ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

# Cleansing, Profiling & Extracting Customer Data for Developing Clusters to Identify Customers Using Data Mining

**D.Sravya**
B.Tech Student,
Department of CSE,
TKR College of Engineering & Technology.

**G.Vihang**
B.Tech Student,
Department of CSE,
TKR College of Engineering & Technology.

**K.Rakesh Reddy**
B.Tech Student,
Department of CSE,
TKR College of Engineering & Technology.

**B.Ranjitha**
Assistant Professor,
Department of CSE,
TKR College of Engineering & Technology.

## Abstract:

Data mining, an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. By knowing the actual need, preference and purchase trend of customers the marketer can make a future business plan to increase the sale and earn more profit. This paper provides a framework for Cleansing, profiling & extracting customer data for developing clusters to identify customers using Data Mining.

## Keywords:

Customers, Clustering, data mining, Analysis

## Introduction:

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data.

It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

## Data mining

## Data mining involves six common classes of tasks:

1.Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

2.Association rule learning (Dependency modelling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

marketing purposes. This is sometimes referred to as market basket analysis.

3.Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

4.Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

5.Regression – attempts to find a function which models the data with the least error.

6.Summarization – providing a more compact representation of the data set, including visualization and report generation.

Data mining in customer management applications can contribute significantly to the bottom line. Rather than randomly contacting a prospect or customer through a call center or sending mail, a company can concentrate its efforts on prospects that are predicted to have a high likelihood of responding to an offer.

More sophisticated methods may be used to optimize resources across campaigns so that one may predict to which channel and to which offer an individual is most likely to respond (across all potential offers). Additionally, sophisticated applications could be used to automate mailing.

Once the results from data mining (potential prospect/customer and channel/offer) are determined, this "sophisticated application" can either automatically send an e-mail or a regular mail. Finally, in cases where many people will take an action without an offer, "uplift modeling" can be used to determine which people have the greatest increase in response if given an offer.

Uplift modeling thereby enables marketers to focus mailings and offers on persuadable people, and not to send offers to people who will buy the product without an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

## Related Work:

Clustering can be defined as the process of grouping a set of physical or abstract objects into classes of similar objects[2]. Clustering is also called unsupervised classification, because the classification is not dictated/ordered by given class labels. There are many clustering approaches, all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).This chapter identified extensive work on customer segmentation with data mining techniques and elaborates as follows:

Samira et al.(2007) applied segmentation of customers of Trade Promotion Organization of Iran using a proposed distance function which measures dissimilarities among export baskets of different countries based on association rules concepts. Later,in order to suggest the best strategy for promoting each segment,each cluster is analyzed using RFM model. Variables used for segmentation criteria are "the value of the group-commodities", "the type of group-commodities" and "the correlation between export group-commodities". Huang, Chang, and Wu (2009) applied K-means method, Fuzzy C-means clustering method and bagged clustering algorithm to analyze customer value for a hunting store in Taiwan and finally concluded that bagged clustering algorithm outperforms the other two methods.

Pramod et al.(2011) elaborates the use of clustering to segment customer profiles of a retail store.The study concluded that the K-Means clustering allows retailers to increase customer understanding and make knowledge-driven decisions in order to provide personalised and efficient customer service. Hosseini et al. (2010) adopted K-means algorithm to classify the customer loyalty based on RFM values. Cheng and Chen (2009) used K-means and rough set theory to segment customer value based on RFM values. Chen et al. (2009) identified purchasing patterns based on sequential patterns. Migueis.V.L et al.(2012) proposed a method for customers segmentation,given by the nature of the products purchased by customers.This method is based on clustering techniques, which enable segmenting customers according to their lifestyles.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The author segmented customers of an European retailing company according to their lifestyle and proposed promotional policies tailored to customers from each segment, aiming to reinforce loyal relationships and increase sales.Kanwal garg et. al(2008) applied clustering and decision tree techniques for identifying the trend of customer investment behavior in life insurance sector in India.This paper analyzed the prediction of customer buying preference over newly launched policies.

