خوشه بندی داده های بانک قرض الحسنه مهر ایران با به کارگرفتن الگوریتم ترکیبی رقابت استعماری و خوشه بندی سلسله مراتبی

عباس محسنی[1]، فاطمه سعادت جو(نویسنده مسئول) [2]

۱.دانشجو،گروه کامپیوتر، واحد بندرعباس ، دانشگاه آزاد اسلامی ، بندرعباس ، ایران

۲.استادیار ،گروه کامپیوتر، واحدیزد ، دانشگاه علم و هنر یزد ،یزد ، ایران

چکیده

اگر چه، این حجم عظیم داده واقعاً برای مردم و شرکت ها می تواند مفید می باشد، اما مشکل ساز نیز است. مشکلی که در راستای این پیشرفت وجود دارد ، آنالیز و تجزیه و تحلیل داده های بزرگ است. با استفاده از تکنیک های داده کاوی می توان اطلاعات مفید و روابط پنهان میان داده ها را استخراج کرد. روش های سنتی داده کاوی به علت سرعت پایین ، نمی توانند بطور مستقیم برروی داده های بزرگ اجرا شوند و ما باید به دنبال راه حلی باشیم که بتوانیم با آنها داده های بزرگ را مورد تجزیه و تحلیل قرار دهیم. در این پژوهش خوشه بندی داده های بزرگ بانکی با استفاده از الگوریتم خوشه بندی سلسله مراتبی و رقابت استعماری بررسی شده و نتایج حاصل از آن با برخی از روش های موجود در این زمینه مقایسه شده است. . الگوریتم پیشنهادی روی چند پایگاه داده مورد بررسی قرار گرفته اند و نتایج مطلوبی را شامل شده اند نتایج در این پژوهش نشان میدهد، ارزیابی الگوریتم پیشنهادی نشان می دهد که روش پیشنهاد شده در این پژوهش نسبت به برخی از روش های موجود در این زمینه پاسخ های بهتری داشته و عملیات خوشه بندی را بصورت بهتر و با سرعت و دقت بالاتر و زمان اجرای کمتری نسب به سایر روش ها انجام می دهد.

کلمات کلیدی: رقابت استعماری ، داده های بزرگ بانکی ، سر خوشه ،روش های خوشه بندی، سلسله مراتبی

۱. مقدمه

همانطور که میدانیم همه افراد نیاز به ذخیره سازی اطلاعات دارند که اطلاعات خود را در محلی ذخیره کنند و طبقه بندی این داده و دسترسی به انها بسیار مهم میباشد . در در دنیای امروزی افراد داده های مختلفی را ذخیره کرده و بعدا از آن ها استفاده می کنند. ذخیره و بازیابی این حجم از داده ها چالش ها و مشکلات خاص خود را ایجاد می کند. یکی از داده هایی که ذخیره سازی آن باید با سرعت و دقت بالایی انجام شود داده های بانکی هستند. از آنجایی که تمامی افراد حساب بانکی دارند، لذا حجم این داده ها به شدت بزرگ بوده و مدیریت آن نیازمند استفاده از روش های مناسب و کارا می باشد. در این پژوهش خوشه بندی داده های بانکی با استفاده از الگوریتم خوشه بندی سلسله مراتبی و رقابت استعماری بررسی شده و نتایج حاصل از آن با برخی از روش های موجود در این زمینه مقایسه شده است. در عصر دیجیتالی کنونی، سرعت در تغییر حجم و تنوع داده ها تفاوت چشمگیری نسبت به دهه های قبل داشته است. با توجه به پیشرفت و توسعه ی زیاد اینترنت و تکنولوژی های جهانی آن لاین از قبیل سرورهای قدرتمند و بزرگ داده، هر روز با حجم عظیمی از اطلاعات و داده ها از منابع و سرویس های متفاوت روبرو می شویم که در دهه های گذشته وجود نداشت[(دانگ و سیرواستاوا*،۲۰۱۳) و (تانگ،کانگ†،۲۰۱۳)].

۲.بیان مساله

با توجه به اینکه ،حجم عظیم داده واقعاً برای مردم و شرکت ها می تواند مفید می باشد، اما مشکل ساز نیز خواهد بود، مساله ای که در این زمینه وجود دارد ، آنالیز و تجزیه و تحلیل داده های بزرگ است. با استفاده از تکنیک های داده کاوی می توان اطلاعات مفید و روابط پنهان میان داده ها را استخراج کرد. روش های سنتی داده کاوی به علت سرعت پایین ، نمی توانند بطور مستقیم برروی داده های بزرگ اجرا شوند و ما باید به دنبال راه حلی باشیم که بتوانیم با آنها داده های بزرگ را مورد تجزیه و تحلیل قرار دهیم(شیرخورشیدی،آقا بزرگی،واه،هروان ،۲۰۱۴).

امروزه حجم انبوه داده‌ها در بسیاری از کاربردهای داده محور، ابزار و روش های تحلیل و مدیریت داده‌ها را نیازمند تغییر نموده است. داده‌های حجیم، اصطلاحی برای مجموعه داده‌های بسیار بزرگ است که از نظر ساختار، پیچیدگی و منابع تولید بسیار متنوع هستند و ذخیره و تحلیل آنها کار پیچیده‌ای است. در دهه‌های اخیر یافتن الگوهای مفید در مجموعه های داده بزرگ بسیار مورد توجه می باشد.

* X. L. Dong and D. Srivastava
† H. Tong and U. Kang

خوشه‌بندی ابزاری قوی جهت پردازش داده‌های تولیدشده توسط برنامه‌های مختلف می‌باشد. این تکنیک به عنوان یکی از روش‌های بدون نظارت تشخیص الگوهای پنهان شناخته می‌شود.در واقع از مسائل مهم و بسیار مورد توجه در مجموعه‌های داده بزرگ، شناسایی خوشه‌ها یا نواحی دارای جمعیت متراکم در مجموعه داده چند بُعدی می‌باشد. در زمینه خوشه‌بندی داده‌های حجیم تکنیک های مختلفی وجود دارند و الگوریتم های مختلفی توسعه داده شده‌اند. یافتن الگوریتم خوشه‌بندی مناسب با بهینه ترین خوشه ها، در مدت زمان معقول از چالش‌های مهم در این حوزه می باشد. برای ارائه الگوریتم با کیفیت مسائلی چون بهینه بودن، عدم افتادن در بهینه محلی و مقاومت در برابر برون نهشت‌ها، به عنوان ویژگی های عمومی الگوریتم خوشه بندی، می‌باید مورد توجه قرار گیرد(شیرخورشیدی،آقا بزرگی،واه،هروان ،۲۰۱۴).

