The International Journal of Digital Accounting Research
Vol. 1, No. 2, pp. 153-179
ISSN: 1577-8517

Statistical Techniques vs. SEE5 Algorithm. An Application to a Small Business Environment

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

Abstract. The aim of this research is to compare the accuracy of a rule induction classifier system –Quinlan's SEE5– with linear discriminant analysis and logit.

The classification task chosen is the differentiation of the most efficient companies from the least efficient ones on the basis of a set of financial variables.

The sample consists of a database containing the annual accounts of the companies located in the Principality of Asturias (Spain), which are mainly small businesses.

The main results indicate that SEE5 outperforms logit, but it is not clearly better than discriminant analysis. However, SEE5 models suffer from bigger increases in error rates when tested with validation samples.

Another interesting finding is that in SEE5 systems both the number of variables selected and the number of rules inferred grow when sample size increases.

Keywords: *Discriminant analysis, logit, SEE5, efficiency, small companies.*

1. INTRODUCTION

Classification techniques hold a very important place in the field of accounting research, as many research problems involve classification of observations into discrete categories. As Elliott and Kennedy (1988, p. 292) point out, these problems include: (1) management's choice of accounting methods, (2) bank loan classification, (3) bankruptcy prediction, (4) prediction of bond ratings, (5) lobbying positions on accounting standards, and (6) prediction of takeover targets.

So, it is a matter of particular interest to determine which classification techniques better suit the special characteristics of accounting data. The most popular classification models are the statistical ones. However, these systems are often based on certain distributional hypotheses that accounting data do not always fit.

In an attempt to overcome these problems, a number of nonparametrical techniques have been developed. Most of them belong to the field of Artificial Intelligence, for example neural networks or rule induction systems. Many papers have dealt with the issue of comparing the results of the traditional statistical techniques with one or more of the new models.

The research we present here can be included in this research line, as we compare the accuracy of two well-known techniques, linear discriminant analysis (LDA) and logistic regression or logit, with a recent development in the field of rule induction systems, the SEE5 algorithm (Quinlan, 1997).

However, our work has several features that make it different from the previous research. First of all, we test the techniques on data from small companies. The vast majority of the previous papers have used accounts of listed companies, probably due to data availability problems.

Second, we have considered a classification problem that former works have seldom paid attention to. The task is the differentiation of the most efficient companies from the least efficient ones on the basis of a set of financial variables. As we will see, most of the previous research has focused on financial distress prediction or on accounting choice problems.

Third, we have split the sample and replicated the study for each one of the branches of economic activity. As each sector has a different number of companies, we have tested how the accuracy of the models is affected by the size of the database used in their estimation.

And, finally, we have chosen the SEE5 algorithm for the inference of the classification rules. Many of the previous works that compare rule induction systems with statistical techniques use early versions of Quinlan's programs (i. e. ID3). These systems underestimate the capabilities of the rules approach, because they lack the possibility of pruning the model, which is a key feature in order to avoid overfitting problems.

The remainder of the paper is structured as follows: in section 2 prior research on the comparison of classification techniques is briefly reviewed. Section 3 is dedicated to the description of the methodology of our study. Section 4 contains the major findings of the research and, finally, in section 5 the main conclusions of the paper are detailed.

2. PRIOR RESEARCH

As stated before, many researchers have been interested in comparing the accuracy of statistical classification techniques with that of the systems developed in the field of Artificial Intelligence.

Within these models, the programs for the inference of decision rules or trees are of particular interest, because the rules or trees can be easily understood and interpreted by human analysts. Other approaches, like, for example, neural networks, are less useful as they are 'black box' models. Therefore, they can only be used as classification devices but not for economic analysis.

The most popular rules and tree induction systems are the Recursive Partitioning Algorithm (RPA), based on theoretical analyses by Friedman (1977) and Gordon and Olshen (1978); and Quinlan's programs ID3, C4.5 and SEE5 (Quinlan, 1979, 1983, 1986, 1988, 1993 and 1997).

More recently, certain developments in the field of Computing have been used in the design of inference engines. Fuzzy sets, rough sets, genetic algorithms and the hybridization of Quinlan's programs with statistical techniques (i.e. NEWQ algorithm) have proven to be valid approaches to the construction of machine learning systems.

Table 1 shows a summary of the most important studies on the comparison of inductive systems with statistical techniques when applied to classification tasks. The task, the tested systems, and the main results are indicated (LDA=Linear

Discriminant Analysis; QDA=Quadratic Discriminant Analysis; NN=Neural Networks).

Author(s)	Task	Techniques	Main results
Marais *et al.* (1984)	Modeling commercial bank loan classifications.	• Probit. • RPA.	RPA is not significantly better, especially when the data do not include nominal variables.
Frydman *et al.* (1985)	Predicting financial distress.	• LDA. • RPA.	Less complex RPA models perform better than DA in terms of actual cross-validated and bootstrapped results.
Braund and Chandler (1987)	Predicting stock market behavior.	• LDA. • ID3.	ID3 achieves better results than discriminant analysis.
Messier and Hansen (1988)	Predicting loan defaults and bankruptcies.	• LDA. • ID3.	ID3 outperforms LDA. Attribute sets included in discriminant models intersect only partially those induced through ID3.
Elliott and Kennedy (1988)	Modeling accounting strategies	• LDA. • QDA. • Logit. • Probit. • RPA.	In the analysis of classification problems with more than two categories, all statistical techniques, with the exception of QDA, are better than RPA.
Garrison and Michaelsen (1989)	Tax decisions.	• LDA. • ID3. • Probit.	ID3 performs better than both linear discriminant analysis and probit.
Parker and Abramowicz (1989)	Tax decisions.	• LDA. • ID3. • Probit.	ID3 performs better than both linear discriminant analysis and probit.
Cronan *et al.* (1991)	Assessing mortgage, commercial, and consumer lending.	• LDA. • Logit. • Probit. • RPA. • ID3.	RPA, which used few variables, provided notably higher accuracy than ID3, which used many variables. RPA also outperformed statistical techniques.
Liang *et al.* (1992)	Modeling FIFO/LIFO decision.	• Probit. • ID3. • NN.	Predictive accuracy of ID3 is lower than that of probit or NN. However, ID3 is less sensitive to reductions in sample size.
Tam and Kiang (1992)	Bank failure prediction.	• LDA • Logit. • K-nearest neighbour. • ID3. • NN.	ID3 performs worse than the statistical techniques and neural networks, but better than k-nearest neighbour. The most accurate procedure is the back propagation neural network.
Hansen *et al.* (1993)	Modeling auditors' going concern decision.	• Logit. • ID3. • NEWQ algorithm.	NEWQ algorithm slightly outperforms logistic regression, and both methods slightly outperform ID3.

