ISSN 2278-3091

**International Journal of Advanced Trends in Computer Science and Engineering**, Vol.3 , No.5, Pages : 493- 497 (2014)
*Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh*

# Review on Classification Techniques in Data Mining

**Hari Krishna Kanagala[1], Dr JayaramakrishnaiahVemula[2] , Prof. Ramchand H Rao Kolasani[3]**

[1]Assistant Professor,Vignan's Lara Institute of Technology & Science, Vadlamudi, India, harikanagala@gmail.com
[2]Associative Professor, ASN Degree College, Tenali, India, jkvemula@gmail.com
[3]Professor, ASN Degree College, Tenali, India, ramkolasani@gmail.com

**Abstract:**Data mining is the knowledge discovery process by analysing the large volumes of data from various perspectives and summarizing it into useful information. Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on analysis of set of training data. . In this paper, we present different kinds of classification techniques. Those techniques are Decision Tree Classification, Bayesian Classification, Bayesian Belief Networks and, nearest neighbour classifier.

**Key words:** Bayes Net, Decision Tree, Naïve Bayes, K-Nearest Neighbour

## INTRODUCTION

There are two forms of data analysis that can be used for extract models describing important classes or predict future data. These two forms are Classification and Prediction[4].

These data analysis help us to provide a better understanding of large data. Classification predicts categorical and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation [4].

*Classification example:*

Following are the examples of cases where the data analysis task is Classification:

- A bank loan officer wants to analyse the data in order to know which customer (loan applicant) are risky or which are safe.
- A marketing manager at a company needs to analyse to guess a customer with a given profile will buy a new computer.

In both of the above examples a model or classifier is constructed to predict categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data [4].

*Prediction example:*

Following are the examples of cases where the data analysis task is Prediction:

Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bother to predict a numeric value. Therefore the data analysis task is example of numeric prediction. In this case a model or predictor will be constructed that predicts a continuous-valued-function or ordered value [4].

## CLASSIFICATION

The Data Classification process includes the two steps:

- Building the Classifier or Model: The model is represented in the form of classification rules, decision trees and mathematical formula.
- Using Classifier for Classification: The model is used for classification. The accuracy of the model is estimated on the test data and is considered acceptable, this model can be used to classify the data tuples for which the class label is not known.

For example, the following table 1 shows the database of customer credit information. The rules can be used to categorize future data samples, as well as provide a better understanding of the database contents [4].

Table 1: Training Data

| Training Data | | | |
|---|---|---|---|
| Name | Age | Income | Credit_rating |
| A | <30 | Low | Fair |
| B | <30 | Low | Excellent |
| C | 30-40 | High | Excellent |
| D | >40 | Medium | Fair |
| E | >40 | Medium | Fair |
| F | 30-40 | High | Excellent |

Now the classification rules are used to identify the customers as having either excellent or fair credit_rating. The rules are as follows.

If age 30-40 AND income = "High" THEN credit_rating= Excellent

These rules can be used to categorize the data samples with in the test data shown in table 2.

ISSN 2278-3091

**International Journal of Advanced Trends in Computer Science and Engineering**, Vol.3 , No.5, Pages : 493- 497 (2014)
*Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh*

Table 2: Test Data

| Test Data | | | |
|---|---|---|---|
| Name | Age | Income | Credit_rating |
| G | >40 | high | Fair |
| H | <30 | low | Fair |
| I | 30-40 | high | Excellent |

If the accuracy of the model on the test data is accepted, the classification rules are used to classify the unknown sample data as follows. If the name is J, age is 30-40 and the income is high then this model classifies the credit_rating as excellent.

In this paper, we present different kinds of classification techniques such as Decision Tree Classification, Bayesian Classification, Bayesian Belief Networks and nearest neighbour classifier [4].

**Decision Tree**

The decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on attribute, each branch denotes the outcome of test and each leaf node holds the class label. The topmost node in the tree is the root node. Different measures can be used to estimate the best splitting of attribute. These measures are Information Gain, Gain ratio, Gini-index, and so on. The following figure 1 is a decision tree for concept buys_computer, that indicates whether a customer at a company is likely to buy a computer or not.

