Usage Behaviour Based Rigorous Clustering for Customer Segmentation and Profiling

[1]Ch.V.V.D. Prasad, [2]R Seeta Sireesha, [3]K Rohini, [4]Ch. Sailaja

Assistant Professor
CSE Department
Gayatri Vidya Parishad College of Engineering For Women

***Abstract*— The main aim of this paper focuses on the K-means and the DBSCAN clustering techniques to identify customer behaviors and group the customers in order to improve the customer service and to remove the outliers that are existing in the dataset. We use data mining techniques for customer profiling and segmentation. Clustering can be used for identifying and grouping similar customers. Finally, our focus is to find a way to divide the objects into clusters so that we can find frequent item in each cluster to improve sales by contacting customers through message, mail or phone based on the clusters.**

***Keywords*— Clustering, K-means clustering, DBSCAN clustering, Profiling, Segmentation.**

I. INTRODUCTION

Data mining is the process of extracting valid, useful, previously unknown, and ultimately comprehensible knowledge from large databases [1]. Data mining is considered as a step in the whole process of knowledge discovery problem statement [2]. Data mining techniques can be used efficiently in any business application that involves data, such as: (i) Increasing the business unit and overall profitability, (ii) Understanding customer desires and needs, (iii) Identifying profitable customers and acquiring new ones, (iv) Retaining customers and increasing loyalty, (v) Cross-selling and up-selling, (vi) Detecting fraud waste and abuse, (vii) Determining credit risks, (viii) Increasing web site profitability. Data mining can help companies in better understanding of the vast volume of data collected by the CRM systems [4]. Data mining can identify products that are often purchased together, which can help build product bundles that are more likely to be successful [5][6]. Today, data mining is being used by several industries including banking and finance, retail, insurance, telecommunications, etc. Other possible applications for data mining include database marketing, sales forecasting, call behaviour analysis and churning management in telecommunications; forecasting of demand for utilities, such as energy and water; simulation of chemical and other process reactions; finding critical factors in discrete manufacturing (aerospace, automobile, electronics); CPU usage and forecasting. It can help organizations better understand their business, be able to better serve their customers, and increase the effectiveness of the organization in the long run [5][6].

1.1 CLUSTERING

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

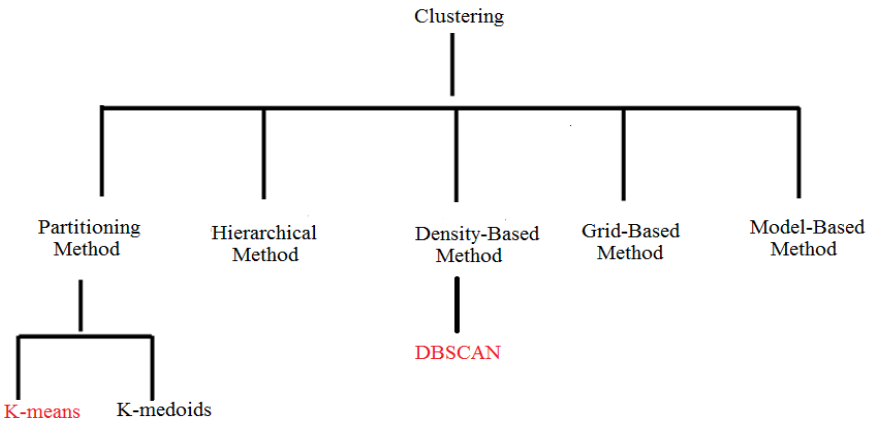

Fig. 1: Types of Clustering Methods

1.1.1 K-Means Algorithm

Input – number of clusters k and data set D containing n objects. Output – A set of k clusters.
Step 1: From D, randomly generate k points as the initial cluster centers.
Step 2: Assign each object to a cluster to which the object is the most similar, based on the cluster mean value and

the object value.
Step 3: Re-compute mean of each cluster from the objects in it and update the cluster means.
Step 4: Repeat steps (2) and (4) till there is no change in clusters.

