Integrated Intelligent Research (IIR) International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 01 June 2014, Page No. 30- 32
ISSN:2278-2419

Analysis of Classification Algorithm in Data Mining

R. Aruna devi[1],K. Nirmala[2]

[1]Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India
[2]Associate Professor, Department of Computer Science, Quaid-E- Millath Government College for Women (A), Chennai, Tamil Nadu, India
Email-arunaa_2008@yahoo.com, nimimca@yahoo.com

Abstract-Data Mining is the extraction of hidden predictive information from large database. Classification is the process of finding a model that describes and distinguishes data classes or concept. This paper performs the study of prediction of class label using C4.5 and Naïve Bayesian algorithm.C4.5 generates classifiers expressed as decision trees from a fixed set of examples. The resulting tree is used to classify future samples .The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. C4.5 uses information gain to help it decide which attribute goes into a decision node. A Naïve Bayesian classifier is a simple probabilistic classifier based on applying Baye's theorem with strong (naive) independence assumptions. Naive Bayesian classifier assumes that the effect of an attribute value on a given class is independent of the values of the other attribute. This assumption is called class conditional independence. The results indicate that Predicting of class label using Naïve Bayesian classifier is very effective and simple compared to C4.5 classifier.

Keywords: Data Mining, Classification, Naïve Bayesian Classifier, Entropy

I. INTRODUCTION

Data mining is the extraction of implicit, previously unknown, and potentially useful information from large databases. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form, which is easily comprehensible to humans. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining task can be classified into two categories: Descriptive and Predictive. Descriptive mining tasks characterize the general properties of the data in the database[1]. Predictive mining tasks perform inference on the current data in order to make prediction. Classification is the process of finding a model that describes and distinguishes data classes / concepts. The goal of data mining is to extract knowledge from a data set in a human-understandable structure and involves database and data management, data preprocessing, model and inference considerations, complexity considerations, post-processing of found structure, visualization and online updating. The actual data-mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. Data mining involves six common classes of tasks.(1)Anomaly detection – The identification of unusual data records, that might be interesting or data errors and require further investigation.(2)Association rule learning-Searches for relationships between variables.(3) Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.(4)Classification – is the task of generalizing known structure to apply to new data.(5)Regression – Attempts to find a function which models the data with the least error.(6)Summarization – providing a more compact representation of the data set, including visualization and report generation.

II. ALGORITHM

C4.5 algorithm is introduced by Quinlan for inducing Classification Models, also called Decision Trees[13]. We are given a set of records. Each record has the same structure, consisting of a number of attribute/value pairs. One of these attributes represents the category of the record. The problem is to determine a decision tree that on the basis of answers to questions about the non-category attributes predicts correctly the value of the category attribute. Usually the category attribute takes only the values {true, false}, or {success, failure}, or something equivalent.

The C4.5 algorithm can be summarized as follows:

Step 1: Given a set of S cases, C4.5 first grows an initial tree using the concept of information entropy. The training data is a set $S=S_1$, S2, of already classified samples. Each sample $S_{i=}X_{1,}$ $X_{2...}$ is a vector where $X_{1,}$ X2, represent attributes or features of the sample. The training data is augmented with a vector $C = c_1, c2,$ where $c_1, c2,$ represent the class to which each sample belongs.

Step 2: At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision.

Step 3: Create a decision tree based on the best node

Step 4: Apply the same procedure recursively

III. NAÏVE BAYESIAN

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 01 June 2014, Page No. 30- 32
ISSN:2278-2419

A Naive Bayesian classifier is a simple probabilistic classifier based on applying Baye's theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model"[2]. In simple terms, a Naive Baye's classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable.

Depending on the precise nature of the probability model, naive Baye's classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Baye's models uses the method of maximum likelihood; in other words, one can work with the naive Baye's model without believing in Bayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified assumptions, naive Baye's classifiers have worked quite well in many complex real-world situations.

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X/C_i)$. In order to reduce computation in evaluating $P(X/C_i)$, the naïve assumption of class conditional independence is made. This presumes that the values of the attributes 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,$P(X_1/C_i)\times\ P(X_2/C_i)\times \ldots \times P(X_n/C_i)$.

easily estimate the probabilities $P(X_1/C_i)$,$P(X_2/C_i \ldots P(X_n/C_i)$ from the training tuples .Recall that here X_k refers to the value of attribute A_k for tuple X.

IV. DATASET DESCRIPTION

The main objective of this paper is to use classification algorithm to predict the class label using C4.5 classifier and Bayesian classifier on the large dataset. The model used in this paper predicts the status of the tuple having the values department has "system" who are 26..30 years, have income 46..50K[2].

Table 1: Department

Department	Status	age	salary
Sales	Senior	31..35	46k..50k
Sales	Junior	26..30	26k..30k
Sales	Junior	31..35	31k..35k
Systems	Junior	21..25	46k..50k
Systems	Senior	31..35	66k..70k
Systems	Junior	26..30	46k..50k
Systems	Senior	41..45	66k..70k
Marketing	Senior	36..40	46k..50k
Marketing	Junior	31..35	41k..45k
Secretary	Senior	46..50	36k..40k
Secretary	Junior	26..30	26k..30k

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this paper we wish to study of C4.5 and Naïve Bayesian classification, given the training data as in Table 1. The data tuple are described by the attributes department, status, age and salary. The class label attribute status, has two distinct values namely { Senior,Junior }.Let C1 corresponds to the class Status="Junior" and C2 correspond to Status="Senior". The tuple we wish to classify is,

X=(dept="system",age="26..30",salary="46..50k")

