*International Journal of Engineering Research and Modern Education (IJERME)*
*Impact Factor: 6.525, ISSN (Online): 2455 - 4200*
*(www.rdmodernresearch.com) Volume 2, Issue 1, 2017*

# AN ALL-EMBRACING USAGE OF ENHANCED K-MEDOID CLUSTERING ALGORITHM IN INSURANCE PRECINCT

**B. Kalaiselvi**
Assistant Professor, Department of Information Technology, NGM College,
Pollachi, Tamilnadu

**Cite This Article:** B. Kalaiselvi, "An All-Embracing Usage of Enhanced K-Medoid Clustering Algorithm in Insurance Precinct", International Journal of Engineering Research and Modern Education, Volume 2, Issue 1, Page Number 170-175, 2017.

**Abstract:**

The insurance domain has the large amount of data to be presented as useful information. But there are no systematic techniques to produce the information. So, we apply the data mining techniques to mine the data, related to customers, in order to provide the new mined information to the insurance company. Most of the insurance companies have the customer related information as a whole. Here clustering, one of the data mining techniques can be used to group the customer related information into different clusters. Each cluster represents the similar group of customers. For example, in our application, the customers are clustered according to their type of the plan, age, marital status, method of premium payment, monthly income in order to identify the potentiality of the policy holders and to improve the benefits of the plans. There are many attributes that can be used as parameters to cluster the customer related data. Some of the sample attributes are mentioned above. In this implementation the data is clustered by using policy number as a parameter with the help of the enhanced k-medoid clustering algorithm.

**Key Words:** K-Medoid, Clusters, Policy Number & Premium Payment

## 1. Introduction:

Data mining refers to extracting or mining knowledge from large amounts of data. Data mining functionalities are used to specify the kinds of patterns to be found in data mining task. In general, data mining tasks can be classified into two categories, Descriptive and Predictive. Descriptive mining task characterize the general properties of the data in the database. Predictive mining task perform inference on the current data in order to make prediction. The partition clustering techniques partition the database into a predefined number of clusters. They attempt to determine k partitions that optimize a certain criterion function. The partition clustering algorithms are of two types: k-means algorithms and k-medoid algorithms. The hierarchical clustering techniques do a sequence of partitions, in which each partition is nestled into the next partition in the sequence. It creates a hierarchy of clusters from small too big or big too small.

## 2. Related Work:

Moh'd Belal Al- Zoubi [2009][10] proposes method based on clustering approaches for outlier detection is presented. It first performs the PAM clustering algorithm. Small clusters are then determined and considered as outlier clusters. The rest of outliers (if any) are then detected in the remaining clusters based on calculating the absolute distances between the medoid of the current cluster and each one of the points in the same cluster. In this paper, a new proposed method based on clustering algorithms for outlier detection is proposed. It first performs the PAM clustering algorithm. Small clusters are then determined and considered as outlier clusters. The rest of outliers are then found (if any) in the remaining clusters based on calculating the absolute distances between the medoid of the current cluster and each of the points in the same cluster.

Lance Parsons, Ehtesham Haque, Huan Liu, [2004][6] explores Clustering techniques often define the similarity between instances using distance measures over the various dimensions of the data. Subspace clustering is an extension of traditional clustering that seeks to find clusters in different subspaces within a dataset. Traditional clustering algorithms consider all of the dimensions of an input dataset in an attempt to learn as much as possible about each instance described. In high dimensional data, however, many of the dimensions are often irrelevant. These irrelevant dimensions confuse clustering algorithms by hiding clusters in noisy data. In very high dimensions it is common for all of the instances in a dataset to be nearly equidistant from each other, completely masking the clusters. Subspace clustering algorithms localize the search for relevant dimensions allowing them to find clusters that exist in multiple, possibly overlapping subspaces. This paper presents a survey of the various subspace clustering algorithms. We then compare the two main approaches to subspace clustering using empirical scalability and accuracy tests.

