Analysis of Various Clustering Algorithms

Sunila Godara, Amita Verma

***Abstract*— Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. This paper reviews four types of clustering techniques- k-Means Clustering, Farther first clustering, Density Based Clustering, Filtered clusterer. These clustering techniques are implemented and analyzed using a clustering tool WEKA. Performance of the 4 techniques are presented and compared.**

***Index Terms*— Data clustering, Density Based Clustering, Farther first clustering, Filtered clusterer, K-Means Clustering**.

I. INTRODUCTION

Clustering is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar between them- selves and dissimilar compared to objects of other groups. Cluster analysis is a very important technology in Data Mining. It divides the datasets into several meaningful clusters to reflect the data sets' natural structure. Cluster is aggregation of data objects with common characteristics based on the measurement of some kind of information. There are several commonly used clustering algorithms, such as K-means, Density based and Hierarchical and so on. [2] Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups.[3]Clustering is an unsupervised classification mechanism where a set of patterns (data), usually multidimensional is classified into groups (clusters) such that members of one group are similar according to a predefined criterion.

Clustering of a set forms a partition of its elements chosen to minimize some measure of dissimilarity between members of the same cluster .Clustering algorithms are often useful in various fields like data mining, pattern recognition, learning theory etc[14].

Terms:

A. Cluster

A cluster is an ordered list of objects, which have some common characteristics. The objects belong to an interval [a, b], in our case [0, 1]

B. Distance between Two Clusters

The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed.

C. Similarity

A similarity measure SIMILAR (Di, Dj) can be used to represent the similarity between the documents. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement.

D. Average Similarity

If the similarity measure is computed for all pairs of documents (Di, Dj) except when i=j, an average value AVERAGE SIMILARITY is obtainable. Specifically, AVERAGE SIMILARITY = CONSTANT SIMILAR (Di, Dj), where i=1, 2….n and j=1, 2….n and i < > j

E. Threshold

The lowest possible input value of similarity required to join two objects in one cluster.

F. Similarity Matrix

Similarity between objects calculated by the function SIMILAR (Di, Dj), represented in the form of a matrix is called a similarity matrix.

G. Dissimilarity Coefficient

The dissimilarity coefficient of two clusters is defined to be the distance between them. The smaller the value of dissimilarity coefficient, the more similar two clusters are.

H. Cluster Seed

First document or object of a cluster is defined as the initiator of that cluster i.e. every incoming object's similarity is compared with the initiator. The initiator is called the cluster seed.[6]

II. RELATED WORK

Comparisons between Data Clustering Algorithms

Osama Abu Abba, Computer Science Department, Yarmouk University, Jordan [2]. This paper is intended to study and compare different data clustering algorithms. The algorithms under investigation are: k-means algorithm, hierarchical clustering algorithm, self-organizing maps algorithm and expectation maximization clustering algorithm. All these algorithms are compared according to the following factors: size of dataset, number of clusters, type of dataset and type of software used. Some conclusions that are extracted belong to the performance, quality, and accuracy of the clustering algorithms.

A. Comparative Study of Various Clustering Algorithms in Data Mining

Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta [3]. This paper reviews six types of clustering techniques- k-Means Clustering, Hierarchical Clustering, DB Scan clustering, Density Based Clustering, Optics, EM Algorithm. These clustering techniques are implemented and analyzed using a clustering tool **WEKA**. Performance of the 6 techniques are presented and compared.

Manuscript published on 30 June 2013.
*Correspondence Author(s)

Asst Prof. Sunila Godara, Computer Science Engineering Department, Guru Jambheshwar University of Science & Technology, Hisar, India.

Ms. Amita Verma, Computer Science Engineering Department, Guru Jambheshwar University of Science & Technology, Hisar, India.

Performance analysis of k-means with different initialization methods for high dimensional data

Tajunisha and Saravanan[4].In this paper, we have analyzed the performance of our proposed method with the existing works. In our proposed method, we have used Principal Component Analysis (PCA) for dimension reduction and to find the initial centroid for k-means. Next we have used heuristics approach to reduce the number of distance calculation to assign the data point to cluster. By comparing the results on iris data set, it was found that the results obtained by the proposed method are more effective than the existing method.

A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set

D.Napoleon, S.Pavalakodi [5].K-means clustering algorithm often does not work well for high dimension, hence, to improve the efficiency, apply PCA on original data set and obtain a reduced dataset containing possibly uncorrelated variables. In this paper principal component analysis and linear transformation is used for dimensionality reduction and initial centroid is computed, then it is applied to K-Means clustering algorithm.

Evolving limitations in K-means algorithm in data mining and their removal"

Kehar Singh,Dimple Malik and Naveen Sharma[6] *K*-means is very popular because it is conceptually simple and is computationally fast and memory efficient but there are various types of limitations in k means algorithm that makes extraction somewhat difficult. In this paper we are discussing these limitations and how these limitations will be removed.

