

خوشه‌بندی ترکیبی با پیشینه‌سازی پراکندگی با

به کارگیری الگوریتم‌های بهینه‌سازی تکاملی



صدرالله عباسی^۱، صمد نجاتیان^{۲*}، حمید پروین^۳، کرم الله باقری فرد^۴ و وحیده رضایی^۵

^۱گروه کامپیوتر، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

^۲گروه برق، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

^۳گروه کامپیوتر، واحد نورآباد ممسنی، دانشگاه آزاد اسلامی، نورآباد ممسنی، ایران

^۵گروه ریاضی، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

چکیده

خوشه‌بندی داده‌ها یکی از مراحل اصلی در داده‌کاوی است که وظیفه کاوش الگوهای پنهان در داده‌های بدون برچسب را بر عهده دارد. به‌خاطر پیچیدگی مسئله و ضعف روش‌های خوشه‌بندی پایه، امروزه بیشتر مطالعات به‌سمت روش‌های خوشه‌بندی ترکیبی هدایت شده‌است. پراکندگی در نتایج اولیه یکی از مهم‌ترین عواملی است که می‌تواند در کیفیت نتایج نهایی اثرگذار باشد. همچنین، کیفیت نتایج اولیه نیز عامل دیگری است که در کیفیت نتایج حاصل از ترکیب مؤثر است. هر دو عامل در تحقیقات اخیر خوشه‌بندی ترکیبی مورد توجه قرار گرفته‌اند. در این‌جا یک چارچوب جدید برای بهبود کارایی خوشه‌بندی ترکیبی پیشنهاد شده‌است که بر پایه به‌کارگیری زیرمجموعه‌ای از خوشه‌های اولیه هستند، روش ارائه‌شده نشان‌می‌دهد که به‌کارگیری زیرمجموعه‌ای از نتایج خوشه‌بندی‌های اولیه می‌تواند بهتر از به‌کارگیری کل نتایج باشد؛ همچنین معیاری را پیشنهاد می‌دهد که چگونه نتایج اولیه نسبت به هم ارزیابی شوند. این پژوهش معیاری ارائه می‌دهد که به‌وسیله آن می‌توان تشخیص داد کدام زیرمجموعه از نتایج اولیه می‌تواند منجر به بهبود عملکرد خوشه‌بندی ترکیبی شود. از آن‌جایی که الگوریتم‌های هوشمند تکاملی توانسته‌اند بیشتر مسائل پیچیده مهندسی را حل کنند، در این مقاله نیز از این روش‌های هوشمند برای انتخاب زیرمجموعه‌ای از خوشه‌های اولیه استفاده شده‌است. این انتخاب به‌کمک سه روش هوشمند (الگوریتم ژنتیک، شبیه‌سازی تبرید و الگوریتم ازدحام ذرات) انجام می‌گیرد. ایده‌های اصلی در روش‌های پیشنهادی برای انتخاب زیرمجموعه‌ای از خوشه‌ها، به‌کارگیری خوشه‌های پایدار به‌کمک الگوریتم‌های جستجوی هوشمند (الگوریتم‌های تکاملی) هستند. برای ارزیابی خوشه‌ها، از معیار پایداری بر پایه اطلاعات متقابل استفاده شده‌است. در پایان نیز خوشه‌های انتخاب‌شده را به‌کمک چندین روش ترکیب نهایی با هم جمع می‌کنیم. نتایج تجربی روی چندین مجموعه‌داده استاندارد و با معیارهای ارزیابی اطلاعات متقابل نرمال‌شده، فیشر و دقت در مقایسه با روش‌های علیزاده، عظیمی، RCESCC، CLWGC، Berikov، CFSFDP، KME و NSC نشان‌می‌دهد که روش‌های پیشنهادی می‌تواند به‌طور مؤثری روش ترکیب کامل را بهبود دهد.

واژگان کلیدی: بهینه‌سازی محلی، پراکندگی، الگوریتم‌های تکاملی، ماتریس همبستگی، پراکندگی.

The ensemble clustering with maximize diversity using evolutionary optimization algorithms

Sadrollah Abbasi¹, Samad Nejatian^{2*}, Hamid Parvin³, Karamollah Bagherifard⁴ & Vahideh Rezaie⁵

^{1,4}Department of Computer Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran

²Department of Electrical Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran

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

⁵Department of Mathematics, Yasooj Branch, Islamic Azad University, Yasooj, Iran

* Corresponding author

* نویسنده عهده‌دار مکاتبات

Abstract

Data clustering is one of the major tasks in data mining, which is responsible for exploring hidden patterns in unlabeled data. Due to the complexity of the problem and the weakness of the basic clustering methods, today most of the studies are directed towards the clustering ensemble methods. Diversity in the initial results is one of the most important factors that can affect the quality of the final results. Also, the quality of the primary results is another factor that is effective in the quality of the results of the ensemble. Both factors have been taken into consideration in the recent researches of clustering ensemble. Both have been the goal of optimization in various researches. Recently, the simultaneous optimization of these two has also been investigated. But most of the optimization of these two is consecutive. Here, a new framework is proposed to improve the efficiency of clustering ensemble based on the use of a subset of primary clusters. The presented method shows that using a subset of the primary clustering results can be better than using the whole results. Also, this article suggests a criterion on how to evaluate the preliminary results. This research provides a criterion by which it can be determined which subset of the primary results can lead to an improvement in the clustering ensemble performance. Since evolutionary optimization algorithms have been able to solve complex engineering problems, in this article these optimization methods have been used to find the optimal subsets of primary clusters. This selection is done with the help of three optimization methods (genetic algorithm, simulated simulation and particle swarm algorithm). The main ideas in the proposed framework for selecting a subset of clusters are the use of stable clusters with the help of optimization search algorithms (evolutionary algorithms). To evaluate the clusters, the stability criterion based on mutual information has been used. Finally, we combine the selected clusters with the help of several final consensus functions. The experimental results are presented in terms of Fisher and accuracy and normalized mutual information evaluation criteria. The proposed methods are compared to Alizadeh, Azimi, Berikov, CLWGC, RCESCC, KME, CFSFDP, DBSCAB, NSC and Chen methods on several standard data sets. The experimental results show that the proposed methods can effectively improve the state of the art methods. A benchmark of 10 real datasets is used in the experiments. PSO method is significantly better than other methods. The proposed method opens a wide field of studies on the future of clustering algorithms. Introducing the concept of fuzzy in clustering ensemble is one of the first ideas that can be investigated in future studies. Data normalization is one of the necessary measures when using Euclidean distance. Since there is no guarantee to improve the quality of clustering when using normalized data, usually the clustering methods present their reports on raw non-normalized data. Therefore, another idea that can be studied in future studies is to find a dynamic method to assign a normalization method to each data set.

Keywords: Clustering Ensemble, local optimization, evolutionary algorithm, correlation matrix, diversity.

هدف، کمینه کردن تفاوت نمونه‌های هر خوشه و بیشینه کردن تفاوت نمونه‌های یک خوشه با خوشه‌های دیگر است. البته کیفیت نتایج خوشه‌بندی به روش اندازه‌گیری شباهت نیز وابستگی دارد. هر یک از الگوریتم‌های خوشه‌بندی، با توجه به این که بر روی جنبه‌های متفاوتی از داده‌ها تأکید می‌کنند، داده‌ها را به صورت‌های متفاوتی خوشه‌بندی می‌کنند. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با به کارگیری ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کند؛ درواقع هدف اصلی خوشه‌بندی ترکیبی جستجوی بهترین خوشه‌ها با به کارگیری ترکیب نتایج الگوریتم‌های دیگر است [1, 2].

۱-۱- بیان مسأله

به طور کلی مسأله خوشه‌بندی را می‌توان به دو صورت بیان کرد. یک نوع به صورت یک ماتریس $n \times n$ بیان می‌شود که به عنوان یک ماتریس شباهت یا بدون شباهت برای داده‌ها تعریف می‌شود و دیگری یک ماتریس $n \times d$ است

۱- مقدمه

خوشه‌بندی یکی از شاخه‌های یادگیری بدون نظارت است و فرآیند خودکاری است که در طی آن، نمونه‌ها به خوشه‌های افراز می‌شوند که اعضای آن مشابه یکدیگر هستند و با نمونه‌های موجود در دیگر خوشه‌ها بیشینه فاصله را داشته باشند. برای اندازه‌گیری مشابه بودن می‌توان معیارهای مختلفی را در نظر گرفت، برای مثال می‌توان معیار فاصله را برای خوشه‌بندی مورد استفاده قرار داد و اشیایی را که به یکدیگر نزدیک‌تر هستند، به عنوان یک خوشه در نظر گرفت که به این نوع خوشه‌بندی، خوشه‌بندی بر پایه فاصله نیز گفته می‌شود. در روش‌های بدون ناظر متغیر هدفی تعریف نمی‌شود و الگوریتم داده‌کاوی همبستگی‌ها و ساختارهای بین تمام متغیرها را جستجو می‌کند. از مهم‌ترین روش‌های داده‌کاوی بدون ناظر، خوشه‌بندی را می‌توان نام برد. مسأله خوشه‌بندی می‌تواند به صورت یک مسأله بهینه‌سازی فرمول‌بندی شود. نکته کلیدی این است که چه‌طور متغیرهای تصمیم‌گیری و توابع هدف را تعریف کنیم. در خوشه‌بندی کلی‌ترین

۳- انتخاب برخی از ویژگی داده‌ها یا ایجاد ویژگی‌های جدید [1, 11].

۴- تقسیم بندی داده‌های اصلی به زیر مجموعه‌هایی متفاوت و مجزا [7-10].

برای الگوریتم ترکیب‌کننده نتایج برای تولید خوشه‌های نهایی، مطالعات گسترده‌ای صورت گرفته است و مقالات پراکندگی چاپ گردیده‌است [12-15]. پراکندگی در رده‌بندی اطلاعات به این معنا است که اگر یک طبقه‌بندی‌کننده دارای خطا در چند نمونه باشد، آن‌گاه ما به دنبال طبقه‌بندی‌کننده دیگری می‌گردیم که دارای خطاهایی در نمونه‌های متفاوتی از خطاهای طبقه‌بندی‌کننده اول در مجموعه یادگیری باشد تا ترکیب این طبقه‌بندی‌ها نتیجه بهتری را حاصل کند. نبود وجود مجموعه یادگیر این قابلیت را از روش‌های خوشه‌بندی اطلاعات سلب کرده است و ما سعی کرده‌ایم به‌نوعی این مفهوم را وارد بحث خوشه‌بندی اطلاعات کنیم [16-18]. مفهوم پراکندگی به‌طور گسترده‌ای در تحقیقات سال‌های اخیر مورد استفاده قرار گرفته است [19-22]. در زمینه خوشه‌بندی ترکیبی کارهای زیادی انجام گرفته که در ادامه به چند مورد اشاره می‌شود. روش‌های خوشه‌بندی ترکیبی سعی می‌کنند تا با ترکیب افزاینده‌های^۶ مختلف تولیدشده از روش‌های خوشه‌بندی پایه، یک افزاینده مستحکم^۷ از داده‌ها را تولید کنند [1, 5, 11]. در بیشتر مطالعات اخیر، همه افزاینده‌ها با وزن برابر در ترکیب نهایی حاضر می‌شوند و همه خوشه‌های موجود در همه افزاینده‌ها نیز با وزن برابر در ترکیب نهایی شرکت می‌کنند. نویسندگان مرجع [5] یک معیار برای انتخاب از میان ترکیبات ممکن ارائه کرده‌اند که بر پایه کیفیت کلی یک خوشه‌بندی است. برای این کار، آنها میزان ثبات بین افزاینده‌های ترکیبی و افزاینده‌های پایه را در نظر گرفته‌اند و با به‌کارگیری یک قاعده ترکیبی ثابت، یک معیار شباهت دو به دو^۸ را روی فضای ویژگی‌های d -بعدی به کار برده‌اند. عظیمی [1] از مفهوم پراکندگی برای هوشمند کردن خوشه‌بندی ترکیبی استفاده کرده است. در این روش که به‌صورت پویا اقدام به انتخاب زیرمجموعه بهینه‌ای از نتایج اولیه در ترکیب نهایی می‌کند، ابتدا یک خوشه‌بندی ترکیبی ساده انجام می‌شود؛ سپس این روش میزان شباهت تمام نتایج خوشه‌بندی‌های اولیه را نسبت به پاسخ به‌دست‌آمده ارزیابی و سعی در طبقه‌بندی^۹ مجموعه داده‌ها به سه

که هر سطر آن یک شیء را توصیف می‌کند. خروجی الگوریتم خوشه‌بندی نیز می‌تواند به سه صورت بیان شود: یکی گروه‌بندی اشیاء به مجموعه‌های مجزا، مجموعه‌های فازی و دیگری خوشه‌بندی سلسله‌مراتبی، که یک درخت برای تقسیم‌بندی اشیاء پیدا می‌کند. الگوریتم‌های دو نوع اول سریع‌تر از نوع سوم هستند. هر یک از الگوریتم‌های خوشه‌بندی، با توجه به اینکه بر روی جنبه‌های متفاوتی از داده‌ها تأکید می‌کند، داده‌ها را به‌صورت‌های متفاوتی دسته‌بندی می‌نماید. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با به‌کارگیری ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کند؛ درواقع هدف اصلی خوشه‌بندی ترکیبی جستجوی بهترین خوشه‌ها با به‌کارگیری ترکیب نتایج الگوریتم‌های دیگر است [1, 2]. خوشه‌بندی ترکیبی می‌تواند جواب‌های بهتری از لحاظ استحکام^۱، پایداری^۲ و انعطاف‌پذیری^۳ تولید کند [1-4]. به‌اختصار خوشه‌بندی ترکیبی شامل دو مرحله اصلی زیر است:

الف- تولید زیرمجموعه‌های متفاوت از همه نمونه‌های داده، به‌عنوان خوشه‌بندی اولیه بر پایه اعمال الگوریتم‌های متفاوت خوشه‌بندی بر روی زیرمجموعه‌های ایجادشده از نمونه‌های اصلی که این مرحله را مرحله ایجاد پراکندگی یا پراکندگی^۴ می‌نامند.

ب- ترکیب نتایج به‌دست‌آمده از خوشه‌بندی‌های متفاوت اولیه برای تولید خوشه نهایی؛ که این کار به‌وسیله تابع توافقی^۵ (الگوریتم ترکیب‌کننده) انجام می‌شود.

به‌طورکلی در خوشه‌بندی ترکیب دو مسأله مهم باید در نظر گرفته شود. یکی پراکندگی الگوریتم‌های خوشه‌بندی مختلف به‌طوری که هر کدام از این الگوریتم‌های خوشه‌بندی بر ویژگی‌های خاصی از داده‌ها تأکید کنند و دوم الگوریتم ترکیب‌کننده نتایج برای تولید خوشه‌های نهایی است. در رابطه با مسأله اول یعنی به‌دست‌آوردن نتایج پراکندگی که هر کدام بر ویژگی خاصی از داده‌ها تأکید کند، می‌توان از چهار روش مختلفی به شرح زیر استفاده کرد:

- ۱- به‌کارگیری الگوریتم‌های متفاوت خوشه‌بندی [5].
- ۲- تغییر مقادیر اولیه و یا سایر پارامترهای الگوریتم خوشه‌بندی انتخاب‌شده [6].

