ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 2|| February – 2018 ||
International Journal of Computational Engineering Research (IJCER)

Critical Analysis ofClustering Algorithms

Atul Kumar Pandey* Sakuntla Dubey** Pankaj Shrivastava*Arti Pandey**
R.K. Katare**

*Department of Computer Science, APS University Rewa (M.P.)-India.***
*Department of Physics & Computer Applications, Govt. PG Science College, Rewa (M.P.)-India.**
*Corresponding author: Atul Pandey**

Abstract:
Clustering is an unsupervised learning problem which is used to determine the intrinsic grouping in a set of unlabeled data. This paper presents a comparative analysis of various clustering algorithms on two different datasets named Pima Diabetics and Cleveland Heart Disease dataset. In this experiment, the effectiveness of algorithms is evaluated by comparing the results on Weka tool.
***Key Words:** Clustering, Weka, K-Means, Farthest First, Make Density and EM.*

I. INTRODUCTION

In Data mining, there are mainly two approaches which is supervised and unsupervised used for the prediction and the description of the datasets. Clustering can be used as the pre-processing step before the classification of dataset. Clustering algorithms are often useful in various fields like data mining, learning theory, pattern recognition to find clusters in a set of data [8]. Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are “similar” and are “dissimilar” to the objects belonging to other clusters. Clustering is an unsupervised learning technique used for grouping elements or data sets in such a way that elements in the same group are more similar (in some way or another) to each other than to those in other groups. These groups are known as clusters. Clustering[1] is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, marketing, libraries, insurance, world wide web and bioinformatics. Cluster analysis was originated in anthropology by Driver and Kroeber in 1932 and introduced to psychology by Zubin in 1938 and Robert Tryon in 1939[2][3]. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently cluster the elements. Generally used scheme used to find similarity between data elements are inter and intra- cluster distance among the cluster elements.

II. CLUSTERING TECHNIQUES

Clustering is a method of data explorations, a technique of finding patterns in the data that of our interest. Clustering is a form of unsupervised learning that means we don’t know in advance how data should be group together [5]. A number of clustering techniques used in data mining tool WEKA have been presented in this section. These are:

2.1 Simple EM:

EM (expectation maximization) assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to create by cross validation, or you may specify apriori how many clusters to generate. EM finds clusters by determining a mixture of Gaussians that fit a given data set. Each Gaussian has an associated mean and covariance matrix. However, since we use spherical Gaussians, a variance scalar is used in place of the covariance matrix. The prior probability for each Gaussian is the fraction of points in the cluster defined by that Gaussian. These parameters can be initialized by randomly selecting means of the Gaussians, or by using the output of K-means for initial centers. The algorithm converges on a locally optimal solution by iteratively updating values for means and variances.

2.2 Farthest Fast:

Farthest first [9][10] is a heuristic based method of clustering. It is a variant of K Means that also chooses centroids and assigns the objects in cluster but at the point furthermost from the existing cluster centre lying within the data area. This places the cluster center at the point further from the present cluster. This point must lie within the data area. The points that are farther are clustered together first. This feature of farthest first clustering algorithm speeds up the clustering process in many situations like less reassignment and adjustment is needed. Fast clustering is provided by this algorithm in most of the cases since less reassignment and adjustment is needed.In the farthest-point heuristic, the point with highest score is selected as the first point, and remaining points are selected in the same manner as that of basic farthest-point heuristic.

2.3 Make Density Based:

Make Density based clustering has been long proposed as another major clustering algorithm [7]. The make density based clustering algorithm can also be used in noise and when outliers are encountered. The points with same density and present within the same area will be connected to form clusters. The density based method a natural and attractive basic clustering algorithm for data streams, because it can find arbitrarily shaped clusters, it can handle noises and is a one-scan algorithm that needs to examine the raw data only once. Furthermore, the density within the areas of noise is lower than the density in any of the clusters.

2.4 K-Mean Clustering:

K-means clustering technique [11] is one of the simplest unsupervised learning techniques that aim to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean value. Initially, k centroids need to be chosen in the beginning. The next step is to take instances or points belonging to a data set and associate them to the nearest centers. After finding k new centroids, a new binding has to be done between the same data set points and the nearest new center. Process is repeated until no more changes are done. Finally, this algorithm aims at minimizing intra cluster distance (cost function also known as squared error function), automatically inter cluster distance will be maximized.

