THE APPLICATION OF NEURAL NETWORK BASED METHODS TO THE EXTRACTION OF KNOWLEDGE FROM ACCOUNTING REPORTS

Duarte Treigueiros Robert Berry

Accountancy and Finance Sector School of Information Systems
University of East Anglia Norwich NR4 7TJ United Kingdom

Abstract

In this study we develop a new approach to the problem of extracting meaningful information from samples of accounting reports. We show that Neural Network-like algorithms are capable of implementing this approach. Such tools are able to automatically build optimal structures similar to financial ratios.

Some results are presented. They suggest that this approach effectively avoids the search of appropriate ratios by the analyst along with some other major drawbacks of the multivariate statistical modelling techniques used in accountancy. The organization of the Neural Network models also outlines internal features of accounting data, providing new insights into the relative importance of variables for modelling a particular relation.

The paper also argues that in the accounting and finance context a major problem of Neural Networks, that of understandability of the resulting parameters, is minimised. Much of the internal operation of the networks involves the construction of generalisations of the ratio concepts with which accountants are familiar.

1 Accounting Decision Models

Accounting reports are an important source of information for managers, investors and financial analysts. Statistical techniques have often been used to extract information from databases where accounting reports and related outcomes are gathered. The goal is to construct models suitable for prediction or for isolating the main features of the firm.

An early model is that of Beaver [5] who used ratios of accounting variables to predict financial distress. Many other researchers followed him, mainly using more sophisticated statistical techniques (see [1] or [26]). Other examples of accounting statistical models are the prediction of bond ratings [15], the relationship between market and accounting risk [4], and the structures of costs and output in various industries [13].

The procedures used to obtain these models are quite similar. The first stage consists of forming a set of ratios from selected items on an accounting report. This selection is typically made in accordance with the beliefs end expectations of researchers. Next, the normality of these ratio variables is discussed and transformations are applied. Finally some linear modelling technique is used to find optimal parameters in the least square sense. Linear Regressions and Fisher's Multiple Discriminant Analysis are the most popular algorithms. However Logistic Regression can also be found in some studies. Foster [10] offers a review of accounting modelling practice.

All such models use ratios as predictors. The use of ratios as input variables in accounting statistical models seems to be an extrapolation of their normal use in accountancy. Ratios are supposed to capture in a simple and standard way some interesting feature of the firm.

However, there are difficulties involved in using ratios. As M meaningful accounting variables can generate up to $M^2 - M$ ratios, some research seems to get lost in a prolific use of all sorts of combinations of variables. It is easy to find in the accounting literature models with forty and more predictors. Consequently, Factor Analysis is often called upon to cope with the mass of variables and their linear dependence.

Our goal is to show that the problem of choosing ratios can be avoided. Using procedures consistent with the generating process of accounting variables, an internal node of a Neural Network can build structures similar to ratios. Such structures are also self-explanatory thus improving our knowledge of the modelled relation.

The next section will describe a generating model for accounting variables. In section 3 we show that Neural Network optimization is consistent with such a generating model. In section 4 we compare the new method with the typical one using a difficult problem of classification with accounting data. The post-processing of outputs is discussed in section 4.2 and improvements in the interpretability of the resulting model in section 4.3.

2 Accounting Variables

2.1 Probability Distribution

Empirical observation suggests that many accounting variables are approximately log-

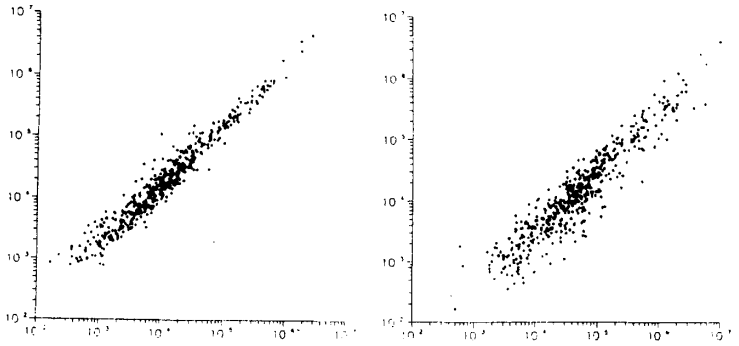

Figure 1: Typical accounting two-variate relations in log space

normally distributed. Log-normality has been observed in accounting variables which are sums of similar transactions with the same sign like Sales, Stocks, Creditors or Current Assets [19].

The statistical distribution of basic accounting variables received little attention in the literature. Ratios, however, have been object of a much bigger effort and, as a secondary product, some evidence has been gathered on the variables themselves. Horrigan [16] (1965) is an early work on this subject. He reports positive skewness on ratios, explaining it as a result of effective lower limits of zero for these variables. Other studies followed, [20] [6] considering skewness as an accident and implicitly suggesting the pruning or winsorizing of distributions. Deakin [8] (1976) reacted by showing that the positive skewness could not be ignored and Buijink [7] reported the persistency of this feature over a large period. Barnes [2] (1982) suggested that skewness on ratios could be the result of deviations from strict proportionality between the numerator and the denominator. Frecka [11] (1983) tried to achieve normality by pruning, proposing such procedure as the standard way of dealing with the problem of deviations from normality.

