Machine Learning Approach for Taxation Analysis using Classification Techniques

R.Deepa Lakshmi
MPhil Research Scholar
PSGR Krishnammal College for women
India

N.Radha
GR Govindarajulu School of Computer Technology
PSGR Krishnammal College for women
India

ABSTRACT
Data mining process discovers useful information from the hidden data, which can be used for future prediction. Machine learning provides methods, techniques and tools, which help to learn automatically and to make accurate predictions based on past observations. The data are retrieved from the real time environmental setup. Machine learning techniques can help in the integration of computer-based systems in predicting the dataset and to improve the efficiency of the system. The main purpose of this paper is to provide a comparison of some commonly employed classification algorithms under same conditions. Such comparison helps to provide the accurate result in algorithms. Hence comparing the algorithms for such a classifier is a tedious task, for real time dataset. The classification models were experimented by using 365 datasets with 24 attributes. The predicted values for the classifiers were evaluated and the results were compared.

General Terms: Algorithms, Experimentation, Performance.

Keywords: Machine-learning Techniques, Audit Selection Strategy, Data Mining, open source tools, Naive bayes, Tax audit, WEKA Classification.

1. INTRODUCTION
Data mining is the nontrivial extraction of implicit previously unknown and potentially useful information from data and science of extracting useful information from large datasets [14]. The main bottleneck of data mining software programming is the data management. Weka, Tanagra, Sipina, is free data mining software. Data mining process involves multiple stages. A simple, but typical process might include preprocessing data, then applying data-mining algorithms, finally processing the mining results. There are many achievements of applying data mining techniques to various areas such as marketing, medical and financial. The amount of data being collected in databases today far exceeds our ability to reduce and analyze data without the use of automated analysis techniques. Audit Division enables the agency to identify non-compliant taxpayers more efficiently and effectively, and to focus auditing resources on the accounts most likely to produce positive tax adjustments [4].

Data mining helps the agency refine its traditional audit selection strategies to produce more accurate results. This paper analyzes the performance of algorithms using different classification methods to make audit process more efficient and effective. Data mining helps the agency refine its traditional audit selection strategies to produce more accurate results. The development of data-mining applications such as classification has shown the need for supervised learning algorithms to be applied to large-scale data [6].

In classification method, many attributes are involved. The reasons for selecting a subset of attributes instead of the whole dataset are (1) It is easier to measure only a reduced set of data of selected dataset, (2) Prediction accuracy may be improved through exclusion of redundant and irrelevant attribute, (3) The predictor to be built is usually simpler and potentially faster when fewer input data are used and (4) Knowing which attributes are relevant can give accurate result of the prediction problem and allows a better understanding of the final classification.

Machine learning provides methods techniques and tools, which help to learn automatically and to make accurate predictions based on past observations [10]. Machine learning is popularly being used in areas of business like data analysis, financial analysis; stock market forecast etc[3]. Classification is used to build classification tree for predicting continuous dependent variables and categorical predictor variables.

For classifying the dataset, Estimate the accuracy of the model using a test dataset. Test dataset is independent of training dataset, otherwise over-fitting will occur. The known label of the test dataset sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model.

2. DATASET DESCRIPTION
Data mining has many existing and potential applications in tax administration. The data under analysis represent the income and taxation particulars of 365 clients of M/s. MSS and Co., chartered Accountants, Tirupur. The income data pertaining to the financial year ending 31st march, 2006 (Assessment year 2006-07) is segregated under various Heads and Gross total Income, Deductions and Net Taxable Income along with Income tax payable thereon and interest, if any, are also listed. Under Income tax Act, 1961, Persons or entities, known as Assessee, earning Income are divided in to various categories as listed below:

1. Individual;
2. Hindu Undivided Family;
3. Unregistered firms;

4. Registered Partnership firms including Professional firms;
5. Limited Companies, both Public and Private;
6. Cooperative Societies;
7. Trusts and Association of Persons;
8. Body of Individuals;
9. Artificial Juridical Persons;
10. Foreign company; and
11. Local authority.

The data analyzed here pertain mostly to the first five categories mentioned above. Similarly, the Income earned by an assessee is segregated for tax purposes under the following Heads:

1. Income from Salary;
2. Income from House Property;
3. Income from Own Business;
4. Share Income from Registered firms;
5. Share Income from Un-Registered firms;
6. Capital Gains;
7. Income from Other Sources; and
8. Agricultural Income.

