ترجمه مقاله:

چکیده:

با پیشرفت فناوری ، اندازه داده ها به سرعت در حال افزایش است. برای تصمیم گیری هوشمند بر اساس داده ها ، روش های تحلیلی کارآمد لازم است. خوشه بندی داده ها ، یک روش تحلیلی برجسته از داده کاوی است که به طور کارآمد در تجزیه و تحلیل داده ها استفاده می شود. برای تجزیه و تحلیل مجموعه داده های عظیم ، نیاز به پیشرفت در روش های پیشین حس میشود. در این مقاله ، یک روش خوشه بندی کارآمد ، بهینه ساز گرگ خاکستری مبتنی بر (MR-EGWO) MapReduce ، برای خوشه بندی مجموعه های داده در مقیاس بزرگ ارائه شده است. روش پیشنهادی یک نوع جدید از بهینه ساز گرگ خاکستری ، بهینه ساز گرگ خاکستری پیشرفته (EGWO) را معرفی کرد ، جایی که استراتژی شکار گرگ خاکستری با کراس اوور دو جمله ای ترکیبی شده و مراحل توزیع لوی برای تقویت توانایی جستجو برای جستجو طعمه استفاده میشود. بعلاوه ، از نوع پیشنهادی برای بهینه سازی فرآیند خوشه بندی استفاده می شود. کارایی خوشه بندی EGWO در هفت مجموعه داده معیار UCI آزمایش شده و با پنج تکنیک خوشه بندی موجود یعنی K-Means ، بهینه سازی ازدحام ذرات (PSO) ، الگوریتم جستجوی گرانشی (GSA) ، الگوریتم خفاش (BA) و بهینه ساز گرگ خاکستری (GWO) مقایسه می شود رفتار و سازگاری EGWO از طریق نمودار همگرایی و جعبه ها تأیید شده است. علاوه بر این ، EGWO پیشنهادی در مدل MapReduce در چارچوب Hadoop موازی شده و MR-EGWO نامگذاری شده است تا مجموعه داده های مقیاس بزرگ را اداره کند. علاوه بر این ، کیفیت خوشه بندی MR-EGWO نیز از نظر F-measure تأیید شده و در مقایسه با چهار روش پیشرفته مبتنی بر MapReduce یعنی:K-Means موازی، K-PSO موازی و بهینه سازی کلنی مصنوعی زنبور عسل مبتنی بر MapReduce( MR-ABC). نتایج تجربی تایید می کند که روش پیشنهادی جایگزین امیدوار کننده و قدرتمندی برای خوشه بندی داده های کارآمد و در مقیاس بزرگ است.

معرفی:

خوشه بندی روش برجسته یادگیری بدون نظارت است و به عنوان بخشی از برنامه های مهندسی داده مانند تقسیم تصویر ، داده کاوی ، سیستم بازیابی اطلاعات ، تشخیص ناهنجاری ، دارو ، بینایی رایانه و مدیریت ساخت در نظر گرفته می شود. طی سالهای گذشته ، الگوریتم های مختلفی برای خوشه بندی داده ها در این زمینه معرفی شده است تا تنوع در داده ها و مجموعه های مختلف برنامه های مورد نیاز را کنترل کند. K-means ، یکی از ساده ترین و محبوب ترین الگوریتم ها ، برای رفع مشکلات مختلف خوشه بندی استفاده شده است. با این حال ، نتایج الگوریتم K-means بسیار وابسته به مرکز خوشه های اولیه است و احتمال گرفتار شدن در بهینه های محلی زیاد است. برای بهبود این مشکلات ، روشهای مختلف خوشه بندی metaheuristic-based در این زمینه ارائه شده است. ماولیک و همکاران از توانایی الگوریتم ژنتیک برای یافتن بهترین مرکز خوشه در فضای ویژگی استفاده کرد تا فشردگی خوشه های حاصل بهینه شود. شارما و همکاران روشی مبتنی بر الگوریتم خفاش برای بهینه سازی فرآیند خوشه بندی پیشنهاد کردند که روش پیشنهادی همچنین با استفاده از MapReduce برای اداره مجموعه های بزرگ داده موازی سازی شد. حاتملو و همکاران روش خوشه بندی مبتنی بر الگوریتم جستجوی گرانشی را معرفی کرد که از الگوریتم K-means برای مقداردهی اولیه مراکز خوشه ها استفاده می کند. کارابوگا و همکاران یک روش مبتنی بر کلنی زنبور عسل مصنوعی جدید برای خوشه بندی داده های چند متغیره ارائه داده است. کومار و همکاران یک الگوریتم خوشه بندی معرفی کرد که از رفتار شکار گرگهای خاکستری تقلید می کند. کورا و همکاران روش خوشه بندی بهینه سازی ازدحام ذرات را توسعه داده و مشکلات برنامه وب را حل کرده است. در سال 2017 ، ابراهیمی و همکاران یک روش مبتنی بر جستجوی meta-heuristic تطبیقی ​​برای خوشه بندی بهینه حسگرها ، مستقر در محیط اینترنت اشیا معرفی کرد. الگوریتم های پیشنهادی در مقایسه با الگوریتم های خوشه بندی سنتی عملکرد بهتری داشته اند.

