مطالعات مهندسی صنایع و مدیریت تولید

دوره ۳، شماره ۳، پاییز ۱۳۹۶

صفحات ۳۷- ۲۴

Online ISSN: ۲۴۷۶-۴۴۵۰

Print ISSN: ۲۵۳۸-۲۱۰۱

www.irijournals.com

ارائه یک روش ترکیبی از الگوریتم ژنتیک برای استفاده در خوشه بندی داده ها

میثم رهنمای فلاح[۱]، مرضیه فریدی ماسوله[۲*]، محمدرضا عسگری پور[۳]

[۱و۲و۳] دانشگاه آزاد اسلامی، واحد الکترونیکی، گروه کامپیوتر، تهران، ایران

[*] نویسنده مسئول مکاتبات

چکیده

داده های بزرگ یا عظیم داده ترجمه اصطلاح Big Data می‌باشد که معمولا به مجموعه از داده‌ها اطلاق می‌شود که اندازه آنها فراتر از حدی است که با نرم افزارهای معمول بتوان آنها را در یک زمان معقول اخذ، دقیق‌سازی، مدیریت و پردازش کرد. مفهوم «اندازه» در داده های بزرگ بطور مستمر در حال تغییر است و به مرور بزرگتر می‌شود. از این رو با رشد روز افزون داده ها و نیاز به بهره برداری و تحلیل از این داده ها، بکارگیری زیرساخت‌های Big Data از اهمیت ویژه ای برخوردار شده است. ما به ارائه خلاصه ای از بررسی کلی برروی مسایل داده‌های حجیم می‌پردازیم که شامل فرصت ها و چالش‌های داده‌های حجیم، تکنیک‌ها و فناوری‌های فعلی می باشند. برای بهینه سازی عملکرد وتحلیل داده‌های بزرگ الگوریتم‌های تکاملی و داده کاوی مطرح می‌شوند. الگوریتم‌های تکاملی (EA) معروف است که کشف کننده های بهتر در فضای جستجو در مقایسه با تکنیک‌های سنتی هستند؛ که روش های قطعی را در برمی‌گیرند و از مکانیزم‌ها و عملیات ابتدایی برای حل مسئله استفاده می‌کنند و طی یک سری از تکرارها به راه حل مناسب برای مسئله می‌رسند. ما در اینجا از ترکیب الگوریتم k- میانگین و الگوریتم ژنتیک برای رسیدن به نتیجه مطلوب استفاده کردیم که الگوریتم ژنتیک یکی از مشهورترین و متداول ترین الگوریتم-ها در بین الگوریتم‌های تکاملی است. این الگوریتم یک روش سراسری کمینه سازی است که برای حل مسائل بهینه سازی کاربردهای بسیار دارد.

واژه‌های کلیدی: داده های بزرگ، الگوریتم های تکاملی، الگوریتم ژنتیک، داده کاوی، k- میانگین.

۱. مقدمه

داده های بزرگ یا عظیم داده ترجمه اصطلاح Big Data می‌باشد که معمولا به مجموعه از داده‌ها اطلاق می‌شود که اندازه آنها فراتر از حدی است که با نرم افزارهای معمول بتوان آنها را در یک زمان معقول اخذ، دقیق‌سازی، مدیریت و پردازش کرد. مفهوم «اندازه» در داده های بزرگ بطور مستمر در حال تغییر است و به مرور بزرگتر می‌شود. داده‌های بزرگ (Big Data) مجموعه از تکنیک‌ها و تاکتیک‌هایی است که نیازمند شکل جدیدی از یکپارچگی هستند تا بتوانند ارزش‌های بزرگی را که در مجموعه های بزرگ، وسیع، پیچیده و متنوع داده پنهان شده اند، آشکار سازند. از این رو با رشد روز افزون داده‌ها و نیاز به بهره برداری و تحلیل از این داده ها، بکارگیری زیرساخت‌های Big Data از اهمیت ویژه ای برخوردار شده است.

عبارت Big Data مدت‌ها است که برای اشاره به حجم‌های عظیمی از داده‌ها که توسط سازمان‌های بزرگی مانند گوگل یا ناسا ذخیره و تحلیل می‌شوند مورد استفاده قرار می‌گیرد؛ اما به تازگی، این عبارت بیشتر برای اشاره به مجموعه‌های داده‌ای بزرگی استفاده می‌شود که به قدری بزرگ و حجیم هستند که با ابزارهای مدیریتی و پایگاه‌های داده سنتی و معمولی قابل مدیریت نیستند. مشکلات اصلی در کار با این نوع داده‌ها مربوط به برداشت و جمع‌آوری، ذخیره‌سازی، جست‌وجو، اشتراک‌گذاری، تحلیل و نمایش آن‌ها است. این مبحث، به این دلیل هر روز جذابیت و مقبولیت بیشتری پیدا می‌کند که با استفاده از تحلیل حجم‌های بیشتری از داده‌ها، می‌توان تحلیل‌های بهتر و پیشرفته‌تری را برای مقاصد مختلف، از جمله مقاصد تجاری، پزشکی و امنیتی، انجام داد و نتایج مناسب‌تری را دریافت‌کرد. بیشتر تحلیل‌های مورد نیاز در پردازش داده‌های عظیم، توسط دانشمندان در علومی مانند هواشناسی، ژنتیک، شبیه‌سازی‌های پیچیده فیزیک، تحقیقات زیست‌شناسی و محیطی، جست‌وجوی اینترنت، تحلیل‌های اقتصادی و مالی و تجاری مورد استفاده قرار می‌گیرد. حجم داده‌های ذخیره‌شده در مجموعه‌های داده‌ای Big Data، عموماً به‌خاطر تولید و جمع‌آوری داده‌ها از مجموعه بزرگی از تجهیزات و ابزارهای مختلف مانند گوشی‌های موبایل، حسگرهای محیطی، لاگ نرم‌افزارهای مختلف، دوربین‌ها، میکروفون‌ها، دستگاه‌های تشخیصRFID ، شبکه‌های حسگر بی‌سیم و غیره با سرعت خیره‌کننده‌ای در حال افزایش است. در اینجا ضمن بررسی مفاهیم پایه‌ای در داده های حجیم، به بررسی راه حل های موجود برای مدیریت و بهره برداری از این نوع داده ها خواهیم پرداخت.

