GRAMMAR-GUIDED GENETIC PROGRAMMING FOR FUZZY RULE-BASED CLASSIFICATION IN CREDIT MANAGEMENT

A. TSAKONAS
Aristotle University of Thessaloniki, Dept. of Informatics, Artificial Intelligence and Information Analysis Lab
BOX 451, 54124, Thessaloniki, Greece,
tel: +2310-996361, fax: +2310-998453 , e-mail: tsakonas@csd.auth.gr

G.DOUNIAS
University of the Aegean, Dept. of Financial and Management Engineering,
31 Fostini st., 82100 Chios, Greece,
tel. +23271-35165, fax: +23271-93464, e-mail: g.dounias@aegean.gr

Summary

The study presents a computational intelligent methodology for fuzzy rule-based classification of enterprises into different categories of credit risk. The presented methodology correspond to an approach to the problem of classifying credit applicants, according to the need for reduction of complexity, higher classification accuracy, and comprehensibility of the acquired decision rules. The data used are both of numerical and linguistic nature and they represent a real world problem, that of deciding whether a loan should be granted or not, in respect to financial details of customers applying for that loan, to a private bank of a southern province of the European Union. The techniques involved in the rule-based categorization task are the inductive machine learning and the type-constrained genetic programming. We examine a two-step model, with a sample of 124 enterprises that applied for a loan, each of which is described by 76 (mainly financial) decision variables, and classified to one of the seven predetermined classes. Special attention is given to the comprehensibility and the ease of use for the acquired decision rules. The application of the proposed methods can make the classification task easier and may minimize significantly the amount of required credit data. We consider that the methodology may also give the chance for the extraction of a comprehensible credit management model or even the incorporation of a related decision support system in banking. The overall architecture of the model can be continuously retrained and reformed, by adding every new credit-risk case, becoming more and more accurate and robust classification models over time.

1. INTRODUCTION AND LITERATURE REVIEW

Recently, there is a rise in the use of different kinds of financial transactions, either in the banking level, or in the private business sector, a fact that makes the automation of decision-making processes imperative. However, this automation has to keep the error rates at a low level during the data processing phase, while it should also be as less time consuming as possible. Furthermore, the decisions taken with the use of an automated process should approximate if possible, the corresponding decision of an expert for the same problem. This is the reason why, methods that imitate human thought and model human expertise and abilities, have become popular in the last two decades. The paper deals with the financial application domain of loans, which is one of the most commonly used financial transactions in banking organizations. A large number of criteria exist, according to each case, that in practice determine the restrictions and the final outcome of the decision process for granting a loan. For the fundamentals of credit scoring the reader should advise (Lewis 1992). The subject of credit scoring is not so popular in literature as one would expect (Thomas 2000). Most of the papers attempt statistical approaches for classifying candidates to be loaned in two classes: positive and negative, see (Boyes *et al.* 1989), (Thomas 1998), (Thomas 2000). They usually perform discriminant analysis in order to produce a discriminant rule for these two classes (Steenackers and Goovaerts 1994). Other approaches use count data models in order to predict the number of times that an applicant for credit will not pay the accorded amount to return the credit (Guillen and Artis 1992). Other classic methods used to face the credit scoring problem, include classification and regression trees, Markov chains, linear models and linear programming, graphical models, logistic regression, and many more (Thomas 1998), (Thomas 2000). On the other hand, a number of intelligent methodologies has appeared during the last years focusing on credit scoring, such as neural networks (Desai *et al.* 1997), genetic algorithms (Desai *et al.* 1997), machine learning (Carter and Catlett 1987), as well as data mining and bayesian networks, see also (Thomas 1998), (Thomas 2000). Note that, in literature there is no objective or unified way of measuring

