Bull. Pure Appl. Sci. Sect. E Math. Stat.
38E(Special Issue)(2S), 71–78 (2019)
e-ISSN:2320-3226, Print ISSN:0970-6577
DOI 10.5958/2320-3226.2019.00082.1
©Dr. A.K. Sharma, BPAS PUBLICATIONS,
387-RPS-DDA Flat, Mansarover Park,
Shahdara, Delhi-110032, India. 2019

Bulletin of Pure and Applied Sciences
Section - E - Mathematics & Statistics
Website : https : //www.bpasjournals.com/

Analysis on pre-processing of heterogeneous dataset for ensemble clustering *

Darsana Prakash[1], S. Saranya[2] and R. Abitha[3]

1,2,3. Department of Information Technology, Women's Christian College,
Affiliated to University of Madras, Chennai-600008, Tamil Nadu, India.
1. E-mail: vihaasree123@gmail.com , 2. E-mail: s.saranya437@gmail.com
3. E-mail: abirae2000@yahoo.co.in

Abstract Cluster analysis is an unsupervised learning which reveals underlying structures in data and organizes them in clusters based on similarities. The approach to the both hard and soft clustering involves the concept of partial membership of the instance in the clusters and distance measure in the cluster. Clustering algorithms that have been analyzed are Fuzzy c-means (FCM), K- Means and K-Medoids etc. All these clustering algorithms do have some successful applications in agriculture, medicine, education, finance and business. Pre-processing is one of the key components in the clustering framework. The main objective is to preprocess heterogeneous dataset for different clustering algorithms and the time complexity is analyzed. In this project heterogeneous dataset that contains missing value is obtained from the UCI repository and it is used for preprocessing. The data pre- processing techniques are applied on the target data set to fill the missing value and attribute reduction to increase the effectiveness of algorithm. The key idea of this project is to preprocess the heterogeneous dataset and apply different clustering algorithms thereby to obtain best clustering result based on the time complexity. Finally the resultant clusters is be validated using silhouette plot and time complexity is also analyzed.

Key words Clustering, Fuzzy c-Means, K-Means, K-Medoids, Pre-processing, Validation.

2010 Mathematics Subject Classification 94A15.

1 Introduction

Data mining is a very well known technique and process for the extraction of desirable knowledge or patterns from large databases for some specific purpose. It is also a process for merging together statistical analysis, machine learning and databases to extract hidden rules and relationships [1]. There are many data mining techniques like classification, association, clustering, etc. Clustering is an unsupervised

* Communicated, edited and typeset in Latex by *Lalit Mohan Upadhyaya* (Editor-in-Chief).
Received March 16, 2019 / Revised October 15, 2019 / Accepted November 26, 2019. Online First Published on June 03, 2020 at https://www.bpasjournals.com/.
Corresponding author R. Abitha, E-mail: abirae2000@yahoo.co.in
Refereed Proceedings of the National Conference on Mathematics and Computer Applications held at the Department of Mathematics, Women's Christian College, Chennai, India from January 29, 2019 to January 30, 2019.

learning form which aims at revealing patterns by partitioning the instances of a dataset into clusters based on similarity. The central idea of clustering is the distribution of the data points into cluster, so that each instance is typically more similar to the instances belonging to the same cluster than to the other clusters, were the intra cluster similarity is high and inter cluster similarity is low. There is a wide application area of clustering including business, science, medicine, agriculture, marketing, genetics and biology. Regarding the membership instances in the cluster, two major types of clustering can be distinguished: classical (hard) and fuzzy (soft) clustering. In hard clustering, each data point either belongs to a cluster completely or not, whereas the soft clustering is a form of clustering in which each data point can belong to more than one cluster [5].

Data preprocessing is one of the steps in the data mining process. Data-gathering methods are often loosely controlled, resulting in out-of-range values (Age: -1000), impossible data combinations (e.g.: college-name,: xyz, theatre: yes), missing values, etc. The data that are analyzed and the ones that have not been carefully screened for such problems can produce misleading results. Data preprocessing is a process performed on raw data, where it transforms the raw data into an understandable format. Data preprocessing prepares raw data for further processing.

