# ترکیب وزن‌دار خوشه‌بندی‌ها با هدف افزایش صحّت خوشه‌بندی نهایی

صدیقه وحیدی فردوسی و حسین امیرخانی*
گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه قم، قم، ایران

## چکیده

**با توجه به ماهیت بدون ناظر مسائل خوشه‌بندی و تأثیرگذاری مؤلفه‌های مختلف از جمله تعداد خوشه‌ها، معیار فاصله و الگوریتم انتخابی، ترکیب خوشه‌بندی‌ها برای کاهش تأثیر این مؤلفه‌ها و افزایش صحت خوشه‌بندی نهایی معرفی شده است. در این مقاله، روشی برای ترکیب وزن‌دار خوشه‌بندی‌های پایه با وزن‌دهی به خوشه‌بندی‌ها بر اساس روش AD ارائه شده است. روش AD برای برآورد صحّت انسان‌ها در مسائل جمع‌سپاری از هماهنگی یا تضاد بین آرای آنها استفاده می‌کند و با پیشنهاد مدلی احتمالاتی، فرآیند برآورد صحّت را به‌کمک یک فرآیند بهینه‌سازی انجام می‌دهد. نوآوری اصلی این مقاله، تخمین صحت خوشه‌بندی‌های پایه با استفاده از روش AD و استفاده از صحت‌های تخمین زده‌شده در وزن‌دهی به خوشه‌بندی‌های پایه در فرآیند ترکیب است. نحوهٔ تطبیق مسألهٔ خوشه‌بندی به روش برآورد صحّت AD و نحوهٔ استفاده از صحّت‌های برآوردشده در فرآیند ترکیب نهایی خوشه‌ها، از چالش‌هایی است که در این پژوهش به آنها پرداخته شده است. چهار روش برای تولید خوشه‌بندی‌های پایه شامل الگوریتم‌های متفاوت، معیارهای فاصله‌ی متفاوت در اجرای k-means، ویژگی‌های توزیع‌شده و تعداد خوشه‌های متفاوت بررسی شده است. در فرآیند ترکیب، قابلیت وزن‌دهی به الگوریتم‌های خوشه‌بندی ترکیبی CSPA و HGPA اضافه شده است. نتایج روش پیشنهادی روی سیزده مجموعه دادهٔ مصنوعی و واقعی مختلف و بر اساس نُه معیار ارزیابی متفاوت نشان می‌دهد که روش ترکیب وزن‌دار ارائه‌شده در بیش‌تر موارد بهتر از روش ترکیب خوشه‌بندی بدون وزن عمل می‌کند که این بهبود برای روش HGPA نسبت به CSPA بیشتر است.**

**واژگان کلیدی: خوشه‌بندی ترکیبی وزن‌دار، یادگیری بدون نظارت، HGPA، CSPA، AD.**

# Weighted Ensemble Clustering for Increasing the Accuracy of the Final Clustering

**Sedigheh Vahidi Ferdosi & Hossein Amirkhani***
Computer Engineering and Information Technology, University of Qom, Qom, Iran

## Abstract

**Clustering algorithms are highly dependent on different factors such as the number of clusters, the specific clustering algorithm, and the used distance measure. Inspired from ensemble classification, one approach to reduce the effect of these factors on the final clustering is ensemble clustering. Since weighting the base classifiers has been a successful idea in ensemble classification, in this paper we propose a method to use weighting in the ensemble clustering problem. The accuracies of base clusterings are estimated using an algorithm from crowdsourcing literature called agreement/disagreement method (AD). This method exploits the agreements or disagreements between different labelers for estimating their accuracies. It assumes different labelers have labeled a set of samples, so each two persons have an agreement ratio in their labeled samples. Under some independence assumptions, there is a closed-form formula for the agreement ratio between two labelers based on their accuracies. The AD method estimates the labelers' accuracies by**

* Corresponding author
*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۶/۰۹/۰۹ • تاریخ پذیرش: ۱۳۹۹/۰۳/۱۹ • تاریخ انتشار: ۱۳۹۹/۰۶/۲۴ • نوع مطالعه: پژوهشی





[ DOI: 10.29252/jsdp.17.2.100 ]

**minimizing the difference between the parametric agreement ratio from the closed-form formula and the agreement ratio from the labels provided by labelers. To adapt the AD method to the clustering problem, an agreement between two clusterings are defined as having the same opinion about a pair of samples. This agreement can be as either being in the same cluster or being in different clusters. In other words, if two clusterings agree that two samples should be in the same or different clusters, this is considered as an agreement. Then, an optimization problem is solved to obtain the base clusterings' accuracies such that the difference between their available agreement ratios and the expected agreements based on their accuracies is minimized. To generate the base clusterings, we use four different settings including different clustering algorithms, different distance measures, distributed features, and different number of clusters. The used clustering algorithms are mean shift, k-means, mini-batch k-means, affinity propagation, DBSCAN, spectral, BIRCH, and agglomerative clustering with average and ward metrics. For distance measures, we use correlation, city block, cosine, and Euclidean measures. In distributed features setting, the k-means algorithm is performed for 40%, 50%,…, and 100% of randomly selected features. Finally, for different number of clusters, we run the k-means algorithm by k equals to 2 and also 50%, 75%, 100%, 150%, and 200% of true number of clusters. We add the estimated weights by the AD algorithm to two famous ensemble clustering methods, i.e., Cluster-based Similarity Partitioning Algorithm (CSPA) and Hyper Graph Partitioning Algorithm (HGPA). In CSPA, the similarity matrix is computed by taking a weighted average of the opinions of different clusterings. In HGPA, we propose to weight the hyperedges by different values such as the estimated clustering accuracies, size of clusters, and the silhouette of clusterings. The experiments are performed on 13 real and artificial datasets. The reported evaluation measures include adjusted rand index, Fowlkes-Mallows, mutual index, adjusted mutual index, normalized mutual index, homogeneity, completeness, v-measure, and purity. The results show that in the majority of cases, the proposed weighted-based method outperforms the unweighted ensemble clustering. In addition, the weighting is more effective in improving the HGPA algorithm than CSPA. For different weighting methods proposed for HGPA algorithm, the best average results are obtained when we use the accuracies estimated by the AD method to weight the hyperedges, and the worst results are obtained when using the normalized silhouette measure for weighting. Finally, among different methods for generating base clusterings, the best results in weighted HGPA are obtained when we use different clustering algorithms to come up with different base clusterings.**

**Keywords: Weighted Ensemble Clustering, Unsupervised Learning, HGPA, CSPA, AD**

## ۱- مقدمه

خوشه‌بندی[1] به فرآیندی گفته می‌شود که در آن مجموعه‌ای از اشیا به زیرمجموعه‌هایی به نام خوشه، بخش‌بندی می‌شوند؛ به‌طوری‌که، اعضای هر خوشه به یکدیگر شبیه و با اعضای دیگر خوشه‌ها متفاوت‌اند. ویژگی متمایزکنندهٔ تحلیل خوشه نسبت به دسته‌بندی[2]، گروه‌بندی بدون نظارت داده‌ها است. این فرآیند سودمند می‌تواند موجب یافتن گروه‌ها و الگوهای ناشناخته‌ای در داده‌ها شود [1].

امروزه استفاده از خوشه‌بندی در بسیاری از حوزه‌ها همچون داده‌کاوی، یادگیری ماشین، جستجوی وب، بازیابی اطلاعات، تشخیص الگو، امنیت و غیره رو به افزایش است. با وجود حجم بالا و گونه‌های مختلفی از داده و داده‌هایی با ابعاد بالا تلاش‌های بسیاری برای یافتن روش‌های کارآمد انجام شده است.

اگرچه روش‌های خوشه‌بندی برای خلاصه‌سازی و فهم حجم انبوه داده‌ها بسیار مناسب هستند، اما همواره پژوهش‌گران با چالش‌های آن نیز روبه‌رو بوده‌اند. از نخستین مسایلی که می‌توان به آن اشاره کرد، انتخاب الگوریتم خوشه‌بندی مناسب از میان روش‌های متفاوت است که معیار ارزیابی یکتایی برای تمایزدادن بین آنها وجود ندارد. همچنین نبود معیار سنجش استانداردی برای نتایج خوشه‌بندی موجب شده بدون هیچ دانش قبلی در مورد بهترین روش، هر نتیجه‌ای معقول به نظر برسد [2].

نتایج متفاوت حاصل از اجرای الگوریتم‌های خوشه‌بندی مختلف بر روی مجموعهٔ یکسانی از داده‌ها، حاکی از تمرکز هر کدام از آنها بر جنبه‌ای متفاوت از بهینگی است. دردسترس‌نبودن خوشه‌بندی ایده‌آل، اعتبارسنجی خوشه‌بندی‌ها را مشکل ساخته است. علاوه‌براین بسیاری از الگوریتم‌ها نیازمند تنظیم پارامترهای ورودی از جانب کاربر هستند که این مسأله کنترل کیفیت خوشه‌بندی را دشوار می‌سازد. ابعاد بالای داده نیز از مواردی است که عملکرد خوشه‌بندی را از نظر پیچیدگی زمانی به چالش کشیده است.

از آن‌جایی‌که پژوهش‌های انجام‌شده در زمینهٔ دسته‌بندی ترکیبی نمایان‌گر بهبود نتایج بوده [3-10]، منطقی است که این روش در خوشه‌بندی نیز مورد بررسی

[1] clustering
[2] classification





[ DOI: 10.29252/jsdp.17.2.100 ]

قرار بگیرد. خوشه‌بندی ترکیبی شامل تجمیع چندین افراز است که این افرازها می‌توانند نتیجهٔ اجرای یک الگوریتم با شرط‌های آغازین گوناگون، یا نمونه‌برداری‌های متفاوت از مجموعهٔ داده‌ها، و یا نتایج اجرای الگوریتم‌های مختلف بر روی مجموعه دادهٔ یکسان باشند.

روش‌های خوشه‌بندی ترکیبی می‌توانند راه‌حل‌های قدرتمند و پایداری برای چالش‌های ذاتی خوشه‌بندی ارائه دهند و بر مشکلات خوشه‌بندی غلبه کنند. این روش‌ها با استفاده از اجماع بین نتایج چندین خوشه‌بندی به متعادل‌کردن ساختارهای ناخواسته‌ی ناشی از تنظیمات متفاوت هر الگوریتم، به‌وسیلهٔ واریانس ایجادشده در اثر نمونه‌گیری‌های گوناگون از داده می‌پردازند [11]. در پایگاه داده‌های توزیع‌شده که امکان اجرای خوشه‌بندی به‌صورت یکجا بر روی داده‌ها وجود ندارد، می‌توان خوشه‌بندی را برای هر پایگاه داده اجرا و تنها برچسب خوشه‌ها را به یک محل مرکزی منتقل کرد؛ سپس با استفاده از الگوریتم‌های خوشه‌بندی ترکیبی، افراز نهایی به‌دست می‌آید [12]. در کنار مزیت‌های خوشه‌بندی ترکیبی که ذکر شد، خود این روش نیز با چالش‌هایی روبه‌رو است و عمده‌ترین مسأله نحوهٔ ترکیب خوشه‌بندی‌ها است. در ترکیب خوشه‌ها ممکن است با مسألهٔ سازگارنبودن نتایج مواجه شویم و همهٔ روش‌ها نظر یکسانی در مورد هم‌خوشه‌بودن دو نمونه نداشته باشند.