Table 1. Prior research on the comparison of machine learning systems and statistical techniques

Author(s)	Task	Techniques	Main results
Kattan *et al.* (1993)	Modeling decisions on account overdrafts.	• LDA. • RPA. • ID3. • NN.	Recursive partitioning and ID3 outperform DA and NN. RPA builts smaller trees than ID3, due to its built-in pruning procedure.
Deal and Edgett (1997)	Assessing the criteria that contribute to successful product development for financial services.	• LDA. • Logit. • RPA.	The approaches yielding the most information are discriminant analysis and RPA. Logit is effective as a supportive technique.
Didzarevich *et al.* (1997)	Prediction of bankruptcy.	• LDA. • Logit. • RPA. • CN2. • Bayesian networks.	CN2 rules and bayesian networks have a severe overfitting problem, but these problems could be solved in the future through the application of criteria that penalize complex structures.
Jeng *et al.* (1997)	Prediction of bankruptcy	• Fuzzy Inductive Learning Algorithm (FILM). • ID3. • LDA.	Induction systems achieve better results than LDA. FILM slightly outperforms ID3.
Slowinsky *et al.* (1997)	Forecasting the acquisition of firms.	• LDA. • Rough sets.	Rough sets perform better than LDA.
Varetto (1998)	Analysis of insolvency risk.	• LDA. • Genetic algorithms.	LDA proved to be slightly better than the generation of linear functions through genetic algorithms and the calculation of scores based on rules obtained using genetic algorithms.
Dimitras *et al.* (1999)	Predicting business failure.	• LDA. • Logit. • Rough sets.	Rough sets perform clearly better than LDA and slightly better than logistic regression.

Table 1 (cont). Prior research on the comparison of machine learning systems and statistical techniques

As can be seen, some authors find that inductive systems outperform statistical techniques, while others conclude the opposite. It depends on the classification task, the variables, the database, and the inference engine. This evidences the need for a replication of the comparison when a new task is considered, a new inference engine is used, and different kinds of companies make up the database.

3. METHODOLOGY

The main objective of this work is to compare the accuracy of LDA, logistic regression and SEE5 algorithm when they are used to predict if a company belongs to a very efficient group or not. A binary variable is defined which equals zero

when the company belongs to the group of the least efficient firms, and is equal to one when the firm is in the group of the most efficient companies.

As a secondary objective, the sensitivity of the accuracy of the systems and the number of variables in the functions/rules when sample size varies is tested.

In the following paragraphs we discuss the sample selection process and the variables. In addition, a brief description of the three techniques is provided, the treatment for the misclassification costs are detailed and the limitations of the methodology are expounded.

3.1. Sample selection

It starts from a database which is made up of the annual accounts of the companies located in the Principality of Asturias[1]. Businesses are required to deposit their financial statements in the "*Registro Mercantil de Asturias*" (Asturian Business Register). The analysed accounts correspond to the year 1995.

A set of filters is applied to the database to guarantee the quality of financial information and also to guarantee that the selected sample really shows the economic activity of each sector. Companies are eliminated if:

- They did not carry out any activity during 1995.
- 1995 was the first year of business.
- They omit data about fixed assets in the notes to the accounts.
- They do not provide any information about their employees.

After the filtering process, the chosen sample is divided into sectors because a separate analysis of every branch of economic activity will be made. The sectorial classification corresponds to the basic level of the European Community's activities nomenclature, which has 17 sections.

Farming, fishing and production and distribution of energy, gas and water sectors are excluded from this analysis because the number of companies is insufficient. Financial companies are also eliminated, as their annual accounts have special characteristics. "Public administration", "Private houses with employees" and "Overseas organisms" are also excluded because they are not corporations.

[1] The principality of Asturias is a province in the northern part of Spain.

To guarantee a sufficient number of companies in every branch, extractive and manufacturing industries have been merged into one sector. For the same reason, "Education", "Health and veterinary activities; social services" and "Other social activities and other services to the community; personal services" have been combined. Table 2 shows the sectors which have really been considered in this study.

N.	Name
1	Extractive and manufacturing industry
2	Building
3	Trade: mending of motor vehicles, motor-cycles. Personal consumer goods
4	Hotel industry
5	Transport, storage and communications
6	Real estate and rent activities; management services
7	Other (includes "Education", "Health …", and "Other social activities …")

Table 2. Analysed sectors

3.2. The formally dependent variable[2]

As settled before, the formally dependent variable is a dichotomic one which is equal to zero if the company is included in the most efficient group and is equal to one if it belongs to the least efficient group.

If we take into account the limitations of the available information, efficiency will be calculated with the economic profitability ratio. This ratio is the quotient between the operating result and total assets. Some authors (Kay, 1976; Long and Ravenscraft, 1984; Kelly and Tippet, 1991; Brief and Lawson, 1992) claim that this is a suitable measure of efficient management, in spite of its limitations.

In order to include both efficient and inefficient groups, the specification is made for each sector by discarding the intermediate quartiles in the economic profitability ratio. In this way, the group which has the most efficient companies will comprise 25% of the firms with the highest economic profitability and the group which has the least efficient companies will comprise 25% of the firms with the lowest value for this ratio.