برای ایجاد یک دید مناسب در خصوص داده های حجیم و اهمیت آن، جامعه ای را تصور کنید که در آن جمعیت بطور نمایی در حال افزایش است، اما خدمات و زیرساخت های عمومی آن نتواند پاسخگوی رشد جمعیت باشد و از عهده مدیریت آن برآید. چنین شرایطی در حوزه داده در حال وقوع است؛ بنابراین نیازمند توسعه زیرساخت های فنی برای مدیریت داده و رشد آن در بخش هایی نظیر جمع آوری، ذخیره سازی، جستجو، به اشتراک گذاری و تحلیل می باشیم. دستیابی به این توانمندی معادل است با شرایطی که مثلا بتوانیم "هنگامی که با اطلاعات بیشتری در حوزه سلامت مواجه باشیم، با بازدهی بیشتری سلامت را ارتقا دهیم"، "در شرایطی که خطرات امنیتی افزایش پیدا می‌کند، سطح امنیت بیشتری را فراهم کنیم"، "وقتی که با رویدادهای بیشتری از نظر آب و هوایی مواجه باشیم، توان پیش بینی دقیقتر و بهتری بدست آوریم"، "در دنیایی با خودروهای بیشتر، آمار تصادفات و حوادث را کاهش دهیم"، "تعداد تراکنش های بانکی، بیمه و مالی افزایش پیدا کند، ولی تقلب کمتری را شاهد باشیم"، "با منابع طبیعی کمتر، به انرژی بیشتر و ارزانتری دسترسی داشته باشیم" و بسیاری موارد دیگر از این قبیل که اهمیت پنهان داده های حجیم را نشان می دهد.

داده‌های حجیم اصطلاحی است که به مجموعه داده‌هایی اطلاق می‌شود که مدیریت، کنترل و پردازش آنها فراتر از توانایی ابزارهای نرم‌افزاری در یک زمان قابل تحمل و مورد انتظار است و چالش‌های بسیاری را منجر می شوند مانند دشواری‌ها و پیچیدگی‌هایی در گرفتن داده‌ها، ذخیره داده‌ها، تجزیه و تحلیل‌داده‌ها و غیره. نـمونه‌هایی از بزرگ‌داده گـزارش‌های وبی، سامانه‌های بازشناسی با امواج رادیویی، شبکه‌های حسگر، شبکه‌های اجتماعی، متون و اسناد اینترنتی، جستجوهای اینترنتی، نجوم، مدارک پزشکی، آرشیو عکس، آرشیو ویدیو، پژوهش‌های زمین‌شناسی و تجارت در مقیاس بزرگ هستند (چن و همکاران[1]، ۲۰۱۴).

[1] Chen et al.

داده‌های عظیم می توانند بر اساس مشخصات زیر تعریف شوند:

حجم : مقدار داده‌های تولید شده در این زمینه بسیار مهم است. اندازه داده‌ها ارزش و پتانسیل داده‌های مورد توجه به آن را تعیین می‌کند تا جایی که می‌توان تصمیم گرفت که داده عظیم محسوب می‌شود یا خیر.

تنوع : جنبه بعدی در داده‌های عظیم تنوع آن است. این بدان معنی است که دسته بندی داده‌های عظیم به ضرورت نیاز شناسایی شده توسط تحلیلگران داده دارد. این به افراد کمک می‌کند تا داده‌ها و ارتباطاتشان را دقیق‌تر تحلیل کنند تا از مزایا و رعایت اهمیت داده‌های عظیم به طور موثر استفاده کنند.

نرخ تولید : اصطلاح 'نرخ تولید' در این موضوع به سرعت تولید داده اشاره دارد و یا چگونگی سرعت تولید و پردازش داده‌ها برای پاسخگویی به خواسته و چالش‌های پیش رو در مسیر رشد و توسعه است.

ما به ارائه خلاصه ای از بررسی کلی برروی مسایل داده‌های حجیم می‌پردازیم که شامل فرصت ها و چالش‌های داده‌های حجیم، تکنیک‌ها و فناوری‌های فعلی می باشند. برای بهینه سازی عملکرد وتحلیل داده‌های بزرگ الگوریتم‌های تکاملی مطرح می‌شوند. الگوریتم‌های تکاملی (EA) معروف است که کشف کننده های بهتر در فضای جستجو در مقایسه با تکنیک‌های سنتی هستند. (باتاچاریا و همکاران[1]، ۲۰۱۶) که روش های قطعی را در برمی‌گیرند و از مکانیزم‌ها و عملیات ابتدایی برای حل مسئله استفاده می‌کنند و طی یک سری از تکرارها به راه حل مناسب برای مسئله می‌رسند. ما در اینجا الگوریتم ژنتیک را مطرح می‌کنیم که یکی از مشهورترین و متداول ترین الگوریتم‌ها در بین الگوریتم‌های تکاملی است. این الگوریتم یک روش سراسری کمینه سازی است که برای حل مسائل بهینه سازی کاربردهای بسیار دارد. (وایتلی[2]، ۱۹۹۴)

۲. اهداف پژوهش

در این پژوهش قصد داریم پارامترهای الگوریتم ژنتیک را برای استفاده درکاربردهای داده های بزرگ با اتومات های یادگیر تنظیم کنیم و همچنین تکنیکی برای حل مشکل در ذخیره‌سازی اطلاعات در داده‌های بزرگ ارائه نمائیم.

۳. فرضیه‌های تحقیق

۱. می‌توان بر اساس ماشین های یادگیر پارامترهای الگوریتم ژنتیک را تنظیم کرد.
۲. می‌توان با استفاده از تکنیک هایی مشکلات داده های بزرگ را حل نمود.
۳. استفاده از الگوریتم بهینه ژنتیک در قیاس با دیگر الگوریتم‌های بهینه سازی در این‌حوزه بهتر جواب می‌دهد.

