یک رویکرد جدید به‌منظور خوشه‌بندی سری های زمانی با استفاده از ترکیب زیرسری‌هایی زمانی

علی قربانیان (دانشجوی دکتری)
حمیده رضوی* (دانشیار)
گروه مهندسی صنایع، دانشکده مهندسی، دانشگاه فردوسی مشهد

چکیده

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 به‌منظور 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 مجموعه‌داده سری زمانی با استفاده از اندازه پنجره ثابت قطعه‌بندی شده و هر قطعه به‌طور جداگانه خوشه‌بندی شده است. سپس با استفاده از معیارهای درونی، بهترین نتایج حاصله انتخاب شده‌است. در مرحله دوم نتایج حاصل از مرحله اول با استفاده از خوشه‌بندی ترکیبی، پردازش شده و برچسب نهایی خوشه‌بندی حاصل شده‌است. نتایج الگوریتم ارائه‌شده 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.

کلمات کلیدی: سری‌های زمانی، خوشه‌بندی، قطعه‌بندی، خوشه‌بندی ترکیبی

A new approach to time series clustering by combination of sub-series

Ali Ghorbanian (Phd Student)
Hamideh Razavi* (Associate Professor)
Department Of Industrial Engineering. Faculty Of Engineering. Ferdowsi University Of Mashhad (FUM)

Abstract

Time series-clustering, defined as deriving trends and archetypes from sequential data, divides time series into groups considering their characteristics. Previous work mainly focused on distance criterion and clustering algorithm to cluster the time series so few researchers have investigated the similarities between the segments of a time series. To address this research gap, we propose a new two-step approach based on sub-time series and combination clustering. In the first step, a time series data set is segmented using a fixed window size, and each segment is clustered by applying a hierarchical clustering algorithm and Euclidean distance. Also, we use a logarithmic relation based on the length of the time series data set to determine the number of components, selecting the best outcomes using various internal criteria including intergroup variance, Kalinsky-Harbaz, and Dunn index. In the second step, the results of the first stage are processed using ensemble clustering and the final clustering label is obtained. We develop two novel algorithms based on different internal criteria for selecting the best segmentations: the first one in which we consider only one internal criterion and the second one in which we consider three internal criteria simultaneously. Moreover, we run various settings on ۸۲ datasets with ۱۰ replications for the two presented algorithms, checking the final precision using an external RAND index. Then, in order to identify the best settings for the proposed algorithms we applied Wilkinson statistical test. Statistical comparison of the results of the two new algorithms on ۸۲ data sets with some algorithms in the related literature indicates significant improvement In terms of error rate and execution time. Finally, the findings acquired based on the best settings of the proposed algorithms indicate that the suggested method has the best RAND index among the previous algorithms in the literature for ۳۲٪ of the dataset tiers.

Key Words: Time series, Clustering, Data mining ،Sub-series.

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

* نویسنده مسئول. آدرس پستی: مشهد- دانشگاه فردوسی مشهد- دانشکده مهندسی- گروه مهندسی صنایع

۱- مقدمه و مرور ادبیات

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 اطلاعی از برچسب اشیا موجود نمی‌باشد. برچسب‌ها اعدادی هستند که از فرایند خوشه‌بندی حاصل گشته و نشان‌دهنده سری‌های زمانی قرارگرفته 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 به‌صورت 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 می باشد. هدف از این کار شناخت الگوهای رفتاری مصرف‌کنندگان به‌منظور مدیریت مصرف متقاضیان در زمان‌های مختلف می‌باشد. همچنین به‌منظور کشف *الگوهای* مشابه در حوزه سلامت مانند همه‌گیری کوید-۱۹ نیز از این تکنیک استفاده می گردد[۱]. به‌صورت غیرمستقیم نیز خوشه‌بندی سری‌های زمانی به‌منظور پیش‌بینی این نوع داده مورد استفاده قرار می‌گیرد. درواقع 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 درگروهای مختلف قرار می‌گیرند. به عنوان مثال در شکل۲ چندین سری زمانی در دو گروه با رنگ آبی و سبز خوشه‌بندی شده‌اند که نشان‌دهنده خوشه‌بندی کل می‌باشد. از روش‌های خوشه‌بندی سری‌های زمانی می‌توان روش‌های نیمه نظارت‌شده [۲, ۳]، رویکردهای ترکیب و جنگل تصادفی [۴]، ترکیب روش‌های خوشه‌بندی بر مبنای فاصله و چگالی [۵]، استفاده از روش‌های خوشه‌بندی وزنی [۶] و استفاده از شبکه‌های عصبی و رویکردهای یادگیری عمیق اشاره نمود [۷, ۸]. در این تحقیق از روش خوشه‌بندی سری زمانی کل استفاده می‌شود که در آن یک مجموعه از سری‌های زمانی با توجه به معیارهای مشابهت خوشه‌بندی می‌گردد [۹].

یکی از راهکارهایی که به‌منظور افزایش دقت خوشه‌بندی موردتوجه قرار گرفته است، 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 موردتوجه قرارگرفته است [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 بردارد.

منظور از دقت خوشه‌بندی شاخص‌های بیرونی خوشه‌بندی مانند خلوص، شاخص رند[۳] و آنتروپی می‌باشد [۱۴]. به عنوان مثال در پژوهشی که توسط رحیم خان و زکریا با ارائه یک الگوریتم مناسب به‌منظور 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 به‌صورت فازی دقت خوشه‌بندی را بهبود بخشیده‌اند ولی در روش ایشان زمان افزایش یافته‌است [۱۲]. همچنین کمال زاده و همکاران یک معیار فاصله جدید به‌منظور خوشه‌بندی سری‌های زمانی با طول بلند با استفاده از روابط هندسی تعریف کرده‌اند [۱۱]. 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 به‌منظور استفاده از مزایای هریک از معیارهای فاصله نظیر DTW، DDTW[۷] و LCSS از ترکیب این معیارها به‌منظور افزایش دقت خوشه‌بندی استفاده شده‌است [۵, ۱۳, ۲۱]. منظور از این مزایا شناسایی نقاط پرت و همچنین شناسایی انتقال داده‌ها در طول زمان در یک سری زمانی می‌باشد. باوجود این که

---

[۱] Dynamic time warping
[۲] Longest common subsequence
[۳] Rand Index
[۴] Weighted dynamic time warping
[۵] Matrix profile
[۶] Maximum shifting correlation distance
[۷] Derivative dynamic time warping

ترکیب معیارها دقت خوشه‌بندی را افزایش می‌دهد به دلیل افزایش حجم محاسبات، زمان بیشتری صرف می‌کند. به‌عنوان نمونه استفاده از معیار DDTW به‌منظور خوشه‌بندی ۸۴ مجموعه داده، زمانی نزدیک به ۸۰ ساعت برای هر مجموعه داده در برداشته است درحالی که متوسط شاخص رند برای هر مجموعه داده نیز برابر با ۶۰ می باشد [۲۲]. علیرغم زمان اجرای طولانی، مشاهده می‌گردد که دقت خوشه‌بندی نیز افزایش چشم‌گیری نداشته است به همین سبب در این پژوهش از فاصله اقلیدسی استفاده می‌گردد.

به‌منظور برخورد با چالش هزینه زمانی نیز چندین روش مورد توجه قرار گرفته‌است. روش اول، استفاده از الگوریتم‌های چندمرحله‌ای به‌منظور کاهش حجم مجموعه‌داده 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 ژانگ و همکاران نیز با استفاده از یک شبکه وزن‌دار جهت در یک الگوریتم دومرحله‌ای، پیچیدگی مسئله خوشه‌بندی را کاهش داده‌اند [۲۴]. ماناکوا و تاچنکو با به‌کارگیری معیار هاپکینز و سیستم نمونه‌گیری، یک الگوریتم دومرحله‌ای را توسعه داده‌اند [۲۵]. در روش دوم، از مشخصه‌های سری زمانی به‌منظور خوشه‌بندی استفاده می‌شود. این مشخصه‌ها می‌تواند به‌صورت مستقیم از سری زمانی استخراج گردد. ازجمله مشخصه‌های پرکاربرد می‌توان به معیارهای واریانس، همبستگی مرتبه اول، خطی بودن، انحنا، فصلی بودن، نقطه اوج و فرورفتگی اشاره نمود [۲۶]. به‌عنوان‌مثال زو و همکاران به‌منظور استخراج مشخص‌های 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 کرده است [۲۸]. همچنین فریرا و ژاوبه به‌منظور خوشه‌بندی یک مجموعه داده از دو رویکرد k-NN و ε-NN به‌منظور 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.

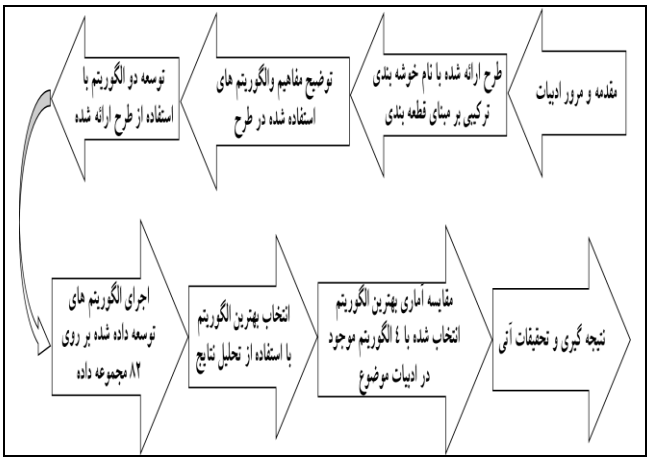
مطالعات اخیر نشان داده‌است که قطعه‌بندی یک سری زمانی می‌تواند نقش مهمی در افزایش دقت خوشه‌بندی سری‌های زمانی داشته باشد. به‌عنوان‌مثال، گویجو و همکاران در مرحله اول یک سری زمانی را قطعه‌بندی نموده و در مرحله بعد با توجه به خصوصیات آماری هر یک از قطعه‌های ایجادشده، مشخصه‌های آن‌ها را استخراج نموده‌اند. سپس از مشخصات استخراج‌شده به‌منظور خوشه‌بندی نهایی استفاده کرده‌اند [۲۲]. همچنین بوناسینا و همکاران به‌منظور خوشه‌بندی از ترکیب رویکردهای تبدیل سری زمانی به شبکه و قطعه‌بندی، استفاده نموده‌اند [۳۰]. عملکرد مناسب الگوریتم‌های ارائه‌شده بر مبنای قطعه‌بندی جایگاه آن را در مباحث خوشه‌بندی تثبیت نموده است.

در پژوهش‌های پیشین در بیشتر موارد به‌منظور بهبود کارایی خوشه‌بندی یک مجموعه‌داده سعی بر استفاده از معیارهای فاصله جدید و یا الگوریتم‌های جدید در حوزه یادگیری ماشین بوده‌است. اما نتایج ارائه‌شده نشان داده است 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 داده است که استفاده از الگوریتم‌های خاص شاید بتواند در بعضی از مجموعه داده‌ها نتایج ملموسی داشته باشد، اما این مورد همه مجموعه‌دادها جامعیت ندارد. با توجه به پژوهش‌های انجام‌گرفته در حوزه داده‌کاوی می‌توان به نقش مهم خوشه‌بندی ترکیبی در افزایش کارایی خوشه‌بندی در داده‌های مختلف اشاره نمود، اما در هیچ‌کدام از پژوهش‌های انجام‌گرفته در حوزه سری‌های زمانی این تکنیک موردتوجه قرار نگرفته است. همچنین پژوهش‌های اخیر نشان داده است که زیر سری‌های ایجادشده با استفاده از قطعه‌بندی یک سری زمانی می‌تواند دارای اهمیت فراوانی به‌منظور خوشه‌بندی یک مجموعه داده باشد. با توجه به موارد اشاره شده هدف از این پژوهش ارائه یک الگوریتم با کارایی بالا و سریع به‌منظور خوشه‌بندی یک مجموعه داده سری زمانی می باشد. در همین راستا در تحقیق فوق نشان داده‌ایم که لزوماً همه قطعات ایجادشده در یک مجموعه داده سری زمانی نمی‌تواند نماینده خوبی به‌منظور استفاده در خوشه‌بندی نهایی باشد، برای رفع این مسئله نشان داده‌ایم که می‌توان با ابتدا با استفاده از معیارهای درونی بهترین قطعات ایجادشده را انتخاب نمود و با استفاده از بهترین قطعات ایجادشده و خوشه‌بندی ترکیبی این قطعات، یک الگوریتم سریع و دقیق به‌منظور خوشه‌بندی داشته باشیم.