پندی و همکاران روش جستجوی ترکیبی cuckko برای خوشه بندی داده های توییتر برای تجزیه و تحلیل احساسات کاربران ارائه شده دادند. با این حال ، الگوریتم های خوشه بندی فوق به دلیل اجرای پی در پی آنها از نظر فضای حافظه و پیچیدگی های زمانی در مجموعه داده های بزرگ بدرستی عمل نمی کنند. برای کاهش هزینه محاسبات در مجموعه داده های بزرگ ، محاسبات موازی و توزیع شده راه حل های خوبی را به نمایش گذاشته اند. با پیشرفت در ابزارهای موازی سازی ، apache hadoop ابزاری است که به طور گسترده در این زمینه مورد استفاده قرار می گیرد. Hadoop یک پلتفرم منبع باز است که توسط Apache برای مدیریت مجموعه های داده بزرگ با استفاده از پردازش توزیع شده توسعه و مدیریت می شود. Hadoop با سیستم فایل خود کار می کند که به عنوان HDFS (سیستم فایل توزیع شده hadoop) نامیده می شود و قادر به پردازش بایت داده های zeta با واحد سخت افزاراست. MapReduce بستر محاسبات موازی را فراهم می کند و با موفقیت از نقاط قوت الگوریتم های meta-heuristic برای تجزیه و تحلیل مجموعه های داده در مقیاس بزرگ استفاده کرده است. گونگ و همکاران مدل های تکاملی توزیع شده مختلف را مطالعه کرده و از سادگی معماری MapReduce برای حل مشکلات مختلف محاسبات قدردانی کرده است. بنابراین ، محققان در پنج سال گذشته بر روی الگوریتم های meta-heuristic موازی مبتنی بر MapReduce کار کرده اند. روش ترکیبی K-PSO با معماری MapReduce توسط وانگ و همکاران برای خوشه بندی مجموعه های داده عظیم پیشنهاد شد. Banharnsakun الگوریتم مستعمره مصنوعی زنبورعسل مبتنی بر MapReduce (MR-ABC) را برای خوشه بندی مجموعه های داده بزرگ پیشنهاد کرد. ژائو و دیگران در استخراج دانش از داده های بزرگ از طریق نسخه موازی الگوریتم K-Means موفق بودند. خضر و همکاران مدلهای مختلف توزیع شده الگوریتمهای الهام گرفته از طبیعت را مطالعه کرد و نتیجه گرفت که مدل Hadoop MapReduce به دلیل سادگی و استحکام ، یکی از پلتفرمهای گسترده برای پردازش موازی مجموعه های داده بزرگ است. در این مقاله از مزایای GWO برای جمع آوری مجموعه داده های عظیم به طور موازی استفاده می شود.

بهینه ساز گرگ خاکستری (GWO) ، یک الگوریتم فرا ابتکاری است که از رفتار شکار گرگهای خاکستری الهام گرفته شده است. این الگوریتم از الگوریتم های meta-heuristic موجود مانند بهینه سازی ازدحام ذرات (PSO) ، استراتژی تکامل (ES) ، الگوریتم جستجوی گرانشی (GSA) ، تکامل افتراقی (DE) بر روی مشکلات استاندارد بهتر عمل کرده است. الگوریتم GWO در سه سال اخیر به طور گسترده ای در تعدادی از برنامه ها مورد استفاده قرار گرفته است. اماری و همکاران از نسخه باینری GWO برای انجام انتخاب بهینه ویژگی استفاده کرده است. کوماکی و کیوانفر با استفاده از GWO وظایف زمان بندی فروشگاه ها را بهینه سازی کردند. Fergany و Hasanien ، کارایی GWO را در مسئله جریان بهینه قدرت (OPF) نشان دادند. در همین زمینه ، GWO برای طراحی تثبیت کننده سیستم قدرت گسترده (WAPSS) استفاده شده است. گوها برای بهینه سازی کنترل فرکانس بار (LFC) از GWO در سیستم های قدرت استفاده کرد. سونگ و همکاران از نقاط قوت GWO برای یافتن پارامترهای بهینه امواج سطحی استفاده کرده است. بهینه ساز GWO همچنین به طور موثر برای آموزش گیرنده های چند لایه استفاده می شود. شیما امیرصدری و همکاران. نوع جدیدی از GWO را با استفاده از پروازهای levy در ترکیب با انتشار مجدد برای آموزش شبکه عصبی پیشنهاد کرد. با وجود کاربرد گسترده ، GWO محدودیت عدم تنوع جمعیتی را دارد. این نتیجه در سرعت همگرایی آهسته و خطر به دام افتادن در بهینه های محلی است. برای بهبود دقت جستجوی خود ، نسخه جدیدی از GWO ، بهینه ساز گرگ خاکستری پیشرفته (EGWO) ، با در نظر گرفتن قابلیت های زیر در این مقاله ارائه شده است.

• مراحل پرواز Lévy: برای بزرگنمایی جستجوی طعمه.

• کراس اوور دو جمله ای: برای تورم حمله به طعمه.

سهم کلی این مقاله به سه برابر تقسیم شده است. اول ، یک روش خوشه بندی جدید بر اساس نوع جدید GWO ارائه شده است. دوم ، کارایی نوع پیشنهادی بر روی مسئله خوشه بندی مطالعه شده است. سوم ، روش پیشنهادی با استفاده از معماری MapReduce موازی می شود و MR-EGWO برای خوشه بندی کارآمد مجموعه های داده بزرگ نامگذاری می شود. تجزیه و تحلیل تجربی EGWO در هفت مجموعه داده UCI انجام شده و در برابر پنج الگوریتم خوشه بندی یعنی K-Means ، PSO ، GSA ، BA و GWO از نظر میانگین و بهترین فاصله درون خوشه ای اعتبارسنجی شده است. رفتار همگرایی EGWO همراه با نمودارهای جعبه ای برای تجسم سازگاری آن مورد بحث قرار گرفته است. علاوه بر این ، عملکرد خوشه بندی MR-EGWO همچنین از نظر اندازه گیری F با مقایسه با چهار حالت موازی مبتنی بر MapReduce معتبر تأیید می شود: PK-Means ، K-PSO موازی ، MR-ABC ، ​​DFBPKBA. برای نشان دادن عملکرد محاسبه موازی ، روش پیشنهادی (MR-EGWO) در چهار مجموعه داده در مقیاس بزرگ مورد تجزیه و تحلیل قرار گرفته و از طریق نمودارهای سریع ادعا شده است.

ادامه این مقاله به شرح زیر است. بخش 2 خلاصه ای از اصول خوشه بندی و الگوریتم GWO است. در بخش 3 ، فرآیند خوشه بندی با استفاده از الگوریتم پیشنهادی همراه با موازی سازی آن با استفاده از MapReduce توضیح داده شده است. بخش 4 تنظیمات محیط آزمایشی و نتایج شبیه سازی را توضیح می دهد. سرانجام ، مقاله در بخش 5 به نتیجه رسید.

