نشریه علمی- پژوهشی علوم و فنون نقشه برداری، دوره پنجم، شماره ۴، اردیبهشت ماه ۱۳۹۵

ارائه‌ی یک روش خوشه‌بندی سری‌های زمانی بر مبنای الگوریتم تکاملی دیفرانسیلی و تبدیل کسینوسی گسسته

زاهده ایزکیان*[۱]، یزدان عامریان[۲]، محمدسعدی مسگری[۲]

[۱]دانشجوی دکتری سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه نصیرالدین طوسی
zahedeh_izakian@yahoo.com

[۲]استادیار دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه نصیرالدین طوسی
{Amerian, mesgari}@kntu.ac.ir

(تاریخ دریافت تیر ۱۳۹۴، تاریخ تصویب بهمن ۱۳۹۴)

چکیده

با پیشرفت روز افزون تکنولوژی‌های جمع آوری اطلاعات و امکان دسترسی به حجم عظیمی از داده همواره نیازمند روش‌هایی برای تجزیه و تحلیل این حجم داده خام و استخراج اطلاعات مفید از آن می‌باشیم. امروزه خوشه‌بندی داده به عنوان یکی از روش‌های آنالیز و ساده سازی مجموعه داده‌های بزرگ، مورد توجه بسیاری از محققین قرار گرفته است. در این میان خوشه‌بندی سری‌های زمانی با دقت مورد قبول، حائز اهمیت بسیاری می‌باشد. در روش پیشنهادی از ترکیب الگوریتم تکاملی دیفرانسیلی و روش خوشه Fuzzy-Cmeans به عنوان یکی از الگوریتم‌های خوشه‌بندی مطرح و شناخته شده، برای خوشه‌بندی سری‌های زمانی استفاده گردید. در این روش برای کاهش مجهولات مسئله و در نتیجه افزایش کارایی الگوریتم، تکنیک‌های مختلف نمایش داده‌های مکانی-زمانی را مورد بررسی قرار دادیم و از این میان روش ضرایب DCT را برای کاهش مجهولات مراکز خوشه‌ها انتخاب کردیم. 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. با در نظر گرفتن تابع فاصله‌ی Dynamic Time Warping و انتخاب تابع بهینه‌سازی مربوط به روش خوشه‌بندی Fuzzy-Cmeans، 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. روش پیشنهادی کندتر از الگوریتم خوشه‌بندی Fuzzy-Cmeans بوده اما به دلیل استفاده از روش تبدیل کسینوسی گسسته برای کاهش مجهولات، سریع‌تر از روش خوشه‌بندی معمول مبتنی بر الگوریتم تکاملی دیفرانسیلی عمل می‌کند. همچنین نتایج حاصل از مقایسه‌ی این سه روش نشان‌دهنده‌ی عملکرد بهتر روش پیشنهادی نسبت به دو روش دیگر می‌باشد.

واژگان کلیدی: سری زمانی، خوشه‌بندی، الگوریتم تکاملی دیفرانسیلی، روش خوشه‌بندی Fuzzy-Cmeans، تبدیل کسینوسی گسسته

* نویسنده رابط

۱- مقدمه

سری‌های زمانی ترتیبی از اعداد حقیقی هستند که مقدار مشاهده شده از یک رویداد را در فواصل زمانی برابر نشان میدهند. تجزیه و تحلیل این سری‌ها در علومی مانند هواشناسی، اقتصاد، زمین‌شناسی، علوم دریایی، پزشکی و مهندسی کاربرد فراوان دارد. به طور کلی تجزیه و تحلیل سری‌های زمانی در این زمینه‌ها موجب برآورد یک مدل مناسب برای نظارت و پیش‌بینی‌های دقیق شده و اطلاعات مفیدی را از چگونگی تاثیر مجموعه‌ای از ورودی‌های مشخص بر روی خروجی‌های مورد نظر مسئله در اختیار قرار می‌دهد [۱].

امروزه با پیشرفت تکنولوژی جمع آوری اطلاعات مانند موبایل، جی پی اس و سنسور امکان جمع‌آوری میزان زیادی از داده‌های مکانی-زمانی از جمله سری‌های زمانی فراهم گردید و تمایل به استفاده از تکنیک‌های داده‌کاوی برای استخراج اطلاعات مفید از داخل انبوهی از داده‌ها بیشتر شد.

یکی از تکنیک‌های تجزیه و تحلیل بر روی سری‌های زمانی که طی سال‌های اخیر مورد توجه دانشمندان قرار گرفته است، تکنیک خوشه‌بندی می‌باشد. عملیات خوشه-بندی عبارت است از گروه بندی اشیاء با ویژگی‌های مشابه به گونه‌ای که اشیاء موجود در یک گروه، بیشترین شباهت را نسبت به هم و بیشترین تفاوت را با اشیاء موجود در سایر گروه‌ها داشته باشند. به طور کلی هدف از خوشه‌بندی نمایش یک مجموعه داده‌ی بزرگ به کمک خوشه‌ها و مراکز آنها می‌باشد که این امر باعث ساده‌سازی و استخراج اطلاعات مفید در مجموعه داده‌های حجیم می‌گردد [۲].

روش خوشه‌بندی FCM[۱] [۳] یکی از مهم‌ترین روش-های خوشه‌بندی کلاسیک می‌باشد که تاکنون در بسیاری از تحقیقات محققان مورد استفاده قرار گرفته است. از مهمترین معایب این روش می‌توان به احتمال بالای به دام افتادن جواب‌ها در بهینه‌ی محلی اشاره کرد که با پیچیده‌تر شدن داده‌ها این احتمال افزایش می‌یابد [۴]. از طرف دیگر تابع فاصله‌ی معمول مورد استفاده در روش خوشه‌بندی FCM تابع فاصله‌ی اقلیدسی می‌باشد در حالی که در بسیاری موارد نیازمند در نظر گرفتن معیار شباهتی دیگر برای خوشه‌بندی داده‌ها هستیم. به عبارتی دیگر در این روش از متوسط گیری داده‌های موجود در یک خوشه برای بدست آوردن مرکز آن خوشه استفاده می‌گردد و در صورتی که معیار شباهت مورد نظر تابع فاصله‌ای مانند DTW[۲] باشد، استفاده از میانگین‌گیری مناسب به نظر نمی‌رسد. با توجه به توضیحات داده شده و به دلیل عملکرد ضعیف الگوریتم خوشه‌بندی FCM در مواجه با داده‌های پیچیده مانند سری‌های زمانی، در این تحقیق به دنبال روشی برای ترکیب با روش خوشه‌بندی FCM برای خوشه‌بندی سری‌های زمانی با طول بلند و به منظور جبران نقاط ضعف روش FCM بوده‌ایم.

