E-ISSN: 2676-6159 | P-ISSN: 2538-5079

Journal of Decisions and Operations Research

www.journal-dmor.ir

J. Decis. Oper. Res. Vol. 8, No. 2 (2023) 507–526.

Paper Type: Original-Application Paper

Developing a Hybrid Model for Comparative Analysis of Financial Data Clustering Algorithms

Mojtaba Movahedi[1], Mahdi Homayounfar[1,*], Mehdi Fadaei[1], Mansour Soufi[1]

[1] Department of Industrial Management, Faculty of Management and Accounting, Rasht Branch, Islamic Azad University, Rasht, Iran; m.movahedi62@gmail.com; homayounfar@iaurasht.ac.ir; fadaei@iaurasht.ac.ir; msoufi@iaurasht.ac.ir.

Citation:

Movahedi, M., Homayounfar, M., Fadaei, M., & Soufi, M. (2023). Developing a hybrid model for comparative analysis of financial data clustering algorithms. *Journal of decisions and operations research, 8*(2), 507-526.

Received: 27/07/2021 Reviewed: 28/08/2021 Revised: 10/10/2021 Accepted: 25/11/2021

Abstract

Purpose: Clustering algorithms are useful tools for understanding data structure and classifying them into different data sets. Due to the importance of using these algorithms in analyzing financial market data that have a high volume and scope, this study in order to select the best clustering algorithm for clustering companies listed on the Tehran Stock Exchange in the field of finance from It has used different clustering algorithms and evaluated the validity of these algorithms and selected the best algorithm.

Methodology: This research is applied in terms of purpose and descriptive in terms of implementation method and is of quantitative type (mathematical modeling). The statistical population of the research includes 403 companies listed on the Tehran Stock Exchange in 2019, whose performance has been evaluated based on four financial criteria.

Findings: After clustering the surveyed companies by five clustering algorithms, namely K-means, EM, COBWEB, density-based algorithm and ward method, seven indicators RS, DB, Dun, SD, Purity, Entropy and Time were used to evaluate the algorithms. Finally, the total performance of the algorithms was analyzed based on TOPSIS, VICOR and DEA methods. Based on the results, K-means has a better performance in clustering based on the financial data sets.

Originality/Value: Since no clustering algorithm can have the best performance in all measurements for each data set, this study uses a combination of multiple criteria to analyze data clustering algorithms related to the field of financial performance appraisal. Companies have provided suggestions and the results of this study have been used effectively for investors in the field of finance, which leads to the optimal choice of investment portfolio.

Keywords: Clustering, Multi-criteria decision making, Financial performance evaluation.

Corresponding Author: homayounfar@iaurasht.ac.ir http://dorl.net/dor/20.1001.1.25385097.1402.8.2.13.7

Journal of Decisions and Operations Research

تصمیم‌گیری و تحقیق در عملیات

دوره ۸، شماره (۲)، (۱۴۰۲)، ۵۲۶-۵۰۷

www.journal-dmor.ir

نوع مقاله: پژوهشی

ارایه یک مدل ترکیبی به‌منظور تحلیل تطبیقی الگوریتم‌های خوشه‌بندی داده‌های مالی

مجتبی موحدی[۱]، مهدی همایون‌فر[۱،*]، مهدی فدایی[۱]، منصور صوفی[۱]

[۱]گروه مدیریت صنعتی، دانشکده مدیریت و حسابداری، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران.

چکیده

هدف: الگوریتم‌های خوشه‌بندی، ابزارهای مفیدی برای درک ساختار داده‌ها و طبقه‌بندی آن‌ها در مجموعه داده‌های مختلف می‌باشند. باتوجه به اهمیت به‌کارگیری این الگوریتم‌ها در تحلیل داده‌های بازارهای مالی که از حجم و گستردگی بالایی برخوردارند، این پژوهش به‌منظور انتخاب بهترین الگوریتم خوشه‌بندی برای خوشه‌بندی شرکت‌های حاضر در بورس اوراق بهادار تهران در حوزه مالی از الگوریتم های خوشه‌بندی مختلف استفاده نموده و به ارزیابی اعتبار این الگوریتم‌ها و انتخاب بهترین الگوریتم پرداخته است.

روش‌شناسی پژوهش: این پژوهش از نظر هدف، کاربردی و از نظر روش اجرا توصیفی و از نوع کمی (مدل‌سازی ریاضی) است. جامعه آماری تحقیق شامل ۴۰۳ شرکت حاضر در بورس اوراق بهادار تهران در سال ۹۸ است که عملکرد آن‌ها براساس چهار معیار مالی ارزیابی شده است.

یافته‌ها: پس از خوشه‌بندی شرکت‌های موردبررسی توسط پنج الگوریتم خوشه‌بندی *K-Means*، *EM*، *COBWEB*، الگوریتم مبتنی‌بر چگالی و روش وارد، از هفت شاخص *RS*، *DB*، دان، *SD*، خلوص، آنتروپی و زمان برای ارزیابی الگوریتم‌های خوشه‌بندی استفاده گردید. در نهایت، عملکرد نهایی الگوریتم‌های به‌کار رفته براساس روش‌های تاپسیس، ویکور و تحلیل پوششی داده‌ها مورد تجزیه‌وتحلیل قرار گرفت. براساس نتایج، روش *K-Means* از عملکرد بهتری در خوشه‌بندی شرکت‌ها براساس مجموعه داده‌های مالی برخوردار است.

اصالت/ارزش افزوده علمی: از آن‌جایی‌که هیچ الگوریتم خوشه‌بندی نمی‌تواند بهترین عملکرد را در تمام اندازه‌گیری‌ها برای هر مجموعه داده داشته باشد، این پژوهش ضمن به‌کارگیری ترکیبی از معیارهای چندگانه به‌منظور تجزیه‌وتحلیل الگوریتم‌های خوشه‌بندی داده‌های مربوط به حوزه ارزیابی عملکرد مالی شرکت‌ها، به ارایه پیشنهاداتی پرداخته و نتایج این پژوهش برای سرمایه‌گذاران حوزه مالی کاربرد موثر داشته که منجر به انتخاب بهینه سبد سرمایه‌گذاری می‌شود.

کلیدواژه‌ها: ارزیابی عملکرد مالی، تصمیم‌گیری چندمعیاره، خوشه‌بندی.

۱- مقدمه

در عصر داده‌های بزرگ، حجم زیادی از داده‌های کلان توسط صنایع مختلف تولید می‌شوند که یکی از ویژگی‌های اصلی آن بالا بودن بعد آن‌ها است. چگونگی کار با این داده‌ها به‌گونه‌ای که ارزش بالقوه آن‌ها از دست نرود و بتوان مدلی مناسب خلق نمود یک مشکل چالش‌برانگیز است [1]. امروزه از راه‌حل‌های مبتنی‌بر تجزیه‌وتحلیل داده‌های بزرگ و محاسبات هوشمند برای کاهش پیچیدگی پردازش مقادیر زیادی از داده‌ها استفاده می‌شود [2]. تجزیه‌وتحلیل داده‌های بزرگ مزایای زیادی به‌دنبال خواهد داشت که هدف از آن‌ها بهبود شفافیت در تصمیم‌گیری است. چراکه تصمیم‌گیری مبتنی‌بر تجزیه‌وتحلیل این‌گونه داده‌ها، عملکرد کل سیستم را باتوجه به ساختار داخلی

* نویسنده مسئول

homayounfar@iaurasht.ac.ir

http://dorl.net/dor/20.1001.1.25385097.1402.8.2.13.7

به حداکثر می‌رساند [1]. پنج ویژگی اصلی داده‌های بزرگ از نظر لی و همکاران [1] عبارت‌اند از: ۱- حجم[1] زیاد داده‌ها (ظرفیت بالا)، ۲- سرعت[2] بالا (داده‌ها با سرعت بالا تولید و به‌روز می‌شوند)، ۳- تنوع[3] (داده‌های تولیدشده توسط منابع مختلف در اشکال مختلف ظاهر می‌شوند.)، ٤- دقت[4] (صحت) بالا و ٥- ارزش[5] بالا (پتانسیل بالقوه پنهان‌شده در داده‌ها).

داده‌های مالی بخشی جدایی‌ناپذیر از اقتصاد است که وضعیت اقتصادی فعلی و آینده را منعکس می‌کند. به‌عنوان‌مثال قیمت سهام یک شرکت منعکس‌کننده فعالیت آن و یک مرجع مهم سرمایه‌گذاری برای سرمایه‌گذاران است. این داده‌ها وضعیت کلی بازار را نشان می‌دهند و برای درک بهتر وضعیت اقتصادی می‌بایست مورد تجزیه‌وتحلیل قرار گیرند [3]. داده‌های مالی معمولا حاوی اطلاعات مهم و دارای ابعاد بزرگ هستند، حجم بزرگ داده‌ها به‌تنهایی به مدیران سازمان‌ها و سرمایه‌گذاران در تصمیم‌گیری کمکی نمی‌کند، بلکه باعث سردرگمی آنان نیز می‌شود؛ بنابراین، مدیریت داده‌های خام و تبدیل داده‌های خارجی و داخلی سازمان به اطلاعات و دانش با استفاده از روش‌های گوناگون، نقش اساسی و محوری دارد [4]. داده‌کاوی به‌عنوان یک توانایی پیشرفته، در تحلیل داده و کشف دانش مورداستفاده قرار می‌گیرد. برای این کار روش‌های متعددی وجود دارد که هر یک از آن‌ها برای اهداف خاصی مورداستفاده قرار می‌گیرند. یکی از مهم‌ترین روش‌های داده‌کاوی، خوشه‌بندی است که کاربرد بسیاری در کشف دانش دارد. خوشه‌بندی ما را قادر می‌سازد تا به‌جای پردازش حجم انبوه اطلاعات، تنها اطلاعاتی را بررسی و تحلیل کنیم که بسیار به هم شبیه هستند و این خود یک گام بلند برای ساده کردن مساله است [5]. اگرچه تحلیل خوشه‌ای اولین بار توسط ترایون استفاده شد [6]، اما امروزه الگوریتم‌های مختلفی ازجمله الگوریتم خوشه‌بندی تفکیکی، سلسله مراتبی، مبتنی‌بر مبنای چگالی و مبتنی‌بر مبنای مدل، برای خوشه‌بندی موجود است. درک ویژگی‌های این روش‌ها و نقاط قوت و ضعف آن‌ها می تواند در تصمیم‌گیری برای انتخاب روش مناسب سودمند باشد. بعضی از روش‌های خوشه‌بندی فقط در مجموعه داده با حجم کم نتیجه‌بخش هستند، دسته‌ای از توابع خوشه‌بندی فقط قادر به استخراج خوشه‌ها با شکل محدب بوده ولی بعضی دیگر، خوشه با هر نوع شکل را استخراج می‌کنند. برخی توابع خوشه‌بندی نیازمند تعیین پارامتر اولیه مثل تعداد خوشه‌ها از طرف کاربر هستند و دسته دیگر به این پارامترها نیاز ندارند. از این‌رو، انتخاب الگوریتم خوشه‌بندی مناسب برای داده‌های مالی چالش‌برانگیز بوده و مقایسه و ارزیابی این الگوریتم‌ها از اهمیت بالایی برخوردار است.

حجم بالای داده‌ها و معیارهای ارزیابی مالی در دنیای امروز حاصل پیشرفت چشم‌گیر تکنولوژی است؛ بنابراین، دغدغه‌ی امروز سرمایه‌گذاران و تصمیم‌گیرندگان حوزه‌های مالی کمبود داده‌ها و معیارها ارزیابی نیست. بلکه استفاده درست از داده‌های موجود و تبدیل آن‌ها به اطلاعات و سپس دانش برای اتخاذ درست‌ترین و کم‌خطاترین تصمیم می‌باشد. تکنیک‌های داده‌کاوی و روش‌های تصمیم‌گیری چندمعیاره همواره کمک‌رسان مدیران در تصمیم‌گیری بوده است. اما گستردگی و نتایج منحصربه‌فرد هر یک از تکنیک‌های داده‌کاوی و روش‌های تصمیم‌گیری چندمعیاره بر لزوم توجه و بررسی نقاط قوت و ضعف آن‌ها تاکید می‌نماید.

از این‌رو همواره محققان به‌دنبال مدل‌هایی هستند که بهترین عملکرد را داشته باشد؛ چراکه استفاده از یک الگوریتم خاص خوشه‌بندی برای داده‌های مختلف، سوال‌هایی را مطرح می‌نماید که پاسخ به آن‌ها از اهمیت بالایی برخوردار است. این سوالات عبارت‌اند از آیا الگوریتم خوشه‌بندی بهتری برای این نوع داده‌ها وجود ندارد؟ آیا خوشه‌بندی داده‌ها به‌درستی انجام‌شده است؟ الگوریتم خوشه‌بندی استفاده‌شده برای این نوع داده کارا است؟ بهترین معیار برای ارزیابی عملکرد الگوریتم‌های خوشه‌بندی کدام است؟ آیا معیارهای مختلف ارزیابی الگوریتم‌های خوشه‌بندی، نتایج یکسانی را ارایه می‌نمایند؟ به‌منظور پاسخ به این سوال‌های طرح‌شده، پژوهش حاضر به‌دنبال آن است که ضمن بررسی تطبیقی الگوریتم‌های خوشه‌بندی بر مبنای داده‌های مالی، به ارزیابی این الگوریتم‌ها و انتخاب الگوریتم مناسب جهت تحلیل‌های سرمایه‌گذاری بپردازد. به‌این‌صورت که با پیاده‌سازی پنج الگوریتم پرکاربرد خوشه‌بندی بر روی مجموعه داده‌های مالی واقعی شرکت‌های حاضر در بورس اوراق بهادار، عملکرد هر یک از این الگوریتم‌های خوشه‌بندی با استفاده از هفت شاخص اعتبارسنجی، ارزیابی‌شده و نهایتا با استفاده از تکنیک‌های شناخته‌شده *MCDM* و تحلیل پوششی داده‌ها[6]، بر مبنای نقطه نظرات مختلف بهترین روش خوشه‌بندی برای این نوع داده‌ها انتخاب شود.

