African Journal of Science, Technology, Innovation and Development

ISSN: 2042-1338 (Print) 2042-1346 (Online) Journal homepage: http://www.tandfonline.com/loi/rajs20

A Decision Tree and Naïve Bayes algorithm for income tax prediction

G. V. Mabe-Madisa

To cite this article: G. V. Mabe-Madisa (2018): A Decision Tree and Naïve Bayes algorithm for income tax prediction, African Journal of Science, Technology, Innovation and Development, DOI: 10.1080/20421338.2018.1466440

To link to this article: https://doi.org/10.1080/20421338.2018.1466440

Published online: 04 Jun 2018.

African Journal of Science, Technology, Innovation and Development, 2018
https://doi.org/10.1080/20421338.2018.1466440

A Decision Tree and Naïve Bayes algorithm for income tax prediction

G. V. Mabe-Madisa *

Department of Decision Sciences, University of South Africa, South Africa
**Corresponding author email: mabemgv@unisa.ac.za*

One of the concerns regarding the tax collection system is incorrect case selection. Manual selection of audit cases by auditors (whose role is to detect individual cases of tax non-compliance) based on their expert knowledge of the taxpayers' behaviour, cannot uncover all patterns of non-compliant behaviour hidden in historical data. In addition, random selection of audit cases is not focused on the highest risks. In other words, manual selection has a high opportunity cost if it is used as the sole selection method. Computational intelligence provides methods, techniques and tools, which have been taught to automatically make accurate income tax predictions based on past observations. The data were retrieved from the real time environmental situation. Application of computational intelligence methods proved to be efficient in learning a classification algorithm to classify compliant and non-compliant taxpayers. The new algorithm was evaluated and validated in empirical tests on the same dataset. Although this algorithm had the same performance measurement as Bagging, it outperformed the other existing multiple classifiers in terms of performance. This illustrates an automated system that replicates the investigative operation of human tax risk auditors.

Keywords: classification accuracy, computational methods, ensemble, performance measures, tax compliance

Introduction

Model development is a complex discipline. However, developing a model is not about complex mathematical formulae or complex design diagrams. Rather, the process, in this case related to tax matters, concerns discovering behaviour patterns in human beings and then utilizing those patterns to collaboratively produce knowledge, which can be translated into an executable code. In the animal kingdom, behaviour patterns can be reasonably predicted, but this does not apply to human beings. The concept of ethics has to be observed and respected. Some data are highly confidential and not easily accessed or obtained. Some cases may produce false negatives (being of low risk while they are actually of high risk) and cannot be addressed until there is concrete evidence that an audit has to be conducted, which might sometimes never happen.

Random selection of non-compliant taxpayers based on auditors' expert knowledge of the taxpayers' behaviour is not focused on high risk. In other words, this method has a high opportunity cost if it is used as the sole selection method. Opportunity cost is the cost paid when giving up one option for another: the cost of passing up the next best choice when making a decision. Risk-based case selection identifies those taxpayers that are most likely to be non-compliant (Vellutini 2011).

Income tax is an important source of revenue for governments in developing countries as well as in developed ones (Hudson and Teera 2004). The amount of revenue to be generated by a government from taxes for its expenditure programme depends, among other things, on the willingness of the taxpayers to comply with the tax laws of a country (Eshag 2006). Failure to follow the tax provisions suggests that a taxpayer is committing an act of non-compliance (Kirchler 2007). Tax non-compliance occurs through failure to file tax returns, misreporting income or misreporting allowable subtractions from taxable income or tax due (Ayuba, Saad, and Ariffin 2015; Shome, Aggarwal, and Singh 1996). Monitoring tax compliance involves collection, processing and interpretation of data relating to the condition of critical non-compliant behaviour. Compliant behaviour involves registering for tax, declaring the appropriate income, paying the tax debt and filing the tax returns.

Data-driven case selection, on which this paper's research is based, requires objective criteria, and does not rely on the discretion of the tax official. Instead, computational methods are used to extract valuable information from the data about the taxpayer's behaviour. Data are trained to learn a target function that can be used to predict values of a discrete class attribute, e.g. correct income tax return or incorrect tax return, and high risk or low risk.

It is necessary to categorize this risk into two groups: high risk and low risk. To be able to classify this risk requires development a model. The goal of a prescriptive data-mining effort is to automate a decision-making process by creating a model capable of making a prediction by assigning a label. An important measure of such a model would be its accuracy (Berry and Linoff 2000). The objective of this research, as presented in this paper, was to develop such a model. A novel classifier model, Decision Tree and Naïve Bayes hybrid algorithm (DTNBC), that uses some optimization capability was proposed. The model ultimately needed to be capable of making an accurate prediction.

Tax compliance is still far from optimal, so that enforcement of tax laws is still deficient and has many loopholes. There is no scientific method which helps to address these loopholes of tax compliance. This paper describes a proposal to predict high-risk taxpayers' characteristics and how to distinguish them from low-risk taxpayers' characteristics. The model needs to assist with the future decision on whether to send a client to be audited. To reiterate, the focus of this paper is on the process of determining a target function to be able to

predict the values of a discrete class attribute, for example, a high-risk or a low-risk taxpayer.