هدف این پژوهش حل مساله ی تجزیه و تحلیل داده های با حجم بالا و داده های بانکی برای خوشه بندی مربوط به بانک قرض الحسنه مهر ایران خواهد بود، که نیاز به تجزیه و تحلیل و برررسی دارند ما در این پژوهش از روش خوشه بندی داده های بانکی با حجم بالای با استفاده از الگوریتم خوشه بندی سلسله مراتبی و الگوریتم رقابت استعماری استفاده میکنیم و با روش های انجام شده در این زمینه مقایسه خواهیم کرد انتظار داریم روش ما کاآمد تر از روش های موجود عمل کند. بنابراین تکنیک های خوشه بندی کمک می کند تا کیفیت خوبی از خوشه ها و خلاصه سازی ها بدست آید .

در خوشه بندی سلسله مراتبی، خوشه ها به عنوان یک درخت نمایش داده می شود که دندروگرام نامیده می شود. این الگوریتم ها می توانند با بالا- پایین (تقسیم کننده) یا پایین- بالا (جمع کننده) باشند.این الگوریتم ها به پارامتر حد آستانه نیاز دارند که تا هنگام توقف جستجوی زیرگروه ها را اعلام کند[(دن،زانگ ،۲۰۱۵)،(کومار،بیزدک،پلانسماوی،راج اسکرار ،لیسکی،هاونس ،۲۰۱۵)]. در خوشه بندی سلسله مراتبی تقسیم کننده، این الگوریتم از خوشه های سراسری شروع می شود که شامل همه ی عناصر است و سپس داده ها به زیرخوشه ها تقسیم می شوند. باید مشخص شود که کدام خوشه به دو قسمت تقسیم می شود و چگونه تقسیم انجام می شود. در حالیکه در خوشه بندی سلسله مراتبی جمع کننده، این الگوریتم از یک خوشه شروع می شود و سپس هر دو خوشه با یکدیگر ادغام می شوند تا یک خوشه ی سراسری حاصل شود(ساجانا ،رانی،نارایانا ،۲۰۱۶).

۳. اهداف تحقیق

اهداف این تحقیق به شرح زیر می‌باشند:

۱.خوشه بندی داده های بزرگ بانکی با استفاده از روش خوشه بندی سلسله مراتبی و الگوریتم رقابت استعماری برای دسته بندی اصولی داده ها

۲.خوشه بندی داده ها از روش خوشه بندی سلسله مراتبی و الگوریتم رقابت استعماری جهت دسترسی سریع به داده های بانکی

۴.چند مورد کار که در این زمینه توسط افراد ایرانی انجام شده

یکی از مهمترین اعمال در داده کاوی خوشه بندی داده های موجود در یک دیتاست می باشد. این تکنیک به دنبال کشف ساختارهایی بوده که منجر به گروه بندی نمونه های موجود در یک پایگاه داده بوده، به گونه ای که نمونه های مشابه درون دسته هایی که بیشترین شباهت را با هم داشته قرار گرفته، در حالی که دارای تفاوتی قابل قبول با نمونه های سایر گروه ها داشته باشند. الگوریتم های خوشه بندی را می توان به چند دسته کلی الگوریتم های خوشه بندی تفکیکی، سلسله مراتبی، مبتنی بر چگالی و مبتنی بر گرید تقسیم نمود. این تحقیق ابتدا به مرور روش های خوشه بندی مطرح پرداخته و چند الگوریتم از هر روش را معرفی کرده است. در ادامه همچنین به معرفی روش های دیگری از جمله روش های خوشه بندی مبتنی بر الگوریتم های فرا ابتکاری و خوشه بندی های آنلاین پرداخته شده و سپس چالش های موجود در خوشه بندی از قبیل انتخاب تعداد بهینه خوشه ها، کاهش ابعاد، مدیریت داده های پرت و روشهای مقابله با آنها معرفی شده اند. در نهایت روش ها و الگوریتم های موجود از نظر برخی از پارامترها مقایسه شده و به بررسی مزایا و معایب هر الگوریتم پرداخته شدهاست(شاکری, محمود و محمد عبدالهی، ۱۳۹۴).داده کاوی شامل فرایند دانش نهفته از میان انبوهی از اطلاعات می باشد. این دانش می تواند ملاک تصمیم گیری های آتی و عملکردهای سیستم باشد. در سالهای اخیر استفاده از تکنیکهای دادهکاوی و الگوریتمهای هوشمند فراگیر شده است .بسیاری از کارهایی که در گذشته با صرف هزینه های زمانی و مالی فراوان حاصل شدهاند، میتوانند توسط این تکنیکها و الگوریتمها انجام گردند. یکی از مهمترین تکنیک های داده کاوی خوشه بندی می باشد. در روش های خوشه بندی با استفاده از معیارهای تشابه و عدم تشابه خوشه بندی داده ها انجام می پذیرد .اغلب داده های درون یک خوشه دارای بیشترین شباهت می باشند؛ درحالیکه میان خود خوشه ها تفاوتهای معنی داری وجود دارد. الگوریتم های متفاوتی برای خوشه بندی ارائه شده است که هر کدام دارای نقاط ضعف و قوت مختص به خود می باشند. در این مقاله مباحث مربوط به خوشه بندی داده ها مطرح شده و چند الگوریتم مشهور خوشه بندی بررسی و با استفاده از معیارهای مختلف مورد ارزیابی قرار می گیرند(بهروزیان نژاد, ابراهیم؛ محمد بهروزیان نژاد و شادی افتخار، ۱۳۹).امروزه دسته بندی داده ها یکی از پر کاربردترین زمینه های مطالعاتی در رشته های گوناگونی نظیر مهندسی، پزشکی، بیولوژی و داده کاوی می باشد. روش های تکاملی از قبیل ذرات، ژنتیک و رقابت استعماری و ترکیب آنها نیز برای حل اینگونه مسائل بکار می روند. در این مقاله به معرفی سیاست جذب جدیدی پرداخته می شود که برای مسائل خوشه بندی بسیار مفید است. در الگوریتم رقابت استعماری ، راه حل کاندید جدید با حرکت راه حل قدیمی بسمت یکی از امپراطوری های انتخاب شده با یک شعاع یکنواخت، تولید می شود. این روش حذب باعث می شود الگوریتم بخوبی از عهده استخراج برنیاید. هدف از این تحقیق معرفی سیاست جذبی است که بتواند

بخوبی از عهده خوشه بندی مسائل مختلف برآید. هدف آن است که تاریخچه ی حرکت هر مستعمره، روش حرکت کنونی اش را تعیین کند. سیاست جذب شامل ترکیب استراتژی Kmeans و Pivot است . الگوریتم رقابت استعماری جدید روی ۵ پایگاه داده مورد بررسی قرار گرفته اند و نتایج مطلوبی را شامل شده اند(جالسیان, حدیث و مهدی یعقوبی، ۱۳۹۲)