Of the above, income earned under the category 'Agricultural Income' is not taxed under Income tax Act; but is included in Gross total Income only and appropriate deduction is given in the tax computation for such inclusion. Income earned under the other heads is taxed as per prevailing law [7]. Salary Income is restricted only for Individuals. It is not necessary that a person earning income under one head should earn under other heads also.

Similarly, for justified reasons, income earned under a particular head in a financial year may not be earned in the subsequent financial year. After making many adjustments under individual heads, Gross Total Income is computed. Some more deductions are also allowed collectively from the Gross total Income and Net Taxable Income is arrived at. Individuals are categorized as Senior Citizens and Others for the purpose of basic exemption. Assessees aged 65 and above are specified as Senior citizens. Non-Senior citizens are further classified as male and female and different exemption limits are adopted. Different rates of Income tax have been prescribed for a different slab of Income and for different status and class of assesses [9]. After arriving at the net taxable income, income tax is computed as per the applicable rate prevailing in force. The assessee is required to pay the income tax as and when the income is earned by way of advance tax.

In some cases, tax is also deducted or collected at the source of income from which they are earned or received. Interest is levied for any short remittance as per the prescribed procedure [12]. Any difference between the total tax and interest due is allowed to remit at the time of filing the Return of income. Should there still be any balance; the same can be paid after assessment, inclusive of interest. Any excess tax paid is refunded after the assessment is completed.

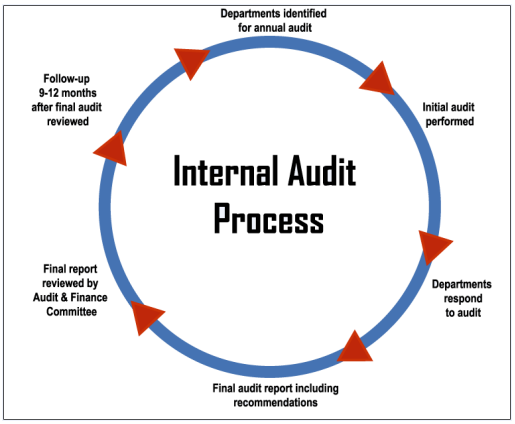

Figure 1. Audit process

The collected data also comprises of income tax due for different asessees and the actual tax paid by them and balance tax due by them or to be refunded to them. Every Office of the Chartered Accountants takes a detailed analysis of these dates every year to enable corrective action being taken in the subsequent years [14]. The collection of data presented here gives rise to varied types of analysis, depending on the requirement.

3. CLASSIFICATION ALGORITHMS AND THEORITICAL BASIS

For analyzing real time dataset and to predict the performance, the supervised learning algorithms were adopted here. The main motivation for different classification algorithms is accuracy improvement. There are two main paradigms for handling different classification algorithms. First is Classifier Selection and second is Classifier Fusion. The first one selects a single algorithm for classifying new instances, while the latter fuses the decisions of all algorithms.

3.1 Classifier Selection

It is a very simple method, which produces Selection or Select Best. This method evaluates each of the classification algorithms on the training set and selects the best one for application on the test set. Although this method is simple, it has been found to be highly effective and comparable to other more complex state-of-the-art methods. Another line of research proposes the selection of a learning algorithm based on its performance on similar learning domains. Several approaches have been proposed for the characterization of learning domain, including general, statistical and information theoretic measures.

Apart from the characterization of each domain, the performance of each learning algorithm on that domain is recorded. When a new domain arrives, the performance of the algorithms has retrieved and the algorithms are ranked according to their average performance. Generally algorithms are ranked based on a measure called Adjusted Ratio of Ratios (ARR). That combines accuracy, learning time of algorithm. The selection of algorithms is based on their local performance, but not around the test dataset itself, and also comprising the predictions of the classification models on the test instance. Training data are produced by recording the predictions of each algorithm, using the full training data both for training and for testing.

3.2 Classifier Fusion

The classifier fusion approach is capable of taking several specialized classifiers as input (possibly along with some raw data) and learning from training data how well they perform and how their outputs should be combined. In addition to data fusion, relies quite heavily on machine learning.

