Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining

Bhavesh Patankar
Sr. Lecturer, M.Sc.(IT) Department,
Kadi sarva Vishwavidyalaya, Gandhinagar, India

Dr. Vijay Chavda
Principal, NPCCSM, Kadi
Kadi sarva Vishwavidyalaya, Gandhinagar, India

Abstract*— Data mining concept is emerging fast in recognition; it is a technology which is amalgamation of diversified fields like machine learning, artificial intelligence, database system and statistics. The main goal of data mining process is to dig out information from a large data into form which could be comprehensible for further application. There are various methods for classification exists like decision trees, bayesian, neural networks etc. Before applying any mining technique, relevant attributes needs to be extracted. This paper is an introductory paper on different techniques used for classification. The paper will demonstrate the strength and precision of each algorithm for classification in term of performance competence and time complexity required. Some algorithms of data mining are used to give solutions to classification problems on a given dataset.

***Keywords*— Data Mining, Classification Techniques, Decision tree, Bayesian classifier, k- NN classifier**

I. INTRODUCTION

Data mining involves the use of refined data analysis tools to ascertain formerly unknown, valid patterns and relationships in large data set. Smart methods are applied in order to extract data pattern, by series of steps. Data mining is an integral part of knowledge discovery in database (KDD), which is the overall process of converting the raw data into useful information [1]. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In common, data mining tasks can be divided into two categories: descriptive and predictive Classification techniques in data mining are competent of processing a huge amount of data. It can foresee categorical class labels and classifies data based on training set and class labels and therefore can be used for classifying newly available data. Thus it can be outlined as a certain part of data mining and is achieving more recognition. [4] In this paper Classification Methods are considered, it focuses on a study on different classification techniques that are most commonly used in data-mining. The study is carried out on three algorithms (K-NN classifier, Bayesian network and Decision tree) to show the strength and accuracy of each algorithm for classification in term of performance efficiency and time complexity. Coming section deals with a study on Algorithm, section III describe algorithm analysis and time and space complexity, in section IV k-nearest neighbor algorithm is explained. Section V describes Decision Tree and section VI comprises with Bayesian network, to end with last section concludes the paper.

II. DECISION TREE

Decision tree is a simple still widely used classification technique. A decision tree is a flow chart like tree structure, where every internal node (nonleaf node) denotes a test on an attribute, each and every branch represents result of the test, and each leaf node (or terminal node) is assigned a class label [2]. The topmost node is the root node. Decision tree is constructed in a divide and conquer approach [2]. Each path in decision tree forms a decision rule. Generally it utilizes greedy approach from top to bottom. Decision tree can be explained with an example as depicted below.

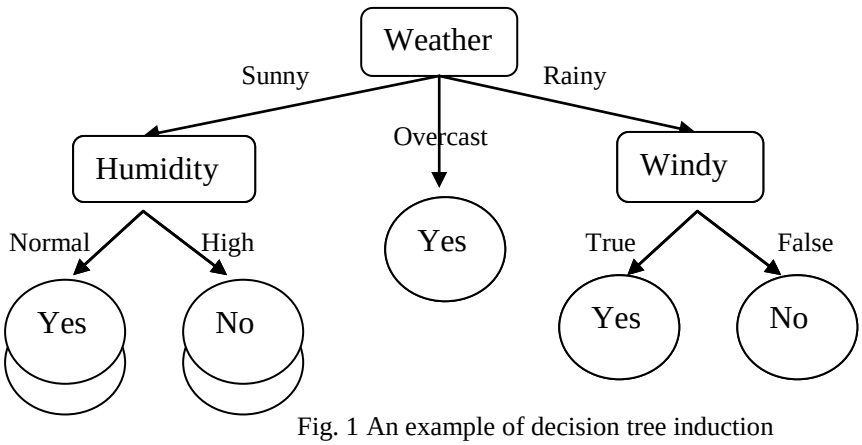

Fig. 1 An example of decision tree induction

Example shows a weather forecasting process which deals with predicting whether it is sunny or rainy and the amount of humidity if it is sunny [3]. Thus this can be applied to determine whether the atmosphere is suitable to play the tennis. So person can easily find the present climate as well as what will be followed by in future and based on that decision can be made whether match can be possible or not. There are various forms of decision tree algorithm available like ID3 (Iterative Dechotomiser), C4.5, CART (Classification and Regression Tree), CHAID (Chi-square automatic interaction detector) etc.

