ارائه روشی جدید برای خوشه‌بندی داده‌های مخلوط بر مبنای تعداد ویژگی مشابه

حمید رضایی و نگین دانشپور*

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

چکیده

خوشه‌بندی، عملیاتی است که در آن مجموعه‌ای از نمونه‌داده‌ها، نسبت به میزان شباهت، دسته‌بندی می‌شوند. نمونه‌داده‌های خوشه‌بندی، عددی یا مخلوطی از عددی و غیرعددی (اسمی) هستند. یافتن میزان شباهت و اندازه‌گیری فاصله، از چالش‌های خوشه‌بندی داده‌های مخلوط است. در این مقاله سعی شده‌است در محاسبه میزان شباهت و تعیین فاصله، به پارامتر "تعداد ویژگی‌های مشابه" توجه شود. در نسبت‌دادن هر نمونه به خوشه در مواردی که فاصله‌ها برابر یا نزدیک باشد، تعداد ویژگی‌های مشترک نمونه‌ها تعیین کننده خوشه مناسب خواهد بود. برای محاسبه فاصله در الگوریتم مورد نظر از تفاضل عددی نرمال‌سازی‌شده برای ویژگی‌های عددی و از فاصله همینگ برای ویژگی‌های غیرعددی استفاده شده‌است. تعیین مرکز خوشه اولیه نیز مانند بسیاری از روش‌ها به‌صورت تصادفی انجام شده‌است و در تکرارهای بعدی الگوریتم، نمونه مناسب‌تر به‌عنوان مرکز خوشه انتخاب می‌شود. الگوریتم مورد نظر با پنج الگوریتم دیگر در پنج مجموعه‌داده مقایسه شده‌است. در بررسی نتایج، از سه معیار Accuracy، RI، F-Measure استفاده شده‌است. طبق نتایج آزمایش‌ها، در سه مجموعه‌داده، الگوریتم موردنظر حداقل دو درصد بهتر از دو الگوریتم و یک درصد بهتر از یکی دیگر از الگوریتم‌ها عمل کرده‌است. در یکی دیگر از مجموعه‌داده‌ها الگوریتم موردنظر نتایج برابر یا نزدیک به یک درصد دقت بهتر نسبت به الگوریتم برتر داشت. در مجموعه‌داده آخر نیز الگوریتم مورد نظر در رتبه دوم از بین پنج الگوریتم قرار داشت.

واژگان کلیدی: خوشه‌بندی، داده مخلوط، فاصله مقادیر، تشابه مقادیر، مرکز خوشه.

Presenting a new method for mixed data clustering based on the number of similar features

Hamid Rezaei And Negin Daneshpour*

Computer Engineering Department, Shahid Rajaee Teacher Training University, Tehran, Iran

Abstract

Clustering is an operation in which a set of data samples is categorized according to the degree of similarity. Examples of clustering data are numerical or a mixture of numerical and non-numerical (nominal) data. Finding similarities and measuring distances is one of the challenges of mixed data clustering. In the related works, to detect the degree of similarity and obtain the distance value, only the parameter of the distance value was considered and the cluster was selected based on its value. Clustering in this way, especially for mixed data, has not had very accurate results.

In this paper, we have tried to pay attention to the parameter "number of similar features" in calculating the degree of similarity and determining the distance. In assigning each sample to a cluster in cases where the distances are equal or close, the number of common features of the samples will determine the appropriate cluster. That is, we will pay attention to the "number of similar features" in addition to the distance to select the cluster. This idea believes that in cases where the distance of the cluster centers is close to the data object, it is better to choose the cluster center that has more features similar to the data object. Logically and also according to the proposed algorithm, the amount of

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۸/۹/۲۸ • تاریخ پذیرش: ۱۴۰۰/۹/۱۵ • تاریخ انتشار: ۱۴۰۳/۵/۱۰ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.1.39]

similarity should be in a larger number of features, not just a few limited features but with high similarity.
The parameter of the "number of similar features" has a specific definition and is obtained with a suitable threshold. If the distance value of two features is less than the threshold, those two features are considered as similar features.
To calculate the distance in the algorithm, the normalized numerical difference for numerical properties and the Hamming distance for non-numerical properties are used. Determining the initial cluster centers, like many methods, is done randomly, and in subsequent iterations of the algorithm, more appropriate samples are selected as the cluster centers. The algorithm is compared with 5 other algorithms in 5 datasets.
In examining the results, three criteria of Accuracy, RI and F-Measure have been used. According to the test results, in the mixed and integer datasets, the algorithm performs at least two percent better than the two algorithms and one percent better than the other algorithm. In another data set, the proposed algorithm had results equal to or close to one percent better accuracy than the superior algorithm. In the last data set, the proposed algorithm was ranked second among 5 algorithms. In general, the proposed algorithm won the top rank in most of the results, and in the rest of the cases, it won the second rank out of the five tested algorithms.

Keywords: **Clustering, Mixed data, Distance of values, Similarity of values, Cluster Center**

۱- مقدمه

خوشه‌بندی یک روش یادگیری ماشین بدون نظارت است که برای گروه‌بندی داده‌های بدون برچسبی است که "شبیه" به یکدیگر و "بی‌شباهت" با سایر خوشه‌ها هستند [5] [2]. بسیاری از الگوریتم‌های خوشه‌بندی فقط می‌توانند داده‌هایی را که دارای ویژگی‌های عددی هستند، مدیریت کنند. ویژگی‌های عددی مانند قد، وزن و سن هستند و ویژگی‌های اسمی (غیرعددی) مانند رنگ، نژاد، شغل و گروه‌خونی [5].

بسیاری از مجموعه‌داده‌های دنیای واقعی دارای ویژگی‌های عددی و اسمی هستند. آن‌ها مجموعه‌داده‌های مخلوط[1] نامیده می‌شوند. مجموعه‌داده‌های مخلوط اغلب در بسیاری از حوزه‌ها مانند سلامت، پزشکی، امور مالی و بازاریابی وجود دارند؛ بنابراین، توسعه الگوریتم‌های خوشه‌بندی که بتوانند چنین داده‌هایی را اداره کنند اهمیت پیدا کرده است [6].

موارد مهم و تأثیرگذار در نتایج خوشه‌بندی عبارتند از مراکز خوشه، تعداد خوشه مشخص‌شده، نوع محاسبه فاصله و شباهت، و روش و الگوریتم خوشه‌بندی. بیشتر الگوریتم‌های خوشه‌بندی با این چالش‌ها مواجه هستند و دقت خوشه‌بندی تحت تأثیر این موارد تغییر می‌کند [7].

برای محاسبه شباهت بین مقادیر ویژگی‌های عددی، عملیات ریاضی (مانند فاصله، زاویه، جمع یا میانگین) روی آن‌ها اعمال می‌شود. معیارهای شباهت مبتنی بر فاصله، بیشتر برای نقاط داده عددی استفاده می‌شوند [9]. مقادیر ویژگی‌های اسمی ذاتاً مرتب نمی‌شوند و امکان محاسبه مستقیم فاصله بین دو مقدار ویژگی اسمی وجود ندارد. یعنی مفهوم شباهت برای داده‌های اسمی به‌طور واضح تعریف نشده‌است؛ بنابراین، محاسبه اندازه‌گیری‌های مشابه مبتنی بر فاصله برای داده‌های غیرعددی یک کار چالش برانگیز است؛ زیرا اعمال مستقیم عملیات ریاضی، مانند جمع‌بندی یا میانگین‌گیری، در مقادیر ویژگی این مجموعه‌داده‌ها دشوار است [8].

کارهایی که تاکنون انجام شده‌است هر چند پیشرفت‌هایی داشته‌اند، اما تا رسیدن به دقت بالا در خوشه‌بندی داده‌های نوع مخلوط فاصله دارند. در این مقاله سعی شده‌است تا این دقت با ارائه راه‌حلی نو افزایش یابد.

به‌طورتقریبی در تمام روش‌های خوشه‌بندی داده‌های عددی و مخلوط تنها به پارامتر فاصله کلی مقادیر ویژگی‌ها توجه شده‌است. بدین صورت که در مقایسه میزان شباهت دو نمونه، فاصله کلی در مجموع ویژگی‌ها محاسبه و به‌دست می‌آید و با مقایسه آن، نمونه به خوشه مناسب تخصیص می‌یابد[9] [10]. درواقع در روش‌های قبلی برای این موضوع نحوه محاسبه فاصله و تعیین یک روش محاسبه فاصله در ویژگی‌های غیرعددی مجموعه‌داده‌ها مورد بررسی و مطالعه قرار گرفته‌است. در روش و الگوریتم این مقاله، پارامتر دیگری با عنوان *"تعداد ویژگی‌های مشابه"* در کنار پارامتر فاصله معرفی شده و بررسی آن در زمان اختصاص نمونه داده به خوشه، مورد توجه قرار گرفته است. در این مقاله ویژگی‌های مشابه به‌الزام ویژگی‌هایی با مقادیر به‌طورکامل مشابه نیستند، بلکه درصورتی‌که دو مقدار در یک ویژگی، فاصله‌ای کمتر از حدآستانهٔ مقداردهی‌شده توسط کاربر (برای مثال ۱۰ یا

[1] - Mixed Data

[DOI: 10.61186/jsdp.21.1.39]

۱۵ درصد و یا غیره) داشته باشند، به‌عنوان ویژگی مشابه در نظر گرفته می‌شوند.

