International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 8 (2017), pp. 2135-2142
© Research India Publications
http://www.ripublication.com

Classification Algorithms on Datamining: A Study

N. Chandra Sekhar Reddy, K. Sai Prasad and A. Mounika

Department of Computer Science Engineering, MLR Institute of Technology, Hyderabad, Telangana, India.

Abstract

Data mining is that the method of analyzing data from completely different views and summarizing it into useful information. Classification could be a data processing technique supported machine learning which is employed to classify each item in a set of data into a group of predefined categories or teams. Classification is method of generalizing the data consistent according to different instances. Several major kinds of classification algorithms including k-nearest neighbor, naïve bays, support vector machines and neural network. This paper provides a comprehensive survey of various classification algorithms and their advantages and disadvantages.

Keywords: Classification, NB, SVM, K-NN.

1. INTRODUCTION

Data mining could be a method of extracting or mining the useful pattern or information and relationships within massive amounts or huge volumes of data. The term data processing is additionally referred to as “Knowledge mining from data” [1]. The general goal of the data mining method is to extract information from a information set and associated it into an comprehensive structure for future use. These tools can include statistical models, mathematical algorithm and machine learning strategies. Consequently, data processing consists of more than collection and managing data, it also includes analysis and prediction. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity [4].

The renowned classification algorithms are k-nearest neighbor, naive bays, SVM and neural network. There are several applications for Machine Learning (ML), the foremost and vital of that is data mining [5]. People are often prone to making mistakes throughout the analyses or, possibly, once making an attempt to ascertain relationships between multiple options. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines [3]. Classification is that the organization of information in given classes and also referred to as supervised classification, the classification uses given class labels to order the objects within the data collection.

2. CLASSIFICATION ALGORITHMS IN DATA MINING:

A. C4.5 Algorithm:

C4.5 is an algorithm accustomed to generate a decision tree developed by Ross Quinlan. C4.5 is an associate extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be often used for classification, and for this reason, C4.5 is usually noted as a statistical classifier One limitation of ID3 is that it is too sensitive to features with massive numbers of values. This should be overcome if you are about to use ID3 as an Internet search agent. I label this problem by borrowing from the C4.5 algorithm, an associate ID3 extension.ID3's sensitivity to options with massive numbers of values is illustrated by Social Security numbers. Since Social Security numbers are distinctive for each individual, testing on its value can always yield low conditional entropy values. However, this is not a useful test. To overcome this problem, C4.5 uses a metric called "information gain," which is defined by subtracting conditional entropy from the base entropy; that is,

$$\text{Gain }(P|X) = E\ (P) - E\ (P|X). \qquad \text{Eq.(1)}$$

This computation does not, in itself, produce anything new. However, it permits you to measure a gain ratio. Gain ratio, defined as

$$\text{Gain Ratio }(P|X) = \text{Gain }(P|X)/E(X), \qquad \text{Eq.(2)}$$

Where (X) is the entropy of the examples relative to the attribute. It has an enhanced method of tree pruning that reduces misclassification errors due noise or excess amount of 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. It uses gain ratio impurity method to evaluate the splitting attribute. Decision trees are built in C4.5 by employing a set of training data or data sets as in ID3. At every node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criteria are that the normalized information gain (difference in entropy) that results from selecting an associate attribute for rendering the data. The attribute with the best normalized data gain is chosen to make the decision. The C4.5 algorithm then resources on the smaller sub lists.

B. Iterative Dichotomiser 3 (Id3) Algorithm:

ID3 algorithm begins with the first set because it is the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set and calculates the entropy IG(A) of that attribute. Then select the attribute that has the small entropy (or largest information gain) worth. The set is S then split by the chosen attribute (e.g. age < 50, 50 <= age < 100, age >= 100) to provide subsets of the data. The algorithm continues to recurse on each subset, considering solely attributes never selected before. Recursion on a subset might stop in one of these cases:

- Every element within the subset belongs to the same class (+ or -), then the node is changed into a leaf and labeled with the class of the examples.
- There are not any additional attributes to be selected, but the examples still does not belong to the same class (some are + and some are -), then the node is changed into a leaf and labeled with the foremost common class of the examples in the subset.
- There are not any examples in the subset, this happens once no example within the parent set was found to be matching a selected value of the chosen attribute, for example if there was no example with age >= 100. Then a leaf is created, and labeled with the most common class of the examples in the parent set.

