ارائه‌ی یک روش خوشه‌بندی توافقی سری‌های زمانی بر اساس روش Fuzzy C-Means و الگوریتم انبوه ذرات

زاهده ایزکیان*[۱]، محمدسعدی مسگری[۲]

[۱] دانشجوی دکتری سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه‌نصیرالدین‌طوسی
zizakian@mail.kntu.ac.ir

[۲] استادیار دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه‌نصیرالدین‌طوسی
mesgari@kntu.ac.ir

(تاریخ دریافت اردیبهشت ۱۳۹۹، تاریخ تصویب تیر ۱۳۹۹)

چکیده

در سال‌های اخیر با پیشرفت فناوری‌های جمع آوری اطلاعات و فراهم شدن حجم عظیمی از داده‌های پیچیده همچون سری‌های زمانی نیاز به روش‌هایی مناسب به منظور تجزیه و تحلیل این نوع داده بیش از پیش احساس می‌شود. از میان روش‌های مختلف داده‌کاوی موجود تکنیک خوشه‌بندی داده‌ها با هدف ساده سازی مجموعه داده‌های بزرگ و استخراج اطلاعات مفید توجه بسیاری از محققین علوم کامپیوتر را به خود جلب کرده است. مسئله‌ی انتخاب تابع فاصله یکی از مهم‌ترین چالش‌هایی است که پیش از آغاز فرآیند خوشه‌بندی سری‌های زمانی مورد توجه قرار می‌گیرد. انتخاب تابع فاصله‌ی مناسب یک مجموعه داده به شناخت ماهیت داده پیش از انجام عملیات خوشه‌بندی وابسته می‌باشد و از این رو امری پیچیده و زمانبر می‌باشد. از سویی دیگر تاکنون توابع فاصله‌ی مختلفی با ویژگی‌ها و نقاط قوت متفاوت به منظور اندازه‌گیری میزان تفاوت/شباهت میان سری‌های زمانی پیشنهاد داده شده است. چگونگی ارائه‌ی یک روش خوشه-بندی با قابلیت بهره جستن از ویژگی‌های توابع فاصله‌ی مختلف به طور همزمان و بدون نیاز به شناخت ماهیت داده‌ها پیش از آغاز فرآیند خوشه‌بندی، چالش اصلی این تحقیق می‌باشد. به منظور حل این مسئله در این تحقیق یک روش خوشه‌بندی با ترکیب روش خوشه‌بندی Fuzzy C-Means (FCM) و الگوریتم شناخته شده‌ی مبتنی بر هوش جمعی انبوه ذرات (PSO) با هدف استفاده از توابع فاصله‌ی مختلف با وزن‌های متفاوت در حین فرآیند خوشه‌بندی پیشنهاد داده شد. انتخاب تابع هدف در این مطالعه به گونه‌ای بوده است که نتیجه‌ی حاصل از خوشه‌بندی بیشترین اشتراک را با نتایج خوشه‌بندی حاصل از توابع فاصله‌ی مختلف داشته باشد. به عبارت دیگر روش خوشه‌بندی ارائه شده در این تحقیق یک روش خوشه‌بندی توافقی می‌باشد که نتیجه حاصل توافق میان توابع فاصله‌ی مختلف می‌باشد. روش پیشنهادی ارائه شده در این تحقیق با در نظر گرفتن سه تابع فاصله‌ی مختلف بر روی هفت سری مجموعه داده‌ی شناخته شده از سری‌های زمانی پیاده‌سازی شد و با پنج روش دیگر مقایسه گردید نتایج حاصل از این مقایسه نشان داد روش ارائه شده در این تحقیق در بیشتر از ۸۵ درصد موارد بهتر از سایر روش‌ها عمل کرده است.

واژگان کلیدی: داده‌کاوی، خوشه‌بندی، تابع فاصله، سری زمانی، اطلاعات مشترک نرمال شده، الگوریتم انبوه ذرات

۱- مقدمه

مقادیر ثبت شده از یک رویداد در فواصل زمانی برابر را می‌توان به صورت دنباله‌ای از اعداد نمایش داد که این دنباله سری زمانی نامیده می‌شود. امروزه تجزیه و تحلیل این نوع داده‌ها در علوم مختلف کاربرد زیادی دارد. تجزیه و تحلیل سری‌های زمانی در هواشناسی اطلاعات مفیدی را از چگونگی تغییرات آب و هوایی در اختیار متخصصان هواشناسی قرار داده و موجب برآورد یک مدل مناسب به منظور پیش‌بینی‌های هواشناسی می‌گردد [۱ و ۲ و ۳ و ۴]. در اقتصاد تجزیه و تحلیل این نوع داده در بررسی تغییرات قیمت کالاها، محاسبه‌ی میزان تورم، پیش بینی قیمت سهام و غیره مورد استفاده قرار می‌گیرد [۵ و ۶ و ۷ و ۸ و ۹]. داده های لرزه نگاری انواع دیگری از سری‌های زمانی می‌باشند که متخصصین حوزه‌ی زلزله‌شناسی با تحلیل این داده‌ها به بررسی چگونگی رفتار لایه‌های داخلی زمین می‌پردازند [۱۰ و ۱۱]. جمعیت شناسان با تحلیل و بررسی تغییرات جمعیت در سال‌های متوالی به اطلاعات ارزشمندی در زمینه‌ی ساختار مهاجرتی انسان‌ها دست می‌یابند[۱۲ و ۱۳]. علاوه بر این تحلیل این نوع داده‌ها در علوم دریایی، پزشکی و مهندسی و غیره نیز کاربردهای فراوان دارد [۱۴ و ۱۵ و ۱۶ و ۱۷ و ۱۸].

امروزه پیشرفت فناوری جمع آوری اطلاعات مانند موبایل و سیستم تعیین موقعیت جهانی میزان زیادی از داده‌های مکانی-زمانی مانند سری‌های زمانی را فراهم کرده است، از این رو تمایل به استفاده از روش‌های داده-کاوی به منظور استخراج الگوهای مفید بیش از گذشته احساس می‌شود.

تکنیک خوشه‌بندی سری‌های زمانی از مهم‌ترین تکنیک‌های داده‌کاوی است که با هدف کشف و گروه‌بندی داده‌هایی با ساختار و روند متشابه استفاده می‌شود [۱۹]. یک روش خوشه‌بندی کارآمد به گونه‌ای عمل می‌کند که داده‌های موجود در یک گروه بیشترین شباهت را نسبت به هم و بیشترین اختلاف را با داده‌های موجود در گروه‌های دیگر داشته باشند [۲۰]. به طور کلی هدف اصلی خوشه-بندی ساده سازی و خلاصه کردن یک مجموعه داده‌ی بزرگ به کمک خوشه ها و مراکز آنها می‌باشد و این تکنیک نقش مهمی را در کشف اطلاعات ارزشمند ایفا می‌کند [۱۹ و ۲۰]. از مهم‌ترین مسائل موجود در خوشه-بندی سری‌های زمانی اندازه‌گیری میزان شباهت میان داده‌ها یا به عبارت دیگر انتخاب تابع فاصله‌ی مناسب می-باشد. تاکنون توابع فاصله‌ی مختلفی به منظور ارزیابی میزان شباهت میان سری‌های زمانی پیشنهاد داده شده است. به طور کلی توابع فاصله مورد استفاده در سری‌های زمانی به سه گروه فواصل Lp-norm، الاستیک[۱] و آماری[۲] تقسیم‌بندی می‌شوند. فواصل منهتن، اقلیدسی و چبیشف از مهم‌ترین تکنیک‌های اندازه‌گیری میزان فاصله میان سری‌های زمانی می‌باشند که در گروه فواصل Lp-norm قرار می‌گیرند [۲۱]. در تحقیقات متعددی از این توابع فاصله به منظور بررسی شباهت میان داده‌ها استفاده شده است. در سال ۲۰۱۰ شریف و همکاران [۲۲] و شاعری و همکاران [۲۳] به مطالعه‌ی توابع فاصله‌ی قابل استفاده برای خطوط سیر همچون فاصله‌ی اقلیدسی پرداخته و نقاط مثبت و منفی هریک را مورد بحث و بررسی قرار دادند. در سال ۲۰۰۳، Vlachos و همکاران ابتدا سری‌های زمانی را با استفاده از تبدیل موجک به فضای جدیدی انتقال داده و سپس با استفاده از الگوریتم K-Means و تابع فاصله‌ی اقلیدسی به خوشه‌بندی داده‌ها پرداختند [۲۴]. D'Urso و همکاران در سال ۲۰۰۹ به منظور خوشه‌بندی سری‌های زمانی ابتدا داده‌ها را با استفاده از اوتوکورولیشن[۳] به فضای جدیدی منتقل کرده و سپس فواصل میان سری‌های زمانی انتقال یافته را با استفاده از تابع فاصله‌ی اقلیدسی محاسبه نمودند [۲۵]. Nanda و همکاران در سال ۲۰۱۰ از تابع فاصله‌ی اقلیدسی به منظور اندازه‌گیری فواصل میان سری‌های زمانی بازار سهام استفاده کردند. در این تحقیق روش‌های Fuzzy C-Means، K-Means و SOM[۴] به منظور خوشه‌بندی این داده‌ها مورد استفاده قرار گرفتند و نتایج حاصل برتری روش K-Means را نشان دادند [۲۶]. Yuan و همکاران در سال ۲۰۱۶ با توجه به ویژگی‌هایی همچون پیچیدگی زمانی، قابلیت حذف نویز، تعداد پارامترهای مورد نیاز و کاربرد به بررسی و مقایسه‌ی مهم‌ترین معیارهای شباهت مورد استفاده در سری‌های زمانی و خطوط سیر پرداختند [۲۷]. معیدی و همکاران در سال ۲۰۱۹ در تحقیقی به

۱ Elastic
۲ Statistical
۳ Autocorrelation
۴ Self-organization Map

بررسی عملکرد توابع فاصله‌ی مختلف در فرآیند خوشه-بندی مبتنی بر چگالی پرداختند. آنها اذعان داشتند که مهم‌ترین عامل در تعیین تابع فاصله نوع داده و ویژگی‌های آن می‌باشد. همچنین نتایج حاصل از تحقیق آنها نشان داد تابع فاصله‌ی اقلیدسی در روش‌های خوشه‌بندی مبتنی بر چگالی بهتر از سایر توابع فاصله عمل می‌کند [۲۸].