Following the literature on Ratio Analysis, accounting statistical models try to obtain improvements in normality by empirically pruning out tails and imposing transformations which are not always the most appropriate ones. The model parameters, after pruning, centering, scaling and rotating become difficult to interpret. The entire routine tends to a broad empiricism.

The gaussian distribution is often interpreted as the result of many independent elementary perturbations. This approximation entails the strong assumption of a constant effect. For example, the probability of getting odds, when tossing a fair coin, is a constant value of $1/2$ no matter the number of games or the size of the coin. And the probability of getting particular proportions of odds when tossing a coin in several sequences of games draw a gaussian distribution.

If, however, the perturbation dx of a variable x is proportional to the value of x itself, the effect is no longer constant. It is a proportionate effect [12] and a gaussian generating model will not be able to explain it.

Several important economic variables obey the proportionate effect generating model rather then the constant one. In general, if the perturbations suffered by a variable lead to an increase in the variable itself, as happens in growth or accumulative processes, the probability distribution of such variable will exhibit skewness and will not be well described by a normal curve.

The log-normal probability distribution can be viewed as a result of a proportionate effect generating model. If dx is proportional to x, the quotient

$$\frac{dx}{x}$$

will be independent of x. So, if we can find a function $z = f(x)$ such that

$$dz = \frac{dx}{x} \tag{1}$$

then the new variable z will obey the assumption of a constant effect, yielding a normal distribution. $f(x)$ is the logarithmic function as $d(\log x) = dx/x$.

2.2 Extending the Notion of Ratio

We considered the proportionate effect as a generating model for accounting variables. Now we shall study the joint variation of more then one accounting item and show that multivariate descriptors lead to definitions of accounting ratios which are a generalisation of the traditional one.

First, we notice that dx/x has the structure of an elementary growth rate. Thus, the perturbations of the logarithm of x, which obey to a constant-effect process, can be viewed as

a growth rate leading to x. In other words, an observed realization of x can be seen as the result of a growth process carried out along a time period. Several realizations of x spread their growth rates normally.

In the particular case of accounting variables, this gaussian growth rate is a sum of two components. A strong component which accounts for perturbations acting over the firm as a whole and therefore is the same for all the variables. And a weak component, particular to each variable.

A strong common effect must exist in order to explain differences in size amongst firms. Observing the contents of a database containing accounting reports of several firms we notice at once that each firm seems to be an isolated case. Accounting figures range from a few thousands of units in one case, to several hundreds of millions in another. Inside a firm, the values are much more similar. It would be strange not to admit the existence of a common effect for in that case the figures in an accounting report would have to be allowed to vary in an independent way, thus breaking accounting identities.

The variability remaining in each growth rate after the common effect has been explained is the weak effect. Inside firms the statistical behaviour of one particular variable is different from the statistical behaviour of others because the internal mechanisms commanding them are different.

Accepting a common effect over $1, \cdots, i, \cdots, M$ variables in the j^{th} firm we can write

$$\frac{dx_1}{x_1} = \frac{dx_2}{x_2} = \cdots = \frac{dx_M}{x_M} \qquad (2)$$

that is, the growth rate will have to be similar for all variables.

Let us analyze first the descriptor for two variables x and y. It is easy to see that 2 leads to the traditional notion of ratio. In fact

$$\frac{dy}{y} = \frac{dx}{x}$$

yields

$$\log(y) = \log(x) + C_j$$

It is clear from the normality of the common growth rate that, when we consider the $1, \cdots, j, \cdots, N$ firms whose reports are to be analysed statistically, the descriptor of the relation between y and x for firm j can be written in the form

$$\log(y_j) = \log(x_j) + \bar{C} + \epsilon_j$$

where $\bar{C}$ represents a central trend and ϵ_j is a residual. This model yields the known ratio form:

$$\frac{y}{x} = \exp \bar{C} \times \exp \epsilon_j = R \times f_j$$

Here, we arbitrarily use natural logarithms. If $\bar{C}$ is the mean of all the C_j, R will be the median of the log-normal distribution of the ratio y/x. f_j is a multiplicative deviation and accounts for the particular case of firm j. A good estimator for $\bar{C}$ is $\overline{\log y} - \overline{\log x}$, the difference between the means of the logarithms of y and x.

Therefore accounting ratios describe the common component of the variability of y and x when both x and y are supposed to be the final result of a growth process. Fieldsend et al. [9] provide empirical evidence on the existence of a strong common effect by showing that in logarithmic space any two-variate regression between accounting variables yields slopes with a constant value of 1.

The notion of ratio can be generalised. For example, we may want to consider two groups of variables instead of two simple variables. Given $y_1, \cdots, y_K$ and $x_1, \cdots, x_M$ the equality 2 leads to the relation

$$\frac{1}{K} \sum_{k=1}^{K} \log y_{jk} = \frac{1}{M} \sum_{m=1}^{M} \log x_{jm} + \bar{C} + \epsilon_j$$

or, in ratio form

$$\frac{\prod_{k=1}^{K} y_{jk}^{1/K}}{\prod_{m=1}^{M} x_{jm}^{1/M}} = \exp \bar{C} \times \exp \epsilon_j = R \times f_j$$

that is, we now have a ratio of geometric means of variables as a descriptor of the strong common effect.

