موضوع: **ارائه متد ترکیبی جدید به منظور خوشه بندی به روشK-Means تقسیم و غلبه** **K-Means Divide And Conquer Clustering**
نویسنده: **میلاد یداللهی- مدیر فنی شرکت آشناایمن-m.yadollahi@ashnasecure.com**
ارائه شده در: **سومین کنفرانس ملی داده کاوی ایران- دانشگاه امیرکبیر- تابستان ۸۸**

چکیده

روش آنالیز خوشه‌ای در حقیقت نوعی کاوش ابتدایی بدون دانش یا با دانش قبلی کم می باشد که تحقیقات زیادی در مورد آن به شکل های متفاوت صورت پذیرفته است. بسیاری از تکنیک‌های خوشه‌ای مواردی همچون اندازه و سطح داده ها را درنظر نمی گیرند و در بسیاری موارد اشیاء یا نمونه های مشابه را به اشتباه و بدون توجه به اینکه ممکن است آنها در سطوح مختلفی قرار داشته باشند، دریک خوشه قرار می دهند. برای گروه‌های داده‌ای بسیار بزرگ و چندین بعدی، قبل از اجرای تکنیک های داده‌کاوی، مرحله حذف برخی داده‌ها می بایست اعمال گردد. معمولا در این بخش بعضی از بعدهای داده‌ای حذف می شوند. ولی سوالی که باقی میماند این است که آیا این داده‌های آماده شده و مورد پردازش قرار گرفته، می توانند نتیجه ای مناسب را به ما بدهند و حذف برخی ابعاد تاثیری بر روی آنها نمی گذارد.

روش خوشه بندی ترکیبی مطرح شده در این مقاله از دو مرحله خوشه بندی به صورت تقسیم و غلبه استفاده می نماید، این روش بدین ترتیب است که در مرحله اول خوشه‌بندی بمنظور گروه بندی نمونه های هم سطح، کل فضا را به زیرفضاهایی تبدیل کرده و در مرحله دوم به خوشه‌بندی هریک از زیرفضاها به روش K-Means می پردازد و مشکل مطرح شده به عنوان عدم هماهنگی را حذف می‌نماید. این روش هیچ بعدی را حذف نخواهد کرد و به جای آن زیرفضاهای کوچکتر را انتخاب می نماید و براساس همین زیرفضاها عمل خوشه‌بندی را انجام می دهد. روش مطرح شده از روش های معمول خوشه‌بندی کارآمدتر بوده و از صحت بالاتری برخوردار است، همچنین ملاحظه می شود که این روش از لحاظ تعداد تکرار برای عمل خوشه‌بندی بهینه‌تر عمل می کند.

کلمات کلیدی

داده‌کاوی، خوشه‌بندی، K-Means، تقسیم و غلبه، آنالیز شخصیتی.

K-Means Divide And Conquer Clustering

Milad Yadollahi

ABSTRACT

Cluster analysis, primitive exploration with little or no prior knowledge, consists of research developed across a wide variety of communities. Most clustering techniques ignore the fact about difference in size or level – where in most cases, clustering are more concern with grouping similar objects or samples together, ignoring the fact that even though they might be similar but they are belong to different levels. For really large data sets, data reduction should be performed prior to applying the data-mining techniques which is usually performing dimension reduction, and the main question is whether some of these prepared and preprocessed data can be discarded without sacrificing the quality of results. Existing clustering techniques would normally merge small clusters with big ones, removing its identity.

The proposed method uses a two-step clustering base on divide and conquer. In this method, first we cluster the objects based on their size and level and actually create some subspaces, then we cluster each of the subspaces that was created by the previous K-means clustering. Our proposed method is not to reduce dimension but to select subspaces by

clustering and perform clustering based on these subspaces. This method has achieved more accurate and efficient results in comparison with the similar methods. Also we can see the numbers of iterations are reduced.

KEYWORDS
Data mining, clustering, K-means, Divide and conquer, Personality analysis.

