انتخاب اعضای ترکیب در خوشه‌بندی ترکیبی با استفاده از رأی‌گیری

علیرضا لطیفی پاکدهی و نگین دانشپور*

دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران

چکیده

خوشه‌بندی ترکیبی، به ترکیب نتایج حاصل از خوشه‌بندی‌های موجود می‌پردازد. پژوهش‌های دهۀ اخیر نشان می‌دهد، چنان‌چه به جای ترکیب همۀ خوشه‌بندی‌ها، تنها دسته‌ای از آن‌ها بر اساس کیفیت و تنوع انتخاب شوند، آن‌چه به‌عنوان خروجی خوشه‌بندی ترکیبی حاصل می‌شود، بسیار دقیق‌تر خواهد بود. این مقاله به ارائه یک روش جدید برای انتخاب خوشه‌بندی‌ها بر اساس دو معیار کیفیت و تنوع می‌پردازد. برای رسیدن به این منظور ابتدا خوشه‌بندی‌های مختلفی با استفاده از الگوریتم k-means ایجاد می‌شود که در هر بار اجرا، مقدار k یک عدد تصادفی است. در ادامه خوشه‌بندی‌هایی که به این نحو تولید شده‌اند، با استفاده از الگوریتم جدیدی که براساس میزان شباهت بین خوشه‌بندی‌های مختلف عمل می‌کند، گروه‌بندی می‌شوند تا آن‌دسته از خوشه‌بندی‌هایی که به یکدیگر شبیه‌اند در یک دسته قرار گیرند؛ سپس از هر دسته، با استفاده از یک روش مبتنی بر رأی‌گیری، با کیفیت‌ترین عضو آن برای ایجاد خوشه‌بندی ترکیبی انتخاب می‌شود. در این مقاله از سه تابع HPGA، CSPA و MCLA برای ترکیب خوشه‌بندی‌ها استفاده شده است. در انتها برای آزمایش این روش جدید از داده‌های واقعی موجود در پایگاه داده UCI استفاده شده است. نتایج نشان می‌دهد که روش جدید کارایی بیشتر و دقیق‌تری نسبت به روش‌های قبلی دارد.

واژگان کلیدی: خوشه‌بندی ترکیبی، انتخاب اعضا، شاخص‌های ارزیابی کیفیت

Cluster ensemble selection using voting

Alireza Latifi-Pakdehi & Negin Daneshpour*

Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran

Abstract

Clustering is the process of division of a dataset into subsets that are called clusters, so that objects within a cluster are similar to each other and different from objects of the other clusters. So far, a lot of algorithms in different approaches have been created for the clustering. An effective choice (can combine) two or more of these algorithms for solving the clustering problem. Ensemble clustering combines results of existing clusterings to achieve better performance and higher accuracy. Instead of combining all of existing clusterings, recent decade researchers show, if only a set of clusterings is selected based on quality and diversity, the result of ensemble clustering would be more accurate. This paper proposes a new method for ensemble clustering based on quality and diversity. For this purpose, firstly first we need a lot of different base clusterings to combine them. Different base clusterings are generated by k-means algorithm with random k in each execution. After the generation of base clusterings, they are put into different groups according to their similarities using a new grouping method. So that clusterings which are similar to each other are put together in one group. In this step, we use normalized mutual information (NMI) or adjusted rand index (ARI) for computing similarities and dissimilarities between the base clustering. Then from each group, a best qualified clustering is selected via a voting based method. In this method, Cluster-validity-indices were used to measure the quality of clustering. So that all members of the group are evaluated by the Cluster-validity-indices. In each group, clustering that optimizes the most number of Cluster-validity-indices is selected. Finally, consensus functions combine all selected clustering. Consensus function is an algorithm for combining

* Corresponding author

* نویسندۀ عهده‌دار مکاتبات • تاریخ ارسال مقاله: ۱۳۹۵/۱۰/۱۱ • تاریخ آخرین بازنگری: ۱۳۹۷/۸/۲۲ • تاریخ پذیرش: ۱۳۹۷/۱۰/۱۹

existing clusterings to produce final clusters. In this paper, three consensus functions including CSPA, MCLA, and HGPA have used for combining clustering. To evaluate proposed method, real datasets from UCI repository have used. In experiment section, the proposed method is compared with the well-known and powerful existing methods. Experimental results demonstrate that proposed algorithm has better performance and higher accuracy than previous works.

Keywords: Ensemble clustering, select member, validity index.

۱- مقدمه

خوشه‌بندی، فرآیندی است که در آن مجموعه‌ای از داده‌ها به گروه‌های متفاوتی (که خوشه نامیده می‌شوند) افراز می‌شوند؛ به‌طوری‌که داده‌های موجود در یک گروه، در ویژگی‌ها یا خصوصیاتی، اشتراک دارند که ممکن است، در داده‌های دیگر گروه‌ها موجود نباشند[1]. خوشه‌بندی به‌طور گسترده در رشتهٔ هوش مصنوعی مورد استفاده قرار گرفته است [2]. در اغلب موارد هدف از خوشه‌بندی فهم بهتر توزیع داده‌هاست؛ ولی در بعضی موارد گام نخست برای پردازش‌های بعدی نظیر اندیس‌گذاری یا فشرده‌سازی است [3]. خوشه‌بندی در بسیاری از زمینه‌ها مثل ستاره‌شناسی، فیزیک، داروسازی و بازاریابی مورد استفاده قرار گرفته است.

خوشه‌بندی ترکیبی[1] نتایج حاصل از چندین خوشه‌بندی را برای به‌دست‌آوردن خوشه‌بندی دقیق‌تر، باکیفیت‌تر و مقیاس‌پذیرتر ترکیب می‌کند. در خوشه‌بندی ترکیبی نیاز است تا به جای خودِ داده فقط به نتایج خوشه‌بندی‌های پایه دسترسی داشته باشیم؛ بنابراین خوشه‌بندی ترکیبی رویکردی مناسب برای تأمین خصوصی‌سازی و استفاده مجدد از دانش است [4]. همچنین خوشه‌بندی ترکیبی این امکان را فراهم می‌کند تا با اجماع نتایج اجرای چندین‌بارهٔ یک الگوریتم خوشه‌بندی، به نتایج پایدارتری دست پیدا کنیم [5]. خوشه‌بندی ترکیبی می‌تواند امکان استفاده از فناوری پردازش موازی را نیز فراهم آورد [6]. خوشه‌بندی ترکیبی کاربردهایی در بیوانفورماتیک، پردازش تصویر و بازاریابی دارد [9-7].

روش‌های سنتی خوشه‌بندی ترکیبی خوشه‌بندی‌های متعددی از مجموعه‌داده تولید می‌کردند و سپس با ترکیب همهٔ آن‌ها، خوشه‌بندی نهایی تولید می‌شد. برای ایجاد خوشه‌بندی‌های متفاوت، رویکردهای متفاوتی وجود دارد [10]:

- **مقداردهی اولیه پارامترها به‌صورت متفاوت:** در هر بار اجرا مقدار پارامتر اولیه از یک مجموعه انتخاب می‌شود [11].
- **الگوریتم‌های خوشه‌بندی متفاوت:** تعدادی از الگوریتم‌های خوشه‌بندی متفاوت برای ایجاد خوشه‌بندی‌های پایه مورد استفاده قرار می‌گیرند [12] و [13].
- **زیرمجموعه‌ای از صفات مختلف:** زیرمجموعه‌های مختلف از صفات برای تولید خوشه‌بندی‌های مختلف انتخاب می‌شوند [11] و [14].
- **زیرمجموعه‌ای از اشیای مختلف:** در هر بار زیرمجموعه‌های متفاوت از داده‌ها برای تولید خوشه‌بندی‌های مختلف انتخاب می‌شوند [15].

برای نخستین‌بار [16] نشان داد که در یادگیری ترکیبی[2]، انتخاب اعضای ترکیب از ترکیب کامل ممکن است بهتر باشد (در ترکیب کامل، همهٔ اعضای یادگیری ترکیب، اما در انتخاب اعضای ترکیب، اعضا قبل از ترکیب انتخاب می‌شوند. اجزای یادگیری نیز با توجه به نوع یادگیری متفاوت است. برای مثال در ترکیب طبقه‌بندی‌ها، اجزای یادگیری، طبقه‌بندی‌های پایه هستند که قرار است ترکیب شوند. در خوشه‌بندی ترکیبی، اجزای یادگیری، خوشه‌بندی‌های پایه هستند که قرار است با یکدیگر ترکیب شوند). در همین اواخر روش‌های خوشه‌بندی جدیدی ایجاد شده که اعضای ترکیب بر اساس تنوع و کیفیت انتخاب می‌شوند. این روش‌ها با عنوان انتخاب اعضای ترکیب(CES)[3] شناخته می‌شوند [17]. CES برای بهبود کارایی خوشه‌بندی ترکیبی پیشنهاد شده است [18] و برای انتخاب زیرمجموعه‌ای از مجموعهٔ بزرگ خوشه‌بندی‌ها به‌کار می‌رود تا مجموعه‌ای کوچک‌تر حاصل شود؛ به‌نحوی که نتایج مجموعهٔ جدید به‌خوبی ترکیب کامل یا حتی بهتر از آن باشد. [19, 20].

