پذیرفته شده در مجله محاسبات نرم

خوشه‌بندی قیمت سهام با استفاده از الگوریتم کوچک‌ترین درخت پوشا

فائزه کلهری[1]، کارشناسی ارشد، سیدمحمد حسینی[2*]، استادیار

[1] گروه ریاضی – دانشکده علوم پایه ـ دانشگاه آیت الله بروجردی(ره) ـ بروجرد ـ ایران

[2] گروه ریاضی – دانشکده علوم پایه ـ دانشگاه آیت الله بروجردی(ره) ـ بروجرد ـ ایران ـ sm.hoseini@abru.ac.ir

چکیده: باتوجه به فعالیت‌های روز افزون 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 استفاده‌ی پژوهش، قیمت بسته‌شدن روزانه‌ی سهام شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران در بازه‌ی زمانی ۱۳۹۸/۷/۱ تا ۱۳۹۹/۸/۲۰ است. در این پژوهش در مرحله‌ی اول، تعدادی زیرخوشه شامل شرکت‌های مشابه از نظر رفتار قیمتی، تشکیل می‌شود پس از آن بر اساس برخی معیارهای تشابه، زیرخوشه‌ها ادغام و نتیجه‌ی مطلوب یعنی داشتن خوشه‌هایی شامل اعضایی با بیشترین مشابهت حاصل می‌شود. از جمله مزایای روش پیشنهادی این است که در این روش معیارهای تشابه به‌صورت موضعی محاسبه می‌شود و بنابراین محاسبات آن نسبت به سایر روش‌ها کمتر خواهد بود. نتایج این پژوهش نشان می‌دهد که فرآیند خوشه‌بندی مجموعه‌های حجیم به‌راحتی و با دقتی مطلوب قابل انجام و استفاده است.

واژه‌های کلیدی: خوشه‌بندی، درخت پوشا، قیمت سهام، نزدیک‌ترین همسایه، الگوریتم کراسکال.

* نویسنده مسئول، **sm.hoseini@abru.ac.ir**

© پیش نویس مجله علمی محاسبات نرم، برای داوری در مجله

Clustering of the stock price using minimum spanning tree

Faezeh Kalhori [1], Graduate Student, Sayyed Mohammad Hoseini [2*], Assistant Professor

[1] Department of Mathematics, Ayatollah Boroujerdi University, Boroujerd, Iran,

[2] Department of Mathematics, Ayatollah Boroujerdi University, Boroujerd, Iran, sm.hoseini@abru.ac.ir

Abstract: Due to the increasing activities of individuals and legal entities in the capital market and the transformation of this market into one of the most important economic drivers of any country, it is concluded that more knowledge regarding the selection of shares will undoubtedly lead to higher profitability. In this paper, clustering of stock price time series using the smallest spanning tree algorithm is suggested. The daily closing prices of the shares of the companies listed on the Tehran Stock Exchange from 09/23/2019 to 11/10/2020 are used as the dataset. In the first stage, we form some sub-clusters that include similar companies in terms of price behavior. Then, based on a similarity criterion, sub-clusters are merged until the desired clusters, which contain members with the most similarity, are achieved. The main advantage of the proposed method is that the similarity measures are calculated locally, resulting in lower computational costs compared to other methods. The results indicate that the method can easily perform the clustering process, especially for large datasets, with favorable accuracy.

Keywords: *Clustering, spanning tree; stock price; nearest neighbor;Kruskal algorithm.*

*** corresponding author, sm.hoseini@abru.ac.ir**

۱. مقدمه

خوشه‌بندی یکی از رویکردهای بااهمیت 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 ذاتی در یک مجموعه داده‌ی مشخص مورد استفاده قرار می‌گیرد. خوشه‌بندی را می‌توان به صورت «فرایند 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 رده‌ی افرازی، 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 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 شیء داده‌ای محاسبه شود. در نتیجه پیچیدگی محاسباتی پیدا کردن شبیه‌ترین شیء داده‌ای برای هر یک از اشیاء داده‌ای عضو مجموعه، از مرتبه‌ی $O(N)$ است و پیچیدگی محاسباتی کل مجموعه‌ی داده‌ای حداقل از مرتبه‌ی $O(N^2)$ 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 شیء داده‌ای به یک شیء داده‌ای مشخص، نیاز نیست تمام اعضای مجموعه‌ی داده‌ای بررسی شوند، بلکه بررسی تنها بخش کوچکی از نقاط اطراف آن کفایت می‌کند. بنابراین اگر بتوان ساختار شباهت اشیاء 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 پایه‌ی کوچک‌ترین درخت پوشا[1] است. این روش خوشه‌بندی شامل سه مرحله‌ی تقسیم، تسخیر[2] و ادغام خوشه‌ها است. در مرحله‌ی تقسیم، مجموعه‌ی 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 مرحله‌ی تسخیر، گراف کوچک‌ترین درخت پوشای متناظر با گراف وزن‌دار مرحله‌ی قبل محاسبه می‌شود، و در مرحله‌ی ادغام، زیرخوشه‌های تولید شده در مرحله‌ی تقسیم که در مرحله‌ی 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، سیاست‌گزاران و سرمایه‌گذاران بوده است. همان‌طور که وجود بازارهای مالی فعال و پویا در رشد اقتصادی یک 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 [۱ و ۲]. 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

[1] Minimum Spanning Tree (MST)
[2] Conquer

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 خوشه‌ی 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 به‌دست‌آوردن اطلاعات مفید درخصوص سهام یک شرکت، تعمیم مناسبی به شرکت‌های هم‌خوشه با آن را لحاظ کنند و تصمیم مناسبی برای کسب بیش‌ترین بازده اتخاذ کنند. مهم‌ترین دستاوردهای مقاله‌ی حاضر عبارتند از:

- معرفی و مقایسه‌ی معیار ادغام جدید و مناسب‌تری نسبت به معیار ادغام پیشنهاد شده توسط میشرا و موهانتی [۳]؛
- 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 به‌کار بستن روش‌های معتبر موجود برای پیش‌بینی یک یا چند نماینده از آن خوشه، رفتار سایر اعضای آن را هم تخمین بزند و به این ترتیب اطلاعات بسیاری در اختیار سهام‌دار خواهد بود.

۲. 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 [۴]. برای ارزیابی تشابه در مجموعه داده‌های سری‌های زمانی از معیارهای فاصله مخصوص سری زمانی، مانند پیچش زمانی پویا[1] و فاصله اقلیدسی بهره گرفته می‌شود.

فرید و پورحمیدی [۲] از داده‌های ۳۳۸ شرکت پذیرفته شده در بورس اوراق بهادار تهران در یک بازه زمانی ۵ ساله از ابتدای سال ۱۳۸۴ تا انتهای سال ۱۳۸۸ استفاده کردند. آنها با روش C-میانگین فازی، داده‌های صورت‌های مالی شرکت‌های مذکور شامل ۱۱ 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 خوشه‌ی سهام ترکیبی قرار می‌گیرند و گرایش آنها به خوشه‌ی سهام رشدی است.

سروش‌یار و اخلاقی [۵] با به‌کارگیری تکنیک‌های داده‌کاوی هم‌چون K-میانگین بر روی داده‌های ۱۰۷ 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 همکارن [۶] در سال ۱۳۹۹ نوسانات شاخص کل بورس اوراق بهادار تهران در بازه‌ی زمانی ۱۰ 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 ۱۳۹۴، اقبال‌نیا و همکاران [۷] از یک روش خوشه‌بندی سه‌مرحله‌ای برای خوشه‌بندی بر مبنای شباهت در روند حرکتی قیمت سهام شرکت‌های بورس اوراق بهادار تهران در دوره زمانی سال ۱۳۹۲ استفاده کردند. در روش مذکور، ابتدا کاهش ابعاد داده‌ها انجام می‌شود و خوشه‌بندی تقریبی توسط روش K-میانگین تعیین می‌شود. سپس خوشه‌های به‌دست آمده به زیرخوشه‌هایی با کیفیت‌تر تقسیم می‌شوند و در انتها زیرخوشه‌های مذکور با یک‌دیگر ادغام می‌شوند تا خوشه‌های نهایی حاصل شوند.

