The 2nd National Conference on
Geospatial Information Technology (NCGIT)

K.N.Toosi University of Technology
Faculty of Geomatics Engineering

18 January 2017

دومین کنفرانس ملی مهندسی فناوری اطلاعات مکانی

دانشگاه صنعتی خواجه نصیرالدین طوسی
دانشکده مهندسی نقشه‌برداری
۲۹ دی ماه ۱۳۹۵

ارائه‌ی یک روش خوشه‌بندی فازی سری‌های زمانی بر مبنای الگوریتم تکاملی دیفرانسیلی

زاهده ایزکیان[1*]، محمدسعدی مسگری[2]

۱- دانشجوی دکتری سیستم اطلاعات مکانی دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه‌نصیرالدین‌طوسی
۲- استادیار دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه‌نصیرالدین‌طوسی

چکیده :

امروزه حجم داده‌های جمع‌آوری شده به مراتب بیشتر از گذشته می‌باشد. چگونگی استخراج اطلاعات مفید از میان این حجم عظیم داده همواره یکی از سوالات اساسی در علوم مختلف بوده و داده‌کاوی یکی از مهمترین روش‌های حل این مسئله است. در این میان تکنیک خوشه‌بندی داده‌ها به عنوان یکی از پرکاربردترین زمینه‌های داده‌کاوی همواره مورد توجه بسیاری از محققین بوده است. در این تحقیق از یک روش خوشه‌بندی مبتنی بر الگوریتم تکاملی دیفرانسیلی به عنوان یکی از الگوریتم‌های بهینه‌سازی کارا و قدرتمند، برای خوشه‌بندی سری‌های زمانی استفاده گردید. روش پیشنهادی به گونه‌ای عمل می‌کند که مراکز خوشه‌ها الزاما از میان داده‌های موجود انتخاب می‌گردند. با در نظر گرفتن تابع فاصله‌ی پیچش زمانی پویا (DTW) و انتخاب تابع بهینه‌سازی مناسب، روش پیشنهادی بر روی چندین مجموعه داده پیاده‌سازی شد و با استفاده از معیارهای ارزیابی مانند دقت (Precision) و معیار F (F-Measure) با روش خوشه‌بندی FKM مقایسه گردید و نتایج حاصل مورد بحث و بررسی قرار گرفت.

واژه‌های کلیدی : سری زمانی، خوشه‌بندی، الگوریتم تکاملی دیفرانسیلی

نویسنده مکاتبه کننده : زاهده ایزکیان
آدرس پستی : تهران - خیابان ولیعصر - تقاطع میرداماد - روبروی ساختمان اسکان-دانشکده نقشه برداری
تلفن : ۰۹۳۳۹۴۹۱۳۷۷
آدرس پست الکترونیک : zahedeh_izakian@yahoo.com

۱- مقدمه

سری‌های زمانی ترتیبی از اعداد حقیقی هستند که نمایانگر مقدار مشاهده شده از یک رویداد در فواصل زمانی برابر می‌باشند. تغییرات قیمت سهام، میزان مصرف انرژی، میزان بارندگی و... از جمله مواردی هستند که با استفاده از سری‌های زمانی نمایش داده می‌شوند. در اقتصاد از سری‌های زمانی برای بررسی و پیش‌بینی قیمت سهام، میزان صادرات، متوسط درآمد و ... استفاده می‌شود. در بازاریابی، تجزیه و تحلیل ارقام فروش در هفته یا ماه‌های متوالی که مسئله‌ای مهم در تجارت است بوسیله‌ی سری‌های زمانی انجام می‌شود. در هواشناسی از سری‌های زمانی برای نمایش تغییرات آب و هوایی و پیش‌بینی وضع هوا استفاده می‌شود. در جمعیت‌شناسی از سری‌های زمانی برای پیش‌بینی تغییرات جمعیت در مدت زمان مشخص استفاده می‌گردد. علاوه بر این تجزیه و تحلیل این سری‌ها در علومی مانند بیوانفورماتیک، زمین‌شناسی، علوم دریایی، پزشکی و مهندسی کاربرد فراوان دارد. به‌طور کلی محققان برای برآورد یک مدل مناسب و بررسی چگونگی تاثیر مجموعه‌ای از ورودی‌های مشخص بر روی خروجی‌های مورد نظر مسئله به تجزیه و تحلیل سری‌های زمانی می‌پردازند که این امر اطلاعات مفیدی را برای نظارت و پیش‌بینی‌های دقیق فراهم می‌آورد. یکی از مهم‌ترین تکنیک‌های تجزیه و تحلیل سری‌های زمانی که طی سال‌های اخیر مورد توجه دانشمندان قرار گرفته است، تکنیک خوشه‌بندی می‌باشد. عمل خوشه‌بندی داده‌های موجود در یک مجموعه داده را با توجه به معیارهای شباهت (توابع فاصله) مورد نظر به گروه‌های متشابه تقسیم می‌کند. گروه‌بندی باید به صورتی انجام شود که اشیاء موجود در یک گروه بیشترین شباهت را نسبت به هم و بیشترین اختلاف را با اشیاء موجود در گروه‌های دیگر داشته باشند. از دیدگاه کاربردی، مهم‌ترین هدف خوشه‌بندی کشف داده‌هایی است که ساختار مشابه دارند که این امر باعث ساده‌سازی و خلاصه کردن داده شده و نقش مهمی را در داده‌کاوی و کشف اطلاعات در مجموعه داده ایفا می‌کند [۱، ۲ و ۳]. بدلیل گستردگی فضای جستجو، ماهیت بهینه‌سازی و عدم وجود اطلاعات قبلی راجع به ساختار داخلی پایگاه داده در مسئله‌ی خوشه‌بندی، و از طرفی توانایی الگوریتم‌های تکاملی در غلبه بر مسائل پیچیده و چندهدفه، عدم نیاز به درک و مدلسازی دقیق پدیده، شکل تصادفی جستجو در فضا، تکنیک‌های خوشه‌بندی به کمک این روش‌ها عملکرد بهتری را نسبت به روش‌های کلاسیک خوشه‌بندی مجموعه داده‌های پیچیده در دنیای واقعی نشان می‌دهند. در این میان کارایی و عملکرد خوب الگوریتم بهینه‌سازی تکاملی دیفرانسیلی[1] (DE) به‌عنوان یکی از روش‌های جستجوی کلی سریع، قوی و کارآمد در خوشه‌بندی داده‌ها به اثبات رسیده است.