The data goes through a series of steps during preprocessing:

(i) Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

(ii) Data Integration: Data with different representations are put together and conflicts within the data are resolved.

(iii) Data Transformation: It transforms the data to forms suitable for mining process.

(iv) Data Reduction: Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume but still contains critical information [1].

2 Literature review

Bedali [2] proposed a detailed study to improve the stability, accuracy of the clustering procedure with the increased computational complexity. They mainly focused on different fuzzy clustering algorithms on the datasets obtaining multiple partitions, which in the later stage were fused into the final consensus matrix. Finally they experimentally evaluated and compared the accuracy of this methodology.

Kavili [4] proposed an objective in which fuzzy clustering was used to cluster people into a number of groups based on their use, tendency and intention.Their paper is about the application of fuzzy clustering on the data of young people's attitude toward tobacco products.

Ganesan [3] analyzed the performance of the three algorithms based on the clustering output criteria. It is proven that the efficiency of k-means is better than that of the Fuzzy c-Means and those obtained by Gustafson [13, 14]. The results were compared with the results obtained from the repository. The results showed that Gustafson–Kessel [13, 14] produces close results to Fuzzy c-Means.

Ban [7] discussed about the fuzzy partition set that is obtained by Fuzzy c-Means Algorithm. The experimental result that is obtained is compared with the result obtained by the traditional Importance Performance Analysis (IPA). The main benefit is related with the deriving of the managerial decisions which become more refined due to the fuzzy approach.

Velmurugan [9] discussed about the paritition-based clustering algorithms like the K-Means and the Fuzzy c-Means. From the experimental analysis they found that the computational time of K-means algorithm is less when compared with that of the Fuzzy c-Means.The computational complexity (execution time) of each algorithm is analyzed and the results are compared with one another by them.

3 Methodology

The figure below (Fig.1) summarizes the proposed work flow where the dataset that are not preprocessed and the dataset that are preprocessed are applied into the clustering algorithm and the time complexity will be obtained. The preprocessed data consumes less time when compared with the data that are not preprocessed and finally the cluster results are evaluated.

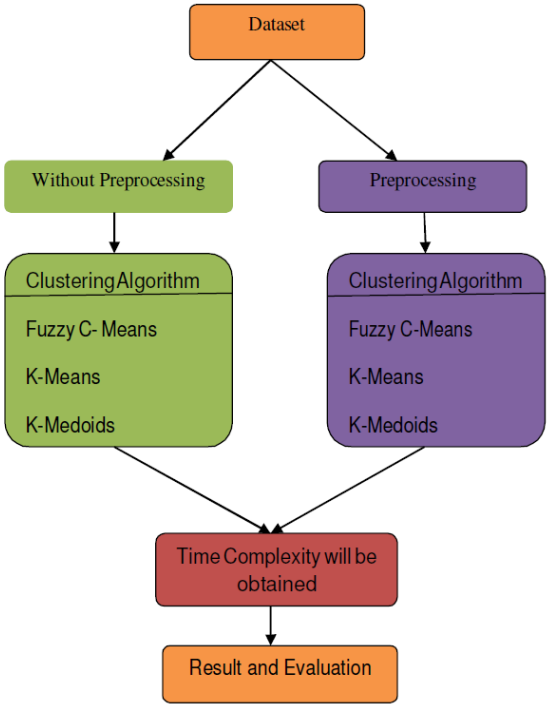

Fig. 1: The Proposed Work Flow.

4 An overview of the clustering algorithm

In this section we briefly describe the clustering algorithms which will be employed in our heterogeneous dataset. These are the Fuzzy c-Means (FCM), the K-Means algorithm and the K-Medoids algorithm.

4.1 Fuzzy c-Means algorithm

The Fuzzy c-means algorithm was developed by Dunn [15] and improved by Bezdek [5] and is one of the most widely used unsupervised learning algorithms. FCM works as an iteration scheme, aiming to achieve the objective function. The FCM algorithm is briefly described by the pseudo-code [7, 8].

