یک روش خوشه‌بندی ترکیبی جدید مبتنی بر خوشه‌بند cmeans فازی با حفظ تنوع در اجماع

فاطمه نجفی[۱]، حمید پروین*[۲]، کمال میرزائی[۳]، صمد نجاتیان[۴] و سیده وحیده رضایی[۵]

[۱و۳]دانشکده فنی و مهندسی، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران

[۲]دانشکده فنی و مهندسی، واحد ممسنی، دانشگاه آزاد اسلامی، نورآباد ممسنی، فارس، ایران

[۲]باشگاه پژوهشگران جوان و نخبگان، دانشگاه آزاد اسلامی، نورآباد ممسنی، فارس، ایران

[۴و۵]باشگاه پژوهشگران جوان و نخبگان، دانشگاه آزاد اسلامی، یاسوج، کهگیلویه و بویراحمد، ایران

[۴]دانشکده مهندسی برق، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

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

چکیده

به‌علت بدون‌ناظربودن مسأله خوشه‌بندی، انتخاب یک الگوریتم خاص جهت خوشه‌بندی یک مجموعه ناشناس امری پرخطر و به‌طورمعمول شکست‌خورده است. به‌خاطر پیچیدگی مسأله و ضعف روش‌های خوشه‌بندی پایه، امروزه بیش‌تر مطالعات به سمت روش‌های خوشه‌بندی ترکیبی هدایت شده است. در خوشه‌بندی ترکیبی ابتدا چندین خوشه‌بندی پایه تولید و سپس برای تجمیع آن‌ها، از یک تابع توافقی جهت ایجاد یک خوشه‌بندی نهایی استفاده می‌شود که بیشینه شباهت را به خوشه‌بندی‌های پایه داشته باشد. خوشه‌بندی توافقی تولیدشده باید با استفاده از بیشترین اجماع و توافق به‌دست آمده باشد. ورودی تابع یادشده همه خوشه‌بندی‌های پایه و خروجی آن یک خوشه‌بندی به‌نام خوشه‌بندی توافقی است. درحقیقت روش‌های خوشه‌بندی ترکیبی با این شعار که ترکیب چندین مدل ضعیف بهتر از یک مدل قوی است، به میدان آمده‌اند. با این‌وجود، این ادعا درصورتی درست است که برخی شرایط همانند تنوع بین اعضای موجود در اجماع و کیفیت آن‌ها رعایت شده باشند. این مقاله یک روش خوشه‌بندی ترکیبی را ارائه داده که از روش خوشه‌بندی پایه ضعیف cmeans فازی به‌عنوان خوشه‌بند پایه استفاده کرده است. همچنین با اتخاذ برخی تمهیدات، تنوع اجماع را بالا برده است. روش خوشه‌بندی ترکیبی پیشنهادی مزیت الگوریتم خوشه‌بندی cmeans فازی را که سرعت آن است، دارد و همچنین ضعف‌های عمده آن را که عدم قابلیت کشف خوشه‌های غیرکروی و غیریکنواخت است، ندارد. در بخش مطالعات تجربی الگوریتم خوشه‌بندی ترکیبی پیشنهادی با سایر الگوریتم‌های خوشه‌بندی مختلف به‌روز و قوی بر روی مجموعه داده‌های مختلف آزموده و با یکدیگر مقایسه شده است. نتایج تجربی حاکی از برتری کارایی روش پیشنهادی نسبت به سایر الگوریتم‌های خوشه‌بندی به‌روز و قوی است.

واژگان کلیدی: یادگیری ترکیبی، خوشه‌بندی ترکیبی، الگوریتم خوشه‌بندی cmeans فازی، اعتبار داده‌ها.

A new ensemble clustering method based on fuzzy cmeans clustering while maintaining diversity in ensemble

Fatemeh Najafi[1], Hamid Parvin*[2], Kamal Mirzaie[3], Samad Nejatian[4] & Vahideh Rezaie[5]

[1,3]Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran

[2]Department of Computer Engineering, Mamasani Branch, Islamic Azad University, Fars, Iran

[4,5]Department of Electrical Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran

Abstract

An ensemble clustering has been considered as one of the research approaches in data mining, pattern recognition, machine learning and artificial intelligence over the last decade. In clustering, the combination first produces several bases clustering, and then, for their aggregation, a function is used to create a final cluster that is as similar as possible to all the cluster bundles. The input of this function is all base clusters

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۷/۱۱/۳۰ • تاریخ پذیرش:۱۳۹۹/۱۰/۲۴ • تاریخ انتشار: ۱۳۹۹/۰۹/۱۵ • نوع مطالعه: پژوهشی

[DOI: 10.29252/jsdp.17.4.103]

and its output is a clustering called clustering agreement. This function is called an agreement function. Ensemble clustering has been proposed to increase efficiency, strong, reliability and clustering stability. Because of the lack of cluster monitoring, and the inadequacy of general-purpose base clustering algorithms on the other, a new approach called an ensemble clustering has been proposed in which it has been attempted to find an agreed cluster with the highest Consensus and agreement. In fact, ensemble clustering techniques with this slogan, the combination of several poorer models, is better than a strong model. However, this claim is correct if certain conditions (such as the diversity between the members in the consensus and their quality) are met. This article presents an ensemble clustering method. This paper uses the weak clustering method of fuzzy cmeans as a base cluster. Also, by adopting some measures, the diversity of consensus has increased. The proposed hybrid clustering method has the benefits of the clustering algorithm of fuzzy cmeans that has its speed, as well as the major weaknesses of the inability to detect non-spherical and non-uniform clusters. In the experimental results, we have tested the proposed ensemble clustering algorithm with different, up-to-date and robust clustering algorithms on the different data sets. Experimental results indicate the superiority of the proposed ensemble clustering method compared to other clustering algorithms to up-to-date and strong.

Keywords: Ensemble Learning, Ensemble Clustering, Fuzzy Cmeans Clustering Algorithm, Data Validity.

۱- مقدمه

یکی از مهم‌ترین مسایل در حوزه‌ی داده‌کاوی و شناسایی الگو، خوشه‌بندی است. خوشه‌بندی نوعی یادگیری بدون ناظر است که به معنی سازمان‌دهی الگوها در چند دسته است؛ به‌طوری‌که اعضای هر دسته، از جنبه‌هایی به هم شبیه باشند. در خوشه‌بندی سعی می‌شود الگوها در چند خوشه چنان تقسیم شوند که اعضای هر خوشه به‌هم شبیه باشند و با اعضای دیگر خوشه‌ها بیشینه تفاوت را داشته باشند. تقسیم‌بندی‌های مختلفی برای روش‌های خوشه‌بندی وجود دارد. این تقسیم‌بندی عبارتند از: سلسله‌مراتبی[1] در برابر افرازبندی[2]، انحصاری[3] در برابر غیرانحصاری، فازی در برابر غیرفازی، جزیی در برابر کامل و ... .

درواقع خوشه‌بندی داده‌ها یک ابزار ضروری برای یافتن گروه‌ها در داده‌های بدون برچسب است. دست کم پنج دلیل اصلی برای اهمیت خوشه‌بندی وجود دارد:

۱) جمع‌آوری و برچسب‌گذاری یک مجموعه بزرگ از الگوهای نمونه می‌تواند بسیار با ارزش باشد.

۲) ممکن است ما به دنبال‌کردن در جهت معکوس علاقمند باشیم؛ یعنی آموزش با مقدار زیاد داده‌های بدون برچسب و سپس تنها استفاده از ناظر برای برچسب‌گذاری خوشه‌های پیدا شده است. این می‌تواند برای کاربردهای داده‌کاوی بزرگ که محتویات یک پایگاه داده از قبل شناخته‌شده نیست، مناسب باشد.

۳) در خیلی از کاربردها، مشخصه‌های الگوها مثل رده‌بندی[4] خودکار مواد غذایی با تغییر فصل می‌توانند به‌آهستگی با زمان تغییر کنند. اگر این تغییرات بتواند با یک رده‌بندی‌کننده[5] به‌صورت بدون‌ناظر[6] ره‌گیری[7] شود، عملکرد بهتری می‌تواند به دست آید.

۴) می‌توانیم از روش‌های بدون‌ناظر (خوشه‌بندی) برای پیداکردن و استخراج ویژگی‌ها استفاده کنیم.

۵) با خوشه‌بندی می‌توانیم یک دید و بینشی از طبیعت و ساختار داده به‌دست آوریم که این می‌تواند برای ما با ارزش باشد. کشف زیررده‌های[8] مجزا یا شباهت‌های بین الگوها ممکن است به‌طور چشم‌گیری در روش طراحی رده‌بندی‌کننده به ما پیشنهاد ارایه کند.

این تنوع در الگوریتم‌های خوشه‌بند خود یک چالش محسوب می‌شود؛ چون هر کدام نقاط ضعف و قوت گوناگونی دارند؛ پس هیچکدام برای همه مجموعه‌داده‌ها مناسب نیست. چالش این است که برای یک مجموعه‌دادهٔ در دست، چطور بهترین و مناسب‌ترین روش خوشه‌بندی را انتخاب کرد. برای مثال الگوریتم خوشه‌بند kmeans که یکی از رویکردهای مسطح است، به‌عنوان یک الگوریتم بسیار سریع و با کارایی به‌نسبه مناسب شناخته می‌شود [4]. همچنین این الگوریتم به‌عنوان یک روشی که (۱) در بهینه محلی گیر می‌افتد، (۲) به انتخاب مراکز اولیه خوشه‌ها حساس است، (۳) هچنین بر این فرض که ساختار خوشه‌ها کروی و یک‌نواخت است، استوار است و (۴) توزیع داده‌ها بر کارایی آن مؤثر است، شناخته می‌شود. این الگوریتم به‌عنوان یک الگوریتم خوشه‌بند ضعیف، یکی از الگوریتم‌های خوشه‌بند پایه مناسب برای مشارکت در ساخت اجماع محسوب می‌شود. در نقطهٔ مقابل الگوریتم‌های خوشه‌بند محلی

[1] Hierarchical
[2] Partitioning
[3] Exclusive
[4] Classification
[5] Classifier
[6] Unsupervised
[7] Tracking
[8] SubClass

[DOI: 10.29252/jsdp.17.4.103]

(local) شبیه به kmeans، الگوریتم‌های خوشه‌بند سراسری قرار می‌گیرد که برخلاف داشتن کارایی خوب، از ضعف‌های واضحی چون داشتن پیچیدگی زمانی بالا رنج می‌برند (چراکه ممکن است برای مثال نیازمند محاسبه فاصله بین همه زوج‌اشیای داده‌ای باشند). برخی از این دسته الگوریتم‌ها شامل kmeans سراسری (gkmeans) [5]، the density-based spatial clustering of applications with noise (DBSCAN) [6]، the clustering by fast search and find of density peaks (CFSFDP) [7] و الگوریتم خوشه‌بندی طیفی [8-9] است. الگوریتم gkmeans از محاسبه تعدادی از فواصل بین داده‌ها برای حصول به یک خوشه‌بندی بهتر استفاده می‌کند. DBSCAN می‌تواند هر ساختاری از خوشه‌ها را در محیط‌های حتی نوفه‌ای با استفاده از تعداد زیادی از محاسبات فاصله‌ای و چگالی تشخیص دهد. خروجی الگوریتم‌های خوشه‌بند سراسری اگرچه قوی و باکیفیت هست؛ اما در مقایسه با الگوریتم‌های خوشه‌بند محلی، هزینه‌های محاسباتی بسیار چشم‌گیری دارد؛ بنابراین، درهمین‌واخر به جای پرداختن به ساخت یک الگوریتم خوشه‌بند سراسری قوی، توجه بیشتر به ساخت چارچوب‌هایی شده است که چندین خوشه‌بند ضعیف را یک‌پارچه کنند. در این راستا "خوشه‌بند ترکیبی" یا "تجمع خوشه‌بندها" [11-10]، برای بهبود استحکام و کیفیت فرایند خوشه‌بندی ارایه شده است.

