Arab J Sci Eng (2018) 43:3979–3992
https://doi.org/10.1007/s13369-017-2761-2

RESEARCH ARTICLE - SPECIAL ISSUE - COMPUTER ENGINEERING AND COMPUTER SCIENCE

An Attributes Similarity-Based *K*-Medoids Clustering Technique in Data Mining

G. Surya Narayana[1] · **D. Vasumathi**[2]

Received: 31 March 2017 / Accepted: 17 July 2017 / Published online: 8 August 2017

Abstract In recent days, mining data in the form of information and knowledge from large databases is one of the demanding and task. Finding similarity between different attributes in a synthetic dataset is an aggressive concept in data retrieval applications. For this purpose, some of the clustering techniques are proposed in the existing works such as *k*-means, fuzzy *c*-means, and fuzzy *k*-means. But it has some drawbacks that include high overhead, less effective results, computation complexity, high time consumption, and memory utilization. To overcome these drawbacks, a similarity-based categorical data clustering technique is proposed. Here, the similarities of inter- and intra-attributes are simultaneously calculated and it is integrated to improve the performance. The dataset loaded as input, where the preprocessing is performed to remove the noise. Once the data are noise free, the similarity between the elements is computed; then, the most relevant attributes are selected and the insignificant attributes are neglected. The support and confidence measures are estimated by applying association rule mining for resource planning. The similarity-based *K*-medoids clustering technique is used to cluster the attributes based on the Euclidean distance to reduce the overhead. Finally, the bee colony (BC) optimization technique is used to select the optimal features for further use. In experiments, the results of the proposed clustering system are estimated and analyzed with respect to the clustering accuracy, execution time (s), error rate, convergence time (s), and adjusted Rand index (ARI). From the results, it is observed that the proposed technique provides better results when compared to the other techniques.

Keywords Data mining · *K*-medoids clustering · Association rule mining (ARM) · Preprocessing · Inter- and intra-similarity estimation, and Euclidean distance

✉ G. Surya Narayana
surya.aits@gmail.com

D. Vasumathi
rochan44@gmail.com

[1] JNTUH, Hyderabad, Telangana State, India

[2] CSE Department, JNTUCEH, Hyderabad, Telangana State, India

1 Introduction

Data mining is defined as the process of extracting some meaningful information from the database for further use. In this domain [1], cluster analysis is an important and essential task in unsupervised machine learning and statistical multivariate analysis. The numerous developments in the field of information and communication (IC) increase the dimensionality of the data, which leads to the difficulties during the knowledge extraction. The data mining aims to cluster the large datasets with diverse attributes of different types. The process of discovering the knowledge from large and sparse database by partitioning it into several disjoint groups is known as clustering. The main aim of clustering is to group the set of objects into a cluster. So, the objects that present in the same cluster have high similarity and in the dissimilar clusters have low similarity. The major steps involved in clustering are preprocessing, mining, and result validation. Preprocessing is defined as the process of removing an unwanted data or filling the missed values. Typically, several tasks are sequentially performed to mine the useful information from the input data. The evaluation of the mined data using various assessment algorithms constitutes the final stage known as result validation.

The good clustering technique [2] must satisfy the following properties:

Scalability—It performs well for the large number of datasets.
Analyze the mixture of attributes—It has the ability to analyze both the single and mixture of attributes.
Find arbitrary-shaped clusters—To determine the shape or the bias is not an easy task. This algorithm has the capability to find the kinds of clusters based on the shape.
Minimum requirement for input parameters—Most of the clustering algorithms need some user-defined parameters like the number of clusters for data analyze.
Handling of noise—The algorithms must handle the deviations, which is defined as the data objects that depart from the norms of behavior in order to enhance the clustering quality.
Sensitivity to the order of input records—The algorithm must insensitive to the order of inputs.
High dimensionality of data—The algorithm must be capable to handle the large datasets. The dataset contains a large number of attributes, so the clustering algorithm cannot able to handle more dimensions.

1.1 Existing Works

Celebi et al. [3] presented an overview of clustering techniques to solve the problem of numerical initialization. Based on the study, the most popular initialization method (IM) was introduced for the *k*-means clustering. In this work, different large-size datasets were used for proving the clustering performance of the IM technique. But, the *K*-means clustering techniques have some of the disadvantages, which includes:

- It required more number of clusters.
- It was capable to detect only the hyper spherical clusters.
- Moreover, it was more sensitive to noise that affects the respective clusters.

Ghosh and Dubey [4] compared the *k*-means and fuzzy *C*-means (FCM) clustering techniques based on their efficiency for identifying the best clustering technique in data mining. Typically, this clustering technique analyzed the data based on the locations between various input data points. The FCM was an unsupervised clustering technique that was mainly applied in the fields of agriculture, astronomy, chemistry, geology, image analysis, classifier design, and clustering. From the investigation, it was analyzed that the FCM clustering technique provides an efficient clustering results compared with the *k*-means clustering technique. Velmurugan [5] analyzed both the *k*-means and FCM clustering techniques for connection-oriented telecommunication data. Here, the performance of these algorithms was evaluated based on the connection-oriented broadband area. This paper also stated that the FCM technique was more accurate and easy to understand, when compared to the *k*-means clustering.

Wang et al. [6] introduced a coupled attribute similarity for objects (CASO) technique for clustering the large size data. In this work, the inter-coupled and intra-coupled attributes are considered for improving the accuracy and reducing the complexity. Based on the attribute types, that were classified into two types such as discrete and continuous. From the paper, it was inferred that the categorical data-based clustering techniques were more suitable for the large size data. Joshi and Kaur [7] compared different clustering techniques in data mining for selecting the best one. The techniques surveyed were as follows:

- *K*-means clustering,
- Hierarchical clustering,
- Density-based spatial clustering,
- Ordering points to identify clustering structure,
- Statistical information grid.

From the analysis, it was analyzed that the existing techniques have both merits and demerits. But, the *K*-means clustering provides the better results compared to the other techniques. Mukhopathy et al. [8] surveyed various multi-objective evolutionary algorithms for the purpose of clustering, association rule mining, and other data mining tasks. The major drawback of the binary encoding scheme was not suitable for clustering, when the number of attributes were large. Moreover, two different multi-objective rule mining algorithms were reviewed in this paper, which includes:

- Multi-objective differential evolution-based numeric association rule mining (MODENAR).
- Multi-objective differential evolution (MODE).

In addition, three different types of data such as categorical, numerical, and fuzzy data clustering techniques are investigated in the work. From the evaluation, it was noticed that the categorical data clustering techniques efficiently cluster the huge size data with large number of attributes.

Arora et al. [9] evaluated two different clustering algorithms that include *k*-means and *k*-medoids on dataset transaction 10k. Here, the drawbacks of *k*-means algorithm were mentioned, which includes:

- Finding the value of K is a difficult task.
- It is not efficient.
- it do not handled the different size and different density clusters.

Also, this paper stated that the *k*-medoids clustering technique overwhelmed the disadvantages of *k*-means and provides the better results concerning with execution time, non-sensitive, reduced noise, and minimized sum of dissimilar objects.