[1] Volume
[2] Velocity
[3] Variety
[4] Veracity
[5] Value
[6] Data Envelopment Analysis (DEA)

در رابطه با نوآوری تحقیق حاضر باید اشاره نمود که: ۱- باتوجه به فرآیند تحقیق، در فاز اول تحقیق مرور جامعی از شاخص‌های مالی ارزیابی عملکرد شرکت‌ها صورت گرفته و در ادامه با طراحی پرسشنامه‌های اهمیت‌سنجی شاخص‌های اولیه و پیاده‌سازی روش دلفی فازی به غربال‌سازی اولیه و ثانویه شاخص‌های شناسایی‌شده پرداخته شد. سپس با استفاده از نقشه نگاشت فازی، ضمن بررسی روابط بازخوردی شاخص‌های غربال‌شده، مهم‌ترین شاخص‌های ارزیابی شرکت‌های بورس مشخص گردیدند که از این حیث، مطالعه‌ای که با این جامعیت به بررسی معیارهای مالی بپردازد، مشاهده نشد، ۲- از سوی دیگر، استفاده از الگوریتم‌های خوشه‌بندی مختلف و به‌کارگیری شاخص‌های اساسی اعتبارسنجی (از دیدگاه خبرگان) در ارزیابی آن‌ها، حاکی از جامعیت تحقیق می‌باشد که مشابه آن در تحقیقات داخلی مشاهده نشده است و ۳- درنهایت از سه روش قدرتمند ویکور، تاپسیس و تحلیل پوششی داده‌ها در جمع‌بندی نتایج رتبه‌بندی الگوریتم‌ها و انتخاب بهترین الگوریتم برای خوشه‌بندی داده‌ها استفاده شد که همراه با موارد ۱ و ۲، در تبیین اهمیت این تحقیق نقش دارد.

نتایج این پژوهش می‌تواند برای شرکت‌های سرمایه‌گذاری، مدیران ارشد، تحلیلگران مالی، صندوق‌های سرمایه‌گذاری و پژوهشگران در حوزه مالی و به‌طورکلی سرمایه‌گذاران، کاربرد موثری داشته باشد. آن‌ها زمانی‌که می‌خواهند عملکرد شرکت‌ها را موردسنجش قرار دهند، می‌توانند از مدل پیشنهادی در جهت انتخاب بهترین روش خوشه‌بندی شرکت‌ها که نهایتا منجر به انتخاب سبد سرمایه‌گذاری مطلوب می‌شود، استفاده نمایند.

۲- مبانی نظری و پیشینه پژوهش

۲-۱- خوشه‌بندی

روش‌های یادگیری بدون نظارت و نظارت‌شده دو تکنیک اصلی هستند که در تحلیل مالی مورداستفاده قرار می‌گیرند [7]. خوشه‌بندی یکی از شاخه‌های مهم شناسایی الگوی بدون نظارت و یادگیری ماشین است که به‌طور گسترده‌ای در زمینه‌های مختلف ازجمله تحلیل مالی، تشخیص پزشکی، برش تصویر و ترکیب اطلاعات و ... مورداستفاده قرار گرفته است [8]. هدف آن یافتن گروه‌ها یا خوشه‌هایی از اشیایی است که مشابه یکدیگر در یک خوشه هستند اما با اشیا در هر خوشه دیگری متفاوت هستند [9]. درزمینه مالی خوشه‌بندی، ابزاری مفید برای درک و تشخیص ساختار، طبقه‌بندی و سلسله‌مراتب در بازارهای مالی است و اطلاعات طبقه‌بندی‌شده و مفیدی از چگونگی ارتباط سهم‌ها، توزیع و اثرات ارتباطی بین خوشه‌های حاصل از سهم‌های مرتبط با هر صنعت را فراهم می‌نمایند [10]. انواع زیادی از الگوریتم‌های خوشه‌بندی وجود دارد که می‌تواند به‌طورتقریبی آن‌ها را به دو گروه خوشه‌بندی سخت و خوشه‌بندی نرم طبقه‌بندی نمود، خوشه‌بندی سخت و نرم هر دو از یک مجموعه C_i برای به تصویر کشیدن خوشه‌ها استفاده می‌نمایند با این تفاوت که خوشه‌بندی سخت هم‌پوشانی بین خوشه‌ها وجود ندارد $(C_i \cap C_j = \emptyset)$ اما خوشه‌بندی‌نرم این‌گونه نیست، $(C_i \cap C_j) \neq \emptyset$ [9]. به‌طورکلی رویکردهای خوشه‌بندی عبارت‌اند از افرازی[1]، سلسله مراتبی[2]، مبتنی‌بر چگالی[3]، مبتنی‌بر مشبک‌کردن فضا، نقشه‌های خودسازمانده[4] و متاهیورستیک[5] [11].

جدول ۱- رویکردهای خوشه‌بندی.

Table 1- Clustering approaches.

ردیف	رویکرد	تعریف
1	افرازی یا تفکیکی	در این روش، براساس n مشاهده و k گروه، عملیات خوشه‌بندی انجام می‌شود. به‌این‌ترتیب تعداد خوشه‌ها یا گروه‌ها از قبل در این الگوریتم مشخص است. با طی مراحل خوشه‌بندی افرازی، هر شی فقط و فقط به یک خوشه تعلق خواهد داشت و هیچ خوشه‌ای بدون عضو باقی نمی‌ماند. دو الگوریتم مهم این روش عبارت‌اند از: K-Means و K-medoids.
2	سلسله مراتبی	برعکس خوشه‌بندی افرازی که اشیا را در گروه‌های مجزا تقسیم می‌کند، در این نوع خوشه‌بندی، هر سطح از فاصله، نتیجه خوشه‌بندی را نشان می‌دهد. این سطوح به‌صورت سلسله مراتبی هستند که برای نمایش نتایج خوشه‌بندی به‌صورت سلسله مراتبی از درخت‌واره[6] استفاده می‌شود.

[1] Partitional clustering
[2] Hierarchical clustering
[3] Density-bases clustering
[4] Self-organizing map
[5] Heuristic-meta
[6] Dendrogram

جدول ۱- ادامه.

Table 1- Continued.

ردیف	رویکرد	تعریف
3	مبتنی‌بر چگالی	بسیاری از روش‌های افرازی اشیا را براساس فاصله آن‌ها نسبت به یکدیگر خوشه‌بندی می‌کنند. برخی روش تنها خوشه‌های کروی شکل را پیدا می‌کنند و در برابر خوشه‌هایی به شکل‌های دلخواه با مشکل مواجه می‌شوند، در الگوریتم‌ها خوشه‌بندی برمبنای چگالی، نقاط با تراکم زیاد شناسایی شده و در یک خوشه قرار می‌گیرند. از الگوریتم‌های معروف در این زمینه می‌توان به DBSCAN اشاره کرد.
4	مبتنی‌بر مشبک کردن فضا	روش مشبک‌سازی فضا به سلول‌های مختلف امکان کار بر روی اطلاعات با درجه تفکیک شفافیت‌های متفاوت را فراهم می‌کند. در این روش ابتدا فضا به سلول‌هایی تقسیم شده و سپس عملیات خوشه‌بندی روی این سلول‌ها انجام می‌گیرد. مهم‌ترین مزیت این روش افزایش سرعت است.
5	نقشه‌های خودسازمانده	یک روش تجسم داده است که توسط پروفسور تو و کوهونن اختراع شده است. در این روش ابعاد داده‌ها را از طریق کاربرد شبکه عصبی خود سازمان دهنده کاهش می‌دهند. نقشه‌های خودسازمانده[1] روشی برای خوشه‌بندی و پیش پرداز اطلاعات می‌باشد.
6	متاهیورستیک	برای مسایلی با ابعاد بزرگ، روش سیمپلکس[2] از کارایی بسیار خوبی برخوردار است، ولی روش شاخه و کرانه کارایی خود را از دست می‌دهد و عملکرد بهتری از شمارش کامل نخواهد داشت. به دلایل فوق، اخیرا تمرکز بیش‌تری بر روش‌های ابتکاری[3] یا فرا ابتکاری یا جست‌وجوی تصادفی[4] صورت گرفته است. روش‌های جست‌وجوی ابتکاری، روش‌هایی هستند که می‌توانند جوابی خوب (نزدیک به بهینه) درزمانی محدود برای یک مساله ارایه کنند. روش‌های جست‌وجوی ابتکاری عمدتا بر مبنای روش‌های شمارشی می‌باشند، با این تفاوت که از اطلاعات اضافی برای هدایت جستجو استفاده می‌کنند. این روش‌ها از نظر حوزه کاربرد، کاملا عمومی هستند و می‌توانند مسایل خیلی پیچیده را حل کنند. عمده این روش‌ها، تصادفی بوده و از طبیعت الهام گرفته شده‌اند.

۲-۲- الگوریتم‌های خوشه‌بندی

براساس مبانی نظری پژوهش به‌منظور بررسی میدانی مدل پیشنهادی از پنج الگوریتم خوشه‌بندی *K-Means*، ماکزیمم امید ریاضی[5]، *COBWEB*، روش مبتنی‌بر چگالی[6] و روش وارد[7] استفاده می‌شود که در پاراگراف‌های زیر به شرح هر یک از آن‌ها پرداخته شده است:

K-Means: این الگوریتم پارامتر k را به‌عنوان ورودی گرفته و مجموعه n شی را به k خوشه افراز می‌کند [12]. علی‌رغم سادگی یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر محسوب می‌شود. برای این الگوریتم شکل‌های مختلفی بیان شده است که روند کلی آن‌ها به‌این‌صورت است که ابتدا نقاطی به‌عنوان مراکز خوشه‌ها به‌دست می آید. این نقاط درواقع همان میانگین نقاط متعلق به هر خوشه هستند. سپس، هر نمونه داده به یک خوشه که آن داده کم‌ترین فاصله تا مرکز آن خوشه را دارا باشد، نسبت داده می‌شود. بهترین خوشه‌بندی آن است که مجموع تشابه بین مرکز خوشه و همه اعضای خوشه را حداکثر و مجموع تشابه بین مراکز خوشه‌ها را حداقل کند [13].

ماکزیمم امید ریاضی: این الگوریتم شامل دو مرحله امید ریاضی و ماکزیمم‌سازی است. در مرحله امید ریاضی، احتمال قرارگیری هر شی در خوشه محاسبه می‌شود و در مرحله ماکزیمم‌سازی پارامترهای توزیع داده‌ها محاسبه شده به‌گونه‌ای‌که تابع درست‌نمایی[8] را به حداکثر مقدار خود می‌رساند. فرمول محاسبه این الگوریتم به‌صورت زیر تعریف شود [14]:

$$Q\ \theta|\theta^t) = E_{\left(Z|X.\ \theta^T\right)}[Log\ L(\theta; X. Z)]\ and\ \theta^{(t+1)} = argmax_\theta Q\ \theta|\theta^t). \qquad (۱)$$

COBWEB: یک الگوریتم، خوشه‌بندی مفهومی است که توسط فیشر [15] معرفی شد و توسط جناری و همکاران [16] توسعه یافت. در این روش اشیا به‌صورت یک درخت طبقه‌بندی سازمان‌دهی می‌شوند که در آن هر گره دارای یک توصیف احتمالاتی از یک مفهوم است.

[1] Self-Organizing Map (SOM)
[2] Simplex
[3] Heuristic
[4] Random method
[5] Expectation–Maximization (EM)
[6] Density-based method
[7] Ward method
[8] Likelihood

این درخت طبقه‌بندی با استفاده از چهار عمل اصلی شامل قرار دادن یک شی در یک کلاس موجود، ایجاد یک گره/کلاس جدید، ادغام دو گره و تقسیم یک گره، ساخته می‌شود [14].

الگوریتم مبتنی بر چگالی: این الگوریتم یکی از روش‌های اصلی برای خوشه‌بندی در داده‌کاوی است که عدم محدودیت به شکل خوشه‌ها، ساده و قابل‌فهم بودن ازجمله مزایای این الگوریتم است [17].

الگوریتم وارد: در این روش خوشه‌بندی، برای کاهش تلفات ناشی از داده‌های دورافتاده از معیار جدیدی برای محاسبه عدم شباهت بین خوشه‌ها استفاده می‌شود؛ بدین منظور از مجموع مربعات تفاضل هر داده از یک خوشه با بردار میانگین آن خوشه، به‌عنوان معیاری برای سنجش یک خوشه استفاده می‌شود. الگوریتم مربوطه بدین‌صورت است که ابتدا هر داده به‌عنوان یک خوشه درنظر گرفته شده و به ازای تمام جفت خوشه‌های ممکن از مجموعه خوشه‌ها، آن دو خوشه‌ای که مجموع مربعات تفاضل داده‌های خوشه حاصل از اجتماع آن‌ها با بردار میانگین خوشه حاصل، کمینه باشد، انتخاب می‌شود. دو خوشه انتخاب‌شده باهم ترکیب می‌شوند و تا زمانی‌که تعداد خوشه‌ها به تعداد موردنظر نرسیده است، مراحل تکرار می‌گردد [18].