Das و همکاران [۴] از الگوریتم DE بهبود یافته برای خوشه‌بندی داده‌هایی با طول نه چندان بلند استفاده کردند. در روش پیشنهادی آنها هر کروموزوم از دو بخش مراکز خوشه‌ها و کدهای فعالسازی تشکیل می‌شود که کدهای فعالسازی وظیفه‌ی تعیین خودکار تعداد خوشه‌ها را بر عهده دارند. آنها در روش خود، به جای در نظر گرفتن مقداری ثابت به‌عنوان فاکتور مقیاس، از عددی متغیر در بازه‌ی (۱و۰.۵) استفاده کردند. نتایج به‌دست آمده از پیاده‌سازی روش DE بهبود یافته بر روی تعدادی مجموعه داده به خوبی کارایی این روش را نشان داد. Das و همکاران [۵] از الگوریتم DE برای خوشه‌بندی خودکار پیکسل‌های موجود در تصاویر استفاده کردند. الگوریتم پیشنهادی آنان به اطلاعات اولیه راجع به تعداد خوشه‌های موجود در مجموعه داده نیازی ندارد. آنها در روش خود پیشنهاد دادند که مقدار نرخ ترکیب با گذشت زمان از اجرای الگوریتم به صورت خطی تغییر کند. بدین صورت که در ابتدای اجرای الگوریتم ماکزیمم مقدار را برای نرخ ترکیب در نظر گرفته و در طول اجرا از مقدار آن کاسته شده تا به مینیمم مقدار برسد. این روش کمک می‌کند که الگوریتم در ابتدا به صورت گسترده ای فضا را جستجو کند و در مراحل بعد با کاهش فضای جستجو دامنه‌ی حرکات نیز کاهش می‌یابد. نتایج حاصل از پیاده سازی این روش قدرت بالای آن را در مقایسه با روش خوشه‌بندی FCM و الگوریتم ژنتیک به خوبی نشان داد. Fuad [۶] در مقاله‌ی خود یک روش جدید

[1] Differential evolution

خوشه‌بندی سری‌های زمانی بر مبنای روش خوشه‌بندی c میانگین[1] پیشنهاد داد که در آن از ترکیب وزن دار چندین تابع فاصله به عنوان معیار شباهت داده‌ها استفاده گردید. در روش پیشنهادی او از الگوریتم DE برای تعیین وزن این توابع فاصله استفاده گردید. روش پیشنهادی او با روش خوشه‌بندی c میانگین مقایسه گردید و برتری آن به اثبات رسید. LI و همکاران [۷] برای جبران نقاط ضعف الگوریتم خوشه بندی c میانگین فازی[2] (FCM) همانند به دام افتادن در بهینه‌ی محلی و حساس بودن به مقداردهی اولیه، روش خوشه‌بندی ترکیبی FCM و DE را پیشنهاد دادند. در روش پیشنهادی آنها با استفاده از الگوریتم ریلیف[3] و با توجه به ویژگی‌های خاص هر داده، یک وزن به آنها نسبت داده شد. آنها در روش خود به جای استفاده از تابع فاصله‌ی اقلیدسی از شباهت مورفولوژی وزندار داده‌ها به عنوان معیار شباهت استفاده کردند. روش پیشنهادی آنها بر روی مجموعه داده‌هایی با ابعاد کم پیاده سازی شد و با روش خوشه‌بندی FCM مقایسه گردید. نتایج حاصل به خوبی برتری روش آنها را به اثبات رساند. ایزکیان و همکاران [۸] در روش پیشنهادی خود از ترکیب الگوریتم DE و روش FCM برای خوشه‌بندی سری‌های زمانی استفاده گردید. آنها در روش خود برای کاهش مجهولات مسئله و در نتیجه افزایش کارایی الگوریتم، از روش ضرایب گسسته‌ی کسینوسی[4] (DCT) به عنوان یکی از روش‌های کاهش بعد داده استفاده کردند. بدین مفهوم که الگوریتم تکاملی دیفرانسیلی انتخابی برای خوشه‌بندی، به جای یافتن تمامی المان‌های مراکز خوشه‌های موجود در مجموعه داده، تنها تعداد محدودی از ضرایب DCT این مراکز را یافته و سپس با استفاده از همین ضرایب محدود مراکز خوشه‌ها بازسازی می‌شوند. روش پیشنهادی بر روی دو مجموعه داده پیاده‌سازی شد و با روش خوشه‌بندی FCM و روش خوشه‌بندی مبتنی بر الگوریتم تکاملی دیفرانسیلی بدون استفاده از ضرایب DCT مقایسه گردید.