Harikumar and Surya [10] suggested a similarity-based clustering technique, namely *k*-medoids for heterogeneous datasets. Here, the distance between the objects is identified with the heterogeneous attribute types by using the suggested technique. The main aim of *k*-medoids clustering technique was to identify a set of non-overlapping clusters named as medoids. Choi and Chung [11] developed a *k*-partitioning algorithm to cluster the large spatio-textual data. The major contributions of this paper were as follows:

- The problem of clustering a spatio-textual data was investigated, where the spatio-clustering has the process of social data analysis, location-based data cleaning and preprocessing of spatial keyword querying.
- Here, the expected pairwise distance was utilized to implement the modified version of *k*-means clustering.

The demerit of this work was it needs to generalize the clustering scheme with different types of textual distances. Mei and Chen [12] recommended a *k*-medoid clustering technique for relational data, where the objects in each fuzzy cluster bring their degrees of representativeness in that cluster. Moreover, the prototype weights and the fuzzy memberships in each cluster were attained by using the quadratic regularization technique. Galluccio et al. [13] introduced a minimal spanning tree (MST) to measure the distance for clustering the high-dimensional data. Here, the non-convex-shaped clusters were separated by implementing the powerful clustering technique. The main advantage of this paper was it reduced the computational complexity of the dual rooted MSTs. Jiang et al. [14] clustered uncertain objects based on the similarity between the probability distributions. The Gauss transform has the linear complexity, so it was equipped with the randomized *k*-medoids in order to scale the large datasets. The advantage of this technique was it performed the scalable clustering tasks with moderate accuracy.

Chatti and Rao [15] investigated various data mining techniques in a dynamic environment by using the fuzzy clustering technique. The main goal of this model was to identify the deviations in the data, based on this, to adjust the input parameters. The major drawback of this paper was it do not consider the issues of complexity, noise, and more accurate results. Sunil Raj et al. [16] surveyed various clustering methods that include partitioning, density-based technique, hierarchical technique, and grid-based technique in data mining. In this paper, the positives and negative attributes of each clustering technique were discussed. Also, the BIRCH and Chameleon techniques were utilized to overcome the problem of object swapping between the clusters. Zadegan et al. [17] introduced a rank-based partitioning algorithm to cluster the large datasets in a fast and accurate manner. Here, two types of validation measures such as internal and external were utilized during the results evaluation. Moreover, the quality of the resulting clusters were improved by capturing the internal cluster structure. Sood and Bansal [18] suggested a combination of Bat algorithm with the *k*-medoids algorithm to improve the efficiency of clustering. The location of the cluster was initialized by using the Bat algorithm to select the initial representative object in the *k*-medoids technique. Moreover, the path was recovered with minimum complexity by using the swarm intelligence technique.

Skabar and Abdalgader [19] suggested a fuzzy relational clustering algorithm to cluster the sentence level text. The main aim of this technique was to identify the semantically overlapping clusters based on the related sentences. Here, the similarities between the objects were estimated in the form of a square matrix. Kulkarni and Kinariwala [20] introduced a fuzzy relational eigenvector centrality-based clustering algorithm (FRECCA) based on the mixture model. In this work, the mixing coefficients and the cluster membership values were optimized by using the expected maximization (EM) algorithm. Here, the importance of object in a network was computed with the help page rank mechanism. Moreover, this algorithm identified the similarity between the documents and text summarization by using the potential application. The major disadvantage of this technique was time complexity. Kameswaran and Malarvizhi [21] surveyed various clustering techniques in data mining to extract the large information from the dataset. In this investigation, the hierarchical algorithm, partitioning algorithm, density-based algorithm, graph-based algorithm, and grid-based algorithms were studied with its advantages and disadvantages. From the survey, it was inferred that the graph-based algorithm outperforms the other clustering algorithms. Ghadiri et al. [22] suggested an active distance-based clustering technique by using the *k*-medoids technique. Here, the unknown distances between the clusters were identified during an active clustering process. In this analysis, the real world and the synthesized datasets were utilized to evaluate the clustering technique. During the utilization, the performance improvement for entity recognition and retrieval depends on the discovery of alternative forms of attribute values. Li et al. [23] proposed the novel compact clustering framework that jointly identified synonyms for the set of attribute values. The integration of signals from the multiple information sources into the similarity function and the optimization of weights of

signals through the unsupervised process assured the effectiveness of mining the entity attribute systems. With the large-scale datasets, the clustering-based recommender system suffered from the accuracy and coverage limitations. Guo et al. [24] employed the support vector regression model that predicted the given item based on the prediction-related features. The insufficiency in data representation affected the clustering performance adversely. Hence, they proposed the probabilistic method that derived the predictions from the views regarding the prediction relationship through the optimization framework. The gathering of the relevant information from the cluster was difficult issue due to the large size data handling. Balabantaray et al. [25] accomplished the K-means or K-medoids algorithms and found the best clustering algorithms for document summarization. On the basis of the sentence weight, the document summarization was executed with the focus of the key points from the whole document that provided easy retrieval performance. The simultaneous minimization of intra-cluster distances and the maximization of inter-cluster distances were the difficult task in textbook machine learning problem. Grossi et al. [26] presented the constraint programming model that made the standard problem as easier one and allowed the generation of interesting variants, respectively. The important aspects of the constraint programming model were density-based clustering and the label propagation approach. The emerging of web tables in the information retrieval applications required the selection of ranking of tables and the summarization of meaningful content. Nguyen et al. [27] formalized the issues as the diversified table selection problem and the structured table summarization problems. They presented the heuristic algorithms to assure the near-optimal, stable, and fairness. The high scalability was achieved by using the web table searching process. During the clustering process, the initial selection of center, accuracy, and the ability of clustering algorithms were the major issues. Zhang et al. [28] provided the clustering algorithm on the basis of the artificial bee colony (ABC) optimization algorithm with the quick convergence property increased the accuracy and reliability of the system. In this survey, the advantages and disadvantages of each and every paper were analyzed. Then, it is identified that the existing clustering techniques have some of the major disadvantages.

1.2 Motivation of the Proposed Work

In order to solve those issues, this research work focuses to propose an efficient clustering technique for large size data in data mining. Based on the problem identification, this work as the following contributions:

1. To cluster the categorical data based on inter- and intra-attribute similarity measure.
2. To improve the performance resource planning by implementing the association rule mining concept with support and confidence estimation.
3. To reduce the computational overhead of clustering process by filtering the data and attributes.

2 Materials and Methods

The clear description about the proposed similarity-based K-medoids clustering technique for extracting the information from large data is presented in this segment. The overall flow of the proposed system is shown in Fig. 1. It includes the following stages:

- Preprocessing
- Similarity computation
- Attributes filtering
- Association rule mining
- K-medoids clustering
- Euclidean distance estimation

At first, the given dataset is preprocessed by eliminating the unwanted attributes in the dataset. Then, the similarities between the inter-attributes and intra-attributes are estimated, where the categorical data are converted into the numerical data. After that, the similarities of both are integrated and it is filtered for further processing. Here, the association rule learning process is applied in which the support and confidence values are estimated. Again, the data are filtered before the starting the clustering process. In this work, the K-medoids clustering technique is applied to group the similar attributes based on the Euclidean distance. After clustering, the evolutionary optimization algorithm, namely BC, is applied to obtained the optimized data that are used for further process.