امروزه الگوریتم بهینه‌سازی تکاملی دیفرانسیلی[۳] به عنوان یکی از روش‌های جستجوی کلی سریع، قوی و کارآمد مطرح است که کارایی این الگوریتم در خوشه‌بندی داده‌ها در تحقیقات محققان به اثبات رسیده است. در ادامه به بررسی نمونه‌هایی از روش‌های خوشه‌بندی مبتنی بر الگوریتم‌های تکاملی میپردازیم که توسط محققان پیشنهاد داده شده است.

Paterlinia و Krink [۵] در مقاله‌ی خود نشان دادند زمانی که با مجموعه داده‌هایی نسبتا پیچیده با بعد زیاد مواجه میشویم که در آنها تعداد خوشه‌های زیاد با همپوشانی بالا وجود دارند، الگوریتم‌هایی مانند PSO، GA و DE کارایی بیشتری نسبت به روش‌های C-means و RS (جستجوی تصادفی) دارند. در این میان الگوریتم DE از نظر دقت و استحکام در نتایج بدست آمده در اجراهای مختلف، بهتر از PSO و GA عمل می‌کند. کار آن‌ها بر روی خوشه‌بندی غیراتوماتیک داده‌ها متمرکز شده بود و تعداد خوشه‌های موجود در مجموعه داده به عنوان معلومات مسئله به شمار می‌رفت. Omran و همکاران [۶] یک روش خوشه‌بندی قطعی[۴] بر اساس الگوریتم DE پیشنهاد دادند که در روش آن‌ها تعداد خوشه‌ها توسط کاربر مشخص می‌شود. آنها با پیاده سازی روش پیشنهادی خود و روش‌های خوشه بندی مانند FCM، PSO و GA بر روی سه مجموعه داده و مقایسه‌ی آن‌ها عملکرد بهتر روش پیشنهاد خود را نشان دادند. Das و همکاران [۱] از

[۱] Fuzzy C-means

[۲] Dynamic Time Warping
[۳] Differential evolution
[۴] Crisp

الگوریتم تکاملی دیفرانسیلی 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 به عنوان فاکتور مقیاس، از عددی متغیر در بازه‌ی (۰٫۵و۱) استفاده کردند. نتایج بدست آمده از پیاده‌سازی روش تکاملی دیفرانسیلی بهبود یافته بر روی تعدادی مجموعه داده به خوبی کارایی این روش را نشان داد. LI و همکاران [۷] برای جبران نقاط ضعف الگوریتم خوشه بندی FCM همانند به دام افتادن در بهینه‌ی محلی و حساس بودن به مقداردهی اولیه، روش خوشه‌بندی ترکیبی FCM و DE را برای خوشه‌بندی داده‌های نقطه‌ای پیشنهاد دادند. در روش پیشنهادی آنها با استفاده از الگوریتم ReliefF 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. Das و همکاران [۸] از الگوریتم تکاملی دیفرانسیلی 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 [۹] در مقاله‌ی خود یک روش جدید خوشه‌بندی سری‌های زمانی بر مبنای روش خوشه‌بندی k-means 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 k-means مقایسه گردید و برتری آن به اثبات رسید.

کار کردن با سری‌های زمانی با طول زیاد فرایندی پیچیده می‌باشد از این رو استفاده از تکنیک‌های کاهش بعد سری‌های زمانی و سپس استفاده از روش‌های خوشه-بندی، روشی کارا و موثر به نظر می‌رسد.

Vlachos و همکاران [۱۰] روش جدیدی را برای خوشه‌بندی سری‌های زمانی پیشنهاد دادند. در روش پیشنهادی آنان عملیات خوشه بندی در چند سطح مختلف و با رزولیشن‌های مختلف انجام می‌شود. بدین منظور آنها در روش خود از تکنیک کاهش بعد تبدیل موجک استفاده کرده و سپس با استفاده از روش خوشه-بندی k-means عملیات خوشه‌بندی انجام می‌شود. بدین صورت که مراکز بدست آمده در هر سطح به عنوان مقادیر اولیه در سطح بالاتر مورد استفاده قرار می‌گیرد. Powell و همکاران [۱۱] در مقاله‌ی خود روش‌های طبقه‌بندی نظارت نشده را با روش‌های طبقه بندی نظارت شده با هدف پیش‌بینی قیمت سهام مقایسه کردند. در مقاله‌ی آنها از تکنیک PCA[۱] برای کاهش بعد سری‌های زمانی و انتخاب مولفه‌های با بیشترین تاثیر استفاده شد. آنها به این نتیجه رسید که روش‌های نظارت نشده کارایی بهتری در پیش‌بینی قیمت سهام دارند. GUO و همکاران [۱۲] در مقاله‌ی خود یک روش جدید برای خوشه‌بندی سری-های زمانی ارائه کردند. در روش پیشنهادی آنها ابتدا سری‌های زمانی با استفاده از روش کاهش بعد ICA[۲] به سری‌هایی با طول کمتر تبدیل می‌شوند و سپس با استفاده از روش خوشه‌بندی بهبود یافته‌ی k-means عملیات خوشه‌بندی آنها صورت می‌گیرد. در آخر روش پیشنهادی آنها بر روی مجموعه داده‌ی واقعی مربوط به قیمت سهام پیاده سازی شده و کارایی آن به اثبات رسید. Yu و همکاران [۱۳] در مقاله‌ی خود یک روش خوشه-بندی برای جریان داده پیشنهاد دادند که در روش آنها داده‌های اصلی با کمک تکنیک تبدیل کسینوسی گسسته کاهش بعد یافته و سپس از یک الگوریتم خوشه‌بندی

۱ Principal component analysis
۲ Independent component analysis

شبکه‌ای مبتنی بر تراکم برای خوشه‌بندی داده‌های انتقال داده شده استفاده کردند.

