Knowledge Discovery in the Prediction of Bankruptcy

R. J. Almeida[1] S. Vieira[2] V. Milea[1] U. Kaymak[1] J. M. C. Sousa[2]

1.Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
2.Technical University of Lisbon, Instituto Superior Técnico, Dept. Mechanical Engineering/CIS-IDMEC
Av. Rovisco Pais 1, 1049-001 Lisbon
Email: {rjalmeida,kaymak,milea}@ese.eur.nl, {susana,jsousa}@dem.ist.utl.pt

Abstract— *Knowledge discovery in databases (KDD) is the process of discovering interesting knowledge from large amounts of data. However, real-world datasets have problems such as incompleteness, redundancy, inconsistency, noise, etc. All these problems affect the performance of data mining algorithms. Thus, preprocessing techniques are essential in allowing knowledge to be extracted from data. This work presents a real world application of knowledge discovery in databases, with the objective of prediction of bankruptcy. For this task fuzzy classification models based on fuzzy clustering are used, which are developed solely from numerical data. This data set has missing values, extreme values and also presents a much smaller bankruptcy class than the not bankruptcy class, which makes it a challenging problem in the scope of KDD.*

Keywords— Knowledge discovery in databases, feature selection, missing data, noisy data, prediction of bankruptcy, fuzzy classification.

1 Introduction

With the increase of economic globalization and evolution of information technology, financial data are being generated and accumulated at an exponential rate. It is used to keep track of companies, business performance, monitor market changes, and support financial decision-making. This rapidly growing volume of data triggered the need for automated approaches that allow effective and efficient utilization of massive financial data to support companies and individuals in strategic planning and investment decision-making.

Knowledge discovery can contribute to solving business problems in finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume of data is too large or is generated too quickly to be screened by experts. Knowledge discovery has already been applied to a number of financial applications, including development of trading models, investment selection, loan assessment, portfolio optimization, fraud detection and bankruptcy prediction, amongst others. The prediction of bankruptcy has been been previously investigated in terms of the likelihood of success for the introduction of fuzzy systems for decision support [1]

The prediction of corporate failure or bankruptcy has been characterized as one of the most important problems facing business and government [2]. It also is a problem that affects the economy of every country. The number of failing firms is important for the economy of a country and it can be considered as an index of the development and robustness of the economy. The high individual, economic, and social costs encountered in corporate failures or bankruptcies make this problem very important to parties such as auditors, management, government policy makers, and investors [3].

There is a long history of research attempting to develop bankruptcy prediction models based on financial variables and other indicators of financial distress, using a wide variety of techniques. The pioneer in predicting business failure ratios is considered to be [4]. The predictive accuracy of the initial approaches has varied from around 65% [5] to around 90% [6]. Higher predictive accuracy is often achieved by using samples concentrated in a few industries, using samples with widely varying bankruptcy/non-bankruptcy company sizes or making inappropriate assumptions about real world bankruptcy/non-bankruptcy frequencies.

The data set used in this work, concerning the bankruptcy, is different from all of the above mentioned, so no direct comparison of results can be made. Also, other works never mentioned that data sets had missing values and extreme values, as in this work.

In this work we present a full KDD process, applied in a real-world problem: the prediction of bankruptcy. The data set used is quite challenging as it has missing values and extreme values. It also presents a much smaller bankruptcy class than the not bankruptcy class. Several possibilities were tested in each step of the KDD process, such as data preparation, feature selection and fuzzy classification, and we discuss them briefly although we give more focus on the best results obtained. Note that the use of fuzzy systems for classification, besides building a numeric prediction model also represents the model behaviour in terms of linguistic rules, making it possible to interpret, which is an important final step in the KDD process.

The outline of the paper is as follows. In Section 2 we briefly present the KDD steps used to obtain a bankruptcy classifier. In Section 3 we present techniques used for data preparation. Feature selection is presented in Section 4 and Section 5 describes the procedures used to derive Takagi-Sugeno fuzzy models by means of fuzzy clustering. The data used in the work and the results are presented in Section 6. Finally, conclusions and future work are given in Section 7.

