ارائهٔ روش مبتنی بر الگوریتم ژنتیک برای مسئلهٔ یافتن پایدارترین خوشه‌ها در خوشه‌بندی ترکیبی

نوید صمیمی بهبهان[۱]، صمد نجاتیان[۲*]، حمید پروین[۳]، کرم‌اله باقری‌فرد[۴]، وحیده رضایی [۵]

گروه مهندسی کامپیوتر، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران [۱]

گروه مهندسی برق، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران [۲]

گروه مهندسی کامپیوتر، واحد نورآباد ممسنی، دانشگاه آزاد اسلامی، نورآباد ممسنی، ایران [۳]

گروه مهندسی کامپیوتر، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران [۴]

گروه ریاضی، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران [۵]

چکیده

خوشه‌بندی نقش حیاتی در روش‌های بازیابی اطلاعات برای سازمان‌دهی مجموعه‌های بزرگ، درونِ تعداد کمی خوشهٔ معنادار دارد. یکی از مهم‌ترین انگیزه‌های استفاده از خوشه‌بندی، تعیین و آشکارکردن ساختار ذاتی و پنهان یک مجموعه‌داده است. کاربران انسانی به‌علت تفاوت در سلیقه و طرز تفکرات مختلف از کشف ساختار ذاتی و درونی مجموعه‌داده‌ای بزرگ متون ناتوان‌اند. الگوریتم‌های خوشه‌بندی ترکیبی چند الگوریتم خوشه‌بندی را با هم ترکیب می‌کنند تا درنهایت به یک سامانهٔ کلی خوشه‌بندی برسند. روش‌های خوشه‌بندی ترکیبی برای یافتن راه‌های بهتری با استفاده از بیرون‌کشیدن اطلاعات از چندین افراز اولیهٔ داده‌هاست. ازآنجاکه الگوریتم‌های خوشه‌بندی مختلف به نقاط مختلف داده نگاه می‌کنند، آن‌ها می‌توانند افرازهای مختلفی را از این‌چنین داده‌هایی تولید کنند؛ با ترکیب افرازهای به‌دست‌آمده از الگوریتم‌های مختلف، ایجاد یک افراز با کارایی بالا ممکن است، حتی اگر خوشه‌ها از هم بسیار متراکم باشند. در این مقاله، روشی جدید معرفی شده‌است که به‌جای استفاده از تمامی خوشه‌های اولیهٔ تولیدشده، از پایدارترین آن‌ها که توسط شش روش مختلف تولید شده‌اند، استفاده می‌کند. برای انتخاب خوشه‌های پایدارتر از تابع توافقی مبتنی بر ماتریس هم‌بستگی استفاده می‌شود. انتخاب خوشه‌های پایدارتر بر اساس معیار پایداری خوشه مبتنی بر معیار فیشر انجام می‌گیرد و سپس خوشه‌های به‌دست‌آمده به‌وسیلهٔ الگوریتم ژنتیک ارزیابی و طبق این الگوریتم پایدارترین خوشه‌ها انتخاب می‌شوند؛ درنهایت ماتریس هم‌بستگی به‌دست‌آمده از اجماع خوشه‌های بهینه، به‌عنوان یک ماتریس مشابهت در نظر گرفته می‌شود. یک الگوریتم خوشه‌بندی سلسله‌مراتبی به‌عنوان تابع جمع‌کنندهٔ نهایی در نظر گرفته می‌شود و ماتریس هم‌بستگی به‌دست‌آمده را به‌عنوان ورودی گرفته و خوشه‌بندی توافقی نهایی را برمی‌گرداند. نتایج تجربی روی چندین مجموعه‌داده نشان می‌دهد که روش پیشنهادی، خوشه‌های متنوع و با پایداری بالا تولید می‌کند. به‌طور مشخص، این روش در معیارهای NMI و ARI به‌ترتیب بهبودهای قابل توجهی به میزان ۱۲٪ و ۵٪ نسبت به بهترین روش‌های پیشین به دست آورده‌است. این، نشان‌دهندهٔ برتری روش خوشه‌بندی ترکیبی پیشنهادی مبتنی بر پایداری خوشه و الگوریتم‌های ژنتیک است.

واژگان کلیدی: خوشه‌بندی ترکیبی، پایداری خوشه، معیار فیشر، ماتریس هم‌بستگی، الگوریتم ژنتیک.

Presenting a Method based on Genetic Algorithm for finding the most Stable Clusters in Ensemble Clustering

Navid Samimi Behbahan[1], Samad Nejatian[2*], Hamid Parvin[3], Karamolah Bagheri Fard[4], Zahra Rezaei[5]

Department of Computer Engineering, Yasuj Branch, Islamic Azad University, Yasuj, Iran[1]

Department of Electrical Engineering, Yasuj Branch, Islamic Azad University, Yasuj, Iran[2]

Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Yasuj, Iran[3]

Department of Computer Engineering, Yasuj Branch, Islamic Azad University, Yasuj, Iran[4]

Department of Mathematic, Yasuj Branch, Islamic Azad University, Yasuj, Iran[5]

[*] Corresponding author

[*] نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۹/۱۲/۲۹ • تاریخ پذیرش: ۱۴۰۳/۶/۱ • تاریخ انتشار: ۱۴۰۳/۱۰/۲۸ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.3.111]

Abstract

Clustering is one of the fundamental tools in data analysis and data mining, enabling the extraction of hidden and meaningful structures from large datasets by grouping data based on intrinsic similarities. However, selecting optimal clusters in conventional clustering algorithms poses challenges, especially when clusters are dense or heterogeneous. In this study, a novel genetic algorithm-based method is proposed to identify the most stable clusters in ensemble clustering. By leveraging cluster stability criteria and a correlation matrix, the proposed approach improves the accuracy and stability of the final clustering results. The proposed method involves generating initial partitions of the data using six different clustering algorithms. Next, the Fisher criterion is applied to identify more stable clusters. These selected clusters are then evaluated and optimized using a genetic algorithm to construct an optimized correlation matrix. This matrix is subsequently fed into a hierarchical clustering algorithm, which produces the final consensus clustering. The proposed method was tested on standard datasets. Results demonstrated improvements of 12% and 5% in NMI and ARI metrics, respectively, compared to previous methods. The use of a genetic algorithm enabled the identification of clusters with higher stability and diversity, reducing the impact of noise and increasing the accuracy of the final clustering. Moreover, the method outperformed individual base clustering algorithms in providing more precise clustering results. Due to its ability to enhance the accuracy and stability of clustering, the proposed method holds potential for applications in domains such as big data analysis, machine learning, and information retrieval. The use of the Fisher criterion for selecting stable clusters and genetic algorithms for optimization are among the strengths of this research. This method not only preserves diversity among clusters but also significantly enhances clustering accuracy. Future studies could explore the combination of this approach with more advanced algorithms to assess its applicability to more complex datasets.

Keywords: Ensemble clustering, Cluster Stability, Fisher Criterion, Correlation matrix, Genetic Algorithm.

۱-مقدمه

امروزه داده‌کاوی به‌عنوان روشی برای استخراج دانش‌های غیربدیهی، ناشناخته و باارزش از حجم عظیمی از داده‌ها شناخته می‌شود[۱و۲]. هدف داده‌کاوی، استخراج مفاهیم یا دانش موجود در این داده‌هاست؛ به‌گونه‌ای که این دانش قابل دسترس و فهم باشد و در تصمیم‌گیری‌های آینده به‌کار رود.

اهداف داده‌کاوی درکل به چند گروه تقسیم می‌شوند:

پیش‌بینی: هدف ساختن مدلی است برای پیش‌بینی مقادیر ویژگی‌های معین. داده‌های ورودی برای مدل‌سازی پیش‌گویانه شامل دو نوع ویژگی یا متغیرند:

۱) متغیرهای توضیحی که ویژگی‌هایی هستند که برای انجام پیش‌بینی استفاده می‌شوند؛

۲) متغیر هدف که ویژگی است که مقدار آن باید پیش‌بینی شود. این وظیفه به دو دستهٔ کوچک‌تر تقسیم می‌شود: رده‌بندی (برای پیش‌بینی مقادیر یک ویژگی گسسته) و برازش[1] (برای پیش‌بینی مقادیر یک ویژگی پیوسته).

درکل، الگوریتم‌های داده‌کاوی به چند دسته تقسیم می‌شوند که دو دستهٔ اصلی آن‌ها عبارت‌اند از:

رده‌بندی[2]: این فرایند شامل پیش‌بینی برچسب‌ها برای داده‌ها بر اساس داده‌های از پیش برچسب‌خورده است[۲]. رده‌بندی به دنبال یافتن مدل (تابعی) است که داده‌ها را توصیف کند و رده‌های آن‌ها را تشخیص دهد. از این مدل برای شناسایی برچسب رده‌های ناشناخته استفاده می‌شود، که بر اساس تحلیل داده‌های آموزشی (داده‌هایی با برچسب رده مشخص) به‌دست می‌آید.

خوشه‌بندی: در مقابل رده‌بندی که برچسب‌های داده‌ها را تحلیل می‌کند، خوشه‌بندی تحلیل داده‌های بدون برچسب است[۲و ۳]؛ در این حالت، برچسب‌های رده در داده‌های آزمایشی به‌سادگی وجود ندارند؛ زیرا ممکن است، ناشناخته باشند. خوشه‌بندی به تولید چنین برچسب‌هایی کمک می‌کند. اشیای خوشه‌بندی‌شده بر اساس اصل بیشترکردن شباهت درون رده‌ها و کمینه‌کردن شباهت بین رده‌ها رده‌بندی می‌شوند. اشیای درون هر خوشه بالاترین شباهت را به یکدیگر، اما بیشترین تفاوت را با اشیا در رده‌های دیگر دارند. هر خوشه می‌تواند به‌عنوان رده آن دسته از اشیا درنظر گرفته شود و قوانین مربوط به آن‌ها از آن استخراج می‌شود. از الگوریتم‌های خوشه‌بندی می‌توان به الگوریتم‌های خوشه‌بندی سلسله‌مراتبی، نگاشت خودسازمان‌ده[3]، K-Means، و k-Medoid اشاره کرد.

خوشه‌بندی درواقع به معنی یافتن ساختار در مجموعه‌های داده‌ای است که هنوز گروه‌بندی یا رده‌بندی نشده‌اند. به عبارت دیگر، خوشه‌بندی فرایندی است که داده‌ها را بر اساس شباهت‌های خاص به گروه‌هایی تقسیم می‌کند؛ در این گروه‌ها، شباهت اعضای داخل یک خوشه بر اساس یک معیار خاص از یک مقدار آستانه بیشتر و شباهت آن‌ها با اعضای سایر خوشه‌ها کمتر از این آستانه

[1] Regression
[2] classification
[3] Self-Organized Mapping

[DOI: 10.61186/jsdp.21.3.111]

است. محاسبهٔ فاصلهٔ بین داده‌ها در خوشه‌بندی بسیار مهم است و تاکنون معیارهای مختلفی برای ارزیابی خوشه‌بندی و عملکرد آن تعیین شده‌است. پژوهش‌گران با استفاده از این معیارها به ارزیابی خوشه‌بندی‌های مختلف پرداخته‌اند. با محاسبهٔ فاصلهٔ بین دو داده می‌توان فهمید چقدر این دو داده به هم نزدیکند و بر این اساس آن‌ها را در یک خوشه قرار می‌دهیم. معیار شباهت می‌تواند فاصله، تقابل اطلاعاتی[1]، کوواریانس[2] و... باشد[۴ و ۵].

روند کلی خوشه‌بندی شامل مراحل زیر است[۳]:

۱- بازنمایی الگوها

۲- تعریف معیار ارزیابی شباهت بر اساس داده‌ها، بررسی هم‌بستگی، فاصله و اطلاعات مشترک بین داده‌ها

۳- خوشه‌بندی یا گروه‌بندی

۴- خلاصه‌سازی داده‌ها (در صورت نیاز)

۵- ارزیابی خروجی

۶- حذف نمونه‌های نوفه‌ای و یا کاهش بُعد (در صورت نیاز)

نمایش الگوها به تعداد رده‌ها، تعداد نمونه‌های موجود، نوع و مقیاس ویژگی‌های موجود در الگوریتم خوشه‌بندی اشاره دارد. برخی از این اطلاعات توسط کاربر قابل کنترل نیستند. مجاورت نمونه‌ها به‌طور معمول با استفاده از یک تابع فاصله بین زوج الگوهای ورودی اندازه‌گیری می‌شود. معیارهای متنوعی برای اندازه‌گیری فاصله در حوزه‌های مختلف استفاده می‌شود[۳، ۶ و ۷]. یک معیار ساده مانند فاصلهٔ اقلیدسی به‌طور معمول برای نمایش عدم‌تشابه بین دو الگو به‌کار می‌رود، اما برای تشخیص شباهت‌های مفهومی بین الگوها می‌توان از معیارهای ارزیابی دیگری نیز استفاده کرد[۸]. ارزیابی خروجی برای تعیین معنی‌داربودن خروجی به‌کار می‌رود و یک ساختار خوشه‌بندی زمانی معتبر است که به‌صورت تصادفی اتفاق نیفتد[۹].

به‌دلیل نبود دانش و اطلاعات اولیه دربارهٔ اشکال خوشه‌ها، انتخاب یک روش خوشه‌بندی برای هر مجموعه‌داده‌ای آسان نیست. با این دانش، ممکن است به فکر ترکیب روش‌های مختلف خوشه‌بندی با اشیای مختلف باشیم و به همین دلیل به استفاده از روش‌های خوشه‌بندی ترکیبی[3] برای یافتن روش‌های بهتر روی می‌آوریم.

چندین روش خوشه‌بندی وجود دارد، یکی از این روش‌ها خوشه‌بندی افرازی[4] است. در این روش، مجموعه‌ای از N نقطه به K خوشهٔ مجزا تقسیم می‌شود به‌گونه‌ای که هر نقطه تنها در یک خوشه قرار می‌گیرد و خوشه‌ها هم‌پوشانی[5] ندارند. روش‌های خوشه‌بندی افرازی به‌دلیل نیاز به محاسبات تاحدودی کم، در بسیاری از پژوهش‌های گذشته مورداستفاده قرار گرفته‌اند. نکتهٔ مهم در این روش‌ها، انتخاب تعداد خوشه‌ها و روش‌های تغییر اعضای خوشه‌ها برای به‌دست‌آوردن نتایج بهتر است؛ به همین دلیل، روش‌های رده‌یافتی برای تخمین تعداد خوشه‌ها ارائه شده‌اند؛ از جملهٔ این روش‌ها، K-Means است که بر اساس کمینه‌سازی جمع مربعات فواصل عمل می‌کند.

روش‌های خوشه‌بندی افرازی با فشرده‌سازی اطلاعات و رقمی‌کردن بردارها ارتباط نزدیکی دارند. الگوریتم خوشه‌بندی افرازی محبوب، K-Means، برای نخستین بار توسط استینهاوس در سال ۱۹۵۶ مطرح شد. الگوریتم‌های خوشه‌بندی افرازی، برخلاف روش‌های سلسله‌مراتبی که یک ساختار خوشه‌بندی را ارائه می‌کنند، تنها یک تقسیم‌بندی از داده‌ها را نشان می‌دهند. این روش‌ها از نظر سرعت اجرا نسبت به سایر روش‌ها بهترند، اما دارای نقاط ضعفی نیز هستند؛ از جمله اینکه نتایج نهایی به انتخاب خوشه‌های اولیه وابسته‌اند و همواره بهترین پاسخ را ارائه نمی‌دهند.

خوشه‌بندی روشی برای رده‌بندی الگوها به گروه‌های مختلف است. این الگوها می‌توانند شامل مشاهدات، اقلام داده‌ای یا برداری از مشخصه‌ها[6] باشند که درنهایت در گروه‌هایی قرار می‌گیرند که به آن‌ها «خوشه‌ها» گفته می‌شود. همان‌طور که در بخش‌های پیشین اشاره شد، خوشه‌بندی، ابزار مهمی در پیمایش[7] و تحلیل داده-هاست و در زمینه‌هایی چون تحلیل الگوهای اکتشافی[8]، تصمیم‌گیری، یادگیری ماشین[9] و دیگر موضوعات مانند داده‌کاوی[10]، بازیابی اسناد[11]، قطعه‌سازی تصاویر[12] و رده‌بندی الگو[13] کاربرد دارد[۳].

گاهی‌اوقات تقسیم‌بندی داده‌ها به چندین گروه مشکل است؛ یکی از دلایل این پیچیدگی این است که برای برخی مجموعه‌های داده‌ای، افراز بدون ابهام وجود ندارد یا اینکه حتی توسط انسان نیز قابل ساخت نیست[۱۰]. حتی در مواردی که این مشکل وجود ندارد، برخی الگوریتم‌های خوشه‌بندی با شکست مواجه می‌شوند. این به‌دلیل آن است که بیشتر الگوریتم‌های خوشه‌بندی موجود بر اساس فقط یک تابع ارزیابی درونی عمل می‌کنند. این تابع هدف

[1] Mutual Information
[2] Covariance
[3] Ensemble Clustering
[4] Partitioning Clustering
[5] Overlapping

[6] Features Vector
[7] Exploration
[8] Exploratory Pattern Analysis
[9] Machine Learning
[10] Data Mining
[11] Document Retrieval
[12] Image Segmentation
[13] Pattern Classification

[DOI: 10.61186/jsdp.21.3.111]

خصوصیات باطنی افراز را مانند جداپذیری بین خوشه‌ها یا تراکم درون هر خوشه اندازه‌گیری می‌کند. با اجرای الگوریتم‌های مختلف خوشه‌بندی بر روی مجموعه‌های داده، افرازهای گوناگونی به‌دست می‌آید که باید بر اساس تابع هدف، مناسب‌ترین افراز انتخاب شود.

