Performance Comparison of Clustering Algorithm On Banking Dataset

Manish Pathak
Student
SCOE, Kharghar
Mumbai University
Manishp930@gmail.com

Jatinder Saini
Student
SCOE,Kharghar
Mumbai University
jsnghsaini1897@gmail.com

Siddhesh Shinde
Student
SCOE,Kharghar
Mumbai University
sid.shinde14@gmail.com

Deepa Parasar
SCOE,Kharghar
deepaparasar@gmail.com

Abstract

Clustering is a technique used in data mining which is used to set data elements into their interrelated groups with no advancement of knowledge regarding grouping of definitions. It is not a particular algorithm but a common task is being solved. There will be three cluster in one algorithm. This cluster will be formed on the basis of the amount of transaction.It evaluates the performance of clustering algorithm based on accuracy, time efficiency and error rates. Analysing this data the additional benefit services can be given to the customer. Clustering is a process of grouping a set of similar data objects within the same group based on similarity criteria (i.e. based on a set of attributes). There are many clustering algorithms. The proposed system shows the comparative analysis of three clustering algorithms namely K-means algorithm, Farthest first algorithm and Density based algorithm. These algorithms are compared in terms of efficiency and accuracy. The data for clustering is used in normalized and as well as un-normalized format. In terms of efficiency and accuracy K-means produces better results as compared to other algorithms. The K-mean algorithm specifies 90% of time accuracy compare to other two algorithm.

Clustering is the process of grouping of data, where the grouping is done by finding similarities between data based on their features or characteristics. Such groups are termed as Clusters. A study of comparison of clustering algorithms across banking customer is performed here. The performance of the various clustering algorithms is compared based on the time taken to form the desired clusters. The experimental results of various clustering algorithms to form clusters are represented as a graph. Identifying customers by a customer behaviour analysis model is helpful characteristics of customer and facilitates marketing strategy development.

Keywords— Datamining, Data processing, Simple k means clustering, Efficient k means clustering, Farthest firs1t clustering, Make density bsed clustering.

I. Introduction

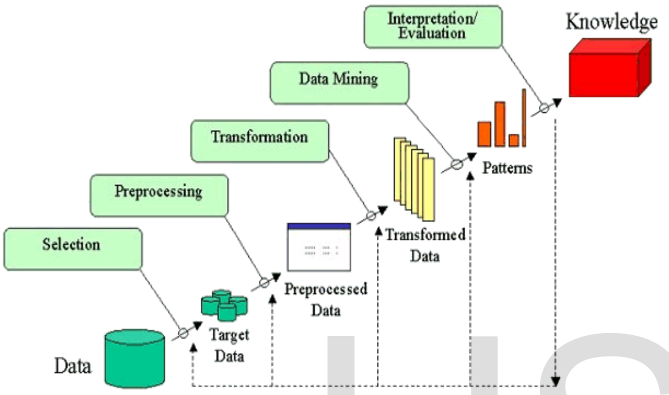
The important resource in contemporary marketing strategies is Customers. Therefore, it is useful for enterprises and organization to successfully acquire new customers and maintain high value customers. To gain these goals, many organization plan to achieve their own customers' data with many of database tools which can be analysed to achieve the customer behavioural and applied to develop new business strategy. Economic theory has discovered that a business derives80% of its income from 20% of its customers. However, organization select only those individuals that meet specified profitability levels where based on previous behaviour or individual needs and assuming there is same pattern between customer behaviours.

Many method have been introduced to achieve better knowing of customer behaviours, the "behavioural scoring models" is one of the most successful technique that help decision makers to realize their customer behaviours. Behavioural scoring models help to analyse purchasing behaviour of customers [1]. These models are based on data mining approaches. For a bank, most existing data mining approaches were discovered rules [2] and predicted bankruptcy probability [3] in a bank database. Few works have focused on the analysing of bank databases from the viewpoint of customer behavioural analyse [4]. More specifically we wanted to look at both the customers profile data and their debit cards transactions. With these data, the aim was to discover applied patterns or rules in the data that could provide information about what incentives a company could offer as better marketing strategies to its customers.

