



دانشگاه کاشان  
University of Kashan

مجله محاسبات نرم

## SOFT COMPUTING JOURNAL

تارنمای مجله: [scj.kashanu.ac.ir](http://scj.kashanu.ac.ir)



### خوشه‌بندی قیمت سهام با استفاده از الگوریتم کوچک‌ترین درخت پوشا

فائزه کلهری<sup>1</sup>، کارشناسی ارشد، سیدمحمد حسینی<sup>1\*</sup>، استادیار

<sup>1</sup> گروه ریاضی، دانشکده علوم پایه، دانشگاه آیت الله بروجردی (ره)، بروجرد، ایران.

#### اطلاعات مقاله

#### چکیده

#### تاریخچه مقاله:

دریافت 11 شهریور ماه 1402

پذیرش 14 فروردین ماه 1403

#### کلمات کلیدی:

خوشه‌بندی

درخت پوشا

قیمت سهام

نزدیک‌ترین همسایه

الگوریتم کراسکال

با توجه به فعالیت‌های روزافزون اشخاص حقیقی و حقوقی در بازار سرمایه و تبدیل شدن این بازار به یکی از مهم‌ترین اهرم‌های اقتصادی هر کشور، هرچه دانش در رابطه با انتخاب سهم و سهام‌داری بیشتر باشد، بی‌شک سودآوری بیشتری به دنبال خواهد داشت. در پژوهش حاضر نظر به اهمیت قیمت سهام، خوشه‌بندی آن با استفاده از الگوریتم کوچک‌ترین درخت پوشا پیشنهاد شده است. داده‌های مورد استفاده پژوهش، قیمت بسته شدن روزانه سهام شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران در بازه زمانی 1398/7/1 تا 1399/8/20 است. در این پژوهش در مرحله اول، تعدادی زیرخوشه شامل شرکت‌های مشابه از نظر رفتار قیمتی، تشکیل می‌شود پس از آن بر اساس برخی معیارهای تشابه، زیرخوشه‌ها ادغام و نتیجه مطلوب یعنی داشتن خوشه‌هایی شامل اعضای با بیشترین مشابهت حاصل می‌شود. از جمله مزایای روش پیشنهادی این است که در این روش معیارهای تشابه به صورت موضعی محاسبه می‌شود و بنابراین محاسبات آن نسبت به سایر روش‌ها کمتر خواهد بود. نتایج این پژوهش نشان می‌دهد که فرآیند خوشه‌بندی مجموعه‌های حجیم به راحتی و با دقتی مطلوب قابل انجام و استفاده است.

© 1403 نویسندگان. مقاله با دسترسی آزاد تحت مجوز CC-BY

#### 1. مقدمه

ذاتی در یک مجموعه داده مشخص مورد استفاده قرار می‌گیرد. خوشه‌بندی را می‌توان به صورت «فرآیند سازمان‌دهی داده‌ها به گروه‌هایی که اعضای آنها بر مبنای معیاری مشخص، شبیه به هم هستند» تعریف کرد. بنابراین خوشه، مجموعه‌ای از داده‌ها است که با یکدیگر بیشترین شباهت و با اعضای دیگر خوشه‌ها بیشترین عدم شباهت را دارند.

تاکنون الگوریتم‌های خوشه‌بندی متنوعی ارائه شده است، که به سه رده افزایی، سلسله مراتبی و مبتنی بر تراکم، طبقه‌بندی می‌شوند. اما بسیاری از این الگوریتم‌ها نسبت به داده‌های مورد استفاده و بعضی پارامترهای تنظیمی الگوریتم حساسیت نشان می‌دهند و نتایج نهایی این الگوریتم‌ها علاوه بر وابستگی به

خوشه‌بندی یکی از رویکردهای با اهمیت در زمینه‌ی داده‌کاوی، یادگیری ماشین و هوش مصنوعی است که در علوم مختلف و زمینه‌های کاربردی متنوعی چون مهندسی، پزشکی، روان‌شناسی، بازاریابی، اقتصاد، رایانه و ... به‌کار گرفته شده است. از خوشه‌بندی در هوش مصنوعی و شناسایی الگو به عنوان یادگیری بدون ناظر نام برده می‌شود و برای کشف شباهت‌های

\* نوع مقاله: پژوهشی

\* نویسنده مسئول

پست(های) الکترونیک: faezehkhalhori96@gmail.com (کلهری)

sm.hoseini@abru.ac.ir (حسینی)

کشور موثر است، نداشتن دانش کافی نسبت به آنها می‌تواند مضر باشد. برای سرمایه‌گذار، بازاری‌های مختلفی همچون بازار سهام، املاک، خودرو، طلا، ارز و مشابه این‌ها وجود دارد که همگی برای افزایش بازدهی و سود در رقابت هستند.

بازارهای مالی مختلف، ریسک‌های متفاوتی دارند و بازار سهام هم ریسک‌های خاص خود را دارد. سرمایه‌گذاری‌های بدون تحلیل در بازار سهام منجر به از دست رفتن فرصت و سرمایه خواهد شد. یکی از دشواری‌های بازار سهام، نوسانات قیمتی و پیچیدگی حرکت آن است. دانستن حرکت قیمت سهام، یکی از دغدغه‌های اصلی سهام‌داران و فعالان بازار سرمایه است تا بیشترین سود را کسب کنند، از این رو همواره به دنبال راه‌کارهای منطقی و دقیق جهت شناسایی روند قیمتی سهام هستند [1]، [2]. با توجه به اینکه بازار سهام، یک سیستم تصادفی و غیرخطی است، بنابراین استفاده از روش‌های تجزیه و تحلیل سنتی برای درک رفتار آن نامفید است. در میان رویکردهای جدید، خوشه‌بندی سهام شرکت‌های موجود در بازار سهام، خوشه‌هایی را متناسب با تغییر رفتار سهام به گونه‌ای ایجاد می‌کند که بیشترین شباهت بین اعضای هر خوشه وجود داشته باشد و نمادهای مشابه از نظر رفتاری، به آسانی قابل تشخیص باشند تا بتوان بهترین تصمیم را در این بازار مالی داشت.

در این پژوهش، خوشه‌بندی شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران مورد توجه قرار گرفته است. از آنجا که مجموعه‌های متنوعی از داده‌های این شرکت‌ها وجود دارد و همچنین با توجه به حجم بالای این داده‌ها، از روش خوشه‌بندی سه‌مرحله‌ای مبتنی بر کوچک‌ترین درخت پوشا استفاده شده است. بعد از خوشه‌بندی، از آنجا که نمادهای موجود در یک خوشه مشخص، مشابه یکدیگر هستند، بنابراین انتظار می‌رود رفتار یکسانی در آینده داشته باشند و این اطلاعات برای سرمایه‌گذار مهم است. خوشه‌بندی شرکت‌ها باعث می‌شود که سهام‌داران با به دست آوردن اطلاعات مفید در خصوص سهام یک شرکت، تصمیم مناسبی به شرکت‌های هم‌خوشه با آن را لحاظ کنند و تصمیم مناسبی برای کسب بیشترین بازده اتخاذ

ساختار داده‌ها، به پارامترهای تنظیمی نیز وابسته هستند و وقتی این پارامترها نادرست لحاظ شوند یا زمانی که بر روی مجموعه داده‌های ناهمگن استفاده شوند، ناکارآمد هستند. در این‌گونه روش‌ها، بایستی شباهت یا عدم شباهت هر دو شی داده‌ای محاسبه شود. در نتیجه پیچیدگی محاسباتی پیدا کردن شبیه‌ترین شی داده‌ای برای هر یک از اشیا داده‌ای عضو مجموعه، از مرتبه  $O(N)$  است و پیچیدگی محاسباتی کل مجموعه داده‌ای حداقل از مرتبه  $O(N^2)$  می‌شود. به بیان دیگر، در روش‌های مذکور از یک گراف کامل وزن‌دار با وزن یالی برای خوشه‌بندی استفاده می‌شود که در آن وزن هر یال، میزان شباهت یا عدم شباهت رئوس آن یال است. اما برای پیدا کردن شبیه‌ترین شی داده‌ای به یک شی داده‌ای مشخص، نیاز نیست تمام اعضای مجموعه داده‌ای بررسی شوند، بلکه بررسی تنها بخش کوچکی از نقاط اطراف آن کفایت می‌کند. بنابراین اگر بتوان ساختار شباهت اشیا داده‌ای را با استفاده از یک گراف موضعی به جای یک گراف کامل نشان داد، پیچیدگی محاسباتی کاهش می‌یابد.

یکی از دیدگاه‌های خوشه‌بندی مبتنی بر گراف موضعی، الگوریتم خوشه‌بندی بر پایه کوچک‌ترین درخت پوشا<sup>1</sup> است. این روش خوشه‌بندی شامل سه مرحله تقسیم، تسخیر<sup>2</sup> و ادغام خوشه‌ها است. در مرحله تقسیم، مجموعه داده‌ها به تعداد زیادی زیرخوشه تقسیم می‌شود و از هر زیرخوشه یک نماینده انتخاب و یک گراف وزن‌دار بر روی نماینده‌های زیرخوشه‌ها ایجاد می‌شود. سپس در مرحله تسخیر، گراف کوچک‌ترین درخت پوشای متناظر با گراف وزن‌دار مرحله قبل محاسبه می‌شود و در مرحله ادغام، زیرخوشه‌های تولید شده در مرحله تقسیم که در مرحله تسخیر مجاور شده‌اند، ادغام می‌شوند و این ادغام براساس معیارهای معینی انجام می‌شود.

مجموعه‌های داده‌ای متنوع و مهمی در بازارهای مالی تولید و نگهداری می‌شود که تجزیه و تحلیل آنها همواره مورد توجه محققان، سیاست‌گذاران و سرمایه‌گذاران بوده است. همان‌طور که وجود بازارهای مالی فعال و پویا در رشد اقتصادی یک

<sup>1</sup> Minimum Spanning Tree (MST)

<sup>2</sup> Conquer

کنند. مهم‌ترین دستاوردهای مقاله حاضر عبارتند از:

- معرفی و مقایسه معیار ادغام جدید و مناسب‌تری نسبت به معیار ادغام پیشنهاد شده توسط میشر و موهانتی [3]؛
  - استفاده از روش پیشنهادی برای خوشه‌بندی شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران.
- نتایج این تحقیق نشان می‌دهند که الگوریتم پیشنهادی، خوشه‌بندی مناسبی از مجموعه داده‌های بورس را ارائه می‌دهد. بنابراین سهام‌داران می‌توانند با بررسی دقیق و بکار بستن روش‌های معتبر موجود برای پیش‌بینی یک یا چند نماینده از آن خوشه، رفتار سایر اعضای آن را هم تخمین بزنند و به این ترتیب اطلاعات بسیاری در اختیار سهام‌دار خواهد بود.

