Multi-Objective Evolutionary Algorithms for Feature Selection: Application in Bankruptcy Prediction

António Gaspar-Cunha[1], Fernandes Mendes[1], João Duarte[2], Armando Vieira[2], Bernardete Ribeiro[3], André Ribeiro[4], and João Neves[4]

[1] IPC/I3N - Institute of Polymers and Composites, University of Minho, Guimarães, Portugal
{agc,fmendes}@dep.uminho.pt

[2] Department of Physics, Instituto Superior de Engenharia do Porto, R. S. Tomé, 4200 Porto, Portugal
{jmmd,asv}@isep.ipp.pt

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

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

Abstract. A Multi-Objective Evolutionary Algorithm (MOEA) was adapted in order to deal with problems of feature selection in datamining. The aim is to maximize the accuracy of the classifier and/or to minimize the errors produced while minimizing the number of features necessary. A Support Vector Machines (SVM) classifier was adopted. Simultaneously, the parameters required by the classifier were also optimized. The validity of the methodology proposed was tested in the problem of bankruptcy prediction using a database containing financial statements of 1200 medium sized private French companies. The results produced shown that MOEA is an efficient feature selection approach and the best results were obtained when the accuracy, the errors and the classifiers parameters are optimized.

Keywords: Multi-Objective, Evolutionary Algorithms, Feature Selection, Bankruptcy Prediction.

1 Introduction

The problem of feature selection from databases with high amount of data is of crucial importance, specially when dealing with problems such as bankruptcy prediction given its consequences for banks, insurance companies, creditors and investors. Thus, the ability to discriminate between faithful customers from potential bad ones is thus crucial for commercial banks and retailers [1].

Traditional methods used to study this problem, such as discriminant analysis [2] and Logit and Probit models [3], 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. In non-linear models the choice of the regression function creates a bias that restricts the outcome, they are very sensitive to exceptions, and most conclusions have an implicit Gaussian distribution on data, which is inappropriate in many cases. To overcome these problems other approaches have been applied recently in the problem of bankruptcy classification, such as Artificial Neural Networks (ANN) [4,5], Evolutionary Algorithms (EA) and Support Vector Machines (SVM) [6]. Usually, complementary tools based on ANN, EA and SVM are used to classify credit risk. In some studies it is shown that ANN outperforms discriminant analysis in bankruptcy prediction [5,6,7,8,9]. Moreover these promising results, it is generally recognized that further research is needed [10].

Due to the large number of variables usually present, and due to the high correlation between these variables, it is of fundamental importance the existence of a feature selection method able to reduce the number of features considered for analysis [11]. A possible approach to deal with this problem consists on the use of Multi Objective Evolutionary Algorithms (MOEA). Bi in [12] 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 in [13] 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 et al. in [14] 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 et al. in [15] used the NSGA-II [16] algorithm to optimize simultaneously the number of features and the global error obtained by a neural network classifier. Alfaro-Cid et al. in [17] applied a MOEA to take into account individually the errors of type I (false positive) and type II (false negative). Finally, Handl and Knowles in [18] studied the problem of unsupervised feature selection by formulating them as a multi-objective optimization problem.

This work follows the main ideas of a previous work proposed by the authors, were a methodology based on MOEA was used to accomplish simultaneously two objectives: the minimization of the number of features used and the maximization of the accuracy of the classifier used [19]. In the present case different accuracy measures, such as maximization of the F measure (F_m) and the minimization of errors (type I and type II), will be tested. Simultaneously, the parameters required by the classifier will be optimized. This is an important issue since parameter tuning is not an easy task [20].

2   Bankruptcy Prediction

The Problem and Dataset. In the bankruptcy prediction problem the aim is to infer the probability that a company will become distressed, over a specified period, given a set of financial statements. This can be done from over one, or several years. In general 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. Often, the database needs some previous treatment, prior to training the classifiers, in order to create a well balanced and unbiased sample. Usually, a full dataset is composed by tenths of accounting features, or ratios, measuring different characteristics of a company (e.g., the profitability, liabilities, cash-flow and equity). These features are often highly correlated and confusing, being important to use just some of them. This will simplify considerably the problem. However, in order not to loose important information, special care must be taken during the process of reducing the number of features. Thus the performance of the classifier will not decrease. It is clear that these ideas can be generalized to other type of classification problems than bankruptcy prediction.

In this 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 (Table 1), of companies covering a wide range of industrial sectors.

