ANUSANDHAN- AISECT University Journal
Vol. 06, Issue No. 13, March-2018, P-ISSN 2278-4187, E-ISSN 2457-0656

Comparative Clustering Algorithms on Integrated Dataset

Atul Kumar Pandey[1]* Prabhat Pandey[2] Ajitesh Singh Bhaghel[3]

[1]Research Scholar, Dept. of Physics, Govt. PG Science College, Rewa (M.P.) India

[2]OSD, Additional Directorate Higher Education, Division, Rewa (M.P.) India

[3]Assistant Professor, Dept. of Computer Science, APSU, Rewa (M.P.) India

Abstract – Clustering is an unsupervised learning problem which is used to determine the intrinsic grouping in a set of unlabeled data and also applied in the preprocessing of datasets resulting further improvement in the next task such as classification. While clustering, grouping of objects is done on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity in such a way that the objects in the same group/cluster share some similar properties/traits. There is a wide range of algorithms available for clustering in various data mining tools. This paper presents a comparative analysis of four clustering algorithms in classes to cluster evaluation mode against the three datasets where one of them is integrated. In experiments, the effectiveness of algorithms is evaluated by comparing the results among the datasets and algorithms.

Key Words: Clustering, K-mean, Farthest First, Density, Simple EM and WEKA.

----------♦----------

I. INTRODUCTION

Clustering algorithms are quite useful in various fields like data mining, learning theory, pattern recognition to find clusters in a data set. Clustering is an unsupervised learning technique which is used for grouping elements or data sets. It is done in such a way that elements in the same group are more similar (in some way or another) to each other than compared to those in other groups. These groups are known as clusters. Clustering is the main task of exploratory data mining, and a common technique for statistical data analysis, which is 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(K. Bailey,1994) (R. C. Tryon,1939).

2. CLUSTERING TECHNIQUES

In this study various clustering techniques for data mining tool WEKA have been presented. These are:

(a) **Simple K-Means Clustering-** K-means clustering technique (J. Han. et.al. 2006). It is one of the simplest unsupervised learning techniques which 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. After this we take instances or points belonging to a data set and associate them to the nearest centers. The next step is to find k new centroids. A new binding has to be done between the same data set points and the nearest new center. Process is kept repeated until no more changes are done. Finally, this algorithm minimizes intra cluster distance (cost function also known as squared error function), automatically inter cluster distance will be maximized.

(b) **Farthest First-** Farthest first (H. Zengyou, 2006) (M. Bilenko 2004) is a heuristic based method of clustering. It is a variant of K Means that also chooses centroids and assigns the objects in a cluster but at the point furthermost from the existing cluster. However centre is within the data area. Fast clustering is provided by this algorithm in most of the cases as it needs less reassignment and adjustment. In the farthest-point heuristic, the first point highest score is selected, and remaining points are selected

in the same manner as that for basic farthest-point heuristic.

(c) **Simple EM-** Simple 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 on number of clusters to create by cross validation, or it may be specified as to 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. The EM algorithm for clustering is described in detail in Witten and Frank (Witten I. 2005).

(d) **Make Density Based Clustered-** Make Density based clustering has been long proposed as one of the major clustering algorithm (Sander J. 1998). The make density based clustering algorithm suits in noise and when outliers are encountered. Cluster will be formed by connecting the points with same density and present within the same area. 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. Also, the density within the areas of noise in this case is lower than the density in any of the clusters. Here the intuitive notion of "clusters" and "noise" in a database D of points of some k-dimensional space S is formalized.

3. TOOLS

WEKA (Waikato Environment for Knowledge Analysis) (E. Frank. et.al. 2005)(M. Hall, et.al. 2009) is an open source, platform independent and easy to use data mining tool portable, issued under GNU General Public License. It comes with Graphical User Interface (GUI) and contains collection of data preprocessing and modeling techniques. It is fully implemented in the Java programming language and therefore runs on almost any modern computing platform.

4. EXPERIMENTS

(a) **Data Source-** The publicly available heart disease database has been used. The Cleveland Heart Disease database consists of 303 records & Statlog Heart Disease database consists of 270 records and it is available at UCI Repository. (Website www.ucirepository.com.)

