A Mechanism for Selecting Appropriate Data Mining Techniques

Rose Tinabo
School of Computing
Dublin institute of technology
Kevin Street
Dublin 8
Ireland
rose.tinabo@mydit.ie

ABSTRACT: *Due to an increase of several data mining techniques, selection of an appropriate data mining technique to use is one of the important steps when undertaking a data mining project. Data miners used to test various numbers of techniques to work out which one will give them the best results every time they have new data, this is costly and time consuming. They can solve this problem by understanding the main objective of mining, strength and weaknesses of the techniques and the features of the data they have, to determine the best technique to use without testing numbers of techniques. Testing different techniques shows that, there is a confusion of what data mining technique will be most appropriate for them than other techniques. Therefore, this paper discus different strength and weaknesses of four different data mining techniques and compare their performance on different features of the datasets by four different evaluation metrics. The summary was theoretically applied to the retail dataset to decide on the most appropriate technique for customer retention as a case study. Thus, users can use the summary provided in this paper to select the most appropriate data mining technique.*

Keywords: Data Mining, Decision Trees, Neural Networks, k-nearest Neighbor, Naïve Bayes

Received: 12 December 2010, Revised 27 January 2011, Accepted 27 Janurary 2011

1. Introduction

The rapid increase of different ways of generating and collecting data, such as common use of bar codes for most commercial products, the computerization of different business and government transactions, and the advances in data collection tools has resulted in the availability of huge amounts of data. This increase in amount of data has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge [1-4]. As a result, data mining has become a research area of increasing importance.

Generally, data mining is the process for discovering significant, nontrivial, previously unknown, and potentially useful knowledge from large amount of data. The main goal of data mining methodologies is to extract hidden patterns from these complex information sources and predict future trends and behaviours. Data mining appears in other articles and documents as knowledge mining from databases, knowledge extraction, data archaeology, data dredging or data analysis [5, 6].

There are number of established data mining techniques, ranging from classical statistical methods, such as linear and logistic regression, to neural network and tree-based techniques, to the most recent support vector machine and random forest. Data mining techniques can be categorised as supervised or unsupervised. Supervised techniques means the training data (observations, measurements) are accompanied by labels indicating the class of the observations. New data is classified based on the training set. While unsupervised techniques means that the class labels of training data is unknown, therefore, given a set of training data with the aim of establishing the existence of classes or clusters in the data [7, 8].

Different data mining techniques need to be employed for different categories of problem. It is therefore important that the users of data mining be able to make decision about which data mining technique to employ. In order to do this, they need to know which data mining technique is the most appropriate than others and in which type of data [9]. Authors propose selection of an appropriate data mining techniques based on understanding of the strength and weaknesses of these techniques, features of the data they have and the objective of mining. Understanding strength and weaknesses of different techniques and type of the mined data is important in data mining, otherwise it would be necessary to test number of techniques every time to find which one will provide the best results with new data, which is both difficult and time consuming.

This paper discusses strength and weaknesses of four leading supervised data mining techniques. The performance of the four techniques were compared to a number of datasets using the features of the data such as number of attributes, number of instances, size of the dataset, data type and presence of missing values as the basis of the comparison. The summary was theoretically applied to the retail dataset to decide on the most appropriate technique for customer retention as a case study. The results presented in this paper will provide data miners with valuable knowledge when faced with the challenge of selecting a data mining technique. This paper is also useful to researchers and students who want to gain an understanding on data mining, data mining techniques and different performance measures for comparing techniques.

2. Data Mining

Data mining is a fast-growing field of research [10]. Its popularity is caused by an ever increasing demand for tools that help in revealing and understand information hidden in huge amount of data that now exist in every type of organisation [11]. Such data are generated on a daily basis in different organisation such as by banks, insurance companies, retail stores, federal agencies and on World Wide Web (www). This explosion came about through the increasing use of computers, scanners, digital cameras, bar code, etc. We are in situation with rich sources of data, stored in databases, warehouses, and other data repositories, are readily available but not easily analysable [2, 5, 12]. This causes pressure from business, industry communities and governments for improvements in the data mining technology. What is needed is a clear and simple methodology for extracting the knowledge hidden in the data [3] .

