PROJECT REPORT ON CONSUMER PROFILING.

**BY**

**SUMAN CHATTERJEE**

**ADMISSION NO.-2015JE001400**

**3rd SEMESTER, DUAL DEGREE, COMPUTER SCIENCE**

**SHUBHAM GUPTA**

**ADMISSION NO.-2015JE001403**

**3RD SEMESTER,B.TECH COMPUTER SCIENCE**

**SIDDHARTH TIU**

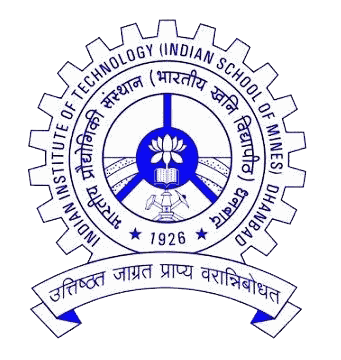
**ADMISSION NO.-2015JE001167**

**3RD SEMESTER, B.TECH , COMPUTER SCIENCE**

**ACADEMIC YEAR-2016-2017 .**

**UNDER THE GUIDANCE OF**

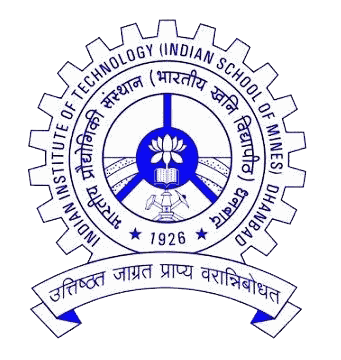
***Dr. CHANDRASHEKHAR RAO.***

******

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**INDIAN INSTITUTE OF TECHNOLOGY(INDIAN SCHOOL OF MINES), DHANBAD.**

**DHANBAD-826004.**

******

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**INDIAN INSTITUTE OF TECHNOLOGY(INDIAN SCHOOL OF MINES) , DHANBAD.**

**DHANBAD-826004.**

**CERTIFICATE**

This is to certify that the report entitled “**Consumer Profiling”** submitted by  **Suman Chatterjee,** Department of Computer Science and Engineering,IIT(ISM) DHANBAD have successfully completed the project in 3rd Semester B.Tech of Academic year 2016-2017.

(**Dr. Chiranjeev Kumar) (Dr. Chandrasekhar Rao)**

**HEAD OF DEPARTMENT, Assistant Professor,**

**Department of CSE, Project Guide,**

**IIT(ISM) DHANBAD. Department of CSE,**

**IIT(ISM) DHANBAD.**

ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude and respect towards my project guide,

Dr. ACS Rao, Assistant Professor,Department of Computer Science And Engineering,IIT(ISM) DHANBAD.

I am very much indebted to him for the generosity , expertise and guidance I have received from him while working on the project.

I wish to express my profound gratitude to Dr. Chiranjeev Kumar,Head of Department Of Computer Science And Engineering , IIT(ISM) DHANBAD for his guidance and inspiration.

DATE-26/10/2016 NAME-SUMAN CHATTERJEE.

ADMISSION NO.-15JE001400.

NAME-SHUBHAM GUPTA

ADMISSION NO.-15JE001403.

NAME-SIDDHARTH TIU

ADMISSION NO.-15JE001167.

ABSTRACT

This report examines the problems of customer relationship management (CRM) particularly customer segmentation and customer profiling, and how data mining tools are used to support the decision making. We first describe the steps towards predicting customer’s behavior, such as collecting and preparing data, segmentation and profile modeling. Then, we present a general overview of most used data mining methods including cluster discovery, decision trees, neural networks, association rules discovery.