در این مقاله، با توجه به نقش مهم قطعه‌بندی سری‌های زمانی به‌منظور خوشه‌بندی، دو الگوریتم خوشه‌بندی بر مبنای قطعه‌بندی و خوشه‌بندی ترکیبی ارائه‌شده است. الگوریتم‌های ارائه‌شده دارای سه گام اصلی قطعه‌بندی، خوشه‌بندی و خوشه‌بندی ترکیبی می‌باشند. در الگوریتم اول به‌منظور خوشه‌بندی ترکیبی نهایی از یک معیار درونی به‌منظور 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.

شکل ۱ فلوچارت 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 استفاده‌شده در طرح به‌صورت کامل توضیح داده می‌شود. در گام بعدی با توجه به طرح ارائه‌شده 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. به‌منظور بررسی کارایی الگوریتم انتخاب‌شده 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.

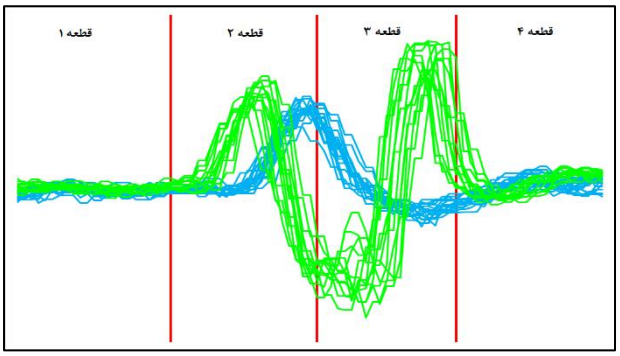


شکل ۱: فلوچارت روش پژوهش

۲- خوشه‌بندی ترکیبی بر مبنای قطعه‌بندی

در مدل این پژوهش به‌جای این که یک مجموعه‌داده سری زمانی به‌صورت مستقیم خوشه‌بندی گردد، ابتدا این مجموعه‌داده به قطعات مساوی تبدیل شده و سپس هر قطعه ایجادشده به‌صورت جداگانه خوشه‌بندی می‌گردد. در پایان نیز با استفاده از الگوریتم‌های خوشه‌بندی ترکیبی نتایج به‌دست‌آمده از قطعات مختلف با یکدیگر ترکیب می‌گردند. الگوریتم دارای سه گام اصلی شامل قطعه‌بندی با اندازه ثابت، خوشه‌بندی هر یک از قطعات ایجادشده و ترکیب نتایج حاصل شده می‌باشد.

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

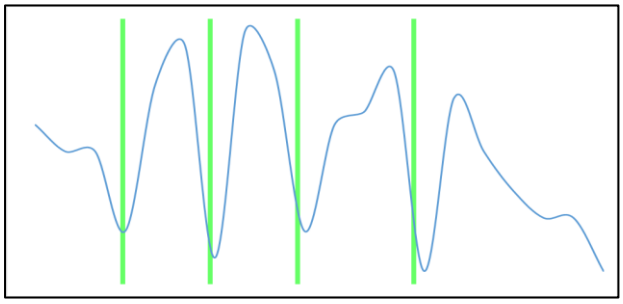


شکل ۲ : قطعه‌بندی یک سری زمانی

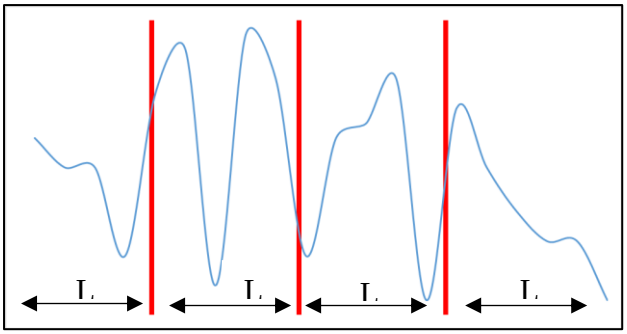
۲-۱ قطعه‌بندی

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 زیر سری‌های، سری‌های زمانی با ابعاد کوچک‌تر از سری زمانی اصلی می‌باشند که در کنار یکدیگر سری زمانی اصلی را تشکیل می‌دهند. به‌عنوان مثال در شکل ۲ چهار زیر سری تشکیل‌شده 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 [۳۱, ۳۲]. 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 می‌گردد [۳۳]. شکل ۳ (ب) این نوع قطعه‌بندی را نمایش می‌دهد.



الف- قطعه‌بندی با طول متغیر



ب- قطعه‌بندی با طول ثابت

شکل ۳: انواع مختلف قطعه‌بندی سری زمانی

۲-۲ خوشه‌بندی قطعات ایجادشده

الگوریتمی که به‌منظور خوشه‌بندی قطعات ایجادشده در گام اول مورداستفاده قرارگرفته 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 به‌منظور پیوند استفاده شده است. روابط شماره (۱) و (۲) به ترتیب این دو نوع فاصله را نمایش می‌دهند. A و B نمایش‌دهنده خوشه‌ها و a و b اعضا هر خوشه می‌باشند همچنین d نشان‌دهنده فاصله بین دو عضو می‌باشد[۱۳].

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

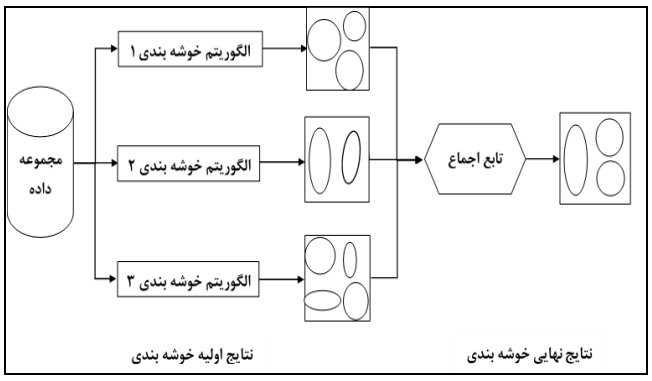
$$d_{mean} = \frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d\left(a,b\right) \quad (۲)$$

به‌منظور محاسبه فاصله بین دو سری زمانی می‌توان از معیارهای فاصله‌ای متفاوتی از قبیل فاصله اقلیدسی، DTW و بزرگ‌ترین زیر دنباله مشترک (LCSS) استفاده نمود [۳۴-۳۶]. با توجه به بالا بودن هزینه زمانی معیارهای DTW و LCSS در این پژوهش فاصله اقلیدسی استفاده شده است [۳۷-۳۹]. 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) قابل‌محاسبه می‌باشد. البته باید توجه داشت که این فاصله صرفاً هنگامی می‌تواند مورداستفاده قرار بگیرد که دو سری زمانی دارای طول یکسان باشند [۳۴].

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

۳-۲ خوشه‌بندی ترکیبی

خوشه‌بندی ترکیبی یکی از موضوعات مطرح‌شده در حوزه خوشه‌بندی می‌باشد که به‌ندرت در حوزه سری‌های زمانی موردتوجه قرار گرفته است. ازآنجایی‌که همه روش‌های خوشه‌بندی لزوماً برای یک مجموعه‌داده دارای نتایج مطلوب نمی‌باشد، لذا می‌توان به‌طور 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[۴۰].



شکل ۴: خوشه‌بندی ترکیبی

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 به‌منظور 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 ایجادشده برای هر شیء محاسبه می‌گردد و سپس از این فواصل و استفاده از یک الگوریتم خوشه‌بندی به‌منظور 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 ایجادشده بر اساس شباهت درون خوشه‌ای و تمایز بین خوشه‌ای می‌باشد. جدول ۱ سه معیار مورداستفاده در این پژوهش را نمایش می‌دهد. در این روابط n و k به ترتیب نشان‌دهنده تعداد اشیاء و خوشه‌ها می‌باشند، همچنین x و y نیز اشیاء خاص داخل خوشه c را نمایش می‌دهند [۴۲].

جدول ۱: معیارهای درونی ارزیابی خوشه‌بندی

ردیف	معیار	نماد	رابطه ریاضی	مقدار مطلوب
۱	واریانس بین گروهی[۱][۱۳]	V	$\frac{1}{n-k}\sum_{i=1}^{k}\sum_{x\in c_i} dist\left(x,c_i\right)$	کمینه

[۱] Inter-group Variance

۲	دان[۱][۴۳]	D	$\min_i \left\{ \min_j \left(\frac{\min_{x \in c_i, y \in c_j} \left(dist(x,y) \right)}{\max_k \left\{ \max_{x,y \in c_k} \left(dist(x,y) \right) \right\}} \right) \right\}$	بیشینه
۳	کالینسکی-هاراباز[۲][۴۴]	CH	$\frac{n-k}{n-1} \frac{\sum_{i=1}^{k} n_i dist^2(x, c_i)}{\sum_{i=1}^{k} \sum_{x \in c_i} dist^2(x, c_i)}$	بیشینه

۲-۵ معیارهای بیرونی

معیارهای بیرونی با فرض مشخص بودن برچسب اشیاء به‌منظور بررسی دقت الگوریتم مورداستفاده قرار می‌گیرند. در این مطالعه به‌منظور 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[۳] نشان‌دهنده تعداد اشیا می‌باشد که در کلاس و خوشه یکسانی قرار دارند. TN[۴] تعداد اشیا هستند که در خوشه‌های متفاوت قرار دارند و کلاس آن‌ها نیز متفاوت می‌باشد.FP[۵] نشان‌دهنده اشیایی می‌باشد که دارای خوشه‌بندی متفاوت می‌باشند درحالی‌که کلاس آن‌ها یکسان باشد و سرانجام FN[۶] نشان‌دهنده تعداد اشیایی است که دارای خوشه یکسان و کلاس‌های متفاوت می باشند.

۲-۶ الگوریتم‌های پیشنهادی

در این پژوهش دو الگوریتم بر پایه قطعه‌بندی با اندازه ثابت و خوشه‌بندی ترکیبی به‌منظور خوشه‌بندی سری‌های زمانی ارائه‌شده است.

۲-۶-۱ الگوریتم اول توسعه داده شده

این الگوریتم دارای ۵ گام به شرح زیر می‌باشد.

گام اول: یک مجموعه‌داده سری زمانی به قطعات با طول مساوی تقسیم می‌گردد. تعداد قطعات بین ۲ تا k قطعه می‌باشد.

گام دوم: برای هر یک از قطعات ایجادشده و برای هر مقدار k با استفاده از الگوریتم خوشه‌بندی سلسله مراتبی، خوشه‌بندی انجام می گیرد. در این گام با توجه به اینکه مجموعه‌داده سری زمانی به ۲ تا k قطعه تبدیل شده‌است، به ترتیب تعداد ۲ تا k برچسب خوشه‌بندی نیز وجود خواهد داشت.

گام سوم: در این گام نتایج خوشه‌بندی موجود با استفاده از یک الگوریتم خوشه‌بندی ترکیبی با یکدیگر ترکیب می‌گردد تا یک نتیجه واحد برای هر مرحله قطعه‌بندی از ۲ تا k قطعه ایجاد گردد. در پایان این گام، تعداد k-۱ برچسب خوشه‌بندی برای مجموعه‌داده سری زمانی وجود خواهد داشت. در انتهای این گام نتایج خوشه‌بندی مجموعه‌داده بدون قطعه‌بندی نیز به این تعداد اضافه می‌گردد و تعداد برچسب‌های موجود برابر با k خواهد بود.