It is important to notice that the above explanatory models have only one free parameter. The adjustment for a common source of variability is made by finding a unique optimal value in logarithmic space, in a similar way to the centering by subtraction of means. The inclusion of more then one variable in each group will not account for more explained variability because the number of free parameters remains equal to one. However, more variables, if conveniently chose, can enhance the accuracy of ratios by self-smoothing undesirable particularities.

In the same way, the strong common effect could be introduced in a statistical model just by using any of the accounting variables available. But it is desirable to build geometric means (in logarithmic space, averages) of several selected variables in order to self-smooth particular components.

If we wish to model the joint behaviour of $1, \cdots, i, \cdots, M$ variables after controlling for the strong effect we must be able to account for differences amongst them. The simplest way of doing this is by introducing one parameter, b_i, individualizing each growth rate.

The introduction of this new parameter allows us to describe, using the same formalism and without loss of generality, the two components of

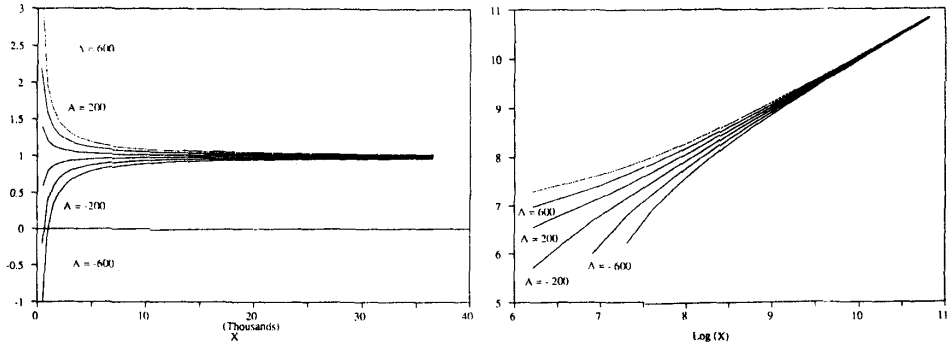

Figure 2: When $Y = A + X$ is transformed, the fact that $A \neq 0$ introduces non-linearity in the resulting relation. Such non-linearity affects only values of Y near A. Left, ratio transform. Right, log transform

the variability of each accounting item. A common effect would have $b_i = 1$ for all variables.

Similarly to 2, there is a growth rate for which

$$b_1 \times \frac{dx_1}{x_1} = b_2 \times \frac{dx_2}{x_2} = \cdots = b_M \times \frac{dx_M}{x_M} \quad (3)$$

holds. Here, b_i is a gain or attenuation factor. When considering several firms the overall growth rate will exhibit normality. The multivariate descriptor for $1, \cdots, i, \cdots, M$ variables has thus a general form like this

$$\sum_{i=1}^{M} w_i \times \log(x_i) = C \quad (4)$$

in which the residual is omitted. w_i is a parameter expressible in terms of b_i. We can write 4 as a linear relation

$$\sum_{i=1}^{M} w_i \times u_i = C$$

where $u_i = \log x_i$. In logarithmic space a simple inner product can account for the particular behaviour of accounting variables. In ratio form,

$$\prod_{i=1}^{M} x_i^{w_i} = \exp C$$

The above descriptors are an extension of the traditional notion of ratio. Plausible assumptions supporting the use of ratios also justify their use. Therefore, inner products can model accounting relations in logarithmic space. And they will be able to describe the particular behaviour of each variable as well as the common effect.

2.3 Sources of Non-Linearity

The relation $dx/x = dz$ can be a simplistic description of growth generating processes. We establish a more realistic basis by admitting that when such growth started, x was already endowed with a value x_0 different from zero. Therefore we should write, instead of formula 1

$$\frac{dx}{x_0 + x} = dz$$

Such processes lead to descriptors which are no longer linear nor easy to describe. Accepting as a rough approximation that the starting value, x_0, act as a parameter of the model, the multivariate descriptor becomes

$$\sum_{i=1}^{M} w_i \times \log(x_{0i} + x_i) = C$$

or, in ratio form,

$$\prod_{i=1}^{M} (x_{0i} + x_i)^{w_i} = \exp C$$

It is convenient to investigate the range of accounting variables and the foundation for the existence of significant x_0. A similar topic has been object of some attention in the accounting literature, but under a different approach. Lev and Sunder [18] and Whittington [27] are the usual references.

It is easy to gain insight into the way x_0 affects the linearity of a two-variate relation like $Y = A + X$ by applying transformations in a spreadsheet or similar tool. Figure 2 shows the results of applying logarithms (right) or ratios (left) in both sides of $Y = A + X$ for growing values of A. Clearly, non-linearity becomes significant only when the order of magnitude of the variables is similar to the order of magnitude of A. Accordingly, non-linearity must be taken into account only when the final realization of a growth process, x, is not far away from its beginning, x_0.

It is possible to foresee that this effect will be present in some accounting variables and

also that it is not an important feature. The examination of two-variate scatter-plots of accounting variables in logarithmic space would detect departures from linearity should they exist. Accounting variables draw very homogeneous highly correlated linear scatters, always with a slope of 45° corresponding to the strong effect (see fig. 1). There are traces of convexity affecting only small values. This convexity is consistent with figure 2 for positive constant x_0 but can not be considered as a clear feature. Therefore we conclude that it seems possible to model the x_0 effect, when present, as a convexity affecting small firms.

