Qualitative Input Conditioning to Enhance RBF Neural Networks Generalization in Financial Rating Classification

Xavier Parra*

ESAII, Universitat Politècnica de Catalunya
Av. Víctor Balaguer, s/n – 08800 Vilanova i la Geltrú – Barcelona (Spain)
Xavier.Parra@upc.es

Núria Agell*

ESADE, Universitat Ramon Llull
Av. Pedralbes, 62 – 08034 Barcelona (Spain)
agell@esade.edu

Abstract

The rating is a qualified assessment about the credit risk of bonds issued by a government or a company. There are specialized rating agencies, which classify firms according to their level of risk. These agencies use both quantitative and qualitative information to assign ratings to issues. The final rating is the judgement of the agency's analysts and reflects the probability of issuer default. Since the final rating has a strong dependency on the experts knowledge, it seems reasonable the application of learning based techniques to acquire that knowledge. The learning techniques applied are neural networks and the architecture used corresponds to radial basis function neural networks. A convenient adaptation of the variables involved in the problem is strongly recommended when using learning techniques. The paper aims at conditioning the input information in order to enhance the neural network generalization by adding qualitative expert information on orders of magnitude. An example of this method applied to some industrial firms is given.

Introduction

The present paper aims at applying Neural Network techniques to predict credit of bonds issued by a government or a company. Predicting the rating of a firm therefore requires a thorough knowledge of the ratios and values that indicate the firm's situation and, also, a thorough understanding of the relationships between them and the main factors that can alter these values (Agell et al. 2000).

The application of learning based techniques to acquire the analyst's knowledge seems reasonable due to the special nature of the problem. The strong dependency on the expert's knowledge is the main reason that brought us to the connectionist approach proposed. However, how to represent the input and output variables of a learning problem in a neural network implementation of the problem is one of the key decisions influencing the quality of the solutions one can obtain. Moreover, this is especially important when qualitative information is available during training.

This application shows how Neural Network and Qualitative Reasoning techniques, and particularly orders of magnitude calculus (Piera 1995, Travé-Massuyès, Dague and Guerrin 1997), can be useful in the financial domain (Goonatilake and Treleaven 1996).

The paper gives a brief introduction to the neural network architecture used. Next, the process to prepare the reference scales of different qualitative variables to be operated is established. An application of this neural network to credit risk evaluation of a firm or an issue of bonds is presented as well. The paper finishes with some conclusions and also with some comments about the implementation of this application and the first results obtained with it.

RBF Architecture

Radial basis function networks (RBF) are especially interesting for the problem proposed since they are universal classifiers (Poggio and Girosi 1990) and the training can be much faster than for other neural architectures (e.g. MLP or SVM). Moreover, it is possible to extract rules from an RBF architecture. RBF have been traditionally associated with a simple architecture of three layers (Broomhead and Lowe 1988) (see Figure 1). Each layer is fully connected to the following one and the hidden layer is composed of a number of nodes with radial activation functions called radial basis functions. Each of the input components feeds forward to the radial functions. The outputs of these functions are linearly combined with weights into the network output. Each radial function has a

* Authors are members of the European Associated Laboratory of Intelligent Systems and Advanced Control (LEA–SICA).

local response (opposite to the global response of sigmoid function) since their output only depends on the distance of the input from a center point.

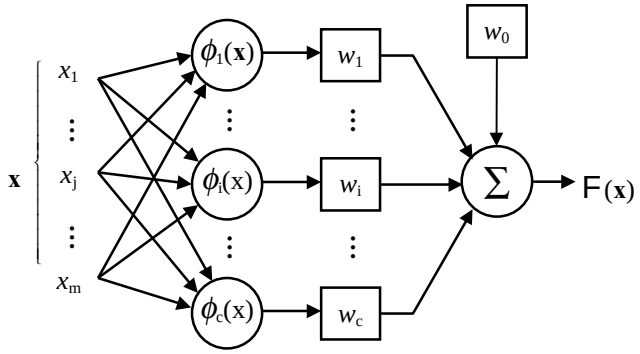

Figure 1: Radial basis function network architecture.

