استفاده از الگوریتم آبکاری فلزات برای بهبود اجماع خوشه‌بندی

سیده فروزان رشیدی[۱]، صمد نجاتیان*[۲]، حمید پروین[۳]، وحیده رضایی[۴]، کرم‌الله باقری فرد[۵]

[۱و۵] گروه کامپیوتر، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

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

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

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

چکیده

خوشه‌بندی داده‌ها یکی از وظایف اصلی داده‌کاوی است که وظیفهٔ کاوش الگوهای پنهان را در داده‌های بدون برچسب بر عهده دارد. به‌خاطر پیچیدگی مسئله و ضعف روش‌های خوشه‌بندی پایه، امروزه بیشتر مطالعات به سمت روش‌های اجماع خوشه‌بندی هدایت شده است. اگرچه برای بیشتر مجموعه داده‌ها، الگوریتم‌های خوشه‌بندی منفردی وجود دارد که نتایج قابل‌قبولی به دست می‌دهند، امّا توانایی یک الگوریتم خوشه‌بندی منفرد محدود است. در واقع، هدف اصلی اجماع خوشه‌بندی جستجوی نتایج بهتر و پایدارتر، با استفاده از ترکیب اطّلاعات و نتایج حاصل از چندین خوشه‌بندی اولیه است. در این مقاله، روشی مبتنی بر اجماع خوشه‌بندی پیشنهاد خواهد شد که مانند بیشتر روش‌های انباشت شواهد دارای دو گام است: ۱- ساختن ماتریس مشارکت هم‌زمان و ۲- تعیین افرازهای نهایی از ماتریس مشارکت پیشنهادی. در روش پیشنهادی، برای ساخت ماتریس مشارکت هم‌زمان، علاوه‌بر هم‌خوشه بودن نمونه‌ها، از برخی اطّلاعات دیگر هم استفاده خواهد شد. این اطّلاعات می‌توانند مربوط به میزان شباهت نمونه‌ها، اندازهٔ خوشه‌های اولیه، میزان پایداری خوشه‌های اولیّه و ... باشد. در این مقاله، مسئلهٔ خوشه‌بندی به‌صورت یک مسئلهٔ بهینه‌سازی صریح توسط الگوی آمیختهٔ گوسی تعریف، و با استفاده از الگوریتم آبکاری فلزات حل می‌شود. همچنین، روشی تکاملی مبتنی بر آبکاری فلزات برای تعیین افراز نهایی از ماتریس مشارکت هم‌زمانِ پیشنهادی ارایه خواهد شد. مهم‌ترین بخش روش تکاملی، تعیین تابع هدفی است که تضمین کند افراز نهایی از کیفیّت بالایی برخوردار است. نتایج تجربی نشان می‌دهد روش پیشنهادی از نظر معیارهای گوناگون ارزیابیِ کیفیّت خوشه‌بندی از سایر روش‌های مشابه بهتر است.

کلیدواژه‌ها: اجماع خوشه‌بندی، الگوی آمیختهٔ گوسی، الگوریتم آبکاری فلزات، ماتریس مشارکت هم‌زمان، پایداری، تابع هدف.

Using Simulated Annealing Algorithm to Improve Ensemble Clustering

Seyedeh Foroozan Rashidi[1], Samad Nejatian*[2], Hamid Parvin[3], Vahideh Rezaie[4], Karamolah Bagherifard[5]

[1,5] Department of Computer Engineering, Yasuj Branch, Islamic Azad University, Yasuj, Iran.
[2]Department of Electrical Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran.
Samad.nej.2007@gmail.com
[3] Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran
[4] Department of Mathematic, Yasooj Branch, Islamic Azad University, Yasooj, Iran

Abstract

Data clustering is one of the data mining main tasks, which is responsible to explor hidden patterns in unlabeled data. Due to the complexity of the problem and the weakness of the basic clustering methods, today most studies are directed towards clustering ensemble methods. Although for most datasets, there are individual clustering algorithms that provide acceptable results, the ability of a single clustering algorithm is limited. In fact, the main purpose of clustering ensemble is to search for better and more stable results, using

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۰/۱/۵ • تاریخ پذیرش: ۱۴۰۲/۳/۱۲ • تاریخ انتشار: ۱۴۰۲/۵/۲۰ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.20.1.99]

the combination of information and results obtained from several initial clustering. In this paper, a clustering ensemble-based method will be proposed, which, like most evidence accumulation methods, has two steps: 1- building a simultaneous participation matrix and 2- determining the final output from the proposed participation matrix. In the proposed method, some other information will be used in addition to the samples clustering to construct the simultaneous participation matrix. This information can be related to the degree of similarity of the samples, the size of the initial clusters, the stability degree of the initial clusters, etc. In this paper, the clustering problem is defined as an explicit optimization problem by the mixed Gaussian model and is solved using the simulated annealing algorithm. Also, an evolutionary method based on simulated annealing will be presented to determine the final output from the proposed simultaneous participation matrix. The most important part of the evolutionary method is to determine the objective function that guarantees the final output will be of high quality. The proposed method uses a new method to determine the initial state in the problem. This method, according to the labels obtained from the initial clustering of data, determines the possible distributions in such a way that the clustering results are reflected in it. The amount of reflection can be controlled through a parameter in the proposed method. It uses a new and innovative mechanism to go to the next state. In this method, which can be controlled through several parameters, by focusing on some regions of the correlation matrix, the probability distributions of samples belonging to different clusters are changed in a controlled and soft manner and a new probability distribution is produced, but close to the previous one. The degree of closeness between the previous and the new state is controlled through several parameters that determine which regions of the correlation matrix should be focused on and how many changes are in the current state. In experimental tests, the proposed method was compared with four individual clustering methods and two combined clustering methods. Experimental results show that this proposed method generally produces higher quality classifications than other methods. Part of the experimental results show that the proposed method is able to identify clusters with normal distribution better than other methods. Overall, the experimental results show that the proposed method is better than other similar methods in terms of different clustering quality evaluation criteria.

Keywords: Clustering ensemble, Gaussian mixture model, simulated annealing algorithm, simultaneous participation matrix, stability, objective function

مقدمه

از جمله مسایل مهمی که در حوزهٔ داده‌کاوی و شناسایی الگو وجود دارد، خوشه‌بندی است. خوشه‌بندی به نوعی یادگیری بدون ناظر گفته می‌شود که در آن سعی می‌شود الگوها به چند دسته تقسیم شوند؛ به طوری که اعضای هر دسته مشابه یکدیگر باشند و با اعضای دیگر دسته‌ها بیشترین تفاوت را داشته باشند[۲ و ۱]. دست کم پنج دلیل اصلی برای اهمیّت خوشه‌بندی وجود دارد:

۱) جمع‌آوری و برچسب‌گذاری یک مجموعهٔ بزرگ از الگوهای نمونه می‌تواند بسیار باارزش باشد.

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

۳) در خیلی از کاربردها ویژگی‌های الگوها می‌توانند به آهستگی با زمان تغییر کنند، مثل رده‌بندی[۱] خودکار موادّ غذایی با تغییر فصل. اگر این تغییرات بتوانند با یک رده‌بندی‌کننده[۲] به‌صورت بدون ناظر[۳] رهگیری[۴] شود، عملکرد بهتری می‌تواند به دست آید.

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

۵) با خوشه‌بندی می‌توانیم بینشی از طبیعت و ساختار داده به دست آوریم که می‌تواند برای ما با ارزش باشد. کشف زیررده‌های[۵] مجزّا یا شباهت‌های بین الگوها ممکن است به‌طور چشمگیری در روش طرّاحی رده‌بندی‌کننده به ما پیشنهاد ارایه کند.

الگوریتم‌های گوناگونی برای خوشه‌بندی وجود دارد. اگرچه برای بیشتر مجموعه داده‌ها، الگوریتم‌های خوشه‌بندی منفردی وجود دارد که نتایج قابل‌قبولی به‌دست می‌دهند، امّا توانایی یک این الگوریتم‌ها محدود است [۳]. بنابراین روش‌هایی به نام اجماع خوشه‌بندی[۶] (خوشه‌بندی ترکیبی) وجود دارند که الگوها را بسیار بهتر از مناسبترین الگوریتم منفرد خوشه‌بندی می‌کنند [۴]. در اجماع خوشه‌بندی، نتایج چند خوشه‌بندی با یکدیگر ترکیب می‌شود و از برآیند آنها افرازهایی جدید به دست می‌آید که به‌طور عموم درک بهتری از الگوها فراهم می‌کند.

[۱] Classification
[۲] Classifier
[۳] Unsupervised
[۴] Tracking
[۵] SubClass
[۶] Ensembling clustering

[DOI: 10.61186/jsdp.20.1.99]

روش‌های اجماع خوشه‌بندی افرازی توافقی را روی یک مجموعه از نمونه‌ها تعریف می‌کنند که برآمده از ترکیب نتایج تعدادی الگوریتم خوشه‌بندی پایه است. رهیافت انباشت شواهد خوشه‌بندی[۷] روی نتایج خوشه‌بندی‌های پایه، ماتریسی به نام ماتریس هم-مشارکت[۸] می‌سازد. این رهیافت از برخورد به مشکل ارتباط برچسب‌ها[۹] اجتناب می‌کند، زیرا برای ساختن ماتریس هم‌مشارکت نیازی به بازبرچسب‌گذاری[۱۰] ندارد.

در این مقاله، روشی پیشنهاد خواهد شد که مانند بیشتر روش‌های انباشت شواهد، دارای دو گام است: ۱- ساختن ماتریس مشارکت هم‌زمان و ۲- تعیین افرازهای نهایی از ماتریس مشارکت پیشنهادی.

در روش پیشنهادی، برای ساخت ماتریس مشارکت هم‌زمان، علاوه‌بر هم‌خوشه بودن نمونه‌ها از برخی اطّلاعات دیگر هم استفاده خواهد شد. این اطّلاعات می‌توانند مربوط به میزان شباهت نمونه‌ها، اندازه خوشه‌های اولیّه، میزان پایداری خوشه‌های اولیّه و ... باشد. همچنین، روشی تکاملی برای تعیین افراز نهایی از ماتریس مشارکت هم‌زمان پیشنهادی ارایه خواهد شد. مهم‌ترین بخش روش تکاملی، تعیین تابع هدفی است که تضمین کند افراز نهایی از کیفیّت بالایی برخوردار است. یکی از مهم‌ترین نوآوری‌های روش پیشنهادی، معرّفی یک روش نوین برای تعیین حالت‌های پسین است. نوآوری روش پیشنهادی به‌طور عمده مربوط به نحوهٔ تعیین حالت آغازین و عملگر حرکت است که در بخش سه با تفصیل بیشتر در مورد آن‌ها صحبت خواهد شد.

در ادامهٔ مقاله، در فصل دو زمینه‌های مرتبط با موضوع مقاله بررسی خواهند شد. این زمینه‌ها شامل خوشه‌بندی و الگوی آمیختهٔ گاوسی می‌شود. بخش سه، به شرح روش پیشنهادی اختصاص دارد. در بخش چهار، آزمایش‌ها و نتایج تجربی ارائه شده است. بخش پنج نیز به نتیجه‌گیری و معرفی کارهای آینده اختصاص دارد.

۲- ادبیات تحقیق و کارهای مرتبط پیشین

این بخش، به مفاهیم پایه خوشه‌بندی اختصاص دارد. باتوجه‌به گستردگی موضوع، تنها بر مطالبی تأکید شده است که ارتباط بیشتری با موضوع مقاله دارند.

۲-۱- خوشه‌بندی منفرد و اجماع خوشه‌بندی

تجزیه و تحلیل خوشه یک ابزار اساسی در زمینه‌های یادگیری ماشین، تشخیص الگو و داده‌کاوی است [۳]، که در آن تجزیه و تحلیل کارآمد، تجسّم و تفسیر داده‌ها ضروری است. این امر به‌طور عمده، به‌دلیل رشد مداوم حجم داده‌ها است. خوشه‌بندی، گروه‌بندی مجموعه‌ای از اشیاست، به‌طوری‌که اشیا در یک گروه (به نام خوشه) در مقایسه با دیگر گروه‌ها (خوشه‌ها) مشابه‌تر هستند [۴]. به‌طور سنّتی، خوشه‌بندی داده‌ها یک کار یادگیری بدون‌نظارت است؛ به این معنی که تعداد خوشه‌ها ناشناخته است و هیچ‌یک از نقاط دادهٔ ورودی، برچسب-گذاری نشده‌اند. از کاربردهای خوشه‌بندی می‌توان تقسیم‌بندی تصویر [۵]، متن‌کاوی [۶]، تجزیه و تحلیل بیان ژن [۷]، تحلیل آلودگی هوا [۸] و تشخیص خطا [۹] را نام برد، که در این‌جا فقط این چند مورد ذکر شده است.

روش‌های اجماع خوشه‌بندی افرازی توافقی را روی یک مجموعه از نمونه‌ها تعریف می‌کنند که برآمده از ترکیب نتایج تعدادی الگوریتم خوشه‌بندی پایه است. رهیافت انباشت شواهد خوشه‌بندی[۱۱] روی نتایج خوشه‌بندی‌های پایه، ماتریسی به نام ماتریس هم-مشارکت[۱۲] می‌سازد. این رهیافت از برخورد به مشکل ارتباط برچسب‌ها[۱۳] اجتناب می‌کند؛ چراکه برای ساختن ماتریس هم-مشارکت نیازی به بازبرچسب‌گذاری[۱۴] ندارد.

هر یک از الگوریتم‌های خوشه‌بندی، باتوجّه‌به‌اینکه بر روی جنبه‌های متفاوتی از داده‌ها تأکید دارد، داده‌ها را به صورت‌های متفاوتی رده‌بندی می‌کند. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با استفاده از ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کند. در واقع، هدف اصلی اجماع خوشه‌بندی جستجوی بهترین خوشه‌ها با استفاده از ترکیب نتایج الگوریتم‌های دیگر است.

تحلیل داده پژوهشی، و به‌طور خاص، خوشه‌بندی داده‌ها می‌تواند به‌طور چشمگیری از ترکیب چندین افراز داده سود ببرد. اجماع خوشه‌بندی می‌تواند جواب‌های بهتری از نظر استحکام[۱۵]، نوآوری [۱۶]، پایداری[۱۷] و انعطاف پذیری[۱۸] ارایه دهد. به‌علاوه، قدرت موازی‌سازی آنها یک

[۷] Evidence Accumulation Clustering
[۸] Co-association
[۹] Label correspondence
[۱۰] Relabeling

[۱۱] Evidence Accumulation Clustering
[۱۲] Co-association
[۱۳] Label correspondence
[۱۴] Relabeling
[۱۵] Robustness
[۱۶] Novelty
[۱۷] Stability
[۱۸] Felixibility

[DOI: 10.61186/jsdp.20.1.99]

انطباق طبیعی با نیاز داده‌کاوی توزیع‌شده دارد. هنوز، به دست آوردن پایداری در ترکیب خوشه‌ها با دشواری روبه‌رو است. ترکیب خوشه‌بندی‌ها کار مشکل‌تری از ترکیب رده‌بندی‌های با ناظر است. در غیاب دادهٔ آموزشی برچسب‌دار، ما با مشکل تناظر بین برچسب‌های خوشه در افرازهای مختلف از یک ترکیب مواجه هستیم. مطالعات اخیر نشان می‌دهند که اجماع خوشه‌بندی، می‌تواند خارج از وضعیّت‌های نوع رأی‌گیری، با استفاده از روش‌های مبتنی بر گراف، آماری یا نظریّهٔ اطّلاعات، بدون حلّ دقیق مشکل تناظر برچسب‌ها انجام شود. همچنین، به توابع توافقی تجربی دیگر توجه شده بود. اگرچه، مسئلهٔ خوشه‌بندی توافقی به‌عنوان $NP-complete$ شناخته شده، روش‌های زیادی برای حل آن پیشنهاد شده است.

