Journal of Data Analysis and Information Processing, 2015, 3, 63-71
Published Online August 2015 in SciRes. http://www.scirp.org/journal/jdaip
http://dx.doi.org/10.4236/jdaip.2015.33008

Mining Profitability of Telecommunication Customers Using K-Means Clustering

Hasitha Indika Arumawadu[1], R. M. Kapila Tharanga Rathnayaka[2,3], S. K. Illangarathne[4]

[1]School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China
[2]School of Economics, Wuhan University of Technology, Wuhan, China
[3]Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Balangoda, Sri Lanka
[4]School of Management, Wuhan University of Technology, Wuhan, China
Email: hasitha87@gmail.com, kapilar@sab.ac.lk, skillangarathne@gmail.com

Received 2 July 2015; accepted 18 August 2015; published 21 August 2015

Abstract

Data mining is the powerful technique, which can be widely used for discovering the customers' behaviors as well as customer's preferences. As a result, it has been widely used in top level companies for evaluating their Customer Relationship Management (CRM) system today. In this study, a new K-means clustering method proposed to evaluate the cluster customers' profitability in telecommunication industry in Sri Lanka. Furthermore, RFM model mainly used as an input variable for K-means clustering and distortion curve used to identify optimal number of initial clusters. Based on the results, telecommunication customers' profitability in Sri Lanka mainly categorized into three levels.

Keywords

K-Means Clustering, Data Mining, RFM Model, Customer Relationship Management

1. Introduction

Customer satisfaction and attraction are one of the significant goals in top level leading companies today. It will directly impact on companies' revenue and income. Customers' profitability is the profit the company makes from serving a customer or customer group over a specified period of time. The customers who provide more profit to the company are called high profitability customers. So, understanding profitability of the customer is the most important factor for the companies' future development. Generally, in Telecommunication Company customer's profitability can be categorized under three different levels. They are revenue (monthly bill value), call duration and total number of calls in given time period.

How to cite this paper: Arumawadu, H.I., Rathnayaka, R.M.K.T. and Illangarathne, S.K. (2015) Mining Profitability of Telecommunication Customers Using K-Means Clustering. *Journal of Data Analysis and Information Processing*, **3**, 63-71. http://dx.doi.org/10.4236/jdaip.2015.33008

The understanding of the nature of customer portfolios assists to make future decisions. So, mining profitability of the customers will make huge advantage for managers to make their future decisions.

1.1. Customer Relationship Management (CRM)

Customer relationship management (CRM) is an approach to managing a company's interactions with current and future customers. It often involves using technology to organize, automate, and synchronize sales, marketing, customer service, and technical support. Since the early 1980s, the concept of customer relationship management in marketing consists under the four different dimensions. They are; customer identification, customer attraction, customer retention and customer development have gained its importance. According to the literature, very few studies can be seen relates to the CRM. It can be describe as a comprehensive strategy and process of acquiring, retaining and partnering with selective customers to create superior value for the company and the customer [1]-[3].

The CRM systems can also give customer-facing staff detailed information on customers' personal information, purchase history, buying preferences and concerns. It is one of the most important divisions in any company. So, CRM directly communicate with customers for managing interaction between company and the customer. The CRM databases include current information and transactions of the customers. It has direct link with Data-ware house. Generally, mining part is handling through Data-ware house. So, Integration between CRM and Data-ware house is most important.

1.2. RFM Variables (Recency, Frequency, Monetary)

The RFM stands for recency, Frequency and Monetary value. RFM analysis is a marketing technique used for analyzing customer behavior such as how recently a customer has purchased (recency), how often the customer purchases (frequency), and how much the customer spends (monetary). It is a useful method to improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions [4] [5].

- Recency refers to the interval between the time, that the latest consuming behavior happens, and present. Many direct marketers believe that most-recent purchasers are more likely to purchase again than less-recent purchasers [6].
- Frequency is the number of transactions that a customer has made within a certain period. This measure is used based on the assumption that customers with more purchases are more likely to buy products than customers with fewer purchases.
- Monetary refers to the cumulative total of money spent by a particular customer.