## 2. پیشینه پژوهش

در این بخش، فعالیت و تحقیقات مرتبط با موضوع خوشه‌بندی سهام از پژوهش‌گران داخلی و خارجی ارائه می‌شود. سری زمانی، مجموعه‌ای از مشاهده‌ها در بازه‌های زمانی مساوی و مشخص است [4]. برای ارزیابی تشابه در مجموعه داده‌های سری‌های زمانی از معیارهای فاصله مخصوص سری زمانی، مانند پیچش زمانی پویا<sup>1</sup> و فاصله اقلیدسی بهره گرفته می‌شود. فرید و پورحمیدی [2] از داده‌های 338 شرکت پذیرفته شده در بورس اوراق بهادار تهران در یک بازه زمانی 5 ساله از ابتدای سال 1384 تا انتهای سال 1388 استفاده کردند. آنها با روش C- میانگین فازی، داده‌های صورت‌های مالی شرکت‌های مذکور شامل 11 شاخص متداول از قبیل نسبت قیمت به درآمد هر سهم، نسبت ارزش بازار به ارزش دفتری، شاخص ریسک سیستماتیک، بازده حقوق صاحبان سهم و ... را خوشه‌بندی کردند. نتایج تحقیق آنها نشان داد که بخش اعظم نمادها در خوشه سهام ترکیبی قرار می‌گیرند و گرایش آنها به خوشه سهام رشدی است.

سروش‌یار و اخلاقی [5] با بکارگیری تکنیک‌های داده‌کاوی همچون K-میانگین بر روی داده‌های 107 شرکت پذیرفته شده در بورس اوراق بهادار تهران، تاثیر متغیرها را در پیش‌بینی بازده

سهام مورد بررسی قرار دادند. قیمت سهام شرکت‌ها پیوسته در حال تغییر است و پیش‌بینی ارزش آینده بازار توسط فروشندگان و خریداران بسیار دشوار است.

شیرازیان و همکاران [6] در سال 1399 نوسانات شاخص کل بورس اوراق بهادار تهران در بازه‌ی زمانی 10 ساله خوشه‌بندی کردند و نتیجه گرفتند که خوشه‌بندی نوسانات در بورس تهران وجود دارد و نامتقارن است. بنابراین در مدیریت ریسک بایستی به آن توجه نمود.

در سال 1394، اقبال‌نیا و همکاران [7] از یک روش خوشه‌بندی سه‌مرحله‌ای برای خوشه‌بندی بر مبنای شباهت در روند حرکتی قیمت سهام شرکت‌های بورس اوراق بهادار تهران در دوره زمانی سال 1392 استفاده کردند. در روش مذکور، ابتدا کاهش ابعاد داده‌ها انجام می‌شود و خوشه‌بندی تقریبی توسط روش K-میانگین تعیین می‌شود. سپس خوشه‌های به دست آمده به زیرخوشه‌هایی با کیفیت‌تر تقسیم می‌شوند و در انتها زیرخوشه‌های مذکور با یکدیگر ادغام می‌شوند تا خوشه‌های نهایی حاصل شوند.

بلاشو و گریفیت [8] در سال 2016 یک سیستم تجزیه و تحلیل که به افراد کمک می‌کند تا با استفاده از رویکردهای داده‌کاوی، شرکت‌های سودآور را شناسایی کنند، ارائه شده است. رویکرد خوشه‌بندی آنها، سهام را بر اساس معیارهای سرمایه‌گذاری خاص طبقه‌بندی می‌کند. نتایج تحقیق ناندا و همکاران [9] نشان می‌دهد که الگوریتم خوشه‌بندی K-میانگین، مرتب‌ترین خوشه‌ها را در مقایسه با روش‌های نگاشت خودسازمانده و C-میانگین فازی برای خوشه‌بندی سهام به دست می‌آورد.

در برخی از روش‌های خوشه‌بندی فرض می‌شود که ترجیحات سرمایه‌گذاران نامشخص است در حالی که در واقع ترجیحات انتخاب سرمایه‌گذاران متفاوت است. برخی از سرمایه‌گذاران بدبین هستند، در حالی که دیگران ممکن است خوش‌بین باشند. باتوجه به این مساله، زین‌العابدین و همکاران [10] روش جدید خوشه‌بندی فازی سازگار با عمل کرد سهام ارائه می‌دهد که بین سرمایه‌گذاران بدبین و خوش‌بین تفاوت قائل می‌شود. این روش، سهام در نظر گرفته شده را بر اساس سطح اعتماد به

<sup>1</sup> Dynamic Time Wrapping (DTW)

کوچک‌ترین درخت پوشای تقریبی را پیشنهاد داده‌اند. این روش در مرحله اول خوشه‌بندی، طولانی‌ترین یالی را که در ساختار کوچک‌ترین درخت پوشا شرکت نمی‌کند، شناسایی می‌کند و نشان داده‌اند که روش مذکور برای به دست آوردن خوشه‌ها محاسبات کمتری نیاز دارد.

تقسیم و ادغام، نمونه دیگری از الگوریتم خوشه‌بندی ترکیبی است که توسط ژونگ و همکاران [17] پیشنهاد شده است. این روش، گراف کوچک‌ترین درخت پوشا را برای نشان دادن مجموعه داده ایجاد می‌کند و بالاترین درجه گره را به عنوان مرکز خوشه انتخاب می‌کند و سپس از الگوریتم K-میانگین برای تقسیم‌بندی مجموعه داده‌ها به تعداد زیادی زیرخوشه استفاده می‌کند.

میشرا و موهانتی یک روش سه‌مرحله‌ای را بر مبنای دیدگاه خوشه‌بندی موضعی و ادغام نتایج آن پیشنهاد دادند و ثابت کرده‌اند که روش پیشنهادی کاهش چشم‌گیری در محاسبات در پی دارد. آنها نشان داده‌اند که پیچیدگی محاسباتی روش سه‌مرحله‌ای مذکور از مرتبه  $O(N^2)$  و بنابراین محاسبات روش پیشنهادی نسبت به روش پیمایش کل شبکه وزن‌دار مربوط به فواصل نقاط داده‌ای کاهش قابل ملاحظه‌ای می‌یابد.

اخیرا، ژاو و همکاران [18] معیار شباهت اختصاصی چند رویکردی پویا<sup>5</sup> را معرفی کردند و از آن برای خوشه‌بندی و پیش‌بینی قیمت 285 شرکت پذیرفته شده در بورس شانگهای استفاده کردند. در این روش، ابتدا سری‌های زمانی مورد نظر قطعه‌بندی می‌شوند و سپس وزن‌دهی مناسب اعمال می‌گردد. سپس فاصله‌ی کانبرا<sup>6</sup> را در پیچش زمانی پویا بجای فاصله اقلیدسی می‌نشانند و از آن برای خوشه‌بندی قطعه‌های سری زمانی مذکور استفاده می‌شود. در انتها نیز پیش‌بینی انجام می‌شود. نتایج این تحقیق نشان داد، روش مذکور در مقایسه با روش‌های دیگر از قبیل معیار اقلیدسی و پیچش زمانی پویا عملکرد مطلوب‌تری دارد.

نفس و ترجیحات سرمایه‌گذاران رتبه‌بندی می‌کند و سپس خوشه‌بندی را بر روی سهام انجام می‌دهد. تحقیق مذکور بر روی 30 شرکت در بورس اوراق بهادار مالزی انجام شده است. گونگور و اوزمن [11] یک روش خوشه‌بندی بدون پارامتر به نام فاصله چگالی گاوسی<sup>1</sup> برای تعیین خوشه‌ها معرفی کردند که با استفاده از هسته گاوسی و بر پایه فاصله و شکل خوشه‌ها عمل می‌کند. این روش پیشنهادی به منظور کمک به سرمایه‌گذاران بورس اوراق بهادار در شناسایی فرصت‌های سودآوری احتمالی و همچنین کمک به درک بهتر در مورد چگونگی استخراج اطلاعات مربوط به داده‌های قیمت سهام است. در مرجع [12] از درخت رگرسیون برای کاهش ابعاد داده‌ها و از نگاشت خودسازمانده<sup>2</sup> برای خوشه‌بندی استفاده شده است. یکی دیگر از الگوریتم‌های خوشه‌بندی ترکیبی برای بررسی مجموعه داده‌های حجیم، الگوریتم خوشه‌بندی به کمک نماینده‌ها<sup>3</sup> است [13]. در این الگوریتم هر خوشه با تعداد مشخصی از اشیا داده‌ای آن نمایش داده می‌شود و شباهت هر دو زیرخوشه به وسیله نزدیک‌ترین فاصله بین نقاط نماینده زیرخوشه‌های مختلف اندازه‌گیری می‌شود. چسبندگی خود-ادغام<sup>4</sup> یک روش خوشه‌بندی قوی و کارآمد است که ابتدا از K-میانگین برای تقسیم‌بندی مجموعه داده به K زیرخوشه استفاده می‌کند و به طور تکراری آنها را در خوشه‌های نهایی ادغام می‌کند [14]. الگوریتم دیگری از خوشه‌بندی ترکیبی که می‌تواند مجموعه داده ناهمگن را خوشه‌بندی کند در مرجع [15] ارائه شده است. این الگوریتم، K-نزدیک‌ترین همسایه را ایجاد می‌کند و از روش تقسیم گراف برای تقسیم‌بندی مجموعه داده به چندین زیرخوشه استفاده می‌کند.

