# ترکیب یک روش خوشه‌بندی تجمعی و یک معیار شباهت جدید برای مدل‌سازی رفتار وراثتی بیماری‌ها

**موسی مجرد*[۱و۲]، حمید پروین[۳و۴]، صمد نجاتیان [۵و۶] و کرم‌الله باقری فرد[۶و۷]**

[۱]گروه کامپیوتر، واحد فیروزآباد، دانشگاه آزاد اسلامی، فیروزآباد، ایران

[۲]باشگاه پژوهش‌گران جوان، واحد فیروزآباد، دانشگاه آزاد اسلامی، فیروزآباد، ایران

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

[۴]باشگاه پژوهش‌گران جوان، واحد نورآباد ممسنی، دانشگاه آزاد اسلامی، نورآباد ممسنی، ایران

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

[۶]باشگاه پژوهش‌گران جوان، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

[۷]گروه کامپیوتر، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

## چکیده

**امروزه تئوری‌های بسیاری در مورد علل بروز بیماری‌های وراثتی وجود دارد، اما پزشکان معتقدند که دو فاکتور ژنتیک و محیط زیست هم‌زمان باهم نقش مهمی در بروز و پیشرفت این بیماری‌ها ایفا می‌کنند، هرچند که چگونگی این اثرگذاری هنوز به‌طور دقیق مشخص نیست. برای اینکه بتوان ژن‌های مؤثر در بروز بیماری‌ها را تشخیص داد، باید ارتباط بین سلول‌ها/بافت‌ها را به‌دست آورد. تعامل بین سلول‌ها یا بافت‌های مختلف را می‌توان، با بیان ژن بین آنها نشان داد. با نمونه‌برداری از کروموزوم‌ها، اطلاعات مفیدی در مورد نوع بیماری و چگونگی انتقال آن استخراج می‌شود. با بررسی این اطلاعات می‌توان اختلالاتی را که منجر به تغییرات به‌شدت پرتکراری شده‌اند شناسایی کرد. در این مقاله تشخیص ارتباط‌های بین سلولی و بین بافتی در بیماری‌های مختلف با توجه به مشخصات ساختار توپولوژیکی گراف و یک روش خوشه‌بندی تجمعی بهبودیافته انجام شده است. روش پیشنهای دو مرحله دارد؛ در مرحله نخست چندین مدل خوشه‌بندی به‌منظور تشخیص ارتباط‌های اولیه بین سلول‌ها یا بافت‌ها در جهت تولید نتایج بهتر نسبت به الگوریتم‌های انفرادی، ترکیب می‌شوند. در مرحله دوم تشابه بین سلول‌ها یا بافت‌ها در هر خوشه با استفاده از یک معیار شباهت مبتنی بر ساختار توپولوژیکی گراف محاسبه و درنهایت از بیشینه شباهت‌های بین سلول‌ها یا بافت‌ها در هر خوشه برای کشف ارتباطات بین بیماری‌ها استفاده می‌شود. به‌منظور ارزیابی عملکرد روش پیشنهادی از چندین مجموعه‌داده UCI و همچنین مجموعه‌داده فانتوم پنج استفاده شده است. نتایج روش پیشنهادی روی مجموعه داده فانتوم پنج، ضریب سیلوئت ۰/۹۰۱ را در ۱۸ خوشه برای سلول‌ها و ۰/۷۶۲ در ۱۳ خوشه برای بافت‌ها را گزارش می‌کند.**

**واژگان کلیدی: ارتباط بین سلولی، خوشه‌بندی بهبودیافته، ساختار توپولوژیکی گراف، مجموعه‌داده فانتوم ۵.**

# Combining an Ensemble Clustering Method and a New Similarity Criterion for Modeling the Hereditary Behavior of Diseases

**Musa Mojarad*[1, 2], Hamid Parvin[3,4], Samad Nejatian[5,6] & Karamollah Bagheri Fard[6,7]**

[1]Department of Computer Engineering, Firoozabad Branch, Islamic Azad University, Firoozabad, Iran

[2]Young Researchers Club, Firoozabad Branch, Islamic Azad University, Firoozabad, Iran

[3]Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran.

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

●تاریخ ارسال مقاله: ۱۳۹۷/۱۲/۷ ●تاریخ پذیرش: ۱۳۹۹/۵/۲۸ ●تاریخ انتشار: ۱۴۰۰/۰۷/۱۷ ●نوع مطالعه: کاربردی

سال ۱۴۰۰ شمارهٔ ۲ پیاپی ۴۸





[ DOI: 10.52547/jsdp.18.2.97 ]

[4]Young Researchers Club, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran
[5]Department of Electrical Engineering, Yasooj Branch, Islamic Azad University, Yasooj, IranYoung Researchers Club, Yasooj Branch, Islamic Azad University, Yasooj, Iran
[7]Department of Computer Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran.

## Abstract

***Background:*** **There are many theories about the causes of hereditary diseases, but physician believe that both the genetic and environmental factors simultaneously play an important role in the development and progression of these diseases, although the extent to which this effect is not yet clear. In order to detect effective genes in the development of diseases, it is necessary to achieve the relationship between cells/tissues.**

***Objective:*** **In fact, inter-cell or inter-tissue communications indicate the hereditary relationships between patients. Detecting these communications help to identify common parts of the body that are influenced by various diseases. The interaction between different cells/tissues can be demonstrated by expressing the gene between them. By sampling chromosomes, useful information is obtained about the type of disease and how it is transmitted. By examining this information, you can identify disorders that have led to highly altered changes. In previous research, various clustering methods have been used to discover the links between diseases based on gene expression data. However, ensembl clustering approaches have not yet been used for this purpose.**

***Method:*** **In this paper, the recognition of intercellular and inter-tissue interactions in various diseases have been done according to the characteristics of the topological structure of the graph and an improved ensembl clustering method. The proposed clustering algorithm uses an agreed similarity function to measure the similarity between objects. The proposed method has two stages; in the first step, several clustering models are combined to identify the initial relationships between cells or tissues in order to produce better results than individual algorithms. In the second stage, the similarity between cells or tissues in each cluster is calculated by using a similarity criterion based on the topological structure of the graph. Eventually, the maximum similarity between cells or tissues in each cluster is used to discover the relationship between diseases. In addition, an algorithm for improving the uncertainty of objects is evaluated by allocating them to other clusters in order to enhance the quality of the final clusters.**

***Results:*** **To evaluate the performance of the proposed method, several UCI datasets and the FANTOM5 dataset have been used. The results of the proposed method on the phantom data set 5 report a silhouette of 0.901 in 18 clusters for cells and 0.762 in 13 clusters for tissues.**

***Conclusion:*** **The conducted evaluations have confirmed the power of the proposed clustering algorithm in terms of accuracy. Clustering of cells or tissues has increased the accuracy and concentration of the topological similarity criterion of the graph in the range of similarity of cells or tissues.**

**Keywords: Intercellular communication, Improved clustering, Graph topological structure, FANTOM5 dataset.**

## ۱- مقدمه

بدن انسان از انواع مختلفی سلول و بافت تشکیل شده است. تعداد سلول‌های بدن انسان حدود ۳۷/۲ تریلیون تخمین زده شده است. سلول‌ها و بافت‌های بدن به‌طور مستقیم یا غیرمستقیم به یکدیگر وابسته هستند و توانایی ارتباط و تأثیر بر روی یکدیگر را دارند؛ بنابراین سازوکاری مؤثر و کارآمد به‌منظور یافتن ارتباط بین این تعداد نجومی از سلول‌ها و بافت‌ها مورد نیاز است. یافتن این ارتباطات به شناسایی بیماری‌ها مختلف و عوامل آن کمک خواهد کرد. سلول‌ها و بافت‌های بیماری می‌تواند بر اثر جهش ایجاد شوند [1]. هسته هر سلول دستورالعمل‌های کدگذاری‌شده لازم را برای هدایت فعالیت‌های سلول و ساخت پروتئین‌های دارد. به یک گروه کامل از این دستورالعمل‌های خام ژنوم اطلاق می‌شود. ژنوم انسان به مجموعه ژنتیکی و ژن‌های داخل هسته سلول‌های انسان گفته می‌شود [2].

ژن‌های موجود در یک کروموزوم حدود چند میلیون است. بیش‌تر این ژن‌ها در تمام انسان‌ها یکسان و ثابت هستند(Fix Gen). به این نوع ژن‌ها دخالتی در تغییرات ژنتیکی و سلولی/بافتی ندارند [3]. تغییرات این نوع ژن‌ها باعث ایجاد ناپایداری‌هایی در انسان خواهد شد (به‌عنوان مثال تولید انسان‌هایی با چهارچشم)؛ بنابراین این نوع ژن‌ها نباید در بدن انسان تغییر کنند؛ اما بخش کوچکی از ژن‌ها ثابت نیستند و مقدار آن‌ها در انسان‌های مختلف متغیر هستند (Variables Gen). متغیربودن این ژن‌ها باعث می شود انسان‌های مختلفی وجود داشته باشد. هر ژن در روی کروموزوم مکان معینی را اشغال می‌کند که در اصطلاح به آن جایگاه ژن (Locus) گفته می‌شود [4].




[ DOI: 10.52547/jsdp.18.2.97 ]

همان‌طور که گفته شد، تمامی کروموزوم‌های یک بخش بدن یکسان هستند. به‌عنوان‌مثال تمام کروموزوم‌های دستگاه دفاعی بدن در تمام انسان‌ها کروموزوم $\Gamma$ است. در ابتدا این کروموزوم برای همه انسان‌ها به‌طورکامل یکسان بودند، ولی در اثر حوادث و بروز بیماری‌های مختلف، درگذر زمان دچار تغییراتی شده‌اند. در یک انسان ممکن است، یکی از کروموزوم‌های دستگاه دفاعی بدن در یکی از سلول‌ها یا بافت‌ها دچار بیماری‌های ویروسی شده باشد. خصوصیات این ویروس باعث تغییرات و انتقال ویژگی‌های آن به سایر ژن‌های همسایه می‌شود. درنتیجه این تغییرات باعث تغییر در DNA آن سلول/بافت شده و اختلالاتی را ایجاد می‌کند [5].

با انجام نمونه‌برداری‌هایی روی این کروموزوم‌ها باعث استخراج اطلاعات مفیدی در مورد نوع بیماری و چگونگی انتقال آن به دست می‌آید. با بررسی کروموزوم‌های نمونه‌برداری‌شده از این سلول‌ها و بافت‌ها، می‌توانیم اختلالات ناشی از عملکرد ویروس را که منجر به تغییرات به‌شدت پرتکراری می‌شوند، شناسایی کنیم؛ اما چالش پیشرو شباهت بخش زیادی از این سلول‌ها/بافت‌ها و عدم توجه به تغییرات و الگوها در بخش‌های مختلف است. هدف این پژوهش ارائه یک روش خوشه‌بندی بهبودیافته به‌منظور شناسایی ارتباطات بین سلولی و بین بافتی با توجه به بیماری‌های مختلف است. وجود بیماری‌های مختلف باعث ایجاد ارتباط‌های چندوجهی بین سلول‌ها/بافت‌ها می‌شود که درنهایت این ارتباطات یک هایپر گراف را تشکیل می‌دهد [6]. به‌منظور شناسایی تغییرات ایجادشده در سلول‌ها/بافت‌ها، ارتباطات بین هر سلول/بافت را به‌صورت یک یال با وزن خاصی در گراف ایجاد می‌شود. در اینجا با استخراج اطلاعات توپولوژیکی گراف و استفاده از یک معیار شباهت ترکیبی احتمال مرتبط‌بودن ژن‌ها با توجه به مجموعه‌ای از بیماری‌ها مشخص می‌شود. به‌منظور تحقق اهداف پژوهش از مجموعه‌داده فانتوم پنج [7] استفاده شده است. اطلاعات مربوط به این مجموعه‌داده در نشانی (http://fantom.gsc.riken.jp/5) موجود است. در این پایگاه داده اطلاعات بیان ژن بیماران تحت عنوان راه‌انداز ژن در دسترس است [8].

ادامه مقاله به‌صورت زیر سازمان‌دهی شده است. بخش ۲ مسأله خوشه‌بندی تجمعی و تئوری گراف معرفی می‌شود. بخش ۳، کارهای مرتبط و بخش ۴، جزئیات روش پیشنهادی به همراه مراحل متفاوت آن شرح داده می‌شود. در بخش ۵ به شرح آزمایش‌های انجام‌شده روی مجموعه داده‌های UCI و همچنین پایگاه داده فانتوم می‌پردازیم. درنهایت نتیجه‌گیری و پیشنهادهای آینده در بخش ۶ ارائه می‌شود.

## ۲- خوشه‌بندی تجمعی و تئوری گراف

خوشه‌بندی یکی از شاخه‌های یادگیری بدون نظارت است که هدف آن یافتن خوشه‌های مشابه از اشیا در بین نمونه‌های ورودی است. می‌توان نشان داد که هیچ معیار مطلقی برای بهترین خوشه‌بندی وجود ندارد، بلکه این بستگی به مسأله و نظر کاربر دارد؛ بااین‌حال معیارهای مختلفی برای خوب‌بودن یک خوشه‌بندی ارائه شده است که می‌تواند کاربر را برای رسیدن به یک خوشه‌بندی مناسب راهنمایی کند. در این مقاله از یک خوشه‌بندی تجمعی بهبودیافته به‌منظور قراردادن سلول‌ها یا بافت‌های مشابه بیماری‌های مختلف در خوشه‌های یکسانی استفاده می‌شود. خوشه‌بندی تجمعی با یک خوشه به‌ازای هر مشاهده شروع شده و در هر مرحله، دو خوشه با بیشترین شباهت ادغام شده و یک خوشه بزرگ‌تر را تشکیل می‌دهند. این کار باعث می‌شود تا در سطوح بالاتر خوشه‌های کمتری وجود داشته باشد [9]. در اینجا از چند مدل مختلف خوشه‌بندی به‌منظور تولید نتایج بهتر نسبت به الگوریتم‌های انفرادی استفاده می‌شود. برای یک مجموعه‌داده با n نمونه به‌صورت $X = \{x_1, x_2, \dots, x_n\}$، نتیجه $P_q = \{c_1^q, c_2^q, \dots, c_k^q\}$ q-مین روش خوشه‌بندی با k خوشه است. خوشه‌بندی $\Phi$ می‌تواند به‌وسیله m روش به‌صورت $\Gamma = \{P_1, P_2, P_3, \dots, P_m\}$ و تابع توافقی F ساخته شود. رابطه (۱) فرم کلی تابع هدف برای خوشه‌بندی $\Phi$ را نشان می‌دهد:

$$\Phi(F, \Gamma)P_1, P_2, P_3, \dots, P_m) = F(\Gamma) \qquad (۱)$$

وظیفه یک خوشه‌بندی تجمعی یافتن افراز $P^*$ از مجموعه‌داده X به‌وسیله ترکیب اعضای P و ارائه یک تابع F مؤثر بدون دسترسی به ویژگی‌های اصلی است [10]. بنابراین $P^*$ به‌احتمال از نظر سازگاری و کیفیت بهتر از اعضای انفرادی در تجمع است. این پژوهش در نظر دارد با توجه به مشخصات ساختار توپولوژی گراف، یک معیار برای محاسبه شباهت بین سلول‌ها/بافت‌ها کاربران ارائه دهد. شباهت بین دو سلول و یا دو بافت احتمال وجود ارتباط بین آن‌ها را نشان می‌دهد که بر اساس خصوصیت‌های موجودیت‌ها و سایر پیوندهای مشاهده‌شده انجام می‌گیرد [11].



