The International Journal of Digital Accounting Research
Vol. 2, No. 4, pp. 131-155
ISSN: 1577-8517

The Sensitivity of Machine Learning Techniques to Variations in Sample Size: A Comparative Analysis

Javier de Andrés. University of Oviedo. Spain.
jdandres@correo.uniovi.es

Pedro Lorca. University of Oviedo. Spain.
plorca@correo.uniovi.es

Elias F. Combarro. University of Oviedo. Spain.
elias@orion.ciencias.uniovi.es

Abstract. A comparative analysis of the performance of some well-known classification techniques (Discriminant Analysis, Quinlan's See5, and Neural Networks) and certain machine learning systems of recent development (ARNI, FAN and SVM) is conducted. The chosen classification task is the forecasting of the level of efficiency of Spanish commercial and industrial companies. Assignment of the firms is made upon the basis of a set of financial ratios, which make a high dimension feature space with low separability degree. In the present research the effects on the accuracy of variations of each technique in the estimation sample size are measured. The main results suggest that ARNI and See5 yield the best results, even with small sample sizes.

Keywords: *Financial Ratios, Machine Learning Algorithms, Efficiency.*

1. INTRODUCTION

Machine Learning Systems (MLS) are proven alternatives to traditional statistical methods for classification purposes. They have been used for the study of important issues in accounting research, such as insolvency forecasting or the choice of accounting methods. Among their advantages, we can highlight the following (Bonsón *et al.*, 1997a):

1. Unlike Discriminant Analysis or Logistic Regression, MLS are nonparametric models which do not require that the feature space meet any property. This is important because, as several authors point out (e. g., Deakin, 1976; Watson, 1990), the distributional properties of financial ratios often lead to violations of the hypotheses of the most popular parametric techniques, and these violations may induce serious biases in the conclusions of the research.
2. Unlike other nonparametric models such as Neural Networks (NN), which are 'black box' devices, they are easy to interpret. The outcome of these systems is a set of rules or a classification tree, which can be understood even by people with no specific Artificial Intelligence (AI) knowledge. So, they are useful tools for the economic analysis, and not only classificatory devices.

In recent years, a wide panoply of rules and trees induction systems has been developed and is now at the researcher's disposal. The main goals of the present research are to test the accuracy of three well known techniques (LDA, See5 and NN) in comparison with that of three newer MLS (ARNI, FAN and SVM) and to measure their sensitivity to variations in sample size. Even though many previous papers have focused on the comparison of classification techniques, our research has several distinctive characteristics, which offer substantial differences to the prior research on this issue. The most important are the following:

1. The three MLS tested (ARNI, FAN and SVM) have never been used for classification tasks in the fields of Accounting and Economics.
2. We focus on business efficiency analysis. This is a task that has not received much attention in the previous literature, as the majority of the papers have dealt with the issue of insolvency forecasting.
3. Our classification problem has a low separability degree. We deliberately do not consider those variables that, at first look,

may seem good predictors for the class indicator, but on closer analysis are redefinitions of the variable to forecast. The inclusion of these ratios may inflate the predictive ability of models by this 'tautological' effect.

4. Unlike other comparative studies, the sensitivity of the accuracy of each technique to variations in the size of the estimation sample is tested. This is important because in accounting research we often have to deal with small data bases or, in case of having a large one, we might be interested in splitting it into smaller subsets (of sector or of company size) in order to increase the validity of the economic analysis.

For the achievement of the aforementioned goals, the remainder of the paper is structured as follows: in section 2, prior research is reviewed, so that the need for the research we propose is shown to be evident. Section 3 comprises the methodology of the study, including the sample selection procedure, the variables and the class indicator, and each of the classification models and the procedure for the measurement of the sensitivity to variations in sample size is briefly described. The main results are shown in section 4, and section 5 is devoted to the summary and conclusions of our research work.

2. PRIOR RESEARCH

Many researchers in the field of modelling economic decisions or phenomena have been interested in comparing the accuracy of different classification techniques[1]. With regard to the rules and trees induction systems, the most tested models have been the Recursive Partitioning Algorithm (RPA), and Quinlan's programs (ID3, C4.5 and See5). More recently, certain developments in the field of Computing as, for example fuzzy sets, rough sets and genetic algorithms have been used in the design of inference engines (see, e.g., Bonsón *et al.*, 1997b;

[1] For a literature review on classification methods see, for example, Zopounidis and Doumpos (2002).

McKee, 1998; Varetto, 1998). The results show, in general, that induction systems are a valid approach to economic classification tasks, even taking into account that sometimes statistical models outperform induction systems in terms of classification accuracy (see, e. g., Marais *et al.*, 1984; Elliott and Kennedy, 1988; McKee, 1995a; McKee, 1995b; Varetto, 1998). The overfitting problems, which are quite common in nonparametric estimation, are in most of the cases the causes of these situations.

However, the bulk of prior research has focused on the comparison of a certain model with the traditional parametric techniques. Only in certain papers two or more of the induction systems that are at the researcher's disposal are compared with each other. Among these exceptions, we can highlight the works of Jeng *et al.* (1997), which consists of the comparison of a fuzzy learning algorithm with Quinlan's ID3; Cronan *et al.* (1991), which compares ID3 and RPA; Didzarevich *et al.* (1997), which tests RPA and CN2, and McKee and Lensberg (2002), which compares rough sets with rough sets in conjunction with genetic algorithms.

Moreover, the majority of the research works has focused on the prediction of insolvency and bankruptcy. Other tasks have received much less attention. The works of Braun and Chandler (1987) for the prediction of the stock market behavior; Liang *et al.* (1992) for the analysis of the FIFO/LIFO decision; Deal and Edgett (1997), on product development decisions; Markham *et al.* (2000), for the setting of the number of kanban cards in a just-in-time production system, and Mak and Munakata (2002), on the product entry decision, are examples.