در این مقاله این ایده استفاده می‌شود که با نظرات خوشه‌بندی‌های مختلف به‌طور یکسان برخورد نشود و برای هر کدام وزنی در نظر گرفته شود. این وزن‌دهی می‌تواند براساس صحت[1] هر یک از الگوریتم‌ها باشد، در این صورت چگونگی محاسبهٔ صحت در خوشه‌بندی‌ها که روش‌های بدون ناظر هستند از مشکلات پیش روست. مرجع [13] روش AD[2] را جهت برآورد صحّت انسان‌ها در مسائل جمع‌سپاری[3] ارائه داده است. این روش، از هماهنگی یا تضاد بین نظرات انسان‌ها به‌عنوان یک منبع مفید اطلاعات استفاده کرده است و با پیشنهاد مدلی احتمالاتی، فرآیند برآورد صحّت را به‌کمک یک فرآیند بهینه‌سازی انجام می‌دهد. در این مقاله، در فرآیند ترکیب خوشه‌بندی‌ها وزن‌دهی به خوشه‌ها براساس روش AD انجام می‌شود تا بتوانیم صحّت خوشه‌بندی نهایی را افزایش دهیم. نحوهٔ تطبیق مسألهٔ خوشه‌بندی به روش برآورد صحّت AD و نحوهٔ استفاده از صحّت‌های برآوردشده در فرآیند ترکیب نهایی خوشه‌ها، از چالش‌هایی است که در این پژوهش به آنها پرداخته شده است.

در ادامه ابتدا ادبیات این موضوع در بخش ۲ تعریف و پژوهش‌های انجام‌شده در این زمینه بررسی شده است. در بخش ۳ الگوریتم‌هایی شرح داده شدند که در روش پیشنهادی مورد استفاده قرار گرفتند. در بخش ۴ به شرح الگوریتم پیشنهادی پرداخته خواهد شد و نتایج تجربی حاصل از آزمایش‌های انجام‌شده در بخش ۵ گزارش می‌شود. درنهایت نتیجه‌گیری و پیشنهادهایی جهت کارهای آینده بیان شده است.

## ۲- ادبیات پژوهش و مرورکارهای پیشین

ایدهٔ اصلی ترکیب مدل‌های آموزشی و منابع داده در رشته‌های متعددی وجود دارد. در روش‌های دسته‌بندی ترکیبی مدلی جهت دسته‌بندی ساخته می‌شود که ترکیب چندین دسته‌بندی است. درواقع، برچسب دستهٔ یک نمونهٔ جدید براساس تجمیع آرای دسته‌بندی‌های اولیه تعیین می‌شود. از جمله روش‌های رایج در دسته‌بندی ترکیبی می‌توان به روش bagging، boosting و جنگل‌های تصادفی اشاره کرد [1].

الگوریتم آرای بیشینهٔ وزن‌دار[4]، در سال ۱۹۹۴ ارائه شده است [14] این روش با استفاده از پیش‌بینی‌های انجام‌شده توسط مجمعی[5] از الگوریتم‌ها با خطای کمتری برچسب مورد نظر را پیش‌بینی می‌کند. روش دیگری مشابه این روش با نام voted-perceptron نیز پیشنهاد شده است [15]. در دسته‌بندی ترکیبی نسبت به خوشه‌بندی ترکیبی مطالعات بیشتری صورت گرفته است [19-16]. در میان روش‌هایی که درهمین‌اواخر در این زمینه مطرح شده است، می‌توان به روش AD اشاره کرد [13]. این روش نیز از مفهوم رأی‌گیری بیشینه در آرای مجمعی از انسان‌ها استفاده می‌کند و برای وزن‌دهی نظرات، روشی را جهت برآورد صحّت انسان‌ها در مسائل جمع‌سپاری معرفی کرده است.

بخش عمدهٔ پژوهش‌های انجام‌شده در زمینهٔ خوشه‌بندی ترکیبی، طی دههٔ اخیر صورت گرفته که نمایان‌گر نوپابودن این بخش است. دو گام اساسی در روش‌های مطرح‌شده برای خوشه‌بندی ترکیبی، گام تولید[6] و گام تابع توافقی[7] است (شکل ۱).

در گام تولید که نخستین مرحله است، مجموعه‌ای از

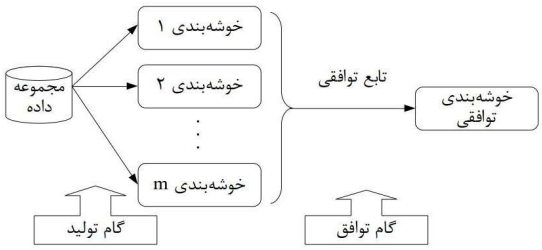
[1] accuracy
[2] agreement/disagreement
[3] crowdsourcing

[4] weighted majority algorithm
[5] ensemble
[6] generation
[7] consensus function




[ DOI: 10.29252/jsdp.17.2.100 ]

خوشه‌بندی‌ها از مجموعهٔ داده تولید می‌شوند؛ سپس در مرحلهٔ دوم باید نتایج حاصل از خوشه‌بندی‌های پایه به‌وسیلهٔ یک تابع توافق با یکدیگر تجمیع شوند تا خوشه‌بندی نهایی به‌دست آید.




**(شکل-۱): نمودار فرآیند کلی خوشه‌بندی ترکیبی [2]**
**(Figure-1): The general process of ensemble clustering [2]**

*استرل* و *گاش*[1] جزء نخستین کسانی بودند که ترکیب را در خوشه‌بندی مورد پژوهش قرار دادند [12]. آنها در پژوهش خود سه روش افرازبندی مبتنی بر گراف، جهت ساخت تابع توافقی پیشنهاد داده‌اند: الگوریتم افرازبندی با شباهت مبتنی بر خوشه (CSPA)[2]، الگوریتم افرازبندی ابرگراف (HGPA)[3] و الگوریتم متا خوشه‌بندی (MCLA)[4]. در این روش‌ها ابتدا مجموعهٔ خوشه‌بندی‌های پایه به‌صورت یک ابرگراف مدل و محاسبات دیگر با استفاده از این ابرگراف انجام می‌شود.

حدود یک سال بعد *فرد* و *جین*[5] به بررسی ایدهٔ جمع‌آوری شواهد برای ترکیب نتایج چندین خوشه‌بندی پرداختند [20]. آنها برای ساخت تابع توافق، افرازهای اشیا را به یک ماتریس همبستگی[6] از الگوهای تکراری نگاشت کردند. روش‌های رأی‌گیری در مراجع [21] و [22] برای ساختن تابع توافقی به‌عنوان جایگزینی برای ماتریس همبستگی مطرح شده‌اند. مرجع [23] تابع توافقی را بر اساس اصل تنگنای اطلاعات[7] می‌سازد.

مرجع [24] دو روش برای تولید افرازهای گوناگون، با عنوان الگوریتم‌های خوشه‌بندی ضعیف[8] معرفی می‌کند. در پژوهش دیگری با استفاده از اجرای الگوریتم k-means با تعداد تصادفی خوشه‌ها، افرازهای گوناگون تولید و به‌وسیلهٔ ماتریس تطابق بین تمام جفت اشیا افرازها ترکیب می‌شوند [25]. *مینایی* و همکارانش روشی مبتنی بر نمونه‌برداری مجدد پیشنهاد داده‌اند [26]. *تاپچی*[9] و همکارانش نیز چارچوبی یکپارچه برای ایجاد افرازهای متعدد ارائه داده‌اند [27].

*علیزاده* و همکاران برای خوشه‌بندی ترکیبی روشی مبتنی بر زیرمجموعه‌ای از خوشه‌های پایه پیشنهاد داده‌اند [28]. در این روش تنها خوشه‌های پایدار در ترکیب نهایی مؤثر هستند و برای ساخت تابع توافقی مبتنی بر ماتریس همبستگی، روش خوشه‌بندی انباشت مدارک را توسعه داده‌اند. مرجع [29] با تعبیهٔ خوشه‌بندی در فضاهای برداری، مسألهٔ پیچیدهٔ خوشه‌بندی ترکیبی را به مسألهٔ شناخته‌شدهٔ میانه‌ی اقلیدسی[10] کاهش می‌دهد و به‌وسیلهٔ الگوریتم Weiszfeld آن را حل می‌کند. روش‌های ذکرشده بر روی خوشه‌بندی مسطح تمرکز دارند؛ اما *میرزایی* در رسالهٔ خود به معرفی ترکیب خوشه‌بندی‌های سلسله‌مراتبی پرداخته است [30].

در میان پژوهش‌هایی که در زمینهٔ خوشه‌بندی ترکیبی وزن‌دار صورت گرفته است، می‌توان به روش خوشه‌بندی زیرفضایی COSA[11] اشاره کرد [31]. این روش به ازای هر داده یک بردار وزن برای ویژگی‌ها تعریف می‌کند که در ابتدا وزن تمام ویژگی‌ها یکسان است. فرآیند محاسبهٔ فواصل داده‌ها با توجه به وزن ویژگی‌ها، اجرای KNN و به‌روزرسانی وزن ویژگی‌ها تا جایی تکرار می‌شود که کمینه تغییر مورد نظر رخ دهد. الگوریتم LAC نیز مشابه COSA از وزن‌دهی استفاده می‌کند، با این تفاوت که بردار وزن را به خوشه‌ها اختصاص می‌دهد [32].

الگوریتم افرازبندی مشابهت وزن‌دار (WSPA)[12] برای حل مشکل ابعاد بالای داده از الگوریتم LAC استفاده می‌کند و افراز توافقی با افرازبندی گرافی به‌دست می‌آید که از ماتریس شباهتی مبتنی بر بردار احتمال نسبت داده‌شده به هر نمونه ساخته شده است [33]. روش WBPA در ابتدا همانند روش WSPA است، اما در ساخت گراف پایانی همان‌طور که از نام آن مشخص است، یک گراف دوبخشی از نمونه‌ها و خوشه‌ها می‌سازد [33]. WSBPA[13] نخستین رویکرد در خوشه‌بندی زیرفضایی با استفاده از خوشه‌بندی ترکیبی است [11]. این روش توسعه‌ای از روش WBPA است که علاوه‌بر افرازبندی نهایی، بردار وزن ویژگی‌ها برای هر خوشه نیز ارائه می‌شود.