III. DATA COLLECTION & PREPROCESSING

In this study, the two sample datasets Pima diabetic's andCleveland Heart Diseaseare collected from uci machine learning repository where the number of instances are 768 and 303 respectively and the number of attributes are 9 and 14 respectively.In Pima diabetic's dataset, there are no any missing values but in Cleveland Heart Disease 7 values are missing which is replaced with the mean/median value of the attribute using unsupervised ReplaceMissingValues filter of Weka. The conversion of dataset is not necessary because it is in default format (.arff) of Weka, so it can be directly implemented on tool. In the overall study, the four clustering algorithms are implemented against the two sample datasets as given above. The four clustering algorithms are EM,Farthest First,Make Density Based Clusters and Simple K-Means which are implemented on Weka tool and then further compare the results to obtain the suitable and appropriate algorithm on the basis of datasets.

IV. INTERPRETATION

While the dataset is implemented on Weka then the resulting accuracies is stored one by one which is illustrated below in table 1 and table 2. During implementation the mode of testing (cluster mode) is "classes to cluster evaluation" that behave like a classification and further creates the confusion matrix where the target attribute is the last attribute named "class" and "num".

Table 1: Prediction Accuracy of Clustering Algorithm on Pima Diabetics Dataset

.Clusters Algorithms	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy %	Time(Seconds)
EM	299	469	38.9323	58.92
Farthest First	505	263	65.7552	0.0
Make Density Based Clusters	514	254	66.9271	0.05
Simple K-Means	513	255	66.7969	0.03

When the clustering algorithms are implemented on**Pima Diabetics Dataset**, the accuracy of the EM algorithm is lowest and it also takes maximum time span (58.92 Seconds).Farthest First perform well as compared to EM and it takes minimum time span (0.0Seconds).Make Density Based Clusters achieved highest accuracy as compared to all and the time span is only 0.03 second that is overall good. Simple K-Means algorithm is also achieved the highest accuracy but less than as compared to Make Density Based Clusters algorithm whereonly one instance is more correctly classified by Make Density Based Clusters as compared to Simple K-Means. Finally it can be observed that, these two algorithms performed well and achieved the highest accuracies which are something equivalent to each other.

Similarly,when these clustering algorithms are implemented on **Cleveland Heart Disease Dataset**, the two algorithms Make Density (81.5182%) and EM (81.5182%) achieved the highest accuracies but the EM algorithm always take maximum time span (3.77 seconds). The next algorithm got the highest accuracy is Simple K-Means (80.8581%).

Table 2: Prediction Accuracy of Clustering Algorithm on Cleveland Heart Disease Dataset [6]

Clusters Algorithms	Correctly Classified Instances	Incorrectly Classified Instances	Prediction Accuracy %	Time(Seconds)
EM	247	56	81.5182	3.77
Farthest First	223	80	73.5974	0.02
Make Density	247	56	81.5182	0.02
Simple K-Means	245	58	80.8581	0.02

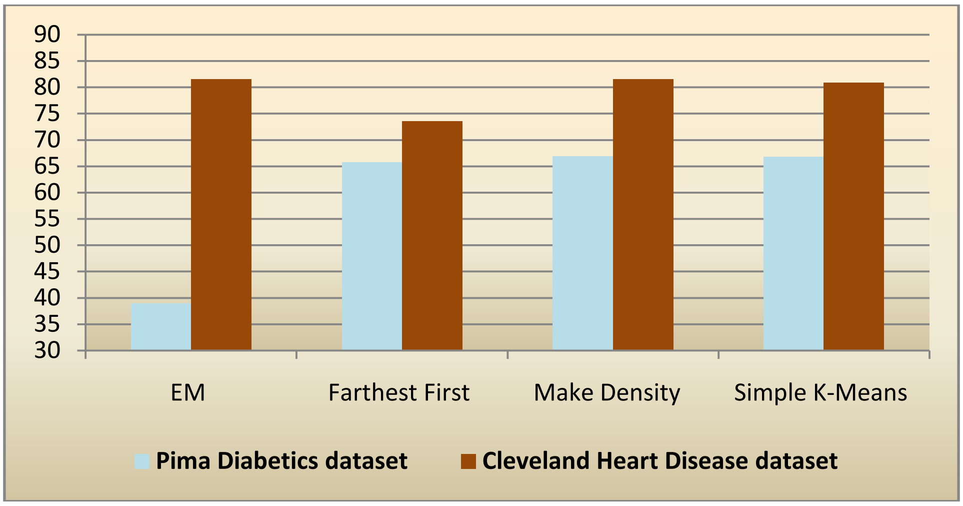

Figure 1: Prediction Accuracy of Pima Diabetics and Cleveland Heart Disease dataset