Comparative Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Data set

N.S.Chandolikar & V.D.Nandavadekar[7]. This paper evaluate performance to two well known classification algorithms for attack classification. Bayes net and J48 algorithm are analyzed The key ideas are to use data mining techniques efficiently for intrusion attack classification.

Compression, Clustering, and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Sets

Mehmet Koyutu¨rk, Ananth Grama, and Naren Ramakrishnan[8].This paper presents an efficient framework for error-bounded compression of high-dimensional discrete-attribute data sets. Such data sets, which frequently arise in a wide variety of applications, pose some of the most significant challenges in data analysis. Sub sampling and compression are two key technologies for analyzing these data sets. The proposed framework, PROXIMUS, provides a technique for reducing large data sets into a much smaller set of representative patterns, on which traditional (expensive) analysis algorithms can be applied with minimal loss of accuracy. We show desirable properties of PROXIMUS in terms of run time, scalability to large data sets, and performance in terms of capability to represent data in a compact form and discovery and interpretation of interesting patterns. We also demonstrate sample applications of PROXIMUS in association rule mining and semantic classification of term-document matrices. Our experimental results on real data sets show that use of the compressed data for association rule mining provides excellent precision and recall values (above 90 percent) across a range of problem parameters while reducing the time required for analysis drastically. We also show excellent interpretability of the patterns discovered by PROXIMUS in the context of clustering and classification of terms and documents. In doing so we establish PROXIMUS as a tool for both preprocessing data before applying computationally expensive algorithms and directly extracting correlated patterns.

A Hierarchical Latent Variable Model for Data Visualization

Christopher M. Bishop and Michael E. Tipping[9]. We introduce a hierarchical visualization algorithm which allows the complete data set to be visualized at the top level, with clusters and sub clusters of data points visualized at deeper levels. The algorithm is based on a hierarchical mixture of latent variable models, whose parameters are estimated using the expectation-maximization algorithm. We demonstrate the principle of the approach on a toy data set, and we then apply the algorithm to the visualization of a synthetic data set in 12 dimensions obtained from a simulation of multiphase flows in oil pipelines, and to data in 36 dimensions derived from satellite images.

A Modified K-Means Algorithm for Circular Invariant Clustering

Dimitrios Charalampidis Member [10].This paper introduces a distance measure and a K-means-based algorithm, namely, Circular K-means (CK-means) to cluster vectors containing directional information, such as Fd, in a circular-shift invariant manner. A circular shift of Fd corresponds to pattern rotation, thus, the algorithm is rotation invariant. An efficient Fourier domain representation of the proposed measure is presented to reduce computational complexity. A split and merge approach (SMCK-means), suited to the proposed CK-means technique, is proposed to reduce the possibility of converging at local minima and to estimate the correct number of clusters. Experiments performed for textural images illustrate the superior performance of the proposed algorithm for clustering directional vectors Fd, compared to the alternative approach that uses the original K-means and rotation-invariant feature vectors transformed from Fd.

Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation

Fasahat Ullah Siddiqui and Nor Ashidi Mat Isa[11]. This paper presents an improved version of the Moving K Means algorithm called Enhanced Moving K-Means (EMKM) algorithm. In the proposed EMKM, the moving concept of the conventional Moving K-Means (i.e. certain members of the cluster with the highest fitness value are forced to become the members of the clusters with the smallest fitness value) is enhanced. Two versions of EMKM, namely EMKM-1and EMKM-2 are proposed. The qualitative and quantitative analyses have been performed to measure the efficiency of both EMKM algorithms over the conventional algorithms (i.e. K-Means, Moving K-Means and Fuzzy C-Means) and the latest clustering algorithms (i.e. AMKM and AFMKM). It is investigated that the proposed algorithms significantly outperform the other conventional clustering algorithms.

A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry

Mu-Chun Su and Chien-Hsing Chou[12]. In this paper, we propose a modified version of the K-means algorithm to cluster data. The proposed algorithm adopts a novel non metric distance measure based on the idea of point symmetry. This kind of point symmetry distance can be applied in data clustering and human face detection. Several data sets are used to illustrate its effectiveness.

An Efficient k-Means Clustering Algorithm: Analysis and Implementation

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y.Wu[13].In this paper, we present a simple and efficient implementation of Lloyd's k means clustering algorithm, which we call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. We establish the practical efficiency of the filtering algorithm in two ways. First, we present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, we present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.

A Modified k-means Algorithm to Avoid Empty Clusters

Malay K. Pakhira [14].This paper presents a modified version of the *k*-means algorithm that efficiently eliminates this empty cluster problem. We have shown that the proposed algorithm is semantically equivalent to the original *k*-means and there is no performance degradation due to incorporated modification. Results of simulation experiments using several data sets prove our claim.