CONTENTS

* INTRODUCTION
* CUSTOMER RELATIONSHIP MANANGEMENT
* ALGORITHMS FOR CLASSIFICATION
* PROGRAM ON CUSTOMER PROFILING USING DECISION TREES.
* CONCLUSION

**INTRODUCTION**

In today’s marketing strategies, customers have a real value to the company. Therefore, it is essential to any company to be successful in acquiring new customers and retain those that have high value. For this, many companies have gathered significant numbers of large and heterogeneous databases and these data need to be analyzed and applied in order to develop new business strategies and opportunities. The problem is that "*we are rich in data and poor in information*". What are the methods that can be used to automatically extract knowledge from data? Recently, new data analysis tools have appeared using various "machine learning" techniques. It is through the use of machine learning that data mining tools emerge. The advantage of data mining is that it can handle large amount of data and "learn" inherent structures and patterns in data; it can also generate rules and models that are useful in replicating or generalizing decisions that can be applied to the future cases. Data mining tools are therefore very useful in market segmentation, customer profiling, risk analysis, and many other applications.

The growing interests in data mining tools have also fostered the growth of the data mining tool market. Nowadays, there are so many vendors offering all range of products. For a person who is new in the field and would like to use those tools, it is essential that he/she understands how each method works in order to be able to choose the adequate ones for his/her problems. This report aims tempts to make the us familiar with the most used data mining methods (clustering, decision trees, neural networks, associations...).

1.CUSTOMER RELATIONSHIP MANNAGEMENT

Customer relationship management (CRM) includes *customer segmentation* and *customer profiling*.

*Customer segmentation* is a term used to describe the process of dividing customers into homogeneous groups on the basis of shared or common attributes (habits, tastes etc.)

*Customer profiling* is describing customers by their attributes, such as age, income, and lifestyles. This is done by building a customer’s behavior model and estimating its parameters. Customer profiling is a way of applying external data to a population of possible customers. Depending on data available, they can be used to prospect new customers or to cut off existing bad customers. The goal is to predict behavior based on the information we have on each customer. Profiling is performed after customer segmentation.

## CUSTOMER SEGEMENTATION

Segmentation is a way to have more targeted communication with the customers. The process of segmentation describes the characteristics of the customer groups (called *segments* or *clusters*) within the data. Segmenting means *putting the population into segments* according to their affinity or similar characteristics.

## 1.2 CUSTOMER PROFILING

Customer profiling provides a basis for marketers to "communicate" with existing customers in order to offer them better services and retaining them. This is done by assembling collected information on the customer such as demographic and behavioral data. Customer profiling is also used to prospect new customers using external sources, such as demographic data purchased from various sources. This data is used to break the database into clusters of customers with shared purchasing traits.

Depending on the goal, one has to select what is the profile that will be relevant to the project. A simple customer profile is a file containing at the least his name, address, city, state, and zip code. And if one needs profiles for specific products, the file would contain product information and/or volume of money spent.

The customer features that can be used for profiling are:

* + - *Geographic*. Are they grouped regionally, nationally or globally?
    - *Cultural and ethnic*. What languages do they speak? Does ethnicity affect their tastes or buying behaviors?
    - *Economic conditions, income and/or purchasing power*. What is the average household income or purchasing power of the customers? Do they have any payment difficulty? How much or how often does a customer spend on each product?
    - For acquired customer, *shopping frequency*, *frequency of complaints*, *degree of satisfaction*, *preferences* may be used to build a purchase profile.
    - *Age*. What is the predominant age group of your target buyers? How many children and of what age are in the family?
    - *Values, attitudes, beliefs*. What is the customers’ attitude toward your kind of product or service?
    - *Life cycle*. How long has the customer been regularly purchasing products?
    - *Knowledge and awareness*. How much knowledge do customers have about a product or service, or industry? How much education is needed? How much brand building advertising is needed to make a pool of customers aware of the offer?
    - *Lifestyle*. How many lifestyle characteristics about purchasers are useful?
    - *Media used*. How do targeted customers learn? What do they read? What magazines do they subscribe to?
    - *Recruitment method*. How was the customer recruited?

## 1.3 Data collection and preparation

There are many ways of collecting the data:

* + - *In-house customer database.* Names can come from direct mailers used in the past, frequent buyer programs, contest, warranty registrations, receipts and membership cards.
    - *External sources.* There are software or databases that can discover lifestyle, demographic information by using for example only zip code. E*.*g*. SuomiCD* can find Finnish demographic data from Finnish zip code. *Mosaic boxes* can also provide lifestyle data on small areas (boxes) of different countries including Finland.
    - *Research survey* either face-to-face, over the telephone, via a postal questionnaire or through Internet.

