How many good fishes are there in our Net ?
Neural Networks as a Data Analysis tool in CDE-Mondadori's Data Warehouse.

Guido Cuzzocrea, Bocconi University of Milan, Italy
Alberto Saccardi, Bocconi University of Milan, Italy
Giovanni Lux, CDE-Gruppo Mondadori, Milan, Italy
Emilia Porta, CDE-Gruppo Mondadori, Milan, Italy
Arianna Benatti, CDE-Gruppo Mondadori, Milan, Italy

Abstract:

In Data Analysis, Neural Networks (NN) are not a universal problem-solver, nor a completely user-friendly tool that anyone might consider to obtain the best and quickest answer to the most complex question. In spite of that, it would be wrong to ignore the possible advantages of NN in analysing real-world databases, when prior hypotheses are poor and linear modelling inadequate.

In classification problems such as discriminating between heavy buyers and non buyers in Direct Marketing, NN seem to be a powerful analysis tool. Are data manipulation and standard statistics therefore useless? Our experience is that only an integrated approach produces the best results: knowledge discovery in large databases means goal identification, sample design, a reduction in data dimensions, variable selection, model building (e.g., NN architecture), simulation and probabilistic error measurement.

The paper is a step by step description of Neural Networks and other Data Analysis tools used to classify CDE-Mondadori's Customers and Prospects: we will show a concrete and successful example of Data Mining with SAS System in a Data Warehouse.

Introduction

The CDE- MONDADORI Group has been selling its publications through mail-order catalogues for years. Currently it is present on the market with three 'book clubs': the `Club degli Editori', created over thirty years ago, that addresses a vast and poorly characterised target; 'Il Circolo', the most prestigious, that promotes a very highly qualified style of literature; the 'Club per Voi', that addresses a smaller group of readers that consider price the most important factor. Recruitment of members is made possible through their acceptance of an offer that, to be capable of capturing the attention of people that generally avoid mail-order purchases, has to be very attractive. This kind of acceptance doesn't allow one to distinguish between a member that is only interested in the initial promotion and one that is really interested in reading. It is probable that only the latter will continue to order from the catalogues he is sent monthly.

The objective of the analysis is to predict the client's behaviour in the shortest possible time and, consequently, evaluate the prospective profitability. Correct indications of this type would allow us, not only to invest upon the desired group of members, but to apply differentiated marketing strategies. In synthesis the study intends to increase the company's profitability .

The introduction of The SAS System into CDE, in 1994, represented quite a turn-around, giving rise to the most advanced statistical techniques. In 1995 neural networks were first considered in an attempt to predict the profitability of the members of `Il Circolo'.

1. How the book clubs work

CDE sends members 13 34-page catalogues per year containing about 100 titles among which 10 new entries and the book of the month, an outstanding work selected by publishing experts. All the books present in the catalogue are briefly introduced and commented to make it easier to choose, they are also priced at least 20% less than their normal book shop price.

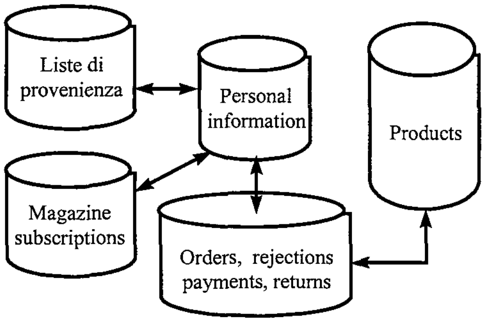
The only obligation imposed on the member is to buy at least three books per year so that his membership isn't suspended; nevertheless, he can cancel his membership at any time.

2. The data base

There are various types of information available to carry out the analysis. It is necessary to distinguish between the variables related simply to the anagraphic information of an individual and those recorded at the moment he joins the club.

The first group contains socio-demographic variables such as: geographic area, post code and sex; variables concerning the origin of the anagraphic information, for instance belonging to several types of lists as well as information relative to subscriptions to Mondadori periodicals. In the second group, the recruitment method (direct mail, press,..); commercial variables (orders, rejections, payments, returns,..); types of reading and finally elasticity to promotions.

Fig. 1 The informartion base



2.1 The sample

The problem of segmentation and classification of clients imposes the analysis of homogeneous cohorts of members: all members in the sample have to have had the same opportunities to purchase, i.e. they have to have received the same number of catalogues.

Only members that joined the club between 1/1/94 and 30/6/94 and had received 8 catalogues by 31/3/95 were considered: In this way as at 31/8/95 (the starting point of the analysis) no payments were pending.

From this population a random sample was extracted at a rate of 1 / 10.

2. 2 The dependent variable

The profitability of each member is evaluated considering a cost/gain balance determined by payments vs. mailing costs and returns/non payments.