بلائو و گریفیت [۸] در سال ۲۰۱۶ 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

[1] Dynamic Time Wraping (DTW)

طبقه‌بندی می‌کند. نتایج تحقیق ناندا و همکاران [۹] 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 سرمایه‌گذارن نامشخص است در حالی که در واقع ترجیحات انتخاب سرمایه‌گذاران متفاوت است. برخی از سرمایه‌گذران بدبین هستند، در حالی که دیگران ممکن است خوش‌بین باشند. باتوجه به این مسئله، زین‌العابدین و همکارن [۱۰] روش جدید خوشه‌بندی فازی سازگار با عمل‌کرد سهام ارائه می‌دهد که بین سرمایه‌گذران بدبین و خوش‌بین تفاوت قائل می‌شود. این روش، سهام در نظر گرفته شده را بر اساس سطح اعتماد به‌نفس و ترجیحات سرمایه‌گذران رتبه‌بندی می‌کند و سپس خوشه‌بندی را بر روی سهام انجام می‌دهد. تحقیق مذکور بر روی ۳۰ شرکت در بورس اوراق بهادار مالزی انجام شده است.

گونگور و اوزمن [۱۱] یک روش خوشه‌بندی بدون پارامتر به نام فاصله چگالی گاوسی[1] برای تعیین خوشه‌ها معرفی کردند که با استفاده از هسته گاوسی و بر پایه‌ی فاصله و شکل خوشه‌ها عمل می‌کند. این روش پیشنهادی به‌منظور 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 [۱۲] از درخت رگرسیون برای کاهش ابعاد داده‌ها و از نگاشت خودسازمانده[2] برای خوشه‌بندی استفاده شده است. یکی دیگر از الگوریتم‌های خوشه‌بندی ترکیبی برای بررسی مجموعه داده‌های حجیم، الگوریتم خوشه‌بندی به کمک نماینده‌ها[3] است [۱۳]. در این الگوریتم هر خوشه با تعداد مشخصی از اشیاء داده‌ای آن نمایش داده می‌شود و شباهت هر دو زیرخوشه به‌وسیله نزدیک‌ترین فاصله بین نقاط نماینده زیرخوشه‌های مختلف اندازه‌گیری می‌شود. چسبندگی خود–ادغام[4] یک روش خوشه‌بندی قوی و کارآمد است که ابتدا از K–میانگین برای تقسیم‌بندی مجموعه داده به K زیرخوشه استفاده می‌کند و به‌طور تکراری آن‌ها را در خوشه‌های نهایی ادغام می‌کند [۱۴]. الگوریتم دیگری از خوشه‌بندی ترکیبی که می‌تواند مجموعه داده ناهمگن را خوشه‌بندی کند در [۱۵] 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 پایه‌ی کوچک‌ترین درخت پوشا است. وانگ و همکاران [۱۶]، روش‌های خوشه‌بندی مبتنی بر کوچک‌ترین درخت پوشا با استفاده از تقسیم و تسخیر برای ایجاد 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 به‌دست آوردن خوشه‌ها محاسبات کمتری نیاز دارد.

تقسیم و ادغام، نمونه‌ی دیگری از الگوریتم خوشه‌بندی ترکیبی است که توسط ژونگ و همکاران [۱۷] 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 مرتبه‌ی $O(N^{\frac{3}{2}})$ و بنابراین محاسبات روش پیشنهادی نسبت به روش پیمایش کل شبکه‌ی وزن‌دار مربوط به فواصل نقاط داده‌ای کاهش قابل ملاحظه‌ای می‌یابد.

اخیراً، ژاو و همکاران [۱۸] معیار شباهت اختصاصی چند رویکردی پویا[5] را معرفی کردند و از آن برای خوشه‌بندی و پیش‌بینی قیمت ۲۸۵ 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[6] را در پیچش زمانی پویا بجای فاصله‌ی 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

[1] Gauss Density Distance (GDD)

[2] Self-Organizing Map (SOM)
[3] Clustering Using Representatives (CURE)
[4] Cohesion Self-Merging (CSM)

[5] Dynamic multi-perspective personalized similarity measurement
[6] Canberra

در مقایسه با روش‌های دیگر از قبیل معیار اقلیدسی و پیچش زمانی پویا عملکرد مطلوب‌تری دارد.

۳. روش تحقیق

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 میشرا و موهانتی [۳] است و شامل سه مرحله‌ی زیر است:

- مرحله تقسیم: در این مرحله با توجه به معیار پراکندگی، مجموعه‌ی داده‌ای به $\sqrt{N}$ زیرخوشه‌ی متراکم افراز می‌شود.
- مرحله تسخیر: در این مرحله نماینده‌ای از هر زیرخوشه‌ی 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 فاصله‌ی اقلیدسی دو راس آن یال است تشکیل می‌شود و در انتها، کوچک‌ترین درخت پوشا به‌دست می‌آید.
- مرحله ادغام: در این مرحله، شاخص ادغام برای هر جفت زیرخوشه‌ی مجاور در کوچک‌ترین درخت پوشای مرحله‌ی تسخیر، محاسبه می‌شود. این شاخص، میزان تمایل دو زیرخوشه برای تبدیل شدن به یک زیرخوشه‌ی 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 خوشه‌ی مطلوب ادامه می‌یابد.

در ادامه هر یک از این مراحل با جزئیات بیشتر ارایه شده است.

مرحله تقسیم

در مرحله‌ی تقسیم، مجموعه‌ی داده‌ای به تعدادی زیرمجموعه افراز می‌شود. هدف از این مرحله، شناسایی زیرخوشه‌های زیاد به کمک محاسبات ساده است. این‌کار براساس میزان پراکندگی[1] نقاط داده‌ای انجام می‌شود. اگر مجموعه‌ی داده‌ای مورد نظر N عضو داشته باشد. مرحله‌ی تقسیم، با مجموعه‌ی داده‌ای مورد نظر به‌عنوان یک زیرخوشه، شروع و تازمانی که تعداد زیرخوشه‌ها به $\sqrt{N}$ تعداد برسد، فرآیند زیر تکرار می‌شود.

فرض کنید C_0، C_1، ...، C_{M-1} زیرخوشه‌های موجود باشند. اکنون بر اساس میزان پراکندگی، یکی از زیرخوشه‌ها برای تقسیم به دو زیرخوشه‌ی جدید انتخاب می‌شود [۳]:

$$m = \underset{0\le i<M}{argmax} \frac{1}{|C_i|}\sum_{x\in C_i}\|x-\mu(C_i)\|_2, \qquad (۱)$$

که در آن $|C_i|$ عدد اصلی خوشه‌ی i ام و $\|\cdot\|_2$ فاصله‌ی (نرم) اقلیدسی است. با توجه به رابطه‌ی (۱)، زیرخوشه‌ی C_m خوشه‌ای است که بیشترین پراکندگی را حول مرکز زیرخوشه نسبت به سایر زیرخوشه‌ها دارد. این زیرخوشه به روش زیر به دو زیرخوشه‌ی جدید C_{m1} و C_{m2} تقسیم می‌شود. فرض کنید $X = \{x_1, x_2, x_3, \cdots, x_N\}$ مجموعه نقاط زیرخوشه‌ی C_m باشد که در آن

$$x_i = \left(x_i^1, x_i^2, \cdots, x_i^d\right) \in \mathbb{R}^d \qquad (۲)$$

یک بردار ویژگی با d مولفه(ویژگی) است. در ادامه، مرکز (میانگین) خوشه‌ی C_m با $\mu(C_m) = \left(\bar{x}^1, \bar{x}^2, \cdots, \bar{x}^d\right)$ و واریانس مجموعه‌ی نقاط داده‌ای X بر روی ویژگی i ام با $Var(X^i)$ نشان داده می‌شود. قرار دهید

$$k = \underset{1\le i\le d}{\operatorname{argmax}} Var(X^i) \qquad (۳)$$

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

$$C_{m1} = \{x_i \in C_m |\, x_i^k < \bar{x}^k\}, \qquad (۴)$$

$$C_{m2} = C_m \backslash C_{m1}. \qquad (۵)$$

که در رابطه‌ی (۴)، $\bar{x}^k$ میانگین ویژگی kام و در رابطه‌ی (۵) نماد $\backslash$ نشان دهنده‌ی تفاضل مجموعه‌ها است و بنابراین C_{m2} شامل اعضایی از C_m است که در C_{m1} نیستند. به این‌ترتیب زیرخوشه‌ی C_m به دو زیرخوشه‌ی جدید تقسیم می‌شود.