¹ Robustness

² Stability

³ Flexibility

⁴ Diversity

⁵ Consensus Function

⁶ Partitions

⁷ Robust

⁸ Pairwise

⁹ Classification

مجموعه داده راحتی^۱، معمولی^۲ و سخت^۳ می‌کند. در این طبقه‌بندی، مجموعه داده راحتی به مجموعه داده ای اطلاق می‌شود که خوشه‌بندی‌های اولیه تفاوت چندانی با خوشه‌بندی ترکیبی به‌دست‌آمده نداشته‌باشند. به این معنی که هر خوشه‌بندی ساده بتواند تقریباً مانند خوشه‌بندی ترکیبی نتایج مشابه‌ای ارائه کند. این رویداد نشان‌می‌دهد که داده‌های مجموعه مورد نظر به‌طور کامل دارای مرزهای مشترک هستند و روش‌های ساده و معمولی خوشه‌بندی همانند روش‌های پیچیده و قدرتمند خوشه‌بندی ترکیبی قادر به جداسازی نمونه‌ها نیستند؛ سپس همه نتایج خوشه‌بندی‌های اولیه به چهار زیرمجموعه متفاوت بر پایه میزان تطبیق دقت‌شان با نتایج خوشه‌بندی ترکیبی ساده تقسیم می‌شوند و بر پایه رده^۴ هر مجموعه داده (راحت، معمولی و سخت) اقدام به انتخاب یکی از این زیرمجموعه‌ها برای ترکیب و به‌دست‌آوردن نتیجه نهایی می‌کنیم. نتایج تجربی^۵ نیز نشان داده‌اند که ترکیب خوشه‌بندی‌های اولیه با بیشترین کمترین و میزان متوسطی از تطبیق با خوشه‌بندی ترکیبی اولیه، نتیجه بهتری را به‌ترتیب، در مجموعه داده‌های راحتی، سخت و متوسط می‌دهد. روش بالا در هر مجموعه داده سعی می‌کند تا نتایج خوشه‌بندی اولیه‌ای را که موجب منحرف‌شدن نتایج نهایی می‌شود، از ترکیب نهایی خارج کند و به این ترتیب خوشه‌بندی‌های ترکیبی اولیه‌ای را که دارای دقت به‌نسبه مناسبی هستند، وارد ترکیب نهایی کند [1]. روش دیگری که بسیار به روش ما نزدیک است، روش [2] است که در آنجا برای همه مجموعه داده‌ها ابتدا خوشه‌ها بر پایه پایداری مرتب و سپس ۳۳ درصد پایدارتر انتخاب می‌شوند. این کار در مقایسه با کار این مقاله یک الگوریتم قطعی است. بامگارتن^۶ و همکاران [23] یک روش بر پایه بازنمونه‌برداری را برای بررسی اعتبارسنجی نتایج خوشه‌بندی فازی ارائه کرده‌اند. در چند سال اخیر، پایداری خوشه به‌عنوان یک معیار ارزیابی خوشه مورد توجه زیادی قرار گرفته است^۸. ایده‌های اولیه برای اعتبارسنجی خوشه با به‌کارگیری بازنمونه‌برداری در [26] ارائه شده و بعدها در [27] کامل‌تر شده‌است. نویسندگان مرجع [28] روشی برای تعیین تعداد خوشه‌ها ارائه کرده‌اند که به‌طور خودکار ویژگی‌ها را وزن‌دهی می‌کند. نویسندگان مرجع [15] روشی برای خوشه‌بندی ترکیبی

¹ Easy

² Intermediate

³ Hard

⁴ Class

⁵ Experimental Results

⁶ Baumgartner

پیشنهاد کرده‌اند که از زیرمجموعه‌ی مؤثرتری از افزازهای اولیه در ترکیب نهایی استفاده می‌کند. در این روش اگر چه تعداد اعضای شرکت‌کننده در ترکیب نهایی کمتر از یک خوشه‌بندی ترکیبی کامل^۷ است، به‌دلیل انتخاب افزازهای با کارایی بالاتر، نتایج نهایی بهبود می‌یابند. پارامترهایی که در این روش مورد توجه قرار گرفته‌اند، عبارتند از: کیفیت و پراکندگی. این روش سعی می‌کند تا زیرمجموعه‌ای از افزازهایی از نتایج اولیه را وارد ترکیب نهایی کند که از بالاترین میزان کیفیت برخوردار بوده و در عین حال نسبت به هم بیشترین پراکندگی را دارا باشند. در این روش از معیار مجموع اطلاعات متقابل نرمال‌شده (SNMI)^۸ (برای یک افزاز در مقایسه با افزازهای دیگر ترکیب) برای اندازه‌گیری کیفیت یک افزاز استفاده شده است.

۱-۲- هدف اصلی و نوآوری روش پیشنهادی

هدف اصلی در اکثر روش‌های اخیر خوشه‌بندی ترکیبی، تنها بررسی مجموعه داده از زوایای مختلف است و این سؤال که "آیا پراکندگی به وجود آمده مفید است یا نه؟" چندان مورد توجه قرار نگرفته‌است. در حقیقت به‌خاطر ماهیت بدون ناظر بودن مسأله خوشه‌بندی مطالعه این امر با دشواری‌های زیادی روبه‌روست. اگرچه نتایج تجربی نشان داده‌اند که ایجاد پراکندگی در خوشه‌بندی‌های اولیه به‌طور معمول موجب بهبود خوشه‌بندی در اکثر مواقع می‌شود اما در مراجع [29, 30] نشان داده‌شده‌است که در بعضی مجموعه داده‌ها، پراکندگی بیشتر لزوماً کمکی به افزایش دقت در نتایج نهایی نمی‌کند.

هدف اصلی اکثر مطالعاتی که تاکنون در زمینه خوشه‌بندی انجام گرفته است، جستجوی روش‌هایی است که بتواند پراکندگی نتایج مجمع رده‌بندها را بهتر کند. در این بررسی‌ها، این اصل که آیا پراکندگی به وجود آمده مفید است یا نه، چندان مورد نظر قرار نگرفته‌است. در حقیقت به‌خاطر ماهیت بدون ناظر بودن مسأله خوشه‌بندی، مطالعه زیادی در این زمینه صورت نگرفته‌است. نتایج تجربی نشان داده است که ایجاد پراکندگی در خوشه‌بندی‌های اولیه به‌طور معمول، موجب بهبود خوشه‌بندی در اکثر مواقع می‌شود [31].

نوآوری اصلی در این مقاله تأکید همزمان بر پراکندگی و کیفیت خوشه‌بندی‌های اولیه برای انتخاب بوده است. عامل دیگری که به‌طور معمول برای بهبود عملکرد خوشه‌بندی ترکیبی از آن استفاده شده‌است،

⁷ Full Ensemble

⁸ Sum of Normalized Mutual Information

کیفیت نتایج اولیه است. نشان داده شده است که هر چه نتایج اولیه علاوه بر داشتن پراکندگی لازم، از کیفیت بالاتری برخوردار باشند، کیفیت خوشه‌های نهایی نیز بهتر خواهد بود [22]. اگرچه نویسندگان [15] نشان داده‌اند که بهینه‌سازی همزمان دو عامل پراکندگی و کیفیت در نتایج اولیه خوشه‌بندی ترکیبی می‌تواند کارایی خوشه‌بندی ترکیبی را به‌طور چشم‌گیری بهبود بخشد، تنظیم و مصالحه بین این دو عامل مسأله‌ای است که حل دقیق آن هنوز با دشواری‌های فراوانی روبه‌روست. در این مقاله تأکید همزمان بر پراکندگی و کیفیت خوشه‌بندی‌های اولیه برای انتخاب است.

۱-۳-۲- روش‌های جستجو مکاشفه

افزایش روز افزون ابعاد و پیچیدگی‌های مسائل بهینه‌سازی مهندسی سبب کاهش کارایی روش‌های معمول و احساس نیاز به روش‌های نوین جستجو گردیده است. از این رو در دهه‌های اخیر، روش‌های تکاملی به‌عنوان یک ابزار جستجو و بهینه‌سازی در بسیاری از حوزه‌ها توسعه یافته و مورد استفاده قرار گرفته‌اند. وسعت دامنه کاربرد، سهولت استفاده و قابلیت دستیابی به جواب نزدیک به بهینه مطلق از جمله دلایل موفقیت این روش‌ها است. در این بخش دو الگوریتم بر پایه روش‌های تکاملی مورد استفاده در این مقاله به‌طور اجمالی بررسی می‌شود.

۱-۳-۱- الگوریتم ژنتیک

الگوریتم ژنتیک^۱ رهیافتی است که تکامل طبیعی موجودات را الگو قرار می‌دهد [32]. این روش تقلیدی از فرایند تکامل با به‌کارگیری الگوریتم‌های رایانه‌ای است. اساسی‌ترین اصل تکامل، وراثت است. مبتکر الگوریتم ژنتیک جان هلند در دهه هفتاد میلادی با الهام گرفتن از ویژگی‌های تئوری تکامل، الگوریتم جستجوی ابداع کرد که در این الگوریتم از همان اصولی که طبیعت فرایند تکامل را روی نمادهای ژنی انجام می‌دهد [32]، برای تکامل جواب‌های مربوط به حل مسأله بهینه‌سازی استفاده می‌کند. دو جنبه مهم در الگوریتم ژنتیک وجود دارند که دائماً جواب‌ها را آشفته کرده و مجال خروج از بهینه‌های موضعی را می‌آورند. یکی از این جنبه‌ها آمیزش است که A را با آن برای تولید جواب استفاده می‌کند. جنبه دیگر که منجر به جهش نام دارد، قادر است مقادیر جدیدی به جواب‌ها بدهد که در گروه والدین وجود نداشته است. عمل جهش کمک

می‌کند که پراکندگی ژنتیک باقی بماند و جستجو به نواحی جدیدی برسد. پارامترهای الگوریتم ژنتیک شامل اندازه جمعیت ۱۰۰۰، تعداد نسل‌های ۵۰۰ و طول کروموزوم برابر با ۱۲۰ برابر تعداد خوشه‌های واقعی به‌علاوه ۱۸۰ است. همچنین از احتمال جهش ۰/۰۱ عملگر تقطیع یک‌نواخت و عملگر انتخاب مرتب‌سازی استفاده شده است.

۱-۳-۲- الگوریتم مورد شبیه‌سازی شده

مورد شبیه‌سازی شده یک روش بهینه‌سازی است که به‌جهت شباهت آن به فرایند حرارت فلزات و سرد کردن آرام آن‌ها به این اسم نامیده می‌شود [33]. این روش برای توابع هدفی مؤثر است که ساده و فقط دارای یک نقطه حدی موضعی باشند (برای مسائل کمینه یا بیشینه‌سازی). برای توابع پیچیده، به‌طور مثال برای مسائل بیشینه‌سازی، این نقطه بهینه موضعی ممکن است با بهینه عمومی به‌طور کامل متفاوت باشد و مدل بهینه‌سازی قادر به ارائه جواب‌های بهینه مورد نظر نباشد. SA با به‌کارگیری روش رهاسازی تصادفی توانایی خارج شدن از نقاط حداقل، موضعی را دارد که در ادامه تشریح می‌گردد.

فرآیند SA از یک جواب امکان‌پذیری مانند q_0 (یک بردار حقیقی که نشان‌دهنده کلیه متغیرهای تصمیم است) و تابع هدف متناظر $J_0 = J(q_0)$ شروع می‌شود. یک جواب جدید q_1 با تابع هدف $J_1 = J(q_1)$ به‌صورت تصادفی از همسایگی جواب اولیه انتخاب و مورد ارزیابی قرار می‌گیرد. میزان تغییر در متغیر تصمیم عموماً مشخص است. ماهیت تصادفی به‌دلیل جهت یا بعد تغییر است (برای مثال، در حالتی ممکن است، مقدار تغییر x مشخص باشد، ولی جهت آن به‌صورت تصادفی تعیین شود). اگر جواب جدید دارای مقدار تابع هدف کمتری باشد $J_1 < J_0$ (برای مسائل کمینه‌سازی)، این جواب پذیرفته شده و عمل جستجو به نقطه q_1 منتقل می‌شود. اگر جواب جدید بهتر از جواب فعلی نباشد ($J_1 \geq J_0$)، جواب جدید ممکن است انتخاب یا رد شود که این امر بستگی به احتمال پذیرش زیر دارد.

$$p_{acc} = e^{-\frac{J_1 - J_0}{T}} \quad (1)$$

الگوریتم مورد شبیه‌سازی شده نیز با $T=0/9$ انجام پذیرفته است. میزان خطای دو جواب پی‌درپی الگوریتم مورد شبیه‌سازی شده نباید کمتر از ۰/۰۰۱ باشد ($\varepsilon=0/001$). این الگوریتم هم از همان نمایش کروموزوم الگوریتم ژنتیک و تابع برازندگی آن استفاده می‌کند.

۱-۳-۳- الگوریتم ازدحام ذرات

الگوریتم ازدحام ذرات (PSO) [34] با یک گروه از جواب‌های تصادفی شروع به کار می‌کند؛ سپس برای یافتن جواب بهینه در فضای مسأله با به‌روزر کردن موقعیت و سرعت هر ذره به جستجو می‌پردازد. هر ذره به‌صورت چندبعدی (بسته به طبیعت مسأله) با دو مقدار x_i^d, v_i^d که به ترتیب معرف مکان و سرعت مربوط به بعد d -ام از i -امین ذره هستند تعریف می‌شود. در هر مرحله از حرکت جمعیت، هر ذره با توجه به دو مقدار بهترین به‌روز می‌شود. نخستین مقدار بهترین، بهترین جواب از لحاظ شایستگی است که تاکنون برای هر ذره به‌طور جداگانه به‌دست آمده است. این مقدار بهترین برای هر فرد^۱ است و $pbest_i$ نامیده می‌شود. مقدار بهترین دیگر که به‌وسیله PSO به‌دست می‌آید، بهترین مقداری است که تاکنون به‌وسیله تمام ذره‌ها در میان جمعیت به‌دست آمده است، این مقدار بهترین سراسری^۲ است و $gbest$ نام دارد. پس از یافتن دو مقدار $pbest_i$ و $gbest$ هر ذره سرعت و مکان جدید خود را با رابطه (۲) به‌روز می‌کند:

$$v_i^d = wv_i^d + n_1r_1(x_{pbest_i}^d - x_i^d) + n_2r_2(x_{gbest}^d - x_i^d) \quad (2)$$

$$x_i^d = x_i^d + v_i^d$$

x_i^d و $x_i'^d$ به ترتیب مکان فعلی و مکان قبلی مربوط به بعد d -ام از i -امین ذره هستند. v_i^d و $v_i'^d$ سرعت فعلی و سرعت قبلی مربوط به بعد d -ام از i -امین ذره هستند؛ به‌طوری که $w \in (0, 1)$ و وزن اینرسی^۳ است، n_1 و n_2 ضرایب شتاب^۴، r_1 و r_2 اعداد تصادفی با توزیع یکنواخت در بازه (۰ و ۱) هستند. برای جلوگیری از واگرایی الگوریتم، مقدار نهایی سرعت هر ذره در بازه $[-V_{MAX}, V_{MAX}]$ محدود می‌شود. w, n_1r_1 و n_2r_2 از پارامترهای PSO هستند و هم‌گرایی الگوریتم وابسته به مقدار این پارامترهاست. به‌طورمعمول مقدار n_1, n_2 عددی بین ۱/۵ تا ۲ است ولی بهترین انتخاب $n_1 = 2.05$ است. هم‌گرایی به‌شدت به مقدار w وابسته و بهتر است به‌صورت پویا تعریف شود.

۲- مروری بر ادبیات موضوع و کارهای گذشته

ترکیب خوشه‌بندی‌ها کار مشکل‌تری نسبت به ترکیب رده‌بندی‌های بانظر است. به عبارت دیگر، برخلاف مسأله

رده‌بندی که دارای ناظر^۵ و یک مجموعه یادگیری^۶ است، در خوشه‌بندی هیچ‌گونه شناختی نسبت به مجموعه‌داده وجود ندارد. عدم وجود ناظر و مجموعه یادگیری، ارائه روش‌های مدرن و هوشمند خوشه‌بندی داده‌ها که دارای کارایی بالا باشند را بسیار مشکل نموده‌است. همچنین، در غیاب داده آموزشی برچسب‌دار، ما با مشکل تناظر بین برچسب‌های خوشه در افرازهای مختلف از یک ترکیب مواجه هستیم. خوشه‌بندی ترکیبی روشی در خوشه‌بندی است که از ترکیب نتایج روش‌های خوشه‌بندی متفاوت به‌دست می‌آید. دو گام اساسی در تولید یک اجماع از خوشه‌بندی‌های اولیه عبارتند از: تولید هر یک از اعضای اجماع و به‌کارگیری یک تابع توافقی یا مکانیسمی (در ساده‌ترین حالت رأی‌گیری) برای جمع بندی نتایج اجماع برای به‌دست آوردن نتیجه نهایی. از آنجایی که نتیجه نهایی خوشه‌بندی ترکیبی از خوشه‌بندی‌های اولیه به‌دست می‌آید، هرچه خوشه‌بندی‌های اولیه نتایج متفاوت‌تری ارائه دهند نتیجه نهایی بهتری حاصل می‌شود؛ درواقع هرچه داده‌ها از جنبه‌های متفاوت‌تری مطالعه و بررسی شوند نتیجه نهایی که از ترکیب این نتایج حاصل می‌شود متعاقباً دارای دقت بالاتری خواهد بود. راه‌های مختلفی برای ایجاد پراکندگی در خوشه‌بندی ترکیبی وجود دارد که عبارتند از: به‌کارگیری الگوریتم‌های متفاوت خوشه‌بندی، تغییر مقادیر اولیه و یا سایر پارامترهای الگوریتم خوشه‌بندی انتخاب‌شده، انتخاب بعضی از ویژگی داده‌ها یا ایجاد ویژگی‌های جدید و تقسیم‌بندی داده‌های اصلی به زیرمجموعه‌هایی متفاوت و مجزا. در روش‌های ارائه‌شده هدف تنها بررسی مجموعه‌داده از زوایای مختلف است و این اصل که آیا پراکندگی به‌وجودآمده مفید است یا مفید نیست چندان مورد نظر نیست.