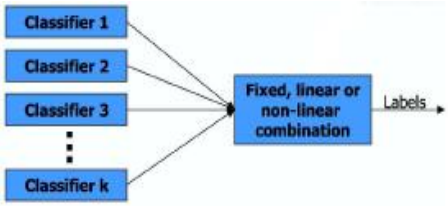

Figure 2. Fusing the classifiers for the dataset

This method assumes that the classifiers in the pool are trying to solve the same classification problem. As a result, only adequate fusing classifiers that can attempt to detect the entire set.

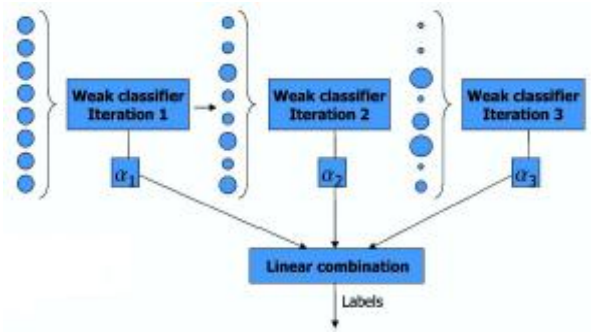

Figure 3. Changing weights over the classifiers

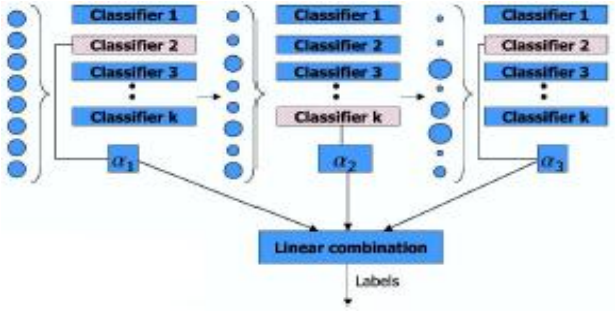

Figure 4. Selecting the best at each iteration

The varying diameters of the blue circles represent the changing weights over the training examples. At each iteration, the weak classifier is trained on the current distribution over the training data. The next figure gives to train all the classifiers in the pool and select the best performing one at each iteration.

3.3 Classification Methods

Various classification models are used in the area of data mining and knowledge discovery. Different classification methods were used such as Bayes, Function, Meta, and Rule. Each method has its own variety of algorithms. Various algorithms of these methods were used to predict the accuracy of the dataset. Each model is associated with a coefficient, usually proportional to its classification accuracy.

3.3.1 Bayes

The Naive Bayes classifier (NB) is a simple but effective classifier that has been used in numerous applications of information processing including, natural language processing, information retrieval, etc.

Naïve Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. The Naive-Bayes inducer computes conditional probabilities of the classes given the instance and picks the class with the highest posterior. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting.

3.3.2 Functions

Function algorithms are classified by type of mathematical equation that represents their relationship.

Logistic regression is a well-known statistical technique that is used for modeling binary outcomes, such as 0 or 1. Logistic Regression Models presents an overview of the full range of logistic models, including binary, proportional, ordered, partially ordered, and unordered categorical response regression procedures. A simple logistic regression is used for prediction of the probability of occurrence of an event by fitting data to a logistic curve. It is a generalized linear model used for binomial regression. Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical.

A radial basis function network (RBF) is an artificial neural network that uses radial basis functions as activation functions. A radial basis function is a real-valued function whose value depends only on the distance. It is a linear combination of radial basis functions. They are used in function approximation, time series prediction, and control. Radial basis function networks typically have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer. In a RBF network there are three types of parameters that need to be chosen to adapt the network for a particular task.

SMO is stands for Sequential Minimal Optimization algorithm. The idea behind SMO is that avoiding the large matrix computation, the SMO can handle very large training sets in between linear and quadratic time with a linear amount of memory in the training set size. The SMO algorithm performs especially well with data sets and mainly used to improve the performance for datasets.

3.3.3 Meta

Meta classifier used to develop the statistical model for given dataset to predict the accuracy.

Bootstrap aggregating (bagging) and boosting are useful techniques to improve the predictive performance of tree models. Boosting is also be useful in connection with many other models, e.g. for additive models with high-dimensional predictors; whereas bagging is most prominent for improving tree algorithms. Boosting is a very different method to generate multiple predictions (function estimates) and combine them linearly.

Boosting provides a ready method for improving existing learning algorithms for classification. Taking a weaker learner as input, boosters use the weak learner to generate weak hypotheses, which are combined into a classification rule more accurate than the weak hypotheses.

