



چکیده

با توجه به افزایش روزافزون اطلاعات و تحلیل دقیق آنها مسأله خوشه‌بندی که برای آشکارسازی الگوهای پنهان موجود در داده‌ها استفاده می‌شود، همچنان از اهمیت بالایی برخوردار است؛ و از طرفی دیگر، خوشه‌بندی داده‌های با ابعاد بالا با استفاده از روش‌های سنتی پیشین محدودیت‌های زیادی دارد. در مقاله حاضر، یک روش خوشه‌بندی گروهی نیمه‌نظارتی برای مجموعه‌ای از داده‌های پزشکی با ابعاد بالا پیشنهاد می‌شود. در فرموله‌سازی مسأله خوشه‌بندی اطلاعات نظارتی اندکی به‌عنوان دانش پیشین با استفاده از اطلاعات مربوط به تشابه و یا عدم تشابه (به‌صورت تعدادی زوج محدودیت‌های دوه‌دو) در نظر گرفته می‌شود. نخست، با استفاده از خاصیت تراگذری، زوج محدودیت‌های دوه‌دو 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 تجميع 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 نتیجه خوشه‌بندی دارد، می‌یابیم. نتایج آزمایش‌های متعدد بر روی چندین داده پزشکی با ابعاد بالا نشان می‌دهد که رویکرد پیشنهادی، عملکرد و کارایی بهتری نسبت به روش‌های پیشین دارد.

واژگان کلیدی: خوشه‌بندی، یادگیری زیرفضا، یادگیری گروهی، یادگیری نیمه‌نظارتی، زوج محدودیت‌های دوه‌دو

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

* نویسنده عهده‌دار مکاتبات



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

روش‌های یادگیری بدون نظارت داشته باشند؛ از طرفی دیگر، این روش‌ها در مقایسه با روش‌های یادگیری بانظارت، زمان و هزینه کمتری جهت برچسب‌دارکردن داده‌ها صرف می‌کنند. هزینه مربوط به فرایند برچسب زدن داده‌ها ممکن است تأثیر منفی زیادی بر یک الگوی آموزشی به‌طور کامل برچسب‌دار (روش یادگیری بانظارت) بگذارد که در عمل الگو بی‌استفاده شود. از طرف دیگر، فراهم کردن اطلاعات 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 کلی زیر است:

۱- مقدمه

یادگیری نیمه‌نظارتی دسته‌ای از روش‌های یادگیری ماشین است که در آن از داده‌های بدون برچسب و داده‌های برچسب‌دار به‌صورت هم‌زمان برای بهبود دقت یادگیری استفاده می‌شود [1]. یادگیری نیمه‌نظارتی میان یادگیری بدون نظارت (بدون هیچ‌گونه اطلاعات آموزشی برچسب‌گذاری‌شده) و یادگیری بانظارت (با داده‌های آموزشی به‌طور کامل برچسب‌گذاری‌شده) قرار می‌گیرد. بسیاری از پژوهش‌گران یادگیری ماشین دریافتند که استفاده از تعداد اندکی داده برچسب‌دار در کنار داده‌های بدون برچسب، می‌تواند بهبود قابل‌توجهی در دقت

- 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 محاسبه فاصله 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 حوزه داده‌کاوی است که موردتوجه و بررسی پژوهشگران بسیاری قرار گرفته است. از دهه ۱۹۵۰ میلادی به بعد مطالعات بسیاری در مورد خوشه‌بندی صورت گرفته است [2-7]. در دو دهه اخیر بحث خوشه‌بندی برای تشخیص خوشه‌هایی با ساختارهای متنوع مطرح شده و روش‌های خوشه‌بندی به‌نسبت زیادی نیز معرفی شده است که هرکدام نقاط قوت و ضعف ویژه خود را دارند. اغلب از خوشه‌بندی برای تقسیم داده‌های یک مجموعه بر مبنای 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 اشیاء است، به‌طوری‌که اشیاء مشابه در یک گروه و اشیاء متفاوت در گروه‌های مختلف قرار بگیرند. از آنجاکه در خوشه‌بندی از اطلاعات برچسب داده‌ها استفاده نمی‌شود، خوشه‌بندی یکی از روش‌های یادگیری بدون‌نظارت محسوب می‌شود. در مقابل آن، روش یادگیری بانظارت است که نیازمند اطلاعات برچسب داده‌ها است. اما یادگیری نیمه‌نظارتی یک روش بسیار کاربردی است که بین این دو رویکرد قرار می‌گیرد. در خوشه‌بندی نیمه‌نظارتی اطلاعات اضافه‌ای به‌صورت تعداد کمی داده برچسب‌خورده، یا تعدادی زوج محدودیت‌های دوه‌دو به‌صورت محدودیت‌های باید-متصل و نباید-متصل داده شده است. نتیجه خوشه‌بندی نیمه‌نظارتی متأثر از محدودیت‌ها تغییر می‌کند. همچنین، خوشه‌بندی نیمه‌نظارتی مبتنی بر شباهت است که اندازه‌گیری شباهت بر اساس محدودیت‌ها تغییر می‌کند. بسیاری از رویکردهای خوشه‌بندی نیمه‌نظارتی جدید، گسترش الگوریتم‌های خوشه‌بندی سنتی با توجه به اطلاعات برچسب یا زوج محدودیت‌ها هستند [8-15].

کاهش ابعاد داده‌ها یکی دیگر از مسائل رایج در حوزه آمار، یادگیری ماشین و نظریه اطلاعات است. کاهش ابعاد داده، فرایند کاهش تعداد متغیرهای تصادفی موردبررسی و استخراج مجموعه‌ای از متغیرهای 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 مسأله نفرین بُعد، پیش‌از به‌کارگیری الگوریتم‌های رگرسیون یا طبقه‌بندی همچون طبقه‌بند K-نزدیک‌ترین همسایگان (K-NN) انجام می‌شود.

۲- پیشینه پژوهش

به‌تازگی، گروه‌های زیادی به یادگیری گروهی در مسائل خوشه‌بندی نیمه‌نظارتی کرده‌اند [19-23]. یادگیری گروهی یک الگوار (پارادایم) یادگیری ماشین است که در آن چندین فراگیر برای حل یک مسأله آموزش داده می‌شوند. روش‌های یادگیری ماشین متداول سعی می‌کنند یک فرضیه از داده‌های آموزشی را یاد بگیرند، اما روش‌های گروهی سعی می‌کنند این روش‌ها را بهبود دهند. باتوجه به قابلیت 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^۱ (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^۲ مبتنی بر گراف لاپلاسی- p ، داده‌ها به‌طور مستقل خوشه‌بندی می‌شوند. نتایج هرکدام از خوشه‌بندها با زوج محدودیت‌های دوه‌دو تطبیق داده می‌شوند و بر اساس میزان تطابق یک درجه 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. از آنجاکه تشکیل زیرفضاها به‌صورت تصادفی صورت گرفته است، در ادامه

^۱ Curse of dimension

^۲ Random Subspace Based Semi-Supervised

Cluster Ensemble

^۳ Spectral clustering

یک الگوریتم جستجو برای یافتن بهترین زیرفضا طراحی شده است. الگوریتم جستجو بدین گونه عمل می کند که در بین تعدادی زیرمجموعه شامل زیرفضاها، که از این به بعد هر زیرمجموعه را یک محیط می نامیم، به دنبال بهترین محیط می گردد. بهترین محیط، محیطی است که بهترین نتایج خوشه بندی را داشته باشد. برای پیاده سازی الگوریتم جستجو از تعدادی عملگر جستجو بهره گرفته ایم که هر کدام به گونه ای در کل فضای جستجو، یا در یک فضای محلی سعی در یافتن بهترین محیط ممکن دارد.

