THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Implementation and Evaluation of Optimal Clustering Algorithm For Banking Application

Robin Prakash Mathur

Asst.Professor, School of Computer Science & Engineering

Lovely Professional University
Punjab, India
mathur.robin@gmail.com

Prabhjot Saini
Research Scholar, School of Computer Science & Engineering
Lovely Professional University
Punjab, India
saini_prabhjot@hotmail.com

***Abstract*— Human life is surrounded by era of computer generation. In every field whether science and technology, entertainment, profession, research, there is no place where we are not dealing with data and extraction of data which leads to data warehousing and data mining. Customer modeling is an important business application which uses data mining techniques for analysis purpose. Clustering is the popular method of mining where group of similar objects also known as clusters are formed which is highly dissimilar to other clusters. The present work compares the performance of clustering algorithms K-means, Self organized maps and Hierarchical clustering algorithm after applying them to banking dataset. Banking systems use cluster analysis to develop a customer's topology to retain the loyal customers by designing the best possible financial solutions to specific clusters. Experiment result will show the best accuracy, higher robustness and generalization ability in one of the algorithm.**

Keywords*— Clustering, Hierarchical clustering algorithm(HAC), Self-Organizing Map (SOM), Performance analysis*

I. INTRODUCTION

Our life is surrounded towards era of computer generation. In every field whether science and technology, entertainment, profession, research, there is no place where we are not dealing with data and extraction of data which leads to data warehousing and data mining. Customer modeling is an important business application which uses data mining techniques for analysis purpose in order to identify the set of customers for applying different set of loyalty schemes. In order to get success in data mining, we have two important key, one is formulation of the problem we are trying to solve and second is using the right data. Data mining is widely used in day to day business application since 1990s and helps the business organization in analyzing their business insights. Data mining is all about the different types of patterns and there are corresponding different types of data mining techniques like classification, clustering, association, estimation, link analysis etc. Data mining is also known as knowledge discovery from data (KDD) process. Several data mining tools are used as analytical tools for exploring data. It permits users to analyze data from several dimensions, performs categorization, and recapitulates the recognized relationships. Clustering

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[1] is the popular method of mining where group of similar objects also known as clusters are formed which is highly dissimilar to other clusters. As we know that clustering [2] [3] has their root in many areas like biology, statistics and also including data mining. The sub field of artificial intelligence is machine learning. The main job of machine learning [7] [8] is to develop and design the algorithms and techniques that implement various kinds of learning mechanisms to induce knowledge from several examples. One of the most important human activities is cluster analysis. Automated clustering concept can be used to identify sparse and dense regions in object space. Mainly clustering analysis is focusing on distance based cluster analysis. In machine learning, two main techniques i.e supervised learning and unsupervised learning prevails along with third as reinforcement learning. The machine learning technique which is used to create a function is called supervised learning. Learning with a teacher is also called supervised learning. One of the important methods of learning is unsupervised learning where a model is fit to its observations. There are two most important terms in cluster analysis that are preprocessing in which normalization the data and remove the outliers is performed and post processing in which removal of small clusters which acts like outliers is done. The focus of this work is on the clustering methods' scalability factor and their usefulness for performing clustering on high dimensional data [13]. Analysis of large databases can turn the passive data into actionable data. Data mining involve various tools which can be used to extract patterns from huge amount of data. In data mining large training set containing so many examples are common. If one needs to check the similarity between data points in a cluster, one need to find the Euclidean distance. Intracluster similarity measures how near the data objects are in the cluster. Inter-cluster distance is measured by within cluster sum of square. As the cluster density can be viewed as the compactness, so it is related to inter individual distances. The good quality clustering method forms group of similar objects also known as clusters which is highly dissimilar to other clusters. Euclidean distance can be calculated from the raw data. Using following standard formulae Euclidean distance D can be calculated as:

$$D((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2} \qquad \text{(i)}$$

II. BACKGROUND

In this section, we have discussed some methods used in comparative analysis of the clustering techniques:
Partitioning Method: Suppose the database contains n objects, this method will construct k- partitioning of corresponding data. Every partition or separation will represent a cluster. Since n denotes total objects count and k is the desired cluster formation required. As one standard is build that number of cluster are always less than or equal to the number of objects.Partitioning must be done according to the following given two conditions:

1. Every cluster holds one object atleast inside it.
2. Every object must belong to single cluster means no object can belong to two clusters.