There are two types of information from the data that should be collected : classification variables and descriptor variables .

### **Classification variables**

Classification variables are used to classify survey respondents into segments. These variables are demographic, geographic, psychographic or behavioral variables.

* + - * *Demographic variables* - Age, gender, income, ethnicity, marital status, education, occupation, household size, length of residence, type of residence, etc.
      * *Geographic variables* - City, state, zip code, census tract, county, region, metropolitan or rural location, population density, climate, etc.
      * *Psychographic variables* - Attitudes, lifestyle, hobbies, risk aversion, personality traits, leadership traits, magazines read, television programs watched, etc.
      * *Behavioral variables* - Brand loyalty, usage level, benefits sought, distribution channels used, reaction to marketing factors, etc.

### **Descriptor variables**

Descriptors are used to describe each segment and distinguish one group from the others.

Descriptor variables must be easily obtainable measures or linkable to easily obtainable measures that exist in or can be appended to customer files. Many of the classification variables can be considered descriptor variables. However, only a small portion of those classification/descriptor variables are readily available from external sources.

### **Data preparation**

Before the data can be introduced to a data mining tool, they need to be cleaned and prepared in a required format . These tasks are:

* + - 1. Resolving inconsistent data formats and resolving inconsistent data encoding, geographic spellings, abbreviations, and punctuation.
      2. Stripping out unwanted fields. Data may contain many fields that are meaningless from an analysis point of view, such as version numbers and formatted production keys.
      3. Interpreting codes into text. This means to augment or replace cryptic codes with textual equivalents written in recognizable words.
      4. Combining data such as customer data from multiple sources under a common key.
      5. Finding multiple used fields. A good way to find it out is to count and perhaps list all the distinct values residing in a field.

## **K-Nearest neighbors**

### **Definition**

*K-nearest neighbor* is a predictive technique suitable for classification models. K represents a number of similar cases or the number of items in a group. With the *k-NN* technique, the training data is the model. When a new case or instance is presented to the model, the algorithm looks at all the data to find a subset of cases that are most similar to it and uses them to predict the outcome.

There are two principal parameters in the *k*-NN algorithm:

1. the number of nearest cases to be used (*k*);
2. a metric to measure the similarity.

Each use of the *k-NN* algorithm requires that a positive integer value for *k* is specified. This determines how many existing cases are looked at when predicting a new case. For example, 4-*NN* indicates that the algorithm will use the four nearest cases to predict the outcome of a new case.

### **Algorithm**

*K-NN* decides into which class to place a new case by examining some number (the *k*) of the most similar cases or neighbors. The algorithm computes the distance from the new case to each case in the training data. The new case is predicted to have the same outcome as the predominant outcome in the *k* closest cases in the training data.

Suppose that we have *n* example feature vectors *x*1, *x*2, ..., *x*n all from the same class, and we know that they fall into *c* compact clusters, *c* < *n*. Let *mi* be the mean of the vectors in *Cluster i*. If the clusters are well separated, we can use a minimum-distance classifier to separate them. That is, we can say that *x* is in *Cluster i* if the distance || *x* - *mi* || is the minimum of all the *k* distances.

*K-NN* is based on a **concept of distance**, and this requires a metric to determine distances. For continuous attributes Euclidean distance can be used, for categorical variables, one has to find a suitable way to calculate the distance between attributes in the data. Choosing a suitable metric is a very delicate task because, different metrics, used on the same training data, can result in completely different predictions. This means that a business expert is needed to help determine a good metric.

### **Advantages/Disadvantages**

With *k-NN*, the number of classes is usually given as an input variable, in some problems it is not always easy to guess.

*K-NN* is a huge model because it uses the entire training set as the model. *K-NN*’s calculation time increases as the factorial of the total number of points because k-NN requires the calculation to be made for every new case.

