IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

# Optimizing Data Pattern of Targeted Customers Using Datamining Techniques: A Review

**B.S. Rawat[1*], K. Kumar [2], R.K. Mishra[3]**

[1,2,3]Dept. Of Computer Applications, IFTM University, Moradabad, INDIA

*[*]Corresponding Author: rawat.bhupender@gmail.com*

**Available online at: www.ijcseonline.org**

Accepted: 16/Sept/2018, Published: 30/Sept/2018

***Abstract*—** The trend of shopping in current scenario has changed a lot, with the evolution of online shopping. Companies can thrive their business by maintaining robust relation with their customers, by keeping information of consumer behaviour to provide them personalised services. This can be done by data pre-processing which involves extracting the database about the needs of customer's pattern (i.e. quality, quantity, price and item), followed by data transformation and data mining techniques. But, to extract the discovered patterns in a huge database is still a tedious task, especially in the field of text mining. This paper unfolds various data mining techniques (clustering, classification, decision trees) to discover useful patterns for improving customer's database accuracy and efficiency taking into consideration, their past performance. Further, we try to present an efficient pattern discovery technique by optimizing the original K-means algorithm, which would perform better in global searching and finding the relevant information.

Keywords— Data mining, Literature review, clustering, K-Means.

## I. INTRODUCTION

Data Mining: Data mining is the process of deriving useful information, unusual occurrences and trends in data stored for decision making or survey. Data mining which emerged about thirty years ago, has been developed to deal with large amount of data with multiple attributes. The work by Han and Kamber provides a comprehensive review of the key components in this discipline and a discussion of the solution algorithms and associated computer technologies [1]. It usually involves dividing existing data into test, training and validation data, such that an appropriate model can be built upon the training data to minimize the error and predict with accuracy the outcome for test data. Thus, the validation dataset is used to check whether the prediction model is within a given error interval, to ensure the precision of the model. To generalize a decision making process for collecting and processing original data fig.(1), shows the process and steps of mining data from a given dataset.

In this paper, we present various data mining algorithms to dig out individual consumers most adaptive products from the customer's transactional data and help enterprise to make better decision for marketing service. The paper is structured as follows: 2) Classical datamining techniques, 3) Literature review, 4) Clustering techniques for data pattern and 5) Conclusion.

## II. CLASSICAL DATAMINING TECHNIQUES

Data mining is the discovery for relationships and patterns that exist in huge database but are 'unseen' among the large amount of data, i.e. relationship between patient data and their medical diagnosis. These associations symbolize important information about the database and it's objects [2].

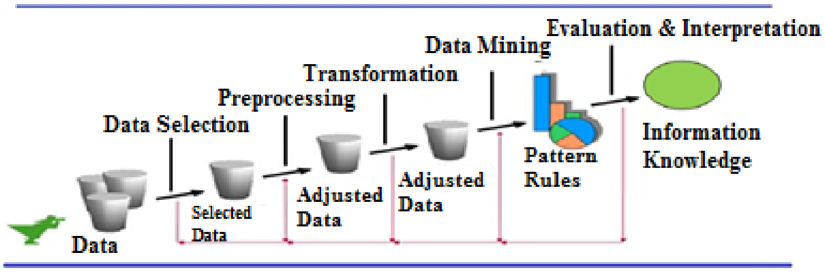

**Fig. (1): Data Mining Model**

A number of key datamining techniques have been developed and used in projects in recent times including association, classification, clustering, prediction and sequential patterns. We look at those data mining techniques with examples to have general idea of them.

**2.1 Association:** It is extraordinary compared to other known datamining technique. In which, a pattern is discovered based on a relationship of a particular item, in the same transaction. For example, the association technique is used to identify the products that customers purchase together in market basket analysis. Businesses can have related marketing campaign to sell maximum products and have maximum profit, based on this data.

**2.2 Classification:** Classification is a standard datamining method in view of machine learning. Basically, classification arranges each item in a set of data into predefined set of classes or groups. Several, mathematical techniques like decision trees, linear programming, neural network and statistics are used by Classification. Classification develops the software that arranges data items into groups.