۱. مقدمه

تکنیک های خوشه بندی با این هدف شروع بکار می کنند که مجموعه از رکوردها را به چند گروه تقسیم کرده و رکوردهای "مشابه" را در خوشه های یکسان قرار دهند و بدین ترتیب چندین زیرمجموعه داده ای را ایجاد می نماید. نمونه ها در یک گروه مشابه قرار می گیرند و بدین ترتیب نمونه هایی که در گروه های متفاوت قرار دارند مشابه نیستند، مثلا یک خوشه ممکن است دسته ای از مشتریان با سابقه خرید یکسان، تراکنش های مشابه و دیگر فاکتورهای همانند و مشابه یکدیگر باشد. نمونه ها در تکنیک خوشه بندی به عنوان یک بردار در یک فضای چندبعدی یا حتی یک نقطه می توانند خود را نشان دهند. شباهت میان دو بردار مانند بردار X و Y ، می تواند با استفاده از معیارهای مختلفی همچون رابطه کسینوسی ۱ محاسبه شود:

رابطه ۱

$$S(\vec{X},\vec{Y}) = \frac{\vec{X}^t \,.\, \vec{Y}}{|\vec{X}|\ |\vec{Y}|}$$

$\vec{X}^t$ ترانهاده بردار $\vec{X}$ و $|\vec{X}|$ اندازه بردار $\vec{X}$ می باشد.

رابطه ۱ را می توان در ابعاد بیش از دو بعد نیز بسط داد و به وسیله آن میزان تشابه را میان دو شی اندازه گیری نمود. نمونه‌هایی که دارای شباهت بیشتری با یکدیگر می باشند، مقادیر کمتری را برای کسینوس ایجاد می نمایند. به عنوان مثال در این موارد زاویه میان بردارهای هریک کوچکتر خواهد بود. در تحقیق صورت گرفته و برای کسب داده های تحقیقاتی، وب سایتی (www.peivand.org) ایجاد گردید که در آن کاربران، اشیای خوشه بندی ما را تشکیل می دادند. در این وب سایت افراد پس از پاسخگویی به چندین سوال روانشناسی، ابعاد شخصیتی خود را به سیستم داده کاوی معرفی می نمایند. سپس خوشه بندی اولیه با توجه به رابطه فوق ایجاد می گردد.

بیشتر تحقیقات صورت گرفته در رابطه با خوشه بندی، بیشتر به فاکتور فاصله توجه کرده اند و کمتر به فاکتورهایی مانند سطح یا اندازه توجه شده است. نمونه هایی ممکن است در یک گروه قرار بگیرند ولی متعلق به سطوح متفاوتی باشند، مثلا یک ماهیگیر و یک شرکت ماهیگیری. در مورد انسانها نیز می توان سطوح مختلف تحصیلی، درآمد، ثروت و موارد دیگر را تعریف نمود. بدین ترتیب درصورتی که بخواهیم افراد یا اشیاء دیگر را تنها با معیار فاصله و بدون توجه به سطح آنها گروه بندی نماییم، ممکن است اشیایی که در سطوح مختلف قرار دارند را در یک گروه قرار دهیم. تحقیقات کمی در مورد دیتاست های بزرگی که دارای ابعاد زیادی می باشند، صورت گرفته است. بیشتر تحقیقات صورت گرفته در مورد داده هایی که ابعاد زیادی را دارا می باشند در ابتدا با کاهش ابعاد داده همراه می باشند که این موارد سوالی را مطرح می سازد که آیا با این عمل صحت و قابلیت اطمینان نتایج حفظ می گردد. در زمانی که با فضاهای کوچکتر کار می کنیم، می توانیم بدون هیچ نگرانی‌ای در مورد پیچیدگی و منابع موجود از فاصله اقلیدوسی استفاده نماییم، دراینصورت قادر خواهیم بود که کارآمدی و صحت تکنیک های خوشه بندی موجود را توسعه دهیم. روشی که مطرح خواهد شد هیچ بعدی را حذف نخواهد کرد و به جای آن زیرفضاهای کوچکتر را انتخاب می نماید و براساس همین زیرفضاها عمل خوشه بندی را انجام می دهد. انتظار می رود که این روش بتواند نمونه های دارای اندازه یکسان و هم سطح را در یک خوشه قرار داده و صحت و بهینگی بالاتری نسبت به خوشه بندی یک بعدی را از خود نشان دهد.