در یادگیری مبتنی بر اجماع به‌عنوان یکی از مباحث داغ پژوهشی در بحث داده‌کاوی، شناسایی الگو، یادگیری ماشین و هوش مصنوعی، چندین یادگیرنده ساده (اغلب ضعیف) را برای حل یک مسأله واحد آموزش می‌دهیم. در این روش یادگیری، به جای یادگیری مستقیم داده‌ها به‌وسیلهٔ یک یادگیر قوی (که به‌طورمعمول کند هستند)، سعی در یادگیری یک مجموعه از یادگیرهای ضعیف (که به‌طورمعمول سریع هستند) و تلفیق نتایج آنها با یک روش تابع توافقی (شبیه رأی‌گیری) دارند [12]. در یادگیری باناظر به‌علت وجود برچسب‌های اشیای‌داده‌ای، ارزیابی هر یک از یادگیرهای ساده سرراست است؛ امّا در یادگیری بدون ناظر چنین نیست و بسیار سخت بتوان راه حلی بدون استفاده از اطلاعات جانبی برای ارزیابی اینکه یک الگوریتم خوشه‌بند روی یک مجموعه‌داده چه ضعف‌ها و نقاط قوتی دارد، یافت [13]. هم‌اکنون، روش‌های متعددی برای خوشه‌بندی ترکیبی به‌منظور بهبود استحکام و کیفیت نتایج خوشه‌بندی، مطرح شده‌اند. هر خوشه‌بندی موجود در خوشه‌بندی ترکیبی، به‌عنوان یک یادگیر پایه در نظرگرفته می‌شود.

خوشه‌بندی ترکیبی هنوز هم به‌عنوان یک ابزار و هم به‌عنوان یک زمینه پژوهشی تئوری مورد مطالعه است. یک مقاله مروری در [14]، برای انواع این روش‌ها ارایه شده است. به‌علت آن‌که دقت در خوشه‌بندی معنای سرراستی همچون رده‌بندی ندارد؛ بنابراین، مفهوم جایگزینی برای آن ارایه شده است که بیان می‌دارد یک خوشه‌بندی دقیق، خوشه‌بندی‌ای است که به خوشه‌بندی‌های دیگرِ شکل‌گرفته بر روی داده‌های مورد نظر، بیشترین شباهت را داشته باشد؛ به‌عبارتی خوشه‌بندی بهتر یعنی خوشه‌بندی پایدارتر. به‌دلیلی مشابه با علت مناسب‌بودن یک مجمع متنوع از رده‌بندها برای رده‌بندی ترکیبی، یک مجمعی از خوشه‌بندی‌ها را مناسب می‌گوییم اگر خوشه‌بندی‌های پایه آن، متنوع باشند. برای متنوع قلمدادشدن یک اجماعی از خوشه‌بندی‌ها، باید یک الگوریتم خوشه‌بند ضعیف بر روی داده‌ها چندین‌بار اعمال شود. برای حل این مسأله، الگوریتم خوشه‌بند cmeans فازی بهبودیافته را به‌عنوان یک خوشه‌بند ضعیف به‌کار می‌بریم. چهار زیرمسأله در خوشه‌بندی ترکیبی در زیر آورده شده است:

۱. مسأله تشخیص برچسب‌های *به‌نسبه صحیح* در خوشه‌بندی: بر خلاف رده‌بندی، هیچ اطلاعات واقعی از برچسب‌ها در خوشه‌بندی وجود ندارد.
۲. مسأله حصول خوشه‌بندی‌های متنوعی که توصیف‌گر کل داده‌ها باشد: در یادگیری ترکیبی، درحالی‌که چندین یادگیرندهٔ ضعیف به‌عنوان یک یادگیرنده قوی ترکیب می‌شوند، هر چه یادگیرنده‌های پایه بیشتر مکمل هم‌دیگر باشند، یادگیرنده ترکیبی بهتر عمل می‌کند. یعنی هر خوشه‌بندی ضعف بقیه خوشه‌بندی‌ها را بپوشاند؛ بنابراین، برای این منظور ما بایستی چندین خوشه‌بندی مکمل با اعمال الگوریتم خوشه‌بند cmeans فازی بهبودیافته تولید کنیم.
۳. مسأله تشخیص تناسب بین خوشه‌ها: بر خلاف رده‌بندی که در آن هر برچسبی فقط به یک رده دلالت دارد، برچسب‌ها در خوشه‌بندی هیچ معنای واحدی ندارد و تنها هم‌خوشه‌بودن داده‌ها را نشان نمی‌دهد و خوشه‌های هم‌نام در دو خوشه‌بندی گوناگون به هیچ حقیقتی دلالت نمی‌کند؛ بنابراین، پیش از هر کاری در خوشه‌بندی ترکیبی، برچسب خوشه‌بندی‌های مختلف باید بر اساس تناظر بازبرچسب‌گذاری شود؛ علاوه‌براین، حتی دو خوشه از یک خوشه‌بندی یکسان نیز احتمال دارد که به یک خوشه واقعی دلالت کند.

[DOI: 10.29252/jsdp.17.4.103]

۴. مسأله ترکیب نتایج خوشه‌بندی‌های پایهٔ همسان‌شده: در خوشه‌بندی‌های گوناگون، هر شیء ممکن است برچسب‌های گوناگونی داشته باشد؛ بنابراین ما باید یک برچسب نهایی را به نام برچسب توافقی تعیین کنیم. در یادگیری ترکیبی، درحالی‌که چندین یادگیرنده ضعیف به‌عنوان یک یادگیرنده قوی ترکیب می‌شوند، هر چه عمل ترکیب مؤثرتر باشد، یادگیرندهٔ ترکیبی بهتر عمل می‌کند.

این مقاله سعی در مرتفع‌سازی همه زیرمسائل یادشده به‌کمک تعریف خوشه‌های محلاً معتبر خواهد داشت. درحقیقت این مقاله داده‌های اطراف یک مرکز خوشه در الگوریتم خوشه‌بند cmeans فازی بهبودیافته را خوشه‌های داده‌ای محلاً معتبر می‌نامد. برای تولید خوشه‌بندی‌های متنوع، از یک استراتژی تکراری تولید خوشه‌بندی‌های ضعیف (به‌کمک الگوریتم خوشه‌بند cmeans فازی بهبودیافته) بر روی داده‌های ظاهرنشده در خوشه‌های محلاً معتبر قبلی، استفاده می‌شود؛ سپس به‌کمک یک معیار شباهت بین خوشه‌ای، شباهت بین خوشه‌های محلاً معتبر تولیدی را اندازه‌گیری می‌کنیم. با تشکیل یک گراف وزن‌دار که رأس‌های آن خوشه‌های محلاً معتبر است و وزن یال‌های آن میزان شباهت بین خوشه‌ها است، گام بعدی الگوریتم پیشنهادی این مقاله انجام می‌شود. در گام بعد الگوریتم، کمینه برش گراف برای افراز این گراف به تعدادی (که از پیش تعیین شده است) خوشه نهایی اعمال می‌شود. در گام نهایی به‌کمک خروجی این خوشه‌ها و میزان اعتبار متوسط و بیشینه‌سازی توافق، خوشه‌های نهایی توافقی تولید می‌شوند. گفتنی است که هر خوشه‌بند پایه دیگر نیز می‌تواند به‌عنوان خوشه‌بند پایه استفاده شود؛ برای مثال می‌توان الگوریتم خوشه‌بند پایه kmeans را به‌کار برد. همچنین از دیگر روش‌های تابع توافقی مرسوم نیز می‌توان به‌عنوان تابع توافقی جهت ترکیب نتایج خوشه‌بندهای پایه بهره برد.

در ادامه مقاله، بخش دوم به کارهای مرتبط می‌پردازد. در بخش سوم روش پیشنهادی ارایه‌شده و در بخش چهارم نتایج تجربی ارایه شده و در بخش نهایی، نتیجه‌گیری و کارهای آینده بحث شده است.

۲- کارهای مرتبط

دو مقوله بسیار مهم در خوشه‌بندی ترکیبی وجود دارد: (۱) این‌که چه‌طور یک اجماعی از خوشه‌بندی‌های پایه معتبر و متنوع تولید کنیم و (۲) این‌که چه‌طور از یک اجماع در دست، بهترین نتایج توافقی را تولید کنیم. البته اگرچه کارایی هر یک از این دو بر کارایی دیگری تأثیر دارد، امّا این دو به‌عنوان دو مسأله به‌طورکامل مستقل شناخته می‌شوند. به‌همین علت، تاحدودی در مقاله‌های پژوهشی، تنها به یکی از این دو مقوله پرداخته شده و کمتر دیده شده است که هر دو توأم در نظر گرفته شوند.

مسأله نخست که مسأله تولید اجماع نیز نام دارد، سعی در تولید یک مجموعه‌ای از خوشه‌بندی‌های پایه معتبر و البته متنوع دارد. این مسأله به‌کمک روش‌های گوناگونی انجام شده است؛ برای مثال می‌تواند توسط اعمال یک الگوریتم خوشه‌بند پایه ناپایدار بر روی مجموعه‌دادهٔ داده‌شده با تغییر در پارامترهای الگوریتم تولید [17-15] و همچنین می‌تواند توسط اعمال الگوریتم‌های خوشه‌بند پایه گوناگون بر روی مجموعه‌دادهٔ داده‌شده تولید شود [11]، [19-18]. راه دیگر می‌تواند با اعمال یک الگوریتم خوشه‌بند پایه بر روی نگاشت‌های گوناگون از مجموعه‌دادهٔ داده‌شده خوشه‌بندی‌های پایه معتبر و متنوع را تولید کند [28-20]. در راه بعدی خوشه‌بندی‌های پایه معتبر و متنوع می‌تواند به‌کمک اعمال یک الگوریتم خوشه‌بند پایه بر روی زیرمجموعه‌های گوناگون که می‌تواند با جای‌گذاری یا بدون جای‌گذاری تولید شده باشد، از مجموعه‌دادهٔ داده‌شده تولید شود [19-18].

برای حل مسأله دوم نیز راه‌کارهای فراوانی مطرح شده‌اند. نخستین راه‌کار، رویکردهای مبتنی بر ماتریس هم‌رخدادی است که در این رویکردها ابتدا تعداد هم‌خوشه‌شدن‌های هر زوج‌داده در یک اجماع را در یک ماتریس به نام ماتریس هم‌رخدادی ذخیره می‌کنیم؛ سپس با درنظرگرفتن این ماتریس به‌عنوان ماتریس شباهت و اعمال یک روش خوشه‌بندی سلسله‌مراتبی، خوشه‌های نهایی توافقی به‌دست می‌آیند. این رویکرد به‌طورتقریبی به‌عنوان مرسوم‌ترین روش قدیمی شناخته می‌شود [32-29]. رویکردی دیگر، رویکرد مبتنی بر برش (ابر)گراف است. در این رویکرد، ابتدا مسأله یافتن خوشه‌بندی توافقی به یک مسأله افراز گراف تبدیل می‌شود؛ سپس به‌کمک الگوریتم‌های افراز یا برش گراف، خوشه‌های نهایی توافقی به‌دست می‌آیند [10]، [35-33]. چهار الگوریتم ترکیبی ابرگراف معروف CSPA، HGPA، MCLA، و HBGF هستند. رویکردی دیگر، رویکرد رأی‌گیری است [22-21]، [36]، [39-38]. برای این منظور باید ابتدا عمل

[DOI: 10.29252/jsdp.17.4.103]

بازبرچسب‌گذاری انجام شود. عمل باز برچسب‌گذاری به‌منظور هم‌سان‌سازی برچسب‌های خوشه‌بندی‌های گوناگون برای تطابق است. از رویکردهای مهم دیگر می‌توان به موارد زیر اشاره کرد [45-40]:

۱- رویکرد درنظرگرفتن خوشه‌بندی‌های اولیه به‌عنوان یک فضای واسط (یا مجموعه‌دادهٔ جدید) و خوشه‌بندی این فضای جدید به‌کمک یک الگوریتم خوشه‌بند پایه شبیه الگوریتم بیشینه‌سازی انتظار [41]،

۲- رویکرد استفاده از الگوریتم‌های تکاملی برای یافتن سازگارترین خوشه‌بندی به‌عنوان خوشه‌بندی توافقی [40]، رویکرد استفاده از الگوریتم خوشه‌بندی kmods برای یافتن خوشه‌بندی توافقی [45-44].

بسیار محتمل است که یک پارتیشن وجود داشته باشد که با استفاده از یک اندازه‌گیری پایداری به‌عنوان یک پارتیشن بد قضاوت شود؛ در حالی‌که دارای یک (یا بیشتر) خوشه با کیفیت بالا است [46]؛ بنابراین، پژوهش‌گران با الهام از ارزیابی پارتیشن‌ها، اقدامات لازم برای ارزیابی خوشه‌ها را تعیین می‌کنند. بسیاری از سنجه‌های پایداری مانند اطلاعات متقابل نرمال برای اعتبارسنجی یک پارتیشن پیشنهاد شده است. سنجه‌های تعریف‌شده مبتنی بر اطلاعات متقابل نرمال است. اشکال رویکرد متداول در این مقاله مورد بحث قرار گرفت و ملاکی برای ارزیابی ارتباط بین یک خوشه و یک پارتیشن ارائه شد که به آن معیار ENMI گویند. معیار ENMI اشکال اندازه‌گیری معمول اطلاعات متقابل نرمال معمولی را جبران می‌کند؛ همچنین، یک روش خوشه‌بندی ترکیبی که مبتنی بر جمع‌کردن زیرمجموعه‌ای از خوشه‌های اولیه است، ارائه شد [46].