در ادامه مقاله، در بخش دوم پژوهش های پیشین انجام شده را مرور می کنیم. در بخش سوم، روش پیشنهادی برای خوشه بندی نیمه نظارتی گروهی برای داده های با ابعاد بالا به تفصیل معرفی می شود. در بخش چهارم، نتایج شبیه سازی روش پیشنهادی مورد بررسی و تحلیل قرار می گیرد و در نهایت، نتیجه به دست آمده جمع بندی می شود.

رشد روزافزون داده ها در سال های اخیر، مسایل و نیازهای متفاوت و متعددی را در زمینه داده کاوی و یادگیری ماشین ایجاد کرده است. یکی از مهم ترین این مسایل انتخاب ویژگی های مرتبط به مجموعه اصلی ویژگی های موجود است که عملکرد یادگیری را نسبت به عملکرد مجموعه اصلی ویژگی ها به میزان بیشتری بهبود می بخشد. خوشه بندی گراف، زیر مجموعه ای از تجزیه و تحلیل خوشه ای است که به دنبال گروه هایی از رأس های مرتبط در یک گراف است. به دلیل کاربرد گسترده آن چندین الگوریتم خوشه بندی گراف در سال های گذشته ارائه شده است. یکی از معروف ترین الگوریتم های خوشه بندی گراف، الگوریتم های خوشه بندی طیفی هستند که بیشتر بر اساس تجزیه ویژه ماتریس های لاپلاسی گراف های وزن دار یا گراف های بی وزن طراحی شده اند. در سال های اخیر مطالعات مختلفی در زمینه خوشه بندی انجام شده است که در ادامه، تعدادی از آنها به طور مختصر آورده شده است. چریسلی و همکارش [1] به منظور بهینه سازی ساختار یک گراف و دستیابی به نتایج خوشه بندی بهتر یک روش خوشه بندی گراف طیفی را بررسی کردند که از الگوریتم ژنتیک استفاده کرده بود. از آنجاکه عملکرد الگوریتم آنها تا حد زیادی به نحوه ایجاد جمعیت اولیه بستگی داشت، بنابراین، نتایج رضایت بخش نبودند. در الگوی دیگر هو و همکاران [2] یک رویکرد خوشه بندی جدید به نام خوشه بندی طیفی غیرمنفی

انعطاف پذیر¹ (SNSC) معرفی کردند. در واقع SNSC، ویژگی غیرمنفی اصلی یک ماتریس شاخص را حفظ می کند، که با یک راه حل دقیق به یک مسئله بهینه سازی با انعطاف پذیری بیشتری منجر شده است. همچنین، هنسر و همکاران [3] در کار پژوهشی خود به بررسی جامع رویکردهای انتخاب ویژگی ها برای خوشه بندی باتوجه به حذف مزایا و ضررهای رویکردهای فعلی، از دیدگاه های متفاوت پرداختند. چائو و همکاران [4] راهبردهای مشترک برای ترکیب چندین دیدگاه از داده ها را مطالعه کردند و طبقه بندی جدیدی را از رویکردهای خوشه بندی چندمنظوره² (MVC) پیشنهاد دادند. به علاوه هی و همکارانش [5] یک روش اندازه گیری شباهت هسته گاوسی تطبیقی را از طریق کمیت میزان اهمیت برای هر رأس از گراف شباهت و هسته گاوسی مقیاس بندی شده، ارائه کردند. سپس، آنها الگوریتمی برای خوشه بندی طیفی تطبیقی بر اساس اهمیت نزدیک ترین همسایگان را پیشنهاد دادند. جیا و همکارانش [6] یک الگوریتم خوشه بندی طیفی ویژه را بر اساس آنتروپی دانش ارائه دادند. قابل ذکر است که به مفهوم آنتروپی در مجموعه ویژگی های بزرگ، برای ارزیابی اهمیت هر ویژگی در الگوریتم آنها توجه شده است. یو و همکاران [7] در پژوهش خود یک چارچوب گروهی برای استخراج ساختار خوشه ها از داده با رویکرد احتمالاتی را (با عنوان GMMSE) طراحی کردند که ساختار خوشه را از مجموعه داده ها شناسایی می کند. نتایج تجربی مطالعه آنها نشان داد که GMMSE روی مجموعه داده های ترکیبی و مجموعه داده های واقعی مخزن UCI کارکرد مناسبی دارد. یادگیری نیمه نظارتی، شاخه ای از یادگیری ماشین است که مانند یادگیری بانظارت فقط از داده های برچسب خورده استفاده نمی کند، بلکه در مرحله آموزش از داده های بدون برچسب نیز برای انجام وظایف یادگیری مشخص بهره می برد. همچنین، یکی از رویکردهای یادگیری نیمه نظارتی استفاده از اطلاعات بسیار ناچیز در مورد رده داده هاست. در این رویکرد، هیچگونه داده برچسب داری مانند رویکرد معمول یادگیری نیمه نظارتی، در دسترس نیست. اما اطلاعات ناچیزی همچون یکسانی رده دو داده (معروف به زوج داده باید-متصل³) و یا تفاوت

¹ scalable nonnegative spectral clustering

² multi-view clustering

³ Must-link

رده دو داده (معروف به زوج داده-متصل^۱) در دسترس است. این اطلاعات، معروف به زوج محدودیت-های دوبه‌دو^۲ هستند که به‌عنوان یک قید سخت^۳ در تابع هدف مسئله یادگیری ماشین در نظر گرفته می‌شوند و باید ارضا شوند.

در سال‌های اخیر پژوهش‌ها در این زمینه، روندهای کلی مشاهده‌شده در یادگیری ماشین را دنبال کرده است. روش یادگیری که در خوشه‌بندی استفاده می‌شود، می‌تواند عملکرد خوشه‌بندی را بهبود بخشد. متأسفانه، 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 انگِلن 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 محدودیت‌ها^۴ (SSCCE) مطالعه کردند. این الگوریتم علاوه بر گسترش مجموعه محدودیت‌های شناخته‌شده، رابطه 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 سنتی را در حوزه یادگیری نیمه‌نظارتی گسترش دادند. آنها با کمک مقدار کمی از اطلاعات نظارتی یک ماتریس مورب غیربلوکی^۵ ایجاد، و از آن در نظم‌دادن به یک مجموعه داده کم‌بعد و انتقال آن استفاده کردند. آنها همچنین، در مطالعه دیگری [11]، یک الگوی فاکتورگیری ماتریس غیرمنفی

نیمه‌نظارتی^۶ با استفاده از الگوسازی اطلاعات برچسب‌دار پیشنهاد کردند. الگوی ارائه‌شده توسط آنها قادر به تولید الگوهای نمایش کم‌بعد، جهت تمایز عملکرد خوشه‌بندی بود. در یک کار پژوهشی دیگر، باغشا 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 روش‌های انتخاب ویژگی نیمه‌نظارتی بود. فایویر و شوانکر [14] روش خوشه‌بندی نیمه‌نظارتی را برای داده‌های حجیم که نخست، با زیرمجموعه کوچکی از داده‌ها و مجموعه داده‌های 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 محدودیت جهت اعمال محدودیت‌های مثبت (باید-متصل)، داده‌ها را با روش K-means هسته‌ای خوشه‌بندی کردند و بیان کردند که محدودیت‌های منفی (نباید-متصل) با ماتریس هسته ارتباط ندارند. سوگیاما و همکاران [15] رویکردی را با استفاده از بیشینه‌سازی اطلاعات برای مسئله خوشه‌بندی نیمه‌نظارتی مطالعه کردند. بر اساس این رویکرد، خوشه‌بندی با بیشینه‌سازی اطلاعات، مسئله انتخاب الگو را مدنظر قرار می‌دهد. آنها از تابع مربع اتلاف اطلاعات متقابل (SMI)^۷ برای بیشینه‌سازی اطلاعات استفاده کردند.

