Reclust: an efficient clustering algorithm for mixed data based on reclustering and cluster validation

Amala Jayanthi Maria Soosai Arockiam[1], Elizabeth Shanthi Irudhayaraj[2]
[1]Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, India
[2]Department of Computer Science, Avinashilingam Institute of Home Science and Higher Studies, Coimbatore, India

Article Info

Article history:

Received Apr 8, 2022
Revised Sep 15, 2022
Accepted Oct 12, 2022

Keywords:

Clustering
Hierarchical cluster
K-means
Mixed data
Numerical and categorical data
Re-clustering

ABSTRACT

Clustering is a significant approach in data mining, which seeks to find groups or clusters of data. Both numeric and categorical features are frequently used to define the data in real-world applications. Several different clustering algorithms are proposed for the numerical and categorical datasets. In clustering algorithms, the quality of clustering results is evaluated using cluster validation. This paper proposes an efficient clustering algorithm for mixed numerical and categorical data using re-clustering and cluster validation. Initially, the mixed dataset is clustered with four traditional clustering algorithms like expectation-maximization (EM), hierarchical cluster (HC), k-means (KM), and self-organizing map (SOM). These four algorithms are validated, and the best algorithm is selected for re-clustering. It is an iterative process for improving the quality of cluster results. The incorrectly clustered data is iteratively re-clustered and evaluated based on the cluster validation. The performance of the proposed clustering method is evaluated with a real-time dataset in terms of purity, normalized mutual information, rand index, precision, and recall. The experimental results have shown that the proposed reclust algorithm achieves better performance compared to other clustering algorithms.

Corresponding Author:

Amala Jayanthi Maria Soosai Arockiam
Department of Computer Science, Kumaraguru College of Technology
Coimbatore, Tamil Nadu, India
Email: research.amala@gmail.com

1. INTRODUCTION

Clustering analysis is one of the most important approaches in data mining, and it seeks to determine the nature of groupings or clusters of data objects in attributes space. Clustering methods are employed in a variety of applications [1], including social network analysis [2], knowledge discovery, image processing, text and sentiment analysis [3]. Clustering analysis seeks to group data objects with similar properties together, and those with distinct characteristics into separate clusters. Hierarchical and partitional clustering methods are the two types of clustering algorithms [4]. Data are dispersed into a dendrogram of layered segments using a split or agglomerative technique in hierarchical clustering algorithms. Data are partitioned into a certain number of clusters by minimizing an objective cost function in partitional clustering algorithms.

For specific kinds of information, clustering algorithms have been developed. Continuous values are used to represent numerical data, whereas categorical data, which is a subset of discrete data, can only have a finite number of values. Many real-world applications use categorical data, such as name, gender, and educational level. Both numerical and category values were present in the mixed datasets. Real-world data is frequently of various sorts. Medical data, for example, includes categorical and numerical values such as age, height, weight, and salary, as well as categorical and numerical values such as nationality, gender, employment, education [5], marital status, and chest pain type [6]. When a dataset comprises both numerical and categorical variables, the issue of determining the similarity of two data becomes more complicated [7]. Splitting the numeric and categorical elements of a mixed dataset and finding the Euclidean distance between two data points for numeric characteristics and the Hamming distance for categorical features is a simple technique for solving the similarity problem [8].

For clustering mixed data, several techniques have been developed. To cluster heterogeneous data, Huang [9] presented the well-known k-prototypes technique, which merged the k-means and k-modes approaches. The k-prototypes algorithm was improved in [10] by incorporating attribute influence and enhancing the cluster center representation. The unsupervised feature learning (UFL) approach was developed by Lam *et al*. [11] by combining the fuzzy adaptive resonance theory (ART) with the UFL. The approach Kay-means for mixed large data sets (KAMILA) introduced by Foss *et al*. [12] can directly deal with multiple types of attributes and requires fewer parameters. Chen and He [13] used the principle of density clustering to present a self-adaptive peak density clustering technique. Most mixed data clustering algorithms have two main goals: to develop new approaches to construct novel measures of similarity between mixed characteristics and to cluster data using previous or new strategies to obtain a local optimum result.

