PREDICTING THE SPENDING CAPACITY OF CUSTOMERS ON CARS USING A COMBINED DATA MINING APPROACH

Manisha Rana*

Geetali Banerji**

Abstract

Data Mining is an analytic process designed to search through a large database in search of consistent patterns or systematic relationships between variables among data. It is an iterative process of selecting, exploring and modeling large amounts of data to identify meaningful, logical patterns and relationships among key variables. It is used to uncover trends, predict future events and assess the merits of various courses of action. In today's scenario there exist various Multinational as well as Indian companies offering a flotilla of cars existing in a wide assortment. It becomes very difficult for the owner of the dealership to identify and project their prospected customers and also to extract the knowledge from the existing voluminous data, which can be used for decision making. This paper is an attempt to analyze various supervised and unsupervised algorithms and to suggest an efficient model that predicts the spending capacity on cars as per the customer's demographic details which further results in future cost reduction as every company now a day is running towards cost cutting and its one of the crucial factor for survival in market in present scenario. The company can make production in accordance with the demand and thereby reducing their maintenance cost. The model is tested using WEKA, a regression and data mining tool. Empirical result shows that the proposed model works better than the traditional model.

Keywords: Data mining, classification, clustering, knowledge discovery in databases (KDD), ZeroR , M5 , Decision Table.

* MCA V Semester, Institute of Information Technology and Management, New Delhi

** Associate Professor, IT Institute of Information Technology and Management, New Delhi

1. Introduction

This paper is related to the prediction of the spending capacity on cars as per customer's demographic details. It analyses various data mining techniques like classification, clustering and then suggests an efficient model which can be proved fruitful for decision making process. It will help the Corporate Executive Board (CEB) of the company to plan for the production of the cars to satisfy the present customer's desires and acquire larger portion of market and thereby adding future customer as well.

This paper discusses about KDD, data mining, data mining algorithm and its application. The aim of this paper is to investigate the performance of different classification or clustering methods for predicting the spending capacity of customer on fleet of cars and suggests a model, which generate efficient rules.

1.1 KDD

The term Knowledge Discovery in Databases (KDD), refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods (Figure 1). It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step. It does this by using data mining methods to extract what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformations of that database.

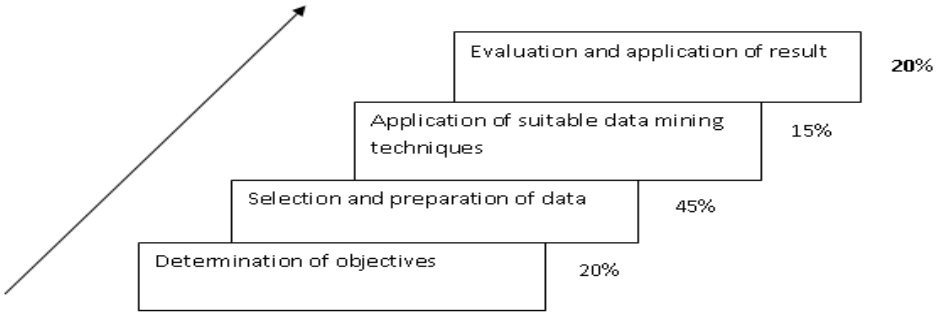

Figure 1 Percentage of total efforts in each step of KDD Process

The process of finding and interpreting patterns from data involves the repeated application of the steps. The steps of KDD process are further explored as follows:

i. Developing an understanding of the application domain ,the relevant prior knowledge, the goals of the end-user

ii. Creating a target data set selecting a data set or focusing on a subset of variables, or data samples, on which discovery is to be performed.

iii. Data cleaning and preprocessing removal of noise or outliers, collecting necessary information to model or account for noise, strategies for handling missing data fields, accounting for time sequence information and known changes.

iv. Data reduction and projection finding useful features to represent the data depending on the goal of the task ,using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration .

v. Choosing the data mining task, deciding whether the goal of the KDD process is classification, regression, clustering, etc.

vi. Choosing the data mining algorithms, selecting methods to be used for searching for patterns in the data, deciding which models and parameters may be appropriate, matching a particular data mining method with complete KDD process.

vii. Data mining, searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering etc.

viii. Interpreting mined patterns.

ix. Consolidating discovered knowledge.