یکی از الگوریتم‌های خوشه‌بندی مبتنی بر گراف، الگوریتم خوشه‌بندی بر پایه کوچک‌ترین درخت پوشا است. وانگ و همکاران [16]، روش‌های خوشه‌بندی مبتنی بر کوچک‌ترین درخت پوشا با استفاده از تقسیم و تسخیر برای ایجاد

<sup>1</sup> Gauss Density Distance (GDD)

<sup>2</sup> Self-Organizing Map (SOM)

<sup>3</sup> Clustering Using Representatives (CURE)

<sup>4</sup> Cohesion Self-Merging (CSM)

<sup>5</sup> Dynamic multi-perspective personalized similarity measurement

<sup>6</sup> Canberra

## 3. روش تحقیق

تعداد زیرخوشه‌ها به  $\sqrt{N}$  تعداد برسد، فرآیند زیر تکرار می‌شود. فرض کنید  $C_0, C_1, \dots, C_{M-1}$  زیرخوشه‌های موجود باشند. اکنون بر اساس میزان پراکندگی، یکی از زیرخوشه‌ها برای تقسیم به دو زیرخوشه‌ی جدید انتخاب می‌شود [3]:

$$m = \operatorname{argmax}_{0 \leq i < M} \frac{1}{|C_i|} \sum_{x \in C_i} \|x - \mu(C_i)\|_2, \quad (1)$$

که در آن  $|C_i|$  عدد اصلی خوشه  $i$  ام و  $\|\cdot\|_2$  فاصله (نرم) اقلیدسی است. با توجه به رابطه (1)، زیرخوشه  $C_m$  خوشه‌ای است که بیشترین پراکندگی را حول مرکز زیرخوشه نسبت به سایر زیرخوشه‌ها دارد. این زیرخوشه به روش زیر به دو زیرخوشه جدید  $C_{m1}$  و  $C_{m2}$  تقسیم می‌شود. فرض کنید مجموعه نقاط زیرخوشه  $C_m$  باشد که در آن:

$$x_i = (x_i^1, x_i^2, \dots, x_i^d) \in \mathbb{R}^d \quad (2)$$

یک بردار ویژگی با  $d$  مولفه (ویژگی) است. در ادامه، مرکز (میانگین) خوشه  $C_m$  با  $(\bar{x}^1, \bar{x}^2, \dots, \bar{x}^d)$  و واریانس مجموعه نقاط داده‌ای  $X$  بر روی ویژگی  $i$  ام با  $\operatorname{Var}(X^i)$  نشان داده می‌شود. قرار دهید:

$$k = \operatorname{argmax}_{1 \leq i \leq d} \operatorname{Var}(X^i) \quad (3)$$

به عبارت دیگر،  $k$  شماره آن ویژگی در بین تمام ویژگی‌های نقاط داده‌ای خوشه  $C_m$  است که بیشترین واریانس را دارد. در صورتی که  $k$  منحصر به فرد نباشد، یکی به دلخواه انتخاب می‌شود. اکنون برای تقسیم‌بندی خوشه  $C_m$  به زیرخوشه‌های  $C_{m1}$  و  $C_{m2}$  به صورت زیر عمل می‌شود:

$$C_{m1} = \{x_i \in C_m \mid x_i^k < \bar{x}^k\}, \quad (4)$$

$$C_{m2} = C_m \setminus C_{m1}. \quad (5)$$

که در رابطه (4)،  $\bar{x}^k$  میانگین ویژگی  $k$ -ام و در رابطه (5) نماد  $\setminus$  نشان‌دهنده تفاضل مجموعه‌ها است و بنابراین  $C_{m2}$  شامل اعضای  $C_m$  است که در  $C_{m1}$  نیستند. به این ترتیب زیرخوشه  $C_m$  به دو زیرخوشه جدید تقسیم می‌شود.

در این مقاله، از روش خوشه‌بندی مبتنی بر کوچک‌ترین درخت پوشا، برای شناسایی خوشه‌های مجموعه‌ی داده‌ای سری زمانی قیمت بسته شدن سهام شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران استفاده می‌شود. این روش مشابه روش معرفی شده در توسط میسرا و موهانتی [3] است و شامل سه مرحله زیر است:

- مرحله تقسیم: در این مرحله با توجه به معیار پراکندگی، مجموعه داده‌ای به  $\sqrt{N}$  زیرخوشه متراکم افزای می‌شود.
- مرحله تسخیر: در این مرحله نماینده‌ای از هر زیرخوشه مرحله تقسیم، به عنوان مرکز هر یک در نظر گرفته می‌شود. سپس گراف کامل وزن‌دار که راس‌های آن مراکز زیرخوشه‌ها و وزن یال‌های آن برابر با فاصله اقلیدسی دو راس آن یال است تشکیل می‌شود و در انتها، کوچک‌ترین درخت پوشا به دست می‌آید.
- مرحله ادغام: در این مرحله، شاخص ادغام برای هر جفت زیرخوشه مجاور در کوچک‌ترین درخت پوشای مرحله تسخیر، محاسبه می‌شود. این شاخص، میزان تمایل دو زیرخوشه برای تبدیل شدن به یک زیرخوشه بزرگ‌تر را تعیین می‌کند. زیرخوشه‌های تولید شده در مرحله تقسیم، بر مبنای کوچک‌ترین درخت پوشا و شاخص ادغام، تبدیل به زیرخوشه‌های بزرگ‌تر می‌شوند. این ادغام تا زمان رسیدن به تعداد خوشه مطلوب ادامه می‌یابد.

در ادامه هر یک از این مراحل با جزئیات بیشتر ارائه شده است. **مرحله تقسیم:** در مرحله تقسیم، مجموعه داده‌ای به تعدادی زیرمجموعه افزای می‌شود. هدف از این مرحله، شناسایی زیرخوشه‌های زیاد به کمک محاسبات ساده است. این کار بر اساس میزان پراکندگی<sup>1</sup> نقاط داده‌ای انجام می‌شود. اگر مجموعه داده‌ای مورد نظر  $N$  عضو داشته باشد. مرحله تقسیم، با مجموعه داده‌ای مورد نظر به عنوان یک زیرخوشه، شروع و تازمانی که

<sup>1</sup> Dispersion

فرمول زیر استفاده می‌شود:

$$T(C_i, C_j) = \frac{|C_{ij}|}{|C_i| + |C_j|} \frac{1}{\sum_{x \in C_{ij}} d(x, \mu_{ij})} \quad (6)$$

که در آن

$$C_{ij} = \{x \in C_i \cup C_j \mid d(x, \mu_{ij}) < d(\mu_{ij}, \mu(C_i))\}, \quad (7)$$

و  $\mu_{ij} = 0.5(\mu(C_i) + \mu(C_j))$  اعضای مجموعه  $C_{ij}$  نقاطی از زیرخوشه‌های مذکور است که داخل دایره‌ای به مرکز نقطه میانه و با شعاع نصف فاصله دو مرکز قرار می‌گیرند [3]. در ادامه به این نقاط، نقاط بین دو زیرخوشه می‌گوییم.

شاخص چسبندگی دو زیرخوشه، یعنی  $T(C_i, C_j)$  بر اساس نقاط بین دو زیرخوشه اندازه‌گیری می‌شود. بخش اول شاخص چسبندگی، یعنی  $\frac{|C_{ij}|}{|C_i| + |C_j|}$  نسبت تعداد نقاط بین دو زیرخوشه به کل نقاط دو زیرخوشه است. بنابراین هر اندازه تعداد نقاط بیشتری بین دو زیرخوشه قرار بگیرد، آن دو زیرخوشه تمایل بیشتری به ادغام شدن دارند و به اصطلاح تمایل چسبندگی بیشتری دارند. مخرج بخش دوم شاخص چسبندگی، یعنی  $\frac{\sum_{x \in C_{ij}} d(x, \mu_{ij})}{|C_{ij}|}$  میانگین فاصله نقاط بین دو زیرخوشه با میانه زیرخوشه‌ها است. هر اندازه این میانگین کوچکتر باشد به این معنا است که نقاط بین دو زیرخوشه به میانه نزدیک‌تر هستند و بنابراین دو زیرخوشه تمایل بیشتری برای ادغام دارند. با توجه به نکاتی که بیان شد، نتیجه می‌شود که دو زیرخوشه‌ای که نقاط بین آنها بیشتر و نزدیک‌تر به میانه متناظرشان دارند زودتر ادغام می‌شوند.

میشرا و موهانتی [3] شاخص چسبندگی را به عنوان معیاری برای سنجش میزان تمایل ادغام دو زیرخوشه، به صورت زیر معرفی کردند:

$$cohesion(C_i, C_j) = \frac{\sum_{x \in C_{ij}} d(x, \mu_{ij})}{|C_i| + |C_j|}, \quad (8)$$

و طبق ادعای آنها، هر اندازه این شاخص کمتر شود چسبندگی دو زیرخوشه و در واقع تمایل برای ادغام آنها بیشتر خواهد شد. اما به چند دلیل که در ادامه اشاره خواهد شد، این معیار دارای

**مرحله تسخیر:** در این مرحله، ابتدا گرافی کامل بر روی زیرخوشه‌های به دست آمده از مرحله تقسیم ساخته شود. در این گراف مرکز زیرخوشه‌ها به عنوان راس و فاصله اقلیدسی بین مراکز به عنوان وزن یال بین دو راس در نظر گرفته می‌شود. پس از تشکیل این گراف کامل، کوچک‌ترین درخت پوشا را می‌توان با استفاده از الگوریتم‌های کراسکال<sup>1</sup> و پرایم<sup>2</sup> به دست آورد [19]. در این پژوهش از الگوریتم کراسکال استفاده شده است که یک الگوریتم حریصانه برای پیدا کردن کوچک‌ترین درخت پوشا است. هر دو زیرخوشه‌ای که در کوچک‌ترین درخت پوشا به وسیله یک یال به هم متصل شده باشند، جفت زیرخوشه مجاور نامیده می‌شوند.

**مرحله ادغام:** در این مرحله، جفت‌های زیرخوشه مجاور در مرحله‌ی تسخیر با یکدیگر ادغام می‌شوند. اما یک زیرخوشه ممکن است چندین مجاور داشته باشد در صورتی که تنها یکی از آنها برای ادغام مناسب است. بنابراین لازم است که برای انتخاب بهترین مجاور هر زیرخوشه یک شاخص ادغام کارآمد پیشنهاد شود. در این پژوهش از یک شاخص ادغام کارآمد به شرح زیر استفاده شده است.

چسبندگی<sup>3</sup>، شباهت یا عدم تشابه دو زیرخوشه را بر اساس تراکم و توزیع نقاط داده در ناحیه میانی مراکز آنها اندازه‌گیری می‌کند. هرچه ناحیه میانی پرجمعیت‌تر باشد، زیرخوشه‌ها منسجم‌تر خواهند بود. سطح چسبندگی بین زیرخوشه‌ها با میزان پراکندگی نقاط داده‌ها نسبت به نقطه میانی مراکز دو زیرخوشه اندازه‌گیری می‌شود.