Radial functions in the hidden layer have a structure that can be represented as follows:

$$\phi_i(\mathbf{x}) = \varphi((\mathbf{x}-\mathbf{c}_i)^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{x}-\mathbf{c}_i)) \qquad (1)$$

where φ is the radial function used, $\{\mathbf{c}_i \mid i = 1, 2,\ldots, c\}$ is the set of radial function centers and **R** is a metric. The term $(\mathbf{x} - \mathbf{c})^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{x} - \mathbf{c})$ denotes the distance from the input **x** to the center **c** on the metric defined by **R**. There are several common types of functions used, though the Gaussian function is the most typical choice, combined with the Euclidean metric. In this case, the output of the RBF network is:

$$F(x) = w_0 + \sum_{i=1}^{c} w_i \exp\left(-\frac{\|\mathbf{x}-\mathbf{c}_i\|^2}{r^2}\right) \qquad (2)$$

where c is the number of basis functions, $\{w_i \mid i = 1, 2,\ldots, c\}$ are the synaptic weights, $\|\cdot\|$ denotes the Euclidean norm and r is the radius of the radial function.

The RBF learning algorithm is an incremental and evolutionary process. Its mathematical foundation is called *subset selection* and consists in comparing models made up of different subsets of elements drawn from the same fixed set of candidates. To find the best subset is usually intractable so heuristics must be used to search for a small but hopefully interesting fraction of the space of all subsets. However, the use of these heuristics does not guarantee that the solutions we get include the least number of elements needed to reduce the approximation error to a fixed value.

The heuristic method called *forward selection* is widely used with RBF networks (Chen, Cowan and Grant 1990). According to this method, the subset that must be determined is the subset of centers that fix the location of the radial functions in the input space. The method begins with an empty subset to which is added one basis function at a time. The center of the radial function added is selected among the whole set of input patterns and is the one that most reduces the approximation error. The learning process continues until some chosen criterion stops decreasing (e.g. generalized cross-validation).

Qualitative Input Conditioning

The problem of having to extract some information from qualitative values represented in heterogeneous references is not unusual. Many qualitative reasoning techniques are used to manage this kind of references. However, when the values have to be processed with a neural network it is not clear how to prepare the values in order to take into account the experts' knowledge.

In general, the performance obtained when using a neural network depends on the problem representation, i.e., the input and output representations. Moreover, when the neural network used is an RBF network this dependency on the representation is more critical since, as it is shown in 2, the radial function output is a function of a distance defined in the input space. Depending on the kind of problem, there may be several different kinds of variables that can be represented. Unfortunately, there is not a unique method to represent all of these variable kinds. However, it is common to use some of the following hints:

- *Real-valued attributes* are usually rescaled by some function that maps the value into the range 0…1 or –1…+1, in a way that makes roughly even distribution within that range. It is also common to rescale the values to mean 0 and standard deviation 1 by using a linear transformation.
- *Nominal attributes* with *m* different values are usually either represented using a 1-of-*m* code or a binary code.
- *Ordinal attributes* with *m* different values are represented by *m*-1 variables of which the leftmost *k* have value 1 to represent the *k-th* attribute value while all others are 0.

As it is pointed in 1, financial problems involve a strong understanding of ratios and values that, somehow, indicate different financial situations. Often, these ratios and values are represented through a real value though quite frequently the expert extracts information not strictly from that numerical value but from a more qualitative representation. For example, in front of a concrete value the expert is more likely to think in terms of *good*, *bad*, *very good*,… than in any other way. Thus, there is the financial information represented with real values and the expert knowledge that treats the same information in a qualitative sense. A way of combining both representations is needed to prepare the variables for a learning process.

