An Approach to CRM: Usage Behavior Analysis with K-Means Algorithm

May Thu Kyaw, Mar Mar Lwin
Computer University (Maubin), Myanmar
maythukyawmoon@gmail.com, dawmarmarlwin@gmail.com

Abstract

Nowadays, information and communication technologies are developing with great speed. In the advent of information era, information technology has developed rapidly and has become significant for every business, particularly the credit card business. The purpose of this paper is to classify the selected customers into clusters using k-Means Algorithms and then decide with the concept of customer relationship management (CRM) to identify high profit, gold customers. The result of this study will benefit the organization in exploiting it to bring about competitive advantage and being able to retain customers as well as attract potential ones. The objective of this paper is to help the managers to segment the customers and identify which customers to target for credit card business services by using behavior analysis. Consequently, the results show that there is a clear distinction between the segments in terms of customer behavior.

Keywords: Data Mining, Customer Relationship Management, k-means algorithm, Credit Card.

1. Introduction

For companies with large numbers of customers, Data Mining and Market Research are often employed to gain intelligence into customer behavior and attitudes respectively. In this paper, we will examine the issues surrounding the convergence of Data Mining for deeper customer understanding [3]. Data Mining is a process that uses a variety of data analysis and modeling techniques to discover patterns and relationships in data that may be used to make accurate predictions. It can help to select the right prospects on whom to focus, offer the right additional products to your existing customers, and identify good customers who may be about to leave you. The result is improved revenue because of a greatly improved ability to respond to each individual contact in the best way, and reduced costs due to properly allocating your resources. Customer relationship management, through which, banks hope to identify the preference of different customer groups, products and services tailored to their liking to enhance the cohesion between credit card customers and the bank, has become a topic of great interest. The systemic application of data mining techniques reinforces the knowledge management processes and allows marketing personnel to know their customers well to provide better services.

In this paper, Section 2 gives the background theory and related work. And then, the overview of the system is described in Section 3. Section 4 presents the implementation of the system. Finally, Section 5 concludes the paper.

2. Background and Related work

There are a few partitioned methods that specifically address the issue of clustering spherically constrained data [5][3][6]. Banerjee et al. and Dortet-Bernadet propose different mixture models, and employ an expectation-maximization (EM) approach. There is also a version of k-means proposed by Dhillon, I. S., and Modha, D. S. using cosine similarity. Called *spkmeans*, this algorithm replaces the Euclidean distance metric in the base k-means algorithm by cosine similarity.

The *spkmeans* algorithm does not inherit the properties of Hartigan's efficient implementation of the k-means algorithm, and can be slow, potentially performing poorly in many datasets. Another difficult issue, unaddressed in Dhillon and Modha or in Banerjee and Dortet-Barnadet, is in determining the optimal number of clusters (k) in a dataset. Dortet-Barnadet and Wicker study some methods [1][2][8][9] in this context, but their recommendation of Akaike's AIC is hardly convincing, given the known tendency of AIC to overestimate the number of clusters by this criterion.

In this paper, we employ the principle of data mining to cluster customer segments by using k-means algorithm and effectively discover the current spending pattern of customers and provide as early as possible services desired by the customer to expand the clientele base and prevent customer attrition. Our algorithm is general enough to cluster the credit card customer and help to improve the decision of business management. This section describes the background of our system. And section 2.1 describes definition of Customer Relationship Management (CRM). About clustering and various clustering methods are described in Section 2.2 and 2.3. In section 2.4 and 2.5, we provide some background of the k-means clustering with k-means algorithm and Euclidean distance.

2.1. Customer Relationship Management

Alex Sbesbunoff defines CRM as the strategy integrating sales, marketing and service, which unites operating procedures and technology to better understand customers from different perspectives. Vince Kellen views CRM as a customer centric initiative that regards customer lifecycle as an important business asset and aims to retain customers and enhance customer satisfaction. According to Caldwell, CRM is not a new concept. Many businesses have practiced it for a long time, for instance, by memorizing customer's background and spending habits and introducing promotions targeting certain customers based on the information obtained.

Customer Relationship Management (CRM) emerged in the last decade to reflect the central role of the customer for the strategic positioning of a company. CRM takes a holistic view over customers. It encompasses all measures for understanding the customers and for exploiting this knowledge to design and implement marketing activities, align production and coordinate the supply-chain. CRM puts emphasis on the coordination of such measures, also implying the integration of customer-related data, meta-data and knowledge and the centralized planning and evaluation of measures to increase customer lifetime value.