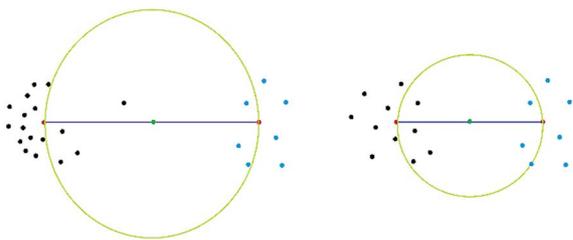
جفت زیرخوشه‌ی مجاور  $C_i$  و  $C_j$  را در نظر بگیرید. فرض کنید  $\mu(C_i)$  و  $\mu(C_j)$  به ترتیب مراکز زیرخوشه‌های  $C_i$  و  $C_j$  و  $\mu_{ij}$  نقطه میانه مراکز باشد. در این صورت نقاطی از دو زیرخوشه مذکور که فاصله آنها تا نقطه میانی در مقایسه با فاصله مراکز و نقطه میانه کمتر است را در مجموعه‌ای به نام  $C_{ij}$  قرار دهید. در این پژوهش، برای محاسبه چسبندگی این جفت زیرخوشه از

<sup>1</sup> Kruskal

<sup>2</sup> Prim

<sup>3</sup> Cohesion

بگیرید، می‌توان بدون تغییر یا با تغییر اندک در صورت کسر، مقدار مخرج کسر را به نحوی افزایش داد که معیار چسبندگی کوچک‌تر شود. برای روشن شدن این نکته نیز شکل (2) را مشاهده و مقایسه نمایید. در شکل (2) سمت راست یک جفت زیرخوشه با تعدادی نقطه داده‌ای در فضای بین دو زیرخوشه نشان داده شده است و در سمت چپ آن، همان دو زیرخوشه با تغییرات زیر را مشاهده می‌نمایید. در جفت زیرخوشه سمت چپ، به نقاط بین دو زیرخوشه تعداد کمی افزوده شده است که طبق دلیل اول باعث افزایش معیار چسبندگی خواهد شد. اما تغییر دیگر جفت زیرخوشه سمت چپ، این است که تعداد زیادی نقطه به‌غیر از نقاط بین دو زیرخوشه به یکی از زیرخوشه‌ها افزوده شده است که اثر خود را در مخرج کسر معیار چسبندگی خواهد داشت و باعث کوچک شدن آن خواهد شد به نحوی که اثر افزایش صورت کسر را از بین می‌برد. علی‌رغم این‌که دو زیرخوشه سمت راست نسبت به زیرخوشه‌های سمت چپ به یکدیگر نزدیک‌تر هستند، اما معیار چسبندگی با توجه به استدلالی که ارائه شد ادغام زیرخوشه‌های سمت چپ را در اولویت قرار می‌دهد چون مقدار کمتری از این معیار را نتیجه می‌دهند. لازم به ذکر است که نقطه افزوده شده در بین دو زیرخوشه سمت چپ، به خاطر تغییر نکردن نقطه مرکز زیرخوشه سمت چپ است.

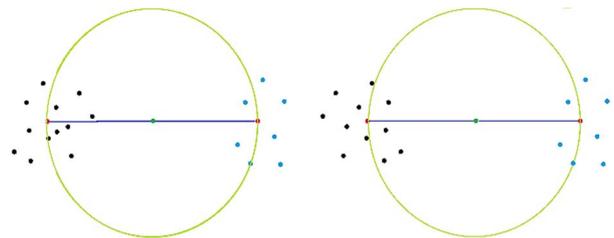


شکل 2: مقایسه شاخص‌های چسبندگی معرفی شده در [3] و پیشنهادی در این مقاله از لحاظ تعداد اعضای دو زیرخوشه. (منبع: محاسبات محقق)

با توجه به مباحث فوق در مجموع، معیار چسبندگی معرفی شده در مرجع [3] میزان چگالی نقاط بین دو زیرخوشه را اندازه‌گیری نمی‌کند و می‌توان نتیجه گرفت که شاخص چسبندگی معرفی شده در مقاله حاضر، در اندازه‌گیری چگالی

نقاط ضعف است. لازم به ذکر است که معیار پیشنهادی در این مقاله، یعنی  $T$ ، هرچه بزرگ‌تر باشد احتمال ادغام دو زیرخوشه بیشتر می‌شود. همچنین معیار  $T$  با درصد نقاط داده‌ای بین دو زیرخوشه رابطه مستقیم دارد و با میانگین فاصله نقاط بین دو زیرخوشه تا میانه مراکز دو زیرخوشه به طور معکوس مرتبط است.

دلیل اول برای ارجحیت معیار  $T$  این است که، با توجه به این که  $d(x_i, \mu_{ij})$  همواره یک مقدار مثبت است، به وضوح اگر تعداد نقاط میان دو زیرخوشه افزایش یابد، مقدار چسبندگی نیز بیشتر می‌شود و بنابراین مقدار آن رابطه مستقیم با تعداد نقاط در این ناحیه خواهد داشت. برای روشن شدن بیشتر دلیل اول، در شکل (1)، دو جفت زیرخوشه نشان داده شده است. جفت زیرخوشه سمت چپ، همان جفت زیرخوشه سمت راست این شکل است، با این تفاوت که زیرخوشه با داده‌های مشکی رنگ نسبت به مرکز آن چرخش 180 درجه داشته است. نتیجه این چرخش باعث افزایش تعداد نقاط بین جفت زیرخوشه شده است. با توجه به نحوه محاسبه چسبندگی مقدار این معیار برای جفت زیرخوشه سمت راست شکل (1) از مقدار آن برای جفت سمت چپ شکل (1) کمتر است و بنابراین جفت خوشه سمت چپ شکل (1) مقدم بر ادغام هستند، در حالی که چگالی نقاط بین دو زیرخوشه سمت راست بیشتر است. اما بر اساس معیار پیشنهادی در این مقاله، جفت زیرخوشه سمت چپ نسبت به جفت سمت راست اولویت ادغام بالاتری دارند.



شکل (1): مقایسه شاخص‌های چسبندگی معرفی شده در [3] و پیشنهادی در این مقاله از لحاظ تعداد نقاط بین دو زیرخوشه. (منبع: محاسبات محقق)

از طرف دیگر مخرج کسر معیار چسبندگی مجموع تعداد دو زیرخوشه است. در نتیجه، اگر یک جفت زیرخوشه را در نظر

علاوه بر معیار چسبندگی، معیار شباهت درون خوشه‌ای نیز مهم است و در بسیاری از تحقیقات از آن همزمان با معیارهای دیگر استفاده می‌شود. این معیار سبب می‌شود که زیرخوشه‌هایی ادغام شوند که تراکم نقاط درونی بالایی داشته باشند. شباهت درونی بر اساس تشابه نقاط داده در زیرخوشه‌ها تعریف می‌شود. رویکرد پیشنهادی با محاسبه میانگین تشابه نقاط داده در زیرخوشه‌ها نزدیکی دو زیرخوشه را اندازه‌گیری می‌کند. به زبان ساده‌تر از تشابه نقاط یک زیرخوشه با یکدیگر برای ارزیابی میزان شباهت این نقاط با نقاط زیرخوشه مجاور استفاده می‌شود.

در این پژوهش از معیار شباهت درون خوشه‌ای معرفی شده در مرجع [3] و [21] استفاده شده است. برای این کار در ابتدا کوچک‌ترین درخت پوشا روی تمام نقاط زیرخوشه تشکیل می‌شود. فرض کنید  $MST_i = (E_i, V_i)$  درخت مینیمم پوشای زیرخوشه  $C_i$  باشد. در این صورت شاخص تشابه این زیرخوشه با  $S(C_i)$  نمایش و به صورت زیر محاسبه می‌شود:

$$S(C_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{D^2}{2}} \quad (9)$$

که در آن

$$D = \frac{1}{|E_i|} \sum_{e_i \in E_i} |e_i| \quad (10)$$

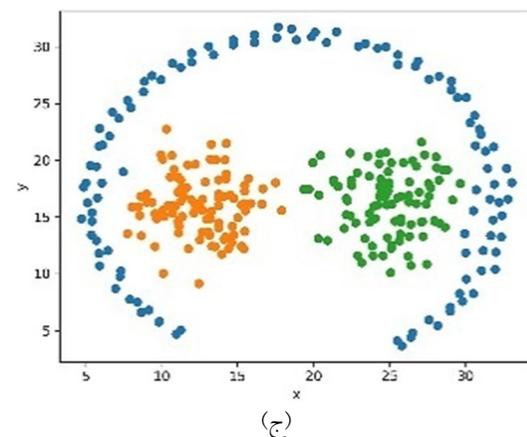
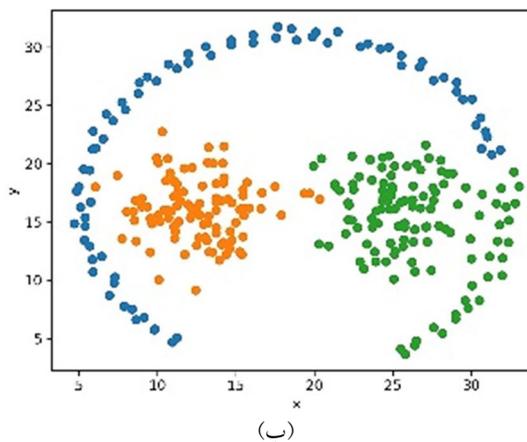
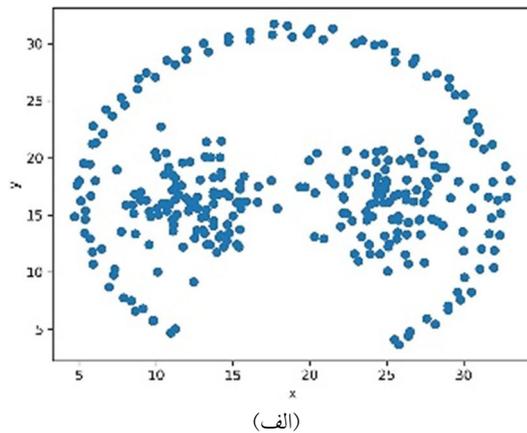
و  $|e_i|$  وزن یال  $e_i$  است. در شاخص تشابه یک زیرخوشه از  $D$  به عنوان میانگین وزنی یال‌های  $MST_i$  زیرخوشه استفاده شده است. همچنین به منظور ادغام زودتر زیرخوشه‌های با تعداد عضو کمتر، از عامل معکوس تعداد اعضای هر یک از دو زیرخوشه مجاور استفاده می‌کنیم.

پس از اندازه‌گیری چسبندگی و شباهت درونی جفت زیرخوشه‌ها شاخص ادغام برای انتخاب بهترین جفت برای ادغام محاسبه می‌شود. با فرض  $C_i$  و  $C_j$  به عنوان یک جفت زیرخوشه شاخص ادغام به شرح زیر محاسبه می‌شود:

$$MI(C_i, C_j) = T(C_i, C_j) \cdot \frac{S(C_i)}{|C_i|} \cdot \frac{S(C_j)}{|C_j|} \quad (11)$$

شاخص ادغام، چسبندگی و شباهت درونی یک جفت زیرخوشه

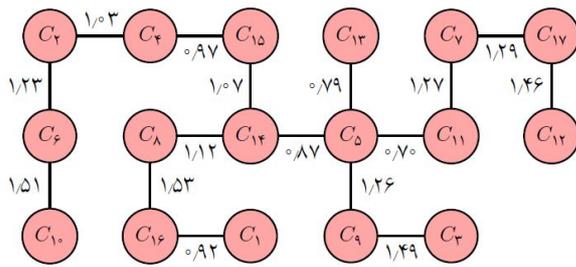
نقاط بین زیرخوشه‌ها بهتر از معیار معرفی شده در مرجع [3] عمل می‌کند. به منظور مقایسه بهتر، یک مجموعه داده‌ای دو بعدی از منبع [20] را که توسط روش پیشنهادی و روش معرفی شده در مرجع [3] خوشه‌بندی شده است، در شکل (3) مشاهده می‌کنید.