Let's suppose that each one of these variables is qualitatively described via a different set of labels, which are intervals of the real line, with an odd number of landmarks given by the experts. This allows having a central point l_i in each set of landmarks. In order to prepare these variables for the neural network training, two steps will first be taken:

Step 1 Transformation of the central landmark l_i to 0, through a translation t_i: $\mathbb{R} \longrightarrow \mathbb{R}$, $t_i(x) = x - l_i$

Step 2 Transformation of the values through a linear transformation f: $\mathbb{R} \longrightarrow \mathbb{R}$,

$$f(x) = \begin{cases} +s \cdot t_i(x) / l_r & \text{if } t_i(x) \geq 0 \\ -s \cdot t_i(x) / l_l & \text{if } t_i(x) < 0 \end{cases}$$

where s is the sign of the linear transformation determined by the expert, and l_r and l_l denotes the right and left landmark, respectively.

After these two steps, all the values of the variables are described in a similar range, but the discretizations of the real line are different since the set of landmark is different for each variable (see Figure 2).

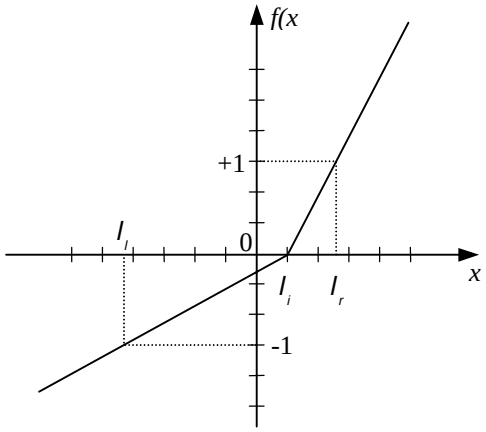

Figure 2: Real line discretization using expert landmarks.

A Credit Risk Prediction: Rating Evaluation

There are specialized rating agencies, the most important of which are Moody's and Standard & Poor's, that classify firms according to their level of risk. For example, Standard & Poor's gives the following labels to assign a rating to the firms: {AAA,AA,A,BBB,BB,B,CCC,CC,C}. 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.

The processes employed by these agencies are highly complex. Decision technologies involved are not based on purely numeric models. On the one hand, the information given by the financial data is used, and the different values included in the problem are also influential. On the other hand, they also analyze the industry and the country or countries where the firm operates, 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.

The classification process in a first approach has already been implemented. The work is currently in the initial process of empirical application: the construction of the financial database for the firms included in the index D.J.500.

Each firm is represented by a set of financial ratios that will be the input variables. The first one is the sector were the firm acts. As it can be seen in next table, the firms are considered in seven different sectors.

	Sector	# Firms
Cyclical consumer	1	71
Non-cyclical consumer	2	80
Technology	3	42
Utilities	4	38
Basic Materials	6	33
Industrial	7	58
Energy	8	31
	Total	**353**

Table 1: Sectors

Initially the quantitative variables being used were: Interest coverage (IC), Market value over debt (MV/DBT), Debt over net worth, (DBT/ATN), Cash flow over debt (CF/DBT), return on assets (ROA), internal financing percent (INTFIN), short term debt over long term debt (DC/DL), sales growth (SALES). The experts agree that some of these variables had a strongly dependence. For that reason and after a statistical study, they were reduced to the following five: V_1=IC, V_2=MV/DEBT, V_3=ROA, V_4=INTFIN, V_5=DC/DL. Each one of the variables has different landmarks, as it can be seen in table 3, according to expert's knowledge.

Experiments and Results

Initially, we started with a database that included a total of 353 patterns. There were 12 input qualitative variables, 1 qualitative variable (sector) for each pattern and 1 output. Since many instances had missing values, all those instances that had one or more missing values were deleted from the database. Following the experts' recommendations and due to the especial peculiarity of a sector of activity (technological sector), the technological companies were also deleted from the set of patterns. Next step was, according to the experts' knowledge, to select those variables that were the most relevant in computing credit risk. The input space was reduced from 12 to 5 variables, and from 495 to 244 instances. All 5 input variables are real-valued while the rating, i.e. the output variable, is a nominal variable with 6 different classes {AAA, AA, A, BBB, BB, B}, and has been represented using a 1-of-6 code (classes CCC, CC and C were not used because there were not firms available on the database for them). At this point, there were at least two options: to train a single RBF network with 5 inputs and 6 outputs, or to train 6 different RBF networks with 5 inputs each one but only 1 output. The former option is not too appropriate because of the low number of patterns available for training. The latter option is more efficient from the point of view of resource optimization. Although the final size of the architecture training will be probably smaller for the