In this, the very first condition is to do the initial partitioning that uses IRT (iterative relocation technique).Performance can be improved by following this technique. This technique can be used to move the objects from one cluster to another cluster. In order to achieve the aim of global optimality, requirement of exhaustive enumeration of all the partitioning is there. Partitioning method consist of two heuristic methods which are k-means and k-medoids. As in k-means clustering algorithm, cluster is allied with a centroid and each object is allocated to the cluster with the nearby centroid.

Hierarchical methods: This method is used to create a level by level decomposition of a given dataset. Every cluster node consists of a child cluster. Hierarchical clustering is also visualized as dendrogram. Clusters can be built gradually with the help of hierarchical algorithm. Hierarchical method is split into two main approaches:

1. Agglomerative approach (AGNES)
2. Divisive approach (DIANA)

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Another name of agglomerative approach (AGNES) is bottom up approach. This approach is basically used to merge the objects that are close to one another, this procedure is carried on until we get a termination condition. Another name of DIANA is top down approach. In this approach, a cluster is being partitioned into small clusters until each object get inside in a single cluster. There are several limitations of this method. One of the most important limitations of this method rollback of an operation i.e cluster formation is not possible once formed. Quality of hierarchical clustering approach can be improved by using following steps severely:

1. Do the careful analysis at the splitting level.

2. Combine all the approaches using agglomerative algorithm to the dataset into micro clusters and then evaluate macro clusters on micro clusters.

Density based method : As the partitioning clustering is only suitable for getting the spherical shaped clusters, as it does not support the arbitrary shapes. In this it is necessary to mention the number of clusters in advance. A new method is made which is purely based on the density. The basic mode of this method is to keep on increasing cluster size till the density in the neighborhood exceeds some threshold. The two main methods of density approach are DBSCAN and OPTICS. DBSCAN needs only single input parameter and support the user in defining an suitable value for it.The main reason of recognizing the cluster is that inside every cluster there are usual density of points which is very higher than the exterior one. As the density in the noise area is very much lower than others.

Grid based method : This method is using dense grid cells to form the clusters. All kind of operations relating with clustering are performed on the grid structure. In this, multi resolution grid data structure is used. A number of exciting methods are used in it like STING, Wave cluster, CLIQUE. In STING, spatial area is divided into a rectangular cells. Wave cluster applies wavelet transformation for doing cluster analysis. The major benefit of this method is optimal time in processing data which does not depend upon the number on clusters.

Model based method: It provide an environment for incorporating our knowledge about a domain. These methods suppose that data were generated from the model and we always try to build an original model from that corresponding data. The model build from the data is used to define clusters. This method supports one important algorithm called Expectation-maximization. Partitioning can be done by combining both the hierarchical and Expectation Maximization algorithm. This approach also gives very better performance than other approaches.EM also provides the results of uncertainity.Probability analysis are performed by the conceptual algorithm. Biological neural networks motivate the neural network approach. Since neural network is a set of connected inputs and outputs and each is having some kind of the weight associated with it.SOM is the most important algorithm which comes under the model based method. In this, the concept of unsupervised learning is supported where no any human intervention is required. The main aim of the SOM is to signify all data points which are in high dimensional source space into a low dimensional space.SOM manages the data in the form of clusters or cells of map in such a way that objects are highly similar in same cells and disimilar in different cells.