CRM plays an importance role for companies that serve multiple groups of customers and exploit different interaction channels for them. This is due to the fact that information about the customers, which can be acquired for each group and across any channel, should be integrated with existing knowledge and exploited in a coordinated fashion.

It should be noted, however, that CRM is a broadly used term, and covers a wide variety of functions, not all of which require data mining. These functions include *marketing automation* (e.g., campaign management, cross- and up-sell, customer segmentation, customer retention), *sales force automation* (e.g., contact management, lead generation, sales analytics, generation of quotes, product configuration), and *contact center management* (e.g., call management, integration of multiple contact channels, problem escalation and resolution, metrics and monitoring, logging interactions and auditing), among others. We focus on how backend data mining and analytics can make these functions more effective.

2.2. Clustering

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters [4].We can show this with a simple graphical example:

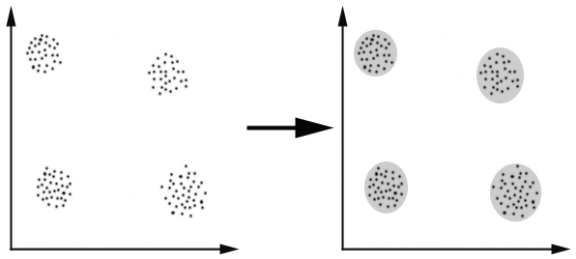

Figure 1. Clustering

2.3. Clustering Methods

The major clustering methods can be classified into the following categories.

1) **Partitioning methods**: Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k<=n. That is, it classifies the data into k groups, which together satisfy the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group. Notice that the second requirement can be relaxed in some fuzzy partitioning techniques.
2) **Hierarchical methods**: A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical 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, starts with each object forming a separate group. The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster.
3) **Density-based methods**: Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold.

4) **Grid-based methods**: Grid-based methods quantize the object space into a finite number of cells that form a grid structure. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.
5) **Model-based methods**: Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points. It also leads to a way of automatically determining the number of clusters based on standard statistics, taking noise or outliers into account and thus yielding robust clustering methods.

2.4. K-means clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem [7]. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

The main idea is to define *k* centroids, one for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate *k* new centroids as barycenters of the clusters resulting from the previous step. After we have these *k* new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the *k* centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

K-Means Algorithm is a follows:

Input: The number of clusters k and a database containing n objects.

Output: A set of k clusters.

Method:

1) arbitrarily choose k objects from D as the initial cluster centers,
2) **repeat**
3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster,
4) update the cluster means, i.e, calculate the mean value of the objects for each cluster,
5) **until** no change.

2.5. Euclidean Distance Measure

Different measures may be used in determining similarities and differences. In this system, Euclidean distance measure is used to calculate the closest centroids for each object. The Euclidean distance measure is frequently used as a distance that would measure on numeric value, and is easy to use in two dimensional planes.

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ...+ |x_{in} - x_{jn}|^2)}$$

3. Overview of the system

In this paper, there are three steps in the process of customer segmentation as shown in Figure 2. These steps are (i) Data preparation and transformation in which the customer data are prepared as the variable data for the use of mining techniques; (ii) Clustering to calculate for customers segment by using *k-means* algorithm; and (iii) Presenting clusters with pie chart for easily understand to the users.

(i) **Data Preparation and Transformation:** The proposed system acquires the input data from users and compiles bank data to the database. In this system, the data of credit card customer from the database are transformed in appropriate forms. For example; when the user put the age for 300, our system say that the error of input data and transform 300 to 30.

(ii) **Clustering step:** While using clustering process, customers' behavioral data from database are applied for evaluation of customer segment. Then the number of cluster is given by the user and fined the centroid randomly. For each of the remaining customer data, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. Note that we produce up to three clusters in our proposed system. By doing this process, the customers who were used the credit card are group according to their characteristics, forming clusters.

(iii) Presentation step: In this step, we produce clusters with the form of pie chart for easily understanding to the users. So, the information of each cluster and appropriate chart can be seen and users can be made correct decisions with the concept of CRM.