This paper proposes an efficient clustering algorithm for mixed numerical and categorical data based on re-clustering and cluster validation called reclust. The proposed method contains three important processes: initial clustering, validation, and re-clustering. The initial clustering process uses four traditional clustering algorithms such as expectation-maximization (EM), hierarchical cluster (HC), k-means (KM), and self-organizing map (SOM). The validation process evaluates the clustering result. The re-clustering process re-clusters the incorrectly clustered data. The validation and the re-clustering process is an iterative process [14]. It improves the quality of cluster results.

The remaining part of this research paper is as follows: section 2 describes the research background including different clustering methods for numerical and categorical data and also explains clustering algorithms used in research. Then, the proposed methodology is explained in section 3. The performance of the proposed work is analyzed in section 4-the conclusion and the future work of this research work are provided in section 5.

2. RESEARCH BACKGROUND

2.1. Mixed data clustering

Clustering mixed data is a difficult process that is rarely accomplished using well-known clustering algorithms developed for a certain type of data. It is common knowledge that converting one type to another is insufficient since it may result in data loss [15]. For clustering mixed datasets, Que *et al*. [16] suggest a similarity measurement using entropy-based weighting. An automatic categorization technique is used to convert numerical data into category data. The relevance of various attributes is then denoted using an entropy-based weighting technique.

Li *et al*. [17] offer a mixed data clustering technique with a noise-filtered distribution centroid and an iterative weight modification strategy. It defines a noise-filtered distribution centroid for categorical attributes. By integrating the mean and noise-filtered distribution centroid, this method displays the cluster centre with mixed properties. The frequency of occurrences for each potential value of the categorical attributes in a cluster is more accurately recorded by the noise-filtered distribution centroid.

Jia and Cheung [18] show how to cluster data using soft subspace clustering with both numerical and categorical features. The model is based on the definition of object-cluster similarity and is attribute-weighted. Using a uniform weighting approach for numerical and categorical qualities, the attribute-to-cluster contribution is measured by accounting for both inter-cluster difference and intra-cluster similarity.

For data with heterogeneous features, D’Urso and Massari [19] suggest a fuzzy clustering model. Different sorts of variables, or qualities, can be considered using the clustering model. This result is obtained by using a weighting system to combine the dissimilarity measurements for each attribute, yielding a distance measure for several attributes. During the optimization phase, the weights are computed objectively. The weights in the clustering findings represent the importance of each attribute type. Rodriguez *et al*. [20] suggest a multipartition clustering process that combines Bayesian network factorization and the variational Bayes framework to efficiently handle mixed data.

2.2. K-means clustering algorithm

Let X=x 1,x 2,...,x n be a data collection in a d-dimensional Euclidean space Rd, and A=a 1,a 2,...,a c be the c cluster centres, with d ik=x i-a k as its euclidean norm. Let U= ik _(nc), where _ik is a binary variable (i.e., _ik 0,1) that indicates whether the data point xi belongs to the kth cluster, k=1,2,...,c. By minimizing the k-means objective function, the k-means clustering method is iterated via the updating equations for cluster centres and memberships [12]: $\mathrm{J(U, A)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \sum_{\mathrm{k=1}}^{\mathrm{c}} \mathrm{\mu_{ik}} \|\mathrm{x_i} - \mathrm{a_k}\|^2$ as $\mathrm{a_k} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{\mu_{ik}}\, \mathrm{x_{ij}} / \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{\mu_{ik}}$ and $\mathrm{\mu_{ik}} = \begin{cases} 1 \text{ if } \|\mathrm{x_i} - \mathrm{a_k}\|^2 - \min\limits_{1 \le \mathrm{k} \le \mathrm{c}} \|\mathrm{x_i} - \mathrm{a_k}\|^2 \\ \qquad 0 \text{ Otherwise} \end{cases}$

2.3. Hierarchical clustering

In Algorithm 1 describe the hierarchical clustering pseudocode. Methods that use hierarchical clustering build a hierarchy of clusters that are arranged from top to bottom (or bottom to up). The hierarchical algorithms require both of the following to build clusters:

- Similarity matrix–this is created by determining how similar each pair of mixed data values are. The shape of the clusters is influenced by the similarity measure used to generate the similarity matrix.
- Linkage criterion–this establishes the distance between sets of observations as a function of pairwise distances.