مرحله تسخیر

در این مرحله، ابتدا گرافی کامل بر روی زیرخوشه‌های به‌دست آمده از مرحله‌ی تقسیم ساخته شود. در این گراف مرکز زیرخوشه‌ها به

[1] Dispersion

پذیرفته شده در مجله محاسبات نرم

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[1] و پریم[2] به‌دست آورد [۱۹]. 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 زیرخوشه‌ی 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 هرزیرخوشه یک شاخص ادغام کارآمد پیشنهاد شود. در این پژوهش از یک شاخص ادغام کارآمد به شرح زیر استفاده شده است.

چسبندگی[3]، شباهت یا عدم تشابه دو زیرخوشه را براساس تراکم و توزیع نقاط داده در ناحیه‌ی میانی مراکز آنها اندازه‌گیری می‌کند. هرچه ناحیه‌ی میانی پرجمعیت‌تر باشد، زیرخوشه‌ها منسجم‌تر خواهند بود. سطح چسبندگی بین زیرخوشه‌ها با میزان پراکندگی نقاط داده‌ها نسبت به نقطه‌ی 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 μ_{ij} نقطه میانه مراکز باشد. در این‌صورت نقاطی از دو زیرخوشه‌ی مذکور که فاصله‌ی آنها تا نقطه میانی در مقایسه با فاصله‌ی 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 فرمول زیر استفاده می‌شود:

$$T(C_i, C_j) = \frac{|C_{ij}|}{|C_i| + |C_j|} \frac{1}{\frac{\sum_{x \in C_{ij}} d(x, \mu_{ij})}{|C_{ij}|}} \qquad (۶)$$

که در آن

$$C_{ij} = \{x \in C_i \cup C_j \mid d(x, \mu_{ij}) < d(\mu_{ij}, \mu(C_i))\}, \qquad (۷)$$

و $\mu_{ij} = 0.5(\mu(C_i) + \mu(C_j))$ اعضای مجموعه‌ی C_{ij} نقاطی از زیرخوشه‌های مذکور است که داخل دایره‌ای به مرکز نقطه‌ی میانه و با شعاع نصف فاصله‌ی دو مرکز قرار می‌گیرند [۳]. 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 فاصله‌ی نقاط بین دو زیرخوشه با میانه‌ی 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 میانه‌ی متناظرشان دارند زودتر ادغام می‌شوند.

میشرا و موهانتی [۳] شاخص چسبندگی را به‌عنوان معیاری برای سنجش میزان تمایل ادغام دو زیرخوشه، به‌صورت زیر معرفی کردند

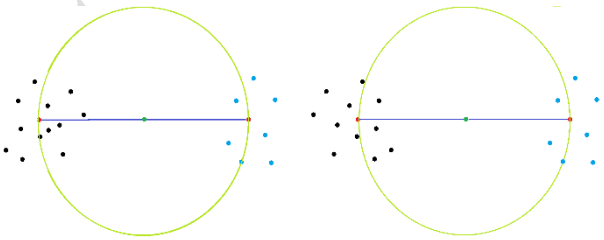
$$cohesion(C_i, C_j) = \frac{\sum_{x \in C_{ij}} d(x, \mu_{ij})}{|C_i| + |C_j|}, \qquad (۸)$$

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 نقاط ضعف است. لازم به‌ذکر 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 رابطه‌ی مستقیم دارد و با میانگین فاصله‌ی نقاط بین دو زیرخوشه تا میانه‌ی مراکز دو زیرخوشه به‌طور معکوس مرتبط است.

دلیل اول برای ارجحیت معیار T این است که، با توجه به این‌که $d(x, \mu_{ij})$ همواره یک مقدار مثبت است، به‌وضوح اگر تعداد نقاط

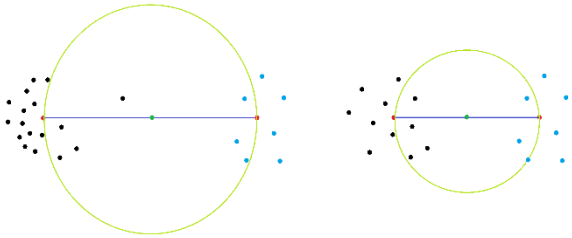
[1] Kruskal
[2] Prim
[3] Cohesion

میان دو زیرخوشه افزایش یابد، مقدار چسبندگی نیز بیشتر می‌شود و بنابراین مقدار آن رابطه‌ی مستقیم با تعداد نقاط در این ناحیه خواهد داشت. برای روشن شدن بیشتر دلیل اول، در شکل ۱ دو جفت زیرخوشه نشان داده شده است. جفت زیرخوشه‌ی سمت چپ، همان جفت زیرخوشه سمت راست این شکل است، با این تفاوت که زیرخوشه‌ی با داده‌های مشکی رنگ نسبت به مرکز آن چرخش ۱۸۰ درجه داشته است. نتیجه‌ی این چرخش باعث افزایش تعداد نقاط بین جفت زیرخوشه شده است. با توجه به نحوه‌ی محاسبه‌ی چسبندگی مقدار این معیار برای جفت زیرخوشه‌ی سمت راست شکل ۱ از مقدار آن برای جفت سمت چپ شکل ۱ کمتر است و بنابراین جفت خوشه‌ی سمت چپ شکل ۱ مقدم بر ادغام هستند، در حالی‌که چگالی نقاط بین دو زیرخوشه‌ی سمت راست بیشتر است. اما براساس معیار پیشنهادی در این مقاله، جفت زیرخوشه‌ی سمت چپ نسبت به جفت سمت راست اولویت ادغام بالاتری دارند.



شکل ۱: مقایسه شاخص‌های چسبندگی معرفی شده در [۳] و پیشنهادی در این مقاله از لحاظ تعداد نقاط بین دو زیرخوشه. (منبع: محاسبات محقق)

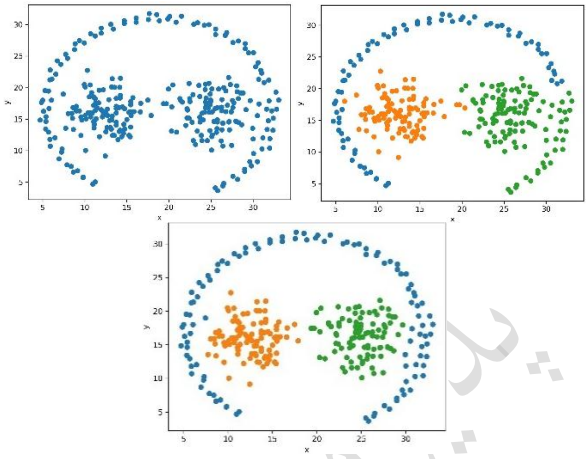
از طرف دیگر مخرج کسر معیار چسبندگی مجموع تعداد دو زیرخوشه است. در نتیجه، اگر یک جفت زیرخوشه را در نظر بگیرید می‌توان بدون تغییر یا با تغییر اندک در صورت کسر، مقدار مخرج کسر را به‌نحوی افزایش داد که معیار چسبندگی کوچک‌تر شود. برای روشن شدن این نکته نیز شکل‌های ۲ را مشاهده و مقایسه نمایید. در شکل ۲ سمت راست یک جفت زیرخوشه با تعدادی نقطه‌ی 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 زیرخوشه‌ی 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 زیرخوشه‌ی 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 به‌نحوی که اثر افزایش صورت کسر را از بین می‌برد. علی‌رغم این‌که دو زیرخوشه‌ی سمت راست نسبت به زیرخوشه‌های سمت چپ به یک‌دیگر نزدیک‌تر هستند، اما معیار چسبندگی با توجه به استدلالی که ارایه شد ادغام زیرخوشه‌های سمت چپ را در اولویت قرار می‌دهد چون مقدار کمتری از این معیار را نتیجه می‌دهند. لازم به‌ذکر است که نقطه‌ی افزوده شده در بین دو زیرخوشه‌ی سمت چپ، به‌خاطر تغییر نکردن نقطه‌ی مرکز زیرخوشه‌ی سمت چپ است.