برخلاف برخی از تلاش‌ها برای بهبود کیفیت روش‌های خوشه‌بندی، به نظر می‌رسد که پژوهش‌های اندکی به رویهٔ انتخاب در خوشه‌بندی ترکیبی فازی اختصاص یافته است؛ علاوه‌براین، کیفیت و تنوع محلی دو عامل مهم در انتخاب خوشه‌بندی‌های پایه است. تعداد کمی از مطالعات، این دو عامل را برای انتخاب بهترین خوشه‌بندی‌های پایه فازی در اجماع در نظر گرفته‌اند. در [47] یک چارچوب خوشه‌بندی ترکیبی فازی جدید براساس یک معیار اندازه‌گیری تنوع فازی جدید پیشنهاد شده تا خوشه‌های پایه را با بهترین عملکرد پیدا کنند. تنوع و کیفیت براساس اطلاعات متقابل عادی فازی بین خوشه‌بندی‌های پایه فازی تعریف شده است.

در [48]، یک چارچوب جدید خوشه‌بندی ترکیبی براساس وزن‌گیری در سطح خوشه ارائه شده است. مقدار اطمینان این اجماع در مورد یک خوشه، به‌عنوان قابلیت اطمینان آن خوشه در نظر گرفته شده است. مقدار قطعیتی که یک اجماع خاص در مورد یک خوشه دارد بر اساس میانگین میزان قابلیت اطمینان آن خوشه توسط اجماع محاسبه و سپس با انتخاب بهترین خوشه‌ها و تعیین وزنی به هر خوشه انتخابی براساس قابلیت اطمینان آن، مجموعه خوشه‌های نهایی ایجاد می‌شود؛ پس از آن، مقاله به‌جای ماتریس توافقی سنتی، ماتریس توافقی وزنی در سطح خوشه را پیشنهاد می‌کند؛ سپس دو رویکرد اجماع برای تولید افراز توافقی معرفی و مورد استفاده قرار گرفته است.

۳- روش پیشنهادی

در این بخش ابتدا تعاریف و علائم لازم ارائه می‌شود؛ سپس مسأله خوشه‌بندی ترکیبی را تعریف می‌کنیم. در گام بعد الگوریتم پیشنهادی ارائه خواهد شد. بالاخره در گام نهایی تحلیل الگوریتم آورده خواهد شد.

۳-۱- تعاریف و علائم

تمامی علائم مورد استفاده در این مقاله در جدول (۱) ارایه شده‌اند.

مجموعه‌داده: یک مجموعه‌داده، یک مجموعه‌ای از اشیای داده‌ای است که هر شیء داده خود یک بردار عددی (یا بردار ویژگی) است. مجموعه‌داده با X و هر شیء داده با x_i نشان داده می‌شود؛ بدیهی است که $x_i \in X$. ویژگی j-ام از شیء داده x_i را با x_{ij} نشان می‌دهیم. اندازه هر مجموعه X را با $|X|$ نشان می‌دهیم. تعداد ویژگی‌های مجموعه‌داده X را با $|x_1|$ نشان می‌دهیم.

خوشه‌بندی: یک مجموعه‌ای را از c زیرمجموعه از داده‌ها خوشه‌بندی یا افراز می‌گوییم اگر اجتماع زیرمجموعه‌ها، کل مجموعه‌داده و اشتراک هر زوج از زیرمجموعه‌ها تهی باشد. به هر زیرمجموعه‌ای از یک مجموعه‌داده خوشه می‌گوییم. یک خوشه‌بندی را با $\pi = \{\pi^1, \pi^2, \dots, \pi^c\}$ نشان می‌دهیم که π^i نشان‌دهنده خوشه i-ام است. بدیهی است که $\forall i, j \in \{1,2,\dots,c\}: \pi^i \cap \pi^j = \emptyset$ و $\mathrm{U}_{i=1}^{c} \pi^i = \mathrm{X}$. مرکز هر خوشه π^i را با C^{π^i} نشان می‌دهیم و j-امین ویژگی آن به‌شکل رابطه (۱) تعریف می‌شود [4].

[DOI: 10.29252/jsdp.17.4.103]

$$C_j^{\pi^i} = \frac{\sum_{k \in \pi^i} x_{kj}}{|\pi^i|} \quad (۱)$$

(جدول-۱): شرح علائم

(Table-1): Description of symbols

نماد	شرح
X	یک مجموعه داده
$\boldsymbol{x_i}$	i-امین شی داده X
$\boldsymbol{x_{ij}}$	j-امین ویژگی از شی داده x_i
$\|X\|$	اندازه یک مجموعه X
$\|x_1\|$	تعداد ویژگی‌های مجموعه داده X
π	یک مجموعه از خوشه‌های اولیه یا یک خوشه‌بندی
π^i	i-امین خوشه در خوشه‌بندی π
$c = \|\pi\|$	تعداد خوشه‌های π
r_{π^i}	زیرخوشه معتبر از یک خوشه π^i
γ	پارامتر شعاع همسایگی خوشه معتبر
C^{π^i}	نقطه مرکز خوشه π^i
$C_w^{\pi^i}$	w-امین ویژگی از نقطه مرکز خوشه π^i
π^*	خوشه‌بندی توافقی
$\|\pi^*\|$	تعداد خوشه‌های خوشه‌بندی توافقی
$sim(\pi^i, \pi^j)$	شباهت بین دو خوشه π^i و π^j
$T_q(\pi^i, \pi^j)$	q-امین خوشه فرضی بین دو خوشه π^i و π^j
$p_q(\pi^i, \pi^j)$	مرکز q-امین خوشه فرضی بین دو خوشه π^i و π^j
Π	خوشه‌بندی پایه $\|\Pi\|$ یک اجماعی از
B	اندازه اجماع؛ $B = \|\Pi\|$
π_i	i-امین خوشه‌بندی در اجماع Π
c_i	تعداد خوشه‌های خوشه‌بندی π_i؛ $c_i = \|\pi_i\|$
$G(\Pi)$	گراف تعریف شده بر روی اجماع Π
$V(\Pi)$	گره‌های گراف تعریف شده بر روی اجماع Π
$E(\Pi)$	یال‌های گراف تعریف شده بر روی اجماع Π
L	خوشه‌های واقعی تعریف‌شده برحسب برچسب‌های داده

زیرخوشه معتبر از یک خوشه: زیرخوشه معتبر از یک خوشه π^i به‌شکل r_{π^i} نمایش داده می‌شود و بنابر رابطه (۲) تعریف می‌شود:

$$r_{\pi^i} = \left\{ x_k : \pi^i \mid \sqrt{\sum_{j=1}^{|x_1|} \left| C_j^{\pi^i} - x_{kj} \right|^2} \le \gamma \right\} - \bigcup_{j=1}^{i-1} r_{\pi^j} \quad (۲)$$

که γ یک پارامتر است. گفتنی است که یک زیرخوشه می‌تواند به‌عنوان یک خوشه در نظر گرفته شود.

شباهت بین دو خوشه: معیارهای فاصله/شباهت بین دو خوشه گوناگونی وجود دارد. در حال حاضر، معیارهای اندازه‌گیری فاصله بین خوشه‌های گوناگونی مطرح هستند [49-52] که ما نیز در این مقاله، به فراخور نیاز یک معیار جدید را معرفی می‌کنیم. بر اساس معیار معرفی‌شده در این مقاله، شباهت بین دو خوشه π^i و π^j را که با $sim(\pi^i, \pi^j)$ نشان می‌دهیم، به‌شکل رابطه (۳) تعریف می‌کنیم:

(۳)

$$\mathrm{sim}(\pi^i, \pi^j) = \begin{cases} \dfrac{\pi^i \cap \pi^j}{\pi^i \cup \pi^j} + \dfrac{\bigcup_{q=1}^{9} \mathrm{T_q}(\pi^i, \pi^j) - (\pi^i \cup \pi^j)}{\sqrt{\sum_{w=1}^{|x_1|} \left| C_w^{\pi^i} - C_w^{\pi^j} \right|^2}} & if \ \sqrt{\sum_{w=1}^{|x_1|} \left| C_w^{\pi^i} - C_w^{\pi^j} \right|^2} \le 4\gamma \\ 0 & ow. \end{cases}$$

که $\mathrm{T_q}(\pi^i, \pi^j)$ از رابطه زیر به‌دست می‌آید:

(۴)

$$\mathrm{T_q}(\pi^i, \pi^j) = \left\{ \mathrm{x_k} : \mathrm{X} \mid \sqrt{\sum_{w=1}^{|x_1|} \left| \mathrm{p_{qw}}(\pi^i, \pi^j) - x_{kw} \right|^2} \le \gamma \right\}$$

که $p_q(\pi^i, \pi^j)$ یک نقطه است که ویژگی w-ام آن بنابر رابطه (۵) تعریف می‌شود:

$$p_{qw}(\pi^i, \pi^j) = \frac{(q) \times C_w^{\pi^i} + (10 - q) \times C_w^{\pi^j}}{10} \quad (۵)$$

اجماع خوشه‌بندی: یک مجموعه‌ای از B خوشه‌بندی از داده‌ها را یک اجماع خوشه‌بندی می‌گوییم و آن را با $\Pi = \{\pi_1, \pi_2, \dots, \pi_B\}$ نشان می‌دهیم که π_i نشان‌دهندهٔ خوشه‌بندی i-ام است. بدیهی است که $\pi_i = \{\pi_i^1, \pi_i^2, \dots, \pi_i^{c_i}\}$ و c_i نشان‌دهنده تعداد خوشه‌های خوشه‌بندی i-ام است. خوشه j-ام از خوشه‌بندی i-ام با π_i^j نشان داده شده است. خوشه‌بندی هدف یا بهترین خوشه‌بندی را با π^* نشان می‌دهیم.

گراف وزن‌دار متناظر با یک اجماع خوشه‌بندی: یک گراف وزن‌دار متناظر با یک اجماع خوشه‌بندی Π را با $G(\Pi)$ نشان می‌دهیم و به شکل $G(\Pi) = (V(\Pi), E(\Pi))$ تعریف می‌شود. مجموعه رئوس این گراف همان زیرخوشه‌های معتبر همه خوشه‌های اجماع است؛ یعنی:

$$V(\Pi) = \left\{ r_{\pi_1^1}, \dots, r_{\pi_1^{c_1}}, r_{\pi_2^1}, \dots, r_{\pi_2^{c_2}}, \dots, r_{\pi_B^1}, \dots, r_{\pi_B^{c_B}} \right\}$$

وزن یال‌های بین رئوس این گراف، یا یال‌های بین خوشه‌ها میزان مشابهت آن‌ها است و مطابق رابطه (۶) به‌دست می‌آید.