سابقه:

رویکرد خوشه بندی داده ها:

خوشه بندی یک مجموعه داده در فضای t-بعدی ، فرایند جمع آوری N داده در K گروه بر اساس شباهت است [26]. خوشه بندی، داده ها را به طور تکراری در گروه های K تایی (خوشه ها) به گونه ای تقسیم می کند که داده های در یک گروه حداکثر شباهت را داشته باشند. بعلاوه ، خوشه بندی داده ها نوعی رویکرد یادگیری بدون نظارت است ، به این معنی که داده ها بر اساس ساختارشان ، بدون هیچ گونه آموزشی گروه بندی می شوند. در حالی که ، در یادگیری نظارت شده مانند طبقه بندی ، داده ها بر اساس مجموعه آموزش با استفاده از داده های دارای برچسب طبقه بندی می شوند. EGWO خوشه بندی پیشنهادی برای تعداد مشخصی خوشه را انجام می دهد. مجموع فاصله درون خوشه ای هر خوشه به عنوان ملاک ارزیابی کیفیت آن خوشه انتخاب شده است. فرض کنید Z = (z1 , z2 , …,zn) مجموعه ی N تایی از داده ها است که در آن هر داده در فضای tبعدی نشان داده می شود. داده ها با ماتریسی از Z با n سطر و t ستون نشان داده می شوند که هر بردار سطری یک داده را توصیف می کند. فرآیند خوشه بندی،مجموعه ای از N داده را به K خوشه اختصاص داده و مجموعه ای از مرکز خوشه C = {C1 , C2 , … , Ck} را با هدف به حداقل رساندن مجموع فاصله اقلیدسی مربع شده بین هر یک از داده های Zi پیدا می کند و مراکز آنها را Ci می نامد. به طور کلی ، فرآیند خوشه بندی خصوصیات زیر را برآورده می کند:

• هر خوشه باید حداقل یک داده داشته باشد ،

به عنوان مثال{1 ,2 , 3 ,…,k} ∀i ∈، Ci = φ

• هر داده قطعاً بخشی از یک خوشه است.

• هیچ داده ای نمی تواند بخشی از بیش از یک خوشه باشد ، یعنی Cq

Cr = φ ∩Cq، ∀q = r و q , r ∈ {1 , 2 ,...,k}.

یک مجموعه داده بر اساس سه شرط فوق گروه بندی شده و کیفیت خوشه بندی از نظر تابع fitness ارزیابی می شود. مجموع مربع فاصله اقلیدسی یکی از توابع معروف مورد استفاده برای ارزیابی کیفیت خوشه بندی است که با استفاده از معادله زیر محاسبه می شود:

که در آن d (Zi,Cl) اندازه گیری تنوع بین داده Zi و مرکز خوشه Cl است. برای محاسبه عدم تشابه بین داده ها ، بسیاری از معیارهای فاصله پیشنهاد شده است. فاصله اقلیدسی یکی از معیارهای محبوب برای محاسبه عدم تشابه بین داده ها است. با توجه به دو داده Zi و Z j با ابعاد t ، فاصله اقلیدسی d (Zj,Zi) را با معادله زیر محاسبه می کنیم:

**2.2. بهینه ساز گرگ خاکستری:**

بهینه ساز گرگ خاکستری (GWO) یک الگوریتم فرا ابتکاری است که توسط میرجلیلی و همکاران پیشنهاد شده است که از مکانیسم شکار گرگهای خاکستری تقلید می کند. در GWO ، گرگهای خاکستری با توجه به سلسله مراتب اجتماعی خود در آلفا (α) ، بتا (β) ، دلتا (δ) و امگا (ω) دسته بندی می شوند. بهترین سه گرگ خاکستری به عنوان آلفا ، بتا ، دلتا تغییر نام می دهند و بقیه گرگهای خاکستری به عنوان امگا در نظر گرفته می شوند. گرگهای آلفا فرمان دهنده هستند و همه گرگهای خاکستری دیگر از دستورات آنها پیروی می کنند. دسته دوم گرگ های متعلق به گروه بتا مسئول کمک به آلفا در تصمیم گیری خود هستند. امگا گرگهای خاکستری کم رتبه هستند. در الگوریتم گرگ خاکستری ، شکار با آلفا ، بتا و دلتا همراه است در حالی که گرگهای امگا وظیفه محاصره طعمه را برای یافتن راه حل بهتر دارند. عملیات محاصره توسط گرگهای خاکستری توسط ریاضیات توسط دو معادله زیر تعریف شده است:

که در آن Xp محل طعمه است ، X (i) محل گرگ خاکستری در تکرار iام است. و بردارهای ضریب هستند و با استفاده از روابط زیر به ترتیب محاسبه می شوند:

که بردار ضریبی است که به طور خطی با افزایش تعداد تکرار از 0 تا 2 کاهش می یابد و r1 ، r2 اعداد تصادفی بین [0 ، 1] هستند.

بعلاوه ، معادلات بالا به ترتيب محدوده تخمين زده شده را به ترتيب براي وضعيت فعلي و آلفا ، بتا و دلتا تعريف مي كنند. پس از برآورد مسافت ، موقعیت نهایی گرگهای ω با اولین معادله زیر مشخص مiی شود. که ، ، بردارهای رندوم را نشان می دهد ، i عدد تکرار فعلی را نشان می دهد و بردارها ، ، ، به ترتیب توسط معادلات دوم تا چهارم زیرتعریف می شوند:

3. روش پیشنهادی:

برای مقابله با مشکلات خوشه بندی مجموعه داده های بزرگ ، یک روش پیشنهادی جدید ، بهینه ساز گرگ خاکستری تقویت شده مبتنی بر Map-reduce (MR-EGWO) است. MR-EGWO از نقاط قوت نوع جدیدی از بهینه ساز گرگ خاکستری ، بهینه ساز گرگ خاکستری تقویت شده (EGWO) برای خوشه بندی کارآمد داده ها بهره می برد. در این بخش ، ابتدا شرح مفصلی از بهینه ساز گرگ خاکستری تقویت شده (EGWO) و سپس نسخه موازی که بهینه ساز گرگ خاکستری بهبود یافته با Map-reduce (MR -EGWO) است ارائه شده است.