گام چهارم: در ادامه با استفاده از یک معیار داخلی مشخص تعداد m برچسب از میان k برچسب ایجادشده، انتخاب می‌شوند ($m \leq k$).

گام پنجم: در گام نهایی m برچسب حاصل‌شده مجدداً با استفاده از یک الگوریتم خوشه‌بندی ترکیبی با یکدیگر ترکیب می‌گردند و یک برچسب خوشه‌بندی نهایی حاصل می‌گردد.

انتخاب مقدار k و m می‌تواند تا حدودی پیچیده باشد اما به نظر می‌رسد می‌توان مقدار k را طوری انتخاب نمود که برای سری‌های زمانی طولانی، بزرگ نباشد و همچنین برای سری‌های زمانی با طول کوتاه نیز بتواند ساختار سری‌زمانی را برای قطعه‌ها حفظ نماید. با توجه به این موارد به نظر می‌رسد استفاده از یک رابطه لگاریتمی 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 شده‌است. در این رابطه، L برابر با طول سری‌زمانی می‌باشد. با توجه به نتایج عددی به دست آمده برای پارامتر m ، مقادیر عددی ۲ یا ۳ مناسب به نظر می‌رسد.

$$k = \log_{10} L \quad (۵)$$

---

۱ Dunns Indices

۲ Calinski-Harabasez

۳ True positive

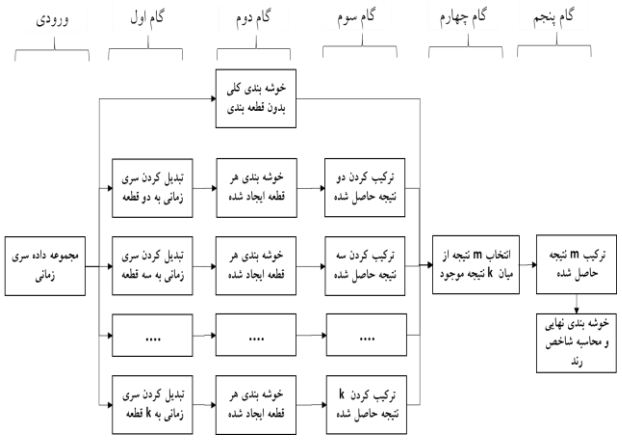
۴ True negative

۵ False positive

۶ False negative

برای انتخاب m برچسب از k برچسب نهایی برای هر سری زمانی، از معیارهای درونی ارزیابی خوشه‌بندی می‌توان استفاده نمود. البته از آنجایی که این معیارها نیز متنوع می‌باشند، می‌توان از سه معیار واریانس بین‌گروهی، کالینسکی- هارباز و شاخص دان که در ادبیات موضوع بیشتر مورداستفاده قرارگرفته‌اند، استفاده نمود. با توجه به اینکه استفاده از واریانس بین گروهی در الگوریتم اول برای نتیجه‌گیری کفایت می‌کند از معیارهای دیگر استفاده نشده است.

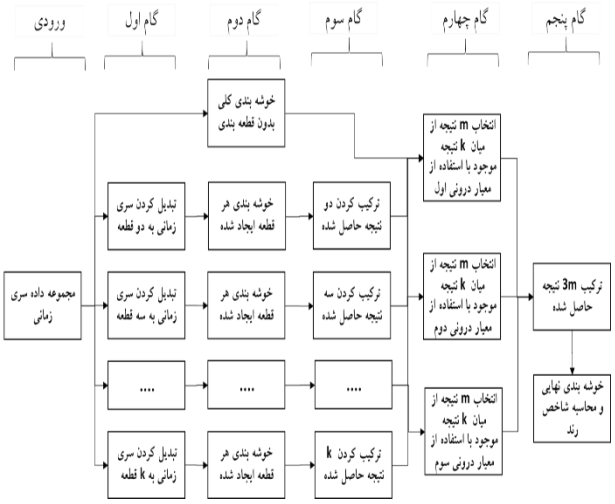
در الگوریتم اول صرفاً از یک معیار برای انتخاب m برچسب از k برچسب ایجادشده، استفاده می‌گردد. الگوریتم اول پیشنهادی در شکل ۵ نمایش داده شده است.



شکل ۵: الگوریتم اول پیشنهادی

۲-۶-۲ الگوریتم دوم توسعه داده شده

الگوریتم دوم نیز مانند الگوریتم اول دارای ۵ گام می‌باشد که سه گام نخست این الگوریتم نیز مانند الگوریتم اول می‌باشد. تفاوت این الگوریتم در گام سوم ایجاد می‌گردد که به نحوه انتخاب m نتیجه مرتبط می‌باشد. در گام چهارم این الگوریتم از سه معیار درونی معرفی شده در بخش ۲-۶-۱ به‌طور هم‌زمان به‌منظور انتخاب m نتیجه استفاده می‌گردد. در مرحله پنجم و نهایی نیز مطابق الگوریتم قبلی تعداد $3m$ برچسب نهایی ایجادشده با استفاده از خوشه‌بندی ترکیبی با یکدیگر ترکیب می‌گردند. در مرحله نهایی در الگوریتم اول ما دارای m برچسب نهایی خواهیم بود اما در این الگوریتم تعداد برچسب‌ها $3m$ می‌باشد. باید توجه داشت که تعدادی از برچسب‌ها می‌تواند تکراری باشد. الگوریتم دوم در شکل ۶ نمایش داده شده است.



شکل ۶: الگوریتم دوم پیشنهادی

۳- اجرای مدل

به‌منظور بررسی کارایی الگوریتم‌های ارائه شده، این الگوریتم‌ها روی ۸۲ مجموعه‌داده از آرشیو UCR اجرا شده است [۴۵]. مجموعه‌داده موردنظر زمینه‌های مختلفی مانند سلامت، مالی، رسانه و مهندسی را پوشش می‌دهد. با توجه به اینکه مجموعه‌داده‌های موردنظر برای کلاس‌بندی سری‌های زمانی ارائه‌شده‌اند و دارای دو بخش آموزش و آزمون می‌باشند، برای خوشه‌بندی با یکدیگر تجمیع شده‌اند. پارامترهای ورودی برای هر الگوریتم شامل الگوریتم خوشه‌بندی استفاده‌شده، نوع پیوند، تعداد معیارهای درونی استفاده‌شده، معیار درونی استفاده‌شده، m و k می‌باشد. همچنین خروجی مدل نیز شامل سری‌های زمانی قرارگرفته در هر خوشه و همچنین شاخص رند نهایی می‌باشد.

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

First 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

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
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

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

جدول ۲: پارامترهای الگوریتم‌های مختلف خوشه‌بندی

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

۳-۱ نتایج

به‌منظور بررسی الگوریتم‌های ارائه‌شده نتایج حاصل روی ۸۲ مجموعه‌داده معرفی‌شده مورد مقایسه قرار گرفته است. الگوریتم‌ها برای هر یک از مجموعه‌داده‌ها ۱۰ بار اجرا شده است و نتایج شاخص رند و زمان اجرا برای میانگین ۱۰ تکرار در جدول ۳ گزارش شده است. مجموعه‌داده‌ها با سایز بزرگ، زمان‌های اجرای طولانی در مقایسه با سایر مجموعه‌داده‌ها دارند.

جدول ۳: نتایج شاخص رند و زمان اجرا برای دو الگوریتم ارائه‌شده

مجموعه‌داده	شاخص رند				زمان اجرا (ثانیه)			
	الگوریتم اول وضعیت اول	الگوریتم اول وضعیت دوم	الگوریتم اول وضعیت سوم	الگوریتم دوم	الگوریتم اول وضعیت اول	الگوریتم اول وضعیت دوم	الگوریتم اول وضعیت سوم	الگوریتم دوم
ITA	۵۰,۰۰٪	۵۱,۳۰٪	۵۰,۱۰٪	۵۰,۸۰٪	۳,۳۷	۴,۱۳	۸,۳۴	۷,۴
SYN	۸۰,۵۰٪	۷۹,۰۰٪	۸۱,۳۰٪	۸۲,۰۰٪	۱,۹۴	۲,۰۸	۳,۵۷	۵,۶۲
SO۲	۵۴,۳۰٪	۵۴,۳۰٪	۶۷,۵۰٪	۶۳,۸۰٪	۴,۹۵	۶,۲۲	۱۰,۵۹	۹,۹
SO۱	۴۹,۹۰٪	۵۰,۱۰٪	۵۲,۶۰٪	۵۰,۱۰٪	۲,۱۷	۲,۶۴	۳,۹۶	۴,۶۶
DPA	۷۱,۲۰٪	۷۱,۲۰٪	۷۳,۷۰٪	۷۲,۹۰٪	۱,۷۱	۱,۸۳	۲,۸۱	۳,۶۴
DPC	۵۳,۰۰٪	۵۳,۰۰٪	۵۲,۴۰٪	۵۳,۵۰٪	۴,۳۱	۴,۹۲	۷,۸۶	۷,۱۳
DPT	۸۴,۳۰٪	۸۱,۵۰٪	۸۶,۳۰٪	۸۷,۸۰٪	۱,۵۸	۱,۸۷	۳,۱۸	۳,۳۲
MPA	۷۰,۴۰٪	۷۰,۵۰٪	۷۲,۸۰٪	۷۳,۰۰٪	۱,۷۴	۲,۳۳	۳,۱	۴,۹۲
MPC	۵۳,۳۰٪	۵۳,۲۰٪	۵۰,۰۰٪	۵۰,۰۰٪	۴,۴۱	۵,۴۱	۷,۴۹	۸,۷۳
MPT	۸۲,۳۰٪	۸۱,۸۰٪	۸۳,۹۰٪	۸۳,۸۰٪	۱,۵	۱,۷۲	۲,۹۸	۴,۸۷
PHA	۵۴,۰۰٪	۵۴,۰۰٪	۵۲,۱۰٪	۵۴,۲۰٪	۴۵,۱۶	۴۹,۴۵	۸۴,۶۹	۷۷,۶۲
PPA	۷۷,۴۰٪	۷۷,۳۰٪	۷۹,۷۰٪	۷۸,۹۰٪	۲,۲۶	۲,۱۹	۳,۵۷	۵,۷۷
PPC	۵۳,۹۰٪	۵۳,۸۰٪	۵۷,۶۰٪	۵۲,۱۰٪	۴,۳۷	۴,۹۶	۷,۳۷	۸,۱۹
PPT	۸۳,۹۰٪	۸۳,۸۰٪	۷۹,۸۰٪	۷۸,۶۰٪	۱,۸۶	۲,۰۵	۳,۷۵	۵,۵۲
TWE	۵۰,۱۰٪	۵۰,۰۰٪	۵۰,۱۰٪	۵۰,۳۰٪	۷,۰۵	۸,۰۴	۱۴,۴۴	۱۴,۰۳
MOT	۵۰,۳۰٪	۵۰,۳۰٪	۶۳,۱۰٪	۶۱,۸۰٪	۹,۷۱	۱۰,۸۲	۲۰,۹۹	۱۷,۴۸
EC۲	۶۲,۳۰٪	۵۴,۰۰٪	۶۱,۲۰٪	۶۲,۳۰٪	۰,۶۶	۰,۷۵	۱,۱۳	۱,۰۵
MED	۶۰,۲۰٪	۶۱,۳۰٪	۶۴,۶۰٪	۶۳,۴۰٪	۱۰,۱۱	۱۰,۸۷	۱۵,۵	۲۱,۵۵