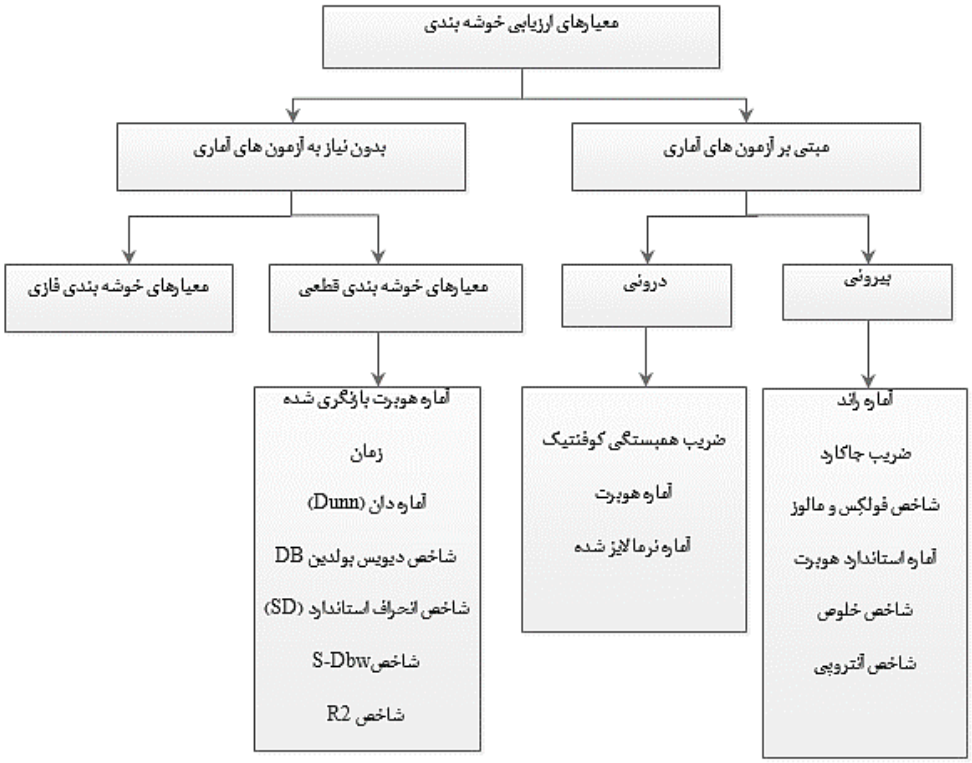
۳-۲- شاخص‌های ارزیابی خوشه‌بندی

همان‌گونه که پیش‌تر اشاره شد خوشه‌بندی داده‌ها فرآیندی بدون ناظر است؛ بدین معنا که کاربر در این فرآیند دخالت نمی‌کند. تعداد طبقات از پیش تعیین‌شده یا مثال‌هایی وجود ندارد که نشان دهد نتایج به‌دست‌آمده، از اعتبار برخوردارند یا خیر؛ بنابراین، می‌بایست به‌دنبال شاخص‌هایی باشیم که اعتبار خوشه‌بندی‌های انجام‌شده را موردسنجش قرار دهد. از این‌رو، ارزیابی نتایج یکی از مهم‌ترین موضوعات در تجزیه‌وتحلیل خوشه‌ها است. شاخص‌های اعتبار خوشه‌بندی[1] کیفیت نتایج خوشه‌بندی را اندازه‌گیری می‌کنند و به دو دسته درونی و بیرونی تقسیم می‌شوند. هدف معیارهای درونی ارزیابی ساختار خوشه‌بندی است که آن ساختار با روشی برای مجموعه داده‌های کمی تولید شده است؛ اما معیارهای بیرونی ساختار به‌دست‌آمده از یک روش خوشه‌بندی، با ساختار از پیش تعیین‌شده‌ای که بر پایه شهودی است مقایسه می‌شود [19]. این شاخص‌ها اغلب از دو معیار فشردگی و جدایش برای ارزیابی خوشه‌بندی انجام‌شده استفاده می‌کنند [20].

این شاخص‌ها سعی در محاسبه فشردگی و جدایش بین خوشه‌ها و در برخی موارد هم‌پوشانی آن‌ها و ساخت ترکیب مناسبی از آن‌ها برای پیداکردن مناسب‌ترین خوشه‌بندی دارند. بسیاری از شاخص‌های اعتبار خوشه‌بندی از تمام اطلاعات موجود در مورد شکل خوشه استفاده نمی‌کنند. همین موضوع باعث می‌شود که آن‌ها در برخی موارد، نتیجه درست را ارایه ندهند و یا برخی از آن‌ها نمی‌توانند در مجموعه داده‌های دارای اغتشاش[2]، تعداد مناسب خوشه‌ها را تشخیص دهند. برخی دیگر از آن‌ها نیز تمام خصوصیات موجود در خوشه‌ها را درنظر نمی‌گیرند [21]. شاخص‌های ارزیابی اعتبار خوشه‌بندی در **شکل ۱** قابل‌مشاهده است.

جهت اعتبارسنجی جامع الگوریتم‌های خوشه‌بندی، شاخص‌های متعددی در مبانی نظری موجود است **(شکل ۱)**. در این پژوهش باتوجه به کسب‌نظر خبرگان دانشگاهی و براساس تکنیک سوآرا[3] هفت شاخص، ضریب تعیین[4]، دیویس بولدین[5]، دان[6]، انحراف استاندارد[7]، خلوص[8]، آنتروپی[9] و زمان[10] مورداستفاده قرار گرفته‌اند که در ادامه به شرح آن‌ها پرداخته شده است.

[1] Clustering Validity Indices (CVIs)
[2] Noise
[3] Stepwise Weight Assessment Ratio Analysis (SWARA)
[4] R-Squared index
[5] Davies-Bouldin index (DB)
[6] Dunn’s index
[7] Standard Deviation (SD)
[8] Purity
[9] Entropy
[10] Time



شکل ۱- شاخص‌های ارزیابی اعتبار خوشه‌بندی [22].

Figure 1- Criteria for evaluating the validity of clustering [22].

شاخص *RS*: همان نسبت مجموع توان دوم انحرافات بین گروه‌ها (SS_b) به مجموع توان دوم انحرافات کل داده‌ها (SS_t) است. به این شاخص، ضریب تعیین یا ضریب تشخیص نیز می‌گویند، دامنه مقادیر بین ۰ و ۱ است. هر چقدر این نسبت به یک نزدیک‌تر باشد، بیان‌گر این است که پراکندگی داده‌ها داخل یک خوشه کم و فاصله بین خوشه‌ها زیاد است [22].

$$RS = \frac{SS_b}{SS_t} = \frac{SS_t - SS_w}{SS_t}, \qquad (۲)$$

که C_j بیان‌گر خوشه j، d تعداد شاخص‌ها (متغیرها) و n نشان‌دهنده تعداد اشیا داده‌ها باشد، در این صورت:

$$SS_w = \sum_{i=1}^{k} \sum_{x \in C_j} \sum_{j=1}^{d} (x_{ij} - x_{ij})^2. \qquad (۳)$$

$$SS_w = \sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - x_j)^2, \qquad (۴)$$

که در آن x_{ij} مقدار شی i از نظر شاخص j، x_j میانگین داده‌های شاخص j و x_{ij} میانگین شاخص j در خوشه i است.

شاخص *DB*: شاخص اعتباری است که به تعداد خوشه‌ها و الگوریتم‌های خوشه‌بندی وابستگی ندارد، مقدار این شاخص هر مقدار کوچک باشد نشان از بهینه‌بودن روش خوشه‌بندی است و با استفاده از **رابطه (۵)** محاسبه می‌شود [22].

$$DB = \frac{1}{k} \sum_{i=1}^{k} R_i, \quad R_i = \max_{i \neq j} R_{ij}, \quad R_{ij} = \frac{(S_i + S_j)}{D_{ij}}. \qquad (۵)$$

$$S_i = \sqrt{\frac{1}{|C_i|} \sum_{x \in C_i} d^2\ x. c_i)}, \qquad (۶)$$

که $|C_i|$ تعداد اعضای خوشه i، D_{ij} فاصله بین مرکز دو خوشه i و j و $d^2\ x. c_i)$ توان دوم فاصله بین دو نقطه x و مرکز خوشه i است.

شاخص دان: هدف این شاخص این است که مشخص کند خوشه‌ها فشرده و مجزا[1] هستند. هرچقدر مقدار این شاخص بیش‌تر باشد اعضای خوشه‌ها فشرده‌تر و خوشه‌ها مجزاتر خواهند بود [23].

$$D = \min_{1\le i\le k}\left\{\min_{i+1\le j\le k}\left(\frac{d(c_i.c_j)}{\max_{1\le l\le k} diam(c_l)}\right)\right\}. \quad (۷)$$

که در آن، k تعداد خوشه‌ها، $d(c_i.c_j)$ فاصله بین خوشه‌های c_i و c_j $diam(c_l)$ قطر خوشه c_l است.

$$d(c_i.c_j) = \min_{x\in c_i.y\in c_j} d(x.y). \quad (۸)$$

$$diam\ c_l) = \max_{x.y\in c_l} d(x.y). \quad (۹)$$

شاخص SD: اساس شاخص اعتبارسنجی SD، میانگین پراکندگی[2] و جدایی کلی[3] خوشه‌ها است. مقدار محاسبه‌شده توسط این معیار هرچه کوچک‌تر باشد به معنی خوشه‌بندی بهتر است [22].

$$SD = \alpha S_a + S_t, \qquad S_a = \frac{1}{k}\sum_{i=1}^{k}\frac{\|\sigma\ v_i)\|}{\|\sigma\ x)\|}. \quad (۱۰)$$

$$S_t = \frac{\max_{1\le i.j\le k}(\|v_i - v_j\|)}{\min_{1\le i.j\le k}(\|v_i - v_j\|)}\sum_{i=1}^{k}(\sum_{i=1.j\ne i}^{k}\|v_i - v_j\|)^{-1}, \quad (۱۱)$$

که در آن $\sigma\ v_i)$ واریانس یک خوشه، $\sigma\ x)$ واریانس کد داده‌ها، v_i مرکز خوشه i و α عمل وزنی برای رابطه است که برابر میزان جدایی خوشه‌ها در صورت داشتن حداکثر تعداد خوشه‌ها می‌باشد.

شاخص خلوص: معیار خلوص برای خوشه‌هایی که دارای یک کلاس واحد هستند که با استفاده از **رابطه (۱۲)** محاسبه می‌شود [14].

$$Purity = \sum_{r=1}^{N}\frac{1}{n}\max_i(n_r^i), \quad (۱۲)$$

که n_r^i تعداد مشاهدات طبقه (کلاس) ith در خوشه rth است.

شاخص آنتروپی: یک مفهوم نظریه اطلاعات است که محتوای اطلاعات پیام‌ها را اندازه‌گیری می‌کند، در ارزیابی خوشه‌بندی، آنتروپی نحوه توزیع طبقات مختلف در هر خوشه را اندازه‌گیری می‌کند [14].

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}\left(-\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}log\frac{n_r^i}{n_r}\right), \quad (۱۳)$$

که در آن q تعداد طبقه‌ها، n_r تعداد اعضای خوشه r، k تعداد خوشه و n_r^iتعداد اعضای طبقه i در خوشه r است. مقدار محاسبه‌شده توسط این معیار هرچه بزرگ‌تر باشد به معنی خوشه‌بندی بهتر است.

شاخص زمان: مدت‌زمان محاسباتی لازم برای تحلیل خوشه‌ای یک الگوریتم خاص می‌باشد. الگوریتمی که زمان محاسبه در آن سریع تر باشد بهینه‌تر خواهد بود. روکاچ [24] پیشنهاد داد که انتخاب الگوریتم را می‌توان به‌عنوان یک مساله تصمیم‌گیری چندمعیاره درنظر گرفت و از تکنیک‌های $MCDM$ برای انتخاب بهترین روش گروهی در یک مساله استفاده نمود. از آن‌جاکه در این پژوهش ارزیابی الگوریتم‌های خوشه‌بندی شامل بیش‌از یک معیار، مانند آنتروپی، شاخص دان و زمان محاسبه است، می‌توان آن را به‌عنوان یک مساله $MCDM$ مدل‌سازی کرد.

۴-۲- شاخص‌های ارزیابی عملکرد مالی

ارزیابی عملکرد یکی از مهم‌ترین رویه‌های سازمانی است که مدیران برای دست‌یابی به اهداف از پیش تعیین‌شده ناشی از استراتژی‌های سازمان، انجام می‌دهند [25] و ازجمله بهترین راه‌های کسب اطلاعات برای تصمیم‌گیری در سازمان‌ها است [26]. تاکنون شیوه‌های مختلفی به‌منظور ارزیابی عملکرد سازمانی ارایه شده‌اند که ارزیابی مالی یکی از مهم‌ترین آن‌ها است. در این پژوهش به‌منظور استخراج داده‌های مالی و آزمون مدل پیشنهادی، عملکرد مالی شرکت‌های حاضر در بورس اوراق بهادار تهران موردبررسی قرارگرفته است.

