A Comparative Study of Decision Tree and Naive Bayesian Classifiers on Medical Datasets

D.Sheela Jeyarani
Research Scholar
Department Of CS
Mother Teresa Women's University
Kodaikanal,Tamil Nadu,

G.Anushya
Research Scholalar
Department Of CS
Manonmaniam Sundaranar University
Tirunelveli,Tamil Nadu,

R.Raja rajeswari
Assistant Professor
Department Of CS
Sri Meenakshi Government Arts College(W)
Madurai,Tamil Nadu,.

A.Pethalakshmi, Ph.D
Head &Associate Professor
Department Of CS
M.V.M Government Arts College(W)
Dindigul,Tamil Nadu,

ABSTRACT

Data Mining is a process to discover valuable patterns from large datasets. Classification is an important data mining functionality and it employs supervised learning to predict class labels for a given sample. This research paper apprises about two important classification algorithms, Decision trees and Naive Bayesian and compares their predictive accuracy on selected medical datasets.

1. INTRODUCTION

Data mining and Data warehousing are two significant domains in the field of knowledge engineering. Data Mining deals with finding patterns, computing categories and predicting values, from vast amounts of data. Data Mining functionalities include Association Rule Mining, Classification, Prediction and Outlier Analysis. Among these, Classification, an important data mining technique builds a learning model using training samples and based on the learnt classification rules categorises a given sample. This research paper discusses two well known classification algorithms Decision trees and Naive Bayesian Algorithms, using MATLAB implements them on selected medical data sets and analyses the experimental results.

2. CLASSIFICATION BY DECISION TRESS

2.1 Decision tree Induction

Decision trees [6,8] are powerful and popular tools for classification and prediction. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution. The top most node in a tree is the root rode. An example follows.

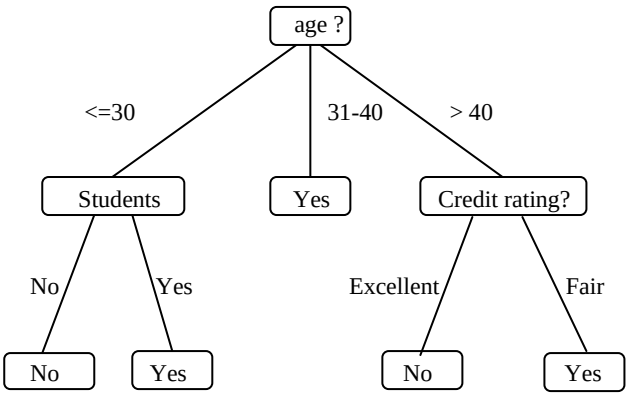

Fig -1 A decision tree

The most popular Decision tree algorithms are CART, CHAID and C4.5 [1]. This section outlines C4.5 algorithm, by first introducing the basic methods of its predecessor, ID3 algorithm.

The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top down recursive divide and conquer manner. The basic algorithm of ID3 [8]is discussed below.

Algorithm: Generate decision tree

Narrative : Generate a decision tree from the given training data

Input : The training samples, samples, represented by discrete – valued attribute; the set of candidate attributes, attribute list.

Output : A decision tree.

Method :

1) create a node N
2) if samples are all of the same class, C then
3) return N as a leaf node labeled with the class C;
4) if attribute – list is empty then
5) return N as a leaf node labeled with the most common class in samples; // majority voting

6) select test – attribute, the attribute among attribute – list with the highest information gain;

7) label node N with test attribute;

8) for each known value a_i of test attribute;

9) grow a branch from node N for the condition test attribute = a_i ;

10) let s_i be the set of samples in samples for which test attribute = a_i // a partition

11) if s_i is empty then

12) attach a leaf labeled with the most common class in samples;

13) else attach the node returned by Generate – decision – tree (s_i, attribute – list – test – attribute);

2.2 Extraction of classification rules

The knowledge represented in decision trees can be extracted and represented in the form of IF- THEN rules. One rule is created for each path from the root to a leaf node. Each attribute – value pair along a given path forms a conjunction in the rule antecedent. The leaf node holds the class prediction, forming the rule consequent. The IF – THEN rules may be easier for us to understood, for larger trees.

C.4.5, a successor algorithm to ID3 proposes mechanism for more types of attribute tests [5]. The information gain measure is biased in that it tends to prefer attributes with many values. C 4.5 proposes gain ratio, which considers the probability of each attribute value.

