Learning Financial Rating Tendencies with Qualitative Trees

Llorenç Roselló[1], Mónica Sánchez[1], Núria Agell[2], Francesc Prats[1]

[1]GREC Group and Ma2-Universitat Politècnica de Catalunya
[2]GREC Group and ESADE-Universitat Ramon Llull

Abstract

Learning financial rating tendencies requires knowledge of the ratios and values that indicate a firm's situation as well as a deep understanding of the relationships between them and the main factors that can modify these values. In this work, the Qualitative Trees provided by the algorithm QUIN are used to model financial rating and to learn its tendencies. Some examples are given to show the system's predictive capabilities. The rating tendencies and the variables that most influence those tendencies are analyzed.

1. Introduction

In this paper, a learning process to induce a qualitative model providing a causal interpretation between the variation of some input variables and the tendency of the output variable is described. To obtain the model, the algorithm QUIN (QUalitative INduction) is used [Šuc and Bratko, 2001, 2003, 2004]. QUIN addresses the problem of the automatic construction of qualitative models across an inductive learning of numerical examples by means of Qualitative Trees. These trees have qualitative functional restrictions inspired in the and-predicates introduced by Forbus [Forbus, 1984] in their writings. A qualitative tree defines a partition in the attributes space in zones with a common behavior of the chosen variable. The algorithm was designed and implemented by Dorian Šuc and Ivan Bratko. This qualitative model is especially suitable for analysing financial rating tendencies because it allows one to analyze how the variables describing the state of a firm at a given moment can modify its valuation rating.

Big data sets containing patterns or examples with many attributes are unmanageable with QUIN because of its algorithmic complexity. Considering that this is a characteristic of the case of study, it has been necessary to reduce the number of variables and to group the sets of data in order to simplify the available data set. Data have been provided by Thomson Financial and Standard & Poor's, and correspond to 1177 firms represented by 46 input variables together with their financial rating given by Standard & Poor's for 2003. The input variables are ratios that try to capture aspects of liquidity, profitability, financial structure, size and turnover or level of activity of the company.

The QUIN algorithm was applied using data for firms operating in Canada, Japan, a group containing European firms, and random samples of firms in the USA.

The structure of the paper is as follows: in Section 2 outlines the preprocessing of data through factorial analysis. Section 3 gives general descriptions of qualitative trees and the QUIN algorithm. Section 4 explains the general approach to financial ratings, the experiments undertaken, and the results obtained. The concluding section sets out findings and suggests new ways for solving the problems presented.

2. Data preprocessing

In the words of QUIN's authors "QUIN cannot efficiently handle large learning sets, neither in terms of examples nor the attributes" [Zabkar et al, 2005] due to its complexity. When either the number of patterns or the number of attributes is too large for the algorithm, data pre-processing is needed.

In this work, the set of patterns has been partitioned following the categories of a nominal qualitative variable. The country where the firm has its headquarters has been used to partition the set of 1177 worldwide firms. In addition, the number of input variables considered in the learning process has been reduced by using factorial analysis. SPSS software has been used to extract principal components for the whole set of 46 variables, turning out 5 principal components to explain 60% of the total variability. QUIN has been applied using data of these 5 principal components corresponding to companies of Canada, Japan, a group containing European firms, and some random samples of firms operating in USA.

3. Qualitative Trees and QUIN

In this section the concept of qualitative tree is outlined as a previous concept to introduce the QUIN algorithm used for the case of study.

Given a set of *N* patterns, each pattern described by *n+1* variables where $X_1,\ldots,X_n$, are the *attributes* and X_{n+1} is the *class*, the goal is learning zones of the space that should present a common behavior of the class variable. These zones are described by means *of qualitative trees*. A qualitative tree is a binary tree with internal nodes called *splits*; its leaves are *qualitatively constrained functions*. From now on these functions will be denoted by QCF. The internal nodes define partitions of the space of attributes. In each node there is an attribute and a value of this attribute. The QCF define qualitative constraints of the class variable in the following way: if $F: R^n \rightarrow R$ is a map that associates to each *n* attributes a value of the class variable, a QCF associated to the function *F* and to a *m*-tuple $(x_1, \ldots, x_m) \in R^m$, with $m \leq n$, is denoted by $F^{s1,\ldots,sm}(x_1,\ldots,x_m)$, where $s_i \in \{+,-\}$, and it means that:

$$\left.\frac{\Delta F}{\Delta x_i}\right|_{\Delta x_j = 0,\, j \neq i} = s_i$$

In other words, $s_i = +$ means that *F* is a strictly increasing function respect to the variable x_i, and it is strictly decreasing when $s_i = -$. Figures 1 and 2 give a simple example of qualitative tree; figure 1 shows the graphical representation of the function $F(x,y) = \dfrac{x}{y^2+1}$