در توابع فاصله‌ی الاستیک به منظور بهینه کردن فاصله بین دو سری زمانی الزاما نقاط متناظر دو سری با هم مقایسه نمی‌شوند و هر نقطه از سری زمانی اول می-تواند با هر نقطه اختیاری از سری دوم مقایسه شود [۲۹]. توابع فاصله‌ی انحراف زمانی پویا[۱] [۳۰]، طولانی‌ترین دنباله‌ی مشترک[۲] [۲۴] و فاصله‌ی ویرایش[۳][۲۱] از جمله توابع فاصله‌ای هستند که در گروه توابع الاستیک قرار می-گیرند. این توابع فاصله نیز تاکنون توجه محققین بسیاری را در حوزه‌ی داده‌کاوی سری‌های زمانی به خود جلب کرده‌اند. Keogh و همکاران در سال ۱۹۹۹ یک روش خوشه‌بندی سلسله‌مراتبی مناسب برای سری‌های زمانی پیشنهاد کردند و از تابع فاصله‌ی انحراف زمانی پویا به منظور اندازه‌گیری فواصل میان داده‌ها بهره جستند [۳۱]. Sakurai و همکاران در سال ۲۰۰۵ نسخه‌ی جستجوی سریع تابع فاصله‌ی انحراف زمانی پویا را معرفی کردند. سرعت بالای نسخه‌ی پیشنهادی نسبت به فاصله‌ی انحراف زمانی پویا از مهم‌ترین مزیت‌های آن می‌باشد [۳۲]. Petitjean و همکاران در سال ۲۰۱۱ نسخه‌ی بهبود یافته-ای از تابع فاصله‌ی انحراف زمانی پویا با توانایی میانگین-گیری پیشنهاد کردند و با استفاده از الگوریتم K-Means به بررسی عملکرد تابع فاصله‌ی پیشنهادی خود پرداختند [۳۳]. Jeong و همکاران در سال ۲۰۱۱ به معرفی نسخه-ای دیگر از تابع فاصله‌ی انحراف زمانی پویا با عنوان فاصله‌ی انحراف زمانی پویای وزن‌دار پرداختند. نتایج حاصل از خوشه‌بندی سری‌های زمانی با استفاده از معیار شباهت پیشنهادی آنها توانایی این معیار را در خوشه‌بندی داده‌ها به خوبی نشان داد [۳۴]. Bankó و همکاران در سال ۲۰۱۲ با ترکیب تابع فاصله انحراف زمانی پویا با معیار اندازه‌گیری PCA[۴] اقدام به حفظ اطلاعات همبستگی موجود در داده‌ها ضمن محاسبه‌ی فاصله‌ی انحراف زمانی پویا میان دو سری زمانی نمودند [۳۵].

سومین گروه از توابع فاصله، توابع فاصله‌ی آماری می-باشند که فاصله‌ی همبستگی پیرسون[۵] [۳۶] مهم‌ترین تابع موجود در این گروه می‌باشد. صبح بیداری و همکاران در سال ۲۰۰۸ با استفاده از الگوریتم‌های خوشه‌بندی Fuzzy C-Means و K-Means و معیار شباهت همبستگی پیرسون اقدام به کشف الگوهای مشابه در مجموعه‌هایی شامل داده‌های رشته‌ای نمودند [۳۷]. Wang و همکاران در سال ۲۰۱۳ به طور مفصل به بررسی نقاط ضعف و قدرت تعدادی از مهم‌ترین توابع فاصله‌ی مورد استفاده در سیگنال‌ها و خطوط سیر پرداخته و تابع فاصله‌ی انحراف زمانی پویا را به عنوان قدرتمندترین تابع فاصله در توابع فاصله‌ی مورد بررسی تشخیص دادند [۳۸]. ایزکیان و همکاران در سال ۲۰۱۵ یک روش خوشه‌بندی مبتنی بر الگوریتم انبوه ذرات پیشنهاد کردند. در روش پیشنهادی آنها از تابع فاصله‌ی همبستگی پیرسون به عنوان معیار شباهت میان سری‌های زمانی استفاده شد. نتایج حاصل توانایی این تابع فاصله را در اندازه‌گیری میزان شباهت میان داده‌ها نشان داد [۳۹]. Kim و همکاران در سال ۲۰۱۹ اذعان داشتند که تابع فاصله‌ی مورد استفاده در فرآیند خوشه‌بندی داده‌های رشته‌ای مربوط به ژن‌های تک سلولی تاثیری مستقیم بر دقت نتایج حاصل دارد. آنها دریافتند که فاصله‌ی همبستگی پیرسون در مقایسه با فاصله‌ی اقلیدسی عملکرد بهتری در خوشه‌بندی این نوع داده‌ها دارند[۴۰]. Shinde و همکاران در سال ۲۰۱۹ به بررسی و مرور روش‌های خوشه‌بندی افزایشی موجود پرداختند. بر اساس نتایج حاصل از این تحقیق روش‌های خوشه‌بندی افزایشی که از توابع فاصله‌ی احتمالاتی به عنوان معیار اندازه‌گیری شباهت داده‌ها استفاده می‌کنند عملکرد ضعیفتری نسبت به روش‌هایی دارند که در آنها از توابع فاصله‌ی غیر احتمالاتی همچون فاصله‌ی همبستگی پیرسون استفاده شده است [۴۱].

با توجه به تحقیقات بررسی شده در بسیاری از مطالعات صورت گرفته توسط محققین شکل و ماهیت داده‌ها از مهمترین دلایل انتخاب یک تابع فاصله‌ی مشخص می‌باشد.

۱ Dynamic Time Warping Distance
۲ Longest Common Subsequence
۳ Edit Distance
۴ Principal Component Analysis

۵ Pearson Correlation Coefficient

از آنجا که بررسی شکل و ماهیت داده‌های موجود در یک مجموعه فرآیندی پیچیده و زمانبر است، انتخاب معیار شباهت مناسب همچنان به عنوان یکی از مهم‌ترین چالش-ها در مبحث خوشه‌بندی سری‌های زمانی مطرح می‌باشد. از سویی دیگر هریک از توابع فاصله معرفی شده برای سری-های زمانی دارای نقاط قوت منحصر به فرد خود هستند و محققان با تکیه بر ویژگی‌های مثبت هریک اقدام به انتخاب آنها در فرآیند خوشه‌بندی می‌نمایند. استفاده‌ی همزمان از توابع فاصله‌ی مختلف در خوشه‌بندی سری‌های زمانی با تمرکز بر دستیابی به نتیجه‌ی مورد توافق تمامی فواصل شرکت کننده، موضوع اصلی این مطالعه می‌باشد که تاکنون در تحقیقات انجام شده در زمینه‌ی خوشه‌بندی سری‌های زمانی مورد توجه قرار نگرفته است. به بیانی دیگر هدف از این تحقیق خوشه‌بندی سری‌های زمانی می‌باشد به گونه‌ای که خوشه‌های حاصل بیشترین سطح توافق را با خوشه‌های بدست آمده از هریک از توابع فاصله‌ی شرکت کننده در فرآیند خوشه‌بندی داشته باشند.

در روش ارائه شده در این تحقیق با استفاده از الگوریتم خوشه‌بندی FCM و با تعریف یک تابع هدف مناسب بر مبنای تابع هدف روش FCM به وزن‌دهی توابع فاصله‌ی مختلف شرکت کننده در فرآیند خوشه‌بندی شامل توابع فاصله‌ی اقلیدسی، انحراف زمانی پویا و ضرایب همبستگی پیرسون پرداخته شد. از آنجا الگوریتم‌های بهینه‌یابی روش‌هایی توانمند در یافتن مقادیر نزدیک به بهینه در مسائل پیچیده می‌باشند [۳۹] در این روش از الگوریتم انبوه ذرات به عنوان یکی از الگوریتم‌های کارآمد در یافتن جواب بهینه در مسائل پیوسته، استفاده شد.

در بخش ۲ این تحقیق سه تابع فاصله‌ی مهم مورد استفاده برای سری‌های زمانی معرفی شده و نقاط ضعف و قوت هریک بررسی شد. در بخش ۳ الگوریتم خوشه‌بندی FCM به عنوان اساس روش خوشه‌بندی پیشنهادی معرفی شده و سپس توضیحات مربوط به روش پیشنهادی ارائه گردید. در بخش ۴ عملکرد روش پیشنهادی مورد بررسی قرار گرفت و در بخش ۵ به نتیجه‌ی حاصل از این مطالعه پرداخته شد.

