ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018

Analysis of Clustering and its Best Algorithms in Data Mining

Kiruthika.V[1], Sampath Kumar. D[2], Megha.K.B[3]

Student, Dept. of Computer Technology, Sri Krishna Arts & Science College, Coimbatore, Tamil Nadu, India[1]

Assistant Professor, Dept. of Computer Technology,

Sri Krishna Arts & Science College, Coimbatore, Tamil Nadu, India[2]

Student, Dept. of Computer Technology, Sri Krishna Arts & Science College, Coimbatore, Tamil Nadu, India[3]

Abstract: In this paper we discuss about the clustering type of retrieving data from the database for data mining and also the algorithms used for clustering. We also analyse its best algorithm and the algorithm's drawbacks if any, thus giving a successful review on it. We here discuss about k-means clustering, Fuzzy-c means clustering and Hierarchial Clustering from the knowledge gathered from various publications in journals given further at the end in the reference section.

Keywords: Data Mining, Clustering, Algorithms, K-means Algorithm

I. INTRODUCTION

Retrieving processed hidden information from databases is called data mining and to retrieve the required information from there, we use various methods called as algorithms. Sure there are many methods to retrieve the data but we have to analyse which one would benefit the users most of the time under majority of conditions given. So we are going to analyse which are the top most algorithms in clustering of data in data mining and explicitly explain the top most algorithm.

II. WHAT IS DATA MINING?

Data could be called as processed information. Whereas Data mining is defined as finding hidden information from a database. Many methodologies and algorithms are present to find these data as desired by the user. Clustering of data is a method one among them.

III. CLUSTERING

Clustering analysis has been an emerging research issue in data mining due its variety of applications in various fields of science. With the arrival of many data clustering algorithms in the recent years and its extensive use in wide variety of applications, including image processing, computational biology, mobile communication, medicine and economics, has been leading to the popularity of these algorithms. Clustering is a process which partitions a given data set into homogeneous set of groups based on features that are similar object wise and are kept in a group whereas dissimilar objects are in different groups.

IV. CONDITIONS FOR CLUSTERING

1) Scalability - Data must be scalable otherwise we might get the wrong result. Fig (I) shows simple graphical example where we may get the wrong result.
2) Clustering algorithm should be able to deal with different types of attributes or data variables.
3) Clustering algorithm should be able to find clustered data with the arbitrary shape.
4) Clustering algorithm must be not sensitive to noise and any other outliers.
5) Interpret-ability and Usability - Result obtained must be interpretable and usable so that maximum knowledge about the input parameters can be obtained.

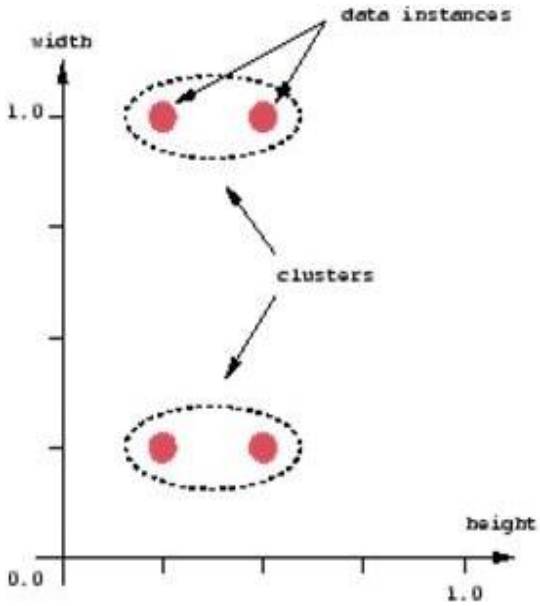
These above conditions should all be met by data heap to consider it under one cluster and display that under the specified data category. The diagram depicting the graphical difference between how data is being added to separate clusters are shown as:

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

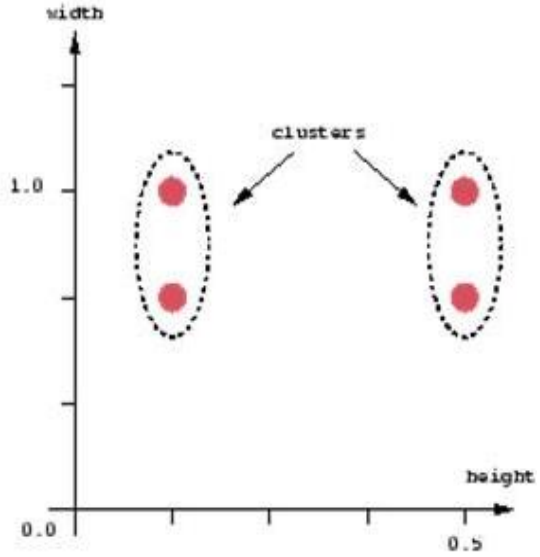
IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018



Before Scaling



After Scaling

Fig I: showing example where scalability may leads to wrong result

6) High dimensionality must be supported by the data to perform

V. TYPES OF ALGORITHMS

- ✓ Centroid based algorithms
- ✓ Connectivity based algorithms
- ✓ Density based algorithms
- ✓ Probabilistic
- ✓ Dimensionality Reduction

There are three algorithms that are the best in the clustering division and work efficiently among all the others even if the others have different uses. The top three algorithms are:

1. K-means clustering
2. Hierarchical clustering
3. Fuzzy means-c clustering

In these, K-means clustering is the most dependable and efficient or comfortable algorithm as it is simple. So, let's take a look into the said algorithm.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018

VI. HIERARCHIAL CLUSTERING

Hierarchical clustering is done using the Agglomerative or Divisive technique depending on whether the hierarchical breakdown is from top-down (Agglomerative) method or by bottoms-up (Divisive) method. In Agglomerative approach, at first one object is selected and is successively merged (agglomerates) with the closest similar pair based on similarity criteria until all the data form the desired cluster.
This algorithm is an agglomerative algorithm that has several variations depending on the metric used to measure the distances among the clusters selected or given.

1. Average linkage clustering: The dissimilarity between clusters are calculated using average values. The average distance is calculated from the distance between each point in a cluster and all the other points in another cluster. The two clusters with lowest average distance are joined together to form the new cluster.
2. Centroid linkage clustering: This variation type uses the group centroid as average. The centroid is defined as the centre of a collection of a cloud of points.
3. Complete linkage clustering (Furthest-Neighbour Method): The dissimilarity between 2 groups is equal to the greatest dissimilarity between a member of cluster l and a member of cluster m. This method tends to produce very tight clusters of similar cases and hence very similar.
4. Single linkage clustering (Nearest-Neighbour Method): The dissimilarity between two clusters is the minimum dissimilarity between the members of the two clusters. This method produces long chains which form loose and messy clusters.
5. Ward's Method: Cluster membership is assigned by calculating total sum of the squared deviations from the mean of a cluster. The criteria for fusion is that it should be able to produce the smallest possible increase in the error sum of squares.

Disadvantages:

1. Inability to make corrections once the splitting/merging decision is made.
2. Lack of interpretability regarding the cluster descriptors while performing algorithm.
3. Vague termination criterion is provided.
4. It is too expensive for high dimensional and massive or huge and huge datasets.
5. Highly ineffective in high dimensional spaces because hierarchy becomes complex and not related at times.

VII. K-MEANS CLUSTERING

K-means performs division of objects into clusters which are “characteristically similar” between them and those which are “dissimilar or not alike” to the objects belonging to another cluster. It is also called as Lloyd’s algorithm.

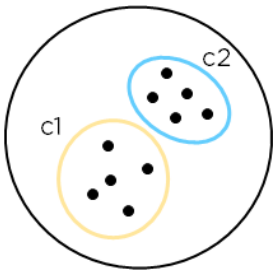
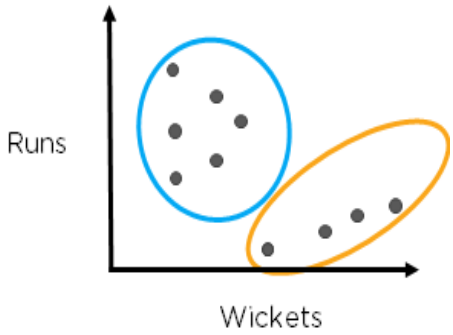


Fig (II) c1 and c2 clusters showing batsmen and bowlers.