Distance based similarity measure cannot cope with high dimensional data, because the notion of neighborhood becomes meaningless. In this case, a more sophisticated method is used. Another problem is related to handling data inaccuracies and cluster overlapping: how to decide in which group to put those data that are on the cluster boundary region. Fuzzy approach can be used. A fuzzy *k*-nearest neighbors (fuzzy *k-NN*) technique is simply a nearest neighbors technique in which the basic measurement technique is fuzzy [Hansen00].

## **Artificial neural networks**

### **Definition**

Artificial neural networks (ANN) are among the most complicated of the classification and regression algorithms. They are often considered as a black box. A neural network require a lot of data for training, thus consuming time, but once trained, it can make predictions for new cases very quickly, even in real time. Moreover, neural networks can provide multiple outputs representing multiple simultaneous predictions. A key feature of neural nets is that they only operate directly on numbers. As a result, any nonnumeric data in either the independent or dependent (output) columns must be converted to numbers,

e.g. variables with "yes/no", "high/low" values must be replaced by "0/1".

### **Terminology**

Neural networks are defined by their **architecture**. The most common type of artificial neural network (ANN) consists of three *layers* of units: A layer of "**input**" units is connected to a layer of "**hidden**" units, which is connected to a layer of "**output**" units.

### **Advantages/Disadvantages**

Neural network is often considered as a black box as it is unable to explain the found relationships. It only works with numeric data, thus this means that non numerical data need to be converted. Moreover, the inputs need also to be normalized between 0 and 1.

Neural network is quick in predicting new cases if it is properly trained. The training phase is quite delicate, while one needs to choose appropriate number of data and control overfitting. In some package, this is done with help of other data mining tool (e.g. genetic algorithm). The drawback is that a neural network can never be exact (only accurate), even if it is trained for ever .

Decision trees

### 

### **Definition and terminology**

Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Decision trees perform many tests and then try to arrive at the best sequence for predicting the target. Each test creates branches that lead to more tests, until testing terminates in a *leaf* node (Figure 13). The path from the root to the target leaf is the *rule* that classifies the target. The rules are expressed in **if-then** form.

***Figure 13****: A decision tree grows from the root node, at each node the data is split to form new branches, until reaching a node that is not splittable any more (leaf node). Traversing the tree from the best leaf node to the root provides the rule that classifies the target variable.*

### **Tree induction**

The training process that creates the decision tree is called **induction** and requires a small number of passes through the training set. Most decision tree algorithms go through two phases: a tree growing (**splitting**) phase followed by a **pruning** phase.

* **Splitting**: The tree growing phase is an iterative process which involves splitting the data into progressively smaller subsets. The first iteration considers the root node that contains all the data. Subsequent iterations work on derivative nodes that will contain subsets of the data. At each split, the variables are analyzed and the best split is chosen. One important characteristic of splitting is that it is *greedy* which means that the algorithm does not look forward in the tree to see if another decision would produce a better overall result.
* **Stopping criteria** : Tree-building algorithms usually have several stopping rules. These rules are usually based on several factors including maximum tree depth, minimum number of elements in a node considered for splitting, or its near equivalent, the minimum number of elements that must be in a new node. In most implementations the user can alter the parameters associated with these rules. Some algorithms, in fact, begin by building trees to their maximum depth. While such a tree can precisely predict all the instances in the training set (except conflicting records), the problem with such a tree is that, more than likely, it has *overfit* the data.
* **Pruning** : After a tree is grown, one can explore the model to find out nodes or subtrees that are undesirable because of overfitting, or rules that are judged inappropriate. Pruning removes splits and the subtrees created by them. Pruning is a common technique used to make a tree more general. Algorithms that build trees to

maximum depth will automatically invoke pruning. In some products users also have

the ability to prune the tree interactively.

### **Understanding the output**

Once trained, a tree can predict a new data instance by starting at the top of the tree and following a path down the branches until encountering a leaf node. The path is determined by imposing the split rules on the values of the independent variables in the new instance. A decision tree can help a decision maker identify which factors to consider and how each factor has historically been associated with different outcomes of the decision.

