Hindawi Publishing Corporation
Scientifica
Volume 2016, Article ID 4273813, 6 pages
http://dx.doi.org/10.1155/2016/4273813

Research Article

Evaluation of Modified Categorical Data Fuzzy Clustering Algorithm on the Wisconsin Breast Cancer Dataset

Amir Ahmad

Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, P.O. Box 344, Rabigh 21911, Saudi Arabia

Correspondence should be addressed to Amir Ahmad; amirahmad01@gmail.com

Received 9 December 2015; Revised 31 January 2016; Accepted 1 February 2016

Academic Editor: Dick de Ridder

The early diagnosis of breast cancer is an important step in a fight against the disease. Machine learning techniques have shown promise in improving our understanding of the disease. As medical datasets consist of data points which cannot be precisely assigned to a class, fuzzy methods have been useful for studying of these datasets. Sometimes breast cancer datasets are described by categorical features. Many fuzzy clustering algorithms have been developed for categorical datasets. However, in most of these methods Hamming distance is used to define the distance between the two categorical feature values. In this paper, we use a probabilistic distance measure for the distance computation among a pair of categorical feature values. Experiments demonstrate that the distance measure performs better than Hamming distance for Wisconsin breast cancer data.

1. Introduction

Breast cancer is the most common form of cancer amongst women [1]. Early and accurate detection of breast cancer is the key to the long survival of patients [1]. Machine learning techniques are being used to improve diagnostic capability for breast cancer [2–4]. Wisconsin breast cancer dataset has been a popular dataset in machine learning community [5]. Various classification techniques such as techniques like decision trees [6], support vector machines [7], and fuzzy-genetic algorithm [8] have been used to study this dataset. In medical datasets, sometimes it is difficult to put some data points in one of the groups. Fuzzy methods are better equipped to handle these kinds of datasets [9–11].

Clustering divides the data points into different groups (clusters) depending upon a similarity measure [12]. The data points in a group (cluster) are similar whereas data points in different groups (clusters) are dissimilar. Clustering algorithms can be divided into two groups [12, 13]: hard clustering algorithms and fuzzy clustering algorithms. In hard clustering, a data point can have a membership to a cluster. However, in fuzzy clustering, a data point has memberships to all the clusters.

K-means algorithm [14] is very popular hard clustering algorithm because of its linear complexity. K-means clustering algorithm is an iterative algorithm which computes the mean of each feature of data points presented in a cluster. This makes the algorithm inappropriate for the datasets that have categorical features. Huang [15] extends the K-mean algorithm for the datasets having categorical features. Instead of mean, mode is used to represent a cluster. Hamming distance is used to calculate the membership of a data point. In Hamming distance if the feature values are same for two data points the distance is taken as 0; otherwise the distance is taken as 1.

Hierarchical clustering algorithms [12] can be applied for categorical datasets; however they have high computation complexity. This makes them less useful for large datasets.

Fuzzy clustering has shown great promise in understanding medical datasets [10, 11]. It has been shown that the fuzzy clustering can be used to improve the classification performance of various classifiers for diagnosis of breast cancer [16]. Fuzzy c-mean (FCM) [17, 18] is one of the most popular clustering techniques. Original FCM clustering technique can only handle numeric features. Using the methodology of FCM, fuzzy K-mode algorithm [19] is

proposed for categorical datasets. This method use Hamming distance and hard cluster centres. Kim et al. [20] propose a fuzzy clustering algorithm that uses fuzzy cluster centres. This algorithm performs better than fuzzy K-mode algorithm [20].

Most of fuzzy clustering algorithms for categorical datasets use Hamming distance. However, Lee and Pedrycz [21] show that the simple matching similarity like Hamming distance cannot capture the correct similarities among categorical feature values; hence an appropriate distance measure should be used to improve the performance of fuzzy clustering algorithm with fuzzy cluster centres.