[1] Compact and well-separated
[2] Avrage scattering
[3] Total sepration

بدین‌صورت که در ابتدا بامطالعه مبانی نظری، حدود ۴۰۰ معیار مالی در زمینه ارزیابی عملکرد شناسایی شدند که با اجماع نظر خبرگان تحقیق ۸۲ معیار به‌عنوان معیارهای دارای اهمیت بیش‌تر انتخاب گردیدند. در ادامه این معیارها در چهار منظر کارت امتیازی متوازن طبقه‌بندی‌شده و جهت غربال‌گری بیش‌تر در قالب پرسشنامه دلفی برای خبرگان ارسال شدند. پس از دو راند برگزاری دلفی و دست‌یابی به هم‌گرایی، ۲۱ معیار به‌عنوان معیارهای کلیدی ارزیابی عملکرد شرکت‌های حاضر در بورس اوراق بهادار تهران انتخاب گردیدند. درنهایت، با استفاده از روش نقشه نگاشت فازی[1] روابط بین معیارها بررسی و چهار شاخص، رشد سود عملیاتی[2]، نسبت قیمت به سود هر سهم[3]، نسبت سود عملیاتی به فروش و نسبت سود ناخالص به فروش برای ارزیابی نهایی انتخاب شدند. در ادامه به تعریف این معیارها پرداخته شده است.

رشد سود عملیاتی *OPG*: سود عملیاتی از تفاوت درآمدهای عملیاتی یک دوره واحد تجاری و هزینه‌های عملیاتی همان دوره حاصل می‌گردد، در این مطالعه رشد سود عملیاتی با استفاده از **رابطه (۱۴)** محاسبه می‌شود [27]:

$$OPG = \frac{EBIT_t - EBIT_{t-1}}{EBIT_{t-1}}. \qquad (۱۴)$$

در **رابطه (۱۴)** $EBIT_t$ و $EBIT_{t-1}$ به ترتیب سود عملیاتی در دوره t و در یک دوره ماقبل آن می‌باشد.

نسبت قیمت به سود هر سهم (*P/E*): نسبت سود به قیمت هر سهم به‌صورت راهبردهای سرمایه‌گذاری مورداستفاده قرار می‌گیرد و توسط تحلیلگران مرتبط با فروشنده، برای توجیه پیشنهادهای قیمت سهام آنان به‌کار برده می‌شود. از این‌رو، جامعه سرمایه‌گذاری این نسبت را به‌عنوان انتظارهای بازار از رشد آتی سود تفسیر می‌کند که از طریق **رابطه (۱۵)** محاسبه می‌شود [28]:

$$P/E = \text{قیمت هر سهم} / \text{سود هر سهم}. \qquad (۱۵)$$

نسبت سود عملیاتی به فروش: برای محاسبه نسبت سود عملیاتی شرکت کافی است که میزان سود عملیاتی آن را بر مبلغ فروش تقسیم نماییم [27].

$$\text{نسبت سود عملیاتی} = \text{سود عملیاتی} / \text{فروش}. \qquad (۱۶)$$

سود ناخالص به فروش: نسبت سود ناخالص به فروش نشان می‌دهد که از هر یک ریال فروش شرکت، چند درصد سود به‌دست‌آمده است. این نسبت، عملیات اجرایی و کسب درآمد شرکت را ارزیابی می‌کند و هم‌چنین توانایی شرکت در کنترل بهای تمام‌شده کالای فروش‌رفته را بررسی کرده و رابطه‌ی بین فروش و هزینه‌های تولید کالای فروخته‌شده را نشان می‌دهد. برای محاسبه نسبت سود ناخالص به فروش کافی است، سود ناخالص (قبل از کسر مالیات) را بر فروش تقسیم کنیم و عدد حاصل را در ۱۰۰ ضرب نماییم:

$$\text{نسبت سود ناخالص به فروش} = (\text{سود ناخالص} / \text{فروش}) * ۱۰۰. \qquad (۱۷)$$

به‌منظور محاسبه سود ناخالص کافی است بهای تمام‌شده کالاهای فروخته شده را از مبلغ فروش شرکت، کسر نماییم.

۵-۲- پیشینه تحقیق

اگرچه درزمینه اعتبارسنجی الگوریتم‌های خوشه‌بندی، مطالعات متعددی صورت گرفته است، با این‌وجود تعداد معدودی به بررسی این الگوریتم‌ها با استفاده از معیارهای مختلف پرداخته‌اند. از آن‌جایی‌که نتایج تحقیقات گذشته نشان‌دهنده آن است که الگوریتم‌های خوشه‌بندی در مجموعه داده‌های مختلف دارای عملکرد متفاوتی هستند و یک الگوریتم بهینه خوشه‌بندی برای مسایل مختلف قابل‌تعریف نیست، بررسی حاضر سعی دارد تا ضمن بررسی تحلیلی الگوریتم‌های خوشه‌بندی به مقایسه تطبیقی آن‌ها با استفاده از مهم‌ترین شاخص‌های اعتبارسنجی در مبانی نظری از نگاه خبرگان بپردازد تا بهره‌وران نتایج تحقیق به جمع‌بندی مناسب‌تری از نتایج دست‌یافته و از آن‌ها در تصمیمات سرمایه‌گذاری استفاده نمایند. برخی از پژوهش‌های انجام‌شده در این حوزه به‌همراه رویکرد هر کدام در **جدول ۲** آورده شده است.

[1] Fuzzy cognitive map
[2] Operating Profit Growth (OPG)
[3] Price/ Earnings Per Share (P/E)

۳- روش‌شناسی پژوهش

۳-۱- فرآیند اجرای تحقیق

این پژوهش به لحاظ روش از نوع توصیفی-پیمایشی است. در این مطالعه به‌منظور آزمون مدل پیشنهادی از اطلاعات مالی یک‌ساله شرکت‌های حاضر در بورس اوراق بهادار تهران در سال ۹۸ به‌دلیل در دسترس بودن کامل اطلاعات در زمان انجام پژوهش استفاده شده است که نشان‌دهنده عملکرد مالی آن‌ها می‌باشد. به‌منظور انتخاب شرکت‌ها باتوجه به چهار معیار: ۱- دوره مالی شرکت‌ها منتهی به پایان اسفندماه باشد، ۲- شرکت انتخابی جزو شرکت‌های سرمایه‌گذاری یا واسطه‌گری مالی، هلدینگ و لیزنگ‌ها نباشد، ۳- اطلاعات مالی شرکت‌ها در دوره زمانی موردمطالعه در دسترس باشد و ٤- معاملات سهام شرکت‌ها به‌طور مدام در بورس اوراق بهادار تهران صورت گرفته و توقف بیش از یک ماه نداشته باشد، از روش حذف نظام‌مند استفاده شد.

جدول ۲- خلاصه پژوهش‌های داخلی و خارجی انجام‌شده.
Table 2- Summary of domestic and foreign research.

پژوهشگران	عنوان پژوهش	نتیجه پژوهش
شاکری و همکاران [29]	مقایسه روش‌های خوشه‌بندی در داده‌های بیان ژنی	براساس شرایط داده‌ها می‌بایست بهترین روش خوشه‌بندی را انتخاب نمود.
فاضل زرندی و همکاران [21]	ارایه یک شاخص اعتبار خوشه‌بندی جدید با استفاده از معیار فاصله جاکار	شاخص ECASJ نسبت به سایر شاخص‌های اعتبارسنجی عملکرد مطلوب‌تری دارد.
شاکری و عبداللهی [30]	خوشه‌بندی داده‌ها، مروری بر روش‌های موجود و مقایسه عملکرد آن‌ها	هر یک از الگوریتم‌ها برای هر نوع داده دارای مزایا و معایبی می‌باشند. الگوریتم‌هایی که نیازی به تعیین تعداد خوشه ندارند، خوشه‌بندی بهتری را انجام می‌دهند.
زاده ده بالایی و همکاران [17]	بررسی مشکلات الگوریتم خوشه‌بندی DBSCAN و مروری بر بهبودهای ارایه شده آن	نتایج نشان داد، الگوریتم MD-DBSCAN نسبت به سایر الگوریتم‌های مورد ارزیابی، از شاخص شباهت بالاتر و میزان خطای پایین‌تری برخوردار است.
دهقان نیری [31]	ارایه یک شاخص جدید اعتبار خوشه‌بندی بر مبنای کاردینالیته فازی	نتایج تحقیق نشان داد شاخص FCI که از مفهوم کاردینالیته در مجموعه‌های فازی بهره می‌برد، دارای عملکرد مطلوبی می‌باشد.
گلبارد [32]	بررسی تنوع روش‌های خوشه‌بندی (یک مقایسه تجربی)	نتایج تحقیق آن‌ها ماهیت غیرقابل‌پیش‌بینی فرآیند خوشه‌بندی را تایید نمود.
هیرانو و تسوموتو [33]	بررسی مقایسه چندمتغیره و خوشه‌بندی معاملات سه‌بعدی مبتنی‌بر حداکثر منحنی	نتایج تحقیق آن‌ها درستی این الگوریتم خوشه‌بندی را برای داده‌های آن‌ها تایید نمود.
کاو و همکاران [14]	ارزیابی عملکرد الگوریتم‌های خوشه‌بندی برای تجزیه‌وتحلیل ریسک مالی با استفاده از روش‌های MCDM	الگوریتم خوشه‌بندی تقسیم مکرر[1] نتیجه بهتری به‌نسبت سایر روش‌ها برای داده‌های ریسک مالی دارد. هم‌چنین روش‌های MCDM رتبه مشابه برای شش الگوریتم خوشه‌بندی انتخابی تولید نمی‌کنند و هیچ الگوریتمی نمی‌تواند بهترین عملکرد را در تمام اندازه‌گیری‌ها برای هر مجموعه داده داشته باشد.
مک‌نیکولاس [34]	بررسی نحوه خوشه‌بندی داده‌ها باتوجه به مدل‌های مختلف و الگوریتم مختلف	روش‌های مختلف خوشه‌بندی مبتنی‌بر مدل مزایا و معایبی دارند که روش‌های جدید برطرف‌کننده نقاط ضعف روش‌های گذشته می‌باشند.
کیمز و همکاران [35]	بررسی تاثیر روش سلسله مراتبی در تجزیه‌وتحلیل خوشه‌بندی	این روش برای تشخیص ساختار خوشه‌ای واقعی و شبیه‌سازی مفید می‌باشد.
لوپز-روبیو و همکاران [36]	یادگیری پیشرفته با استفاده از بهینه‌سازی کیفیت خوشه‌بندی	الگوریتم استاندارد K-Means به‌عنوان یک مورد خاص می‌تواند باعث بهینه‌سازی کیفیت خوشه‌ها شود.
لونا-رومرا و همکاران [19]	شاخص اعتبار خوشه‌بندی خارجی براساس آزمون آماری خی ۲	بررسی شاخص اعتبار جدید خوشه‌بندی و مقایسه آن با ۱۵ شاخص اعتبار بیرونی، نشان داد که شاخص خی ۲ نسبت به سایر روش‌ها دارای عملکرد بهتری است.

[1] Repeated-bisection

جدول ۲- ادامه.

Table 2- Continued.

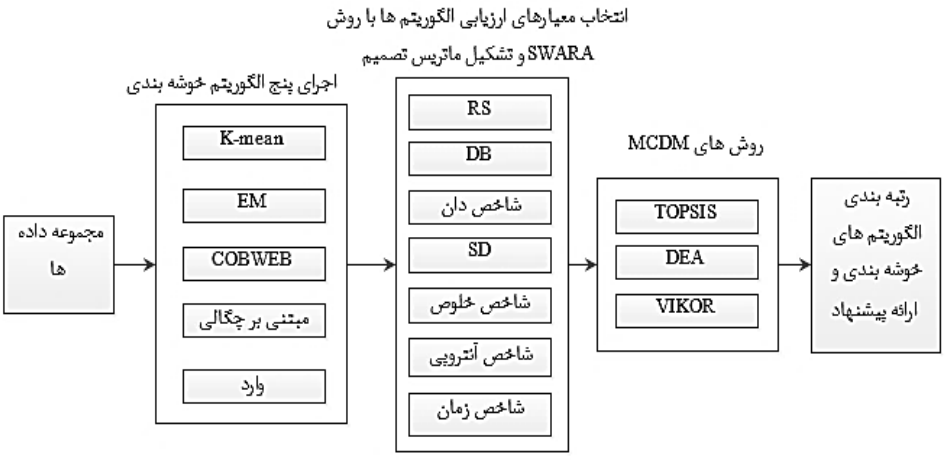
پژوهشگران	عنوان پژوهش	نتیجه پژوهش
رنجیث و همکاران [37]	بررسی عملکرد الگوریتم‌های خوشه‌بندی برای داده‌ها با بعد بالا	نتایج نشان داد برای این نوع داده‌ها روش‌های خوشه‌بندی k-Mean، CLARA و CLARANS دارای عملکرد بهتری است. آن‌ها به‌منظور بررسی الگوریتم‌های انتخابی از روش زمان چرخش استفاده نمودند.
حسن و همکاران [38]	نتایج ارزیابی عملکرد الگوریتم خوشه‌بندی تکاملی ستاره (ECA*) برای خوشه‌بندی مجموعه داده‌های ناهمگن	آن‌ها با استفاده از داده‌های موجود عملکرد ECA* را با پنج الگوریتم خوشه‌بندی مدرن و سنتی ازجمله K-Means و EM مقایسه نمودند. نتایج نشان داد الگوریتم ECA برای داده‌های ناهمگن به نسبت سایر الگوریتم‌ها دارای عملکرد بهتری است.
لوسیو ونتوریا و همکاران [39]	ارزیابی خوشه‌بندی و روش‌های مدل‌سازی موضوعی بر روی توییت ها و ایمیل‌های مربوط به سلامت	دلیل کوتاه بودن متون، رویکردهای خوشه‌بندی و مدل‌سازی موضوع موجود عملکرد غیر بهینه‌ای دارند و مقایسه آن‌ها دشوار است.

جدول ۳- ساختار مجموعه داده‌ها.

Table 3- Data set structure.