In addition, the sensitivity of AI techniques to changes in data structure has seldom been analysed by researchers in the field of Economics. We can only highlight the works of Kattan and Cooper (2000), Pavur (2002), and Pendharkar (2002), which conclude that factors like data distribution, class proportions, and the position of outliers have a certain degree of influence upon the accuracy of AI systems. These researchers, however, used data generated by simulations instead of the actual figures from company financial statements.

Furthermore, it is interesting to point out that the accuracy of each system is prone to vary depending on the classification task, the variables, the database, and the inference engine. This evidences the need for a replication of the comparisons when a new task is considered, a new inference engine is used, and different kinds of companies make up the database.

Thus, taking into account the gaps in the prior research, the need for the research that we propose is clearly justified, for several reasons. Firstly, we focus on the issue of business efficiency, to which, despite its importance, not much attention has been paid in the literature. Secondly, our work is not a comparison of a certain induction system with Logit or Discriminant Analysis, but rather a testing of the accuracy of several inference engines. The systems used as a benchmarks are Quinlan's See5, Linear Discriminant Analysis (LDA) and perceptron NN. Finally, we test the effect of sample size variations on the accuracy of each system. This is especially important, because if we were able to develop a model that performs well with small samples, the significance of economic analysis could be enhanced (e. g., by splitting a big sample into branch of activity or/and company size, and analyzing separately each partition).

3. METHODOLOGY

As indicated above, in the present research we focus on an analysis of the performance of several MLS (ARNI, FAN and SVM) in comparison with some well known techniques that we use as benchmarks. In addition, the sensitivity of the accuracy of the systems when sample size varies is tested.

In the following paragraphs we discuss the sample selection process and the variables. A brief description of the techniques is also provided, and the procedure for the measurement of the sensitivity of the accuracy of each system to variations in the size of the estimation sample is expounded.

3.1. The Data Base

For the purpose of this article we start from a data base elaborated from the financial statements of the commercial and industrial firms located in Spain[2]. In accordance with Spanish legislation, limited liability companies are required to deposit their annual accounts in the *Registro Mercantil* (Commercial Register), whose files are publicly available for every user of financial information. The analysed accounts correspond to the year 1999.

We only considered companies with more than 100 employees available in SABE data base[3]. A set of filters were applied to guarantee not only the quality of financial information but also that the selected sample really shows the economic activity of each sector. Companies were eliminated if they did not carry out any activity during 1999, if 1999 was the first year of business, or if they did not offer enough information to compute the selected ratios. Thus, after those eliminations, the remainder of the data base was made up of the accounts of 5671 companies. For these firms we have considered the consolidated statements, when available.

3.2. The Class Inductor

In the present study we focus on the identification of the financial variables which are more related to high levels of efficiency. The level of efficiency is represented using a dichotomous variable, which is equal to one if the company is included in the most efficient group and is equal to zero if it belongs to the least efficient group. Therefore, our classification task is a dichotomous one.

For the measurement of efficiency, and taking into account the limitations of the available information (only annual accounts), we have chosen the financial profitability ratio. This ratio is the quotient between the net profit and equity capital. Many authors (e. g., Kelly and Tippet, 1991; Brief and Lawson, 1992)

[2] Appendix A shows the sectors which have been considered in the study according to the NACE (rev. 1).

[3] *Sistema de Análisis de Balances Españoles* (System for the Analysis of Spanish Balance Sheets) is a financial data base elaborated by Bureau van Dijk with includes the majority of the Spanish commercial and industrial firms.

claim that this is a suitable measure of efficient management, in spite of its limitations.

Once that financial profitability has been computed for each firm, in order to avoid distortions caused by the sector effect, we have divided the ratio by the median of the profitability for each branch of activity.

For the definition of both efficient and inefficient groups, the final specification is made by discarding the eighty intermediate percentiles of the financial profitability ratio. In this way, the group which has the most efficient companies will comprise 10% of the firms with the highest financial profitability and the group which has the least efficient companies will comprise 10% of the firms with the lowest value for this ratio. Appendix A indicates the number of companies included in each group and each sector.

We must state that an alternate version of the present research was carried out considering the 25% most profitable firms and the 25% less profitable companies. The results are not included here because they are close to those obtained with the '10%' criterion. This suggests that other definitions of the class indicator would have not changed the qualitative conclusions of this paper, only showing small differences in the performance of the tested MLS. Since we pursued the 'stylised facts' rather than a very precise estimation of error rates, we have considered our expedient sufficient.

3.3. The Financial Variables

With the aim of describing the financial situation of the firms, we take as reference the set of ratios proposed by López (2000) for the analysis of the financial statements drawn up according to Spanish GAAP. We have excluded from the analysis the variables that, at first sight, may be highly related to the class indicator, but which on closer analysis simply result in redefinitions of the variable to forecast, and thus may inflate the models' predictive ability simply by a 'tautological' effect.

Therefore, ratios which are intrinsically connected with financial profitability are excluded. In this way, our task is a ‘low separability’ problem, for which a high percentage of correct classifications is not expected. The final considered indicators are the set of nine variables that can be seen in appendix B. We must state that in order to guarantee the absence of the aforementioned ‘tautological’ effect, we ran a correlation analysis between each predictor and the financial profitability, with the result of no significant correlations.

For all these variables, as was done for the financial profitability ratio, the distortions caused by the sector effect were corrected by dividing the ratio by the median for each branch of activity.

