ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/102065, October 2016*

Classification Algorithms of Data Mining

K. Deeba* and B. Amutha

Department of Computer Science and Engineering, SRM University, Kattankulathur, Chennai - 603203, Tamil Nadu, India;
deeba.k@ktr.srmuniv.ac.in, amutha.b@ktr.srmuniv.ac.in

Abstract

Objectives: To make a comparative study about different classification techniques of data mining. **Methods**: In this paper some data mining techniques like Decision tree algorithm, Bayesian network model, Naive Bayes method, Support Vector Machine and K-Nearest neighbour classifier were discussed. **Findings**: Each algorithm has its own advantages and disadvantages. Decision tree technique do not perform well if the data have smooth boundaries. The Naive Bayesian classifier works with both continuous and discrete attributes and works well for real time problems. This method is very fast and highly scalable. The drawback of this technique is when a data set which has strong dependency among the attribute is considered then this method gives poor performance. KNN can perform well in many situations and it particularly suits well for multi-model classes as well as applications in which an object can have many labels. The drawback of KNN is it involves lot of computation and when the size of training set taken is large then the process will become slow. Support vector machine suites well when the data need to be classified into two groups. **Application**: This paper is to provide a wide range of idea about different classification algorithms..

Keywords: Bayesian Network, Data Mining, Decision Tree, K-Nearest Neighbour Classifier, Naive Bayes, Support Vector Machine

1. Introduction

Data mining[1,2] is the process of analysing the existing data and extracting useful information out of it. Sometimes data mining is used to predict the future based on the existing data. It finds the relation between the existing data and based on the relation it predicts the outcome of the remaining data. There are several methodologies used for these problems like classification, clustering, regression, rule generation etc. Classification is a type of data mining algorithm used to predict class labels and classify the data to a particular class based on training data sample and then is used to classify the new testing data sets. This paper focuses on various techniques used for classification of data that are commonly used in the field of data mining.

2. Decision Tree Algorithm

Decision tree[3,4] algorithm adopts a super incumbent model to create a tree structure from the given data set where each node represents attributes test or conditions and final leaf node represents the test results or classes. The construction of decision tree is done by divide and conquer strategy. The attribute used here should be of categorical and if it is of continuous values then it had to be converted to discrete values before starting the process. Initially all samples are on the single root node and then

**Author for correspondence*

the remaining nodes are created based on the attribute partitioning condition. This is a recursive process and the stopping condition for this are:

1. If all the samples in each node belongs to same class
2. If there are no more samples to be portioned
3. If there is no further division to be made in the given sample

The algorithm for decision tree is given as:

Input for the method: Set of training samples S with their class attributes, Attribute List, Splitting criteria and Attribute selection method

Output of the method: A Decision tree

Decision tree method:

1. Initially create a new node N
2. If all the data samples are in same category C, then return N and mark it as class C
3. If attribute list of node is empty, the return N and label with class C whichever is major in that set of samples in the node
4. Using attribute selection method find out the best Splitting criteria
5. Use Splitting criteria to mark N
6. In attribute list remove the splitting attributes
7. Create new node N with the resultant set of attributes and continue the process.
8. Return N.

In this method all the attributes should be classified hence the continuous values have to be discretized before starting the process. This method is easy to understand, execute and validate. Validation of the algorithm can be done using simple statistical tests. The major drawback of the algorithm is that the algorithm does not works well if the data have smooth boundaries or if the data have lot of un correlated values.

3. Bayesian Network

Bayesian network[5,6] is a kind of probabilistic method of rule generation. This method will derive a directed acyclic graph that describes the dependency relationship among the variables. The learning process is divided into two parts:

1. Construction of directed acyclic graph between the dependent variables
2. Deriving conditional probability distribution among the dependent variables