Algorithm 4.1. Step 1: Fix the cluster c and select the value for the fuzziness parameter m,
Step 2: Initialize the partition matrix(μ),
Step 3: Calculate cluster centre for each step

$$v_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^2 x_{kj}}{\sum_{k=1}^{n} (\mu_{ik})^2},$$

Step 4: Calculate distance of each data points from each cluster centre using Euclidean Distance:

$$d_i = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2},$$

Step 5: Update partition membership metrics for each iteration

$$\mu_{ik} = \left[\sum_{j=1}^{c} (d_{ik}/d_{jk})^2 / m - 1 \right]^{-1}$$

Step 6: Check for convergence, if $\max \|U_{k+1} - U_k\| \leq \varepsilon$
stop;

else,
Go to the Step 3.

Here c represents the number of clusters, m represents the fuzziness parameter where the value of fuzziness parameter ranges from 1.25 to 2 and ε represents the threshold value.

4.2 *K*-Means algorithm

The term K-Means was first used by James Macqueen in 1967 [16, 17], though the idea goes back to Hugo Steinhaus in 1957 [16, 17]. K- Means clustering is a type of unsupervised learning (i.e., data without defined categories or groups). The key idea of this algorithm is to find groups in the data, with the number of groups represented by the variable K. Due to its ubiquity it is often called the K-Means algorithm; it is also referred to as the Lloyd's algorithm [18]. The K-means algorithm is briefly described by the pseudo-code [6].

Algorithm 4.2. Step 1: Initialize the K number of the cluster,

Step 2: Randomly select c, the cluster center,

Step 3: Calculate Euclidean distance using the given equation:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2},$$

Step 4: Recalculate the new cluster centers using:

$$c_i = (x_1 + \ldots + x_n)/n, (y_1 + \ldots + y_n)/n,$$

Step 5: Recalculate the Euclidean distance between each data point and new obtained cluster centers.

Step 6: Go back to the Step 3, unless the centroid are not changing or when no more new assignments are left.

Here K represents the number of cluster, c represents the cluster centre, $x_1 + \ldots + x_n$ and $y_1 + \ldots + y_n$ are the set of data points.

4.3 *K*-Medoids algorithm

The K-Medoids algorithm is a partition clustering algorithm which is a slightly modified from the K-Means algorithm. Both the algorithms the K-Means and the K-Medoids attempt to minimize the squared-error but the K-Medoids algorithm is more robust to noise than the K-Means algorithm. In the K-Means algorithm, the means are chosen as the centroids but in the K-Medoids, data points are chosen to be the medoids. The key idea of this algorithm is to first compute the K representative objects which are called the medoids. After finding the set of medoids, each object of the data set is assigned to the nearest medoids [6].

Algorithm 4.3. Step 1: Initially select k random points as the medoids from the given n data points of the data set,

Step 2: Associate each data point to the closest medoid by using any of the most common distance metrics,

Step 3: For each medoid m, for each non-medoid data point o:

Swap m and o, recompute the cost (sum of distances of points t their medoid),

If the total cost of the configuration increased in the previous step, undo swap.

Step 4: Repeat the Steps 2 and 3 until there is no change of the medoids.

5 Results and discussion

In order to estimate the accuracy of the clustering algorithm, the time complexity of the different clustering algorithms for different dataset is evaluated for the preprocessed dataset and for the data without preprocessing. The preprocessed dataset gives a more accurate result when compared with the dataset that is not preprocessed. Thus these findings exhibit the significance of preprocessing. Further the clustered data being validated by silhouette plot, where it validates by computing the value that falls between -1 to 1. It is proved that the Fuzzy c–Means algorithm4.1 gives the best clustering result, when compared with the K-Means algorithm 4.2 and the K-Medoids algorithm 4.3. The dataset that have been used in our analysis were from UCI repository namely hepatitis, dermatology, and the vote dataset. A brief description about the these datasets now follows:

1. The Hepatitis dataset contains information about liver disease with the class attributes, where the class attribute contains value – live, die [10].
2. The Dermatology dataset contains information about patients' clinical features 33 of which are linear valued and one of them is nominal [11].
3. The Vote dataset contains information about the U.S Congress (267 democrats, 168 republicans), with Class Distribution: two classes – 45.2 percent are democrats and 54.8 percent are republican [12].