خوشه‌بندی طیفی، برای خوشه‌بندی داده‌ها از مسئله افراز گراف بهره می‌برد. برش چگیر^۸ یک معیار بهینه برای تقسیم‌بندی نمودار است. برای ربه کمترین حد رساندن تابع هدف برش چگیر، به تجزیه مجازی ماتریس لاپلاسی- p ^۹ نیاز است. بوهر و هین [16] یک نسخه کلی از خوشه‌بندی طیفی را با استفاده از گراف لاپلاسی- p ، به‌عنوان تعمیمی غیرخطی از گراف لاپلاسی استاندارد ارائه دادند. آنها ارتباطی بین برش چگیر و دومین بردار ویژه ماتریس لاپلاسی- p یافتند. پژوهشگران زیادی از ماتریس لاپلاسی- p در پژوهش‌ها استفاده کردند. جوست و همکاران [17] ماتریس لاپلاسی- p را برای گراف‌ها و

^۶ nonnegative matrix factorization

^۷ Squared-loss Mutual Information

^۸ Cheeger cut

^۹ P-Laplacian

^۱ Cannot-link

^۲ Pairwise constraints

^۳ Hard constraint

^۴ semi-supervised spectral clustering based on constraints expansion

^۵ Anti-block-diagonal

همچنین، عملگر رأس لاپلاسی^۱ را برای تنظیم کلی 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 اجازه 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. از آنجاکه، 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 براینند این خوشه‌بندی‌های ضعیف است. دانگ و همکارانش [20] پیشرفت تحقق 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 مرحله نخست، زیرمجموعه‌های تصادفی از زیرفضاهای^۲ تولید می‌شود و در مرحله دوم، بر اساس خاصیت تراگذاری محدودیت‌های دوبره‌دو را گسترش داده، سپس، با در نظر گرفتن هریک از زیرمجموعه‌های تصادفی، یک خوشه‌بندی متفاوت باتوجه‌به روش‌های

تراگذاری متفاوت، ارائه می‌دهند. سپس، یک مرحله تطبیقی برای افزایش اطمینان هرکدام از الگوریتم‌های خوشه‌بندی اجرا می‌کنند که از یک فاکتور اطمینان طراحی شده استفاده می‌کند. همچنین، آنها در مطالعه‌ای دیگر [23]، اعضای گروه را به‌عنوان ویژگی‌ها در نظر گرفتند و نحوه 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 طبقه‌بندی 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^۳ (ISSCE) پیشنهاد می‌شود، که از یک فرایند جدید انتخاب عضو برای تولید یک زیرمجموعه بهینه از اعضا استفاده کرده است.

۳- روش پیشنهادی

در رویکرد پیشنهادی، نخست، زوج محدودیت‌های باید-متصل و نباید-متصل را روی کل داده‌ها انتشار می‌دهیم. 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 کل فضای ویژگی m -بعدی باشد، هر زیرفضا m' -بعدی خواهد بود. ما در کار پیشین [25]، جهت سادگی در تولید زیرفضاها ابعاد زیرفضاها را برابر در نظر گرفته بودیم، اما در پژوهش جاری این محدودیت 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 محدودیت‌های به‌روزرسانی‌شده،

^۳ incremental semi-supervised clustering ensemble

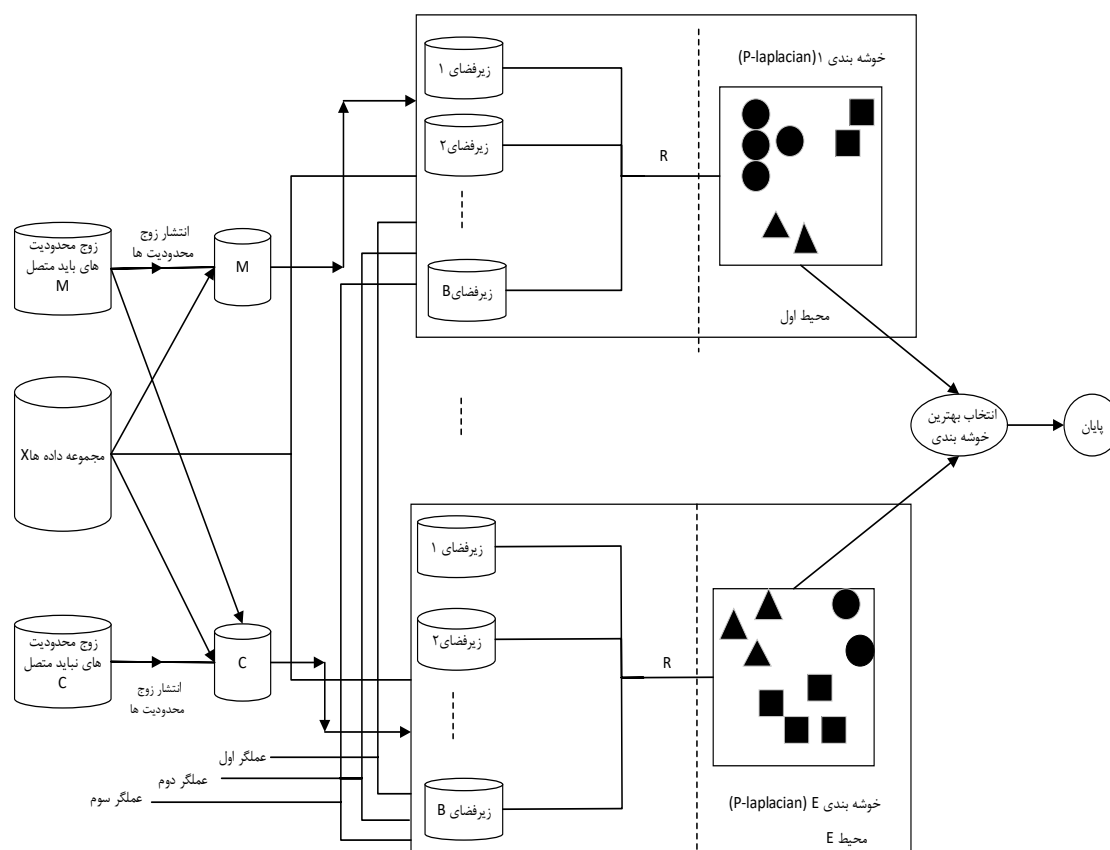
^۱ vertex Laplace

^۲ subspaces

بنابراین، با استفاده از چندین عملگر جستجو روی محیط‌های متشکل از زیرفضاهای متفاوت، بهترین مجموعه 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.

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. از آنجاکه 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 مجموعه زیرفضاها نیاز داریم.



