SEGMENTATION OF CUSTOMERS FOR A MESSAGE TRANSMISSION SYSTEM

[1]Dhaval N. Gajjar
M.E.(EC-pursuing), B.Tech.(EC)
Sarvajanik College of Engineering & Technology, Surat, 395001, India
Email:[1]dngajjar.007@gmail.com

***Abstract*—The magnitude of data which has accumulated by Telecommunications Company is so huge that manual analysis of data is not feasible. However, this data holds valuable information that can be applied for operational and strategical purposes. Therefore, in order to extract such information from this data, automatic analysis is essential, by means of advanced data mining techniques. This paper will address the question how to perform customer segmentation with data mining techniques. In our context, 'customer segmentation' is a term used to describe the process of dividing customers into homogeneous groups on the basis of shared or common attributes. Having this component, managers can decide which marketing actions to take for each segment. In this paper, the customer segmentation is based on usage call behavior, i.e. the behavior of a customer measured in the amounts of incoming or outgoing communication of whichever form. K-means clustering algorithm is applied for creating customer segments. An optimality criterion is used in order to measure their performance and to find the optimum algorithm.**

I. INTRODUCTION

This paper deals the issue , namely customer segmentation. Customer segmentation is a term used to describe the process of dividing customers into homogeneous groups on the basis of shared or common attributes (habits, tastes, etc.). Customer selection is done by various strategy, i.e. after segmentation is done, particular service message should broadcast to segment of customers which are having higher mean compare to other segment. A detailed description is given below. Having this component, marketers can decide which marketing actions to take for each segment and then allocate scarce resources to segments in order to meet specific business objectives. In sort Customer Segmentation is applied on Customer Database.

A. Customer Segmentation:

A basic way to perform customer segmentation is to define segmentations in advance with knowledge of an expert, and dividing the customers over these segmentations by their best fits. This paper will deal with the problem of making customer segments without knowledge of an expert and without defining the segments in advance. The segmentations will be determined based on (call) usage behavior. To realize this, one data mining technique [2][5], called clustering technique[7], will be tested, validated and compared to each other. In this paper, the principals of the clustering techniques will be described and the process of determining the best technique will be discussed.

Customer segmentation is a term used to describe the process of dividing customers into homogeneous groups on the basis of shared or common attributes (habits, tastes, etc) [5][6]. Segmentation is a way to have more targeted communication with the customers. The process of segmentation describes the characteristics of the customer groups (called segments or clusters) within the data. Segmenting means putting the population in to segments according to their similar characteristics. Customer segmentation is a preparation step for classifying each customer according to the customer groups that have been defined. Various Dataminig techniques can be

applied to perform segmentation. Detail discussion is given in following section.

II. DATABASE AND DATAMINING

A. *Call Detail Record*

Every time a call is placed on the telecommunications network , descriptive information about the call is saved as a call detail record [3][8]. Call detail records include sufficient information to describe the important characteristics of each call. At a minimum, each call detail record will include the originating and terminating phone numbers, the date and the time of the call and the duration of the call. Call detail records cannot be used directly for data mining, since the goal of data applications is to extract knowledge at the customer level, not at the level of individual phone calls. Thus, the call detail records associated with a customer must be summarized into a single record that describes the customer's calling behavior. To define the features, one can think of the smallest set of variables that describe the complete behavior of a customer. Keywords like what, when, where, how often, who, etc. can help with this process. A list of features that can be used as a summary description of a customer based on the calls over time period of a day is given below.

- average call duration
- average # calls received per day
- average # calls originated per day
- % of Roaming call within same operator
- % of Roaming call with different operator
- % of calls to mobile phones
- average # sms received per day
- average # sms originated per day
- international calls per day
- % of local outgoing calls
- # Amount of 2G data used
- # Amount of 3G data used

These features can be used to build customer segments and such segments can be useful to describe a certain behavior of group of customers. For examples, customers who travel more often could be in a different segment then users they travel less. In that case, the segmentation will be based on the % of roaming calls and % of local outgoing calls accordingly.