۵.الگوریتم پیشنهادی

در این بخش به تشریح الگوریتم پیشنهادی میپردازیم. برای اینکه بتوانیم روش پیشنهادی را مورد بررسی قرار دهیم ابتدا از تعدادی دیتاست های استاندارد موجود در زمینه های بانکی و پزشکی و آموزشی میپردازیم و الگوریتم پیشنهادی طبق این داده ها بررسی میگردد.روش کار به این صورت میباشد که در ابتدا الگوریتم پیشنهادی دیتاست مورد نظر را دریافت کرده و سپس با استفاده از خوشه بندی سلسله مراتبی و رقابت استعماری اقدام به خوشه بندی داده ها می نماید. همانطور که میدانیم در خوشه بندی ما در اولین مرحله از الگوریتم پیشنهادی مسئله مورد نظر (بانکی،پزشکی،آموزشی)به بخش های کوچکتر شکسته میشود و سپس خوشه بندی داده ها شروع می گردد. عملیات خوشه بندی باید به گونه ای انجام شود که فاصله درون خوشه ای کم و فاصله مابین خوشه ها ماکزیمم گردد. داده های موجود در دیتاست ها مورد استفاده بصورت تصادفی به دو بخش تقسیم می شوند که از یک بخش برای آموزش سیستم و از بخش دیگر برای تست آن استفاده می شود. هنگام آموزش بهترین مراکز خوشه تعیین می شوند. سپس در هنگام تست بررسی می شود که آیا مراکز خوشه تعیین شده به درستی انتخاب شده اند یا خیر. اگر خوشه بندی دیتا ست ها بصورت صحیح تعیین شده باشند پس از ورود داده های تست نباید مقدار فاصله درون خوشه ای تغییرات زیادی داشته باشد که در این صورت خوشه بندی به درستی انجام شده است ولی اگر مراکز خوشه به درستی تعیین نشده باشند آنگاه فرایند خوشه بندی بصورت مجدد انجام می گیرد تا مراکز خوشه مناسب تعیین شوند و به خوشه بندی مناسب دست یابیم.

- تقسیم داده به دو قسمت آموزشی و تست اصلی خوشه بندی

برای بررسی الگوریتم پیشنهادی داده ها مورد استفاده که داده های بانکی ،آموزشی یا هر نوع داده دیگر که باشد را به دو بخش آموزش و تست تقسیم می کنیم. از داده های آموزش برای آموزش سیستم و از داده های تست برای بررسی کیفیت آموزش استفاده می شود. در انتخاب داده های آموزش و تست هیچ گونه عملیات گزینشی انجام نشده و انتخاب بصورت تصادفی انجام گرفته است. از کل داده های موجود در دیتاست، هشتاد درصد از داده ها به عنوان داده آموزش و مابقی به عنوان داده های تست در نظر گرفته شده اند. علت استفاده از هشتاد درصد داده ها به عنوان داده آموزش این است که در اکثر مقالات مورد مطالعه به این زمینه از این مقدار از داده ها برای آموزش سیستم استفاده شده است ولی به راحتی می توان این مقدار را تغییر داده و نسبت داده های آموزش به تست را کم یا زیاد نمود. در صورتی که

تعداد داده های آموزش کم باشد، یا به عبارتی درصد کوچکی از داده های موجود در دیتاست را به بحث آموزش سیستم اختصاص دهیم ممکن است فرایند آموزش به خوبی انجام نشده و خروجی های حاصل از آموزش سیستم بهینه نباشند در این صورت می توان با تکرار فرایند آموزش به نتایج بهینه دست یافت. بدیهی است که تکرار فرایند آموزش باعث می شود تا این فرایند زمان زیادی به طول انجامد که این مسئله باعث کند شدن الگوریتم پیشنهادی خواهد شد. علاوه بر این هزینه اجرای الگوریتم پیشنهادی را نیز افزایش می دهد. به همین منظور معمولا تعداد داده های آموزش را بیشتر از تعداد داده های تست در نظر می گیرند تا به تکرار چند نسل از الگوریتم پیشنهادی، نتایج مطلوب حاصل شده و نیاز به آموزش بیشتر سیستم مرتفع شود.

- تعیین جمعیت اولیه

معمولا هر یک از الگوریتم های بهینه سازی نیازمند استفاده از یک جمعیت اولیه هستند که عملیات بهینه سازی را با توجه به آن شروع کنند. در الگوریتم پیشنهادی این پژوهش جمعیت اولیه بصورت مجموعه بردارهایی شامل جواب احتمالی تعریف می شوند. طول هر یک از این بردار ها برابر با تعداد داده های موجود تقسیم بر تعداد خوشه ها می باشد و تعداد بردارها بسته به نظر کاربر و در قالب یک سوال در طول اجرای الگوریتم پیشنهادی پرسیده خواهد شد. تعداد خوشه ها با توجه به نوع دیتاست و توضیحات مربوط به دیتاست مشخص می شود. در ابتدای کار داده های موجود در هر بردار مقدار دهی می شوند و با اجرای الگوریتم بهینه سازی و تغییر هر نسل*، بردار بهینه شناسایی و مابقی بردار های غیر بهینه حذف خواهند شد. هر عنصر بردار شامل فاصله اقلیدسی داده مورد نظر تا سایر داده های موجود در جمعیت اولیه می باشد. این فاصله با استفاده از مقدار هر یک از داده ها و اندیس قرار گیری آن در بردار مشخص می شود. ساختار بردار ها را با استفاده از یک مثال تشریح می کنیم. فرض کنید دیتاستی شامل یکصد داده مختلف وجود دارد که باید در دو خوشه بخش بندی شود. ضمن فرض می کنیم که تعداد داده های موجود در هر یک از دو خوشه با هم برابر باشند. در این صورت طول هر بردار را برابر پنجاه در نظر می گیریم و به ازای هر داده موجود در دیتاست برداری را می سازیم که الگوریتم پیشنهادی با توجه به مقادیر آن عملیات بهینه سازی را انجام می دهد.

۶.بررسی داده های تست

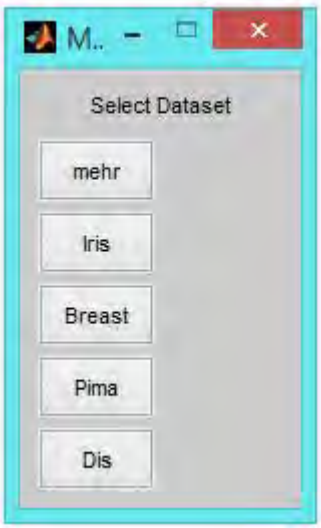
مراحل آموزش الگوریتم پیشنهادی در بخش های قبلی تشریح شد. برای ارزیابی کیفیت آموزش الگوریتم پیشنهادی داده های تست را به سیستم اعمال می کنیم. با اعمال این داده ها احتمال وقوع دو حالت مختلف وجود دارد. اولین حالت بهینه بودن کار را نتیجه می دهد که در این صورت فرایند اجرای الگوریتم پیشنهادی خاتمه یافته و نتایج بهینه تولید شده اند. این حالت زمانی رخ می دهد که با اعمال داده های تست به الگوریتم پیشنهادی، مراکز خوشه تغییر نکرده و اختلاف زیادی در مراکز خوشه

* Iteration

مشاهده نمی شود. حالت دوم زمانی رخ می دهد که با اعمال دادهای تست به الگوریتم پیشنهادی، مراکز خوشه تغییر کرده و نیاز به تعیین مراکز خوشه جدید و تعیین سرخوشه های جدید احساس شود. در این حالت می توان بیان نمود که آموزش الگوریتم پیشنهادی به نحو صحیح انجام نگرفته و لازم است تا فرایند آموزش بطور مجدد طی شود.

