Mathematical Researches
Print ISSN: 2588-2546
Online ISSN: 2588-2554
Kharazmi University
Homepage: https://mmr.khu.ac.ir/

Improving clustering by detecting and removeing outliers based on Euclidean and exponential distance approach

Atiyeh Ehsani[1], **Nafiseh Alemohammad**[2]✉

1. Department of Mathematic, College of Basic Sciences, Shahed University, Iran.
   E-mail: atiyeh.ehsani@shahed.ac.ir
2. Department of Mathematic, College of Basic Sciences, Shahed University, Iran.
   ✉E-mail: n.alemohammad@shahed.ac.ir

Article Info

Article type:
Research Article

Article history:
Received:
9 November 2019
Received in revised form:
30 January 2021
Accepted:
2 May 2021
Published online:
20 June 2023

Keywords:
Clustering,
GARCH Model,
Outlier time series,
Exponential
Distance Approach.

ABSTRACT

Introduction

Clustering is a powerful tool for depicting and finding out the structure of data, and it is also an unsupervised learning method that divides the data set into clusters where the elements in each cluster have the maximum similarity and the elements of the two clusters have the minimum similarity. Clustering and outlier detection are interdependent processes. The main goal in this article is to detect outliers to improve clustering.

Material and methods

In this paper, a new method for identifying outlier time series based on GARCH model by exponential distance approach is presented in three steps: first fuzzy and hard clustering methods are implemented on time series, then the outlier time series are detected and removed from the dataset. After removeing outlier time series, clustering algorithms are applied for dataset again.

Results and discussion

The 30 stocks of the top active, lucrative and profitable stocks in the Iranian stock market are used to evaluate the presented methods. By computing the two Silhouette and Xe-Beni indexes, the accuracy of the

clustering methods are compared. The numerical results show that the accuracy of presented method is improved.

Conclusion

The following conclusions were drawn from this research.

- The paper Introduced a clustering method for financial data based on unconditional volatility and time-varying volatility.
- Both fuzzy and hard clustering methods using Euclidean and exponential distance have been implemented on the studied data.
- The clustering results indicated that after removing the outlier time series, the values of the silhouette and Xe-Beni indices have improved.

How to cite: Ehsani, A., Alemohammad, N. (2023). Improving clustering by detecting and removeing outliers based on Euclidean and exponential distance approach. *Mathematical Researches*, 9 (1), 1-29.

Publisher: Kharazmi University

Mathematical Researches
Print ISSN: 2588-2546
Online ISSN: 2588-2554
Homepage: https://mmr.khu.ac.ir/

بهبود خوشه‌بندی از طریق تشخیص و حذف سری‌های زمانی پرت بر اساس رویکرد فاصله‌ای اقلیدسی و نمایی

عطیه احسانی[۱]، سیده نفیسه آل‌محمد[۲]✉

۱. گروه ریاضی، دانشکدهٔ علوم پایه، دانشگاه شاهد، تهران، ایران. رایانامه: atiyeh.ehsani@shahed.ac.ir
۲. نویسندهٔ مسئول، گروه ریاضی، دانشکدهٔ علوم پایه، دانشگاه شاهد، تهران، ایران. رایانامه: n.alemohammad@shahed.ac.ir

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۳۹۸/۰۸/۱۸
تاریخ بازنگری: ۱۳۹۹/۱۱/۱۱
تاریخ پذیرش: ۱۴۰۰/۰۲/۱۲
تاریخ انتشار: ۱۴۰۲/۰۳/۳۰

واژه‌های کلیدی:
خوشه‌بندی،
مدل گارچ،
سری‌های زمانی پرت،
رویکرد فاصله‌ای نمایی.

چکیده

در این مقاله یک روش جدید جهت شناسایی سری زمانی پرت بر اساس مدل گارچ بر اساس رویکرد فاصله‌ای نمایی ارائه می‌شود که در سه مرحله انجام می‌گیرد: ابتدا روش‌های خوشه‌بندی فازی و غیر فازی بر سری‌های زمانی پیاده سازی شده، سپس سری‌های زمانی پرت تشخیص داده و از مجموعه داده‌ها حذف می‌شود. پس از حذف سری‌های زمانی پرت دوباره روش‌های خوشه بندی بر مجموعه داده‌ها اعمال خواهد شد. برای ارزیابی روش‌های ارائه شده از سهام ۳۰ شرکت برتر، فعال و پر سود موجود در بازار بورس ایران استفاده می‌شود. با محاسبه دو شاخص سیلهوت و زایبنی دقت روش‌های خوشه‌بندی به کار گرفته شده با هم مقایسه می‌شوند و در پایان نشان داده می‌شود با حذف داده پرت روش خوشه‌بندی بر اساس مدل گارچ بر اساس رویکرد فاصله‌ای نمایی عملکرد بهتری دارد.

استناد: احسانی، عطیه؛ آل‌محمد، نفیسه؛ (۱۴۰۲). بهبود خوشه‌بندی از طریق تشخیص و حذف سری‌های زمانی پرت بر اساس رویکرد فاصله‌ای اقلیدسی و نمایی. *پژوهش‌های ریاضی*، ۹ (۱)، ۲۹-۱.

ناشر: دانشگاه خوارزمی

۱. مقدمه

خوشه‌بندی وسیله‌ای قدرتمند برای به تصویر کشیدن و معلوم کردن ساختار داده‌ها است و همچنین روش یادگیری بدون نظارتی است که مجموعه داده‌ها را به خوشه‌هایی تقسیم‌بندی می‌کند که عناصر در هر خوشه با هم حداکثر شباهت و عناصر دو خوشه با هم حداقل شباهت را داشته باشند. خوشه‌بندی به دو نوع تفکیک می‌شود: خوشه‌بندی سخت و خوشه‌بندی فازی. اگر در خوشه‌بندی، هر عنصر فقط به یک خوشه تعلق داشته باشد آن خوشه‌بندی را خوشه‌بندی سخت می‌گویند، در غیر این‌صورت اگر هر عنصر به بیش از یک خوشه متعلق باشند آن خوشه‌بندی، خوشه‌بندی فازی نام دارد. خوشه‌بندی و تشخیص داده پرت فرایندهای وابسته به هم هستند [۱,۹]. تشخیص داده پرت جنبه‌ی مهمی در تحلیل داده‌ها دارد زیرا با حذف داده پرت دقت خوشه‌بندی بهبود می‌یابد. مشهورترین تعریف داده پرت[1] که توسط هاوکینز ارائه شده است، به شرح زیر است:

«سری زمانی پرت به مشاهده‌ای گفته می شود که با دیگر مشاهدات موجود در یک مجموعه فاصله‌ی زیادی داشته باشد به طوری‌که این فرض به وجود می‌آید که آن مشاهده توسط مکانیزم دیگری ساخته شده است [۲].»

روش‌های خوشه‌بندی سری‌های زمانی می‌تواند به ۳ کلاس طبقه بندی شود:

۱- خوشه‌بندی بر اساس مشاهدات: این رویکرد شامل روش‌های خوشه‌بندی بر اساس سری‌های زمانی واقعی است که به طور ویژه برای سری‌های زمانی کوتاه مدت مفید هستند. علاوه بر این سایر روش‌های خوشه‌بندی بر اساس مشاهدات با استفاده از روش اندازه‌گیری DTW که یک روش اندازه‌گیری شناخته شده و وابسته به زمان است، به دست می‌آیند [۷].

۲- خوشه‌بندی بر اساس ویژگی: این رویکرد بر اساس ویژگی‌های گرفته شده از سری‌های زمانی هست. کایادو و همکارانش در سال (۲۰۰۷) بیان کردند که "اگر یک سری زمانی شامل تعداد زیادی مشاهدات باشد ، خوشه‌بندی بر اساس این سری‌های به دلیل وجود نویز[2] گزینه مناسبی نیست. " با تحقیقات انجام شده به چند روش خوشه‌بندی بر اساس ویژگی برای سری‌های زمانی دارای نویز در زیر به طور مختصر اشاره می‌شود:

۱. روش‌های بر اساس ویژگی‌های دامنه زمانی.

۲. روش‌های بر اساس ویژگی‌های دامنه فرکانسی.

۳.روش‌های بر اساس ویژگی موجک‌ها [۳,۷].

۳- خوشه‌بندی بر اساس مدل: این رویکرد بر اساس ویژگی‌های مدل‌هایی که برای سری‌های زمانی در نظر گرفته می‌شود، است. منظور از ویژگی‌های مدل، میانگین و واریانس هست. در این رویکرد فرض شده است که مجموعه سری‌های زمانی که از یک مدل تولید شده‌اند، الگوهای مشابهی دارند.

[1] Outlier
[2] Noise

اگر ویژگی‌های مدل بر اساس میانگین پارامترهای تخمین زده شده یا به وسیله‌ی میانگین باقی‌مانده‌ی مدل‌های اختصاص داده شده به سری‌های زمانی به دست آید، از مدل آریما[3] استفاده می‌شود ولی اگر واریانس شرطی مورد توجه باشد از مدل گارچ[4] استفاده می‌شود. خوشه‌بندی برای سری‌های زمانی مالی، با توجه به تغییر نوسانات آن‌ها در طول زمان احتیاج به تعریف اندازه‌گیری مناسب دارد [۷].

هوتاماکی و همکارانش در سال (۲۰۰۵) الگوریتمی را معرفی کردند که هم‌زمان با تشخیص سری‌های زمانی پرت، خوشه‌بندی نیز انجام می‌دهد و بدین‌ترتیب مراکز خوشه دقیق‌تری به دست می‌آید. این روش شامل دو مرحله است: مرحله اول خوشه بندی

K- میانگین[5] انجام می‌گیرد و مرحله دوم مشاهداتی که از مراکز خوشه مربوطه دور هستند، باتوجه به فاکتور پرت‌افتادگی و مقدار آستانه، حذف می‌شوند. آن‌ها همچنین با انجام آزمایش بر سه مجموعه داده مصنوعی نشان دادند که روش پیشنهادی آنها خطای کمتری نسبت به سایر روش ها دارد [۱۰].

کایادو و کرتو در سال (۲۰۰۷) روش هایی بر اساس نوسانات برای تجزیه و تحلیل سری‌های زمانی مالی معرفی کردند. آنها با استفاده از مدل گارچ فاصلهٔ بین نوسانات بازده سهام را تخمین زدند و مشکل تفاوت طول در سری‌های زمانی را حل کردند [۳].

پارابجات و همکارانش در سال (۲۰۱۱) روش خوشه‌بندی قوی بر اساس چگالی ارائه دادند که سری‌های زمانی پرت را قبل از خوشه بندی بر اساس چگالی مجموعه داده‌ها در هر خوشه تشخیص می‌دهد. بنا به این روش نقطه‌ای سری‌های زمانی پرت به حساب می‌آید که در قسمت‌هایی با چگالی بیشتر، قرار نداشته باشد [۱۲].

دورسو و همکارانش در سال (۲۰۱۶) مدل های خوشه‌بندی فازی قوی بر اساس مدل پارامتری گارچ برای طبقه‌بندی نوسانات سری‌های زمانی با سه رویکرد ارایه دادند. این سه رویکرد (اصلاح و متریک و نویز) تأثیر منفی ناشی از سری‌های زمانی پرت را بر مرکز خوشه‌ها از بین می‌برند [۷].

جان و اوک-پو نج در سال (۲۰۱۷) با گسترش الگوریتم K- میانگین توانستند، هم‌زمان با خوشه‌بندی سری‌های زمانی پرت را با معرفی یک خوشه اضافه برای نگه‌داری داده‌های پرت شناسایی کنند و با طراحی الگوریتم تکراری تابع هدف روش ارائه شده را بهینه سازی کرده‌اند. در آخر با انجام آزمایش‌های عددی بر داده های مصنوعی و داده‌های واقعی کارایی الگوریتم پیشنهادی را نشان دادند [۸].

