International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

The Data Mining Model that Predicts the Customer Decision-Making in Buying or Renting a Home

Myint Myint Than

University of Computer Studies, Kalay, Myanmar

Abstract: ***For the last few decades, World Populations become larger and larger. All country in the world become more populated explosion. It is difficult for people to own a house. To solve this problem, they have to make a decision for Buying or Renting a house. This paper is aimed at people without a house to make the decision whether they would buy a house or rent. The main objective to this prediction is to advance the knowledge of what are factors that influences the people's choice for living. The frequent item sets are minded and associated from the market basket database using the efficient algorithm and hence the association rules are generated. The prediction can be constructed using data mining approaches: j48 Decision Tree classifier and Naïve Bayesian classifier. Weka software is used one of Data Mining techniques in this paper.***

Keywords: Decision Tree, J48, Naïve Bayes, Weka

1. Introduction

Data mining discovers the hidden knowledge from large data sets. It is divided into six steps such as data cleaning, data transformation, data selection, data integration, pattern evaluation and knowledge presentation.

Now a days, it is becoming a necessity of a person to own a home than rend. One of the challenges for the people that have heavily invested in decision making is how to extract the important information from the market basket databases. Associations rules are derived as threshold levels from the frequent set using support and confidence. Frequent item set is the sets of items, which have minimum support, are known as. The proportion of transactions in the data set, which contain the item set, is called the support count of an item set. Data mining contains of many classification algorithms such as classification by decision tree induction, Bayesian classification, Rule-Based classification, and contains many prediction algorithms such as Linear Regression, non linear Regression. These algorithms are used to predict the customer Decision-Making in buying or renting a Home

2. Related Work

The research work used classification algorithm such as classification by decision tree induction and Bayesian classification. This paper focused on J48 decision tree algorithms for data analysis with an experimental approach and proved that J48 has more accurate result compared with Naïve Bayesian Classifier [2].

There was another research study described that decision tree was widely used learning method and did not require any prior knowledge of data distribution in [3]. It worked well on noisy data and it had been applied to classify based on the data set. It discovered classification rules for home data set using the decision tree algorithm [4]. The proposed study [5] comprised data pre-processing to remove noisy data in home data that offers better accuracy. Naïve Bayesian Classifier are applied for filtering. The study [6] and [7] presented Analysis of Classification Algorithms and assessment of Decision Tree Algorithms. The study [8] presented detail of supervised learning algorithms used with the aim of selecting the algorithms that give the best.

2.1. Machine Learning Tool

Here we used WEKA 3.9 tool for analyzing our dataset. WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. In a real-world datamining problems WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their applications. It is a collection of different machine learning algorithms for data mining tasks and predictions. The algorithms are applied directly to a dataset. WEKA implements algorithms for data association, preprocessing, classification, clustering, regressing rules; it has another advantages as visualization tools. This package of WEKA used for developing new machine learning schemes. Under the GNU general Public License, WEKA is a open source software.

3. Data Preprocessing

The data collected from the real world is incomplete, inconsistent, inadequate and it consisting of noise, redundant groups. The Knowledge discovery using the training data with such an irrelevant, inconsistent and redundant data will reduce the mining quality. Data preprocessing is very important and a prerequisite step in the data mining process. Low quality data will lead to low quality mining results. Thus, the data can be preprocessed in order to improve the quality of the data. Data quality can be accessed in terms of accuracy, completeness and consistency. Preprocessing reconstructs the data into a format that will be very easy and effective for further processing.

To improve the quality of mining, the data preprocessing techniques are applied. The main objective of data preprocessing is cleaning of noise, filling up of missing values, reduce the redundancy and normalize the data. The preprocessing steps are data cleaning, data integration, data

transformation, data reduction and data discretization. After preprocessing has been done the data will be complete, noise free and ready for classification. Any classification algorithm can be applied for classifying the data.