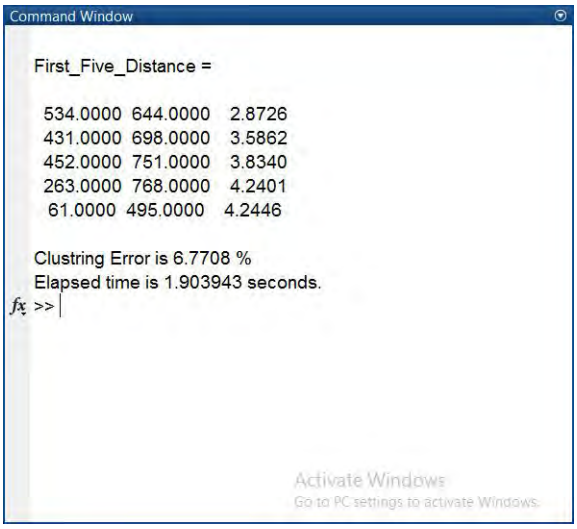
۷.شبیه سازی الگوریتم پیشنهادی

در بخش های قبلی ساختار الگوریتم پیشنهادی و دیتاست های مورد استفاده در آن تشریح شدند. در ادامه به شبیه سازی الگوریتم پیشنهادی با استفاده از متلب پرداخته و نتایج حاصل از آن را نمایش می دهیم. شکل (۲) پنجره انتخاب دیتاست مورد نظر را نمایش می دهد. با هر بار اجرای الگوریتم ارائه شده در این تحقیق پنجره موجود در شکل (۲) نشان داده شده و امکان انتخاب دیتاستی که قصد داریم با استفاده از روش الگوریتم رقابت استعماری و خوشه بندی سلسله مراتبی خوشه بندی آن را انجام دهیم فراهم می کند.



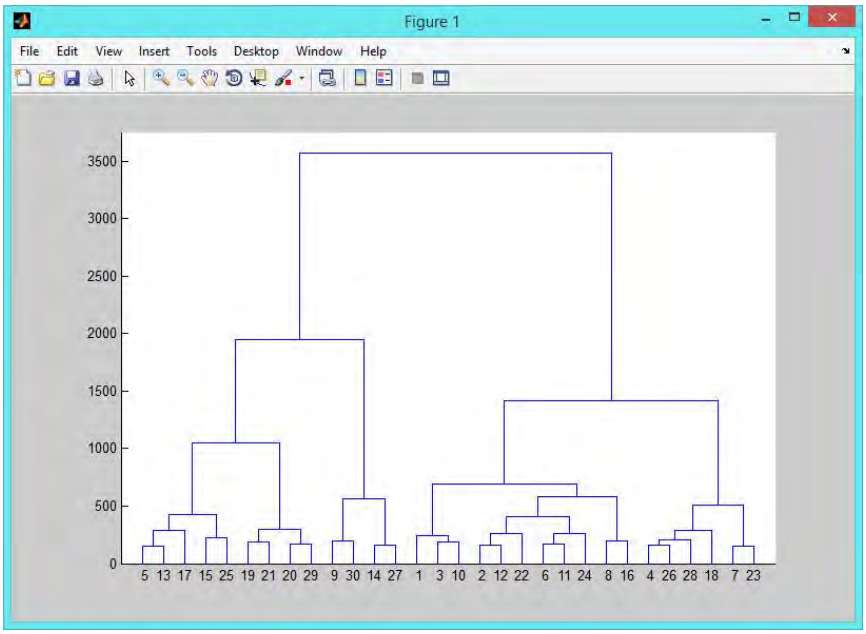
شکل ۲- پنجره انتخاب دیتاست

انتخاب دکمه mehr باعث خوشه بندی داده ها با استفاده از الگوریتم رقابت استعماری و خوشه بندی سلسله مراتبی خواهد شد و همچنین انتخاب iris باعث می شود تا فرایند خوشه بندی برای دیتاست iris انجام شود و انتخاب دکمه های breast و pima خوشه بندی را به ترتیب برای دیتاست های breast و pima انجام می دهند. پس از تعیین نوع دیتاست فرایند خوشه بندی آغاز شده و فاصله اقلیدسی داده ها را محاسبه می کند. شکل (۴-۳) فاصله اقلیدسی بین پنج داده اول موجود در دیتاست mehrرا نشان می دهد.

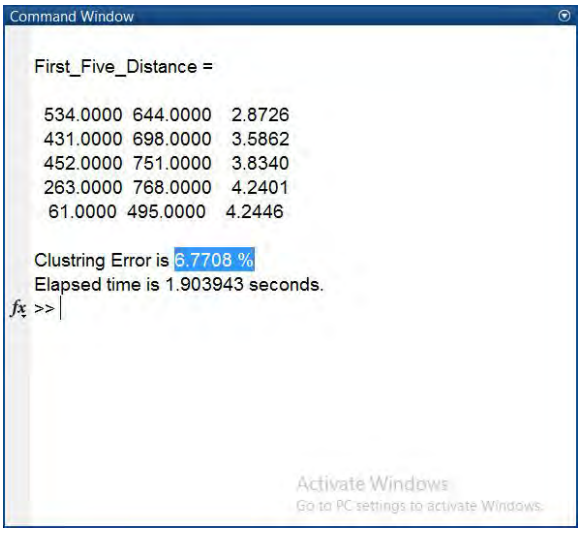


شکل ۳- فاصله اقلیدسی بین پنج داده اول موجود در دیتاست mehr

پس از آن تعداد خوشه های لازم جهت خوشه بندی داده ها محاسبه شده و عملیات خوشه بندی با توجه به شرایطی که در بخش های قبلی گفته شد انجام می گردد. با تکمیل فرایند خوشه بندی باید درخت حاصل از انجام این عملیات نمایش داده شود. شکل (۴) درخت ناشی از خوشه بندی دیتاست mehr را نشان می دهد. محور عمودی این شکل معرف معیار خوشه بندی یا همان فاصله اقلیدسی است. در صورتیکه تعداد داده های موجود در دیتاست کمتر از سی داده باشد محور افقی داده های گفته شده را نشان می دهد اما اگر تعداد داده های بیشتر از سی باشد محور افقی ارزش برگ ها را با استفاده از روالی که در قبل گفته شد، محاسبه کرده در محور افقی نشان می دهد.پس از نمایش درخت و خوشه بندی آن باید مقدار خطای خوشه بندی مشخص گردد. بدیهی است که کم بودن مقدار خطا نشان دهنده بهینه بودن الگوریتم پیشنهادی بوده و توانایی آن در خوشه بندی صحیح اطلاعات را نشان می دهد. مقدار خطای ناشی از خوشه بندی در شکل (۵) نشان داده شده است. پس از نمایش مقدار خطا باید پراکندگی داده ها بر حسب گروهی که داده در آن قرار می گیرد نشان داده شود. شکل (۴-۶) پراکندگی داده های موجود در دیتاست mehr را نشان می دهد. محور افقی شکل (۴-۶) طول کاسبرگ و محور عمودی آن عرض کاسبرگ های گل زنبق را نمایش داده و ملاک دسته بندی نیز نوع گل می باشد.