Table 3 indicates the number of companies included in each group and each sector, when we apply the selection process. Further details on the size and efficiency of the firms are shown in appendix A.

[2] From now on we will use the terms 'formally dependent variable' and 'formally independent variables' because the techniques only determine the strength of the relation among them without specifying which is cause and which effect.

Sector	Low Efficiency	High Efficiency	Total
Industry	145	145	290
Building.	120	120	240
Trade: mending of motor vehicles, motor-cycle. Personal consumer goods	283	283	566
Hotel industry	53	53	106
Transport, storage and communications	63	63	126
Real estate and rent; management services	99	99	198
Other	55	55	110

Table 3. Composition of the sample

For each sector, this sample is split into two sub-samples, each containing the same number of low efficiency and high efficiency companies. One of the subsamples is used for the estimation of the models, while the other is the holdout sample.

As can be seen, some sectors have only a few companies while others have a high number. Thus, the sensitivity of each technique to changes in sample size will be tested.

3.3. The formally dependent variables

As formally independent variables, we chose a set of indicators including those proposed by Prado (1997) and López (2000) for the analysis of the financial statements drawn up according to Spanish GAAP. The ratios which are intrinsically connected to the economic profitability are excluded. In this way, the set of variables, which can be seen on table 4, includes neither profitability ratios nor variables that measure the productivity of the factors.

With the information in the annual accounts, we consider that this set of variables is sufficient to describe the economic and financial situation of each company, as a previous study on the variance-covariance matrix indicates that there are high correlations among a large number of them (De Andrés, 1998).

On appendix B is shown some descriptive statistical information on each variable. The marked positive skewness that characterizes the distribution of most of the variables is noticeable.

3.4. The classification techniques

As previously discussed, three techniques are compared, LDA, logistic regression and the SEE5 algorithm. A brief description of their foundations and the specific application of these systems follows.

Code	Variable
V01	Average number of employees
V02	Average total assets
V03	$\frac{\text{Turnover sales year 1995} - \text{Turnover sales year 1994}}{\text{Turnover sales year 1994}}$
V04	$\frac{\text{Net turnover sales}}{\text{Total employment}}$
V05	$\frac{\text{Total Debt}}{\text{Equity capital}}$
V06	$\frac{\text{Current liabilities}}{\text{Total Debt}}$
V07	$\frac{\text{Net fixed assets}}{\text{Total liabilites} - \text{Current liabilities}}$
V08	$\frac{\text{Net fixed assets}}{\text{Net total assets}}$
V09	$\frac{\text{Net turnover sales}}{\text{Current assets} - \text{Current liabilities}}$
V10	$\frac{\text{Tangible fixed assets + intangible fixed assets}}{\text{Total employment}}$
V11	$\frac{\text{Accumulated depreciation fixed assets. (tang. And intang.)}}{\text{Gross tangible and fixed unliquid assets}}$
V12	$\frac{\text{Depreciation fixed assets}}{\text{Gross tangible and intangible fixed assets}}$
V13	$\frac{\text{Fixed assets entry}}{\text{Gross fixed asset + fixed asset provision}}$
V14	$\frac{\text{Fixed assets outflow}}{\text{Gross fixed assets + fixed assets provision}}$
V15	$\frac{\text{Total assets}}{\text{Total debt}}$
V16	$\frac{\text{Current assets}}{\text{Current liabilities}}$
V17	$\frac{\text{Current assets} - \text{Inventory}}{\text{Current liabilities}}$
V18	$\frac{\text{Available assets + Financial contingent assets}}{\text{Current liabilities}}$
V19	$\frac{\text{Financial expenses}}{\text{Total long-term liabilities}}$
V20	$\frac{\text{Total employment year 1995} - \text{Total employment year 1994}}{\text{Fixed employment year 1994}}$
V21	$\frac{\text{Not fixed employment}}{\text{Total employment}}$
V22	$\frac{\text{Labor costs}}{\text{Total employment}}$
V23	$\frac{\text{Inventory}}{\text{Operating consumption/365}}$
V24	$\frac{\text{Current liabilities}}{(\text{Operating expenses} + \text{Others operating expenses})/365}$
V25	$\frac{\text{Debtors}}{\text{Net turnover sales/365}}$

Table 4. Formally independent variables

3.4.1 LINEAR DISCRIMINANT ANALYSIS (LDA)

With this technique, the classification is made by estimating a discriminant function which takes the form $Z=v_0+v_1x_1+\quad+v_nx_n$, where $x_1...x_n$ are the formally independent variables and $v_0...v_n$ 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).

From the seminal paper by Altman (1968), many researchers have used LDA, most of them on the topic of financial distress prediction. The following are noteworthy: Edmister (1972), Blum (1974), Altman *et al.* (1977), Taffler (1983), Laffarga *et al.* (1988), Houghton and Woodliff (1987), Laitinen (1991), López *et al.* (1994), Lizárraga (1997), and Calvo-Flores and Arqués (1997).

In order to avoid multicollinearity and instability in the resulting functions, we have applied an iterative stepwise procedure for the selection of the variables. In each iteration, some variable is added or deleted according to statistical significance criteria[3]. The process ends when further additions or deletions of variables do not improve the explanatory capacities of the model. The software used to estimate the models was SPSS 10.0.

3.4.2 LOGISTIC REGRESSION

This technique classifies by calculating the probability of each individual belonging to each group. This probability is a function of a linear combination of the explanatory variables. The functional form chosen is the logistic one, which takes the following expression:

$$p = \frac{e^{\beta}}{1+e^{\beta}} = \frac{1}{1+e^{-\beta}}$$

[3] The criterion for the inclusion or deletion of a variable is its Wilk's lambda. A variable enters in the models if its significance is under 0.05 and is eliminated if it is above 0.10 (See Uriel, 1995).