B. *Dataming*

Various data mining techniques [2][7] that allow extracting unknown relationships among the data items from large data collection that are useful for decision making. The wide-spread use of distributed information systems leads to the construction of large data collections in business, science and on the Web. These data collections contain a wealth of information, which however needs to be discovered. Businesses can learn from their transaction data more about the behavior of their customers and therefore can improve their business by exploiting this knowledge. Web usage information can be analyzed and exploited to optimize information access. Thus data mining generates novel, unsuspected interpretations of data. In practice the two fundamental goals of data mining tend to be [2][6]: prediction and description.

- *Prediction* makes use of existing variables in the database in order to predict unknown or future values of interest. I.e. Classification, Regression, Time series analysis etc.
- *Description* focuses on finding patterns describing the data and the subsequent presentation for user interpretation. The relative emphasis of both prediction and description differ with respect to the underlying application and the technique. I.e. Clustering, Summarization, Association, Sequence Discovery.

Among all this method segmentation lies in Descriptive part and under Segmentation, 'Clustering' Technique is used since it gives segments as a output. Detail discussion on which clustering technique to select is described below.

III. CLUSTERING TECHNIQUE:

A. *Introduction: Clustering analysis*

The objective of cluster analysis [3][7] is the organization of objects into groups, according to similarities among them. Clustering can be considered the most important unsupervised learning method. As every other unsupervised method, it does not use prior class identifiers to detect the underlying structure in a collection of data. A cluster can be defined as a collection of objects which are "similar" between them and "dissimilar" to the objects belonging to other clusters. Fig. 1 shows this with a simple graphical example. In this case the 3 clusters into which the data can be divided were easily identified.

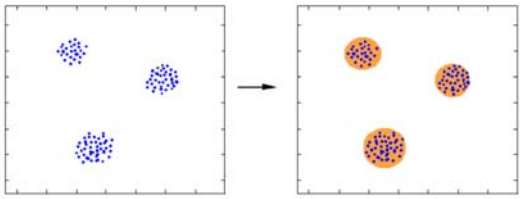

Fig. 1 Clustering

B. The data

The data, as described in above section, are typically summarized observations of a physical process (call behavior of a customer). Each observation of the customers calling behavior consists of n measured values, grouped into an n-dimensional row vector $x_k = [x_{k1}, x_{k2}, ..., x_{kn}]^T$, where $x_k \in R^n$. A set of N observations is denoted by $X = \{x_k | k = 1, 2, ..., N\}$, and is represented as an N x n matrix :

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nc} \end{pmatrix} \quad (1)$$

In pattern recognition terminology, the rows of X are called patterns or objects, the columns are called the features or attributes, and X is called the pattern matrix. In this paper, X will be referred to the data matrix. The rows of X represent the customers, and the columns are the feature variables of their behavior.

C. The Clusters [3][7]

In general, one can accept the definition that a cluster is a group of objects that are more similar to another than to members of other clusters. The term "similarity" can be interpreted as mathematical similarity, measured in some well-defined sense. In metric spaces, similarity is often defined by means of a distance norm, or distance measure. Distance can be measured in different ways. The first possibility is to measure among the data vectors themselves. A second way is to measure the distance from the data vector to some prototypical object of the cluster. The cluster centers are usually not known a priori, and will be calculated by the clustering algorithms simultaneously with the partitioning of the data. The cluster centers may be vectors of the same dimensions as the data objects.

D. Clustering algorithms

1. The K-means algorithm[1]

K-means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The procedure follows an easy way to classify a given N x n data set through a certain numbers of c clusters defined in advance. The K-means algorithm allocates each data point to one of the c clusters to minimize the within sum of squares:

$$\sum_{i=1}^{c} sum_{k \in A_i} ||x_k - v_i||^2 \quad (2)$$

A_i represents a set of data points in the i^{th} cluster and v_i is the average of the data points in cluster i. Note that $||x_k - v_i||^2$ is actually a chosen distance norm. Within the cluster algorithms, v_i is the cluster center of cluster i:

$$v_i = \frac{\sum_{k=1}^{N_i} x_k}{N_i}, x_k \in A_i, \quad (3)$$

where N_i is the number of data points in A_i.

