

یک رویکرد جدید به منظور خوشه‌بندی سری‌های زمانی با استفاده از ترکیب زیرسری‌های زمانی

علی قربانیان (دانشجوی دکتری)

حمیده رضوی* (دانشیار)

گروه مهندسی صنایع، دانشکده مهندسی، دانشگاه فردوسی مشهد

مهندسی صنایع و مدیریت شریف (تابستان ۱۴۰۳)
دوری ۴۰، شماره ۱، ص. ۲۷-۴۱، (پژوهشی)

خوشه‌بندی سری‌های زمانی فرایندی است که سری‌های زمانی را با توجه به خصوصیات آن‌ها گروه‌بندی می‌کند. در پژوهش‌های پیشین به شباهت موجود بین قطعات یک سری زمانی به منظور خوشه‌بندی توجه کمتری شده است. در این مقاله یک رویکرد جدید دومرحله‌ای بر اساس قطعه‌بندی سری زمانی و خوشه‌بندی ترکیبی ارائه شده است. در مرحله اول یک مجموعه داده سری زمانی با استفاده از اندازه پنجره ثابت قطعه‌بندی شده و هر قطعه به طور جداگانه خوشه‌بندی شده است. سپس با استفاده از معیارهای درونی، بهترین نتایج حاصله انتخاب شده است. در مرحله دوم نتایج حاصل از مرحله اول با استفاده از خوشه‌بندی ترکیبی، پردازش شده و برچسب نهایی خوشه‌بندی حاصل شده است. نتایج الگوریتم ارائه شده نشان‌دهنده افزایش کارایی خوشه‌بندی به میزان ۲/۹۲ درصد و رسیدن به عدد ۶۷/۲۵ می‌باشد. همچنین بررسی عملکرد الگوریتم با بهترین نتایج ادبیات نیز نشان‌دهنده بهترین کارایی با حداقل هزینه زمانی می‌باشد.

واژگان کلیدی: سری‌های زمانی، خوشه‌بندی، قطعه‌بندی، خوشه‌بندی ترکیبی.

ali.ghorbanian@mail.um.ac.ir
h-razavi@um.ac.ir

۱. مقدمه و مرور ادبیات

می‌باشد. هدف از این کار شناخت الگوهای رفتاری مصرف‌کنندگان به منظور مدیریت مصرف متقاضیان در زمان‌های مختلف است. همچنین، به منظور کشف الگوهای مشابه در حوزه سلامت مانند همه‌گیری کوید - ۱۹ نیز از این تکنیک استفاده می‌گردد.^[۱] به صورت غیرمستقیم نیز خوشه‌بندی سری‌های زمانی به منظور پیش‌بینی این نوع داده مورد استفاده قرار می‌گیرد. در واقع تحقیقات نشان داده است که پیش‌بینی سری‌های زمانی در خوشه‌های یکسان می‌تواند نتایج بهتری را داشته باشد.

خوشه‌بندی سری زمانی شامل سه نوع خوشه‌بندی، خوشه‌بندی سری زمانی کل، خوشه‌بندی توالی و خوشه‌بندی نقطه‌ای می‌باشد. در خوشه‌بندی سری زمانی کل برخلاف دو نوع دیگر مجموعه از سری‌های زمانی با توجه به معیارهای مشابهت مانند داشتن کمترین فاصله از یکدیگر در گروه‌های مختلف قرار می‌گیرند. از روش‌های خوشه‌بندی سری‌های زمانی می‌توان روش‌های نیمه نظارت‌شده،^[۲] رویکردهای ترکیب و جنگل تصادفی،^[۳] ترکیب روش‌های خوشه‌بندی بر مبنای فاصله و چگالی،^[۴] استفاده از روش‌های خوشه‌بندی وزنی^[۵] و استفاده از شبکه‌های عصبی و رویکردهای یادگیری عمیق اشاره نمود.^[۶] در این تحقیق از روش خوشه‌بندی سری زمانی کل

خوشه‌بندی سری‌های زمانی در علوم مختلفی مانند ستاره‌شناسی، بیولوژی و آب‌وهوا استفاده می‌گردد. این نوع خوشه‌بندی مانند خوشه‌بندی سایر داده‌ها یک نوع یادگیری بدون نظارت می‌باشد که در آن اطلاعاتی از برچسب اشیا موجود نمی‌باشد. برچسب‌ها اعدادی هستند که از فرایند خوشه‌بندی حاصل گشته و نشان‌دهنده سری‌های زمانی قرار گرفته در یک خوشه می‌باشند. تعریف معیارهای فاصله‌ای مناسب برای سری‌های زمانی، موضوع پژوهش‌های متعددی بوده است. علت آن است که با توجه به ماهیت داده‌های سری زمانی، معیارهای فاصله‌ای ویژه‌ای مورد نیاز است. دقت خوشه‌بندی و زمان اجرای الگوریتم، دو چالش اصلی در خوشه‌بندی سری‌های زمانی می‌باشد. خوشه‌بندی سری‌های زمانی می‌تواند به صورت مستقیم و یا غیرمستقیم در صنایع و خدمات مورد استفاده قرار گیرد. یکی از کاربردهای اصلی خوشه‌بندی سری‌های زمانی کشف الگوهای رفتاری در مورد تقاضا مصرف به‌عنوان نمونه در حوزه انرژی الکتریکی، مصرف گاز طبیعی، مصرف آب و غیره

* نویسنده مسئول

تاریخ: دریافت ۱۴۰۳/۳/۲۵، اصلاحیه ۱۴۰۳/۸/۲۸، پذیرش ۱۴۰۳/۹/۲۱

استناد به این مقاله:

قربانیان، علی و رضوی، حمیده. ۱۴۰۳. یک رویکرد جدید به منظور خوشه‌بندی سری‌های زمانی با استفاده از ترکیب زیرسری‌های زمانی. مهندسی صنایع و مدیریت شریف، ۴۰(۲)،

صص. ۲۷-۴۱. DOI:10.24200/J65.2022.60405.2303

استفاده می‌شود که در آن یک مجموعه از سری‌های زمانی با توجه به معیارهای مشابهت خوشه‌بندی می‌گردد.^[۹]

یکی از راهکارهایی که به منظور افزایش دقت خوشه‌بندی مورد توجه قرار گرفته‌است، استفاده از معیارهای فاصله‌ای مخصوص سری‌های زمانی می‌باشد. معمولاً معیارهای فاصله‌ای متفاوتی برای سری‌های زمانی با توجه به ماهیت این نوع داده‌ها استفاده می‌گردد. تعدادی از رایج‌ترین معیارهای فاصله عبارت است از هاسدورف، همینگ، DTW^۱، اقلیدسی و LCSS^۲.^[۱۰]

در مطالعات قبلی دو رویکرد شامل تعریف، بهبود و ترکیب معیارهای فاصله مورد توجه قرار گرفته‌است.^[۱۱-۱۳] در رویکرد اول با معرفی معیار فاصله جدید و یا بهبود معیارهای موجود سعی در بهبود خوشه‌بندی داشته‌اند. اگرچه توسعه معیارهای فاصله‌ای جدید، بهبود و یا ترکیب آن‌ها می‌تواند تا حدودی دقت خوشه‌بندی این نوع داده‌ها را افزایش دهد اما هزینه زمانی (زمان اجرای الگوریتم) زیادی را نیز در بر دارد.

منظور از دقت خوشه‌بندی شاخص‌های بیرونی خوشه‌بندی مانند خلوص، شاخص رند^۳ و آنتروپی می‌باشد.^[۱۴] به‌عنوان مثال در پژوهشی که توسط رحیم خان و زکریا با ارائه یک الگوریتم مناسب به منظور معیار LCSS انجام شد تا حدود ۵۰ درصد زمان محاسبات کاهش یافت ولی هم‌زمان باعث کاهش دقت گردید.^[۱۵] به همین دلیل، سلیمانی و عابسی با تغییر این معیار به صورت فازی دقت خوشه‌بندی را بهبود بخشیده‌اند ولی در روش ایشان زمان افزایش یافته‌است.^[۱۶] همچنین کمال‌زاده و همکاران یک معیار فاصله جدید به منظور خوشه‌بندی سری‌های زمانی با طول بلند با استفاده از روابط هندسی تعریف کرده‌اند.^[۱۷] برای سری‌های زمانی با طول متفاوت، وانگ و همکاران از یک معیار فاصله بر مبنای اختلاف سطح زیر منحنی دو سری زمانی استفاده نموده‌اند.^[۱۶] در مطالعات دیگر در همین راستا با معرفی معیار WDTW^۴، MP^۵ و MSCD^۶ دقت خوشه‌بندی افزایش یافته‌است.^[۱۷-۲۰]

در رویکرد دوم به منظور استفاده از مزایای هر یک از معیارهای فاصله نظیر DTW، DDTW^۷ و LCSS از ترکیب این معیارها به منظور افزایش دقت خوشه‌بندی استفاده شده‌است.^[۲۱، ۲۲] منظور از این مزایا شناسایی نقاط پرت و همچنین شناسایی انتقال داده‌ها در طول زمان در یک سری زمانی می‌باشد. با وجود این که ترکیب معیارها دقت خوشه‌بندی را افزایش می‌دهد به دلیل افزایش حجم محاسبات، زمان بیشتری صرف می‌کند. به‌عنوان نمونه استفاده از معیار DDTW به منظور خوشه‌بندی ۸۴ مجموعه داده، زمانی نزدیک به ۸۰ ساعت برای هر مجموعه داده در بر داشته‌است در حالی که متوسط شاخص رند برای هر مجموعه داده نیز برابر با ۶۰ می‌باشد.^[۲۲] علیرغم زمان اجرای طولانی، مشاهده می‌گردد که دقت خوشه‌بندی نیز افزایش چشم‌گیری نداشته‌است بنابراین، در این پژوهش از فاصله اقلیدسی استفاده می‌گردد. به منظور برخورد با چالش هزینه زمانی نیز چندین روش مورد توجه قرار گرفته‌است. روش اول، استفاده از الگوریتم‌های چندمرحله‌ای به منظور کاهش حجم مجموعه داده و روش دوم استفاده از مشخصه‌های سری‌های زمانی به‌عنوان متغیرها برای خوشه‌بندی می‌باشد. به‌عنوان نمونه در روش اول آقابزرگی و همکاران، با استفاده از یک الگوریتم خوشه‌بندی در فاز اول توانستند زمان خوشه‌بندی را کاهش دهند.^[۲۳] همچنین وانگ و همکاران نیز با استفاده از یک شبکه وزن‌دار جهت در یک الگوریتم دوم‌مرحله‌ای، پیچیدگی مسئله خوشه‌بندی را کاهش داده‌اند.^[۲۴] ماناکوا و تاجنکو با به‌کارگیری معیار هایپکینز و سیستم نمونه‌گیری، یک الگوریتم دوم‌مرحله‌ای را توسعه داده‌اند.^[۲۵] در روش دوم، از مشخصه‌های سری زمانی به منظور خوشه‌بندی استفاده می‌شود. این مشخصه‌ها می‌تواند به صورت مستقیم از سری زمانی استخراج گردد. از جمله مشخصه‌های پرکاربرد می‌توان به معیارهای واریانس،

همبستگی مرتبه اول، خطی بودن، انحنای فصلی بودن، نقطه اوج و فرورفتگی اشاره نمود.^[۲۶] به‌عنوان مثال زو و همکاران به منظور استخراج مشخصه‌های سری زمانی از تبدیل سه نوع گراف بازگشتی، پدیداری و شبکه‌های انتقال استفاده نمودند.^[۲۷] داسیلوا با استخراج مشخصه‌های متوسط درجه، متوسط طول مسیر، تعداد اجتماعات، ضریب خوشه‌بندی و چگالی، از تبدیل سری زمانی به سه نوع گراف NVG، HVG و QG جهت خوشه‌بندی استفاده کرده‌است.^[۲۸] همچنین فریرا و ژاوبه به منظور خوشه‌بندی یک مجموعه داده از دو رویکرد NN-k و NN-ε به منظور تبدیل مجموعه داده به یک شبکه پیچیده بدون وزن استفاده نمودند.^[۲۹] استفاده از این رویکرد اگرچه موجب کاهش زمان محاسبات می‌گردد اما دقت خوشه‌بندی را کاهش می‌دهد.

مطالعات اخیر نشان داده‌است که قطعه‌بندی یک سری زمانی می‌تواند نقش مهمی در افزایش دقت خوشه‌بندی سری‌های زمانی داشته باشد. به‌عنوان مثال، گویجو و همکاران در مرحله اول یک سری زمانی را قطعه‌بندی نموده و در مرحله بعد با توجه به خصوصیات آماری هر یک از قطعه‌های ایجاد شده، مشخصه‌های آن‌ها را استخراج نموده‌اند. سپس از مشخصات استخراج شده به منظور خوشه‌بندی نهایی استفاده کرده‌اند.^[۳۰] همچنین یوناسینا و همکاران به منظور خوشه‌بندی از ترکیب رویکردهای تبدیل سری زمانی به شبکه و قطعه‌بندی، استفاده نموده‌اند.^[۳۱] عملکرد مناسب الگوریتم‌های ارائه شده بر مبنای قطعه‌بندی جایگاه آن را در مباحث خوشه‌بندی تثبیت نموده‌است.