3.3.3 Rules

Rule classifier algorithms can be used for classification of datasets with nominal class labels.

Conjugative rule generates the initial rule set and prune two variants of each rule from the randomized data by using grow phase and prune phase procedures. Only one variant is generated from an empty rule while the other is generated, by adding antecedents to the original rule. After all the rules in have been examined and if there are still residual positives, then more rules are generated [2].

OneR classifier; uses the minimum error attribute for prediction, discretizing numeric attributes. Part algorithm uses separate-and-conquer. Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule. ZeroR predicts the mean (for a numeric class) or the mode (for a nominal class).

3.3.4 Trees

Decision trees are a classic way to represent information from a machine-learning algorithm, and offer a fast and powerful way to express structures in data. Decision Tree uses the concept of information gain to make a tree of classificatory decisions with respect to a previously chosen target classification. Decision tree is information produced by data mining techniques that can be represented in many different ways.

J48 builds decision trees from a set of training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class.

J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features. A decision-tree model is built by analyzing training data and the model is used to classify unseen data.

4. EXPERIMENTAL SETUP

The data mining method used to build the model is classification. The data analysis was carried out using WEKA for machine learning environment [11].

The WEKA (Waikato Environment for Knowledge Analysis), Open Source, Portable, GUI-based workbench is a collection of state-of-the-art machine learning algorithms and data pre processing tools [1]. It has been used for conducting the machine learning process that supports several data mining tasks specifically preprocessing classification, clustering, regression, visualization and feature selection. It also consists of multiple interfaces, including an interactive command line interface, a graphical Explorer environment, and a graphical Knowledge Flow environment [5].

In classification, the classes are known and given by so- called class label attributes. The goal of classification is to determine rules on the other attributes that allows predicting the class label attribute. The training data set consists of 180 instances with 24 different attributes. Independent instances have been used for predicting the accuracy for the result. The performance of the classifiers is evaluated and their results are analyzed.

In general, tenfold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier. The 10-fold cross validation was performed to test the performance of the dataset. The purpose of running multiple cross-validations is to obtain more reliable estimates of the risk measures.

In order to determine the quality of the rules derived from the training dataset, the test dataset is used. If rules are of sufficient quality, then it is used in order to classify data that has not been seen before. Since the reliability of the rule has been evaluated by testing it against the test set and assuming that the test set is a representative sample of all data, then the reliability of the rule applied to the dataset should be same.

5. RESULTS AND DISCUSSION

The data set was separated into two parts, one part is used as training data set to produce the prediction model, and the other part is used as test data set to test the accuracy of our model. Two learning performance evaluators are included in WEKA [13]. The Training data set contains feature values as well as classification of each record. Testing is done by 10-fold cross validation method. Dataset was divided into training and testing set by choosing one-fourth records as test cases. These test data were not used for training purpose. Testing was carried out until every data appeared in the test set. Since a clear decision could not be made during the first set of dataset, the train phase was conducted again for a different data to analyze the performance of the various algorithms. A confusion matrix is computed for every test [8].

Usually it is very tough to predict large dataset due to randomness data. Hence testing for larger datasets would give us the flexibility to analyze each algorithm's real effectiveness in prediction. The results of the various classification methods are given below.

Table 1: Predictive Performance of the Classifiers

Evaluation Criteria	Classifiers					
Classification Methods	Bayes		Functions		Meta	
Algorithms	BN	NB	RBF	SMO	Bagging	Attribute Selected Classifier
Correctly classified instances	332	182	254	353	323	309
Incorrectly classified instances	33	183	111	12	42	56
Time taken to build the model (in secs)	0.11	0.03	2.16	0.03	4.69	0.11
Prediction Accuracy (%)	90.95	49.86	69.58	96.71	88.49	84.65

Table 2: Predictive Performance of the Classifiers

Evaluation Criteria	Classifiers			
Classification Methods	Rules		Trees	
Algorithms	JRip	PART	J48	LAD Tree
Correctly classified instances	350	349	350	352
Incorrectly classified instances	15	16	15	13
Time taken to build the model (in secs)	0.22	0.09	0.05	4.99
Prediction Accuracy (%)	95.89	95.61	95.89	96.43

The predictive performance of the classifiers for a whole dataset is shown in the TABLE I and TABLE II. The various criteria that have been used for evaluation of the classifiers are Time taken to build model, correctly classified instances, incorrectly classified instances, and predictive accuracy.