۲-۲- الگوریتم خوشه‌بندی منفرد

به‌طور کلی، الگوریتم‌های خوشه‌بندی را می‌توان به دو دسته کلی تقسیم کرد [۱۰ و ۱۱]:

۱- الگوریتم‌های سلسله‌مراتبی

۲- الگوریتم‌های افرازبندی

الگوریتم‌های سلسله‌مراتبی، یک روال برای تبدیل یک ماتریس مجاورت به یک دنباله از افرازهای تو در تو، به صورت یک درخت است. در این روش‌ها، به‌طور مستقیم با داده‌ها سروکار داریم و از روابط بین آنها برای به دست آوردن خوشه‌ها استفاده می‌کنیم.

در نقطهٔ مقابل الگوریتم‌های سلسله‌مراتبی، الگوریتم‌های افرازبندی قرار دارند. هدف این الگوریتم‌ها، تقسیم داده‌ها در خوشه‌ها، به گونه‌ای است که داده‌های درون یک خوشه دارای بیشترین شباهت را به همدیگر باشند؛ و در عین حال، بیشترین فاصله و اختلاف را با داده‌های خوشه‌های دیگر داشته باشند.

مهم‌ترین روش‌های خوشه‌بندی سلسله‌مراتبی که در این بخش بررسی می‌شوند، عبارتند از[۱۰ و ۱۱]:

۱- اتّصال منفرد[۱۹]

۲- اتّصال کامل[۲۰]

۳- اتّصال میانگین[۲۱]

هر سه روش به صورت گام به گام انجام می‌شوند و در هر گام دو خوشه‌ای که کمترین فاصله را با یکدیگر دارند، با هم ادغام می‌شوند. تفاوت بین روش‌ها ریشه در تفاوت بین نحوهٔ محاسبه فاصله در آنها دارد. در اتّصال منفرد، فاصلهٔ بین دو خوشه برابر با فاصلهٔ بین نزدیک‌ترین دو نمونه از دو خوشهٔ متفاوت است. در اتّصال کامل، فاصله بین دو خوشه به صورت فاصلهٔ بین دورترین دو نمونه از دو خوشه محاسبه می‌شود. همچنین، فاصلهٔ بین دو خوشه در اتّصال میانگین به‌صورت میانگین فاصله بین همهٔ جفت‌نمونه‌های ممکن محاسبه می‌شود، که هر جفت شامل یک نمونه از خوشه اول و یک نمونه از خوشهٔ دیگر است.

مهم‌ترین روش‌های خوشه‌بندی افرازبندی عبارتند از[۱۰ و ۱۱]:

۱- الگوریتم Forgy

۲- الگوریتم $K-means$

۳- الگوریتم Isodata

۲-۳- الگوریتم الگوی آمیختهٔ گاوسی

هر الگوی آمیخته[۲۲] الگوی آماری است که جمعیّتی از مشاهده‌ها را به صورت تعدادی زیرجمعیّت بازنمایی می‌کند. پس از ایجاد یک الگوی آمیخته روی یک مجموعه از مشاهده‌ها دیگر با تک‌تک مشاهده‌ها سر و کار نداریم، بلکه تعدادی زیرجمعیت داریم که هرکدام نمایندهٔ تعدادی از مشاهده‌های اولیّه است. شاید توزیع آمیخته نام مناسب‌تری باشد؛ چراکه به‌طور معمول، برای بازنمایی هر زیرجمعیت از یک توزیع آماری استفاده می‌شود. در واقع، هر توزیع، ویژگی‌های یک زیرجمعیت را بازنمایی می‌کند، بدون آنکه به توصیف جزءبه‌جزء جمعیت و تک‌تک مشاهدات بپردازد.

الگوی آمیختهٔ گاوسی[۲۳] [۷] نوع خاصی از الگوهای آمیخته است که در آن هر زیر جمعیت با یک توزیع گاوسی بازنمایی می‌شود. به‌طور کلی تابع توزیع چگالی احتمال گاوسی یک متغیّر تصادفی نرده‌ای[۲۴] به‌صورت رابطه (۱) است[۷]:

$$(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-0.5(\frac{x-\mu}{\sigma})^2} = N(\mu, \sigma^2) \quad (1)$$

که در رابطه (۱)، x بردار تصادفی، μ بردار میانگین متغیّرهای تصادفی و σ انحراف معیار متغیّرهای تصادفی است.

شکل توسعه‌یافتهٔ این رابطه برای بردار تصادفی مانند x

[۱۹] Single Linkage (SL)

[۲۰] Complete Linkage (CL)

[۲۱] Average Linkage (AL)

[۲۲] Mixture Model

[۲۳] Gaussian Mixture Model

[۲۴] Scalar

[DOI: 10.61186/jsdp.20.1.99]

به‌صورت رابطه (۲) است [۷]:

$$p(x) = \frac{1}{D/2\sqrt{2\pi}|\Sigma|^{1/2}} e^{-0.5(x-\mu)^T \Sigma^{-1}(x-\mu)} \quad (۲)$$

که در رابطه (۲)، x بردار تصادفی، μ بردار میانگین متغیّرهای تصادفی و Σ ماتریس کوواریانس متغیّرهای تصادفی است.

در بسیاری از روش‌های اجماع خوشه‌بندی از الگوریتم $K-Means$ به‌عنوان روش خوشه‌بندی پایه استفاده می‌شود. در الگوریتم $K-Means$ هر خوشه با مرکز آن نمایندگی می‌شود که میانگین نمونه‌های عضو آن خوشه و یا همان مرکز ثقل خوشه است. این روش توصیف خوشه دارای یک نقص بنیادین است که مفروض بر کروی بودن خوشه‌هاست. در واقع، کشیدگی خوشه‌ها در جهات مختلف نادیده گرفته می‌شود و فرض می‌شود نمونه‌ها پیرامون مرکز خوشه با فواصل یک‌نواخت پراکنده شده‌اند. شکل توسعه‌یافتهٔ خوشه‌های کروی، خوشه‌های بیضوی است. در خوشه‌های بیضوی، آنها می‌توانند در جهات مختلف فضای ویژگی کشیدگی داشته باشند. الگوی آمیختهٔ گاوسی که بررسی شد، فرض را بر بیضوی بودن خوشه‌ها می‌گیرد و می‌توان آن را شکل توسعه‌یافته‌ای از الگوریتم $K-Means$ دانست. در الگوی آمیختهٔ گاوسی یا GMM، هر خوشه با یک توزیع nبعدی طبیعی نمایندگی می‌شود. این الگو از آن جهت آمیخته نامیده می‌شود که از ترکیب چند توزیع طبیعی تشکیل شده است، که هرکدام یک خوشه را نمایندگی می‌کنند. برچسب خوشهٔ هر نمونه با توجه به مقدار تابع چگالی احتمال هریک از توزیع‌ها که در نقطه‌ای نمونه در آن واقع شده، تعیین می‌شود. در واقع، یک نمونه متعلّق به خوشه‌ای است که در توزیع طبیعی مربوط به آن خوشه بیشترین مقدار چگالی احتمال را نسبت به سایر خوشه‌ها دارد.

یکی از مهم‌ترین ملاحظات در خوشه‌بندی به کمک الگوی آمیختهٔ گاوسی، بهره‌گیری از یک روش مناسب برای بهینه‌سازی الگو است. در $K-Means$ از یک روش تکراری برای بهینه کردن یا آموزش الگو استفاده می‌شود که مبتنی بر تعیین مرکز خوشه‌ها در هر تکرار بر مبنای نمونه‌های نسبت‌داده‌شده به آن خوشه است. در الگوی آمیختهٔ گاوسی، علاوه‌بر به‌روزرسانی مراکز خوشه، باید ماتریس کواریانس و وزن هریک از خوشه‌ها را نیز به‌روزرسانی کند. ماتریس کواریانس کشیدگی خوشه در راستای ابعاد مختلف فضای ویژگی را مشخّص می‌کند.

یکی از رایج‌ترین روش‌های آموزش الگوی آمیختهٔ گاوسی استفاده از الگوریتم بیشینه‌سازی امید ریاضی[۲۵] یا به اختصار EM [۷] است. در این الگوریتم، نخست به شاخص‌های مقداردهی اولیّه می‌شوند. به‌این‌ترتیب، می‌توان بر مبنای شاخص‌های اولیّه (که مقادیر اولیّه به بردارهای میانگین، ماتریس‌های کواریانس و وزن خوشه‌هاست) نمونه‌ها را به خوشه‌های نسبت داد. سپس، با توجه به خوشه‌های شکل گرفته، شاخص‌ها به‌روزرسانی می‌شود. این فرآیند دوگانهٔ تعیین خوشه‌ها و به‌روزرسانی شاخص‌ها بارها و بارها تکرار می‌شود تا الگوریتم همگرا شود.

در الگوریتم EM، مقداردهی اولیّه به شاخص‌ها به کمک الگوریتم خوشه‌بندی $K-Means$ انجام می‌گیرد. الگوی آمیختهٔ گوسی را با Θ نشان می‌دهند که منظور از Θ مجموعهٔ همهٔ شاخص‌هایی است که الگو را توصیف می‌کنند. آموزش GMM عبارت است از تخمین Θ یا همان شاخص‌های الگو، یعنی $\Theta = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{M}$ (در این روابط w_k وزن خوشه kام، μ_k بردار میانگین kام، Σ_k ماتریس کواریانس خوشه‌ی kام) به‌وسیلهٔ نمونه‌های آموزش $X = \{x_1, ..., x_T\}$. در الگوریتم EM هدف بیشینه کردن لگاریتم احتمال تعلّق نمونه‌ها به خوشه‌هاست. بیشینه کردن لگاریتم احتمال منجر به روابطی می‌شود که مشخّص می‌کند شاخص‌ها چگونه باید با استفاده از نمونه به‌روزرسانی شوند. فرمول الگوریتم EM برای به‌روزرسانی بردارهای میانگین به‌صورت رابطه (۳) است[۷]:

$$\mu_k = \frac{\sum_{t=1}^{T} x_t p(k \mid x_t; \Theta)}{\sum_{t=1}^{T} p(k \mid x_t; \Theta)} \quad (۳)$$

فرمول به‌روزرسانی ماتریس کواریانس هم به‌صورت رابطه (۴) است[۷]:

$$\Sigma_k = \frac{\sum_{t=1}^{T} p(k \mid x_t; \Theta)(x_t - \mu_t)(x_t - \mu_t)^T}{\sum_{t=1}^{T} p(k \mid x_t; \Theta)} \quad (۴)$$

همچنین، برای به‌روزرسانی وزن هر خوشه از رابطه (۵) استفاده می‌شود[۷]:

$$w_k = \frac{1}{T} \sum_{t=1}^{T} p(k \mid x_t; \Theta) \quad (۵)$$

[۲۵] Expectation-Maximazation

[DOI: 10.61186/jsdp.20.1.99]

در این روابط w_k وزن خوشهٔ kام، μ_k بردار میانگین kام، Σ_k ماتریس کواریانس خوشه‌ی kام، T تعداد نمونه‌ها و x_t بردار ویژگی نمونه‌ی tام است. محاسبهٔ احتمال $p(k|x_t;\Theta)$ با استفاده از رابطه (۶) انجام می‌شود[۷]:

$$p(k \mid x_t;\Theta) = \frac{w_k N(x_t;\mu_k,\Sigma_k)}{\sum_{i=1}^{M} w_i N(x_t;\mu_i,\Sigma_i)} \qquad (۶)$$

که در رابطه (۶)، M تعداد خوشه‌ها و $N(x_t;\mu_k,\Sigma_k)$ مقدار تابع چگالی توزیع طبیعی با میانگین و واریانس μ_k و Σ_k در نقطه x_t است.

۲-۴- اجماع خوشه‌بندی

هر یک از الگوریتم‌های خوشه‌بندی، باتوجّه‌به‌این‌که بر روی جنبه‌های متفاوتی از داده‌ها تأکید دارد، داده‌ها را به صورت‌های متفاوتی رده‌بندی می‌کند. به‌همین‌دلیل، نیازمند روش‌هایی هستیم که بتواند با استفاده از ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری تولید کند. در واقع، هدف اصلی اجماع خوشه‌بندی جستجوی بهترین خوشه‌ها با استفاده از ترکیب نتایج الگوریتم‌های دیگر است[۴].

اجماع خوشه‌بندی‌ها کار مشکل‌تری از ترکیب رده‌بندی‌های با ناظر است. در غیاب دادهٔ آموزشی برچسب‌دار، ما با مشکل تناظر بین برچسب‌های خوشه در افرازهای مختلف از یک ترکیب مواجه هستیم. مطالعات اخیر نشان می‌دهند که اجماع خوشه‌بندی، می‌تواند خارج از وضعیّت‌های نوع رأی‌گیری، با استفاده از روش‌های مبتنی بر گراف، آماری یا تئوری اطّلاعات، بدون حل دقیق مشکل تناظر برچسب‌ها انجام شود [۵]. همچنین، به توابع توافقی تجربی دیگری توجه شده بود. اگرچه، مسئلهٔ خوشه‌بندی توافقی به‌عنوان NP – complete شناخته شده، روش‌های زیادی برای حل آن پیشنهاد شده است[۵].

به‌طور خلاصه، اجماع خوشه‌بندی شامل دو مرحلهٔ اصلی زیر است:

- تولید نتایج متفاوت از خوشه‌بندی‌ها، به‌عنوان نتایج خوشه‌بندی اولیه بر اساس اعمال روش‌های مختلف که این مرحله را، مرحلهٔ ایجاد تنوع یا پراکندگی[۲۶] می‌نامند.
- ترکیب نتایج به‌دست‌آمده از خوشه‌بندی‌های متفاوت اولیّه برای تولید خوشه نهایی؛ که این کار توسط تابع توافقی[۲۷] (الگوریتم ترکیب‌کننده) انجام می‌شود.

۲-۵- الگوریتم تابکاری شبیه‌سازی‌شده

الگوریتم‌های بهینه‌سازی کاربردهای متنوعی دارند [۶ و ۷]. الگوریتم تابکاری فلزات (تابکاری شبیه‌سازی‌شده) (SA)[۲۸] [۸]، یک الگوریتم بهینه‌سازی فراابتکاری ساده و اثربخش در حل مسائل بهینه‌سازی است. روش تابکاری تدریجی، به‌وسیلهٔ متالورژیست‌ها برای رسیدن به حالتی که در آن مادهٔ جامد، به‌خوبی مرتّب و انرژی آن کمینه شده باشد، استفاده می‌شود. این روش شامل قرار دادن مادّه در دمای بالا و سپس کاهش تدریجی این دماست.