1.1.2 DBSCAN Algorithm

DBSCAN is a density-based algorithm. Density is the number of points within a specified radius r (Eps). A point is a core point if it has more than a specified number of points (MinPts) within Eps. These are points that are at the interior of a cluster. A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point. A noise point is any point that is not a core point or a border point. Any two core points are close enough– within a distance *Eps* of one another – are put in the same cluster. Any border point that is close enough to a core point is put in the same cluster as the core point. Noise points are discarded.
Step 1: Select a point ***p***
Step 2: Retrieve all points density-reachable from ***p*** with respect to ε and ***MinPts***.
Step 3: If ***p*** is a core point, a cluster is formed.
Step 4: If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.
Step 5: Continue the process until all of the points have been processed.

1.2 Profiling and Segmentation

Customer profiling integrates several aspects of customers into a rational evaluation, such as customer details, historical records and contact details, customer attractiveness, or customer satisfaction [7][8][9]. Segmenting customers provides approaches to better understand their preferences and to more efficiently allocate resources based on the information. The benefit is twofold: (1) First, it enables companies to differentiate themselves by providing appropriate and suitable services for their customers' needs; therefore, building up a competitive advantage. (2) Second, it guides the companies to where their most valuable customers are located and helps allocate major capital, effort and time to generate the most profit [5]. Market segmentation is one of the central concepts in marketing and customer profitability as a segmentation criterion is a newer phenomenon which has become increasingly prevalent in many industries, leading to differential treatment of customers [14].
One of the most important points for customer profiling is targeting valued customer and having special attention to them . Xu and Walton [8][9] distinguished four criteria for segmenting customers: (1) customer profitability score, (2) retention score, (3) satisfaction and loyalty score,(4) response to the promotion. There are two basic approaches to segmentation: (1) market driven and (2) data driven.
Marketers use direct marketing campaigns to communicate a message to their customers through mail, the Internet, e-mail, phone, and other direct channels. The stages of direct marketing campaigns are illustrated in Fig. 2 and explained below :

1. Collect and clean the necessary data from different sources.
2. Customer analysis and segmentation (clustering) into different groups.
3. Development of targeted marketing campaigns in order to select the right customers.
4. Campaign execution by choosing the appropriate channel, the appropriate time, and the appropriate offer for each campaign.
5. Campaign evaluation through the use of test and control groups. The evaluation involves the partition of the population into test and control groups and comparison of the positive responses.
6. Analysis of campaign results in order to improve the campaign for the next round in terms of targeting, time, offer, product, communication, and so on.

Fig. 2: The stages of direct marketing campaigns

II. PROPOSED WORK

Flowchart for proposed work

As shown in Fig.3, the procedural steps for predicting the customer profiling and segmentation for electronic products. The three basic steps are:

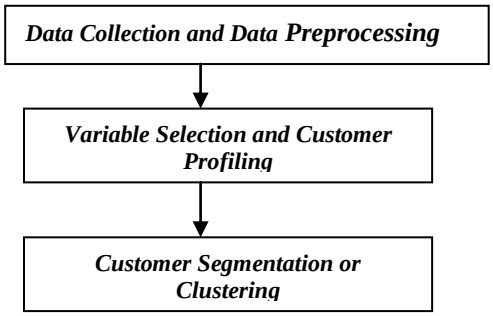

Fig. 3: Procedures steps for predicting the profile and segmentation for electronic products.
(Flowchart for proposed work)

(1) Data Collection and Data Preprocessing:

In the process of data collection we simply collected the reports from Electronic Gadgets Store and after that MIS department prepared one excel data sheet. In data preprocessing process the excel sheet was converted into .arff file that is compatible with WEKA. This dataset consists of 279 records with attributes like CustomerId,Name, Gender, Age, Salary, ItemName, MobileNum and MailId.

(2) Variable Selection and Customer Profiling:

In this process the Salary attribute was selected as input and MobileNum /MailId are used as output for contacting the profitable customers to improve the sales.

(3) Customer Segmentation or Clustering:

In this process the specific profiled data from data set was partitioned in different segments and clusters using K-means algorithm and identifying the outliers using DBSCAN algorithm.

III. RESULTS

Attribute ItemName is selected from the data set and the frequency count of each purchased item based on customer profiling is plotted in the Fig.4 and the corresponding items with frequency count is shown in Fig.5.