IV. VALIDATION

Cluster validation [4] refers to the problem whether a found partition is correct and how to measure the correctness of a partition. A clustering algorithm is designed to parameterize clusters in a way that it gives the best fit. However, this does not apply that the best fit is meaningful at all. The number of clusters might not be correct or the cluster shapes do not correspond to the actual groups in the data. In the worst case, the data cannot be grouped in a meaningful way at all. One can distinguish two main approaches to determine the correct number of clusters in the data:

- Start with a sufficiently large number of clusters, and successively reducing this number by combining clusters that have the same properties.
- Cluster the data for different values of c and validate the correctness of the obtained clusters with validation measures.

To be able to perform the second approach, validation measures have to be designed. Different validation methods [6] have been used in the paper; however, none of them is perfect by oneself. Therefore, in this paper are used several indexes, which are described below:

1. Partition Coefficient (PC):

It measures the amount of "overlapping" between clusters. It is defined as follows:

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^2 \quad (4)$$

Where u_{ij} is the membership of data point j in cluster i. The main drawback of this validity measure is the lack of direct connection to the data itself. The optimal number of clusters can be found by the maximum value.

2. Classification Entropy (CE):

It measures only the fuzziness of the cluster, which is a slightly variation on the Partition Coefficient.

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij} \log u_{ij} \quad (5)$$

3. Partition Index (SC):

It expresses the ratio of the sum of compactness and separation of the clusters. Each individual cluster is measured with the cluster

validation method. This value is normalized by dividing it by the fuzzy cardinality of the cluster. To receive the Partition index, the sum of the value for each individual cluster is used.

$$SC(c) = \sum_{i=1}^{c} \frac{\sum_{j=1}^{N}(u_{ij})^m \|x_j - v_i\|^2}{N_i \sum_{k=1}^{c} \|x_k - v_i\|^2} \quad (6)$$

PI is mainly used for the comparing of different partitions with the same number of clusters. A minor value of a SC indicates a better partitioning.

4. Separation Index (SI):

In contrast with the partition index (PI), the separation index uses a minimum-distance separation to validate the partitioning.

$$SI(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^2 \|x_j - v_i\|^2}{N \min_{i,k} \|x_k - v_i\|^2} \quad (7)$$

5. Xie and Beni's Index (XB):

It aims to quantify the ratio of the total variation within the clusters and the separations of the clusters.

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^m \|x_j - v_i\|^2}{N \min_{i,j} \|x_j - v_i\|^2} \quad (8)$$

The lowest value of the XB index should indicate the optimal number of clusters.

6. Dunn's Index (DI):

This index was originally designed for the identification of hard partitioning clustering. Therefore, the result of the clustering has to be recalculated.

$$DI(c) = \min_{i \in c} \left\{ \min_{j \in c, i \neq j} \left\{ \frac{\min_{x \in C_i, y \in C_j} d(x,y)}{\max_{k \in c} \left\{ \max_{x,y \in c} d(x,y) \right\}} \right\} \right\} \quad (9)$$

7. Alternative Dunn Index (ADI):

To simplify the calculation of the Dunn index, the Alternative Dunn Index was designed. This will be the case when the dissimilarity between two clusters, measured with $\min_{x \in C_i, y \in C_j} d(x,y)$, is rated in under bound by the triangle-inequality:

$$d(x,y) \geq |d(y,v_j) - d(x,v_j)| \quad (10)$$

Were v_j represents the cluster center of the j^{th} cluster.

$$ADI(c) = \min_{i \in c} \left\{ \min_{j \in c, i \neq j} \left\{ \frac{\min_{x_i \in C_i, x_j \in C_j} |d(y,v_j) - d(x_i,v_j)|}{\max_{k \in c} \left\{ \max_{x,y \in C} d(x,y) \right\}} \right\} \right\} \quad (11)$$

Note, that the Partition Coefficient and the Classification Entropy are only useful for fuzzy partitioned clustering. In case of fuzzy clusters the values of the Dunn's Index and the Alternative Dunn Index are not reliable. This is caused by the repartitioning of the results with the hard partition method.