۲- توابع فاصله

در این بخش به معرفی توابع فاصله‌ی Lp-norm، انحراف زمانی پویا و ضرایب همبستگی پیرسون می‌پردازیم.

۲-۱- فاصله‌ی Lp-norm

برای دو سری زمانی X و Y با طول یکسان K، فاصله‌ی L_p -norm از رابطه‌ی (۱) محاسبه می‌شود:

$$L_p - norm(X,Y) = \sqrt[p]{\sum_{i=1}^{k}(x_i - y_i)^p} \qquad (۱)$$

به ازای p=1، p=2 و p=∞ به ترتیب فاصله‌ی منهتن، فاصله‌ی اقلیدسی و ماکزیمم فاصله نامیده می‌شود.

$$L_1 = \sum_{i=1}^{k} |x_i - y_i| \qquad (۲)$$

$$L_2 = \sqrt[2]{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (۳)$$

$$L_\infty = max\{|x_i, y_i|, i = 1,2, \dots, k\} \qquad (۴)$$

فاصله‌ی اقلیدسی به عنوان اولین تابع فاصله‌ای است که به عنوان معیار اندازه‌گیری شباهت بین سری‌های زمانی مطرح شد. از مزایای این روش می‌توان به سادگی محاسباتش اشاره کرد. اما از بزرگترین معایب این تابع فاصله این است که برای بدست آوردن فاصله اقلیدسی، دو سری زمانی باید طول یکسانی داشته باشند. لازم به ذکر است که این تابع فاصله از شیفت زمانی محلی[1] حمایت نمی‌کند [۸].

۲-۲- فاصله‌ی انحراف زمانی پویا

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 کمینه شود. در این تابع فاصله سری‌های زمانی با ساختار مشابه که در پریودهای زمانی متفاوت بوجود آمده باشند، متشابه قلمداد می‌شوند. بنابراین می‌توان گفت در این روش داده-ها بر اساس شکل گروه‌بندی می‌شوند [۱۲]. فاصله‌ی پیچش زمانی پویا (DTW) بین دو سری زمانی X و Y با طولهای m و k از رابطه‌ی زیر محاسبه می‌شود: (۵)

$$DTW(X,Y) = \begin{cases} \infty & if\ m = 0\ or\ k = 0 \\ dist(x_1, y_1) + \min\{DTW(Rest(X), Rest(Y)), \\ DTW(Rest(X), (Y)), DTW((X), Rest(Y))\} & o.w \end{cases}$$

[1] Local Time Shifting

در رابطه‌ی (۵) تابع فاصله‌ی dist میتواند یکی از L_p– norm ها باشد. در اینجا از L_1 –norm استفاده شده است.

$$dist(x_1, y_1) = |(x_1 - y_1)| \quad (۶)$$

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

$$Rest(X) = [(t_2, x_2), \dots, (t_m, x_m)] \quad (۷)$$

از مزایای این تابع فاصله می‌توان به قابلیت آن در اندازه‌گیری فاصله میان دو سری زمانی با طول‌های متفاوت و حمایت از شیفت زمانی محلی و از معایب این روش می‌توان به زمان‌گیر بودن آن اشاره نمود. زمان محاسبات در این روش به صورت $O(m*k)$، از درجه‌ی دوم و تابعی از طول سری‌ها می‌باشد. بنابراین زمانی که حجم پایگاه داده بالا باشد، به زمان زیادی برای بدست آوردن فاصله انحراف زمانی پویا بین داده‌ها نیاز است[۲].

۲-۲- فاصله‌ی ضرایب همبستگی پیرسون

روشی است برای تخمین اینکه دو سری زمانی با چه درجه ای به هم همبستگی دارند. برای دو سری زمانی X و Y با میانگین $\bar{x}$ و $\bar{y}$ و طول m ضریب همبستگی می‌تواند به صورت زیر تعریف شود:

$$\rho_{X,Y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2}} \quad (۸)$$

در رابطه‌ی بالا $\rho_{x,y}$ مقداری است در بازه‌ی [۱ ,۱-]. در صورتی که $\rho_{x,y}$ =۱ به این معنی است که دو سری زمانی همبستگی کامل مثبتی دارند. $\rho_{x,y}$ =۱- نشانگر همبستگی کامل منفی دو سری زمانی است و $\rho_{x,y}$ =۰ یعنی دو سری زمانی هیچ همبستگی با هم ندارند [۱۸].

در جدول (۱) توابع فاصله‌ی اقلیدسی، انحراف زمانی پویا و همبستگی پیرسون با هم مقایسه شدند.

جدول ۱- مقایسه‌ی توابع فاصله‌ی L_2–norm، انحراف زمانی پویا و همبستگی پیرسون

تابع فاصله	طولهای متفاوت	شیفت زمانی	زمان محاسبات	متریک
L_2 –norm	خیر	خیر	O(N)	بله
انحراف زمانی پویا	بله	بله	$O(N^2)$	خیر
همبستگی پیرسون	خیر	خیر	O(N)	خیر

منظور از متریک بودن تابع فاصله در جدول (۱)، پیروی از خاصیت نامساوی مثلثی[۱] می‌باشد. با توجه به این جدول زمان محاسبات در فاصله‌ی انحراف زمانی پویا به صورت توان دو بوده و این تابع فاصله توانایی محاسبه‌ی فاصله بین سری‌های زمانی با طول‌های مختلف را دارا می‌باشد.

۳- ارائه‌ی یک روش خوشه‌بندی توافقی بر مبنای الگوریتم Fuzzy C- Means با ترکیب توابع فاصله‌ی مختلف

در این بخش یک روش خوشه‌بندی بهبود یافته‌ی Fuzzy C- Means با توانایی استفاده همزمان از چندین تابع فاصله‌ی مختلف ارائه شده است. روش خوشه‌بندی Fuzzy C- Means یکی از معروف‌ترین مدل‌های خوشه-بندی تفکیکی فازی می‌باشد که در این روش هر شیء می‌تواند به چند خوشه با درجه عضویت بین صفر و یک تعلق داشته باشد. اگر $r_1, r_2, \dots, r_n$ نشان‌دهنده‌ی n شی مختلف باشد، در روش خوشه‌بندی Fuzzy C- Means تابع هدف رابطه‌ی (۹) کمینه می‌گردد[۲۳].

$$J = \sum_{i=1}^{C}\sum_{k=1}^{n} u_{ik}^m d^2(v_i, r_k) \quad (۹)$$

در رابطه‌ی (۹)، m ضریب فازی کننده و $d(.)$ تابع فاصله-ی در نظر گرفته شده می‌باشد. در گام اول تعداد c داده‌ی مختلف به صورت تصافی از میان مجموعه داده انتخاب می-شوند. با درنظر این داده‌ها به عنوان مراکز خوشه ماتریس درجه عضویت طبق رابطه‌ی (۱۰) محاسبه می‌گردد.

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d(v_i, r_k)}{d(v_j, r_k)}\right)^{\left(\frac{2}{m-1}\right)}} \quad (۱۰)$$

در گام بعدی به منظور کمینه کردن تابع هدف رابطه-ی (۹) مراکز خوشه‌های جدید با استفاده از رابطه‌ی (۱۱) بدست می‌آیند.

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m r_k}{\sum_{k=1}^{n} u_{ik}^m} \quad (۱۱)$$

۱ Triangle inequality

الگوریتم با محاسبه‌ی ماتریس درجه‌ی عضویت و مراکز خوشه‌های جدید در تکرارهای متوالی ادامه می‌یابد تا زمانی که تغییر محسوسی در مراکز خوشه‌های حاصل مشاهده نگردد.

در این تحقیق یک روش خوشه‌بندی با قابلیت استفاده از توابع فاصله‌ی مختلف در عملیات خوشه‌بندی به طور همزمان پیشنهاد داده می‌شود. ماتریس درجه عضویت در روش پیشنهادی طبق رابطه‌ی (۱۲) محاسبه می‌گردد.

$$U = w_1U_1 + w_2U_2 + \cdots + w_pU_p \qquad (۱۲)$$

در این رابطه U_i نشان‌دهنده‌ی ماتریس درجه عضویت حاصل از در نظر گرفتن iامین تابع فاصله به عنوان معیار شباهت در فرآیند خوشه‌بندی می‌باشد. همچنین w_i وزن مربوط به تابع فاصله‌ی iام در فرآیند خوشه‌بندی را نشان می‌دهد. هریک از این وزن‌ها به گونه‌ای انتخاب می‌شوند که رابطه‌ی (۱۳) میان آنها برقرار باشد.

$$w_1 + w_2 + \cdots + w_p = 1 \quad , \quad 0 \le w_i \le 1 \qquad (۱۳)$$

با در نظر گرفتن w_i= ۰ اثر تابع فاصله‌ی iام از فرآیند خوشه‌بندی حذف خواهد شد در حالی که w_i= ۱ اثر تمامی توابع فاصله را حذف کرده و تنها تابع فاصله‌ی iام را در الگوریتم خوشه‌بندی در نظر می‌گیرد. به طور کلی مقادیر بزرگتر w_i اثر تابع فاصله iام را در فرآیند خوشه-بندی افزایش داده در حالی که اثر سایر توابع فاصله کاهش می‌یابند.