The descriptive statistical information on each variable, displayed on appendix C, shows a marked positive skewness and a high level of kurtosis that characterizes the distribution of most of the variables. The Lilliefors test was carried out resulting in the rejection of the normality assumption for every indicator. In addition, we must state that all the variables make up a relatively high dimensional feature space, with almost uncorrelated components, which makes dimension reduction strategies based on some kind of principal components analysis unfeasible.

3.4. The Tested Techniques

As stated above, in the present research we test the accuracy of several rules and trees induction systems. Three of them (LDA, See5 and NN) are world-class standards, which we use as benchmarks. The other three (ARNI, FAN and SVM) are newer models that have not been previously employed for economic classification tasks. In the following lines a brief description of each of the tested systems is provided.

3.4.1. LDA

LDA classifies using a function which takes the form $Z=v_0+v_1x_1+ \dots +v_nx_n$, where $x_1 \dots x_n$ are the formally independent variables and $v_0 \dots v_n$ the discriminant coefficients, computed through a differential calculus procedure (see Jobson, 1992).

Individuals are assigned to either one or the other group depending on their estimated Z values. LDA procedure assumes that formally independent variables are multivariate normally distributed and that the group dispersion matrices are equal across all groups. This can lead to non-optimum results, as violation of these assumptions is the rule rather than the exception, at least in economics and finance (see Eisenbeis, 1977).

3.4.2. Quinlan's See5

This algorithm is the latest version of the induction systems developed by Quinlan (1979, 1983, 1986, 1988, 1993 and 2000). These systems use the entropy criterion, which means that the classification tree grows, if we choose, at each step, the variable which has the highest entropy or amount of information. Entropy is calculated through the following expression:

$$Entropy = -\sum_{j=1}^{k} \frac{n_j}{N} \times Log_2\left(\frac{n_j}{N}\right)$$

where N is the total number of observations, k the number of classes and n_j is the number of observations belonging to each class. See5 algorithm uses a more sophisticated version of this criterion, and includes additional functions, the most important being the possibility of changing the obtained tree into a simpler set of classification rules.

Quinlan's programs are recognized worldwide as a standard in classification. So, there are a lot of research papers in Accounting and Finance on the topic of the application of Quinlan's induction systems. Most of them use the previous versions of See5 (ID3 and C4.5) and compare their accuracy with those of parametric statistical techniques. The results show, in general, that Quinlan's systems are a valid approach to economic classification tasks[4].

[4] For a literature review on this issue see, for example, De Andrés (2001).

In this paper, we have followed two steps in the application of See5. First, a classification tree was inferred from the original data and, second, the tree was simplified into a set of simpler rules which have the following structure:

If <condition> then assign the case to class <x>

where the conditions are logical expressions which involve the input variables of the model.

For the generation of the rule set See5 considers all the possible rules that can be constructed with the nodes and the variables of the tree. These rules are then ordered according to a measure of their classification performance and the best ones are selected. The algorithm selects also a classification by default to assign to the cases which do not satisfy the conditions of any rules. This class by default will be calculated so that classification mistakes are minimum. The software used to develop these models is SEE5 by RULEQUEST, Inc.

3.4.3. ARNI

ARNI (Ranilla *et al.*, 1999) is a MLS which constructs decision trees following the procedure used by ID3 and C4.5 (Quinlan, 1993) but using a measure of the quality of the rules which is called the *Impurity Level* (IL) as heuristic instead of entropy. The process of pruning is also replaced by the one used in FAN.

The resulting algorithm usually shows, when applied to benchmark tasks, an accuracy level slightly better than C4.5, but produces a number of rules that is considerably smaller.

There exists also a version of the algorithm that produces classification rules and that is called ARNI-rules. This system operates in a way that is analog to See5-rules. The results of this procedure when applied to our data base are also discussed in section 4.

3.4.4. FAN

FAN is the acronym of *Finding Accurate Inductions*, a machine learning system which combines the advantages of both instance-based algorithms and rule inducers. It has been developed by Ranilla and Bahamonde (2000) and it has been shown to produce, in exemplary data bases, fewer rules than C4.5 with no less accuracy than other algorithms of the same type, like RISE (Domingos, 1996).

When presented with a collection of already classified examples, FAN produces successive sets of rules which then enter a process of pruning. The main tool used in this process is the IL. This is a heurisitic estimation of the classification performance of the induced rules and is applied to select the reduced set of rules which better generalize the given examples.

To classify an unseen case, the distance of the new example to all the induced rules is computed, and the closest one is then applied to conclude the predicted category.

3.4.5. SVM

Support Vector Machines (SVM) are universal learners based on the *Structural Risk Minimization* principle from computational learning theory (Vapnik, 1995). They are able to find out linear or non-linear threshold functions to separate the examples of a certain category from the rest by means of *support vectors*.

One of the most important properties of SVM is that they can deal with very large feature spaces, independently of their dimensionality.

3.4.6. Perceptron NN

NN are algorithms inspired on the structure and behavior of neurons in the human brain[5]. They can be used to recognize and categorize patterns of data. A NN is formed of individual neurons which are connected one to the others. Each neuron receives input from other neurons, processes these signals and sends an output to other neurons. Each connection has a different weight associated to it and each neuron has a function (usually sigmoid or threshold functions) which, together, determine the response to the input signals. These weights and functions define the function of the NN and are iteratively modified during the process of training (in which the classified examples are presented to the network).

[5] For more details see, for example, Bishop (1995) and Haykin (1999).

The neurons of the NN are organized into layers. The structure and number of these layers define the different models of NN. One of the most common (and the one we use in this study) is the Multilayer Perceptron trained with the backpropagation algorithm. This system has an input layer that receives the external data, one or more hidden layers and one output layer which gives the result of the processing. In the present reseach we have used only one hidden layer, as previous papers (i.e. Altman *et al.*, 1994) suggest that additional layers do not increase the explanatory power of the models.