Algorithm 1. Hierarchical clustering pseudocode

```
C = {Ci – {xi} | xi ϵ D}
Δ = {δ(xi, xj); xi, xj ϵ D}
Repeat
 Find the closest pair of clusters Ci, Cj ϵ C
 Cij = Ci U Cj
 C = C \ {{ Ci} U {Cj}} U {Cij}
 Update distance matrix Δ to reflect new clustering
Until |C| = k
```

2.4. Expectation maximization

The EM algorithm in Algorithm 2 finds maximum likelihood parameter estimates in probabilistic models. The iterative technique of expectation maximisation (EM) alternates between two steps: expectation (E) and maximum (M). To cluster data, EM employs the finite Gaussian mixtures model, which iteratively estimates a set of parameters until the desired convergence value is obtained. Each of the K probability distributions in the mixture corresponds to a single cluster. A membership probability is assigned to each instance by each cluster [21].

Algorithm 2. EM clustering pseudocode

```
1. Initialize estimates for $\theta \coloneqq \pi, \mu_1, \sigma_{1,}\mu_2, \sigma_2$
2. (Expectation) Compute the responsibilities for each data point
                    $\gamma_i = \frac{\pi\varphi(x_i; \mu_{2,}\sigma_2)}{(1-\pi)\varphi(x_i; \mu_1, \sigma_1) + \pi\varphi(x_i; \mu_2, \sigma_2)}$
3. (Maximization) Update the estimates for the parameters using the maximum likelihood
estimator formula. All sums are taken across the data indexed by i and are just
means/standard deviations weighted by the responsibilities $\gamma$
                    $\mu_2 = \frac{\sum \gamma_i x_i}{\sum \gamma_i} \sigma_2 = \frac{\sum \gamma_i (x_i - \mu_2)^2}{\sum \gamma_i} \pi = \frac{1}{n} \sum \gamma_i$
4. Repeat steps 2 and 3 until the parameters converge to a local optimum.
```

2.5. Self organization map

The SOM algorithm in Algorithm 3 is a classic unsupervised learning neural network model that clusters input data with similarities. It employs an unsupervised learning methodology and used a competitive learning algorithm to train its network. In order to minimise complex issues for straightforward interpretation, SOM is utilised for clustering and mapping (or dimensionality reduction) procedures to map multidimensional data onto lower-dimensional spaces. The input layer and the output layer are the two layers that make up SOM. The SOM merges the clustering and projection operations (reduce the dimensionality of information).

Algorithm 3. Self organization map

```
1. Initialize the weight wj, neighborhood parameter Np, k = 0, and learning rate μ=1.0;
2. Select random vector x from input data
3. Compute and select the winning neuron i.e Best Matching Unit based on a distance measure
and neighborhood function. The empirical index of the winning neurons is determined as
follows:
                    $\mathrm{i(x)} = \arg\min\limits_{1 \le \mathrm{j} \le \mathrm{d}} \|\mathrm{x} - \mathrm{w_j}\|$
4. Update the weight vector of winning neurons
          $\mathrm{w_i(k+1)} = \begin{cases} \mathrm{w_i(k)} + \mu[\mathrm{x(k)} - \mathrm{w_i(k)}]\ \mathrm{i} \in \mathrm{N_p(k)} \\ \qquad \mathrm{w_i(k)}\ \mathrm{i} \notin \mathrm{N_p(k)} \end{cases}$
5. Update the parameters
6. Repeat Steps 2, 3, and 4 until the stopping criteria are met.
```

3. PROPOSED METHOD

This section explains the proposed clustering algorithm for mixed numerical and categorical data [22] based on re-clustering and cluster validation called reclust. The proposed method contains three important processes: initial clustering, validation, and re-clustering. The initial clustering process uses four traditional clustering algorithms such as EM, HC, KM, and SOM. The validation process evaluates the clustering result. The re-clustering process re-clusters the incorrectly clustered data. The validation and the re-clustering process is an iterative process. It improves the quality of cluster results.