The main characteristics of the dataset are provided in the following Table 1 (Fig.2), while the Table 2 (Fig.3) summarizes the outcome of the time complexity obtained in analyzing the Fuzzy c-Means algorithm 4.1 for the different dataset like hepatitis, dermatology, and vote. In the Table 3 (Fig.4) the outcome of the time complexity obtained in analyzing the K-Means algorithm 4.2 for the different dataset like hepatitis, dermatology and vote is presented and finally the Table 4 (Fig.5) presents the same analysis for the same attributes for the K- Medoids algorithm 4.3.

Dataset	Number of attributes	Number of instances
Hepatitis	20	155
Dermatology	34	366
Vote	17	435

Fig. 2: **Table 1:** The main characteristics of the used dataset.

Dataset	Missing value	Irrelevant Attribute	After Pre-processing	Time Complexity	
				Without Pre-processing	With Pre-processing
Hepatitis	no	yes	reduced to 11 attribute	22	21
Dermatology	yes	no	filled with mean value	13	12
Vote	yes	no	filled with mean value	15	12

Fig. 3: **Table 2:** The time complexity for the Fuzzy c- Means algorithm 4.1.

The outcomes of the time complexity obtained for hepatitis dataset when applied to the different clustering algorithms like the Fuzzy c-means algorithm 4.1, the K-Means algorithm 4.2 and the K-Medoids algorithm 4.3 are summarized in Fig.6. Similarly the corresponding outcomes for these algorithms are summarized in Fig.7 and Fig.8 for the Dermatology chart and the Vote chart respectively.

6 Conclusion

The analysis is done on pre-processing of heterogeneous dataset for ensemble clustering. In this analysis three different dataset hepatitis, dermatology, vote were taken from UCI repository for clustering purpose using three different clustering algorithms viz., the Fuzzy c-Means algorithm 4.1, the K-Means algorithm 4.2 and the K-Medoids algorithm 4.3. After analysis the results showed the importance of preprocessing of dataset before applying the clustering technique to it. When applying clustering algorithm on the dataset without preprocessing it does not lead to the cluster formation, rather clusters are formed for the preprocessed dataset, while applying all the above mentioned three algorithms. The

Dataset	Missing value	Irrelevant Attribute	After Pre-processing	Time Complexity	
				Without Pre-processing	With Pre-processing
Hepatitis	no	yes	reduced to 11 attribute	27	26
Dermatology	yes	no	filled with mean value	30	14
Vote	yes	no	filled with mean value	35	13

Fig. 4: **Table 3:** The time complexity for the K- Means algorithm 4.2.

Dataset	Missing value	Irrelevant Attribute	After Preprocessing	Time Complexity	
				Without Pre-processing	With Pre-processing
Hepatitis	no	yes	reduced to 11 attribute	80	74
Dermatology	yes	no	filled with mean value	63	50
Vote	yes	no	filled with mean value	80	55

Fig. 5: **Table 4:** The time complexity for the K- Medoids algorithm 4.3.

time complexity for all these three algorithms are analyzed for hepatitis, dermatology and vote dataset. The results show that the Fuzzy c- Means works faster than the K-Means and the K-Medoids algorithm. We also observe that the K-Medoids algorithm takes much larger amount of time when compared to the other two algorithms. The resultant clusters are validated using silhouette plot. The comparison is done on resultant cluster formed with and without preprocessing of data. Clusters formed from unprocessed dataset are not accurate as it shows values which are not in the range -1 to $+1$, since it contains missing values. But the cluster formed from the preprocessed data are accurate and the values fall between -1 to $+1$.Hence this project concludes that the preprocessed data forms accurate clusters for hepatitis, dermatology and vote dataset using the above said three different clustering algorithm. It is shown from this analysis that the Fuzzy c-Means algorithm 4.1 works faster and forms accurate clusters than the K-Means algorithm 4.2 and the K-Medoids algorithm 4.3.