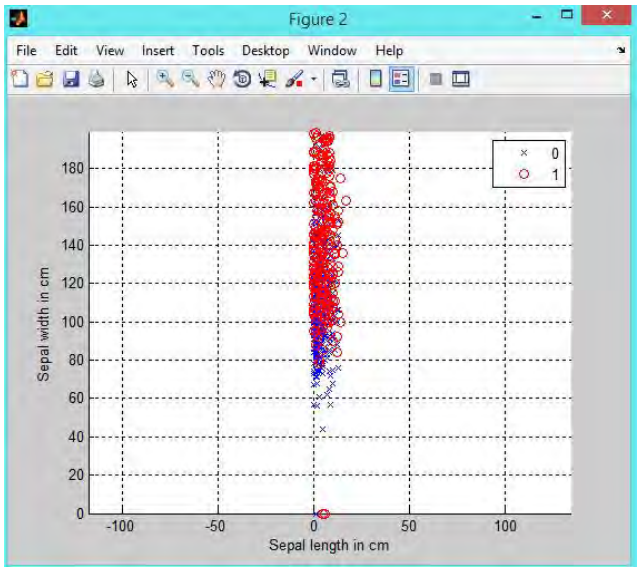
شکل ۴-درخت ناشی از خوشه بندی دیتاست mehr



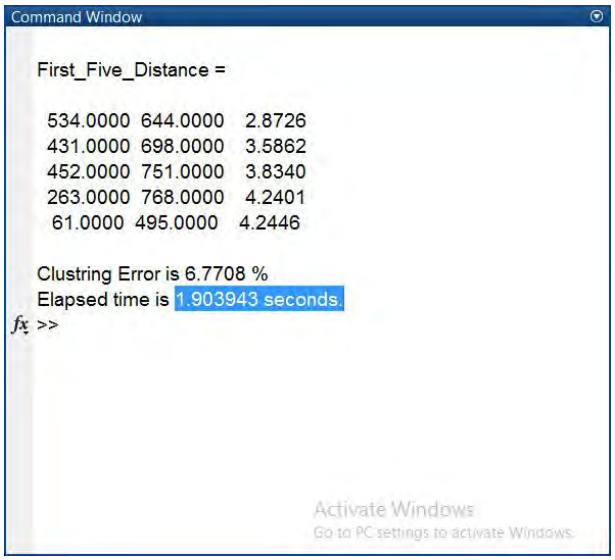
شکل ۵- نمایش مقدار خطای ناشی از خوشه بندی دیتاست mehr

انجام هر گونه عملیاتی در کامپیوتر مستلزم صرف مقداری هزینه زمانی است. بدیهی است که الگوریتم هایی که زمان کمتری را صرف پردازش اطلاعات می کنند سریع تر بوده و در نزد کاربران از محبوبیت بیشتری برخوردارند. شکل (۶) زمان لازم جهت خوشه بندی با استفاده از الگوریتم رقابت استعماری و خوشه بندی سلسله مراتبی دیتاست mehr را بر حسب واحد ثانیه نمایش می دهد. این ارزیابی زمانی برای اجرای الگوریتم پیشنهادی بر ماشینی است که دارای چهار گیگا بایت حافظه Ram بوده و از پردازنده intel(R) Core(TM) i7-3687Uاستفاده می کند. ماشین مورد نظر از سیستم عامل

ویندوز 8.1 استفاده کرده و شبیه سازی الگوریتم پیشنهادی با استفاده از نرم افزار متلب نسخه 2017 انجام شده است.



شکل ۶- نمایش پراکندگی طول و عرض کاسبرگ در دیتاست mehr



شکل ۷- زمان لازم جهت خوشه بندی دیتاست mehr

۸- معرفی ملاک های ارزیابی

برای ارزیابی الگوریتم پیشنهادی ملاک های مختلفی وجود دارد که می توان از آن ها استفاده نمود. یکی از ملاک های موجود جهت ارزیابی الگوریتم پیشنهادی بررسی مقدار خطای ناشی از خوشه بندی داده ها است. ملاک دیگری که می توان از آن جهت ارزیابی الگوریتم پیشنهادی استفاده نمود زمان مورد نیاز جهت انجام خوشه بندی می باشد. بدیهی است که هرچه مقدار این زمان کمتر باشد، سرعت اجرا بیشتر

بوده و بهینه بودن الگوریتم پیشنهادی مشهود تر خواهد بود. ملاک های دیگری نیز برای ارزیابی و مقایسه الگوریتم پیشنهادی و سایر الگوریتم های موجود در این زمینه وجود دارند که از بررسی آن ها صرف نظر می کنیم. در ادامه ابتدا تاثیر تغییر در برخی از پارامترها و ویژگی های الگوریتم پیشنهادی بر زمان اجرای و درصد خطای آن را بررسی کرده و در نهایت الگوریتم پیشنهادی را با برخی از الگوریتم های موجود در زمینه خوشه بندی داده ها استفاده مقایسه می کنیم.

۹. نتایج حاصل از الگوریتم پیشنهادی

نتایج حاصل از شبیه سازی الگوریتم پیشنهادی برای دیتاست mehr در قسمت قبلی ارائه شدند اما می توان ملاک های ارزیابی را برای سایر دیتاست ها نیز بررسی کرده و نتایج را مشاهده نمود. در ادامه ملاک های ارزیابی را برای خوشه بندی سایر دیتاست های موجود بررسی کرده و در نهایت الگوریتم پیشنهادی را با سایر الگوریتم های موجود در این زمینه مقایسه می کنیم. جدول (۴-۴) زمان اجرا و مقدار خطای خوشه بندی را برای هر یک از دیتاست های موجود در الگوریتم پیشنهادی نمایش می دهد.