شکل ۲: مقایسه شاخص‌های چسبندگی معرفی شده در [۳] و پیشنهادی در این مقاله از لحاظ تعداد اعضای دو زیرخوشه. (منبع: محاسبات محقق)

با توجه به مباحث فوق در مجموع، میعار چسبندگی معرفی شده در [۳] میزان چگالی نقاط بین دو زیرخوشه را اندازه‌گیری نمی‌کند و می‌توان نتیجه گرفت که شاخص چسبندگی معرفی شده در مقاله‌ی حاضر، در اندازه‌گیری چگالی نقاط بین زیرخوشه‌ها بهتر از معیار معرفی شده در مرجع [۳] عمل می‌کند. به‌منظور مقایسه‌ی بهتر، یک مجموعه داده‌ای دو بعدی از منبع [۲۰] را که توسط روش پیشنهادی و روش معرفی شده در [۳] خوشه‌بندی شده است، در شکل ۳ مشاهده می‌کنید.

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.



شکل ۳: مقایسه‌ی روش خوشه‌بندی پیشنهادی و روش [۳] بر روی مجموعه داده‌ای معرفی شده در [۲۱]، شامل داده‌های خام (بالا چپ)، روش [۳] (بالا راست) و روش پیشنهادی (پایین). (منبع: مرجع [۲۱] و محاسبات محقق)

در این پژوهش از معیار شباهت درون خوشه‌ای معرفی شده در [۳ و ۲۱] 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 زیرخوشه‌ی C_i باشد. در این‌صورت شاخص تشابه این زیرخوشه با $S(C_i)$ نمایش و به صورت زیر محاسبه می‌شود:

$$S(C_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{D^2}{2}} \quad (۹)$$

که در آن

$$D = \frac{1}{|E_i|} \sum_{e_i \in E_i} |e_i| \quad (۱۰)$$

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 زیرخوشه‌ی 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:

$$MI(C_i, C_j) = T(C_i, C_j) \cdot \frac{S(C_i)}{|C_i|} \cdot \frac{S(C_j)}{|C_j|}. \quad (۱۱)$$

شاخص ادغام ، چسبندگی و شباهت درونی یک جفت زیرخوشه را به طور همزمان در نظر می‌گیرد و یک جفت با بیش‌ترین مقدار MI برای ادغام انتخاب می‌شود و جفت بعدی براساس 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 به‌دست 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 میشرا و موهانتی [۳] می‌باشد، بنابراین پیچیدگی محاسباتی آن نیز از مرتبه‌ی $O(N^{3/2})$ است.

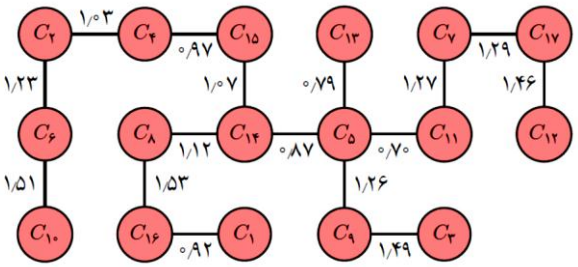
۴. جامعه و نمونه آماری

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 بسته‌شدن 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 (بسته‌شدن) سهام شرکت‌های موجود در بورس تهران در بازه زمانی از ۱۳۹۸/۰۷/۰۱ تا ۱۳۹۹/۰۸/۲۰ را شامل می‌شود. در بازه‌ی زمانی مذکور، تعداد ۳۵۴ شرکت در بورس اوراق بهادار تهران پذیرفته شده‌اند و همچنین هر شرکت در طول بازه‌ی زمانی موردنظر، شامل ۲۷۳ مقدار است. با توجه به اینکه برای اجرای الگوریتم تعداد داده همه شرکت‌ها باید برابر باشند، شرکت‌هایی که به هردلیلی این تعداد داده نداشتند، از لیست شرکت‌های مورد بررسی حذف شدند. در نهایت تعداد ۲۹۵ شرکت برای خوشه‌بندی مورد استفاده قرار گرفت. لازم به‌ذکر است که در این پژوهش از نرم‌افزار R برای تجزیه، تحلیل و کدنویسی الگوریتم پیشنهادی استفاده شده است.

خوشه‌بندی قیمت سهام به روش پیشنهادی

بر اساس روش پیشنهادی، ابتدا ۲۹۵ شرکت منتخب به ۱۷ زیرخوشه تقسیم شدند. برای تقسیم‌بندی شرکت‌ها به زیرخوشه‌ها، از روابط (۱)–(۵) استفاده شد. سپس مراکز زیرخوشه‌ها به عنوان

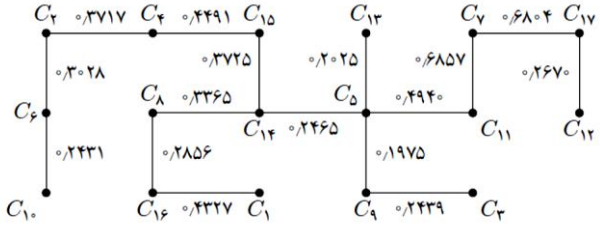
راس گراف در نظر گرفته می‌شوند. بنابراین گرافی کامل شامل ۱۷ راس و ۱۳۶ یال ایجاد می‌شود که هر یال آن وزن دارد که همان فاصله‌ی اقلیدسی مرکز دو زیرخوشه‌ی راس‌های متناظر است. پس از تشکیل این گراف کامل، الگوریتم کراسکال بر روی این گراف اجرا می‌شود تا کوچک‌ترین درخت پوشا حاصل شود، که در شکل ۴ رسم شده است. در شکل ۴، زیرخوشه‌ها شماره‌گذاری شده‌اند و راس‌های درخت پوشا را تشکیل می‌دهند. همچنین فاصله‌ی اقلیدسی مراکز زیرخوشه‌های مجاور بر روی یال‌های درخت پوشا نوشته شده است.



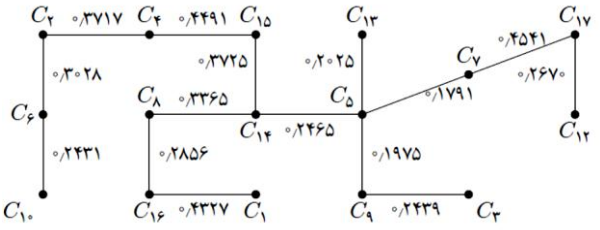
شکل ۴: نمودار کوچک‌ترین درخت پوشا اولیه بر روی مراکز زیرخوشه‌ها. (منبع: محاسبات محقق)

در مرحله ادغام، برای هر جفت زیرخوشه‌ی مجاور طبق رابطه‌ی (۶) ابتدا چسبندگی و سپس از رابطه‌ی (۹) تشابه درونی هر زیرخوشه محاسبه شد. در انتها با داشتن چسبندگی و شباهت درونی هر جفت زیرخوشه مجاور، شاخص ادغام آن‌ها از رابطه‌ی (۱۱) به‌دست آمد. شاخص ادغام میان جفت‌های زیرخوشه مجاور بر روی درخت پوشای شکل ۵ نشان داده شده است.

در ادامه‌ی روش پیشنهادی، در هر گام جفت زیرخوشه‌ی مجاوری که بیش‌ترین مقدار شاخص ادغام را دارند در یک‌دیگر ادغام می‌شوند و یک زیرخوشه‌ی بزرگ‌تر را به‌وجود می‌آورند و این‌کار تا زمانی که تعداد خوشه‌ی مطلوب حاصل شود ادامه می‌یابد. با توجه به شکل ۵، بیش‌ترین شاخص ادغام متعلق به جفت زیرخوشه مجاور C_7 و C_{11} با مقدار ۰٫۶۸۵۷ است و بنابراین C_7 و C_{11} اولین دو زیرخوشه‌ای هستند که با یک‌دیگر ادغام می‌شوند و یال میان آن‌ها حذف می‌شود. نام‌گذاری زیرخوشه‌ی جدید براساس کوچک‌ترین شماره‌ی دو زیرخوشه‌ی ادغام شده انجام می‌شود و بنابراین نام زیرخوشه‌ی جدید C_7 گذاشته شده است که اکنون با دو زیرخوشه‌ی C_5 و C_{17} مجاور می‌شود. این فرآیند به‌اختصار، ادغام زیرخوشه‌ی C_{11} در زیرخوشه‌ی C_7 گفته می‌شود. با توجه به تغییرات به‌وجود آمده، بایستی شاخص ادغام بین زیرخوشه‌ی جدید و زیرخوشه‌های C_5 و C_{17} محاسبه شود. شکل ۶، نتیجه‌ی این ادغام و مقادیر جدید شاخص ادغام را نشان می‌دهد.