Algorithm I – Categorical data clustering

D ← load dataset;
Nod ← no of data;
A ← attributes;
for i = 0: Nod do
  if Nod_i.Name == null then
    remove Nod_i.Data;
  end if;
end for;
Dda ← Distinct data in attributes

for i = 0: A do
  Dda = A.i;
end for;
Nc ← Numerical conversion
for i = 0: A do
  Nc = A.i;
end for;
S ← Support
for i = 0: Dda do
  S.i = Calculate support;
end for;
Iaas ← Intra attribute similarity
for i = 0: S do
  for j = 0: S do
    $Iaas = \frac{S.i*S.j}{S.i+S.j+(S.i*S.j)}$
  end for;
end for;
Oiaas ← Overall intra attribute similarity
for i =0: S do
  $Oiaas = \sum_0^{S.i.size} Iaas.i$
end for;
Ieas ← Inter attribute similarity
For I = 0: A do
  For j = 0: A do
    If A.i != A.j then
      $Ieas = \frac{count(A.i.i==A.j.j)}{distinct(A.i)}$
    End if
  End for
End for
Oieas ← Overall inter attribute similarity
For i = 0: S do
  $Oieas = \sum_0^{S.i.size} Ieas.i$
End for
Is ← Integrated similarity
For i = 0: S do
  $Is = Iaas.i + Ieas.i$
End for;
Nis ← Normalized integrated similarity
For i = 0: Is do
  $Nis = \frac{(Is.i-Is.min)}{Is.max-Is.min}$
End for;
Af ← Attribute filtering
T ← Threshold
For i = 0: Nis do
  if Nis.i < T then
    Af = Nis.i
  End if
End for
S ← Support
Soa ← Set of attributes
$S = \frac{supp(Soa)}{N}$
C ← confidence
$C = \frac{supp(Soa_x \cup Soa_y)}{Soa_x}$
Df ← Data filtering
For i = 0: C do
  If C.i > T then
    Df = C.i;
  End if;
End for;
Noc ← Number of clusters
Soch ← Set of cluster heads
Ch ← Cluster heads
Dis ← Distance
Cl ← Cluster label
Md ← Minimum distance
Sc ← Selected Clusters
While true do
For i = 0: Soch do
  For j = 0: Ch do
    For k = 0: D do
      $Dis = \sqrt{(Ch_1 - D_1)^2 + (Ch_2 - D_2)^2 + \cdots + (Ch_n - D_n)^2}$
    End for
  End for
CI ← Assign cluster labels
End for
If Dis < Md then
  Md = Dis
  Sc ← Selected clusters
End if;
End while;
Eb ← Employed bees
Fit ← Fitness calculation
Ob ← Onlooker bees
For i = 0: Noc do
  Eb ← assign employeed bees
  For j = 0: Nod do
    Fit ← Assign fitness values
  End for;
  Ob ← Subset information
End for;
Pr ← Probability
Bs ← Best subset
For i = 0:Ob do
  Bs ← Get best subset
End for;

2.1 Preprocessing

Preprocessing is an imperative and essential task in many data mining applications. It efficiently removes the unwanted attributes in a dataset for improving the performance of clustering. In this stage, the adult dataset is given as the input, which contains the attributes of age, work class, final weight, education number, marital status, occupation, relationship, race, gender, capital gain, capital loss, hours per week, native country, and salary. Once the dataset is noise free, the distinct attributes are identified for similarity computation.

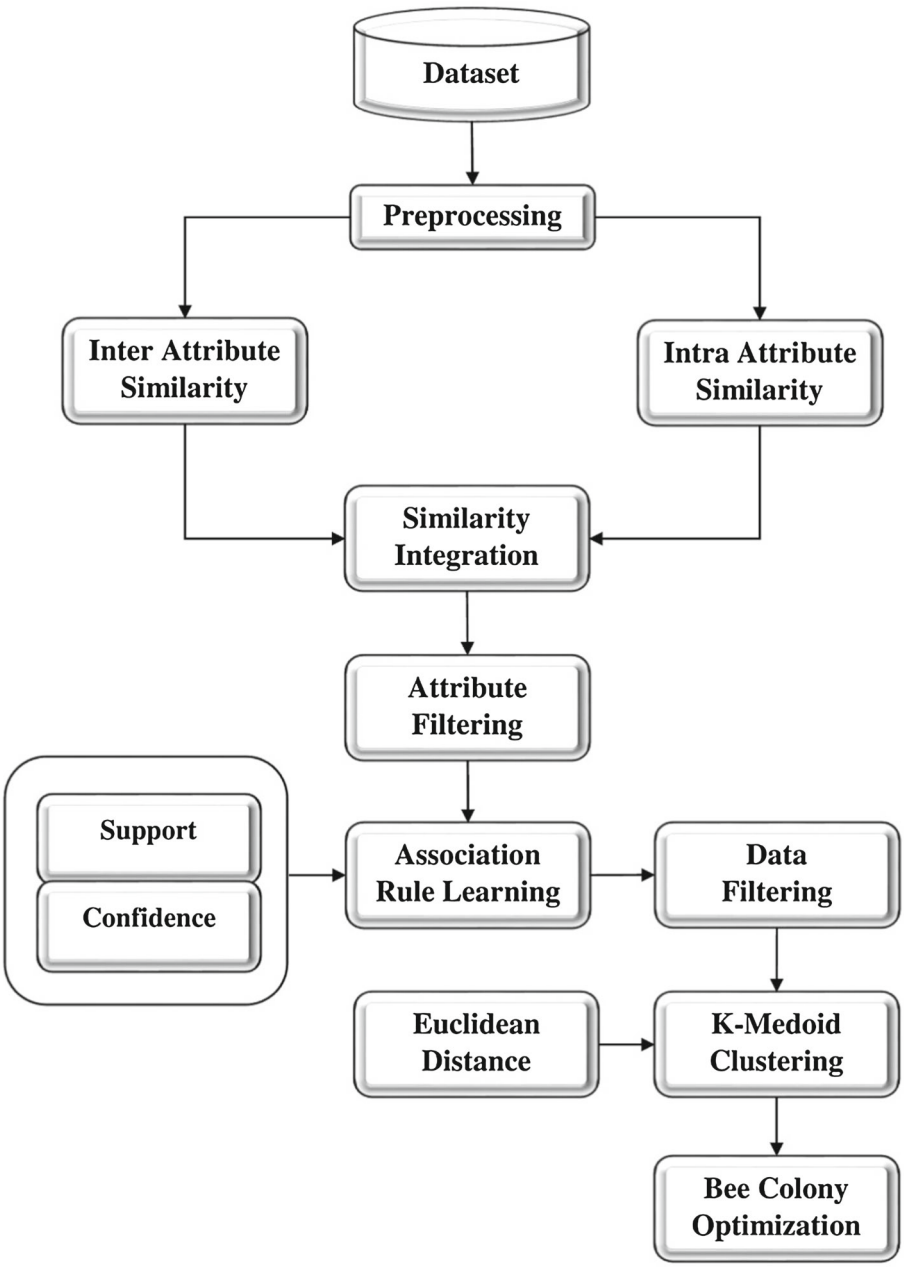

Fig. 1 Flow of the proposed system

2.2 Similarity Computation

In the recent days, similarity analysis is a tedious process in various domains. Typically, the similarity-based clustering is performed based on the followings:

- Between attributes,
- Between clusters,
- Between data objects,
- Between attribute values.

