Mining on K-way Means Clustered Streaming Data

[1]J.Gitanjali, [2]P.Dinesh Kumar, [3]B.Suresh Babu, [4]S.B.Prabakaran
[1,2,3,4] School of Information Technology and Engineering
[1] Assistant Professor
VIT University, Vellore-632014, Tamil Nadu, India

ABSTRACT
Data mining is the process of finding patterns or correlations from a different data set and changing it into the useful information. Clustering is dividing the data into groups that are similar in behavior among the data sets in a group and distinct across the groups. Data Stream mining is very important and challenging problem, because in business transactions we need to make better managerial choices and extract the essence of this streaming data where the data streams are temporally ordered, fast changing, large and continuous concurrent flow of data. Our objective in this paper is to propose a model using data mining, with the help key performance indicators (variables) found for each customer, clustering will be done using K-means clustering technique on real time basis with streaming data.

INDEX TERMS:
Data Mining, Clustering, Data Stream Mining, K-way Means, Key Performance Indicators, RFM.

1. INTRODUCTION

1.1 Literature survey
Nowadays many applications are generating streaming data, unlike traditional data sets these data streams [2] are difficult to mine because these streams are large in size and rapid moving, so it becomes difficult to store and cluster these kinds of data. Clustering [2], [3] acts as pre processing step in over all data mining [1] process, because clustering perform outlier analysis and classify data into interesting groups. Data streams are Pervasive; these can be found in many application domains from e-commerce [9] to business and hospital management systems. Now due to the advancement in computing and global connectivity, many new applications are adding every day. So there is need to effectively cluster and manage these kinds of data.

In this paper we are using k-way means clustering algorithm to cluster these kind of streaming data based on key performance indicators [4]. K-mean clustering is of vector quantization initially from signal processing. The main aim of this k-means clustering is to divide n samples into k clusters in which each sample fit in to the cluster with the nearby mean value, which gives a prototype of the cluster. Here a set of samples (z1, z2, … Zn) where each sample is a v-dimensional real vector, k- mean clustering targets to divide the x samples into k sets ($k \leq x$) A={A1, A2, … , Ak} to reduce the within-cluster sum of squares (WCSS):

$$\sum_{j=1}^{k} \sum_{zi \in Ai} pow(zi - Mi, 2)$$

Where, M_i is the mean of samples in A_i.
After removing collinear and insignificant variables, M variables will come out that are significant which will be used to cluster streamed data to find interesting information from large data.

2. EXISTING SYSTEM
RFM [8] is an analysis technique which scans the database for three decisive factors namely Recency, Frequency, Monetary value. To decide the best consumers by based on their purchases. RFM segmentation is generally used for direct marketing and has received meticulous interest in retail and proficient services industries.

Understanding these terms:
Recency – how lately did the consumer buy?
Frequency – how regularly do consumers buy?
Monetary value – to what extent do consumer spend?
On the other hand, we can create sub categorize the decisive factors. For example, the recency factor might be sub divided into three sub categories: consumers within the last 6 months; between 6 months to a year; and longer than a year. By using a data mining technique or business strategies these sub categories may be divided. Segmentation was done by intersection of the values from these categories. If each decisive factor has three sub categories then the resultant set would have 27 feasible combination sets. Recognizing the most important segments can exploit the risk relationships in the information utilized for dissection. Hence, it is remarkably proposed that an alternative set of information be utilized to accept the effects of RFM segmentation process.

In the traditional approach, clustering is done on static data by identifying objects which are having similar behavior and they are organized in to groups. This technique is easy because of following reasons

1. For traditional clustering data sets are static
2. Data can be easily stored in memory and can be scanned any number of times
3. The clustering results are static and don't change over time

In this traditional approach clustering can be done on static data sets. This clustering technique was done based on the means of adjacent data attributes. The similar mean distanced attributes will come under the appropriate clusters.