CBF	۶۰٫۶۰٪	۶۳٫۱۰٪	۶۴٫۷۰٪	۶۶٫۰۰٪	۶٫۲۵	۶٫۹	۱۰٫۵۵	۱۰٫۹۴
SWE	۳۹٫۰۰٪	۳۷٫۹۰٪	۸۸٫۷۰٪	۸۷٫۲۰٪	۱۲٫۰۶	۱۳٫۶	۱۸٫۶۷	۲۳٫۴۲
TWP	۶۲٫۶۰٪	۶۲٫۱۰٪	۶۲٫۹۰٪	۶۲٫۸۰٪	۲۴۰٫۷۳	۲۵۸٫۳۶	۳۸۳٫۸۵	۳۶۱٫۴۶
FAA	۷۳٫۷۰٪	۸۱٫۶۰٪	۸۵٫۰۰٪	۸۳٫۳۰٪	۳۹٫۶۳	۷۵٫۵	۶۲٫۱	۶۳٫۱
ECF	۴۹٫۹۰٪	۵۰٫۳۰٪	۵۰٫۵۰٪	۵۰٫۰۰٪	۶٫۵۳	۷٫۹۵	۱۰٫۷۷	۱۰٫۵۵
EC۵	۸۹٫۶۰٪	۹۰٫۶۰٪	۷۴٫۶۰٪	۷۴٫۱۰٪	۲۵۰٫۰۶	۲۷۸٫۷۱	۴۶۳٫۸۳	۳۲۸٫۳۳
PLA	۹۲٫۸۰٪	۹۲٫۵۰٪	۹۴٫۰۰٪	۹۴٫۰۰٪	۱٫۰۵	۱٫۰۵	۱٫۸۷	۲٫۹۹
GUN	۴۹٫۷۰٪	۵۰٫۹۰٪	۴۹٫۷۰٪	۵۳٫۷۰٪	۰٫۶۶	۱٫۹۸	۱٫۲۷	۱٫۶
WAF	۵۳٫۴۰٪	۵۳٫۴۰٪	۵۳٫۴۰٪	۵۳٫۴۰٪	۵۸۰٫۶۹	۶۱۸٫۹۸	۱۳۹۵٫۹۴	۱۵۷۱٫۷۶
CHL	۵۰٫۵۰٪	۵۱٫۵۰٪	۵۲٫۲۰٪	۵۳٫۱۰٪	۲۱۰٫۲۹	۲۰۸٫۳۱	۲۹۰٫۳۴	۲۸۱٫۷۲
ADI	۴۷٫۵۰٪	۴۸٫۷۰٪	۹۲٫۸۰٪	۹۲٫۲۰٪	۷٫۵۷	۸٫۲۵	۱۱٫۸۲	۱۱٫۵۶
کوتاه	۶۲٫۴۳٪	۶۲٫۵۲٪	۶۷٫۱۵٪	۶۶٫۸۷٪	۵۰	۵۵	۹۸	۹۹
WIN	۴۹٫۹۰٪	۴۹٫۹۰٪	۴۹٫۸۰٪	۴۹٫۹۰٪	۰٫۹۳	۲٫۵۲	۳٫۸۲	۵٫۳۷
STR	۵۱٫۵۰٪	۵۱٫۵۰٪	۵۰٫۴۰٪	۵۰٫۴۰٪	۱۳٫۶۸	۱۵٫۱۳	۲۲٫۹۱	۲۴٫۰۱
ARR	۳۵٫۸۰٪	۳۵٫۳۰٪	۶۶٫۸۰٪	۶۰٫۵۰٪	۱٫۱۷	۱٫۵۷	۱٫۸۷	۱٫۸۹
INS	۸۶٫۵۰٪	۸۶٫۸۰٪	۸۶٫۰۰٪	۸۶٫۸۰٪	۶۰٫۳۲	۶۱٫۴	۹۶٫۶	۱۱۸٫۸۷
W۵۰	۹۴٫۶۰٪	۹۴٫۴۰٪	۹۴٫۶۰٪	۹۴٫۵۰٪	۱۰٫۲۳	۹٫۴۲	۲۲٫۶۳	۳۲٫۸۵
WOS	۸۳٫۹۰٪	۸۷٫۶۰٪	۸۸٫۷۰٪	۸۸٫۶۰٪	۱۱٫۸۶	۱۴٫۰۳	۱۹٫۹۲	۵۴٫۰۸
TRA	۷۵٫۰۰٪	۷۴٫۹۰٪	۷۵٫۰۰٪	۷۵٫۱۰٪	۱٫۱۸	۳٫۱۶	۱٫۹۲	۳٫۲۴
TO۱	۴۹٫۹۰٪	۵۰٫۱۰٪	۴۹٫۹۰٪	۵۰٫۰۰٪	۱٫۳۶	۱٫۵۳	۲٫۳۱	۳٫۰۶
COF	۴۹٫۸۰٪	۴۹٫۱۰٪	۷۰٫۱۰٪	۷۹٫۵۰٪	۰٫۶۵	۱٫۰۷	۰٫۹۶	۰٫۸۶
CRX	۷۱٫۲۰٪	۷۷٫۹۰٪	۸۱٫۶۰٪	۸۳٫۴۰٪	۹٫۴۸	۹٫۴	۱۳٫۲۹	۱۴٫۷۹
CRY	۷۷٫۴۰٪	۷۷٫۴۰٪	۸۲٫۴۰٪	۸۴٫۰۰٪	۹٫۳۲	۹٫۹۲	۱۳٫۱۵	۱۷٫۲۶
CRZ	۶۹٫۶۰٪	۷۷٫۱۰٪	۸۴٫۶۰٪	۸۴٫۳۰٪	۸٫۴۸	۹٫۳۵	۱۳٫۴۸	۱۵٫۷
UWX	۷۱٫۵۰٪	۷۳٫۶۰٪	۸۱٫۹۰٪	۸۱٫۴۰٪	۳۲۶٫۹۸	۳۳۶٫۹۳	۴۷۱٫۹۶	۵۰۱٫۱۵
UWY	۸۰٫۵۰٪	۷۹٫۸۰٪	۸۲٫۲۰٪	۸۲٫۱۰٪	۳۰۳٫۸۷	۳۲۳٫۱۷	۴۴۷٫۹۶	۵۰۹٫۲۳
UWZ	۷۹٫۹۰٪	۸۳٫۳۰٪	۸۱٫۹۰٪	۸۱٫۹۰٪	۳۰۹٫۹۶	۳۳۰٫۶۵	۴۶۰٫۵۵	۵۸۹٫۵۳
LI۷	۷۴٫۱۰٪	۷۵٫۸۰٪	۷۱٫۴۰٪	۷۵٫۶۰٪	۱٫۱۵	۱٫۳۵	۲٫۱۴	۲٫۷۸
TO۲	۴۹٫۷۰٪	۴۹٫۹۰٪	۴۹٫۷۰٪	۵۱٫۱۰٪	۰٫۹۲	۱٫۰۵	۱٫۵۹	۲٫۰۲
DIA	۳۰٫۶۰٪	۳۰٫۶۰٪	۲۹٫۶۰٪	۳۰٫۶۰٪	۲٫۳۹	۲٫۵۷	۳٫۱۹	۴٫۱۵
FAF	۷۲٫۱۰٪	۷۳٫۸۰٪	۶۹٫۲۰٪	۶۹٫۵۰٪	۱٫۰۲	۱٫۵	۱٫۴۴	۱٫۶
SYM	۸۹٫۹۰٪	۸۹٫۹۰٪	۸۹٫۶۰٪	۸۹٫۲۰٪	۱۹٫۶۳	۱۸٫۸۶	۲۷٫۴۷	۴۰٫۴۵
YOG	۵۰٫۰۰٪	۵۰٫۰۰٪	۵۰٫۱۰٪	۵۰٫۰۰٪	۲۷۲٫۷۷	۲۹۵٫۰۱	۴۲۰٫۳۷	۴۸۱٫۲۲
OSU	۶۷٫۸۰٪	۶۲٫۷۰٪	۷۱٫۴۰٪	۷۳٫۳۰٪	۴٫۸۹	۵٫۳۵	۶٫۸۷	۱۲٫۰۴
HAM	۵۰٫۱۰٪	۵۰٫۲۰٪	۵۱٫۵۰٪	۵۱٫۴۰٪	۱٫۶۸	۲٫۰۵	۲٫۴۹	۷٫۱۳
MEA	۷۲٫۸۰٪	۷۷٫۰۰٪	۷۲٫۰۰٪	۷۲٫۰۰٪	۱٫۱۲	۱٫۲۳	۱٫۶۶	۲٫۶۲
FIS	۶۴٫۷۰٪	۷۱٫۶۰٪	۷۲٫۰۰٪	۷۱٫۴۰٪	۳٫۷۸	۴٫۲۹	۵٫۴۸	۶٫۴۴
BEE	۶۵٫۶۰٪	۶۵٫۵۰٪	۶۳٫۸۰٪	۶۳٫۸۰٪	۱٫۰۴	۱٫۱۵	۳٫۵۹	۱٫۵۲
FOA	۵۰٫۲۰٪	۵۰٫۲۰٪	۵۰٫۰۰٪	۵۰٫۰۰٪	۷۹۶٫۰۲	۱۰۴۸٫۸۶	۹۹۲٫۹۸	۱۱۱۹٫۴۹
FOB	۵۰٫۰۰٪	۵۰٫۰۰٪	۵۰٫۰۰٪	۵۰٫۰۰٪	۶۵۹٫۷۴	۶۹۴٫۹۴	۸۰۷٫۴۳	۹۵۸٫۵۸