Let D be the mixed dataset consisting of n instances, indicates as {d1, d2, …, dn}. The dataset D has ac categorical attributes and au numerical attributes. Then $\mathrm{d_i}(1 \leq \mathrm{i} \ \leq \mathrm{n})$ can be denoted as $[\mathrm{d_i^c}, \mathrm{d_i^u}]$ with $\mathrm{d_i^c} = \left[\mathrm{d_{i1}^c}, \mathrm{d_{i2}^c}, \dots, \mathrm{d_{i,a_c}^c}\right]$ and $\mathrm{d_i^u} = \left[\mathrm{d_{i1}^u}, \mathrm{d_{i2}^u}, \dots, \mathrm{d_{i,a_u}^u}\right]$. Cluster the dataset D into k clusters C= {C1, C2, …., Ck}. $\mathrm{C_i} \cap \mathrm{C_j} = \emptyset$, $\bigcup_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{C_i} = \mathrm{C}(\mathrm{i}, \mathrm{j} = 1,2, \dots, \mathrm{k}, \mathrm{i} \neq \mathrm{j})$. Algorithm 4 explains the reclust clustering algorithm.

Algorithm 4. Reclust

```
Input: Dataset D = { d1, d2, …, dn}, Number of Cluster k
Output: Clustering Result
1. Initial Clustering
 1a. EMcls = Apply EM(D, k)
 1b. HCcls = Apply HC(D, k)
 1c. KMcls = Apply KM(D, k)
 1d. SOMcls = Apply SOM(D, k)
2. Cluster Validation
 2a. EMeval = evaluateCluster(EMcls, D)
 2b. HCeval = evaluateCluster(HCcls, D)
 2c. KMeval = evaluateCluster(KMcls, D)
 2d. SOMcls = evaluateCluster(SOMcls, D)
 2e. minCls = Min (EMeval, HCeval, KMeval, SOMeval)
 2f. incorrectD = incorrectlyClusteredData(D)
3. Reclustering
 3a. While (the stop criterion is not met)
 3b. recls= Cluster incorrectD using minCls
 3c. reclsEval = evaluateCluster (recls)
 3d. subD = incorrectlyClusteredData(incorrectD)
 3e. incorrect =subD
 3f. End While
```

In this algorithm, step 1 applies four traditional clustering algorithms. Step 2 evaluates the cluster results. The evaluateCluster uses classes to cluster evaluation method. It builds clustering after ignoring the class attribute. It then allocates classes to the clusters during the test phase, depending on the majority value of the class feature within each cluster. The classification error is then calculated based on this assignment. Step 2e finds the minimum error value of four traditional clustering algorithms. Step 2f extracts the incorrectly clustered data from the evaluation results. Step 3 is an iterative re-clustering, which clusters the incorrect data and evaluates the clustering result. The stop criterion for the re-clustering step is either a minimum error value or a minimum number of instances in incorrect clustered data.

4. EXPERIMENTAL RESULT

This section evaluates the performance of the proposed work through experiments. Three publicly available data sets and students' data with seven questionnaires are used to analyze the cluster results. Table 1 shows the summary of the dataset used for experiments. The following metrics are used to evaluate the clustering results: rand index (RI), precision (Pre), and recall (Rec). These evaluation metrics are computed using the classes to cluster assignment (CCA) table shown in Table 2.

Let D={D1, D2, D3,…, Dn} be the dataset contains n number of instances, C={C1, C2, …,Ck} denotes set of k clusters generated from D using clustering algorithm and P = {P1, P2, …, Pc}denotes set of c true classes of D. In table 2, aij represents the number of common instances between Pi and Cj i.e aij = |Pi ∩ Cj|. SPi and SCj denote the number of instances in Pi and Cj.