II. LITERATURE SURVEY

A. DATA MINING

Data mining (the analysis step of the "knowledge discovery in databases" process, or kdd), is a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an easy structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing model and inferenceconsiderations,interestingness metrics, complexity, considerations, post-processing of discovered structures, visualization, and online updating.



B. CONSUMER BEHAVIOR

Consumer behavior is very complex phenomenon, which is considered primarily in marketing decisions. It has been rightly said "understand, you do not understand, you will not understand, you cannot understand all your customers but still you have to do your best to understand them."

In consumer behavior this is very difficult to make a uniform theory that may suggest that a particular individual or group will behave in a particular manner. Consumer behavior is dynamic and to be studied regularly. Increasing awareness, living standards and urbanization has led to increase in the changing preferences and the same has forced the marketers to change their product features, packaging styles, distribution channels and so on. There is a famous saying the "success has a simple formula-do your best and people must like it". Similarly, for marketers the advice is- offer the best and customers must like it'. Identical products always have their life cycle the product life cycle suggests that there is a level of maturity of the product and after that no more consumers can be attracted for that. The case is very same with preferences of consumers that they always like some innovative and different products to use. The study of consumer behavior is compulsory to know about the basic requirement of consumers from time to time so that the products and services can be offered accordingly. Customers have their own unique needs, demands and preferences in a particular segment. Marketers have to study customers in particular required area. Really interesting it is, the study of consumer behavior can make it possible that after observing and examining the behavior of consumer a marketer can present his product in such a way that the product can capture the market. However it was very difficult to sell that product earlier. Consumer behavior indeed gives every possible answer to the complex questions concerned with consumer's buying reasons.

Analyzing the Consumer Behavior

Banks seeking newer and better ways to differentiate themselves from their competitors, customer clustering one of important way to rich this result; Customer clustering is the use of past transaction data to divide customer to the similar groups. The results produced are based on the assumptions that the customer behavior follows patterns similar to past pattern and repeats in the future. Therefore, there could not be a better time than now to analyze the importance of an effective new marketing strategy using the customer behavior analyze. The decisions to be made include which target groups of customers will be encouraged to use more, what terminal type to assign, how estimated probability of acceptance new products, whether to promote new products to target groups of customers, and, how to manage groups of customers to rich the customer satisfaction and direct marketing. However, attempts to make good customer behavior analysis may be limited by the poor quality of data, poor relevant of data, or the volume of data needing to be processed. Database marketing (DM) is a systematic approach to the gathering, consolidation, and processing of customer data to help the marketers' better target their markets efforts to existing customers [5].Additional DM analyzes customer data to look for patterns to use these patterns for a more targeted selection of the consumers [6].Over the decades, many database marketing tools were developed and used in various stages of marketing. Some of the most popular tools include: the RFM (recency, frequency, monetary), Formula, the behavior segment of the existing customers and the lifetime value of a customer [7]. Customer clustering is a process that divides customers into smaller groups; Clusters are to be homogeneous within and desirably heterogeneous in between [8].

As shown in the following equation, this study employs " Customer Power" (CP) as a customer behavior variable to model Customer behavior,

$$\text{Customer Power} = \frac{\text{Sum of the customer score on a term}}{\text{Number of months on a term}} \quad \ldots(1)$$

The default observation range is assumed to be 12 months, and CP is computed as the 'sum of the customer score on a term' divided by the ‘number of the month on a term’. Score of each customer in a month calculated by flowing table.