**2.3 Clustering:** Datamining technique that makes meaningful or useful group of objects and have similar characteristic is known as clustering. To make the concept clearer, we take an example of shopping mall. In a shopping mall, products have a wide range of brands available. The challenge is how to keep those products in a way that customers can take several products in a specific brand without hassle. Hence, we keep the products that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If customers want to grab products in a particular brand, he or she would only go to that shelf instead of looking in the whole shopping mall.

**2.4 Prediction:** Prediction is a datamining technique that derives a relationship between information we know and information we predict for future reference. For e.g. prediction analysis technique can be used in result analysis of student's to predict pass percentage of students for future, if we consider result analysis as an independent variable, pass percentage could be a dependent variable. Then based on the historical result analysis and pass percentage data, we can draw a fitted regression curve that is used for pass percentage prediction.

**2.5 Sequential Patterns:** To find out related patterns in data operation over a business period or transaction is done by sequential patterns analysis. The discovered patterns are used for further business analysis to identify associations between data.

## III. LITERATURE REVIEW

*Y.Y.Yao & Ning Zhong* [3] proposed a linear model which focused on targeted marketing problem characterized by identifying potential new members based on the characteristics of existing members. The interpretation of linear market value functions was adopted from disciplines such as information retrieval, case-based reasoning, and multi-criteria decision making. Various methods for estimating parameters of market value functions were suggested. They are based on probability related interpretations of utility functions, and information-theoretic measures for attributes weighting. Each of the suggested methods seems intuitively appealing, and captures different aspects of our perception of the utilities of attribute values and importance of attributes. The theoretical investigation is only the first step towards the specific type of targeted marketing.

*Mohammed J. Zaki* [4] proposed a new algorithm "SPADE: An Efficient Algorithm for Mining Frequent Sequences" for fast discovery of Sequential Patterns using repeated database scans, and use complex hash structures which have poor locality.SPADE utilizes combinatorial properties to decompose the original problem into smaller sub-problems that can be independently solved in main-memory using efficient lattice search techniques, and using simple join operations. All sequences are discovered in only three database scans. SPADE usually makes three database scans, one for frequent 1-sequences, second for frequent 2-sequences, and third for generating all other frequent sequences. If the support of 2-sequences is available then only one scan is required. SPADE uses simple temporal join operations, for direct integration with DBMS. It also has excellent scale up properties with respect to a number of parameters such as the number of input-sequences, the number of events per input-sequence, the event size, and the size of potential maximal frequent events and sequences. It was observed that simple mining of frequent sequence produces an overwhelming number of patterns, many of them trivial or useless. However, the mined set does contain all potentially useful patterns.

*Hye-Chung et.al* [5] proposed ApproxMAP approximate sequential pattern mining to identify patterns approximately shared by many sequences. ApproxMAP, mine patterns from large sequence databases in two steps i.e first by clustering and then consensus patterns are mined directly from each cluster through multiple alignment. ApproxMAP uses clustering as a preprocessing step to group similar sequences, and then mines the underlying patterns in each cluster directly through multiple alignment. The method is applicable to many interesting problems, such as business analysis, security, and complex bio-sequences analysis.

*Ivancsy Renata and Vajk Istvan* [6] discussed frequent pattern discovery methods in Web log data. The aim of discovering frequent patterns in Web log data is to obtain information about the navigational behavior of the users. The association rule mining was accomplished using the ItemsetCode algorithm with different minimum support and minimum confidence threshold values. Rules generated from msnbc.com data at a minimum support threshold of 0.1% and at a minimum confidence threshold of 85%. This can be used for advertising purposes, for creating dynamic user profiles etc. Three pattern mining approaches i.e. page sets, page sequences and page graphs are investigated from the Web usage mining point of view.

