ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

Employing Descriptive Methods for Customer Segmentation

R.Sabitha[1], Dr.S.Karthik[2]

[1]Asst.Professor, Dept of IT, Info Institute of Engineering, Coimbatore, India

[2]Professor &Dean, Dept of CSE, SNS College of Technology, Coimbatore, India

ABSTRACT: Customer segmentation is the process of finding homogenous sub-groups within a heterogeneous aggregate market. This approach is used in direct marketing to target and focus on increasingly well-defined and profitable market segments. The process of segmentation begins with observing customer actions and continues with learning about the demographic and psychographic characteristics of these customers. This intelligence can be made available to the customer facing teams which may be a great tool to increase cross selling and up selling capability of a company. Data Mining is the process of discovering knowledge from huge volumes of data. It is widely used in customer segmentation where various classes can be formed based on the customers buying behaviour. This paper deals with two such methods: K-Means and Hierarchical Clustering. K-Means groups the customers into K-clusters. It is an Iterative algorithm which repeatedly calculates the distance and reforms the centroids based on the distance. The Hierarchical clustering employed here uses divisive method where it begins with just only one cluster that contains all sample data and it slits into two or more clusters that have higher dissimilarity between them. Both the techniques were experimented on NORTHWIND database and the results were analyzed based on the execution time and iteration count. The results concurred that K-Means performs well comparably to Hierarchical clustering.

KEYWORDS: Customer Segmentation, Clustering, K-Means, Hierarchical Clustering.

I. INTRODUCTION

Customer segmentation, also referred to as market segmentation, is the process of finding homogenous sub-groups within a heterogeneous aggregate market. Typically this approach is used in direct marketing to target and focus on increasingly well-defined and profitable market segments. The process of segmentation begins with observing customer actions and continues with learning about the demographic and psychographic characteristics of these customers. Detecting these sub-groups within the market enables an organization to better understand its customers. Learning about clusters within the customer base allows for customized marketing plans to cater specifically to the needs of a particular group. Market segments can be used to find the most profitable groups of customers, allowing the company to focus on maintaining these valuable customers. Another market segment may show a high risk of losing these customers. A cost versus benefit study would help determine how aggressively these customers should be pursued. Generally, a customer database for a marketing study is quite large, possibly containing millions of records and hundreds if not thousands of variables. Due to the size of the data and complexities found within, data mining tasks can be the most appropriate for uncovering information from the data. [1]

Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data. These patterns and trends can be collected and defined as a data mining model.

Mining models can be applied to specific scenarios, such as:

Forecasting: Estimating sales, predicting server loads or server downtime
Risk and probability: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes
Recommendations: Determining which products are likely to be sold together, generating recommendations
Finding sequences: Analyzing customer selections in a shopping cart, predicting next likely events
Grouping: Separating customers or events into cluster of related items, analyzing and predicting affinities. [2].

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

Cluster analysis is commonly used for customer segmentation. In cluster analysis, the goal is to organize observed data into a meaningful structure. This type of analysis is different from traditional statistical approaches such as linear regression in that cluster analysis does not have a dependent variable. Both continuous and categorical variables are used to find sub-groups/clusters. These clusters should consist of observations that are both similar to other members of the group and different from other cluster members. Once clusters are found, characteristics of those clusters can be explored, providing insight into its members, and new observations can be assigned to clusters.[1]

This paper deals with two such techniques: K-Means and Hierarchical Clustering to deal with real world retail database. The comparisons of these techniques are made based on the execution time and the number of iteration to produce the clusters.

The rest of the paper is modeled as follows:Section 2 deals with Clustering Techniques in customer segmentation, Section 3 and 4 focuses on K-Means and Hierarchical Clustering, the experimental results are explained in Section 5 with the Conclusion and future enhancements in Section 6.

II. CLUSTERING TECHNIQUES IN CUSTOMER SEGMENTATION

a. Overview

Clustering is a statistical technique much similar to classification. It sorts raw data into meaningful clusters andgroups of relatively homogeneous observations. The objects of a particular cluster have similar characteristics and properties but differ with those of other clusters. The grouping is accomplished by finding similarities among data according to characteristics found in raw data [3]. The main objective was to find optimum number of clusters. There aretwo basic types of clustering methods, hierarchical andnon-hierarchical. Clustering process is not one time task but iscontinuous and an iterative process of knowledge discoveryfrom huge quantities of raw and unorganized data [4]. For aparticular classification problem, an appropriate clusteringalgorithm and parameters must be selected for obtainingoptimum results. [5].