به منظور محاسبه‌ی درجه‌ی توافق، یک معیار ارزیابی بر اساس تابع هدف الگوریتم FCM تعریف گردید. اگر U ماتریس درجه‌ی عضویت توافقی حاصل از خوشه‌بندی داده‌ها با در نظر گرفتن توابع فاصله‌ی مختلف باشد (رابطه‌ی ۱۲)، کیفیت خوشه‌های بدست آمده با استفاده از رابطه‌ی (۱۴) محاسبه می‌گردد.

$$E = \sum_{i=1}^{p} \frac{J(U)_{Dist_i}}{J_{Dist_i}} - p \qquad (۱۴)$$

در رابطه‌ی (۱۴)، $J(U)_{Dist_i}$ نشان‌دهنده‌ی مقدار تابع هدف الگوریتم FCM می‌باشد در صورتی که $Dist_i$ به عنوان تابع فاصله و U به عنوان ماتریس درجه‌ی عضویت در نظر گرفته شوند. همچنین J_{Dist_i} نشان‌دهنده‌ی مقدار تابع هدف الگوریتم FCM به ازای تابع فاصله‌ی $Dist_i$ می‌باشد. از آنجا که مقادیر $J(U)_{Dist_i}$ به ازای توابع فاصله‌ی مختلف متفاوت می‌باشند، به منظور نرمالسازی این مقادیر از عبارت J_{Dist_i} در مخرج رابطه‌ی (۱۴) استفاده کردیم. توجه به این نکته ضروری است که زمانی که $Dist_i$ به عنوان تابع فاصله‌ی مورد استفاده در فرآیند خوشه‌بندی در نظر گرفته شود، J_{Dist_i} مقدار بهینه‌ی حاصل از اجرای الگوریتم می‌باشد بنابراین $J(U)_{Dist_i} \ge J_{Dist_i}$ و یا به عبارت دیگر E ≥ ۰. بر اساس رابطه‌ی (۱۴)، E سطح تناسب ماتریس درجه‌ی عضویت توافقی (U) را با توجه به توابع فاصله‌ی شرکت کننده در فرآیند خوشه‌بندی و با استفاده از تابع هدف الگوریتم FCM محاسبه می‌کند. در مواردی که ماتریس‌های درجه عضویت حاصل از هریک از توابع فاصله‌ی مختلف در سطح بالایی از توافق با نتیجه‌ی حاصل از روش پیشنهادی باشند مقدار کمتری برای E بدست خواهد آمد. همچنین E=0 بیانگر این نکته است که ماتریس درجه‌ی عضویت توافقی بدست آمده از روش پیشنهادی در توافق کامل با ماتریس‌های درجه‌ی عضویت حاصل از هریک از توابع فاصله‌ی شرکت کننده در فرآیند خوشه‌بندی می‌باشد. به عبارت دیگر ساختار کشف شده با استفاده از روش پیشنهادی مورد توافق تمامی توابع فاصله‌ی شرکت کننده می‌باشد. با توجه به روابط (۱۲) و (۱۴) با یافتن مقادیر بهینه برای وزن‌های $w_1, w_2 \ldots, w_p$ مقدار بهینه‌ی E بدست می‌آید. از آنجا که یافتن مقادیر بهینه‌ی وزن‌ها با روش‌های ریاضی زمانبر می‌باشد، این مسئله می‌تواند به عنوان یک مسئله‌ی بهینه‌سازی مطرح شود. الگوریتم‌های مبتنی بر هوش جمعی تکنیک‌های بهینه‌یابی و جستجوی تصادفی هستند که خصوصیت اصلی آنها رفتار گروهی و خود سازماندهی شده‌ی اعضای موجود در اجتماع می‌باشد از این رو این الگوریتم‌ها روش‌هایی کارآمد و قدرتمند برای یافتن راه‌حل‌های نزدیک به بهینه می‌باشند. در این تحقیق از الگوریتم انبوه ذرات به منظور یافتن وزن‌های بهینه‌ی توابع فاصله‌ی مختلف استفاده شده است. الگوریتم انبوه ذرات به عنوان یکی از شناخته‌شده‌ترین الگوریتم‌های مبتنی بر جمعیت در حل مسائل پیوسته همواره مورد توجه محققان بوده و توانایی این روش در تحقیقات مختلف به اثبات رسیده است [۲۰ و ۳۹ و ۴۲ و ۴۳].

برای یافتن وزن‌های بهینه با استفاده از الگوریتم انبوه ذرات ابتدا مجموعه‌ای از ذرات تولید می‌شوند به طوری که هر ذره شامل تعداد p المان می‌باشد و هر المان نماینده‌ی

یک وزن مشخص می‌باشد. سپس وزن‌های موجود در هر ذره برای محاسبه‌ی ماتریس درجه‌ی عضویت توافقی (U) طبق رابطه‌ی (۱۲) مورد استفاده قرار می‌گیرند. در مرحله‌ی بعد با محاسبه‌ی مقدار E، میزان کیفیت خوشه-های حاصل محاسبه می‌گردد. در حقیقت E تعریف شده در رابطه‌ی (۱۴) به عنوان تابع بهینگی الگوریتم انبوه ذرات در نظر گرفته خواهد شد. در گام بعدی ذرات با توجه به مقادیر تابع هدف و با استفاده از رابطه‌ی (۱۵) بروزرسانی می‌شوند [۴۲].

$$X_i(t+1) = X_i(t) + V_i(t+1) \qquad (۱۵)$$

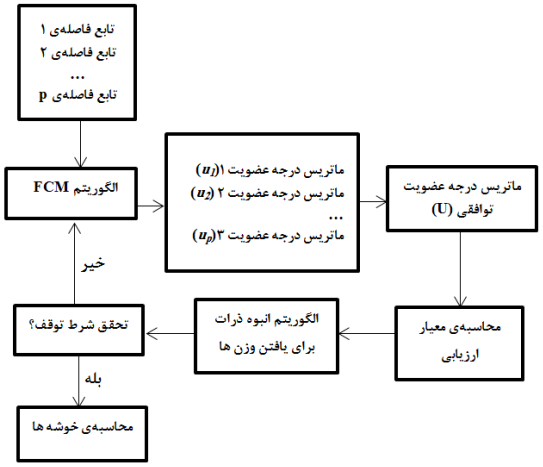
در رابطه‌ی (۱۵)، $V_i(t+1)$ بردار حرکت[۱] ذره در زمان $t+1$ است که با استفاده از رابطه‌ی (۱۶) محاسبه می‌گردد [۴۲].

$$V_i(t+1) = w.V_i(t) + C_1 r_1\big(pbest_i(t) - X_i(t)\big) + C_2 r_2\big(Gbest_i(t) - X_i(t)\big) \qquad (۱۶)$$

در رابطه‌ی (۱۶)، $Pbest_i(t)$ بهترین مکانی است که ذره i از ابتدای الگوریتم تاکنون با آن مواجه شده است در حالی که $Gbest_i(t)$ بهترین مکانی است که همسایگان ذره i از ابتدا تاکنون با آن مواجه بوده‌اند. همچنین c_1 و c_2 ضرایب شتاب نامیده می‌شوند و این ضرایب تعیین می-کنند که هر ذره با چه شتابی به سمت $pbest$ یا $Gbest$ حرکت کند. r_1 و r_2 نیز اعدادی تصادفی در بازه [۱ و۰] می‌باشند که با استفاده از توزیع یکسان، تولید می‌شوند. در نهایت متغیر w در این رابطه وزن اینرسی[۲] نامیده می‌شود و برای تضمین همگرایی ذرات به کار می‌رود.

لازم به ذکر است فرآیند بروزرسانی ذرات تا بدست آمدن مقادیر بهینه‌ی وزن‌ها ادامه خواهد یافت. در شکل (۱) مراحل روش پیشنهادی نشان داده شده است.

به طور خلاصه در مرحله‌ی نخست روش پیشنهادی با استفاده از الگوریتم خوشه‌بندی FCM و با در نظر گرفتن توابع فاصله‌ی مختلف به صورت جداگانه اقدام به خوشه-بندی داده‌های موجود در مجموعه داده می‌کند. در پایان این مرحله ماتریس‌های درجه‌ی عضویت مختلف منتسب به توابع فاصله‌ی مختلف حاصل می‌گردد. در مرحله‌ی دوم ماتریس درجه‌ی عضویت توافقی (U) با استفاده از مجموع وزن‌دار ماتریس‌های درجه‌ی عضویت حاصل در مرحله‌ی قبل محاسبه می‌گردد. لازم به ذکر است که در ابتدا وزن‌ها به صورت تصادفی مقداردهی می‌شوند. در گام سوم با استفاده از ماتریس درجه‌ی عضویت توافقی و همچنین توابع فاصله‌ی مختلف به محاسبه‌ی معیار ارزیابی انتخابی طبق رابطه‌ی (۱۴) می‌پردازیم. در نهایت با در نظر گرفتن معیار ارزیابی انتخابی به عنوان تابع هدف از الگوریتم انبوه ذرات به منظور یافتن مقدار بهینه‌ی وزن هر تابع فاصله استفاده شد.



شکل ۱- فلوچارت مراحل روش خوشه‌بندی توافقی پیشنهادی

۴- پیاده‌سازی و اجرا

در این بخش چگونگی عملکرد روش پیشنهادی با استفاده از چندین مجموعه از سری‌های زمانی مورد تحلیل و بررسی قرر می‌گیرد.