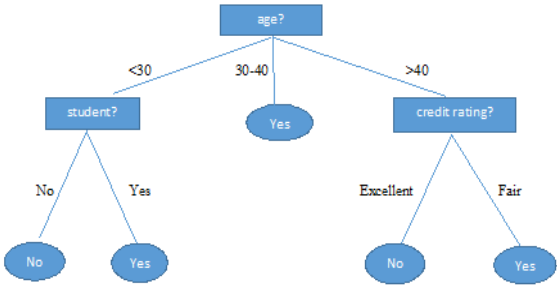


Fig 1: Decision Tree

*Advantages* Rules can be generated that are easy to interpret and understand. It is scalable for large database because the tree size is independent of the database size. Each tuple in the database must be filtered through the tree, and time is proportional to the height of the tree [3].

*Disadvantages* It is does not handle continuous data. Handling missing data is difficult because correct branches in tree could not be taken the labels [3].

In this classification method, different types algorithms are used to classify the data sets such as ID3(Iterative Dichotomiser), C4.5(a Successor of ID3), and Classification and Regression Trees(CART). J. Ross Quinlan in 1980 developed a decision tree algorithm. This Decision Tree Algorithm is known as ID3 (Iterative Dichotomiser). Later he gave C4.5 which was successor of ID3. ID3 and C4.5 adopt a greedy approach. In this algorithm there is no backtracking, the trees are constructed in a top down recursive divide-and-conquer manner.

*ID3 Algorithm*

Iterative Dichotomiser 3 is a simple decision tree learning algorithm introduced in 1986 by Quinlan Ross. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. ID 3 uses the Information Gain as the measure for best split of the node. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise and it is serially implemented. Thus pre-processing of data is carried out before building a decision tree model with ID3 [1].

*C4.5 Algorithm*

It is an improvement of ID3 algorithm developed by Quilan Ross in 1993. It is also like ID3. Pruning takes place in C4.5 by replacing the internal node with a leaf node thereby reducing the error rate. It accepts both continuous and categorical attributes in building the decision tree. It has an enhanced method of tree pruning that reduces misclassification errors due to noise and too many details in the training data set. Like ID3 the data is sorted at every node of the tree in order to determine the best splitting attribute. C4.5 uses gain ratio impurity measure to evaluate the splitting of attribute [1].

The algorithm C4.5 has following advantages.

- Handling attributes with different costs.
- Handling training data with missing attribute values-C4.5 allows attribute values to be marked as „?" for missing. Missing attribute values are simply not used in gain and entropy calculations.
- Handling both continuous and discrete attributes- in order to handle continuous attributes,C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Pruning trees after creation- C4.5 goes back through the tree once it has been created and

ISSN 2278-3091

**International Journal of Advanced Trends in Computer Science and Engineering**, Vol.3 , No.5, Pages : 493- 497 (2014)
*Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh*

attempts to remove branches that do not help by replacing them with leaf nodes.

*CART Algorithm*

It stands for classification and regression trees and was introduced by Breiman in 1984.It builds both classifications and regression trees. The classification tree construction by CART is based on binary splitting of the attributes. CART uses gini index as the impurity measure to select the splitting attribute. CART also used for regression analysis with the help of the regression trees. The regression analysis feature is used in forecasting a dependent variable given a set of predictor variables over a given period of time. It uses many single-variable splitting criteria like gini index, symginietc and one multi-variable in determining the best split point and data is stored at every node to determine the best splitting point [1].

**Bayesian Classification**

Bayesian classification is based on Baye's Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifier are able to predict class membership probabilities such as the probability that a given tuple belongs to a particular class [7].

Equation (1) refers the Bayes theorem allows us to express the posterior probability in terms of the prior probability $P(C_i)$, the class-conditional probability $P(X|C_i)$, and the evidence $P(X)$:

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})} \quad (1)$$

When comparing the posterior probability for different values of $C_i$, $P(X)$ is always constant and can be ignored. The prior probability $P(C_i)$ can be estimated from the training set by computing the fraction of training records that belong to each class. To estimate the class-conditional probability $P(X|C_i)$, naïve bayes classifier is used.