در این مقاله برخی از روش‌های خوشه‌بندی بر روی سری‌های زمانی مالی بر اساس مدل گارچ معرفی می‌شود، که قادر است عملکرد روش‌های خوشه‌بندی فازی و غیرفازی را در از بین بردن تأثیر منفی ناشی از سری‌های زمانی پرت، بهبود بخشد.

[3] ARIMA
[4] GARCH
[5] K-Means

مزیت استفاده از روش‌های خوشه‌بندی پیشنهادی این است که می‌توان به سرمایه‌گذار در بازار بورس در تشخیص سهام‌هایی با الگوهای مشابه کمک کرد، تا انتخاب مناسب‌تری برای سبد سهام خود داشته باشد. در این صورت ریسک معامله پایین آمده و احتمال از دست دادن سرمایه کمتر خواهد شد.

در این مقاله سری‌های زمانی مالی بر اساس نوسان غیر شرطی و نوسان زمان متغیر مدل گارچ نشان داده می‌شوند و همچنین روش‌های خوشه‌بندی K-مدوید و فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصله اقلیدسی و خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی معرفی می‌شوند. در فصل اول مفاهیم و تعاریف اولیه ارائه می‌شود. فصل دوم به تعریف مدل گارچ و اندازه‌گیری فاصله برای سری‌های زمانی می‌پردازد. در فصل سوم به تعریف خوشه‌بندی‌ها با استفاده از فواصل تعیین شده بر اساس مدل گارچ پرداخته می‌شود. فصل چهارم تشخیص سری‌های زمانی پرت با استفاده از میزان پرت افتادگی و شاخص‌های ارزیابی بیان می‌شود. در فصل پنجم نتایج به‌دست آمده از پیاده‌سازی روش‌های خوشه‌بندی مطرح شده در مقاله بر سهام ۳۰ شرکت برتر، فعال و پر سود موجود در بازار بورس ایران مورد تجزیه و تحلیل قرار داده می‌شوند. در فصل ششم جمع‌بندی و نتیجه‌گیری ارائه می‌شود.

۲- مدل گارچ و اندازه‌گیری فاصله برای سری‌های زمانی

فرض کنید y_t یک سری زمانی باشد که t شاخص زمانی هست و μ دارای مقدار ثابت است. ε_t یک فرایند تک متغیر با میانگین صفر و واریانس نا همگن است به طوری که:

$$y_t = \mu + \varepsilon_t$$
$$\varepsilon_t = u_t\sqrt{h_t}$$

u_t فرآیند نوفه سفید با میانگین صفر و واریانس ۱ است. h_t واریانس شرطی که از مدل $GARCH(p,q)$ پیروی می‌کند.

$$h_t = \gamma + \alpha_1\varepsilon_{t-1}^2 + ... + \alpha_p\varepsilon_{t-p}^2 + \beta_1 h_{t-1} + ... + \beta_q h_{t-q} \qquad (۱,۲)$$

شروط مانایی مدل به شرح زیر است:

$$\gamma > 0, \quad 0 \le \alpha_i, \beta_j \le 1, \quad \begin{cases} i = 1,2,...,p \\ j = 1,2,...,q \end{cases}, \quad \sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j < 1$$

$\eta_t = \varepsilon_t^2 - h_t$ فرض شود که و η_t به اطلاعات قبل بستگی ندارد و میانگین خطاهای صفر است، در این صورت با یک جبر ساده از معادله 2.1، ε_t^2 را می‌توان به صورت زیر نوشت:

$$\varepsilon_t^2 = \eta_t + h_t = \gamma + \alpha_1\varepsilon_{t-1}^2 + ... + \alpha_p\varepsilon_{t-p}^2 + \beta_1 h_{t-1} + ... + \beta_q h_{t-q} + \eta_t$$
$$\Rightarrow \varepsilon_t^2 = \gamma + \sum_{i=1}^{p^*}(\alpha_i + \beta_i)\varepsilon_{t-i}^2 - \sum_{j=1}^{q}\beta_j\eta_{t-j} + \eta_t \qquad (۲,۲)$$

که همان فرایند آرما[6] (p^*,q) است و $p^* = max(p,q)$ و اگر $i > p$ در این صورت $\alpha_i = 0$ و اگر $j > q$ در این صور ت $\beta_j = 0$ است. با توجه به شرایط معمول مانایی و برگشت پذیری بر روی ریشه‌های $(1-(\alpha_1+\beta_1)z-...-(\alpha_i+\beta_i)z^i-...-(\alpha_{p*}+\beta_{p*})z^{p*})$ و $(1-\beta_1 z-...\beta_j z_j-...-\beta_q z_q)$ ، ε_t^2 را در فرمول 2.2 می‌توان به صورت مدل اتورگرسیو مرتبه بی‌نهایت $AR(\infty)$ از طریق رابطه بازگشتی به صورت زیر بیان کرد [۱۲]

$$\varepsilon_t^2 = \frac{\gamma}{1-\sum_{j=1}^{q}\beta_j} + \sum_{k=1}^{\infty}\pi_k\varepsilon_{t-k}^2 + \eta_t \qquad (۲,۳)$$

و از طرفی اگر در مدل $ARMA(p,q)$، φ_i پارامتر مدل AR و θ_j پارامتر مدل MA باشد رابطه بازگشتی زیر به دست می‌آید [۶,۷]:

$$\pi_k - \sum_{j=1}^{q}\theta_j\pi_{k-j} = \phi_k, k = 0,1,... \qquad (۲.۴)$$

در رابطهٔ (۲.۴) دنباله از پارامترهای π_k تولید می‌شود. $\phi_i = 0, \phi_0 = 1$ اگر $i > p$ و $\pi_k = 0$ اگر $k < 0$ باشد. بر اساس فرمول ۲.۲و ۲.۴ پارامتر π_k به صورت زیر حاصل می‌شود [۶,۷]:

$$\pi_k = (\alpha_k + \beta_k) - \sum_{j=1}^{q}\beta_j\pi_{k-j}$$

با توجه به فرمول ۲,۳ نوسانات مورد انتظار در زمان $t+1$ ، مشروط به موجود بودن اطلاعات در زمان t ، می‌تواند به یک قسمت غیر شرطی و یک قسمت زمان متغیر تقسیم شود [۱۱]. نوسان غیر شرطی یک نوسان کلی است که اطلاعات اضافی در آن وجود ندارد. برای به‌دست آوردن نوسان غیرشرطی معادلهٔ ۲.۳، از این معادله امید ریاضی گرفته می‌شود:

$$E(\varepsilon_{t+1}^2) = \frac{\gamma}{(1-\sum_{j=1}^{q}\beta_j)} + \sum_{k=1}^{\infty}\pi_k E(\varepsilon_{t-k}^2) + E(\eta_t)$$

چون در شرایط مانایی $E(\varepsilon_{t+1}^2) = E(\varepsilon_{t-k}^2)$ و از طرفی $E(\eta_{t-j}) = E(\eta_t) = 0$ بنابراین نوسان غیرشرطی معادلهٔ ۲,۳ به صورت زیر نوشته می‌شود:

[6] ARMA

$$E(\varepsilon_{t+1}^2) = \frac{\gamma}{(1-\sum_{j=1}^{q}\beta_j)} + \sum_{k=1}^{\infty}\pi_k E(\varepsilon_{t+1}^2) \Rightarrow E(\varepsilon_{t+1}^2)(1-\sum_{k=1}^{\infty}\pi_k) = \frac{\gamma}{(1-\sum_{j=1}^{q}\beta_j)}$$

$$\Rightarrow uv = E(\varepsilon_{t+1}^2) = \frac{\gamma}{(1-\sum_{j=1}^{q}\beta_j)(1-\sum_{k=1}^{\infty}\pi_k)}$$

منظور از نوسان زمان متغیر، نوساناتی است که به تغیرات نوسان در بازه‌های زمانی مختلف وابسته است. نوسان زمان متغیر از مجموع وزن نامتناهی متغیر تصادفی مشاهده نشده به‌دست می‌آید و به نوسانات گذشته بستگی دارد. نوسان زمان متغیر از رابطهٔ زیر به دست می‌آید [۱۱]:

$$tvv = (\sum_{k=1}^{\infty}\pi_k^2)^{\frac{1}{2}}$$

هر سری زمانی دارای دو مؤلفه می‌باشد که منظور از دو مؤلفه یک سری زمانی نوسان غیر شرطی و زمان متغیر سری زمانی است. برای به دست آوردن فاصله بین دو سری زمانی از مجموع فواصل بین نوسانات غیر شرطی دو سری زمانی و نوسانات زمان متغیر دو سری زمانی که برای هر کدام از این فواصل وزن‌های w_1 و w_2 در نظر گرفته می‌شود، استفاده شده است [۷]. بنابراین فاصله‌ی iامین و i'امین سری زمانی به صورت زیر اندازه‌گیری می‌شود:

$$d_{ii'} = [w_1^2(uv_i - uv_{i'})^2 + w_2^2(tvv_i - tvv_{i'})^2]^{\frac{1}{2}} \qquad (۲.۵)$$

که در معادلهٔ ۲.۵ قسمت $(uv_i - uv_{i'})^2$ مربع فاصله‌ی اقلیدسی نوسانات غیرشرطی دو سری زمانی و قسمت $(tvv_i - tvv_{i'})^2$ مربع فاصله‌ی اقلیدسی نوسانات زمان متغیر دو سری زمانی و همچنین w_1 و w_2 وزن مؤلفه‌ها مناسب برای نوسانات غیر شرطی و زمان متغیر هستند. وزن مؤلفه‌ها می‌توانند به دو صورت به دست بیایند: با استفاده از سیستم وزن‌گیری داخلی به صورت ذهنی از قبل ثابت باشند یا از طریق سیستم وزن‌گیری خارجی با استفاده از روش خوشه‌بندی مناسب محاسبه شوند [۷]. شرایط زیر برای وزن مؤلفه‌ها در نظر گرفته می‌شود:

۱. در شرایط نرمال $w_1 + w_2 = 1$، منظور از شرایط نرمال این است که طبق این شرط مجموع احتمالات باید برابر یک شود، بنابراین می‌توان در نظر گرفت: $w_1 = 1 - w_2$.

۲. وزن مؤلفه‌ها همواره مثبت هستند $w_1, w_2 \geq 0$.

برای انتخاب وزن‌ها، شرایط ذهنی در نظر گرفته می‌شود. (در این مقاله، $w_1 = 0.1, w_2 = 0.9$ در نظر گرفته شده‌اند [۷].)

اندازه‌گیری فاصله به روش معادلهٔ ۲.۵ دارای ویژگی‌های زیر است:

۱- این فاصله متریک است. (یعنی ویژگی های غیر منفی بودن، تقارن و هویت را دارا هستند [۷].)

۲- از نظر محاسباتی آسان و از نظر تئوری قابل فهم است.

سری‌های زمانی که از نظر نوسانات غیر شرطی مشابه هستند می‌توانند، از نظر نوسانات زمان متغیر متفاوت باشند و بالعکس. اگر دو سری زمانی نوسان غیر شرطی و زمان متغیر یکسانی داشته باشند نمی‌توان نتیجه گرفت که از یک فرآیند گارچ هستند. در واقع تساوی گارچ دو سری زمانی در صورتی ممکن است که در فرآیند تولید پارامتر ثابت γ و پارامترهای α و β برای دو مدل گارچ یکسان باشد، بنابراین در صورت تساوی پارامترها می‌توان تساوی گارچ دو سری‌زمانی را نتیجه گرفت. همچنین طول سری‌های زمانی میتوانند با هم متفاوت باشند زیرا نتایج و تحلیل‌های خوشه‌بندی تحت تأثیر مؤلفه‌های به دست آمده از مدل گارچ سری‌های زمانی هست و طول‌ها بر این نتایج تأثیرگذار نیستند.