In this work, the inter-attribute and intra-attribute similarity analysis are performed after preprocessing. In intra-attribute similarity, the similarity is estimated between the same column attributes. It reveals that the greater similarity is assigned to the particular attribute based on the approximation of equal frequencies. The similarity between the closer two values is identified, if the frequency is high. Different frequencies indicate distinct levels of attribute value significance. Then, the inter-attributes are selected by finding the similarity between the rows. It does not involve the couplings between attribute values during the calculation of attribute values. If the values occur with the same relative frequency, this estimation determines that the values are similar. After calculating inter- and intra-similarity attributes, it is integrated into a single attribute similarity. Then, it is filtered for further process-

ing. The attribute couplings are categorized into two, namely intra-coupled and inter-coupled in this paper. The discrepancies occurred in attribute value estimation reflected the similarity in terms of the frequency distribution. The relationship between the attribute value frequencies is proposed as the intra-coupled similarity with the satisfaction of above principles. The intra-coupled attribute similarity for values (IaASV) among the values corresponding to the attribute are defined as follows:

$$\text{Iaas} = \frac{\left|G_j\left(v_j^x\right)\right| . \left|G_j\left(v_j^y\right)\right|}{\left|G_j\left(v_j^x\right)\right| + \left|G_j\left(v_j^y\right)\right| + \left|G_j\left(v_j^x\right)\right| . \left|G_j\left(v_j^y\right)\right|} \tag{1}$$

where v_i^x, v_j^y—specific values for objects

G—mapping function of value of attribute to the dependent object set

In proposed work, the attribute relationship is defined through the support values (S), and hence, Eq. (1) is reformulated as follows:

$$\text{Iaas} = \frac{S.i * S.j}{S.i + S.j + (S.i * S.j)} \tag{2}$$

From Eqs. (1) and (2), it is observed that the problematic issue arises if the values for attribute have the same frequency. To overcome this issue, the inter-coupled similarity estimation takes place. The inter-coupled similarity defines the interaction between the values within the attribute $\left(a_j\right)$ without considering the couplings between the attributes. The definition of inter-coupled similarity is expressed as

$$\text{Ieas} = \min_{V_k' \subseteq V_k}\left(2 - P_{k|j}\left(V_k' | v_j^x\right) - P_{k|j}\left(V_{k|}' v_j^x\right)\right) \tag{3}$$

The conditional probabilities of value of attributes depend on the count of similar and distinct values in the attribute set. Hence, the equation (3) formulated as

$$\text{Ieas} = \frac{\text{count}\left(A.i.i == A.j.j\right)}{\text{distinct}\left(A.i\right)} \tag{4}$$

The overall similarity is the combination of inter- and intra-similarity measures as follows:

$$\text{Is} = \text{Iaas} + \text{Ieas} \tag{5}$$

The normalization of similarity measures contributes to the further processing with support and confidence value as

$$\text{Nis} = \frac{(\text{Is.i} - \text{Is.min})}{\text{Is.max} - \text{Is.min}} \tag{6}$$

where Nis—normalized similarity

Is.max—maximum similarity value.

Is.min—minimum similarity value.

The attributes that having equal frequency distributions are assigned to the maximum similarity. Hence, the similar frequencies indicate the closeness and the dissimilar frequencies indicate the distinct levels. The similarity results of inter- and intra-attributes are graphically shown in Fig. 2.

2.3 Association Rule Mining

After filtering the similarity attributes, the association rule mining is applied to mine the attributes, where the support and confidence values are estimated. Typically, the association rule mining satisfies the minimum support and confidence constraints by discovering the small set of rules in a database. It handles the weighted association rule mining

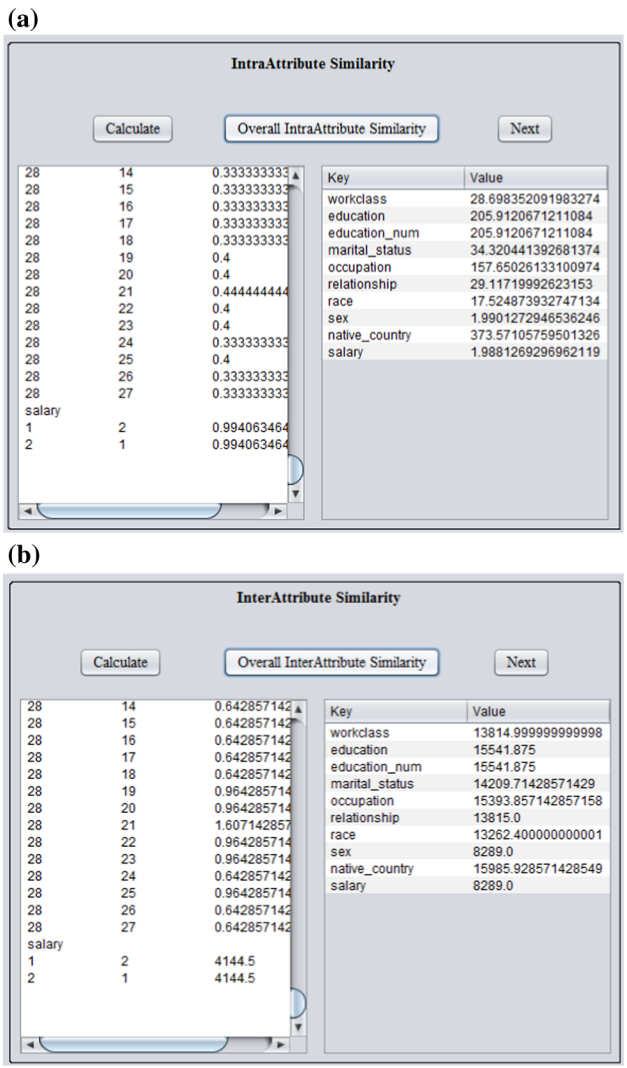

Fig. 2 **a** Intra-attribute similarity. **b** Inter-attribute similarity

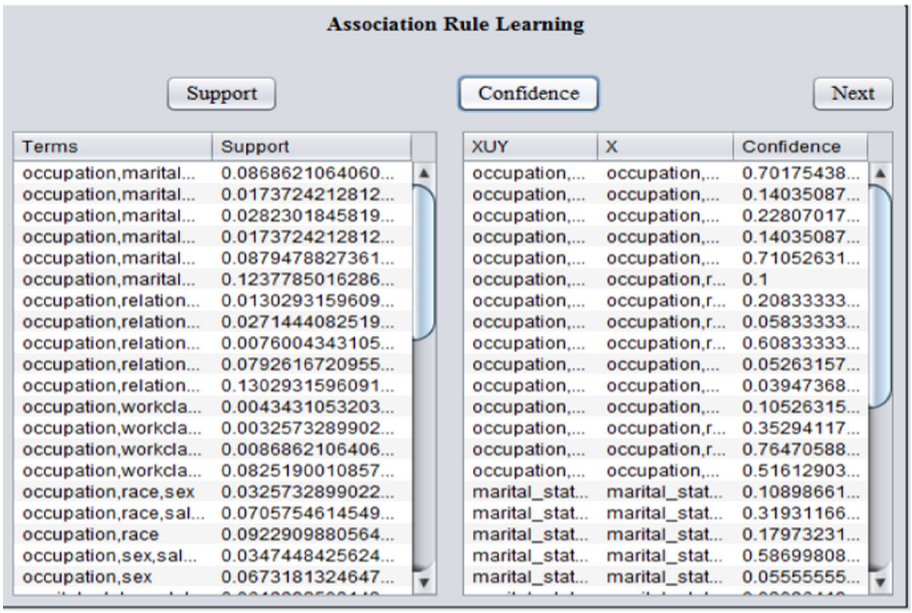

Fig. 3 Support and confidence estimation

problems, and it assumes that the items have the same significance. Here, the target of mining is not predetermined and it is adapted to mine all the classes by satisfying the constraints. It works based on the following steps:

- It satisfies the minimum support value by generating the large item set.
- It satisfies the minimum confidence value by generating all the association rules using the large item sets.