single RBF network and the training faster, its generalization will be worse, and to get a good generalization is more important than the size or the training time. Thus, the experiments had been performed considering that the initial problem of classifying a pattern in 1-of-6 classes has been transformed in 6 different problems of classifying a pattern in a single class. The network will say whether the pattern is or is not of the class for which the network has been trained.

Simulations have been carried out following the PROBEN1 standard rules (Prechelt 1994). The data set available has been sorted by the company name before partitioning it into three subsets: training set, validation set and test set, with a relation of 50%, 25% and 25% respectively. Table 2 shows the pattern distribution in each data subset. Note that for class AAA there are no patterns available in the test subset, and for class B there are no patterns to perform the training or the validation.

Rating	AAA	AA	A	BBB	BB	B	Total
Training	5	18	53	41	5	0	122
Validation	2	10	28	18	3	0	61
Test	0	7	23	27	3	1	61
Total	7	35	104	86	11	1	244

Table 2: Pattern distribution over data subsets.

To study and analyze the effect that qualitative input conditioning had over RBF generalization, two different kinds of training have been done. First training (referred to as *blind* training) rescale all the input values to mean 0 and standard deviation 1, but do not take into account the experts' knowledge. Second training (*expert* training) performs the input transformation described previously, i.e. it consider the information on orders of magnitude to rescale the values (see Table 3).

	l_1	l_2	l_3	s
V_1	1	4	10	+1
V_2	1	2	8	+1
V_3	0.02	0.07	0.15	+1
V_4	0.2	1	10	–1
V_5	0.0	0.1	0.3	+1

Table 3: Expert landmarks and signs.

Initially, networks are trained on the training set while the validation set is used to adjust the radial function width (*r*). To perform this adjustment of the radial width, a total of 4000 simulations have been done for each class. The widths checked are the following:

- from 0,0001 to 0,1 with increments of 0,0001
- from 0,101 to 1,1 with increments of 0,001
- from 1,11 to 11,1 with increments of 0,01
- from 11,2 to 111,1 with increments of 0,1

The final width (see Table 4) is selected among the 4000 widths trained by applying the next criteria:

(a) Choose the width that maximizes the classification accuracy for the validation subset.
(b) Choose the width that, with criterion (a), produces the smallest network.
(c) Choose the width that, with criterion (b), minimizes the mean squared error for the validation subset.
(d) Choose the width that, with criterion (c), minimizes the mean squared error for the training subset
(e) Choose the width that, with criterion (d), maximizes the classification accuracy for the training subset.

	Blind training		*Expert* training	
	r	CA	*r*	CA
AAA	1,055	96,7%	51,2	98,4%
AA	11,2	82,0%	68,1	85,2%
A	0,831	73,7%	4,7	70,5%
BBB	80,6	63,9%	32,5	75,4%
BB	5,11	95,1%	19,9	95,1%
B	111,1	100,0%	111,1	100,0%

Table 4: Radial function width (*r*) and classification accuracy (CA) for validation subset.

Once the radial width is determined, networks are trained on training and validation sets while the test set is used to assess the generalization ability of the final solution. The same radial widths shown in Table 4 are used and the results are presented in Table 5.

	Blind training		*Expert* training	
	r	CA	*r*	CA
AAA	1,055	100,0%	51,2	100,0%
AA	11,2	88,5%	68,1	90,2%
A	0,831	50,8%	4,7	59,0%
BBB	80,6	57,4%	32,5	59,0%
BB	5,11	95,1%	19,9	95,1%
B	111,1	98,4%	111,1	98,4%

Table 5: Radial function width (*r*) and classification accuracy (CA) for test subset.