In accountancy, apart from this starting value or base-line effect, non-linearity can occur due to two other factors.

- Real non-linear relations between variables. This is not common. However Whittington [27] reports quadratic relations in variables connected with profitability and gives a good explanation for this phenomena. Basically, such occurrence would be due to saturation effects.

- Higher order interactions between groups present in the sample. It occurs, for example, when modelling financial risk. Leveraged and non-leveraged groups of firms behave in opposite ways if they belong to some specific industries. A statistical version of the "exclusive-OR" problem can arise. Linear Discriminant techniques would not be able to separate such risk groups.

Existing accounting statistical models seem not to be aware of potential sources of non-linearity. Base-line and saturation effects are not very imposing and the second source of non-linearity can be avoided by increasing the dimension of the input space, which accounting models implicitly do.

3 Modelling Accounting Variables with Neural Networks

3.1 Introduction

"Neural Network" is the the name of several modelling heuristics, having in common a topology inspired in the way neurons are organized in the brain and the use of non-analytical algorithms.

If a sample containing input and related outcome variables represent an unknown relation, a Neural Network will model this relation by successive approximations using interactive algorithms. Such process is known as learning or training of the net.

Topologically, Neural Networks are lattice structures of simple computational elements called "nodes" connected in a specific way. The connections between nodes (known as weights) can be strengthen or weakened during the training process, by means of iterative heuristics, causing the net, as a whole, to become a model of the data.

A typical node will sum M weighted inputs and will pass the result as an input for several other nodes. Variables labeled $x_1, \cdots, x_M$ are applied through a set of associated weights $w_1, \cdots, w_M$ to the node. The weighted sum of the inputs $y = \sum_{i=1}^{M} w_i \times x_i$ is the output of the node.

Several nodes form a layer. There, all M inputs are fully connected to the N nodes, that is, a weight w_{ij} exists, linking each input x_i, $i = 1, M$ with each node j, $j = 1, N$. In this layer outputs of nodes are $y_j = \sum_{i=1}^{M} w_{ij} \times x_i$.

The Multilayer Perceptron is a class of Neural Network directed to classification tasks. Topologically it is feed-forward: Nodes are arranged in layers but each node's output is connected only to next layer's inputs. No intra-layer connections exist, nor any feedback paths from an output to earlier layers. In this kind of Neural Network the output produced by each node is not a simple weighted sum of inputs. After the usual summation $s_j = \sum_{i=1}^{M} w_{ij} \times x_i$ the result s_j is submitted to some continuous differentiable non-linearity $F(s_j)$ known as the transfer function, before being sent to the next layer.

The learning algorithm, known as the Generalized Delta Rule, [22] is an enhanced version of the stochastic gradient-descent optimization procedure able to propagate deviations backwards through more then one layer of nodes. The Multilayer Perceptron creates an internal model of the relation input-outcome by adjustment of its weights with steady improvements in the direction which minimizes the deviations observed between the produced output and the desired outcome.

Minimum Least Squares deviation, as a success criterion, is just one of the possible criteria. Likelihood maximization is also used. In this case, the model learns to maximize the probability of having obtained the set of input-output pairs which were actually observed in the training set.

If the number of nodes (and therefore, connections) is big compared with the number of classifying features of the data, the Multilayer Perceptron behaves just like a storage device. No generalization can be expected. An opposite situation, very few nodes, would make the Perceptron recognize only main features. Using hidden layers with variable number of nodes it is possible to control the number of free parameters used in classification.

It is the back-propagation of deviations towards more then one layer of nodes which makes the Multilayer Perceptron potentially attractive

as a statistical modelling tool. The outputs of intermediate nodes, considered as new variables, can eventually bear interesting information about the modelling process or the features of the relation. Such new variables, known as internal representations, along with the net topology, can make the modelling process self-explanatory and therefore attractive as a form of knowledge acquisition.

3.2 Node Outputs as Extended Ratios

As seen in section 2.2 a descriptor able to account for both common and particular components of accounting variability is the one who approach an outcome using extended ratios of the form

$$R = \prod_{j=1}^{M} x_j^{w_j}$$

as input. In logarithmic space

$$\log R = \sum_{j=1}^{M} w_j \times \log(x_j) \qquad (5)$$

Notice that this expression, an inner product, is functionally the same as a Neural Network node's output.

Our approach consists on letting w_j be the adjustable connections or weights of a Multilayer Perceptron whose inputs are the logarithms of accounting variables. Doing so, we are creating in each node of the first hidden layer an internal representation with the form of an extended ratio. Such extended ratios will be the inputs for the next layer, where they are linearly combined in order to model the relation.

In this way, the first hidden layer of the net is dedicated to the building of appropriate ratios. Next layers will use such ratios to approach the targets. If the model is optimal in some sense, it seems reasonable to expect that the discovered ratios will also represent an optimal choice of combinations of variables.

The problem of forming ratios given a set of accounting variables considered as preeminent can thus be avoided. The best ratios to be used are not imposed by the analyst. Instead, they are discovered by the modelling algorithm. We show in section 4.3 that by using an appropriate training scheme these extended ratios often assume a simple and interpretable form.