Directed acyclic graph consists of set of nodes that represent the random variables and edges connecting these nodes represent the probabilistic dependency between the variables. The important thing is, the directed graph contains of nodes X_i that are assumed to be independent of their parent node Pa_i, i.e. If $A(X_i)$ indicates any nodes constitute the non Xi offspring node set, $\prod_i$ represented by variable Xi parent node set, π_i (or Pa_i) represent $\prod_i$ configuration , Pa_i indicating a specific configuration. For every X_i will have a subset $\prod_i \subseteq \{X_i,...,_{Xi-1}\}$, and makes every X_i and $A(X_i)=\{ X_i,...,X_{i-1}\}/\prod_i$ are a piece of independent in the given $\prod_i$ promise. So, there will be $P(x_i|x_i,...,x_{i-1})=P(x_i/\pi_i)$ for any of the X. Therefore,

$$P(x) = \prod_{i=1}^{n} P(x_i/\text{parent of } x_i(\pi_i)) \quad (1)$$

where, the set of variables ($\prod_i$,..., $\prod_n$)corresponding to the Bayesian network parent nodes (Pa_1,...,Pa_n).

Conditional probability distribution among the dependent variables mapped to a conditional probabilistic table. It use $P(X_i/\pi_i)$ which describes the relationship between any node to its parent node conditional probability. Because of the mutual relationship between the nodes in conditional probability table, Bayesian network can be expressed as joint probability distribution.

4. Naive Bayesian Classification

The naive Bayesian classifier[7,8] is a type of probabilistic classifier. This method uses Bayes' theorem and also assumes that each and every features in a class are highly independent, that is the appearance of a feature in a particular category is not connected with to the presence of any other feature. The naive Bayesian classification done based on the prior probability and likelihood of a tuple to a class.

The working of naive Bayesian classifier[6,10] is given as follows:

1. The training data samples are partitioned based on class labels. Each data partition id denoted using class labels Ci where i=1,2,...,m and each class will have set of tuples represented as Xj where j=1,2,...,n.
2. After training process, if an unknown tuple X is given for classification then the classifier will find the posterior probability of Ci for give tuple X and assign X to class Ci if and only if posterior probability of Ci given X is greater than posterior probability of Cj given X where 1<=j>=m and j not equal to i. The posterior probability of Ci for give tuple X can be calculated as,

Posterior probability (Ci given X)=(Likelihood of tuple X with class Ci * Class prior probability Ci)/ Prior probability of tuple X. (2)

The Naive Bayesian classifier works with both continuous and discrete attributes and works well for real time problems. This method is very fast and highly scalable. The drawback of this technique is when a data set which has strong dependency among the attribute is considered then this method gives poor performance.

5. K-Nearest Neighbour (KNN)

KNN[9,10] is a classification method which very simple and works practically. The training process is very simple. The training sample consists of set of tuples and class labels associated with that. This algorithm works for arbitrary number of classes. KNN uses distance function to map the samples with classes.

Classification process of KNN will find the distance between the given test instance X with that of existing samples y1,y2,...,yk. The nearest neighbours are found and based on the voting of neighbours, the majority neighbourhood class is assigned to the test samples. The distance between the samples can be found by Euclidean method or Manhattan method or Minkowski method, if the values are continuous. If the value used is categorical the Hamming distance is used. The probability of assigning a sample X to that of a class C is based on the number of neighbours considered, denoted as k.

$$\text{Probability of X to a class C} = \frac{\sum_{i=1}^{k} dis\tan ce(c, c(yi))}{k} \quad (3)$$

This will improve the probability estimates of KNN which in turn will improve the performance of classification. KNN can perform well in many situations and it particularly suits well for multi-model classes as well as applications in which an object can have many labels. The drawback of KNN is it involves lot of computation and when the size of training set taken is large then the process will become slow.

