مجله علمی-پژوهشی
رایانش نرم و فناوری اطلاعات
جلد۴، شماره۴، سال ۱۳۹۴

یک الگوریتم ترکیبی خوشه‌بندی جدید در رویکرد داده‌های دسته‌ای

مریم نبی‌لو[۱]، نگین دانشپور[۲]
[۱] کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجائی،mnabiloo13@yahoo.com
[۲] استادیار، دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجائی، ndaneshpour@srttu.edu

چکیده - خوشه‌بندی داده‌ها یک ابزار پایه موجود برای درک ساختار مجموعه داده‌ها است. فرایندی که داده‌ها را در گروه‌هایی از اشیاء شبیه به هم قرار می‌دهد خوشه‌بندی نام دارد. خوشه‌بندی یکی از مهم‌ترین مسائل بدون ناظر برای یافتن ساختار در یک مجموعه داده‌های بدون برچسب است. 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 به دلیل ماهیت و کاربرد این داده‌ها نسبت به الگوریتم‌های خوشه‌بندی داده‌های عددی از اهمیت بالایی برخوداراند. 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 می‌کنیم و در انتها، روشی ترکیبی، بر پایه ترکیب دو الگوریتم خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی تفکیکی برای خوشه‌بندی بهتر این نوع داده‌ها ارائه می‌دهیم. آزمایشات نشان می‌دهد که روش ارائه شده در این مقاله نتایج حاصل از خوشه‌بندی را بهبود می‌بخشد.

کلید واژه‌ها- داده‌کاوی، خوشه‌بندی، داده‌های دسته‌ای، الگوریتم خوشه‌بندی k-modes، الگوریتم خوشه‌بندی سلسله مراتبی.

۱. مقدمه

امروزه، داده‌کاوی به عنوان یک ابزار قوی برای تولید اطلاعات و دانش از داده‌های خام، شناخته شده و همچنان با سرعت در حال رشد و تکامل است[1]. خوشه‌بندی[۱] [2, 3] به عنوان مرحله اصلی در هر سیستم داده‌کاوی امروزه کاربرد فراوان پیدا کرده و مورد توجه محققان قرار گرفته است. خوشه‌بندی یک تکنیک مهم برای آنالیز اکتشافی داده‌ها است. خوشه‌بندی مسئله‌ای است که در حوزه‌های زیادی کاربرد دارد. اکثر الگوریتم های خوشه‌بندی امروزه روی داده‌های عددی تمرکز دارند و خوشه‌بندی آن‌ها مبتنی بر معیار فاصله اقلیدسی است. اما داده‌های جهان واقعی و کاربردهای داده‌کاوی اغلب روی داده‌های دسته‌ای[۲] تمرکز دارند [4]. داده‌های دسته‌ای شامل مجموعه صفاتی است که دامنه آن‌ها عددی نیست و هر صفت یک مقدار (از مجموعه مقادیری که هیچ نظم خاصی ندارند) را به خود می‌گیرد. مهم‌ترین ویژگی این نوع داده‌ها این است که معیار فاصله طبیعی در مورد آن‌ها کاربردی ندارد[5]. تبدیل داده‌های

[۱] Clustering

[۲] Categorical data

عددی به داده‌های دسته‌ای اغلب نتایج مفیدی را به همراه ندارد. تغییر شکل داده‌های دسته‌ای به داده‌های عددی معنای داده را از بین می‌برد و زمانی که دامنه صفات دسته‌ای بزرگ می‌شود، ماهیت این نوع داده‌ها را نیز از بین می‌برد. این نوع داده‌ها عبارت‌اند از ویژگی‌های افراد، اشیاء، واحدها و غیره که می‌توانند مقادیر کیفی یا کمی اختیار کنند. به‌طورکلی داده‌ها به دو دسته تقسیم می‌شوند: داده‌های دسته‌ای: مانند رنگ چشم یا جنسیت، داده‌های عددی: مانند طول، قد، وزن و بهره هوشی.

دو دسته‌ی کلی از الگوریتم‌های خوشه‌بندی، الگوریتم‌های سلسله‌مراتبی[3] و الگوریتم‌های تفکیکی[4] هستند. الگوریتم‌های سلسله‌مراتبی خوشه‌ها را به‌تدریج می‌سازند (مانند کریستال‌ها رشد می‌کنند) ولی الگوریتم‌های تفکیکی مستقیماً خوشه‌بندی را انجام می‌دهند. آن‌ها سعی می‌کنند که خوشه‌ها را با جایگذاری مجدد نقطه‌ها بین زیرمجموعه‌ها بسازند [6]. در این مقاله ابتدا به تعریف و تبیین مفهوم خوشه‌بندی، به خصوص خوشه‌بندی داده‌های دسته‌ای پرداخته شده است. در ادامه، روش ترکیبی جدیدی بر پایه ترکیب دو الگوریتم خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی تفکیکی برای خوشه-بندی داده دسته‌ای ارائه شده است. آزمایشات نشان می‌دهد که روش یاد شده موجب بهبود در نتایج خوشه‌بندی داده‌ها می‌گردد. تمام الگوریتم‌های خوشه‌بندی ترکیبی در حوزه داده‌های عددی قرار دارند و از روش‌های ترکیبی در زمینه 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. پیشینه پژوهش

الگوریتم‌های متنوعی برای خوشه‌بندی داده‌ها ارائه شده است. الگوریتم k-means یکی از اولین الگوریتم‌های خوشه‌بندی تفکیکی بود که در کاربردهای داده‌های واقعی به صورت وسیع و مؤثر در خوشه‌بندی مجموعه داده‌ها مورد استفاده قرار گرفت [7]. انواع فازی از الگوریتم k-means نیز ارائه شده است [8]. در مدل فازی هر داده می‌تواند با یک درجه عضویت عضوی از تمام خوشه‌ها باشد [9]. استفاده از انواع الگوریتم‌های k-means فقط محدود به داده‌های عددی است، اما بیشتر داده‌های جهان واقعی از نوع داده‌های دسته‌ای هستند. همان‌طور که پیش‌تر هم بیان شد خوشه‌بندی داده‌های دسته‌ای به دلیل ماهیت و کاربرد این داده‌ها نسبت به خوشه‌بندی داده‌های عددی از اهمیت بالاتری برخوردار است. به منظور خوشه‌بندی داده‌های دسته‌ای الگوریتم k-means به الگوریتم k-modes [10] گسترش یافت. در این الگوریتم محدودیت در مورد داده‌های عددی از بین رفت و این الگوریتم قادر بود که به‌طور مؤثر روی داده‌های دسته‌ای اجرا شود. از زمان اولین انتشار الگوریتم k-modes این الگوریتم به یک تکنیک کارا برای حل مسائل خوشه‌بندی داده‌های دسته‌ای بدل شد. در این الگوریتم معیار شباهت جدیدی برای داده‌های دسته‌ای معرفی شد. معیار شباهت بین دو داده دسته‌ای برابر با تعداد مقادیر صفات مشابه میان آن دو داده است[11] . الگوریتم خوشه‌بندی k-modes میانگین خوشه‌ها را با مفهوم mode جایگزین کرد و از متدهای مبتنی بر فرکانس برای به روز رسانی 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 می شود تا پایداری راه حل خوشه‌بندی بدست آید. اما این الگوریتم همانند الگوریتم k-means توانایی مدیریت داده‌های پرت و نویز را در فرایند خوشه‌بندی

[3] Hierarchiacal algorithm

[4] Partitional algorithm

ندارد و همچنین جواب نهایی به انتخاب مراکز اولیه خو شه‌ها وابسته است. علاوه بر این، الگوریتم fuzzy k-modes نیز ارائه شد که در این الگوریتم داده‌ها می‌توانند با یک درجه عضویت ع ضوی از تمام خو شه‌ها با شند [4]. این امر مهمترین مزیت این الگوریتم ا ست و موجب شده این الگوریتم برای داده‌های دارای نویز مناسب باشد. اما این الگوریتم نیز مانند الگوریتم k-modes توانایی مدیر یت داده های پرت را ندارد. در ادامه ت عدادی از الگوریتم های خوشه‌بندی مختص داده های دسته‌ای را بررسی می‌کنیم:

الگوریتم خوشه‌بندی weighting k-modes [12]، در واقع همان الگوریتم k-modes ا ست با این تفاوت که به هر صفت وزنی اختصاص داده ا ست. دلیل وزن دادن به صفات در این الگوریتم این است که ممکن است بعضی از صفات در مجموعه داده‌ها از اهمیت بی شتری برخوردار با شند. این‌گونه صفات در فرایند خو شه‌بندی نقش بی شتری را ایفا می‌کنند. وزن یک صفت معمولاً عددی بین ۰ تا ۱ است و صفتی که از اهمیت بالاتری برخوردار باشد وزن بیشتری دارد. مکانیزم وزن دهی به صفات در این الگوریتم موجب بهبود کیفیت خوشه‌بندی می‌گردد. ا ما این الگوریتم توا نایی مدیریت داده‌های پرت و دارای نویز را ندارد.