Taxation

Taxation is a means through which governments finance their expenditure by imposing charges on citizens and corporate entities (Jenkins, Kuo, and Shukla 2000). There are different types of tax, but just the major one, that this paper focuses on, i.e. national tax (personal income tax), is briefly discussed. Personal income tax (PIT) provides an overview of assessed PIT revenues of registered individual taxpayers. It provides information on taxable income by income categories, age, gender, source of income, fringe benefits, allowances and other deductions. In other words, unlike other taxes, it provides the necessary attributes for classification. It should be noted that a large percentage of income taxpayers (standard income tax on employees (SITE)-only taxpayers) – those with taxable income below a certain amount) are not considered in this paper. SITE-only taxpayers do not have to register with the Revenue Service and are not required to file annual tax returns. Although larger in number, their contribution to total PIT revenues is small (S.A. Tax Statistics 2016).

A profile of the customer is continually developed by being informed through collecting and analyzing data. This information is used to create a knowledge base of the customer's interactive behaviour. Based on the insight gained about the customer, a subdivision that clusters the customers is created. Attributes such as risk, complexity of revenue, customs affairs, geography and so forth, are used to cluster the customers into groups to facilitate cost-effective differentiation. This is a fundamental shift from matching customers to products to a position where clusters (irrespective of entity type) are identified and offerings are developed according to their profiles and behaviours.

Having a comprehensive view of the customer, the customer contact agents continually add to and build up the customer profile across offerings or products (income tax, pay as you earn (PAYE), value-added tax (VAT), licenses, duties, capital gains tax, trade taxes, etc.). The customer history and profile, and real-time updating of information continually gathered, facilitates the ability to identify risks proactively and take the necessary actions. The necessary actions include applying risk mitigation techniques and tax compliance enforcement by laws. However, this does not allow identification of all the applicable risks and therefore does not reduce the possibility of further risk and non-compliance.

Related works

Most recent research (Farid, Rahman, and Tani 2012; Karim and Rahman 2013; Kohavi 2011) conducted, indicates that ensemble classifiers have improved classification performance compared to individual or constituent classifiers. In addition, for example, Keegan et al. (2016) showed that the Decision-Tree-based support vector machine which combines support vector machines and Decision Trees, could be an effective way for solving multi-class problems in Intrusion Detection Systems. Moreover, examining the performance of the base classifiers, Molale, Seeletse, and Twala (2013) found that Decision Trees, support vector machines and Artificial Neural Networks have the highest accuracy. Their findings further showed that all the multi-stage systems significantly outperform the baseline classifiers.

Several studies (Altincay 2007; Agrawal and Bala 2011; Gjorgjevikj, Madzarov, and Tomche 2010) contend that ensemble methods are able to improve the predictive performance of many base classifiers. Base classifiers refer to individual classifiers used to construct the ensemble classifiers. Twala (2009) also explored the predicted behaviour of classifiers for different types of noise in terms of credit risk prediction accuracy, and the way such accuracy could be improved by using classifier ensembles. His findings indicated that the classifier ensembles improve accuracy relative to individual classifiers.

Characteristics of data like size and correlation, among others, have an effect on model performance (Sakizadeh 2015). Different models may perform well in different conditions (Mojirsheibani 1999). Combining models in such a way that strengths of individual model form a synergy, as well as offsetting weakness of individual methods is regarded as an effective way to synthesize individual models for combination. Based on the above findings, this paper addresses the said weakness by describing how this research developed a combined classifier or hybrid algorithm based on optimal classifiers for the data using computational intelligence methods.

Classifier construction

Data

Dealing with risk requires skilful use of information. Information is useful data processed to create and increase knowledge (Hoffer, Ramesh, and Topi 2015). Knowledge in the context of this study is increased by extracting new, unexplored and interesting patterns using statistical and mathematical techniques from large sets of data stored electronically in databases. Many businesses have large databases of customer information. They use information to make useful prediction models about the market and the customer. Data mining and knowledge discovery are used for extracting valuable information from these databases.

The records of a dataset have K attributes: $A_1, A_2, \ldots, A_k$, and each example is labelled with a predefined class. The goal is to learn a classification function from the data that can be used to predict the classes of new examples/cases/instances.

A receiver of revenue receives many thousands of tax returns. Each return contains information (attributes / variables) about the taxpayer, e.g. identity, age, gender, taxable income, registered contractors tax (RCT), industry code, outstanding returns, and the like. Some of these variables are explained as follows. Taxable income is the income from employment such as salaries, wages, bonuses, overtime pay, taxable (fringe) benefits, allowances and certain lump sum benefits. The abbreviation, RCT, refers to the statuses of the contractors who are registered under other contractors when they register for tax. For example, if the contract is a contract of employment, the

subcontractor is an employee and the earnings are subject to PAYE.

The Standard Industry Classification (SIC) is part of the international convention that allocates a unique classification number to the different types of economic activities that take place in a normal national economy. The system initially divides the economy into three broad categories: the primary (agriculture, forestry, fishing and mining and quarrying), secondary (manufacturing) and tertiary (construction, packaging, distribution and provision of services.

Sources, adequacy and accuracy of data

A revenue authority requires adequate quality data to successfully administer the taxation system (Hasseldine et al. 2001). In its raw and in its processed form, the data informs the risk identification, assessment and prioritization processes. In identification, the data are aggregated to inform the strategic risk identification. In assessment, it is used to determine the extent of the risk whereas in prioritization it is used to select the cases.

