خوشه‌بندی گروهی طیفی لاپلاسی–P نیمه‌نظارتی برای داده‌های با ابعاد بالا

صدیقه صفری، فاطمه افسری*

بخش مهندسی کامپیوتر، دانشکدهٔ فنی و مهندسی، دانشگاه شهید باهنر کرمان، کرمان، ایران

چکیده

با توجه به افزایش روزافزون اطّلاعات و تحلیل دقیق آنها مسألهٔ خوشه‌بندی که برای آشکارسازی الگوهای پنهان موجود در داده‌ها استفاده می‌شود، همچنان از اهمیّت بالایی برخوردار است؛ و از طرفی دیگر، خوشه‌بندی داده‌های با ابعاد بالا با استفاده از روش‌های سنّتی پیشین محدودیّت‌های زیادی دارد. در مقالهٔ حاضر، یک روش خوشه‌بندی گروهی نیمه‌نظارتی برای مجموعه‌ای از داده‌های پزشکی با ابعاد بالا پیشنهاد می‌شود. در فرموله‌سازی مسألهٔ خوشه‌بندی اطّلاعات نظارتی اندکی به‌عنوان دانش پیشین با استفاده از اطّلاعات مربوط به تشابه و یا عدم تشابه (به‌صورت تعدادی زوج محدودیّت‌های دوبه‌دو) در نظر گرفته می‌شود. نخست، با استفاده از خاصیّت تراگذری، زوج محدودیّت‌های دوبه‌دو را بر روی تمام داده‌ها تعمیم می‌دهیم. سپس، با تقسیم فضای ویژگی به‌صورت تصادفی به چندین زیرفضای نابرابر ابعاد داده‌ها کاهش داده می‌شوند. خوشه‌بندی طیفی نیمه‌نظارتی مبتنی بر گراف لاپلاسی–p در هر زیرفضا به‌طور مستقل انجام می‌شود. سپس، با استفاده از نتایج هرکدام یک ماتریس مجاورت، حاصل از تجمیع نتایج هرکدام (مبتنی بر یادگیری گروهی) ایجاد می‌شود. در نهایت، با استفاده از چند عملگر جستجو روی زیرفضاها، بهترین زیرفضا، یعنی زیرفضایی را که بهترین نتیجهٔ خوشه‌بندی دارد، می‌یابیم. نتایج آزمایش‌های متعدّد بر روی چندین دادهٔ پزشکی با ابعاد بالا نشان می‌دهد که رویکرد پیشنهادی، عملکرد و کارآیی بهتری نسبت به روش‌های پیشین دارد.

واژگان کلیدی: خوشه‌بندی، یادگیری زیرفضا، یادگیری گروهی، یادگیری نیمه‌نظارتی، زوج محدودیّت‌های دوبه‌دو

Semi-Supervised Ensemble p-Laplacian Spectral Clustering for High Dimensional Data

Sedighe Safari, Fatemeh Afsari*

Department of Computer Engineering, Faculty of Engineering,
Shahid Bahonar University of Kerman, Kerman, Iran.

Abstract

Due to information increasing and the detailed analysis of them, the clustering problems that detect the hidden patterns lie in the data, are still of a great importance. On the other hand, clustering of high-dimensional data using previous traditional methods has many limitations. In this study, a semi-supervised ensemble clustering method is proposed for a set of high-dimensional medical data. In the proposed method of this study, little information is available as prior knowledge using the information on similarity or dissimilarity (as a number of pairwise constraints). Initially, using the transitive property, we generalize the pairwise constraints to all data. Then, we divide the feature space into a number of sub-spaces, and to find the optimal clustering solution, the feature space is divided into an unequal number of sub-spaces randomly. A semi-supervised spectral clustering based on the p-Laplacian graph is performed at each sub-space independently. Specifically, to increase the accuracy of spectral clustering, we have used the spectral clustering method based on the p-Laplacian graph. The p-Laplacian graph is a nonlinear generalization of the Laplacian graph. The results of any clustering

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۹/۹/۲۶ • تاریخ پذیرش: ۱۴۰۱/۷/۶ • تاریخ انتشار: ۱۴۰۲/۵/۲۰ • نوع مطالعه: پژوهشی

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]

[DOI: 10.61186/jsdp.20.1.39]

solutions are compared with the pairwise constraints and according to the level of matching, a degree of confidence is assigned to each clustering solution. Based on these degrees of confidence, an ensemble adjacency matrix is formed, which is the result of considering the results of all clustering solutions for each sub-space. This ensemble adjacency matrix is used in the final spectral clustering algorithm to find the clustering solution of the whole sub-space. Since the sub-spaces are generated randomly with an unequal number of features, clustering results are strongly influenced by different initial values. Therefore, it is necessary to find the optimal sub-space set. To this end, a search algorithm is designed to find the optimal sub-space set. The search process is initialized by forming several sets (we call each set an environment) consisting of several numbers of sub-spaces. An optimal environment is the one that has the best clustering results. The search algorithm utilized three search operators to find the optimal environment. The search operators search all the environments and the consequent sub-spaces both locally and globally. These operators combine two environments and/or replace an environment with a newly generated one. Each search operator tries to find the best possible environment in the entire search space or in a local space.

We evaluate the performance of our proposed clustering schema on 20 cancer gene datasets. The normalized mutual information (NMI) criterion and the adjusted rand index (ARI) are used to evaluate the performance evaluation. We first examine the effect of a different number of pairwise constraints. As expected, with increasing the number of pairwise constraints, the efficiency of the proposed method also increases. For example, the NMI value increases from 0.6 to 0.9 on the Khan-2001 dataset, when the number of pairwise constraints increases from 20 to 100. More number of pairwise constraints means more information is available, which helps to improve the performance of the clustering algorithm. Furthermore, we examine the effect of the number of random subspaces. It is observed that increasing the number of random subspaces has a positive effect on clustering performance with respect to the NMI value. In most datasets, when the number of sub-spaces reaches 20, the performance of the proposed method does not change much and is stable. Examining the effect of sampling rate for random subspace generation shows that the proposed method has the best performance in most cancer datasets, such as Armstrong-2002-v3, and Bredel-2005 datasets, when the random subspace generation rate is 0.5, and by deviating the rate from 0.5, the level of satisfaction decreases. Then, the results of the proposed idea are compared with the results of the method proposed in the reference [22] according to ARI and we see that our proposed method has performed better in 12 data sets out of 20 data sets than the method proposed in the reference [22]. Finally, the proposed idea is compared with some metric learning approaches with respect to NMI. We have observed that the proposed method obtained the best results compared to other compared methods on 11 datasets out of 20 datasets. It also achieved the second-best result on 6 out of 20 datasets. For example, the value NMI obtained in the proposed method is 0.1042 more than the reference [22] and it is 0.1846 more than RCA and it is 0.4 more than ITML and also it is 0.468 more than DCA on the Bredel-2005 dataset.

Utilizing ensemble clustering methods besides the confidence factor improves the ability of the proposed algorithm to achieve better results. Also, utilizing the transitive operators as well as the selection of random subspaces of unequal sizes play an important role in achieving better performance for the proposed algorithm. Using the p-Laplacian spectral clustering method produces a better, more balanced, and normal volume of clusters compared to the standard spectral clustering. Another effective approach to the performance of the proposed method is to use search operators to find the best subspace, which leads to better results.

Keywords: Clustering, Subspace Learning, Ensemble Learning, Semi-supervised Learning, Pairwise Constraints

۱- مقدمه

یادگیری نیمه‌نظارتی دسته‌ای از روش‌های یادگیری ماشین است که در آن از داده‌های بدون برچسب و داده-های برچسب‌دار به‌صورت هم‌زمان برای بهبود دقّت یادگیری استفاده می‌شود [1]. یادگیری نیمه‌نظارتی میانِ یادگیری بدون نظارت (بدون هیچ‌گونه اطّلاعات آموزشی برچسب‌گذاری‌شده) و یادگیری بانظارت (با داده‌های آموزشیِ به‌طورکامل برچسب‌گذاری‌شده) قرار می‌گیرد. بسیاری از پژوهش‌گران یادگیری ماشین دریافتند که استفاده از تعداد اندکی دادهٔ برچسب‌دار در کنار داده‌های بدون برچسب، می‌توانند بهبود قابل‌توجهی در دقّت روش‌های یادگیری بدون‌نظارت داشته باشند؛ از طرفی دیگر، این روش‌ها در مقایسه با روش‌های یادگیری بانظارت، زمان و هزینهٔ کمتری جهت برچسب‌دارکردن داده‌ها صرف می‌کنند. هزینهٔ مربوط به فرایند برچسب زدن داده‌ها ممکن است تأثیر منفی زیادی بر یک الگوی آموزشی به‌طور کامل برچسب‌دار (روش یادگیری بانظارت) بگذارد که درعمل الگو بی‌استفاده شود. از طرف دیگر، فراهم کردن اطّلاعات بدون برچسب، به‌نسبت ارزان است. بنابراین، در چنین شرایطی، یادگیری نیمه‌نظارتی می‌تواند از نقطه‌نظر عملی بسیار ارزشمند باشد. یکی از رویکردهای یادگیری نیمه‌نظارتی، رویکرد مبتنی بر گراف و شامل مراحل کلّی زیر است:

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

۱- پیش‌پردازش داده‌ها؛ شامل استخراج ویژگی‌ها، کاهش بعد، حذف نوفه و سایر موارد.
۲- تشکیل گراف همسایگی وزن‌دار مناسب با استفاده از داده‌ها که به‌طور معمول، مقدار وزن از طریق محاسبهٔ فاصلهٔ بین دو داده انجام می‌شود.
۳- تخمین برچسب داده‌های بدون برچسب با استفاده از یکی از روش‌های رایج برچسب‌گذاری.

خوشه‌بندی به‌عنوان یک روش یادگیری بدون‌نظارت، یکی از شاخه‌های رایج در حوزهٔ داده‌کاوی است که موردتوجّه و بررسی پژوهشگران بسیاری قرار گرفته است. از دههٔ ۱۹۵۰ میلادی به بعد مطالعات بسیاری در مورد خوشه‌بندی صورت گرفته است [7-2]. در دو دههٔ اخیر بحث خوشه‌بندی برای تشخیص خوشه‌هایی با ساختارهای متنوع مطرح شده و روش‌های خوشه‌بندی به‌نسبت زیادی نیز معرفی شده است که هرکدام نقاط قوّت و ضعف ویژهٔ خود را دارند. اغلب از خوشه‌بندی برای تقسیم داده‌های یک مجموعه برمبنای شباهت‌هایی که با هم دارند و همچنین، تحلیل آنها در محیط‌های مختلف دانشگاهی، صنعتی و... استفاده می‌شود. خوشه‌بندی اغلب برای گروه‌بندی داده‌ها بر اساس یک معیار مجاورت، استفاده می‌شود. به عبارتی، خوشه‌بندی به معنی یافتن گروهی از اشیا است، به‌طوری‌که اشیای مشابه در یک گروه و اشیای متفاوت در گروه‌های مختلف قرار بگیرند. ازآنجاکه در خوشه‌بندی از اطّلاعات برچسب داده‌ها استفاده نمی‌شود، خوشه‌بندی یکی از روش‌های یادگیری بدون‌نظارت محسوب می‌شود. در مقابل آن، روش یادگیری بانظارت است که نیازمند اطّلاعات برچسب داده-ها است. امّا یادگیری نیمه‌نظارتی یک روش بسیار کاربردی است که بین این دو رویکرد قرار می‌گیرد. در خوشه‌بندی نیمه‌نظارتی اطّلاعات اضافه‌ای به‌صورت تعداد کمی دادهٔ برچسب‌خورده، یا تعدادی زوج محدودیّت‌های دوبه‌دو به‌صورت محدودیّت‌های باید-متّصل و نباید-متّصل داده شده است. نتیجهٔ خوشه‌بندی نیمه‌نظارتی متأثّر از محدودیّت‌ها تغییر می‌کند. همچنین، خوشه‌بندی نیمه‌نظارتی مبتنی بر شباهت است که اندازه‌گیری شباهت بر اساس محدودیّت‌ها تغییر می‌کند. بسیاری از رویکردهای خوشه‌بندی نیمه‌نظارتی جدید، گسترش الگوریتم‌های خوشه‌بندی سنّتی با توجه به اطّلاعات برچسب یا زوج محدودیّت‌ها هستند [15-8].

کاهش ابعاد داده‌ها یکی دیگر از مسائل رایج در حوزهٔ آمار، یادگیری ماشین و نظریّهٔ اطّلاعات است. کاهش ابعاد داده، فرایند کاهش تعداد متغیّرهای تصادفی موردبررسی و استخراج مجموعه‌ای از متغیّرهای اصلی است. اغلب، تجزیه و تحلیل داده‌ها مانند رگرسیون یا طبقه‌بندی در فضای کاهش‌یافته دقیق‌تر از فضای اصلی انجام می‌شود. به‌طور معمول، برای مجموعه‌داده‌های با ابعاد بزرگ (به‌عنوان مثال با تعداد ابعاد بیش از ۱۰) کاهش ابعاد به منظور جلوگیری از مسألهٔ نفرین بُعد[1]، پیش‌از به‌کارگیری الگوریتم‌های رگرسیون یا طبقه‌بندهایی همچون طبقه‌بند K-نزدیک‌ترین همسایگان (K-NN) انجام می‌شود.