در روش SA، هر نقطهٔ s در فضای جستجوی مشابه یک حالت از یک سیستم فیزیکی است، و تابع $E(s)$ که باید کمینه شود، مشابه با انرژی داخلی سیستم در آن حالت است. در این روش، هدف انتقال از حالت اولیّهٔ دلخواه، به حالتی است که سیستم در آن کمترین انرژی را داشته باشد.

برای حل یک مسألهٔ بهینه‌سازی، الگوریتم SA، نخست، از یک جواب اولیّه آغاز، و سپس، در یک حلقهٔ تکرار به جواب‌های همسایه حرکت می‌کند. اگر جواب همسایه بهتر از جواب فعلی باشد، الگوریتم آن را به‌عنوان جواب کنونی قرار می‌دهد (به آن حرکت می‌کند)، درغیراین‌صورت، الگوریتم آن جواب را با احتمال $e^{-\Delta E/T}$ به‌عنوان جواب فعلی می‌پذیرد. در این رابطه، ΔE تفاوت بین تابع هدف جواب فعلی و جواب همسایه است و T یک شاخص به نام دما. در الگوریتم دما به‌آرامی کاهش داده می‌شود. در گام‌های اولیّه دما خیلی بالا قرار داده می‌شود تا احتمال بیشتری برای پذیرش جواب‌های بدتر وجود داشته باشد. با کاهش تدریجی دما، در گام‌های پایانی احتمال کمتری برای پذیرش جواب‌های بدتر وجود خواهد داشت و بنابراین الگوریتم به‌سمتِ یک جواب خوب همگرا می‌شود. الگوریتم SA درصورتی‌که دما در آن به اندازهٔ کافی به‌آهستگی تغییر کند، می‌تواند جواب‌های بسیار خوبی پیدا کند.

در هر مرحله، الگوریتم تابکاری شبیه‌سازی‌شده، چند حالت را در همسایگی حالت کنونی s در نظر می‌گیرد، و به‌طور احتمالی تصمیم می‌گیرد که سیستم را از حالت s منتقل کند یا در همین حالت باقی بماند. این احتمالات در نهایت سیستم را به حالت با انرژی کمتر میل می‌دهد.

[۲۶] Diversity

[۲۷] Consensus Function

[۲۸] Simulated Annealing

[DOI: 10.61186/jsdp.20.1.99]

همسایه‌های یک حالت، حالت‌های جدیدی از مسئله هستند که با تغییر در حالت کنونی و با توجه به روشی ازپیش‌تعیین‌شده ایجاد می‌شوند. برای مثال، در مسئلهٔ فروشندهٔ دوره‌گرد، هر حالت به‌طور کلی یک جایگشت خاص از شهرهایی است که باید ملاقات شوند. همسایهٔ یک جواب، جایگشت‌هایی هستند که با انتخاب یک جفت از شهرهای هم‌جوار، از کل مجموعه جایگشت‌ها، و جابه‌جا کردن آن دو شهر ایجاد می‌شوند. عمل تغییر در جواب فعلی و رفتن به جواب‌های همسایه «حرکت»[۲۹] خوانده می‌شود و حرکت‌های متفاوت، همسایه‌های گوناگون را ارائه می‌دهد.

۲-۴- مروری بر کارهای گذشته

یکی از روش‌های متداول برای اجماع خوشه‌بندی استفاده از روش انباشت شواهد است. فرد و جین، پایه‌گذار این روش هستند [۵ و ۷] که در آن نتایج خوشه‌بندی‌های پایه در ماتریسی به نام ماتریس مشارکت هم‌زمان[۳۰] ذخیره می‌شود. این ماتریس مشخّص می‌کند که هر جفت نمونه چند بار هم‌خوشه بوده‌اند. پس از ساختن ماتریس مشارکت هم‌زمان می‌توان هر سطر از این ماتریس را یک نمونه فرض کرد و با استفاده از هریک از الگوریتم‌های خوشه‌بندی پایه به‌ویژه، روش‌های سلسله‌مراتبی به افرازبندی نهایی دست یافت. در [۹] از الگوریتمی برای یافتن درخت پوشای کمینه برای استخراج خوشه‌های نهایی از ماتریس مشارکت هم‌زمان استفاده شده است. همچنین، در [۱۰] دو روش خوشه‌بندی سلسله‌مراتبی اتّصال منفرد و اتّصال میانگین روی ماتریس مشارکت هم‌زمان اعمال شده‌اند تا افرازبندی نهایی مشخّص شود.

در ماتریس مشارکت هم‌خوشه شدن نمونه‌ها شمارش می‌شود و اطّلاعاتی همچون اندازهٔ خوشه‌ها در نظر گرفته نمی‌شود. به عبارت دیگر، روش‌هایی که تنها بر پایهٔ استخراج ماتریس مشارکت هم‌زمان کار می‌کنند، دارای این کمبود هستند؛ که برخی ویژگی‌های مهم و مفید را نادیده می‌گیرند. در [۱۱] روشی پیشنهاد شده است که آن را انباشت احتمال[۳۱] نام‌گذاری کرده‌اند. در این روش، به‌جای توجّه انحصاری به هم‌خوشه بودن نمونه‌ها، احتمالی محاسبه می‌شود که اندازهٔ خوشه‌های اولیّه نیز در محاسبهٔ مقدار آنها نقش دارد.

در [۱۲] روشی پیشنهاد شده است که به‌جای اعمال الگوریتم خوشه‌بندی روی ماتریس مشارکت هم‌زمان، گرافی به نام گراف شباهت از آن استخراج می‌شود. سپس پردازش‌هایی روی گراف انجام می‌شود که هدف آنها تعیین خوشه‌های نهایی است. در این کار ادعا شده است که روش مذکور مقید به توزیع‌های داده‌ای خاص نیست (مانند الگوریتم K – Means که برای خوشه‌های دایروی مناسب است). همچنین، تعیین خودکار تعداد خوشه‌ها از دیگر ویژگی‌های این الگوریتم است.

در [۱۳] روشی پیشنهاد شده است که هنگام پر کردن ماتریس مشارکت هم‌زمان، در محاسبهٔ مقداری که برای دو نمونه هم‌خوشه ثبت می‌کند، میزان شباهت آن دو نمونه را تأثیر می‌دهد. به‌این‌ترتیب، میزان تأثیرگذاری خوشه‌بندی‌های پایه با یکدیگر متفاوت خواهد شد. پس از ساختن ماتریس مشارکت هم‌زمان به شیوهٔ گفته‌شده، ماتریس مشارکت هم‌زمان به ماتریس دیگری تبدیل می‌شود که ماتریس شباهت مسیرمحور[۳۲] نامیده می‌شود.

در [۱۴] پیشنهاد شده است که به‌جای ساختن یک ماتریس مشارکت هم‌زمان ساده، ماتریس مشارکت وزن‌دار ساخته شود. ماتریس مشارکت وزن‌دار حاوی اطّلاعات جامع‌تری در مورد خوشه‌بندی‌های پایه است و می‌تواند نقش بهتری در تعیین خوشه‌بندی نهایی ایفا کند.

یک دسته از الگوریتم‌های خوشه‌بندی، الگوریتم‌های خوشه‌بندی مبتنی بر چگالی هستند. در [۱۵] روشی دومرحله‌ای برای استخراج ماتریس همبستگی پیشنهاد شده است. در مرحلهٔ نخست، یک روش جدید برای تخمین چگالی و استخراج ماتریس نزدیکی[۳۳] است که در برخی روش‌های خوشه‌بندی مبتنی بر چگالی استخراج می‌شود. سپس، برای مرحلهٔ دوم، روشی برای استخراج نسخه‌ای طبیعی‌شده از ماتریس شباهت مبتنی بر چگالی پیشنهاد شده است که معادل با ماتریس همبستگی در روش خوشه‌بندی انباشت شواهد است.

در [۱۶] روشی پیشنهاد شده است که هم‌زمان تعداد خوشه‌ها و جواب نهایی را به کمک یک روش ترکیبی مشخّص می‌کند. در این روش، ماتریس شباهت که برآمده از مجموعه‌ای از خوشه‌بندی‌هاست، برای مقادیر مختلف اندازهٔ خوشه ساخته می‌شود. برای تعیین تعداد خوشه‌ها، نخست، ماتریس توافقی[۳۴] M به‌ازای مقادیر

[۲۹] Move
[۳۰] Co-association matrix
[۳۱] Probability accumulation
[۳۲] Path-based similarity matrix
[۳۳] Affinity matrix
[۳۴] Consensus matrix

[DOI: 10.61186/jsdp.20.1.99]

مختلف تعداد خوشه محاسبه می‌شود. سپس، مقادیر ویژهٔ ماتریس P استخراج می‌شود که بر مبنای ماتریس M ساخته می‌شود.

در برخی از روش‌های جدید، از راهکارهای انتخاب ویژگی و همچنین، استخراج ویژگی در کنار خوشه‌بندی جمعی استفاده شده است. روش‌های انتخاب ویژگی، زیرمجموعه‌ای از ویژگی‌های اصلی را تولید می‌کنند، در-حالی‌که روش‌های استخراج ویژگی، ویژگی‌های جدیدی را بر اساس ویژگی‌های اصلی ایجاد می‌کنند. هدف نهایی این روش‌ها حذف ویژگی‌های زائد، نوفه‌ای یا بی‌ربط برای کاهش یادگیری الگوریتم‌های تحت‌نظارت یا نظارت‌نشده است. برای مروری بر تاریخ و بررسی الگوریتم‌های پیشرفته برای انتخاب/ استخراج ویژگی، به [۱۷- ۲۰] مراجعه کنید.

در زمینهٔ اجماع خوشه‌ای، روش‌های جدیدی ارائه شده است که شامل آگاهی از نوع داده [۲۱] و الگوریتم‌های انتخاب ویژگی جدید بدون‌نظارت [۲۲ و ۲۳] است. توابع اجماع مبتنی بر انباشت شواهد برای مجموعه داده‌های عظیم به‌دلیل درجه دوم بودن و پیچیدگی فضا در تعداد نقاط داده هنگام ساخت ماتریس هم‌زمانی مقیاس خوبی ندارند [۲۴ و ۲۵].

برخی از روش‌هایی که در ادامه معرفی می‌شوند از افرازبندی و وزن‌دهی خوشه‌ها برای اجماع استفاده می-کنند. روش‌های مختلفی برای تعیین وزن افراز در مجمع وجود دارد: وزن‌دهی بر اساس خواصّ ذاتی خوشه‌ای، تحلیل قابلیّت اطمینان بر اساس شاخص‌های اعتبار خوشهٔ خارجی و تحلیل ارتباط پارتیشن با استفاده از شاخص‌های اعتبار خوشه داخلی. با در نظر گرفتن اندازه خوشه‌ها، الگوریتم اصلی تجمیع شواهد در [۲۶] ارائه شده است. این روش خوشه‌بندی احتمالی احتمالی (PAC) نامیده می‌شود و ایدهٔ یک ماتریس همبستگی غیردودویی (غیرباینری) را مطرح می‌کند، به طوری‌که یک شباهت جفتی بین اشیای داده، یک عدد واقعی بین صفر و یک است. بنابراین، روابط بین اشیای داده مجدد و آموزش دیده می‌شوند. یک راه‌حل جایگزین به‌نام معنی-دار بودن همکاری و در این‌جا به‌عنوان تجمیع وزن ثابت (WEA) در [۲۷] ارائه شده است. همچنین، تعداد خوشه‌های افراز و اندازه‌گیری نقاط شباهت داده‌ها را در نظر می‌گیرد.

راهبرد پالایش ماتریس هم‌زمانی شباهت مبتنی بر اتّصال سه‌گانه (CTS) نامیده می‌شود و بخشی از الگوریتم اجماع خوشه مبتنی بر پیوند (LCE) است [۲۸]. در [۲۸]، اگر خوشه i شبیه خوشه j و خوشه j شبیه خوشه k باشد، فرض می‌شود خوشه‌های i و k نیز تا حدی شبیه هستند؛ حتّی اگر هیچ نقطه دادهٔ مشترکی بین آن‌ها وجود نداشته باشد. نویسندگان در [۲۸] رویکرد سه‌گانهٔ متصل را با در نظر گرفتن شباهت بین خوشه‌ها و قابلیّت اطمینان افراز خوشه با استفاده از اعتبار خوشه خارجی گسترش دادند. آن‌ها قابلیت اطمینان افراز را به‌عنوان یک شباهت متوسط به سایر افرازهای خوشه تعریف کردند. به‌عنوان یک معیار شباهت، از یک شاخص اعتبار خارجی به نام اطّلاعات متقابل طبیعی‌شده استفاده کردند. هرچه توافق بیشتری در مجمع بر روی یک افراز وجود داشته باشد، اطمینان بیشتری نیز در نظر گرفته می‌شود. برتری چنین رویکردی عدم نیاز به داده‌های اصلی برای ارزیابی اعضای مجمع است. با این‌حال، اگر افرازهای با کیفیّت پایین در گروه وجود داشته باشند، تجزیه و تحلیل قابلیت اطمینان از این طریق، نتایج بی‌ربطی به‌همراه خواهد داشت [۲۹-۳۲].

۳- روش پیشنهادی

روش پیشنهادی دارای چند ویژگی است، که عبارتند از: مسألهٔ خوشه‌بندی به‌صورت یک مسألهٔ بهینه‌سازی صریح تعریف می‌شود. مهم‌ترین مؤلّفهٔ یک مسأله، بهینه‌سازی معرفی یک تابع هدف مناسب است که این کار هم‌ارز با یافتن پاسخ برای مسألهٔ اصلی است. در روش پیشنهادی از تابع هدفی که در رابطه (۷) مشخّص شده است، استفاده می‌شود. تابع هدف موردنظر یک تابع احتمالی است. می‌توان فرض کرد که تعلّق یک نمونه به خوشه‌های ممکن یک تابع توزیع احتمال است که در آن برای هر نمونه یک بردار مانند y_i وجود دارد که مشخّص می‌کند احتمال تعلّق نمونهٔ موردنظر به هریک از خوشه‌ها چقدر است. اکنون می‌توان تابع هدف را بر پایهٔ بردارهای توزیع احتمال، ماتریس همبستگی و ماتریس مشارکت در افرازبندی تعریف کرد.

$$Y^* = \arg\min\left\{ \sum_{for\ each\ i\ and\ j} N_{i,j}\left(\frac{c_{i,j}}{N_{i,j}} - y_i^T y_j \right)^2 \right\} \quad (۷)$$

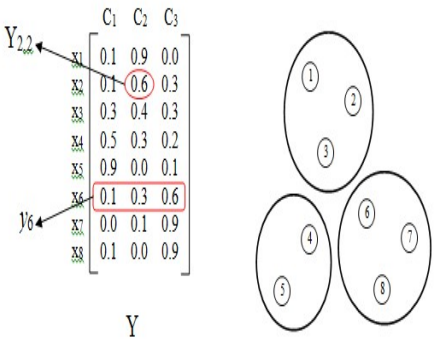
از ابزار بهینه‌سازی تابکاری فلزات برای حل مسأله بهینه‌سازی استفاده می‌شود. به‌کارگیری الگوریتم تابکاری فلزات الزاماتی دارد. به بیان دقیق‌تر، باید چند چیز به-خوبی روشن شود. نخست این‌که باید مشخّص شود هر حالت مسأله چگونه بازنمایی می‌شود. دوم اینکه، حالت

[DOI: 10.61186/jsdp.20.1.99]

آغازین چگونه تعیین می‌شود. مورد سوم، نحوهٔ تعیین همسایه‌ها یا عملگر حرکت است. نوآوری روش پیشنهادی به‌طور عمده مربوط به نحوهٔ تعیین حالت آغازین و عملگر حرکت است که در ادامه با تفصیل بیشتر در مورد آنها صحبت خواهد شد.