در پژوهش‌های پیشین در بیشتر موارد به منظور بهبود کارایی خوشه‌بندی یک مجموعه داده سعی بر استفاده از معیارهای فاصله جدید و یا الگوریتم‌های جدید در حوزه یادگیری ماشین بوده‌است. اما نتایج ارائه شده نشان داده‌است که معمولاً معیارهای فاصله‌ای جدید مورد استفاده، می‌توانند بسیار زمان‌بر باشند و استفاده عملی از این نوع معیارهای فاصله کاربردی نمی‌باشد. همچنین نتایج حاصل از این پژوهش‌ها نشان داده‌است که استفاده از الگوریتم‌های خاص شاید بتواند در بعضی از مجموعه داده‌ها، نتایج ملموسی داشته باشد، اما این مورد همه مجموعه داده‌ها جامعیت ندارد. با توجه به پژوهش‌های انجام گرفته در حوزه داده‌کاوی می‌توان به نقش مهم خوشه‌بندی ترکیبی در افزایش کارایی خوشه‌بندی در داده‌های مختلف اشاره نمود، اما در هیچ‌کدام از پژوهش‌های انجام گرفته در حوزه سری‌های زمانی این تکنیک مورد توجه قرار نگرفته‌است. همچنین پژوهش‌های اخیر نشان داده‌است که زیرسری‌های ایجاد شده با استفاده از قطعه‌بندی یک سری زمانی می‌تواند دارای اهمیت فراوانی به منظور خوشه‌بندی یک مجموعه داده باشد. با توجه به موارد اشاره شده هدف از این پژوهش ارائه یک الگوریتم با کارایی بالا و سریع به منظور خوشه‌بندی یک مجموعه داده سری زمانی می‌باشد. در همین راستا در تحقیق فوق نشان داده‌ایم که لزوماً همه قطعات ایجاد شده در یک مجموعه داده سری زمانی نمی‌تواند نماینده خوبی به منظور استفاده در خوشه‌بندی نهایی باشد، برای رفع این مسئله نشان داده‌ایم که می‌توان ابتدا با استفاده از معیارهای درونی بهترین قطعات ایجاد شده را انتخاب نمود و با استفاده از بهترین قطعات ایجاد شده و خوشه‌بندی ترکیبی این قطعات، یک الگوریتم سریع و دقیق به منظور خوشه‌بندی داشته باشیم.

در این مقاله، با توجه به نقش مهم قطعه‌بندی سری‌های زمانی به منظور خوشه‌بندی، دو الگوریتم خوشه‌بندی بر مبنای قطعه‌بندی و خوشه‌بندی ترکیبی ارائه شده‌است. الگوریتم‌های ارائه شده دارای سه گام اصلی قطعه‌بندی، خوشه‌بندی و خوشه‌بندی ترکیبی می‌باشند. در الگوریتم اول به منظور خوشه‌بندی ترکیبی نهایی از یک معیار درونی به منظور انتخاب تعداد مشخص از نتایج به دست آمده استفاده می‌گردد، اما در الگوریتم دوم به‌طور هم‌زمان از سه معیار درونی برای این انتخاب استفاده می‌گردد.

رنگ سبز و آبی مشخص شده است، مشاهده می‌شود که قطعات شماره ۱ و ۴ خوشه‌بندی را به خوبی نمایش نمی‌دهند درحالی‌که قطعات شماره ۲ و ۳ خوشه‌ها به خوبی متمایز شده‌اند. با توجه به شکل اگر بتوان در خوشه‌بندی مجموعه داده از قطعاتی که خوشه‌بندی را به درستی نمایش داده‌اند استفاده نمود نتایج خوشه‌بندی کل مجموعه داده بهبود خواهد داشت، به همین سبب استفاده از خوشه‌بندی قطعات و ترکیب آن‌ها می‌تواند نتایج دقیق‌تری داشته باشد.

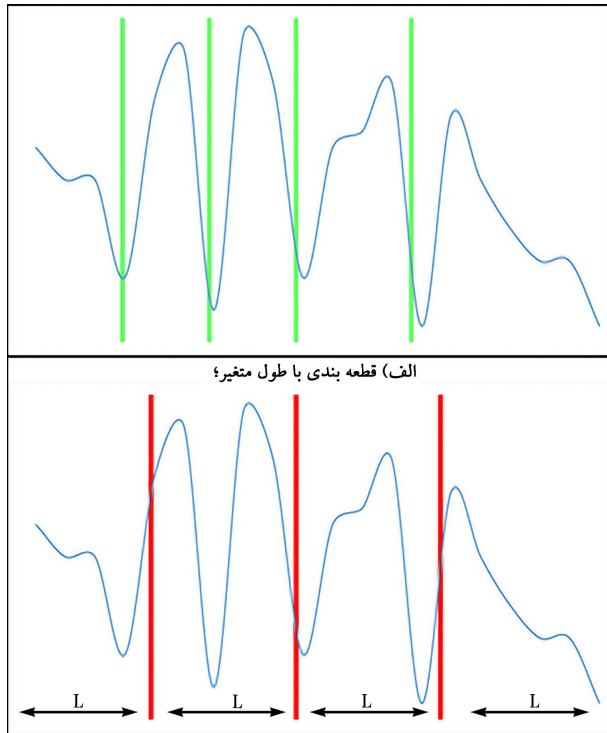
۱.۲. قطعه‌بندی

قطعه‌بندی، یکی از روش‌های تحلیل سری‌های زمانی می‌باشد که در آن به دو روش متفاوت می‌توان توالی از زیرسری‌ها را ایجاد نمود. منظور از توالی از زیرسری‌ها، سری‌های زمانی با ابعاد کوچک‌تر از سری زمانی اصلی می‌باشند که در کنار یکدیگر سری زمانی اصلی را تشکیل می‌دهند. به عنوان مثال در شکل ۱ چهار زیرسری تشکیل شده با استفاده از قطعه‌بندی در کنار یکدیگر سری زمانی اولیه را شکل می‌دهند.

در روش اول یک سری زمانی با استفاده از الگوریتم‌های مشخصی می‌تواند به زیرمجموعه‌هایی با طول‌های غیر یکسان تبدیل گردند. [۳۲، ۳۱] شکل ۳ الف) این نوع قطعه‌بندی را نمایش می‌دهد. در روش دوم که در این پژوهش نیز مورد استفاده قرار گرفته است ابتدا پنجره‌ای به طول L تعریف گردیده و سپس سری زمانی به زیرمجموعه‌هایی با طول‌های برابر با L تقسیم می‌گردند. [۳۳] شکل ۳ ب) این نوع قطعه‌بندی را نمایش می‌دهد.

۲.۲. خوشه‌بندی قطعات ایجاد شده

الگوریتمی که به منظور خوشه‌بندی قطعات ایجاد شده در گام اول مورد استفاده قرار



شکل ۳. انواع مختلف قطعه‌بندی سری زمانی.

شکل ۱ فلوجارت کلی روش پژوهش را نمایش می‌دهد. بعد از مقدمه و مرور ادبیات طرحی بانام خوشه‌بندی ترکیبی بر مبنای قطعه‌بندی ارائه می‌شود، سپس مفاهیم و الگوریتم‌های استفاده شده در طرح به صورت کامل توضیح داده می‌شود. در گام بعدی با توجه به طرح ارائه شده دو الگوریتم توسعه می‌یابد. در مرحله بعد الگوریتم‌های موردنظر روی چندین مجموعه داده اجرا می‌گردد و با توجه به معیارهای مختلف مانند کارایی و زمان بهترین الگوریتم انتخاب می‌گردد. به منظور بررسی کارایی الگوریتم انتخاب شده نتایج آن با ۴ الگوریتم موجود در ادبیات با استفاده از تست آماری مقایسه می‌گردد. در انتها نیز نتیجه‌گیری و مطالعات پیش رو بیان می‌گردد.

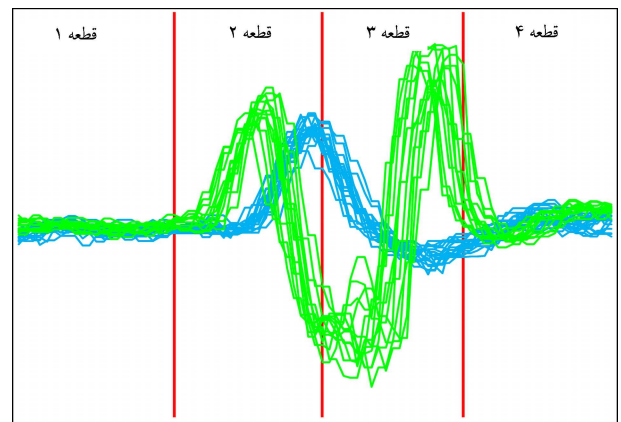
۲. خوشه‌بندی ترکیبی بر مبنای قطعه‌بندی

در مدل این پژوهش به جای این‌که یک مجموعه داده سری زمانی به صورت مستقیم خوشه‌بندی گردد، ابتدا این مجموعه داده به قطعات مساوی تبدیل شده و سپس هر قطعه ایجاد شده به صورت جداگانه خوشه‌بندی می‌گردد. در پایان نیز با استفاده از الگوریتم‌های خوشه‌بندی ترکیبی نتایج به دست آمده از قطعات مختلف با یکدیگر ترکیب می‌گردند. الگوریتم دارای سه گام اصلی شامل قطعه‌بندی با اندازه ثابت، خوشه‌بندی هر یک از قطعات ایجاد شده و ترکیب نتایج حاصل شده می‌باشد.

شکل ۲ یک مجموعه داده سری زمانی با دو خوشه را نمایش می‌دهد که با



شکل ۱. فلوجارت روش پژوهش.



شکل ۲. قطعه‌بندی یک سری زمانی.

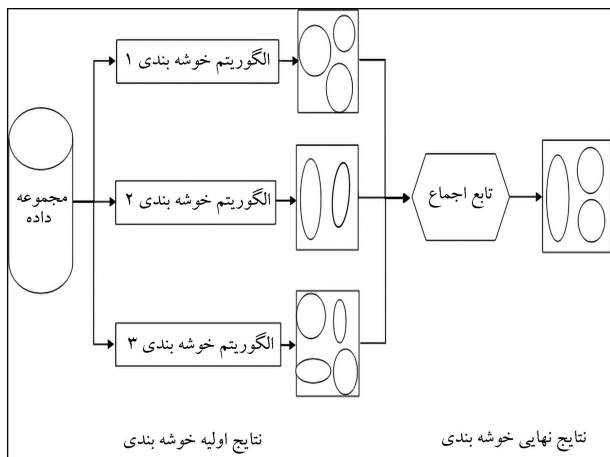
جدول ۱. معیارهای درونی ارزیابی خوشه‌بندی.

ردیف	معیار	نماد	رابطه ریاضی	مقدار مطلوب
۱	واریانس بین گروهی ^[۱۳]	V	$\frac{1}{n-k} \sum_{i=1}^k \sum_{x \in c_i} dist(x, c_i)$	کمینه
۲	دان ^[۲۳]	D	$\min_i \left\{ \min_j \left(\frac{\min_{x \in c_i, y \in c_j} (dist(x, y))}{\max_k \left\{ \max_{x, y \in c_k} (dist(x, y)) \right\}} \right) \right\}$	بیشینه
۳	کالینسکی - هاراباز ^[۴۴]	CH	$\frac{n-k}{n-1} \sum_{i=1}^k \sum_{x \in c_i} dist^2(x, c_i)$	بیشینه

^۱Inter - Group Variance

^۲Dunns Indices

^۳Calinski - Harabasez



شکل ۴. خوشه‌بندی ترکیبی.