Tab 1. Table of customer score

Sum of Transaction Amount at month (STAM)	Customer Score (CS)
0< STAM <2000	1
2000<= STAM <7000	2
700<=STAM <12000	3
12000<=STAM	4

As shown in the upper table each customer score (CS) calculated:

1. Sum of Transaction Amount on a Month (STAM)
2. Score elicit according to table

For example, a customer has STAM=9000 as Customer Score (CS) elicit according to table as 3. For instance, a customer has (CS) during 12 month as (3,2,2,3,2,2,2,3,2,1,2,2) and then the degree of CP is computed as:

$$CP = \frac{3+2+2+3+2+2+2+3+2+1+2+2}{12} = 2.166$$

For each customer, if CP is between 3 & 4, then the behavior of that customer is considered a pamper user. Meanwhile, if CP is between 2 & 3 then the behavior of that customer is considered a transactor user. Finally, if the value of CP is between zero and one then the behavior of that customer is considered a raring user.

C. Existing System

The Current system used for analysis is proved to be beneficial but requires some more upgrading as it involves limited analysis of clean data**.**

D. Proposed System

Clustering is mainly needed to organise the results provided by a search engine. Clustering can also be viewed as a special type of classification. The clusters formed as a result of clustering can be defined as a set of like elements. But the elements from different clusters are not alike. Clustering is similar to database segmentation, where like tuples in a database are grouped together. When clustering is applied to a real world database, many problems occur there such as: handling outlier is difficult; interpreting the semantics of each cluster is difficult, no correct answer for a clustering problem and what data should be used for clustering.

The problem of clustering can also be defined as below: Given a collection of data objects, the work of clustering is to divide the data objects into groups such that objects in the same group are similar. Objects in different groups should be dissimilar. Data belonging to one cluster are the most similar; and data belonging to different clusters are the most dissimilar. Clustering algorithms can be viewed as hierarchical and partitional. With hierarchical clustering, a nested set of clusters is created. The hierarchy is divided into various levels. In the lowest level, each item will have its own cluster. In the highest level, all the items will be belonging to a single cluster. With partitional clustering, only one set of cluster is created. Hierarchical clustering is represented using a tree structure called dendrogram. Examples of hierarchical clustering algorithms are agglomerative and divisive clustering algorithms. Examples of partitional clustering algorithms are K-Means, nearest neighbor and PAM. Clustering can be done on large databases also. Most popular clustering algorithms like BIRCH (balanced iterative reducing and clustering using hierarchies) [9], DBSCAN (density based spatial clustering of applications with noise) and CURE (Clustering using representatives) [10]. Clustering can also be performed with categorical attributes. Optimization based partitioning algorithms are represented by its prototype.

Objects of similar prototype are clustered together. An iterative control strategy is used to optimize the clustering. If the clusters are of convex shape, same size, density, and if their number *k* can be reasonably estimated, then the clustering algorithm can be selected correctly. *K*-means, *k*-modes and *k*-medoid algorithms can be differentiated based on their prototypes. The *k*-means method has been shown to be effective in producing good clustering results for many practical applications. However, the *k*-means algorithm requires time proportional to the product of number of patterns and number of clusters per iteration. This computationally may be expensive especially for large datasets. Among clustering formulations that are based on minimizing a formal objective function, perhaps the most widely used and studied is *k*-means clustering. Given a set of *n* data points in real *d*-dimensional space, Rd, and an integer *k*, the problem is to determine a set of *k* points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest centre. A comparative study between various clustering algorithms based on the time taken to form the clusters is considered. The various clustering algorithms taken into consideration are simple K-Means, overlapped K-Means, enhanced K-Means are compared against filtered clusterer, make density based clusterer and farthest first clustering.