3.1 بهینه ساز گرگ خاکستری پیشرفته (EGWO):

کارایی یک الگوریتم فرا ابتکاری به تعادل بین exploration و explotation بستگی دارد. الگوریتم GWO به دلیل کمبود تنوع در گرگها برای موارد خاص محدودیت هایی مانند کندی در همگرایی و خطر به دام افتادن در بهینه های محلی را دارد. این محدودیت ها می توانند با افزایش تنوع و تشدید فضای جستجو برطرف شوند. بنابراین ، در این مقاله نسخه جدیدی از GWO به نام بهینه ساز گرگ خاکستری پیشرفته (EGWO) پیشنهاد شده است. روش پیشنهادی با استفاده از مزایای lévy flights و کراس اوور دوجمله ای برای بهبود قابلیت های exploration و explotation برخوردار است. EGWO دو مرحله جدید برای رفع مشکل ذکر شده در بالا معرفی می کند.

3.1.1 حمله به طعمه با استفاده از کراس اوور دو جمله ای:

از آنجا که تشدید جمعیت پیرامون بهترین راه حل فعلی ، تولید راه حل های بهینه را افزایش می دهد،با استفاده از یکی از اپراتورهای کراس اوور دو جمله ای محبوب و پرکاربرد در نسخه پیشنهادی قدرت explotation در مدل افزایش می یابد. همانطور که گرگ آلفا بهترین موقعیت فعلی را تعریف می کند ، می توان از موقعیت آن برای تعریف موقعیت بهتر گرگ های دیگر استفاده کرد. از این رو ، عملگر متقاطع دوجمله ای بین آلفا و X (i) انجام می شود تا تورم را به سمت طعمه بکشاند. موقعیت به روز شده (UP) گرگ های خاکستری در معادله تعریف شده است:

که موقعیت iامین گرگ خاکستری در jامین بعد است. K عدد تصادفی بین [0 ، 1] است و [0، 1] ∈ C ثابت کراس اوور است.

3.1.2. جستجو بزرگ شده ی طعمه با استفاده از lévy flight :

در GWO ، هنوز مشکل رکود در برخی موارد غالب است ، زیرا به روزرسانی موقعیت یک گرگ فقط توسط موقعیت گرگهای رهبر ، آلفا ، بتا و دلتا تعیین می شود. به همین ترتیب ، نتایج GWO منجر به همگرایی رشد نیافته می شود. برای ارتقا قابلیت exploration ، EGWO پیشنهادی از مفهوم توزیعflight lévy برای به روزرسانی موقعیت هر گرگ استفاده می کند. همانطور که flight lévy مراحل طول تصادفی ترسیم شده از توزیع levy را تعریف می کند ، احتمال کاوش در فضای جستجو افزایش می یابد. در این مقاله از الگوریتم Mantegna برای تولید مراحل با طول تصادفی استفاده می شود. معادله زیر فرمول بندی طول گام z را که توسط الگوریتم Mantega تعریف شده است ، به تصویر می کشد:

که در آن ، β ∈ (0,2] ، شاخص لوی است و r و s به ترتیب متغیرهای توزیع طبیعی وهستند. توسط معادله زیر محاسبه می شود. در حالی که همیشه 1 است:

که تابع گاما نامیده می شود و توسط معادله زیر تعریف می شود:

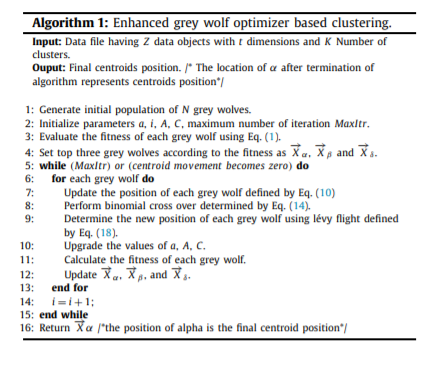
در مرحله پیشنهادی ، هر گرگ خاکستری برای جستجوی طعمه ، از levy flight پیروی می کند و موقعیت خود را با استفاده از رابطه زیر به روز می کند:

جایی که ، موقعیت گرگ خاکستری در تکرار tام است ، موقعیت گرگ آلفا و در همان تکرار tام اندازه گام lecy flight را تعریف می کند و توسط معادله زیر محاسبه می شود:

3.2 خوشه بندی مبتنی بر EGWO*:*

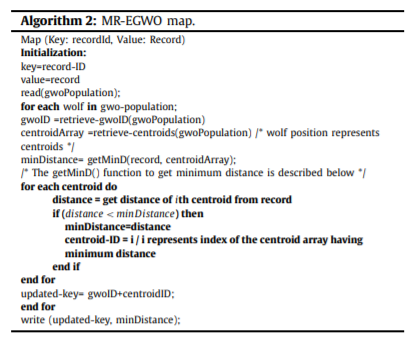
*در ادامه ، بهینه ساز پیشرفته ی گرگ خاکستری (EGWO) برای مسئله خوشه بندی توضیح داده شده است. در خوشه بندی مبتنی بر EGWO ، موقعیت X هر گرگ خاکستری نشان دهنده مجموعه ای از مراکز خوشه ای (C1 ، C2 ، C3 ، ··· ، C K) برای k خوشه است. به حداقل رساندن فاصله درون خوشه ای به عنوان تابع هزینه در نظر گرفته شده و در معادله زیر معادل فرموله شده است:*

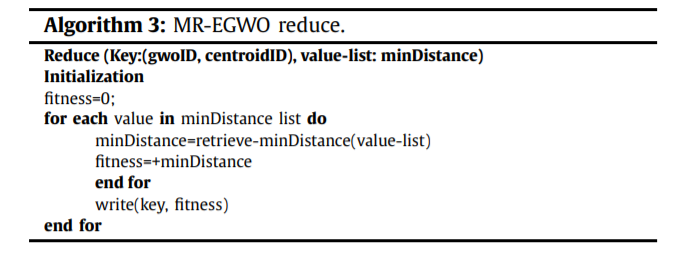
*خوشه های مطلوب با موقعیت گرگ آلفا مطابقت دارد. شبه کد روش خوشه بندی مبتنی بر EGWO در الگوریتم 1 شرح داده شده است. زمان محاسبه خوشه بندی مبتنی بر EGWO متناسب با اندازه و تعداد خوشه های مجموعه داده است. در این مقاله ، EGWO با حجم محاسباتی O (N × K × t) برای t تکرار، مراکز بهینه خوشه ها را تولید می کند ، که N تعداد داده ها و K مربوط به تعداد خوشه های مورد نیاز است. بنابراین ، برای اندازه جمعیت P ، پیچیدگی کل روش خوشه بندی پیشنهادی O (P × N × K × t) است.*