Generally, an item set contains a set of items that has the transaction support above minimum support. The key element of this mining is to prune the search space by limiting the number of generated rules. In some of the applications, the frequently occurred items on the dataset may vary.

The set of attributes is denoted as (Soa) is used to measure those measures as follows:

Support The measure of the frequency of rule within the transactions refers to support and such rule $(A \Rightarrow B)$ involves the great part of the dataset for high support values.

$$\text{supp}\,(A \Rightarrow B) = p\,(A \cup B) \tag{7}$$

In this paper, the support value for the sensitive attributes is formulated as

$$S = \frac{\text{supp}\,(\text{Soa})}{N} \tag{8}$$

Confidence The measure of the percentage of transactions containing A which contain also B refers the confidence value. The mathematical formulation for the confidence estimation is conditional probability estimation that is represented as

$$\text{Confidence}\,(C) = P\left(\frac{B}{A}\right) = \text{supp}(A, B)/\text{supp}\,(A) \tag{9}$$

This formulation is modified with sensitive attributes as follows:

$$C = \frac{\text{supp}\,(\text{Soa}_A \cup \text{Soa}_B)}{\text{Soa}_B} \tag{10}$$

The estimation of support and confidence measures are shown in Fig. 3.

2.4 *K*-Medoids Clustering

Normally, clustering the categorical data is a demanding and critical task, because of many fields that dealt with categorical data. For this purpose, some of the techniques are developed in the existing work for clustering the categorical data. The traditional clustering techniques including fuzzy-C-means (FCM), k-means, and hierarchical techniques are suitable for clustering the categorical data. Due to its limitations, the k-medoids clustering technique is employed in this work. It forms the cluster based on the distance between the data points, and the cluster heads are selected for each group. In this algorithm, a medoid is used as a reference point of a mostly centrally located object in a cluster. Based on the principles, the sum of dissimilarities between the objects are minimized. Moreover, it finds the medoids of

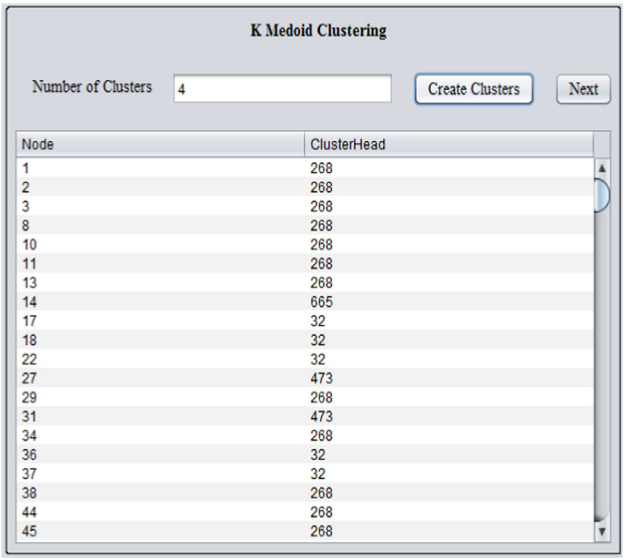

Fig. 4 Output of K-medoids clustering

each cluster by identifying the k number of clusters in an object. Instead of using the mean value of the object, this algorithm uses the representative objects in each cluster. The major advantages of k-medoids clustering technique are as follows:

- It has no limitations on attribute types.
- The medoids are selected based on the location of a predominant fraction of the points in the cluster.
- It is less sensitive to outliers.

Due to these advantages, the k-medoids technique provides the better results, when compared to the traditional clustering techniques. The operating phases in the K-medoids clustering are initialization, iterative, and clustering. The prediction of potential set of medoids by using the greedy approach constitutes the initialization phase. In this phase, the sample points are selected randomly and apply the greedy technique to obtain the small set of sample points with the size is equal to B.K. Here, B = small integer and K = clusters. The iterative phase determines the quality of the cluster formed by replacing the bad medoids through the measure of Davies–Bouldin Index (DBI) defined as follows:

$$\mathrm{DB} = 1/K \sum_{j=1}^{K} \max D_{i,j} \tag{11}$$

where, $D_{i,j}$ − − − distance ratio

$$= \sqrt{(Ch_1 - D_1)^2 + (Ch_2 - D_2)^2 + \cdots + (Ch_n - D_n)^2} \tag{12}$$

The clustering results of k-medoid technique are visualized in Fig. 4.

2.5 Bee Colony for Optimized Dataset

After clustering the data, it is further optimized by implementing an evolutionary-based optimization algorithm. In this work, the bee colony (BC) optimization algorithm is employed to select the optimized data for further use. It is a type of swarm-based optimization algorithm that is inspired by the intelligent foraging behavior of the honey bees. It is mainly used to solve more complex optimization problems in the search space based on the fitness value of the solution. Also, it provides the better final solutions from the constructive moves, when compared to the other optimization algorithms such as artificial bee colony optimization (ABC) [29], bee swarm optimization (BSO) [30], and bees algorithm [31]. The agents in the BCO are termed as artificial bees that are located in the initial stage of search process. Furthermore, each agent performs the local moves to construct the solution for optimization. The quality of the obtained final solution is improved by using the parallelization strategy. In which, the meta-heuristics are used to represent the searching solutions even in the repeated sequential iterations. The parallel execution provides an efficient search in the solutions by improving the quality of final solution with minimum execution time. Due to these reasons, it is considered as an effective technique for solving the non-standard optimization problems based on multiple criteria. The main intention of using this algorithm is to develop a multi-agent system for solving combinatorial optimization problems.