با توجه به توانمندی الگوریتم تکاملی دیفرانسیلی در خوشه‌بندی داده‌های پیچیده و با توجه به نقاط قوت این الگوریتم از جمله پیاده سازی آسان و تعداد کم پارامترها برای مقداردهی اولیه نسبت به الگوریتم‌هایی مانند PSO و GA [۱و۹]، در این تحقیق با ترکیب روش خوشه‌بندی FCM و الگوریتم تکاملی دیفرانسیلی سعی در بهبود عملکرد روش FCM داشته‌ایم. روش پیشنهادی برای خوشه‌بندی سری‌های زمانی با طول زیاد معرفی شده است و از آنجا که در این نوع مجموعه داده‌ها با تعداد زیادی المان‌های مجهول مربوط به مراکز خوشه‌ها مواجه‌ایم، برای کاهش مجهولات مسئله و افزایش کارایی الگوریتم تکاملی دیفرانسیلی استفاده شده، از یکی از روش‌های کاهش بعد سری‌های زمانی استفاده کردیم.

۲- روش تحقیق

در این بخش به معرفی روش خوشه‌بندی مشهور FCM پرداخته و سپس الگوریتم تکاملی دیفرانسیلی را به عنوان اساس روش خوشه‌بندی پیشنهادی معرفی کردیم. در ادامه توضیحاتی از تابع فاصله DTW ارائه شد و یکی از روش‌های کاهش بعد سری‌های زمانی معرفی شده و در آخر به معرفی روش پیشنهادی پرداختیم.

۲-۱- روش خوشه‌بندی Fuzzy C-means

یکی از معروف‌ترین مدل‌های خوشه‌بندی تفکیکی فازی روش FCM [۳] می‌باشد که درواقع نسخه‌ی بهبود-یافته‌ی روش C-Means است. در روش C-Means هر شیء تنها به یک خوشه تعلق می‌گیرد در حالی که در FCM هر شیء می‌تواند به چند خوشه با درجه عضویت بین صفر و یک تعلق داشته باشد. الگوریتم FCM تابع هدف زیر را بهینه می‌کند [۱۴]:

$$J(U,V)=\sum_{i=1}^{C}\sum_{j=1}^{n}u_{ij}^{m}d^{2}(v_i,x_j)\quad 1\le m\le\infty \qquad (۱)$$

در رابطه ی ۱، m پارامتر فازی کننده نامیده می شود که معمولا برابر با عدد دو انتخاب می شود و C نشان دهنده‌ی تعداد خوشه‌ها می‌باشد. uij نشان دهنده ی درجه عضویت xi در خوشه ی j است و با استفاده از رابطه‌ی زیر بدست می آید.

$$u_{ij}=\frac{1}{\sum_{k=1}^{C}\left(\frac{d(x_j,v_i)}{d(x_j,v_k)}\right)^{2/(m-1)}} \qquad (۲)$$

همچنین ماتریس مربوط به مراکز خوشه‌ها از رابطه‌ی (۳) حاصل می‌شود.

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}} \qquad (۳)$$

۲-۲- الگوریتم تکاملی دیفرانسیلی

الگوریتم تکاملی دیفرانسیلی یا DE یکی از الگوریتم-های بهینه سازی مبتنی بر جمعیت می‌باشد. در این الگوریتم، i امین عضو (کروموزوم) جمعیت در مرحله‌ی زمانی (نسل) tام، شامل d مولفه (بعد) می‌باشد.

$$\vec{Z_i}(t)=[z_{i,1}(t),z_{i,2}(t),...,z_{i,d}(t)] \qquad (۴)$$

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 trial offspring طبق رابطه‌ی (۵) محاسبه می‌شود.

$$\vec{U_k}(t+1)=\begin{cases}z_{m,n}(t)+F(z_{i,n}(t)-z_{j,n}(t))\\ \text{if } rand_n(0,1)\prec C_r\\ z_{k,n}(t)\ otherwise\end{cases} \qquad (۵)$$

$C_r\in[0,1]$ پارامتر عددی الگوریتم می‌باشد که نرخ ترکیب[۱] نامیده می‌شود و F پارامتر مقیاس است که معمولا مقداری بین صفر و یک است. اگر فرزند جدید دارای مقدار بهینگی بهتری نسبت به والدینش باشد، در نسل بعدی جایگزین آنها خواهد شد. در غیر اینصورت والدین در جمعیت باقی می‌مانند.

۱ Crossover Rate

$$\overrightarrow{Z_i}(t+1)=\begin{cases}\overrightarrow{U_i}(t+1)\ , if\quad f(\overrightarrow{U_i}(t+1))\succ f(\overrightarrow{Z_i}(t))\\ \overrightarrow{Z_i}(t)\ , if\quad f(\overrightarrow{U_i}(t+1))\preceq f(\overrightarrow{Z_i}(t))\end{cases} \qquad (۶)$$

۲-۳- توابع فاصله

انتخاب معیار شباهت یا به عبارتی تابع فاصله‌ی مناسب داده‌ها از جمله مواردی است که قبل از شروع فرآیند خوشه‌بندی باید مورد توجه قرار گیرد. یک تابع فاصله معیاری است که از آن برای تعیین میزان شباهت بین داده‌های موجود در یک مجموعه داده استفاده می‌شود. تابع فاصله‌ی DTW یکی از شناخته شده ترین توابع فاصله می‌باشد. در این تابع فاصله سری‌هایی با ساختار مشابه که در پریودهای زمانی متفاوت بوجود آمده باشند، متشابه قلمداد می‌شوند. بنابراین می‌توان گفت در این روش سری‌های زمانی بر اساس شکل گروه‌بندی می‌شوند [۱۵]. فاصله‌ی DTW بین دو سری زمانی X و Y با طولهای m و k از رابطه‌ی (۷) محاسبه می‌شود:

$$DTW(x,y)=\begin{cases}\infty \text{ if } m=0 \text{ or } k=0\\ dist(x_1,y_1)+\\ \min\{DTW(rest(X),rest(Y)),\\ DTW(rest(X),(Y)),\\ DTW((X),rest(Y))\}\ otherwise\end{cases} \qquad (۷)$$

در فرمول بالا تابع فاصله‌ی *dist* به طور معمول از رابطه‌ی (۸) محاسبه می‌گردد.

$$dist(x_1,y_1)=\left|(x_1-y_1)\right| \qquad (۸)$$

منظور از $Rest(X)$ دنباله‌ی X بدون المان اول است.

$$rest(X)=[(t_2,x_2),...,(t_m,x_m)] \qquad (۹)$$

۲-۴- روش‌های نمایش سری‌های زمانی

روش‌های نمایش سری‌های به سه دسته‌ی روش‌های سازگار با داده، روش‌های ناسازگار با داده و روش‌های آماری دسته‌بندی می‌شوند.