شکل (3): مقایسه روش خوشه‌بندی پیشنهادی و روش مرجع [3] بر روی مجموعه داده‌ای معرفی شده در مرجع [21]. (الف) شامل داده‌های خام، (ب) روش مرجع [3] و (ج) روش پیشنهادی (پایین). (منبع: مرجع [21] و محاسبات محقق)

#### 1.4. خوشه‌بندی قیمت سهام به روش پیشنهادی

بر اساس روش پیشنهادی، ابتدا 295 شرکت منتخب به 17 زیرخوشه تقسیم شدند. برای تقسیم‌بندی شرکت‌ها به زیرخوشه‌ها، از روابط (1)-(5) استفاده شد. سپس مراکز زیرخوشه‌ها به عنوان راس گراف در نظر گرفته می‌شوند. بنابراین گرافی کامل شامل 17 راس و 136 یال ایجاد می‌شود که هر یال آن وزن دارد که همان فاصله اقلیدسی مرکز دو زیرخوشه راس‌های متناظر است. پس از تشکیل این گراف کامل، الگوریتم کراسکال بر روی این گراف اجرا می‌شود تا کوچک‌ترین درخت پوشا حاصل شود، که در شکل (4) رسم شده است. در شکل (4)، زیرخوشه‌ها شماره‌گذاری شده‌اند و راس‌های درخت پوشا را تشکیل می‌دهند. همچنین فاصله اقلیدسی مراکز زیرخوشه‌های مجاور بر روی یال‌های درخت پوشا نوشته شده است.



شکل (4): نمودار کوچک‌ترین درخت پوشا اولیه بر روی مراکز زیرخوشه‌ها. (منبع: محاسبات محقق)

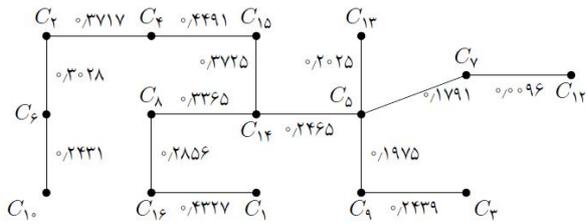
در مرحله ادغام، برای هر جفت زیرخوشه مجاور طبق رابطه (6) ابتدا چسبندگی و سپس از رابطه (9) تشابه درونی هر زیرخوشه محاسبه شد. در انتها با داشتن چسبندگی و شباهت درونی هر جفت زیرخوشه مجاور، شاخص ادغام آنها از رابطه (11) به دست آمد. شاخص ادغام میان جفت‌های زیرخوشه مجاور بر روی درخت پوشای شکل (5) نشان داده شده است. در ادامه روش پیشنهادی، در هر گام جفت زیرخوشه مجاور که بیشترین مقدار شاخص ادغام را دارند در یکدیگر ادغام می‌شوند و یک زیرخوشه بزرگ‌تر را به وجود می‌آورند و این کار تا زمانی که تعداد خوشه مطلوب حاصل شود ادامه می‌یابد. با توجه به شکل (5)، بیشترین شاخص ادغام متعلق به جفت

را به طور همزمان در نظر می‌گیرد و یک جفت با بیشترین مقدار  $MI$  برای ادغام انتخاب می‌شود و جفت بعدی بر اساس  $MI$  جدید در نظر گرفته می‌شود. یعنی نتیجه ادغام یک جفت زیرخوشه به عنوان زیرخوشه‌ای جدید در نظر گرفته می‌شود. تکرار ادغام زیرخوشه‌ها تا زمانی که تعداد خوشه‌های مطلوب به دست آیند و یا با محاسبه  $MI$  جدید، هیچ‌گونه بهبودی در خوشه‌بندی ایجاد نشود ادامه می‌یابد.

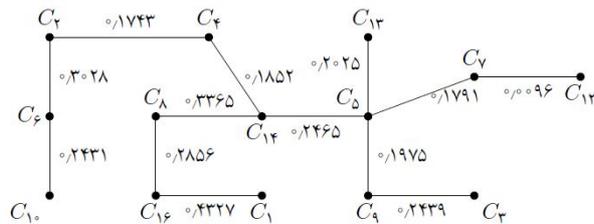
از آنجا که ساختار روش پیشنهادی در این مقاله مشابه روش معرفی شده توسط میشررا و موهانتی [3] می‌باشد، بنابراین پیچیدگی محاسباتی آن نیز از مرتبه  $O(N^{3/2})$  است.

#### 4. جامعه و نمونه آماری

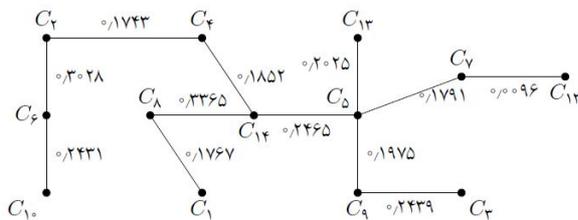
در این پژوهش، بورس اوراق بهادار تهران به عنوان جامعه آماری در نظر گرفته شده است. اگرچه متغیرهای فراوانی در بورس اوراق بهادار در مورد سهام شرکت‌ها موجود است اما با توجه به هدف پوشش سهام اکثر شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران، در این پژوهش متغیر اصلی مورد استفاده، قیمت تعدیل یافته بسته شدن سهام هر شرکت است. دلیل استفاده از قیمت تعدیل شده سهام این است که افزایش، کاهش و همچنین سود نقدی سرمایه در این قیمت لحاظ می‌شود. نمونه آماری پژوهش حاضر، سری زمانی قیمت پایانی (بسته شدن) سهام شرکت‌های موجود در بورس تهران در بازه زمانی از 1398/07/01 تا 1399/08/20 را شامل می‌شود. در بازه زمانی مذکور، تعداد 354 شرکت در بورس اوراق بهادار تهران پذیرفته شده‌اند و همچنین هر شرکت در طول بازه زمانی مورد نظر، شامل 273 مقدار است. با توجه به اینکه برای اجرای الگوریتم تعداد داده همه شرکت‌ها باید برابر باشند، شرکت‌هایی که به هر دلیلی این تعداد داده نداشتند، از لیست شرکت‌های مورد بررسی حذف شدند. در نهایت تعداد 295 شرکت برای خوشه‌بندی مورد استفاده قرار گرفت. لازم به ذکر است که در این پژوهش از نرم‌افزار R برای تجزیه، تحلیل و کدنویسی الگوریتم پیشنهادی استفاده شده است.



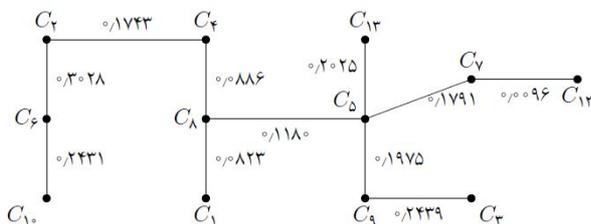
(آ) درخت پوشا بعد از ادغام زیرخوشه  $C_{17}$  در زیرخوشه  $C_{12}$ .



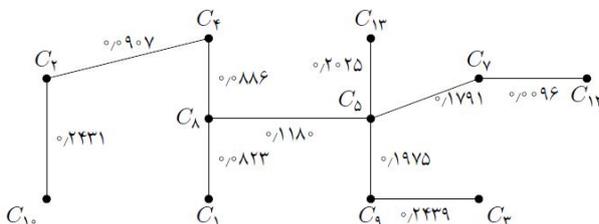
(ب) درخت پوشا بعد از ادغام زیرخوشه  $C_{15}$  در زیرخوشه  $C_4$ .



(ج) درخت پوشا بعد از ادغام زیرخوشه  $C_{16}$  در زیرخوشه  $C_8$ .



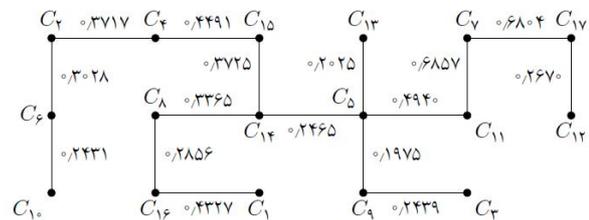
(د) درخت پوشا بعد از ادغام زیرخوشه  $C_{14}$  در زیرخوشه  $C_8$ .



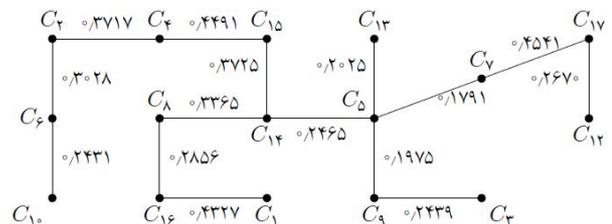
(ه) درخت پوشا بعد از ادغام زیرخوشه  $C_6$  در زیرخوشه  $C_2$ .

شکل (7): درخت پوشای اصلاح شده بعد از ادغام هر جفت از زیرخوشه‌ها. (منبع: محاسبات محقق)

زیرخوشه مجاور  $C_7$  و  $C_{11}$  با مقدار  $0/6857$  است و بنابراین  $C_7$  و  $C_{11}$  اولین دو زیرخوشه‌ای هستند که با یکدیگر ادغام می‌شوند و یال میان آنها حذف می‌شود. نام‌گذاری زیرخوشه جدید بر اساس کوچک‌ترین شماره دو زیرخوشه ادغام شده انجام می‌شود و بنابراین نام زیرخوشه جدید  $C_7$  گذاشته شده است که اکنون با دو زیرخوشه  $C_5$  و  $C_{17}$  مجاور می‌شود. این فرآیند به اختصار، ادغام زیرخوشه  $C_{11}$  در زیرخوشه  $C_7$  گفته می‌شود. با توجه به تغییرات به وجود آمده، بایستی شاخص ادغام بین زیرخوشه جدید و زیرخوشه‌های  $C_5$  و  $C_{17}$  محاسبه شود. شکل (6)، نتیجه این ادغام و مقادیر جدید شاخص ادغام را نشان می‌دهد.