[DOI: 10.29252/jsdp.17.4.103]

$$E\left(v_i, v_j\right) = sim\left(v_j, v_i\right) \quad (۶)$$

۲-۳- الگوریتم خوشه‌بندی برای تولید خوشه‌های پایه فازی متنوع

پیچیدگی زمانی الگوریتم cmeans فازی $O(|X|cI)$ است؛ به‌طوری که I تعداد تکرارها است. گفتنی است که الگوریتم cmeans فازی یک یادگیرنده ضعیف است که عملکرد آن تحت تأثیر عوامل بسیاری قرار دارد. به‌عنوان مثال، الگوریتم بسیار حساس به مراکز اولیه خوشه است. به‌طوری‌که انتخاب مراکز اولیه مختلف، اغلب منجر به نتایج خوشه‌بندی متفاوتی می‌شود؛ علاوه‌براین، الگوریتم cmeans فازی تمایل به کشف خوشه‌های کروی با اندازه‌های به‌نسبه یک‌نواخت دارد که برای دیگر توزیع داده‌ها، مناسب نیست؛ بنابراین، ما تلاش خواهیم کرد تا خوشه‌بندی‌های چندگانی تولیدشده به‌وسیلهٔ الگوریتم cmeans فازی را برای ایجاد یک نتیجه خوشه‌بندی خوب در مجموعه‌داده‌ها، با توزیع داده‌های مختلف، به‌جای یک خوشه‌بند قوی داشته باشیم. به‌دلیل اهمیت موضوع تنوع در میان خوشه‌بندی‌های پایه، ما در ابتدا در مورد یک الگوریتم خوشه‌بندی محلی‌معتبر بحث می‌کنیم. تولید خوشه‌بندی‌های پایه بر اساس الگوریتم شکل (۱) انجام می‌پذیرد. در الگوریتم پیشنهادی که مبتنی بر شکل (۱) است، B افراز با کمک الگوریتم خوشه‌بندی پایه که در شکل (۲) آورده شده است، تولید می‌شود. هر افراز تعداد خوشه‌هایی بین c تا $\min\left(\sqrt{|X|}, 100\right)$ دارد. بعد از اجرای الگوریتم خوشه‌بندی پایه، ممکن است الگوریتم پایه تعداد بیشتر خوشه تولید کند (درحقیقت به‌اندازه ضریبی از تعداد خوشه‌های که به‌عنوان ورودی دریافت می‌کند، خوشه تولید می‌کند). الگوریتم ارائه‌شده در شکل (۲)، یعنی الگوریتم تولید خوشه‌های محلی‌معتبر، هر بار یک الگوریتم خوشه‌بندی فازی cmeans بهبودیافته را با تعداد خوشه‌های ورودی فراخوانی و به آن اندازه خوشه اولیه محلی‌معتبر تولید می‌کند؛ پس این الگوریتم یک ضریبی از تعداد خوشه‌های ورودی، خوشه اولیه محلی‌معتبر تولید می‌کند. این روش یادگیری افزایشی را برای حل مسأله تولید خوشه‌های محلی‌معتبر استفاده می‌کند. با تبدیل مسأله تولید خوشه‌های محلی‌معتبر به یک مسأله افزایشی، این روش به‌تدریج خوشه‌بندی‌های پایه را در هر مرحله تولید می‌کند. در روش یادگیری افزایشی، در هر مرحله به‌طور تصادفی، c شیء داده را از بین داده‌های هنوزخوشه‌بندی‌نشده به‌عنوان مراکز خوشه‌های اولیه انتخاب و از آن در الگوریتم خوشه‌بندی فازی cmeans بهبودیافته استفاده می‌کنیم. این روال تا زمانی‌که تعداد داده‌های هنوزخوشه‌بندی‌نشده کمتر از c^2 نشود، ادامه داده می‌شود [53-54]. پیچیدگی زمانی الگوریتم خوشه‌بندی فازی cmeans بهبودیافته که در شکل (۳) آورده شده، $O(|X|I\min(c, 100))$ است؛ به‌طوری‌که I تعداد تکرارها است. پیچیدگی زمانی کلی الگوریتم تولید افرازهای اولیه که در شکل (۱) آورده شده، $O\left(|X|I\sum_{i=1}^{B} c_i\right)$ است به‌طوری که I تعداد تکرارها است.

The algorithm for generating a diverse clustering ensemble
Input: X, B, c Output: Π 01. $\Pi = \emptyset$; 02. γ =rand× 0.5; 03. For i=1 to B 04. $n = \|X\|$; 05. c_i=rand($[c, \ldots, \min(\sqrt{X}, 100)]$); 06. $[\pi_i, c_i]$=BaseClustering(X, c_i, γ); 07. $\Pi = \Pi \cup \{\pi_i\}$; 08. EndFor 09. Return Π

(شکل-۱): شبه‌کد تولید اجماعی از خوشه‌بندی‌های محلی‌معتبر پایه

(Figure-1): A pseudocode of consensus production of local clustering - valid base

The base clustering algorithm
Input: X, c, γ Output: π, c 01. $\pi = \emptyset$; 02. $TempX = \emptyset$; 03. $counter = 0$; 04. $CN = 0$;

[DOI: 10.29252/jsdp.17.4.103]

05. While $\left((|X| - |TempX|) \geq c^2\right)$
06. $[\pi^{CN+1}, \dots, \pi^{CN+c}]$=modifiedfuzzycmeans($X$, c, γ, $TempX$);
07. $CN = CN + c$;
08. For k=1 to c
09. If $(\left|\pi^{counter \times c+k}\right| \leq c)$
10. $\pi^{counter \times c+k} = \emptyset$; $counter = counter + 1$;
11. EndIf
12. $TempX = TempX \cup \pi^{counter \times c+k}$;
13. EndFor
14. EndWhile
15. $c = CN - counter$;
16. Remove all empty clusters from π;
17. Return π, c;

(شکل-۲): شبه‌کد الگوریتم تولید یک خوشه‌بندی محلی‌معتبر پایه

(Figure-2): Pseudo-code algorithm for producing a valid local-clustering base

The modified fuzzy c-means clustering algorithm

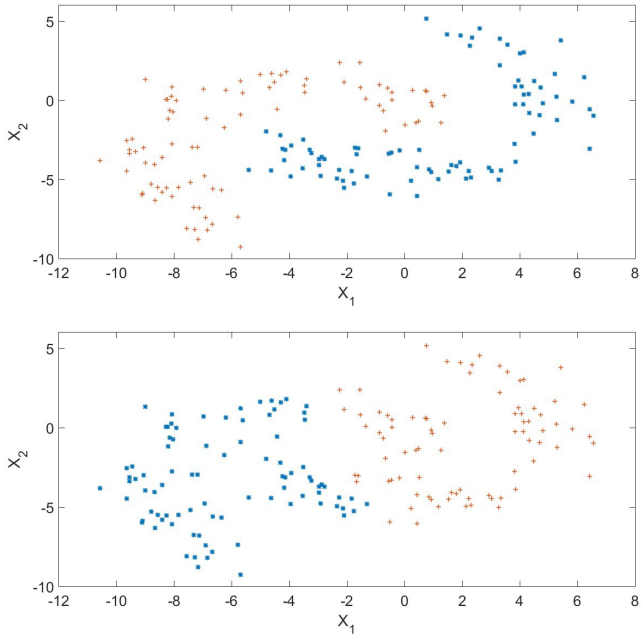
Input: X, c, γ, RDI % Removed Data Index
Output: $[r_{\pi^1}, r_{\pi^2}, \dots, r_{\pi^c}]$
01. $\forall i \in \{1,2, \dots, c\}$: $\pi^i = \emptyset$;
02. $counter$=0;
03. $Y = \emptyset$;
04. $Temp = \{1, \dots, |X|\} - RDI$;
05. For each j in $Temp$
06. $counter = counter + 1$; $y_{\text{counter}} = x_j$;
07. If $(counter \leq c)$
08. $C_{counter} = y_{counter}$;
09. EndFor
10. For j=1 to $MaxIteration$
11. For p=1 to $|X|$
12. For k=1 to c
13. $dis_{kp} = |x_i - C_k|$;
14. $DIR_{kp} = \begin{cases} 1 & dis_{kp} < \gamma \\ 0 & \text{o. w.} \end{cases}$;
15. If $(p \in RDI)$ $DIR_{kp} = -DIR_{kp}$;
16. EndFor
17. $ADIR_k = \sum_{k=1}^{c} |DIR_{kp}|$;
18. cln_p=arg $\min_{k \in \{1,\dots,c\}} dis_{kp}$;
19. If $\left((ADIR_k > 1) \& (p \notin RDI)\right)$
20. For p=1 to c
21. $DIR_{kp} = -DIR_{kp}$;
22. $DIR_{k,cln_p} = 1$;
23. EndIf
24. EndFor
25. For k=1 to c
26. $\Delta C = \frac{\left(DIR \times \left(X - \begin{bmatrix} C_k \\ \vdots \\ C_k \end{bmatrix}_{|X| \times f}\right)\right)}{ADIR_k}$;
27. EndFor
28. $C = C + \Delta C$;
29. EndFor
30. For p=1 to $|X|$
31. For k=1 to c
32. If $(DIR_{kp} == 1)$
33. $r_{\pi^k} = r_{\pi^k} \cup \{p\}$;
34. EndIf
35. EndFor
36. EndFor
37. Return $[r_{\pi^1}, r_{\pi^2}, \dots, r_{\pi^c}]$

(شکل-۳): شبه‌کد الگوریتم خوشه‌بندی فازی cmeans بهبودیافته

(Figure-3): Pseudo-coding of fuzzy clustering algorithms improved cmeans

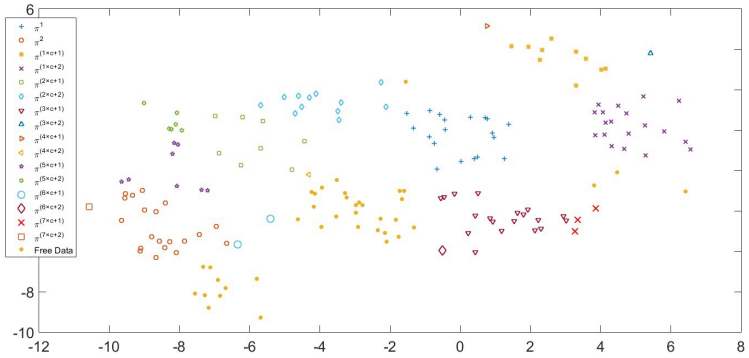
[DOI: 10.29252/jsdp.17.4.103]

مثال زیر را در شکل (۴) ببینید. یک مجموعه داده تصنعی با دو خوشه و ۱۷۰ داده که در هر خوشه ۸۵ شیءداده موجود است، در شکل (۴-الف) نشان داده شده است. خروجی الگوریتم خوشه‌بندی فازی cmeans بر روی این داده‌ها در شکل ۴-ب قابل مشاهده است. چنان که در شکل (۴) مشهود است، الگوریتم خوشه‌بندی فازی cmeans بر روی این داده‌ها ناکارامد است. در شکل (۵) سه خروجی الگوریتم خوشه‌بندی فازی cmeans بهبودیافته (ارائه‌شده در شکل ۲) بر روی این داده‌گان به تصویر کشیده شده است. در هر یک از دو اجرای نخست و دوم از این الگوریتم (نشان‌داده‌شده در دو تصویر بالاتر در شکل ۵)، هشت بار حلقه موجود در سطر پنج الگوریتم پیشنهادی (ارایه‌شده در شکل ۲) تکرار می‌شود که منجر به تولید هشت زوج خوشه یعنی ۱۶ خوشه می‌شود؛ و درنهایت تعدادی داده که بدون خوشه مانده‌اند، به‌عنوان داده‌های Free Data مشخص شده‌اند؛ ولی در اجرای سوم از این الگوریتم (نشان‌داده‌شده در دو تصویر پایینی در شکل ۵)، هفت‌بار حلقه موجود در سطر پنج الگوریتم پیشنهادی (ارایه‌شده در شکل ۲) تکرار می‌شود که منجر به تولید هفت زوج خوشه یعنی چهارده خوشه می‌شود. از خروجی‌های این الگوریتم مشاهده می‌کنیم که این خوشه‌های پایه تا اندازه‌ای متنوع بوده، که این حقیقت برای خوشه‌بندی اجماعی بسیار مفید است. توجه داشته باشید که تعداد خوشه‌های محلی‌معتبر در یک خوشه‌بندی پایه بستگی به پارامتر اندازه همسایگی γ دارد. هنگامی‌که این مقدار کم باشد، تعداد خوشه‌های محلی‌معتبر در یک خوشه‌بندی پایه زیاد است و بالعکس؛ بنابراین، اگر این مقدار را به مقادیر کوچک‌تر تنظیم کنیم، نیاز به خوشه‌بندی‌های پایه بیشتری در اجماع داریم. تنظیم این پارامتر بسته به نیاز می‌توانند تغییر کند.

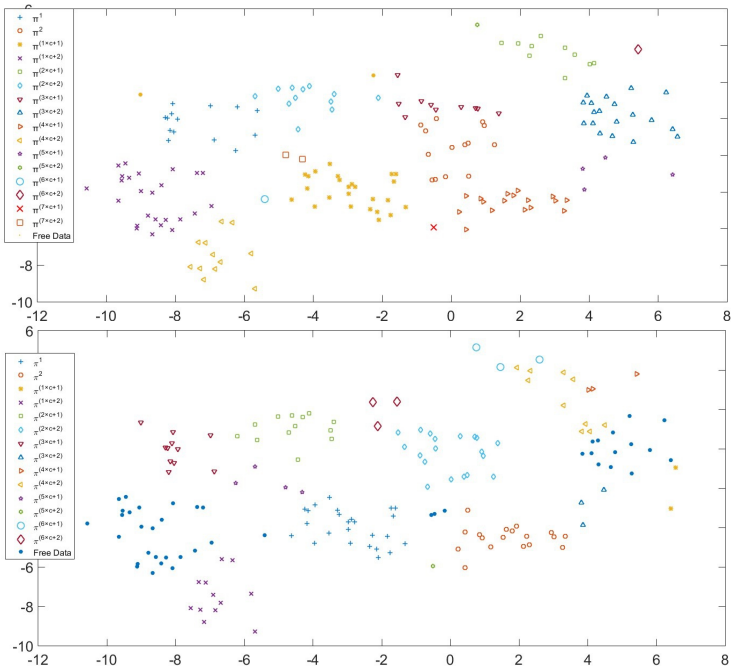


(شکل-۴): (الف) یک مجموعه‌داده و برچسب‌های خوشه واقعی (بالا). (ب) نتیجه اعمال الگوریتم خوشه‌بندی cmeans فازی (پایین).