The dependent variable is the profitability cumulated over a period of 8 catalogues.

Even though profitability is a continuous variable, the aim of the analysis is to predict in which of four profitability classes a member might fall. Therefore, depending on the neural network architecture, the dependent variable has been introduced either as continuous or categorical.

The four profitability classes have been constructed on the basis of economic considerations set out by the CDE Mondadori's marketing management.

2.3 The input variables

To be able to use neural networks within the SAS system not only does the information base need to be organised in such a way as to have one statistic unit (member) per record but each categorical variable needs to be transformed into a series of dummy variables.

3. The development of the project

3.1 Why neural networks

Apart from the curiosity of experimenting with a relatively new tool which is well known for its potential in data analysis, the following are basically the motives that guided our decision:

1. the necessity to solve a problem of forecasting and classification having at our disposal a consistent an complex information base;
2. the availability of software in the SAS environment.

The complexity of the classification problem stems from the numerous variables and their ambiguous relationships. This type of complexity and lack of prior hypothesis, brings us to consider traditional statistical modelling insufficient.

As far as software availability is concerned, the %TNN family of macros allows one to train MLP (multilayer perceptrons) neural network architectures by specifying all the characteristics or just accepting default parameters. The launch of the macros can be as simple as the declaration of input and target variables. In this case the system trains a network with no hidden layer and activation, link and loss function determined by the nature of the variables.

Alternatively it is possible to declare the number of neurons in as many hidden layers as necessary, the activation function of each neuron, link functions between layers and the loss function upon which the optimisation of the iterative algorithm is based.

3.2 Training and testing

Neural networks may be considered to be iterative algorithms capable of learning to recognise empirical laws on the basis of a vast number of relevant examples. With that in mind it is opportune to separate data into two sub-samples: the first will be the training set and the second the test set upon which the predictive ability of a network is gauged.

The statistic units belonging to the training set are the examples from which the network should learn. The network will attempt to fit the data iteractively, reproducing the target variable's response to the behaviour of the input variables. Goodness of fit is measured by a loss function which confronts observed value of the target variable and the predicted value generated by the network.

After the training phase the network has to be validated over the test set. It is extremely important to verify whether or not the knowledge gained by the network can be satisfactorily generalised. In other words, it is necessary to verify how well the network performs in the prediction of unobserved cases.

The homogeneous sample is randomly divided into two sub-samples using the SAS/BASE RANUNI (Random Uniform) function.

3.3 An *almost* automatic approach

It is often affirmed that Neural Networks are a powerful mathematical tool, that anyone, regardless of their specific skills, can use to find the best and most immediate solution to the most complex problem. Because of that, after the *preliminary* steps of organising the information base, separating the training and test subsamples, defining the target variable, including the whole set of possible input variables, we have considered a nearly automatic approach.

We then tried to train some different network architectures considering all available input variables.

The results obtained by the *almost* automatic approach were quite disappointing, confirming our initial supposition. In fact, the input variables are extremely numerous (over 100) and they often have a very similar meaning, with a very low contribution to the explanation of the target variable's dynamics.

In our application it seems evident that the large number of weights and redundant parameters to be estimated determine a lack of robustness of the network: fit coefficients measured for the training set might be considered satisfactory, but error level in the test set is quite relevant.

The results are even worse as the complexity of the network architectures increases: entering more hidden layers, or more neurons in the same hidden layer, we obtain a better fit for the training set and a lower capability to generalise (higher error level in the test set).

The *almost* automatic approach is even more unsatisfactory if we keep in mind how important the *preliminary* steps are: data analysts know that goal identification, sample design, data manipulation and cleaning are often, if not always, the greatest part of the job.

3.4 Fit and generalisation

The relationship that exists between goodness of fit and the capability to generalise of neural networks corresponds to the concept of robustness in traditional statistical modelling: under equal conditions, the greater the number of variables introduced, the less robust the model.

In the case of neural networks the problem is even more complex: one should not only avoid the introduction of non significant input variables, which only cause noise, but it is equally important to minimise the complexity of the algorithm. If one introduces too many neurons and hidden layers or runs too high a number of iterations to perfectly fit the training data, then the neural network learns the patterns off by hart heart. The neural network becomes less and less capable of generalising: you can rarely lend a tailor-made suit to a friend.

3.5 The selection of the input variables

The first step is therefore to reduce the number of input variables. Traditional statistics aids us in the accomplishment of this task through various stepwise selection techniques.

It could be said that these techniques are only available for Generalised Linear Models, while non linear relationships can be considered in neural networks. Variables that are not linearly significant, might be significant in a more general context. From an operative point of view, in a multivariate analysis, if the number of input variables is sufficiently large, this risk is greatly reduced. A slightly linearly significant variable can gain importance considering non linear relationships, though this is rarely the case for variables that are not linearly significant at all.

