Feature Selection for Bankruptcy Prediction: A Multi-Objective Optimization Approach

A. Gaspar-Cunha[1], F. Mendes[1], J. Duarte[2], A. Vieira[2], B. Ribeiro[3], A. Ribeiro[4], J. Neves[4],

[1] Institute for Polymers and Composites/I3N, University of Minho, Campus de Azurém 4800-058 Guimarães, Portugal, e-mail: agc@dep.uminho.pt and fmendes@dep.uminho.pt

[2] Department of Physics, Instituto Superior de Engenharia do Porto, R. S. Tomé, 4200 Porto, Portugal, e-mail: jmmd@isep.ipp.pt and asv@isep.ipp.pt

[3] Department of Informatics Engineering, Center of Informatics and Systems, University of Coimbra, Coimbra 3030-290, Portugal, e-mail: bribeiro@dei.uc.pt

[4] ISEG School of Economics and Management, Technical University of Lisbon, Portugal, e-mail: andremsr@mail.pt and jcneves@iseg.utl.pt

Abstract

In this work a Multi-Objective Evolutionary Algorithm (MOEA) was applied for feature selection in the problem of bankruptcy prediction. The aim is to maximize the accuracy of the classifier while keeping the number of features low. A two-objective problem - minimization of the number of features and accuracy maximization – was fully analyzed using two classifiers, Logistic Regression (LR) and Support Vector Machines (SVM). Simultaneously, the parameters required by both classifiers were also optimized. The validity of the methodology proposed was tested using a database containing financial statements of 1200 medium sized private French companies. Based on extensive tests it is shown that MOEA is an efficient feature selection approach. Best results were obtained when both the accuracy and the classifiers parameters are optimized. The method proposed can provide useful information for the decision maker in characterizing the financial health of a company.

Keywords: feature selection, bankruptcy prediction, multi-objective optimization, evolutionary algorithms, support vector machines, logistic regression.

1 Introduction

Financial bankruptcy prediction is of high importance for banks, insurance companies, creditors and investors. One of the most important threats for business is the credit risk associated with counterparts. The rate of bankruptcies have increased in recent years and its becoming harder to estimate as companies become more complex and develop sophisticated schemes to hide their real situation. Due to the recent financial crisis and regulatory concerns, credit risk assessment is a very active area both for academic and business community. The ability to discriminate between faithful customers from potential bad ones is thus crucial for commercial banks and retailers (Atiya, 2001).

Different approaches have been used to analyze this problem, like discriminant analysis (Eisenbeis, 1977) and Logit and Probit models (Martin, 1977). However, most of these methods have important limitations. Discriminant analysis is limited due to its linearity, restrictive assumptions, for treating financial ratios as independent variables and can only be used with continuous independent variables. Furthermore, the choice of the regression function creates a bias that restricts the outcome and they are also very sensitive to

exceptions, while has an implicit Gaussian distribution on data, which is inappropriate in many cases.

More recently other approaches have been applied for bankruptcy classification, such as Artificial Neural Networks (ANN) (Atiya, 2001; Charitou & al., 2004; Neves & Vieira, 2006), Evolutionary Algorithms (EA) and Support Vector Machines (SVM) (Fan & Palaniswami, 2000). ANN, EA and SVM are used as complementary tools to classify credit risk. Some of the studies performed show that ANN outperforms discriminant analysis in bankruptcy prediction (Neves & Vieira 2006; Coats & Fant, 1993; Yang, 1999; Tan & Dihardjo, 2001). Huang & al. (2008) conclude that financial ratios are important tools in prediction of business failures and that they are commonly used to develop the models or classifiers. In their work failed firms are targeted aiming to seek out relevant features of their financial ratios. To this end, automatic clustering techniques are employed to automatically divide targeted failed firms into some clusters according to characteristics of financial ratios. In order to simplify the task of analysis, as well as to increase the classification accuracy, feature selection techniques are used to reduce the overall number of financial ratios analyzed. Also, in their paper the authors, particularly emphasizes the importance of both expert knowledge and data mining techniques in feature selection. This means that it is preferable to conduct the analysis task using not only the data mining technique but also the expert knowledge, and to compare their performances of classification accuracies in terms of the feature selection. In this way, more accurate results and practical insights can be obtained. More recently, Wu (2010) proposed a method which directly explores the features of failed firms rather than researching pairs of failed and non-failed firms. To this end, automatic clustering techniques and feature selection techniques are employed for this study. Taking these conclusions into account, it is generally recognized that further research is needed to achieve higher predictive capabilities, which is the avenue of the present research (Vieira & al., 2009).

Banks collect large amounts of data available from companies and other creditors. These data is often inconsistent and redundant and needs considerable manipulation to make it useful for problems like credit risk analysis. First, it is necessary to build a set of ratios that may be appropriated for the problem. Then, is necessary to further restrict the number of these ratios, or attributes with higher information content in order to reduce the complexity of the problem. Finally, these reduced set of attributes, or features, are used to train any classification algorithm designed to predict the company financial health.

Due to the large number of variables and the fact that some of these variables are highly correlated, it is crucial to have a feature selection algorithm to reduce the complexity of the problem (Guyon & al., 2006). As pointed out by many evidences, feature selection plays an important role in classification in terms of improving the predictive accuracy and decreasing the complexity of models. Additionally, the resultant predictive model is somewhat dependent on the parameters employed.

Considerable efforts have been put in Feature Reduction (FR) for forecasting bankruptcy prediction in financial problems. The two main approaches that have hitherto been pursued use Feature Selection (FS) and Nonlinear Dimension Reduction (NLDR) by projection methods. Examples of research concerning feature selection (which is the aim of the present paper) are presented by Atiya (2001), Kumar & Ravi (2007), Verikas & al. (2009), Shin & al. (2005) and Thomas (2007). For example, Kumar & Ravi (2007) and Verikas & al. (2009) performed a complete review of methods used for the prediction of business failure and introduced new trends in this area.

Concerning NLDR by projection methods, Rekba & al. (2004) tested a linear pre-processing stage using principal component analysis (PCA) for dimensionality reduction purposes. However, nonlinear projection methods (e.g. ISOMAP) have been successfully used by Ribeiro & al. (2009) making them more suitable for this problem. With the same goal, non-negative matrix factorization (NMF) is used by Ribeiro & al. (2009b) for extracting the most discriminative features.

Evolutionary Algorithms (EAs) are an excellent tool to deal with this problem, since they are able to provide the resource to simultaneously optimize the factors with potential impact in the performance, including subset of features and structure of network. Chen & al. (2010) proposed a genetic algorithm-based approach to integrate the connection weight optimization, network structure optimization and feature selection in the evolutionary procedure. The preference cost is directly incorporated into the fitness function of the genetic algorithm.

Therefore, since various objectives are to be pursued simultaneously, one possible approach to deal with this problem consists on the use of Multi Objective Evolutionary Algorithms (MOEA). Bi (2003) proposed a framework for SVM based on multi-objective optimization with the aim of minimize the risk of the classifier and the model capacity (or accuracy). Igel (2005) followed an identical approach, but replaced the objective concerning the minimization of the risk by the minimization of the complexity of the model (i.e., the number of features). Oliveira & al. (2006) used an hierarchical MOEA operating at two levels:

performing a feature selection to generate a set of classifiers (based on artificial neural networks) and selecting the best set of classifiers. Hamdani & al. (2007) used the NSGA-II (Deb & al., 2002) algorithm to optimize simultaneously the number of features and the global error obtained by a neural network classifier. Alfaro-Cid & al. (2008) applied a MOEA to take into account individually the errors of type I (false positive) and type II (false negative). Finally, Handl & Knowles (2006) studied the problem of unsupervised feature selection by formulating it as a multi-objective optimization problem.