(Figure-4): (a) A dataset and true cluster labels (high). (B) The result of the fuzzy cmeans clustering algorithm (bottom).



[DOI: 10.29252/jsdp.17.4.103]



(شکل-۵): نتیجه اعمال سه بار الگوریتم خوشه‌بندی cmeans فازی بهبود یافته بر روی مجموعه داده شکل (۴) (دقت شود c = 2 است).

(Figure-5): The result of applying the fuzzy cmeans clustering algorithm improved on the data set in Figure 4. (Note c = 2).

۳-۳- تابع توافقی پیشنهادی

الگوریتم خوشه‌بندی ترکیبی پیشنهادی در شکل (۶) ارایه شده است. پیچیدگی کلی این الگوریتم برابر با:

$O\left(|X|(I\sum_{i=1}^{B}c_i+\sum_{i=1}^{B}c_i+(\sum_{i=1}^{B}c_i)^2+B)\right)$ است.

قابل مشاهده است که رابطه پیچیدگی زمانی الگوریتم با تعداد اشیای خطی است و برای یادگیری ترکیبی، بیشتربودن تعداد B خوشه‌های پایه به‌منزله بهتربودن نیست؛ بنابراین، به‌طورکلی ما می‌توانیم فرض کنیم که عبارت:

$$\sum_{i=1}^{B}c_i \le \sum_{i=1}^{B} min\left(\sqrt{|X|},100\right) \le 100B \ll |X|$$

(به‌خصوص برای داده‌های خیلی بزرگ) صحیح است؛ به‌طوری این نشان‌دهنده آن است که الگوریتم پیشنهادی برای مقابله با مجموعه‌داده‌های با اندازه بسیار بزرگ نیز مناسب است.

ورودی: X، c، B، γ

خروجی: π^*

۰۱. ایجاد یک مجموعه خوشه‌بندی پایه به کمک الگوریتم شکل ۲ وقتی تعداد خوشه‌ها برابر یک است c

۰۲. ساخت یک گراف وزن‌دار از اجماع به دست آمده از مرحله قبل

۰۳. افراز گراف وزن‌دار به دست آمده از اجماع

۰۴. به دست آوردن یک خوشه‌بندی نهایی π^* از طریق خروجی افراز گراف

۰۵. بازگرداندن خوشه‌بندی نهایی π^*

(شکل-۶): شبه‌کد الگوریتم خوشه‌بندی ترکیبی پیشنهادی

(Figure-6): Pseudo-code of proposed ensemble clustering algorithm

۴- مطالعات تجربی

در بخش جاری، الگوریتم خوشه‌بندی ترکیبی پیشنهادی را بر روی چهار مجموعه‌داده مصنوعی و پنج مجموعه‌داده واقعی محک می‌زنیم و برحسب کارایی و زمان اجرا با الگوریتم‌های به‌روز مقایسه می‌کنیم.

(جدول-۲): شرح مجموعه‌داده‌ها: تعداد اشیاء داده ($|X|$)، تعداد صفات ($|x_1|$)، تعداد خوشه‌ها (c)

(Table-2): Description of the data set: The number of data objects ($|X|$), the number of traits ($|x_1|$), the number of clusters (c)

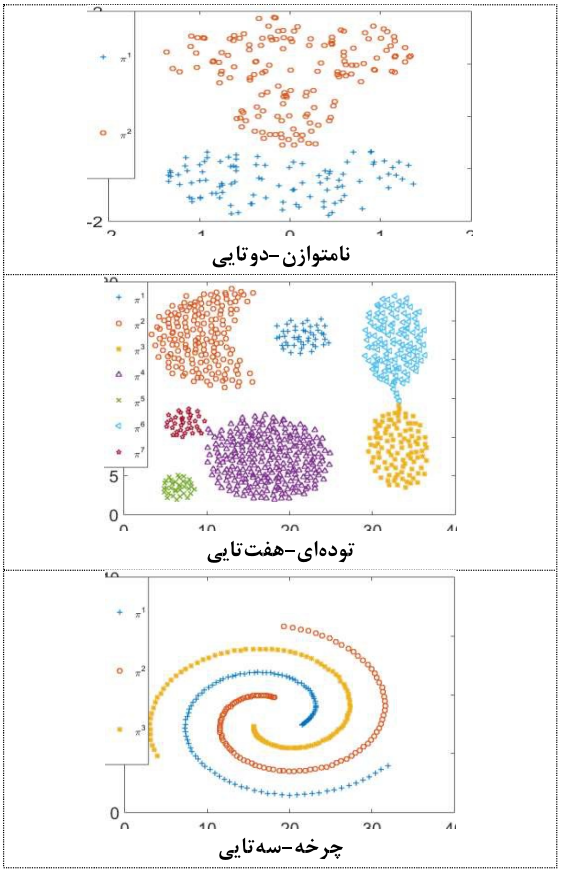
مجموعه داده	$\|X\|$	$\|x_1\|$	c
چرخه-سه‌تایی (شکل ۷-پایین)	600	2	3
موزی-دوتایی (شبیه شکل ۴-بالا، فقط وقتی تعداد داده‌ها ۳۰۰ تا باشد)	300	2	2
توده‌ای-هفت‌تایی (شکل ۷-وسط)	788	2	7
نامتوازن-دوتایی (شکل ۷-بالا)	300	2	2
Iris	150	4	3
Wine	178	13	3
Breast	569	30	2
Digits	5620	63	10
KDD-CUP'99	1048576	39	2

۴-۱- توصیف مجموعه‌داده‌ها

ارزیابی‌های تجربی بر روی نُه مجموعه‌داده انجام شده است. جدول (۲) جزئیات این مجموعه داده‌ها را نشان می‌دهد.

[DOI: 10.29252/jsdp.17.4.103]

توزیع خوشه‌ای این مجموعه‌داده‌های مصنوعی در شکل (۷) نشان داده شده است. مجموعه‌داده‌های واقعی برگرفته از مجموعه‌داده‌های UCI هستند [55].



(شکل-۷): توزیع سه داده مصنوعی
(Figure-7): Distribution of three artificial datasets

۲-۴- معیارهای ارزیابی

ما به‌طور گسترده از دو معیار خارجی استفاده کردیم تا شباهت بین نتیجه خوشه‌بندی و تقسیم درست بر روی مجموعه‌داده‌ها را اندازه‌گیری کنیم. با توجه به یک مجموعه‌داده X و دو افراز (یا خوشه‌بندی) از این مجموعه‌اشیا، یعنی $\pi_p = \{\pi_p^1, \pi_p^2, \dots \pi_p^{c_p}\}$ (که نتیجه یک الگوریتم خوشه‌بندی پایه است) و $L = \{L^1, L^2, \dots L^c\}$ (که افراز هدف واقعی برای مجموعه‌داده است)، n_{ij} را به‌شکل زیر تعریف می‌کنیم:

$$n_{ij} = \left|\pi_p^i \cap L^j\right| \qquad (۷)$$

همچنین $n_{i.}$ به‌شکل زیر تعریف می‌شود:

$$n_{i.} = \left|\pi_p^i\right| \qquad (۸)$$

همچنین $n_{.j}$ به‌شکل زیر تعریف می‌شود:

$$n_{.j} = \left|\pi_p^j\right| \qquad (۹)$$

n نیز تعداد داده‌های مجموعه‌داده را نشان می‌دهد. شاخص ARI تنظیم شده [52] به‌صورت زیر تعریف می‌شود:

(۱۰)

$$\mathrm{ARI}(\pi_p, L) = \frac{\binom{n}{2}\times\sum_{ij}\binom{n_{ij}}{2} - \sum_i\binom{n_{i.}}{2}\times\sum_j\binom{n_{.j}}{2}}{0.5\times\binom{n}{2}\times\left[\sum_i\binom{n_{i.}}{2}+\sum_j\binom{n_{.j}}{2}\right] - \sum_i\binom{n_{i.}}{2}\times\sum_j\binom{n_{.j}}{2}}$$

اطلاعات متقابل نرمال‌شده [56] به‌صورت زیر تعریف می‌شود:

(۱۱)

$$\mathrm{NMI}(\pi_p, L) = -\frac{2\times\sum_i\sum_j n_{ij}\times\log\frac{n_{ij}\times n}{n_{i.}\times n_{.j}}}{\sum_{ij} n_{i.}\times log\frac{n_{i.}}{n} + \sum_{ij} n_{.j}\times\log\frac{n_{.j}}{n}}$$

هر دو معیار، عددی بین صفر و یک را بر می‌گردانند که هر چه عدد یادشده بزرگ‌تر باشد، به‌منزله این است که افراز نتیجه، نزدیک‌تر به افراز واقعی است.

۳-۴- روش‌های به‌روز مقایسه‌شده

برای این‌که درستی عملکرد الگوریتم پیشنهادی را بررسی کنیم، آن را با الگوریتم های خوشه‌بندی ترکیبی زیر مقایسه کردیم.

۱. الگوریتم‌های شباهت دوگانه شامل ماتریس هم‌رخدادی: روش‌های مبتنی بر ماتریس همرخدادی [15] و سه ماتریس شباهت مبتنی بر پیوند WCT، WTQ و CSM [31] از جمله روش‌های این دسته محسوب می‌شوند. همچنین الگوریتم‌های سلسله‌مراتبی پیوندتکی (SL) و پیوندمتوسط (AL) برای به‌دست‌آوردن خوشه‌بندی توافقی نهایی در این مورد استفاده شده است.

۲. الگوریتم‌های مبتنی بر ابرگراف: این روش‌ها شامل CSPA، HGPA و MCLA [10] هستند.

۳. الگوریتم‌های مبتنی بازبرچسب‌گذاری: این روش‌ها عبارتند از الگوریتم‌های SV و SWV [22].

۴. الگوریتم‌های مبتنی بر فضای واسط: این روش‌ها شامل الگوریتم بیشینه انتظار (EM) [41] و الگوریتم توافق رأی‌گیری تکراری (IVC) [44] است؛ علاوه‌بر این، روش

[DOI: 10.29252/jsdp.17.4.103]

پیشنهادی را با دیگر الگوریتم‌های خوشه‌بند قوی پایه مقایسه خواهیم کرد. الگوریتم‌های خوشه‌بند قوی پایه مورد مقایسه شامل الگوریتم خوشه‌بندی طیفی نرمال (NSC) [9]، خوشه‌بندی فضایی مبتنی بر تراکم در شرایط نوفه‌ای (DBSCAN) [6] و خوشه‌بندی با جستجوی سریع و پیداکردن قله‌های چگالی (CFSFDP) [41] هستند.

۴-۴- تنظیمات پارامترهای الگوریتم‌های گوناگون

برای روشن‌بودن مقایسات تجربی، تمامی پارامترهای الگوریتم‌های گوناگون در این بخش ارایه شده‌اند تا بتوان نتایج روش‌های گوناگون را در صورت نیاز دوباره بازتولید کرد؛ همچنین برای اطمینان از منصفانه‌بودن نتایج به‌دست‌آمده برای روش‌های گوناگون، نتایج ارایه‌شده میانگین پنجاه‌بار اجرای مجزای هر الگوریتم خواهد بود. تعداد خوشه‌های موجود در هر خوشه‌بندی پایه برابر است با عددی تصادفی بین تعداد واقعی خوشه‌ها در هر یک از مجموعه داده‌های مورد نظر تا دست‌کم صد یا جذر تعداد داده‌های آن مجموعه داده؛ همچنین تعداد خوشه‌های خوشه‌بندی توافقی برابر تعداد واقعی خوشه‌ها در هر یک از مجموعه داده‌های مورد نظر در نظر گرفته می‌شود. همچنین الگوریتم خوشه‌بند cmeans فازی بهبودیافته به‌عنوان مولد خوشه‌بندی‌های پایه استفاده شده است.

دو روش برای خوشه‌بندی‌های پایه وجود دارد:

۱) اجرای kmeans به تعداد B بار مجزا، که هر کدام دارای مقادیر موقعیت اولیه خوشه‌های متفاوتی هستند.

۲) اجرای الگوریتم پیشنهادی (شکل ۱) برای تولید خوشه‌بندی‌های پایه.

برای ارائه نتایج روش‌های مقایسه‌شده، ما بر اساس پیشنهادها نویسندگان، پارامترهای آن‌ها را تعیین می‌کنیم. کیفیت هر یک از الگوریتم‌های خوشه‌بندی مبتنی بر اجماع با توجه به تنظیمات آن اجماع خاص، با میانگین پنجاه اجرا می‌شوند.

