Mining Customer's Data for Vehicle Insurance Prediction System using k-Means Clustering - An Application

S. S. Thakur[1], and J. K. Sing[2]

[1]*MCKV Institute of Engineering/Department of Computer Science & Engineering, Liluah, Howrah, Kolkata, West Bengal, 711204, India*
[2]*Jadavpur University/Department of Computer Science & Engineering,, Jadavpur, Kolkata, West Bengal, 700032, India*

subroto_thakur@yahoo.com, jk_koustav@yahoo.com

Abstract **— Data mining or mining customer's data helps to discover the key characteristics from the customer's data, and possibly use those characteristics for future prediction. The problem of selecting the "best" algorithm/parameter setting is a difficult one. However k-Means Clustering is an algorithm helps to classify or to group the objects based on attributes/features into k number of groups. A good clustering algorithm ideally should produce groups with distinct non-overlapping boundaries, although a perfect separation cannot typically be achieved in practice. In this paper, an approach has been made by collecting data samples from customers, and then applying clustering on optimized data for Vehicle Insurance Prediction System. To determine which algorithm is good is a function of the type of data available and the particular purpose of analysis.**

Keywords- **Data Mining, Vehicle Insurance, k – Means Clustering, Prediction, databases.**

I. INTRODUCTION

Vehicle insurance (also known as auto insurance, GAP insurance, car insurance, or motor insurance) is insurance purchased for cars, trucks, motorcycles, and other road vehicles. The specific terms of vehicle insurance vary with legal regulations in each region. To a lesser degree vehicle insurance may additionally offer financial protection against theft of the vehicle and possibly damage to the vehicle, sustained from things other than traffic collisions. Car Insurance is mandatory by law. Comprehensive car insurance protects your car from any man made or natural calamities like terrorist attacks, theft, riots, earth quake, cyclone, hurricane etc in addition to third party claims/damages. There are certain guidelines that should be followed by the Car Insurance buyers while choosing the policy. Car insurance [1] acts like a great friend at the time of crisis.

There are certain general insurance companies who also offer online insurance service for the vehicle. The insurance companies [1, 2] have tie-ups with leading automobile manufacturers. They offer their customers instant auto quotes. Auto premium is determined by a number of factors and the amount of premium increases with the rise in the price of the vehicle. The claims of the Auto Insurance in India can be accidental, theft claims or third party claims. Certain documents are required for claiming Auto Insurance in India, like duly signed claim form, RC copy of the vehicle, Driving license copy, FIR copy, Original estimate and policy copy.

There are different types of Auto Insurance in India:
I. Private Car Insurance
II. Two Wheeler Insurance
III. Commercial Vehicle Insurance

The auto insurance generally includes:

- Loss or damage by accident, fire, lightning, self ignition, external explosion, burglary, housebreaking or theft, malicious act.
- Liability for third party injury/death, third party property and liability to paid driver.
- On payment of appropriate additional premium, loss or damage to electrical/electronic accessories.

The auto insurance does not include:

- Consequential loss, depreciation, mechanical and electrical breakdown, failure or breakage
- When vehicle is used outside the geographical area
- War or nuclear perils and drunken driving.

This paper outlines the implementation of Vehicle Insurance Prediction system using k-Means clustering algorithm.

II. Proposed Approach

Application of k- Means Clustering for prediction of Vehicle Insurance system emphasizes some key areas. These are as follows:

- Our approach for Prediction of Online Vehicle Insurance system has been dealt in section 2.
- Algorithm for Vehicle Insurance Prediction system using k – Means Clustering has been dealt in section 3.
- Implementation Methodology for prediction using k – Means Clustering, its working principle and Pseudocode of traditional k-means has been dealt in section 4.
- Experimental evaluation and Results has been dealt in section 5.
- Design issues and future work has been dealt in section 6.

Figure 1 shows the block diagram of the complete system, in which the prediction to customers for purchasing Online Insurance is explained in detail.

