International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 172 - 175, 2016, ISSN:- 2319–8656

# A Comparative Study of Various Data Mining Techniques: Statistics, Decision Trees and Neural Networks

Balar Khalid
Department of Computer Science
Hassan II University-FMPC
Casablanca, Morocco

Naji Abdelwahab
Department of Computer Science
Hassan II University-ENSET
Mohammedia, Morocco

**Abstract**: In this paper we focus on some techniques for solving data mining tasks such as: Statistics, Decision Trees and Neural Networks. The new approach has succeed in defining some new criteria for the evaluation process, and it has obtained valuable results based on what the technique is, the environment of using each techniques, the advantages and disadvantages of each technique, the consequences of choosing any of these techniques to extract hidden predictive information from large databases, and the methods of implementation of each technique. Finally, the paper has presented some valuable recommendations in this field.

**Keywords**: Data mining, Statistics, Logistic Regression, Decision Trees and Neural Networks.

## 1. INTRODUCTION

Extraction useful information from data is very far easier from collecting them. Therefore many sophisticated techniques, such as those developed in the multi- disciplinary field data mining are applied to the analysis of the datasets. One of the most difficult tasks in data mining is determining which of the multitude of available data mining technique is best suited to a given problem. Clearly, a more generalized approach to information extraction would improve the accuracy and cost effectiveness of using data mining techniques.

Therefore, this paper proposes a new direction based on evaluation techniques for solving data mining tasks, by using three techniques: Statistics, Decision Tree and Neural Networks.

The aim of this new approach is to study those techniques and their processes and to evaluate data mining techniques on the basis of: the suitability to a given problem, the advantages and disadvantages, and the consequences of choosing any technique, [5].

## 2. DATA MINING TOOLS

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses [6]. Data mining tools predict future trends and behaviors allowing businesses to make proactive knowledge driven decisions. Data mining tools can answer business question that traditionally were too time consuming to resolve.

They scour database for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

## 3. SELECTED DATA MINING TECHNIQUES

A large number of modeling techniques are labeled "data mining" techniques [7]. This section provides a short review of a selected number of these techniques. Our choice was guided the focus on the most currently used models. The review in this section only highlights some of the features of different techniques and how they influence, and benefit from. We do not present a complete exposition of the mathematical details of the algorithms, or their implementations.

Although various different techniques are used for different purposes those that are of interest in the present context [4]. Data mining techniques which are selected are Statistics, Decision Tree and Neural Networks.

### 3.1 Statistical Techniques

By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models.

Today people have to deal with up to terabytes of data and have to make sense of it and glean the important patterns from it. Statistics can help greatly in this process by helping to answer several important questions about their data: what patterns are there in database?, what is the chance that an event will occur?, which patterns are significant?, and what is a high level summary of the data that gives some idea of what is contained in database?

In statistics, prediction is usually synonymous with regression of some form. There are a variety of different types of regression in statistics but the basic idea is that a model is created that maps values from predictors in such a way that the lowest error occurs in making a prediction.

The simplest form of regression is S*imple Linear Regression* that just contains one predictor and a prediction. The relationship between the two can be mapped on a two dimensional space and the records plotted for the prediction values along the Y axis and the predictor values along the X axis. The simple linear regression model then could be viewed as the line that minimized the error rate between the actual prediction value and the point on the line [2].

Adding more predictors to the linear equation can produce more complicated lines that take more information into

account and hence make a better prediction, and it is called multiple linear regressions.

## 3.2 Decision Tree Techniques

The decision tree is a predictive model that, as its name implies, can be viewed as a decision tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes in an object. Induction decision tree can be used for exploration analysis, data preprocessing and prediction work.

The process in induction decision tree algorithms is very similar when they build trees. These algorithms look at all possible distinguishing questions that could possibly break up the original training dataset into segments that are nearly homogeneous with respect to the different classes being predicted. Some decision tree algorithms may use heuristics in order to pick the questions. As example, CART (Classification And Regression Trees) picks the questions in a much unsophisticated way as it tries them all. After it has tried them all, CART picks the best one, uses it to split the data into two more organized segment and then again ask all possible questions on each of these new segment individually [4].