از میان این روش‌های نمایش، روش ضرایب DCT[۱] به عنوان یکی از سریع‌ترین و کاراترین آن‌ها مطرح است که از دیگر مزایای آن می‌توان به مواجه بودن با اعداد حقیقی، سادگی محاسبات و امکان بازسازی داده با دقت بالا در آن اشاره کرد.

تبدیل کسینوسی گسسته (DCT) روشی برای مدل کردن سری‌های زمانی با استفاده از مجموعه‌ای از امواج کسینوسی می‌باشد. برای سری زمانی با طول m ضرایب DCT طبق رابطه‌ی (۱۰) و (۱۱) بدست می‌آیند.

$$y_k=w(k)\sum_{i=0}^{m-1}x_i\cos\left(\frac{\pi(2i+1)k}{2m}\right) \qquad (۱۰)$$
$$k=0,1,\cdots,m-1$$

$$w(k)=\begin{cases}\dfrac{1}{\sqrt{m}} & \text{if } k=0\\ \sqrt{\dfrac{2}{m}} & \text{if } 1\le k\le m-1\end{cases} \qquad (۱۱)$$

سری اولیه با استفاده از تبدیل کسینوسی معکوس (IDCT) طبق رابطه‌ی (۱۲) بازسازی می‌شود.

$$x_i=w(k)\sum_{k=0}^{m-1}y_k\cos\left(\frac{\pi(2i+1)k}{2m}\right) \qquad (۱۲)$$
$$i=0,1,\cdots,m-1$$

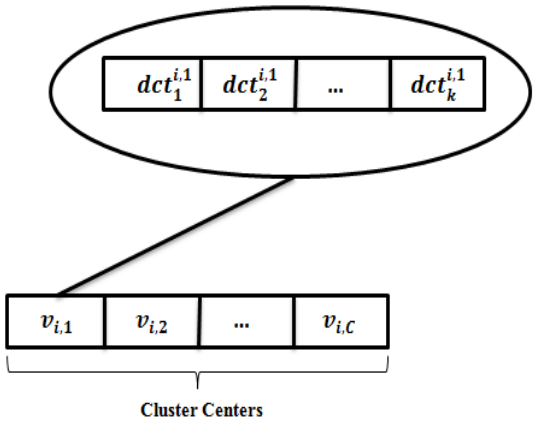
اولین ضرایب DCT یک سری زمانی بیانگر مهم‌ترین ویژگی‌های آن سری هستند و یک سری زمانی تنها با از دست دادن مقدار کمی اطلاعات با استفاده از این ضرایب بازسازی می‌شود.

۲-۵- روش پیشنهادی

در روش پیشنهادی از ترکیب روش خوشه‌بندی FCM با الگوریتم تکاملی دیفرانسیلی برای خوشه‌بندی سری‌های زمانی با در نظر گرفتن تابع فاصله‌ی DTW به عنوان معیار شباهت داده‌ها استفاده گردید. در روش پیشنهادی هر مرکز خوشه با توجه به مجموعه داده، تعداد مشخصی المان دارد که همان ضرایب DCT مربوط به مراکز خوشه‌ها هستند. از آنجا که داده‌های مورد استفاده در این تحقیق سری‌های زمانی با طول بلند هستند، هر ذره به جای یافتن تمامی المان‌های مراکز خوشه‌ها تنها مهمترین ضرایب DCT آن‌ها را می‌یابد و بدین ترتیب فضای جستجو کاهش و کارایی الگوریتم افزایش می‌یابد. اگر مجموعه

[۱] Discrete Cosine Transform

داده شامل N سری زمانی بوده و تعداد k ضریب اول DCT به عنوان مهم‌ترین ضرایب برای بازسازی سری زمانی در نظر گرفته شود و همچنین تعداد خوشه های موجود در مجموعه برابر با C باشد، هر کروموزوم از C×k المان تشکیل می‌شود که این المان‌ها، مهم‌ترین ضرایب DCT مراکز خوشه ها می باشند. شکل (۱) یک کروموزوم را در روش پیشنهادی نشان می‌دهد.



شکل ۱- نمایش یک کروموزوم در روش پیشنهادی

در شکل ۱، i و k و C به ترتیب نشان دهنده‌ی شماره‌ی کروموزوم، تعداد ضرایب DCT مناسب برای بازسازی سری‌های زمانی موجود در مجموعه داده و تعداد خوشه‌های موجود در مجموعه داده می‌باشند. همچنین *v* و *dct* به ترتیب بیانگر مراکز خوشه‌ها و ضرایب DCT مجهول مربوط به مراکز خوشه‌ها می‌باشند. در گام اول کروموزوم‌ها به صورت تصادفی تولید می‌شوند. هر کروموزوم مطابق شکل (۱) شامل مهم‌ترین ضرایب DCT مراکز خوشه‌ها می‌باشد. به عنوان مثال با در نظر گرفتن k ضریب DCT اول مراکز خوشه ها به عنوان مجهولات مسئله، المان اول تا k ام هر کروموزوم نمایانگر ضرایب DCT مربوط به مرکز خوشه‌ی اول می‌باشد. همچنین المان k+1 تا 2×k نماینده‌ی ضرایب DCT دومین مرکز خوشه می‌باشد. در مرحله‌ی بعد با توجه به ضرایب DCT بدست آمده برای مراکز خوشه‌ها، با استفاده از دستور معکوس DCT (IDCT) مراکز بازسازی می‌شوند. سپس با استفاده از تابع فاصله‌ی DTW ماتریس فاصله‌ی میان مرکز خوشه‌ها و داده‌ها محاسبه می‌گردد و بعد از آن درجه عضویت‌ها را با استفاده از U مربوط به الگوریتم FCM بدست می‌آوریم.