Clustering is a type of explorative datamining used in many application oriented areas such asmachine learning, classification and pattern recognition [6].In recent times, data mining is gaining much faster momentumfor knowledge based services such as distributed and gridcomputing. Cloud computing is yet another example offrontier research topic in computer science and engineering.

b. Distance Measure

For clustering method, the most important property is that a *tuple* of particular cluster is more likely to be similar to the other *tuples* within the same cluster than the *tuples* of other clusters.

For classification, the similarity measure is defined as *sim*(*ti*, *tl*), between any two *tuples*, *ti*,*tj*□ □*D*. For a given cluster, *Km* of *N* points {*t*ml, *t*m2 ... *t*mN}, the centroid is defined as the *middle* of the cluster. Many of the clustering algorithms assume that the cluster is represented by centrally located one object in the cluster, called a *medoid*. The radius is the square root of the average mean squared distance from any point in the cluster to the centroid. We use the notation *Mm* to indicate the medoid for cluster *Km*. For given clusters *Ki* and *Kj*, there are several ways to determine the distance between the clusters. A natural choice of distance is Euclidean distance measure [7].

Single link is defined as smallest distance between elements in different clusters given by *dis*(*Ki, Kj*) = *min*(*dist*(*ti*1, *tjm*)) ∀ *til*∈ *Ki* ¢ *Kjand* ∀ *tjm*∈ *Ki* ¢ *Kj*. The complete link is defined as the largest distance between elements in different clusters given by *dis* (*Ki, Kj*) = *max* (*dis*(*til, tjm*)), ∀ *til*∈ *Ki* ¢ *Kjand* ∀ *tjm*∈ *Kj*¢ *Kj*. The average link is the average distance between elements in different clusters. We thus have, *dis*(*Ki, Kj*) = *mean*(*dis*(*til, tjm*)), ∀ *til*∈ *Ki* ∉ *Kj*,∀ *tjm*∈ *Kj*¢ *Kj*. If clusters are represented by centroids, the distance between two clusters is the distance between their respective centroids. We thus have, *dis* (*Ki, Kj*)= *dis* (*Ci, Cj*), where *Ci*and *Cj*are the centroid for *Ki* and *Kj*respectively. If each cluster is represented by its medoid then the distance between the cluster can be defined as the distance between medoids which can be given as *dis* (*Ki* , *Kj*)=*dis* (*Mi, Mj*), where *Mi*and *Mj*are the Medoid for *Ki* and *Kj*respectively. [7]

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

c. Market Segmentation

The market segmentation is a process to divide customers into homogeneous groups which have similar characteristics such as buying habits, life style, food preferences etc. [8]. Market segmentation is one of the most fundamental strategic planning and marketing concepts wherein grouping of people is done under different categories such as the keenness, purchasing capability and the interest to buy. The segmentation operation is performed according to similarity in people in several dimensions related to a product under consideration. The more accurately and appropriately the segments performed for targeting customers by a particular organization, the more successful the organization is in the marketplace.

The main objective of market segmentation is accurately predicting the needs of customers and thereby intern improving the profitability by procuring or manufacturing products in right quantity at time for the right customer at optimum cost. To meet these stringent requirements *k*-means clustering technique may be applied for market segmentation to arrive at an appropriate forecastingand planning decisions [9]. It is possible to classify objects such as brands, products, utility, durability, ease of use etc with cluster analysis [10]. For example, which brands are clustered together in terms of consumer perceptions for a positioning exercise or which cities are clustered together in terms of income, qualification etc. [11].