Where *p* is the probability of belonging to the group called 1 (the other group is 'group 0') and β is a linear combination of the explanatory variables. The determination of coefficients is carried out by an iterative maximum likelihood estimation method (see Hosmer and Lemeshow, 1989).

Some authors (Efron, 1975; Lo, 1986) are of the opinion that the logistic regression achieves optimum results when the explanatory variables fit one distribution of the exponential family. Because this hypothesis is less restrictive than those which affect other statistical techniques (i.e. LDA), logit models have been widely used to solve several classification problems in the accounting field, especially those related to prediction of financial failure. The following papers are noteworthy: Martin (1977), Ohlson (1980), Mensah (1983), Gentry *et al.* (1985), Casey and Bartczak (1985), Peel and Peel (1987), Lau (1987), Laffarga *et al.* (1987), Pina Martínez (1989), Mora (1994), Lizárraga (1997) and López *et al.* (1998).

As well as for LDA, and for the same reasons, in the estimation of logit models we have also applied an iterative stepwise process for the selection of the variables[4] . The software used has also been SPSS 10.0.

3.4.3 SEE5 ALGORITHM

This algorithm is the latest version of the induction systems developed by Quinlan (1979, 1983, 1986, 1988, 1993 and 1997). These systems use the entropy criterion, which means that the classification tree grows if we choose, at each step, the variable which has the biggest entropy or amount of information. This variable 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.

[4] A variable enters if its significance in the model is under 0.05 and is deleted if is above 0.10. The used method is "conditional forward" (see Peña, 1987).

There are a lot of research papers in Accounting on the topic of the application of induction systems to classification tasks. Most of them use also statistical techniques and compare the results. The most relevant papers of this kind have been reviewed in section 2. From the works that do not consist of comparisons, we must highlight McKee (1995), López *et al.* (1998), Correa (1999), González *et al.* (1999, 2000), which use the Quinlan algorithms; McKee (1998) which uses rough sets; and Bonsón *et al.* (1996, 1997), who mix genetic algorithms with the entropy criterion, using the commercial application XPERT RULE.

In this paper, we have followed two steps in the application of the 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. The algorithm selects 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.

3.5. Costs of misclassifications

For most of the classification tasks, the different types of misclassifications do not have the same costs, and this fact must be borne in mind in the estimation of the functions/rules. However, for economic analysis purposes the calculation of the costs is not a simple process. As Mensah (1984, p. 240) points out, the determination of the misclassification costs is an admittedly arbitrary choice, since they are likely to be user and situation specific.

We have chosen to estimate the misclassification costs from the investors' point of view, and therefore the system for the calculation is the following: the cost of classifying a high efficiency company as a low efficiency one (type I error) is the loss of profitability due to the 'wrong not investing decision'. This loss is measured through the difference between the median of the economic profitability for the high efficiency firms and the yields of the riskless investment (Spanish government bonds), which was 9% for the year 1995.

In the same way, the cost of classifying a low efficiency company as a high efficiency one (type II error) is the loss due to the 'wrong investing decision', that

is, the difference between the yields of the riskless investment and the median of the economic profitability for the low efficiency firms.

The misclassification costs for each branch of activity can be seen on table 5. It is remarkable that, with the exception of "Real estate and rent; management services", the costs of type II error are higher than that of type I.

Sector	Type I error (high eff. as low)	Type II error (low eff. as high)	Type I / Type II
Industry	9.029%	24.393%	0.37017
Building	12.455%	25.614%	0.48628
Trade: mending of motor vehicles, motor-cycle. Personal consumer goods	6.842%	21.173%	0.32318
Hotel industry	11.156%	25.316%	0.44069
Transport, storage and communications	16.290%	23.964%	0.67979
Real estate and rent activities; management services	18.520%	18.454%	1.00357
Other.	22.950%	40.969%	0.56017

Table 5. Misclassification costs

3.6. Research limitations

As defined above, the methodology of this study has several limitations that are either impossible or not cost-effective to overcome. First of all, the decision of analyzing small companies 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 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.

Other limitations are the aforementioned arbitrariness of the determination of the misclassification costs and the scarce number of firms in each sector after splitting the data into estimation and holdout samples. Nevertheless, this last feature can be understood as a 'test in extreme conditions' for the techniques. It is very interesting to know which techniques perform better when the sample size is very small, as this allows researchers to consider very specific branches of economic activity.

4. RESULTS

In this section, the results of the research are discussed, having taken into account the methodology explained above and its limitations. First of all, the sets

of variables in the functions/rules are exposed, and the sensitivity of the size of each set when the number of sector companies varies is tested. Second, the accuracy of each system is compared with the others and the sensitivity of the accuracy of the models when the sample size varies is tested.

4.1. Variables in the functions/rules

The sets of variables in the logit and discriminant functions and in the rules of the SEE5 algorithm are shown in table 6. This table also displays the number of rules for each rule system.

Sector	Sets of variables			N. of SEE5 rules
	LDA	Logit	SEE5 Rules	
Industry	V15, V19	V15, V19	V01, V02, V04, V13, V17, V18, V19, V22, V25	11
Building	V01, V04, V19, V20, V24	V01, V04, V19, V20, V24	V01, V04, V06, V11, V19	5
Trade: mending of motor vehicles, motor-cycles. Personal consumer goods	V4, V10, V19	V4, V5, V10, V11, V19	V02, V03, V04, V07, V08, V09, V13, V17, V19, V20, V21, V24	10
Hotel industry	V13, V14, V16	V16	V03, V13, V17, V18, V19	5
Transport, storage and communications	V02, V13, V19	V02, V09, V19	V01, V03, V07, V13, V14, V16, V17, V23	6
Real estate and rent activities; management services	V25	V25	V02, V03, V05, V07, V13, V15, V19	9
Other	V04	V04, V22	V04, V17, V24	4

Table 6. Sets of variables and number of rules

The following conclusions can be drawn:

- The selected sets for LDA and logit are very similar. The number of relevant variables, except in the building sector, is always bigger if SEE5 is used. This can be an indication of its superior sensitivity in detecting regularities in the values of the different variables.
- There seems to be a direct relation between the number of sector companies and the number of variables chosen and rules inferred. Testing this through the Spearman's coefficient of correlation (Rho)[5], we found that this correlation is significant (at a level of 0.05) only for the rules and variables of the SEE5 algorithm. The explanation of this fact could be that the aforementioned regularities in the variables become more evident when sample size increases.