[1] Strehl and Ghosh
[2] cluster-based similarity partitioning algorithm
[3] hypergraph partitioning algorithm
[4] meta-clustering algorithm
[5] Fred and Jain
[6] co-association matrix
[7] information bottleneck principle
[8] weak clustering algorithms
[9] Topchy
[10] euclidean median
[11] clustering objects on subsets of attributes
[12] weighted similarity partitioning algorithm
[13] weighted subspace bipartite partitioning algorithm




[ DOI: 10.29252/jsdp.17.2.100 ]

در روش WKF[1] گام ارزیابی کیفیت افرازبندی‌ها میان گام تولید و تابع توافقی اضافه شده است که موجب تأثیر کمتر افرازهای نوفه با وزن کمتر، در افراز نهایی می‌شود [34]. این روش سپس به GWKF[2] ارتقا پیدا کرد [35] که در آن اینکه کدام افراز دو نمونه را در یک خوشه قرار می‌دهد، حائز اهمیت است و در گام ارزیابی کیفیت افرازبندی‌ها دیگر احتیاجی به دسترسی به داده‌های اصلی وجود ندارد.

در زمینهٔ وزن‌دهی به خوشه‌بندی‌ها مرجع [36] ثابت می‌کند تنظیم $L_1$[3] از مسألهٔ LASSO معادل بهینه‌سازی وزن برای روشی است که جهت خوشه‌بندی ترکیبی وزن‌دار پیشنهاد می‌دهد. *گولو*[4] و همکارانش سه طرح برای تجمیع خوشه‌بندی‌ها بر اساس دو معیار تنوع مبتنی بر F-measure و NMI[5] معرفی کردند: وزن‌دهی تکی[6]، وزن‌دهی گروهی[7] و وزن‌دهی دندروگرام[8] [37].

پژوهش [38] مسألهٔ خوشه‌بندی ترکیبی را به یک مسألهٔ برنامه‌نویسی خطی دودویی[9] تبدیل کرده و راه حلی مبتنی بر گراف فاکتور[10] پیشنهاد داده است. این پژوهش از روشی بدون ناظر برای اعتبار سنجی خوشه‌بندی‌های پایه استفاده کرده و نمایش ممتاز-شیء[11] را برای داده‌های حاصل تجمیع پیشنهاد داده است.

مرجع [39] خوشه‌بندی ترکیبی طیفی[12] را برای استفادهٔ بیشینه‌ای از مزایای ماتریس همبستگی پیشنهاد می‌دهد. در این مقاله نشان داده شده که این خوشه‌بندی معادل خوشه‌بندی k-means وزن‌دار است که پیچیدگی آن به‌طور چشم‌گیری کاهش یافته است. در مقالهٔ [40] از وزن‌دهی به اشیا در فرآیند تجمیع خوشه‌بندی استفاده شده است. این وزن‌دهی به‌وسیلهٔ ماتریس همبستگی صورت می‌گیرد که نتایج خوشه‌بندی پایه را خلاصه می‌کند.

مرجع [41] یک مدل هوش گروهی[13] جدید به نام بهینه‌سازی سرگردان صدف‌ها[14] را با خوشه‌بندی k-means ترکیب کرده و روشی با عنوان K-MWO ارائه داده است. این روش از وزن‌دهی به نقاط داده استفاده می‌کند. در میان پژوهش‌های داخلی نیز پروین در رسالهٔ خود روش جدیدی ارائه داده که علاوه‌بر وزن‌دهی خوشه‌های پایه، از وزن‌دهی ویژگی‌های هر خوشه نیز استفاده کرده است [42].

در پژوهش [43] پس از گروه‌بندی خوشه‌بندی‌ها با استفاده از معیارهای شباهت، با کیفیت‌ترین خوشه‌بندی از هر گروه بر اساس شاخص‌های اعتبار خوشه انتخاب می‌شود؛ سپس خوشه‌بندی‌های انتخاب‌شده با استفاده از سه تابع HGPA، CSPA و MCLA ترکیب می‌شوند. مرجع [44] هم از توافق/عدم توافق خوشه‌بندی‌ها برای سنجش اعتبار آنها استفاده می‌کند. از میان پژوهش‌های اخیر نیز می‌توان به مراجع [48-45] اشاره کرد.

با توجه به پژوهش‌های انجام‌شده، برآورد صحّت خوشه‌بندی‌ها با توجه به ماهیت بدون ناظر آنها یکی از چالش‌های روش‌های خوشه‌بندی ترکیبی است. از آنجایی که در روش AD وزن‌دهی صرفاً با استفاده از آرای دسته‌بندها انجام می‌شود، بررسی این روش در زمینهٔ خوشه‌بندی ترکیبی نیز می‌تواند سودمند باشد. در این مقاله، علاوه‌بر معرفی یک روش خوشه‌بندی ترکیبی وزن‌دار مبتنی بر AD، تأثیر عواملی مانند اندازهٔ خوشه‌ها و ارزیابی خوشه‌بندی‌ها با استفاده از یک معیار داخلی نیز بررسی می‌شود.

## ۳- الگوریتم‌های مورد استفاده

### ۳-۱- الگوریتم CSPA

الگوریتم CSPA یکی از روش‌های خوشه‌بندی ترکیبی مبتنی بر گراف است که در مرجع [12] ارائه شده است. اگر r خوشه‌بندی، n داده و k خوشه داشته باشیم، برای بردار برچسب هر یک از خوشه‌بندی‌ها $\lambda^{(q)} \in N^n$، یک ماتریس عضویت دودویی ($H^{(q)}$) می‌سازیم که سطرهای آن داده‌ها یا همان رئوس ابرگراف و ستون‌ها، خوشه‌ها یا ابریال‌ها هستند. مقدار یک در هر ستون بیان‌کنندهٔ عضویت رأس متناظر با آن سطر در این خوشه (ابریال) است. با کنار هم قراردادن ماتریس تمام خوشه‌بندی‌های پایه، بلوکی از ماتریس‌ها به‌دست می‌آید ($H = h^{(1, \dots, r)} = | H^{(1)}, \dots H^{(r)} |$) که خود ماتریس مجاورت یک ابرگراف با n رأس و $\sum_{q=1}^{r} K(q)$ ابریال است. بدین ترتیب هر خوشه به یک ابریال و هر خوشه‌بندی به یک ابرگراف نگاشت می‌شود.

[1] weighted cluster ensemble using a kernel consensus function
[2] generalized WKF
[3] regularization
[4] Gullo
[5] normalized mutual information
[6] single weighting
[7] group weighting
[8] dendrogram weighting
[9] binary linear programming
[10] factor graph
[11] super-object
[12] spectral ensemble clustering
[13] swarm intelligence model
[14] mussels wandering optimization




[ DOI: 10.29252/jsdp.17.2.100 ]

بنابراین، اگر شباهت دو شیء را که در خوشهٔ یکسان هستند یک و غیر این صورت صفر تعریف کنیم، می‌توانیم یک ماتریس شباهت n×n برای هر خوشه‌بندی ایجاد کنیم. در این صورت با محاسبهٔ میانگین بین درایه‌های متناظر ماتریس هر خوشه‌بندی، می‌توان ماتریس شباهتی تعریف کرد که از طریق معادلهٔ (۱) قابل محاسبه است.

$$S = \frac{1}{r} HH^t \quad (۱)$$

رئوس گراف شباهت حاصل‌شده از این ماتریس با اشیا و وزن یال‌ها با مقادیر شباهت دوبه‌دوی اشیا، متناظر است. درنهایت این گراف برای دوباره خوشه‌بندی‌کردن اشیا با استفاده از METIS [49] افرازبندی می‌شود و نتیجهٔ این افراز، تجمیع خوشه‌بندی‌های مختلف را به‌دست می‌دهد.

## ۲-۳- الگوریتم HGPA

این الگوریتم نیز جزء سه روش افرازبندی مبتنی بر گراف است که در مرجع [12] ارائه شده است. HGPA از ابرگرافی که مشابه ابرگراف الگوریتم CSPA به‌دست می‌آید، برای ترکیب خوشه‌بندی‌ها استفاده می‌کند. این الگوریتم سعی در یافتن کمترین تعداد از ابریال‌هایی دارد که قطع آنها ابرگراف را به k بخش مجزا با اندازه‌های به‌طورتقریبی یکسان افرازبندی می‌کند. در این روش تمام یال‌ها و تمام رأس‌ها وزن یکسانی دارند. در افرازبندی ابرگراف از بستهٔ HMETIS [50] استفاده می‌شود که بسیار مقیاس‌پذیر است و افرازهای با کیفیت بالایی تولید می‌کند.

## ۳-۳- الگوریتم AD

این الگوریتم جهت برآورد صحّت انسان‌ها در مسائل جمع‌سپاری ارائه شده است که بر روی مسائل دودسته‌ای تمرکز دارد [13]. فرض کنید $N$ نمونهٔ آموزشی هستند که برچسب‌های درست آنها $\{y_i\}_{i=1}^{N}$ نامشخص است. به جای این برچسب‌های درست، نظرات $R$ انسان را در مورد بعضی از این نمونه‌ها داریم. نظر برچسب‌گذار $j$ در مورد نمونهٔ $i$ با $O_i^j \in \{-1,0,1\}$ نشان داده می‌شود که $-1$ یعنی نظری داده نشده است. صحّت برچسب‌گذار $j$ با $\alpha^j$ نمایش داده می‌شود.

AD شامل دو گام است:

۱. برآورد صحّت برچسب‌گذارها با استفاده از توافقات/اختلافات موجود در نظرات جمع‌آوری شده،

۲. یکی‌کردن نظرات با استفاده از صحّت‌های برآوردشده جهت برآورد برچسب‌های درست.

با توجه به استفاده از گام نخست در این پژوهش، تنها به شرح این گام اکتفا می‌کنیم. برآورد صحّت برچسب‌گذارها در سه بخش انجام می‌شود. در بخش نخست روشی برای برآورد احتمال اختلاف نظر بر حسب برچسب‌های موجود مطرح می‌شود. این احتمال که برای هر جفت برچسب‌گذار محاسبه و $P_i$ نامیده می‌شود، نسبت اختلاف‌نظرها به تعداد کل نمونه‌هایی است که هر دو برچسب‌گذار در مورد آن نظر داده‌اند.

در بخش دوم احتمال اختلاف نظر برحسب صحّت‌های فرضی محاسبه می‌شود. زمانی دو فرد $i$ و $j$ در مورد نمونهٔ $k$ اختلاف نظر پیدا می‌کنند که یکی از دو حالت زیر رخ دهد:

۱. فرد $i$ درست و فرد $j$ غلط برچسب‌گذاری کرده است.

۲. فرد $j$ درست و فرد $i$ غلط برچسب‌گذاری کرده است.