الگوریتم خوشه‌بندی weighting fuzzy k-modes [13]، مدل فازی الگوریتم weighting k-modes است. این الگوریتم در محاسبه شباهت به هر صفت بسته به اهمیت آن یک وزن در بازه [۰،۱]می‌دهد. در این الگوریتم هر داده می‌تواند با یک درجه عضویت عضوی از تمام خوشه‌ها باشد. مزیت این الگوریتم نسبت به 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 همچنان مشکل انتخاب مراکز اولیه خوشه‌ها در این الگوریتم به قوه خود باقی است.

الگوریتم خوشه‌بندی k-modes based on entropy [14]، مشابه الگوریتم خو شه‌بندی k-modes ا ست با این تفاوت که این الگوریتم به جای ا ستفاده از معیار شباهت overlap برای محاسبه شباهت میان اشیاء داده و مراکز خوشه از معیار آنتروپی برای محاسبه شباهت استفاده می‌کند و با مینیمم سازی آنتروپی کلی مورد انتظار، خو شه‌بندی را انجام می‌دهد. این الگوریتم که از جدیدترین نسخه‌های ارائه شده از الگوریتم k-modes است؛ قابلیت بسیار بالایی در فرایند خوشه‌بندی دارد. این الگوریتم همچنین توانسته است مشکل مدیریت داده‌های پرت و دارای نویز را که الگوریتم k-modes در آن ناتوان بود، به خوبی حل کند. اما همچنان مشکل انتخاب مراکز اولیه خوشه‌ها در این الگوریتم نیز به قوه خود باقی است.

در الگوریتم خوشه‌بندی fuzzy k-prototype [15]، k عدد prototype از میان اشیاء داده به وسیله مینیمم‌سازی تابع فاصله انتخاب می‌شود. این الگوریتم مانند الگوریتم خو شه‌بندی fuzzy k- modes عمل می‌کند، با این تفاوت که مفهوم mode با مفهوم prototype جایگزین شده است. انگیزه این الگوریتم بررسی اهمیت صفات مختلف در فرایند خوشه‌بندی است. این الگوریتم هم صفات عددی و هم صفات دسته‌ای را با وزنی متفاوت در فرایند خوشه‌بندی سهیم می‌کند. این امر موجب می‌شود که این الگوریتم برای داده‌ها با صفات ترکیبی (صفات عددی و دسته‌ای) مناسب عمل کند. اما همچنان توانایی مدیریت داده های پرت را ندارد و همچنان مشکل انتخاب مراکز اولیه خو شه‌ها در این الگوریتم به قوه خود باقی است.

الگوریتم خو شه‌بندی ROCK [16] یک الگوریتم خو شه‌بندی سلسله مراتبی تجمیعی است که به بررسی مفاهیم پیوند میان داده‌های دسته‌ای می‌پردازد. الگوریتم خوشه‌بندی ROCK از مفهوم همسایگی نقاط داده استفاده می‌کند. این الگوریتم هر داده را به عنوان یک خو شه جدا در نظر می‌گیرد و سپس خو شه‌ها را بر ح سب نزدیکی بین خو شه‌ها در هم ادغام می‌کند. نزدیکی بین خوشه‌ها به وسیله مجموع تعداد پیوندهای بین جفت داده‌ها بررسی می‌شود. تعداد پیوندها برابر تعداد همسایه‌های مشترک بین دو داده است. این الگوریتم برای خو شه‌بندی هر نوع صفت مناسب ا ست. اما به

دلیل مرتبه زمانی بالایی که دارد قابل استفاده برای خوشه‌بندی مجموعه داده‌های بزرگ نیست.

الگوریتم خوشه‌بندی Squeezer [17] شباهت بین داده‌ها را با یک معیار شباهت ویژه بررسی می‌کند تا داده‌ی مورد نظر را در یکی از خوشه‌های موجود یا در یک خوشه جدید قرار دهد و برای این کار از یک مقدار آستانه استفاده می‌کند. اگر میزان شباهت یک شیء به یک خوشه بیشتر از مقدار آستانه بود آن شیء در آن خوشه قرار می‌گیرد و در غیر این صورت در یک خوشه جدید قرار می‌گیرد. الگوریتم خوشه‌بندی Squeezer برای داده‌های دسته‌ای استفاده می‌شود و با دادن وزن بیشتر به صفات غیرمشابه، خوشه‌بندی را انجام می‌دهد. این الگوریتم تنها با یک‌بار بررسی مجموعه داده‌ها می‌تواند به نتیجه خوشه‌بندی خوبی دست یابد و به منظور افزایش کیفیت خوشه‌ها می‌توان این عمل را چندین بار تکرار نمود. این الگوریتم داده‌های پرت و نویز دار را نیز به خوبی مدیریت می‌کند. عیب این الگوریتم تعیین مقدار آستانه برای مقایسه شباهت است که این مقدار به صورت تجربی بدست می‌آید.

الگوریتم خوشه‌بندی LIMBO [18] یک الگوریتم خوشه‌بندی سلسله مراتبی است که از مفهوم BI[۵] برای محاسبه فاصله بین صفات دسته‌ای استفاده می‌کند. مزیت الگوریتم LIMBO تولید خوشه‌هایی با اندازه‌های مختلف در یک اجرای الگوریتم است. الگوریتم خوشه‌بندی LIMBO مجموعه داده‌های بزرگ را با تولید مدل حافظه محدود، مدیریت می‌کند. این الگوریتم داده‌های پرت و دارای نویز را نیز به خوبی مدیریت می‌کند.

الگوریتم خوشه‌بندی COOLCAT [19] از مفهوم آنتروپی برای گروه‌بندی رکوردها استفاده می‌کند. این الگوریتم، یک الگوریتم افزایشی با هدف مینیمم سازی آنتروپی مورد انتظار برای خوشه‌ها است. با داشتن یک مجموعه از خوشه، الگوریتم COOLCAT نقطه بعدی در مجموعه نقاط داده را با مینیمم سازی آنتروپی کلی مورد انتظار، خوشه‌بندی می‌کند. الگوریتم خوشه‌بندی COOLCAT بدون هیچ پیش پردازشی روی مجموعه داده‌ها، خوشه‌بندی را انجام می‌دهد. بنابراین الگوریتم COOLCAT برای داده‌های جریانی مناسب است. اما توانایی مدیریت داده‌های پرت را ندارد.

الگوریتم خوشه‌بندی CLOPE [20] برای هر خوشه یک نمودار ستونی ترسیم می‌کند و بر مبنای مینیمم سازی نسبت طول به عرض مربوط به آن خوشه، خوشه‌بندی را انجام می‌دهد. الگوریتم خوشه‌بندی CLOPE سریع و مقیاس‌پذیر است و برای جداسازی تراکنش‌های با ابعاد بالا مناسب است. این الگوریتم به ترتیب داده‌های ورودی حساس نیست، اما توانایی مدیریت داده‌های پرت را ندارد.

در این بخش الگوریتم‌های مختلفی که برای خوشه‌بندی داده‌های دسته‌ای معرفی شده بود، بررسی شد. هر یک از این الگوریتم‌ها از یک معیار شباهت برای خوشه‌بندی داده‌ها استفاده می‌کند و برای شرایط خاصی ارائه شده‌است. همان‌طور که بیان شد، الگوریتم k-modes ابتدایی‌ترین الگوریتم در فرایند خوشه‌بندی داده‌های دسته‌ای است این الگوریتم نسبت به سایر الگوریتم‌ها، خوشه‌های متراکم‌تری تولید می‌کند به خصوص هنگامی که خوشه‌ها به صورت کروی باشند؛ و در صورت زیاد بودن تعداد متغیرها، این روش نسبت به سایر روش‌های خوشه‌بندی دارای سرعت بالاتری است. این الگوریتم با تمام مزایایی که در بردارد توانایی مدیریت داده‌های پرت و دارای نویز را ندارد و همچنین جواب نهایی به انتخاب مراکز اولیه خوشه‌ها وابسته است و روالی مشخص برای محاسبه اولیه مراکز خوشه‌ها وجود ندارد. در این مقاله روشی ترکیبی برای خوشه‌بندی بهتر داده‌های دسته‌ای بر مبنای الگوریتم k-modes و بهبود این روش ارائه

[۵] Bioinformatics Clustering

می شود. در روش پیشنهادی، مشکل انتخاب مراکز اولیه که در خانواده الگوریتم‌های خوشه‌بندی تفکیکی وجود داشت، از بین رفت.

