International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Research Article

Clustering Algorithms: Study and Performance Evaluation Using Weka Tool

Bhoj Raj Sharma[a*] and Aman Paul[a]

[a]Department of Computer Science, Eternal University, Baru Sahib, Sirmour (HP)

Accepted 02 August 2013, Available online 05 August 2013, **Vol.3, No.3 (August 2013)**

Abstract

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Clustering is a procedure to organizing the objects in to groups or clustered together, based on the principle of maximizing the intra-class similarity and minimizing the inter class similarity. The various clustering algorithms are analyzed and compare the performance of clustering algorithms on aspect for time taken to build the model, Epsilon, minpts. The aim is to judge the efficiency of different data mining algorithms on diabetic dataset and determine the optimum algorithm. The performance analysis depends on many factors encompassing test mode, distance function and parameters.

Key words*: Data mining, cluster analysis, clustering algorithms, distance function, Weka 3.6.9 tools, Performance analysis*

1. Introduction

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group is called cluster which are more similar (in some sense or another) to each other than to those in other groups. It 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, and bioinformatics. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data pre-processing and model parameters until the result achieves the desired properties.

2. Clustering algorithm

- **Simple K-Means**

The K-Means clustering algorithms is a classical and well known clustering algorithm and its discovers K(non-overlapping) clusters by finding K centroids" Center Points" and then assigning each point to the cluster associated with its nearest centroid (Gupta A *et al*..2011). The K-Means method aims to minimize the sum of squared distances between all points and the cluster centre. This procedure consists of the following steps, as described below:

K-Means clustering algorithms (Tiwari and Singh, 2012)

1. Choose K cluster centre to coincide with K randomly chosen patterns or K randomly define inside the hyper volume containing the pattern set.
2. Assign each pattern to the closest cluster centre.
3. Recomputed the cluster centres' using the current cluster membership.
4. If a convergence criterion is not met step 2.Typical convergence criteria are: no reassignment of patterns to new cluster centres, or minimal decrease in squard error (Tiwari and Singh, 2012).

- **Connectivity based Clustering (Hierarchical clustering)**

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. Hierarchical methods are usually classified into Agglomerative and Divisive methods depending on how the hierarchy is constructed. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form. These algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. (Ranjini and Rajalinngum, 2011)

*Correspoding author **Bhoj Raj Sharma** is a Research Scholar and **Aman Paul** is working as Assistant Professor

- **Density-based clustering (DB SCAN)**

The most popular density based clustering method is DBSCAN(Ester *et al.*, 1996) In contrast to many newer methods, it features a well-defined cluster model called "density-reach ability". Similar to linkage based clustering; it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects which can form a cluster of an arbitrary shape, in contrast to many other methods and plus all objects that are within these objects range.

- **Optics Algorithms**

The Ordering points to identify the clustering structure (OPTICS) (Ankerst *et al.*, 1999) algorithms is procedurally identical to that of the previously mentioned DB SCAN. The OPTICS technique builds upon DBSCAN by introducing values that are stored with each data object; and attempt to overcome the necessity to supply different input parameters. Specifically, these are referred to as the core distance, the smallest epsilon value that makes a data object a core object, and the reach ability-distance, which is a measure of distance between a given object and another. The reach ability-distance is calculated as the greater of either the core-distance of the data object or the Euclidean distance between the data object and another point. These newly introduced distances are used to order the objects within the data set. Cluster are defined based upon the reach ability information and core distances associated with each object; potentially revealing more relevant about the attributes of each cluster.

- **Filtered Algorithms**

Selection of information or pattern relevant to a query from an incoming stream is called filtering.it is the filtering system that decides whether the new information or pattern is relevant instantly, without waiting for other information to arrive.

The filtering algorithm is used for the purpose of filtering the information or pattern. In this the user supplies the keywords or a sample set of relevant information. On the arrival of new information they are comparing against the filtering profile and the information matching the keywords is presented to the user. Filtering profile can be corrected by the user by providing relevant feedback on the retrieved information. The user is not provided with the details of filtering algorithm used by the system. An appropriate clustering threshold is selected by the system based on the filtering profile and relevant information (Aslam *et al.,* 2000). The filtering algorithm as follows:

1. Find pre-filtering threshold θ.
2. Cluster the pre-filtered set.
3. Select clustering threshold σ, on the basis of the keyword and initial relevant document set.
4. For each new information or pattern α within the distance θ from the filtering profile:

Add the information α to the clustering using the above procedure.

