INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Fuzzy clustering of data mining: A survey paper

Shivani Singh
vishen2894shivani@gmail.com
Suyas Institute of Information Technology, Gorakhpur, Uttar Pradesh

Dr. A. K. Singh
Ajeet11@gmail.com
Suyas Institute of Information Technology, Gorakhpur, Uttar Pradesh

ABSTRACT

Clustering is the main and essentially used method for the automatic information taking out from huge amounts of data. Its task is to recognize groups, its called clusters, of indistinguishable objects in a data set. Clustering methods are used in a broad area, including database shopping, web inspects, information acceptance, bio technology, and broad others. Whenever, if clustering methods are used on real data, a problem that often produces up is that absent values shown in the data sets. For established clustering methods were developed to inspect complete data, there is a need for data clustering method pickup incomplete data. Proceed towards recommending in the literature to modifying the clustering algorithms to incomplete data effort better on data set with similar scattered clusters. In this thesis we are description a new proceed towards for suitable for a new use the fuzzy c-means clustering algorithm to incomplete data that takes get hold the scatters of clusters into a statement. In this experiment on made by a human being and real data sets with various scattered clusters .we show that we proceed out to accomplish another clustering technique for defective data. , We explain various cluster reliability functions and modify them to defective data according to the "present-case" proceed it. We inspect the original and the modify cluster reliability functions using the separate results of various artificial and real data produce by various fuzzy clustering algorithms for imperfect data. Therefore both the clustering algorithms and the cluster reliability functions are modified to imperfect data, we should target are finding the factor that is searching for determining the optimal number of clusters on defected data: the modify of the clustering algorithms, the modify of the cluster reliable functions or the miss of information in the data self. In this research, we present that our capacity is capable of right to check the clusters nearly situated to every other and cluster within broad density substance.

Keywords— *Data Mining, Fuzzy clustering, Clustering methods*

1. INTRODUCTION

World Wide Web is a very famous and attractive medium to distributing information till. According to the huge, and effective nature of web its lift up the expansible, multimedia data and globally issues. The construction of the web has lifted up to a broad amount of data that is freely present for user received by different users effectively and systematic. The research of web mining method to experiments helpful knowledge have increasing precious within the magnitude and a range of different thing of available information from the internet, this information to satisfy the need of the people within many backgrounds. In this a topic that attraction much notice is construct the web client's browsing design and doing guide best.

Web mining is the capacity to research web pages, a text document, multimedia file, images, and different content from the web.

2. CLUSTERING

Clustering defined as to find a group of things such that the things in one group must be similar to one another and different from objects in another group. Clustering is unsupervised learning. There are following types of clustering:

(a) Hierarchical: Agglomerative, Divisive
(b) Partitioning: k-means, nearest neighbour, PAM, MST
(c) Large-database: Sampling, Compression
(d) Categorical: Density-based

3. KNOWLEDGE DISCOVERY PROCESS

Knowledge Discovery Process (KDP) is shown in figure 1.

4. SIMPLE K-MEANS CLUSTERING

(a) K-means is a nonhierarchical cluster
(b) First, you have to decide how many numbers of clusters you needed i.e. firstly decide k. where k is a number of clusters you needed.

(c) Suppose if I select the k=8 cluster that I have needed then the k-mean algorithm will be implemented in 4 steps.