حال با استفاده از توضیحاتی که در مورد فاصله بین دو سری زمانی داده شد، دو به دو فاصله سری‌های زمانی به دست آورده می‌شود و روش‌های خوشه‌بندی بر اساس این فواصل تعریف شده بر سری‌های زمانی پیاده سازی می‌شوند.

۳- روش‌های خوشه‌بندی با استفاده از فواصل تعیین شده بر اساس مدل گارچ

در این فصل ابتدا روش‌های خوشه‌بندی فازی و غیر فازی و رویکرد نمایی معرفی شده و سپس این روش‌ها بر اساس مدل گارچ با استفاده از فاصله نمایی و اقلیدسی تعریف می‌شوند. هدف از معرفی روش‌های خوشه‌بندی فازی و غیر فازی بر اساس مدل گارچ با استفاده از فاصله نمایی نشان دادن عملکرد بهتر این روش‌ها در برابر داده پرت و دقت بالاتر آن‌ها زمانی که داده پرت از بین داده‌ها حذف می‌شود، نسبت به روش‌های خوشه‌بندی با فاصله اقلیدسی است.

۳,۱- روش خوشه‌بندی K- مدوید: در روش خوشه‌بندی K- مدوید به جای در نظر گرفتن مقدار میانگین داده‌های موجود در هر خوشه، هر داده می‌تواند به عنوان نماینده یک خوشه در نظر گرفته شود و داده‌هایی که به این نماینده شباهت دارند، به عبارتی فاصله کمتری دارند، در یک خوشه قرار می‌گیرند. هدف در این خوشه‌بندی کم کردن فاصله بین هر داده و نماینده خوشه‌ای که به آن تعلق دارد، می‌باشد. تابع هدف روش خوشه‌بندی K- مدوید توسط فرمول زیر به دست می‌آید:

$$minf_{K-medoids} = \sum_{i=1}^{N}\sum_{c=1}^{C}\|x_i - v_c\|^2$$

الگوریتم روش خوشه‌بندی K- مدوید به صورت زیر انجام می‌گیرد:

۱- ابتدا از مجموعه داده‌ها k داده را به عنوان نماینده اولیه خوشه‌ها در نظر گرفته می‌شود. سپس با استفاده از فاصله اقلیدسی، $\|x_i - v_c\|$، داده‌هایی که به این نماینده‌ها نزدیک‌ترند در یک خوشه قرار داده می‌شوند.

۲- حال از بین داده‌های موجود در هر خوشه داده‌ای به عنوان نماینده‌ی جدید انتخاب می‌شود.

۳- سپس با استفاده از تابع هدف روش خوشه‌بندی K- مدوید مقدار تابع هدف نماینده جدید به دست می‌آید. اگر مقدار تابع هدف نماینده جدید از تابع هدف نماینده قبلی کوچکتر باشد نماینده جدید به عنوان نماینده خوشه در نظر گرفته می‌شود و به مرحلهٔ ۲ برمی‌گردد، در غیر این صورت خوشه‌بندی پایان می‌یابد [۱,۹].

۳,۲- روش خوشه‌بندی فازی C- مدوید: خوشه‌بندی فازی C- مدوید توسط کاریشناپوران و همکارانش در سال ۱۹۹۳ ارائه شده است. تابع هدف مدل خوشه‌بندی فازی C- مدوید به صورت زیر می‌باشد:

$$minf_{fc-medoids} = \sum_{i=1}^{N}\sum_{c=1}^{C} u_{ic}^{m} \left\| x_i - v_c \right\|^2$$

الگوریتم این روش خوشه‌بندی به صورت زیر می‌باشد:

۱- مقدار ضریب فازیm ، تعداد خوشه‌ها، بیشترین تعداد تکرار، میزان عضویت اولیه‌ی هر داده به هر خوشه، C داده به عنوان نمایندهٔ اولیه خوشه‌ها انتخاب می‌شوند.

۲- حال از بین داده‌های موجود در هر خوشه داده‌ای به عنوان نماینده‌ی جدید انتخاب می‌شود و سپس مقدار درجه عضویت از طریق فرمول $u_{ic} = \frac{1}{\sum_{c'=1}^{C} [\frac{\left\| x_i - v_c \right\|^2}{\left\| x_i - v_{c'} \right\|^2}]^{\frac{2}{m-1}}}$ به دست می‌آید.

۳- سپس با استفاده از تابع هدف مدل خوشه‌بندی فازی C- مدوید، مقدار تابع هدف نماینده جدید به دست می‌آید. اگر مقدار تابع هدف نماینده جدید از تابع هدف نماینده قبلی کوچکتر باشد نماینده جدید به عنوان نماینده خوشه در نظر گرفته می‌شود و به مرحله ۲ بر می‌گردد، در غیر این صورت خوشه‌بندی پایان می‌یابد [۱,۹].

۳,۳- فاصلهٔ نمایی: پرکاربردترین معیار اندازه‌گیری شباهت، معیار اندازه‌گیری فاصله است که میزان شباهت و یا تفاوت بین دو نقطه را اندازه‌گیری می‌کند. فاصله‌ی اقلیدسی شناخته شده‌ترین و رایج‌ترین فاصله‌ی متریک می‌باشد، با این حال تابع هدف به دست آمده از فاصله‌ی اقلیدسی ممکن است در فضاهای دارای داده‌های پرت به خوبی عمل نکند. یکی از رویکردهای اندازه‌گیری فاصله که در برابر داده پرت عملکرد بهتری دارد، رویکرد فاصله‌ای نمایی می‌باشد که به صورت زیر تعریف می‌شود:

$$d(x, y) = 1 - \exp(-\beta \left\| x - y \right\|^2)$$

از طرفی فاصله نمایی نیز مانند فاصله اقلیدسی یک فاصله متریک است و می‌توان از این فاصله برای اندازه‌گیری شباهت در روش‌های خوشه‌بندی استفاده نمود [۱۴].

۳,۴- خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی: دورسو و همکارانش در سال ۲۰۱۶ مدل‌های خوشه‌بندی فازی بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی[۷] برای طبقه‌بندی نوسانات سری‌های زمانی ارایه دادند [۷]. w_1 و w_2 در این مقاله ثابت فرض شده‌اند. فرض کنید y یک ماتریس $N \times T$ بعد باشد که N تعداد سری‌های زمانی و T طول این سری‌ها هستند.

[7] Garch-E-fc-medoids

$$y=\begin{pmatrix} y_{11} & \cdots & y_{1T} \\ . & . & . \\ . & . & . \\ . & . & . \\ y_{N1} & \cdots & y_{NT} \end{pmatrix}_{N\times T} \quad (۳,۱)$$

$$v=\begin{pmatrix} uv_1 & tvv_1 \\ . & . \\ . & . \\ . & . \\ uv_N & tvv_N \end{pmatrix}_{N\times 2} \quad (۳,۲)$$

در ماتریس ۳.۲ ستون اول نوسانات غیرشرطی uv و ستون دوم نوسانات زمان متغیر tvv هر سری زمانی هستند. با توجه به ماتریس ۳.۱ در مدل خوشه‌بندی خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی، فاصله بر اساس معادلهٔ

$$1-e^{(-\beta[w_1^2 (uv_i-uv_{i'})^2+w_2^2 tvv(v_i-tvv_{i'})^2} \quad (۳,۳)$$

محاسبه می شود. تابع هدف این مدل به صورت زیر به دست می‌آید:

$$minf_{Garch-E-fc-medoids}=\sum_{i=1}^{N}\sum_{c=1}^{C}u_{ic}^{m}[1-e^{-\beta[w_1^2(uv_i-uv_c)^2+w_2^2(tvv_i-tvv_c)^2]}] \quad (۳,۴)$$

در این خوشه‌بندی میزان عضویت داده‌ی iام به خوشه‌ی cام توسط u_{ic} نشان داده می‌شود و N تعداد کل داده‌ها، C تعداد کل خوشه‌ها و m ضریب فازی و (uv_c, tvv_c) مرکز خوشه‌ی cام هستند. β پارامتر مثبت و ثابت است که از طریق معکوس واریانس داده‌ها به دست می‌آید. نحوه محاسبه پارامتر β به شرح زیر است:

ابتدا میانگین uv و tvv سری‌های زمانی را به دست آورده سپس از طریق فرمول ۲.۴ فاصله uv و tvv سری‌های زمانی را با میانگین آنها محاسبه کرده و در پایان مقدار به دست آمده بر تعداد کل سری‌های زمانی تقسیم می‌شود [۱۴].

مقدار β بر فاصله و میزان عضویت (در خوشه‌بندی های فازی) تأثیرگذار است. تابع لاگرانژ فرمول ۳.۴ به صورت زیر نوشته است:

$$L_m(u_{ic},\lambda,w_1,w_2)=\sum_{i=1}^{N}\sum_{c=1}^{C}u_{ic}^{m}[1-e^{-\beta[w_1^2(uv_i-uv_c)^2+w_2^2(tvv_i-tvv_c)^2]}]-\lambda(\sum_{c=1}^{C}u_{ic}-1) \quad (۳,۵)$$

ابتدا از فرمول ۳,۵ نسبت به u_{ic} و λ مشتق گرفته می‌شود و سپس با صفر قرار دادن رابطه‌ها مقدار بهینهٔ u_{ic} به دست می‌آید.

$$\frac{\partial L_m(u_{ic},\lambda,w_1,w_2)}{\partial u_{ic}}=0\Leftrightarrow mu_{ic}^{m-1}[1-e^{-\beta[w_1^2(uv_i-uv_c)^2+w_2^2(tvv_i-tvv_c)^2]}]-\lambda=0$$

$$\frac{\partial L_m(u_{ic},\lambda,w_1,w_2)}{\partial \lambda}=0\Leftrightarrow \sum_{c=1}^{C}u_{ic}-1=0 \qquad (۳.۶$$

$$u_{ic}=(\frac{\lambda}{m})^{\frac{1}{m-1}}[\frac{1}{1-e^{-\beta[w_1^2(uv_i-uv_c)^2+w_2^2(tvv_i-tvv_c)^2]}}]^{\frac{1}{m-1}} \qquad (۳.۷$$

بر اساس دو فرمول ۳.۶ و ۳.۷ فرمول زیر به دست می‌آید:

(۳.۸)

$$(\frac{\lambda}{m})^{\frac{1}{m-1}}\sum_{c=1}^{C}[\frac{1}{1-e^{-\beta[w_1^2(uv_i-uv_c)^2+w_2^2(tvv_i-tvv_c)^2]}}]^{\frac{1}{m-1}}=1\Rightarrow(\frac{\lambda}{m})^{\frac{1}{m-1}}=\frac{1}{\sum_{c=1}^{C}[\frac{1}{1-e^{-\beta[w_1^2(uv_i-uv_c)^2+w_2^2(tvv_i-tvv_c)^2]}}]^{\frac{1}{m-1}}}$$

از طرفی با توجه به دو فرمول ۳.۷ و ۳.۸ مقدار بهینه u_{ic} به صورت زیر نوشته می‌شود:

$$u_{ic}=\frac{1}{\sum_{c'=1}^{C}[\frac{1-e^{-\beta[w_1^2(uv_i-uv_c)^2+w_2^2(tvv_i-tvv_c)^2]}}{1-e^{-\beta[w_1^2(uv_i-uv_{c'})^2+w_2^2(tvv_i-tvv_{c'})^2]}}]^{\frac{1}{m-1}}} \qquad (۳.۹)$$

$c'=1,...,C$ نشان‌دهنده‌ی تعداد خوشه‌ها است. از طرفی میزان عضویت هر داده به خوشه‌های مختلف باید از محدودیت رو به رو پیروی کند: $\sum_{c=1}^{C}u_{ic}=1, 0<u_{ic}<1$.