در گام بعدی طبق رابطه‌ی (۱) مقدار تابع هدف مربوط به هر کروموزوم محاسبه می‌شود. در مرحله‌ی بعد به ازای هر کروموزوم بردار Trial محاسبه می‌گردد و سپس نوبت به ارزیابی و انتخاب بین کروموزوم اولیه و بردار Trial با توجه به مقدار تابع هدف می‌رسد. در صورت بهتر بودن مقدار تابع هدف به ازای هرکدام از این دو بردار (کروموزوم اولیه و بردار Trial)، بردار مذکور به نسل بعدی منتقل می‌شود و بدین ترتیب مراکز خوشه‌های جدید از کروموزوم انتخابی بدست می‌آیند. از آنجا که عملیات میانگین‌گیری در توابع فاصله‌ای مانند DTW بی معنیست، بنابراین بر خلاف روش خوشه‌بندی FCM که از میانگین-گیری برای محاسبه‌ی مراکز خوشه‌ها استفاده می‌شود، در این روش الگوریتم تکاملی دیفرانسیلی عملیات بدست آوردن مراکز خوشه‌ها را بر عهده دارد. شبه کد روش پیشنهادی در جدول (۱) آورده شده است.

جدول ۱- شبه کد روش پیشنهادی

1: Create and initialize P chromosomes with *C* cluster centers(DCT coeficients of cluster centers)

2: **FOR** iteration count = 1 to maximum iterations **DO**

3: **FOR** each chromosome *i* **DO**

4: Calculate cluster centers with the use of IDCT E.q.(12)..

5: Calculate distance matrix using E.q.(7).

5: Calculate partition matrix (U) using E.q.(2).

6: Calculate the fitness function using E.q.(1).

7: **END**

8: **FOR** each chromosome *i* **DO**

10: Generate a trial vector using E.q.(5).

11: Evaluate trial vector using E.q.(6).

12: Selection

13: **END**

13: **END**

14: Extract the cluster centers corresponding to best chromosome.

۳- پیاده‌سازی و اجرا

در این بخش به معرفی مجموعه داده‌های استفاده شده و تحلیل و بررسی نتایج حاصل از مقایسه‌ی روش‌ها می‌پردازیم.

۳-۱- مجموعه داده‌ها

برای ارزیابی عملکرد روش پیشنهادی، آن را بر روی دو سری از مجموعه داده‌ی شناخته شده از سری‌های زمانی به نام-های CBF و Trace پیاده کردیم [۱۶،۱۷]. خصوصیات این مجموعه داده‌ها شامل طول سری‌های زمانی، تعداد سری‌های زمانی و تعداد خوشه‌های موجود در این مجموعه‌ها به طور مختصر در جدول (۲) آورده شده است.

جدول ۲- خصوصیات مجموعه داده‌های استفاده شده

مجموعه داده	طول سری-های زمانی	اندازه‌ی مجموعه داده	تعداد خوشه‌های موجود
CBF	۱۶۸	۹۳۰	۳
Trace	۲۷۵	۲۰۰	٤

۳-۲-تنظیم پارامترها

مقادیر پارامترهای موجود در روش‌های مورد مقایسه در جدول (۳) آورده شده است. این مقادیر با استفاده از روش سعی و خطا بدست آمدند.

جدول ۳- مقادیر پارامترهای استفاده شده در روش‌ها

پارامتر	FCM	DE	DE پیشنهادی
m	۲	۲	۲
P	-	۵۰	۵۰
Itr	۱۰۰	۲۰۰	۲۰۰
Cr	-	۰،۵	۰،۵
F	-	۱	۱
dct	-	-	۲۰

در جدول (۳)، *m* پارامتر فازی کننده، *P* اندازه‌ی جمعیت، *Itr* تعداد تکرارها، dct، تعداد ضرایب DCT استفاده شده برای بازسازی مراکز خوشه‌ها و *Cr* و *F* مقدار نرخ ترکیب[۱] و فاکتور مقیاس می‌باشند. همچنین لازم به ذکر است که بیشترین و کمترین مقدار موجود در المان‌های سری‌های زمانی موجود در مجموعه داده را به عنوان محدوده‌ای برای کنترل المان‌های موجود در مراکز خوشه‌های مربوط به هر کروموزوم بکار بردیم بدین معنا که این المان‌ها نمی‌توانند از مقدار مرزی تعیین شده برای آنها تجاوز کنند.

[۱] Crossover Rate

با این توضیحات در ادامه به بررسی و مقایسه‌ی عملکرد هرکدام از روش‌ها می‌پردازیم.

۳-۳- معیارهای ارزیابی نتایج

برای ارزیابی و مقایسه‌ی نتایج حاصل از روش های خوشه‌بندی مختلف، معیارهای متفاوتی وجود دارد. در این تحقیق ما از چهار معیار مقدار تابع هدف، دقت یا Precision خوشه‌بندی، مقدار F-measure [۱۸] و زمان اجرا برای مقایسه‌ی دو روش استفاده کردیم. از آنجا که تابع هدف استفاده شده در این روش، تابع بهینگی مربوط به روش خوشه‌بندی FCM می‌باشد، کمتر بودن مقدار تابع بهینگی نشان‌دهنده‌ی فشردگی بیشتر میان اعضای موجود در یک خوشه و پراکندگی میان خوشه‌های متفاوت می-باشد. به عبارت دیگر هرچه مقدار بدست آمده برای تابع بهینگی کمتر باشد، روش خوشه‌بندی عملکرد بهتری داشته است. در مورد زمان اجرا نیز واضح است که سرعت بالاتر یک روش از امتیازات آن روش محسوب می‌شود. یکی دیگر از معیارهای بررسی عملکرد یک روش خوشه-بندی محاسبه‌ی مقدار دقت یا Precision مربوط به آن روش است. برای توضیح مفهوم دقت، ابتدا با مفهوم ماتریس درهم ریختگی (Classification Matrix) آشنا می‌شویم. این ماتریس چگونگی عملکرد الگوریتم خوشه-بندی را با توجه به مجموعه داده و به تفکیک انواع دسته‌های موجود نمایش می‌دهد. جدول (۴) این ماتریس را نشان می‌دهد.