[5] We use Spearman's Rho instead of the more popular Pearson's coefficient to measure correlations because the distributional properties of the variables are unknown (see McPherson, 1990). The detailed results of these and subsequent Spearman's tests are not shown due to space restrictions.

4.2. Accuracy of the systems

The error percentage for each technique is detailed separately for every type of error. As settled before, we shall call *type I error* the error which classifies the most effective companies in the group of the least effective ones. Therefore, a *type II error* is to classify a company belonging to the least effective group in the most effective firms. Finally, global accuracy is indicated.

4.2.1 TYPE I ERROR

Table 7 shows the error percentage for the estimation and holdout samples and the variation from estimation to holdout. Table 8 displays the results of the application of the Mann-Whitney test for the comparison of the error levels (estimation and holdout) and the variations.

Differences between LDA and Logit are not significant. SEE5 is not better than LDA, neither for the estimation nor for the holdout sample (at a 0.05 level). SEE5 performs better than logit only for the estimation sample. Regarding the loss of accuracy when the system is tested in the validation sample, SEE5 is more sensitive than the statistical techniques, while there is no difference between LDA and logit.

Sector	Error percentage in the holdout (estimation) sample			Increase (decrease) from estimation to holdout		
	LDA	Logit	SEE5	LDA	Logit	SEE5
Industry	76.39 (72.22)	63.88 (62.5)	45.83 (26.38)	4.17	1.30	19.45
Building	41.66 (38.33)	63.33 (31.66)	31.66 (28.33)	3.33	31.67	3.33
Trade: mending of motor vehicles, motor-cycle. Personal consumer goods	87.23 (85.91)	71.63 (69.71)	63.38 (59.86)	1.32	1.92	3.52
Hotel industry	81.48 (65.38)	62.96 (61.53)	40.74 (42.30)	16.10	1.43	(1.56)
Transport, storage and communications	59.37 (58.06)	71.87 (41.93)	25.80 (6.45)	1.31	29.94	19.35
Real estate and rent; management services	18.00 (18.36)	22.00 (20.40)	28.00 (4.08)	(0.36)	1.60	23.92
Other	85.71 (85.18)	64.28 (40.74)	14.28 (14.81)	0.53	23.54	(0.53)

Table 7. Percentage of type I error

Comparison	Estimation sample	Validation Sample	Variation from estimation to holdout
LDA – Logit	45.6%	90.2%	80.5%
LDA – SEE5	7.3%	9.7%	0.1%
Logit – SEE5	0.4%	7.3%	0.2%

Table 8. Critical significance level on the Mann-Whitney test for type I error

Testing through Spearman's Rho the sensitivity of the accuracy when sample size varies, we found that none of the techniques, neither in the estimation nor in the validation samples, is affected by this parameter. The loss of accuracy from estimation to validation remains also unaffected.

4.2.2 TYPE II ERROR

The error percentages for the two sub-samples (estimation and validation) and the variations are detailed in table 9. The number of type II misclassifications is smaller than that of type I, mainly because the procedure to estimate error costs assigns lower costs to this kind of error, with the exception of "Real estate and rent; management services".

Table 10 shows the results of the Mann-Whitney test for the comparison of the error levels and the variations between the different techniques. As in the case of type I error, for type II there are no significant differences between the two statistical techniques. With a significance level of 0.05, SEE5 is better than LDA only in the estimation sample, and better than logit only in the holdout sample. The loss of accuracy when the system is tested in the validation sample is not significantly bigger for SEE5 algorithm.

Sector	Error percentage in the holdout (estimation) sample			Increase (decrease) from estimation to holdout		
	LDA	Logit	SEE5	LDA	Logit	SEE5
Industry	5.55 (2.77)	11.11 (9.72)	20.83 (0.00)	2.78	1.39	20.83
Building	23.33 (8.33)	11.66 (8.33)	23.33 (6.66)	15.00	3.33	16.67
Trade: mending of motor vehicles, motor-cycle. Personal consumer goods	4.25 (2.11)	10.63 (7.04)	9.21 (0.00)	2.14	3.59	9.21
Hotel industry	7.69 (3.70)	19.23 (3.70)	15.38 (0.00)	3.99	15.53	15.38
Transport, storage and communications	19.35 (12.50)	6.45 (15.62)	53.12 (0.00)	6.85	(9.17)	53.12
Real estate and rent; management services	51.02 (62.00)	42.85 (56.00)	22.44 (2.00)	(10.98)	(13.15)	20.44
Other	0.00 (0.00)	0.00 (3.57)	51.85 (3.57)	0.00	(3.57)	48.28

Table 9. Percentage of type II error

Comparison	Estimation sample	Validation Sample	Variation from estimation to holdout
LDA – Logit	31.8%	53.5%	31.8%
LDA – SEE5	2.6%	5.3%	38.3%
Logit – SEE5	7.3%	3.8%	90.2%

Table 10. Critical significance level on Mann-Whitney test for type II error

The results of the Spearman's rank correlation test on the sensitivity of the accuracy to variations in sample size are the same as those obtained for type I error, that is, the number of sector companies is not significantly correlated with the error rates.

4.2.3 GLOBAL PERCENTAGE

Tables 11 and 12 show respectively the error percentages for each one of the branches of economic activity and the result of the Mann-Whitney test when applied to the comparison of the techniques.