Various tools and techniques are used for preprocessing which includes:

- *Data cleaning*: It can be applied to remove noise and correct inconsistencies in data.
- *Data integration*: It merges data from multiple sources into a coherent data store such as a data warehouse.
- *Data reduction*: It can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering.
- *Data transformation*: It may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0.
- *Data discretization*: It can also be useful, where row data values for attributes are replaced by ranges or higher conceptual levels. For example, raw values for age may be replaced by higher-level concepts, such as youth, adult, or senior.

3.1. Customer Dataset

The customer data sets is provided by the Machine learning Competitions platform for the analysis of this work. The data set has seven input attributes namely customer-age, job, salary, martial-status, wife-salary, Home-status and Choose-Type. Training data are analyzed by a classification algorithm. In this article the class label attribute is Choose-Type. It has two attribute values, rent and buy. This paper predicts the Choose-Type by using the naïve Bayesian classification algorithm and Decision Tree classification algorithm. The Weka software Version 3.9 is used to predict the customer decision making. Figure 1 shows all attributes visualization.

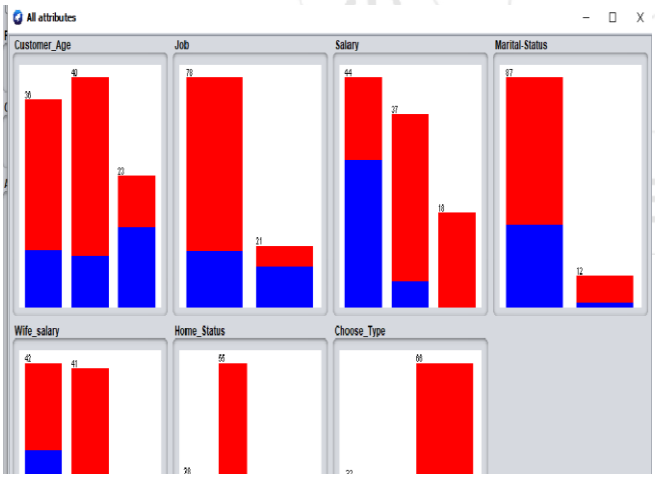

Figure 1: Attributes Visualization

Loading Data set:
Select Explorer > Open file. Then choose the Home dataset.

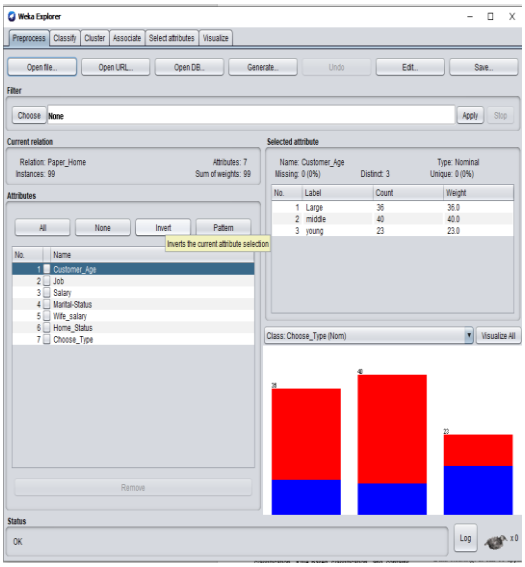

Figure 2: Loading Dataset

3.2. Classification Analysis

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels.

Many classification methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large amounts of disk-resident data. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis.

Classification analysis is used to map data sets into predefined groups and classes. As the classes are determined before examining the datasets, it is considered to be the supervised learning. The learning of classifier is " supervised" in that it is told to which class each training tuple belongs. It constracts with unsupervised learning (or) clustering, in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.

This first step of the classification process can also be viewed as the learning of a mapping or function, y = f(X), that can predict the associated class label y of a given tuple X. In this view, we wish to learn a mapping or function that separates the dataclasses. Typically, this mapping is represented in the form of classification rules, decision trees or mathematical formulae. The rules can be used to categorized future data tuples, as well as provide deeper insight into the data contents. They also provide a compressed data representation.

