Data Mining: Analysis and Comparative Study of Supervised Techniques

Balar Khalid
Department of Economic Science and Management
Hassan II University
Casablanca, Morocco

Chaabita Rachid
Department of Economic Science and Management
Hassan II University
Casablanca, Morocco

Boumhamdi Mounir
Department of Economic Science and Management
Hassan II University
Casablanca, Morocco

ABSTRACT
Data mining techniques are used more and more in the economic field. Such as the prediction of certain economic indicators, the discovery of hidden information, problems or finding problems in the industrial sector, as well as in relations with customers through the study of their data and behaviors to improve the cost-effectiveness of customer relationships or attract new customers.

Data Mining techniques are classified into two categories: supervised and unsupervised.

This paper focus only on the first techniques for solving Data Mining tasks such as: Decision Trees, Regression, Neural Networks and Support vector machines (SVM). 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, Regression, Decision Trees, Neural Networks, SVM, Supervised learning techniques

1. INTRODUCTION
The field of data mining is an emerging research area with important applications in Engineering, Science, Economic, Medicine, Business and Education.

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 four techniques: Decision Trees, Regression, Neural Networks and Support vector machines (SVM). 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, [1].

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 [2]. 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 SUPERVISED TECHNIQUES
A large number of modeling techniques are labeled "data mining" techniques [3]. 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. This study 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 supervised learning techniques which are selected are Statistics, Decision Trees, Regression, Neural Networks and Support vector machines (SVM).

In supervised or predictive modeling, the goal is to predict an event or to estimate the values of a continuous numerical attribute. In these models, there are fields where the input attributes and an output area or target. Input fields are also called predictors because they are used by the model to identify an output field prediction function. Can consider that predictors as the X part of the function and the target area as the Y part, the result.

Predictive models are subdivided into classification and estimation models:

• **The classification model:** In these models the groups or target classes are known from the beginning. The goal is to categorize cases in these predefined groups; in other words, to predict an event. The generated template can be used as a tagging engine for assigning new cases for predefined classes. He also estimates a propensity score for each case. The propensity score denotes the probability of occurrence of the target group or event.

• **Estimation models:** These models are similar to classification models, but with one major difference. They are used to predict the value of a continuous field based on the

observed values of the input attributes.

3.1 Decision Tree Technique

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].

Decision tree algorithms are based on speed and scalability. Algorithms available are:

- C5.0

- CHAID

- Classification and regression trees

- QUEST.

3.2 Regression

Regression is the method used to estimate continuous values. Its purpose is to find the best model that describes the relationship between a continuous output variable and one or more input variables. It is a matter of finding a function f which is as close as possible to a given scenario of inputs and outputs.

3.3 Neural Network Technique

The area of neural networks probably belongs to the border line between the artificial intelligence and approximation algorithm. A neural network is a collection of neurons like processing units with weighted connection between the units. It composes of many elements, called nodes which are connected in between. The connection between two nodes is weighted and by the adjustment of this weight, the training of the network is performed[5]. A classification model can be represented in different forms like neural network and decision tree which is shown in fig. 1.

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],[10]. 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.

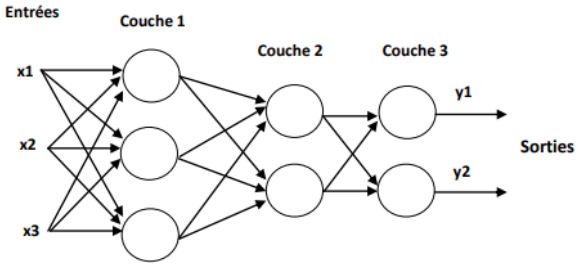

Fig 1: Artificial Neural Network

3.4 Support vector machines (SVM)

SVM is a classification algorithm that can model highly complex non-linear data profiles, and avoid over-learning, that is, the situation in which a model stores the models, only for specific cases analyzed. SVM works in map data at a large, characteristic space in which records become more easily separable from target categories.

The input drive data is appropriately transformed by the non-linear kernel functions and this transformation is followed by a search for simpler functions, that is, linear functions, which register optimally distinctly. Analysts usually experiment with different transformation functions and compare the results. Overall SVM is an efficient and demanding algorithm, in terms of memory resources and processing time. In addition, it lacks transparency since the forecasts are not explained and only the importance of the predictors is summarized.

4. EVALUATION OF DATA MINING TECHNIQUES

In this section compare the selected techniques with the five criteria [1]: 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 Decision Tree Technique

* Identification of Decision Tree "A decision tree is a predictive model that, as its name implies, can be viewed as a tree " [6].

* 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.

* 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.

* 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 Regression Technique

There are two main advantages to analyzing data using a multiple regression model. The first is the ability to determine the relative influence of one or more predictor variables to the criterion value. The real estate agent could find that the size of the homes and the number of bedrooms have a strong correlation to the price of a home, while the proximity to schools has no correlation at all, or even a negative correlation if it is primarily a retirement community.

The second advantage is the ability to identify outliers, or anomalies. For example, while reviewing the data related to management salaries, the human resources manager could find that the number of hours worked, the department size and its budget all had a strong correlation to salaries, while seniority did not. Alternatively, it could be that all of the listed predictor values were correlated to each of the salaries being examined, except for one manager who was being overpaid compared to the others.

Any disadvantage of using a multiple regression model usually comes down to the data being used. The example of this is using incomplete data and falsely concluding that a correlation is a causation.

In the example of management salaries, suppose there was one outlier who had a smaller budget, less seniority and with fewer personnel to manage but was making more than anyone else. The HR manager could look at the data and conclude that this individual is being overpaid. However, this conclusion would be erroneous if he didn't take into account that this manager was in charge of the company's website and had a highly coveted skillset in network security.

4.3 Neural Network Technique

Identification of Neural Network "A neural network is given a set of inputs and is used to predict one or more outputs". [7]. "Neural networks are powerful mathematical models suitable for almost all data mining tasks, with special emphasis on classification and estimation problems" [8], [9].

* 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.

* 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.

* 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.

* 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.

4.4 Support vector machines (SVM)

SVM doesn't give us the probability, it directly gives us the resultant classes.

Usual methods of validation like sensitivity, specificity, cross validation, ROC and AUC are the validation methods.

SVM Advantages

SVM's are very good when we have no idea on the data and Works well with even unstructured and semi structured data like text, Images and trees.

The kernel trick is real strength of SVM. With an appropriate kernel function, we can solve any complex problem.

Unlike in neural networks, SVM is not solved for local optima.

It scales relatively well to high dimensional data. SVM models have generalization in practice, the risk of over-fitting is less in SVM.

SVM is always compared with ANN. When compared to ANN models, SVMs give better results.

SVM Disadvantages

Choosing a "good" kernel function is not easy and Long training time for large datasets.

Difficult to understand and interpret the final model, variable weights and individual impact.

Since the final model is not so easy to see, we can not do small calibrations to the model hence its tough to incorporate our business logic.

The SVM hyper parameters are Cost -C and gamma. It is not that easy to fine-tune these hyper-parameters. It is hard to visualize their impact

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

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

[3] 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.

[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] Gajendra Sharma, "Data mining and Data Warehousing and OLAP", Published by S.K. Kataria & Sons, New Delhi, India.

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

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

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

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

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

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