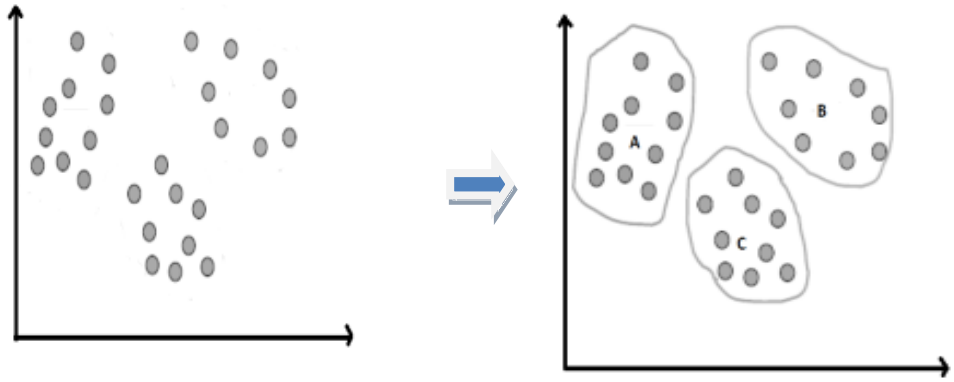

Figure 1: Traditional clustering process

3. CHALLENGES FOR DATA STREAMS

Streaming – It is a continuous flow of data from external sources in less time intervals that is frequently changing. Streaming data is very vigorous and it becomes indurate to cluster these kinds of data.

Large in size – Storing and processing of data streams is very expensive and difficult procedure, because of its large size and they move rapidly in and out of databases.

Global view not achievable – In traditional data clustering we can generate good clusters using predefined models, but for streaming data it's not possible. So we lack global view of data.

Outliers – Outliers may become big problem for clustering is neglected, since the data is changing rapidly interested data may become outlier and vice-versa.

Multi dimensional in nature – Due to the multi dimensional feature of data streams all points, looks to be in equal distance with other.

4. PROPOSED SYSTEM

Using RFM Technique, segmentation of streaming data is not possible as it is fast changing and also even using traditional clustering approach, the given challenges cannot be achieved. To overcome this problem, we adopted k-way means clustering technique to cluster streaming data to achieve business needs.

Here, as per the problem definition the streaming data sets will be continuously gathered and processed. On these data sets, k way means clustering technique was applied as a preprocess step before going with data mining. Data mining process can be done to get interesting information which is useful. This information was further analyzed to know the business status and can be used to improve their business by providing some discounts or increases in cost of products based on the demand.

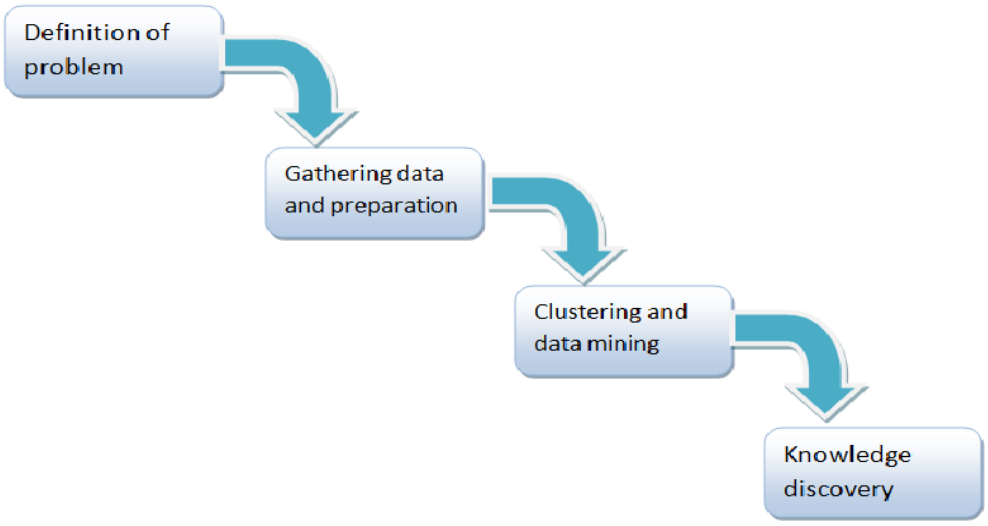

Figure 2: System architecture

5. IMPLEMENTATION

5.1 Import data into database

a. Installed Xampp software using "xampp-win32-1.7.4-VC6-installer".
b. After installation Xampp control needs to be executed to start Apache service and MYSQL service, in order to use the database.
c. Collected transaction level data from internet and imported by creating table in a database.
d. Transaction level data contains, bill level details like bill number, bill date, bill amount, telephone number of a customer, store name in which a customer has made transaction, discount amount availed on bill, tax amount paid on bill, Cash Return etc.,
e. Just by looking at transaction level data, we can't draw any conclusions about the behavior of any customer. So this data needs to be converted into information using techniques like data mining.