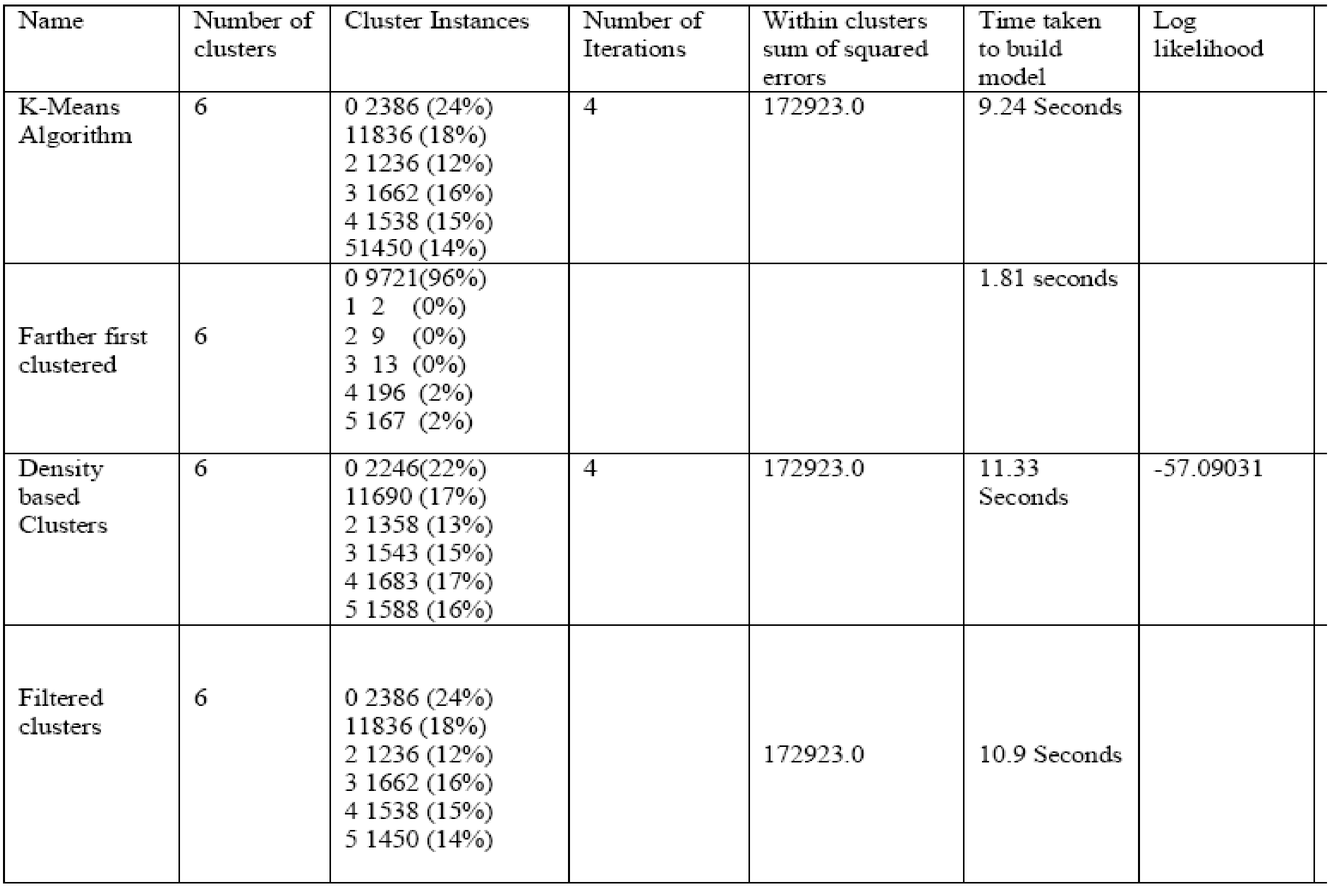
Comparison the various clustering algorithms of weka tool

Narendra Sharma, Aman Bajpai, Mr.Ratnesh Litoriya [15].In this paper we are studying the various clustering algorithms. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Our main aim to show the comparison of the different-different clustering algorithms on WEKA and find out which algorithm will be most suitable for the users.