نوآوری اساسی در این مقاله، ارائهٔ یک روش بهینه‌سازی برای بهبود کارایی خوشه‌بندی ترکیبی معرفی شده‌است. هدف این مقاله، رفع مشکلی است که در بند قبلی توضیح داده شد و شامل ارائهٔ یک روش بهینه‌سازی خوشه‌بندی ترکیبی بر روی مجموعه‌داده‌هاست. اهداف زیر در این مقاله دنبال می‌شوند:

- انتخاب پایدارترین خوشه‌ها با استفاده از معیار فیشر: در این پژوهش، به جای استفاده از تمامی خوشه‌های اولیهٔ تولیدشده، از پایدارترین آن‌ها که به‌وسیلهٔ شش روش مختلف تولید شده‌اند، استفاده می‌شود. معیار فیشر برای انتخاب خوشه‌های پایدارتر به‌کار گرفته می‌شود که از طریق ارزیابی پایداری هر خوشه، خوشه‌هایی را که پایداری بالاتری دارند انتخاب می‌کند. این نوآوری منجر به افزایش دقت و پایداری خوشه‌بندی نهایی می‌شود.
- استفاده از الگوریتم ژنتیک برای بهبود انتخاب خوشه‌ها: پس از انتخاب اولیهٔ خوشه‌ها با استفاده از معیار فیشر، خوشه‌های به‌دست‌آمده به‌وسیلهٔ الگوریتم ژنتیک مورد ارزیابی قرار می‌گیرند. الگوریتم ژنتیک با بهینه‌سازی مجموعه خوشه‌ها، پایدارترین خوشه‌ها را انتخاب می‌کند. این مرحله بهبود قابل توجهی در کیفیت خوشه‌بندی نهایی ایجاد می‌کند.
- استفاده از ماتریس هم‌بستگی به‌عنوان ورودی برای خوشه‌بندی سلسله‌مراتبی: ماتریس هم‌بستگی به‌دست‌آمده از اجماع خوشه‌های بهینه، به‌عنوان یک ماتریس مشابهت در نظر گرفته می‌شود. یک الگوریتم خوشه‌بندی سلسله‌مراتبی به‌عنوان تابع جمع‌کنندهٔ نهایی در نظر گرفته می‌شود که این ماتریس را به‌عنوان ورودی گرفته و خوشه‌بندی توافقی نهایی را برمی‌گرداند. این روش به حفظ تنوع و افزایش پایداری خوشه‌های نهایی کمک می‌کند.
- بهبود معیارهای ارزیابی خوشه‌بندی: نتایج تجربی نشان می‌دهد که روش پیشنهادی در معیارهای NMI و ARI به ترتیب بهبودهای قابل توجهی به میزان ۱۲٪ و ۵٪ نسبت به بهترین روش‌های پیشین به‌دست آورده‌است؛ این بهبودها نشان‌دهندهٔ افزایش دقت و کارایی روش پیشنهادی در مقایسه با روش‌های موجود است.

ساختار مقاله: این مقاله به پنج بخش تقسیم شده‌است: در بخش دو خلاصه‌ای از مفاهیم خوشه‌بندی و خوشه‌بندی ترکیبی و الگوریتم‌های این حوزه معرفی، در بخش سه الگوریتم‌های پیشنهادی مطرح و مجموعه‌داده‌های به‌کاررفته معرفی شده‌است و در بخش چهار، نتایج کلی حاصل از پژوهش توضیح داده می‌شود؛ همچنین نتیجه‌گیری و کارهای آینده در بخش پنج آمده‌اند.

۲-مرور منابع

پیشرفت‌های اولیه در خوشه‌بندی اسناد: برای نخستین‌بار در سال ۱۹۷۱، گاردین و ون ریجزبرگر[1] با استفاده از چندین آزمایش نشان دادند که خوشه‌بندی اسناد می‌تواند کارایی سامانه‌های بازیابی اطلاعات را بهبود ببخشد. انتظار می‌رود که کارایی این سامانه‌ها با استفاده از خوشه‌بندی افزایش یابد؛ زیرا پژوهش‌های ویلت[2] در سال ۱۹۸۵ نشان داد که خوشه‌بندی توجه به ارتباط بین اسناد مجموعه را افزایش داده و اسناد مرتبطی را که در جست‌وجوهای معمول در انتهای فهرست قرار می‌گیرند، به هم نزدیک می‌کند. این امر باعث می‌شود بازیابی اسناد در رتبه‌های بالاتر انجام شود و درنتیجه کارایی سامانه افزایش یابد.

فرضیهٔ اساسی خوشه‌بندی بیان می‌کند که اسناد مرتبط با یک پرسش، نسبت به اسناد غیرمرتبط، تمایل بیشتری به شبیه‌بودن دارند؛ بنابراین در یک خوشه قرار می‌گیرند. اعمال این فرضیه بر یک مجموعه اسناد خاص، تفکیک خوبی بین اسناد مرتبط و غیرمرتبط ایجاد می‌کند[۱۱ و ۱۲].

پیشرفت‌های بعدی: تامبروس[3] [۴۴] در سال ۲۰۰۲ فرمولی برای محاسبهٔ شباهت‌های بین اسناد به‌دست آورد و آن را معیار شباهت حساس به پرسش نامید. این معیار در سال ۲۰۰۶ توسط ولی‌زاده و ذوالقدری [۴۵] ارتقا یافت. آن‌ها با استفاده از آزمون N نزدیک‌ترین همسایگی[4] ثابت کردند که معیار پیشنهادی‌شان نسبت به دیگر معیارها عملکرد بهتری دارد.

۲-۱-الگوریتم‌های خوشه‌بندی تک‌منظوره

الگوریتم‌های خوشه‌بندی تک‌منظوره را می‌توان از چند جنبهٔ مختلف تقسیم‌بندی کرد:

الف) بر اساس نوع تعلق داده‌ها به خوشه‌ها:

خوشه‌بندی انحصاری[5]: در این روش، هر داده پس از خوشه‌بندی به‌طور دقیق به یک خوشه تعلق می‌گیرد. مثالی از این روش، خوشه‌بندی K-Means است.

[1] Jardin & Van Rijsbergen
[2] Willett
[3] Tombros
[4] N-Nearest neighbor
[5] Exclusive or Hard Clustering

[DOI: 10.61186/jsdp.21.3.111]

خوشه‌بندی با هم‌پوشانی[1]: در این روش، به هر داده یک درجهٔ تعلق نسبت به هر خوشه داده می‌شود؛ به‌طوری‌که یک داده می‌تواند به نسبت‌های متفاوتی به چندین خوشه تعلق داشته باشد. نمونه‌ای از این روش، خوشه‌بندی فازی است[۱۳].

ب) بر اساس رویکرد ساختاری خوشه‌ها:

خوشه‌بندی سلسله‌مراتبی[2]: خوشه‌ها در سلسله‌مراتب سازماندهی می‌شوند که از سطوح بالا به پایین یا برعکس شکل می‌گیرند.

خوشه‌بندی افرازی: داده‌ها به‌صورت مجزا به خوشه‌ها تقسیم می‌شوند.

خوشه‌بندی مبتنی بر چگالی[3]: خوشه‌ها بر اساس تراکم داده‌ها در نواحی مختلف تشکیل می‌شوند.

خوشه‌بندی مبتنی بر گرید: داده‌ها در یک فضای گریدبندی‌شده تقسیم می‌شوند.

خوشه‌بندی مبتنی بر کرنل: خوشه‌بندی بر اساس تبدیل فضای ویژگی به فضایی با بعد بالاتر به‌وسیلهٔ توابع کرنل انجام می‌گیرد.

۲-۲-روش‌های خوشه‌بندی ترکیبی

الگوریتم‌های خوشه‌بندی ترکیبی، چندین الگوریتم خوشه‌بندی را با هم ترکیب می‌کنند تا به یک سامانهٔ خوشه‌بندی جامع برسند. این الگوریتم‌ها با هدف ترکیب چندین روش مختلف، می‌توانند به‌عنوان خوشه‌بندی چندمنظوره[4] در نظر گرفته شوند. در کاری که توسط توپچی و همکاران [۴۶] در سال ۲۰۰۵ ارائه شد، چندین الگوریتم خوشه‌بندی برای شناسایی خوشه‌های متفاوت به‌کار برده شده‌است و درنهایت خروجی‌های این الگوریتم‌ها به‌صورت افرازهای ترکیبی استفاده می‌شوند؛ در این روش، هر خوشه به‌وسیلهٔ یک الگوریتم خوشه‌بندی جداگانه تولید و خوشه‌های نهایی بر اساس توابع ارزیابی مستقل انتخاب می‌شود؛ درنتیجه، این الگوریتم بهترین تابع هدف را برای قسمت‌های مختلف فضای مشخصه انتخاب می‌کند[۱۴].

نجفی و همکاران [۶۲] یک روش خوشه‌بندی ترکیبی را ارائه داده‌اند که از روش خوشه‌بندی پایهٔ ضعیف C-Means فازی به‌عنوان خوشه‌بند پایه استفاده کرده‌است؛ همچنین با اتخاذ برخی تمهیدات، تنوع اجماع را بالا برده‌است. روش خوشه‌بندی ترکیبی پیشنهادی مزیت الگوریتم خوشه‌بندی C-Means فازی را دارد که این مزیت سرعت آن است؛ همچنین ضعف‌های عمدهٔ آن را که عدم قابلیت کشف خوشه‌های غیرکروی و غیریکنواخت است، ندارد.

الگوریتم‌های خوشه‌بندی ترکیبی به دنبال بهبود روش‌های استخراج اطلاعات از چندین افراز اولیهٔ داده‌ها هستند؛ زیرا هر الگوریتم خوشه‌بندی به نقاط مختلفی از داده نگاه می‌کند و افرازهای متفاوتی تولید می‌کنند. با ترکیب این افرازها، امکان ایجاد یک افراز با کارایی بالا وجود دارد، حتی اگر خوشه‌ها از یکدیگر بسیار متفاوت و متراکم باشند. درکل، دو گام اصلی در خوشه‌بندی ترکیبی وجود دارد:

گام نخست: استفاده از تعدادی الگوریتم‌های خوشه‌بندی پایه یا تکرار یک الگوریتم با مقادیر متفاوت برای ایجاد مجموعه‌ای از افرازهای اولیه که به‌عنوان یک ترکیب نامیده می‌شود. هر افراز یکتا، جنبه‌ای متفاوت از داده‌ها را نمایان می‌کند.

گام دوم: استفاده از اطلاعات جمع‌آوری‌شده در ترکیب، برای انجام خوشه‌بندی ترکیبی در مرحلهٔ بعدی که به بهینه‌سازی بیشتر کمک می‌کند.

این رویکردها به‌خصوص در مواردی که داده‌ها پیچیدگی بالایی دارند و نیاز به استخراج اطلاعات دقیق‌تری است، مفید واقع می‌شوند.

درکل الگوریتم‌های خوشه‌بندی ترکیبی را می‌توان به گروه‌های زیر تقسیم بندی کرد[۱۵ و ۱۶]:

- روش‌های مبتنی بر ماتریس هم‌بستگی
- روش‌های تئوری (نظری) اطلاعات
- روش‌های ابرگراف
- مدل مخلوطی
- خوشه‌بندی متراکم عمومی (GAC-GEO)
- ترکیب بیضی
- روش‌های رأی‌گیری

۲-۳-مروری بر روش‌های خوشه‌بندی ترکیبی

۲-۳-۱-روش‌های مبتنی بر ماتریس هم‌بستگی[5]

برای استفاده از توابع مبتنی بر هم‌بستگی ابتدا باید خوشه‌بندی ترکیبی را به ماتریس هم‌بستگی N×N تبدیل کرد. از این ماتریس هم‌بستگی نیز می‌توان به‌عنوان یک ماتریس مشابهت یادکرد که تشابه بین دو عضو x و y را محاسبه می‌کند[۱۵ و ۱۶].

در توابع مبتنی بر هم‌بستگی، می‌توان از الگوریتم‌های خوشه‌بندی سلسله‌مراتبی مبتنی بر مشابهت مانند تک‌پیوندی به‌عنوان توابع اجماع استفاده کرد. استفاده از چندین الگوریتم سلسله‌مراتبی متراکم به همراه ماتریس مشابهت، امکان دستیابی به یک افراز نهایی را فراهم

[1] Overlapping or Soft Clustering
[2] Hierarchical Clustering
[3] Density-Based Methods
[4] Multi-Objective Clustering
[5] Co-Association Matrix

[DOI: 10.61186/jsdp.21.3.111]

می‌آورد. روش‌های مختلف برای ترسیم ماتریس هم‌بستگی به چندین دسته کلی تقسیم می‌شوند؛ از جمله: تک‌پیوندی/درخت کمینه پوشا، پیوند کامل، پیوند متوسط و دیگر الگوریتم‌های مرتبط[۱۵ و ۱۶].

۲-۳-۲-روش‌های تئوری (نظری) اطلاعات[1]

معروف‌ترین الگوریتم‌های این روش، الگوریتم‌های بر پایهٔ اطلاعات متقابل درجه دوم[2] هستند. در این روش‌ها، افراز اجماع به‌عنوان یک ویژگی هدف در نظر گرفته شده و هر یک از افرازهای پایه به‌عنوان ویژگی ورودی تلقی می‌شود. پژوهشی که توسط مجرد و همکاران [۱۳] انجام شد نشان می‌دهد که افراز اجماع باید به‌گونه‌ای باشد که اطلاعات متقابل[3] بین ویژگی هدف و ویژگی‌های ورودی را به حداکثر برساند. ازآنجاکه تمامی افرازهای پایه مستقل از یکدیگرند، اطلاعات متقابل به‌صورت جفت‌های متقابل بین ویژگی هدف و افرازهای اولیه محاسبه می‌شود؛ به این ترتیب، مسئلهٔ خوشه‌بندی ترکیبی به یافتن افراز اجماعی تبدیل می‌شود که تابع هدف را بیشینه می‌کند.

اطلاعات متقابل درجه دوم یا روش‌های مبتنی بر ویژگی می‌توانند به‌وسیلهٔ الگوریتم K-Means در فضای ویژگی‌های تبدیل‌شده‌ای که از برچسب‌های خوشه‌ای ترکیبی استفاده می‌کند، به بیشترین مقدار برسند؛ در این رویکرد، خروجی هر الگوریتم خوشه‌بندی به‌عنوان یک ویژگی دسته‌بندی‌شده عمل می‌کند و مجموعه‌ای از افرازهای پایه به‌عنوان فضای ویژگی متوسط در نظر گرفته می‌شود که بر روی آن سایر الگوریتم‌های خوشه‌بندی اجرا می‌شوند.

یک تابع اجماع مبتنی بر خوشه‌بندی K-Means در فضای ویژگی‌های استانداردشده می‌تواند به‌طور مؤثر یک تعریف کلی از اطلاعات متقابل را به بیشترین مقدار برساند. پیچیدگی زمانی این تابع اجماع $O(KNB)$ است، که در آن B تعداد افرازهای اولیه، K تعداد خوشه‌ها و N تعداد داده‌هاست؛ اگرچه الگوریتم اطلاعات متقابل درجه دوم ممکن است، بالقوه در یک بهینهٔ محلی قرار گیرد، پیچیدگی محاسباتی کمابیش پایین آن امکان می‌دهد تا چندین بار بازآغاز شود تا یک راه حل اجماع با کیفیت و با کاهش واریانس درون خوشه انتخاب شود[۱۷ و ۱۸].

۲-۳-۳-روش‌های ابرگراف[4]

استرل و قوش[5] [۱۹] در سال ۲۰۰۲ مفهوم اجماع خوشه‌بندی را معرفی کردند و سه تابع اجماع مبتنی بر ابرگراف را پیشنهاد دادند. در گام نخست، افرازهای داده‌ای به ابرگراف تبدیل می‌شوند که در آن، رئوس ابرگراف نمایندهٔ نقاط داده‌ای و یال‌ها نمایندهٔ خوشه‌ها هستند؛ سپس، برای تقسیم و افراز رئوس یا نقاط داده‌ای، از الگوریتم‌های ابرگراف حداقل برش استفاده می‌شود. این فرایند که شامل حداقل k برش ابرگراف است، به k افراز منجر می‌شود که درنهایت افراز اجماع را تشکیل می‌دهد.

۲-۳-۴-مدل مخلوطی[6]

خوشه‌بندی زیرفضا، توسعه‌یافته‌ای از الگوریتم‌های انتخاب ویژگی است. این روش در تلاش است تا خوشه‌ها را در زیرفضاهای مختلفی از مجموعه‌داده‌ها شناسایی کند[۲۰ و ۲۱]. مشابه با الگوریتم‌های انتخاب ویژگی، خوشه‌بندی زیرفضا نیز به یک روش جست‌وجو و یک معیار ارزیابی نیاز دارد. درواقع، این چالش در شناسایی زیرفضاها با مشکل یافتن قوانین رابطه‌ای مرتبط است که مناطق جذاب ویژگی‌های مختلف را مشخص می‌کند[۲۰ و ۲۱]. اگر ما یک زیرمجموعه از ویژگی‌های موجود در مجموعه‌داده‌ها را در نظر بگیریم، یعنی یک زیرفضای داده‌ها، خوشه‌هایی که در طی پیشرفت یک الگوریتم خوشه‌بندی کشف می‌شوند، می‌توانند به‌شدت متفاوت باشند با آن‌هایی که در دیگر زیرفضاها پیدا می‌شوند.

۲-۳-۵-خوشه‌بندی متراکم عمومی[7]

یکی از اصول کلیدی الگوریتم GAC-GEO، بازآفرینی الگوریتم‌های خوشه‌بندی موجود با کمینه تلاش انسانی با هدف بهبود نتایج خوشه‌بندی است. GAC-GEO در روش خوشه‌بندی خود از دو فاز استفاده می‌کند: نخست، فاز پیش‌پردازش که در آن مجموعه‌ای از خوشه‌های اولیه و روابط همسایگی فضایی بین این خوشه‌ها شکل می‌گیرد؛ دوم، فاز متراکم‌سازی که در آن خوشه‌های همسایه به صورت حریصانه با یکدیگر ادغام می‌شوند تا تابع برازندگی به حداکثر برسد[۲۳].

۲-۳-۶-روش‌های مبتنی بر رأی‌گیری[8]

در الگوریتم‌های خوشه‌بندی قبلی، نیازی به حل صریح مسئلهٔ تطابق بین برچسب‌های خوشه‌های شناخته‌شده و مشتق‌شده وجود ندارد. به‌جای آن، روش رأی‌گیری برای حل این مسئلهٔ تطابق به‌کار رفته‌است که در آن از بیشترین مقدار آرا برای تعیین افراز اجماع نهایی استفاده می‌شود[۲۴]. ایدهٔ اصلی این است که برچسب‌های

[1] Information Theory Approach
[2] Quadratic Mutual Information (QMI)
[3] Mutual Information (MI)
[4] Hyper-graph methods
[5] Strehl and Ghosh
[6] Mixture model
[7] Generic Agglomerative Clustering(GAC-GEO)
[8] Voting approaches

[DOI: 10.61186/jsdp.21.3.111]

خوشه‌ها به‌گونه‌ای تغییر یابند که بهترین توافق بین برچسب‌های دو افراز به‌دست آید؛ درنهایت، تمام افرازهای ترکیبی باید بر اساس یک افراز مرجع ثابت برچسب‌گذاری شوند.