Various dissimilarity measures have been proposed for categorical feature values [23]. Ahmad and Dey [22] present a dissimilarity measure for categorical features. Ahmad and Dey [22] show that K-mode clustering algorithm can be improved with this dissimilarity measure. Ahmad and Dey [24] use this distance measure to propose a clustering algorithm for datasets having numerical and categorical features. Ahmad and Dey [25] also suggest a subspace clustering algorithm with this dissimilarity measure. Motivated by the success of the dissimilarity measure for clustering categorical data, Ji et al. [26] use the distance measure for fuzzy clustering of mixed datasets. Ahmad and Dey [27] presented a fuzzy clustering method that uses a distance measure that calculates distances for each iteration.

Wisconsin breast cancer dataset has been studied extensively in machine learning field [25, 30–32]. Each feature of Wisconsin breast dataset has ten categories (1 to 10). It has been a popular dataset for analysing clustering algorithms for categorical datasets [25, 30–32]. In this paper, we show the application of the clustering algorithm proposed by Ji et al. [26] for Wisconsin breast cancer dataset. This way we will show the applicability of the distance measure proposed by Ahmad and Dey [22] for the analysis of categorical breast cancer dataset.

This paper has the following organization. We will discuss fuzzy c-mean clustering algorithm in Section 2. Section 3 reviews the method that computes the distance between two categorical feature values. Section 4 discusses the method to compute the fuzzy centroid for categorical datasets and the distance between a data point and a cluster centre [26]. Experimental results are presented in Section 5. Section 6 has conclusion and future work.

2. Fuzzy c-Mean Clustering Algorithm

Fuzzy c-mean (FCM) [17, 18] is a popular clustering algorithm. In this section, we will discuss FCM.

The following information is given:

(i) A dataset is with n data points.

(ii) Each data point is defined by s features.

(iii) The desired number of clusters is K.

(iv) Fuzzy membership matrix $U = (u_{ij})_{n\times K}$.

FCM clusters a set of n data points, $X = \{X_1, X_2, \ldots, X_n\}$, into K clusters, where $X_i = \{x_{i,1}, \ldots, x_{i,s}\}$ is the ith data point for $i = 1, \ldots, n$.

FCM compute the cluster centres v_j $(j = 1, 2, \ldots, K)$, where $v_j = \{v_{j1}, v_{j2}, \ldots, v_{js}\}$, and the fuzzy membership matrix U. It is done by minimizing an objective function, J, presented below iteratively:

$$J = \sum_{j=1}^{K}\sum_{i=1}^{n}\left(u_{ij}\right)^m d_{ij} \tag{1}$$

(m is used as defined real number which controls the fuzziness)

$$\text{subject to } \sum_{j=1}^{K} u_{ij} = 1, \quad i = 1, 2, \ldots, n, \tag{2}$$

where d_{ij} is the distance between data point X_i and cluster centre v_j.

For numeric data, v_j and u_{ij} are computed as follows:

$$v_{jp} = \frac{\sum_{i=1}^{n}\left(u_{ij}\right)^m X_{ip}}{\sum_{i=1}^{n}\left(u_{ij}\right)^m},$$
$$u_{ij} = \frac{1}{\left(\sum_{k=1}^{K}\left(d_{ij}/d_{kj}\right)^{1/(m-1)}\right)}. \tag{3}$$

The steps for FCM based algorithm are presented as follows.

Step 1. Select a stopping value ε. Initialize the fuzzy membership matrix U. It is done by creating $n \times K$ random numbers; these numbers are in the interval $[0, 1]$.

do

Step 2. Compute cluster centres.

Step 3. Compute distances from centres and use these distances for updating fuzzy membership matrix U.

Step 4. Calculate the objective function J.

While (the difference between two subsequent computed values of J is more than the given stopping value ε).

3. The Distance between Two Categorical Feature Values

Ahmad and Dey [22] propose an algorithm to calculate the distance between two categorical feature values in an unsupervised framework. Unlike Hamming distance, this distance measure does not take binary measure for the distance between two categorical values. The distance is calculated by computing the cooccurrence of the feature values (for which the distance is calculated) with feature values of other features.