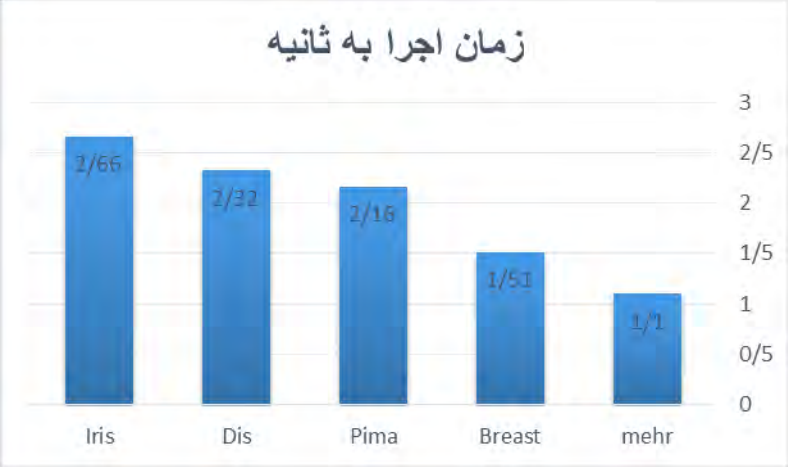
جدول ۱- زمان اجرا و مقدار خطای خوشه بندی در دیتاست های مختلف

نام دیتاست	مقدار خطا به درصد	زمان اجرا به ثانیه
mehr	**1.10333**	۱,۹۰۱۰۹۶
Breast	**1.5177**	۲,۲۱۷۱۴۴
Pima	**2.1608**	**2.712603**
Dis	**2.32333**	**3.907551**
Iris	**2.6667**	**2.028409**

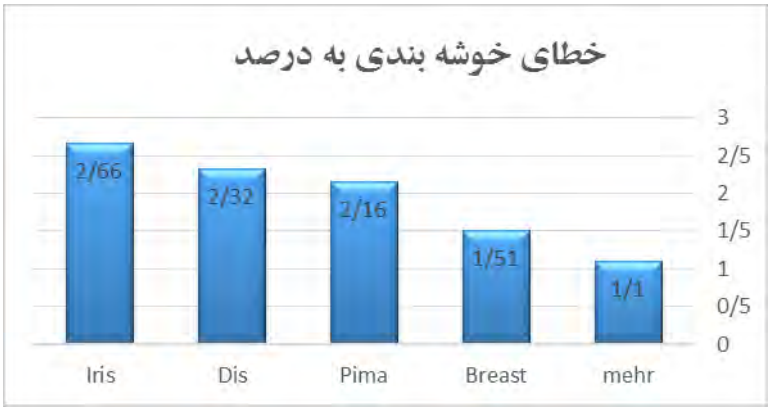
بررسی جدول (۱) نشان می دهد استفاده از دیتاست Iris بیشترین خطا را در خوشه بندی ایجاد نموده است و این در حالی است که دیتاست mehr علی رغم حجم داده های کلان موجود در خود کمترین خطا را ایجاد کرده است. از طرفی چون حجم داده های موجود در دیتاست mehr بسیار زیاد است بنابراین اجرای خوشه بندی با استفاده از الگوریتم رقابت استعماری و خوشه بندی سلسله مراتبی با توجه به سرعت الگوریتم پیشنهادی در آن به زمان قابل قبولی را برای اجرای در الگوریتم پیشنهادی ارایه داده است این نتیجه بر اساس چندین دفعه اجرا به دست امده است با توجه به حجم سنگین داده ممکن است زمان بیشتری برای اجرای الگوریتم نیاز باشد که علت این مسئله حجم زیاد پردازش مورد نیاز جهت اجرای عملیات خوشه بندی می باشد. خطای موجود در هر دیتاست به علت پراکندگی داده های موجود در آن ایجاد می شود. وجود پراکندگی باعث می شود تا دقت خوشه بندی کاهش یافته و مقدار خطا

افزایش یابد. بررسی داده های موجود در دیتاست mehr پراکندگی داده ها را نشان داده و در نتیجه در جدول (۲) بیشترین خطای ناشی از خوشه بندی را نمایش داده است. در صورتیکه داده ها پراکندگی کمتری داشته باشند فاصله اقلیدسی که برای هر داده محاسبه می شود کاهش یافته و در نتیجه انسجام خوشه ها افزایش می یابد که این انسجام باعث کاهش مقدار خطا خواهد شد. نتایج حاصل از جدول (۸) در شکل های (۹) و (۱۰) بصورت گرافیکی نشان داده شده اند.

تا اینجا دیتاست های مختلفی را با استفاده از الگوریتم پیشنهادی خوشه بندی کرده ایم در ادامه الگوریتم پیشنهادی این پژوهش را با دو الگوریتم دیگری که در این زمینه وجود دارند مقایسه می کنیم. اولین الگوریتمی که برای مقایسه الگوریتم پیشنهادی این تحقیق از آن بهره می گیریم روشی است که خوشه بندی داده های بزرگ را با استفاده از شبکه عصبی انجام می دهد. این الگوریتم از دیتاست های Dis و Breast استفاده کرده و نتایج حاصل از آن در ادامه ارائه خواهد شد. الگوریتم دیگری که نتایج آن را با روش پیشنهادی این پژوهش مقایسه می کنیم روشی است که خوشه بندی داده ها را با استفاده از ترکیب روش عصبی و الگوریتم ممتیک انجام می دهد. تمامی الگوریتم ها از دیتاست مشابه استفاده کرده و بر روی ماشین یکسانی اجرا شده اند.



شکل ۸- مقایسه زمان اجرای الگوریتم پیشنهادی در دیتاست های مختلف



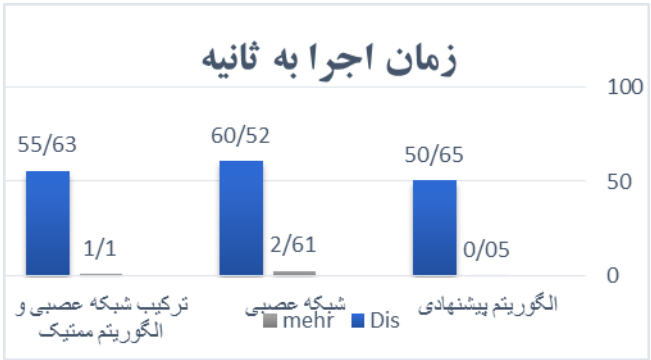
شکل ۹- مقایسه خطای خوشه بندی مربوط به الگوریتم پیشنهادی در دیتاست های مختلف

نام دیتاست	الگوریتم پیشنهادی		شبکه عصبی		ترکیب شبکه عصبی و الگوریتم ممتیک	
	مقدار خطا به درصد	زمان اجرا به ثانیه	مقدار خطا به درصد	زمان اجرا به ثانیه	مقدار خطا به درصد	زمان اجرا به ثانیه
mehr	**1.1745**	**0.050382**	**14.7363**	**2.61455**	**1.27462**	**1.10172**
Dis	**0.320001**	**50.655802**	**12.62468**	**60.5215**	**0.17573**	**55.6347346**

جدول (۲) ملاک های ارزیابی را در الگوریتم های مختلف مقایسه می کند. داده های موجود در این جدول میانگین نتایج حاصل از یکصد بار اجرای الگوریتم های گفته شده می باشند و مقدار خطای نمایش داده شده در این جدول به درصد بوده و زمان اجرا به واحد ثانیه سنجیده شده است.