3.5. The Measurement of the Sensitivity of Each Technique to Changes in Sample Size

We have evaluated the variations of each model's performance as the size of the estimation sample increases. For each sample size, and for each technique, not only one sample is used. A number of them are randomly selected from the data base, and therefore a number of models are estimated, so average error rates can be computed.

Due to data availability restrictions, we have chosen 9 different sizes for the estimation sample (*n*=100, 200, 300, 400, 500, 600, 700, 800, 900). For each sample size *n*, the following two-steps procedure is repeated 50 times: first, from the data base are randomly selected a training set of size *n*. And, second, once the model is estimated from the training set, its performance is evaluated at an independent test set. Once the 50 iterations have been completed, average error rates on the test sample are computed.

Regarding the independent test set, its size is 134, as the original data base contains 1134 examples and 1000 of them are used for the selection of the training sets. We decided to use as estimation set, along all the evaluation process, a fixed separate partition of the original sample in order to guarantee that all the training sets are strictly independent from the test set used.

The whole process is designed with the purposes of (1) neutralizing as far as possible the potential effects of excessive data mining on one single test set, and (2) averaging the performances of many models, with a view to neutralizing the effect of random variability as well as local minima in training processes. Analysis of average results for a reasonably high number of replications seems to us more adequate for a comparative analysis than the study of one single estimation task, or only a best or worse case analysis.

4. RESULTS

Tables 1 and 2 show the percentage error rates in the test samples for the selected estimation sample sizes. We call 'type I error' the one which consists of assigning a high efficiency company to the low efficiency group. Therefore, 'type II error' is classifying a low efficiency firm in the high efficiency group.

Size	TOTAL ERROR							
	LDA	Arni	Arni rules	See5	See5 rules	FAN	SVM	NN
100	42,30	30,27	31,00	29,43	30,12	57,27	49,52	38,81
200	41,57	26,43	26,85	26,52	26,64	41,49	49,01	33,12
300	41,13	24,96	25,75	25,40	25,19	40,91	49,33	30,75
400	40,19	24,49	24,30	24,00	24,21	42,01	49,16	28,52
500	39,87	23,22	23,78	23,06	23,30	41,94	48,64	27,99
600	39,70	23,93	23,67	21,16	21,52	42,48	48,81	27,57
700	38,67	23,24	22,79	21,51	21,07	42,48	48,34	27,66
800	39,13	22,49	22,18	20,73	20,30	42,88	47,57	25,51
900	38,01	22,54	22,88	20,87	20,61	42,97	46,94	25,73

Table 1. Total percentage of errors in the test sample

Size	TYPE I ERROR								TYPE II ERROR							
	LDA	Arni	Arni rules	See5	See5 rules	FAN	SVM	NN	LDA	Arni	Arni rules	See5	See5 rules	FAN	SVM	NN
100	41,83	30,33	31,37	30,75	30,78	42,21	56,75	39,20	42,72	30,21	30,63	28,12	29,46	43,25	42,3	38,38
200	40,86	26,09	26,42	24,84	25,67	38,09	49,1	32,88	42,17	26,78	27,28	28,21	27,61	44,9	48,93	33,35
300	40,42	26,12	24,81	25,4	25,7	39,43	49,85	30,12	41,75	23,79	26,69	25,4	24,69	42,39	48,81	31,34
400	38,71	24,48	23,43	24,33	25,31	41,34	57,04	28,61	41,33	24,51	25,16	23,67	23,1	42,69	41,28	28,44
500	37,98	24,09	23,85	22,93	23,91	42,51	60,57	27,51	41,24	22,36	23,7	23,19	22,69	41,37	36,72	28,44
600	37,78	24,99	24,18	21,01	21,4	43,34	50,09	26,77	41,10	22,87	23,16	21,31	21,64	41,61	47,52	28,32
700	36,49	22,9	21,64	20,99	20,6	40,06	46,96	26,85	40,25	23,58	23,94	22,03	21,55	44,9	49,73	28,41
800	37,22	22,2	20,69	20,15	19,19	40,84	47,49	25,02	40,55	22,87	23,67	21,31	21,4	44,93	47,64	25,97
900	35,74	22,42	23,01	20,21	20,66	37,1	44,72	24,97	39,67	22,66	22,75	21,52	20,57	48,84	49,16	26,45

Table 2. Percentage of type I errors and percentage of type II errors in the test sample

Firstly, we can appreciate the relatively high error rates, which confirm that this is a low separability problem. This is the result of the variables selection process. It must be remembered that we have excluded from the analysis the variables that, at first sight, may be highly related to the class indicator.

Another interesting conclusion for all techniques is the reduction of total errors when the sample size grows (this reduction is bigger in See5 rules). This implies that, if possible, it is better to work with big sample sizes but, any case, it is necessary to carry out a cost benefit analysis on the use of large databases.

It is clearly shown that SVM, FAN and LDA have higher total percentages of error in the test sample for all selected sample sizes than ARNI, See5 and NN. The advantage of See5 and See5-rules over ARNI and ARNI-rules when the sample size increases is also prominent. These results can be seen more clearly in Figures 1, 2 and 3.