Algorithm: Generate decision tree. Generate a decision from the given training data set.

Input: The data partition D of data set, which is training data items and their class labels, attribute list and a procedure to find the splitting standard.

Output: A decision tree.

Method:

(1) Create a node N;
(2) if test samples in T are all of the matching class, C then
(3) return N as a leaf node labeled as the class C;
(4) if attribute list is blank then
(5) return N as a leaf node labeled with the greater part class in T; // majority voting
(6) apply Attribute selection method(T, attribute list) to
find the "best" splitting standard;
(7) label node N with splitting standard;
(8) if splitting attribute is discrete-valued and multiway splits allowed then // not restricted to binary trees
(9) attribute list attribute list _ splitting attribute; // remove splitting attribute
(10) for each outcome k of splitting criterion// partition the tuples and grow subtrees for each partition
(11) let Tk be the set of data tuples in T satisfying outcome k; // a partition
(12) if Tk is empty then
(13) attach a leaf labeled with the majority class in T to node N;
(14) else attach the node returned by Generate decision tree(Tk, attribute list) to node N;
endfor
(15) return N

Decision tree is useful in generating the model from the input data set and represented in the form of IF-THEN rules. The leaf nodes represent the class label. The IF-THEN rules are easier to understand.

III. NAIVE BAYSIAN CLASSIFICATION

A. Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given data item belongs to a particular class label.

Bayesian classification [1] is based on Bayes Theorem as stated below:

Let X is a data sample whose class label is not known and let H be some hypothesis, such that the data sample X belongs to a specified class.

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

B. Naive Bayesian Classifier

The naive Bayesian classifier [1] works as follows:

1. Let D be the data partition of the dataset forming a training set and their associated class labels. Every tuple is represented by an n-dimensional element vector, X=(x1, x2, x3,.....,xn)
2. Suppose that there are m classes C1, C2, C3...., Cm. Given an unknown tuple, 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 $i \neq j$, above posterior probabilities are computed using Bayes Theorem.

The Naive Bayesian classifier is fast and incremental can deal with discrete and continuous attributes, has exceptional performance in real world problems and can explain its decisions as the sum of informational gains. However, its naivety may lead in poor performance in domains with strong dependencies among attributes.

IV. K-NEAREST NEIGHBOR

Nearest-neighbor classifiers are based on learning by resemblance, that is, by comparing a given test sample with training sample that are similar to it. One of the simplest algorithms described in 1950s, is a mechanism that is used to identify the unknown data point based on the nearest neighbor whose value is already recognized, easy to understand but has an implausible work in fields and practice especially in classification. Nearest- neighbor classifiers are based on learning by analogy.

For a data sample x to be classified, its k-nearest neighbors are searched, and this makes a neighborhood of x. If k is too small, then the nearest-neighbor classifier may be susceptible to over fitting because of noise in the training data. On the other hand, if k is too large, the nearest-neighbor classifier may misclassify the test instance because its list of nearest neighbors may include data points that are located far away from its neighborhood.

k-nn fundamentally works on the postulation that the data is connected in a feature space. Hence all the points are contained in it, in order to find out the distance among the points Euclidian distance or Hamming distance is used according to the data type of data classes used [4]. Here a single number k is given which is used to find the total number of neighbors that determine the classification. If the value of k=1, then it is simply called as nearest neighbor. k-nn requires:

- An integer k
- A training data samples
- A metric to measure closeness

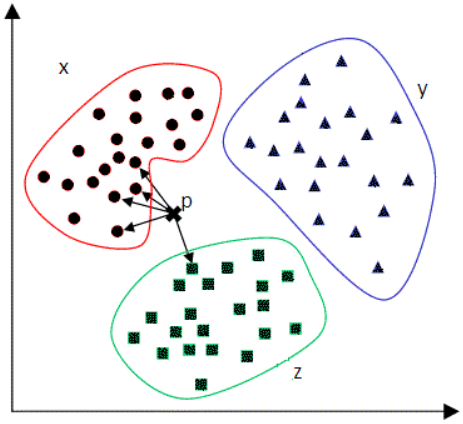

Fig. 2 An example of k-nn classifier

Above example shows three classes are shown x,y and z. Now it is required to find out the class for sample p, which has five tuples. Here k=5 and the Euclidean distance is measured and it is found that four of the tuples are falling in the class label x, while single tuple belongs to z. So the sample p is assigned to class x, the principal class for that sample.

k-nn classifier is relatively easy to use and so it makes the utilization process faster. So the major advantage is that training model can be built faster. Large training data can be determined and so is a strong mechanism [5].

TABLE I

Method/ Parmeter	**Decision Tree**	**Naïve Bayesian**	**KNN**
Understandibility	Simple to understand and generate	Easy to Understand and build	Easy to understand
Data Type	Numerical and categorical	Numerical and categorical	Numerical and categorical
Determinstic/ Non deterministic	Deterministic	Non Deterministic	Non Deterministic
Effectiveness on	Large data	Huge data	Large data
Applicable	Pattern Recognition, Sequence Recognition, Financial Applications	Text Classification, Spam Filtering	Text Classification, Decision Making

Above table shows the comparison among Decion Tree, Naive Bayesian and k-nn techniques of data mining.

V. CONCLUSIONS

Owing to our study on comparison among data mining classification's algorithms (k-nn, Bayesian and Decision tree) and reviewing the time complexity of the above mentioned algorithms we conclude that all the decision Tree's algorithms have less error rate and it is the easier algorithm as compared to k-nn and Bayesian. The acquaintance in Decision Tree represented in the form of [IF-THEN] rules which are easier for humans to comprehend. The disadvantages of decision tree algorithm are typically requiring certain knowledge of statistics and experience to complete the process accurately. It

is also complicated to include variables on the decision tree, exclude redundant information. As stated earlier, there are many specific decision-tree algorithms. CART algorithm of decision tree is the greatest algorithm for classification of data, which has shortest execution time. The outcome to predictive technique on the same dataset showed that Decision Tree supersede others and Bayesian classification having the same accuracy as of decision tree but other predictive algorithms like k-nn, Neural Networks, Classification based on clustering are not giving good results. From our study based on the previously researches we conclude the fact that among (Decision tree, k-nn, Bayesian) algorithms in data mining, k-nn is having lesser accuracy whereas Decision tree and Bayesian are equal. But if Decision tree algorithm has merged with genetic algorithm then in this way the accuracy of the Decision tree algorithm will improve and become more powerful and it will come up to be the best model approach among the other two algorithms. The effectiveness of results using k-nn can be improved by raising the number of data sets and for Bayesian classifier by raising the attributes. For time issue, researches statistics we conclude that the faster algorithm for classifier respectively is: Naive Bayes algorithm, Decision tree and lastly k-nn algorithm that mean the last one is the slowest algorithm for classifier.

REFERENCES

[1] introduction to data mining – pearson

[2] Jiawei Han and Nicheline Kamber, Data Mining Concepts and Techniques, 2nd ed., San Francisco, CA, Elsevier, 2006.

[3] Top 10 algorithms in data mining XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · *Dan Steinberg © Springer-Verlag London Limited 2007*

[4] RAJ, M. A. 2012. Mrs. Bincy G, Mrs. T. Mathu. Survey on common data mining classification Technique. International Journal of Wisdom Based Computing, 2.

[5] Survey of Classification Techniques in Data Mining: Thai Nu Phyu