Clustering formation of high dimensional data

As the most of the algorithms are made for handling only low dimensional data and not suitable for handling high dimensional data. The main cause of this is when there came the concept of high dimensional data, only a trivial amount of dimensions are appropriate for definite clusters. Data which is having the improper dimensions may create a lot of noise and cover the real cluster. When the dimensionality increases, it is obvious that the data will become very sparse.As the data become highly sparse, data points located at different dimensions become equally distanced. Measurement of the distances in case of cluster analysis is having great importance.When the distances become equal, it became difficult or meaningless to count the distance as the distances are equal. In order to overcome this demerit, the research chooses the concept of PCA i.e. Principal Component Analysis. With the help of this technique, transformation can be done. Research undergoes transformation by transforming the data onto minor space while conserving the original distances between the objects. CA algorithm is having a lot of advantages like using this algorithm

quality of graphics also increased. There are other methods also which can easily handle the high dimensional data like Attribute subset selection of feature subset selection. This method is mainly having the purpose of reducing dimensions by removing the redundancy and irrelevant information. In this, basically the concept of supervised learning is used. As the supervised learning is learning with a teacher, it is also known as classsification problem. The training data is used to train the model and validation data is used to compute the accuracy of the classifier. The work has used the high dimensional dataset having 1000 entries.

K-means clustering algorithm: K-means is the most popular in unsupervised algorithm. This algorithm partition the feature vector into k clusters such that within sum of square is minimized-means method is a centroid based technique. In this at the initial stage number of clusters must be known to us. K-means [4] [5] algorithm initially started the value K, an input parameter which defines the number of cluster and partition set of objects of size n objects into k clusters so that the subsequent intracluser similarity is high but intercluster similarity is low. As the similarity can be dignified by calculating the mean value of the objects in the cluster. This mean can also be viewed as a centroid of a cluster. The following are main steps in this algorithm:

Input: The number of cluster required K and D which is the dataset containing n objects.

Output: A set of k number of clusters

Process:

1. It accepts the number of clusters and groups the data into it.

2. In the second step, it generates the first K preliminary clusters where K is the number of clusters.

3. K-means clustering algorithm will find the arithmetic mean of every cluster made in the dataset. As the cluster mean is the mean of all individual record in the cluster.

4. As the dataset contain several records, K-means allocate every record to only single of the preliminary cluster.

5. K-means reassign every record in the dataset to the most alike cluster and recalculation of mean of each cluster is performed.

6. When the new cluster will be formed, K-means will reassign every record in the dataset to only single of the fresh cluster which is already built.

Self-organizing Map: Research has also used the model based approach that is self-organizing map. SOM [9] [10] [11] is purely related with the concept of neural network as neural network is the study of neurons. Neural network can be extremely fast and efficient. This is sometimes referred as Kohonen SOM .SOM can also be observed as a visualization method. SOM permits the users to envisage in a low dimensional representation space. SOM is also very much suitable for the data survey .It generates a set of prototype vectors demonstrating the dataset and conveys out a topology stabilizing projection of the prototypes. Since two level approaches are used in case of clustering where the dataset is first clustered using SOM and then the SOM is clustered. In this way the computational load will get decrease. This approach is very much beneficial when we have a noisy, irregular and that kind of data which contain outliers. Research has applied this SOM [11] [12] on the dataset just to change the high dimensional data into low dimensional data. SOM is having the feed forward structure with a single layer of neurons having corresponding rows and column. Nodes in the SOM acts like inputs. Every neuron is fully linked with the source units in corresponding input layer. Since a one dimensional map is having single row and single column. SOM follows one rule that is winner takes all neuron. This is also called feature map in which one input vector live to the low dimensional space which is formed by arranging all the neurons in the form of grid. After applying the k-means algorithm, research has applied the SOM algorithm on that because to transform the high dimensional data into low dimensional data. Following are some steps which come under the SOM algorithm:

1. Weight of each node is prepared.

2. A vector is designated at random from the set of training data and presented to the network.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