The tax data may be sourced from the tax administrative data, which is necessarily in the form of annual tax returns or declarations of income. In other words, this is taxpayer history data from the databases and data warehouses. The data used for this study is the taxpayer history data from the databases. For risk identification purposes, the tax return and industry classification code data are required. A tax return is the declared taxable income. This enables the Receiver of Revenue to better understand taxpayers' behaviour, identify specific economic activities that reveal unique taxpayer behaviour and which reinforces the comprehensive compliance approach.

To prevent incorrect identification of risk, the data must be accurate. For example, revenue authorities and employee defects should be avoided. These are mistakes made, particularly by the employees either during data validation and correction processes in call centres or in income tax return capturing or assessments.

The data collected for this research study comprises the personal income tax data for the fiscal years 2006/2007, 2007/2008, 2008/2009, and 2009/2010. To preserve the privacy of the taxpayers, all personal identifiable information was removed. Predictive models are built, or trained, using historical income tax data stored in databases. A sample of tax returns belonging to taxpayers who have outstanding returns as well as those whose submission of income tax returns is up to date, was extracted. Just personal income tax data in the form of tax returns was extracted and used for prediction. Personal income tax was chosen because it accounts for the largest percentage of national tax returns and it is not complex since its features are easily identifiable. The sample consisted of 7 890 tax returns against which the learning algorithms were trained and tested.

The constituent classifiers

The algorithm to learning a Decision Tree (DT) and Naïve Bayes classifiers (NBC) was developed. The said algorithm uses the advantages of both the DT and NBC, that is, splitting or branching and evidence gathered from multiple attributes through probabilities. This combination of algorithms into a hybrid learning algorithm classifies the tax data into no outstanding return (NOR) and outstanding return (OR) classes.

The Decision Tree

The important decision to make about Decision Trees is which decision node (attribute) to choose in order to branch out. The objective is to reduce the impurity or uncertainty in data as much as possible. Data are regarded as pure if all instances or examples belong to the same class. The heuristic intention in this algorithm is to choose the attribute with the maximum information gain based on information theory.

Information theory

In information theory, entropy is used as a measure of impurity or disorder of information. As an example, considering a tossed coin, the question is whether this coin would come up heads or not. The answer to this question is less informative if one has a good guess about it. If one already knows that the coin is rigged so that it will come up heads with a probability of 0.99 for instance, then this information about a flip is worth less than it would be for a fair coin that comes up heads with a probability of 0.5. With this probability, there is no prior information about the outcome, which therefore makes this information valuable. Information theory uses this same intuition but measures information contents in bits. One bit of information is sufficient to answer a question that needs a yes or no answer. This theory provides a mathematical basis for measuring the information content (Liu 2007).

$$\text{Entropy (S)} = -\sum_{j=1}^{|c|} \Pr\ (c_j)\ \log_2 \Pr\ (c_j), \text{ and } \sum_{j=1}^{|c|} \Pr\ (c_j) = 1. \tag{1}$$

$\Pr\ (c_j)$ is the probability of class c_j in dataset S. As the data become purer, the entropy decreases. If attribute A_j, with $A_{|A|}$ values, is made the root of the current tree, it will partition S into $A_{|A|}$ subsets $S_1, S_2, \ldots, S_v$, and have the entropy:

$$\text{Entropy}_{A_i}(S) = \sum_{j=1}^{v} \frac{|S_j|}{|S|} \times \text{entropy}\ (S_j) \tag{2}$$

Information gained by selecting attribute A_j to partition the data is

$$\text{gain}\ (S, A_j) = \text{entropy}\ (S) - \text{entropy}_{A_j}(S) \tag{3}$$

The attribute with the highest gain is chosen to split the tree (Liu 2007).

The Naïve Bayes classifier

Bayesian classification is a learning from a probabilistic point of view. For example, having attributes A_1 through

A_k, and a test example d with observed attribute values a_1 through a_k. Classification is basically to compute the posterior probability. The prediction is the class c_j such that (Liu 2007)

$$\Pr(C = c_j | A_1 = a_1, \ldots, A_{|A|} = a_{|A|}) \text{ is maximal}$$

$$\Pr(C = c_j | A_1 = a_1, \ldots, A_{|A|} = a_{|A|}) = \frac{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} | C = c_j) \Pr(C = c_j)}{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|})}$$

$$= \frac{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} | C = c_j) \Pr(C = c_j)}{\sum_{r=1}^{|c|} \Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} | C = c_r) \Pr(C = c_r)}.$$

$\Pr(C = c_j)$ is the class prior probability and is estimated from the training data.

Computing probabilities

The denominator $(\Pr(A_1 = a_1, \ldots, A_{|k|} = a_{|k|})$ is irrelevant for decision making since it is the same for every class. The term needed is $\Pr(A_1 = a_1, \ldots, A_{|k|} = a_{|k|} | C = c_j)$, and can be written as

$$\Pr(A_1 = a_1 | A_2 = a_2, \ldots, A_k = a_k, C = c_j) * \Pr(A_2 = a_2, \ldots, A_k = a_k | C = c_j).$$

The second factor above can be written recursively in the same way, and so on. All attributes are conditionally independent given the class $C = c_j$. It is assumed that (Liu 2007)