الگوریتم روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصله نمایی به صورت زیر است:

۱- مقدار ضریب فازی، تعداد خوشه‌ها، مقدار بتا، بیشترین تعداد تکرار، میزان عضویت اولیه‌ی هر داده به هر خوشه و c داده به عنوان نماینده‌ی اولیه خوشه‌ها و $\tilde{H}$ زیر ماتریس شامل C نماینده و مقدار w_1 و w_2 توسط محقق در نظر گرفته می‌شود.

۲- در این مرحله ماتریس $H^{(s)}$، جایی‌که $s\geq 1$ نشان‌دهندهٔ تعداد تکرار است، از طریق فرمول

$$q=\underset{1\leq i'\leq N}{argmin}\sum_{i=1}^{N}u_{ic}^{m}[1-e^{-\beta[w_1^2(uv_i-uv_{i'})^2+w_2^2(tvv_i-tvv_{i'})^2]}] \qquad (۳.۱۰)$$

برای هر $c=1,...,C$ به روز می‌شود و سپس با استفاده از فرمول ۳.۹ میزان عضویت داده به خوشه‌ها با توجه به نماینده جدید هر خوشه به دست می‌آید.

۳- اگر مقدار ماتریس $H^{(s)}=H^{(s-1)}$ باشد، یا تعداد تکرار s بیشترین باشد الگوریتم متوقف می‌شود، در غیر این صورت به مرحلهٔ ۲ بر می‌گردد [۷].

۳,۵- خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصله‌ی اقلیدسی: تابع هدف مدل خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصلهٔ اقلیدسی[8]

$$((uv_i - uv_{,})^2 + (tvv_i - tvv_{,})^2)^{\frac{1}{2}} \qquad (۳.۱۱)$$

به صورت زیر به دست می‌آید:

$$minf_{Garch-fc-medoids} = \sum_{i=1}^{N}\sum_{c=1}^{C} u_{ic}^{m}((uv_i - uv_c)^2 + (tvv_i - tvv_c)^2)^{\frac{1}{2}}$$

سپس از بین داده‌های موجود در هر خوشه یک داده به عنوان نماینده جدید هر خوشه انتخاب می‌شود، اگر مقدار تابع هدف نماینده جدید از تابع هدف نماینده قبلی کوچکتر باشد نماینده جدید به عنوان مرکز خوشه در نظر گرفته می‌شود، در غیر این صورت خوشه‌بندی پایان می‌یابد [۱].

۳,۶- خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از فاصله‌ی اقلیدسی: در این خوشه‌بندی که از جمله خوشه‌بندی‌های سخت به شمار می‌رود، از مجموعه داده‌ها k داده به عنوان نماینده اولیه خوشه انتخاب می‌شود. سپس با استفاده از فاصله اقلیدسی ۳.۱۱ داده‌هایی که به این نماینده‌ها نزدیک‌ترند در یک خوشه قرار گرفته می‌شوند. تابع هدف روش خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از فاصله‌ی اقلیدسی[9] توسط فرمول زیر به دست می‌آید:

$$minf_{Garch-K-Medoids} = \sum_{i=1}^{N}\sum_{c=1}^{C}((uv_i - uv_c)^2 + (tvv_i - tvv_c)^2)^{\frac{1}{2}} \qquad (۳.۱۲)$$

uv_i نوسانات غیرشرطی و tvv_i نوسانات زمان متغیر سری‌های زمانی هستند. سپس از بین داده‌های موجود در هر خوشه داده‌ای به عنوان نمایندهٔ جدید انتخاب می‌شود. اگر مقدار تابع هدف نماینده جدید از تابع هدف نماینده قبلی کوچکتر باشد نمایندهٔ جدید به عنوان مرکز خوشه در نظر گرفته می‌شود، در غیر این صورت خوشه‌بندی پایان می‌یابد.

۳,۷- خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی: در این خوشه‌بندی با استفاده از فاصلهٔ ۳.۳ تابع هدف روش خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی[10] به صورت زیر به دست می‌آید:

$$minf_{Garch-E-k-means} = \sum_{i=1}^{N}\sum_{c=1}^{C}[1 - e^{-\beta[w_1^2(uv_i - uv_c)^2 + (w_2)^2(tvv_i - tvv_c)^2]}] \qquad (۳.۱۳)$$

[8] Garch-fc-medoids
[9] Garch-K-Medoids
[10] Garch-E-k-medoids

الگوریتم روش خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی به صورت زیر است:

۱- بیشترین تعداد تکرار، c داده به عنوان نمایندهٔ اولیهٔ خوشه‌ها، مقدار بتا و $\tilde{H}$ زیر ماتریس شامل C نماینده و مقدار w_1 و w_2 توسط محقق در نظر گرفته می‌شود.

۲- در این مرحله ماتریس $H^{(s)}$، جایی‌که $s \geq 1$ نشان‌دهنده‌ی تعداد تکرار است، از طریق فرمول

$$q = \underset{1 \leq i' \leq N}{argmin} \sum_{i=1}^{N} 1 - e^{(-\beta[w_1^2 (uv_i - uv_{i'})^2 + (w_2)^2 (tvv_i - tvv_{i'})^2])} \qquad (۳.۱۴)$$

برای هر $c = 1,...,C$ به روز می‌شود و سپس با استفاده از فاصلهٔ ۳.۳ نشان داده می‌شود هر داده به خوشه‌ای تعلق دارد که کمترین فاصله را تا نمایندهٔ آن خوشه داشته باشد.

۳- اگر مقدار ماتریس $H^{(s)} = H^{(s-1)}$ باشد، یا تعداد تکرار s بیشترین باشد الگوریتم متوقف می‌شود، در غیر این صورت به مرحلهٔ ۲ برمی‌گردد.

۴- تشخیص سری‌های زمانی پرت

در این مقاله برای تشخیص سری‌های زمانی پرت در نوسانات زمان متغیر و غیر شرطی مدل گارچ از میزان پرت افتادگی هر داده استفاده شده است. در این روش هدف به دست آوردن خوشه‌بندی دقیق‌تر به منظور تجزیه و تحلیل داده‌ها است. روند پیاده‌سازی این الگوریتم به صورت زیر به دست می‌آید:

۱- ابتدا هر کدام از روش‌های خوشه‌بندی را بر مجموعه داده‌ها برای به دست آوردن مراکز نهایی هر خوشه پیاده‌سازی کرده، سپس با استفاده از فاصلهٔ ۳.۳، فاصله هر داده تا مرکز خوشه مربوطه به دست آورده می‌شود. در خوشه‌بندی فازی داده به خوشه‌ای متعلق است که بیشترین میزان عضویت را به آن خوشه داشته باشد [۲].

۲- از بین فواصل موجود در هر خوشه بیشترین فاصله انتخاب و آن با d_{cmax} (بیشترین فاصله در خوشهٔ cام) نشان داده می‌شود.

۳- سپس برای به‌دست آوردن فاکتور پرت افتادگی فاصله به دست آمده برای هر داده در مرحلهٔ اول را بر ماکزیمم فاصله موجود در آن خوشه تقسیم می‌کنیم:

۱- برای روش‌های خوشه‌بندی با فاصلهٔ اقلیدسی

$$o_i = \frac{[(uv_i - uv_c)^2 + (tvv_i - tvv_c)^2]^{\frac{1}{2}}}{d_{cmax}}$$

۲- برای روش‌های خوشه‌بندی با فاصلهٔ اقلیدسی

$$o_i = \frac{1-e^{(-\beta[w_1^2(uv_i-uv_c)^2+w_2^2(tvv_i-tvv_c)^2])}}{d_{cmax}}$$

۴- در این مرحله محقق آستانه‌ای در محدودهٔ o_i در نظر می‌گیرد [۱۲]. (در این مقاله مقدار آستانه برای تمامی روش‌های خوشه‌بندی ۰.۸ در نظر گرفته شده است.) در این صورت تشخیص سری‌های زمانی پرت به صورت زیر انجام می‌گیرد:

۱- برای روش خوشه‌بندی K- مدوید بر اساس گارچ با استفاده از فاصله اقلیدسی و با استفاده از رویکرد فاصله‌ای نمایی

$$\forall i = 1, \dots, N \quad u_{ic} = \begin{cases} 1 & \text{اگر} \quad o_i \le T \\ 0 & \text{اگر} \quad o_i > T \end{cases} \qquad (۴,۱)$$

۲- برای روش‌های خوشه‌بندی فازی C- مدوید بر اساس گارچ با استفاده از فاصلهٔ اقلیدسی

(۴,۲)

$$u_{ic} = \begin{cases} \dfrac{1}{\sum_{\acute{c}=1}^{C}[\frac{[(uv_i-uv_c)^2+(tvv_i-tvv_c)^2]^{\frac{1}{2}}}{[(uv_i-uv_{\acute{c}})^2+(tvv_i-tvv_{\acute{c}})^2]^{\frac{1}{2}}}]^{\frac{1}{m-1}}} & \text{اگر} \quad o_i \le T \\ 0 & \text{اگر} \quad o_i > T \end{cases}$$

۳- برای روش‌های خوشه‌بندی فازی C- مدوید بر اساس گارچ با استفاده از رویکرد فاصله‌ای نمایی

$$u_{ic} = \begin{cases} \dfrac{1}{\sum_{\acute{c}=1}^{C}[\frac{1-e^{-\beta[w_1^2(uv_i-uv_c)^2+w_2^2(tvv_i-tvv_c)^2]}}{1-e^{-\beta[w_1^2(uv_i-uv_{\acute{c}})^2+w_2^2(tvv_i-tvv_{\acute{c}})^2]}}]^{\frac{1}{m-1}}} & \text{اگر} \quad o_i \le T \\ 0 & \text{اگر} \quad o_i > T \end{cases} \qquad (۴.۳)$$

در این روش اگر میزان پرت افتادگی از مقدار آستانهٔ تعیین شده بیشتر باشد، آن نقطه سری‌های زمانی پرت در نظر گرفته می‌شود و میزان درجهٔ عضویت صفر به آن نقطه تعلق می‌گیرد. در واقع از طریق روش ارائه شدهٔ فوق سری‌های زمانی پرت در حین فرآیند خوشه‌بندی تشخیص و از مجموعه داده‌ها حذف می‌شود [۱۲]. هدف از ارزیابی روش‌های خوشه‌بندی، تجزیه و تحلیل ساختار خوشه‌های ایجاده شده توسط روش‌های خوشه‌بندی است. برای سنجش صحت نتایج خوشه‌بندی، ملاک‌ها و معیارهای متفاوتی معرفی شده است. این شاخص‌ها سعی در اندازه‌گیری میزان شباهت اعضای درون خوشه و عدم شباهت بین خوشه‌ها دارند. بنابراین روشی که بیشترین شباهت درون خوشه یا بیشترین تمایز بین خوشه‌ها را ایجاد کند، روش

مناسبی در نظر گرفته می‌شود. در این مقاله از دو شاخص زایبنی[11] و سیلهوت[12] برای ارزیابی روش‌های خوشه‌بندی استفاده شده است.

