International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 4, April - 2013

K-Mean Clustering Algorithm Implemented To E-Banking

Kanika Bansal
Banasthali University

Anjali Bohra
Banasthali University

Abstract

As the nations are connected to each other, so is the banking sector. Due to this globalization, the banks struggle to gain an edge over others. To overcome this, banks has rapidly increased the ability to generate, manipulate and storage of data of the customers because the information contained in this data can be highly useful. The E-banking industry around the globe has undergone a remarkable change in the way business is conducted. The banking industry has realized the need of the techniques like data mining, web mining, etc which can help them to compete in the market and to make the prediction and decision for the industry for growth. Banks can use data mining tools like clustering on customer's profile. This paper provides an overview of the concept of K-mean clustering algorithm and highlights the applications of e-banking to increase the customer's satisfaction

Keywords K-mean algorithm, Clustering, Data mining

I. Introduction

Technology and innovation changes the world. The E-banking increases the competition edge in the market and increases the tendency of how to do business in the competitive market. Now-a-days, the bank customer can perform many tasks like viewing account balance, doing transaction, downloading bank statements, ordering cheque books, download periodic account statements and many more tasks. The bank customer can perform this multiple task by logging into the financial institution's website. A profile of bank customer is created by banks by verifying their username and password given to them and manipulated by customer.

Techniques of data mining bring abundant changes in the banking sector. Currently, electronic data are being analyzed, stored, and manipulated by banks. The behemoth size of these data make impossible for the bank's employee to come up with useful information which help in the decision making process for the satisfaction of bank's customer. The data mining techniques enhance the performance by knowing the details of customer profile. As the growth of customer's increases day by day, it is very important to analyze their profile. The profile can also analyzed by seeing their occurring date, occurring time, transaction status, transaction time and many more attributes.

II. Data Mining

There is a need for a new generation of computational theories and tools to help humans, software industry in extracting useful information knowledge from the rapidly growing very large volume of digital data. The KDD process is interactive and iterative, involving numerous steps with many decisions

made by the user. Brachman and Anand (1996) give a practical view of the KDD process, emphasizing the interactive nature of the process. [1] KDD refers to the overall process of discovering useful knowledge from data.[2] The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages: (1) Selection (2) Pre-processing (3) Transformation (4) Data Mining (5) Interpretation/Evaluation. [4]

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns over the data. The term data mining is used by statisticians, data analysts, the management information system communities, and software companies. It has also gained popularity in the database. [1]

Data mining has two primary objectives: prediction and description. Prediction involves using some variables in data sets in order to predict unknown values of other relevant variables (e.g. classification, regression, and anomaly detection). Description involves finding human understandable patterns and trends in the data (e.g. clustering, association rule learning, and summarization) [3].

Fayyad et.al. (1996) define six main functions of data mining:

1. Classification is finding models that analyze and classify a data item into several predefined classes.

2. Regression is mapping a data item to a real-valued prediction variable.

3. Clustering is identifying a finite set of categories or clusters to describe the data.

4. Dependency Modeling (Association Rule Learning) is finding a model which describes significant dependencies between variables.

5. Deviation Detection (Anomaly Detection) is discovering the most significant changes in the data.

6. Summarization is finding a compact description for a subset of data.

III Clustering

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). [5] It is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. [**6]**

3.1Cluster analysis is classified as follows:

Distance Based Clustering - Method according to the distance between data. It is sensitive to noise data and isolated. [7]

Density Based Clustering - These clustering deals with connecting region with the same density. Therefore the density clustering needs the scanning of the entire data set and data will be divided into different small squares, which is approximately express as clusters. [7]

Link Based clustering - This analysis puts the clustering object into a graph model. Clustering web usage data is different from the traditional clustering due to the data. Therefore, there is a need to develop specialized techniques for clustering analysis based on Web usage data. [7]

To better understand the difficulty of deciding what constitutes a cluster, consider figures 1 (a) through 1 (b), which show fifteen points and three different ways that they can be divided into clusters. If we allow clusters to be

nested, then the most reasonable interpretation of the structure of these points is that there are two clusters, each of which has three sub clusters. However, the division of the two larger clusters into three sub clusters may simply be an artifact of the human visual system. Finally, it may not be unreasonable to say that the points form four clusters. Thus, we stress once again that the definition of what constitutes a cluster is imprecise, and the best definition depends on the type of data and the desired results. [8]

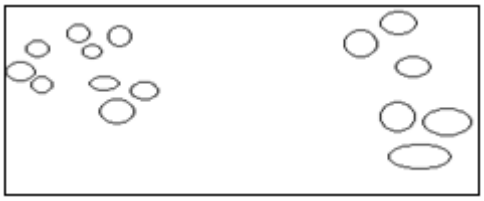

Figure 1 (a)