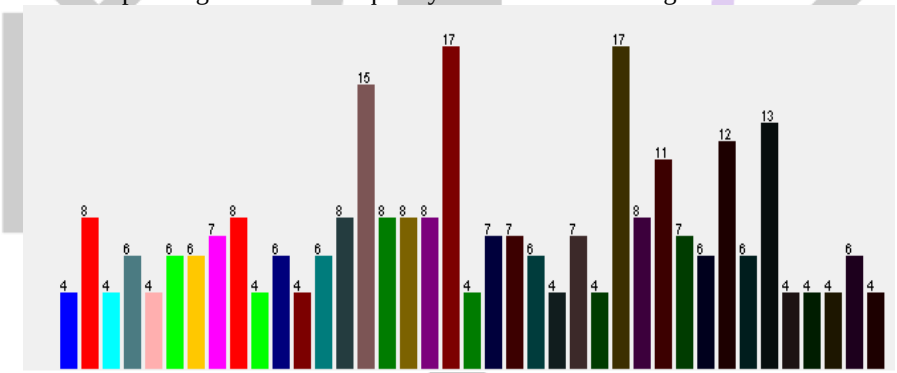

Fig. 4: Graph representing frequency count of purchased items

Selected attribute

Name: itemname Missing: 0 (0%) Distinct: 39 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	Acer	4
2	Apple	8
3	Asus	4
4	Bajaj	6
5	Benq	4
6	Birla	6
7	Braun	6
8	Canon	7
9	Compaq	8
10	Dell	4
11	Ereka	6

Selected attribute

Name: itemname Missing: 0 (0%) Distinct: 39 Type: Nominal Unique: 0 (0%)

No.	Label	Count
29	Panasonic	11
30	Phillips	7
31	Prestiage	6
32	Samsung	12
33	Segway	6
34	Sony	13
35	Tata	4
36	Toshiba	4
37	Vediocon	4
38	Voltas	6
39	Whirlpoll	4

No.	Label	Count
12	Fujitsu	4
13	Garmin	6
14	Godrej	8
15	HCL	15
16	HP	8
17	Hilkop	8
18	Kenstar	8
19	LG	17
20	Lenovo	4
21	Lexibook	7
22	Logitech	7

No.	Label	Count
23	Microsoft	6
24	Moto	4
25	Nikon	7
26	Nokia	4
27	Onida	17
28	Oster	8
29	Panasonic	11
30	Phillips	7
31	Prestiage	6
32	Samsung	12
33	Segway	6

Fig. 5: Representation of purchased items with frequency count

K-means clustering technique to identify customer behaviours and group the customers in order to improve the customer service. In this regard dataset is provided to the WEKA tool to find the instances in the four specified clusters using Euclidean Distance Similarity measure. The results of K-means algorithm are in Fig.6.

```
=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 238.47697272589647

Cluster centroids:
                              Cluster#
Attribute        Full Data           0           1           2           3
                    (278)        (60)        (84)        (88)        (46)
=============================================================================
age                41.8417     44.0833     30.369     52.8295     38.8478
                  +/-11.296   +/-6.4079   +/-7.7689   +/-6.8753   +/-5.5895

salary          37568.5252  42358.3333  35184.5238  36821.5909  37103.2609
              +/-20938.3793 +/-42070.1182 +/-8915.1049 +/-8517.1558 +/-6986.4106

itemname                LG         Onida          Sony      Panasonic         Godrej
  Acer             4 (  1%)      0 (  0%)      2 (  2%)      1 (  1%)      1 (  2%)
  Apple            8 (  2%)      2 (  3%)      3 (  3%)      2 (  2%)      1 (  2%)
  Asus             4 (  1%)      1 (  1%)      2 (  2%)      1 (  1%)      0 (  0%)
  Bajaj            6 (  2%)      2 (  3%)      1 (  1%)      2 (  2%)      1 (  2%)
  Benq             4 (  1%)      1 (  1%)      2 (  2%)      1 (  1%)      0 (  0%)
  Birla            6 (  2%)      0 (  0%)      2 (  2%)      3 (  3%)      1 (  2%)
  Braun            6 (  2%)      0 (  0%)      3 (  3%)      3 (  3%)      0 (  0%)
  Canon            7 (  2%)      2 (  3%)      2 (  2%)      2 (  2%)      1 (  2%)
  Compaq           8 (  2%)      1 (  1%)      3 (  3%)      2 (  2%)      2 (  4%)
  Dell             4 (  1%)      1 (  1%)      1 (  1%)      1 (  1%)      1 (  2%)


Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       60 ( 22%)
1       84 ( 30%)
2       88 ( 32%)
3       46 ( 17%)
```

Fig. 6: Results of K-means Clustering Technique

The DBSCAN clustering technique is used to improve the customer service and to remove the 48 outliers that are existing in the dataset of 278 instances , which could not be done in K-means.The results of DBSCAN algorithm are shown in Fig.7.