$\Pr(A_1 = a_1 | A_2 = a_2, \ldots, A_{|A|} = a_{|A|}, C = c_j) = \Pr(A_1 = a_1 | C = c_j)$ and so on for A_2 through $A_{|A|}$. i.e. $\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} | C = c_i) = \prod_{i=1}^{|A|} \Pr(A_i = a_i | C = c_j)$.

$$\Pr(C = c_j | A_1 = a_1, \ldots, A_{|A|} = a_{|A|}) = \frac{\Pr(C = c_j) \prod_{i=1}^{|A|} \Pr(A_i = a_i | C = c_j)}{\sum_{r=1}^{|C|} \Pr(C = c_r) \prod_{i=1}^{|A|} \Pr(A_i = a_i | C = c_r)}.$$

$$c = \underset{C_j}{\operatorname{argmax}} \Pr(c_j) = \prod_{i=1}^{|A|} \Pr(A_i = a_i | C = c_j).$$

The design cycle

In stage 1, the DT evaluates which variable to use for splitting by applying the information theory. DT selects the feature with the highest information gain as a root node. It continues selecting other additional possible features one after the other and the process continues with the classification process until the entire feature vector, $A^T = [A_1, A_2, \ldots, A_n]$ is completely used up for grouping. The NBC exploits context input dependent information using probabilities to improve performance. In stage 2, just the DT is used for classification. Figure 1 illustrates the process.

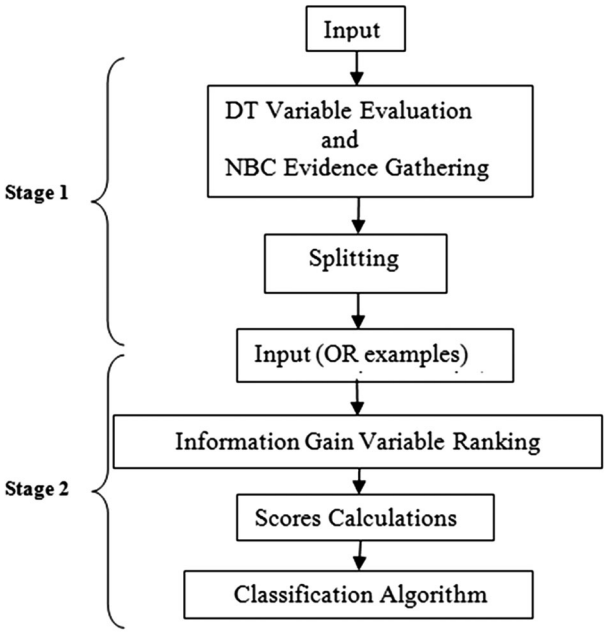


Figure 1: The design cycle.

Stage 1 classification

Firstly, the dimensions along which the tax returns are maximally different were determined, and data extracted accordingly. The DT is used in selecting the most suitable attributes among the rest of the attributes and eliminating the redundant ones using the information theory already explained. A DT is built with univariate splits at each node and NBC probabilities at the leaves. An example is assigned to the class which has the largest posterior probability if it is a suitable one; otherwise it is assigned to the other class.

The following is a description of the process at the leaves, given the training data S with variables $\{A_1, A_2, \ldots, A_n\}$ and each variable containing the following variables $\{v_1, v_2, \ldots, v_m\}$. The variables can be discrete or continuous. This training data S also contains the set of classes $C = \{C_1, C_2, \ldots, C_n\}$. Each example in the training data S has a particular C_j, $(j = 1, \ldots, n)$. The posterior probabilities can be defined as $\Pr(C = c_j | A_1 = a_1, \ldots, A_{|A|} = a_{|A|})$. The algorithm first searches for the multiple copies of the same example in the training data S; for example, suppose all variable values of two examples are equal then the two examples are similar.

The algorithm then discretizes the continuous variables in the training data S by finding each adjacent pair of continuous variable values that are not classified into the same class value for that continuous variable. The algorithm then calculates the prior probabilities and the posterior probabilities in the training data S. The prior probability for each class is estimated by counting how frequently each class occurs in the training data S. For each variable the number of occurrences of each variable value can be estimated by counting how frequently each variable occurs in the class in the training data S. The algorithm classifies all the examples in the training data S with these prior probabilities and posterior probabilities to make the prediction.