2 KDD process

The search for knowledge in large data sets, with the use of different hypothesis spaces, is the central and necessary phase within the discovery process. A large number of methods have been developed that handle many search tasks, but hypotheses

ISBN: 978-989-95079-6-8

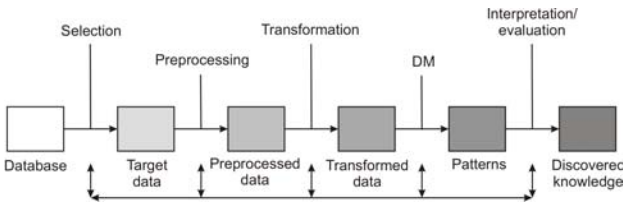

Figure 1: Phases in the KDD process, adapted from [8]

inference and verification is only a part of the whole process of knowledge discovery. As any other process, it has its environment, its phases, and runs under certain assumptions and constraints. The process undertakes many phases, namely [7]:

1. Definition and analysis of the problem;
2. Understanding and preparation of data;
3. Setup of the search for knowledge;
4. The actual search for knowledge;
5. Interpreting mined patterns;
6. Deployment and practical evaluation of the solutions.

The KDD process, activities and phases, is shown in Figure 1.

Compared to the traditional manual analysis, KDD provides a much higher degree of system autonomy, especially in processing large hypotheses spaces. However, at the current state of the art, a human analyst still makes many decisions in the course of a discovery process.

The KDD process starts from specification of a given problem and data understanding, and ends with actionable conclusions from the discovered knowledge. The output of DM is, in general, a set of patterns, some of which possibly represent discovered knowledge. In the next sections we will briefly explain each step of the KDD process.

3 Data Preparation

Real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size. Thus, data preparation consumes most of the time needed to mine data [9], and can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining [10].

This section is a brief overview of some of the concepts regarding data preparation to yield the best possible model. For more details refer to [9, 10].

3.1 Missing Data

When applying data analysis methods to real problems, we often find that the data sets contain many missing elements. There are two forms of randomly missing data: missing completely at random (MCAR) and missing at random (MAR) [11]. Missing values MCAR, behave like a random sample and their probability does not depend on the observed data or the unobserved data [12, 13]. MAR exists when missing values are not randomly distributed across all observations but are randomly distributed within one or more subsamples.

A possible approach to deal with the missing values is to discard all incomplete data, and then execute the data analysis method on the remaining data. However, if missing values are frequent, the data set size may be considerably reduced, yielding unreliable or distorted results. A way to minimize this extreme data reduction problem is presented in [14]. If values are missing completely at random they can be imputed. Widely used imputation methods use the variables mean, median or the most probable value as a replacement [11, 10].

3.2 Noisy Data

Noise is considered to be a random error or variance in a measured variable [10]. A specific type of noise, which the data in this study has, is extreme values, which in some cases can be labelled as outliers. In this work numerically sensitive mining algorithms are used, and it is recommended to normalize the individual variable distributions [15].

Outliers are defined as a large deviation from the mean value of the rest of the data. Usually outliers can be thrown out of the data set, as they bias the analytical results. Distribution normalization deals with the problems presented by valid outliers. An outlier is valid if it represents an accurate measurement and still falls well outside the range of the majority of values. These should not be discarded.

The distribution normalization can be achieved using transformations. In cases where the data have strong asymmetry, many outliers or batches at different levels with widely differing spreads a power transformation may alleviate this problem, without violating the necessary transformation properties [16].

4 Feature Selection

One of the great challenges in classification is selecting the important input variables from all possible input variables. Classification problems involve a large number of potential inputs. The number of inputs actually used by the model must be reduced to the necessary minimum, especially when dealing with fuzzy models that are, presumably, nonlinear and contain many parameters. Therefore, it is necessary to select carefully the variables that are relevant for the feature class.

Feature selection is a process that chooses a subset of M features from the original set of n features ($M \leq n$), so that the feature space is optimally reduced according to a certain criteria.

Even when a good criteria exists for model selection, there is no guarantee that a model based on a given set of variables is optimal unless all possible combinations of variables have been explored. The problem is known to be NP-hard [17]. Hence, finding an optimal solution requires building a model for each possible combination of input variables, which becomes computationally prohibitive for problems involving even a moderate number of candidate input variables.

Feature selection algorithms, essentially divide into wrappers and filters [18]. In this work we use wrappers that make use of specific learning algorithms to evaluate variables in the context of the learning problem, rather than independently. Wrappers share strengths and weaknesses of the learning algorithms and have the advantage of using the actual hypothesis accuracy as a measure of subset quality. Furthermore, wrapper methods do tend to outperform filter methods [18].