1.3. Clustering

Clustering basically deals with grouping of objects such that each group consists of similar or related objects. The main idea behind clustering is to maximize the intra-cluster similarities and minimize the inter cluster similarities. Very common methods of clustering involve computing distance, density and interval or a particular statistical distribution. Depending on the requirements and data sets we apply the appropriate clustering algorithm to extract data from them. The Clustering has a broad spectrum and the methods of clustering on the basis of their implementation can be grouped as follows [7]-[9]. **Figure 1** clearly shows the clustering methods based on several criterions.

Partitioning Method: Given a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k ≤ n. That is, it divides the data into k groups such that each group must contain at least one object. e.g.: K-means, K-medoids.

Hierarchical Method: While partitioning methods meet the basic clustering requirement of organizing a set of objects into a number of exclusive groups, in some situations we may want to partition our data into groups at different levels such as in a hierarchy. A hierarchical clustering method works by grouping data objects into a hierarchy or "tree" of clusters.

e.g.: Agglomerative, divisive, BIRCH.

Density-Based Method: Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty in discovering clusters of arbitrary

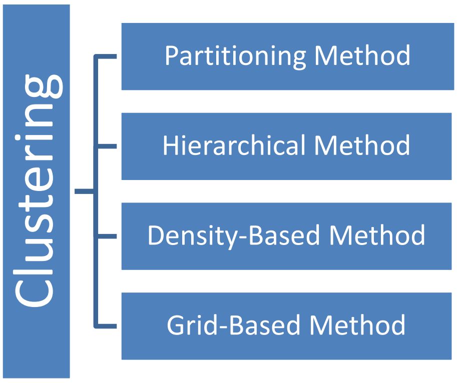

Figure 1. Clustering methods [1] [3] [4].

shapes. Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing a given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold. For example, 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 or outliers and discover clusters of arbitrary shape.

e.g.: DBSCAN, DENCLUE.

Grid-Based Method-Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All the clustering operations are performed on the grid structure (*i.e.*, on the quantized space). 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.

e.g.: STING, CLIQUE.

K-Means Clustering

The K-Means Clustering is a method used to classify semi structured or unstructured data sets. This is one of the most common and effective method to classify data because of its simplicity and ability to handle voluminous data sets. Generally, it accepts the number of clusters and the initial set of centroids as parameters. The distance of each item in the data set is calculated with each of the centroids of the respective cluster. The item is then assigned to the cluster with which the distance of the item is the least [10]. The centroid of the cluster to which the item was assigned is recalculated. One of the most important and commonly used methods for grouping the items of a data set using K-Means Clustering is calculating the distance of the point from the chosen mean. This distance is usually the Euclidean Distance [11].

2. Literature Review

The Telecom operators' activity consists of gathering and managing a large amount of information and data. Thus, millions of people, in millions of places can perform tens or hundreds of transactions in a short period resulting in billions of events to be recorded. In order to handle such an enormous quantity of data, special analyses methods need to be involved. These have appeared and grown at the same pace with the information technology [12] [13].

CRM segmentation is a fundamental component in the companies' strategic marketing planning in industrialized countries because goods and services can no longer be produced and retailed without taking into consideration the customers' needs and wishes and the fact that they differ [6]. The segmentation concept relies on the fact that it is more likely for persons grouped according to common behaviors and needs to have a homogenous response to marketing actions [14].

CRM segmentation refers to the process of splitting current or potential customers into several groups both as homogenous as possible and as heterogeneous among themselves [15] [16]. The perspective of segmentation being just an organizational tool is shallow because it bears both strategic and tactical marketing implications. Strategically speaking, segmentation allows the identification of profitable customers, the stabilization of market decisions and market segments and placing the product or service on the market. At operational level it drives the companies to lay more stress on enhanced customer understanding to develop more efficient relations with them [15].

According to Swift [17], companies can gain many benefits from CRM implementation. He states that the benefits are commonly found in one of these areas:

- Lower cost of recruiting Customers;
- No need to acquire so many customers to preserve a steady volume of business;
- Reduced cost of sales;
- Higher Customer Profitability;
- Increased Customer retention & Loyalty;
- Evaluation of customers Profitability.