۴,۱- زایبنی: در شاخص زایبنی صورت کسر کل فاصلهٔ درون خوشه‌ای است که به آن فشردگی می‌گویند. مخرج که مینیمم مربع فاصله بین دو مرکز است را جدایی می‌گویند. برای تعداد مشخص خوشه و داده، هر چه قدر مقدار زایبنی به صفر نزدیکتر باشد خوشه‌بندی بهتری انجام گرفته است [۴].

$$XB = \frac{\sum_{c=1}^{C}\sum_{i=1}^{N} u_{ic}^{m} \times 1 - e^{-\beta[w_1^2(uv_i - uv_c)^2 + w_2^2(tvv_i - tvv_c)^2]}}{N \times min_{p \neq q}[1 - e^{-\beta[w_1^2(uv_{c_p} - uv_{c_q})^2 + w_2^2(tvv_{c_p} - tvv_{c_q})^2]}} \qquad (۴.۴)$$

c_p مرکز خوشهٔ p ام و c_q مرکز خوشهٔ q ام هستند که $p \neq q$. با این‌که شاخص زایبنی، شاخصی برای روش‌های خوشه‌بندی فازی است اما برای روش‌های خوشه‌بندی سخت نیز کاربرد دارد. شاخص زایبنی برای روش‌های خوشه‌بندی سخت به صورت میانگین فواصل هر داده تا مرکز خوشه‌ی مربوطه تعریف می‌شود [۵].

$$XB = \frac{\sum_{c=1}^{C}\sum_{i=1}^{N} 1 - e^{-\beta[w_1^2(uv_i - uv_c)^2 + w_2^2(tvv_i - tvv_c)^2]}}{N} \qquad (۴.۵)$$

۴,۲- شاخص سیلهوت: برای به دست آوردن مقدار سیلهوت در خوشه‌بندی‌های سخت فاصله هر داده تا مراکز خوشه‌ها را اندازه‌گیری کرده، داده به خوشه‌ای تعلق دارد که کمترین فاصله را تا مرکز آن خوشه داشته باشد. در خوشه‌بندی‌های فازی مقدار عضویت داده به تمامی خوشه‌ها را به دست آورده، داده به خوشه‌ای تعلق دارد که بیشترین میزان عضویت را به آن خوشه داشته باشد. سپس میانگین فاصله هر داده با داده‌هایی که در خوشه p ام قرار دارند را به دست آورده و آن را با a_{ip} نشان داده می‌شود. سپس میانگین فاصله هر داده با داده‌هایی که در یک خوشه قرار ندارند را به دست آورده و آن را با d_{iq} که $p \neq q$ نشان داده می‌شود. b_{ip} کمترین مقدار d_{iq} است که نشان دهنده عدم تشابه داده iام با نزدیک‌ترین خوشهٔ همسایه است. برای تعداد مشخص خوشه و داده، هرچه قدر مقدار سیلهوت به یک نزدیک باشد خوشه‌بندی بهتری انجام گرفته است. مقدار سیلهوت هر شی به صورت زیر به دست می‌آید:

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \qquad (۴,۶)$$

مقدار سیلهوت کل خوشه‌بندی در خوشه‌بندی‌های سخت به صورت زیر به دست می‌آید:

[11] Xe-Beni
[12] silhoutte

$$s_i = \frac{\sum_{i=1}^{N} s_i}{N} \qquad (۴,۷)$$

مقدار سیلهوت کل خوشه‌بندی در خوشه‌بندی‌های فازی به صورت زیر به دست می‌آید:

$$s_i = \frac{\sum_{i=1}^{N} (u_{ip} - u_{iq})^{\alpha} s_i}{\sum_{i=1}^{N} (u_{ip} - u_{iq})^{\alpha}} \qquad (۴,۸)$$

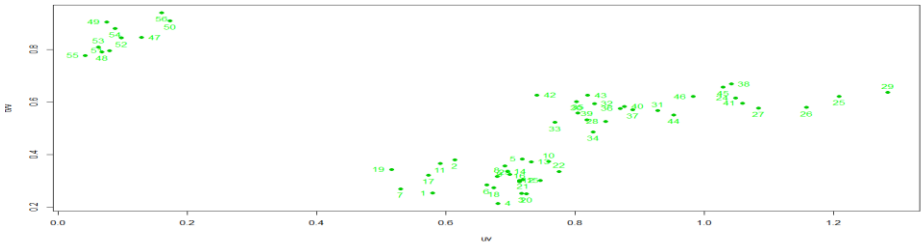
در واقع u_{ip} و u_{iq} اولین و دومین میزان عضویت بزرگ در ستون iام ماتریس عضویت هستند و آلفا برابر یک است [۴,۵]

۵- شبیه‌سازی

در این قسمت به شبیه‌سازی مطالعاتی جهت ارزیابی عملکرد و دقت روش‌های خوشه‌بندی مطرح شده در این مقاله پرداخته می‌شود. مجموعه داده‌ها از ۴۶ سری زمانی با طول ۱۰۰۰ تشکیل شده‌اند. سری‌های زمانی موجود در این مجموعه داده‌ها بر اساس یک حالت تولید شده‌اند. این حالت شامل دو خوشه جدا از سری‌های زمانی می‌باشد که با دو فرآیند گارچ مختلف شبیه‌سازی شده‌اند. روند تولید در این حالت بدین صورت است که نخستین خوشه توسط گارچ با پارامترهای (۰.۴,۰.۳,۰.۲) و روند تولید دومین خوشه توسط GARCH (۰.۴,۰.۶,۰.۲) است. به منظور ارزیابی دقت روش‌های خوشه‌بندی مطرح شده در این مقاله ۱۰ داده پرت (منظور از داده همان سری‌های زمانی است) به مجموعه داده‌ها اضافه می‌شود. داده‌های تولید شده توسط شبیه‌سازی مطالعاتی در جدول ۱ و مقدار uv و tvv ۵۶ سری زمانی تولید شده توسط این حالت در شکل ۱ نمایش داده شده است.

جدول۱: مجموعه داده‌ها و داده‌های پرت تولید شده در شبیه‌سازی مطالعاتی

-	$GARCH(\gamma, \alpha, \beta)$	uv	tvv
خوشه۱ ($i = ۱,..., \frac{N}{۲}$) خوشه۲ ($i = \frac{N}{۲} + ۱,..., N$)	GARCH (۰.۴,۰.۳,۰.۲) GARCH (۰.۴,۰.۶,۰.۲)	۰.۶۲۲۹ ۰.۸۵۸۳	۰.۳۰۱۹ ۰.۵۶۰۵
پارامترهای داده پرت در این حالت	$\gamma \sim N$ (۰.۰۲,۰.۰۰۵) $\alpha \sim N$ (۰.۸۵,۰.۰۰۱) $\beta \sim U$ (۰.۱,۱−α)	۰۱۰۵۲	۰.۸۸۳۵



شکل ۱: مقدار uv و tvv سری‌های زمانی تولید شده

برای تولید داده پرت در حالت ذکر شده مقدار پارامتر γ از توزیع گوسی $N(۰.۰۲,۰.۰۰۵)$، مقدار پارامتر α از توزیع $N(۰.۸۵,۰.۰۰۱)$ و مقدار پارامتر β از توزیع یکنواخت $U(۰.۱,۱-\alpha)$ $(\alpha+\beta<۱)$ تولید شده‌اند. برای ارزیابی روش‌های خوشه‌بندی از شاخص‌های سیلهوت و زایبنی استفاده می‌شود. مقدار شاخص‌های زایبنی و سیلهوت در جدول ۲ نشان داده شده است.

جدول ۲: مقدار شاخص‌های ارزیابی روش‌های خوشه‌بندی

-	داده پرت	حالت	
		سیلهوت	زای بنی
$GARCH-K-medoids$	۰ ۱۰	۰.۶۷۲۱ ۰.۶۳۱۶	۰.۱۰۹۶ ۰.۴۵۶۶
$GARCH-fc-medoids$	۰ ۱۰	۰.۸۶۴۳ ۰.۵۶۸۹	۰.۰۷۰۲ ۲.۲۹۹۲
$GARCH-E-K-medoids$	۰ ۱۰	۰.۷۹۰۹ ۰.۶۷۸۹	۰.۱۰۳۲ ۰.۱۷۶۳
$GARCH-E-fc-medoids$	۰ ۱۰	۰.۸۱۲۱ ۰.۷۵۴۲	۰.۱۰۰۴ ۰.۱۲۶۷

هرچه در یک روش، مقادیر شاخص سیلهوت بزرگتر و زایبنی کوچکتر باشد، آن روش، مناسبتر خواهد بود. با توجه به جدول ۲، اگرچه روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصله اقلیدسی در عدم حضور داده پرت بیشترین مقدار شاخص سیلهوت و کمترین مقدار زایبنی را داراست، اما با افزودن داده پرت مقدار شاخص سیلهوت این روش کمترین و شاخص زایبنی بیشترین شده است، که نشان می‌دهد این روش نسبت به حضور داده پرت حساسیت زیادی داشته و در نتیجه یک روش قوی محسوب نمی‌شود. در صورتی که در روش‌های خوشه‌بندی K- مدوید و فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصله نمایی، چه در حضور داده پرت و چه در عدم حضور داده پرت، مقادیر سیلهوت بزرگ و زایبنی کوچکی دارند و از طرفی تأثیرات داده پرت در مقادیر این دو شاخص بسیار اندک است، که نشان از قوی بودن این دو روش می‌باشد.

۶- مطالعات تجربی

در این بخش روش‌های خوشه‌بندی مطرح شده در این مقاله با استفاده از فاصله اقلیدسی و رویکرد فاصله‌ای نمایی با سری‌های زمانی پرت را با همان روش‌های خوشه‌بندی بدون سری‌های زمانی پرت برای سهام ۳۰ شرکت برتر، فعال و پر سود موجود در بازار بورس ایران مقایسه می‌کنیم. این سهام‌ها از فروردین سال ۱۳۹۴ تا فروردین سال ۱۳۹۸ در نظر گرفته شده‌اند. تمامی سری‌های زمانی به کار رفته در این مقاله در سایت بازار بورس[13] قرار دارند. در این مقاله از نرم افزار R و شاخص سیلهوت و زایینی برای مقایسهٔ دقت خوشه‌بندی استفاده شده است. مؤلفه‌های وزنی $w_1 = 0.1, = w_2 = 0.9$ و در روش‌های فازی، ضریب فازی $m = 1.5$ در نظر گرفته شده است. ابتدا برای هر کدام از سری‌های زمانی آزمون آرچ[14] گرفته می‌شود. فرض صفر در این آزمون بیانگر عدم وجود نوسانات شرطی (اثرات آرچ) است. در بیشتر نرم‌افزارهای آماری مانند R برای سهولت در تصمیم‌گیری نسبت به نتیجه آزمون فرض آماری، شاخصی به نام مقدار احتمال ارائه می‌شود. در این آزمون اگر مقدار احتمال[15] کمتر از مقدار آلفا (۰.۱,۰.۰۱,۰.۰۵) باشد، فرض صفر رد شده و در نتیجه می‌توان به آن سری زمانی فرآیند گارچ برازش داد. سپس از دو نوع متفاوت فرآیند GARCH(۱,۱) و GARCH(۲,۲) بر اساس فرآیند آرما (۱,۱) استفاده شده و AIC[16] و BIC[17] دو فرآیند گارچ به دست آمده است. هرکدام از فرآیندهای گارچ که مقدار AIC و BIC کمتری داشته باشد آن فرآیند گارچ برای سری زمانی مناسب است. مقدار AIC و BIC ۳۰ سری زمانی در هر دو فرآیند گارچ و مقدار احتمال به دست آمده از آزمون آرچ در جدول ۱ نمایش داده شده است. با توجه به جدول ۳ به دلیل آن که مقدار احتمال به دست آمده از آزمون آرچ تمامی سری‌های زمانی از مقدار α کمتر هستند، می‌توان مدل گارچ را برای سری‌های زمانی پیاده‌سازی کرد. از طرفی سری‌های زمانی (۱,۳,۵,۶,۱۶,۱۷,۱۹,۲۲,۲۳,۲۵,۲۷,۳۰) از GARCH(۱,۱) و سری‌های زمانی (۲,۴,۷,۸,۹,۱۰,۱۱,۱۲,۱۳,۱۴,۱۵,۱۸,۲۰,۲۱,۲۴,۲۶,۲۸,۲۹) از GARCH(۲,۲) پیروی می‌کنند. به مقدار ضرایب گارچ سری زمانی‌های در جدول ۴ و مقدار uv و tvv سری‌های زمانی در جدول ۵ و در شکل ۲ نمایش داده شده است.