From the above table it is seen that five classification methods are compared. It is important to note that the time taken for total number of instances have been varied and increased to a higher amount.

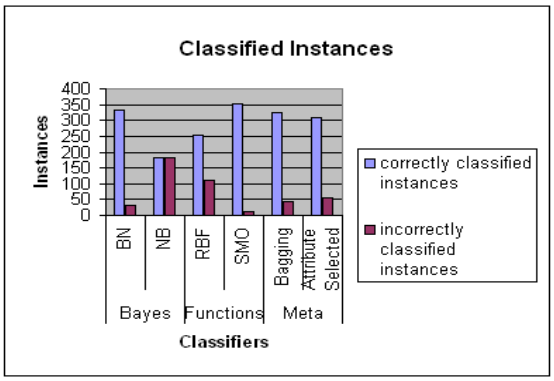

Figure 5. Classified Instance

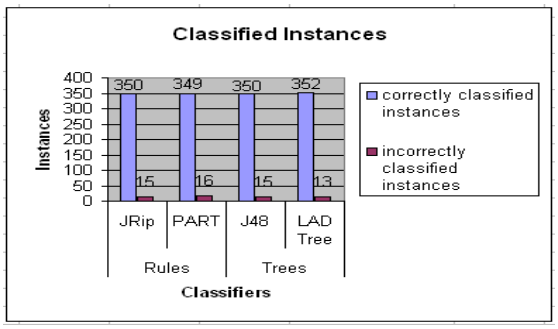

Figure 6. Classified Instances

The Fig. 5 and 6 clearly describes the comparison of number of correctly classified and incorrectly classified instances. Every classification method has a high percentage of correctly classified instances. From this figure it is seen that SMO algorithm has 353 instances is correctly classified and 12 instances is incorrectly classified which gives highest classified instance when compared to other algorithms.

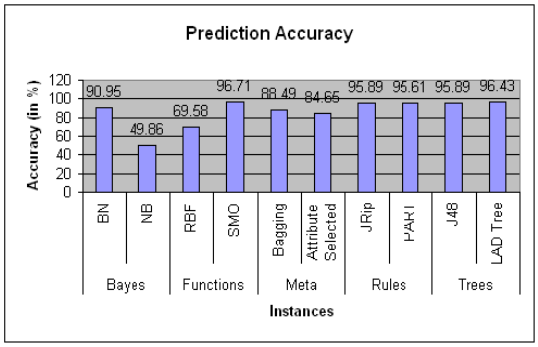

Figure 7. Prediction Accuracy

The prediction accuracy is a parameter that delineates how accurate an algorithm predicts the required data. The predictive accuracy was calculated using the formula shown below.

$$\text{Prediction accuracy} = \frac{\text{Number of Correctly Classified Instances}}{\text{Total Number of Instances}} \quad (1)$$

$$\text{Total Number of Instances} = \text{Correctly Classified Instances} + \text{Incorrectly Classified Instances} \quad (2)$$

The predictive accuracy for various algorithms is shown in the above graph. From the above figure it is seen that SMO has the best predictive accuracy.

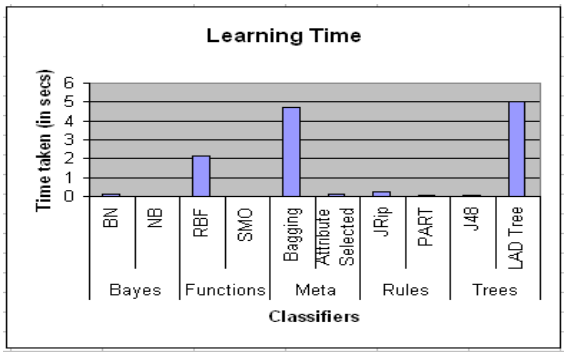

Figure 8. Learning Time

The time taken to build the model gives an idea on how fast the classifier works on he given dataset. In the above figure, the time taken to build model is plotted in the shape of a bar graph and compared for various algorithms. From implementation, it can be understood that since the data set is large it takes quite some time for the algorithm to build. For this criterion Naive Bayes and SMO took the least time and hence it is the useful in time critical applications where the time required to build the model plays a significant role in its efficiency.

6. CONCLUSION

This paper presents leading edge results in the field of data classification obtained with five different classification methods of weka tool like Function, Bayes, Rule, Meta and Trees. Two different algorithms were selected in each method to predict the accuracy of dataset.