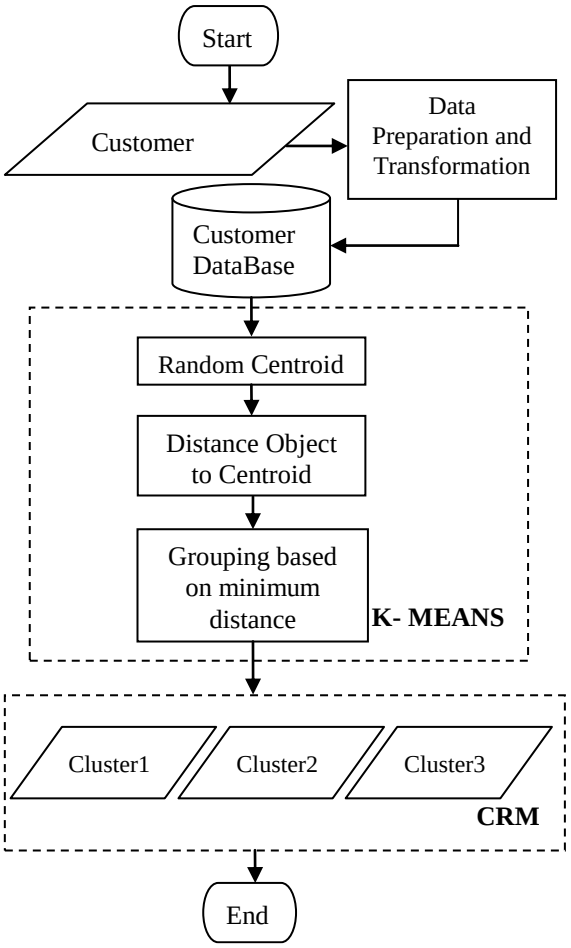

Figure 2. Design of Proposed System

4. Implementation of the System

The intelligent and effective of clustering system with the aid of *k-means* algorithm is presented in this section. This method primarily based on the information that is collected from bank and these data is stored in large database. This system is used Microsoft Access database and C# programming language. Table 2 shows the sample data for the system. The analysis of this data requires highly scalable algorithms that churn through the data looking for common aggregate patterns. We show the clustering result with pie chart. Customer understanding is derived from interpreting behavioral patterns. Intentions are then inferred from actions. Finally, the insights gained through k-means algorithm are presented in the form of 'chart' that can be used to score databases and real-time applications.

4.1. Data Set Description of the System

This system uses the *K-MEANS* algorithm and makes available the managers for the decision making. The data for the empirical study are credit card customer data and spending records in 2003 & 2004 from Bank. Each of this data set consists of customer record with eleven attributes such as age, sex, region, income, married, car, save_act, current_act, mortgage, pep, and type_of_card.

Table 1. Data Set of the System

Attributes	**Data Type**
Id	Numeric
Age	Numeric
Sex	Male, female
Region	Inner-city, town, rural, suburban
Income	Numeric
Married	Yes, No
Car	Yes, No
Save_act	Yes, No
Current_act	Yes, No
Mortgage	Yes, No
Pep	Yes, No
Type_of_card	Diamond, Gold, Silver

Table 2. Example of customer data

Id	Age	Sex	Region	Income (00)	Married	Car	Save act	Current act	Mortgage	Pep	Type_of_card
1	20	Male	Rural	100	Yes	Yes	No	Yes	Yes	No	Diamond
2	70	Female	Inner City	70	Yes	No	No	No	No	No	Gold
3	32	Female	Rural	120	No	No	Yes	No	No	Yes	Diamond
4	25	Male	Town	80	Yes	Yes	Yes	Yes	No	Yes	Gold
5	20	Female	Inner City	50	No	Yes	No	Yes	Yes	Yes	Silver
6	21	Female	Town	70	No	No	No	No	Yes	No	Silver

4.2 Clustering Results for the System

In this section, we will report the result of the experiment. In this example, the k-means algorithm of data mining task clusters the six points of customer data into three clusters.

Initially, we assign three clusters 2, 4 and 6 as the center of each cluster for the number of user specified clusters (k=3). Let the distance function d be Euclidean distance and the function name Sqrt be square root.