۳. تعاریف پایه

هدف از آنالیز خوشه‌بندی، تقسیم مجموعه داده‌ها داخل خوشه‌های بامعنی است [2]. الگوریتم‌های زیادی در این زمینه معرفی شده است. در ادامه به شرح مفهوم خوشه‌بندی و دو روش خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی k-modes می‌پردازیم و سپس در بخش بعد، روش ترکیبی پیشنهادی بیان می‌شود.

۳,۱. خوشه‌بندی

پدیده‌ی خوشه‌بندی که یکی دیگر از اهداف داده‌کاوی می‌باشد، به فرآیند تقسیم مجموعه‌ای از داده‌ها (یا اشیا) به زیرکلاس‌هایی با مفهوم خوشه اتلاق می‌شود. به‌این‌ترتیب یک خوشه، مجموعه داده‌های مشابه می‌باشد که همانند یک گروه واحد رفتار می‌کنند [21]. لازم به ذکر است خوشه‌بندی همان کلاسه‌بندی[6] است، با این تفاوت که کلاس‌ها از پیش تعریف شده و معین نمی‌باشند. در خوشه‌بندی عمل گروه‌بندی داده‌ها بدون نظارت انجام می‌گیرد [22, 23].

فرض کنید که مجموعه داده های X مورد نظر ما از نقاط داده‌ای (یا مترادف آن اشیا، موارد، الگوها، تراکنش‌ها، گروه‌ها یا رکوردها)، در فضای ویژگی A تشکیل شده باشند. یعنی $x_i = (x_{i1},...,x_{id}) \in A$ است که در آن $i = 1,...,N$ و هر جز $x_{il} \in A_{il}$ یک داده ویژگی طبقه‌بندی‌شده‌ی دسته‌ای باشد. این فرمت داده-ویژگی مفهوما متناظر با یک ماتریس $N \times D$ است. هدف خوشه‌بندی پیدا کردن درایه‌هایی در ماتریس فوق می‌باشد، که اجتماع همه‌ی آن‌ها تمام ماتریس باشد و دوبه‌دوی آن‌ها نقطه اشتراکی نداشته باشند که این مفهوم در رابطه ۱ نشان داده شده است:

$$X = C_1 \cup C_2 \cup ... \cup C_k \qquad C_{j1} \cap C_{j2} = \varnothing \qquad (1)$$

برخلاف کلاسه‌بندی که اشیا داده را براساس کلاس‌ها تحلیل می‌کند، خوشه‌بندی اشیا داده‌ها را بدون در نظر گرفتن برچسب‌های کلاس، تحلیل و آنالیز می‌نماید. عمدتاً برچسب کلاس‌ها در داده‌های آموزشی[7] به‌آسانی مشخص نیست زیرا این کلاس‌ها شناخته‌شده نمی‌باشند. خوشه‌بندی گاهی برای تعیین و تولید چنین برچسب‌هایی بکار می‌رود. اشیای خوشه‌بندی شده براساس اصل ماکزیمم شباهت بین اعضای هر کلاس و مینیمم شباهت بین کلاس‌های مختلف گروه‌بندی می‌شوند، یعنی خوشه‌ها به‌گونه‌ای تنظیم می‌شوند که اشیای داخل هر خوشه بیشترین شباهت را با یکدیگر داشته باشند[23] . هر خوشه به عنوان یک کلاس می‌باشد که قوانین از آن مشتق می‌شوند. امروزه، اکثر مفاهیم خوشه‌بندی روی داده‌های عددی تمرکز دارند. برای داده‌های عددی متدهای عمومی برای محاسبه فاصله بین دو نقطه چند متغیره وجود دارد. فاصله منهتن [24] و فاصله اقلیدسی [24] دو تا از وسیع‌ترین متدهایی هستند که برای اندازه‌گیری فاصله دو داده عددی استفاده می‌شود. مشاهدات نشان می‌دهد که این اندازه‌گیری‌ها مستقل از مجموعه داده‌ای است که نقاط داده به آن تعلق دارند. اما داده‌های جهان واقعی و کاربردهای داده‌کاوی اغلب روی داده‌های دسته‌ای تمرکز دارد [4]. داده‌های دسته‌ای شامل مجموعه صفاتی هستند که دامنه آن‌ها عددی نیستند و هر صفت یک مقدار از مجموعه مقادیری که هیچ نظم خاصی ندارند، می‌گیرد.

[6] Classification

[7] Training data

مهم‌ترین ویژگی این داده‌ها این است که معیار فاصله طبیعی در مورد آن‌ها کاربردی ندارد. در مورد داده‌های دسته‌ای، محاسبه شباهت و فاصله برای داده‌های دسته‌ای به راحتی داده‌های عددی نیست. نکته مهم در مورد داده‌های دسته‌ای این است که مقادیر مختلفی که یک صفت دسته‌ای می‌تواند بگیرد ترتیب مشخصی ندارد. بنابراین، مقایسه دو مقدار دسته‌ای مختلف با معیارهای عددی ممکن نیست. در ادامه سه مورد از مهمترین معیارهای شباهت که مربوط به داده‌های دسته‌ای هستند، بررسی می‌شود.

ساده‌ترین راه برای پیدا کردن شباهت بین دو صفت دسته‌ای این است که دو داده اگر یکسان باشند شباهتشان ۱ منظور شود و در غیر این صورت شباهتشان صفر درنظر گرفته شود [24]. برای داده‌های دسته‌ای چند متغیره، شباهت بین آن‌ها برابر با تعداد صفات یکسان میان آن‌ها است. این روش overlap نام دارد [11]. عیب بزرگ روش overlap این است که فرقی بین مقادیر مختلف یک صفت قائل نیست. در واقع همه صفات یکسان و غیر یکسان میان دو داده رفتار برابر دارند [24] . فاصله overlap بر مبنای عدم تطابق خصیصه‌های متناظر، از رابطه ۲ محاسبه می‌شود:

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j) \qquad (۲)$$
$$\delta(x_j,y_j)=\begin{cases}0 & (x_j=y_j)\\ 1 & (x_j\neq y_j)\end{cases}$$

معیار دیگری که برای تعیین تشابه داده‌های دسته‌ای به کار می‌رود، معیار چگالی است [25]. معیار چگالی فرکانس تکرار صفات را برای هر داده محاسبه می‌کند. عیب بزرگ این معیار این است که اگر تنها معیار چگالی را در نظر بگیریم، بیشتر خوشه‌ها در یک ناحیه از مجموعه داده‌ها انتخاب می‌شوند. برای محاسبه چگالی از رابطه ۳ استفاده می‌شود.

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

که d(x,y) فرکانس تکرار دو داده را محاسبه می‌کند و در نهایت برای نرمال سازی چگالی بدست آمده آن را به |U| که برابر تعداد داده‌ها در مجموعه داده است تقسیم می‌کند [25].

اما معمول‌ترین معیاری که برای محاسبه شباهت میان دو داده استفاده می شود معیار شباهت Jaccard است [26]. در این مقاله نیز از همین معیار برای محاسبه شباهت میان اشیاء داده استفاده شده است. این معیار براساس رابطه ۴ محاسبه می‌گردد:

$$Sim(X_k,Y_k)=\frac{|X_k\cap Y_k|}{|X_k\cup Y_k|} \qquad (۴)$$

که $|X_k|$ تعداد عناصر X_k است. رابطه ۲ نزدیکی بین عناصر دو داده را محاسبه می‌کند و عناصری که در هر دو داده X_k,Y_k مشترکند به صورت $|X_k\cap Y_k|$ محاسبه می شوند و بعد از نرمال سازی با پارامتر θ مقایسه می شوند [26].