ISBN: 978-989-95079-6-8

The variable selection procedure can be used with various performance criteria for model selection. In real-world databases sometimes one of the classes is more difficult to classify then the others. This can happen, for instance, when one of the classes is much bigger than the other or the interest of the problem is a specific class. To cope with this problem, we use a criterion, that assigns specific weights to each class in the model evaluation of the feature selection algorithm [19].

In this work we compare the well known sequential forward selection (SFS) and the sequential backward elimination (SBE) against the newly proposed ant feature selection algorithm (AFS). The SFS and SBE search algorithms may not be the best search methods, nor guarantee an optimal solution, but they are popular because they are simple, fast, provide a very reasonable solution and are much more efficient than exhaustive search.

Sequential forward selection and sequential backward elimination were first used in the context of feature selection for pattern classification. SFS was first used in [20] and was later used in [21]. It has also been used to determine input variables for fuzzy models in [22] and [23]. SBE was first used in [20] and in [24] it was used with simple linear models to provide a first-round elimination of the input variables for a fuzzy model.

4.1 Ant Feature Selection

Ant algorithms were first proposed by Dorigo [25] as a multi-agent approach to difficult combinatorial optimization problems, such as traveling salesman problem, quadratic assignment problem or supply chain management [26]. Here we present an implementation of ACO applied to feature selection, where the best number of features is determined automatically.

In this approach, two objectives are considered: minimizing the number of features and minimizing the error classification. Two cooperative ant colonies optimize each objective. The first colony determines the number (cardinality) of features and the second selects the features based on the cardinality given by the first colony. Thus, two pheromone matrices and two different heuristics are used. The heuristic value is computed using the Fisher discriminant criterion for feature selection [27], which ranks the features giving them a given relative importance.

The best number of features is called *features cardinality* N_f. The determination of the *features cardinality* is addressed in the first colony sharing the same minimization cost function J^τ with the second colony, which in this case aggregates both the maximization of the classification accuracy and the minimization of the features cardinality. Hence, the first colony determines the size of the subsets of the ants in the second colony, and the second colony selects the features that will be part of the subsets.

The objective function of this optimization algorithm aggregate both criteria, the minimization of the classification error rate and the minimization of the features cardinality:

$$J^\tau = w_1 \frac{N_e^\tau}{N_n} + w_2 \frac{N_f^\tau}{n} \tag{1}$$

where $\tau = 1, \ldots, g$, g being the number of ants, w_1 and w_2 are weights, N_n is the number of used data samples, n is the total number of features, N_e is the number of errors produced by the solution and N_f is the features cardinality.

To evaluate the classification error, a fuzzy classifier is built for each solution following the procedure described in Section 5. This approach was presented in [28].

5 Fuzzy Classification

In this section we outline the basics of the adopted fuzzy reasoning scheme for pattern classification problems. Let us consider a n-dimensional classification problem for which N patterns $\vec{x}_p = (x_p^1, \ldots, x_p^n)$, $p = 1, 2, \ldots, N$ are given from κ classes $C_1, C_2, \ldots, C_\kappa$. The task of a pattern classifier is to assign a given pattern $\vec{x}$ to one of the κ possible classes based on its features values. Thus, a classification task can be represented as a mapping $\psi : X \subset \Re^n \rightarrow \{0,1\}^\kappa$ where $\psi(\vec{x}) = \vec{c} = (c_1, ..., c_\kappa)$ such that $c_k = 1$ and $c_j = 0 (j = 1, \ldots, \kappa, j \neq k)$.

Assuming that there is an arbitrary ordering of the classes, one way to solve this classification problem is to consider classifiers with a continuous output, e.g., a Takagi-Sugeno affine system [29]. The output of an affine Takagi-Sugeno fuzzy rule is

$$y_k = \vec{a}\vec{x} + b_k \tag{2}$$

where y_k is the output for rule k, $\vec{a}_k$ is a parameter vector and b_k is a scalar offset. In classification problems the output should be a discrete value corresponding to one of the classes to be identified. So a threshold T_l can be used on the output y_k, to decide which class that it belongs to, as $x_k \in C_l$ if $y_k \in T_l$.