[ DOI: 10.52547/jsdp.18.2.97 ]

## ۳- کارهای مرتبط

رده‌یافت‌های ریاضی برای بررسی رفتار سلولی بیشینه طولانی دارد [12]، [13]. برای بررسی رفتار سلولی نظریه گراف سهم گسترده‌ای به دانش بیولوژی اضافه کرده است [14]،[15]. نظریه گراف به بررسی تعامل بر روی توپولوژی‌های با مقیاس بزرگ از یک شبکه تعامل تمرکز می‌کند. ترکیب دو رده‌یافت به‌منظور بررسی تأثیر تئوری بازی روی توپولوژی سلولی بافت مسطح توسط [16] انجام شد. در این پژوهش تئوری بازی و نظریه گراف برای درمان بافت به‌عنوان یک گراف دینامیکی به‌روزشده بر مبنای تعاملات محلی سلول‌های مجزا با هم ترکیب می‌شوند. وجود سازوکارهای ارتباط سلول‌به‌سلول و بافت‌به‌بافت ممکن است، قادر به طراحی دستگاه‌های شبکه ارتباطی کنترل‌شده بین نانوماشین‌ها باشد [18]. فنّاوری مهندسی مولکولی اخیر ممکن است اجازه اصلاح اجزا بیولوژیکی به اجزای معماری را که برای دستگاه‌های شبکه ارتباطی مناسب می‌باشند، بدهند. توصیف یک شبکه ارتباط مولکولی مبتنی بر سلول توسط [19] ارائه شد. در این شبکه یک ردهکلاس از دستگاه‌های ارتباط مولکولی که در آن‌ها نانوماشین‌ها با استفاده از شبکه ارتباطی سلول‌به‌سلول با یکدیگر ارتباط برقرار می‌کنند، مورد توجه است. یک مدل محتوا بر مبنای تصمیم‌گیری و یک روش استخراج گراف توسط [20] ارائه شد. مدل پیشنهادی دو مرحله دارد؛ در مرحله نخست گروه‌های مشابه در گراف تشخیص داده شده سپس ادغام می‌شوند. تولید مجموعه‌ای از یال‌های تکراری بر مبنای درجه پشتیبانی در مرحله دوم انجام می‌شود. به‌منظور درک بهتر سلول‌ها و شبکه‌های ارتباطی خود جمعی و میزان پیشرفت بیماری یک مدل ریاضی یک‌پارچه سلولی سیتوکین متنوع ایجاد شد [21]. در این پژوهش از یک مدل بهینه‌سازی‌شده برای سازگاری بهتر با مجموعه‌داده بهره گرفته شده است. یک روش نمونه‌گیری تصادفی یکنواخت یال‌ها برای الگوریتم استخراج گراف توسط [22]، استفاده شد. در این روش علاوه‌بر مقادیر وزن یال‌ها از نام یال‌ها، درجه و وابستگی نیز بهره گرفته شده است. به‌کارگیری استراتژی جستجوی محلی برای تحلیل استخراج گراف برای کشف الگوهای رفتاری بیماری‌ها توسط [24] پیشنهاد شد. برای نخستین‌بار نقشه‌ای از ارتباطات بین ۱۴۴ نوع سلول بدن انسان ارائه شد [25]. اطلاعات این پژوهش نشان می‌دهد که بیش‌تر سلول‌ها یک بیانی از ده‌ها تا صدها گیرنده برای ایجاد یک شبکه سیگنالی از طریق مسیرهای گیرنده چندگانه دارند. تشخیص ارتباطات ژنی در شبکه‌های چندلایه جهت آشکارسازی سرطان با استفاده از اطلاعات ژنومی برای شناسایی دارندگان بیماری سرطان نیز بررسی شده است. شبکه‌های چندلایه در نظر گرفته‌شده در [26] ترکیب فاکتور هدف‌گیری رونویسی و هدف‌گیری همکاری MICRO RNA، تعامل پروتئین-پروتئین و شبکه‌های ژنی است. در این پژوهش جهت استخراج اطلاعات بیولوژیکی مرتبط به هم از یک الگوریتم خوشه‌بندی ترکیبی استفاده شده است. ترکیبی از لایه‌های مختلف اطلاعات اجازه استخراج الگوهای نظارتی و نقش کارکردی هر دو ژن شناخته‌شده و ژن‌های نامزد جدید را می‌دهد. مدل‌ها و راه‌حل‌هایی به‌منظور استخراج خواص پنهان ویژگی‌های شبکه در [27] مورد بحث قرار گرفته است. این نتایج حاصل از این پژوهش به درک بهتر مفاهیم بیولوژیکی کمک خواهد کرد. شبکه‌های نظارتی حاوی اطلاعات مربوط به کنترل بیان ژن در سلول‌ها است. شبکه‌های انتقال سیگنال اغلب از چندین گراف برای نشان دادن یک‌سری فعل‌وانفعالات بین خواص زیستی بهره می‌گیرند. الگوریتم خوشه‌بندی طیفی برای پیداکردن خوشه‌ها در گراف‌ها مورداستفاده قرار می‌گیرد؛ به‌طوری‌که گروه‌ها با یال‌های بسیار مشابه به هم متصل و ارتباط بین آن‌ها ضعیف است. تراکم شبکه تعداد اتصالات در هر گره را مطابق رابطه (۲) نشان می‌دهد:

$$densiy = \frac{2|E|}{|V|(|V|-1)} \qquad (۲)$$

استدلال می‌شود که شبکه‌های بیولوژیکی به‌طورکلی کم‌اتصال هستند. در یک گراف بدون جهت مرکزیت به‌صورت رابطه (۳) تعریف می‌شود.

$$G_{CLO}CIL = \frac{1}{\sum_{|t\epsilon v|}^{|V|} dist(i,j)} \qquad (۳)$$

درهمین‌اواخر کنسرسیوم فانتوم پنج از تحلیل پوششی بیان ژن cage برای ایجاد ارتقادهنده اطلس استفاده کرده است [25]. در این پژوهش به معرفی نتخستین نقشه سیگنال‌دهی سلول‌به‌سلول از طریق ارائهٔ شبکه در مقیاس کلان پرداخته شده است. الگوریتم افرازبندی ابر گراف (HGPA) و همچنین الگوریتم فراخوشه‌بندی (MCLA) توسط [29] ارائه شدند. یک ابر گراف در HGPA و MCLA ساخته می‌شود و در آن هر خوشه به‌صورت یک فرایال ارائه می‌شود. HGPA به‌طورمستقیم فراگراف را به‌وسیله برش یک تعداد کمینه




[ DOI: 10.52547/jsdp.18.2.97 ]

ممکن از فرایال به K گره متصل از تقریب همان اندازه با استفاده از افرازبندی فراگراف HMETIS [30] افرازبندی می‌کند. MCLA ابتدا شباهت بین زوج‌خوشه‌ها را از جنبه اشتراک اشیای بین آن‌ها تعریف می‌کند. این کار از طریق توسعه شاخص Jaccard انجام، سپس گراف طوری ساخته می‌شود که گره‌ها، خوشه‌ها و یال‌ها، شباهت بین زوج خوشه‌ها را بیان می‌کنند.

یک توسعه دیگر توسط Iam-on [31] با هدف تعریف مجدد ماتریس همکاری ارائه شده است و ارتباط بین خوشه‌های تخمین‌شده از یک مدل پیوند شبکه و ترجمه این ماتریس به‌عنوان ویژگی‌های بردارها و یا یک گراف دوبخشی را استفاده می‌کند. این روش از یک الگوریتم خوشه‌بندی معمولی درون یک ماتریس مشابهت استفاده می‌کند تا نتایج خوشه نهایی را تولید کند. روش دیگر خوشه بندی تقسیم تجمعی (DICLENS) است که به‌وسیله [32] ارائه شده است. این روش بر اساس درخت مشابهت پوشای کمینه است. در این روش هر گره یک خوشه و هر یال یک مشابهت درون‌خوشه‌ای بین خوشه‌ها را نشان می‌دهد. این روش، یال‌های با کمینه شباهت، برش داده می‌شوند تا ازهم گسیختگی متاخوشه‌ها فراهم شود. به‌طورکامل واضح است که اعضای ساخته‌شده دارای یک تأثیر مستقیم و قوی روی کارایی اجماع دارند؛ بنابراین یک الگوریتم خوشه‌بندی برای ایجاد اعضای خوشه‌ها به‌صورت خودکار به کار می‌رود.

## ۴- روش پیشنهادی

نرمال‌سازی، گام بسیار مهم در تحلیل ژن‌های متمایزکننده بیان‌شده است. روش‌های نرمال‌سازی متفاوتی با توجه به نوع داده‌ها، معرفی شده‌اند. همانند RPKM، میانگین متوسط ترمیم‌شده مقادیر (TMM)، میانگین کوانتوم، RLE و غیره [33]. TMM و RLE دو روش متفاوت برای محاسبه مقیاس‌های پوششی [34] و این‌که روش RLE یک روش قابل مقایسه است [35]. روش نرمال‌سازی RLE، جهت نرمال‌سازی داده‌های میکروآرایه و داده‌های مربوط به مقادیر بیان ژن سلول‌ها یا بافت‌ها با ابعاد بسیار بالا به‌کار برده می‌شود. این داده‌ها به‌طورمعمول از نمونه‌های انسانی متعددی به‌دست آمده‌اند. در این پژوهش برای نرمال‌سازی از روش RLE استفاده شده است [36]. در این روش برای هر ژن J میانگین را برای هر نمونه محاسبه می‌کند؛ یعنی $med(y_{*j})$. به‌طوری‌که $y_{*j}$ مقدار j امین ستون از ماتریس $[y_{ij}]$ است؛ سپس انحراف از میانگین محاسبه می‌شود. به عبارتی $y_{ij} - med(y_{*j})$، جایی‌که $y_{ij}$ لگاریتم بیان‌شده هر ژن j از نمونه i است [37].

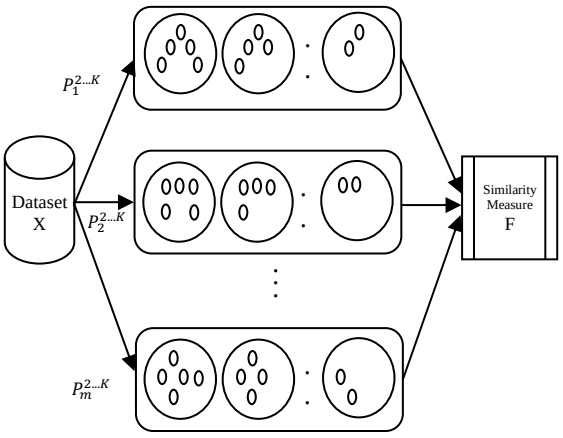
یک روش خوشه‌بندی مناسب نقش کلیدی در یافتن ارتباطات بین سلولی و بین بافتی به عهده دارد. هدف ما طراحی یک روش خوشه‌بندی تجمعی بهبودیافته است که کارایی و اثربخشی بیشتری نسبت به روش‌های انفرادی داشته باشد.

در پژوهش‌های قبلی یک چگالی مبتنی بر الگوریتم خوشه‌بندی برای دسته‌بندی مجموعه‌داده‌های دودویی پیشنهاد شد که تعامل بین سلول‌ها و بافت‌های مختلف را می‌تواند با بیان ژن بین سلول‌ها و یا بافت‌های مختلف نشان دهد. مشکل پژوهش قبلی این بود که تنها از یک معیار شباهت برای خوشه‌بندی داده‌های بیان ژن استفاده می‌شد. این مشکل در این پژوهش با درنظرگرفتن چند روش خوشه‌بندی انفرادی حل‌شده است. ایده اصلی روش پیشنهادی این است که به‌جای محاسبه شباهت بین یک زوج از اشیا در یک روش انفرادی، مشابهت بین اعضای ایجادشده در چندین روش محاسبه و سپس این مشابهت در عضویت بین خوشه‌های جدید مشتق شود. عاقلانه آن است که ما شباهت بین یک زوج از اشیا را با توجه به چندین روش خوشه‌بندی انفرادی ایجاد کنیم. فرایند تجمیع به‌منظور محاسبه شباهت به‌وسیله m روش انفرادی و یک تابع توافقی F ساخته می‌شود؛ بنابراین واضح است که اثربخشی بیشتر و امکان تأثیرگذاری بیشتر برای درنظر گرفتن مشابهت بین خوشه‌های اولیه به‌جای مشابهت اشیا را دارد؛ سپس ما می‌توانیم مفهوم اطلاعات همسایه اشتراکی را از سطح شیء به سطح خوشه ارتقا دهیم؛ بنابراین دو خوشه به‌خوبی به یکدیگر مرتبط هستند اگر اشیای آن‌ها در روش‌های انفرادی بیشتری یکسان باشند. اگر دو خوشه از دو روش خوشه‌بندی انفرادی مختلف دارای اعضای مشترکی باشند، درجه قطعیت اعضای آن‌ها به یک خوشه بیشتر خواهد بود. با توجه به نامشخص‌بودن تعداد خوشه‌های نهایی، اگر دو خوشه دارای درجه بالایی از قطعیت نیز باشند، آن‌ها باید جدا نگه‌داشته شوند. درهرحال، به‌جای استفاده از میزان شباهت تابع توافقی در تشکیل خوشه‌ها، از آن برای تولید خوشه‌های اولیه و ادغام آن‌ها در جهت یافتن خوشه‌بندی پایانی بهره می‌گیریم. در روش پیشنهادی تعداد خوشه‌های بهینه در مرحله ادغام و با درنظرگرفتن یک حد آستانه به‌صورت خودکار ایجاد می‌شود.