الگوریتم بدین صورت عمل می‌کند که هنگام خوشه‌بندی علاوه‌بر فاصله محاسبه‌شده کلی، به مقدارِ تعداد ویژگی‌های مشابه توجه می‌کند. در مقایسه فاصله نمونه داده موردنظر با مراکز خوشه، اگر فاصله‌ها برابر یا نزدیک به هم باشد، خوشه‌ای انتخاب می‌شود که تعداد ویژگی مشترک بیشتری داشته باشد. بدین صورت این الگوریتم در انتخاب خوشه برای نمونه داده با دقت بیشتر، خوشه مناسب‌تر را انتخاب می‌کند. به‌طور کلی می‌توان گفت در راه‌کار ارائه‌شده در این مقاله سعی شده‌است:

۱) روش خوشه‌بندی جدیدی معرفی شود تا علاوه‌بر معیار فاصله، به معیار دیگری با عنوان *"تعداد ویژگی‌های مشابه"* توجه شود و با بررسی آن دقت خوشه‌بندی را افزایش دهد.

۲) علاوه بر این از ماتریس فاصله برای محاسبه و یافتن میزان فاصله داده‌ها در تکرارهای الگوریتم خوشه‌بندی استفاده شده‌است. این کار برای یافتن مرکز خوشه مناسب، سرعت محاسبه فواصل نقاط را افزایش داده است.

الگوریتم موردنظر با الگوریتم پایه k-prototypes و چهار الگوریتم SKC [4]، Multiview [5] ، DPC-MD [12] و Equi-based k-prototypes [6] مقایسه شده‌است. در انجام آزمایش‌ها از پنج مجموعه‌داده از UCI-Repository [24] استفاده شده‌است. معیار سنجش در آزمایش‌ها عوامل Accuracy، RI، F-Measure بودند که طبق نتایج، الگوریتم ارائه شده بهتر یا نزدیک به روش‌های ارائه شده قبلی عمل کرده است.

در ادامه، بخش دوم این مقاله به تشریح کارهای انجام‌شده می‌پردازد. در بخش سوم، الگوریتم مورد نظر با جزئیات معرفی می‌شود. نتایج مقایسه‌ها و آزمایش‌های انجام گرفته‌شده الگوریتم پیشنهادی و سایر روش‌ها در بخش چهارم به‌تفصیل بیان شده‌است. در انتها و در بخش پنجم نیز به نتیجه‌گیری و پیشنهادهایی برای توسعه الگوریتم مقاله پرداخته شده‌است.

۲-کارهای پیشین

مهم‌ترین موضوعات در پژوهش‌های مرتبط با خوشه‌بندی داده‌های مخلوط عبارتند از: نحوه محاسبه فاصله و میزان تشابه بین دو مقدار، روش و الگوریتم خوشه‌بندی، نحوه انتخاب مراکز خوشه و درنظرگرفتن وزن برای میزان اهمیت هر ویژگی. روش‌های متفاوتی برای خوشه‌بندی داده‌های مخلـوط وجود دارد که بیشتر بر مبنای k-prototypes یا k-modes هستند. نحوه کار این الگوریتم‌های پایه بدین صورت است که ابتدا مراکز خوشه به‌صورت تصادفی تعیین می‌شوند. نمونه‌داده‌ها با مرکزی که فاصله نزدیک‌تری داشته باشند، تشکیل خوشه می‌دهند. در مرحله بعد مرکز هر خوشه به نمونه‌ای که کمترین فاصله با سایر نمونه‌ها را داشته باشد، تغییر می‌کند و خوشه‌ها مجددا شکل می‌گیرند. این کار تا زمان عدم تغییر مراکز و یا تشکیل خوشه‌های بهینه ادامه می‌یابد [1]. برای محاسبه فاصله در داده‌های مخلوط، در بخش صفات عددی، عمده روش‌های خوشه‌بندی از فاصله اقلیدسی استفاده می‌کنند [2] و در بخش غیرعددی نیز مهم‌ترین روش‌ها عبارتند از [17]:

۱- تطبیق ساده

۲- بر مبنای فرکانس یا میزان رخداد در کل مجموعه

۳- بر مبنای رابطه با مقدار مقایسه شونده در کل مجموعه

روش نخست باعث ازدست‌رفتن اطلاعات می‌شود [2]، چون داده‌ها را در دو حالت صفر و یک بررسی می‌کند. درصورتی‌که داده‌ها به‌طورکامل مانند هم باشند فاصله صفر (تشابه کامل) و در غیر این‌صورت فاصله یک (بیشینه فاصله) را در نظر می‌گیرد. یعنی این روش میزان تشابه نسبی مقادیر صفات غیرعددی را درنظر نمی‌گیرد [15] [14].

روش دوم هر مقدار صفت غیرعددی را به‌صورت مستقل از دیگر صفات بررسی می‌کند و میزان مشاهده مقدار صفت در کل مجموعه را به‌عنوان پارامتر ارزیابی و مقایسه با دیگر مقادیر در نظر می‌گیرد [12]. روش سوم به ارتباط مقادیر توجه می‌کند و هرچند این میزان ارتباط، غیرمستقیم به‌دست می‌آید و محاسبه می‌شود؛ اما می‌توان آن را نوعی مزیت نسبت به دو دوش قبل در نظر گرفت و باید نتیجه آن در آزمایش‌ها، بیشتر بررسی شود [18] [3].

با توجه به محورهای مطالعاتی گفته‌شده، پژوهش‌های زیادی انجام شده‌است که به تعدادی از آن‌ها اشاره می‌شود. احمد و دی برای انجام خوشه‌بندی داده‌های مخلوط، الگوریتم k-means را بهبود دادند. آن‌ها در کار خود برای محاسبه فاصله ویژگی‌های عددی از فاصله اقلیدسی استفاده کردند و برای ویژگی‌های غیرعددی فرمول توزیع مقادیر در کل مجموعه‌داده را پیشنهاد دادند. فاصله‌ها بر اساس میزان تکرار و توزیع مقادیر در کل مجموعه‌داده به‌دست می‌آمد. همچنین اصلاحی در انتخاب

[DOI: 10.61186/jsdp.21.1.39]

مراکز خوشه انجام دادند. نتایج کار آن‌ها پیشرفت خوبی را نشان داد [23].

سال‌ها بعد آن‌ها در پژوهشی دیگر ساخت پارتیشن‌های اولیه را نسبت به پارتیشن‌های تصادفی پیشنهاد دادند. آن‌ها بر اساس هر ویژگی، پارتیشن‌هایی را تشکیل دادند و سپس پارتیشن‌ها را با هم ادغام کردند. خروجی، تعدادی پارتیشن اولیه برای شروع عملیات خوشه بندی بود. با الگوریتم initKmix دقت خوشه‌ها در داده‌ای مخلوط نسبت به حالت تصادفی بهبود خوبی داشت. آن‌ها ساخت پارتیشن‌های اولیه برای خوشه‌بندی فازی را نیز برای کارهای آتی پیشنهاد داده‌اند [2].

کومار و کاناوالی پیش‌پردازش اولیه بر روی مجموعه‌داده را برای خوشه‌بندی مطرح کردند. داده‌های نوفه‌ای و ناقص کنار گذاشته می‌شدند و سپس خوشه‌بندی بر اساس روش k-means انجام می‌گرفت. آزمایش‌ها دقت بالا و تشکیل خوشه‌های بهتر توسط این روش را نشان می‌داد. همچنین در روش آن‌ها دو مقدار غیرعددی در صورتی با هم مشابه بودند که میزان تکرار و توزیع آن‌ها در کل مجموعه‌داده یکسان و یا نزدیک به هم باشد [4].

سنگام و اوم الگوریتم k-prototypes را برای خوشه‌بندی داده‌های مخلوط بهبود دادند. آن‌ها ابتدا نرمال‌سازی مقادیر را برای رسیدن به دقت بهتر در نظر گرفتند. آن‌ها برای ویژگی‌ها وزن درنظر گرفتند تا ارزش هر ویژگی بهتر مشخص شود. بیش‌ترین کار آن‌ها مربوط به همین قسمت بود؛ یعنی دخالت‌دادن وزن در محاسبه فاصله‌ی بین دو ویژگی. مقدار وزن در کار آن‌ها با یک فرمول مشخص که طبق میزان تکرار و توزیع مقادیر در کل مجموعه‌داده بود، به‌دست می‌آمد. همچنین آن‌ها نشان دادند خوشه‌بندی داده‌های مخلوط، طبق فاصله کلی که از همه ویژگی‌ها به‌دست می‌آید، نسبت به حالتی که فقط چند ویژگی غیرعددی در نظر گرفته شود، به‌طورمعمول نتایج بهتری به دنبال دارد [6].

دوی تای و همکاران پژوهش‌های خود را برای مراکز خوشه انجام دادند. آن‌ها با الگوریتم k-pbc نسبت به حالت تصادفی، مراکز خوشه بهتری برای شروع خوشه‌بندی انتخاب می‌کردند که برای انواع خوشه‌بندی فازی و غیرفازی هم کاربرد داشت [21].

جی و همکاران در کار خود به اهمیت چگالی داده‌ها در انتخاب مراکز خوشه اشاره کردند. آن‌ها به‌جای انتخاب مراکز خوشه به‌صورت تصادفی از یک الگوریتم برای پیداکردن مراکز خوشه‌ی دقیق‌تر استفاده کردند. در روش آن‌ها با درنظرگرفتن میزان چگالی، مراکز خوشه مناسب پیدا می‌شدند. آن‌ها در کار خود به این نتیجه رسیدند بهترین مراکز خوشه آن‌هایی هستند که طبق میزان چگالی در کل مجموعه‌داده و میزان فاصله داده‌ها نسبت به هم انتخاب شوند. آزمایش‌های مختلف اهمیت این دو پارامتر برای تعیین مراکز خوشه را به‌خوبی تأیید می‌کرد [16]. جین یین و همکاران در پژوهشی دیگر با استفاده از روش و فرمولی متفاوت به همین نتیجه رسیدند [19]. مشابه این روش، پژوهشی توسط مینگ جینگ و همکاران انجام شد و آن‌ها نیز به اهمیت چگالی‌های محلی برای تعیین مراکز خوشه اشاره کردند [12]. ولدن و همکاران در پژوهش‌های خود فاصله‌ها را طبق فرمول گاور به‌دست آوردند [17].