This work proposes a methodology based on MOEA to accomplish simultaneously two objectives: the minimization of the number of features used and the maximization of the accuracy of the classifier used. Simultaneously, the parameters required by the classifier will be optimized. The evaluation of the potential solutions proposed by the MOEA during the successive generations will be made using two different classifiers, LR and SVM. This methodology has the great advantage of using simultaneously more than one criterion for the selection of the features, as no consensus exits about the best objective (measure) to use (Provost and Fawcett, 1997; Kupinski and Anastasio, 1999). The possibility of using multiple objectives constitutes the main difference concerning the traditional method used for this purpose, such as the filter approach. An important advantage of MOEAs resides on the fact that the search, for the best combination of features, is done by testing the sensitivity of the model to the value of features in an automatic way.

In this work a large database of French companies, DIANE, was used. This database is very detailed containing information on a wide set of financial ratios spanning over a period of several years. It contains up to three thousands distressed companies and about sixty thousands healthy ones.

This text is organized as follows. In section 2 the problem to solve will be explained in more detail, as well the classification methods employed and the main characteristics of the database used. In section 3 the MOEA used will be presented and described in detail. The method proposed will be applied to a case study and the results will be presented and discussed in section 4. Finally, the conclusion will be established in section 5.

2 Bankruptcy Prediction

The Problem

The bankruptcy prediction problem can be stated as follows: given a set of financial statements from a company over one, or several years, predict the probability that it will

become distressed over a given period, normally the next year or two ahead. Normally this task is performed by dividing the data into two groups: healthy and bankrupted companies, and then training a binary classifier, either supervised or unsupervised, to learn the pattern that discriminate between the two cases. Priori to train the classifiers, the database has to be "cleaned up" in order to create a well balanced and unbiased sample. Normally, a full dataset is composed by tenths of accounting features, or ratios, that measures the profitability, liabilities, cash-flow and equity of a company. These features are often correlated or confusing, so it is important to use just a handful of them. These reduced set will simplify the problem while not discharging important information. Care must be taken so that this reduction does not decrease the performance of the classifier.

The Dataset

In the present work a sample obtained from the DIANE database was selected. The initial database consisted of financial ratios of about 60 000 industrial French companies, for the years of 2002 to 2006, with at least 10 employees. From these companies, about 3000 were declared bankrupted in 2007 or presented a restructuring plan ("Plan de Redressement") to the court for approval by the creditors. No distinction between these two categories has been made since both categories signals companies in financial distress.

The dataset includes information about 30 financial ratios, as defined by COFACE, of companies covering a wide range of industrial sectors. This database contains many instances with missing values, especially concerning defaults companies. For this reason the default cases were sorted by the number of missing values and the examples with 10 missing values at most were selected. A final set of 600 default examples was obtained. In order to obtain a balanced dataset, 600 random non-default examples were selected, thus resulting in a set of 1200 examples.

The 30 financial ratios produced by COFACE are described in Table 1. These ratios allow a very comprehensive financial analysis of the firms including the financial strength, liquidity, solvability, productivity of labour and capital, margins, net profitability and return on investment. Although, in the context of linear models, some of these variables have small discriminatory capabilities for default prediction, some non-linear approaches may extract relevant information contained in these ratios to improve the classification accuracy without compromising generalization.

It is not common to consider such a large number of ratios. By construction we know that some of these ratios contain information that is highly correlated. However, it was decided to include all this information and let the feature selection algorithm decide which the best combinations of feature to achieve good accuracy are.

Table 1- Set of features considered.

Feature	**Designation**
F1	Number of employees
F2	Capital Employed / Fixed Assets
F3	Financial Debt / Capital Employed (%)
F4	Depreciation of Tangible Assets (%)
F5	Working capital / current assets
F6	Current ratio
F7	Liquidity ratio
F8	Stock Turnover days
F9	Collection period
F10	Credit Period
F11	Turnover per Employee (thousands euros)
F12	Interest / Turnover
F13	Debt Period days
F14	Financial Debt / Equity (%)
F15	Financial Debt / Cashflow
F16	Cashflow / Turnover (%)
F17	Working Capital / Turnover (days)
F18	Net Current Assets/Turnover (days)
F19	Working Capital Needs / Turnover (\)
F20	Export (%)
F21	Value added per employee
F22	Total Assets / Turnover
F23	Operating Profit Margin (%)
F24	Net Profit Margin (%)
F25	Added Value Margin (%)
F26	Part of Employees (%)
F27	Return on Capital Employed (%)
F28	Return on Total Assets (%)
F29	EBIT Margin (%)
F30	EBITDA Margin (%)

Classifiers and Methodology

The methodology proposed in this work uses different classifiers to obtain the accuracy on each set of features, while a MOEA is used to determine the best compromise between the two conflicting objectives. Two classifier algorithms will be applied: Logistic Regression (LR) and Support Vector Machines (SVM).

Logistic Regression is a well known generalized linear method, allowing the prediction of a discrete outcome (generally binary, such as success/failure), from a set of variables that may be continuous, discrete, binary, or a mix of any of these (Agresti, 1996). In the present case the LR was trained by Stochastic Gradient Descent method, which is able to estimate the maximum likelihood logistic regression coefficients from sparse input data.

Support Vector Machines (SVMs) are a set of supervised learning methods based on the use of a kernel, which can be applied to classification and regression. In the SVM a hyper-plane or set of hyper-planes is (are) constructed in a high-dimensional space. The initial step consists in transforming the data points, through the use of a non-linear mapping, into the high-dimensional space. In this case, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class. Thus, the generalization error of the classifier is lower when this margin is larger. SVMs can be seen an extension to nonlinear models of the generalized portrait algorithm developed by Vapnik (1995). In this work the SVM from LIBSVM was used (Chang, 2000).

3 Multi-Objective Optimization

Algorithm

MOEAs have been recognized in the last decade as good methods to explore and find an approximation to the Pareto-optimal front for multi-objective optimization problems. This is due to the difficulty of traditional exact methods to solve this type of problems and by their capacity to explore and combine various solutions to find the Pareto front in a single run. A MOEA must provide a homogeneous distribution of the population along the Pareto frontier, together with an improvement of the solutions along successive generations (Deb, 2001; Gaspar-Cunha & Covas, 2004). In this work, the Reduced Pareto Set Genetic Algorithm (RPSGA) is adopted (Gaspar-Cunha et al., 1997; Gaspar-Cunha, 2000; Gaspar-Cunha & Covas, 2004), where a clustering technique is applied to reduce the number of solutions on the efficient frontier. Detailed information about this algorithm can be found elsewhere (Gaspar-Cunha, 2000; Gaspar-Cunha & Covas, 2004).

Methodology for Feature Selection

In the present study the RPSGA algorithm was adapted to deal with the feature selection problem, so it can be considered as a combinatory optimization task. Concerning the definition of the decision variables, two possibilities were considered. Initially, a pure feature selection problem was analysed. In this case the parameters of the classifiers, such as type of training (holdout method or k-fold cross validation), learning rate and training fraction, for both LR and SVM, kernel type and other SVM parameters, were initially set. In a second approach, these parameters were also included as variables to be optimized. The latter approach has the advantage of obtaining in a single run the best features and, simultaneously fine tuning the classifier parameters.

For that purpose the solutions proposed by the RPSGA, each one consisting in the features selected (the initial population or generation is obtained randomly), will be evaluated by the LR or the SVM algorithms. This information is returned to the RPSGA to generate a new population of solutions based on the performance of the previous generation. More possibility of surviving is given to the fittest solutions. This approach will be illustrated in the next section.

4 Results and Discussion

Case Studies

The use of the MOEA methodology presented above is illustrated by solving the problem of finding the minimum number of features that keeps the classifiers accuracy near maximum. Accuracy is defined as the number of companies correctly classified as either bankrupted or healthy divided by the total number of companies in the test set. Table 1 presents the features and their definitions used in the DIANA database. Based of data from a given year, the classifiers are trained to predict whether the company will survive over the following year.