تاکنون تعداد کمی مطالعه روی این مسأله انجام شده که چگونه زیرمجموعه‌ای از خوشه‌بندی‌ها باید براساس کیفیت و تنوع انتخاب شوند [23-21]. روش‌هایی که تاکنون در این زمینه ارائه شده‌اند، نیازمند یک پارامتر ورودی جهت تعیین تعداد خوشه‌بندی‌های انتخاب‌شده از مجموع کل تعداد خوشه‌بندی‌ها هستند. یک راه برای عبور از چنین معضلی، آزمایش بازهٔ وسیعی از این پارامتر و نشان‌دادن خروجی به‌ازای

[1] Ensemble clustering

[2] Ensemble learning

[3] Cluster Ensemble Selection

هر مقدار از این پارامتر است؛ اما این یک راه حل کلی محسوب نمی‌شود؛ بنابراین نیاز به حذف این پارامتر، یک نیاز جدّی است که روش پیشنهادی این مقاله، این معضل را برطرف کرده است.

این مقاله یک روش ترکیبی جدید بر اساس تنوع و کیفیت ارائه می‌دهد. در نخستین گام، خوشه‌بندی‌های متعددی از مجموعه‌داده تولید می‌شود. این خوشه‌بندی‌ها با استفاده از اجرای چندین بار الگوریتم k-means ایجاد می‌شوند که در هر بار اجرا مقدار k، خروجی یک تابع تصادفی است؛ سپس این خوشه‌بندی‌ها گروه‌بندی می‌شوند و در هر گروه نیز با کیفیت‌ترین خوشه انتخاب می‌شود. در این روش نیازی به بیان تعداد خوشه‌بندی‌های انتخاب‌شده نیست و گروه‌بندی و انتخاب بدون نیاز به این عدد انجام می‌شود. در روش ارائه‌شده، برای انتخاب از هر گروه، راه‌کار جدیدی براساس رأی‌گیری روی شاخص‌های ارزیابی کیفیت[1] پیشنهاد شده است. بدین ترتیب اندازهٔ مجموعه‌داده با حذف خوشه‌بندی‌هایی که درعمل در نتیجهٔ نهایی تأثیر مثبتی نداشتند، کاهش می‌یابد و درنهایت این مجموعه‌دادهٔ کاهش‌یافته برای تولید خوشه‌بندی نهایی در اختیار توابع اجماع[2] قرار می‌گیرد.

بقیهٔ مطالب مقاله به این‌صورت است: در بخش دو به تبیین مفاهیم پرکاربرد در مقاله و در بخش ۳ پیشینهٔ پژوهش و کارهای مشابه تشریح شده و در بخش ۴ به بیان روش پیشنهادی و چارچوب کلی راه حل پرداخته شده است. دربخش ۵ نتایج راه حل پیشنهادی روی داده‌های واقعی برگرفته از UCI نشان داده و درنهایت نتیجه‌گیری در بخش ۶ ذکر شده است.

۲- مفاهیم پایه

در این بخش به شرح دو مفهوم پراستفاده در این مقاله یعنی تابع اجماع و معیارهای شباهت و تفاوت پرداخته و سپس در بخش بعد، روش پیشنهادی بیان می‌شود.

۲-۱- تابع اجماع

تابع اجماع کار اصلی در خوشه‌بندی ترکیبی را بر عهده دارد. تابع اجماع خوشه‌بندی‌های متعددتولیدشده را به‌عنوان ورودی می‌گیرد و با ترکیب آن‌ها، خوشه‌بندی نهایی را تولید می‌کند. در [4] سه مورد توابع اجماع مبتنی بر گراف معرّفی شده که عبارتند از: الگوریتم بخش‌بندی شباهت مبتنی بر خوشه[3] (CSPA)، الگوریتم بخش‌بندی ابر گراف[4] (HGPA)، الگوریتم ابر خوشه‌بندی[5] (MCLA). به‌طوراساسی اگر دو شیء در یک خوشه باشند، مشابه در نظر گرفته می‌شوند و در غیر این‌صورت غیر مشابه هستند. بر اساس این دیدگاه در CSPA یک ماتریس شباهت برای هر یک از خوشه‌بندی‌ها ساخته می‌شود. در ماتریس شباهت اگر عنصر مربوط به دو اندیس در یک خوشه باشند، درایه متناظر یک و در غیر این‌صورت صفر خواهد بود. بعد از محاسبه تمامی ماتریس‌های شباهت مربوط به تمام خوشه‌بندی‌ها، ماتریس نهایی از میانگین درایه‌های ماتریس شباهت به‌دست می‌آید؛ سپس از الگوریتم METIS [24] استفاده می‌شود تا خوشه‌بندی نهایی حاصل شود.

در تابع اجماع HGPA مسأله خوشه‌بندی ترکیبی به مسأله بخش‌بندی ابرگراف به‌وسیلهٔ حذف کمترین تعداد ابریال تبدیل می‌شود. بنابراین لازم است، ابرگراف را تعریف کنیم. یک ابر گراف، گرافی است که رأس‌هایش از مجموعه‌ای از اشیای مجموعه‌داده تشکیل شده است و یال‌هایش ارتباط بین رأس‌ها را (با درنظرگرفتن خوشه‌بندی) معین می‌کنند. این یال‌ها دارای وزنی یکسان هستند؛ سپس از الگوریتم HMETIS [25] برای شکستن ابرگراف و تولید خوشه‌بندی نهایی استفاده می‌شود.

در تابع اجماع MCLA بر خلاف CSPA که در آن اشیا رأس‌های گراف را می‌ساختند، خوشه‌ها به‌عنوان رأس در نظر گرفته می‌شوند و وزن یال‌ها با استفاده از مقیاس Jaccard دودویی به‌دست می‌آید [4]. به این صورت که وزن بین دو رأس h1 و h2 با رابطهٔ (۱) محاسبه می‌شود:

$$w(C_x, C_y) = \frac{C_x' C_y}{\|C_x\|_2^2 + \|C_y\|_2^2 - C_x' C_y} \qquad (۱)$$

به‌طوری‌که C_x و C_y بردارهایی هستند که دو خوشه h1 و h2 را نمایش می‌دهند و C_x' نیز ترانهاده C_x است. هر عنصر بردار، نمایان‌گر یک شیء است. اگر خوشه، حاوی شیئی باشد، عنصر معادل برابر یک و در غیراین‌صورت صفر خواهد بود؛ سپس برای تعیین خوشه‌بندی نهایی از الگوریتم METIS [24] استفاده می‌شود.

[1] Validity index
[2] Consensus function
[3] Cluster-based Similarity Partitioning Algorithm
[4] Hyper Graph Partitioning Algorithm
[5] Meta-CLustering Algorithm

۲-۲- اندازه‌گیری میزان شباهت و تفاوت:

برای تعیین میزان شباهت دو خوشه‌بندی معیار[1]های متفاوتی وجود دارد [21, 22, 26] که از پرکاربردترین آنها می‌توان به NMI[2] [4] و ARI[3] [27] اشاره کرد. NMI یک روش برای محاسبهٔ میزان شباهت یا تفاوت دو خوشه‌بندی است. فرض کنید π_a و π_b دو خوشه‌بندی روی مجموعه‌داده‌ای با N نمونه باشد، در این‌صورت NMI بین دو خوشه‌بندی به‌صورت رابطهٔ (۲) است:

$$NMI(\pi_a,\pi_b)= \frac{\sum_{h=1}^{k_a}\sum_{l=1}^{k_b} n_{h,l}\log_2\left(\frac{N.n_{h,l}}{n_h.n_l}\right)}{\sqrt{\left(\sum_{h=1}^{k_a} n_h \log_2\frac{n_h}{N}\right)\left(\sum_{l=1}^{k_b} n_l \log_2\frac{n_l}{N}\right)}} \qquad (۲)$$

به‌طوری‌که k_a تعداد خوشه‌های π_a، k_b تعداد خوشه‌های π_b، n_{hl} تعداد نمونه‌های مشترک h اُمین خوشه π_a و l اُمین خوشه π_b است. N_h تعداد نمونه‌های h اُمین خوشه π_a و n_l تعداد نمونه‌های l اُمین خوشه π_b است. NMI میزان اطلاعات مشترک بین دو خوشه‌بندی را اندازه می‌گیرد و می‌تواند مقداری بین صفر تا یک را بپذیرد. اگر مقدار خروجی NMI برابر یک باشد، دوخوشه‌بندی به‌طورکامل مشابه‌اند و هرچقدر از مقدار یک کاسته شود، نشان می‌دهد که دو خوشه‌بندی تفاوت بیشتری دارند و اگر صفر باشد دو خوشه‌بندی به‌طورکامل متفاوت هستند.