Barry Senensky, Jonathan Polon, [2007][12] proposes that most tools used to detect potential fraud in health insurance claims are rules-based. They analyze at the level of a single claim, in isolation, and identify forms of abuse that are either pre-determined or already known. The analysis proved to be very effective in identifying dentists whose portfolio of claims differed significantly from the norm. As demonstration of its effectiveness, the study includes an analysis of a wide-ranging sample of 14 dentists who were identified as having atypical claims activity. As an exploration of how predictive modeling can be used in fraud detection, this study delivered both specific examples and a strong overall indication of the increased capabilities of this approach.

Dilbag Singh, Pradeep Kumar, [2012] [23] proposed that insurance industry contributes largely to the economy therefore risk management in this industry is very much necessary. In the insurance parlance, the risk management is a tool identifying business opportunities to design and modify the insurance products. Risk can have severe impact in case not managed properly and timely. The mapping of risk management with data mining will help organizations to analyze risks and formulate risk mitigation and prevention techniques more efficiently and effectively. This paper aims to study the conceptual mapping of various task of insurance risk management to data mining. A new paradigm has been suggested for insurance risk management using the main attributes and key aspects of data mining.

V. Sree Hari Rao , Murthy V. Jonnalagedda [2012][24] proposed that extraction of customer behavioral patterns is a complex task and widely studied for various industrial applications under different heading viz., customer retention management, business intelligence and data mining. In this paper, authors experimented to extract the behavioral patterns for customer retention in Health care insurance. Initially, the customers are classified into three general categories – stable, unstable and oscillatory. To extract the patterns the concept of Novel index tree (a variant of K- tree) clubbed with K-Nearest Neighbor algorithm is proposed for efficient classification of data, as well as outliers and the concept of insurance dynamics is proposed for analyzing customer behavioral patterns.

## 3. Methodology:

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within in the same cluster and are dissimilar to the objects in other clusters. Partitional Clustering The partition clustering techniques partition the database into a predefined number of clusters. They attempt to determine k partitions that optimize a certain criterion function. The partition clustering algorithms are of two types: k-means algorithms and K-Medoid algorithms. Partitioning algorithms construct partitions of a database on N objects into a set of k clusters. The construction involves determining the optimal partition with respect to an objective function. There is approximately kN/k! Ways of partitioning a set of N data points into k subsets.

**K-Medoid Algorithm:** The basic strategy of K-Medoids clustering algorithms is to find k clusters in objects by first arbitrarily finding a representative object for each cluster. Each remaining object is clustered with the Medoid to which it is the most similar. K-Medoid method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. The algorithm takes the input parameter k, the number if clusters to be partitioned among a set of n objects. A typical KMedoids algorithm for partitioning based on Medoid or central objects is as follows:

**Procedure for K-Medoid Clustering:**

Input: k: number of clusters D: the data set containing n items

Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoids.

Z = ∑ k i-1 ∑ |x-mi | (2) Z: Sum of absolute error for all items in the data set x: the data point in the space representing a data item mi: is the medoid of cluster Ci

Step 1: Arbitrarily choose k data items as the initial medoids.

Step 2: Assign each remaining data item to a cluster with the nearest medoid.

Step 3: Randomly select a non-medoid data item and compute the total cost of swapping old medoid data item with the currently selected non-medoid data item.

Step 4: If the total cost of swapping is less than zero, then perform the swap operation to generate the new set of k-medoids.

Step 5: Repeat steps 2, 3 and 4 till the medoids stabilize their locations.

**The Loop Hole in the Classical K-Medoid Algorithm:** The original k-medoid algorithm stops working when the previously calculated cost is lesser than the currently calculated cost. It is not checking whether the non-selected non-medoids may also be the best medoid and can hold the data point with minimum cost. Because it stops the iteration, when it find the minimum cost from the currently calculated costs.

