A New Approach for Evaluation of Data Mining Techniques

Moawia Elfaki Yahia[1], Murtada El-mukashfi El-taher[2]

[1]College of Computer Science and IT
King Faisal University
Saudi Arabia, Alhasa 31982

[2]Faculty of Mathematical Sciences
University of Khartoum
Sudan, Khartoum 11115

Abstract

This paper tries to put a new direction for the evaluation of some techniques for solving data mining tasks such as: Statistics, Visualization, Clustering, Decision Trees, Association Rules 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 Evaluation, Statistics, Visualization, Clustering, Decision Trees, Association Rules, Neural Networks.*

1. Introduction

Extracting 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 six techniques: Statistics, Visualization, Clustering, Decision Tree, Association Rule 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, the consequences of choosing any technique, and the methods of implementation [5].

2. Data Mining Overview

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. Review of Selected 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, Visualization, Clustering, Decision Tree, Association Rules 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 Visualization Techniques

Visualization techniques are a very useful method of discovering patterns in data sets, and may be used at the beginning of a data mining process. There is a whole field of research dedicated to the search for interesting projections of datasets – this is called *Projection Pursuit*. For example, clusters are usually numerical represented. Also, a large set of rules is easier to understand when structured in a hierarchical fashion and graphically viewed such as in the form of a decision tree [8]. Many data mining method will discover meaningful patterns for "good" data but none of them or just few can produce meaningful patterns for "poor" data. One of the goals of Visualization is to transform "poor" data into "good" data permitting a wide variety of data mining methods to be used successfully to discover hidden patterns [10].

3.3 Clustering Techniques

Clustering techniques examine data to find groups of items that are similar. For example, an insurance company might group customers according to income, age, types of policy purchased or prior claims experience [3]. One of popular clustering algorithms is *Nearest Neighbor*, which is a prediction technique that is quite similar to clustering in order to predict what a prediction value is in one record look for records with similar predictor values in historical database and use the prediction value from the record that it is nearest to the unclassified record. The nearest neighbor prediction algorithm simply stated is: objects that are near to each other will have similar prediction values as well. Thus if you know the prediction value of one of the objects you can predict it for its nearest neighbors. One of the improvements of the basic nearest neighbor algorithm is to take a vote from the nearest neighbors rather than relying on the sole nearest neighbor to the unclassified record.

3.4 Induction Decision Tree Techniques

An induction 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].

3.5 Association Rule Techniques

An association rule tells us about the association between two or more items. For example, If we are given a set of items where items can be referred as books and a large collection of transactions (i.e., issue/return) which are subsets (baskets) of these items/books. The task is to find relationship between the presence of various items within these baskets [4]. In order for the rules to be useful there are two pieces of information that must be supplied as well as the actual rule: *Support* is how often does the rule apply? and *Confidence* is How often is the rule is correct.

In fact association rule mining is a two-step process: Find all frequent itemsets / booksets - by definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, and then generate strong association rules from the frequent itemsets - by definition, these rules must satisfy minimum support and minimum confidence.

3.6 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 six 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.

4.1.6 Implementation of Statistical Technique Process

There are many different parts of statistics of collecting data and counting it, the most notable ways of doing this are: histograms and linear regression.

4.2 Visualization Technique

4.2.1 Identification of Visualization

Visualization is to transform "poor" data into "good" data permitting a wide variety of data mining methods to be used successfully to discover hidden patterns [6].

4.2.2 The Environment of using Visualization Technique

It is used at the beginning of a data mining process. And it is used as the famous remark and picture of data mining process.

4.2.3 The Advantages of Visualization Technique

It is very useful for the method of discovering patterns in data sets. It holds for the exploration of large data sets. It transforms "poor" data into "good" data. It permits a wide variety of data mining methods to be used successfully to discover hidden patterns.

4.2.4 The Disadvantages of Visualization Technique

The visualization can be built for famous remark, but the numbers are not easy to be overlooked by humans. It is difficult to understand when structured in a hierarchical fashion and graphically viewed such as in the form of a decision tree.

4.2.5 Consequences of choosing of Visualization Technique

Numbers are not easy to be overlooked by humans, so that the summarization of these data into a proper graphical representation may give humans a better insight into the data. Interactive visualization techniques are also successfully applied for data exploration.

4.2.6 Implementation of Statistical Visualization process

Scatter diagrams can be used to implementation the visualization of the data sets, so that we can focus on the rest of the data mining process.