(b) **Cleaned & Integrated Datasets-** The missing values are replaced with the un-supervised filter and maintaining the consistency, datasets are made ready for the further critical investigation. The datasets so obtained after cleaning is Cleveland and Stat log which contain 303 and 270 instances with 14 features. After cleaning, the dataset Cleveland and Statlog are named H1 and H2 respectively. After that integrated dataset is created by combining the datasets Cleveland (H1) and Statlog (H2) named H11 containing all the 14 features where the number of instances is 573.

(c) **Comparison of Clustering Algorithms-**Four clustering algorithms namely EM, Farthest Fast, Make Density Based Cluster and Simple K-means were implemented to observe their performances. While clustering, the choice of testing mode for the algorithms is "classes to cluster" evaluation mode and the cluster value is two, where this mode performs clustering on classification basis resulting the two clusters 0 and 1 against the predicted (targeted) features.

The datasets with all the 14 attributes are H1, H2 and H11 (H1+H2) where H11 is an integrated dataset. The accuracy achieved and time span taken by the clustering algorithms were observed. The table 1 present the accuracy (with time span) of the four clustering algorithms against the datasets H1, H2 and H11.

Table 1 Accuracy of the Clustering Algorithms against all the 14 Features

Datasets →	H1		H2		H1+H2→H11	
Algorithms ↓	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
EM	81.5182	3.88	79.2593	3.16	48.6911	25.08
Farthest First	73.5974	0.0	72.5926	0.0	80.2792	0.02
Make Density	81.5182	0.03	71.4815	0.02	82.548	0.05
Simple K-Means	80.8581	0.02	59.2593	0.0	80.6283	0.05

For the dataset H1 (Cleveland), the algorithms Make Density Based (81.51%) and EM (81.51%) have obtained the same and highest prediction accuracy, but EM ranked in the second position due to their

more time span (3.88 seconds) taken to build the model. If EM is not considered due to their time span, next algorithm perform well is simple K-Means clustering.

Similarly for the dataset H2 (Statlog), EM (79.25%) got the highest prediction accuracy among them, but it took more time again as H1 and the Farthest Fast (72.59%) got the second position. When EM is not considering, then next algorithm performed well is Make Density Based Clusters (71.48%).

At last, for the dataset H11, Make Density Based algorithm (82.54%) has the highest prediction accuracy among all the clustering algorithms where Simple K-Means is lower at (80.62%) and third one is Farthest Fast (80.27%).

Moreover, the algorithm Make Density Based Clusters performed outstanding against the datasets H1 and especially for integrated dataset H11. EM also performed well in two datasets H1 and H2, but it took more time span against all the datasets and among all the algorithms. Datasets need to be strong for the reliable prediction. Farthest First loses its performance strength on datasets H1, H11. Accuracy differs due to the nature of datasets.

5. CONCLUSION

Various clustering algorithms made on non-integrated and integrated dataset have been compared and analysed. The results have been validated using integrated datasets that ensured the reliability of analysis. It is observed that datasets are successfully clustered with quite good accuracy. Few of the clustering techniques have better accuracy, others take less time, and many others have a trade-off between accuracy and time taken. Appropriate methods can be used according to their usage and the nature of datasets. Specifically, the algorithm Make Density Based Clusters has performed better against the datasets H1 and especially for integrated dataset H11.

6. REFERENCES

E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, (2005).“Weka,” in Data Mining and Knowledge Discov., Springer, pp. 1305–1314.

H. Zengyou, (2006). “Farthest-point heuristic based initialization methods for K-modes clustering,”.

J. Han, M. Kamber, and J. Pei, (2006). Data mining: concepts and techniques. Morgan kaufmann.

K. Bailey, (1994). “Numerical taxonomy and cluster analysis,” Typol. Taxon., vol. 34, pp. 24.

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

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, (2009). “The WEKA data mining software: an update,” ACM SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10–18.

M. Mor, P. Gupta, and P. Sharma, “A Genetic Algorithm Approach for Clustering.”

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

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

Witten I. H. & Frank E., (2005). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems): Morgan Kaufmann Publishers Inc, San Francisco.

Corresponding Author

Atul Kumar Pandey*

Research Scholar, Dept. of Physics, Govt. PG Science College, Rewa (M.P.) India

E-Mail – atul.pandey.it2009@gmail.com