In the case of LR, several runs were performed using the gradient descent method and various combinations of Training Method (TrainM) - holdout method and 5-fold and 10-fold validation, Learning Rate (LearnR) and Training Fraction (TrainF), as shown in Table 2. In the case of the holdout method only part of the data (TrainF) is used to generate the classification model and the remaining set of data is used to test the classifier. While in the case of k-fold validation, all the set of the data is divided in k sets, and successively, k-1 of these sets are used to build the model and one of them is used to evaluate it. In this case the final result is an average of these k evaluations. Experiments identified as Log1 to Log6 were

used to test the influence of learning rate (comparing Log1, Log2, Log3 and Log4) and training fraction (comparing Log2, Log5 and Log6) using holdout method, while experiments Log11 to Log15, using the k-fold validation, were used to test the influence of Learning Rate (LearnR) and the number of folds (5-fold for run Log15). In the experiment Log20 the learning rate and the training fraction are considered as decision variables (i.e., they are parameters to be optimized) using the holdout validation, while in Log21 experiment (10-fold validation) only the learning rate was considered as decision variable. For these two experiments the range of variation allowed for LearnR and TrainF are shown on Table 2.

Table 2- Set of optimization for LR classifier (H: holdout; K: 10-fold validation).

Experiment	TrainM	LearnR	TrainF
Log1	H	**0.001**	0.(6)
Log2	H	**0.01**	0.(6)
Log3	H	**0.02**	0.(6)
Log4	H	**0.1**	0.(6)
Log5	H	0.01	**0.5**
Log6	H	0.01	**0.8**
Log11	K (10)	**0.001**	NA
Log12	K (10)	**0.01**	NA
Log13	K (10)	**0.02**	NA
Log14	K (10)	**0.1**	NA
Log15	**K (5)**	0.01	NA
Log20	H	[0.001; 0.1]	[0.2, 0.9]
Log21	K (10)	[0.001; 0.1]	NA

NA: Not applicable

Similarly, for the case of SVMs two different types were tested, C-SVC (Cortes and Vapnik, 1995) and μ-SVC (Schölkopf et al., 2000) using, in both cases, the Radial Basis Function (RBF) as a kernel. Different combinations of training method, training fraction and other kernel parameters (such as: γ – RBF kernel parameter for both methods; C – penalty term for C-SVM and ν – for ν-SVM) were varied, as shown in Tables 3 and 4. In these tables when the values were not shown means that the reference values (second row) are used.

Table 3- Set of optimization runs for C-SVM (H: holdout; K: 10-fold validation).

Experiment	γ	C	TrainM	TrainF
Ref. Values	**0.01**	**1**		**0.(6)**
C-svc01	0.01	-	H	-
C-svc02	0.1	-	H	-
C-svc03	1.0	-	H	-
C-svc04	10	-	H	-
C-svc07	-	10	H	-
C-svc08	-	100	H	-
C-svc09	-	1000	H	-
C-svc21	0.01	-	K	NA
C-svc22	0.1	-	K	NA
C-svc23	1.0	-	K	NA
C-svc24	10	-	K	NA
C-svc27	-	10	K	NA
C-svc28	-	100	K	NA
C-svc29	-	1000	K	NA
C-svc50	-	-	H	[0.2,0.9]
C-svc51	-	-	K	NA
C-svc52	[0.005, 10]	[1, 1000]	H	[0.2,0.9]
C-svc53	[0.005, 10]	[1, 1000]	K	NA

NA: Not applicable

Table 4- Set of optimization runs ν-SVM (H: holdout; K: 10-fold validation).

Experiment	γ	ν	TrainM	TrainF
Ref. Values	**0.01**	**0.05**		**0.(6)**
ν-svc01	0.01	-	H	-
ν -svc02	0.1	-	H	-
ν -svc03	1.0	-	H	-
ν -svc04	10	-	H	-
ν -svc10	-	0.01	H	-
ν -svc11	-	0.1	H	-
ν -svc12	-	0.5	H	-
ν -svc21	0.01	-	K	NA
ν -svc22	0.1	-	K	NA
ν -svc23	1.0	-	K	NA
ν -svc24	10	-	K	NA
ν -svc30	-	0.01	K	NA
ν -svc31	-	0.1	K	NA
ν -svc32	-	0.5	K	NA
ν-svc50	-	-	H	[0.2,0.9]
ν -svc51	-	-	K	NA
ν -svc52	[0.005, 10]	[0.01, 0.5]	H	[0.2,0.9]
ν -svc53	[0.005, 10]	[0.01, 0.5]	K	NA

NA: Not applicable

The RPSGAe was applied using the following parameters: 100 generations, crossover rate of 0.8, mutation rate of 0.05, internal and external populations with 100 individuals, limits of the

clustering algorithm set at 0.2 and the number of ranks (NRanks) at 30. These values resulted from an analysis made previously (Gaspar-Cunha, 2000; Gaspar-Cunha & Covas, 2004). Due to the stochastic nature of the initial tentative solutions several runs have to be performed (in the present case 16 runs) for each experiment. Thus, a statistical method based on attainment functions was applied to compare the final population for all runs (Fonseca & Fleming, 1996; Knowles & al., 2006). This method attributes to each objective vector a probability that this point is attaining in one single run (Fonseca & Fleming, 1996). It is not possible to compute the true attainment function, but it can be estimated based upon approximation set samples, i.e., different approximations obtained in different runs, which is denoted as Empirical Attainment Function (EAF) (Fonseca & al., 2001). The differences between two algorithms can be visualized by plotting the points in the objective space where the differences between the empirical attainment functions of the two algorithms are significant (Lopez & al., 2006).

Finally, the features selection results obtained with the method proposed here are compared with some of the features selection methods provided by the WEKA software (http://www.cs.waikato.ac.nz/~ml/weka/). This system works by providing simultaneously an attribute (feature) evaluator and a search method.

Logistic Regression

Initially, a simple example will be presented for illustration purposes. Figure 1 despite the results obtained for experiment Log 6 (Table 2) using the LR with a gradient descent and holdout method, learning rate and training fraction of 0.01 and 0.8, respectively. In this Figure the entire initial random population and the Pareto front after 100 generations can be seen. The evolution lead to a considerable gain in accuracy while decreasing significantly the number of features needed. The final population has only 4 non-dominated solutions having respectively 2, 3, 5 and 6 features. These features are (see Figure 2) F12 and F16 for the case with two features, F12, F16 and F11 for the case with 3 features, F12, F16, F11, F1 and F19 for the case with 5 features and for the case with 6 features F12, F16, F11, F1, F19 and F30. This is one of the simplest cases as same features are maintained for the solutions present in the final Pareto front.

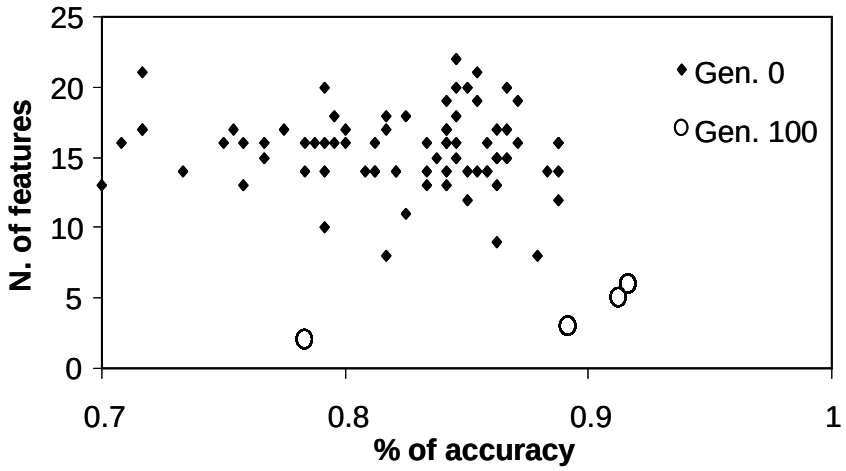

Figure 1- Initial population and Pareto front after 100 generations for experiment Log6 in Table 2.