In iteration (1): The Euclidean distance is measured on numeric values, so the attribute values of age and income are only used in this measure. The distance measure of attribute age and income are as follows:

$d(2,1)=Sqrt[(70-20)^2 + (70-100)^2]=58.31$
$d(4,1)=Sqrt[(25-20)^2 + (70-120)^2]=20.61$
$d(6,1)=Sqrt[(21-20)^2 + (70-100)^2]=30.01$
$d(2,3)=Sqrt[(70-32)^2 + (70-120)^2]=62.801$
$d(4,3)=Sqrt[(25-32)^2 + (80-120)^2]=40.61$
$d(6,3)=Sqrt[(21-32)^2 + (70-120)^2]=51.19$
$d(2,5)=Sqrt[(70-20)^2 + (70-50)^2]=53.85$
$d(4,5)=Sqrt[(25-18)^2 + (80-50)^2]=30.81$
$d(6,5)=Sqrt[(21-18)^2 + (70-50)^2]=20.22$

The K-Means algorithm clusters the points into three clusters according to the input parameter k based on the above resulting distance measures as shown in Table 3.

Table 3. Final three clusters

Cluster no	1	2	3
Id	2	4, 1,3	6,5
Age	70	25.67	20.5
Sex	Female	Male	Female
Region	Inner-city	Rural	Inner-city
Income(00)	70	100	60
Married	Yes	Yes	No
Car	No	Yes	No
Save_act	No	Yes	No
Current_act	No	Yes	No
Mortgage	No	No	Yes
Pep	No	Yes	Yes
Type_of_card	Gold	Diamond	Silver

According to the data shown in Table 3, at most three points are included in clusters. So the algorithm terminates the process and presents the result as a pie chart to the user as shown in Figure 3 so that the information of each customer can be seen and manager can easily check the data of customers for their business.

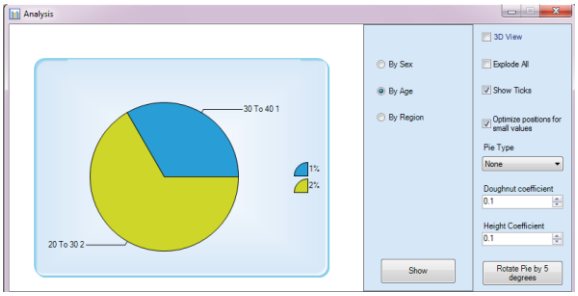

Figure 3. Result of the System

The pie chart in Figure 3 shows the information about the customers who use the diamond card in cluster 2. In this system, the decision makers or users can see the result based on three types (By Sex, By Age and By Region). Here, we use the type (By Age) to show the result. According to this type, it results two slices: the blue slice represents the customers of age between 30 and 40, and the yellow slice represents the customers of age between 20 and 30. As the result of these two slices, the number of customers of age between 20 and 30 are more than the number of customers of age between 30 and 40. So, decision makers understand customer needs and target their marketing strategy for the customers of age between 20 and 30.

5. Conclusion

With the increasing possibility of collecting data in business applications, there is a rising demand to utilize the available information. In this system, a data mining approach has been presented for clustering data from large amount of data to support manager to segment the customers and identify which customers to target for their business. For future research, another theory may be used to segment the same data of credit card customer usage such as neural network, or another data set may be used with k-means algorithm.

6. References

[1] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," *Second international symposium on information theory*, pp. 267–281.

[2] Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

[3] Berkhin, "Survey of Clustering Data Mining Techniques".

[4] "Data Mining" by HEIDI KUZMA AND SHEILA VAIDYA

[5] Dhillon, I. S., and Modha, D. S. (2001), "Concept Decompositions for Large Sparse Text Data Using Clustering," *Machine Learning*, 42, 143–175.

[6] Dortet-Bernadet, J., and Wicker, N. (2008), "Model-based clustering on the unit sphere with an illustration using gene expression profiles," *Biostatistics*, 9(1), 66–80.

[7] Hartigan, J. A., and Wong, M. A. (1979), "A K-means clustering algorithm," *Applied Statistics*, 28, 100–108.

[8] Schwarz, G. (1978), "Estimating the dimensions of a model," *Annals of Statistics*, 6, 461–64.

[9] Tibshirani, R. J., Walther, G., and Hastie, T. J. (2003), "Estimating the number of custers in a dataset via the gap statistic," *Journal of the Royal Statistical Society*, 63920, 411-423.

[10] Vince Kellen, "CRM Measurement Frameworks", *Blue Wolf White Paper*, p.4, 2002.