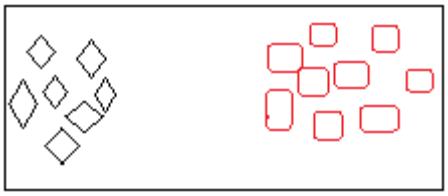

Figure 1 (b)

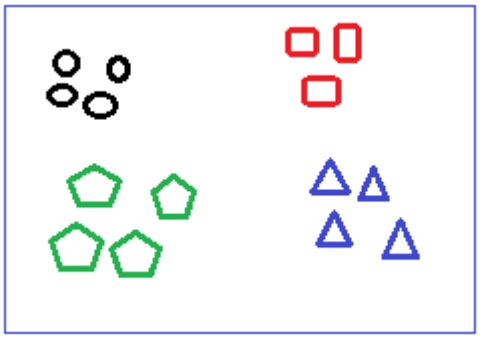

Figure 1 (c)

Categorization of major Clustering methods

1) Partitioning Method

To achieve global optimality in partitioning-based clustering would require the exhaustive enumeration of all of the possible partitions. Instead, most applications adopt one of two popular heuristic methods:

1. k-means algorithm - where each cluster is represented by the mean value of the objects in the cluster.

2. k-medoids algorithm - where each cluster is represented by one of the objects located near the center of the cluster.

To find clusters with complex shapes and for clustering very large data sets, partitioning-based methods need to be extended. [9]

2) Hierarchical Method

A hierarchical method creates a hierarchical decomposition of the given set of data objects. This method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach and the divisive approach called the top-down approach. [9]

3) Density Based Method

Their general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold; that is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise (outliers) and discover clusters of arbitrary shape. Examples are DBSCAN and OPTICS. [9]

4) Grid Based Methods

Grid based methods quantize the object space into a finite number of cells that form a grid structure. [9]

5) Model Based Model

Model Based Model is to hypothesize a model for each of the clusters and find the best fit of the data to the given model.[9]

III. Proposed Method

Choosing a K-mean algorithm for the system as it has two advantage of using this technique:

- With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).
- K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular. [5]

K-means algorithm was first introduced by Llyod and MacQueen for partitioning methods. It is the simplest clustering algorithm and widely used. K-means requires an input which is a predefined number of clusters. This input is named k.

The steps of the K-means algorithm are given below.

1. Select randomly k points to be seeds for the centroids of k clusters.

2. Assign each point to the centroids closest to the point.

3. After all points have been assigned, recalculate new centroids of each cluster.

4. Repeat step 2 and step 3 until the centroids no longer move

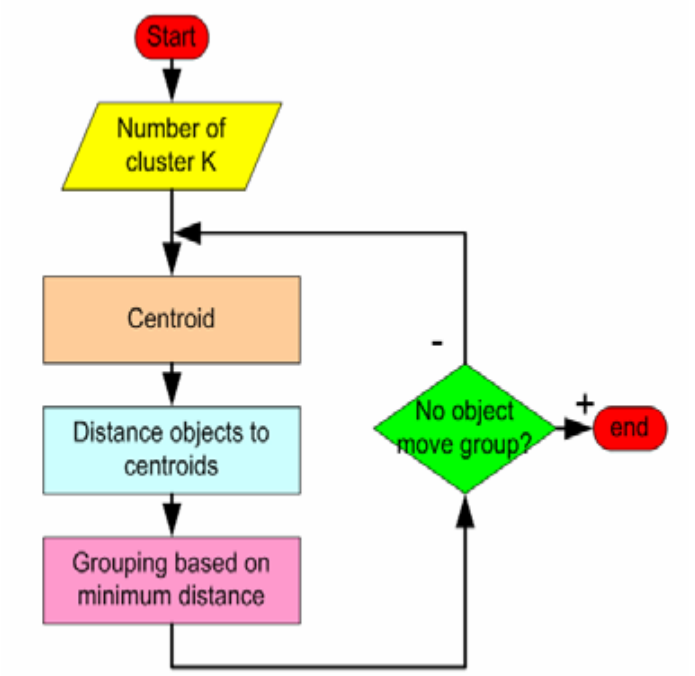

Figure 2[6]

A systematic method was used to collect our own representative sample data of similar to internet banking transactions of private bank. The representative sample data is a dummy data created similarly as internet banking transactions from web log file of bank customer profile. Each session of when customer login into the system give details of web usage including customer accounts, requested web pages and their transaction order, and the period of time they operate and the pages they viewed. This data were used as the basis for analysis in this study. We are choosing four attributes in our system: (1) Operating Date, (2) Operating Time (3) Transaction time (4) Transaction Status.