۳-۱- تعریف حالت مسأله

حالت مسأله شامل مجموعهٔ همه شاخص‌هایی است که باید مقدار مناسب آنها مشخّص شود. در رابطهٔ (۷)، از سه ماتریس C، N و Y استفاده می‌شود که به‌ترتیب ماتریس همبستگی، ماتریس مشارکت هم‌زمان در افرازبندی و ماتریس احتمال تعلّق نمونه‌ها به خوشه‌های نهایی است. ماتریس همبستگی و ماتریس مشارکت هم‌زمان در افرازبندی، مقادیر معلوم دارند که با اجرای الگوریتم انباشت شواهد به دست می‌آیند. به عبارت دیگر، الگوریتم انباشت شواهد یک بار در آغاز روش پیشنهادی باید اجرا شود تا ماتریس همبستگی و ماتریس مشارکت هم‌زمان در افرازبندی از روی آن ساخته شود. پس از ساخت این دو ماتریس، مقدار آن‌ها مشخّص است و تنها چیزی که باید مشخّص شود، ماتریس Y است. در واقع، ماتریس Y حالت مسأله است و یافتن مقدار بهینه برای آن یعنی Y^* هدف مسأله است؛ به عنوان مثال، شکل ۱ یک حالت از مسأله را نشان می‌دهد که هشت نمونهٔ x_1 تا x_8 با احتمال‌های مختلف به سه خوشه C_1 تا C_3 نسبت داده شده‌اند. ماتریس Y برای این مسألهٔ خوشه‌بندی یک ماتریس هشت در سه است که سطرهای آن مربوط به نمونه‌ها و ستون‌های آن مربوط به خوشه‌ها است. درایه $Y_{2,2}$ نشان می‌دهد که نمونهٔ دوم با احتمال ۰٫۶ به خوشهٔ دو تعلّق دارد. بردار y_6 نیز بیانگر احتمال تعلّق نمونه ششم به هریک از خوشه است. ماتریس Y نشان‌داده‌شده در شکل، یک جواب خوب برای مسأله است، چون احتمال تعلّق نمونه‌ها به خوشه‌ها به نحوی است که برای هر نمونه، خوشهٔ درست بیشترین احتمال را دارد.



(شکل-۱): مثال برای حالت مسأله

Figure 1: Example for problem state

۳-۲- تعیین حالت آغازین در روش پیشنهادی

نخستین گام در روش پیشنهادی، تعیین حالت آغازین است. نحوه تعیین حالت آغازین در روش پیشنهادی در شکل ۲ نشان داده شده است. تعیین حالت آغازین در روش پیشنهادی نیازمند یک تخمین اولیّه از برچسب نمونه‌هاست. برای این کار، نخست، باید یک الگوریتم خوشه‌بندی اجرا شود و نمونه‌ها را برچسب‌گذاری کند. پس از آن، می‌توان در ماتریس Y باتوجّه‌به برچسب‌ها مقادیر اولیّه را تعیین کرد (شکل ۲ ماتریس a). ازآن-جاکه مقداردهی اولیّه با برچسب‌ها ممکن است باعث هم‌گرایی زودهنگام الگوریتم بهینه‌سازی شود، باید به ماتریس قبلی مقداری نوفه اضافه کرد (شکل ۲ ماتریس b). برای این کار، یک ماتریس از اعداد تصادفی ساخته می‌شود. سپس دو ماتریس a و b با هم جمع می‌شود (شکل ۲ ماتریس c). در آخرین گام، باتوجّه‌به‌این‌که هر سطر از ماتریس Y باید یک توزیع احتمال باشد، ضروری است که جمع درایه‌ها در سطرها برابر با یک شود، باید هر سطر را بر مجموع درایه‌های آن تقسیم کرد (شکل ۲ ماتریس d).

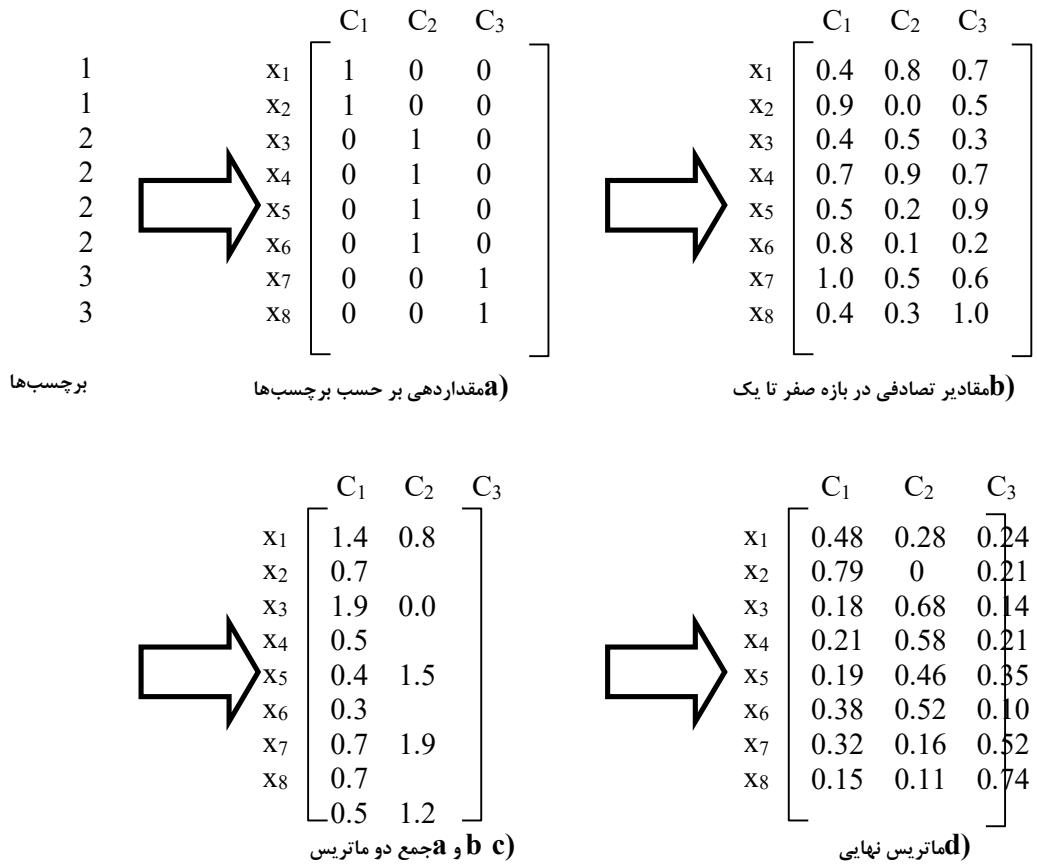
الگوریتم تابکاری فلزات نیاز مبرم به پیمانه‌ای دارد که حالت‌های پسین هر حالت را مشخّص کند. حالت‌های پسین همان حالت‌های همسایه هستند و وجود چنین پیمانه‌ای کمک می‌کند تا الگوریتم بتواند در فضای حالت مسأله جستجوی محلی انجام دهد. یکی از ساده‌ترین حالت‌های پسین که به ذهن می‌رسد، ایجاد تغییرات جزیی در ماتریس حالت مسألهٔ Y است؛ به عنوان مثال، می‌توان به هر درایه یک مقدار تصادفی اضافه کرد و سپس باتوجّه‌به‌این‌که هر سطر باید معرف یک توزیع

[DOI: 10.61186/jsdp.20.1.99]

احتمال باشد، درایه‌های سطر را تقسیم بر مجموع آن‌ها کرد (کم و بیش مشابه کاری که در تعیین حالت اولیّه انجام شد).

یکی از مهم‌ترین نوآوری‌های روش پیشنهادی، معرّفی یک روش نوین برای تعیین حالت‌های پسین است. شبه-رمز نحوهٔ حرکت به حالت پسین در روش پیشنهادی در شکل ۳ آمده است.

[DOI: 10.61186/jsdp.20.1.99]



(شکل- ۲): تعیین حالت آغازین در روش پیشنهادی

Figure 2: Determining the initial state in the proposed method

```
NextState = GetNextState(CurrentState, Min, Max, ChangeNum, Step)
NextState = CurrentState;
MinVal = GetMinElementOf(CurrentState);
MaxVal = GetMaxElementOf(CurrentState);
Range = MaxVal – MinVal;
MinRange = MinValue + Min ×Range;
MinRange = MinValue + Max ×Range;
[Rows, Cols] = find(MinRange < CM & CM < MaxRange);
for i = 1: ChangeNum
  rnum = RandomNumber(1);
  NextState (Rows(rnum, 1)) = NextState(Rows(rnum, 1),1) +
                              tep ∗ NextState (Cols(rnum, 1));
end
 for i = 1: n
  NextState(i, : ) = NextState(i, : ) / sum(NextState(i, : ));
End
```

(شکل- ۳): شبه‌رمز حرکت به حالت پسین

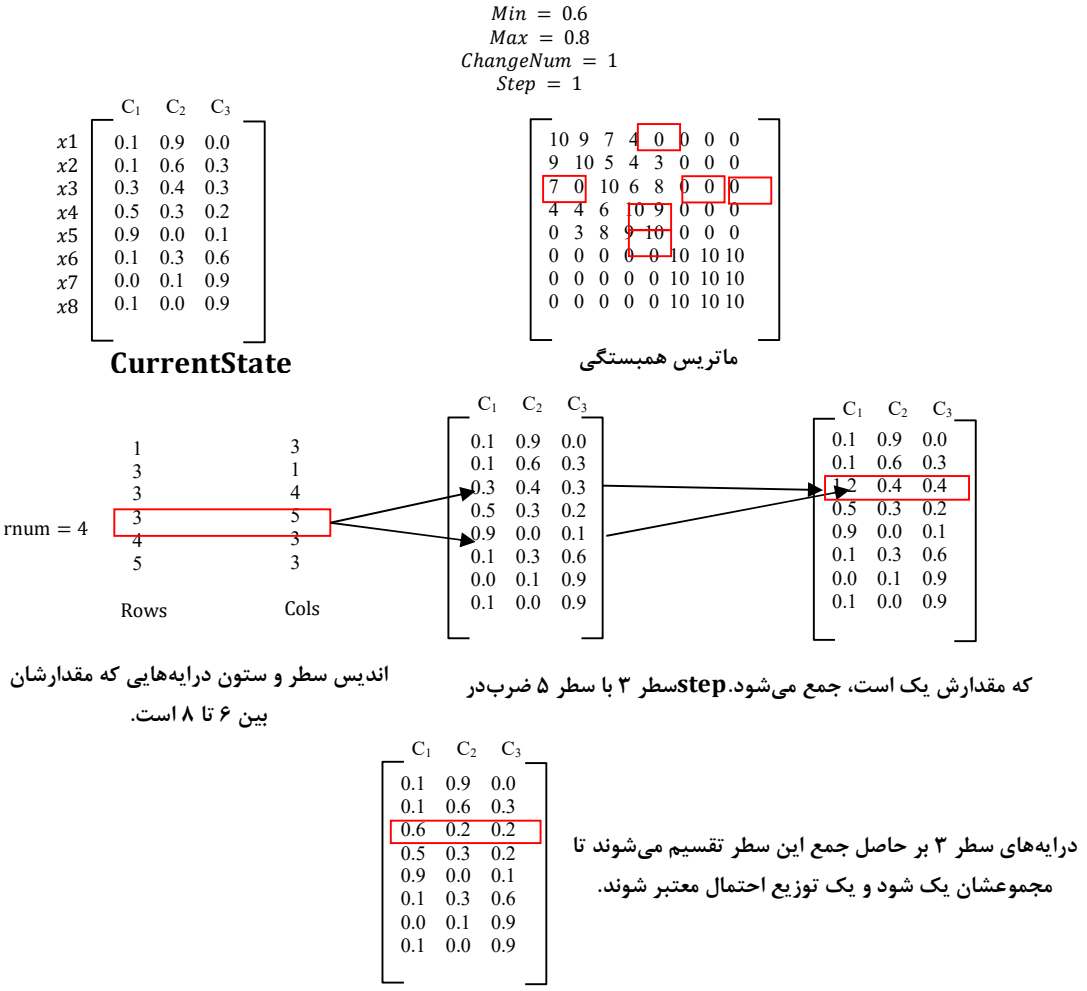
Figure 3: Pseudo-code to move backwards state

۳-۳- حرکت به حالت پسین

در ماتریس همبستگی، مقدار درایه‌ها عددی بین صفر تا تعداد خوشه‌بندی‌ها است. این روش، روی درایه‌هایی از ماتریس همبستگی تمرکز می‌شود که مقدار آنها در یک بازهٔ خاص است. این بازهٔ خاص توسط دو شاخص ورودی Min و Max مشخّص می‌شود که هر دو عددی بین صفر

[DOI: 10.61186/jsdp.20.1.99]

و یک هستند و Min کوچک‌تر از Max است؛ به عنوان مثال، فرض کنیم تعداد خوشه‌بندی‌ها ۱۰۰ باشد. به‌این‌ترتیب، بزرگ‌ترین مقدار در ماتریس همبستگی ۱۰۰ است و کمترین آن ۰. اگر مقدار دو شاخص Min و Max به ترتیب ۰٫۶ و ۰٫۸ باشد، درایه‌هایی از ماتریس همبستگی برای ما جذاب است که مقدار بین ۶۰ تا ۸۰ داشته باشد. آغاز و پایان بازهٔ موردنظر با دو شاخص MinRange و MaxRange مشخّص می‌شود. پس از تعیین این بازه، همهٔ درایه‌هایی را که در این بازه هستند، پیدا و اندیس سطر و ستون آنها را در دو ماتریس Rows و Cols ذخیره می‌کنیم. گام بعدی، تغییر برخی سطرها در ماتریس Y یا همان حالت کنونی CurrentState است. تعداد تغییرها یکی از شاخص‌های ورودی است که با ChangeNum مشخّص می‌شود. به‌طور تصادفی یکی از درایه‌ها که در بازهٔ MinRange و MaxRange وجود دارد، انتخاب می‌شود. فرض کنیم این درایه rnumام درایه باشد که اندیس سطر و ستون آن به‌ترتیب، در Rows(rnum) و Cols(rnum) آمده است. اکنون باید سطر Rows(rnum) از ماتریس Y را با توجه به سطر Cols(num) از ماتریس Y تغییر دهیم. معنای این کار آن است که توزیع احتمال نمونه Rows(rnum)ام را به توزیع احتمال نمونه Cols(rnum) نزدیک می‌کنیم. در واقع، سطر Rows(rnum) با ضرب Cols(rnum) در Step جمع می‌شود. شاخص Step مشخّص می‌کند که سطر اول، به چه میزان از سطر دوم تأثیر بگیرد. شکل ۴ با یک مثال، نحوهٔ حرکت به حالت پسین را نشان می‌دهد.