CLUSTERING ALGORITHMS

A. *Simple K-Means Clustering*

K-Means is an iterative clustering algorithm [11], [12] in which items are moved among set of clusters until the desired set is reached. This can be viewed as a type of squared error algorithm. The cluster mean of Ki= {ti1,ti2,….. ,tim} is defined as,

$$m_i = \frac{1}{m}\sum_{j=1}^{m} \text{tij} \qquad \ldots\ldots (2)$$

***Algorithm 1**: K-Means Clustering*

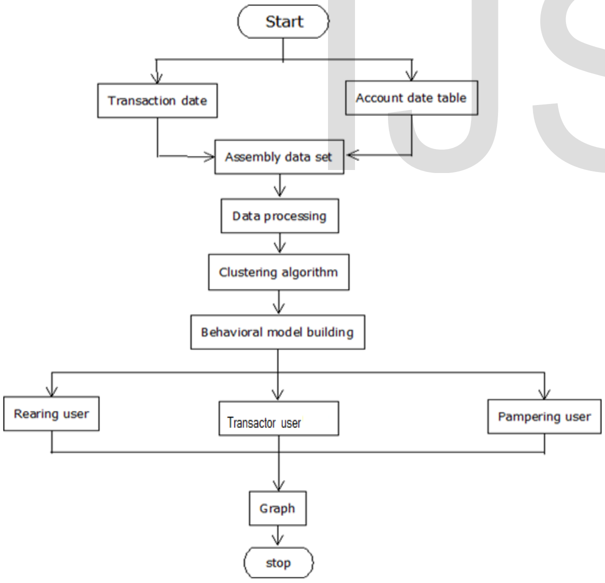
Input: D= {a1, t2,….., tm} // set of elements.
K // number of desired clusters.
Output: K // set of clusters.
Procedure:

```
Assign initial values for means a1, a2 …., .ak;
Repeat
Assign each item ai to the cluster which has the closest mean;
Calculate new mean for each cluster; until convergence criteria is met;
```

In almost all cases, the simple K-Means clustering algorithm [13] takes more time to form clusters. So it is not suitable to be employed for large datasets.

B. *Efficient K-Means Clustering*

In each iteration, the *k*-means algorithm computes the distances between data point and all centers; this is computationally very expensive especially for huge datasets. For each data point, the distance can be kept to the nearest cluster. At the next iteration, compute the distance to the previous nearest cluster. By comparing the old distance with new distance, and if it is less than or equal, then the point will be in the same cluster. This saves the time required to compute distances to *k*−1 cluster centers. Two functions are written to implement efficient K-Means clustering algorithm [14].[15] The first is the simple K-Means, which calculates the nearest point of center. This is done by computing the distances to all centers. Each data point keeps its distance to the nearest center.



***Algorithm 2**: Efficient K-Means Clustering*

```
Function distance ()
// each point is assigned to the cluster nearby
For i=1 to b
For j=1 to k
Compute squared Euclidean distance d2(ri,aj);
endfor
Find the closest centroid aj to ri;
aj=aj+ri; bj=bj+1;
MSE=MSE+d2(ri, aj);
Clusterid[i]=number of the closest centroid;
endfor
For j=1 to k
aj=aj/bj;
endfor
```

```
Function distance_new ()
// each point is assigned to the cluster nearby
For i=1 to b
 Compute squared Euclidean distance
d2(ri, Clusterid[i]);
If (d2(ri, Clusterid[i])<=Pointdis[i])
Point stay in its cluster;
Else
For j=1 to k
Compute squared Euclidean distance d2(ri,aj);
endfor
Find the closest centroid aj to ri;
aj=aj+ri; bj=bj+1;
MSE=MSE+d2(ri, aj);
Clustered[i]=number of the closest centroid;
Pointdis[i]=Euclidean distance to closest centroid;
endfor
For j=1 to k
aj=aj/bj;
endfor
```

C. Farthest First Clustering

Farthest first is a variant of K Means. This places the cluster center at the point further from the present cluster. This point must lie within the data area. The points that are farther are clustered together first. This feature of farthest first clustering algorithm speeds up the clustering process in many situations like less reassignment and adjustment is needed.

D. Make Density Based Clustering

A cluster is a dense region of points that is separated by low density regions from the tightly dense regions. This clustering algorithm can be used when the clusters are irregular. The make density based clustering algorithm can also be used in noise and when outliers are encountered. The points with same density and present within the same area will be connected to form clusters.

