

## **A Novel Enhancing Process Model Using Data Mining Techniques**

**Rekha Arun<sup>1</sup>, Dr. J. Jebamalar Tamilselvi<sup>2</sup>**

### **Abstract**

With an unbridled increase in international and domestic forms of business, Decision Making (DM) has become a leading business strategy in highly competitive business environments. Clustering customers provides an in-depth understanding of their behavior. Clustering is one of the most important and useful technologies in data mining methods. For analyzing the customer behavior, the important attributes in the customer database are first chosen and then they are segmented into groups using clustering algorithm based on those attribute values. In this paper we propose k- means clustering algorithm is used to grouping the customer and support vector machine (SVM) is used to identify the customer behavior. For analyzing the customer behavior, the important attributes in the customer database are first chosen and then they are segmented into groups using clustering algorithm based on those attribute values. This process model was investigated the large amount of data Medical health care based data. Our proposed techniques are providing better solutions and customer satisfaction by taking good decisions.

**Keywords:** Decision Making (DM), K-Means Clustering Algorithm, Support Vector Machine (SVM)

### **Introduction**

Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. The use of data mining techniques in this process is essential to analyze, understand and predict behaviors of an organization [1]. Clustering is one of the important topics in data mining. Clustering is one of the most important research areas in the field of data mining. Clustering means creating groups of objects based on their features in such a way that the objects belonging to the same groups are similar and those belonging to different groups are dissimilar. Clustering is an unsupervised learning technique [2]-[5]. The main advantage of clustering is that interesting patterns and structures can be found directly from very large data sets with little or none of the background knowledge. Clustering algorithms can be applied in many domains.

In a Modern health centers comprise not only doctors, patients and medical staff but also various processes, including the patient's treatment. In recent years modern systems and techniques have been introduced in health-care institutions to facilitate their operations. A huge amount of medical records are stored in databases and data warehouses [6]. Such databases and applications differ from one another. The basic ones store only primary information about patients such as name, age, address, blood type, etc. The more advanced ones let the medical staff record patients' visits and store detailed information concerning their health condition. Some systems also facilitate patients' registration, units' finances and scheduling of visits. Recently a new type of a medical system has emerged: medical decision support system [7]- 19]. It originates in the business intelligence and is to support medical decisions. The data which is stored in such a system may contain valuable knowledge hidden in medical records. Appropriate processing of this information has potential of enriching every medical unit by providing it with experience of many specialists who contributed their knowledge to building the system.

The situation described above is the reason for a close collaboration between computer scientists and medical staff. It aims at development of the most suitable method of data processing which would enable discovering nontrivial rules and dependencies in data. The results may improve the process of diagnosing and treatment as well as reduce the risk of a medical mistake or the time of a diagnosis delivery [10]. This may turn out to be critical especially in emergency incidents. The research area which seeks for methods of knowledge extraction from data is called knowledge discovery or data mining. It utilizes various data mining algorithms to analyze databases.

This paper focuses on clustering in data mining. Data mining adds to clustering the complications of very large datasets with very many attributes of different types. This imposes unique computational requirements on relevant clustering algorithms [14]. A variety of algorithms have recently emerged that meet these requirements and were successfully applied to real-life data mining problems. Therefore it is required to analyze these algorithms to select best solution for Decision Making System.

## **Literature Review**

Decision making comprises a set of processes and enabling systems supporting a business strategy to build long term, maintain patient profiles. It is an important technology in every business because all the businesses are customer centric. It consists of identifying, maintaining, retaining and developing patient details. Identification includes patient data analysis and segmentation.

Data analysis is to analyze the customer characteristics to seek segments of customers [Woo et.al. (2005)]. Customer segmentation is the process of dividing customers into homogeneous groups on the basis of common attributes [Zeling Wang and Xinghui Lei (2010)]. Customer segmentation is typically done by applying some form of cluster analysis to obtain a set of segments [Mirko Bottcher et.al. (2009)] the customer identification is followed by customer attraction which motivates each segment of customers in different way. Customer retention and customer development

deals with retaining the existing customers and maximizing the customer purchase value respectively [Ngai et.al. (2009)]. Segmentation as a technique for forming customer groups for effective targeting is a widely researched area in marketing (Simkin, 2008). Cluster analysis is a popular tool to segment markets. Simply stated, it is a technique for separation of customers into different groups such that each group of customers is collectively different from the customers in the other groups.

In this paper we evaluated an behavior based segmentation methodology using k means clustering algorithm and Support Vector Machine (SVM) data mining techniques for Decision making process.

## **Related Work**

### **1. Clustering**

Two of the most common applications of data mining models are for behavioral segmentation and classification. In behavioral segmentation, clustering models are used to analyze the behavioral patterns of the customers and identify actionable groupings with differentiated characteristics. Classification models are applied to predict the occurrence of an event and estimate the event's propensity. Classification (or propensity) models are typically used to optimize direct communication with customers and that relationship with the customers.