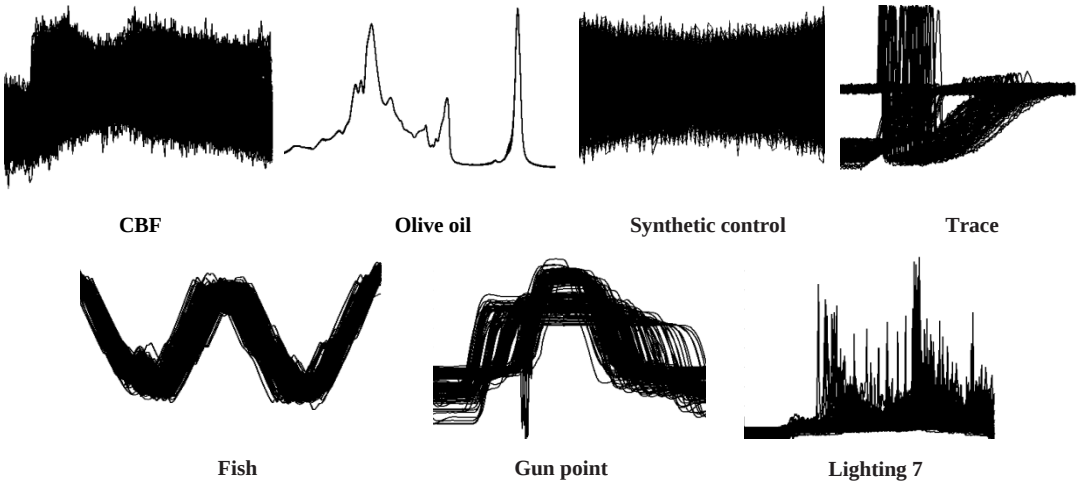
۴-۱- مجموعه داده‌ها

به منظور ارزیابی، روش پیشنهادی بر روی هفت سری مجموعه داده‌ی شناخته شده از سری‌های زمانی که در تحقیقات مختلف مورد استفاده قرار گرفتند [۴۲ و ۴۴]، پیاده‌سازی شد. توضیحات مربوط به هریک از این داده‌ها در سایت UCR به طور مفصل آورده شده است. (https://www.cs.ucr.edu/~eamonn/time_series_data/) 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. همچنین در شکل (۲) نمایی از مجموعه داده-ها نشان داده شده است.

[۱] Velocity Vector
[۲] Inertia Weight

جدول ۲- خصوصیات مجموعه داده‌های استفاده شده در این تحقیق

مجموعه داده	طول سری‌های زمانی	اندازه‌ی مجموعه داده	تعداد خوشه‌های موجود
CBF	۱۲۸	۹۳۰	۳
Trace	۲۷۵	۲۰۰	۴
Synthetic control	۶۰	۶۰۰	۶
Gun point	۱۵۰	۲۰۰	۲
Lighting 7	۳۱۹	۱۴۳	۷
Olive oil	۵۷۰	۶۰	۴
Fish	۴۶۳	۳۵۰	۷



شکل ۲- نمایش مجموعه سری‌های زمانی استفاده شده در این تحقیق

۴-۲- تنظیم پارامترها

با در نظر گرفتن مقادیر مختلف برای پارامترهای الگوریتم انبوه ذرات و با استفاده از روش سعی و خطا، مقدار مناسب هر پارامتر با توجه به جدول (۳) حاصل شده است. در جدول (۳)، P اندازه‌ی جمعیت، Itr تعداد تکرارها، c_1 و c_1 ضرایب شتاب و w_{min} و w_{max} کمینه و بیشینه وزن اینرسی استفاده شده در مسئله می‌باشد.

جدول ۳- مقادیر پارامترهای استفاده شده در الگوریتم انبوه ذرات

پارامترها	مقادیر
p	۱۰
Itr	۴۰
c_1	۲/۰۲
c_2	۲/۰۲
w_{min}	۰/۴
w_{max}	۰/۹

مقادیر بیشینه و کمینه‌ی وزن اینرسی در این جدول نشان می‌دهد که در ابتدای الگوریتم وزن اینرسی برابر با ۰/۹ در نظر گرفته شده است اما در تکرارهای بعدی از مقدار این پارامتر کاسته و در نهایت به مقدار ۰/۴ ختم می‌شود.

۴-۳- مقایسه و ارزیابی نتایج

با توجه به هدف تعیین شده از ارائه‌ی یک روش خوشه‌بندی تاکنون روش‌های مختلفی به منظور ارزیابی و بررسی عملکرد روش‌های خوشه‌بندی پیشنهاد داده شده است. بررسی مقادیر بدست آمده برای تابع هدف روش خوشه‌بندی استفاده شده[۲۱ و ۴۳]، استفاده از شاخص-های اعتبارسنجی خوشه[۴۵ و ۴۶]، محاسبه‌ی میزان فشردگی اعضای یک خوشه و جدایی و تفکیک پذیری بین خوشه‌های متفاوت [۲۰ و ۴۵] محاسبه و تحلیل مقادیر معیارهای دقت[۱]، معیار فیشر[۲] و معیار فراخوان[۳] [۳۹] از

۱ Precision
۲ F-measure
۳ Recall

جمله معیارهای ارزیابی روش‌های خوشه‌بندی می‌باشند. از آنجا که هدف این تحقیق ارائه‌ی یک روش خوشه‌بندی با تمرکز بر بدست آوردن حداکثر میزان توافق بین توابع فاصله‌ی مختلف بوده است در این مطالعه از متوسط اطلاعات مشترک نرمال‌شده [۴۲] به عنوان معیار ارزیابی نتایج استفاده شد. متوسط اطلاعات مشترک نرمال‌شده معیاری است که میزان توافق میان خوشه‌بندی‌های مختلف را محاسبه می‌کند و در تحقیقات متعددی مورد استفاده قرار گرفته است [۴۲ و ۴۷ و ۴۸ و ۴۹].

اطلاعات مشترک نرمال شده با استفاده از رابطه‌ی (۱۷) محاسبه می‌شود.

$$\emptyset(C_1, C_2) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} n_{ij}\log(\frac{n\, n_{ij}}{n_i n_j})}{\sqrt{(\sum_{i=1}^{c} n_i \log\frac{n_i}{n})(\sum_{j=1}^{c} n_i \log\frac{n_j}{n})}} \qquad (۱۷)$$

در این معادله C_1 و C_2 نتایج حاصل از خوشه‌بندی با استفاده از دو تابع فاصله‌ی اول و دوم می‌باشد. همچنین c، تعداد خوشه‌ها و n تعداد داده‌های شرکت کننده در فرآیند خوشه‌بندی را نشان می‌دهند. علاوه بر این n_i نشان‌دهنده‌ی تعداد داده‌های موجود در iمین خوشه‌ی حاصل از تابع فاصله‌ی اول و n_j نشان‌دهنده‌ی تعداد داده‌های موجود در jمین خوشه‌ی حاصل از تابع فاصله‌ی دوم می‌باشند. در نهایت n_{ij} تعداد داده‌های موجود در iمین خوشه‌ی حاصل از تابع فاصله‌ی اول و jمین خوشه‌ی حاصل از تابع فاصله‌ی دوم می‌باشد. با در نظر گرفتن تعداد p تابع فاصله‌ی مختلف اگر C نتیجه‌ی حاصل از خوشه‌بندی داده‌ها با استفاده از روش پیشنهادی باشد متوسط اطلاعات مشترک نرمال‌شده با استفاده از رابطه‌ی (۱۸) بدست می‌آید.

$$\emptyset\left(C, C_{1,2,\dots,p}\right) = \frac{1}{p}\sum_{i=1}^{p} NMI(C, C_i) \qquad (۱۸)$$

مقدار حاصل از متوسط اطلاعات مشترک نرمال‌شده در بازه‌ی ۰ تا ۱ قرار می‌گیرد و مقدار مطلوب به عدد یک نزدیکتر می‌باشد. به عبارت دیگر هرچقدر مقدار متوسط اطلاعات مشترک نرمال شده بیشتر باشد C در توافق بیشتری با C_1، C_2 و غیره می‌باشد. با در نظر گرفتن سه تابع فاصله‌ی اقلیدسی، انحراف زمانی پویا و ضرایب پیرسون عملکرد روش پیشنهادی با پنج روش مختلف به شرح زیر مقایسه و نتایج حاصل در جدول (۴) ثبت گردید.

روش ۱: در این روش خوشه‌بندی با استفاده از الگوریتم FCM و تابع فاصله‌ی اقلیدسی انجام شده و نتیجه‌ی حاصل به عنوان C در رابطه‌ی (۱۸) در نظر گرفته می‌شود. به عبارت دیگر در این روش اثر توابع فاصله‌ی انحراف زمانی پویا و ضرایب پیرسون صفر در نظر گرفته می‌شود.

روش ۲: در روش ۲ تابع فاصله‌ی انحراف زمانی پویا را به عنوان معیار شباهت داده‌ها در نظر گرفته و از الگوریتم FCM به منظور خوشه‌بندی داده‌ها استفاده می‌شود. نتیجه‌ی حاصل به عنوان خوشه‌بندی نهایی حاصل در رابطه‌ی (۱۸) قرار می‌گیرد.

روش۳: در این روش تابع فاصله‌ی ضرایب پیرسون به عنوان معیار شباهت داده‌ها در نظر گرفته می‌شود.

روش ۴: در این روش به هریک از توابع فاصله‌ی شرکت کننده در عملیات خوشه‌بندی وزن‌های یکسان تعلق میگیرد. به عبارت دیگر در این مطالعه به هریک از توابع فاصله‌ی اقلیدسی، انحراف زمانی پویا و ضرایب پیرسون وزنی برابر با ۱/۳ اختصاص می‌یابد.