به‌طورمعمول بیشتر روش‌های خوشه‌بندی ترکیبی از الگوریتم k-means جهت خوشه‌بندی اولیه خود استفاده می‌کنند [6, 19, 35]. اما در روش‌های ارائه‌شده نشان داده‌شده‌است که با توجه به رفتار هر مجموعه‌داده گاهی اوقات یک روش خوشه‌بندی خاص پیدا می‌شود که دقت بهتری از k-means برای بعضی از مجموعه‌داده‌ها می‌دهد [11]. اما الگوریتم k-means به دلیل سادگی و توانایی مناسب در خوشه‌بندی همواره به‌عنوان انتخاب نخست مطالعات خوشه‌بندی ترکیبی مورد مطالعه قرار گرفته است. یکی دیگر از راه‌های افزایش پراکندگی تغییر پارامترهای اولیه الگوریتم‌های خوشه‌بندی است. برای مثال تغییر تعداد خوشه‌ها در الگوریتم K-means و یا

⁵ Supervisor

⁶ Train

¹² Personal Best

² Global Best

³ Inertia Wight

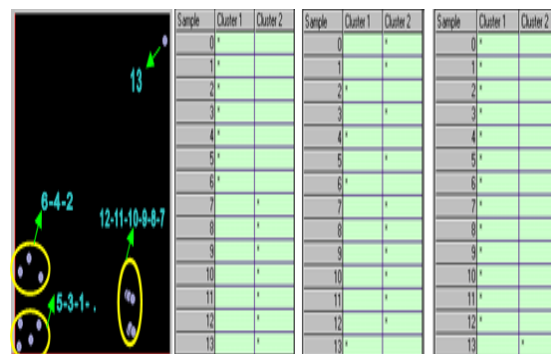
⁴ Acceleration Coefficients

رویکردی که در چندین سال اخیر بسیار به آن پرداخته‌شده، رویکرد بر پایه برش (ابر) گراف است. در این رویکرد، ابتدا مساله یافتن خوشه‌بندی توافقی به یک مساله افراز گراف تبدیل می‌شود؛ سپس به‌کمک الگوریتم‌های افراز یا برش گراف، خوشه‌های نهایی توافقی به‌دست می‌آیند [38-36]. چهار الگوریتم ترکیبی ابرگراف معروف CSPA، HGPA، MCLA و HBGF هستند. رویکردی دیگر، رویکرد رأی‌گیری است، [41-39، 45]. برای این منظور باید ابتدا عمل بازبرچسب‌گذاری انجام شود. عمل باز برچسب‌گذاری به‌منظور هم‌سان‌سازی برچسب‌های خوشه‌بندی‌های گوناگون برای تطابق است. از رویکردهای مهم دیگر می‌توان به موارد روبه‌رو اشاره کرد [42-47]: (۱) رویکرد در نظر گرفتن خوشه‌بندی‌های اولیه به‌عنوان یک فضای واسط (یا مجموعه‌داده جدید) و خوشه‌بندی این فضای جدید به‌کمک یک الگوریتم خوشه‌بند پایه شبیه الگوریتم پیشینه‌سازی انتظار [7]، (۲) رویکرد به‌کارگیری الگوریتم‌های تکاملی برای یافتن سازگارترین خوشه‌بندی به‌عنوان خوشه‌بندی توافقی [44]، رویکرد به‌کارگیری الگوریتم خوشه‌بند kmods برای یافتن خوشه‌بندی توافقی [48-49].

بسیار محتمل است که یک پارتیشن وجود داشته‌باشد که با به‌کارگیری یک اندازه‌گیری پایداری به‌عنوان یک پارتیشن بد قضاوت شود در حالی‌که دارای یک (یا بیشتر) خوشه با کیفیت بالا است [50]؛ بنابراین، پژوهش‌گران با الهام از ارزیابی پارتیشن‌ها، اقدامات لازم برای ارزیابی خوشه‌ها را تعیین می‌کنند. بسیاری از اقدامات پایداری مانند اطلاعات متقابل نرمال برای اعتبارسنجی یک پارتیشن پیشنهاد شده‌است. اقدامات تعریف‌شده بر پایه اطلاعات متقابل نرمال است. اشکال رویکرد متداول در این مقاله مورد بحث قرار خواهد گرفت و ملاکی برای ارزیابی ارتباط بین یک خوشه و یک پارتیشن ارائه شده‌است که به آن معیارهای ENMI گویند. معیار ENMI اشکال اندازه‌گیری معمول اطلاعات متقابل نرمال (NMI) را جبران می‌کند. همچنین، یک روش گروه‌بندی خوشه‌ای که بر پایه جمع‌کردن زیر مجموعه‌ای از خوشه‌های اولیه است، ارائه می‌دهد [50].

برخلاف برخی از تلاش‌ها برای بهبود کیفیت روش‌های خوشه‌بندی، به نظر می‌رسد که پژوهش‌های اندکی به رویه انتخاب در گروه خوشه‌بندی فازی اختصاص یافته‌است؛ علاوه‌براین، کیفیت و پراکندگی محلی بسترهای اصلی دو عامل مهم در انتخاب خوشه‌بندی‌های پایه است. تعداد کمی از مطالعات، این دو عامل را برای انتخاب بهترین خوشه‌بندی‌های پایه فازی در گروه در

تغییر نمونه‌های اولیه^۱ الگوریتم تأثیر به‌سزایی در افزایش پراکندگی در خوشه‌بندی دارند و در خوشه‌بندی اطلاعات نقش مهمی ایفا می‌کند. در شکل (۱) اثر نمونه‌های اولیه در خوشه‌بندی نهایی به وضوح قابل مشاهده‌است. در شکل (۱) در سمت چپ ابتدا نحوه توزیع نمونه‌ها نمایش داده‌شده‌است و سپس نتایج سه اجرای مختلف الگوریتم با سه نمونه شروع مختلف نمایش داده‌شده‌است.



(شکل-۱): نمونه‌های اولیه در نتایج الگوریتم k-means

شکل‌ها به ترتیب از چپ به راست عبارتند از: ۱- نمایش فضایی ۱۴ نمونه پراکنده در فضا. ۲- نتایج به‌دست آمده با نمونه‌های اولیه ۱ و ۸. ۳- نتایج به‌دست آمده با نمونه‌های اولیه ۲ و ۳. ۴- نتایج به‌دست آمده با نمونه‌های اولیه ۱ و ۱۳.

(Figure-1): The prototypes in the k-means algorithm results for the figures, from left to right, as follows: 1. Spatial representation of 14 scattered samples in space. 2- Results obtained with the initial samples 1 and 8. 3- Results obtained with the initial samples 2 and 3. 4- Results obtained with the initial samples 1 and 13.

انتخاب ویژگی‌ها نیز می‌تواند در خوشه‌بندی ترکیبی به‌عنوان یک منبع تولید پراکندگی در نظر گرفته شود؛ پس یکی دیگر از راه‌کارهای افزایش پراکندگی در خوشه‌بندی ترکیبی به‌کارگیری برخی از ویژگی‌های کل فضای مجموعه‌داده و یا تولید ویژگی‌های جدید است؛ ولی در خوشه‌بندی اطلاعات به‌دلیل ماهیت بدون ناظر بودن مسئله، انتخاب زیر مجموعه‌ای از ویژگی‌ها کمتر مورد توجه بوده و بیشتر سعی در تولید ویژگی‌های جدید بوده است. روش‌های گوناگونی برای تولید ویژگی و به‌کارگیری آن در خوشه‌بندی ترکیبی وجود دارد که ساده‌ترین آن‌ها نرمال‌سازی داده‌ها است. در واقع نشان داده‌شده‌است که هر مجموعه‌داده‌ای با یک روش نرمال‌سازی رفتار بهتری نشان می‌دهد و به همین دلیل در خیلی از روش‌های ارائه‌شده در خوشه‌بندی اطلاعات، نتایج بر طبق داده‌های خام گزارش می‌شوند. در ادامه به برخی از جدیدترین روش‌های ارائه‌شده در چندین سال اخیر اشاره شده‌است.

¹ Seed Points

نظر گرفته‌اند. در [51] یک چارچوب گروه‌بندی خوشه‌بندی فازی جدید بر اساس یک اندازه‌گیری پراکندگی فازی جدید و اندازه‌گیری کیفیت فازی پیشنهاد شده تا خوشه‌های پایه را با بهترین عملکرد پیدا کنند. پراکندگی و کیفیت بر اساس اطلاعات متقابل عادی فازی بین خوشه‌بندی‌های پایه فازی تعریف شده‌است.

در [52]، یک چارچوب جدید گروه خوشه‌بندی بر اساس وزن‌گیری در سطح خوشه ارائه شده‌است. مقدار اطمینان این گروه در مورد یک خوشه، به‌عنوان قابلیت اطمینان آن خوشه در نظر گرفته شده‌است. مقدار قطعیتی که گروه خاص در مورد یک خوشه دارد با میزان جمع‌آوری آن خوشه توسط گروه محاسبه؛ سپس با انتخاب بهترین خوشه‌ها و تعیین وزنی به هر خوشه انتخابی بر اساس قابلیت اطمینان آن، مجموعه نهایی ایجاد می‌شود. پس از آن، مقاله به‌جای ماتریس توافقی سنتی، ماتریس توافقی وزنی در سطح خوشه را پیشنهاد می‌کند؛ سپس دو عملکرد اجماع برای تولید پارتیشن اجماع معرفی و مورد استفاده قرار گرفته‌است.

در این مقاله [53] سعی شده‌است یک تابع تجمعی، به‌نام خوشه‌بندی جمعی مقاوم، بر پایه نمونه‌برداری و خوشه‌بندی خوشه‌ای (RCESCC) ارائه شود؛ سپس، یک ماتریس شباهت خوشه خوشه‌ای از خوشه‌های فازی به‌دست می‌آورد. پس از آن، با به‌کارگیری یک الگوریتم خوشه‌بندی سلسله‌مراتبی بر روی ماتریس شباهت خوشه خوشه‌ای، خوشه‌های فازی را تقسیم می‌کند. در مرحله بعدی، الگوریتم RCESCC نقاط داده را به خوشه‌های ادغام‌شده اختصاص می‌دهد.

در این مطالعه [54]، یک رویکرد جدید خوشه‌بندی با به‌کارگیری یک رویکرد وزنی ارائه شده‌است. در این مقاله روشی برای انجام خوشه‌بندی جمعی با بهره‌برداری از مفهوم عدم اطمینان خوشه ارائه شده‌است؛ درواقع، هر خوشه براساس عدم وابستگی آن محاسبه شده‌است. همه برچسب‌های خوشه‌ای پیش‌بینی‌شده موجود در گروه برای ارزیابی یک عدم وابستگی خوشه‌ای، از اندازه‌گیری بر پایه تئوری اطلاعات استفاده می‌کنند. در این مقاله دو روش بر پایه عدم وابستگی یا عدم اطمینان از خوشه برای برآورد قابلیت اطمینان ارائه شده‌است. این مقاله دو رویکرد ارائه شده‌است: جمع‌آوری شواهد با وزن خالص و تقسیم‌بندی گراف با وزن خوشه.

در [55] یک رویکرد جدید ترکیب خوشه‌بندی‌های فازی با به‌کارگیری یک رویکرد وزنی ارائه شده‌است. در این مقاله وزن هر خوشه متناظر با قابلیت اتکای هر خوشه در مجموعه خوشه‌بندی‌های پایه اندازه‌گیری شده‌است.

در [56] یک مدل احتمالی از مجموعه خوشه‌بندی فازی بر اساس ماتریس مشارکتی وزن‌دار توسط بریکوف

پیشنهاد شده‌است. در این مدل، هر یک از خوشه‌بندی‌های پایه توسط الگوریتم‌های خوشه‌بندی متفاوتی تولید می‌شوند.

۳- روش پیشنهادی

فرض می‌شود X ، مجموعه داده شامل N نمونه باشد:

$$X = \{x_1, x_2, \dots, x_N\} \quad (2)$$

اگر $\pi_j(x_i)$ خروجی متناظر با j -امین الگوریتم خوشه‌بندی پایه بر روی نمونه x_i باشد، داریم:

$$x_i \rightarrow \{\pi_1(x_i), \pi_2(x_i), \dots, \pi_H(x_i)\} \quad (3)$$

$\pi_j(x_i)$ ، $j = 1, 2, \dots, H$ و $i = 1, 2, \dots, N$ خروجی‌های حاصل از اجرای k -means هوشمند بر روی x_i است. فضای ویژگی جدید دارای H بعد^۱ است. هر بار اجرای الگوریتم خوشه‌بندی پایه متناظر با یک بعد در فضای ویژگی جدید خواهد بود. اگر مجموعه X شامل N نمونه با m ویژگی باشد، مجموعه جدید ایجادشده، X' ، مجموعه‌ای متشکل از N نمونه با H ویژگی خواهد بود:

$$\begin{aligned} X &\equiv \{x_1, x_2, \dots, x_N\} & X &= \{x'_1, x'_2, \dots, x'_N\} \\ x_i &= \{x_{i1}, x_{i2}, \dots, x_{im}\} & \Rightarrow & x'_i = \{x'_{i1}, x'_{i2}, \dots, x'_{iH}\} \\ i &= 1, 2, \dots, N & i &= 1, 2, \dots, N \end{aligned} \quad (4)$$

مقادیر ویژگی‌ها در فضای جدید با به‌کارگیری مکانیزمی انجام می‌گیرد که در بخش بعد معرفی می‌گردد. بعد از اینکه نمونه‌ها در فضای ویژگی جدید ایجاد شدند، خوشه‌بندی نهایی با به‌کارگیری ساده یکی از روش‌های خوشه‌بندی پایه انجام می‌گیرد.

۳-۱- تابع توافقی

جمع‌بندی خوشه‌بندی‌های اولیه و دستیابی به نتیجه نهایی یکی از مهم‌ترین مراحل خوشه‌بندی ترکیبی است. روش‌های گوناگونی برای ترکیب نتایج خوشه‌بندی‌های اولیه مختلف و ایجاد خوشه‌بندی نهایی وجود دارد که در زیر به معرفی چند روش جدید و معروف در این زمینه می‌پردازیم و در ادامه اقدام به معرفی روش پیشنهادی ارائه شده در این قسمت می‌کنیم.

۳-۱-۱- روش بر پایه ابر گراف

در روش بر پایه ابر گراف^۲ ما ابتدا مسأله خوشه‌بندی ترکیبی را به یک مسأله افراز گراف تبدیل، سپس آن را به کمک الگوریتم‌های افراز گراف حل می‌کنیم. خوشه‌ها با

¹ Dimention

² Hyper Graph Partitioning

شماره خوشه‌ها در اجراهای متفاوت است که سربار^۸ محاسباتی سنگینی را به الگوریتم تحمیل می‌کند. این سربار اضافی محاسباتی عاملی می‌شود تا این روش را یک روش کم مصرف در بین روش‌های توابع توافقی گوناگون تبدیل کند.

۳-۱-۳- ماتریس همبستگی

فرض کنید مجموعه داده D شامل N نقطه (نمونه) در فضای d بعدی است. داده‌های ورودی را می‌توان به صورت یک ماتریس الگوی $N \times d$ و یا یک ماتریس عدم تشابه $N \times N$ در نظر گرفت. فرض کنید $X = \{X_1, X_2, \dots, X_{B1}\}$ مجموعه زیرمجموعه نمونه‌های در دسترس است که از نمونه‌های اولیه استخراج شده‌اند. هر یک از الگوریتم‌های انتخابی هنگامی که بر روی زیر مجموعه نمونه‌های موجود در X اجرا شوند نتایج $P = \{P_1, P_2, \dots, P_{B1}\}$ را تولید می‌کنند. هر P_i مجموعه‌ای از خوشه‌ها است؛ یا به عبارتی دیگر $C_{k(i)}^i = \{C_1^i \cup C_2^i \dots \cup C_{k(i)}^i\}$ و $P_i = \{C_1^i \cup C_2^i \dots \cup C_{k(i)}^i\}$ به طوری که $k(i)$ تعداد خوشه‌ها در i -امین خوشه‌بندی است.

نخستین الگوریتم پایه مورد استفاده الگوریتم k -means است. در نخستین گام، الگوریتم k -means را بر روی $X = \{X_1, X_2, \dots, X_{B1}\}$ اجرا می‌کنیم تا بتوانیم با به‌کارگیری P_i ‌های تولیدشده ماتریس همبستگی^۹ را به صورت زیر به دست آوریم:

که در این رابطه داریم:

$$Co - association(xy) = \sum_{i=1}^{B1} \lambda(P_i(x), P_i(y)) \quad (5)$$

$$\lambda(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \quad (6)$$

تابع $\lambda(P_i(a), P_i(b))$ در صورتی که دو عنصر a و b در خوشه‌بندی P_i در یک خوشه قرار گرفته باشند، مقدار یک و در غیر این صورت مقدار صفر را بر می‌گرداند. مقدار پارامتر $B1$ نمایانگر تعداد زیر مجموعه‌ها و یا به بیان دیگر تعداد دفعات تکرار الگوریتم پایه k -means است. بعد از به دست آوردن ماتریس همبستگی با به‌کارگیری الگوریتم ساده سلسله‌مراتبی نظیر AG^{10} (پیوند میانگین) اقدام به استخراج خوشه‌های نهایی از ماتریس همبستگی می‌شود.