با توجه به مرور روش‌های مختلف خوشه‌بندی ترکیبی، مشخص شد که هر کدام از این روش‌ها مزایا و معایب خاص خود را دارند؛ بااین‌حال، روش‌های مبتنی بر ماتریس هم‌بستگی به‌دلیل توانایی آن‌ها در استفاده از اطلاعات جامع‌تر و انعطاف‌پذیری بالا، به‌خصوص در ترکیب با الگوریتم‌های سلسله‌مراتبی نشان داده‌اند که می‌توانند نتایج بهتری در خوشه‌بندی ترکیبی ارائه دهند. این روش‌ها قادرند تا با استفاده از معیارهای مشابهت دقیق‌تر، افرازهای بهینه‌تری را به‌دست آورند؛ بنابراین، مقالهٔ حاضر بر روی بهبود و استفاده از روش‌های مبتنی بر ماتریس هم‌بستگی متمرکز شده‌است. با استفاده از معیار فیشر برای انتخاب خوشه‌های پایدارتر و الگوریتم ژنتیک برای بهینه‌سازی، روشی ارائه شده‌است که نتایج بهتری در معیارهای NMI و ARI نسبت به روش‌های پیشین دارد. این انتخاب بر اساس نتایج تجربی و تحلیل دقیق مزایا و معایب هر روش بوده‌است.

۴-۲-روش‌های جدید در خوشه‌بندی ترکیبی

دو گرایش جدید در روش‌های خوشه‌بندی ترکیبی شامل انتخاب خوشه‌بندی ترکیبی و بهینه‌سازی خوشه‌بندی ترکیبی است. در روش نخست، ایده این است که زیرمجموعه‌ای از خوشه‌بندی‌های اولیه انتخاب شود؛ به‌گونه‌ای که افراز اجماع حاصل از آن زیرمجموعه، بهتر از ترکیب کامل خوشه‌ها باشد. در مطالعات قبلی، تمام خوشه‌ها و افرازبندی‌ها در ترکیب، دارای وزن یکسانی بوده‌اند، به این معنا که هر عضو ترکیبی دارای ارزش یکسانی در تصمیم‌گیری نهایی است.

فرن و لین[1] [۴۷] از معیار اطلاعات متقابل نرمال‌شده استفاده کرده‌اند که ابتدا توسط استرل و قوش [۱۹] تعریف و سپس توسط فرد و جین[2] [۴۸] برای ارزیابی افرازبندی‌های اولیه تکمیل شده‌است. آن‌ها با نتایج تجربی نشان داده‌اند که انتخاب یک زیرمجموعه از افرازبندی‌ها می‌تواند به نتایج بهتری نسبت به کل افرازبندی‌ها در ترکیب کلی دست یابد. فرن و لین روش خوشه‌بندی ترکیبی را پیشنهاد داده‌اند که مجموعه‌ای از افرازها را بر اساس کیفیت و پراکندگی انتخاب می‌کند؛ دو پارامتر که نشان داده‌اند بر کارایی خوشه‌بندی ترکیبی تأثیرگذارند. روش آن‌ها برای انتخاب زیرمجموعه‌های افرازهای اولیه که هم‌زمان دارای بالاترین کیفیت و بیشترین پراکندگی‌اند، طراحی شده‌است.

پروین و همکاران [۱۵] روش نوینی برای خوشه‌بندی داده‌ها ارائه داده‌اند که شامل اختصاص یک بردار وزن به فضای ویژگی می‌شود. در این روش، واریانس داده‌ها بر اساس هر ویژگی محاسبه می‌شود و ویژگی‌هایی که دارای واریانس بالاتری هستند، در ترکیب نهایی با وزن بیشتری مشارکت می‌کنند؛ همچنین، هم‌گرایی الگوریتم پیشنهادی توسط آن‌ها تأیید شده‌است. در رویکرد دوم، افراز اجماع از طریق حل یک مسئلهٔ بهینه‌سازی به‌دست می‌آید که هدف آن یافتن افراز بهینه با استفاده از یک تابع هدف مشخص است. این تابع هدف به‌طور معمول بر اساس ترکیب خوشه‌ها تنظیم می‌شود.

یکی از روش‌های رایج در این زمینه، مدل‌سازی مسئلهٔ خوشه‌بندی ترکیبی به‌صورت یک گراف است که شامل n گره (معادل تعداد داده‌ها) و چندین لبه است که تشابه‌های محاسبه‌شده بین دو گره را نمایش می‌دهد. برخلاف پیچیدگی الگوریتم به‌کاررفته، گراف نمایش‌داده‌شده ممکن است، باعث نتایج با درجهٔ مطلوبیت پایین شود.

پژوهش‌های اخیر در این زمینه بیشتر به سمت فرموله‌کردن مسئله به‌عنوان یک وظیفهٔ بهینه‌سازی و حل آن توسط حل‌کننده‌های ریاضی یا حتی حل‌کننده‌های بهینه‌سازی هوشمند معطوف شده‌اند. این رویکرد به دست‌یابی به نتایج بهینه و کارآمد کمک می‌کند که می‌تواند درجهٔ مطلوبیت بالاتری داشته باشد.

کریستوی[3] [۴۹] فرمولی مبتنی بر بهینه‌سازی را برای خوشه‌بندی ترکیبی ارائه داده که مناسب برای رده مسئله‌های دارای معیارهای درون خوشه، مانند خوشه‌بندی کمینه مجموع مربعات است. وی فرمول افرازبندی مجموعهٔ مسئلهٔ خوشه‌بندی اصلی را برای دستیابی به یک الگوریتم خوشه‌بندی ترکیبی ساده و کارآمد تغییر داده‌است. کریستوی تأکید کرده که تحت فرضیات و تسهیلات فرمول اصلی، امکان یافتن راه حل‌های بهتر نسبت به ترکیبات موجود تضمین شده‌است.

سینگ و همکاران [۵۰] فرمول بهینه‌سازی جدیدی ارائه داده‌اند که هدف آن تشکیل خوشه‌های نهایی برای بیشترکردن مقدار توافقات و کمترکردن مقدار اختلافات در نتیجهٔ اجماع است. این رویکرد، با توجه هم‌زمان به اعضای گروه، بهینه‌سازی می‌کند تا اطمینان حاصل شود که نتایج نهایی هماهنگی بیشتری با اهداف موردنظر دارند.

[1] Fern and Lin
[2] Fred and Jain
[3] Christou

[DOI: 10.61186/jsdp.21.3.111]

در پژوهش [۵۱] توجه به دو چالش عمده در الگوریتم‌های خوشه‌بندی مجموعه‌ای یعنی ساخت یک معیار ارزیابی قوی و درنظرگرفتن ارتباطات محلی بین خوشه‌ها هنگام ساخت ماتریس هم‌بستگی موردتوجه قرار گرفته‌است. جهت این چالش، یک چارچوب خوشه‌بندی گروهی جدید به نام خوشه‌بندی گروه وزنی تصادفی چندمتغیره و استراتژی پیاده‌روی تصادفی[1] ارائه می‌کند.

مزایا: بهبود استحکام و پایداری، کاربرد در مجموعه‌داده‌های متنوع

معایب: چالش‌های بالقوه در تنظیم پارامتر برای عملکرد بهینه و کاهش قابلیت تفسیر به‌دلیل پیچیدگی روش و تجمیع نتایج خوشه‌بندی چندگانه.

در مقالهٔ [۵۲] تقویت ماتریس هم‌بستگی، یک نمایش جامع‌تر از روابط داده‌ها و تخصیص وزن به نتایج خوشه‌بندی پایه که منعکس‌کنندهٔ سهم آن‌ها در نتیجهٔ مجموعه است بررسی شده‌است. روش پیشنهادی یک ماتریس جدید هم‌بستگی فازی را معرفی می‌کند که هم ارتباط هم‌بستگی و هم روابط بین خوشه‌ای را در بین نقاط داده نشان می‌دهد.

مزایا: تنظیم تأثیر نتایج خوشه‌بندی پایه بر اساس ارتباط نمونه‌ها به‌طور پویا، گروه‌بندی مستقیم (ایجاد نتیجهٔ خوشه‌بندی مجموعه‌ای بهینهٔ مستقیم از ماتریس هم‌بستگی)

معایب: پیچیدگی محاسباتی بالا، چالش در تفسیرپذیری به‌دلیل ماهیت پیچیدهٔ روش پیشنهادی

در مرجع [۵۳] به محدودیت‌های روش‌های خوشه‌بندی موجود اشاره شده است که شامل ملاحظاتی پیرامون اندازهٔ خوشه‌ها و ابتکارات جدید برای انتخاب داده‌ها می‌شود. این محدودیت‌ها در نظر گرفته شده و منجر به افزایش استحکام و دقت نتایج خوشه‌بندی شده‌اند. روشی که ارائه شد خوشه‌بندی سلسله‌مراتبی تجمعی با یادگیری نیمه‌نظارتی است که با استفاده از داده‌های محدودیتی پیوندی و معیارهای فاصلهٔ ابتکاری، عملکرد خوشه‌بندی را بهبود بخشیده است.

مزایا: دست‌یابی به نتایج خوشه‌بندی قوی، به‌ویژه در حضور داده‌های نوفه، سازگاری الگوریتم با ساختارهای دادهٔ مختلف

معایب: حساسیت نسبت به انتخاب پارامترها و حد آستانه، پیچیدگی محاسباتی بالا

پژوهش [۵۴] با هدف یادگیری ارتباط بین ویژگی‌های داده‌ها و وزن‌های خوشه‌بندی انجام شده‌است. تنظیم وزن‌ها برای مجموعه‌داده‌های مختلف به منظور بهبود استحکام و دقت نتایج خوشه‌بندی است. برای رسیدن به این هدف، از یک استراتژی فرایادگیری استفاده می‌شود که قادر است به‌طور خودکار وزن‌های بهینه را برای خوشه‌بندی‌های پایه در یک مجموعه‌داده، با درنظرگرفتن ویژگی‌های منحصربه‌فرد داده‌ها تعیین کند.

مزایا: جایگزینی خوشه‌بندی اجماع مبتنی بر دانش (نظارت‌شده) مبتنی بر اکتشاف، اعمال ضریب جهت رتبه‌بندی خوشه‌ها

معایب: امکان انتخاب نامناسب ابرداده، پیچیدگی محاسباتی و نیاز به پردازش فراداده

پژوهش [۵۵] به ارائهٔ یک روش خوشه‌بندی برای داده‌های پیچیده می‌پردازد که از اطلاعات ساختاری سراسری و محلی به‌منظور بهبود نتایج خوشه‌بندی بهره می‌برد؛ همچنین، از دو مدل تکمیلی استفاده می‌شود: یک مدل خودتوصیف برای یادگیری ساختار کلی و یک مدل خودافزایش ماتریس هم‌بستگی برای یادگیری جزئیات ساختار محلی؛ علاوه‌براین، از رگرسیون کمینه مربعات به‌منظور بیشینه‌سازی تشابه‌ها بین ساختارهای سراسری و محلی استفاده می‌شود. تابع هدف از طریق روش جهت متناوب ضرایب[2] بهینه‌سازی می‌شود.

مزایا: ترکیب ساختارهای سراسری و محلی و ارائهٔ نمای دقیق‌تر و کامل‌تری از روابط داده‌ها

معایب: پیچیدگی تفسیر نتایج خوشه‌بندی، حساسیت به مقادیر اولیه

در مرجع [۵۶] با هدف بهبود استحکام، ثبات و دقت نتایج خوشه‌بندی در مجموعه‌ای از داده‌ها با استفاده از اندازه‌گیری عدم قطعیت دو لایه انجام شده‌است. این اندازه‌گیری شامل ارزیابی‌های در سطح خوشه و در سطح پایه خوشه‌بندی می‌شود. برای دست‌یابی به این هدف، رویکردی برای عدم قطعیت خوشه‌ای اتخاذ شده‌است که با استفاده از معیارهای آنتروپی اصلاح‌شده، برچسب‌های خوشه‌بندی پایه را شامل می‌شود. این فرایند به افزایش دقت ارزیابی‌ها در سطح خوشه و سطح خوشه‌بندی پایه کمک می‌کند؛ علاوه‌براین، دو تابع اجماع جدید، تجمع شواهد وزن‌دار دو سطحی[3] و تقسیم‌بندی نمودار وزن‌دار دوسطحی[4] توسعه یافته‌اند که به‌منظور دستیابی به خوشه‌بندی نهایی و بهبودبخشیدن به نتایج استفاده می‌شوند.

مزایا: اعمال تمایز در خوشه‌بندهای پایه با احتساب ضرایب وزنی

معایب: ناپایداری در مجموعه‌داده‌های با ویژگی کم و مقیاس کوچک

در مقالهٔ [۵۷] هدف اصلی بهبود ماتریس هم‌بستگی به کمک ماتریس لاپلاسین ذکر شده‌است. ارائهٔ نظریهٔ اتصال مرتبهٔ بالاتر در خوشه‌بندی ترکیبی به‌عنوان یک رویکرد یادگیری ماتریس لاپلاسی بهینهٔ رتبه پایین و

[1] Weighted Ensemble Clustering for Multivariate Randomness

[2] Alternating Direction Method Of Multipliers

[3] Two-Level Weighted Evidence Accumulation

[4] Two-level weighted graph Partitioning

[DOI: 10.61186/jsdp.21.3.111]

سپس خوشه‌بندی تجمعی سلسله‌مراتبی پیوند متوسط برای به‌دست‌آوردن نتیجهٔ نهایی مورد توجه قرار گرفته است. ماتریس هم‌بستگی تقویت‌شده توانایی نمایش بهتری نسبت به ماتریس‌های سنتی داشته و درنهایت نتایج خوشه‌بندی گروهی را بهبود داده‌است.

مزایا: بهبود نتایج خوشه‌بندی ترکیبی به کمک ماتریس هم‌بستگی تقویت‌شده

معایب: عدم توجه به کیفیت خوشه‌های پایه، پیچیدگی در تفسیر نتایج

در مرجع [۵۸] تقویت تکنیک‌های خوشه‌بندی مبتنی بر اجماع با انتخاب زیرمجموعه‌ای از راه‌حل‌های خوشه‌بندی که کیفیت، تنوع و اندازه را به طور بهینه متعادل می‌کند و منجر به راه‌حل خوشه‌بندی اجماع دقیق‌تر شده‌است؛ بنابراین مقایسهٔ افرازها با یکدیگر جهت حذف افرازهای ضعیف و اعمال حد آستانه برای معیارهای اندازه، پوشش و تنوع جهت انتخاب از بین افرازهای باقیمانده مورد توجه قرار گرفته‌است.

مزایا: اعمال روش‌های متنوع جهت اجماع که امکان انطباق بر اساس نیازها یا محدودیت‌های خاص داده‌ها را فراهم می‌کند.

معایب: عدم توجه به کیفیت خوشه‌های منفرد در سطح افراز

در مرجع [۵۹] ایدهٔ اصلی اعمال وزن به نمونه‌هایی که به‌درستی خوشه‌بندی شده‌اند جهت ساخت ماتریس هم‌بستگی بهینه است. در این پژوهش یک مفهوم جدید از اشیای داده با توجه به شاخص نفوذ و مدل مخلوط گاوسی تعریف شده‌است. روش KNN به‌منظور کشف میزان شباهت جفت نمونه‌ها و یک الگوریتم خوشه‌بندی چندنمایه با استفاده از ماتریس هم‌بستگی پیشنهاد شده‌است.

مزایا: توانایی پردازش اشیای داده‌های مختلف به‌طور جداگانه انعطاف‌پذیری را افزایش داده‌است.

معایب: KNN بر اساس فاصلهٔ اقلیدسی زمان‌بر بوده و برای مجموعه‌های با ابعاد زیاد نامطلوب است.

در پژوهش [۶۰] با پیشنهاد یک چارچوب خودافزاینده که ماتریس هم‌بستگی را بهبود می‌بخشد و درنتیجه به عملکرد خوشه‌بندی بهبود می‌یابد، به چالش ایجاد ماتریس هم‌بستگی با کیفیت پایین پرداخته است. این روش با استخراج اطلاعات با اطمینان بالا از خوشه‌بندی‌های پایه برای تشکیل یک ماتریس خوشه‌بندی انباشتهٔ سلسله‌مراتبی پراکنده که نشان‌دهندهٔ روابط جفتی قابل اعتماد بین نمونه‌هاست، شروع شده و از آن برای اصلاح ماتریس هم‌بستگی با انتشار اطلاعات با اطمینان بالا استفاده شده‌است.

مزایا: معرفی یک چارچوب جدید خودافزایش جهت ماتریس هم‌بستگی که به اطلاعات خارجی یا پارامترهای اضافی متکی نیست

معایب: ماهیت تکراری الگوریتم می‌تواند منجر به هزینه‌های محاسباتی بالا شود.

رضایی و دانشپور [۶۱] برای محاسبهٔ میزان شباهت و تعیین فاصله، به پارامتر «تعداد ویژگی‌های مشابه» توجه کرده‌اند. در نسبت‌دادن هر نمونه به خوشه در مواردی که فاصله‌ها برابر یا نزدیک باشد، تعداد ویژگی‌های مشترک نمونه‌ها تعیین‌کنندهٔ خوشهٔ مناسب بوده‌است. برای محاسبهٔ فاصله در الگوریتم مورد نظر از تفاضل عددی نرمال‌سازی‌شده برای ویژگی‌های عددی و از فاصلهٔ همینگ برای ویژگی‌های غیرعددی استفاده شده‌است. تعیین مرکز خوشهٔ اولیه نیز مانند بسیاری از روش‌ها تصادفی انجام شده و در تکرارهای بعدی الگوریتم، نمونهٔ مناسب‌تر به‌عنوان مرکز خوشه انتخاب می‌شود.