تعداد شرکت‌ها	تعداد صنایع	تعداد معیارهای ارزیابی عملکرد
403	43	4

به‌منظور ارزیابی الگوریتم‌های خوشه‌بندی از هفت شاخص، ضریب تعیین[1]، دیویس بولدین، دان، *SD*، خلوص، آنتروپی و شاخص زمان که براساس روش *SWARA* دارای اهمیت بیش‌تری شناخته شدند، استفاده گردیده است. هم‌چنین، برای ارزیابی عملکرد مالی شرکت‌ها از چهار معیار، *OPG*، نسبت قیمت به سود هر سهم، نسبت سود عملیاتی به فروش و نسبت سود ناخالص به فروش استفاده شد. **جدول ۳** نشان‌دهنده اطلاعات مربوط به مجموعه داده‌های مورداستفاده می‌باشد.



شکل ۲- فرآیند ارزیابی الگوریتم‌های خوشه‌بندی.

Figure 2- Evaluation process of clustering algorithms.

فرآیند اجرای این پژوهش که شامل پنج الگوریتم خوشه‌بندی، هفت معیار ارزیابی الگوریتم‌های خوشه‌بندی، یک روش غربال‌گری معیارهای ارزیابی و سه روش *MCDM* است که بر روی مجموعه داده‌های مالی پیاده شده‌اند، در **شکل ۲** نمایش داده شده است. گام‌های پیاده‌سازی تحقیق عبارت‌اند از:

گام ۱- آماده‌سازی مجموعه داده‌ها که شامل نمونه‌گیری با روش حذف نظام‌مند است و استخراج مقادیر چهار معیار ارزیابی عملکرد شرکت‌های حاضر در بورس اوراق بهادار تهران در سال ۹۸ است.

[1] R-Squared index

گام ۲- واردکردن اطلاعات استخراج‌شده در نرم‌افزار *Weka 3.9* و *SPSS 21* به‌منظور خوشه‌بندی داده‌ها با استفاده از پنج الگوریتم خوشه‌بندی منتخب.

گام ۳- غربال‌گری معیارهای ارزیابی عملکرد الگوریتم‌های خوشه‌بندی با استفاده از روش سوآرا و تعیین معیارهای نهایی.

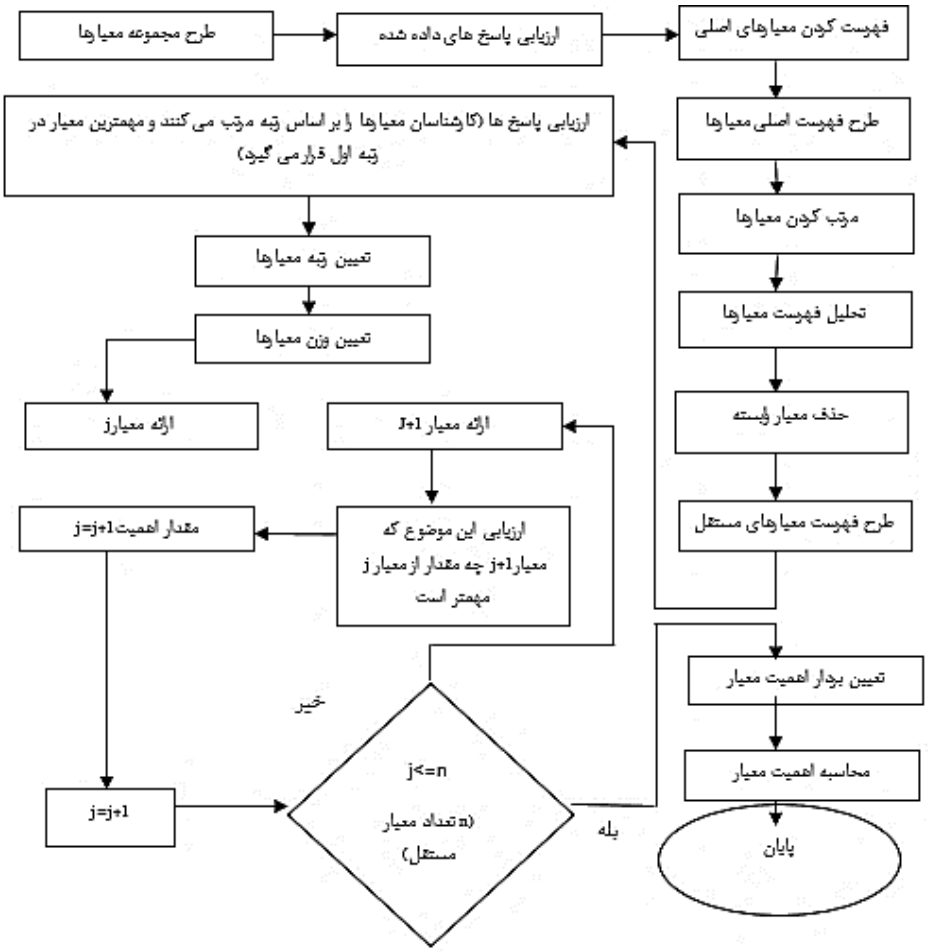
گام ۴- تعیین مقادیر عددی معیارهای ارزیابی عملکرد الگوریتم‌های خوشه‌بندی برای هر یک از پنج الگوریتم به‌کار رفته با استفاده از نرم‌افزار *MATLAB* و تشکیل ماتریس تصمیم ۷×۵.

گام ۵- رتبه‌بندی الگوریتم‌های خوشه‌بندی با استفاده از روش‌های تاپسیس، ویکور و تحلیل پوششی داده‌ها.

۲-۳- روش‌های به‌کار رفته در ارزیابی الگوریتم‌های خوشه‌بندی

روش سوآرا

روش سوارا یکی از روش‌های جدید تصمیم‌گیری چندمعیاره برای تعیین وزن شاخص‌ها است که توسط کرسولاین و تورسکیس [40] معرفی شد. از این روش برای محاسبه وزن معیارها استفاده می‌شود. روش سوارا یکی از روش‌های وزن‌دهی است که دیدگاه خبرگان در آن اهمیت بالایی دارد، جامعه روش سوارا شامل خبرگان حوزه موردمطالعه است. آن‌ها معتقد هستند که بهتر است گروهی از خبرگان گرد هم قرار گیرند و به‌صورت گروهی دیدگاه خود را مطرح کنند و پژوهش‌گر با یادداشت و جمع‌بندی دیدگاه خبرگان، ضمن رتبه‌بندی معیارها نسبت به تعیین وزن نسبی آن‌ها اقدام کند.



شکل ۳- چارچوب پیاده‌سازی سوآرا.

Figure 3- SWARA implementation framework.

در **گام ۱** تکنیک سوآرا، معیارهای موردنظر براساس میزان اهمیت به ترتیب نوشته می‌شوند. مهم‌ترین معیارها در رده‌های بالاتر و معیارهای کم‌اهمیت‌تر در رده‌های پایین‌تر قرار می‌گیرند. در **گام ۲** اهمیت نسبی هر معیار نسبت به معیارهای قبلی مشخص می‌شود.

در فرآیند روش سوآرا این مقدار با Sj نشان داده می‌شود. سپس ضریب Kj که تابعی از مقدار اهمیت نسبی هر معیار است محاسبه می‌شود. در **گام ۴** وزن اولیه معیارها Qj تعیین می‌گردد. نهایتا وزن نهایی معیارها با نرمال‌سازی به روش فراوانی نسبی به‌دست می‌آید.

روش تاپسیس[1]

این تکنیک بر این مفهوم استوار است که گزینه انتخابی، باید کم‌ترین فاصله را با راه‌حل ایده‌آل مثبت (بهترین حالت ممکن) و بیش‌ترین فاصله را با راه‌حل ایده‌آل منفی (بدترین حالت ممکن) داشته باشد. پس از تشکیل ماتریس تصمیم که گزینه‌ها و معیارهای انتخاب است مراحل انجام این تکنیک به‌صورت زیر می‌باشد:

گام ۱- محاسبه ماتریس تصمیم نرمال‌سازی شده.

$$r_{ij} = \frac{X_{ij}}{\sqrt{\sum_{i=1}^{J} X_{ij}^2}}, \quad j = 1,2,\dots,J, \quad i = 1,2,\dots,n. \qquad (۱۸)$$

گام ۲- وزن‌دار کردن ماتریس تصمیم نرمال شده.

$$\vartheta_{ij} = W_i r_{ij}, \quad j = 1,2,\dots,J, \qquad i = 1,2,\dots,n, \quad \sum_{i=1}^{n} W_i = 1. \qquad (۱۹)$$

گام‌های ۳ و ۴- تعیین نقاط ایده‌آل مثبت و منفی.

$$S^+ = \{\vartheta_1^+, \dots, \vartheta_n^+\} = \left\{\left(\max_j \vartheta_{ij} | i \in I_1\right), \left(\min_j \vartheta_{ij} | i \in I_2\right)\right\}. \qquad (۲۰)$$

$$S^- = \{\vartheta_1^-, \dots, \vartheta_n^-\} = \left\{\left(\min_j \vartheta_{ij} | i \in I_1\right), \left(\max_j \vartheta_{ij} | i \in I_2\right)\right\}. \qquad (۲۱)$$

در رابطه فوق I_1 و I_2 به ترتیب نشان‌دهنده مجموعه‌هایی هست که شاخص‌های مثبت و منفی را در خود دارند.

گام ۵- به‌دست آوردن اندازه فاصله‌ها از ایده‌آل مثبت و منفی.

$$D_j^+ = \sqrt{\sum_{i=1}^{n} (\vartheta_{ij} - \vartheta_1^+)^2}, \quad j = 1,2,\dots,J. \qquad (۲۲)$$

$$D_j^- = \sqrt{\sum_{i=1}^{n} (\vartheta_{ij} - \vartheta_1^-)^2}, \quad j = 1,2,\dots,J. \qquad (۲۳)$$

گام ۶- محاسبه فاصله نسبی از ایده‌آل مثبت.

$$R_j^+ = \frac{D_j^-}{D_j^+ + D_j^-}, \quad j = 1,2,\dots,J. \qquad (۲۴)$$

گام ۷- رتبه‌بندی گزینه‌های موجود با ماکزیمم مقدار نسبت R_j^+.

روش تحلیل پوششی داده‌ها

تحلیل پوششی داده‌ها، یک رویکرد ناپارامتریک برای اندازه‌گیری کارایی واحدهای تصمیم‌گیرنده[2] متجانس است که اولین بار توسط چارنز و همکاران [41] ارایه گردید [25]. این روش مبتنی‌بر برنامه‌ریزی خطی ریاضی است که به‌منظور ارزیابی کارایی واحدهای تصمیم‌گیرنده از طریق شناسایی مرز کارایی و مقایسه هر DMU با آن عمل می‌کند. DEA از ابزارهای مناسب و کارا در زمینه سنجش و ارزیابی بهره‌وری است [42] و به دلیل برآورد کارایی با حداقل پیش‌فرض‌ها به نسبت سایر روش‌ها، مانند تحلیل رگرسیون دارای مزیت نسبی است [14]. مدل اصلی که توسط چارنز و همکاران [41] ارایه شده بود CCR نامیده می‌شد که از نسبت خروجی‌ها به ورودی‌ها برای اندازه‌گیری کارایی $DMUs$ استفاده می‌کند. فرض کنید J واحد تصمیم‌گیرنده با m ورودی و s خروجی وجود دارد. x_{ij} و y_{rj} به ترتیب مقدار ورودی ith و

[1] Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)

[2] Decision Making Units (DMUs)

خروجی rth برای DMU_J $(J = 1,2, \dots, J)$ را نشان می‌دهد. در این پژوهش برای بررسی کارایی الگوریتم‌های خوشه‌بندی، علاوه‌بر تکنیک‌های تاپسیس و ویکور از مدل‌های CCR و BCC تحلیل پوششی داده‌ها استفاده شده است. مدل پوششی CCR در **معادله (۲۵)** آورده شده است.

$$\begin{aligned} & max\ \varphi_o, \\ & s.t. \\ & \sum_{j=1}^{n} \lambda_j x_{ij} \le x_{io}, \quad i = 1,2, \dots, m, \\ & \sum_{j=1}^{n} \lambda_j y_{rj} \ge \varphi\, y_{ro}, \quad r = 1,2, \dots, s, \\ & \lambda_j \ge 0, \ \ j = 1,2, \dots, n. \end{aligned} \qquad \textbf{(۲۵)}$$

در معادله فوق u_r و v_i متغیرهای تصمیم و y_{r0} و x_{i0} مقادیر خروجی r و ورودی i مربوط به DMU تحت ارزیابی هستند. بنکر و همکاران [43] با اضافه نمودن قید $\sum_{j=1}^{J} \lambda_j = 1$ به مدل CCR، مدل BCC را ارایه نمودند. مدل پوششی BCC در **معادله (۲۶)** آورده شده است.

$$\begin{aligned} & max\ \varphi_o, \\ & s.t. \\ & \sum_{j=1}^{n} \lambda_j x_{ij} \le x_{io}\,, \quad i = 1,2, \dots, m, \\ & \sum_{j=1}^{n} \lambda_j y_{rj} \ge \varphi\, y_{ro}, \quad r = 1,2, \dots, s, \\ & \sum_{j=1}^{n} \lambda_j = 1, \\ & \lambda_j \ge 0, \ \ j = 1,2, \dots, n. \end{aligned} \qquad \textbf{(۲۶)}$$

براساس مقدار تابع هدف، $DMUs$ با مقدار **تابع هدف (۱)** کارا و با سایر مقادیر ناکارا هستند. مدل‌های بازده‌به‌مقیاس ثابت، محدودکننده‌تر از مدل‌های بازده‌به‌مقیاس متغیر می‌باشند [44] و این مدل‌ها از قدرت تفکیک‌پذیری بیش‌تری در ارزیابی $DMUs$ برخوردارند. از سوی دیگر مدل‌های بازده‌به‌مقیاس متغیر دارای پایداری بیش‌تری در خروجی‌ها هستند و در مقایسه با مدل‌های دسته اول کاربرد بیش‌تری در مبانی نظری دارند؛ بنابراین، در پژوهش حاضر از مدل‌های CCR و BCC تحلیل پوششی داده‌ها به‌طور هم‌زمان در ارزیابی کارایی استفاده شده است. قابل‌ذکر است که از میان ۷ شاخص اعتبارسنجی الگوریتم‌های خوشه‌بندی، شاخص‌های دارای ماهیت مثبت به‌عنوان خروجی‌ها و شاخص‌های دارای ماهیت منفی به‌عنوان ورودی‌های مدل‌های DEA درنظر گرفته شده‌اند.