۲. بررسی دیگر تحقیقات

تحقیقات بسیار زیادی در سال های اخیر برای توسعه خوشه بندی صورت گرفته است. یک دسته بندی عمومی برای خوشه بندی دیتاست های دارای ابعاد زیاد وجود دارد که بصورت زیر می باشد:

۱- کاهش ابعاد
۲- مدل های صرفه جویی
۳- خوشه بندی زیرفضایی [۱] .

انتخاب مشخصه و استخراج مشخصه، معروفترین تکنیک های مربوط به روش کاهش ابعاد می باشند، البته واضح است که در هر دو تکنیک بخشی از اطلاعات را از دست خواهیم داد و طبیعتا این موضوع بر روی صحت نتایج تاثیرگذار خواهد بود. یک بررسی مناسب از مدل های صرفه جویی در [۲] موجود می باشد. همچنین مدل های گوس از ساده ترین تا پیچیده ترین شکل موجود می باشند که از لحاظ بهینگی تقریبا با روش K-Means یکسان هستند. باید توجه داشت که در جایی که دارای فضاهایی با ابعاد کم می باشیم، این روش ها دارای عملکرد مناسبی نمی‌باشند. دو رویکرد اصلی برای روش های زیرفضایی وجود دارد، اولا می بایست مراکز کلاس در زیرفضای یکسان قرار گرفته شوند و ثانیا هر کلاس در یک زیرفضای خاص باشد [۳] . روش استفاده از زیرموضوعات و زیرگروه ها در روش هایی همچون روش خوشه بندی متون و متن کاوی و موارد مشابه مناسب می باشند [۴] . همچنین روش فاکتورگیری تنشی یا tensor factorization به عنوان یک روش مناسب در [۵] مورد استفاده قرار گرفته است. ناسازگاری و عدم یکپارچگی به عنوان یک مشکل در چنین دیتاست های عمومی نشان داده شده است.

در [۶] ما شاهد روشی هستیم که انتخاب زیرفضاها را با خوشه بندی ترکیب می نماید. تعادل میان خوشه بندی K-Means و انتخاب مکرر زیرفضاها نیز نشان داده شده است. همچنین در [۷] یک الگوی توسعه خوشه بندی عمومی وجود دارد که در آن دو مرحله اصلی مشاهده می‌شود. در مرحله اول از مشخصه های باطنی دیتاست برای کاهش ابعاد استفاده می شود و بعد از آن چندین بار الگوریتم های خوشه بندی با پارامترهای متفاوت اجرا می شوند و سپس براساس معیار BIC بهترین نتیجه انتخاب می گردد. البته ضعف هایی در این روش وجود دارد، مثلا ازآنجاییکه BIC یک مدل را برای توزیع دیتاست خاصی مورد استفاده قرار می دهد، نمی توان از آن برای مقایسه مدل های دیتاست های مختلف استفاده نمود. همچنین روشی در [۸] ارائه شده است که خوشه بندی نیمه نظارتی را با استفاده از روش K-Means کروی برای کار با داده های چند بعدی اسپارس مورد استفاده قرار می دهد. در [۹] روشی موجود است که خوشه بندی براساس شی ترکیبی و خوشه بندی تئوری گراف را تلفیق می نماید.

- روش اول : چندین محدوده را برحسب اهمیت و میزان محوری بودن به عملکردی بر اساس شی مربوط می سازد و این مسئله را با استفاده از برنامه نویسی بهینه غیرخطی حل می نماید.
- روش دوم: شامل دو رویه پشت سرهم می باشد:
  - یک خوشه بندی عملکرد بر اساس شی برای ایجاد نتایج اولیه
  - یک سیستم افزودنی ارتباط خودکار بر اساس خوشه بندی تئوری گراف برای اصلاح نتایج اولیه.

در [۱۰] روشی مطرح شده که می توان توسط آن، ترکیب روش های خوشه بندی ترتیبی را برای خوشه بندی داده ها استفاده نمود، برای توسعه عملکرد خوشه بندی می توان از بیش از یک روش خوشه بندی واحد استفاده نمود. سپس به این نتیجه خواهیم رسید که ترکیب روش های خوشه بندی ترتیبی از روش ترتیبی ساده‌تر می باشد و می توان بدون داشتن هزینه های اضافی مزایایی را کسب نمود.