گرفته است از نوع سلسله مراتبی تجمیعی می‌باشد. در این الگوریتم از دو نوع فاصله دورترین (d_{max}) و میانگین (d_{mean}) برای فاصله بین خوشه‌ها به منظور پیوند استفاده شده است. روابط شماره ۱ و ۲ به ترتیب این دو نوع فاصله را نمایش می‌دهند. A و B نمایش دهنده خوشه‌ها و a و b اعضا هر خوشه می‌باشند، همچنین d نشان دهنده فاصله بین دو عضو می‌باشد.^[۱۳]

$$d_{max} = \max \{d(a, b) : a \in A, b \in B\} \quad (۱)$$

$$d_{mean} = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \quad (۲)$$

به منظور محاسبه فاصله بین دو سری زمانی می‌توان از معیارهای فاصله‌ای متفاوتی از قبیل فاصله اقلیدسی، DTW و بزرگ‌ترین زیر دنباله مشترک (LCSS) استفاده نمود.^[۲۶-۲۴] با توجه به بالا بودن هزینه زمانی معیارهای DTW و LCSS در این پژوهش فاصله اقلیدسی استفاده شده است.^[۲۹-۲۷] اگر دو سری زمانی $X = \{x_t\}_{t=1}^n$ و $Y = \{y_t\}_{t=1}^n$ با طول N موجود باشد در این صورت فاصله اقلیدسی (ED) با رابطه ۳ قابل محاسبه می‌باشد. البته باید توجه داشت که این فاصله صرفاً هنگامی می‌تواند مورد استفاده قرار بگیرد که دو سری زمانی دارای طول یکسان باشند.^[۲۴]

$$ED(X, Y) = \sqrt{\sum_{t=1}^n (x_t - y_t)^2} \quad (۳)$$

۳.۲. خوشه‌بندی ترکیبی

خوشه‌بندی ترکیبی یکی از موضوعات مطرح شده در حوزه خوشه‌بندی می‌باشد که به ندرت در حوزه سری‌های زمانی مورد توجه قرار گرفته است. از آنجایی که همه روش‌های خوشه‌بندی لزوماً برای یک مجموعه داده دارای نتایج مطلوب نمی‌باشد، لذا می‌توان به طور هم‌زمان از چند روش خوشه‌بندی استفاده نمود. سپس با استفاده از تکنیک‌های مشخصی خوشه‌بندی نهایی را انتخاب نمود. شکل ۴ یک نمایش کلی از این رویکرد را نشان می‌دهد.^[۴۴]

خوشه‌بندی ترکیبی شامل دو قسمت تولید و ترکیب می‌باشد. در قسمت تولید می‌توان از دو رویکرد همگن و ناهمگن جهت تولید برچسب‌های متفاوت استفاده نمود. روش‌های ترکیب را می‌توان به چهار دسته اصلی شامل مستقیم، مبتنی بر خصوصیات، تشابهات زوجی و رویکردهای مبتنی بر گراف تقسیم‌بندی نمود. در این پژوهش به منظور تولید برچسب‌های مختلف از قطعه‌بندی و در بخش ترکیب نیز از یک الگوریتم مبتنی بر گراف با نام LWGP استفاده شده است. در این روش ابتدا فاصله بین خوشه‌های ایجاد شده برای هر شیء محاسبه و سپس از این فواصل و استفاده از یک الگوریتم خوشه‌بندی به منظور برچسب نهایی هر یک از اشیا استفاده می‌گردد.^[۴۱]

۴.۲. معیارهای درونی

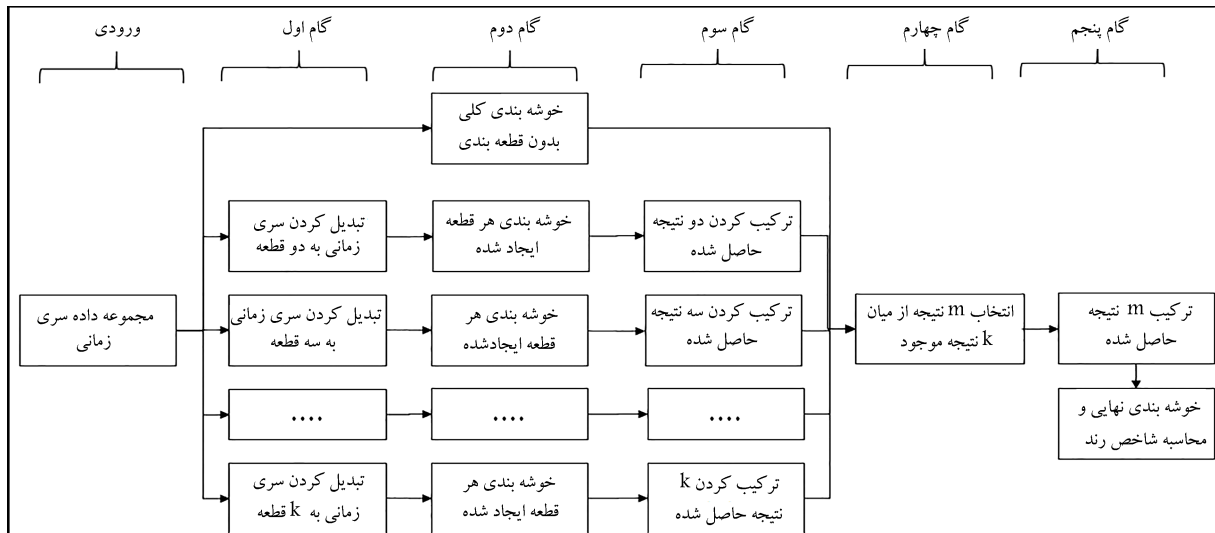
با توجه به اینکه در خوشه‌بندی از برچسب هر یک از اشیا اطلاعی در دست نیست، معیارهای درونی می‌توانند اهمیت ویژه‌ای داشته باشند. هدف از این معیارها، ارزیابی ساختار خوشه‌های ایجاد شده براساس شباهت درون خوشه‌ای و تمایز بین خوشه‌های می‌باشد. جدول ۱ سه معیار مورد استفاده در این پژوهش را نمایش می‌دهد. در این روابط n و k به ترتیب نشان‌دهنده تعداد اشیا و خوشه‌ها می‌باشند، همچنین x و y نیز اشیا خاص داخل خوشه c را نمایش می‌دهند.^[۴۲]

۵.۲. معیارهای بیرونی

معیارهای بیرونی با فرض مشخص بودن برچسب اشیا به منظور بررسی دقت الگوریتم مورد استفاده قرار می‌گیرند. در این مطالعه به منظور مقایسه با پژوهش‌های انجام‌گرفته از معیار بیرونی رند (RI) استفاده شده است که رابطه ۴ آن را نمایش می‌دهد.

$$RI = \frac{TP + TN}{TP + TN + FN + FP} \quad (۴)$$

در رابطه ۴، TP نشان‌دهنده تعداد اشیا می‌باشد که در کلاس و خوشه یکسانی قرار دارند. TN تعداد اشیا هستند که در خوشه‌های متفاوت قرار دارند و کلاس آن‌ها نیز متفاوت می‌باشد. FP نشان‌دهنده اشیا می‌باشد که دارای خوشه‌بندی متفاوت می‌باشند درحالی‌که کلاس آن‌ها یکسان باشد و سرانجام FN نشان‌دهنده تعداد اشیا می‌باشد که دارای خوشه یکسان و کلاس‌های متفاوت می‌باشند.



شکل ۵. الگوریتم اول پیشنهادی.

بتواند یک مقدار k مناسب را برای استفاده در الگوریتم ایجاد نماید. بنابراین، رابطه ۵ برای نحوه محاسبه مقدار k در این پژوهش به کار گرفته شده است. در این رابطه، L برابر با طول سری زمانی می باشد. با توجه به نتایج عددی به دست آمده برای پارامتر m ، مقادیر عددی ۲ یا ۳ مناسب به نظر می رسد.

$$k = \log_3 L \quad (5)$$

برای انتخاب m برچسب از k برچسب نهایی برای هر سری زمانی، از معیارهای درونی ارزیابی خوشه بندی می توان استفاده نمود. البته از آنجایی که این معیارها نیز متنوع می باشند، می توان از سه معیار واریانس بین گروهی، کالینسکی -- هارباز و شاخص دان که در ادبیات موضوع بیشتر مورد استفاده قرار گرفته اند، استفاده نمود. با توجه به اینکه استفاده از واریانس بین گروهی در الگوریتم اول برای نتیجه گیری کفایت می کند از معیارهای دیگر استفاده نشده است.

در الگوریتم اول صرفاً از یک معیار برای انتخاب m برچسب از k برچسب ایجاد شده، استفاده می گردد. الگوریتم اول پیشنهادی در شکل ۵ نمایش داده شده است.

۲.۶.۲. الگوریتم دوم توسعه داده شده

الگوریتم دوم نیز مانند الگوریتم اول دارای ۵ گام می باشد که سه گام نخست این الگوریتم نیز مانند الگوریتم اول می باشد. تفاوت این الگوریتم در گام سوم ایجاد می گردد که به نحوه انتخاب m نتیجه مرتبط می باشد. در گام چهارم این الگوریتم از سه معیار درونی معرفی شده در بخش ۱.۶.۲. به طور هم زمان به منظور انتخاب m نتیجه استفاده می گردد. در مرحله پنجم و نهایی نیز مطابق الگوریتم قبلی تعداد $3m$ برچسب نهایی ایجاد شده با استفاده از خوشه بندی ترکیبی با یکدیگر ترکیب می گردند. در مرحله نهایی در الگوریتم اول ما دارای m برچسب نهایی خواهیم بود اما در این الگوریتم تعداد برچسب ها $3m$ می باشد. باید توجه داشت که تعدادی از برچسب ها می تواند تکراری باشد. الگوریتم دوم در شکل ۶ نمایش داده شده است.

۳. اجرای مدل

به منظور بررسی کارایی الگوریتم های ارائه شده، این الگوریتم ها روی ۸۲ مجموعه داده

۶.۲. الگوریتم های پیشنهادی

در این پژوهش دو الگوریتم بر پایه قطعه بندی با اندازه ثابت و خوشه بندی ترکیبی به منظور خوشه بندی سری های زمانی ارائه شده است.

۱.۶.۲. الگوریتم اول توسعه داده شده

این الگوریتم دارای ۵ گام به شرح زیر می باشد.

گام اول. یک مجموعه داده سری زمانی به قطعات با طول مساوی تقسیم می گردد. تعداد قطعات بین ۲ تا k قطعه می باشد.

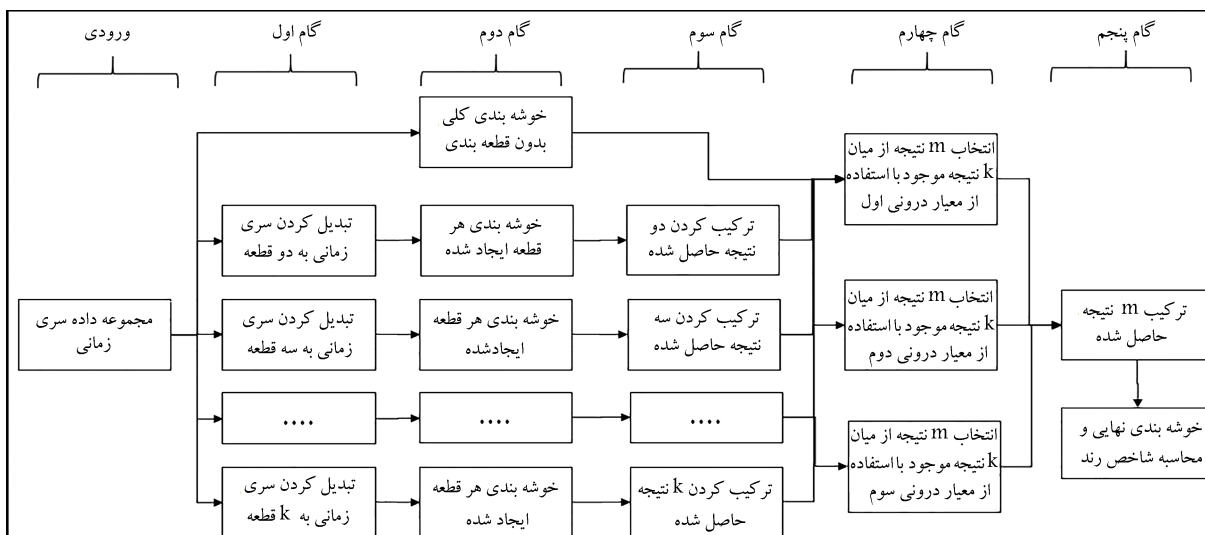
گام دوم. برای هر یک از قطعات ایجاد شده و برای هر مقدار k با استفاده از الگوریتم خوشه بندی سلسله مراتبی، خوشه بندی انجام می گیرد. در این گام با توجه به اینکه مجموعه داده سری زمانی به ۲ تا k قطعه تبدیل شده است، به ترتیب تعداد ۲ تا k برچسب خوشه بندی نیز وجود خواهد داشت.

گام سوم. در این گام نتایج خوشه بندی موجود با استفاده از یک الگوریتم خوشه بندی ترکیبی با یکدیگر ترکیب می گردد تا یک نتیجه واحد برای هر مرحله قطعه بندی از ۲ تا k قطعه ایجاد گردد. در پایان این گام، تعداد $k - 1$ برچسب خوشه بندی برای مجموعه داده سری زمانی وجود خواهد داشت. در انتهای این گام نتایج خوشه بندی مجموعه داده بدون قطعه بندی نیز به این تعداد اضافه می گردد و تعداد برچسب های موجود برابر با k خواهد بود.