ARI [27] روش دیگری برای تعیین میزان شباهت دو خوشه‌بندی است. ابتدا همانند بالا فرض می‌کنیم π_a و π_b دو خوشه‌بندی روی مجموعه‌داده‌ای با N نمونه باشد. در این‌صورت ARI بین دو خوشه‌بندی به‌صورت رابطه (۳) محاسبه می‌شود:

$$ARI(\pi_a,\pi_b)=\frac{\sum_{h=1}^{k_a}\sum_{l=1}^{k_b}\binom{n_{h,l}}{2}-t_3}{1/2(t_1+t_2)-t_3} \qquad (۳)$$

به‌طوری‌که:

$$t_1=\sum_{h=1}^{k_a}\binom{n_h}{2},\ t_2=\sum_{l=1}^{k_b}\binom{n_l}{2},\ t_3=\frac{2t_1t_2}{n(n-1)}$$

متغیرهای k_a، k_b، $n_{h,l}$، n_h و n_l نیز مشابه موارد گفته‌شده در معیار NMI است.

[1] Measure
[2] Normalized Mutual Information
[3] Adjusted Rand Index

۳- پیشینهٔ پژوهش

ایدهٔ اولیهٔ ترکیب خوشه‌بندی‌های مختلف جهت به‌دست‌آوردن خوشه‌بندی بهتر، تحت روش‌های مختلف مورد بررسی قرار گرفته است [28, 29]؛ امّا چارچوب رسمی خوشه‌بندی ترکیبی برای نخستین‌بار در [4] معرّفی شد. فرض کنید $X=(x_1,x_2,\ldots,x_n)$، نمایان‌گر مجموعه‌داده‌ای باشد که n نمونه داده دارد. مجموعه‌داده H بار با الگوریتم‌های خوشه‌بندی، بخش‌بندی می‌شوند تا H نتیجه $P=(p_1,p_2,\ldots,p_H)$ حاصل شود، به‌طوری‌که $p_k(k=1,2,\ldots,H)$ نتیجهٔ خوشه‌بندی در k اُمین اجرا باشد. درنهایت تابع اجماع Γ ، مجموعه خوشه‌بندی‌های تولیدشده را ترکیب می‌کند تا خوشه‌بندی نهایی تولید شود. توابع اجماع به‌کار بسته‌شده در این روش HPGA، CSPA و MCLA هستند که همگی مبتنی بر گرافند.

در [17] سه روش بر پایه کیفیت و تنوع ارائه شده است. نخستین روش که معیار مشترک[4] نام دارد، یک تابع هدف[5] ارائه که کیفیت و تنوع را ترکیب می‌کند. دومین روش CAS[6] نامیده شده است. این روش، ابتدا یک ماتریس شباهت می‌سازد که هر عنصر آن، فاصلهٔ دو خوشه‌بندی مربوط به دو اندیس آن عنصر است؛ سپس این ماتریس شباهت به‌عنوان تابع وزنی به الگوریتم خوشه‌بندی طیفی داده می‌شود تا خوشه‌بندی‌ها گروه‌بندی شود؛ سپس از هر گروه بهترین آن را با استفاده از معیارهای کیفیت، شناسایی کرده و در اختیار توابع اجماع مبتنی بر گراف قرار می‌دهد. روش آخر که پوسته محدب[7] نامیده شده، ابتدا دیاگرام کیفیت-تنوع برای مجموعه خوشه‌بندی‌ها می‌سازد. به‌طوری‌که هر نقطه، متناظر با یک جفت خوشه‌بندی در مجموعهٔ خوشه‌بندی‌هاست؛ سپس پوستهٔ محدب همه نقاط، مجموعهٔ کاهش‌یافته را برای توابع اجماع به‌وجود می‌آورد.

در [19]، یک روش تطبیقی ارائه شده است که بر اساس آن داده‌ها به دو دسته پایدار و ناپایدار تقسیم می‌شوند. نویسندگان نشان دادند که در داده‌های ناپایدار، انتخاب خوشه‌بندی‌های با تنوع زیاد، نتایج را بهبود می‌دهد.

در الگوریتم SELSCE [30]، خوشه‌بندی‌های اولیه با استفاده از الگوریتم خوشه‌بندی طیفی ساخته می‌شوند. در این روش ابتدا T رتبه‌بندی مختلف از خوشه‌بندی‌های پایه ایجاد می‌شود و سپس از این T رتبه‌بندی، یک رتبه‌بندی واحد تولید می‌شود. در این روش برای ایجاد هر رتبه‌بندی از

[4] Joint Criterion
[5] Objective function
[6] Cluster And Select
[7] Convex Hull

خوشه‌بندی‌ها، ابتدا نیمی از آنها به‌صورت تصادفی انتخاب و خوشه‌بندی ترکیبی آنها تولید می‌شود؛ سپس با استفاده از معیارهای ارتباط[1] (معیارهای اندازه‌گیری شباهت و تفاوت)، یک رتبه‌بندی ایجاد و برای ایجاد T رتبه‌بندی، این مرحله T مرتبه تکرار می‌شود. درنهایت براساس تعدادی که کاربر در پارامتر ورودی مشخص می‌کند، خوشه‌بندی از ابتدای رتبه‌بندی انتخاب و به توابع اجماع داده می‌شود تا فرآیند خوشه‌بندی ترکیبی کامل شود.

در الگوریتم similarity-based [31] یک روش جدید برای انتخاب اعضای ترکیب ارائه می‌شود که بر پایهٔ قاعده نزدیک‌ترین همسایه است. بدین‌ترتیب که در ابتدا مجموعه اولیه خوشه‌بندی‌ها و تعداد خوشه‌بندی‌هایی که می‌بایست درنهایت ترکیب شوند از ورودی دریافت می‌شوند؛ سپس این مجموعه بر اساس قاعدهٔ نزدیک‌ترین همسایه آن‌قدر کاهش داده می‌شوند تا به تعداد مورد نظر دست یابند. برای این منظور ابتدا فاصله دوبه‌دوی تمام خوشه‌بندی‌ها بر اساس معیارهای شباهت به‌دست می‌آیند؛ سپس نزدیک‌ترین خوشه‌بندی به هر خوشه‌بندی تعیین و فاصله بین آن‌ها به‌عنوان فاصله تا نزدیک‌ترین خوشه‌بندی ذخیره می‌شود. درنهایت آن خوشه‌بندی که کم‌ترین فاصله را تا نزدیک‌ترین خوشه‌بندی خود دارد، حذف می‌شود و این فرآیند ادامه می‌یابد تا تعداد مورد نظر حاصل شود.

الگوریتم ESDF [32] تلاش می‌کند تا نشان دهد با اولویت‌بندی خوشه‌بندی‌ها بر اساس فرکانس[2] و تنوع، و انتخاب خوشه‌بندی‌ها به‌صورت حریصانه[3] و با ترتیب نزولی اولویت‌ها، خوشه‌بندی ترکیبی بهتری ایجاد می‌شود؛ چون تعداد خوشه‌بندی‌های کم‌تری در خوشه‌بندی ترکیبی شرکت می‌کنند. هم‌چنین خوشه‌بندی‌هایی در خوشه‌بندی ترکیبی شرکت می‌کنند که اختلاف بیش‌تری با دیگر خوشه‌بندی‌ها دارند (با این عقیده که دورترین عضو به‌احتمال دارای اطلاعات مفیدی است) و تعداد تکرارشان بیشتر است.

در [21] نویسنده روش‌های متعددی ارائه می‌کند که یکی از بهترین روش‌های آن SR است. در این روش، برای نخستین‌بار شاخص‌های ارزیابی کیفیت در موضوع انتخاب اعضای ترکیب در خوشه‌بندی ترکیبی، مورد استفاده قرار گرفته است (شاخص‌های ارزیابی خوشه‌بندی برای دهه‌های متمادی، برای ارزیابی کیفیت خوشه‌بندی مورد استفاده قرار گرفته‌اند). در این روش با استفاده از چندین شاخص اعتبار[4]، کیفیت هریک از خوشه‌بندی‌ها را محاسبه می‌کند. به‌عبارت دیگر به‌ازای هر شاخص اعتبار یک رتبه‌بندی از خوشه‌بندی‌ها ایجاد می‌شود. در ادامه مجموع رتبه‌ها برای هر خوشه‌بندی محاسبه شده و سپس آنها از کوچک به بزرگ مرتب می‌شوند. درنهایت با توجه به تعداد خوشه‌بندی‌هایی که از قبل برای انتخاب تعیین شده است، خوشه‌بندی‌هایی با کمترین رتبه انتخاب می‌شود.