حالت نخست را $E_1$ و حالت دوم را $E_2$ می‌نامیم. بنابراین:

$$P(O_k^i \neq O_k^j) = P(E_1) + P(E_2) = \quad (۲)$$
$$\mathrm{P}(\mathrm{O_k^i} = \mathrm{y_k}, \mathrm{O_k^j} = 1 - \mathrm{y_k}) +$$
$$\mathrm{P}(\mathrm{O_k^i} = 1 - \mathrm{y_k}, \mathrm{O_k^j} = \mathrm{y_k})$$

اگر صحّت فرضی $i$ و $j$ به‌ترتیب $\hat{\alpha}^i$ و $\hat{\alpha}^j$ باشد، پس می‌توان احتمال $E_1$ و $E_2$ را به‌صورت نیز نوشت:

$$P(O_k^i = y_k, O_k^j = 1 - y_k) = \quad (۳)$$
$$P(O_k^i = y_k)P(O_k^j = 1 - y_k) = \hat{\alpha}^i(1 - \hat{\alpha}^j)$$

$$P(O_k^i = 1 - y_k, O_k^j = y_k) = \quad (۴)$$
$$P(O_k^i = 1 - y_k)P(O_k^j = y_k) = (1 - \hat{\alpha}^i)\hat{\alpha}^j$$

در نتیجه معادلهٔ (۲) به این ترتیب ساده‌سازی می‌شود:

$$P(O_k^i \neq O_k^j) = \hat{\alpha}^i + \hat{\alpha}^j - 2\hat{\alpha}^i\hat{\alpha}^j \quad (۵)$$

ماتریس این احتمالات $P_2$ نامیده می‌شود.

بخش سوم در برآورد صحّت برچسب‌گذارها، حل مسألهٔ بهینه‌سازی است. در این بخش با استفاده از یک تابع بهینه‌سازی اختلاف بین $P_1$ و $P_2$ کمینه می‌شود. این تابع، مقدار دلخواهی برای صحّت یک برچسب‌گذار درنظر می‌گیرد و با جای‌گذاری در معادلهٔ $P_2 - P_1 = 0$ صحّت بقیه را برآورد می‌زند و این فرآیند را برای مقادیر مختلف صحّت یک برچسب‌گذار تکرار می‌کند. خروجی تابع بهینه‌سازی بردار صحّت‌های برآوردشده برای برچسب‌گذاران است.





[ DOI: 10.29252/jsdp.17.2.100 ]

با توجه به معادلهٔ (۵) برای برآورد صحّت برچسب‌گذارها، این مسأله دو نقطهٔ بهینه دارد:

$$\hat{\alpha}^i + \hat{\alpha}^j - 2\hat{\alpha}^i\hat{\alpha}^j = (1-\hat{\alpha}^i) + (1-\hat{\alpha}^j) - 2(1-\hat{\alpha}^i)(1-\hat{\alpha}^j) \quad (۶)$$

بنابراین، با جایگزین کردن بردار $(\hat{\alpha}^1,\dots,\hat{\alpha}^R)$ با بردار $(1-\hat{\alpha}^1,\dots,1-\hat{\alpha}^R)$ مقدار تابع هدف تغییر نمی‌کند. در نتیجه، اگر خروجی الگوریتم بهینه‌سازی بردار $\vec{\alpha} = (\hat{\alpha}^1,\dots,\hat{\alpha}^R)$ باشد، بردار $1-\vec{\alpha} = (1-\hat{\alpha}^1,\dots,1-\hat{\alpha}^R)$ هم می‌تواند پاسخ درست باشد. به عبارت دیگر، یکی از این نقطه‌ها بهینهٔ درست و دیگری غلط است. برای مسألهٔ بحث‌شده در این مقاله، روشی جهت تشخیص بهینهٔ درست طراحی و بررسی شده است که در بخش بعد معرفی خواهد شد.

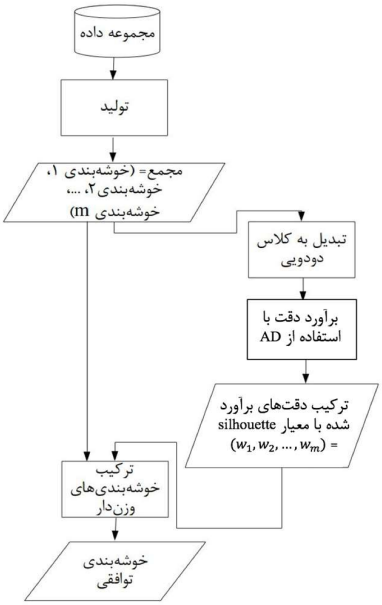
## ۴- روش پیشنهادی

ایدهٔ مطرح‌شده در روش پیشنهادی، استفاده از صحّت‌های خوشه‌بندی‌های پایه به‌عنوان وزن هر یک از آنها در فرآیند ترکیب است. در طراحی روش پیشنهادی سه گام طی می‌شود: تولید خوشه‌بندی‌های پایه، برآورد صحّت هر خوشه‌بندی با استفاده از روش AD و معیار silhouette و ترکیب خوشه‌بندی‌های پایه مبتنی بر وزن‌های برآوردشده. شکل (۲) این مراحل را به تصویر کشیده است.

### ۴-۱- گام نخست: تولید خوشه‌بندی‌های پایه

در گام نخست روش‌های متفاوتی برای تولید خوشه‌بندی‌های پایه مورد بررسی قرار داده شده است. با اجرای هر روش بر روی مجموعه‌ای از داده‌ها، خوشه‌بندی‌های اولیه به‌دست می‌آیند. در ادامه روش‌های مورد استفاده در آزمایش‌ها برای تولید خوشه‌بندی‌های پایه بیان می‌شوند:

- اجرای k-means با kهای متفاوت: در این روش، ترکیب بر روی خوشه‌بندی‌های حاصل از اجرای الگوریتم k-means برای ۲ خوشه و ۵۰٪، ۷۵٪، ۱۰۰٪، ۱۵۰٪ و ۲۰۰٪ تعداد واقعی خوشه‌ها صورت می‌گیرد. خوشه‌بندی توافقی برای تعداد خوشه‌هایی معادل تعداد واقعی به‌دست می‌آید.
- استفاده از الگوریتم‌های متفاوت: در این روش مجموعه‌ای شامل ۹ الگوریتم خوشه‌بندی داریم. هر یک از الگوریتم‌های انتقال میانگین[1]، خوشه‌بندی تجمیعی[2] با دو معیار ward و average، k-means، Mini-batch k-means، DBSCAN، Affinity Propagation، Spectral و BIRCH به طور جداگانه بر روی داده‌ها اجرا می‌شوند؛ سپس نتایج این خوشه‌بندی‌ها برای به‌دست‌آوردن خوشه‌بندی‌نهایی ترکیب می‌شوند.




**(شکل-۲): چارچوب روش پیشنهادی برای ترکیب وزن‌دار خوشه‌بندی‌ها**

**(Figure-2): The proposed method for weighted ensemble clustering**

- خوشه‌بندی ویژگی‌ه
- ای توزیع‌شده (FDC[3]): در روش ویژگی‌های توزیع‌شده، خوشه‌بندی‌های پایه به‌وسیلهٔ اجرای k-means برای زیرمجموعه‌هایی از ویژگی‌ها تولید می‌شوند. این زیرمجموعه‌ها با انتخاب تصادفی ۴۰٪، ۵۰٪، ۶۰٪، ۷۰٪، ۸۰٪، ۹۰٪ و ۱۰۰٪ از ویژگی‌ها به‌دست می‌آیند. درنهایت هفت خوشه‌بندی ایجادشده ترکیب و نتیجهٔ توافقی حاصل می‌شود.
- استفاده از معیارهای فاصلهٔ متفاوت در اجرای k-means: برای استفاده از معیارهای فاصلهٔ متفاوت در تولید خوشه‌بندی‌های پایه، الگوریتم k-means را به‌ازای هر یک از فاصله‌های کسینوسی[4]، اقلیدسی[5]، همبستگی[6] و بلوک

[1] mean shift
[2] agglomerative clustering
[3] feature-distributed clustering
[4] cosine
[5] euclidean
[6] correlation




[ DOI: 10.29252/jsdp.17.2.100 ]

شهری[1] دو بار به‌طور مستقل اجرا می‌کنیم (جدول ۱). سپس الگوریتم ترکیب بر روی ده خوشه‌بندی به‌دست‌آمده اجرا می‌شود. مستند جزئیات پیاده‌سازی‌های این معیارها در پایتون از طریق پایگاه scipy[2] قابل دسترسی است.

**(جدول-۱): معیارهای فاصله مورد استفاده**
**(Table-1): The used distance metrics**

| نام معیار | نحوهٔ محاسبه (برای دو شی $i = (x_{i1}, \dots, x_{ip})$ و $j = (x_{j1}, \dots, x_{jp})$ با $p$ ویژگی عددی) |
|---|---|
| اقلیدسی [1] | $\sqrt{(x_{i1} - x_{j1})^2 + \dots + (x_{ip} - x_{jp})^2}$ |
| بلوک شهری [1] | $\lvert x_{i1} - x_{j1}\rvert + \dots + \lvert x_{ip} - x_{jp}\rvert$ |
| کسینوسی [51] | $d(i,j) = 1 - \frac{i^T.j}{\|i\|.\|j\|}$, $\|i\| = \sqrt{x_{i1}^2 + \dots + x_{ip}^2}$ |
| همبستگی [51] | $d(i,j) = 1 - \frac{(i-\bar{i})^T.(j-\bar{j})}{\|i-\bar{i}\|.\|j-\bar{j}\|}$, $\bar{i} = \frac{x_{i1} + \dots + x_{ip}}{p}$ |

## ۲-۴- گام دوم: برآورد صحت هر خوشه‌بندی با استفاده از روش AD

از آنجایی که روش AD بر روی مسائل دودسته‌ای کار می‌کند باید نتایج خوشه‌بندی‌های پایه را به شکل مناسبی تبدیل کنیم. شکل (۳) الگوریتم این تبدیل را نمایش می‌دهد که در آن ورودی، مجموعه داده (Dataset) و ماتریس خوشه‌بندی‌های پایه در گام تولید (Ensemble) است. در این الگوریتم ماتریسی دودویی (ADinput) در نظر می‌گیریم که سطرها متناظر با خوشه‌بندی‌ها (برچسب‌گذاران) هستند. ستون‌ها نیز، زوج مرتب‌هایی هستند که از مجموعه‌ی داده‌ها ایجاد شده‌اند. مقدار یک برای یک زوج مرتب $(i,j)$ در سطر $k$ به این معنی است که خوشه‌بندی $k$ دو نمونهٔ $i$ و $j$ را در خوشهٔ یکسان قرار داده است و صفر یعنی در خوشهٔ یکسان نیستند. به سبب متقارن‌بودن این ماتریس تنها یک نیمهٔ ماتریس محاسبه می‌شود و عناصر روی قطر اصلی که همگی مقدار یک دارند، به‌منظور کاهش هزینهٔ محاسباتی حذف می‌شوند؛ سپس این ماتریس به‌عنوان ورودی الگوریتم AD قرار داده می‌شود تا صحّت‌های هر خوشه‌بندی برآورد شود.