3. Each node in the network is inspected properly. The captivating node is called the best matching unit (BMU).

4. The calculation of the radius of the neighborhood of the BMU is performed.

5. Any node found within the radius of BMU is adjusted

6. Repeat step 2 for n number of iterations.

Hierarchical Algorithm: In this method [6], group of the data objects are made into tree structure that is level by level structure. In the hierarchical database the set of records are connected with each other with the help of links. As each record consists of attributes, each of which encompasses only single data value. Since Hierarchical clustering is having two major types that are agglomerative approach and other one is divisive approach. In this, splitting of the whole data is done in two manners that are top down approach and another is bottom up approach. Bottom up hierarchical clustering is called hierarchical agglomerative clustering. Top down clustering require a method for splitting the data. In our research we only have done the analysis using hierarchical agglomerative approach. This is also called as AGNES (Agglomerative nesting).Using this method each objects are placed into the clusters. Then the merging of the clusters are done step by step. Single linkage method can easily detect the clusters having the arbitrary shapes. This method is having complexity of O (n^2). HAC begins with one point cluster and merges the most similar pairs of clusters. A dendrogram is a tree structure which is used to represent the hierarchical clustering. Dendrogram is basically used to show the way of grouping the objects together. When an algorithm uses the smallest distance, d min (Ci,Cj), to measure the intercluster distance, this is at times called a nearest-neighbor clustering algorithm. If the clustering process executed by using the minimum distance between them, then it is called single linkage algorithm. After analyzing the results of k-means and SOM by applying on dataset, at the end HAC algorithm is applied and the dendrogram is build. The following are the steps followed by the algorithm:

1. Start by assigning each item to its own cluster such that if there are n objects, then there must be n number of clusters and each cluster contains at least one object.

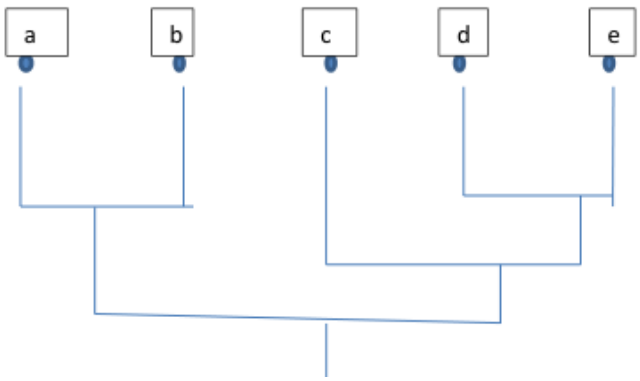
2. Find the neighboring pair of clusters and amalgamate them into a single cluster

3. Calculate the distances between the new clusters and the old clusters.

4. Repeat the step number 2 and step number 3 until we got a single cluster having size n.

The graphical representation of the cluster formation in hierarchical algorithm is called the dendrogram and it is cut at the desired level to form the desired number of clusters.

Figure 1: Dendrogram representation for hierarchical clustering



III. COMPARITIVE ANALYSIS OF K-MEANS , SOM AND HAC

Data exploration helps to get the insights of data at initial level of analysis. In this work, dataset of bank has been used for analysis. The dataset used in this work consists of various attributes such as interest on tax, qualified rebate, rate of interest, compound interest, bonus percentage, withdrawal restrictions, loan against

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

deposit, payment of return, nomination facility, maturity period, premature closure, payment rule, transferability, minimal deposit, maximum banking service etc. The dataset contain 1000 entries of data. Tanagra is used as tool for performing the comparative analysis of data by applying the K-means, Hierarchical clustering algorithm (HAC), Self-Organizing Map (SOM).Cluster analysis has been performed using the tool and error ratio for all three algorithm has been calculated.

Table 1: Table showing cluster centroid in SOM

Attributes	Cluster 1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
Qualified for rebate	1.601852	0.921739	2.300000	1.439456	1.236842	1.729565
Rate of interest	1.620370	3.878261	3.738806	3.360544	3.488722	3.773913
withdrawal restriction	2.796296	2.867391	2.679104	5.200680	2.637218	2.865218
Loan/advance against deposit	4.039815	3.244348	5.268657	4.352381	5.936090	0.926087
Minimal deposit	2.796296	2.867391	2.679104	5.200680	2.633459	2.865217

Table 1 shows the centroid formation in SOM algorithm. The various centroids are formed during the course of SOM algorithm.