۳-معیارهای ارزیابی و خوشه‌بندی توافقی

نتایج حاصل از اعمال الگوریتم‌های خوشه‌بندی بر روی یک مجموعه‌داده می‌تواند به‌شدت تحت‌تأثیر انتخاب پارامترهای الگوریتم باشد[۲۵]. این تفاوت‌ها می‌توانند چالش‌های عمده‌ای برای پژوهش‌گران در تخمین خوشه‌های بهینه ایجاد کنند، که به‌تازگی به ارائهٔ شاخص‌های اعتبار خوشه‌بندی متعدد منجر شده‌است[۲۵]. این مسئله، به‌طور معمول تحت عنوان مسئلهٔ اعتبار خوشه‌ها شناخته می‌شود و در پژوهش‌های جدید مورد توجه ویژه‌ای قرار گرفته است[۲۶].

برای یافتن خوشه‌های بهینه، می‌توان از الگوریتم‌های بهینه‌سازی مختلف استفاده کرد؛ در این روند، ابتدا باید تابع برازندگی ایجاد و سپس توسط الگوریتم‌های بهینه‌سازی بهینه شود. در مطالعه‌ای که مورد بررسی قرار گرفته، الگوریتم ژنتیک به‌عنوان ابزار بهینه‌سازی انتخاب شده‌است. الگوریتم‌های ژنتیک که بر پایهٔ اصول تکاملی و ژنتیکی طراحی شده‌اند، می‌توانند در حل مسائل بهینه‌سازی که دارای فضاهای جست‌وجوی بزرگ و پیچیده‌اند، بسیار مؤثر واقع شوند.

۳-۱-بررسی تکنیک‌های اندازه‌گیری اعتبار خوشه‌ها

این شاخص‌ها اغلب بر اساس دو معیار کلیدی ارزیابی می‌شوند: تراکم و جدایی. تراکم به میزان تجمع داده‌ها درون هر خوشه اشاره دارد و نشان‌دهندهٔ کیفیت داخلی خوشه است. جدایی به میزان تفکیک خوشه‌ها از یکدیگر اشاره می‌کند و نشان‌دهندهٔ توانایی الگوریتم در تفکیک

[DOI: 10.61186/jsdp.21.3.111]

خوشه‌های متمایز از داده‌هاست[۲۷]. اعتبارسنجی خوشه‌ها برای تضمین اینکه خوشه‌های انتخاب‌شده بهترین نمایندگی را از داده‌ها ارائه می‌دهند، ضروری است. بدون این شاخص‌ها، دشوار خواهد بود تا به درکی واضح از کارایی واقعی الگوریتم‌های خوشه‌بندی دست یافت؛ بنابراین، استفاده از شاخص‌های معتبر اعتبارسنجی به تحلیل‌گران کمک می‌کند تا تصمیمات مستند و مبتنی بر داده در مورد انتخاب مدل خوشه‌بندی بگیرند.

معیارهای اعتبارسنجی خوشه‌بندی:

۱. تراکم: داده‌های متعلق به یک خوشه باید تا حد ممکن به یکدیگر نزدیک باشند. معیار رایج برای تعیین میزان تراکم داده‌ها، واریانس داده‌ها در درون خوشه‌هاست. این معیار نشان‌دهندهٔ فشردگی داده‌ها و کوچک‌بودن فاصله‌های درون خوشه‌ای است.

۲. جدایی: خوشه‌ها باید به اندازهٔ کافی از یکدیگر جدا باشند. سه روش برای سنجش میزان جدایی خوشه‌ها مورد استفاده قرار می‌گیرد:
- فاصلهٔ بین نزدیک‌ترین داده‌ها از دو خوشه
- فاصلهٔ بین دورترین داده‌ها از دو خوشه
- فاصلهٔ بین مراکز خوشه‌ها

این شاخص‌ها تلاش می‌کنند تا فشردگی و جدایی بین خوشه‌ها را محاسبه کنند و در برخی موارد، هم‌پوشانی بین آن‌ها را نیز ارزیابی کنند تا ترکیب مناسبی از آن‌ها برای یافتن مناسب‌ترین خوشه‌بندی به‌دست آید.

چالش‌های شاخص‌های اعتبارسنجی: بسیاری از شاخص‌های اعتبار خوشه‌بندی از تمام اطلاعات موجود در مورد شکل خوشه استفاده نمی‌کنند که ممکن است باعث شود در برخی موارد، نتایج دقیق نباشند. برخی شاخص‌ها نمی‌توانند در مجموعه‌داده‌های دارای نوفه، خوشه‌های مناسب را تشخیص دهند؛ همچنین، برخی از آن‌ها نظیر شاخص ارائه‌شده توسط کیم [۲۸ و ۲۹]، فشردگی خوشه‌ها را در نظر نمی‌گیرند.

روش‌های ارزیابی خوشه‌های حاصل از خوشه‌بندی تقسیم می‌شوند به:

۱. معیارهای خارجی: این معیارها کیفیت خوشه‌ها را با استفاده از دانش خاصی که توسط کاربران و از داده‌های برچسب‌دار به‌دست می‌آید، ارزیابی می‌کنند. آن‌ها به میزان تطابق خوشه‌های تولیدشده با برچسب‌های پیش‌فرض خارجی می‌پردازند و به‌طور معمول در مواردی که داده‌های آزمایشی واقعی موجود است، استفاده می‌شوند.

۲. معیارهای درونی: این معیارها کیفیت خوشه‌ها را بر اساس اطلاعات استخراج‌شده مستقیم از داده‌های خوشه‌بندی‌شده ارزیابی می‌کنند. این ارزیابی‌ها شامل محاسبهٔ فشردگی داده‌ها درون خوشه‌ها و جدایی بین خوشه‌ها هستند و نیاز به دانش قبلی یا برچسب‌های خارجی ندارند.

۳. معیارهای نسبی: این معیارها بر پایهٔ مقایسه بین چندین الگوریتم خوشه‌بندی با پارامترهای مختلف است که بر روی یک مجموعه‌داده اجرا می‌شوند. انجام چندین بار آزمایش با روش‌های مختلف خوشه‌بندی با پارامترهای مختلف به انتخاب بهترین شمای خوشه‌بندی از بین تمام گزینه‌ها کمک می‌کند. شاخص‌های اعتبارسنجی به‌عنوان مبنای این مقایسه‌ها عمل می‌کنند.

تعداد زیادی شاخص ارزیابی برای خوشه‌بندی پیشنهاد شده‌اند که برخی از رایج‌ترین آن‌ها شامل شاخص‌های داویس-بولدین، سیلهوت و دان‌دهی هستند. این شاخص‌ها به تحلیل‌گران کمک می‌کنند تا به درک بهتری از کارایی مدل‌های خوشه‌بندی دست یابند و تصمیمات مستندتری در مورد انتخاب مدل بگیرند.

شاخص داویس-بولدین معیاری برای اندازه‌گیری فشردگی داخلی خوشه‌ها و جدایی بین آن‌هاست. این شاخص برای هر خوشه نسبت میانگین فاصله‌های داخل خوشه به فاصلهٔ خوشه تا نزدیک‌ترین خوشهٔ دیگر را محاسبه می‌کند.

شاخص سیلهوت معیار میزان شباهت یک شی به خوشهٔ خودش (انسجام) در مقایسه با خوشه‌های دیگر (جداسازی شده) است. این شاخص به طور معمول بین ۱- تا ۱ متغیر است که مقادیر نزدیک به یک نشان‌دهندهٔ جدایی خوب بین خوشه‌هاست.

شاخص دان‌دهی اغلب برای تعیین تعداد بهینهٔ خوشه‌ها در داده‌ها استفاده می‌شود و بر اساس نسبت میان فاصلهٔ داخل خوشه‌ای به فاصلهٔ بین خوشه‌ای محاسبه می‌شود.

استفاده از این شاخص‌ها امکان تحلیل دقیق‌تری از کیفیت خوشه‌بندی را فراهم می‌کند و به تحلیل‌گران اجازه می‌دهد تا با اطمینان بیشتری در مورد انتخاب مدل‌های خوشه‌بندی تصمیم‌گیری کنند.

۳-۲-روش‌های مختلف ارزیابی خوشه‌بندی

در چرخهٔ خوشه‌بندی، به‌طور معمول هیچ اطلاعاتی از برچسب‌های رده‌ها وجود ندارد و به همین دلیل این فرایند را یادگیری بدون ناظر می‌نامند[۳۰، ۳۱]. خوشه‌بندی تلاش می‌کند تا داده‌ها را بر اساس شباهت‌های درونی بین نمونه‌ها در گروه‌هایی قرار دهد که هر گروه خواص مشترکی دارد، بدون آنکه از پیش تعیین شود که هر نمونه

[DOI: 10.61186/jsdp.21.3.111]

باید به کدام رده تعلق داشته باشد؛ در مقابل، رده‌بندی فرایندی است که در آن نمونه‌های دیده نشده را به دسته‌ای ازپیش تعیین‌شده از رده‌ها اختصاص می‌دهد[۳۲]. این روش، که جزو دستهٔ یادگیری با ناظر محسوب می‌شود، بر اساس مدلی که از داده‌های آموزشی با برچسب یاد می‌گیرد، عمل می‌کند و سپس این دانش را برای تخصیص برچسب به داده‌های جدید به‌کار می‌برد. این دو روش، خوشه‌بندی و رده‌بندی، در بسیاری از کاربردهای پردازش داده و تحلیل داده‌های بزرگ نقش کلیدی دارند، هر کدام با هدف و رویکردی متفاوت برای کشف الگوها و دسته‌بندی اطلاعات.

اساسی‌ترین مراحل یک الگوریتم خوشه‌بندی عبارت‌اند از:

- **انتخاب ویژگی:** هدف از این مرحله جمع‌آوری ویژگی‌هایی است که تا حد ممکن، حداکثر اطلاعات ممکن از توزیع داده‌ها را نمایش دهند. این مرحله برای بهینه‌سازی کارایی الگوریتم خوشه‌بندی بسیار مهم است.
- **الگوریتم خوشه‌بندی:** این مرحله شامل انتخاب یک الگوریتم مناسب با مجموعه‌داده‌های در اختیار است که قادر به ارائهٔ یک شمای خوب از خوشه‌ها باشد. معیار نزدیکی و خوشه‌بندی دو فاکتور اساسی در این مرحله‌اند که تأثیر مستقیمی بر کارایی الگوریتم دارند و منجر به تقسیماتی از خوشه‌ها می‌شوند که تا حد ممکن با مجموعه‌داده تنظیم شده‌اند.
- **معیار نزدیکی:** این معیار میزان شباهت یا نزدیکی بین دو نمونه را اندازه‌گیری می‌کند. ویژگی‌های انتخاب‌شده در مرحلهٔ قبلی به‌طور معمول با یک وزن در این معیار شرکت داده می‌شوند تا دقت خوشه‌بندی را افزایش دهند.
- **اعتبارسنجی نتایج:** صحت و درستی الگوریتم خوشه‌بندی با استفاده از تکنیک‌ها و معیارهای مختلف صورت می‌گیرد. ازآنجاکه خوشه‌بندی تقسیماتی از خوشه‌ها را تولید می‌کند که اطلاعاتی در مورد چگالی و شکل این خوشه‌ها نیست، روش‌های اعتبارسنجی باید قادر به ارزیابی دقیق از نتایج با توجه به این فضای ناشناخته باشند[۳۳].
- **تفسیر نتایج:** در بیشتر موارد، یک متخصص، به طور معمول انسانی در حوزهٔ کاربردی، باید نتایج خوشه‌بندی را با شواهد آزمایشی دیگر تجزیه‌وتحلیل کند تا بتواند استنتاج و نتیجهٔ نهایی را ارائه دهد.

یکی از مهم‌ترین پارامترها در خوشه‌بندی، یافتن خوشه‌های بهینه است. برای تخمین خوشه‌های بهینه، ممکن است تعداد مختلفی از خوشه‌بندی‌ها روی مجموعه‌داده‌ها آزمایش شوند و نوعی تقسیم‌بندی انتخاب شود که بنا بر معیار خاصی، بهترین عملکرد را داشته باشد.

روش‌های خوشه‌بندی بر پایهٔ مدل سعی می‌کنند یک مدل ترکیبی را بر روی داده‌ها جور کنند، به طوری که هر خوشه نمایندهٔ یکی از اجزای ترکیب باشد. این روش‌ها برای هر داده توزیعی روی توابع عضویت خوشه‌ها تولید می‌کنند که این توزیع‌ها به طور معمول با نرمال‌کردن احتمالاتی که داده به هر یک از اجزای تولید شده‌است، به‌دست می‌آیند. این نحوهٔ تخصیص، مشابه عملکرد الگوریتم‌های نرم فازی است و یک روش معمول برای ارزیابی این است که مدل ترکیبی تا چه حد توزیع داده‌ها را منعکس می‌کند.

۳-۲-۱-اطلاعات متقابل نرمال‌شده[1] (NMI):

یک معیار مبتنی بر اطلاعات متقابل است که برای سنجش میزان توافق بین دو تقسیم‌بندی مختلف استفاده می‌شود. این شاخص، اندازه‌گیری میزان اطلاعات مشترک استفاده‌شده بین دو متغیر تصادفی را انجام می‌دهد و در شرایطی که تعداد خوشه‌ها زیاد است بسیار مفید واقع می‌شود.

در حالی که معیارهای خلوص و آنتروپی نیز برای سنجش کیفیت خوشه‌بندی استفاده می‌شوند، هر دو معیار ممکن است با افزایش تعداد خوشه‌ها (k) دچار انحراف شوند. با افزایش k، احتمال اینکه هر خوشه به طور انحصاری شامل نمونه‌های یک رده باشد افزایش می‌یابد، که می‌تواند منجر به بالارفتن خلوص اما کاهش کارایی تفسیر خوشه‌بندی در تعمیم‌پذیری شود.

NMI از آنتروپی هر خوشه و هر رده استفاده می‌کند تا اطلاعات متقابل را نرمال‌سازی کند که به بررسی کیفیت خوشه‌بندی بدون تأثیر تعداد خوشه‌ها کمک می‌کند. این معیار به‌ویژه برای تحلیل‌هایی که در آن‌ها تعداد خوشه‌ها نسبت به تعداد کل نمونه‌ها بالاست، بسیار مناسب است.

فرض کنید P(Ci) احتمال این باشد که یک نمونه به رده Ci تعلق داشته باشد، و P(Cj) احتمال این باشد که یک نمونه به خوشهٔ Cj تعلق داشته باشد؛ همچنین، P(i,j) احتمال مشترک این است که یک نمونه هم به رده Ci و هم به خوشه Cj تعلق داشته باشد.

میانگین کاهش عدم قطعیت روی تمام اشیا می‌تواند به صورت معادلهٔ (۱) بیان شود:

$$I(C',C)=\sum_{i-1}^{k'}\sum_{j-1}^{k}P(i,j)\log\frac{P(i,j)}{P'(i)P(j)} \qquad (۱)$$

$I(C',C)$ مقداری بین صفر و $\min(E(c'),E(c))$ می‌گیرد. به‌طوری‌که که بیشینه مقدار آن کمینه مقدار آنتروپی (بی‌نظمی) برای دو خوشه‌بندی است. اطلاعات

[1] Normalized Mutual Information (NMI)

[DOI: 10.61186/jsdp.21.3.111]

متقابل نرمال‌شده بین دو خوشه‌بندی با توجه به میانگین هندسی آنتروپی آن‌ها نرمال شده و به‌صورت معادلهٔ (۲) تعریف می‌شود[۳۴]:

$$NI(C'',C)=\frac{I(C',C)}{\sqrt{E(C')E(C)}} \qquad (۲)$$

درعمل، یک تخمین برای این مقدار، بر اساس تخصیص خوشه به صورت معادلهٔ (۳) محاسبه می‌شود:

$$\mathrm{NM}\,(C',C)=\frac{\sum_{t-1}^{k'}\sum_{j-1}^{k} n_y \log\frac{m_q}{n_t' n_j}}{\sqrt{\left(\sum_{i=1}^{k'} n_i' \log\frac{n_q'}{n}\right)\left(\sum_{j-1}^{k} n_j \log\frac{n_f}{n}\right)}} \qquad (۳)$$

۳-۲-۲-شاخص دقت[1]:

با توجه به تکنیک‌های ارزیابی سنتی در یادگیری با نظارت و بدون نظارت، پژوهش‌گران پیشنهاد داده‌اند که کیفیت یک خوشه با توجه به تخصیص یک رده غالب به هر خوشه و شمارش تعداد اشیایی که به صورت صحیح به خوشهٔ درست تخصیص یافته‌اند محاسبه شود. برای انجام این کار، ابتدا بزرگ‌ترین فصل مشترک N_{ij} را که نتیجهٔ تطابق بین C_i و P_i است را پیدا می‌کند. مقدار تطابق بعدی بر اساس بزرگ‌ترین تطابق N_{ij} از جفت‌های باقی‌مانده انتخاب می‌شود. این رویه تا زمانی که $\min(k_p,k_c)$ پیدا شود، ادامه پیدا خواهد کرد. توجه شود که هیچ خوشه‌ای با بیش از یک رده تطابق داده نمی‌شود. پس از اینکه هر خوشه با یک رده مطابقت داده شد، دقت خوشه‌بندی برای C با استفاده از معادلهٔ (۴) بیان می‌شود.

$$AC(C,P)=\frac{1}{n}\sum_{j\in \mathrm{march}(H)} N_{qf} \qquad (۴)$$

به طوری که $match(j)$ اندیس ردهی که برای تطابق با خوشهٔ C_j انتخاب شده است.

این معیار یک تخمین ساده از کیفیت تقسیم‌بندی تهیه می‌کند، درحالی‌که مقادیر بیشتر برای خلوص، نماینده‌های تعیین بهترین خوشه‌بندی‌اند؛ بااین‌حال، این شاخص تمایل به خوشه‌های کوچک‌تر دارد.

۳-۲-۳-معیار F:

یکی از معیارهای ارزیابی اصلی یک خوشه‌بند، مقدار صحت آن بر روی مجموعه‌آزمون است. این معیار یک معیار کارایی منطقی است؛ زیرا درصد موفقیت یک مدل را در برابر نمونه‌هایی که رده‌های آن‌ها در دست ما نیست بیان می‌کند.