برای الگوریتم NSC، از هسته گاوسی استفاده شده است و پارامتر هسته را از مجموعه $\{0.1,0.2,\dots,1.9,2.0\}$ انتخاب کرده‌ایم. در بین همه این مقادیر، مقداری که به بهترین نتیجه خوشه‌بندی منتج شده، برای مقایسه انتخاب شده است.

الگوریتم‌های DBSCAN و CFSFDP نیز به پارامتر ورودی ε نیاز دارند. مقدار ε استفاده‌شده با استفاده از $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d(X_i, \bar{X})$ تخمین زده شده که به‌صورت $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ است. با این‌حال، هر یک از این الگوریتم‌ها ممکن است، نیاز به مقادیر مختلف ε داشته باشند؛ بنابراین، هر یک از این الگوریتم‌ها با ده مقدار مختلف آزمایش شده او پارامتری که منتج به بهترین نتیجه خوشه‌بندی شده است، برای مقایسه انتخاب شده است. مقادیر مختلف ε پژوهش‌شده در این مقاله شامل مجموعه زیر است:

$$\{\bar{d},\ \bar{d}/2, \bar{d}/3, \bar{d}/4, \bar{d}/5, \bar{d}/6, \bar{d}/7, \bar{d}/8, \bar{d}/9, \bar{d}/10\}$$

۵-۴- نتایج تجربی

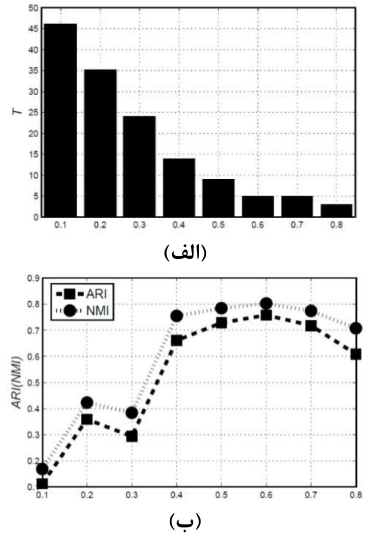
مقایسه با روش‌های ترکیبی دیگر: براساس معیارهای اعتباری ARI و NMI، شکل (۱۰ و ۱۱) نشان‌دهنده مقایسه عملکرد الگوریتم‌های خوشه‌بندی مختلف بر روی مجموعه داده‌های مصنوعی و واقعی هستند. بر طبق شکل (۱۰)، مشاهده می‌کنیم که الگوریتم خوشه‌بندی ترکیبی پیشنهادی دارای دقت خوشه‌بندی بسیار بالا در این مجموعه داده‌های مصنوعی نسبت به دیگر الگوریتم‌های موجود است. نتایج تجربی نشان می‌دهد که الگوریتم ترکیبی پیشنهادشده می‌تواند به‌طور مؤثر خوشه‌های مختلف‌الشکل را کشف و عملکرد الگوریتم خوشه‌بندی پایه را افزایش دهد؛ همچنین در شکل (۱۱) مشاهده می‌کنیم که کارایی الگوریتم خوشه‌بندی ترکیبی پیشنهادی بهتر از الگوریتم‌های دیگر در مجموعه‌داده‌های واقعی است؛ با این‌حال، بهبود دقت الگوریتم خوشه‌بندی ترکیبی پیشنهادی در مجموعه‌داده‌های واقعی از آن‌چه در مجموعه‌داده‌های مصنوعی است، کم‌تر است. دلیل اصلی آن، این است که پیچیدگی مجموعه‌داده‌های واقعی بزرگ‌تر از مجموعه‌داده‌های مصنوعی است؛ علاوه‌براین، با توجه به مقایسه‌ها، مشاهده می‌کنیم که بیش‌تر الگوریتم‌های موجود در سازوکار تولید خوشه‌بندی‌های پایه تصادفی از روش تولید خوشه‌بندی‌های پایه پیشنهادی بهتر عمل می‌کنند؛ زیرا هر خوشه‌بندی پایه تولیدشده به‌وسیلهٔ روش تولید خوشه‌بندی‌های پایه پیشنهادی، قابل فهم کامل به‌وسیلهٔ آن‌ها نیستند؛ بنابراین، آن‌ها نمی‌توانند یک نتایج خوشه‌بندی خوب را در روش تولید خوشه‌بندی‌های پایه پیشنهادی به‌دست آورند. الگوریتم خوشه‌بندی ترکیبی پیشنهادی عملکرد بهتری در

[DOI: 10.29252/jsdp.17.4.103]

روش تولید خوشه‌بندی‌های پایه پیشنهادی در مقایسه با الگوریتم‌های دیگر دارد. توجه داشته باشید که الگوریتم خوشه‌بندی ترکیبی پیشنهادی فقط در روش تولید خوشه‌بندی‌های پایه پیشنهادی اجرا می‌شود، به این دلیل که روش تولید خوشه‌بندی‌های پایه پیشنهادی بخشی از آن است. مشاهده می‌کنیم که الگوریتم خوشه‌بندی ترکیبی پیشنهادی در روش تولید خوشه‌بندی‌های پایه پیشنهادی نیز بر اساس الگوریتم‌های دیگر در روش تولید خوشه‌بندی‌های پایه تصادفی، از نظر ARI و NMI بهتر عمل می‌کند.

مقایسه با الگوریتم‌های پایه مقاوم: شکل (۱۲) 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 $(standard\ diviation)$ اعتبار خوشه‌بندی هر الگوریتم برای این مجموعه داده‌ها نیز بیان شده است. مشاهده می‌کنیم که اعتبار خوشه‌بندی به‌دست‌آمده با الگوریتم خوشه‌بندی ترکیبی پیشنهادی برتر یا نزدیک به بهترین نتایج حاصل از سه الگوریتم دیگر است. این آزمایش‌ها به ما می‌گویند که الگوریتم پیشنهادی می‌تواند نتایج مقاوم به‌دست‌آمده به‌وسیلهٔ الگوریتم‌های خوشه‌بندی مقاوم را تولید کند.

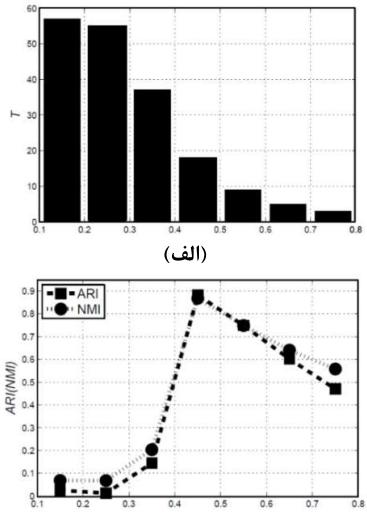
تجزیه و تحلیل پارامتر: نحوه تنظیم پارامتر شعاع همسایگی خوشه معتبر (γ) در الگوریتم خوشه‌بندی ترکیبی پیشنهادی یک چالش مهم محسوب می‌شود. ما در بخش ۳ بحث کردیم که انتخاب این پارامتر بر روی تعداد خوشه‌بندی‌های پایه‌ای تأثیر مستقیم دارد. اینک با بررسی اثر این پارامتر بر روی عملکرد الگوریتم خوشه‌بندی ترکیبی پیشنهادی با آزمایش بر روی داده‌های Iris و Wine در شکل‌های (۸-الف و ۹-الف) مشاهده می‌کنیم که تعداد خوشه‌های پایه تولیدشده به‌وسیلهٔ الگوریتم خوشه‌بندی ترکیبی پیشنهادی با افزایش مقدار این پارامتر کاهش می‌یابد. با این‌حال، شکل‌های (۸-ب و ۹-ب) نشان می‌دهند که دقت خوشه‌بندی افزایش نمی‌یابد؛ بنابراین مقدار این پارامتر بایستی تا حد مشخصی رشد کند. این آزمایش‌ها به ما می‌گویند که تعداد خوشه‌بندی‌های پایه برای به‌دست‌آوردن یک نتیجه خوب، داشتن اجماع‌های بزرگ‌تر از یک آستانه و کوچک‌تر از یک آستانه دیگر هستند؛ بنابراین، ما باید یک مقدار مناسب از این پارامتر را برای کنترل تعداد خوشه‌بندی‌های پایه بر روی هر مجموعه‌داده انتخاب کنیم.



(الف)

(ب)

(شکل-۸): اثر پارامتر γ بر روی داده‌های Iris.
(Figure-8): Effect of the parameter s on the Iris data



(الف)

(ب)

(شکل-۹): اثر پارامتر γ بر داده‌های Wine.
(Figure-9): Effect of the parameter s on the Wine data

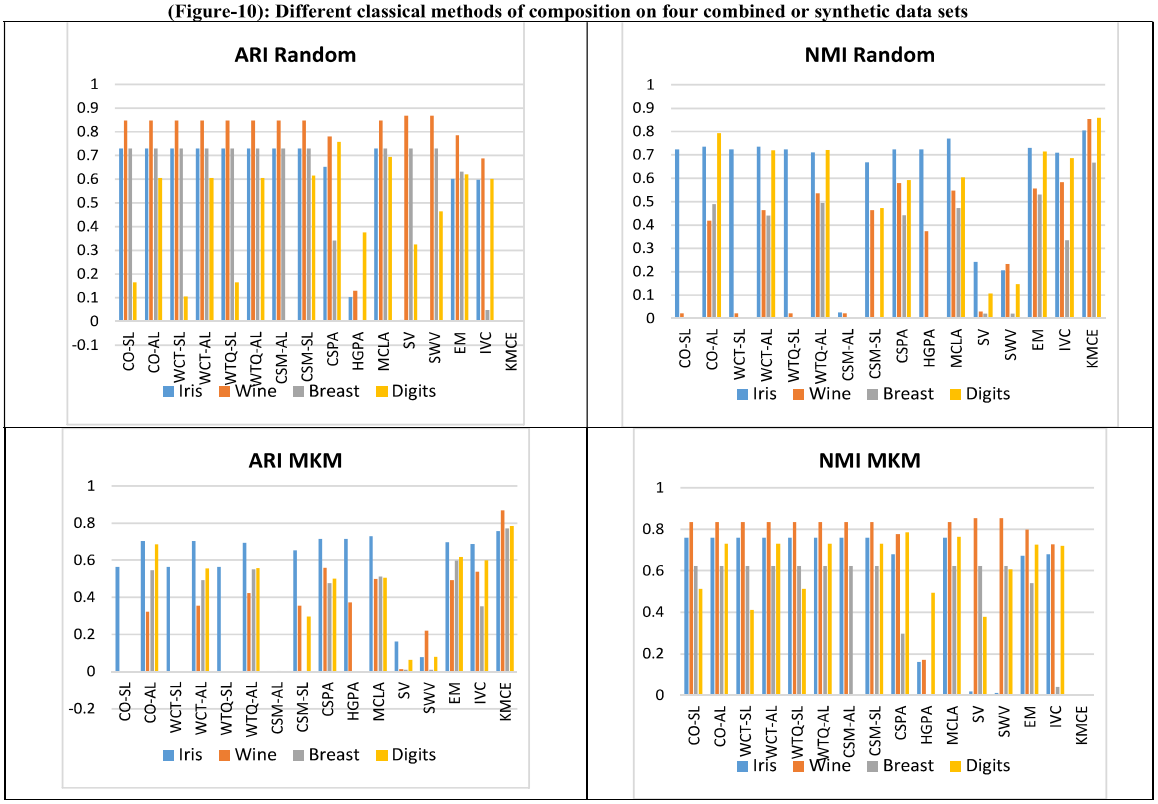
بررسی زمان محاسباتی روش پیشنهادی: در پایان بخش نتایج تجربی، کارایی الگوریتم خوشه‌بندی ترکیبی پیشنهادی را بر روی مجموعه‌دادهٔ بسیار بزرگ KDD-CUP'99 محک می‌زنیم. ما در این‌جا دو خوشه (خوشه حمله و غیرحمله) داریم و $\gamma = 0.14$ را تنظیم کرده‌ایم. تمامی الگوریتم‌ها بر روی یک دستگاه رایانه‌ای واحد اجرا شده است. جدول (۳) زمان اجرای الگوریتم پیشنهادی با تعداد شیءداده‌های مختلف را نشان می‌دهد. همان‌طورکه مشاهده می‌کنیم، با افزایش تعداد اشیا، تعداد خوشه‌های پایه، یعنی $\sum_{i=1}^{B} c_i$، نیز افزایش می‌یابد. با توجه به پیچیدگی

[DOI: 10.29252/jsdp.17.4.103]