5.2 Data mining before clustering

a. Using data mining procedure with the help of MYSQL queries created key performance indicators of each customer using transaction level data.
b. Key performance indicators include Number of bills made by a customer, last purchase date, first purchase date, life time purchase, number of days since the last purchase of a customer, average gap of days between each purchase of the customer, average bill amount made, repeat bills, last purchase amount etc.,
c. Using all these key performance indicators, a customer level table is created where characteristics of each customer is explained.
d. By looking at these indicators behavior of a customer can be explained individually.

5.3 Clustering using k means algorithm

a. Customer table created using data mining will be used as input for clustering.
b. We have used data from a food industry for research study.

Clustering Methodology:

- Up to 20 Behavioral metrics were used based on various hypothesis
- Variables like – Average Bill Value, Lifetime Spend, Visits, Longevity, Latency, Group Size, # of SKUs, # of categories, Repeat Bills etc. were used for significance
- After removing collinear and insignificant variables, 5 variables came out significant which were - ABV, Latency, Recency, Total Visits and Life time purchase
- Having identified significant variables, K-Means methodology of clustering was adopted to cluster customers
- Plotting within-outside variance chart using multiple iterations, data was statistically assessed to have either K = 4 clusters or K = 5 clusters
- With K = 5, the clusters did not have any business significance and with 4 clusters clear patterns were observed and model was frozen at 4 clusters

Clustering Procedure:

- Using customer table created as input to the clustering procedure, after various iterations we have ended up with 5 key performance indicators out of 20 indicators.
- These 5 indicators which are non collinear and significant variables were used to cluster the customers.
- After deciding significant variables, K-means methodology was used and divided all customers into 4 clusters who have same behavior among clusters and behave differently across the clusters.
- Model is saved and applied on customers table. Output table is created with telephone number and cluster id for each customer.

5.4 *Data mining after clustering*

After getting cluster id for each customer, average key performance indicators were calculated for each cluster which explains the characteristics of each segment.

