IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 3, March 2016.
www.ijiset.com
ISSN 2348 – 7968

A Comparative Study on Classification Algorithms in Data Mining

Dr.B.Srinivasan [1] and K.Pavya [2]
[1]Department of Computer Science, Gobi Arts & Science College,
Bharathiar University, Tamil Nadu, India
[2]Department of Computer Science, Vellalar College for Women,
Bharathiar University, Tamil Nadu, India

Abstract

Databases are rich with unknown information that can be used for intelligent decision making. Classification and prediction are two form of data analysis that can be used to extract models describing important data classes or to predict future trends. Such analysis can help provide us with a better understanding of the data at a large. Whereas classification predicts definite labels, prediction models continuous valued functions. Several major kinds of classification algorithms including C4.5, ID3, k-nearest neighbor, Naive Bayes, ANN and SVM are used for classification. This paper provides a Comparative study of different classification algorithms and their advantages and disadvantages.

Keywords: *C4.5, ID3, k-nearest neighbor classifier, SVM, ANN, Naive Bayes*

1. Introduction:

Classification maps data into predefined groups or classes. It is frequently referred to as supervised learning because the classes are determined before examining the data. Many classification methods have been proposed by researchers in machine learning, pattern recognition, and statistic. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, increasing scalable classification and prediction techniques capable of handling large disk-resident data. Classification and Prediction have numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis.

2. ID3 Algorithm:

The basic strategy used by ID3 is to choose splitting attributes with the highest information gain first. The quantity of information associated with an attribute value is related to the probability of event. The concept used to quantify information is called entropy. Entropy is used to measure the amount of uncertainty or chance in a set of data. When all data in a set belong to a single class, there is no uncertainty. In this case the entropy is zero(0).The objective of decision tree classification is to iteratively partition the given data set into subsets belong to the same class.

Definition: Given probabilities $p_{1,}p_{2,.....}p_s$ where $\sum_{i=1}^{s} p_i=1$, entropy is defined as

$$H(_{p1,p2,.....ps}) = \sum_{i=1}^{s}(p_i \log(1/p_i))$$

Given a database state, D, H(D) finds the amount of order in that state. when that state is split into s new states $S=\{D_1,D_{2,}\ldots\ldots,D_s\}$,we can again look at the entropy of those states. Each step in ID3 chooses the state that orders splitting the most. A database state is completely ordered if all the tuples in it are in the similar class.ID3 chooses the splitting attribute with the highest gain in information, where gain is defined as the difference between how much information is needed to make a correct classification before the split versus how much information is needed after the split.

The ID3 algorithm calculates the gain of a particular by the following formula:

$$\text{Gain } (D,S)= H(D) - \sum_{i=1}^{s} p(D_i)H(D_i)$$

3. C4.5 Algorithm:

The C4.5 algorithm improves ID3 in the following ways:

Missing data: When the decision tree is built, missing data are simply unobserved. That is the gain ratio is calculated by the other records that have a value for that attribute.

Continuous data: The basic idea is to divide the data into ranges based on the attribute values for that item that are found in the training sample.

Pruning: There are two primary pruning strategies proposed in C4.5

1) Subtree replacement, a subtree is replaced by a leaf node if this replacement results in an error rate close to that of the original tree.

ISSN 2348 – 7968

2) Subtree raising , replaces a subtree by its most used subtree. Here a subtree is raised from its current location to a node higher up in the tree.

Rules: C4.5 allows classification via either decision tree or rules generated from them.

Spliting: This approach uses the GainRatio as opposed to gain.The GainRatio is defined as

GainRatio(D,S) = Gain(D,S)/ H(|D_1|/|D|,……|D_s|/|D|)

For splitting purposes, C4.5 uses the largest GainRatio that ensures a larger than average information gain.

4. k-nearest neighbor Algorithm:

K- nearest neighbors (KNN),a classification scheme based on the distance measure. The KNN technique assumes that the entire training set includes not only the data in the set but also the preferred classification for each item. When classification is to be made for a new item, the distance to each item in the training set must be determined. Only the K closest entries in the training set are considered further. The new item is next placed in the class that contains the most items from this set of K closest items.

5. Naive Bayes Algorithm:

A simple classification method called navie Bayes classification is based on Bayes rule of conditional probability. This approach is called "navie" because it assumes the independence between the various attribute values. Given a data value x_i the probability that is related tuple t_i , is in class C_j is described by

$P(C_j \mid x_i)$.Training data can be used to determine $P(x_i)$,$P(x_i \mid C_j)$, and $P(C_j)$.From these values Bayes theorem allows to estimate the posterior probability $P(C_j \mid x_i)$ and then $P(C_j \mid t_i)$.

Given a training set, the navie Bayes algorithm first estimate the prior probability $P(C_j)$ for each class by counting how often each class occurs in the training data. For each attribute , x_i, the number of occurrences of each attribute value x_i can be counted to determine $P(x_i)$.similarly the probability $P(x_i \mid C_j)$ can be estimated by counting how often each value occurs in the class in the training data.