۲- پیشینهٔ پژوهش

به‌تازگی، گروه‌های زیادی به یادگیری گروهی در مسائل خوشه‌بندی نیمه‌نظارتی کرده‌اند [23-19]. یادگیری گروهی یک الگووارهٔ (پارادایم) یادگیری ماشین است که در آن چندین فراگیر برای حل یک مسأله آموزش داده می‌شوند. روش‌های یادگیری ماشین متداول سعی می‌کنند یک فرضیّه از داده‌های آموزشی را یاد بگیرند، امّا روش‌های گروهی سعی می‌کنند این روش‌ها را بهبود دهند. باتوجّه‌به قابلیّت روش‌های یادگیری گروهی، همچنین، لزوم کاهش ابعاد برای خوشه‌بندی داده‌های با ابعاد بزرگ، در این پژوهش از رویکرد یادگیری گروهی خوشه‌بندی نیمه‌نظارتی مبتنی بر زیرفضاهای تصادفی[2] (RSEMICE) استفاده می‌شود. نخست، فضای ویژگی را به تعدادی زیرفضای نابرابر که در هرکدام تعدادی از ویژگی‌ها به‌طور تصادفی انتخاب شده‌اند، تقسیم می‌کنیم. سپس، در هر زیرفضا، با یک روش خوشه‌بندی طیفی[3] مبتنی بر گراف لاپلاسی-p، داده‌ها به‌طور مستقل خوشه‌بندی می‌شوند. نتایج هرکدام از خوشه‌بندها با زوج محدودیّت‌های دوبه‌دو تطبیق داده می‌شوند و بر اساس میزان تطابق یک درجهٔ اطمینان به هر خوشه‌بند نسبت داده می‌شود؛ سپس، بر اساس این درجات اطمینان یک ماتریس مجاورت که حاصل اجماع تمامی الگوریتم‌های خوشه‌بندی است، تشکیل می‌شود. ازآنجاکه تشکیل زیرفضاها به‌صورت تصادفی صورت گرفته است، در ادامه

[1] Curse of dimension
[2] Random Subspace Based Semi-Supervised Cluster Ensemble
[3] Spectral clustering

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

یک الگوریتم جستجو برای یافتن بهترین زیرفضا طراحی شده است. الگوریتم جستجو بدین‌گونه عمل می‌کند که در بین تعدادی زیرمجموعه شامل زیرفضاها، که از این به بعد هر زیرمجموعه را یک محیط می‌نامیم، به‌دنبال بهترین محیط می‌گردد. بهترین محیط، محیطی است که بهترین نتایج خوشه‌بندی را داشته باشد. برای پیاده‌سازی الگوریتم جستجو از تعدادی عملگر جستجو بهره گرفته‌ایم که هرکدام به گونه‌ای در کل فضای جستجو، یا در یک فضای محلی سعی در یافتن بهترین محیط ممکن دارد.

در ادامهٔ مقاله، در بخش دوم پژوهش‌های پیشین انجام‌شده را مرور می‌کنیم. در بخش سوم، روش پیشنهادی برای خوشه‌بندی نیمه‌نظارتی گروهی برای داده‌های با ابعاد بالا به‌تفصیل معرفی می‌شود. در بخش چهارم، نتایج شبیه‌سازی روش پیشنهادی مورد بررسی و تحلیل قرار می‌گیرد و در نهایت، نتیجهٔ به‌دست‌آمده جمع‌بندی می‌شود.

رشد روزافزون داده‌ها در سال‌های اخیر، مسایل و نیازهای متفاوت و متعددی را در زمینهٔ داده‌کاوی و یادگیری ماشین ایجاد کرده است. یکی از مهم‌ترین این مسایل انتخاب ویژگی‌های مرتبط به مجموعهٔ اصلی ویژگی‌های موجود است که عملکرد یادگیری را نسبت‌به عملکرد مجموعهٔ اصلی ویژگی‌ها به میزان بیشتری بهبود می‌بخشد. خوشه‌بندی گراف، زیر مجموعه‌ای از تجزیه و تحلیل خوشه‌ای است که به‌دنبال گروه‌هایی از رأس‌های مرتبط در یک گراف است. به‌دلیل کاربرد گستردهٔ آن چندین الگوریتم خوشه‌بندی گراف در سال‌های گذشته ارائه شده است. یکی از معروف‌ترین الگوریتم‌های خوشه‌بندی گراف، الگوریتم‌های خوشه‌بندی طیفی هستند که بیشتر بر اساس تجزیهٔ ویژهٔ ماتریس‌های لاپلاسی گراف‌های وزن‌دار یا گراف‌های بی‌وزن طراحی شده‌اند. در سال‌های اخیر مطالعات مختلفی در زمینه خوشه‌بندی انجام شده است که در ادامه، تعدادی از آنها به‌طور مختصر آورده شده است. چریسلی و همکارش [1] به‌منظور بهینه-سازی ساختار یک گراف و دستیابی به نتایج خوشه‌بندی بهتر یک روش خوشه‌بندی گراف طیفی را بررسی کردند که از الگوریتم ژنتیک استفاده کرده بود. ازآنجاکه عملکرد الگوریتم آنها تا حدّ زیادی به نحوهٔ ایجاد جمعیّت اولیه بستگی داشت، بنابراین، نتایج رضایت‌بخش نبودند. در الگوی دیگر هو و همکاران [2] یک رویکرد خوشه‌بندی جدید به نام خوشه‌بندی طیفی غیرمنفی انعطاف‌پذیر[1](SNSC) معرّفی کردند. در واقع SNSC، ویژگی غیرمنفی اصلی یک ماتریس شاخص را حفظ می‌کند، که با یک راه‌حلّ دقیق به یک مسئلهٔ بهینه‌سازی با انعطاف‌پذیری بیشتری منجر شده است. همچنین، هنسر و همکاران [3] در کار پژوهشی خود به بررسی جامع رویکردهای انتخاب ویژگی‌ها برای خوشه‌بندی باتوجّه‌به حذف مزایا و ضررهای رویکردهای فعلی، از دیدگاه‌های متفاوت پرداختند. چائو و همکاران [4] راهبردهای مشترک برای ترکیب چندین دیدگاه از داده‌ها را مطالعه کردند و طبقه‌بندی جدیدی را از رویکردهای خوشه‌بندی چندنمایه[2](MVC) پیشنهاد دادند. به‌علاوه هی و همکارانش [5] یک روش اندازه‌گیری شباهت هسته گاوسی تطبیقی را از طریق کمیّت میزان اهمیّت برای هر رأس از گراف شباهت و هسته گاوسی مقیاس‌بندی‌شده، ارائه کردند. سپس، آنها الگوریتمی برای خوشه‌بندی طیفی تطبیقی بر اساس اهمیّت نزدیک‌ترین همسایگان را پیشنهاد دادند. جیا و همکارانش [6] یک الگوریتم خوشه‌بندی طیفی ویژه را بر اساس آنتروپی دانش ارائه دادند. قابل ذکر است که به مفهوم آنتروپی در مجموعه ویژگی‌های بزرگ، برای ارزیابی اهمیّت هر ویژگی در الگوریتم آنها توجّه شده است. یو و همکاران [7] در پژوهش خود یک چارچوب گروهی برای استخراج ساختار خوشه‌ها از داده با رویکرد احتمالاتی را (با عنوان GMMSE) طرّاحی کردند که ساختار خوشه را از مجموعهٔ داده‌ها شناسایی می‌کند. نتایج تجربی مطالعهٔ آنها نشان داد که GMMSE روی مجموعه‌داده‌های ترکیبی و مجموعه‌داده‌های واقعی مخزن UCI کارکرد مناسبی دارد.

یادگیری نیمه‌نظارتی، شاخه‌ای از یادگیری ماشین است که مانند یادگیری بانظارت فقط از داده‌های برچسب‌خورده استفاده نمی‌کند، بلکه در مرحلهٔ آموزش از داده‌های بدون برچسب نیز برای انجام وظایف یادگیری مشخص بهره می‌برد. همچنین، یکی از رویکردهای یادگیری نیمه‌نظارتی استفاده از اطّلاعات بسیار ناچیز در مورد ردهٔ داده‌هاست. در این رویکرد، هیچگونه دادهٔ برچسب‌داری مانند رویکرد معمول یادگیری نیمه‌نظارتی، در دسترس نیست. امّا اطّلاعات ناچیزی همچون یکسانی ردهٔ دو داده (معروف به زوج دادهٔ باید-متّصل[3]) و یا تفاوت

[1] scalable nonnegative spectral clustering
[2] multi-view clustering
[3] Must-link

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

ردهٔ دو داده (معروف به زوج دادهٔ نباید-متّصل[1]) در دسترس است. این اطّلاعات، معروف به زوج محدودیّت-های دوبه‌دو[2] هستند که به‌عنوان یک قید سخت[3] در تابع هدف مسألهٔ یادگیری ماشین در نظر گرفته می‌شوند و باید ارضا شوند.

در سال‌های اخیر پژوهش‌ها در این زمینه، روندهای کلّی مشاهده‌شده در یادگیری ماشین را دنبال کرده است. روش یادگیری که در خوشه‌بندی استفاده می‌شود، می‌تواند عملکرد خوشه‌بندی را بهبود بخشد. متأسّفانه، جمع‌آوری داده‌های دارای برچسب در بسیاری از داده‌های دنیای واقعی دشوار است، درحالی‌که داده‌های بدون برچسب به‌آسانی در دسترس هستند. این موضوع، پژوهشگران را به استفاده از روش‌های یادگیری نیمه-نظارتی مبتنی بر ویژگی‌هایی که از داده‌های دارای برچسب و فاقد برچسب برای ارزیابی ارتباطی ویژگی‌ها استفاده می‌کنند، ترغیب می‌کند. در همین راستا انگِلِن و هوس [8] مروری جامع و به‌روز در مورد روش‌های یادگیری نیمه‌نظارتی ارائه دادند. آنها در گام نخست بر طبقه‌بندی نیمه‌نظارتی، که بسیاری از پژوهش‌های یادگیری نیمه‌نظارتی در این محدوده انجام می‌شود، تمرکز کردند. سپس، یک رویکرد جدید طبقه‌بندی نیمه‌نظارتی پیشنهاد دادند که در رویکردهای مختلف برای ترکیب داده‌های غیرمجاز در فرایند آموزش کاربرد دارد. دینگ و همکارانش [9] نیز الگوریتم خوشه‌بندی جدیدی را به نام خوشه‌بندی طیفی نیمه‌نظارتی بر اساس گسترش محدودیّت‌ها[4] (SSCCE) مطالعه کردند. این الگوریتم علاوه‌بر گسترش مجموعهٔ محدودیّت‌های شناخته‌شده، رابطهٔ شباهت نقاط نمونه را از طریق مسیر حساس به چگالی تغییر، سپس، با استفاده از خوشه‌بندی طیفی نیمه نظارتی عمل خوشه‌بندی را انجام داد. همچنین، جیا و همکارانش [10]، خوشه‌بندی طیفی سنّتی را در حوزهٔ یادگیری نیمه‌نظارتی گسترش دادند. آنها با کمک مقدار کمی از اطّلاعات نظارتی یک ماتریس مورّب غیربلاکی[5] ایجاد، و از آن در نظم‌دادن به یک مجموعه‌دادهٔ کم‌بُعد و انتقال آن استفاده کردند. آنها همچنین، در مطالعهٔ دیگری [11]، یک الگوی فاکتورگیری ماتریس غیرمنفی نیمه‌نظارتی[6] با استفاده از الگوسازی اطّلاعات برچسب‌دار پیشنهاد کردند. الگوی ارائه‌شده توسّط آنها قادر به تولید الگوهای نمایش کم‌بُعد، جهت تمایز عملکرد خوشه‌بندی بود. در یک کار پژوهشی دیگر، باغشاه و همکاران [12] یک روش خوشه‌بندی طیفی نیمه‌نظارتی مقیاس‌پذیر ارائه دادند که قادر به خوشه‌بندی داده‌های با ابعاد بالا بود. شیخ‌پور و همکارانش [13] در کار پژوهشی خود روش‌های انتخاب ویژگی نیمه‌نظارتی را به‌طور کامل بررسی کردند. سپس، دو طبقه‌بندی از این روش‌ها بر اساس دو دیدگاه مختلف ارائه دادند که بیانگر ساختار سلسله‌مراتبی روش های انتخاب ویژگی نیمه‌نظارتی بود. فایوبر و شوانکر [14] روش خوشه‌بندی نیمه‌نظارتی را برای داده‌های حجیم که نخست، با زیرمجموعهٔ کوچکی از داده‌ها و مجموعهٔ‌داده‌های برچسب‌دار شروع می‌شود، بررسی کردند. در این روش داده‌های جدید به‌تدریج به خوشه‌ها اضافه می‌شوند؛ به‌طوری‌که در نهایت، تمام داده‌ها در نظر گرفته شوند. آن‌ها با افزودن یک ماتریس محدودیّت جهت اعمال محدودیّت‌های مثبت (باید-متّصل)، داده‌ها را با روش K-means هسته‌ای خوشه‌بندی کردند و بیان کردند که محدودیّت‌های منفی (نباید-متّصل) با ماتریس هسته ارتباط ندارند. سوگیاما و همکاران [15] رویکردی را با استفاده از بیشینه‌سازی اطّلاعات برای مسألهٔ خوشه‌بندی نیمه‌نظارتی مطالعه کردند. بر اساس این رویکرد، خوشه‌بندی با بیشینه‌سازی اطّلاعات، مسأله انتخاب الگو را مدنظر قرار می‌دهد. آنها از تابع مربّع اتلاف اطّلاعات متقابل (SMI[7]) برای بیشینه‌سازی اطّلاعات استفاده کردند.