میزان مشابه‌بودن دو افراز به‌وسیلهٔ معیار فیشر اندازه گیری می‌شود. حال باید بگوییم که معیار فیشر[2] چگونه محاسبه می‌شود. این معیار که در این مقاله برای ارزیابی یک افراز در نظر گرفته شده‌است، به‌صورت معادلهٔ (۵) است.

$$FM(P,L)=\max_{\tau}\sum_{i=1}^{K_P}\frac{2\times N_i^P\times\left(\frac{N_{i\tau(i)}^{PL}}{N_i^P}\times\frac{N_{i\tau(i)}^{PL}}{N_{\tau(i)}^{L}}\right)}{N\times\left(\frac{N_{i\tau(i)}^{PL}}{N_i^P}+\frac{N_{i\tau(i)}^{PL}}{N_{\tau(i)}^{L}}\right)} \qquad (۵)$$

که K_P تعداد خوشه‌های افراز P، N_i^P نشان‌دهندهٔ تعداد داده‌های موجود در خوشهٔ i-ام از افراز P، N_j^L نشان‌دهندهٔ تعداد داده‌های موجود در خوشهٔ j-ام از افراز L، N_{ij}^{PL} نشان‌دهندهٔ تعداد داده‌هایی که مشترک در خوشهٔ i-ام از افراز Pو در خوشهٔ j-ام از افراز Lقرار دارد، Nتعداد کل داده‌ها را نشان می‌دهد و τ یک جای‌گشت از اعداد یک تا N است. اگر دو افراز P و برچسب L کامل مشابه باشند آن‌گاه FM مقدار بیشینه یعنی یک و اگر دو افراز به‌طور کامل متفاوت از یکدیگر باشند، مقدار صفر را برمی‌گرداند.

۳-۳-خوشه‌بندی توافقی

یکی از روش‌های خوشه‌بندی ترکیبی که در پژوهش‌های جدید مورد توجه قرار گرفته، روش خوشه‌بندی توافقی[3] است. این نوع خوشه‌بندی در مقالات با نام‌های دیگری نظیر «گروه‌های خوشه‌بندی»[4] و «اجتماع خوشه‌بندی‌ها»[5] نیز شناخته می‌شود. در روش خوشه‌بندی توافقی، نتایج حاصل از چندین خوشه‌بندی با هم ترکیب می‌شوند تا درنهایت به یک خوشه‌بندی واحد دست یابیم.

الگوریتم‌های خوشه‌بندی توافقی بیشتر اوقات خوشه‌بندی بهتری تولید می‌کنند؛ این روش‌ها خوشه‌بندی ترکیبی را می‌یابند که به‌تنهایی به‌وسیلهٔ هر الگوریتم خوشه‌بندی دیگری قابل تولید نیست. این الگوریتم‌ها حساسیت کمتری نسبت به نوفه دارند و قادر به یک‌پارچه‌سازی نتایج از منابع توزیع شده‌اند. به‌دلیل این قابلیت‌ها، خوشه‌بندی توافقی برای استفاده در داده‌هایی که از چندین منبع مختلف جمع‌آوری شده‌اند یا داده‌هایی که دارای نوفهٔ زیادی هستند، ایده‌آل است.

۴-روش کار

در بخش ابتدایی (۱-۴) خوشه‌های اولیه ایجاد می‌شوند که می‌تواند با اجرای چندین الگوریتم خوشه‌بندی روی یک مجموعه‌داده یا اجرای مکرر یک الگوریتم روی مجموعه‌داده صورت گیرد. به‌دلیل این‌که خوشه‌های اولیه ممکن است از پایداری مناسبی نداشته نباشند، باید با استفاده از یکی از معیارهای ارزیابی خوشه‌بندی مانند فیشر (مطابق الگوریتم شکل (۱)) بررسی شوند تا میزان پایداری خوشه‌ها مشخص شود. در فاز دوم (۴-۲)، خوشه‌های پایدارتر به‌کمک

[1] Accuracy
[2] F-measure
[3] Consensus Clustering
[4] Clustering Ensemble
[5] Clustering Aggregation

[DOI: 10.61186/jsdp.21.3.111]

الگوریتم ژنتیک با یکدیگر ترکیب می‌شوند تا بهترین خوشه‌بندی ترکیبی نهایی حاصل شود. برای ترکیب نتایج، از ماتریس هم‌بستگی استفاده می‌شود.

در فاز سوم (۳-۴)، ماتریس هم‌بستگی به یک الگوریتم خوشه‌بندی سلسله‌مراتبی داده می‌شود و خوشه‌بندی توافقی نهایی را برمی‌گرداند.

ماتریس هم‌بستگی در اینجا نشان‌دهندهٔ شباهت یا میزان ارتباط بین خوشه‌های مختلف است که در مراحل قبلی تولید شده‌اند. استفاده از این ماتریس در الگوریتم سلسله‌مراتبی به این معنی است که خوشه‌های با بیشترین شباهت (با بیشترین میزان هم‌بستگی) با یکدیگر ترکیب می‌شوند.

۴-۱-محاسبهٔ پایداری خوشه‌ها

یک خوشهٔ پایدار، خوشه‌ای است که اگر چندین روش خوشه‌بندی دیگر روی آن مجموعه‌داده اجرا شود، با احتمال زیاد این خوشه باز هم دیده خواهد شد؛ به‌عبارت‌دیگر، خوشه‌های پایدار به خوشه‌هایی اطلاق می‌شوند که در خوشه‌بندی‌های مختلف روی زیرمجموعه‌های به‌دست‌آمده از نمونه‌برداری‌های مختلف بیشترین تکرار را داشته باشند.

برای ارزیابی پایداری خوشه‌ها، ابتدا الگوریتم خوشه‌بندی را بر روی همان مجموعه‌داده چندین بار اعمال کرده تا داده‌ها را به خوشه‌های مختلف تقسیم کنیم. سپس، خوشه‌های حاصل از هر اجرا را مورد بررسی قرار داده و پایداری هر خوشه را بر اساس تکرار آن در نتایج مختلف محاسبه می‌کنیم.

این فرایند اجازه می‌دهد تا خوشه‌هایی که بیشترین پایداری را دارند شناسایی شوند و به‌عنوان خوشه‌های نهایی در نظر گرفته شوند. خوشه‌های پایدار اغلب نشان‌دهندهٔ ساختارهای واقعی در داده‌ها هستند و بنابراین، توانایی بالایی در تفسیر و تعمیم داده‌های به‌دست‌آمده از آن‌ها وجود دارد.

با توجه به الگوریتم ارائه‌شده، میزان پایداری هر خوشه را با تمامی خوشه‌های ترکیب دوم به‌دست آورده، میانگین مقادیر به‌دست آمده برابر با پایداری خوشه مربوطه است.

معادلهٔ (۶) پایداری هر خوشه را برحسب معیار فیشر نمایش می‌دهد:

(۶)

$$Stability(C_i) = \frac{1}{|RefSet|}\sum_{j=1}^{|RefSet|} F-measure(C_i, RefSet_j)$$

RefSet. تعداد خوشه‌بندی‌ها در مجموعهٔ جدید است.

Input:
D — a dataset {x1, x2, … , xn}
K — maximum number of Ensemble Clusters
|RS|
— numbers of clusters in new partition (Refrence S
C_i — cluster i from Esemble
RSj- cluster j from RefrenceSet clusterig
Output:
Stability(C_i) — Stability of C_i
Require:
Resample D to obtain the perturbed data set D′;
Run K
— Means or other clustering algorithms (cluster(D,
over D to obtain P′(D);
Re — labeling P′(D) to P(D);

۱. stability(C_i)=0;
۲. For j:=1 to RS do
۳. stability(C_i)=Fmeasure(c_i, RS_j)+ stability (C_i);
۴. end for
۵. $stability(C_i) = \frac{stability(C_i)}{|RS|}$;

(شکل-۱): الگوریتم محاسبهٔ پایداری خوشهٔ Ci به‌عنوان تابع برازندگی

(Figure-1): Algorithm for calculating the stability of Ci cluster as a fitness function

پس از اینکه عمل انتخاب خوشه‌بندی‌ها با توجه به مقدار پایداری خوشه انجام شد، آنگاه یک روش مبتنی بر الگوریتم‌های تکاملی انتخاب می‌کنیم. الگوریتم تکاملی که در این مقاله مورد بررسی قرار گرفته‌است، الگوریتم ژنتیک است.

۴-۲-روند کلی الگوریتم ژنتیک

پیش از اینکه یک الگوریتم ژنتیکی بتواند اجرا شود، ابتدا باید کدگذاری یا نمایش مناسبی برای مسئلهٔ مورد نظر پیدا شود. معمول‌ترین شیوهٔ نمایش کروموزوم‌ها در الگوریتم ژنتیک به شکل رشته‌های دودویی است، که در آن هر متغیر تصمیم‌گیری به صورت دودویی درآمده و سپس با کنار هم قرارگرفتن این متغیرها، کروموزوم ایجاد می‌شود؛ گرچه این روش گسترده‌ترین شیوهٔ کدگذاری است، شیوه‌های دیگری مانند نمایش با اعداد حقیقی نیز در حال گسترش‌اند.

در بیشتر الگوریتم‌های فراابتکاری، از جمله الگوریتم ژنتیک، جواب‌های اولیه به صورت تصادفی انتخاب می‌شوند. در این مورد، یک خوشه‌بندی ترکیبی است که جواب‌های اولیه یک رشته باینری به تعداد خوشه‌های پایدار است.

هر کروموزوم بیانگر یک جواب از فضای جست‌وجو است و به‌عنوان فرد شناخته می‌شود. مجموعهٔ این افراد، جمعیت یا نسل فعلی نامیده می‌شوند. به هر فرد، برازندگی

[DOI: 10.61186/jsdp.21.3.111]

بر اساس مقدار تعیین‌شده به‌وسیلهٔ تابع هدف تعلق می‌گیرد و از این مقدار برای هدایت فرایند انتخاب به سمت اشخاص مناسب‌تر استفاده می‌شود. افراد با برازندگی بالا نسبت به کل جمعیت، احتمال بیشتری برای انتخاب‌شدن برای تولید مثل دارند و در مقابل، اشخاص با برازندگی کمتر احتمال انتخاب کمتری دارند.

همچنین یک تابع برازندگی نیز باید ابداع شود تا به هر راه حل کدگذاری‌شده ارزشی را نسبت دهد. در طی اجرا، والدین برای تولید مثل انتخاب می‌شوند و با استفاده از عمل‌گرهای آمیزش و جهش با هم ترکیب می‌شوند تا فرزندان جدیدی تولید کنند. این فرایند چندین بار تکرار می‌شود تا نسل بعدی جمعیت تولید شود. سپس این جمعیت بررسی می‌شود و در صورتی که ضوابط هم‌گرایی برآورده شوند، فرایند خاتمه می‌یابد.

استفاده از الگوریتم ژنتیک در پژوهش حاضر به دلایل متعددی بوده‌است که در زیر به آن‌ها اشاره می‌کنیم:

قابلیت تطبیق‌پذیری: الگوریتم‌های ژنتیک به خاطر قابلیت‌های تطبیق‌پذیری بالا در شرایط و مسائل مختلف معروف‌اند. این ویژگی به ما امکان داد تا یک مدل بهینه‌سازی قوی و قابل انعطاف داشته باشیم که قادر است در مواجهه با مسائل متنوع خوشه‌بندی، پاسخ‌های کارآمد ارائه دهد.

کارایی در فضاهای جست‌وجوی بزرگ: در مسائلی که فضای جست‌وجوی وسیع و پیچیده‌ای دارند، الگوریتم‌های ژنتیک می‌توانند با استفاده از روش‌های جهش و ترکیب، به حل مسئله بپردازند، در حالی که دیگر الگوریتم‌های بهینه‌سازی ممکن است در یافتن جواب‌های بهینه با محدودیت مواجه شوند.

قابلیت اکتشاف و بهره‌برداری: الگوریتم‌های ژنتیک توانایی بالایی در اکتشاف[1] و بهره‌برداری[2] دارند. این ویژگی امکان می‌دهد که الگوریتم هم‌زمان به دنبال راه‌های جدید در فضای جست‌وجو باشد و هم از راه‌حل‌های موجود به بهترین شکل ممکن استفاده کند.

مناسب برای مسائل بهینه‌سازی چندمنظوره: در مواردی که خوشه‌بندی باید به تعدادی از اهداف متفاوت جواب دهد، الگوریتم‌های ژنتیک با توانایی تنظیم پارامترهای متنوع، انتخاب ایدئالی هستند.

با توجه به این دلایل، الگوریتم ژنتیک به‌عنوان یک انتخاب مناسب برای این پژوهش شناخته و پیاده‌سازی شده‌است.

[1] Exploration
[2] Exploitation

۴-۳-توابع هدف و برازندگی[3]

تابع هدف به منظور ارزیابی عملکرد و کارایی اعضای جمعیت در الگوریتم‌ها استفاده می‌شود. مسائل بهینه‌سازی می‌توانند به‌صورت کمینه‌سازی یا بیشینه‌سازی باشند؛ در مسائل بهینه‌سازی نوع کمینه‌سازی، برازنده‌ترین اعضا باید دارای کمترین مقدار عددی تابع هدف باشند. به‌طور مشابه، در مسائل بهینه‌سازی بیشینه‌سازی، برازنده‌ترین اعضا باید دارای بیشترین مقدار عددی تابع هدف باشند. تابع برازندگی، که مقادیر ارزیابی‌شدهٔ اعضا به‌وسیلهٔ تابع هدف را به مقادیری تبدیل می‌کند که به‌عنوان مقدار برازندگی کروموزوم‌ها شناخته می‌شوند، از طریق توابع تبدیل به‌کار گرفته می‌شود. در بسیاری از الگوریتم‌های فراابتکاری، از جمله الگوریتم ژنتیک، جواب‌های اولیه به‌صورت تصادفی انتخاب می‌شوند. در این مورد، خوشه‌بندی ترکیبی به‌کار رفته‌است که در آن جواب‌های اولیه به‌صورت یک رشته دودویی برابر با تعداد خوشه‌های پایدارند؛ مقدار یک در ژن نشان‌دهندهٔ شرکت آن خوشه در تابع برازندگی و مقدار صفر نشان‌دهندهٔ عدم‌شرکت خوشهٔ مربوطه در ترکیب نهایی است. فرایند انتخاب خوشه‌بندی‌ها در دو فاز صورت می‌پذیرد: در فاز نخست، یک الگوریتم تکاملی به دنبال یافتن زیرمجموعه‌ای از خوشه‌بندی‌ها با بیشترین پایداری است. این هدف به‌صورت ضمنی با انتخاب نیمه‌پایدارتر در گام پیش از الگوریتم تکاملی دنبال می‌شود. فاز دوم، که به دنبال انتخاب متنوع‌ترین خوشه‌بندی‌هاست، به‌صورت صریح در تابع کارایی الگوریتم‌های تکاملی ظاهر می‌شود. این الگوریتم‌های تکاملی دارای کروموزوم‌های بیتی هستند که طول آن‌ها برابر با تعداد کل خوشه‌بندی‌های موجود در اجماع نهایی است. در این کروموزوم‌ها، هر ژن می‌تواند مقدار یک یا صفر را اتخاذ کند. مقدار یک نشان‌دهندهٔ انتخاب‌شدن خوشه‌بندی متناظر با شمارهٔ آن ژن و مقدار صفر به معنای عدم انتخاب آن خوشه‌بندی است. برای محاسبهٔ تابع برازندگی در این الگوریتم تکاملی، میزان پراگندگی خوشه‌بندی‌های انتخاب‌شده بررسی می‌شود؛ برای مثال، اگر طول کروموزوم شش باشد، کروموزوم حاصل از اجرای الگوریتم ژنتیک می‌تواند به صورت شکل (۲) باشد:

۱	۱	۱	۱	۰	۱

(شکل-۲): نمایش یک راه حل نامزد (کروموزوم)
(Figure-2): Representation of a candidate solution (chromosome)

[3] The objective and fitness function

[DOI: 10.61186/jsdp.21.3.111]

در این کروموزوم خوشه‌هایی که مقدار یک دارند در ماتریس هم‌بستگی نهایی شرکت داده می‌شوند؛ بنابراین تنها پنجمین خوشه‌بندی شرکت نمی‌کند.

تابع برازندگی به صورت ضابطهٔ (۷) محاسبه می‌شود.

$$\text{FitnessFurction} = 0.5 - \frac{\sum_{xy} abs(Co(x,y) - 0.5)}{N^2} \qquad (۷)$$

که N تعداد نمونه‌ها و Co(x,y) نشان‌دهندهٔ هم‌بستگی بین دو خوشهٔ x و y است. این هم‌بستگی به‌صورت نسبت حداقل مقادیر مشترک به حداکثر مقادیر مشترک بین دو خوشه تعریف می‌شود که می‌تواند اطلاعاتی در مورد میزان شباهت یا نزدیکی بین این دو خوشه ارائه دهد.

$$Co(x,y) = \frac{Co_{min}(x,y)}{Co_{max}(x,y)} \qquad (۸)$$

$$Co_{min}(x,y) = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{k_i} \min\left(p_j(x), p_j(y)\right) \qquad (۹)$$

$$Co_{max}(x,y) = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{k_i} \max\left(p_j(x), p_j(y)\right) \qquad (۱۰)$$

همچنین B به‌عنوان یک پارامتر برای تعریف تعداد بلاک‌ها یا تقسیم‌بندی‌هایی است که در آن‌ها مقادیر داده‌ها برای محاسبهٔ میانگین‌های Co_{min} و Co_{max} استفاده می‌شوند؛ این روش به ما کمک می‌کند تا تأثیر نوفه و تغییرات جزئی در داده‌ها را کاهش دهیم و تصویر دقیق‌تری از هم‌بستگی بین خوشه‌ها به‌دست آوریم. k_i تعداد نمونه‌ها در خوشهٔ i-ام و $p_j(x)$ و $p_j(y)$ نشان‌دهندهٔ مقدار مشخصی در داده‌ها برای خوشه‌های x و y در جایگاه j-ام هستند.