در ادامه‌ی مقاله ابتدا به معرفی الگوریتم مشهور FKM[5] پرداخته و سپس الگوریتم تکاملی دیفرانسیلی را به‌عنوان اساس روش خوشه‌بندی پیشنهادی معرفی کردیم. در ادامه توضیحاتی از توابع فاصله ارائه شد و در آخر به معرفی روش پیشنهادی پرداختیم.

۲- الگوریتم خوشه‌بندی FKM

الگوریتم FKM [۹] یکی از مهم‌ترین الگوریتم‌های خوشه‌بندی تفکیکی می‌باشد که در این الگوریتم با توجه به تعداد خوشه‌ها دسته‌بندی آغاز می‌شود و با استفاده از تکنیک جابجایی تکراری، تلاش می‌کند که با جابجایی اشیا از یک خوشه به خوشه‌ی دیگر، دسته‌بندی را بهبود بخشد و در انتها هر داده به یک خوشه‌ی معین با درجه عضویت بین صفر و یک تعلق می‌گیرد. تفاوت این روش با روش خوشه‌بندی شناخته‌شده‌ی FCM در این است که که در این روش مرکز خوشه الزاما از بین داده‌های موجود انتخاب می‌شود.

اگر $X = \{x_i | i = 1,2,\dots,n\}$ مجموعه‌ای از n شیء، $r(x_i, x_j)$ نشان دهنده‌ی میزان تفاوت x_i و x_j و $V = \{v_1, v_2, \dots v_C\}$ زیر مجموعه‌ای از X با C شیء باشند، FKM تابع هدف زیر را بهینه می‌کند:

$$J(U,V) = \sum_{i=1}^{C} \sum_{j=1}^{n} u_{ij}^{m}\, r(v_C, x_j) \qquad 1 \le m < \infty$$ رابطه (۱)

در رابطه‌ی (۱)، m پارامتر فازی‌کننده نامیده می‌شود که معمولا برابر با عدد دو انتخاب می‌شود. u_{ij} نشان‌دهنده‌ی درجه عضویت x_i در خوشه‌ی j است و با روش‌های مختلفی می‌تواند تعریف شود، برای مثال می‌توان از مدل درجه عضویت الگوریتم FCM استفاده کرد.

---
[1] c-means
[2] Fuzzy C-means
[3] Relief
[4] Discrete Cosine Transform
[5] Fuzzy k-medoids

رابطه (۲)
$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{r(x_j,v_i)}{r(x_j,v_k)}\right)^{\left(\frac{1}{m-1}\right)}}$$

مراحل الگوریتم FKM در زیر آورده شده است:

۱. تعیین تعداد ثابت خوشه‌ها (C)
۲. انتخاب اولیه‌ی مراکز خوشه‌ها از بین مجموعه داده ($V=\{v_1,v_2,\ldots,v_c\}$)
۳. محاسبه‌ی درجه عضویت (u_{ij}) به ازای $i=1,2,\ldots,C$ و $j=1,2,\ldots,n$ با توجه به رابطه‌ی (۲)
۴. محاسبه مراکز خوشه‌های جدید از بین مجموعه داده با استفاده از روابط (۳و۴)

رابطه (۳)
$$q = \arg min_{1\le C\le n} \sum_{j=1}^{n} u_{ij}^{m}\ r(x_C, x_j)$$

رابطه (۴)
$$v_i = x_q$$

۵. تکرار مراحل ۳ و ۴ تا تحقق یافتن شرط توقف برنامه

۳- الگوریتم تکاملی دیفرانسیلی

الگوریتم تکاملی دیفرانسیلی[1] (DE) یکی از الگوریتم‌های بهینه سازی مبتنی بر جمعیت[2] می‌باشد [۹]. در این الگوریتم، i امین عضو (کروموزوم) جمعیت در مرحله‌ی زمانی (نسل) tام، شامل d مولفه (بعد) می‌باشد.

رابطه (۵)
$$\vec{Z}_i(t) = [Z_{i,1}(t), Z_{i,2}(t), \ldots, Z_{i,d}(t)]$$

برای هر بردار $\vec{Z}_k(t)$ که متعلق به جمعیت کنونی است الگوریتم به طور تصادفی سه عضو دیگر $\vec{Z}_i(t)$، $\vec{Z}_j(t)$ و $\vec{Z}_m(t)$ را از همان نسل و به ازای مقادیر متفاوت i، j و m انتخاب می‌کند و مولفه‌ی nام بردار $\vec{U}_k(t+1)$ به نام بردار فرزندان آزمایشی[3] طبق رابطه‌ی (۶) محاسبه می‌شود.

رابطه (۶)
$$\vec{U}_k(t+1) = \begin{cases} Z_{m,n}(t) + F\left(Z_{i,n}(t) - Z_{j,n}(t)\right), if\ rand_n(0,1) < Cr \\ Z_{k,n}(t) \qquad otherwise. \end{cases}$$

$Cr \in [0,1]$ پارامتر عددی الگوریتم می‌باشد که نرخ ترکیب[4] نامیده می‌شود و F پارامتر مقیاس است که معمولا مقداری بین صفر و یک است. اگر فرزند[5] جدید دارای فیت نس بهتری نسبت به والدینش باشد، در نسل بعدی جایگزین آنها خواهد شد. در غیر اینصورت والدین در جمعیت باقی می‌مانند.