The logistic function

$$f(x) = \frac{1}{1 + \exp(-x + w_0)} \qquad (6)$$

which is standard in Multilayer Perceptrons as a transfer function, will bring back the extended ratios from logarithmic space and will also provide a controlled amount of non-linearity for the lower values of R.

$$f(R) = \frac{1}{1 + \exp(-\log R - w_0)} = \frac{R}{R + \exp(-w_0)}$$

w_0 is another free parameter, a bias, adjustable during the training process. Large negative values of w_0 yield a linear relation between R and the output of the node. Smaller values introduce a concavity affecting small R.

In order to model difficult non-linear relations one additional hidden layer of nodes is required. Depending on the number of nodes allowed, this layer will be able to apportion as much piecewise non-linearity as necessary and, hopefully, not more then the necessary. The logistic transfer function seems able to tightly control the amount of non-linearity required to model a relation. These simple, continuously increasing functions, with constant values out of a very limited range, are ideal for interpolation.

3.3 Applicability of the Log Transform

Although many accounting variables are well suited for a logarithmic transformation, some important cases deserve a further discussion. As an example, the long term debt is frequently zero and earnings can be negative.

In order to transform accounting variables having negative values, we suggest the rule

$$\begin{aligned} x &\mapsto \quad \log(x), \quad x > 0 \\ x &\mapsto -\log(-x), \quad x < 0 \end{aligned}$$

which corresponds to the assumption that negative cases are affected by the proportionate effect in a negative direction. Losses should be related with size too.

We also avoid the problem of variables having zero values, like Debt, simply by using, instead of $\log 0$, a very small number: $\log 1 = 0$.

Such criteria will work if the unity measures are not far away from the normal values. For instance, if a unity of millions of pounds is required instead of the usual unity of thousands, it is very likely to have in the sample cases with values near zero, both positive and negative. In order for this approximation to work properly over all situations, it is desirable that the sample contains no values little then several units in absolute value.

In our case this has never been a problem, also because the variability of some negative valued accounting variables can be introduced in a statistical relation by using other variables appropriate to logarithmic transformations. The variability of Earnings can be brought to the model by the introduction of Sales and Expenses.

The base of logarithms may be selected in a way that avoids further scaling. If we adopt natural logs the transformed values will range from about 2 to 18 approximately. Base 10 logs are specially attractive because they can be directly interpreted as powers of 10. Accounting figures, after transformed, range between 3 and 7, and a

Group	Name	N. Cases	Pr.
1	Building Mat.	31	6.2%
2	Metallurgy	19	3.8%
3	Paper, Pack	46	9.2%
4	Chemicals	45	9.0%
5	Electrical	34	6.8%
6	Industrial Pl.	17	3.4%
7	Machine Tools	21	4.2%
8	Electronics	79	15.7%
9	Motor Comp.	23	4.6%
10	Clothing	42	8.4%
11	Wool	19	3.8%
12	Misc. Text.	30	6.0%
13	Leather	16	3.2%
14	Food	80	15.9%

Table 1: Industry groups and number of cases in the one-year (1984) training set

simple translation is enough to bring them to a range acceptable for Neural Network training.

4 An Application: Modelling Industry Homogeneity

4.1 Description of the problem

In this section we apply our framework to a known accounting statistical problem, the test of the homogeneity of a particular industry grouping. We compare our procedure with the traditional one and we extract some conclusions. Sections 4.2 and 4.3 explain departures from usual techniques associated with the Multilayer Perceptron.

All companies quoted on the London Stock Exchange are classified into different industry groups according to the Stock Exchange Industrial Classification (SEIC) which aims to group together companies whose results are likely to be affected by the same economic, political and trade influences [21]. Although the declared criteria are ambitious, the practice seems to be more trivial, consisting of classifying firms mainly on a end-product basis. The SEIC classifies firms according to a perception of groupings of firms. Extended Ratios are used here to model such groupings, should they exist, using accounting variables.

The data for the training sets were drawn from the Micro-EXSTAT database of company financial information provided by EXTEL Statistical Services Ltd, which covers the top 70% of UK industrial companies. We selected 14 manufacturing groups according to the SEIC criteria. After discarding some firms (see below) we got accounting information on 297 firms covering a six year period (1982 - 1987) and a bigger sample (500 cases) for only one year (1984).

The input variables received two different types of processing. The first, usual in finance research, consisted of

> forming 18 financial ratios chosen as to reflect a broad range of important characteristics relating to the economic, financial and trade structure of industries (...) [25]

and extracting from them the eight bigger principal components. These new variables were then used as inputs for a Multiple Discriminant Analysis. A description of these ratios and the modelling procedure can be found in [25].

The new approach consisted on using eight accounting variables directly, not in the form of ratios. The selected items were Fixed Assets (FA), Inventory (I), Debtors (D), Creditors (C), Long Term Debt (DB), Net Worth (NW), Wages (W) and Sales less Operating Profit less Wages (EX). All this variables were present in the original 18 ratios, along with others like Earnings, Value Added, Total Capital Employed, Total Assets, Operating Profit, which we did not use in the new approach.

Logarithmic transformations were applied according to the procedure outlined below. Then, a Multilayer Perceptron were used to capture the observed relation between such input variables and industry grouping. Our aim was to compare both approaches in order to evaluate the ability of Multilayer Perceptrons to form interpretable hidden representations similar to ratios.