۴. مبانی نظری پژوهش

۴.۱ داده های حجیم

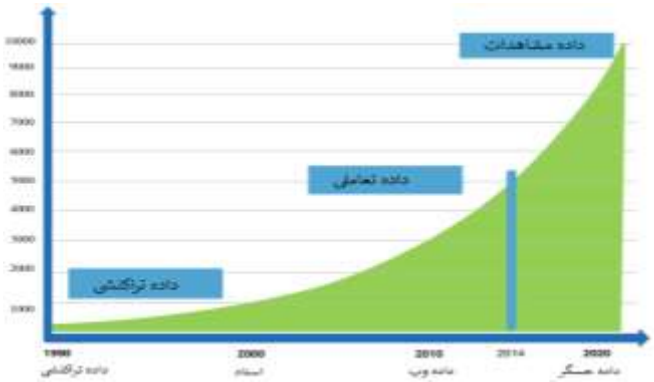
دانشمندان و محققان بر این باورند که داده‌های حجیم امروزه به اصطلاحی فراگیر و مبحثی مهم تبدیل شده در ادبیات فناوری اطلاعات است که مورد توجه دانشگاه ها، دولت ها و صنایع قرار گرفته شده است. همان طور که سازمان ها بزرگ می‌شوند، اطلاعات مرتبط با آنها نیز به صورت تصاعدی رشد می‌نماید و به تبع پیچیدگی‌های مرتبط با داده‌ها افزایش می یابد. سرعت داده‌های جدید در حال تولید سرسام آور است. بسیاری از سازمان های بزرگ در برنامه های کاربردی متفاوتشان داده‌های زیادی در فرمت های مختلفی را دارند. (کندی[3]، ۱۹۹۵) به همان میزان که داده‌ها گسترش می یابند، دسته بندی آنها با یک الگوریتم یا منطق مشخص بسیار دشوار می‌گردد. سازمان های بزرگ در واقع با این چالش مواجه هستند که تمامی اطلاعات را در یک پلتفرم نگهداری نمایند و یک دیدگاه ثابت به آن ها داشته باشند. این چالش منحصر به فرد برای درک تمامی داده‌هایی

[1] Bhattacharya et al.
[2] Whitley
[3] Kennedy

که از منابع متفاوت به دست می آیند و استخراج اطلاعات علمی مفید از آن ها، انقلاب داده‌های حجیم جهانی نامیده می شود. از سویی انجام کسب و کار برخلاف گذشته نیازمند تصمیم‌گیری هوشمند است. (آیبینو و همکاران[1]، ۲۰۱۶)

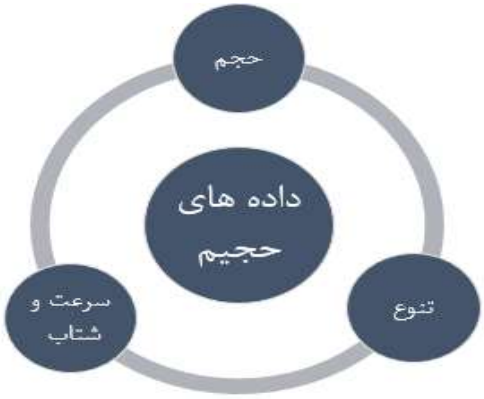
۴.۲ خصوصیات داده های حجیم

موسسه گارتنر (گروه متا) در سال ۲۰۰۱، سه بُعد از چالش ها و فرصت های پیش رو در حوزه رشد داده‌ها را مطرح کرد. این ویژگی ها به عنوان ویژگی های اصلی و معرف داده‌های بزرگ مطرح شدند. در واقع این سه واژه هستند که داده‌های حجیم را تعریف می کنند که در اصطلاح عامیانه به آن ها 3V گفته می شود:



شکل۱: حجم در مقابل تنوع

سرعت و شتاب (Velocity)/ افزایش سرعت تولید داده‌های ورودی و خروجی: اشاره به سرعت انتقال داده‌ها دارد. سرعت با افزایش تعداد منابع افزایش یافته است. داده‌ها از طریق برنامه های کاربردی و سنسورهای بسیار زیادی که در محیط وجود دارند با سرعت بسیار زیاد و به صورت بلادرنگ تولید می شوند که اغلب باید در لحظه پردازش و ذخیره شوند. رشد داده‌ها و انفجار رسانه‌های اجتماعی، نگاه ما را به داده‌ها تغییر داده است. امروزه مردم در شبکه های اجتماعی آخرین رخدادها را برای استفاده دیگران به روز رسانی می‌کنند. (چانگ و همکاران[2]، ۲۰۱۶)

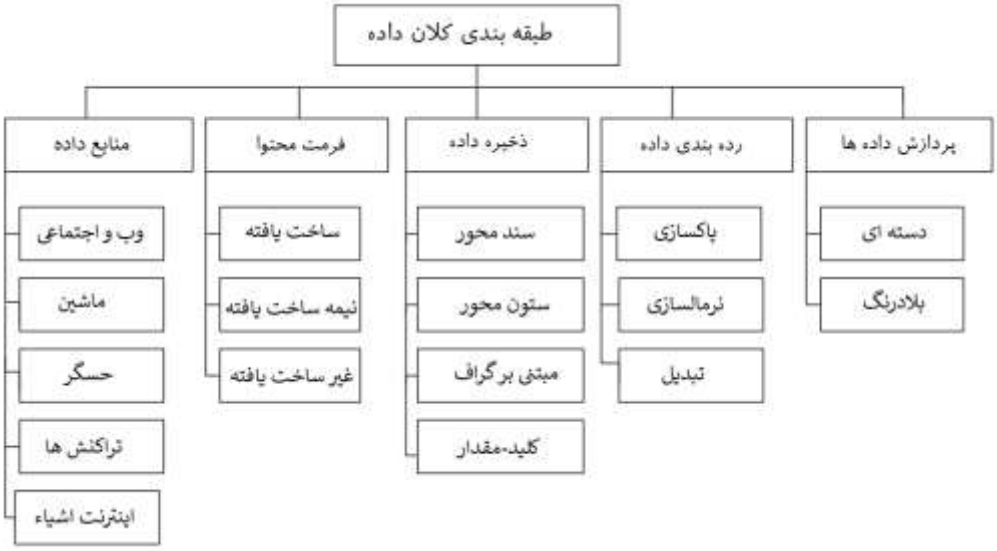


شکل ۲: خصوصیات داده های حجیم

۴.۳ طبقه بندی داده های حجیم

داده‌های بزرگ برای درک بهتر ویژگی هایشان به دسته‌های مختلف طبقه بندی می‌شوند. شکل ۳ دسته بندی های متعدد از داده‌های بزرگ را نشان می دهد.

[1] Aibinu et al.
[2] Chang et al.