References

[1] Agrawal, Rakesh, Imielinski, Tomasz and Swami, Arun (1993). Mining association rules between sets of items in large databases, 3rd edition, IBM Almaden Research Center, San Jose, California.

[2] Bedali, Erin (2016). A heterogeneous cluster ensemble model for improving the stability of fuzzy cluster analysis, *Procedia Computer Science*, 102, 129–136.

[3] Ganesan, G. (2016). Performance comparison of fuzzy and non-fuzzy classification methods, *Procedia Computer Science*, 17, 183–188.

[4] Kavili, Hazel (2016). An application of fuzzy clustering on the prevalence of youth tobacco survey, *Procedia Computer Science*, 38, 70–76.

[5] Bezdek, J.C. (1981). Pattern recognition with fuzzy objective function algoritm, Plenum Press, New York.

[6] Jain, A.K., Dubes, R.C. (1988). Algorithms for clustering data, Prentice Hall, New Jersey.

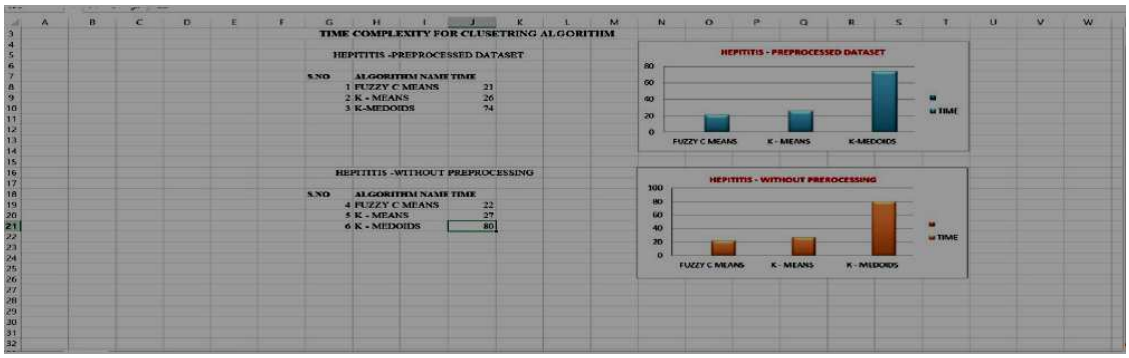

Fig. 6: The Hepatitis chart.

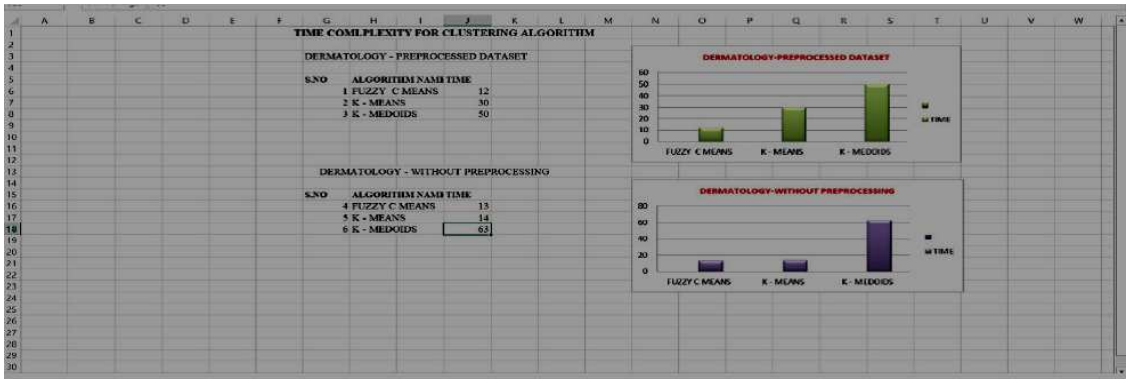

Fig. 7: The Dermatology chart.