A major methodological difference between our approach and the usual one is that in general univariate normality criteria is used to prune the original sample of ratios down to an acceptable number of standard deviations. We followed a case-wise method for discarding outliers not based on normality considerations. Only cases known as distressed firms, non-manufacturing representatives of foreign companies, merged or highly diversified firms were excluded.

Results concerning two sets of data are reported. The first (1984) represents a cross-sectional view. The second (SIX YEARS) checks the regularity of firm grouping.

Table 1 displays the proportions of cases in the 1984 set. Notice how groups are dissimilar in size, the smallest one having 16 firms and the biggest 80. These proportions entail no prior knowledge of any classification.

4.2 Post - Processing of Outputs

Discrimination, when overlapping distributions are present, implies a probabilistic interpretation of outputs.

In Accountancy, Bayesian considerations are in general independent of the proportions observed in the sample. Neural Network application to other sciences can be misleading. There, proportions observed in the sample are generally taken as acceptable prior probabilities.

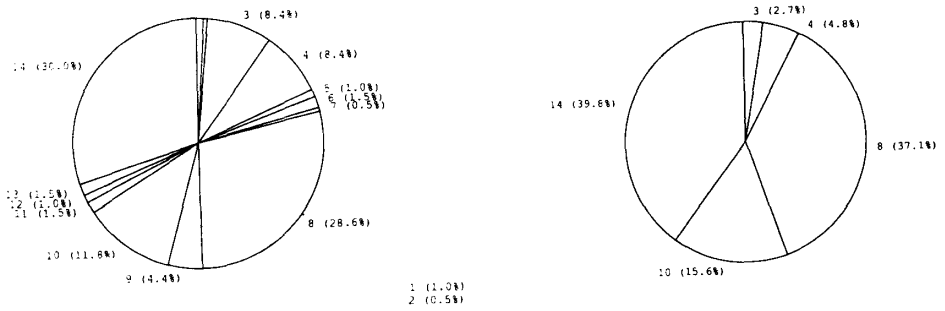

Figure 3: On the left, classification results after post-processing (1984 sample and prior probabilities proportional to the size of the group). On the right, the same with direct interpretation.

Following suggestions like those of Baum and Wilczek [3] several authors advocate a direct interpretation of outputs as probabilities [14] [24] and show how the usual squared error criterion can be corrected to achieve likelihood maximization. In such case, the weights are corrected in the gradient direction of the log-likelihood rather then on the gradient of the squared error.

We found that node outputs when interpreted directly as probabilities produce a clear reduction in accuracy. The final result is a severe loss of ability to distinguish small groups.

Thus, we decided to interpret outputs of a Multilayer Perceptron as a multi-dimensional measure of distance to targets. If departures from normality are not severe, this interpretation can be carried out by using conventional statistics like Chi-Square, Penrose or Mahalanobis distances. Such measures can be regarded as scores and conditional probabilities can be deduced from them, allowing further Bayesian corrections, independent of proportions observed in the sample. Of course, a Bayesian correction could be done directly over the outputs interpreted as probabilities. However, due to the observed lack of accuracy, a direct correction would lead to a very bold classification.

Using a Multilayer Perceptron and the data described in section 4, and implementing training schemes as described by Hopfield [14] and Solla et alt. [24] we tested the direct interpretation of node outputs as probabilities, and also the usual correction of node outputs based on the way linear discriminant analysis, for example, corrects scores. Results are reported in figure 3 when prior probabilities are taken as equal to the size of the group. On the left we can see the result of using post-processing. On the right, the corresponding result derived by directly interpreting node outputs as probabilities.

The post-processing gives detailed classifications. Direct interpretation ignores 9 of the 14 groups, the small ones, but finally achieves a better global performance by classifying the remaining 5 groups, which are the bigger ones, very well. Therefore, although for the sake of efficiency of convergence we adopted the likelihood cost function, node outputs were post-processed as distances.

A short description of this post-processing follows. For a training set with N cases, consider o_{ik}, the output produced in node $m, 1 \leq m \leq M$ by case $i, 1 \leq i \leq N$. Compute K square deviations, d_{kim}, between that node's output and all possible targets for that node: $d_{kim} = (t_{km} - o_{im})^2$, with $k, 1 \leq k \leq K$. The mean sum of squares in node m for the whole sample will be: $\sigma^2_{km} = \sum_{ik} d_{ikm}/(N-1)$ and the standardized distances between a node's output and all possible targets can now be added over all nodes:

$$D_{ki} = \sum_{m=1}^{M} \frac{d_{kim}}{\sigma^2_{km}}$$

The minimum of these distances would identify the classification if no Bayesian corrections were needed, that is, if the assumption of equal prior probabilities is acceptable. This distance has been compared with a more elaborated measure, the Mahalanobis distance, and it was found that the latter would not achieve a more accurate performance. In order to introduce Bayesian considerations, D_{ki} ought to be computed as a Chi-Square distance to targets. The significance of this distance is the desired conditional probability.

4.3 Generalization and Interpretability of Results

Two major goals of this research were the evaluation of the generalization capacity and the interpretability of Multilayer Perceptron models.