SHS	٥١,٤٠٪	٥١,٤٠٪	٥٠,٠٠٪	٤٩,٨٠٪	١,٦٧	٢,٠٧	٢,٥	٣,١٤
متوسط	٦٤,٣٤٪	٦٥,٤٢٪	٦٧,٨٠٪	٦٨,٢٨٪	٩٨	١١١	١٣٤	١٥٦
BFL	٥٥,٤٠٪	٥٦,٩٠٪	٥١,٧٠٪	٥٠,٩٠٪	٠,٨٧	٠,٨٢	١,٢٧	١,١٦
BIR	٤٨,٩٠٪	٤٩,٢٠٪	٤٨,٨٠٪	٥٣,٣٠٪	٠,٧٦	٠,٧٧	١,٥٨	١,٢
EAR	٥٦,٨٠٪	٥٦,٨٠٪	٥٢,٤٠٪	٥٢,٣٠٪	٦,٩٤	٧,٥٦	٩,٦	١٠,٠٦
HER	٥١,٧٠٪	٥١,٧٠٪	٥٠,٦٠٪	٥٠,٥٠٪	١,٤٤	١,١٩	١,٦٣	١,٩٧
SHA	٨٣,٥٠٪	٨٤,٤٠٪	٩٦,١٠٪	٩٤,٨٠٪	٣٦,٢١	٣٦,٩	٥٣,٣	٦٨,٨
OLI	٨٨,٧٠٪	٨٦,١٠٪	٨٦,٠٠٪	٨٥,٨٠٪	١,١٤	١,٠٦	١,٦٤	٢,٣٦
CAR	٦٥,٣٠٪	٥٩,٥٠٪	٦٣,٣٠٪	٦١,٣٠٪	١,٢٦	١,٤	٢,٣٣	١,٨٥
LI٢	٥٧,٩٠٪	٥٨,٥٠٪	٥٠,٤٠٪	٦٠,٠٠٪	١,٢٣	١,٣٧	١,٧٧	٢,١٨
COM	٤٩,٩٠٪	٤٩,٩٠٪	٥٠,٣٠٪	٥٠,٤٠٪	١٠,٥٢	١٠,٤٤	١٢,٤٧	١٤,٢
LAR	٥٤,١٠٪	٥٤,٢٠٪	٥٠,٩٠٪	٥٢,٢٠٪	٢١,١٩	٢٢,٠١	٢٧,٣٧	٣٨,٦٩
REF	٤٥,٨٠٪	٤٥,٥٠٪	٥٤,٩٠٪	٥٥,٢٠٪	٢١,٦	٢٢,٨٤	٢٧,١٨	٣٥,٧٨
SCR	٤٠,٣٠٪	٤٠,٢٠٪	٥٥,٠٠٪	٥٥,٢٠٪	٢٢,٨	٢٢,٦٤	٢٩,١٦	٣٧,٣٥
SMA	٣٣,٨٠٪	٣٤,٣٠٪	٤٦,٦٠٪	٤٣,٠٠٪	٢٢,٦٢	٢٣,٢٧	٣١,٠١	٣٨,٩
NO١	٨٦,٧٠٪	٨٥,٤٠٪	٩٥,١٠٪	٩٤,٦٠٪	٥٣٢,٨١	٥٥٩,٣٧	٦٣٠,٧٤	٩٩٢,١٨
NO٢	٩١,٣٠٪	٩١,١٠٪	٩٥,٩٠٪	٩٥,٦٠٪	٥٠٤,٨٨	٤٩١,٢٦	٦٤٤,١٩	٩٨٠,٥٢
WOR	٦٢,٧٠٪	٥٩,٦٠٪	٦٣,٠٠٪	٦٣,٥٠٪	٥,٧١	٥,١٥	٦,٠٨	١٢,٩
WOT	٤٩,٧٠٪	٤٩,٩٠٪	٤٩,٣٠٪	٤٩,٨٠٪	٤,٧٩	٤,١١	٥,٩٨	٨,٧
UWA	٨٧,٦٠٪	٨٥,٩٠٪	٨٦,٣٠٪	٨٥,٦٠٪	٨٦٥,٩٨	٩٠٧,٤٩	١١٨٠,٠٢	١٧٥٣,٩٢
MAL	٩٠,٣٠٪	٩٠,٢٠٪	٩٢,٢٠٪	٩٣,٩٠٪	٢٤٠	٢٥٣,١٧	٣٣٦,٢١	٥٠٩,٠٦
PHO	٩٠,١٠٪	٩٠,١٠٪	٩٢,٠٠٪	٩٢,٦٠٪	٢٨١,٧	٢٧٧,٢٨	٣٣٤,١٢	٤٣٢,٧٥
HAP	٣٦,٦٠٪	٣٩,٤٠٪	٦٧,١٠٪	٦٧,١٠٪	١٥,٧٩	١٤,٨٧	١٦,٥١	٢٥,٦٤
CIN	٦١,٤٠٪	٦١,٨٠٪	٦٦,١٠٪	٦٣,٨٠٪	١٦٧,٩٣	١٧٠,٣٩	١٩٠,٥٧	٢٦٢,٥٦
INL	٦١,٧٠٪	٦٤,٦٠٪	٦٦,٣٠٪	٦٨,٠٠٪	٤١,٩٧	٤٣,١٨	٥٣,٨٥	٨١,٨٨
HAN	٥٧,١٠٪	٥٤,٥٠٪	٥٠,٠٠٪	٥٦,١٠٪	٣٤٤,٦	٣٦٣,٦٩	٤٣٩,٨٢	٥٥٢,٠١
بلند	٦٢,٨٠٪	٦٢,٤٩٪	٦٥,٨٥٪	٦٦,٤٨٪	١٣١	١٣٥	١٦٨	٢٤٤
متوسط	٦٣,٢٢٪	٦٣,٥٤٪	٦٧,٠٠٪	٦٧,٢٪	٩٠,٩٣	٩٨,٢١	١٣١,١٣	١٦١

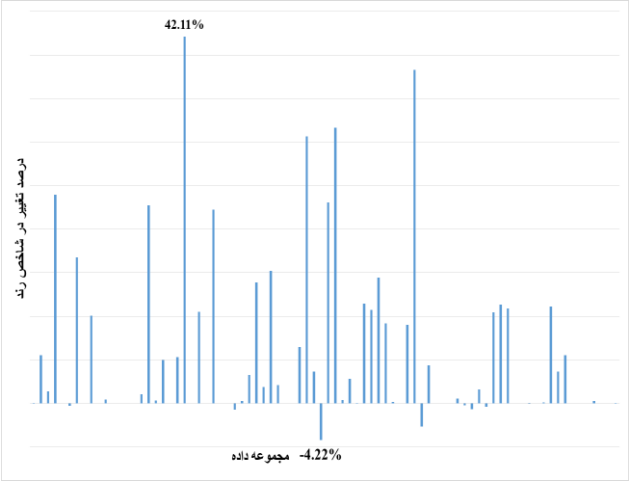
جدول ۴ به‌صورت مختصر و با توجه به معیارهای مختلف عملکرد دو الگوریتم ارائه‌شده را نمایش می‌دهد منظور از الگوریتم اولیه در این جدول، خوشه‌بندی بدون قطعه‌بندی با استفاده از الگوریتم سلسه‌مراتبی می‌باشد.

جدول ۴: نتایج الگوریتم‌های مختلف خوشه‌بندی

معیار	الگوریتم اول			الگوریتم دوم
	وضعیت اول	وضعیت دوم	وضعیت سوم	
میانگین شاخص رند	٦٣,٢٢٪	٪٦٣,٥۴	٦٧,٠٠٪	٦٧,٢٥٪
انحراف از معیار	٠,٠٠٥٨	٠,٠١٠۴	٠,٠٠٩١	٠,٠٠٦٥
میانگین بهبود شاخص رند نسبت به حالت اولیه	٥,۴۴٪	٪٥,٧۴	٢,٦٩٪	٢,٩٢٪
حداکثر مقدار بهبود در یک مجموعه‌داده	۴٢,١١٪	٪۴٣,١٢	٣٧,٩۶٪	٣٧,١٣٪
حداکثر میزان تنزل شاخص رند در یک مجموعه‌داده	-۴,٢٢٪	-٣,٠٠٪	-١٠,٢٧٪	-١٠,٧٨٪
نسبت مجموعه‌داده با بهبود شاخص رند	٥٩٪	٪٧٣	٥٥٪	٥٩٪
نسبت مجموعه‌داده بدون تغییر شاخص رند	٣٢٪	٪١٠	١٠٪	٣٠٪

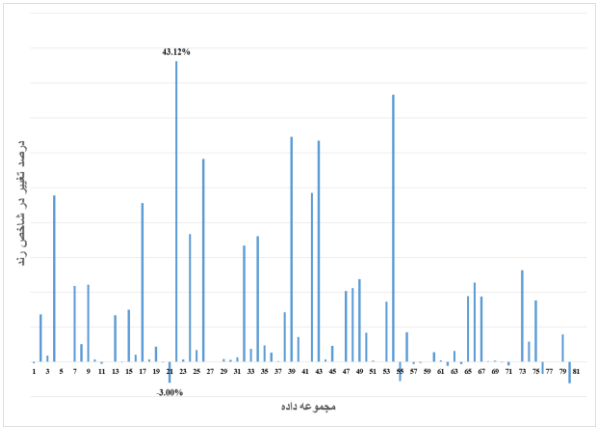
نسبت مجموعه‌داده با افت شاخص رند	۹٪	٪۱۷	۳۵٪	۱۱٪
نسبت ایجاد بهترین پاسخ در پژوهش‌های قبلی	۱۸٪	۲۰٪	۳۰٪	۳۲٪

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 پیدا نموده‌است، همین‌طور در ۳۰ درصد حالات، الگوریتم مورداشاره توانسته‌است بهترین شاخص رند را در بین ادبیات موضوع داشته‌باشد. همچنین الگوریتم اول وضعیت اول توانسته‌است در ۵۹ درصد مجموعه‌داده‌ها 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 مجموعه‌داده بهبود دهد و از سوی دیگر می‌تواند نتایج اولیه خوشه‌بندی را تا ۴ درصد با افت روبرو نماید.



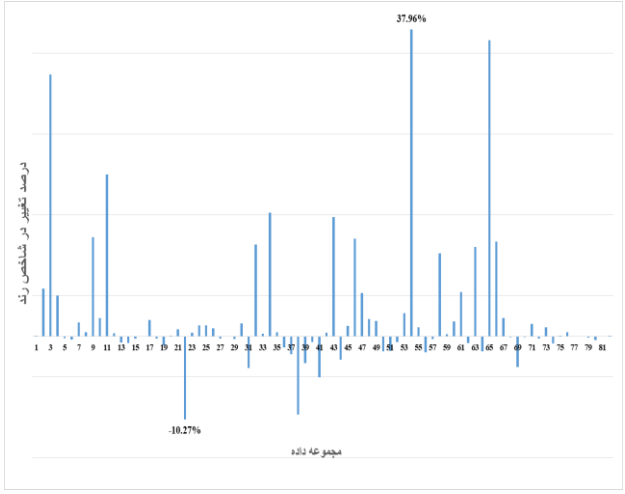
شکل ۸: اختلاف میانگین شاخص رند الگوریتم اول وضعیت اول

وضعیت دوم الگوریتم اول با مقدار پارامتر $m=3$ اجرا شده است و مانند حالت قبل برای الگوریتم خوشه‌بندی سلسله مراتبی، فاصله برابر با میانگین در نظر گرفته شده است. 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 داشته است و در ۷۳ درصد موارد نیز جواب خوشه‌بندی ابتدایی را بهبود داده است، همچنین در ۱۰ درصد مجموعه‌داده نیز این مقدار بدون تغییر باقی مانده است. بر اساس اطلاعات شکل ۷ در وضعیت دوم الگوریتم اول در بهترین حالت خود توانسته است تا ۴۳ درصد نتایج اولیه خوشه‌بندی را برای یک مجموعه‌داده خاص افزایش دهد و از طرفی در بدترین حالت نیز باعث کاهش ۳ درصدی شاخص رند برای یک مجموعه‌داده شده است.



شکل ۹: اختلاف میانگین شاخص رند الگوریتم اول وضعیت دوم

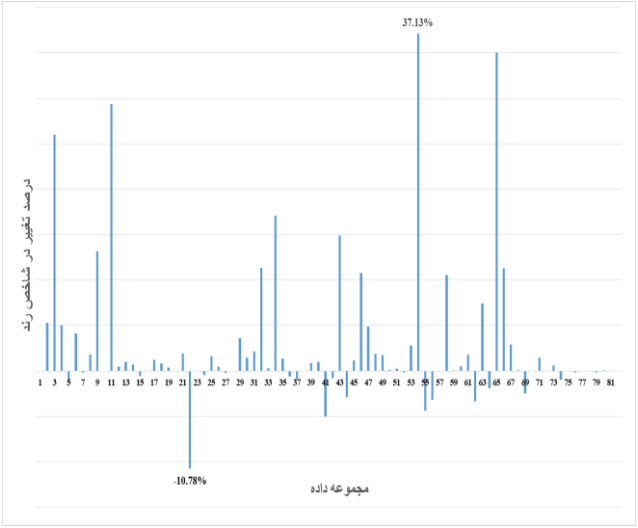
وضعیت سوم الگوریتم اول با پارامتر $m=3$ و فاصله حداکثر برای الگوریتم خوشه‌بندی سلسله مراتبی اجرا شده است. با توجه به اطلاعات جدول ۴ می‌توان مشاهده نمود که میانگین شاخص رند در ۸۲ مجموعه‌داده به ۶۷ درصد رسیده است و نتایج اولیه خوشه‌بندی در این وضعیت ۲,۶۹ درصد به‌طور میانگین افزایش داشته است. همچنین در ۳۰ درصد حالات، وضعیت سوم الگوریتم اول توانسته است بهترین شاخص رند را در بین ادبیات موضوع داشته باشد. بر اساس اطلاعات شکل ۸، وضعیت سوم الگوریتم اول در بهترین و بدترین حالت به ترتیب توانسته است ۳۸ درصد نتیجه اولیه را بهبود ببخشد و ۱۰ درصد نتیجه اولیه را ضعیف‌تر نماید.