گام چهارم. در ادامه با استفاده از یک معیار داخلی مشخص تعداد m برچسب از میان k برچسب ایجاد شده، انتخاب می شوند ($m \leq k$).

گام پنجم. در گام نهایی m برچسب حاصل شده مجدداً با استفاده از یک الگوریتم خوشه بندی ترکیبی با یکدیگر ترکیب می گردند و یک برچسب خوشه بندی نهایی حاصل می گردد.

انتخاب مقدار k و m می تواند تا حدودی پیچیده باشد اما به نظر می رسد می توان مقدار k را طوری انتخاب نمود که برای سری های زمانی طولانی، بزرگ نباشد و همچنین برای سری های زمانی با طول کوتاه نیز بتواند ساختار سری زمانی را برای قطعه ها حفظ نماید. با توجه به این موارد به نظر می رسد استفاده از یک رابطه لگاریتمی



شکل ۶. الگوریتم دوم پیشنهادی.

جدول ۲. پارامترهای الگوریتم‌های مختلف خوشه‌بندی.

مقدار پارامتر m	معیار درونی	تعداد معیار درونی	نوع پیوند	الگوریتم خوشه‌بندی	نوع
۱	واریانس بین گروهی	۱	میانگین	سلسله‌مراتبی	وضعیت اول
۳	واریانس بین گروهی	۱	میانگین	سلسله‌مراتبی	وضعیت دوم
۳	واریانس بین گروهی	۱	حداکثر	سلسله‌مراتبی	وضعیت سوم
۲	واریانس بین گروهی، کالیسنسکی - هاربا سو و دان	۳	حداکثر	سلسله‌مراتبی	الگوریتم دوم

First Algorithm	Second Algorithm
Input: Time series dataset, m, k , internal measure Output: Best quality clustering, RI Clustering dataset without segmentation for each k do Apply time series segmentation for each segment do Cluster the segment end for Ensemble clustering all clustered segments end for Select m results from k results using an internal measure Ensemble clustering m selected results Evaluate the goodness of the clustering Return Best quality clustering, RI	Input: Time series dataset, m, k , internal measure Output: Best quality clustering, RI Clustering dataset without segmentation for each k do Apply time series segmentation for each segment do Cluster the segment end for Ensemble clustering all clustered segments end for for 3 internal measure do Select m results from k results using an internal measure end for Ensemble clustering $3m$ Selected results Evaluate the goodness of the clustering Return Best quality clustering, RI

شکل ۷. شبه کد الگوریتم اول و دوم.

گزارش شده است. مجموعه داده‌ها با سایز بزرگ، زمان‌های اجرای طولانی در مقایسه با سایر مجموعه داده‌ها دارند.

جدول ۴ به صورت مختصر و با توجه به معیارهای مختلف عملکرد دو الگوریتم ارائه شده را نمایش می‌دهد. منظور از الگوریتم اولیه در این جدول، خوشه‌بندی بدون قطعه‌بندی با استفاده از الگوریتم سلسله‌مراتبی می‌باشد.

با توجه به اطلاعات جدول ۴ می‌توان مشاهده نمود که در الگوریتم اول وضعیت اول شاخص رند ۵/۴۴ درصد نسبت به الگوریتم اولیه بهبود یافته‌است، همین‌طور

از آرشیو UCR اجرا شده‌است.^[۲۵] مجموعه داده مورد نظر زمینه‌های مختلفی مانند سلامت، مالی، رسانه و مهندسی را پوشش می‌دهد. با توجه به اینکه مجموعه داده‌های مورد نظر برای کلاس‌بندی سری‌های زمانی ارائه شده‌اند و دارای دو بخش آموزش و آزمون می‌باشند، برای خوشه‌بندی با یکدیگر جمع شده‌اند. پارامترهای ورودی برای هر الگوریتم شامل الگوریتم خوشه‌بندی استفاده شده، نوع پیوند، تعداد معیارهای درونی استفاده شده، معیار درونی استفاده شده، m و k می‌باشد. همچنین خروجی مدل نیز شامل سری‌های زمانی قرار گرفته در هر خوشه و همچنین شاخص رند نهایی می‌باشد.

شبه کد الگوریتم اول و دوم توسعه داده شده شامل ورودی‌ها، خروجی‌ها و روند اجرای هر الگوریتم توسط شکل ۷ نمایش داده شده‌است.

برای الگوریتم اول سه وضعیت و برای الگوریتم دوم یک وضعیت با توجه به پارامترهای مختلف در نظر گرفته شده‌است. جدول ۲ پارامترهای استفاده شده برای سه وضعیت الگوریتم اول و الگوریتم دوم نمایش می‌دهد. در تمام وضعیت‌ها به منظور خوشه‌بندی قطعات از الگوریتم سلسله‌مراتبی استفاده شده‌است.

۱.۳. نتایج

به منظور بررسی الگوریتم‌های ارائه شده نتایج حاصل روی ۸۲ مجموعه داده معرفی شده مورد مقایسه قرار گرفته‌است. الگوریتم‌ها برای هر یک از مجموعه داده‌ها ۱۰ بار اجرا شده‌است و نتایج شاخص رند و زمان اجرا برای میانگین ۱۰ تکرار در جدول ۳

جدول ۳. نتایج شاخص رند و زمان اجرا برای دو الگوریتم ارائه شده.

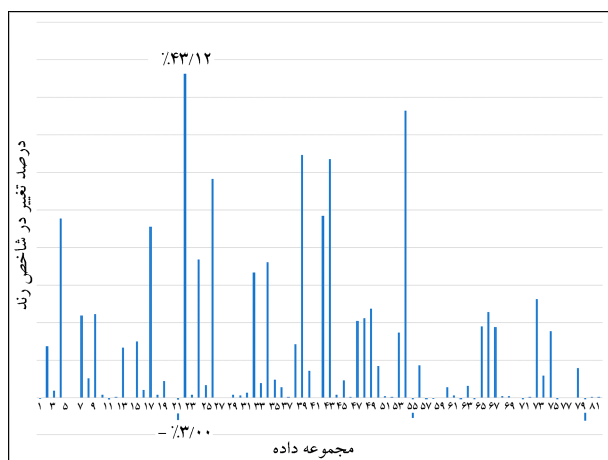
الگوریتم	زمان اجرا (ثانیه)			شاخص رند (%)				مجموعه داده
	الگوریتم اول			الگوریتم اول				
	دوم	وضعیت سوم	وضعیت دوم	وضعیت اول	دوم	وضعیت سوم	وضعیت دوم	
۷,۴	۸,۳۴	۴,۱۳	۳,۳۷	۵۰,۸۰	۵۰,۱۰	۵۱,۳۰	۵۰,۰۰	ITA
۵,۶۲	۳,۵۷	۲,۰۸	۱,۹۴	۸۲,۰۰	۸۱,۳۰	۷۹,۰۰	۸۰,۵۰	SYN
۹,۹	۱۰,۵۹	۶,۲۲	۴,۹۵	۶۳,۸۰	۶۷,۵۰	۵۴,۳۰	۵۴,۳۰	SO۲
۴,۶۶	۳,۹۶	۲,۶۴	۲,۱۷	۵۰,۱۰	۵۲,۶۰	۵۰,۱۰	۴۹,۹۰	SO۱
۳,۶۴	۲,۸۱	۱,۸۳	۱,۷۱	۷۲,۹۰	۷۳,۷۰	۷۱,۲۰	۷۱,۲۰	DPA
۷,۱۳	۷,۸۶	۴,۹۲	۴,۳۱	۵۳,۵۰	۵۲,۴۰	۵۳,۰۰	۵۳,۰۰	DPC
۳,۳۲	۳,۱۸	۱,۸۷	۱,۵۸	۸۷,۸۰	۸۶,۳۰	۸۱,۵۰	۸۴,۳۰	DPT
۴,۹۲	۳,۱	۲,۳۳	۱,۷۴	۷۳,۰۰	۷۲,۸۰	۷۰,۵۰	۷۰,۴۰	MPA
۸,۷۳	۷,۴۹	۵,۴۱	۴,۴۱	۵۰,۰۰	۵۰,۰۰	۵۳,۲۰	۵۳,۳۰	MPC
۴,۸۷	۲,۹۸	۱,۷۲	۱,۵	۸۳,۸۰	۸۳,۹۰	۸۱,۸۰	۸۲,۳۰	MPT
۷۷,۶۲	۸۴,۶۹	۴۹,۴۵	۴۵,۱۶	۵۴,۲۰	۵۲,۱۰	۵۴,۰۰	۵۴,۰۰	PHA
۵,۷۷	۳,۵۷	۲,۱۹	۲,۲۶	۷۸,۹۰	۷۹,۷۰	۷۷,۳۰	۷۷,۴۰	PPA
۸,۱۹	۷,۳۷	۴,۹۶	۴,۳۷	۵۲,۱۰	۵۷,۶۰	۵۳,۸۰	۵۳,۹۰	PPC
۵,۵۲	۳,۷۵	۲,۰۵	۱,۸۶	۷۸,۶۰	۷۹,۸۰	۸۳,۸۰	۸۳,۹۰	PPT
۱۴,۰۳	۱۴,۴۴	۸,۰۴	۷,۰۵	۵۰,۳۰	۵۰,۱۰	۵۰,۰۰	۵۰,۱۰	TWE
۱۷,۴۸	۲۰,۹۹	۱۰,۸۲	۹,۷۱	۶۱,۸۰	۶۳,۱۰	۵۰,۳۰	۵۰,۳۰	MOT
۱,۰۵	۱,۱۳	۰,۷۵	۰,۶۶	۶۲,۳۰	۶۱,۲۰	۵۴,۰۰	۶۲,۳۰	EC۲
۲۱,۵۵	۱۵,۵	۱۰,۸۷	۱۰,۱۱	۶۳,۴۰	۶۴,۶۰	۶۱,۳۰	۶۰,۲۰	MED
۱۰,۹۴	۱۰,۵۵	۶,۹	۶,۲۵	۶۶,۰۰	۶۴,۷۰	۶۳,۱۰	۶۰,۶۰	CBF
۲۳,۴۲	۱۸,۶۷	۱۳,۶	۱۲,۰۶	۸۷,۲۰	۸۸,۷۰	۳۷,۹۰	۳۹,۰۰	SWE
۳۶,۱,۴۶	۳۸۳,۸۵	۲۵۸,۳۶	۲۴۰,۷۳	۶۲,۸۰	۶۲,۹۰	۶۲,۱۰	۶۲,۶۰	TWP
۶۳,۱	۶۲,۱	۷۵,۵	۳۹,۶۳	۸۳,۳۰	۸۵,۰۰	۸۱,۶۰	۷۳,۷۰	FAA
۱۰,۵۵	۱۰,۷۷	۷,۹۵	۶,۵۳	۵۰,۰۰	۵۰,۵۰	۵۰,۳۰	۴۹,۹۰	ECF
۳۲۸,۳۳	۴۶۳,۸۳	۲۷۸,۷۱	۲۵۰,۰۶	۷۴,۱۰	۷۴,۶۰	۹۰,۶۰	۸۹,۶۰	EC۵
۲,۹۹	۱,۸۷	۱,۰۵	۱,۰۵	۹۴,۰۰	۹۴,۰۰	۹۲,۵۰	۹۲,۸۰	PLA
۱,۶	۱,۲۷	۱,۹۸	۰,۶۶	۵۳,۷۰	۴۹,۷۰	۵۰,۹۰	۴۹,۷۰	GUN
۱۵۷۱,۷۶	۱۳۹۵,۹۴	۶۱۸,۹۸	۵۸۰,۶۹	۵۳,۴۰	۵۳,۴۰	۵۳,۴۰	۵۳,۴۰	WAF
۲۸۱,۷۲	۲۹۰,۳۴	۲۰۸,۳۱	۲۱۰,۲۹	۵۳,۱۰	۵۲,۲۰	۵۱,۵۰	۵۰,۵۰	CHL
۱۱,۵۶	۱۱,۸۲	۸,۲۵	۷,۵۷	۹۲,۲۰	۹۲,۸۰	۴۸,۷۰	۴۷,۵۰	ADI
۹۹	۹۸	۵۵	۵۰	۶۶,۸۷	۶۷,۱۵	۶۲,۵۲	۶۲,۴۳	کوتاه
۵,۳۷	۳,۸۲	۲,۵۲	۰,۹۳	۴۹,۹۰	۴۹,۸۰	۴۹,۹۰	۴۹,۹۰	WIN
۲۴,۰۱	۲۲,۹۱	۱۵,۱۳	۱۳,۶۸	۵۰,۴۰	۵۰,۴۰	۵۱,۵۰	۵۱,۵۰	STR
۱,۸۹	۱,۸۷	۱,۵۷	۱,۱۷	۶۰,۵۰	۶۶,۸۰	۳۵,۳۰	۳۵,۸۰	ARR
۱۱۸,۸۷	۹۶,۶	۶۱,۴	۶۰,۳۲	۸۶,۸۰	۸۶,۰۰	۸۶,۸۰	۸۶,۵۰	INS
۳۲,۸۵	۲۲,۶۳	۹,۴۲	۱۰,۲۳	۹۴,۵۰	۹۴,۶۰	۹۴,۴۰	۹۴,۶۰	W۵۰
۵۴,۰۸	۱۹,۹۲	۱۴,۰۳	۱۱,۸۶	۸۸,۶۰	۸۸,۷۰	۸۷,۶۰	۸۳,۹۰	WOS
۳,۲۴	۱,۹۲	۳,۱۶	۱,۱۸	۷۵,۱۰	۷۵,۰۰	۷۴,۹۰	۷۵,۰۰	TRA
۳,۰۶	۲,۳۱	۱,۵۳	۱,۳۶	۵۰,۰۰	۴۹,۹۰	۵۰,۱۰	۴۹,۹۰	TO۱
۰,۸۶	۰,۹۶	۱,۰۷	۰,۶۵	۷۹,۵۰	۷۰,۱۰	۴۹,۱۰	۴۹,۸۰	COF
۱۴,۷۹	۱۳,۲۹	۹,۴	۹,۴۸	۸۳,۴۰	۸۱,۶۰	۷۷,۹۰	۷۱,۲۰	CRX
۱۷,۲۶	۱۳,۱۵	۹,۹۲	۹,۳۲	۸۴,۰۰	۸۲,۴۰	۷۷,۴۰	۷۷,۴۰	CRY
۱۵,۷	۱۳,۴۸	۹,۳۵	۸,۴۸	۸۴,۳۰	۸۴,۶۰	۷۷,۱۰	۶۹,۶۰	CRZ
۵۰,۱,۱۵	۴۷۱,۹۶	۳۳۶,۹۳	۳۲۶,۹۸	۸۱,۴۰	۸۱,۹۰	۷۳,۶۰	۷۱,۵۰	UWX
۵۰,۹,۲۳	۴۴۷,۹۶	۳۲۳,۱۷	۳۰۳,۸۷	۸۲,۱۰	۸۲,۲۰	۷۹,۸۰	۸۰,۵۰	UWY