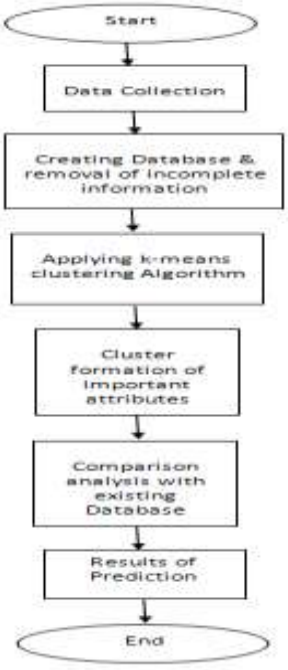

Fig 1. Block diagram of the complete system

In this work, at first we collected data samples/ dataset (Table 1) by asking 4 questions, to owners/drivers of the cars, motorcycles, who parked their vehicles in the parking, in different Shopping Complex namely City Centre I – Salt Lake, City Centre II – Rajarhat, Mani square – Eastern Metropolitan Bye Pass and South City Mall at Jadavpur, Kolkata.

TABLE 1. DATA SET

Vehicle Owner (Y/N)	Qualification	Age	Online Insurance (Y/N)
Yes	Higher Secondary	60	Yes
No	Graduate	60	Yes
No	Madhyamik	30	Yes
Yes	Graduate	60	Yes
No	Post Graduate	55	No
No	Graduate	20	Yes
Yes	Post Graduate	60	Yes
No	Higher Secondary	40	No
No	Graduate	35	Yes
No	Higher Secondary	45	No

The dataset shown in Table 1, contains 4 attributes namely Vehicle Owner (Yes/No), Qualification, Age and if Vehicle Owner are interested in Online Insurance (Yes/No). This dataset is very important as the prediction in this work is concerned.

III. Algorithm – Vehicle Insurance Prediction System

Step 1: Collect the data set
Step 2: Creation of data base from the collected data
Step 3: Apply Optimization technique for Step 2, to remove inconsistency
Step 4: Formation of modified database
Step 5: Apply k-means algorithm for Clustering of Selective Attributes on modified database
  a: Accept the number of clusters to group data into and the dataset to cluster as input values
  b: Initialize the first K clusters
  - Take first k instances or
  - Take Random sampling of k elements
  c: Calculate the arithmetic means of each cluster formed in the dataset.
  d: K-means assigns each record in the dataset to only one of the initial clusters
  e: Each record is assigned to the nearest cluster using a measure of distance (e.g. Euclidean distance).
  f: K-means re-assigns each record in the dataset to the most similar cluster and re-

calculates the arithmetic mean of all the clusters in the dataset.

Step 6: Formation of Cluster Set {Cluster1, $Cluster_2$,……, $Cluster_n$ }

Step 7: Repeat Step 8 for i=1 to n, where n is the number of clusters

Step 8: Check If (Cluster(i) = Cluster(i+1)) Then Group Cluster (i) and Cluster (i+1) into same category

Step 9: Apply Predicate Logic to get the final result

IV. IMPLEMENTATION METHODOLOGY

Vehicle Insurance Prediction System: We collected data samples/ dataset by asking 4 questions, to owners/drivers of the cars, who parked their vehicles in the parking, in different Shopping Complex in Kolkata. We asked for 4 attribute namely whether they are vehicle owner or not, their qualification, their age and whether they will opt for Online Insurance if such facility is available to them. Based on the data collected a database is created using MySql with all the information available.