لیو و گیوان برعکس بیشتر پژوهش‌هایی که از فاصله اقلیدسی برای فاصله ویژگی‌های عددی استفاده می‌کردند، از فاصله کسینوسی نرمال‌سازی شده‌استفاده کرده‌اند. به عقیده آن‌ها این کار باعث افزایش دقت در بازه‌های بزرگ عددی می‌شود. همچنین برای محاسبه فاصله در ویژگی‌های غیرعددی ماتریس فاصله را تعریف کردند تا فاصله‌ی دو مقدار ویژگی را به کمک این ماتریس و در فضای اقلیدسی به‌دست آوردند [13]. یوآن و همکاران در محاسبه میزان شباهت مقادیر ویژگی‌های غیرعددی از تئوری خشن استفاده کردند. تئوری خشن برای بیان مسائلی که در آن‌ها عدم قطعیت و ابهام وجود دارد، مورد استفاده قرار می‌گیرد. روش آن‌ها برای خوشه‌بندی، k-modes بود [7].

سونگ و جیا الگوریتم k-prototypes را با وزن‌دارکردن مقادیر، بهبود دادند. آن‌ها برای محاسبه وزن از قضیه آنتروپی اطلاعات استفاده کردند. در این روش نیز مانند بیشتر روش‌ها، وزنی که با فرمول آنتروپی اطلاعات محاسبه می‌شد، طبق میزان تکرار و توزیع مقادیر ویژگی‌های غیرعددی در کل مجموعه‌داده بود [8]. جیا و همکاران در محاسبه فاصله مقادیر ویژگی‌های غیرعددی با الگوریتم k-prototypes از مفهوم وضعیت برابر[1] استفاده کردند [9]. جین چائو و همکاران در پژوهش‌های خود برای خوشه‌بندی از الگوریتم زنبورعسل استفاده کردند [10].

آندرو اسکابار با استفاده از الگوریتم Randomwalk خوشه‌بندی داده‌های مخلوط را انجام داد. او در روش خود همه مقادیر ویژگی‌ها را به حالت گسسته تبدیل کرد و سپس میزان شباهت داده‌ها و فاصله آن‌ها را به‌دست آورد. البته او به‌جای گسسته‌کردن معمولی از گسسته‌کردن

[1] - Equal Situation

[DOI: 10.61186/jsdp.21.1.39]

فازی استفاده کرد تا در کنار افزایش دقت، میزان از دست رفتن اطلاعات نیز به کمینه برسد [11].

در روش آقای جی و همکارانش خوشه‌بندی داده‌های مختلف بر اساس دید[1]های مختلف انجام می‌گرفت. هر دید از خوشه‌بندی بر اساس یک یا تعداد محدودی از ویژگی‌ها تشکیل می‌شد. هر دید برای کاربرد و موضوع خاصی می‌توانست استفاده شود و یا حتی امکان ادغام دیدهای مختلف برای رسیدن به یک خوشه‌بندی جامع و کلی وجود داشت. در کار آن‌ها برای تشکیل دیدها از روش خوشه‌بندی k-prototypes استفاده می‌شد. برای دقت بالای محاسبات، فاصله‌های ویژگی‌های عددی، نرمال‌سازی می‌شد. فاصله ویژگی‌های غیرعددی نیز از روش تطبیق ساده محاسبه می‌شد. همچنین برای درنظرگیری میزان اهمیت ویژگی‌ها، از یک وزن در محاسبات استفاده می‌شد [5].

ژانگ و یو خوشه‌بندی را کمی متفاوت‌تر از بقیه انجام دادند. در روش آن‌ها هر خوشه از دو بخش اصلی و بخش فرعی تشکیل می‌شد. میزان تعلق هر داده به یک خوشه در سه حالت تعلق کامل، تعلق نامطمئن و عدم تعلق مشخص می‌شد. کاملا روشن است که اگر داده به یک خوشه تعلق کامل داشت در بخش اصلی خوشه قرار می‌گرفت. در حالت تعلق نامطمئن، در بخش فرعی خوشه قرار می‌گرفت. طبق نتایج روش آن‌ها برای خوشه‌بندی فازی بسیار مناسب بود [20].

آقای سو و همکاران در روش خود برای خوشه‌بندی داده‌های مخلوط از نوعی درخت سلسله‌مراتبی استفاده کردند. در این روش یک درخت برای ویژگی‌های غیرعددی تشکیل می‌شد. ویژگی‌ها و مقادیر، سلسله‌مراتب درخت را تشکیل می‌دادند. در درخت تشکیل شده، ریشه ویژگی است و گره‌ها همه مقادیری که آن ویژگی در مجموعه‌داده دارد. مقادیر گره‌ها طبق میزان توزیع و اهمیت در سطوح مختلف درخت قرار می‌گرفتند. به هر گره یک مقدار عددی طبق میزان توزیع آن مقدار، داده می‌شد که از آن برای محاسبه فاصله گره‌ها یا همان مقادیر غیرعددی استفاده می‌کردند [22]. سحر بهزادی و همکاران هم برای خوشه‌بندی داده‌های مخلوط مانند روش قبل، درخت سلسله مراتبی ساختند؛ با این تفاوت که برای هر ویژگی درختی ساخته می‌شد و مقادیر ویژگی طبق میزان نزدیکی و فاصله در برگ‌های درخت و سلسله مراتب هم قرار می‌گرفتند و بر طبق میزان توزیع در مجموعه‌داده یک ارزش عددی دریافت می‌کردند و فاصله دو ویژگی از روی همان ارزش عددی محاسبه می‌شد. خوشه‌بندی بدون داده‌های پرت و نویزی از ویژگی‌های این روش بود [3].

بیشتر کارهای انجام گرفته شده در زمینه خوشه‌بندی داده‌های مخلوط بر روی معیار فاصله و نحوه‌ی محاسبه میزان شباهت داده‌ها متمرکز هستند و الگوریتم‌های پایه‌ای خوشه‌بندی را با فرمول جدیدی برای محاسبه پارامتر وزن یا فاصله بهبود می‌دهند. پژوهشگران این حوزه معتقدند برای رسیدن به روش دقیق خوشه‌بندی و دقت بالا برای تعیین فاصله مقادیر داده‌های مخلوط مطالعات بیشتری باید انجام گیرد. مطالعه برای رسیدن به روش‌های جدید در روش خوشه‌بندی داده‌های مخلوط شاید بتواند دقت مورد ارزیابی در این حوزه را افزایش دهد. در جدول (۱) تعدادی از روش‌های خوشه‌بندی مطالعه‌شده در سال‌های اخیر همراه با ویژگی‌های مهم نمایش داده شده‌است.

۳- الگوریتم پیشنهادی برای خوشه‌بندی داده‌های مخلوط

داده‌های مخلوط از دو بخش عددی و غیرعددی تشکیل می‌شوند. در الگوریتم‌های خوشه‌بندی داده‌های مخلوط، فاصله در ویژگی‌های عددی محاسبه و با فاصله محاسبه‌شده در بخش ویژگی‌های غیرعددی جمع می‌شود و یک مقدار عددی به عنوان فاصله کلی بین دو نمونه داده به‌دست می‌آید. در بیشتر این الگوریتم‌ها مراکز خوشه اولیه به‌صورت تصادفی انتخاب و در تکرارهای بعدی نمونه‌داده‌ی مناسب‌تر به عنوان مرکز خوشه انتخاب می‌شود. الگوریتم حالت تکرار شونده دارد و تا رسیدن به حالتی که تغییری در خوشه‌ها ایجاد نشود ادامه دارد [3] [2].

بیشتر روش‌های خوشه‌بندی، اعم از نوع داده عددی یا مخلوط و یا غیره برای اختصاص داده به خوشه موردنظر تنها به فاصله‌ای که از محاسبات مقادیر در ویژگی‌ها به‌دست می‌آید اکتفا می‌کنند. یعنی فاصله محاسبه شده در مجموع ویژگی‌های دو نمونه داده به عنوان ملاک ارزیابی برای تعیین خوشه لحاظ می‌شود [2] [3]. اشکالی که وارد می‌شود این است که در این روش‌ها، مقدار فاصله‌ی کلی، تنها ملاک ارزیابی شباهت دو نمونه داده در نظر گرفته می‌شود. این فاصله می‌تواند در یک یا دو ویژگی کم و حداقل باشد و در بقیه ویژگی‌ها به مقدار بزرگتری برسد [9]. درنهایت مجموع فاصله ویژگی‌ها برای انتخاب خوشه‌ی نزدیک‌تر بررسی می‌شود [11].