Here is the sample data:

1. Date was divided in 2 groups:

Date 1 - Between 1st to 15th of month

Date 2 - Between 16th to the last day of the month

2. Time was divided in 4 groups:

Interval 1 – Between 00.00 hrs to 06:00 hrs

Interval 2 - Between 06.00 hrs to 12:00 hrs

Interval 3 - Between 12.00 hrs to 18:00 hrs

Interval 4 - Between 18.00 hrs to 00:00 hrs

3. Transaction Status of was divided in 2 groups:

Status1 - Real-Time Transaction

Status2 - Schedule Transaction

4. Transactions Kind of were divided in 4 groups:

Type1 - Balance Inquiry

Type2 - Statement and Bill Payment Report

Type3 - Money Transfer

Type4 - Payment Transaction

These sample data is implemented on the K-mean algorithms with the help of the Windows forms (Visual Studio 2012), SQL Server 2012 and C# is used as a programming language.

IV EXPERIMENTAL RESULTS

As discussed, the project is implemented using Windows forms (Visual Studio 2012), SQL Server 2012 and C# is used as language.

I. Tables design and architecture

The design of the tables has been in such manner that it can filter data very easily and quickly. Figure 3 shows the design and relationship between tables.

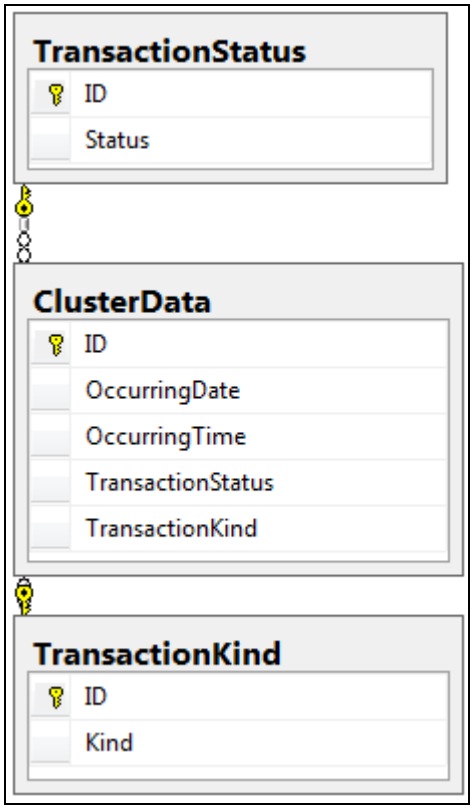

Figure 3

The above figure shows three tables. First table (TransactionStatus) shows an auto-generated unique id and the name of the transaction status. Second table (TransactionKind) shows an auto-generated unique id and the name of the transaction kind. Third and final table also contains an auto-generated unique id along with date, time, and foreign /referenced keys for Transaction Status and Transaction Kind.

II.Windows Forms

The application is programmed using C# and Windows forms. Windows form is used so that a desktop application can be developed and an analyst can use it without having an internet connection.

On loading of the application, a database connection is made to get all the data from the database. All data is extracted, so that various charts can be made quickly.

Figure 4 shows the screenshot of the application.

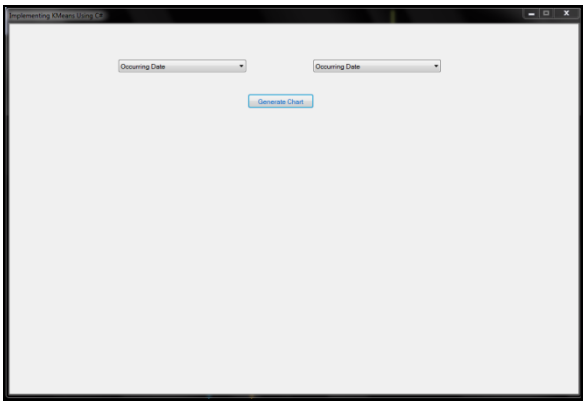

Figure 4

The screenshot shows two dropdowns and one button. Each dropdown has the following values (Figure 5).

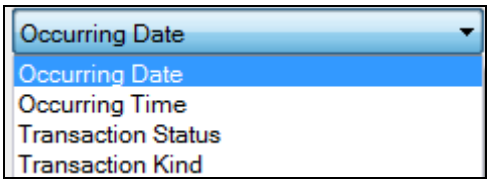

Figure 5

To generate a chart, different values had to be selected in the dropdowns. If same value will be selected, a message box will be opened (Figure 6) and no chart will be displayed.