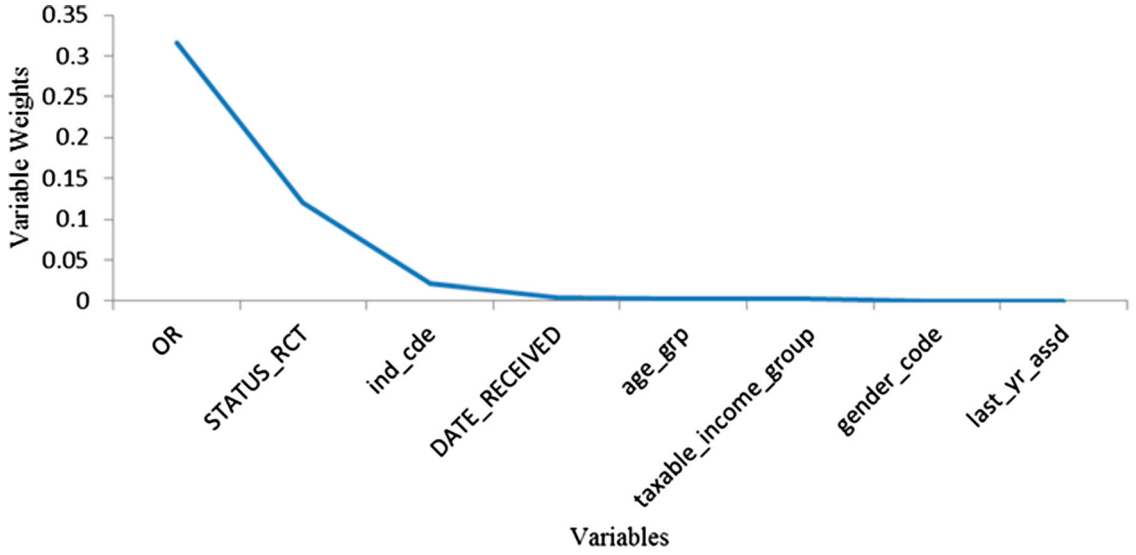


Figure 2: Ranked variables.

Referring briefly to the learning algorithms that are used as base learners, the most commonly used C4.5 algorithm (Quinlan 1993) is representative of the Decision Trees in this study. The NBC is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions.

Stage 2 classification

All the examples classified as having outstanding returns in stage 1, are now used as input data to stage 2 classification. On developing the stage 2 algorithm, consideration was given to the fact that the variables (features) are not of the same weight. For example, the variable, 'taxable income', does not carry the same weight as the variable 'outstanding returns'. A rank for each variable A_i was derived using the information theory's information gain where the weights are selected proportionally to an estimate of the amount of discriminative information. The graph of the ranked variables is presented in Figure 2. A score for each variable A_i was derived as a ratio of an estimate of the amount of discriminative information to the weighted sum of all the variables. These scores are the probabilities and their sum is equal to one. A decision rule for high risk or low risk, for each record or case, was derived as a probability of that record being greater than or less than a determined threshold.

This model classifies income tax returns of taxpayers who had outstanding returns in the previous years. In stage 2, the modification of a DT decision rule of a learning algorithm considers classifying tax return documents by their content. The ranked variables are used to determine the likelihood that the new case belongs to a specific class. A scoring method derived from the weights of the ranked variables is used. A tax return document containing these ranked variables with a score greater than a determined threshold (in this case, 0.009) is considered to be of high risk. The following explains how the scores were determined.

OR = *N*, score 0 otherwise 0.3160704
STATUS_RCT = 11, score 0 otherwise 0.1204764
ind_cde = 22, score 0 otherwise 0.0206188
DATE_RECEIVED = month {Oct, Nov, Dec}, score 0 otherwise 0.0044002
age_grp = 65+, score 0 otherwise 0.0024728
taxable_income_group = {a-m}, score 0 otherwise 0.0024633

The algorithm DTNBC

*Stage 1: Classification Rule 1 (*Liu 2007*)*

If S contains only training examples of the same class $c_j \in C$ **then**
make N a leaf node labelled with class c_j;
else if $A = \emptyset$ make N a leaf node labelled with class c_j, which is the most frequent class in S
// S contains examples belonging to a mixture of classes. Select a single
// variable to partition S into subsets so that each subset is purer
for each variable $A_i \in \{A_1, A_2, \ldots, A_n\}$ do
gain (S, A_j) = entropy (S) − entropy$_{A_j}(S)$
end
Select $A_g \in \{A_1, A_2, \ldots, A_n\}$ that gives the greatest gain,
if gain < threshold **then** //A_g does not significantly reduce impurity
make N a leaf node labelled with c_j, the most frequent class in S
else //A_g is able to reduce impurity
Make N a decision node on A_g;
Let the possible values of A_g be $\{v_1, v_2, \ldots, v_m\}$,
Partition S into m
disjoint subsets $S_1, S_2, \ldots, S_m$ based on the m values of A_g.
for each S_j *in* $\{S_1, S_2, \ldots, S_m\}$ **do**
if $S_j \neq \emptyset$ **then**
create a branch node N_j for v_j as a child node of N,
DT(S_j, $A - \{A_g\}$, N_j)//A_g is pruned
end
end
end
end

Stage 1: Classification Rule 2

Classify example to class 1 if

$$\Pr(C = c_j | A_1 = a_1, \ldots, A_{|A|} = a_{|A|}) \text{ is maximal}$$

$$\text{where: } \Pr(C = c_j | A_1 = a_1, \ldots, A_{|A|} = a_{|A|}) = \frac{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} | C = c_j) \Pr(C = c_j)}{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|})}.$$

//Assign example to the class, which has the largest posterior probability//

End

Stage 2: Classification Rule

// Rank the n variables using information theory's information Gain or Gain ratio
Get the n variables' scores / ranks
Add the n scores / ranks (*SumOfScores*)
Name the ranked variables $SCORE_i, \ldots, SCORE_n$ //

Determine the Score

//For variables $A_i, \ldots, A_n$//
If example has A_1, $SCORE = SCORE_1$ otherwise $SCORE = 0$.
If example has A_n, $SCORE = SCORE_n$ otherwise $SCORE = 0$

TotalScore = 0;
for each example **do**
//Repeat the following steps for each example

(a) Get Account_ID // taxID given the last seven digits of the ID //
(b) Use the algorithm above to determine the score

Determine the Probability of each example in the file as follows //

$$\text{Probability (ratio)} = \frac{\text{TotalScore}}{\text{SumOfScores}}$$

If Probability > threshold, example is of High Risk otherwise example is of Low Risk.