CRM segmentation has become more efficient in the recent years due to the development of database marketing techniques. Profiling techniques provide marketers with superior tools for customer segmentation and adaptation of marketing strategies to the specific needs of each consumer segment [11].

In general, in order to perform customer segmentation, companies use criteria that relate to geographical, demographical, psycho-graphical, socio-economic, behavioral characteristics and psychological attitudes toward the respective product or service [18]. Most companies use segmentation based on demographics. In the case of markets that feature high competition, such as the telecom market, this approach is not enough. These companies also need to consider the information related to the customers' needs, consumer behavior, service or payment preferences, perception of product, probability of leaving the network, growth potential and customer migration. To become subject to segmentation, a market needs to be heterogeneous segmentation criteria need to fulfil the following conditions [19]:

- Segments will be measurable so that the size, buying power and other characteristics can be quantified and determined;
- Segments will be substantial and have a potential for profit so that they justify the creation of special later created marketing programs;
- Segments are homogenous, each being distinct in terms of clients' profiles and needs;
- Segments are accessible and differentiable so that they can be differently approached marketing-wise;
- Segments enable action, namely they allow the formulation of efficient programs to attract and serve customers;

According to Bayer [14] mobile operators use basically the following segmentation types: subscriber value-based segmentation, subscriber behavior-based segmentation, subscriber lifecycle-based segmentation and subscriber (possible) migration-based segmentation. They are used for different situations and focus on different aspects.

Data Mining (DM) is a powerful new technique to help companies discover the patterns and trends in their customers' preferences. It is also a well-known tool for customer relationship management (CRM). Data mining methodology has made a tremendous contribution for researchers wanting to extract hidden knowledge and information [3]. Proposed a new procedure, based on an expanded RFM model, by including two additional parameters D and C. It constructs a model for clustering customer value based on RFMDC attributes and K-means algorithm.

The major DM process uses data exploration technology to extract data, create predictive models using decision trees, and test and verify the stability and effectiveness of the models. The K-means method segments customers into clusters based on billing, loyalty and payment behaviors to create decision tree-based models. Determining the number of k clusters in a data set with limited prior knowledge of the appropriate value is a common problem that is distinct from solving data clustering issues [20] [21]. Two methods for selecting the initial centroids that save computation iterations in K-means clustering: 1) Carrying centroids forward; 2) Minimum impact. Both approaches are designed to expedite K-means computing and the identification of K.

3. Methodology

Here we going to apply K-means clustering algorithm for cluster telecommunication customer into different levels according to their profit level. The methodology of this research is organized under the five different phases. In the first phase, use RFM model as dimensions of input data. Then divide under different Segment's according to the RFM data. Next determine the best K value by applying distortion curve and determining weighted RFM values for each cluster. Finally, examine the profitability of the customers according to the RFM values in each cluster.

3.1. RFM Analysis

RFM analysis depends on Recency (R), Frequency (F), and Monetary (M) measures which are three important purchase-related variables that influence the future purchase possibilities of the customers.

According to the telecommunication industry here we take RFM as,

- R (Recency)—Average time duration between two calls for 1 month (in hours).
- F (Frequency)—Average calls taken per day for 1 month.
- M (Monetary)—Customers bill value for 1 month.

Here we have taken 100 customers as a sample data.

3.2. K-Means Clustering

Based on concept of K-means clustering below discussed the steps.

Step 1: Select K initial cluster centroids, C_1, C_2, C_3, …, C_k. K number of observations is randomly selected from all N number of observations, according to the number of clusters, and these become centers of the initial clusters. In here we use different K values and repeat the process.

Step 2: Proceed through the list of items. For each of the remaining N-K observations, assign an item to the cluster whose centroid is nearest (distance is computed by using Euclidean) and re-calculate the centroid for the cluster receiving the new item or for the cluster losing the item.

$$d_{euc} = \sum_{i=0}^{n} \sqrt{\left(x_i - c_i\right)^2} \qquad (1)$$

d_{euc}—Euclidean distance.

x_i—i^{th} point in cluster.

i—Number of points in cluster.

Step 3: Repeat Step 2 until no more reassigning. Rather than starting with a partition of all items into K preliminary groups in Step 1, we could specify K initial centroids (seed points) and then proceed to Step 2.