In order to obtain an estimate of the generalization capacity associated with every model, the original samples were divided randomly into two sub-samples of approximately equal size. All models were constructed twice, first with one half of the sample and a check carried out with the

	Node Number				
Variable	2	3	4	5	6
DB			-6		
NW	8				
W	1			-6	
I	8				
D	2				-2
C				3	
FA	-9	-4		6	-4
EX	-10	4	8	-2	3

Table 2: Approximate values of weights connecting input variables with nodes in the first hidden layer after training with random penalization

other half, and again reversing the roles of the two half data sets. Results were considered conclusive if both models, when validated with the half-sample not used to build them, produced consistent results.

All classification results reported here concern the test set, not the training set. That is, they were obtained by measuring the rate of correct classification the model would produce when evaluated by the half-set not used to train it. The classification performance on the set used for the training depends solely on the number of free parameters and can be increased simply by introducing more nodes on the net. Therefore such results are uninteresting and are not presented here.

The normal approach to test a model, by deleting a single observation and predicting its value with the model estimated on the rest of the data set, and repeating this procedure N times, is infeasible. This is because the training of a Neural Network is time consuming. The procedure adopted will however, with a large enough data set, produce unbiased estimates.

We found the generalization capacity very much dependent on the topology of the net. The number of nodes in a hidden layer seems to determine, not only the dimension of the relation, but also the ability of these tools to properly generalize. Persistently, we obtained good generalizations whenever a hidden layer would have six nodes. Both the 1984 and the SIX YEARS data set exhibit such feature. Figure 4 shows some classification results for different number of nodes in the first hidden layer when using the SIX YEARS data. Similar patterns, though not so contrasted, were observed when using the 1984 set.

Other major goal of this study was to evaluate the power of Neural Networks as knowledge acquisition devices. Multilayer Perceptrons are often considered as not ideal in applications where self-explanatory power is required. However, in the case of accounting variables, it seems possible to interpret the way the relation has been modelled by looking at the weights connecting input variables with the first hidden layer's nodes. These weights are the exponents of the extended ratios which provide an optimal relation.

In order to enhance interpretability we introduced during training a random penalization of weights with inhibitory values. In a Neural Network each node acts as a modelling unit with a certain amount of free parameters. The same output can be obtained with very different combinations of such parameters. Inhibitory weights connecting inputs with the first hidden layer appear when the node tries to weaken the contribution of a variable. Therefore, if we randomly introduce small penalizations of inhibitory weights along the training, as the correction of weights is proportional to the input variables, the inhibitory weights will tend to remain inhibitory. In the same way, the not-inhibitory weights will tend to have their values strengthen.

The final result is a contrasted set of weights. If the relation to be modelled is consistent with such a contrast, then there is no reason to expect that the described manipulation will damage the performance of the model.

The procedure we follow to achieve interpretability involves the following steps

- let one of the nodes in the first hidden layer model the strong common effect and introduce it in subsequent layers. Input variables not convenient for the modelling of this effect (Debt is an example) have weights connecting to this node set to zero. The others have fixed and equal weights.
- During training, and whenever a new presentation of the entire training set is to begin, one of the remaining nodes of the first layer is randomly selected. Their weights are examined and those with inhibitory weights are penalized by a small factor, typically 0.98.
- Before the end of training, all the weights connecting inputs to the first layer and exhibiting strong inhibitory values are set to zero and fixed.

Such procedure is applied only after the discovery of the topology yielding the best results.

Just by dedicating one node of the first hidden layer to the modelling of the strong effect we notice an improvement in speed of convergence and in the final generalization capacity. Adding the random penalization of inhibitory weights, both speed and generalization receive a further, significant, improvement. However, when the topology of the net is not the best, this procedure can in some cases worsen the generalization.

The interesting result is that, when training finishes, the number of variables to consider in

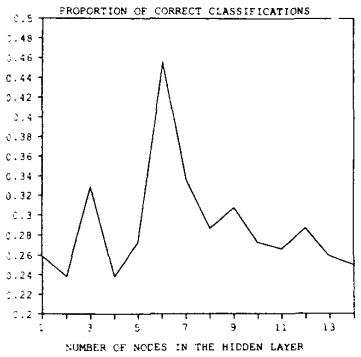

Figure 4: Classification results on the test set versus number of nodes in the hidden layer for SIX YEARS data: Six nodes provide the best generalization

each node is very small and characteristic. Looking at the non-zero weights it is possible to understand, in accounting terms, what the extended ratios formed in each node are doing.

Table 2 shows the extended ratios formed in a net with 8 inputs, 6 nodes in one hidden layer and 14 output nodes, trained with 1984 data. The emerging organization reproduces the way an expert in ratio analysis chooses variables. It is usual to build several ratios around one or two variables judged as important to capture a relation. As an example, efficiency is modelled around capital turnover, stock turnover and so on. Profitability is built around profit margin, return on equity, etc.

Experts put together several points of view around a few significant variables. And extended ratios seem to be trying the same sort of procedure.

Finally, apart from these non-standard features, our Multilayer Perceptron also implemented two enhancements described in the literature:

- a learning rate particular to each weight [17] in a version slightly modified by Silva and Almeida [23].
- likelihood maximization instead of squared deviations minimization, as mentioned.

4.4 Discussion of Results

Hitherto expectations about Neural Networks are related with the modelling of difficult relations (pattern recognition) or the mimicking of brain functions. However, some specific statistical problems requiring self-explanatory power can take advantage from the existence of meaningful internal representations. Numerical, continuous-valued observations such those found in stock returns, or data organized in accounting reports, can not be efficiently used by actual expert systems as a source of knowledge. We showed that Neural Networks can provide an intelligent and self-explanatory tool, along with improvements in performance.