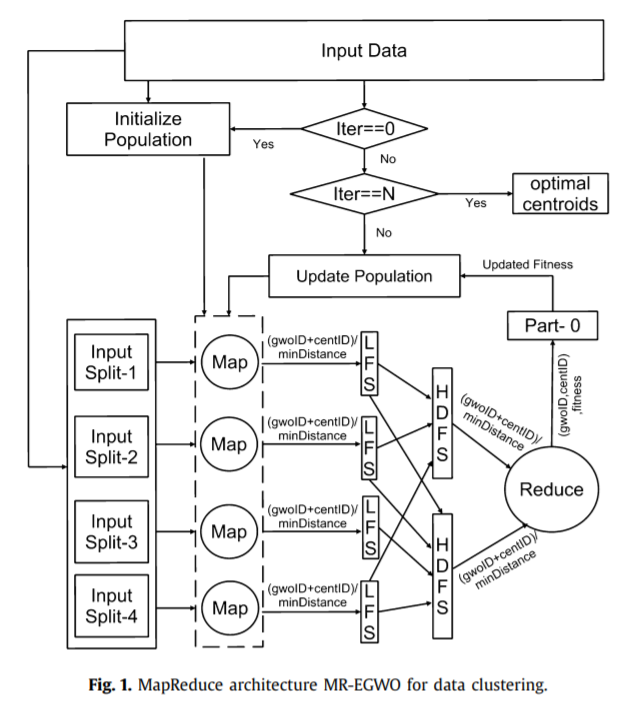


*3*.*3 موازی سازی EGWO با استفاده از معماری* MapReduce*:*

*برای* *نشان* *دادن* *کاربرد* EGWO *در* *مجموعه* *داده* *های* *بزرگ* ، *نسخه* *موازی* *الگوریتم* EGWO *با* *استفاده از* *چارچوب* Hadoop MapReduce ، EGWO *مبتنی* *بر* MapReduce (MR-EGWO) ، *ارائه* *شده* *است*. MR-EGWO *در دو مرحله کار می کند*: EGWO-Map *و* EGWO-Reduce*. در ابتدا ، چارچوب* MapReduce *مجموعه داده های بزرگ را به قطعات کوچکتر تقسیم کرده و به طور یکنواخت بین گره های* hadoop *توزیع می کند*. *بعلاوه ، هر نمونه داده توسط* record reader *به جفت* key/value *تبدیل* *می* شود. *بنابراین* ، *مرحله* map در MR-EGWO ، جفت key/value *ورودی را با مرکز خوشه به طور* *موازی پردازش می کند و شاخص مرکز هر داده را پیدا می کند. شبه کد مرحله* map در MR-EGWO *در الگوریتم 2 ارائه شده* *است. خروجی این فاز مجموعه دیگری از جفت* key/value *است* *که* key از *{*gwoID ، *شناسه* centroid*}* *تشکیل می شود ، در* *حالی که فاصله هر داده با* ID *مربوط به* centroid ، *مولفه* value *را تعریف می کند. علاوه بر این ، مرحله* reduceدر MR-EGWO *تمام مقادیر محاسبه شده را با کلیدهای یکسان ادغام می کند و مقدار* fitness function *مربوطه را برای هر گرگ* *خاکستری را* *محاسبه* می کند. *الگوریتم* *3* شبه *کد* *تابع* reduce در EGWO *را* *ارائه* *می* دهد*. گرگهای آلفا ، بتا و دلتا همراه با موقعیت هر گرگ خاکستری مطابق با* EGWO 1 *به روز می شوند. این یک تکرار* MR-EGWO *را نشان می دهد و این روند تا رسیدن به* *معیار توقف ادامه می یابد. معماری کامل* MR-EGWO *برای خوشه بندی داده ها در شکل 1 نشان داده شده است.*







4. نتایج تجربی:

کار پیشنهادی در دو مرحله ارزیابی می شود. ابتدا ، EGWO از نظر فاصله درون خوشه ای و رفتار همگرایی برای خوشه بندی ارزیابی می شود. مقایسه با k-means و چهار الگوریتم فرا ابتکاری برای خوشه بندی انجام می شود: GSA ، PSO ، BA و GWO. دوم ، اثربخشی MR-EGWO مبتنی بر MapReduce از نظر F-measure در برابر چهار روش خوشه بندی مبتنی بر MapReduce یعنی PKmeans، K-PSO ،KPSO موازی، بهینه سازی کلنی زنبور عسل مصنوعی مبتنی بر MapReduce برای خوشه بندی داده ها در مقیاس بزرگ (MR-ABC) و الگوریتم موازی K-Bat مبتنی بر فرکانس پویا (DFBPKBA) بررسی می شود. رفتار افزایش سرعت MR-EGWO نیز با افزایش تعداد گره ها در هر اجرا مورد مطالعه قرار می گیرد.