3. NAIVE BAYESIAN CLASSIFICATION

3.1 Bayesian Classification

The Bayesian classification [4] represents a supervised learning method as well as a statistical method for classification Assuming an underlying probabilistic model, it allows to capture an certainty about the model in a principled way by determining probabilities of the outcomes.

Bayesian classification [2] is based on Bayes Theorem which is stated below.

Let X be a data sample whose class label is unknown. Let H be some hypothesis, such as that the data sample X belongs to a specified class.

$$P(H/X) = \frac{P(X/H).\,P(H)}{P(X)}$$

3.2 Naive Bayesian Classifier

The naive Bayesian classifier [2] works as follows.

1. Each data sample is represented by an n dimensional feature vector, X = (*x*1, *x*2.... *x*n)

2. Suppose that there are m classes C_1, C_2.... C_m .Given an unknown data sample, X, the classifier will predict that X belongs to the class having the higher posterior probability, conditioned on X. That is, the naive Bayesian classifier assigns an unknown sample X to the class C_i if and only if

$$P(C_i / X) > P(C_j / X)$$

for $1 \leq j \leq m$ and the above posterior probabilities are computed using Bayes theorem. In other words an unknown sample X is assigned to the class C_i for which the $P(C_i/X)$ is the maximum.

4. EXPERIMENTAL ANALYSIS

The above stated algorithms were implemented and evaluated using MATLAB on selected datasets of UCI machine learning repository. Accuracy of classification is used as the metric for deciding the best suited model. Accuracy [3,7,9] is determined as the ratio of records correctly classified during testing to the total number of records tested. The details of the tested data sets are as follows:

Data Sets	**No. of Instances**	**No. of Attributes**
Haberman	306	3
Ecoli	331	7
Wine	178	13
Yeast	1484	8
Wisconsin breast cancer	699	9
Dermatology	362	33
Mammographic masses	830	5
PIMA Indian diabetics	768	8
Lung cancer	32	56
Heart	297	13

Table 1.Data set Description

The following table shows, for each dataset, the computed classification accuracy [7] of the two algorithms, Decision trees and Naive Bayes classifier.

Data sets	**Instances**	Decision trees	Naive Bayesian
Haberman	**306**	76.2500	80.1385
Ecoli	**331**	73.8279	82.0970
Wine	**178**	79.7830	80.3445
Yeast	**1484**	81.8850	75.7342
Wisconsin breast cancer	**699**	79.2110	79.2561
Dermatology	**362**	79.6150	82.3229
Mammographic masses	**830**	78.9620	82.5176
PIMA Indian diabetics	**768**	85.9090	82.0698
Lung cancer	**32**	90.5880	82.3077
Heart	**297**	80.7130	**80.7572**

Table. 2 Performance Analysis

5. FINAL REMARK

From the experiment results it can be inferred that, classification accuracy of Decision trees and Naive Bayesian algorithms tends to be the same, for most data sets. Though Naive Bayesian Algorithm, has a proven theoretical frame work, Decision tree algorithm suits datasets with more number of instances compared to Naive Bayesian.

This study suggests to build hybrid algorithms by combining Decision tree / Naive Bayesian with optimization techniques like Genetic algorithms and Ant colony optimization.

6. REFERENCES

[1] Berry. M.J et.al, "Data Mining Techniques for Marketing, Sales and Customer Support", John Wiley of Sons Inc, USA, 1997.

[2] Han, Jiawei, Kamber, Micheline, "Data Mining Concepts & Techniques", Morgen Kaufmann publications, USA, 2001.

[3] Nitu Mathuriya, et.al, Comparison of K.means and Back propagation Data Mining Algorithms, International Journal of Computer Technology and Electronics Engineering, Volume 2 Issue 2, 2012.

[4] Patrick Ozer, Data Mining Algorithms for classification, B.Sc Thesis, Redbound University Nijimegan, 2008.

[5] 5Quinlan, J.Ross, C4.5; Programs for Machine Learning, Morgan Kaufmann Publication, USA, 1993.

[6] Raj kumar et.al,Classifiction algorithms for Data Mining :A Survey,International Journl of Innovations in Engineering and Technology,Vol.1 Issue 2 ,2012.

[7] Sampson Adu Poku,Comparing Classification algorithms in Data mining,M.Sc Thesis,Centrl Connecticut State University,2012.

[8] Veronical, S.Moetini, Towards the use of C4.5 Algorithm for classifying Banking Data set, Integral, Vol.8. No.2, 2003.

[9] Xin dong wu et.al, "Top 10 Algorithms of Data Mining", Springs – Verlag London,2007.