6. RESULTS:

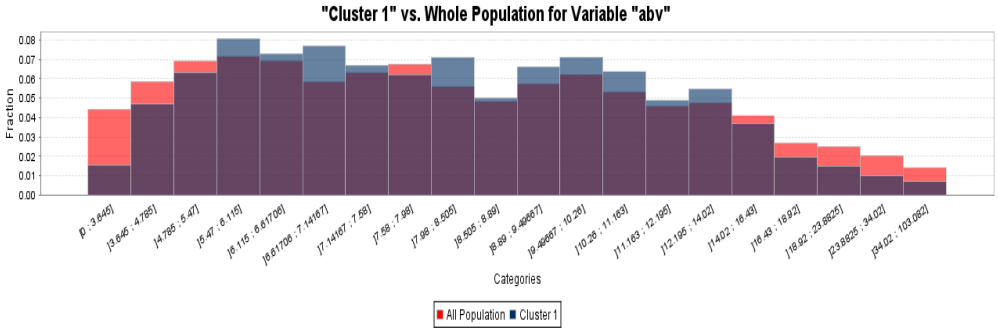

Figure 3.1

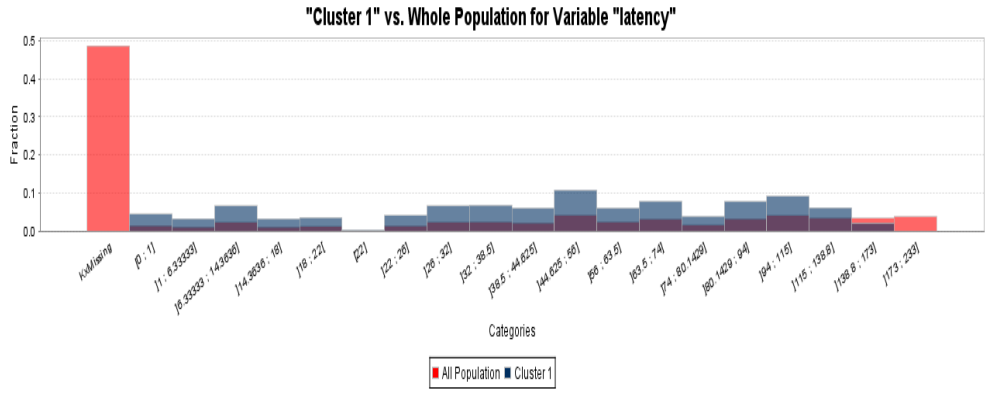

Figure 3.2

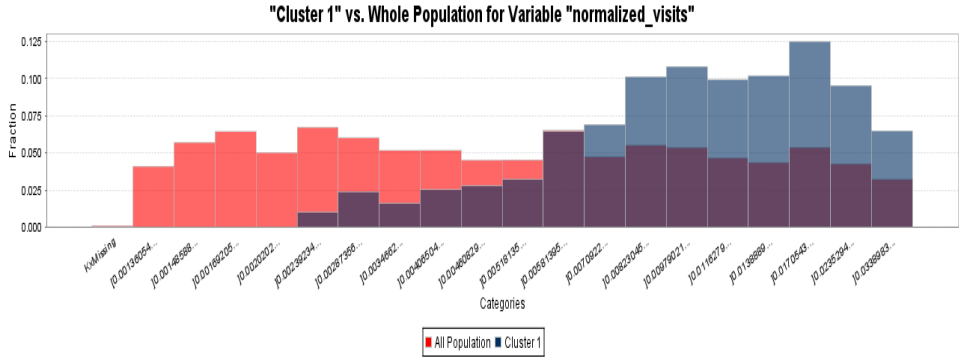

Figure 3.3

In fig 3.1, 3.2, 3.3, the abv, latency and normalized visits respectively are moderate, when Cluster 1 was compared with these key performance indicators. So we named this cluster as value consumers. These category consumers will come in average intervals compare to population as shown in fig 3.1 and have consciousness on their spending.