Experiments

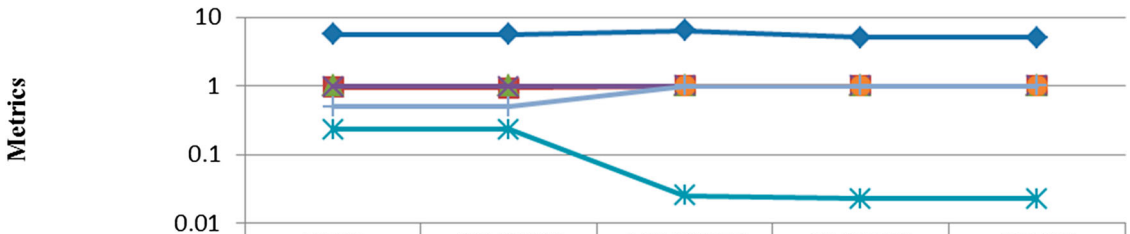
The application of classifiers to the problem of tax non-compliance prediction was investigated. Four supervised learning multiple classifiers were considered. These were: Vote, Stacking, Boosting and Bagging. The performance of these multiple classifiers and the combined classifiers algorithm (DTNBC) were empirically evaluated in terms of their ability to correctly predict high risk and low risk in terms of non-compliance outcome, outstanding returns (OR) using personal income tax data. The results from all systems were compared with one another.

Classifiers were constructed using a ten-fold cross-validation method as the sampling method. The multiple classification methods, together with the DTNBC were each applied to the training, validation and the test samples. Classifiers were constructed using the Waikato Environment for Knowledge Analysis (WEKA) free and open software that uses the Java™ language (Frank, Witten, and Hall 2011) and a 2.60 GHz CPU microcomputer. WEKA is a collection of algorithms for data mining tasks. All algorithms were implemented in WEKA Release 3.6.9. Their performances were compared using different performance metrics such as accuracy, error rate, precision, recall, F-value and squared error. The performance of DTNBC was compared with that of the existing multiple classifiers.

DTNBC and multiple classifiers comparison

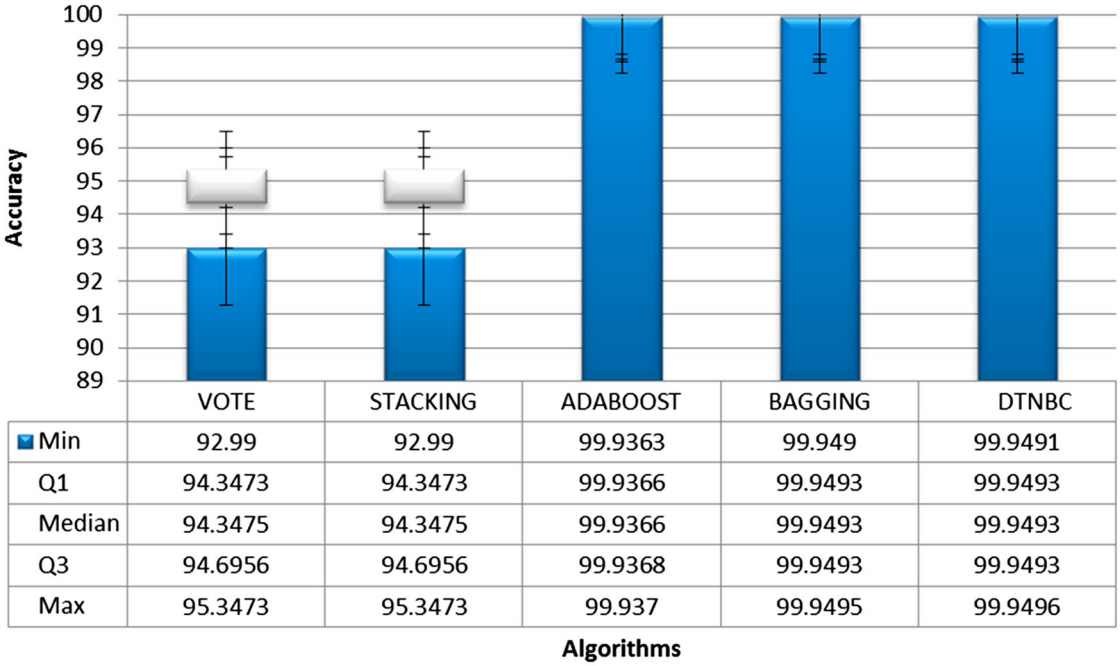
The graph of the comparison, in terms of Error rate, Precision, Recall, F-value, Squared Error, Kappa Statistics and Receiver operating curve (ROC) area, is illustrated in Figure 3.

The performance in terms of accuracy is shown in Figure 4 as box plots. Although the DTNBC algorithm had the same performance measure as Bagging, it outperformed the other learning algorithms in terms of accuracy. The DTNBC and Bagging achieve the highest accuracy rate, with an error rate of 0.0507%. These are followed by AdaBoost with an error rate of 0.0634%, while Vote and Stacking both have an error rate of 5.65%. The performances between some of these supervised learning methods were found to be significantly different at the 5% level. Filing of tax returns (NOR) /



	VOTE	STACKING	ADABOOST	BAGGING	DTNBC
Error rate	5.65	5.65	6.34	5.07	5.07
Precision	0.944	0.943	0.989	0.991	0.991
Recall	1	1	1	0.996	0.996
F-value	0.971	0.971	0.994	0.995	0.995
Squared Error	0.2309	0.2309	0.0251	0.0227	0.0227
Kappa statistic	0	0	0.9941	0.9953	0.9953
ROC Area	0.5	0.5	1	1	1

Figure 3: The performance of DTNBC and different multiple classifiers.