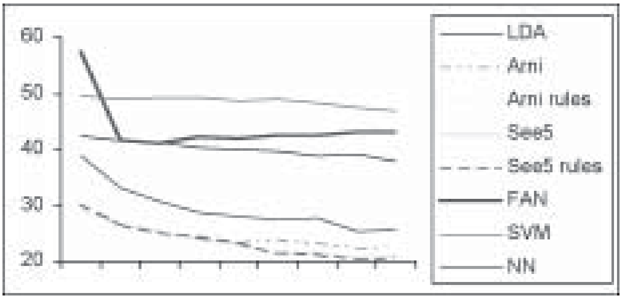


Figure 1. Total percentage of errors in the test (estimation) sample

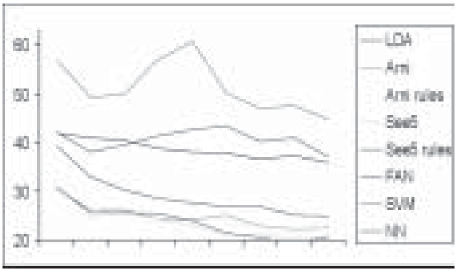


Figure 2. Percentage of type I errors in the test (estimation) sample

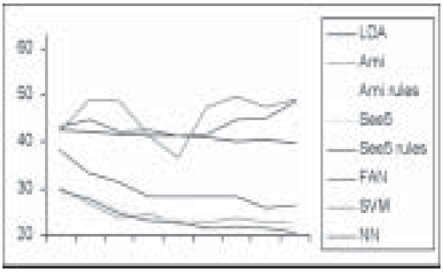


Figure 3. Percentage of type II errors in the test (estimation) sample

In addition, it must be highlighted that these results were tested using the Mann Withney statistic (detailed results are not shown due to space restrictions), so we are able to corroborate the previous statements.

For the percentages of type I and type II errors the results are similar, that is, FAN, SVM and LDA present higher values in all the sample sizes. It is interesting to emphasize that while for the percentage of type I errors in FAN and SVM the values decrease with the sample size, for percentage of type II errors values do not decrease when sample size increases. Another noticeable result is the higher value of percentage type II errors in relation to type I in the FAN technique for all sample sizes. In the other techniques values are very similar.

With the aim of examining the interpretability of the results of each system, it is interesting to comment on the number of nodes, rules or vectors that are generated (see Table 3)[6]. For all the techniques, this number increases with sample size, except for FAN, because in this model the rules do not cover the whole space, since they are applied using a minimum distance criterion. According to table 3, FAN is the easiest to interpr*et al*gorithm, although it has a high percentage of total errors.

In addition, although according to the theoretical specifications ARNI-rules should produce less rules than See5-rules, in our case the results are the opposite for all the sample sizes. The reasons for this are surely the specific features of accounting ratios (positive skewness, high levels of kurtosis, unequality of dispersion matrices), as in other environments rulesets inferred with ARNI are smaller than those generated using See5 (see Ranilla *et al.*, 1999).On the other hand, and regarding to the classification trees, ARNI employs less nodes than See5 for the majority of the sample sizes.

[6] LDA and NN are not included in table 3 because these systems have a fixed topology (nine discriminant coefficients for LDA and three layers for perceptron NNs).

Size	Arni[1]	Arni rules[2]	See5[1]	See5 rules[2]	FAN[2]	SVM[3]
100	6,20	21,04	6,22	16,40	4,70	89,84
200	8,24	35,32	7,74	24,16	5,32	187,82
300	9,60	49,12	9,76	34,24	4,88	286,18
400	10,18	61,60	10,82	39,24	4,18	385,46
500	9,92	71,84	11,66	48,24	3,58	482,42
600	10,08	85,88	13,28	54,56	3,38	580,58
700	10,82	100,16	13,52	60,60	2,90	680,70
800	10,38	109,12	14,68	63,04	2,72	781,50
900	10,40	117,24	14,82	73,24	2,34	881,74

[1] Nodes [2] Classification rules [3] Vectors

Table 3. Average number of nodes, rules or vectors.

5. RESEARCH LIMITATIONS

As defined above, this research has certain limitations that are either impossible or not cost-effective to overcome. These limitations must be stated in order to allow the reader to achieve a clearer understanding of the results. First of all, the decision of including medium-sized companies in the sample implies assuming the risk that a certain number of firms could be in the highest quartile due to 'cooking the books', as according to Spanish legislation small and medium-sized businesses do not have the obligation to submit their annual accounts to the auditor's judgment. However, it could be argued that both efficient and inefficient companies have incentives to carry out creative accounting practices.

Another drawback is that in nonparametric contexts the set of possible probabilistic behaviour is by nature much wider. The most habitual strategy is to focus on the analysis of more or less exemplary cases. Because of this, the validity of our analysis is, strictly speaking, limited to our case. For a different population other conclusions might be obtained. However, there are some reasons that suggest that the kind of analysis we perform here may be more illuminating than a priori expected: (1) Available literature (e. g. De Andrés, 2000) suggests that the Spanish case may be representative of most European countries, and may be used as a satisfactory benchmark. (2) The results of the study are interesting in their own right for Spanish researchers in the field of efficiency. (3) As indicated before, prior research comparing classificatory devices has never focused on efficiency or on the analysis of ARNI, FAN, and SVM systems. And, of course, (4) partially valid conclusions often seem better than a complete absence of knowledge.

Finally, we must state that the misclassification costs have not been considered in the present study. The main reason for this is that the two newer MLS (ARNI and FAN) do not have the feature to deal with dissimilar misclassification costs. In addition, the estimation could be rather an arbitrary and unreliable process, having into account that these costs are sector-specific, and our methodology considers all branches of activity as a whole. Nevertheless, as previous research clearly demonstrates (De Andrés, 2001), in the analysis of business efficiency through MLS, misclassification costs are very similar for each type of error.

6. SUMMARY AND CONCLUSIONS

Rules and trees induction systems play an increasing role in economic research. So, it is worth determining which of them achieves better results. In this paper we have compared the performance of three well known models (LDA, Quinlan's See5 and NN) with three newer inference engines (ARNI, FAN and SVM). In addition, we have tried to measure the sensitivity of each system to changes in sample size.