Moreover, it contains a population of solutions in which the fitness is evaluated based on the quality of the food source. The major reasons for using this algorithm are as follows:

- It handling complex problems with large dataset.
- It identified the high quality optimal solutions,
- It provides the balance between the performance and complexity,

Algorithm II – Bee Colony Optimization

```
Eb ← Employed bees;
Fit ← Fitness Calculation;
Ob ← Onlooker bees;
for i = 0: Noc do
    Eb ← Assign employed bees;
    for j = 0: Nod do
        Fit ← Assign fitness values;
    end for;
    Ob ← Subset information;
end for;
Pr ← Probability;
Bs ← Best Subset;
for I = 0: Ob do
    Bs ← Get best subset;
end for;
```

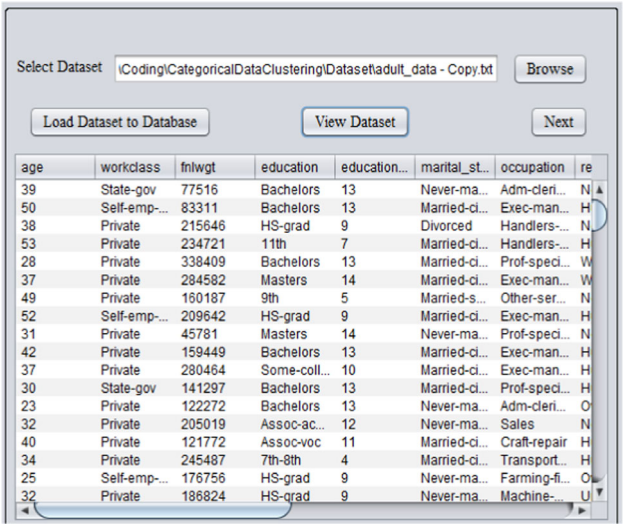

Fig. 5 Dataset loading

3 Results and Discussion

This section evaluates the performance results of both existing and proposed techniques for proving the superiority of the proposed system. Here, the results are analyzed and evaluated in terms of clustering error, convergence time, accuracy, ARI, execution time, and clustering accuracy. The datasets used in this work are adult dataset [32], chess dataset [33], and connect-4 dataset [34], which are the benchmark datasets obtained from the UCI machine learning repository. The adult dataset contains the fields of age, work class, final weight, education, education number, marital status, occupation, relationship, race, gender, capital gain, capital loss, hours per week, native country, and salary, which is shown in Fig. 5. The adult dataset is a kind of multivariate that contains both the categorical and integer attributes, in which the number of instances 48,842 is used, and the number of attributes 14 is used. Then, the chess dataset contains the fields of white king file (column), white king rank (row), white rook file, white rook rank, black king file, black king rank, and optimal depth of win for white. It is also a type of multivariate dataset that contains both integer and categorical attributes, in which 28,056 number of instances are used, 6 number of attributes are used, and there is no missing values in the dataset. The connect-4 dataset contains the fields of *x* player, *o* player, and blank, and this dataset has the characteristics of multivariate and spatial, and it contains only the categorical attributes. In which, 67,557 number of instances are used, 42 number of attributes are used, and it does not have any missing values.

Table 1 Clustering error versus different clustering algorithms

K-means	*K*-prototype	OCIL	Similarity-based *K*-medoids clustering
0.3869	0.3855	0.249	0.2231

Table 2 Clustering error versus different datasets

Chess	Connect-4	Mushroom	Adult dataset
0.2543	0.2175	0.243	0.2231

3.1 Clustering Error

Clustering error is the error rate that is occurred due to the process of clustering. Table 1 shows the error rate of both existing and proposed clustering techniques, and Table 2 shows the clustering error with respect to various datasets. The existing algorithms considered in this analysis are *K*-means, *K*-prototype, and OCIL. The clustering error is estimated as follows:

$$\text{Error} = \frac{\sum_{i=1}^{N} \delta\left(a_i, \text{map}\left(m_i\right)\right)}{N} \tag{13}$$

where N represents the number of instances in the dataset, a_i represents the provided label, m_i indicates the mapping function that maps the obtained cluster label. The clustering error is computed as $e = 1 - \text{Error}$. From the results, it is observed that the proposed similarity-based *k*-medoids clustering technique reduced the error rate to 0.23, when compared to the other techniques. In *k*-means algorithm, it is difficult to predict the *k* value, and it does not provide the better clustering results with global cluster. Also, different initial partitions result in varying final clusters, and it does not work well with the clusters of different size and separate density. Then, in the *k*-prototype algorithm, the process converges not to a global minimum, but to a local minimum. Furthermore, the similarity computation processes that performed in the OCIL algorithm are low. Due to these drawbacks, the proposed similarity-based *k*-medoids clustering technique provides the better results.

3.2 Convergence Time

The convergence time of the existing [35] and proposed techniques is evaluated on the categorical data clustering, which is represented in Table 3 and, the convergence time with respect to different datasets is shown in Table 4. The *k*-modes and OCIL techniques require more convergence time, due to its computational cost. The superiority of the proposed system is proved by taking the real time adult dataset and consider-

Table 3 Convergence time (s) versus different clustering algorithms

K-prototype	OCIL	Similarity-based *K*-medoids clustering
15.2795	3.5447	3.475

Table 4 Convergence time versus different datasets

Chess	Connect-4	Mushroom	Adult dataset
4.36	5.78	3.512	3.475

Table 5 Accuracy of the proposed technique with different datasets

Adult	Chess	Connect-4	Mushroom
95.2	94.8	94.1	94.3

Table 6 Clustering accuracy of existing and proposed techniques

K-modes	Chan's	Wk-modes	Similarity-based *K*-medoids clustering
0.6583	0.6583	0.6583	0.7042

ing the convergence time for this dataset. From the analysis, it is evaluated that the proposed similarity-based *k*-medoids technique provides the minimized convergence time (s). Due to the drawbacks of the existing *k*-means, *k*-prototype, and OCIL techniques, the proposed similarity-based *k*-medoids clustering provides the reduced clustering error and convergence time.

3.3 Accuracy

The performance of the clustering algorithms is evaluated based on the measure of clustering accuracy. Table 5 shows the clustering accuracy of the proposed similarity-based *k*-medoids clustering technique with respect to different datasets. Table 6 shows the clustering accuracy of both existing and proposed clustering techniques. The cluster C is partitioned into a set of clusters $\{c_1, c_2, \ldots c_{k3}\}$ on a dataset O with n number of objects, and the clustering accuracy is calculated as follows:

$$\text{Clustering accuracy} = \frac{\sum_{i=1}^{k} c_l}{|O|} \tag{14}$$

where k is the number of clusters desired, c_l is the number of objects that occurred in cluster C_i, and $|O| = n$ represents the number of objects in the dataset. From the analysis, it is evaluated that the proposed similarity-based *k*-medoids clustering technique provides better clustering accuracy compared to

Table 7 Clustering accuracy of the proposed technique with and without optimization process

Approaches	Clustering accuracy
With optimization	95.8
Without optimization	95.2

Table 8 Clustering accuracy with respect to different datasets

Chess	Connect-4	Mushroom	Adult dataset
0.692	0.654	0.659	0.7042

the other techniques for the datasets including adult, chess, connect-4, and mushroom.

Table 7 shows the clustering accuracy analysis without and with optimization techniques in detail. The accuracy without optimization is 95.2 and 95.8% with optimization. The provision of optimization to the clusters derived from the *K*-medoids algorithm improved the accuracy considerably by 0.6% that assures the effectiveness of the proposed work.

Table 8 shows the clustering accuracy of the proposed technique with respect to different datasets for proving the betterment of the proposed technique.