مشکل اصلی زمانی است که نقاط خوشه بندی برای ایجاد یک معیار تعیین شباهت، در فضاهای چندین بعدی قرار می گیرند. تحقیقات اخیر نشان داده است که در فضاهای چندین بعدی، پیدا کردن فواصل مختلف در ابعاد متفاوت مناسب نمی باشد، زیرا ممکن است دورترین همسایه در یک بعد، در واقعیت نزدیک ترین همسایه باشد [۱۱] .

برای محاسبه خوشه ها در زیرفضاهایی که ابعاد کمتری دارند، تحقیقات اخیر بر روی خوشه بندی افکنشی یا projective clustering متمرکز شده است، این روش بدین صورت است که دسته ها به زیرمجموعه هایی تقسیم می گردند که بوسیله تابعی بر اساس شی به فضاهایی با ابعاد کمتر تقسیم می شوند. بجای اینکه تمامی نقاط را در یک زیرفضا قرار دهیم، با این روش می توانیم برای هر خوشه چندین زیرفضای مرتبط داشته باشیم [۱۲] . مدلی که ارائه خواهد شد کارایی بالاتری دارد و روش قدیمی خوشه بندی را به صورت مناسبی ارتقا داده است. این روش قادر خواهد بود که نمونه هایی که دارای اندازه و سطح یکسانی می باشند را به صورت کاراتر و با صحت بالاتری در مقایسه با روش خوشه بندی یکباره، خوشه بندی نماید. محدودیت هایی نیز برای این روش وجود دارد. اولا فضا می بایست قائمه باشد، بدین معنا که نباید هیچ گونه وابستگی ای میان صفات یک شی وجود داشته باشد. ثانیا براساس برنامه ای که مورد استفاده قرار می گیرد، تمامی صفاتی که در یک شی وجود دارد می بایست نوع های داده ای یکسانی داشته باشند. با حفظ کلیات این روش، می توان به نوعی عمل نمود که بتوان از نوع های داده ای دیگر هم استفاده نمود. ازطرفی نمونه ها می بایست دارای ابعاد برابری باشند. در این مقاله با انتخاب زیرفضاها و انجام عمل خوشه بندی بر اساس این فضاها، روش تقسیم و غلبه K-Means ارائه خواهد شد.

در این مقاله مزایا و فرض های زیر موجود می باشد:

فرض۱ : روش مطرح شده قادر خواهد بود که نمونه های دارای اندازه یکسان را گروه بندی نموده و شباهت میان آنها را پیدا نماید.

فرض۲ : روش مطرح شده دارای سرعت بالاتری در مقایسه با خوشه بندی یک مرحله‌ای توسط روش تقسیم و غلبه می باشد.

فرض۳ : روش مطرح شده دارای دقت بیشتری از خوشه بندی یک مرحله‌ای توسط روش تقسیم و غلبه می باشد.

فرض۴ : روش مطرح شده اجازه می دهد که فاصله اقلیدوسی در داده های دارای ابعاد زیاد مورد استفاده قرار گیرد.

محدودیت هایی نیز برای استفاده از این روش وجود دارد که عبارتند از: ۱- فضا می بایست قائمه باشد، ۲- ابعاد، برای تمامی اشیاء یکسان است و ما از نوع داده ترتیبی استفاده می کنیم. در ادامه چارچوب کاری و متدولوژی مورد استفاده در این روش توضیح داده خواهد شد. نهایتا نتایج تحقیق ارائه شده و بر روی آن بحث خواهیم کرد.

۳. مدل تحقیق

در این تحقیق از چارچوب کاری ای استفاده شده است که متشکل از پنج مرحله اصلی می باشد. شکل ۱ مراحل انجام شده در این چارچوب کاری را نشان می دهد.