این روش از محاسبه به ما امکان می‌دهد تا تشابه بین دو خوشه را با توجه به توزیع داده‌هایشان ارزیابی کنیم. با استفاده از میانگین وزن‌دار، تأثیر نوفه و تفاوت‌های کوچک در داده‌ها کاهش می‌یابد؛ بنابراین نتایج به‌دست‌آمده می‌تواند بیانگر یک تصویر واقعی‌تر از شباهت‌های بین خوشه‌ها باشد. میزان تابع براندگی برای ماتریس هم‌بستگی فوق برابر ۲۹ است، اما ماکزیمم مقدار برای این تابع برابر 29.14 است. در گام آخر ماتریس هم‌بستگی به‌دستآمده از اجماع ثالث بهینه به‌عنوان یک ماتریس مشابهت در نظر گرفته می‌شود. در این صورت یک الگوریتم خوشه‌بندی سلسله‌مراتبی به‌عنوان تابع جمع‌کنندهٔ نهایی در نظر گرفته می‌شود و ماتریس هم‌بستگی به‌دست‌آمده را به‌عنوان ورودی گرفته و خوشه‌بندی توافقی نهایی را بر می‌گرداند. در خوشه‌بندی سلسله‌مراتبی با توجه به مقادیر ماتریس هم‌بستگی، خوشه‌هایی که دارای کمترین فاصله (بیشترین شباهت) اند با هم ادغام می‌شوند و خوشهٔ جدیدی می‌سازند. در مرحلهٔ بعد باز هم فاصلهٔ بین خوشه‌های موجود و جدید، توسط ماتریس شباهت که به‌روزرسانی شده، محاسبه و کار ادغام ادامه پیدا می‌کند تا تنها K خوشه باقی بماند.

۵- آزمایش‌ها

در این بخش گزارش نتایج روش پیشنهادی بر روی مجموعه‌داده‌هایی[1] که مقالات معتبر دنیا نتایج روش خود را بر روی آن‌ها گزارش می‌کنند، آورده شده‌است. چهار مجموعه‌داده که در بیشتر مقالات مورد استفاده قرار می‌گیرند در این قسمت مورد بررسی قرار گرفته‌است که با این کار قادر خواهیم بود روش ارائه‌شده را به‌راحتی با روش‌های دیگر مقایسه کنیم. این مجموعه‌های داده‌ای از UCI Machine Learning Repository هستند.

۵-۱-مجموعه‌های داده‌ای

ما الگوریتم‌های پیشنهادی را بر روی چهار مجموعه‌داده به نام‌های Iris، Halfring، Wine و Nglass ارزیابی کرده‌ایم. مجموعهٔ Wine شامل نمونه‌هایی با سیزده ویژگی است. مجموعهٔ Iris دارای ۱۵۰ نمونه است که در سه ردهٔ پنجاه تایی تقسیم شده و هر رده مربوط به یک نوع گیاه است؛ هر نمونه در این مجموعه دارای چهار ویژگی است که منجر به تشکیل سه خوشه می‌شود. مجموعهٔ Nglass شامل داده‌هایی در مورد ترکیبات شیمیایی شش نوع شیشه است، که هر نمونه از نُه ویژگی برخوردار است. این مجموعه‌های داده از مخزن یادگیری ماشین UCI تهیه شده‌اند. مشخصات این چهار مجموعه‌داده در جدول (۱) آورده شده‌است. ارائهٔ داده‌ها و نتایج به صورت تابعی از پارامترهای مستقل انجام می‌شود.

(جدول-۱): مشخصات مجموعه‌های داده‌ای
(Table-1): Specifications of data sets

نام مجموعه‌ها	تعداد داده‌ها	تعداد ویژگی‌ها	تعداد خوشه‌ها
Iris	۱۵۰	۴	۳
wine	۱۷۸	۱۳	۳
Halfring	۴۰۰	۲	۲
Nglass	۲۱۴	۹	۶

۵-۲-نتایج

اجتماع خوشه‌بندی‌های اولیه در این پژوهش شامل پنجاه خوشه‌بندی برای هر مجموعه‌داده است؛ این خوشه‌بندی‌ها با استفاده از شش روش مختلف شامل K-Means، Single

[1] Data Sets

[DOI: 10.61186/jsdp.21.3.111]

Linkage، Average Linkage، Complete Linkage، FCM و EAC در نرم‌افزار MATLAB ایجاد شده‌اند. برای تولید این پنجاه خوشه‌بندی، هر یک از این شش روش با پارامترهای اولیهٔ متفاوتی پیکربندی شده‌اند. در ارزیابی نتایج، از میانگین بیست بار اجرا استفاده شده‌است؛ به این معنی که هر نتیجه از میانگین بیست اجرای متفاوت به‌دست آمده‌است. برای ارزیابی خوشه‌بندی نهایی، معیار ارزیابی فیشر استفاده شده‌است. نتایج این ارزیابی‌ها در جدول (۲) آورده شده‌اند.

بر اساس نتایج حاصل از اجرای الگوریتم‌های مختلف (خوشه‌بندهای پایه) بر روی چندین مجموعه‌داده، می‌توان به نمودارهای شکل‌های (۳ تا ۷) اشاره کرد؛ برای مثال، در مجموعه‌دادهٔ Iris، نمودار شکل (۳) حاصل شده‌است. نمودار شکل (۴) نشان می‌دهد که با استفاده از معیار ارزیابی فیشر اصلاح‌شده، الگوریتم Single Linkage در مجموعهٔ Wine نتایج بهتری نسبت به سایر الگوریتم‌ها داشته‌است. نمودار شکل (۵) نتایج الگوریتم‌های مختلف را برای مجموعه‌دادهٔ Halfring نمایش می‌دهد که در آن الگوریتم EAC با معیار ارزیابی فیشر اصلاح‌شده نتایج بهتری را ارائه داده‌است. در این سه مجموعه‌داده، معیار ارزیابی فیشر اصلاح‌شده نتایج بهتری نسبت به دو معیار دیگر ارائه داده‌است. به طور غیرمنتظره، نتایج برای مجموعه‌دادهٔ Nglass با استفاده از معیار فیشر ماکزیمم و الگوریتم EAC بهترین بوده‌است. نمودار شکل (۷)، نمودار کلی برای تمامی مجموعه‌های داده را نمایش می‌دهد.

به منظور بررسی عملکرد الگوریتم پیشنهادی، ما آن را با الگوریتم‌های خوشه‌بندی ترکیبی پیشرفته از جمله روش‌های ذکر شده در ادامه مقایسه می‌کنیم (اشکال ۸ تا ۱۰):

(۱) خوشه‌بندی انباشت شواهد (EAC) همراه با الگوریتم خوشه‌بندی پیوند یکپارچه به‌عنوان تابع توافق ($EAC + SL$)،

(۲) الگوریتم خوشه‌بندی($EAC + AL$) [۳۵]،

(۳) اتصال سه‌گانه (WCT) همراه الگوریتم خوشه‌بندی پیوند تکی به‌عنوان تابع اجماع ($WCT + SL$) [۳۶]،

(۴) الگوریتم خوشه‌بندی به‌عنوان تابع اجماع ($WCT + AL$) [۳۵]،

(۵) کیفیت سه‌گانهٔ وزن‌دار (WTQ) همراه با الگوریتم خوشه‌بندی پیوند یکپارچه به‌عنوان تابع توافق ($WTQ + SL$) [۳۶]،

(۶) الگوریتم خوشه‌بندی پیوند متوسط به‌عنوان تابع اجماع ($WTQ + AL$) [۳۶]،

(۷) اندازه‌گیری تشابه ترکیبی (CSM) همراه با الگوریتم خوشه‌بندی پیوند یکپارچه به‌عنوان تابع اجماع ($CSM + SL$) [۳۶]،

(۸) الگوریتم خوشه‌بندی پیوند متوسط به‌عنوان تابع توافق ($CSM + AL$) [۳۶]،

(۹) الگوریتم تقسیم مشابهت خوشه‌ای ((CSPA) [۳۷]،

(۱۰) الگوریتم ($HGPA$) [۳۷]،

(۱۱) الگوریتم خوشه‌بندی ($MCLA$) [۳۷]،

(۱۲) رأی‌گیری انتخابی بدون وزن (SUW) [۳۸]،

(۱۳) رأی‌گیری وزنی انتخابی (SWV) [۳۸]،

(۱۴) بیشینه‌سازی انتظار (EM)

(۱۵) اجماع رأی‌گیری تکراری (IVC)[۳۹].

علاوه‌براین، ما روش پیشنهادی را با الگوریتم‌های خوشه‌بندی پایهٔ «قوی» دیگر از جمله:

(۱) الگوریتم خوشه‌بندی طیفی طبیعی (NSC) [۴۰]،

(۲) الگوریتم «خوشه‌بندی فضایی مبتنی بر تراکم» ($DBSCAN$)[۴۱]

(۳) الگوریتم «خوشه‌بندی با جست‌وجوی سریع و یافتن قله‌های متراکم» ($CFSFDP$) [۴۲] مقایسه کرده‌ایم. هدف از این مقایسه این است که آیا روش پیشنهادی یک خوشه‌بندی «مقاوم» است یا خیر.

جدول (۳) مشخصات ۸ مجموعه‌داده‌ای آورده شده‌است. این مجموعه جهت بررسی دقت روش پیشنهادی (براساس معیارهای ARI و NMI) در مقایسه با سایر روش‌ها مورد توجه قرار گرفته‌است.

نتایج الگوریتم خوشه‌بندی جمعی پیشنهادی در مقایسه با چهار الگوریتم خوشه‌بندی «قوی» در مجموعه‌داده‌های معیار مختلف در اشکال (۸ و۹) نشان داده شده‌است؛ در این بخش برای ارزیابی روش پیشنهادی با سایر روش‌های رقیب از دو معیار معروف NMI و ARI [۴۳] استفاده شده‌است. در مجموعه‌داده‌های مختلف مشاهده می‌شود که خوشه‌بندی پیشنهادی معتبر بوده و برتری چشم‌گیری نسبت به چهار الگوریتم رقیب دارد. نتایج این آزمایش‌ها نشان می‌دهند که الگوریتم پیشنهادی نه تنها می‌تواند با الگوریتم‌های خوشه‌بندی «قوی» رقابت کند، بلکه بر روی بسیاری از مجموعه‌داده‌های استاندارد برتری مطلوبی را نشان دهد. شکل‌های (۳ تا ۶) مقایسهٔ انواع مختلف معیار فیشر جهت ارزیابی پایداری خوشه‌ها در مجموعه‌های داده‌ای جدول (۱) هستند. در نهایت در شکل (۷) جهت ارزیابی بهتر هر چهار شکل در هم ادغام شده‌است. معیار ارزیابی فیشر اصلاح‌شده، در مقایسه با دو معیار دیگر، عملکرد بهتری را روی بیشتر مجموعه‌داده‌ها نشان می‌دهد؛ همچنین مطالعهٔ الگوریتم‌های مختلف نشان داد که با استفاده از معیار فیشر اصلاح‌شده، نتایج بهتری نسبت به دیگر معیارها به‌دست می‌آید. درنهایت در شکل

[DOI: 10.61186/jsdp.21.3.111]

(۱۰) میزان میانگین NMI روش پیشنهادی با الگوریتم‌های خوشه‌بندی ترکیبی پیشرفته مقایسه شده‌است که نتایج حاصله نشان از برتری روش پیشنهادی دارد. ذکر این نکته نیز ضروریست که هر الگوریتم خوشه‌بندی قوی ممکن است بر روی برخی مجموعه‌های داده‌ای به دلایل زیر عملکرد مطلوبی نداشته باشد (مانند NMI مجموعه‌داده‌ای A7):

- ✓ حساسیت به نوفه و داده‌های پرت
- ✓ پیچیدگی ذاتی داده‌ها
- ✓ انتخاب معیارهای شباهت
- ✓ تفاوت در توزیع داده‌ها (مثلاً تفاوت در تراکم داده‌ها)

(جدول-۲): نتایج تجربی به‌دست‌آمده از خوشه‌بندهای پایه جهت سه نوع معیار ارزیابی فیشر

(Table-2): Experimental results obtained from the clusters of basic clauses for 3 types of Fisher's evaluation criteria

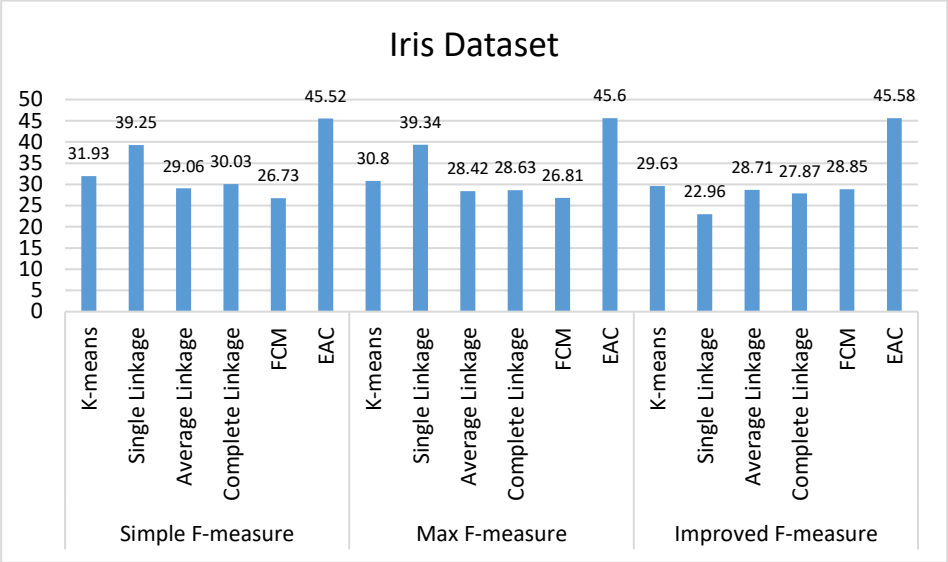
	داده ها / الگوریتم ها	Iris	Wine	Halfring	Nglass
Simple F-measure	K-means	31.93	35.15	11.02	44.98
	Single Linkage	39.25	46.07	41.18	44.21
	Average Linkage	29.06	42.49	42.76	40.06
	Complete Linkage	30.03	40.25	29.47	42.11
	FCM	26.73	33.26	11.02	41.74
	EAC	45.52	45.94	43.24	39.27
Max F-measure	K-means	30.8	31.22	11.4	44.5
	Single Linkage	39.34	46.28	42.11	46.76
	Average Linkage	28.42	42.26	44.18	48.19
	Complete Linkage	28.63	39.08	31.86	47.99
	FCM	26.81	31.76	11.14	43.87
	EAC	45.6	46.17	44.25	48.29
Improved F-measure	K-means	29.63	27.6	11.64	33.35
	Single Linkage	22.96	46.3	42.24	42.31
	Average Linkage	28.71	27.9	15.55	15.42
	Complete Linkage	27.87	31.23	26.07	26.8
	FCM	28.85	26.91	11.62	11.44
	EAC	45.58	46.17	44.37	44.34

(جدول-۳): مشخصات مجموعه‌های داده‌ای استفاده شده برای مقایسه مقاوم‌بودن روش پیشنهادی

(Table-3): Characteristics of data sets used to compare the robustness of the proposed method

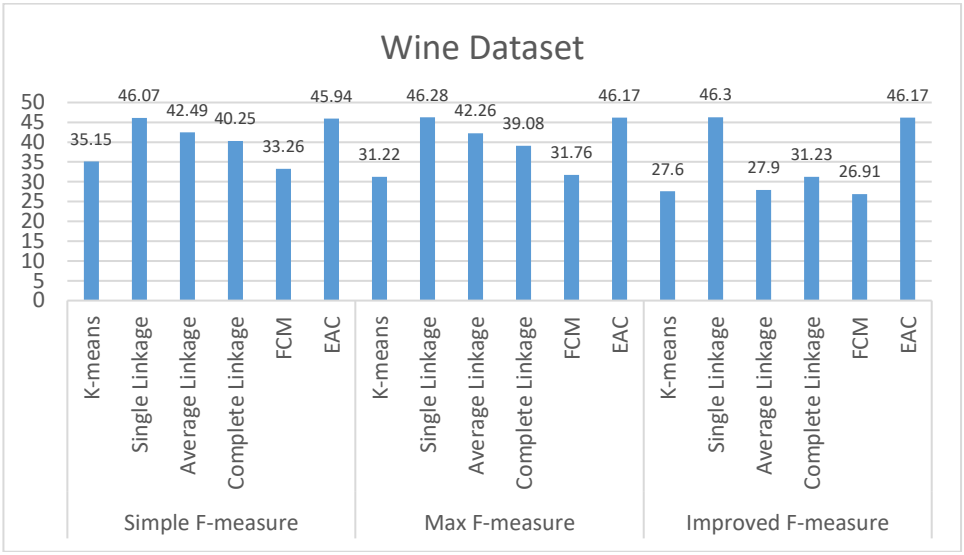
	Dataset	The size of dataset	The number of features in dataset	The number of the consensus clusters in given dataset
Artificial dataset	Ring3 (R3)	1500	2	3
Artificial dataset	Banana2 (B2)	2000	2	2
Artificial dataset	Aggregation7 (A7)	788	2	7
UCI dataset	Imbalance2 (I2)	2250	2	2
UCI dataset	Iris (I)	150	4	3
UCI dataset	Wine (W)	178	13	3
UCI dataset	Breast (B)	569	10	2
UCI dataset	Digits (D)	5620	63	10

[DOI: 10.61186/jsdp.21.3.111]



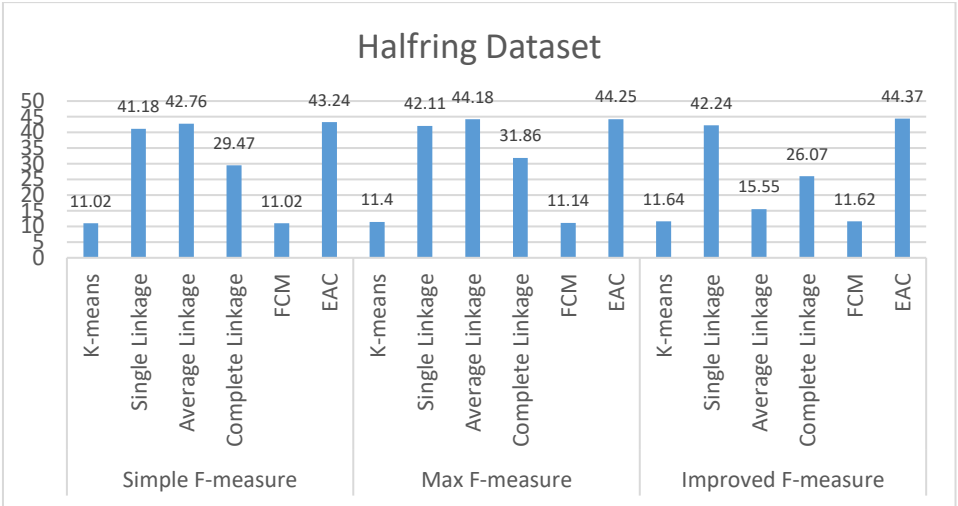
(شکل-۳): اعمال الگوریتم‌های مختلف بر مجموعه‌داده Iris