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

دانش پیشین داده شده باشد (که y_i و y_j برچسب رده داده‌های x_i و x_j هستند). سپس، یک ماتریس مربعی $n \times n$ به نام R معرفی می‌کنیم که اطلاعات تمام زوج محدودیت‌ها را در خود جمع کرده است:

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

۳-۱- انتشار زوج محدودیت‌ها

فرض کنید $X = \{x_1, x_2, \dots, x_n\}$ مجموعه داده‌ها که شامل n نمونه است، داده شده باشد. همچنین، فرض کنید که مجموعه‌های زوج محدودیت‌ها: $M = \{(x_i, x_j) | y_i = y_j, 1 \leq i, j \leq n\}$ که نشان‌دهنده مجموعه زوج محدودیت‌های باید-متصل و $C = \{(x_i, x_j) | y_i \neq y_j, 1 \leq i, j \leq n\}$ که نشان‌دهنده مجموعه زوج محدودیت‌های نباید-متصل است، به‌عنوان

پس از آن، با استفاده از خاصیت تراگذری که به صورت زیر تعریف می شود، زوج محدودیت ها به تمامی داده ها تعمیم داده می شوند.

مطابق خاصیت تراگذری، برای هر $1 \leq i, j, k \leq n$ ، چنانچه زوج داده $(x_i, x_j) \in M$ و زوج داده $(x_j, x_k) \in M$ باشند، در نتیجه، زوج داده $(x_i, x_k) \in M$ خواهد بود. بنابراین، براساس رابطه هم ارزی سه داده x_i, x_j و x_k در یک رده یکسان قرار می گیرند و به این ترتیب، تعداد بیشتری زوج داده در مجموعه زوج محدودیت های باید-متصل قرار می گیرند. مجموعه به روزرسانی شده را با $\tilde{M}$ نمایش می دهیم. به روش مشابه، می توانیم زوج محدودیت های باید-متصل را نیز تعمیم دهیم. با توجه به تعریف روابط باید-متصل و نباید-متصل، چنانچه زوج داده (x_i, x_j) عضو مجموعه M و زوج داده (x_j, x_k) عضو مجموعه C باشند، در نتیجه زوج داده (x_i, x_k) عضو مجموعه C خواهد بود؛ بنابراین، اگر دست کم یک زوج محدودیت به مجموعه زوج محدودیت های باید-متصل در مرحله پیش اضافه شده باشد که پیشتر یکی از داده های آن در دست کم، یک زوج محدودیت نباید-متصل ظاهر شده باشد، آنگاه زوج محدودیت های باید-متصل نیز به روزرسانی می شود؛ که آن را با $\tilde{C}$ نمایش می دهیم. قابل ذکر است که باتوجه به مجموعه های به روزرسانی شده زوج محدودیت ها، ماتریس R نیز به روزرسانی می شود. جهت انتشار زوج محدودیت های دوبه دو، از روش های مبتنی بر گراف های طیفی استفاده می کنیم، به این صورت که نخست، مجموعه داده ها را با یک گراف وزن دار نمایش می دهیم و داده ها به عنوان رئوس گراف و شباهت مابین داده ها به عنوان وزن هر یال گراف در نظر گرفته می شود. بنابراین، ماتریس وزن به صورت زیر تعریف می شود:

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

که در آن، $\mathcal{N}_k$ مجموعه k همسایه داده $x_i(x_j)$ و σ به عنوان یک شاخص بیانگر شعاع همسایگی است. پس از به روزرسانی زوج محدودیت های باید-متصل و نباید-متصل ماتریس وزن ها نیز باید به روزرسانی شود. به این منظور، از یک ماتریس مربعی $n \times n$ به نام $\tilde{R}$ استفاده می کنیم که پس از انتشار زوج محدودیت ها با استفاده از خاصیت تراگذری، اطلاعات مربوط به دانش پیشین تعمیم یافته را نمایش می دهد. بنابراین، اگر هیچ

عنصری به مجموعه های باید-متصل و نباید-متصل افزوده نشده باشد، در این صورت $\tilde{R} = R$ است و در غیراین صورت، مشابه مسئله برش کمینه^۱ به همراه عبارت تنظیم کننده رابطه محاسبه ماتریس U ، به صورت زیر نوشته می شود:

$$\min_{\tilde{R}} \frac{1}{2} \|\tilde{R} - R\|_F^2 + \frac{\lambda}{2} \text{tr}(\tilde{R}^T L \tilde{R} + \tilde{R} L \tilde{R}^T) \quad (3)$$

که در آن $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ ماتریس لاپلاسی گراف^۲ و $d_{ii} = \sum_j w_{ij}$ ماتریس درجه^۳ گراف است. معادله بالا یک جواب بسته و یکتا دارد که با استفاده از مشتق گیری نسبت به متغیر مجهول مسئله، مقدار بهینه ماتریس $\tilde{R}$ مطابق رابطه زیر محاسبه می شود:

$$\tilde{R} = \left(\frac{1}{1+\lambda} \right)^2 \left(I - \frac{1}{1+\lambda} \bar{L} \right)^{-1} R \left(I - \frac{1}{1+\lambda} \bar{L} \right)^{-1} \quad (4)$$

که در آن $\bar{L} = I - L$ است.

هر دره ماتریس $\tilde{R}$ (یعنی $\tilde{r}_{ij}$) میزان احتمال هم رده بودن دو داده x_i و x_j را نشان می دهد؛ بنابراین، اگر مقدار درایه $\tilde{r}_{ij}$ کم باشد، به این معنی است که میزان احتمال هم رده بودن دو داده x_i و x_j نیز کم است. در نتیجه، میزان وزن (شباهت) بین دو داده باید کاهش پیدا کند؛ پس وزن به روزرسانی شده برابر است با:

$$\tilde{w}_{ij} = (1 + \tilde{r}_{ij}) w_{ij} \quad (5)$$

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

$$\tilde{w}_{ij} = 1 - (1 - \tilde{r}_{ij})(1 - w_{ij}) \quad (6)$$

پس از انتشار زوج محدودیت های باید-متصل و نباید-متصل و به روزرسانی ماتریس وزن، می توانیم با استفاده از خوشه بندی طیفی کل داده ها را خوشه بندی کنیم.

۳-۲- تولید محیط های جستجو و

خوشه بندی گروهی در هر محیط

همان طور که در شکل (۱) نشان داده شده است، برای این که بتوانیم بهترین خوشه بندی در مجموعه

^۱ Min Cut

^۲ Laplacian matrix

^۳ Degree matrix

زوج محدودیت‌ها را به‌صورت زیر معرفی کرده، سپس، فاکتور اطمینان را به دست می‌آوریم:

$$\hat{a}_{ij} = \begin{cases} 1 & \text{if } (x_i, x_j) \in M \\ 0 & \text{if } (x_i, x_j) \in C \\ -1 & \text{otherwise} \end{cases} \quad (8)$$

که M مجموعه محدودیت‌های باید-متصل و C مجموعه محدودیت‌های نباید-متصل است. با استفاده از ماتریسهای مجاورت معرفی‌شده در روابط (۷) و (۸) نرخ درستی خوشه‌بندی زوج‌داده‌های مربوط به محدودیت‌ها را محاسبه می‌کنیم:

$$l^b = \frac{\sum_i \sum_{j \text{ and } j > i} \{a_{ij}^b = \hat{a}_{ij}\}}{\sum_i \sum_{j \text{ and } j > i} 1 \{ \hat{a}_{ij} \neq -1 \}} \quad (9)$$