A naïve bayes classifier estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label $C_i$. It can be stated in (2).

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i) \quad (2)$$

For a categorical attribute $X_i$, the conditional probability $P(X_i = x_i| C_i=c)$ is estimated according to the fraction of training instances in class c that take on a particular attribute value $x_i$ [7].

For continuous attribute, a Gaussian distribution is chosen to represent the class-conditional probability.

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

Where μ is the sample mean of $X_i$ and the σ is the sample variance of the training records.

To predict the class label of X, $P(X|C_i)$ $P(C_i)$ evaluated for each class Ci. The classifi0er predicts that the class label of tuple X is the class Ci if and only if $P(X|C_i)$ $P(C_i) > P(X|C_J)$ $P(C_J)$ for $1 \le j \le n$, $j \ne i$. In other words, the predicted class label is the class Ci for which $P(X|C_i)$ $P(C_i)$is the maximum [5].

*Advantage:* It only requires a small amount of training data to estimate the parameters (means and variances of the variables) 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 [3]. It is easy to implement. It generates the good results in most of the cases. Naïve bayes classifiers are robust to isolated noise points because such points are averaged out when estimating conditional probabilities from data. Naïve Bayes classifiers can handle missing values by ignoring the example during model building and classification.

*Disadvantage:* If the class-conditional probability for one of the attribute is zero, then the overall posterior probability for the class vanishes. Because of class conditional independence assumption, there is a loss of accuracy. Practically more dependencies exist among the variables. For example salary and age. In other words, Correlated attributes can degrade the performance of naïve bayes classifiers because the conditional independence assumption no longer holds for such attributes [7].

**Bayesian Belief Networks**

A Bayesian belief network provides a graphical representation of the probabilistic relationships among a set of random variables [7]. The two key elements are:

1. A directed acyclic graph encoding the dependence relationships among a set of variables.
2. A Probability table associating each node to its immediate parent node.

Conditional independence property: A node in a Bayesian network is conditionally independent of its non-descendants, if its parents are known.

*Advantage:* Instead of requiring all the attributes to be conditionally independent given the class, the

ISSN 2278-3091

**International Journal of Advanced Trends in Computer Science and Engineering**, Vol.3 , No.5, Pages : 493- 497 (2014)
*Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh*

Bayesian belief networks allows us to specify which pair of attributes are conditionally independent.

Bayesian networks are well suited to dealing with incomplete data [7].

*Disadvantage:* Constructing the network can be time consuming and requires a large amount of effort [7].

**K-Nearest Neighbour Classifier**

This approach is used to find all the training examples that are relatively similar to the attributes of the test example. These examples, which are known as nearest neighbours, can be used to determine the class label of the test example. The nearest neighbour classifier represents each example as a data point in a d-dimensional space, where d is

the number of attributes. Given the test example, Proximity can be computed to the rest of the data points in the training data. The K-nearest neighbours of a given example z refers to the k points that are closest to z. The data point is classified based on the class labels of its neighbours. If the neighbours have more than one label, the data point is assigned to the majority class of its nearest neighbours. If neighbours have a tie between the classes, one of the class can be randomly choose to classify the data point [7].

*Advantage:* Nearest Neighbour classifiers do not require model building.

*Disadvantage:* If k is too small, then the nearest-neighbour classifier may be susceptible to overfitting because of noise in the training data. If k is too large, the nearest-neighbour classifier may misclassify the test instance because its list of nearest neighbours may include data points that are located far away from its neighbourhood. Classifying a test example can be expensive because we need to compute the proximity values individually between the training and test examples. Because the classification decisions made locally, nearest neighbour classifiers susceptible to noise.