جدول ۴- ماتریس درهم ریختگی

پیش بینی شده / واقعی	Positive	Negative
Positive	True Positive (TP)	True Negative (TN)
Negative	False Positive (FP)	False Negative (FN)

در این ماتریس منظور از TP، تعداد اشیایی است که متعلق به خوشه‌ی i بوده‌اند و توسط الگوریتم به خوشه‌ی i تخصیص داده شده‌اند. منظور از TN، تعداد اشیایی است که متعلق به خوشه‌ی i نبوده‌اند و توسط الگوریتم به خوشه‌ی i تخصیص داده نشده‌اند. منظور از FP، تعداد اشیایی است که متعلق به خوشه‌ی i نبوده‌اند و توسط

الگوریتم به خوشه‌ی i تخصیص داده شده‌اند و منظور از FN، تعداد اشیایی است که متعلق به خوشه‌ی i بوده‌اند و توسط الگوریتم به خوشه‌ی i تخصیص داده نشده‌اند. با توجه به جدول (۴) دقت از رابطه‌ی (۱۳) محاسبه می‌گردد.

$$p = \frac{TP}{TP + FP} \qquad (۱۳)$$

واضح است که هرچه مقدار دقت بیشتر باشد، الگوریتم عملکرد بهتری داشته است. علاوه بر این با توجه به جدول (۴) معیار دیگری به نام فراخوانی تعریف می‌شود که از رابطه‌ی (۱۴) محاسبه می‌گردد.

$$R = \frac{TP}{TP + FN} \qquad (۱۴)$$

مقدار F-measure از میانگین هارمونیک دقت و فراخوانی بدست می‌آید.

$$\frac{1}{F} = \frac{1}{2}(\frac{1}{R} + \frac{1}{P}) \qquad (۱۵)$$

با دقیق شدن در رابطه‌ی (۱۵) نیز می‌توان دریافت که بیشتر بودن میزان F-measure نشان‌دهنده‌ی عملکرد بهتر الگوریتم می‌باشد.

۳-۴- نتایج حاصل و تفسیر آن

برای ارزیابی چگونگی عملکرد روش پیشنهادی، در جدول (۵) نتایج حاصل از پیاده‌سازی روش پیشنهادی، الگوریتم خوشه‌بندی FCM و روش خوشه‌بندی مبتنی بر الگوریتم تکاملی دیفرانسیلی و بدون استفاده از روش کاهش بعد داده بر روی دو مجموعه داده، نشان داده شده است.

جدول ۵- نتایج حاصل از پیاده‌سازی دو روش بر روی دو مجموعه داده

مجموعه داده	روش	Fitness	Pr	Fm	Time(sec)
CBF	Proposed DE	۹۵۷۳,۹۶۸	۰,۷۵	۰,۶۲	۵۷۲۵
	FCM	۱۴۲۵۸,۴۵۹۰	۰,۷۳	۰,۶۱	۴۵
	DE	۱۰۱۲۸,۱۳۱	۰,۷۳	۰,۶۲	۶۲۳۵
Trace	Proposed DE	۱۹۷,۴۹۶	۰,۵۴	۰,۶۳	۷۹۴۰
	FCM	۱۷۲۶,۸۸۶۴	۰,۵۰	۰,۵۳	۶۹,۲۸۰
	DE	۲۰۵,۷۶۹	۰,۵۲	۰,۵۸	۸۷۵۶

همانگونه که از جدول (۵) مشهود است، مقدار تابع هدف (Fitness) حاصل از روش ترکیبی پیشنهادی در هر دو مجموعه داده کمتر (بهتر) از دو روش FCM و DE بوده است. همچنین با مقایسه‌ی مقدار دقت (Pr) و F-measure (Fm) حاصل از سه روش مورد مقایسه، می‌توان به برتری روش پیشنهادی نسبت به دو روش دیگر پی برد. در مورد زمان اجرا، الگوریتم‌های تکاملی معمولا به زمان زیادی برای اجرا نیازمندند و این نکته با مقایسه‌ی زمان اجرای دو روش تکاملی دیفرانسیلی و FCM به خوبی مشهود است. اما از آنجا که در روش پیشنهادی از ضرایب DCT برای کاهش مجهولات مسئله بهره بردیم و الگوریتم DE تنها مهم‌ترین ضرایب DCT هر مرکز خوشه را می‌یابد، الگوریتم تکاملی دیفرانسیلی پیشنهادی نسبت به الگوریتم تکاملی دیفرانسیلی که سعی در یافتن تمامی المان‌های مراکز خوشه‌ها دارد، سریع‌تر عمل می‌نماید.

