ISSN (Online) 2456-1290
International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 2, Issue 3, March 2017

A Survey on Non-Parametric Classification Methods

[1]Hema Sree Kotari [2] Ravikumar Penugonda,
[1] Student At RGUKT IIIT-RKVALLEY, [2] Assistant Profesor AT RGUKT IIIT

***Abstract:*— In this modern era, along with the rapid development of computer hardware – increase processing speed and the size of the memory available – there has been a rapid development of data analysis methods that previously were not possible for practical use due to the required computing power. This paper will provide a survey of non-parametric techniques. It does not focus greatly on the technical aspects of each method beyond the headline advantages and disadvantages.**

***Keywords:*----- Data mining, classification, non-parametric, Neural Networks.**

I. INTRODUCTION

Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data pattern analysis [1]. It is a wide area that combines techniques from various fields such as machine learning, artificial intelligence, statistics and pattern recognition, for the analysis of large volumes of data. There have been a large no of data mining algorithms embedded in these fields to perform different data analysis tasks. A lot of time is spent searching for the most performing data mining algorithms. Among the major three approaches (One and Two way tables, interactive analysis, Non parametric data mining) to understand the nuisance of data, Non parametric is the fastest approach .The study set out to get an overview of non-parametric data mining algorithms. The Classification is the one of the major role in Data mining. Basically classification is a 2-step process; the first step is supervised learning for the sake of the predefined class label for training data set. Second step is classification accuracy evaluation [2]. A Large number of classification techniques were developed by the means of statistical method, logical method, symbolic method, perceptron based method, fuzzy based concept, neural network based and rule based etc. Classification techniques consist of following steps to perform the mining process.

a. Data Acquisition
b. Data pre-processing
c. Data Presentation
d. Decision Making
e. Performance Evaluation

Supervised Learning

- 1. Learning from examples, concept learning
- Step 1: Using learning algorithms to extract rules from the training data. The training data are the precalssified examples.
- Step 2: Evaluate the rules on the test data. Usually split known data into training sample and test sample.
- Step 3: Apply the rules to new data.
- 2. Instance – based learning: Predict class label of new sample using the trained data directly.
- Nearest neighbor
- Bayesian classification
- 3. Regression: Learning functions or predicting numeric value.

Unsupervised Learning

No class label, finding common patterns, grouping similar examples.

- Statistical clustering
- Agglomerative hierarchical clustering
- Conceptual clustering

Non-parametric classification

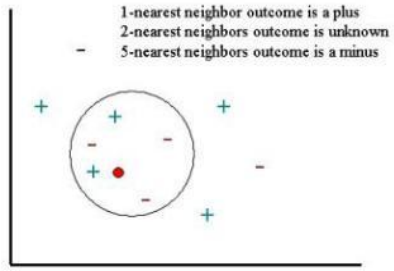
K- nearest Neighbor

The nearest neighbor (NN) algorithm distinguishes the classification of test sample on the basis of its nearest neighbor whose label is already known. M. Cover and P. E. Hart purpose k nearest neighbour (KNN) in which nearest neighbor is computed on the basis of estimation of k that indicates how many nearest neighbors are to be considered to assign label of a test sample. It uses more than one closest neighbor to determine the class in which the given test sample belongs to and consequently it is called as KNN. The training points are assigned weights according to their distances from sample data point. But at the same time the

ISSN (Online) 2456-1290
International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 2, Issue 3, March 2017

computational complexity and space complexity remain the primary concern dependably. To overcome memory limitation size of data set is reduced. For this the repeated patterns which don't include additional data are also eliminated from training data set. To further enhance the information focuses which don't influence the result are additionally eliminated from training data set. Using the algorithms we can expand the speed of basic KNN algorithm. Consider that an object is sampled with a set of different attributes. Assuming its group can be determined from its attributes; different algorithms can be used to automate the classification process. In pseudo code k-nearest neighbor classification algorithm can be expressed,

K ← number of nearest neighbors
For each object *X*in the test set do
calculate the distance D(*X*,*Y*) between *X* and every object *Y* in the training set
neighborhood ← the *k* neighbors in the training set closest to *X*
X.class ← SelectClass (neighborhood)
End