روش ویکور

یکی دیگر از روش‌های تصمیم‌گیری چندمعیاره برای حل یک مساله تصمیم‌گیری گسسته با معیارهای نامتناسب واحدهای اندازه‌گیری مختلف و متعارض توسط اپروکویک و تیزنگ ایجاد شده است. کلمه ویکور، برگرفته از نام صربستانی[1]، به معنای بهینه‌سازی چندمعیاره و حل سازشی[2] است. این روش، یک مجموعه رتبه‌بندی شده از گزینه‌های موجود را باتوجه به شاخص‌های متضاد تعیین می‌کند، به‌طوری‌که رتبه‌بندی گزینه‌ها براساس این هدف صورت می‌گیرد. در این پژوهش مراحل پیاده‌سازی الگوریتم ویکور پس از تشکیل ماتریس تصمیم و نرمال‌سازی آن به‌صورت زیر است:

گام ۱- تعیین بهترین f_i^+ و بدترین f_i^- مقدار در میان همه گزینه‌ها برای هر یک از معیارها.

$$f_i^+ = \begin{Bmatrix} \max\limits_j f_{ij}, & \text{برای معیارهای مثبت} \\ \min\limits_j f_{ij}, & \text{برای معیارهای منفی} \end{Bmatrix}, \quad i = 1,2, \dots, n, \qquad j = 1,2, \dots, J. \qquad \textbf{(۲۷)}$$

[1] VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR)

[2] Multi-criteria optimization and compromise solution

$$f_i^- = \begin{Bmatrix} \min\limits_j f_{ij}, & \text{برای معیارهای مثبت} \\ \max\limits_j f_{ij}, & \text{برای معیارهای منفی} \end{Bmatrix}, \quad i = 1,2,\dots,n, \qquad j = 1,2,\dots,J. \tag{۲۸}$$

گام ۲- محاسبه مقادیر S_j (شاخص مطلوبیت) و R_j (شاخص نارضایتی) برای $j = 1,2,\dots,J$ به‌صورت زیر:

$$S_j = \sum_{i=1}^{n} w_i(f_i^+ - f_{ij})/(f_i^+ - f_i^-). \tag{۲۹}$$

$$R_j = \max_i \left[w_i(f_i^+ - f_{ij})/(f_i^+ - f_i^-) \right], \tag{۳۰}$$

که در آن w_i وزن معیار ith است.

گام ۳- محاسبه مقادیر Q_j برای $j = 1,2,\dots,J$ به‌صورت زیر:

$$Q_j = \frac{\vartheta(S_j - S^*)}{(S^- - S^*)} + (1-\vartheta)\frac{(R_j - R^*)}{(R^- - R^*)},$$

$$S^* = \min_j S_j, \quad S^- = \max_j S_j. \tag{۳۱}$$

$$R^* = \min_j R_j, \quad R^- = \max_j R_j.$$

در رابطه فوق پارامتر ϑ باتوجه به میزان توافق گروه تصمیم‌گیرنده تعیین می‌گردد. در این پژوهش مقدار آن ۰/۵ درنظر گرفته‌شده است.

گام ۴- در گام پایانی، گزینه‌ها براساس مقادیر S، R و Q در سه گروه از کوچک به بزرگ مرتب می‌شوند. بهترین گزینه آن است که کوچک‌ترین Q را داشته باشد به‌شرط آن‌که دو شرط زیر برقرار باشد:

۱. شرط اول: اگر گزینه A_1 و A_2 در میان J گزینه رتبه اول و دوم را داشته باشند، باید رابطه $Q(A_2) - Q(A_1) \geq 1/(J-1)$ برقرار باشد.

۲. شرط دوم: گزینه A_1 باید حداقل در یکی از گروه‌های R و S به‌عنوان رتبه برتر شناخته شود.

۴- تجزیه‌وتحلیل داده‌ها

۴-۱- اولویت‌بندی معیارها با استفاده از روش سوآرا

در ادامه از روش سوآرا به‌منظور تعیین وزن معیارها استفاده شد. بدین‌صورت که در ابتدا خبرگان معیارها را به ترتیب اهمیت مرتب نمودند و میانگین رتبه‌های اعلام‌شده مبنای مرتب‌سازی معیارها به ترتیب نزولی قرار گرفت. سپس اهمیت نسبی هر معیار نسبت به معیارهای قبلی (Sj) و ضریب Kj که تابعی از مقدار اهمیت نسبی هر معیار است محاسبه گردید. در این مقاله به‌منظور ایجاد تمایز بیش‌تر بین شاخص‌ها، از توان دوم Kj استفاده شده است. ستون ٤ و ٥ در **جدول ۳** وزن اولیه و نهایی معیارها را نشان می‌دهند. از آن‌جایی‌که وزن معیار هشتم (هوبرت) و معیارهای بعدازآن کم‌تر از ۰/۰۱ است و باتوجه به تفاوت فاحش وزن شاخص ۱ تا ۷ نسبت به سایر شاخص‌ها، بنابر نظر خبرگان از شاخص ۱ و شاخص ۷ در ارزیابی الگوریتم‌ها استفاده شد و سایر شاخص‌ها از ادامه ارزیابی کنار گذاشته شدند. در آخرین ستون نیز وزن نهایی معیارها با استفاده از فراوانی نسبی به‌دست آمده است.

جدول ۳- رتبه‌بندی معیارهای ارزیابی با روش سوآرا.

Table 3- Ranking of evaluation criteria by Soara method.

شاخص	ضریب S_j	ضریب $k_j = S_j + 1$	وزن اولیه	وزن نهایی
انحراف استاندارد (SD)	-	1	1	0.4575
دان	0.946	1.946	0.514	0.2350
DB	0.848	1.848	0.278	0.1272
آنتروپی	0.786	1.786	0.156	0.0712
خلوص	0.750	1.750	0.089	0.0407
ضریب تعیین	0.714	1.714	0.052	0.0237
زمان	0.598	1.598	0.032	0.0149
هوبرت بازنگری شده	0.545	1.545	0.021	0.0096

موحدی و همکاران / تصمیم‌گیری و تحقیق در عملیات، دوره ۸، شماره ۲، تابستان ۱۴۰۲، صفحه: ۵۲۶-۵۰۷

جدول ۳- ادامه.

Table 3- Continued.

شاخص	ضریب S_j	ضریب $k_j = S_j + 1$	وزن اولیه	وزن نهایی
S-Dlow	0.501	1.501	0.014	0.0064
هوبرت	0.491	1.491	0.009	0.0043
آماره نرمالایز شده	0.446	0.441	0.006	0.0030
ضریب همبستگی کوفنتیک	0.375	1.375	0.005	0.0022
راند	0.366	1.366	0.003	0.0016
ضریب جاکارد	0.339	1.339	0.003	0.0012
آماره استاندارد هوبرت	0.304	1.304	0.002	0.0009
شاخص فولکسو مالوز	0.259	1.259	0.002	0.0007

آماده‌سازی داده‌ها یکی از مراحل اصلی است که داده‌ها را برای مراحل بعد آماده می‌کند؛ بنابراین، ابتدا براساس شروط بیان‌شده در بخش ۳-۱، با استفاده از روش نمونه‌گیری حذف نظام‌مند شرکت‌هایی که اطلاعات کامل آن‌ها در بانک اطلاعاتی بورس اوراق بهادار تهران در سال ۹۸ موجود بود انتخاب شد، سپس متغیرهای تحقیق با استفاده از **رابطه (۱۴)** و **رابطه (۱۷)** محاسبه شدند. براساس ادبیات نظری، روش‌های متعددی برای خوشه‌بندی داده‌ها وجود دارد. در این پژوهش به‌منظور ارزیابی الگوریتم‌های خوشه‌بندی از پنج الگوریتم خوشه‌بندی، *K-Means*، ماکزیمم امید ریاضی، *COBWEB*، روش مبتنی‌بر چگالی و الگوریتم وارد استفاده و داده‌ها خوشه‌بندی شدند، نتایج تعداد خوشه‌های درنظر گرفته‌شده توسط هر یک از الگوریتم‌ها در **جدول ۴** ارایه شده است.

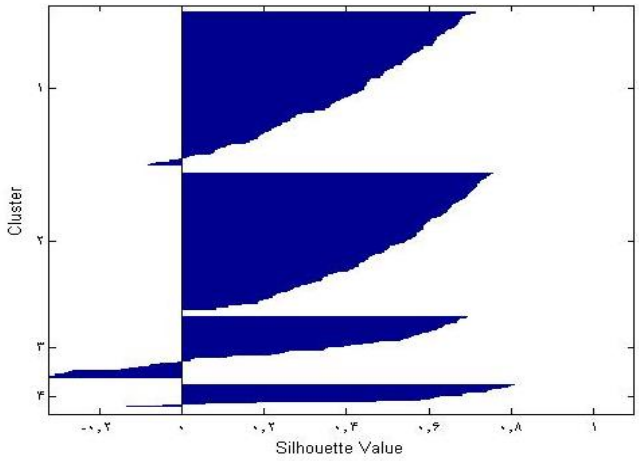
جدول ۴- تعداد خوشه‌ها با استفاده از هر روش خوشه‌بندی.

Table 4- Number of clusters using each clustering method.

روش خوشه‌بندی	تعداد خوشه
EM	7
COBWEB	302
مبتنی‌بر چگالی	2
K-Means	2
وارد	2

همان‌گونه که در مبانی نظری اشاره شد برخی توابع خوشه‌بندی نیازمند تعیین پارامتر اولیه مثل تعداد خوشه‌ها از طرف کاربر هستند و دسته دیگر به این پارامترها نیاز ندارند، الگوریتم‌های *EM*، *COBWEB*، مبتنی‌بر چگالی و الگوریتم وارد به تعداد خوشه‌ها نیاز ندارند و تعداد خوشه‌ها در این الگوریتم‌ها براساس نوع داده و طی فرآیند خوشه‌بندی تعیین می‌گردد، اما در الگوریتم *K-Means* می‌بایست تعداد خوشه‌ها از پیش تعیین شود، در این پژوهش به‌منظور تعیین تعداد بهینه خوشه‌ها در الگوریتم *K-Means* از نمودار و متوسط ضریب نیمرخ[1] استفاده نمودیم، نمودار نیمرخ که بر پایه ماتریس عدم تشابه بنانهاده شده، ابزاری برای کیفیت خوشه‌بندی است که به‌وسیله روسیوف [45] ارایه شد، مقدار شاخص نیمرخ بین ۱- و ۱+ قرار دارد، هرچقدر این شاخص به ۱+ نزدیک‌تر باشد نشان می‌دهد که شی i به خوشه خودش نزدیک‌تر است تا به همسایه‌اش؛ بنابراین، در این حالت طبقه‌بندی به‌درستی انجام شده است. زمانی‌که مقدار شاخص به ۱- نزدیک شود بدان معناست که فاصله شی i با خوشه خودی زیاد و با خوشه همسایه کم است، به‌عبارت‌دیگر طبقه‌بندی نامناسب است. متوسط ضریب نیمرخ، متوسط ضرایب همه اشیا است که نشان‌گر خوبی برای تعداد بهینه خوشه‌ها می‌باشد. مقادیر بالای ۰/۵ نشان‌دهنده اعتبار قابل‌پذیرش و بیان‌گر ساختار قابل‌قبول خوشه‌بندی است [22]. باتوجه به نمودار ضریب نیمرخ (**شکل ٤**) برای داده‌های پژوهش حاضر، زمانی‌که تعداد خوشه‌ها ۲ درنظر گرفته شد، نمودار تماما مثبت بود؛ اما وقتی تعداد خوشه‌ها بیش‌تر از ۲ درنظر گرفته شد، شاهد مقادیر کم‌تر از صفر بودیم؛ بنابراین، می‌توان نتیجه گرفت برای این نوع داده در روش خوشه‌بندی *K-Means* تعداد بهینه خوشه‌ها ۲ می‌باشد. نتایج حاصل از محاسبه متوسط ضرایب نیمرخ در **جدول ۵** آمده است، نتایج نشان می‌دهد زمانی‌که تعداد خوشه‌ها برابر ۲ است، مقدار این شاخص برابر ۰/۵۷۳٤ بوده که بیان‌گر ساختار قابل‌قبول خوشه‌بندی است؛ بنابراین، در این پژوهش به‌منظور خوشه‌بندی داده به روش *K-Means* از ۲ خوشه استفاده شد.

[1] Silhouette



شکل ۴- نمودار ضریب نیمرخ.
Figure 4- Profile coefficient diagram.

جدول ۵- متوسط ضرایب نیمرخ.
Table 5- Average profile coefficients.