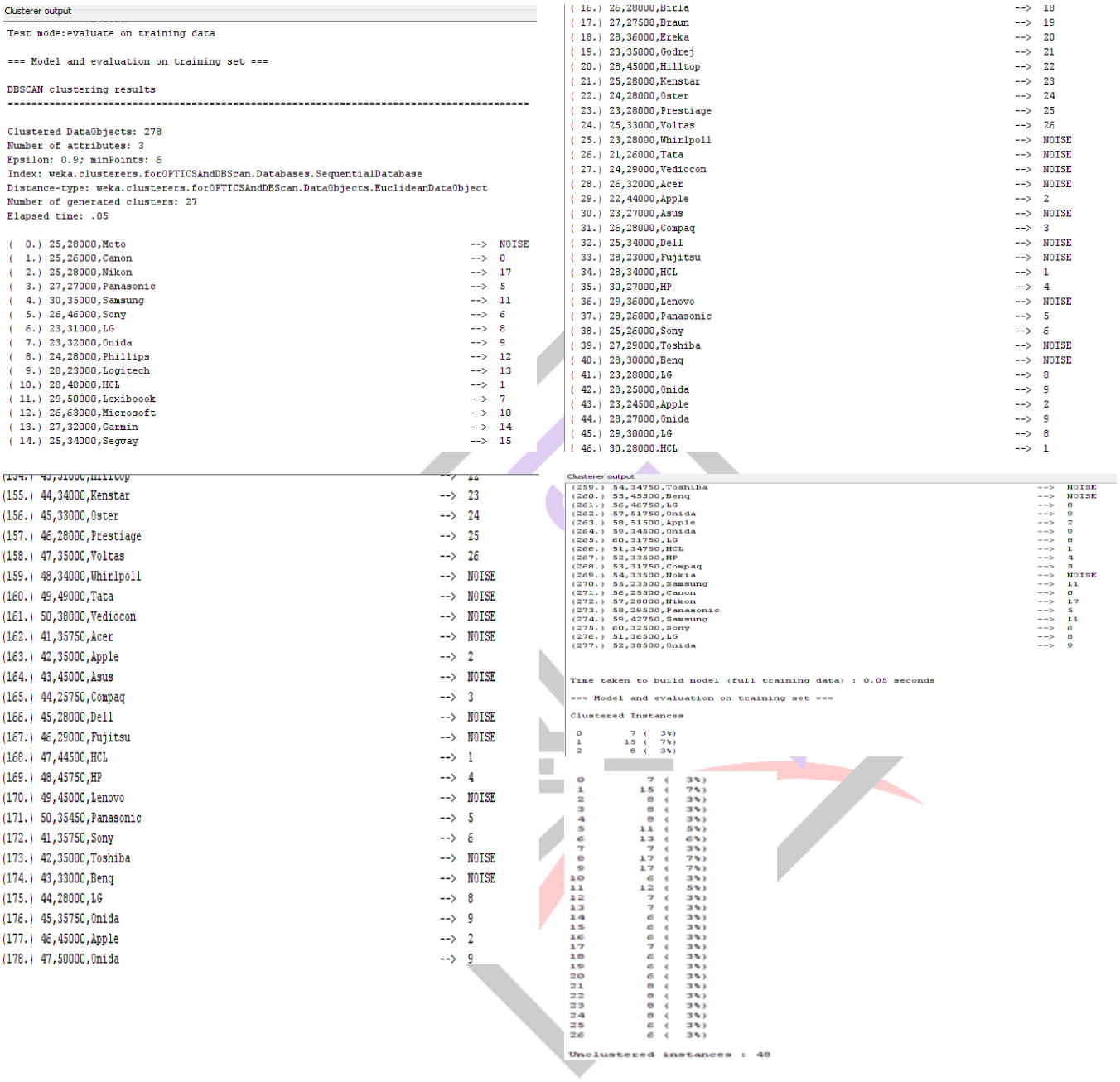

Fig. 7: Results of DBSCAN Clustering Technique

Fig. 8: Visualization of Instances w.r.t. Salary on Y-axis and Item on X-axis

Finally, we could divide the objects into clusters so that we can find frequent item in each cluster to improve sales by contacting customers through message, mail or phone based on the clusters.