الگوریتم HCES [10] از الگوریتم‌های خوشه‌بندی سلسله‌مراتبی برای گروه‌بندی خوشه‌بندی‌ها بهره می‌برد. در این روش ابتدا خوشه‌بندی‌های متعددی با استفاده از الگوریتم k-means ایجاد می‌شود. در هر بار اجرای k-means مقدار k یک عدد تصادفی است. در گام بعدی کلیه این خوشه‌بندی‌ها با استفاده از توابع اجماع HPGA و CSPA ترکیب می‌شوند تا h^* که همان خوشه‌بندی ترکیب کامل نام دارد، حاصل شود. در این روش برای به‌دست‌آوردن فاصله بین دو خوشه‌بندی Aو B، ابتدا فاصله بین هریک از A و B با h^* به‌دست می‌آید و قدرمطلق تفاضل مقدارهای به‌دست‌آمده، فاصلهٔ بین A و B در نظر گرفته می‌شود؛ سپس ماتریس شباهت محاسبه می‌شود که هر مؤلفهٔ ماتریس، فاصله بین دو خوشه‌بندی را نشان می‌دهد. در گام بعدی با استفاده از خوشه‌بندی سلسله‌مراتبی، خوشه‌بندی‌ها گروه‌بندی می‌شوند و از هر گروه با استفاده از معیار کیفیت NMI، بهترین خوشه‌بندی انتخاب می‌شود (در هر لایه از *درخت سلسله‌مراتب*[5] و در اختیار توابع اجماع قرار می‌گیرد. درنهایت خوشه‌بندی با بالاترین کیفیت، انتخاب می‌شود.

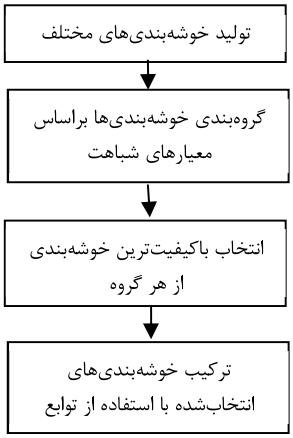
در این قسمت از مقاله، الگوریتم‌های متعددی که برای انتخاب اعضای ترکیب طراحی شده بودند، مورد بررسی قرار گرفتند. این الگوریتم‌ها سعی داشتند تا با کاهش تعداد خوشه‌بندی‌های پایه (که به‌عنوان ورودی به تابع اجماع داده می‌شوند)، کیفیت خوشه‌بندی ترکیبی را بهبود ببخشند. بعضی الگوریتم‌ها از روش‌های مبتنی بر شباهت و تفاوت استفاده کرده بودند؛ نظیر[10, 17, 31, 32] و برخی نیز با استفاده از روش رتبه‌بندی اقدام به انتخاب خوشه‌بندی‌ها کرده بودند؛ نظیر [21, 30]. به‌طورتقریبی در همه الگوریتم‌های ذکرشده نیاز به تعیین تعداد خوشه‌بندی‌هایی که باید انتخاب شوند، یک نیاز اساسی است و جزو ورودی‌های آن الگوریتم محسوب می‌شود؛ بنابراین ارائه یک الگوریتم جدید که در آن نیازی به تعیین این عدد نباشد، ضروری است.

[1] Relevance measure
[2] Frequency
[3] Greedy
[4] Validity index
[5] Dendogram

۴- روش پیشنهادی

روش پیشنهادی، برای ایجاد خوشه‌بندی ترکیبی شامل چهار گام است. در گام نخست خوشه‌بندی‌های متعددی روی مجموعه‌داده انجام می‌شود. در گام دوم با استفاده از الگوریتم گروه‌بندی، خوشه‌بندی‌های تولیدشده در مرحله قبل گروه‌بندی می‌شود. در گام سوم با کمک الگوریتم انتخاب، از هر گروه بهترین خوشه‌بندی انتخاب می‌شود و درنهایت در گام آخر، خوشه‌بندی‌های انتخاب‌شده با استفاده از توابع اجماع ترکیب می‌شود. در ادامهٔ این بخش، مراحل ذکرشده با جزئیات کامل بیان خواهند شد.

در خوشه‌بندی ترکیبی سنتی، همهٔ خوشه‌بندی‌های ایجادشده، با استفاده از توابع اجماع ترکیب می‌شوند؛ اما در CES (رویکرد جدید خوشه‌بندی ترکیبی) این‌گونه نیست؛ در این رویکرد، خوشه‌بندی‌ها بر اساس تنوع و کیفیت انتخاب می‌شوند تا نتیجه ترکیب بهبود یابد. در نتیجهٔ این انتخاب، توابع اجماع درگیر افزونگی نمی‌شوند؛ چون خوشه‌بندی‌های مشابه درعمل تأثیری در خوشه‌بندی نهایی ندارند. برای رسیدن به این هدف لازم است بعد از تولید خوشه‌بندی‌های متعدد، ابتدا خوشه‌بندی‌ها گروه‌بندی شوند و سپس در هر گروه با استفاده از رأی‌گیری روی شاخص‌های ارزیابی کیفیت، تنها یک مورد باکیفیت آن انتخاب شود و در اختیار توابع اجماع قرار گیرد. شکل (۱) این مراحل را بیان می‌کند.



(شکل-۱): چارچوب خوشه‌بندی ترکیبی ارائه‌شده
(Figure-1): Framework of the proposed ensemble clustering

۴-۱- تولید خوشه‌بندی‌های مختلف

نخستین مرحلهٔ چارچوب خوشه‌بندی ترکیبی، ایجاد خوشه‌بندی‌های متفاوت است. خوشه‌بندی‌های متفاوت می‌تواند یا با الگوریتم‌های متفاوت انجام شوند، نظیر [12, 13] و یا می‌تواند شامل تکرار یک الگوریتم مشخص مانند k-means، DBSCAN و Spectral با پارامترهای مختلف باشد [11]. در روش پیشنهادی از روش دوم استفاده شده است.

۴-۲- الگوریتم گروه‌بندی

پس از تولید خوشه‌بندی‌های اولیه، نوبت به گروه‌بندی می‌رسد. هدف الگوریتم گروه‌بندی این است که خوشه‌بندی‌های مشابه در یک گروه قرار گیرند تا در مرحله بعدی (الگوریتم انتخاب)، از هر گروه یک خوشه‌بندی باکیفیت انتخاب شود. شکل (الگوریتم گروه‌بندی ارائه‌شده را نشان می‌دهد.

```
Algorithm 1: Grouping
Input: matrix GSC;
Output: Vector group_label;
Begin
Vector Obtained_NMI;
for i←1 to n              //n is number of clustering
   for j← i +1 to n
      Current_NMI←NMI(GSC (j) , GSC (i));
      If Current_NMI>=Obtained_NMI(j) and j!=i
         Obtained_NMI(j) ←Current_NMI;
         Index ← j;
      end if
      i and index are given same group_label;
   end for
End for
End Algorithm
```

(شکل-۲): الگوریتم گروه‌بندی
(Figure-2): Grouping algorithm

ورودی این الگوریتم، ماتریس GSC، حاوی خوشه‌بندی‌های تولیدشده از مجموعه‌داده اولیه است. هر سطر این ماتریس حاوی یک خوشه‌بندی از مجموعه‌داده اولیه است. اگر تعداد خوشه‌بندی‌های اولیه n باشد، تعداد سطرهای ماتریس هم n خواهد بود. خروجی این الگوریتم یک بردارستونی بنام group_label است که نشان می‌دهد هر خوشه‌بندی به کدامیک از خوشه‌بندی‌ها نزدیک‌تر است. در اینجا نیز اگر تعداد خوشه‌بندی‌های اولیه n باشد، تعداد سطرهای این بردار نیز n خواهد بود. خوشه‌بندی‌های دارای group_label یکسان، در یک گروه قرار دارند.

هم‌چنین در این الگوریتم یک بردار ستونی به نام Obtained_NMI در نظر گرفته شده که نشان می‌دهد هر خوشه‌بندی با چه میزان شباهت با دیگر خوشه‌بندی‌ها در یک گروه قرار گرفته است. این بردار ابتدا با صفر مقداردهی شده است. متغیر index نیز درواقع، اندیس شبیه‌ترین خوشه‌بندی را به خوشه‌بندی مورد بررسی در خود قرار می‌دهد.

وظیفهٔ الگوریتم گروه‌بندی این است که مشابه‌ترین خوشه‌بندی‌ها را در یک گروه قرار دهد. برای یافتن شبیه‌ترین خوشه‌بندی به هر خوشه‌بندی می‌بایست میزان شباهت آن با

سایر خوشه‌بندی‌ها سنجیده شود. به این‌صورت که در هر مرحله، ابتدا میزان شباهت دو خوشه‌بندی با استفاده از معیار NMI سنجیده می‌شود و درون متغیر Current_NMI قرار می‌گیرد؛ سپس اگر Current_NMI میان دو خوشه‌بندی مورد بررسی بزرگ‌تر از Obtained_NMI باشد، در این بردار جایگزین می‌شود و اندیس آن درون متغیر index قرار می‌گیرد. (Current_NMI، مقدار NMI فعلی به‌دست‌آمده بین دو خوشه‌بندی است و Obtained_NMI، بالاترین مقدار NMI را که تاکنون برای هر خوشه‌بندی به‌دست آمده است، نشان می‌دهد). این روال ادامه پیدا می‌کند تا شبیه‌ترین خوشه‌بندی به خوشه‌بندی مورد نظر یافت شود و اندیس آن درون متغیر index قرار گیرد. درنهایت، خوشه‌بندی مورد نظر و خوشه‌بندی با اندیس index درون یک گروه قرار می‌گیرند(برچسب یکسانی در بردار group_label خواهند گرفت).