the success of different methodological approaches in the classification and decision support concerning credit problems, as a variety of different decision variables, categories of credit risk, or other adjustments and hypotheses, is involved in each proposed attempt. According to the above, the present paper is not of course, the first attempt of facing credit risk with the aid of computational intelligence techniques, still, it proposes a strong and robust computational intelligence methodology, i.e. combination of inductive learning and genetic programming for the production of either competitive or cooperative rules. This methodology consists of two steps. First step is the feature selection. This procedure reduces the input data that will be used in the second step. Feature selection is performed using inductive machine learning. Second step is the production of a fuzzy rule-based system for classification and for knowledge extraction. The approach uses a fuzzy rule-based system, built also by genetic programming. This model usually obtains the highest accuracy and generalization degree, due to the use of fuzzy concepts, i.e. due to its ability to handle ambiguity and fuzziness contained in real world data sets. On the other hand, the outcome is less comprehensible, as the involvement of the membership functions of the variables are needed in order for the experts to apply the decision rules manually.

The paper is organized as follows: A brief reference to concepts related to credit management and a short presentation of the data that constitute the specific problem to be solved is given in section 2. In section 3 we briefly present the computational intelligent approach. In section 4 we discuss the acquired results from the application of the model. Finally, in section 5 we draw concluding remarks.

2. CREDIT SCORING

The problem analyzed in this paper, lies in the decision making process from the bank’s viewpoint, on whether to grant a loan to an enterprise and with what amount of certainty, based only on the data gathered from the application of the potential client.

Our experimental data was gathered from 4 different sources, which had both numerical data (in the form of percentages, indexes or simple numbers), and qualitative ones (consisted of terms). In the classic system that banks use to date, the qualitative data are transformed into numerical with the use of certain transformation scales. For the needs of our modeling, we further normalized these data to a range between zero and one.

As far as the application area is concerned, it is the relevant department of a financial organization that grants loans to firms. In this area, highly qualified experts are used, that have a firm knowledge of the problem to deal with, i.e., the granting of a loan to firms that issue specific accounting processes and rules, and have a lot of field experience.

In order to achieve the goal of managing credit scoring with the use of computational intelligence techniques, our data fulfilled the criteria below:

- They represent the general population
- They are homogenous
- They have the least possible noise
- There is a substantial amount of cases (all the recent cases were included in the sample)
- All the different (posterior) classification categories are covered in the sample

The particular European bank of the private sector examined in this paper, acts in a highly tourist region of the EU, a region with rather insignificant industrial or agricultural activities. The specific bank has already been using a classic application credit scoring system during the last years. This system keeps as input, 76 fields in total. For each of the 124 potential customers that had applied for receiving a loan during the past from the examined bank, there were 76 attributes that covered an area of up to three years before granting the loan. The bank classifies these cases in one of the following classes:

- AC: risk-free (or low risk)
- AV: average risk
- UO: under observation (surveillance)

- HR: high risk
- IW: (in) weak (i.e. impossible to receive a loan)

The challenge for computational intelligence is to provide models that can be used either for knowledge extraction –in the form of rules- or even to work as automated decision-making assistants. Note also that, although a standard process already exists within the examined bank, for deciding on whether to grant a requested loan or not, very often appear other forms of pressure applied to the managerial staff of the bank in order for a "weak" applicant to get the loan. In this sense, the proposed systems through this paper can only perform competitively to the experts' performance. Their extra advantage could be that a computer program that is difficult to be interfered, is always considered a non-deviating judge for humans, compared to the "flexibility" of human experts. So its adoption can absorb any informal attempt to affect the decision making process. Our approach is presented in the next section, discussing the methodology we selected to apply for this task.

3. THE COMPUTATIONAL INTELLIGENCE MODEL

This section presents the computational intelligent components involved in our approach and discusses the proposed architecture. The reader should advise (Zimmermann *et al*., 2001) and (Nilsson 1998), for a variety of modern methods and applications related to computational intelligence in the new era. The model consists of a two-step procedure. The first step is the feature selection. This procedure is covered in sub-section 3.1. The second step is classification and knowledge extraction. The proposed model uses type-constrained genetic programming. This model is presented in sub-section 3.2. In Figure 1, we show graphically the operation of our computational intelligent model.