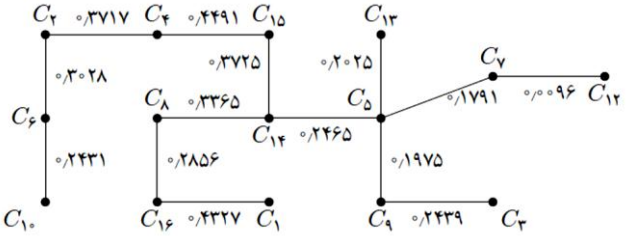
شکل ۵: گراف کوچک‌ترین درخت پوشا حاصل از مرحله اول همراه با وزن‌های مربوط به شاخص ادغام. (منبع: محاسبات محقق)



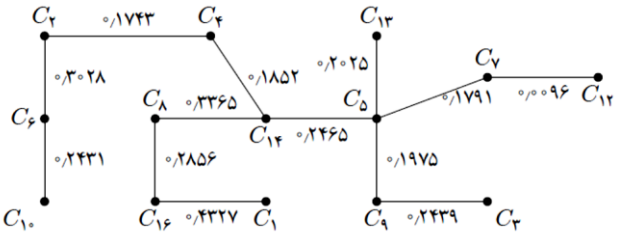
شکل ۶: نمودار درخت پوشا بعد از ادغام زیرخوشه‌ی C_{11} در زیرخوشه‌ی C_7. (منبع: محاسبات محقق)

به منظور خلاصه‌سازی، گام‌های طی شده تا رسیدن به تعداد مناسب زیرخوشه‌ها، در شکل ۷ از بالا به پایین نشان داده شده است. شکل ۷ (آ)، بعد از ادغام زیرخوشه‌ی C_{17} در C_{12} به‌دست می‌آید. شکل ۷ (ب)، بعد از ادغام زیرخوشه C_{15} در C_4؛ شکل ۴ (ج) بعد از ادغام زیرخوشه‌ی C_{16} در C_8؛ شکل ۷ (د) بعد از ادغام زیرخوشه‌ی C_{14} در C_8؛ شکل ۷ (ه) بعد از ادغام زیرخوشه‌ی C_6 در C_2؛ و در نهایت شکل ۷ (و) بعد از ادغام زیرخوشه‌ی C_9 در C_5 حاصل می‌شوند. این فرآیند تا رسیدن به تعداد مناسب خوشه ادامه می‌یابد.

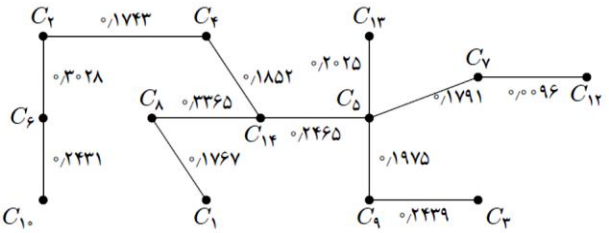
در این پژوهش، ملاک زیر برای شناسایی تعداد مناسب خوشه‌ها لحاظ شده است. بعد از هر بار ادغام، ابتدا میانگین فاصله اعضای هر زیرخوشه‌ی موجود تا مرکز متناظر محاسبه می‌شود، و پس از آن کم‌ترین، بیش‌ترین و مقدار متوسط میانگین‌های مذکور از زیرخوشه‌ها به‌دست می‌آید.



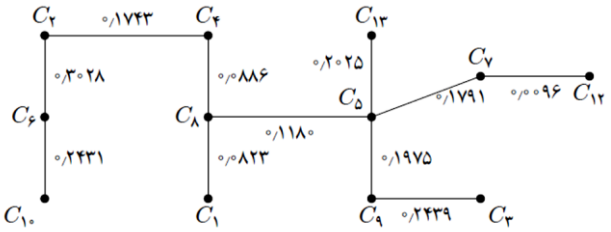
(آ) درخت پوشا بعد از ادغام زیرخوشه‌ی C_{17} در زیرخوشه‌ی C_{12}.



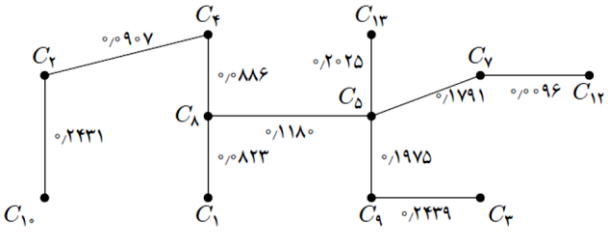
(ب) درخت پوشا بعد از ادغام زیرخوشه‌ی C_{15} در زیرخوشه‌ی C_4.



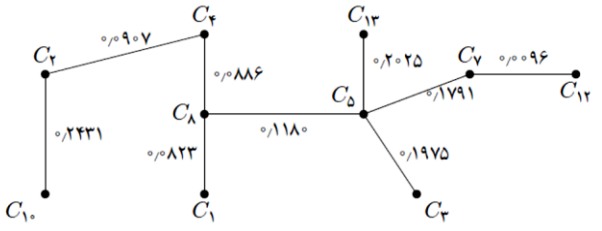
(ج) درخت پوشا بعد از ادغام زیرخوشه‌ی C_{16} در زیرخوشه‌ی C_8.



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



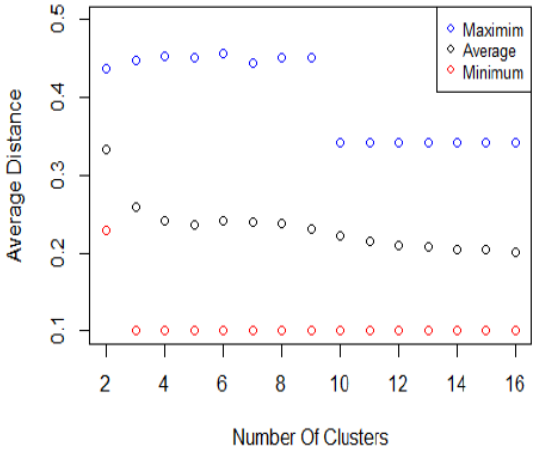
(ه) درخت پوشا بعد از ادغام زیرخوشه‌ی C_6 در زیرخوشه‌ی C_2.