The evaluation metrics are computed as shown in:

$$Purity = \frac{1}{n}\sum_k \max_c |a_{kc}|$$

$$ARI = \frac{\sum_{ij}\binom{a_{ij}}{2} - \left[\sum_i \binom{SP_i}{2} \sum_j \binom{SC_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{SP_i}{2} + \sum_j \binom{SC_j}{2}\right] - \left[\sum_i \binom{SP_i}{2} \sum_j \binom{SC_j}{2}\right] / \binom{n}{2}}$$

here $\binom{n}{2} = n(n-1)/2$,

$$NMI = \frac{\sum_{i=1}^{c}\sum_{j=1}^{k} a_{ij}\log\left(\frac{a_{ij}{*}n}{SP_i{*}SC_j}\right)}{\sqrt{\sum_{i=1}^{c} SP_i\log\left(\frac{SP_i}{n}\right)\sum_{j=1}^{k} SC_j log\left(\frac{SC_j}{n}\right)}}$$

$$Pre = \frac{1}{c}\sum_{i=1}^{c}\frac{\max_k a_{ki}}{SP_i}\mathrm{S}$$

$$Rec = \frac{1}{k}\sum_{i=1}^{k}\frac{\max_c a_{ci}}{SC_i}$$

In this experiment, the number of clusters to be found was equal to the number of classes in the data set i.e., c = k. Larger values of RI, Pre, and Rec indicate better clustering results. Table 3 shows the Classes for Cluster Assignment for the emotional intelligence dataset. Most of the classes are correctly clustered.

Tables 4-9 shows CCA for EPQ, GSE, EHQ, PNA, RSE [23], SDS datasets. Table 10 shows the comparison of evaluation metrics for different datasets. The metrics RI, Precision, and Recall is compared with ABC-K-Prototypes [24], CCS-K-Prototypes [1], and Multi-view K-Prototype [25]. Table 11 and Figure 1 shows the Rand Index comparison. Table 12 and Figure 2 depict the precision comparison. Table 13 and Figure 3 depicts the recall comparison.

Table 1. Dataset summary

Dataset Type	Dataset	# Instances	# Numerical Features	# Categorical Features	# Classes
Student Info with Question Response	Emotional Intelligence (EIQ)	1000	2	11	3
	Eysenck Personality (EPQ)	1000	2	11	3
	General Self Efficacy (GSE)	1000	2	11	2
	Emotional Happiness (EHQ)	1000	2	11	3
	Positive /Negative Attitude (PNA)	1000	2	11	3
	Self Esteem(RSE)	1000	2	11	3
	Self Determination (SDS)	1000	2	11	2
Medical	Heart	293	7	6	5
	Dermatology	358	1	33	6
Credit Card	Credit	653	6	9	2

Table 2. Classes to cluster assignment table

	C_1	C_2	….	C_k	Sum
P_1	a_{11}	a_{12}	….	a_{1k}	SP_1
P_2	a_{21}	a_{22}	….	a_{2k}	SP_2
⋮	⋮	⋮	⋮	⋮	⋮
P_c	a_{c1}	a_{c2}	….	a_{ck}	SP_c
Sum	SC_1	SC_2	….	SC_k	

Table 3. CCA for emotional intelligence dataset

CCA	Assigned Cluster			
		High	Average	Low
Actual Classes	High	700	0	0
	Average	15	265	10
	Low	8	0	2

Table 4. CCA for eysenck personality dataset

CCA	Assigned Cluster			
		Extroversion	Psychoticism	Neuroticism
Actual Classes	Extroversion	665	0	0
	Psychoticsm	7	25	3
	Neuroticism	20	0	280

Table 5. CCA for self efficacy

CCA	Assigned Cluster		
		High	Low
Actual Classes	High	960	10
	Low	4	26

Table 6. CCA for emotional happiness

CCA		Assigned Cluster		
		Happy	Moderately_happy	Unhappy
Actual Classes	Happy	238	12	0
	Moderately_happy	39	692	0
	Unhappy	0	4	15