۳-۲- توضیحات مربوط به روش پیشنهادی

در این بخش ما از دل مشکلات قبل، یک رویکرد ارائه می‌دهیم که هم پراکندگی را بهینه کند و هم دقت را در نظر داشته باشد

ابریه‌های^۱ یک گراف نمایش داده می‌شوند. رأس‌های^۲ گراف معادل نمونه‌هایی هستند که باید خوشه‌بندی شوند. مسأله تکه‌تکه کردن این گراف و ایجاد k قسمت منفک است که هر قطعه مربوط به یک خوشه می‌شود. سه نوع الگوریتم متفاوت در این خانواده وجود دارد که عبارتند از^۳ CPSA،^۴ HGPA،^۵ MCLA،^۶ [5, 11].

الف - CPSA

در CPSA فضای ویژگی نقاط داده‌ای در ابتدا به فضای ویژگی همبستگی ابرگراف نگاشت می‌شود. سپس یک الگوریتم کمینه برش ابرگراف شبیه METIS بر نقاط داده‌ای به تازگی فاصله دار شده به کار برده می‌شود. همانند قبل این روش فرض می‌کند که هر چه نقاط داده‌ای بیشتری در یک خوشه در افراز اولیه باشد، احتمال بیشتری دارد که آن نقاط داده‌ای ذاتاً متعلق به یک خوشه باشند. CPSA ساده‌ترین مکاشفه است. پیچیدگی محاسباتی آن $O(kN^2M)$ است که k تعداد خوشه‌ها، N تعداد نقاط داده‌ای و M تعداد نواحی است. دو روش بعدی پیچیدگی محاسباتی کمتری دارند.

ب - HGPA

الگوریتم HGPA فرض می‌کند که رأس‌ها نقاط داده‌ای و خوشه‌هایی که از افراز اولیه بیرون آمده‌اند، ابريال‌های آن هستند. حال دوباره یک الگوریتم ابرگراف کمینه برش شبیه METIS بر روی آن ابرگراف جهت جداسازی رأس‌ها یعنی نقاط داده‌ای ابرگراف به k مولفه متفاوت به کار برده می‌شود. پیچیدگی محاسباتی آن $O(kNM)$ است که دوباره k تعداد خوشه‌ها، N تعداد نقاط داده‌ای و M تعداد نواحی است.

ج - MCLA

الگوریتم MCLA ابتدا خوشه به دست آمده از افراز اولیه را افراز و سپس از یک سازوکار بر پایه رأی‌گیری جهت تولید افرازهای مجمع استفاده می‌کند. خوشه‌بندی خوشه با به‌کارگیری METIS انجام شده است. پیچیدگی محاسباتی آن $O(k^2NM^2)$ است که k ، N و M مانند قبل هستند. جهت جزئیات بیشتر در مورد روش‌های بر پایه ابرگراف به [5] مراجعه کنید.

۳-۱-۲- روش رأی‌گیری

روش رأی‌گیری^۶ این روش همان روش رأی پیشینه^۷ است. به این صورت که خوشه هر نمونه بر اساس رأی پیشینه تعیین می‌شود. مشکل اساسی این روش مسأله تطبیق

¹ Hyper Edges

² Vetexes

³ Cluster-based Similarity Partitioning Algorithm

⁴ Hyper Graph-Partitioning Algorithm

⁵ Meta-CLustering Algorithm

⁶ Voting

⁷ Majority of Vote

⁸ Overhead

⁹ Co-association Matrix

¹⁰ Average Link

نشان‌دهنده تعداد داده‌هایی که مشترکاً در خوشه i -ام از افراز P و در خوشه j -ام از افراز L قرار دارد، N تعداد کل داده‌ها را نشان می‌دهد و τ یک جایگشت از اعداد یک تا N است. اگر دو افراز P و L به‌طور کامل مشابه باشند، آنگاه FM مقدار بیشینه یعنی یک و اگر دو افراز به‌طور کامل متفاوت از یکدیگر باشند، مقدار صفر را برمی‌گرداند. گفتنی است که پایداری افراز $Ensemble_i$ را به‌شکل زیر محاسبه می‌کنیم.

(۸)

$$Stability(Ensemble_i) = \frac{1}{|RefSet|} \sum_{j=1}^{|RefSet|} FM(Ensemble_i, RefSet_j)$$

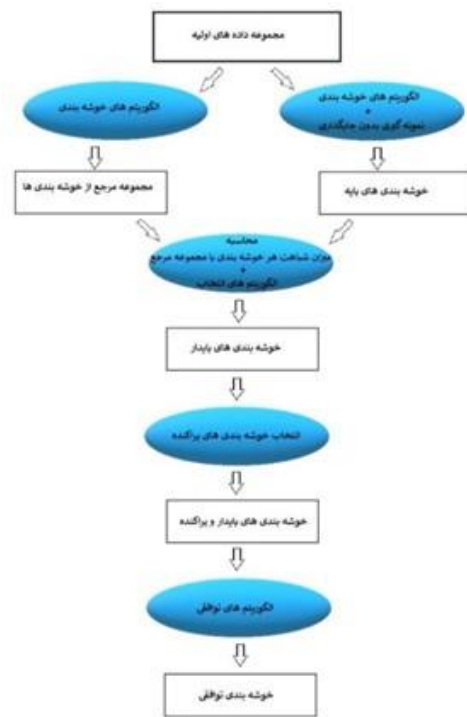
سپس خوشه‌بندی‌ها را بر اساس پایداری و البته پراکندگی مورد کاوش قرار دهیم تا خوشه‌بندی‌های پایدارتر و م پراکندگی‌تر انتخاب شوند. پس از اینکه پایداری هر خوشه محاسبه شد، در گام بعد، عمل انتخاب خوشه‌ها با توجه به مقدار پایداری خوشه انجام می‌شود. یک روش بر پایه الگوریتم‌های تکاملی برای انتخاب زیرمجموعه‌ای از خوشه‌های اولیه ارائه شده‌است که در این بخش به تشریح این الگوریتم می‌پردازیم.

در این جا عمل انتخاب خوشه‌ها در دو مرحله انجام می‌پذیرد. ابتدا در مرحله یک، یک الگوریتم تکاملی سعی در یافتن زیرمجموعه‌ای از خوشه‌ها دارد که بیشترین پایداری را داشته باشند. این الگوریتم تکاملی دارای یک کروموزوم بیتی به طول تعداد کل خوشه‌های تولیدشده در بخش تولید خوشه‌بندی‌های گوناگون است. هر کدام از ژن‌های این کروموزوم می‌تواند عدد یک یا صفر را به خود بگیرد. عدد یک نشان‌دهنده آن است که خوشه‌ای به شماره آن ژن در بین خوشه‌های انتخاب‌شده باشد و عدد صفر در یک ژن شماره m یعنی خوشه‌ی m -ام در بین خوشه‌های انتخاب‌شده نباشد. برای محاسبه تابع برازندگی این الگوریتم تکاملی، اختلاف میزان پایداری میانگین خوشه‌های انتخاب‌شده از عدد یک (بیشینه میزان پایداری میانگین خوشه‌های انتخاب‌شده است)، محاسبه می‌شود. برای انجام این کار، ابتدا مثالی را مطرح می‌کنیم. فرض کنید که ۱۳ داده زیر را داریم:

(جدول-۱): یک مجموعه داده با ۱۳ داده فرضی
(Table-1): A data set with 13 hypothetical data

دیس	1	2	3	4	5	6	7	8	9	10	11	12	13
ژگی	1	1	2	2	3	4	5	4	3	2.5	3	2.5	3.5
ژگی	1	2	1	2	3	5	4	4	2	2	2.5	3	3.5

این داده‌ها در شکل (۳) در فضای ویژگی‌ها ارائه شده‌اند.



(شکل-۲): چارچوب پیشنهادی

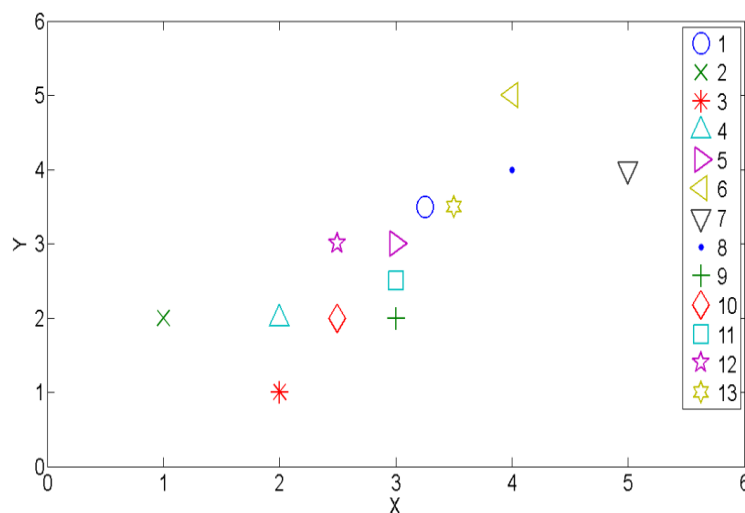
(Figure-2): The proposed framework.

روندنامی روش پیشنهادی در شکل (۲) نشان داده شده‌است. در این روش ما ابتدا یک مجموعه (اجماع) از خوشه‌بندی‌های اولیه به نام مجموعه مرجع یا RefSet تولید می‌کنیم. اجماع RefSet با اندازه $|RefSet|$ است که تعداد عناصر این مجموعه را نشان می‌دهد. گفتنی است که RefSet نشان‌دهنده i -امین عضو از این اجماع است؛ سپس یک اجماع دیگر به نام اجماع اصلی یا Ensemble تولید می‌کنیم. گفتنی است که Ensemble نشان‌دهنده i -امین عضو از این اجماع است. حال برای هر یک از Ensemble که i از یک تا B تغییر کند، میزان پایداری را محاسبه می‌کنیم. پایداری افراز Ensemble متوسط میزان مشابه بودن این افراز در مجموعه مرجع است. میزان مشابه بودن دو افراز از رابطه معروف معیار فیشر اندازه گیری می‌شود. حال باید بگوییم که معیار فیشر چگونه محاسبه می‌شود. این معیار که در این مقاله برای ارزیابی یک افراز در نظر گرفته شده‌است، معیار فیشر یا (F - Measure) است:

$$FM(P, L) \quad (Y)$$

$$= \max_{\tau} \sum_{i=1}^{K_P} \frac{2 \times N_i^P \times (\frac{N_{tr(i)}^{PL}}{N_i^P} \times \frac{N_{tr(i)}^{PL}}{N_{\tau(i)}^L})}{N \times (\frac{N_{tr(i)}^{PL}}{N_i^P} + \frac{N_{tr(i)}^{PL}}{N_{\tau(i)}^L})}$$

که K_P تعداد خوشه‌های افراز P ، N_i^P نشان‌دهنده تعداد داده‌های موجود در خوشه i -ام از افراز P ، N_j^L نشان‌دهنده تعداد داده‌های موجود در خوشه j -ام از افراز L ، N_{ij}^{PL}



(شکل-۳): نمایش مجموعه داده جدول در فضای ویژه‌گی‌ها
(Figure-3): Displays the table data set in the properties space

الگوریتم k -means اجرا شده است، تعداد خوشه‌های مفروض سه بوده است.

جدول (۲) یک اجماع $RefSet$ با اندازه دلخواه ۷ را نشان می‌دهد. در جدول (۲) مقدار مستطیل خاکستری رنگ، نشان‌گر آن است که در خوشه‌بندی $RefSet_2$ داده شماره ۱۱ (چرا که در سطر ۱۱ است) به خوشه ۱ متعلق شده است.

فرض کنید می‌خواهیم که این سیزده داده را به سه خوشه افراز کنیم. ابتدا باید یک مجموعه $RefSet$ با اندازه دلخواه (که در این مثال ۷ را اختیار کرده‌ایم) تولید کنیم. با اعمال روش خوشه‌بندی k -means بر روی این مجموعه داده، ابتدا یک مجموعه $RefSet$ با اندازه ۷ تولید می‌کنیم. برای تولید این اجماع الگوریتم k -means را ۷ بار بر روی این داده‌ها اجرا کرده‌ایم. در همه بارهایی که

(جدول-۲): یک اجماع $RefSet$ با اندازه ۷

(Table-2): A $RefSet$ ensemble with size 7

اندیس	$RefSet_1$	$RefSet_2$	$RefSet_3$	$RefSet_4$	$RefSet_5$	$RefSet_6$	$RefSet_7$
1	2	1	3	3	3	3	3
2	1	2	1	1	1	1	2
3	1	2	1	1	1	1	2
4	1	2	1	1	1	1	2
5	2	1	2	2	2	3	1
6	3	3	3	3	3	2	3
7	3	3	3	3	3	2	3
8	3	3	3	3	3	2	3
9	2	1	2	2	2	3	1
10	1	2	2	2	2	1	1
11	2	1	2	2	2	3	1
12	2	1	2	2	2	3	1
13	2	1	3	3	3	3	3

می‌کنیم. به دلیل آنکه تعداد داده‌ها ۱۳ است، تعداد داده‌های نمونه‌برداری شده برابر ۱۰ خواهد شد و در اجرای الگوریتم k -means اول و ساخت $Ensemble_1$ فقط ۱۰ داده تصادفی انتخاب شده دخالت خواهند داشت (چرا که $round(13 \times 0.8) = 10$). حال فرض کنید برای اجرای اول الگوریتم و تولید $Ensemble_1$ ، داده‌های ۱، ۲، ۴، ۵، ۶، ۷، ۹، ۱۰، ۱۱ و ۱۳ انتخاب شوند. پس از اجرای الگوریتم خوشه‌بندی k -means بر روی این داده‌ها با تعداد خوشه معلوم ۳، افراز به صورت جدول (۳) به دست می‌آید.