Although there are other comparisons in the literature, the classification task we have chosen, that is, the analysis of the most/least efficient companies, the rule induction system we have tested, and the measure of the sensitivity to changes in sample size we have carried out, are significant features that distinguish this work from previous research papers.

The application of MLS techniques to the analysis of business efficiency has corroborated the low separability of this task and, for this reason, high error percentages are present in almost every case. The comparison of the techniques is unfavourable in respect of SVM, FAN and LDA. Another noticeable finding is that ARNI, See5 and NN yield the best results and See5 is the best for big sample sizes. These results are caused by the low separability of this problem, where the two classes intersect in wide regions of the feature space, and therefore nonlinear methods such as SEE5, ARNI or NN can capture the features defining each class in a better way than those operating through linear partitions (LDA, FAN and SVM).

Regarding type I and type II errors, FAN, SVM and LDA present higher values in all the sample sizes. Another noticeable result is that while for FAN and SVM type I error is decreasing with the sample size, for type II errors values do not decrease when sample size increases. Therefore, the reduction in total error is due to the decrease of the type I error.

The number of generated nodes, rules and vectors, as indicator of the ease of interpretation of the results, shows an increase as we increase the sample size. That is to say, as we increase the sample size we manage to decrease the total error, but the interpretation becomes more difficult. It is necessary to carry out a cost-benefit analysis on the use of large databases.

One final consideration is that the results that we have obtained are only for this database and we cannot be sure that they will be seen to be the same if we use other databases. So, this indicates the area of future research. Other lines of investigation that we propose for further studies are: other classification problems (financial distress prediction, analysis of management decisions, etc.), more than two classes in the classification problem and the use of other techniques (genetic algorithms, mathematical programming, etc.).

7. REFERENCES

ALTMAN, E.I.; MARCO, G.; VARETTO, F. (1994): “Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience)”, *Journal of Banking and Finance*, n.18:505-529.

BISHOP, C. (1995): *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK.

BONSÓN, E.; ESCOBAR, T.; MARTÍN, M.P. (1997a): “Decision tree induction systems. Applications in Accounting and Finance”, in *Intelligent Technologies in Accounting and Business*, edited by G. Sierra and E. Bonsón. Huelva. Spain.

BONSÓN, E.; ESCOBAR, T.; MARTÍN, M.P. (1997b): *Sistemas de inducción de árboles de decisión: utilidad en el análisis de crisis bancarias*. http://ciberconta.unizar.es/ Biblioteca/Biblioteca.html.

BRAUN, H.; CHANDLER, J. S. (1987): “Predicting stock market behavior through rule induction: an application of the learning-from-examples approach”, *Decision Sciences*, Summer:415-429.

BRIEF, R.P.; LAWSON, R.A. (1992): “The role of the accounting rate of return in financial statement analysis”, *Accounting Review*, vol. 67, n.2:411-426.

CRONAN, T.P.; GLORFELD, L.W.; PERRY, L.G. (1991): “Production system development for expert systems using a recursive partitioning induction approach: an application to mortgage, commercial and consumer lending”, *Decision Sciences*, vol. 22, n. 4:812-845.

DE ANDRÉS, J. (2000): “Los parámetros característicos de las empresas manufactureras de alta rentabilidad. Una aplicación del análisis discriminante”, *Revista Española de Financiación y Contabilidad*, n. 104:443-482.

DE ANDRÉS, J. (2001): “Statistical techniques vs. SEE5 algorithm. An application to a small business environment”, *International Journal of Digital Accounting Research*, vol. 1, n. 2:157-184.

DEAKIN, E.B. (1976): “Distribution of financial accounting ratios: some empirical evidence”, *Accounting Review*, vol. 51, n. 1:90-96.

DEAL, K.; EDGETT, S.J. (1997):“Determining success criteria for financial products: A comparative analysis of CART, logit and factor/discriminant analysis”, *Service Industries Journal*, vol. 17, n. 3:489-506.

DIDZAREVICH, S.; LIZARRAGA, F.; LARRAÑAGA, P.; SIERRA, B.; GALLEGO, M.J. (1997): "Statistical and machine learning methods in the prediction of bankruptcy", in *Intelligent Technologies in Accounting and Business*, edited by G. Sierra and E. Bonsón. Huelva, Spain.

DOMINGOS, P. (1996): "Unifying instance-based and rule-based induction", *Machine Learning*, n. 24:141-168.

EISENBEIS, RA. (1977): "Pitfalls in the application of discriminant analysis in business, finance, and economics", *Journal of Finance*, vol. 32, n. 3:875-900.

ELLIOTT, J.A.; KENNEDY, D.B. (1988): "Estimation and prediction of categorical models in Accounting research", *Journal of Accounting Literature*, n. 7:202-242.

HAYKIN, S. (1999):*Neural Networks - A Comprehensive Foundation (2nd Ed.)*, Prentice Hall, New York, USA.

JENG, B.; JENG, Y.M.; LIANG, T.P. (1997): "FILM: A fuzzy inductive learning method for automated knowledge acquisition", *Decision Support Systems*, vol. 21, n. 2:61-73.

JOBSON, JD. (1992): *Applied multivariate data analysis*, Springer-Verlag. New York, USA.

KATTAN, M.W.; COOPER, R.B. (2000): "A simulation of factors affecting machine learning techniques: an examination of partitioning and class proportions", *Omega*, n. 28:501-512.

KELLY, G.; TIPPET, M. (1991): "Economic and accounting rates of return: a statistical model", *Accounting & Business Research*, vol. 21, n. 84:321-329.