4.1 تجزیه و تحلیل عملکرد خوشه بندی مبتنی بر EGWO:

الگوریتم پیشنهادی EGWO در هفت مجموعه داده معتبر از مخزن UCI آزمایش شده و نتایج با K-means ، PSO ، GSA ، BA و GWO مقایسه می شود. Table1 خلاصه ای از هفت مجموعه داده معتبر در نظر گرفته شده است. این شبیه سازی برای 30 بار اجرا بر روی سیستم با Matlab 2015a ، Intel core i3 pro-cessor ، فرکانس 2.80GHz ، 4 گیگابایت رم و 500 گیگابایت هارد دیسک انجام شده است. Table2 جزئیات پارامتر آزمایش را تعیین می کند. جدول 3 بهترین و میانگین مقادیر fitness function بدست آمده توسط روشهای پیشنهادی و در نظر گرفته شده در طول 30 اجرا را تعریف می کند. می توان آن را از جدول 3 دریافت کرد ، زیرا EGWO از نظر بهترین مقدار fitness function ، از هر پنج روش در تمام مجموعه های داده بهتر عمل کرده است. از نظر ارزش متناسب بودن ، EGWO از نتایج دیتاست های wine، seeds ، glass و cancer پیشی گرفته است. با این حال ، GWO نتایج رقابتی در مجموعه داده های Iris و Balance دارد در حالی که PSO در مجموعه داده های Haberman عملکرد خوبی داشت. علاوه بر این ، برای تأیید تفاوت عملکرد در روشهای پیش بینی شده و آزمایش شده ، یک آزمون آماری غیر پارامتریک ، آزمون جمع آوری رتبه ویلکاکسون ، در سطح 5٪ از میزان سرطان انجام می شود. جدول 4 شامل مقدار p و SGFT (معنی دار بودن) هر روش است. اگر p-value <0.05 و همگام سازی شده توسط "+" یا "-" فرضیه صفر رد شود ، در غیر این صورت ، این نماد پذیرفته شده و با نماد "=" نشان داده می شود. "+" نشان می دهد که روش متفاوت و به طور قابل توجهی خوب است در حالی که "-" نشان می دهد متفاوت و به طور قابل توجهی ضعیف است. در جدول 4 مقدار p <0/5 در تمام مجموعه های داده مشاهده می شود. به همین ترتیب ، اطمینان حاصل می شود که EGWO به طور قابل توجهی با روشهای در نظر گرفته شده به جز GSA برای مجموعه داده های تعادل متفاوت است.

بهبود explotation و exploration ، رفتار همگرایی EGWO و روشهای در نظر گرفته شده در دو مجموعه داده ، یعنی wine و glassدر شکل 2 نشان داده شده است. محور افقی اعداد تکرار را نشان می دهد در حالی که مقادیر fitness function مربوطه در امتداد محور عمودی تراز می شوند. می توان از شکل 2 تجسم کرد که EGWO exploration را در مرحله اولیه تکرار ترجیح می دهد و سپس برای انجام explotation ، میزان exploration آن را کاهش می دهد. در مرحله بعد ، این کاهش فضای جستجو را به خوبی برای یافتن راه حل بهینه استفاده می کند. از این رو ، از نمودارهای همگرایی واضح است که EGWO توانایی های exploration و explotation را برخلاف GWO بهبود می بخشد. علاوه بر این ، نمودارهای جعبه ای در شکل 3 نشان دهنده سازگاری نتایج خوشه بندی گزارش شده توسط EGWO و سایر روش های در نظر گرفته شده است. خطوط عمودی جعبه ها نشان دهنده تنوع بهترین مقدار fitness finction تاکنون در بیش از 30 اجرا است. شکل 3 الف ، 3 به روشنی نشان می دهد که درجه پراکندگی در EGWO در مقایسه با PSO ، GSA ، BA و GWO حداقل است. بنابراین ، از نتایج تجربی می توان نتیجه گرفت که EGWO یک گزینه کارآمد برای انجام کارهای خوشه بندی است.

4.2 تجزیه و تحلیل عملکرد EGWO مبتنی بر MapReduce (MR-EGWO):

در بخش 4.1 ، EGWO نشان داده است که یک گزینه کارآمد برای کار خوشه بندی است. بنابراین ، عملکرد EGWO موازی (MR EGWO) ، مورد تجزیه و تحلیل قرار می گیرد. چهار مجموعه داده مصنوعی در مقیاس بزرگ با تکثیر107 بار از هر رکورد مجموعه داده اصلی استفاده می شود. Table5 این مجموعه داده ها را از نظر سه پارامتر ، یعنی تعداد خوشه های واقعی (#C) ، تعداد ابعاد (#D) و تعداد نقاط داده (#N) خلاصه می کند. تنظیم پارامتر مورد نیاز همه روشها همان است که در جدول 2 آورده شده است. برای شبیه سازی ، یک خوشه Hadoop از پنج گره طراحی شده است که هر گره از پردازنده Intel Corei3-4570 با 3.20 گیگاهرتز ، 4 گیگابایت حافظه و 500 گیگابایت دیسک سخت تشکیل شده است. Apache Hadoop نسخه 2.6.2 ، java نسخه 1.8.0 برای پیاده سازی همه روش ها استفاده می شود و سیستم عامل اوبونتو نسخه 14.04 است. جدول 6 میانگین F-measure و زمان محاسبه را برای 4 مجموعه داده مصنوعی در مقیاس بزرگ نشان می دهد که با اجرای هر روش روی خوشه ای از 5 ماشین بدست آمده است. مقایسه F-measureچهار روش مبتنی بر MapRedcue همانطور که در جدول 6 آورده شده تایید می کند که MR-EGWO پیشنهادی بهتر از همه روشهای مقایسه شده عمل می کند در حالی که K-means کمترین عملکرد را در بین تمام روشهای در نظر گرفته شده دارد. با این حال ، زمان محاسبه K-means در مقایسه با روشهای خوشه بندی مبتنی بر متا-ابتکار کمتر است. بنابراین ، می توان نتیجه گرفت که روش پیشنهادی می تواند برای خوشه بندی کارآمد مجموعه های داده بزرگ استفاده شود.