اکنون یک اجماع هشت‌عضوی تولید می‌کنیم. برای تولید کردن اجماع باید فاکتور پراکندگی را در نظر داشته باشیم. برای آنکه مجمع ما تا حد ممکن پراکنده باشد، از روش نمونه‌برداری استفاده خواهد شد. با کمک روش نمونه‌برداری بدون جای‌گذاری، کیسه‌هایی به اندازه دلخواه $sampling_rate$ (که نشان‌گر نرخ نمونه برداری است) که در اینجا ۸۰٪ در نظر گرفته شده است، از داده‌ها را پر می‌کنیم. به طور ساده بر روی هر کیسه، یک الگوریتم خوشه‌بندی k -means با تعداد خوشه مفروض ۳ اجرا

(جدول ۳-): اجرای الگوریتم خوشه‌بندی k-means بر روی

۸۰ درصد از داده‌های شکل (۲)

(Table-3): Implementation of k-means clustering algorithm on 80% of Figure 2 data

Ensemble ₁
3
1
-
1
3
2
2
-
1
1
3
-
3

دقت شود که بعضی داده‌ها (داده‌هایی با اندیس ۳، ۸ و ۱۲) در این افراز مقدار "-" دارند. مقدار "-" یعنی خوشه این داده مشخص نیست؛ چراکه این داده در نمونه‌برداری موجود نبود. این فقدان بعضی از عناصر در این جدول در آینده باعث به‌وجود آمدن مشکلاتی خواهد شد. پس بهتر است به جای آنکه این مسأله را در گام‌های بعدی مدیریت کنیم، همین جا مسأله فقدان برچسب بعضی داده‌ها را مرتفع کرد. پس ابتدا مرکز خوشه ۱ در این افراز Ensemble₁ را به‌دست می‌آوریم. برای به‌دست آوردن مرکز خوشه ۱ در این افراز، میانگین داده‌های شماره ۲، ۴، ۹ و ۱۰ (همان داده‌هایی که در این افراز Ensemble₁ به خوشه ۱ متعلق شده‌اند) را مرکز خوشه ۱ در نظر می‌گیریم.

$$C_1 = \frac{(1,2) + (2,2) + (3,2) + (2.5,2)}{4} = \frac{(8.5,8)}{4} = (2.125,2) \quad (9)$$

که C_1 مرکز خوشه اول در افراز Ensemble₁ است. حال مراکز دیگر را نیز محاسبه می‌کنیم.

$$C_2 = (4.5,4.5) \quad (10)$$

$$C_3 = (2.625,2.5) \quad (11)$$

اکنون داده‌هایی که خوشه‌های آنها موجود نیستند، یعنی داده‌های ۳، ۸ و ۱۲ را باید به یکی از خوشه‌ها منسوب کنیم. ابتدا داده ۳ را در نظر بگیریم. فاصله داده ۳ تا مرکز خوشه C_1 از رابطه زیر برابر ۱/۰۱ خواهد بود.

$$\begin{aligned} |C_1 - X_3| &= |(2.125,2) - (2,1)| \\ &= |(0.125,1)| \\ &= \sqrt{0.125^2 + 1^2} \end{aligned} \quad (12)$$

فاصله داده ۳ تا مرکز خوشه‌های C_2 و C_3 نیز به‌طور مشابه ۴/۳۰ و ۱/۶۳ خواهد بود. پس این داده به خوشه ۱ منتسب می‌شود؛ چرا که مقدار فاصله داده یادشده با خوشه نخست از همه کمتر است. به همین ترتیب متوجه خواهیم شد که داده شماره ۸ به خوشه ۲ و داده شماره ۱۲ به خوشه ۳ باید منتسب شوند. حال بقیه اجماع ۸ هشت‌عضوی به‌صورت جدول (۴) تولید می‌شوند. پس از انتساب‌دهی مقادیر مفقود، جدول به‌صورت جدول (۵) بازسازی خواهد شد. این اجماع را اجماع اول (First Ensemble) می‌گوییم (FE_1) یعنی خوشه‌بندی نخست از اجماع نخست که همان Ensemble₁ است بعد از عمل مدیریت برچسب‌های مفقود). حال معیار فیشر را برای افراز FE_1 و $RefSet_1$ محاسبه می‌کنیم:

$$\begin{aligned} FM(FE_1, RefSet_1) & \quad (13) \\ &= \max_{\tau} \sum_{i=1}^{K_P} \frac{2 \times N_i^P \times \left(\frac{N_{it(i)}^{PL}}{N_i^P} \times \frac{N_{\tau(i)}^{PL}}{N_{\tau(i)}^L} \right)}{N \times \left(\frac{N_{it(i)}^{PL}}{N_i^P} + \frac{N_{\tau(i)}^{PL}}{N_{\tau(i)}^L} \right)} \\ &= 0.9389 \end{aligned}$$

پس از محاسبه تمام مقادیر $FM(FE_1, RefSet_j)$ که j از ۱ تا ۷ تغییر خواهد کرد، $Stability(FE_1)$ را محاسبه می‌کنیم:

$$Stability(FE_1) = 0.8214 \quad (14)$$

سپس $Stability(FE_i)$ را برای سایر خوشه‌بندی‌ها به‌دست می‌آوریم:

$$Stability(FE_2) = 0.8873 \quad (15)$$

$$Stability(FE_3) = 0.8873 \quad (16)$$

$$Stability(FE_4) = 0.9155 \quad (17)$$

$$Stability(FE_5) = 0.8873 \quad (18)$$

$$Stability(FE_6) = 0.7405 \quad (19)$$

$$Stability(FE_7) = 0.8214 \quad (20)$$

$$Stability(FE_8) = 0.8873 \quad (21)$$

(جدول-۴): یک اجماع با ۸ اندازه

(Table-4): A ensemble of 8 sizes

Ensemble ₈	Ensemble ₇	Ensemble ₆	Ensemble ₅	Ensemble ₄	Ensemble ₃	Ensemble ₂	Ensemble ₁
1	3	3	1	-	2	2	3
2	1	2	2	1	1	3	1
2	1	2	2	1	-	3	-
-	1	2	2	1	1	3	1
1	3	-	1	-	2	-	3
3	2	-	3	3	3	1	2
3	-	1	3	3	3	-	2
-	2	3	-	3	3	1	-
-	-	2	-	-	2	2	1
2	1	2	-	2	-	-	1
1	-	2	1	2	-	2	3
1	3	-	1	2	2	2	-
1	3	3	1	3	2	2	3

(جدول-۵): اجماع اول (Ensemble₁) برای اجماع جدول (۴)(Table-5): The first Ensemble (Ensemble₁) for consensus Table (4)

FE ₈	FE ₇	FE ₆	FE ₅	FE ₄	FE ₃	FE ₂	FE ₁
1	3	3	1	3	2	2	3
2	1	2	2	1	1	3	1
2	1	2	2	1	1	3	1
2	1	2	2	1	1	3	1
1	3	3	1	2	2	2	3
3	2	3	3	3	3	1	2
3	2	1	3	3	3	1	2
3	2	3	3	3	3	1	2
1	1	2	1	2	2	2	1
2	1	2	2	2	1	3	1
1	3	2	1	2	2	2	3
1	3	2	1	2	2	2	3
1	3	3	1	3	2	2	3

(جدول-۶): اجماع جدول پس از مرحله دو

(Table-6): Table ensemble after step two

SE ₄	SE ₃	SE ₂	SE ₁
1	3	2	2
2	1	1	3
2	1	1	3
2	1	1	3
1	2	2	2
3	3	3	1
3	3	3	1
3	3	3	1
1	2	2	2
2	2	1	3
1	2	2	2
1	2	2	2
1	3	2	2

انتخاب شود، سه تا از این چهار افراز (یعنی افرازهای ۲، ۳، ۵ و ۸) به‌طور تصادفی انتخاب می‌شوند. حال اجماع دوم (یا Second Ensemble) به‌صورت جدول (۶) خواهد بود (SE₁ یعنی عضو نخست از اجماع دوم).

دقت شود که خوشه‌بندی‌های SE₁، SE₂، SE₃ و SE₄ به‌ترتیب همان خوشه‌بندی‌های FE₂، FE₃، FE₄ و FE₅ بوده است. پس از اینکه پایداری هر خوشه‌بندی محاسبه شد، و در گام بعد، عمل انتخاب خوشه‌بندها با توجه به مقدار پایداری خوشه انجام شد، آنگاه یک روش بر پایه الگوریتم‌های تکاملی برای انتخاب زیرمجموعه‌ای از خوشه‌بندی‌های اولیه ارائه شده‌است که در این بخش به

حال خوشه‌بندی‌های گوناگون FE_i را بر اساس مقادیر پایداری آنها مرتب می‌کنیم. بنا به پژوهش‌های قبلی (علیزاده و همکاران، ۲۰۱۴؛ پروین و مینایی، ۲۰۱۵) با به‌کارگیری ۵۰٪ از پایدارترین خوشه‌بندی‌ها بهترین نتایج را می‌توان به‌دست‌آورد. پس ۵۰٪ برتر خوشه‌بندی‌ها را با مرتب‌کردن این ۸ خوشه‌بندی به‌دست می‌آوریم. خوشه‌بندی‌های ۲، ۳، ۴ و ۵ پایدارترین خوشه‌بندی‌ها هستند. دقت شود که خوشه‌بندی ۴ پایدارترین خوشه‌بندی‌ها و خوشه‌بندی‌های ۲، ۳، ۵ و ۸ در رده بعدی پایدارترین خوشه‌بندی‌ها هستند که در این صورت به‌علت آنکه به‌طور دقیق باید ۴ (۵۰٪ از ۸ افراز برابر ۴ افراز است) افراز

$$FitnessFunction = 0.5 - \frac{\sum \sum abs(Co(x,y) - 0.5)}{N^2} \quad (22)$$

که N تعداد داده‌ها و $Co(x,y)$ میزان همبستگی داده‌های x و y در اجماع ثالث (یا $Third\ Ensemble$) که با TE نیز نمایش داده می‌شود) است که از رابطه زیر محاسبه می‌شود.

$$Co(x,y) = \frac{\sum_{i=1}^{|TE|} \lambda(TE_i(x), TE_i(y))}{|TE|} \quad (23)$$

(جدول-۷): اجماع جدول پس از مرحله سه

(Table-7): Table ensemble after step three

TE_3	TE_2	TE_1
1	3	2
2	1	3
2	1	3
2	1	3
1	2	2
3	3	1
3	3	1
3	3	1
1	2	2
2	2	3
1	2	2
1	2	2
1	3	2

که TE_i افراز i -ام در اجماع ثالث است و تابع λ در زیر آورده شده است.

$$\lambda(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \quad (24)$$

حال اجازه دهید که این معادله را برای محاسبه میزان کارایی در مثال قبل به کار بگماریم تا ببینیم که کارایی کروموزوم مثال قبل چقدر خواهد شد. از آنجایی که بنا به کروموزوم اجماع ثالث فقط شامل خوشه‌بندی‌های ۱، ۳ و ۴ از اجماع ثانی است، پس اجماع ثالث یا $Third\ Ensemble$ به شکل زیر است (TE_i یعنی خوشه‌بندی i -ام از اجماع ثالث).

تشریح این الگوریتم می‌پردازیم. این الگوریتم‌های تکاملی دارای یک کروموزوم بیتی به طول تعداد کل خوشه‌بندی‌ها موجود در اجماع ثانی در بخش تولید خوشه‌بندی‌های گوناگون، است. هر کدام از ژن‌های این کروموزوم می‌تواند عدد یک یا صفر را به خود بگیرد. یعنی کروموزوم ما از جنس بیتی هستند. عدد یک نشان‌دهنده آن است که خوشه‌بندی به شماره آن ژن در بین خوشه‌بندی‌های انتخاب‌شده باشد و عدد صفر یعنی خوشه‌بندی به شماره آن ژن در بین خوشه‌بندی‌ها انتخاب‌شده نباشد. به عبارتی عدد یک در ژن i -ام نشان‌دهنده آن است که خوشه‌بندی SE_i در بین خوشه‌های انتخاب‌شده باشد و عدد صفر در یک ژن شماره m یعنی خوشه‌بندی m -ام در بین خوشه‌بندی‌ها انتخاب‌شده نباشد. برای محاسبه تابع برازندگی این الگوریتم تکاملی، میزان پراکندگی خوشه‌بندی‌ها انتخاب‌شده محاسبه می‌شود. برای تشریح بیشتر، به مثال قبل برمی‌گردیم. شکل (۴) می‌تواند یک کروموزوم الگوریتم تکاملی در نظر گرفته شود.

۱	۱	۰	۱
---	---	---	---

(شکل-۴): نمایش یک راه حل کاندید (کروموزوم)

(Figure-4): Showing a solution (chromosome)

طول کروموزوم نشانگر تعداد خوشه‌بندی‌های اجماع ثانی است. در این کروموزوم، به دلیل آنکه ژن‌های ۱، ۳ و ۴ مقدار یک دارند، فقط خوشه‌بندی‌های ۱، ۳ و ۴ از اجماع ثانی انتخاب می‌شوند و در اجماع ثالث گذاشته می‌شوند و خوشه‌بندی‌های ۲ به دلیل آنکه ژن متناظر (یعنی ژن دوم) در کروموزوم صفر است از اجماع کنار گذاشته می‌شود در اجماع ثالث استفاده نمی‌شود. حال برای محاسبه کارایی این کروموزوم از رابطه (۲۲) که میزان پراکندگی بودن آرا را نشان می‌دهد، استفاده می‌کنیم.

(جدول-۸): ماتریس نهایی همبستگی افرازهای جدول (۷)

(Table-8): The final correlation matrix of table partitions (7)

1	0.67	0.67	0	0.67	0.33	0.33	0.33	0.67	0	0	0	1
0	0	0	0.67	0	0	0	0	0	1	1	1	0
0	0	0	0.67	0	0	0	0	0	1	1	1	0
0	0	0	0.67	0	0	0	0	0	1	1	1	0
0.67	1	1	0.33	1	0	0	0	1	0	0	0	0.67
0.33	0	0	0	0	1	1	1	0	0	0	0	0.33
0.33	0	0	0	0	1	1	1	0	0	0	0	0.33
0.33	0	0	0	0	1	1	1	0	0	0	0	0.33
0.67	1	1	0.33	1	0	0	0	1	0	0	0	0.67
0	0.33	0.33	1	0.33	0	0	0	0.33	0.67	0.67	0.67	0
0.67	1	1	0.33	1	0	0	0	1	0	0	0	0.67
0.67	1	1	0.33	1	0	0	0	1	0	0	0	0.67
1	0.67	0.67	0	0.67	0.33	0.33	0.33	0.67	0	0	0	1

پس از جستجوی سراسری خواهیم دید که بهترین کروموزوم، کروموزوم زیر است. میزان تابع کارایی بر روی این کروموزوم ۲۱ است.

۰	۰	۱	۱
---	---	---	---

(شکل-۶): نمایش بهترین راه‌حل (کروموزوم)
(Figure-6): Showing the best solution (chromosome)

ماتریس نهایی همبستگی بهینه که با کروموزوم شکل (۶) تعریف شده‌است، در جدول (۹) آورده شده‌است. در گام آخر ماتریس همبستگی به‌دست‌آمده از اجماع ثالث بهینه، به‌عنوان یک ماتریس مشابهت در نظر گرفته می‌شود. در این‌صورت یک الگوریتم خوشه‌بندی سلسله‌مراتبی به‌عنوان تابع جمع‌کننده نهایی در نظر گرفته می‌شود و ماتریس همبستگی به‌دست‌آمده را به‌عنوان ورودی گرفته و خوشه‌بندی توافقی نهایی را بر می‌گرداند. حالت دیگر این است که ماتریس همبستگی به‌دست‌آمده را به‌عنوان یک مجموعه داده جدید در نظر گرفت و یک خوشه‌بندی در این فضا انجام داد. در این مجموعه داده جدید، هر ستون را به‌عنوان یک ویژگی و هر سطر را به‌عنوان یک داده در نظر می‌گیریم. این مجموعه داده جدید را **فضای واسط** می‌نامیم؛ سپس در این فضای واسط یک الگوریتم خوشه‌بندی k-means یا fuzzy k-means را انجام می‌دهیم.

ابتدا ماتریس همبستگی را تولید می‌کنیم. همان‌طور که می‌بینید، داده ۱ و ۲ هیچ‌گاه هم خوشه نبوده‌اند. یعنی نه در خوشه‌بندی TE_1 و نه در خوشه‌بندی TE_2 و نه در خوشه‌بندی TE_3 این دو داده هم خوشه نبوده‌اند. پس برای این دو داده در ماتریس همبستگی صفر داریم (یعنی درایه (۲و۱) و (۱و۲) ماتریس همبستگی صفر است؛ به عبارتی $Co(1,2) = Co(2,1) = 0$). به تبع داده ۱۲ و ۱۳ در دو خوشه‌بندی TE_1 و خوشه‌بندی TE_3 این دو داده هم خوشه بوده‌اند و در خوشه‌بندی TE_2 این دو داده هم خوشه نبوده‌اند. پس برای این دو داده در ماتریس همبستگی $\frac{2}{3}$ داریم (یعنی درایه (۱۳و۱۲) و (۱۲و۱۳) ماتریس همبستگی $\frac{2}{3}$ است؛ به عبارتی $Co(12,13) = Co(13,12) = 0$ ؛ چراکه در دو مورد از سه مورد این دو داده هم خوشه بوده‌اند. ماتریس نهایی همبستگی در جدول (۸) آورده شده‌است. پس از انجام کلیه مراحل خواهیم داشت:

$$FitnessFunction([1,1,0,1]) = 14 \quad (25)$$

به‌طوری مشابه می‌توان به راحتی محاسبه کرد و دید که برانزنگی کروموزوم نشان داده شده در شکل (۵) صفر است.

۱	۰	۱	۱
---	---	---	---

(شکل-۵): نمایش یک راه‌حل نامزد (کروموزوم)

(Figure-5): Displays a candidate solution (chromosome)

پس $FitnessFunction([1,0,1,1]) = 0$ و از طرفی

(جدول-۹): ماتریس نهایی همبستگی بهینه برای افزایش جدول (۷)

(Table-9): The final optimal correlation matrix for table partitions

1	0.5	0.5	0	0.5	0.5	0.5	0.5	0.5	0	0	0	1
0	0	0	0.5	0	0	0	0	0	1	1	1	0
0	0	0	0.5	0	0	0	0	0	1	1	1	0
0	0	0	0.5	0	0	0	0	0	1	1	1	0
0.5	1	1	0.5	1	0	0	0	1	0	0	0	0.5
0.5	0	0	0	0	1	1	1	0	0	0	0	0.5
0.5	0	0	0	0	1	1	1	0	0	0	0	0.5
0.5	0	0	0	0	1	1	1	0	0	0	0	0.5
0.5	1	1	0.5	1	0	0	0	1	0	0	0	0.5
0	0.5	0.5	1	0.5	0	0	0	0.5	0.5	0.5	0.5	0
0.5	1	1	0.5	1	0	0	0	1	0	0	0	0.5
0.5	1	1	0.5	1	0	0	0	1	0	0	0	0.5
1	0.5	0.5	0	0.5	0.5	0.5	0.5	0.5	0	0	0	1

شماره خطاب کنیم تا بتوانیم شباهت دو مجموعه را اندازه‌گیری کنیم. برای مثال ۲ مجموعه شکل (۷) با ۹ نمونه را در نظر بگیرید که هر کدام دارای ۳ خوشه و هر خوشه دارای ۳ نمونه است. این دو مجموعه به‌طور کامل با هم مشابه خوشه‌بندی شده‌اند در حالی که برچسب‌های خوشه‌های نظیر به‌طور کامل متفاوت است.