As in the case of type I and type II errors, for the overall rates there are no significant differences between LDA and Logit. In the estimation sample, SEE5 is better than both statistical techniques. Nevertheless, when tested with the holdout samples, SEE5 outperforms only logistic regression. Regarding the loss of accuracy in the validation samples, the differences are only significant between Logit and SEE5.

Sector	Error percentage in the holdout (estimation) sample			Increase (decrease) from estimation to holdout		
	LDA	Logit	SEE5	LDA	Logit	SEE5
Industry	40.97 (37.5)	37.50 (36.11)	33.33 (13.19)	3.47	1.39	20.14
Building	32.50 (23.33)	37.50 (20.00)	27.50 (17.50)	9.17	17.50	10.00
Trade: mending of motor vehicles, motor-cycle. Personal consumer goods	44.87 (44.52)	40.28 (39.22)	36.39 (30.03)	0.35	1.06	6.36
Hotel industry	45.28 (33.96)	41.50 (32.07)	28.30 (20.75)	11.32	9.43	7.55
Transport, storage and communications	39.68 (34.92)	39.68 (28.57)	39.68 (3.17)	4.76	11.11	36.51
Real estate and rent; management services	34.34 (40.40)	32.32 (38.38)	25.25 (3.03)	(6.06)	(6.06)	22.22
Other	43.63 (41.81)	32.72 (21.81)	32.72 (9.03)	1.82	10.91	23.69

Table 11. Global percentage of error

Comparison	Estimation sample	Validation Sample	Variation from estimation to holdout
LDA – Logit	16.5%	20.9%	53.5%
LDA – SEE5	0.7%	7.3%	9.7%
Logit – SEE5	0.1%	1.7%	1.1%

Table 12. Critical significance level on the Mann-Whitney test for the global percentage of error

When testing the relation between the accuracy of the systems and the number of sector companies, we found that no significant correlations exist, neither for any of the techniques nor for any kind of sample (estimation or holdout), nor for the differences.

5. SUMMARY AND CONCLUSIONS

In the present research we have tested the performance of a rule induction classifier system –the SEE5 algorithm by Quinlan– with that of two commonly used statistical models, LDA and logistic regression. There are other comparisons in the literature, but the topic of the characterization of the most/least efficient companies, the use of a sample of small companies, and the rule induction system tested (SEE5) are features that distinguish this work from other research papers.

The main results indicate that SEE5 outperforms Logit in terms of accuracy. The superiority of SEE5 over LDA is less clear. In contrast, SEE5 seems to suffer bigger increases in the error rates when the systems are tested using holdout samples. This finding suggests that the built-in procedures incorporated in the SEE5 code in order to prevent the overfitting of the models are not as effective as they could be.

Regarding the sensitivity of the systems to variations in the number of sector companies, only SEE5 seems to be affected in terms of the number of selected variables and the number of rules, as they are significantly bigger in the bigger sectors. Neither the accuracy of any of the models nor the loss of accuracy when validating with holdout samples is affected by changes in the number of sector companies.

As directions for future research, we can mention the study of classification problems with more than two classes, the use of other machine learning systems (i.e. induction of fuzzy rules or those based on genetic algorithms), and the extension of the present research using a validation sample formed on the basis of annual accounts of big companies.

6. REFERENCES

Altman, EI. (1968): Financial ratios, discriminant analysis and the prediction of the corporate bankruptcy. *Journal of Finance* 23 (4):589-609.

Altman, EI, RG Haldeman, and P Narayanan. (1977): ZETA analysis: a new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance* 1:29-54.

Blum, M. (1974): Failing company discriminant analysis. *Journal of Accounting Research* 12 (1):1-26.

Bonsón, E, MP Martín, and G Sierra. (1996): Induction of decision trees as a methodology for financial analysis. An application to crisis in banking. In *Intelligent Systems in Accounting and Finance*, edited by G Sierra and E Bonsón. Huelva, Spain.

Bonsón, E, and G Sierra. (1996): Intelligent Accounting: impact of Artificial Intelligence on accounting research and accounting information. Paper read at International Conference on Intelligent Technologies in Human-Related Sciences, León, Spain.

Bonsón, E, and G Sierra. (1997): Intelligent technologies in accounting research. In *Intelligent Technologies in Accounting and Business*, edited by G Sierra and E Bonsón. Huelva, Spain.

Braun, H, and JS Chandler. (1987): Predicting stock market behavior through rule induction: an application of the learning-from-examples approach. *Decision Sciences* Summer:415-429.

Brief, RP, and RA Lawson. (1992): The role of the accounting rate of return in financial statement analysis. *Accounting Review* 67 (2):411-426.

Calvo-Flores, A, and A Arqués. (1997): Factores discriminantes del riesgo financiero en la industria manufacturera española. In *Predicción de la Insolvencia Empresarial*, edited by AECA. Madrid, Spain.

Casey, C, and N Bartczak. (1985): Using operating cash flow data to predict financial distress: some extensions. *Journal of Accounting Research* 23 (1):384-401.

Correa, A. (1999): *Factores determinantes del crecimiento empresarial*. Ph. D. Dissertation, University of La Laguna, Spain.

Cronan, TP, LW Glorfeld, and LG Perry. (1991): Production system development for expert systems using a recursive partitioning induction approach: an application to mortgage, commercial and consumer lending. *Decision Sciences* 22 (4):812-845.

De Andrés, J. (1998): *Caracterización económico-financiera de los sectores de la economía asturiana en función del nivel de rentabilidad de las empresas*. Ph. D. Dissertation, University of Oviedo, Spain.

Deal, K, and SJ Edgett. (1997): Determining success criteria for financial products: A comparative analysis of CART, logit and factor/discriminant analysis. *Service Industries Journal* 17 (3):489-506.

Didzarevich, S, F Lizarraga, P Larrañaga, B Sierra, and MJ Gallego. (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.