To form the fuzzy system model from the data set with N data samples, given by $X = [\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N]^T$, $Y = [y1, y2, \ldots, yN]^T$ where each data sample has a dimension of n ($N >> n$), the structure is first determined and afterwards the parameters of the structure are identified. The number of rules characterizes the structure of a fuzzy system. Fuzzy clustering in the Cartesian product-space $X \times Y$ is applied to partition the training data. The partitions correspond to the characteristic regions where the systems behavior is approximated by local linear models in the multidimensional space. Given the training data X_T and the number of clusters K, a suitable clustering algorithm is applied.

The fuzzy clustering algorithms used in this work are based on the optimization of an objective function. In particular we use the fuzzy c-means (FCM) [15], the Gustafson-Kessel (GK) [30], the possibilistic c-means (PCM) [31], fuzzy possibilistic c-means (FPCM) [32] and the recent possibilistic fuzzy c-means (PFCM) [33].

The FCM functional uses a probabilistic constraint which states that the sum of membership degrees must equal one [15]. Problems arise in situations, where the total membership of a data point to all the clusters does not equal one, as in the presence of outliers. Clearly, one would like the memberships for representative feature points to be as high as possible, while unrepresentative points should have low membership in all clusters. The PCM objective function relaxes this constraint [31].

Gustafson-Kessel extended the standard fuzzy c-means algorithm by employing an adaptive distance norm, in order to detect clusters of different geometrical shapes in one data set [30]. As the GK algorithm is based on an adaptive distance

ISBN: 978-989-95079-6-8

measure, it is less sensitive to scaling (normalization, standardization) of the data.

Fuzzy-possibilistic c-means, simultaneously produces both memberships and possibilities. FPCM tries to solve the noise sensitivity defect of FCM, and also overcomes the coincident clusters problem of PCM. Note that FCM and FPCM will not generate the same membership values, even if both algorithms are started with the same initialization [32].

FPCM imposes a constraint on the typicality values. PFCM relaxes this constraint but retains the column constraint on the membership values. The PFCM functional has two constants that define the relative importance of fuzzy membership and typicality values.

Memberships and typicalities are both considered important for correct interpretation of data substructure. When the objective is to classify a data point, membership may be a better choice as it is natural to assign a point to that cluster whose representative vector is closest to the data point. On the other hand, when seeking the clusters, i.e., while estimating the centroids, typicality is an important means for alleviating the undesirable effects of outliers.

Depending on the clustering algorithm used in this work, a fuzzy partition matrix $U = [\mu_{ik}]$ and/or the typicality matrix $T = [t_{ik}]$ will be obtained. The fuzzy sets in the antecedent of the rules are identified by means of the matrix U and T which have dimensions $[N \times K]$. One dimensional fuzzy sets A_{ij} are obtained from the multidimensional fuzzy sets by projections onto the space of the input variables x_j. This is expressed by the point-wise projection operator of the form $\mu_{A_{ij}}(x_{jk}) = \text{proj}_j(\mu_{ik})$ The antecedent membership functions can now be obtained from the fuzzy partition matrix or from the typicality matrix. The point-wise defined fuzzy sets A_{ij} are then approximated by appropriate parametric functions. The consequent parameters for each rule are obtained by means of linear least square estimation, which concludes the identification of the classification system.

6 KDD Applied to Prediction of Bankruptcy

The data set used in this work contains data from 1817 companies, each one described by 52 features (including the class feature), containing financial, behavioral and qualitative features (as perceived by the account manager).

Some of the features contain extreme values (EV) as their value is 10000 larger then the other values. In this case, the extreme values are believed to be different values for different type and size of companies, containing valuable information.

The data set contains about 10% of missing values. Only 18.38% (334) of the companies have all the features complete and there are only 5 features without missing values, one of those features being the status of the company.

Each company in the data set has two possible status (classes): status 0 (bankrupt) and status 1 (not bankrupt). The distribution of the classes is uneven: only 4.3% (78) of the companies have the status 0, and the remaining 95.7% (1739) have a status 1. This distribution skewness is common in bankruptcy data.

The KDD steps taken in this work to obtain a compact fuzzy classification model are the following:

1. Manual selection of relevant data from the available data.
2. Preprocessing of the data to deal with extreme values, using a power transformation followed by a linear transformation.
3. Preprocessing of the data to deal with missing values, by replacement of missing values using the most probable value.
4. Searching and selecting the relevant features using search algorithms.
5. Obtaining a fuzzy classification model by using only the selected features in the previous step.