خوشه‌بندی طیفی، برای خوشه‌بندی داده‌ها از مسألهٔ افراز گراف بهره می‌برد. برش چیگر[8] یک معیار بهینه برای تقسیم‌بندی نمودار است. برای ربه کمترین حد رساندنِ تابع هدف برش چیگر، به تجزیهٔ مجازی ماتریس لاپلاسی-$\boldsymbol{p}$[9] نیاز است. بوهر و هِین [16] یک نسخهٔ کلّی از خوشه‌بندی طیفی را با استفاده از گراف لاپلاسی-$\boldsymbol{p}$ ، به‌عنوان تعمیمی غیرخطی از گراف لاپلاسی استاندارد ارائه دادند. آنها ارتباطی بین برش چیگر و دومین بردار ویژه ماتریس لاپلاسی-$\boldsymbol{p}$ یافتند. پژوهشگران زیادی از ماتریس لاپلاسی-$\boldsymbol{p}$ در پژوهش‌ها استفاده کردند. جوست و همکاران [17] ماتریس لاپلاسی-$\boldsymbol{p}$ را برای گراف‌ها و

[1] Cannot-link
[2] Pairwise constraints
[3] Hard constraint
[4] semi-supervised spectral clustering based on constraints expansion
[5] Anti-block-diagonal
[6] nonnegative matrix factorization
[7] Squared-loss Mutual Information
[8] Cheeger cut
[9] P-Laplacian

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

همچنین، عملگر رأس لاپلاسی[1] را برای تنظیم کلّی نمونه‌های شیمیایی، تعمیم دادند. سایتو و همکارانش [18] قیاس بین گراف لاپلاسی و هندسه دیفرانسیلی را به تنظیمات هایپرگراف تعمیم دادند و یک هایپرگراف لاپلاسی-p را پیشنهاد کردند. برخلاف گراف‌های لاپلاسی استاندارد، این تعمیم امکان تجزیه و تحلیل روی هایپرگراف‌ها را که در آن لبه‌ها اجازهٔ اتصال هر تعداد گره دارند، فراهم کرد. دینگ و همکاران [19] با استفاده از گرانول ویژگی همسایگی، خوشه‌بندی طیفی-p را بهبود و الگوریتم NAG-PSC را پیشنهاد دادند. مجموعه‌های ناهموار همسایگی می‌توانند داده‌های پیوسته را به‌طور مستقیم پردازش کنند.

یادگیری گروهی، یکی از روش‌های یادگیری در مواجهه با داده‌های حجیم است. ازآنجاکه، آموزش یک الگو با این حجم از داده در عمل امکان‌پذیر نیست، باید حجم داده کم شود تا بتوانیم الگو را آموزش دهیم. از طرفی دیگر، با کاهش حجم داده، تمام فضای ویژگی برای الگو قابل‌شناسایی نخواهد بود. بنابراین، روش‌های یادگیری گروهی اغلب زیرمجموعه‌ای از ویژگی‌های مختلف (زیرفضاهای مختلف) را استخراج می‌کنند و سپس، از الگوریتم‌های یادگیری چندگانه برای تولید نتایج پیش‌بینی (که بیشتر الگوریتم‌های یادگیری ضعیف هستند)، استفاده می‌کنند. خوشه‌بندی گروهی به رویکردی اشاره دارد که در آن تعدادی از خوشه‌های پایه (به‌طور معمول ضعیف) ایجاد می‌شوند و خوشه‌بندی نهایی برایند این خوشه‌بندی‌های ضعیف است. دانگ و همکارانش [20] پیشرفت تحققِ رویکردهای اصلی یادگیری گروهی را بررسی، و آنها را براساس ویژگی‌های مختلف طبقه‌بندی کردند. نیو و همکارانش [21] نیز، یک چارچوب خوشه‌بندی گروهی ارائه دادند که از یک الگوریتم خوشه‌بندی ساده مانند الگوریتم خوشه‌بندی kmedoids استفاده می‌کند. یو و همکاران [22] رویکردی گروهی را پیشنهاد کردند و نام آن را چارچوب گروهی خوشه‌بندی مبتنی بر دانش نامیدند. روش آنان به این صورت است که در مرحلهٔ نخست، زیرمجموعه‌های تصادفی از زیرفضاها[2] تولید می‌شود و در مرحلهٔ دوم، بر اساس خاصیّت تراگذری محدودیّت‌های دوبه‌دو را گسترش داده، سپس، با در نظر گرفتن هریک از زیرمجموعه‌های تصادفی، یک خوشه‌بندی متفاوت باتوجّه‌به روش‌های تراگذری متفاوت، ارائه می‌دهند. سپس، یک مرحلهٔ تطبیقی برای افزایش اطمینان هرکدام از الگوریتم‌های خوشه‌بندی اجرا می‌کنند که از یک فاکتور اطمینان طرّاحی‌شده استفاده می‌کند. همچنین، آنها در مطالعه‌ای دیگر [23]، اعضای گروه را به‌عنوان ویژگی‌ها در نظر گرفتند و نحوهٔ استفاده از روش‌های انتخاب ویژگی مناسب را برای انتخاب اعضای گروه بررسی کردند. بیشتر رویکردهای گروهی خوشه‌بندی تنها با استفاده از یک امتیاز شباهت یا روش انتخاب ویژگی برای حذف اعضای اضافی گروه عمل می‌کنند و تعداد کمی از آن‌ها از یک روش بهینه‌سازی برای یافتن زیرمجموعه مناسب از اعضای گروه استفاده کردند. روش‌های یادگیری گروهی علاوه‌بر مسائل خوشه‌بندی، در بسیاری از مسائل طبقه-بندی نیز به کار گرفته شده‌اند. گالار و همکاران [24] یک الگوریتم گروهی طبقه‌بندی جدید به نام EUSBoost طراحی کردند که زیرنمونه‌های تصادفی را با افزودن نمونه به گروه برای رسیدگی به مشکل عدم تعادل داده‌ها ترکیب کرد.

در این مطالعه، یک چارچوب گروهی خوشه‌بندی نیمه‌نظارتی افزایشی[3] (ISSCE) پیشنهاد می‌شود، که از یک فرایند جدید انتخاب عضو برای تولید یک زیرمجموعهٔ بهینه از اعضا استفاده کرده است.

۳- روش پیشنهادی

در رویکرد پیشنهادی، نخست، زوج محدودیّت‌های باید-متّصل و نباید-متّصل را روی کل داده‌ها انتشار می-دهیم. سپس، فضای ویژگی را به‌صورت تصادفی به چندین زیرفضای نابرابر تقسیم می‌کنیم. تولید زیرفضاها به این صورت است که هر زیرفضا بخشی از ویژگی‌ها را پوشش می‌دهد؛ به‌عنوان مثال، اگر کلّ فضای ویژگی m-بُعدی باشد، هر زیرفضا m'-بُعدی خواهد بود. ما در کار پیشین [25]، جهت سادگی در تولید زیرفضاها ابعاد زیرفضاها را برابر در نظر گرفته بودیم، امّا در پژوهش جاری این محدودیّت حذف شد تا بتوانیم بهترین زیرفضای شامل تأثیرگذارترین ویژگی‌ها را بیابیم. در واقع، رویکرد پیشنهادی با کمک یادگیری گروهی و خوشه‌بندی طیفی لاپلاسی-p به نوعی از مزایای روش‌های انتخاب ویژگی نیز برای خوشه‌بندی داده‌های حجیم با ابعاد بالا استفاده می‌کند. با استفاده از زوج محدودیّت‌های به‌روزرسانی‌شده،

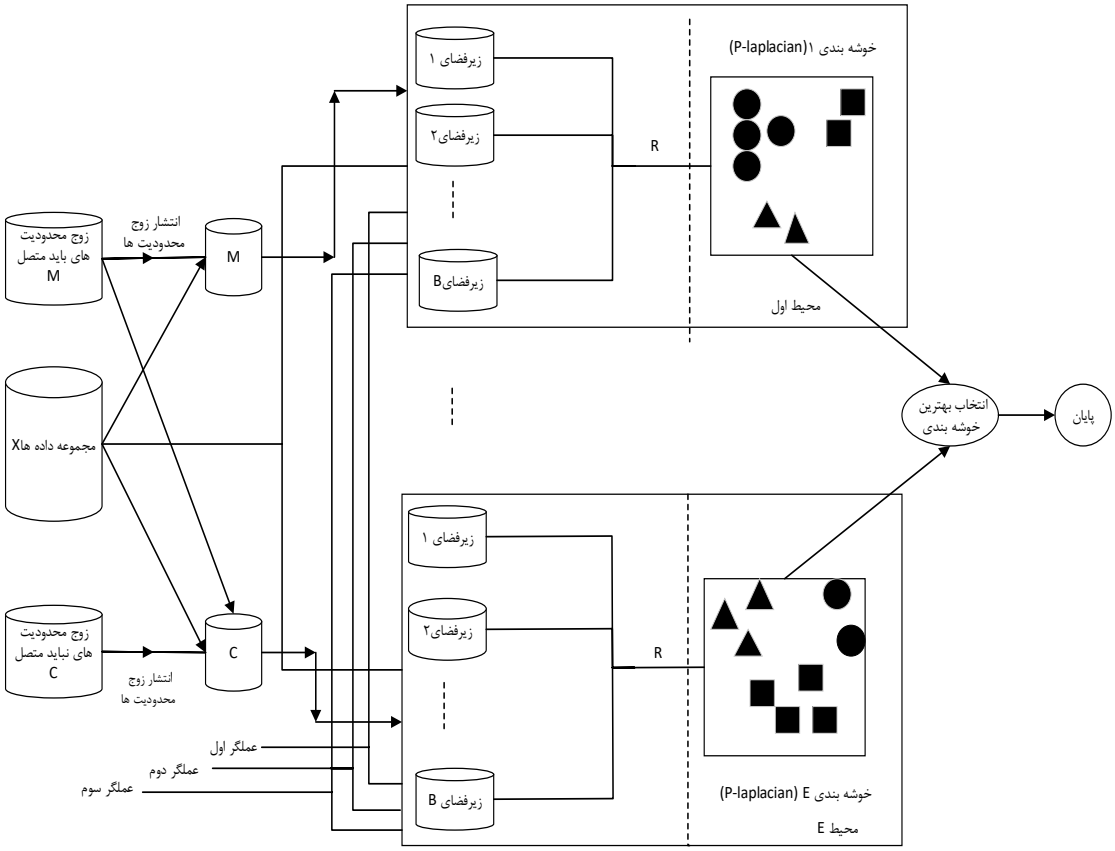
[1] vertex Laplace
[2] subspaces
[3] incremental semi-supervised clustering ensemble

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

خوشه‌بندی طیفی نیمه‌نظارتی در هر زیرفضا را به‌طور مستقل انجام داده، سپس، با استفاده از نتایج هرکدام یک ماتریس مجاورت، حاصل از برآیند نتایج هرکدام (مبتنی بر یادگیری گروهی) ایجاد می‌کنیم. ازآنجاکه تقسیم فضای ویژگی به تعدادی زیرفضا به‌طور تصادفی انجام شده است، بنابراین، به یک روند بهینه‌سازی جهت جستجوی بهترین تقسیم‌بندی یا به عبارتی دیگر، بهترین محیط حاصل از مجموعهٔ زیرفضاها نیاز داریم.

بنابراین، با استفاده از چندین عملگر جستجو روی محیط‌های متشکل از زیرفضاهای متفاوت، بهترین مجموعهٔ زیرفضاها، یعنی محیطی از زیرفضاها که بهترین نتیجه خوشه‌بندی را دارد، می‌یابیم.

مراحل روش پیشنهادی در شکل شماره (۱) به تصویر کشیده شده است. در ادامه، هریک از مراحل روش پیشنهادی را جداگانه توضیح می‌دهیم.



(شکل- ۱): نمودار (دیاگرام) روش پیشنهادی خوشه‌بندی گروهی داده‌های با ابعاد بالا
(figure-1): The diagram of the proposed clustering method of high-dimensional data

۳-۱-انتشار زوج محدودیّت‌ها

فرض کنید $X = \{x_1, x_2, \dots, x_n\}$، مجموعهٔ داده‌ها که شامل n نمونه است، داده شده باشد. همچنین، فرض کنید که مجموعه‌های زوج محدودیّت-ها؛ $M = \{(x_i, x_j) | \, y_i = y_j, 1 \le i, j \le n\}$ که نشان-دهندهٔ مجموعهٔ زوج محدودیّت‌های باید-متّصل و $C = \{(x_i, x_j) | \, y_i \ne y_j, 1 \le i, j \le n\}$ که نشان‌دهندهٔ مجموعهٔ زوج محدودیّت‌های نباید-متّصل است، به‌عنوان دانش پیشین داده شده باشند (که y_i و y_j برچسب ردهٔ داده‌های x_i و x_j هستند). سپس، یک ماتریس مربّعی $n \times n$ به نام R معرفی می‌کنیم که اطّلاعات تمام زوج محدودیّت‌ها را در خود تجمیع کرده است:

$$r_{ij} = \begin{cases} 1 & if \ (x_i, x_j) \in M \\ -1 & if \ (x, x_j) \in C \\ 0 & otherwise \end{cases} \qquad (۱)$$

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

پس از آن، با استفاده از خاصیّت تراگذری که به‌صورت زیر تعریف می‌شود، زوج محدودیّت‌ها به تمامی داده‌ها تعمیم داده می‌شوند.