Most of them led to a class nearby their actual classes. These techniques have been implemented using WEKA and the independent trained models were generated. The performance of the learned models was evaluated based on their predictive

accuracy and ease of learning. Based on the experimental results the classification accuracy has found to be better using function classifier than other two classifiers. From the above results it has been observed that the function classifier algorithms play a major role in determining better classification accuracy in the dataset. Thus from all perspectives, the SMO could be deemed to be the most efficient.

7. REFERENCES

[1] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, Sally Jo Cunningham, "WEKA: Practical Machine Learning Tools and Techniques with Java Implementations,".

[2] Tulai, A., Oppacher, F., 2004. Maintaining Diversity and Increasing the Accuracy the Accuracy of Classification Rules through Automatic Speciation. Congress of Evolutionary Computation, Portland, USA, 2241-2248

[3] Introduction to Machine Learning and Data Mining: Peng Du, Wenxiang Yao.

[4] Alm, J., (1999). "Tax Compliance and Administration,In *Handbook on Taxation*; eds. Hildreth, W. B.,Richardson, J. A.,pp. 741-768. Marcel Dekker, Inc.

[5] Weka 3: Data Mining Software in Java http:// www.cs .waikato .ac.nz/ml/weka/.

[6] Jiawei Han and Micheline Kamber (2001). *Data Mining: concepts and techniques*. Academic Press,San Diego, California.

[7] Cecil, Wayne H. (1998) Assuring Individual Taxpayer Compliance: Audit rates, Selection Methods, and Electronic Auditing. *The CPA Journal*, 68, (12), available at http://www.nysscpa.org/cpajournal/ 1998/ 1198/ Departments/D661198.html, last accessed 27 Sep 2007

[8] Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementation, by I. H. Witten and E. Frank, Morgan Kanfmann Publishers, 2000.

[9] Murray, Mathew N. (1995) Sales Tax Compliance and Audit Selection. *National Tax Journal*. 48, (4), 515-30.

[10] Michalski RS, Kaufman K. Learning patterns in noisy data: the AQ approach. In: Paliouras G, Karkaletsis V,Spyropoulos C, editors. Machine learning and its applications. Berlin: Springer-Verlag; 2001. p. 22–38.

[11] M. Pazzani and D. Kibler, The Utility of Knowledge in Inductive Learning, Machine Learning, Vol. 9, No. 1, 1992, pp. 57-94.

[12] Micci-Barreca Daniele, Ramachandran Satheesh.(2006) Analytics Elite. Predictive Tax Compliance Management.

[13] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Second edition, 2005. Morgan aufmann

[14] GfAlbrecht, C.C., Albrecht, W.S. and Dunn, J.G. (2001), "Can auditors detect fraud: a review of the research evidence", Journal of Forensic Accounting, Vol. 2 No. 1, pp. 1-12.

[15] Kalousis, A., Theoharis, T., " NDesign, Iimplementation and performance results of an intelligent assistant for classifier selection", In: Intelligent Data Analysis, (1999).

[16] U. M. Fayyad, G. Piatetsky-Shapiro, P.Smyth, and R. G. R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press / The MIT Press, Menlo Park, CA. 1996.

[17] T. Mitchell, "Machine learning", Ed. Mc Graw-Hill International Editions, 1997.

[18] Teknomo, Kardi. K-Nearest Neighbors Tutorial.

[19] Chen S., "Nonlinear time series modeling and prediction using Gaussian RBF networks with enhanced clustering and RLS learning", Inst. Elect. Eng. Electron. Lett.31:17– 118, 1995.

[20] Tulai.A., Oppacher, F., "Multiple Species Weighted Voting – a Genetic-Based Machine Learning System", Genetic and Evolutionary Computation Conference, Seattle, USA, 1263-1274.

[21] Watts, R. L., and J. L. Zimmerman, 1986, Positive Accounting Theory.Prentice-Hall.

[22] Inza I., Larranaga P. and Sierra B.," Feature Subset Selection by Bayesian Networks: A Comparison with Genetic and Sequential Algorithms", International Journal of Approximate Reasoning 27, pp143-164, 2001.

[23] Brazdil, P.B., Soares, C., Da Costa, J.P.: Ranking Learning Algorithms: Using IBL and Meta-Learning on Accuracy and Time Results. Machine Learning 50 (2003) 251-277.