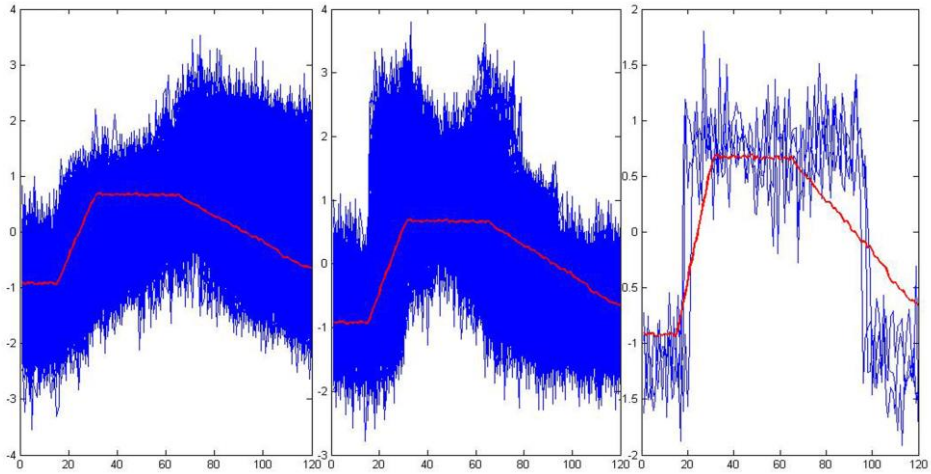
روش خوشه‌بندی FCM در مواجه با داده‌های پیچیده همچون سری‌های زمانی با طول زیاد عملکرد نسبتا ضعیفی دارد. در این روش خوشه‌بندی معمولا احتمال به دام افتادن در بهینه‌ی محلی با مواجه شدن با داده‌های حجیم و پیچیده افزایش می‌یابد. علاوه بر این روش خوشه بندی FCM زمانی که معیار شباهت در نظر گرفته برای داده‌ها تابع فاصله‌ای به جز تابع اقلیدسی باشد، کارایی خوبی ندارد. به عنوان مثال مفهوم میانگین‌گیری برای تابع فاصله‌ای مانند DTW تعریف نشده و بی معنی می‌باشد در حالی روش FCM از مفهوم میانگین‌گیری برای بدست آوردن مراکز خوشه‌ها استفاده می‌کند. بنابراین روش FCM در برخورد با داده‌هایی مانند سری‌های زمانی و در صورت استفاده از تابع فاصله‌ای مانند DTW به عنوان معیار شباهت، ضعیف عمل می‌کند.

در روش خوشه‌بندی DE، مراکز خوشه‌ها که مجهولات مسئله می‌باشند هم طول با داده‌های موجود در مجموعه داده در نظر گرفته شدند و از الگوریتم‌های تکاملی DE

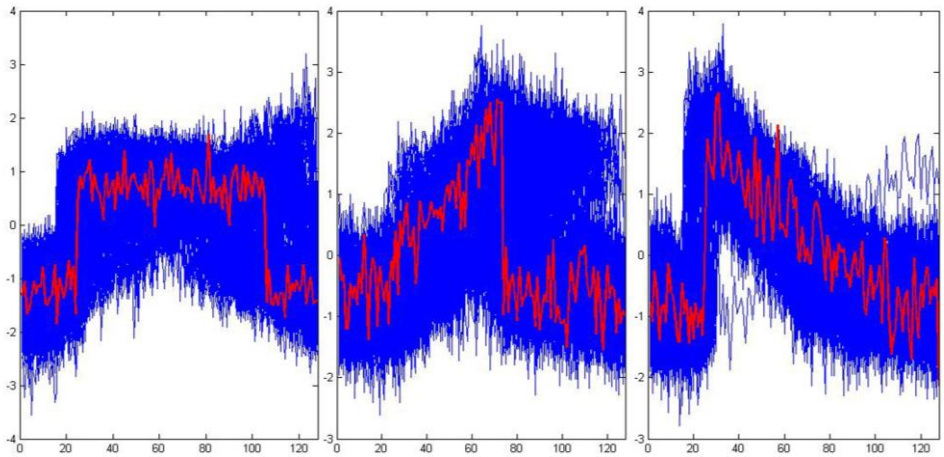
برای یافتن المان‌های تمامی مراکز استفاده گردید. تعداد زیاد مجهولات باعث کاهش کارایی الگوریتم مورد استفاده شده و نتایج مورد انتظار را بدست نخواهد داد، بنابراین این روش‌ها تنها برای پیاده‌سازی بر روی مجموعه داده‌هایی با طول کم مناسب به نظر می‌رسند.

در روش خوشه‌بندی تکاملی دیفرانسیلی پیشنهادی تنها اولین و مهم‌ترین ضرایب DCT خوشه‌ها به عنوان مجهولات مسئله شناخته شده‌اند و پس از مشخص شدن این مقادیر، با استفاده از معکوس DCT مراکز خوشه‌ها بازسازی می‌گردند. بنابراین مجهولات مسئله کاهش چشمگیری پیدا کرده و این امر علاوه بر کاهش زمان محاسبات، باعث افزایش دقت و دستیابی به جواب‌های بهتر نیز می‌گردد. بنابراین روش پیشنهادی با تعداد مجهولات کمتر قابلیت پیاده‌سازی بر روی مجموعه داده‌هایی شامل سری‌های زمانی طویل را دارا بوده که این امر یکی از بزرگترین مزیت‌های این روش به شمار می‌رود.

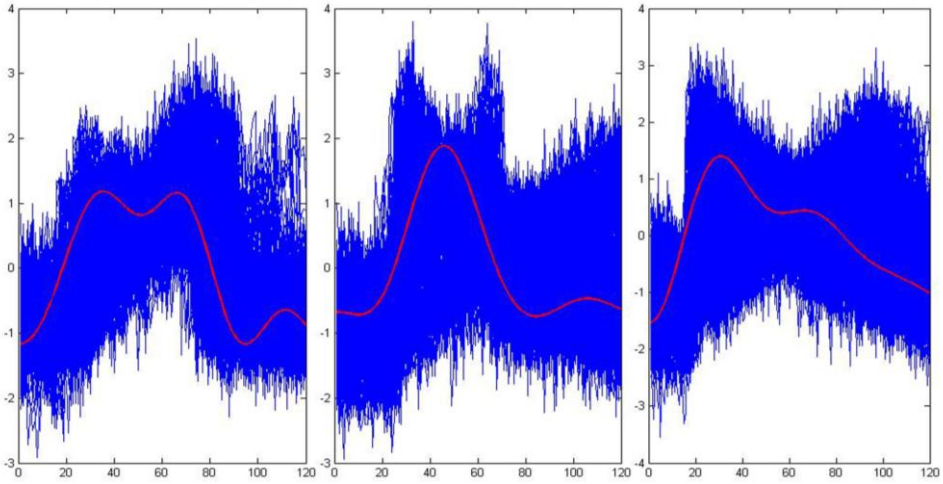
در شکل‌های ۲ و ۳ و ۴ به ترتیب خوشه‌ها و مراکز بدست آمده از روش FCM، روش مبتنی بر الگوریتم تکاملی دیفرانسیلی بدون استفاده از روش‌های کاهش بعد سری زمانی و روش پیشنهادی، نشان داده شده است. در شکل ۲ به خوبی استفاده از مفهوم میانگین‌گیری در الگوریتم FCM مشهود است در حالی که همانگونه که قبلا به آن اشاره شد، زمانی که معیار شباهت داده‌ها تابع فاصله‌ای مانند DTW باشد، استفاده از مفهوم میانگین-گیری درست نمی‌باشد. همچنین با مقایسه‌ی شکل ۳ و ۴ می‌توان به اهمیت استفاده از روش کاهش بعد تبدیل سینوسی گسسته در روش پیشنهادی پی برد. با در نظر گرفتن مهم‌ترین ضرایب DCT خوشه‌ها به عنوان مجهولات مسئله و به عبارتی با کاهش مجهولات مسئله این امکان را برای الگوریتم ایجاد می‌شود که با جستجوی بهتر در فضای مسئله به دنبال جوابی بهینه باشد.