Classification and Metrics. In the methodology proposed a SVM classifier will be used while a MOEA is used to determine the best compromise between the two and/or the three conflicting objectives. 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. In this case, a

Table 1. Set of features considered (as defined by COFACE)

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	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
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

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 in [21]. In this work the SVM from LIBSVM was used [22]. The selection of the right kernel, as well the definition of the best kernel parameters, is of primordial importance for the SVM performance [13]. In the present study only the C-SVC method using as kernel the Radial Basis Function (RBF) was tested [22]. Thus, two different SVM parameters are to be selected carefully: the regularization parameter (C) and the kernel parameter (γ). Simultaneously, another important parameter for training the SVM is the Learning Rate (LR), which was also taking into account in this study. Another important issue concerns the performance metrics used to evaluate the learning methods [23,24,25]. The most straightforward way is to use the accuracy given by the ratio between the number instances correctly evaluated and the total number of instances, i.e.:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where, TP are the positives correctly classified, TN are the negatives correctly classified, FP are the positives incorrectly classified and FN are the negative incorrectly classified.

It is also important to know the level of the errors accomplished by the classifier, mainly on problems where the existence of errors is critical. Two different error types can be defined, type I and type II, given respectively by:

$$e_I = \frac{FP}{FP + TN} \tag{2}$$

and

$$e_{II} = \frac{FN}{FN + TP} \tag{3}$$

Another traditional way to evaluate the information is using the sensitivity or recall (R) and the precision (P) of the classifier:

$$R = \frac{TP}{TP + FN} \tag{4}$$

and

$$P = \frac{TP}{TP + FP} \tag{5}$$

In the present work F_m, which represents the harmonic mean of R and P, was adopted here to evaluate globally the classifier:

$$F_m = \frac{2PR}{P + R} \tag{6}$$

The selection of the best learning algorithm to use and the best performance metric to measure the efficiency of the classifier is nowadays the subject of many studies [23,25].

3   Multi-Objective Evolutionary Algorithms

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 [26,27]. In this work, the Reduced Pareto Set Genetic Algorithm (RPSGA) is adopted [27,28], 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 [27,28]. In the present study the RPSGA algorithm was adapted to deal with the features 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 analyzed. In this case the parameters of the classifier, such as type of training (in the present study only k-fold cross validation was used) and learning rate and the SVM parameters (C and γ), 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. This approach will be illustrated in the next section.

4   Results and Discussion

4.1   Case Studies

The MOEA methodology presented above will be used in a problem of finding the minimum number of features while maximizing F_m and minimizing e_I. Based on the data from a given year, the classifier is trained to predict whether the company will survive over the following year. Table 2 shows the different experiments tested. In all cases the C-SVC method using as kernel the Radial Basis Function (RBF) and 10-fold validation training method were used. First four experiments, using only as decision variables the features, were performed (experiments c-svc1 to c-svc04 in Table 2). In this case the Learning Rate (LR), C and γ are set to 0.01, 1 and 10, respectively. In experiments c-svc11 to c-svc14 and c-svc21, LR, C and are also considered as decision variables (i.e., they are parameters to be optimized). The range of variation allowed for these variables is shown on Table 2. 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 (N_Ranks) at 30. Due to the stochastic nature of the initial population several runs were performed (in the present case 16 runs) for each experiment. Thus, a statistical method based on attainment functions

Table 2. Experimental setup

Exp.	LR	C	γ	Objectives
c-svc1	0.01	1	10	$NF + F_m$
c-svc2	0.01	1	10	$NF + e_I$
c-svc3	0.01	1	10	$NF + e_I I$
c-svc4	0.01	1	10	$NF + F_m + e_I$
c-svc11	[0.001,0.1]	[1,1000]	[0.005,10]	$NF + F_m$
c-svc12	[0.001,0.1]	[1,1000]	[0.005,10]	$NF + e_I$
c-svc13	[0.001,0.1]	[1,1000]	[0.005,10]	$NF + e_I I$
c-svc14	[0.001,0.1]	[1,1000]	[0.005,10]	$NF + F_m + e_I$

was applied to compare the final population for all runs [29,30]. This method attributes to each objective vector a probability that this point is attaining in one single run [29]. 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) [31]. 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 [32].