The blue circle of cluster is identified as people with higher wickets and orange cluster – people with low runs.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018

Advantages:

1. Easy to implement.
2. Works with any of the standard norms.
3. Allows straightforward parallelization.
4. Insensitive to data ordering.
5. Computationally faster than hierarchical clustering.
6. Produces tighter clusters than hierarchical clustering.
7. An instance can belong to one or more clusters.
8. Along with neural networks, this method was implemented for applying it in heart diseases predictions using data.

Disadvantages:

1. Difficult to predict no. of clusters.
2. Initial constituents have impact on final result.
3. Order of the data also has an impact on final result.
4. Sensitive to scale.
5. We have to spend extra time to scale the data on which the algorithm is to be performed.

VIII. FUZZY-C MEANS ALGORITHM

This algorithm is done by assigning membership to each data point corresponding to each cluster centre on the basis of distance between the cluster centre and the data point. More the data is near to the cluster's centre point, the more is its membership towards the particular cluster centre. So the summation of membership of each data point should be equal to one.

Advantages;

1. Gives best result for overlapped data set and comparatively better then k-means algorithm.
2. Unlike k-means where data point must exclusively belong to atleast one cluster centre here data point is assigned a membership to each cluster centre as a result of which data point may belong to more than one cluster centre at the time.

Disadvantages;

1. A priori specification of the number of clusters is a disadvantage here.
2. With lower value of β we get the better result but only at the expense of more number of iterations to receive data.
3. Euclidean distance measures can unequally weight underlying factors at various circumstances.

CONCLUSION

The algorithm has a loose relationship to the k-nearest neighbour algorithm, a popular machine learning technique for classification that is often confused with k-means algorithm due to k in the name. Aside that k-means algorithm has a really great advantage over other algorithms in Clustering because it is simple and easy to implement. Even if there are some disadvantages that have to be taken care of, k-means algorithm is the best algorithm amongst the others in clustering technique of retrieving data from the database. Clustering is a means to collect and organise data for displaying accordingly to any graphical user interface by performing operations for their accuracy. K-means clustering technique is the most effective algorithm in clustering analysis that is applied on data. It gives the most accurate result and more convenient as it is very faster in fetching details from the server or a mainframe server.

REFERENCES

[1]. Advantages and disadvantages of k-means and hierarchical clustering (Unsupervised Learning), Marina Santini.
[2]. Introduction to clustering techniques, Leo Warner.
[3]. Fast efficient clustering algorithm for balanced data, Adel.A.Sewisy, M.H.Marghny, Rasha.M Abd Elaziz, Shmed.I.Tabola.
[4]. An Analysis on Clustering Algorithms in Data Mining, Mythili S1 , Madhiya E.
[5]. A. Jain , M. Murty , and p. Flynn " Data clustering: A review.," ACM Computing Surveys, vol. 31, pp. 264-323, 1999.
[6]. Shi Na,Liu Xumin, "Research on k-means Clustering Algorithm",IEEE Third International Conference on Intelligent Information Technology and Security Informatics,2010.
[7]. K. J. Cios, W. Pedrycz, and R. M. Swiniarsk, "Data mining methods for knowledge discovery," IEEE Transactions on Neural Networks, vol. 9, pp. 1533-1534, 1998.
[8]. Clustering Algorithms – A Literature Review, B. Ramesh , K. Nandhini.
[9]. Fuzzy c-means by Balaji K and Juby N Zacharias.
[10]. Fast and Robust Fuzzy C-Means Clustering Algorithms Incorporating Local Information for Image Segmentation by Weiling Cai, Songcan Chen and Daoqiang Zhang.
[11]. Guha, Meyerson, A. Mishra, N. Motwani, and O. C.."Clustering data streams: Theory and practice ." IEEETransactions on Knowledge and Data Engineering, vol. 15,pp. 515-528, 2003.