مطابق خاصیّت تراگذری، برای هر $1 \le i,j,k \le n$، چنانچه زوج‌داده $(x_i, x_j) \in M$، و زوج‌داده $(x_j, x_k) \in M$ باشند، در نتیجه، زوج‌داده $(x_i, x_k) \in M$ خواهد بود. بنابراین، براساس رابطهٔ هم‌ارزی سه دادهٔ x_i، x_j و x_k در یک ردهٔ یکسان قرار می‌گیرند و به این ترتیب، تعداد بیشتری زوج‌داده در مجموعهٔ زوج محدودیّت‌های باید-متّصل قرار می‌گیرند. مجموعهٔ به‌روزرسانی‌شده را با $\widetilde{M}$ نمایش می‌دهیم. به روش مشابه، می‌توانیم زوج محدودیّت‌های نباید-متّصل را نیز تعمیم دهیم. با توجه به تعریف روابط باید-متّصل و نباید-متّصل، چنانچه زوج‌داده (x_i, x_j) عضو مجموعه M، و زوج‌داده (x_j, x_k) عضو مجموعه C باشند، در نتیجه زوج‌داده (x_i, x_k) عضو مجموعه C خواهد بود؛ بنابراین، اگر دست‌کم یک زوج‌محدودیّت به مجموعهٔ زوج محدودیّت-های باید-متّصل در مرحلهٔ پیش اضافه شده باشد که پیشتر یکی از داده‌های آن در دست‌کم، یک زوج‌محدودیّت نباید-متّصل ظاهر شده باشد، آنگاه زوج محدودیّت‌های نباید-متّصل نیز به‌روزرسانی می‌شود؛ که آن را با $\widetilde{C}$ نمایش می‌دهیم. قابل ذکر است که باتوجّه‌به مجموعه‌های به‌روزرسانی‌شدهٔ زوج محدودیّت‌ها، ماتریس R نیز به‌روزرسانی می‌شود. جهت انتشار زوج محدودیّت‌های دوبه‌دو، از روش‌های مبتنی بر گراف‌های طیفی استفاده می‌کنیم، به‌این‌صورت‌که نخست، مجموعهٔ داده‌ها را با یک گراف وزن‌دار نمایش می‌دهیم و داده‌ها به‌عنوان رئوس گراف و شباهت مابین داده‌ها به‌عنوان وزن هر یال گراف در نظر گرفته می‌شود. بنابراین، ماتریس وزن به‌صورت زیر تعریف می‌شود:

$$w_{ij} = \begin{cases} \exp\left(-\dfrac{\|x_i - x_j\|^2}{2\sigma^2}\right) & \text{if } x_j(x_i) \in \mathcal{N}_k\big(x_i(x_j)\big) \\ 0 & \text{otherwise} \end{cases} \quad (۲)$$

که در آن، $\mathcal{N}_k$ مجموعه k همسایه داده x_i (x_j) و σ به‌عنوان یک شاخص بیانگر شعاع همسایگی است. پس از به‌روزرسانی زوج محدودیّت‌های باید-متّصل و نباید-متّصل ماتریس وزن‌ها نیز باید به‌روزرسانی شود. به این منظور، از یک ماتریس مربّعی $n \times n$ به نام $\widetilde{R}$ استفاده می‌کنیم که پس از انتشار زوج محدودیّت‌ها با استفاده از خاصیّت تراگذری، اطّلاعات مربوط به دانش پیشین تعمیم‌یافته را نمایش می‌دهد. بنابراین، اگر هیچ عنصری به مجموعه‌های باید-متّصل و نباید-متّصل افزوده نشده باشد، در این صورت $\widetilde{R} = R$ است و در غیراین‌صورت، مشابه مسئلهٔ برش کمینه[1] به‌همراه عبارت تنظیم‌کنندهٔ رابطه محاسبهٔ ماتریس U، به‌صورت زیر نوشته می‌شود:

$$\min_{\tilde{R}} \frac{1}{2}\|\tilde{R} - R\|_F^2 + \frac{\lambda}{2} tr\left(\tilde{R}^T L \tilde{R} + \tilde{R} L \tilde{R}^T\right) \quad (۳)$$

که در آن $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ ماتریس لاپلاسی گراف[2] و $d_{ii} = \sum_j w_{ij}$ ماتریس درجهٔ گراف[3] است. معادلهٔ بالا یک جواب بسته و یکتا دارد که با استفاده از مشتق‌گیری نسبت به متغیّر مجهول مسئله، مقدار بهینهٔ ماتریس $\widetilde{R}$ مطابق رابطهٔ زیر محاسبه می‌شود:

$$\tilde{R} = \left(\frac{1}{1+\lambda}\right)^2 \left(I - \frac{1}{1+\lambda}\bar{L}\right)^{-1} R \left(I - \frac{1}{1+\lambda}\bar{L}\right)^{-1} \quad (۴)$$

که در آن $\bar{L} = I - L$ است.

هر دراهٔ ماتریس $\widetilde{R}$ (یعنی $\tilde{r}_{ij}$) میزان احتمال هم‌رده بودن دو دادهٔ x_i و x_j را نشان می‌دهد؛ بنابراین، اگر مقدار درایه $\tilde{r}_{ij}$ کم باشد، به این معنی است که میزان احتمال هم‌رده بودن دو دادهٔ x_i و x_j نیز کم است. درنتیجه، میزان وزن (شباهت) بین دو داده باید کاهش پیدا کند؛ پس وزن به‌روزرسانی شده برابر است با:

$$\widetilde{w}_{ij} = (1 + \tilde{r}_{ij}) w_{ij} \quad (۵)$$

همچنین، اگر مقدار درایه $\tilde{r}_{ij}$ زیاد باشد، به این معنی است که میزان احتمال هم‌رده بودن دو دادهٔ x_i و x_j نیز زیاد است؛ در نتیجه میزان وزن (شباهت) بین دو داده باید افزایش پیدا کند؛ پس وزن به‌روزرسانی‌شده برابر است با:

$$\widetilde{w}_{ij} = 1 - (1 - \tilde{r}_{ij})(1 - w_{ij}) \quad (۶)$$

پس از انتشار زوج محدودیّت‌های باید-متّصل و نباید-متّصل و به‌روزرسانی ماتریس وزن، می‌توانیم با استفاده از خوشه‌بندی طیفی کل داده‌ها را خوشه‌بندی کنیم.

۳-۲- تولید محیط‌های جستجو و خوشه‌بندی گروهی در هر محیط

همان‌طورکه در شکل (۱) نشان داده شده است، برای‌این‌که بتوانیم بهترین خوشه‌بندی در مجموعهٔ

[1] Min Cut
[2] Laplacian matrix
[3] Degree matrix

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

زیرفضاها را بیابیم، تعدادی زیرمجموعهٔ متفاوت از زیرفضاها (محیط) ایجاد کرده، سپس، با استفاده از چندین عملگر جستجو بهترین زیرفضا را می‌یابیم. به این منظور، نخست تعدادی محیط متشکل از تعدادی زیرمجموعه از زیرفضاهای تصادفی تولید می‌کنیم. فرض کنید شاخص $\boldsymbol{E}$ نشان‌دهندهٔ تعداد محیط‌ها باشد و شاخص $\boldsymbol{B}$ تعداد زیرفضاهای هر محیط را نشان دهد؛ که این دو شاخص توسط کاربر معرفی می‌شوند. در این صورت، $\left\{\boldsymbol{\Phi}^{\boldsymbol{1}},\dots,\boldsymbol{\Phi}^{\boldsymbol{E}}\right\}$ مجموعهٔ محیط‌ها هستند و $\{\boldsymbol{S}^{\boldsymbol{1}},\dots,\boldsymbol{S}^{\boldsymbol{B}}\}$ مجموعهٔ زیرفضاها که به‌طور کامل تصادفی تولید شده‌اند. نخست، خوشه‌بندی طیفی نیمه‌نظارتی مبتنی بر ماتریس وزن به‌روزرسانی‌شده در رابطهٔ (۶) از زیربخش ۳-۱ برای داده‌های هر یک از زیرفضاهای هر محیط به‌طور جداگانه انجام می‌شود.

در روش پیشنهادی برای‌این‌که بتوانیم دقّت خوشه‌بندی طیفی را به‌طور قابل ملاحظه‌ای افزایش دهیم، از روش خوشه‌بندی طیفی مبتنی بر گراف لاپلاسی-$\boldsymbol{p}$ استفاده شده است. همان‌گونه‌که پیش‌تر ذکر شد، گراف لاپلاسی-$\boldsymbol{p}$ یک تعمیم غیرخطی از گراف لاپلاسی است. در خوشه‌بندی طیفی از مقادیر درایه‌های دومین بردار ویژگی ماتریس لاپلاسی گراف استفاده می‌شود؛ که در روش پیشنهادی نیز از خوشه‌بندی طیفی مبتنی بر ماتریس لاپلاسی-$\boldsymbol{p}$ استفاده شده است. در پژوهشی که بوهلر و هِین [۱۵] انجام دادند، نشان داده شد که ارتباطی بین دومین بردار ویژه ماتریس لاپلاسی-$\boldsymbol{p}$ و برش چیگر وجود دارد و در واقع، دومین بردار ویژه ماتریس لاپلاسی-$\boldsymbol{p}$ زمانیکه $\boldsymbol{p}\to\boldsymbol{1}$، به‌سمت جواب بهینهٔ برش چیگر میل می‌کند. در روش خوشه‌بندی طیفی مبتنی بر گراف لاپلاسی-$\boldsymbol{p}$ بجای نُرم-$\boldsymbol{2}$ از نُرم-$\boldsymbol{p}$[1] در محاسبات گراف استفاده می‌شود. پس از انجام خوشه-بندی، نتایج هر خوشه‌بند برای هرکدام از زیرفضاها را در یک ماتریس مجاورت به نام $\boldsymbol{A^b}$ که $\boldsymbol{b}=\boldsymbol{1},\dots,\boldsymbol{B}$، قرار می‌دهیم که درایه‌های آن به‌صورت زیر مقداردهی می-شوند:

$$a_{ij}^{b} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ is in the same cluster,} \\ 0 & \text{otherwise} \end{cases} \qquad (۷)$$

سپس، یک فاکتور اطمینان برای نمایش میزان تطابق خوشه‌بندی حاصل با زوج محدودیّت‌ها محاسبه می‌شود. به این منظور، یک ماتریس مجاورت برای نمایش زوج محدودیّتها را به‌صورت زیر معرفی کرده، سپس، فاکتور اطمینان را به دست می‌آوریم:

$$\hat{a}_{ij} = \begin{cases} 1 & \text{if } (x_i, x_j) \in M \\ 0 & \text{if } (x_i, x_j) \in C \\ -1 & \text{otherwise} \end{cases} \qquad (۸)$$

که $\boldsymbol{M}$ مجموعه محدودیّت‌های باید-متّصل و $\boldsymbol{C}$ مجموعه محدودیّت‌های نباید-متّصل است. با استفاده از ماتریسهای مجاورت معرفی‌شده در روابط (۷) و (۸) نرخ درستی خوشه‌بندی زوج‌داده های مربوط به محدودیّت‌ها را محاسبه می‌کنیم:

$$\iota^b = \frac{\sum_i \sum_{j \text{ and } j>i} \{a_{ij}^b = \hat{a}_{ij}\}}{\sum_i \sum_{j \text{ and } j>i} 1\{\hat{a}_{ij} \neq -1\}} \qquad (۹)$$

سپس، نرخ $\boldsymbol{\iota^b}$، مطابق رابطهٔ زیر طبیعی‌سازی و مقیاس‌بندی می‌شود:

$$\bar{\iota}^b = \xi \cdot \frac{\iota^b - \iota_{min}}{\iota_{max} - \iota_{min}} \qquad (۱۰)$$

که در آن $\boldsymbol{\iota_{min}} = \min_{\boldsymbol{b}} \boldsymbol{\iota^b}$ و $\boldsymbol{\iota_{max}} = \max_{\boldsymbol{b}} \boldsymbol{\iota^b}$ و $\boldsymbol{\xi}$ ضریب مقیاس است. در نهایت، فاکتور اطمینان مربوط به هر زیرفضا به‌صورت زیر محاسبه می‌شود:

$$\pi^b = \frac{B^{\bar{\iota}^b}}{max_b\, B^{\bar{\iota}^b}} \qquad (۱۱)$$

هرچه میزان فاکتور اطمینان یک خوشه‌بند عدد بزرگتری باشد، به این معناست که این خوشه‌بند تعداد بیشتری از محدودیّت‌ها را ارضا کرده است.

در آخرین مرحله، پس از محاسبهٔ فاکتور اطمینان، یک ماتریس مجاورت برای تمامی داده‌ها در همه‌ی زیرفضاها با اجماع نتایج تمامی خوشه‌بندها و فاکتور اطمینان تولید می‌کنیم. این ماتریس مجاورت گروهی با استفاده از روابط (۷) و (۱۱) به‌صورت زیر به دست می‌آید:

$$\mathcal{A} = \frac{1}{B} \sum_{b=1}^{B} (\pi^b A^b) \qquad (۱۲)$$

اکنون با استفاده از روش خوشه‌بندی طیفی مبتنی بر گراف لاپلاسی-$\boldsymbol{p}$ که در آن ماتریس لاپلاسی $\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{D}}^{-\frac{1}{2}} \boldsymbol{\mathcal{A}} \boldsymbol{\mathcal{D}}^{-\frac{1}{2}}$ و $\boldsymbol{d_{ii}} = \sum_{\boldsymbol{j}} \boldsymbol{a_{ij}}$ است، تمامی داده‌ها در هرکدام از محیط‌ها به‌طور جداگانه خوشه‌بندی می‌شوند.