Table 7. CCA for positive/negative attitude

Table 8. CCA for self esteem

CCA		Assigned Cluster		
		Positive	Negative	Neutral
Actual Classes	Positive	756	35	14
	Negative	12	113	0
	Neutral	0	4	66

CCA		Assigned Cluster		
		High	Normal	Low
Actual Classes	High	890	0	16
	Normal	6	52	0
	Low	6	0	30

Table 9. CCA for self determination

CCA		Assigned Cluster	
		High	Low
Actual Classes	High	745	15
	Low	20	220

Table 10. Evaluation metrics comparison

Data Set	EIQ	EPQ	GSE	EHQ	PNA	RSE	SDS	Heart	Dermatology	Credit Card
Purity	0.975	0.97	0.986	0.945	0.935	0.972	0.965	0.724	0.911	0.928
RI	0.908	0.901	0.821	0.798	0.774	0.838	0.859	0.899	0.865	0.947
NMI	0.823	0.83	0.606	0.689	0.645	0.715	0.735	0.817	0.886	0.9
Pre	0.905	0.883	0.928	0.896	0.929	0.904	0.948	0.777	0.919	0.983
Rec	0.934	0.983	0.859	0.946	0.851	0.88	0.955	0.684	0.936	0.986

Table 11. RI comparison

Dataset	ABC-K	CCS-K	Multi-View	Reclust
Heart	0.667	0.680	0.684	0.899
Dermatology	0.689	0.694	0.691	0.865
Credit Card	0.673	0.674	0.695	0.947

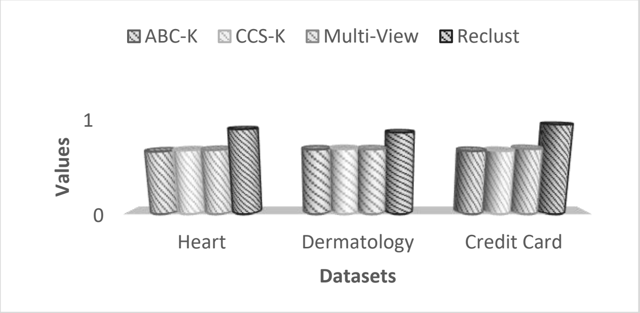

Figure 1. Rand index comparison

Table 12. Precision comparison

Dataset	ABC-K	CCS-K	Multi-View	Reclust
Heart	0.658	0.675	0.637	0.777
Dermatology	0.808	0.812	0.809	0.919
Credit Card	0.792	0.814	0.810	0.983

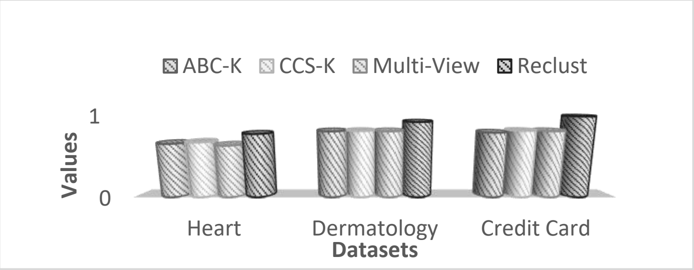

Figure 2. Precision comparison

Table 13. Recall comparison

Dataset	ABC-K	CCS-K	Multi-View	Reclust
Heart	0.379	0.388	0.398	0.684
Dermatology	0.806	0.809	0.807	0.936
Credit Card	0.795	0.796	0.810	0.986