Behavioral segmentation models are based only on the most recent view of the customer. However, since the objective is to identify a segmentation solution founded on consistent and not on random behavioral patterns, the included data should cover a sufficient time period of at least 6 months. Classification models on the other hand, require the splitting of the modeling dataset in different time periods. To identify data patterns associated with the occurrence of an event, the model should analyze the customer profile before the event occurrence. Therefore, analysts should focus on a past moment and analyze the patient view before them taking medicine.

Generally, a customer database for a health care 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 tools can be the most appropriate for uncovering information from the data. Following are descriptions of data mining techniques commonly used for customer and segmentation. K-means clustering algorithm analysis is a technique 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.

## Proposed Work

### 1. K-Means Clustering Algorithm

The categorization of objects into various groups or the partitioning of data set into subsets so that the data in each of the subset share a general feature, frequently the proximity with regard to some defined distance measure is known as Clustering. The clustering problem has been identified in numerous contexts and addressed being proven beneficial in many medical applications. Clustering the medical data into small with meaningful data can aid in the discovery of patterns by supporting the extraction of numerous appropriate features from each of the clusters thereby introducing structure into the data and aiding the application of conventional data mining techniques.

Defines K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. Build to classify or grouping objects based on features into „k“ number of group. K is positive integer number and the grouping is done by mining the sum of squares of distance between data and the corresponding cluster centroid. The cluster centroid is the average point in the multidimensional space defined by the dimensions [9]. There are a lot of applications of the K-mean clustering, range from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligent, image processing, machine vision, etc. In principle, we have several objects and each object have several attributes and we want to classify the objects based on the attributes, then we can apply this algorithm. There are commonly four steps followed for K-mean idea;

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centers.
2. Assign each object to the group that has the closest centered.
3. When all objects have been assigned, recalculate the positions of the K centered.
4. Repeat Steps 2 and 3 until the centered no longer change. This produces a separation of the objects into groups from which the metric to be minimized can be calculated as follows.

$$\sum_{j=1}^k \sum_{i=1}^{n_j} \|x_i^j - c_j\|^2 \dots\dots\dots (1)$$

$$c_j = \frac{\sum_{k=1}^{N_i} x_k}{N_i}, x_k \in A_i \dots\dots\dots (2)$$

K: number of clusters

nj: number of points in jth cluster

xij: ith point in jth cluster

### Procedures Code For K-Means Algorithm

#### Input:

D = {t1,t2,...,tn} // set of elements

A // Adjacency matrix showing distance between elements.

K // Number of desired clusters.

**Output:**

K // Set of clusters.  
 K-means Algorithm:  
 assign initial values for means  $m_1, m_2, \dots, m_k$  ;  
 repeat  
     assign each item  $t_i$  to the cluster which has the closest mean  
     calculate new mean for each cluster;  
 until all the data to be clustered  
 by using this algorithm medical data are segmented. To improve our result using support vector machine.

**2. Behavior Based SVM Analysis**

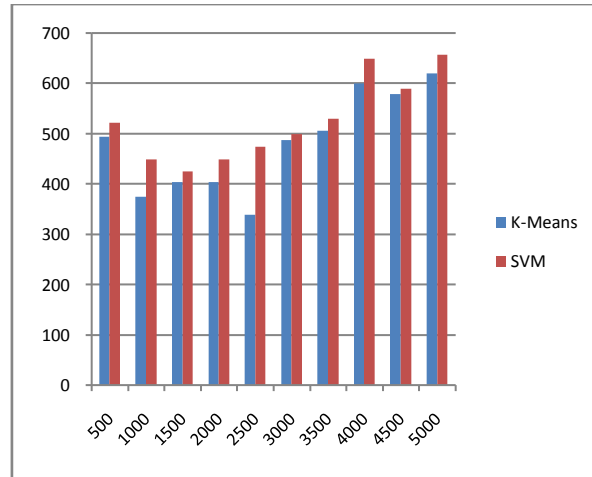
SVM was first introduced by Vapnik [6] and has been very effective method for regression, classification and large data segmentation. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the “best” classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyperplane  $f(x)$  that passes through the middle of the two classes, separating the two.