The distance between categorical feature values x and y of feature A_i against the feature A_j, for a subset w of feature A_j values, is defined as follows:

$$\delta_w^{ij}(x, y) = p\left(\frac{w}{x}\right) + p\left(\sim \frac{w}{y}\right) - 1. \tag{4}$$

The distance $\delta^{ij}(x, y)$ between the feature values x and y for A_i against feature A_j is presented by $\delta^{ij}(x, y)$ and is defined by $\delta^{ij}(x, y) = p(\omega/x) + p(\sim \omega/y) - 1$, where ω is the subset w of feature values of A_j that maximizes the quantity $(p(w/x) + p(\sim w/y))$. To compute the distance between x and y, we compute the distances between x and y against every other feature. The average distance is taken as the distance, $\delta(x, y)$, between x and y in the dataset. Distances between every pair of feature values are employed to calculate the distance between a data point and a cluster centre.

4. Modified Centre and the Distance from the Modified Centre

For categorical datasets, the mode is used to calculate the centre of clusters [19]. However, taking only one feature value to represent a cluster centre does not capture the cluster centre well; hence loss of information takes place. Ji et al. [26] use the fuzzy centroid [20] concept with distance measure suggested by Ahmad and Dey [22] for fuzzy clustering of categorical datasets.

The fuzzy centroid for a cluster, C, for a categorical dataset is defined as

$$\left(\frac{1}{N_c}\right)\left\langle\left(N_{1,1,c}, N_{1,2,c}, \ldots, N_{1,p1,c}\right), \ldots, \left(N_{l,1,c}, N_{l,2,c}, \ldots, N_{l,k,c}, \ldots, N_{l,pl,c}\right), \ldots, \left(N_{s,1,c}, N_{s,2,c}, \ldots, N_{s,ps,c}\right)\right\rangle. \quad (5)$$

Assume that lth feature has pl different values.

Thus,

$$N_c = \sum_{i=1}^{n} \left(u_{ij}\right)^m \quad (\text{where } j\text{th cluster is } C), \quad (6)$$

where $N_{l,k,c}$ is the association of value $A_{l,k}$ (kth feature value for the lth feature) with cluster C:

$$N_{l,k,c} = \sum_{i=1}^{n} L\left(X_{il} = A_{l,k}\right)\left(u_{ij}\right)^m \quad (j\text{th cluster is } C), \quad (7)$$

where $L(X_{il} = A_{l,k})$

= 1 for a data point X_i having lth feature value = $A_{l,k}$,

= 0 for a data point X_i having lth feature value $\neq A_{l,k}$.

The distance between a data point having lth categorical feature value Z in the lth dimension and the centre of cluster C is defined as

$$\Omega(Z, C) = \left(\frac{N_{l,1,c}}{N_c}\right) * \delta\left(Z, A_{l,1}\right) + \left(\frac{N_{l,2,c}}{N_c}\right) * \delta\left(Z, A_{l,2}\right) + \cdots + \left(\frac{N_{l,t,c}}{N_c}\right) * \delta\left(Z, A_{l,t}\right) + \cdots + \left(\frac{N_{l,pl,c}}{N_c}\right) * \delta\left(Z, A_{l,pl}\right), \quad (8)$$

where $A_{l,t}$ is the tth feature value of lth categorical feature.

$\delta(x, y)$ is calculated by the method discussed in Section 3. For dataset having s features, the distance is calculated for each feature value of the data point and the summation of these distances is the distance between the data point and the centre. In FCM, the distances between data points and cluster centres are used to calculate fuzzy membership matrix. Hence, this distance measure will be employed to compute the fuzzy membership matrix.

The cluster centre definition and distances between cluster centre and data points discussed in this section can be used with FCM algorithm discussed in Section 2 to create fuzzy clustering algorithm for categorical datasets [26]. The steps of fuzzy clustering algorithm for categorical data are as follows.