شکل ۱۰: اختلاف میانگین شاخص رند الگوریتم اول وضعیت سوم

در انتها الگوریتم دوم با پارامتر $m=2$ و فاصله حداکثر برای الگوریتم خوشه‌بندی سلسله مراتبی اجرا شده است. در الگوریتم دوم همان‌گونه که قبلاً بیان‌شده است از سه معیار داخلی به‌صورت هم‌زمان استفاده‌شده است. با توجه به اطلاعات جدول ۴ الگوریتم

دوم توانسته است ۲,۹ درصد به‌صورت میانگین دقت خوشه‌بندی را افزایش دهد دقت شاخص رند را به ۶۷,۲۵ درصد برساند که بهترین نتیجه ممکن می‌باشد. در الگوریتم دوم نیز نتایج مربوط به ۶۰ درصد مجموعه‌داده‌ها به‌طور میانگین بهبود داشته است و در ۳۰ درصد موارد نیز تغییر نداشته است. همین‌طور در ۳۲ درصد حالات، الگوریتم دوم توانسته است بهترین شاخص رند را در بین ادبیات موضوع داشته باشد. بر اساس اطلاعات شکل ۹ الگوریتم دوم در بهترین حالت توانسته است ۳۷ درصد نتیجه شاخص رند را برای یک مجموعه‌داده بهبود ببخشد و در بدترین حالت نیز نزدیک به ۱۰ درصد نتیجه اولیه را کاهش دهد.



شکل ۱۱: اختلاف میانگین شاخص رند الگوریتم دوم

همچنین به‌منظور تحلیل دقیق‌تر نتایج به دست آمده مجموعه داده‌های مورداستفاده 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 ارائه‌شده 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 ارائه‌شده بسیار ناچیز می‌باشد که این نشان دهند پایداری خوب طرح ارائه شده می‌باشد.

۳-۲ مقایسه الگوریتم‌ها

در این قسمت آزمون آماری به‌منظور بررسی بهترین الگوریتم در دو مرحله انجام گرفته است در مرحله اول الگوریتم‌های ارائه‌شده با یکدیگر مقایسه شده‌اند و در گام بعدی بهترین الگوریتم ارائه‌شده با چهار الگوریتم ادبیات در این حوزه مقایسه آماری شده‌است. یکی از روش‌های پیشنهادشده به‌منظور ارزیابی الگوریتم‌های خوشه‌بندی و طبقه‌بندی، استفاده از آزمون ویلکانسون می‌باشد[۴۶]. با توجه به مقدار P-value 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 می‌باشد. 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 ازلحاظ 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.

جدول ۵: مقادیر P-value الگوریتم‌های ارائه‌شده

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

به‌منظور بررسی دقت و سرعت الگوریتم‌های ارائه‌شده، بهترین نتایج که مربوط به الگوریتم دوم می‌باشد، با چهار الگوریتم در ادبیات موضوع بررسی‌شده است. چهار الگوریتم مذکور عبارت‌اند از DD_{DTW} که در آن از ترکیب دو معیار فاصله به همراه الگوریتم سلسله مراتبی استفاده‌شده است [۱۳]، الگوریتم KSC که در آن از الگوهای زمانی مشخص به‌منظور خوشه‌بندی استفاده‌شده است [۴۷] و الگوریتم‌های $TS3C_{\mathrm{MV}}$ و $TS3C_{\mathrm{CH}}$ که بر اساس مشخصات قطعه‌بندی سری‌های زمانی توسعه داده‌شده‌اند [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 ارائه‌شده عملکرد بهتری نسبت به این الگوریتم‌ها داشته‌است. در مورد مقایسه هزینه زمانی الگوریتم ارائه‌شده نیز با توجه به اطلاعات جدول ۶ می‌توان مشاهده نمود که الگوریتم ارائه‌شده دارای زمان اجرای بسیار کمتری می‌باشد. نتایج جدول ۷ نیز این ادعا را تائید می‌نماید. همچنین با توجه به اطلاعات جدول ۶ مشاهده می‌شود که الگوریتم ارائه‌شده در ۳۰ مورد از ۸۲ مجموعه‌داده‌ها توانسته است بهترین نتایج بین چهار الگوریتم را داشته باشد.

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

مجموعه داده	شاخص رند					زمان اجرا				
	الگوریتم دوم	TS^3C_{CH}	TS^3C_{MV}	DD_{DTW}	KSC	الگوریتم دوم	TS^3C_{CH}	TS^3C_{MV}	DD_{DTW}	KSC
50W	۹۵٪	۹۴٪	۹۴٪	۹۲٪	۶۶٪	۳۲,۸۵	۱۴۷۸	۱۵۰۳	۱۰۴۲۸۵	۱۷۶۶۷
ADI	۹۲٪	۹۲٪	۹۲٪	۶۸٪	۹۵٪	۱۱,۵۶	۸۸۷	۹۰۴	۳۲۲۲۱	۵۳۷۴
ARR	۶۰٪	۶۲٪	۶۳٪	۳۵٪	۶۳٪	۱,۸۹	۳۶۱	۳۶۲	۴۶۰۳	۸۴
BEE	۶۴٪	۶۸٪	۶۸٪	۵۸٪	۷۱٪	۱,۵۲	۱۸۸	۱۸۸	۱۲۴۷	۶۳
BFL	۵۱٪	۴۹٪	۴۹٪	۵۹٪	۵۰٪	۱,۱۶	۱۳۶	۱۳۶	۶۴۲	۲۱
BIR	۵۳٪	۴۹٪	۴۹٪	۵۰٪	۵۴٪	۱,۲۰	۱۱۷	۱۱۷	۶۸۸	۱۹
CAR	۶۱٪	۶۵٪	۶۵٪	۵۰٪	۶۸٪	۱,۸۵	۳۹۲	۳۹۲	۸۳۷۱	۱۴۲
CBF	۶۶٪	۶۷٪	۶۷٪	۷۸٪	۵۶٪	۱۰,۹۴	۹۷۶	۹۷۸	۲۴۸۳۰	۵۸۴
CHL	۵۳٪	۴۹٪	۴۷٪	۴۰٪	۵۳٪	۲۸۱,۷۲	۶۵۵۶	۶۶۰۱	۷۶۳۵۸۷	۲۶۲۳
CIN	۶۴٪	۶۴٪	۶۴٪	۵۶٪	۶۹٪	۲۶۲,۵۶	۱۱۶۴۸	۱۱۶۵۳	۷۶۳۵۸۷	۲۷۷۷۲
COF	۷۹٪	۵۱٪	۵۱٪	۴۹٪	۷۵٪	۰,۸۶	۹۸	۹۸	۳۸۵	۱۰
COM	۵۰٪	۵۰٪	۵۱٪	۵۰٪	۵۰٪	۱۴,۲۰	۲۶۹۹	۲۷۰۰	۲۱۰۱۱۳	۴۱۳
CRX	۸۳٪	۸۵٪	۸۵٪	۷۸٪	۴۱٪	۱۴,۷۹	۱۷۰۸	۱۷۱۵	۸۲۹۷۰	۴۳۵۶
CRY	۸۴٪	۸۴٪	۸۴٪	۶۹٪	۵۳٪	۱۷,۲۶	۱۷۵۲	۱۷۵۹	۸۳۷۱۰	۴۴۸۸
CRZ	۸۴٪	۸۴٪	۸۵٪	۷۱٪	۴۱٪	۱۵,۷۰	۱۶۸۰	۱۶۸۷	۷۸۵۶۳	۳۸۰۷
DIA	۳۱٪	۷۲٪	۷۲٪	۳۰٪	۹۶٪	۴,۱۵	۶۲۳	۶۲۴	۲۰۴۳۳	۳۰۳
DPA	۷۳٪	۶۰٪	۶۰٪	۷۱٪	۷۲٪	۳,۶۴	۳۱۲	۳۱۴	۳۰۹۱	۸۳
DPC	۵۴٪	۵۱٪	۵۱٪	۵۳٪	۵۰٪	۷,۱۳	۴۸۴	۴۸۷	۸۵۳۶	۴۵
DPT	۸۸٪	۶۸٪	۶۶٪	۸۶٪	۶۶٪	۳,۳۲	۳۰۳	۳۰۶	۳۲۳۱	۱۹۴
EAR	۵۲٪	۵۳٪	۵۳٪	۵۴٪	۶۲٪	۱۰,۰۶	۲۲۴۶	۲۲۴۷	۹۳۲۶۲	۳۰۸
EC۲	۶۲٪	۵۰٪	۵۰٪	۵۴٪	۶۱٪	۱,۰۵	۱۲۲	۱۲۲	۶۵۱	۱۸
EC۵	۷۴٪	۶۴٪	۶۰٪	۸۹٪	۵۹٪	۳۲۸,۳۳	۵۰۱۹	۵۰۸۸	۷۳۶۰۶۵	۱۶۲۱۴
EFC	۵۰٪	۵۰٪	۵۰٪	۵۰٪	۸۱٪	۱۰,۵۵	۹۴۲	۹۴۵	۱۹۵۷۸	۱۳۶
FAA	۸۳٪	۸۵٪	۸۵٪	۶۰٪	۳۰٪	۶۳,۱۰	۱۸۷۶	۱۸۹۳	۷۶۳۵۸۷	۵۲۳۶
FAF	۶۹٪	۵۷٪	۵۷٪	۵۵٪	۳۸٪	۱,۶۰	۲۷۰	۲۷۰	۲۴۴۹	۹۰
FIS	۷۱٪	۷۳٪	۶۴٪	۱۸٪	۷۹٪	۶,۴۴	۹۳۰	۹۳۲	۴۲۷۰۷	۱۲۶۵
FOA	۵۰٪	۵۲٪	۵۱٪	۵۴٪	۵۰٪	۱۱۱۹,۴۹	۱۶۲۶۵	۱۶۳۱۶	۷۶۳۵۸۷	۳۶۸۲۳
FOB	۵۰٪	۵۰٪	۵۰٪	۵۰٪	۵۰٪	۹۵۸,۵۸	۱۲۱۸۷	۱۲۲۲۲	۷۶۳۵۸۷	۲۸۶۳۷
GUN	۵۴٪	۵۴٪	۵۴٪	۵۰٪	۵۱٪	۱,۶۰	۱۳۴	۱۳۴	۲۰۰۹	۱۰
HAM	۵۱٪	۵۲٪	۵۲٪	۵۰٪	۵۳٪	۷,۱۳	۵۵۷	۵۵۸	۱۴۰۳۹	۱۵۲
HAN	۵۶٪	۵۰٪	۵۵٪	۵۵٪	۶۹٪	۵۵۲,۰۱	۲۰۴۰۴	۲۰۴۰۸	۷۶۳۵۸۷	۳۴۰۵۲
HAP	۶۷٪	۶۰٪	۶۰٪	۳۹٪	۶۹٪	۲۵,۶۴	۲۳۰۲	۲۳۰۳	۷۶۳۵۸۷	۳۸۸۱
HER	۵۰٪	۵۰٪	۵۰٪	۵۱٪	۵۰٪	۱,۹۷	۲۷۳	۲۷۳	۷۴۷۵	۳۸
INL	۶۸٪	۷۱٪	۷۱٪	۵۴٪	۷۴٪	۸۱,۸۸	۵۵۹۸	۵۶۰۰	۷۶۳۵۸۷	۱۱۲۳۴
INS	۸۷٪	۸۱٪	۸۱٪	۵۵٪	۶۹٪	۱۱۸,۸۷	۳۶۶۱	۳۶۸۵	۴۷۴۹۰۵	۱۰۴۷۳
ITA	۵۱٪	۵۰٪	۵۰٪	۵۱٪	۶۴٪	۷,۴۰	۲۴۲	۲۴۶	۱۲۳۳	۲۰