(Figure-3): Applying different algorithms to the Iris dataset



(شکل-۴): اعمال الگوریتم‌های مختلف بر مجموعه‌دادهWine

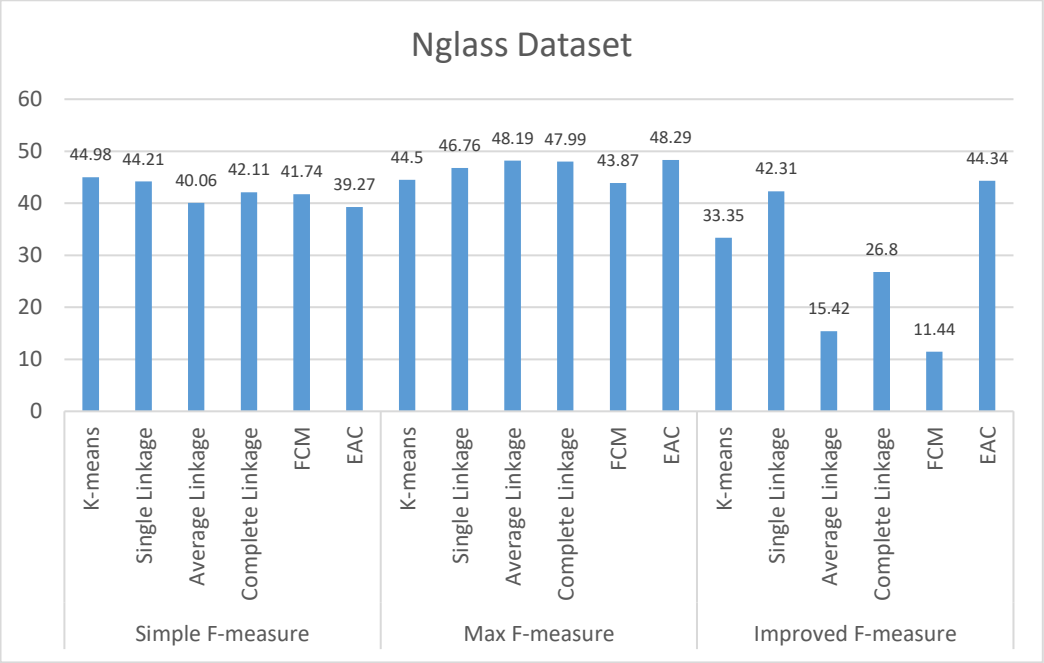
(Figure-4): Applying different algorithms to the Wine dataset



(شکل-۵): اعمال الگوریتم‌های مختلف بر مجموعه‌داده Halfring

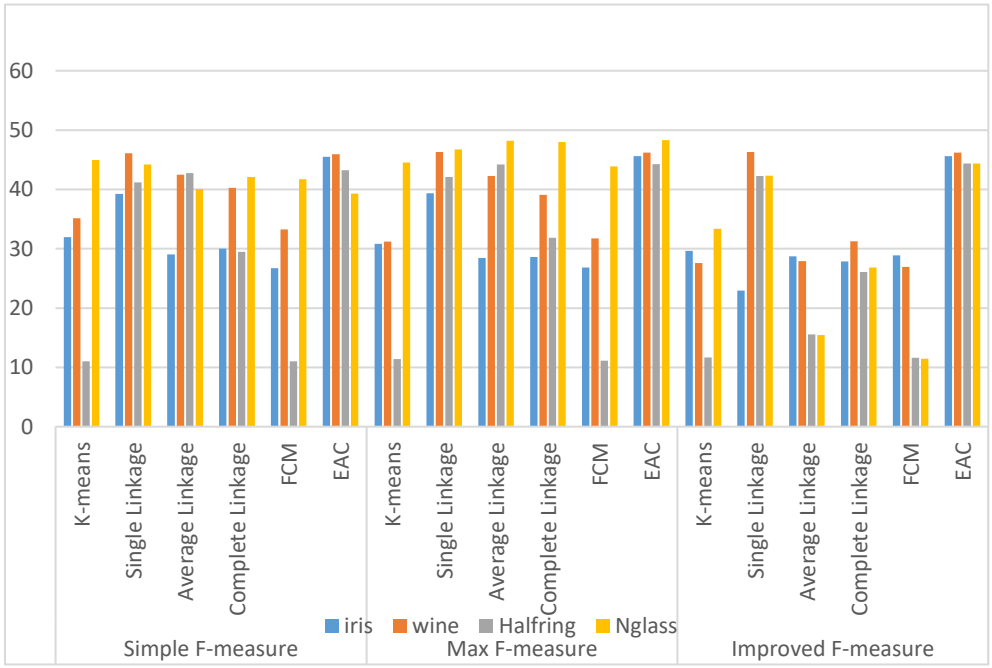
(Figure-5): Applying different algorithms to the Halfring dataset

[DOI: 10.61186/jsdp.21.3.111]



(شکل-۶): اعمال الگوریتم‌های مختلف بر مجموعه‌داده Nglass

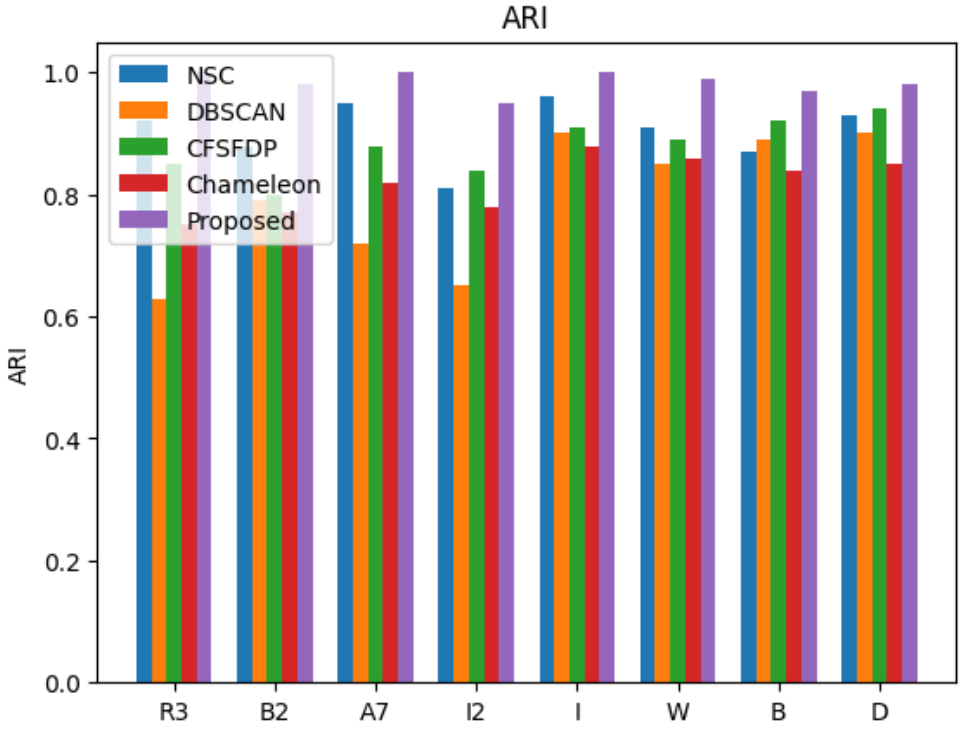
(Figure-6): Applying different algorithms to the Nglass dataset



(شکل-۷): اعمال الگوریتم‌های مختلف بر روی ۴ مجموعه‌داده‌ها با احتساب ۳ نوع معیار فیشر

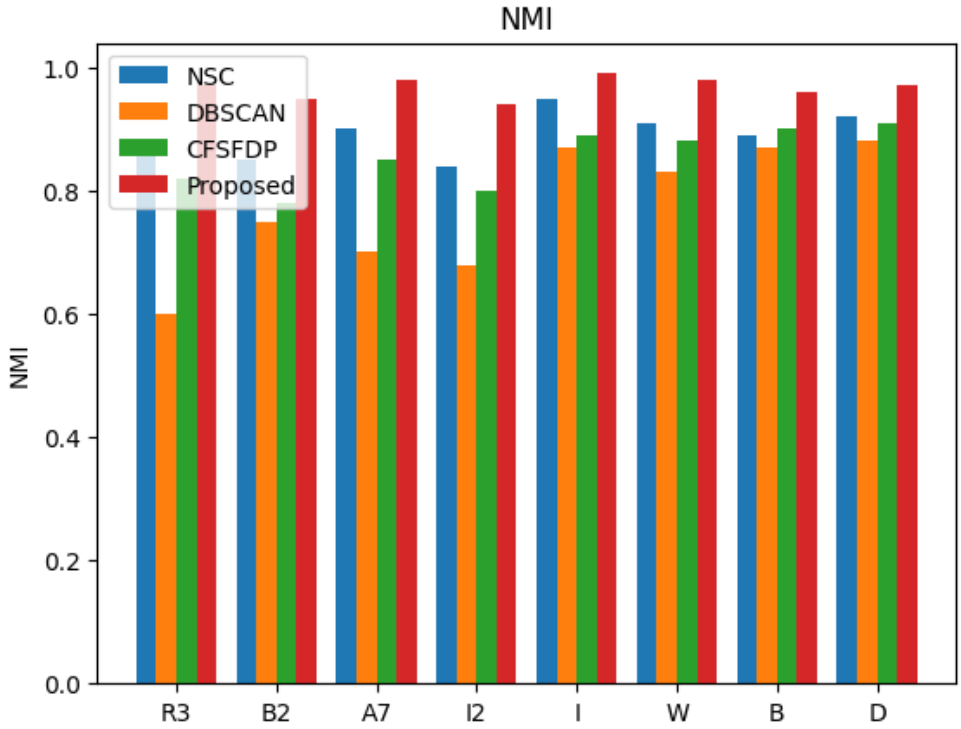
(Figure-7): Applying different algorithms on 4 data sets including 3 types of Fisher's criteria

[DOI: 10.61186/jsdp.21.3.111]



(شکل-۸): نتایج تجربی روش‌های مختلف خوشه‌بندی قوی در مقایسه با نتایج الگوریتم خوشه‌بندی پیشنهادی در مجموعه‌داده‌های مختلف از نظر **ARI**

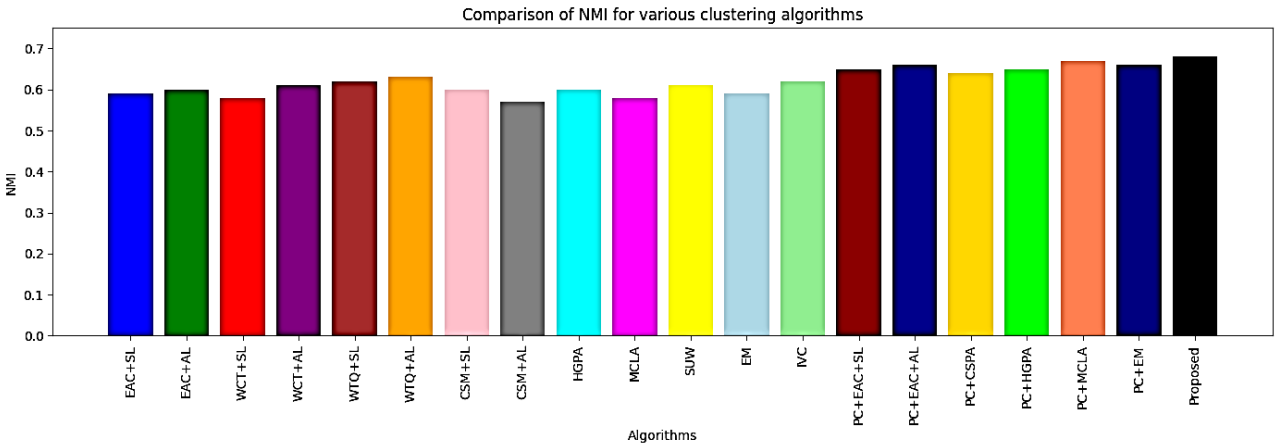
(Figure-8): Experimental results of different robust clustering methods compared to the results of the proposed clustering algorithm in different datasets in terms of ARI



(شکل-۹): نتایج تجربی روش‌های مختلف خوشه‌بندی قوی در مقایسه با نتایج الگوریتم خوشه‌بندی اجماع پیشنهادی در مجموعه‌داده‌های مختلف از نظر **NMI**

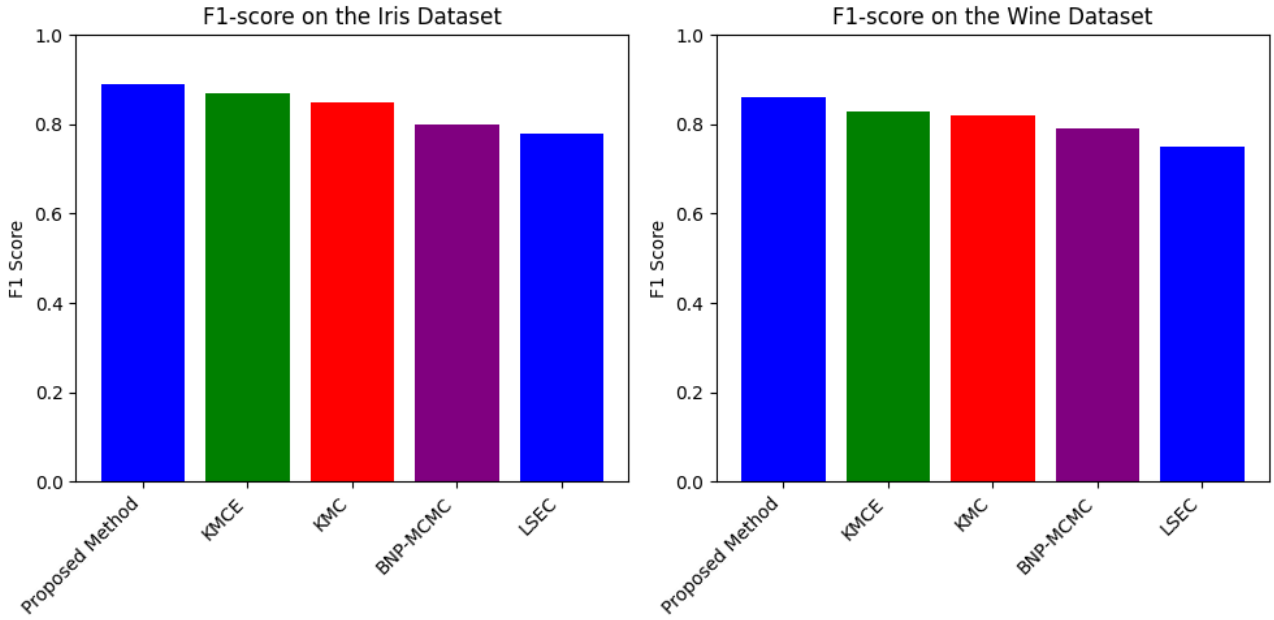
(Figure-9): Experimental results of different robust clustering methods compared to the results of the proposed consensus clustering algorithm in different datasets in terms of NMI

[DOI: 10.61186/jsdp.21.3.111]



(شکل-۱۰): نتایج تجربی روش‌های مختلف خوشه‌بندی قوی در مقایسه با نتایج الگوریتم خوشه‌بندی اجماع پیشنهادی در مجموعه‌داده‌های مختلف از نظر شاخص NMI

(Figure-10): Experimental results of different robust clustering methods compared to the results of the proposed consensus clustering algorithm in different datasets in terms of NMI index



(شکل-۱۱): مقایسهٔ نموداری امتیاز F1 برای روش پیشنهادی در مقابل روش‌های خوشه‌بندی KMCE، KMC، BNP-MCMC، و LSEC بر روی دو مجموعه‌دادهٔ Iris و Wine

(Figure-11): Graphical comparison of F1 score for the proposed method against KMCE, KMC, BNP-MCMC, and LSEC clustering methods on two datasets Iris and Wine

نتایج نمودار شکل (۱۱) نشان می‌دهد، روش پیشنهادی توانایی بالایی در شناسایی خوشه‌های معنی‌دار را دارد این روش، با درنظرگرفتن روابط پیچیده بین نقاط داده، به بهبود جداسازی خوشه‌ها و درنتیجه، به‌دست‌آوردن امتیاز F1 بالاتر کمک کرده‌است. به‌خصوص، در مجموعه‌داده‌های Iris و Wine، روش پیشنهادی به ترتیب امتیازهای F1 برابر با ۰.۸۹ و ۰.۸۶ را کسب کرد که نسبت به سایر روش‌های مورد بررسی مانند KMCE، KMC، BNP-MCMC، و LSEC، عملکرد بهتری داشته‌است. این نتایج نشان می‌دهند که ترکیب چندین خوشه‌بندی در یک راه‌حل خوشه‌بندی نهایی توافقی، محدودیت‌های خوشه‌بندی‌های فردی را کاهش داده و منجر به بهبود کارایی می‌شود.

۶-جمع‌بندی و کارهای آینده

این مقاله به بررسی روش‌های پیشرفتهٔ خوشه‌بندی ترکیبی و به‌کارگیری الگوریتم ژنتیک برای شناسایی و انتخاب پایدارترین خوشه‌ها از میان چندین خوشهٔ اولیه می‌پردازد. با استفاده از تابع توافقی مبتنی بر ماتریس هم‌بستگی و معیار فیشر برای پایداری خوشه، این روش توانسته‌است به ساختار ذاتی و پنهان مجموعه‌های داده‌های بزرگ نفوذ کند و نتایج بهینه‌ای را ارائه دهد.