4.2 Correspondence between Optimization Objectives

First, in order to have an idea about the shape of the Pareto fronts a population of 1000 individuals generated randomly was initially evaluated using the classifier and the different metrics (equations 1 to 6). Figure 1 shows these results. As can be seen it is possible to obtain identical value for F_m with different number and combinations of features. However, when approaching the top left corner (maximization of Fmeasure and minimization of the number of features) this does not happen, as expected. In the case of the graph e_I versus F_m the best location is the bottom right corner, but, as expected, the tendency is to go to the top right corner. This means that these objectives are conflicting.
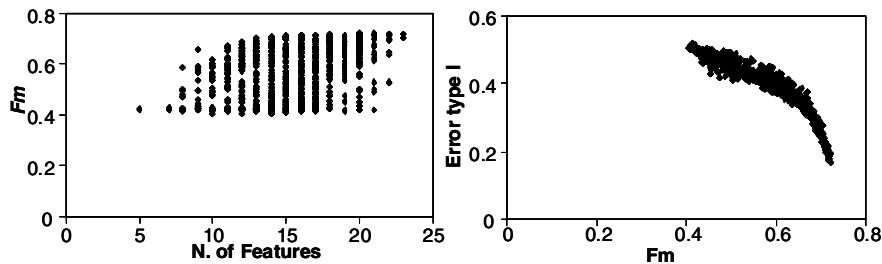
Fig. 1. Pareto plots for 1000 solutions generated randomly

Table 3. Results for Run 1 of experiment C-SVC14

N.F.	F_m	e_I	LR	γ	C	Features
2	0.885	0.0467	0.0953	9.99	985	F11, F28
3	0.962	0.0054	0.0992	9.98	983	F11, F18, F28
3	0.967	0.0173	0.0970	9.99	975	F 8, F11, F28
4	0.997	0.0000	0.0941	9.98	959	F 8, F11, F16, F28
4	0.998	0.0017	0.0960	9.99	967	F 8, F11, F13, F28
5	1.000	0.0000	0.0986	9.99	967	F 1, F 8, F11, F13, F2
5	1.000	0.0000	0.0937	9.98	971	F 3, F 8, F11, F16, F28
5	1.000	0.0000	0.0971	9.94	905	F 3, F 8, F11, F13, F28
5	1.000	0.0000	0.0969	9.97	986	F 3, F 8, F11, F28, F30
5	1.000	0.0000	0.0917	9.98	968	F 8, F11, F13, F16, F28
5	1.000	0.0000	0.0943	9.98	975	F 4, F 8, F11, F16, F28
5	1.000	0.0000	0.0908	9.99	965	F 8, F 9, F11, F16, F28
5	1.000	0.0000	0.0962	9.96	977	F 6, F 8, F11, F13, F28
5	1.000	0.0000	0.0958	9.99	969	F 8, F11, F13, F22, F28
5	1.000	0.0000	0.0984	9.95	905	F 5, F 8, F11, F13, F28
5	1.000	0.0000	0.0979	9.99	980	F 8, F11, F18, F22, F28

Table 4. Results with 3 features for Runs of experiment C-SVC14

Run	F_m	e_I	LR	γ	C	Features
1	0.967	0.0173	0.0970	9.99	975	F8, F11, F28
2	0.963	0.0151	0.0805	9.93	737	F8, F16, F22
3	0.953	0.0168	0.0443	9.95	942	F8, F9, F28
4	0.962	0.0194	0.0739	9.95	990	F6, F8, F23
5	0.956	0.0135	0.0529	9.90	961	F8, F14, F16
6	0.964	0.0226	0.0182	9.81	984	F8, F11, F29