شکل ۳: طبقه بندی داده های حجیم

۴.۴ چالش ها در داده های بزرگ

خدمات داده های حجیم، چه در بستر سخت افزاری و چه نرم افزاری، دارای محدودیت‌هایی است. ما در اینجا برخی از این محدودیت‌های مهم که به طور مستقیم احساس می‌شود را لیست می‌کنیم: دستگاه‌های ذخیره سازی به یک محدودیت بزرگ تبدیل شده اند. هارد دیسک با تکنولوژی دسترسی تصادفی برای ذخیره سازی داده‌ها دارای محدودیت به ویژه برای انتقال ورودی/خروجی سریع هستند که برای پردازش داده‌های بزرگ نیاز است. از سایر محدودیت‌های ذخیره سازی، می‌توان محدودیت‌های طراحی الگوریتم از نظر تعریف ساختمان داده مناسب که جوابگوی دسترسی سریع برای مدیریت داده‌هاست نام برد. نیاز به طراحی بهینه و پیاده سازی نمایه سازی برای دسترسی سریع به داده‌ها وجود دارد. ایده جدید دخیره سازی کلید-مقدار و آرایش سیستم فایل پایگاه داده چالش‌هایی برای مدیریت داده‌های بزرگ هستند. (دینگ[1]، ۲۰۱۶)

۴.۵ تجزیه تحلیل در داده های بزرگ

تجزیه و تحلیل در داده‌های بزرگ شامل شش حوزه می باشد که عبارتند از:

- ارتباط: حسگرها و شبکه ها
- ابر: محاسبات و داده‌ها برخط
- سایبر: مدل و حافظه
- محتوا و زمینه: معانی و ارتباطات
- جوامع: اشتراک و همکاری
- سفارشی سازی: شخصی سازی و ارزش تکنولوژی های داده‌های بزرگ

۴.۶ داده کاوی

امروزه دیگر فقدان اطلاعات یک مشکل نیست، بلکه عدم توانایی در استخراج اطلاعات مفید از داده‌ها یک مسئله مهم است. در دو دهه قبل توانایی‌های فنی بشر برای تولید و جمع آوری داده‌ها به سرعت افزایش یافته است. بطور کلی استفاده همگانی از وب و اینترنت به عنوان یک سیستم اطلاع‌رسانی جهانی ما را مواجه با حجم زیادی از داده و اطلاعات می‌کند. این رشد انفجاری در داده‌های ذخیره شده، نیاز مبرم وجود تکنولوژی‌های جدید و ابزارهای خودکاری را ایجاد کرده که به صورت

[1] Ding

هوشمند به انسان یاری رسانند تا این حجم زیاد داده را به اطلاعات و دانش تبدیل کند؛ داده کاوی به عنوان یک راه حل برای این مسائل مطرح می باشد. (لیل و همکاران[1]، ۲۰۱۷)

۴.۷ دسته بندی

دسته بندی از رایج‌ترین متدهای داده کاوی است که با عنوان تکنیک‌های یادگیری نظارتی[2] شناخته می‌شود. دسته بندی مقادیر بعضی متغیرها را محاسبه می‌کند و بر مبنای نتایج حاصله، طبقه بندی می‌کند. الگوریتم‌های مورد استفاده در کلاس بندی عبارتند از: درخت تصمیم، شبکه عصبی و... .

۴.۸ خوشه بندی

خوشه‌بندی از متدهای یادگیری بدون نظارت[3] است که مجموعه‌ای از الگوها را در گروه‌ها (یا خوشه‌ها) بخش بندی می‌کند. تحلیل‌های خوشه ای به گروه بندی مجموعه اشیاء داده ای به خوشه‌ها برمی‌گردد. به طور خاص، هیچ کلاس از پیش تعریف شده ای اختصاص داده نشده است. (جین و همکاران[4]، ۲۰۱۵)

۴.۹ کاربردهای خوشه‌بندی

از آنجا که خوشه‌بندی یک روش یادگیری بدون نظارت محسوب می‌گردد، در موارد بسیاری می‌تواند کاربرد داشته باشد.

در بازاریابی (Marketing): دسته‌بندی مشتری‌ها به دسته‌هایی بر حسب رفتارها و نیازهای آنها از طریق مجموعه زیادی از ویژگی‌ها و آخرین خریدهای آنها.

زیست‌شناسی (Biology): دسته‌بندی حیوانات و گیاهان از روی ویژگی‌های آنها

کتابداری: دسته‌بندی کتاب ها

نقشه‌برداری شهری (Planning-City): دسته‌بندی خانه‌ها بر اساس نوع و موقعیت جغرافیایی آنها.

۴.۱۰ خوشه بندی داده های حجیم

خوشه‌بندی به عنوان یکی از روش‌های داده کاوی می تواند به تشخیص الگوهای پنهان در داده‌ها و تحلیل حجم عظیم داده‌های شبکه بپردازد. کاربرد خوشه بندی داده‌ها تقریبا به وسعت همه حوزه‌های زندگی انسان است. خوشه‌بندی داده یک تکنیک شناخته شده در زمینه های مختلف از علوم کامپیوتر و حوزه مرتبط است. می‌توان گفت هرکجا داده ای ذخیره و استفاده می‌شود، پتانسیل فراوانی برای ورود داده کاوی و خوشه بندی داده در آن مشاهده می شود. (مائولیک[5]، ۲۰۰۰) بنابراین، روش های خوشه بندی داده با طیف وسیعی از انواع داده و دانش نهفته در داده‌ها روبرو است. داده‌های بیشتر به تحلیل‌های دقیق‌تر می‌انجامد؛ تحلیل‌های دقیق تر نیز منجر به تصمیم گیری‌های مطمئن‌تر شده و در پایان تصمیمات بهتر، می‌تواند به معنای کارایی بیشتر عملیات و کاهش هزینه ها و ریسک ها باشد. 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. (برتینو و همکاران، ۲۰۱۱)

[1] Leal et al
[2] Supervised
[3] Unsupervised
[4] Jin
[5] Maulik

۴.۱۱ روش های خوشه‌بندی داده های حجیم

به طور کلی، تکنیک های خوشه‌بندی داده‌های بزرگ را می توان به دو دسته عمده طبقه بندی کرد: تکنیک های خوشه‌بندی تک ماشین و تکنیک های خوشه‌بندی چند ماشینه. اخیرا تکنیک های خوشه‌بندی چند ماشینه به خاطر انعطاف پذیر بودن در مقیاس پذیری و ارائه زمان پاسخ سریع به کاربران، توجه بیشتری را بخود جذب کرده اند. (آنچالیا و همکاران[1]، ۲۰۱۳)