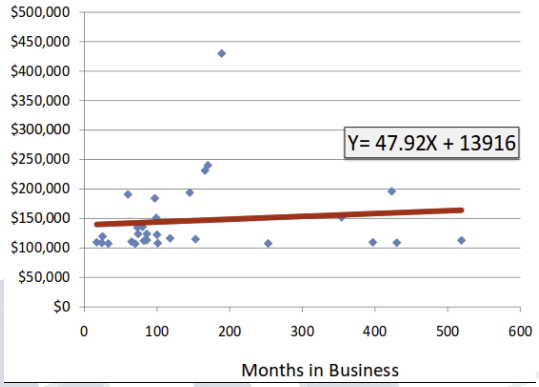


[2] Figure 1: k- nearest neighbour model

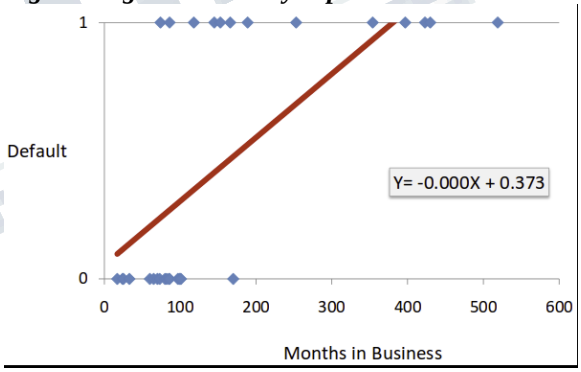
Logistic Regression

The regression is explained as an analyzing relation between a dependent variable and one or more independent variable. Regression can be defined by two types: Linear regression and logistic regression[3]. Logistic regression is a generalization of linear regression. It is basically used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modelled directly by linear regression i.e. discrete variable changed into continuous value. Logistic regression basically is used to classify the low dimensional data having nonlinear boundaries. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance. Logistic regression is also known as logistic model/ logit model that provide categorical variable for target variable with two categories such as light or dark, slim/ healthy, 0/1, Y/N, T/F. Following graph gives idea about logistic regression.

Logistic Regression-continuous dependent variable



Logistic Regression-binary dependent variable



Neural Network

The field of Neural Networks has arisen from diverse sources ranging from understanding and emulating the human brain to broader issues of copying human abilities such as speech and can be used in various fields such as banking, legal, medical, news, in classification program to categorize data as intrusive or normal. Generally neural networks consist of layers of interconnected nodes where each node producing a non-linear function of its input and input to a node may come from other nodes or directly from the input data. Also, some nodes are identified with the output of the network. There are wide variety of neural

ISSN (Online) 2456-1290
International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 2, Issue 3, March 2017

networks and their architectures. They range from simple Boolean networks to complex self-organizing networks. [5]The most important class of neural networks for real world problems solving includes Multilayer Perceptron, Radial Basis Function Networks, and Kohonen Self Organizing Feature Maps.

On the basis of this example there are different applications for neural networks that involve recognizing patterns and making simple decisions about them. In airplanes we can use a neural network as a basic autopilot where input units reads signals from the various cockpit instruments and output units modifying the plane's controls appropriately to keep it safely on course. Inside a factory we can use a neural network for quality control.

Characteristics of Neural Networks:

- Exhibit mapping capabilities, that is, they can map input patterns to their associated output patterns.
- Learn by examples. Thus, NN architectures can be 'trained' with known examples of a problem before they are tested for their 'inference' capability on unknown instances of the problem. They can, therefore, identify new objects previously untrained.
- Possess the capability to generalize. Thus, they can predict new outcomes from past trends.
- Robust systems and are fault tolerant. They can, therefore, recall full patterns from incomplete, partial or noisy patterns.

Artificial NN

Artificial neural networks (ANNs) are types of computer architecture inspired by biological neural networks (Nervous systems of the brain) and are used to approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are presented as systems of interconnected "neurons" which can compute values from inputs and are capable of machine learning as well as pattern recognition due their adaptive nature.

An artificial neural network operates by creating connections between many different processing elements each corresponding to a single neuron in a biological brain. These neurons may be actually constructed or simulated by a digital computer system. Each neuron takes many input signals then based on an internal weighting produces a single output signal that is sent as input to another neuron. The neurons are strongly interconnected and organized into different layers. The input layer receives the input and the output layer produces the final output. In general one or more hidden layers are sandwiched in between the two. This structure makes it impossible to forecast or know the exact flow of data.