LIANG, T.P.; CHANDLER, J.S.; HAN, I.; ROAN, J. (1992): "An empirical investigation of some data effects on the classification accuracy of probit, ID3 and neural networks", *Contemporary Accounting Research*, vol. 9 (Fall):306-328.

LÓPEZ, A. (2000): *Análisis económico-financiero de las empresas de Asturias por sectores de actividad 1994-1995*. Principado de Asturias. Oviedo. Spain.

MAK, B.; MUNAKATA, T. (2002): "Rule extraction from expert heuristics: a comparative study of rough sets with neural networks and ID3", *European Journal of Operational Research*, n. 136:212-229.

MCKEE, T.E. (1995a):“Predicting bankruptcy via induction”, *Journal of Information Technology*, vol. 10:26-36.

MCKEE, T.E. (1995b): “Predicting bankruptcy via an inductive inference algorithm: an extension”, in *Artificial Intelligence in Accounting, Finance and Tax*, edited by G. Sierra and E. Bonsón. Huelva. Spain.

MCKEE, T.E. (1998): “A mathematically derived rough set model for bankruptcy prediction”, in *Emerging Technologies in Accounting and Finance*, edited by E. Bonsón and M. Vasarhelyi. Huelva. Spain.

MCKEE, T.E.; LENSBERG, T. (2002): “Genetic programming and rough sets: A hybrid approach to bankruptcy classification”, *European Journal of Operational Research*, n. 138:436-451.

MARAIS, M.L.; PATELL, J.M.; WOLFSON, M.A. (1984): “The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications”, *Journal of Accounting Research*, vol. 22 (supplement):87-114.

MARKHAM, I.S.; MATHIEU, R.G.; WRAY, B.A. (2000): “Kanban setting through artificial intelligence: a comparative study of artificial neural networks and decision trees”, *Integrated Manufacturing Systems*, vol. 11, n.4:239-259.

PAVUR, R. (2002): “A comparative study of the effect of the position of outliers on classical and nontraditional approaches to the two group classification problem”, *European Journal of Operational Research*, n. 136:603-615.

PENDHARKAR, P.C. (2002):“A computational study on the performance of artificial neural networks under changing structural design and data distribution”, *European Journal of Operational Research*, n. 138:155-177.

QUINLAN, J.R. (1979): “Discovering rules by induction from large collections of examples”, in *Expert systems in the microelectronic age*, edited by D. Michie.

QUINLAN, J.R. (1983): “Learning efficient classification procedures”, in *Machine learning: an Artificial Intelligence approach*. Tioga Press. Palo Alto. USA.

QUINLAN, J.R. (1986): “Induction of decision trees”, *Machine Learning*, vol. 1, n. 1:81-106.

QUINLAN, J.R. (1988): “Decision trees and multivalued attributes”, *Machine Intelligence*, n. 11:305-318.

QUINLAN, J.R. (1993): *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, California.

QUINLAN, J.R. (2000): *See5*. http://www.rulequest.com/See5-info.html.

RANILLA, J.; GARCÍA, M.; BAHAMONDE, A. (1999): "Construcción de árboles de decisión usando el Nivel de Impureza. CAEPIA'99-TTIA'99", *VIII Conferencia de la Asociación Española para la Inteligencia Artificial - III Jornadas de Transferencia Tecnológica de Inteligencia Artificial* - Libro de Actas - Vol. I:34-41, Murcia, Spain.

RANILLA, J.; BAHAMONDE, A. (2000): *FAN: Finding Accurate Inductions*. Technical report. Centro de Inteligencia Artificial de la Universidad de Oviedo.

VARETTO, F. (1998): "Genetic algorithms applications in the analysis of insolvency risk", *Journal of Banking and Finance*, n. 22:1421-1439.

VAPNIK, V. (1995): *The Nature of Statistical Learning Theory*, Springer, New York.

WATSON, C.J. (1990): "Multivariate distributional properties, outliers, and transformation of financial ratios", *Accounting Review*, vol. 65, n. 3:662-695.

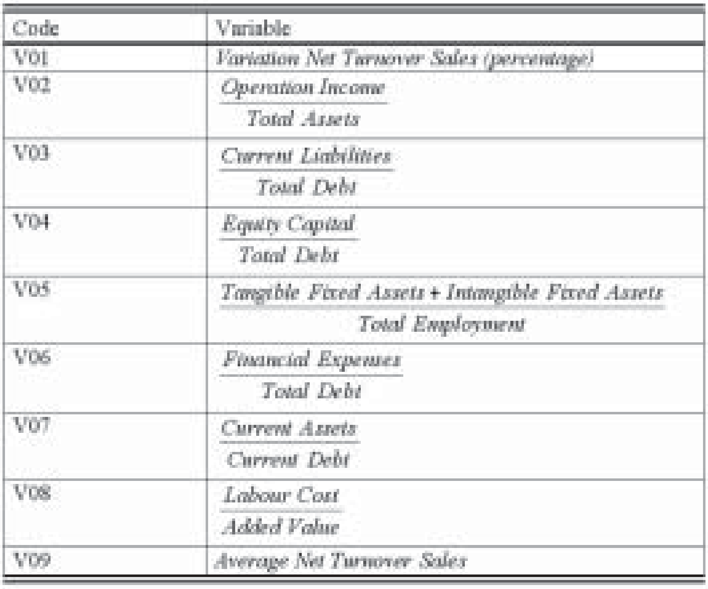
ZOPOUNIDIS, C.; DOUMPOS, M. (2002): "Multicriteria classification and sorting methods: a literature review", *European Journal of Operational Resarch*, n. 138:229-246.