3.1. FEATURE SELECTION

Inductive machine learning works as a feature selection mechanism, due to its ability to cope with mixed-mode data domains. This makes the possibility of applying a genetic programming technique easier and faster –a greedy in resources approach but proved as effective in long-term generalization according to literature- on the examined data. We selected to apply the inductive learning approach while standard feature selection techniques (i.e. based on similarity norms) would require that the credit data is all of numerical nature.

3.2. CLASSIFICATION AND KNOWLEDGE EXTRACTION

We first apply feature selection by the inductive machine learning is used, as mentioned above, to reduce the size of input data. Second, we apply type constraints in the genetic programming trees, in order to improve the readability of the solutions and speed up the search by reducing the number of possible node combinations - thus the search space. In the following sub-section, the theoretical background of the approach is briefly discussed.

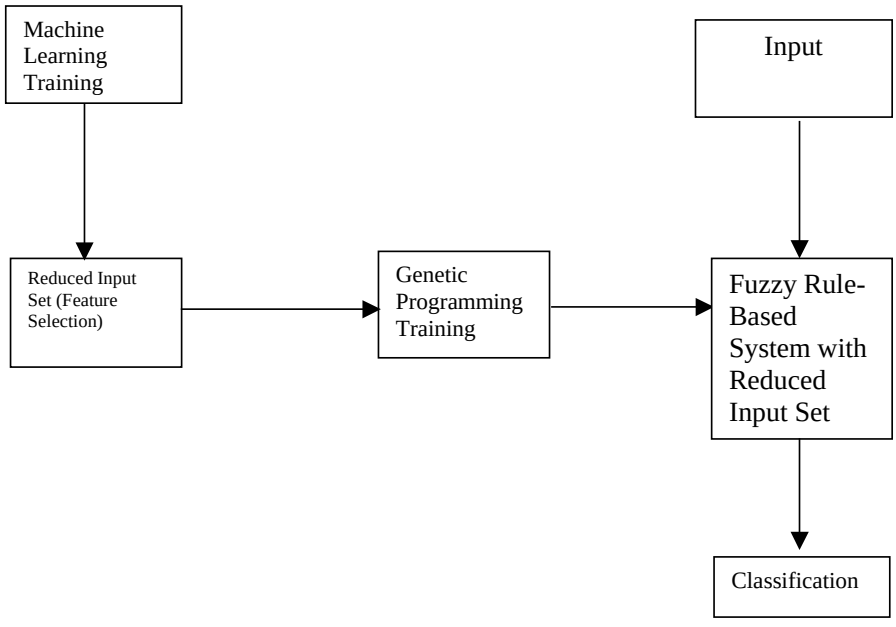

Figure 1: The Model Architecture

3.2.1. Fuzzy Rule-Based Systems by Type-Constrained Genetic Programming

As previously stated, the methodology examined in this work, uses fuzzy rule-based systems produced by genetic programming. Hence, a brief introduction to fuzzy sets follows. These sets are an extension to the classic sets that have a crisp boundary. According to (Zadeh 1965), fuzzy sets play an important role in human thinking. In fuzzy sets, the transition for a value from belonging to a set and not belonging to the set is gradual and characterized be the membership function (Jang et al. 1997). A fuzzy set is defined as:

$$A = \{(x, \mu_A(x)) \mid x \in X\}$$

where the $\mu_A(x)$ is a *membership function* for the fuzzy set *A*. The membership functions are described as mathematical formulas. The *X* is called the *universe of discourse*, and it may be comprised of discrete or continuous values. When the universe of discourse X is a continuous space, several fuzzy sets are used, most times covering the X uniformly. These fuzzy sets often are given linguistic terms such as "Small" or "Medium", thus they are called *linguistic variables*. These linguistic variables are used in fuzzy rules, which are interpreted as fuzzy relations using *fuzzy reasoning*. Fuzzy reasoning contains inference rules, which derive conclusions from a set of fuzzy rules and input data. A fuzzy if-then rule can be in the form:

if x is A then y is B

where the "x is A" is the antecedent (or premise) set, and "y is B" is the consequent (or conclusion) set. In fuzzy reasoning, the traditional two-valued logic, the *modus ponens*, is used in a generalized form. Namely, a fact may be more or less true, based on the truth of another fact. In the well-known *Mamdani*-classifier model using the *max-min* composition, see also (Jang et al. 1997), several steps are followed to perform fuzzy reasoning. Firstly, we compare the input data with the antecedent sets of the fuzzy rules and we get the degrees of compatibility (called *weights*) with respect to these antecedent sets. Then, we combine these degrees using fuzzy “AND” or “OR” to obtain a *firing strength*, which shows the degree that a rule is satisfied. The *max-min* criterion -when only “AND” operators are used- will assign as firing strength the smaller (*min*) of the antecedent degrees of compatibility. Finally, we obtain the

overall output between the consequent sets of the rules. The rule with the larger (*max*) firing strength will be the system's output. The genetic programming procedure is used to create and test such fuzzy rule-based models. As stated in the previous sections, a BNF-grammar is used to enable only valid trees. The definition, of the constructed BNF-grammar, is shown in Table 1.

Table 1: BNF grammar of program trees for Fuzzy Rule-Based Systems. The trees are described in a prefix notation. Words in bold denote valid program nodes.

Grammar used for the GP Tree	
<TREE>	::=<RL> |<RULE>
<RL>	::=**RL** <TREE> <TREE>
<RULE>	::=**RULE** <COND> <CLASS>
<COND>	::=<IF_A> | <IF_B> | <AND>
<IF_A>	::=**IF_A** <INP_A> <FS_A>
<IF_B>	::=**IF_B** <INP_B> <FS_B>
<AND>	::=**AND** <COND> <COND>
<CLASS>	::=**THEN** <OUT> <CLASS_VALUE>
<FS_A>	::=**S_A | M_A | L_A**
<FS_B>	::=**VS_B | S_B | M_B | L_B | VL_B**
<INP_A>	::=**X1**
<INP_B>	::=**X2 | X3 | X4**
<CLASS_VALUE>	::=**CLASS1 | CLASS2 | CLASS3** …
<OUT>	::=**Y**

4. RESULTS

The proposed methodology was applied to a specific data set, supplied from a regional European private bank. The data set corresponds to a real credit management problem that the above bank faces during the last years. The success of the bank through these years in granting loans to reliable applicants and avoiding the unreliable customers is relatively poor. The bank would like (a) to obtain an objective computer assisted methodology for deciding on loans

granting, and (b) to explore knowledge and possible interrelations hidden inside the past data that the bank has stored in its files through the years. The sample consists of 124 firms (cases) applying for a loan, which contains a total number of 76 requested financial attributes for each of the applicants, both of numerical and linguistic nature. In the following paragraphs we first show the feature selection results using the machine learning approach. Then, we present the classification results.

4.1 FEATURE SELECTION

The application of an inductive learning methodology, in the examined set of credit scoring data, reduced the number of important parameters from 76 to 16. A brief description of the attributes is given and their range of values so that the reader can understand the meaning of the outcome of this application of inductive learning in the specific field of credit scoring. The promoted attributes are:

- Debt / Equity a year ago: Index in the form of percentage (Continuous).
- Annual change in Sales a year ago: Percentage (Continuous).
- Products – Services (Quality): Bad, Average, Good, Exceptional.
- Sector's Net Profit Margin: Index average (%) for the sector (Continuous).
- Net Profit Margin a year ago: Index (%) (Continuous).
- Geographical Coverage: Certain Areas, Local, Widely Local, National.
- Years in business: Years that the firm is in business (Integer).
- Net Income a year ago: The firm's net income (Continuous).
- Sector's Average Inventory (Continuous).
- Number of Products: Number of products-services the firm provides (Cont.).
- Business's Future: How the executives of the bank foresee the firm's future in the sector. Range: Insufficient to Adequate, Adequate, Good, Exceptional.
- Sector's Debt / Equity: Sector index average in percentage (Continuous).

- Security Margin 2 years ago: Index in the form of percentage (Continuous).
- Sector's Accounts Receivable: (Continuous).
- Accounts Receivable a year ago: (Continuous).
- Quick Ratio: Index in the form of percentage (Continuous).