Even though defined differently by different researchers by using different words, there is agreement that data mining can be defined as the process of searching and analysing large quantities of data in order to discover meaningful information [5, 12] and that its main goal is to extract hidden patterns from complex information sources and predict future trends and behaviours. Prediction and description are the main goals of data mining [13]. In prediction-oriented data mining, the discovered patterns are used for predicting future value of some variables. Description-oriented data mining concentrate on identifying patterns for the purpose of being presented in understandable way to users. To achieve these goals can be expanded to different functions such as classification, clustering, association, visualization and some basic statistical analyse, such as regression, estimation and hypothesis testing. Different kinds of data mining methods and algorithms have been proposed, each of which has its own advantages and suitable application domains [14].

Classification is an important and the most common problem for which data mining is used. It is the process of finding a set of general features and models that describe and distinguish data classes or concepts [15]. These models are used to predict the class of attribute whose class label is unknown. Therefore, classification deals with examining the features of new attribute and assigning to it a predefined class label [2]. Classifying a credit customer as low, medium, or high risk and assigning customers to predefined customer segments are examples of classification tasks.

3. An Overview of Data-Mining Techniques

To help achieve goals of data mining, a variety of techniques are available. A data mining technique is the function that assigns an unlabelled instance to a label using internal data structure [16]. Each data mining technique is used for different purposes, and offers its own advantages and disadvantages. Categories of the most used data mining techniques are; Decision Trees, Artificial Neural Networks, Rule Induction, Case-Based Reasoning, Bayesian Belief Networks, and Genetic Algorithms and Evolutionary programming [17].

For the purpose of this paper; four of the most widely used data mining techniques are discussed [18]. The criteria used to guide the selection of the techniques to investigate in this paper based on the different type of explanation (rules or functions) provided by the techniques and the fact that they represent different categories. Decision tree represents techniques whose explanations are represented by tree format, nearest neighbors represent lazy techniques from Case-based reasoning, for Bayes there is a Naïve Bayes and for techniques representing explanation by functions are represented by Neural Network from Artificial Neural Networks category.

3.1 Decision Tree

A decision tree is a repeated and special form of tree structure for expressing classification rules. The main strength of using decision tree is clarity and conciseness. The results are presented symbolically in a tree form which is simple and easy to understand. That is, decision tree algorithms have the capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution that is often easier to interpret, as they are straight forward if-then-else rules which most business analysts can follow and interpret (unlike the black box issue of neural networks). Other strength of decision trees is that the learning time is much faster than training a neural network [19].

However, decision tree has several drawbacks such as it tends to over fit the data. In addition, since one big tree is grown, it is hard to account for additive effects. Decision trees are usually univariate since they are using single attribute to split at each internal node. That is, each test is limited to a straightforward division based on a single attribute and most of the decision tree algorithms cannot perform well with problems that require crosswise partitioning [18]. The simplicity of decision trees cause the cost of not scaling well to larger data sets from a number of angles. First, decision trees cannot be modified for incremental learning (i.e., updating a model with additional data cannot be able without restarting the learning process from scratch). Secondly, performance does not scale linearly or sub-linearly with size. Thirdly, the results are hard to visualize beyond very simple examples and that data ordering affects its performance. Also decision trees cannot uncover rules based on combinations of variables.

In summary, decision tree tends to perform better when dealing with discrete/categorical features, and the strength of the decision tree is due to its ability to create understandable explanation of the results. In the absence of the noisy data, decision trees are accurate, deals extremely well with raw data, they are scalable to any data size, and they work fast and can cope with multi-dimensional data.

3.2 Neural network

Neural networks are networks of simple computational units operating in a parallel and highly interconnected way [20]. Neural networks are also popular for predictive capability and ability to learn patterns from real-life data that is noisy, imprecise and incomplete [14]. However, neural networks have some significant limitations in terms of training time, clarity, and dimensionality.

Mostly neural networks are very useful in predicting the future rather than explaining the past as there are no enough explanations as to why a particular conclusion is reached [21, 22]. When used for prediction, neural networks provide highly accurate results in a wide variety of different problem domains.

The main weakness of neural networks is that they are extremely slow, not only in the training but also in application phase. Also, one of the problems of neural network is the determination of the size of the hidden layer, because underestimate of the number of neurons can leads to poor approximation and generalization capabilities, while increasing nodes can result in over fitting that will result to search for the global optimum more difficult. Also, with neural networks it is difficult to determine how the network is making its decision due to lack of explanations, therefore, it is hard to determine which of the input features being used are important and useful for classification and which are worthless [21, 22].