4.3   Optimization Results

EAFs graphs were vused to compare the performance between experiments c-svc1 and c-svc11, c-svc2 and c-svc12 and c-svc3 and c-svc13, were objective 1 is the number of features and objective 2 is F_m, e_I or e_{II}, respectively. These plots were not presented here due to a lack of space. The analysis of these plots allows concluding that the best performance is always obtained when the classifier parameters are optimized simultaneously (i.e., experiments c-svc11 to c-svc13). This indicates that the optimization algorithm is able to find the best classifier parameters for the case under study. The same is true for the experiments with three objectives (c-svc4 and c-svc14), to which is more difficult to obtain the EAFs graphs and, thus, they are not present. Figure 2 plots the Pareto-fronts after 100 generations of a single run of experiments c-svc4 (left) and c-svc14 (right). It is clear from these plots that the run corresponding to experiment c-svc is more efficient. In this case fewer solutions are found since the MOEA was able to reach to solutions were the Fmeasure and eI converge to its best value (1 and 0, respectively).
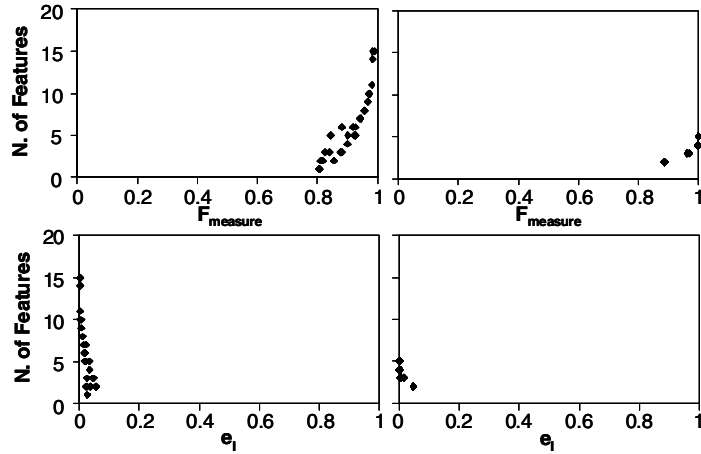
Fig. 2. Pareto optimal fronts for a single run of experiment c-svc14 (Table 2)

The results obtained for a single run of experiment c-svc14 (i.e., the solutions shown in the right graphs of Figure 2) are presented on Table 3. As can be seen there are several solutions were the number of features obtained are the same, mainly in the case of the solutions with five features. This is due to small changes produced in the classifier parameters values (LR, C and γ) also optimized. If considered that a good solution will be the one with an F_m higher than 90%, the two solutions with three features were selected: the solution with F11, F18 and F28 and the solution with F8, F11 and F28 features, respectively. The difference between these two solutions is due to features F11 and F8. When F8 is present both F_m and e_I increase.

Table 4 presents for ilustrating purposes the best results accomplished only with three features selected for some of the runs of experiment c-svc14. In this case the MOEA converged to a different set of features. Also, in this case, the features selected are able to cluster the companies. Therefore, for this set of data there is more than one solution able to attain the objectives defined.

5   Conclusion

In this work a MOEA was used for feature selection in the bankruptcy prediction problem using Support Vector Machines classifier. The methodology proposed was able not only to reduce the features necessary but is able also to provide relevant information to the decision maker. The algorithm does not only provide the best features to be used but, also, with the best parameters of the classifier. The best performance only is attained when the classifier parameters are optimized simultaneously with the features to be selected, since the classifier performance is strongly dependent on these parameters. Finally, the MOEA was able to provide more than one set of features able to optimize the objectives defined.

References

1. Atiya, F.: Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Transactions on Neural Networks 12, 12–16 (2001)
2. Eisenbeis, R.A.: Pitfalls in the Application of Discriminant Analysis in Business, Finance and Economics. J. of Finance 32, 875–900 (1997)
3. Martin, D.: Early Warning of Bank Failure: A Logit Regression Approach. J. of Banking and Finance 1, 249–276 (1977)
4. Charitou, A., Neophytou, E., Charalambous, C.: Predicting corporate failure: empirical evidence for the UK. European Accounting Review 13, 465–497 (2004)
5. Neves, J.C., Vieira, A.S.: Improving Bankruptcy Prediction with Hidden Layer Learning Vector Quantization. European Accounting Review 15, 253–271 (2006)
6. Fan, A., Palaniswami, M.: Selecting bankruptcy predictors using a support vector machineapproach. In: Proceedings of IJCNN 2000, pp. 354–359 (2000)
7. Coats, P.K., Fant, L.F.: Recognizing Financial Distress Patterns Using a Neural Network Tool. Financial Management 22, 142–155 (1993)
8. Yang, D.T.: Urban-biased policies and rising income inequality in China. American Economic Review Papers and Proceedings 89, 306–310 (1999)
9. Tan, C.N.W., Dihardjo, H.: 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, 56–77 (2001)
10. Vieira, A.S., Duarte, J., Ribeiro, B., Neves, J.C.: Accurate Prediction of Financial Distress of Companies with Machine Learning Algorithms. In: Kolehmainen, V., Toivanen, P., Beliczynski, B. (eds.) ICANNGA 2009. LNCS, vol. 5495, pp. 569–576. Springer, Heidelberg (2009)
11. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.: Feature Extraction Foundations and Applications. Springer, Heidelberg (2006)
12. Bi, J.: Multi-Objective Programming in SVMs. In: Proceedings of the Twentieth International Conference on Machine Learning, ICML 2003, Washington, DC (2003)
13. Igel, C.: Multi-Objective Model Selection for Support Vector Machines. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 534–546. Springer, Heidelberg (2005)
14. Oliveira, L.S., Morita, M., Sabourin, R.: Feature Selection for Ensembles Using the Multi-Objective Optimization Approach. SCI, pp. 49–74 (2006)
15. Hamdani, T.M., Won, J.-M., Alimi, A.M., Karray, F.: Multi-objective Feature Selection with NSGA II. In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) ICANNGA 2007. LNCS, vol. 4431, pp. 240–247. Springer, Heidelberg (2007)
16. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Transaction on Evolutionary Computation 6, 181–197 (2002)
17. Alfaro-Cid, E., Castillo, P.A., Esparcia, A., Sharman, K., Merelo, J.J., Prieto, A., Mora, A.M., Laredo, J.L.J.: Comparing Multiobjective Evolutionary Ensembles for Minimizing Type I and II Errors for Bankruptcy Prediction. In: CEC 2008, Washington, USA, pp. 2907–2913 (2008)