۴- معیارهای اندازه‌گیری کیفیت

یکی از اشکالات اندازه‌گیری شباهت در بین دو مجموعه مشکل برچسب خوشه‌ها است. برای تطبیق برچسب خوشه‌ها در ۲ مجموعه باید جایگشت‌های متفاوتی را امتحان کنیم تا خوشه‌های مشابه را به یک نام و یک

$$MI(A, B) \quad (27)$$

$$= \frac{-2 \sum_{i=1}^{c_A} \sum_{j=1}^{c_B} \frac{N_{ij}}{N} \log \left(\frac{N_{ij} N}{N_i N_j} \right)}{\sum_{j=1}^{c_A} N_i \log \left(\frac{N_i}{N} \right) + \sum_{j=1}^{c_B} N_j \log \left(\frac{N_j}{N} \right)}$$

اگر ۲ مجموعه A و B به طور کامل مشابه باشند آنگاه NMI مقدار بیشینه یک و اگر دو مجموعه به طور کامل متفاوت از یکدیگر باشند مقدار صفر را برمی گرداند. برای مثال مجموعه A و B ارائه شده در جدول (۱۰) را در نظر بگیرید. در مجموعه زیر NMI به صورت زیر محاسبه می شود:

(جدول-۱۰): یک مجموعه داده نوعی

Object	۱	۲	۳	۴	۵	۶	۷	۸
Partition A	۱	۱	۲	۲	۳	۳	۴	۴
Partition B	۲	۱	۳	۳	۲	۱	۲	۲

ماتریس تداخل ۲ مجموعه بالا به صورت شکل (۸) است. و مقدار NMI برابر است با:

$$NMI(A, B) = \frac{-2 \times 5.5452}{-11.0904 - 8.3178} \quad (28)$$

$$\cong 0.5714$$

مشاهده می شود که مقدار NMI به دست آمده برابر با همان برچسب گذاری مناسب بر روی دو مجموعه بالا است.

A↓	B→			Total
	1	2	3	
1	1	1	0	2
2	0	0	2	2
3	1	1	0	2
4	0	2	0	2
Total	2	4	2	

(شکل-۸): مثال ارائه شده ماتریس تداخل

(Figure-8): An example of a Confuse matrix

رند ۲ یک روش ساده دیگری برای اندازه گیری تشابه بین دو مجموعه را به صورت زیر بیان می کند. ابتدا تعاریف زیر را ببینید:

n_{11} : تعداد جفت نمونه هایی که هم در مجموعه A و هم در مجموعه B در یک خوشه قرار دارند.

n_{00} : تعداد جفت نمونه هایی که هم در مجموعه A و هم در مجموعه B در دو خوشه متفاوت قرار دارند و در هیچ کدام از این دو مجموعه با هم در یک خوشه قرار ندارند.

Object	۱	۲	۳	۴	۵	۶	۷	۸	۹
Partition 1	۱	۱	۱	۲	۲	۲	۳	۳	۳
Partition 2	۳	۳	۳	۱	۱	۱	۲	۲	۲

(شکل-۷): یک مثال از مشکل تطابق برچسب ها

(Figure-7): An example of a label matching problem

برای حل مشکل برچسب خوشه ها می توانیم برچسب مجموعه نخست را ثابت در نظر بگیریم و برچسب خوشه های مجموعه دوم را آن قدر تغییر دهیم تا بیشینه تشابه با مجموعه نخست به دست آید و در آن شرایط میزان مطابقت (یا صحت) آن دو افراز را به عنوان مقدار تشابه دو مجموعه معرفی کنیم. یک راه برای جلوگیری از تطابق برچسب ها به کارگیری روش MI (Mutual Information) است. ماتریس تداخل برای مجموعه های A و B را به این صورت در نظر بگیرید که سطرها معرف خوشه های مجموعه A است و ستون ها معرف خوشه های مجموعه B است. تعاریف زیر را در نظر بگیرید:

N_{ij} : مقدار درایه (i, j) ماتریس است که معرف تعداد نمونه هایی است که در خوشه i در مجموعه A و در خوشه j در مجموعه B قرار دارد.

N_i : مجموع مقادیر ردیف i یا همان مجموع نمونه ها در خوشه i در مجموعه A .

N_j : مجموع مقادیر ستون j یا همان مجموع نمونه ها در خوشه j در مجموعه B .

C_A : تعداد خوشه های مجموعه A .

C_B : تعداد خوشه های مجموعه B .

شایان ذکر است که هیچ لزومی ندارد که تعداد خوشه های ۲ مجموعه برابر باشد.

با توجه به تعریف بالا رابطه MI به صورت زیر تعریف می شود:

$$MI(A, B) = \sum_{i=1}^{c_A} \sum_{j=1}^{c_B} \frac{N_{ij}}{N} \log \left(\frac{N_{ij} N}{N_i N_j} \right) \quad (26)$$

از آنجایی که هیچ کران بالایی بر مقدار MI نیست، رابطه اطلاعات متقابل نرمال شده را (که نرمال شده MI است) تعریف می کنیم. رابطه نرمال شده بالا که NMI نام دارد به صورت زیر است:

² Rand

¹ Normal Mutual Information

معیاری دیگر در ارزیابی یک خوشه‌بندی، معیار فیشتر است که در بخش روش پیشنهادی معرفی شد. رابطه این معیار در زیر آورده شده‌است.

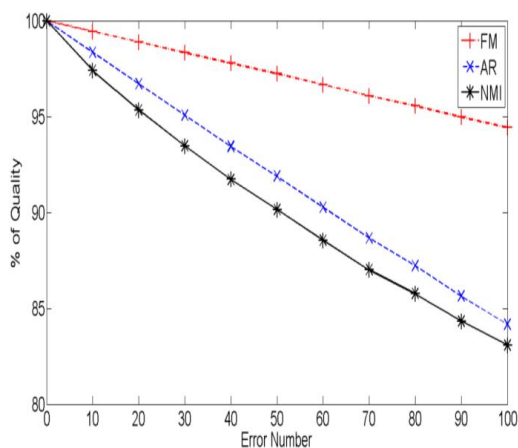
$$FM(P, L) \quad (34)$$

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

$$FM(P, L) = 0.5147 \quad (35)$$

که K_P تعداد خوشه‌های افزاز P است؛ N_i^P نشان‌دهنده تعداد داده‌های موجود در خوشه i -ام از افزاز P است؛ N_j^L نشان‌دهنده تعداد داده‌های موجود در خوشه j -ام از افزاز L است؛ N_{ij}^{PL} نشان‌دهنده تعداد داده‌هایی که مشترکاً در خوشه i -ام از افزاز P و در خوشه j -ام از افزاز L قرار دارد؛ N تعداد کل داده‌ها را نشان می‌دهد؛ τ یک جای‌گشت از اعداد یک تا N است.

اگر دو افزاز P و برچسب L به‌طور کامل مشابه باشند آنگاه FM مقدار ماکزیمم یعنی یک و اگر دو افزاز به‌طور کامل متفاوت از یکدیگر باشند مقدار صفر را برمی‌گرداند. برای مثال بالا مقدار FM برگشتی در زیر محاسبه شده‌است.



(شکل-۹): نتیجه تطبیق بین نتایج یک الگوریتم خوشه‌بندی فرضی و برچسب‌های واقعی نمونه‌ها در یک مجموعه داده که

دارای ۱۵۰۰ نمونه و ۳ خوشه

(Figure-9): The result of matching the results of a hypothetical clustering algorithm and the actual tags of the samples in a data set containing 1500 samples and 3 clusters.

• مقایسه سه روش FM، AR و NMI

سه روش FM، AR و NMI به‌طور تقریبی نتایج بهتری از روش‌های دیگر می‌دهند و نیازی به تطبیق برچسب خوشه‌های مجموعه‌های متفاوت ندارند. در شکل زیر سعی در ارائه مقایسه‌ای بین سه روش داریم تا روش مناسب‌تر را برای مطالعات آتی بیشتر مد نظر قرار دهیم. برای ارائه

n_{10} : تعداد جفت نمونه‌هایی که در مجموعه A در یک خوشه قرار دارند، ولی در مجموعه B در ۲ خوشه متفاوت قرار دارند.

n_{01} : تعداد جفت نمونه‌هایی که در مجموعه B در یک خوشه قرار دارند، ولی در مجموعه A در ۲ خوشه متفاوت قرار دارند. به‌طور کامل مشخص است که n_{00} و n_{11} مواردی را نشان می‌دهد که دو مجموعه نتیجه یکسانی در مورد یک جفت می‌دهند و n_{01} و n_{10} مواردی را بیان می‌کند که دو مجموعه نتایج متفاوتی در مورد یک جفت خاص می‌دهند. به‌طور کلی $\frac{N(N-1)}{2}$ جفت نمونه متفاوت در یک مجموعه N عضو وجود دارد؛ لذا:

$$n_{00} + n_{01} + n_{10} + n_{11} = \frac{N(N-1)}{2} \quad (29)$$

و رابطه Rand به‌صورت زیر تعریف می‌شود:

$$r(A, B) = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} = \frac{2 \times (n_{00} + n_{11})}{N(N-1)} \quad (30)$$

رابطه Rand به‌ازای دو مجموعه به‌طور کامل متشابه مقدار ۱ را بر می‌گرداند ولی اگر ۲ مجموعه متفاوت باشند هیچ مقدار منطقی نمی‌دهد.

رابطه $1RA$ عیب رابطه رند در زمانی که ۲ مجموعه نامتشابه باشند را رفع می‌کند و یک مقدار منطقی قابل دفاع بر می‌گرداند. فرض کنید که دو مجموعه و با تعدادی خوشه با تعداد نمونه برابر داریم. اگر دو مجموعه به‌طور کامل نامتشابه باشند الگوریتم AR همانند رابطه NMI مقداری نزدیک ۰ را بر می‌گرداند [1, 11] همانند تعاریف مربوط به قسمت NMI رابطه محاسبه AR به‌صورت زیر بیان می‌شود:

$$AR(A, B) = \frac{\sum_{i=1}^{c_A} \sum_{j=1}^{c_B} \binom{N_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \quad (31)$$

که

$$t_1 = \sum_{i=1}^{c_A} \binom{N_i}{2}; \quad t_2 = \sum_{j=1}^{c_B} \binom{N_j}{2}; \quad t_3 = \frac{2t_1 t_2}{N(N-1)} \quad (32)$$

برای نمونه در مثال قبل مقدار AR برابر می‌شود با:

$$AR(A, B) = \frac{2 - \frac{8}{7}}{\frac{1}{2}(4 + 8) - \frac{8}{7}} = \frac{3}{17} \quad (33)$$

¹ Adjusted Rand

دارد. مثلاً اگر $Error = 3$ باشد، سه مقدار تصادفی p, r و q را بین ۱ تا ۱۵۰۰ انتخاب، سپس A_p, A_q و A_r را با مقادیری دیگر جایگزین می‌کنیم؛ یعنی مقدار آنها یکی اضافه می‌کنیم. برای مثال A_q را یکی زیاد می‌کنیم؛ اگر یک باشد، آن را به دو تغییر می‌دهیم، اگر دو باشد، آن را به سه تغییر می‌دهیم و اگر سه باشد، آن را به یک تغییر می‌دهیم. به‌طور خلاصه می‌توان گفت:

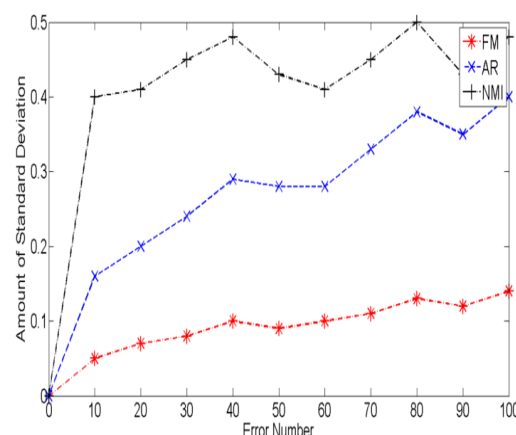
$$A_q = (A_q + 1) \bmod 3 \quad (36)$$

$$A_p = (A_p + 1) \bmod 3 \quad (37)$$

$$A_r = (A_r + 1) \bmod 3 \quad (38)$$

حال که در A سه مقدار (در موقعیت‌های تصادفی p, r و q) با T متفاوت است، FM, AR یا NMI به‌ترتیب محاسبه می‌کنیم. اگر $r = 782, p = 1467$ و $q = 1186$ باشد، FM, AR و NMI به‌ترتیب $۹۹/۵۰, ۹۹/۸۳$ و $۹۹/۱۴$ درصد خواهند بود. اگر $r = 772, p = 905$ و $q = 1262$ باشد، FM, AR و NMI به‌ترتیب $۹۹/۸۷, ۹۹/۶۰$ و $۹۹/۴۳$ درصد خواهند بود. به همین ترتیب صد مرتبه این روال با صد مجموعه مقادیر برای موقعیت‌های تصادفی p, r و q کار را تکرار کرده و صد مجموعه مقادیر برای FM, AR و NMI محاسبه می‌کنیم. در نهایت میانگین این صد مقدار را که به‌ترتیب با μ_{AR}^3, μ_{FM}^3 و μ_{NMI}^3 نشان می‌دهیم، به‌عنوان مقادیر FM, AR و NMI با سه خطا ($Error = 3$) در نظر می‌گیریم. همچنین انحراف‌معیار همان صد مقدار را که به‌ترتیب با $\sigma_{AR}^3, \sigma_{FM}^3$ و σ_{NMI}^3 نشان می‌دهیم، به‌عنوان خطای مقادیر FM, AR و NMI وقتی که سه خطا ($Error = 3$) داریم در نظر می‌گیریم. اکنون با تغییر مقدار خطا از صفر تا صد، مقادیر $\mu_{AR}^{Error}, \mu_{FM}^{Error}, \mu_{NMI}^{Error}, \sigma_{AR}^{Error}, \sigma_{FM}^{Error}, \sigma_{NMI}^{Error}$ را محاسبه می‌کنیم و در شکل‌های (۹ و ۱۰) نمایش می‌دهیم.

نتایج این شکل ابتدا سعی در بیان پاره‌ای مطالب می‌نماییم. ابتدا یک مجموعه داده مصنوعی دست‌ساز را که دارای ۱۵۰۰ نمونه و ۳ خوشه است در نظر بگیرید. در این مجموعه داده مصنوعی توزیع داده‌ها در سه خوشه یکسان است؛ یعنی در هر خوشه ۵۰۰ داده وجود دارد. فرض کنید برچسب‌های این مجموعه داده را با T نشان دهیم. همچنین فرض کنید T_i برچسب داده i -ام است. پس T_i برابر ۱ است اگر $1 \leq i \leq 500$ باشد؛ ۲ است اگر $501 \leq i \leq 1000$ باشد؛ و ۳ است اگر $1001 \leq i \leq 1500$ باشد.