In summary, neural networks are better in predicting the feature than describing the past as they have no enough explanation ability (i.e., it is a black box approach). They are capable of dealing with large amounts of noisy data and robust in dealing with missing data but they are mostly suitable for expert users. They need a lot of pre-processing that cause them to be slow compared to decision trees. But, once a neural network has been trained, it has good predictive capabilities. Neural network has been the most commonly used data mining technique in financial applications.

3.3 Nearest neighbour

A case-based reasoning (CBR) is a simple classification method that tries to solve a given problem by making direct use of past experiences and solutions [23]. The *k*-nearest neighbour (*k*-NN) technique is the simplest form of case based reasoning.

The main advantage of nearest neighbour is that they are easy to understand. They can provide good result if the features are chosen carefully and weighted carefully in the computation of the distance. Nearest neighbor does well with good features, but tends to degrade with many poor features. That is despite its simplicity, the nearest neighbor technique has many advantages over other methods. For example, it can learn from a small set of examples, can incrementally add new information at runtime, and can give competitive performance with the more modern methods such as decision tree or neural networks [24].

Weakness of case-based reasoning includes; they are like neural network, do not simplify the distribution of objects in parameter space to an understandable set of parameters. Instead, the training set is retained in its entirety as a description of the object distribution. This method is also rather slow if the training set has many examples [25]. The most serious shortcoming of case-based reasoning methods is that they are very sensitive to the presence of unrelated parameters. Adding a single parameter that has a random value for all objects (so that it does not separate the classes) can cause these methods to fail miserably.

In summary Nearest Neighbour Classification is one of the easiest classification techniques to understand which is easy to implement as no training involved, therefore is a lazy learning; new data can be added during runtime; it has some explanation capabilities and is robust to noisy data by averaging *k*-nearest neighbours. But k-nearest neighbor is not the most powerful classification and it is a slow classification and is affected by dimensionality.

3.4 Naïve Bayes

Naïve Bayes is one of the simplest and effective techniques for classification problem and is based on the so called Bayesian theorem. Naïve Bayes has been used as an effective classifier for many years due to its simple structure that have resulted in to several advantages. In particular, the construction of naive Bayes is very simple therefore can be trained in a short computational time. Furthermore, the inference (classification) is achieved in a linear time. Finally, naïve Bayes can be easily updated as it is always easy to consider and take into consideration new cases in hand; this results into incremental construction of naïve Bayes [18].

The main problem of Naïve Bayes is its strongly unrealistic assumption that all attributes are independent of each other given the background of the class and all other variables are directly dependent on the classification variable assumes that attributes are independent given the class which is almost always wrong in most of the real-world tasks

4. Experiment

In order to perform the experimentation, four data mining analysis tools were evaluated in selecting appropriate data analysis tool. These tools include: SAS Enterprise Miner [26], and Oracle Data Miner [27] ,few are available as open source such as WEKA [28], RapidMiner; [29]. Four main categories of criteria for evaluating data mining tools: performance, functionality, usability, and support of additional activities were used [30]. After evaluating the four data mining analysis software tools, tutorials and guidelines for the usage of such suites, RapidMiner was selected as it is the open-source, simple and easy to understand.

Therefore, a series of experiments were conducted by using RapidMiner data mining analysis tool. These experiments were designed to study the performance of the selected data mining techniques performing the classification data mining tasks using different size of the datasets and four evaluation metrics (predictive accuracy, prediction, recall and Area under the Receiver Operating Characteristics (ROC) curve or simply Area under the curve (AUC)) were used, as it is possible for a technique that performs well on one metric may not perform well on other metrics [31]. Experiments were conducted by varying the inputs, collecting the output, and analysing the results that will support in determining performance of the data mining techniques under the study. The flow of the experiment is summarized by Figure 1.

5. Results of the Experiment

A summary of the performance of the four techniques in different features are presented in Table 1. This summary based on the evidence of the existing empirical and theoretical studies and the experiments done. Performances range from low, medium, high to very high categories. Data mining technique with very high category of the mentioned feature means completely handles the problem; technique with high category handles more than 70% of the problem, technique with medium handles less than 70% but greater than 40% while technique with low handles less than 40% of the problem.