سپس، نرخ l^b ، مطابق رابطه زیر طبیعی‌سازی و مقیاس‌بندی می‌شود:

$$l^b = \xi \cdot \frac{l^b - l_{\min}}{l_{\max} - l_{\min}} \quad (10)$$

که در آن $l_{\max} = \max_b l^b$ و $l_{\min} = \min_b l^b$ و ξ ضریب مقیاس است. در نهایت، فاکتور اطمینان مربوط به هر زیرفضا به‌صورت زیر محاسبه می‌شود:

$$\pi^b = \frac{B^{l^b}}{\max_b B^{l^b}} \quad (11)$$

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

در آخرین مرحله، پس از محاسبه 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:

$$\mathcal{A} = \frac{1}{B} \sum_{b=1}^B (\pi^b A^b) \quad (12)$$

اکنون با استفاده از روش خوشه‌بندی طیفی مبتنی بر گراف لاپلاسی- p که در آن ماتریس لاپلاسی $\mathcal{L} = \mathcal{I} - \mathcal{D}^{-\frac{1}{2}} \mathcal{A} \mathcal{D}^{-\frac{1}{2}}$ 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 تحت تأثیر مقیاس‌دهی‌های اولیه متفاوت، قرار می‌گیرند. بنابراین، یافتن بهترین زیرفضای راه‌حل امری

زیرفضاها را بیابیم، تعدادی زیرمجموعه 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 E نشان‌دهنده تعداد محیط‌ها باشد و شاخص B تعداد زیرفضاهای هر محیط را نشان دهد؛ که این دو شاخص توسط کاربر معرفی می‌شوند. در این صورت، $\{\Phi^1, \dots, \Phi^E\}$ مجموعه محیط‌ها هستند و $\{S^1, \dots, S^B\}$ مجموعه زیرفضاها که به‌طور کامل تصادفی تولید شده‌اند. نخست، خوشه‌بندی طیفی نیمه‌نظارتی مبتنی بر ماتریس وزن به‌روزرسانی‌شده در رابطه (۶) از زیربخش ۱-۳ برای داده‌های هر یک از زیرفضاهای هر محیط به‌طور جداگانه انجام می‌شود.

در روش پیشنهادی برای این که بتوانیم دقت خوشه‌بندی طیفی را به‌طور قابل ملاحظه‌ای افزایش دهیم، از روش خوشه‌بندی طیفی مبتنی بر گراف لاپلاسی- p استفاده شده است. همان‌گونه که پیش‌تر ذکر شد، گراف لاپلاسی- 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- p استفاده شده است. در پژوهشی که بوهلر و هین [۱۵] انجام دادند، نشان داده شد که ارتباطی بین دومین بردار ویژه ماتریس لاپلاسی- p و برش چپگر وجود دارد و در واقع، دومین بردار ویژه ماتریس لاپلاسی- p زمانیکه $p \rightarrow 1$ ، به سمت جواب بهینه برش چپگر میل می‌کند. در روش خوشه‌بندی طیفی مبتنی بر گراف لاپلاسی- p بجای نرم-2 از نرم- p ^۱ در محاسبات گراف استفاده می‌شود. پس از انجام خوشه‌بندی، نتایج هر خوشه‌بند برای هرکدام از زیرفضاها را در یک ماتریس مجاورت به نام A^b که $b = 1, \dots, B$ قرار می‌دهیم که درایه‌های آن به‌صورت زیر مقداردهی می‌شوند:

$$a_{ij}^b = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ is in the same cluster,} \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

سپس، یک فاکتور اطمینان برای نمایش میزان تطابق خوشه‌بندی حاصل با زوج محدودیت‌ها محاسبه می‌شود. به این منظور، یک ماتریس مجاورت برای نمایش

^۱ p-norm

$$Sim(\Phi^i, \Phi^j) = \frac{1}{\sum_{k=1}^K \left(\frac{\|v^i - v^j\|}{\|v^i - v^k\|} \right)^{\frac{2}{q-1}}} \quad (14)$$

که در آن K و q ، دو شاخص هستند که به ترتیب تعداد همسایه‌های نزدیک و فازی‌ساز را نمایش می‌دهند. هرچه مقدار Sim بیشتر باشد، احتمال انتخاب محیط j به عنوان همسایه نزدیک و مشابه محیط i بیشتر می‌شود. سپس، بر اساس میزان شباهت محیط‌ها، یک درجه عضویت تجمعی معرفی می‌کنیم:

$$K = \sum_{h=1}^k Sim(\Phi^i, \Phi^h) \quad (15)$$

با استفاده از این دو مقدار، یعنی میزان شباهت محیط‌ها و درجه عضویت تجمعی، اندازه فاصله احتمالی هر محیط با محیط‌های دیگر محاسبه می‌شود:

$$\rho_l = \left[\frac{K_{l-1}}{K} - \frac{K_l}{K} \right] \quad (16)$$

$$K_l = K_{l-1} + Sim(\Phi^i, \Phi^j) \quad (17)$$

مقدار $K_0 = 0$ در نظر می‌گیریم و $l \in \{1, \dots, K\}$ است. در ادامه، نحوه عملکرد هرکدام از عملگرها را جداگانه معرفی می‌کنیم:

عملگر جستجوی محلی: در این الگو دو محیط انتخاب می‌شوند که محیط نخستین تصادفی و محیط دوم از بین همسایه‌های نزدیک نخستین با استفاده از رابطه (۱۶) انتخاب می‌شود. فرض کنید محیط اول Φ^i و محیط دوم Φ^j باشد. پس از انجام خوشه‌بندی در هرکدام از زیرفضاهای محیط‌های i و j ، آنها را بر اساس میزان ارضای محدودیت‌های دوبه‌دو رتبه‌بندی می‌کنیم. معیار رتبه‌بندی که بر اساس میزان ارضای محدودیت‌ها ساخته شده است، به صورت زیر تعریف می‌شود:

$$\eta(S^b) = \frac{N_s}{N_t}, \quad b=1, \dots, B \quad (18)$$

که در آن N_s ، تعداد محدودیت‌های ارضاشده توسط خوشه‌بند مربوط به زیرفضای S^b و N_t ، تعداد کل محدودیت‌ها است. محیط‌های Φ^i و Φ^j پس از رتبه‌بندی بر اساس میزان ارضای محدودیت‌ها به صورت زیر تعریف می‌شوند:

$$\Phi^i = \{S_{\Phi^i}^{i1}, S_{\Phi^i}^{i2}, \dots, S_{\Phi^i}^{iB}\} \quad (19)$$

$$\Phi^j = \{S_{\Phi^j}^{j1}, S_{\Phi^j}^{j2}, \dots, S_{\Phi^j}^{jB}\} \quad (20)$$

ضروری است که طی 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 راه حل را می‌یابیم. عملگرهای جستجو به دو صورت محلی و سراسری تمامی محیط‌ها و زیرفضاهای هر محیط را جستجو می‌کنند. نحوه عملکرد این عملگرها به صورت 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 اولیه 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 به صورت محلی یا سراسری، لازم است محدوده جستجو را با تعیین میزان نزدیکی یا شباهت محیط‌ها تعیین کنیم. همان‌گونه که پیشتر، بیان کردیم، تعدادی زیرفضای تصادفی نابرابر در چند محیط مجزا تولید شده است. باتوجه به 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 نشان‌دهنده ویژگی‌های انتخاب‌شده هر زیرفضای درون هر محیط است، برای هر محیط به صورت زیر تعریف می‌کنیم:

$$v^i = \sum_b v_b \quad (13)$$

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

بندی نیمه‌نظارتی نمی‌تواند نتایج رضایت‌بخشی در این مجموعه داده به دست آورند.