The following computational protocols were used: $\varepsilon = 0.0001$, maximum number of iterations 100, fuzzy exponent $m = 2$ and $\eta = 2$, and the Euclidean norm is used. For both PCM and PFCM we first run FCM to termination. All trials terminated with the convergence criteria after a few iterations.

6.1 Data Preprocessing

When analysing the data we found that seven features can be discarded because they contain company descriptive features that may not be relevant to this research.

Extreme Values Linear transformations of re-expressed data present little additional difficulty in interpretation. Since power transformations are monotonic for positive data values, we transform the data so that it has only positive values, maintaining the missing values, using the linear transformation,

$$z_{jk} = z^*_{jk} + |\min z_j| \tag{3}$$

where the asterisk denotes the unscaled data. After this linear transformation, which does not alter the shape of the data, we apply a power transformation $T(x) = \log(x)$, because this transformation alters the distribution of the data and bring the extreme values to a value closer to the other values [16]. We chose a matched transformation, presented as:

$$z^1 = a \; + \; bT(x), \tag{4}$$

where a and b are chosen, by using a point x_0 and require that:

$$z_0 = a \; + \; bT(x) = x_0, \tag{5}$$

and, furthermore, that the derivative of z with respect to x, evaluated at x_0, to be 1. That is,

$$\left.\frac{dz}{dz}\right|_{x_0} = \frac{d[a + bT(x)]}{dx} = b \left.\frac{dT(x)}{dx}\right|_{x_0} = 1. \tag{6}$$

This method relies on the linearity of the transformation near the center of the range x. For sake of simplicity, we chose x_0 as the median value of each feature. Other point could have been chosen. Furthermore, we normalized the obtained values of the matched transformation so that all the features are contained between the interval $[0, 1]$, obtaining the data set z^2.

Missing Values Almost every company has missing values for their features, and so discarding the companies with missing values cannot be considered a feasible approach for this research. Filling in the missing value during the data preprocess, was used. After the data transformation, we imputed the missing values using probable value, calculated using the Expectation-Maximization (EM) algorithm [34].

ISBN: 978-989-95079-6-8

6.2 Fuzzy Classification

One of the most important advantages of the, rather complex, transformation described, can be seen in Table 1. This table shows the obtained accuracy obtained with the data before the transformation, using 6 clusters, and the obtained results after the transformation with only 3 clusters. Since the class 0 is so small, we used 4-fold cross validation. As can be seen the results obtained with the data before transformation are more disperse. This does not happen with the transformed data. The GK algorithm could not be used in the original data because of the extreme values. For this reason we decided not to included it in this comparison. Furthermore the converge of the clustering algorithms is improved and less number of clusters are needed to obtain good results.

Table 1: Accuracy obtained for bankruptcy data with 6 clusters (Raw data) and 3 clusters (Processed data).

	Raw data			Processed data		
Alg.Cl.	Max.	Mean	Min.	Max.	Mean	Min.
FCM	0.928	0.885	0.800	0.972	0.946	0.919
PCM	0.925	0.894	0.733	0.969	0.945	0.919
FPCM-U	0.924	0.877	0.800	0.953	0.944	0.939
PFCM-U	0.926	0.894	0.733	0.955	0.941	0.927
FPCM-T	0.925	0.891	0.800	0.939	0.939	0.939
PFCM-T	0.925	0.891	0.800	0.939	0.925	0.919

6.3 Feature Selection

After the fuzzy classification models with all the features were extracted, we look to further simplify the model, using feature selection. We applied each algorithm of feature selection (SFS,SBE and AFS), using the weighted accuracy for each feature class present [19], and all of the clustering algorithms in study (GK,FCM, FFCA, PCM, FPCM and PFCM), to both the data normalized and not normalized.

In this case, if no weights were used, the algorithms would choose the features that would maximize the class no bankruptcy as this is the dominant class. Usually after one or two iterations the process would stop. In this case, without the use of the weights the results are highly biased towards the bigger class.

The features which were selected throughout our tests varied between 2 with the sequential forward search and 37 with the sequential backward search. This result was expected as different types of features selection algorithms choose features in a different manner, and we tested a number of different clustering algorithms. Good results were obtained using the ants feature selection. The number of features chosen was 15, 3 rules were derived and the obtained accuracy was 78.9% for companies that are bankrupt and 94.9% for companies that are not bankrupt. The fuzzy clustering algorithm used was PCM.