Step 1. Select a stopping value ε. Initialize the fuzzy membership matrix U. It is done by creating $n \times K$ random numbers; these numbers are in the interval $[0, 1]$.

do

Step 2. Compute cluster centres by using (5).

Step 3. Compute distances from centres by using (8). Hamming distance/distances discussed in Section 3 will be used in this step. Use these distances for updating fuzzy membership matrix U.

Step 4. Calculate the objective function J.

While (the difference between two subsequent computed values of J is more than the given stopping value ε).

5. Results and Discussion

The experiments were carried out on Wisconsin breast cancer data. This dataset has 699 data points. Each data point is represented by 9 features. 16 data points have missing values. Missing feature values were replaced by the mode of that feature. The information about these features is given in Table 1. These are two groups in this dataset: benign and malignant. Benign group has 458 data points whereas malignant group has 241 data points. Each feature has categories (0–10). We ran fuzzy clustering with fuzzy centroid with Hamming distance and the distance measure proposed by Ahmad and Dey [22] to see how the incorporation of the distance measure affects the quality of the clustering.

Clustering error = the number of data points not in desired clusters/the number of data points.

To assess the quality of clustering, it is assumed that a preclassified dataset is provided and the "overlap" between an achieved clustering and the ground truth classification is measured. Experiments were carried out at different values of m: 1.1, 1.5, and 1.9. The random initialization was used for both clustering algorithms. Clustering algorithms were run 100 times in each setting (different m values) and average results are presented in Table 2. We also presented the performance of various clustering algorithms on Wisconsin breast cancer

Table 1: Information about features.

Feature number	Feature
1	Clump thickness
2	Uniformity of cell size
3	Uniformity of cell shape
4	Marginal adhesion
5	Single epithelial cell size
6	Bare nuclei
7	Bland chromatin
8	Normal nucleoli
9	Mitoses

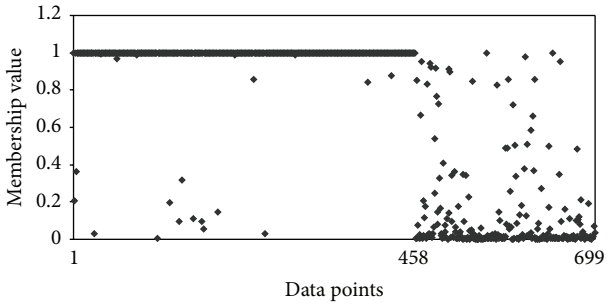

Figure 1: Membership for different data points for the benign cluster by using Ahmad and Dey [22] distance measure for $m = 1.1$. The first 458 data points are benign and next 241 data points are malignant.

dataset. We performed the experiments for fuzzy K-modes clustering algorithm. The average result of 10 runs with $m = 1.1$ is presented. Other clustering results are taken from [30]. Results are presented in Table 3. Confusion matrices for different setups are presented in Tables 4–9.

Clustering results suggest that for all values of m the fuzzy clustering algorithm with Ahmad and Dey [22] distance measure performed better; for example, for $m = 1.1$, the average clustering error for the proposed algorithm was 5.0%, whereas the average clustering error with Hamming distance was 10.4%. This shows that the application of the distance measure improved the clustering results. Table 3 suggests that the fuzzy clustering algorithm with Ahmad and Dey [22] distance measure performed better than other clustering algorithms.