As per our observation we found that Vehicle owners who are females are bit hesitant in telling their age. In those cases we marked our sample collection data sheet and in creation of database all the samples are included. Later on we apply optimization to remove inconsistent or incomplete data, and its becomes our modified database which contains 112 records, and before optimization the dataset size was 176 samples for 4 wheelers. Similarly we had 102 records, before optimization the dataset size was 160 samples for 2 wheelers and all the data are real data. Then we apply k-means algorithm for clustering of selective attributes on modified database. The attributes are qualification and age. In case of qualification we had data samples for Madhyamik, Higher Secondary, Graduate and Post graduate. Similarly we divide the age into 4 age groups namely Age>20 & Age<25, Age>=25 & Age<=30, Age>30 & Age<=45, and Age>45 & Age<=65.

Simply speaking k-Means is an algorithm to classify or to group your objects based on attributes/features into K number of group where K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of k-Means clustering [3, 4] is to classify the data.

Example: Suppose we have 4 objects as your training data point and each object have 2 attributes. Each attribute represents coordinate of the object (Table 2).

TABLE 2. DATA SET EXAMPLE

Object	Attribute1 (X): weight index	Attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Thus, we also know beforehand that these objects belong to two groups of medicine (cluster 1 and cluster 2). The problem now is to determine which medicines belong to cluster 1 and which medicines belong to the other cluster [5]. Each medicine represents one point with two components coordinate. Now we explain how we had applied k-Means clustering to our database. Given a dataset of n data points x_1, x_2, ..., x_n such that each data point is in R^d, the problem of finding the minimum variance clustering of the dataset into k clusters is that of finding k points $\{m_j\}$ (j=1, 2, …, k) in $\mathbf{R}^d$ such that is minimized, where $d(x_i, m_j)$ denotes the Euclidean distance between x_i and m_j. The points $\{m_j\}$ (j=1, 2, …,k) are known as cluster centroids. The problem in Eq.(1) is to find k cluster centroids, such that the average squared Euclidean distance (mean squared error, MSE) between a data point and its nearest cluster centroid is minimized [5,6].

$$1/n \sum_{i=1}^{n} [min_j \; d^2(x_i m_j)] \tag{1}$$

The k-means algorithm provides an easy method to implement approximate solution to Eq.(1). The reasons for the popularity of k-means are ease and simplicity of implementation, scalability, speed of convergence and adaptability to sparse data. The k-means algorithm can be thought of as a gradient descent procedure, which begins at starting cluster centroids, and iteratively updates these centroids to decrease the objective function in Eq.(1). The k-means always converge to a local minimum. The particular local minimum found depends on the starting cluster centroids. The problem of finding the global minimum is NP-complete. The k-means algorithm updates cluster centroids till local minimum is found.

$$1/n \left(\sum_{i=1}^{n} \left(\frac{1}{n \sum_{i=1}^{n} X_i} \right) \right) \tag{2}$$

Before the k-means algorithm converges, distance and centroid calculations are done while loops are executed a number of times, say l, where the positive integer l is known as the number of k-means iterations.

The precise value of l varies depending on the initial starting cluster centroids even on the same dataset. So the computational time complexity of the algorithm is $O(nkl)$, where n is the total number of objects in the dataset, k is the required number of clusters we identified and l is the number of iterations, $k \leq n$, $l \leq n$ [6].

In our work initially we had taken four clusters as shown in Table 3 based on qualification e.g. Madhyamik, Higher Secondary, Graduate and Post graduate and later on reduced the same to two clusters only Madhyamik and Higher Secondary in cluster 1, Graduate and Post graduate in two clusters (Table 4).

TABLE 3: DATA WITH 4 CLUSTERS FOR CARS

Age → / Qualification ↓	Age>20 & Age<25	Age>=25 & Age<=30	Age>30 & Age<=45	Age>45 & Age <=65
Madhyamik	3	1	6	9
Higher Secondary	3	6	8	13
Graduate	6	6	12	16
Post Graduate	2	3	4	14

It can be observed from Table 3, as there are four clusters in which there are total 14 customers having Age>20 & Age<25, 16 customers having Age>=25 and Age<= 30, 30 customers having Age>30 and Age<= 45 and 52 customers having Age>45 and Age<= 65.