به‌عنوان مثال فرض کنیم با استفاده از روشی که در پاراگراف قبلی به آن اشاره شد، شبیه‌ترین خوشه‌بندی به خوشه‌بندی نخست، خوشه‌بندی nام باشد. این دو خوشه‌بندی، یعنی خوشه‌بندی نخست و nام درون یک گروه قرار می‌گیرند(برچسب یکسان خواهند گرفت). پس از آن نوبت به خوشه‌بندی دوم می‌رسد. شبیه‌ترین خوشه‌بندی به خوشه‌بندی دوم نیز با الگوریتم بالا یافت می‌شود و برچسب آن را به خود می‌گیرد (اگر خوشه‌بندی یافت‌شده برچسب نداشت، یعنی عضو هیچ گروهی نبود، این دو خوشه‌بندی گروهِ جدیدی را تشکیل خواهند داد). همین روال برای سایر خوشه‌بندی‌ها تکرار می‌شود؛ درنتیجه خوشه‌بندی‌های شبیه به هم دارای یک برچسب یکسان در بردار group_label خواهند شد و در یک گروه قرار می‌گیرند.

در این الگوریتم می‌توان برای محاسبهٔ میزان شباهت دو خوشه‌بندی از معیار ARI نیز استفاده کرد؛ به این ترتیب که هرجا نام NMI آمده، با ARI جایگزین می‌شود.

۳-۴- الگوریتم انتخاب

بعد از گروه‌بندی باید در هر گروه با کیفیت‌ترین آن انتخاب شود. هدف از این الگوریتم پیداکردن باکیفیت‌ترین خوشه‌بندی از هر گروه است. در بخش آزمایش‌ها نشان خواهیم داد که این کاهش افزونگی در بسیاری موارد باعث افزایش کارایی خوشه‌بندی ترکیبی خواهد شد.

برای سنجش کیفیت خوشه‌بندی می‌توان از شاخص‌های اعتبار خوشه بهره گرفت. بدین ترتیب که همه اعضای گروه‌ها با شاخص‌ها مورد سنجش قرار می‌گیرند. در هر گروه، خوشه‌بندی‌ای که تعداد بیشتری از شاخص‌ها را بهینه کند، انتخاب می‌شود. DB[1] [33]، CH[2] [34] و SI[3] [35] از جمله این شاخص‌ها هستند که در ادامه به توضیح آنها می‌پردازیم:

شاخص DB [33] تابعی است که نسبت جمع پراکندگی داخل خوشه‌ای را به پراکندگی بین خوشه‌ای محاسبه می‌کند. رابطه (۴)، این تابع ارزیابی را بیان می‌کند:

$$DB(k)=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq i'}\frac{S_n(c_i)+S_n(c_{i'})}{S(c_i,c_{i'})} \qquad (۴)$$

به‌طوری‌که k تعداد خوشه‌ها، S_n میانگین فاصلهٔ اشیای خوشه c_i از مرکز این خوشه و $S(c_i,c_{i'})$ فاصلهٔ بین مراکز خوشه‌های c_i و $c_{i'}$ است. اگر خوشه‌ها فشرده باشند و نسبت به یکدیگر فاصله داشته باشند، این نسبت کوچک است. بنابراین DB مقدار کوچکی برای خوشه خوب دارد [36].

شاخص CH [34] به‌صورت رابطه (۵) تعریف می‌شود:

$$CH(k)=\frac{B/(k-1)}{W/(n-k)} \qquad (۵)$$

به‌طوری‌که k تعداد خوشه‌ها، B مجموع مربعات فاصله بین خوشه‌ها، W مجموع مربعات فاصله در میان خوشه‌ها و n تعداد داده‌هاست. در مورد گروه‌هایی که اندازهٔ یکسان دارند CH اغلب معیار خوبی برای مشخص‌کردن تعداد صحیح گروه‌هاست. بهترین خوشه‌بندی، بالاترین مقدار CH را دارد.

علاوه‌بر شاخص‌های اعتبار می‌توان از SNMI [4] در ارزیابی کیفیت خوشه‌بندی‌ها بهره برد. فرض کنیم که مجموعه E شامل k خوشه‌بندی باشد که به‌صورت $E=(p_1,p_2,\ldots,p_k)$ نمایش داده شود، آن‌گاه معیار SNMI به‌صورت رابطه (۶) تعریف می‌شود:

$$SNMI(p,E)=\sum_{i=1}^{k}NMI(p,p_i) \qquad (۶)$$

امّا از آن‌جایی که هر شاخص ممکن است، برای دادهٔ مشخصی مناسب باشد، در روش پیشنهادی از روش رأی‌گیری شاخص‌ها بهره گرفته شده است تا این نقص پوشش داده شود. شکل (۳) الگوریتم انتخاب خوشه‌بندی را نمایش می‌دهد.

[1] Davies Bouldin
[2] Calinski Harabasz
[3] Silhouette

Algorithm 2: Selection
Input: **Matrix** GSC, **Vector** group_label;
Output: **Matrix** RGSC;
Begin
For each group **repeat**
$y_1 \leftarrow$ index of clustering that obtains min DB value
$y_2 \leftarrow$ index of clustering that obtains max CH value
$y_3 \leftarrow$ index of clustering that obtains max SI value
$y_4 \leftarrow$index of clustering that obtains max SNMI value
$Y \leftarrow$mode (y_1, y_2, y_3, y_4);
Add GSC(Y) to RGSC
end For
end Algorithm

(شکل-۳): الگوریتم انتخاب
(Figure-3): Selection algorithm

ورودی الگوریتم انتخاب، ماتریس GSC است که حاوی خوشه‌بندی‌های اولیه است. ورودی دیگر این الگوریتم، بردار group_label است که نشان می‌دهد هر خوشه‌بندی با کدام خوشه‌بندی در یک گروه قرار گرفته است (خوشه‌بندی-هایی که برچسب یکسانی در این بردار دارند، دریک گروه قرار دارند). درواقع بردار group_label خروجی الگوریتم گروه‌بندی است.

خروجی الگوریتم انتخاب، ماتریس RGSC است که مجموعه خوشه‌بندی‌ها بعد از فرآیند انتخاب است. این مجموعهٔ کاهش‌یافته، به‌منظور انجام عمل ترکیب در اختیار توابع اجماع قرار می‌گیرد.

بدنهٔ این الگوریتم یک حلقه است که در هر بار تکرار، شاخص‌های ارزیابی کیفیت (DB، CH، SI و SNMI) را روی یک گروه از خوشه‌بندی‌ها اعمال و باکیفیت‌ترین خوشه‌بندیِ هر گروه را انتخاب کرده و به ماتریس RGSC اضافه می‌کند.

عمل انتخاب باکیفیت‌ترین خوشه‌بندی از هر گروه شکل (۳) به این‌صورت است که ابتدا در هر گروه، اندیس خوشه‌بندی‌ای که شاخص‌ها را بهینه کرده، درون متغیرهای y_1 تا y_4 قرار می‌گیرد؛ به این ترتیب که اندیس خوشه‌بندی‌ای که کمترین مقدار DB را داشته درون y_1، اندیس خوشه‌بندی‌ای که بیشترین مقدار CH را داشته درون y_2 ، اندیس خوشه‌بندی‌ای که بیشترین مقدار SI را داشته درون y_3 و درنهایت اندیس خوشه‌بندی‌ای که بیش‌ترین مقدار SNMI را داشته درون y_4 قرار می‌گیرد (قابل ذکر است که باکیفیت‌ترین خوشه‌بندی، کمترین مقدار را در DB و بیشترین مقدار را در سایر شاخص‌ها خواهد داشت)؛ سپس با انجام عمل رأی‌گیری (یا همان شاخص مد[1] در آمار) روی متغیرهای y_1 تا y_4، نتیجه‌ای که بیش‌ترین تکرار را داشته انتخاب می‌شود (اندیس آن درون متغیر Y قرار می‌گیرد). درنهایت، این خوشه‌بندی انتخاب‌شده (خوشه‌بندی با اندیس Y) به ماتریس RGSC اضافه می‌شود.

مهم‌ترین مزیت روش رأی‌گیری این است که چنان‌چه نتیجهٔ معیارهای مختلف هم‌خوانی نداشته باشد، آن نتیجه‌ای که تکرار بیشتری دارد، انتخاب می‌شود. در روش رأی‌گیری به‌کار گرفته‌شده به هر معیار وزن یکسانی تعلّق گرفته است؛ چون درصد موفقیت معیارها به‌طورتقریبی یکسان است. بنابراین منطقی است که به معیارها وزن یکسانی اختصاص داده شود. روال پیداکردن باکیفیت‌ترین خوشه‌بندی از هر گروه (پاراگراف بالا)، برای سایر گروه‌ها انجام می‌شود و از هر گروه باکیفیت‌ترین خوشه‌بندی انتخاب و به ماتریس RGSC اضافه می‌شود. در انتهای این الگوریتم، این ماتریس یا همان مجموعهٔ خوشه‌بندی‌های کاهش‌یافته، آمادهٔ عمل ترکیب می‌شود. به‌عنوان مثال اگر در یک گروه از خوشه‌بندی‌ها تعداد C خوشه‌بندی مختلف وجود داشته باشد، الگوریتم انتخاب برای هر یک از این C خوشه‌بندی، چهار شاخص یادشده را اجرا می‌کند. به‌عنوان مثال شاخص DB و بقیه شاخص‌ها هر کدام به تشخیص خود یک خوشه‌بندی را به‌عنوان باکیفیت‌ترین انتخاب می‌کنند. درنهایت بین تشخیص آنها رأی‌گیری به‌عمل آمده و بهترین مورد انتخاب می‌شود. همین کار برای دیگر گروه‌ها نیز انجام می‌شود تا بهترین خوشه‌بندی سایر گروه‌ها انتخاب شود و بهترین خوشه‌بندی‌ها به تابع اجماع اعمال می‌شود.