V. RESULT AND ANALYSIS

A. Database

A sample of CDR data is given below, which is synthetically generated on MATLAB. For a simplicity total four features have been taken and number of customers are 10.

Table 1 synthetically generated CDR data

Sr. No	Average call duration per day	Average # calls received per day	Average # roaming call originated per day	Average # sms originated per day
1	14.09777	4	14	20
2	61.25957	13	8	11
3	29.55091	10	2	8
4	12.07112	6	9	10
5	19.36854	17	4	1
6	27.02982	5	6	17
7	60.13652	3	15	16
8	54.47436	2	7	20
9	20.62533	12	14	16
10	9.975946	20	3	14

B. Visulization

To understand the data and the results of the clustering methods, it is useful to visualize the data and the results. However, the used data set is a high dimensional data set, which cannot be plotted and visualized directly. For simplicity graph is plotted between two features only, (beyond three features it is hard to plot as well as visualize). Example of K-Means algorithms is given below.

1) K-Means

First database (Call Detail Record) of 50 customers was generated synthetically on Matlab and then K-means was applied on them using (refer Fig.2). Points represented using small circle are cluster centers and plot was

between two feature, average # calls received per day and average # roaming call originated per day.

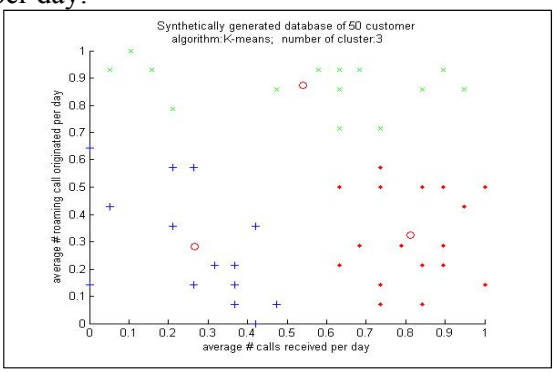

Fig. 2 Visualization of K-means algorithm

C. Cluster Validity:

For case of K-Means algorithm all 7 indices values are plotted and elbow criterion is applied. For all indices, graphs are shown below (fig 3 & 4).

- Now elbow criterion is applied on all indices and optimum number of cluster is found. The elbow criterion says that one should choose a number of clusters so that adding another cluster does not add sufficient information.
- Take a case of Partition Index (Fig.4), it can be found that beyond c=4 there is no significant change in index value, therefore optimum number of cluster for this index is 4.
- Similarly optimum number of cluster is found for all indices and mean value is taken as final optimum number of cluster.
- In this case PC: 7 , CE: 5, PI: 4 , S: 4, XB: 3, DI: 5, ADI: 5
- And mean value is 4.71~5, so optimum number of cluster should be 5.

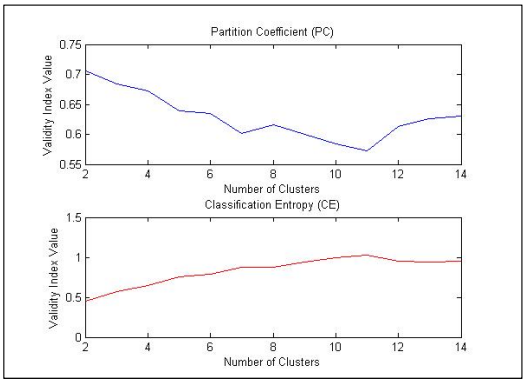

Fig. 3 Plot of PC and CE indices

VI. SUMMARY

The objective of paper was to perform automatic customer segmentation based on usage behavior, without the direct intervention of a human specialist. The customer segments were created by applying K-means clustering algorithms. In this paper, the feature values were selected in such a way that it would describe the customer's behavior as complete as possible. To seek out the optimum number of clusters, elbow criterion was applied. Sadly, this criterion could not always be unambiguously identified. Another problem was that the location of the elbow could differ between the validation measures for the same algorithm.