شکل ۲- خوشه‌ها و مراکز خوشه‌های بدست آمده از روش خوشه‌بندی FCM



شکل ۳- خوشه‌ها و مراکز خوشه‌های بدست آمده از روش تکاملی دیفرانسیلی بدون استفاده از روش کاهش بعد سری زمانی



شکل ۴- خوشه‌ها و مراکز خوشه‌های بدست آمده از روش تکاملی دیفرانسیلی پیشنهادی

۶- نتیجه‌گیری

در این مقاله یک روش خوشه‌بندی از ترکیب الگوریتم تکاملی دیفرانسیلی و روش خوشه بندی FCM و با استفاده از روش کاهش بعد DCT، برای گروه‌بندی سری‌های زمانی پیشنهاد داده شد. با در نظر گرفتن تابع فاصله‌ی DTW به عنوان معیار شباهت داده‌ها و همچنین تابع بهینگی الگوریتم FCM به عنوان تابع هدف، روش پیشنهادی بر روی دو مجموعه داده از سری‌های زمانی پیاده سازی شده و با الگوریتم خوشه‌بندی FCM و روش خوشه‌بندی مبتنی بر الگوریتم تکاملی دیفرانسیلی بدون استفاده از روش کاهش بعد داده، مقایسه گردید. از نظر زمان اجرا روش خوشه‌بندی پیشنهاد داده شده کندتر از الگوریتم FCM و سریع‌تر از روش دیگر بوده اما از نظر دقت و کارایی، روش پیشنهادی بهتر از سایرین عمل می‌کند. به دلیل توانایی الگوریتم‌های تکاملی در یافتن جواب‌هایی نزدیک به بهینه در مسائلی با فضای جستجوی گسترده، و از طرفی کاهش کارایی روش خوشه‌بندی FCM در مواجه با مجموعه داده‌های حجیم شامل داده‌هایی با بعد زیاد، ترکیب این روش‌ها در حل مسئله‌ی خوشه‌بندی سری‌های زمانی روش خوبی برای بهبود عملکرد الگوریتم FCM می‌باشد. در این تحقیق نشان داده شد که الگوریتم تکاملی دیفرانسیلی که امروزه به عنوان یکی از قوی‌ترین الگوریتم های تکاملی مطرح است، روشی مناسب برای جبران نقاط ضعف الگوریتم خوشه‌بندی FCM می‌باشد.

مراجع

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

[2] Abraham, A., Das, S., Roy, S. (2007). “Swarm Intelligence Algorithms for Data Clustering,” Soft Computing for Knowledge Discovery and Data Mining, O. Maimon and L. Rokach (Eds.), Springer Verlag, Germany, pp. 279-313.

[3] Pedrycz, W., Gomide, F. (2007). “Fuzzy Systems Engineering: Toward Human-Centric Computing,” Wiley-Interscience.

[4] Winkler, R., Klawonn, F., Kruse, R. (2012). “Problems of Fuzzy c-Means Clustering and Similar Algorithms with High Dimensional Data Sets”, Challenges at the Interface of Data Analysis, Computer Science, and Optimization Studies in Classification, Data Analysis, and Knowledge Organization, pp. 79-87.

[5] Paterlinia, S., Krink, T. (2006). “Differential evolution and particle swarm optimisation in partitional clustering,”Comput. Stat. Data Anal., vol. 50, no. 5, pp. 1220–1247.

[6] Omran, M., Engelbrecht, A. P., Salman, A. (2005). “Differential evolution methods for unsupervised image classification,” inProc. 7th CEC, pp. 966–973.

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

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

[9] Fuad, M. (2014). “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.

[10] Vlachos, M., Lin, J., Keogh, E., Gunopulos, D. (2003). “A waveletbased anytime algorithm for k-means clustering of time series,” Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, May 1–3.

[11] Powell, N., Foo, S.Y., Weatherspoon, M. (2008). “Supervised and Unsupervised Methods for Stock Trend Forecasting,” System Theory, 2008. SSST 2008. 40th Southeastern Symposium on , vol., no., pp. 203-205.

[12] Guo, C., Jia, H., Zhang, N. (2008). “Time Series Clustering Based on ICA for Stock Data Analysis,” Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on , vol., no., pp.1-4, 12-14.

[13] Yu, F., Oyana, D., Hou, W.C., Wainer, M. (2010). “Approximate clustering on data streams using discrete cosine transform,” J. Inform. Process. Syst.6(1), pp. 67–78.

[14] Sepehr, R., Moradi, M.H., Mashayekhi, GH., Kardar, L., (2007).“ Review and compare different methods of partitional fuzzy clustering based on standard FCM,” First joint Congress on Fuzzy and Intelligent Systems, mashhad, Iran.

[15] Berndt, D., Clifford, J. (1994).“Using Dynamic Time Warping to Find Patterns in Time Series,” Proc. AAAI-94 Workshop Knowledge Discovery in Databases, pp. 359-370.

[16] Keogh, E., Smyth, P. (1997). “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.

[17] Zhang, H., Ho, T.B., Zhang, Y., Lin, M.S. (2006). “Unsupervised Feature Extraction for Time Series Clustering Using Orthogonal Wavelet Transform,” Informatica, pp. 305-319.

[18] Izakian, Z., Mesgari, M. S. (2015). “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.