"What about classification accuracy?" In the second step, the model is used for classification. First, the predictive accuracy of the classifier is estimated. If we were to use the training set to measure the classifier's accuracy, this estimate

There are various classification algorithm used for the implementation of the classification analysis such as Logistic Regression, Random Forest classifiers, Naïve Bayesian classification, Decision Tree classifier etc. For this paper, we choose two of widely used classifiers, namely Naïve Bayesian and Decision Tree classification algorithms. The classifiers classifies the customer into two groups as- customer can buy a home and can rent a home on the basis of the characteristics mentioned in the datasets.

3.2.1. Bayes classification Method

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probabilities that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem.

Naïve bayes are group of probabilistic classifiers built on the Bayes's theorem. The naïve Bayesian classifier is simple Bayesian classifier.

Which states that: Consider X and H
X : is an evident, X = x1, x2,……. Xn
H: is the hypothesis. P(H\X) is the posteriori probability, of H conditioned on X.
Then P (H/X) = (P (X/H) × P (H)) / P (X)

3.2.2. Naïve Bayesian Classification

The naïve Bayesian Classifier, or simple classifier, works as follows:

1) Let D be a training set of tuples and their associated class labels. As usual, each tuple is represent by an n-dimensional attribute vector, X = (x_1, x_2, ---, x_n), depicting n measurements made on the tuple from n attributes, respectively, A_1, A_2, ---, A_n.

2) Suppose that there are m classes , C_1,C_2, --- C_m. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is , the naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and only if
P (Ci\X) > P (Cj\X) for 1<=j <= m, j != i.
Thus, we maximize P(Ci\X). The class Ci for which P(Ci\X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem
P (Ci\X) = (P (X\Ci) × P (Ci)) / P (X)

3) As P (X) is constant for all classes, only P(X\Ci) × P (Ci) needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1)=P(C_2)$= ---- = P(Cm), and we would therefore maximize P(X\Ci). Otherwise, we maximize (P (X\Ci) × P (Ci)). Note that the class prior probabilities may be estimated by P (Ci)= $|C_{i,D}|$ / |D|, where $|C_{i,D}|$ is the number of training tuples of class Ci in D.

4) Given data sets with many attributes , it would be extremely computationally expensive to compute P(X\Ci). To reduce computation in evaluating P(X\Ci), the naïve assumption of class-conditional independence is made. This presums that the attributes' values are conditionally independent of one another, given the class label of the tuple (i.e, that there are no dependence relationships among the attributes). Thus,

$$P(X\backslash Ci) = \prod_{k=1}^{n} P(X_k\backslash Ci)$$

$$= P(X_1\backslash C_i) \times P(X_1\backslash C_i) \times \text{----} \times P(X_n\backslash C_i)$$

We can easily estimate the probabilities P ($X_1\backslash C_i$) × P ($X_2\backslash C_i$) × ---- × P ($X_n\backslash C_i$) from the training tuples. Recall that here xk refers to the value of attribute Ak for tuple X. For each attribute, we look at whether the attribute is categorical or continuous valued. For instance, to compute P (X\Ci) , we considered the following:

(a) If A_k is categorical, then P (Xk\ Ci) is the number of tuples of class Ci in D having the value xk for Ak, divided by |C i,D|, the number of tuples of class Ci in D.

(b) If Ak is continuous-valued, then we need to do a bit more work, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gussian distribution with a mean μ and standard deviation σ, defined by

$$g(x, \mu, \sigma) = 1 / (2\pi\sigma)^{\frac{1}{2}} \times e^{-(x-\mu)^2/2\sigma^2}$$

So that P (Xk \Ci) = g (x_k, μ_{Ci}, σ_{Ci})
These equations may appear daunting, but hold on! We need to compute μ_{Ci} and $\sigma_{Ci,}$ which are the mean (i.e., average) and standard deviation, respectively, of the values of attribute Ak for training tuples of class Ci.

5) To predict the class label of X, P (X\Ci) × P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if
P(X \Ci) × P(Ci) > P(X\Cj) × P (Cj) for 1<= j<= m, j!=1
In other words, the predicted class label is the class Ci for which P (X\Ci) × P(Ci) is the maximum.

3.2.3 Experiment using Naïve Bayes classifier

Select Classify > Choose >Classifiers > Bayes > NaiveBayes. Then we set the Percentage Split of data, 75% of training and 25% for testing. Figure 3 shows the obtained results from applying Naïve classifier on the selected dataset which which shows 78.374 % accuracy, 29 correctly classified instances, and with Recall value of 0.784 and Precision of 0.792.