[ DOI: 10.52547/jsdp.18.2.97 ]

یکی از مهم‌ترین ویژگی‌های گراف‌های پیچیده وجود ساختارهای اجتماعی است. به‌طور مشخص شناسایی این ساختارها در گراف‌های پیچیده به تحلیل ویژگی‌های ساختاری گراف کمک می‌کند. ساختار کلی روش پیشنهادی در نظر دارد با توجه به اندازه‌گیری مشابهت بین سلول‌ها/بافت‌های بدن انسان، سلول‌ها/بافت‌هایی با بالاترین ارتباط را ارائه دهد. سامانه پیشنهادی شامل دو مرحله کلی است: در مرحله نخست، تمام سلول‌ها یا بافت‌ها با توجه به اندازه‌گیری شباهت در خوشه‌های یکسانی قرار می‌گیرند. در مرحله دوم، معیار شباهت توپولوژیکی به هر یک از خوشه‌ها اعمال شده و تشابه هر سلول/بافت با سایر سلول‌ها/بافت‌ها محاسبه می‌شود. درنهایت، ارتباط هر سلول/بافت با تعدادی از سلول‌ها/بافت‌ها با توجه به بالاترین نمره شباهت ارائه می‌شود. کارایی روش مطرح‌شده با استفاده از مجموعه‌داده‌های واقعی و همچنین مجموعه‌داده فانتوم پنج مورد ارزیابی قرار می‌گیرد. الگوریتم ارائه‌شده صرفاً برای مجموعه‌داده‌های واقعی طراحی‌شده است و حضور نوفه را به‌حساب نمی‌آورد؛ بنابراین در مورد وجود نوفه و تداخل آن در تعیین تعداد خوشه‌ها نگرانی وجود ندارد، چون تعداد خوشه‌ها به‌طور خودکار طی فرایند خوشه‌بندی به‌وسیله ادغام چندین خوشه کوچک استنتاج می‌شود. جزئیات الگوریتم مطرح‌شده به شرح زیر است:




**(شکل-۱): محاسبه شباهت بین هر زوج از اشیا.**
**(Figure-1): Calculates the similarity between each pair of objects.**

## ۴-۱- محاسبه شباهت

ارائه یک تابع توافقی مناسب در تعیین میزان شباهت بین هر زوج از اشیا و تشکیل خوشه‌های نهایی بسیار مؤثر است. شباهت بین دو شئ بر مبنای حضور بیشتر در خوشه‌بندی‌های مختلف محاسبه می‌شود. با توجه به نامشخص‌بودن تعداد خوشه‌ها، مجموعه‌داده به تعداد K-1 خوشه مختلف انجام می‌شود. به‌طوری‌که k = 2,3,...,K و K = N است (N تعداد کل اشیا در مجموعه‌داده). تابع توافقی F یک ماتریس N×N متقارن است که شباهت بین تمام زوج از اشیا را مطابق شکل (۱) نشان می‌دهد.

$f(x_i, x_j)$ به‌منظور محاسبه شباهت بین دو شیء $x_i$ و $x_j$ است و به‌صورت رابطه (۴) تعریف می‌شود:

$$f(x_i, x_j) = \sum_{P=1}^{m} \sum_{k=2}^{N} \sum_{\substack{z=1 \\ \in X}}^{k} \varphi_{x_i,x_j}^{z}(P,k)\,, \quad (x_i, x_j) \qquad (۴)$$

در این رابطه m روش خوشه‌بندی انفرادی به‌منظور افزایش دقت، تابع توافقی روی مجموعه اشیای $X$ به‌کار گرفته می‌شود. تابع $\varphi_{x_i,x_j}^{z}(P,k)$ شباهت بین دو شیء $x_i$ و $x_j$ را در zاُمین خوشه از روش خوشه‌بندی انفرادی P با k خوشه نشان می‌دهد. قرارگرفتن دو شئ در یک خوشه در مراحل مختلف خوشه‌بندی، باعث افزایش شباهت دو شئ خواهد شد. این تابع به‌صورت رابطه (۵) خواهد بود:

$$\varphi_{x_i,x_j}^{z}(P,k) = \begin{cases} \beta^{(N-k)} & if\ (x_i, x_j) \in C_z^P \\ 1 & otherwise \end{cases} \qquad (۵)$$

در این رابطه $\beta$ یک ثابت عددی بین [۰-۱] است که میزان تأثیر شباهت را نسبت به تعداد خوشه‌ها نشان می‌دهد. اهمیت شباهت بین دو شئ در روش‌های خوشه‌بندی انفرادی با تعداد خوشه‌های بالا، بیشتر است. لذا $\beta^{(N-k)}$ به همسایگی اشیا در یک خوشه با تعداد خوشه‌های بیشتر، اهمیت بیشتری می‌دهد. $C_z^P$ اعضاء z-مین خوشه از Pاُمین روش خوشه‌بندی انفرادی را نشان می‌دهد.

به‌منظور استفاده بهتر از شباهت بین زوج‌اشیا در تابع توافقی از نرمال شده ماتریس F به‌صورت رابطه (۶) استفاده می‌شود.

$$f(x_i, x_j) = \frac{f(x_i, x_j)}{\max(F)}\,, (x_i, x_j) \in \mathrm{X} \qquad (۶)$$

شکل (۲) شبه‌کد الگوریتم پیشنهادی در مرحله محاسبه شباهت را نشان می‌دهد.

## ۴-۲- ایجاد خوشه‌های اولیه

به‌منظور ایجاد خوشه‌های اولیه از شباهت‌های بین اشیا در تابع F و یک مقدار آستانه استفاده می‌شود. در این مرحله




[ DOI: 10.52547/jsdp.18.2.97 ]

اشیا با بیشینه شباهت در خوشه‌های یکسانی قرار داده می‌شوند. برای محاسبه بیشینه شباهت اشیا، هر سطر در ماتریس F موردتوجه است. رابطه (۷) به‌منظور محاسبه حداکثر شباهت‌ها ارائه‌شده است:

$$F_{max}^{th_1} = \begin{bmatrix} f_{max}^1 \\ f_{max}^2 \\ \cdots \\ \cdots \\ f_{max}^N \end{bmatrix} = \begin{bmatrix} \max(f_{1,1}, f_{1,2}, \dots, f_{1,N}) \\ \max(f_{2,1}, f_{2,2}, \dots, f_{2,N}) \\ \vdots\vdots\vdots \\ \max(f_{N,1}, f_{N,2}, \dots, f_{N,N}) \end{bmatrix}_{N\times N} \quad (۷)$$

تابع $F_{max}^{th_1}$ شاخصی برای تعریف بیشینه مقادیر شباهت با توجه به مقدار آستانه $th$ است؛ بنابراین $f_{max}^i$ فهرستی از اشیای سطر i-ام را نشان می‌دهد که $f_\Gamma^i \geq th_1$ جایی‌که $\Gamma = 1,2,\dots,N$ است. محتوای نخستین سطر شامل نخستین شیء است؛ بنابراین اشیایی که به $f_{max}^1$ رسیده‌اند، به خوشه نخست $C_1$ تعلق دارند. برای سایر اشیا ($x_i$) از ۲ تا N به‌صورت زیر عمل می‌شود:

۱) تعیین می‌کنید که آیا شئ $x_i$ در خوشه قبلی دیده‌شده است یا خیر؟ اگر چنین است پس شئ در این خوشه قرار نمی‌گیرد.

۲) در غیر این صورت شئ $x_i$ در خوشه بعدی $C_z$ قرار داده می‌شود؛ بنابراین اشیایی که به بیشینه شاخص شباهت برای شئ $x_i$ رسیده‌اند، به خوشه $C_z$ تعلق دارند؛ بنابراین خوشه‌های اولیه $C_z$، برای همه $z \in (1,2,\dots,Z)$ به‌طوری‌که $Z \leq N$ ایجاد می‌شوند.

```
Similarity Calculation Algorithm :
Input : X dataset with N samples and M features; m clustering method and β.
Output : Similarity matrix F_{N×N}
For P = 1 to m Do
  For k = 2 to N Do
    For z = 1 to k Do
      If (x_i, x_j) ∈ Cluster_z from p method Then
        F(x_i, x_j) = β^(N−k)
      Else
        F(x_i, x_j) = 1
      End If
    End For
  End For
End For
F ← Normalized F;
```

**(شکل-۲): الگوریتم محاسبه شباهت.**

**(Figure-2): Similarity Calculation Algorithm**

## ۳-۴- ترکیب خوشه‌های اولیه

این مرحله با Z خوشه اولیه شروع و ترکیب خوشه‌ها بر اساس مشابهت آن‌ها به‌صورت سلسله‌مراتبی تا رسیدن به یک حد آستانه تکرار می‌شود. در هر تکرار دو خوشه با بالاترین مشابهت بین اعضا ترکیب می‌شوند. فرض کنید $C_i$ و $C_j$ دو نمونه از خوشه اولیه با مقادیر $C_i = \{a_1, a_2, \dots, a_{z1}\}$ و $C_j = \{b_1, b_2, \dots, b_{z2}\}$ هستند. فرایند ترکیب دو خوشه بر اساس رابطه (۸) انجام می‌شود:

$$if\ \sigma(C_i, C_j) \geq th_2 \Rightarrow \begin{cases} hence\ merged & true \\ not\ merged & false \end{cases} \quad (۸)$$

پارامتر $th_2$ حد آستانه برای ترکیب خوشه‌ها است و تأثیر و حساسیت بالایی روی کیفیت نتیجه خوشه‌های پایانی و همچنین تعداد خوشه‌ها دارد؛ بنابراین مقدار آن به‌وسیله یک الگوریتم بهینه‌سازی تپه‌نوردی موردمطالعه قرارگرفته است. برای $\sigma$، در هر تکرار تنها دو خوشه با بالاترین شباهت که معیار داده‌شده در رابطه (۸) را ارضا می‌کند، باید به‌وسیله جایگزین‌کردن آن‌ها با یک خوشه جدید $C_k$ ترکیب شوند. میانگین شباهت درون‌خوشه‌ای بین $C_i$ و $C_j$ به‌صورت رابطه (۹) تعریف می‌شود:

$$\sigma(C_i, C_j) = \frac{\sum_{i=1}^{z1}\sum_{j=1}^{z2} F(a_i, b_j)}{z_1 \times z_2} \quad (۹)$$

به‌طوری‌که $F(a_i, b_j)$ شباهت دو شئ $a_i$ از خوشه $C_i$ و $b_j$ از خوشه $C_j$ را نشان می‌دهد. برای خوشه نخست، شباهت میان خوشه با سایر خوشه‌ها به‌صورت $\sigma(C_i) = [\sigma(C_i, C_1), \sigma(C_i, C_2), \dots, \sigma(C_i, C_Z)]$ تعیین می‌شود. بالاترین مقدار از نمونه‌های $\sigma(C_i)$ نشان‌دهنده بیشترین شباهت بین خوشه $C_1$ و $C_z$ برای همه $z \in (2,\dots,Z)$ است؛ بنابراین با فرض بالاترین شباهت نمونه $\sigma(C_i, C_j)$، خوشه‌های $C_i$ و $C_j$ به‌وسیلهٔ جایگزین‌شدن با خوشه جدید $C_k = \{a_1, a_2, \dots, a_{z1}, b_1, b_2, \dots, b_{z2}\}$ ترکیب می‌شوند. ترکیب خوشه‌ها که درنهایت به‌صورت یک خوشه بزرگ نشان داده می‌شوند، راه‌حل نهایی مسأله است. فرایند ترکیب خوشه‌ها تا زمانی که هیچ زوج‌خوشه‌ای شبیه به هم وجود نداشته باشد، ادامه می‌یابد.

## ۴-۴- بهبود خوشه‌های نهایی

آنچه در یک روش خوشه‌بندی حائز اهمیت است، کمینه داده‌هایی است که به‌اشتباه به خوشه‌ای تخصیص داده‌شده است. هدف اصلی از این مرحله برطرف‌کردن خوشه اشیایی است که در مرحله ترکیب خوشه‌ها به‌اشتباه به یک خوشه تخصیص داده شده‌اند. برای شیء $x_i$ از خوشه $z$، اگر میانگین مقدار شباهت $x_i^z$ نسبت به سایر اعضای خوشه $z$ از مقدار آستانه $th_3$ کمتر باشد $f(x_i, x_j) \leq th_3\ (\forall j = 1,2,\dots,z_1)$، شئ از خوشه حذف و یک خوشه جدید به آن تخصیص داده می‌شود. $\psi_z$ اندیس





[ DOI: 10.52547/jsdp.18.2.97 ]

شئ از خوشه $z$ را با کمترین مشابهت نسبت به سایر اعضا نشان می‌دهد و به‌صورت رابطه (۱۰) تعریف می‌شود:

$$\psi_z = arg\ \underset{1\le j\le z_1}{min}\left\{\frac{1}{z_1}\sum_{i=1}^{z_1} f(x_i,x_j)\right\}\ (\forall z = 1,\dots,Z) \qquad (۱۰)$$

به‌طوری‌که $z_1$ تعداد اعضای خوشه $z$ و $arg\ \underset{*}{min}\{*\}$ اندیس شئ با کمترین مقدار شباهت را نشان می‌دهند. شئ با اندیس $\psi_z$ در صورت کمتربودن نسبت به پارامتر حد آستانه $th_3$ از خوشه $z$ حذف می‌شود.

برای تعیین خوشه‌های مناسب برای اشیای حذف‌شده مرحله ترکیب خوشه‌ها دوباره تکرار می‌شود. دو مرحله بهبود خوشه‌های نهایی و ترکیب خوشه‌ها، به‌صورت سلسله‌مراتبی تکرار شده تا زمانی که هیچ تغییری در خوشه‌های نهایی حاصل نشود. شکل (۳) شبه‌کد الگوریتم خوشه‌بندی پیشنهادی را نشان می‌دهد.