In the overall comparison,theaccuraciesof both datasets are differing too much that is mainly dependent on nature of dataset. It is observed the accuracies of all four clustering algorithm against the Cleveland datasetare better as compared to Pima diabetics dataset. The result of Make density and Simple K-Means algorithm is something equivalent on Pima Diabetics dataset whereas the result of EM and Make density algorithm are something equivalent against the Cleveland Heart disease dataset. The overall comparative results of both sample datasets against the four clustering algorithms are represented in figure1.

V. CONCLUSIONS

In this paper, the four clustering algorithms have been implemented on Weka tool against the two sample datasets named Pima diabetics and Cleveland Heart disease. These clustering algorithms are Expectation Maximization (EM), Farthest Fast, Make density Based cluster and Simple K-Means. In the entire study, it is observed thatthe Cleveland Heart disease dataset achieved the highest accuraciesoverall as compared to Pima diabetic's dataset.Make density based (66.9271%) and Simple K-Means (66.7969%) clustering algorithmsobtained the highest accuracies for the Pima diabetics dataset. These two clustering algorithms EM (81.5182%) and Make density based (81.5182%) got the highest accuracies for the Cleveland Heart disease dataset but EM algorithm took more time span (3.77 seconds). Similarly for the Pima diabetic's dataset, clustering Algorithm EM performed too badly and it took much more time span (58.92 seconds) as before. Therefore, the nature of dataset is also play an important in the selection of clustering algorithms.

REFERENCES

[1]. M. Mor, P. Gupta, and P. Sharma, "A Genetic Algorithm Approach for Clustering."

[2]. K. Bailey, "Numerical taxonomy and cluster analysis", Typol. Taxon, vol. 34, p. 24, 1994.

[3]. R. C. Tryon, "Cluster analysis: correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality", Edwards brother, Incorporated, lithoprinters and publishers, 1939.

[4]. Sander J., Ester M., Kriegel H., and Xu X., "Density-based clustering in spatial databases: The algorithm dbscan and its applications", Data Mining Knowledge Discovering, vol. 2, no. 2, pp. 169–194, 1998.

[5]. Saurabh Shah & Manmohan Singh, "Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm", International Conference on Communication Systems and Network Technologies, 2012.

[6]. Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal, Ashish Kumar Sen, "Data Mining Clustering Technique in the Prediction of Heart Disease using Attribute Selection Method", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 2 Issue 10, pp. 2003-08, October 2013.

[7]. Sander J., Ester M., Kriegel H., and Xu X., "Density-based clustering in spatial databases: The algorithm dbscan and its applications", Data Mining Knowledge Discovering, vol. 2, no. 2, pp. 169–194, 1998.

[8]. Ashish Kumar Sen, Prabhat Pandey, Atul Kumar Pandey, "Data Mining Techniques for the Prediction of Thyroid Disease: A Review", International Journal of Trend in Research and Development (IJTRD), Volume 4, No. 4, pp. 136-145, July-Aug 2017.

[9]. H. Zengyou, "Farthest-point heuristic based initialization methods for K-modes clustering", 2006.

[10]. M. Bilenko, S. Basu, and R. J. Mooney, “Integrating constraints and metric learning in semi-supervised clustering”, in Procs. of the twenty-first international conference on Machine learning, p. 11, 2004.

[11]. J. Han, M. Kamber, and J. Pei, “Data mining: concepts and techniques. Morgan Kaufmann”, 2006.

Shakuntla Dubey* “Critical Analysis of Clustering Algorithms” International Journal of Computational Engineering Research (IJCER), vol. 08, no. 02, 2018, pp. 19–22.