تعداد خوشه	متوسط ضریب نیمرخ
2	0.5734
3	0.4332
4	0.4630

پس از خوشه‌بندی داده‌ها، به‌منظور سنجش کیفیت و ارزیابی هر یک از الگوریتم‌های خوشه‌بندی ارایه شده از هفت شاخص، *RS*، *DB*، دان، *SD*، خلوص، آنتروپی و زمان محاسبه استفاده شده است. برای محاسبه این شاخص‌ها از نرم‌افزار متلب استفاده گردید که مقدار هر یک از شاخص‌ها برای هر یک از الگوریتم‌های خوشه‌بندی در **جدول ۶** ارایه شده است.

جدول ۶- ماتریس نتایج معیارهای ارزیابی عملکرد الگوریتم‌های خوشه‌بندی.
Table 6- Matrix of results of performance evaluation criteria of clustering algorithms.

معیارها/الگوریتم	دان	آنتروپی	خلوص	RS	SD	DB	زمان
EM	0.0006	0.6805	0.1693	0.4199	9464	38.05	3.27
COBWEB	0.0000	0.2379	0.9561	0.9818	40675	6.08	0.18
مبتنی‌بر چگالی	0.1131	0.1752	0.1034	0.0881	0.0473	2.45	0.01
K-Means	0.3452	0.1765	0.1191	0.4667	0.0255	1.0118	0.0001
وارد	0.3073	0.0702	0.1066	0.0124	0.0180	0.5008	0.21

نتایج **جدول ۶** که براساس **رابطه (۲)** و **رابطه (۱۳)** و پیاده‌سازی در نرم‌افزار متلب محاسبه شده‌اند، نشان می‌دهد که شاخص‌های ارزیابی عملکرد الگوریتم‌های خوشه‌بندی نتایج یکسانی ندارند، مقدار محاسبه‌شده برای شاخص‌های دان و زمان، بیانگر آن است که الگوریتم *K-Means* به نسبت سایر روش‌ها دارای عملکرد بهتری است. شاخص‌های آنتروپی، *SD* و *DB*، روش وارد را بهترین الگوریتم خوشه‌بندی برای این مجموعه داده می‌دانند؛ اما دو شاخص خلوص و *RS* نشان می‌دهند که الگوریتم *COBWEB* دارای بهترین عملکرد است. از این‌رو با قاطعیت نمی‌توان گفت کدام الگوریتم خوشه‌بندی برای این مجموعه داده‌ها دارای بهترین عملکرد است؛ چراکه شاخص‌های مختلف نشان‌دهنده نتایج مختلف هستند.

به‌منظور رفع این مشکل، باتوجه به این‌که در این پژوهش از هفت شاخص برای ارزیابی اعتبار الگوریتم‌های خوشه‌بندی استفاده شده است، یک ماتریس تصمیم شکل می‌گیرد که گزینه‌های آن الگوریتم‌های خوشه‌بندی و شاخص‌های آن معیارهای اعتبارسنجی هستند. در ادامه از سه روش تاپسیس، تحلیل‌پوششی‌داده‌ها و ویکور برای تجزیه‌وتحلیل ماتریس تصمیم و رتبه‌بندی الگوریتم‌های خوشه‌بندی استفاده شده است. نتایج این ارزیابی در **جدول ۷** ارایه شده است.

جدول ۷ - نتایج رتبه‌بندی الگوریتم‌های خوشه‌بندی توسط روش‌های MCDM.

Table 7- Ranking results of clustering algorithms by MCDM methods.

	TOPSIS		VIKOR		DEA	
الگوریتم	مقدار	رتبه	مقدار	رتبه	CCR	BCC
EM	1.2238	5	0.3477	5	0.291	0.260
COBWEB	0.8261	3	0.5910	4	1	1
مبتنی‌بر چگالی	0.8712	4	0.5917	3	0.854	1
K-Means	0.2238	1	0.6754	1	1	1
وارد	0.7856	2	0.6214	2	1	1

در این تحقیق، وزن هر یک از هفت شاخص ارزیابی اعتبار الگوریتم‌های خوشه‌بندی یکسان درنظر گرفته شده است. براساس نتایج، روش‌های تاپسیس و ویکور در انتخاب بهترین الگوریتم خوشه‌بندی نتایج یکسانی ارایه می‌دهند و الگوریتم *K-Means* را برای خوشه‌بندی شرکت‌هایی که عملکرد آن‌ها از منظر مالی در حال ارزیابی است، بهترین الگوریتم می‌دانند. این دو روش پس از الگوریتم *K-Means*، الگوریتم وارد را بهترین دانستند. به‌منظور بررسی کارایی الگوریتم‌های خوشه‌بندی نیز از مدل‌های *CCR* **رابطه (۲۵)** و *BCC* **رابطه (۲۶)** استفاده شده است. نتایج مدل *CCR* نشان می‌دهد که الگوریتم‌های *K-Means*، *COBWEB* و وارد برای مجموعه داده‌های پژوهش کارا و الگوریتم‌های مبتنی‌بر چگالی و *EM* ناکارا هستند. به‌علاوه نتایج مدل *BCC* نیز بیان‌گر ناکارایی الگوریتم *EM* و کارایی ۴ الگوریتم دیگر است. لازم به ذکر است که در مدل‌های *DEA*، شاخص‌های آنتروپی، *SD*، *DB* و زمان به دلیل ماهیت منفی (کاهشی) به‌عنوان متغیرهای ورودی و شاخص‌های دان، *RS* و خلوص به دلیل ماهیت مثبت (افزایشی) به‌عنوان متغیرهای خروجی درنظر گرفته شدند.

۵- بحث و نتیجه‌گیری

در دنیای امروز مساله تصمیم‌گیری از مسایل پراهمیتی است که ذهن بسیاری از مدیران شرکت‌ها در سطوح مختلف و سرمایه‌گذاران در حوزه‌های گوناگون به‌خصوص حوزه مالی را به خود مشغول کرده است، در حل مسایل تصمیم‌گیری، روش‌های بسیاری وجود دارد که روش‌های تصمیم‌گیری چندمعیاره و تکنیک‌های داده‌کاوی از روش‌های پرکاربرد در این عرصه است. از آن‌جایی‌که نتایج بسیاری از تحقیقات گذشته نشان‌دهنده این مطلب بود که هیچ‌کدام از الگوریتم‌های خوشه‌بندی نمی‌تواند بهترین عملکرد را در تمام اندازه‌گیری‌ها برای هر مجموعه داده داشته باشد و هم‌چنین به دلیل این‌که هر یک از معیارهای ارزیابی اعتبار الگوریتم‌های خوشه‌بندی موجود داری نقاط قوت و ضعف هستند و نمی‌توانند به‌تنهایی در انتخاب بهترین الگوریتم موفق باشند، در پژوهش حاضر یک رویکرد ارزیابی جدید پیشنهادشده است و از ترکیب روش‌های *MCDM* و *DEA* برای ارزیابی اعتبار الگوریتم‌های خوشه‌بندی در حوزه ارزیابی عملکرد مالی شرکت‌ها بهره می‌برد. براساس یافته‌های محققان به نظر می‌رسد تکنیک‌های ترکیبی نسبت به نمونه‌های غیرترکیبی، از موفقیت بیش‌تری در حل مسایل حساس و پیچیده برخوردارند [46]. به‌منظور تایید روش پیشنهادی و هم‌چنین آماده‌سازی مجموعه داده‌های پژوهش در ابتدا از میان شرکت‌های حاضر در بورس اوراق بهادار تهران در سال ۹۸ با استفاده از روش نمونه‌گیری حذف نظام‌مند، ۴۰۳ شرکت که در ۴۳ صنعت مختلف دسته‌بندی‌شده بودند، انتخاب شدند. در مرحله بعد به‌منظور سنجش عملکرد شرکت‌های منتخب از میان تعداد زیاد شاخص‌های ارزیابی عملکرد شرکت‌ها، چهار شاخص، رشد سود عملیاتی (*OPG*)، نسبت قیمت به سود هر سهم (*P/E*)، نسبت سود عملیاتی به فروش و نسبت سود ناخالص به فروش، باتوجه به روابط بازخوردی میان منظرهای مختلف روش کارت امتیاز متوازن (*BSC*) مرتبط با حوزه ارزیابی عملکرد و با استفاده از ترکیب روش‌های دلفی فازی و نقشه نگاشت فازی و بهره گرفتن از نظر خبرگان انتخاب شدند.

پس از استخراج عملکرد مالی شرکت‌ها توسط چهار شاخص فوق و آماده شدن مجموعه داده‌های پژوهش، شرکت‌ها با استفاده از پنج الگوریتم خوشه‌بندی پرتکرار در تحقیقات گذشته شامل *K-Means*، *EM*، *COBWEB*، الگوریتم مبتنی‌بر چگالی و الگوریتم وارد، خوشه‌بندی شدند. مساله دیگری که در هنگام خوشه‌بندی شرکت‌ها وجود داشت تعداد خوشه‌ها در هر الگوریتم بود، چراکه اگر تعداد خوشه‌ها بهینه نباشد، فرآیند خوشه‌بندی به‌درستی انجام نگرفته است و در نتایج تحقیق خلل ایجاد می‌نماید. از آن‌جایی‌که تنها در الگوریتم خوشه‌بندی *K-Means* می‌بایست تعداد خوشه‌ها از قبل تعیین شوند و در سایر الگوریتم‌ها طی فرآیند خوشه‌بندی براساس مجموعه داده‌ها تعداد بهینه خوشه‌ها تعیین می‌گردد، از روش‌های نمودار و متوسط ضریب نیمرخ برای تعیین تعداد خوشه‌ها استفاده گردیده است. هر دو روش اشاره‌شده تعداد بهینه خوشه‌ها را برای مجموعه داده‌های پژوهش در الگوریتم *K-Means*، ۲ خوشه دانستند.

در مرحله بعد، پس از خوشه‌بندی شرکت‌ها، اعتبار هر یک از این پنج الگوریتم خوشه‌بندی با استفاده از مجموعه‌ای از معیارهای ارزیابی اعتبار داخلی و خارجی شامل ۷ معیار *RS*، *DB*، دان، *SD*، خلوص، آنتروپی و مدت‌زمان محاسبه، اندازه‌گیری شدند. دلیل انتخاب ۷ معیار برای ارزیابی الگوریتم‌های خوشه‌بندی را می‌توان در اهمیت تعیین‌شده برای آن‌ها براساس روش سوآرا دانست.

نتایج اولیه با تطبیقی الگوریتم‌های خوشه‌بندی در شاخص‌های ارزیابی نشان داد که هیچ‌یک از الگوریتم‌ها نمی‌توانند بهترین عملکرد را در تمام اندازه‌گیری‌ها برای مجموعه داده به‌دست آورد. همان‌طور که در **جدول ٦** مشاهده می‌شود، الگوریتم *EM* در مقایسه با *COBWEB* در شاخص‌های دان، خلوص، *RS* و *SD* عملکرد بهتری دارد، درحالی‌که در شاخص‌های آنتروپی، *DB* و زمان از وضعیت بدتری برخوردار است. در مقایسه الگوریتم *EM* و الگوریتم مبتنی‌بر چگالی مشاهده می‌شود که الگوریتم *EM* در شاخص‌های آنتروپی و خلوص دارای عملکرد بهتر و در شاخص‌های دان، *RS* و *SD*، *DB* و زمان دارای عملکرد بدتری است. هم‌چنین، الگوریتم *EM* در شاخص‌های آنتروپی، *RS* و خلوص عملکرد دارای عملکرد بهتر و در شاخص‌های دان، *SD*، *DB* و زمان دارای عملکرد بدتری نسبت به الگوریتم *K-Means* می باشد. نهایتا، در مقایسه الگوریتم *EM* و الگوریتم وارد مقایسه می‌شود که الگوریتم *EA* در شاخص‌های آنتروپی و خلوص دارای عملکرد بهتر و در شاخص‌های دان، *RS* و *SD*، *DB* و زمان دارای عملکرد بدتری است. الگوریتم *COBWEB* نیز نسبت به الگوریتم‌های مبتنی‌بر چگالی و *K-Means* در شاخص‌های آنتروپی و خلوص دارای عملکرد بهتر و در شاخص‌های دان، *RS* و *SD*، *DB* و زمان دارای عملکرد بدتری است. این الگوریتم در مقایسه با الگوریتم وارد نیز در شاخص‌های آنتروپی، خلوص و زمان دارای عملکرد بهتر و در شاخص‌های دان، *RS* و *SD* و *DB* دارای عملکرد بدتری است. با بررسی صورت گرفته، الگوریتم مبتنی‌بر چگالی نیز نسبت به الگوریتم *K-Means* در شاخص‌های *RS* و *SD* و *DB* دارای عملکرد بهتر و در شاخص‌های دان، آنتروپی، خلوص و زمان دارای عملکرد بدتری است. از سوی دیگر مقایسه الگوریتم‌های مبتنی‌بر چگالی و وارد نشان می‌دهد که الگوریتم مبتنی‌بر چگالی در شاخص‌های آنتروپی و زمان دارای عملکرد بهتر و در شاخص‌های دان، خلوص، *RS* و *SD* و *DB* دارای عملکرد بدتری است. نهایتا، الگوریتم *K-Means* در مقایسه با الگوریتم وارد در شاخص‌های دان، آنتروپی، خلوص و زمان دارای عملکرد بهتر و در شاخص‌های *RS* و *SD* و *DB* دارای عملکرد بدتری است.