۴.۱۲ الگوریتم ژنتیک

در علوم کامپیوتر و ریاضیات الگوریتم جستجو الگوریتمی است که یک مساله را به عنوان ورودی می‌گیرد و پس از ارزیابی راه حل های ممکن یک راه حل برای آن مساله بر می‌گرداند. هنگامی که یک مساله را حل می‌کنیم بدنبال راه حل بهینه و یا به عبارتی بهترین پاسخ از بین پاسخ های ممکن هستیم. (آرورا[2]، ۲۰۱۴) در الگوریتم ژنتیک با الهام از طبیعت، هدف حل یک مساله جستجو و رسیدن به پاسخ بهینه است. الگوریتم ژنتیک به دلیل تقلید نمودن از طبیعت دارای چند اختلاف اساسی با روش‌های جستجوی مرسوم می‌باشد که در زیر به تعدادی از آنها اشاره می‌کنیم. (برمن[3]، ۲۰۱۳)

- الگوریتم ژنتیک با رشته‌های بیتی کار می‌کند که هر کدام از این رشته‌ها کلّ مجموعه متغیرها را نشان می‌دهد حال آنکه بیشتر روش‌ها به طور مستقل با متغیرهای ویژه برخورد می‌کنند.
- الگوریتم ژنتیک برای راهنمایی جهت جستجو، انتخاب تصادفی انجام می‌دهد که به این ترتیب به اطلاعات مشتق نیاز ندارد.

۴.۱۳ عملگرهای یک الگوریتم ژنتیک

به طور خلاصه الگوریتم ژنتیک از عملگرهای زیر تشکیل شده است:

- **کدگزاری**

این مرحله شاید مشکل‌ترین مرحله حل مسأله به روش الگوریتم باشد. الگوریتم ژنتیک به جای اینکه بر روی پارامترها یا متغیرهای مسأله کار کند، با شکل کد شدهٔ آنها سروکار دارد. یکی از روشه‌ای کد کردن، کد کردن دودویی می باشد که در آن هدف تبدیل جواب مسأله به رشته‌ای از اعداد باینری (در مبنای ۲) است. (بهبه و همکاران[4])

- **ارزیابی**

تابع برازندگی را از اِعمال تبدیل مناسب بر روی تابع هدف یعنی تابعی که قرار است بهینه شود به دست می‌آورند. این تابع هر رشته را با یک مقدار عددی ارزیابی می‌کند که کیفیت آن را مشخص می‌نماید. هر چه کیفیت رشته جواب بالاتر باشد مقدار برازندگی جواب بیشتر است و احتمال مشارکت برای تولید نسل بعدی نیز افزایش خواهد یافت. (بوزا و همکاران[5]، ۲۰۱۴)

- **ترکیب**

مهمترین عملگر در الگوریتم ژنتیک، عملگر ترکیب است. ترکیب فرآیندی است که در آن نسل قدیمی کروموزوم‌ها با یکدیگر مخلوط و ترکیب می‌شوند تا نسل تازه‌ای از کروموزوم‌ها بوجود بیاید. جفت‌هایی که در قسمت انتخاب به عنوان والد در نظر گرفته شدند در این قسمت ژن‌هایشان را با هم مبادله می‌کنند و اعضای جدید بوجود می‌آورند. ترکیب در الگوریتم ژنتیک باعث از بین رفتن پراکندگی یا تنوع ژنتیکی جمعیت می‌شود زیرا اجازه می‌دهد ژن‌های خوب یکدیگر را بیابند. (چانگ و همکاران، ۲۰۰۸)

[1] Anchalia et al.
[2] Arora
[3] Berman
[4] Bhih et al.
[5] Buza et al.

- **جهش**

جهش نیز عملگر دیگری هست که جواب‌های ممکن دیگری را متولد می‌کند. در الگوریتم ژنتیک بعد از اینکه یک عضو در جمعیت جدید بوجود آمد هر ژن آن با احتمال جهش، جهش می‌یابد. در جهش ممکن است ژنی از مجموعه ژن‌های جمعیت حذف شود یا ژنی که تا به حال در جمعیت وجود نداشته است به آن اضافه شود. جهش یک ژن به معنای تغییر آن ژن است و وابسته به نوع کدگذاری روش‌های متفاوت جهش استفاده می‌شود. (چن و همکاران[1]، ۲۰۱۲)

- **رمزگشایی**

رمزگشایی، عکسِ عمل رمزگذاری است. در این مرحله بعد از اینکه الگوریتم بهترین جواب را برای مسأله ارائه کرد لازم است عکس عمل رمزگذاری روی جواب‌ها یا همان عمل رمزگشایی اعمال شود تا بتوانیم نسخه واقعی جواب را به وضوح در دست داشته باشیم. (سوکس و همکاران[2]، ۱۹۹۷)

۵. روش تحقیق

در روش خوشه بندی k-میانگین، تعداد خوشه ها باید توسط کاربر مشخص شود. تشخیص صحیح تعداد خوشه های یک مجموعه داده اغلب برای کاربر کار دشواری است. به همین دلیل روش های خوشه بندی که به طور خودکار قادر به تعیین تعداد خوشه های مجموعه داده باشند، مورد توجه قرار می گیرند. به منظور غلبه بر این نقاط ضعف، ایده استفاده از الگوریتم های بهینه سازی برای تعیین مراکز اولیه ی خوشه ها در الگوریتم k-میانگین مطرح شد. یکی از الگوریتم های بهینه سازی کارا در این زمینه، الگوریتم ژنتیک است که به میزان قابل ملاحظه ای نتایج خوشه بندی را بهبود می بخشد اما به دلیل زمان مصرفی بالای الگوریتم ژنتیک، استفاده از آن برای یافتن مراکز اولیه خوشه ها در k-میانگین باعث بالا رفتن زمان این الگوریتم به ویژه برای مجموعه داده های بزرگ می شود. همچنین، از آنجا که اغلب مجموعه های داده ای که روش های خوشه بندی به منظور تحلیل آنها ایجاد شده‌اند، شامل داده های مختلط (داده های شامل ویژگی های اسمی و عددی) می باشند، روش خوشه‌بندی که قابلیت کار بر روی داده های مختلط را دارا باشد بسیار مورد توجه خواهد بود. در این بخش روشی برای خوشه بندی مجموعه داده های مختلط بزرگ با استفاده از k- میانگین و الگوریتم ژنتیک ارایه شده است. الگوریتم ژنتیک می تواند تعداد و مراکز اولیه خوشه ها را به طور خودکار پیدا کند. استفاده از الگوریتم ژنتیک برای یافتن مراکز خوشه های اولیه در k-میانگین تا حد زیادی باعث بهبود نتایج خوشه‌بندی می شود. اما با توجه به زمان مصرفی بالای الگوریتم ژنتیک، خوشه‌بندی مجموعه داده‌های بزرگ به این روش امکان پذیر نمی باشد. برای غلبه بر این مشکل و با توجه به ایده مطرح شده در اینجا ابتدا زیر مجموعه ای از مجموعه داده ی مورد نظر را انتخاب می‌کنیم. با اجرای الگوریتم ژنتیک روی این زیر مجموعه، بهترین مراکز اولیه ی خوشه ها را پیدا کرده و با ارسال آنها به k- میانگین و اجرای آن روی کل داده ها به خوشه‌بندی مجموعه ی داده مورد نظر می پردازیم. اجرای الگوریتم ژنتیک روی زیرمجموعه ای از داده ها، به جای کل آن ها، تا حد زیادی زمان مصرفی را کاهش می دهد. در ادامه به توضیح کامل مراحل کار می پردازیم.