[1] - View

[DOI: 10.61186/jsdp.21.1.39]

(جدول- ۱): دسته‌بندی تعدادی از روش‌های خوشه‌بندی در سال‌های ۲۰۱۷ تا ۲۰۲۱

(Table-1): Classification of a number of clustering methods in 2017 to 2021

الگوریتم	سال	روش	ویژگی
Multi-view EM k-Prototype	۲۰۲۱	بر مبنای k-Prototype	خوشه‌بندی بر اساس ویژگی‌های انتخابی انجام می‌شود. فرایندی زمانبر است. [5]
initkmix	۲۰۲۰	بر مبنای k-Prototype	روش مورد استفاده ساده است و محاسبه فاصله دقیق نیست. [2]
SKC Technique	۲۰۲۰	بر مبنای k-means	با عملیات پیش پردازش، داده‌های نویزی حذف می‌شوند تا نتایج دقیق شوند. محاسبه فاصله بر اساس معیار واگرایی است. بیشترین شباهت را مقادیری دارند که تعداد تکرارشان برابر باشد. [4]
WKPCA	۲۰۲۰	بر مبنای k-Prototype	با درنظر گرفتن وزن ویژگی‌ها، نوآوری و دقت بالا در محاسبه فاصله دارد. با درنظرگیری تعداد خوشه‌های زیاد دقت کم می‌شود. [8]
ABC k-Prototype	۲۰۱۹	k-Prototype, artificial bee colony strategy	با تلفیق الگوریتم زنبور عسل مراکز خوشه راحت‌تر و دقیق‌تر شناسایی می‌شوند. [9]
Equi-biased k-Prototype	۲۰۱۸	بر مبنای k-Prototype	با یک وزن میزان مشارکت ویژگی‌ها در خوشه‌بندی قابل تنظیم است. مجموع فاصله در ویژگی‌های عددی و غیرعددی را لحاظ می‌کند. [5]
Clustering using Random Walk	۲۰۱۷	بر مبنای Random Walk	این روش به تعریف صریح اندازه‌گیری فاصله احتیاج ندارد. مناسب خوشه‌بندی فازی است. [10]

یعنی به این نکته که بهتر است فاصله در همه‌ی ویژگی‌ها یا تعداد بیشتری از ویژگی‌ها به‌صورت یکنواخت کم باشد توجه نمی‌شود. روش پیشنهادی این است که در انتخاب خوشه برای نمونه موردنظر، تنها به فاصله اکتفا نکرده و به تعداد ویژگی‌های مشابه نیز توجه کرد و این دو در کنار هم به عنوان معیار خوشه‌بندی انتخاب شوند. در ادامه ابتدا تعاریف و بخش‌های مقدماتی معرفی می‌شوند و سپس با جزئیات بیشتر الگوریتم خوشه‌بندی پیشنهادی تشریح می‌شود. جدول (۲) علائم و نشان‌های به‌کاررفته در فرمول‌ها و تعاریف را شرح می‌دهد.

تعریف ۱– مجموعه‌داده مخلوط مورد استفاده در الگوریتم پیشنهادی به‌صورت زیر تعریف می‌شود:

$$X = \{ x_1 , x_2 , x_3 , ..., x_n \}$$
$$x_i = \{ x_i^1 , x_i^2 , \dots , x_i^m , x_i^{m+1} , x_i^{m+2} , \dots , x_i^p \}$$
$$0 \le i \le n$$

تعریف ۲ – حد آستانه

در الگوریتم پیشنهادی برای سنجش میزان فاصله و میزان تشابه مقادیر، به یک حد آستانه احتیاج است. مقدار حدآستانه باید توسط کاربر و در ابتدای الگوریتم به‌عنوان یک پارامتر، مشخص و مقداردهی شود. در این مقاله این پارامتر را Tr می‌نامیم.

(جدول-۲): شرح علائم و اختصارات

(Table-2): Symbols and Abbreviations

علائم	شرح
X	مجموعه داده مخلوط (شامل مقادیر عددی و غیرعددی)
x	یک نمونه داده
n	تعداد نمونه داده در مجموعه
a	یک ویژگی عددی یا غیرعددی متعلق به نمونه داده
l	شماره‌ی ویژگی
x_2^l	ویژگی l ام از نمونه داده‌ی دوم
m	تعداد ویژگی‌های عددی در مجموعه داده
p-m	تعداد ویژگی‌های غیر عددی در مجموعه داده
Tr	حد آستانه
dist	میزان فاصله
C	مرکز خوشه

تعریف ۳ – ویژگی‌های مشابه

در روش پیشنهادی مفهومی با عنوان "*ویژگی مشابه*" ارائه شده‌است. دو مقدار در یک ویژگی را مشابه می‌دانیم درصورتی‌که فاصله آن‌ها کوچکتر یا مساوی حد آستانه باشد.

$\text{dist}(x_i^l, x_j^l) \le \text{Tr}$ is similar

$\text{dist}(x_i^l, x_j^l) > \text{Tr}$ is not similar

بدیهی است که برای مقایسه، دو مقدار باید در یک نوع و بازه باشند. چون حدآستانه به‌صورت درصد تعیین می‌شود، فاصله به‌دست‌آمده نیز با توجه به بازه مقادیر ویژگی، به درصد تبدیل و سپس با حدآستانه مقایسه می‌شود. نکته مهم این است که نحوه محاسبه فاصله مهم نیست؛ یعنی فرقی نمی‌کند فاصله مقادیر عددی با فاصله اقلیدسی و یا با معیار دیگر محاسبه شود ؛ همین‌طور نحوه محاسبه فاصله مقادیر غیرعددی مهم نیست. بلکه مقدار به‌دست‌آمده برای فاصله، مورد سنجش با حدآستانه قرار خواهد گرفت.

مثال زیر نحوه تشخیص ویژگی مشابه در ویژگی‌های عددی و غیرعددی را نشان می‌دهد. با فرض اینکه حدآستانه بیست درصد تعیین شده‌است، جدول (۳) دو نمونه داده با تعدادی ویژگی و مقادیر آن را نشان می دهد.

[DOI: 10.61186/jsdp.21.1.39]

(جدول-۳): ویژگی‌های مشابه در دو نمونه داده
(Table-3): Similar features in two data objects

عنوان	قد	وزن	جنسیت	تحصیلات	...
نمونه۱	۱۶۰	۶۰	زن	کارشناسی	...
نمونه۲	۱۷۵	۹۰	مرد	کارشناسی ارشد	...

بازه ویژگی‌ها به‌صورت زیر است:
قد: ۱۰۰ تا ۲۰۰ سانتیمتر، وزن: ۴۰ تا ۱۰۰ کیلوگرم، جنسیت: زن و مرد، تحصیلات: ۰ بیسواد، ۱ خواندن و نوشتن، ۲ سیکل، ۳ زیردیپلم، ۴ دیپلم، ۵ فوق دیپلم، ۶ کارشناسی، ۷ کارشناسی ارشد، ۸ دکترا، ۹ فوق دکترا.

بررسی تشابه ویژگی قد:
(۱۷۵-۱۶۰) / ۱۰۰ = ۰/۱۵ < ۰/۲۰
با توجه به مقادیر ویژگی قد، این ویژگی در این دو نمونه "ویژگی مشابه" در نظر گرفته می‌شود.

بررسی تشابه ویژگی وزن:
(۹۰-۶۰) / ۶۰ = ۰/۵۰ > ۰/۲۰
مقدار به‌دست آمده، بیشتر از حدآستانه است و"ویژگی مشابه" در نظر گرفته نمی‌شود.

بررسی تشابه ویژگی جنسیت:
با فرض فاصله همینگ برای فاصله در جنسیت و عدم تشابه کامل مقادیر، "ویژگی مشابه" در نظر گرفته نمی‌شود.

بررسی تشابه ویژگی تحصیلات:
(۷-۶) / ۱۰ = ۰/۱۰ < ۰/۲۰
"ویژگی مشابه" در نظر گرفته می‌شود.
"تعداد ویژگی مشابه" این دو نمونه در ۴ ویژگی، ۲ عدد است.

۳-۱- فاصله در ویژگی‌های عددی

محاسبه فاصله در ویژگی‌های عددی پیچیدگی خاصی ندارد. تفاضل ساده و فاصله اقلیدسی نمونه‌ای از روش‌ها برای به‌دست آوردن این فاصله هستند. ویژگی‌ها در مجموعه‌داده‌ها معمولا نوع متفاوت و بازه‌های متفاوتی برای مقادیر دارند. در صورتی که به این بازه‌ها دقت نشود عدد به‌دست آمده در مجموع فاصله می‌تواند تحت تاثیر یک ویژگی که بازه‌ی خیلی بزرگتر و یا خیلی کوچکتر از بقیه دارد قرار گیرد [13]. به منظور رفع این مشکل، برای اندازه‌گیری فاصله در ویژگی‌های عددی از نرمال‌سازی استفاده شده‌است. نرمال‌سازی، روش‌های مختلفی دارد که روش Gmax در این مقاله استفاده شده‌است [6] [13]. با انجام نرمال‌سازی، مقدار فاصله همه ویژگی‌ها در محدوده ۰ تا ۱ به‌دست می‌آید. یعنی وزن و اولویت ناخواسته‌ای که در حالت قبل (بدون نرمال‌سازی) در محاسبه فاصله کلی اعمال می‌شد، از بین می‌رود. رابطهٔ (۱) نحوه محاسبه فاصله در بخش عددی دو نمونه داده را نمایش می‌دهد. در رابطه (۲) نیز به‌صورت جزئی‌تر نحوه محاسبه فاصله دو مقدار از یک ویژگی در بخش عددی نمایش داده شده‌است که در آن از تفاضل ساده به همراه نرمال‌سازی، مقدار فاصله به‌دست می‌آید:

$$dist_{numerical}(x_i, x_j) = \sum_{l=1}^{m} dist(x_i^l, x_j^l) \qquad (۱)$$

$$dist(x_i^l, x_j^l) = \frac{|x_i^l - x_j^l|}{Gmax_l - Gmin_l} \qquad (۲)$$

در رابطه (۲)، $Gmax_l$ بیشینه مقدار و $Gmin_l$ کمینه مقدار در بازه مقادیر ویژگی l ام است.