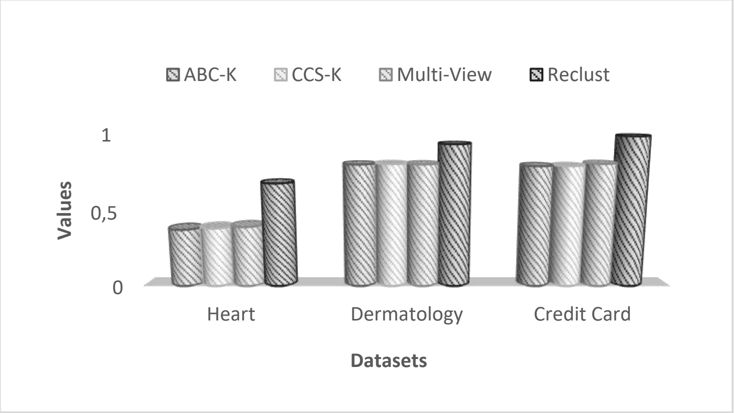

Figure 3. Recall comparison

5. CONCLUSION

Clustering is a typical data mining technique, and clustering mixed datasets into meaningful groups is possible since mixed items are ubiquitous in real-world datasets. This research presents an effective clustering approach for grouping mixed numerical and categorical datasets. Furthermore, iterative re-clustering and cluster validation enhance the clustering results. In terms of clustering purity, NMI, rand index, precision, and recall, the suggested reclust algorithm was tested on several datasets. The results of the experiments confirm the reclust algorithm's superior performance.

REFERENCES

[1] J. Ji, W. Pang, Z. Li, F. He, G. Feng, and X. Zhao, "Clustering mixed numeric and categorical data with cuckoo search," *IEEE Access*, vol. 8, pp. 30988–31003, 2020, doi: 10.1109/ACCESS.2020.2973216.

[2] L. K. Ramasamy, S. Kadry, Y. Nam, and M. N. Meqdad, "Performance analysis of sentiments in Twitter dataset using SVM models," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2275–2284, 2021, doi: 10.11591/ijece.v11i3.pp2275-2284.

[3] Amala Jayanthi M. and E. Shanthi I., "Role of Educational data mining in student learning processes with sentiment analysis," *International Journal of Knowledge and Systems Science*, vol. 11, no. 4, pp. 31–44, Oct. 2020, doi: 10.4018/IJKSS.2020100103.

[4] J. Han, J. Pei, and H. Tong, *Data Mining*, 3rd ed. Elsevier, 2012.

[5] M. A. Jayanthi, R. L. Kumar, A. Surendran, and K. Prathap, "Research contemplate on educational data mining," in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Oct. 2016, pp. 110–114, doi: 10.1109/ICACA.2016.7887933.

[6] D.-T. Dinh, V.-N. Huynh, and S. Sriboonchitta, "Clustering mixed numerical and categorical data with missing values," *Information Sciences*, vol. 571, pp. 418–442, Sep. 2021, doi: 10.1016/j.ins.2021.04.076.

[7] L. K. Ramasamy, S. Kadry, and S. Lim, "Selection of optimal hyper-parameter values of support vector machine for sentiment analysis tasks using nature-inspired optimization methods," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 290–298, 2021, doi: 10.11591/eei.v10i1.2098.

[8] Ahmad Amir and Khan Shehroz, "Survey of state-of-the-art mixed data clustering algorithms," *IEEE Access*, vol. 8, pp. 318883–31902, 2020.

[9] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD)*, 1997, pp. 21–34, [Online]. Available: http://reference.kfupm.edu.sa/content/c/l/clustering_large_data_sets_with_mixed_nu_362883.pdf.

[10] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, pp. 590–596, Nov. 2013, doi: 10.1016/j.neucom.2013.04.011.

[11] D. Lam, M. Wei, and D. Wunsch, "Clustering data of mixed categorical and numerical type with unsupervised feature learning," *IEEE Access*, vol. 3, pp. 1605–1613, 2015, doi: 10.1109/ACCESS.2015.2477216.

[12] A. Foss, M. Markatou, B. Ray, and A. Heching, "A semiparametric method for clustering mixed data," *Machine Learning*, vol. 105, no. 3, pp. 419–458, 2016, doi: 10.1007/s10994-016-5575-7.

[13] J. Y. Chen and H. H. He, "A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data," *Information Sciences*, vol. 345, pp. 271–293, 2016, doi: 10.1016/j.ins.2016.01.071.