The other interesting observation is that, with Hamming distance, the clustering algorithm was putting malignant data points in benign clusters. In other words, it had difficulty in assigning malignant data points correctly, whereas the clustering algorithm with Ahmad and Dey [22] distance measure had better assignment of malignant data points. To understand this point more, we compared the membership of different data points for these two algorithms. Figures 1 and 2 show the membership of different data points for benign cluster. It shows that with Hamming distance even the malignant data points have high memberships for benign cluster. However, with the distance measure proposed by Ahmad and Dey [22], we have better membership relationship. Figures 3 and 4 show the membership of different data points

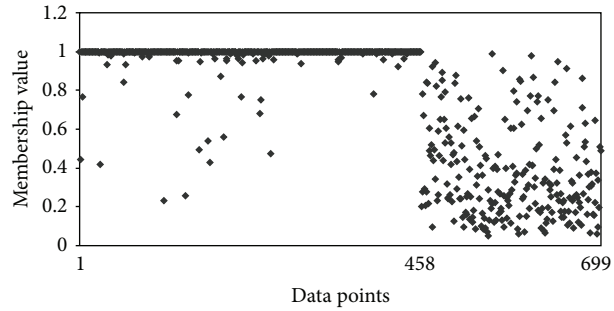

Figure 2: Membership for different data points for the benign cluster by using the Hamming distance for $m = 1.1$. The first 458 data points are benign and next 241 data points are malignant.

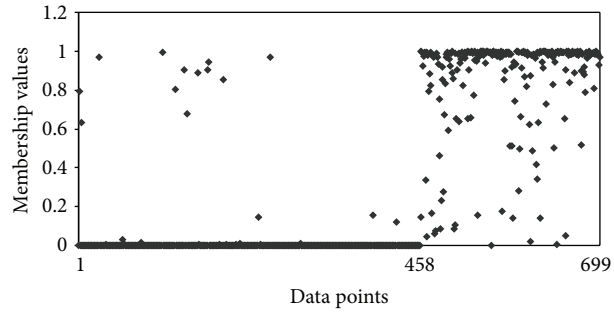

Figure 3: Membership for different data points for the malignant cluster by using Ahmad and Dey [22] distance measure for $m = 1.1$. The first 458 data points are benign and next 241 data points are malignant.

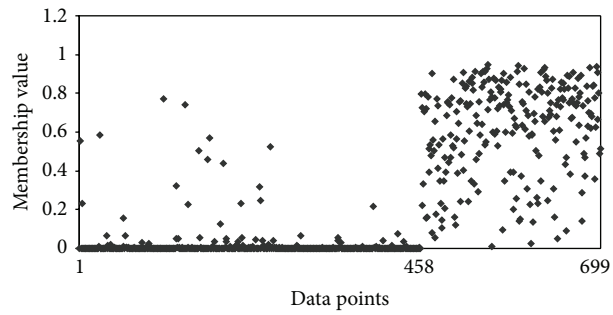

Figure 4: Membership for different data points for the malignant cluster by using the Hamming distance for $m = 1.1$. The first 458 data points are benign and next 241 data points are malignant.

for malignant cluster. This suggests that with Hamming distance membership values of benign data points are low; however, membership values of malignant data points are not as high as that with the proposed algorithm. These observations demonstrate that, with the distance measure proposed by Ahmad and Dey [22], better membership values were achieved.

Table 2: Average clustering results with standard deviation, in brackets, for Wisconsin breast cancer dataset. For all values of m, Ahmad and Dey [22] distance measure performed better.

The value of m (the measure of fuzziness)	Clustering error with Ahmad and Dey [22] distance measure algorithm in %	Clustering error with Hamming distance in %
1.1	5.0 (0.3)	10.4 (0.8)
1.5	4.7 (0.4)	10.3 (0.7)
1.9	5.0 (0.6)	9.7 (1.1)

Table 3: Comparative clustering results for Wisconsin breast cancer dataset.

Clustering algorithm	Clustering error in %
Fuzzy clustering with Ahmad and Dey [22] distance measure for $m = 1.1$	5.0
Fuzzy clustering with Hamming distance for $m = 1.1$	10.4
Fuzzy K-modes clustering for $m = 1.1$	13.9
Squeezer [28]	13.2
GAClust [29]	18.4
CcdByEnsemble [30]	15.0

Table 4: Confusion matrix for the average clustering result with Ahmad and Dey [22] distance measure, for $m = 1.1$.