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

در رویکرد پیشنهادی، از دو معیار اطلاعات متقابل طبیعی شده^۱ (NMI) [27] و شاخص تصادفی تنظیم شده^۲ (ARI) [28] برای بررسی ارزیابی عملکرد استفاده می‌شود. معیار NMI اطلاعات آماری اشتراکی متغیرهای تصادفی مربوط به خوشه‌بندی به دست آمده و خوشه‌بندی هدف را اندازه‌گیری می‌کند. فرض کنید $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_k\}$ به ترتیب خوشه‌بندی هدف به دست آمده باشند، و $|\mathcal{C}_i| = n_i$ ، $|\hat{\mathcal{C}}_j| = n'_j$ ، معیار NMI به صورت زیر تعریف می‌شود:

$$NMI(\mathcal{C}, \hat{\mathcal{C}}) \quad (29)$$

$$= \frac{\sum_{i=1}^c \sum_{j=1}^k n_{ij} \log \left(\frac{n \times n_{ij}}{n_i \times n'_j} \right)}{\sqrt{\left(\sum_{i=1}^c n_i \log \frac{n_i}{n} \right) \left(\sum_{j=1}^k n'_j \log \frac{n'_j}{n} \right)}}$$

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

$$ARI(\mathcal{C}, \hat{\mathcal{C}}) = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \sum_i \binom{n_i}{2} \sum_j \binom{n'_j}{2} / \binom{n}{2}}{\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n'_j}{2} \right] / 2 - \sum_i \binom{n_i}{2} \sum_j \binom{n'_j}{2} / \binom{n}{2}} \quad (30)$$

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

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

$$\eta(S_{\Phi_i}^{i_1}) \geq \eta(S_{\Phi_i}^{i_2}) \geq \dots \geq \eta(S_{\Phi_i}^{i_B}) \quad (21)$$

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 محدوده $[0, 1]$ ، زیرفضای $S_{\Phi_i}^{i_{[r_3 B]}}$ انتخاب شده و یک زیرفضای جدید جایگزین آن می‌شود:

$$\Phi'^i = \{S_{\Phi_i}^{i_1}, \dots, S_{\Phi_i}^{i_{[r_3 B] + 1}}, \dots, S_{\Phi_i}^{i_B}\} \quad (28)$$

این مراحل بر روی محیط‌های مسأله، تا زمانی که تغییر در محیط‌ها قابل ملاحظه باشد، یا این که 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 مسئله خوشه‌بندی روی آن محیط برتر اعمال می‌شود.

۴- نتایج آزمایش‌ها

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

مجموعه داده‌های معرفی شده در جدول (۱)، مجموعه داده‌های چالش برانگیزی هستند که در آنها ابعاد داده‌ها بسیار بالا و اندازه نمونه کوچک است؛ به عنوان مثال، مجموعه داده Garber-2001 شامل ۶۶ نمونه و هر نمونه دارای ۴۵۵۳ بُعد است. رویکردهای متداول خوشه-

^۱ Normalized Mutual Information

^۲ Adjusted Rand Index

ایده پیشنهادی با نتایج روش پیشنهاد شده در مرجع [22] مقایسه می‌شود و سرانجام، ایده پیشنهادی با رویکردهای مختلف خوشه‌بندی نیمه‌نظارتی رایج و معروف بر روی مجموعه‌داده‌های دنیای واقعی مقایسه و نتایج آزمایش‌ها بحث می‌شود.

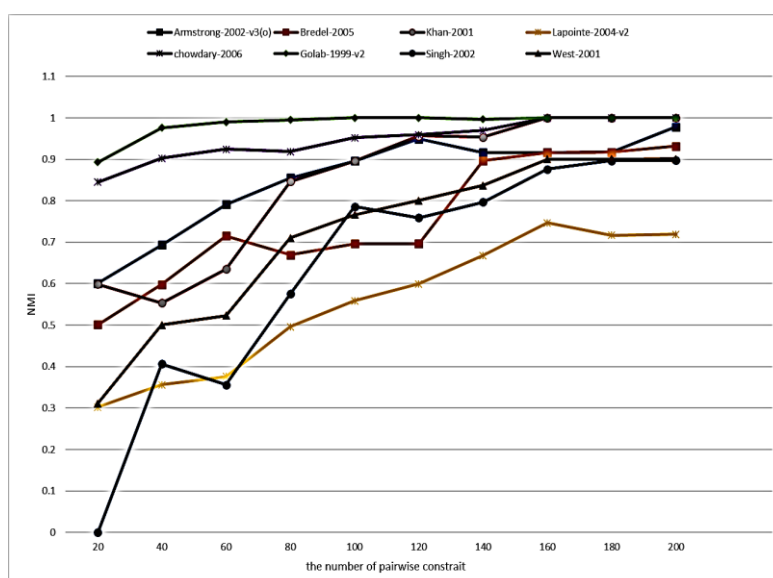
تعداد دفعات تکرار الگوریتم پیشنهادی، چهل‌بار در نظر گرفته شده است و زوج محدودیت‌ها در تمام اجراها یکسان است. نخست، تأثیر تعداد متفاوت زوج محدودیت‌ها و تأثیر تعداد زیرفضای تصادفی و تأثیر نرخ تولید زیرفضای تصادفی را بررسی می‌کنیم. سپس، نتایج

(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.

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



(شکل-۲): تأثیر تعداد متفاوت زوج محدودیت‌ها

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

۴-۱- تأثیر تعداد متفاوت زوج محدودیت‌ها

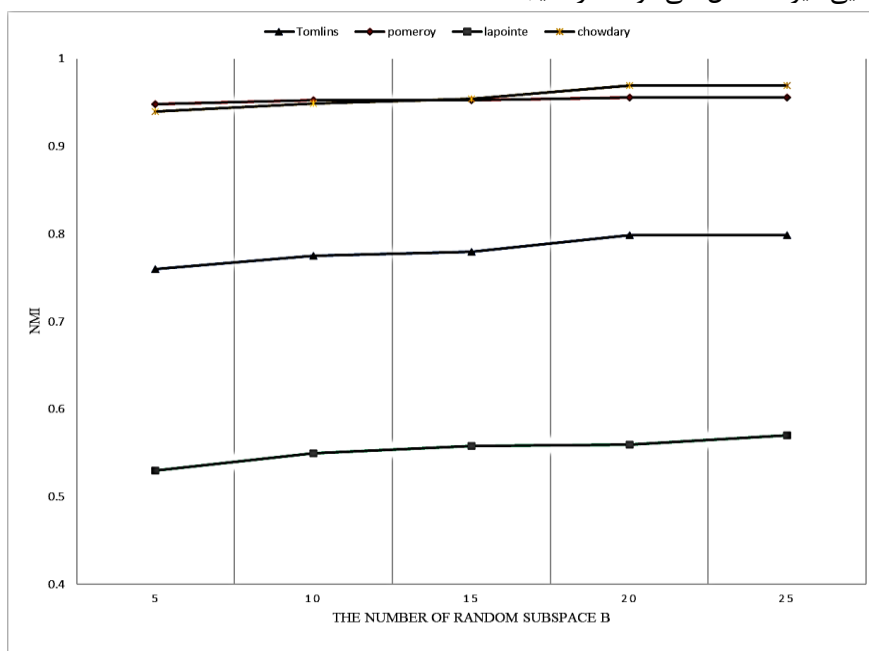
شکل (۲) تأثیر تعداد متفاوت زوج محدودیت‌ها را بر عملکرد روش پیشنهادی بر روی ۸ مجموعه داده سرطان با استفاده از معیار سنجش NMI نشان می‌دهد. منحنی-های مقادیر معیار NMI نمایش داده شده در شکل (۲) را بیان می‌کنند که همان‌گونه که انتظار داریم، با افزایش تعداد زوج محدودیت‌ها، کارایی روش پیشنهادی نیز افزایش می‌یابد؛ به عنوان مثال، در مجموعه داده Khan-2001 وقتی که تعداد زوج محدودیت‌ها از ۲۰ تا ۱۰۰ افزایش می‌یابد، مقدار NMI نیز از ۰/۶ تا ۰/۹ افزایش می‌یابد. همچنین، در مجموعه داده West-2001 مقدار NMI از ۰/۳۱ به ۰/۷۶ افزایش می‌یابد. این نشان می‌دهد که روش پیشنهادی به تعداد زوج محدودیت‌ها حساس است؛ که یک نتیجه‌گیری منطقی است و دور از انتظار نیست. تعداد بیشتر زوج محدودیت‌ها به معنای در دسترس بودن اطلاعات بیشتر است، که به بهبود عملکرد الگوریتم خوشه‌بندی کمک می‌کند. قابل ذکر است که زوج محدودیت‌ها به طور معمول، به کمک دانش پیشین کارشناسان به دست می‌آید و اغلب برای به دست آوردن این دانش هزینه بالایی نیز متحمل 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 محدودیت‌ها 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 مجموعه داده که به طور تصادفی انتخاب شدند، انجام شد.