شکل ۱- چارچوب کاری برای خوشه بندی دیتاست های چندین بعدی

در این چارچوب بعد از پیش پردازش و نرمال‌سازی، برای هر شی رابطه $\sum x^2$ اعمال می‌گردد. این رابطه مجموع تمامی صفات هر شی را محاسبه می کند. پس از اینکه اندازه اشیاء را محاسبه کردیم، آنها را بر اساس اندازه، خوشه بندی کرده و در نهایت هر گروه را به صورت جداگانه خوشه بندی می نماییم.

۱.۳. پیش پردازش

- داده ها می بایست قبل انجام داده کاوی پاکسازی شوند.
- از آنجایی که داده های تحقیقاتی از وب سایت مربوطه گرفته شده اند، برخی نام ها می بایست تغییر نمایند.
- همچنین مقادیری وجود دارد که در بازه مجاز قرار نمی گیرند. بازه مجاز اعداد صحیح میان ۱ تا ۵ می باشند:
  - در صورتی که مقدار کمتر از ۱ بود آن را به ۱ تغییر می دهیم.
  - در صورتی که بزرگتر از ۵ بود آن را به ۵ تغییر می دهیم.
- ممکن است مقادیری موجود نباشند یا از بین رفته باشند:
  - در این صورت شی را حذف کرده و از آن چشم پوشی می نماییم.
  - عدد ۱ را به آن اختصاص می دهیم..

۲.۳. نرمال‌سازی

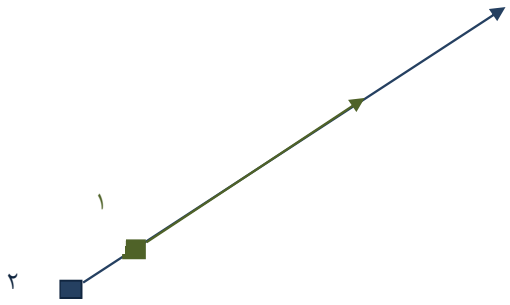
داده ها بصورت ترتیبی می باشند، بنابراین می بایست با استفاده از رابطه ۲ آنها را نرمال کرد:

رابطه ۲- رابطه نرمال‌سازی

$$z_{if} = \frac{r_{if} - 1}{m_f - 1}$$

۳.۳. روش ترکیبی برای خوشه بندی

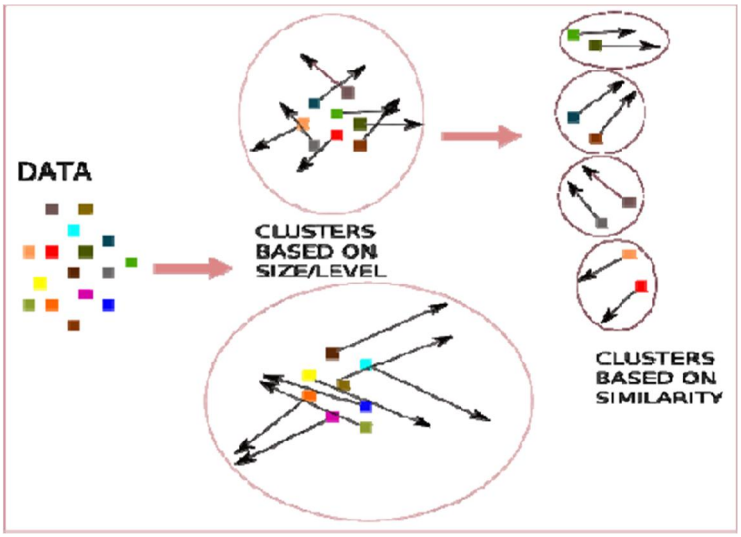
رابطه $\sum x^2$ برای دسته بندی اشیائی که دارای اندازه یکسانی می باشند بکار می رود، بدین ترتیب ما دارای نمونه هایی با اندازه و یا سطح برابر خواهیم بود.



شکل ۲- دو بردار مشابه ولی متفاوت در سطح

شکل۲ دو بردار را نشان می دهد که از لحاظ معیار شباهت کسینوسی، مشابه یکدیگر می باشند ولی واضح است که با توجه به ابعاد متفاوتی که هرکدام از آنها دارند، از لحاظ اندازه و سطح با یکدیگر متفاوت می باشند. با این روش به عنوان مثال ممکن است یک دانشجوی دکترا با یک کودک دبستانی در یک خوشه قرار بگیرند.