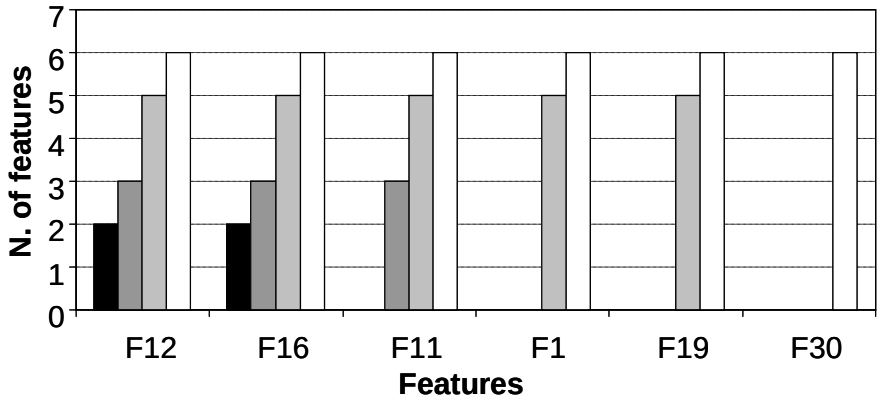

Figure 2- Features obtained for the non-dominated solutions after 100 generations (Log6, Table 2).

The different runs presented in Table 2 were compared using the EAF statistical methodology. Figure 3 shows the comparisons between experiments Log1 to Log4, *i.e.*, when the learning rate varies between 0.001 and 0.1, using the holdout methdo and a training fraction of 0.(6). Objective 1 is the accuracy and Objective 2 the number of features. First row compares Log 1 with Log 2. In this case Log2 is slightly better since more black dots appear in the graph located at right, since these points indicate that Log1 attain points with an higher frequency, being the the amount of the difference encoded in grey-scale, i.e., the darker the points the stronger are the observed differences. The two extreme lines indicated in the plot connect the best points ever found by the two algorithms compared (grand best) and the points dominated in any run (grand worst); the line in the middle corresponds to the boundary

of the region that is obtained in 50% of the runs of each algorithm, that is, it represents the median attainment function surface (on the right side for Log1, on the left side the median for the Log2). Second row compares Log2 with Log3. In this case Log2 is also slightly better. Finally, the last row allows one to conclude that the best value for Learning Rate is 0.01 (*i.e.*, Log2).

Best performance is attained for experiment Log20 where the training method is holdout and all parameters are optimized simultaneously. As shown in Figure 4, Log6 is better than Log2 but lags Log20 (which is better than Log21) in performance. The full set of results can be found at www.dep.uminho.pt/agc/results.

Figures 5 to 7 shows the results obtained for a single run (of 16 runs) of experiment Log20 (Table 2). Figure 5 shows the Pareto fronts at generations 0, 50 and 100. Again, the evolution leads to a considerable gain in accuracy while reducing the number of features. As can be seen the final population identified 9 non-dominated solutions (or optimal solutions) having, respectively 1, 2, 3, 4, 9, 11, 12, 13 and 14 features. The features selected for each one of these situations are identified in Figure 6. For example, the solution with one feature selected feature F11 with accuracy approximately equal to 67% (Figure 5). The solution with 2 features selected F11 and F30, but now the accuracy is much better (88%). The best accuracy (94%) is accomplished for the solution with the higher number of features selected (14 features: F1, F2, F3, F4, F6, F8, F9, F11, F12, F15, F16, F21, F26 and F30). However, the decision maker can select a compromise solution with 9 features and accuracy of 92.6%, or a solution with 4 features (F3, F8, F11 and F30) and accuracy equal to 92%.

The approach proposed can be extremely useful to the analyst as, usually, he does not have access to such large number of features. Moreover, contrary to other feature selection approaches, the present method provides extra information about the usefulness of using extra features.

Figure 7 shows the values obtained for the learning rate and the training fraction for the same 9 non-dominated solutions. For example, for the solution with 4 features these values are 0.0017 and 0.895, respectively.

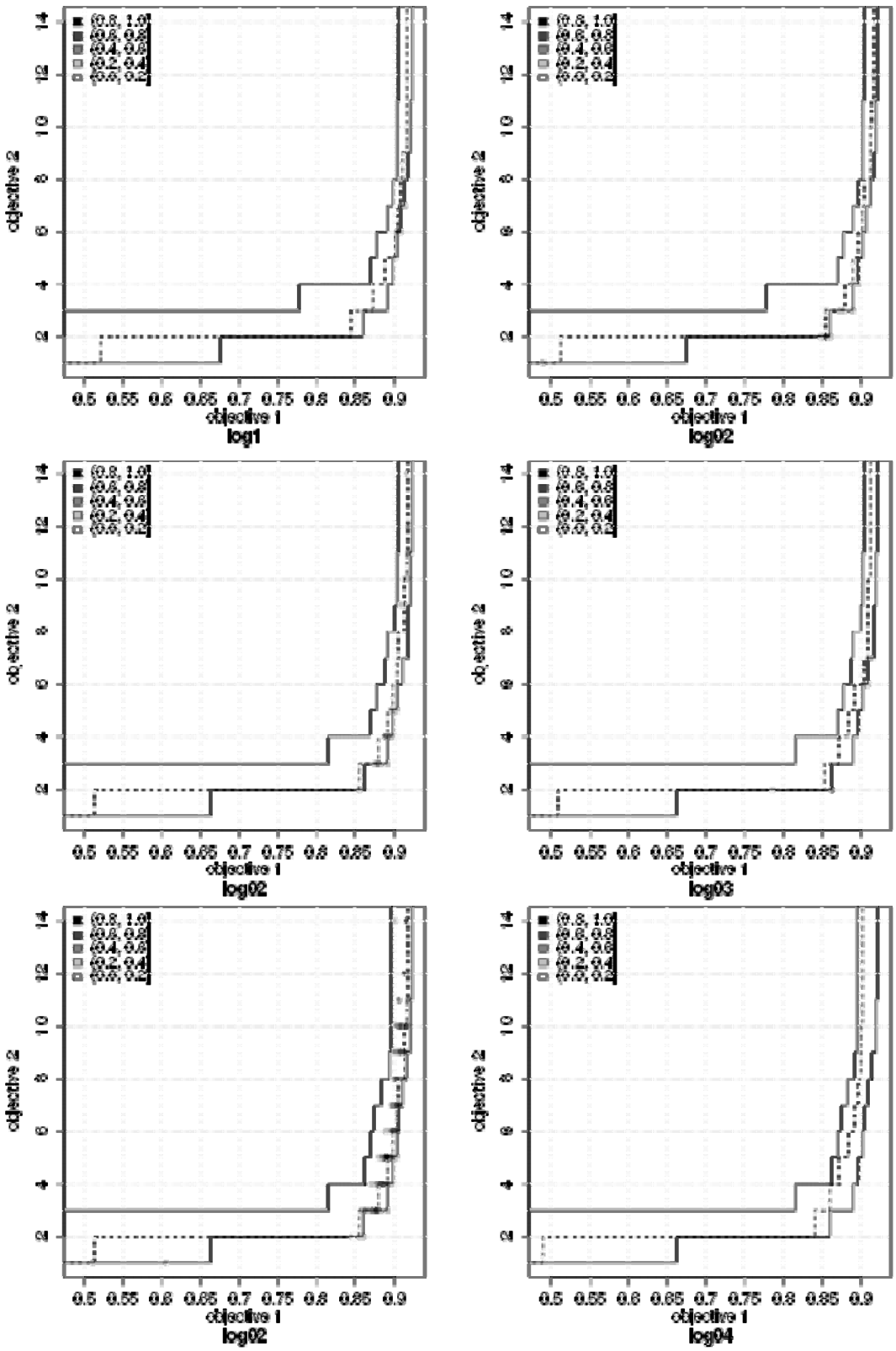

Figure 3- EAFs differences between experiments: top: Log1 and Log2; middle: Log2 and Log3; bottom: Log2 and Log4.