Dimitras, AI, R Slowinsky, R Susmaga, and C Zopounidis. (1999): Business failure prediction using rough sets. *European Journal of Operational Research* 114 (2):263-280.

Edmister, RO. (1972): An empirical test of financial ratio analysis for smaller business failure prediction. *Journal of Financial & Quantitative Analysis* March:1477-1497.

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

Elliott, JA, and DB Kennedy. (1988): Estimation and prediction of categorical models in Accounting research. *Journal of Accounting Literature* 7:202-242.

Friedman, J. (1977): A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers* C-26 (4):404-408.

Frydman, H, EI Altman, and DL Kao. (1985): Introducing recursive partitioning for financial classification: the case of financial distress. *Journal of Finance* 40 (1):269-291.

Garrison, LR, and RH Michaelsen. (1989): Symbolic concept acquisition: A new approach to determining underlying tax law constructs. *The Journal of the American Tax Association* Fall:77-91.

Gentry, JA, P Newbold, and DT Whiteford. (1985): Classifying bankrupt firms with funds flow components. *Journal of Accounting Research* 23 (1):146-160.

González, AL, A Correa, and JA Blázquez. (1999): Perfil del fracaso empresarial para una muestra de pequeñas y medianas empresas. Paper read at Tenth AECA Congress, Zaragoza, Spain.

González, AL, A Correa, JA Blázquez. (2000): Análisis de la rentabilidad de la pequeña y mediana empresa. Paper read at Ninth ASEPUC Congress, Las Palmas de Gran Canaria, Spain.

Gordon, L, and R Olshen. (1978): Asymptotically efficient solutions to the classification problem. *Annals of Statistics* 6 (3):515-533.

Hansen, JV, GJ Koehler, WF Messier, and JF Mutchler. (1993): Developing knowledge structures: A comparison of a qualitative-response model and two machine-learning algorithms. *Decision Support Systems* 10 (2):235-243.

Hosmer, DW, and S Lemeshow. (1989): *Applied logistic regression*. New York, USA: J Wiley and Sons.

Houghton, KA, and R Woodliff. (1987): Financial ratios: the prediction of corporate 'success' and failure. *Journal of Business Finance & Accounting* Winter: 537-554.

Jeng, B, YM Jeng, and TP Liang. (1997): FILM: A fuzzy inductive learning method for automated knowledge acquisition. *Decision Support Systems* 21 (2):61-73.

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

Kattan, MW, DA Adams, and MS Parks. (1993): A comparison of machine learning with human judgment. *Journal of Management Information Systems* 9 (4):37-57.

Kay, J. (1976): Accountants, too, could be happy in a golden age: the accountant's rate of profit and the internal rate of return. *Oxford Economic Papers* November:447-460.

Kelly, G, and M Tippet. (1991): Economic and accounting rates of return: a statistical model. *Accounting & Business Research* 21 (84):321-329.

Laffarga, J, JL Martín, and MJ Vázquez. (1987): Predicción de la crisis bancaria española: la comparación entre el análisis logit y el discriminante. *Cuadernos de Investigación Contable* 1 (1):103-110.

Laffarga, J, JL Martín, and MJ Vázquez. (1988): Forecasting bank failures: the Spanish case. *Studies in Banking and Finance* 7:127-139.

Laitinen, EK. (1991): Financial ratios and different failure processes. *Journal of Business, Finance & Accounting* 18 (5):649-673.

Lau, AHL. (1987): A five state financial distress prediction model. *Journal of Accounting Research* 25 (1):127-138.

Liang, TP, JS Chandler, I Han, and J Roan. (1992): An empirical investigation of some data effects on the classification accuracy of probit, ID3 and neural networks. *Contemporary Accounting Research* 9 (Fall):306-328.

Lizárraga, F. (1997): Utilidad de la información contable en el proceso de fracaso: análisis del sector industrial de la mediana empresa española. *Revista Española de Financiación y Contabilidad* 92:871-915.

Long, WF, and DJ Ravenscraft. (1984): The misuse of accounting rates of return: comment. *American Economic Review* 74 (3):494-500.

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

López, A, J De Andrés, and E Rodríguez. (1998): La opinión emitida por el auditor en el informe y su relación con determinadas variables. Paper read at Eigth ASEPUC Congress, Alicante, Spain.

López, J, JL Gandía, and R Molina. (1998): La suspensión de pagos en las PYMES: una aproximación empírica. *Revista Española de Financiación y Contabilidad* 94:71-97.

López, D, J Moreno, and P Rodríguez. (1994): Modelos de previsión del fracaso empresarial: aplicación a entidades de seguros en España. *ESIC-Market* 84:83-125.

Marais, ML, JM Patell, and MA Wolfson. (1984): The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications. *Journal of Accounting Research* 22 (supplement):87-114.

Martin, D. (1977): Early warning of bank failure: a logit regression approach. *Journal of Banking & Finance* 1:249-276.

Mc Kee, TE. (1995): 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.

Mc Kee, TE. (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.

Mc Pherson, G. (1990): *Statistics in scientific investigation. Its basis, application, and, interpretation*. New York, USA: Springer-Verlag.

Mensah, YM. (1983): The differential bankruptcy predictive ability of specific price level adjustments: some empirical evidence. *Accounting Review* 58 (2): 228-246.

Messier, WF, and JV Hansen. (1988): Inducing rules for expert system development: an example using default and bankruptcy data. *Management Science* 34 (12):1403-1415.

Mora, A. (1994): Los modelos de predicción del fracaso empresarial: una aplicación empírica del logit. *Revista Española de Financiación y Contabilidad* 78:203-233.

Ohlson, JA. (1980): Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18 (1):109-132.

Parker, JE, and KF Abramowicz. (1989): "Predictive abilities of three modelling procedures", *The Journal of American Taxation Association*, Fall: 37-53.

Peel, M.J.; Peel, D.A. (1987): Some further empirical evidence on predicting private company failure. *Accounting & Business Research* Winter: 57-66.