در شکل ۳ همانطور که نشان داده شده است، ما در هر خوشه زیرگروه‌هایی را تشکیل می دهیم که اشیا را بر اساس ویژگی های آنها خوشه بندی می نمایند، باید توجه داشت که در خوشه بندی مرحله اول تنها معیار عملکرد، اندازه بود و به ویژگی ها و صفات توجه نشد.

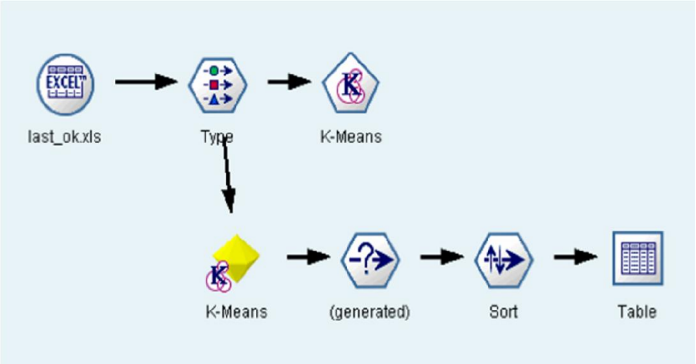


شکل ۳- مراحل روش خوشه بندی ارائه شده

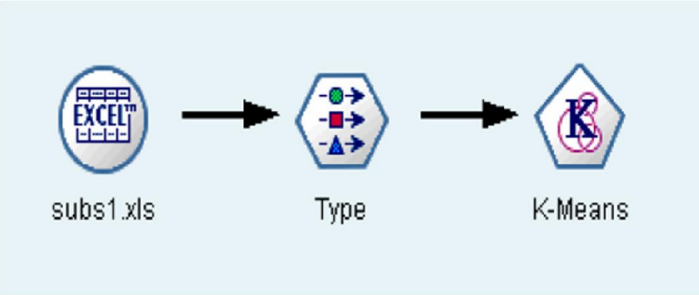
بعد از انجام دومین عملیات خوشه بندی، تقریبا اشیائی که در یک سطح قرار دارند در یک خوشه قرار می گیرند. بدین ترتیب هرکدام از زیرفضاهای ایجاد شده توسط خوشه بندی K-Means خوشه بندی می گردند.

۴. متدولوژی

یک طرح تحقیقاتی با استفاده از چارچوب کاری‌ای که بدان اشاره شد، ایجاد گردید. در ابتدا ابزارهای متفاوتی برای داده کاوی به منظور پیاده سازی مورد بررسی قرار گرفتند. در نهایت تصمیم بر این شد که از SPSS Clementine 12.0 در Oracle Data Mining 10.2 و WEKA 3-5-6 استفاده شود. SPSS Clementine 12.0 پشتیبانی خوبی را برای تحلیل تصویری ارائه می دهد و همچنین در آن قادر خواهیم بود که از رکوردهایی که به خوشه‌های مختلف تعلق دارند، گزارش بگیریم. SPSS Clementine 12.0 از الگوریتم های بیشتری پشتیبانی نموده و مرحله آماده سازی داده ها را بخوبی انجام می دهد.