In the next paragraph, for the production the fuzzy rule-based model, the exclusive use of these parameters is assumed.

4.2 CLASSIFICATION AND KNOWLEDGE EXTRACTION

Before the presentation of the acquired decision rules for credit management, we comment on the selected shape and style of the membership functions used to construct the fuzzy genetic programming approach for each variable, see Figure 2. A bell-type membership function was selected for its similarity to the normal distribution, which corresponds to the most frequently appearing distribution in nature and phenomena of the real world. There are 9 different linguistic areas within each fuzzy set. In terms of comprehensibility, 3 or 5 areas would be ideal. On the other hand, dividing the whole range of a parameter to many different linguistic areas, we may lose completely the comprehensibility issue, but we gain in accuracy of the solution obtained. Thus, our choice of 9 linguistic areas in the membership functions corresponds to an average choice between comprehensibility and accuracy.

The best solution obtained for the fuzzy genetic programming approach had an accuracy of 76.19% on the training set (32 out of 42 cases correctly classified), and almost 60% in the test set (50 out of 84 cases correctly classified). However, it had a very high degree of generalization as well as of comprehensibility. Note, that now the split of the cases to training and testing ones is even smaller, nevertheless the difference between classification accuracy of the training and test set has been reduced compared to the previous two approaches. In other words, the reader should observe that this model can make a considerable generalization from a training phase that took place only in one third of the data set, i.e. only from a small set 42 cases. In contrast,

inductive learning was using 90% of the data for training and only 10% of the data for testing. As we state in the last section, there are many different aspects to be taken into account, when comparing results of competitive methods for classification to similar domains of application but not alike amounts of data and experimental conditions.

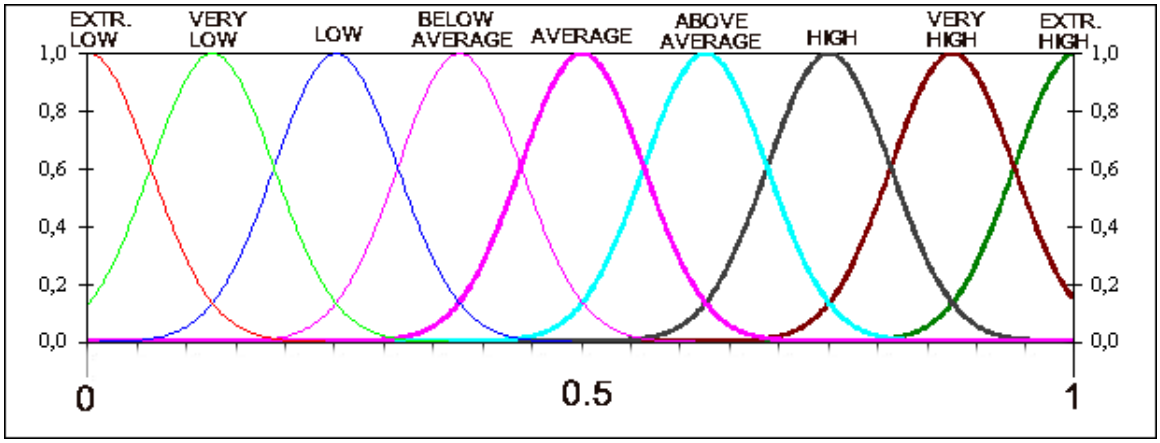

Figure 2: Membership functions for the fuzzy rule-based model

Note that the produced rules are competitive, i.e. one of them is activated each time, but their order does matter in the final decision. The outcome is very easy to automate and apply it directly to new data, given that the membership function range has been pre-determined (Table 2) with the aid of credit experts. In Figure 3 we present the rules acquired.

The rule set shown in Figure 3, does not contain reference to the risk-free category. We consider that this may happen because of two reasons. First, it could happen due to a very small number of cases in the set that made generalization for this class impossible. Second, it may happen due to the application of the fuzzy set modeling, which was focused on the average-to-high risk cases. Thus, this solution results in the judgment that a risk free case is risky, instead of accepting it as risk-free, actually risky cases. However, this bias is not seriously troublesome. According to (Wuthrich 1997) a 5 times higher penalty should be given to a misclassification of a very risky case as risk-free, than the opposite type of fault, i.e. the misclassification of a risk-free case as very risky.