رابطه (۷)
$$\vec{Z}_i(t+1) = \begin{cases} \vec{U}_i(t+1) \ , \ if\ f(\vec{U}_i(t+1)) > f(\vec{Z}_i(t)) \\ \vec{Z}_i(t) \ , \ if\ f(\vec{U}_i(t+1)) \le f(\vec{Z}_i(t)) \end{cases}$$

[1] Differential Evolution
[2] Population-Based
[3] trial offspring
[4] Crossover Rate
[5] Offspring

۴- توابع فاصله

یکی از مهمترین مسائلی که در خوشه‌بندی به آن توجه می‌شود، انتخاب تابع فاصله‌ی مناسب می‌باشد. یک تابع فاصله معیاری است که از آن برای تعیین میزان شباهت بین داده‌های موجود در یک مجموعه داده استفاده می‌شود. تابع فاصله‌ی اقلیدسی [۱۰]، پیچش زمانی پویا[1] (DTW) [۱۱]، طولانی ترین توالی مشترک[2] (LCSS) [۱۲] از جمله مهم‌ترین توابع برای اندازه‌گیری میزان شباهت میان سری‌های زمانی می‌باشند. در این تحقیق از تابع فاصله‌ی DTW به عنوان معیار شباهت میان سری‌های زمانی استفاده گردید. بر خلاف فاصله اقلیدسی در این روش هر نقطه از سری زمانی اول می‌تواند با هر نقطه اختیاری از سری دوم مقایسه شود، به شرطی که اختلافات آنها مینیمم شود. در این تابع فاصله مسیرهایی با ساختار مشابه که در پریودهای زمانی متفاوت بوجود آمده باشند، متشابه قلمداد می‌شوند. به عبارتی دیگر استفاده از تابع فاصله‌ی DTW در خوشه‌بندی داده‌ها به منزله‌ی گروهبندی آن‌ها بر اساس شکل داده می‌باشد. شبه کد مربوط به DTW برای دو سری زمانی x و y با طول m و k در شکل (۱) آورده شده است.

```
DTW( x = x_1,...,x_m , y = y_1,...,y_k )
C=array[0…m][0…k];
For i=0,1,…,m    C[i][0]=∞ ; End
For j=0,1,…,k    C[0][j]=∞ ; End
C[0][0]=0;
For i=1,2,…,m
    For j=1,2,…k
        C[i][j]= { (D(x_i, y_j))^2 + (min(C[i-1][j], C[i][j-1], C[i-1][j-1])^2 }^(1/2) ;
End
End
return C[m][k]
```

شکل (۱): شبه کد تابع فاصله‌ی DTW [۱۱]

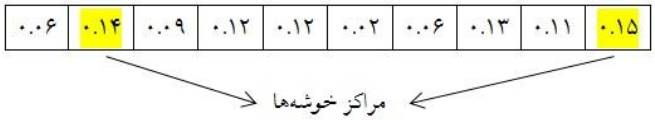
از مزایای این تابع فاصله این است که قابلیت اندازه گیری فاصله بین دو سری با طول‌های متفاوت را دارا بوده و از شیفت زمانی محلی حمایت می‌کند و از معایب این روش می‌توان به زمان‌گیر بودن آن اشاره کرد.

۵- روش پیشنهادی

در روش پیشنهادی از الگوریتم تکاملی دیفرانسیلی برای خوشه‌بندی سری‌های زمانی استفاده شد. امروزه الگوریتم تکاملی دیفرانسیلی به‌عنوان یکی از روش‌های جستجوی کلی سریع، قوی و کارآمد مطرح بوده و نتایج خوبی در استفاده از این الگوریتم برای خوشه‌بندی داده‌ها بدست آمده است [۱]. از مهم‌ترین مزایای این الگوریتم پیاده‌سازی آسان و نیاز به تنظیم تعداد کم پارامترها در مقایسه با الگوریتم‌هایPSO و GA برای رسیدن به جواب مناسب می‌باشد.

[1] Dynamic Time Warping
[2] Longest Common Subsequence

چگونگی تعریف کروموزوم‌ها مهم‌ترین بخش را در این روش به خود اختصاص می‌دهد. اگر مجموعه داده شامل N سری زمانی باشد، هر کروموزوم از N المان تشکیل می‌شود که این المان‌ها به صورت تصادفی مقادیر بین صفر و یک را می‌پذیرند و مجموع این N عدد صحیح برابر با عدد یک می‌شود. به عبارتی به هر سری زمانی موجود در مجموعه داده یک احتمال نسبت داده می‌شود که هرچه این مقدار به عدد یک نزدیک‌تر باشد، احتمال مرکز خوشه بودن آن داده بیشتر است. بنابراین اگر یک مجموعه داده شامل C مرکز خوشه باشد، در هر کروموزوم تعداد C داده با بیشترین عدد احتمال به‌عنوان مراکز خوشه‌ها در آن کروموزوم شناخته می‌شود. به عنوان مثال با فرض وجود ده داده در مجموعه‌ای شامل دو خوشه، یک کروموزوم می‌تواند به صورت شکل (۲) نشان داده شود.