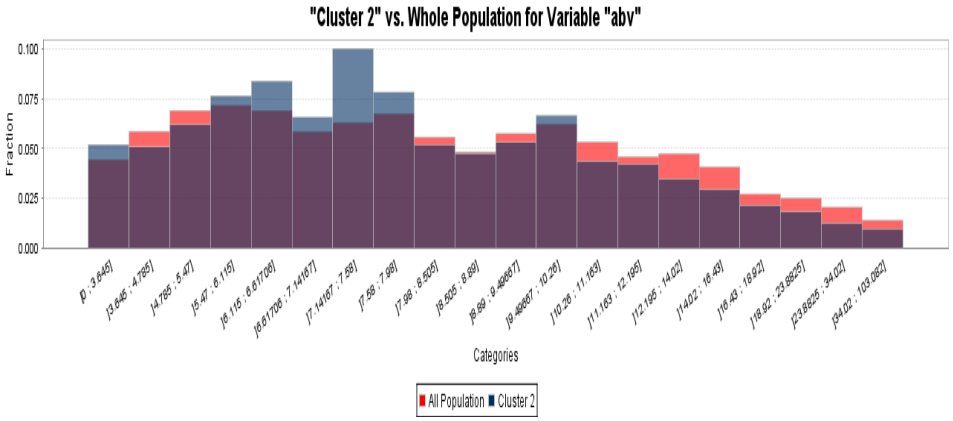

Figure: 4.1

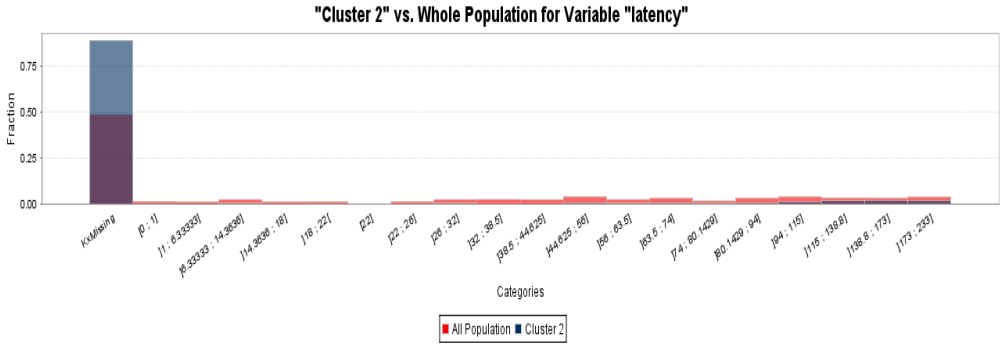

Figure 4.2

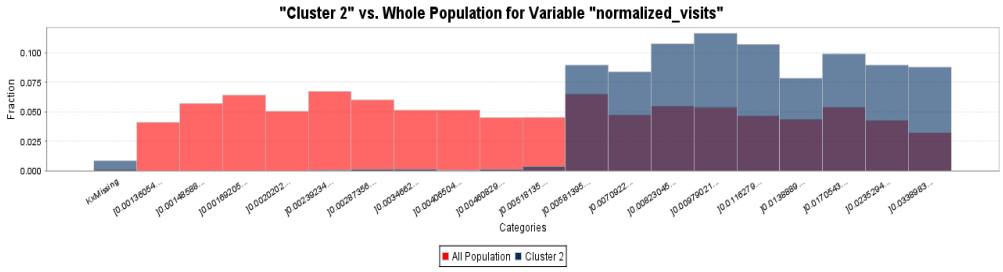

Figure 4.3

In fig 4.1, 4.2, 4.3, the abv, latency, and normalized visits respectively, this cluster having high bill value, less latency period i.e. they visit for many times and consumes more. This cluster can be named as Royal consumers.