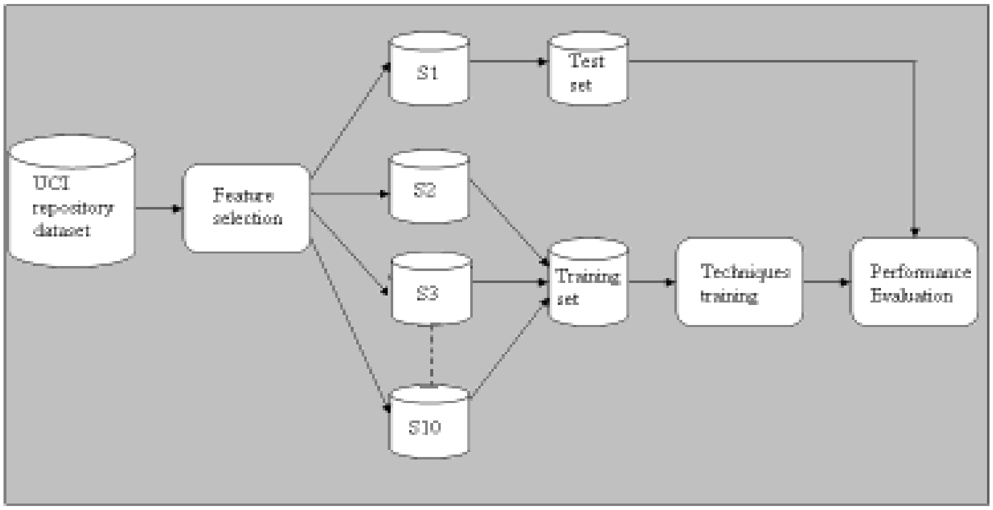
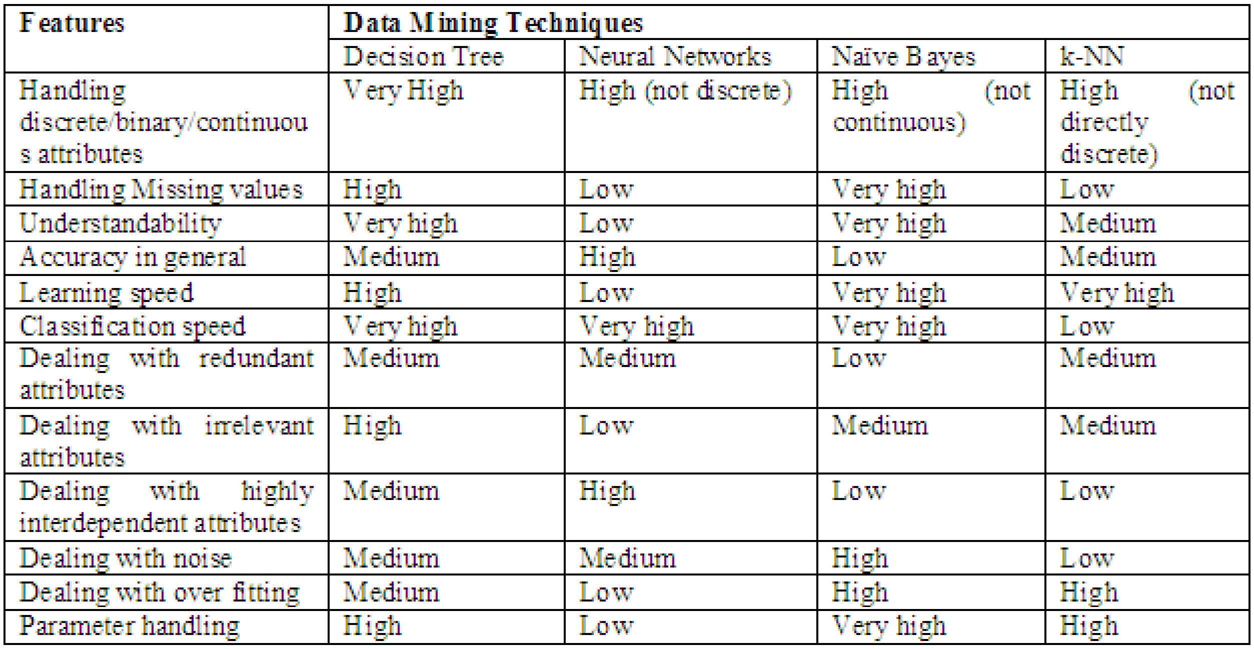