[13] tsetmc.com
[14] Arch Test
[15] P-Value
[16] Akaike Information Criteria
[17] Bayes Information Criteria

جدول ۳: مقدار AIC و BIC فرآیندهای گارچ و مقدار احتمال به دست آمده از آزمون آرچ سری‌های زمانی

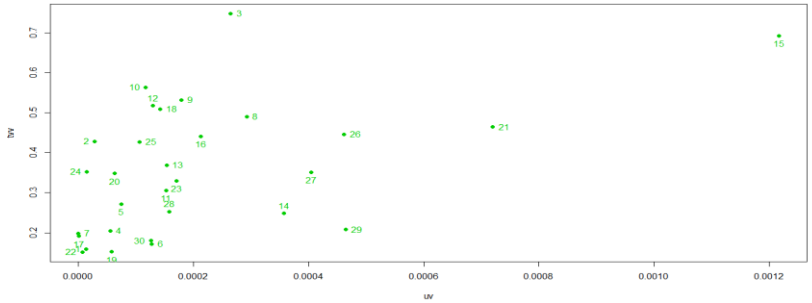
ردیف	نماد سهام	GARCH (۱,۱)		GARCH (۲,۲)		مقدار احتمال
		AIC	*BIC*	*AIC*	*BIC*	
۱	کگل	−۴.۵۰۸۵	−۴.۴۷۹۴	−۴.۴۲۱۶	−۴.۳۸۲۷	۰
۲	کچاد	−۵.۰۰۶۲	−۴.۹۷۷۱	−۵.۰۱۶۶	−۴.۹۷۷۸	۰
۳	پارسان	−۵.۲۴۲۷	−۵.۲۱۳۶	−۵.۲۴۰۷	−۵.۲۰۱۹	۰
۴	ومعادن	−۵.۲۳۳۹	−۵.۲۰۴۷	−۵.۲۳۷۵	−۵.۱۹۸۶	۰
۵	اخابر	−۵.۶۷۷۱	−۵.۶۴۸۰	−۵.۶۷۲۹	−۵.۶۳۴۱	۰
۶	خودرو	−۴.۷۷۸۹	−۴.۷۴۹۷	−۴.۷۸۶۷	−۴.۷۴۷۹	۰.۰۰۰۰۰۰۳
۷	وانصار	−۶.۸۱۳۲	−۶.۷۸۴۱	−۷.۷۴۱۹	−۷.۷۰۳۱	۰.۰۰۰۰۸۱۲
۸	شاراک	−۴.۶۷۷۸	−۴.۶۴۸۷	−۴.۸۰۶۳	−۴.۷۶۷۵	۰.۰۰۰۰۲۴۹
۹	سفارس	−۴.۹۵۱۰	−۴.۹۲۱۹	−۵.۰۲۹۴	−۴.۹۹۰۶	۰
۱۰	شسپا	−۵.۶۰۹۵	−۵.۵۸۰۴	−۵.۶۵۷۰	−۵.۶۱۸۲	۰.۰۰۰۰۰۲۴
۱۱	ستران	−۴.۹۷۰۴	−۴.۹۴۱۳	−۴.۹۹۴۴	−۴.۹۵۵۶	۰.۰۰۷۲۴۴
۱۲	وخاور	−۵.۶۲۰۴	−۵.۵۹۱۳	−۵.۶۲۳۴	−۵.۵۸۴۶	۰.۰۰۶۳۲۷
۱۳	لبوتان	−۵.۱۸۵۰	−۵.۱۵۵۹	−۵.۲۱۶۴	−۵.۱۷۷۵	0
۱۴	بتایر	−۴.۴۵۷۲	−۴.۴۲۸۱	−۴.۴۵۹۴	−۴.۴۲۰۶	۰.۰۰۰۰۰۱۸
۱۵	کرماشا	−۵.۷۱۳۶	−۵.۶۸۴۵	−۵.۷۴۱۰	−۵.۷۰۲۲	0
۱۶	شرانل	−۵.۰۸۲۸	−۵.۰۵۳۷	−۵.۰۷۸۰	−۵.۰۳۹۱	۰.۰۰۰۱۶۳۵
۱۷	کساوه	−۴.۲۲۶۵	−۴.۱۹۷۴	−۳.۷۲۷۸	−۳.۶۸۹۰	0
۱۸	هرمز	−۵.۱۶۳۰	−۵.۱۳۳۸	−۵.۱۶۶۱	−۵.۱۲۷۲	۰.۰۰۰۰۰۰۲
۱۹	پکرمان	−۴.۷۴۷۷	−۴.۷۱۸۶	−	−	0
۲۰	وبیمه	−۴.۶۱۲۸	−۴.۵۸۳۷	−۴.۶۳۶۰	−۴.۵۹۷۲	0
۲۱	رتاپ	−۴.۱۷۸۷	−۴.۱۴۹۵	−۴.۱۹۲۶	−۴.۱۵۳۷	0
۲۲	رتاپ	−۴.۶۱۵۳	−۴.۵۸۶۲	−۴.۲۸۸۱	−۴.۲۴۹۳	۰.۰۰۴۰۶۴

۲۳	وتوکا	−۴.۷۹۰۱	−۴.۷۶۱۰	−۴.۷۸۵۸	−۴.۷۴۷۰	0
۲۴	وخارزم	−۵.۴۱۶۷	−۵.۳۸۷۶	−۵.۴۳۱۱	−۵.۳۹۲۳	۰.۰۰۰۰۷۱۳
۲۵	البرز	−۵.۶۸۱۵	−۵.۶۵۲۳	−۵.۶۷۷۹	−۵.۶۳۹۰	0
۲۶	واحیا	−۴.۰۶۰۷	−۴.۰۳۱۶	−۴.۰۹۰۸	−۴.۰۵۱۹	۰.۰۰۱۱۹۶
۲۷	شصفها	−۳.۷۸۸۸	−۳.۷۵۹۶	−۳.۷۸۸۲	−۳.۷۴۹۴	۰.۰۰۹۷۹۱
۲۸	آسیا	−۵.۱۰۹۷	−۵.۰۸۰۶	−۵.۱۱۳۶	−۵.۰۷۴۸	0
۲۹	ساذری	−۳.۵۵۵۹	−۳.۵۲۶۷	−۴.۰۶۳۵	−۴.۰۲۴۶	0
۳۰	کروی	−۴.۶۵۶۵	−۴.۶۲۷۴	−۴.۶۵۳۳	−۴.۶۱۴۵	۰.۰۰۰۷۰۷۴

جدول ۴: مقدار ضرایب گارچ سری‌های زمانی

ردیف	نماد سهام	γ	α_1	α_2	β_1	β_2
۱	ککل	۰.۰۰۰۰۰۰۱	۰.۶۱۱۵۷۴۹	–	۰.۹۲۲۹۴۰۵	–
۲	کچاد	۰.۰۰۰۰۰۷۹	۰.۳۵۵۸۹۴۳	۰	۰.۵۸۵۶۱۷۱	۰.۰۵۷۴۸۸۵
۳	پارسان	۰.۰۰۰۰۸۳۶	۰.۷۱۷۶۹۹۷	–	۰.۲۸۱۳۰۰۳	–
۴	ومعادن	۰.۰۰۰۰۱۵۱	۰.۲۰۴۵۰۵۱	۰.۱۳۷۲۰۲۶	۰.۶۵۱۷۶۹۶	۰.۰۰۳۸۱۹۲
۵	اخابر	۰.۰۰۰۰۰۹۷	۰.۱۳۸۲۲۶۱	–	۰.۸۶۰۷۷۳۹	–
۶	خودرو	۰.۰۰۰۰۱۱۱	۰.۰۷۱۵۲۹۶۸	–	۰.۹۰۹۷۰۱۰	–
۷	وانصار	۰	۰.۰۸۳۳۵۲۸۳	۰.۰۶۰۳۹۳۸۳	۰.۰۰۰۰۸۵۵	۰.۸۵۵۱۶۷۶
۸	شاراک	۰.۰۰۰۰۸۶۳	۰	۰.۳۸۸۴۶۱۴	۰	۰.۶۱۰۵۳۸۲
۹	سفارس	۰.۰۰۰۰۶	۰.۰۴۳۸۹۶۵۸	۰.۴۵۹۶۹۸۵	۰	۰.۴۹۵۴۰۴۶
۱۰	شسپا	۰.۰۰۰۰۴۰۱	۰.۱۰۱۴۶۲۰	۰.۵۰۸۸۲۸۲	۰	۰.۳۸۸۷۰۹۸
۱۱	ستران	۰.۰۰۰۰۴۸۱	۰.۰۹۳۵۱۵۹۹	۰.۲۹۱۰۶۲۹	۰.۵۸۴۶۰۳۰	۰.۰۰۰۰۰۰۱
۱۲	وخاور	۰.۰۰۰۰۴۵۲	۰.۴۹۲۹۹۳۱	۰.۰۰۰۰۰۰۶	۰.۲۳۲۴۴۵۵	۰.۲۴۳۴۶۵۰
۱۳	لبوتان	۰.۰۰۰۰۴۵۲	۰.۱۲۲۵۷۴۷	۰.۲۶۵۰۱۱۶	۰	۰.۶۱۱۴۱۳۶
۱۴	بتایر	۰.۰۰۰۱۳۲۹	۰.۰۷۹۲۴۹۶۲	۰.۱۹۱۸۹۱۷	۰	۰.۵۴۹۰۵۳۴
۱۵	کرماشا	۰.۰۰۰۰۶۸۷	۰.۴۱۶۵۱۸۱	۰.۵۵۳۰۶۱۲	۰.۰۰۰۰۰۰۴	۰.۰۲۹۴۲۰۳۶
۱۶	شرانل	۰.۰۰۰۰۸۷۵	۰.۳۹۸۴۹۱۱	–	۰.۴۳۰۴۹۲۲	–
۱۷	کساوه	۰	۰.۰۷۲۴۷۶۰۱	–	۰.۹۲۵۷۵۰۲	–
۱۸	هرمز	۰.۰۰۰۰۴۶۷	۰.۳۴۲۸۷۴۲	۰.۳۴۲۹۳۱۴	۰.۰۰۳۹۴۶۱۸۴	۰.۳۰۹۲۴۸۲

۱۹	پکرمان	۰.۰۰۰۰۰۲۹	۰.۰۴۹۰۴۰۹۲	–	۰.۹۴۷۳۲۸۶	–
۲۰	وبیمه	۰.۰۰۰۰۱۶۶	۰.۰۸۳۷۰۰۵۳	۰.۲۴۳۹۹۰۷	۰	۰.۶۷۱۳۰۸۷
۲۱	کیسون	۰.۰۰۰۲۷۶۸	۰.۱۹۲۸۸۷۹	۰.۴۲۲۴۸۵۴	۰.۰۰۰۰۰۰۳	۰.۲۳۲۴۶۹
۲۲	رتاپ	۰.۰۰۰۰۰۰۵	۰.۰۵۹۸۲۸۲۱	–	۰.۹۱۹۲۰۳۵	–
۲۳	وتوکا	۰.۰۰۰۰۴۱۴	۰.۲۲۹۱۸۳۸	–	۰.۷۱۹۴۹۲۳	–
۲۴	وخارزم	۰.۰۰۰۰۰۴۱	۰.۱۱۶۷۴۹۰	۰.۲۴۵۹۹۹۶	۰.۰۰۰۰۰۰۰۱	۰.۶۳۶۲۵۱۳
۲۵	البرز	۰.۰۰۰۰۳۶	۰.۳۵۲۷۰۳۳	–	۰.۵۶۳۴۰۸۵	–
۲۶	واحیا	۰.۰۰۰۱۶۷۳	۰.۱۶۹۹۸۳۱	۰.۳۶۸۶۸۹۹	۰	۰.۴۱۵۰۹۱۶
۲۷	شصفها	۰.۰۰۰۰۷۸۵	۰.۲۲۰۶۶۰۱	–	۰.۷۷۸۳۳۹۹	–
۲۸	آسیا	۰.۰۰۰۰۵۰۹	۰.۰۸۷۰۱۵۵۹	۰.۱۷۹۸۲۸۹	۰.۰۰۰۰۰۰۰۱	۰.۶۱۲۷۴۹۴
۲۹	ساذری	۰.۰۰۰۰۵۱۲	.۰۰۰۰۰۰۰۱	۰.۰۹۷۴۷۳۳۰	۰	۰.۸۸۴۲۶۱۲
۳۰	کروی	۰.۰۰۰۰۱۰۸	۰.۰۷۴۵۰۸۵۵	–	۰.۹۱۰۳۸۶۵	–