	VOTE	STACKING	ADABOOST	BAGGING	DTNBC
Min	92.99	92.99	99.9363	99.949	99.9491
Q1	94.3473	94.3473	99.9366	99.9493	99.9493
Median	94.3475	94.3475	99.9366	99.9493	99.9493
Q3	94.6956	94.6956	99.9368	99.9493	99.9493
Max	95.3473	95.3473	99.937	99.9495	99.9496

Figure 4: Box plots of accuracies of multiple classifiers and DTNBC.

non-filing (OR) attribute was found to be the most highly significant at the 5% level.

Summary

The combined approach provides a spectrum of solutions to learning. The Naïve Bayes method is at one end of the continuum, while the Decision Tree lies at the other end. The two algorithms are combined by learning a Decision Tree using pruning methods and including a Naïve Bayes method at the leaf nodes consisting of the remaining variables that have not been branched upon. The second stage of DTNBC ensures that there are no false positives and negatives by classifying the output data from stage one into high risk (OR) and low risk (NOR).

In stage 2, the modification of a decision rule of a learning algorithm, to the tax return, considers classifying tax return documents by their content. The ranked variables are used to determine the likelihood that the new case belongs to a class. For ethical reasons, stage 2 ascertains whether a taxpayer has to be audited or not. In this way the cost of labelling a taxpayer as being of high risk if he/she in fact not, is avoided. The cost of labelling a taxpayer as of low risk is better than the opposite. In this stage, the tree learns, from a training set, the individual importance of each variable relation by computing the gain ratio or score (Quinlan 1993) of each variable. Based on these computed ratios (scores), a sum of scores is derived, which is the arithmetical sum of these n ratios (scores), where n is the number of variables.

For each example, a ratio (score) is determined on whether the example has a value for a particular variable or not; for example, does s/he have an outstanding return or not or whether the taxpayer is a pensioner or not and so on. A total score for this example is computed as an arithmetical sum of these ratios (scores). Ultimately a ratio, TotalScore/SumOfScores is determined for each example and this ratio establishes whether an example or taxpayer is of high risk or not depending upon whether the ratio is greater than or less than a determined threshold.

The new decision rule determines the scores based on information gain and computes the sum of these scores. It then determines the ratio of the total score a taxpayer acquires to the sum of scores of the variables. Based on this ratio, a high-risk or low-risk taxpayer is identified. The classification is performed by some ratio that exceeds or does not exceed a certain threshold.

The gain ratio DT learns from a training set the individual importance of each variable by computing the gain ratio (Quinlan 1993). Based on this ratio, a binary tree is constructed where a leaf indicates a class, and a decision node chooses between two subtrees based on the presence of a variable relation. The more important the variable relation is for the classification task at hand, the closer it is located near the root of the tree. In other words, the classification is carried out by the branching of a tree. This study illustrates an industrial implementation of an adaptive risk-case building system that uses computational intelligence methods to conduct the search and decision-making process.

Discussion

A novel approach, that is a hybrid or a combination between a Decision Tree method and the Naïve Bayes method (DTNBC), was analyzed to determine why this method differs from the existing multiple classifiers in improving generalization. This approach takes advantage of the Naïve Bayes' fast condition and Decision Tree to break the Naïve Bayes' dependence condition. Pruning is implemented to avoid over-fitting and memory problems of a Decision Tree. As the data used in this study is large, pruning helps by keeping the tree to a manageable size. In the case where some of the data is lost due to pruning, Naïve Bayes is there with its predictions to recover the lost data. Because it is fast and does not consume memory, it is a better classifier.

In addition, the success of the DTNBC classification algorithm is due to the fact that its base classifiers were first tested to establish whether they were suitable for the data. The 'no free lunch' theorem is the key to constructing reliable classifiers. In the first stage, the two techniques, DT and NBC are combined by teaching a Decision Tree using pruning methods and including an NBC model at the leaf nodes to gather evidence before the final classification. This method takes advantage of the simplicity of NBC. As the number of variables increases, it is not possible to use the NBC technique. This is due to the explosion of nodes if branching occurs on every variable.

This limitation is overcome by pruning. Pruning ensures that the final tree has not learnt unwanted patterns and it restricts the amount of resources consumed. A DT is pruned due to memory consumption and to avoid overfitting. As pruning a DT results in the loss of data, NBC helps by recovering some of the lost data, which leads to better predictions. This is also the reason why the variables were initially extracted based on domain knowledge to make certain that relevant variables are included. DT alone cannot handle continuous attributes. This limitation is overcome by the presence of NBC which is used to handle continuous attributes.

Conclusion

This paper described the proposed new hybrid algorithm approach to teaching the Decision Tree and Naïve Bayes classifiers. This algorithm was thought of as the classification algorithm that could improve generalization on the tax dataset. Human behaviour keeps changing; the knowledge base therefore has to be updated to keep track of the new behaviour patterns or information. While high-risk case selection does not allow identification of all the applicable risks and therefore does not completely reduce the possibility of further risk and tax non-compliance behaviour, a great deal of non-compliance behaviour is reduced.