(شکل- ۴): مثال برای حرکت به حالت پسین

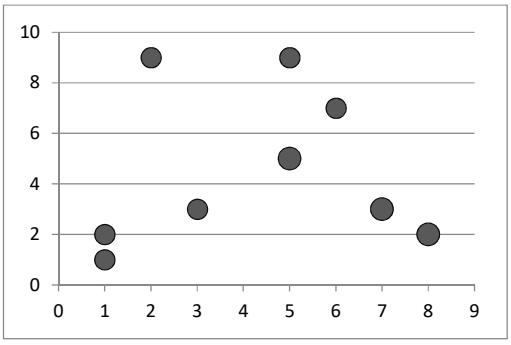
Figure 4: Example for moving backwards

[DOI: 10.61186/jsdp.20.1.99]

در شکل ۴، حالت کنونی، شاخص‌های ورودی و ماتریس همبستگی مفروض گرفته شده‌اند. نخستین گام، تعیین درایه‌هایی است که در بازهٔ موردنظر هستند. این بازه شامل اعداد شش تا هشت است. تعداد شش درایه از ماتریس در این بازه وجود دارند که سطر و ستون آنها در دو ماتریس Rows و Cols مشخّص شده است. در گام بعدی، باتوجّه‌به‌این‌که مقدار ChangeNum برابر با یک است، یک تغییر باید در حالت کنونی انجام شود. عدد تصادفی rnum تولید می‌شود که مقدار آن برابر با چهار است. این بدان معناست که سطر سوم باید با سطر پنجم جمع، و حاصل در سطر سوم ذخیره شود. در نهایت نیز، درایه‌های سطر سوم بر مجموع آنها تقسیم می‌شود تا یک توزیع احتمال معتبر ایجاد شود. یک رمز ساده برای نمایش چگونگی عملکرد برخی از بخش‌های روش پیشنهادی در زیر ارایه شده است:

```
clc
clear
close all;
X = [1,1; 1,2; 3,3; 5,9; 8,2; 5,5; 6,7; 7,3; 2,9];
k = 3;
[~,LabelSet,SampleIndicesSet] = ConstructingKMeansEnsembleWithSampling(X, k, 5, 7/8);
[CM,N] = ExtractMatricesCMandN(SampleIndicesSet,LabelSet);
Z = linkage(CM,'single');
L = cluster(Z,'maxclust',k);
plot(X(L==1,1),X(L==1,2),'or');
hold on;
plot(X(L==2,1),X(L==2,2),'*g');
plot(X(L==3,1),X(L==3,2),'+b');
Y = RandomY(size(X,1),k,L,1);
[ObjectiveValue] = ObjectiveFunc2(CM, N, Y);
[NY] = NextY2(Y,CM,0.7,1,1,1);
```

مجموعه داده‌ای را به صورت زیر در نظر می‌گیریم که شامل نه نمونه است (خط چهار ساخته می‌شود):



سپس یک مجمع از خوشه‌بندی‌ها ساخته می‌شود شامل ۵ خوشه‌بندی با سه خوشه با نرخ نمونه‌برداری ۰٫۸۷۵ (یا همان ۷/۸) که در خط شش رمز آمده است. نتایج خوشه‌بندی به‌صورت زیر است (هر جا برچسب یک نمونه صفر است، به معنی عدم حضور در نمونه‌برداری است. در هر خوشه‌بندی تنها هفت نمونه از هشت نمونه مشارکت دارد).

شمارهٔ نمونه	شمارهٔ خوشه‌بندی					
		1	2	3	4	5
	1	2	1	3	2	2
	2	2	1	3	2	2
	3	2	1	0	2	2
	4	0	3	1	1	1
	5	1	2	2	0	0
	6	3	2	2	3	3
	7	3	0	1	3	3
	8	1	2	2	3	3
	9	3	3	1	1	1

سپس ماتریس‌های همبستگی (CM) و ماتریس حضور هم‌زمان در نمونه‌برداری (N) استخراج می‌شود (خط هفت رمز). که به قرار زیر هستند:

ماتریس CM

	۱	۲	۳	۴	۵	۶	۷	۸	۹
۱	5	5	4	0	0	0	0	0	0
۲	5	5	4	0	0	0	0	0	0
۳	4	4	4	0	0	0	0	0	0
۴	0	0	0	4	0	0	1	0	4
۵	0	0	0	0	3	2	0	3	0
۶	0	0	0	0	2	5	3	4	1
۷	0	0	0	1	0	3	4	2	2
۸	0	0	0	0	3	4	2	5	0
۹	0	0	0	4	0	1	2	0	5

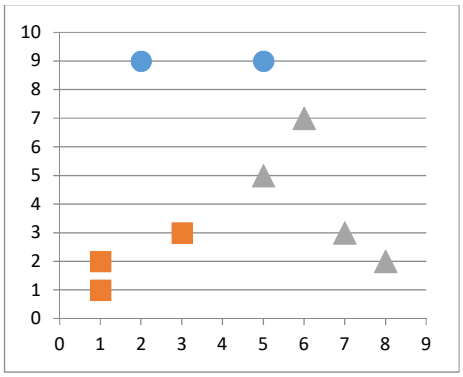
ماتریس N

	۱	۲	۳	۴	۵	۶	۷	۸	۹
۱	5	5	4	4	3	5	4	5	5
۲	5	5	4	4	3	5	4	5	5
۳	4	4	4	3	2	4	3	4	4
۴	4	4	3	4	2	4	3	4	4
۵	3	3	2	2	3	3	2	3	3
۶	5	5	4	4	3	5	4	5	5
۷	4	4	3	3	2	4	4	4	4
۸	5	5	4	4	3	5	4	5	5
۹	5	5	4	4	3	5	4	5	5

سپس با اعمال یک الگوریتم اتّصال منفرد روی CM خوشه‌بندی نهایی استخراج می‌شود (خط هشت و نه)، تا از آن برای مقداردهی اولیّهٔ Y استفاده شود. برچسب نهایی نمونه‌ها:

نمونه	برچسب
۱	3
۲	3
۳	3
۴	1
۵	2
۶	2
۷	2
۸	2
۹	1

[DOI: 10.61186/jsdp.20.1.99]



حاصل خوشه‌بندی روی شکل

سپس، نوبت به تعیین اولیّهٔ ماتریس Y می‌رسد (خط ۱۴ کد $Y = RandomY(size(X,1), k, L, 1)$).

باتوجّه‌به‌این‌که مقدار شاخص آخر یک است، وزن برچسب‌های گام قبل در مقداردهی اولیّه به Y برابر با یک است. ماتریس Y حاصل به‌صورت زیر است:

	۱	۲	۳
۱	0.145652	0.35705	0.497298
۲	0.064247	0.339909	0.595845
۳	0.186866	0.242993	0.570141
۴	0.79636	0.065065	0.138575
۵	0.04725	0.56609	0.38666
۶	0.387276	0.611055	0.001669
۷	0.20549	0.711528	0.082982
۸	0.307244	0.548141	0.144615
۹	0.530406	0.291546	0.178048

سپس، این ماتریس ارزیابی می‌شود که مقدار تابع هدف برای آن ۳۱,۸۳ است. مشاهده می‌شود که برای هر نمونه، احتمال مربوط به خوشهٔ درست آن بیشتر است. در گام آخر مثال، Y بعدی به کمک تابع NextY ساخته می‌شود که ماتریس زیر به دست می‌آید. به‌جز شاخص Y و CM که مقدار Y پیشین و ماتریس همبستگی هستند، چهار شاخص دیگر، به‌ترتیب، مشخّص می‌کنند که دنبال کدام درایه‌ها از ماتریس همبستگی بر ایجاد تغییر در Y هستیم (MinRange و MaxRange) و تعداد تغییرات (ChangeNum) و چگونگی آنها (Step) چگونه است.

$[NY] = NextY2(Y, CM, 0.7,1,1,1);$

$[NY] = NextY2(Y, CM, MinRange, MaxRange, ChangeNum, Step)$

	۱	۲	۳
۱	0.145652	0.35705	0.497298
۲	0.064247	0.339909	0.595845
۳	0.31489	0.204735	0.480375
۴	0.79636	0.065065	0.138575
۵	0.04725	0.56609	0.38666
۶	0.387276	0.611055	0.001669
۷	0.20549	0.711528	0.082982
۸	0.307244	0.548141	0.144615
۹	0.530406	0.291546	0.178048

مقدار تابع هدف برای این Y جدید ۳۳,۲۲ است. با دقّت در ماتریس جدید مشاهده می‌شود که احتمال‌ها برای نمونهٔ سه تغییر کرده است.

۴- آزمایش‌ها و ارزیابی نتایج تجربی

این بخش به بررسی نتایج پیاده‌سازی روش پیشنهادی اختصاص دارد. در ادامه، نخست، مجموعه داده‌های استفاده‌شده در آزمایش‌های تجربی معرفی خواهد شد. سپس، بعد از بررسی شرایط پیاده‌سازی، نتایج آزمایش‌های طرّاحی‌شده برای مقایسهٔ روش پیشنهادی با روش‌های دیگر و به نمایش گذاشتن جنبه‌های مختلف روش پیشنهادی بررسی خواهد شد.

۴-۱- مجموعهٔ داده‌ها

جدول ۱ ویژگی‌های هریک از مجموعه‌داده‌های استاندارد مورد استفاده را نشان می‌دهد. در همه نتایج تجربی، از مجموعه داده‌های طبیعی‌شده با میانگین صفر و واریانس یک استفاده شده است.

جدول ۱: مشخصات مجموعه داده‌های استاندارد استفاده‌شده

Table 1: Specifications of the standard data set used

	نام مجموعه داده	تعداد نمونه‌ها	تعداد ویژگی‌ها	تعداد رده‌ها
1	Breast Tissue	106	9	6
2	Bupa	345	6	2
3	Dolphins	62	62	5
4	Ecoli	336	7	8
5	Galaxy	323	4	7
6	Glass	214	9	6
7	Haberman	306	3	2
8	Halfring	400	2	2
9	Ionosphere	351	34	2
10	Iris	150	4	3
11	Planning	182	12	2
12	SAHeart	462	9	2
13	Seeds	210	7	3
14	Vertebral3C	310	6	3
15	Wine	178	13	3

همان‌طورکه در جدول مشاهده می‌شود، مجموعه داده‌های انتخابی از نظر تعداد نمونه‌ها تعداد ویژگی‌ها و تعداد رده‌ها از تنوّع خوبی برخوردارهستند.

روش پیشنهادی در نرم‌افزار Matlab R2013a پیاده‌سازی شده است. الگوریتم‌های خوشه‌بندی پایه مورد استفاده در آزمایش‌های تجربی شامل K – Means، GMM و اتّصال منفرد همان توابع پیش‌فرض تعبیه‌شده

[DOI: 10.61186/jsdp.20.1.99]

در این نرم‌افزار هستند. همچنین، نتایج اجرای الگوریتم‌های منفرد تصادفی که شاخص‌های اولیّه روی کارایی آنها تأثیر دارد، به‌ازای ۲۰ اجرای مستقل از هم گزارش شده است.

۴-۲- روش اعتبارسنجی نتایج

نتایج تجربی روش پیشنهادی و روش‌های دیگر با معیارهای معتبر جهانی نظیر معیار فیشر، اطّلاعات متقابل طبیعی‌شده و با دقّت گزارش خواهد شد.

روش اول اعتبارسنجی یک افراز، دقّت است. در این حالت، ما برچسب‌های واقعی را می‌دانیم، و میزان دقّت افراز موردنظر و برچسب‌های واقعی را محاسبه می‌کنیم.

معیار فیشر و اطّلاعات متقابل طبیعی‌شده طبق رابط زیر محاسبه می‌شود:

$$NMI(P,L) = \frac{-2\sum_{i=1}^{K_P}\sum_{j=1}^{K_L}\frac{N_{ij}^{PL}}{N}\log\frac{N_{ij}^{PL}N}{N_i^P N_j^L}}{\sum_{i=1}^{K_P}N_i^P\log\frac{N_i^P}{N}+\sum_{j=1}^{K_L}N_j^L\log\frac{N_j^L}{N}} \quad (۸)$$

اگر دو افراز P و برچسب L به‌طور کامل مشابه باشند، آنگاه NMI مقدار ماکزیمم، یعنی یک و اگر دو افراز به‌طور کامل متفاوت از یکدیگر باشند، مقدار صفر را برمی‌گرداند.

معیار دیگری که در این مقاله برای ارزیابی یک افراز در نظر گرفته شده است، معیار فیشر (FM) است؛

$$FM(P,L) = \max_{\tau}\sum_{i=1}^{K_P}\frac{2\times N_i^P\times\left(\frac{N_{i\tau(i)}^{PL}}{N_i^P}\times\frac{N_{i\tau(i)}^{PL}}{N_{\tau(i)}^{L}}\right)}{N\times\left(\frac{N_{i\tau(i)}^{PL}}{N_i^P}+\frac{N_{i\tau(i)}^{PL}}{N_{\tau(i)}^{L}}\right)} \quad (۹)$$

که K_P تعداد خوشه‌های افراز P، N_i^P نشان‌دهندهٔ تعداد داده‌های موجود در خوشهٔ i-ام از افراز P، N_j^L نشان‌دهندهٔ تعداد داده‌های موجود در خوشهٔ j-ام از افراز L، N_{ij}^{PL} نشان‌دهندهٔ تعداد داده‌هایی که به‌طور مشترک، در خوشهٔ i-ام از افراز P و در خوشهٔ 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 برچسب L به‌طور کامل مشابه باشند، آنگاه FM مقدار ماکزیمم یعنی یک؛ و اگر دو افراز به‌طور کامل متفاوت از یکدیگر باشند، مقدار صفر را برمی‌گرداند.

۴-۳- مقایسهٔ روش پیشنهادی با روش‌های دیگر

در این بخش، نتایج پیاده‌سازی روش پیشنهادی با چند روش دیگر روی مجموعه داده‌های معرفی‌شده در ۵-۲ گزارش خواهد شد. علاوه‌بر چهار الگوریتم خوشه‌بندی پایه $K-Means$، GMM، اتّصال منفرد (SL) و $Fuzzy\ C-Means$ (FCM)، روش پیشنهادی با روش ترکیبی انباشت شواهد در دو حالت مختلف مقایسه شده است. در حالت نخست، الگوریتم خوشه‌بندی پایه $K-Means$ (EAKM) و در حالت دوم GMM (EAGMM) در نظر گرفته شده است. نتایج مقایسه در جدول نشان داده شده است. در همه آزمایش‌ها اندازهٔ مجمع برای روش‌های ترکیبی ۱۰۰ در نظر گرفته شد.