شکل (۲): نمایی از یک کروموزوم در روش پیشنهادی

در این کروموزوم عدد احتمال متناظر با داده‌های دوم و دهم بیشترین مقدار را به خود اختصاص داده‌اند بنابراین این دو داده به عنوان مراکز خوشه‌های این کروموزوم در نظر گرفته می‌شوند.

تابع هدف در نظر گرفته شده در این روش تابع هدف الگوریتم FKM طبق رابطه (۱) می‌باشد. در این تابع هدف اصول یک خوشه‌بندی درست (شرط فشردگی اعضای یک خوشه و پراکندگی خوشه‌های مختلف) به خوبی رعایت شده است. ماتریس مربوط به مقادیر درجه عضویت‌ها نیز طبق رابطه (۲) محاسبه می‌گردد.

برای تعیین میزان شباهت داده‌ها در عملیات خوشه‌بندی انتخاب تابع فاصله‌ی مناسب امری بدیهی می‌باشد. علاوه بر اینکه ویژگی‌های مثبت و منفی یک تابع فاصله در انتخاب آن به عنوان معیار شباهت مسیرها در عملیات خوشه‌بندی نقش مهمی دارد، هدف و کاربرد خوشه‌بندی داده‌های موجود در یک مجموعه نیز از دیگر عوامل موثر در انتخاب تابع فاصله‌ی مناسب می‌باشد. برای مثال اگر در یک مجموعه داده شامل سری‌های زمانی هدف از خوشه‌بندی، گروه‌بندی سری‌های زمانی بر اساس شکل آنها و با نادیده گرفتن شیفت زمانی باشد بهتر است از توابع فاصله‌ی DTW یا LCSS به عنوان معیار شباهت استفاده شود. در این روش از تابع فاصله‌ی DTW به عنوان معیار شباهت داده‌ها استفاده گردید که در بخش (۴) توضیحات مربوط به این تابع فاصله ارائه شد. شبه کد روش پیشنهادی در شکل (۳) آورده شده است.

```
1: Create and initialize P chromosomes with N probability value (N is the size of data
   set)
2: FOR iteration_count = 1 to maximum_iterations DO
3:     FOR each chromosome i DO
4:         Select C data in each chromosome with maximum probability value as
           cluster centers
5:         Calculate DTW distance matrix.
5:         Calculate partition matrix (U) using E.q.(2).
Calculate the fitness function f (Z_i) using E.q.(1). 6:
```

شکل(۳): شبه کد روش پیشنهادی

در گام اول روش پیشنهادی کروموزوم‌ها به صورت تصادفی تولید می‌شوند. در مجموعه داده‌ای با تعداد N داده، هر کروموزوم شامل تعداد N تعداد عدد احتمال است که این اعداد، احتمال مرکز خوشه بودن داده‌ی متناظر را نشان می‌دهند. در مرحله‌ی بعد در هر کروموزوم تعداد C (تعداد خوشه‌های موجود در مجموعه) سری زمانی با بیشترین عدد احتمال را به عنوان مراکز خوشه‌ها در نظر گرفته و فاصله‌ی DTW این مراکز را از داده‌های موجود در مجموعه داده طبق شبه کد شکل (۱) محاسبه می‌کنیم. سپس ماتریس درجه عضویت را با استفاده از معادله‌ی (۲) بدست آورده و در گام بعدی مقدار تابع هدف مربوط به کروموزوم مورد نظر طبق رابطه‌ی (۱) محاسبه می‌گردد. در مرحله ی بعد بردار آزمایشی[1] متناظر با هر کروموزوم را با استفاده از رابطه‌ی (۶) محاسبه کرده و سپس با توجه به مقدار بهینگی مربوط به کروموزوم اولیه و بردار آزمایشی متناظر با آن، بهترین کروموزوم را برای انتقال به نسل بعدی انتخاب می‌کنیم. بعد از تکرار مراحل بالا به تعداد کافی، با توجه به بهترین کروموزوم حاصل، مراکز خوشه‌ها استخراج می‌گردد.

۶- پیاده‌سازی و اجرا

در این بخش ابتدا به معرفی مجموعه داده‌های استفاده شده در این تحقیق، مقادیر تنظیم شده برای پارامترها و معیارهای در نظر گرفته برای ارزیابی نتایج پرداخته و سپس نتایج بدست آمده را مورد بحث و بررسی قرار می‌دهیم.

۶-۱- مجموعه داده‌ها

روش پیشنهادی بر روی هشت مجموعه داده‌ی شناخته شده [۱۴،۱۳ و ۱۵] از سری‌های زمانی پیاده شد که خصوصیات این مجموعه‌ها به طور مختصر در جدول (۱) آورده شده است.

جدول ۱: خصوصیات مجموعه داده‌های استفاده شده

مجموعه داده	طول سری‌های زمانی	اندازه‌ی مجموعه داده	تعداد خوشه‌های موجود در مجموعه
Beef	۴۷۰	۶۰	۵
FaceFour	۳۵۰	۱۱۲	۴
FISH	۴۶۳	۳۵۰	۷
Lighting2	۶۳۷	۱۲۱	۲
Lighting7	۳۱۹	۱۴۳	۷
OliveOil	۵۷۰	۶۰	۴
OSULeaf	۴۲۷	۴۴۲	۶
Synthetic_control	۶۰	۶۰۰	۶

۶-۲- تنظیم پارامترها

مقادیر پارامترهای موجود در هر روش در جدول (۲) آورده شده است. این پارامترها با روش سعی و خطا به‌دست آمده‌اند.