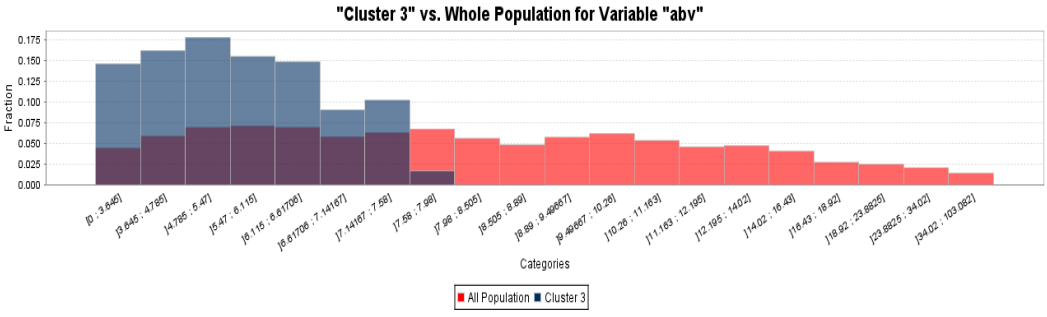

Figure 5.1

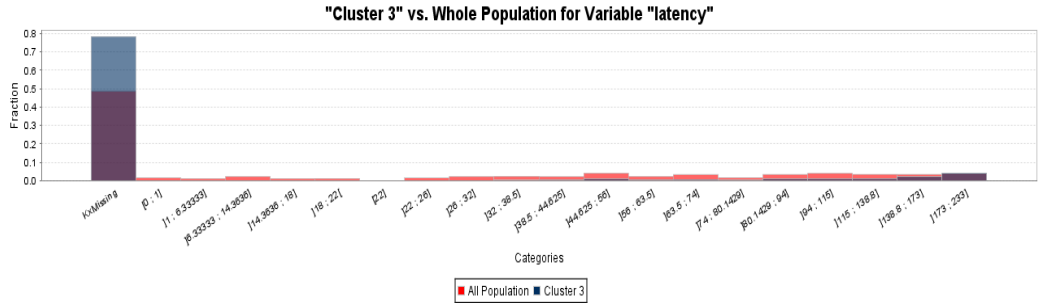

Figure 5.2

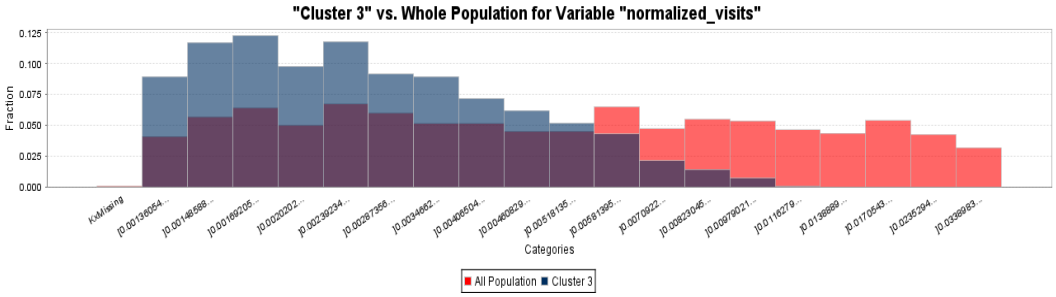

Figure 5.3

In fig 5.1, 5.2, 5.3, abv percentage is more, latency period are more i.e. they visiting with much time intervals and less normalized visits. So this cluster can be named as high dollar consumer.