## 3.3 Neural Network Technique

Artificial neural network derive their name from their historical development which started off with the premise that machines could be made to think if scientists found ways to mimic the structure and functioning of the human brain on the computer. There are two main structures of consequence in the neural network: The node - which loosely corresponds to the neuron in the human brain and the link - which loosely corresponds to the connections between neurons in the human brain [4].

Therefore, a neural network model is a collection of interconnected neurons. Such interconnections could form a single layer or multiple layers. Furthermore, the interconnections could be unidirectional or bi-directional. The arrangement of neurons and their interconnections is called the architecture of the network. Different neural network models correspond to different architectures. Different neural network architectures use different learning procedures for finding the strengths of interconnections.

Therefore, there are a large number of neural network models; each model has its own strengths and weaknesses as well as a class of problems for which it is most suitable.

# 4. EVALUATION OF DATA MINING TECHNIQUES

In this section, we can compare the selected techniques with the five criteria [5]: The identification of technique, the environment of using each technique, the advantages of each technique, the disadvantages of each technique, the consequences of choosing of each technique, and the implementation of each technique's process.

## 4.1 Statistical Technique

### *4.1.1 Identification of Statistics*

"Statistics is a branch of mathematics concerning the collection and the description of data" [2].

### *4.1.2 The Environment of Using Statistical Technique*

Today data mining has been defined independently of statistics though "mining data" for patterns and predictions is really what statistics is all about. Some of the techniques that are classified under data mining such as CHAID and CART really grew out of the statistical profession more than anywhere else, and the basic ideas of probability, independence and causality and over fitting are the foundation on which both data mining and statistics are built. The techniques are used in the same places for the same types of problems (prediction, classification discovery).

### *4.1.3 The Advantages of Statistical Technique*

Statistics can help greatly in data mining process by helping to answer several important questions about your data. The great values of statistics is in presenting a high level view of the database that provides some useful information without requiring every record to be understood in detail. As example, the histogram can quickly show important information about the database, which is the most frequent.

### *4.1.4 The Disadvantages of Statistical Technique*

Certainly statistics can do more than answer questions about the data but for most people today these are the questions that statistics cannot help answer. Consider that a large part of data the statistics is concerned with summarizing data, and more often than not, the problem that the summarization has to do with counting.

Statistical Techniques cannot be useful without certain assumptions about data.

### *4.1.5 The Consequences of choosing The Statistical Technique*

Statistics is used in the reporting of important information from which people may be able to make useful decisions. A trivial result that is obtained by an extremely simple method is called a naïve prediction, and an algorithm that claims to learn anything must always do better than the naïve prediction.

## 4.2 Decision Trees Technique

### *4.2.1 Identification of Decision Trees*

"A decision tree is a predictive model that, as its name implies, can be viewed as a tree " [2].

### *4.2.2 The Environment of using Decision Trees Technique*

Decision trees are used for both classification and estimation tasks. Decision trees can be used in order to predict the outcome for new samples. The decision tree technology can be used for exploration of the dataset and business problem. Another way that the decision tree technology has been used is for preprocessing data for other prediction algorithms.

### *4.2.3 The Advantages of Decision Trees Technique*

The Decision trees can naturally handle all types of variables, even with missing values. Co-linearity and linear-separability problems do not affect decision trees performance. The representation of the data in decision trees form gives the illusion of understanding the causes of the observed behavior of the dependent variable.

### *4.2.4 The Disadvantages of Decision Trees Technique*

Decision trees are not enjoying the large number of diagnostic tests. Decision trees do not impose special restrictions or requirements on the data preparation procedures. Decision trees cannot match the performance of that of linear regression.

#### *4.2.5 Consequences of choosing of Decision Trees Technique*

The decision trees help to explain how the model determined the estimated probability (in the case of classification) or the mean value (in the case of estimation problems) of the dependent variable. Decision trees are fairly robust with respect to a variety of predictor types and it can be run relatively quickly. Decision trees can be used on the first pass of a data mining run to create a subset of possibly useful predictors that can then be fed into neural networks, nearest neighbor and normal statistical routines.