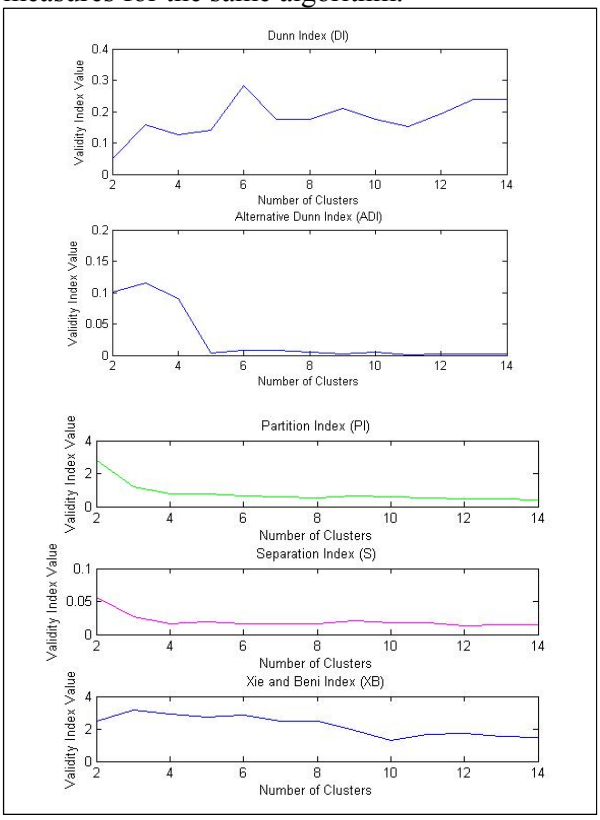

Fig. 4. Plot of remaining 5 validity indices

REFERENCES

[1] Kanungo, Tapas; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y., "An efficient k-means clustering algorithm: analysis and implementation," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.24, no.7, pp.881,892, Jul 2002

[2] RanshulChaudhary, Prabhdeep Singh, Rajiv Mahajan,"A Survey On Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering,Vol. 3, Issue 1, January 2014

[3] Lin, Qining ,"Mobile Customer Clustering Analysis Based on Call Detail Records," Communications of the IIMA: Vol. 7: Iss. 4, Article 11

[4] WeinaWanga,YunjieZhanga ,"On fuzzy cluster validity indices", International Journal of Fuzzy Sets and Systems, Elsevier, Volume 158, Issue 19, 1 October 2007, Pages 2095-2117.

[5] Bascacov, A.; Cernazanu, C.; Marcu, M., "Using data mining for mobile communication clustering and characterization," Applied Computational Intelligence and Informatics (SACI), 2013 IEEE 8th International Symposium on , vol., no., pp.41,46, 23-25 May 2013.

[6] I. O. Folasade, "Computational Intelligence in Data Mining and Prospects in Telecommunication Industry", Journal of Emerging Trends in Engineering and Applied Sciences, vol. 2, no. 4, pp. 601–6051, Aug. 2011.

Books:

[7] P. Berkhin,"A Survey of Clustering Data Mining Techniques", Springer Book-Grouping Multidimensional Data, 2006, pp 25-71, ISBN: 978-3-540-28348-5 (Print) 978-3-540-28349-2 (Online)

[8] Horak, Ray ,"Telecommunications and data communications handbook", Hoboken, N.J.: Wiley-Interscience. pp. 110–111. ISBN 0470127228.