[1] trial

جدول ۲: مقادیر پارامترها

پارامتر	FKM	DE
m	۲	۲
P	NA	۵۰
Itr	۴۰	۲۰۰
Cr	NA	۰.۵
F	NA	۱

در جدول (۲)، m پارامتر فازی کننده، P اندازه‌ی جمعیت، Itr تعداد تکرارها و Cr و F مقدار نرخ ترکیب[1] و فاکتور مقیاس می‌باشند.

۶-۳- معیارهای ارزیابی نتایج

برای ارزیابی و مقایسه‌ی روش پیشنهادی با روش خوشه‌بندی FKM علاوه بر تابع هدف، از دو معیار دقت[2] و معیار F[3] استفاده شد. اگر n_i، n_j و n_{ij} به ترتیب برابر با تعداد داده‌های کلاس i، تعداد داده‌های خوشه‌ی j (به‌دست آمده توسط الگوریتم خوشه‌بندی) و تعداد داده‌های کلاس i در خوشه‌ی j باشند، معیارهای دقت و F به شرح زیر می‌باشند[۱،۸].

دقت: این معیار طبق رابطه‌ی (۸) بدست می‌آید. بیشتر بودن این معیار در یک روش بالاتر بودن دقت آن روش را نشان می‌دهد.

$$P(i,j) = \frac{n_{ij}}{n_j}$$ رابطه (۸)

معیار F: این معیار مقادیری بین صفر و یک می‌پذیرد که مقادیر نزدیک به یک مقادیری مطلوب هستند.

$$F(i,j) = \frac{2\times P(i,j)\times R(i,j)}{P(i,j)+R(i,j)}$$ رابطه (۹)

در رابطه‌ی (۹)، مقدار R(i,j) از رابطه‌ی (۱۰) محاسبه می‌شود.

$$P(i,j) = \frac{n_{ij}}{n_i}$$ رابطه (۱۰)

۶-۴- نتایج حاصل و تفسیر آن

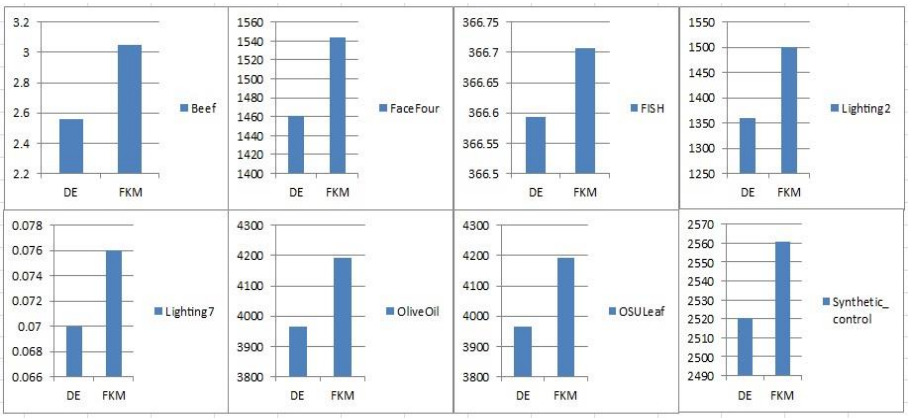
در جدول (۳) میانگین و انحراف معیار مقادیر تابع هدف ، دقت و معیار F حاصل از پیاده‌سازی روش پیشنهادی بر روی هشت مجموعه داده در ده بار تکرار مستقل نشان داده شده و با نتایج حاصل از روش خوشه‌بندی FKM مقایسه گردید.

[1] Crossover Rate
[2] Precision
[3] F-measure

جدول ۳: نتایج حاصل از پیاده‌سازی روش مبتنی بر الگوریتم DE و FKM بر روی هشت مجموعه داده