۳-۲- فاصله در ویژگی‌های غیرعددی

ساده‌ترین روش که در الگوریتم k-prototypes [8] (الگوریتم پایه‌ای خوشه‌بندی داده‌های مخلوط) نیز برای محاسبه فاصله ویژگی با مقادیر غیرعددی استفاده شده‌است، استفاده از فاصله همینگ یا روش تطبیق ساده است. در فاصله همینگ درصورتی‌که دو مقدار در یک ویژگی برابر باشند، فاصله یک و در غیر این‌صورت فاصله صفر خواهد بود [3]. بدیهی است که در نظرگیری میزان فاصله در این حالت خیلی دقیق نخواهد بود و اشکالاتی مانند دقت پایین و ازدست‌رفتن اطلاعات دارد [4]. در بسیاری از مجموعه‌داده‌های واقعی و در ویژگی‌های مختلف، تشابه صددرصدی مقادیر برای رسیدن به فاصله صفر شاید به دفعات کمی اتفاق افتاده باشد. رابطه‌ی (۳) [3]فاصله کلی بخش غیرعددی دو نمونه داده را به‌صورت فرمول نمایش داده است. در رابطهٔ (۴) [3] نیز نحوه به‌دست‌آوردن فاصله با روش همینگ برای دو مقدار غیرعددی ذکر شده‌است:

$$dist_{non-numerical}(x_i, x_j) = \sum_{l=m+1}^{p} dist(x_i^l, x_j^l) \qquad (۳)$$

$$dist(x_i^l, x_j^l) = 0 \quad : \quad x_i^l = x_j^l \text{ اگر}$$
$$dist(x_i^l, x_j^l) = 1 \quad : \quad x_i^l \neq x_j^l \text{ اگر} \qquad (۴)$$

[DOI: 10.61186/jsdp.21.1.39]

۳-۳- فاصله در کل مجموعه‌داده

همان‌طورکه گفته‌شد، فاصله کلی از مجموع فاصله در بخش عددی و غیرعددی به‌دست می‌آید. رابطه (۵) نحوه محاسبه این فاصله را نمایش می دهد.

$$dis(x_i, x_j) = \sum_{l=1}^{m} dist_{numerical}(x_i^l, x_j^l) + \sum_{l=m+1}^{p} dist_{non-numerical}(x_i^l, x_j^l) \quad (۵)$$

۳-۴- مراکز خوشه و ماتریس فاصله

مراکز خوشه نقش بسیار مهمی در خوشه‌بندی دارند و خوشه‌های تشکیل شده تا حدی بستگی به مراکز خوشه انتخاب شده دارند [3] [7]. در اکثر روش‌های خوشه‌بندی، مراکز خوشه در ابتدای الگوریتم به‌صورت تصادفی انتخاب می‌شوند. خوشه‌بندی عملیاتی تکرار شونده است و تا رسیدن به موقعیت مطلوب الگوریتم تکرار می‌شود [7]. الگوریتم‌ها در هر تکرار، مراکز خوشه را برای رسیدن به نمونه‌داده‌ی مناسب‌تر، تغییر می‌دهند. برخی از پژوهشگران نمونه‌داده‌ی مناسب را با توجه به چگالی هر نمونه انتخاب می‌کنند [12]، برخی دیگر نیز با توجه به میزان همگرایی یا واگرایی و سایر پارامترها انتخاب می‌کنند [16]. در بیشتر روش‌ها نمونه مناسب، نمونه‌ای است که کمترین فاصله را با بقیه داشته باشد. بدین صورت که بعد از اولین مرحله و طبق اولین خوشه‌هایی که به‌دست می‌آید، در هر خوشه نمونه‌داده‌ای که کمترین فاصله با بقیه را دارد به عنوان مرکز خوشه جدید انتخاب می‌شود [3]. بدیهی است در هر تکرار تغییری در تعداد خوشه‌ها و مراکز خوشه ایجاد نمی‌شود. فقط مراکز خوشه مناسب‌تر جایگزین مراکز خوشه قبلی می‌شوند. در روش پیشنهادی، مراکز خوشه‌ی جدید طبق الگوی کمترین فاصله انتخاب می‌شوند. برای پیدا کردن ساده‌تر نمونه‌داده‌ی مناسب، طبق شکل ۱ ماتریس فاصله تشکیل می‌شود. در این ماتریس، فاصله‌ی هر داده با سایر داده‌های همان خوشه ذخیره می‌شود. به دلیل متقارن بودن ماتریس فاصله، وقتی فاصله x_1 با x_2 محاسبه می‌شود و در ماتریس قرار می‌گیرد، برای محاسبه فاصله x_2 با x_1 دیگر محاسبه‌ای انجام نمی‌شود و از همان مقدار استفاده می‌شود و این باعث افزایش سرعت این بخش از الگوریتم خواهد شد. جمع سطری ماتریس نشانگر مجموع فاصله یک نمونه داده با سایر داده‌های آن خوشه است. نمونه‌ای که کمترین مقدار را داشته باشد به معنای فاصله‌ی کمتر با بقیه است و به عنوان مرکز خوشه جدید در تکرار بعدی انتخاب می‌شود.

	x_1	x_2	x_3	..	..	x_n	
x_1	0	$dist_{12}$	$dist_{13}$	..	..	..	$dist_{12} + dist_{13}$+... = totlal dist x_1
x_2	$dist_{12}$	0	$dist_{23}$	..	..	..	$dist_{12} + dist_{23}$+... = totlal dist x_2
x_3	$dist_{13}$	$dist_{23}$	0	..	..	..	$dist_{13} + dist_{23}$+... = totlal dist x_3
..	..	..	..	0	..	..	
..	..	..	..	..	0	..	
x_n	..	..	..	..	..	0	

(شکل-۱): ماتریس فاصله.

(Figure-1): Distance Matrix

۳-۵- الگوریتم پیشنهادی

روش پیشنهادی این است که در انتخاب خوشه برای نمونه‌داده موردنظر، تنها به فاصله اکتفا نکرده و به تعداد ویژگی‌های مشابه نیز توجه کرد و این دو در کنار هم معیار خوشه‌بندی باشند. یعنی در حالتی که فاصله نمونه‌داده با چند مرکز خوشه برابر یا نزدیک باشد، خوشه‌ای انتخاب می‌شود که تعداد ویژگی مشابه بیشتری داشته باشد. در این روش اشیایی که تعداد ویژگی‌های مشابه زیادی داشته باشند به هم شبیه‌تر هستند و فاصله به تنهایی معیار نخواهد بود. بدیهی‌ست احتمال تشابه صد درصدی مقادیر ویژگی‌ها در بسیاری از موارد کم است (البته بستگی به نوع داده‌ها نیز دارد). این تشابه را می‌توان به کمک یک حدآستانه[1] (برای مثال ده درصدی) با توجه به بازهٔ هر ویژگی تعریف کرد. شرایطی را در نظر بگیرید که یک شیء داده با دو خوشهٔ یک و دو مقایسه[2]می‌شود و فاصله در خوشه نخست کمی کمتر از خوشه دوم است، اما تعداد ویژگی‌های مشابه در خوشه دوم نسبت به خوشه نخست بیشتر است؛ در این حالت خوشه مناسب برای این نمونه داده، خوشه دوم در نظر گرفته می‌شود.

با یک مثال، نحوه خوشه‌بندی الگوریتم پیشنهادی به‌وضوح بیان می‌شود. در این مثال حدآستانه پانزده درصد فرض شده‌است. جدول (۴)، میزان فاصله و تعداد ویژگی مشابه نمونه‌ای با دو مرکز خوشه را نشان می‌دهد.

(جدول-۴): فاصله نمونه داده با مراکز خوشه

(Table-4): The distance of the data object to the cluster centers

عنوان	میزان فاصله کلی	تعداد ویژگی مشابه
مرکز خوشه ۱	۲۰۰	۵
مرکز خوشه ۲	۲۱۰	۷

[1] - Threshold
[2] - Compare

[DOI: 10.61186/jsdp.21.1.39]

در حالت عادی، نمونه باید به خوشه ۱ ملحق شود؛ چون فاصله کمتری دارد؛ اما در الگوریتم پیشنهادی خوشه ۲ به‌عنوان خوشه مناسب‌تر شناسایی می‌شود.

الگوریتم ابتدا به سراغ خوشه با بیشترین تعداد ویژگی مشابه می‌رود. خوشه۲ تعداد ویژگی مشابه بیشتری دارد. حال میزان اختلاف فاصله خوشه۲ با نزدیک‌ترین خوشه سنجیده می‌شود:

(۲۱۰-۲۰۰) / (۲۱۰-۰)= ۰/۰۴۷ < ۰/۱۵

مقدار اختلاف فاصله (نرمالسازی شده با Gmax) خوشه۲ (خوشه با بیشترین تعداد ویژگی مشابه) با خوشه۱ (نزدیک‌ترین خوشه) کمتر از حدآستانه است، پس خوشه۲ خوشه منتخب شناخته می‌شود.

تأکید می‌شود که این حالت با توجه به فاصله و در کنار سنجش فاصله انجام می‌شود. یعنی این‌طور نیست در مواردی که فاصله نسبت به خوشه‌های مقایسه‌شونده اختلاف زیادی داشته باشد، به صرف تعداد ویژگی مشابه زیاد، انتخاب خوشه انجام گیرد. مثال زیر این حالت را به‌خوبی نشان می‌دهد.

(جدول-۵): فاصله نمونه‌داده با مراکز خوشه
(Table-5): The distance of the data object to the cluster centers

عنوان	میزان فاصله کلی	تعداد ویژگی مشابه
مرکز خوشه ۱	۲۰۰	۵
مرکز خوشه ۲	۳۰۰	۷

در جدول (۵)، اختلاف خوشه ۲ و خوشه ۱ بیش از حدآستانه است و شرط محقق نخواهد شد. در این حالت، با وجود اینکه خوشه ۲ تعداد ویژگی مشابه بیشتری دارد، اما خوشه ۱ (خوشه با نزدیک‌ترین فاصله) انتخاب می‌شود. یعنی همانند الگوریتم های مرسوم.

(۳۰۰-۲۰۰) / (۳۰۰-۰)= ۰/۳۳ > ۰/۱۵

در الگوریتم پیشنهادی، همان حد آستان‌های مرحله قبل استفاده می‌شود. همچنین فاصله دو نمونه‌داده در ویژگی‌ها، طبق رابطه (۵) به‌دست می‌آید؛ اما مهم‌ترین بخش الگوریتم، محاسبه "*تعداد ویژگی مشابه*" است. محاسبه تعداد ویژگی‌های مشابه بدین‌صورت است که در هنگام محاسبه فاصله چنانچه طبق تعریف ۳، میزان فاصله کوچکتر یا مساوی حد آستانه باشد این ویژگی یک ویژگی مشابه در نظر گرفته می‌شود. در پایان محاسبه فاصله بین دو نمونه داده، تعداد ویژگی‌های مشابه دو نمونه‌داده نیز به‌دست می‌آید. الگوریتم (۱) فرآیند روش پیشنهادی خوشه‌بندی را نشان می‌دهد.