**Enhanced K-Medoid Algorithm:** This new enhanced k-medoid algorithm assumes all the data points as medoids and calculates the costs for individual points. After calculating the total cost of all the data points it specifies the number of clusters in which the original data to be grouped. Since k-medoid algorithm is an unsupervised algorithm, we specify the number of clusters. The medoids are selected from the data points in which that data point scored the least minimum cost. For example, we need ten clusters, the first 10 least minimum cost points are selected as medoids. This algorithm overcomes the problem of possibility to check all the data points as medoids. Manhattan distance metric is used to calculate the distance between the cluster points.

**Attributes in Insurance Domain:** The sample attributes used in this algorithm for clustering are

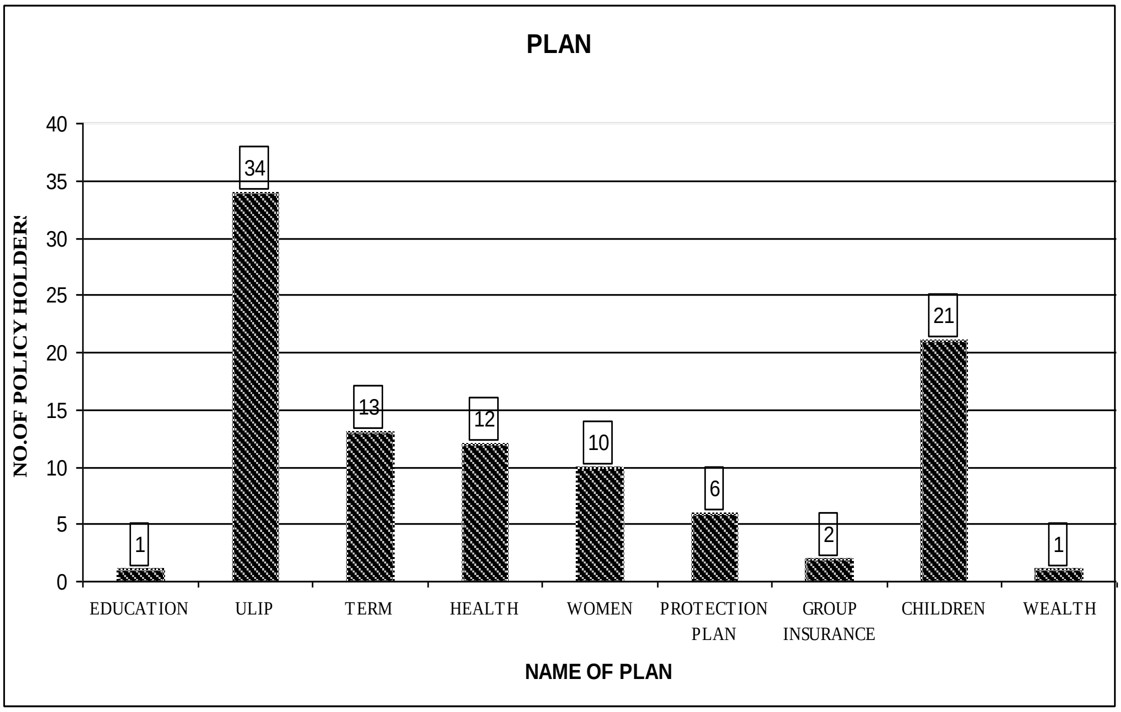
SNO: This field contains the serial number to know about the sequence of data.

Name: This field contains the name of the person, who are already a policy holder for the specified policy.
Sex: This field represents the gender of the person, whether the policy holder is male or female.
Personal Address: This field contains the personal address of the policy holders.
Official Details: This field contains the official address of the policy holders.
Father Name: This field contains father's name of the policy holders.
Age: This field contains age of the policy holders.
Educational Qualification: This field contains the educational qualification of the policy holders.
Marital Status: This field contains the marital status of the policy holders.
Area of Residence: This field contains the area of residence of the policy holders.
Occupation: This field contains the occupation of the policy holders.
Monthly Income: This provides the information about the monthly income of the policy holders.
Policy Number: This provides the unique policy number for each policy holder.
Plan: This gives the information about the various plans available in the organization.
Members Covered: This provides the information about the no. of family members covered in the particular policy.
Period of Policy: This gives the information about the maturity date of the policy.
Sum Assured: This contains the total sum assured at the maturity date.
Premium Amount: This gives the information about the premium amount should be paid.
Method of Payment: This provides the information about the method of payment like cheque, demand draft, or cash etc.
Period of Premium Payment: This gives the information about period of premium payment.