Clustering Algorithm :
Input : $F_{N\times N}$; $th_1$; $th_2$; $th_3$
Output : Cluster members, $C_z$ $(\forall j = 1,\dots,Z)$

//Creation of initial clusters :
$F_{max}^{th_1}$= Maximum similarity in the F for each $x_i$ sample, $(\forall i = 1,\dots,N)$.
$z = 1$, $C_z = \{1^{st}\ sample\ and\ f_{max}^1\}$;
**For** i = 2 to N **Do**
  **If** sample $x_i$ has not included in any previous clusters **Then**
    $z = z + 1$;
    $f_{\overline{max}}^i$ = samples of $f_{max}^i$ has not included in any previous clusters.
    $C_z = \{i^{st}\ sample\ and\ f_{\overline{max}}^i\}$;
  **End If**
**End For**
**Repeat** until no change in the final cluster
  //Merging of the initial clusters :
  Z ← Total number of clusters;
  $\sigma^{max} = \infty$;
  **While** $\sigma^{max} \ge th^2$ **Do**
    **For** z = 1 to Z **Do**
      $\sigma(C_z, C_j) = eq.(9),\ (\forall j = 1,\dots,Z)$
    **End For**
    $[C_i, C_j, \sigma^{max}] = \max_{1\le i\le Z} \sigma(C_i, C_j),\ (\forall j = 1,\dots,Z)$
    $C_{new} = \{C_i, C_j\}$;
    Remove $C_i$ and $C_j$ clusters.
  **End While**
  //Improve the final cluster :
  **For** z = 1 to Z **Do**
    $\psi = eq.(10)$;
    **If** $\psi \le th_3$ **Then**
      Remove $\psi$ sample from $z$ cluster.
      $C_{new} = \{C_\psi\}$;
    **End If**
  **End For**
**End Repeat**

**(شکل-۳): الگوریتم خوشه‌بندی پیشنهادی**
**(Figure-3): Proposed Clustering Algorithm**

## ۵-۴- تئوری گراف در تشخیص ارتباطات بین سلولی/بافتی

یک گراف $G = (V, E, W)$ مجموعه‌ای از رئوس V و مجموعه‌ای از یال‌های E است؛ به‌طوری‌که یال به یک جفت از رئوس می‌پیوندد. $W(e)$ برای هر یال $e \in E$ متناظر با وزن است. در اینجا، به‌طورکلی G همیشه یک گراف بدون جهت وزن‌دار خواهد بود؛ بنابراین $e(V_i, V_j)$ و $e(V_j, V_i)$ نشان‌دهنده یال مشابه در گراف هستند. رئوس گراف نماینده سلول‌ها/بافت‌ها و یال‌ها ارتباط بین سلول‌ها/بافت‌ها در گراف G را نشان می‌دهند. میزان شباهت در گراف وزن یال‌ها را مشخص می‌کند؛ بنابراین $W = F$.

ماتریس مجاورت A از گراف G یک ماتریس $N \times N$ متقارن با برچسب‌های رئوس گراف و با مقادیر صفر و یک است. مقدار صفر و یک در موقعیت $e(V_i, V_j)$ با توجه به این‌که آیا $V_i$ و $V_j$ مرتبط هستند یا نه تعیین می‌شود. ارتباط بین رئوس (سلول‌ها/بافت‌ها) با توجه به معیار تعریف‌شده در رابطه (۱۱) تعیین می‌شود:

$$A_{i,j} \Rightarrow \begin{cases} 1 & W_{i,j} \ge th_4 \\ 0 & otherwise \end{cases} \qquad (۱۱)$$




[ DOI: 10.52547/jsdp.18.2.97 ]

به‌طوری‌که $W_{i,j}$ وزن بین رئوس $V_i$ و $V_j$ را نشان می‌دهد و $th_4$ حد آستانه وزن برای تشکیل ماتریس مجاورت گراف G است. هدف این مرحله یافتن فهرستی از سلول‌ها یا بافت‌های مشابه برای یک سلول یا بافت $V_c$ ($\forall c = 1,2,\dots,N$) است.

سامانه‌های توصیه‌گر دوست (FRS) برنامه‌های بسیار محبوب در خدمات شبکه‌های اجتماعی برخط هستند که به کاربران کمک می‌کند تا دوستان جدید با علایق مشابه بیابند. ساختار کلی سامانه در نظر دارد با توجه به اندازه‌گیری مشابهت برای محاسبه شباهت بین کاربران و هر کاربر، دوستان با بالاترین شباهت را ارائه دهد. در این مقاله، یک FRS با استفاده از ترکیب «شباهت توپولوژیکی ساختار گراف» و «اطلاعات خوشه‌بندی سلول‌ها/بافت‌ها» به‌منظور معرفی سلول‌ها/بافت‌های مرتبط با ویژگی‌ها و کاربردهای مشابه معرفی شده است. درواقع از این روش برای شناسایی جوامع (سلول‌ها یا بافت‌های مشابه) بهره می‌گیریم.

اگرچه تعریف کامل و مستقلی برای یک جامعه وجود ندارد، اما بیشتر شبکه‌های دنیای واقعی ساختارهای جامعه را نشان می‌دهند. یک جامعه (که همچنین خوشه، ماژول یا گروه نامیده می‌شود) یک زیرمجموعه از گره‌ها است که وجه تشابه زیادی با یکدیگر دارند [38]. جوامع در شبکه‌ها گروهی از گره‌ها هستند که بین آن‌ها ارتباطات کم است؛ درحالی‌که درون آن اتصالات چگال است [39]. تشخیص ساختار جامعه در شبکه‌های پیچیده توجه زیادی از پژوهش‌گران را به خود جلب کرده است؛ زیرا ساختار جامعه در شبکه‌های دنیای واقعی مثل ارتباط بین بیان ژن‌ها، پروتئین‌ها [40] و راه‌انداز ژن برای تشخیص انواع بیماری، شبکه‌های اجتماعی [39]؛ تجزیه‌وتحلیل سند، که در آن رأس‌ها نشان‌دهنده نویسندگان یا اسناد هستند [41] و غیره به‌کار برده می‌شوند.

سامانهٔ پیشنهادی این پژوهش با استفاده از یک معیار شباهت جدید که ترکیبی از الگوریتم FriendLink [42] و Katz [43] به هر یک از خوشه‌ها اعمال شده و تشابه هر سلول/بافت با تمام سلول‌ها/بافت‌های غیر مجاور محاسبه و درنهایت، به هر سلول/بافت با توجه به بالاترین نمره شباهت، تعدادی سلول/بافت به‌عنوان سلول‌ها/بافت‌های مرتبط ارائه می‌شود. الگوریتم FriendLink و Katz یک روش برای پیش‌بینی پیوندهای جدید در گراف است که با پیمودن تمام‌مسیرها با طول محدود عمل می‌کند، به‌طوری که از همه جهت‌ها با طول‌های مختلف در شبکه استفاده نمی‌کند و به‌جای آن از مسیرهایی با بیشینه طول $l \in [2, L]$ بین رئوس استفاده می‌کند. برای محاسبه شباهت بین سلول‌ها یا بافت‌ها در داخل هر خوشه، ابتدا ماتریس مجاورت برای هر خوشه ایجاد می‌شود. رابطه (۱۲) ماتریس مجاورت خوشه $z$ را نشان می‌دهد:

$$A_{i,j}(z) \Rightarrow \begin{cases} A_{i,j} & V_i \in C_z \text{ and } V_j \in C_z \\ 0 & otherwise \end{cases} \quad (۱۲)$$

به‌طوری‌که $C_z$ فهرست رئوس گراف در خوشه $z$ را مشخص می‌کند.

ماتریس مجاورت گراف G هنگامی‌که به توان n می‌رسد، تعدادی از مسیرهای با طول n بین دو رأس را نشان می‌دهد. در این مقاله، با توجه به در نظر گرفتن $L = 4$، برای هر خوشه، ماتریس مجاورت به توان‌های ۲، ۳ و ۴ می‌رسد. برای مثال $A^2_{i,j}(z)$ تعداد مسیرهای غیر مجاور بین دو رأس $V_i$ و $V_j$ با طول مسیر ۲ را در خوشه $z$ مشخص می‌کند.

در مرحله بعد برای تعیین شباهت بین سلول‌ها/بافت‌ها، یک معیار ترکیبی جدید به نام $KFL$ پیشنهاد می‌دهیم که ترکیبی از دو معیار شباهت Katz و FriendLink است. با درنظرگرفتن این دو معیار و نیز گره‌های مشترک رأس میانی با دو رأس اصلی محاسبه می‌شود. معیار شباهت پیشنهادی بین دو گره $v_x$ و $v_y$ به‌صورت رابطه (۱۳) محاسبه می‌شود.

$$KFL(v_x, v_y) = \sum_{i=2}^{l} \frac{1 - (\left|paths^i_{v_x,v_y}\right| \times \alpha^i)}{i-1} \cdot \frac{\left|paths^i_{v_x,v_y}\right|}{\prod_{j=2}^{i}(n-j)} \quad (۱۳)$$

در این رابطه $l$ بیشینه طول مسیر موردبررسی است. α پارامتر ثابتی بین [0,1] است که با توجه به اینکه به توان i می‌رسد، مقدار آن به نسبت طول ارتباطات متغیر است. هرچه طول ارتباط‌ها بزرگ‌تر شود، ارزش این ترم به‌دلیل کم‌اهمیت‌شدن و کاهش احتمال ارتباط، کم خواهد شد. $N$ تعداد کل گره‌ها و $\left|paths^i_{v_x,v_y}\right|$ تعداد مسیرهای بین $v_x$ و $v_y$ با طول i است. در ارزیابی‌های انجام‌شده معیار $KFL$ با مقدار پارامترهای $\alpha = 0.05$ و $l = 3$ عملکرد بهتری دارد؛ درنهایت، با استفاده از ماتریس شباهت $sim$ مبتنی بر الگوریتم ابتکاری $KFL$، $k$ تعداد سلول/بافت با بالاترین میزان شباهت به‌عنوان سلول‌ها/بافت‌های مرتبط به سلول/سلول هدف توصیه می‌شوند. رابطه (۱۴) سلول‌ها/بافت‌های مرتبط با سلول/بافت $x_i$ را نشان می‌دهد. به‌طوری‌که $rank[*]$ رتبه شباهت را با توجه به سلول/بافت $x_i$ و تمام سلول‌ها/بافت‌های غیر مجاور آن را در خوشه





[ DOI: 10.52547/jsdp.18.2.97 ]

جاری نشان می‌دهد. شکل (۴) شبه‌کد الگوریتم پیشنهادی تئوری گراف در تشخیص ارتباط‌های بین سلولی را نشان می‌دهد. تشخیص ارتباط‌های بین بافتی نیز به‌صورت مشابه انجام می‌شود:

$$com(x_i, x_j) \Rightarrow \begin{cases} 1 & rank[sim(x_i, x_j)] \geq k \\ 0 & otherwise \end{cases}, (\forall j = 1,2, \dots, \mathrm{N}) \quad (۱۴)$$

## ۵- نتایج و آزمایش‌ها

در این بخش عملکرد الگوریتم پیشنهادی به‌منظور کشف روابط بین سلول‌ها/بافت‌ها با استفاده از پنج مجموعه‌داده واقعی از مخزن یادگیری ماشین UCI مورد ارزیابی قرارگرفته است. Iris، Wine، Vowel، Glass و Seed مجموعه‌داده‌هایی هستند که ما استفاده کرده‌ایم. جدول (۱) جزئیات این مجموعه‌داده‌ها را نشان می‌دهد. برخی از مجموعه‌داده‌ها دارای مقادیر گم‌شده در برخی از اشیا هستند که حذف شده است.

Cell-to-Cell Communication with Graph topological Algorithm
Input : Cluster members, $C_z\ \forall j = 1, \dots, Z$ and Adjacency matrix $A$
Output : $com(x_i, x_j)\ \forall x_i, x_j \in X$

$A(z)$ : Adjacency matrix calculate for the $z$ cluster $\forall z = 1, \dots, Z$.
**For** z = 2 to Z **Do**
  **For** i = 2 to N **Do**
    $sim(v_i, v_j)$ $= eq.(13),$ $(\forall j = 1,2, \dots, \mathrm{N})$ //KFL algorithm is applied to the A(z).
    $com(x_i, x_j) = eq.(14)$ // For each cell k number of cells with the highest similarity score obtained are offered as a cell-to-cell communication.
  **End For**
**End For**

**(شکل-۴): الگوریتم پیشنهادی تئوری گراف در تشخیص ارتباط‌های بین سلولی و بین بافتی.**

**(Figure-4): The proposed algorithm for graph theory in detecting intercellular and interconnected relationships.**

**(جدول-۱): مجموعه داده‌های استفاده‌شده در آزمایش‌ها.**

**(Table-1): The data set used in experiments**

| مجموعه داده | تعداد نمونه‌ها | تعداد ویژگی‌ها | تعداد خوشه‌ها | تعداد نمونه در هر خوشه | نوع ویژگی‌ها |
|---|---|---|---|---|---|
| Iris | ۱۵۰ | ۴ | ۳ | ۵۰-۵۰-۵۰ | (۴ R) |
| Wine | ۱۷۸ | ۱۳ | ۳ | ۴۸-۷۱-۵۹ | (۱۱ R), (۲ I) |
| Thyroid | ۲۱۵ | ۵ | ۳ | ۳۰-۳۵-۱۵۰ | (۵ R) |
| Glass | ۲۱۴ | ۹ | ۶ | -۱۳-۱۷-۷۸-۷۰ ۲۹-۹ | (۹ R) |
| Bcw | ۶۸۳ | ۹ | ۲ | ۲۳۹-۴۴۴ | (۹ I) |

خوشه‌بندی تجمعی، روشی جدید در خوشه‌بندی است که از ترکیب نتایج روش‌های خوشه‌بندی مختلف به‌دست می‌آید. صحت، درستی و پایداری از مشخصه‌های مهم یک سامانه خوشه‌بندی تجمعی در مقایسه با روش‌های کلاسیک خوشه‌بندی است. در این مقاله از سه الگوریتم K-Means، FCM و K-Medoids برای خوشه‌بندی اسناد استفاده شده است. الگوریتم پیشنهادی با ترکیب و استفاده از نتایج این سه روش سعی در بهبود نتایج بر روی مجموعه‌داده‌ها دارد.