3.4 Adjusted Rand Index (ARI)

The adjusted rand index is defined as an external criterion that measures the similarity between two partitions of an objects in the same dataset. Let $X = \{X_1, X_2, \ldots X_k\}$ and $X' = \{X_1', X_2' \ldots X_k'\}$ be two partitions on a dataset O with n objects. Then N_{ij} be the number of objects in a cluster X_i in partition X and in cluster X_i in partition X', $H_{ij} = \{X_i \cap X_j'\}$. Based on this, the rand index is calculated as follows:

$$\text{Adjusted Rand Index}\left(X, X'\right) = \frac{b_0 - b_3}{0.5(b_1 + b_2) - b_3} \tag{15}$$

where, $b_0 = \sum_{i=1}^{k}\sum_{j=1}^{k\prime}\binom{H_{ij}}{2}$, $b_1 = \sum_{i=1}^{k}\binom{X_i}{2}$, $b2 = \sum_{i=1}^{k\prime}\binom{X_j'}{2}$, $r3 = \frac{2b_1 b_2}{H(H-1)}$, $\binom{n}{m}$ represents the binomial coefficient. The value of ARI is high, if the clustering result is more close to the true class distribution. Table 9 shows the ARI of both existing [36] and proposed techniques. In this analysis, it is proved that the proposed technique provides highest ARI value. In the *k*-modes algorithm, the information gain and entropy calculation lead to increased time consumption, then in the Chan's algorithm, if the same attribute values in some dimension in the cluster is labeled as 1, this means that the rest of the attributes are ignored. Also, the wk-modes algorithm requires more computation

Table 9 ARI of existing and proposed techniques

K-modes	Chan's	Wk-modes	Similarity-based *K*-medoids clustering
0.0016	−0.0026	0.0019	0.0023

Table 10 Cluster data dimensionality analysis

No. of clusters	Data dimensionality	
	Without optimization	With optimization
Cluster 1	209	189
Cluster 2	96	87
Cluster 3	97	85
Cluster 4	89	72

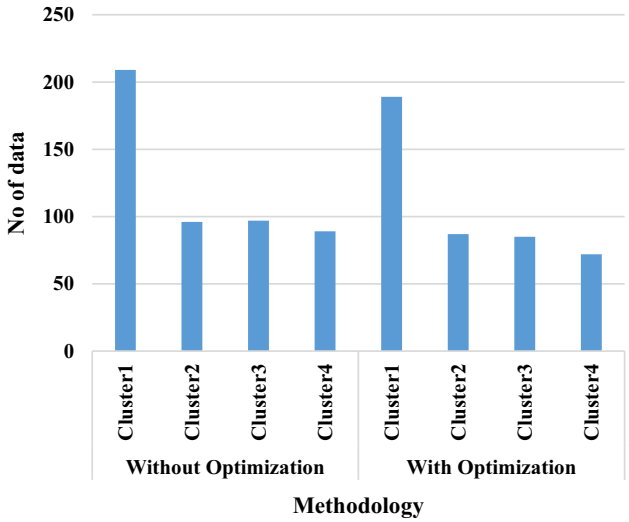

Fig. 6 Data dimensionality analysis

for weight calculation, and its iterations are also high. Due to these drawbacks, the proposed technique has an increased ARI value.

3.5 Data Dimensionality

If the size of the data occupied in the cluster is minimum, then the manipulation of attributes is easy task in mining applications. The prediction of interrelationship between the data within the cluster through the proposed work consumes minimum time effectively. Table 10 and Fig. 6 show the dimensionality variation for each cluster corresponding to without and with optimization.

The overall count of data within the cluster 1 is 209 without optimization, and it is reduced to 189 due to the evolutionary optimization. The comparative analysis of proposed system without and with optimization states that the proposed work reduced the dimension by 9.57%. Similarly, dimensionality of other clusters (2, 3, and 4) is 11.46, 12.37 and 19.01%, respectively.

3.6 Execution Time

Execution time is defined as the amount of time taken for clustering the data. Table 11 shows the execution time of the proposed similarity-based *k*-medoids clustering technique with respect to four different datasets that include adult, chess, connect-4, and mushroom. Typically, the execution time is evaluated in terms of seconds. From the analysis, it is observed that the proposed technique requires the minimum execution time for processing all the datasets.

Table 12 shows the analysis of the execution time variations without and with optimization. The execution time without optimization is 28 s, and it is 31 s with optimization. The fitness formulation and the iterative process of updating the positions consume additional time for the relevant attribute selection.

Table 11 Execution time

Adult	Chess	Connect-4	Mushroom
28	22	31	25

Table 12 Execution time analysis without and with optimization

Approaches	Execution time (s)
With optimization	31
Without optimization	28

3.7 Average Best

Average best is defined as a measure of identifying the best technique that a better optimization solutions. Here, the convergence curve values of the PSO [37], GSA, WOA [38], and bee colony optimization techniques are illustrated in Table 13. The PSO can easily fall into local optimum in a high-dimensional space and has a low convergence rate during the iterative process, then the GSA used only one mathematical formulate to update the position of search agents, which increases the likeliness of stagnation in the local optima. Also, the WSO requires more computation time for optimization. Due to these problems, the proposed bee colony optimization provides the better convergence results.

Table 13 Convergence graph

Iteration	PSO	GSA	WOA	Bee colony
1	600	600	400	350
250	300	200	150	124
500	200	200	90	76

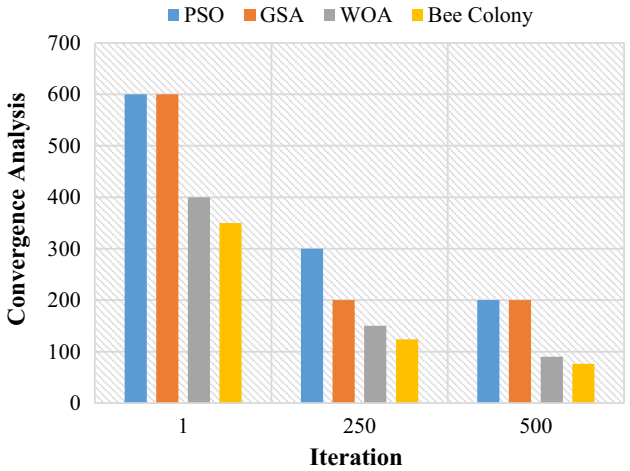

Fig. 7 Convergence analysis between the optimization techniques

4 Conclusion and Future Work

This paper proposed a similarity-based *K*-medoids clustering technique for processing the large datasets. The intention of this technique is to reduce the clustering overhead during the process of clustering and similarity computation. Initially, the adult dataset is given as the input of preprocessing, which eliminates the unwanted attributes in the dataset. After noise removal, inter- and intra-similarity values are estimated between the attributes in the dataset. The inter-attribute similarity is estimated between two different rows, and the intra-attribute similarity is estimated the attributes in the same column. After that, the similarity values are integrated and filtered for further processing. Then, the association rule mining is applied to mine the data based on the minimum support and confidence values. Again, it is filtered and clustered by using the *k*-medoids clustering technique. It efficiently groups the cluster based on its similarity and Euclidean distance. After clustering, the bee colony optimization algorithm is implemented to select the optimized data for further processing. The major advantages of the proposed techniques are low computational complexity, reduced time consumption, and highly efficient. The effectiveness of the proposed technique is proved by comparing it with the existing techniques. Different datasets are also used to prove the superiority of the proposed similarity-based *K*-medoids clustering technique. When compared to the other techniques, the proposed technique provides the best results.

In future, this work will be enhanced by developing a new clustering technique to increase the performance of the system.