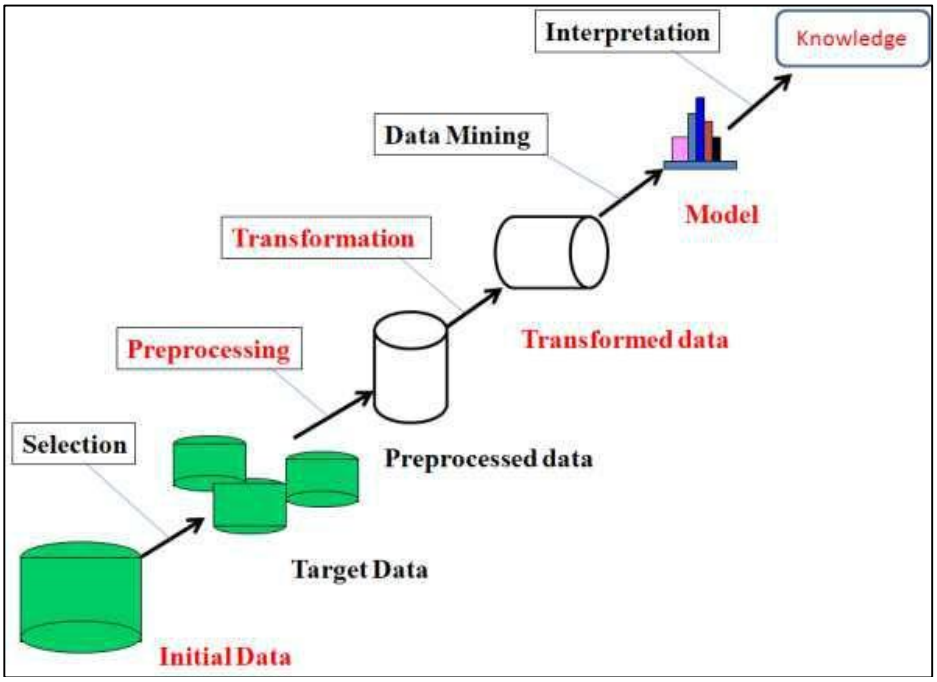


Fig. 1: KDD process

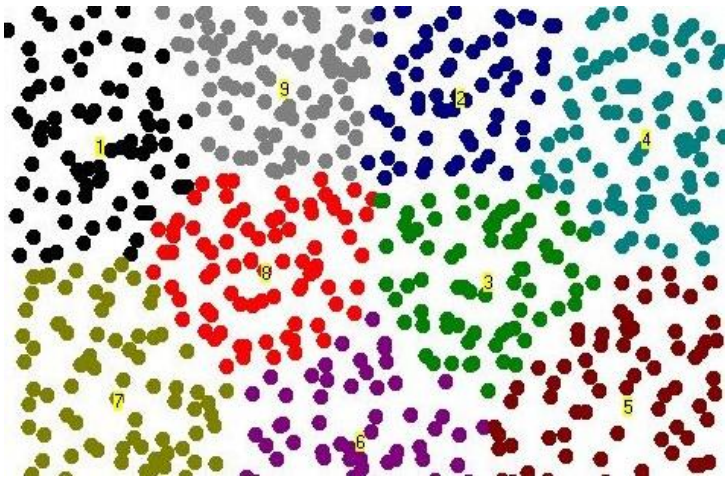
(1) Partition the whole dataset into k non-empty dataset randomly.
(2) Compute seed points as centroids of clusters of current partitioning, centroids will be centre i.e. mean point of clusters**.**

First centroids of parameter = x1+x2+x3+……+x15/15
Second centroids of parameter = y1+y2+y3+……+y15/15

(3) Assign each object to cluster with nearest seed point.
(4) Go back to step 2 and stop only when the assignment does not change.

5. RELATED WORK

The different types algorithm was K-Mean Clustering, thickness related clustering, the centre of mass based clustering, Distribution related clustering, Hierarchical rank clustering, and DBSCAN algorithm. Hard clustering was much followed in the base of conceptual sense and answered the query "Yes" or "Not" with respect to object classifying which is soft clustering or fuzzy clustering give into this detail picture an answer to question "maybe".



The upper picture describes an example of hard clustering. The clustering here is related to the colour. That is the best definition of the entity. The available of an individual object/feature belonging to a specific class is related to colour and hence the value is either 0 or 1.

The FCM algorithm is first described for a set of data points and then the algorithm is increase to a grey-scale image. Let us consider we has a group of data points $x = \{ x_1 , x_2, x_3, \ldots x_m\}$where every point $x_i = \{ x_1 , x_2, x_3, \ldots x_n\}$ is an *n*-dimensional vector.

Mathematically, it constructs form an *n*-dimensional vector and so an $m * n$ matrix. We should consider that the matrix is *A[m][n]*; then, we have 'K' clusters. So the problem should be done dividing the data into *k* clusters such as the distance between the centre of mass and every point is less amount. Here again, the distance should be Euclidean distance (distinguish between the square of

- Allocate a beginning made the centre of mass to each cluster (Group).

- Counting the distance between every point and the cluster centre uses a simple algorithm.
- It is related to the distance between each point and the cluster centre, again counting the membership function.
- It is related to the new membership function, again computing the centre of mass.
- Let's assume that difference between the original centroid and the next one is under a certain entry value say, €, then the algorithm finish, else it proceeds till that condition is true.

The steps for the FCM algorithm are illustrated as follows:

Step 1: Allocate an beginning made centre of mass to each cluster, say

$$c[0] = k_1,\ c[1]= k_2,\ c[2]= k_3 ,\ldots.. c[k] = k_n. \quad (1)$$

Step 2: Counting the distance between every point and the cluster centre uses a simple algorithm

$$Dist = | x_i - C_j| \quad (2)$$

Step 3: It is related to the distance between each point and the cluster centre, again counting the membership function.

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

Step 4: Its centre and every point in the image/ matrix.

Step 5: Counting the centre of mass from the matrix

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

$$J(u,C) = SSE(C1, C_2, C_3, \ldots\ldots\ldots\ldots, C_K) = \sum\sum u_{ij}^{p}\, dist(x_i - C_j)$$

The route for large the scope of the FCM algorithm is many. These can be modified in the associate ship function, algorithm design, setting an optimum number of clusters etc. The modification of the parameters is based on need. It is a better profile research topic with different applications. Image segmentation is the step of profile the Region of Interest (ROI) in the image for forwarding processing. For any time if we want to check the head tumour of the pituitary. In this the pituitary or the hypothalamus gland in the brain, then MRI scan image of the brain is further segmented to get the field of interest so that the continuity of the image can be done.

6. CONCLUSION

This thesis makes an investigation of fuzzy clustering techniques and their application in image segmentation problem. Conventional image segmentation techniques and Fuzzy c-Means (FCM) algorithm are applied in the segmentation of microscopic Pap smear images. Based on the shortcoming of these applications, a modified FCM algorithm is proposed which tries to rectify the shortcomings. The statistical data can be used to construct a database containing cell characteristics, which can be analyzed with various data analysis techniques. One of the major drawbacks of this method that it can only handle images with less complexity where there is no overlapping of cells. The adaptive histogram thresholding may fail to identify the right threshold if the contrast variation is very suitable. Despite these problems, the method is effective in case of a very simple image and the extracted features can be of great help for the pathologists.

7. REFERENCES

[1] Langhnoja, Shaily G., Mehul P. Barot, and Darshak B. Mehta. "Web usage mining using Total association rule mining on clustered data for pattern discovery." International Journal of Data Mining Techniques and Applications 2, no. 01 (2013).

[2] http://www.cs.waikato.ac.nz .

[3] Solanki A.V, Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 5857- 5860,2014.

[4] In Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on, vol. 2, pp. 159-164. IEEE, 1999. Masseglia, Florent, Maguelonne Teisseire, and Pascal Poncelet. "Real-time web.