۳،۲. الگوریتم خوشه‌بندی سلسله مراتبی

روش خوشه‌بندی سلسله‌مراتبی نیز به عنوان پرکاربردترین روش تحلیل خوشه‌ای مطرح است که برای داده‌های کم حجم به کار می‌رود. روش خوشه‌بندی سلسله‌مراتبی در ابتدا با در نظر گرفتن برخی معیارها به تجزیه‌ی سلسله‌مراتبی داده‌ها می‌پردازد و سپس با روش‌های اجماع و تقسیم تغییراتی در دسته‌بندی اولیه ایجاد می‌نماید. الگوریتم‌های سلسله‌مراتبی به دو گونه تقسیم می‌شوند: الگوریتم‌های تجمیعی[8] (پایین به بالا) و تقسیمی[9] (بالا به پایین) [27, 28, 23]. در خوشه‌بندی تجمیعی، کار با خوشه‌هایی با یک داده شروع می‌شود (تعداد خوشه‌ها در ابتدا

[8] Agglomorative

[9] Divise

به اندازه‌ی تعداد داده‌های موجود می‌باشد). در هر مرحله دو یا چند خوشه‌ی مناسب با هم ترکیب شده و خوشه‌ی جدیدی را به وجود می‌آورند. در خوشه‌بندی تقسیمی عمل خوشه‌بندی با یک خوشه شروع می‌شود. این خوشه به صورت بازگشتی به دو یا چند خوشه تقسیم می‌گردد و به همین ترتیب عمل خوشه‌بندی ادامه پیدا می‌کند[30 ,29 ,23] . برای هر دو نوع از الگوریتم‌های سلسله مراتبی، نیاز به یک شرط پایانی است. این شرط اغلب رسیدن به k خوشه‌ی مورد نظر می‌باشد. شکل۱ الگوریتم خوشه‌بندی سلسله‌مراتبی تجمیعی را بیان می‌کند.

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)= $\max_{1\le i,j\le d} sim(x_i, x_j)$;
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 ۱: الگوریتم خوشه‌بندی سلسله مراتبی تجمیعی

۳،۳. الگوریتم خوشه‌بندی k-modes

الگوریتم k-means یکی از الگوریتم‌های خوشه‌بندی تفکیکی است که در کاربردهای داده‌های واقعی به صورت وسیع و مؤثر در خوشه‌بندی مجموعه داده‌های بزرگ مورد استفاده قرار می‌گیرد. الگوریتم k-modes تعمیم داده شده الگوریتم k-means [31] برای داده‌های دسته‌ای است. از زمان اولین انتشار الگوریتم k-modes این الگوریتم یک تکنیک کارا برای حل مسائل داده‌های دسته‌ای بوده است. در این الگوریتم محدودیت در مورد داده‌های عددی از بین رفت و این الگوریتم قادر است که به طور مؤثر روی داده های دسته‌ای اجرا شود. ایده اصلی این الگوریتم این است که داده‌های دسته‌ای را بر مبنای مد آن‌ها به چند خوشه از پیش تعیین شده اختصاص می‌دهد [31]. اگر X و Yدو داده دسته‌ای که با m خصیصه تعریف شده‌اند باشند؛ فاصله (عدم شباهت) این دو متغیر بر مبنای عدم تطابق خصیصه‌های متناظر، از رابطه ۱ محا سبه می شود. روند این الگوریتم مانند k-means است. تابع هزینه برای این الگوریتم به صورت رابطه ۵ تعریف می‌شود:

$$E = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} d(x_{ij}, q_{lj}) \qquad (۵)$$

که n تعداد داده‌ها، m ابعاد داده، k تعداد خوشه‌ها و q مد یک خوشه را نشان می‌دهد. مانند الگوریتم k-means الگوریتم k-modes هم مستعد در ایجاد کمینه محلی است که این امر بستگی به انتخاب مدهای اولیه دارد[11 ,5] . شکل۲ الگوریتم خوشه‌بندی k-modes را نشان می‌دهد.

The K-modes Algorithm.
Input: Data set D, Number of Clusters k, Dimensions d:
Output: results.
Begin
Select k initial modes Q = {q_1, q_2, . . . , q_k};
for i = 1 to n do
Find a cluster l that dsim(x_i , q_l)= $\min_{1\le t\le k} dsim(x_i, q_t)$;
Allocate x_i to cluster l;
Update the mode q_l for cluster l;
end for
repeat
for i = 1 to n do
Let l_0 be the index of the cluster to which x_i
belongs;
Find an cluster l_1 that dsim(x_i , q_{l1}) = $\min_{1\le t\le k} dsim(x_i, q_t)$;
if dsim(x_i , q_{l1}) < dsim(x_i , q_{l0}) then
Reallocate x_i to cluster l_1;
Update q_{l0} and q_l1 ;
end if
end for
until No changes in cluster membership
End

شکل ۲: الگوریتم خوشه‌بندی k-modes

۴. روش پیشنهادی

الگوریتم خوشه‌بندی k-modes همان طور که پیش‌تر بیان شد، دارای مزایا و معایب مربوط به خود است. از جمله این معایب این است که این الگوریتم توانایی مدیریت داده‌های پرت و دارای نویز را در فرایند خوشه‌بندی ندارد و همچنین جواب نهایی به انتخاب مراکز خوشه‌های اولیه وابسته است و روالی مشخص برای محاسبه اولیه مراکز خوشه‌ها وجود ندارد. برای حل این مشکلات در این مقاله از یک الگوریتم سلسله-مراتبی با معیار همگرایی Jaccard استفاده می‌نماییم و سپس نتایج بدست آمده را به یک الگوریتم خوشه‌بندی k-modes می‌دهیم تا مجدد عمل خوشه‌بندی روی آن‌ها انجام شود. این امر سبب می‌شود که داده‌های پرت و دارای نویز در الگوریتم k-modes مدیریت شود. انتخاب مراکز اولیه خوشه‌ها نیز به وسیله نتایج تولید شده از الگوریتم سلسله مراتبی صورت می‌گیرد. با روش پیشنهادی، برای انتخاب مراکز اولیه خوشه، مشکل انتخاب مراکز اولیه که در خانواده الگوریتم‌های خوشه-بندی تفکیکی وجود داشت، از بین رفت. این دو نوع الگوریتم از متداولترین و رایج‌ترین الگوریتم‌های خوشه‌بندی داده‌ها (داده‌های عددی و داده‌های دسته‌ای) می‌باشند. به منظور تست الگوریتم پیشنهادی، دو مجموعه داده soybean data و mushroom data [32] از مجموعه داده‌های دسته‌ای انتخاب شده و الگوریتم پیشنهادی روی آن‌ها اجرا شده است. در ادامه روند الگوریتم ترکیبی شرح داده می‌شود.

۴،۱. الگوریتم ترکیبی

الگوریتم خوشه‌بندی سلسله‌مراتبی، به عنوان ورودی، داده های دسته‌ای را دریافت می‌کند و خوشه‌بندی سلسله‌مراتبی را به عنوان خروجی تولید می‌کند. در گام اول داده‌ها به یک سری زیرمجموعه‌ها خوشه‌بندی می‌شود. در الگوریتم خوشه‌بندی سلسله‌مراتبی تجمیعی ابتدا هر شیء به عنوان یک خوشه در نظر گرفته می شود. سپس نزدیک‌ترین خوشه‌ها با هم ادغام می‌گردند و این کار تا رسیدن به شرایط پایانی ادامه می‌یابد. الگوریتم k-modes نیز به عنوان ورودی داده‌های دسته‌ای و عددی را دریافت می‌کند و بخش‌بندی مناسب را به عنوان خروجی تولید می‌کند. الگوریتم k-modes در گام دوم اعمال می‌شود تا مجموعه داده‌ها را خوشه‌بندی کند. الگوریتم k-modes به صورت تصادفی یا با روش مخصوصی (به عنوان مثال، قرار دادن داده‌هایی با صفات پرتکرار به عنوان مرکز خوشه) k شیء را از میان مجموعه داده‌ها انتخاب می‌کند (در اینجا از روش تصادفی استفاده شده است). الگوریتم k-modes اشیاء را به صورت مکرر با استفاده از معیار فاصله در نزدیک‌ترین خوشه قرار می‌دهد. سپس مد اشیاء موجود در خوشه را به عنوان مرکز جدید خوشه قرار می‌دهد. این کار تا رسیدن به یک معیار توقف ادامه می یابد. معیار توقف می‌تواند عدم تغییر مراکز اولیه خوشه‌ها یا تعداد تکرار مشخص باشد.