Table 2: Bounds for each variable (before normalization) used in the fuzzy rule-based system.

Var	**x1**	**X2**	**x3**	**X4**	**X5**	**x6**	**x7**	**x8**	**x9**	**x10**	**x11**	**x12**	**x13**	**x14**	**x15**	**x16**
Low	-56,08	-1	1	0	-0,44	0	1	1	-742	0	0	-3,36	0	0	-15,01	0
High	149,31	1,41	3	0,19	0,53	115	4	3	829,3	332,55	6,58	5,90	136,96	32	629,91	43,21

1. IF DEBT/EQUITY A YEAR AGO is EXTREMELY LOW **THEN** IN WEAK
2. IF NET PROFIT MARGIN A YEAR AGO is LOW **THEN** IN WEAK
3. IF ACCOUNTS RECEIVABLE A YEAR AGO is BELOW AVERAGE and SECTOR'S ACCOUNTS RECEIVABLE is AVERAGE and ACCOUNTS RECEIVABLE A YEAR AGO is EXTREMELY LOW **THEN** HIGH RISK
4. IF SECTOR'S DEBT/EQUITY is ABOVE AVERAGE and NET PROFIT MARGIN A YEAR AGO is LOW and ACCOUNTS RECEIVABLE A YEAR AGO is BELOW AVERAGE and DEBT/EQUITY A YEAR AGO is EXTREMELY HIGH and SECTOR'S ACCOUNTS RECEIVABLE is AVERAGE **THEN** HIGH RISK
5. IF NET PROFIT MARGIN A YEAR AGO is LOW **THEN** UNDER OBSERVATION
6. IF DEBT/EQUITY A YEAR AGO is AVERAGE **THEN** UNDER OBSERVATION
7. IF SECTOR'S DEBT/EQUITY is ABOVE AVERAGE and NET PROFIT MARGIN A YEAR AGO is LOW and NET PROFIT MARGIN A YEAR AGO is LOW and NET PROFIT MARGIN A YEAR AGO is LOW **THEN** UNDER OBSERVATION
8. IF SECTOR'S ACCOUNTS RECEIVABLE is AVERAGE and NET PROFIT MARGIN A YEAR AGO is LOW **THEN** UNDER OBSERVATION
9. IF NET PROFIT MARGIN A YEAR AGO is LOW and PRODUCTS-SERVICES (QUALITY) is BELOW AVERAGE and SECTOR'S DEBT/EQUITY BELOW is AVERAGE and NET PROFIT MARGIN A YEAR AGO is LOW and PRODUCTS-SERVICES (QUALITY) is BELOW AVERAGE and SECTOR'S DEBT/EQUITY ABOVE is AVERAGE **THEN** UNDER OBSERVATION
10. IF SECTOR'S ACCOUNTS RECEIVABLE is AVERAGE and SECTOR'S DEBT/EQUITY is BELOW AVERAGE and ACCOUNTS RECEIVABLE A YEAR AGO is EXTREMELY LOW **THEN** UNDER OBSERVATION
11. IF DEBT/EQUITY A YEAR AGO is HIGH **THEN** AVERAGE RISK
12. IF GEOGRAPHICAL COVERAGE is AVERAGE **THEN** AVERAGE RISK
13. IF YEARS IN BUSINESS is EXTREMELY HIGH **THEN** AVERAGE RISK

IF ACCOUNTS RECEIVABLE A YEAR AGO is BELOW AVERAGE and PRODUCTS-SERVICES (QUALITY) is BELOW AVERAGE and SECTOR'S DEBT/EQUITY BELOW is AVERAGE **THEN** AVERAGE RISK

Figure 3. Fuzzy Rule-Based System, derived by genetic programming.