**Algorithm 1:** Convert ensemble to binary opinions
**Parameters:** $N$ (number of instances), $R$ (number of clusterings)
**Input:** Dataset, Ensemble ($R \times N$)
**Output:** ADinput ($R \times N^2$)
(1) **for** clusterings **in** Ensemble **do**
(2) Index← 0
(3) **for** (i,j) **in** Dataset[2] **do**
(4) **If** clusterings (i) = clusterings (j) **then**
(5) ADinput (labeler)(index) ← 1
(6) **else do**
(7) ADinput (labeler)(index) ← 0
(8) Index← Index+1
(9) **end for**
(10)**end for**

**(شکل-۳): الگوریتم تبدیل مجموعه‌ای از خوشه‌بندی‌ها به نظرات دودویی قابل استفاده در الگوریتم AD**
**(Figure-3): Converting an ensemble of clusterings to binary opinions which can be used in AD algorithm**

### ۱-۲-۴- انتخاب جواب درست از میان دو بهینهٔ روش AD

برای تشخیص بهینهٔ درست از غلط آخرین نمونهٔ موجود در مجموعهٔ داده را بیست مرتبه تکرار و سپس الگوریتم پیشنهادی را یک مرتبه با استفاده از صحّت‌های $\vec{\alpha}$ و بار دیگر با استفاده از $1 - \vec{\alpha}$ مطابق توضیحات بخش قبل بر روی این مجموعه داده اجرا می‌کنیم. با توجه به یکسان‌بودن بیست نمونهٔ آخر، نتایج به‌دست‌آمده تنها برای این بیست نمونه با استفاده از معیار Fowlkes-Mallows[3] مورد سنجش قرار می‌گیرد.

برای تمام مجموعه‌داده‌ها، روش‌های تولید خوشه‌بندی پایه و معیارهای ارزیابی مطرح‌شده، با مقایسهٔ نتایج خوشه‌بندی نهایی و نتایج بیست نمونهٔ تکراری، مشاهده شد هر یک از وزن‌دهی‌ها که برای بیست نمونهٔ تکراری بهتر عمل کرده، در خوشه‌بندی نهایی نیز عملکرد بهتری داشته است. بنابراین بیشترین مقدار معیار Fowlkes-Mallows برای این بیست نمونه نشان‌دهندهٔ استفاده از وزن‌های برآوردی درست است. تمام مواردی که در این فصل به‌عنوان نتیجهٔ وزن‌دهی با استفاده از صحّت‌های برآوردی AD گزارش شده است، وزن‌دهی با جواب درست برمبنای این روش است.

## ۳-۴- گام سوم: ترکیب خوشه‌های پایه مبتنی بر وزن‌های برآوردشده

در این گام خوشه‌بندی‌های پایه با استفاده از روش CSPA و HGPA وزن‌دار ترکیب می‌شوند که در ادامه وزن‌دارکردن آنها را شرح می‌دهیم. در روش CSPA وزن‌دار از صحّت‌های برآوردشده توسط AD به‌عنوان وزن خوشه‌بندی‌ها استفاده می‌شود. در روش HGPA وزن‌دار، برای وزن ابریال‌ها در

[1] city block
[2] http://docs.scipy.org/doc/scipy/reference/spatial.distance.html
[3] http://scikit-learn.org/stable/modules/clustering.html#fowlkes-mallows-scores




[ DOI: 10.29252/jsdp.17.2.100 ]

حالت‌های مختلف ترکیب، سه پارامتر صحّت‌های برآوردشده به‌وسیلهٔ AD، معیار silhouette و اندازهٔ خوشه‌ها بررسی می‌شود. جدول (۲) انواع وزن‌دهی‌های مورد آزمایش برای HGPA را نشان می‌دهد.

**(جدول-۲): شیوه‌های مختلف وزن‌دهی ابریال‌های در HGPA**
**(Table-2): Different ways of weighting hyperedges in HGPA algorithm**

| نماد | وزن ابریال‌ها |
|---|---|
| $w_1$ | $\text{int}(accuracy \times 1000)$ |
| $w_2$ | $\text{int}(accuracy \times 1000) \times Vertices$ |
| $w_3$ | $\text{int}(accuracy \times 1000) \times \text{int}(\frac{Vertices}{N} \times 1000)$ |
| $w_4$ | $\text{int}(silhouette \times 1000) \times \text{int}(\frac{Vertices}{N} \times 1000)$ |
| $w_5$ | $\text{int}(silhouette \times 1000)$ |
| $w_6$ | $\text{int}(accuracy \times silhouette \times 1000)$ |
| $w_7$ | $\text{int}(accuracy \times silhouette \times 1000) \times \text{int}(\frac{Vertices}{N} \times 1000)$ |

پارامتر accuracy در جدول (۲) با جای‌گذاری صحّت‌های برآوردشده توسط AD به‌طور جداگانه آزمایش شده است. از آنجایی‌که HMETIS وزن‌ها را به‌صورت عدد صحیح می‌پذیرد، صحّت را که عددی بین صفر و یک است در هزار ضرب و سپس به عدد صحیح تبدیل کرده‌ایم. پارامتر Vertices تعداد رأس‌های هر ابریال یا به‌عبارتی دیگر تعداد نمونه‌های عضو هر خوشه است؛ زیرا در این الگوریتم هر خوشه به یک ابریال نگاشت می‌شود. $N$ تعداد کل نمونه‌ها است که برای نُرمال‌سازی وزن‌های $w_3$، $w_4$ و $w_7$ استفاده شده است. با توجه به این‌که این پارامتر پس از نُرمال‌سازی بین صفر و یک قرار می‌گیرد، مانند صحّت آن را به عدد صحیح تبدیل می‌کنیم. silhouette یک معیار ارزیابی داخلی خوشه‌بندی در بازهٔ $[-1,1]$ است [52] و همان‌طور که پیش‌تر توضیح داده شد به عدد صحیح تبدیل می‌شود. در وزن‌دهی به‌وسیلهٔ $w_1$، $w_5$ و $w_6$ وزن تمام ابریال‌ها یا خوشه‌های یک خوشه‌بندی یکسان است؛ زیرا صحّت و معیار silhouette برای کل خوشه‌بندی محاسبه می‌شود و برای خوشه‌های داخل خوشه‌بندی متفاوت نیست.

### ۱-۳-۴- وزن‌دارکردن CSPA و HGPA

با اندک تغییری در CSPA، امکان ترکیب وزن‌دار خوشه‌بندی‌ها را به این الگوریتم خوشه‌بندی ترکیبی اضافه کرده‌ایم. اگر $r$ خوشه‌بندی داشته باشیم، برای ساخت ماتریس شباهت، ابتدا برای هر زوج نمونهٔ $(x_i, x_j)$ امتیاز هر خوشه‌بندی $\lambda^l$ با وزن $w_l$ $(0 \le w_l \le 1)$ را مطابق (۷) محاسبه می‌کنیم.

$$\text{score}(\lambda^t) = \begin{cases} w_l & \text{if label}(x_i) = \text{label}(x_j) \\ 1 - w_l & \text{otherwise.} \end{cases} \quad (۷)$$

به عبارت دیگر اگر خوشه‌بندی با وزن $w$ دو نمونه را در یک خوشه قرار نداده است، پس با وزن $1 - w$ به هم‌خوشه‌بودن آنها رأی می‌دهد. در این صورت درایهٔ $i$ و $j$ از ماتریس شباهت از معادلهٔ (۸) به‌دست می‌آید.

$$S(i,j) = \sum_{l=1}^{r} \frac{score(\lambda^l)}{r} \quad (۸)$$

برای وزن‌دارکردن HGPA کافی است به جای یکسان درنظرگرفتن وزن ابریال‌ها، وزن‌های محاسبه‌شده برای هر خوشه‌بندی (طبق جدول ۲) را به‌عنوان وزن ابریال‌های معادل آن در ابرگراف در نظر بگیریم.

## ۵- نتایج تجربی

در این بخش به بیان نتایج حاصل از آزمایش روش پیشنهادی می‌پردازیم. به‌منظور پیاده‌سازی این روش از زبان برنامه‌نویسی Python استفاده شده است. از جمله کتابخانه‌های مورد استفاده در این پژوهش می‌توان به Scipy[1] و Numpy[2] اشاره کرد که برای انواع محاسبات عددی و مهندسی ارائه شده‌اند. همچنین ماژول یادگیری ماشین Scikitlearn[3] ابزارهای ساده و کارایی جهت تحلیل داده و داده‌کاوی فراهم کرده است. در ادامه ابتدا مجموعه داده‌های مورد بررسی و معیارهای ارزیابی مورد استفاده بیان می‌شوند و سپس به بیان نتایج می‌پردازیم.

### ۱-۵- مجموعه‌داده‌های مورد آزمایش

کارایی الگوریتم پیشنهادی در مقایسه با الگوریتم‌های دیگر در چندین مجموعه‌دادهٔ واقعی و دست‌ساز بررسی می‌شود. مجموعه‌دادهٔ دست‌ساز داده‌هایی هستند که توسط استرل و گاش تهیه شده‌اند و از طریق (http://strehl.com) قابل دسترسی است. مجموعه داده‌های واقعی مورد استفاده، تعدادی از مجموعه داده‌های استاندارد UCI[4] و دو مجموعه داده از scikitlearn[5] هستند. بسیاری از مطالعات اخیر با استفاده از این مجموعه داده‌ها گزارش شده‌اند. مشخصات داده‌های مورد آزمایش در جدول (۳) ارائه شده است.

[1] http://www.scipy.org
[2] http://www.numpy.org
[3] http://www.scikit-learn.org
[4] https://archive.ics.uci.edu/ml/datasets.html
[5] http://scikit-learn.org/stable/datasets/\#toy-datasets




[ DOI: 10.29252/jsdp.17.2.100 ]

## ۵-۲- معیارهای ارزیابی نتایج

برای ارزیابی روش پیشنهادی از ۹ معیار استفاده شده است. جدول (۴) خلاصه‌ای از معیارهای ارزیابی استفاده‌شده را نشان می‌دهد. مستند جزئیات این معیارها در پایتون از طریق پایگاه scikitlearn[1] قابل دسترسی است.