The following section discusses about data mining.

1.2 Data mining

Data mining, a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes 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. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating [2].

Data mining refers to the application of algorithms for extracting patterns from data. It is the exploration and analysis of large quantities of data in order to discover *valid*, *novel*, potentially *useful*, and ultimately *understandable* patterns in data (Figure 2). These terms are defined as follows:

Valid: The patterns hold in general.

Novel: We did not know the pattern beforehand.

Useful: We can devise actions from the patterns.

Understandable: We can interpret and comprehend the patterns.

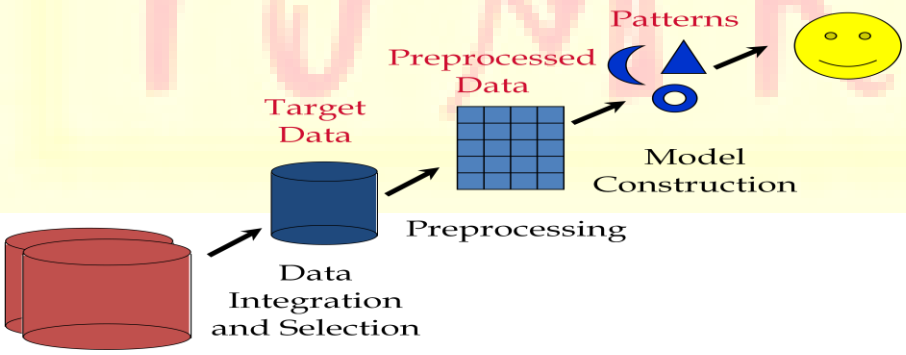

Figure 2 Preprocessing and Mining

1.3 Data mining techniques

Data mining techniques can be broadly classified into following two categories (Figure 3).

- Supervised learning: discover patterns in the data that relate data attributes with a target (class) attribute.

  These patterns are then utilized to predict the values of the target attribute in future data instances.
- Unsupervised learning: The data have no target attribute [3].