الگوریتم ژنتیک در این پژوهش نقش مهمی در ارزیابی و انتخاب خوشه‌ها داشته و به‌عنوان مکانیزمی برای

[DOI: 10.61186/jsdp.21.3.111]

رسیدن به خوشه‌بندی توافقی نهایی با کارایی بالا عمل کرده‌است.

نتایج تجربی روی داده‌های متنوع نشان داده‌اند که این روش نه تنها در بهبود پایداری خوشه‌ها موفق عمل کرده، بلکه در مقایسه با روش‌های موجود، دقت قابل توجهی را در معیارهای NMI و ARI به‌دست آورده‌است.

طبق نتایج به‌دست‌آمده، معیار ارزیابی فیشر اصلاح‌شده، در مقایسه با دو معیار دیگر، عملکرد بهتری روی بیشتر مجموعه‌داده‌ها نشان می‌دهد. مطالعهٔ الگوریتم‌های مختلف نشان داد که با استفاده از معیار فیشر اصلاح‌شده، نتایج بهتری نسبت به دیگر معیارها به‌دست می‌آید و الگوریتم پایهٔ EAC عملکرد بهتری روی داده‌ها ارائه می‌دهد.

نرمال‌سازی داده‌ها یکی از اقدامات ضروری در زمانی است که از فاصلهٔ اقلیدسی استفاده می‌شود؛ بااین‌حال، هیچ تضمینی برای بهبود کیفیت خوشه‌بندی با استفاده از الگوریتم‌های نرمال‌سازی داده‌ها وجود ندارد؛ بنابراین، به‌طور معمول روش‌های خوشه‌بندی بر اساس داده‌های خام و غیر نرمال‌سازی‌شده ارزیابی می‌شوند؛ ازاین‌رو، یکی از ایده‌های قابل بررسی در مطالعات آتی می‌تواند پیداکردن یک روش پویا برای اختصاص یک روش نرمال‌سازی به هر مجموعه‌داده باشد[۱۶].

نتایج تجربی الگوریتم خوشه‌بندی جمعی پیشنهادی با چندین الگوریتم خوشه‌بندی جمعی موجود و سه الگوریتم خوشه‌بندی بنیادی قوی بر روی مجموعه‌ای از داده‌های معیار مصنوعی و واقعی مورد مقایسه قرار گرفت. عملکرد الگوریتم خوشه‌بندی جمعی پیشنهادی به‌مراتب کارآمدتر از سایر روش‌های پیشرفته خوشه‌بندی جمعی است؛ علاوه‌براین، توانایی این الگوریتم در مقابله با مجموعه‌های داده در مقیاس بزرگ مورد بررسی قرار گرفته‌است؛ این روش به‌دلیل توانایی در یافتن ساختارهای محلی در خوشه‌های کوچک و ادغام مؤثر آن‌ها متمرکز و کارآمد شناخته شده‌است؛ همچنین، این الگوریتم از ناپایداری‌های الگوریتم‌های خوشه‌بندی پایه برای ایجاد یک مجموعهٔ متنوع بهره می‌برد.

از آنجا که خوشه‌های نهایی ممکن است ضعف‌هایی داشته باشند، باید محدودیت‌ها و قیدهایی را بر آن‌ها اعمال کرد؛ برای مثال، ممکن است در خوشه‌بندی نهایی، یک داده در هیچ خوشه‌ای قرار نگیرد یا خوشه‌ای بدون داده باشد. این محدودیت‌ها باید بر روی خوشه‌بندی اعمال شود تا از صحت و کارایی کار انجام‌شده اطمینان حاصل شود.

۷-مراجع 7-Reference

[1] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: An overview", AI Magazine, vol. 13, no. 3, p. 57, 1992.

[2] D. J. Hand, H. Mannila, and P. Smyth, “Principles of Data Mining”, MIT Press, 2001.

[3] I. H. Witten and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques”, Morgan Kaufmann, 2005.

[4] A. K. Jain and R. C. Dubes, “Algorithms for Clustering”, Englewood Cliffs, NJ: Prentice Hall, 1988.

[5] H. Frigui and R. Krishnapuram, "A robust competitive clustering algorithm with applications in computer vision," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 21, pp. 455–465, 1999.

[6] M. R. Anderberg, "Cluster Analysis for Applications”, Academic Press, Inc., New York, 1973.

[7] E. Diday and J. C. Simon, "Clustering analysis", Digital Pattern Recognition, K. S. Fu, Ed., Springer-Verlag, New York, pp. 47–94, 1976.

[8] R. S. Michalski and R. E. Stepp, "Automated construction of classification: Conceptual clustering versus numerical taxonomy", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 5, pp. 396–409, 1983.

[9] R. C. Dubes, "Cluster analysis and related issues", Handbook of Pattern Recognition & Computer Vision,World Scientific, 1993.

[10] S. Saha and S. Bandyopadhyay, "Application of multiobjective optimization for data clustering", Seminar (Machine Intelligence Unit, Indian Statistical Institute), Kolkata, India, 2008.

[11] N. Jardin and R. Sibson, “Mathematical Taxonomy”, Wiley, New York, 1971.

[12] N. Jardin and C. J. Van Rijsbergen, "The use of hierarchical clustering in information retrieval", Information Storage and Retrieval, vol. 7, pp. 217–240, 1971.

[13] M. Mojarad, S. Nejatian, H. Parvin, and M. Mohammadpoor, "A fuzzy clustering ensemble based on cluster clustering and iterative fusion of base clusters", Applied Intelligence, vol. 49, pp. 2567–2581, 2019.

[14] M. N. Ghaemi, "A survey: Clustering ensembles techniques", World Academy of Science, Engineering and Technology, pp. 636–646, 2009.

[15] H. Parvin, B. Miaei-Bidgoli, H. Alinejad-Rokny, and W. H. Punch, "Data weighing mechanisms for clustering ensembles",

[DOI: 10.61186/jsdp.21.3.111]

Computers & Electrical Engineering, vol. 39, no. 5, pp. 1433–1450, 2013.

[16] H. Parvin and B. Miaei-Bidgoli, "A clustering ensemble framework based on elite selection of weighted clusters," Advances in Data Analysis and Classification, vol. 7, no. 2, pp. 181–208, 2013.

[17] A. Topchy, B. Minaei-Bidgoli, and W. F. Punch, "Ensembles of partitions via data resampling," Proceedings of the International Conference on Information Technology, ITCC 04, Las Vegas, 2004.

[18] A. Topchy, A. K. Jain, and W. F. Punch, "Combining multiple weak clusterings", Proceedings of the 3rd IEEE International Conference on Data Mining, 331–338, 2003.

[19] A. Strehl and J. Ghosh, "Cluster ensembles - A knowledge reuse framework for combining multiple partitions", Journal of Machine Learning Research, vol. 3, pp. 583–617, 2002.

[20] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables", Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, 1996.

[21] J. W. Chang and D. S. Jin, "A new cell-based clustering method for large-high dimensional data in data mining applications", Proceedings of the ACM Symposium on Applied Computing, 503–507, 2002.

[22] R. Miller and Y. Yang, "Association rules over interval data", Proceedings of the ACM SIGMOD International Conference on Management of Data, 452–461, 1997.

[23] R. Jiamthapthaksin, C. F. Eick, and S. Lee, "GAC-GEO: A generic agglomerative clustering framework for geo-referenced datasets," Knowledge and Information Systems, vol. 29, no. 3, pp. 597–628, 2011.

[24] S. Dudoit and J. Fridlyand, "Bagging to improve the accuracy of a clustering procedure", Bioinformatics, vol. 19, no. 9, pp. 1090–1099, 2003.

[25] M. H. F. Zarandi, M. R. Faraji, and M. Karbasian, "An exponential cluster validity index for fuzzy clustering with crisp and fuzzy data," Scientia Iranica, vol. 17, no. 2, pp. 95–110, 2010.

[26] J. C. Bezdek, "Cluster validity with fuzzy sets", Journal of Cybernetics, vol. 3, no. 3, pp. 58–73, 1973.

[27] F. Kovács, C. Legány, and A. Babos, "Cluster validity measurement techniques", Department of Automation and Applied Informatics, Budapest University of Technology and Economics, 2003.

[28] Y. Zhang, W. Wang, X. Zhang, and Y. Li, "A cluster validity index for fuzzy clustering", Information Sciences, vol. 178, no. 4, pp. 1205–1218, 2008.

[29] W. Wang and Y. Zhang, "fuzzy cluster validity indices", Fuzzy Sets and Systems, vol. 158, no. 19, pp. 2095–2117, 2007.

[30] M. J. A. Berry and G. Linoff, "Data Mining Techniques for Marketing", Sales and Customer Support, 3rd ed., John Wiley & Sons, Inc., USA, 1996.

[31] M. Berry, S. Dumais, G. O'Brien, "Using linear algebra for intelligent information retrieval," SIAM Review, vol. 37, no. 4, pp. 573–595, 1995.

[32] M. U. Fayyad, G. Piatesky-Shapiro, and P. Smuth, "Advances in Knowledge Discovery and Data Mining", AAAI Press, 1996.

[33] M. R. Razme Rezaee, B. P. F. Leleiveldt, and J. H. C. Reiber, "A new cluster validity index for the fuzzy c-mean", Pattern Recognition Letters, vol. 19, no. 3–4, pp. 237–246, 1998.

[34] A. Strehl and J. Ghosh, "Cluster ensembles - A knowledge reuse framework for combining multiple partitions", Journal of Machine Learning Research, vol. 3, no. 1, pp. 583–617, 2003.

[35] A. Fred and A. K. Jain, "Combining multiple clustering using evidence accumulation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835–850, 2005.

[36] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 11, pp. 2396–2409, 2011.

[37] A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions", Journal of Machine Learning Research, vol. 3, pp. 583–617, 2002.

[38] Z. Zhou and W. Tang, "Clusterer ensemble", Knowledge-Based Systems, vol. 19, pp. 77–83, 2006.

[39] N. Nguyen and R. Caruana, "Consensus clusterings," Proceedings of the Seventh IEEE International Conference on Data Mining, Omaha, NE, USA, 28–31 October 2007, pp. 607–612.

[40] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm", Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, vol. 14, 2002.

[41] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", KDD-96: Proceedings of the Second International Conference on Knowledge

[DOI: 10.61186/jsdp.21.3.111]

Discovery and Data Mining, AAAI Press, Menlo Park, CA, USA, 1996, pp. 226–231.

[42] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks", Science, vol. 344, no. 6191, pp. 1492–1496, 2014.

[43] T. Alqurashi and W. Wang, "Clustering ensemble method," International Journal of Machine Learning and Cybernetics, vol. 10, no. 6, pp. 1227–1246, 2019.

[44] A. Tombros, R. Villa, and C. Rijsbergen, "The effectiveness of query-specific hierarchic clustering in information retrieval", Information Processing & Management, vol. 38, no. 4, pp. 559–582, 2002.

[45] M. R. Valizadeh and M. Zolghadri-Jahromi, "A proposed query-sensitive similarity measure for information retrieval", Iranian Journal of Science & Technology, Transaction B, Engineering, vol. 30, no. B2, 2006.

[46] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1689–1704, 2005.

[47] X. Z. Fern and W. Lin, "Cluster ensemble selection", Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 1, no. 3, pp. 128–141, 2008.

[48] A. L. Fred and A. K. Jain, "Combining multiple clustering using evidence accumulation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835–850, 2005.

[49] I. T. Christou, "Coordination of cluster ensembles via exact methods", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 2, pp. 279–293, 2011.

[50] V. Singh, L. Mukherjee, J. Peng, and J. Xu, "Ensemble clustering using semi-definite programming with applications", Machine Learning, vol. 79, no. 1–2, pp. 177–200, 2010.

[51] Sh. Zhou, R. Duan, Zh. Chen, W. Song, "Weighted ensemble clustering with multivariate randomness and random", Applied Soft Computing Journal, vol. 150, p. 111015, 2024.

[52] Z. Bian, J. Qu, J. Zhou, Zh. Jiang, Sh. Wang, "Weighted adaptively ensemble clustering method based on fuzzy Co-association matrix", Information Fusion, vol. 103, p. 102099, 2024.

[53] B. Shen, J. Jiang, F. Qian, D. Li, Y. Ye, Gh. Ahmadi, "Semi-supervised hierarchical ensemble clustering based on an innovative distance metric and constraint information", Engineering Applications of Artificial Intelligence, vol. 124, p. 106571, 2023.

[54] Y. Wu, R. Wu, J. Liu, X. Tang, "MetaWCE: Learning to Weight for Weighted Cluster Ensemble", Information Sciences, vol. 629, pp. 39–61, 2023.

[55] J. Xu, T. Li, D. Zhang, J. Wu, "Ensemble clustering via fusing global and local structure information", Expert Systems with Applications, vol. 237, p. 121557, 2024.

[56] Q. Gu, Y. Wang, P. Wang, X. Li, L. Chen, N. N. Xiong, D. Liu, "An improved weighted ensemble clustering based on two-tier uncertainty measurement", Expert Systems With Applications, vol. 238, p. 121672, 2024.

[57] J. Xua, T. Lia, "Ensemble clustering with low-rank optimal Laplacian matrix learning", Applied Soft Computing, 2023, doi: 10.1016/j.asoc.2023.111095.

[58] D. Aktaş, B. Lokman, T. İnkaya, G. Dejaegere, "Cluster ensemble selection and consensus clustering: A multi-objective optimization approach", European Journal of Operational Research, 2023, doi: 10.1016/j.ejor.2023.10.029.

[59] X. Niu, Ch. Zhang, X. Zhao, L. Hu, J. Zhang, "A multi-view ensemble clustering approach using joint affinity matrix", Expert Systems With Applications, vol. 216, p. 119484, 2023.

[60] Y. Jia, S. Tao, R. Wang, Y. Wang, "Ensemble Clustering via Co-Association Matrix Self-Enhancement", IEEE Transactions on Neural Networks and Learning Systems, 2023, ISSN: 2162-2388.

[۶۱] رضایی، حمید و دانشپور، نگین، "ارائه روشی جدید برای خوشه بندی داده های مخلوط بر مبنای تعداد ویژگی مشابه"، مجله پردازش علائم و داده ها، جلد ۲۱ شماره ۱ صفحات ۵۲-۳۹، ۱۴۰۳.

[۶۲] نجفی، فاطمه و پروین، حمید و میرزایی، کمال و نجاتیان، صمد و رضایی، سیده وحیده ، "یک روش خوشه‌بندی ترکیبی جدید مبتنی بر خوشه‌بند cmeansفازی با حفظ تنوع در اجماع"، مجله پردازش علائم و داده ها، جلد ۱۷ شماره ۴ صفحات ۱۰۳-۱۲۲، ۱۳۹۹.

نوید صمیمی بهبهان کارشناسی خود را در سال ۱۳۸۶ از دانشگاه صنعت آب و برق (شهید عباسپور) تهران در زمینهٔ مهندسی کامپیوتر گرایش نرم‌افزار، همچنین کارشناسی‌ارشد و دکترا را در سال‌های ۱۳۸۹ و ۱۴۰۳ از دانشگاه آزاد اسلامی در زمینهٔ مهندسی کامپیوتر- نرم‌افزار دریافت کرده‌است. زمینهٔ پژوهشی وی داده‌کاوی است.

نشانی رایانامهٔ وی عبارت است از:

nvd.samimi@gmail.com

[DOI: 10.61186/jsdp.21.3.111]

صمد نجاتیان تحصیلات خود را در مقطع کارشناسی در رشتهٔ مهندسی برق-الکترونیک دانشگاه سیستان و بلوچستان در سال ۱۳۸۲ به پایان رساند. ایشان مدرک کارشناسی‌ارشد و دکترای خود را به‌ترتیب در سال‌های ۱۳۸۶ و ۱۳۹۳ از دانشگاه‌های فردوسی مشهد و UMT مالزی در رشتهٔ برق گرایش مخابرات دریافت کرد. وی هم‌اکنون عضو هیئت علمی دانشگاه آزاد اسلامی واحد یاسوج است.

نشانی رایانامهٔ وی عبارت است از:
nejatian@iauYasuj.ac.ir

حمید پروین تحصیلات خود را در مقطع کارشناسی در دانشگاه شهید چمران اهواز به پایان رساند. همچنین مدرک کارشناسی ارشد و دکتری را از دانشگاه علم‌وصنعت تهران با گرایش هوش مصنوعی دریافت کرد و سپس به عضویت هیئت علمی دانشگاه آزاد اسلامی در آمد. زمینهٔ پژوهشی وی مباحثی نظیر الگوریتم‌های بهینه‌سازی، داده‌کاوی، هوش مصنوعی و یادگیری عمیق است. تاکنون وی بیش از بیست مقاله با نمایهٔ SCIE (JCR) و چند کتاب به چاپ رسانیده‌است.

نشانی رایانامهٔ وی عبارت است از:
parvin@iust.ac.ir

کرم‌اله باقری فرد مدرک کارشناسی خود را در سال ۱۳۸۴ در رشتهٔ کامپیوتر گرایش نرم‌افزار از دانشگاه اصفهان همچنین مدرک کارشناسی‌ارشد و دکترای خود را به‌ترتیب در سال‌های ۱۳۸۷ و ۱۳۹۵ از دانشگاه نجف‌آباد و اراک در رشتهٔ کامپیوتر با گرایش نرم‌افزار دریافت کرد. وی از سال ۱۳۸۵ تاکنون عضو هیئت علمی دانشگاه آزاد اسلامی واحد یاسوج است. زمینه‌های تخصصی ایشان داده‌کاوی، یادگیری ماشین و سامانه‌های پیشنهاددهنده است. وی تاکنون بیش از هشتاد مقاله در مجلات علمی معتبر و کنفرانس‌های خارجی و داخلی به چاپ رسانیده است.

نشانی رایانامهٔ وی عبارت است از:
k.bagheri@iauYasuj.ac.ir

سیده وحیده رضایی دارای مدرک دکترای ریاضی است. ایشان هم اکنون عضو هیئت علمی دانشگاه آزاد اسلامی واحد یاسوج بوده و بیش از هفتاد مقالهٔ علمی در نشریات و کنفرانس معتبر داخلی و خارجی به چاپ رسانیده است.

نشانی رایانامهٔ وی عبارت‌است از:
vahidehrezaei80@gmail.com

[DOI: 10.61186/jsdp.21.3.111]