*Zhang & Lu* [7] discussed an efficient approach for data preprocessing for mining Web based customer survey data in order to speed up the data preparation process. The proposed approach is based on a unified data model which is used as a standard representation for the incoming data and derived from analysis of the characteristics of the customer survey data so that it can be mined. It can improve significantly the efficiency of data processing by reducing the number of data transformation from m*n in the traditional way to m+n. It also provides flexibility and adaptability for data processing for different data mining tools.

*Ke Jun Fu* [8] proposed a 5-step data mining model integrated with attribute relevance analysis, decision tree, classification and rules extraction. This datamining model can serve as an efficient vehicle for firms not only to predict the product or services that should be provided or improved for their target customers group, but also to identify the right customers for a specific product family or service. Fu, focused on the datamining methods that can be applied to predict the kind of products or services that will be needed by group of customers. It indicates that this datamining based procedure bypass the limitations and disadvantages of traditional forecasting methods and helps to understand the market and customer more accurately.

*Vijayalakshmi et.al* [9] proposed a new frequent sequence pattern technique called Adaptive Web Access Pattern Tree (AWAPT), for FSP mining. This method is efficient and runs considerably faster than both based WAP Tree and FS-Tree algorithms. It uses the pre-order linking of header nodes to store all events in the same suffix tree loosely together in the linkage, making the search process more efficient. A simple technique for assigning position codes to nodes of any tree has also emerged, which can be used to decide the relationship between tree nodes without repetitive traversals. The AWAPT algorithm quickly determine the suffix of any frequent pattern prefix under consideration, by comparing the assigned binary position codes of nodes of the tree. AWAPT totally eliminates the need to engage in numerous reconstructions of intermediate WAP-trees during mining and considerably reduces execution time.

*Guilllem Lefait* [10] proposed a data mining architecture based on clustering techniques to help experts to segment customer based on their purchase behaviors, researched on a real world data set of 10000 customers over 60 weeks for 6 products and presented a generic architecture to perform customer segmentation on purchase log data by 1) transforming the data, 2) generating diverse models, 3) selecting the most adequate models and 4) creating a visual representation of the segmentations. To increase the segmentation diversity, different clustering algorithms such as density based methods, k-means were used. This architecture has been used on a real world data set and numerous diverse segmentation models have been produced, the best of them have been retained and graphically represented.

*Xiaopin qin* [11] focused on K-means algorithm, based on the triangle inequality theorem in order to improve the efficiency of clustering. The improved algorithm was implemented on Iris Data and compares the performance of two algorithms, showing the effectiveness of the improved method. This algorithm is used to carry out customer segmentation on the customer dataset employed from a company in communication industry. With the continuous increase of customer data amount, improved K-Means algorithm can be used to carry out customer segmentation effectively, so as to provide accurate decision support for enterprises. In future research, other kinds of datasets can be employed to assess this procedure, such as services industry or healthcare industry.

*Nazanin Shahrokhi* [12] used classification techniques in his research, to discover new patterns and improve their performance .Classification is one of the major techniques to discover the patterns in huge amounts of data. Using this approach, the performance patterns can be discovered from an existing data set of the San Francisco airport, including the information of its passengers. This paper presented two concepts : generalization of lifetime values(GLTV) and lifetime values(LTV).The algorithm of classification used in this paper is C5.0, that gave the desired result with exact accuracy and improved performance.

*Ning Zhong et.al* [13] proposed the processes of pattern deploying and pattern evolving, to improve the effectiveness and updating discovered patterns for finding relevant and interesting information. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance. In this research, an effective pattern discovery technique has been proposed which uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms in pure data mining-based methods, concept based model and term-based state-of-the-art models, such as BM25 and SVM-based models.

*Wu and Li* [14] proposed "Pattern-Based Web Mining Using Data Mining Techniques". In this paper a comparison of data mining methods based on the use of several types of discovered patterns is made. The performance of the pattern mining algorithms is investigated on the Reuters dataset RCV1 for completing Web mining tasks. The experimental results show that the closed pattern methods, such as SCPM and NSCPM, have better performance due to the use of pruning mechanism in the pattern discovery stage.