## 4. Results and Discussions:

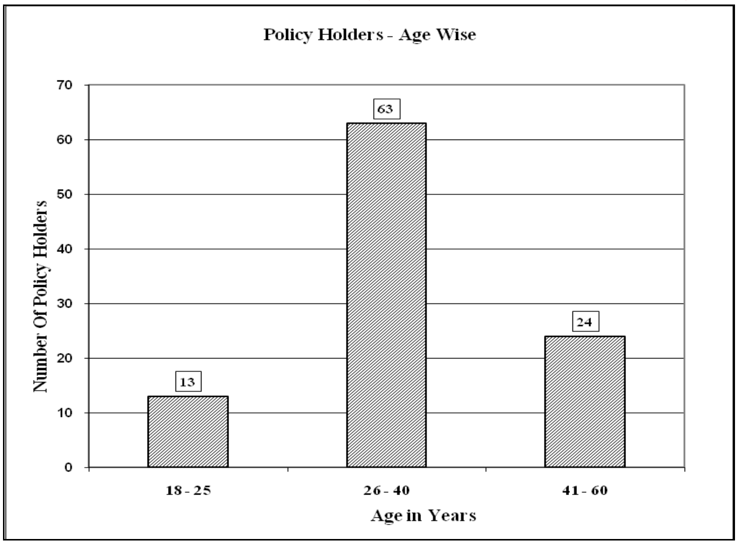
**Analysis of Clustering In Insurance Domain:** The insurance domain has the large amount of data to be presented as useful information. But there are no systematic techniques to produce the information. So, we apply the data mining techniques to mine the data related to customers in order to provide the new mined information to the insurance company. Most of the insurance companies have the customer related information as a whole. Here clustering, one of the data mining techniques can be used to group the customer related information into different clusters. Each cluster represents the similar group of customers. For example, in our application, the customers are clustered according to their type of the plan, age, marital status, method of premium payment, monthly income. There are many attributes that can be used as parameters to cluster the customer related data. Some of the sample attributes are mentioned above. In this implementation the data is clustered by using policy number as a parameter with the help of the enhanced k-medoid clustering algorithm.

Graphical View of Customer's Data



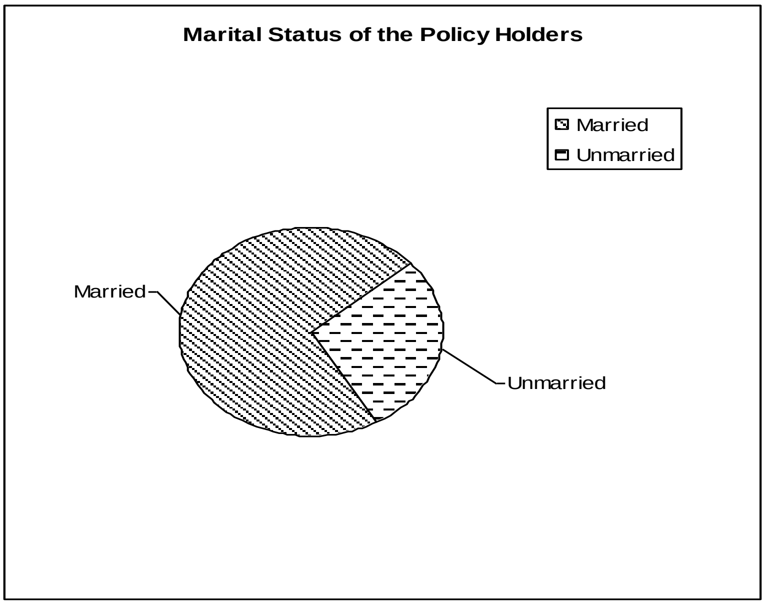
Graph 4.1: No. of Policy holders according to the plan

The following graph represents the range of age of the policy holders.