LAR	۵۲٪	۵۵٪	۵۵٪	۴۱٪	۳۸,۶۹	۳۹۰۶	۳۹۰۸	۳۶۹۳۷۹	۷۱۱
LI۲	۶۰٪	۵۰٪	۵۴٪	۵۰٪	۲,۱۸	۵۴۳	۵۴۴	۹۴۷۸	۱۲۸
LI۷	۷۶٪	۷۵٪	۷۵٪	۵۹٪	۲,۷۸	۳۲۳	۳۲۴	۳۱۲۵	۱۸۰
MAL	۹۴٪	۸۰٪	۸۰٪	۹۲٪	۵۰۹,۰۶	۱۴۵۱۰	۱۴۵۳۰	۷۶۳۵۸۷	۱۸۳۸۸
MEA	۷۲٪	۷۱٪	۴۰٪	۷۶٪	۲,۶۲	۲۷۹	۲۷۹	۴۷۱۸	۸۶
MED	۶۳٪	۶۵٪	۶۵٪	۴۷٪	۲۱,۵۵	۶۴۶	۶۵۱	۱۹۹۵۴	۱۰۶۲
MPA	۷۳٪	۵۶٪	۷۳٪	۷۳٪	۴,۹۲	۲۵۰	۲۵۱	۴۰۸۶	۷۰
MPC	۵۰٪	۵۱٪	۵۱٪	۵۰٪	۸,۷۳	۳۹۵	۳۹۶	۹۳۰۶	۶۵
MPT	۸۴٪	۷۴٪	۸۲٪	۸۱٪	۴,۸۷	۲۸۶	۲۸۹	۳۳۱۹	۲۰۴
MOT	۶۲٪	۵۰٪	۵۰٪	۵۸٪	۱۷,۴۸	۷۱۷	۷۲۱	۲۰۴۵۶	۴۹۶
NO۱	۹۵٪	۹۴٪	۹۵٪	۹۵٪	۹۹۲,۱۸	۱۷۱۱۵	۱۷۲۲۶	۷۶۳۵۸۷	۲۱۸۶۵۰
NO۲	۹۶٪	۹۵٪	۹۵٪	۹۷٪	۹۸۰,۵۲	۱۳۵۹۴	۱۳۶۷۶	۷۶۳۵۸۷	۲۰۸۴۱۶
OLI	۸۶٪	۷۷٪	۷۷٪	۸۵٪	۲,۳۶	۱۳۹	۱۳۹	۲۴۴۴	۶۶
OSU	۷۳٪	۷۳٪	۷۳٪	۲۹٪	۱۲,۰۴	۱۰۹۷	۱۱۰۰	۶۰۷۶۵	۶۷۰
PHA	۵۴٪	۵۱٪	۵۱٪	۵۱٪	۷۷,۶۲	۱۴۹۹	۱۵۱۵	۷۷۵۵۸	۲۰۰
PHO	۹۳٪	۹۳٪	۹۳٪	۵۱٪	۴۳۲,۷۵	۱۵۴۷۷	۱۵۵۲۵	۷۶۳۵۸۷	۳۶۴۷۰۳
PLA	۹۴٪	۸۳٪	۸۰٪	۹۲٪	۲,۹۹	۱۴۹	۱۴۹	۲۰۴۷	۵۳
PPA	۷۹٪	۷۶٪	۷۶٪	۷۶٪	۵,۷۷	۲۳۵	۲۳۶	۴۹۰۹	۸۲
PPC	۵۲٪	۵۶٪	۵۶٪	۵۳٪	۸,۱۹	۴۱۸	۴۲۱	۸۱۹۵	۳۲
PPT	۷۹٪	۷۸٪	۷۸٪	۸۱٪	۵,۵۲	۳۱۹	۳۲۱	۳۸۵۱	۲۲۱
REF	۵۵٪	۵۶٪	۵۴٪	۳۹٪	۳۵,۷۸	۳۹۶۹	۳۹۷۲	۴۲۲۸۴۳	۹۰۰
SCR	۵۵٪	۵۳٪	۵۳٪	۴۵٪	۳۷,۳۵	۴۰۲۲	۴۰۲۴	۳۵۵۱۰۹	۱۳۵۸
SHS	۵۰٪	۹۹٪	۹۹٪	۵۰٪	۳,۱۴	۱۰۰۵	۱۰۰۶	۱۳۸۷۹	۱۱۱
SHA	۹۵٪	۹۷٪	۹۷٪	۶۳٪	۶۸,۸۰	۳۷۳۸	۳۷۷۷	۵۴۶۵۸۵	۵۱۱۰۷
SMA	۴۳٪	۵۹٪	۵۹٪	۵۴٪	۳۸,۹۰	۴۰۴۸	۴۰۵۰	۳۷۹۸۶۹	۱۴۸۷
SO۱	۵۰٪	۵۱٪	۵۲٪	۷۵٪	۴,۶۶	۳۴۹	۳۵۰	۲۷۷۰	۷۴
SO۲	۶۴٪	۶۰٪	۵۳٪	۶۶٪	۹,۹۰	۶۰۷	۶۱۰	۹۰۴۰	۱۴۹
STR	۵۰٪	۵۰٪	۵۲٪	۵۰٪	۲۴,۰۱	۱۳۳۱	۱۳۳۴	۹۲۰۵۲	۳۶۶
SWE	۸۷٪	۸۸٪	۸۸٪	۶۳٪	۲۳,۴۲	۹۸۵	۹۹۶	۳۶۰۹۲	۱۶۲۰
SYM	۸۹٪	۸۱٪	۸۱٪	۶۰٪	۴۰,۴۵	۲۷۷۴	۲۷۷۹	۲۴۸۴۵۴	۲۸۶۷
SYN	۸۲٪	۷۸٪	۷۸٪	۳۸٪	۵,۶۲	۳۶۱	۳۶۵	۲۱۱۲	۲۳۵
TO۱	۵۰٪	۵۱٪	۵۱٪	۵۳٪	۳,۰۶	۵۲۷	۵۲۸	۸۳۱۵	۱۲۱
TO۲	۵۱٪	۵۰٪	۵۰٪	۵۳٪	۲,۰۲	۳۳۸	۳۳۹	۵۲۵۸	۷۹
TRA	۷۵٪	۸۴٪	۸۴٪	۷۲٪	۳,۲۴	۳۲۵	۳۲۵	۴۹۸۶	۱۱۴
TWP	۶۳٪	۶۴٪	۶۴٪	۴۶٪	۳۶۱,۴۶	۶۴۶۵	۶۵۳۲	۵۸۱۰۵۰	۴۶۰۲
TWE	۵۰٪	۶۴٪	۶۴٪	۵۴٪	۱۴,۰۳	۷۰۳	۷۰۷	۱۱۹۰۳	۲۴۰
UWZ	۸۱٪	۷۸٪	۷۵٪	۵۱٪	۵۰۱,۱۵	۷۴۷۰	۷۵۲۳	۷۶۳۵۸۷	۴۷۴۲۳
UWY	۸۲٪	۷۸٪	۷۶٪	۵۴٪	۵۰۹,۲۳	۸۹۵۱	۹۰۱۴	۷۶۳۵۸۷	۴۸۷۵۱
UWZ	۸۲٪	۸۰٪	۸۰٪	۵۴٪	۵۸۹,۵۳	۷۲۶۴	۷۳۱۵	۷۶۳۵۸۷	۴۷۵۵۴
UWA	۸۶٪	۷۶٪	۷۶٪	۴۵٪	۱۷۵۳,۹۲	۱۸۷۵۴	۱۸۷۹۹	۷۶۳۵۸۷	۱۶۷۰۷۹

WAF	۵۳٪	۵۰٪	۶۶٪	۵۳٪	۵۹٪	۱۵۷۱,۷۶	۴۰۸۷	۴۱۵۸	۷۶۳۵۸۷	۱۶۸۲
WIN	۵۰٪	۵۷٪	۵۰٪	۵۰٪	۵۹٪	۵,۳۷	۱۰۵	۱۰۵	۱۲۷۳	۳۱
WOS	۸۹٪	۸۷٪	۸۷٪	۵۰٪	۵۰٪	۵۴,۰۸	۱۴۶۳	۱۴۷۶	۹۸۲۳۳	۸۲۱۲
WOR	۶۳٪	۶۰٪	۵۸٪	۶۲٪	۵۳٪	۱۲,۹۰	۱۶۰۵	۱۶۰۶	۸۲۵۲۷	۱۰۸۵
WOT	۵۰٪	۵۱٪	۵۱٪	۵۰٪	۵۰٪	۸,۷۰	۱۶۲۰	۱۶۲۱	۷۷۴۹۷	۶۷۶
YOG	۵۰٪	۵۱٪	۵۰٪	۵۰٪	۵۰٪	۴۸۱,۲۲	۷۹۵۹	۷۹۸۳	۷۶۳۵۸۷	۴۸۴۸
متوسط	۶۷,۲۵٪	۶۶,۱۰٪	۶۵,۶۸٪	۶۰,۵۵٪	۶۰,۲۶٪	۱۶۲	۳۲۷۰	۳۲۸۴	۲۳۶۴۶۷	۱۷۳۶۵

جدول ۷: نتایج آزمون آماری الگوریتم ارائه‌شده با چهار الگوریتم منتخب ادبیات در دو شاخص رند و زمان اجرا

الگوریتم	معیار مقایسه	DD_{DTW}	KSC	$TS\mathrm{۳}C_{\mathrm{MV}}$	$TS\mathrm{۳}C_{\mathrm{CH}}$
p-value	شاخص رند	۰,۰۰۰	۰,۰۰۲	۰,۰۱۸	۰,۰۲۱
p-value	زمان	۰,۰۰۰	۰,۰۰۰	۰,۰۰۰	۰,۰۰۰

۴- نتیجه‌گیری

در این مقاله یک رویکرد جدید به‌منظور خوشه‌بندی سری‌های زمانی بر مبنای قطعه‌بندی و خوشه‌بندی ترکیبی در ۵ گام ارائه شد. در این الگوریتم‌ها قطعاتی که می‌توانند نتایج بهتری را در خوشه-بندی 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 شده‌است. الگوریتم دوم ارائه‌شده تا گام سوم کاملاً شبیه الگوریتم اول می‌باشد، اما در گام سوم به‌صورت هم‌زمان از سه معیار داخلی به‌منظور انتخاب $\mathrm{۳}m$ برچسب استفاده می‌گردد و در گام پنجم این برچسب‌ها 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 به ترتیب برابر ۶۳,۲۲٪، ۶۳,۵۲٪، ۶۷٪ و ۶۷,۲۵٪ برای ۸۲ مجموعه‌داده بوده است. همچنین آزمون آماری ویلکانسون به همراه دیگر معیارها نشان داد که الگوریتم دوم می‌تواند عملکرد بهتری در خوشه‌بندی سری‌های زمانی داشته باشد. همچنین عملکرد الگوریتم دوم به لحاظ شاخص رند و متوسط زمان اجرای الگوریتم با ۴ الگوریتم DD_{DTW}، KSC، $TS\mathrm{۳}C_{\mathrm{MV}}$ و $TS\mathrm{۳}C_{\mathrm{CH}}$ مقایسه شده است.

در مطالعات اخیر سعی بر آن شده است با استفاده از تعریف معیارهای جدید فاصله و یا استفاده از الگوریتم‌های خاص کارایی خوشه‌بندی افزایش یابد. اما با این وجود نتایج مطالعات اخیر نشان داده است که استفاده از این روش‌ها علی رقم بهبود دقت خوشه‌بندی می‌تواند هزینه زمانی بسیار بالایی داشته باشد. به‌عنوان مثال در چهار الگوریتم موجود در ادبیات حتی می‌تواند به عددی نزدیک به ۲۳۶ هزار ثانیه برسد که عملاً استفاده از این الگوریتم‌ها در بخش صنعت و خدمات را ناممکن می‌نماید. در طرح ارائه‌شده سعی بر ان شده است علاوه بر افزایش دقت خوشه‌بندی بتوان این کار را با حداقل هزینه زمانی انجام پذیرد. با توجه به نتایج ارائه‌شده در جدول ۶ به خوبی می‌توان مشاهده نمود که طرح ارائه‌شده علاوه برافزایش دقت خوشه‌بندی می‌تواند این امر را در کمترین زمان ممکن انجام دهد. همچنین نشان داده شده است که طرح ارائه‌شده در ۳۲٪ از مجموعه داده‌های مورد استفاده توانسته است بهترین دقت را در بین الگوریتم‌های موجود داشته باشد. نتایج آزمون آماری به‌خوبی نشان داده است که الگوریتم دوم ارائه‌شده به لحاظ شاخص رند و همچنین زمان اجرا از همه الگوریتم‌های مقایسه شده کارایی بیشتری دارد. با توجه به‌دقت بالا و هزینه زمانی بسیار ناچیز طرح ارائه‌شده جهت خوشه‌بندی، الگوریتم مورد نظر می‌تواند به‌راحتی در مواردی مانند تشخیص الگوهای مصرف انواع حامل‌های انرژی، تشخیص الگوهای رفتاری بیماری‌های واگیردار و غیره مورد استفاده قرار گیرد. همچنین با توجه به اینکه یکی از کاربردهای اصلی خوشه‌بندی استفاده در پیش‌بینی سری‌های زمانی می‌باشد. با افزایش دقت خوشه‌بندی در زمان کم می توان عملا دقت پیش‌بینی این نوع داده‌ها را نیز افزایش داد.