با توجه به استفاده از پارامترهای حد آستانه در الگوریتم پیشنهادی و تأثیر آن‌ها در تخمین تعداد صحیح خوشه‌ها به‌صورت خودکار، انتخاب بهترین مقادیر برای این پارامترها یک چالش بزرگ برای ما است. راه‌حلی که برای این مشکل ارائه شده، بهینه‌سازی مقدار پارامترها با استفاده از الگوریتم تپه‌نوردی است. تپه‌نوردی یک فن بهینه‌سازی متعلق به خانواده الگوریتم‌های جستجوی محلی است؛ یک فن تکرارشونده که با یک راه‌حل دلخواه شروع به کارکرده و سپس تلاش می‌کند تا با تغییر بر روی یک عنصر از راه حل، به پاسخ بهتری دست پیدا کند. اگر این تغییر منجر به ایجاد یک‌راه حل بهتر شود، تغییر دیگری بر روی این راه حل جدید انجام خواهد گرفت. این روال تا زمانی که بهبود بیشتری در راه حل میسر نباشد ادامه می‌یابد.

با توجه به استفاده از چندین مجموعه‌داده واقعی در آزمایش‌ها، $fitness_{HC}$ عملکرد تخمین مقادیر آستانه $th_1$، $th_2$، $th_3$ و $th_4$ به‌وسیله الگوریتم تپه‌نوردی را نشان می‌دهد و با رابطه (۱۵) تعریف می‌شود. این رابطه میانگین دقت خوشه‌بندی انجام‌شده را به‌وسیله روش پیشنهادی روی تمام مجموعه‌داده‌های استفاده‌شده محاسبه می‌کند:

$$fitness_{HC} = \frac{1}{H}\sum_{h=1}^{H} FC(th_i, ds_h)\ , (\forall i = 1,2,3,4) \quad (۱۵)$$

به‌طوری‌که $H$ تعداد مجموعه‌داده‌های استفاده‌شده و $ds_h$ hاُمین مجموعه‌داده را نشان می‌دهد. تابع $FC$ نقش الگوریتم پیشنهادی را دارد و دقت خوشه‌بندی را با ورودی‌های تعیین‌شده محاسبه می‌کند.

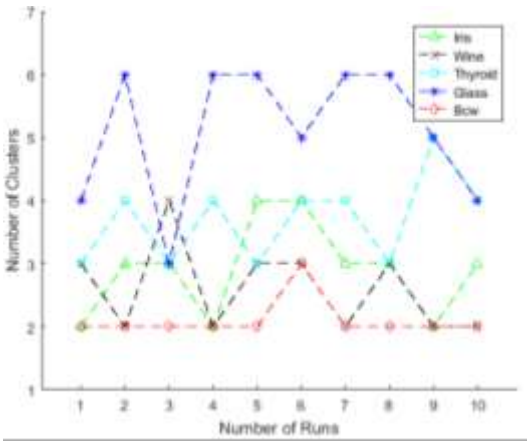
در آزمایش‌ها مقادیر زیر برای پارامترها در نظر گرفته شده و سپس چارچوب خوشه‌بندی تجمعی پیشنهادی دنبال می‌شود:

$th_1 = 0.75, th_2 = 0.0034, th_3 = 0.13, th_4 = 0.5, L = 4, \beta = 0.95$



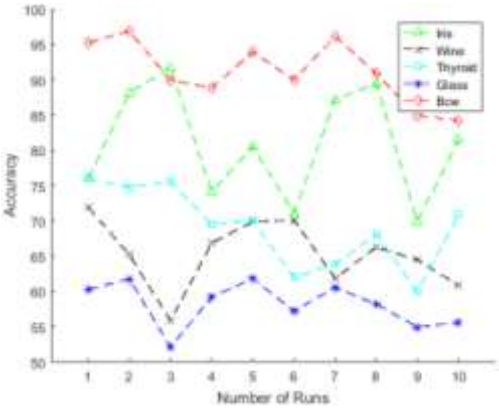
[ DOI: 10.52547/jsdp.18.2.97 ]

در تمام آزمایش‌های انجام‌شده از همه نمونه‌ها در مجموعه‌داده‌ها استفاده شده است. هدف اصلی این آزمایش‌ها، آزمایش‌کردن کارایی الگوریتم خوشه‌بندی پیشنهادی و همچنین مشاهده چگونگی تأثیر الگوریتم ما و مقایسه آن با سایر الگوریتم‌های رقیب است. شکل (۵) تعداد خوشه‌های کشف‌شده به‌وسیله الگوریتم پیشنهادی در همه مجموعه‌داده‌های آزمایش‌شده در ده اجرا را نشان می‌دهد. نتیجه به‌دست‌آمده حاکی از آن است که خوشه‌ها در بیشتر مجموعه‌داده‌ها درست و یا نزدیک به مقدار واقعی تخمین زده شده اما همچنان ناپایدار و قابل‌تغییر در ده اجرا است.




**(شکل-۵): تعداد خوشه‌های تولیدشده در ده اجرای مجزا.**
**(Figure-5): The number of clusters in 10 separate runs**

دقت نتایج روش پیشنهادی مطابق با شکل (۵) از ده اجرای مجزا برای همه مجموعه‌داده‌های آزمایش‌شده در شکل (۶) نشان داده شده است. نتیجه به‌دست‌آمده حاکی از عملکرد خوب روش پیشنهادی در تعیین خوشه‌های درست در بیشتر مجموعه‌داده‌ها است. به‌منظور بررسی دقیق‌تر روش پیشنهادی ما نتایج را با چندین روش مشابه مقایسه و عملکرد آن را تجزیه‌وتحلیل می‌کنیم.




**(شکل-۶): دقت خوشه‌بندی تولیدشده در ده اجرای مجزا.**
**(Figure-6): The accuracy of the cluster in 10 separate runs.**

این الگوریتم‌ها شامل روش همکاری (CO) با استفاده از پیوند میانگین [28] و ONCE که آن‌هم از پیوند میانگین استفاده می‌کند [23] و MCLA [29] است.

روش همکاری پیوندی (CO) از مشکل تطبیق برچسب به‌وسیله نگاشت اعضای اجماع به داخل یک نمایش جدید که در آن ماتریس مشابهت بین یک زوج از اشیا با توجه به چگونگی تعداد اعضای یک زوج مخصوص با یکدیگر خوشه می‌شوند، جلوگیری می‌شود. در این کار، افراز نهایی به‌وسیله به‌کارگیری ادغام الگوریتم‌های خوشه‌بندی سلسله‌مراتبی پیوند منفرد یا میانگین برای ماتریس تطبیق به‌دست می‌آید. در روش ONCE یک مسأله توافق غیرقطعی بین اعضا مطرح شده است. روش جدید خوشه‌بندی تجمعی اشیای همسایه یا خوشه‌بندی تجمعی نامیده شده است. هدف از این روش استفاده از رابطه همسایگی بین زوج‌اشیا است که این ارتباط بین زوج‌ها به‌خوبی برای محاسبه ماتریس مشابهت به کار می‌رود. در MCLA ابتدا شباهت بین زوج خوشه‌ها از جنبه اشتراک اشیای بین آن‌ها را تعریف می‌کند. این کار از طریق توسعه شاخص Jaccard انجام، سپس گراف طوری ساخته می‌شود که گره‌ها، خوشه‌ها و یال‌ها، شباهت بین زوج‌خوشه‌ها را بیان می‌کنند. ما الگوریتم را ده مرتبه اجرا و هر مرتبه کارایی را به‌وسیله MMI, ARI اندازه‌گیری کردیم. هنگامی‌که NMI, ARI برای ارزیابی نتایج خوشه‌بندی به کار می‌روند، یکی از افرازهای خوشه‌بندی باید به‌عنوان افراز درست داده‌ها در نظر گرفته شود که با $p^t$ نمایش داده می‌شود. درعمل به‌طورمعمول برچسب‌های رده در نظر گرفته می‌شود که برای بقیه پاسخ‌هایی که می‌تواند برای تأیید کیفیت (دقت) نتیجه خوشه‌بندی استفاده شود، وجود ندارد. افراز دیگر نتیجه خوشه‌بندی تجمعی است که نیاز به محاسبه $p^*$ دارد. ARI نسخه وابسته به شاخص (RI) است که به‌صورت رابطه (۱۶) تعریف می‌شود:

$$ARI(P^*, P^t) = \frac{RI(P^*, P^t) - Expected[RI]}{1 - Expected[RI]} \qquad (16)$$

که در آن RI از رابطه (۱۷) محاسبه می‌شود:

$$RI(P^*, P^t) = \frac{n_{11} + n_{00}}{n_{00} + n_{11} + n_{10} + n_{01}} \qquad (17)$$

در این رابطه $n_{11}$ تعداد زوج شیء اختصاص‌داده به خوشه‌های مشابه در هردوی $p^*$ و $p^i$ است. $n_{00}$ تعداد زوج شیء اختصاص داده‌شده به خوشه‌های متفاوت در $p^*$ و $p^i$ است. $n_{10}$ تعداد زوج‌اشیای اختصاص داده‌شده به همان




[ DOI: 10.52547/jsdp.18.2.97 ]

خوشه در $p^*$ و خوشه‌های متفاوت در $p^t$ است. $n_{01}$ تعداد زوج‌اشیای اختصاص داده‌شده به خوشه‌های متفاوت $p^*$ و همان خوشه در $p^t$ است. بیشینه مقدار ARI برابر با یک است. این بدان معنا است که $p^*$ با $p^i$ یکسان (همگون) است و یک مقدار مورد انتظار صفر برای خوشه‌بندی مستقل دارد.

NMI بر اساس میانگین اطلاعات دوبه‌دو بین هر زوج از خوشه‌ها و برچسب‌های رده، محاسبه می‌شود. در نظر بگیرید که $p^*$ نتیجه آخر خوشه‌بندی اجماع و$p^t$ خوشه‌بندی ground-truth برای مجموعه‌داده X باشد. NMI دو افراز به‌صورت رابطه (۱۸) محاسبه می‌شود:

$$NMI(P^*,P^t)=\frac{\sum_{i=1}^{k^*}\sum_{j=1}^{k^t}n_{ij}log\left(\frac{n.n_{ij}}{n_i.n_j}\right)}{\left(\sum_{i=1}^{k^*}n_i\ log\left(\frac{n_i}{n}\right)\right)\left(\sum_{i=1}^{k^t}n_i\ log\left(\frac{n_j}{n}\right)\right)} \qquad (۱۸)$$

**(جدول-۲): میانگین کارایی و انحراف استاندارد معیار ARI برای ۱۰ اجرا.**

**(Table-2): Average performance and standard deviation of ARI standard for 10 executions.**

| مجموعه داده | Co-Average | ONCE-Average | MCLM | روش پیشنهادی |
|---|---|---|---|---|
| Iris | ۰/۶۶۹±۰/۰۶۵ | ۰/۶۷۴±۰/۰۵۷ | ۰/۷۲۲±۰/۰۴۳ | ۰/۶۸۱±۰/۰۴۲ |
| Wine | ۰/۳۲۴±۰/۰۴۵ | ۰/۳۴۴±۰/۰۶۰ | ۰/۳۹۳±۰/۰۰۸ | ۰/۴۲۲±۰/۰۱۹ |
| Thyroid | ۰/۲۲۵±۰/۱۷۵ | ۰/۱۸۹±۰/۱۲۱ | ۰/۴۴۸±۰/۱۱۹ | ۰/۳۸۳±۰/۰۴۵ |
| Glass | ۰/۲۶۵±۰/۰۰۶ | ۰/۲۵۶±۰/۰۰۸ | ۰/۱۵۲±۰/۰۲۲ | ۰/۲۷۱±۰/۰۰۹ |
| Bcw | ۰/۸۶۶±۰/۰۱۸ | ۰/۸۶۰±۰/۰۱۶ | ۰/۸۶۴±۰/۰۱۴ | ۰/۸۶۸±۰/۰۱۲ |
| Ave-P | ۰/۴۱۴ | ۰/۴۰۳ | ۰/۴۳۳ | ۰/۴۱۷ |
| Ave-C | ۰/۰۴۵ | ۰/۰۳۷ | ۰/۰۳۵ | ۰/۰۴۲ |

بیشینه مقدار NMI برابر یک است که به این معنی است که $p^*$ همگون با $p^t$ و کمینه مقدار آن برابر صفر است، که مربوط به زمانی است که $p^*$ به‌طورکامل با $p^t$ متفاوت است. جدول (۲) میانگین کارایی اندازه‌گیری‌شده به‌وسیله شاخص ARI از طریق انحراف استاندارد هر مجموعه‌داده و میانگین کارایی اعضای تولیدشده را نشان می‌دهد. این اندازه‌گیری شامل میانگین کارایی (Ave-P) هر یک از روش‌های اجماع روی پنج مجموعه‌داده و همچنین میانگین سازگاری (Ave-C) است. نتایج نشان می‌دهد که الگوریتم پیشنهادی اغلب بهتر از الگوریتم‌های خوشه‌بندی تجمعی بررسی‌شده کار می‌کند. این موضوع به‌خصوص در مورد مجموعه‌داده‌های Wine و Bcw صحیح است. این در حالی است که برای Iris و Thyroid نتایجی نزدیک به بالاترین کارایی در این مجموعه‌داده‌ها ارائه می‌دهد. هرچند نتیجه روی Bcw نشان می‌دهد که روش پیشنهادی برای مجموعه‌داده‌های به‌نسبه بزرگ مناسب است. درخصوص سازگاری روش پیشنهادی، در مورد دو مجموعه‌داده Bcw, Glass بسیار سازگارتر است، درحالی‌که روی Wine در مقایسه با سایر الگوریتم‌ها از نظر سازگاری در مکان دوم قرار دارد. به‌طور میانگین، دو الگوریتم نتایج بسیار نزدیکی از نظر سازگاری داشتند که عبارت‌اند از MCLA، ONCE. نتایج این دو الگوریتم به‌ترتیب عبارت‌اند از ۰/۰۳۵ و ۰/۰۳۷.