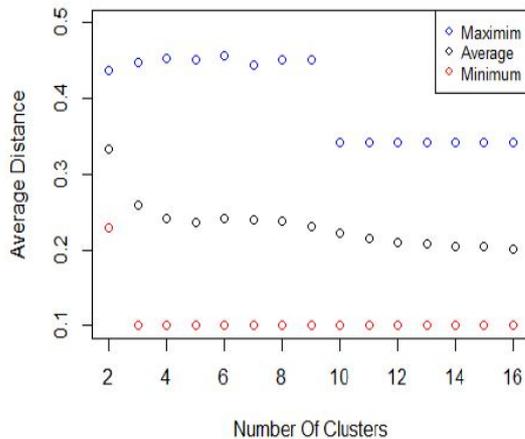
شکل (5): گراف کوچک‌ترین درخت پوشا حاصل از مرحله اول همراه با وزن‌های مربوط به شاخص ادغام. (منبع: محاسبات محقق)



شکل (6): نمودار درخت پوشا بعد از ادغام زیرخوشه  $C_{11}$  در زیرخوشه  $C_7$ . (منبع: محاسبات محقق)

به منظور خلاصه‌سازی، گام‌های طی شده تا رسیدن به تعداد مناسب زیرخوشه‌ها، در شکل (7) از بالا به پایین نشان داده شده است. شکل (7) - (آ)، بعد از ادغام زیرخوشه  $C_{17}$  در  $C_{12}$  به دست می‌آید. شکل (7) - (ب)، بعد از ادغام زیرخوشه  $C_{15}$  در  $C_4$ ؛ شکل (7) - (ج) بعد از ادغام زیرخوشه  $C_{16}$  در  $C_8$ ؛ شکل (7) - (د) بعد از ادغام زیرخوشه  $C_{14}$  در  $C_8$ ؛ شکل (7) - (ه) بعد از ادغام زیرخوشه  $C_6$  در  $C_2$ ؛ و در نهایت شکل (7) - (و) بعد از ادغام زیرخوشه  $C_9$  در  $C_5$  حاصل می‌شوند. این فرآیند تا رسیدن به تعداد مناسب خوشه ادامه می‌یابد.

به بقیه خوشه‌ها پراکندگی داده‌های بیشتری دارد.



شکل (8): روند فاصله‌ی اعضای خوشه‌ها بر اساس تعداد خوشه‌های تشکیل شده. (منبع: محاسبات محقق)

بررسی نتایج حاصل از روش پیشنهادی نشان می‌دهد که نمادهای فعال زیادی در زمینه غذا، ماشین آلات و برق، دارو و سرمایه‌گذاری، بانک و بیمه در خوشه‌های  $C_3$  و  $C_6$  مشاهده می‌شوند. همچنین خوشه‌های  $C_4$  و  $C_7$  بیشترین نمادهای فعال در زمینه فلزات اساسی، بانکی و سرمایه‌گذاری را در خود دارند. بیشترین نمادهای شیمیایی در خوشه‌های  $C_2$  و  $C_7$  قرار دارند و در خوشه  $C_2$  تعداد زیادی نماد خودروبی و فلزات اساسی وجود دارد که نشان از شباهت سری‌های زمانی این صنایع با صنایع شیمیایی دارد.

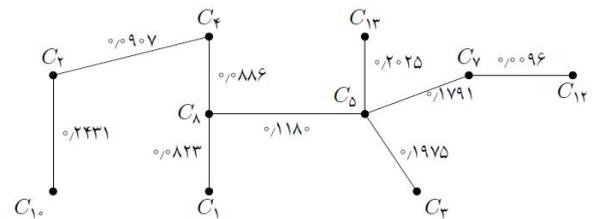
به منظور اعتباربخشی به نتایج حاصل از روش پیشنهادی، در ادامه با نتایج خوشه‌بندی به روش K-میانگین که بر مبنای فاصله اقلیدسی است مقایسه‌ای ارائه شده است. فرض کنید:

$$P = \{P_1, P_2, \dots, P_k\}, \quad (12)$$

مجموعه خوشه‌بندی روش پیشنهادی و

$$C = \{C_1, C_2, \dots, C_l\}, \quad (13)$$

مجموعه خوشه‌بندی حاصل از الگوریتم K-میانگین باشد. در این صورت معیار زیر در مرجع [3] مطرح شده است که در ادامه از این معیار برای ارزیابی و مقایسه عملکرد روش پیشنهادی استفاده می‌شود:

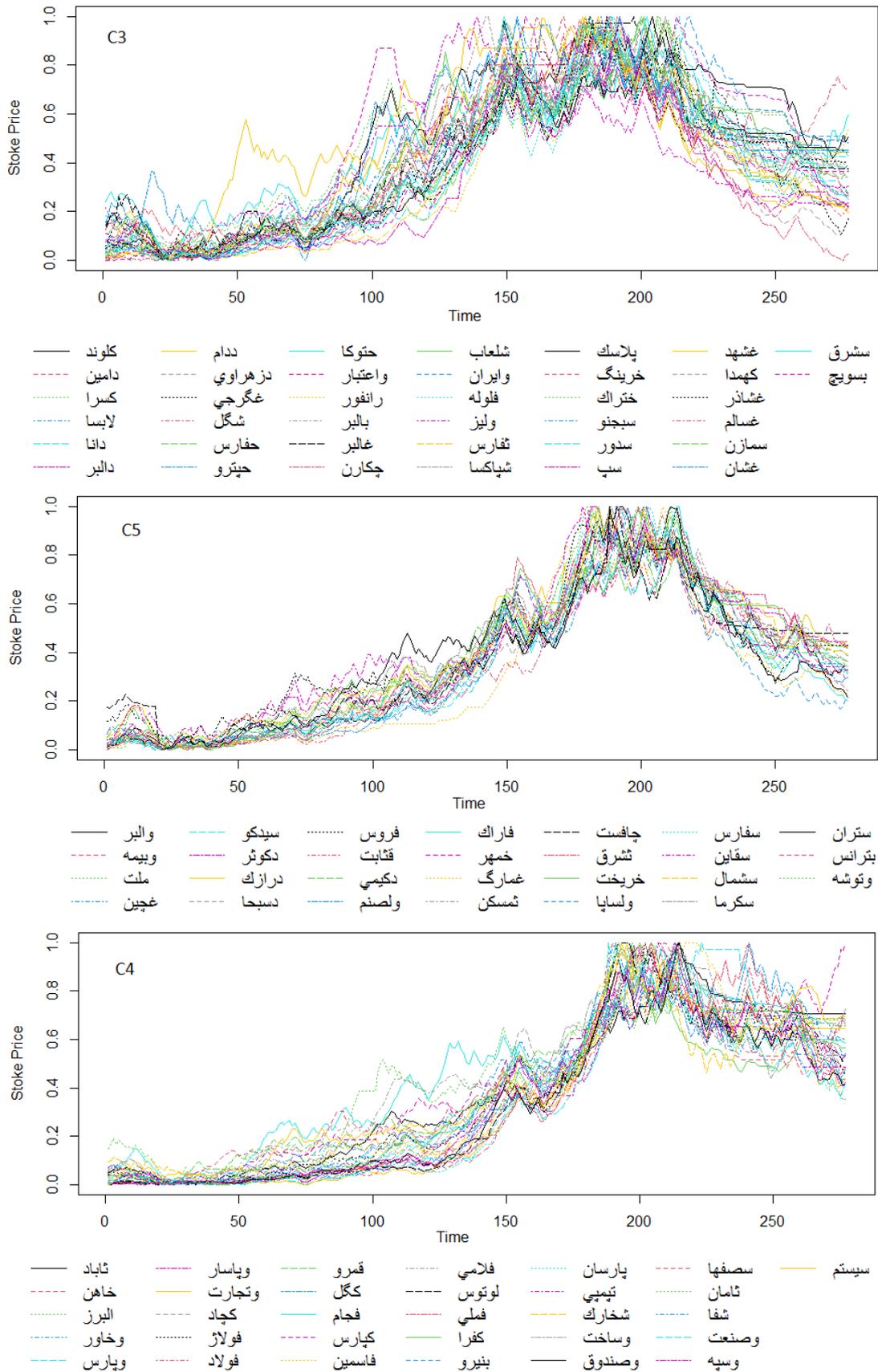


(و) درخت پوشا بعد از ادغام زیرخوشه  $C_9$  در زیرخوشه  $C_5$ .

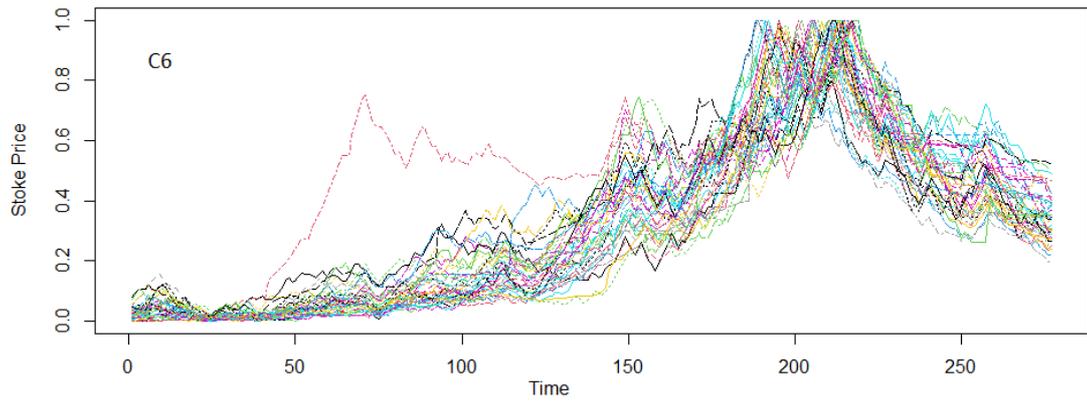
ادامه شکل (7): درخت پوشای اصلاح شده بعد از ادغام هر جفت از زیرخوشه‌ها. (منبع: محاسبات محقق)

در این پژوهش، ملاک زیر برای شناسایی تعداد مناسب خوشه‌ها لحاظ شده است. بعد از هر بار ادغام، ابتدا میانگین فاصله اعضای هر زیرخوشه موجود تا مرکز متناظر محاسبه می‌شود و پس از آن کمترین، بیشترین و مقدار متوسط میانگین‌های مذکور از زیرخوشه‌ها به دست می‌آید. در شکل (8)، این مقادیر در هر بار ادغام تا رسیدن به دو زیرخوشه نشان داده شده است. در شکل (8) مشاهده شد که میانگین فاصله اعضا با کم شدن تعداد زیرخوشه‌ها، بیشتر می‌شود و در واقع با ادغام شدن زیرخوشه‌ها، اعضای زیرخوشه جدید فاصله بیشتری از یکدیگر می‌گیرند. شکل (8) نشان می‌دهد که تعداد مناسب برای توقف مرحله ادغام، 10 زیرخوشه است زیرا با کاهش تعداد زیرخوشه‌ها به 9 تفاوت قابل توجهی مشاهده می‌شود. بنابراین تعداد خوشه مطلوب در این مورد 10 خوشه است و شکل (7) - (و) خوشه‌های نهایی را نشان می‌دهد. اکنون می‌توان شماره‌گذاری خوشه‌های نهایی را به  $C_1, C_2, \dots, C_{10}$  تغییر داد. برای این منظور، نام خوشه  $C_{12}$  در شکل (7) - (و) به  $C_6$  و نام خوشه  $C_{13}$  در شکل (7) - (و) به  $C_9$  تغییر یافت.