$P(c|x) = P(x_1|c) \times P(x_2|c) \times \ldots\ldots \times P(x_n|c) \times P(c)$

- P(c|x) is the posterior probability of the class given predictor of class.
- P(c) is called the prior probability of the class.
- P(x|c) is the likelihood which is the probability of predictor of given class.
- P(x) is the prior probability of predictor of class

6. ANN (Artificial Neural Network) Algorithm:

Neural networks are often referred to as artificial neural networks (ANN). The NN is an information processing system that consist of a graph representing the processing system and various algorithm that access that graph. The NN can be viewed as a directed graph with source (output), (input), s ink (output), and internal (hidden) nodes. To perform the data mining task, a tuple is input through the input nodes and the output node determines what the prediction is. Unlike decision trees,which has only one input node, the NN has one input node for each attribute value to be examined to solve the data mining function.

A neural network (NN) model is a computational model consisting of three parts:

1. Neural network graph that defines the data structure of the neural network.
2. Learning algorithm that indicates how learning takes place.
3. Recall techniques that determine how information is obtained from the network.

NNs have been used in speech recognition and synthesis, pattern recognition, medical applications, fault detection, problem diagnosis robot control and computer vision.

7. SVM (Support Vector Machine) Algorithm:

Support Vector Machine(SVM) is based on statistical learning theory and is increasingly becoming useful in data mining. The main idea is to non-linearly map the dataset into a high-dimensional feature space and use a linear discriminator to classify the data. Its success has been demonstrated in the areas of regression, classification and decision tree construction. Unlike other classification techniques, which attempt to minimize error of classification, SVMs incorporate structured risk minimization which minimizes an upper bound on the generalized error. Consider two sets A and B are linearly separable. The idea is to determine from an infinite number of planes correctly separating A and B, the one which will have the smallest generalization error. SVMs select with which minimizes the margin separating the two classes. The margin is defined as the distance between the separating hyperplane to the nearest point of a, plus

ISSN 2348 – 7968

the distance from the hyperplane to the nearest point in B.

Table 1: Advantages and Disadvantages of classification algorithms

S. NO	***ALGORITHMS***	***ADVANTAGES***	***DISADVANTAGES***
1.	ID3	1. It produces the more accuracy result than the C4.5 algorithm. 2. ID3 algorithm generally uses nominal attributes for classification with no missing values. 3.It produces false alarm rate and omission rate decreased, increasing the detection rate and reducing the space Consumption	1 It has long searching time. 2. It takes the more memory than the C4.5 to large program execution.
2.	C4.5	1. It produces the accurate result. 2. It takes the less memory to large program execution. 3. It has short searching time. 4. It takes less model build time	1. Empty branches. 2. Over fitting. 3. Insignificant branches.
3.	k-nearest neighbour	1. It is an easy to understand and easy to implement classification technique. 2. Robust to noisy training data. 3. Training is very fast. 4.It is particularly well suited for smultimodal classes	1. It is sensitive to the local structure of the data. 2. Being a supervised learning lazy Algorithm i.e., runs slowly. 3. Memory limitation.
4.	Naive bayes	1.To improves the classification performance by removing the irrelevant features. 2. Good performance. 3.It is short computational time	1. The naive Bayes classifier requires a very large number of records to obtain good results. 2. It is instance-based or lazy in that they store all of the training samples
5.	Support vector machine	1. Produce very accurate classifiers. 2. Memory-intensive. 3. Especially popular in text classification problems where very high-dimensional spaces are the norm. 4. Less over fitting, robust to noise.	1. SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used (one class against all others, for all classes). 2. Computationally expensive, thus runs slow.
6.	Artificial neural network	1.Requires less formal statistical training, 2.Ability to implicitly detect complex nonlinear relationships between dependent and independent variables 3.Ability to detect all possible interactions between predictor variables 4.The availability of multiple training algorithms	1.It is “black box” nature 2.Greater computational burden 3.Proneness to over fitting 4.The empirical nature of model development

Conclusion:

This paper deals with different classification techniques used in data mining and a study of them.

ISSN 2348 – 7968

These classification algorithms can be implemented on variety of data sets like patient data set, share market data set etc. Hence these classification techniques show how data can be grouped and determined when a new set of data is available. Each technique has both advantages and disadvantages which are given in the paper. Based on the conditions each one as needed can be selected.

References

[1] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers,Second Edition, 2008.
[2] Arun K.Pujari,"Data Mining Techniques, Universities Press(India)Pvt. Ltd.,Third Edition , 2013.
[3] Margaret H.Dunhan, "Data Mining: Introductory and Advanced Topics",Pearson Education,2013
[4] H.Bhavsar, A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231 -2307, 2(4); (2012).
[5] S.Archana et al, "Survey of Classification Techniques in Data Mining",International Journal of Computer Science and Mobile Applications,Feb-2014, Vol.2 Issue.2, Pg.65-71.
[6] Sagar S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms" Oriental Journal of Computer Science & Technology, April.2015,Vol.8,No (1), Pg.13-19.
[7] M . Aly, "Survey on Multiclass Classification Methods", November (2005).
[8] V. Vapnik, "Statistical Learning Theory", Wiley, New York, (1998).
[9] T.Joachims, "Making large-scale support vector machine learning practical", In Advances in Kernel Methods: Support Vector Machines, (1999).
[10] D. Michie, D.J. Spiegelhalter, C.C. Taylor "Machine Learning, Neural and Statistical Classification", February 17, (1994).
[11] Delveen Luqman Abd Al.Nabi, Shereen Shukri Ahmed, "Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)" (ISSN 2222-2863)4(8); (2013).