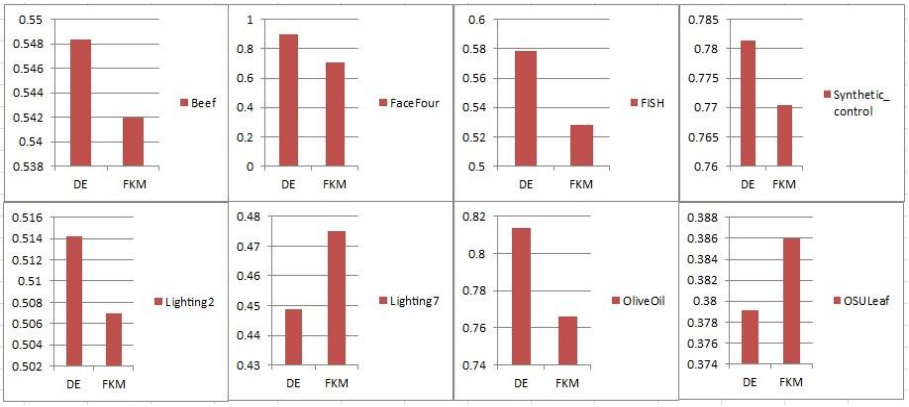
مجموعه داده	روش	مقدار تابع هدف	Pr	Fm
Beef	DE	۰.۰۰۱۷±۲.۵۵۸	۰±۰.۵۴۸۴	۰.۰۰۱۹±۰.۴۹۱۳
	FKM	۰.۲۲۳۵±۳.۰۴۷	۰.۰۱۷۳±۰.۵۴۲۱	۰.۰۳۱۵±۰.۵۱۷۳
FaceFour	DE	۵.۳۵۱۱±۱۴۶۱.۰۳	۰.۰۱۶۲±۰.۸۹۹	۰.۰۱۷۱±۰.۸۷۳۲
	FKM	۵۱.۶۹۲۲±۱۵۴۴.۱۲۷	۰.۰۹۹۲±۰.۷۰۹۸	۰.۰۸۲۹±۰.۶۹۰۷
FISH	DE	۰.۲۱۴۷±۶۸.۹۵۳۱	۰.۰۱۷۹±۰.۵۷۸۴	۰.۰۱۹±۰.۵۳۲۱
	FKM	۱.۱۰۱±۱۰۱.۳۸۳	۰.۰۵۶۹±۰.۵۲۸۱	۰.۰۵۰۵±۰.۴۹۲۹
Lighting2	DE	۱۶.۲۴۴۶±۶۹۲۹.۳۱	۰.۰۱۱±۰.۵۱۴۲	۰.۰۰۲۱±۰.۳۸۲۱
	FKM	۱۹۱.۷۱±۷۱۸۴.۶۱	۰.۰۴۸±۰.۵۰۷	۰.۰۳±۰.۳۷۹
Lighting7	DE	۲.۲۴۲۳±۱۳۶۰.۹۲	۰.۰۴۲۹±۰.۴۴۸۸	۰.۰۴۴۷±۰.۴۳۸۳
	FKM	۷۷.۴۴±۱۵۰۰.۲۸	۰.۰۴۹۵±۰.۴۷۵	۰.۰۵±۰.۴۴۹
OliveOil	DE	۰±۰.۰۷	۰.۰۴۹۸±۰.۸۱۴۷	۰.۰۲۵۳±۰.۷۹۸۹
	FKM	۰.۰۰۸±۰.۰۷۶	۰.۰۸۶±۰.۷۶۶	۰.۱۱±۰.۷۲۸
OSULeaf	DE	۳۳.۴۳۳۲±۳۹۶۵.۹	۰.۰۱۵۶±۰.۳۷۹۱	۰.۰۱۴۲±۰.۳۷۶۴
	FKM	۸۱.۷±۴۱۹۱.۰۶	۰.۰۶±۰.۳۸۶	۰.۰۳±۰.۳۵۹
Synthetic_control	DE	۲۱.۹۲۷۸±۲۵۲۰.۱۷	۰.۰۹۰۴±۰.۷۸۱۴	۰.۰۸۴۷±۰.۷۷۱۸
	FKM	۸۱.۶۹±۲۵۶۰.۸۳	۰.۰۸۵±۰.۷۷۰۴	۰.۰۷۲±۰.۷۶۷

برای مقایسه‌ی بهتر مقادیر به‌دست آمده برای تابع هدف با استفاده از دو روش پیشنهادی مبتنی بر DE و روش خوشه‌بندی FKM، تفاوت‌ها را با استفاده از نمودارهای میله‌ای در شکل (۴) نمایش دادیم.



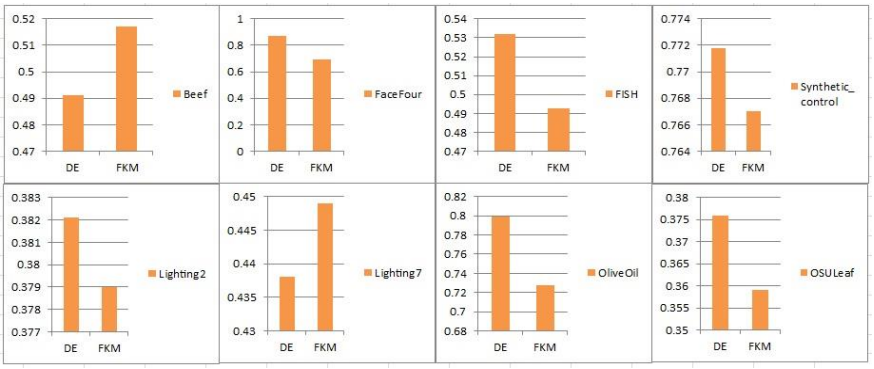
شکل (۴): مقایسه‌ی مقادیر بدست آمده برای تابع هدف

همانگونه که از شکل (۴) مشهود است، مقدار بدست آمده برای تابع هدف در روش پیشنهادی در تمامی مجموعه داده‌ها کوچکتر از روش خوشه‌بندی FKM می‌باشد که این امر دلیلی بر عملکرد بهتر روش پیشنهادی می‌باشد. همچنین در شکل (۵) مقادیر بدست آمده برای دقت در دو روش با استفاده از نمودار باهم مقایسه گردیدند.



شکل ۵: مقایسه‌ی مقادیر بدست آمده برای دقت

با توجه به شکل (۵) مقدار Precision بدست آمده از روش پیشنهادی در شش سری مجموعه داده از مجموع هشت مجموعه داده‌ی استفاده شده در این تحقیق، بیشتر (بهتر) از مقدار حاصل با استفاده از روش FKM می‌باشد. همچنین شکل (۶) بیانگر این نکته است که با توجه به مقدار معیار F حاصل از دو روش مورد مقایسه، روش پیشنهادی در شش مجموعه داده عملکرد بهتری نسبت به روش دیگر داشته است.