زمانی الگوریتم پیشنهادی، رابطه هزینه زمانی با تعداد خوشه‌های پایه از نوع درجه دوم است؛ با این‌حال، با توجه به این واقعیت که می‌دانیم عبارت $\sum_{i=1}^{B} c_i$ در مقابل $|X|$ در مجموعه‌داده‌های بزرگ قابل چشم‌پوشی است و همچنین این که $\sum_{i=1}^{B} c_i$ به آرامی در مقایسه با رشد $|X|$ افزایش می‌یابد، افزایش هزینه محاسباتی به‌خاطر این افزایش در $\sum_{i=1}^{B} c_i$ چشم‌گیر نیست؛ همچنین از آن‌جایی که می‌دانیم رابطه هزینه محاسباتی اجرای الگوریتم پیشنهادشده، با تعداد اشیا مجموعه‌داده خطی است، انتظار داریم که الگوریتم پیشنهادی بتواند به‌سرعت، خوشه‌بندی نهایی را در یک مجموعه‌داده بسیار بزرگ به‌دست آورد. این مهم به‌وسیلهٔ نتایج ارایه‌شده در جدول (۳)، تایید می‌شود

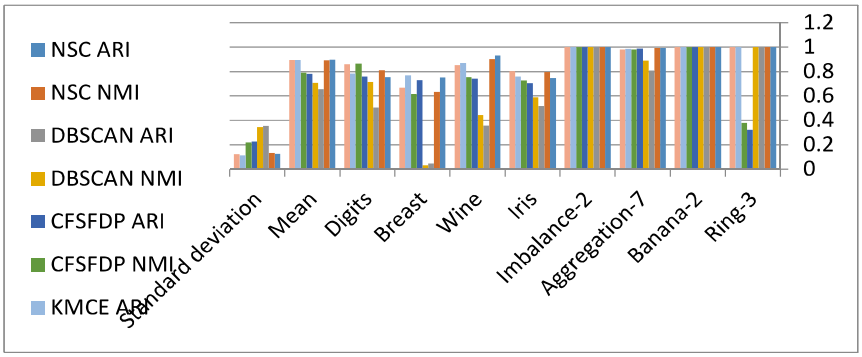
(شکل-۱۰): روش‌های مختلف کلاسیک ترکیبی بر روی چهار مجموعه‌داده‌های ترکیبی یا مصنوعی

(Figure-10): Different classical methods of composition on four combined or synthetic data sets



(شکل-۱۱): روش‌های مختلف کلاسیک ترکیبی بر روی چهار مجموعه‌داده‌های مصنوعی یا ترکیبی

(Figure-11): Different combinations of classical methods on four synthetic or combined data sets



(شکل-۱۲): مقایسه با الگوریتم‌های خوشه‌بندی "قوی"

(Figure-12): Comparison with Strong Clustering Algorithms

[DOI: 10.29252/jsdp.17.4.103]

(جدول-۳): اثر تعداد داده‌ها بر هزینه محاسباتی الگوریتم خوشه‌بندی ترکیبی پیشنهاد

(Table-3): Effect of number of data on computational cost of proposed hybrid clustering algorithm

$\|X\|$	$\sum_{i=1}^{B} c_i$	Time (in sec.)
10K	91	11.23
20K	213	51.29
30K	225	80.11
40K	232	114.06
50K	233	138.91
60K	242	178.71
70K	245	197.62
80K	353	331.02
90K	461	516.96
100K	472	576.58

آزمون آماری معناداری:

برای مقایسه آماری الگوریتم‌ها از آزمون Wilcoxon-Signed-Rank استفاده شده است. در این آزمون از ۹۸ درجه آزادی با سطح معنی‌دار برابر با ۰/۵ استفاده کرده‌ایم. بعد از انجام آزمایش نتایج بدین صورت نمایش داده می‌شوند که:

علامت + نشان می‌دهد که KMCE به‌صورت معناداری از الگوریتم رقیب معین‌شده بهتر است. علامت – نشان‌دهنده برتری معنادار الگوریتم رقیب مقابل KMCE بوده و علامت ~ نشان‌دهنده عدم وجود تفاوت معنادار بین KMCE و رقیب آن است. جداول (۴ تا ۷) نتایج آماری را برای روش پیشنهادی و سایر روش‌ها بر روی مجموعه داده‌های استاندارد مختلف نشان می‌دهند.

(جدول-۴): نتیجه آزمون آماری KMCE در مقابل سایر روش‌ها بر روی مجموعه‌داده Iris

(Table-4): The result of KMCE statistical test against other methods on the *Iris* data set

datasets	*methods*	نتیجه آزمون آماری KMCE در مقابل:	
		ARI	NMI
Iris	**NSC**	+	–
	DBSCAN	+	+
	CFSFDP	+	+

(جدول-۵): نتیجه آزمون آماری KMCE در مقابل سایر روش‌ها بر روی مجموعه‌داده *Wine*

(Table-5): The result of KMCE statistical test against other methods on the Wine data set

datasets	*methods*	نتیجه آزمون آماری KMCE در مقابل:	
		ARI	NMI
Wine	**NSC**	–	–
	DBSCAN	+	+
	CFSFDP	+	+

(جدول-۶): نتیجه آزمون آماری KMCE در مقابل سایر روش‌ها بر روی مجموعه‌داده *Breast*

(Table-6): The result of KMCE statistical test against other methods on the Breast data set

datasets	*methods*	نتیجه آزمون آماری KMCE در مقابل:	
		ARI	NMI
Breast	**NSC**	+	+
	DBSCAN	+	+
	CFSFDP	~	~

(جدول-۷): نتیجه آزمون آماری KMCE در مقابل سایر روش‌ها بر روی مجموعه‌داده *Digits*

(Table-7): The result of KMCE statistical test against other methods on the Digits data set

datasets	*methods*	نتیجه آزمون آماری KMCE در مقابل:	
		ARI	NMI
Digits	**NSC**	+	+
	DBSCAN	+	+
	CFSFDP	+	–

ارزیابی روش پیشنهادی در مقایسه با روش‌های رقیب بر روی مجموعه‌داده‌های پیچیده:

در آزمایش‌های ما از ۱۰ مجموعه داده‌های واقعی استفاده شده است[55]:

Semeion (با ۱۵۹۳ نقطه داده ، ۲۵۶ ویژگی و ده طیقه)، چندین ویژگی (MF) (با دوهزار نقطه‌داده، ۶۴۹ ویژگی و ده طبقه)، تقسیم‌بندی تصویر (IS) (با ۲۳۱۰ نقطه‌داده، نوزده ویژگی و هفت طبقه)، Forest-Cover-Type (FCT) (با ۳۷۸۰ نقطه‌داده، ۵۴ ویژگی و طبقه)، MNIST (با پنج‌هزار نقطه‌داده، ۷۸۴ ویژگی و ده طبقه)، تشخیص رقمی نوری (ODR) (با ۵۶۲۰ نقطه داده، ۶۴ ویژگی و ده طبقه)، ماهواره لندست (LS) (با ۶۴۳۵ نقطه‌داده، ۳۶ ویژگی و شش طبقه)، SOLET (با ۷۷۹۷ نقطه داده، ۶۱۷ ویژگی، و ۲۶ طبقه)، USPS خوشه‌بندی‌های پایه توسط الگوریتم‌های خوشه‌بندی k-means و fuzzy c-means تولید می‌شوند. گفتنی است که خوشه‌های تولیدشده به‌وسیلهٔ الگوریتم خوشه‌بندی fuzzy c-means ابتدا به خوشه‌های واضح تبدیل می‌شوند؛ سپس در ادامه، کل فرآیند برای هر دو الگوریتم، خوشه‌بندی پایه یکسان است. تعداد خوشه‌ها در هر خوشه‌بندی پایه به‌طور تصادفی از محدوده $[2,4K]$ انتخاب می‌شود [48].

در [57] سعی شده است یک تابع تجمعی، به‌نام خوشه‌بندی جمعی مقاوم، مبتنی بر نمونه‌برداری و خوشه‌بندی خوشه‌ای (RCESCC) ارائه شود؛ سپس، یک

[DOI: 10.29252/jsdp.17.4.103]

ماتریس شباهت خوشه خوشه‌ای از خوشه‌های فازی به‌دست می‌آورد؛ پس از آن، با استفاده از 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 [58]، یک رویکرد جدید خوشه‌بندی با استفاده از یک رویکرد وزنی ارائه شده است. در این مقاله روشی برای انجام خوشه‌بندی جمعی با بهره‌برداری از مفهوم عدم اطمینان خوشه ارائه شده است. درواقع، هر خوشه براساس عدم وابستگی آن محاسبه شده است. همه برچسب‌های خوشه‌ای پیش‌بینی‌شده موجود در اجماع برای ارزیابی عدم وابستگی خوشه‌ای، از معیار مبتنی بر تئوری اطلاعات استفاده می‌کنند. در این مقاله، دو روش مبتنی بر عدم وابستگی یا عدم اطمینان از خوشه برای برآورد قابلیت اطمینان ارائه شده است. در این مقاله دو رویکرد ارائه شده است: جمع آوری شواهد با وزن خالص و تقسیم‌بندی گراف با وزن خوشه. جدول (۸) روش پیشنهادی را با چهار روش جدید که دو روش نخست از دو نوع خوشه‌بندی k-means و fuzzy c-means به‌عنوان الگوریتم خوشه‌بندی پایه استفاده می‌کنند، مقایسه کرده است. نتایج مقایسه حاکی از برتری روش پیشنهادی نسبت به چهار روش دیگر است.

(جدول-۸): مقایسه عملکرد الگوریتم خوشه‌بندی پیشنهادی با CLWGC در صورت استفاده خوشه‌بندی k-means و fuzzy c-means به‌عنوان الگوریتم خوشه‌بندی پایه

(Table-8): Comparison of the performance of the proposed clustering algorithm with CLWGC using k-means and fuzzy c-means clustering as the basic clustering algorithm.

Dataset	CLWGC with k-means	CLWGC with fuzzy c-means	π_{GND}	RCESCC	PROPOSED
Semeion	66.81	66.43	68.59	67.43	**69.40**
MF	68.89	69.13	69.28	70.15	**72.10**
IS	67.05	67.26	62.71	67.22	**68.54**
FCT	23.31	23.38	26.19	30.23	**34.30**
MNIST	65.26	64.16	66.16	67.02	**68.10**
ODR	83.12	82.57	85.50	84.04	**85.62**
LS	63.28	61.42	65.60	63.15	**63.48**
ISOLET	76.38	75.51	77.71	77.19	**78.55**
USPS	65.68	65.56	67.12	67.20	**68.89**
LR	41.47	41.69	**47.07**	43.42	45.85
Average	62.12	61.71	63.59	63.70	**65.48**

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

۵- نتیجه‌گیری

الگوریتم خوشه‌بند فازی cmeans یک نوع الگوریتم خوشه‌بند است که به‌علت محاسبات کم و سرعت بالا بسیار پرکاربرد است. با این‌حال، به‌علت حساسیت بسیار بالا به انتخاب اولیه مراکز خوشه‌ها، شکل خوشه‌ها و توزیع داده‌ها، الگوریتم خوشه‌بند فازی cmeans یک روش خوشه‌بند ضعیف نیز محسوب می‌شود. در این مقاله، ما یک الگوریتم جدید خوشه‌بندی ترکیبی را با استفاده از چندین خوشه‌بندی به‌دست‌آمده از الگوریتم خوشه‌بند فازی cmeans بهبودیافته، ارایه کرده‌ایم. در این راستا به‌منظور افزایش تنوع در اجماع، الگوریتم خوشه‌بند فازی cmeans نیز بهبود داده شده است. در نتایج تجربی، نتایج الگوریتم پیشنهادی با نتایج چندین الگوریتم خوشه‌بندی ترکیبی به‌روز و نتایج سه الگوریتم خوشه‌بندی پایه مقاوم بر روی مجموعه‌داده‌های متنوعی مقایسه شده است. نتایج مقایسات نشان می‌دهند که الگوریتم پیشنهادی سایر الگوریتم‌های به‌روز را تحت شعاع قرار می‌دهد؛ علاوه‌براین، ما کارایی الگوریتم پیشنهادی را که برای مقابله با مجموعه‌داده‌های بزرگ، مناسب است، مورد بررسی قرار داده‌ایم. در بخش نتایج تجربی از آزمون آماری معناداری استفاده شد و نشان داده‌شده که روش پیشنهادی دارای بالاترین سطح معناداری نسبت به مجموعه داده‌های استاندارد است.

۶- مراجع 6- References

[1] J. Han, M. Kamber, *Data Mining: Concepts and Techniques,* Morgan Kaufmann, 2001.

[2] A.K. Jain, R.C. Dubes, *Algorithms for Clustering Data,* Prentice Hall, 1988.

[3] A.K. Jain, ''Data clustering: 50 years beyond Kmeans'', *Pattern Recognition Letters,* vol. 31, no. 8, pp. 651-666, 2010.