III. K-MEANS CLUSTERING

a. Methodology

The algorithm is called *k*-means due to the fact that theletter *k* represents the number of clusters chosen. Anobservation is assigned to a particular cluster for which itsdistance to the cluster mean is the smallest. The principalfunction of algorithm involves finding the *k*-means. First, aninitial set of means is defined and then subsequentclassification is based on their distances to the centres [12].Next, the clusters' mean is computed again and thenreclassification is done based on the new set of means. This isrepeated until cluster means don't change much betweensuccessive iterations [13]. Finally, the means of the clustersonce again calculated and then all the cases are assigned to thepermanent clusters.Given a set of observations ($x1$, $x2$, $\ldots\ldots$, xn), where eachobservation xi is a d-dimensional real vector. The *k*-meansclustering algorithm aims to partition the n observations into $k$groups of observations called clusters where $k \leq n$, so as tominimize the sum of squares of distances betweenobservations within a particular cluster [14].

b. Algorithm: k-means

The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- k: the number of clusters,
- D: a data set 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;

A drawback of the k-means algorithm is that the number of clusters k is an input parameter. An inappropriate choice of k may yield poor results. The algorithm also assumes that the variance is an appropriate measure of cluster scatter. [15]

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

IV. HIERARCHICAL CLUSTERING

The divisive type of hierarchical clustering is used in this paper. Divisive algorithms begin with just only one cluster thatcontains all sample data. Then, the single cluster splits into 2 ormore clusters that have higher dissimilarity between them untilthe number of clusters becomes number of samples or asspecified by the user. The following algorithm is one kind ofdivisive algorithms using splinter party method.

Divisive algorithm using splinter party method:

1. Start with one cluster that contains all samples.
2. Calculate diameter of each cluster. Diameter is the maximaldistance between samples in the cluster. Choose one cluster Chaving maximal diameter of all clusters to split.
3. Find the most dissimilar sample x from cluster C. Let x departfrom the original cluster C to form a new independent cluster N(now cluster C does not include sample x). Assign all membersof cluster C to MC.
4. Repeat step 6 until members of cluster C and N do not change.
5. Calculate similarities from each member of MC to cluster Cand N, and let the member owning the highest similarities in MCmove to its similar cluster C or N. Update members of C and N.
6. Repeat the step 2, 3, 4, 5 until the number of clusters becomesthe number of samples or as specified by the user. [16].

V. EXPERIMENTAL ANALYSIS

The K-Means and Hierarchical clustering algorithm were implemented on NORTHWIND retail dataset [17].

a. NORTHWIND Database

The database diagram for NORTHWIND is given in figure 5.1.1.

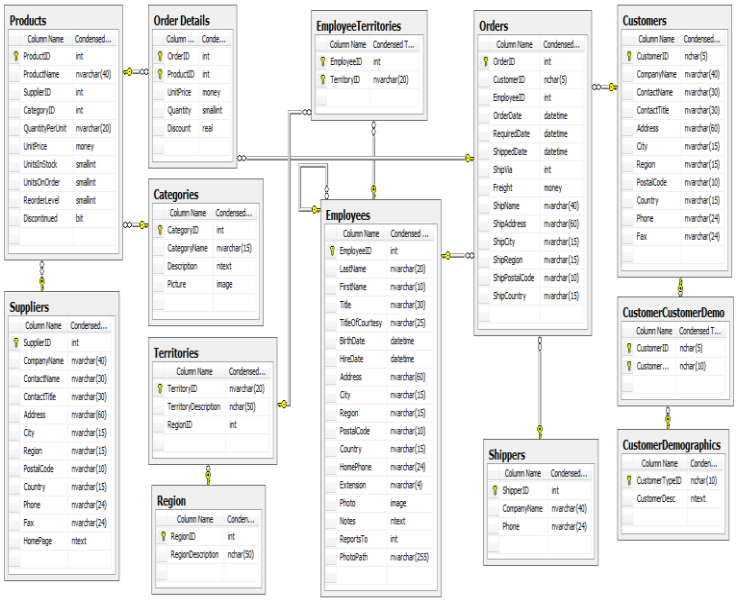

Figure 5.1.1 – Schema of NORTHWIND DB

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

The Database was preprocessed and transformed so that it can be applied for clustering techniques.Total of 2155 records were sub-grouped and consolidated into 830 records. The Customer_ID and Total_purchase_cost were used as major attributes in the experimental process.

b. Experimental Comparison

The K-Means and Hierarchical algorithms were coded in JAVA. The comparisons were made based on the execution time and the number of iteration to produce the clusters.

i. Execution Time

The graph comparing the two algorithms for various Distance / K- values against the execution time is shown in figure 5.2.1. It states that initially for Hierarchical clustering as the distance value increases the execution time increases largely. Later when the distance measure increases beyond a point the execution time confines to a limit and does not change widely. But for K-Means the scenario is different, the execution time increases as the cluster size increases. It is obvious that K-Means works quickly when the size of K is small though both the algorithms take more time to cluster as the cluster size increases.