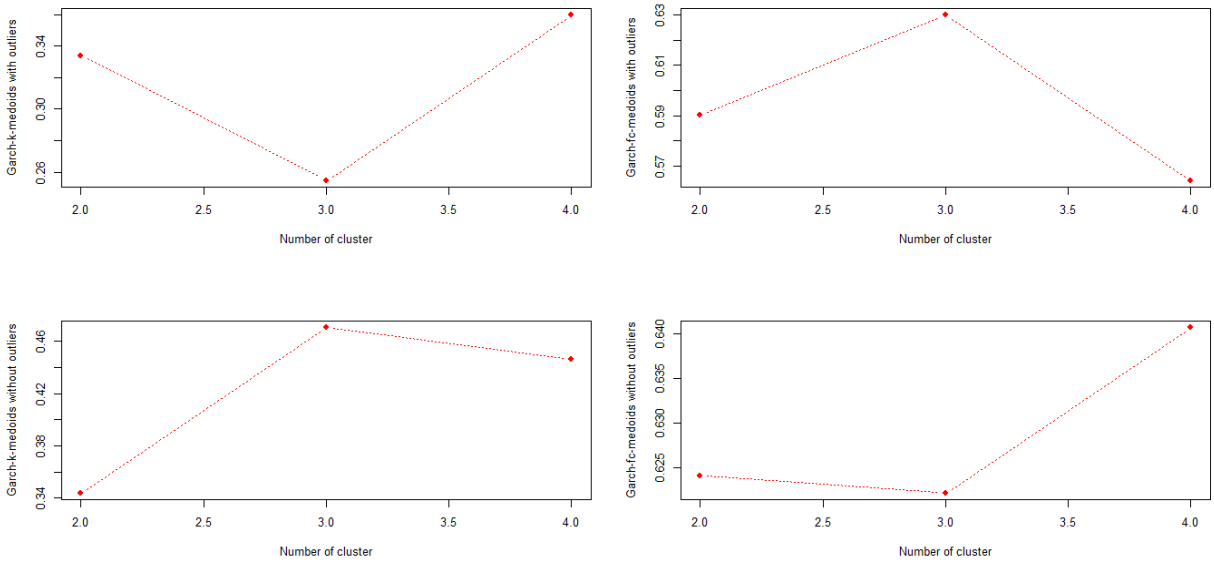
شکل ۲: مقدار uv و tvv سری‌های زمانی

جدول ۵: مقدار uv و tvv سری‌های زمانی

ردیف	نماد سهام	*uv*	*tvv*	ردیف	نماد سهام	*uv*	*tvv*
۱	کگل	۰.۰۰۰۰۱۳۴	۰.۱۵۸۷۴۴۲	۱۶	شرانل	۰.۰۰۰۲۱۲۹۸۸۹	۰.۴۴۱۴۹۵۳
۲	کچاد	۰.۰۰۰۰۲۸۴	۰.۴۲۸۱۲۵۳	۱۷	کساوه	۰.۰۰۰۰۰۰۴	۰.۱۹۱۴۶۶۶
۳	پارسان	۰.۰۰۰۲۶۴۵۴۶۶	۰.۷۴۷۹۰۰۱	۱۸	هرمز	۰.۰۰۰۱۴۲۲۶۱۳	۰.۵۰۹۱۶۷۳
۴	ومعادن	۰.۰۰۰۰۵۵۱	۰.۲۰۴۵۸۹۱	۱۹	پکرمان	۰.۰۰۰۰۵۷۱	۰.۱۵۲۱۱۴۲
۵	اخابر	۰.۰۰۰۰۷۴۹	۰.۲۷۱۵۶۹۹	۲۰	وبیمه	۰.۰۰۰۰۷۱۹	۰.۱۲۰۳۷۶۳
۶	خودرو	۰.۰۰۰۱۲۷۱۹۱۳	۰.۱۷۲۲۰۶۸	۲۱	کیسون	۰.۰۰۰۷۱۹۵۵۲	۰.۴۶۴۴۳۴۷
۷	وانصار	۰	۰.۱۹۸۳۸۲۱	۲۲	رتاپ	۰.۰۰۰۰۰۰۷	۰.۱۵۱۸۴۲۱
۸	شاراک	۰.۰۰۰۲۹۲۱۸۴۱	۰.۴۹۰۴۸۹۷	۲۳	وتوکا	۰.۰۰۰۱۷۰۱۹۱	۰.۳۲۹۹۹۸۱
۹	سفارس	۰.۰۰۰۱۷۹۱۶۷۸	۰.۵۳۱۶۱	۲۴	وخارزم	۰.۰۰۰۰۱۴۶	۰.۳۵۲۹۵۴۳
۱۰	شسپا	۰.۰۰۰۱۱۶۹۷۷	۰.۵۶۳۱۲۹۹	۲۵	البرز	۰.۰۰۰۱۰۶۴۱۲۷	۰.۴۲۶۹۰۹۳
۱۱	ستران	۰.۰۰۰۱۵۲۹۰۲۷	۰.۳۰۶۰۰۸۲	۲۶	واحیا	۰.۰۰۰۴۶۱۷۱۷۳	۰.۴۴۶۲۴۹۱

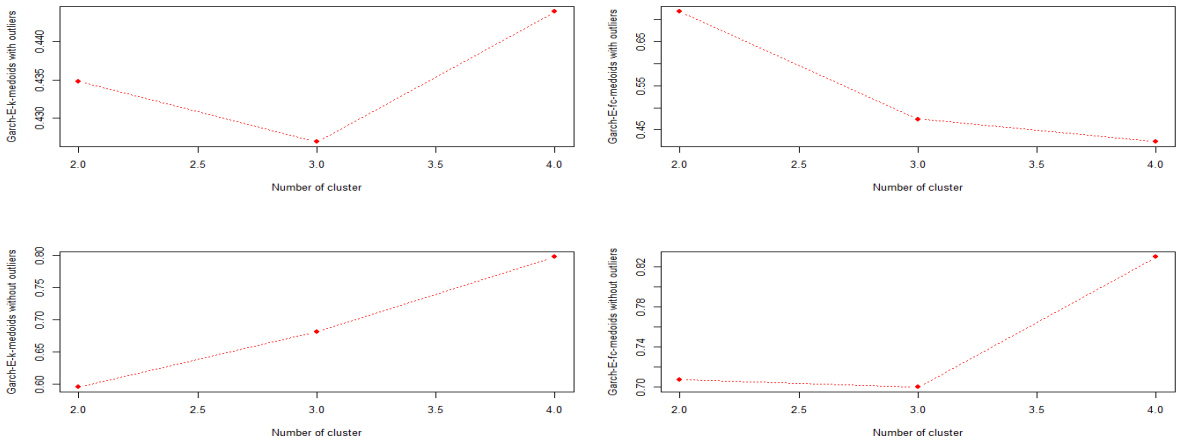
۱۲	وخاور	۰.۰۰۰۱۲۹۳۹۲۳	۰.۵۱۷۴۰۸۱	۲۷	شصفها	۰.۰۰۰۴۰۴۲۴۱۴	۰.۳۵۱۴۵۷۴
۱۳	لبوتان	۰.۰۰۰۱۵۳۰۲۵	۰.۳۶۸۹۸۹۸	۲۸	آسیا	۰.۰۰۰۱۵۷۳۴۹۵	۰.۲۵۲۷۹۱۳
۱۴	بتایر	۰.۰۰۰۳۵۷۱۶۹۹	۰.۲۴۸۴۰۳۵	۲۹	ساذری	۰.۰۰۰۴۶۴۰۳۱	۰.۲۰۷۹۶۲۵
۱۵	کرماشا	۰.۰۰۱۲۱۶۹۲۸	۰.۶۹۲۶۶۰۹	۳۰	کروی	۰.۰۰۰۱۲۵۸۶۰۹	۰.۱۸۰۰۲۸

با توجه به تعریفی که در مورد شاخص سیلهوت در فصل ۴ بیان شد، اگر شاخص سیلهوت روش خوشه‌بندی به یک نزدیک باشد روش خوشه‌بندی عملکرد بهتری دارد. ابتدا داده‌ها به ۲، ۳ و ۴ خوشه با استفاده از روش‌های گفته شده در این مقاله تقسیم شده و از طریق فرمول ۴.۶ و ۴.۷ و ۴.۸ مقدار شاخص سیلهوت به دست آورده می‌شود. سپس در هر روش خوشه‌بندی تعداد خوشه‌ای در نظر گرفته می‌شود، که مقدار شاخص سیلهوت آن به یک نزدیک‌تر است.



شکل ۳: ضرایب سیلهوت برای روش‌های خوشه‌بندی بر اساس فاصله‌ی اقلیدسی

با توجه به شکل ۳ تعداد دقیق خوشه‌ها در روش خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از فاصله اقلیدسی با داده پرت ۴ خوشه و همین روش بدون سری‌های زمانی پرت ۳ خوشه و و روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصله اقلیدسی با سری‌های زمانی پرت ۳ خوشه و همین روش بدون داده پرت ۴ خوشه است.



شکل ۴: ضرایب سیلهوت برای روش‌های خوشه‌بندی بر اساس رویکرد فاصله‌ای نمایی

با مشاهده شکل ۴ تعداد دقیق خوشه‌ها در روش‌های خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی با سری‌های زمانی پرت و همچنینی همین روش بدون سری‌های زمانی پرت ۴ خوشه و روش‌های خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی بدون سری‌های زمانی پرت ۴ خوشه و همین روش با سری‌های زمانی پرت ۲ خوشه هستند. بیشترین مقدار سیلهوت هر کدام از روش‌های خوشه‌بندی در جدول ۶ نشان داده شده است.

جدول۶ : مقدار ضریب سیلهوت

با سری‌های زمانی پرت			
$Garch-E-fc-medoids$	$Garch-E-k-medoids$	$Garch-fc-medoids$	$Garch-k-medoids$
۰.۷۲	۰.۴۴	۰.۶۳	۰.۳۶
بدون سری‌های زمانی پرت			
$Garch-E-fc-medoids$	$Garch-E-k-medoids$	$Garch-fc-medoids$	$Garch-k-medoids$
۰.۸۳	۰.۸۰	۰.۶۴	۰.۴۷

با مشاهدهٔ جدول ۶ می‌توان نتیجه گرفت که هر کدام از روش‌های خوشه‌بندی بدون سری‌های زمانی پرت نسبت به همان روش با داده پرت مقدار سیلهوت بالاتری دارند، از طرفی از آن‌جا که هرچه قدر مقدار ضریب سیلهوت به یک نزدیک باشد خوشه‌بندی بهتری انجام گرفته است، با توجه به جدول ۶ در روش‌های خوشه‌بندی K- مدوید و فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی بدون سری‌های زمانی پرت نسبت به سایر روش‌های خوشه‌بندی مقدار ضریب سیلهوت بالاتری دارند که نشان می‌دهد خوشه‌بندی بر اساس رویکرد نمایی و بدون سری‌های زمانی پرت عملکرد بهتری

نسبت به روش‌های خوشه‌بندی دیگر دارد. مقدار عضویت هر داده به هر خوشه در روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی با سری‌های زمانی پرت در جدول ۷ و روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی بدون سری‌های زمانی پرت در جدول ۸ نشان داده شده است.