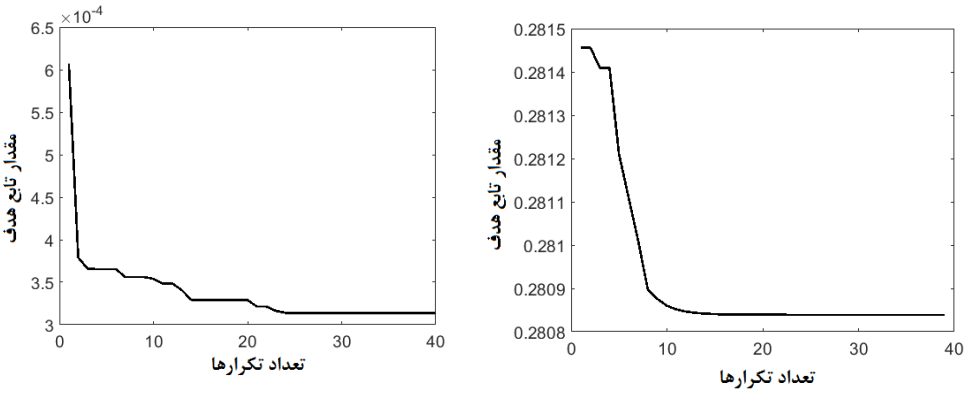
روش ۵: در این حالت یک تابع فاصله کلی از مجموع وزندار سه تابع فاصله‌ی انتخابی طبق رابطه‌ی (۱۹) محاسبه می‌گردد.

$$d = \sum_{i=1}^{3} w_i d_i \qquad (۱۹)$$

در رابطه‌ی (۱۹)، d_i هریک از توابع فاصله‌ی در نظر گرفته شده در روش پیشنهادی و w_i وزن‌های مربوط به هریک از توابع فاصله می‌باشد. در مرحله‌ی نخست از این روش برای هریک از این توابع فاصله وزنی دلخواه انتخاب می‌کنیم به طوری که مجموع وزن‌ها برابر با عدد یک شود. در مراحل بعد هریک از این وزن‌ها با استفاده از الگوریتم انبوه ذرات به‌گونه‌ای بهینه می‌گردند که تابع هدف الگوریتم FCM کمینه گردد. نتیجه‌ی خوشه‌بندی نهایی به عنوان C در رابطه‌ی (۱۸) جایگزین می‌گردد. این روش حالت بهبود یافته‌ی روشی است که در سال ۲۰۰۸ توسط Zhang و همکارانش پیشنهاد داده شد [۵۰].

در شکل (۳) چگونگی همگرایی الگوریتم انبوه ذرات در دو مجموعه‌ی داده‌ی Trace و CBF نشان داده شده است. از این اشکال می‌توان دریافت میزان بهبود ذرات در

اجراهای نخست بیشتر بوه است. به عبارتی دیگر ذرات در تکرارهای اولیه با سرعت بیشتری به جواب بهینه نزدیک شدند. همچنین این نمودارها نشان می‌دهند تغییرات مقدار تابع هدف حدودا از تکرار ۳۰ به بعد به صفر نزدیک شده است.



مجموعه داده‌ی CBF　　مجموعه داده‌ی Trace

شکل ۳- چگونگی همگرایی الگوریتم انبوه ذرات در یافتن وزن‌های بهینه

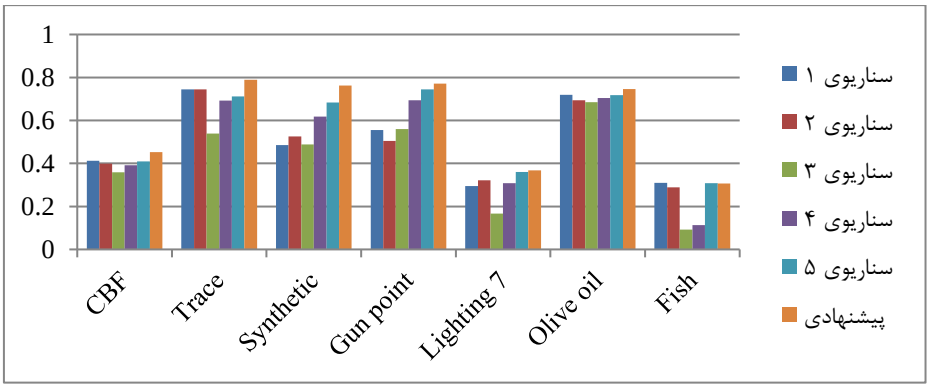
در جدول (۴) مقادیر معیار متوسط اطلاعات مشترک نرمال‌شده در روش پیشنهادی و پنج روش اشاره شده برای هفت مجموعه داده‌ی انتخابی آورده شده است. لازم به ذکر است هریک از روش‌ها ۱۰۰ بار به طور مستقل اجرا شده و میانگین نتایج حاصل به منظور مقایسه‌ی عملکرد روش‌ها مورد استفاده قرار گرفته است.

جدول ۴- مقادیر متوسط اطلاعات مشترک نرمال‌شده در ۱۰۰ اجرا

مجموعه داده	روش ۱	روش ۲	روش ۳	روش ۴	روش ۵	روش پیشنهادی
CBF	۰/۴۱۱۹	۰/۳۹۸۹	۰/۳۵۸۲	۰/۳۹۱۸	۰/۴۰۸۷	**۰/۴۵۳۶**
Trace	۰/۷۴۵۱	۰/۷۴۴۴	۰/۵۳۸۴	۰/۶۹۲۳	۰/۷۱۲۳	**۰/۷۸۹۵**
Synthetic control	۰/۴۸۴۶	۰/۵۲۵۹	۰/۴۸۷۷	۰/۶۱۷۹	۰/۶۸۳۹	**۰/۷۶۳۱**
Gun point	۰/۵۵۶۱	۰/۵۰۴۵	۰/۵۵۹۶	۰/۶۹۴۵	۰/۷۴۵۳	**۰/۷۷۱۴**
Lighting 7	۰/۲۹۴۵	۰/۳۲۲۰	۰/۱۶۵۹	۰/۳۰۷۹	۰/۳۵۹۷	**۰/۳۶۸۳**
Olive oil	۰/۷۱۹۸	۰/۶۹۳۸	۰/۶۸۴۶	۰/۷۰۴۱	۰/۷۱۸۲	**۰/۷۴۶۶**
Fish	**۰/۳۰۹۴**	۰/۲۸۸۹	۰/۰۹۱۷	۰/۱۱۳۴	۰/۳۰۷۵	۰/۳۰۶۳

در شکل (۴) مقادیر بدست آمده برای معیار متوسط اطلاعات مشترک نرمال شده حاصل از شش روش در هفت مجموعه داده‌ی مختلف با هم مقایسه شدند. همانگونه که در شکل (۴) به خوبی مشهود است، نمودار مربوط به روش پیشنهادی ما در شش مورد از مجموع هفت مورد، بیشترین (بهترین) مقدار معیار در نظر گرفته شده را به خود اختصاص داده است. به عبارت دیگر در ۸۵ درصد موارد روش پیشنهادی بهتر از سایر روش‌ها عمل کرده است حال آنکه تنها در مجموعه داده‌ی Fish (۱۵ درصد موارد) با استفاده از روش اول جواب بهتری حاصل شده است.

علاوه بر این مقایسه‌ی نتایج حاصل از پنج روش اول نشان داد که در چهار مجموعه داده‌ی CBF، Trace، Olive oil و Fish از مجموع هفت مجموعه یا به عبارتی دیگر در ۵۷ درصد موارد نتایج حاصل از روش اول بهتر از چهار روش دیگر بوده است حال آنکه در ۴۳ درصد موارد روش پنجم بهتر از روش‌های اول تا چهارم عمل کرده است. علاوه بر این مقایسه‌ی روش‌های دوم، سوم و چهارم نشان می‌دهد در مجموع روش دوم قدرتمندتر از روش چهارم ظاهر شده است حال آنکه عملکرد سناریوی سوم ضعیف‌تر از سایر روش‌ها بوده است.



شکل ۴- مقایسه‌ی مقادیر متوسط اطلاعات مشترک نرمال‌شده حاصل از شش روش مختلف

میزان اختلاف نتایج حاصل از روش‌های مورد مقایسه امری مهم و قابل توجه می‌باشد و تنها در صورت معنی‌دار بودن این اختلافات می‌توان به برتری عملکرد یک روش نسبت به روش دیگر اذعان داشت. از این رو در ادامه به منظور مقایسه‌ی بهتر روش‌ها در این مطالعه از روش مقایسه‌ی چندگانه‌ی توکی (Tukey) با حد آستانه‌ی $p < ۰/۰۵$ با توانایی مقایسه‌ی دو به دوی روش‌ها استفاده شد [۵۱].

در جدول (۵) نتایج حاصل از آزمون مقایسه‌ی توکی آورده شده است. بر اساس اطلاعات موجود در این جدول در مجموعه داده‌های CBF، Synthetic control، Gun point و Olive oil روش پیشنهادی نسبت به پنج روش دیگر به طور معناداری بهتر عمل کرده است ($p <۰/۰۵$). همچنین در مجموعه‌ی داده‌ی Trace برتری روش پیشنهادی نسبت به روش دوم، سوم، چهارم و پنجم قابل توجه بوده است حال آنکه اختلاف نتایج حاصل از روش پیشنهادی و روش اول در این مجموعه داده معنادار نبوده است. به بیانی دیگر در مجموعه داده‌ی Trace با در نظر گرفتن درجه اطمینان ۹۵ درصد میزان برتری عملکرد روش پیشنهادی نسبت به روش اول قابل توجه نمی‌باشد.