4.3 Clustering Technique

4.3.1 Identification of Clustering

“Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters ” [3].

4.3.2 The Environment of using Clustering Technique

Clustering techniques is used by the end user to tag the customers in their database. Once this is done the business user can get a quick high level view of what is happening within the cluster. Clustering can be used for discovery or prediction.

4.3.3 The Advantages of Clustering Technique

Clustering is done to give the end user a high level view of what is going on in the database. Clustering techniques are very fast to compute on the database. There is no direct notion of generalization to a test data set.

4.3.4 The Disadvantages of Clustering Technique

The interpretation of how interesting a clustering is will inevitably be application-dependent and subjective to some degree. Clustering techniques suffer from the fact that once a merge or a split is committed, it cannot be undone or refined. Sometimes clustering is performed not so much to keep records together as to make it easier to see when one record sticks out from the rest.

4.3.5 Consequences of choosing of Clustering Technique

To build groupings information of the data set, clustering is consolidating data into a high-level view and general grouping of records into like behaviors. Clustering is statistically separating records into smaller unique groups based on common similarities.

4.3.6 Implementation of Clustering Technique process

There are two main types of clustering techniques, those that create a hierarchy of clusters and those that do not. Those techniques are: *Hierarchical Clustering Techniques and Partitional Clustering Techniques*.

4.4 Decision Trees Technique

4.4.1 Identification of Decision Trees

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

4.4.2 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.

4.4.3 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.

4.4.4 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.

4.4.5 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.

4.4.6 Implementation of Decision Trees Technique process

There are two popular approaches for building decision tree technology [7] : *CART* stands for Classification and Regression Trees and *CHAID* stands Chi-Square Automatic Interaction Detector.

4.5 Association Rule Technique

4.5.1 Identification of Association Rule

“Association rules are a set of techniques, as the name implies, that aim to find association relationships in the data” [1].

4.5.2 The Environment of using Association Rule Technique

As example, the associations rule can be used to discover of the HTML documents to give insight in the user profile of the website visitors and retrieved in connection with other HTML documents.

4.5.3 The Advantages of Association Rule Technique

Association rule algorithms can be formulated to look for sequential patterns. The methods of data acquisition and integration, and integrity checks are the most relevant to association rules.

4.5.4 The Disadvantages of Association Rule Technique

Association rules do not show reasonable patterns with dependent variable and cannot reduce the number of independent variables by removing. Association rules cannot be useful if the information do not provide support and confidence of rule are correct.

4.5.5 Consequences of choosing of Association Rule Technique

Association rule algorithms can be formulated to look for sequential patterns. The data usually needed for association analysis algorithms is the transaction.

4.4.6 Implementation of Association Rule Technique process

Association rules can be implemented in a two-step process: find all frequent item sets: by definition, each of these item sets will represent the transaction, and then generate strong association rules from the frequent item sets: by definition, these rules must satisfy minimum support and minimum confidence.

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.

4.3.6 Implementation of Neural Networks Technique process

A typical neural network application requires consideration of the following issues: model selection; input-output encoding; and learning rate. The choice of the model depends upon the nature of the predictive problem, its expected complexity, and the nature of the training examples. Generally there are three popular models to implement neural networks applications. These are: Single Layer Perception Network, Multiple Layer Feed Forward Network and Self Organizing Feature Map.

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.

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.

Moawia Elfaki Yahia received his B.Sc and M.Sc in computer science from University of Khartoum in 1989, 1995, and his PhD degree in Artificial Intelligence from Putra University, Malaysia in 1999. Currently, he is an Associate Professor of computer science at King Faisal University, Saudi Arabia. Dr. Yahia is member of editorial board and reviewer of many journals in computing science. He is a Member of ACM, IEEE CS, Arabic Computer Society, and International Rough Sets Society. Dr. Yahia received TWAS award for Young Scientists in Developing Countries in 2006. He published over 25 articles in international journals and proceedings. His research interests are: intelligent hybrid systems, data mining, Arabic text miming, rough set theory, and information retrieval.

Murtada El-mukashfi El-taher obtained his B.Sc and M.Sc in Computer Science from Faculty of Mathematical Science, University of Khartoum in 2006 and 2009. He accumulated several experiences in teaching and research in computer science during the last five years. His area of interest includes data mining applications and IT applications to business and e-government.