همان‌گونه که در شکل (۳) مشاهده می‌کنید، افزایش تعداد زیرفضاهای تصادفی، تأثیر مثبتی روی عملکرد خوشه‌بندی باتوجه به مقدار NMI دارد. نکته قابل تأمل این است که در بیشتر مجموعه داده‌ها، 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

رود، مقدار NMI افزایش پیدا می‌کند و با تعداد ۲۰ زیرفضا، مقدار NMI برابر با ۰/۷۸۵ و ثابت باقی می‌ماند.

به عنوان مثال، در مجموعه داده Tomlins-2006-v1 زمانی که مقدار B برابر ۵ است، مقدار NMI به دست آمده برابر با ۰/۷۷ است و با افزایش مقدار B، همان‌طور که انتظار می‌-

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 باتوجه به مقدار NMI را نشان می‌دهد. روش پیشنهادی در بیشتر مجموعه داده‌های سرطانی، مانند مجموعه داده‌های Khan-2001، Bredel-2005، Armstrong-2002-v3(o) 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 رضایت بخشی نتایج حاصل کمتر می‌شود. یکی از دلایل احتمالی این است که افزایش نرخ تولید زیرفضاهای تصادفی، باعث ایجاد نمونه داده‌های نوفه دار و زاید بیشتری در زیرفضا می‌شود که بر عملکرد روش پیشنهادی تأثیر منفی می‌گذارد.

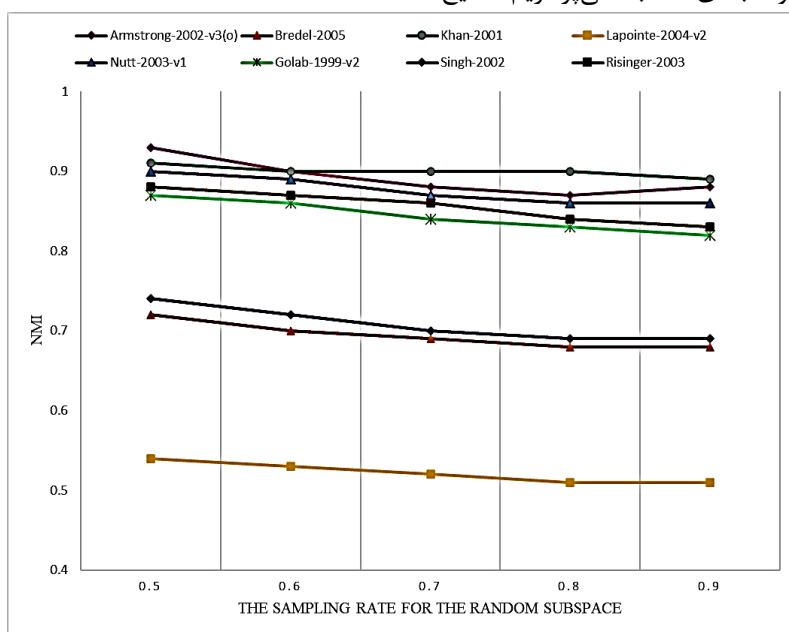
۴-۴- مقایسه روش پیشنهادی با روش مرجع [22]

در این زیر بخش به تحلیل و مقایسه روش پیشنهادی با یکی از روش‌های خوشه بندی مشابه می‌پردازیم. نتایج

عملکرد روش خوشه بندی ایده پیشنهادی نسبت به روش ارائه شده در مرجع [22] با استفاده از معیار سنجش ARI، در جدول (۲) نشان داده شده است. با بررسی و مقایسه نتایج به دست آمده، باتوجه به معیار سنجش ARI، روش پیشنهادی در ۱۲ مجموعه داده از ۲۰ مجموعه داده از روش مرجع [22] پیشی گرفته و عملکرد بهتری داشته است. به عنوان مثال، در مجموعه داده Pomeroy-2002-v2 مقدار ARI برای روش پیشنهادی، به میزان ۰/۳۵۷ بیشتر از روش مرجع [22] به دست آمده است.

به عنوان مثالی دیگر، در مجموعه داده Khan-2001 مقدار ARI روش پیشنهادی به میزان ۰/۶۰۷ بیشتر از روش مرجع [22] است.

برای تحلیل بهتر نتایج و مقایسه روش پیشنهادی، عملکرد روش پیشنهادی و روش مرجع [22] در شکل (۵) با استفاده از نمودار جعبه ای رسم شده است. نمودار جعبه ای، نموداری برای توصیف تغییرات، توزیع و پراکندگی داده ها بر اساس چندین شاخص مرکزی و پراکندگی آماری است. از این نمودار می‌توان برای مقایسه ویژگی های چند روش مختلف بر روی مجموعه داده های یکسان به طور هم زمان استفاده کرد.



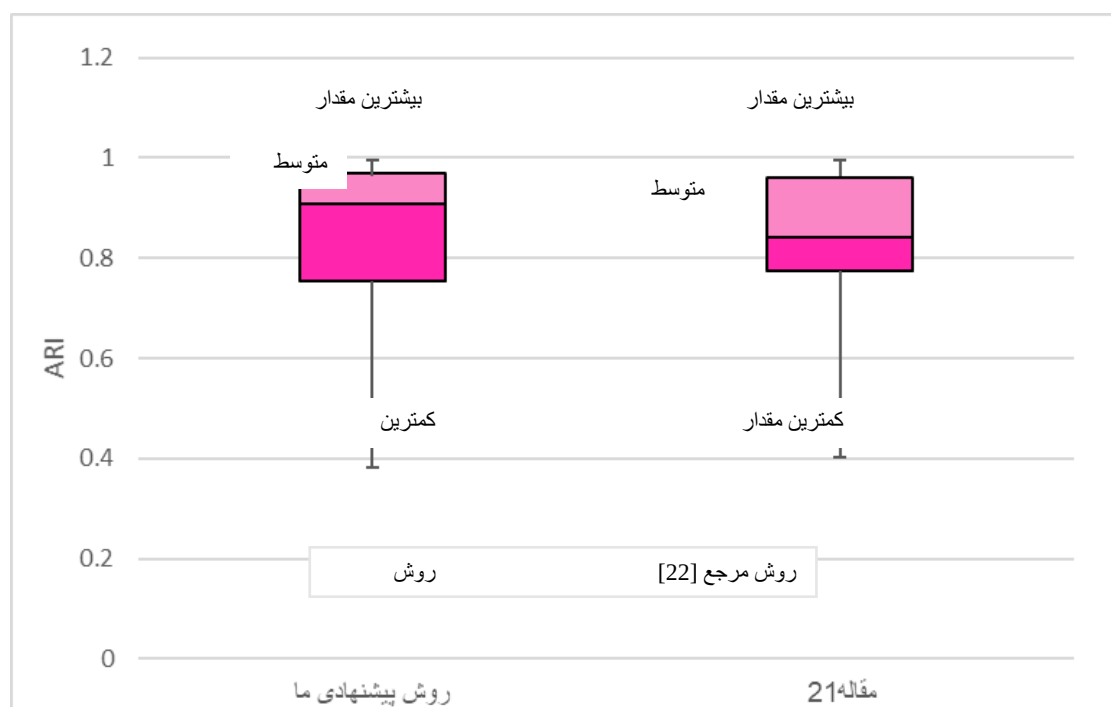
(شکل-۴): تأثیر نرخ تولید زیرفضاهای تصادفی