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

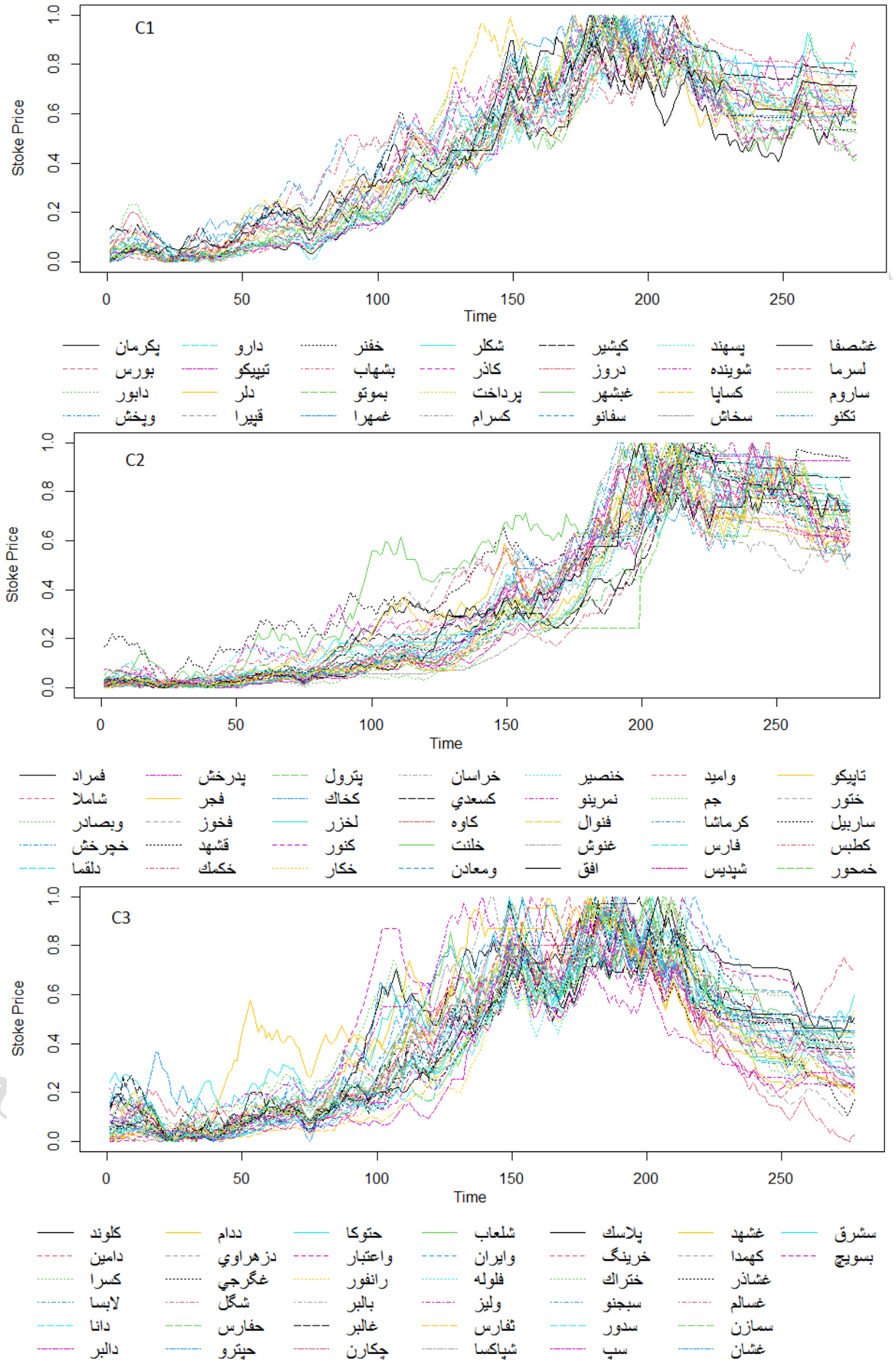
شکل ۷: درخت پوشای اصلاح شده بعد از ادغام هر جفت از زیرخوشه‌ها. (منبع: محاسبات محقق)

در شکل ۸، این مقادیر در هر بار ادغام تا رسیدن به دو زیرخوشه نشان داده شده است. در شکل ۸ مشاهده شد که میانگین فاصله اعضا با کم شدن تعداد زیرخوشه‌ها، بیش‌تر می‌شود و در واقع با ادغام شدن زیرخوشه‌ها، اعضای زیرخوشه‌ی جدید فاصله‌ی بیشتری از یکدیگر می‌گیرند. شکل ۸ نشان می‌دهد که تعداد مناسب برای توقف مرحله‌ی ادغام، ۱۰ زیرخوشه است زیرا با کاهش تعداد زیرخوشه‌ها به ۹ تفاوت قابل توجهی مشاهده می‌شود. بنابراین تعداد خوشه‌ی مطلوب در این مورد ۱۰ خوشه است و شکل ۷ (و) خوشه‌های نهایی را نشان می‌دهد. اکنون می‌توان شماره‌گذاری خوشه‌های نهایی را به C_1، C_2، ... و C_{10} تغییر داد. برای این منظور، نام خوشه‌ی C_{12} در شکل ۴ (و) به C_6 و نام خوشه‌ی C_{13} در شکل ۷ (و) به C_9 تغییر یافت.

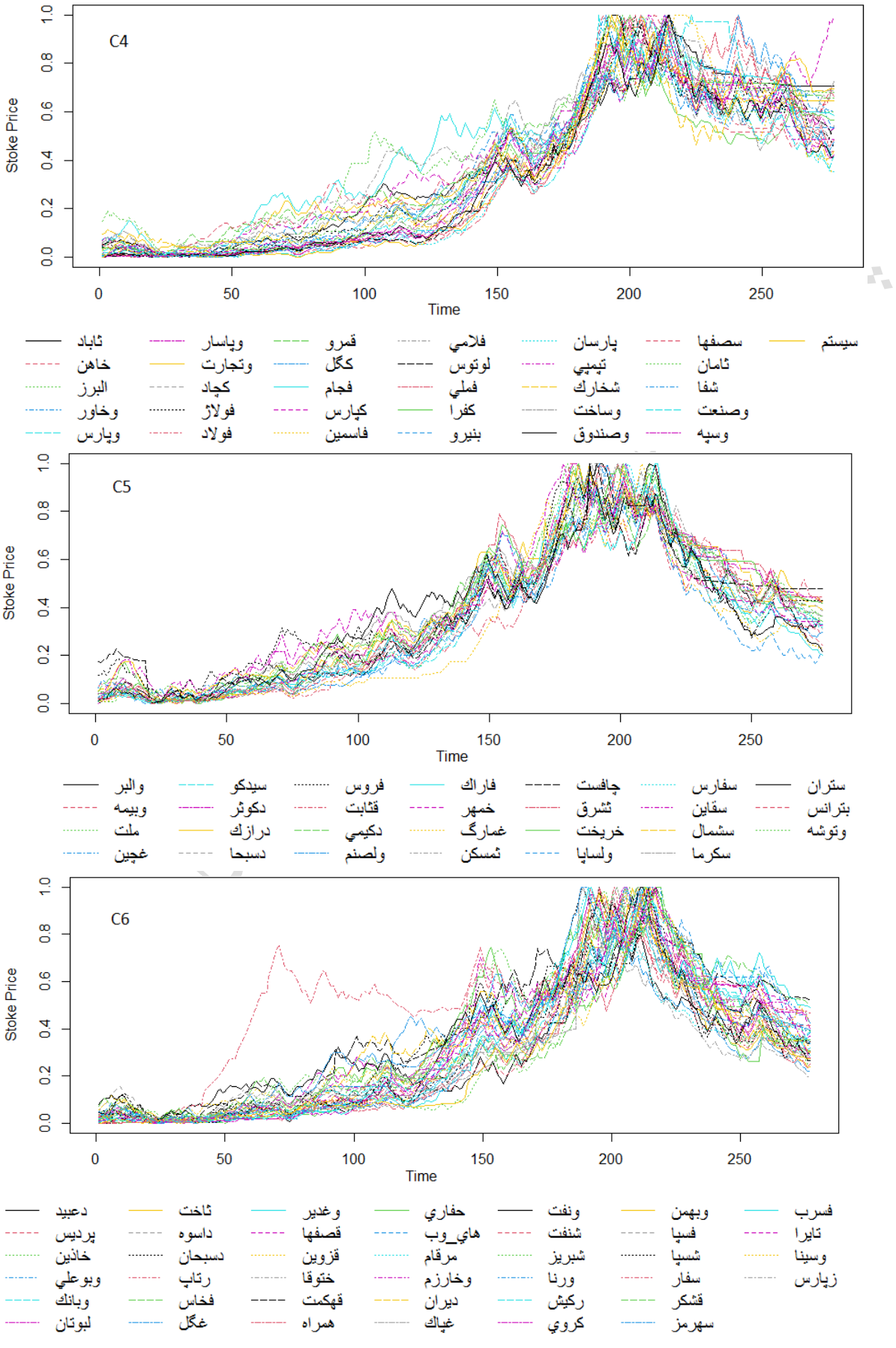


شکل ۸: روند فاصله‌ی اعضای خوشه‌ها بر اساس تعداد خوشه‌های تشکیل شده. (منبع: محاسبات محقق)