## PERFORMANCE EVALUATION

The results obtained after running of the classification techniques for Iris data set which consists of 5 attributes and 150 instances. To construct the algorithms, we use Waikato Environment for Knowledge Analysis (WEKA version 3.6.10), an open source data mining tool which was developed at University of Waikato New Zealand. WEKA is an open source application that is freely available under the GNU general public license agreement. This experiment is performed on Duo Core with 2.10 GHz CPU and 4G RAM. The result for each classification algorithms are shown and described below.

The following table 3 is the comparison of different classification algorithm accuracy results for Iris data set. The results were analysed on data using 10-fold cross validation to test the accuracy.

Table 3: Accuracy results

| Classifier (150 Instances) | Correctly Classified Instances | Incorrectly Classified Instances | Time taken (sec) |
|---|---|---|---|
| J48 | 96% | 4% | 0 |
| Cart | 95.3333 % | 4.6667 % | 0.01 |
| Naïve Bayes | 96% | 4% | 0 |
| Bayes net | 92.6667 % | 7.3333 % | 0 |
| KNN | 95.3333 % | 4.6667 % | 0 |

The following table 4 shows the comparison of error rates of different classification techniques on the Iris data set.

Table 4: Error Rates

| Classifier (150 Instances) | MAE | RMSE | RAE | RRSE |
|---|---|---|---|---|
| J48 | 0.035 | 0.1586 | 7.8705 % | 33.6353 % |
| Cart | 0.0437 | 0.1752 | 9.8273 % | 37.1656 % |
| Naïve Bayes | 0.0342 | 0.155 | 7.6997 % | 32.8794 % |
| Bayes net | 0.0454 | 0.1828 | 10.2111 % | 38.7793 % |
| KNN | 0.0399 | 0.1747 | 8.9763 % | 37.0695 % |

MAE: Mean Absolute Error
RMSE: Root mean squared error
RAE: Relative absolute error
RRSE: Root relative squared error

From the tables 3 and 4, for the Iris data set, Naïve Bayes classifier and J48 classifier are more accurate and Naïve Bayes classifier results less error rate.

## CONCLUSION

We can conclude that different classification algorithms are suitable at different situations. Decision trees having the less error rate when compared to other classification algorithms. In specific to a particular decision tree algorithm, CART algorithm is more accurate for the large volume of data sets. If the attributes are conditionally independent to each other, Naïve Bayes classification is more accurate. Bayes belief networks are more accurate with incomplete data. Nearest neighbour algorithm have less accurate when compared to the other classification techniques.

## REFERENCES

[1] AnjuRathee, Robin prakashmathur,"Survey on Decision Tree Classification algorithms for the Evaluation of Student Performance" International

Journal of Computers & Technology Volume 4 No. 2, March-April, 2013, ISSN 2277-3061.

[2] DelveenLuqmanAbd AL-Nabi, ShereenShukri Ahmed, "Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)", Computer Engineering and Intelligent Systems, Vol.4, No.8, 2013ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online).

[3]Dr.A.Bharathi, E.Deepankumar, "Survey on Classification Techniques in Data Mining" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 2 Issue: 7 1983 – 1986 ISSN: 2321- 8169.

[4] Jiawei Han,MichelineKambar, JianPei,"Data Mining Concepts and Techniques" Elsevier Second Edition.

[5] K.S.Thirunavukkarasu, Dr.S.Sugumaran "Analysis of Classification Techniques in Data Mining" International Journal of Engineering Sciences & Research Technology ISSN: 2277-9655

[6] MohdFauzi bin Othman,ThomasMoh Shan Yau "Comparison of Different Classification Techniques Using WEKA for Breast Cancer" IFMBE Proceedings Vol. 15, pp. 520-523, 2007.

[7] Pang-Ning Tan, Vipin Kumar, Michael Steinbach, " Introduction to Data Mining" Pearson.

[8] SyedaFarhaShazmeen, Mirza Mustafa Ali Baig, M.ReenaPawar, " Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727Volume 10, Issue 6 (May. – Jun. 2013), PP 01-06.

[9] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining" Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I March 18 - 20, 2009, Hong Kong.

.