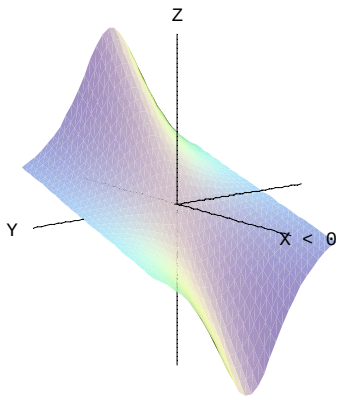

Figure 1. The plot of $F(x,y) = \dfrac{x}{y^2+1}$

And figure 2 shows the induced qualitative tree of this function.

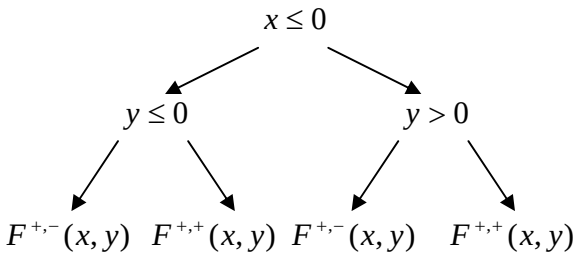

Figure 2. The qualitative tree of $F(x,y) = \dfrac{x}{y^2+1}$

One should note that in general the explicit expression of the function *F* is unknown.

In order to decide which QCF is better adjusted to a given set of patterns, the qualitative changes q_i of variables x_i are used, where $q_i \in \{pos, neg, zero\}$, in such a way that if $\Delta x_i > 0$ then $q_i = pos$ and if $\Delta x_i = 0$ then $q_i = zero$.

Let us define de QCF-prediction $P(s_i, q_i)$, $(s_i \in \{+,-\})$ as

$$P(s_i,q_i) = \begin{cases} \text{pos}, & \text{if } (s_i = + \wedge q_i = \text{pos}) \vee (s_i = - \wedge q_i = \text{neg}) \\ \text{neg}, & \text{if } (s_i = + \wedge q_i = \text{neg}) \vee (s_i = - \wedge q_i = \text{pos}) \\ \text{zero}, & \text{otherwise} \end{cases}$$

Then, for any pair of patterns (e,f) a *qualitative change vector* is formed, being each component of this vector $q_{(e,f),i}$ defined by:

$$q_{(e,f),i} = \begin{cases} \text{pos}, & \text{if } x_{f,i} > x_{e.i} + \varepsilon \\ \text{neg}, & \text{if } x_{f,i} > x_{e.i} - \varepsilon \\ \text{zero}, & \text{otherwise} \end{cases}$$

Where $x_{f,i}$ is the *i*-th component of *f*. The parameter ε is introduced to solve the cases with tiny variations: 1% of the difference between maximal and minimal value of the *i*-th attribute. Once these concepts have been introduced, the method to choose the QCF that better describes data will be explained. What constitutes `better' in this context will be discussed later.

A QCF, $F^{s_1,\ldots,s_n}$, that describes the behavior of the class X_{n+1} is *consistent* with a vector of qualitative changes if all QCF-predictions $P(s_i, q_{n+1})$ are non negative with at least one positive. In other words, it is consistent when the vector of qualitative changes does not contradict the QCF. If there are simultaneously positive and negative QCF-predictions or when all the predictions are zero, then *there*

is an ambiguity in the prediction of the QCF with the vector of qualitative changes. Finally, a QCF is *inconsistent* with a vector of qualitative changes if it is neither consistent nor ambiguous.

For each QCF an error-cost is defined from the number of consistent vectors of qualitative changes and the number of ambiguous vectors of qualitative changes (this has to be verified for all possible qualitative change vectors for the problem under consideration). This error-cost gives a measurement of the suitability of the QCF function to describe data.

The QUIN algorithm constructs the qualitative tree with a greedy algorithm that goes from top to bottom similar to the ID3 [Quinlan, 1986]. Given a set of patterns, QUIN computes the error-cost for each one of the QCF found for each partition, and chooses the partition that minimizes the error-cost of the tree. The error-cost of a leaf is the error-cost of the QCF that there is in this leaf. The error-cost of a node is the error-cost of each one of the sub-trees plus the cost of the division.

4. Case of Study: Financial Rating

The case below falls within the development frame of the AURA research project, which sets out to adapt soft-computing techniques to the study of the financial rating tendencies by using qualitative reasoning.