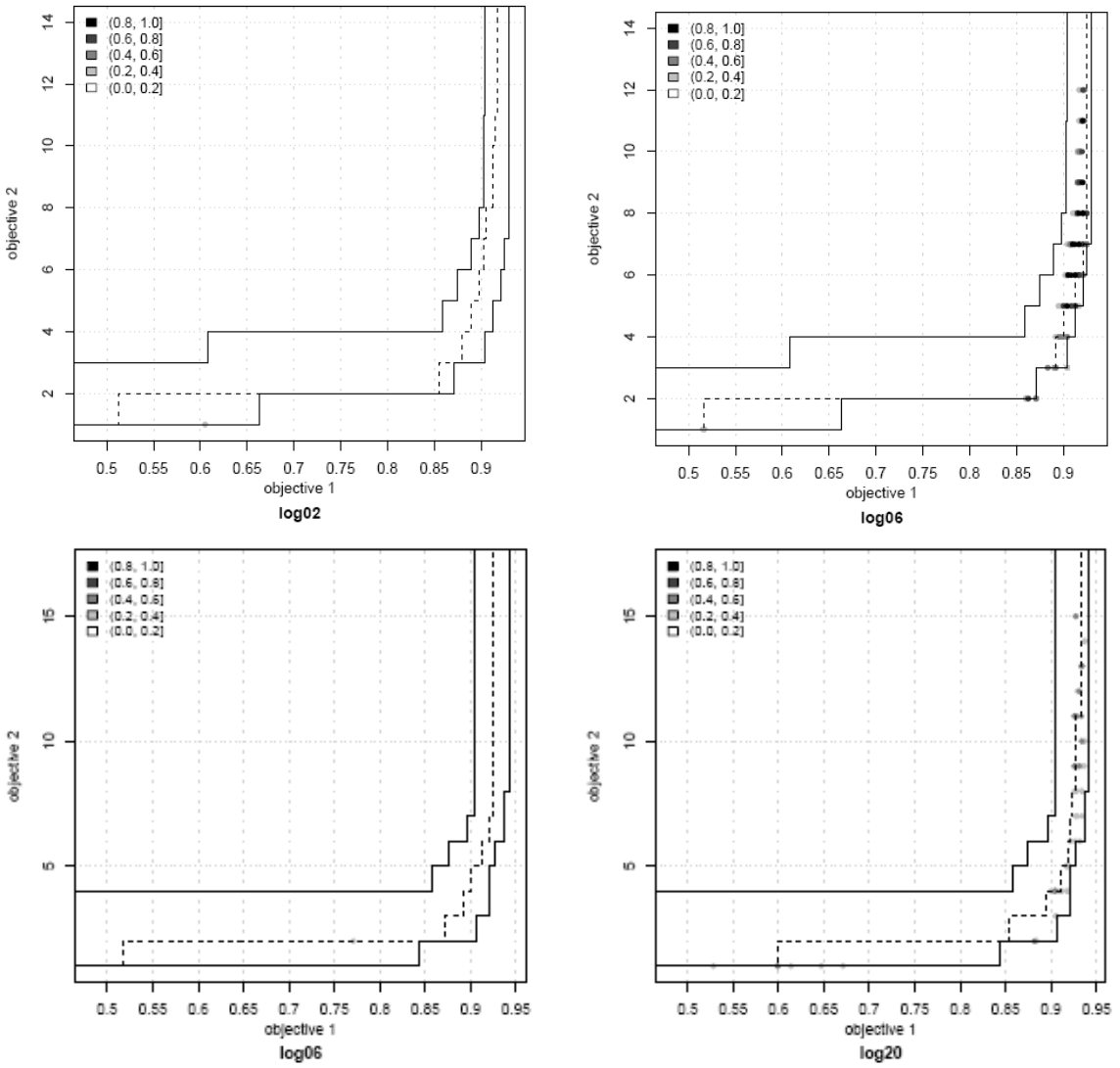

Figure 4- EAFs differences between experiments: top: Log2 and Log6; bottom: Log6 and Log20.

Finally, an analysis about the significance of the features selected will be made. For the present case the accuracy of 92% accomplished for the solution with 4 features seems to be sufficient, since adding more variables does not increase accuracy significantly. The selected features correspond to:

F3: Financial Debt / Capital Employed (%). This measures the capital structure ratio, i.e., the amount of financial debt in relation to the total amount of capital invested in the

firm. The higher the ratio the closest to failure the company is. Several author used this ratio as a good predictor of bankruptcy (Alfaro-Cid & Castillo, 2008);

F8: Stock Turnover days. This is a ratio that measures the number of days invested in inventories. It measures the efficiency of the firm in the conversion of inventories into revenues. Companies with a low ratio may denote difficulty in selling their stock in comparison with other of the same industry. Inventories are part of the total assets and some author may prefer to use sales/total assets ratio such as Atiya (2001) and Altman & al. (1968, 1977);

F11: Turnover per Employee (thousands Euros) is a measure of employee profitability. If all the other ratios are constant, a higher productivity decreases the probability of bankruptcy. This has been used as an indicator for bankruptcy prediction in several studies (Neves & Vieira, 2006);

F30: EBITDA Margin (%) measures the Earnings Before Interest, Taxes, Depreciation and Amortization for the total revenues. This is a measure of operational profitability of the firm. This ratio is commonly used by financial analysts and investors to benchmark profitability within a given industry and to understand the effects of competition in operating profitability.

Therefore, all components of the financial management structure of a firm are included to make bankruptcy prediction: the capital-debt structure, measured by financial debt/capital employed ratio; liquidity, measured by stock turnover days; activity, measured by turnover per employee; profitability, measured by EBITDA margin

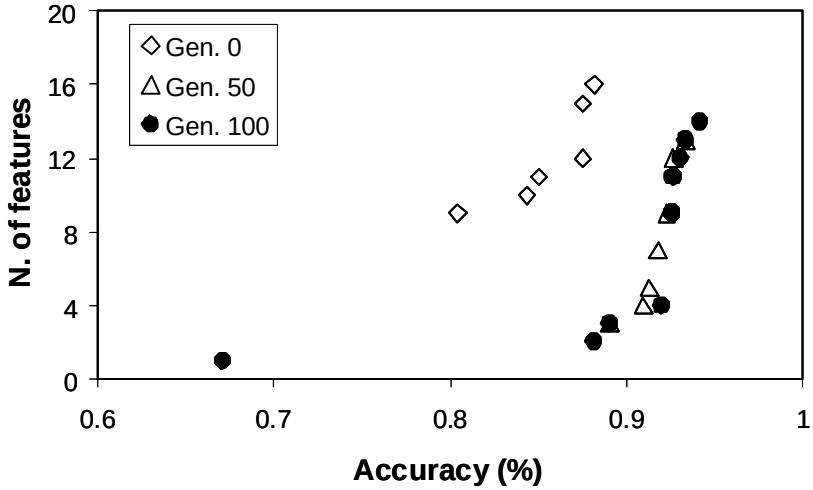

Figure 5- Evolution of the Pareto front along the successive generations for a single run of experiment Log20.

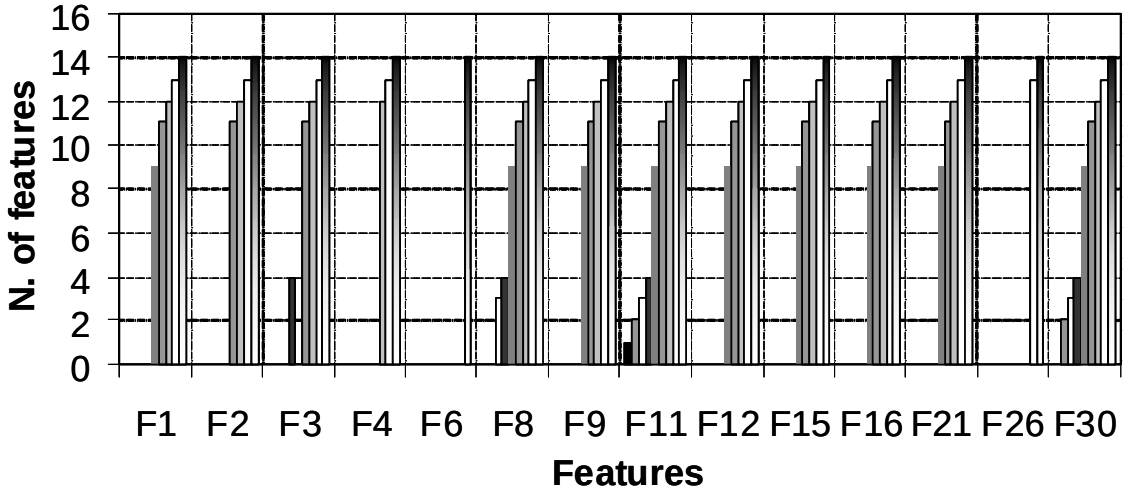

Figure 6- Features obtained for the non-dominated solutions after 100 generations for a single run of experiment Log20.