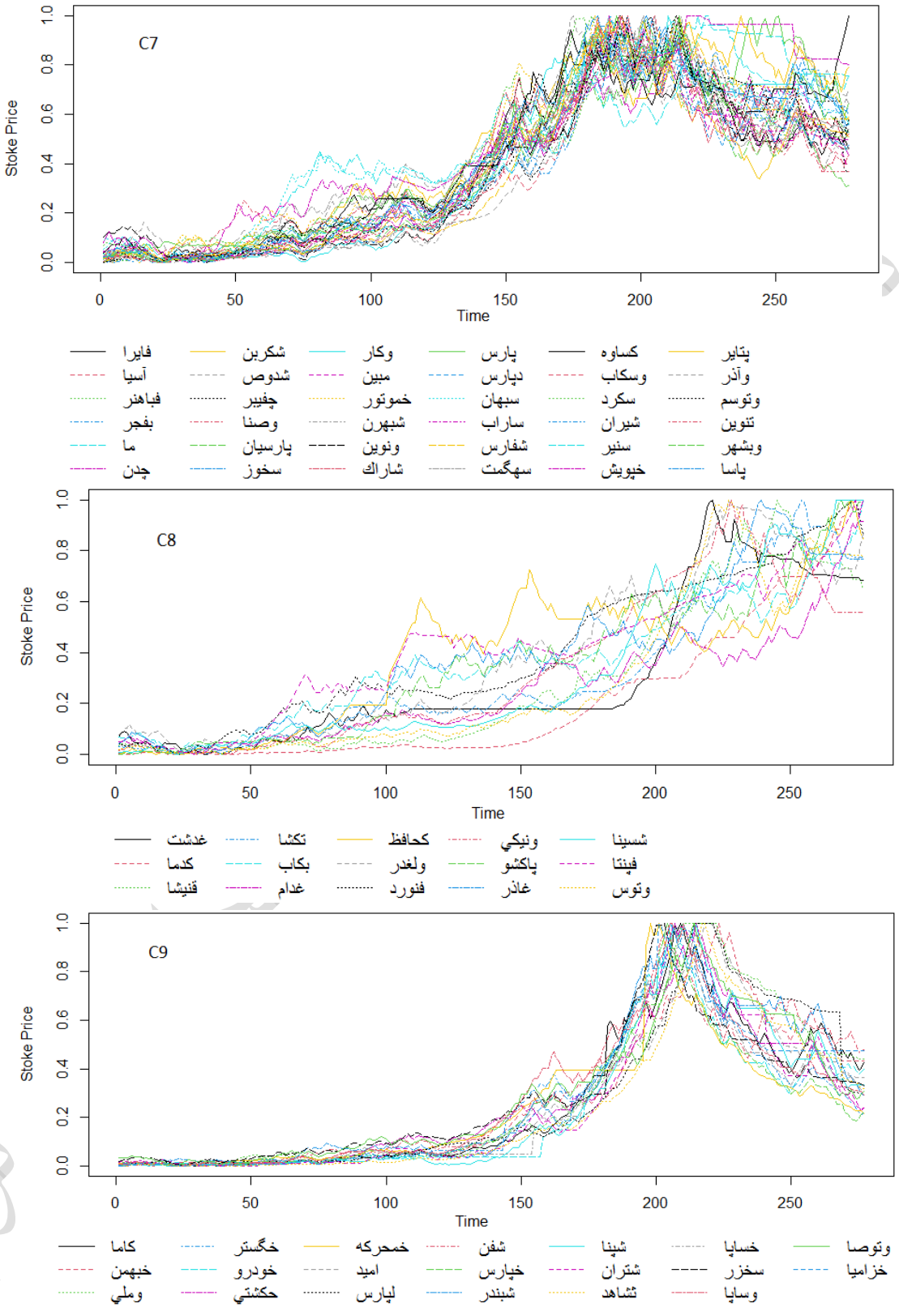
در شکل ۹ سری زمانی نمادهای هر خوشه نشان داده شده است. در این شکل مشاهده می‌شود که خوشه‌های C_5، C_7، C_9 و C_{10} خوشه‌های خوبی هستند و سری‌های زمانی آنها شباهت زیادی به یکدیگر دارند. به‌عبارت دیگر داده‌های آنها در اطراف سری زمانی میانگین خوشه‌ها و نزدیک به آن نوسان یا هم‌حرکتی دارند. خوشه‌های C_1 تا 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 خوشه‌ی C_8 نسبت به بقیه‌ی خوشه‌ها پراکندگی داده‌های بیشتری دارد.



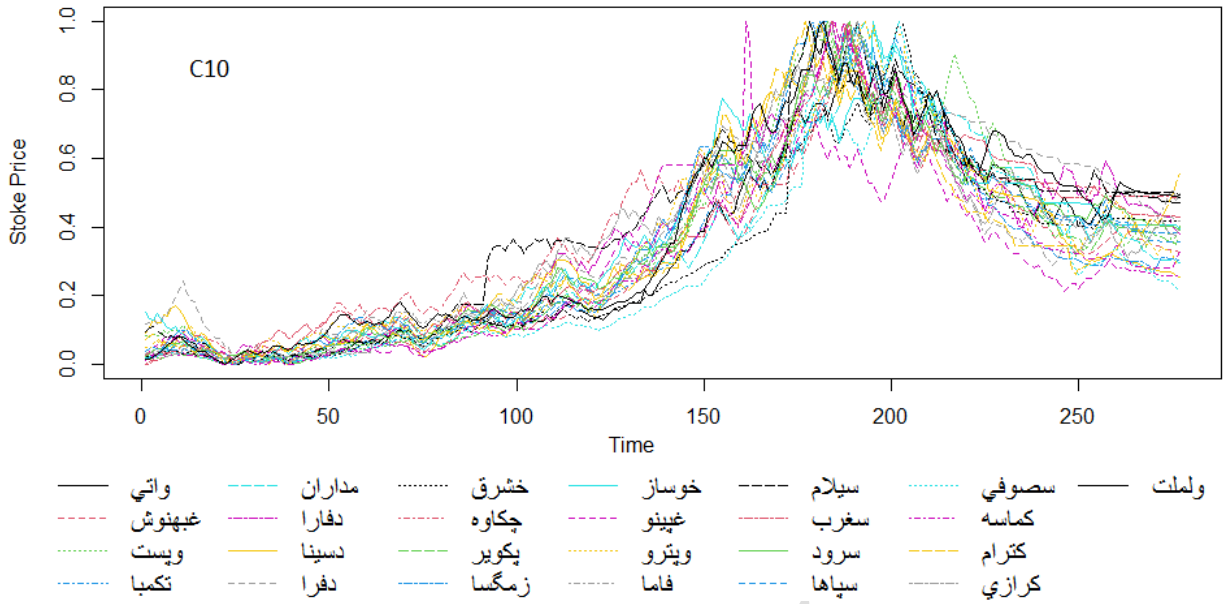
شکل ۹: نمودار سری زمانی خوشه‌های نهایی تشکیل شده بعد از اجرای روش پیشنهادی. (منبع: محاسبات محقق)



شکل ۹ (ادامه): نمودار سری زمانی خوشه‌های نهایی تشکیل شده بعد از اجرای روش پیشنهادی. (منبع: محاسبات محقق)



شکل ۹ (ادامه): نمودار سری زمانی خوشه‌های نهایی تشکیل شده بعد از اجرای روش پیشنهادی. (منبع: محاسبات محقق)



شکل ۹ (ادامه): نمودار سری زمانی خوشه‌های نهایی تشکیل شده بعد از اجرای روش پیشنهادی. (منبع: محاسبات محقق)

بررسی نتایج حا

فعال زیادی در زمینه غذا، ماشین آلات و برق، دارو و سرمایه‌گذاری، بانک و بیمه در خوشه‌های C_3 و C_6مشاهده 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 خوشه‌ی C_2 تعداد زیادی نماد خودرویی و فلزات اساسی وجود دارد که نشان از شباهت سری‌های زمانی این صنایع با صنایع شیمیایی دارد.

به‌منظور اعتباربخشی به نتایج حاصل از روش پیشنهادی، در ادامه با نتایج خوشه‌بندی به روش K-میانگین که بر مبنای فاصله‌ی اقلیدسی است مقایسه‌ای ارایه شده است. فرض کنید

$$P = \{P_1, P_2, \cdots, P_k\}, \quad (۱۲)$$

مجموعه‌ی خوشه‌بندی روش پیشنهادی و

$$C = \{C_1, C_2, \cdots, C_l\}, \quad (۱۳)$$

مجموعه‌ی خوشه‌بندی حاصل از الگوریتم K-میانگین باشد. در این‌صورت معیار زیر در مرجع [۳] مطرح شده است که در ادامه از این معیار برای ارزیابی و مقایسه‌ی عملکرد روش پیشنهادی استفاده می‌شود:

$$RI = \frac{P_{00} + P_{11}}{P_{00} + P_{01} + P_{10} + P_{11}}, \quad (۱۴)$$

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 به‌ازای تعداد خوشه‌ی مختلف در جدول ۱ گزارش شده است. همان‌گونه که مشاهده می‌شود، به‌عنوان نمونه اگر تعداد خوشه‌ی مطلوب در روش پیشنهادی و روش K-میانگین، ۱۰ در نظر گرفته شود معیار RI بیش از ۸۸ درصد می‌شود. بنابراین به‌ازای ۱۰ خوشه، روش پیشنهادی و روش RI دارای تشابه عملکردی بیش از ۸۸ درصد خواهند بود.