The main goal of the project is to use these techniques to extract knowledge and allow prognosis. In particular, in this paper, a qualitative system based on QUIN is considered to represent the factors that are relevant in computing credit risk. Using factorial analysis, five principal components have been extracted and used to study tendencies of the level of risk. QUIN has been applied to several sets of firms (characterized for these five components and their Standard & Poor's rating) corresponding to different countries.

4.1. Financial rating

The rating is an attempt to measure the financial risk of a given company's bond issues. The specialized rating agencies, such as Standard & Poor's, classify firms according to their level of risk, using both quantitative and qualitative information to assign ratings to issues. Learning the tendency of the rating of a firm therefore requires the knowledge of the ratios and values that indicate the firms' situation and, also, a deep understanding of the relationships between them and the main factors that can modify these values.

The processes employed by these agencies are highly complex and are not based on purely numeric models. Experts use the information given by the financial data, as well as some qualitative variables, such as the industry and the country or countries where the firm operates, and, at the same time, they forecast the possibilities of the firm's growth, and its competitive position. Finally, they use an abstract global evaluation based on their own expertise to determine the rating. Standard & Poor's ratings are labeled AAA, AA, A, BBB, BB, B, CCC, CC, C and D. From left to right these rankings go from high to low credit quality, i.e., the high to low capacity of the firm to return debt.

4.2. Learning Financial Rating Tendencies

The problem of classifying firms by using their descriptive variables has already been tackled by several authors [Ammer, J.M. and Clinton, N., 2004]. The goal of this paper is to analyze the variables that influence variations in ratings and how this influence is expressed. Data are the financial results presented by 1177 companies worldwide and the rating that Standard & Poor's granted in reference to year 2003. Each firm is considered as a pattern, described by 46 input variables, and the variable class is the rating. The QUIN algorithm is used to learn which the qualitative tree associated to this problem is.

4.3. Experimental Results

The experiment began with preprocessing of the data. The algorithmic complexity of the QUIN, especially when the number of patterns, as well as the number of attributes, is considerable, making it advisable to start with the following two steps:

1. To limit the number of patterns, by grouping the companies, in particular: Canada (83 patterns), Japan (26 patterns), a group containing all European firms (129 patterns), and some random samples of firms operating in USA (between 60 and 80 patterns).
2. To reduce the number of variables treated, by using factorial analysis extracting principal components.

Several tests have been carried out for firms in the above selected groups. The 46 input variables are grouped into five groups, each group describing a certain financial characteristic. Using SPSS software for the whole set of 46 variables, turned up 5 sufficient principal components to explain 60% of the whole set of patterns variability (and thus learn the financial rating tendency) - the corresponding results are commented upon below. It has to be pointed out that, in addition, experiments with more principal components were carried out; specifically, with 7 principal components, explaining 63% of variability, with 9, explaining 75% of variability, and with 13 principal components (98%). It has been seen that if the number of components increases, then the qualitative trees corresponding to Europe and to the American groups become very complex, making the observation of behavioral patterns difficult. By contrast, the qualitative tree corresponding to Japan with $n \geq 7$ principal components becomes more simplified than in the case $n < 7$, and invariant from 7 on.

The obtained results show certain common trends in the rating tendencies in the European and American groups of firms, whereas Japanese firms exhibit different behavior.

When applying QUIN using data in the case of 5 principal components corresponding to the sets of data of the mentioned groups, the obtained qualitative trees show the rating tendencies and reveal the most relevant variables. The five principal components are named F_1, …, F_5. Three of them, in particular F_1, F_4 and F_5 are related to liquidity, F_2 is related to financial structure and F_3 is related to profitability. The considered order for the rating has been the order given by Standard & Poor's, but from less to more risk, i.e., $AAA \prec AA \prec \ldots \prec D$. The results are represented in the following figures.