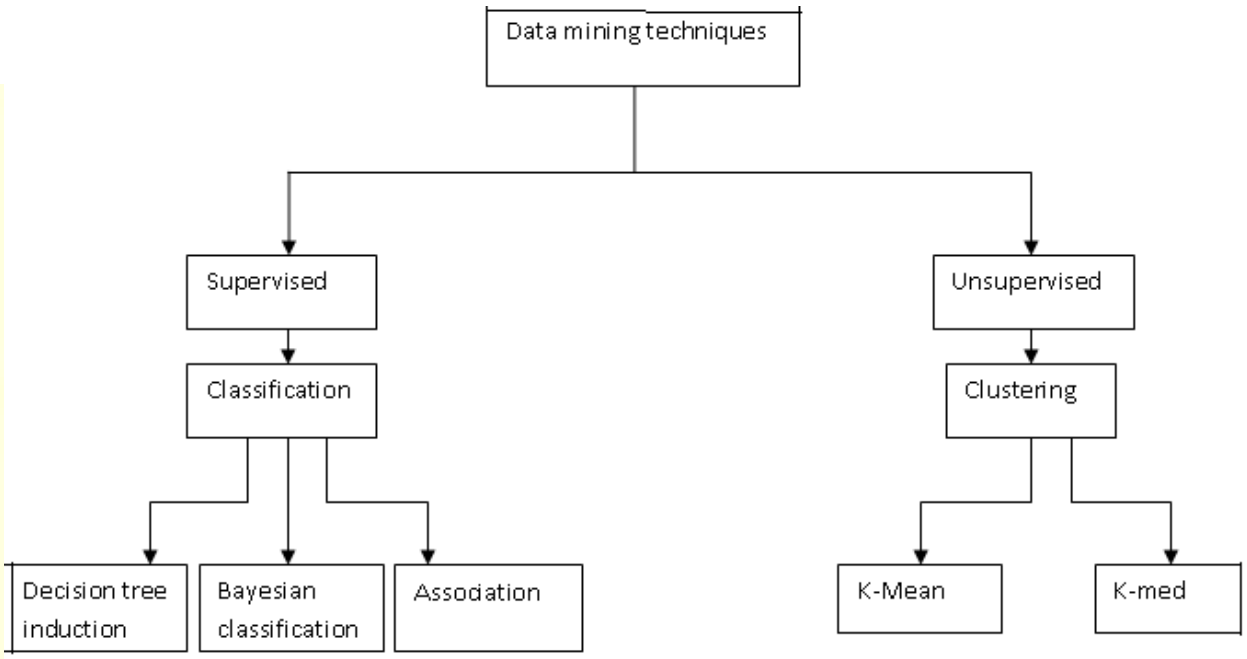

Figure 3 Data mining techniques

1.4 Applications of Data Mining

Some of the most popular areas where data mining is being used is defined as follows:

i. Customer Segmentation

- Understanding of customers by organization
- Technique is used in cluster detection

ii. Market based analysis

- Uncover affinities between product bought together in a retail market
- Customer can be found whom high value item can be sold

iii. Risk management

- Uncover risk associated with potential customers in insurance market

iv. Fraud detection

- Expose fraudulent use of cards based on abnormal customers spending pattern

v. Delinquency tracking

- Track customers who are likely to default on their payments [4].

2. Data

We have collected the primary data of 20 persons using questionnaire. The questionnaire contains details like marital status , sex, age group, region, source of income, income group, no of cars possess and spending capacity on cars in future The data has being collected from various categories, classes like middle level class, business class etc.. It has been further discretized for applying various classification techniques. Table 1 depicts the attributes along with their values.

Table 1 Attributes of Customer

Attribute	Status	Value
Marital status	Married	1
	Unmarried	2
Sex	Male	1
	Female	2
Region	East	1
	West	2
	North	3
	South	4
	NCR	5
Income from	Business	3
	Government	1
	Private	2
Income	Less than three	1
	Between 3-6 lakhs	2
	between 6-10 lakhs	3
	Above 10 lakhs	4
No of cars	One	1
	Two	2
	Three	3
	Doesn't have	0
Future Interest	Less than 10 lakhs	1
	Between 10-20 lakhs	2
	Between 20-50 lakhs	3
	Not intrested	4
Age Group	Below 20	1
	Between 20-35	2
	Between 36-50	3
	Above 50	4

3. Problem Statement

The problem is to model and predict the spending capacity of customers on car keeping account of details like age, source of income, region, marital status, sex, no of cars already possessed and

income range. For this data mining techniques like M5 classifier, ZeroR classifier and decision trees are applied on traditional model (Figure 7) and Proposed model (Figure 8).

4. Concepts Used

4.1. Principle of Data mining using Cluster analysis

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. That is the greater similarity (or homogeneity) within a group, and the greater the difference between groups, the "better" or more distinct the clustering. To better understand the difficulty of deciding what constitutes a cluster, consider figures 4a through 4d, which show twenty points and three different ways that they can be divided into clusters. If we allow clusters to be nested, then the most reasonable interpretation of the structure of these points is that there are two clusters, each of which has three sub clusters. However, the apparent division of the two larger clusters into three sub clusters may simply be an artifact of the human visual system.

Figure a: Initial points.

Figure b: Two clusters.

Figure c: Six clusters

Figure d: Four clusters.