The use of fuzzy systems for classification, besides building a numeric prediction model also represents the model behaviour in terms of linguistic rules, which is very natural for human to understand. Interpretability is considered to be the main advantage of fuzzy systems over other non-fuzzy alternatives like statistical models or neural networks.

The simplest classification model obtained with the data normalized, only has 2 features and 3 rules and was derived using the PCM algorithm. The obtained accuracy is 57.9% for the companies that are bankrupt, whereas the percentage of companies that are not bankrupt is 97.7%. If we compare the possibility of interpretability of this model against the model derived in [3], that had 70 rules and 35 features then it can be considered that these results are good. Also bare in mind that this data set has missing values, extreme values and one class which is much smaller than the other.

The rule base model is relatively simple, as only 3 rules and 2 features are used, the understanding of the consequents of the model is not a simple task. The obtained rules, for this simple model are:

1. **If** $PROFIAT$ is Low **and** $CFEQ$ is Low **then** $y_1 = 0.35PROFIAT + 0.16CFEQ - 0.16$
2. **If** $PROFIAT$ is $Medium$ **and** $CFEQ$ is $Medium$ **then** $y_2 = -24.52PROFIAT - 18.65CFEQ + 39.12$
3. **If** $PROFIAT$ is $High$ **and** $CFEQ$ is $High$ **then** $y_3 = 5.54PROFIAT + 8.49CFEQ - 11.85$

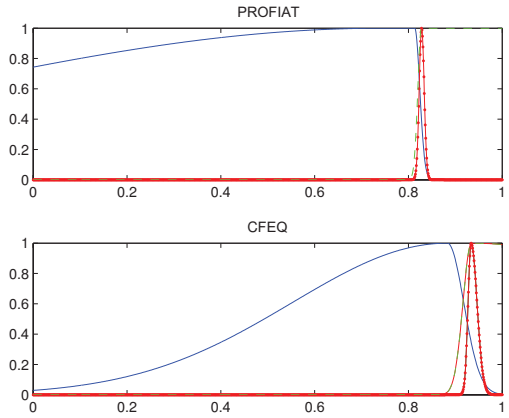

Figure 2: Membership Functions bankruptcy Model. solid - Low, dotted - Medium, dashed - High.

The obtained membership functions are shown in Fig. 2. The features we have used are the profit after tax (PROFIAT), and cash flow to equity (CFEQ). Note that the former is the company's net operating after tax profit for investors, while the latter is the cash that can be paid to the equity shareholders after the company expenses. It is interesting to note that the membership functions are skewed towards the higher values, which indicates that the difference between company defaulting or not is quite small. According to our simple model, a small value on either of these features is an indication that the company is performing badly and it is likely to default. It is interesting to note that the *medium* membership function are located in high values. Therefore, even companies with high values of both features can still default.

7 Conclusions

In this work we present a real world application of knowledge discovery for prediction of bankruptcy using a databases that has noisy, missing, and inconsistent data. We present for each step of the KDD process several possibilities that we tested in order to obtain good fuzzy classification models. Before fuzzy models could be extracted from the data, it is necessary to represent the real-world objects of interest in the data

ISBN: 978-989-95079-6-8

in a way that this specific method can access the data. With data preparation, although time consuming, it was possible to derive compact fuzzy models with only a few features that predicted the bankruptcy with an high accuracy rate.

Acknowledgments

This work is partially supported by the Portuguese Government and FEDER under the programs: Programa de financiamento Plurianual das Unidades de I&D da FCT (POCTI-SFA-10-46-IDMEC) and by the FCT grant SFRH/25381/2005, Fundação para a Ciência e a Tecnologia, Ministério do Ensino Superior, da Ciência e da Tecnologia, Portugal.

References

[1] M. Setnes, U. Kaymak, and H.R. van Nauta Lemke. Fuzzy target selection in direct marketing. *IEEE Transactions on Fuzzy Systems*, 7(1):368–369, 1999.

[2] D.E. O'Leary. Using neural networks to predict corporate failure. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 7(3):187–197, 1998.

[3] J. Stefanowski and S. Wilk. Evaluating business credit risk by means of approach-integrating decision rules and case-based learning. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 10(2):97–114, June 2001.

[4] E. I. Altman. Financial ratios, discriminate analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4):589–609, 1968.

[5] D.H. Lindsay and A. Campbell. A chaos approach to bankruptcy prediction. *Journal of Applied Business Research*, 12(4):1–9, 1996.