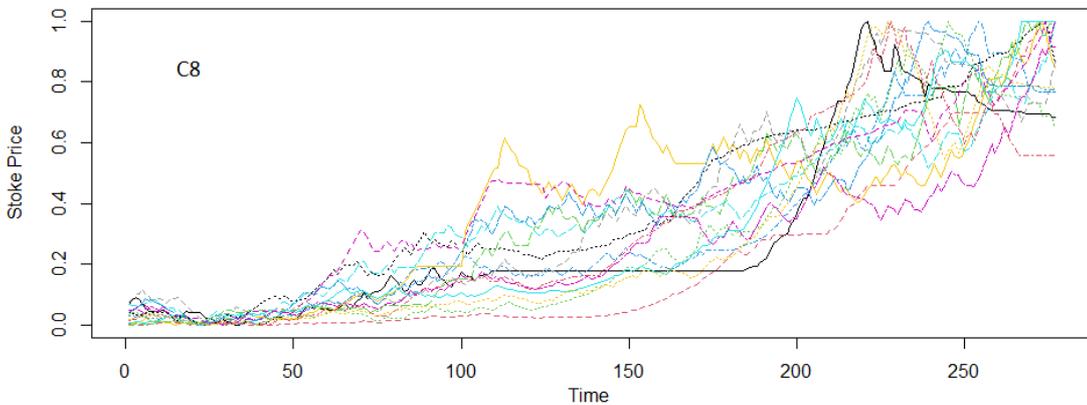
در شکل (9) سری زمانی نمادهای هر خوشه نشان داده شده است. در این شکل مشاهده می‌شود که خوشه‌های  $C_5, C_7, C_9$  و  $C_{10}$  خوشه‌های خوبی هستند و سری‌های زمانی آنها شباهت زیادی به یکدیگر دارند. به عبارت دیگر داده‌های آنها در اطراف سری زمانی میانگین خوشه‌ها و نزدیک به آن نوسان یا هم‌حرکتی دارند. خوشه‌های  $C_1$  تا  $C_4$  و  $C_6$  نسبت به خوشه‌های بالا نوسان بیشتر مشاهده می‌شود اما الگوی نسبتاً مناسبی بین داده‌های آنها همچنان وجود دارد. در نهایت خوشه  $C_8$  نسبت



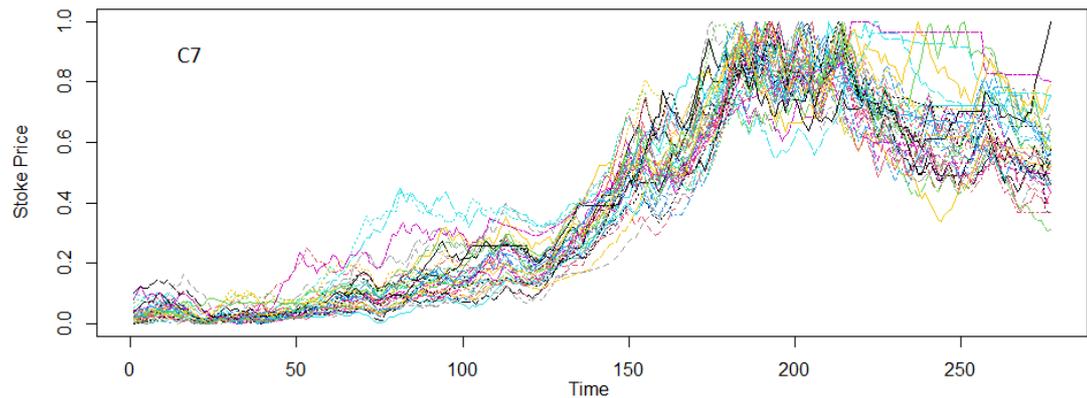
شکل (9): نمودار سری زمانی خوشه‌های نهایی تشکیل شده بعد از اجرای روش پیشنهادی. (منبع: محاسبات محقق)



- |        |        |       |        |       |       |       |
|--------|--------|-------|--------|-------|-------|-------|
| دعبید  | تاخت   | وغدیر | حفاری  | ونفت  | وبهن  | فسرب  |
| پردیس  | داسوه  | قصفها | های_وب | شنفت  | فسیا  | تایرا |
| خادین  | دسیحان | قزوین | مرقام  | شیریز | شسپا  | وسینا |
| وبوعلی | رتاب   | ختوقا | وخارزم | ورنا  | سفار  | زیارس |
| وبانک  | فخاس   | قهکمت | دیران  | رکیش  | قشکر  |       |
| لیوتان | عگل    | همراه | عیاک   | کروی  | سهرمز |       |

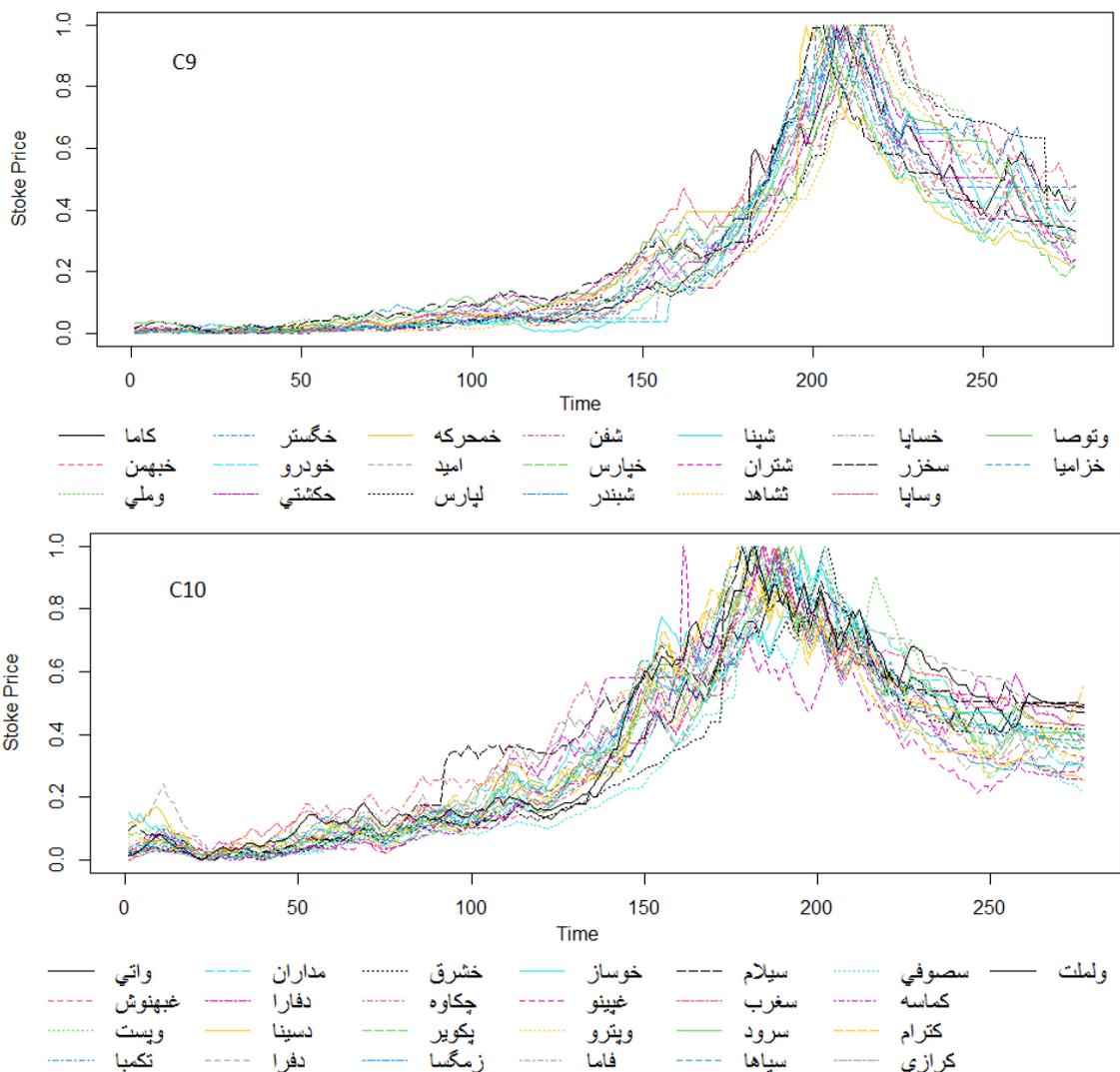


- |       |      |       |       |       |
|-------|------|-------|-------|-------|
| غدشت  | تکشا | کحافظ | ونیکی | شسینا |
| کدما  | بکاب | ولغدر | پاکشو | فینتا |
| فنیشا | غدام | فنورد | غاندر | وتوس  |



- |       |         |        |       |       |       |
|-------|---------|--------|-------|-------|-------|
| فایرا | شکرین   | وکار   | پارس  | کساوه | پتایر |
| آسینا | شدوص    | مبین   | دپارس | وسکاب | وآذر  |
| فباهر | چفیر    | خموتور | سبهان | سکرد  | وتوسم |
| بفجر  | وصنا    | شبهرن  | ساراب | شیران | تنوین |
| ما    | پارسیان | ونوین  | شفارس | سنیر  | وبشهر |
| چدن   | سخوز    | شاراک  | سهگمت | خیویش | پاسا  |

ادامه شکل (9): نمودار سری زمانی خوشه‌های نهایی تشکیل شده بعد از اجرای روش پیشنهادی. (منبع: محاسبات محقق)



ادامه شکل (9): نمودار سری زمانی خوشه‌های نهایی تشکیل شده بعد از اجرای روش پیشنهادی. (منبع: محاسبات محقق)

معیار  $RI$  بیش از 88 درصد می‌شود. بنابراین به ازای 10 خوشه، روش پیشنهادی و روش  $RI$  دارای تشابه عملکردی بیش از 88 درصد خواهند بود.