	Cluster 1	Cluster 2
Benign	11	447
Malignant	217	24

Table 5: Confusion matrix for the average clustering result with Hamming distance, for $m = 1.1$.

	Cluster 1	Cluster 2
Benign	9	449
Malignant	177	64

Table 6: Confusion matrix for the average clustering result with Ahmad and Dey [22] distance measure, for $m = 1.5$.

	Cluster 1	Cluster 2
Benign	12	446
Malignant	220	21

Table 7: Confusion matrix for the average clustering result with Hamming distance, for $m = 1.5$.

	Cluster 1	Cluster 2
Benign	14	444
Malignant	183	58

Table 8: Confusion matrix for the average clustering result with Ahmad and Dey [22] distance measure, for $m = 1.9$.

	Cluster 1	Cluster 2
Benign	12	446
Malignant	218	23

Table 9: Confusion matrix for the average clustering result with Hamming distance, for $m = 1.9$.

	Cluster 1	Cluster 2
Benign	9	449
Malignant	182	59

6. Conclusion and Future Work

Early and correct detection is the key for the cure of breast cancer. Machine learning techniques are important diagnostic tools for breast cancer. Fuzzy clustering algorithms have shown great promise in analysis of breast cancer. Wisconsin breast cancer dataset has been treated as a categorical dataset in different studies because its features have categories (1–10). Ahmad and Dey [22] suggested a distance measure that has been successfully used in many clustering algorithms for categorical datasets. We used this distance measure for fuzzy clustering of Wisconsin breast cancer dataset. Our results suggest that we got better results as compared to the fuzzy clustering algorithm with Hamming distance. Experiment results also suggest that the membership values achieved by the distance measure proposed by Ahmad and Dey [22] better matched the given information. In future, we will apply this distance measure to other medical datasets. Various other fuzzy clustering algorithms for categorical datasets have been suggested [33–35]; in future, we will study the applicability of the distance measure proposed by Ahmad and Dey [22] for these algorithms. A comparative study of other distance measures will also be carried out [36]. The cluster centre initialization is a problem as different random initialization leads to different clustering results [37]. In future, we will apply different cluster centre initialization methods to overcome this problem.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

References

[1] *World Cancer Report*, International Agency for Research on Cancer, Lyon, France, 2008.

[2] M. Muthu Rama Krishnan, S. Banerjee, C. Chakraborty, C. Chakraborty, and A. K. Ray, "Statistical analysis of mammographic features and its classification using support vector machine," *Expert Systems with Applications*, vol. 37, no. 1, pp. 470–478, 2010.

[3] J. Ren, D. Wang, and J. Jiang, "Effective recognition of MCCs in mammograms using an improved neural classifier," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 4, pp. 638–645, 2011.

[4] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 38, no. 7, pp. 9014–9022, 2011.

[5] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original).

[6] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77–90, 1996.

[7] K. P. Bennet, J. Blue, and R. P. I. Math, "A support vector machine approach to decision trees," Report 97-100, Rensselaer Polytechnic Institute, Troy, NY, USA, 1997.

[8] C. A. Pena-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," *Artificial Intelligence in Medicine*, vol. 17, no. 2, pp. 131–155, 1999.

[9] J. Fuentes-Uriarte, M. Garcia, and O. Castillo, "Comparative study of fuzzy methods in breast cancer diagnosis," in *Proceedings of the IEEE Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS '08)*, pp. 1–5, New York, NY, USA, May 2008.

[10] P. R. Innocent, R. I. John, and J. M. Garibaldi, "Fuzzy methods for medical diagnosis," *Applied Artificial Intelligence*, vol. 19, no. 1, pp. 69–98, 2005.

[11] G. Berks, D. G. Keyserlingk, J. Jantzen, M. Dotoli, and H. Axer, "Fuzzy clustering—a versatile mean to explore medical database," in *Proceedings of the European Symposium on Intelligent Techniques (ESIT '00)*, Aachen, Germany, September 2000.