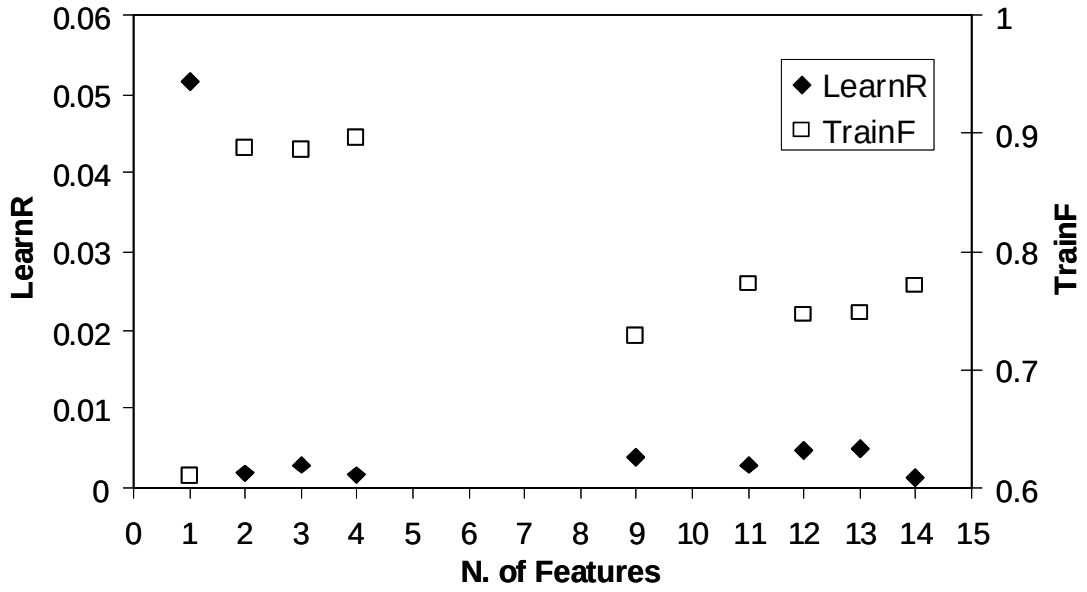

Figure 7- Training fraction and accuracy for the 5 non-dominated solutions for a single run of experiment Log20.

Support Vector Machines (SVM)

Identical analysis was made for the case where both types of SVM classifiers (C-SVC and μ-SVC) are used (Tables 3 and 4). Again, details of the comparative results can be found at www.dep.uminho.pt/agc/results.

The comparison between the results obtained for the experiments with C-SVC (experiments of Table 3), using the EAFs, allow to conclude that the best results is obtained when the classifier parameters are optimized simultaneously (i.e., experiment c-svc53). A similar analysis was carried out for the case of ν-SVC type (experiments of Table 4). Identical results are obtained, i.e., the best solutions are obtained when the classifier parameters are optimized simultaneously (i.e., experiment ν-svc53). Figure 8 compares the performance between the best experiments for each type of SVM. It is possible to conclude that their performance is very similar.

Tables 5 to 7 show the optimal results obtained for a single run of experiments c-svc50, c-svc53 and ν-svc53, respectively. For experiment c-svc50 (Table 5) accuracy above 90% only is accomplished using 8 features (F1, F3, F7, F9, F11, F14, F16 and F25) and the training fraction equal to 0.477. In the case of experiments c-svc53 (Table 6) and ν-svc53 is possible to obtain accuracy higher than 90% using only three features. This is possible at expenses of optimizing simultaneously the classifier parameters. Therefore, best results; i.e., the simultaneous increase of accuracy and decrease of the number of features, are accomplished when the classifier parameters are optimized simultaneously with the features to be selected.

Table 5- Optimal results for a single run of experiment c-svc50 (γ=0.01, C=1).

N. Features	**Accuracy (%)**	**TrainF**	**Features**
3	49.9	0.269	F3, F9, F11
4	85.2	0.527	F1, F3, F9, F30
5	87.2	0.479	F1, F3, F9, F11, F30
6	87.8	0.428	F1, F3, F7, F9, F11, F30
7	88.7	0.446	F1, F3, F7, F9, F11, F16, F25
8	**90.4**	**0.477**	**F1, F3, F7, F9, F11, F14, F16, F25**
9	90.8	0.475	F1, F3, F7, F9, F11, F13, F14, F16, F25
10	91.4	0.496	F1, F3, F7, F9, F11, F13, F14, F15, F16, F25
11	91.6	0.487	F1, F3, F7, F9, F10, F11, F13, F14, F15, F16, F25

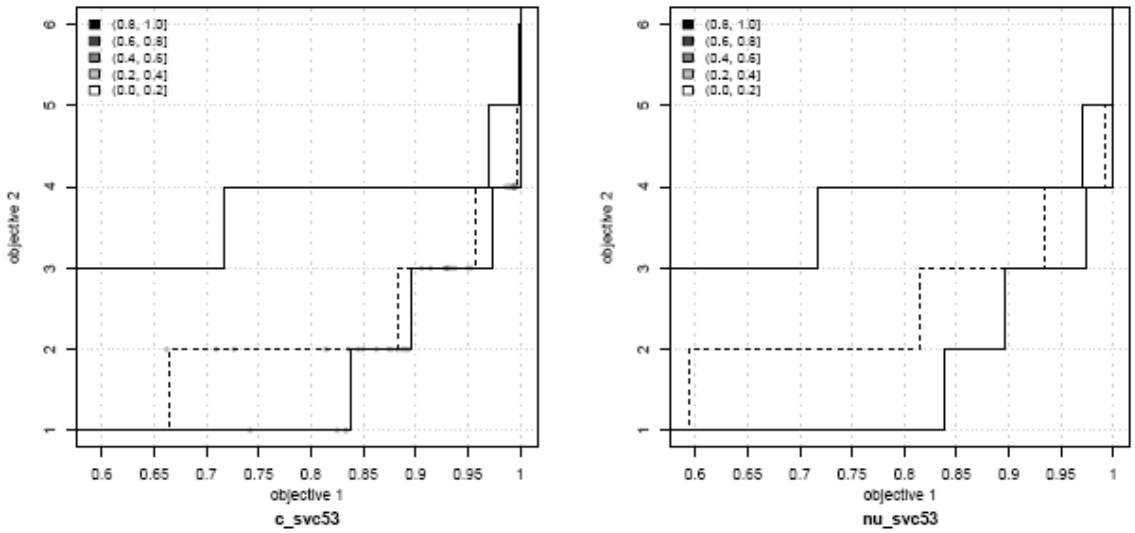

Figure 8- EAFs differences between experiments using C-SVC and ν-SVC.

Table 6- Optimal results for a single run of experiment c-svc53.

N. Features	Accuracy (%)	γ	C	Features
1	62.0	7.2	875	F4
2	84.5	7.7	905	F4, F30
3	93.0	10.0	995	F4, F7, F30
4	98.9	10.0	977	F4, F7, F22, F30
5	100.0	9.3	959	F4, F7, F11, F22, F30

Table 7- Optimal results for a single run of experiment ν-svc53.