از آنجاکه زیرفضاها به‌طور تصادفی و با تعداد ویژگی‌های نابرابر تولید شده‌اند، نتایج خوشه‌بندی به‌شدت تحت‌تأثیر مقداردهی‌های اولیّهٔ متفاوت، قرار می-گیرند. بنابراین، یافتن بهترین زیرفضای راه‌حل امری

[1] p-norm

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

ضروری است که طیّ یک فرایند جستجو با استفاده از چند عملگر، باید بهترین زیرفضای ممکن را بیابیم. بنابراین، با استفاده از محیط‌های گوناگون چندین زیرفضای راه حل تولید کرده و سپس، در یک روش متناوب، بهترین زیرفضای راه‌حل را می‌یابیم. عملگرهای جستجو به دو صورت محلی و سراسری تمامی محیط‌ها و زیرفضاهای هر محیط را جستجو می‌کنند. نحوهٔ عملکرد این عملگرها به‌صورت ترکیب و به‌روزرسانی دو محیط و یا به‌روزرسانی فقط یک محیط است. هدف از ترکیب دو محیط این است که محیط‌های جدیدی حاصل از ترکیب بهترین زیرفضاهای هرکدام از دو محیط اولیّه به دست بیاید. همچنین، با به‌روزرسانی فقط یک محیط، سعی می‌کنیم بهره‌وری یک محیط با تغییر یکی از زیرفضاهای آن محیط را افزایش دهیم. بنابراین، عملگر نخست، روی دو محیط که اولی یک محیط دلخواه و دومی یکی از محیط‌های مشابه و نزدیک به آن است، عمل جستجو و به‌روزرسانی را انجام می‌دهد که آن را عملگر محلی می‌نامیم. عملگر دوم، دو محیط را به‌طور کامل تصادفی انتخاب کرده و آن دو را به‌روزرسانی می‌کند؛ که آن را عملگر سراسری می‌نامیم. و در نهایت، یک عملگر نیز فقط روی یک محیط اعمال می‌شود.

۳-۲-۱- پیاده‌سازی روش پیشنهادی

جهت پیاده‌سازی عمل جستجو به‌صورت محلی یا سراسری، لازم است محدودهٔ جستجو را با تعیین میزان نزدیکی یا شباهت محیط‌ها تعیین کنیم. همان‌گونه‌که پیشتر، بیان کردیم، تعدادی زیرفضای تصادفی نابرابر در چند محیط مجزّا تولید شده است. باتوجّه‌به ویژگی‌هایی که هر زیرفضا بر اساس آنها تولید شده است، می‌توان محیط‌های نزدیک (مشابه) به هم را تعیین کرد. برای تعیین میزان شباهت هر دو محیط از یک معیار شباهت فازی استفاده می‌کنیم. به این منظور، نخست یک بردار شاخص را که نشان‌دهندهٔ ویژگی‌های انتخاب‌شدهٔ هر زیرفضای درون هر محیط است، برای هر محیط به‌صورت زیر تعریف می‌کنیم:

$$v^i = \sum_b v_b \qquad (۱۳)$$

سپس، میزان شباهت دوبه‌دوی همه‌ی محیط‌ها را می‌یابیم. میزان شباهت محیط‌های $\boldsymbol{i}$ و $\boldsymbol{j}$ با استفاده از معیار شباهت فازی زیر محاسبه می‌شود:

$$Sim(\Phi^i, \Phi^j) = \frac{1}{\sum_{k=1}^{K}\left(\frac{\|v^i - v^j\|}{\|v^i - v^k\|}\right)^{\frac{2}{q-1}}} \qquad (۱۴)$$

که در آن $\boldsymbol{K}$ و $\boldsymbol{q}$، دو شاخص هستند که به‌ترتیب تعداد همسایه‌های نزدیک و فازی‌ساز را نمایش می‌دهند. هرچه مقدار $\boldsymbol{Sim}$ بیشتر باشد، احتمال انتخاب محیط $\boldsymbol{j}$ به‌عنوان همسایهٔ نزدیک و مشابه محیط $\boldsymbol{i}$ بیشتر می‌شود.

سپس، بر اساس میزان شباهت محیط‌ها، یک درجهٔ عضویّت تجمّعی معرفی می‌کنیم:

$$\mathrm{K} = \sum_{h=1}^{k} \boldsymbol{Sim}(\boldsymbol{\Phi^i}, \boldsymbol{\Phi^h}) \qquad (۱۵)$$

با استفاده از این دو مقدار، یعنی میزان شباهت محیط‌ها و درجهٔ عضویّت تجمّعی، اندازهٔ فاصلهٔ احتمالی هر محیط با محیط‌های دیگر محاسبه می‌شود:

$$\boldsymbol{\rho_l} = \left[\frac{\boldsymbol{\kappa_{l-1}}}{\boldsymbol{\kappa}} - \frac{\boldsymbol{\kappa_l}}{\boldsymbol{\kappa}}\right] \qquad (۱۶)$$

$$\mathrm{K}_{\boldsymbol{l}} = \mathrm{K}_{\boldsymbol{l-1}} + \boldsymbol{Sim}(\boldsymbol{\Phi^i}, \boldsymbol{\Phi^j}) \qquad (۱۷)$$

مقدار $\mathrm{K}_{\mathbf{0}} = \mathbf{0}$ در نظر می‌گیریم و $\boldsymbol{l} \in \{\mathbf{1}, \dots, \boldsymbol{k}\}$ است. در ادامه، نحوهٔ عملکرد هرکدام از عملگرها را جداگانه معرفی می‌کنیم:

عملگر جستجوی محلّی: در این الگو دو محیط انتخاب می‌شوند که محیط نخستین تصادفی و محیط دوم از بین همسایه‌های نزدیک نخستین با استفاده از رابطهٔ (۱۶) انتخاب می‌شود. فرض کنید محیط اول $\boldsymbol{\Phi^i}$ و محیط دوم $\boldsymbol{\Phi^j}$ باشد. پس از انجام خوشه‌بندی در هرکدام از زیرفضاهای محیط‌های $\boldsymbol{i}$ و $\boldsymbol{j}$، آنها را بر اساس میزان ارضای محدودیّت‌های دوبه‌دو رتبه‌بندی می‌کنیم. معیار رتبه‌بندی که بر اساس میزان ارضای محدودیّت‌ها ساخته شده است، به‌صورت زیر تعریف می‌شود:

$$\boldsymbol{\eta}(\boldsymbol{S^b}) = \frac{\boldsymbol{N_s}}{\boldsymbol{N_t}} \qquad , \mathrm{b}=1, \dots, \mathrm{B} \qquad (۱۸)$$

که در آن $\boldsymbol{N_s}$، تعداد محدودیّت‌های ارضاشده توسّط خوشه‌بند مربوط به زیرفضای $\boldsymbol{S^b}$ و $\boldsymbol{N_t}$، تعداد کل محدودیّت‌ها است. محیط‌های $\boldsymbol{\Phi^i}$ و $\boldsymbol{\Phi^j}$ پس از رتبه‌بندی بر اساس میزان ارضای محدودیّت‌ها به‌صورت زیر تعریف می‌شوند:

$$\boldsymbol{\Phi^i} = \{\boldsymbol{S^{i_1}_{\Phi^i}}, \boldsymbol{S^{i_2}_{\Phi^i}}, \dots, \boldsymbol{S^{i_B}_{\Phi^i}}\} \qquad (۱۹)$$

$$\boldsymbol{\Phi^j} = \{\boldsymbol{S^{i_1}_{\Phi^j}}, \boldsymbol{S^{i_2}_{\Phi^j}}, \dots, \boldsymbol{S^{i_B}_{\Phi^j}}\} \qquad (۲۰)$$

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]

[DOI: 10.61186/jsdp.20.1.39]

که $S_{\Phi_i}^{i_1}$ ، نشان‌دهندهٔ زیرفضای i_1 از محیط Φ^i است و ترتیب قرارگیری زیرفضاها در هر محیط به‌صورت زیر است:

$$\eta\left(S_{\Phi^i}^{i_1}\right) \geq \eta\left(S_{\Phi^i}^{i_2}\right) \geq \dots \geq \eta\left(S_{\Phi^i}^{i_B}\right) \quad (۲۱)$$

دوباره محیط‌های مسأله به‌روزرسانی می‌شوند.

عملگر جستجوی تصادفی: در این مرحله یک محیط تصادفی به نام Φ^i انتخاب می‌شود و با استفاده از یک عدد تصادفی به نام r_3 در محدودهٔ $[0,1]$، زیرفضای $S_{\Phi^i}^{i_{[r_3B]}}$ انتخاب شده و یک زیرفضای جدید جایگزین آن می‌شود:

$$\Phi'^i = \left\{S_{\Phi^i}^{i_1}, \dots, S'_{\Phi^i}, S_{\Phi^i}^{i_{[r_5B]}+1}, \dots, S_{\Phi^i}^{i_B}\right\} \quad (۲۸)$$

این مراحل بر روی محیط‌های مسأله، تا زمانی که تغییر در محیط‌ها قابل‌ملاحظه باشد، یا این‌که به بیشترین تعداد تکرار الگوریتم نرسیده باشد، تکرار می‌شوند و سپس، از بین تمام محیط‌های موجود در مسأله، محیط بهینه (محیطی که دارای کمترین خطای خوشه‌بندی است) انتخاب و مسئلهٔ خوشه‌بندی روی آن محیط برتر اعمال می‌شود.

۴- نتایج آزمایش‌ها

در این کار پژوهشی، ما کارایی عملکرد ایدهٔ پیشنهادی خود را بر روی ۲۰ مجموعه دادهٔ ژنتیکی بیماری سرطان ارزیابی کرده‌ایم [26]. ویژگی‌های این ۲۰ مجموعه‌داده در جدول (۱) نشان داده شده‌اند. در جدول (۱)، n نشان‌دهندهٔ تعداد نمونه داده‌ها است، mتعداد ویژگی‌ها و k تعداد رده‌ها را نشان می‌دهد. نتایج به‌دست‌آمده در این کار پژوهشی با استفاده از زبان برنامه‌نویسی متلب به دست آمده است. برای آزمایش ایدهٔ پیشنهادی، زوج محدودیّت‌ها را به‌صورت تصادفی با استفاده از مجموعهٔ داده‌ها تولید کرده‌ایم. تعداد زوج محدودیّت‌ها برای تمامی مجموعه‌داده‌های آزمایشی یکسان در نظر گرفته شده است. همچنین، مجموعهٔ زوج محدودیّت‌ها در تمامی اجراهای هرکدام از مجموعهٔ داده‌ها یکسان و بدون تغییر هستند.

مجموعه‌داده‌های معرفی‌شده در جدول (۱)، مجموعه‌داده‌های چالش برانگیزی هستند که در آنها ابعاد داده‌ها بسیار بالا و اندازهٔ نمونه کوچک است؛ به‌عنوان مثال، مجموعه‌داده Garber-2001 شامل ۶۶ نمونه و هر نمونه دارای ۴۵۵۳ بُعد است. رویکردهای متداول خوشه‌بندی نیمه‌نظارتی نمی‌توانند نتایج رضایت‌بخشی در این مجموعه‌داده به دست آورند.

به‌عنوان مثال دیگر، مجموعه‌داده Ramaswamy-2001، ۱۹۰ نمونه با ابعاد بسیار بالا دارد که به ۱۴ رده اختصاص یافته است. برای این مورد، روش‌های خوشه‌بندی سنّتی به دلیل تعداد زیاد رده‌ها، نمی‌توانند به طور مؤثّر بر این مجموعه‌داده اعمال شوند. بنابراین، می‌توان از این مجموعه‌های داده برای بررسی دقیق‌تر مرزهای عملکرد روش‌های خوشه‌بندی نیمه‌نظارتی استفاده کرد.