**(جدول-۳): مجموعه داده‌های مورد آزمایش**
**(Table-3): Datasets used in the experiments**

| نام مجموعه داده | تعداد خوشه‌ها | تعداد ویژگی‌ها (ابعاد) | تعداد نمونه‌ها |
|---|---|---|---|
| (دست‌ساز) 8d5k | 5 | 8 | 1000 |
| Iris (scikitlearn) | 3 | 4 | 150 |
| Digits (scikitlearn) | 10 | 64 | 1797 |
| Seeds (UCI) | 3 | 6 | 210 |
| Data User Modeling Int (UCI) | 4 | 4 | 258 |
| Wholesale customers data (UCI) | 3 | 6 | 440 |
| Movement libras (UCI) | 15 | 90 | 360 |
| Zoo (UCI) | 7 | 16 | 110 |
| Wine (UCI) | 3 | 13 | 178 |
| Ionosphere (UCI) | 2 | 34 | 351 |
| Ecoli (UCI) | 8 | 7 | 336 |
| Image segmentation (UCI) | 7 | 19 | 210 |
| Glass (UCI) | 7 | 10 | 214 |

**(جدول-۴): معیارهای ارزیابی روش پیشنهادی**
**(Table-4): Evaluation measures**

| معیار | | بازه | متقارن | نُرمال شده در برابر برچسب‌های تصادفی |
|---|---|---|---|---|
| adjusted rand index (ARI) | | [-۱,۱] | ✓ | ✓ |
| Fowlkes-Mallows (FMI) | | [۰,۱] | × | ✓ |
| مبتنی بر اطلاعات متقابل | اطلاعات متقابل (MI) | × | ✓ | ✓ |
| | اطلاعات متقابل تطبیق داده شده (AMI) | Max=1 | ✓ | × |
| | اطلاعات متقابل نُرمال شده (NMI) | [۰,۱] | ✓ | × |
| همگونی (Homogeneity) | | [۰,۱] | × | × |
| جامعیّت (Completeness) | | [۰,۱] | × | × |
| V-measure | | [۰,۱] | ✓ | × |
| خلوص (Purity)[2] | | [۰,۱] | × | × |

## ۵-۳- نتایج آزمایش‌ها

در این بخش نتایج هر یک از روش‌های تولید خوشه‌بندی‌های پایه و ترکیب وزن‌دار آنها با استفاده از CSPA و HGPA وزن‌دار گزارش شده است. برای اینکه نتایج بیشتر قابل اعتماد باشد، میانگینی از ده بار اجرای مجزای فرآیند تولید خوشه‌های پایه و سپس ترکیب وزن‌دار ارائه شده است. در نمودارهای این بخش خط افقی معیارهای ارزیابی را نشان می‌دهد و چهار میلهٔ رنگی روش‌های تولید خوشه‌های پایه را مشخص کرده است که میلهٔ آبی‌رنگ استفاده از الگوریتم‌های مختلف، میلهٔ قرمز استفاده از معیارهای فاصلهٔ متفاوت در اجرای k-means، میلهٔ خاکستری خوشه‌بندی ویژگی‌های توزیع‌شده و میلهٔ زرد اجرای k-means با $k$های متفاوت است.

هر میله در نمودار، میانگین تفاضل CSPA (یا HGPA) بدون وزن از CSPA (یا HGPA) وزن‌دار برای هر یک از معیارهای ارزیابی گزارش شده است. بدین‌ترتیب اعداد مثبت نمایان‌گر بهبود عملکرد CSPA (یا HGPA) با استفاده از وزن‌دهی است. به‌دلیل حجم بالای آزمایش‌ها، در این بخش خلاصه‌ای از نتایج به‌صورت میانگین نتایج تمام مجموعه داده‌ها گزارش شده است. گزارش آزمایش تمام مجموعه داده‌ها از طریق فایل موجود در وب‌سایت آزمایشگاه داده‌کاوی و یادگیری ماشین دانشگاه قم[3] در دسترس است.

در نمودار (۱)، نتایج برای تولید خوشه‌بندی‌های پایه بر مبنای ویژگی‌های مختلف با استفاده از الگوریتم CSPA نمایش داده شده است. همان‌طور که مشاهده می‌شود، در این آزمایش بهبود قابل توجهی مشاهده نمی‌شود؛ به‌طوری‌که بهترین بهبود ایجادشده کمتر از ۰.۰۱ است. به‌علاوه، استفاده از ویژگی‌های توزیع‌شده برای تولید خوشه‌بندی‌های پایه باعث کاهش دقت در روش CSPA وزن‌دار شده است.

نمودارهای (*نمودار* تا(نمودار تفاضل الگوریتم HGPA بدون وزن از HGPA وزن‌دار را با وزن‌دهی توسط $w_1$ تا$w_7$ (طبق (*جدول*) زمانی که accuracy برابر صحّت‌های برآورد شده توسط AD است نشان می‌دهند.

برای مقایسهٔ روش‌های تولید خوشه‌بندی‌های پایه توجه کنید که عملکرد روش تولید به روش وزن‌دهی وابسته است. به‌عنوان مثال، روش استفاده از تعداد متفاوت خوشه‌ها (میلهٔ زرد) هرچند برای روش وزن‌دهی $w_1$ و $w_2$ و $w_7$

[1] http://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation
[2] http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html
[3] http://dml.qom.ac.ir/wp-content/uploads/2017/05/Results_Vahidi.pdf

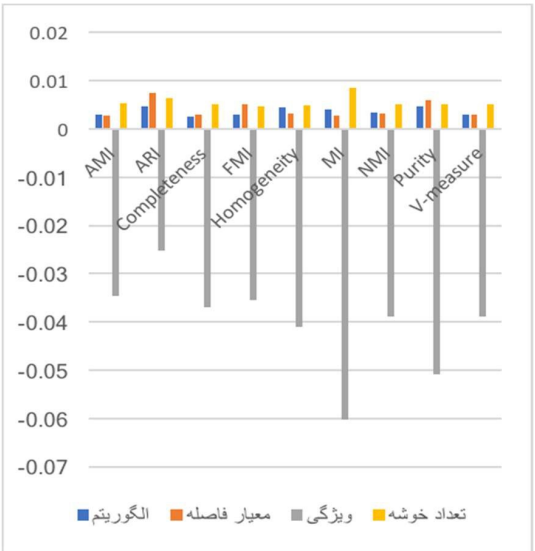



[ DOI: 10.29252/jsdp.17.2.100 ]

بهترین نتایج را حاصل کرده است، برای روش‌های $w_3$ و $w_4$ ضعیف‌تر از بقیهٔ روش‌ها عمل کرده است. به همین صورت، روش استفاده از الگوریتم‌های مختلف (میلهٔ آبی) برای روش وزن‌دهی $w_3$ و $w_6$ عملکرد بسیار خوبی داشته است، درحالی‌که در روش $w_5$ باعث کاهش دقت روش تجمیعی شده است. روش استفاده از معیارهای فاصلهٔ متفاوت برای تولید خوشه‌بندی‌های پایه نیز در روش وزن‌دهی $w_5$ بهترین نتایج را به‌دست آورده است؛ درنهایت روش ویژگی‌های توزیع شده که در بسیاری از روش‌های وزن‌دهی عملکرد قابل قبولی نشان داده است، در روش $w_5$ ضعیف‌تر از بقیهٔ روش‌ها عمل کرده و باعث کاهش دقت الگوریتم ترکیب شده است.

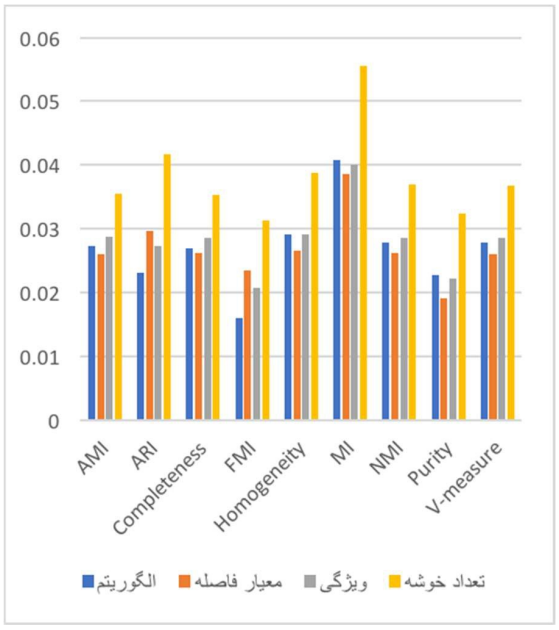
مقایسهٔ روش‌های وزن‌دهی مختلف نشان می‌دهد که در مجموع استفاده از $w_1$، $w_2$، $w_6$ و $w_7$ نتایج را بهبود داده است و به‌طور متوسط بهترین نتایج مربوط به زمانی است که وزن‌دهی به وسیله $w_1$ یعنی تنها با استفاده از صحّت‌های برآوردی AD است. از مقایسهٔ نمودار (۳) و نمودار (۴) این نتیجه حاصل می‌شود که نرمال‌سازی کارایی را کاهش داده است. نمودار (۶) نیز نشان می‌دهد استفاده از silhouette نیز کارایی پایینی دارد.

در نمودار (۵) ضعیف‌ترین عملکرد مشاهده می‌شود ($w_4$) که وزن‌دهی تحت تأثیر silhouette و نرمال‌سازی است و AD در آن تأثیری ندارد. درحالی‌که در دیگر وزن‌دهی‌ها که AD حضور داشته است، دست‌کم برای دو مورد از چهار روش تولید خوشه‌بندی‌های پایه بهبود مشاهده می‌شود که نشان‌دهندهٔ مؤثربودن روش ارائه‌شده در این مقاله است.