Below algorithm used for cluster above details using R programming.

```
data<-read.csv('clustering.csv',header=TRUE) #Import the data file
clus<-kmeans(data,centers=4) # Divided into 4 clusters
plot(data,col=clus$cluster) # Plot the graph
points(clus$centers,col=1:20,pch=8) # Mark the centers of each clusters
```

Figure 2 shows the final output after clustering into 4 clusters. Above graph is a combination of 6 graphs. x, y coordinates represents Recency, Frequency and Monetory as shown in the graph. But the problem is to determine initial number of clusters. So, determine optimal number of clusters we used Elbow method.

3.3. Determine Best K Value by Applying Distortion Curve

We used distortion curve method to find an appropriate value for K by running a standard K-means operation on all k values between 1 and K_{max}. Here we take K_{max} as 15. In Elbow curve (or distortion curve) x axis represents number of cluster and y axis represents within groups sum of squares in clusters [22].

Below algorithm used for plot the elbow curve.

```
wss <- (nrow(data)-1)*sum(apply(data,2,var))

for (i in 2:15){
wss[i] <- sum(kmeans(data,centers=i)$withinss)
}
plot(1:15, wss, type="b", xlab="Number of Clusters",ylab="Within groups sum of squares")
```

We then computed the resulting clustering distortion to find a specific range that features a minimal decrease

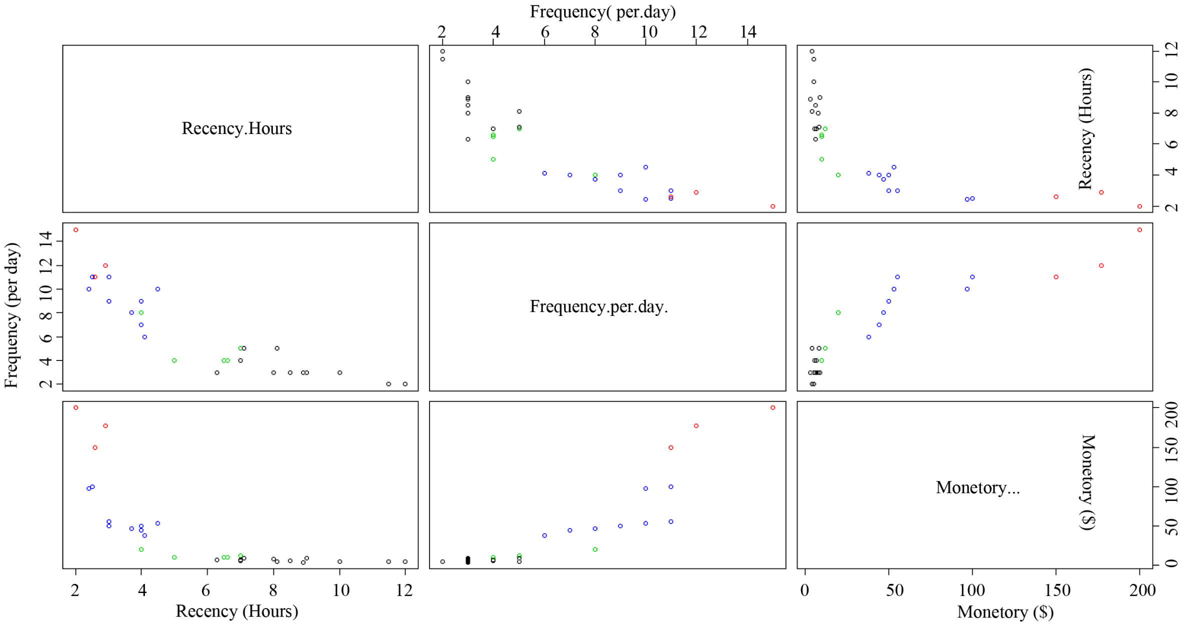

Figure 2. Multi-dimensional K-means clustering.

in average diameter (**Figure 3**). According to the above distortion curve optimal number of cluster is 3. Hence we take initial number of cluster as 3.