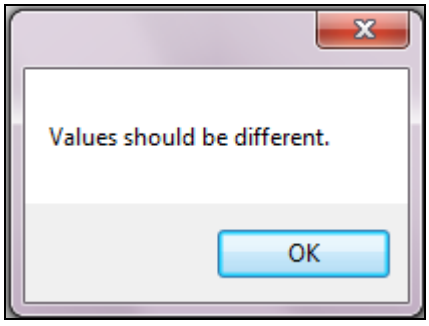

Figure 6

On selecting different parameters in the dropdowns, a chart will be displayed. This chart will contain all the points entered in the database along with 4 clustered points specific to the shown chart. The points will be displayed in square colored box and the clustered point will be displayed as bubbles. The screenshot of the chart looks similar to figure 7.

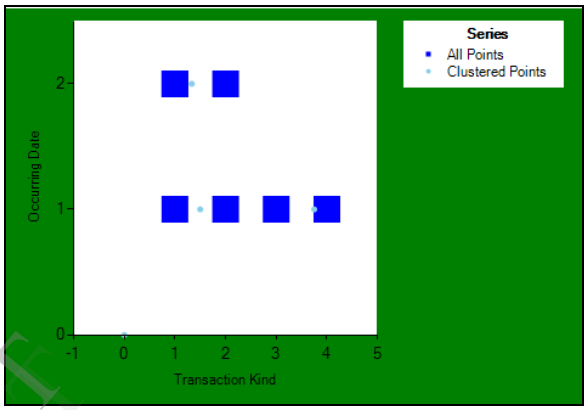

Figure 7

For ease of the analyst, the x and y axis of the 4 clustered points is also displayed on the right of the chart. The below screenshot will show the point at which the clustered point is formed and the total number of points contained in each cluster points (Figure 8).

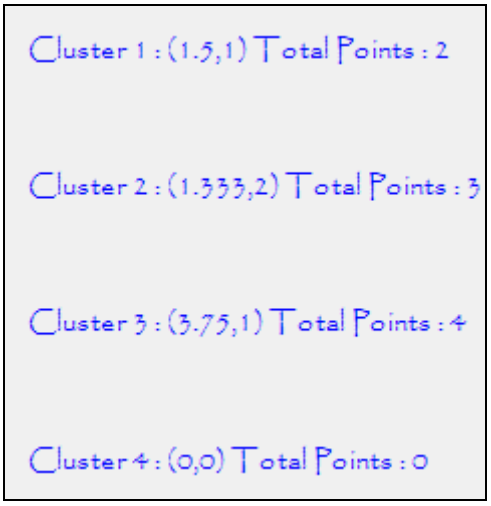

Figure 8

The complete layout of the application after generating chart will look similar to figure 9.

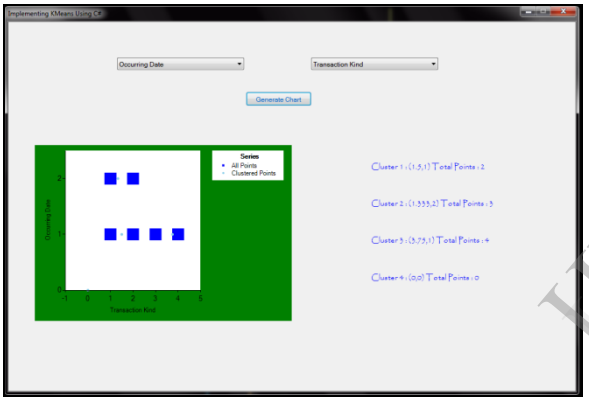

Figure 9

V. CONCLUSION

This paper focuses on clustering with the help of K-Mean algorithm, applied to internet banking dummy database of customer to analyze customer characteristics and behaviors with appropriated criteria of operating time, operating date, transaction status, transaction time. The benefits are valuable for all the banks to improve services of e-banking. This application and concept can be used for other online e-commerce websites too.

VI. REFERENCES

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, 1996, From Data Mining to Knowledge Discovery in Databases.

[2] Tipawan Silwattananusarn, Dr. KulthidaTuamsuk, 2012, Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012.

[3]Gorunescu, F., 2011, Data Mining: Concepts, Models, and Techniques. India: Springer.

[4] http://en.wikipedia.org/wiki/Data_mining.

[5] A. K. Jain, M. N. Murty. J. Flynn, 1999, Data clustering: a review

[6] Sapna Jain, M Afshar Aalam, M. N Doja, 2010, K-means clustering using weka interface.

[7] V.chitraa, Dr.antony selvadoss thanamani, 2012, An Enhanced Clustering Technique for Web Usage Mining.

[8] Narendra Sharma 1, Aman Bajpai 2, Mr. Ratnesh Litoriya, 2012, comparison the various clustering algorithms of weka tools.

[9]Jiawer Han, Micheline Kamber; Data Mining Concepts and Techniques; Morgan Kaufmann Publishers, Urbana-Champaign