Figure 1. The experiment workflow example for the first iteration by 10-fold cross-validation (amended from [12])



Features	Data Mining Techniques			
	Decision Tree	Neural Networks	Naïve Bayes	k-NN
Handling discrete/binary/continuous attributes	Very High	High (not discrete)	High (not continuous)	High (not directly discrete)
Handling Missing values	High	Low	Very high	Low
Understandability	Very high	Low	Very high	Medium
Accuracy in general	Medium	High	Low	Medium
Learning speed	High	Low	Very high	Very high
Classification speed	Very high	Very high	Very high	Low
Dealing with redundant attributes	Medium	Medium	Low	Medium
Dealing with irrelevant attributes	High	Low	Medium	Medium
Dealing with highly interdependent attributes	Medium	High	Low	Low
Dealing with noise	Medium	Medium	High	Low
Dealing with over fitting	Medium	Low	High	High
Parameter handling	High	Low	Very high	High

Table 1. Performance of data mining techniques in different features

6. Case Study: Customer Retention in Retail Sector

As a result of the intensive literature review and the experiment done in this research, decision tree is the proposed data mining technique for customer retention in retail sector not only because of its better performance but also due to different characteristic it posses and objective and features of the retail data.

In the retail sector, organisations often wish to explore data and explain the results, and hence they are interested in applying a quick technique to provide an insight to understand the relations between complex data. Therefore, this concludes that when comparing the performance of the techniques in retail sector, the technique with better explanatory power is more preferred. Decision tree is very good in providing explanation than other techniques which are not very good in explanation and also there slow.

A typical retail datasets has several thousands of features that show a high degree of redundancy. In order to reduce dimensionality the data must be pre-processed, and peaks, which roughly correspond to individual information. This can be well done by decision tree as decision trees are not sensitive to differences of scale between the inputs, or to outliers. This means that data preparation is less of a trouble with decision trees than it is with other techniques such as neural networks

The use of decision tree can make organisation to quickly gain a better understanding of the variables that influence customer churn and allowing them to determine not only which customers are likely to leave, but why are they leaving, as they do provide description of the variables.

Decision tree is one of the mostly used techniques in data mining for searching prediction information. Due to its characteristics which are suitable for parallelism, it has been widely adopted in high performance field and developed into various parallel decision tree algorithms to deal with huge data and complex computation.

Unlike decision tree based techniques, other classification techniques such as neural networks, nearest neighbours and naïve Bayes can determine a probability for a prediction with its likelihood. However, comparing with decision tree based algorithms; these techniques do not explicitly express the uncovered patterns in a symbolic, easily understandable form (e.g., if-then rules).

But all in all, performance of the data mining techniques depends on the problems they are required to solve and features of the available data. Thus, there is no one data mining technique which is the best for all problems. For example, if the problem is only to predict what will happen to the future and not description of how the technique did that, neural network could be best techniques than decision tree as neural network is good in prediction than description.

7. Conclusion

Different data mining techniques need to be employed in different categories of the problem. It is therefore important for users of data mining to be able to make decision about which data mining technique to employ. Therefore users need to know which data mining technique is most appropriate than others and in which type of data. A mechanism for selecting data mining technique is one of the important steps in data mining, otherwise it would be necessary to test number of techniques every time to find which one will provide the best results with new data, which is both difficult and time consuming. After comprehensive literature review and experimentation, this paper propose selection of an appropriate data mining techniques based on understanding the strength and weaknesses of these techniques, features of the data they have and the objective of data mining. Therefore, this paper discus different strength and weaknesses of four different data mining techniques and compare their performance on different features of the datasets by four different evaluation metrics. The summary was theoretically applied to the retail dataset to decide on the most appropriate technique for customer retention as a case study. Thus, users can use the summary provided in this paper to select appropriate data mining technique to use depending on the features of the data and different strengths and weaknesses of the technique.

References

[1] Ahmed, S.R.2005, *Applications of data mining in retail business*, in *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*. . p. 455-459.

[2] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth,1996. *From data mining to knowledge discovery in databases*. AI magazine, 17(3): p. 37.

[3] Pal, S.K. and P. Mitra, 2004: *Pattern Recognition Algorithms for Data Mining*. Chapman & Hall/CRC.

[4] Quinlan, J.R., 1993*C4. 5: programs for machine learning*.: Morgan Kaufmann.