در مجموعه داده‌ی Lighting 7 تفاوت میان نتایج حاصل از روش پیشنهادی و روش‌های اول، دوم، سوم و چهارم معنادار بوده است حال آنکه برتری روش پیشنهادی در مقایسه با روش پنجم قابل توجه نمی‌باشد زیرا مقدار p حاصل از مقایسه‌ی روش پیشنهادی و روش پنجم بزرگتر از ۰/۰۵ می‌باشد. در نهایت در مجموعه داده‌ی Fish روش پیشنهادی نسبت به روش سوم و چهارم به طور قابل توجه‌ای بهتر عمل کرده است. همچنین مقدار p حاصل در این مجموعه داده نشان می‌دهد برتری روش اول نسبت به روش پیشنهادی معنادار نبوده است ($p >۰/۰۵$).

با توجه به اطلاعات موجود در این جدول اختلاف نتایج حاصل از روش‌های اول و پنجم تنها در مجموعه داده‌های Synthetic control، Trace، Lighting 7 و Gun point معنی‌دار می‌باشد حال آنکه در سایر موارد این اختلاف قابل توجه نیست. علاوه بر این نتایج حاصل از آزمون توکی نشان می‌دهد برتری عملکرد دو روش دوم و چهارم نسبت به هم تنها در چهار مجموعه داده قابل توجه می‌باشد و در مجموعه‌های دیگر اختلاف عملکرد آنها ناچیز می‌باشد.

جدول ۵- مقادیر p حاصل از روش مقایسه‌ی چندگانه‌ی توکی

روش	CBF	Trace	Synthetic control	Gun point	Lighting 7	Olive oil	Fish
روش ۱- روش ۲	۰/۹۶۵۲	۰/۸۶۴۴	۰/۰۳۶۶	۰/۰۰۰۰	۰/۸۰۳۷	۰/۰۸۱۸	۰/۸۵۲۹
روش ۱- روش ۳	۰/۰۰۰۱	۰/۰۰۰۰	۰/۹۹۹۹	۱,۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
روش ۱-روش ۴	۰/۰۳۳۵	۰/۰۰۰۱	۰/۰۰۰۰	۰/۰۰۰۰	۰/۳۸۷۲	۰/۵۹۰۶	۰/۰۰۰۰
روش ۱-روش ۵	۰/۹۹۵۹	۰/۰۰۲۵	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۶۴۰۷	۰/۸۶۵۱
روش ۱- روش پیشنهادی	۰/۰۰۷۱	۰/۱۶۵۶	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۱	۰/۰۴۳۲	۰/۸۰۶۳
روش ۲- روش ۳	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۲۵۱	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۴۷۶	۰/۰۰۰۰
روش ۲-روش ۴	۰/۱۶۰۸	۰/۰۰۰۱	۰/۰۰۰۰	۰/۰۰۰۰	۰/۸۸۰۱	۰/۹۶۱۱	۰/۰۰۰۰

روش	CBF	Trace	Synthetic control	Gun point	Lighting 7	Olive oil	Fish
روش ۲-روش ۵	۰/۹۹۸۵	۰/۰۳۷۵	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۳	۰/۰۷۵۵	۱/۰۰۰۰
روش ۲- روش پیشنهادی	۰/۰۴۰۲	۰/۰۱۶۷	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۱۹	۰/۰۰۸۶	۱/۰۰۰۰
روش ۳- روش ۴	۰/۸۹۰۱	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۹۹۹۲	۰/۷۲۸۵
روش ۳- روش ۵	۰/۰۱۱۳	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۹۸	۰/۰۰۰۰
روش ۳- روش پیشنهادی	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۱	۰/۰۰۰۰
روش ۴- روش ۵	۰/۱۶۲۸	۰/۱۸۹۰	۰/۰۰۰۷	۰/۰۰۰۰	۰/۰۰۱۳	۰/۰۲۶۲	۰/۰۰۰۰
روش ۴- روش پیشنهادی	۰/۰۰۰۱	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۵۵	۰/۰۰۰۶	۰/۰۰۰۰
روش ۵- روش پیشنهادی	۰/۰۱۹۵	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۱۸۱	۰/۹۸۱۳	۰/۰۴۳۵	۰/۹۹۹۹

نتایج حاصل از این تحقیق نشان می‌دهد روش پیشنهادی به دلیل قابلیت ترکیب نتایج خوشه‌بندی حاصل از توابع فاصله‌ی مختلف، در مجموع بهتر از روش-های اول، دوم و سوم در تولید نتیجه‌ای با بیشترین توافق میان توابع فاصله‌ی مختلف عمل می‌کند. دلیل اصلی عملکرد بهتر روش پیشنهادی در مقایسه با روش چهارم استفاده از الگوریتم انبوه ذرات به عنوان یکی از شناخته-شده‌ترین الگوریتم‌های مبتنی بر جمعیت به منظور یافتن وزن‌های بهینه می‌باشد. دلیل استفاده از این گروه از الگوریتم‌ها در این مسائل، توانایی آنها در غلبه بر مسائل پیچیده و چندهدفه، عدم نیاز به درک و مدلسازی دقیق پدیده، شکل تصادفی جستجو در فضا و همچنین مدلسازی پدیده‌های دینامیک به علت امکان تولد، فعالیت و مرگ ذرات می‌باشد. از دیگر نقاط قوت روش ارائه شده در این مطالعه تعریف و استفاده از یک تابع هدف مناسب و با توانایی نرمالسازی اثر توابع فاصله‌ی مختلف می‌باشد. جستجوی وزن هریک از معیارهای شباهت در این فضای نرمال شده سبب کسب نتایج بهتر می‌گردد. برتری روش پیشنهادی نسبت به روش پنجم به روشنی گویای این مطلب می‌باشد.

۵- نتیجه‌گیری و پیشنهادات

در این مقاله یک روش خوشه‌بندی توافقی بر اساس روش خوشه‌بندی FCM و الگوریتم انبوه ذرات به منظور خوشه‌بندی سری‌های زمانی پیشنهاد داده شد. از آنجا که انتخاب تابع فاصله‌ی مناسب به عنوان یکی از پیچیده‌ترین و زمانبرترین چالش‌های موجود در فرآیند خوشه‌بندی مطرح می‌باشد، در این تحقیق یک روش خوشه‌بندی با قابلیت استفاده از چندین تابع فاصله به طور همزمان ارائه شد. به عبارتی دیگر در این روش به هر تابع فاصله یک وزن اختصاص می‌یابد که مقدار بهینه‌ی این وزن‌ها با توجه به تابع هدف تعریف شده مشخص می‌گردد. از آنجا که تابع هدف انتخابی تاثیری مستقیم بر مقادیر وزن‌ها و کیفیت نتایج دارد، با استفاده از تابع هدف روش خوشه‌بندی FCM به تولید تابعی متناسب با مسئله‌ی موجود پرداخته و از الگوریتم انبوه ذرات به منظور بهینه‌سازی تابع هدف تعریف شده استفاده شد. به منظور ارزیابی روش ارائه شده، با در نظر گرفتن سه تابع فاصله‌ی اقلیدسی، انحراف زمانی پویا و ضرایب پیرسون، نتایج حاصل از پیاده‌سازی شش روش مختلف بر روی هفت مجموعه داده‌ی سری‌های زمانی با هم مقایسه شدند. با بررسی مقادیر بدست آمده برای معیار متوسط اطلاعات مشترک نرمال شده در شش روش مورد مقایسه و همچنین با توجه به نتایج حاصل از آزمون توکی برتری روش پیشنهادی به اثبات رسید.

برخی از توابع فاصله مانند تابع پیچش زمانی پویا به زمان بالایی برای محاسبه‌ی میزان شباهت میان دو داده نیازمند است که این مسئله یکی از چالش‌های موجود در این تحقیق بوده است. پیشنهاد می‌شود در تحقیقات آینده با بررسی روش‌های سریع محاسبه‌ی این توابع فاصله سرعت اجرای برنامه‌ها را افزایش داد.

یکی از مهم‌ترین چالش‌های موجود در مسئله‌ی خوشه‌بندی داده تعیین تعداد بهینه‌ی خوشه‌ها در مجموعه داده‌هایی با تعداد خوشه‌های نامعلوم می‌باشد. توسعه‌ی روش پیشنهای به منظور افزایش قابلیت آن در تعیین تعداد خوشه‌ها به عنوان یکی از کارهای آینده مطرح می‌باشد.

از آنجا که خطوط سیر داده‌های پیچیده‌تری نسبت به سری‌های زمانی می‌باشند توسعه‌ی روش ارائه شده در این مطالعه با هدف استفاده بر روی مجموعه داده‌های خط سیر می‌تواند به عنوان یکی از پیشنهادات برای کارهای آینده مورد توجه قرار گیرد.

مراجع

[1] Duchon, C., Hale, R. (2012). “Time Series Analysis in Meteorology and Climatology”. Financial Engineering and Portfolio Management. An Introduction. Oxford: Wiley.

[2] Craddock, JM. (1973). “Problems and prospects for eigenvector analysis in meteorology”, The Statistician. 22: 133–145.