۶. یافته های پژوهش

۶.۱ مجموعه داده Iris

این مجموعه در واقع مجموعه‌ای از داده‌ها می‌باشد که شامل سه نمونه گل زنبق است که توسط فیشر[3] در سال ۱۹۳۶ برای نشان دادن تکنیک‌های خطی تفکیک‌پذیر معرفی گردید. از این رو به نام مجموعه داده گل زنبق فیشر نیز خوانده می‌شود. از

[1] Chen et al.
[2] Cox et al.
[3]-Sir.Ronald Aylmer Fisher

طرف دیگر، به دلیل این‌که ادگار اندرسون[1] نیز این مجموعه را به‌دلیل کیفیت تنوع جغرافیایی در شبه جزیره گاسپه، گردآوری کرده است، به مجموعه داده زنبق اندرسون نیز مشهور می‌باشد.

جدول ۱: مراکز خوشه به دست آمده با اجرای الگوریتم Kmeans-GA روی مجموعه دادهIris

مراکز خوشه مجموعه دادهIris		
مرکز خوشه اول	مرکز خوشه دوم	مرکز خوشه سوم
۵.۸۸۹۰	۶.۷۷۴۹	۵.۰۰۳۶
۲.۷۶۱۲	۳.۰۵۲۴	۳.۴۰۳۰
۴.۳۶۴۰	۵.۶۴۶۶	۱.۴۸۵۰
۱.۳۹۷۳	۲.۰۵۳۵	۰.۲۵۱۵

بنا به بهترین مرکز خوشه به‌دست آمده روی مجموعه داده های Irish، مقدار حداقل تابع هزینه الگوریتم Kmeans-GA، ۶۰.۵۷۵۹۵۹ گردیده است (جدول ۲). ادعا بر برتری روش پیشنهادی با توجه به نتیجه حاصله امکان‌پذیر می‌باشد.

۶.۲ مجموعه داده Wine

این مجموعه که از آزمایشگاه MCI گرفته شده؛ از فعل و انفعالات شیمیایی شراب به‌دست آمده است. این مجموعه داده، نتیجه‌ی جمع‌آوری ۳ نمونه شراب در مکان‌های مختلف ایتالیا می‌باشد. در این مجموعه پارامترها (N=۱۷۸، d=۱۳، k=۳) می‌باشند. از این ۱۷۸ مورد، ۱۰۶ مورد مربوط به آموزش، ۳۶ مورد برای اعتبارسنجی و ۳۶ مورد باقی مانده جهت تست انتخاب گردیده‌اند. لازم به ذکر است که تمام ۱۳ ویژگی، پیوسته می‌باشند و هیچ‌کدام مقدار تهی ندارند. هم‌چنین، توزیع داده-ها در هر خوشه، ۵۹، ۷۱، ۴۸ می‌باشد. ۱۳ ویژگی این مجموعه داده بدین شرح می‌باشد: الکل، اسید مالیک، خاکستر، خاصیت قلیایی خاکستر، منیزیم، اسید فنیک، فلاونوئید، فنول‌غیرفلاونوئید، پرانسوسیانینز، شدت رنگ، شکل، OD280/OD315 از شراب رقیق و پرولین.

با توجه به جدول ۲ برای مجموعه داده Wine، جواب بهینه الگوریتم پیشنهادی ۱۷۹۶۰۸۲.۷۵۹۵ می‌شود.

جدول ۲: پاسخ الگوریتم های موجود بر روی مجموعه دادهWine

الگوریتم‌های تکاملی	تابع هزینه		
	حداقل هزینه	متوسط هزینه	حداکثر هزینه
Kmeans-GA	۱۷۹۶۰۸۲.۷۵۹۵	۱۷۹۶۰۸۲.۷۵	۱۷۹۶۰۸۲.۷۵۹۵
FCM	۱۷۹۶۲۰۶.۳۳۲۰	۱۷۹۶۳۱۴.۸۴	۱۷۹۶۴۲۳.۳۵۳۹
PSO	۱۷۹۷۰۰۱.۷۹۵۴	۱۸۰۷۱۵۳.۹۹	۱۸۱۷۳۰۶.۱۸۳۰
SA	۱۷۹۹۵۶۴.۷۲۳	۱۸۳۱۹۹۴.۳۳	۱۸۶۴۴۲۳.۹۴۵۳
TS	۱۸۱۷۳۰۶.۱۸۳۰	۱۸۲۴۰۶۲.۷	۱۸۳۰۸۱۹.۲۰۹۷
ACO	۱۸۰۹۹۵۳.۳۴۶۰	۱۸۱۰۶۳۸.۸۵	۱۸۱۱۳۲۴.۳۴۶۰
K-means	۱۸۰۵۲۱۴.۴۷۵۸	۱۸۱۱۹۸۰.۳۳	۱۸۱۸۷۴۶.۱۸۳۰

[1]-EdgarAnderson

۶.۳ مجموعه داده CMC

این مجموعه زیر مجموعه‌ای از بررسی‌های به انجام آمده جهت پیشگیری از شیوع بارداری خانم‌ها در اندونزی در سال ۱۹۸۷ می‌باشد. نمونه‌های مورد مطالعه خانم‌هایی بودند که در زمان مصاحبه یا باردار بودند و یا این‌که هنوز این موضوع برای آنها مشخص نشده بود. مشکلی که در این زمینه وجود داشت این بود که محققان نمی‌دانستند از کدام روش پیشگیری از بارداری خانم‌ها بر اساس خصوصیـــات دموگرافی و اجتماعی- اقتصادی او باید استفاده نمایند. با این فرض که از روش‌های کوتاه مدت و بلنـد مدت نیز نباید استفاده کنند. پارامتـــرهای این مجموعه به این صورت (N=۱۴۷۳، d=۱۰، k=۳) معرفی گردیدند که سن، تحصیلات، تحصیلات همسر، تعداد فرزندان به‌دنیا آورده، مذهب، شغل، شغل همسر، شاخص‌های زندگی استاندارد، دسترسی به رسانه‌های جمعی، استفاده از روش‌های جلوگیری از بارداری جزء شاخصه‌های این مجموعه می‌باشد.
در جدول ۳ بهـترین مرکز خوشـه به‌دست آمده از اجرای برنامه بر روی مجموعه داده CMC آورده شده است.