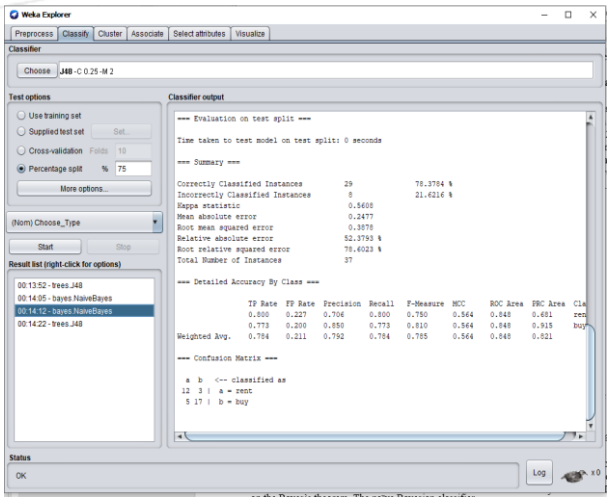

Figure 3: Naïve Bayes Classification Results

3.3 Decision Tree Induction

The decision tree is one of the classification algorithms. It is frequently used by the researchers to classify the data. The decision tree is very popular because it is easy to build and require less domain knowledge. Also the decision tree method is scalable for large database.

The first decision tree algorithm is developed in early 1980s is Iterative Dichotomiser (ID3). Quinlan and Kaufmann (1993) presented the C4.5 which is the successor of ID3. In 1984 Classification And Regression Tree (CART) is introduced. It is mainly support for the binary tree classification. All the three algorithms adopt the greedy approach and construct the decision tree in top down, recursive, divide and conquer manner.

The following are the basic steps used for decision tree algorithm:
Input: Data partition, D, which is a set of training tuples and their associated class labels;
Attribute list: The set of candidate attributes;

Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.

Output: A decision tree:
(1) Create a node N
(2) If tuples in 'D' are belongs to same class C, then return N as a leaf node labeled with the class C
(3) If attribute list is empty then return N as a leaf node labeled with the majority class in D;
(4) Apply Attribute selection method (D, attribute list) to find the "best" splitting criterion
(5) label node N with splitting criterion
(6) If splitting attribute is discrete-valued and multiday splits allowed then//not restricted to binary trees
(7) Attribute list ← attribute list-splitting attribute;
//remove splitting attribute
(8) For each outcome j of splitting criterion
//partition the tuples and grow subtrees for each partition
(9) Let Dj be the set of data tuples in D satisfying outcome j;
//a partition
(10) If Dj is empty then attach a leaf labeled with the majority class in D to node N(11) Else attach the node returned by Generate decision tree (Dj, attribute list) to node N;
End for
(12) Return N

The decision tree algorithm is very robust and learning efficiency with its learning time complexity of O(nlog2n). The outcome of a decision tree that can
be easily represented as a set of symbolic rules (IF….THEN). This rule can be directly interpreted and compared with available knowledge and provide useful information.

A decision tree comprises of node and leaves, where nodes represent a test on the values of an attribute and leaves represent the class of an instance that satisfies the conditions. The outcome is 'true' or 'false'.

Rules can be derived from the path starting from the root node to the leaf and utilizing the nodes along the way as preconditions for the rule, to predict the class at the leaf. The tree Pruning has to be carried out to remove unnecessary preconditions and duplications.

The nodes represent genes and branches represent the expression conditions. The leaves of the tree represent the decision outcome. The brace under a leaf denotes the number of instances correctly and incorrectly classified by the leaf and the equivalent decision rules are derived from the decision trees.