علاوه بر این ، عملکرد سرعت MR-EGWO در مجموعه داده های iris و CMC تحلیل می شود. اندازه گیری سرعت این روش توسط معادله زیر تعیین می شود:

*S p* = *Tbase*/*TN*

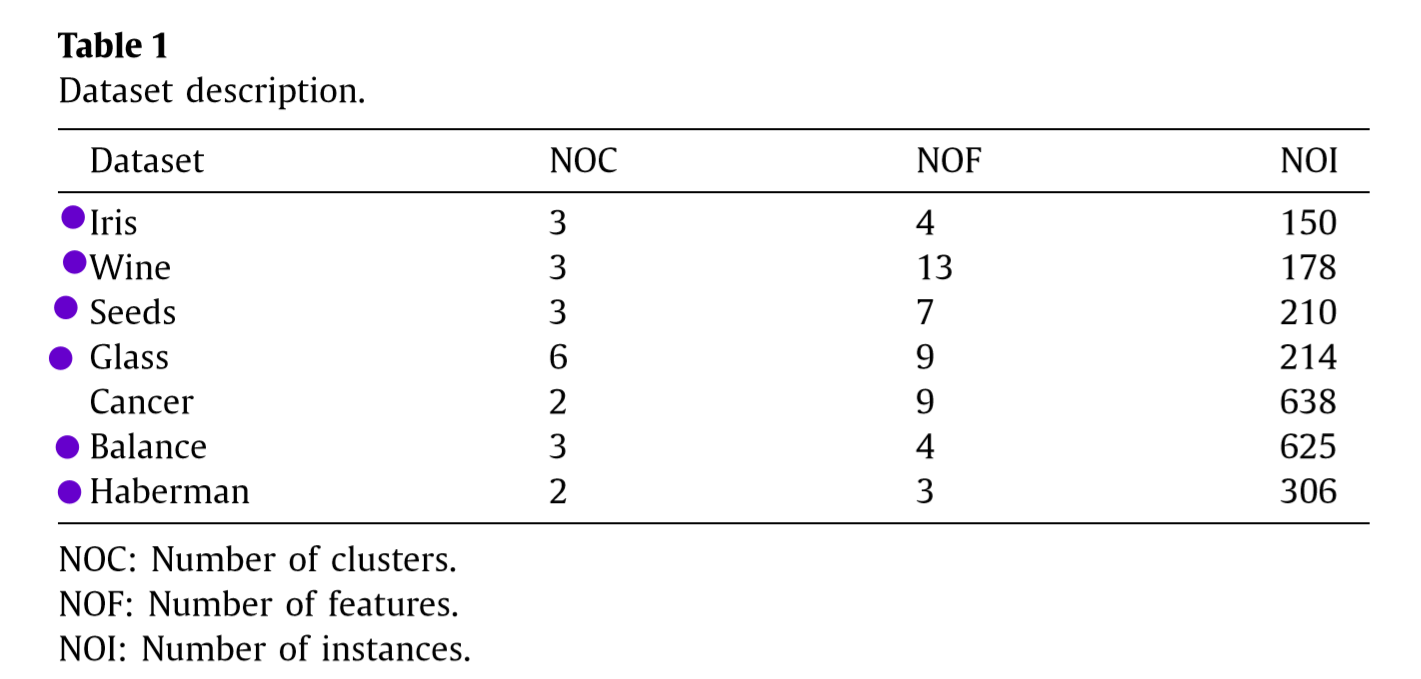
که در آن ، Tbase زمان اجرا است که روش p بر روی یک ماشین اجرا می شود و TN زمان اجرای همان روش است که در یک خوشه با N ماشین اجرا می شود. برای اندازه گیری عملکرد سرعت MR-EGWO ، در هر بار اجرا یک دستگاه در خوشه افزایش می یابد. عملکرد سرعت MR-EGWO در شکل 4 نشان داده شده است. از شکل 4 می توان نتیجه گرفت که زمان کار MR-EGWO با افزایش ماشین آلات در خوشه Hadoop به تدریج کاهش می یابد. روش پیشنهادی با مجموعه خوشه 5 دستگاه به ترتیب در مجموعه داده 1 و 2 به ترتیب 4.6754 ، 4.3457 را نشان داده است. بنابراین ، تأیید شده است که MR-EGWO پیشنهادی برای داده های مقیاس بزرگ سودمند است.

5. نتیجه گیری:

در این مقاله ، یک روش خوشه بندی مبتنی بر MapReduce جدید ارائه شده است. روش پیشنهادی دارای سه ویژگی است ، (i) یک نوع کارآمد بهینه ساز گرگ خاکستری به نام بهینه ساز خاکستری گرگ پیشرفته (EGWO) برای بهبود کیفیت خوشه بندی معرفی شده است (ii) عملکرد نسخه پیشنهادی (EGWO) در هفت مجموعه داده معیار برای مشکل خوشه بندی مقبول است. روش پیشنهادی از نظر خواص میانگین و بهترین fitness function ، از پنج روش خوشه بندی بهتر عمل کرده است: K-means ، PSO ، GSA ، BA و GWO. قابلیت های exploration و explotation از نوع پیشنهادی نیز با استفاده از نمودار همگرایی تحلیل می شود. نمودار های جعبه ای برای مطالعه سازگاری نتایج در طول 30 اجرا کشیده شده اند. (iii) ، یک روش جدید به نام MR-EGWO با موازی سازی EGWO با استفاده از MapReduce برای خوشه بندی مجموعه های داده در مقیاس بزرگ پیشنهاد شده است. روش پیشنهادی ، MR-EGWO ، از مزیت EGWO برای کاهش کیفیت خوشه بندی و معماری MapReduce برای کنار آمدن با مجموعه داده های مقیاس بزرگ استفاده می کند. علاوه بر این ، برای اطمینان از کارآیی MR-EGWO در محیط موازی ، روش پیشنهادی در خوشه Hadoop از پنج گره برای چهار مجموعه داده مصنوعی در مقیاس بزرگ یعنی iris ، CMC ، wine و vovel اجرا می شود. نتایج شبیه سازی از نظر F-measure از چهار روش پیشرفته خوشه بندی مبتنی بر MapReduce بهتر عمل کرد. علاوه بر این ، بازده سرعت MR-EGWO با تغییر تعداد گره های خوشه Hadoop در دو مجموعه داده مصنوعی (iris و CMC) مورد مطالعه قرار می گیرد. نتایج تسریع نشان می دهد که MR-EGWO برای تحلیل مجموعه داده های بزرگ با عملکرد سریع فوق العاده و کیفیت خوشه بندی بهتر مناسب است. بنابراین ، نتیجه گرفته می شود که MR-EGWO یک روش رقابتی برای مشکلات خوشه بندی در مقیاس بزرگ است. در آینده ، ممکن است ابزارهای موازی سازی اخیر مانند spark برای کاهش زمان محاسبه روش پیشنهادی مورد آزمایش قرار گیرند. علاوه بر این ، روش پیشنهادی می تواند در برخی از برنامه های خوشه بندی در دنیای واقعی با مجموعه داده های بزرگ مانند تجزیه و تحلیل توییتر ، تجزیه و تحلیل ویدئو و تجزیه و تحلیل تصویر ماهواره ای گسترش یابد.