در ادامه به‌منظور استفاده هم‌زمان از تمامی معیارهای هفت‌گانه و انتخاب بهترین الگوریتم خوشه‌بندی برای مجموعه داده‌های پژوهش، از دو روش پرکاربرد *MCDM* یعنی تاپسیس و ویکور به‌همراه روش *DEA* استفاده شده است. دلیل استفاده از روش *DEA* در کنار روش‌های تصمیم‌گیری چندمعیاره مانند تاپسیس و ویکور را می‌توان رویکرد متفاوت این روش در ارزیابی عملکرد دانست. باتوجه به این‌که مدل‌های تحلیل پوششی داده‌ها به ارزیابی عملکرد از منظر کارایی نسبی می‌پردازند، در مبانی نظری به تعدد در کنار تکنیک‌های *MCDM* به‌کار گرفته شده‌اند.

بنابراین، براساس نتایج حاصله می‌توان استفاده از روش‌های *MCDM* را یک ابزار مفید برای ارزیابی الگوریتم‌های خوشه‌بندی دانست که مبتنی‌بر ترکیبی از معیارهای ارزیابی بااهمیت می‌باشند. بررسی یافته‌های حاصل نشان می‌دهد که رتبه‌بندی ارایه‌شده توسط روش‌های تاپسیس و ویکور با درنظر گرفتن هفت شاخص ارزیابی اعتبار برای انتخاب بهترین الگوریتم خوشه‌بندی یکسان است و هر دو روش با قاطعیت الگوریتم خوشه‌بندی *K-Means* را نسبت به سایر الگوریتم‌های منتخب بهتر دانسته‌اند. به‌علاوه براساس جمع‌بندی مدل‌های *CCR* و *BCC* تحلیل پوششی داده‌ها، الگوریتم‌های *K-Means*، وارد و *COBWEB* در خوشه‌بندی داده‌های کارا شناخته شدند و دو الگوریتم *EM* و مبتنی‌بر چگالی ناکارا بودند. درمجموع، براساس یافته‌های پژوهش می‌توان نتیجه گرفت بهترین و کاراترین الگوریتم برای خوشه‌بندی شرکت‌ها زمانی‌که عملکرد مالی آن‌ها مورد ارزیابی قرار می‌گیرد، الگوریتم *K-Means* با ۲ خوشه است.

نتایج حاصل از پژوهش به تصمیم‌گیرندگان، تحلیلگران، پژوهشگران و سرمایه‌گذاران حوزه‌های مالی به‌منظور تحلیل بهتر شرکت‌ها موردبررسی کمک خواهد نمود، چراکه استفاده از الگوریتم مناسب و خوشه‌بندی صحیح شرکت‌ها منجر به انتخاب سبد سرمایه‌گذاری مطلوب می‌شود. در راستای مطالعه انجام‌شده برای پژوهش‌های آتی پیشنهاد می‌گردد، تعمیم‌پذیری نتایج این پژوهش را با استفاده از سایر روش‌های خوشه‌بندی، ارزیابی اعتبار خوشه‌بندی، روش‌های *MCDM* و داده‌ها با بعد بالاتر مورد آزمون قرار گیرد و هم‌چنین پیشنهاد می‌گردد قبل از خوشه‌بندی هر نوع داده، مناسب‌ترین الگوریتم خوشه‌بندی، توسط ترکیبی از شاخص‌های ارزیابی مورد آزمون قرار گیرد.

منابع

[1] Li, C., Chen, Y., & Shang, Y. (2022). A review of industrial big data for decision making in intelligent manufacturing. *Engineering science and technology, an international journal*, *29*, 101021. https://www.sciencedirect.com/science/article/pii/S2215098621001336

[2] Iqbal, R., Doctor, F., More, B., Mahmud, S., & Yousuf, U. (2020). Big data analytics and computational Intelligence for cyber--physical systems: recent trends and state of the art applications. *Future generation computer systems*, *105*, 766–778.

[3] Huang, Y., Gao, Y., Gan, Y., & Ye, M. (2021). A new financial data forecasting model using genetic algorithm and long short-term memory network. *Neurocomputing*, *425*, 207–218.

[4] Iqbalnia, M., Pouyanfar, A., & Maleki, M. (2015). Equilibrium modeling of stocks in Tehran stock exchange using a three-stage clustering approach. *Financial management perspective*, *5*(11), 133–158. (**In Persian**). https://jfmp.sbu.ac.ir/article_94651.html?lang=fa

[5] Salehi Vaziri, S. M., & Barzaghi Khaneghah, J. (2020). Investigating the effect of different data clustering methods on the accuracy of models related to accounting estimates by comparing traditional and classical clustering methods. *Management accounting*, *13*(44), 165–178. (**In Persian**). https://jma.srbiau.ac.ir/article_15515_8892be7c2957d2bad0b53a712c54f5ca.pdf

[6] Rahman, S. H. (2003). Modelling of international market selection process: a qualitative study of successful Australian international businesses. *Qualitative market research: an international journal*, *6*(2), 119–132.

[7] Nachev, A., Hill, S., Barry, C., & Stoyanov, B. (2010). Fuzzy, distributed, instance counting, and default artmap neural networks for financial diagnosis. *International journal of information technology & decision making*, *9*(06), 959–978.

[8] Zhang, Z., Liu, Z., Martin, A., Liu, Z., & Zhou, K. (2021). Dynamic evidential clustering algorithm. *Knowledge-based systems*, *213*, 106643. https://www.sciencedirect.com/science/article/pii/S0950705120307723

[9] Yu, H., Chen, L., & Yao, J. (2021). A three-way density peak clustering method based on evidence theory. *Knowledge-based systems*, *211*, 106532. https://www.sciencedirect.com/science/article/pii/S0950705120306614

[10] Sadeghi, H., & Forooghi Dehnavi, S. (2017). Codification of dendrograms portfolio based on Euclidean distance measure (a comparison between different methods of hierarchical clustering). *Financial knowledge of security analysis (financial studies)*, *10*(34), 89–105. (**In Persian**). https://jfksa.srbiau.ac.ir/article_10606.html?lang=en

[11] Serafrazi, A. (2018). Half a century after clustering; investigation and evaluation of clustering approaches and methods with multi-criteria decision analysis. *Research in science, engineering and technology*, *4*(2), 65–84. (**In Persian**). https://www.noormags.ir/view/fa/articlepage/

[12] Adel, A., Mahdavi Rad, A., & Mousakhai, M. K. (2015). Designing a combined model of data mining and multi-criteria decision making (case study: Iran statistics center subsidies database). *Journal of operational research and its applications*, *12*(1), 95-111. (**In Persian**). http://jamlu.liau.ac.ir/article-1-1045-fa.html

[13] Mirakbari, Z., Mojavarian, S. M., Rafiei, H., & Amirnejad, H. (2020). Clustering of Iran pistachio export target countries based on combined hyper-innovative algorithms. *Research in economics and agricultural development of iran*, *51*(3), 413–427. (**In Persian**). DOI:10.22059/ijaedr.2018.263272.668633

[14] Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information sciences*, *275*, 1–12.

[15] Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine learning*, *2*, 139–172. https://link.springer.com/article/10.1007/BF00114265

[16] Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial intelligence*, *40*(1–3), 11–61.

[17] Zadedehbalaei, A., Bagheri, A., & Afshar, H. (2017). A study on DBSCAN clustering algorithm issues and a survey on its improvements. *Soft computing journal*, *6*(1), 2–37. (**In Persian**). https://scj.kashanu.ac.ir/article_111412_en.html?lang=fa

[18] Kazemi, R., & Porhemmat, J. (2018). Investigating the effect of hierarchical clustering methods on accurately modeling of runoff coefficient in Karkheh Basin. *Watershed engineering and management*, *10*(1), 81–94. (**In Persian**). https://jwem.areeo.ac.ir/article_115724.html?lang=en

[19] Luna-Romera, J. M., Martínez-Ballesteros, M., García-Gutiérrez, J., & Riquelme, J. C. (2019). External clustering validity index based on chi-squared statistical test. *Information sciences*, *487*, 1–17. https://doi.org/10.1016/j.ins.2019.02.046

[20] Wang, W., & Zhang, Y. (2007). On fuzzy cluster validity indices. *Fuzzy sets and systems*, *158*(19), 2095–2117.

[21] Fazel Zarandi, M. H., Ghazanfar Ahri, S., & Ghafari Nasab, N. (2012). A new exponential cluster validity index using Jaccard distance. *Industrial management studies*, *10*(27), 22–43. (**In Persian**). https://jims.atu.ac.ir/article_1901.html

[22] Momeni, M. (2018). *Data clustering (cluster analysis)*. Mansoor Momeni Publication. (**In Persian**). https://www.gisoom.com/book/1761036/

[23] Sumathi, S., & Grace, H. G. (2020). Withdrawn: a novel distance measure for microarray dataset using entropy. *Materials today: proceedings*. https://doi.org/10.1016/j.matpr.2020.10.520

[24] Rokach, L. (2010). Ensemble-based classifiers. *Artificial intelligence review*, *33*, 1–39.

[25] Homayounfar, M., & Amirteimoori, A. R. (2019). Balanced evaluation of suppliers performance by applying a hybrid DEMATEL-DEA approach in presence of undesirable factors. *Journal of new researches in mathematics*, *5*(18), 31–48. (**In Persian**). https://jnrm.srbiau.ac.ir/article_14279.html?lang=fa

[26] Shariati, R., & Afkhami Ardakani, M. (2016). Identifying and prioritizing the performance evaluation indicators of research and development centers based on the balanced scorecard model. *Scientific-promotional monthly of oil and gas exploration and production*, *137*, 25–32. (**In Persian**). https://ekteshaf.nioc.ir/article-1-1920-fa.html

[27] Ergul, N., & Seyfullahogullari, C. A. (2012). The ranking of retail companies trading in ISE. *European journal of scientific research*, *70*(1), 29–37.

[28] Nikbakht, M. reza, Firooznia, A., & Kalhornia, H. (2019). The relationship between earnings per share to price ratio (E / P) and future earnings growth. *Empirical studies in financial accounting*, *16*(61), 55–78. (**In Persian**). DOI:10.22054/qjma.2019.22686.1621

[29] Shakeri, M. T., Sabaghian, E., & Esmaeili, H. (2012). CCK (clustering-classification-kappa) a new validation index to assessing clustering results of gene expression data. *North khorasan university of medical sciences*, *3*(5), 67–78. DOI:10.29252/jnkums.3.5.S5.67

[30] Shakri, M., & Abdulahi, M. (2015). *Investigating the impact of different data clustering methods on the accuracy of models related to accounting estimates by comparing traditional and classical clustering methods* [presentation]. International conference on applied research in information technology, computer and communication.

[31] Dehghan Nayeri, M. (2017). A new cluster validity index based on fuzzy cardinality. *Modern research in decision making*, *2*(3), 99–122.

[32] Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Professional.

[33] Hirano, S., & Tsumoto, S. (2010). Multiscale comparison and clustering of three-dimensional trajectories based on curvature maxima. *International journal of information technology & decision making*, *9*(6), 889–904.

[34] McNicholas, P. D. (2016). Model-based clustering. *Journal of classification*, *33*, 331–373.

[35] Kimes, P. K., Liu, Y., Neil Hayes, D., & Marron, J. S. (2017). Statistical significance for hierarchical clustering. *Biometrics*, *73*(3), 811–821. DOI:10.1111/biom.12647

[36] López-Rubio, E., Palomo, E. J., & Ortega-Zamorano, F. (2018). Unsupervised learning by cluster quality optimization. *Information siences*, *436*, 31–55.

[37] Renjith, S., Sreekumar, A., & Jathavedan, M. (2020). Performance evaluation of clustering algorithms for varying cardinality and dimensionality of data sets. *Materials today: proceedings*, *27*, 627–633.

[38] Hassan, B. A., Rashid, T. A., & Mirjalili, S. (2021). Performance evaluation results of evolutionary clustering algorithm star for clustering heterogeneous datasets. *Data in brief*, *36*, 107044. https://www.sciencedirect.com/science/article/pii/S2352340921003280

[39] Lossio-Ventura, J. A., Gonzales, S., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., & Bian, J. (2021). Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artificial intelligence in medicine*, *117*, 102096. DOI:10.1016/j.artmed.2021.102096

[40] Keršuliene, V., & Turskis, Z. (2011). Integrated fuzzy multiple criteria decision making model for architect selection. *Technological and economic development of economy*, *17*(4), 645–666.

[41] Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, *2*(6), 429–444.

[42] Bagheri Mazraeh, N., Daneshvar, A., & Madanchi Za, M. (2022). Development a new ensemble learning approach for stock portfolio selection using multiclass SVM and genetic algorithm. *Journal of fnancial engineering and securities management*, *13*(50), 282–305. (**In Persian**). https://fej.ctb.iau.ir/article_692412.html?lang=en

[43] Banker, R., Chen, J. Y. S., & Klumpes, P. (2016). A trade-level DEA model to evaluate relative performance of investment fund managers. *European journal of operational research*, *255*(3), 903–910.

[44] Hamidizadeh, M. R., & Shahab Al-Dini, M. (2015). Explanation of efficiency and analysis of returns in relation to the scale of the country's electricity industry. *Business management quarterly*, *26*. (**In Persian**). https://journals.iau.ir/article_525559_1853b9c5aa21ebea0279b2d3401be5be.pdf

[45] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

[46] Zavadskas, E. K., Mardani, A., Turskis, Z., Jusoh, A., & Nor, K. M. (2016). Development of TOPSIS method to solve complicated decision-making problems - an overview on developments from 2000 to 2015. *International journal of information technology and decision making*, *15*(3), 645–682. DOI:10.1142/S0219622016300019