N. Features	Accuracy (%)	γ	ν	Features
1	59.3	7.91	0.492	F2
2	66.2	7.55	0.486	F2, F22
3	91.4	9.91	0.232	F2, F16, F22
4	99.3	9.88	0.079	F2, F8, F16, F22
5	100.0	6.63	0.043	F1, F2, F8, F16, F22

Comparative Study

In this section the results obtained above were compared with the combinations of features evaluation and search methods provided in the WEKA software (version 3.6.2) shown in Table 8. In all cases 10-fold cross validation is used, together with the default parameters for each one of the methods tested as provided by the WEKA software. The description of each one of the methods is presented below (Witten & Frank, 2005):

For the search methods:

- GreedyStepwise: Greedy hill-climbing without backtracking;
- Ranker: Rank individual attributes (not subsets) according to their evaluation;
- GeneticSearch: Search using a simple genetic algorithm;

For the feature evaluators:

- CfsSubset: Consider the predictive value of each subset evaluator attribute individually, along with the degree of redundancy among them;
- ConsistencySubset: Project training set onto attribute set and measure consistency in class values;
- ChiSquared: Compute the chi-squared statistic of each attribute evaluator attribute with respect to the class;
- GainRatio: Evaluate attribute based on gain ratio;
- SVM: Use a linear support vector machine to determine the value of attributes.

Table 8- Combination of feature evaluation and search methods.

Identification	Search Method	Feature Selection
W1	Greedy Stepwise	CfsSubset
W2	Greedy Stepwise	ConsistencySubset
W3	Ranker	ChiSquared
W4	Ranker	GainRatio
W5	Ranker	SVM
W6	Genetic Search	CfsSubset
W7	Genetic Search	ConsistencySubset

Table 9 presents a summary of the results obtained using these combinations (W1 to W7, of Table 8) and the results obtained using the methodolgy proposed in this work (i.e., the results shown in Tables 5 to 7), which are identified in Table 9 as T5 to T7. In this latter case, only the solutions with 5 features were considered. The search methods Greedy Stepwise and Genetic Search select the best features without categorizing the importance of each one (identified by an X in Table 9) while the Ranker search method ranks the features by order of importance (identified by a number in Table 9). Clearly features F1, F11 and F16 are the most frequently selected. The significance of features F1 and F11 was already identified above.

Cash flow to turnover (F16) is a ratio that measures the overall financial performance of the company. Cash flow is considered the lifeblood of any business. As a consequence, in order

to improve performance, the management team must develop programs that improve the cash flow performance while ensuring that operations are aligned with the strategic objectives. Thus, the higher this ratio is, the better the performance of the company, and the lower the probability of bankruptcy or financial distress.

Finally, the accuracy of these solutions was calculated using two different sets of SVM parameters: the reference parameters values identified in Table 3 (C=1.0 and γ=0.01) and the optimized values resulted from run c-svc53 and shown in the last row of Table 7 (C=959.0 and γ=9.3). The accuracy values of the set of features selected by the method proposed here (T5 to T7) are very similar to that of solutions W1, W2 and W6, but in these latter cases the number of features is higher. The comparison with the cases where the number of features is equal (W3, W4, W5 and W7) shows that generally the accuracy values for solutions T5 to T7 are higher. However, a particular attention must be paid to solution W3, since in this case the accuracy when parameters C=1.0 and γ=0.01 are used is higher than that of all the other solutions. This allows, in fact, an important conclusion. The performance achieved by a search method using a particular feature evaluator (in the present case the SVM) can depend strongly on the sensitivity of the evaluator to its own parameters. By other words, the solutions obtained after a search process must be robust concerning changes in decision variables values, since in the present case the evaluator parameters are also considered parameters to optimize. Robustness can be seen as the inverse of the sensitivity. For the concept of robustness, as presented here, see for example the work of (Gaspar-Cunha & Covas, 2008; Ferreira & al., 2008). It is, certainly, true that the same type of robustness analysis can be applied to changes on the features values. This is an important opportunity for further research and to improve the performance of feature selection methodologies.

Table 9- Features selected using the WEKA software and runs c-svc50 (T5-Table 5), c-svc3 (T6-Table 6) and ν-svc53 (T7-Table 7) and corresponding accuracy.

Feature	Designation	W1	W2	W3	W4	W5	W6	W7	T5	T6	T7
F1	Number of employees	X	X			3	X		X		X
F2	Capital Employed / Fixed Assets										X
F3	Financial Debt / Capital Employed (%)								X		
F4	Depreciation of Tangible Assets (%)	X	X				X			X	
F5	Working capital / current assets										
F6	Current ratio										
F7	Liquidity ratio	X					X			X	
F8	Stock Turnover days							X			X
F9	Collection period		X					X	X		
F10	Credit Period										
F11	Turnover per Employee (thousands €)	X	X			2	X	X	X	X	
F12	Interest / Turnover		X					X			
F13	Debt Period days										
F14	Financial Debt / Equity (%)	X					X				
F15	Financial Debt / Cashflow	X			1	4	X				
F16	Cashflow / Turnover (%)	X	X	1	2	1	X				X
F17	Working Capital / Turnover (days)										
F18	Net Current Assets/Turnover (days)										
F19	Working Capital Needs / Turnover (\)										
F20	Export (%)										
F21	Value added per employee		X					X			
F22	Total Assets / Turnover									X	X
F23	Operating Profit Margin (%)			3	5						
F24	Net Profit Margin (%)										
F25	Added Value Margin (%)										
F26	Part of Employees (%)	X									
F27	Return on Capital Employed (%)	X					X				
F28	Return on Total Assets (%)	X		2	3	5					
F29	EBIT Margin (%)			5							
F30	EBITDA Margin (%)			4	4				X	X	
N. of features selected		**10**	**7**	**5**	**5**	**5**	**8**	**5**	**5**	**5**	**5**
Accuracy (%) – C=1.0; γ=0.01		**56.1**	**56.0**	**66.3**	**59.0**	**55.9**	**55.9**	**54.6**	**55.8**	**55.9**	**56.0**
Accuracy (%) – C=959; γ=9.3		**100**	**100**	**95.3**	**95.7**	**99.8**	**100**	**99.7**	**100**	**100**	**99.9**

5 Conclusions

In this study MOEA were used to optimize the bankruptcy prediction problem. Two different and complementary, classifier algorithms have been used: Logistic Regression and Support Vector Machines. The proposed methodology provides a powerful solution, not only reducing the necessary features but also enhancing the representation of the solution by making

available to the decision maker relevant information. The algorithm does not only provide the best features to be used but, also, with the best parameters of the classifier.

The most important characteristic of the MOEA strategy is the possibility of the decision maker to have multiple Pareto optimal solutions to perform the final analysis.

An important conclusion from this work is that the best performance is only attained when the classifier parameters are optimized simultaneously with the features selection, since the classifier performance is strongly dependent on these parameters.

Finally, further work is needed to take into account the robustness of the solutions obtained against changes on the decision variables values, i.e., features and/or classifier parameters.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley.

Alfaro-Cid, E., Castillo, P.A., Esparcia, A., Sharman, K., Merelo, J.J., Prieto, A., Mora, A.M., & Laredo, J.L.J. (2008). Comparing Multiobjective Evolutionary Ensembles for Minimizing Type I and II Errors for Bankruptcy Prediction, *Congress on Evolutionary Computation - CEC'2008* (pp. 2907-2913). Washington, USA.

Altman, E.I. (1968). Financial Ratios, Discriminant Analysis and the prediction of Corporate Bankruptcy, *Journal of Finance*, 23(4), 589-609.

Altman, E.I., Haldeman, R., & Narayanan, P. (1977). Zeta Analysis: A New Model to Identify Bankruptcy Risk of Corporations, *Journal of Banking & Finance*, 1(1), 29-54.

Atiya, F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks*, 12(4), 12-16.

Bi, J. (2003). Multi-Objective Programming in SVMs. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC.

Chen, N., Ribeiro, B., Vieira, A., Duarte, J. & Neves, J. (2010). Hybrid Genetic Algorithm and Learning Vector Quantization modeling for Cost-Sensitive Bankruptcy Prediction, *Proceedings of the International Conference on Machine Learning and Computing (ICMLC 2010)*, Bangalore, India.

Chang, C.-C. & Lin, C.-J. (2000). *LIBSVM a library for support vector machines* (Tech. Rep.). Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

Charitou, A., Neophytou, E., & Charalambous, C. (2004). Predicting corporate failure: empirical evidence for the UK, *European Accounting Review*, 13(3), 465–497.