جدول ۳: مراکز خوشه به دست آمده با اجرای الگوریتم پیشنهادی روی مجموعه داده CMC

مراکز خوشه مجموعه داده CMC		
مرکز خوشه اول	مرکز خوشه دوم	مرکز خوشه سوم
۲۴.۰۳۳۴	۴۳.۹۹۱۵	۳۳.۵۳۸۷
۲.۹۸۳۱	۲.۸۵۰۷	۳.۰۸۹۸
۳.۴۶۲۷	۳.۳۵۲۱	۳.۵۱۱۶
۱.۷۶۴۲	۴.۸۲۳۴	۳.۶۳۰۲
۰.۹۲۸۲	۰.۸۱۴۴	۰.۷۹۰۱
۰.۷۹۴۷	۰.۷۶۹۴	۰.۶۹۷۱
۲.۳۱۸۴	۱.۸۸۳۲	۲.۰۹۹۱
۲.۹۱۲۷	۳.۳۴۱۷	۳.۲۶۸۱
۰.۰۴۵۴	۰.۱۱۲۳	۰.۰۶۷۳
۱.۹۸۴۹	۱.۶۲۴۰	۲.۰۷۳۰

۶.۴ مجموعه داده Vowel

این مجموعه داده‌ها شامل ۸۷۱ نمونه می‌باشند که برخلاف سه مجموعه‌ی قبل، مجموعه داده‌ها که صداهای آوادار تلوگو هندی می‌باشند، به ۶ خوشه تقسیم می‌گردند و ویژگی‌های مورد بررسی در این مجموعه ۳ شاخصه می‌باشد. به‌عبارت دیگر، این مجموعه داده نمونه‌هایی که دارای بعد کم، متوسط و زیاد می‌باشند؛ را پوشش می‌دهد. در واقع، این ۳ معیار شامل متکلـم، آوا و ورودی می‌شود. محدوده‌ی تغـیرات هر نطق مقادیر صحیح را به خود اختصاص می‌دهد که برای متکلم‌ها، بین ۰ تا ۸۹، برای آواها بین ۰ تا ۱۰ و برای مقادیر ورودی بین ۰ تا ۹ تخمین زده شده است؛ به عبارت دیگر، پارامترهای این مجموعه را می‌توان به صورت (N=۸۷۱، d=۳، k=۶) نشان داد.

جدول ۴: مراکز خوشه به دست آمده با اجرای الگوریتم پیشنهادی روی مجموعه داده Vowel

مراکز خوشه مجموعه داده Vowel					
مرکز خوشه اول	مرکز خوشه دوم	مرکز خوشه سوم	مرکز خوشه چهارم	مرکز خوشه پنجم	مرکز خوشه ششم
۵۱۰.۱	۴۱۵.۴	۴۰۹.۹	۳۶۱.۲	۴۴۲.۶	۶۴۴.۴
۱۷۷۳	۱۰۲۷.۵	۲۰۹۴	۲۲۹۳.۹	۹۹۷.۲	۱۲۹۰.۷
۲۵۲۰.۸	۲۳۴۵.۹	۲۶۵۳.۵	۲۹۷۰.۲	۲۶۶۵.۴	۲۲۹۸.۱

۷. نتیجه گیری

امروزه کاربرد داده‌کاوی دراکثر علوم به‌طورچشم‌گیر مشاهده می‌گردد. واضح است در صورتی‌که بستر مناسبی جهت استفاده از این علم مهیا نگردد، از تکنولوژی روز و بهره‌گیری از پیشرفت‌های به‌دست آمده دور خواهیم ماند. خوشه‌بندی یکی از ابزار داده‌کاوی محسوب می‌گردد. از این رو، سهم به‌سزایی از تحقیقات اخیر معطوف به این روش می‌باشد. الگوریتم Kmeans یک روش یادگیری بدون نظارت است که در آن تعداد خوشه‌ها از قبل تعیین نگردیده و علاوه بر آن یک داده می‌تواند به صورت همزمان به چندین خوشه تعلق داشته باشد. متأسفانه این الگوریتم علی‌رغم پیاده‌سازی آسان، با مشکلاتی چون وابستگی به شرایط اولیه، همگرایی زودرس و گیرکردن در بهینه محلی روبه‌رو است.

در صورتی‌که مقدار اولیه مناسب برای الگوریتم Kmeans انتخاب گردد، امکان همگرایی به نقاط بهینه وجود دارد. لذا، با بهره‌گیری از الگوریـتم ژنتیک به‌عنوان روش بهینه‌سازی نوین، این محدودیت مرتفع گردید. قابل توجه است که نسخه‌ی اصل الگوریتم ژنتیک به تنهایی در رفع این مشکل کارساز نبود. به همین دلیل، از روش تولید عدد متضاد جهت انتخاب جمعیت اولیه و روش استراتژی خود تطبیق، برای بهبود عملکرد این الگوریتم استفاده شد. این استراتژی از دو قانون جهش استفاده می‌کند. قانون اول از همگرایی زودرس و قانون دوم از سکون و گیر کردن در بهینه‌ی محلی جلوگیری می‌کند.

به این ترتیب، نتایج به‌دست آمده از ترکیب این دو الگوریتم، علاوه بر نوآوری، محدودیت‌ها را نیز پوشش داد و با توجه به مقایسه‌ی نتایج، مقاوم بودن و کارایی این روش تضمین گردید.

با توجه به شبیه‌سازی‌های انجام گرفته در اغلب موارد، روش ترکیبی جدید بهترین جواب را با حداکثر دقت نتیجه می‌دهد. در این روش ترکیبی، علاوه بر این‌که محدودیت وابستگی به شرایط اولیه تا حدود بسیار زیادی خنثی گردیده، مشکل همگرایی دائمی به پاسخ بهینه‌ی محلی نیز تا حد قابل قبولی از بین رفته است.