(الگوریتم- ۱): روش پیشنهادی خوشه‌بندی
Algorithm1: Proposed Algorithm

Input: Dataset X, the number of cluster k
1. Choose k data objects in a random manner from the dataset X as the initial cluster centers.
2. For each data object in X:
   Calculate distance with Equation (5)
   Calculate number of common attributes with definition (3) and threshold
   If the nearest cluster center has maximum number of common attributes, select it as cluster for this object
   Else if the cluster center with the maximum number of common attributes, has a difference less than the threshold compared to the closest cluster center, it is selected as cluster for this object.
   Else
      assign data object to nearest cluster center.
   End For
3. Calculate similarity matrix between data objects with figure (1).
4. Update cluster centers (The data objects in the similarity matrix that have the shortest distance from other data objects are selected)
5. Repeat Step 2 until no data object changes its cluster.
6. Output: the clustering result

پیچیدگی زمانی الگوریتم، مانند بیشتر الگوریتم‌های خوشه‌بندی، O(mnki) است. که در آن m تعداد ویژگی‌ها، n تعداد نمونه‌داده‌ها، k تعداد خوشه‌ها و i تعداد تکرارهای اجرای الگوریتم است؛ اما الگوریتم پیشنهادی نسبت به سایر الگوریتم‌ها زمان محاسبه کمتری خواهد داشت. همه الگوریتم‌های خوشه‌بندی که به‌طور تصادفی مراکز خوشه را انتخاب می‌کنند، برای یافتن نمونه‌های بهتر به‌عنوان مرکز، در هر تکرار الگوریتم، فاصله هر نمونه با سایر نمونه‌ها را محاسبه می‌کنند که پردازش به‌نسبه سنگینی است. روش پیشنهادی مراکز خوشه مناسب را ابتدای الگوریتم شناسایی می‌کند؛ یعنی تعداد اجرای الگوریتم برای یافتن مراکز خوشه مناسب و ساخت خوشه، یک‌بار و یا در تعداد تکرار کم خواهد بود.

۴-آزمایش‌ها و نتایج

برای پیاده‌سازی الگوریتم‌ها از زبان برنامه‌نویسی Python 3.7 در سیستم‌عامل Windows 10 x64 استفاده شد. برای ارزیابی الگوریتم در آزمایش‌ها از مجموعه‌داده‌های Iris، Adult، Car، Credit و Contraceptive Method Choice (CMC) متعلق به UCI-Repository [24] استفاده شد. مجموعه‌داده Iris یک مجموعه‌داده تمام عددی با تعداد پنج ویژگی و شامل ۱۵۰ نمونه‌داده است. مجموعه‌داده

[DOI: 10.61186/jsdp.21.1.39]

بعدی، مجموعه Car انتخاب شد. نمونه‌داده‌های این مجموعه شامل ۶ ویژگی هستند و برخلاف مجموعه‌داده قبلی هیچ ویژگی عددی ندارد. این مجموعه‌داده نیز شامل ۱۷۲۸ نمونه داده است. برای مشاهده و ارزیابی بهتر الگوریتم، مجموعه‌داده سوم حجیم‌تر و متنوع‌تر از دو مجموعه‌داده قبلی انتخاب شد. برای آزمایش سوم مجموعه‌داده مخلوط Adult با ۳۲۵۶۱ نمونه داده انتخاب شد که هر نمونه داده شامل ۱۴ ویژگی عددی و غیرعددی است. مجموعه‌داده مخلوط Credit شامل ۶۹۰ نمونه‌داده و مجموعه‌داده مخلوط CMC شامل ۱۴۷۳ نمونه داده است. جدول شماره ۶، مشخصات مجموعه‌داده‌ها را با جزئیات بیشتری نشان می‌دهد.

(جدول-۶): مشخصات مجموعه داده‌ها

(Table-6): Datasets specifications

مجموعه داده	تعداد شی داده	تعداد ویژگی عددی	تعداد ویژگی غیرعددی	تعداد کلاس	داده ناقص
Iris	۱۵۰	۴	۰	۳	ندارد
Car	۱۷۲۸	۰	۶	۴	ندارد
adult	۳۲۵۶۱	۶	۸	۲	دارد
Credit	۶۹۰	۶	۹	۲	دارد
CMC	۱۴۷۳	۲	۷	۳	ندارد

مجموعه‌داده‌های منتخب نوفه ندارند و یا نوفه آن‌ها در مقایسه با نمونه‌های سالم بسیار ناچیز است. در آزمایش انجام‌شده این تعداد اندک نوفه‌ها نادیده گرفته می‌شود. درصورتی‌که مجموعه دارای نوفه قابل توجهی باشد، به‌حتم قبل از اجرای الگوریتم، باید عملیات پیش‌پردازش انجام گیرد.

برای ارزیابی میزان کارایی نتایج خوشه‌بندی، در آزمایش‌ها از معیارهای ACC، RI و F-measure استفاده شده‌است. در این معیارها هر چه عدد به‌دست آمده بالاتر باشد به معنای کارایی بهتر روش خوشه‌بندی است. بیشینه مقدار این معیارها برابر با ۱ است و نشان‌دهنده بهترین کارایی است.

معیار ارزیابی ACC طبق بررسی کلاس نمونه‌داده‌های هر خوشه با کلاسی که هر نمونه داده پیش از خوشه‌بندی داشت، دقت الگوریتم را ارزیابی می‌کند. یعنی هر چقدر تعداد نمونه‌داده‌های بیشتری از یک خوشه، یک کلاس مشترک و واحد داشته باشند به منزله تشکیل خوشه بهتر است. این معیار طبق رابطه (۶) [2] تعریف می‌شود:

$$\mathrm{ACC} = \frac{\sum_{i=1}^{k} a_i}{n} \qquad (۶)$$

در این رابطه k تعداد خوشه‌ها است و a_i تعداد نمونه‌هایی است که کلاس واحدی قبل و بعد از خوشه‌بندی دارند. برای سنجش دقیق و صحیح این مورد بهتر است هر خوشه را نمایندهٔ کلاسی دانست (بدیهی است کلاسی که بیشتر نمونه‌ها دارند، باید انتخاب شود). پارامتر n نیز تعداد کل نمونه‌داده‌ها است.

معیار ارزیابی(Rand Index) RI نیز یکی دیگر از معیارهای ارزیابی کارایی خوشه‌بندی است که همانند معیار قبل، با بررسی برچسب کلاس‌ها در خوشه‌های تشکیل‌شده و بررسی آن با پیش از خوشه‌بندی، میزان کارایی الگوریتم خوشه‌بندی را به‌دست می‌آورد و به‌صورت رابطه (۷) [9] تعریف می‌شود:

$$\mathrm{RI} = \frac{TP+TN}{TP+FP+TN+FN} \qquad (۷)$$

در این رابطه TP تعداد نمونه‌هایی است که با توجه به برچسب کلاس در خوشه درست قرار گرفته‌اند. TN نیز تعداد نمونه‌های متعلق به یک کلاس است که درست خوشه‌بندی شده‌اند؛ اما متعلق به خوشه هدف نبوده‌اند. FP تعداد نمونه‌های خوشه است که کلاس آن با خوشه متفاوت است و FN تعداد نمونه‌هایی است که کلاس آن‌ها با خوشه یکسان است، اما در خوشه درست قرار نگرفته‌اند. معیار ارزیابی سوم F-measure نیز همانند دو معیار قبل به‌صورت رابطه (۸)[11]، کارایی الگوریتم را در خوشه‌های تشکیل‌شده ارزیابی می‌کند:

$$\text{F-measure} = \frac{2PR}{(P+R)} \qquad (۸)$$

$$\mathrm{P} = \frac{TP}{(TP+FP)} \quad , \quad \mathrm{R} = \frac{TP}{(TP+FN)}$$

پارامتر k یا تعداد خوشه‌ها در هر آزمایش برابر با تعداد کلاس‌های هر مجموعه‌داده تنظیم شد تا با توجه به تعداد کلاس‌های مجموعه‌داده‌ها بتوان ارزیابی دقیقی برای الگوریتم خوشه‌بندی انجام داد. با توجه به نقش پارامتر حدآستانه در الگوریتم، نباید مقدار آن را عددی بزرگ تنظیم کرد. عدد بزرگ برای حدآستانه به این معنی است که مقادیر با فواصل زیاد هم به‌عنوان ویژگی مشابه شناخته شوند. در این‌صورت احتمال برابری تعداد ویژگی مشابه نمونه‌ها در مقایسه‌ها بالا می‌رود و به نوعی تأثیر این عامل در مقایسه نمونه‌داده‌ها به‌دلیل برابری، نادیده گرفته می‌شود؛ اما برای تعیین حدآستانه مناسب، الگوریتم از پنج درصد به بالا مقداردهی شد و مشاهده شد با افزایش مقدار این پارامتر تا پانزده درصد، معیارهای سنجش هم بهبود می‌یابد و در مقادیر بالاتر از آن منجر به کاهش دقت و معیارها شد. نتایج آزمایش الگوریتم با

[DOI: 10.61186/jsdp.21.1.39]

حدآستانه‌های متفاوت در مجموعه‌داده‌های متفاوت در جداول (۷) تا (۹) آمده است. در مقایسه الگوریتم با سایر روش‌ها این پارامتر با اعداد ۵ درصد، ۱۰ درصد و ۱۵ درصد مقداردهی شده‌است. گفتنی است چون برای سایر روش‌ها پارامتری برای تنظیم در الگوریتم وجود نداشت، در جدول مقایسات، مقادیر مختلف پارامتر فقط برای الگوریتم پیشنهادی مشخص شده‌است.