## Cluster analysis:

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, botryology (from Greek βότρυς "grape") and typological analysis. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goals. There is no objectively "correct" clustering algorithm, but as it was noted, "clustering is in the eye of the beholder." The most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally,

unless there is a mathematical reason to prefer one cluster model over another. It should be noted that an algorithm that is designed for one kind of model has no chance on a data set that contains a radically different kind of model. For example, k-means cannot find non-convex clusters.

## Connectivity based clustering:

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

## Centroid-based clustering:

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

## Distribution-based clustering:

The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution. While the theoretical foundation of these methods is excellent, they suffer from one key problem known as overfitting, unless constraints are put on the model complexity.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.

## Data Mining and Customer Management :

Customer relationship management (CRM) is a process that manages the interactions between a company and its customers. The primary users of CRM software applications are database marketers who are looking to automate the process of interacting with customers. To be successful, database marketers must first identify market segments containing customers or prospects with high-profit potential. They then build and execute campaigns that favorably impact the behavior of these individuals. The first task, identifying market segments, requires significant data about prospective customers and their buying behaviors. In theory, the more data the better. In practice, however, massive data stores often impede marketers, who struggle to sift through the minutiae to find the nuggets of valuable information. Recently, marketers have added a new class of software to their targeting arsenal. Data mining applications automate the process of searching the mountains of data to find patterns that are good predictors of purchasing behaviors.

After mining the data, marketers must feed the results into campaign management software that, as the name implies, manages the campaign directed at the defined market segments. In the past, the link between data mining and campaign management software was mostly manual. In the worst cases, it involved "sneaker net," creating a physical file on tape or disk, which someone then carried to another computer and loaded into the marketing database. This separation of the data mining and campaign management software introduces considerable inefficiency and opens the door for human errors. Tightly integrating the two disciplines presents an opportunity for companies to gain competitive advantage.

## The proposed framework has three phases:

1. Data preparation phase

2. Data clustering phase

3. Customer preference analysis

A part of the first phase includes collection of data from data store and the subsequent data cleansing. Second phase generates Behavioral segmentation based clusters and profile of the clusters. Third phase is concerned with identification of customers preferences over products and the risk levels for diseases diagnosed,treated and subsequent process of claim settlement.

## Conclusion:

Clustering technique is critically important step in data mining process. It is a multivariate procedure quite suitable for segmentation customers in the market.For data mining to impact a business, it needs to have relevance to the underlying business process. Data mining is part of a much larger series of steps that takes place between a company and its customers. The way in which data mining impacts a business depends on the business process, not the data mining process. Data mining helps marketing users to target marketing campaigns more accurately; and also to align campaigns more closely with the needs, wants, and attitudes of customers and prospects. In this paper we have implemented and discussed a Cleansing, profiling & extracting customer data for developing clusters to identify customers using Data Mining.

## References:

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

[2] Duc Thang Nguyen, Lihui Chen and Chee Keong Chan, "Clustering with Multi-Viewpoint based Similarity Measure", IEEE Transactions on Knowledge and Data Engineering, 2011.

[3] I. S. Dhillon and D. M. Modha, "Concept decompositions for largesparse 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 clusteringalgorithm," IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 24, pp. 881-892, 2002.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

[5] MacKay and David, "An Example Inference Task: Clustering,"Information Theory, Inference and Learning Algorithms, CambridgeUniversity 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] S. Dasgupta and Y. Freund, "Random Trees for Vector Quantization," IEEE Trans. on Information Theory, vol. 55, pp. 3229-3242, 2009.

[9] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The Planar K-Means Problem is NP-Hard," LNCS, Springer, vol. 5431, pp. 274-285, 2009.

[10] A. Vattani, "K-means exponential iterations even in the plane," Discrete and Computational Geometry, vol. 45, no. 4, pp. 596-616, 2011.

[11] C. Elkan, "Using the triangle inequality to accelerate K-means," in Proc. the 12th International Conference on Machine Learning (ICML), 2003.

[12] H. Zha, C. Ding, M. Gu, X. He, and H. D. Simon, "Spectral Relaxation for K-means Clustering," Neural Information Processing Systems, Vancouver, Canada, vol.14, pp. 1057-1064, 2001.