***Algorithm 3**: Density based Clustering*

1. Compute the ε-neighborhood for all objects in the data space.

2. Select a core object CO.

3. For all objects co Ɛ CO, add those objects y to CO which are density connected with co. Proceed until no further y are encountered.

4. Repeat steps 2 and 3 until all core objects have been processed.

III. CONCLUSION

A comparative study of clustering algorithms across two different data items is performed here.The main conclusion of this paper is to make a comparative performance analysis of Simple K-means, DBSCAN, Efficient K-Means Clustering.It is important to remember that cluster analysis is an exploratory tool. While hundreds of clustering algorithms are available and new ones continue to appear, we compare only four of them. All the algorithms have some ambiguity in some (noisy) data when clustered. Simple K-means make clusters with minimum amount of time DBSCAN is not suitable for data with high variance in density. In terms of time complexity and dataset used, K-means produces better results in comparison to all explained algorithms.

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.

Reference

[1] Setiono, R., Thong, J.Y.L., Yap, C.S., “Symbolic Rule Extraction from Neural Networks-an Application to Identifying Organizations Adopting IT”. Information and Management, 34(2), 1998, pp 91–101.

[2] Au, W.H,Chan, K.C.C., Mining Fuzzy Association rules in a Bank-Account Database, IEEE Transactions on Fuzzy systems 2003, Vol. 11.

[3] Donato, J.C., Schryver, G.C., Hinkel, R.L., Schmoyer,J., Leuze, M.R., Grandy, N.W., Mining Multi-Dimensional Data for Decision Support, Future Generation Computer Systems, Vol. 15, 1999, pp.433-441.

[4] Sharda, R., Wilson, R., Neural Network Experiments in Business Failures Predication: a Review of Predictive Performance Issues. International Journal of Computational Intelligence and Organizations, 1(2), 1996, pp 107–117.

[5] Tao, Y.H. Yeh, C.C.R., Simple Database Marketing Tools in Customer Analysis and Retention, International Journal of Information Management, Vol. 23, 2003,pp.291-301.

[6] Qizhong, Zhang, An Approach to Rough Set Decomposition of Incomplete Information Systems. 2nd IEEE Conference on Industrial Electronics and Applications, ICIEA, 2007, pp. 2455-2460.

[7] Feelders, A.J., Credit Scoring and Reject Inference with Mixture Models, International Journal of Intelligent Systems in Accounting, Finance and Management, Vol. 9, 2000, pp.1-8.

[8] Anil, K.J., Data Clustering: 50 Years Beyond K-Means, International Journal of Pattern Recognition Letters 2009.

[9] Zhang, T., Ramakrishnan, R., Linvy, M., 1996. BIRCH:”An Efficient Data Clustering Method for Very Large databases”. Proc. ACM SIGMOD Int. Conf. on Management of Data. ACM Press, New York, p.103-114.

[10] Guha, S., Rastogi, R., Shim, K., 1998. CURE: An Efficient Clustering Algorithms for Large Databases. Proc. ACM SIGMOD Int. Conf. on Management of Data. Seattle, WA, p.73-84.

[11] Jain, A.K., Dubes, R.C., 1988. “Algorithms for Clustering Data”. Prentice-Hall Inc.

[12] Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li, Automated Variable Weighting in k-Means Type Clustering”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, VOL. 27, NO. 5, PP. 657-668, 2005.

[13] Shi Na, L. Xumin, G. Yong, “Research on K-Means clustering algorithm-An Improved K-Means Clustering Algorithm”, “IEEE Third International Symposium on Intelligent Information Technology and Security Informatics”, pp.63-67, Apr.2010.

[14] 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. 881-892, 2002.

[15] Fahim AM, Salem AM, Torkey FA, Ramadan MA (2006) An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University SCIENCE A7:1626–1633. Available online at www.zju.edu.cn/jzus.