نتایج آزمایش‌های مشابه با استفاده از شاخص NMI در جدول (۳) نشان داده شده است. روش پیشنهادی دارای بالاترین کارایی روی دو مجموعه‌داده Iris و Bcw است. هرچند در Wine و Glass این الگوریتم نتایجی بسیار نزدیک به بالاترین کارایی دارد. در اینجا، چندین نتیجه جالب وجود دارد. نخست این‌که کارایی روش پیشنهادی بهتر از ONCEو CO در پنج مجموعه‌داده است، درحالی‌که عملکرد آن روی سایر مجموعه‌داده‌ها بسیار نزدیک به آن‌ها است. دوم این‌که روش پیشنهادی همان کارایی الگوریتم‌های CO و MCLA را در مجموعه‌داده Bcw دارد و این‌که بالاترین دقت مربوط به همین مجموعه‌داده است. درخصوص سازگاری اندازه‌گیری‌شده به‌وسیله انحراف استاندارد، روش ما در مقایسه با سایر الگوریتم‌ها روی مجموعه‌داده thyroid دارای بیشترین سازگاری است و نتایج آن بسیار نزدیک به سازگارترین الگوریتم در مجموعه داده‌های دارای بیشترین استفاده نظیر Wine و Bcw است. با ملاحظه میانگین کارایی اعضای ایجادشده، خواهیم یافت که عملکرد همه روش‌های تجمعی به‌اندازه میانگین اعضای روی همه مجموعه‌داده‌ها است.

**(جدول-۳): میانگین کارایی و انحراف استاندارد معیار NMI برای ۱۰ اجرا.**

**(Table-3): Average performance and standard deviation of NMI criteria for 10 executions.**

| مجموعه داده | Co-Average | ONCE-Average | MCLM | روش پیشنهادی |
|---|---|---|---|---|
| Iris | ۰/۷۵۳±۰/۰۱۷ | ۰/۷۴۹±۰/۰۲۷ | ۰/۷۵۵±۰/۰۳۷ | ۰/۷۶۱±۰۲/۰۹ |
| Wine | ۰/۴۰۶±۰/۰۱۰ | ۰/۱۴۵±۰/۰۲۲ | ۰/۴۱۵±۰/۰۰۵ | ۰/۴۱۹±۰/۰۱۵ |





[ DOI: 10.52547/jsdp.18.2.97 ]

| | | | | |
|---|---|---|---|---|
| Thyroid | ۰/۲۹۳±۰/۰۷۷ | ۰/۲۵۰±۰/۰۶۶ | ۰/۳۵۶±۰/۰۴۸ | ۰/۳۱۶±۰/۰۴۹ |
| Glass | ۰,۴۴۱±۰,۰۱۸ | ۰/۴۴۹±۰/۰۱۶ | ۰/۳۰۷±۰/۰۳۲ | ۰/۴۵۰±۰/۰۱۷ |
| Bcw | ۰/۷۷۳±۰/۰۲۴ | ۰/۷۶۵±۰/۰۲۱ | ۰/۷۷۰±۰/۰۱۹ | ۰/۷۷۵±۰۲/۰۰ |
| Ave-P | ۰/۴۸۸ | ۰/۴۸۱ | ۰/۴۷۷ | ۰/۵۰۱ |
| Ave-C | ۰/۰۲۳ | ۰۲۱/۰ | ۰/۰۲۲ | ۰/۰۳۷ |

درهرحال، الگوریتم ما بهترین الگوریتم از نظر مقایسه با سایر الگوریتم‌ها است. با توجه به نتایج شاخص NMI، این نتایج بسیار نزدیک به نتایج ارائه‌شده به‌وسیلهٔ ARI است. نتایج حاصل از ارزیابی الگوریتم خوشه‌بندی پیشنهادی، کارایی این الگوریتم را نشان می‌دهد. در ادامه، از این الگوریتم در جهت خوشه‌بندی مجموعه‌داده فانتوم و استخراج ارتباطات بین سلولی یا بین بافتی استفاده می‌شود. در این مقاله از داده‌های فانتوم۵ استفاده شده است. در داده‌های فانتوم ۱۸۲۹ نمونه با ۲۰۱۸۰۲ راه‌انداز وجود دارد. ویژگی‌ها در فانتوم راه‌اندازها هستند که درواقع حاوی اطلاعاتی در مورد ژن‌ها است که منجر به تولید آن شده است. هدف این پژوهش این است که مشخص شود چه ژن‌هایی برای چه بیماری‌هایی Expression شده‌اند. درواقع هدف شناسایی سلول‌ها یا بافت‌هایی است که در یک یا چند بیماری Gene Expression مشابهی دارند؛ بنابراین ویژگی‌های مجموعه‌داده راه‌اندازها هستند که از قسمت خاصی از یک کروموزوم گرفته شده است و میزان Gene Expression را نشان می‌دهد. بررسی‌های انجام‌شده نشان می‌دهد که مقادیر برخی از راه‌اندازها (در برخی از ژن‌ها) در تمام سلول‌ها/بافت‌ها به میزان بسیار کمی Gene Expression شده‌اند؛ بنابراین با درنظرگرفتن یک حد آستانه می‌توان تنها راه‌اندازهایی از مجموعه‌داده را انتخاب کرد که در آن‌ها Gene Expression به میزان مناسبی باشد. یک مقدار حد آستانه در نظر گرفته‌شده است و تنها راه اندازهایی که دست‌کم در یکی از سلول‌ها یا بافت‌ها دارای مقدار Gene Expression بالاتر از حد آستانه باشد، انتخاب می‌شوند. تعداد کل راه‌اندازها با درنظرگرفتن آستانه هزار (فیلتر) به ۸۶۴۲۷ کاهش می‌یابد که در ۷۰۲ نمونه برای سلول‌ها و ۱۲۵ نمونه برای بافت‌ها قرار دارند. از این تعداد نمونه، ۱۰۸ سلول منحصربه‌فرد و چهل بافت منحصربه‌فرد وجود دارد که از مجموعه‌داده انتخاب می‌شوند؛ بنابراین مجموعه‌داده فانتوم مورد آزمایش دارای ۷۰۲ نمونه برای سلول‌ها، ۱۲۵ نمونه برای بافت‌ها با ۸۶۴۲۷ راه‌انداز است که برای انجام آزمایش‌ها مشخص‌شده‌اند.

راه‌اندازهایی از مجموعه‌داده که تمامی مقادیر Expression آن‌ها از عدد آستانه کوچک‌تر است، برای یافتن ارتباط بین سلول‌ها/بافت‌ها مفید نیست. دلیل این امر نزدیک‌بودن میانگین مقادیر این راه‌اندازها در تمام نمونه‌ها با بیشینه مقدار راه‌انداز در همان نمونه است (انحراف معیار). این امر نشان می‌دهد که میزان Expression در این بخش از کروموزوم در تمام سلول‌ها/بافت‌ها به‌طورتقریبی یکسان است. ارتباط دو سلول یا دو بافت زمانی به وجود می‌آید که در تعداد از سلول‌ها یا بافت‌ها مقدار راه‌اندازها به میزان قابل توجهی Expression شده باشند. تجزیه‌وتحلیل نشان می‌دهد که مقادیر Expression کمتر از هزار با توجه به بیشینه مقدار Expression در کل مجموعه‌داده که گاهی به حدود پنج‌هزار نیز می‌رسد برای یافتن ارتباط بین سلول‌ها یا بافت‌ها مفید نیست. جدول (۴) ارتباطات بین سلولی را با توجه به بخش ۴-۵ نشان می‌دهد. ارتباط بین سلول‌ها با توجه به ده مورد از قوی‌ترین ارتباط‌ها بیان شده است. نماد $C_i$ معرف سلول i-ام است که در پیوست یک آورده شده است. میزان شباهت بین دو سلول با رابطه (۶) محاسبه شده است. تعداد راه‌اندازهای بیان‌شده در دو سلول با توجه به تعداد مشابهت‌ها بین دو نمونه (سلول) اندازه‌گیری شده است. ژنی با بیشترین راه‌اندازهای بیان‌شده نشان‌دهنده منطقه(مکان) ارتباط بین دو سلول است. به‌منظور بررسی حد آستانه مناسب جهت فیلترکردن برخی از ارتباطات راه اندازها و استخراج ارتباط‌های بین سلولی مناسب ما با چهار حد آستانه مختلف ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ نتایج را بررسی می‌کنیم.

**(جدول-۴): ارتباط‌های بین سلولی استخراج‌شده از مجموعه داده فانتوم با استفاده از الگوریتم پیشنهادی.**

**(Table-4): Intercellular connections extracted from the phantom dataset using the proposed algorithm.**

| نام سلول | میزان شباهت | تعداد راه‌اندازهای بیان‌شده | ژن با بیشترین راه‌انداز بیان‌شده | نام سلول | نام سلول | میزان شباهت | تعداد راه‌اندازهای بیان‌شده | ژن با بیشترین راه‌انداز بیان‌شده | نام سلول |
|---|---|---|---|---|---|---|---|---|---|
| حد آستانه ۵۰۰ | | | | | حد آستانه ۱۰۰۰ | | | | |
| $C_7$ | ۰/۹۱۱ | ۶۴۵۸۰ | ABLIM1 | $C_1$ | $C_7$ | ۰/۹۸۱ | ۶۴۵۸۰ | ABLIM1 | $C_1$ |
| $C_9$ | ۰/۹۰۹ | ۶۴۱۶۶ | ABLIM1 | $C_1$ | $C_8$ | ۰/۹۷۸ | ۶۴۳۲۵ | ABLIM1 | $C_1$ |
| $C_{12}$ | ۰/۹۰۹ | ۵۷۸۵۲ | TACC2 | $C_5$ | $C_9$ | ۰/۹۶۳ | ۶۴۱۶۶ | ABLIM1 | $C_1$ |
| $C_{17}$ | ۰/۸۹۶ | ۵۸۱۴۳ | KIAA1217 | $C_2$ | $C_6$ | ۰/۹۴۳ | ۶۱۸۲۳ | ABLIM1 | $C_2$ |





[ DOI: 10.52547/jsdp.18.2.97 ]

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $C_3$ | ABLIM1 | ۶۱۶۱۲ | ۰/۹۳۰ | $C_6$ | $C_2$ | ABLIM1 | ۶۱۸۲۳ | ۰/۸۹۰ | $C_6$ |
| $C_4$ | ABLIM1 | ۶۱۲۹۱ | ۰/۹۰۸ | $C_6$ | $C_6$ | ABLIM1 | ۶۰۸۸۳ | ۰/۸۶۵ | $C_{11}$ |
| $C_5$ | TACC2 | ۶۱۰۹۶ | ۰/۸۹۶ | $C_2$ | $C_9$ | ABLIM1 | ۴۶۵۹۲ | ۰/۸۶۱ | $C_{11}$ |
| $C_6$ | ABLIM1 | ۶۰۹۸۲ | ۰/۸۹۱ | $C_{10}$ | $C_{18}$ | ABLIM1 | ۵۷۲۵۰ | ۰/۸۵۸ | $C_{10}$ |
| $C_6$ | ABLIM1 | ۶۰۸۸۳ | ۰/۸۸۷ | $C_{11}$ | $C_{15}$ | KIAA1217 | ۵۷۵۰۴ | ۰/۸۵۷ | $C_{11}$ |
| $C_2$ | TACC2 | ۶۰۸۲۳ | ۰/۸۸۰ | $C_{12}$ | $C_{15}$ | KIAA1217 | ۵۲۲۰۹ | ۰/۸۵۳ | $C_{16}$ |
| حد آستانه ۲۰۰۰ | | | | | حد آستانه ۱۵۰۰ | | | | |
| $C_1$ | ABLIM1 | ۶۴۵۸۰ | ۰/۹۷۰ | $C_7$ | $C_1$ | ABLIM1 | ۶۴۵۸۰ | ۰/۹۷۹ | $C_7$ |
| $C_1$ | ABLIM1 | ۶۴۱۶۶ | ۰/۹۶۱ | $C_9$ | $C_1$ | ABLIM1 | ۶۴۳۲۵ | ۰/۹۷۴ | $C_8$ |
| $C_3$ | ABLIM1 | ۶۱۶۱۲ | ۰/۹۴۳ | $C_6$ | $C_1$ | ABLIM1 | ۶۴۱۶۶ | ۰/۹۶۸ | $C_9$ |
| $C_5$ | TACC2 | ۶۱۰۹۶ | ۰/۹۲۱ | $C_2$ | $C_2$ | ABLIM1 | ۶۱۸۲۳ | ۰/۹۴۰ | $C_6$ |
| $C_2$ | ABLIM1 | ۶۰۸۲۳ | ۰/۹۰۹ | $C_{12}$ | $C_3$ | ABLIM1 | ۶۱۶۱۲ | ۰/۹۲۲ | $C_6$ |
| $C_9$ | ABLIM1 | ۴۶۵۹۲ | ۰/۸۸۱ | $C_{11}$ | $C_4$ | ABLIM1 | ۶۱۲۹۱ | ۰/۸۹۵ | $C_6$ |
| $C_{12}$ | TACC2 | ۵۴۷۶۵ | ۰/۸۸۰ | $C_{14}$ | $C_5$ | TACC2 | ۶۱۰۹۶ | ۰/۸۸۶ | $C_2$ |
| $C_{15}$ | KIAA1217 | ۵۲۲۰۹ | ۰/۸۳۴ | $C_{16}$ | $C_{11}$ | MUC6 | ۵۷۱۰۸ | ۰/۸۵۸ | $C_{13}$ |
| $C_{15}$ | KIAA1217 | ۵۷۵۰۴ | ۰/۸۱۳ | $C_{11}$ | $C_{15}$ | KIAA1217 | ۵۷۵۰۴ | ۰/۸۵۷ | $C_{11}$ |
| $C_{15}$ | KIAA1217 | ۵۴۱۸۷ | ۰/۷۹۸ | $C_8$ | $C_{15}$ | KIAA1217 | ۵۲۲۰۹ | ۰/۸۵۳ | $C_{16}$ |