### 4.3 Neural Networks Technique

#### *4.3.1 Identification of Neural Network*

"A neural network is given a set of inputs and is used to predict one or more outputs". [3]. "Neural networks are powerful mathematical models suitable for almost all data mining tasks, with special emphasis on classification and estimation problems" [9].

#### *4.3.2 The Environment of using Neural Networks Technique*

Neural network can be used for clustering, outlier analysis, feature extraction and prediction work. Neural Networks can be used in complex classification situations.

#### *4.3.3 The Advantages of Neural Networks Technique*

Neural Networks is capable of producing an arbitrarily complex relationship between inputs and outputs.
Neural Networks should be able to analyze and organize data using its intrinsic features without any external guidance. Neural Networks of various kinds can be used for clustering and prototype creation.

#### *4.3.4 The Disadvantages of Neural Networks Technique*

Neural networks do not work well when there are many hundreds or thousands of input features. Neural Networks do not yield acceptable performance for complex problems. It is difficult to understand the model that neural networks have built and how the raw data affects the output predictive answer.

#### *4.3.5 Consequences of choosing of Neural Networks Technique*

Neural Networks can be unleashed on your data straight out of the box without having to rearrange or modify the data very much to begin with. Neural Networks is that they are automated to a degree where the user does not need to know that much about how they work, or predictive modeling or even the database in order to use them.

## 5. CONCLUSION

In this paper we described the processes of selected techniques from the data mining point of view. It has been realized that all data mining techniques accomplish their goals perfectly, but each technique has its own characteristics and specifications that demonstrate their accuracy, proficiency and preference.
We claimed that new research solutions are needed for the problem of categorical data mining techniques, and presenting our ideas for future work.
Data mining has proven itself as a valuable tool in many areas, however, current data mining techniques are often far better suited to some problem areas than to others, therefore it is recommend to use data mining in most companies for at least to help managers to make correct decisions according to the information provided by data mining.
There is no one technique that can be completely effective for data mining in consideration to accuracy, prediction, classification, application, limitations, segmentation, summarization, dependency and detection. It is therefore recommended that these techniques should be used in cooperation with each other.

## 6. REFERENCES

[1] Adamo, J. M, Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms Springer-Verlag, New York, 2001.

[2] Berson, A, Smith, S, and Thearling, K. Building Data Mining Applications for CRM, 1st edition - McGraw-Hill Professiona, 1999.

[3] Bramer, M. Principles of Data Mining, Springer-Limited, 2007.

[4] Dwivedi, R. and Bajpai, R. Data Mining Techniques for dynamically Classifying and Analyzing Library Database Convention on Automation of Libraries in Education and Research Institutions, CALIBER, 2007.

[5] El-taher, M. Evaluation of Data Mining Techniques, M.Sc thesis (partial-fulfillment), University of Khartoum, Sudan, 2009.

[6] Han, J and Kamber, M. Data Mining , Concepts and Techniques, Morgan Kaufmann , Second Edition, 2006.

[7] Lee, S and Siau, K. A review of data mining techniques, Journal of Industrial Management & Data Systems, vol 101, no 1, 2001, pp.41-46.

[8] Perner, P. Data Mining on Multimedia - Springer- Limited , 2002.

[9] Refaat, M. Data Preparation for Data Mining Using SAS, Elsevier, 2007.

[10] Vityaev, E and Kovalerchuk, B. Inverse Visualization In Data Mining, in International Conference on Imaging Science, Systems, and Technology CISST'02, 2002.

**BALAR Khalid**[1], PhD in Computer Science, Hassan II University-Faculty of Medicine and Pharmacy of Casablanca, 19 Rue Tarik Ibnou Ziad, B.P. 9154, Casablanca, Morocco. Email: balarkhalid@gmail.com

**NAJI Abdelwahab**[2]**,** Assistant Prof in Computer Science, Hassan II University- Superior Normal School of Technical Education, Rue Boulevard Hassan II Mohammedia, Morocco. Email: abdelwahab.naji@gmail.com