A. Algorithm

C4.5 uses a statistical property, called information gain. Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information is selected as test attribute. In order to define gain, we first borrow an idea from information theory called entropy. Entropy measures the amount of information in an attribute.

$$E(S) = -\sum_{j=1}^{n} f_S(j) \log_2 f_S(j)$$

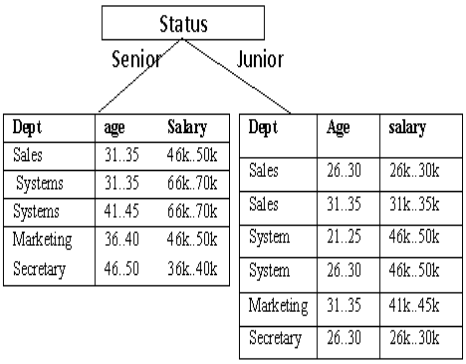
Where:

- $E(S)$ is the information entropy of the set S ;
- n is the number of different values of the attribute in S (entropy is computed for one chosen attribute)
- $f_S(j)$ is the frequency (proportion) of the value j in the set S
- $\log_2$ is the binary logarithm
- Entropy of 0 identifies a perfectly classified set.

$$G(S,A) = E(S) - \sum_{i=1}^{m} f_S(A_i) E(S_{A_i})$$

Where:

- $G(S,A)$ is the gain of the set S after a split over the A attribute
- $E(S)$ is the information entropy of the set S
- m is the number of different values of the attribute A in S
- $f_S(A_i)$ is the frequency (proportion) of the items possessing A_i as value for A in S
- A_i is i^{th} possible value of A
- S_{A_i} is a subset of S containing all items where the value of A is A_i



Dept	age	Salary
Sales	31..35	46k..50k
Systems	31..35	66k..70k
Systems	41..45	66k..70k
Marketing	36..40	46k..50k
Secretary	46..50	36k..40k

Dept	Age	salary
Sales	26..30	26k..30k
Sales	31..35	31k..35k
System	21..25	46k..50k
System	26..30	46k..50k
Marketing	31..35	41k..45k
Secretary	26..30	26k..30k

Fig-1

B. Naïve Bayesian Classifier

Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X/C_i)$.We need to maximize $P(X/C_i)*P(C_i)$ for i=1,2.$P(C_i)$, the prior probability of each class, can be computed based on the training tuples.

P(status="senior")=5/11=0.455

P(status="junior")=6/11=0.545

To compute $P(X/C_i)$ for i=1to 2, We compute the following conditional probabilities

P(dept="system"/status="senior")=2/5=0.4

P(dept="system"/status="junior")=2/6=0.33

P(age="26..30"/status="senior")=0/5=0

P(age="26..30"/status="junior")=3/6=0.5

P(salary="46k..50k"/status="senior")=2/5=0.4

P(salary="46k..50k"/status="junior")=2/6=0.33

Using the above probabilities, we obtain

P(X/status="junior") =0.33 X 0.5 X 0.33=0.054

Similarly,

P(X/status="Senior") =0.4 X 0 X 0.4 = 0

To find the class $C_{i,}$ that maximizes ,

$P(X/ C_i) X P(C_i)$, We compute

P(X/status="senior") X P(status="Senior") =

0 X 0.455 = 0

P(X/status="Junior") X P(status="Junior") =

0.054 X 0.545 = 0.0294

To find class C_i that maximizes the naïve Bayesian classification predicts status="junior" for tuple X

C. Comparison And Reults

For the comparison of our study, first we used a C4.5 classification algorithm derives its classes from a fixed set of training instances. The classes created by C4.5 are inductive, that is, given a small set of training instances, the specific classes created by C4.5 are expected to work for all future instances. Secondly we used a Naïve Bayesian classification algorithm. The Naïve Bayesian classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. It handles missing values by ignoring the instance. It handles quantitative and discrete data. Naïve Bayesian algorithm is very fast and space efficient.

VI. CONCLUSION AND FUTURE DEVELOPMENT

In this paper, the comparative study of two classification algorithms is compared. The Naïve Bayesian model is tremendously appealing because of its simplicity, elegance, and robustness. The results indicate that Naïve Bayesian classifier is very effective and simple compared to C4.5. A large number of modifications have been introduced, by the statistical, data mining, machine learning, and pattern recognition communities, in an attempt to make it more flexible.

REFERENCES

[1] A.K.Pujari, "*Data Mining Techniques*", University Press, India 2001.
[2] Jiawei Han and Micheline Kamber "Data *Mining Concepts and Techniques*"
[3] S.N.Sivanandam and S.Sumathi, "*Data Mining Concepts Tasks and Techniques*", Thomson , Business Information India Pvt.Ltd.India 2006
[4] H. Wang, W. Fan, P. Yu, and J. Han."*Mining concept-drifting data streams using ensemble Classifiers"*.
[5] V. Ganti, J. Gehrke, R. Ramakrishnan, andW. Loh. "*Mining data* streams *under block evolution"*.
[6] Friedman N, Geiger D, Goldsmith M (1997) "*Bayesian network classifiers"*.
[7] Jensen F., "*An Introduction to Bayesian Networks"*.
[8] Murthy, "Automatic Construction of Decision Trees from Data"
[9] Website:www.cs.umd.edu/~samir/498/10Algorithms-08.pdf
[10] Website:www.hkws.org/seminar/sem433-2006-2007-no69.pdf
[11] Website:en.wikipedia.org/wiki/Data_mining
[12] http://en.wikipedia.org/wiki/ID3_algorithm
[13] Quinlan JR(1993) C4.5: Programs for machine learning.Morgan Kaufmann Publichers, San Mateo.