**جدول-۵): ارتباط‌های بین بافتی استخراج‌شده از مجموعه داده فانتوم با استفاده از الگوریتم پیشنهادی.**

**(Table-5): Interconnecting extracted from the phantom dataset using the proposed algorithm.**

| نام بافت | ژن با بیشترین راه‌انداز بیان‌شده | تعداد راه‌اندازهای بیان‌شده | میزان شباهت | نام بافت | نام بافت | ژن با بیشترین راه‌انداز بیان‌شده | تعداد راه‌اندازهای بیان‌شده | میزان شباهت | نام بافت |
|---|---|---|---|---|---|---|---|---|---|
| حد آستانه ۱۰۰۰ | | | | | حد آستانه ۵۰۰ | | | | |
| $T_7$ | ACSL | ۷۵۳۵۶ | ۰/۹۷۰ | $T_{11}$ | $T_7$ | ACSL | ۷۵۳۵۶ | ۰/۹۷۰ | $T_{11}$ |
| $T_5$ | CELF | ۷۴۵۷۳ | ۰/۹۶۱ | $T_{11}$ | $T_5$ | CELF | ۷۴۵۷۳ | ۰/۹۶۱ | $T_{11}$ |
| $T_4$ | TCF7L | ۷۴۰۳۰ | ۰/۹۱۵ | $T_7$ | $T_7$ | ACSL | ۷۴۳۵۸ | ۰/۹۳۹ | $T_{14}$ |
| $T_7$ | NEURL | ۷۳۵۷۲ | ۰/۸۸۱ | $T_{13}$ | $T_6$ | TCF7L | ۷۴۰۸۴ | ۰/۸۹۶ | $T_{11}$ |
| $T_4$ | ACSL | ۷۳۵۲۶ | ۰/۸۵۸ | $T_5$ | $T_4$ | TCF7L | ۷۴۰۳۰ | ۰/۸۵۰ | $T_7$ |
| $T_{10}$ | TCF7L | ۷۳۴۸۳ | ۰/۸۴۱ | $T_{11}$ | $T_7$ | NEURL | ۷۳۸۷۲ | ۰/۸۴۹ | $T_{13}$ |
| $T_1$ | ADD | ۷۳۴۹۵ | ۰/۸۰۸ | $T_{11}$ | $T_5$ | TCF7L | ۷۳۸۴۰ | ۰/۸۲۲ | $T_{14}$ |
| $T_2$ | SCD | ۷۳۴۵۸ | ۰/۸۰۰ | $T_{11}$ | $T_4$ | ACSL | ۷۳۵۲۶ | ۰/۸۰۶ | $T_5$ |
| $T_6$ | NEURL | ۷۳۴۰۲ | ۰/۷۸۰ | $T_{14}$ | $T_3$ | ABLIM1 | ۷۳۵۱۷ | ۰/۷۹۸ | $T_{11}$ |
| $T_7$ | TACC2 | ۷۳۲۷۴ | ۰/۷۷۱ | $T_9$ | $T_{10}$ | TCF7L | ۷۳۴۸۳ | ۰/۷۸۵ | $T_{11}$ |
| حد آستانه ۲۰۰۰ | | | | | حد آستانه ۱۵۰۰ | | | | |
| $T_7$ | ACSL | ۷۵۳۵۶ | ۰/۹۷۰ | $T_{11}$ | $T_7$ | ACSL | ۷۵۳۵۶ | ۰/۹۷۰ | $T_{11}$ |
| $T_5$ | CELF | ۷۴۵۷۳ | ۰/۹۶۱ | $T_{11}$ | $T_5$ | CELF | ۷۴۵۷۳ | ۰/۹۶۱ | $T_{11}$ |
| $T_7$ | NEURL | ۷۳۸۷۲ | ۰/۹۴۴ | $T_{13}$ | $T_4$ | TCF7L | ۷۴۰۳۰ | ۰/۹۳۲ | $T_7$ |
| $T_5$ | ACSL | ۷۳۸۴۰ | ۰/۹۱۸ | $T_{14}$ | $T_4$ | ACSL | ۷۳۵۲۶ | ۰/۹۰۰ | $T_5$ |
| $T_{10}$ | TCF7L | ۷۳۴۸۳ | ۰/۸۹۱ | $T_{11}$ | $T_{10}$ | TCF7L | ۷۳۴۸۳ | ۰/۸۹۵ | $T_{11}$ |
| $T_1$ | ADD | ۷۳۴۹۵ | ۰/۸۷۳ | $T_{11}$ | $T_2$ | SCD | ۷۳۴۵۸ | ۰/۸۳۲ | $T_{11}$ |
| $T_7$ | TACC2 | ۷۳۲۷۴ | ۰/۸۶۲ | $T_9$ | $T_7$ | TACC2 | ۷۳۲۷۴ | ۰/۸۳۲ | $T_9$ |
| $T_7$ | ABLIM1 | ۷۳۲۴۵ | ۰/۷۸۹ | $T_{12}$ | $T_7$ | SCD | ۷۳۲۴۵ | ۰/۷۶۳ | $T_{12}$ |
| $T_5$ | SCD | ۷۳۰۵۵ | ۰/۷۶۵ | $T_{13}$ | $T_8$ | NEURL | ۷۳۱۶۲ | ۰/۷۵۰ | $T_{11}$ |
| $T_4$ | ACSL | ۷۳۰۳۲ | ۰/۷۸۵ | $T_6$ | $T_5$ | ABLIM1 | ۷۳۰۵۵ | ۰/۷۴۱ | $T_{13}$ |

با توجه به نتایج به‌دست‌آمده میزان شباهت بین سلول‌ها با درنظرگرفتن آستانه هزار در بهینه‌ترین حالت نسبت به موارد دیگر است. نتایج با این آستانه نشان‌دهنده بیشترین ارتباط بین دو سلول $C_1$ و $C_7$ در بیماری‌های





[ DOI: 10.52547/jsdp.18.2.97 ]

مختلف است. تعداد راه‌اندازهای Expression شده در این دو سلول ۶۴۵۸۰ است. این تعداد رفتار مشابه این دو سلول را در مواجه با بیماری‌های مختلف نشان می‌دهد. همچنین سلول $C_1$ با سلول $C_8$ نیز دارای ارتباط بالایی هستند. این اطلاعات می‌تواند در استخراج الگوهای رفتاری از یک ویروس خاص کمک کند. به‌طورکلی بیشتر ارتباطات در ژن ABLIM1 بیان شده‌اند.

جدول (۵) ارتباطات بین بافتی را با توجه به ۱۰ مورد از قوی‌ترین ارتباط‌ها بیان‌شده نشان می‌دهد. نماد $T_i$ معرف بافت i-ام است که در پیوست ۲ آورده شده است. ژنی با بیشترین بیان ژن بیان‌شده نشان‌دهنده منطقه(مکان) ارتباط بین دو بافت است. به‌منظور بررسی حد آستانه مناسب جهت فیلتر کردن برخی از ارتباطات راه اندازها و استخراج ارتباط‌های بین بافتی مناسب ما با چهار حد آستانه مختلف ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ نتایج را بررسی می‌کنیم.

نتایج به‌دست‌آمده میزان شباهت بین بافت‌ها با درنظرگرفتن آستانه پانصد در بهینه‌ترین حالت نسبت به موارد دیگر را نشان می‌دهد که بیشترین ارتباط مربوط به دو بافت $T_7$ و $T_{11}$ در بیماری‌های مختلف است. تعداد راه‌اندازهای Expression شده در این دو بافت ۷۵۳۵۶ است. به‌طورکلی بیشتر ارتباطات در ژن ACSL بیان شده‌اند.

در مجموعه‌داده فانتوم ممکن است از یک سلول یا بافت چند نمونه و از چند شخص بیماری گرفته شده باشد. همچنین ممکن است از یک سلول یا بافت در بیماری‌های مختلفی نمونه گرفته شده باشد؛ بنابراین هر سلول/بافت می‌تواند با چند یال به سایر سلول‌ها/بافت‌ها مرتبط باشد. این ارتباط‌ها درنهایت تشکیل یک ابر گراف را می‌دهد. با توجه به نبود ردهٔ هدف در داده‌های بیان ژن (فانتوم۵)، نیاز به شاخص‌های اعتبارسنجی برای سنجش میزان صحت نتایج خوشه‌بندی داریم. در این پژوهش از شاخص درونی Silhouette [17] برای این منظور استفاده شده است. این معیار عمل ارزیابی خوشه‌ها را با استفاده از مقادیری که از خوشه‌ها و نمای آن‌ها محاسبه می‌شود، انجام می‌دهد.

ارتباط دو سلول یا دو بافت زمانی به‌وجود می‌آید که در تعدادی از سلول‌ها/بافت‌ها (نمونه‌ها)، مقدار راه‌اندازها به میزان قابل‌توجهی بیان‌شده باشند. تجزیه‌وتحلیل نشان می‌دهد که مقادیر بیان ژن با حد آستانه‌های مختلف مقادیر متفاوتی در ضریب Silhouette ایجاد می‌کند. برای بررسی این موضوع، نتایج خوشه‌بندی در جدول (۶) برای بررسی حد آستانه‌های مختلف گزارش‌شده و برای ارزیابی نتایج با پژوهش‌های قبلی و خوشه‌بندی دودویی مقایسه شده است. حد آستانه‌ها با توجه به سایر مقالات مشخص شده است و ضریبی از صد است.

**(جدول-۶): مقایسه نتایج خوشه‌بندی با حد آستانه‌های مختلف.**

**(Table-6): Compare clustering results with different thresholds.**

| | روش‌ها | حد آستانه | ۰۵/۰ | ۱/۰ | ۱۵/۰ | ۲/۰ | ۲۵/۰ | ۳/۰ | ۳۵/۰ | ۰/۴ |
|---|---|---|---|---|---|---|---|---|---|---|
| سلول‌ها | خوشه‌بندی باینری | تعداد خوشه‌ها | ۱۶ | ۱۹ | ۱۹ | ۱۲ | ۱۰ | ۱۲ | ۸ | **۱۱** |
| | | ضریب Silhouette | ۰/۶۴۳ | ۰/۷۵۸ | ۰/۶۰۱ | ۰/۷۵۴ | ۰/۶۹۹ | ۰/۴۳۵ | ۰/۳۸۱ | **-۰/۵۳۱** |
| | خوشه‌بندی تجمعی | تعداد خوشه‌ها | ۱۵ | ۱۸ | ۱۵ | ۱۴ | ۱۱ | ۱۱ | ۱۶ | **۱۴** |
| | | ضریب Silhouette | ۰/۷۱۱ | ۰/۹۰۱ | ۰/۵۹۸ | ۰/۶۷۶ | ۰/۶۳۲ | ۰/۶۴۳ | ۰/۵۴۳ | **۰/۳۴۲** |
| بافت‌ها | خوشه‌بندی باینری | تعداد خوشه‌ها | ۹ | ۸ | ۱۱ | ۱۰ | ۱۱ | ۹ | ۱۳ | **۱۳** |
| | | ضریب Silhouette | ۰/۴۹۸ | ۰/۵۱۶ | ۰/۴۷۸ | ۰/۶۰۰ | ۰/۶۵۴ | ۰/۶۷۵ | ۰/۶۶۱ | **۰/۶۷۳** |
| | خوشه‌بندی تجمعی | تعداد خوشه‌ها | ۱۱ | ۱۱ | ۹ | ۹ | ۱۳ | ۱۳ | ۱۴ | **۱۵** |
| | | ضریب Silhouette | ۰/۵۹۸ | ۰/۵۲۰ | ۰/۶۱۹ | ۰/۶۰۴ | ۰/۷۲۳ | ۰/۷۶۲ | ۰/۷۰۰ | **۰/۷۳۷** |

با توجه به نتایج آزمایش، حد آستانه ۱/۰ دارای ضریب Silhouette با بهترین مقدار ۰/۸۵۸ در ۱۹ خوشه برای روش خوشه‌بندی دودویی در سلول‌ها است. این در حالی است که روش خوشه‌بندی تجمعی پیشنهادشده در این پژوهش با این حد آستانه به ضریب Silhouette ۰/۹۰۱ در ۱۸ خوشه رسیده است و عملکرد بهتری دارد. همچنین برتری روش پیشنهادی در بهترین حالت با ضریب Silhouette ۰/۷۶۲ در سیزده خوشه نسبت به روش دودویی با ضریب Silhouette ۰/۶۷۵ در نُه خوشه با حد آستانه ۲/۵ در بافت‌ها به‌وضوح قابل‌مشاهده است.

## ۶- نتیجه‌گیری

ارتباطات بین سلول‌ها و بافت‌ها به شناسایی بیماری‌ها مختلف و عوامل آن‌ها کمک خواهد کرد. درواقع ارتباط بین





[ DOI: 10.52547/jsdp.18.2.97 ]

سلول‌ها/بافت‌ها روابط وراثتی بین بیماران را نشان می‌دهد. این روابط به شناسایی نقاط مشترک بدن که تحت تأثیر بیماری‌های مختلفی قرار می‌گیرد، کمک می‌کند. در این مقاله تشخیص ارتباط‌های بین سلولی و بین بافتی در بیماری‌های مختلف با توجه به یک معیار شباهت توپولوژیکی گراف و یک روش خوشه‌بندی تجمعی بهبودیافته ارائه شده است. ارزیابی عملکرد الگوریتم خوشه‌بندی پیشنهادی دقت بالای آن را به اثبات رسانده است. خوشه‌بندی سلول‌ها/بافت‌ها باعث افزایش دقت و تمرکز معیار شباهت توپولوژیکی گراف در محدوده‌های مشابهت سلول‌ها/بافت‌ها شده است. الگوریتم خوشه‌بندی ارائه‌شده از یک تابع توافقی شباهت به‌منظور محاسبه است مشابهت بین اشیا استفاده می‌کند. ایجاد خوشه‌های اولیه با دقت بالایی با استفاده از مشابهت بین اشیا تولید و سپس مشابهت بین خوشه‌های اولیه محاسبه‌شده و با خوشه‌هایی با بالاترین ارتباط با یکدیگر ادغام می‌شوند. به‌منظور اصلاح و افزایش کیفیت خوشه‌های نهایی، یک الگوریتم بهبود عدم قطعیت اشیا را به‌وسیله تخصیص آن‌ها به سایر خوشه‌ها بررسی می‌کند. ما روش خود را روی پنج مجموعه‌داده دنیای واقعی آزمایش کردیم. نتایج نشان می‌دهد که به‌طور میانگین الگوریتم خوشه‌بندی پیشنهادی ما بهتر از سایر الگوریتم‌های مشابه کار می‌کند. از مزیت‌های روش پیشنهادی تعیین تعداد بهینه خوشه‌ها به‌صورت خودکار است. به‌جای محاسبه مشابهت بین اشیا به‌وسیلهٔ یک روش خوشه‌بندی انفرادی، از چندین روش بهره می‌گیرد. همچنین استفاده از الگوریتم تپه‌نوردی در تعیین مقادیر حد آستانه باعث کاهش پارامترهای ورودی مسأله و افزایش دقت در تعیین تعداد خوشه‌های بهینه‌شده است. ارزیابی کیفیت نتایج به‌وسیله یک استراتژی تطبیقی برای مقادیر آستانه می‌تواند در پژوهش‌های آینده مورد بررسی قرار بگیرد و میزان تأثیر و حساسیت روی کیفیت نتیجه خوشه‌های پایانی با روش حاضر مقایسه شود.