*Pranata Ilung and, SkinnerGeoff* [15] describes the use of machine learning clustering technique to segment and target customers of a wholesale distributor. It describes the selection, analysis, and interpretation of clusters for evaluating customers annual spending on the products. Several clusters were created using k-means clustering algorithm and an in-depth analysis on these clusters were performed using several techniques to carefully select the best cluster. To determine the best number of clusters, three validation measures, i.e. Elbow Method, Davies-Bouldin Index and Silhouette Width, were used. This work can improve datamining and customer segmentation techniques for businesses in retail, logistic, e-commerce and many other areas.

The above reviews are targeting on the customer pattern, sequence, model and behavior for mining the database and extracting the result. Here, we will use some classical techniques of datamining to efficiently extract the customer data from the database.

## IV. CLUSTERING TECHNIQUES FOR DATA PATTERN

Clustering is a process of dividing a group of data objects into smaller groups. Each smaller group is a cluster , such that objects in a cluster/group are same and different from objects in another clusters/groups. The set of clusters emerging from a cluster analysis is known as a clustering[1][16][17]. Clustering techniques can be applied in vast number of application such as data analysis, pattern recognition, image processing and information retrieval. Several clustering techniques along with their application areas are shown in Fig. (2).

**4.1 Partitional** methods conduct one-level partitioning on data sets. In a set of n objects, partitioning method constructs k partitions of data, where each partition represents a cluster and k ≤ n. That is, it divides the data into k groups such that each group must contain at least one object.

**4.2** Hierarchical algorithms create a hierarchical decomposition of the objects. They are either agglomerative (bottom-up) or divisive (top-down):
*4.2.1 Agglomerative* algorithms start with each object being a separate cluster itself, and successively merge groups according to a distance measure. The clustering may stop when all objects are in a single group or at any other point the user wants.

*4.2.2 Divisive* algorithms start with one group of objects and successively split groups into smaller ones, until each object falls in one cluster or as desired.

**4.3 Density-Based** algorithms group objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighborhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the

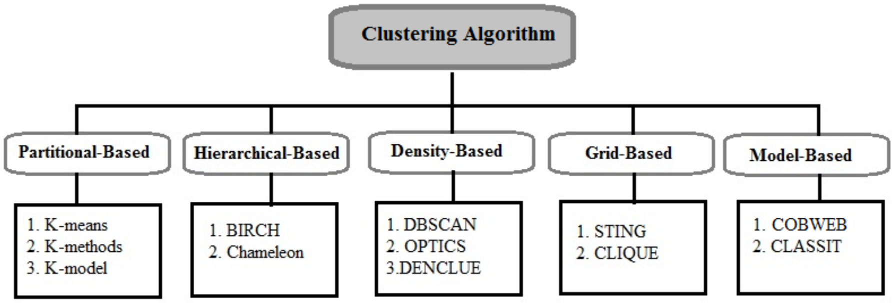

**Fig. (2): Types of Clustering Techniques**

neighborhood exceeds some parameter. This is considered to be different from the idea in partitional algorithms that use iterative relocation of points given a certain number of clusters.

**4.4 Grid-Based** focus on spatial data, i.e., data that model the geometric structure of objects in space, their relationships, properties and operations. These algorithms quantize the data set into a number of cells and then work with objects belonging to these cells. They do not relocate points but rather build several hierarchical levels of groups of objects. Hence, they are closer to hierarchical algorithms but the merging of grids and consequently clusters, does not depend on a distance measure but it is decided by a predefined parameter.

**4.5 Model-Based** clustering algorithms find high-quality approximations of model parameters that best fit the data. They can be either partitional or hierarchical, depending on the structure or model they hypothesize about the data set and the way they refine this model to identify partitionings. They grow particular clusters to improve the preconceived model, hence, closer to density-based algorithms. They start with a fixed number of clusters and do not use the same concept of density.

## V. CONCLUSION

Data mining techniques were used on customer's database focusing on optimization of: marketing, development, identification and targeting customers. We further, found that there was no specific method that would work effectively on the datasets and give us the accurate results.