۵- مطالعات آینده

طرح ارائه‌شده دارای پارامترها و مشخصات مختلفی می‌باشد که در این مطالعه صرفاً حالت‌ها اندکی از آن مورد بررسی و تحلیل قرارگرفته است. در مطالعات آینده می‌توان انواع دیگر الگوریتم‌های خوشه‌بندی، 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 می‌توان به‌منظور

کشف الگوهای مختلف در زمین‌های انرژی و درمان نیزاستفاده عملی نمود.

مراجع

[۱] M. Maleki, H. Bidram, and D. Wraith, "Robust clustering of COVID-۱۹ cases across US counties using mixtures of asymmetric time series models with time varying and freely indexed covariates," *Journal of Applied Statistics*, pp. ۱-۱۵, ۲۰۲۲.

[۲] E. Bair, "Semi-supervised clustering methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. ۵, no. ۵, pp. ۳۴۹-۳۶۱, ۲۰۱۳.

[۳] H. A. Dau, N. Begum, and E. Keogh, "Semi-supervision dramatically improves time series clustering under dynamic time warping," in *Proceedings of the ۲۵th ACM International on Conference on Information and Knowledge Management*, ۲۰۱۶, pp. ۹۹۹-۱۰۰۸.

[٤] L. Alhusain and A. M. Hafez, "Cluster ensemble based on Random Forests for genetic data," *BioData Mining*, vol. ۱۰, no. ۱, p. ۳۷, ۲۰۱۷.

[٥] R. Ma and R. Angryk, "Distance and density clustering for time series data," in ۲۰۱۷ *IEEE International Conference on Data Mining Workshops (ICDMW)*, ۲۰۱۷: IEEE, pp. ۲۵-۳۲.

[٦] S. Mehrmolaei and M. R. Keyvanpour, "A comparative study on weighting-based clustering techniques: Time series data," in ۲۰۱۸ ۸*th Conference of AI & Robotics and* ۱۰*th RoboCup Iranopen International Symposium (IRANOPEN)*, ۲۰۱۸: IEEE, pp. ٦٥-۷۲.

[۷] N. Tavakoli, S. Siami-Namini, M. Adl Khanghah, F. Mirza Soltani, and A. Siami Namin, "An autoencoder-based deep learning approach for clustering time series data," *SN Applied Sciences*, vol. ۲, pp. ۱-۲۵, ۲۰۲۰.

[۸] B. Lafabregue, J. Weber, P. Gançarski, and G. Forestier, "End-to-end deep representation learning for time series clustering: a comparative study," *Data Mining and Knowledge Discovery*, vol. ۳٦, no. ۱, pp. ۲۹-۸۱, ۲۰۲۲.

[۹] S. Zolhavarieh, S. Aghabozorgi, and Y. W. Teh, "A review of subsequence time series clustering," *The Scientific World Journal*, vol. ۲۰۱٤, ۲۰۱٤.

[۱۰] C. A. Ralanamahatana, J. Lin, D. Gunopulos, E. Keogh, M. Vlachos, and G. Das, "Mining time series data," in *Data mining and knowledge discovery handbook*: Springer, ۲۰۰۵, pp. ۱۰٦۹-۱۱۰۳.

[۱۱] H. Kamalzadeh, A. Ahmadi, and S. Mansour, "Clustering time-series by a novel slope-based similarity measure considering particle swarm optimization," *Applied Soft Computing*, p. ۱۰٦۷۰۱, ۲۰۲۰.

[۱۲] G. Soleimani and M. Abessi, "DLCSS: A new similarity measure for time series data mining," *Engineering Applications of Artificial Intelligence*, vol. ۹۲, p. ۱۰۳٦٦٤, ۲۰۲۰.

[۱۳] M. Łuczak, "Hierarchical clustering of time series data with parametric derivative dynamic time warping," *Expert Systems with Applications*, vol. ٦۲, pp. ۱۱٦-۱۳۰, ۲۰۱٦.

[۱٤] L. Rokach and O. Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*: Springer, ۲۰۰۵, pp. ۳۲۱-۳۵۲.

[۱۵] M. A. Rahim Khan and M. Zakarya, "Longest common subsequence based algorithm for measuring similarity between time series: a new approach," *World Applied Sciences Journal*, vol. ۲٤, no. ۹, pp. ۱۱۹۲-۱۱۹۸, ۲۰۱۳.

[۱٦] X. Wang, F. Yu, W. Pedrycz, and J. Wang, "Hierarchical clustering of unequal-length time series with area-based shape distance," *Soft Computing*, vol. ۲۳, no. ۱۵, pp. ٦۳۳۱-٦۳٤۳, ۲۰۱۹.

[۱۷] K. K. W. Chu and M. H. Wong, "Fast time-series searching with scaling and shifting," in *Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART*

Engineering (TCSET), ٢٠٢٠: IEEE, pp. ٦٥٣-٦٥٨.

[٢٦] R. J. Hyndman, E. Wang, and N. Laptev, "Large-scale unusual time series detection," in ٢٠١٥ *IEEE international conference on data mining workshop (ICDMW)*, ٢٠١٥: IEEE, pp. ١٦١٦-١٦١٩.

[٢٧] Y. Zou, R. V. Donner, N. Marwan, J. F. Donges, and J. Kurths, "Complex network approaches to nonlinear time series analysis," *Physics Reports*, vol. ٧٨٧, pp. ١-٩٧, ٢٠١٩.

[٢٨] V. A. F. da Silva, "Time Series Analysis based on Complex Networks," ٢٠١٨.

[٢٩] L. N. Ferreira and L. Zhao, "Time series clustering via community detection in networks," *Information Sciences*, vol. ٣٢٦, pp. ٢٢٧-٢٤٢, ٢٠١٦.

[٣٠] F. Bonacina, E. S. Miele, and A. Corsini, "Time Series Clustering: A Complex Network-Based Approach for Feature Selection in Multi-Sensor Data," *Modelling*, vol. ١, no. ١, pp. ١-٢١, ٢٠٢٠.

[٣١] A. Koski, M. Juhola, and M. Meriste, "Syntactic recognition of ECG signals by attributed finite automata," *Pattern Recognition*, vol. ٢٨, no. ١٢, pp. ١٩٢٧-١٩٤٠, ١٩٩٥.

[٣٢] E. J. Keogh and M. J. Pazzani, "An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback," in *Kdd*, ١٩٩٨, vol. ٩٨, pp. ٢٣٩-٢٤٣.

[٣٣] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "Segmenting time series: A survey and novel approach," in *Data mining in time series databases*: World Scientific, ٢٠٠٤, pp. ١-٢١.

[٣٤] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," *ACM Sigmod Record*, vol. ٢٣, no. ٢, pp. ٤١٩-٤٢٩, ١٩٩٤.

[٣٥] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information symposium on Principles of database systems*, ١٩٩٩, pp. ٢٣٧-٢٤٨.

[١٨] S. Gharghabi, S. Imani, A. Bagnall, A. Darvishzadeh, and E. Keogh, "Matrix profile xii: Mpdist: a novel time series distance measure to allow data mining in more challenging scenarios," in ٢٠١٨ *IEEE International Conference on Data Mining (ICDM)*, ٢٠١٨: IEEE, pp. ٩٦٥-٩٧٠.

[١٩] G. Jiang, W. Wang, and W. Zhang, "A novel distance measure for time series: Maximum shifting correlation distance," *Pattern Recognition Letters*, vol. ١١٧, pp. ٥٨-٦٥, ٢٠١٩.

[٢٠] D. Hong, Q. Gu, and K. Whitehouse, "High-dimensional time series clustering via cross-predictability," in *Artificial Intelligence and Statistics*, ٢٠١٧: PMLR, pp. ٦٤٢-٦٥١.

[٢١] T. Górecki, "Classification of time series using combination of DTW and LCSS dissimilarity measures," *Communications in Statistics-Simulation and Computation*, vol. ٤٧, no. ١, pp. ٢٦٣-٢٧٦, ٢٠١٨.

[٢٢] D. Guijo-Rubio, A. M. Durán-Rosal, P. A. Gutiérrez, A. Troncoso, and C. Hervás-Martínez, "Time-Series Clustering Based on the Characterization of Segment Typologies," *IEEE Transactions on Cybernetics*, ٢٠٢٠.

[٢٣] S. Aghabozorgi, T. Ying Wah, T. Herawan, H. A. Jalab, M. A. Shaygan, and A. Jalali, "A hybrid algorithm for clustering of time series data based on affinity search technique," *The Scientific World Journal*, vol. ٢٠١٤, ٢٠١٤.

[٢٤] X. Zhang, J. Liu, Y. Du, and T. Lv, "A novel clustering method on time series data," *Expert Systems with Applications*, vol. ٣٨, no. ٩, pp. ١١٨٩١-١١٩٠٠, ٢٠١١.

[٢٥] N. Manakova and V. Tkachenko, "Two-stage time-series clustering approach under reducing time cost requirement," in ٢٠٢٠ *IEEE ١٥th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer*

[٣٦] M. Djukanovic, G. R. Raidl, and C. Blum, "Finding Longest Common Subsequences: New anytime A∗ search results," *Applied Soft Computing,* vol. ٩٥, p. ١٠٦٤٩٩, ٢٠٢٠.

[٣٧] M. Paterson and V. Dančík, "Longest common subsequences," in *International symposium on mathematical foundations of computer science*, ١٩٩٤: Springer, pp. ١٢٧-١٤٢.

[٣٨] R. Lin, A. King-lp, and H. S. S. K. Shim, "Fast similarity search in the presence of noise, scaling, and translation in time-series databases," in *Proceeding of the* ٢١*th International Conference on Very Large Data Bases*, ١٩٩٥: Citeseer, pp. ٤٩٠-٥٠١.

[٣٩] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories," in *Proceedings* ١٨*th international conference on data engineering*, ٢٠٠٢: IEEE, pp. ٦٧٣-٦٨٤.

[٤٠] T. Boongoen and N. Iam-On, "Cluster ensembles: A survey of approaches with recent extensions and applications," *Computer Science Review,* vol. ٢٨, pp. ١-٢٥, ٢٠١٨.

[٤١] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE transactions on cybernetics,* vol. ٤٨, no. ٥, pp. ١٤٦٠-١٤٧٣, ٢٠١٧.

[٤٢] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in ٢٠١٠ *IEEE international conference on data mining*, ٢٠١٠: IEEE, pp. ٩١١-٩١٦.

[٤٣] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics,* vol. ٤, no. ١, pp. ٩٥-١٠٤, ١٩٧٤.

[٤٤] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods,* vol. ٣, no. ١, pp. ١-٢٧, ١٩٧٤.

[٤٥] H. A. Dau *et al.*, "The UCR time series archive," *IEEE/CAA Journal of Automatica Sinica,* vol. ٦, no. ٦, pp. ١٢٩٣-١٣٠٥, ٢٠١٩.

[٤٦] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research,* vol. ٧, pp. ١-٣٠, ٢٠٠٦.

[٤٧] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ٢٠١١, pp. ١٧٧-١٨٦.