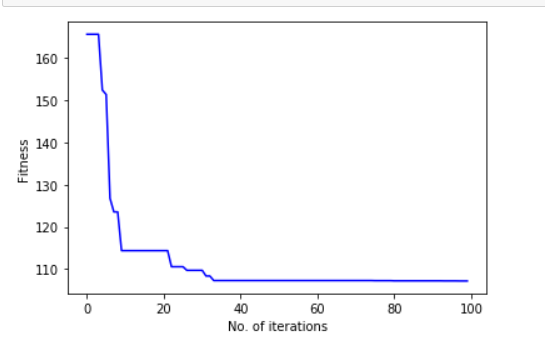
نتایج بدست آمده:

راهکار پیشنهادی در مقاله برای کلاسبندی توسط کد پایتون پیاده سازی شد ابتدا این راهکار بر روی دیتاست iris پیاده سازی شد و سپس بر روی باقی دیتاست های معرفی شده در مقاله مطابق جدول زیر تست شد:

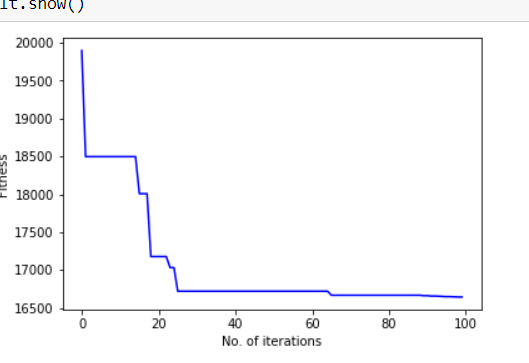


نمودار cost function هر دیتاست به صورت زیر بدست آمده است:

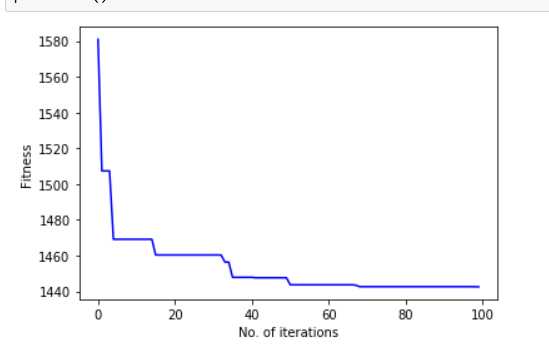
Iris:



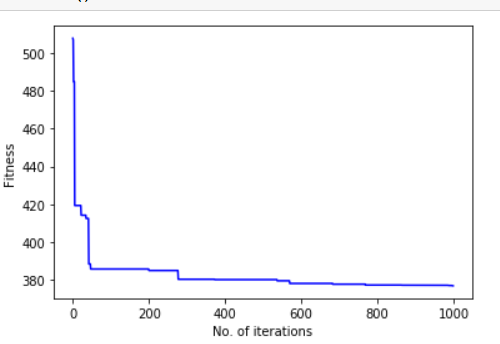
Wine:



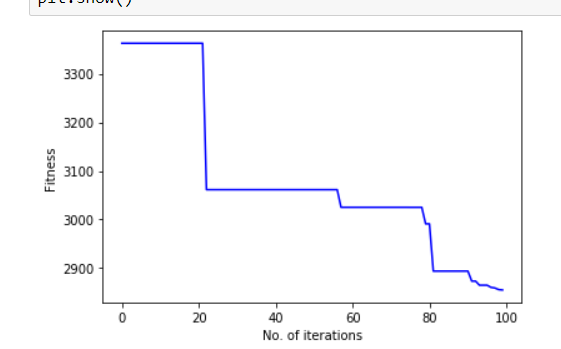
Balance:



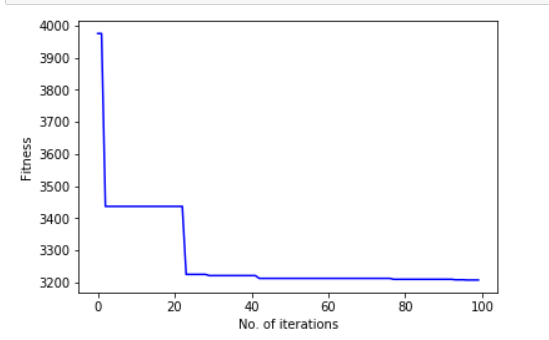
Seed:



Glass:



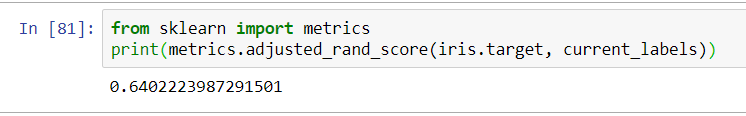
Haberma:



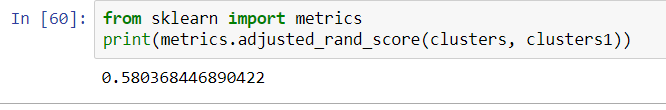
برای بررسی کارایی روش پیشنهادی آن را با 2 روش خوشه بندی مبتنی بر هوش جمعی اولی الگوریتم pso و دیگری graywolf اصلی مقایسه کردم.

در این مقایسه برای هر 3 روش از یک دیتاست مشابه (iris) استفاده شد و برای بررسی میزان کارآمدی روش ها از معیار metrics.adjusted\_rand\_score استفاده کردم که نتایج زیر حاصل شدند:

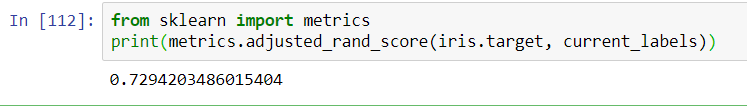
Simple graywolf:



Pso:



Enhanced graywolf:



همانطور که مشاهده میشود در روش ارائه شده در این مقاله 73 درصد از نمونه ها بدرستی کلاسبندی شده اند که در مقایسه با دو روش دیگر عملکرد بهتری از خود نشان داده است.