6. Support Vector Machine

Support vector machine[11,12] is a classification model based on supervised training method for binary classification. Here the training samples belong to any one of two classes. Based on the training data samples, Support Vector Machine builds a prediction model that will classify the new sample properly to any one of the two classes. Here input to the system is the training samples $x_1, x_2 ... x_n$ with the class labelled as y, and the samples are mapped in the plane. Then a proper hyper plane is identified that optimize the classification result. The SVM uses mapping

function called kernel function ϕ. There are four basic kernel function given as,

Linear function:

$$K(x_i,x_j)= \phi(x_i^{Transpose}).\ \phi(x_j) \quad (4)$$

Polynomial function:

$$K(x_i,x_j)=(\gamma .\ \phi(x_i^{Transpose}).\ \phi(x_j)+r)^d,\ \gamma>0 \quad (5)$$

Radial basis function (RBF):

$$K(x_i,x_j) = \exp(-\gamma ||x_i-x_j||^2),\ \gamma>0 \quad (6)$$

Sigmoid function:

$$K(x_i,x_j) = \tanh(\gamma .\ \phi(x_i^{Transpose}).\ \phi(x_j)+r) \quad (7)$$

where, γ ,r and d are kernel parameters. This prediction model is simple and very accurate. This is used to model real time problems which are more complex and have many attributes. The major drawback of this model is it suits only for binary classification.

7. Results

The aim of classification algorithm is to generate a system which classifies the data accurately based on the training data set. In this paper some data mining techniques like Decision tree, Bayesian network, Naive Bayes, Support Vector Machine and K-Nearest neighbour classifier were discussed. Decision tree method easy to understand, execute and validate. Validation of the algorithm can be done using simple statistical tests. The major drawback of the algorithm is that the algorithm does not works well if the data have smooth boundaries or if the data have lot of un correlated values.

The Bayesian network performance is equal to that of the decision tree algorithm. The Naive Bayesian classifier works with both continuous and discrete attributes and works well for real time problems. This method is very fast and highly scalable. The drawback of this technique is when a data set which has strong dependency among the attribute is considered then this method gives poor performance. KNN can perform well in many situations and it particularly suits well for multi-model classes as well as applications in which an object can have many labels. The drawback of KNN is it involves lot of computation and when the size of training set taken is large then the process will become slow. Support vector machine suites well when the data need to be classified into two groups. Thus, this paper gives a comparative study about different classification techniques of data mining.

8. References

1. Tan P-N, Steinbach M, Kumar V. Introduction to data mining. Pearson New International Edition; 2006 Mar.
2. Han J, Kamber M, Pei J. Data mining: Concepts and techniques. 3rd ed. Morgan Kaufmann Publishers; 2011 Jul.
3. Gao Y-Y, Ren N-P. Data mining and analysis of our agriculture based on the decision tree. ISECS International Colloquium on Computing, Communication, Control, and Management. 2009; 2:134–8.
4. Goltz E, Arcoverde GFB, de Aguiar DA, Rudorff BFT, Maeda EE. Data mining by decision tree for object oriented classification of the sugar cane cut kinds. IEEE International Geoscience and Remote Sensing Symposium. 2009; 5:v405–8.
5. Zhang L, Zhang J, Sun Y. The construction and application of Bayesian network in data mining. 6th International Conference on Information Management, Innovation Management and Industrial Engineering IEEE. 2013; 3:501–3.
6. Patankar B, Chavda V. A comparative study of decision tree, Naive Bayesian and KNN classifier in data mining. International Journal of Advanced Research in Computer Science and Software Engineering. 2014 Dec; 4(12).

7. Jiang L, Zhang H, Su J. Learning K-Nearest Neighbour Naive Bayes for ranking. Proceedings of First International Conference ADMA 2005, Wuhan: China; 2005 Jul 22–24. p. 175–85
8. Su J, Zhang H. Full Bayesian network classifiers. Proceedings of the 23rd International Conference on Machine Learning (ICML) ACM; 2006. p. 897–904.
9. Gao S, Li H. Breast cancer diagnosis based on Support vector machine. 2012 International Conference on Uncertainity Reasoning and Knowledge Engineering IEEE; 2012. p. 240–3.
10. Rajeswari V, Arunesh K. Analysing soil data using data mining classification techniques. Indian Journal of Science and Technology. 2016 May; 9(19).
11. Purusothaman G, Krishnakumari P. A survey of data mining techniques on risk prediction: Heart disease. Indian Journal of Science and Technology. 2015 Jun; 8(12).
12. Vedanayaki M. A study of data mining and social network analysis. Indian Journal of Science and Technology. 2014 Nov; 7(S7):185–7.