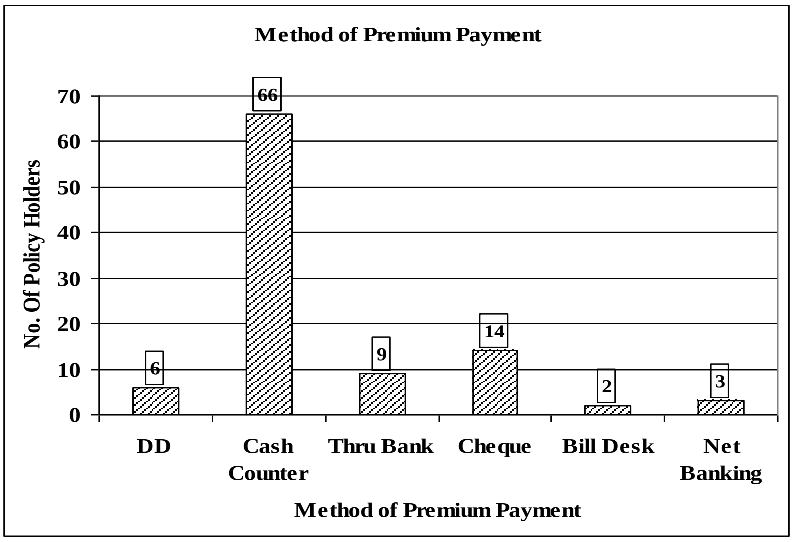
Graph 4.2: Number of policy holders age-wise

The following graph represents the marital status of the policy holders.



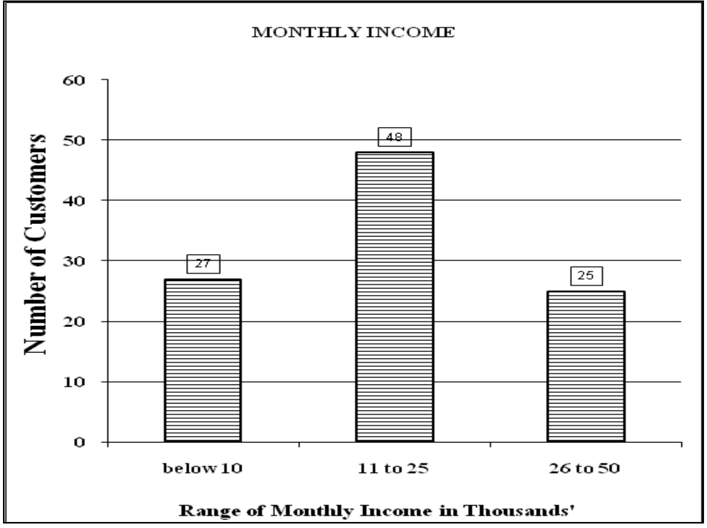
Graph 4.3: Percentage of marital status of policy holders

The following graph represents the method of premium payment done by the policy holders.



Graph 4.4: Method of premium Payment

The following graph represents the monthly income of the policy holders.



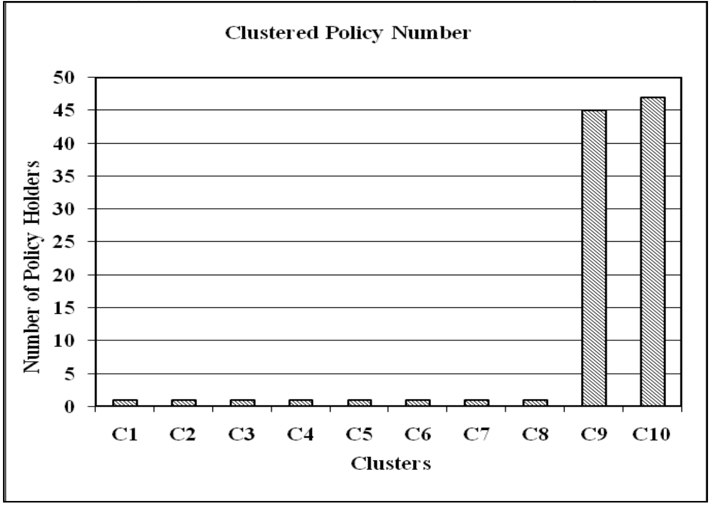
Graph 4.5: Range of Monthly Income of Policy Holders