جدول ۷: مقدار عضویت

ردیف	نماد سهام	$Garch-E-Fc-medoids \quad with \quad outliers$	
		c_1	c_1
۱	کگل	۰.۰۱	۰.۹۹
۲	کچاد	۰.۹۹	۰.۰۱
۳	پارسان	۰.۵۳	۰.۴۷
۴	ومعادن	۰	۱
۵	اخابر	۰.۰۳	۰.۹۷
۶	خودرو	۰	۱
۷	وانصار	۰	۱
۸	شاراک	۱	۰
۹	سفارس	۰.۹۹	۰.۰۱
۱۰	شسپا	۰.۹۴	۰.۰۶
۱۱	ستران	۰.۱۷	۰.۸۳
۱۲	وخاور	۱	۰
۱۳	لبوتان	۰.۷۵	۰.۲۵
۱۴	بتایر	۰.۰۱	۰.۹۹
۱۵	کرماشا	۰.۵۹	۰.۴۱
۱۶	شرانل	۱	۰
۱۷	کساوه	۰	۱
۱۸	هرمز	۱	۰
۱۹	پکرمان	۰.۰۱	۰.۹۹
۲۰	وبیمه	۰.۵۵	۰.۴۵
۲۱	کیسون	۱	۰
۲۲	رتاپ	۰.۰۱	۰.۹۹
۲۳	وتوکا	۰.۳۷	۰.۶۳
۲۴	وخارزم	۰.۶	۰.۴
۲۵	البرز	۰.۹۹	۰.۰۱
۲۶	واحیا	۱	۰
۲۷	شصفها	۰.۵۹	۰.۴۱
۲۸	آسیا	۰.۰۱	۰.۹۹
۲۹	ساذری	۰	۱
۳۰	کروی	۰	۱

حال با استفاده از فرمول های ۳,۴ در فصل ۴ سری‌های زمانی پرت را در روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی شناسایی کرده و از مجموعه دادها حذف و سپس روش خوشه‌بندی بر مجموعه داده‌های جدید پیاده‌سازی می‌شوند. با توجه به جدول ۸ داده‌های (۲,۳,۵,۱۴,۲۱,۲۵,۲۸) در روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی به عنوان سری‌های زمانی پرت شناسایی شده‌اند و مقدار عضویت صفر به آن‌ها تعلق گرفته است.

جدول ۸: مقدار عضویت

ردیف	نماد سهام	$Garch-E-Fc-medoids \quad with \quad outliers$			
		c_1	c_2	c_3	c_4
۱	کگل	۰	۰	۱	۰
۲	کچاد	–	–	–	–
۳	پارسان	–	–	–	–
۴	ومعادن	۰	۰	۱	۰
۵	اخابر	–	–	–	–
۶	خودرو	۰	۰	۱	۰
۷	وانصار	۰	۰	۱	۰
۸	شاراک	۰.۳۳	۰.۰۱	۰	۰.۶۵
۹	سفارس	۰	۰	۰	۱
۱۰	شسپا	۰.۰۱	۰	۰	۰.۹۹
۱۱	ستران	۰.۰۲	۰.۹۶	۰.۰۲	۰
۱۲	وخاور	۰	۰	۰	۱
۱۳	لبوتان	۰.۰۱	۰.۹۹	۰	۰
۱۴	بتایر	–	–	–	–
۱۵	کرماشا	۰.۲۱	۰.۱۹	۰.۱۸	۰.۴۲
۱۶	شرانل	۱	۰	۰	۰
۱۷	کساوه	۰	۰	۱	۰
۱۸	هرمز	۰.۰۲	۰	۰	۰.۹۸
۱۹	پکرمان	۰	۰	۱	۰
۲۰	وبیمه	۰	۱	۰	۰
۲۱	کیسون	–	–	–	–
۲۲	رتاپ	۰	۰	۱	۰
۲۳	وتوکا	۰	۱	۰	۰

۲۴	وخارزم	۰	۱	۰	۰
۲۵	البرز	–	–	–	–
۲۶	واحیا	۱	۰	۰	۰
۲۷	شصفها	۰	۱	۰	۰
۲۸	آسیا	–	–	–	–
۲۹	ساذری	۰	۰	۱	۰
۳۰	کروی	۰	۰	۱	۰

از طرفی مانند روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصله نمایی، سری‌های زمانی پرت را در تمامی روش‌های خوشه‌بندی مطرح شده در مقاله شناسایی کرده و بعد از حذف سری‌های زمانی پرت از مجموعه داده‌ها دوباره خوشه‌بندی انجام می‌گیرد. در روش خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از فاصله اقلیدسی با سری‌های زمانی پرت داده‌های (۱,۱۸,۱۱,۲۱) و در روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصله اقلیدسی با سری‌های زمانی پرت داده‌های (۱۸,۲۶,۴) مراکز نهایی خوشه‌ها می‌باشند. از طرفی در روش خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از فاصله اقلیدسی بدون سری‌های زمانی پرت داده‌های (۵,۱۸,۲۰) و در روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از فاصله اقلیدسی بدون سری‌های زمانی پرت داده‌های (۱,۱۱,۱۸,۲۱) به عنوان مراکز نهایی خوشه‌ها تعیین می‌شوند. همین‌طور مراکز نهایی خوشه‌ها در روش خوشه‌بندی K- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی با سری‌های زمانی پرت داده‌های (۲۶,۲۳,۷,۱۰) و در روش خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی با سری‌های زمانی پرت داده‌های (۸,۵)و همچنین در روش‌های خوشه‌بندی K- مدوید و فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی بدون سری‌های زمانی پرت داده‌ها (۱۶,۲۰,۳۰,۹) می‌باشند. با توجه به جدول ۵ در خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی با سری زمانی پرت داده‌های (۳,۱۱,۱۳,۱۵,۲۰,۲۳,۲۴,۲۷) با مقدار عضویت مختلف به دو خوشه یا بیشتر متعلق هستند و بقیه دادها با میزان عضویت بالا تنها به یک خوشه تعلق دارند، این در حالی است که با توجه به جدول ۸ در خوشه‌بندی فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی بدون سری زمانی پرت تنها سهام شاراک به دو خوشه‌ی اول و چهارم و سهام کرماشا با مقدار عضویت مختلف به تمامی خوشه‌ها متعلق هستند و بقیه داده‌ها با مقدار عضویت بالا تنها به یک خوشه تعلق دارند. برای ارزیابی عملکرد روش‌های خوشه‌بندی مطرح شده با استفاده از فرمول ۴.۴ و ۴.۵ مقدار شاخص زایبنی به دست آورده می‌شود و کمترین مقدار زایبنی هر کدام از روش‌های خوشه‌بندی در جدول ۹ نشان داده شده است [۴,۵].

جدول ۹: مقدار ضریب زایبنی

با سری‌های زمانی پرت			
$Garch-E-fc-medoids$	$Garch-E-k-medoids$	$Garch-fc-medoids$	$Garch-k-medoids$
۰.۱۵	۰.۰۸	۰.۴۰	۰.۰۴
بدون سری‌های زمانی پرت			
$Garch-E-fc-medoids$	$Garch-E-k-medoids$	$Garch-fc-medoids$	$Garch-k-medoids$
۰.۰۸	۰.۰۴	۰.۳۵	۰.۰۳

با مشاهده جدول ۹ می‌توان نتیجه گرفت که هر کدام از روش‌های خوشه‌بندی بدون سری‌های زمانی پرت نسبت به همان روش با سری‌های زمانی پرت مقدار زایبنی کمتری دارند، از آن جا که هرچه قدر شاخص زایبنی کوچکتر باشد خوشه‌بندی بهتری انجام گرفته است، با توجه به جدول ۹ روش‌های خوشه‌بندی K- مدوید و فازی C- مدوید بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی بدون سری‌های زمانی پرت نسبت به سایر روش‌های خوشه‌بندی مطرح شده مقدار عملکرد بهتری دارند.

۷- بحث و نتیجه‌گیری

در این مقاله روش‌های خوشه‌بندی K- مدوید و فازی C- مدوید بر اساس فاصلهٔ اقلیدسی و رویکرد فاصله‌ای نمایی با سری‌های زمانی پرت و بدون سری‌های زمانی پرت معرفی شد. از آن جا که حذف سری‌های زمانی پرت از خوشه‌ها باعث بهبود دقت خوشه‌بندی می‌شود، روش جدید شناسایی سری‌های زمانی پرت بر اساس مدل گارچ با استفاده از رویکرد فاصله‌ای نمایی ارائه شد که در سه مرحله انجام می‌گیرد ابتدا به پیاده سازی روش‌های خوشه‌بندی مطرح شده و سپس به تشخیص و حذف سری‌های زمانی پرت و پس از آن مجدأ روش‌های خوشه‌بندی بر مجموعه داده‌ها بدون سری‌های زمانی پرت اعمال می‌شود. از طرفی با استفاده از شاخص‌های ارزیابی (سیلهوت و زایبنی) به تجزیه و تحلیل روش‌های خوشه‌بندی K- مدوید و فازی C- مدوید بر اساس فاصله اقلیدسی و رویکرد فاصله‌ای نمایی با سری‌های زمانی پرت و بدون سری‌های زمانی پرت پرداخته شد و با توجه به نتایج به دست آمده در بخش شبیه‌سازی مطالعاتی و مطالعات تجربی نشان داده شد که روش خوشبندی بدون سری‌های زمانی پرت و با استفاده از رویکرد فاصله‌ای نمایی از دقت بالاتری برخوردار است.

References

۱. ژیاوی هان.، میشلن کامبر، ژان پی، داده کاوی مفاهیم و تکنیک‌ها، دکتر مهدی اسماعیلی، نیاز دانش، تهران، (۱۳۹۶).

2. Ahmed, M., Mahmood, A., Islam, M., ‘‘A survey of anomaly detection techniques in financial domain’’, Future Generation Computer Systems, 55 (2016) 278–288.

3. Caiado, J., Crato, N. , "A GARCH-based method for clustering of financial time series: International stock markets evidence", Recent Advances in Stochastic Modeling and Data Analysis, World Scientific Publishing, New Jersey, (2007) 542–551.

4. Campello, R.J.G.B., Hruschka. E.R., "A fuzzy extension of the silhouette width criterion for cluster analysis", Fuzzy Sets and Systems, 157 (2006) 2858 – 2875.

5. Desgraupes, B., "Clustering Indices", University of Paris Ouest - Lab Modal'X, 1 (2017) 1-34.

6. D'Urso, P., Cappelli, C., Di Lallo, D., Massari, R., "Clustering of financial time series", Physica A, 392 (2013) 2114-2129.

7. D'Urso, P., DeGiovanni, L., Massari, R., "GARCH-based robust clustering of time series", Fuzzy Sets and Systems, 305 (2016) 1-28.

8. Gan, G., Kwok-Po Ng, M., "k -means clustering with outlier removal", Pattern Recognition Letters, 90 (2017) 8-14.

9. Gosain, A., Dahiya, S., "Performance Analysis of Various Fuzzy Clustering Algorithms:A Review", 7th International Conference on Communication, Computing and Virtualization 2016, 79 (2016) 100-111.

10. Hautamaki, V., Cherednichenko, S., Karkkainen, I., Kinnunen, T., Franti, P., "Improving K-Means by Outlier Removal", The 14th In Scandinavian Conference on Image Analysis, 3540 (2005) 978-987.

11. Otranto, E,. "Clustering heteroskedastic time series by model-based procedures", Computational Statistics & Data Analysis, Elsevier, 52(2008) 4685-4698.

12. Prabhjot, K., I. M. S, L., Anjana, G., "DOFCM: A Robust Clustering Technique Based upon Density", IACSIT International Journal of Engineering and Technology, 3 (2011) 297-303.

13. TSAY, R.S., Analysis of financial time series, John Wiley & Sons: New York, (2002).

14.Wu, K.-L., Yang, M.-S., "Alternative c-means clustering algorithms", Pattern Recognition, 35(2002) 2267–2278.