Figure 4 Formations of Clusters

4.2 Principle of Data mining using Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This

approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

4.3 WEKA – testing tool

Waikato Environment for Knowledge Analysis (WEKA) is open source free software available under the GNU General Public License for data mining tasks. It is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. The original non-Java version of Weka was a TCL/TK front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Make file-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Wake include:

- free availability under the GNU General Public License
- portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform
- a comprehensive collection of data preprocessing and modeling techniques
- ease of use due to its graphical user interfaces

Figure 5 WEKA LOGO

4.4 Modeling

In traditional model (Figure 6), we have applied the classification techniques on complete data set. In case of proposed model (Figure 7), we partitioned the complete data set into four clusters. ***Cluster 1***, contains the customers willing to spend less than 10 lakhs, ***Cluster 2***, contains customers who can spend between10 to 20 lakhs , ***Cluster 3***, contains customer who can spend between 20 to 50 lakhs, and ***Cluster 4,*** contains customer who are not interested in spending on cars in future. Followed by the same classification methods and the results were compared.

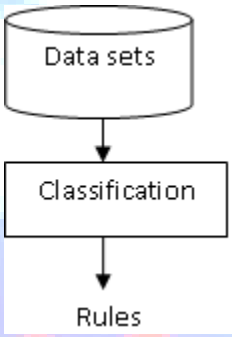

Figure 6 Traditional model

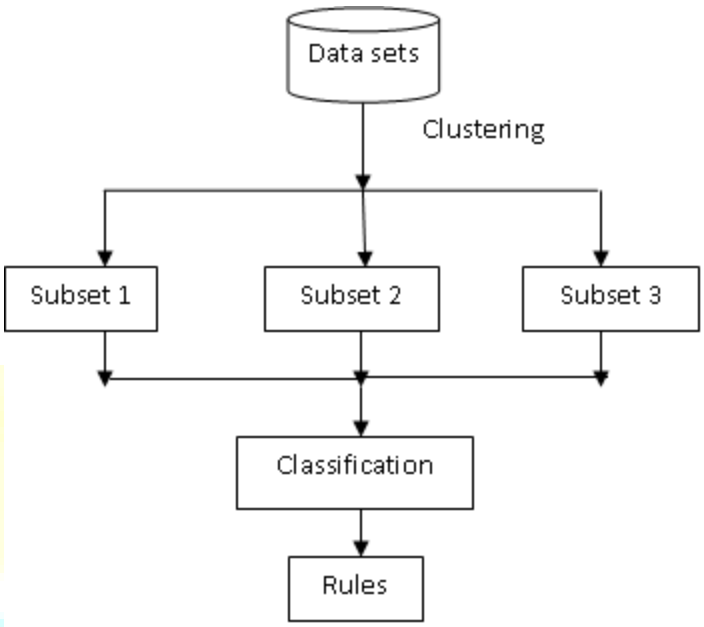

Figure 7 Proposed Model

5. Testing

Testing was conducted separately on ZeroR, M5 (decision tree) and decision table. We have used the complete data set as well as clustered data set (four) for all the methods. The methods are discussed below.

5.1 ZeroR

ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. In WEKA Zero-R is a simple classifier. Zero-R is a trivial classifier, but it gives a lower bound on the performance of a given dataset which should be significantly improved by more complex classifiers. As such it is a reasonable test on how well the class can be predicted without considering the other attributes. It can be used as a Lower Bound on Performance.

5.2 M5

Model trees are a sub-class of regression trees , having linear models at the leaf node. In comparison with classical regression trees, model trees deliver better compactness and prediction accuracy. These advantages issue from the ability of model trees to leverage potential linearity at leaf nodes. The model tree algorithm that is used in this work is based on M5, an optimized,

open-source implementation of the classical M5 algorithm. The input space is recursively partitioned until the data at the leaf nodes constitute relatively homogeneous subsets such that a linear model can explain the remaining variability. This divide-and-conquer approach partitions the training data and provides rules for reaching the models at the leaf nodes. The linear models are then used to quantify, in a statistically rigorous way, the contribution of each attribute (e.g., micro-architectural events here) to the overall predicted value (e.g., performance in this case). A powerful aspect of the prediction model arrived at in this way is that it is interpretable, in contrast with other machine learning approaches, such as neural networks[6].