Based on the results of review papers, possible future works on large data sets of customers are association of products and customer segmentations for cross-selling (selling new products) and up-selling (selling more of what customers currently buy), which can be optimized by using some of the best classical techniques of datamining i.e Association , Classification , Clustering , Prediction and Sequential Patterns. Hence, in our future work we will be using various classical techniques of clustering, such as K-means, density based, classification to improve the performance of customer data pattern in large data sets of online shopping.

## VI.REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pie, "Datamining concepts and techniques", Morgan Kaufmann Series in Data Management Systems,2012.

[2] Sumathi S and Sivanandam S.N "Introduction to Data Mining and Its Applications", Data Mining Book, Springer.

[3] Y.Y. Yao, Ning Zhong, "*Mining Market Value Functions for Targeted Marketing"*, vol. 00, no. , pp. 517, 2001, IEEE ,doi:10.1109/CMPSAC.2001.960662

[4] Mohammed J. Zaki *"SPADE: An Efficient Algorithm for Mining Frequent Sequences"* , in Machine Learning, 42, 31–60, 2001, Kluwer Academic Publishers. Manufactured in The Netherlands.

[5] Hye-Chung Kum, Jian. Pei, Wei. Wang, and Dean Duncan. "*Approx MAP : Approximate Mining of Consensus Sequential Patterns",* Technical Report TR02-031, UNC-CH, 2002.

[6] Renáta Iváncsy "*Frequent Pattern Mining in Web Log Data"*, Vol. 3, No. 1, 2006, Acta Polytechnica Hungarica

[7] N.Zang *"An efficient preprocessing method for mining customer survey data*" ,2007 , IEEE computer society.

[8] Ke-jun Fu "*Using the data mining approach to determine the product preference of target customers*" 2007, IEEE computer society.

[9] S.Vijayalakshmi "*Mining of User's Access Behaviour For Frequent Sequential Pattern from Web Logs*" International Journal of Database Management Systems ( IJDMS ) Vol.2, No.3, August 2010.

[10] Guilllem Lefait "*Customer segmentation architecture based on clustering techniques*", 2010, IEEE computer society.

[11] Xiaopin qin , "*Improved K-means algorithm and application in customer segmentation*" , 2010, IEEE computer society.

[12] Nazanin Shahrokhi ,"*Targeting customers with data mining techniques: classification"*, 2011 IEEE computer society.

[13] Ning Zong, Yuefeng Li, Sheng-Tang Wu, " *Effective Pattern Discovery for Text Mining*", IEEE Transactions on Knowledge and Data Engineering ( Volume: 24, Issue: 1, Jan. 2012 ).

[14] Sheng-Tang Wu , Yuefeng Li , "*Pattern-Based Web Mining Using Data Mining Techniques*" International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 3, No. 2, April 2013, DOI: 10.7763/IJEEEE.2013.V3.215.

[15]Ilung Pranata, Geoff Skinner ,"*Segmenting and targeting customers through clusters selection & analysis*", 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS).

[16]Garima, Hina Gulati and P.K Singh "Clustering techniques in datamining: A Comparison", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).

[17]Anil K. Jain and Richard C. Dubes., "Algorithms for Clustering Data." Prentice-Hall, 1988.

[18]Investigating datamining in Matlab, Douglas Trewartha, November 2006.

[19] WEKA,http://www.cs.waikato.ac.nz/ml/weka.

[20] Rapidminer Studio-V6 manual, 2014 by RapidMiner.

### Author Profile

Mr. B.S. Rawat done MCA from MJP Ruhailkhand University, Bareilly(UP)-INDIA in 2005. He is currently pursuing Ph.D. In Computer Science & Applications from IFTM University Moradabad (UP) INDIA. His main research work focuses on Data Mining, Database Management System, Machine Learning, He has 10 years of teaching experience and 5 years of Research Experience. He has certified OCA from Oracle in the year 2006.