جدول (1): مقایسه روش پیشنهادی با روش  $K$ -میانگین به ازای تعداد خوشه گوناگون. (منبع: محاسبات محقق)

معیار $RI$	$P_{01}$	$P_{10}$	$P_{00}$	$P_{11}$	تعداد خوشه
0/6431	6758	8520	12723	15164	11
0/8804	5188	8722	12182	17273	10

## 2.4. ارزیابی الگوریتم پیشنهادی

در این بخش، روش پیشنهادی بر روی داده‌هایی که خوشه‌های

$$RI = \frac{P_{00} + P_{11}}{P_{00} + P_{01} + P_{10} + P_{11}} \quad (14)$$

که در آن  $P_{11}$  تعداد جفت داده‌هایی است که در هر دوی  $P$  و  $C$  در یک خوشه قرار دارند،  $P_{00}$  تعداد جفت داده‌هایی است که در خوشه‌های متفاوتی در هر دوی  $P$  و  $C$  قرار می‌گیرند،  $P_{10}$  تعداد جفت داده‌هایی که در  $P$  هم‌خوشه هستند اما در  $C$  هم‌خوشه نیستند و  $P_{01}$  تعداد جفت داده‌هایی که در  $C$  هم‌خوشه هستند اما در  $P$  هم‌خوشه نیستند. نتایج معیار  $RI$  به ازای تعداد خوشه مختلف در جدول (1) گزارش شده است. همان‌گونه که مشاهده می‌شود، به عنوان نمونه اگر تعداد خوشه مطلوب در روش پیشنهادی و روش  $K$ -میانگین، 10 در نظر گرفته شود

شرکت‌های پذیرش شده در بورس اوراق بهادار تهران استفاده شد. در روش پیشنهادی، ابتدا بر مبنای یک معیار پراکندگی مجموعه داده‌ای مذکور به تعداد زیادی زیرخوشه تقسیم شد. سپس بر اساس مرکز هر خوشه و فاصله اقلیدسی بین آنها به عنوان وزن، گراف کامل وزنداری ساخته شد و بر روی آن کوچک‌ترین درخت پوشا به کمک الگوریتم کراسکال به دست آمد. در انتها، بر مبنای معیار ادغام معرفی شده، زیرخوشه‌ها با یکدیگر ادغام شدند تا زمانی که تعداد مناسبی از خوشه‌ها به دست آیند. عملکرد الگوریتم پیشنهادی و نتایج آن با روش K- میانگین مقایسه شد، و نشان داده شد که انطباق خوبی بین دو روش وجود دارد. استفاده از نتایج الگوریتم پیشنهادی به منظور تشکیل صندوق‌های سهامی سودآور می‌تواند مورد استفاده قرار گیرد زیرا شرکت‌های عضو یک خوشه، رفتار قیمتی مشابهی دارند و تنها با بررسی نمایندگان آن خوشه، رفتار قیمتی سایر اعضای آن خوشه را تقریب زد. همچنین می‌توان با انتخاب سهام از خوشه‌های متفاوت، سبکی طراحی کرد که از تنوع بالایی برخوردار باشد. استفاده از شاخص‌های الگوریتم پیشنهادی توسط اشخاص و یا سازمان بورس اوراق بهادار تهران به منظور ارزیابی روند بازار نیز می‌تواند مورد توجه قرار گیرد. با توجه به اینکه در این تحقیق از معیار فاصله اقلیدسی برای محاسبه فاصله بین داده‌ها استفاده شد، پیشنهاد می‌شود در تحقیقات آینده از سایر معیارهای فاصله همچون پیچش زمانی پویا استفاده شود.

تعارض منافع: نویسندگان اعلام می‌کنند که هیچ تعارض منافی ندارند.

آن از قبل مشخص هستند اعمال می‌شود. با توجه به این که در این مجموعه‌های داده‌ای، خوشه‌های مطلوب تعیین شده هستند می‌توان از معیار RI که در رابطه (14) ارائه شده است برای ارزیابی عملکرد روش خوشه‌بندی پیشنهادی بهره برد. برای این منظور، در رابطه (12)،  $P$  خوشه‌های مطلوب و در رابطه (13)،  $C$  خوشه‌های به دست آمده توسط روش پیشنهادی است. مجموعه داده‌های مورد استفاده در این بخش از برخی مجموعه داده‌های بررسی شده توسط میسرا و موهانتی [3] اقتباس شده است. اطلاعات مربوط به این مجموعه‌های داده‌ای به همراه نتایج حاصل از اعمال روش پیشنهادی، در جدول (2) گزارش شده‌اند. همچنین به منظور مقایسه بهتر، در جدول (2)، نتایج حاصل از روش فاصله چگالی گاوسی که در مرجع [3] گزارش شده است و همچنین روش سه‌مرحله‌ای میسرا و موهانتی [3] را مشاهده می‌کنید. همان‌گونه که در جدول (2) مشاهده می‌شود، در مورد سه مجموعه داده‌ای Wine، Iris و Glass، روش پیشنهادی در مقایسه با روش [3] و روش فاصله چگالی گاوسی عملکرد بهتری داشته است و در مورد مجموعه داده‌ای Tumor، روش پیشنهادی با اختلاف کمی نسبت به روش [3] و بهتر از روش فاصله چگالی گاوسی عمل کرده است.

جدول (2): مقایسه عملکرد روش پیشنهادی و روش [3] بر روی مجموعه‌های داده‌ای مختلف بر مبنای معیار RI. (منبع: محاسبات محقق و [3])

مجموعه داده‌ای	تعداد خوشه مطلوب	تعداد نقاط داده‌ای	ابعاد داده‌ها	GDD	روش [3]	روش پیشنهادی
Iris	3	150	4	0/7719	0/8247	0/8514
Wine	3	178	13	0/6776	0/7495	0/7847
Glass	7	214	9	0/2598	0/7492	0/8301
Tumor	22	339	17	0/8916	0/9129	0/8973

## 5. نتیجه‌گیری

در این پژوهش از الگوریتم سه مرحله‌ای مبتنی بر کوچک‌ترین درخت پوشا برای خوشه‌بندی سری‌های زمانی قیمت سهام

- [1] M. Saeidi Kousha and S. Mohebbi, "Optimizing stock portfolios by comparing different technical patterns," *Financ. Eng. Portfolio Manage.*, vol. 12, no. 49, pp. 104-125, 2021, dor: 20.1001.1.22519165.1400.12.49.5.7 [In Persian].
- [2] D. Farid and M. Pourhamidi, "Classifying stocks of listed companies on Tehran Stock Exchange using fuzzy cluster analysis," *J. Financ. Account. Res.*, vol. 4, no. 3, pp. 105-128, 2012, dor: 20.1001.1.23223405.1391.4.3.8.8 [In Persian].
- [3] G. Mishra and S.K. Mohanty, "A fast hybrid clustering technique based on local nearest neighbor using minimum spanning tree," *Expert Syst. Appl.*, vol. 132, pp. 28-43, 2019, doi: 10.1016/j.eswa.2019.04.048.
- [4] J.D. Cryer and K.S. Chan, *Time Series Analysis, with Applications in R*. New York, NY, USA: Springer, 2008, doi: 10.1007/978-0-387-75959-3.
- [5] A. Soroushyar and M. Akhlaghi, "The comparative assessment of data mining methods effectiveness to forecasting return and risk of stock in companies listed in Tehran stock exchange," *J. Financ. Account. Res.*, vol. 9, no. 1, 2017, doi: 10.22108/far.2017.21746 [In Persian].
- [6] Z. Shirazian, H. Nikoumaram, and T. Torabi, "Clustering of volatility and its asymmetry in Tehran Stock Exchange," *J. Invest. Knowl.*, vol. 9, no. 35, pp. 1-19, 2020. [In Persian]
- [7] K. Ghanaei, M. Ghanbari, B. Jamshidinavid, and A. Baghfalaki, "Modeling the Co-Movement of Stocks Between Returns with Negative and Positive Shocks of Sentiment Arising from the Imbalance of Orders Using a Tree-Stage Clustering Approach," *Adv. Math. Financ. Appl.*, vol. 10, no. 1, pp. 113-129, 2024, doi: 10.71716/amfa.2025.23011843.
- [8] B.M. Blau and T.G. Griffith, "Price clustering and the stability of stock prices," *J. Bus. Res.*, vol. 69, no. 10, pp. 3933-3942, 2016, doi: 10.1016/j.jbusres.2016.06.008.
- [9] S.R. Nanda, B. Mahanty, and M.K. Tiwari, "Clustering Indian stock market data for portfolio management," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8793-8798, 2010, doi: 10.1016/j.eswa.2010.06.026.
- [10] S.N. Zainol Abidin, S.H. Jaaman, M. Ismail, and A.S. Abu Bakar, "Clustering stock performance considering investor preferences using a fuzzy inference system," *Symmetry*, vol. 12, p. 1148, 2020, doi: 10.3390/sym12071148.
- [11] E. Gungor and A. Ozmen, "Distance and density based clustering algorithm using gaussian kernel," *Expert Syst. Appl.*, vol. 69, pp. 10-20, 2017, doi: 10.1016/j.eswa.2016.10.022.
- [12] B.B. Nair, P.K.S. Kumar, N.R. Sakthivel, and U. Vipin, "Clustering stock price time series data to generate stock trading recommendations: An empirical study," *Expert Syst. Appl.*, vol. 70, pp. 20-36, 2017, doi: 10.1016/j.eswa.2016.11.002.
- [13] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," *SIGMOD Rec.*, vol. 27, no. 2, pp. 73-84, 1998, doi: 10.1145/276305.276312.
- [14] C.R. Lin and M.S. Chen, "Combining partitional and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 145-159, 2005, doi: 10.1109/TKDE.2005.21.
- [15] G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68-75, 1999, doi: 10.1109/2.781637.
- [16] X. Wang, X. Wang, and D.M. Wilkes, "A divide-and-conquer approach for minimum spanning tree-based clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 7, pp. 945-958, 2009, doi: 10.1109/TKDE.2009.37.
- [17] C. Zhong, D. Miao, and P. Franti, "Minimum spanning tree based split-and-merge: A hierarchical clustering method," *Inf. Sci.*, vol. 181, no. 16, pp. 3397-3410, 2011, doi: 10.1016/j.ins.2011.04.013.
- [18] F. Zhao et al., "A similarity measurement for time series and its application to the stock market," *Expert Syst. Appl.*, vol. 182, p. 115217, 2021, doi: 10.1016/j.eswa.2021.115217.
- [19] R. Balakrishnan and K. Ranganathan, *A Textbook of Graph Theory*. New York, NY, USA: Springer, 2012.
- [20] P. Franti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," *Appl. Intell.*, vol. 48, no. 12, pp. 4743-4759, 2018.
- [21] A.K. Das and J. Sil, "Cluster validation using splitting and merging technique," in *Proc. Int. Conf. Comput. Intell. Multimedia Appl. (ICCIMA)*, 2007, pp. 56-60.