نتایج در دو جدول ۲ و ۳ آمده است که جدول نخست روش‌های نام‌برده را به‌لحاظ دقّت مقایسه می‌کند و جدول دوم روش پیشنهادی را نه‌تنها بر حسب دقت، بلکه معیار ارزیابی NMI را نیز، با روش انباشت شواهد مقایسه کرده است. همان‌طورکه در جدول نشان داده شده است، دقّت روش پیشنهادی به‌طور تقریبی، روی تمام مجموعه داده‌ها از روش‌های دیگر بیشتر است. قابل ذکر است که در روش پیشنهادی شاخص‌های MinRange، MaxRange، Step و ChangeNum به ترتیب ۰,۷، ۰,۹، ۰,۵ و ۱ در نظر گرفته شده است. این مقادیر ثابت برای شاخص‌ها روی مجموعه داده‌ها نتایج قابل‌قبولی حاصل می‌کنند. برای انتخاب بهینهٔ مقدار شاخص‌ها نیاز به آزمون و خطا برای هر مجموعه داده است.

(جدول- ۲): مقایسهٔ دقّت روش پیشنهادی و پنج روش دیگر

Table 2: Comparison of the accuracy of the proposed method and five other methods

# of data set	KMeans	SL	FCM	EAGMM	EAKM	Proposed
1	0.412	0.44	0.43	0.432	0.443	0.621
2	0.533	0.54	0.52	0.535	0.545	0.665
3	0.459	0.45	0.42	0.453	0.468	0.687
4	0.72	0.74	0.76	0.719	0.714	0.892
5	0.275	0.29	0.28	0.28	0.272	0.423
6	0.442	0.44	0.43	0.444	0.453	0.571
7	0.545	0.51	0.53	0.511	0.516	0.559
8	0.88	0.90	0.85	0.876	0.88	0.938
9	0.707	0.65	0.70	0.695	0.707	0.836
10	0.833	0.82	0.82	0.842	0.833	0.901
11	0.534	0.55	0.52	0.534	0.555	0.703
12	0.625	0.621	0.62	0.623	0.63	0.892
13	0.919	0.91	0.89	0.91	0.919	0.993
14	0.462	0.46	0.47	0.474	0.468	0.726
15	0.966	0.88	0.96	0.966	0.966	0.989

چنان که جدول نشان می‌دهد، روش پیشنهادی با یک پیکره‌بندی نوعی برای شاخص‌ها روی ۱۱ مجموعه داده دقّت بالاتری نسبت‌به پنج روش دیگر دارد. نزدیک‌ترین رقیب آن روش انباشت شواهد با خوشه‌بندی پایه‌ی $K-Means$ است. روی برخی مجموعه داده‌ها به‌دلیل طبیعت مشترک برخی روش‌های خوشه‌بندی و همچنین، ویژگی‌های مجموعه داده نتایج مشابهی به دست آمده است. برای نمونه روی مجموعه داده نه روش‌هایی که خوشه‌بندی پایه در آنها خوشه‌های کروی تولید می‌کنند، نتایج یکسانی داشته‌اند.

همچنین، برای مقایسه بهتر دو روش انباشت شواهد با

[DOI: 10.61186/jsdp.20.1.99]

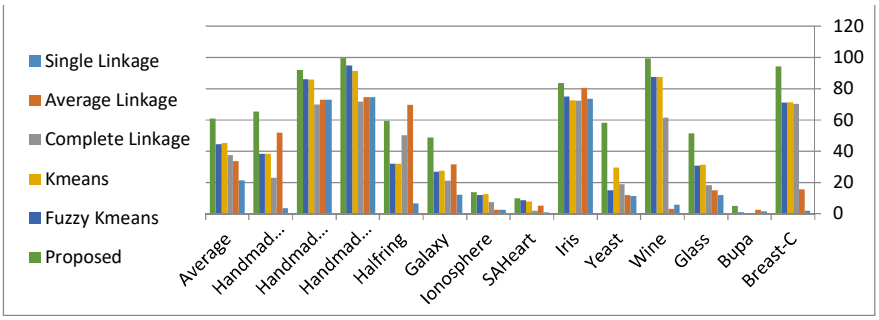
K – Means و روش پیشنهادی که عملکرد بهتری نسبت‌به بقیه دارند، در جدول ۳ معیار NMI نیز گزارش شده است. نتایج جدول ۳ حاکی از برتری کامل روش پیشنهادی از لحاظ دقّت و معیار NMI نسبت‌به روش انباشت شواهد است.

نتایج اطّلاعات متقابل طبیعی‌شده روش‌های پایه خوشه‌بندی و روش پیشنهادی بر روی مجموعه داده‌های گوناگون نیز در شکل ۵ ارایه شده است. معادل این جدول، برای معیار فیشر در شکل ۶ آورده شده است.

(جدول- ۳): مقایسهٔ دقّت و NMI روش پیشنهادی و روش انباشت شواهد

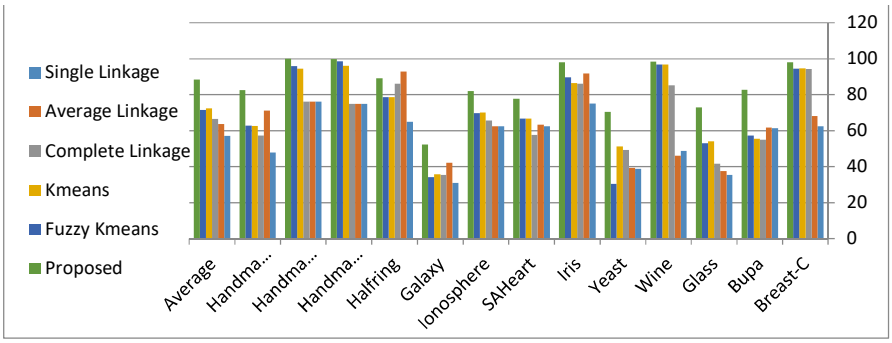
Table 3: Comparison of accuracy and NMI proposed method and evidence accumulation method

	EAKM		Proposed	
	Acc	NMI	Acc	NMI
1	0.443	0.497	0.527	0.665
2	0.545	0.002	0.648	0.201
3	0.468	0.394	0.563	0.558
4	0.714	0.65	0.967	0.794
5	0.272	0.196	0.425	0.401
6	0.453	0.315	0.621	0.466
7	0.516	0.001	0.722	0.2004
8	0.88	0.506	0.901	0.706
9	0.707	0.125	0.911	0.325
10	0.833	0.659	0.936	0.853
11	0.555	0.001	0.725	0.203
12	0.63	0.074	0.745	0.279
13	0.919	0.728	0.983	0.828
14	0.468	0.303	0.764	0.409
15	0.966	0.876	0.985	0.976



(شکل- ۵): اطّلاعات متقابل طبیعی‌شدهٔ روش‌های پایهٔ خوشه‌بندی بر روی مجموعه داده‌های استفاده‌شده

Figure 4: Normalized Interactive Cluster-Based Interaction Methods on the Data Collection Used



(شکل- ۶): معیار فیشر روش‌های خوشه‌بندی پایه بر روی مجموعه داده‌های مورد استفاده

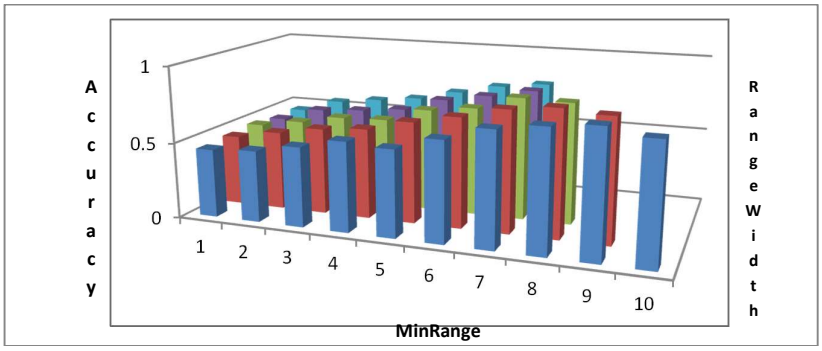
Figure 6: Fisher's criterion of basic clustering methods based on the data set used

۴-۵-بررسی اثر شاخص‌های MinRange و MaxRange

برای بررسی اثر شاخص‌های MinRange و MaxRange نخست، بهتر است تغییر کوچکی در آنها ایجاد کرد. این تغییر بدان صورت است که به‌جای استفاده از دو عدد برای تعیین آغاز و پایان بازه، می‌توان از دو عدد دیگر برای تعیین آغاز و طول بازه استفاده کرد. آغاز بازه که مانند قبل است؛ و برای اشاره به طول بازه نیز از متغیّر RangeWidth استفاده خواهد شد. شکل ۶ دقّت را به‌ازای این دو شاخص گزارش می‌کند. باتوجه‌به این، به‌ازای برخی ترکیب‌های دو شاخص، بازه‌های نامعتبر ایجاد می‌شود. برخی خانه‌های جدول خالی هستند. برای نمونه، نمی‌توان بازه‌ای با شروع از ۰,۷ و به طول ۰,۴ داشت، چون پایان بازه از یک تجاوز می‌کند و این در الگوریتم مجاز نیست. همان طورکه مشاهده می‌شود، بهترین دقّت به ازای MinRange = 0.7 و RangeWidth = 0.2 حاصل شده است. به نظر می‌رسد بازه باید نزدیک به یک انتخاب شود امّا نه چسبیده به آن. دلیل نزدیک بودن بازه به ۱ آن است که باید توزیع احتمال هر نمونه به نمونه‌ای نزدیک شود که با یکدیگر

[DOI: 10.61186/jsdp.20.1.99]

زیاد هم‌خوشه بوده‌اند. همچنین، برتری طول بازهٔ ۰٫۲ بدان سبب است که بازه باید به نحوی انتخاب شود که علاوه‌بر حفظ تنوّع که نسبت مستقیم با طول بازه دارد، از کیفیّت بازه نیز کاسته نشود.



(شکل- ۶): دقّت روی مجموعه داده Ecoli به‌ازای مقادیر مختلف MinRange و RangeWidth

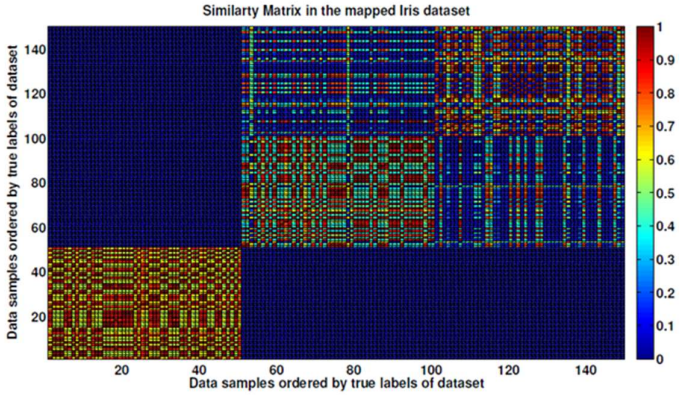
Figure 6: Accuracy on the Ecoli data set for different MinRange and RangeWidth values

برای نشان دادن کاربردی بودن چارچوب خوشه‌بندی پیشنهادی، شکل ۷ و شکل ۸ را در نظر بگیرید. در شکل ۷ و شکل ۸، یک مثال جامع‌تر از ماتریس‌های شباهت در نقاط دادهٔ اصلی و نگاشت‌شده Iris مجموعه داده ارائه شده است. شکل ۷ ماتریس شباهت نقاط داده را در مجموعه دادهٔ اصلی Iris نشان می‌دهد.

شایان ذکر است که ترتیب نقاط داده در هر ماتریس تشابه ارائه‌شده بر اساس برچسب‌های واقعی مجموعه داده برای درک بهتر است. شکل ۸ ماتریس شباهت نقاط داده را در مجموعه داده Iris نشان می‌دهد. همان‌طورکه از شکل ۷ و شکل ۸ استنباط می‌شود، ماتریس شباهت در مجموعه داده‌های Iris متفاوت از ماتریس شباهت در مجموعه داده اصلی Iris است. برای تعمیم نتیجه‌گیری، آزمایش‌های مشابهی روی مجموعه داده‌های wine تکرار شده است. شکل ۹ ماتریس شباهت نقاط داده را در مجموعه دادهٔ اصلی wine نشان می‌دهد.

(شکل- ۷). ماتریس شباهت نقاط داده در مجموعه دادهٔ اصلی Iris

Figure 7: Similarity matrix of data points in original Iris dataset.



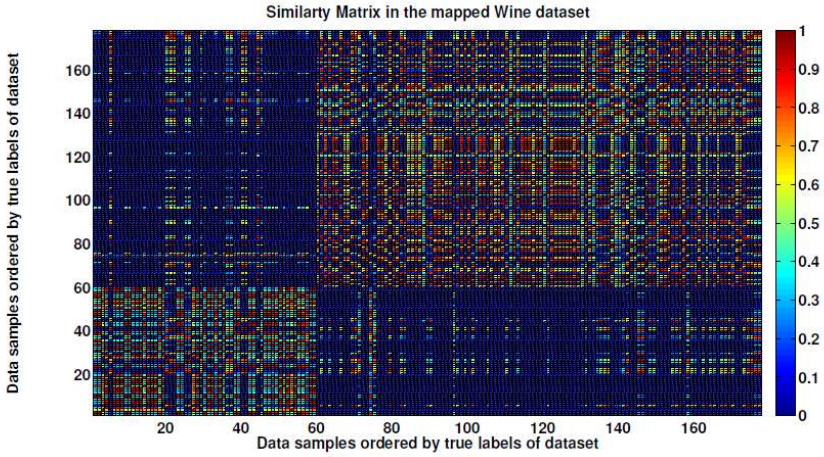
(شکل- ۸): ماتریس شباهت نقاط داده در مجموعه داده Iris

[DOI: 10.61186/jsdp.20.1.99]

Figure 8: Similarity Matrix of data points in mapped Iris dataset

(شکل- ۹): ماتریس شباهت نقاط داده در مجموعه دادهٔ اصلی **Wine**

Figure 9: Similarity matrix of data points in original Wine dataset.



(شکل- ۱۰). ماتریس شباهت نقاط داده در مجموعه داده‌های نگاشت‌شده Wine

Figure 10: Similarity Matrix of data points in mapped Wine dataset

شکل ۱۰ همچنین، ماتریس تشابه نقاط داده را در مجموعه داده‌های نگاشت‌شده wine نشان می‌دهد. همان‌طورکه می‌توان دوباره از شکل ۹ و شکل ۱۰ استنباط کرد، ماتریس شباهت در مجموعه داده‌های نگاشت‌شده از ماتریس شباهت در مجموعه داده اصلی نگاشت‌شده تفاوت دارد.

(جدول- ۴): توابع توافقی در اجماع به‌همراه توضیحات هر روش

Table 4: Consensus functions in ensemble with a description of each method

#	method	description
1	EAC	evidence accumulation clustering
2	CSPA	cluster-based similarity partitioning alg
3	HGPA	hypergraph partitioning algorithm
4	MCLA	meta-clustering algorithm
5	JWEAC	joint weighted EAC
6	PAC	probability accumulation clustering
7	WEA	weighted evidence accumulation using coassociation signi_cance
8	LCE	link-based cluster ensemble using connectedtriple based similarity [31],
9	DICLENS	divisive clustering ensemble with automatic cluster number [54],
10	EAC-W	EAC using PRA,
11	CSPA-W	CSPA using PRA,
12	DICLENS-W	DICLENS using PRA,
13	EAC-Wr	EAC using PRAr,
14	CSPA-Wr	DICLENS using PRAr.
15	DICLENS-Wr	DICLENS using PRAr.