[3] Hsieh, WW. Tang, B. (1998). “Applying neural network models to prediction and data analysis in meteorology and oceanography”, Bulletin of the American Meteorological Society 79: 1855–1870.

[4] Das, S. and Politis, D. N. (2017). “Predictive inference for locally stationary time series with an application to climate data”, arXiv preprint arXiv:1712.02383

[5] Wei, W. (1990). “Time series analysis”, Addison-Wesley, Redwood City.

[6] Quan, L., and Schaub, D. (2004). “Economic globalization and transnational terrorist incidents: A pooled time series analysis”. Journal of Conflict Resolution 48 (2): 230-258.

[7] Granger, C. W. J. (2004). “Time Series Analysis, Cointegration, and Applications, American Economic Review”. 94(3), pp 421-425.

[8] Shirian, Z., Nikoomaram, H., Rahnamay Roodposhti, F. and Torabi, T. (2018). “Volatility clustering in financial markets based on the agent based model.” Financial Engineering and Portfolio Management. Vol.9, No.36, PP. 201-224.

[9] Asgari, S., Yazdani, N., Nazem Bokaei, M. (2017). “Predicting Index of Stock Exchange by Hidden Markov Model and K-Mean Algorithm.” Financial Knowledge of Securities Analysis. Vol.10, No.36, PP. 71-82.

[10] Paillard, D., Labeyrie, L. & Yiou, P. (1996). “Macintosh program performs time-series analysis”. Eos 77, 379.

[11] Li, Z., Fielding, E.J., Cross, P. (2009). “Integration of InSAR time-series analysis and water-vapor correction for mapping postseismic motion after the 2003 Bam (Iran) earthquake”. IEEE Trans. Geosci. Remote Sens, 47, 3220–3230.

[12] Ryder, N.B. (1964). The process of demographic translation. Demography 1, 74–82.

[13] Wah, W., Das, S., Earnest, A., Lim, L., Chee, C., Cook, A., Wang, Y., Win, K., Ong, M., Hsu, L. (2014). Time series analysis of demographic and temporal trends of tuberculosis in Singapore. BMC Public Health, 14(1):1–10.

[14] Martin, L., Zarzalejo, L., Polo, J., Navarro, A., Marchante, R., and Cony, M. (2010). "Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning", Sol. Energy, vol. 84, no. 10, pp. 1772-1781.

[15] Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. (1980). Interrupted time-series analysis and its application to behavioral data. Journal of Applied Behavior Analysis, 13, 543–559.

[16] Alami, M.T, Nourani, V., Daneshvar Vousoughi, F. (2016). “Using Spatial Clustering in Forecasting Groundwater Quality Parameters by ANFIS.” Journal of Water and Wastewater. Vol.27, No.3, PP. 62-74.

[17] Murthy, S.K., Begum, J., Benchimol, E.I, et al. (2019). Introduction of anti-TNF therapy has not yielded expected declines in hospitalisation and intestinal resection rates in inflammatory bowel diseases: a population-based interrupted time series study.

[18] Verma, M., Srivastava, M., Chack, N., Kumar Diswar, A., Gupta, N. (2012), A comparative study of various clustering algorithms in data mining. Int. J. Eng. Res. Appl. (IJERA) 2 1379–1384.

[19] Mennis, J., Guo, D., 2009. Spatial data mining and geographic knowledge discovery- An introduction. Computers, Environment and Urban Systems, 33(6), 403-408.

[20] Izakian Z, Mesgari M.S, Abraham A. (2016). “Automated clustering of trajectory data using a particle swarm optimization“. Comput. Environ. Urban. Syst, 55: 55–65.

[21] Izakian H, Pedrycz W, Jamal I. (2013). “Clustering spatio-temporal data: an augmented fuzzy C-means “, IEEE Transactions on Fuzzy Systems, 25 (5), 855–868.

[22] Sharif M, Alesheikh A. A. (2016). “A Review on the Process of Point Objects’ Movement and Their Trajectory Similarity Measurement Approaches“. GEJ; 7 (1):41-54

[23] Shaeri M, Abbaspour R. A. (2015). “Comparison of Distance Functions for Similarity Measurement in Spatial Trajectories“. JGST, 4 (3):201-212.

[24] Vlachos M, Gunopulos D, Kollios G. (2002). “Similar multidimensional trajectories“. 18th Int. Conf. on Data Engineering, 673-684.

[25] D'Urso, P., Maharaj, E.A. (2009). “Autocorrelation-based fuzzy clustering of time series“. Fuzzy Sets Syst. 160 (24), 3565–3589.

[26] Nanda, S.R., Mahanty, B., Tiwari, M.K. (2010). “Clustering Indian stock market data for portfolio management“. Expert Syst. Appl. 37 (12), 8793–8798.

[27] Yuan, G., Sun, P., Zhao, J. et al. (2017). “A review of moving object trajectory clustering algorithms”. Artif Intell Rev 47, 123–144 (2017).

[28] Moayedi A, Abbaspour RA, Chehreghan A. (2019). “An evaluation of the efficiency of similarity functions in density-based clustering of spatial trajectories”, Annals of GIS- Taylor & Francis.

[29] Abanda, A., Mori, U., and Lozano, J. A. (2019). "A review on distance based time series classification", Data Mining and Knowledge Discovery, vol. 33, no. 2, pp. 378-412.

[30] 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

[31] Keogh, E., Pazzani, M.J., 1999. Scaling up dynamic time warping to massive datasets. In: Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery, pp. 1–11.

[32] Sakurai Y, Yoshikawa M, Faloutsos C. (2005). “FTW: fast similarity search under the time warping distance”. In: Proceedings of the 24th ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems, pp. 326–337.

[33] Petitjean F, Gançarski P. (2012). “Summarizing a set of time series by averaging: from Steiner sequence to compact multiple alignment”. Theor. Comput. Sci. 414, 76–91.

[34] Jeong, Y.-S., Jeong, M.K., Omitaomu, O.A., (2011). Weighted dynamic time warping for time series classification. Pattern Recognit. 44 (9), 2231–2240.

[35] Bankó, Z., Abonyi, J., (2012). Correlation based dynamic time warping of multivariate time series. Expert Syst. Appl. 39 (17), 12814–12823.

[36] Pearson K. (1896). “Regression, Heredity and Panmixia,” Philosophical Trans. Royal Soc., vol. 187, pp. 253-318.

[37] Sobhe Bidari P, Manshaei R, Lohrasebi T, Feizi A, Malboobi M.A, Alirezaie J.(2008). "Time series gene expression data clustering and pattern extraction in Arabidopsis thaliana phosphatase-encoding genes,"BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on, vol., no., pp.1-6, 8-10 Oct.

[38] Wang H, Han S, Zheng K, Sadiq S, and Zhou X. (2013). “An Effectiveness Study on Trajectory Similarity Measures.” Paper presented at the Proceedings of the Twenty-Fourth Australasian Database Conference-Volume, 137, Adelaide, Australia

[39] Izakian, Z. & Mesgari, M. S. (2015). Fuzzy clustering of time series data: A particle swarm optimization approach. Journal of AI and Data Mining, vol. 3, no. 1, pp. 39-46.

[40] Kim, T. et al. (2018). Impact of similarity metrics on single-cell RNA-seq data clustering. Briefings in Bioinformatics, 343, pp.776–bby076 VL – IS.

[41] Shinde, K., & Mulay, P. (2017). Correlation based incremental clustering algorithm, a new approach. In Convergence in Technology (I2CT), 2017 2nd International Conference for (pp. 291-296). IEEE.

[42] Izakian H, Pedrycz W (2014) Agreement-based fuzzy C-means for clustering data with blocks of features. Neurocomputing 127:266–280.

[43] Mahdizadeh, E., Sadinezhad, S., Tavakoli Moghaddam, R. (2008). Optimization of fuzzy clustering criteria by a hybrid pso and fuzzy c-means clustering algorithm, Iranian journal of fuzzy systems, Volume 5 , Number 3; Page(s) 1 To 14.

[44] Keogh, E., Folias, T. (2002), The UCR Time Series Data Mining Archive, Available: http://www.cs.ucr.edu/»eamonn/TSDMA/index.html

[45] Xie. X.L., Beni, G., (1991). A validity measure for fuzzy clustering, IEEE Trans. Pattern Analysis and Machine Intelligence, 13(8), 841 -847.

[46] Pakhira, M. K., Bandyopadhyay, S., Maulik, U., (2004). Validity index for crisp and fuzzy clusters, Pattern Recognition Letters, vol. 37, no. 3, pp.487 -501

[47] He, Z., Xu, X., and Deng, S. (2008). "K-ANMI: A mutual information based clustering algorithm for categorical data", Information Fusion, vol. 9, pp. 223-233.

[48] Knops, ZF. Maintz, J.B., Viergever, M.A., Pluim, J.P., (2006). Normalized mutual information based registration using k-means clustering and shading correction. Med Image Anal 10(3):432–439

[49] McDaid, A.F., Greene, D., Hurley, N. (2011). Normalized mutual information to evaluate overlapping community finding algorithms. http://arxiv.org/abs/1110.2515

[50] Zhang Q, Izquierdo E. (2006). Optimizing metrics combining low-level visual descriptors for image annotation and retrieval, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 405-408.

[51] Izakian Z, Mesgari M.S, Weibel R (2020) A feature extraction based trajectory segmentation approach based on multiple movement parameters. Engineering Applications of Artificial Intelligence, Volume 88.