## ۷- مراجع 7- References


[1] I. H. Park, N. Arora, H. Huo, N. Maherali, T. Ahfeldt, A. Shimamura,G. Q. Daley, "Disease-specific induced pluripotent stem cells", *cell*, no. 134(5), 877-886, 2008.

[2] D. R. Bentley, S. Balasubramanian, H. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, J. M. Boutell, "Accurate whole human genome sequencing using reversible terminator chemistry", *nature*, vol. 456(7218), pp.53, 2008.

[3] W. Cook, and J. Palsberg, "A denotational semantics of inheritance and its correctness", Vol. 24. No. 10. ACM, 1989.

[4] H. C. Lukaski, "Methods for the assessment of human body composition: traditional and new", *The American journal of clinical nutrition*, vol. 46 (4), pp. 537-556, 1987.

[5] R. A. Lewis, B. Otterud, D. Stauffer, J. M. Lalouel, & M. Leppert, "Mapping recessive ophthalmic diseases: linkage of the locus for Usher syndrome type II to a DNA marker on chromosome 1q", Genomics, vol. 7(2), pp.250-256, 1990

[6] A. Bretto, H. Cherifi, D. Aboutajdine, "Hypergraph imaging: an overview", *Pattern Recognition*, vol. 35(3), pp. 651-658, 2002.

[7] A. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. De Hoon, V. Haberle, R. Andersson, "A level mammalian expression atlas", *Nature*, vol. 507(7493), pp. 462, 2012.

[8] M. E. Hegi, A. C. Diserens, S. Godard, , P. Y. Dietrich, , L. Regli, , S. Ostermann, R. Stupp, "Clinical trial substantiates the predictive value of O-6-methylguanine-DNA methyltransferase promoter methylation in glioblastoma patients treated with temozolomide", *Clinical cancer research*, vol.10(6), pp. 1871-1874, 2008.

[۹] کریمی عباس، حسینی لیلا سادات، "الگوریتم بهینه تقسیم‌بندی تصاویر میکروسکوپی خون برای تشخیص سلول‌های لوسمی حاد لنفوبلاست با استفاده از الگوریتم FCM و بهینه‌سازی ژنتیک"، پردازش علائم و داده‌ها، ۱۵ (۲)، ۴۵-۵۴، ۱۳۹۷.

[9] Karimi A, Hoseini L S. An Optimal Algorithm for Dividing Microscopic Images of Blood for the Diagnosis of Acute Pulmonary Lymphoblastic Cell Using the FCM Algorithm and Genetic Optimization. *JSDP. 2018; 15 (2) :45-54*

[۱۰] وحیدی فردوسی صدیقه، امیرخانی حسین، "ترکیب وزن‌دار خوشه‌بندی‌ها با هدف افزایش صحّت خوشه‌بندی نهایی"، پردازش علائم و داده‌ها، ۱۷ (۲)، ۸۵-۱۰۰، ۱۳۹۹.

[10] Vahidi Ferdosi S, Amirkhani H. Weighted Ensemble Clustering for Increasing the Accuracy of the Final Clustering. *JSDP. 2020; 17 (2) :100-85.*

[11] M. A. Ahmad, Z. Borbora, J. Srivastava, & N. Contractor, "Link prediction across multiple social networks. In Data Mining Workshops (ICDMW)", *2010 IEEE International Conference on*, pp. 911-918, 2010.

[12] H. M. Byrne, "Dissecting cancer through mathematics: from the cell to the animal




[ DOI: 10.52547/jsdp.18.2.97 ]

[24] W. Cui, Y. Xiao, H. Wang, W. Wang, "Local search of communities in large graphs", *In Proceedings of the 2014 ACM SIGMOD international conference on Management of data* , 2014, pp. 991-1002.

[25] J. A. Ramilowski, T. Goldberg, J. Harshbarger, E. Kloppman, M. Lizio, V. P. Satagopam, A. Forrest, "A draft network of ligand–receptor-mediated multicellular signalling in human", *Nature communications,* 2015.

[26] L. Cantini, E. Medico, S. Fortunato, M. Caselle, "Detection of gene communities in multi-networks reveals cancer drivers", *Scientific reports*, vol. 5, srep17386, 2015.

[27] G. A. Pavlopoulos, M. Secrier, C. Moschopoulos, N. Soldatos, S. Kossida, J. Aerts, P. G. Bagos, "Using graph theory to analyze biological networks", *BioData mining*, vol. 4(1), no.10, 2010.

[28] A. L. Fred, A. K. Jain, "Combining multiple clusterings using evidence accumulation", *IEEE transactions on pattern analysis and machine intelligence*, vol. 27(6), pp. 835-850, 2005.

[29] A. Strehl, J. Ghosh, "Relationship-based clustering and visualization for high-dimensional data mining", *INFORMS Journal on Computing*, vol. 15(2), pp. 208-230, 2003.

[30] G. Karypis, V. Kumar, A hypergraph partitioning package, 1998.

[31] N. Iam-On, T. Boongoen, S. Garrett, C. Price, "A link-based approach to the cluster ensemble problem", *IEEE transactions on pattern analysis and machine intelligence*, vol.33(12), pp. 2396-2409, 2011.

[32] S. Mimaroglu, E. Aksehirli, "Diclens: Divisive clustering ensemble with automatic cluster number", *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 9(2), pp. 408-420, 2012.

[33] M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, G. Guernec, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis", *Briefings in bioinformatics*, vol. 14(6), pp. 671-683. 2013.

[34] R. Reddy, "A comparison of methods: normalizing high-throughput RNA sequencing data", bioRxiv, 026062, 2015.

[35] S. Hawinkel, "Evaluation of normalization and analysis methods for microbiome data", 2015.

[36] S. Noguchi, T. Arakawa, S. Fukuda, M. Furuno, A. Hasegawa, F. Hori, T. Kawashima, "FANTOM5 CAGE profiles of human and mouse samples", *Scientific data*, vol. 4, pp. 112-170, 2017.

model", *Nature reviews. Cancer*, vol.10 (3), pp. 221, 2010.

[13] A. Csikász-Nagy, "Computational systems biology of the cell cycle", *Briefings in bioinformatics*, vol. 10(4), pp. 424-434, 2004.

[14] J. Hofbauer, K. Sigmund, "Evolutionary game dynamics", *Bulletin of the American Mathematical Society*, vol. 40(4), pp.479-519, 2003.

[15] R. P. Araujo, D. S. McElwain, "A history of the study of solid tumour growth: the contribution of mathematical modelling", *Bulletin of mathematical biology*, vol. 66(5), pp. 1039, 2004.

[16] A. Csikász-Nagy, M. Cavaliere, S. Sedwards, "Combining game theory and graph theory to model interactions between cells in the tumor microenvironment", *In New Challenges for Cancer Systems Biomedicine* Springer Milan, pp. 3-18, 2012.

[17] S.Aranganayagi, K. Thangavel, "Clustering categorical data using Silhouette coefficient as a relocating measure", *In Conference on Computational Intelligence and Multimedia Applications, 2007*, International Conference on vol. 2, pp. 13-17, 2007.

[18] T.Nakano, T. Suda, M. Moore, R. Egashira, A. Enomoto, K. Arima, "Molecular communication for nanomachines using intercellular calcium signaling. In Nanotechnology," 5th IEEE Conference on, pp. 478-481, 2005.

[19] T. Nakano, M. Moore, A. Enomoto, T. Suda, T.Koujin, T. Haraguchi, Y. Hiraoka, A cell-based molecular communication network, *In Proceedings of the 1st international conference on Bio inspired models of network, information and computing systems* , pp. 23, 2006.

[20] H. Jiang, J. Liu, Z. Zhao, Graph mining based knowledge discovery in designing decision-making context models, *In Systems and Informatics (ICSAI), 2014 2nd International Conference on* , pp. 948-953, 2014.

[21] H. Shao, Y. He, K. C. Li, X. Zhou, "A system mathematical model of a cell–cell communication network in amyotrophic lateral sclerosis", *Molecular BioSystems*, vol. 9(3), pp.398-406. 2013.

[22] R.Gao, H. Xu, P. Hu, W. C. Lau, "Accelerating graph mining algorithms via uniform random edge sampling", *In Communications (ICC), 2016 IEEE International Conference on*, pp. 1-6, 2016.

[23] T. Alqurashi, W. Wang, "Object-neighbourhood clustering ensemble method", *In International Conference on Intelligent Data Engineering and Automated Learning*, pp. 142-149, 2014.



[ DOI: 10.52547/jsdp.18.2.97 ]

[37] L. C. Gandolfo, T. P. Speed, RLE Plots: Visualising Unwanted Variation in High Dimensional Data. arXiv preprint arXiv:1704.03590, 2017.

[38] M. Arab, M. Hasheminezhad, Limitations of quality metrics for community detection and evaluation, In Web Research (ICWR), 2017 3th International Conference, pp. 7-14, 2017.

[39] Y. Qiao, H. Wang, D. Wang, "Parallelizing and optimizing overlapping community detection with speaker-listener Label Propagation Algorithm on multi-core architecture", *In Cloud Computing and Big Data Analysis (ICCCBDA)*, 2017 IEEE 2nd International Conference, 2017 , pp. 439-443.

[40] S. Mohammadi, J. Davila-Velderrain, M. Kellis, A. Grama, DECODE-ing sparsity patterns in single-cell RNA-seq. bioRxiv, 241646, 2018

[41] D. J. D. S.Price, Networks of scientific papers. Science, pp. 510-515, 1965.

[42] A. Papadimitriou, P. Symeonidis, Y. Manolopoulos, "Fast and accurate link prediction in social networking systems", *Journal of Systems and Software*, vol. 85(9), pp. 2119-2132, 2012.

[43] L. Katz, "A new status index derived from sociometric analysis", *Psychometrika*, vol. 18 (1), pp. 39-43, 1953.

## پیوست ۱

Full names of cells ($C_i$) in Table(4-19).
$C_1$: (hes.gfp.embryonic.stem.cells), $C_2$: (basophils), $C_3$: (cd14..monocytes...treated.with.lipoplysaccharide), $C_4$: (h9.embryonic.stem.cells), $C_5$: (lens.epithelial.cells), $C_6$: (ciliary.epithelial.cells), $C_7$: (cd14.cd16..monocytes.2), $C_8$: (fibroblast...aortic.advantitial.donor2..cytoplasmic.fraction), $C_9$: (cd14.cd16..monocytes.1), $C_{10}$: (b.lymphoblastoid.cell.line..gm12878.encode), $C_{11}$: (cd14..monocytes), $C_{12}$: (astrocyte...cerebral.cortex), $C_{13}$: (mast.cell), $C_{14}$: (melanocyte), $C_{15}$: (mesothelial.cells), $C_{16}$: (fibroblast...villous.mesenchymal), $C_{17}$: (melanocyte...dark), $C_{18}$: (melanocyte...dark).

## پیوست۲

Full names of tissues ($T_i$) in Table(4-20).
$T_1$: (amygdala), $T_2$: (caudate.nucleus), $T_3$: (hippocampus), $T_4$: (kidney), $T_5$: (locus.coeruleus), $T_6$: (medial.frontal.gyrus), $T_7$: (medulla.oblongata), $T_8$: (occipital.cortex), $T_9$: (small.intestine), $T_{10}$: (temporal.lobe), $T_{11}$: (testis), $T_{12}$: (throat), $T_{13}$: (thymus), $T_{14}$: (trachea).

**موسی مجرد،** در سال ۱۳۹۷ دکترای خود را در رشته رایانه گرایش سیستم‌های نرم‌افزاری از دانشگاه آزاد اسلامی واحد یاسوج دریافت کرد. ایشان هم‌اکنون عضو هیأت علمی دانشگاه آزاد اسلامی واحد فیروزآباد و زمینه‌های پژوهشی مورد علاقه ایشان داده‌کاوی، الگوریتم‌های تکاملی، گراف کاوی، شبکه‌های اجتماعی و بهینه‌سازی است.
نشانی رایانامه‌های ایشان عبارتند از:
**m.mojarad@iauf.ac.ir,**
**mosa.mojarad@gmail.com**

**حمید پروین،** در سال ۱۳۹۰ دکترای خود را در رشته رایانه گرایش هوش مصنوعی از دانشگاه علم و صنعت دریافت کرد و هم‌اکنون عضو هیأت علمی دانشگاه آزاد اسلامی واحد نورآباد ممسنی است. زمینه‌های پژوهشی مورد علاقه ایشان یادگیری ماشین، داده‌کاوی و الگوریتم‌های تکاملی است.
نشانی رایانامه ایشان عبارت است از:
**parvinhamid@gmail.com**

**صمد نجاتیان،** در سال ۱۳۹۳ دکترای خود را در رشته برق (مخابرات) از دانشگاه یو تی ام مالزی دریافت کرد و هم‌اکنون عضو هیأت علمی دانشگاه آزاد اسلامی واحد یاسوج است. زمینه‌های پژوهشی مورد علاقه ایشان سامانه‌های هوشمند، تشخیص الگو، ارتباطات بی‌سیم، نرم‌افزار تعریف رادیو، رادیو شناختی و آماری است.
نشانی رایانامه ایشان عبارت است از:
**nejatian@iauyasooj.ac.ir**

**کرم‌الله باقری‌فرد،** مدرک دکترای خود را در سال ۱۳۹۶ در گرایش نرم‌افزار در دانشگاه اراک دریافت کرد و هم اکنون عضو هیأت علمی دانشگاه آزاد اسلامی واحد یاسوج است. زمینه‌های پژوهشی مورد علاقه ایشان یادگیری ماشین، داده‌کاوی و سامانه‌های پیشنهاددهنده است.
نشانی رایانامه ایشان عبارت است از:
**k.bagheri@iauyasooj.ac.ir**



[ DOI: 10.52547/jsdp.18.2.97 ]