ادامه جدول ۳. نتایج شاخص رند و زمان اجرا برای دو الگوریتم ارائه شده.

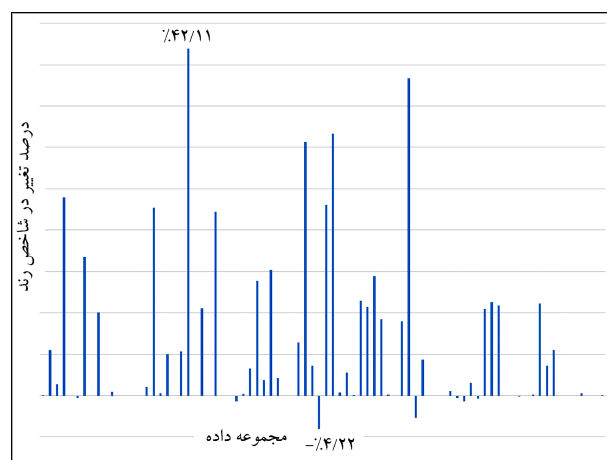
الگوریتم	زمان اجرا (ثانیه)			شاخص رند (%)				مجموعه داده
	الگوریتم اول			الگوریتم دوم	الگوریتم اول			
	وضعیت سوم	وضعیت دوم	وضعیت اول		وضعیت سوم	وضعیت دوم	وضعیت اول	
۵۸۹,۵۳	۴۶۰,۵۵	۳۳۰,۶۵	۳۰۹,۹۶	۸۱,۹۰	۸۱,۹۰	۸۳,۳۰	۷۹,۹۰	UWZ
۲,۷۸	۲,۱۴	۱,۳۵	۱,۱۵	۷۵,۶۰	۷۱,۴۰	۷۵,۸۰	۷۴,۱۰	LIV
۲,۰۲	۱,۵۹	۱,۰۵	۰,۹۲	۵۱,۱۰	۴۹,۷۰	۴۹,۹۰	۴۹,۷۰	TO۲
۴,۱۵	۳,۱۹	۲,۵۷	۲,۳۹	۳۰,۶۰	۲۹,۶۰	۳۰,۶۰	۳۰,۶۰	DIA
۱,۶	۱,۴۴	۱,۵	۱,۰۲	۶۹,۵۰	۶۹,۲۰	۷۳,۸۰	۷۲,۱۰	FAF
۴۰,۴۵	۲۷,۴۷	۱۸,۸۶	۱۹,۶۳	۸۹,۲۰	۸۹,۶۰	۸۹,۹۰	۸۹,۹۰	SYM
۴۸۱,۲۲	۴۲۰,۳۷	۲۹۵,۰۱	۲۷۲,۷۷	۵۰,۰۰	۵۰,۱۰	۵۰,۰۰	۵۰,۰۰	YOG
۱۲,۰۴	۶,۸۷	۵,۳۵	۴,۸۹	۷۳,۳۰	۷۱,۴۰	۶۲,۷۰	۶۷,۸۰	OSU
۷,۱۳	۲,۴۹	۲,۰۵	۱,۶۸	۵۱,۴۰	۵۱,۵۰	۵۰,۲۰	۵۰,۱۰	HAM
۲,۶۲	۱,۶۶	۱,۲۳	۱,۱۲	۷۲,۰۰	۷۲,۰۰	۷۷,۰۰	۷۲,۸۰	MEA
۶,۴۴	۵,۴۸	۴,۲۹	۳,۷۸	۷۱,۴۰	۷۲,۰۰	۷۱,۶۰	۶۴,۷۰	FIS
۱,۵۲	۳,۵۹	۱,۱۵	۱,۰۴	۶۳,۸۰	۶۳,۸۰	۶۵,۵۰	۶۵,۶۰	BEE
۱۱۱۹,۴۹	۹۹۲,۹۸	۱۰۴۸,۸۶	۷۹۶,۰۲	۵۰,۰۰	۵۰,۰۰	۵۰,۲۰	۵۰,۲۰	FOA
۹۵۸,۵۸	۸۰۷,۴۳	۶۹۴,۹۴	۶۵۹,۷۴	۵۰,۰۰	۵۰,۰۰	۵۰,۰۰	۵۰,۰۰	FOB
۳,۱۴	۲,۵	۲,۰۷	۱,۶۷	۴۹,۸۰	۵۰,۰۰	۵۱,۴۰	۵۱,۴۰	SHS
۱۵۶	۱۳۴	۱۱۱	۹۸	۶۸,۲۸	۶۷,۸۰	۶۵,۴۲	۶۴,۳۴	متوسط
۱,۱۶	۱,۲۷	۰,۸۲	۰,۸۷	۵۰,۹۰	۵۱,۷۰	۵۶,۹۰	۵۵,۴۰	BFL
۱,۲	۱,۵۸	۰,۷۷	۰,۷۶	۵۳,۳۰	۴۸,۸۰	۴۹,۲۰	۴۸,۹۰	BIR
۱۰,۰۶	۹,۶	۷,۵۶	۶,۹۴	۵۲,۳۰	۵۲,۴۰	۵۶,۸۰	۵۶,۸۰	EAR
۱,۹۷	۱,۶۳	۱,۱۹	۱,۴۴	۵۰,۵۰	۵۰,۶۰	۵۱,۷۰	۵۱,۷۰	HER
۶۸,۸	۵۳,۳	۳۶,۹	۳۶,۲۱	۹۴,۸۰	۹۶,۱۰	۸۴,۴۰	۸۳,۵۰	SHA
۲,۳۶	۱,۶۴	۱,۰۶	۱,۱۴	۸۵,۸۰	۸۶,۰۰	۸۶,۱۰	۸۸,۷۰	OLI
۱,۸۵	۲,۳۳	۱,۴	۱,۲۶	۶۱,۳۰	۶۳,۳۰	۵۹,۵۰	۶۵,۳۰	CAR
۲,۱۸	۱,۷۷	۱,۳۷	۱,۲۳	۶۰,۰۰	۵۰,۴۰	۵۸,۵۰	۵۷,۹۰	LI۲
۱۴,۲	۱۲,۴۷	۱۰,۴۴	۱۰,۵۲	۵۰,۴۰	۵۰,۳۰	۴۹,۹۰	۴۹,۹۰	COM
۳۸,۶۹	۲۷,۳۷	۲۲,۰۱	۲۱,۱۹	۵۲,۲۰	۵۰,۹۰	۵۴,۲۰	۵۴,۱۰	LAR
۳۵,۷۸	۲۷,۱۸	۲۲,۸۴	۲۱,۶	۵۵,۲۰	۵۴,۹۰	۴۵,۵۰	۴۵,۸۰	REF
۳۷,۳۵	۲۹,۱۶	۲۲,۶۴	۲۲,۸	۵۵,۲۰	۵۵,۰۰	۴۰,۲۰	۴۰,۳۰	SCR
۳۸,۹	۳۱,۰۱	۲۳,۲۷	۲۲,۶۲	۴۳,۰۰	۴۶,۶۰	۳۴,۳۰	۳۳,۸۰	SMA
۹۹۲,۱۸	۶۳۰,۷۴	۵۵۹,۳۷	۵۳۲,۸۱	۹۴,۶۰	۹۵,۱۰	۸۵,۴۰	۸۶,۷۰	NO۱
۹۸۰,۵۲	۶۴۴,۱۹	۴۹۱,۲۶	۵۰۴,۸۸	۹۵,۶۰	۹۵,۹۰	۹۱,۱۰	۹۱,۳۰	NO۲
۱۲,۹	۶,۰۸	۵,۱۵	۵,۷۱	۶۳,۵۰	۶۳,۰۰	۵۹,۶۰	۶۲,۷۰	WOR
۸,۷	۵,۹۸	۴,۱۱	۴,۷۹	۴۹,۸۰	۴۹,۳۰	۴۹,۹۰	۴۹,۷۰	WOT
۱۷۵۳,۹۲	۱۱۸۰,۰۲	۹۰۷,۴۹	۸۶۵,۹۸	۸۵,۶۰	۸۶,۳۰	۸۵,۹۰	۸۷,۶۰	UWA
۵۰۹,۰۶	۳۳۶,۲۱	۲۵۳,۱۷	۲۴۰	۹۳,۹۰	۹۲,۲۰	۹۰,۲۰	۹۰,۳۰	MAL
۴۳۲,۷۵	۳۳۴,۱۲	۲۷۷,۲۸	۲۸۱,۷	۹۲,۶۰	۹۲,۰۰	۹۰,۱۰	۹۰,۱۰	PHO
۲۵,۶۴	۱۶,۵۱	۱۴,۸۷	۱۵,۷۹	۶۷,۱۰	۶۷,۱۰	۳۹,۴۰	۳۶,۶۰	HAP
۲۶۲,۵۶	۱۹۰,۵۷	۱۷۰,۳۹	۱۶۷,۹۳	۶۳,۸۰	۶۶,۱۰	۶۱,۸۰	۶۱,۴۰	CIN
۸۱,۸۸	۵۳,۸۵	۴۳,۱۸	۴۱,۹۷	۶۸,۰۰	۶۶,۳۰	۶۴,۶۰	۶۱,۷۰	INL
۵۵۲,۰۱	۴۳۹,۸۲	۳۶۳,۶۹	۳۴۴,۶	۵۶,۱۰	۵۰,۰۰	۵۴,۵۰	۵۷,۱۰	HAN
۲۴۴	۱۶۸	۱۳۵	۱۳۱	۶۶,۴۸	۶۵,۸۵	۶۲,۴۹	۶۲,۸۰	بلند
۱۶۱	۱۳۱,۱۳	۹۸,۲۱	۹۰,۹۳	۶۷,۲	۶۷,۰۰	۶۳,۵۴	۶۳,۲۲	متوسط

جدول ۴. نتایج الگوریتم‌های مختلف خوشه‌بندی.

الگوریتم دوم	الگوریتم اول			معیار
	وضعیت سوم	وضعیت دوم	وضعیت اول	
۶۷٫۲۵	۶۷٫۰۰	۶۳٫۵۴	۶۳٫۲۲	میانگین شاخص رند (%)
۰٫۰۰۰۶۵	۰٫۰۰۰۹۱	۰٫۰۱۰۴	۰٫۰۰۰۵۸	انحراف از معیار
۲٫۹۲	۲٫۶۹	۵٫۷۴	۵٫۴۴	میانگین بهبود شاخص رند نسبت به حالت اولیه (%)
۳۷٫۱۳	۳۷٫۹۶	۴۳٫۱۲	۴۲٫۱۱	حداکثر مقدار بهبود در یک مجموعه داده (%)
-۱۰٫۷۸	-۱۰٫۲۷	-۳٫۰۰	-۴٫۲۲	حداکثر میزان تنزل شاخص رند در یک مجموعه داده (%)
۵۹	۵۵	۷۳	۵۹	نسبت مجموعه داده با بهبود شاخص رند (%)
۳۰	۱۰	۱۰	۳۲	نسبت مجموعه داده بدون تغییر شاخص رند (%)
۱۱	۳۵	۱۷	۹	نسبت مجموعه داده با افت شاخص رند (%)
۳۲	۳۰	۲۰	۱۸	نسبت ایجاد بهترین پاسخ در پژوهش‌های قبلی (%)



شکل ۹. اختلاف میانگین شاخص رند الگوریتم اول وضعیت دوم.