Peña, D. (1987): *Estadística. Modelos y métodos*. Madrid, Spain: Alianza.

Pina, V. (1989): La información contable en la predicción de la crisis bancaria 1977-1985. *Revista Española de Financiación y Contabilidad* 58:309-338.

Prado, J.M. (1997): *Análisis económico-financiero de las empresas ed Castilla y León por sectores de actividad 1993-1994*. Valladolid, Spain: Junta de Castilla y León.

Quinlan, JR. (1979): Discovering rules by induction from large collections of examples. In *Expert systems in the microelectronic age*, edited by D Michie.

Quinlan, JR. (1983): Learning efficient classification procedures. In *Machine learning: an Artificial Intelligence approach*. Palo Alto, USA: Tioga Press,.

Quinlan, JR. (1986): Induction of decision trees. *Machine Learning* 1 (1):81-106.

Quinlan, JR. (1988): Decision trees and multivalued attributes. *Machine Intelligence* 11:305-318.

Quinlan, JR. (1993): *C4.5: Programs for machine learning*. California, USA: Morgan Kaufmann Publishers, Inc.

Quinlan, JR. (1997): *See5*. Available form http://www.rulequest.com/See5-info.html.

Slowinsky, R, C Zopounidis, and AI Dimitras. (1997): Prediction of company acquisition in Greece by means of the rough set approach. *European Journal of Operational Research* 100 (1):1-15.

Taffler, RJ. (1983): The assessment of company solvency and performance using a statistical model. *Accounting & Business Research* Autumn:295-305.

Tam, KY, and MY Kiang. (1992): Managerial applications of neural networks: The case of bank failure predictions. *Management Science* 38 (7):926-947.

Uriel, E. (1995): *Análisis de datos. Series temporales y análisis multivariante*. Madrid, Spain: Editorial AC.

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

APPENDIX A

	1st quartile	Median	3rd quartile	Mean	Std. Dev.
Industry	-2.44%	6.34%	12.39%	1.03%	0.4075
Building	-7.20%	6.17%	13.34%	-1.94%	0.7732
Trade	-1.49%	4.88%	10.23%	1.57%	0.2487
Hotel industry	-3.48%	5.13%	12.35%	1.04%	0.2772
Transport …	-1.11%	7.25%	15.19%	3.85%	0.2924
Real estate …	-2.87%	6.99%	16.67%	6.94%	0.3021
Other	-7.84%	6.18%	19.27%	3.39%	1.0216

Efficiency of the companies (before discarding the intermediate quartiles)

Sector	High efficiency			Low efficiency		
	1st quartile	Median	3rd quartile	1st quartile	Median	3rd quartile
Industry	6	11	18	3	6	14
Building	4	7.395	15.5	2	4	8
Trade	3	5	9	2	3	5
Hotel industry	3	5	10.5	2	3	5
Transport …	2	4	12	1.25	3	6
Real estate …	2	3.30	7	1	2.15	4
Other	1	4	9	2	3.15	5.12

Size of the companies in the sample (measured by the average number of employees)

APPENDIX B

VAR	High efficiency					Low efficiency				
	1 quart	Med.	3 quart	Mean	Std dev	1 quart	Med.	3 quart	Mean	Std dev
V01	3.000	5.150	12.000	14.053	38.049	2.000	3.250	6.000	27.458	379.43
V02	8669	21791	46929	75277	588856	5247	12518	31229	53005	263190
V03	0.024	0.165	0.542	2.322	38.252	-0.197	0.004	0.303	1.539	18.454
V04	6019	10287	19994	18745	39622	3427	5707	10360	10099	20654
V05	0.448	1.367	3.6443	1.244	50.867	-3.072	0.650	3.361	29.372	706.98
V06	0.852	0.991	1.000	0.899	0.166	0.860	1.000	1.000	0.897	0.180
V07	0.246	0.633	1.053	0.650	4.980	-0.099	0.511	1.208	1.049	43.796
V08	0.119	0.278	0.490	0.328	0.337	0.127	0.305	0.560	0.357	0.269
V09	-7.758	4.186	14.207	19.465	496.66	-6.490	-1.367	4.366	1.290	109.93
V10	403.65	1167	2828	2519	4136	327.01	933.50	2505	3515	16924
V11	0.137	0.292	0.479	0.328	0.231	0.147	0.297	0.518	0.345	0.248
V12	0.057	0.098	0.141	0.117	0.158	0.057	0.100	0.149	0.139	0.374
V13	0.010	0.100	0.368	0.262	0.484	0.000	0.019	0.206	0.455	7.951
V14	0.000	0.000	0.011	0.073	0.306	0.000	0.000	0.000	0.128	1.530
V15	1.219	1.599	2.401	3.039	25.462	0.852	1.172	1.652	1.971	6.180
V16	0.876	1.166	1.727	2181	61126	0.525	0.834	1.246	30.072	806.05
V17	0.506	0.907	1.429	1910	53451	0.212	0.458	0.889	29.360	804.22
V18	0.200	0.447	0.934	668.64	18239	0.098	0.269	0.593	4.367	96.602
V19	0.022	0.069	0.115	0.093	0.164	0.005	0.033	0.073	0.047	0.051
V20	0.000	0.000	0.166	0.222	0.930	-0.063	0.000	0.000	0.097	0.687
V21	0.000	0.400	0.714	0.424	0.350	0.000	0.333	0.672	0.391	0.373
V22	1798	2510	3270	2762	1712	1609	2252	3028	2736	2813
V23	0.000	23.838	80.320	225.38	3069	0.000	56.916	160.07	458.50	5899
V24	93.336	147.43	262.45	259.12	562.73	136.36	235.87	499.54	521.70	1775
V25	20.196	48.417	87.838	75.264	283.62	19.139	56.598	120.79	-1730	55171

Descriptive statistical information about the formally independent variables