$$R^{-,+,-,-,+}(F_3, F_5, F_4, F_1, F_2)$$

Figure 3: Canadian firms' induced qualitative tree

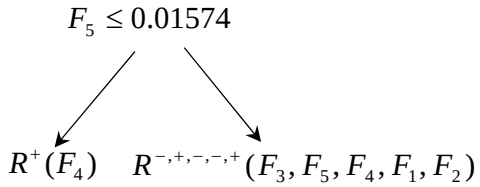

Figure 4: European firms' induced qualitative tree

$$R^{-,+,-,-}(F_3, F_5, F_4, F_1)$$

Figure 5: Random samples 1, 3 and 4 of USA firms' induced qualitative tree

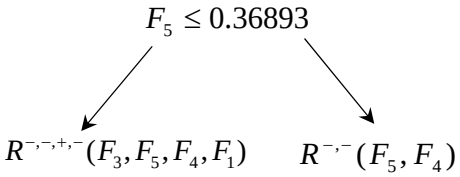

Figure 6: Random sample 2 of USA firms' induced qualitative tree

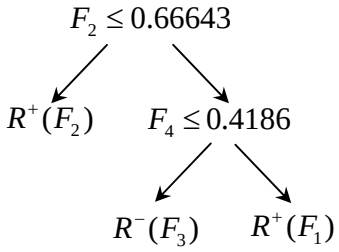

Figure 8: Japanese firms' induced qualitative tree

The induced qualitative tree s obtained for the different countries, though not identical, show certain common characteristics that provide useful information about the problem. With these five principal components, the rating dependency is the same for Canada, Europe and the random samples of USA. Mostly, trees with only one leave are obtained. Within the set of variables corresponding to liquidity, rating always increases with respect to F_1 and F_4, and decreases with respect to F_5. Rating always increases with respect to the profitability-related component. The component related to financial structure appears in few trees, and, at this level, it can be concluded that this component does not give much qualitative information.

One should note that in general trees obtained from different data sets, even though deduced from the *same function*, are not the same. For instance, in the case considered in figures 1 and 2, the qualitative tree corresponding to $(x,y) \in [-1,1] \times [-1,1]$ is totally different to the qualitative tree corresponding to the domain $(x,y) \in [3,7] \times [3,7]$. In addition, when the explicit expression of the function F is unknown, the complexity of the problem increases.

In the presented case of study, one possible explanation of the difference between trees is that each tree has been constructed over a different domain. The examination of the numerical data shows that, for example, the range of the third factor in Japan is approximately [-0.066, 0.095], whereas in Canada, as well as in the first, third and fourth random sets of USA firms, the range is approximately [-0.3, 0.3].

Therefore, it is perfectly natural that firms from very different countries present different features, whereas firms operating in countries under similar economical conditions (as it can be the case of Canada and USA) show similar features. On the other hand, the more factors are considered the more differences are able to appear. To sum up, in this case of study, induced qualitative tree s of different areas are being compared, and, in these different areas the rating behavior must not be neither the same nor very similar.

Conclusion and Future Work

This paper presents on-going work, which provides new strategies for credit risk prediction. The choice and definition of the variables involved, as well as study of the influence of each variable on the final result, have been analyzed.

The induced qualitative tree s provided by QUIN lead to a useful model for learning rating tendencies and studying to what extent ratings depend on several variables representing different financial features. When using the five principal components, the qualitative trees provided by QUIN algorithm for different sets of European and American firms show internal common trends.

In the case studied, the QUIN algorithm was used for an output qualitative variable described on an ordinal scale. In

general, due to the non deterministic intrinsic nature of the problem, the expected results are a probability function for the rating corresponding to different values of the input variables. However, use of the QUIN algorithm provides qualitative information about the monotonic behavior of rating with respect to financial features.
Future work will cover the speed of rating tendencies (i.e., how "fast" or "slow" ratings change) by using orders of magnitude descriptions.
The particular evolution of the rating of a given firm and its prediction from the previous rating and the values of its present financial ratios is currently being studied.

Acknowledgement

This work has been partially financed by MEC (Ministerio de Educación y Ciencia): AURA project (TIN2005-08873-C02-01 and TIN2005-08873-C02-02)

References

Ammer, J.M. and Clinton, N: *Good news is no news? The impact of credit rating changes on the pricing of asset-backed securities.* International Finance Discussion Paper, no 809, Federal Reserve Board, (2004) July

Forbus, K: *Qualitative Process Theory*, Artificial Intelligence (1984) 24:85-168

Quinlan, J.R. *Induction of decision trees.* Machine Learning, 1, pp. 81-106, 1986.

Šuc, D., Bratko, I. *Induction of Qualitative Trees.* In L. De Raedt and P. Flach editors, Proc. 12th European Conference on Machine Learning, pages 442-453. Springer, 2001. Freiburg, Germany.

Šuc, D., Vladusic, D., Bratko, I.: *Qualitative Faithful Quantitative Prediction*, Proc. ŠŠŠ18th In. Joint Conf. on Artificial Intelligence, IJCAI.2003, pp. 1052-1057

Šuc, D., Vladusic, D., Bratko, I.: *Qualitative Faithful Quantitative Prediction.* Artificial Intelligence 158 (2004) 189–214

Zabkar, J., Vladusic, D., Zabkar, R., Cemas, D., Šuc, D., Bratko, I.: *Using Qualitative Constraints in Ozone Prediction,* Proc. 19th In. Workshop on Qualitative Reasoning QR05, pp. 149-156.