The New Algorithm.
Input: Data set D, Number of Clusters k, Dimensions d:
Output: results.
Begin
  Applying Agglomerative Hierarchical Algorithm
  to group data set into i cluster
  Calculate the centroid of every formed cluster.
  Applying K-modes Algorithm to group data set into k cluster
End

شکل ۳: الگوریتم خوشه‌بندی ترکیبی

الگوریتم k-modes معایب چندی نظیر تأثیر نویز و داده‌های پرت در فرایند خوشه‌بندی را به همراه دارد. بنابراین با این متد دو مرحله‌ای تأثیر نویز و داده‌های پرت در فرایند خوشه‌بندی بسیار کاهش می‌یابد. در متد پیشنهادی ابتدا در الگوریتم سلسله مراتبی مجموع داده‌ها به i زیرمجموعه تقسیم می‌شود و سپس هر زیرمجموعه به عنوان یک ورودی به الگوریتم k-modes داده می‌شود. مراکز زیرمجموعه برای معرفی زیرمجموعه محاسبه می‌گردد. فواید اصلی این متد دو مرحله‌ای این است که اشیاء شبیه به هم داخل یک زیرمجموعه قرار می‌گیرند و این زیرمجموعه‌ها در گام دوم به عنوان ورودی به الگوریتم خوشه‌بندی k-modes داده می‌شود. بنابراین، این متد دو مرحله‌ای باعث می‌شود که نویز و داده های پرت تأثیر کمی روی الگوریتم k-modes در فرایند خوشه‌بندی بگذارد. این دو نوع الگوریتم از متداولترین و موفقترین الگوریتم‌های خوشه‌بندی داده‌ها (داده‌های عددی و داده‌های دسته‌ای) می‌باشند. هر یک از این دو الگوریتم

زمانی که پراکندگی داده‌ها نرمال بوده و داده‌های نویز و پرت در مجموعه داده‌ها بسیار اندک است، بسیار مناسب عمل می‌کنند. از این رو در اینجا با ترکیب هر دو الگوریتم سعی در بهبود نتایج حاصل نمودیم. شکل۳ این خوشه‌بندی را نشان می‌دهد.

۴،۲. تحلیل پیچیدگی زمانی

پیچیدگی زمانی الگوریتم پیشنهادی برابر ترکیب خطی پیچیدگی زمانی دو الگوریتم k-modes و الگوریتم سلسله مراتبی تجمیعی است. پیچیدگی زمانی الگوریتم k-modes برابر O(TKN) است، که T برابر تعداد تکرار الگوریتم، K برابر تعداد خوشه‌های تولید شده و N برابر تعداد اشیاء داده است. پیچیدگی زمانی الگوریتم سلسله مراتبی تجمیعی در بهترین حالت برابر O(N2)، که N برابر تعداد اشیاء داده است. بنابراین، پیچیدگی زمانی الگوریتم ترکیبی برابر O(TKN+N2) است. جدول ۱ مقایسه‌ای بین پیچیدگی زمانی الگوریتم ترکیبی با الگوریتم‌های ارائه شده در پیشینه پژوهش ارائه می‌دهد. همان طور که در جدول ۱ نشان داده شده است پیچیدگی زمانی الگوریتم ترکیبی نسبت به سایر الگوریتم‌ها در سطحی متوسط قرار دارد. این الگوریتم در مقایسه با الگوریتم هایی نظیر ROCK، COOLCAT و LIMBO مرتبه زمانی بهتری دارد اما باز هم در مقایسه با الگوریتم‌هایی نظیر K-modes، CLOPE مرتبه زمانی بالاتری دارد.

۴،۳. نتایج پژوهش

در این بخش به بررسی نتایج حاصل از اجرای الگوریتم ترکیبی برای خوشه‌بندی داده‌های دسته‌ای می‌پردازیم. این پیاده‌سازی روی دو مجموعه داده soybean و zoo انجام شده است که شرح این داده‌ها به صورت زیر است:

soybean data : مجموعه داده soybean شامل ۴۷ رکورد است که هر کدام توسط ۳۶ صفت توصیف می شوند. هر رکورد با چهار ویژگی برچسب زده شده Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, Phytophthora Rot به غیر از Phytophthora Rot که شامل ۱۷ رکورد است[32].

جدول۱: پیچیدگی زمانی الگوریتم‌های خوشه‌بندی

الگوریتم	سال ارائه	پیچیدگی زمانی
Merge algorithm	2015	$O(TKN+N^2)$ T - No. of iteration, K - No. of cluster, N - No .of object
K-modes	1997	O(TKN) T - No. of iteration, K - No. of cluster, N - No .of object
Fuzzy K-modes	2002	O(TKMN) T - No. of iterations, K - No. of clusters, M - No. of attributes , N - No. of objects
Weighting K-modes	2013	O(TKN) T - No. of iteration, K - No. of cluster, N - No .of object
Fuzzy K-prototype	2012	O((T+1)KN) T - No. of iteration, K - No. of cluster, N - No .of object
Squeezer	2002	O(NKPM) N - No .of object, K - No. of cluster, M - No. of attribute, P - Distinct attribute values
ROCK	2012	$O(N^2 \log N)$ N - No .of object
CLOPE	2002	O(NKD) N - No .of object, K - No. of cluster, D – No. of attribute
COOLCAT	2002	$O(N^2KlogN)$ N - No .of object, K - No. of cluster
LIMBO	2004	$O(N^2d^2 \log N)$ N - No .of object, d – No. of attribute

zoo data: مجموعه داده zoo شامل ۱۰۱ نمونه ا ست که هر کدام توسط ۱۷ صفت باینری توصیف می شوند. داده‌ها در ۷ خوشه طبقه‌بندی می‌شوند [32].

برای ارزیابی کارایی یک الگوریتم خوشه‌بندی، فاکتورهایی ارائه شده است که به قرار زیر هستند:

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 رابطه ۶ اندازه‌گیری می‌شود [11]:

$$AC=\frac{\sum_{i=1}^{K} a_i}{n} \qquad (۶)$$

Recall: 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 c_i برابر تعداد داده‌هایی است که به صورت نادرست در خوشه C_i قرار نگرفته‌اند؛ مجموعه داده دارای n شیء و K تعداد خوشه‌ها باشد؛ میزان Recall با رابطه ۷ اندازه‌گیری می‌شود [11]:

$$RE=\frac{\sum_{i=1}^{K}\left(\frac{a_i}{a_i+c_i}\right)}{K} \qquad (۷)$$

Precision: 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 b_i برابر تعداد داده‌هایی است که در خوشه نادرست C_i قرارگرفته‌اند، است. مجموعه داده دارای n شیء و K تعداد خوشه‌ها است. میزان Precision با رابطه ۸ اندازه‌گیری می‌شود [11]:

$$PR=\frac{\sum_{i=1}^{K}\left(\frac{a_i}{a_i+b_i}\right)}{K} \qquad (۸)$$

(ARI) Adjusted randindex: فاکتور ارزیابی ARI نشان دهنده ارتباط بین خوشه‌بندی صحیح و خوشه‌بندی ایجاد شده توسط الگوریتم خوشه‌بندی است. با فرض اینکه T و C به ترتیب خوشه‌بندی صحیح (برچسب خوشه)و خوشه‌بندی ایجاد شده توسط الگوریتم خوشه‌بندی باشند، ARI با رابطه ۹ اندازه‌گیری می‌شود [33]:

$$ARI(T,C)=\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \qquad (۹)$$

که a به معنی تعداد جفت‌های متعلق به خوشه‌های یکسان در دو مجموعه T و C، b به معنی تعداد جفت‌های متعلق به خوشه مشابه در مجموعه T اما متفاوت با خوشه‌های موجود در مجموعه C، c به معنی تعداد جفت‌های متعلق به خوشه مشابه در مجموعه C اما متفاوت با خوشه‌های موجود در مجموعه T، و d به معنی تعداد جفت‌های متعلق به خوشه‌های متفاوت در دو مجموعه T و C است. مقادیر فاکتور ARI بین صفر و یک تغییر می‌کند. مقدار بالاتر نشان می‌دهد که خوشه‌بندی تولید شده به خوشه‌بندی واقعی نزدیک‌تر است [33].

Percentage of Correct Pair (%CP): فاکتور ارز یابی CP نشان دهنده درصد جفت عناصری است که به درستی خوشه‌بندی شده‌اند. CP با رابطه ۱۰ اندازه‌گیری می‌شود [33]:

$$CP=\frac{\text{number of pairs correctly clustered into the same cluster}}{\text{pairs actually in the same cluster}} \qquad (۱۰)$$

Minkowski Score (MS): با فرض اینکه T و S به ترتیب خوشه‌بندی صحیح (برچسب خوشه)و خوشه‌بندی ایجاد شده توسط الگوریتم خوشه‌بندی باشند؛ n11 تعداد جفت‌هایی است که در هر دو مجموعه Tو S قرار دارد و n01 و n10 به ترتیب برابر تعداد جفت‌هایی که فقط در مجموعه S و T قرار دارد؛ MS با رابطه ۱۱ اندازه‌گیری می‌شود [33]:

$$MS=\sqrt{\frac{n_{01}+n_{10}}{n_{11}+n_{10}}} \qquad (۱۱)$$

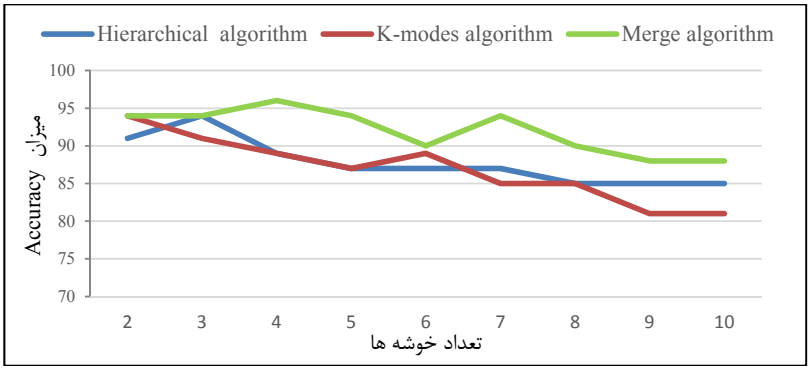
در این مقاله برای ارزیابی کارایی الگوریتم‌های خو شه‌بندی از فاکتور Accuracy، ARI، و Precision که از متداولترین معیارها هستند، استفاده شده است [17].