(شکل-۱۰): میزان ناپایداری در نتیجه تطبیق بین نتایج یک

الگوریتم خوشه‌بندی فرضی

(Figure-10): The degree of instability as a result of matching the results of a hypothetical clustering algorithm

حال فرض کنید که یک الگوریتم خوشه‌بندی X بر روی این مجموعه داده مصنوعی اجرا کرده‌ایم. خروجی فرضی این الگوریتم بر فرض با A نشان داده شود. همچنین فرض کنید A_i برچسب داده i -ام است. همچنین فرض کنید A_i برابر ۲ است اگر $1 \leq i \leq 500$ باشد؛ ۳ است اگر $501 \leq i \leq 1000$ باشد؛ و ۱ است اگر $1001 \leq i \leq 1500$ باشد. حال معیار فیشر (FM)، اندیس رند (AR) یا اطلاعات متقابل نرمال شده (NMI) بین A و T را محاسبه کنیم، پی خواهیم برد که دقت این الگوریتم ۱۰۰٪ است. اکنون فرض کنید به اندازه $Error$ (که این مقدار عددی دلخواه است)، این الگوریتم خطا

(جدول-۱۱): خلاصه‌ای از مشخصه‌های مجموعه داده‌های استاندارد مورد استفاده

(Table-11) Summary of the characteristics of the standard data set used

مجموعه داده	تعداد کلاس	تعداد ویژگی	تعداد نمونه
wine	3	13	178
breast-cancer	2	9	683
bupa	2	6	345
galaxy	7	4	323
glass	6	9	214

halfring	2	2	400
iris	3	4	150
ionosphere	2	34	351
saheart	2	9	462
yeast	10	8	1484

(جدول-۱۲): میانگین دقت بر روی تمام مجموعه‌داده‌ها در مقابل روش‌های گوناگون

(Table-12): Average accuracy of all data sets versus different methods

توابع توافقی	الگوریتم‌ها	دقت
Single Linkage	مجمع کامل	68.35
	GA	68.86
	SA	68.42
	PSO	68.88
Fuzzy k-means	مجمع کامل	68.35
	GA	68.86
	SA	68.42
	PSO	68.88
HGPA	مجمع کامل	52.20
	GA	53.12
	SA	53.32
	PSO	53.38
MCLA	مجمع کامل	63.20
	GA	63.21
	SA	63.20
	PSO	63.25
CSPA	مجمع کامل	65.60
	GA	65.60
	SA	64.98
	PSO	65.54

(جدول-۱۳): میانگین اطلاعات متقابل نرمال شده بر روی تمام مجموعه‌داده‌ها در مقابل روش‌های گوناگون

(Table-13): Average normalized mutual information on all data sets versus different methods

توابع توافقی	الگوریتم‌ها	اطلاعات متقابل نرمال شده
Single Linkage	مجمع کامل	43.15
	GA	43.18
	SA	42.98
	PSO	43.15
Fuzzy k-means	مجمع کامل	41.35
	GA	44.52
	SA	42.89
	PSO	44.67
HGPA	مجمع کامل	35.10
	GA	35.17
	SA	33.87
	PSO	35.38
MCLA	مجمع کامل	38.56
	GA	39.85
	SA	39.87
	PSO	39.82
CSPA	مجمع کامل	39.60
	GA	41.25
	SA	38.99
	PSO	41.31

(Table-14): Fisher's average criterion for all data sets versus different methods

توابع توافقی	الگوریتم ها	معیار فیشر
Single Linkage	مجمع کامل	70.10
	GA	71.09
	SA	70.85
	PSO	71.15
Fuzzy k-means	مجمع کامل	67.78
	GA	69.36
	SA	69.38
	PSO	69.41
HGPA	مجمع کامل	58.15
	GA	60.25
	SA	56.50
	PSO	59.69
MCLA	مجمع کامل	68.75
	GA	68.72
	SA	68.50
	PSO	68.70
CSPA	مجمع کامل	69.87
	GA	70.25
	SA	70.08
	PSO	70.46

سهولت در تحلیل، میانگین نتایج بر روی ۱۰ مجموعه داده در جداول (۱۲)، (۱۳) و (۱۴) آورده شده است. مجموعه داده های استفاده شده مجموعه داده های استاندارد UCI است که به طور تقریبی نتایج تمام مطالعات اخیر دنیا در زمینه خوشه بندی با به کارگیری این مجموعه داده ها گزارش می شوند. روش پیشنهادی بر روی ۱۰ مجموعه داده استاندارد نرمال شده مورد آزمایش قرار گرفته است. برای انجام آزمایش ها سعی شده است تا مجموعه داده ها از لحاظ تعداد رده ها، تعداد ویژگی ها و همچنین تعداد نمونه ها از بیشینه پراکندگی برخوردار باشند تا نتایج آزمایش ها تا حد ممکن دارای استحکام و قابل تعمیم باشد. برای عملیات نرمال سازی، هر کدام از ویژگی های این مجموعه داده ها با میانگین صفر و واریانس یک، $N(0,1)$ ، نرمال شده اند. برای همه این مجموعه داده ها، تعداد خوشه ها و برچسب واقعی نمونه ها از قبل معلوم هستند؛ بنابراین، درصد نمونه هایی که درست تشخیص داده شده اند، به عنوان معیار کارایی روش خوشه بندی مورد استفاده قرار گرفته است. درواقع بعد از حل مسأله تناظر بین برچسب های به دست آمده و خوشه های واقعی، می توان نرخ خطا را تعیین کرد. در تمامی روش های مورد به کارگیری الگوریتم K-means به عنوان الگوریتم پایه استفاده شده است. تعداد نتایج اولیه تولید شده نیز در تمام روش ها ثابت و برابر با ۱۲۰ است. درواقع این تعداد با دستکاری پارامتر k از الگوریتم K-means به دست آمده است. به این صورت که چهار گروه

شکل (۸) بیان گر نتیجه تطبیق بین نتایج یک الگوریتم خوشه بندی فرضی X و برچسب های واقعی نمونه ها در یک مجموعه داده که دارای ۱۵۰۰ نمونه و ۳ خوشه است. با بررسی شکل بالا سعی در شناسایی روشی داریم که نتیجه دقیق تر و نزدیک تری به دقت واقعی خوشه بندی (به دست آمده بر اساس میزان خطا)، دارد. در این بخش نتایج به کارگیری روش پیشنهادی روی مجموعه داده های مختلف و پارامترهای مورد استفاده گزارش شده است. روش پیشنهادی در محیط MATLAB7.1 پیاده سازی و مورد آزمایش قرار گرفته است. نتایج آزمایش ها روی میانگین ۱۰ بار اجرای مستقل برنامه گزارش شده است. عملکرد روش های مختلف خوشه بندی با سه معیار، دقت، NMI و $F - Measure$ محاسبه شده است. جداول (۱۲)، (۱۳) و (۱۴) بیانگر این موضوع است. چنانچه مشاهده می شود، نتایج نه تنها کاهش نداشته، بلکه در اغلب موارد بهبود نیز یافته است. انتخاب خوشه های اولیه با دو الگوریتم ژنتیک و ذوب فلزات به بهبود عملکرد روش پیشنهادی در انتخاب بهینه ترین خوشه های اولیه برای ترکیب در خوشه بندی نهایی کمک شایانی می کند. درعمل اعمال الگوریتم تکاملی برای انتخاب خوشه ها، دو مجموعه خوشه های پایدار و خوشه های غیر پایدار به دست می آید. نتایج مجمع کامل و مجمعی که خوشه های را با به کارگیری الگوریتم تکاملی انتخاب می کند بر حسب NMI و $F - Measure$ و دقت بر روی مجموعه داده های گوناگون محاسبه و برای

۲.۵ خواهند بود. در ادامه این روش به تفصیل آمده است. فرضیه صفر روش فریدمن بیان می‌دارد که روش‌ها تفاوت با معنی ندارند. برای رد این فرضیه نشان‌دادن اینکه روش‌ها تفاوت با معنی دارند، باید به‌شکل زیر اقدام کنیم: ابتدا فرض کنید r_i^j نشان دهنده رتبه روش i -ام در مجموعه داده j -ام باشد. میانگین رتبه روش j -ام از رابطه (۳۹) محاسبه می‌شود.

$$R_j = \frac{1}{N} \sum_{i=1}^N r_i^j \quad (39)$$

روش‌ها است. چون شش روش داریم، درجه آزادی مسأله ۵ است. حال با رابطه ذیل مقدار χ_F^2 را از رابطه (۴۰) محاسبه می‌کنیم.

$$\chi_F^2 = \frac{12N}{k(k+1)} \left(\sum_{j=1}^k R_j^2 - \frac{k(k+1)^2}{4} \right) \quad (40)$$

با محاسبه رابطه (۱۳)، مقدار χ_F^2 برابر ۱۰/۱۶ خواهد شد که این مقدار از مقدار مورد انتظار ما در جدول توزیع چای با درجه آزادی ۵ که برابر ۹/۴۸۹ است بیشتر است. پس فرض صفر رد شده و به این نتیجه می‌رسیم که اختلاف‌های بین روش‌ها با معنی است. از آن‌جا که رتبه متوسط هر یک از روش‌های مجمع کامل، GA، SA، PSO، علیزاده و عظیمی به ترتیب ۲.۱، ۲.۴، ۲.۰، ۳.۵ و ۳.۸ است. پس روش PSO به‌شکل با معنایی از سایر روش‌ها بهتر است.

مقایسه با الگوریتم‌های پایه مقاوم: شکل (۱۱) نتایج مقایسه الگوریتم خوشه‌بندی ترکیبی پیشنهادی را با سه الگوریتم خوشه‌بندی پایه مقاوم بر اساس مجموعه داده‌های مورد نظر نشان می‌دهد. این شکل، میانگین ($mean$) و انحراف معیار ($standard deviation$) اعتبار خوشه‌بندی هر الگوریتم برای این مجموعه داده‌ها نیز بیان شده است. مشاهده می‌کنیم که اعتبار خوشه‌بندی به‌دست آمده با الگوریتم خوشه‌بندی ترکیبی پیشنهادی برتر یا نزدیک به بهترین نتایج حاصل از سه الگوریتم دیگر است. این آزمایشات به ما می‌گویند که الگوریتم پیشنهادی می‌تواند نتایج مقاوم به‌دست آمده توسط الگوریتم‌های خوشه‌بندی مقاوم را تولید کند.

سی‌تایی از نتایج اولیه، با در نظر گرفتن تعداد خوشه‌های مورد استفاده توسط این الگوریتم با اندازه‌های k ، $k+1$ ، $k+2$ و $k+3$ حاصل شده است. همچنین، برای ایجاد پراکندگی بیشتر در نتایج اولیه از نمونه‌برداری بدون جاگذاری با نرخ ۵۰٪ استفاده شده است. همچنین، برای ساختن افراز نهایی از روش اتصال منفرد^۱ بر روی ماتریس همبستگی، روش Fuzzy K-means و روش‌های بر پایه گراف HGPA، MCLA و CSPA استفاده شده است. جدول (۱۱) خلاصه‌ای از مجموعه داده استاندارد مورد استفاده در آزمایش‌ها را نشان می‌دهد.

چنان‌چه در میانگین مشاهده می‌شود، در اغلب موارد بهبود کارایی حاصل شده است؛ لذا نتیجه می‌گیریم که نه تنها کاهش خوشه‌های انتخاب‌شده باعث کاهش کارایی نشده است بلکه افزایش کارایی را نیز در بیشتر موارد موجب شده است.

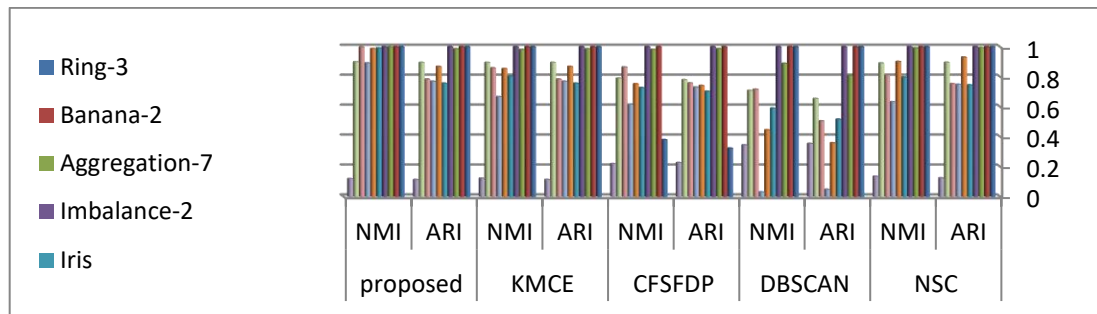
همچنین از آن جهت که این کار در ادامه کاری است که در قبل توسط علیزاده و مینایی انجام شده است، مقایسه‌ای بین این کار با کار آنها و همچنین کار پیشتر از آنکه توسط آقای عظیمی انجام شده است، صورت پذیرفته است.

اگر چه روش پیشنهادی از نظر دقت در جدول (۱۵)، بهتر از سایر روش‌ها بوده اما هنوز نمی‌توان ادعایی دال بر اینکه بهترین روش، روش پیشنهادی است، نمی‌توان داشت. باید دید که آیا این نتایج تصادفی نبوده باشد که با تغییر دوباره پارامترها و مقداردهی اولیه متفاوت الگوریتم‌ها، نتایج به‌گونه‌ای دیگر رقم نخواهد خورد. برای بررسی دقیق‌تر و پی بردن به این نکته که آیا این برتری با معنی است، باید به یکی از روش‌های راستی‌سنجی آماری پناه برد. در اینجا از روش راستی‌سنجی آماری فریدمن استفاده می‌کنیم. این روش را به این دلیل بر می‌گزینیم که مناسب مقایسه چندین روش به‌طور هم‌زمان است. این روش هر یک از روش‌ها را بر اساس کارایی آنها در یک مجموعه داده مرتب می‌کند و رتبه روشی با بیشترین کارایی را ۱ در نظر می‌گیرد و رتبه روشی با کمترین کارایی را M (M تعداد روش‌ها است) در نظر می‌گیرد. در مواردی که چند روش با رتبه‌های یکسان باشد، میانگین رتبه برای آنها در نظر گرفته می‌شود. برای مثال اگر روش A و B دارای دومین و سومین کارایی در بین روش‌ها باشند، یعنی رتبه‌های آنها ۲ و ۳ باشد، ولی کارایی آنها برابر باشد، رتبه‌های آنها به ترتیب برابر ۲/۵ و

¹ Single Linkage

(با ۳۷۸۰ نقطه داده، ۵۴ ویژگی و رده)، MNIST (با ۵۰۰۰ نقطه داده، ۷۸۴ ویژگی و ۱۰ رده)، تشخیص رقمی نوری (ODR) (با ۵۶۲۰ نقطه داده، ۶۴ ویژگی و ۱۰ رده)، ماهواره لندست (LS) (با ۶۴۳۵ نقطه داده، ۳۶ ویژگی و ۶ رده)، SOLET (با ۷۷۹۷ نقطه داده، ۶۱۷ ویژگی، و ۲۶ رده)، USPS، خوشه‌بندی‌های پایه به‌وسیله الگوریتم‌های خوشه‌بندی k-means و fuzzy c-means تولید می‌شوند.

ارزیابی روش پیشنهادی در مقایسه با روش‌های رقیب بر روی مجموعه‌داده‌های پیچیده: در آزمایش‌های ما از ۱۰ مجموعه‌داده معیار واقعی استفاده شده‌است [55]: Semeion (با ۱۵۹۳ نقطه داده، ۲۵۶ ویژگی و ۱۰ رده)، چندین ویژگی (MF) (با ۲۰۰۰ نقطه داده، ۶۴۹ ویژگی و ۱۰ رده)، تقسیم بندی تصویر (IS) (با ۲۳۱۰ نقطه داده، ۱۹ ویژگی و ۷ رده)، Forest – Cover – Type (FCT)



(شکل-۱۱): مقایسه با الگوریتم‌های خوشه‌بندی "قوی"
(Figure-11): Comparison with Strong Clustering Algorithms

(جدول-۱۵): مقایسه عملکرد الگوریتم خوشه‌بندی پیشنهادی با CLWGC در صورت استفاده خوشه‌بندی k-means و fuzzy c-means به‌عنوان الگوریتم خوشه‌بندی پایه و سایر روش‌های رقیب

(Table-15): Comparison of the performance of the proposed clustering algorithm with CLWGC in case of using k-means clustering and fuzzy c-means as the basic clustering algorithm and other competing methods.

Dataset	CLWGC with k-means	CLWGC with fuzzy c-means	π_{GND}	RCESCC	Chen	Berikov	PROPOSED
Semeion	66.81	66.43	68.59	67.43	69.22	56.30	70.40
MF	68.89	69.13	69.28	70.15	71.32	58.04	73.10
IS	67.05	67.26	62.71	67.22	69.11	54.05	68.59
FCT	23.31	23.38	26.19	30.23	33.35	22.21	34.89
MNIST	65.26	64.16	66.16	67.02	69.44	55.43	68.23
ODR	83.12	82.57	85.50	84.04	86.21	64.23	85.76
LS	63.28	61.42	65.60	63.15	64.45	55.54	63.89
ISOLET	76.38	75.51	77.71	77.19	76.92	72.18	78.90
USPS	65.68	65.56	67.12	67.20	68.28	62.17	69.50
LR	41.47	41.69	47.07	43.42	48.11	38.17	47.85
Average	62.12	61.71	63.59	63.70	65.64	53.83	65.90

ترکیبی اولیه انجام می‌دهد و سپس بر اساس پراکندگی بین نتایج الگوریتم‌های خوشه‌بندی اولیه و خوشه‌بندی ترکیبی اولیه اقدام به کاوش در انتخاب‌های ممکن در هر مجموعه‌داده می‌کند. در ادامه با بهترین انتخاب زیرمجموعه‌ای از نتایج اولیه که بر مبنای الگوریتم مکاشفه‌ای یافته شده، می‌کند؛ سپس خوشه‌بندی نهایی را بر روی زیرمجموعه انتخاب‌شده برای به‌دست‌آوردن خوشه‌های نهایی انجام می‌دهد. روش ارائه‌شده به‌خاطر اینکه گلچینی از خوشه‌بندی‌های اولیه انجام‌شده را برپایه هر مجموعه‌داده وارد خوشه‌بندی ترکیبی نهایی می‌کند،

جدول (۸) نتایج میانگین عملکرد روش‌های مختلف را در بیش از ۳۰ اجرای مختلف از نظر NMI مقایسه کرده و نشان داده روش پیشنهادی عملکرد بهتری نسبت به سایر رقبا دارد.