If relevant tag of α is true then retrieve that information (α) and correct its relevancy if needed.

Filtering of information or pattern by collaboration of multiple agents, viewpoints and data sources is called collaborative filtering.

3. Related work

The purpose of clustering algorithm is to organize a collection of data items in to clusters, such items within a cluster are more similar to each other then they are in other clusters. They used k-means & k-mediod clustering algorithms and compare the performance evaluation of both with IRIS data on the basis of time and space complexity. In this investigation, it can be said that portioning based clustering methods are suitable for spherical shaped clusters in small to medium size data set. K-means and k-Medoids both methods find out clusters from the given data. The advantage of k-means is its low computation cost and drawback is sensitive to nosy data while k-mediod has high computation cost and not sensitive to noisy data. (Tiwari and Singh, 2012).

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. Mainly we try to show the comparison of the different-different clustering algorithms of Weka and find out which algorithm will be most suitable for the users. For perform the clustering we used the promise data repository. It is providing the past project data for analysis. Every algorithm has their own importance and we use them on the behaviour of the data, but on the basis of thesis research we found that K-means clustering algorithms are simplest algorithms as compared to other algorithms (Sharma *et al.,* 2012).

A comparative study of clustering algorithms across two different data items is performed here. The performance of the various clustering algorithms is compared based on the time taken to form the estimated clusters. The experimental results of various clustering algorithms to form clusters are depicted as a graph. As the number of cluster increases gradually, the time to form the clusters also increases. The farthest first clustering algorithm takes very few seconds to cluster the data items whereas the simple K-Means takes the longest time to perform clustering. Thus it is very difficult to use simple K-Means clustering algorithm for very large datasets. This proposal can be used in future for similar type of research work (Revethi and Nalini, 2013).

The proposed work is to analyse the three major clustering algorithms: K-means, farthest first and Hierarchical clustering algorithm. The result are tested on three data sets namely Wine, Haberman and Iris dataset using Weka interface A Performance percentage has been

Table 1: Summary of selected reference with goal

Reference	Goal	Data base	Data mining Algorithms	Software
Sharma N et al. (IJETAE)	Comparision the various clustering algorithms of weka.	ISBSG and PROMISE repository	DB Scan, EM, Cobweb, Optics, Farthest First, Simple K-Means	Weka 3.6.9
Revathi R et al. (IJARCSSE)	Performance Comparison of Various clustering Algorithms	Abalone and Letter image from UCI repository	Simple k-Means, Enhanced K-means, Farthest first, Make density based, Filtered	Weka 3.6.6
Pallavi, G sunila (IJERA)	A comparative Analysis of clustering Algorithms	Iris, Haberman, Wine from UCI repository	K-means, Hierarchical Clustering alogrithms	Weka 3.6.4
Tiwari et al., (IOSRJCE)	Performance of Computer Engineering	Zoo, Labour, super market from UCI Machine Learning repository	DB Scan, EM, Hierarchical, K-means	Weka 3.6.6
Tiwari and Singh,(IJERD)	Comparative Investigation of K-Means and K-Medoid Algorithm of Iris Data	Iris Data Set from UCI Machine Learning Repository	K-Means, K-Medoid	Mat Lab

calculated on dataset taking principal component analysis as another parameter of comparison. The result analysis shows that K-means algorithms performs well without inserting the principle component analysis filter as compared to the hierarchical clustering algorithm and Farthest first clustering since it have less instances of incorrectly clustered objects on the basis of class clustering. Hierarchical clustering as compared to Farthest Fast clustering gives better performance and farthest first clustering though gives a fast analysis when taken an account of time domain, makes comparatively high error rate. Using principle component analysis filter with this approach, this shows better results which are comparable with Hierarchical clustering algorithm (Pallavi and Godra, 2011).

The letter image recognition is a challenging problem for knowledge worker, the basic objective of this data set is to identify each of a large number of black and white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. The data describes the horizontal and vertical position of box, width and height of box and mean and correlation of box etc. In the result, the prediction accuracy that DB Scan clustering algorithms is weaker than others in generating cluster instances (Tiwari and Jha, 2012).

4. Datasets and tool used

1. Hardware

We conduct our evaluation on Intel Pentium P6200 platform which consist of 1 GB memory and 320 GB hard disk.

2. Software

In this experiment, we used Weka 3.6.9 and window 8 to evaluate the performance of clustering algorithms using time taken to build the model according to respective no of clusters. Weka is machine learning/data mining software written in Java language (distributed under the GNU Public License).

Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for developing new machine learning schemes. It can be used for Pre-processing, Classification, Clustering, Association and Visualization.

3. Data Set

The input data set is an integral part of data mining application. The data used in my experiment is either real world data obtained from UCI machine learning repository and widely accepted data set available in Weka toolkit. Diabetes data set comprises 768 instances and 9 attributes in the area of Health Science and some of them contain missing value.

4. Experiments result and discussion

To evaluate the selected tool using Diabetes dataset and comparisons are performed in two parts. In first Comparison, I have applied these Clustering algorithms(Simple K-Means, Filtered Cluster, Make density Based cluster, Hierarchical clustering) by using two distance function namely Euclidean Distance and Manhattan Distance in three different Test Modes namely Training Mode, Supplied Test Set and Percentage Split in Weka. By doing so I found the most efficient algorithm

Table 2: The UCI datasets used for the experiments and their properties

Data Set	Diabetes
Instance	768
Attributes	9
Area	Health Science
Missing values	0

among three algorithms and in second parts, DB Scan and Optics clustering algorithms are compared by using vary the parameters E (epsilon) and M(minpts).

Applying Simple k-means, Filtered Clustered, Make Density Based Cluster, Hierarchical Clustering algorithms on diabetes dataset using Euclidean Distance were found to be 0.390, 0.205, 0.248 and 1.483 respectively in Training mode 0.225, 0.196, 0.226 and 1.566 with supplied test set while those in percentage split set were 0.131, 0.128, 0.143 and 0.311.

Performing Simple k-means, Filtered Clustered, Make Density Based Cluster, Hierarchical Clustering algorithms on diabetes dataset using Manhattan Distance were found to be 0.248, 0.243, 0.271 and 1.530 respectively in Training mode 0.215, 0.205, 0.221 and 1.501 with supplied test set while those in percentage split set were 0.090, 0.091, 0.121 and 0.352.

DB Scan and OPTICS algorithms are two other algorithms that were applied on the dataset. Two parameters- epsilon and minpts were considered. One of the parameter was kept constant with varying the second parameter and the efficiency of the above two algorithms was measure in term of time taken to build the model. When epsilon kept constant (has been assigned a constant value) i.e. 0.9 and minpts were assigned three different values- 2.0, 7.0 and 9.0. then resulting time taken to build the model was 0.53, 0.45 and 0.47 respectively with the mean value of 0.48 which shows random alternative pattern with increase in values of minpts and when minpts were kept constant i.e. 6.0 with varying epsilon value- 0.3, 0.7 and 0.9 then, time taken was 0.52, 0.45 and 0.44 respectively with the mean value of 0.47, it shows decrease in time to build model with increasing epsilon. Applying OPTICS algorithms on the diabetes dataset with constant epsilon value i.e. 0.9 and varying minpts as 2.0, 5.0 and 7.0 the corresponding time taken was- 0.64, 0.63 and 0.67 and the average come outs to be 0.64 which shows random alternative pattern with increase in values of minpts. When the same algorithm was applying on the same data set assigning different value as 8.0, 3.0 and 1.0 keeping minpts constant as 6.0 and the corresponding time taken was 0.64, 0.63 and 0.67 then average was found to be 0.64 which shows random alternative pattern with increase in values of epsilon.

5. Conclusion

By using Euclidean Distance function, the minimum time taken to build the model was 0.205 in Training mode, obtained by Filtered Clustered. In Supplied test set and in percentage split as well Filtered clustered was attaining least value among all the four algorithms i.e. 0.196 and 0.128 respectively.

When the algorithms were applied using Manhattan Distance function, the minimum time taken to build the model was by Filtered clustered in both Training mode and Supplied test mode i.e. 0.243 and 0.205. While when test was applied to Percentage split the minimum time was 0.090 obtained by Simple K-means.

When compared the DB Scan and OPTICS algorithms with parameters (epsilon and minpts) on diabetic dataset it is analysed that in DB Scan when epsilon was kept constant with minpts as variants the mean value was 0.48 while when epsilon was variant and minpts was constant the mean value was 0.47.

In OPTICS algorithm when Epsilon was kept constant with minpts as variants and when epsilon was variant and minpts was constant as well the mean value was 0.64.

In conclusion, Filtered Clustered has the highest efficiency using Euclidean Distance and Manhattan Distance method in Training mode and Supplied test set as well. Also Filtered Clustered has the highest efficiency using Euclidean Distance in Percentage split but Simple K-means in Manhattan Distance.