شکل ۸. اختلاف میانگین شاخص رند الگوریتم اول وضعیت اول.

وضعیت سوم الگوریتم اول با پارامتر $m = 3$ و فاصله حداکثر برای الگوریتم خوشه‌بندی سلسله‌مراتبی اجرا شده‌است. با توجه به اطلاعات جدول ۴ می‌توان مشاهده نمود که میانگین شاخص رند در ۸۲ مجموعه داده به ۶۷ درصد رسیده‌است و نتایج اولیه خوشه‌بندی در این وضعیت ۲٫۶۹ درصد به‌طور میانگین افزایش داشته‌است. همچنین در ۳۰ درصد حالات، وضعیت سوم الگوریتم اول توانسته است بهترین شاخص رند را در بین ادبیات موضوع داشته‌باشد. براساس اطلاعات شکل ۱۰، وضعیت سوم الگوریتم اول در بهترین و بدترین حالت به‌ترتیب توانسته است ۳۸ درصد نتیجه اولیه را بهبود بخشد و ۱۰ درصد نتیجه اولیه را ضعیف‌تر نماید.

در انتها الگوریتم دوم با پارامتر $m = 2$ و فاصله حداکثر برای الگوریتم خوشه‌بندی سلسله‌مراتبی اجرا شده‌است. در الگوریتم دوم همان‌گونه که قبلاً بیان شده‌است از سه معیار داخلی به‌صورت هم‌زمان استفاده شده‌است. با توجه به اطلاعات جدول ۴ الگوریتم دوم توانسته است ۲٫۹ درصد به‌صورت میانگین دقت خوشه‌بندی را افزایش دهد و دقت شاخص رند را به ۶۷٫۲۵ درصد برساند که بهترین نتیجه ممکن می‌باشد. در الگوریتم دوم نیز نتایج مربوط به ۶۰ درصد مجموعه داده‌ها به‌طور میانگین بهبود داشته‌است و در ۳۰ درصد موارد نیز تغییر نداشته‌است. همین‌طور در ۳۲ درصد حالات، الگوریتم دوم توانسته است بهترین شاخص رند را در بین ادبیات موضوع داشته‌باشد. براساس اطلاعات شکل ۱۱ الگوریتم دوم در بهترین

در ۳۰ درصد حالات، الگوریتم مورد اشاره توانسته‌است بهترین شاخص رند را در بین ادبیات موضوع داشته‌باشد. همچنین الگوریتم اول وضعیت اول توانسته‌است در ۵۹ درصد مجموعه داده‌ها دقت خوشه‌بندی را افزایش دهد و در ۳۲ درصد نیز دقت خوشه‌بندی تغییری نکرده‌است. با توجه به شکل ۸ در بهترین حالت، الگوریتم اول وضعیت اول می‌تواند تا ۴۲ درصد شاخص رند را برای یک مجموعه داده بهبود دهد و از سوی دیگر می‌تواند نتایج اولیه خوشه‌بندی را تا ۴ درصد با افت روبرو نماید.

وضعیت دوم الگوریتم اول با مقدار پارامتر $m = 3$ اجرا شده است و مانند حالت قبل برای الگوریتم خوشه‌بندی سلسله‌مراتبی، فاصله برابر با میانگین در نظر گرفته شده‌است. با توجه به اطلاعات جدول ۴ می‌توان مشاهده نمود که وضعیت دوم الگوریتم اول میانگین شاخص رند برابر با ۶۳٫۵۴ درصد می‌باشد و الگوریتم توانسته است دقت را در حدود ۵٫۷۴ درصد بهبود دهد. مانند حالت قبل وضعیت دوم الگوریتم اول در ۲۰ درصد موارد بهترین جواب را داشته‌است و در ۷۳ درصد موارد نیز جواب خوشه‌بندی ابتدایی را بهبود داده‌است، همچنین در ۱۰ درصد مجموعه داده نیز این مقدار بدون تغییر باقی مانده‌است. براساس اطلاعات شکل ۹ در وضعیت دوم الگوریتم اول در بهترین حالت خود توانسته است تا ۴۳ درصد نتایج اولیه خوشه‌بندی را برای یک مجموعه داده خاص افزایش دهد و از طرفی در بدترین حالت نیز باعث کاهش ۳ درصدی شاخص رند برای یک مجموعه داده شده‌است.

جدول ۵. مقادیر P-value الگوریتم‌های ارائه شده.

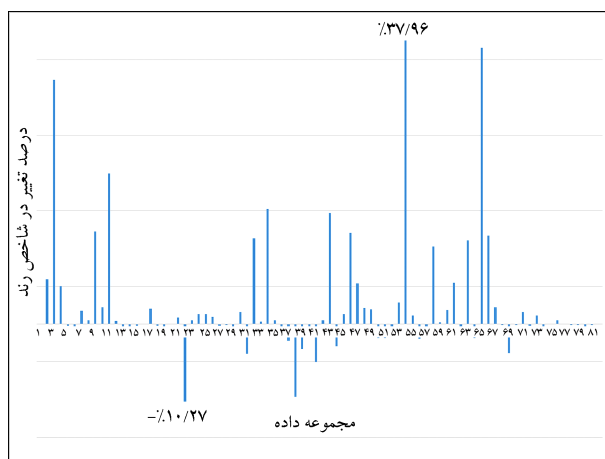
الگوریتم اول		وضعیت اول	وضعیت دوم	وضعیت سوم
وضعیت اول	وضعیت دوم			
الگوریتم اول	وضعیت دوم	۰٫۰۸۸	--	--
الگوریتم اول	وضعیت سوم	۰٫۰۰۱	۰٫۰۰۲	--
الگوریتم دوم	وضعیت سوم	۰٫۰۰۰	۰٫۰۰۰	۰٫۲۸۳

در جدول ۵ می‌توان مشاهده نمود که شاخص رند برای الگوریتم اول وضعیت دوم، از حالت اول با توجه به مقدار آلفای ۱۰ درصد بهتر می‌باشد. الگوریتم اول وضعیت سوم با توجه به مقدار آلفای ۵ درصد از هر دو الگوریتم اول وضعیت اول و دوم با توجه به آزمون بهتر می‌باشد و دارای شاخص رند بهتری است. الگوریتم دوم با توجه به مقدار P-value به دست آمده و در سطح آلفای ۵ درصد از حالت اول و دوم الگوریتم اول بهتر می‌باشد اما در مورد مقایسه الگوریتم دوم و الگوریتم اول وضعیت سوم با توجه به مقدار p-value که برابر با ۰٫۲۸۳ می‌باشد از لحاظ آماری مساوی بودن شاخص رند برای این دو وضعیت را نمی‌توان رد کرد. اما باید توجه داشت که میانگین شاخص رند برای الگوریتم دوم برابر با ۶۷٫۲۵ و برای الگوریتم اول وضعیت سوم برابر با ۶۷ می‌باشد. همچنین با توجه به دیگر معیارها در جدول ۴ می‌توان مشاهده نمود که الگوریتم دوم می‌تواند از الگوریتم اول وضعیت سوم عملکرد بهتری را داشته باشد.

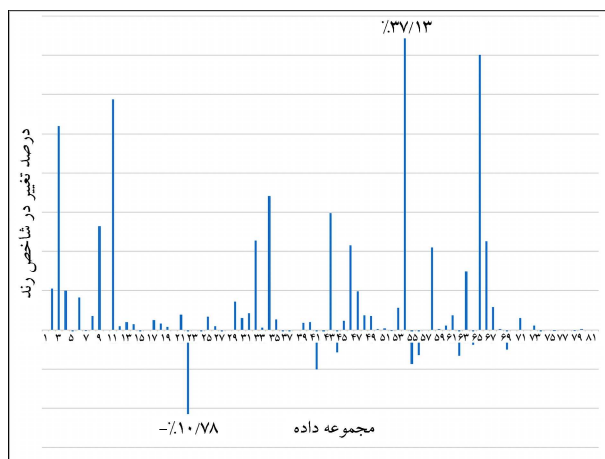
به منظور بررسی دقت و سرعت الگوریتم‌های ارائه شده، بهترین نتایج که مربوط به الگوریتم دوم می‌باشد، با چهار الگوریتم در ادبیات موضوع بررسی شده است. چهار الگوریتم مذکور عبارت‌اند از $DDDTW$ که در آن از ترکیب دو معیار فاصله به همراه الگوریتم سلسله مراتبی استفاده شده است،^[۱۳] الگوریتم KSC که در آن از الگوهای زمانی مشخص به منظور خوشه‌بندی استفاده شده است^[۴۷] و الگوریتم‌های TSC_{CH} و TSC_{MV} که بر اساس مشخصات قطعه‌بندی سری‌های زمانی توسعه داده شده‌اند.^[۲۲] جدول ۶ شاخص رند و زمان اجرای مربوط به چهار الگوریتم منتخب و همچنین الگوریتم دوم را نمایش می‌دهد. با توجه به اطلاعات جدول ۶ و جدول ۷ می‌توان مشاهده نمود که با توجه به مقدار P-value و آلفای ۵ درصد، الگوریتم ارائه شده عملکرد بهتری نسبت به این الگوریتم‌ها داشته است. در مورد مقایسه هزینه زمانی الگوریتم ارائه شده نیز با توجه به اطلاعات جدول ۶ می‌توان مشاهده نمود که الگوریتم ارائه شده، دارای زمان اجرای بسیار کمتری می‌باشد. نتایج جدول ۷ نیز این ادعا را تأیید می‌نماید. همچنین با توجه به اطلاعات جدول ۶ مشاهده می‌شود که الگوریتم ارائه شده در ۳۰ مورد از ۸۲ مجموعه داده‌ها توانسته است بهترین نتایج بین چهار الگوریتم را داشته باشد.

۴. نتیجه‌گیری

در این مقاله یک رویکرد جدید به منظور خوشه‌بندی سری‌های زمانی بر مبنای قطعه‌بندی و خوشه‌بندی ترکیبی در ۵ گام ارائه شد. در این الگوریتم‌ها قطعاتی که می‌توانند نتایج بهتری را در خوشه‌بندی داشته باشند با استفاده از یک یا چند معیار داخلی انتخاب و در مرحله نهایی با یکدیگر ترکیب می‌گردند. الگوریتم اول بر اساس معیارهای داخلی و مقدار پارامتر m در سه وضعیت اجرا شده است. الگوریتم دوم ارائه شده تا گام سوم کاملاً شبیه الگوریتم اول می‌باشد، اما در گام سوم به صورت



شکل ۱۰. اختلاف میانگین شاخص رند الگوریتم اول وضعیت سوم.



شکل ۱۱. اختلاف میانگین شاخص رند الگوریتم دوم.

حالت توانسته است ۳۷ درصد نتیجه شاخص رند را برای یک مجموعه داده بهبود ببخشد و در بدترین حالت نیز نزدیک به ۱۰ درصد نتیجه اولیه را کاهش دهد. همچنین به منظور تحلیل دقیق‌تر نتایج به دست آمده مجموعه داده‌های مورد استفاده در سه کلاس با طول کوتاه (کمتر از ۲۰۰)، طول متوسط (بین ۲۰۰ تا ۵۰۰) و طول بلند (بیشتر از ۵۰۰) تقسیم شده‌اند. با توجه به نتایج جدول ۳ می‌توان مشاهده نمود که در حالت طول کوتاه الگوریتم اول - حالت سوم می‌تواند نتایج بهتری را داشته باشد با این وجود در دو کلاس متوسط و بلند الگوریتم دوم ارائه شده توانسته است نتایج بهتری را در برداشته باشد. با توجه به اطلاعات جدول ۴ می‌توان مشاهده نمود که انحراف از معیار ۱۰ تکرار برای هر دو الگوریتم ارائه شده بسیار ناچیز می‌باشد که این نشان دهند پایداری خوب طرح ارائه شده است.

۲.۳. مقایسه الگوریتم‌ها

در این قسمت آزمون آماری به منظور بررسی بهترین الگوریتم در دو مرحله انجام گرفته است. در مرحله اول الگوریتم‌های ارائه شده با یکدیگر مقایسه شده‌اند و در گام بعدی، بهترین الگوریتم ارائه شده با چهار الگوریتم ادبیات در این حوزه مقایسه آماری شده است. یکی از روش‌های پیشنهاد شده به منظور ارزیابی الگوریتم‌های خوشه‌بندی و طبقه‌بندی، استفاده از آزمون ویلکسون می‌باشد.^[۴۶] با توجه به مقدار P-value

جدول ۶. مقایسه شاخص رند و زمان اجرای الگوریتم دوم با چهار الگوریتم منتخب.