IV. CONCLUSION AND DISCUSSIONS

The work discussed here proposes an approach for Customer profiling and segmentation using clustering methods. As it is evident from the results, this system provides better profiling and understanding of customers. A successful profiling and segmentation process demands that a company should define its business objectives. At the start of any segmentation process, management should agree on and clearly state their goals using language that reflects targeting and measurement. Business objectives can be (1) new account, sales, or usage driven; (2) new product driven; (3) profitability driven; or (4) product or service positioning driven. There are three segmentation methods that could be employed: predefined segmentation, statistical segmentation or hybrid segmentation. In this case, the data is known, the work involves a limited number of variables, and a limited number of segments are determined.

REFERENCES

[1] Rajeshri Lanjewar and Om Prakash Yadav, "Understanding of Customer Profiling and Segmentation Using K-Means Clustering Method for Raipur Sahkari Dugdh Sangh Milk Products", International Journal of Research in Computer and Communication Technology, Vol 2, Issue 3, March-2013

[2] Ebtessam Mohamed Desouky & Mahmoud Mohamed El-Khouly, "A Clustering Internet Search Agent for User Assistance", International Journal of Computing Academic Research (IJCAR), ISSN 2305-9184 Volume 1, Number 1 (October 2012), pp. 45-50.

[3] Vahid Golmah and Golsa Mirhashemi, "Implementing A Data Mining Solution To Customer Segmentation For Decayable Products – A Case Study For A Textile Firm", International Journal of Database Theory and Application Vol.5, No. 3, September, 2012.

[4] Mohammad Safari Kahreh & Zahra Safari Kahreh, "An Empirical Analysis to Design Enhanced Customer Lifetime Value Based on Customer Loyalty: Evidences from Iranian Banking Sector", Iranian Journal of Management Studies (IJMS), Vol.5, No.2, July 2012 pp: 145-167.

[5] Mohammad Safari Kahreh, Mohammad Haghighi, and Mostafa Hesan, "How can a Business Best Dealing with Profitable Customers? Analysis a New Model for Customer Lifetime", International Journal of Innovation, Management and Technology, Vol. 2, No. 4, August 2011.

[6] Dr. Sankar Rajagopal, "Customer Data Clustering Using Data Mining Technique", International Journal of Database Management Systems (IJDMS) Vol.3, No.4, November 2011.

[7] Pramod Prasad & Dr. Latesh G. Malik, "Generating Customer Profiles for Retail Stores Using Clustering Techniques", International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 3 No. 6 June 2011, 2506.

[8] D. D. S. Garla and G. Chakraborty, "Comparison of Probabilistic-D and k-Means Clustering in Segment Profiles for B2B Markets," SAS Global Forum 2011, Management, SAS Institute Inc., USA.

[9] H.-B. Wang, D. Huo, J. Huang, Y.-Q. Xu, L.-X. Yan, W. Sun, X.-L. Li, and Jr. A. R. Sanchez, "An approach for improving K-means algorithm on market segmentation," in Proc. International Conference on Systme Science and Engineering (ICSSE), IEEE Xplore, 2010.

[10] Manojit Chattopadhyay, Pranab K Dan, Sitanath Majumdar & Partha Sarathi Chakraborty, "Application of Artificial Neural Network in Market Segmentation : A Review On Recent Trends", 2010.

[11] Parviz Ahmadi &Fardis Samsami, "Pharmaceutical Market Segmentation using GA K-means", European Journal of Economics, Finance and Administrative Sciences , ISSN 1450-2275 Issue 22 (2010)., pp-72-83.

[12] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hard Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245-249, 2009.

[13] S. Dasgupta and Y. Freund, "Random Trees for Vector Quantization," IEEE Trans. on Information Theory, vol. 55, pp. 3229-3242, 2009. [7] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The Planar K-Means Problem is NP-Hard," LNCS, Springer, vol. 5431, pp. 274-285, 2009.

[14] Nan-Chen Hsieh & Kuo-Chung Chu, "Enhancing Consumer Behavior Analysis by Data Mining Techniques" International Journal of Information and Management Sciences, 20 (2009), 39-53 , "M20N14" — 2009/2/17 — 1:00