۵- نتیجه‌گیری

در این مقاله ما روشی در خوشه‌بندی ترکیبی ارائه کردیم که کیفیت انتخاب خود را بر اساس نوع هر مجموعه‌داده تغییر می‌دهد. روش ارائه شده ابتدا یک خوشه‌بندی

- multiple partitions". Journal of Machine Learning Research, 3(Dec):583-617, 2002.
- [6] Fred, A. and Jain, A.K. "Data Clustering Using Evidence Accumulation", Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR02, Quebec City, pp. 276 - 280, 2002.
- [7] Topchy, A., Jain, A.K. and Punch, W.F., "Combining Multiple Weak Clusterings", Proc. 3d IEEE Intl. Conf. on Data Mining, pp. 331-338, 2003.
- [8] Fred A. and Lourenco A. (2008), "Cluster Ensemble Methods: from Single Clusterings to Combined Solutions", Studies in Computational Intelligence (SCI), 126, 3-30.
- [9] Ayad H.G. and Kamel M.S., Cumulative Voting Consensus Method for Partitions with a Variable Number of Clusters, IEEE Trans. on Pattern Analysis and Machine Intelligence, VOL. 30, NO. 1, 160-173, 2008.
- [10] Minaei-Bidgoli B., Topchy A. and Punch W.F., "Ensembles of Partitions via Data Resampling", in Proc. Intl. Conf. on Information Technology, ITCC 04, Las Vegas, 2004.
- [11] Parvin H., Minaei-Bidgoli B. "A clustering ensemble framework based on selection of fuzzy weighted clusters in a locally adaptive clustering algorithm". Pattern Anal. Appl. 18(1): 87-112, 2015.
- [12] Alizadeh H., Minaei-Bidgoli B., Parvin H. Optimizing Fuzzy Cluster Ensemble in String Representation. IJPRAI 27(2), 2013.
- [13] Parvin H., Minaei-Bidgoli B., Alinejad-Rokny H., Punch W.F. "Data weighing mechanisms for clustering ensembles". Computers & Electrical Engineering 39(5): 1433-1450, 2013.
- [14] Barthelemy J.P. and Leclerc B., The median procedure for partition, In Partitioning Data Sets, AMS DIMACS Series in Discrete Mathematics, Cox, I. J. et al eds., 19, pp. 3-34, 1995.
- [15] Fern X.Z., and Lin W., "Cluster Ensemble Selection". Statistical Analysis and Data Mining 1(3): 128-141, 2008.
- [16] Parvin H., Mirnababoli M., Alinejad-Rokny H. "Proposing a classifier ensemble framework based on classifier selection and decision tree". Eng. Appl. of AI 37: 34-42, 2015.
- [17] Dudoit S. and Fridlyand, J., Bagging to improve the accuracy of a clustering procedure, Bioinformatics, 19 (9), pp. 1090-1099, 2003.
- [18] Fischer B. and Buhmann J.M., "Bagging for path-based clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1411-1415, 2003.

روشی پویا است. نتایج تجربی به‌دست‌آمده نشان‌گر کارایی و توانایی روش ارائه‌شده در خوشه‌بندی اطلاعات است.

۶- کارهای آینده

روش ارائه‌شده زمینه مطالعاتی وسیعی به روی آینده الگوریتم‌های خوشه‌بندی باز می‌کند. واردکردن مفهوم فازی در خوشه‌بندی ترکیبی یکی از نخستین ایده‌هایی است که می‌تواند در مطالعات آینده مورد بررسی قرار گیرد. نرمال‌سازی داده‌ها یکی از اقدامات ضروری در زمان به‌کارگیری فاصله اقلیدسی است. از آنجا که هیچ تضمینی برای بهبود کیفیت خوشه‌بندی در هنگام به‌کارگیری الگوریتم‌های نرمال‌سازی داده‌ها وجود ندارد، به‌طورمعمول روش‌های خوشه‌بندی ارائه‌شده گزارش‌های خود را بر روی داده‌های خام و غیر نرمال ارائه می‌دهند؛ لذا یکی دیگر از ایده‌هایی که می‌تواند در مطالعات آینده مورد مطالعه قرار گیرد، پیداکردن یک روش پویا برای اختصاص‌دادن یک وش نرمال‌سازی به هر مجموعه‌داده است؛ اما مهم‌ترین عاملی که به‌نظر می‌رسد می‌تواند روش ارائه‌شده را به‌نحو مطلوبی بهبود بخشد، ارائه روش هوشمندی است که تولید نتایج اولیه را نیز هدایت کند. به این معنی که سعی کند تا نتایج اولیه‌ای تولید کند که ضعف‌های پوشش داده‌نشده نتایج اولیه دیگر را پوشش دهد.

7- References

۷- مراجع

- [1] Azimi J., The investigation of the Ensemble Clustering Diversity. MSc Thesis. Iran University of Science and Technology, 2006.
- [۱] عظیمی ج، " بررسی پراکندگی در خوشه‌بندی ترکیبی"، پایان‌نامه کارشناسی‌ارشد، دانشگاه علم و صنعت ایران، خرداد ۱۳۸۶.
- [2] Alizadeh A., Minaei-Bidgoli B., Parvin H. Cluster ensemble selection based on a new cluster stability measure. Intell. Data Anal. 18(3): 389-408, 2014.
- [3] Jain A., Murty M. N., and Flynn P. (1999), Data clustering: A review. ACM Computing Surveys, 31(3):264-323.
- [4] Faceli K., Marcilio C.P. Souto d., Multi-objective Clustering Ensemble, Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS'06), 2006.
- [5] Strehl A. and Ghosh J., "Cluster ensembles - a knowledge reuse framework for combining

- Press, Cambridge, Massachusetts. London, England, Fifth printing, 1999.
- [33] Aarts E. H. L. and Korst J. Simulated Annealing and Boltzmann Machines, John Wiley & Sons, Essex, U.K, 1989.
 - [34] Kennedy J and Eberhart R.C., "Particle Swarm Optimization", Proceedings of IEEE International Conference on Neural Networks", Piscataway, NJ, pp. 1942-1948, 1995.
 - [35] Fred A. and Jain A.K., "Learning Pairwise Similarity for Data Clustering", In Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06), 2006.
 - [36] Fridlyand J. and Dudoit S. "Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method". Stat. Berkeley Tech Report. No. 600, 2001.
 - [37] X. Fern, C. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning", Proc. of the 21st International Conference on Machine Learning, 2004.
 - [38] D. Huang, J. Lai, C. D. Wang, "Ensemble clustering using factor graph", Pattern Recognition, vol. 50, pp. 131-142, 2016.
 - [39] M. Selim, E. Ertunc, "Combining multiple clusterings using similarity graph", Pattern Recognition, vol. 44, no. 3, 694-703, 2011.
 - [40] C. Boulis, M. Ostendorf, "Combining multiple clustering systems", Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases, 2004.
 - [41] A. Topchy, B. Minaei-Bidgoli, A. Jain, "Adaptive clustering ensembles", Proc. the 17th International Conference on Pattern Recognition, 2004.
 - [42] P. Hore, L. O. Hall, B. Goldgo, "A scalable framework for cluster ensembles", Pattern Recognition, vol. 42, no. 5, 676-688, 2009.
 - [43] B. Long, Z. Zhang, P. S. Yu, "Combining multiple clusterings by soft correspondence", Proc. the 4th IEEE International Conference on Data Mining, 2005.
 - [44] D. Cristofor, D. Simovici, "Finding median partitions using information theoretical based genetic algorithms", J. Universal Computer Science, vol. 8, no. 2, pp. 153-172, 2002.
 - [45] A. Topchy, A. Jain, W. Punch, "Clustering ensembles: Models of consensus and weak partitions", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, 1866-1881, 2005.
 - [46] H. Wang, H. Shan, A. Banerjee, "Bayesian cluster ensembles", Statistical Analysis and Data Mining, vol. 4, no. 1, pp. 54-70, 2011.
 - [19] Fred A. and Jain A.K., "Robust data clustering", in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR ,USA, vol. II, pp. 128-136, 2003.
 - [20] Fred A.L. and Jain A.K. "Combining Multiple Clusterings Using Evidence Accumulation". IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(6):835-850, 2005.
 - [21] Kuncheva L.I. and Hadjitodorov S. "Using diversity in cluster ensembles". In Proc. of IEEE Intl. Conference on Systems, Man and Cybernetics, pages 1214-1219, 2004.
 - [22] Kuncheva L.I. and Whitaker C. J., "Measures of diversity in classifier ensembles", Machine Learning, 2003.
 - [23] Baumgartner R., Somorjai R., Summers R., Richter W., Ryner L., and Jarmasz M., Resampling as a Cluster Validation Technique in fMRI, JOURNAL OF MAGNETIC RESONANCE IMAGING 11: pp. 228-231, 2000.
 - [24] Breckenridge J., Replicating cluster analysis: Method, consistency and validity, Multivariate Behavioral research, 1989.
 - [25] Shamir O., Tishby N., "Cluster Stability for Finite Samples", 21st Annual Conference on Neural Information Processing Systems (NIPS07), 2007.
 - [26] Roth V., Braun M.L., Lange T., and Buhmann J.M., "Stability-Based Model Order Selection in Clustering with Applications to Gene Expression Data", ICANN 2002, LNCS 2415, pp. 607-612, 2002a.
 - [27] Roth V., Lange T., Braun M., and Buhmann J., A "Resampling Approach to Cluster Validation", Intl. Conf. on Computational Statistics, COMPSTAT, 2002b.
 - [28] Saha A., Das S. "Categorical fuzzy k-modes clustering with automated feature weight learning". Neurocomputing 166: 422-435, 2015.
 - [29] Law M.H.C., Topchy A.P., and Jain A.K. "Multiobjective data clustering". In Proc. of IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 424-430, Washington D.C, 2004.
 - [30] Akbari E., Dahlan H.M., Ibrahim R., Alizadeh H.: Hierarchical cluster ensemble selection. Eng. Appl. of AI 39: 146-156 2015.
 - [31] Iam-On, N. and T. Boongoen, "Diversity-driven generation of link-based cluster ensemble and application to data classification", Expert Systems with Applications, 42(21): p. 8259-8273, 2015.
 - [32] Melanie M., "An Introduction to Genetic Algorithms", A Bradford Book The MIT



صدراله عباسی دانش‌آموخته کارشناسی‌ارشد رشته نرم‌افزار از دانشگاه علوم و تحقیقات است. وی هم‌اکنون دانشجوی دکتری دانشگاه آزاد اسلامی بوده و در چندین واحد دانشگاهی در رشته کامپیوتر مشغول به تدریس است. نشانی رایانامه ایشان عبارت است از:

s.abbasi680@gmail.com



صمد نجاتیان مدرک کارشناسی خود را در سال ۱۳۸۲ در رشته مهندسی برق گرایش الکترونیک از دانشگاه سیستان و بلوچستان و مدرک کارشناسی‌ارشد خود را در سال ۱۳۸۶ در رشته مهندسی برق گرایش مخابرات از دانشگاه مشهد و در سال ۱۳۹۳ مدرک دکتری خود را در رشته مهندسی برق گرایش مخابرات از دانشگاه صنعتی مالزی دریافت کرد. وی هم‌اکنون دانشیار و عضو هیئت علمی گروه برق دانشگاه آزاد اسلامی واحد یاسوج است. حوزه‌های تخصصی ایشان برق-مخابرات، الگوریتم‌های بهینه‌سازی، طبقه‌بندی و خوشه‌بندی داده‌ها و هوش مصنوعی است. وی تاکنون بیش از ۱۱۰ مقاله علمی در نشریات و کنفرانس‌های معتبر داخلی و خارجی به چاپ رسانیده‌است.

نشانی رایانامه ایشان عبارت است از:

samad.nej.2007@gmail.com



حمید پروین مدرک کارشناسی خود را در سال ۱۳۸۵ در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه شهید چمران اهواز و مدرک کارشناسی‌ارشد و دکتری خود را در سال ۱۳۸۷ و ۱۳۹۲ در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه علم و صنعت ایران دریافت کرد. وی تاکنون بیش از ۱۵۰ مقاله علمی در نشریات و کنفرانس‌های معتبر داخلی و خارجی به چاپ رسانیده‌است و چندین کتاب چاپ کرده‌اند. نشانی رایانامه ایشان عبارت است از:

parvin@iust.ac.ir

- [47] Z. He, X. Xu, S. Deng, "A cluster ensemble method for clustering categorical data", Information Fusion, vol. 6, no. 2, pp. 143C151, 2005.
- [48] N. Nguyen, R. Caruana, "Consensus Clusterings", Proc. IEEE Intl Conf. Data Mining, pp. 607-612, 2007.
- [49] Z. Huang, "Extensions to the kmeans algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery, vol. 2, no. 3, pp. 283-304, 1998.
- [50] S. Abbasi, S. Nejatian, H. Parvin, V. Rezaie & K. Bagherifard, "Clustering ensemble selection considering quality and diversity, " Artificial Intelligence Review, vol. 52, PP. 1311-1340, Springer Nature B.V. 2018, <https://doi.org/10.1007/s10462-018-9642-2>.
- [51] A. Bagherinia, B. Minaei-Bidgoli, M. Hossinzadeh, H. Parvin, "Elite fuzzy clustering ensemble based on clustering diversity and quality measures, " Springer Science+Business Media, LLC, part of Springer Nature, Applied Intelligence, 49, PP. 1724-1747, 2019. <https://doi.org/10.1007/s10489-018-1332-x>.
- [52] A. Nazari, A. Dehghan, S Nejatian, V. Rezaie, H. Parvin, "A comprehensive study of clustering ensemble weighting based on cluster quality and diversity, " Pattern Analysis and Applications, vol. 22, pp.133-145, 2019.
- [53] M. Mojarad, S. Nejatian, H. Parvin, M. Mohammadpoor, "A fuzzy clustering ensemble based on cluster clustering and iterative Fusion of base clusters, " The International Journal of Research on Systems for Real Life Complex Problems, Applied Intelligence vol. 49, pp. 2567-2581, 2019.
- [54] Z. Chen, A. Bagherinia B. Minaei-Bidgoli, H. Parvin, Pho KH. Fuzzy Clustering Ensemble Considering Cluster Dependability. International Journal on Artificial Intelligence Tools. 2021 Mar 26;30(02):2150007
- [55] V. Berikov, "A probabilistic model of fuzzy clustering ensemble." Pattern Recognition and Image Analysis 28, no. 1 (2018): 1-10.
- [56] moradi M, nejatian S, parvin H, bagherifard K, rezaei V. Clustering and Memory-based Parent-Child Swarm Meta-heuristic Algorithm for Dynamic Optimization. JSDP 2021; 18 (3) :127-146
- Omidvar M, Nejatian S, Parvin H, Bagherifard K, Rezaie V. Providing an algorithm for solving general optimization problems based on Domino theory. JSDP 2022; 19 (2) :87-106



کرم الله باقری فرد مدرک کارشناسی

خود را در سال ۱۳۸۴ در رشته مهندسی کامپیوتر گرایش نرم افزار از دانشگاه اصفهان و مدرک کارشناسی ارشد و دکترای خود را به ترتیب در سال های

۱۳۸۷ و ۱۳۹۵ از دانشگاه نجف آباد و اراک در رشته مهندسی کامپیوتر گرایش نرم افزار دریافت کرد. وی از سال ۱۳۸۵ تاکنون عضو هیئت علمی بخش مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد یاسوج است. حوزه های تخصصی ایشان داده کاوی، یادگیری ماشین و سامانه های پیشنهاددهنده است. وی تاکنون بیش از ۸۰ مقاله علمی در نشریات و کنفرانس های معتبر داخلی و خارجی به چاپ رسانیده است.

نشانی رایانامه ایشان عبارت است از:

k.bagheri@iauyasooj.ac.ir



سیده وحیده رضایی دارای مدرک

دکترای ریاضی است. ایشان هم اکنون عضو هیئت علمی دانشگاه آزاد اسلامی واحد یاسوج است. وی تاکنون بیش از ۷۰ مقاله علمی در نشریات و

کنفرانس های معتبر داخلی و خارجی به چاپ رسانیده است.

نشانی رایانامه ایشان عبارت است از:

vahidehrezaie80@gmail.com