شکل ۶: مقایسه‌ی مقادیر بدست آمده برای معیار F

۷- نتیجه‌گیری

امروزه مسئله‌ی خوشه‌بندی داده‌ها و مخصوصا داده‌هایی با طول زیاد مانند سری زمانی که شامل پیچیدگی‌های خاص خودش است، از مهمترین مسائل موجود در مبحث داده‌کاوی می‌باشد. گستردگی فضای جستجو، پیچیدگی، چند هدفه بودن و انباشت زیاد اطلاعاتی به‌عنوان برخی از مهم‌ترین ویژگی‌های مسئله‌ی خوشه‌بندی داده مطرح می‌باشد. از طرف دیگر خصوصیت اصلی سری‌های زمانی که طول بالای این نوع داده‌ها می‌باشد، مشکلات مربوط به خوشه‌بندی این نوع داده را دوچندان کرده است. الگوریتم‌های تکاملی به دلیل توانایی غلبه بر مسائل پیچیده و چندهدفه و شکل تصادفی جستجو در فضا توان یافتن جواب‌هایی نزدیک به بهینه را در مسائلی مانند خوشه‌بندی دارا می‌باشند. از مهم‌ترین ویژگی‌های الگوریتم‌های تکاملی مقابله با بهینه‌ی محلی با نگهداری، ترکیب و مقایسه‌ی چندین راه حل کاندید می‌باشد. در حالی که در روش‌های جستجوی قطعی[1] مانند روش خوشه‌بندی FKM همواره احتمال افتادن در بهینه‌ی محلی وجود دارد. با توجه به عملکرد قابل قبول الگوریتم تکاملی دیفرانسیلی در خوشه‌بندی داده‌های پیچیده و با توجه به نقاط قوت این الگوریتم از جمله پیاده‌سازی آسان و تعداد کم پارامترها برای مقداردهی اولیه نسبت به سایر الگوریتم‌ها مانند PSO و GA، در این تحقیق یک روش خوشه‌بندی فازی مبتنی بر الگوریتم تکاملی DE

[1] Deterministic local search

به منظور خوشه‌بندی مجموعه داده شامل سری‌های زمانی پیشنهاد داده شد. روش پیشنهادی بر روی تعداد هشت مجموعه داده پیاده سازی شده و با روش FKM مقایسه گردید. در این تحقیق نشان داده شد که الگوریتم تکاملی DE به شرط تعریف درست کروموزوم‌ها نتایج مطلوبی را در خوشه‌بندی بدست می‌دهد.

مراجع

[1] Z.Izakian, M.S.Mesgari, "Fuzzy clustering of time series data: A particle swarm optimization approach, " Journal of Artificial Intelligence and Data Mining, vol. 3, no. 1, pp. 39–46, 2015.

[2] A. Ahmadyfard, H. Modares, "Combining PSO and K-means to enhance data clustering," In: International symposium on telecommunications. Tehran, Iran, pp.688–691, 2008.

[3] N. I. Ghali, N. El-Dessouki, Mervat A. N., L. Bakrawi, "Exponential Particle Swarm Optimization Approach for Improving Data Clustering, International Journal of Electrical," Computer, and Systems Engineering, pp. 208-212, 2009.

[4] S. Das, A. Abraham, A. Konar, "Automatic clustering using an improved differential evolution algorithm, " IEEE Trans. Syst. Man Cybern. Part A, vol. 38, no. 1, pp.218 -236, 2008.

[5] S. Das, A. Konar, "Automatic image pixel clustering with an improved differential evolution", Appl. Soft Comput., vol. 9, no. 1, pp.226 -236, 2009.

[6] M. Fuad, "Differential evolution-based weighted combination of distance metrics for k-means clustering", In: Dediu, A.-H., Lozano, M., Martín-Vide, C. (eds.) TPNC 2014. LNCS, vol. 8890, pp. 193–204. Springer, Heidelberg 2014.

[7] K. Li, C. Zhang, Z. Chen, Y. Chen, "Development of a weighted fuzzy c-means clustering algorithm based on JADE," International Journal of Numerical Analysis and modeling, Vol.5, pp. 113–122, 2014.

[8] Z. Izakian, Y. Amerian, M.S. Mesgari , "Time Series Data Clustering Based on Differential Evolution Algorithm and Discrete Cosine Transform, " JGST, 5 (4), pp. 199-209, 2016.

[9] Z.Izakian, M.S.Mesgari, A. Abraham," Automated clustering of trajectory data using a particle swarm optimization, " Computers, Environment and Urban Systems, vol. 55, pp. 55–65, 2016

[10] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang and E. Keogh, "Querying and mining of time series data: experimental comparison of representations and distance measures," VLDB, 2008.

[11] D. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," Proc. AAAI-94 Workshop Knowledge Discovery in Databases, pp. 359-370, 1994.

[12] M. Vlachos, D. Gunopulos, and G. Kollios, " Discovering similar multidimensional trajectories," In ICDE, 2002.

[13] E.Keogh, P.Smyth, "A probabilistic approach to fast pattern matching in time series databases," In: Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining, pp. 24–20 (1997).

[14] H. Zhang, T.B. Ho, Y. Zhang, and M.S. Lin, "Unsupervised Feature Extraction for Time Series Clustering Using Orthogonal Wavelet Transform," Informatica, pp. 305-319, 2006.

[15] M.chis, S.Banerjee and A.E.Hassanien, "Clustering Time Series Data: An Evolutionary Approach," Foundations of Computational Intelligence, Volume 206, pp.193-207: Springer , 2008.