Coats, P.K., & Fant, L.F. (1993). Recognizing Financial Distress Patterns Using a Neural Network Tool, *Financial Management*, 22(3), 142-155.

Cortes, C and Vapnick, V. (1995). Support-vector Network, *Machine Learning*, 20(), 273-297.

Deb, K. (2001). *Multi-Objective Optimization using Evolutionary Algorithms*, New York, Wiley.

Deb, K., Pratap. A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multi-objective genetic algorithm: NSGA-II, *IEEE Transaction on Evolutionary Computation*, 6(2), 181-197.

Eisenbeis, R.A. (1997). Pitfalls in the Application of Discriminant Analysis in Business, Finance and Economics, *Journal of Finance*, 32 (3), 875-900.

Fan, A., & Palaniswami, M. (2000). Selecting bankruptcy predictors using a support vector machineapproach, In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, 2000. IJCNN 2000, Vol.6 (pp. 354-359).

Ferreira, F., Fonseca, C., Covas, J.A., & Gaspar-Cunha, A. (2008). Evolutionary Multi-Objective Robust Optimization, *in Advances in Evolutionary Algorithms*, ISBN 978-3-902613-32-5, Witold Kosińsk (Ed), I-Tech Education and Publishing, Vienna, Austria, (http://www.intechweb.org/book.php?id=68&content=title&sid=1), 261-278,

Fonseca, C., & Fleming, P.J. (1996). On the performance assessment and comparison of stochastic multiobjective optimizers. *Parallel Problem Solving from Nature-PPSN IV*, Lectures Notes in Computer Science, Springer-Verlag, 584-593.

Fonseca, V.G., Fonseca, C., & Hall, A. (2001). Inferential performance assessment of stochastic optimisers and the attainment function. *Evolutionary Multi-Criterion Optimization, Lecture Notes in Computer Science*, Springer-Verlag, 213-225.

Gaspar-Cunha, A., Oliveira, P., & Covas, J.A. (1997). Use of Genetic Algorithms in Multicriteria Optimization to Solve Industrial Problems, Seventh International Conference on Genetic Algorithms, Michigan, USA.

Gaspar-Cunha, A., &. Covas, J.A. (2004). - RPSGAe - A Multiobjective Genetic Algorithm with Elitism: Application to Polymer Extrusion. In X. Gandibleux, M. Sevaux, K. Sörensen & V. T'kindt (Eds.), Metaheuristics for Multiobjective Optimisation: Vol. 535 in Lecture Notes in Computer Science, (pp. 221-249). Berlin, Springer Verlag.

Gaspar-Cunha, A. (2000). *Modelling and Optimization of Single Screw Extrusion*, Published doctoral dissertation, In Gaspar-Cunha, A. (2009), Modelling and Optimization of Single Screw Extrusion: Using Multi-Objective Evolutionary Algorithms, Köln, Germany: Lambert Academic Publishing.

Gaspar-Cunha, A. & Covas, J.A. (2008). Robustness in Multi-Objective Optimization using Evolutionary Algorithms, *Computational Optimization and Applications*, 1(39), 75-96.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. (2006). *Feature Extraction Foundations and Applications*, Springer.

Hamdani, T.M., Won, J.-M., Alimi, A.M., & Karray, F. (2007). Multi-objective Feature Selection with NSGA II, In B. Beliczynski, A. Dzielinski, M. Iwanowski, B. Ribeiro (Eds.), *Adaptive and Natural Computing Algorithms*, 8th International Conference, ICANNGA 2007, Part I, Springer-Verlag. Lecture Notes in Computer Science Vol. 4431, pp. 240-247.

Handl, J. and Knowles, J. (2006) Feature subset selection in unsupervised learning via multiobjective optimization. *International Journal of Computational Intelligence Research*, 2 (3): 217-238.

Igel, C. (2005). Multi-Objective Model Selection for Support Vector Machines. In C.A. Coello Coello et al. (Eds.), *EMO 2005*, LNCS 3410 (pp. 534-546). Springer-Verlag Berlin Heidelberg.

Knowles, J.D., Thiele, L., & Zitzler, E. (2006). A tutorial on the performance assessment of stochastic multiobjective optimizers. TIK-Report No. 214.

Kumar, P.R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. European Journal of Operational Research, Vol. 180, 1–28.

Kupinski, M.A. and Anastasio, M.A. (1999). Multiobjective Genetic Optimization of Diagnostic Classifiers with Implications for Generating Receiver Operating Characteristics Curves, *IEEE, Transations on Medical Imaging, 18(8): 675-685.*

López-Ibañez, M., Paquete, L., & Stützle,T. (2006). Hybrid population based algorithms for the bi-objective quadratic assignment problem. *Journal of Mathematical Modelling and Algorithms*, 5(1), 111-137.

Martin, D. (1977). Early Warning of Bank Failure: A Logit Regression Approach, *Journal of Banking and Finance*, 1(3), 249-276.

Neves, J.C. & Vieira, A.S. (2006). Improving Bankruptcy Prediction with Hidden Layer Learning Vector Quantization, *European Accounting Review*, 15(2), 253-271.

Oliveira, L.S., Morita, M., & Sabourin, R. (2006). Feature Selection for Ensembles Using the Multi-Objective Optimization Approach. *Studies in Computational Intelligence* (SCI), Vol. 16, 49-74.

Provost, F and Fawcett, T. (1997). Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions. Proceedings of Third International Conf. on Knowledge Discovery and Data Mining, AAAO Press, Menlo Park, CA.

Rekba, G.A., Annapoorani, R. & Vijayalakshmi, G.A. (2004). Performance analysis of a statistical and an evolutionary neural network based classifier for the prediction of industrial bankruptcy. *In IEEE Conference on Cybernetics and Intelligent Systems*, Vol. 2, 1033–1038.

Ribeiro, B., Vieira, A., Duarte, J., Silva, C., Neves, J.C., Liu, Q. & Sung, A.H (2009). Learning manifolds for bankruptcy analysis. In M. Köppen & al. (Eds.), editors, *Int. Conf. on Neural Information Processing*, Vol. 5506, 722–729, Berlin Heidelberg. Lecture Notes in Computer Science (LNCS), Springer-Verlag.

Ribeiro, B., Silva, C., Vieira, A., & Neves, J.C. (2009b). Extracting discriminative features using non-negative matrix factorization in financial distress data. In M. Kolehmainen et al. (Eds.), editor, *Int Conf on Adaptive and Natural Computing Algorithms*, Vol. 4432, 537–547, Berlin Heidelberg, April 2009. Lecture Notes in Computer Science (LNCS), Springer-Verlag.

Shin, K.S., Lee, T.S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, Vol. 28, 127–135.

Tan, C.N.W., & Dihardjo, H. (2001). A Study on Using Artificial Neural Networks to Develop an Early Warning Predictor for Credit Union Financial Distress with Comparison to the Probit Model, *Managerial Finance*, 27(4), 56-77.Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer Verlag.

Thomas, E.M. (2007). Altman's 1968 bankruptcy prediction model revisited via genetic programming: New wine from an old bottle or a better fermentation process? *Journal of Emerging Technologies in Accounting*, Vol. 4, 87–101.

Verikas, A., Kalsyte, Z., Bacauskiene, M. & Gelzinis, A. (2009). Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. *Soft computing - a fusion of foundations, methodologies and applications, (Online)*, September 2009.

Vieira, A.S., Duarte, J., Ribeiro, B. & Neves, J.C. (2009). Accurate Prediction of Financial Distress of Companies with Machine Learning Algorithms. Proceedings of the *ICANNGA 2009* Conference, Kuopio, Springer Lecture Notes on Computer Science.

WEKA Software (2010). http://www.cs.waikato.ac.nz/~ml/weka/.

Witten, I.H. & Frank, E. (2005). Data mining : practical machine learning tools and techniques. 2nd ed., San Francisco, Morgan Kaufmann.

Yang, D.T. (1999). Urban-biased policies and rising income inequality in China. *American Economic Review Papers and Proceedings*, 89, (2)306–310.