جدول ۱: مقایسه روش پیشنهادی با روش K-میانگین به‌ازای تعداد خوشه‌ی گوناگون. (منبع: محاسبات محقق)

تعداد خوشه	P_{11}	P_{00}	P_{10}	P_{01}	معیار RI
۱۱	۱۵۱۶۴	۱۲۷۲۳	۸۵۲۰	۶۷۵۸	۰٫۶۴۳۱
۱۰	۱۷۲۷۳	۱۲۱۸۲	۸۷۲۲	۵۱۸۸	۰٫۸۸۰۴

۴٫۱ ارزیابی الگوریتم پیشنهادی

در این بخش، روش پیشنهادی بر روی داده‌هایی که خوشه‌های آن از قبل مشخص هستند اعمال می‌شود. با توجه به‌این‌که در این مجموعه‌های داده‌ای، خوشه‌های مطلوب تعیین شده هستند می‌توان

از معیار RI که در رابطه‌ی (۱۴) ارایه شده است برای ارزیابی عملکرد روش خوشه‌بندی پیشنهادی بهره برد. به‌این منظور، در (۱۲)، P خوشه‌های مطلوب و در (۱۳)، C خوشه‌های به‌دست آمده توسط روش پیشنهادی است.

مجموعه داده‌های مورد استفاده در این بخش از برخی مجموعه داده‌های بررسی شده توسط میشرا و موهانتی [۳] اقتباس شده است. اطلاعات مربوط به این مجموعه‌های داده‌ای به‌همرا نتایج حاصل از اعمال روش پیشنهادی، در جدول ۲ گزارش شده‌اند. همچنین به‌منظور مقایسه‌ی بهتر، در جدول ۲، نتایج حاصل از روش فاصله‌ی چگالی گاوسی که در [۳] گزارش شده است و همچنین روش سه‌مرحله‌ای میشرا و موهانتی [۳] را مشاهده می‌کنید. همان‌گونه که در جدول ۲ مشاهده می‌شود، در مورد سه مجموعه‌ی داده‌ای Iris، Wine و Glass، روش پیشنهادی در مقایسه با روش [۳] و روش فاصله‌ی چگالی گوسی عملکرد بهتری داشته است و در مورد مجموعه‌ی داده‌ای Tumor، روش پیشنهادی با اختلاف کمی نسبت به روش [۳] و بهتر از روش فاصله‌ی چگالی گوسی عمل کرده است.

جدول ۲: مقایسه عملکرد روش پیشنهادی و روش [۳] بر روی مجموعه‌های داده‌ای مختلف بر مبنای معیار RI. (منبع: محاسبات محقق و [۳])

نام مجموعه داده‌ای	تعداد خوشه مطلوب	تعداد نقاط داده‌ای	ابعاد داده‌ها	GDD	روش [۳]	روش پیشنهادی
Iris	۳	۱۵۰	۴	۰٫۷۷۱۹	۰٫۸۲۴۷	۰٫۸۵۱۴
Wine	۳	۱۷۸	۱۳	۰٫۶۷۷۶	۰٫۷۴۹۵	۰٫۷۸۴۷
Glass	۷	۲۱۴	۹	۰٫۲۵۹۸	۰٫۷۴۹۲	۰٫۸۳۰۱
Tumor	۲۲	۳۳۹	۱۷	۰٫۸۹۱۶	۰٫۹۱۲۹	۰٫۸۹۷۳

۵. قدردانی

نویسندگان از داوران محترم که نقش به‌سزایی در بهبود و ارتقای مقاله حاضر داشتند تشکر و قدردانی می‌نمایند.

۶. نتیجه‌گیری

در این پژوهش از الگوریتم سه مرحله‌ای مبتنی بر کوچک‌ترین درخت پوشا برای خوشه‌بندی سری‌های زمانی قیمت سهام شرکت‌های پذیرش شده در بورس اوراق بهادار تهران استفاده شد. در روش پیشنهادی، ابتدا بر مبنای یک معیار پراکندگی مجموعه‌ی داده‌ای مذکور به تعداد زیادی زیرخوشه تقسیم شد. سپس بر اساس مرکز هر خوشه و فاصله‌ی اقلیدسی بین آن‌ها به‌عنوان وزن، گراف کامل وزن‌داری ساخته شد و بر روی آن کوچک‌ترین درخت پوشا به‌کمک الگوریتم کراسکال به‌دست آمد. در انتها، بر مبنای معیار ادغام معرفی شده، زیرخوشه‌ها با یک‌دیگر ادغام شدند تا زمانی که تعداد مناسبی از خوشه‌ها به‌دست 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 نمایندگانی از هر خوشه، رفتار قیمتی سایر اعضای آن خوشه را تقریب زد. همچنین می‌توان با انتخاب سهام از خوشه‌های متفاوت، سبدی 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.

مراجع

[1] M. Saeidi Kousha, and S. Mohebbi, "Optimizing stock portfolios by comparing different technical patterns." Financial Engineering and Portfolio Management, Vol. 12(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", *Journal of Financial accounting research,* vol. 4, no. 3 (13), 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 Systems with Applications,* 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", Springer, New York, 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", *Journal of Financial Accounting Research*, 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", *Journal of Investment Knowledge*, vol. 9, no. 35, pp. 1-19, 2020. [In Persian]

[7] M. Iqbalnia, A. Pouyanfar, and M. Maleki, "Modeling the co-movement of stocks in Tehran Stock Exchange using a three-phase clustering approach", *Journal of Financial Management Perspective*, vol. 5, no. 11, pp. 133-158, 2016.

[8] B. M. Blau and T. G. Griffith, "Price clustering and the stability of stock prices", *Journal of Business Research*, vol. 69(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 Systems with Applications*, 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, pp. 1148, 2020. doi: 10.3390/sym12071148

[11] E. Güngör and A. Özmen, "Distance and density based clustering algorithm using gaussian kernel." *Expert Systems with Applications*, vol. 69, pp. 10–20, 2017. doi: 10.1016/j.eswa.2016.10.022

[12] B. B. Nair, P. K. Saravana Kumar, N. R. Sakthivel, and U. Vipin, "Clustering stock price time series data to generate stock trading recommendations: An empirical study", *Expert Systems with Applications*, 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(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 Transactions on Knowledge and Data Engineering*, 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(8), pp. 68–75, 1999. doi: 10.1109/2.781637

[16] X. Wang, X. Wang, and D. Mitchell Wilkes, "A divide-and-conquer approach for minimum spanning tree-based clustering", *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 7, pp. 945-958, 2009. doi: 10.1109/TKDE.2009.37

[17] C. Zhong, D. Miao, and P. Fränti, "Minimum spanning tree based split-and-merge: A hierarchical clustering method", *Information Sciences*, vol. 181, no. 16, pp. 3397–3410, 2011. doi: 10.1016/j.ins.2011.04.013

[18] F. Zhao, Y. Gao, X. Li, Z. An, S. Ge, and C. Zhang, "A similarity measurement for time series and its application to the stock market", *Expert Systems with Applications*, vol. 182, 115217, 2021. doi: 10.1016/j.eswa.2021.115217

[19] R. Balakrishnan and K. Ranganathan, *A Textbook of Graph Theory*, Springer, New York, 2012.

[20] P. Fänti and S. Sieranoja, "K-means properties on six clustering benchmark datasets", *Applied Intelligence*, vol. 48, no. 12, pp. 4743-4759, 2018. https://cs.uef.fi/sipu/datasets/

[21] A. K. Das and J. Sil, "Cluster Validation Using Splitting and Merging Technique", *International Conference on Computational Intelligence and Multimedia Applications* (ICCIMA 2007), Sivakasi, India, 2007, pp. 56-60.

پیش نویس مقاله در مجله محاسبات نرم