[4] J.B. MacQueen, "Some methods for classification and analysis of multivariate observations". *Proc. of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, vol. 1, pp. 281-297, 1967.

[5] A. Likas, M. Vlassis, J. Verbeek, "The global fc-means clustering algorithm", *Pattern Recognition,* vol. 35, no. 2, pp. 451-461, 2003.

[6] M. Ester, H. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", *In Evangelos Simoudis, Jiawei Han, Usama M.*

Transactions on Computational Biology and Bioinformatics, vol. 10, no. 3, pp. 657670, 2013.

[20] B. Fischer, J. Buhmann, "Bagging for path-based clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1411-1415, 2003.

[21] A. Topchy, B. Minaei-Bidgoli, A. Jain, "Adaptive clustering ensembles", *Proc. the 17th International Conference on Pattern Recognition,* 2004.

[22] Z. Zhou, W. Tang, "Clusterer ensemble", *Knowledge-Based Systems,* vol. 19, no. 1, pp. 77-83, 2006.

[23] Y. Hong, S. Kwong, H. Wang, Q. Ren, "Resampling-based selective clustering ensembles", *Pattern Recognition Letters,* vol. 41(9), pp. 2742-2756, 2009.

[24] X. Fern, C. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach", *Proc. International Conference on Machine Learning*, 2003.

[25] P. Zhou, L. Du, L. Shi, H. Wang et al., "Learning a robust consensus matrix for clustering ensemble via kullback-leibler divergence mini-mization", *Proc. the 25th International Joint Conference on Artificial Intelligence*, 2015.

[26] Z. Yu, L. Li, J. Liu et al., "Adaptive noise immune cluster ensemble using affinity propagation", *IEEE Transactions on Knowledge and Data Engineering,* vol. 27, no. 19, pp. 3176-3189, 2015.

[27] F. Gullo, C. Domeniconi, "Metacluster-based projective clustering ensembles", *Machine Learning,* vol. 98, no. 1-2, pp. 1-36, 2013.

[28] Y. Yang, J. Jiang, "Hybrid Sampling-Based Clustering Ensemble with Global and Local Constitutions", *Ieee Transactions on Neural Networks and Learning Systems,* vol. 27, no. 5, pp. 952-965, 2016.

[29] A. Fred, A. K. Jain, "Data clustering using evidence accumulation", *Proc. the 16th International Conference on Pattern Recognition,* , 2002, pp. 276-280.

[30] Y. Yang, K. Chen, "Temporal data clustering via weighted clustering ensemble with different representations", *IEEE Transactions on Knowl-edge and Data Engineering,* vol. 23, no. 2, pp. 307-320, 2011.

[31] N. Iam-On, T. Boongoen, S. Garrett, C. Price, "A link-based approach to the cluster ensemble problem", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 12, pp. 2396-2409, 2011.

[32] N. Iam-On, T. Boongoen, S. Garrett, C. Price, "A *Fayyad, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96),* AAAI Press, 1996, pp. 226-231.

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

[8] J. Shi, J. Malik, "Normalized cuts and image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, 2000.

[9] A.Y. Ng, M.I. Jordan, Y. Weiss, ''*On Spectral Clustering: Analysis and an Algorithm,* in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.)'', *Advances in Neural Information Processing Systems*, vol. 14, MIT Press, Cambridge, MA, 2002.

[10] A. Strehl, J. Ghosh, "Cluster ensembles: a knowledge reuse framework for combining multiple partitions", *Journal on Machine Learning Research,* vol. 3, pp. 583-617, 2002.

[11] A. Gionis, H. Mannila, P. Tsaparas, "Clustering aggregation, ACM Transactions on Knowledge Discovery from Data", vol. 1, no. 1, pp. 1-30, 2007.

[12] Z. Zhou, *Ensemble Methods: Foundations and Algorithms*, CRC Press, 2012.

[13] E. Gonzlez, J. Turmo, "Unsupervised ensemble minority clustering", *Machine Learning*, vol.98, pp. 217-268, 2015.

[14] N. Iam-On, T. Boongoen, "Comparative study of matrix refinement approaches for ensemble clustering", *Machine Learning*, vol. 98, pp. 269-300, 2015.

[15] A. Fred, A. Jain, "Combining multiple clusterings using evidence accumulation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp.835-850, 2005.

[16] L. Kuncheva, D. Vetrov, "Evaluation of stability of kmeans cluster ensembles with respect to random initialization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 17981808, 2006.

[17] X. Zhang, L. Jiao, F. Liu, L. Bo, M. Gong. "Spectral clustering ensemble applied to SAR image segmentation", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2126-2136, 2008.

[18] M. Law, A. Topchy, A. Jain, "Multiobjective data clustering", *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.

[19] Z. Yu, H. Chen, J. You, et al, "Hybrid fuzzy cluster ensemble framework for tumor clustering from bio-molecular data", *IEEE/ACM*

[DOI: 10.29252/jsdp.17.4.103]

Intelligence Review, vol. 52, pp. 1311–1340, Springer Nature B.V. 2018, https://doi.org/10.1007/s10462-018-9642-2.

[47] 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*, vol.49 , PP. 1724–1747, 2019. https://doi.org/10.1007/s10489-018-1332-x.

[48] 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.

[49] S. Guha, R. Rastogi, K. Shim, "Cure: an efficient clustering algorithm for large databases", *Proc. of the Conference on Management of Data (ACM SIGMOD)*, pp.73-84, 1998.

[50] P.H.A. Sneath, R.R. Sokal, *Numerical Taxonomy,* Freeman, San Francisco, London, 1973.

[51] B. King, "Step-wise clustering procedures", *Journal of the American State Association*, vol. 69, pp. 86-101, 1967.

[52] G. Karypis, E.-H.S. Han, V. Kumar, "Chameleon: ahierarchical clustering algorithm using dynamic modeling", *IEEE Computer*, vol. 32, no. 8, pp. 68-75, 1999.

[53] J.C. Bezdek, N. R. Pal, "Some new indexes of cluster validity", *IEEE Transactions on Systems Man and Cybernetics Part B*, vol. 28, no. 3, pp. 301-15, 1998.

[54] N.R. Pal, J.C. Bezdek, "On cluster validity for the fuzzy c-means model", *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370-379, 1995.

[55] UCI Machine Learning Repository, http://www.ics.uci.edu /mlearn /ML-Repository.html, 2016.

[56] T. S. A. V. W. T. Press, W. H. and B. P. Flannery, *Conditional Entropy and Mutual Information. Numerical Recipes: The Art of Scientific computing (3rd ed),* New York: Cambridge University Press, 2007.

[57] F. Rashidi, S. Nejatian, H. Parvin, V. Rezaie, "Diversity based cluster weighting in cluster ensemble: an information theory approach," *Artificial Intelligence Review*, vol. 52, pp.1341–1368, 2019.

[58] F. Najafi, H. Parvin, K. Mirzaie, S. Nejatian, V. Rezaie, "Dependability-based cluster weighting in clustering ensemble," Stat Anal Data Min: *The ASA Data Sci Journal*, vol. 13, pp. 151-164, 2020.

link-based cluster ensemble approach for categorical data clustering", *IEEE Transactions on Knowledge and Data Engineering,* vol. 24, no. 3, pp. 413-425, 2012.

[33] X. Fern, C. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning", *Proc. of the 21st International Conference on Machine Learning*, 2004.

[34] D. Huang, J. Lai, C. D. Wang, "Ensemble clustering using factor graph", *Pattern Recognition,* vol. 50, pp. 131-142, 2016.

[35] M. Selim, E. Ertunc, "Combining multiple clusterings using similarity graph", *Pattern Recognition,* vol. 44, no. 3, 694-703, 2011.

[36] C. Boulis, M. Ostendorf, "Combining multiple clustering systems", Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases, 2004.

[37] A. Topchy, B. Minaei-Bidgoli, A. Jain, "Adaptive clustering ensembles", *Proc. the 17th International Conference on Pattern Recognition*, 2004.

[38] P. Hore, L. O. Hall, B. Goldgo, "A scalable framework for cluster ensembles", *Pattern Recognition*, vol. 42, no. 5, 676-688, 2009.

[39] B. Long, Z. Zhang, P. S. Yu, "Combining multiple clusterings by soft correspondence", *Proc. the 4th IEEE International Conference on Data Mining*, 2005.

[40] 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.

[41] 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.

[42] H. Wang, H. Shan, A. Banerjee, "Bayesian cluster ensembles", *Statistical Analysis and Data Mining,* vol. 4, no. 1, pp. 54-70, 2011.

[43] Z. He, X. Xu, S. Deng, "A cluster ensemble method for clustering categorical data", *Information Fusion,* vol. 6, no. 2, pp. 143-151, 2005.

[44] N. Nguyen, R. Caruana, "Consensus Clusterings", *Proc. IEEE Intl Conf. Data Mining,* 2007, pp. 607-612.

[45] 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.

[46] S. Abbasi, S. Nejatian, H. Parvin, V. Rezaie &K. Bagherifard, "Clustering ensemble selection considering quality and diversity," *Artificial*

[DOI: 10.29252/jsdp.17.4.103]

فاطمه نجفی تحصیلات خود را در مقطع کارشناسی ارشد در رشته مهندسی رایانه در دانشگاه علوم و تحقیقات و مقطع دکترای تخصصی را در رشته مهندسی رایانه در دانشگاه آزاد اسلامی واحد میبد به پایان رساند. وی هم‌اکنون عضو هیأت علمی دانشگاه آزاد اسلامی واحد ایذه است. زمینه‌های پژوهشی ایشان، طبقه‌بندی و خوشه‌بندی داده‌ها است.

نشانی رایانامه ایشان عبارت است از:

najafi.un@gmail.com

حمید پروین تحصیلات خود را در مقطع کارشناسی در دانشگاه چمران اهواز به پایان رساند. ایشان مدرک کارشناسی ارشد و دکترا را از دانشگاه علم و صنعت دریافت کردند و پس از آن به عضویت هیأت علمی دانشگاه آزاد واحد نورآباد ممسنی در آمدند. وی هم‌اکنون در چندین واحد دانشگاهی در رشته رایانه مشغول تدریس است. زمینهٔ پژوهشی وی مباحثی نظیر الگوریتم‌های بهینه‌سازی، طبقه‌بندی و خوشه‌بندی داده‌ها است.

نشانی رایانامه ایشان عبارت است از:

parvin@iust.ac.ir

کمال میرزائی مدرک کارشناسی خود را در رشته مهندسی رایانه در سال ۱۳۸۰ از دانشگاه علم و صنعت تهران و مدرک کارشناسی ارشد را در رشته مهندسی کامپیوتر در سال ۱۳۸۳ از دانشگاه اصفهان دریافت کرد. ایشان در سال ۱۳۹۰ موفق به کسب درجه دکترا در رشته مهندسی رایانه از دانشگاه علوم و تحقیقات تهران شد. زمینه‌های پژوهشی مورد علاقه ایشان، علوم شناختی، الگوریتم‌های تکاملی، شبکه‌های پیچیده پویا، محاسبات نرم، داده‌کاوی پزشکی، پردازش تصویر و شناسایی الگو بوده و در حال حاضر عضو هیأت علمی با مرتبه استادیار در دانشگاه آزاد اسلامی واحد میبد است.

نشانی رایانامه ایشان عبارت است از:

k.mirzaie@maybodiau.ac.ir

صمد نجاتیان تحصیلات خود را در مقطع کارشناسی در رشته مهندسی برق (الکترونیک) دانشگاه سیستان و بلوچستان در سال ۱۳۸۲ به پایان رساند. ایشان مدرک کارشناسی ارشد خود را در رشته برق (مخابرات)، از دانشگاه مشهد در سال ۱۳۸۶ و مدرک دکترای تخصصی خود را در رشته برق (مخابرات)، در سال ۱۳۹۳ از دانشگاه UMT مالزی اخذ کردند. وی هم‌اکنون عضویت هیأت علمی دانشگاه آزاد واحد یاسوج است.

نشانی رایانامه ایشان عبارت است از:

nejatian@iauyasooj.ac.ir

سیده وحیده رضایی دارای مدرک تحصیلی در مقطع دکترای تخصصی رشته ریاضیات هستند. وی هم‌اکنون عضو هیأت علمی دانشگاه آزاد اسلامی واحد یاسوج است. زمینه‌های پژوهشی ایشان بهینه‌سازی ریاضی، متن‌کاوی، پردازش سیگنال، داده‌کاوی و خوشه‌بندی داده‌ها است.

نشانی رایانامه ایشان عبارت است از:

v.rezaie@iauyasooj.ac.ir

[DOI: 10.29252/jsdp.17.4.103]