در رویکرد پیشنهادی، از دو معیار اطّلاعات متقابل طبیعی‌شده[1] (NMI) [27] و شاخص تصادفی تنظیم‌شده[2] (ARI) [28] برای بررسی ارزیابی عملکرد استفاده می‌شود. معیار NMI اطّلاعات آماری اشتراکی متغیّرهای تصادفی مربوط به خوشه‌بندی به‌دست‌آمده و خوشه‌بندی هدف را اندازه‌گیری می‌کند. فرض کنید $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_c\}$ و $\widehat{\mathcal{C}} = \{\widehat{\mathcal{C}}_1, \dots, \widehat{\mathcal{C}}_k\}$ به ترتیب خوشه‌بندی هدف به دست آمده باشند، و $|\mathcal{C}_i| = n_i$، $|\widehat{\mathcal{C}}_j| = n'_j$، $|\mathcal{C}_i \cap \widehat{\mathcal{C}}_j| = n_{ij}$ و n تعداد کل داده‌ها باشد. معیار NMI به‌صورت زیر تعریف می‌شود:

$$NMI(\mathcal{C}, \widehat{\mathcal{C}}) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{k} n_{ij} \log\left(\frac{n \times n_{ij}}{n_i \times n'_j}\right)}{\sqrt{\left(\sum_{i=1}^{c} n_i \log\frac{n_i}{n}\right)\left(\sum_{j=1}^{k} n'_j \log\frac{n'_j}{n}\right)}} \quad (۲۹)$$

همچنین، معیار ARI یک حالت خاص از شاخص تصادفی است که برای گروه بندی تصادفی عناصر تنظیم شده و به‌صورت زیر تعریف می‌شود:

$$ARI(\mathcal{C}, \widehat{\mathcal{C}}) = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i\binom{n_i}{2}\sum_j\binom{n'_j}{2}/\binom{n}{2}}{\left[\sum_i\binom{n_i}{2} + \sum_j\binom{n'_j}{2}\right]/2 - \sum_i\binom{n_i}{2}\sum_j\binom{n'_j}{2}/\binom{n}{2}} \quad (۳۰)$$

در آزمایش‌های حاضر، تعداد زیرفضاهای تصادفی (B) را، ۲۰ و تعداد زوج محدودیّت‌ها را ۱۴۰ و همچنین، تعداد محیط‌ها را، ۲۰ در نظر گرفته‌ایم. مقدار شاخص q در فرمول (۱۴) برابر با ۰/۲ و مقدار شاخص ς در فرمول (۲۷)، ۰/۲۵، همچنین، نرخ تولید زیرفضاهای تصادفی ۰/۵ تنظیم شده است.

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

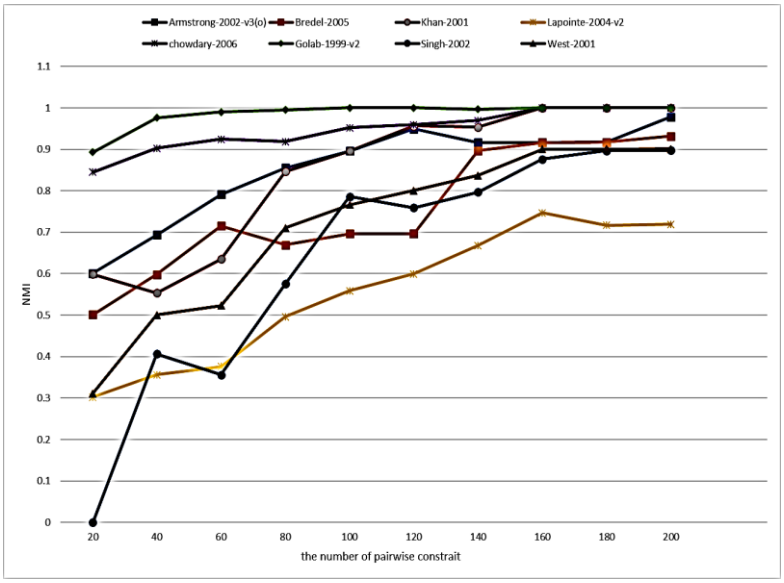
[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

تعداد دفعات تکرار الگوریتم پیشنهادی، چهل‌بار در نظر گرفته شده است و زوج محدودیّت‌ها در تمام اجراها یکسان است. نخست، تأثیر تعداد متفاوت زوج محدودیّت‌ها و تأثیر تعداد زیرفضای تصادفی و تأثیر نرخ تولید زیرفضای تصادفی را بررسی می‌کنیم. سپس، نتایج ایدهٔ پیشنهادی با نتایج روش پیشنهاد شده در مرجع [22] مقایسه می‌شود و سرانجام، ایدهٔ پیشنهادی با رویکردهای مختلف خوشه‌بندی نیمه‌نظارتی رایج و معروف بر روی مجموعه‌داده‌های دنیای واقعی مقایسه و نتایج آزمایش‌ها بحث می‌شود.

(جدول-۱) : مشخصات مجموعه‌داده‌های مربوط به سرطان در دنیای واقعی، که n تعداد نمونه‌های داده، m تعداد ویژگی‌ها و k تعداد خوشه‌هاست [22].

(Table-1): Specifications of the real-world cancer datasets, where *n* is the number of data samples, *m* is the number of features, and *k* is the number of clusters.

مجموعه‌داده	اندیس	n	m	k
Alizadeh-2000-v3(o)	D1	۶۲	۴۰۹۶	۴
Armstrong-2002-v2	D2	۷۲	۲۱۹۴	۳
Bredel-2005	D3	۵۰	۱۷۳۹	۳
Chen-2002	D4	۱۷۹	۸۵	۲
Chowdary-2006	D5	۱۰۴	۱۸۲	۲
Dyrskjot-2003	D6	۴۰	۱۲۰۳	۳
Garber-2001	D7	۶۶	۴۵۵۳	۴
Golab-1999-v2	D8	۷۲	۱۸۷۷	۳
Khan-2001	D9	۸۳	۱۰۶۹	۴
Lapointe-2004-v1	D10	۶۹	۱۶۲۵	۳
Lapointe-2004-v2	D11	۱۱۰	۲۴۹۶	۴
Nutt-2003-v1	D12	۵۰	۱۳۷۷	۴
Pomeroy-2002-v2	D13	۴۲	۱۳۷۹	۵
Ramaswamy-2001	D14	۱۹۰	۱۳۶۳	۱۴
Risinger-2003	D15	۴۲	۱۷۷۱	۴
Singh-2002	D16	۱۰۲	۳۳۹	۲
Su-2001	D17	۱۷۴	۱۵۷۱	۱۰
Tomlins-2006-v1	D18	۱۰۴	۲۳۱۵	۵
Tomlins-2006-v2	D19	۹۲	۱۲۸۸	۴
West-2001	D20	۴۹	۱۱۹۸	۲



(شکل-۲): تأثیر تعداد متفاوت زوج محدودیّت‌ها

(Figure-2): The effect of the different number of pairwise constraints

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]

[DOI: 10.61186/jsdp.20.1.39]

۴-۱- تأثیر تعداد متفاوت زوج محدودیّت‌ها

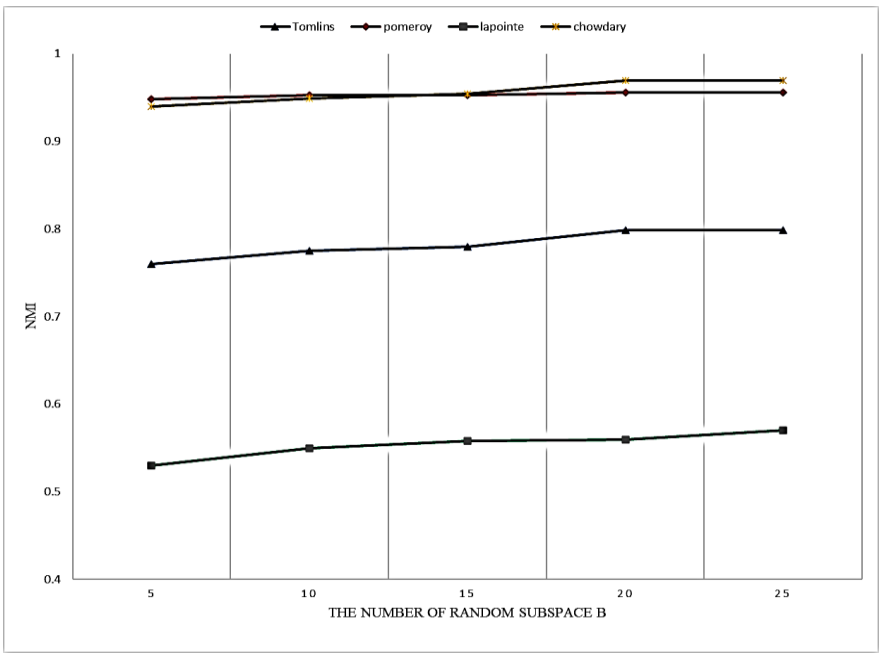
شکل (۲) تأثیر تعداد متفاوت زوج محدودیّت‌ها را بر عملکرد روش پیشنهادی بر روی ۸ مجموعه دادهٔ سرطان با استفاده از معیار سنجش NMI نشان می‌دهد. منحنی-های مقادیر معیار NMI نمایش‌داده‌شده در شکل (۲) را بیان می‌کنند که همان‌گونه که انتظار داریم، با افزایش تعداد زوج محدودیّت‌ها، کارایی روش پیشنهادی نیز افزایش می‌یابد؛ به‌عنوان مثال، در مجموعه‌داده Khan-2001 وقتی‌که تعداد زوج محدودیّت‌ها از ۲۰ تا ۱۰۰ افزایش می‌یابد، مقدار NMI نیز از ۰/۶ تا ۰/۹ افزایش می-یابد. همچنین، در مجموعه‌داده West-2001 مقدار NMI از ۰/۳۱ به ۰/۷۶ افزایش می‌یابد. این نشان می‌دهد که روش پیشنهادی به تعداد زوج محدودیّت‌ها حساس است؛ که یک نتیجه‌گیری منطقی است و دور از انتظار نیست.

تعداد بیشتر زوج محدودیّت‌ها به معنای در دسترس بودن اطّلاعات بیشتر است، که به بهبود عملکرد الگوریتم خوشه‌بندی کمک می‌کند. قابل ذکر است که زوج محدودیّت‌ها به‌طور معمول، به کمک دانش پیشین کارشناسان به دست می‌آید و اغلب برای به دست آوردن این دانش هزینهٔ بالایی نیز متحمّل می‌شوند. در نتیجه، باید بین عملکرد الگوریتم و هزینه برای به دست آوردن دانش پیشین یک مصالحه انجام شود. بنابراین، ما در آزمایش‌های خود تعداد زوج محدودیّت‌ها را حداقل ۱۰۰ در نظر می‌گیریم تا یک تعامل بین عملکرد/ هزینه داشته باشیم.

۴-۲- تأثیر تعداد زیرفضاهای تصادفی

برای بررسی تأثیر تعداد زیرفضاهای تصادفی بر روی عملکرد روش پیشنهادی، آزمایش‌های بی‌شماری با تعداد متفاوت شاخص **B** (تعداد زیرفضاها) از ۵ تا ۲۵ بر روی چهار مجموعه‌داده که به‌طور تصادفی انتخاب شدند، انجام شد.

همان‌گونه‌که در شکل (۳) مشاهده می‌کنید، افزایش تعداد زیرفضاهای تصادفی، تأثیر مثبتی روی عملکرد خوشه-بندی باتوجّه‌به مقدار NMI دارد. نکتهٔ قابل‌تأمّل این است که در بیشتر مجموعه‌داده‌ها، زمانی که مقدار **B** به عدد ۲۰ می‌رسد، تغییر چندانی در عملکرد روش پیشنهادی حاصل نمی‌شود، این بدین معنی است که الگو به یک پایداری و ثبات رسیده است.



(شکل-۳): تأثیر تعداد زیرفضاهای تصادفی
(Figure-3): The effect of the number of random subspaces

به‌عنوان مثال، در مجموعه‌داده Tomlins-2006-v1 زمانی که مقدار B برابر ۵ است، مقدار NMI به‌دست‌آمده برابر با ۰/۷۷ است و با افزایش مقدار B، همان‌طورکه انتظار می-رود، مقدار NMI افزایش پیدا می‌کند و با تعداد ۲۰ زیرفضا، مقدار NMI برابر با ۰/۷۸۵ و ثابت باقی می‌ماند.

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

بنابراین، تنظیم تعداد زیرفضاها روی عدد ۲۰، یک مقدار مناسب برای دستیابی به عملکرد بهتر است.

۴-۳-تأثیر نرخ تولید زیرفضاهای تصادفی

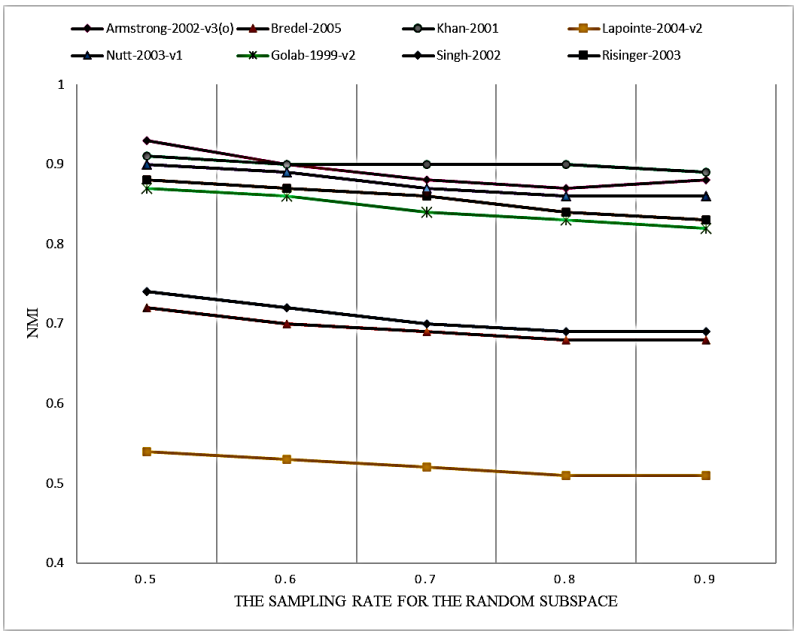
شکل(۴) تأثیر مقدار نرخ تولید زیرفضاهای تصادفی باتوجّه‌به مقدار NMI را نشان می‌دهد. روش پیشنهادی در بیشتر مجموعه‌داده‌های سرطانی، مانند مجموعه‌داده‌های Armstrong-2002-v3(o)، Bredel-2005، Khan-2001 و Singh-2002، زمانی که نرخ تولید زیرفضاهای تصادفی ۰/۵ است، بهترین عملکرد را دارد و با انحراف نرخ از مقدار ۰/۵، میزان رضایت‌بخشی نتایج حاصل کمتر می‌شود. یکی از دلایل احتمالی این است که افزایش نرخ تولید زیرفضاهای تصادفی، باعث ایجاد نمونه داده‌های نوفه‌دار و زاید بیشتری در زیرفضا می‌شود که بر عملکرد روش پیشنهادی تأثیر منفی می‌گذارد.