TABLE 4: DATA WITH 2 CLUSTERS FOR CARS

Age → / Qualification ↓	Age>=20 & Age<=30	Age>30 & Age<=65
Madhyamik	4	15
Higher Secondary	9	21
Graduate	12	28
Post Graduate	5	18

It can be observed from Table 4, as there are two clusters in which there are total 30 customers having Age>=20 and Age<=30, and total 82 customers having Age>30 and Age<= 65 .

TABLE 5: DATA WITH 4 CLUSTERS FOR MOTORCYCLES

Age → / Qualification ↓	Age>20 & Age<25	Age>=25 & Age<=30	Age>30 & Age<=45	Age>45 & Age <=65
Madhyamik	18	14	6	4
Higher Secondary	13	11	4	1
Graduate	8	6	2	3
Post Graduate	5	3	3	1

It can be observed from Table 5, as there are four clusters in which there are total 44 customers having Age>20 & Age<25, 34 customers having Age>=25 and Age<= 30, 15 customers having Age>30 and Age<= 45 and 9 customers having Age>45 and Age<= 65.

TABLE 6: DATA WITH 2 CLUSTERS FOR MOTORCYCLES

Age → / Qualification ↓	Age>=20 & Age<=30	Age>30 & Age<=65
Madhyamik	32	10
Higher Secondary	24	5
Graduate	14	5
Post Graduate	8	4

It can be observed from Table 6, as there are two clusters in which there are total 78 customers having Age>=20 and Age<=30, and total 24 customers having Age>= 30, and Age<= 65. Then we check for another attribute i.e. whether the customers will go for online insurance or not, which is a very important attribute in our work. Finally we check the last attribute whether the customer I vehicle owner or not which helps us for doing the prediction as shown in Table 7, 8.

V. EXPERIMENTAL EVALUATION & RESULTS

There have been some promising results from applying k-means clustering algorithm with the Euclidean distance measure, where the distance is computed by finding the square of the distance between each scores, summing the squares and finding the square root of the sum [6].

TABLE 7: RESULTS OBTAINED FOR CARS

Data → / Qualification ↓	Total number of customers (irrespective of Age)	Online Insurance		Vehicle Owner	
		Yes (% data)	No (% data)	Yes (% data)	No (% data)
Madhyamik	19	52.63	47.36	27.31	73.68
Higher Secondary	30	53.33	46.66	30.00	70.00
Graduate	40	77.50	22.50	72.50	27.50
Post Graduate	23	78.26	21.74	62.51	34.78

TABLE 8: RESULTS OBTAINED FOR MOTORCYCLES

Data → / Qualification ↓	Total number of customers (irrespective of Age)	Online Insurance		Vehicle Owner	
		Yes (% data)	No (% data)	Yes (% data)	No (% data)
Madhyamik	42	52.40	47.60	71.45	28.55
Higher Secondary	29	58.60	41.40	75.80	24.20
Graduate	19	63.10	36.90	89.40	10.60
Post Graduate	12	66.66	33.34	75.00	25.00