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

(جدول-۲): مقایسه 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

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



(شکل - ۵): نمودارهای جعبه‌ای مقایسه روش پیشنهادی با روش مرجع [22] با استفاده از معیار ARI

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

نمودارهای جعبه‌ای خلاصه پنج ویژگی عددی از مجموعه داده‌ها را نشان می‌دهد. این مقادیر شامل کمترین مقدار^۱، یک چهارم اول^۲، متوسط^۳، یک چهارم سوم^۴ و بیشترین مقدار^۵ است. همان‌طور که 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 جعبه آن بزرگتر از ارتفاع جعبه روش مرجع [22] است. یکی از دلایل احتمالی این افزایش مقدار، می‌تواند انتخاب فرایند خوشه‌بندی 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

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 مورد مقایسه عبارتند از روش تجزیه و تحلیل مؤلفه‌های مرتبط^۱ (RCA) [29]، روش یادگیری متریک اطلاعات نظری^۲ (ITML) [30]، روش تجزیه و تحلیل مؤلفه‌های تمیزکننده^۳ (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 زیر خط مشخص شده‌اند.

همان‌طور که در جدول (۳) مشاهده می‌کنید، روش پیشنهادی در ۱۱ مجموعه داده از ۲۰ مجموعه داده، بهترین نتایج را در مقایسه با سایر روش‌های مقایسه‌شده به دست آورده است. همچنین، در شش مجموعه داده از ۲۰ مجموعه داده، دومین مجموعه، بهترین نتیجه را به دست آورده است؛ به‌عنوان مثال، در مجموعه داده Bredel-2005 مقدار به‌دست‌آمده در روش پیشنهادی ۰/۱۰۴۲ بیشتر از مرجع [22] و مقدار ۰/۱۸۴۶ بیشتر از RCA و مقدار ۰/۴۶۸ بیشتر از ITML و همچنین، مقدار ۰/۴۶۸ بیشتر از DCA است. همچنین، در مجموعه داده 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 خوشه‌بندی باشد، زیرا گاهی اوقات دقیق‌تر از فضای اصلی انجام می‌شود و همچنین، دلیل دیگر می‌تواند این موضوع باشد که روش پیشنهادی نتیجه چندین خوشه‌بندی

به‌دست‌آمده از زیرفضاهای مختلف تصادفی را در یک نتیجه کلی ادغام می‌کند؛ که این باعث بهبود عملکرد خوشه‌بندی نسبت به روش‌های یادگیری متریک است.

۵- نتیجه‌گیری

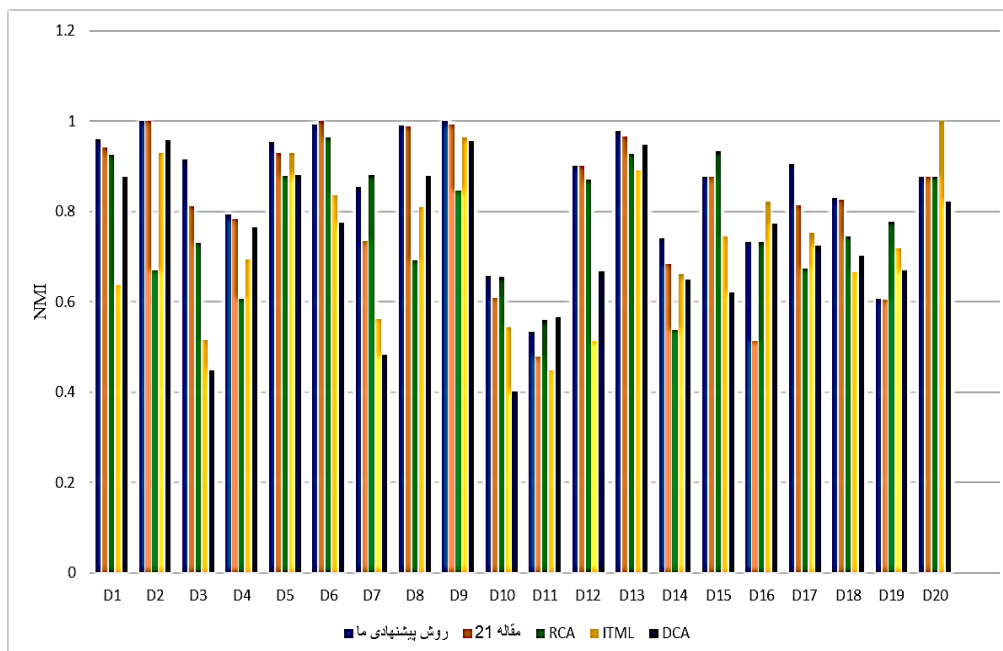
در این مقاله، یک چارچوب خوشه‌بندی گروهی بر مبنای خوشه‌بندی لاپلاسی- p برای خوشه‌بندی داده‌ها با ابعاد بالا، پیشنهاد شد. در مقایسه با رویکردهای خوشه‌بندی نیمه‌نظارتی سنتی، روش پیشنهادی این مقاله، چند ویژگی مهم داشت. از جمله، از روش پیشنهادی برای چگونگی استفاده از تمام جفت محدودیت‌های باید-متصل و نباید-متصل از خاصیت تراگذاری و انتشار جفت محدودیت‌ها استفاده کرد. همچنین، در روش پیشنهادی، خوشه‌بندی گروهی با استفاده از تولید تعدادی زیرفضای تصادفی پیاده‌سازی شد، به این ترتیب که راه‌حل‌های هر کدام از خوشه‌بندها را جمع کرده و وظیفه خوشه‌بندی نهایی را با استفاده از اطلاعات جمع‌شده به یک خوشه‌بند نهایی یک‌پارچه حاصل از زیرفضای بهینه واگذار کرد.

جهت افزایش عملکرد روش پیشنهادی از روش خوشه‌بندی 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 تراگذاری و انتخاب زیرفضاهای تصادفی با اندازه نابرابر، نقش مهمی 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 مجموعه داده‌های 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 محدودیت‌ها و نحوه حذف محدودیت‌های زائد را می‌توان به‌عنوان یکی از محدودیت‌های روش پیشنهادی در نظر گرفت.

¹ Relevant Component Analysis

² Information Theoretic Metric Learning

³ Discriminative Component Analysis



(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

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

- [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
- [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.

- 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 ریاضیات از دانشگاه شهید باهنر کرمان است. زمینه‌های پژوهشی وی طراحی 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 دانش‌آموخته رشته 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 بین‌المللی چاپ کرده است. نشانی رایانامه ایشان عبارت است از:

afsari@uk.ac.ir