These observations are not to be considered as being theoretically rigorous, however they allow us to formulate a simple rule of thumb: the stepwise selection techniques can be extremely useful, to avoid the loss of precious input information it is preferable to apply them in a manner that is not too restrictive.

In this project we applied two parallel selection processes: one considering the target variable as continuous (SAS/STAT's PROC REG SELECTION=MAXR), the other considering it as categorical (SAS/STAT's PROC LOGISTIC SELECTION=STEPWISE SLE=0.15 SLS=0.15). We also considered the precious information supplied by the Servizio Analisi di Marketing of CDE Mondadori. The result of the selections produced a subset of 15-20 variables.

We also attempted to use Principal Component Analysis to transform the original variables (PROC FACTOR, METHOD=P), to then select the most discriminant principal components with reference to the profitability classes (PROC STEPDISC).

3.6 The choice of network architecture

After the variable selection one should choose the best network architecture. The choice is between structures of varying complexity, depending on the number of neurons and hidden layers, activation, link and loss functions.

The %TNN family of SAS macros allows one to automatically determine the optimum number of neurons in all networks with a single hidden layer. The %SELHID macro trains networks containing one, two, three or more neurons and then confronts the obtainable results.

We used %SELHID both with the continuous and categorical target variable. In the first case the number of neurons selected was one, while in the second two was the optimal choice.

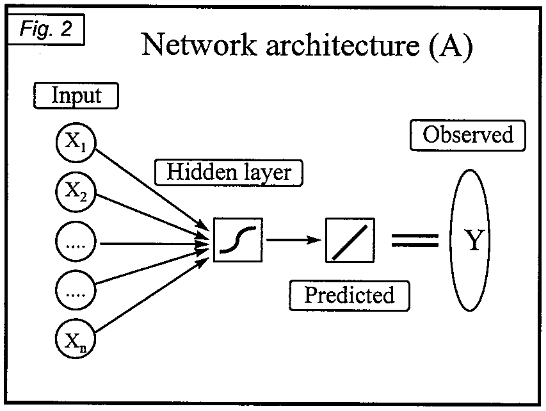


Fig. 2

Network architecture (A)

After that we evaluated architectures with more than one hidden layer. The degree of fit increased slightly, while the predictive ability drastically worsened. Again, the excessive complexity of the network caused a marked lack of robustness.

As far as activation, link and loss functions are concerned it is important to note that their choice depends on the nature of the input and target variables and on the type of problem faced.

Therefore, the network architecture design should be guided by both applicative and statistical considerations. To justify this statement, the following is how we built the network that produced the best results.

First of all, if the target variable is a classification variable and the input variables are quantitative, one should note that he is faced with a discriminant analysis problem.

The introduction of hidden layers allows us to surmount linearity. Given that we have to assign each member to one of four possible profitability classes, two parallel neurons acting as green traffic lights maximise the discriminant power of the tool. As a consequence the activation functions of each neuron should only be logistic (ACT=LOGISTIC), while the activation function of the target variable is multinomial logistic (ACT=MLOGISTIC). As far as the *loss function* of the profitability classes is concerned, the assignment result can be success or not (LOSS=MBERNOULLI).

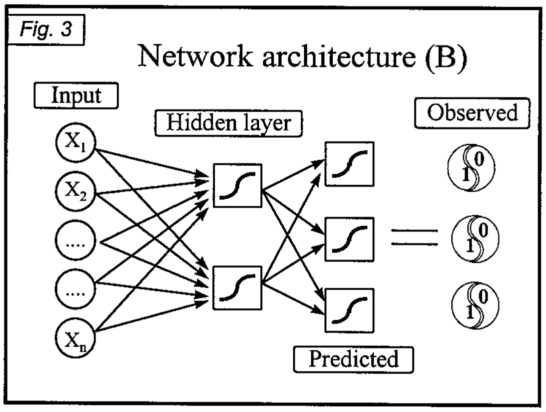


Fig. 3

Network architecture (B)

3.6.1 Measures of the misclassification error

From an operative point of view, the goal of the analysis is mostly to decide, in advance, the exclusion of non profitable book club members: given that, it is extremely important to carefully define the error measures.

To evaluate the predictive power of our tools of analysis, we ran each network over the test sample and we considered the error level with respect to a 4x4 contingency table: the cross tabulation of observed profitability by class and predicted profitability by class.

As it is commonly know in statistics there are several summary measures of association that one can consider to validate the quality of the predictive classification (PROC FREQ, CHISQ). The total misclassification error is given by the number of pairs for which observed and predicted values are different over the total number of pairs.