۴-۴- ترکیب خوشه‌بندی‌های مختلف

در آخرین گام چارچوب خوشه‌بندی ترکیبی(مرحله چهارم)، مجموعهٔ خوشه‌بندی‌های کاهش یافته در اختیار توابع اجماع قرار می‌گیرد تا عمل ترکیب انجام شود.

به‌اختصار چارچوب کلی روش پیشنهادی به‌صورت زیر خواهد بود:

۱. روی یک مجموعه‌داده خوشه‌بندی‌های مختلف تولید می‌شود؛
۲. خوشه‌بندی‌های تولیدشده با استفاده از الگوریتم (۱) گروه‌بندی می‌شود؛
۳. از هر گروه با توجه به الگوریتم (۲) بهترین خوشه‌بندی انتخاب می‌شود؛
۴. توابع اجماع روی مجموعه‌های کاهش‌یافته به‌کار گرفته می‌شود تا خوشه‌بندی نهایی تولید شود.

[1] Mode

به‌عنوان مثال اگر m خوشه‌بندی پایه داشته باشیم و الگوریتم گروه‌بندی، این m خوشه‌بندی را به n گروه تقسیم کرده باشد، از هر گروه باکیفیت‌ترینِ آن با توجه به رأی‌گیری روی شاخص‌ها انتخاب می‌شود (به عبارت دیگر خوشه‌بندی‌ای که تعداد بیشتری از شاخص‌ها را بهینه کرده، به‌عنوان باکیفیت‌ترین خوشه‌بندیِ گروه فعلی انتخاب می‌شود) و درنهایت n خوشه‌بندی در اختیار تابع اجماع قرار می‌گیرد تا خوشه‌بندی نهایی تولید شود.

در بخش بعدی روش پیشنهادی را روی مجموعهٔ متنوعی از داده‌ها آزمایش می‌کنیم.

۵- آزمایش‌ها

در این بخش میزان کارایی الگوریتم ارائه‌شده مورد بررسی و آزمایش قرار می‌گیرد. ابتدا لازم است ویژگی‌های مجموعه‌داده‌ها بیان شوند و سپس به بررسی سایر تنظیمات مورد نیاز و مقایسهٔ کارایی پرداخته خواهد شد.

۵-۱- مجموعه‌داده‌ها

در بخش آزمایش‌ها از داده‌های واقعی استفاده‌شده که همهٔ آنها از سایت UCI [37] گرفته شده‌اند. این مجموعه‌داده‌ها در بسیاری از مقالات ازجمله [10, 21, 30] استفاده شده‌اند. ویژگی‌های این مجموعه‌داده‌ها در جدول (۱) بیان شده است. از جملهٔ ویژگی‌های این داده‌ها، بُعد بالای آنها است. داده‌های با ابعاد بالا[1] به‌دلیل برخی ویژگی‌هایشان نظیر پیچیدگی بیشتر، به معضل جدی در دنیای داده‌کاوی تبدیل شده و توجه زیادی را به خود جلب کرده‌اند.

۵-۲- کارایی روی مجموعه‌داده‌ها

در این قسمت الگوریتم ارائه‌شده در بخش قبل، روی مجموعه‌داده‌های معرفی‌شده آزمایش می‌شود. به‌ازای هر مجموعه‌داده می‌بایست خوشه‌بندی‌های متعددی تولید شود. برای این منظور از الگوریتم k-means استفاده شده است؛ زیرا بیش‌تر کارهای قبلی این حوزه، از جمله [10, 17, 21, 32, 38]، از الگوریتم k-means برای تولید خوشه‌بندی‌های پایه استفاده کرده‌اند. با استفاده از الگوریتم k-means و دادن مقدارهای متفاوت تصادفی به پارامتر k، خوشه‌بندی‌های متعددی می‌توان ایجاد کرد. در هر بار اجرا مقادیر متفاوت k از بین k_{min} و k_{max} به‌طور تصادفی انتخاب می‌شوند، به‌طوری‌که k_{min} برابر ۲ و k_{max} برابر ۲k است. اگر اطلاعی از k در دسترس نباشد، k_{max} را برابر $\sqrt{n}$ قرار می‌دهیم، به‌طوری‌که n برابر تعداد نمونه‌های موجود در مجموعه‌داده است [10, 17, 38]. در بخش آزمایش‌ها، جهت تولید خوشه‌بندی‌های اولیه، تعداد خوشه‌ها به‌صورت تصادفی از یک بازهٔ مشخص، انتخاب شده است؛ اما در روش پیشنهادی نیازی به تعیین تعداد خوشه‌بندی‌هایی که درنهایت به تابع اجماع داده می‌شود، نیست.

(جدول-۱): ویژگی‌های مجموعه‌داده‌ها
(Table-1): Feature of datasets

ردیف	مجموعه‌داده	تعداد خوشه‌ها	تعداد بعد	تعداد نمونه
1	Wine	3	13	178
2	Heart	2	13	270
3	Sonar	2	60	208
4	Soybean	4	35	47
5	Breast tissue	6	9	106
6	Glass	7	9	214
7	WDBC	2	30	569
8	Ecoli	8	7	336
9	Vehicle	4	18	846
10	Segmentation	7	19	2310
11	Sat. image	6	36	6435

با استفاده از روش بالا تعداد حداکثر یکصد خوشه‌بندی تولید و سپس از روش ارائه‌شده در بخش روش پیشنهادی استفاده می‌شود تا از بین خوشه‌بندی‌های تولیدشده، فرآیند انتخاب بر اساس معیارهای کیفیت و تنوع صورت گیرد. پس از انتخاب خوشه‌بندی‌ها، یک تابع اجماع نیاز است تا آنها را ترکیب کند. در این مقاله از توابع اجماع مبتنی بر گراف CSPA،MCLA و HGPA بهره گرفته شده است. جدول (۲) نتایج مربوط به آزمایش‌ها بر اساس تابع اجماع CSPA، جدول (۳) نتایج مربوط به آزمایش‌ها بر اساس تابع اجماع MCLA و جدول (۴) نتایج مربوط به آزمایش‌ها بر اساس تابع اجماع HGPA است. هر عدد در هر خانه از این جداول، نتیجه میانگین ده‌بار اجرا است. برای هر مجموعه‌داده، بهترین مقدار دقّت[2] به‌دست آمده و مقادیر نزدیک به آن به شکل پررنگ، مشخص شده‌اند.

روش ارائه‌شده در این مقاله براساس معیارهای شباهت ARI وNMI که به‌ترتیب ARISelective و NMISelective نامیده شده‌اند با روش Full که در آن همهٔ خوشه‌بندی‌ها ترکیب می‌شوند (ترکیب کامل)، روش CAS [17] و روش SR [21] براساس میزان دقّت مقایسه شده است.

[1] High dimensional data

[2] Accuracy

(جدول-۲): مقایسه دقّت روی CSPA

(Table-2): Accuracy comparison on CSPA

ردیف	مجموعه‌داده	Full	NMISelective	ARISelective	CAS	SR
1	Wine	70.16	**71.12**	70.39	71.04	70.65
2	Soybean	70.85	**73.61**	**73.19**	71.23	70.31
3	Heart	59.77	**60.22**	**60.55**	59.79	**60.44**
4	Sonar	56.44	**57.59**	**57.69**	56.29	54.25
5	WDBC	**84.20**	**84.35**	**84.35**	82.34	80.01
6	Glass	40.74	40.98	40.93	40.60	**41.01**
7	Breast tissue	42.26	42.54	**43.49**	42.70	**43.61**
8	Ecoli	**48.06**	**48.48**	**48.77**	47.79	47.96
9	Vehicle	38.78	**39.48**	38.93	38.70	**39.01**
10	Segmentation	59.65	**60.13**	**60.77**	59.76	57.98
11	Sat. image	**64.45**	**64.87**	**64.98**	62.63	60.76

(جدول-۳): مقایسه دقّت روی MCLA

(Table-3): Accuracy comparison on MCLA

ردیف	مجموعه‌داده	Full	NMISelective	ARISelective	CAS	SR
1	Wine	**72.47**	**72.47**	**72.47**	**72.34**	71.97
2	Soybean	72.97	73.40	**74.68**	72.63	71.29
3	Heart	60.22	**60.77**	**60.96**	60.08	**60.51**
4	Sonar	56.73	**57.45**	55.86	55.94	56.15
5	WDBC	80.35	**82.35**	**82.39**	81.07	79.43
6	Glass	47.42	**50**	49.67	47.25	45.64
7	Breast tissue	39.24	**41.69**	40.47	40.38	**41.77**
8	Ecoli	**53.33**	**53.51**	**53.60**	**53.10**	51.26
9	Vehicle	43.38	**44.23**	43.97	43.34	42.79
10	Segmentation	60.65	**62.34**	**62.96**	60.71	61.48
11	Sat. image	67.30	67.62	**68.88**	65.45	**68.34**