شکل ۴- خوشه بندی مرحله اول



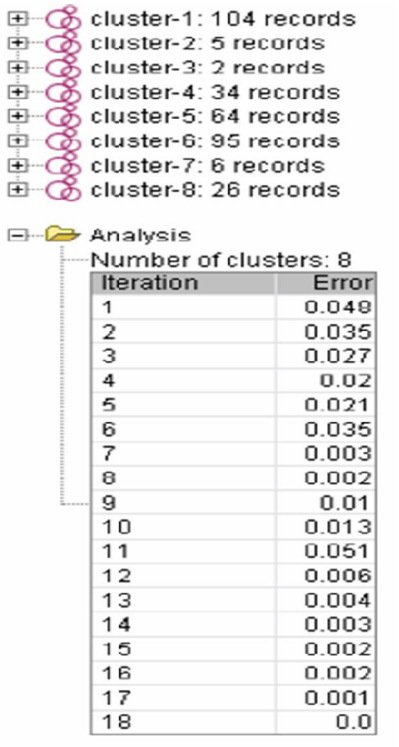
شکل ۵- خوشه بندی مرحله دوم

۵. رویه تحقیق

در وب سایت پروژه، سوالاتی بر اساس اصول روانشناسی طرح گردیده است که افراد پس از ثبت نام به آنها پاسخ می دهند. براساس پاسخ‌هایی که کاربر می دهد یک جدول ایجاد می گردد که ابعاد مختلف شخصیتی هر فرد را نشان می دهد. هر کدام از این ابعاد مقداری میان ۱ تا ۵ را به خود اختصاص می دهند. با استفاده از این مجموعه داده ای، در اولین خوشه بندی افرادی را که در یک سطح قرار دارند، ولی ممکن است دارای ابعاد شخصیتی متفاوتی باشند را در یک خوشه قرار می دهیم. در خوشه بندی دوم، افرادی که از لحاظ شخصیتی با یکدیگر شباهت دارند پیدا می‌شوند.

۶. نتایج و بررسی

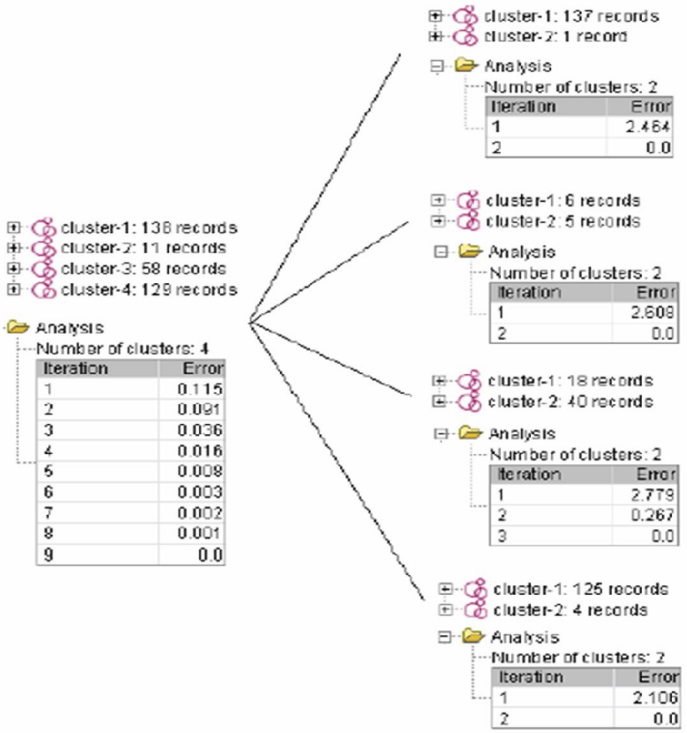
در این بخش به ارائه و بررسی نتایج بدست آمده از این تحقیق می پردازیم. در تحقیق اول خوشه بندی، بدون گروه بندی صورت گرفت و نتایج به صورت نشان داده شده در شکل ۶ حاصل گردید.



Iteration	Error
1	0.048
2	0.035
3	0.027
4	0.02
5	0.021
6	0.035
7	0.003
8	0.002
9	0.01
10	0.013
11	0.051
12	0.006
13	0.004
14	0.003
15	0.002
16	0.002
17	0.001
18	0.0

شکل ۶- نتایج بدست آمده در تحقیق اول

همچنین در این تحقیق مقدار$K=8$ مورد استفاده قرار گرفت. در تحقیق دوم روش توضیح داده شده مورد استفاده قرار گرفت که نتایج آن در شکل ۷ نشان داده شده است.