[14] A. Murugesan, B. Saminathan, F. Al-Turjman, and R. L. Kumar, "Analysis on homomorphic technique for data security in fog computing," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 9, Sep. 2021, doi: 10.1002/ett.3990.

[15] S. Behzadi, N. S. Müller, C. Plant, and C. Böhm, "Clustering of mixed-type data considering concept hierarchies: problem specification and algorithm," *International Journal of Data Science and Analytics*, vol. 10, no. 3, pp. 233–248, Sep. 2020, doi: 10.1007/s41060-020-00216-2.

[16] X. Que, S. Jiang, J. Yang, and N. An, “A Similarity measurement with entropy-based weighting for clustering mixed numerical and categorical datasets,” *Algorithms*, vol. 14, no. 6, p. 184, Jun. 2021, doi: 10.3390/a14060184.

[17] X. Li, Z. Wu, Z. Zhao, F. Ding, and D. He, “A mixed data clustering algorithm with noise-filtered distribution centroid and iterative weight adjustment strategy,” *Information Sciences*, vol. 577, pp. 697–721, Oct. 2021, doi: 10.1016/j.ins.2021.07.039.

[18] H. Jia and Y. M. Cheung, “Subspace clustering of categorical and numerical data with an unknown number of clusters,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3308–3325, Aug. 2018, doi: 10.1109/TNNLS.2017.2728138.

[19] P. D’Urso and R. Massari, “Fuzzy clustering of mixed data,” *Information Sciences*, vol. 505, pp. 513–534, 2019, doi: 10.1016/j.ins.2019.07.100.

[20] F. Rodriguez-Sanchez, C. Bielza, and P. Larrañaga, “Multipartition clustering of mixed data with Bayesian networks,” *International Journal of Intelligent Systems*, vol. 37, no. 3, pp. 2188–2218, Mar. 2022, doi: 10.1002/int.22770.

[21] K. P. Sinaga and M. S. Yang, “Unsupervised K-means clustering algorithm,” *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[22] X. Jin and J. Han, “Expectation maximization clustering,” in *Encyclopedia of Machine Learning and Data Mining*, Boston, MA: Springer US, 2016, pp. 1–2.

[23] M. Amala Jayanthi, S. Swathi, and R. Lakshmana Kumar, “Investigation on association of self-esteem and students’ performance in academics,” *International Journal of Grid and Utility Computing*, vol. 9, no. 3, p. 211, 2018, doi: 10.1504/IJGUC.2018.10015144.

[24] J. Ji, Y. Chen, G. Feng, X. Zhao, and F. He, “Clustering mixed numeric and categorical data with artificial bee colony strategy,” *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 2, pp. 1521–1530, 2019, doi: 10.3233/JIFS-18146.

[25] J. Ji, R. Li, W. Pang, F. He, G. Feng, and X. Zhao, “A multi-view clustering algorithm for mixed numeric and categorical data,” *IEEE Access*, vol. 9, pp. 24913–24924, 2021, doi: 10.1109/ACCESS.2021.3057113.

BIOGRAPHIES OF AUTHORS

Amala Jayanthi Maria Soosai Arockiam iD SC is an assistant professor at Kumaraguru College of Technology in Coimbatore, where she works in the Department of Computer Applications. She has 9 years of academic experience. She earned her master's degree in computer science from Bharathiyar University's St. Joseph's College (Autonomous) in Tiruchirappalli, Tamil Nadu, India. Kalasalingam University awarded her aMCA degree. Data mining is her current study focus. She has 16 research papers published in reputable publications. She can be contacted at email: research.amala@gmail.com.

Dr. Elizabeth Shanthi Irudhayaraj iD SC is an Associate Professor of Computer Science at Coimbatore's Avinashilingam University for Women. She has more than 25 years of teaching experience, as well as ten years of research experience. She has a long list of publications to her name. Data mining, information retrieval, objectoriented data bases, cloud computing, and soft computing are some of her areas of interest. She can be contacted at email: elizabeth_cs@avinuty.ac.in.