۴-۴-مقایسه روش پیشنهادی با روش مرجع [22]

در این زیر بخش به تحلیل و مقایسهٔ روش پیشنهادی با یکی از روش‌های خوشه‌بندی مشابه می‌پردازیم. نتایج عملکرد روش خوشه‌بندی ایدهٔ پیشنهادی نسبت به روش ارائه‌شده در مرجع [22] با استفاده از معیار سنجش ARI، در جدول (۲) نشان داده شده است. با بررسی و مقایسهٔ نتایج به‌دست‌آمده، باتوجّه‌به معیار سنجش ARI، روش پیشنهادی در ۱۲ مجموعه‌داده از ۲۰ مجموعه‌داده از روش مرجع [22] پیشی گرفته و عملکرد بهتری داشته است. به‌عنوان مثال، در مجموعه‌داده Pomeroy-2002-v2 مقدار ARI برای روش پیشنهادی، به میزان ۰/۰۳۵۷ بیشتر از روش مرجع [22] به دست آمده است.

به‌عنوان مثالی دیگر، در مجموعه دادهٔ Khan-2001 مقدار ARI روش پیشنهادی به میزان ۰/۰۶۰۷ بیشتر از روش مرجع [22] است.

برای تحلیل بهتر نتایج و مقایسهٔ روش پیشنهادی، عملکرد روش پیشنهادی و روش مرجع [22] در شکل (۵) با استفاده از نمودار جعبه‌ای رسم شده است. نمودار جعبه-ای، نموداری برای توصیف تغییرات، توزیع و پراکندگی داده‌ها بر اساس چندین شاخص مرکزی و پراکندگی آماری است. از این نمودار می‌توان برای مقایسه ویژگی-های چند روش مختلف بر روی مجموعه‌داده‌های یکسان به‌طور هم‌زمان استفاده کرد.



(شکل-۴):تأثیر نرخ تولید زیرفضاهای تصادفی

(Figure-4): The effect of the production rate of random subspaces

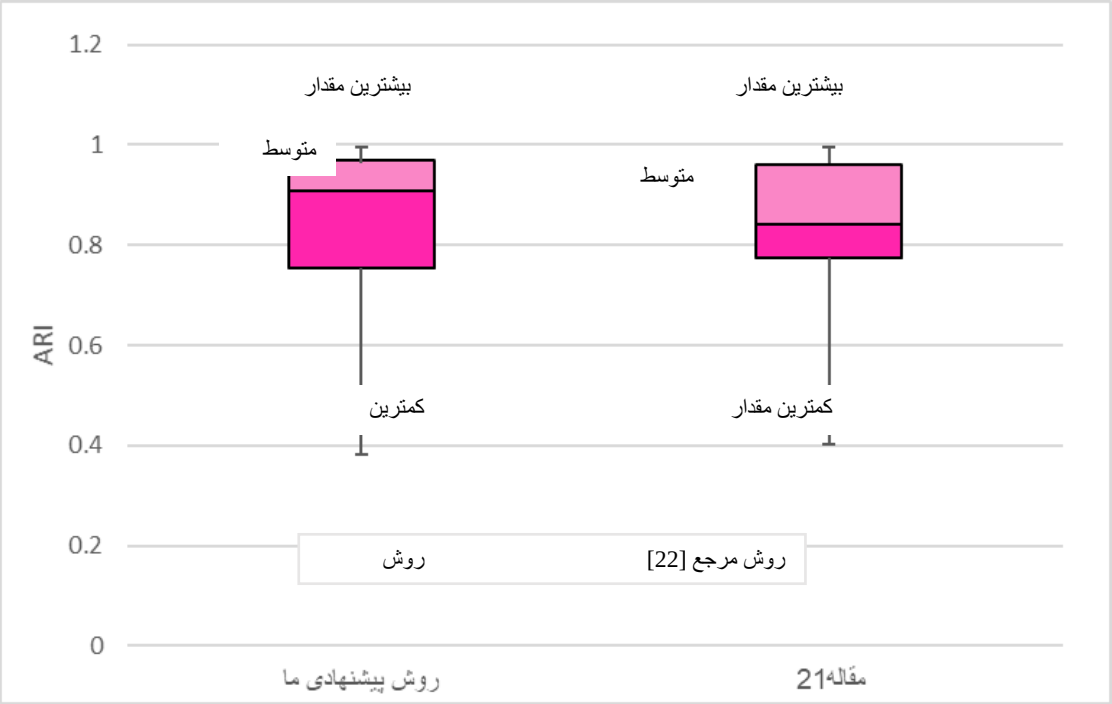
[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]

[DOI: 10.61186/jsdp.20.1.39]

(جدول-۲): مقایسهٔ عملکرد روش پیشنهادی با روش مرجع [22] با استفاده از معیار ARI

(Table-2): The performance of our proposed method in comparison to the reference [22] using ARI criterion

مجموعه‌داده	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
روش پیشنهادی	۰/۹۹۱۷	۰/۹۶۸۳	۰/۸۲۵۵	۰/۹۰۸۸	۱/۰۰۰۰	۱/۰۰۰۰	۰/۷۵۴۹	۰/۹۳۱۵	۰/۹۵۳۹	۰/۷۱۹۱
روش مرجع [22]	۰/۹۸۵۲	۰/۹۲۳۱	۰/۸۲۵۷	۰/۸۴۲۶	۰/۹۸۲۵	۱/۰۰۰۰	۰/۷۷۳۷	۰/۹۰۳۲	۰/۸۹۳۲	۰/۶۷۳۱
مجموعه‌داده	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
روش پیشنهادی	۰/۶۳۷۹	۰/۹۹۰۱	۰/۹۲۱۵	۰/۳۸۱۲	۰/۹۹۷۴	۰/۸۲۸۸	۰/۸۸۷۲	۰/۷۷۸۰	۰/۵۰۹۹	۰/۹۴۱۹
روش مرجع [22]	۰/۵۸۷۴	۰/۹۹۷۰	۰/۸۸۵۸	۰/۴۰۱۴	۰/۹۶۰۴	۰/۷۹۲۴	۰/۸۲۷۱	۰/۸۰۷۸	۰/۵۲۶۹	۰/۹۵۹۶



(شکل- ۵): نمودارهای جعبه‌ای مقایسهٔ روش پیشنهادی با روش مرجع [۲۲] با استفاده از معیار ARI

(Figure-5): Box plot comparison of the proposed method with reference [22] using ARI criterion

نمودارهای جعبه‌ای خلاصهٔ پنج ویژگی عددی از مجموعه‌داده‌ها را نشان می‌دهد. این مقادیر شامل کمترین مقدار[۱]، یک چهارم اول[۲]، متوسط[۳]، یک چهارم سوم[۴] و بیشترین مقدار[۵] است. همان‌طورکه از نمودار جعبه‌ای شکل (۵) مشخص است، مقدار متوسط در روش پیشنهادی بالاتر از مقدار متوسط روش مرجع [22] قرار گرفته است. همچنین، پراکندگی میانگین برای روش پیشنهادی بیشتر است، زیرا ارتفاع جعبهٔ آن بزرگتر از ارتفاع جعبه روش مرجع [22] است. یکی از دلایل احتمالی این افزایش مقدار، می‌تواند انتخاب فرایند خوشه-بندی لاپلاسی-p باشد که باعث خوشه‌بندی متعادل‌تر و ایجاد زیرفضاهای بهینه می‌شود.

[۱] minimum score
[۲] first (lower) quartile
[۳] median
[۴] third (upper) quartile
[۵] maximum score

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]

۴-۵- مقایسه روش پیشنهادی با تعدادی از روش‌های یادگیری متریک فاصله جهت خوشه‌بندی

از آنجاکه روش‌های یادگیری متریک فاصله بر عملکرد روش‌های خوشه‌بندی تأثیر بسیار مثبتی دارد، در این زیربخش روش پیشنهادی را با چندین روش یادگیری متریک فاصله مقایسه می‌کنیم. روش‌های یادگیری متریک موردمقایسه عبارتند از روش تجزیه و تحلیل مؤلّفه‌های مرتبط[1](RCA) [29]، روش یادگیری متریک اطّلاعات نظری[2](ITML) [30]، روش تجزیه و تحلیل مؤلّفه‌های تمییزکننده[3] (DCA) [31]. همچنین، نتایج روش ارائه‌شده در مرجع [22] نیز برای بررسی و مقایسه ارائه شده‌اند. جدول (۳) نتایج روش پیشنهادی در مقایسه با سایر روش‌های ذکر شده را نمایش می‌دهد. بهترین مقادیر در جدول با قلم پررنگ و دومین بهترین مقدار به‌صورت زیرخط مشخص شده‌اند.

همان‌طورکه در جدول (۳) مشاهده می‌کنید، روش پیشنهادی در ۱۱ مجموعه‌داده از ۲۰ مجموعه‌داده، بهترین نتایج را در مقایسه با سایر روش‌های مقایسه‌شده به دست آورده است. همچنین، در شش مجموعه‌داده از ۲۰ مجموعه‌داده، دومین مجموعه، بهترین نتیجه را به دست آورده است؛ به‌عنوان مثال، در مجموعه‌داده Bredel-2005 مقدار به‌دست‌آمده در روش پیشنهادی ۰/۱۰۴۲ بیشتر از مرجع [22] و مقدار ۰/۱۸۴۶ بیشتر از RCA و مقدار ۰/۴ بیشتر از ITML و همچنین، مقدار ۰/۴۶۸ بیشتر از DCA است. همچنین، در مجموعه‌داده Garber-2001 مقدار به‌دست‌آمده در روش پیشنهادی ۰/۱۱۹۸ بیشتر از مرجع [22] و مقدار ۰/۰۲۶۵ کمتر از RCA و مقدار ۰/۲۹۱۶ بیشتر از ITML و همچنین، مقدار ۰/۳۷۰۲ بیشتر از DCA است. شکل (۶)، نمودار میله‌ای مقایسه روش پیشنهادی با سایر روش‌های یادگیری متریک فاصله را نشان می‌دهد. یکی از دلایل بهتر شدن جواب‌های حاصل از روش پیشنهادی می‌تواند استفاده از فضای کاهش‌یافته در تجزیه و تحلیل داده‌ها مانند خوشه-بندی باشد، زیرا گاهی اوقات دقیق‌تر از فضای اصلی انجام می‌شود و همچنین، دلیل دیگر می‌تواند این موضوع باشد که روش پیشنهادی نتیجهٔ چندین خوشه‌بندی به‌دست‌آمده از زیرفضاهای مختلف تصادفی را در یک نتیجهٔ کلّی ادغام می‌کند؛ که این باعث بهبود عملکرد خوشه‌بندی نسبت به روش‌های یادگیری متریک است.

۵- نتیجه‌گیری

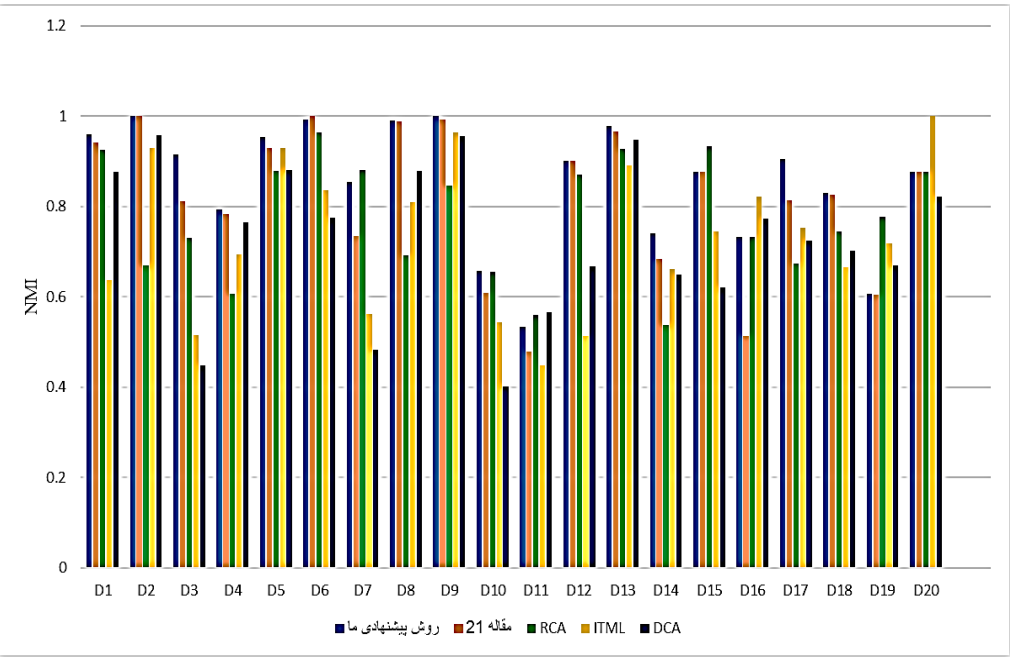
در این مقاله، یک چارچوب خوشه‌بندی گروهی بر مبنای خوشه‌بندی لاپلاسی-p برای خوشه‌بندی داده‌ها با ابعاد بالا، پیشنهاد شد. در مقایسه با رویکردهای خوشه‌بندی نیمه‌نظارتی سنّتی، روش پیشنهادی این مقاله، چند ویژگی مهم داشت. از جمله، از روش پیشنهادی برای چگونگی استفاده از تمام جفت محدودیّت‌های باید-متّصل و نباید-متّصل از خاصیّت تراگذری و انتشار جفت محدودیّت‌ها استفاده کرد. همچنین، در روش پیشنهادی، خوشه‌بندی گروهی با استفاده از تولید تعدادی زیرفضای تصادفی پیاده‌سازی شد، به این ترتیب که راه‌حل‌های هرکدام از خوشه‌بندها را تجمیع کرده و وظیفهٔ خوشه-بندی نهایی را با استفاده از اطّلاعات تجمیع‌شده به یک خوشه‌بند نهایی یک‌پارچه حاصل از زیرفضای بهینه واگذار کرد.