Throughout the algorithm, the decision tree is formed with every non-terminal node representing the chosen attribute on it the data was split, and terminal nodes representing the class label of the ultimate subset of the branch.

C. Naive Bayes

A Naive Thomas Bayes classifier assumes that the presence or absence of a specific feature is unrelated to the presence or absence of the other feature, given the class variable. [1]

The naive bays classifiers are a family of easy probabilistic classifiers supported applying bays theorem with durable independence assumptions between the choices. Bayesian classifier is functioning on the dependent events and therefore the probability of an happening occurring within the future that may be detected from the previous occurring of the constant event [2].

The naive bays classifier may be a easily applied statistical algorithm providing astonishingly higher results. Bayesian filter has been used widely in building spam filters. The Naïve Bays classifier is predicted on the Bays' rule of conditional probability. It makes use of all the attributes contained within the data, and analyses them on an individual basis like they are equally important and independent of each other. The rule for conditional probability is as follows [2]:

$$P\,(H \mid E) = P\,(E \mid H)\,P\,(H) \,/\, P\,(E)$$

Where P(H|E) is that the conditional probability that hypothesis H is true given an evidence E; P(E|H) the conditional probability of E given H, P(H) the prior probability of H, P(E) the previous probability of E[2].

Evidence split into two parts they are:

$$P(B/A)= \frac{P(A1/B)P(A2/B)\ldots\ldots.P(An/B)}{P(B)} \qquad \text{Eq.(3)}$$

where

A1,A2,A3…….An are totally independent priori.

D. Support Vector Machine:

SVM has develop as one of the most standard and helpful techniques for data classification [7]. It will be used for classify the both linear and non linear data. [3] The target of SVM is to supply a model that predicts the target value of data occurrence in the testing set in which only attributes are given .[8] The classification goal in SVM is to separate the two classes by means of a function prepare from available data and thereby to produce a classifier that will work well on further unseen data. [8] The simplest form of SVM classification is the maximal margin classifier. It is accustomed to solve the foremost classification drawback, namely the case of a binary classification with linear separable training data. [1]

The aim of the maximal margin classifier is to find the hyper plane with the largest margin, i.e., the maximal hyper plane, in real-world problems, training data are not always linear separable. In order to handle the nonlinearly severable cases some slack variables are introduced to SVM so on to tolerate some training errors, with the influence of the noise in training data thereby decreased. This classifier with slack variables is noted to as a soft-margin classifier.

E. K- Nearest Neighbor:

Nearest neighbor classifiers is a lazy learner’s method and is based on learning by analogy. It is a supervised classification technique which is used widely. Unlike the previously described methods the nearest neighbor method waits until the last minute before doing any model construction on a given tuple. In this method the training tuples are represented in N-dimensional space. When given an unknown tuple, k-nearest neighbor classifier searches the k training tuples that are closest to the unknown sample and places the sample in the nearest class.

The K nearest neighbor method is simple to implement when applied to small sets of data, but when applied to large volumes of data and high dimensional data it results in slower performance. The algorithm is sensitive to the value of k and it affects the performance of the classifier. New Field Programmable Gate Arrays (FGPA) architectures of KNN classifiers have been proposed in [6] to overcome this difficulty of classifier to easily adapt to different values of k.

Accuracy in data classification is a major issue in data mining and in order to improve the accuracy of classification, improvements have been made to the K nearest neighbor method. Weighted nearest neighbor classifier (wk-NNC) is one such method which adds a weight to each of the neighbors used for classification. Hamamoto's bootstrapped training set can also be used instead of the training patterns where training pattern is replaced by a weighted mean of a few of its neighbors from its own class of training patterns.

This method proves to improve the accuracy of classification. However the time to create the bootstrapped set is O (n2) where n is the number of training patterns. K-Nearest Neighbor Mean Classifier (k-NNMC) proposed in [9] finds k nearest neighbors for each class of training patterns separately. The classification is done based to the nearest mean pattern. This improvisation proves to show better accuracy of classification in when compared to other techniques using Hamamoto's bootstrapped training set.