**Findings Based on Clusters in Insurance Domain:**

In the below table M represents the Medoid.

| Clus. No. | M | DATA POINTS | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| C1 | 50 | 75 | | | | | | | |
| C2 | 51 | 25 | | | | | | | |
| C3 | 49 | 26 | | | | | | | |
| C4 | 52 | 76 | | | | | | | |
| C5 | 48 | 74 | | | | | | | |
| C6 | 53 | 24 | | | | | | | |
| C7 | 47 | 27 | | | | | | | |
| C8 | 54 | 77 | | | | | | | |
| C9 | 46 | 50 | 51 | 49 | 52 | 48 | 53 | 47 | 54 |
| | | 46 | 55 | 45 | 56 | 44 | 57 | 43 | 58 |
| | | 42 | 59 | 41 | 60 | 40 | 61 | 39 | 62 |
| | | 38 | 63 | 37 | 64 | 36 | 65 | 35 | 66 |
| | | 67 | 34 | 33 | 68 | 32 | 69 | 31 | 70 |
| | | 30 | 71 | 29 | 72 | 28 | 73 | | |
| C10 | 55 | 23 | 78 | 22 | 79 | 21 | 80 | 20 | 81 |
| | | 19 | 82 | 18 | 83 | 17 | 84 | 16 | 85 |
| | | 15 | 86 | 14 | 87 | 13 | 88 | 12 | 89 |
| | | 11 | 90 | 10 | 91 | 9 | 92 | 8 | 93 |
| | | 7 | 94 | 6 | 95 | 5 | 96 | 4 | 97 |
| | | 3 | 98 | 2 | 99 | 1 | 100 | | |

The clustered sample policy numbers was presented in the following graphical form.



## 5. Conclusion:

The existing k-medoid clustering algorithm has the limitations in selection of medoids, so that it leads to lacking of accuracy. This was found through the literature review and the algorithm was enhanced through this thesis. It is applied for Insurance sector and found the results are best suite. This algorithm clusters the Insurance policy holders' details with one parameter (Policy Holder number). In future, all the attributes in our insurance domain will be considered as a parameter for this algorithm and it will be helpful for constructing a new plan. It will be implemented in concerned sector.

## 6. References:

1. Lance Parsons, Ehtesham Haque, Huan Liu, "Evaluating Subspace Clustering Algorithms", Supported in part by grants from Prop 301 (No. ECR A601) and CEINT, 2004.
2. Moh'd Belal Al- Zoubi, "An Effective Clustering-Based Approach for Outlier Detection", published in European Journal of Scientific Research ISSN 1450-216X Vol.28 No.2, pp.310-316, 2009.
3. Barry Senensky, Jonathan Polon, "Dental Insurance Claims Identification of Atypical Claims Activity", published in BSc, FSA, April 2007.
4. Dilbag Singh, Pradeep Kumar, "Conceptual Mapping Of Insurance Risk Management To Data Mining" published in International Journal of Computer Applications (0975 – 8887) Volume 39– No.2, February 2012 -13 .
5. V. Sree Hari Rao*, Murthy V. Jonnalagedda ,"Insurance Dynamics – A Data Mining Approach For Customer Retention In Health Care Insurance Industry" published in Bulgarian Academy Of Sciences Cybernetics And Information Technologies • Volume 12, No 1 Sofia • 2012.