References

1. Verma, A.; Kaur, I.; Kaur, A.: Algorithmic approach to data mining and classification techniques. Indian J. Sci. Technol. (IJST) (Association rule mining, classification, clustering, data, data mining, decision tree, neural network) **9**(28), 1–22 (2016)
2. Gayathri, S.; Mary, Metilda M.; Sanjai, Babu S.: A shared nearest neighbour density based clustering approach on a Proclus method to cluster high dimensional data. Indian J. Sci. Technol. (IJST) (Density based approach, high dimensional data, Proclus, SNN algorithm) **8**(22), 1–6 (2015)
3. Celebi, M.E.; Kingravi, H.A.; Vela, P.A.: A comparative study of efficient initialization methods for the *k*-means clustering algorithm. Expert Syst Appl. **40**(1), 200–210 (2013)
4. Ghosh, S.; Dubey, S.K.: Comparative analysis of k-means and fuzzy c-means algorithms. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **4**(4), 35–39 (2013)
5. Velmurugan, T.: Performance based analysis between *k*-means and fuzzy *C*-means clustering algorithms for connection oriented telecommunication data. Appl. Soft Comput. **19**, 134–46 (2014)
6. Wang, C.; Dong, X.; Zhou, F.; Cao, L.; Chi, C.-H.: Coupled attribute similarity learning on categorical data. IEEE Trans. Neural Netw. Learn. Syst. **26**(4), 781–97 (2015)
7. Joshi, A.; Kaur, R.: A review: comparative study of various clustering techniques in data mining. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **3**(3), 67–70 (2013)
8. Mukhopadhyay, A.; Maulik, U.; Bandyopadhyay, S.; Coello, C.A.C.: Survey of multiobjective evolutionary algorithms for data mining: part II. IEEE Trans. Evolut. Comput. **18**(1), 20–35 (2014)
9. Arora, P.; Varshney, S.: Analysis of *K*-means and *K*-medoids algorithm for big data. Proced. Comput. Sci. **78**, 507–512 (2016)
10. Harikumar, S.; Surya, P.: *K*-medoid clustering for heterogeneous datasets. Proced. Comput. Sci. **70**, 226–37 (2015)
11. Choi, D.-W.; Chung, C.-W.: A *K*-partitioning algorithm for clustering large-scale spatio-textual data. Inf. Syst. **64**, 1–11 (2017)
12. Mei, J.-P.; Chen, L.: Fuzzy clustering with weighted medoids for relational data. Pattern Recognit. **43**(5), 1964–74 (2010)
13. Galluccio, L.; Michel, O.; Comon, P.; Kliger, M.; Hero, A.O.: Clustering with a new distance measure based on a dual-rooted tree. Inf. Sci. **251**, 96–113 (2013)
14. Jiang, B.; Pei, J.; Tao, Y.; Lin, X.: Clustering uncertain data based on probability distribution similarity. IEEE Trans. Knowl. Data Eng. **25**(4), 751–63 (2013)
15. Subbalakshmi, G.R.; Rao, S.K.M. (eds.) Evaluation of data mining strategies using fuzzy clustering in dynamic environment. In: Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics. Springer, Berlin (2016)
16. Raj, Y.S.; Rajan, A.P.; Charles, S.; Raj, S.A.J.: Clustering methods and algorithms in data mining: concepts and a study. J. Comput. Technol. **4**(7), 8–11 (2015)
17. Zadegan, S.M.R.; Mirzaie, M.; Sadoughi, F.: Ranked *k*-medoids: a fast and accurate rank-based partitioning algorithm for clustering large datasets. Knowl. Based Syst. **39**, 133–43 (2013)
18. Sood, M.; Bansal, S.: *K*-medoids clustering technique using bat algorithm. Int. J. Appl. Inf. Syst. **5**(8), 20–2 (2013)
19. Skabar, A.; Abdalgader, K.: Clustering sentence-level text using a novel fuzzy relational clustering algorithm. IEEE Trans. Knowl. Data Eng. **25**(1), 62–75 (2013)

20. Kulkarni, B.M.; Kinariwala, S.: Review on fuzzy approach to sentence level text clustering. Int. J. Sci. Res. Educ. **3**(06), 3845–3850 (2015)
21. Kameshwaran, K.; Malarvizhi, K.: Survey on clustering techniques in data mining. Int. J. Comput. Sci. Inf. Technol. **5**(2), 2272–6 (2014)
22. Ghadiri, M.; Aghaee, A.; Baghshah, M.S.: Active distance-based clustering using *K*-medoids. (2015). arXiv preprint arXiv:1512.03953 [cs.LG]
23. Li, Y.; Hsu, B.-J.P.; Zhai, C.; Wang, K. (eds.) Mining entity attribute synonyms via compact clustering. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, ACM (2013)
24. Guo, G.; Zhang, J.; Yorke-Smith, N.: Leveraging multiviews of trust and similarity to enhance clustering-based recommender systems. Knowl. Based Syst. **74**, 14–27 (2015)
25. Balabantaray, R.C.: Sarma, C.; Jha, M.: Document clustering using *K*-means and *K*-medoids. (2015). arXiv preprint arXiv:1502.07938 [cs.IR]
26. Grossi, V.; Monreale, A.; Nanni, M.; Pedreschi, D.; Turini, F. (eds.) Clustering formulation using constraint optimization. In: International Conference on Software Engineering and Formal Methods. Springer (2015)
27. Nguyen, T.T.; Nguyen, Q.V.H.; Weidlich, M.; Aberer, K. (eds.) Result selection and summarization for web table search. In: 2015 IEEE 31st International Conference on Data Engineering (ICDE), IEEE (2015)
28. Zhang, D.; Luo, K.: Clustering algorithm based on artificial bee colony optimization. In: International Conference on Applied Science and Engineering Innovation (ASEI) (2015)
29. Ozturk, C.; Hancer, E.; Karaboga, D.: Dynamic clustering with improved binary artificial bee colony algorithm. Appl. Soft Comput. **28**, 69–80 (2015)
30. Djenouri, Y.; Drias, H.; Habbas, Z.: Bees swarm optimisation using multiple strategies for association rule mining. Int. J. Bioinspir. Comput. **6**(4), 239–49 (2014)
31. Karaboga, D.; Gorkemli, B.; Ozturk, C.; Karaboga, N.: A comprehensive survey: artificial bee colony (ABC) algorithm and applications. Artif. Intell. Rev. **42**(1), 21–57 (2014)
32. Becker, B.: Adult data set (2015). https://archive.ics.uci.edu/ml/datasets/adult
33. Bain, M.; Hoff, A.V. Chess (King-Rook vs. King) data set (2015). https://archive.ics.uci.edu/ml/datasets/Chess+(King-Rook+vs.+King)
34. Tromp J.: Connect-4 data set (2015). https://archive.ics.uci.edu/ml/datasets/Connect-4
35. Cheung, Y.-M.; Jia, H.: Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. Pattern Recognit. **46**(8), 2228–38 (2013)
36. Cao, F.; Liang, J.; Li, D.; Zhao, X.: A weighting *k* modes algorithm for subspace clustering of categorical data. Neurocomputing **108**, 23–30 (2013)
37. Vora, P.; Oza, B.: A survey on k mean clustering and particle swarm optimization. Int. J. Sci. Mod. Eng. (IJISME) **1**(3), 24–26 (2013)
38. Mirjalili, S.; Lewis, A.: The whale optimization algorithm. Adv. Eng. Softw. **95**, 51–67 (2016)