F. Neural Network:

A neural network could be a set of connected input/output units within which every association contains a weight related to it. During the training phase, the network learns by adjusting the weights thus an able to predict the right class label of the input. Neural Network learning is also referred to as connectionist learning due to the connections layers units [2]. Error Back Propagation Network (EBPN) cold be a quite feed forward network (FFN) in which Error Back Propagation Algorithm (EBPA) is employed for learning which one in every of the foremost is wide used training algorithm where training is perform in 2 phases:

- Forward phase: During this phase input is presented and output is calculated supported on activation function and at last error is calculated at outer layer.
- Backward phase: During this phase error is send back to the inner layers to regulate the weights.

3. ADVANTAGES AND DISADVANTTAGES OF CLASSIFICATION ALGORITHMS

ALGORITHMS	ADVANTAGES	DISADVANTAGES
C4.5 Algorithm	1. It produces the correct result. 2. It takes the less memory to massive program execution. 3. It takes less model build time. 4. It has short searching time.	1. Empty branches. 2. Insignificant branches. 3. Over fitting.
ID3	1. It produces the high accuracy result than the C4.5 algorithm. 2.ID3 algorithm typically 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.
Naive Bays Algorithm	1. To improves the classification performance by removing the unrelated options. 2.Good Performance 3. It is short computational time	1.The naive bays classifier requires a very large number of records to obtain good results. 2. Threshold value must be tuned

Support Vector Machine Algorithm	1. Less over fitting, robust to noise. 2. Especially popular in text classification problems	1.SVM is a binary classifier. To do a multi-class classification, pair wise classifications can be used 2. Computationally expensive, thus runs slow.
K-Nearest Neighbor Algorithm	1. It is an easy to understand 2. Training is very fast. 3. Robust to noisy training data.	1. Memory limitation. 2. Being a supervised learning lazy algorithm
Neural Network	1.Capable of producing an arbitrarily complex relationship between input and output	1.Do not work well when there are many hundreds or thousands of input features and difficult to understand the model

4. CONCLUSION

This paper deals with numerous classification techniques employed in data mining and a study on every one of them. Data mining is a wide area that integrates techniques from various fields including machine learning, artificial intelligence, statistics and pattern recognition, for the analysis of large volumes of data. Classification methods are typically strong in modeling interactions Each of these methods can be used in various situations as needed where one tends to be useful while the other may not and vice-versa. Each technique has got its own pros and cons as given in the paper. Based on the needed conditions each one as needed can be selected On the basis of the performance of these algorithms, these algorithms can even be wont to observe the detect the natural disasters like cloud explosive, earth quake, etc..

REFERENCES

[1] Jiawei Han, Micheline Kambar, Jian Pei, “Data Mining Concepts and Techniques” Elsevier Second Edition.

[2] Savita Pundalik Teli and Santoshkumar Biradar,— Effective Email Classification for Spam and Non-Spam‖, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, June 2014.

[3] Galit Shmueli, Nitin R.Patel, Peter C.Bruce, “Data Mining Business Intelligence” Wiley India Edition.

[4] S.Archana , Dr. K.Elangovan,‖ Survey of Classification Techniques in Data Mining‖, International Journal of Computer Science and Mobile Applications,

Vol.2 Issue. 2, pg. 65-71 ISSN: 2321-8363, February- 2014.

[5] Thair Nu Phyu, ―Survey of Classification Techniques in Data Mining‖, IMECS,18-20, vol 1,hong kong, March 2009.

[6] Hussain, H.M. ; Benkrid, K. ; Seker, H. ,"An adaptive implementation of a dynamically reconfigurable K-nearest neighbor classifier on FPGA" , Adaptive Hardware and Systems (AHS), 2012 NASA/ESA Conference on June 2012

[7] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power Utility Nontechnical Loss Analysis With Extreme Learning Machine Method", VOL. 23, NO. 3, AUGUST 2008

[8] Vipin Kumar , J. Ross Quinlan, Joy deep Ghosh ,Qiang Yang ,Hiroshi Motoda, Geoffrey J. McLachlan , Angus Ng , Bing Liu, "Survey paper on Top 10 Algorithms in Data Mining", 4 December 2007© Springer-Verlag London Limited 2007.

[9] Viswanath, P. ; Sarma, T.H.,"An improvement to k-nearest neighbor classifier", Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE , Sept. 2011