(جدول-۷): مقدار معیار ACC با حدآستانه‌های متفاوت
(Table-7): ACC with different thresholds

مجموعه داده	۵	۱۰	۱۵	۲۰	۲۵	۳۰	۵۰
Iris	۰.۸۵	۰.۸۸	۰.۸۹	۰.۸۴	۰.۸۳	۰.۷۹	۰.۷۵
Car	۰.۸۳	۰.۸۵	۰.۸۶	۰.۸۰	۰.۷۷	۰.۷۵	۰.۶۸
Adult	۰.۸۱	۰.۸۳	۰.۸۴	۰.۷۹	۰.۷۵	۰.۷۲	۰.۶۸
Creadit	۰.۸۱	۰.۸۳	۰.۸۵	۰.۸۰	۰.۷۸	۰.۷۵	۰.۷۱
CMC	۰.۶۴	۰.۶۶	۰.۶۸	۰.۶۳	۰.۶۰	۰.۵۸	۰.۵۲

(جدول-۸): مقدار معیار RI با حدآستانه‌های متفاوت
(Table-8): RI with different thresholds

مجموعه داده	۵	۱۰	۱۵	۲۰	۲۵	۳۰	۵۰
Iris	۰.۸۸	۰.۹۰	۰.۹۲	۰.۸۷	۰.۸۵	۰.۸۱	۰.۷۵
Car	۰.۸۴	۰.۸۶	۰.۸۷	۰.۸۲	۰.۸۰	۰.۷۷	۰.۷۲
adult	۰.۷۷	۰.۷۹	۰.۸۶	۰.۷۵	۰.۷۲	۰.۶۹	۰.۶۵
Creadit	۰.۷۸	۰.۷۹	۰.۸۱	۰.۷۵	۰.۷۲	۰.۷۱	۰.۶۳
CMC	۰.۶۶	۰.۶۸	۰.۷۰	۰.۶۳	۰.۶۲	۰.۶۰	۰.۵۷

(جدول-۹): مقدار معیار F-Measure با حدآستانه‌های متفاوت
(Table-9): F-Measure with different thresholds

مجموعه داده	۵	۱۰	۱۵	۲۰	۲۵	۳۰	۵۰
Iris	۰.۷۴	۰.۸۳	۰.۸۱	۰.۷۳	۰.۷۱	۰.۶۸	۰.۶۵
Car	۰.۷۶	۰.۷۹	۰.۸۱	۰.۷۴	۰.۷۲	۰.۷۰	۰.۶۵
adult	۰.۷۴	۰.۷۴	۰.۸۳	۰.۷۲	۰.۷۰	۰.۶۸	۰.۶۲
Creadit	۰.۷۴	۰.۷۵	۰.۸۲	۰.۷۲	۰.۷۰	۰.۶۸	۰.۶۲
CMC	۰.۵۹	۰.۶۴	۰.۶۵	۰.۵۷	۰.۵۳	۰.۵۲	۰.۴۸

در بخش مقدمه اشاره شد که نتیجه عملیات خوشه‌بندی و به دنبال آن میزان دقت خوشه‌بندی ممکن است در هر اجرا متفاوت باشد و بستگی به مراکز خوشه اولیه دارد. یعنی در هر بار اجرا ممکن است نتایج متفاوتی داشته باشد و این اصل در مورد همه روش‌های خوشه‌بندی صادق است. لذا همانند اکثر روش‌ها برای مشخص شدن نتایج، الگوریتم پیشنهادی به دفعات اجرا شد و نتیجه‌ای که بیشترین تکرار را داشت، به عنوان مقدار معیار سنجش، مشخص شد. میزان تفاوت نتایج بین ۰.۰۱ تا ۰.۰۵ بود.

الگوریتم پیشنهادی با پنج الگوریتم k-prototypes [8]، SKC [4]، Multiview [5]، DPC-MD [12] و Equi-based k-prototypes [6] مقایسه شد. نتایج ارزیابی الگوریتم پیشنهادی با مجموعه‌داده Iris در جدول ۱۰ قابل مشاهده است. همانطور که مشاهده می‌شود کارایی الگوریتم پیشنهادی با حدآستانه ۱۵ درصد در دو معیار نزدیک به یکی از الگوریتم‌ها بود و از سایر الگوریتم‌ها حداقل ۰.۰۲ بهتر عمل کرد. در معیار سوم نتیجه الگوریتم پیشنهادی از همه روش‌ها بهتر است. این روش با حدآستانه پایین‌تر، ضعیف‌تر از الگوریتم Multiview و Equi-based عمل کرده است و همچنین نتیجه تقریبا برابری با الگوریتم SKC دارد. در این آزمایشات اهمیت تنظیم حد آستانه مناسب برای الگوریتم پیشنهادی قابل توجه است.

(جدول-۱۰): نتایج خوشه‌بندی با مجموعه داده Iris
(Table-10): Clustering Results with Iris Dataset

الگوریتم		ACC	RI	F-measure
الگوریتم پیشنهادی	Tr = 5%	۰.۸۵	۰.۸۸	۰.۷۴
	Tr = 10%	۰.۸۸	۰.۹۰	۰.۸۳
	Tr = 15%	**۰.۹۱**	**۰.۹۲**	**۰.۸۴**
K-Prototype		۰.۹۰	۰.۸۴	۰.۷۷
SKC		۰.۸۸	۰.۸۸	۰.۷۸
Multiview		**۰.۹۱**	۰.۹۰	**۰.۸۴**
DPC-MD		۰.۸۳	۰.۹۰	۰.۷۵
Equi-Based K-Prototype		۰.۹۰	۰.۸۴	۰.۷۷

نتایج ارزیابی الگوریتم پیشنهادی با مجموعه‌داده Car در جدول (۱۱) قابل مشاهده است. همانطور که مشاهده می‌شود کارایی الگوریتم پیشنهادی در دو حدآستانه‌ی مقداردهی شده، نتیجه بهتری از دیگر روش‌ها در معیار RI داشته است. در دو معیار دیگر نیز الگوریتم پیشنهادی نسبت به سه روش بهتر و نزدیک به الگوریتم برتر بوده است.

(جدول-۱۱): نتایج خوشه‌بندی با مجموعه‌داده Car
(Table-11): Clustering Results with Car Dataset

الگوریتم		ACC	RI	F-measure
الگوریتم پیشنهادی	Tr = 5%	۰.۸۳	۰.۸۴	۰.۷۶
	Tr = 10%	۰.۸۵	۰.۸۶	۰.۷۹
	Tr = 15%	۰.۸۶	**۰.۸۷**	**۰.۸۱**
K-Prototype		۰.۷۴	۰.۵۱	۰.۵۰
SKC		**۰.۸۷**	۰.۸۵	**۰.۸۱**
Multiview		۰.۷۶	۰.۵۷	۰.۵۸
DPC-MD		۰.۷۰	۰.۵۳	۰.۵۲
Equi-Based K-Prototype		۰.۸۶	۰.۶۶	۰۵۸

[DOI: 10.61186/jsdp.21.1.39]

نتایج ارزیابی الگوریتم پیشنهادی با مجموعه‌داده‌ی مخلوط و حجیم Adult در جدول ۱۲ قابل مشاهده است. همانطور که مشاهده می‌شود الگوریتم پیشنهادی با حدآستانه‌ی بالا نتیجه‌ی بهتری نسبت به چهار الگوریتم داشته است و نسبت به الگوریتم برتر در هر سه معیار با حداقل اختلاف ضعیف‌تر عمل کرده است.

(جدول-۱۲): نتایج خوشه‌بندی با مجموعه‌داده Adult
(Table-12): Clustering Results with Adult Dataset

F-measure	RI	ACC	الگوریتم	
۰.۷۴	۰.۷۷	۰.۸۱	Tr = 5%	الگوریتم پیشنهادی
۰.۷۴	۰.۷۹	۰.۸۳	Tr = 10%	
۰.۸۳	۰.۸۶	۰.۸۴	Tr = 15%	
۰.۷۱	۰.۷۵	۰.۷۸	K-Prototype	
۰.۹۲	**۰.۹۳**	**۰.۹۵**	SKC	
۰.۷۹	۰.۷۸	۰.۸۲	Multiview	
۰.۸۲	۰.۷۹	۰.۸۰	DPC-MD	
۰.۷۵	۰.۷۸	۰.۷۹	Equi-Based K-Prototype	

نتایج ارزیابی الگوریتم پیشنهادی با مجموعه‌داده‌ی مخلوط Credit در جدول (۱۳) قابل مشاهده است. در هر سه معیار سنجش، الگوریتم پیشنهادی نتایج بهتری داشته است. حتی در حد آستانه‌های پایین، الگوریتم پیشنهادی باز هم نتیجه بهتر یا نزدیکی ثبت کرده است.

(جدول-۱۳): نتایج خوشه‌بندی با مجموعه داده Credit
(Table-13): Clustering Results with Credit Dataset

F-measure	RI	ACC	الگوریتم	
۰.۷۴	۰.۷۸	۰.۸۱	Tr = 5%	الگوریتم پیشنهادی
۰.۷۵	۰.۷۹	۰.۸۳	Tr = 10%	
۰.۸۲	**۰.۸۱**	**۰.۸۵**	Tr = 15%	
۰.۵۸	۰.۵۰	۰.۶۶	K-Prototype	
۰.۷۶	۰.۷۷	۰.۸۱	SKC	
۰.۷۱	۰.۶۹	۰.۸۱	Multiview	
۰.۸۰	۰.۷۹	۰.۸۴	DPC-MD	
۰.۷۶	۰.۷۰	۰.۷۴	Equi-Based K-Prototype	

نتایج ارزیابی الگوریتم پیشنهادی با مجموعه‌داده‌ی مخلوط CMC در جدول ۱۴ قابل مشاهده است. الگوریتم پیشنهادی در دو معیار نتیجه کاملا برتر و در یک معیار برتر از چهار الگوریتم و نزدیک به الگوریتم دیگر عمل کرده است. با توجه به نتایج می‌توان گفت الگوریتم پیشنهادی برای مجموعه‌داده‌های مخلوط که حجم داده زیادی نداشته باشند بهتر عمل می‌کند.