18. Handl, J., Knowles, J.: Feature subset selection in unsupervised learning via multiobjective optimization. Int. J. of Computational Intelligence Research 2, 217–238 (2006)
19. Gaspar-Cunha, A., Mendes, F., Duarte, J., Vieira, A., Ribeiro, B., Ribeiro, A., Neves, J.: Feature Selection for Bankruptcy Prediction: A Multi-Objective Optimization Approach. Int. J. of Natural Computing Research 1, 71–79 (2010)
20. Kulkarni, A., Jayaraman, V.K., Kulkarni, B.D.: Support vector classification with parameter tuning assisted by agent-based technique. Computers and Chemical Engineering 28, 311–318 (2008)
21. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20, 273–297 (1995)
22. Chang, C.-C., Lin, C.-J.: LIBSVM a library for support vector machines (Tech. Rep.). Dept. of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan (2000)
23. Caruana, R., Niculescu-Mizil, A.: Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria. In: KDD 2004, Seattle, Washington, pp. 69–78 (2004)
24. Provost, F., Fawcet, T.: Analysis and Verification of Classifier Performance: Classification under Imprecise Class and Cost Distributions. In: KDD 1997, Menlo Park, CA, pp. 43–48 (1997)
25. Fawcet, T.: An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874 (2006)
26. Deb, K.: Multi-Objective Optimization using Evolutionary Algorithms. Wiley, New York (2001)
27. Gaspar-Cunha, A., Covas, J.A.: RPSGAe - A Multiobjective Genetic Algorithm with Elitism: Application to Polymer Extrusion. In: Dorigo, M., Birattari, M., Blum, C., Gambardella, L.M., Mondada, F., Stützle, T. (eds.) ANTS 2004. LNCS, vol. 3172, pp. 221–249. Springer, Heidelberg (2004)
28. Gaspar-Cunha, A.: Modelling and Optimization of Single Screw Extrusion. Published doctoral dissertation, 2000. Lambert Academic Publishing, Köln (2009)
29. Fonseca, C., Fleming, P.J.: On the performance assessment and comparison of stochastic multiobjective optimizers. In: Ebeling, W., Rechenberg, I., Voigt, H.-M., Schwefel, H.-P. (eds.) PPSN 1996. LNCS, vol. 1141, pp. 584–593. Springer, Heidelberg (1996)
30. Knowles, J.D., Thiele, L., Zitzler, E.: A tutorial on the performance assessment of stochastive multiobjective optimizers. TIK-Report No. 214 (2006)
31. Fonseca, V.G., Fonseca, C., Hall, A.: Inferential performance assessment of stochastic optimisers and the attainment function. In: Zitzler, E., Deb, K., Thiele, L., Coello Coello, C.A., Corne, D.W. (eds.) EMO 2001. LNCS, vol. 1993, pp. 213–225. Springer, Heidelberg (2001)
32. López-Ibañez, M., Paquete, L., Stützle, T.: Hybrid population based algorithms for the bi-objective quadratic assignment problem. J. of Math. Modelling and Algorithms 5, 111–137 (2006)