جدول ۲ و شکل ۴ مقایسه بین الگوریتم ارائه شده در این مقاله با دو الگوریتم خوشه‌بندی سلسله مراتبی و k-modes، با معیار Accuracy روی مجموعه داده soybean data با تعداد خوشه‌های متفاوت را نشان می‌دهد. جدول ۳ و شکل ۵ مقایسه بین الگوریتم ارائه شده در این مقاله با دو الگوریتم خوشه‌بندی سلسله مراتبی و k-modes، روی مجموعه داده zoo data با تعداد خو شه‌های متفاوت را ن شان می‌دهد. اعداد جدول نتایج حا صل از خو شه‌بندی الگوریتم‌ها با استفاده از معیار ارزیابی Accuracy را نشان می‌دهد و عدد بزرگتر نشان دهنده خو شه‌بندی بهتر ا ست. همان‌طور که در جداول ۲ و ۳ نشان داده شده است، با افزایش تعداد خوشه‌ها کارایی خوشه‌بندی افت پیدا می‌کند. همان‌طور که در جداول ۲ و ۳ قابل مشاهده است، میزان Accuracy خوشه‌بندی الگوریتم پیشنهادی نسبت به هر دو الگوریتم خوشه‌بندی سلسله مراتبی و k-modes، رشد قابل ملاحظه‌ای داشته است.

جدول ۴ و شکل ۶ مقایسه بین الگوریتم ارائه شده در این مقاله با دو الگوریتم خوشه‌بندی سلسله مراتبی و k-modes، با معیار ARI روی مجموعه داده soybean data با تعداد خو شه‌های متفاوت را ن شان می‌دهد. جدول ۵ و شکل ۷ همین مقایسه را روی مجموعه داده zoo data با تعداد خوشه‌های متفاوت نشان می‌دهد. اعداد جدول نتایج حاصل

جدول۲: نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار Accuracy برحسب درصد٪

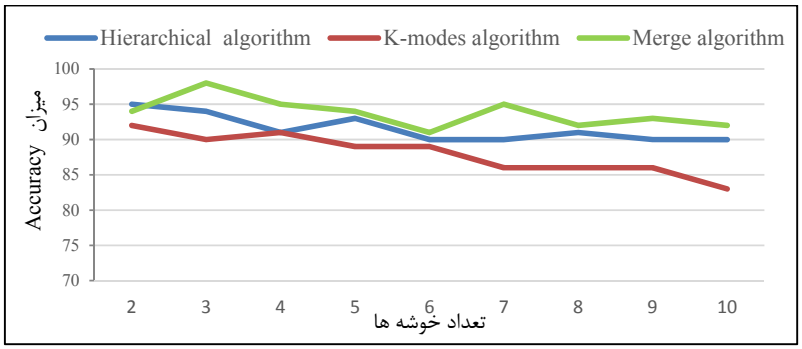
الگوریتم	تعداد خوشه‌ها								
	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical algorithm	۹۱٪	۹۴٪	۸۹٪	۸۷٪	۸۷٪	۸۷٪	۸۵٪	۸۵٪	۸۵٪
K-modes algorithm	۹۴٪	۹۱٪	۸۹٪	۸۷٪	۸۹٪	۸۵٪	۸۵٪	۸۱٪	۸۱٪
Merge algorithm	۹۴٪	۹۴٪	۹۶٪	۹۴٪	۹۰٪	۹۴٪	۹۰٪	۸۸٪	۸۸٪



شکل۴: نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار Accuracy برحسب درصد٪

جدول ۳: نتایج خوشه‌بندی داده‌هایzoo با استفاده از معیار Accuracy برحسب درصد٪

الگوریتم	تعداد خوشه‌ها								
	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical algorithm	۹۵٪	۹۴٪	۹۱٪	۹۳٪	۹۰٪	۹۰٪	۹۱٪	۹۰٪	۹۰٪
K-modes algorithm	۹۲٪	۹۰٪	۹۱٪	۸۹٪	۸۹٪	۸۶٪	۸۶٪	۸۶٪	۸۳٪
Merge algorithm	۹۴٪	۹۸٪	۹۵٪	۹۴٪	۹۱٪	۹۵٪	۹۲٪	۹۳٪	۹۲٪

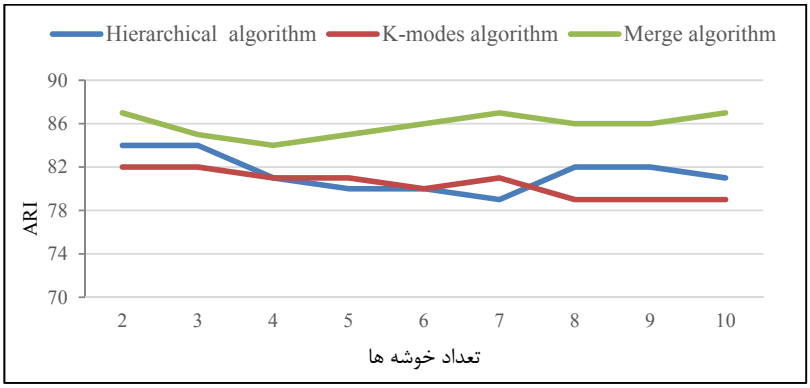


شکل ۵: نتایج خوشه‌بندی داده‌های zoo با استفاده از معیار Accuracy برحسب درصد٪

از خوشـــه‌بندی الگوریتم‌ها با اســـتفاده از معیار ارزیابی ARI را نشان می‌دهد و عدد بزرگتر نشان دهنده خوشه‌بندی بهتر اســت. همان‌طور که در جداول ۴ و ۵ نشــان داده شــده ا ست، با افزایش تعداد خو شه‌ها کارایی خو شه‌بندی افت پیدا می‌کند. همان‌طور که در جداول ۴ و ۵ قابل مشــاهده اســت میزان ARI خو شه‌بندی الگوریتم پیشنهادی نسبت به هر دو الگوریتم خو شه‌بندی سلسله مراتبی و k-modes، ر شد قابل ملاحظه‌ای داشته است.

جدول۴: نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار ARI برحسب درصد٪

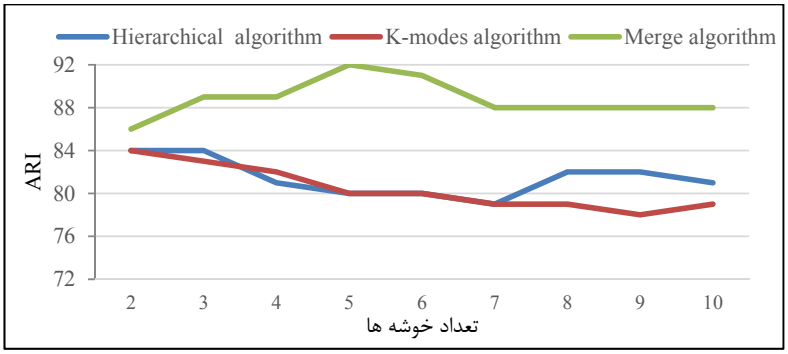
الگوریتم	تعداد خوشه‌ها								
	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical algorithm	۸۴٪	۸۴٪	۸۱٪	۸۰٪	۸۰٪	۷۹٪	۸۲٪	۸۲٪	۸۱٪
K-modes algorithm	۸۲٪	۸۲٪	۸۱٪	۸۱٪	۸۰٪	۸۱٪	۷۹٪	۷۹٪	۷۹٪
Merge algorithm	۸۷٪	۸۵٪	۸۴٪	۸۵٪	٪۸۶	۸۷٪	۸۶٪	۸۶٪	۸۷٪