3.4. Determining Weighted RFM Values for Each Clusters

Table 1 clearly shows final clustering results. In this step we want to determine weights of the RFM values. Because we can't treat RFM values as a same weights. It is not reasonable. The weight of variables can be shown as, W_R, W_F, W_M. In addition $W_R + W_F + W_M = 1$.

Then we have to calculate the value of each cluster by using below equation. Value of the R is low mean that customer takes calls very recently. R is high mean he not take calls very recently. Hence we used $1/R_k$ value for calculating values of each clusters.

$$V_k = W_R\left(1/\overline{R_k}\right) + W_F\overline{F_k} + W_M\overline{M_k} \tag{2}$$

V_k—value of k^{th} cluster.

$\overline{R_k}, \overline{F_k}, \overline{M_k}$—average RFM.

Here we take $W_R\overline{R_k} + W_F\overline{F_k} + W_M\overline{M_k}$ values as, $W_R = 0.2$, $W_F = 0.2$, $W_M = 0.6$.

3.5. Examine the Profitability of the Customers According to the RFM Values in Each Clusters

By examine the values of the each cluster we can determine the profitability of the customer. By analyzing above table (**Table 2**) we can determine cluster 2 is the most profitable customers. Cluster 1 include low profitable customers. So, we can decide:

Cluster 2: High profitable customers.
Cluster 4: Profitable customers.
Cluster 3: Medium profitable customers.
Cluster 1: Low profitable customers.

4. Discussion

Here we used K-means clustering method for determine profitability of telecommunication customers. K-means clustering is the better way to segment customers into different levels. The major drawback of the K-means

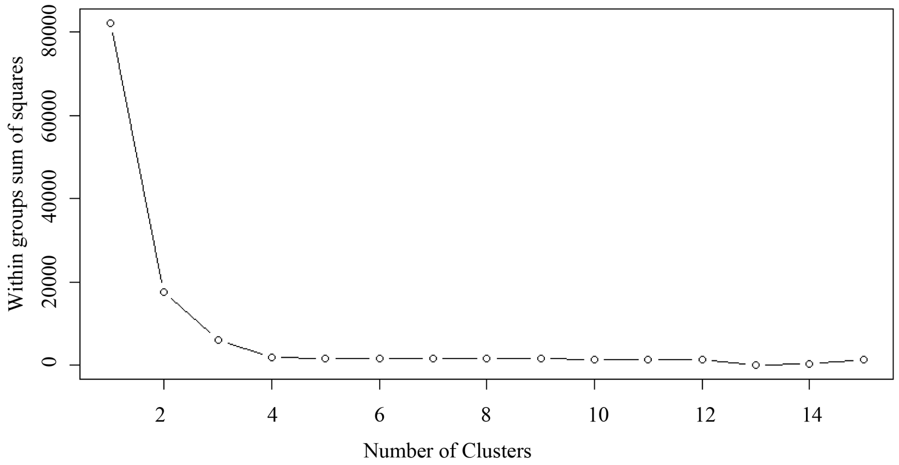

Figure 3. Distortion curve.

Table 1. RFM values for each clusters.

Cluster	Average R	Average F	Average M
1	8.62	3.33	5.98
2	2.50	12.67	175.67
3	5.82	5.00	12.38
4	3.47	9.00	59.33

Table 2. Final value for each clusters.

Cluster	V_k
1	4.277
2	108.016
3	8.462
4	37.456

clustering is to determine initial centroids. Here we used distortion curve to find the optimal number of initial centroids. According to the above we computed the resulting clustering distortion to find a specific range that features a minimal decrease in average diameter. So, we found the optimal number of initial clusters as 3. Before clustering the data we taken RFM model as an input variable. RFM model is the best way to describe telecommunication customers' data. According to the experiment we determine RFM values for each clusters. To improve efficiency we used weighted RFM values for determine final value of the each clusters. According to the final values we divided customers into three profitable levels. High profitable customers, Medium profitable customers and Low profitable customers.