(جدول- ۵): مجموعه داده استفاده‌شده در این بخش

Table 5: The dataset used in this section

dataset name	# of samples	# of features	# of clusters
breastCancerWisconsin	683	9	2
contractions	98	27	2
fertility	100	9	2
glass	214	9	6
ionosphere	351	33	2
iris	150	4	3
laryngeal3	353	16	3
letterRecognitionABC	2291	16	3
respiratory	85	17	2
segmentationTrain	210	18	7
voice_3	238	10	3

(جدول- ۶): میانگین و انحراف معیار رتبه AMI در مجموعه داده‌های REAL به‌طور متوسط بیش از ۳۰ اجرا برای توابع توافقی در اجماع. بهترین نمره برای هر مجموعه داده پررنگ مشخّص شده است.

Mean and standard deviation of AMI scores on REAL datasets averaged over 30 runs for consensus functions. The best score for each dataset is highlighted in bold.

DATASET	DICLENS	LCE	EAC	PAC	CSPA	HGPA	MCLA	WEA	JWEAC	PROPOSED
breastCancerWisconsin	0.323 ±0.305	0.001 ±0.002	0.001 ±0.001	0.000 ±0.001	0.655 ±0.023	0.687 ±0.046	0.628 ±0.030	0.001 ±0.002	0.001 ±0.001	**0.746 ±0.059**
contractions	0.188 ±0.088	0.021 ±0.061	0.000 ±0.003	0.005 ±0.024	0.326 ±0.072	0.311 ±0.092	0.328 ±0.126	0.000 0.000	0.011 ±0.040	**0.571 ±0.132**
fertility	0.008 ±0.024	−0.006 ±0.013	−0.008 ±0.003	−0.008 ±0.003	0.002 ±0.007	0.005 ±0.010	−0.001 ±0.007	−0.006 ±0.012	−0.008 ±0.003	**0.009 ±0.029**
glass	0.254 ±0.023	0.127 ±0.104	0.084 ±0.093	0.142 ±0.107	0.257 ±0.042	0.255 ±0.024	0.224 ±0.064	0.127 ±0.092	0.147 ±0.104	**0.289 ±0.076**
ionosphere	0.201 ±0.136	0.002 0.000	0.002 0.000	0.002 0.000	0.106 ±0.015	0.150 ±0.052	0.074 ±0.023	0.001 ±0.002	0.002 ±0.002	**0.293 ± 0.149**
iris	0.635 ±0.107	0.576 ±0.012	0.564 ±0.020	0.574 ±0.054	0.665 ±0.063	0.693 ± 0.082	0.692 ±0.076	0.566 ±0.018	0.564 ±0.015	**0.740 ±0.090**
laryngeal3	0.058 ±0.047	0.001 ±0.006	0.000 ±0.003	0.000 ±0.003	0.195 ±0.013	0.191 ±0.012	0.203 ±0.019	−0.002 ±0.001	0.000 ±0.006	**0.311 ± 0.026**
letterRecognitionABC	0.138 ±0.173	0.186 ±0.237	0.049 ±0.125	0.080 ±0.170	0.489 ±0.131	0.424 ±0.176	0.515 ±0.198	0.100 ±0.190	0.082 ±0.169	**0.639 ±0.106**
respiratory	0.186 ±0.224	0.030 ±0.085	0.018 ±0.022	0.015 ±0.026	0.598 ±0.082	0.469 ±0.197	0.546 ±0.129	0.001 ±0.004	0.033 ±0.083	**0.619 ± 0.097**
segmentationTrain	0.546 ±0.106	0.352 ±0.104	0.359 ±0.086	0.383 ±0.088	0.653 ±0.049	0.562 ±0.069	0.605 ±0.070	0.368 ±0.067	0.368 ±0.084	**0.698 ±0.056**
voice_3	0.105 ±0.047	0.033 ±0.002	0.032 ±0.001	0.032 ±0.001	0.117 ±0.035	0.108 ±0.034	0.115 ±0.038	0.028 ±0.012	0.033 ±0.002	**0.183 ±0.0473**

در این بخش روش پیشنهادی با ۱۵ روش اجماع که به‌شرح زیر آورده شده‌اند، مقایسه خواهد شد. به‌غیر از الگوریتم DICLENS، همهٔ روش‌های اجماع ذکرشده به تعداد K خوشه به‌عنوان یک شاخص نیاز دارند. به‌منظور مقایسه‌ای منصفانه، در این مقاله مطابق با روش [۲۳] عمل شده است. توابع اجماع EAC، JWEAC، PAC، WEA، LCE، $EAC-W$ و $EAC-Wr$، پارتیشن اجماع را با استفاده از یک الگوریتم خوشه‌بندی واحدی به ماتریسی که از الگوریتم سلسله‌مراتبی استفاده شده است، محاسبه می‌کنند. LCE به یک شاخص اضافی به نام ضریب تنزّل[۱] نیاز دارد که ما از مقدار ۰,۹ همانند [۳۳] استفاده کرده‌ایم.

برای اعتبارسنجی خارجی افرازهای تولیدشده با روش‌های اجماع، از یک شاخص اعتبار خوشه خارجی به‌نام AMI[۲] استفاده می‌شود [۲۳]. یک شاخص خارجی تطابق

[۱] decay

[۲] Adjusted Mutual Information

[DOI: 10.61186/jsdp.20.1.99]

بین راه‌حل خوشه‌بندی و اجماع بهینهٔ شناخته‌شده را اندازه‌گیری می‌کند، که باید به‌عنوان یک حقیقت پایه یا استاندارد برای یک مجموعه داده خاص ارائه شود. AMI یک اندازهٔ شباهت است که بر اساس اطّلاعات متقابل طبیعی (NMI) معروف [۲۳] ساخته شده است. ازآن‌جا-که، تعداد نقاط داده در مقایسه با تعداد خوشه‌ها به‌نسبت کم است، تنظیم شانس هنگام کار با داده‌های ریزآرایه ضروری است. بنابراین، ما AMI را به‌عنوان رتبهٔ عملکرد الگوریتم‌ها به‌جای NMI یا شاخص رند تنظیم شده [۲۳] انتخاب کرده‌ایم. مقدار AMI زمانی است که دو افراز به‌طور کامل با هم مطابقت داشته باشند، درحالی‌که افرازهای تصادفی به‌طور متوسط دارای AMI حدود صفر هستند؛ بنابراین AMI می‌تواند منفی باشد. رتبه‌ها با استفاده از میانگین بیش از ۳۰ اجرای مستقل جمع بندی شده‌اند. میانگین و انحراف معیار رتبه AMI در مجموعه داده‌های واقعی (مجموعه داده جدول ۵) به‌طور متوسط بیش از ۳۰ اجرا برای روش‌های اجماع در مقایسه با روش پیشنهادی در جدول ۶ آورده شده است. بهترین رتبه برای هر مجموعه داده پررنگ مشخّص شده است. نتایج جدول ۶ حاکی از برتری روش پیشنهادی نسبت‌به سایر روش‌های رقیب است.

در این بخش روش پیشنهادی از نظر معیار دقّت با روش‌های جدید مقایسه شده است. مجموعه داده‌های استفاده‌شده در این بخش در جدول ۷ آورده شده‌اند.

نتایج جدول ۸ نشان می‌دهد که روش پیشنهادی نسبت‌به روش‌های رقیب کارآیی مناسب‌تری را از خود نشان داده است. مقایسه بر اساس آزمون Wilcoxon انجام شده است. در این‌جا، روش‌های مبتنی بر پیوند شامل پیوند تک، پیوند متوسّط، پیوند مرکزی و پیوند کامل، و همچنین، الگوریتم‌های خوشه‌بندی جمعی پیشرفته از جمله HMM [۳۴]، DSPA [۳۵] و WHAC [۳۶] برای مقایسه استفاده می‌شوند. آزمون جمع رتبهٔ ویلکاکسون برای تأیید عملکرد و عدم‌تشابه طرح پیشنهادی با سایر الگوریتم‌های قابل‌مقایسه انجام می‌شود. برای این آزمایش، تمام روش‌های مبتنی بر پیوند و همچنین، الگوریتم‌های HMM، DSPA و WHAC برای تحلیل p-value مقایسه می‌شوند. ما یک آستانهٔ قابل‌توجّه مشابه مطالعات قبلی را ۰،۰۵ در نظر می گیریم [۳۷]. بنابراین، آزمون مجموع رتبهٔ ویلکاکسون با استفاده از فرضیه صفر در سطح معناداری ۰،۰۵٪ (یعنی سطح اطمینان ۹۵٪) تجزیه و تحلیل می‌شود. ما مقدار p همهٔ الگوریتم‌ها را بر اساس شاخص Silhouette محاسبه می‌کنیم [۳۸]، که معیاری از ضریب کیفیّت خوشه‌بندی بر اساس تفاوت زوجی بین و درون‌خوشه‌ای است. نتایج این آزمون در $\alpha = 0.05$ برای طرح پیشنهادی در مقایسه با سایر الگوریتم ها در جدول ۸ گزارش شده است.

(جدول- ۸): مقایسه روش پیشنهادی با سایر الگوریتم‌ها در آزمون مجموع رتبه Wilcoxon با $\alpha = 0.05$

Table 8: Comparison of MCEMS with other algorithms in Wilcoxon rank sum test with$\alpha = 0.05$.

Dataset	Single		Average		Centroid		Complete		HMM		DSPA		WHAC	
	p-value	SG	*p*-value	SG	*p*-value	SG	*p*-value	SG	*p*-value	SG	*p*-value	SG	*p*-value	SG
DS1	0.0020	-	0.0021	-	0.0023	-	0.0156	-	0.0227	-	0.0163	-	0.0181	-
DS2	0.0010	-	0.0013	-	0.0006	-	0.0008	-	0.0055	-	0.0013	-	0.0020	-
DS3	0.0001	-	0.0001	-	0.0001	-	0.0000	-	0.0015	-	0.0013	-	0.0001	-
DS4	0.0008	-	0.0006	-	0.0004	-	0.0008	-	0.0016	-	0.0028	-	0.0008	-
DS5	0.0009	-	0.0005	-	0.0017	-	0.0019	-	0.0858	+	0.0413	-	0.0021	-
DS6	0.0002	-	0.0001	-	0.0001	-	0.0002	-	0.0018	-	0.0014	-	0.0002	-
DS7	0.0005	-	0.0005	-	0.0005	-	0.0005	-	0.0010	-	0.0035	-	0.0008	-
DS8	0.0001	-	0.0001	-	0.0001	-	0.0001	-	0.0015	-	0.0013	-	0.0001	-
DS9	0.0005	-	0.0005	-	0.0007	-	0.0008	-	0.0070	-	0.0013	-	0.0027	-
DS10	0.0006	-	0.0006	-	0.0006	-	0.0008	-	0.0011	-	0.0188	-	0.0085	-
DS11	0.0003	-	0.0009	-	0.0005	-	0.0004	-	0.0010	-	0.0016	-	0.0011	-
DS12	0.0001	-	0.0004	-	0.0004	-	0.0003	-	0.0010	-	0.0120	-	0.0051	-
DS13	0.0002	-	0.0000	-	0.0002	-	0.0001	-	0.0010	-	0.0013	-	0.0002	-
DS14	0.0016	-	0.0011	-	0.0014	-	0.0011	-	0.0126	-	0.0017	-	0.0527	+
DS15	0.0010	-	0.0008	-	0.0011	-	0.0005	-	0.0010	-	0.0059	-	0.0020	-
DS16	0.0002	-	0.0002	-	0.0002	-	0.0002	-	0.0011	-	0.0021	-	0.0002	-
DS17	0.0011	-	0.0015	-	0.0016	-	0.0013	-	0.0033	-	0.0368	-	0.0186	-
DS18	0.0008	-	0.0009	-	0.0004	-	0.0007	-	0.0405	-	0.0013	-	0.0245	-
DS19	0.0001	-	0.0004	-	0.0002	-	0.0002	-	0.0018	-	0.0020	-	0.0005	-
DS20	0.0049	-	0.0042	-	0.0032	-	0.0067	-	0.0294	-	0.0073	-	0.0169	-
DS21	0.0014	-	0.0021	-	0.0024	-	0.0017	-	0.0370	-	0.0025	-	0.0183	-
DS22	0.0459	-	0.0552	-	0.0484	-	0.0488	-	0.3307	+	0.3270	+	0.2174	+
DS23	0.0004	-	0.0000	-	0.0002	-	0.0003	-	0.0010	-	0.0201	-	0.0091	-
DS24	0.0023	-	0.0022	-	0.0023	-	0.0024	-	0.0111	-	0.0075	-	0.0028	-
DS25	0.0023	-	0.0020	-	0.0031	-	0.0029	-	0.0256	-	0.0034	-	0.0118	-
Average	0.0020	-	0.0019	-	0.0021	-	0.0024	-	0.0251	-	0.0209	-	0.0166	-

[DOI: 10.61186/jsdp.20.1.99]

(جدول- ۷): شرح مجموعه دادهٔ استفاده‌شده

Table 7: Description of the dataset used.

Index	Dataset	#Instances	#Features	#Classes
DS1	Wine	178	13	3
DS2	Vehicle	846	18	4
DS3	Seeds	210	7	3
DS4	Pima-diabetes	768	8	2
DS5	Landsat-satellite	6435	36	6
DS6	Image-segmentation	2310	19	7
DS7	Mammographic	961	5	2
DS8	Glass	214	10	7
DS9	Bupa	323	4	7
DS10	Breast	683	9	2
DS11	Avila	20,867	10	12
DS12	Yeast	1484	8	10
DS13	Ecoli	336	8	8
DS14	Digits	5620	63	10
DS15	Banana	2000	2	2
DS16	Ring3	1500	2	3
DS17	Imbalance	2250	2	2
DS18	Bupa	345	6	2
DS19	Aggregation	788	2	7
DS20	SAHeart	462	9	2
DS21	Ionosphere	351	34	2
DS22	Galaxy	323	4	7
DS23	Half-Ring	400	2	2
DS24	Hand-made1	300	2	3
DS25	Letter-recognition	20,000	16	26

۵- نتیجه‌گیری و پیشنهادهای آینده

در این بخش به بررسی نتایج به‌دست‌آمده از اجرای روش پیشنهادی و مقایسهٔ آن با روش‌های پیشین و در نهایت به جمع‌بندی و ارائه پیشنهاد، پرداخته خواهد شد.