[12] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, USA, 1988.

[13] B. Everitt, L. Sabine, L. Morven, and M. Leese, *Cluster Analysis*, Hodder Arnold, London, UK, 2001.

[14] J. B. MacQuuen, "Some methods for classification and analysis of multivariate observation," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, University of California Press, Berkeley, Calif, USA, 1967.

[15] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[16] B. Verma, P. McLeod, and A. Klevansky, "Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3344–3351, 2010.

[17] J. C. Dunn, "Some recent investigations of a new fuzzy partitioning algorithm and its application to pattern classification problems," *Journal of Cybernetics*, vol. 4, no. 2, pp. 1–15, 1974.

[18] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, NY, USA, 1981.

[19] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 4, pp. 446–452, 1999.

[20] D.-W. Kim, K. H. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1263–1271, 2004.

[21] M. Lee and W. Pedrycz, "The fuzzy C-means algorithm with fuzzy P-mode prototypes for clustering objects having mixed features," *Fuzzy Sets and Systems*, vol. 160, no. 24, pp. 3590–3600, 2009.

[22] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," *Pattern Recognition Letters*, vol. 28, no. 1, pp. 110–118, 2007.

[23] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: a comparative evaluation," in *Proceedings of the 8th SIAM International Conference on Data Mining (SDM '08)*, pp. 243–254, Atlanta, Ga, USA, April 2008.

[24] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data and Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2007.

[25] A. Ahmad and L. Dey, "A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets," *Pattern Recognition Letters*, vol. 32, no. 7, pp. 1062–1069, 2011.

[26] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowledge-Based Systems*, vol. 30, pp. 129–135, 2012.

[27] A. Ahmad and L. Dey, "Algorithm for fuzzy clustering of mixed data with numeric and categorical attributes," in *Distributed Computing and Internet Technology*, vol. 3816 of *Lecture Notes in Computer Science*, pp. 561–572, Springer, Berlin, Germany, 2005.

[28] Z. He, X. Xu, and S. Deng, "Squeezer: an efficient algorithm for clustering categorical data," *Journal of Computer Science and Technology*, vol. 17, no. 5, pp. 611–624, 2002.

[29] D. Cristofor and D. Simovici, "Finding median partitions using information-theoretical-based genetic algorithms," *The Journal of Universal Computer Science*, vol. 8, no. 2, pp. 153–172, 2002.

[30] Z. He, X. Xu, and S. Deng, "A cluster ensemble method for clustering categorical data," *Information Fusion*, vol. 6, no. 2, pp. 143–151, 2005.

[31] Z. He, X. Xu, and S. Deng, "k-ANMI: a mutual information based clustering algorithm for categorical data," *Information Fusion*, vol. 9, no. 2, pp. 223–233, 2008.

[32] G. Gan and J. Wu, "Subspace clustering for high dimensional categorical data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 2, pp. 87–94, 2004.

[33] I. Saha and U. Maulik, "Incremental learning based multiobjective fuzzy clustering for categorical data," *Information Sciences*, vol. 267, pp. 35–57, 2014.

[34] I. Saha, J. P. Sarkar, and U. Maulik, "Ensemble based rough fuzzy clustering for categorical data," *Knowledge-Based Systems*, vol. 77, pp. 114–127, 2015.

[35] I.-K. Park and G.-S. Choi, "Rough set approach for clustering categorical data using information-theoretic dependency measure," *Information Systems*, vol. 48, pp. 289–295, 2015.

[36] T. R. L. dos Santos and L. E. Zárate, "Categorical data clustering: what similarity measure to recommend?" *Expert Systems with Applications*, vol. 42, no. 3, pp. 1247–1260, 2015.

[37] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for K-modes clustering," *Expert Systems with Applications*, vol. 40, no. 18, pp. 7444–7456, 2013.