As can be seen in Table 5, classification accuracy for the *expert* training is better, or at least equal, than for the *blind* training. Since the only difference is the use of the expert landmarks during the input conditioning, it seems that the use of this kind of information during training can be useful. However, the initial problem was not to make six classifications independently, but just one. Moreover, each one of the six RBF networks trained say if a pattern is or is not of the class for which the network has been trained. Thus, the networks output can be: *Yes, it is* or *No, it is not*. Unfortunately, if we combine the classification we get from each one of the six networks, the answer is not necessary one of the six classes, but could be more than one or even none of them. This means that each input pattern could be correctly or incorrectly classified or even not classified. Table 6 collects this triple classification for the test set and, as can be seen, the *expert* training is again better than the *blind* training. The number of patterns

correctly classified is almost a 40% better for the *expert*. At the same time, the *expert* training has a lower indetermination in the classification (41,0% in front of the 47,5% of the *blind* training) and the same could be said for the number of patterns incorrectly classified (29,5% for 31,2%).

	Blind training		*Expert* training	
Correctly classified	13	21,3%	18	29,5%
Incorrectly classified	19	31,2%	18	29,5%
Not classified	29	47,5%	25	41,0%

Table 6: Final classification for the test data subset.

Conclusion and Future Work

This paper presents an on-going work, which provides strategies for synthesising qualitative information from variables, each one of which is qualitatively described in a different way. It has been proved that using the expert information we enhance the network generalisation.

The system is applied in the financial domain to evaluate and simulate credit risk. But this approach may also be applicable to problems in other areas where the involved variables are described in terms of orders of magnitude.

The limitations of the method presented cannot be evaluated until the implementation is completed and sufficiently tested. The proposed method is currently being implemented to be applied to available data referring to the most important American and European firms, whose Moody's rating is known.

Some of the future tasks consist in:

- Using the landmarks given by the experts to codify the input variables to use orders of magnitude labels.
- Using the experts landmarks to define a qualitative distance in order to build qualitative gaussian density functions.
- Discovering alternative methods for building a homogenised reference that takes advantage of expert's knowledge.
- It is also intended to compare the obtained results with the results furnished by other classifiers used in artificial intelligence.

Our final words are to note that this work is only the first experiment with a new and simple idea, though one we are convinced is promising: the idea of extracting qualitative information from experts to homogenise references.

Acknowledgments

We would like to thank Carmen Ansotegui and Xari Rovira from ESADE Business School for their valuable help in all the financial concepts.

References

Agell, N.; Ansotegui, C.; Prats F.; Rovira, X.; and Sánchez, M. 2000. Homogenising References in Orders of Magnitude Spaces: An Application to Credit Risk Prediction. In *Proceedings of the Fourteenth International Workshop on Qualitative Reasoning*. Morelia, Mexico.

Piera, N. 1995. Current Trends in Qualitative Reasoning and Applications. Technical Report, CIMNE–3, International Center for Numerical Methods in Engineering, Barcelona.

Travé-Massuyès, L.; Dague, Ph.; and Guerrin, F. 1997. *Le Raisonnement Qualitatif pour les Sciences de l'Ingenieur*. Ed. Hermès, Paris.

Goonatilake, S.; and Treleaven, Ph. 1996. *Intelligent Systems for Finance and Business*, John Wiley & Sons.

Poggio, T.; and Girosi, F. 1990. Networks for Approximation and Learning. *Proceedings of the IEEE* 78: 1481–1497.

Broomhead, D.S.; and Lowe, D. 1988. Multivariable Functional Interpolation and Adaptive Network. *Complex Systems* 2:321–355.

Chen, S.; Cowan, C.F.N.; and Grant, P.M. 1991. Orthogonal Least Squares Learning for Radial Basis Function Networks. *IEEE Transactions on Neural Networks* 2(2):302–309.

Prechelt, L. 1994. PROBEN1: Set of Neural Network Benchmark Problems and Benchmarking Rules. Technical Report, 21/94, Department of Informatics, University of Karlsruhe.