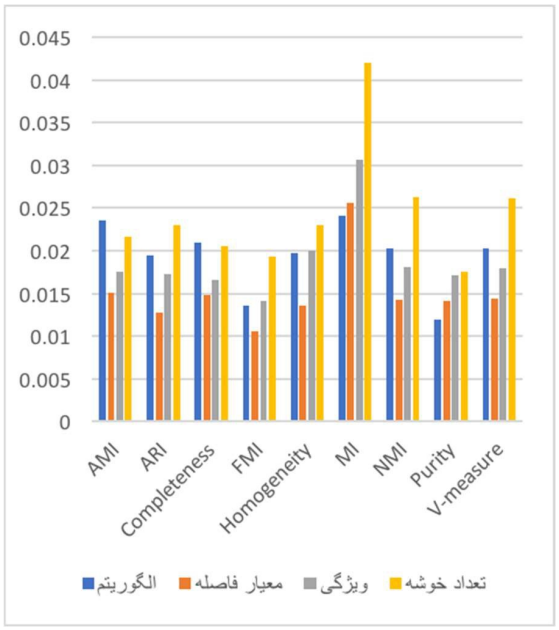

(نمودار-۱): تفاضل الگوریتم CSPA بدون وزن از CSPA وزن‌دار با وزن‌دهی به وسیلهٔ صحّت‌های برآوردشده توسط AD

(Chart-1): Subtraction of unweighted CSPA from weighted CSPA using accuracies estimated by AD method




(نمودار-۲): تفاضل الگوریتم HGPA بدون وزن از HGPA وزن‌دار با وزن‌دهی توسط $w_1$ و صحّت‌های برآوردشده توسط AD

(Chart-2): Subtraction of HGPA from weighted HGPA by using $w_1$ and AD estimated



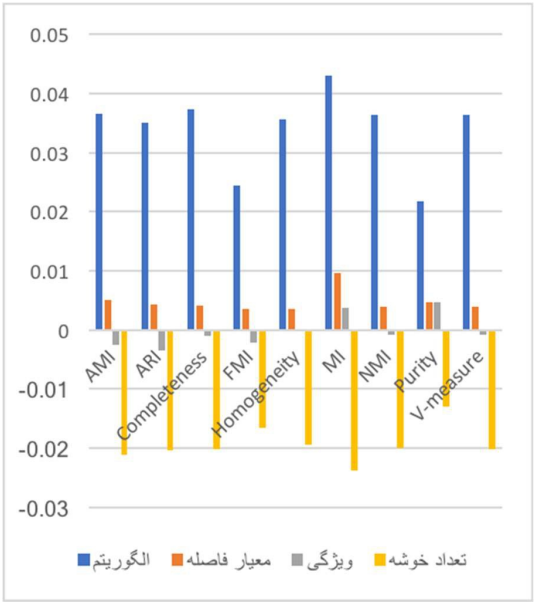

(نمودار-۳): تفاضل الگوریتم HGPA بدون وزن از HGPA وزن‌دار با وزن‌دهی توسط $w_2$ و صحّت‌های برآوردشده توسط AD

(Chart-3): Subtraction of HGPA from weighted HGPA by using $w_2$ and AD estimated accuracies



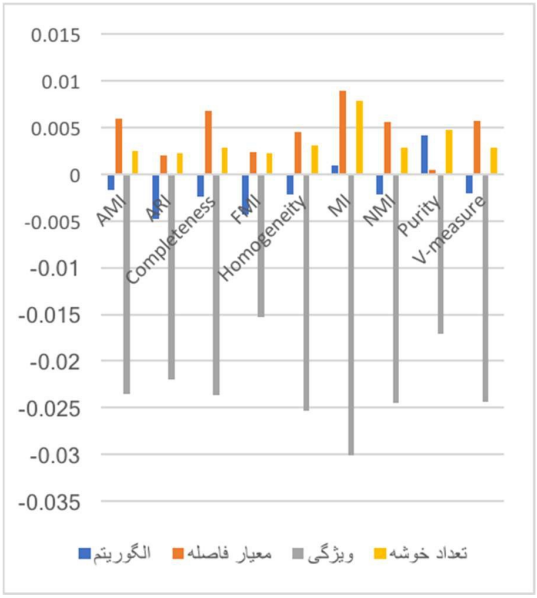


[ DOI: 10.29252/jsdp.17.2.100 ]



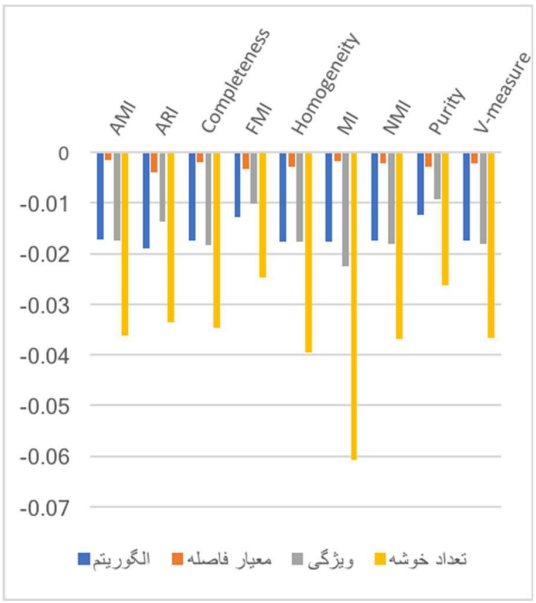

(نمودار-۴): تفاضل الگوریتم HGPA بدون وزن از HGPA وزن‌دار
با وزن‌دهی توسط $w_3$ و صحّت‌های برآوردشده توسط AD

(Chart-4): Subtraction of HGPA from weighted HGPA by using $w_3$ and AD estimated accuracies



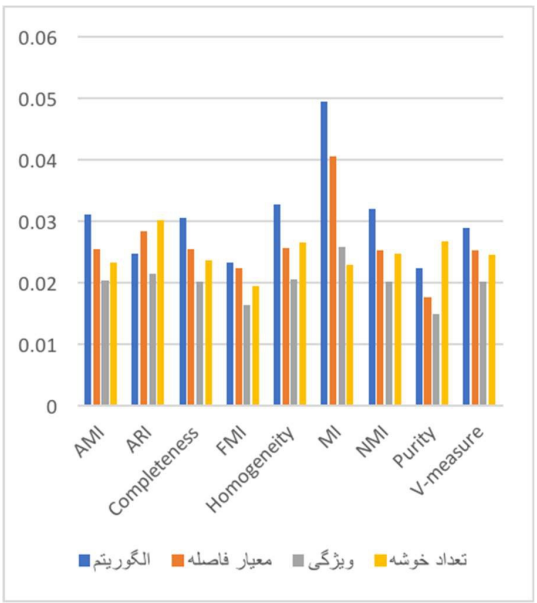

(نمودار-۶): تفاضل الگوریتم HGPA بدون وزن از HGPA وزن‌دار
با وزن‌دهی توسط $w_5$ و صحّت‌های برآوردشده توسط AD

(Chart-6): Subtraction of HGPA from weighted HGPA by using $w_5$ and AD estimated accuracies




(نمودار-۵): تفاضل الگوریتم HGPA بدون وزن از HGPA وزن‌دار
با وزن‌دهی توسط $w_4$ و صحّت‌های برآوردشده توسط AD

(Chart-5): Subtraction of HGPA from weighted HGPA by using $w_4$ and AD estimated accuracies

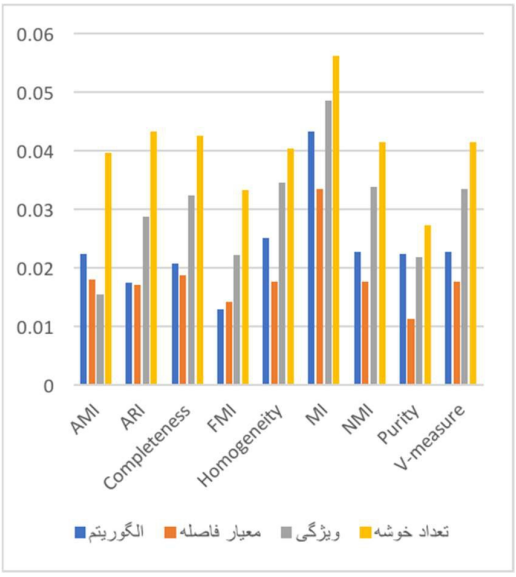



(نمودار-۷): تفاضل الگوریتم HGPA بدون وزن از HGPA وزن‌دار
با وزن‌دهی توسط $w_6$ و صحّت‌های برآوردشده توسط AD

(Chart-7): Subtraction of HGPA from weighted HGPA by using $w_6$ and AD estimated accuracies



[ DOI: 10.29252/jsdp.17.2.100 ]




**(نمودار-۸): تفاضل الگوریتم HGPA بدون وزن از HGPA وزن‌دار با وزن‌دهی توسط $w_7$ و صحّت‌های برآوردشده توسط AD**

**(Chart-8): Subtraction of HGPA from weighted HGPA by using $w_7$ and AD estimated accuracies**

## ۶- نتیجه‌گیری و کارهای آینده

در این مقاله روشی برای ترکیب خوشه‌بندی‌های پایه با وزن‌دهی مبتنی بر الگوریتم AD پیشنهاد شد و براساس نُه معیار ارزیابی مختلف روی سیزده مجموعه‌دادهٔ متفاوت نشان داده شد که روش ترکیب وزن‌دار پیشنهادی بهتر از روش‌های ترکیب بدون وزن قادر به خوشه‌بندی داده‌ها است. از نتایج به‌دست‌آمده مشاهده می‌شود که استفاده از AD در برآورد صحّت خوشه‌بندی‌ها مناسب‌تر است. از میان روش‌های وزن‌دهی مختلفی که برای ابریال‌ها پیشنهاد شد، نتایج هر یک از وزن‌دهی‌ها بر روی مجموعه دادهٔ خاصی توانستند بهبود قابل توجهی ایجاد کنند که حاکی از وابستگی نوع وزن‌دهی به ویژگی‌های ساختاری هر مجموعه داده است؛ اما به‌طور متوسط بهترین نتایج مربوط به زمانی است که وزن‌دهی به وسیله $w_1$ یعنی تنها با استفاده از صحّت‌های برآوردی AD است. ضعیف‌ترین عملکرد را زمانی مشاهده می‌کنیم که وزن‌دهی تحت تأثیر silhouette و نرمال‌سازی می‌باشد و AD در آن تأثیری ندارد.

در این پژوهش HGPA وزن‌دار با روش‌های متفاوت تولید خوشه‌های پایه مورد بررسی قرار گرفت. نتایج آزمایش HGPA با هفت نوع وزن‌دهی متفاوت با صحّت‌های AD نشان داد که در بیش‌تر موارد بیشترین بهبود برای تولید خوشه‌های پایه با استفاده از الگوریتم‌های متفاوت رخ داده است.

با توجه به بهبود صحّت خوشه‌بندی نهایی با وزن‌دهی توسط صحّت‌های برآوردی در شرایط خاصی، برای پژوهش‌های آینده می‌توان به ساخت ماژولی جهت ابریادگیری پرداخت. به‌طوری‌که به‌وسیلهٔ ابرویژگی‌های[1] استخراج‌شده از داده‌ها و خوشه‌بندی‌های پایه، میزان بهبود خوشه‌بندی نهایی را پیش‌بینی کند. برای این منظور باید روی مجموعه داده‌های بیشتری آزمایش صورت بگیرد تا ابرویژگی‌های مؤثر متفاوت، مانند ابعاد داده، تعداد داده، صحّت خوشه‌های پایه، نوع خوشه‌بندی و غیره شناسایی شوند.

## ۷- مراجع 7- References


[1] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Elsevier, 2011.

[2] S. Vega-Pons and J. Ruiz-Shulcloper, "a Survey of Clustering Ensemble Algorithms," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, no. 03, pp. 337–372, 2011.

[3] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

[4] S.-B. Cho and J. H. Kim, "Combining multiple neural networks by fuzzy integral for robust classification," *Syst. Man Cybern. IEEE Trans.*, vol. 25, no. 2, pp. 380–384, 1995.