(جدول-۴): مقایسه دقّت روی HGPA

(Table-4): Accuracy comparison on HGPA

ردیف	مجموعه‌داده	Full	NMISelective	ARISelective	CAS	SR
1	Wine	**72.69**	72.47	72.47	**72.55**	72.12
2	Soybean	72.97	**74.04**	73.40	72.70	73.08
3	Heart	59.14	**60.62**	**60.66**	60.28	60.02
4	Sonar	**58.75**	**58.94**	57.74	57.76	58.11
5	WDBC	83.30	82.79	**84.48**	83.02	81.88
6	Glass	37.66	**38.27**	**38.45**	37.70	**38.34**
7	Breast tissue	38.39	39.52	**40.37**	39.87	**40.14**
8	Ecoli	**52.17**	**52.85**	**52.41**	49.72	51.23
9	Vehicle	40.25	40.84	**41.56**	39.26	39.54
10	Segmentation	**61.34**	**61.74**	**61.97**	**61.12**	60.97
11	Sat. image	64.43	**66.97**	**66.45**	65.36	65.97

CAS، یکی از قوی‌ترین کارهای این حوزه است و اغلب کارهای مشابه، روش خود را با آن مقایسه کرده‌اند. روش CAS برای گروه‌بندی از الگوریتم خوشه‌بندی Spectral استفاده می‌کند؛ اما به‌نظر می‌رسد برای گروه‌بندی

خوشه‌بندی‌ها نیازی به خوشه‌بندی مجدد نیست و در روش پیشنهادی ما برای گروه‌بندی روش ساده‌تری بر مبنای پیداکردن نزدیک‌ترین خوشه‌بندی پیشنهاد شده است. به‌علاوه روش CAS برای انتخاب زیرمجموعه‌ای از مجموع خوشه‌بندی‌های پایه نیازمند یک پارامتر ورودی است و از قبل از اجرای الگوریتم، تعداد خوشه‌بندی‌هایی که قرار است، انتخاب شوند، باید مشخص باشد. یک راه برای عبور از این چنین معضلی، آزمایش بازهٔ وسیعی از این پارامتر و نشان‌دادن خروجی به‌ازای هر مقدار از این پارامتر است. اما این یک راه حل کلی محسوب نمی‌شود. بنابراین نیاز به حذف این پارامتر، یک نیاز جدی است. در روش پیشنهادی ما، چنین پارامتری نیاز نیست. همچنین در روش CAS برای انتخاب از هر گروه از معیار SNMI استفاده می‌کند که در [4] برای نخستین‌بار معرفی شده است. اما در روش پیشنهادی ما، علاوه‌بر SNMI از BD ، CH و SI برای ارزیابی کیفیت استفاده شده است و روی نتایج آنها رأی‌گیری به‌عمل آمده است که این خود موجب افزایش دقّت خواهد شد. همچنین روش پیشنهادی با روش SR نیز مقایسه شده که یکی از جدیدترین و قوی‌ترین روش‌های موجود است.

دو روش CAS و SR نیازمند این هستند که تعداد خوشه‌بندی‌هایی که باید انتخاب شوند، به آنها داده شود. ازاین‌رو در جداول (۲ تا ۴) میانگین دقّت این روش‌ها به‌ازای انتخاب تعداد ۹۰ ... و ۳۰ و ۲۰ و ۱۰ خوشه‌بندی از مجموعهٔ اولیه محاسبه شده و در جدول قرار داده شده است.

همان‌طوری‌که جدول (۲) نشان می‌دهد، برای تابع اجماع CSPA، بهترین روش NMISelective است که به جز دو مجموعه‌داده Glass و Breast tissue در همه موارد بالاترین میزان دقّت را به‌دست آورده است. البته مقدار به‌دست‌آمده برای این دو مجموعه‌داده نیز از روش ترکیب کامل بیشتر است. این نتایج نشان می‌دهد که با استفاده از روش NMISelective هرگز میزان دقّت به‌دست‌آمده کمتر از ترکیب کامل نخواهد بود و در اغلب موارد مقدار دقّت بیشتری نسبت به CAS و SR ایجاد شده است. روش ARISelective نیز در همه مجموعه‌داده‌ها مقدار دقّت کمتر از ترکیب کامل نداشته است.

برطبق جدول (۳)، بهترین روش برای تابع اجماع MCLA نیز روش NMISelective است، چرا که به‌جز مجموعه‌داده Soybean و Ecoli و Sat. image برای تمامی مجموعه‌داده‌ها بالاترین مقدار را به‌دست آورده و برای این مجموعه‌داده‌ها نیز مقداری بالاتر از ترکیب کامل دارد. این نتایج نیز نشان می‌دهد که با استفاده از روش NMISelective هرگز میزان دقّت به‌دست‌آمده کمتر از ترکیب کامل نخواهد بود و در همه موارد مقداری بیشتر از CAS و به‌جز مجموعه‌داده Breast tissue از SR مقدار دقّت بیشتری را ایجاد کرده است. روش ARISelective نیز به جز در مورد مجموعه‌داده Sonar مقدار دقّت کمتر از ترکیب کامل نداشته است و در اغلب موارد در سطح بالاترین مقدار قرار دارد.

نتایج جدول (۴) نشان می‌دهد که برای تابع اجماع HGPA نیز روش‌های NMISelective و ARISelective بهترین روش‌ها هستند. روش NMISelective برای همهٔ مجموعه‌داده‌ها به جز دو مجموعه‌داده Wine و WDBC مقداری بالاتر از ترکیب کامل داشته و در اغلب موارد در سطح بالاترین مقدار است. روش ARISelective نیز به‌جز دو مجموعه‌داده Sonar و Wine مقدار دقّت بالاتری نسبت به ترکیب کامل داشته و در اغلب موارد در سطح بالاترین مقدار قرار دارد.

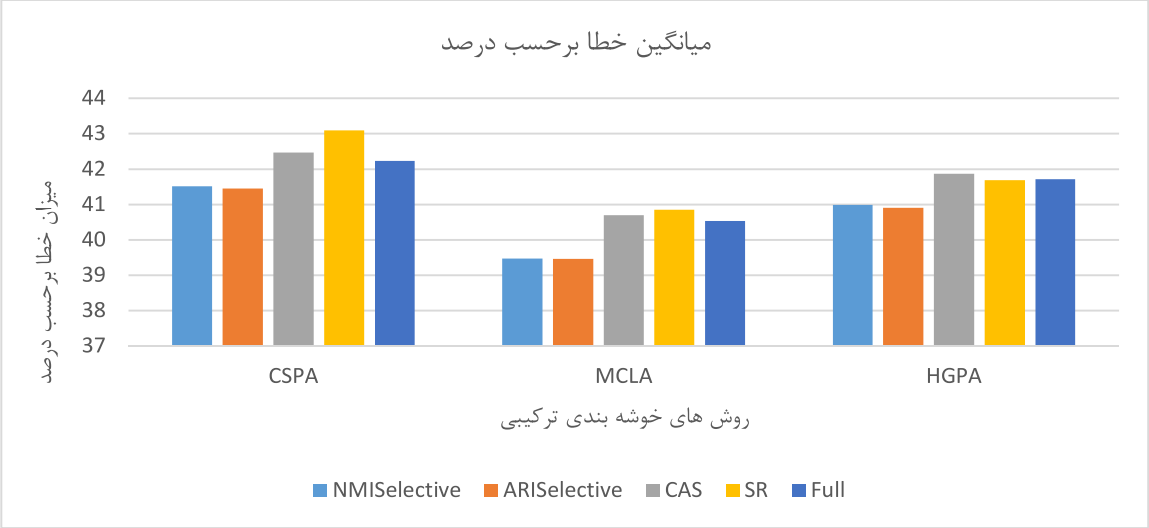
همان‌طوری‌که آزمایش‌ها نشان می‌دهد با کاهش اعضای ترکیب، نه تنها کیفیت و دقّت خوشه‌بندی کاهش نمی‌یابد، بلکه در موارد بسیاری افزایش دقّت نیز رخ می‌دهد. آزمایش‌ها نشان می‌دهد که روش پیشنهادی در اغلب موارد از ترکیب کامل بهتر بوده و رفتاری قابل مقایسه با روش CAS و SR داشته و در بیشتر مواقع عملکردی بهتر از آن‌ها دارد.

یکی دیگر از روش‌های مقایسه عملکرد خوشه‌بندها، مقایسه میزان خطای خوشه‌بندهاست. شکل (۴) میزان میانگین خطای خوشه‌بندی ۱۱ مجموعه‌دادهٔ بالا را نشان می‌دهد.