مجموعه داده	شاخص رند (%)				زمان اجرا				الگوریتم دوم	مجموعه داده
	KSC	DD _{DTW}	TS ³ C _{MV}	TS ³ C _{CH}	KSC	DD _{DTW}	TS ³ C _{MV}	TS ³ C _{CH}		
W50	95	94	94	92	66	92	94	94	32,85	17667
ADI	92	92	92	68	95	68	92	92	11,56	5374
ARR	60	62	63	35	63	35	63	62	1,89	84
BEE	64	68	68	58	71	58	68	68	1,52	63
BFL	51	49	49	59	50	59	49	49	1,16	21
BIR	53	49	49	50	54	50	49	49	1,20	19
CAR	61	65	65	50	68	50	65	65	1,85	142
CBF	66	67	67	78	56	78	67	67	10,94	584
CHL	53	49	47	40	53	40	47	49	281,72	2623
CIN	64	64	64	56	69	56	64	64	262,56	27772
COF	79	51	51	49	75	49	51	51	0,86	10
COM	50	50	51	50	50	50	51	50	14,20	413
CRX	83	85	85	78	41	78	85	85	14,79	4356
CRY	84	84	84	69	53	69	84	84	17,26	4488
CRZ	84	84	84	71	41	71	84	84	15,70	3807
DIA	31	72	72	30	96	30	72	72	4,15	303
DPA	73	60	60	71	72	71	60	60	3,64	83
DPC	54	51	51	53	50	53	51	51	7,13	45
DPT	88	68	66	86	66	86	66	68	3,32	194
EAR	52	53	53	54	62	54	53	53	10,06	308
EC2	62	50	50	54	61	54	50	50	1,05	18
EC5	74	64	60	89	59	89	60	64	328,33	16214
EFC	50	50	50	50	81	50	50	50	10,55	136
FAA	83	85	85	60	30	60	85	85	63,10	5236
FAF	69	57	57	55	38	55	57	57	1,60	90
FIS	71	73	64	18	79	18	64	73	6,44	1265
FOA	50	52	51	54	50	54	51	52	1119,49	36823
FOB	50	50	50	50	50	50	50	50	958,58	28637
GUN	54	54	54	50	51	50	54	54	1,60	10
HAM	51	52	52	50	53	50	52	52	7,13	152
HAN	56	50	55	55	69	55	55	50	552,01	34052
HAP	67	60	60	39	69	39	60	60	25,64	3881
HER	50	50	50	51	50	51	50	50	1,97	38
INL	68	71	71	54	74	54	71	71	81,88	11234
INS	87	81	81	55	69	55	81	81	118,87	10473
ITA	51	50	50	51	64	51	50	50	7,40	20
LAR	52	55	55	34	41	34	55	55	38,69	711
LIY	60	50	50	50	50	50	50	50	2,18	128
LIV	76	75	75	60	59	60	75	75	2,78	180
MAL	94	80	80	93	92	93	80	80	509,06	18388
MEA	72	71	40	77	76	77	40	71	2,62	86
MED	63	65	65	64	47	64	65	65	21,55	1062
MPA	73	56	56	73	73	73	56	56	4,92	70

ادامه جدول ۶. مقایسه شاخص رند و زمان اجرای الگوریتم دوم با چهار الگوریتم منتخب.

زمان اجرا					شاخص رند (%)				الگوریتم دوم	مجموعه داده
KSC	DDDTW	TSCMV	TSCCH	الگوریتم دوم	KSC	DDDTW	TSCMV	TSCCH		
۶۵	۹۳۰۶	۳۹۶	۳۹۵	۸,۷۳	۵۰	۵۰	۵۱	۵۱	۵۰	MPC
۲۰۴	۳۳۱۹	۲۸۹	۲۸۶	۴,۸۷	۸۱	۸۰	۸۲	۷۴	۸۴	MPT
۴۹۶	۲۰۴۵۶	۷۲۱	۷۱۷	۱۷,۴۸	۵۸	۵۰	۵۰	۵۰	۶۲	MOT
۲۱۸۶۵۰	۷۶۳۵۸۷	۱۷۲۲۶	۱۷۱۱۵	۹۹۲,۱۸	۹۵	۷۰	۹۵	۹۴	۹۵	NO۱
۲۰۸۴۱۶	۷۶۳۵۸۷	۱۳۶۷۶	۱۳۵۹۴	۹۸۰,۵۲	۹۷	۸۵	۹۵	۹۵	۹۶	NO۲
۶۶	۲۴۴۴	۱۳۹	۱۳۹	۲,۳۶	۸۵	۷۶	۷۷	۷۷	۸۶	OLI
۶۷۰	۶۰۷۶۵	۱۱۰۰	۱۰۹۷	۱۲,۰۴	۲۹	۶۲	۷۳	۷۳	۷۳	OSU
۲۰۰	۷۷۵۵۸	۱۵۱۵	۱۴۹۹	۷۷,۶۲	۵۱	۵۴	۵۱	۵۱	۵۴	PHA
۳۶۴۷۰۳	۷۶۳۵۸۷	۱۵۵۲۵	۱۵۴۷۷	۴۳۲,۷۵	۵۱	۴۵	۹۳	۹۳	۹۳	PHO
۵۳	۲۰۴۷	۱۴۹	۱۴۹	۲,۹۹	۹۲	۱۰۰	۸۰	۸۳	۹۴	PLA
۸۲	۴۹۰۹	۲۳۶	۲۳۵	۵,۷۷	۷۶	۷۸	۷۶	۷۶	۷۹	PPA
۳۲	۸۱۹۵	۴۲۱	۴۱۸	۸,۱۹	۵۳	۵۴	۵۶	۵۶	۵۲	PPC
۲۲۱	۳۸۵۱	۳۲۱	۳۱۹	۵,۵۲	۸۱	۸۸	۷۸	۷۸	۷۹	PPT
۹۰۰	۴۲۲۸۴۳	۳۹۷۲	۳۹۶۹	۳۵,۷۸	۳۹	۳۵	۵۴	۵۶	۵۵	REF
۱۳۵۸	۳۵۵۱۰۹	۴۰۲۴	۴۰۲۲	۳۷,۳۵	۴۵	۳۵	۵۳	۵۳	۵۵	SCR
۱۱۱	۱۳۸۷۹	۱۰۰۶	۱۰۰۵	۳,۱۴	۵۰	۵۰	۹۹	۹۹	۵۰	SHS
۵۱۱۰۷	۵۴۶۵۸۵	۳۷۷۷	۳۷۳۸	۶۸,۸۰	۶۳	۸۴	۹۷	۹۷	۹۵	SHA
۱۴۸۷	۳۷۹۸۶۹	۴۰۵۰	۴۰۴۸	۳۸,۹۰	۵۴	۳۴	۵۹	۵۹	۴۳	SMA
۷۴	۲۷۷۰	۳۵۰	۳۴۹	۴,۶۶	۷۵	۵۰	۵۲	۵۱	۵۰	SO۱
۱۴۹	۹۰۴۰	۶۱۰	۶۰۷	۹,۹۰	۶۶	۵۳	۵۳	۶۰	۶۴	SO۲
۳۶۶	۹۲۰۵۲	۱۳۳۴	۱۳۳۱	۲۴,۰۱	۵۰	۵۰	۵۲	۵۰	۵۰	STR
۱۶۲۰	۳۶۰۹۲	۹۹۶	۹۸۵	۲۳,۴۲	۶۳	۳۵	۸۸	۸۸	۸۷	SWE
۲۸۶۷	۲۴۸۴۵۴	۲۷۷۹	۲۷۷۴	۴۰,۴۵	۶۰	۸۹	۸۱	۸۱	۸۹	SYM
۲۳۵	۲۱۱۲	۳۶۵	۳۶۱	۵,۶۲	۳۸	۸۸	۷۸	۷۸	۸۲	SYN
۱۲۱	۸۳۱۵	۵۲۸	۵۲۷	۳,۰۶	۵۳	۵۱	۵۱	۵۱	۵۰	TO۱
۷۹	۵۲۵۸	۳۳۹	۳۳۸	۲,۰۲	۵۳	۶۷	۵۰	۵۰	۵۱	TO۲
۱۱۴	۴۹۸۶	۳۲۵	۳۲۵	۳,۲۴	۷۲	۸۷	۸۴	۸۴	۷۵	TRA
۴۶۰۲	۵۸۱۰۵۰	۶۵۳۲	۶۴۶۵	۳۶۱,۴۶	۴۶	۸۵	۶۴	۶۴	۶۳	TWP
۲۴۰	۱۱۹۰۳	۷۰۷	۷۰۳	۱۴,۰۳	۵۴	۵۰	۶۴	۶۴	۵۰	TWE
۴۷۴۲۳	۷۶۳۵۸۷	۷۵۲۳	۷۴۷۰	۵۰,۱,۱۵	۵۱	۸۰	۷۵	۷۸	۸۱	UWZ
۴۸۷۵۱	۷۶۳۵۸۷	۹۰۱۴	۸۹۵۱	۵۰۹,۲۳	۵۴	۸۲	۷۶	۷۸	۸۲	UWY
۴۷۵۵۴	۷۶۳۵۸۷	۷۳۱۵	۷۲۶۴	۵۸۹,۵۳	۵۴	۷۴	۸۰	۸۰	۸۲	UWZ
۱۶۷۰۷۹	۷۶۳۵۸۷	۱۸۷۹۹	۱۸۷۵۴	۱۷۵۳,۹۲	۴۵	۵۹	۷۶	۷۶	۸۶	UWA
۱۶۸۲	۷۶۳۵۸۷	۴۱۵۸	۴۰۸۷	۱۵۷۱,۷۶	۵۹	۵۳	۶۶	۵۰	۵۳	WAF
۳۱	۱۲۷۳	۱۰۵	۱۰۵	۵,۳۷	۵۹	۵۰	۵۰	۵۷	۵۰	WIN
۸۲۱۲	۹۸۲۳۳	۱۴۷۶	۱۴۶۳	۵۴,۰۸	۵۰	۸۷	۸۷	۸۷	۸۹	WOS
۱۰۸۵	۸۲۵۲۷	۱۶۰۶	۱۶۰۵	۱۲,۹۰	۵۳	۶۲	۵۸	۶۰	۶۳	WOR
۶۷۶	۷۷۴۹۷	۱۶۲۱	۱۶۲۰	۸,۷۰	۵۰	۵۰	۵۱	۵۱	۵۰	WOT
۴۸۴۸	۷۶۳۵۸۷	۷۹۸۳	۷۹۵۹	۴۸۱,۲۲	۵۰	۵۰	۵۰	۵۱	۵۰	YOG
۱۷۳۶۵	۲۳۶۶۶۷	۳۲۸۴	۳۲۷۰	۱۶۲	۶۰,۲۶	۶۰,۵۵	۶۵,۶۸	۶۶,۱۰	۶۷,۲۵	توسط

هم‌زمان از سه معیار داخلی به‌منظور انتخاب $3m$ برجسب استفاده می‌گردد و در گام پنجم این برجسب‌ها مجدداً به‌منظور نتیجه نهایی با یکدیگر ترکیب می‌گردند. شاخص رند در نتیجه ۱۰ تکرار برای سه وضعیت مختلف الگوریتم اول و یک وضعیت الگوریتم دوم به‌ترتیب برابر $۰.۶۳/۰.۵۲$ ، $۰.۶۳/۰.۶۷$ و $۰.۶۷/۰.۲۵$ برای ۸۲ مجموعه داده بوده و همچنین آزمون آماری ویلکانسون به‌همراه دیگر معیارها نشان داد که الگوریتم دوم می‌تواند عملکرد بهتری در خوشه‌بندی سری‌های زمانی

جدول ۷. نتایج آزمون آماری الگوریتم ارائه شده با چهار الگوریتم منتخب ادبیات در دو شاخص رند و زمان اجرا.

الگوریتم	معیار مقایسه	DD_{DTW}	KSC	TS^3C_{MV}	TS^3C_{CH}
p-value	شاخص رند	۰/۰۰۰	۰/۰۰۰۲	۰/۰۰۱۸	۰/۰۰۲۱
p-value	زمان	۰/۰۰۰	۰/۰۰۰	۰/۰۰۰	۰/۰۰۰

مقایسه شده کارایی بیشتری دارد. با توجه به دقت بالا و هزینه زمانی بسیار ناچیز طرح ارائه شده جهت خوشه بندی، الگوریتم مورد نظر می تواند به راحتی در مواردی مانند تشخیص الگوهای مصرف انواع حامل های انرژی، تشخیص الگوهای رفتاری بیماری های واگیردار و غیره مورد استفاده قرار گیرد. همچنین با توجه به اینکه یکی از کاربردهای اصلی خوشه بندی استفاده در پیش بینی سری های زمانی می باشد. با افزایش دقت خوشه بندی در زمان کم می توان عملاً دقت پیش بینی این نوع داده ها را نیز افزایش داد.