جهت افزایش عملکرد روش پیشنهادی از روش خوشه-بندی لاپلاسی-p در هر قسمت از عملیات استفاده شد. پس از انجام آزمایش‌های متعدد دریافتیم که بهره‌گیری از روش‌های خوشه‌بندی گروهی با کمک فاکتور اطمینان، توانایی الگوریتم پیشنهادی برای دستیابی به نتایج بهتر را بهبود می‌بخشد. همچنین، استفاده از عملگرهای تراگذری و انتخاب زیرفضاهای تصادفی با اندازهٔ نابرابر، نقش مهمّی در دستیابی به عملکرد بهتر برای الگوریتم پیشنهادی داشت. یکی دیگر از رویکردهای تأثیرگذار در عملکرد روش پیشنهادی استفاده از عملگرهای جستجو جهت یافتن بهترین زیرفضا بود که باعث دستیابی به نتایج بهتر شد. الگوریتم پیشنهادی در مواجهه با مجموعه‌داده‌های با ابعاد بسیار بالا، از بسیاری از پیشرفته‌ترین روش‌ها بهتر عمل کرد. و در نهایت، بهره‌گیری از روش خوشه‌بندی طیفی لاپلاسی-p ، حجم بهتر و متعادل‌تر و طبیعی‌تری از خوشه‌ها در مقایسه با نمونه استاندارد طیفی تولید کرد.

در پایان، قابل ذکر است که میزان اثربخشی زوج محدودیّت‌ها و نحوهٔ حذف محدودیّت‌های زائد را می‌توان به‌عنوان یکی از محدودیّت‌های روش پیشنهادی در نظر گرفت.

[1] Relevant Component Analysis
[2] Information Theoretic Metric Learning
[3] Discriminative Component Analysis

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]
[DOI: 10.61186/jsdp.20.1.39]



(شکل-۶): نمودار میله‌ای مقایسه روش پیشنهادی با سایر روش‌های یادگیری متریک فاصله

(Figure-6): Histogram diagram comparison of the proposed method with other distance metric learning methods

(جدول-۳): عملکرد روش پیشنهادی در مقایسه با چند روش خوشه‌بندی با استفاده از معیار سنجش NMI

(Table 3): Performance of our proposed method compared to several clustering methods using NMI measurement criteria

مجموعه‌داده	روش پیشنهادی	روش مرجع[22]	RCA	ITML	DCA
D1	۰/۹۶۰۱	۰/۹۴۱۸	۰/۹۲۵۴	۰/۶۳۶۱	۰/۸۷۶۶
D2	۱/۰۰۰۰	۱/۰۰۰۰	۰/۶۶۹۶	۰/۹۲۹۳	۰/۹۵۷۲
D3	۰/۹۱۵۴	۰/۸۱۱۲	۰/۷۳۰۸	۰/۵۱۵۴	۰/۴۴۷۴
D4	۰/۷۹۳۲	۰/۷۸۲۹	۰/۶۰۶۹	۰/۶۹۴۸	۰/۷۶۵۴
D5	۰/۹۵۳۴	۰/۹۲۹۸	۰/۸۷۸۵	۰/۹۲۹۸	۰/۸۷۹۶
D6	۰/۹۹۱۵	۱/۰۰۰۰	۰/۹۶۳۸	۰/۸۳۵۰	۰/۷۷۴۴
D7	۰/۸۵۳۶	۰/۷۳۳۸	۰/۸۸۰۱	۰/۵۶۲۰	۰/۴۸۳۴
D8	۰/۹۹۰۰	۰/۹۸۸۸	۰/۶۹۲۰	۰/۸۰۹۴	۰/۸۷۹۱
D9	۱/۰۰۰۰	۰/۹۹۲۲	۰/۸۴۵۵	۰/۹۶۴۸	۰/۹۵۵۳
D10	۰/۶۵۷۱	۰/۶۰۷۶	۰/۶۵۶۲	۰/۵۴۴۲	۰/۴۰۰۵
D11	۰/۵۳۳۳	۰/۴۷۸۸	۰/۵۶۰۵	۰/۴۴۷۴	۰/۵۶۵۷
D12	۰/۹۰۰۰	۰/۹۰۰۳	۰/۸۷۰۹	۰/۵۱۳۸	۰/۶۶۶۶
D13	۰/۹۷۸۹	۰/۹۶۶۱	۰/۹۲۸۲	۰/۸۹۱۷	۰/۹۴۷۲
D14	۰/۷۴۱۵	۰/۶۸۳۶	۰/۵۳۷۹	۰/۶۶۱۵	۰/۶۴۹۸
D15	۰/۸۷۶۹	۰/۸۷۶۶	۰/۹۳۲۹	۰/۷۴۵۴	۰/۶۲۰۹
D16	۰/۷۳۱۶	۰/۵۱۳۲	۰/۷۳۳۳	۰/۸۲۲۷	۰/۷۲۲۰
D17	۰/۹۰۵۹	۰/۸۱۳۸	۰/۶۷۴۰	۰/۷۵۲۳	۰/۷۲۵۲
D18	۰/۸۲۹۹	۰/۸۲۵۴	۰/۷۴۴۱	۰/۶۶۶۴	۰/۷۰۲۷
D19	۰/۷۵۸۴	۰/۶۰۴۴	۰/۷۷۷۵	۰/۷۱۷۵	۰/۶۶۹۸
D20	۰/۸۷۶۵	۰/۸۷۶۲	۰/۸۷۶۲	۱/۰۰۰۰	۰/۸۲۲۱

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]

[DOI: 10.61186/jsdp.20.1.39]

6- Refrence ۶- مراجع

[1] C. Chrysouli and A. Tefas, "Spectral clustering and semi-supervised learning using evolving similarity graphs," Applied Soft Computing, vol. 34, pp. 625-637, 2015.

[2] W. Hu, C. Chen, F. Ye, Z. Zheng, and G. Ling, "Nonnegative Spectral Clustering for Large-Scale Semi-supervised Learning," in International Conference on Database Systems for Advanced Applications, 2019: Springer, pp. 287-291.

[3] E. Hancer, B. Xue, and M. Zhang, "A survey on feature selection approaches for clustering," Artificial Intelligence Review, pp. 1-27, 2020.

[4] G. Chao, S. Sun, and J. Bi, "A survey on multi-view clustering," arXiv preprint arXiv:1712.06246, 2017.

[5] X. He, S. Zhang, and Y. Liu, "An adaptive spectral clustering algorithm based on the importance of shared nearest neighbors," Algorithms, vol. 8, no. 2, pp. 177-189, 2015.

[6] H. Jia, S. Ding, H. Zhu, F. Wu, and L. Bao, "A Feature Weighted Spectral Clustering Algorithm Based on Knowledge Entropy," JSW, vol. 8, no. 5, pp. 1101-1108, 2013.

[7] Z. Yu et al., "Probabilistic cluster structure ensemble," Information Sciences, vol. 267, pp. 16-34, 2014.

[8] J.E. Van Engelen and H.H. Hoos, "A survey on semi-supervised learning," Machine Learning, vol. 109, no. 2, pp. 373-440, 2020.

[9] S. Ding, B. Qi, H. Jia, H. Zhu, and L. Zhang, "Research of semi-supervised spectral clustering based on constraints expansion," Neural Computing and Applications, vol. 22, no. 1, pp. 405-410, 2013.

[10] Y. Jia, S. Kwong, and J. Hou, "Semi-supervised spectral clustering with structured sparsity regularization," IEEE Signal Processing Letters, vol. 25, no. 3, pp. 403-407, 2018.

[11] Y. Jia, S. Kwong, J. Hou, and W. Wu, "Semi-supervised non-negative matrix factorization with dissimilarity and similarity regularization," IEEE Transactions on Neural Networks and Learning Systems, 2019.

[12] M.S. Baghshah, F. Afsari, S. B. Shouraki, and E. Eslami, "Scalable semi-supervised clustering by spectral kernel learning," Pattern Recognition Letters, vol. 45, pp. 161-171, 2014.

[13] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised feature selection methods," Pattern Recognition, vol. 64, pp. 141-158, 2017.

[14] S. Faußer and F. Schwenker, "Semi-supervised clustering of large data sets with kernel methods," Pattern recognition letters, vol. 37, pp. 78-84, 2014.

[15] M. Sugiyama, G. Niu, M. Yamada, M. Kimura, and H. Hachiya, "Information-maximization clustering based on squared-loss mutual information," Neural Computation, vol. 26, no. 1, pp. 84-131, 2014.

[16] T. Bühler and M. Hein, "Spectral clustering based on the graph p-Laplacian," in Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 81-88.

[17] J. Jost, R. Mulas, and D. Zhang, "p-Laplace Operators for Chemical Hypergraphs," arXiv preprint arXiv:2007.00325, 2020.

[18] S. Saito, D. P. Mandic, and H. Suzuki, "Hypergraph p-Laplacian: A Differential Geometry View," arXiv preprint arXiv:1711.08171, 2017.

[19] S. Ding, H. Jia, M. Du, and Q. Hu, "p-Spectral Clustering Based on Neighborhood Attribute Granulation," in International Conference on Intelligent Information Processing, 2016: Springer, pp. 50-58.

[20] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," Frontiers of Computer Science, pp. 1-18, 2020.

[21] H. Niu, N. Khozouie, H. Parvin, H. Alinejad-Rokny, A. Beheshti, and M. R. Mahmoudi, "An Ensemble of Locally Reliable Cluster Solutions," Applied Sciences, vol. 10, no. 5, p. 1891, 2020.

[22] Z. Yu, Z. Kuang, J. Liu, H. Chen, J. Zhang, J. You, H.S. Wong and G. Han, "Adaptive ensembling of semi-supervised clustering solutions. IEEE Transactions on Knowledge and Data Engineering", vol. 29, no. 8, pp. 1577-1590, 2017.

[23] Z. Yu, P. Luo, J. You, H.S. Wong, H. Leung, S. Wu, J. Zhang, and G. Han, "Incremental semi-supervised clustering ensemble for high dimensional data clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 701-714, 2015.

[24] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," Pattern recognition, vol. 46, no. 12, pp. 3460-3471, 2013.

[25] S. Safari, and F. Afsari. "Ensemble P-spectral Semi-supervised Clustering." In 2020 International Conference on Machine Vision and Image Processing (MVIP), pp. 1-5. IEEE, 2020.

[26] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," BMC bioinformatics, vol. 9, no. 1, p. 497, 2008.

[27] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]

[DOI: 10.61186/jsdp.20.1.39]

clusterings comparison: Variants, properties, normalization and correction for chance," The Journal of Machine Learning Research, vol. 11, pp. 2837-2854, 2010.

[28] L. Hubert and P. Arabie, "Comparing partitions," Journal of classification, vol. 2, no. 1, pp. 193-218, 1985.

[29] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a mahalanobis metric from equivalence constraints," Journal of Machine Learning Research, vol. 6, no. Jun, pp. 937-965, 2005.

[30] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in Proceedings of the 24th international conference on Machine learning, 2007, pp. 209-216.

[31] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, "Learning distance metrics with contextual constraints for image retrieval," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, vol. 2: IEEE, pp. 2072-2078.

صدیقه صفری، دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیک از دانشگاه شهید باهنر کرمان است. زمینه‌های پژوهشی وی طرّاحی الگو و یادگیری ماشین است. نشانی رایانامه ایشان عبارت است از:

s.safari@eng.uk.ac.ir

فاطمه افسری، استادیار بخش مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه شهید باهنر کرمان است. ایشان دانش‌آموختهٔ رشتهٔ مهندسی کامپیوتر از دانشگاه شیراز و سپس، دانشگاه شهید باهنر کرمان است، و از سال ۱۳۸۳ فعالیت تمام وقت به‌عنوان عضو هیأت علمی در دانشگاه شهید باهنر کرمان را آغاز کرد. زمینه‌های پژوهشی وی یادگیری ماشین، شناسایی الگو، یادگیری عمیق و پردازش تصاویر پزشکی هستند و در حوزه‌های مرتبط، مقالات زیادی در کنفرانس‌های معتبر داخلی و خارجی و همچنین، در مجلات معتبر بین-المللی چاپ کرده است. نشانی رایانامه ایشان عبارت است از:

afsari@uk.ac.ir

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]

[DOI: 10.61186/jsdp.20.1.39]

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-07]

[DOI: 10.61186/jsdp.20.1.39]