Decision trees have obvious value as both predictive and descriptive models. Prediction can be done on a case-by-case basis by navigating the tree. More often, prediction is accomplished by processing multiple new cases through the tree or rule set automatically and generating an output file with the predicted value or class appended to the record for each case. Many implementations offer the option of exporting the rules to be used externally or embedded in other applications.

### **Advantages/Disadvantages**

Decision trees have unique advantages. They produce models that are easy to understand and they are unaffected by missing values in data.

Decision trees impose certain restrictions on the data that is analyzed. First, decision trees permit only a single dependent variable. In order to predict more than one dependent variable, each variable requires a separate model. Also, most decision tree algorithms require that continuous data are grouped or converted to categorical data.

## **Association rules discovery**

### **Definition and terminology**

Association and sequencing tools analyze data to **discover rules that identify patterns of behavior**, e.g. what products or services customers tend to purchase at the same time, or later on as follow-up purchases. The process of using an association or sequencing algorithm to find such kinds of rules is frequently called **market basket analysis.**

### 

### 2.4.2 **Algorithm**

To discover associations, we assume that we have a set of transactions, each transaction being a list of items (e.g. list of books). A user might be interested in finding all associations which have *s*% support with *c*% confidence such that :

* + - * all associations satisfying user constraints are found,
      * associations are found efficiently from large databases.

To find such associations, the following algorithm is given. This algorithm is also known as "*APrioriAll algorithm*" from IBM’s group: here are the steps to follow :

1. Discover all frequent items that have support higher than the minimum support required;
2. Use the set of frequent items to generate the association rules that have high enough confidence level;
3. Scan all transactions and find all items that have transaction support above *s*%. Let these be L1;
4. Build item pairs from L1. This is the candidate set C2;
5. Scan all transactions and find all frequent pairs in C2. Let this be L2;

The general rule for steps 4 and 5 is:

* 1. Build sets of *k* items from Lk-1. This is set Ck.
  2. Scan all transactions and find all frequent sets in Ck. Let this be Lk.

Step 1 is computationally expensive because it has to find all the possible associations.

### **Illustration**

Consider an example [WWW8] with the following set of transactions:

**Trans. ID Product**

1. B, M, T, Y
2. B, M
3. T, S, P
4. A, B, C, D
5. A, B
6. T, Y, E
7. A, B, M

Assume that we wish to find associations with at least 30% support and 60% confidence.

The list of frequent items is now computed. Only the following three items are qualified as frequent since they appear in more than 30% of the transactions. This is set L1.

|  |  |
| --- | --- |
| **Item** | **Frequency** |
| A | 3 |
| B | 5 |
| M | 3 |

These three items form three pairs {A, B}, {B, M}, and {A, M}. This set is C2. Now we find the frequency of these pairs, which is :

|  |  |
| --- | --- |
| **Pair** | **Frequency** |
| {A, B} | 3 |
| {B, M} | 3 |
| {A, M} | 1 |

The first two pairs have more than 30% support. Their confidence level is:

*c*(**A****B**) = 100%

*c*(**B****A**) = 60%

*c*(**B****M**) = 60%

*c*(**M****B**) = 100%

All are therefore acceptable.

The frequent item pairs (that is L2) are:

**Pair**

{A, B}

{B, M}

**Frequency**

3

3

These pairs are now used to generate a set of three items (i.e. C3). In this example only one such set is possible which is {A, B, M}. The frequency of this set is only 1 which is below 30% support and therefore this set of three items is not qualified.

The algorithm to construct the candidate set for large item sets is crucial to the performance of the algorithm. It is the generation of the large 2-item sets that is the key to improving the performance of the algorithm.

CONCLUSION

This report is about reviewing some data mining methods that can be used for customer segmentation and profiling. We have defined some requirements for segmentation and profiling. Then some methods on data mining are described. Data mining is a vast field, we had to limit the scope of this report to the most common methods like k-nearest neighbor, neural networks, association rule discovery.