۵. مطالعات آینده

طرح ارائه شده دارای پارامترها و مشخصات مختلفی می باشد که در این مطالعه صرفاً حالت ها اندکی از آن مورد بررسی و تحلیل قرار گرفته است. در مطالعات آینده می توان انواع دیگر الگوریتم های خوشه بندی، معیارهای درونی دیگر و همچنین پارامترهای ورودی دیگر را نیز مورد بررسی و تحلیل قرار داد. همچنین با توجه به کارایی بالای الگوریتم در زمان نسبتاً اندک از الگوریتم فوق می توان به منظور کشف الگوهای مختلف در زمین های انرژی و درمان نیز استفاده عملی نمود.

داشته باشد. همچنین عملکرد الگوریتم دوم به لحاظ شاخص رند و متوسط زمان اجرای الگوریتم با ۴ الگوریتم TS^3C_{CH} ، TS^3C_{MV} ، KSC، DD_{DTW} مقایسه شده است.

در مطالعات اخیر سعی بر آن شده است با استفاده از تعریف معیارهای جدید فاصله و با استفاده از الگوریتم های خاص، کارایی خوشه بندی افزایش یابد. اما با این وجود نتایج مطالعات اخیر نشان داده است که استفاده از این روش ها علی رقم بهبود دقت خوشه بندی می تواند هزینه زمانی بسیار بالایی داشته باشد. به عنوان مثال در چهار الگوریتم موجود در ادبیات حتی می تواند به عددی نزدیک به ۲۳۶ هزارم ثانیه برسد که عملاً استفاده از این الگوریتم ها در بخش صنعت و خدمات را ناممکن می نماید. در طرح ارائه شده، سعی بر آن است علاوه بر افزایش دقت خوشه بندی این کار با حداقل هزینه زمانی انجام پذیرد. با توجه به نتایج ارائه شده در جدول ۶ به خوبی می توان مشاهده نمود که طرح ارائه شده علاوه بر افزایش دقت خوشه بندی می تواند این امر را در کمترین زمان ممکن انجام دهد. همچنین نشان داده شده است که طرح ارائه شده در ۳۲٪ از مجموعه داده های مورد استفاده توانسته است بهترین دقت را در بین الگوریتم های موجود داشته باشد. نتایج آزمون آماری به خوبی بیانگر این است که الگوریتم دوم ارائه شده به لحاظ شاخص رند و همچنین زمان اجرا از همه الگوریتم های

پانویس ها

1. Dynamic Time Warping
2. Longest Common Subsequence
3. Rand Index
4. Weighted Dynamic Time Warping
5. Matrix Profile
6. Maximum Shifting Correlation Distance
7. Derivative Dynamic Time Warping

منابع (References)

1. Maleki, M., Bidram, H. and Wraith, D., 2022. Robust clustering of COVID-19 cases across US counties using mixtures of asymmetric time series models with time varying and freely indexed covariates. *Journal of Applied Statistics*, 50(11), pp. 2648-2662.
2. Bair, E., 2013. Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(5), pp. 349-361. <https://doi.org/10.48550/arXiv.1307.0252>.
3. Dau, H.A., Begum, N. and Keogh, E., 2016. Semi-supervision dramatically improves time series clustering under dynamic time warping. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 999-1008.
4. Alhusain, L. and Hafez, A.M., 2017. Cluster ensemble based on Random Forests for genetic data. *BioData Mining*, 10(1), p. 37. <https://doi.org/10.1186/s13040-017-0156-2>.
5. Ma, R. and Angryk, R., 2017. Distance and density clustering for time series data. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, pp. 25-32.

6. Mehrmolaei, S. and Keyvanpour, M.R., 2018. A comparative study on weighting-based clustering techniques: Time series data. *2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN)*, IEEE, pp. 65-72.
7. Tavakoli, N., Siami-Namini, S., Adl Khanghah, M., Mirza Soltani, F. and Siami Namin, A., 2020. An autoencoder-based deep learning approach for clustering time series data. *SN Applied Sciences*, 2(1), pp. 1-25. <https://doi.org/10.1007/s42452-020-2584-8>.
8. Lafabregue, B., Weber, J., Gancarski, P. and Forestier, G., 2022. End-to-end deep representation learning for time series clustering: A comparative study. *Data Mining and Knowledge Discovery*, 36(1), pp. 29-81. <https://doi.org/10.1007/s10618-021-00796-y>.
9. Zolhavarieh, S., Aghabozorgi, S. and Teh, Y.W., 2014. A review of subsequence time series clustering. *The Scientific World Journal*, 2014(1), pp. 1-19. <https://doi.org/10.1155/2014/312521>.
10. Ralanamahatana, C.A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M. and Das, G., 2005. *Mining Time Series Data*. Data Mining and Knowledge Discovery Handbook: Springer pp. 1069-1103. Springer, US.
11. Kamalzadeh, H., Ahmadi, A. and Mansour, S., 2020. Clustering time-series by a novel slope-based similarity measure considering particle swarm optimization. *Applied Soft Computing*, 96(1), p. 106701. <https://doi.org/10.1016/j.asoc.2020.106701>.
12. Soleimani, G. and Abessi, M., 2020. DLCSS: A new similarity measure for time series data mining. *Engineering Applications of Artificial Intelligence*, 92(1), p. 103664. <https://doi.org/10.1016/j.engappai.2020.103664>.
13. Luczak, M., 2016. Hierarchical clustering of time series data with parametric derivative dynamic time warping. *Expert Systems with Applications*, 62(1), pp. 116-130. <https://doi.org/10.1016/j.eswa.2016.06.012>.
14. Rokach, L. and Maimon, O., 2005. *Clustering Methods*. Data Mining and Knowledge Discovery Handbook: Springer, pp. 321-352.
15. Rahim Khan, M.A. and Zakarya, M., 2013. Longest common subsequence based algorithm for measuring similarity between time series: A new approach. *World Applied Sciences Journal*, 24(9), pp. 1192-1198. <https://doi.org/10.11648/j.ajdmkd.20190401.16>.
16. Wang, X., Yu, F., Pedrycz, W. and Wang, J., 2019. Hierarchical clustering of unequal-length time series with area-based shape distance. *Soft Computing*, 23(15), pp. 6331-6343. <https://doi.org/10.1007/s00500-018-3287-6>.
17. Chu, K.K.W. and Wong, M.H., 1999. Fast time-series searching with scaling and shifting. *Proceedings of the Eighteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 237-248.
18. Gharghabi, S., Imani, S., Bagnall, A., Darvishzadeh, A. and Keogh, E., 2018. Matrix profile xii: Mpdist: A novel time series distance measure to allow data mining in more challenging scenarios. *2018 IEEE International Conference on Data Mining (ICDM)*, IEEE, pp. 965-970.
19. Jiang, G., Wang, W. and Zhang, W., 2019. A novel distance measure for time series: Maximum shifting correlation distance. *Pattern Recognition Letters*, 117(1), pp. 58-65. <https://doi.org/10.1016/j.patrec.2018.11.013>.
20. Hong, D., Gu, Q. and Whitehouse, K., 2017. High-dimensional time series clustering via cross-predictability. *Artificial Intelligence and Statistics*, PMLR, pp. 642-651.
21. Górecki, T., 2018. Classification of time series using combination of DTW and LCSS dissimilarity measures. *Communications in Statistics-Simulation and Computation*, 47(1), pp. 263-276. <https://doi.org/10.1080/03610918.2017.1280829>.
22. Guijo-Rubio, D., Durrán-Rosal, A.M., Gutiérrez, P.A., Troncoso, A. and Hervás-Martínez, C., 2020. Time-series clustering based on the characterization of segment typologies. *IEEE Transactions on Cybernetics*, 51(11), pp. 5409-5422. <https://doi.org/10.1109/tcyb.2019.2962584>.
23. Aghabozorgi, S., Ying Wah, T., Herawan, T., Jalab, H.A., Shaygan, M.A. and Jalali, A., 2014. A hybrid algorithm for clustering of time series data based on affinity search technique. *The Scientific World Journal*, 2014(1), pp. 1-12. <https://doi.org/10.1155/2014/562194>.
24. Zhang, X., Liu, J., Du, Y. and Lv, T., 2011. A novel clustering method on time series data. *Expert Systems with Applications*, 38(9), pp. 11891-11900. <https://doi.org/10.1016/j.eswa.2011.03.081>.
25. Manakova, N. and Tkachenko, V., 2020. Two-stage time-series clustering approach under reducing time cost requirement. *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, IEEE, pp. 653-658.
26. Hyndman, R.J., Wang, E. and Laptev, N., 2015. Large-scale unusual time series detection. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, IEEE, pp. 1616-1619.
27. Zou, Y., Donner, R.V., Marwan, N., Donges, J.F. and Kurths, J., 2019. Complex network approaches to nonlinear time series analysis. *Physics Reports*, 787(1), pp. 1-97. <https://doi.org/10.1016/j.physrep.2018.10.005>.
28. Silva, V.F., Silva, M.E., Ribeiro, P. and Silva, F., 2021. Time series analysis via network science: Concepts and algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(3), p. e1404. <https://doi.org/10.1002/widm.1404>.
29. Ferreira, L.N. and Zhao, L., 2016. Time series clustering via community detection in networks. *Information Sciences*, 326(1), pp. 227-242. <https://doi.org/10.1016/j.ins.2015.07.046>.
30. Bonacina, F., Miele, E.S. and Corsini, A., 2020. Time series clustering: A complex network-based approach for feature selection in multi-sensor data. *Modelling*, 1(1), pp. 1-21. <https://doi.org/10.3390/modelling1010001>.
31. Koski, A., Juhola, M. and Meriste, M., 1995. Syntactic recognition of ECG signals by attributed finite automata. *Pattern Recognition*, 28(12), pp. 1927-1940. [https://doi.org/10.1016/0031-3203\(95\)00052-6](https://doi.org/10.1016/0031-3203(95)00052-6).
32. Keogh, E.J. and Pazzani, M.J., 1998. An enhanced representation of time series which allows fast and accurate classification. *Clustering and Relevance Feedback, Kdd*, 98, pp. 239-243.

33. Keogh, E., Chu, S., Hart, D. and Pazzani, M., 2004. Segmenting time series: A survey and novel approach. *Data Mining in Time Series Databases: World Scientific*, pp. 1-21.
34. Faloutsos, C., Ranganathan, M. and Manolopoulos, Y., 1994. Fast subsequence matching in time-series databases. *ACM Sigmod Record*, 23(2), pp. 419-429. <https://doi.org/10.1145/191843.191925>.
35. Keogh, E. and Ratanamahatana, C.A., 2005. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3), pp. 358-386. <https://doi.org/10.1007/s10115-004-0154-9>.
36. Djukanovic, M., Raidl, G.R. and Blum, C., 2020. Finding longest common subsequences: New anytime A* search results. *Applied Soft Computing*, 95(1), p. 106499. <https://doi.org/10.1016/j.asoc.2020.106499>.
37. Paterson, M. and Dančák, V., 1994. Longest common subsequences, *International Symposium on Mathematical Foundations of Computer Science*, pp. 127-142.
38. Lin, R., King-Ip, A. and Shim, H.S.S.K., 1995. Fast similarity search in the presence of noise, scaling, and translation in time-series databases, *Proceeding of the 21th International Conference on Very Large Data Bases*, Cite-seer, pp. 490-501.
39. Vlachos, M., Kollios, G. and Gunopulos, D., 2002. Discovering similar multidimensional trajectories, *Proceedings 18th International Conference on Data Engineering*, pp. 673-684.
40. Boongoen, T. and Iam-On, N., 2018. Cluster ensembles: A survey of approaches with recent extensions and applications. *Computer Science Review*, 28(1), pp. 1-25. <https://doi.org/10.1016/j.cosrev.2018.01.003>.
41. Huang, D., Wang, C.-D. and Lai, J.-H., 2017. Locally weighted ensemble clustering. *IEEE Transactions on Cybernetics*, 48(5), pp. 1460-1473. <https://doi.org/10.1109/tycb.2017.2702343>.
42. Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J., 2010. Understanding of internal clustering validation measures, *2010 IEEE International Conference on Data Mining*, IEEE, pp. 911-916.
43. Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1), pp. 95-104. <https://doi.org/10.1080/01969727408546059>.
44. Caliński, T. and Harabasz, J., 1974. A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, 3(1), pp. 1-27. <https://doi.org/10.1080/03610917408548446>.
45. Dau, H.A., Bagnall, A., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., and Keogh, E., 2019. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6), pp. 1293-1305. <https://doi.org/10.1109/jas.2019.1911747>.
46. Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7, pp. 1-30.
47. Yang, J. and Leskovec, J., 2011. Patterns of temporal variation in online media, *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 177-186.