۵-۱- نتیجه‌گیری

در این مقاله، یک روش جدید برای خوشه‌بندی ترکیبی پیشنهاد شد که از یک توزیع احتمال برای توصیف احتمال تعلّق نمونه‌ها به خوشه‌ها استفاده می‌کرد. هدف، تعیین مقدار صحیح برای این توزیع‌های احتمال بود که در همین راستا یک تابع هدف تعیین و مسأله خوشه‌بندی تبدیل به یک مسألهٔ بهینه‌سازی شد. برای حل مسألهٔ بهینه‌سازی، از روش تابکاری شبیه‌سازی شده استفاده شد. روش پیشنهاد دارای چند ویژگی مهم است که عبارتند از:

- از تابکاری فلزات در آن استفاده می‌شود. این الگوریتم، یکی از الگوریتم‌های مؤثّر در حوزهٔ جستجوی محلی است که به‌طور موثرتری نسبت‌به بسیاری از روش‌های مشابه کار می‌کند و برخلاف برخی از آن‌ها مانند تپه‌نوردی، می‌توان از نقاط بهینهٔ محلی بگریزد.
- از یک روش جدید برای تعیین حالت آغازین در الگوریتم پیشنهادی استفاده می‌کند. این روش، با توجّه به برچسب‌های حاصل از خوشه‌بندی اولیّهٔ داده‌ها، توزیع‌های احتمالی را چنان تعیین می‌کند که تا حدودی نتایج خوشه‌بندی در آن منعکس شده باشد. اینکه میزان انعکاس چقدر باشد، از طریق یک شاخص در روش پیشنهادی قابل‌کنترل است.
- از یک روش جدید و هوشمندانه برای رفتن به حالت بعدی استفاده می‌کند. در این روش که از طریق چند شاخص قابل کنترل است، سعی می‌شود با تمرکز روی برخی درایه‌ها از ماتریس همبستگی، توزیع‌های احتمال تعلق نمونه‌ها به خوشه‌ها به‌صورت کنترل‌شده و نرم تغییر کنند و یک توزیع احتمال جدید، امّا نزدیک به قبلی تولید شود. میزان نزدیکی حالت قبلی و حالت جدید از طریق چند شاخص کنترل می‌شود که مشخّص می‌کنند روی کدام درایه‌های ماتریس همبستگی باید تمرکز شود و تعداد تغییرات در حالت کنونی چندتا است.

در آزمایش‌های تجربی، روش پیشنهادی با چهار روش خوشه‌بندی منفرد و دو روش خوشه‌بندی ترکیبی مقایسه شد. نتایج تجربی نشان می‌دهند که روش پیشنهادی به‌طور عموم افرازبندی‌های با کیفیت‌تری نسبت به سایر روش‌ها تولید می‌کند. بخشی از نتایج تجربی نشان می‌دهد که روش پیشنهادی بهتر از سایر روش‌ها قادر به شناسایی خوشه‌هایی با توزیع نرمال است.

۵-۲- ارائهٔ پیشنهاد

در روش پیشنهادی، خوشه‌بندی به‌صورت یک مسأله بهینه‌سازی تعریف شد. مهم‌ترین مؤلّفه‌ها در یک مسألهٔ بهینه‌سازی، تابع هدف و ابزار بهینه‌سازی است. به‌این‌ترتیب، می‌توان توابع هدفی مشابه تعریف کرد که نقطهٔ اشتراک آنها با تابع مورداستفاده در این مقاله، استفاده از توزیع احتمال تعلّق نمونه‌ها به خوشه‌ها باشد. در روش پیشنهادی از الگوریتم تابکاری فلزات برای بهبود

[DOI: 10.61186/jsdp.20.1.99]

خوشه‌بندی‌ها استفاده شد. ازآن‌جاکه روش‌های تکاملی بسیار گوناگون هستند، به‌کارگیری سایر روش‌های تکاملی از جمله الگوریتم ازدحام ذرات و کلونی زنبور مصنوعی نیز می‌تواند در دستور کارهای آینده قرار گیرد.

قدردانی

این مقاله مستخرج از رساله دکتری خانم سیده فروزان رشیدی با راهنمایی آقای دکتر صمد نجاتیان و آقای دکتر حمید پروین و مشاوره خانم دکتر سیده وحیده رضایی و آقای دکتر کرم‌الله باقری‌فرد است.

۶- مراجع 6-Refrence

[۱] ف. نجفی، یک روش خوشه‌بندی ترکیبی جدید مبتنی بر خوشه‌بند Cmeans فازی با حفظ تنوع در اجماع، نشریه علمی-پژوهشی پردازش علائم و داده‌ها، شماره ۴، پیاپی ۴۶، صفحات ۱۲۱-۱۰۳، ۱۳۹۹.

[۲] ص. عباسی، خوشه‌بندی ترکیبی با بیشینه‌سازی پراکندگی با به کارگیری الگوریتم‌های بهینه‌سازی تکاملی، نشریه علمی-پژوهشی پردازش علائم و داده‌ها، شماره ۴، پیاپی ۵۴، صفحات ۱۲۰-۹۵، ۱۴۰۱ .

S. Abbasi, S. Nejatian, H. Parvin, Karamollah Bagherifard and V. Rezaie, " The ensemble clustering with maximize diversity using evolutionary optimization algorithms", Signal and Data Processing, Vol. 4, No. 54, pp. 95-120, 2023.

[3] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. D. F. Costa, and F. A. Rodrigues, ``Clustering algorithms:Acomparative approach," PLoS ONE, vol. 14, no. 1, Jan. 2019, Art. no. e0210236.

[4] A. Ahmad and S. S. Khan, ``Survey of state-of-the-art mixed data clustering lgorithms," IEEE Access, vol. 7, pp. 31883_31902, 2019.

[5] J. Kim, J. Yoon, E. Park, and S. Choi, ``Patent document clustering with deep embeddings," Scientometrics, vol. 123, no. 2, pp. 563_577, May 2020.

[6] E. Zanboori, M. Rostamy-Malkhalifeh, G.R. Jahanshahloo and N. Shoja, Calculating Super Efficiency of DMUs for Ranking Units in Data Envelopment Analysis Based on SBM Model, The Scientific World Journal, 2014.

[7] E. Zanboori, F. Hosseinzadeh Lotfi, M. Rostamy-Malkhalifeh and G.R. Jahanshahloo, Computing Relative weights in AHP and Ranked Units in the Presence of Large Dimensionality of data set based on Orthogonal Gram Schmidt Technique. Adv. Environ. Biol., 8(21), 78-81, 2014.

[8] Bertsimas, Dimitris, and John Tsitsiklis. "Simulated annealing." Statistical science 8.1 (1993): 10-15.

[9] P. Govender and V. Sivakumar, ``Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980_2019)," Atmos. Pollut. Res., vol. 11, no. 1, pp. 40_56, Jan. 2020.

[10] C. Zong, S. Huang, E. Liu, Y. Yao, and S.-Q. Tang, ``Nowhere to hide methodology: Application of clustering fault diagnosis in the nuclea power industry," IEEE Access, vol. 7, pp. 179864_179879, 2019.

[11] Seni, G. and J. Elder, Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. 2010: Morgan and Claypool Publishers. 126.

[12] Fred, A.L.N. and A.K. Jain, Combining multiple clusterings using evidence accumulation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005. 27(6): p. 835-850.

[13] Fred, A.L.N. and A.K. Jain. Data clustering using evidence accumulation. in Pattern Recognition, 2002. Proceedings. 16th International Conference on. 2002.

[14] Fred, A. and A. Jain, Evidence Accumulation Clustering Based on the K-Means Algorithm, in Structural, Syntactic, and Statistical Pattern Recognition, T. Caelli, et al., Editors. 2002, Springer Berlin Heidelberg. p. 442-451.

[15] Jain, A.K. and R.C. Dubes, Algorithms for clustering data. 1988: Prentice-Hall, Inc. 320.

[16] Duda, R.O., P.E. Hart, and D.G. Stork, Pattern Classification (2nd Edition). 2000: Wiley-Interscience.

[17] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, ``A review of unsupervised feature selection methods," Artif. Intell. Rev., vol. 53, no. 2, pp. 907_948, Feb. 2020.

[18] I. Guyon and A. Elisseeff, ``An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157_1182, Jan. 2003.

[19] G. Chandrashekar and F. Sahin, ``A survey on feature selection methods,"Comput. Electr. Eng., vol. 40, no. 1, pp. 16_28, Jan. 2014.

[20] E. Hancer, B. Xue, and M. Zhang, ``A survey on feature selection approaches for clustering," Artif. Intell. Rev., pp. 1_27, doi: 10.1007/s10462-019-09800-w.

[21] B. Kim, H. Lee, and P. Kang, ``Integrating cluster validity indices based on data envelopment analysis," Appl. Soft Comput., vol. 64, pp. 94_108, Mar. 2018.

[22] D. Huang, X. Cai, and C.-D. Wang, ``Unsupervised feature selection with multi-subspace randomization and collaboration," Knowl. Based Syst., vol. 182, Oct. 2019, Art. no. 104856.

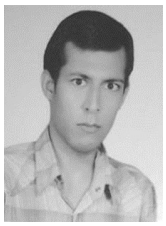
[23] H. Zhang, R. Zhang, F. Nie, and X. Li, ``An ef_cient framework for unsupervised feature selection," Neurocomputing, vol. 366, pp. 194_207, Nov. 2019.

[DOI: 10.61186/jsdp.20.1.99]

[24] D. Huang, C.-D. Wang, J.-S. Wu, J.-H. Lai, and C.-K. Kwoh, ``Ultrascalable spectral clustering and ensemble clustering," IEEE Trans. Knowl. Data Eng., vol. 32, no. 6, pp. 1212_1226, Jun. 2020.
[25] S.-O. Abbasi, S. Nejatian, H. Parvin, V. Rezaie, and K. Bagherifard, ``Clustering ensemble selection considering quality and diversity," Artif. Intell. Rev., vol. 52, no. 2, pp. 1311_1340, Aug. 2019.
[26] X. Wang, C. Yang, and J. Zhou, ``Clustering aggregation by probability accumulation," Pattern Recognit., vol. 42, no. 5, pp. 668_675, May 2009.
[27] S. Vega-Pons and J. Ruiz-Shulcloper, ``Clustering ensemble method for heterogeneous partitions," in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (Lecture Notes in Computer Science), vol. 5856, E. Bayro-Corrochano and J.-O. Eklundh, Eds. Berlin, Germany: Springer, 2009, pp. 481_488.
[28] N. Iam-on, T. Boongoen, and S. Garrett, ``LCE: A link-based cluster ensemble method for improved gene expression data analysis,"Bioinformatics, vol. 26, no. 12, pp. 1513_1519, Jun. 2010.
[29] D. Huang, J.-H. Lai, and C.-D.Wang, ``Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis," Neurocomputing, vol. 170, pp. 240_250, Dec. 2015.
[30] H. Wang, Y. Yang, B. Liu, H. Fujita, "A study of graph-based system for multi-view clustering". Knowl-Based Syst 163:1009–1019, 2019.
[31] C. Fahy, S. Yang, M. Gongora, "Ant Colony stream clustering: a fast density clustering algorithm for dynamic data streams". IEEE Trans Cybern, vol. 49, no. 6, pp. 2215–2228, 2019.
[32] M. Mojarad, S. Nejatian, H. Parvin, M. Mohammadpoor, "A fuzzy clustering ensemble based on cluster clustering and iterative fusion of base clusters". Appl Intell, vol. 49, no.7, pp. 2567–2581, 2019.
[33] T. Lai, R. Chen, C. Yang, Q. Li, H. Fujita, A. Sadri, H. Wang, "Efficient robust model fitting for multistructure data using global greedy search". IEEE Trans Cybern, vol. 50, no. 7, pp. 3294–3306, 2020.
[34] Y. Yang, J. Jiang, Adaptive bi-weighting toward automatic initialization and model selection for HMM-based hybrid meta-clustering ensembles. IEEE Trans. Cybern, vol. 49, no. 5, pp. 1657–1668, 2018a.
[35] Y. Yang, J. Jiang., Bi-weighted ensemble via HMM-based approaches for temporal data clustering. Pattern Recogn. vol. 76, pp. 391–403. 2018b.
[36] A. Banerjee, A.K. Pujari, C. Rani Panigrahi, B. Pati, S. Chandan Nayak, T.H. Weng A new method for weighted ensemble clustering and coupled ensemble selection. Connect. Sci. vol. 33, no. 3 pp. 623–644, 2021.
[37] M. Jafarzadegan, F. Safi-Esfahani, Z. Beheshti, Combining hierarchical clustering approaches using the PCA method. Expert Syst. Appl. vol. 137, pp. 1–10, 2019.
[38] M. Mojarad, F. Sarhangnia, A. Rezaeipanah, H. Parvin, S. Nejatian, Modeling hereditary disease behavior using an innovative similarity criterion and ensemble clustering. Curr. Bioinform, vol. 16, no. 5, pp. 749 764, 2021

سیده فروزان رشیدی تحصیلات خود را در مقطع کارشناسی ارشد، در رشته مهندسی کام پیوتر در دانشگاه علوم و تحقیقات و مقطع دکترای تخصصی را در رشته مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد یاسوج به پایان رساند. وی هم اکنون عضو هیأت علمی دانشگاه آزاد اسلامی واحد امیدیه است. زمینه‌های پژوهشی ایشان، طبقه‌بندی و خوشه‌بندی داده‌هاست.
نشانی رایانامهٔ ایشان عبارت است از:
foroozan.rashidi@iau.ac.ir

صمد نجاتیان دکترای تخصصی خود را در رشتهٔ برق (مخابرات)، در سال ۱۳۹۳ از دانشگاه UTM مالزی گرفت. وی دانشیار و عضو هیئت علمی دانشگاه آزاد اسلامی واحد یاسوج است . از جمله سوابق ایشان معاونت پژوهش و فناوری دانشگاه آزاد اسلامی واحد یاسوج و ریاست باشگاه پژوهشگران جوان دانشگاه آزاد اسلامی واحد یاسوج است. زمینه‌های کاری ایشان برق، مخابرات، متن¬کاوی، پردازش سیگنال، خوشه‌بندی داده‌ها و هوش مصنوعی است .
نشانی رایانامهٔ ایشان عبارت است از:
samad.nejatian.ir@ieee.org



حمید پروین تحصیلات خود در مقطع کارشناسی را در دانشگاه چمران اهواز به پایان رساند. ایشان مدرک کارشناسی ارشد و دکترا را در دانشگاه علم و صنعت گرفت و پس از آن به عضویّت هیئت علمی دانشگاه آزاد اسلامی واحد نورآباد ممسنی درآمد. وی هم‌اکنون در چندین واحد دانشگاهی در رشته کامپیوتر مشغول به تدریس است. زمینه‌های پژوهشی وی مباحثی

[DOI: 10.61186/jsdp.20.1.99]

نظیر الگوریتم‌های بهینه‌سازی، طبقه‌بندی و خوشه‌بندی داده‌هاست.

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

وحیده رضایی دارای مدرک تحصیلی در مقطع دکترای تخصصی رشته ریاضیات است. وی هم اکنون عضو هیئت علمی دانشگاه آزاد اسلامی واحد یاسوج است. زمینه¬های پژوهشی وی، بهینه‌سازی ریاضی، متن‌کاوی پردازش سیگنال و خوشه‌بندی داده‌هاست.

نشانی رایانامهٔ ایشان عبارت است از:
vahidehrezaie@gmail.com

کرم الله باقری فرد مدرک کارشناسی خود را در رشتهٔ مهندسی کامپیوتر از دانشگاه اصفهان و مدرک کارشناسی ارشد خود را به‌ترتیب از دانشگاه نجف آباد و اراک دریافت کرده است. وی هم اکنون عضو هیئت علمی دانشگاه آزاد اسلامی واحد یاسوج است. زمینه‌های پژوهشی ایشان داده‌کاوی، یادگیری ماشین و سامانه‌های پیشنهاددهنده است.

نشانی رایانامهٔ ایشان عبارت است از:
k.bagherifard@iauyasooj.ac.ir

[DOI: 10.61186/jsdp.20.1.99]