5.3 Decision Table

Decision table for data set *D* with *n* attributes *A1*, *A2*, ..., *An* is a table with schema *R* (*A1*, *A2*,..., *An* , *Class*, *Sup*, *Conf*). A row *Ri* = (a1i, *a2i*, ..., *ani* , *ci*, *supi*, *confi*) in table *R* represents a classification rule, contains all possible entries in the decision table. We call this table a *candidate decision table*. In addition to the original columns in the data set, the candidate decision table has one more column, *count*, whose value is the number of tuples in the training data covered by the corresponding row in the table. That is, the candidate decision table has schema (*A1*, *A2*, ..., *An*, *Class*, *Count*). The computation in this phase is rather straightforward. Tuples in the training data set are grouped based on their attribute values and class labels. For each grouping, the number of tuples that belong to each class is counted and recorded in the *count* column. For those non-grouping attributes, we use a special value *ANY* in the candidate decision table.

Classification Using Decision Tables

The decision table generated is to be used to classify unseen data samples. To classify an unseen data sample, *u* (*a1u*, *a2u*, ..., *anu*), the decision table is searched to find rows that matches *u*. That is, to find rows whose attribute values are either *ANY* or equal to the corresponding attribute values of *u*. Unlike decision trees where the search will follow one path from the root to one leaf node, searching for the matches in a decision table could result in *none*, *one* or *more* matching rows[7].

One matching row is found:

If there is only one row, *ri* (*a1i*, *a2i*, ..., *ani* , *ci*, *supi*, *confi*) in the decision table that matches *u* (*a1u*, *a2u*, ..., *anu*), then the class of *u* is *ci* .

More than one matching row is found:

When more than one matching rows found for a given sample, there are a number of alternatives to assign the class label. Assume that *k* matching rows are found and the class label, support and confidence for row *i* is *ci*, *supi* and *confi*, respectively. The class of the sample, *cu*, can be assigned in one of the following ways.

(1) Based on confidence and support:

$$c_u = \{c_i | conf_i = max_{j=1}^{k} \quad conf_j\}$$

If there are ties in confidence, the class with highest support will be assigned to *cu*. If there are still ties, one randomly picked from them will be assigned to *cu*.

(2) Based on weighted confidence and support:

$$c_u = \{c_i | conf_i * supi = max_{j=1}^{k} \quad (conf_j * sup_j)\}$$

The ties are treated similarly.Note that, if the decision table is sorted on (*Conf*, *Sup*), it is easy to implement the first method. We can simply assign the class of the first matching row to the sample to be classified. In fact, our experiments indicated that this simple method provides no worse performance than others.

6. Empirical Evaluation

The results generated from various classifiers in case of traditional and proposed models are depicted in Table 2.

Table 2 Results of classifiers on traditional and proposed models

Measures	Proposed Model												Traditional Model		
	Cluster 1			Cluster 2			Cluster 3			Cluster 4			Complete Data set		
	DT	M5	ZeroR	DT	M5	ZeroR	DT	M5	ZeroR	DT	M5	ZeroR	DT	M5	ZeroR
No of Instances	6	6	6	5	5	5	5	5	5	4	4	4	20	20	20
No of Rules	3	1	-	3	1	-	4	1	-	2	1	-	17	1	-
No of Subsets	24	-	-	28	-	-	23	-	-	25	-	-	30	-	-
Merit of best subset	0.14	-	-	0	-	-	0.6	-	-	0	-	-	0.846	-	-
Feature set	4,7	-	-	3,7	-	-	4,5,7	-	-	1,7	-	-	2,3,4,7,8	-	-
Time taken	0	0	-	0	0.02	0	0	0	0	0	0.02	0	0.02	0	0
Correlation Coefficient	1	0	0	1	0	0	1	0.8709	0	1	1	0	0.9898	0.8138	0
Mean Absolute Error	0	0.444	0.444	0	0.32	0.32	0	0.2941	0.64	0	0	0.5	0.05	0.506	0.985
Root Mean Squared Error	0	0.471	0.471	0	0.4	0.4	0	0.3678	0.75	0	0	0.5	0.1581	0.6438	1.108
Relative Absolute Eerror	0	100%	100%	0	100%	100%	0	45.96%	100%	0	0	100%	5.08%	51.37%	100%
Root Relative Squared Error	0	100%	100%	0	100%	100%	0	49.15%	100%	0	0	100%	14.27%	58.11%	100%

From Table 2, it is found that Decision Table is the best method among the rest in this domain with error rate zero and correlation coefficient 1. The result of proposed model is found to be far better than traditional one.