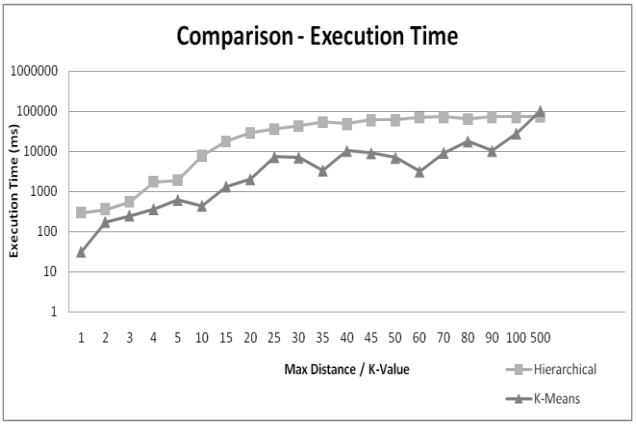

Figure 5.2.1 – Execution Time of both algorithms

ii. Iteration Count

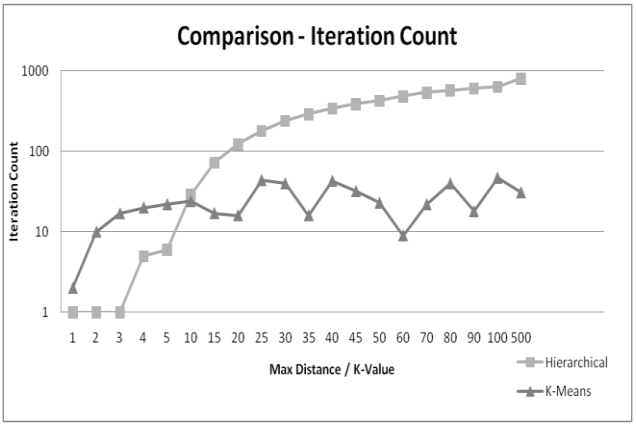

Figure 5.2.2 – Iteration Count of both algorithms

The figure 5.2.2 shows the graph comparing the two algorithms for various Distance / K- values against the iteration count. It is clearly shown that as the distance value increases the iteration count of hierarchical clustering increases exponentially whereas for K-Means it fluctuates around the same range of values. This clearly indicates that K-Means is effective for all possible K-values.

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

The algorithms were also tested using WEKA tool [18]. The algorithms were experimented on NORTHWIND dataset and the results were compared. The result shows that the execution time taken by K-Means is less when compared to Hierarchical clustering. The comparison graph is shown in figure 5.2.3

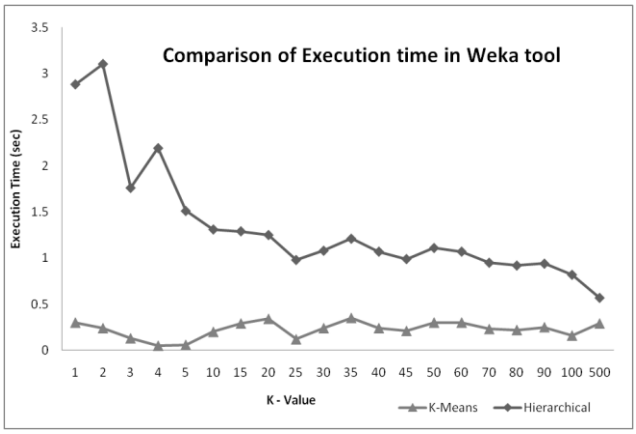

Figure 5.2.3 – Execution time in Weka tool

It is inferred that the K-Means algorithm clusters the data-points based on the distance measure. The data-points are evenly distributed, but this is not the case when hierarchical clustering is used. This scenario is best explained in the results shown in figure 5.2.4 and 5.2.5. The algorithms were tested for K-value = 10 and the cluster assignments were investigated. K-Means proved to be good when compared to Hierarchical in case of cluster assignments