Once this function is determined, new data instance  $f(x_n)$  can be classified by simply testing the sign of the function  $f(x_n)$ ;  $x_n$  belongs to the positive class if  $f(x_n) > 0$ . Because there are many such linear hyper planes, SVM guarantee that the best such function is found by maximizing the margin between the two classes. Intuitively, the margin is defined as the amount of space, or separation between the two classes as defined by the hyper plane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyper plane. To ensure that the maximum margin hyper planes are actually found, an SVM classifier attempts to maximize the following function with respect to  $a$  and  $b$

$$L_p = -\frac{1}{2} \|\vec{w}\|^2 - \sum_{i=1}^t \alpha_i Y_i \left( \vec{w} \cdot \vec{x}_i + b \right) + \sum_{i=1}^t \alpha_i \dots \dots \dots \quad (3)$$

Where  $t$  is the number of training examples, and  $i, i= 1, \dots, t$ , are non-negative numbers such that the derivatives of  $L_p$  with respect to  $i$  are zero.  $\alpha_i$  are the Lagrange multipliers and  $L_p$  is called the Lagrangian. In this equation, the vectors and constant  $b$  define the hyper plane. A learning machine, such as the SVM, can be modeled as a function class based on some parameters. Different function classes can have different capacity in learning, which is represented by a parameter  $h$  known as the VC dimension. The VC dimension measures the maximum number of training examples where the function class can still be used to learn perfectly, by obtaining zero error rates on the training data, for any assignment of class labels on these points. It can be proven that the actual error on the future data is bounded by a sum of two terms. The first term is the training error, and the second term if proportional to the square root of

the VC dimension  $h$ . Thus, if we can minimize  $h$ , we can minimize the future error, as long as we also minimize the training error, SVM can be easily extended to perform numerical calculations.



**Figure 1: SVM Based Segmentation Data**

Let us put these observations to some use. As mentioned above, training an SVM Behavior classifier will automatically give values for the data weights, number of centers, center positions, and threshold. Can we find the optimal value for that too, by choosing that - which minimizes Figure 1, shows a series of experiments done, with 10,000 training data and 60,000 test data for medical domain. Our proposed techniques are provides well satisfied data segmentation and grouping. The process iterates until all the segmentation problems are solved successfully. Finally our results show behavior based SVM and k-means clustering algorithms are more flexible for large amount of data clustering.

## Conclusion

This paper examines the Segmentation techniques in data mining for decision making and shows the performance of Segmentation among Large amount of medical data. We are analyzing the data using K-means clustering and Support Vector Machine Techniques. Our techniques are provide feasible results in a large amount of data's and comparing our existing approaches behavior based SVM is an more suitable for decision making process. And proposed techniques are active more accuracy comparing than existing approaches.

## References

- [1] Ahmed El Seddawy, Ayman Khedr, Turkey Sultan, "Adapted Framework for Data Mining Technique to Improve Decision Support System in an

- Uncertain Situation”, *International Journal of Data Mining & Knowledge Management Process (IJDKP)* Vol.2, No.2, Jun 2012
- [2] E. Papagergiou, et al., "Data mining: a new technique in medical research," *International Journal of Endocrinology and Metabolism*, pp. 189-191, 2005.
  - [3] J. Chen, et al. (2007). A comparison of four data mining models: bayes, neural network, SVM and decision trees in identifying syndromes in coronary heart disease. 4491/2007
  - [4] Support Vector Machine Solvers L'eon Bottou NEC Labs America, Princeton, NJ 08540, USA Chih-Jen Lin cjlin@csie.ntu.edu.tw Department of Computer Science National Taiwan University, Taipei, Taiwan
  - [5] S. Chao and F. Wong, "An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining", (2009)
  - [6] S. B. Patil and Y. S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", *European Journal of Scientific Research* ISSN 1450-216X, © EuroJournals Publishing, Inc., vol. 31, no. 4, (2009), pp. 642-656.
  - [7] L. Duan, W. N. Street & E. Xu, *Healthcare information systems: data mining methods in the creation of a clinical recommender system*, *Enterprise Information Systems*, 5:2, pp169-181 , 2011.
  - [8] D. S. Kumar, G. Sathyadevi and S. Sivanesh, "Decision Support System for Medical Diagnosis Using Data Mining", (2011)
  - [9] W. L. Zuo, Z. Y. Wanga, T. Liua and H. L. Chenc, "Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach", *Biomedical Signal Processing and Control*, Elsevier, (2013), pp. 364-373.
  - [10] S. W. Fei, "Diagnostic study on arrhythmia cordis based on particle swarm optimization based support vector machine", *Expert Systems with Applications*, Elsevier, vol. 37, (2010), pp. 6748-6752.
  - [11] S. W. Fei, "Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine", *Expert Systems with Applications*, Elsevier, vol. 37, (2010), pp. 6748-6752.
  - [12] S. Belciug, F. Gorunescu, A. Salem and M. Gorunescu, "Clustering-based approach for detecting breast cancer recurrence", *10th International Conference on Intelligent Systems Design and Applications*, (2010)
  - [13] M. E. Celebi, Y. A. Aslandogan and R. P. Bergstresser, "Mining Biomedical Images with Density-based Clustering", *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05)*, (2005).

- [14] Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. “A support vector clustering method”. In International Conference on Pattern Recognition, 2000
- [15] Iyigun, C., & Israel, A. (2010). Semi-supervised probabilistic distance clustering and the uncertainty of classification in A Fink et al., Advances in Data Analysis, Data Handling and Business Intelligence, Studies in Classification, Data Analysis, and Knowledge Organization. Berlin: Springer.