APPENDIX A: COMPANIES IN THE DATA BASE DETAILED BY BRANCH OF ACTIVITY

Nº	Name	Low effic.	High effic.	Total
01	Agriculture, hunting and related service activities	18	16	34
02	Forestry, logging and related service activities	1	0	1
05	Fishing, operation of fish hatcheries and fish farms;			
	service activities incidental to fishing	5	5	10
10	Mining of coal and lignite; extraction of peat	4	2	6
11	Extraction of crude petroleum and natural gas; service activities incidental to oil and gas extraction, excluding surveying	1	1	2
13	Mining of metal ores	1	2	3
14	Other mining and quarrying	1	2	3
15	Manufacture of food products and beverages	26	29	55
17	Manufacture of textiles	15	13	28
18	Manufacture of leather clothes	4	4	8
19	Tanning and dressing of leather	3	4	7
20	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials	1	1	2
21	Manufacture of pulp, paper and paper products	10	9	19
22	Publishing, printing and reproduction of recorded media	12	16	28
24	Manufacture of chemicals and chemical products	24	17	41
25	Manufacture of rubber and plastic products	8	11	19
26	Manufacture of other non-metallic mineral products	21	7	28
27	Manufacture of basic metals	10	4	14
28	Manufacture of fabricated metal products, except machinery and equipment	13	8	21
29	Manufacture of machinery and equipment n.e.c.	7	10	17
30	Manufacture of office machinery and computers	0	1	1
31	Manufacture of electrical machinery and apparatus n.e.c.	8	6	14
32	Manufacture of radio, television and communication equipment and apparatus	5	0	5
33	Manufacture of medical, precision and optical instruments,			
	watches and clocks	5	1	6
34	Manufacture of motor vehicles, trailers and semi-trailers	18	9	27
35	Manufacture of other transport equipment	5	4	9
36	Manufacture of furniture; manufacturing n.e.c.	6	0	6
37	Recycling	2	1	3
40	Electricity, gas, steam and hot-water supply	1	0	1
41	Collection, purification and distribution of water	3	1	4
45	Construction	34	59	93
50	Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel	4	8	12
51	Wholesale trade and commission trade, except of motor vehicles and motorcycles	56	68	124
52	Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods	26	22	48
55	Hotels and restaurants	27	32	59
60	Land transport; transport via pipelines	26	17	43
61	Water transport	2	2	4

62	Air transport	1	1	2
63	Supporting and auxiliary transport activities; activities of travel agencies	13	9	22
64	Post and telecommunications	8	1	9
65	Financial intermediation, except insurance and pension funding	2	1	3
67	Activities auxiliary to financial intermediation	1	4	5
70	Real estate activities	5	10	15
71	Renting of machinery and equipment without operator and of personal and household goods	1	0	1
72	Computer and related activities	9	9	18
73	Research and development	0	1	1
74	Other business activities	67	79	146
75	Public administration and defence; compulsory social security	1	1	2
80	Education	3	10	13
85	Health and social work	16	16	32
90	Sewage and refuse disposal, sanitation and similar activities	3	5	8
92	Recreational, cultural and sporting activities	19	24	43
93	Other service activities	5	4	9
	TOTAL	**567**	**567**	**1134**

APPENDIX B: FINANCIAL VARIABLES



Code	Variable
V01	*Variation Net Turnover Sales (percentage)*
V02	$\frac{\textit{Operation Income}}{\textit{Total Assets}}$
V03	$\frac{\textit{Current Liabilities}}{\textit{Total Debt}}$
V04	$\frac{\textit{Equity Capital}}{\textit{Total Debt}}$
V05	$\frac{\textit{Tangible Fixed Assets} + \textit{Intangible Fixed Assets}}{\textit{Total Employment}}$
V06	$\frac{\textit{Financial Expenses}}{\textit{Total Debt}}$
V07	$\frac{\textit{Current Assets}}{\textit{Current Debt}}$
V08	$\frac{\textit{Labour Cost}}{\textit{Added Value}}$
V09	*Average Net Turnover Sales*

APPENDIX C: DESCRIPTIVE STATISTICAL INFORMATION ABOUT THE FINANCIAL VARIABLES

VAR	Low efficiency						
	1 quart.	Med.	3 quart.	Mean	Std dev	Skewness	Kurtosis
V01	-0,650	0,645	2,168	32,710	421,555	16,921	314,462
V02	0,603	0,914	1,349	1,103	0,884	3,302	17,823
V03	0,706	0,973	1,053	0,878	0,288	-0,617	0,517
V04	0,259	0,603	1,027	0,675	0,766	-3,495	43,623
V05	0,445	1,029	2,220	4,156	36,795	22,813	534,385
V06	0,610	1,180	1,830	1,440	1,320	2,460	9,746
V07	0,639	0,837	1,079	0,974	0,686	4,547	35,841
V08	1,063	1,233	1,528	1,488	6,267	-4,405	158,716
V09	0,380	0,790	2,190	2,660	6,290	6,418	55,597

VAR	High efficiency						
	1 quart.	Med.	3 quart.	Mean	Std dev	Skewness	Kurtosis
V01	0,298	1,539	4,029	10,859	120,926	22,628	528,126
V02	0,890	1,308	1,942	1,573	1,174	3,684	26,586
V03	0,829	1,036	1,069	0,946	0,260	-0,844	1,069
V04	0,185	0,545	1,000	0,510	1,247	-9,986	170,038
V05	0,207	0,623	1,444	2,773	19,164	17,220	329,088
V06	0,330	0,930	1,530	1,280	1,850	8,875	134,430
V07	0,676	0,895	1,143	0,995	0,636	5,279	60,652
V08	0,727	0,956	1,127	0,622	6,865	-19,769	432,711
V09	0,390	0,970	2,530	6,500	61,870	18,682	380,890