Following are the rules generated from various clusters and complete data set.

Rules

Cluster 1 :

If ((age between 20-50) and (income range less than 6 lakhs) and (source of income is from business or private) and (no of cars possessed one or doesn't have)) then customers are likely to spend less than 10 lakhs

It means that those who are from middle income group and may or may not possess a car are likely to spend less than 10 lakhs.

Cluster 2:

If ((age is 20 and above) and (income is below 10 lakhs) and (no of cars possessed are less than or equal to 2)) then customers are likely to spend 10 to 20 lakhs

This indicates that those who are in higher middle class, and already possess car are likely to spend between 10-20 lakhs.

Cluster 3:

If ((age between 20-50) and (income is above 20 lakhs) and (no of cars possessed are at least one or more car)) then customers are likely to spend between 20-50 lakhs

From the above mentioned rule it is clear that those who belong to higher class with more that one car are likely to spend up to 50 lakhs.

Cluster 4:

If ((age is below or equal to 50) and (income range is below 6 lakhs) and (source of income is government) and (no of cars possessed are one or two)) then customer are not interested in spending on cars in future.

It means that customer belonging to lower middle group, with government job and already possessing a car is not likely to invest more on cars in future.

Complete data set.

Spending on Cars in future = 0.6594 * Age Group=Above 50, 20-35, above 50, below 20 + 0.3389 * region + 1.0677 * no. of cars possessed - 0.4967

6. Conclusions and Future Scope

This paper analyses various data mining techniques like classification, clustering and then suggests an efficient model which can be proved fruitful for decision making process to the prediction of the spending capacity on cars as per customer's demographic details.

In case of traditional model, the pattern depicts that the customer who possess more number of cars are willing to spend more in future. Marital status, source of income, region, and sex doesn't play any significant role in deciding the spending capacity. In case of proposed model, it is found that as per the cluster rules are very specific. It is found that those who are in higher income group, owning more than one car, are more likely to spend the maximum. Whereas, customers belonging to higher middle class, are likely to spend less. In case of middle class there exists two scenario , one with customers from government job owning one car are not interested in spending on cars in future where as in another scenario customer belonging to private and business class are likely to spend. It is also found that belonging to clusters, it is found that decision table is best among the rest of the classifier in case our domain.

References

[1] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

[2] CS583-unsupervised-learning.ppt .

[3] Book Han and Kamber, Data Mining: Concepts and Techniques, Second edition The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6.

[4] Inamdar S. A (School of Computational Science)Swami Ramanand Teerth, (Marathwada University, Nanded) Preprocessor Agent Approach to Knowledge Discovery Using Zero-R Algorithm, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 12, 2011.

[5] ElMoustapha Ould-Ahmed-Vall, James Woodlee, Charles Yount, Kshitij A. Doshi Intel Corporation, On the Comparison of Regression Algorithms for Computer Architecture Performance Analysis of Software Applications, 5000 W Chandler Blvd Chandler, AZ 85226 eouldahm@ece.gatech.edu and fjim.woodlee,chuck.yount,kshitij.a.doshig@intel.com..

[6] Hongjun Lu Hongyan Liu, Decision Tables: Scalable Classification Exploring RDBMS Capabilities, Department of Computer Science School of Economics and Management Hong Kong University of Science & Technology Tsinghua University Hong Kong, China Beijing, China luhj@cs.ust.hk liuhy@em.tsinghua.edu.cn .