-687, 2002.

[25] Johnson, A., Wichern, D., Applied Multivariate Statistical Analysis, New Jersey: Prentice Hall, Englewood Cliffs, pp. 34-67, 1988.

[26] Kaufman, L., J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, New York: Wiley, pp. 324-367, 1990.

[27] Zhang, X., Xu, Z., Hesitant fuzzy agglomerative hierarchical clustering algorithms, *International Journal of Systems Science*, Vol. 46, No. 3, pp. 562-576, 2015.

[28] Cesario, E., Manco, G., Ortale, R., Top-down parameter-free clustering of high-dimensional categorical data, *IEEE Transactions on Knowledge and Data Engineering*., Vol. 19, No. 12, pp. 1607–1624, 2007.

[29] Boriah, S., Chandola, V., Kumar, V., Similarity Measures for Categorical Data: A Comparative Evaluation, *Proceedings of International Conference SIAM on Data Mining*, pp. 212-220, 2008.

[30] Kou, G., Peng, Y., Wang, G., Evaluation of clustering algorithms for financial risk analysis using MCDM methods, *International Journal of Information Sciences*, Vol. 275, No. 1, pp. 1–12, 2014.

[31] Guha, S., Rastogi, R., Shim, K., ROCK: a robust clustering algorithm for categorical attributes,

International Journal of Information Systems, Vol. 25, No. 5, pp. 345-366, 2000.

[32] Bai, L., Liang, J., Dang, C., A cluster centers initialization method for clustering categorical data, *International Journal of Expert Systems with Applications*, Vol. 39, No. 8, pp. 8022–8029, 2012.

[33] "http://archive.ics.uci.edu/ml/datasets.html.," [Online].

[34] Bai, L., Liang, J., The k-modes type clustering plus between-cluster information for categorical data, *International Journal of Neurocomputing*, Vol. 133, No. 20, pp. 111–121, 2014.