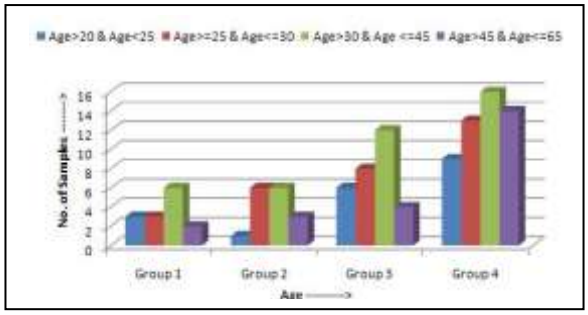

Fig. 2 - Results with 4 clusters for Cars

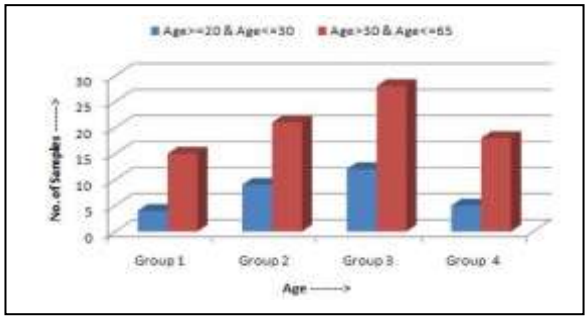

Fig. 3 – Results with 2 clusters for Cars

Further from Fig. 2 we observed that numbers of customers with qualification graduate are dominating ones, irrespective to age criteria. At the same time we observed that age is also an important criteria that requires further attention. From Fig. 3 we observed that, when we reduce the number of clusters from four to two clusters, we found that in addition to qualification the customers having Age>30 and Age <=65 needs special attention.

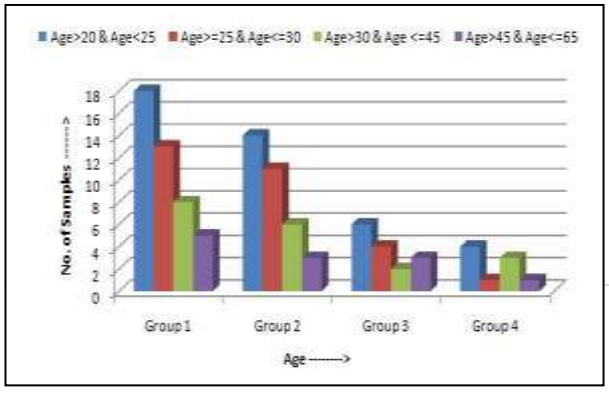

Fig. 4 - Results with 4 clusters for Motorcycles

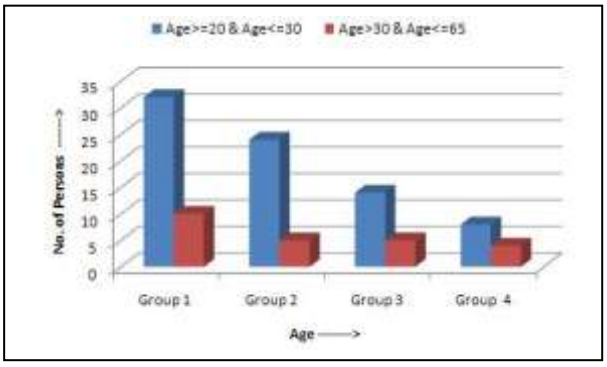

Fig. 5 - Results with 2 clusters for Motorcycles

Further from Fig. 4 we observed that numbers of customers with qualification Madhyamik, Higher Secondary are dominating ones, irrespective to age criteria. At the same time we observed that age is also an important criterion that requires further attention. From Fig. 5 we observed that, when we reduce the number of clusters from four to two clusters, we found that in addition to qualification the customers having Age>=20 and Age<=30 needs special attention.

The results obtained as shown in Table 7 indicates that more than 75% customers with qualification Graduate and Post Graduate are interested in Online Insurance. Additionally we observed that Graduates are higher in numbers as vehicle ownership of 4 wheelers is concerned followed by postgraduates. Similarly results obtained as shown in Table 8 indicates that more than 60% customers with qualification Graduate and Post Graduate are interested in Online Insurance. Additionally we observed that Graduates are higher in numbers as vehicle ownership of 2 wheelers is concerned followed by Higher Secondary as qualification..

Hence if online Insurance System is made available in our country, not just for paying the amount of insurance, but also different insurance schemes provided by the companies, it will benefit both the customers as well as Insurance companies. As customers are concerned they will be having an option/choice to select Insurance for their vehicles at competitive prices. At the same time Insurance Company can tap the customers based on their qualification and age and can make more benefits.