شبیه‌سازی بر روی مجموعه داده‌های متنوع اجرا گردیده است. پاسخ‌های نهایی به‌دست آمده گویای این مطلب می‌باشدکه وابستگی به تعداد نمونه‌ها، تعداد خوشه‌ها و تعداد پارامترها وجود ندارد و الگوریتم ترکیبی عملکرد مناسبی دارد؛ بنابراین، این الگوریتم را می‌توان بر مجموعه داده‌ها مشابه اعمال نمود و نتیجه مناسب را به‌دست آورد.

عدم وابستگی الگوریتـم به پارامترهای موجود، α، γ،N،f_{max} ، یکـــی دیگر از برتری‌های روش ارائه شده می‌باشد.

۸. پیشنهادها

با توجه به تلاش‌های انجام شده می‌توان، ترکیب دو الگوریتم فازی C-means و الگوریتم قورباغه را به‌عنوان کار آینده پیشنهاد نمود.

الگوریتم خوشه بندی C میانگین (Fuzzy C-Means): یکی از مهمترین و پرکاربردترین الگوریتم های خوشه بندی، الگوریتم C میانگین می باشد. در این الگوریتم نمونه ها به C خوشه تقسیم می شوند و تعداد C از قبل مشخص شده است. در نسخه فازی این الگوریتم نیز تعداد خوشه ها (c) از قبل مشخص شده است. الگوریتم (FCM) یکی از معروف ترین تکنیک های استفاده شده برای خوشه بندی است.

الگوریتم جهش ترکیبی قورباغه (SFLA): یک الگوریتم تکاملی و مبتنی بر جمعیت متاهیورستیک جدید است. این الگوریتم سریع است و قابلیت جستجوی سراسری بسیار خوبی دارد.

منابع

1. Aibinu, Abiodun Musa, et al. "A novel Clustering based Genetic Algorithm for route optimization." Engineering Science and Technology, an International Journal 19.4 (2016): 2022-2034.
2. Anchalia, Prajesh P., Anjan K. Koundinya, and N. K. Srinath. "MapReduce design of K-means clustering algorithm." Information Science and Applications (ICISA), 2013 International Conference on. IEEE, 2013.
3. Arora, Saurabh, and Inderveer Chana. "A survey of clustering techniques for big data analysis." Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-. IEEE, 2014.
4. Berman, Jules J. Principles of big data: preparing, sharing, and analyzing complex information. Newnes, 2013.
5. Bertino, Elisa, et al. "Challenges and Opportunities with Big Data." (2011).
6. Bhattacharya, Maumita, Rafiqul Islam, and Jemal Abawajy. "Evolutionary optimization: a big data perspective." Journal of network and computer applications 59 (2016): 416-426.
7. Bhih, Amhmed A., Princy Johnson, and Martin Randles. "EM Clustering Approach for Multi-Dimensional Analysis of Big Data Set."
8. Buza, Krisztian, Gábor I. Nagy, and Alexandros Nanopoulos. "Storage-optimizing clustering algorithms for high-dimensional tick data." Expert Systems with Applications 41.9 (2014): 4148-4157.
9. Chang, Dongxia et al. "A dynamic niching clustering algorithm based on individual-connectedness and its application to color image segmentation." Pattern Recognition 60 (2016): 334-347.
10. Chang, Fay, et al. "Bigtable: A distributed storage system for structured data." ACM Transactions on Computer Systems (TOCS) 26.2 (2008): 4.
11. Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information Sciences 275 (2014): 314-347.
12. Chen, Yanpei, Sara Alspaugh, and Randy Katz. "Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads." Proceedings of the VLDB Endowment 5.12 (2012): 1802-1813.
13. Cox, Michael, and David Ellsworth. "Managing big data for scientific visualization." ACM Siggraph. Vol. 97. 1997.
14. D. Whitley, A genetic algorithm tutorial, Statistics and computing, 4 (1994) 65-85.
15. J. Kennedy, R, Eberhart, Particle swarm optimization, 1 (1995)
16. Jin, Xiaolong, et al. "Significance and challenges of big data research." Big Data Research 2.2 (2015): 59-64.
17. Leal, Emilcy J. Hernández, Néstor D. Duque Méndez, and Julián Moreno Cadavid. "Big Data: an exploration of research, technologies and application cases." TecnoLógicas 20.39 (2017).

18. R.J. Kuo, C.H. Mei, F.E. Zulvia, C.Y. Tsai, An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation, Neurocomputing, 205 (2016) 116
19. U. Maulik, S. Bandyopadhyay, Genetic algorithm-based clustering technique**,** Pattern Recognition, 33 (2000) 1455-1465.
20. Y. Ding, X. Fu, Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm, Neurocomputing, 188 (2016) 233-238.

مطالعات مهندسی صنایع و مدیریت تولید

دوره ۳، شماره ۳، پاییز ۱۳۹۶، صفحات ۳۷-۲۴

Providing a Hybrid Approach of Genetic algorithms for Use in Data Clustering

Meysam Rahnamay Fallah[1], Marzieh Faridi Masouleh[2] *, Mohammad Reza Askaripour[3]

[1, 2, 3] *Islamic Azad University, Electronic Unit, Computer Department, Tehran, Iran*

** corresponding author*

Abstract

Big Data is usually referred to as a set of data that exceeds the extent which can be obtained, managed and processed using standard software within a reasonable time. The concept of size in big data is constantly changing and becoming gradually bigger. Hence, with the increasing growth of the data and the need to exploit and analyze these data, it is particularly important to apply the Big Data infrastructures. We provide a summary of a comprehensive review of big data issues, including the opportunities and challenges of big data and the current techniques and technologies. Evolutionary algorithms (Eas) and data mining are used for optimizing the performance and analyzing the big data. EAs are known as better discoverers in search space than the traditional techniques. They include definitive methods and use basic mechanisms and operations to solve the problem and, provide a suitable solution to the problem through a series of repetitions. We have used a combination of the k-mean algorithm and the genetic algorithm in order to arrive at a desirable result, and the genetic algorithm is one of the most popular and commonly used algorithms among the evolutionary algorithms. This algorithm is a global minimization method with many uses for solving optimization problems.

Keywords: big data, evolutionary algorithms, genetic algorithm, data mining, K-mean.