بررسی جدول (۲) نشان می دهد الگوریتم پیشنهادی هم از لحاظ مقدار خطا و هم از لحاظ زمان اجرا نسبت به روشی که از شبکه عصبی استفاده می کند بهینه است ولی در صورتیکه شبکه عصبی با استفاده از الگوریتم ممتیک بهینه شود نسبت به الگوریتم پیشنهادی مقدار خطای کمتری را خواهد داشت ولی باز هم سرعت اجرای الگوریتم پیشنهادی بهتر می باشد. علت این مسئله بهینه سازی ناشی از الگوریتم ممتیک است و این در حالی است که استفاده از الگوریتم ممتیک مستلزم انجام حجم زیادی از پردازش است که این مسئله باعث کند شدن آن می گردد. نتایج موجود در جدول (۴-۴) بصورت گرافیکی در شکل های (۱۰) و (۱۱) قابل مشاهده هستند.

جدول (۲) مقایسه روش های مختلف خوشه بندی داده ها



شکل ۱۰- مقایسه زمان اجرای الگوریتم های مختلف

شکل ۱۱- مقایسه مقدار خطای خوشه بندی در الگوریتم های مختلف

۱۰. جمع بندی

در این بخش به پیاده سازی الگوریتم پیشنهادی و استخراج خروجی های مساله بر اساس زمان اجرا و مقدار خطا به درصد پرداخته شد و در نهایت الگوریتم پیشنهادی با شبکه های عصبی و ترکیب شبکه عصبی و الگوریتم ممتیک مورد بررسی قرار گرفت. ما در این پژوهش ابتدا مقدمه ای بر خوشه بندی بیان شده و انواع مدل های خوشه بندی به اختصار معرفی شده اند. سپس برخی از الگوریتم های موجود و کاربرد آنها را بیان کردیم سپس در فصل دوم این پژوهش راجع به مسایل کلی خوشه بندی و تکنیک های خوشه بندی ،الگوریتم های خوشه بندی و الگوریتم رقابت استعماری و راهکار های گذشته خوشه بندی سلسله مراتبی در بانک ها ،خوشه بندی داده های پزشکی و آموزشی و .. مورد بررسی قرار گرفت بعد تشریح مباحث عمومی و راهکار گذشته نوبت به توضیح روش تحقیق خواهد رسید که ما اینکار را در فصل سوم انجام دادیم که در این فصل پژوهش روش خوشه بندی سلسله مراتبی و رقابت استعماری و ترکیب آنها به صورت کامل توضیح داده شد سپس مکانیزم و شرح عملیات موجود در آن بیان شده است.که بتوانیم در فصل چهارم الگوریتم پیاده سازی شده با متلب را تست نماییم همچنین ما اینکار را در فصل چهارم انجام دادیم که در این فصل روال کار الگوریتم پیشنهادی بطور کامل معرفی شده و سپس شبیه سازی الگوریتم پیشنهادی این پژوهش انجام گرفت و نتایج حاصل از شبیه سازی روش پیشنهادی این تحقیق نیز در این فصل ارائه شده است.در نهایت در فصل پنجم به جمع بندی نتایج پرداخته و پس از مقایسه نتایج حاصل از الگوریتم پیشنهادی با برخی از الگوریتم های موجود در این زمینه، پیشنهادات و ایده هایی را برای کارهای آتی ارائه دادیم.

۱۱. کارهای آینده

در این پژوهش خوشه بندی داده های با استفاده از روش سلسله مراتبی و رقابت استعماری بررسی شده و نتایج حاصل از این روش در قالب جداول و شکل های مختلف نمایش داده شد. ارزیابی الگوریتم پیشنهادی نشان می دهد که روش پیشنهاد شده در این پژوهش نسبت به برخی از روش های موجود در این زمینه پاسخ های بهتری داشته و عملیات خوشه بندی را بصورت بهتر و با سرعت بالاتر نسب به

کارهای قبلی انجام می دهد. هر چند الگوریتم پیشنهادی ارائه شده در این پژوهش فرایند خوشه بندی داده ها را بهبود داده است اما انتظار داریم احتمالا با استفاده از ترکیب خوشه بندی سلسله مراتبی و روش فازی و الگوریتم های بهینه سازی دیگر بتوان نتایج بهتری را بدست آورد. استفاده از روش فازی و الگوریتم های فرااکتشافی با توجه به قابلیت هایی که این روش در مواجهه به عدم قطعیت دارد می تواند مفید واقع شده و فرایند خوشه بندی داده ها را به صورت بهتری انجام دهند و کار را بهبود بخشند.

۱۲.منابع و مراجع

A. S. Shirkhorshidi, S. Aghabozorgi ,T. Y. Wah, and T. Herawan, "Big data clustering: a review," in *International Conference on Computational Science and Its Applications*, 2014, pp. 707-720: Springer.

A.-E. Hassanien, A. T. Azar, V. Snasel, J. Kacprzyk, and J. H. Abawajy, *Big data in complex systems: challenges and opportunities*. Springer, 2015.

A. Jain, V. Bhatnagar, and P. Sharma, "Collaborative and clustering based strategy in big data," *Collaborative filtering using data mining and analysis,* pp. 140-158, 2017.

A. Sehgal and A. K. Sharma, "Enhancement of Big Data Using Clustering Mechanism," *International Journal Of Scientific Research And Education,* vol. 5, no. 06, 2017.

B. Furht and F. Villanustre, "Introduction to Big Data," in *Big Data Technologies and Applications*: Springer, 2016, pp. 3-11.

B. Chun-Hsien and V. Honavar, "A neural-network architecture for syntax analysis," *IEEE Transactions on Neural Networks,* vol. 10, no. 1, pp.۱۹۹۹ ,۱۱٤-۹٤ .

C.-W. Tsai, C.-F. Lai, H.-C .Chao, and A. V. Vasilakos, "Big data analytics: a survey," *Journal of Big Data,* vol. 2, no. 1, p. 21, 2015.

C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big Data Analytics," in *Big Data Technologies and Applications*: Springer, 2016, pp. 13-52.

C. D. Kumar, J. C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, and T. C. Havens, "A hybrid approach to clustering in big data," 2015.

D. D. Kumar, J. C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, and T. C. Havens, "A hybrid approach to clustering in big data," *IEEE transactions on cybernetics,* vol. 46, no. 10, pp. 2372-2385, 2016.

E. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering,* vol. 60, no. 1, pp. 208-221, 2007

E. Januzaj, H.-P. Kriegel, and M. Pfeifle, "DBDC: Density based distributed clustering," in *International Conference on Extending Database Technology*, 2004, pp. 88-105: Springer.

E. R. Hruschka, R. J. Campello, and A. A. Freitas, "A survey of evolutionary algorithms for clustering," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 39, no. 2, pp. 133-155, 2009.

F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 2, no. 1, pp. 86-97, 2012.

G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer,* vol. 32, no. 8, pp. 68-75, 1999.

G. George, E. C. Osinga, D. Lavie, and B. A. Scott, "Big data and data science methods for management research," *Academy of Management Journal,* vol ,۵۹ .no. 5, pp. 1493-1507, 2016.

G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Information Fusion,* vol. 28, pp. 45-59, 2016.

G. Karypis, "METIS and ParMETIS," in *Encyclopedia of Parallel Computing*: Springer, 2011, pp. 1117-1124.

H. Karypis, K. Schloegel, and V. Kumar, "Parmetis," *Parallel graph partitioning and sparse matrix ordering library. Version,* vol. 2, 2003.