شکل۶: نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار ARI برحسب درصد٪

جدول ۵: نتایج خوشه‌بندی داده‌هایzoo با استفاده از معیار ARI برحسب درصد٪

الگوریتم	تعداد خوشه‌ها								
	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical algorithm	۸۴٪	۸۴٪	۸۱٪	۸۰٪	۸۰٪	۷۹٪	۸۲٪	۸۲٪	۸۱٪
K-modes algorithm	۸۴٪	۸۳٪	٪۸۲	٪۸۰	٪۸۰	٪۷۹	٪۷۹	٪۷۸	٪۷۹
Merge algorithm	۸۶٪	٪۸۹	٪۸۹	۹۲٪	۹۱٪	۸۸٪	۸۸٪	۸۸٪	۸۸٪

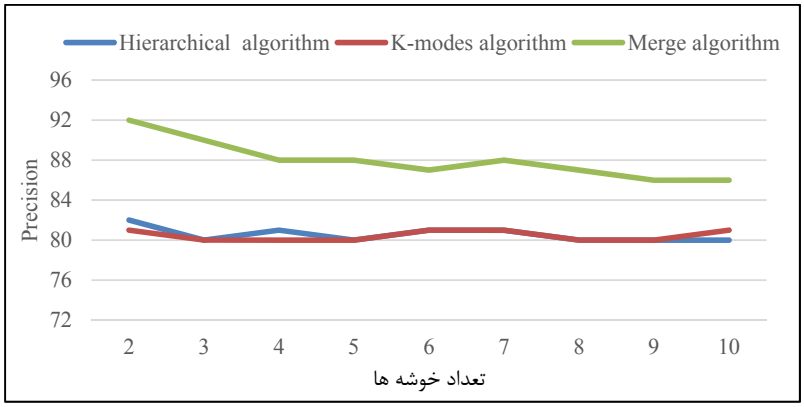


شکل ۷: نتایج خوشه‌بندی داده‌های zoo با استفاده از معیار ARI برحسب درصد٪

جدول ۶ و شکل ۸ مقایسه بین الگوریتم ارائه شده در این مقاله با دو الگوریتم خوشه‌بندی سلسله مراتبی و k-modes، با معیار Precision روی مجموعه داده soybean data با تعداد خوشه‌های متفاوت را نشان می‌دهد. جدول ۷ و شکل ۹ همین مقایسه را روی مجموعه داده 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 نشان دهنده خوشه‌بندی بهتر است. همان‌طور که در جداول ۶ و ۷ نشان داده شده است، با افزایش تعداد خوشه‌ها کارایی خوشه‌بندی افت پیدا می‌کند. همان‌طور که در جداول ۶ و ۷ قابل مشاهده است، میزان Precision خوشه‌بندی الگوریتم پیشنهادی نسبت به هر دو الگوریتم خوشه‌بندی سلسله مراتبی و k-modes رشد قابل ملاحظه‌ای داشته است.

جدول۶: نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار Precision برحسب درصد٪

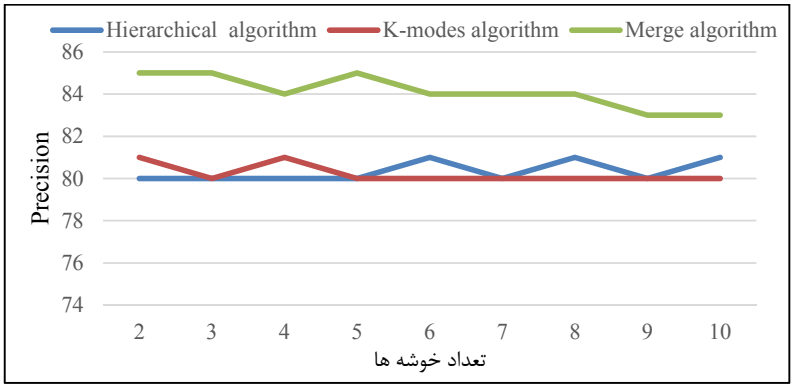
الگوریتم	تعداد خوشه‌ها								
	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical algorithm	۸۲	۸۰	۸۱	۸۰	۸۱	۸۱	۸۰	۸۰	۸۰
K-modes algorithm	۸۱	۸۰	۸۰	۸۰	۸۱	۸۱	۸۰	۸۰	۸۱
Merge algorithm	۹۲	۹۰	۸۸	۸۸	۸۷	۸۸	۸۷	۸۶	۸۶



شکل۸: نتایج خوشه‌بندی داده‌های soybean با استفاده از معیار Precision برحسب درصد٪

جدول ۷: نتایج خوشه‌بندی داده‌هایzoo با استفاده از معیار Precision برحسب درصد٪

الگوریتم	تعداد خوشه‌ها								
	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Hierarchical algorithm	۸۰	۸۰	۸۰	۸۰	۸۱	۸۰	۸۱	۸۰	۸۱
K-modes algorithm	۸۱	۸۰	۸۱	۸۰	۸۰	۸۰	۸۰	۸۰	۸۰
Merge algorithm	۸۵	۸۵	۸۴	۸۵	۸۴	۸۴	۸۴	۸۳	۸۳



شکل ۹: نتایج خوشه‌بندی داده‌های zoo با استفاده از معیار Precision برحسب درصد٪

جدول ۸: نتایج خوشه‌بندی سه الگوریتم با استفاده از معیار Accuracy، Precision، ARI

داده	معیار	الگوریتم		
		Squeezer algorithm	Weighting K-modes	Merge algorithm
zoo data	Accuracy	۹۰٪	۸۵٪	۹۵٪
	Precision	۸۰٪	۸۲٪	۸۵٪
	ARI	۸۰٪	۸۰٪	۸۹٪
soybean data	Accuracy	۸۲٪	۹۰٪	۹۴٪
	Precision	۸۶٪	۸۲٪	۸۸٪
	ARI	۸۰٪	۸۳٪	۸۶٪

جدول ۸ مقایسه‌ای بین الگوریتم پیشنهادی در این مقاله با دو الگوریتم بیان شده در پیشینه پژوهش (الگوریتم Weighting k-modes و الگوریتم Squeezer) روی مجموعه داده zoo data و soybean data[32] با تعداد خوشه معین را بیان می‌کند. دو الگوریتم Weighting k-modes و Squeezer از جمله الگوریتم‌هایی هستند که برای خوشه‌بندی مجموعه داده‌های دسته‌ای در سال‌های اخیر ارائه شده‌اند و از کارایی بالایی برخوردارند [12, 17]. چون شرایط الگوریتم‌های ذکر شده با هم متفاوت است، متوسط میزان Accuracy، ARI، و Precision هر الگوریتم روی این دو مجموعه داده با تعداد خوشه ۵ آمده است. از آنجا که شرایط الگوریتم‌ها متفاوت است (منظور از شرایط متفاوت این است که هر یک از الگوریتم‌ها یاد شده روش مخصوصی برای خوشه‌بندی داده‌ها دارند به عنوان مثال الگوریتم Squeezer نیاز به تعیین یک مقدار آستانه برای اندازه‌گیری شباهت دارد که این مقدار آستانه نیز به صورت تجربی بدست می‌آید)، نمی‌توان مانند ارزیابی قبلی در اینجا هم از نمودار و جدول‌های مشابه قبلی استفاده کرد. بدین منظور متوسط میزان هر یک از معیارها محاسبه شده و در جدول ۸ قرار گرفته است.

در جدول ۸، هر کدام از مقادیر، حاصل متوسط ۱۰ بار اجرای الگوریتم می‌باشند. همان‌طور که در جدول ۸ قابل مشاهده است، میزان Accuracy، ARI، و Precision الگوریتم پیشنهادی، در مقایسه با دو الگوریتم Weighting K-modes و Squeezer، بالاتر است.

۵. نتیجه‌گیری

در این مقاله ترکیب دو الگوریتم خوشه‌بندی سلسله‌مراتبی با الگوریتم k-modes، برای داده‌های دسته‌ای ارائه شد. در اینجا ابتدا الگوریتم خوشه‌بندی سلسله‌مراتبی روی داده‌های دسته‌ای اعمال می‌شود. سپس، خوشه‌بندی تولید شده توسط این الگوریتم به عنوان ورودی به الگوریتم خوشه‌بندی k-modes داده می‌شود. در این متد دو مرحله‌ای، اشیاء شبیه به هم داخل یک زیرمجموعه قرار می‌گیرند و در گام دوم به عنوان ورودی به الگوریتم خوشه‌بندی k-modes داده می‌شود. فایده اصلی روش پیشنهادی این است که تأثیر نویز و داده‌های پرت در فرایند خوشه‌بندی بسیار کاهش می‌یابد و روالی مشخص برای محاسبه اولیه مراکز خوشه‌های برای الگوریتم k-modes ارائه می‌دهد. نتایج حاصل از خوشه‌بندی این الگوریتم نشان می‌دهد که میزان Accuracy، ARI، و Precision فرایند خوشه‌بندی نسبت به دو الگوریتم سلسله مراتبی و k-modes و سایر الگوریتم‌های مطرح در زمینه خوشه‌بندی داده‌های دسته‌ای افزایش یافته است. میزان Accuracy در این خوشه‌بندی ترکیبی با افزایش تعداد خوشه‌ها روند یکنواخت‌تری دارد و نتایج دقیق‌تری را حاصل می‌دهد.