So, understanding profitability and behavior of customers, managers can make decisions to improve their service and increase revenue of the company. High profitable customers also called loyalty customers, because they are very important to the company. They directly affect to the company's income. Companies need serve better service handle above discussed customer category. They have to handle them with care. If one of the high profitable customer leave company it will effect to the company. So, most important customer category is high profitable customers. Companies can use different strategies to satisfy them and keep them with company longer time. They can give some discounts, promotions, give some gifts on their special days like birthdays, anniversaries etc. In telecommunication industry people call onto hotline if they have any problem. So, sometimes they have to wait more time over the telephone for contact customer officers. If loyalty customer called and wait more time he will definitely dissatisfy on service. As a strategy they can arrange special hotline service for them and assign more skilled and senior customer officers to handle them. Sometimes customers visit to company. That time the customer officer who handle the customer should have better communication skills and good personality. So, they can arrange special service department to handle loyalty customers. They can arrange some special identity card to identify them so they can directly get the service by showing that special identity.

We can't ignore medium and low profitable customers. Because they are part of the companies' profit. One day they can be loyal customer. So, companies have to concern about them. They can motivate them for increase their income. They can arrange promotions, discounts to satisfy them. They can send some letters by remembering promotions and discounts.

K-means clustering is better way to segment customers. So, after segmenting customers into profitable level, we can use decision trees to make better decisions and it will be useful for senior management in the company. Other thing is we can cluster them monthly wise and then we can identify customers' behaviors like moving through the clusters. Sometimes high profitable customers can be decreased their level to medium or low. Some can be increased their level from low, medium into high. So, predicting customers' behavior is another aspect of mining. It will be great advantage for the company to stop the decrease moving of the customers.

Normally companies use customers revenue for decide profitability of the customers. But it is not a good way for make decisions. They have to concern more criteria. So, here we used RFM model to address above drawback. RFM is the better way to handle different behaviors of customers' profit. In telecommunication industry especially mobile telecommunication industry by considering only the revenue of the customer is not a better way. So, here we used Recency and Frequency of the customers to address above drawback. Some customers have higher revenue but they are not calling so frequently. Some customers have lower revenue but they are calling very frequently. If they call frequently they always keep in touch with company. So we have to concern about them. So, above reason we used weighted RFM model for improve it.

5. Conclusions

The categorizations of customer analysis under the miscellaneous type of requirements are challenging task in telecommunication industry today. An increasing number of customers make this problem more and more complicated. K-means clustering with RFM model is a better way to address the above problem. Main problem of the K-means clustering is to find initial cluster centroids. Here we proposed distortion curve to identify optimal initial centroids.

Increasing the number of customers is the main challenge in modern telecommunication industry. Day by day customers are increasing due to improvement of modern technology. By using this method to huge customer database the process is become slowly. As a future work we can propose concept of big data to handle huge customers' data. We can use parallel computing like Map Reduce to improve efficiency of the K-means algorithm. By the way we can improve RFM model into more complicated model like RFMDC model. We can add additional two variables into RFM model for improving efficiency of the input variable of the K-means algorithm. D and C stand for Diversity and Continuousness. The varieties of services we can use for evaluations are, especially, volume of customer's IDD calls and ROAMING calls. Continuousness is the continuous following sequence in a particular period.

References

[1] Mohammad, S., Hosseini, S., Maleki, A. and Gholamian, M.R. (2010) Expert Systems with Applications Cluster Analysis Using Data Mining Approach to Develop CRM Methodology to Assess the Customer Loyalty. *Expert Systems with Applications*, **37**, 5259-5264. http://dx.doi.org/10.1016/j.eswa.2009.12.070

[2] Rathnayaka, R.M.K.T., Wei, J.G. and Seneviratne, D.M.K.N. (2014) Geometric Brownian Motion with Ito Lemma Approach to Evaluate Market Fluctuations: A Case Study on Colombo Stock Exchange. *International Conference on Behavioral, Economic, and Socio-Cultural Computing* (*BESC*'2014-*IEEE*), Shanghai.

[3] Liang, Y. (2010) Expert Systems with Applications Integration of Data Mining Technologies to Analyze Customer Value for the Automotive Maintenance Industry. *Expert Systems with Applications*, **37**, 7489-7496. http://dx.doi.org/10.1016/j.eswa.2010.04.097

[4] Hajiha, A., Radfar, R. and Malayeri, S.S. (2011) Data Mining Application for Customer Segmentation Based on Loyalty: An Iranian Food Industry Case Study. *IEEE International Conference on Industrial Engineering and Engineering Management*, Singapore, 6-9 December 2011, 504-508.