	1984		Six Y.	
INPUT	MDA	MLP	MDA	MLP
18 ratios	29%		30%	
8 variables		38%		45%

Table 3: The best classification results of MLP (Multilayer Perceptron) compared with MDA (Multiple Discriminant Analysis)

Table 3 shows the best generalization results achieved with the traditional methodology (ratios) and also with Neural Networks. As can be seen Neural Networks achieved a better performance, with half the number of input variables and within a much simpler framework. Namely, the need for forming appropriate ratios was avoided as well as the blind pruning, and the extraction of a somehow arbitrary number of factors. Several accounting variables used to form the 18 original ratios were not present in our 8 variable set.

Hidden units were able to form more appropriate ratios then the original ones. The examination of such ratios shed light into the importance of input variables to model the relation.

Finally, Back-propagation also shows a useful ability to take advantage of the topology of the net to improve generalization. Even with a large number of free parameters, if the number of nodes in a hidden layer is in resonance with some internal feature of the data, high generalization can arise.

References

[1] E. Altman, R. Haldeman, and P. Narayanan. Zeta analysis: a new model to identify bankrupcy risk of corporations. *Journal of Banking and Finance*, 29–54, June 1977.

[2] P. Barnes. Methodological implications of non-normally distributed financial ratios. *Journal of Business, Finance and Accounting*, 51–62, Spring 1982.

[3] E. Baum and F. Wilkzek. Supervised learning of probability distributions by neural networks. pages 52–61, IEEE Conference on Neural Information Processing Systems - Natural and Synthetic, 1987. Denver.

[4] H. Beaver, W. Kettler and M. Sholes. The association between market-determined and accounting-determined risk measures. *The Accounting Review*, 654–682, October 1970.

[5] W. Beaver. Financial ratios as predictors of failure. *Journal of Accounting Research*, 71–111, 1966. Suplement.

[6] R. Bird and A. McHugh. Financial ratios - an empirical study. *Journal of Business Finance and Accounting*, 29–45, Spring 1977.

[7] W. Buijink and M. Jegers. *Cross-Sectional Distributional Properties of Financial ratios in Belgian Manufacturing Industries: Some Empirical Evidence.* Technical Report, University of Antwerp, Belgium, 1984.

[8] E. Deakin. Distributions of financial accounting ratios: some empirical evidence. *The Accounting Review*, 90–96, January 1977.

[9] S. Fieldsend, N. Longford, and S. McLeay. Industry effects and the proportionality assumption in ratio analysis: a variance component analysis. *Journal of Business Finance and Accounting*, 14(4):497–517, Winter 1987.

[10] G. Foster. *Financial Statement Analysis.* Prentice-Hall, 1986.

[11] T. Frecka and W. Hopwood. The effect of outliers on the cross-sectional distributional properties of financial ratios. *The Accounting Review*, 115–128, January 1983.

[12] R. Gibrat. *Les Inegalites Economiques.* Librairie du Recueil Sirey, 1931.

[13] Z. Griliches. Cost allocation in railroad regulation. *Bell Journal of economics and Management Science*, 26–41, Spring 1972.

[14] J. Hopfield. Learning algorithms and probability distributions in feed-forward and feed-back networks. In *Proceedings of the National Academy of Science USA*, pages 8429–8433, USA Academy of Science, 1987. Volume 84.

[15] J. Horrigan. The determination of long-term credit standing with financial ratios. *Journal of Accounting Research, Suplement. Empirical Research in Accounting: Select Studies*, 44–68, 1966.

[16] J. Horrigan. Some empirical bases of financial ratio analysis. *The Accounting Review*, 558–568, July 1965.

[17] R. Jacobs. Increased rates of convergence through learning rate adaptation. *Neural Networks*, 1:295–307, 1988.

[18] B. Lev and S. Sunder. Methodological issues in the use of financial ratios. *Journal of Accounting and Economics*, 187–210, December 1979.

[19] S. Mcleay. The ratio of means, the mean of ratios and other benchmarks. *Finance, Journal of the French Finance Society*, 7(1):75–93, 1986.

[20] M. O'Connor. On the usefulness of financial ratios to the investor in common stock. *The Accounting review*, 339–352, April 1973.

[21] J. Plymen. Classification of stock exchange securities by industry. *Journal of the Institute of Actuaries*, 97(406), 1971.

[22] D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing*, MIT Press, 1986.

[23] F. Silva and L. Almeida. Speeding up backpropagation. 1990. INESC, R. Alves Redol, Lisbon, Portugal.

[24] S. Solla, E. Levin, and M. Fleisher. Accelerated learning in layered neural networks. *Complex Systems*, 2:625–640, 1988.

[25] P. Sudarsanam and R. Taffler. Industrial classification in u.k. capital markets: a test of economic homogeneity. 1984. University of Leeds.

[26] R. Taffler. Forecasting company failure in the u.k. using discriminant analysis and financial ratios data. *Journal of the Royal Statistical Society*, 145:342–358, 1982.

[27] G. Whittington. Some basic properties of accounting ratios. *Journal of Business, Finance and Accounting*, 7(2):219–232, 1980.