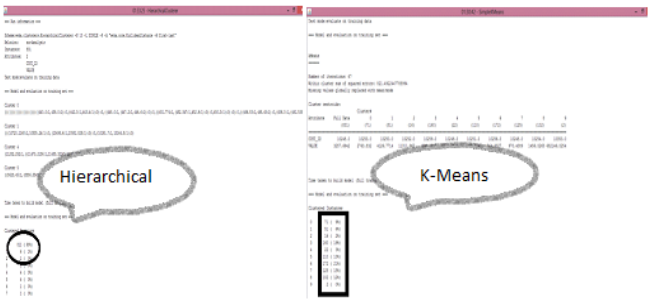

Figure 5.2.4 – Cluster Assignments for both algorithms

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

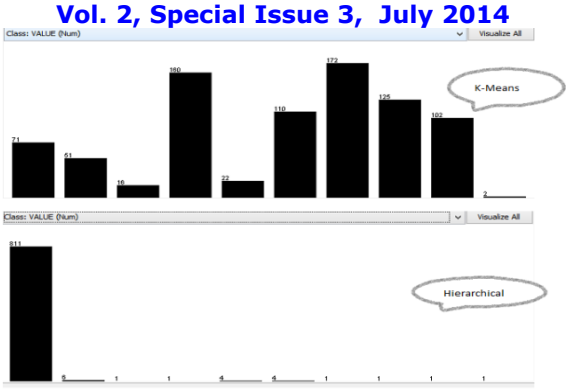

Figure 5.2.4 – Cluster Visualization for both algorithms

VI. CONCLUSION AND FUTURE ENHANCEMENTS

From the experimental results, it is concurredthat K-Means performs well comparably to Hierarchical clustering on NORTHWIND database. K-Means is a popular standard used in clustering. Incorporating business intelligence along with clusteringmechanism to manage a retail business will provide retailers with means tounderstandtheir behaviour and needsand segment customers in a better way. Currently this paper focuses only on segmentation. In the next level enhanced business decisions, profiling and marketing can be done once the clusters are investigated. Standard customer segmentation coupled with few data mining techniques can better predict and model the behavior of the customers. This intelligence can be made available to the customer facing teams which may be a great tool toincrease cross selling and up selling capability of a company.

REFERENCES

[1]https://www.statsoft.com/Textbook/CustomerSegmentation
[2] http://technet.microsoft.com/en-us/library/ms174949.aspx
[3]I. S. Dhillon and D. M. Modha, "Concept decompositions for large sparse text data using clustering," Machine Learning, vol. 42, issue 1,pp. 143-175, 2001.
[4] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R.Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm," IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 24, pp. 881-892, 2002.
[5] MacKay and David, "An Example Inference Task: Clustering,"Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284-292, 2003.
[6] M. Inaba, N. Katoh, and H. Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering," in Proc.10th ACM Symposium on Computational Geometry, 1994, pp. 332-339.
[7] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hard Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245-249, 2009.
[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] H. Hruschka and M. Natter, "Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation," European Journal of Operational Research, Elsvier Science, vol. 114, pp. 346-353, 1999.
[11] P. Ahmadi, "Pharmaceutical Market Segmentation Using GAK- means," European Journal of Economics, Finance and Administrative Sciences, issue 22, 2010.
[12] S. Dasgupta and Y. Freund, "Random Trees for Vector Quantization," IEEE Trans. on Information Theory, vol. 55, pp. 3229-3242, 2009.
[13] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The Planar K-Means Problem is NP-Hard," LNCS, Springer, vol. 5431, pp. 274-285, 2009.
[14] A. Vattani, "K-means exponential iterations even in the plane," Discrete and Computational Geometry, vol. 45, no. 4, pp. 596-616, 2011.
[15] K. Alsabti, S. Ranka, and V. Singh, "An Efficient k-means Clustering Algorithm," Proc. First Workshop High Performance Data Mining, Mar. 1998.
[16] "Hierarchical Clustering Algorithms" in http://home.dei. polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html.
[17]http://northwinddatabase.codeplex.com/
[18] http://www.cs.waikato.ac.nz/~ml/weka