Iteration	Error
1	0.115
2	0.091
3	0.036
4	0.016
5	0.008
6	0.003
7	0.002
8	0.001
9	0.0

Iteration	Error
1	2.464
2	0.0

Iteration	Error
1	2.609
2	0.0

Iteration	Error
1	2.779
2	0.267
3	0.0

Iteration	Error
1	2.106
2	0.0

شکل ۷- نتایج بدست آمده در تحقیق دوم

در مرحله اول اشیاء به چهار گروه بر اساس اندازه تقسیم شدند و اشیاء متعلق به هر گروه دوباره بر اساس شباهت و رابطه‌ی مطرح شده خوشه بندی شدند. با مقایسه این دو تحقیق ملاحظه می شود که تعداد تکرارها کاهش یافته است. این بدین معنا است که با وجود خوشه‌های برابر در هر دو تحقیق، تکرار کمتری در روش ارائه شده دیده می شود.

از سوی دیگر، از آنجایی که خوشه‌ها با خوشه‌های بزرگتر در خوشه بندی معمولی K-Means ادغام شده اند، در این روش قادر خواهیم بود که خوشه هایی که دارای اعضای کمتری می باشند را نیز شناسایی نماییم. روش مطرح شده از روش خوشه بندی معمولی K-Means کارآمدتر بوده و از صحت بالاتری برخوردار می باشد.

۷. نتیجه

در این مقاله با بررسی تحقیقات صورت گرفته، روشی جدید برای خوشه بندی فضاهایی که دارای ابعاد زیاد می باشند مطرح گردید. استفاده از این روش خوشه بندی که از دو مرحله تشکیل شده است، در مجموعه های داده ای چندین بعدی با توجه به اندازه شی، باعث افزایش کارآمدی و صحت نتایج در مقایسه با روش خوشه بندی معمولی K-Means شد. زمانی که در مرحله اول اشیاء براساس معیار اندازه خوشه بندی می شوند، درواقع ما از زیرفضاهایی برای خوشه بندی استفاده کرده ایم. این باعث می شود که از نتایج صحیح تر و مناسب تری برخوردار شویم. به همین دلیل ما از فضای قائمه استفاده کردیم، بدین معنی که هیچگونه همپوشانی میان صفت های یک شی وجود نداشت و ابعاد می بایست دارای اندازه یکسان در تمامی اشیاء باشند. همچنین همان طور که اشاره شد این روش به دلیل انجام تکرارهای کمتر از بهینگی بیشتری نسبت به روش های معمول برخوردار می باشد.

برای پروژه های آینده سعی داریم که این روش را در محدوده های دیگر داده کاوی مانند متن کاوی اعمال نماییم. همچنین تاثیر پارامترهای مختلف در خوشه بندی مانند K-Means و همچنین تعداد ابعاد نیز باید مورد بررسی قرار گیرد.

۸. مراجع

[۱] C.Bouveyran; S. Girard; C.Schmid "*Technical Report 1083M"*, LMC-IMAGE 2008

[۲] C.Fraley; A. Raftery "*Model based clustering discriminate analysis and density estimation"*, journal of American statistics association, 97, 611-631 ,2002

[۳] H. Bock; "*On the interface between cluster analysis, principal component clustering and multidimensional scaling"*, 1987

[۴] Jeffery L. Solka; "*Text data mining Theory and methods*", vol 2, 94-112, 2008

[۵] H. Huang; C. Ding; D. Luo; T. Li "*Simultaneous tensor subspace selection and clustering*", KDD08, lasvegas USA, 2008.

[۶] J. Ye; Z. Zhao; M. Wu, discriminative; "*K-Means for clustering*", proceeding of the annual conference, 2007

[۷] R. Varshavsky; D. itorn; M. Linial "*Cluster algorithm optimizer: A framework for large datasets*" ISBRA pp 85-96,springer 2007.

[۸] W. Tang; H. Xiong; S. Zhang; J. Wu; "*Enhancing semi-supervised clustering*", KDD07 california USA , ACM 2007

[۹] Yuntao Qian; Ching; Y. Suen "*Clustering Combination Method*", IEEE 2000

[۱۰] Yuntao Qian; Ching; Y. Suen "*Sequential combination methods for data clustering"* Journal of computer science, 2002

[۱۱] A. Raftery; N. Dean "*Variable selection for model based clustering*", journal of American statistical association, 101(473):168-178, 2006

[۱۲] Pankaj K. Agarwal; Nabil H. Mustafa "*K-Means Projective clustering*", PODS Paris France, ACM 2004