III. TECHNIQUES

There are four types of Clustering Techniques are analyzed and implemented:

- Simple k-means clustering
- Farther first clustering
- Making density based clustering
- Filtered clusterer



Name	Number of clusters	Cluster Instances	Number of Iterations	Within clusters sum of squared errors	Time taken to build model	Log likelihood
K-Means Algorithm	6	0 2386 (24%) 11836 (18%) 2 1236 (12%) 3 1662 (16%) 4 1538 (15%) 51450 (14%)	4	172923.0	9.24 Seconds	
Farther first clustered	6	0 9721(96%) 1 2 (0%) 2 9 (0%) 3 13 (0%) 4 196 (2%) 5 167 (2%)			1.81 seconds	
Density based Clusters	6	0 2246(22%) 11690 (17%) 2 1358 (13%) 3 1543 (15%) 4 1683 (17%) 5 1588 (16%)	4	172923.0	11.33 Seconds	-57.09031
Filtered clusters	6	0 2386 (24%) 11836 (18%) 2 1236 (12%) 3 1662 (16%) 4 1538 (15%) 5 1450 (14%)		172923.0	10.9 Seconds	

IV. COMPARISON AND RESULT

Above section involves the study of each of the four techniques introduced previously and testing each one of them using **Weka Clustering** Tool on a set of internet usage dataset related to internet user information. The whole dataset consists of 72 attributes and 10108 instances. Clustering of the data set is done with each of the clustering algorithm using Weka tool and the conclusion is:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

V. CONCLUSION

After analyzing the results of testing the algorithms and running them under different factors and situations, we can obtain the following conclusions:

- Performance of K-Means algorithm increases as the RMSE decreases and the RMSE decreases as the number of cluster increases.
- The performance of K-Means algorithm is better than Dansity based Clustering algorithm.
- All the algorithms have some ambiguity in some (noisy) data when clustered.
- The quality of all algorithms become very good when using huge dataset.
- DBSCAN and OPTICS does not perform well on small datasets.
- K-Means is very sensitive for noise in dataset. This noise makes it difficult for the algorithm to cluster data into suitable clusters, while affecting the result of the algorithm.
- K-Means algorithm is faster than other clustering algorithm and also produces quality clusters when using huge dataset.
- Result of Farther First algorithm is not well because there are some empty clusters.
- Running the clustering algorithm using any software produces almost the same result even when changing any of the factors because most of the clustering software uses the same procedure in implementing any algorithm.

REFERENCES

1. Johannes Grabmeier, Fayyad, Mannila, Ramakrishnan, “Techniques of Cluster Algorithms in Data Mining,”May 23 2001.
2. Osama Abu Abbas, Jordan, **“**Comparisons Between Data Clustering Algorithms**, ”**The International Arab Journal of Information Technology, vol. 5, no. 3, pp.320-326,Jul. 2008.
3. Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta, “A Comparative Study of Various Clustering Algorithms in Data Mining, ”International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com, vol. 2, Issue 3, pp.1379-1384,May-Jun. 2012.
4. Tajunisha and Saravanan, “Performance analysis of k-means with different initialization methods for high dimensional datasets, ”International Journal of Artificial Intelligence & Applications (IJAIA), vol. 1, no.4, pp.44-52,Oct. 2010.
5. D.Napoleon, S. Pavalakodi, ”A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set, ”International Journal of Computer Applications (0975– 8887),vol. 13, no.7, pp.41-46, Jan 2011.
6. Kehar Singh, Dimple Malik and Naveen Sharma, “Evolving limitations in K-means algorithm in data mining and their removal, **”**IJCEM International Journal of Computational Engineering &Management, vol. 12, pp.105-109,Apr. 2011.
7. N. S. Chandolikar, V. D. Nandavadekar, “Comparative Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Dataset**, ”**International Journal of Computer Science and Engineering(IJCSE),vol.1,pp.81-88,Aug 2012.
8. Mehmet Koyutu¨rk, Ananth Grama and Naren Ramakrishnan, ”Compression, Clustering and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Sets, ”IEEE Transactions on Knowledge and A Data Engineering”, vol. 17, no. 4, pp.447-461, Apr 2005.
9. Christopher M. Bishop and Michael E. Tipping, “A Hierarchical Latent Variable Model For Data Visualization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp.281-293, Mar. 1998.
10. Dimitrios CharalampidisI,“A Modified K-Means Algorithm for Circular Invariant Clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp.1856-1865, Dec 2005.
11. Fasahat Ullah Siddiqui and Nor Ashidi Mat Isa, ”Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation, ”IEEE, pp.833-841.
12. Mu-Chun Su and Chien-Hsing Chou, “A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry, ”IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp.674-680, Jun. 2001.
13. Tapas Kanungo, David M. Mount,Nathan S. Netanyahu, Christine D.Piatko, Ruth Silverman, and Angela Y. Wu, “An Efficient k-Means Clustering Algorithm: Analysis and Implementation, ”IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-891,Jul 2002.
14. Malay K. Pakhira, ”A Modified *k*-means Algorithm to Avoid Empty Clusters,” International Journal of Recent Trends in Engineering, vol. 1, no. 1, pp.220-226, May 2009.
15. Narendra Sharma, Aman Bajpai, Mr.Ratnesh Litoriya, ”Comparison the various clustering algorithms of weka tools,” International Journal of Emerging Technology and Advanced Engineering, vol. 2, pp.73-80, May 2012.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)