[7] Ban, Olympia I. (2016). Importance-performance analysis by fuzzy *c*-means algorithm, *Procedia Computer Science*, 5, 9–16.

[8] Punera, K. and Ghosh, J. (2007). Soft cluster ensembles, *Advances in Fuzzy Clustering and its Applications*, 20, 307–318.

[9] Velmurugan, T. (2014). Performance based analysis between K–means and fuzzy c-means clustering algorithms for connection oriented telecommunication data, *Procedia Computer Science*, 19, 134–146.

[10] UCI (2012). Dermatology dataset http://archive.ics.uci.edu/ml/machine-learningdatabases/dermatology-disease.

[11] UCI (2009). Hepatitis dataset http://archive.ics.uci.edu/ml/machine-learningdatabases/hepatitis-disease.

[12] UCI (2016). Vote dataset http://archive.ics.uci.edu/ml/machine-learningdatabases/UScongressdataset.

[13] Gosain, Anjana and Dahiya, Sonika (2016). Performance analysis of various fuzzy clustering algorithms: a review, *Procedia Computer Science*, 79, 100–111.

[14] Malhotra, Virender Kumar, Kaur, Harleen and Alam, M. Afshar (2014). An analysis of fuzzy clustering methods, *International Journal of Computer Applications*,94, 105–111.

[15] Grover, Nidhi (2014). A study of various fuzzy clustering algorithms, ıInternational Journal of Engineering Research (IJER), 3, 177–181.

[16] Shukia, Shraddha and Naganna, S. (2014). A review on K-means data clustering approach, *International Journal of Information and Computation Technology*, 4, 1847–1860.

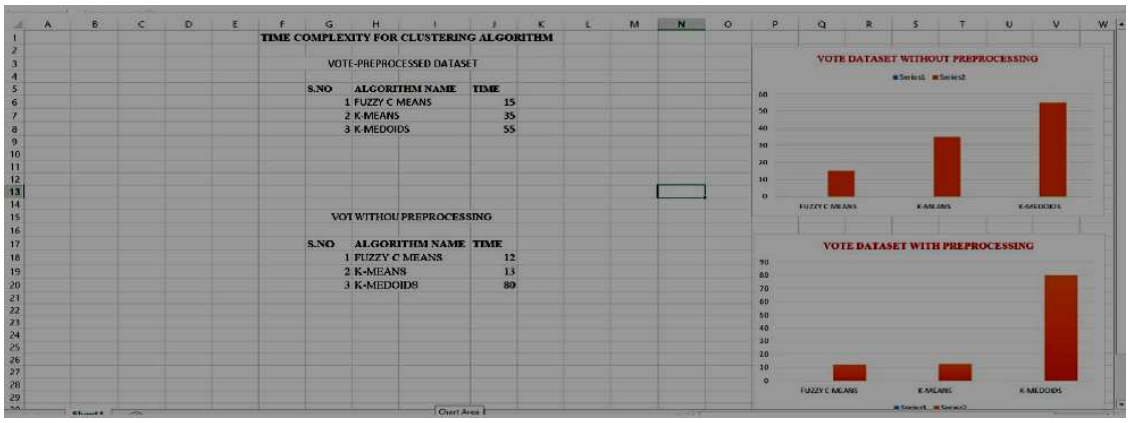

Fig. 8: The Vote chart.

[17] Bora, Dibya Jyoti and Gupta, Anil Kumar (2014).Effect of different distance measures on the performances of K-means algorithm, *International Journal Of Computer Science and Information Technology IJCSIT)*, 5, 2501–2506.

[18] Kanungo, Tapas, Wu, Angela Y. and Netanayhu, Nathan S. (2002). An efficient K-means clustering algorithm, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 881–892.