Table 2: R-Square for each axis in K-MEANS

Trails	R-Square
1	0.646422
2	0.581973
3	0.659957
4	0.638804
5	0.648223

R-square values generated during the execution of K-means algorithm

Table 3: Analysis of K-Means algorithm

Cluster	Description	Size	WSS(within sum of square)
Cluster1	c_kmeans_1	176	260.7051
Cluster2	c_kmeans_2	226	240.0426
Cluster3	c_kmeans_3	206	236.5840
Cluster4	c_kmeans_4	134	173.0984
Cluster5	c_kmeans_5	110	293.6515
Cluster6	c_kmeans_6	148	496.1309

Sum of square error has been computed for clusters using K-Means algorithm.

Table 4: Analysis of HAC algorithm

Clusters	BSS ratio	Gap
1	0.0000	0.0000
2	0.3250	1.0335
3	0.4434	0.1971
4	0.5223	0.0041
5	0.6003	0.1945
6	0.6395	0.0591

Analysis of HAC algorithm has been performed and BSS ratio and Gap are calculated for the clusters.

Table 5: Analysis of HAC, SOM and K-MEANS

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Algorithms	Error ratio	Number of cluster
SOM	0.6382	6
K-MEANS	0.6600	6
HAC	0.6395(BSS)	6

Table 5 show the error ratio of HAC, SOM and K-MEANS algorithm. Total numbers of cluster made in the execution are six.

III. CONCLUSION

In this work, the comparative study has been performed on large amount of high dimensional data. Comparison has been done on three clustering algorithms that are SOM, HAC and K-means algorithms. These algorithms are applied on the dataset of banking which contain high dimensional data. Analysis of each algorithm has been done and after applying the K-Means, SOM algorithm and HAC algorithm, it is found that error ratio of SOM is 0.6382 which is comparatively less as compared to K-means and HAC algorithm.

REFERENCES

[1] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” in KDD Work-shop on Text Mining, 2000.
[2] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Inc., UpperSaddleRiver, NJ,USA,1988
[3] Anil K Jain, "Data clustering: 50 years beyond k-means", Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010.
[4] A Hassan, M Kingravi, "Emre Celebi and Patricio A Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm", Expert Systems with Applications, vol. 40, no. 1, pp. 200-210, 2013.
[5] José Manuel Pena, Jose Antonio Lozano, Pedro Larranaga, "An empirical comparison of four initialization methods for the k-means algorithm", Pattern Recognition Letters, vol. 20, no. 10, pp. 1027-1040, 1999.
[6] Y. Zhao, G. Karypis, "Hierarchical clustering algorithms for document datasets" in Data Mining and Knowledge Discovery, Springer Science + Business Media, Inc, vol. 10, pp. 141-142, 2005.
[7] O. Maimon, L. Rokach, The Data Mining and Knowledge Discovery Handbook, Springer Science+Business Media, Inc, pp. 321-340, 2005.
[8] E. Alpydin, Introduction to Machinel Learning, The MIT Press, pp. 143-158, 2010.
[9] T. Kohonen, "The self-organizing map", Proceedings of the IEEE, vol. 78, no. 9, 1990 Sep
[10] Günter and Bunke, 2002, Self-organizing map for clustering in the graph domain, Pattern Recognition Lett. 23 (2002), pp. 401-417.
[11] Kohonen, T., Somervuo, P., 2002.How to make large self-organizing maps for nonvectorial data. Neural Networks 15 (2002) 945-952
[12] A. K. Mann, N. Kaur, "Survey paper on clustering techniques", International Journal of Science Engineering and Technology Research, vol. 2, no. 4, pp. 0803, 2013.
[13] L.Parsons, E.Haque, H.Liu. "Subspace Clustering for High Dimensional Data: A Review", ACM SIGKDD Explorations Newsletter, 2004, 6(1): pp 90-105.