5. CONCLUDING REMARKS AND FURTHER RESEARCH

The paper presented a computational intelligence based approach, for managing the problem of credit risk. The comparison of the presented intelligent architectures and of the classic statistical models for credit management that is currently used by most banks, results in the following advantages for the use of our approaches:

1. There is no need for a huge amount of cases for the extraction of the classifiers. Only a few representative cases of each class are enough to form a reliable classifier.
2. No mathematical knowledge is required of the user.
3. The processing time of a database in order to extract a classifier is rather limited even for large databases.
4. Due to the above fact it is easy to incorporate new cases to the database in the short future, in order to extract improved or adapted classifiers.
5. A much smaller amount of attributes, for the classification process, is finally used for decision support regarding credit management problems.
6. The acquired rule sets can also very easily be encoded to simple computer programs and used immediately as a decision assistant for risk problems.

Current research of the authors is directed in the following three ways:

First, construction of alternative computational intelligence methodologies for credit management, involving inductive learning as feature selection technique, neural networks and fuzzy or neuro-fuzzy rule-based systems. The primary concern of the authors, for architectures involving neural networks, is dealt with comprehensibility issues. Neural networks operate as black-box systems, hence could be considered as among the less-comprehensible modules in hybrid architectures. Neuro-fuzzy systems ensure some degree of comprehensibility nevertheless their results remain less interpretable than those of the techniques presented here.

Secondly, the use of standard feature selection techniques, instead of inductive machine learning, in the hybrid architecture.

Finally, collection of additional credit data, in order to cross-validate our conclusions on the application of hybrid intelligent techniques to the problem of credit management, as well as to further check the robustness and adaptivity issues of the proposed methodologies.

References

Boyes W.J., Hoffman D.L. and Law S.A. 1989. An Econometric Analysis of the Bank Credit Scoring Problem. *Journal of Econometrics*, Vol. 40, 3-14.

Carter C. and Catlett J. 1987. Assessing credit card applications using machine learning. *IEEE Expert* Vol. 2, 71-79.

Desai V.S., Convay D.G., Crook J.N. and Overstreet J.A. 1997. Credit scoring models in the credit union environment using neural networks and genetic algorithms. *IMA Journal of Mathematics Applied to Business & Industry* Vol. 8, 323-346.

Guillen M., Artis M. 1992. Count Data Models for a credit scoring system. *Third Meeting on the European Series in Quantitative Economics and Econometrics on "Econometrics of Duration, Count and Transition Models, Paris, Dec.10-11, 1992*, 1-9.

J-S.R. Jang, 1998, "Neuro-fuzzy modeling for nonlinear dynamic system identification", in E.H. Ruspini, P.P. Bonissone, W. Pedrycz (eds.), *"Handbook of Fuzzy Computation", Institute of Physics Publishing*, Dirak House, Temple Back, Bristol BS1 6BE UK, 1998

Koza J. R. 1992. *Genetic Programming – On the Programming of Computers by Means of Natural Selection*. The MIT Press.

Koza J. R., Forrest H.Bennett III, David Andre, Martin A. Keane 1999. *Genetic Programming III*, Morgan Kaufmann Publishers, Inc.,

Lewis E.M.1992. *An Intoduction to Credit Scoring*. Athena Press, San Rafael, CA, USA.

Nilsson N. 1998. *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann.

Steenackers A. and Goovaerts M.J. 1994. A Credit Scoring Model for Personal Loans. Insurance*: Mathematics and Economics* 8, 31-34.

Thomas L.C. 1998. Methodologies for classifying applicants for credit. In *Hand D.J. and Jacka, S.D. (Eds.), Statistics in Finance*. Arnold, UK, 83-103.

Thomas L.C. 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting* 16, 149-172.

Wuthrich B. 1997. Discovering probabilistic decision rules. International Journal of Intelligent Systems in Accounting, Finance and Management 6: 4, 269-277.

Zadeh L.A. 1965. Fuzzy Sets. *Information and Control* 8, 338-353.

Zimmermann H-J., Tselentis G., Van Someren M., Dounias G. (Eds.). 2001. *Advances in Computational Intelligence and Learning: Methods and Applications.* Kluwer Academic Publishers.