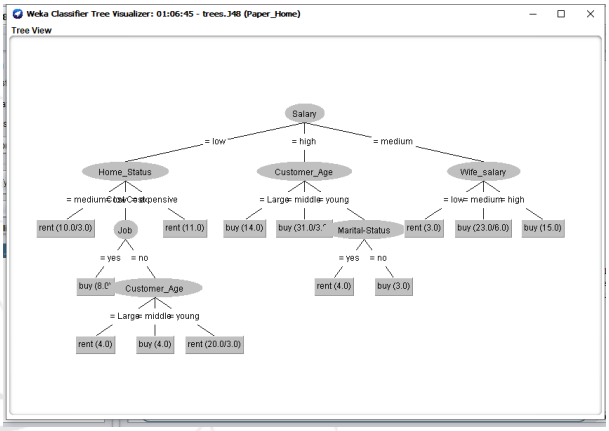

Figure 4: Visualize Decision Tree Classification Results

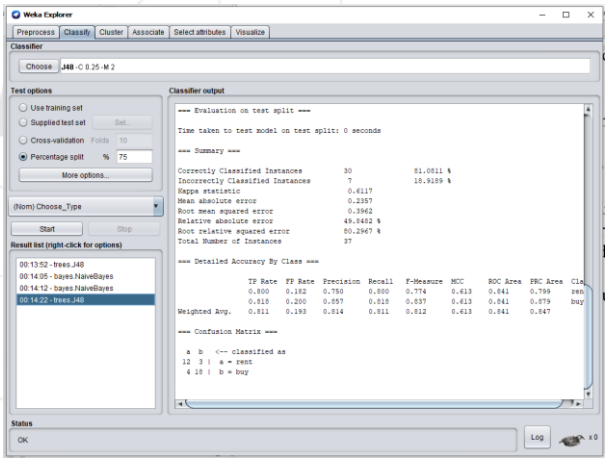

Figure 5: Decision Tree Classification Results

3.3.1 Experiment using Decision Tree classifier

Select Classify > Choose >Classifiers > Trees > J48. Then we set the Percentage Split of data, 75% of training and 25% for testing. Figure 5 shows the obtained results from applying Decision Tree classifier on the selected dataset which which shows 81.0811 % accuracy, 30 correctly classified instances, and with Recall value of 0.811 and Precision of 0.814.

4. Conclusions

In this paper, a comparison for Naïve Bayesian and Decision Tree calssifiers are shown on the frequent item set is

developed for the analysis of Customer-Home data. The Customer-Home data set is taken and analyzed to know which analysis will be better for prediction. On the basic of resulting factors, the performance of Decision Tree classifier is more accuracy and more precision than Naïve Bayes classifier.

References

[1] Machine Learning Pepositary : https:// archive.ics.uci.edu/ml/dataset.php.
[2] WEKA Machine Learning Project, at https:// www.waikato.ac.nz/ml/weka/downloading.html
[3] K.VEMBANDASAMY, IJISET -International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015 Heart Diseases Detection Using Naïve Bayes Algorithm.
[4] Kaveri Kar, Rashmi Patel, The Dat Mining Model that predicts the Customer Decision-Making in Buying a Car; International Journal of Computer Science and Mobile Computing IJCSMC, Vol 8, Issue.4, April 2019, pg.177—181.
[5] Novakovic, J., "The Impact of Feature Selection on the Accuracy of Naïve Bayes Classifier", 18th Telecommunications forum TELFOR 2010, November 2010, pg: 1113-1116.
[6] Vanaja, S. and K. Rameshkumar, "Performance Analysis of Classification Algorithms on Medical diagnoses-a Survey", Journal of Computer Science,2015
[7] Arundthathi A, Ms. K. Glory Vijayaselvi, Dr. V. Savithri, "Assessment of Decision Tree Algorithms on Student's Recital", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 03 | Mar -2017 www.irjet.net , p-ISSN: 2395-0072© 2017, IRJET | Impact Factor value: 5.181 | ISO 9001:2008 Certified Journal | Page 2342
[8] jiawei Han, Micheline Kamber, Jian Pei " DATA MINING Concepts and Techniques", Third Edition, pg 330-355.

Author Profile

Myint Myint Than received the M.I.Sc. (degrees in Master of Information Science) from Mandalay Computer University (Myanmar) in 2001. Now she is a lecturer at University of Computer Studies, Kalay city, Sagaing Division, Myanmar.