همان‌طوری‌که نمودار نشان می‌دهد، روش‌های NMISelective و ARISelective به‌ازای هر سه تابع اجماع، میانگین درصد خطای کمتری دارد.

۶- نتیجه‌گیری

در این مقاله یک روش جدید برای انتخاب اعضای ترکیب در خوشه‌بندی ترکیبی ارائه و نشان داده شد که انتخاب درست خوشه‌بندی‌ها بر اساس کیفیت و تنوع آنها می‌تواند نتایجی بهتر از ترکیب کامل ایجاد کند. در روش پیشنهادی از یک الگوریتم جدید برای گروه‌بندی خوشه‌بندی‌ها استفاده شد. ضمن عمل گروه‌بندی می‌توان با حذف خوشه‌بندی‌های مشابه، معیار تنوع را تضمین کرد. هم‌چنین یک روش جدید برای انتخاب خوشه‌بندی‌ها در هر گروه پیشنهاد شد که براساس شاخص‌های ارزیابی کیفیت عمل می‌کرد. آزمایش‌ها روی مجموعه‌داده‌های مختلف برگرفته از UCI نشان داد که روش پیشنهادی عملکرد مناسبی از لحاظ دقّت داشته و نتایج به‌مراتب بهتری نسبت به ترکیب کامل، CAS و SR دارد.



(شکل-۴): خطای خوشه‌بندی به ازای توابع اجماع مختلف

(Figure-4): Cluster error rate for different consensus functions

۷-مراجع

[۱] فضل ارثی، احسان و کاظمی نوقابی، مسعود، "خوشه‌بندی داده‌ها بر پایه شناسایی کلید" *فصلنامه پردازش علائم و داده‌ها*؛ ۱۴ (۴): ۴۲-۳۱ ؛ ۱۳۹۶.

7- References

[1] Fazl Ersi, Ehsan and Kazemi Noghabi, Masoud, "Clustering of Data Based on Key Identification",*Journal of Signals and Data Processing (JSDP)*; 14 (4): 31-42; 2017.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR),* vol. 31, pp. 264-323, 1999.

[3] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol ۳, pp. 1, 2009.

[4] A. Strehl and J. Ghosh, "Cluster ensembles---a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research,* vol. 3, pp. 583-617, 2002.

[5] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, pp. 91-118, 2003.

[6] C. C. Aggarwal and C. K. Reddy, *Data clustering: algorithms and applications*: CRC Press, 2013.

[7] R. Avogadri and G. Valentini, "Fuzzy ensemble clustering based on random projections for DNA microarray data analysis," *Artificial Intelligence in Medicine,* vol. 45, pp. 173-183, 2009.

[8] S. Mimaroglu and E. Erdil, "Obtaining better quality final clustering by merging a collection of clusterings," *Bioinformatics*, vol. 26, pp. 2645-2646, 2010.

[9] X. Ma, W. Wan, and L. Jiao, "Spectral clustering ensemble for image segmentation," in *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, 2009, pp. 415-420.

[10] E. Akbari, H. M. Dahlan, R. Ibrahim, and H. Alizadeh, "Hierarchical cluster ensemble selection," *Engineering Applications of Artificial Intelligence*, vol. 39, pp. 146-156, 2015.

[11] A. L. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, pp. 835-850, 2005.

[12] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, pp. 1866-1881, 2005.

[13] V. Berikov, "Weighted ensemble of algorithms for complex data clustering," *Pattern Recognition Letters*, vol. 38, pp. 99-106, 2014.

[14] Y. Hong, S. Kwong ,Y. Chang, and Q. Ren, "Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm," *Pattern Recognition*, vol. 41, pp. 2742-2756, 2008.

[15] B. Minaei-Bidgoli, A. Topchy, and W. F. Punch, "Ensembles of partitions via data re-sampling," in *Information Technology: Coding and Computing, 2004: Proceedings. ITCC 2004. International Conference on, 2004*, pp. 188-192.

[16] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial intelligence*, vol. 137, pp. 239-263, 2002.

[17] X. Z. Fern and W. Lin, "Cluster ensemble selection," *Statistical Analysis and Data Mining*, vol. 1, pp. 128-141, 2008.

[18] X. Wang, D. Han, and C. Han, "Rough set based cluster ensemble selection," in *Informa-tion Fusion (FUSION), 2013 16th International Conference on*, 2013, pp. 438-444.

[19] J. Azimi and X. Fern, "Adaptive Cluster Ensemble Selection," in *IJCAI*, 2009, pp. 992-997.

[20] L. I. Kuncheva and S. T. Hadjitodorov, "Using diversity in cluster ensembles," in *Systems, man and cybernetics, 2004 IEEE international conference on*, 2004, pp. 1214-1219.

[21] M. C. Naldi, A. Carvalho, and R. J. Campello, "Cluster ensemble selection based on relative validity indexes," *Data Mining and Knowledge Discovery*, vol. 27, pp. 259-289, 2013.

[22] H. Alizadeh, B. Minaei-Bidgoli, and H. Parvin, "To improve the quality of cluster ensembles by selecting a subset of base clusters," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 26, pp. 127-150, 2015.

[23] B. Minaei-Bidgoli, H. Parvin, H. Alinejad-Rokny, H. Alizadeh, and W. F. Punch, "Effects of resampling method and adaptation on clustering ensemble efficacy," *Artificial Intelligence Review*, vol. 41, pp. 27-48, 2014.

[24] G. Karypis and V .Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM Journal on scientific Computing*, vol. 20, pp. 359-392, 1998.

[25] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel hypergraph partitioning: applications in VLSI domain," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 7, pp. 69-79, 1999.

[26] X. Lu, Y. Yang, and H. Wang, "Selective clustering ensemble based on covariance," in *International Workshop on Multiple Classifier Systems* , pp. 179-189, 2013.

[27] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193-218, 1985.

[28] D. A. Neumann and V. T. Norton, "Clustering and isolation in the consensus problem for partitions," *Journal of classification*, vol. 3, pp. 281-297, 1986.

[29] F. Yang, X. Li, Q. Li, and T. Li, "Exploring the diversity in cluster ensemble generation: Random sampling and random projection," *Expert Systems with Applications*, vol. 41, pp. 4844-4866, 2014.

[30] J. Jia, X. Xiao, B .Liu, and L. Jiao, "Bagging-based spectral clustering ensemble selection," *Pattern Recognition Letters*, vol. 32, pp. 1456-1467, 2011.

[31] J. Jia, X. Xiao, and B. Liu, "Similarity-based spectral clustering ensemble selection," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, 2012, pp. 1071-1074.

[32] A. Banerjee, "Leveraging frequency and diversity based ensemble selection to consensus clustering," in *Contemporary Computing (IC3), 2014 Seventh International Conference on*, 2014, pp. 123-129.

[33] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, pp. 224-227, 1979.

[34] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, pp. 1-27, 1974.

[35] W. S. Sarle, "Finding Groups in Data: An Introduction to Cluster Analysis," *Journal of the American Statistical Association*, vol. 86, pp. 830-833, 1991.

[36] M. Charrad, Y. Lechevallier, M. B. Ahmed, and G. Saporta, "On the Number of Clusters in Block Clustering Algorithms," in *FLAIRS Conference*, 2010.

[37] K. Bache and M. Lichman, "UCI machine learning repository," 2013.

[38] A. L. Fred and A. K. Jain, "Data clustering using evidence accumulation," in *Pattern Recognition, 2002. Proceedings. 16th Inter-na-tional Conference on*, 2002, pp. 276-280.

علیرضا لطیفی پاکدهی مدرک کارشناسی خود را در رشتهٔ مهندسی کامپیوتر-نرم‌افزار در سال ۱۳۹۰ از دانشگاه بین‌المللی امام خمینی (ره) و در سال ۱۳۹۵ مدرک کارشناسی ارشد خود را در دانشگاه تربیت دبیر شهید رجایی اخذ کرده است. موضوع پایان‌نامه ایشان، خوشه‌بندی داده‌های با ابعاد بالا با استفاده از ترکیب الگوریتم‌ها بوده است.
نشانی رایانامهٔ ایشان عبارت است از:
Alireza.latifi@yahoo.com

نگین دانشپور استادیار دانشکدهٔ مهندسی کامپیوتر دانشگاه تربیت دبیر شهید رجایی است. نامبرده تحصیلات خود را در مقطع کارشناسی مهندسی کامپیوتر-سخت‌افزار در سال ۱۳۷۸ در دانشگاه شهیدبهشتی و کارشناسی ارشد مهندسی کامپیوتر-نرم‌افزار در سال ۱۳۸۱ در دانشگاه صنعتی امیرکبیر به پایان رسانده و در سال ۱۳۸۹ دکترای خود را در رشته مهندسی کامپیوتر-نرم‌افزار از دانشگاه صنعتی امیرکبیر اخذ کرده است. زمینه‌های پژوهشی مورد علاقهٔ ایشان عبارتند از: پایگاه داده، پایگاه دادهٔ تحلیلی، سیستم‌های تصمیم‌یار، پیش‌پردازش داده و داده‌کاوی.
نشانی رایانامهٔ ایشان عبارت است از:
ndaneshpour@sru.ac.ir