مراجع

[1] C. Li, J. Zhou, P. Kou and J. Xiao, "A novel chaotic particle swarm optimization based fuzzy clustering algorithm," Neurocomputing, vol. 83, p. 98–109, April 2012.

[2] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data," Englewood Cliffs, NJ: Prentice-Hall, 1988.

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

[4] L. Bai, J. Liang, C. Dang and F. Cao, "A novel fuzzy clustering algorithm with between-cluster information for categorical data," Fuzzy Sets and Systems, vol. 215, p. 55–73, 2013.

[5] F. Cao, J. Liang, D. Li, L. Bai and C. Dang, "A dissimilarity measure for the k-Modes clustering algorithm," Knowledge-Based Systems, vol. 26, p. 120–127, February 2012.

[6] H.-L. Chen, K.-T. Chuang and M.-S. Chen, "On Data Labeling for Clustering Categorical Data," Transactions on knowledge and data engineering, vol. 20, no. 11, pp. 1458 - 1472, 2008.

[7] G. Wiederhold, Advances in Knowledge Discovery in Databases, California: AAAI/MIT Press, 1996.

[8] Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining," in Proceedings of ACM SIGMOD Workshop on Research Issues on data Mining and knowledge Discovery, 1997.

[9] C.-H. Yun, K.-T. Chuang and M.-S. Chen, "Adherence clustering: an efficient method for mining market-basket clusters," Information Systems, vol. 31, no. 2, pp. 170-186 , May 2006.

[10] V. Ganti, J. Gehrke and R. Ramakrishnan, "CACTUS: Clustering Categorical Data Using

Summaries," in Proceedings ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 1999.

[11] L. Bai and J. Liang, "The k-modes type clustering plus between-cluster information for categorical data," Neurocomputing, vol. 133, p. 111–121, June 2014.

[12] F. Cao, J. Liang, D. Li and X. Zhao, "A weighting k-modes algorithm for subspace clustering of categorical data," Neurocomputing, vol. 108, p. 23–30, 2013.

[13] A. Saha and S. Das, "Categorical fuzzy k-modes clustering with automated feature weight learning," Neurocomputing, vol. 166, p. 422–435, 2015.

[14] R. Sankar Sangam and H. Om, "The k-modes algorithm with entropy based similarity coefficient," Procedia Computer Science, vol. 50, pp. 93-98, 2015.

[15] J. Ji, W. Pang, C. Zhou, X. Han and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," Knowledge-Based Systems, vol. 30, p. 129–135, 2012.

[16] A. Tyagi and S. Sharma, "Implementation Of ROCK Clustering Algorithm For The Optimization Of Query Searching Time," International Journal on Computer Science and Engineering (IJCSE), vol. 4, no. 5, May 2012.

[17] H. Zengyou, X. Xiaofei and D. Shengchun, "Squeezer: An Efficient Algorithm for Clustering Categorical Data," J. Computer Science and Technology, vol. 17, no. 5, pp. 611-624, 2002.

[18] P. Andritsos, P. Tsaparas, R. J. Miller and K. C. Sevcik, "LIMBO: Scalable Clustering of Categorical Data," in 9th International Conference on Extending Database Technology, 2004.

[19] D. Barbar'a, J. Couto and Y. Li, "COOLCAT: An entropy-based algorithm for categorical clustering," in Proceedings of ACM CIKM International Conference on Information and Knowledge Management, 2002.

[20] Y. Yang, X. Guan and J. You, "CLOPE: a fast and effective clustering algorithm for transactional data," in eighth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, 2002.

[21] A. K. Jain and J. Mao, "Statistical Pattern Recognition: A Review," Transactions on pattern analysis and machine intelligence, vol. 22, no. 1, p. 4–37, 2000.

[22] F. Hosseininezhad and A. Salajegheh, "Study and Comparison of Partitioning Clustering Algorithms," Iranian Journal of Medical Informatics, vol. 2, no. 1, 2012.

[23] M. Verma, M. Srivastava, N. Chack, A. Kumar Diswar and N. Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Research and Applications (IJERA), vol. 2, no. 3, pp. 1379-1384, May-Jun 2012.

[24] R. Giancarlo, G. Lo Bosco and L. Pinello, "Distance Functions, Clustering Algorithms and Microarray Data Analysis," Learning and Intelligent Optimization, vol. 6073, pp. 125-138, 2010.

[25] S. Boriah, V. Chandola and V. Kumar, "Similarity Measures for Categorical Data: A Comparative Evaluation," in Proceedings SIAM International Conference on Data Mining, 2008.

[26] L. Bai, J. Liang and C. Dang, "A cluster centers initialization method for clustering categorical data," Expert Systems with Applications, vol. 39, p. 8022–8029, 2012.

[27] J. Lee and Y.-J. Lee, "An effective dissimilarity measure for clustering of high-dimensional categorical data," Knowledge and Information Systems, vol. 38, no. 3, pp. 743-757, March 2014.

[28] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, New York: Wiley, 1990.

[29] X. Zhang and Z. Xu, "Hesitant fuzzy agglomerative hierarchical clustering algorithms," International Journal of Systems Science, vol. 46, no. 3, pp. 562-576, 2015.

[30] M. R. Ackermann, J. Blömer, D. Kuntze and C. Sohler, "Analysis of Agglomerative Clustering," Algorithmica, vol. 69, no. 1, pp. 184-215, May 2014.

[31] K. Vinothkumar and M. P. Selvan, "Hierarchical Agglomerative Clustering Algorithm method for distributed generation planning," International Journal of Electrical Power & Energy Systems, vol. 56, p. 259–269, March 2014.

[32] H. Zhexue, "(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, September 1998.

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

[34] I. Saha, J. Prasad Sarkar and U. Maulik, "Ensemble based rough fuzzy clustering for categorical data," Knowledge-Based Systems, vol. 77, pp. 114-127, 2015.

[35] H.-J. Do and J.-Y. Kim, "Categorical Data Clustering Using the Combinations of Attribute Values.," in International Conference (ICCSA), Perugia, 2008.

[36] L. Bai, J. Liang and C. Dang, "An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data," Knowledge-Based Systems, vol. 24, p. 785–795, 2011.

[37] M.-Y. Shih, J.-W. Jheng and L.-F. Lai, "A Two-Step Method for Clustering Mixed Categroical and

Numeric Data," Tamkang Journal of Science and Engineering, vol. 13, no. 1, pp. 11-19, 2010.

[38] R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis, New Jersey: Prentice Hall, Englewood Cliffs, 1988.

[39] S. S. Mesakar and M. S. Chaudhari, "Review Paper On Data Clustering Of Categorical Data," International Journal of Engineering Research & Technology (IJERT), vol. 1, no. 10, December 2012.

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

[41] D. Gibson, J. Kleinberg and P. Raghavan, "Clustering categorical data: an approach based on dynamical systems," The International Journal on Very Large Data Bases, vol. 8, no. 3-4, pp. 222-236, February 2000.

[42] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, Advances in knowledge discovery and data mining, CA, USA: American Association for Artificial Intelligence Menlo Park, 1996.

[43] J. Han and M. Kamber, Data Mining Concepts And Techniques, Harcourt India Private Limited, 2010.

[44] M. Kim and R. S. Ramakrishna, "Projected clustering for categorical datasets," Pattern Recognition Letters, vol. 27, no. 12, p. 1405–1417, September 2006.

[45] M. H. Dunham, Data Mining Introductory and Advanced Topics, Prentice Hall/Pearson Education, 2003.

[46] P. Pendharkar, Managing Data Mining Technologies in Organizations: Techniques and Applications, Idea Group Publishing, 2003.

[47] G. Punj and D. W. Steward, "Cluster analysis in marketing research: review and suggestions for research," Journal of Marketing Research, vol. 20, pp. 134-148, 1983.

[48] B. Babcock, S. Babu, M. Datar, R. Motwani and J. Widom, "Models and issues in data stream systems," in Proceedings of PODS, 2002.

[49] M. J. Zaki and M. Peters, "Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques," in Proceedings International Conference. Data Eng. (ICDE), 2005.