This paper presents k-means clustering algorithm as a simple and efficient tool to do the prediction for

Customers, which enables the customer to purchase Insurance policies with many benefits available for their 4 wheelers i.e. cars, 2 wheelers i.e. motorcycles as shown in Figure 2, 3 and Figure 4, 5. Figure of merit measures (indices) such as the silhouette width or the homogeneity index can be used to evaluate the quality of separation obtained using a clustering algorithm [6]. The concept of stability of a clustering algorithm was considered in [5]. The idea behind this validation approach is that an algorithm should be rewarded for consistency.

VI. Discussion And Conclusion

In this paper, we implemented traditional k-Means clustering algorithm [5] and Euclidean distance measure of similarity was chosen to be used in the analysis of the Insurance Prediction system. We demonstrated our technique using k - Means clustering algorithm. This model improved on some of the limitations of the existing methods, such as model developed by [7] and [8, 9]. Also the research work by [10, 11, 12] only provides Data Mining framework for Students' academic performance.

However, the results obtained from customers data shows that more than 75% customers whose qualification is either graduate and post graduate has shown interest in Online Vehicle Insurance system in case of 4 wheelers i.e. Cars, whereas customers with the same qualification is above 60% who has shown interest in Online Vehicle Insurance system in case of 2 wheelers i.e. motorcycles. We predict and suggest that insurance company can tap the customers of this qualification, and also customers of age between 30 to 65 years for 4 wheelers, and customers of the age between 20 and 30 for 2 wheelers and provide them the services, which will benefit both the customers and also the insurance companies, hence win-win situation for both of them.

Acknowledgment

The authors are thankful to Mr. Reetam Nath, Mr. Monoj Mondal, students of Final Year, CSE Deptt, of MCKV Institute of Engineering, Liluah for their involvement in data collection for the said Research work. The authors are also thankful to Prof. Puspen Lahiri and Prof. Abhisek Saha, Assistant Professor in CSE Deptt, of MCKV Institute of Engineering, Liluah for his valuable suggestions for the implementation of the algorithm in the said research work. The authors are also thankful to Prof. Parasar Bandyopadhyay, Principal, MCKVIE, Liluah for giving permission to use the labs. for carrying out the research work.

References

[1] "What determines the price of my policy?". Insurance Information Institute Retrieved 11 May 2006.

[2] "Am I covered?". Accident Compensation Corporation. Retrieved 23 December 2011

[3] Susmita Datta and Somnath Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data," Bioinformatics, vol. 19, pp.459–466, 2003.

[4] Sharmir R. and Sharan R., "Algorithmic approaches to clustering gene expression data," In current Topics in Computational Molecular Biology MIT Press; pp. 53-65, 2002.

[5] Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., "An efficient enhanced k-means clustering algorithm," Journal of Zhejiang University Science A., pp. 1626–1633, 2006

[6] N. V. Anand Kumar and G. V. Uma, "Improving Academic Performance of Students by Applying Data Mining Technique,"European Journal of Scientific Research, vol. 34(4), 2009.

[7] Varapron P. et al., "Using Rough Set theory for Automatic Data Analysis," 29th Congress on Science and Technology of Thailand, 2003.

[8] Anderberg, M.R., Cluster Analysis for Applications, Academic Press, New York, 1973, pp. 162-163.

[9] J. O. Omolehin, A. O. Enikuomehin, R. G. Jimoh and K. Rauf, "Profile of conjugate gradient method algorithm on the performance appraisal for a fuzzy system," African Journal of Mathematics and Computer Science Research," vol. 2(3), pp. 030-037, 2009.

[10] Cherkassky, V. and Mulier, F. Learning from Data: Concepts, Theory, and Methods. Wiley Interscience, 1998.

[11] Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, 2001

[12] Hand, D. J., Mannila, H., and Smyth, P. Principles of Data Mining. MIT Press, 2001