[5] Berry, M.J. and G. Linoff, 1997:*Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc. New York, NY, USA.

[6] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, 1996.*The KDD process for extracting useful knowledge from volumes of data*. Communications of the ACM, 39(11): p. 27-34.

[7] Cristianini, N. and J. Shawe-Taylor, 2004).*Kernel methods for pattern analysis*. : Cambridge University Press.

[8] Perlich, C., F. Provost, and J.S. Simonoff, 2003*Tree induction vs. logistic regression: a learning-curve analysis*. The Journal of Machine Learning Research, . **4**: p. 211-255.

[9] Kerdprasop, N. and K. Kerdprasop, 2003.*Data partitioning for incremental data mining*, in *The 1st International Forum on Information and Computer Science*. p. 114-118.

[10] Deogun, J.S., et al.,1996 *Data mining: trends in research and development*, in *in Rough Sets and Data Mining: Analysis for Imprecise Data*. c

[11] Fayyad, U.M., G. Piatetsky-Shapiro, and P. Smyth, 1996*From data mining to knowledge discovery: An overview*, in *Advances in knowledge discovery and data mining*. . p. 1-34.

[12] Han, J. and M. Kamber, 2006:Data mining: concepts and techniques. Morgan Kaufmann.

[13] Fayyad, U., D. Haussler, and P. Stolorz, 1996 Mining scientific data. Communications of the ACM,. 39(11): p. 57.

[14] Lee, J.H., S.J. Yu, and S.C. Park, 2001Design of intelligent data sampling methodology based on data mining. IEEE Transactions on Robotics and Automation, . 17(5): p. 637.

[15] Zhang, D. and L. Zhou, 2004Discovering golden nuggets: data mining in financial application. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, . 34(4): p. 513-522.

[16] Kohavi, R.,1995 A study of cross-validation and bootstrap for accuracy estimation and model selection, in International joint Conference on artificial intelligence. . p. 1137-1145.

[17] Braha, D., 2001. Data mining for design and manufacturing: methods and applications.

[18] Kotsiantis, S.,2007: Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies, p. 3.

[19] Kleissner, C.,2002. Data mining for the enterprise, in System Sciences, 1998., Proceedings of the Thirty-First Hawaii International Conference on. p. 295-304.

[20] Lippmann, R.P.,1987 An introduction to computing with neural nets. ARIEL, . 209: p. 115-245.

[21] Vellido, A., P.J.G. Lisboa, and J. Vaughan, 1999. Neural networks in business: a survey of applications (1992-1998). Expert Systems with Applications,17(1): p. 51.

[22] Zhang, G.P.,2002. Neural networks for classification: a survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 30(4): p. 451-462.

[23] Maher, M.L., G. de Silva Garza, and others,2002 Case-based reasoning in design. IEEE Expert, . 12(2): p. 34-41.

[24] Bay, S.D., 1999Nearest neighbor classification from multiple feature subsets. Intelligent Data Analysis, . 3(3): p. 191-209.

[25] Bao, Y., X. Du, and N. Ishii, 2002. Improving performance of the k-nearest neighbor classifier by tolerant rough sets, in Cooperative Database Systems for Advanced Applications, 2001. CODAS 2001. The Proceedings of the Third International Symposium on. p. 167-171.

[26] Cerrito, P.B.,2006: Introduction to data mining using SAS Enterprise Miner. 2006: SAS Publishing.

[27] Milenova, B.L., J.S. Yarmus, and M.M. Campos, 2005. SVM in oracle database 10g: removing the barriers to widespread adoption of support vector machines, in Proceedings of the 31st international conference on Very large data bases. p. 1152-1163.

[28] Holmes, G., A. Donkin, and I.H. Witten, 2002Weka: A machine learning workbench, in Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on. . p. 357-361.

[29]. Alcal'a-Fdez, J., et al.,2009. KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 13(3): p. 307-318.

[30] Abbott, D.W., I.P. Matkovsky, and J.F. Elder 1998 An evaluation of high-end data mining tools for fraud detection, in Systems, Man, and Cybernetics, 1998. IEEE International Conference on. 2002. p. 2836-2841.

[31] Huang, J. and C.X. Ling, 2005. Using AUC and accuracy in evaluating learning algorithms. Knowledge and Data Engineering, IEEE Transactions on, 17(3): p. 299-310.