Among all algorithms considering Euclidean Distance and Manhattan Distance methods using three modes; simple K-means achieves the highest efficiency with average time of 0.09 in percentage split mode using Manhattan Distance methods and DB Scan is analysed more efficient when epsilon is variant by keeping minpts value constant. Whereas OPTICS algorithm was efficient in same proportion in both cases with varying or constant any of the two parameters. DB Scan takes less time to build the model when compared with OPTICS algorithm.

References

Tiwari M and Singh R (2012) Comparative Investigation of K-Means and K-Medoid Algorithm of IRIS Data.*In the International Journal of Engineering Research and Development*, 4: 69-72

Sharma N, Bajpai A and Litoriya R (2012) Comparision the various Clustering algorithms of Weka.*In International Journal of Emerging Technology and Advanced Engineering*, 2:73-80

Revathi S and Nalini T (2013) Performance Comparison of Various Clustering Algorithm.*In International Journal of Advanced Research in Computer Science and Software Engineering*, 3: 67-72

Tiwari M and Jha MB (2012) Enhancing the performance of Data Mining Algorithms in Letter Image Recognition Data.*In International Journal of Computer Application in Engineering Sciences*, 2: 217-220

Cai Z, Li Q and ZhengX (2005) An Experimental Comparison of Three Kinds of Clustering Algorithms. *In International Conference on Neural Network and Brain (IEEE)*, 767-771

Defays D (1977) An efficient algorithm for a complete link method. *In the Computer Journal (British Computer Society)*, 20: 364–366

Dutta V, Sharma KK and Gahalot D (2012) Performance Comparison of Hard and Soft Approaches for Document

Clustering. *In the International Journal of Computer Application*, 41: 44-48

Gupta A, Gupta A and Mishra A (2011) Research Paper on Cluster Techniques of data Variations.*In International Journal of Advance Technology & Engineering Research*, 1: 39-47

Godara S and Yadav R (2013)Performance analysis of lustering algorithms for character recognition using Weka tool.*International Journal of Advanced Computer and Mathematical Sciences*, 4: 119-123

Goddard J and Martinez AE (2000) A Comparison of Different Clustering Algorithms for Speech Recognition.*In Midwest Symposium on Circuit and Systems (IEEE)*, 3: 1222-1225

Jain S and Gajbhiye S(2012) A Comparative Performance Analysis of Clustering Algorithms.*In International Journal of Advanced Research in Computer Science and Software Engineering*, 2: 441-445

Jain S, Aalam MA and Doja MN (2010) K-Means Clustering Using Weka Interface. *Proceedings of the 4th National Conference, INDIACom-2010 Computing For Nation Development*, 1-6, Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi

Javed Aslam, Katya Pelekhov and Daniela Rus (2000) Using Star Custers for Filtering. CIKM '00 *Proceedings of the ninth international conference on Information and knowledge management*, 306-313,Departmetn of Computer Science, Dartmouth College, Hanover, NH

Latiff NMA, Tsimendis CC and Sharif BS (2007) Performance Comparison of Optimization Algorithms for Clustering in Wireless Sensor Networks. *In International Conference on Mobile Adhoc and Sensor Systems(IEEE)*, 1-4

Lau TK and King I (1998) Performance Analysis of Clustering Algorithms for Information Retrieval in Image Databases. *IEEE*, 932-937

Meena K, Subramaniam KR and Gomathy M (2012) Performance Analysis of Gender Clustering and Classification Algorithms. *In International Journal of Computer Science and Engineering*, 4: 442-457

Nathiya G, Punitha SC and Punithavalli M (2010) An Analytical Study on Behavior of Clusters Using KMeans, EM and K* Means Algorithm. *In International Journal of Computer Science and Information Security*, 7: 185-190

Pallavi and Godara S (2011) A Comparative Performance Analysis of Clustering Alogrithms. In International Journal of Engineering Research and Application, 1: 441-445

Ranjini K and Rajalinngum N (2011) Performance Analysis of Hierarchical Clustering Algorithm. *In International Journal of Advanced Networking and Applications*, 3: 1006-1011

Tan P, Steinbach M and Kumar V (2006) Introduction to Data Mining. *Addison Wesley*, 1: 157-169

Velmurugan T and Santhanam T (2010) Clustering Mixed Data Points Using Fuzzy C- Means Clustering Algorithms for Performance Analysis. *In International Journal on Computer Science and Engineering*, 2: 3100-3105