[6] T.B. Bell, G.S. Ribar, and J. Verchio. Neural nets versus logistic regression: a comparison of each models ability to predict commercial bank failures. In *Proceedings of the 1990 Deloitte and Touche University of Kansas Symposium on Auditing Problems*, pages 29–54, 1990.

[7] G. Piatetsky-Shapiro U. Fayyad and P. Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996.

[8] P. Smyth U.M. Fayyad, G. Piatetsky-Shapiro and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA, USA, 1996.

[9] D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, San Francisco, CA, 1999.

[10] J. Han and M. Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers Inc, San Francisco, CA, 2000.

[11] H. Timm, C. Döring, and R. Kruse. Differentiated treatment of missing values in fuzzy clustering. In Taner Bilgiç, Bernard De Baets, and Okyay Kaynak, editors, *Proceedings of the 10th International Fuzzy Systems Association World Congress (IFSA 2003), Lecture Notes in Artifical Intelligence*, volume 2715.

[12] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data, Second Edition*. Wiley and Sons, New York, 2002.

[13] J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.

[14] M. Setnes and U. Kaymak. Fuzzy modeling of a client preference from large data sets: an application to target selection in direct marketing. *IEEE Transactions on Fuzzy Systems*, 9(1):153–163, 2001.

[15] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[16] M.A. Stoto J.D. Emerson. Transforming data. In F. Mosteller D.C. Hoaglin and J.W. Tukey, editors, *Understanding Robust and Exploratory Data Analysis*, chapter 4. Wiley-Interscience, New York, London, Sydney, Toronto, 2000.

[17] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1–2):237–260, 1998.

[18] R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[19] C. A. Silva R.J. Almeida and J.M.C. Sousa. A new feature selection criterion for fuzzy classification. In *Proc. of the 2006 IEEE International Conference on Fuzzy Systems, WCCI 06*, pages Session Classification, pp. 1513–1520, Vancouver, Canada, July 2006.

[20] T. Marill and D.M. Green. On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, 9(1):11–17, 1963.

[21] A.W. Whitney. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, 20(9):1100–1103, 1971.

[22] M. Sugeno and G.T. Kang. Structure identification of fuzzy model. *Fuzzy Sets and Systems*, 28(1):15–33, 1988.

[23] M. Sugeno and T. Yasukawa. A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on Fuzzy Systems*, 1(1):7–31, February 1993.

[24] H. Watanabe K. Tanaka, M. Sano. Modeling and control of carbon monoxide concentration using a neurofuzzy technique. *IEEE Transactions on Fuzzy Systems*, 3(3):271–279, 1992.

[25] Marco Dorigo, M. Birattari, and T. Stützle. Ant colony optimization. *IEEE Computational Intelligence Magazine*, 1(4):28–39, 2006.

[26] C. A. Silva, J. M. C. Sousa, T. A. Runkler, and J. M. G. Sá da Costa. Distributed optimization of a logistic system and its suppliers using ant colony optimization. *International Journal of Systems Science*, 37(8):503–512, 2006.

[27] R. O. Duda, P. E. Hart, and D. G Stork. *Pattern Classification, Second Edition*. Wiley–Interscience Publication, 2001.

[28] S. M. Vieira, J. M. C. Sousa, and T. A. Runkler. Fuzzy classification in ant feature selection. In *Proc. of 2008 IEEE World Congress on Computational Intelligence, WCCI 2008*, pages 1763–1769, Hong Kong, China, June 2008.

[29] T. Takagi and M. Sugeno. Fuzzy identification of systems and its aplication to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics*, 15(1):116–132, 1985.

[30] D. Gustfson and W. Kessel. Fuzzy clustering with a fuzzy covariance matrix. In *Proceedings of IEEE Conference on Decision and Control, CDC*, pages 761–766, San Diego, CA, USA, 1979.

[31] R. Krishnapuram and J.M. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110, May 1993.

[32] N.R Pal, K. Pal, and J.C Bezdek. A mixed c-means clustering model. In *Proceedings of the Sixth IEEE International Conference on Fuzzy Systems*, volume 1, pages 11–21, Barcelona, Catalonia, July 1997.

[33] J.M. Keller N.R. Pal, K. Pal and J.C.Bezdek. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions of Fuzzy Systems*, 13(4):517–530, August 2005.

[34] T. Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14:853–871, October 2001.