The performances of the developed algorithm and the existing multiple classifiers were compared. The new algorithm was evaluated and validated in empirical tests on the same dataset. Although this algorithm has the same performance measurement as Bagging, it outperforms the other existing multiple classifiers in terms of performance.

The original idea behind the development of this novel model, – the DTNBC, – was to develop a thinking tool that could be used to identify the tax non-compliance risk. A gap was realized in the existing tool that was currently used to generate cases or identify risk. The proposed DTNBC model was then improved to meet this need. The study addressed the problem of choosing the most suitable risk identification methodology practices.

Acknowledgements

I thank the University of South Africa, CEMS research department, for their support in this research.

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

G. V. Mabe-Madisa http://orcid.org/0000-0002-3678-2177

References

Agrawal, R. K., and M. Bala. 2011. "Optimal Decision Tree Based Multi-Class Support Vector Machine." *Informatica* 35: 197–209.

Altincay, H. 2007. "Decision Trees Using Model Ensemble-Based Nodes." *Pattern Recognition* 40: 3540–3551.

Ayuba, A., N. Saad, and Z. Z. Ariffin. 2015. "Interacting Role of Perceived Service Orientation on Work Family Conflict, Fuel Subsidy Removal and Tax Compliance Behavior: Evidence From Nigerian SMEs." *Asian Social Science* 11 (28): 1911–2025. doi:10.5539/ass.v11n28p226.

Berry, M. J. A., and G. S. Linoff. 2000. *Mastering Data Mining: the art and Science of Customer Relationship Management*. New York: Wiley and Sons.

Eshag, E. 2006. *Fiscal and Monetary Policies and Problems in Development Countries*. Cambridge: Cambridge University Press. Pp 287.

Farid, D. M., M. Z. Rahman, and F. Y. Tani. 2012. "Ensemble of Decision Tree Classifiers for Mining Web Data Streams." *International Journal of Applied Information Systems* 1 (2): 2249–0868.

Frank, E., I. H. Witten, and M. Hall. 2011. *Data Mining, Practical Machine Learning Tools and Techniques*. 3rd ed. Burlington: Morgan Kaufmann.

Gjorgjevikj, D., G. Madzarov, and D. Tomche. 2010. *Ensembles of Binary SVM Decision Trees. ICT Innovations 2010 Web Proceedings* ISSN 1857-7288.

Hasseldine, J., S. James, P. White, and M. Toumi. 2001. "Developing a Tax Compliance Strategy for Revenue Services." *Bulletin From the International Bureau of Fiscal Documentation* 55 (4): 158–164.

Hoffer, J. A., V. Ramesh, and H. Topi. 2015. *Modern Database Management*. 10th ed. Reading: Addison-Wesley.

Hudson, J., and J. M. Teera. 2004. "Tax Performance: A Comparative Study." *Journal of International Development* 16: 785–802.

Jenkins, G. P., C. Kuo, and G. P. Shukla. 2000. *Tax Analysis and Revenue Forecasting: Issues and Techniques*. http://www.queensjdiexec.org/publications/qed_dp_169.pdf/.

Karim, M., and R. M. Rahman. 2013. "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing." *Journal of Software Engineering and Applications* 6: 196–206.

Keegan, N., S. Ji, A. Chaudhary, C. Concolato, B. Yu, and D. Jeong. 2016. "A Survey of Cloud-Based Network Intrusion Detection Analysis." *Human-centric Computing and Information Sciences* 6: 19.

Kirchler, E. 2007. *The Economic Psychology of tax Behaviour*. Cambridge: Cambridge University Press.

Kohavi, R. 2011. "Scaling Up the Accuracy of Naïve-Bayes Classifiers: A Decision Tree Hybrid." *Data mining and visualization*. Silicon graphics, Inc. Mountain View.

Liu, B. 2007. *Web Data Mining, Data-Centric Systems and Applications Book Series, 55-116*: *Supervised Learning Lecture, Chapter 3, CS583*, UIC: https://www.cs.uic.edu/~liub/teach/cs583-fall-06/CS583-supervised-learning.ppt.

Mojirsheibani, M. 1999. "Combining Classifiers via Discretization." *Journal of the American Statistical Association* 94 (446): 600–609. doi:10.1080/01621459,1999.10474154.

Molale, P., S. Seeletse, and B. Twala. 2013. "Fingerprint Prediction Using Statistical and Machine Learning Methods." *ICIC Express Letters* 7 (2): 311–316.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Sakizadeh, M. 2015. "Assessment the Performance of Classification Methods in Water Quality Studies, A Case Study in Karaj River." *Environmental Monitoring and Assessment* 187 (9): 573.

Shome, P., Aggarwal, P., and K. Singh. 1996. *The System of tax Deduction at Source (TDS): Coverage, Functioning and Suggestions for Reform*. New Delhi: National Institute of Public finance and policy.

Tax Statistics. 2016. A Joint Publication between National Treasury and the South African Revenue Service.

Twala, B. 2009. "Combining Classifiers for Credit Risk Prediction." *Journal of Systems Science and Systems Engineering* Springer 18: 292–311.

Vellutini, C. 2011. *Risk Based Tax Audit Selection Methods*. Washington: World Bank.