G. Andrade, G. Ramos, D. Madeira, R .Sachetto, R. Ferreira, and L. Rocha, "G-dbscan: A gpu accelerated algorithm for density-based clustering," *Procedia Computer Science,* vol. 18, pp. 369-378, 2013.

H. Tong and U. Kang, "Big Data Clustering," ed, 2013.

H. Felzmann, T. Beyan, M. Ryan, and O. Beyan, "Implementing an ethical approach to big data analytics in assistive robotics for elderly with dementia," *ACM SIGCAS Computers and Society,* vol. 45, no. 3, pp. 280-286, 2016

.

H. Kashyap, H. A. Ahmed, N. Hoque, S. Roy, and D .K. Bhattacharyya, "Big data analytics in bioinformatics: architectures, techniques, tools and issues,"

I. Ekbia *et al.*, "Big data, bigger dilemmas: A critical review," *Journal of the Association for Information Science and Technology,* vol. 66, no. 8, pp. 1523-1545, 2015.

I. Triguero, D. Peralta, J. Bacardit, S. García, and F. Herrera, "MRPR: a MapReduce solution for prototype reduction in big data classification," *Neurocomputing,* vol. 150, pp. 331-345, 2015.

J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM,* vol. 51, no. 1, pp. 107-113, 2008.

J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm gdbscan and its applications," *Data mining and knowledge discovery,* vol. 2, no. 2, pp. 169-194, 1998.

J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE transactions on neural networks,* vol. 11, no. 3, pp. 586-600, 2000.

K. Kothari and O. Kale, "'Survey of various clustering techniques for big data in data mining'," *Int. J. of Innovative Research In Technology,* vol. 1, no. 7, pp. 68-71, 2014

K. Shim, "MapReduce algorithms for big data analysis," *Proceedings of the VLDB Endowment,* vol. 5, no. 12, pp. 2016-2017, 2012.

L. Duan and Y. Xiong, "Big data analytics and business analytics," *Journal of Management Analytics,* vol. 2, no. 1, pp. 1-21, 2015.

L. Ma, L. Gu, B. Li, S. QIAO, and J. Wang, "G-DBSCAN: An improved DBSCAN clustering method based on grid," *Adv Sci Technol Lett,* vol. 74, pp. 23-28, 2014.

M. Alamuri, B. R. Surampudi, and A. Negi, "A survey of distance/similarity measures for categorical data," in *Neural Networks (IJCNN), 2014 International Joint Conference on*, 2014, pp. 1907-1914: IEEE.

M. Mas, M. Monserrat, J. Torrens, and E. Trillas, "A Survey on Fuzzy Implication Functions ", *IEEE Transactions on Fuzzy Systems,* vol. 15, no. 6, pp. 1107-1121, 2007.

Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 5, no. 1, p. 28, 2016.

P. K. Jana and A. Naik, "An efficient minimum spanning tree based clustering algorithm," in *Methods and Models in Computer Science, 2009. ICM2CS 2009. Proceeding of International Conference on*, 2009, pp. 1-5: IEEE.

R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural Networks, 1989. IJCNN., International Joint Conference on*, 1989, pp. 593-605: IEEE.

R. L. d. Mantaras and L. Godo, "From fuzzy logic to fuzzy truth-valued logic for expert systems: a survey," in *Fuzzy Systems, 1993., Second IEEE International Conference on*, 1993, pp. 750-755 vol.2.

S. Arora and I. Chana, "A survey of clustering techniques for big data analysis," in *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-*, 2014, pp. 59-65: IEEE.

Serra, P., A. F.Stanton, and S. Kais,pivot *method for global optimization.* The American Physical Society, 1997. **55**: p. 4.

S. Guha, R. Rastogi, and K. Shim, "CURE: an efficient clustering algorithm for large databases," in *ACM Sigmod Record*, 1998, vol. 27, no. 2, pp. 73-84: ACM.

S. V. Kalinin, B. G. Sumpter, and R. K. Archibald, "Big-deep-smart data in imaging for guiding materials design," *Nature materials,* vol. 14, no. 1 ,۰pp. 973-980, 2015.

S. Mitra and Y. Hayashi, "Neuro-fuzzy rule generation: survey in soft computing framework," *IEEE Transactions on Neural Networks,* vol. 11, no. 3, pp. 748-768, 2000.

S. Naz, H. Majeed, and H. Irshad, "Image segmentation using fuzzy clustering: A survey," in *Emerging Technologies (ICET), 2010 6th International Conference on*, 2010, pp. 181-18 ۲: IEEE.

T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: A new data clustering algorithm and its applications," *Data Mining and Knowledge Discovery,* vol. 1, no. 2, pp. 141-182, 1997.

T. Sajana, C. S. Rani, and K. Narayana, "A survey on clustering techniques for big data mining," *Indian Journal of Science and Technology,* vol. 9, no. 3, 2016.

T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," in *ACM Sigmod Record*, 1996, vol. 25, no. 2, pp. 103-114: ACM.

T. W. Cheng, D. B. Goldgof, and L. O. Hall, "Fast fuzzy clustering," *Fuzzy sets and systems,* vol. 93, no. 1, pp. 49-56, 1998.

V. Ganti, R. Ramakrishnan, J. Gehrke, A. Powell, and J. French, "Clustering large datasets in arbitrary metric spaces," in *Data Engineering, 1999. Proceedings., 15th International Conference On*, 1 ,۹۹۹pp. 502-511: IEEE.

W. Fan and A. Bifet, "Mining big data: current status, and forecast to the future," *ACM sIGKDD Explorations Newsletter,* vol. 14, no. 2, pp. 1-5, 2013.

W. Yan, U. Brahmakshatriya, Y. Xue, M. Gilder, and B. Wise, "p-PIC: Parallel power iteration clustering for big data," *Journal of Parallel and Distributed computing,* vol. 73, no. 3, pp. 352-359, 2013.

X. L. Dong and D. Srivastava, "Big data integration," in *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, 2013, pp. 1245-1248: IEEE.

X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *ieee transactions on knowledge and data engineering* ,vol. 26, no. 1, pp. 97-107, 2014.

X. Cai, F. Nie, and H. Huang, "Multi-View K-Means Clustering on Big Data," in *IJCAI*, 2013: Citeseer.

Yang, S.-d., Y.-l. Yi, and Z.-y. Shan.Gbest *guided Imperialist Competitive Algorithm for GlobalnNumerical Optimization*. in *Computer Distributed*n *Control and Intelligent Environmental Monitoring (CDCIEM), 2012 International Conference on*.2012.

Y. He, H. Tan, W. Luo, S .Feng, and J. Fan, "MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data," *Frontiers of Computer Science,* vol. 8, no. 1, pp. 83-99, 2014.

Z. Du, Y. Wang, and Z. Ji, "PK-means: A new algorithm for gene clustering," *Computational Biology and Chemistry,* vol. 32, no. 4, pp. 243-247, 200۸