[5] Rathnayaka, R.M.K.T. and Seneviratne, D.M.K.N. (2014) G M (1, 1) Analysis and Forecasting for Efficient Energy Production and Consumption. *International Journal of Business*, *Economics and Management Works*, **1**, 6-11.

[6] Anchalia, P.P. (2013) Map Reduce Design of K-Means Clustering Algorithm.

[7] Ren, D., Zheng, D., Huang, G., Zhang, S. and Wei, Z. (2013) Parallel Set Determination and K-Means Clustering for Data Mining on Telecommunication Networks.

[8] Wedel, M. and Kamacura, W.A. (2000) Market Segmentation: Conceptual and Methodological Foundations. Kluver Academic Publishers, Boston. http://dx.doi.org/10.1007/978-1-4615-4651-1

[9] Jayathileke, P.M.B. and Rathnayaka, R.M.K.T. (2013) Testing the Link between Inflation and Economic Growth: Evidence from Asia. *Modern Economy*, **4**, 87-92. http://dx.doi.org/10.4236/me.2013.42011

[10] Dibb, S. (1999) Criteria Guiding Segmentation Implementation: Reviewing the Evidence. *Journal of Strategic Marketing*, **7**, 107-129. http://dx.doi.org/10.1080/096525499346477

[11] McDonalds, M. and Dunbar, I. (2004) Marketing Segmentation: How to Do It, How to Profit from It. Elsevier Butterworth-Heinemann, Oxford.

[12] Dibb, S. (2001) New Millennium, New Segments: Moving towards the Segment of One? *Journal of Strategic Marketing*, **9**, 193-213.

[13] Rathnayaka, R.M.K.T. and Wang, Z.-J. (2012) Prevalence and Effect of Personal Hygiene on Transmission of Helminthes Infection among Primary School Children Living in Slums. *International Journal of Multidisciplinary Research*, **2**, 1-13.

[14] Swift, R. (2001) Accelerating Customer Relationship Using CRM and Relationship Technologies. Prentice Hall Inc., New York.

[15] McCarty, J.A. and Hastak, M. (2007) Segmentation Approaches in Data-Mining: A Comparison of RFM, CHAID, and Logistic Regression. *Journal of Business Research*, **60**, 656-662.

[16] Rathnayaka, R.M.K.T., Seneviratne, D.M.K.N. and Wang, Z.-J. (2014) An Investigation of Statistical Behaviors of the Stock Market Fluctuations in the Colombo Stock Market: ARMA & PCA Approach. *Journal of Scientific Research & Reports*, **3**, 130-138. www.sciencedomain.org
http://dx.doi.org/10.9734/JSRR/2014/5409

[17] Tsiptsis, K. and Corianopoulos, A. (2009) Data Mining Techniques in CRM: Inside Customer Segmentation. John Wiley & Sons Ltd., Chicester.

[18] Dibb, S. and Simkin, L. (2010) Judging the Quality of Customer Segments: Segmentation Effectiveness. *Journal of Strategic Marketing*, **18**, 113-131. http://dx.doi.org/10.1080/09652540903537048

[19] Rathnayaka, R.M.K.T. and Wang, Z.-J. (2013) Influence of Family Status on the Dietary Patterns and Nutritional Levels of Children. *Food and Nutrition Sciences*, **3**, 1055-1059. http://www.SciRP.org/journal/fns

[20] Tonks, D.G. (2009) Validity and the Design of Market Segments. *Journal of Marketing Management*, **25**, 341-356. http://dx.doi.org/10.1362/026725709X429782

[21] Rathnayaka, R.M.K.T. (2014) Cross-Cultural Dimensions of Business Communication: Evidence from Sri Lanka. *International Review of Management and Business Research*, **3**, 1579-1587. www.irmbrjournal.com

[22] Rathnayaka, R.M.K.T., Seneviratna, D.M.K.N. and Wei, J.G. (2015) Grey System Based Novel Approach for Stock Market Forecasting. *Grey Systems*: *Theory and Application*, **5**, 178-193.