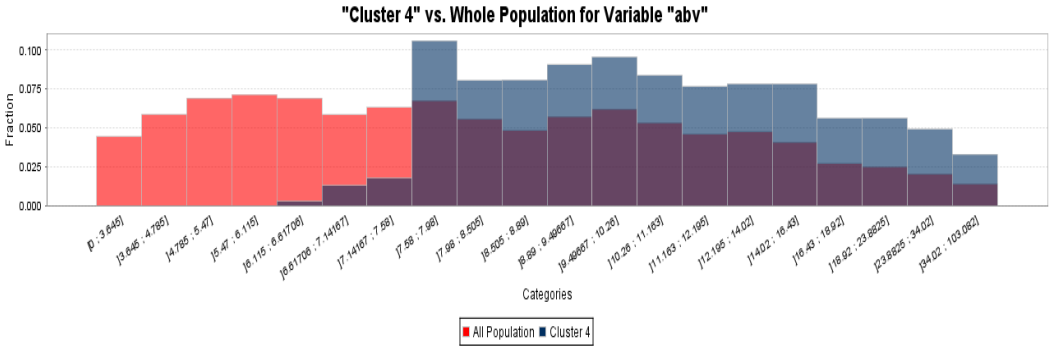

Figure 6.1

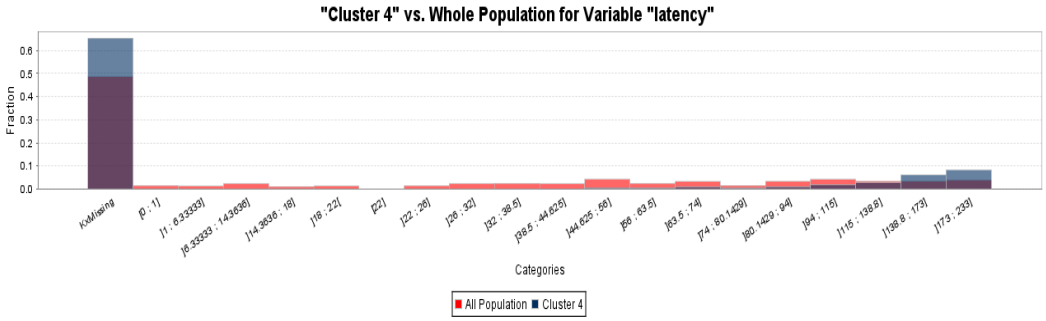

Figure 6.2

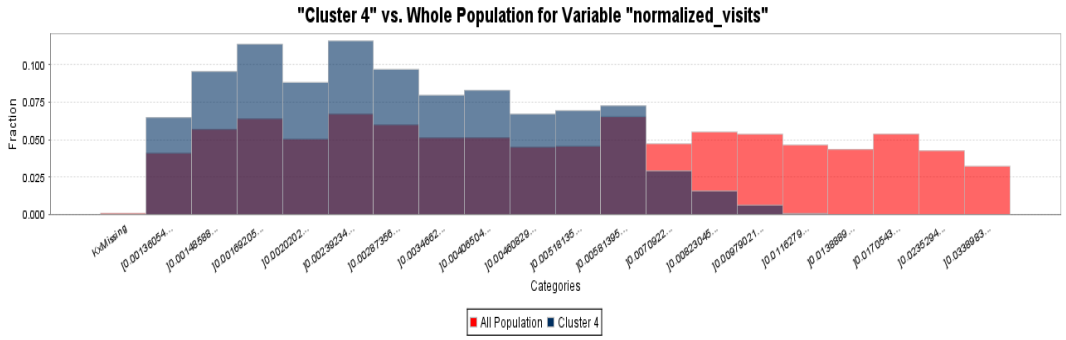

Figure 6.3

Table 1: Final output after clustering and mining process

phonenumber	kc_clusterId
11111111	3
20003561	1
21640584	1
22000504	2
22000517	2
22001210	4
22001226	2
22001233	2
22002222	1
22004006	1
22004302	3
22004333	1
22004345	4
22004396	2
22004412	1

In Table 1, it gives the mobile numbers of consumers and their appropriate cluster id. So, that we can easily differentiate the consumers based on their bill values, latency and their normalized visits in to the different clusters.

7. CONCLUSION

The massive and fast changing streaming data has captured and prepared for the data mining process with the use key performance indicators. By adopting the k means algorithm to generate the clusters of the data sets to make it available to the business bodies to get understand about their transactions.

In this analysis, we studied that the algorithm of k-mean clustering is not more flexible on some data set. In exact, the factor k is known to be tough to opt, when not given by exterior control. Compare to different algorithms, k-means cannot be used on non-numerical data and not be used with random distance functions. For these scenarios, many different algorithms have been developed. The time complexity of this technique is more. So, instead we can use some incremental models to achieve this.

8. ACKNOWLEDGEMENT

The authors are studying at Vellore Institute of Technology, Vellore. We would specially like to thank our guide Prof. In this above figures, abv value is more based on the discounts on products; latency is more i.e. they come at weekends and normalized visits are more. So we named this cluster as deal making consumers.

Mrs. J.Gitanjali (faculty at VIT University) for support given throughout the project work.

9. REFERENCES

[1] M, Kamber and J, Han. (2006). Data Mining: Concepts and Techniques, vol. 54, pp 212-225.Second edition.

[2] L. Callaghan. Et.all. 2001 “Streaming-Data Algorithms for High-Quality Clustering,” Proceedings of IEEE International Conference on Data Engineering.

[3] T. Zhang. Et.all (2006) “Birch: an efficient data clustering method for very large databases,” ACM SIGMOD international conference on Management of data, New York, NY, USA, pp. 103–114.

[4] Y. Chen and L. Tu, 2007. “Density-based clustering for real-time stream data,” in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD ’07, New York, NY, USA, pp. 133–142.

[5] A.Moga, I. Botan, and N. Tatbul. 2011 Upstream: Storage-centric Load Management for Streaming Applications with Update Semantics. VLDB Journal.

[6] B. Babcock. Et.all. 2002 Models and issues in data stream systems. In PODS, pages 1–16.

[7] Li Hengjie, Yang Dingxin, "Study on Data Mining and Its Application in E-business," Journal of Gansu Lianhe University (Natural Science), April 2006, pp 30-33.

[8] Young Sung Cho. Et.al. Implementation of personalized recommendation system using k-means clustering of item category based on RFM. KSCI, 13th-2 Vol, pp 1-5, Mar, 2008

[9] Lu Chuiwei, "Research and Application of Data Mining in E- commerc,"Market Modernization, April 2006, p 87.

[10] Chongsheng, Zhang. 2012. Modeling and Clustering Users with Evolving Profiles in Usage Streams. Temporal Representation and Reasoning. Pp 66-75.

[11] Yogita. 2011. Clustering Techniques for Streaming Data. Indian Institute of Technology, Roorkee.