Artificial neural networks typically start out with randomized weights for all their neurons. This means that initially they must be trained to solve the particular problem for which they are proposed. A back-propagation ANN is trained by humans to perform specific tasks. During the training period, we can evaluate whether the ANN's output is correct by observing pattern. If it's correct the neural weightings that produced that output are reinforced; if the output is incorrect, those weightings responsible can be diminished.

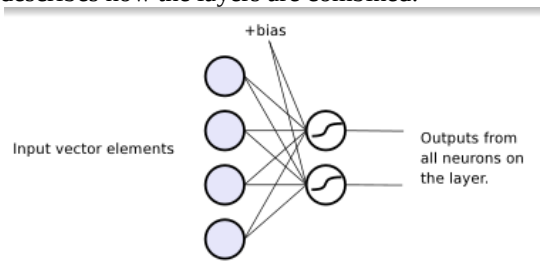
Implemented on a single computer, an artificial neural network is normally slower than more traditional solutions of algorithms. The ANN's parallel nature allows it to be built using multiple processors giving it a great speed advantage at very little development cost. The parallel architecture allows ANNs to process very large amounts of data very efficiently in less time. When dealing with large continuous streams of information such as speech recognition or machine sensor data ANNs can operate considerably faster as compare to other algorithms. An artificial neural network is useful in a variety of real-world applications such as visual pattern recognition and speech recognition that deal with complex often incomplete data. In addition, recent programs for text-to-speech have utilized ANNs. Many handwriting analysis programs (such as those used in popular PDAs) are currently using ANNs.

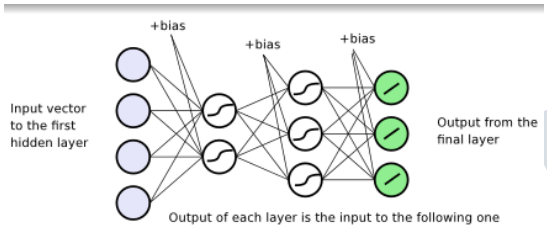
Multilayer Perceptron

The Multi-Layer Perceptron (MLP) or Feed-forward network is a type of artificial neural network that consists of a non-linear activation function in hidden layer [4]. MLP network provide nonlinear mapping between input and output vectors. Neural networks have two important functions i.e. pattern classifiers and nonlinear adaptive filters. A general framework of neural network consist of three layer architecture i.e. an input layer that define the input value, one or more hidden layers define the mathematical function and an output layer define final

ISSN (Online) 2456-1290
International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 2, Issue 3, March 2017

outcome . Each layer consists of a large number of neurons that are interconnected through weights. Each neuron has mathematical function (also known as activation function) that accepts input from previous layer and produced output for next layer. So, in neural networks the prediction is defined by the activation function. The following figures describes how the layers are combined.





General characteristics of MLP: A MLP

1. has any number of inputs.
2. has one or more hidden layers with any number of units.
3. uses linear combination functions in this layers.
4. uses generally sigmoid activation functions in the hidden layer.
5. has any number of outputs with any activation function.
6. has connections between the input layer and the first hidden layer, between the hidden layers, and between the last hidden layer and the output layer.

CONCLUSION

In this paper, we have observed supervised and unsupervised learning techniques. As per our observation all these traditional techniques are using some kind of input parameters to complete any kind task, which was given to those algorithms. After that, we have observed several Non-Parametric classification techniques which will give equivalent results when compared with parametric classification algorithms.

REFERENCES

[1] Ed Colet , "CLUSTERING AND CLASSIFICATION: DATA MINING APPROACHES".
[2] S.Neelamegam, Dr.E.Ramaraj , "Classification algorithm in Data mining: An overview" .
[3] Yugal Kumar , "Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA".
[4] Sarle, Warren S. (1994), "Neural Networks and Statistical Models," Proceedings of the Nineteenth Annual SAS Users Group International Conference, April, pp 1-13.
[5] Amrender Kumar , "ARTIFICIAL NEURAL NETWORKS FOR DATA MINING" .