[5] J. Franke and E. Mandler, "A comparison of two approaches for combining the votes of cooperating classifiers," in *Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on*, 1992, pp. 611–614.

[6] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 10, pp. 993–1001, 1990.

[7] S. Hashem and B. Schmeiser, "Improving model accuracy using optimal linear combinations of trained neural networks," *Neural Networks, IEEE Trans.*, vol. 6, no. 3, pp. 792–794, 1995.

[8] J. Kittler, "Improving recognition rates by classifier combination: A theoretical framework," *DAC and IS, editors, Progress in Handwriting Recognition*, pp. 231-248, 1997.

[9] D. J. Miller and L. Yan, "Critic-driven ensemble classification," *Signal Process. IEEE Trans.*, vol. 47, no. 10, pp. 2833–2844, 1999.


[1] meta-feature





[ DOI: 10.29252/jsdp.17.2.100 ]

[10] K. W. De Bock, K. Coussement, and D. Van den Poel, "Ensemble classification based on generalized additive models," *Comput. Stat. Data Anal.*, vol. 54, no. 6, pp. 1535–1546, 2010.

[11] C. Domeniconi and M. Al-Razgan, "Weighted cluster ensembles: Methods and analysis," *ACM Trans. Knowl. Discov. from Data*, vol. 2, no. 4, pp. 17, 2009.

[12] A. Strehl and J. Ghosh, "Cluster ensembles---a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, no. Dec, pp. 583–617, 2002.

[13] H. Amirkhani and M. Rahmati, "Agreement/disagreement based crowd labeling," *Appl. Intell.*, vol. 41, no. 1, pp. 212–222, Jul. 2014.

[14] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," in *Foundations of Computer Science, 1989., 30th Annual Symposium on*, 1989, pp. 256–261.

[15] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques." *Emerging Artificial Intelligence Applications in Computer Engineering*, vol. 160, pp. 3-24, 2007.

[16] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, 2000, pp. 1–15.

[17] T. Windeatt, "Vote counting measures for ensemble classifiers," *Pattern Recognit.*, vol. 36, no. 12, pp. 2743–2756, 2003.

[18] S. Wang, A. Mathew, Y. Chen, L. Xi, L. Ma, and J. Lee, "Empirical analysis of support vector machine ensemble classifiers," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6466–6476, 2009.

[19] A. J. C. Sharkey, *Combining artificial neural nets: ensemble and modular multi-net systems*. Springer Science & Business Media, 2012.

[20] A. L. N. Fred and A. K. Jain, "Data clustering using evidence accumulation," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2002, vol. 4, pp. 276–280.

[21] A. Topchy, A. K. Jain, and W. Punch, "A mixture model for clustering ensembles," in *Society for Industrial and Applied Mathematics. Proceedings of the SIAM International Conference on Data Mining*, 2004, pp. 379.

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

[23] D. Gondek and T. Hofmann, "Non-redundant clustering with conditional ensembles," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 70–77.

[24] A. Topchy, A. K. Jain, and W. Punch, "Combining multiple weak clusterings," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003, pp. 331–338.

[25] L. I. Kuncheva, S. T. Hadjitodorov, and Others, "Using diversity in cluster ensembles," in *Systems, man and cybernetics, 2004 IEEE international conference on*, 2004, vol. 2, pp. 1214–1219.

[26] B. Minaei-Bidgoli, A. Topchy, and W. F. Punch, "Ensembles of partitions via data resampling," in *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, 2004, vol. 2, pp. 188–192.

[27] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 27, no. 12, pp. 1866–1881, 2005.

[۲۸] علیزاده حسین، مشکی محسن، پروین حمید و مینایی بیدگلی بهروز، "خوشه‌بندی ترکیبی مبتنی بر زیرمجموعه‌ای از خوشه‌های اولیّه"، *پردازش علائم و داده‌ها*، ۱۳۸۹، ۷ (۱): ۱۹-۳۲.

[28] H. Alizadeh, M. Moshki, H. Parvin, B. Minaei Bidgoli, "Clustering ensemble based on combination of subset of primary clusters," *Signal and Data Processing,* vol. 7, no. 1, pp. 19-32, 2010.

[29] L. Franek and X. Jiang, "Ensemble clustering by means of clustering embedding in vector spaces," *Pattern Recognit.*, vol. 47, no. 2, pp. 833–842, 2014.

[۳۰] میرزایی عبدالرضا، "ترکیب خوشه‌بندی‌های سلسله‌مراتبی با تأکید بر حفظ اطلاعات ساختاری خوشه‌بندی‌های پایه"، دانشگاه صنعتی امیر کبیر، ۱۳۸۸.

[30] A. Mirzaei, "Combining hierarchical clusterings with emphasis on retaining the structural contents of the base clusterings," *PhD dissertation*, Amirkabir University of Technology, 2009.

[31] J. H. Friedman and J. J. Meulman, "Clustering objects on subsets of attributes," *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, vol. 66, no. 4, pp. 815–849, 2004.

[32] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos, "Locally adaptive metrics for clustering high dimensional



[ DOI: 10.29252/jsdp.17.2.100 ]

data," *Data Min. Knowl. Discov.*, vol. 14, no. 1, pp. 63–97, 2007.

[33] M. Al-Razgan and C. Domeniconi, "Weighted clustering ensembles," in *Proceedings of the 2006 SIAM International Conference on Data Mining*, 2006, pp. 258–269.

[34] S. Vega-Pons, J. Correa-Morris, and J. Ruiz-Shulcloper, "Weighted cluster ensemble using a kernel consensus function," in *Iberoamerican Congress on Pattern Recognition*, 2008, pp. 195–202.

[35] S. Vega-Pons and J. Ruiz-Shulcloper, "Clustering ensemble method for heterogeneous partitions," in *Iberoamerican Congress on Pattern Recognition*, 2009, pp. 481–488.

[36] T. Li and C. Ding, "Weighted consensus clustering," in *Proceedings of the 2008 SIAM International Conference on Data Mining*, 2008, pp. 798–809.

[37] F. Gullo, A. Tagarelli, and S. Greco, "Diversity-Based Weighting Schemes for Clustering Ensembles," in *Proceedings of the 2009 SIAM International Conference on Data Mining*, 2009, pp. 437–448.

[38] J. W. C.-D. Huang Dong; Lai, "Ensemble clustering using factor graph," *Pattern Recognit.*, vol. 50, pp. 131-142, 2015.

[39] H. Liu, J. Wu, T. Liu, D. Tao, and Y. Fu, "Spectral Ensemble Clustering via Weighted K-Means: Theoretical and Practical Evidence," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1129–1143, 2017.

[40] Y. Ren, C. Domeniconi, G. Zhang, and G. Yu, "Weighted-object ensemble clustering: methods and analysis," *Knowl. Inf. Syst.*, pp. 1–29, 2016.

[41] Q. Kang, S. Liu, M. Zhou, and S. Li, "A weight-incorporated similarity-based clustering ensemble method based on swarm intelligence," *Knowledge-Based Syst.*, vol. 104, pp. 156–164, 2016.

[۴۲] پروین حمید، "خوشه بندی ترکیبی با وزن دهی توام خوشه ها و ویژگی ها"، دانشگاه علم و صنعت ایران، ۱۳۹۲.

[42] H. Parvin, " Clustering ensembles with weighting clusters and features," *PhD dissertation,* Iran University of Science & Technology, 2013.

[۴۳] لطیفی پاکدهی علیرضا، دانشپور نگین، "انتخاب اعضای ترکیب در خوشه‌بندی ترکیبی با استفاده از رأی‌گیری،" *پردازش علائم و داده‌ها*، ۱۳۹۷، ۱۵ (۴): ۱۷-۳۰.

[43] A. Litifi Pakdehi, N. Daneshpour, "Cluster ensemble selection using voting," *Signal and Data Processing*, vol. 15, no. 4, pp. 17-30, 2019.

[44] Z. Yu and H. S. Wong, "Class discovery from gene expression data based on perturbation and cluster ensemble," *IEEE Trans. Nanobioscience*, vol. 8, no. 2, pp. 147–160, 2009.

[45] H. Liu, M. Shao, S. Li, and Y. Fu, "Infinite ensemble clustering," *Data Min. Knowl. Discov.*, vol. 32, no. 2, pp. 385–416, 2018.

[46] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1460–1473, 2018.

[47] C. Zhong, L. Hu, X. Yue, T. Luo, Q. Fu, and H. Xu, "Ensemble clustering based on evidence extracted from the co-association matrix," *Pattern Recognit.*, vol. 92, pp. 93-106, 2019.

[48] F. Li, Y. Qian, J. Wang, C. Dang, and L. Jing, "Clustering ensemble based on sample's stability," *Artif. Intell.*, vol. 273, pp. 37–55, 2019.

[49] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 359–392, 1998.

[50] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel hypergraph partitioning: applications in VLSI domain," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 7, no. 1, pp. 69–79, 1999.

[51] L. Rokach and O. Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*, Springer, 2005, pp. 321–352.

[52] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

**صدیقه وحیدی فردوسی،** در سال ۱۳۹۳ مدرک کارشناسی خود را در رشته مهندسی کامپیوتر با گرایش نرم‌افزار از دانشگاه قم دریافت و مدرک کارشناسی ارشد خود را نیز از همان دانشگاه در رشته مهندسی فناوری اطلاعات با گرایش تجارت الکترونیک در سال ۱۳۹۵ اخذ کرد. موضوع پایان‌نامه کارشناسی ارشد ایشان ترکیب وزن‌دار خوشه‌بندی‌ها با هدف افزایش دقت خوشه‌بندی نهایی بوده است. زمینه‌های پژوهشی مورد علاقه وی یادگیری ماشین و داده‌کاوی است.

نشانی رایانامه ایشان عبارت است از:

**s.vahidi@stu.qom.ac.ir**





[ DOI: 10.29252/jsdp.17.2.100 ]

**حسین امیرخانی**، مدرک کارشناسی خود را در رشته مهندسی کامپیوتر با گرایش نرم‌افزار در سال ۱۳۸۶ از دانشگاه اصفهان دریافت کرده است. ایشان در سال ۱۳۸۸ مدرک کارشناسی ارشد خود را در رشته مهندسی کامپیوتر با گرایش هوش مصنوعی از دانشگاه صنعتی امیرکبیر و در سال ۱۳۹۳ نیز مدرک دکترای خود را از همان دانشگاه و در همان رشته کسب کرده است. وی هم‌اکنون استادیار گروه مهندسی کامپیوتر و فناوری اطلاعات دانشگاه قم و زمینه‌های پژوهشی مورد علاقه ایشان یادگیری ماشین و پردازش زبان‌های طبیعی است.

نشانی رایانامه ایشان عبارت است از:

**amirkhani@qom.ac.ir**



[ DOI: 10.29252/jsdp.17.2.100 ]