(جدول-۱۴): نتایج خوشه‌بندی با مجموعه داده CMC
(Table-14): Clustering Results with CMC Dataset

F-measure	RI	ACC	الگوریتم	
۰.۵۹	۰.۶۶	۰.۶۴	Tr = 5%	الگوریتم پیشنهادی
۰.۶۴	۰.۶۸	۰.۶۶	Tr = 10%	
۰.۶۵	**۰.۷۰**	**۰.۶۸**	Tr = 15%	
۰.۵۱	۰.۶۷	۰.۵۵	K-Prototype	
۰.۶۴	۰.۶۹	۰.۶۷	SKC	
۰.۶۰	۰.۶۸	۰.۵۷	Multiview	
۰.۶۰	۰.۶۸	۰.۶۴	DPC-MD	
۰.۶۴	**۰.۷۰**	۰.۶۵	Equi-Based K-Prototype	

۵-جمع‌بندی

در اکثر روش‌های خوشه‌بندی، صرف نظر از نوع مجموعه‌داده ساده یا مخلوط، فاصله نمونه‌داده‌ها محاسبه می‌شود و مبنای اختصاص به خوشه‌ها درنظر گرفته می‌شود [5]. عوامل مختلفی در خوشه‌بندی تاثیرگذار هستند [13]. مهمترین ویژگی روش پیشنهادی این است که شباهت دو نمونه داده را تنها میزان فاصله آن‌ها در نظر نمی‌گیرد. این روش نشان می‌دهد توجه به پارامترهای دیگر در کنار فاصله می‌تواند میزان شباهت دقیق‌تری از دو نمونه داده را به‌دست آورد که به کمک آن، کارایی الگوریتم و دقت تشکیل خوشه‌ها افزایش پیدا می‌کند. برای این منظور پارامتری با عنوان تعداد ویژگی مشابه تعریف شد. سپس خوشه‌بندی با الگوریتم پیشنهادی با رعایت ملاحظاتی علاوه بر فاصله نمونه‌داده‌ها، تعداد ویژگی‌های مشابه را نیز بررسی می‌کرد و این دو در کنار هم خوشه‌ی مناسب را انتخاب می‌کرد. طبق آزمایشات، این روش بهبودهایی را در مقایسه با برخی روش‌ها نشان داد و دقت حاصل از خوشه‌بندی با الگوریتم پیشنهادی افزایش پیدا کرد. همچنین با تغییرات و مطالعات بیشتر در بخش‌هایی از روش پیشنهادی، دقت الگوریتم می‌تواند بهبود بیشتری یابد. بطور مثال استفاده از روش‌هایی برای تعیین مراکز اولیه بجای مراکز تصادفی یکی از این بخش-هاست. یعنی در شروع کار الگوریتم، مراکز اولیه با یک هوشمندی و الگوریتم مشخص شوند، زیرا طبق مطالعات انجام گرفته شده یکی از عوامل تاثیرگذار در کیفیت خوشه‌های تشکیل شده، مراکز اولیه خوشه‌ها هستند. همچنین استفاده از نوعی الگوریتم با دقت بالاتر برای محاسبه فاصله در بخش غیرعددی داده‌های مخلوط بجای فرمول ساده‌ی همینگ نیز باعث افزایش دقت و نتایج خوشه‌بندی خواهد شد. با این کار نوعی هوشمندی در محاسبه میزان فاصله و شباهت مقادیر غیرعددی بکار

[DOI: 10.61186/jsdp.21.1.39]

گرفته می‌شود که منجر به دستیابی به فاصله دقیق‌تر در مجموعه‌داده‌های مخلوط خواهد شد. مطالعه و پژوهش برای تعیین حدآستانه هوشمند و افزودن وزن به ویژگی‌ها از دیگر مواردی است که باعث بهبود و افزایش دقت در الگوریتم پیشنهادی خواهد شد.

6-Reference ۶-مراجع

[1]Ahmad, Amir, and Shehroz S. Khan. "Survey of state-of-the-art mixed data clustering algorithms." Ieee Access 7 (2019): 31883-31902.

[2] Ahmad, Amir, and Shehroz S. Khan. "initKmix-A novel initial partition generation algorithm for clustering mixed data using k-means-based clustering." Expert Systems with Applications 167 (2021): 114149.

[3] Behzadi, Sahar, et al. "Clustering of mixed-type data considering concept hierarchies: problem specification and algorithm." International Journal of Data Science and Analytics 10.3 (2020): 233-248.

[4] Kumar, Pradeep, and Anita Kanavalli. "A Similarity based K-Means Clustering Technique for Categorical Data in Data Mining Application." International Journal of Intelligent Engineering and Systems 14.2 (2021): 43-51.

[5] Ji, Jinchao, et al. "A Multi-View Clustering Algorithm for Mixed Numeric and Categorical Data." IEEE Access 9 (2021): 24913-24924.

[6] Sangam, Ravi Sankar, and Hari Om. "An equi-biased k-prototypes algorithm for clustering mixed-type data." Sādhanā 43.3 (2018): 1-12.

[7] Yuan, Fang, Youlong Yang, and Tiantian Yuan. "A dissimilarity measure for mixed nominal and ordinal attribute data in k-Modes algorithm." Applied Intelligence 50.5 (2020): 1498-1509

[8] Jia, Ziqi, and Ling Song. "Weighted k-Prototypes Clustering Algorithm Based on the Hybrid Dissimilarity Coefficient." Mathematical Problems in Engineering 2020 (2020).

[9] Jia, Hong, Yiu-ming Cheung, and Jiming Liu. "A new distance metric for unsupervised learning of categorical data." IEEE transactions on neural networks and learning systems 27.5 (2015): 1065-1079.

[10] Ji, Jinchao, et al. "Clustering mixed numeric and categorical data with artificial bee colony strategy." Journal of Intelligent & Fuzzy Systems 36.2 (2019): 1521-1530.

[11 Skabar, Andrew. "Clustering Mixed-Attribute Data using Random Walk." Procedia Computer Science 108 (2017): 988-997.

[12] Du, Mingjing, Shifei Ding, and Yu Xue. "A novel density peaks clustering algorithm for mixed data." Pattern Recognition Letters 97 (2017): 46-53

[13] Qian, Yuhua, et al. "Space structure and clustering of categorical data." IEEE transactions on neural networks and learning systems 27.10 (2015): 2047-2059.

[14] dos Santos, Tiago RL, and Luis E. Zárate. "Categorical data clustering: What similarity measure to recommend?. " Expert Systems with Applications 42.3 (2015): 1247-1260.

[15] Ahmad, Amir, and Sarosh Hashmi. "K-Harmonic means type clustering algorithm for mixed datasets." Applied Soft Computing 48 (2016): 39-49.

[16] Ji, Jinchao, et al. "An initialization method for clustering mixed numeric and categorical data based on the density and distance." International Journal of Pattern Recognition and Artificial Intelligence 29.07 (2015): 1550024.

[17] van de Velden, Michel, Alfonso Iodice D'Enza, and Angelos Markos. "Distance-based clustering of mixed data." Wiley Interdisciplinary Reviews: Computational Statistics 11.3 (2019): e1456.

[18] Caruso, Giulia, et al. "Cluster analysis: An application to a real mixed-type data set." Models and Theories in Social Systems. Springer, Cham, 2019. 525-533.

[19] Jinyin, Chen, et al. "A novel cluster center fast determination clustering algorithm." Applied Soft Computing 57 (2017): 539-555

[20] Xiong, Jing, and Hong Yu. "An adaptive three-way clustering algorithm for mixed-type data." International Symposium on Methodologies for Intelligent Systems. Springer, Cham, 2018.

[21] Dinh, Duy-Tai, and Van-Nam Huynh. "k-PbC: an improved cluster center initialization for categorical data clustering." Applied Intelligence (2020): 1-23.

[22] Hsu, Chung-Chian, and Yan-Ping Huang. "Incremental clustering of mixed data based on distance hierarchy." Expert systems with applications 35.3 (2008): 1177-1185.

[23] Ahmad, Amir, and Lipika Dey. "A k-mean clustering algorithm for mixed numeric and categorical data." Data & Knowledge Engineering 63.2 (2007): 503-527.

[24] UCI Repository. https://archive.ics.uci.edu/ml/datasets.html. (September 6, 2021).

حمید رضایی کاردانی را در دانشکده فنی شهید شمسی پور و کارشناسی را در دانشگاه شهید رجایی و در رشته مهندسی نرم‌افزار گذرانده است. در حال حاضر دانشجوی مقطع کارشناسی ارشد رشته مهندسی نرم افزار دانشگاه شهید رجایی و زمینه مطالعاتی ایشان داده‌کاوی و پایگاه‌داده است.

نشانی رایانامه ایشان عبارت است از:

hid.rezaei@gmail.com

[DOI: 10.61186/jsdp.21.1.39]

نگین دانشپور درجه کارشناسی ارشد و دکترای خود را در رشته مهندسی کامپیوتر-نرم‌افزار از دانشگاه صنعتی امیرکبیر در سال‌های ۱۳۸۰ و ۱۳۸۹ اخذ کرده‌است. وی در حال حاضر عضو هیئت علمی و دانشیار دانشکده مهندسی کامپیوتر دانشگاه تربیت دبیر شهید رجایی است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: داده‌کاوی و پیش‌پردازش و مدیریت داده‌ها، پایگاه داده تحلیلی و سیستم‌های تصمیم‌یار.

نشانی رایانامه ایشان عبارت است از:

ndaneshpour@sru.ac.ir

[DOI: 10.61186/jsdp.21.1.39]