The operative use of the predictive tool suggests a different way to measure the error level. The department of the Servizio di Analisi di Marketing of CDE knows that members who guarantee a profitability level over the median value are, indeed, those who generate profit. In this way, the predictive tool should allow to correctly classify members above and below the present median profitability value.

In our application the error measure should answer to the following question: if CDE stops investing on members classified as non profitable by the network, how high is the risk of losing good members?

On the other hand, an important piece of information we should obtain from our analysis is the percentage of truly profitable members among those who are predicted to be the best by the network.

3.6.2 Simulation trials

Within the customised definition of error measures, we obtained a significant justification of the choice of the network architecture and, at the same time, we could validate the whole network design by running numerous simulation trials based on resampling techniques. At each trial the training and test subsamples are randomly reallocated. Then the network is trained and tested, monitoring step by step the error percentages.

For each kind of network we considered the summary results of at least 100 simulation trials.

4. The analysis of the results

Before presenting the main results obtained, we have to bare in mind that the project's goal is to build a decision procedure both for the current members and for the prospects.

In the case of the current members, the classification algorithm should predict a profitability category, using early behavioural evidence: a good classification rule would allow to more precisely plan catalogue promotional strategies.

In the case of the prospects, the number of available input variables is greatly reduced. Using a poorer information base and having even less *a priori* hypotheses, we wanted to obtain good suggestions for the best prospect selection by direct mail.

In conclusion, reaching the goals of this project would guarantee an increase in the company profit level.

4.1 The classification of members on the basis of early behaviour

When we considered profitability as a continuous variable, we obtained the best results through a very simple network. This network consisted of 15 input variables and one single hidden layer with only one neuron: a network architecture equivalent to non linear multiple regression, as stated by W.S. Sarle (1994).

Analysing the contingency table in figure 4 it is possible to evaluate the quality of the classification rule. The value of the Phi coefficient (0.884) is undoubtedly satisfactory, but the most interesting observation is the following: if CDE stops investing on those members classified as non profitable by the network, the expected error is less than 8%.

Fig. 4 REDDITIVITA' CONTINUA - RETE CON 1 NEURONE
SINTESI DI 100 PROVE DI SIMULAZIONE SUL TEST

Percent Row Pct Col Pct	1	2	3	4	Total
1	3.82 39.39 72.69	4.92 50.74 12.44	0.73 7.47 2.3	0.23 2.4 0.99	9.7
2	1.06 2.61 20.23	27.35 67.11 69.1	11.81 28.98 37.45	0.53 1.3 2.24	40.75
3	0.21 0.83 3.9	5.88 23.69 14.86	12.91 51.97 40.92	5.84 23.51 24.72	24.83
4	0.17 0.68 3.18	1.42 5.76 3.59	6.1 24.67 19.33	17.02 68.89 72.06	24.71
Total	5.26	39.58	31.54	23.62	100

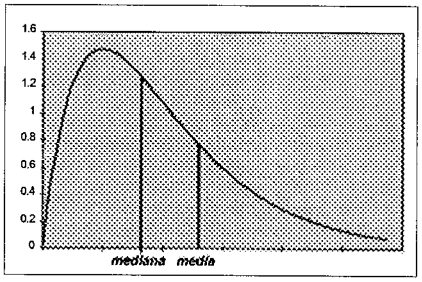
When we considered profitability as categorical, we obtained the best results from a network with a single hidden layer and two neurons, equivalent to a non linear discriminant model. The contingency table in figure 5 has a higher value of the Phi coefficient (0.914), i.e. the total misclassification error is lower. But the risk of eliminating profitable members increases to 14%.

Fig. 5 REDDITIVITA' CATEGORICA - RETE CON 2 NEURONI
SINTESI DI 100 PROVE DI SIMULAZIONE SUL TEST

Percent Row Pct Col Pct	1	2	3	4	Total
1	4.74 47.11 71.64	4.8 47.64 8.82	0.26 2.63 1.91	0.27 2.63 1.05	10.07
2	1.33 3.27 20.1	36.07 88.74 66.37	2.64 6.49 19.11	0.61 1.5 2.41	40.65
3	0.36 1.48 5.5	10.38 42.31 19.1	7.04 28.67 50.95	6.76 27.54 26.8	24.54
4	0.18 0.74 2.77	3.1 12.52 5.7	3.87 15.65 28.02	17.59 71.09 69.74	24.74
Total	6.62	54.35	13.81	25.22	100

Economic considerations make the choice easier and the first network can be considered the best applicative solution (there is a lower probability of incorrectly eliminating a member). From a statistical point of view the question is why a continuous target brings about a more prudential classification rule (in the sense that the network is weakly biased towards overestimation). The pattern of the profitability distribution function gives an immediate answer: there are very large profitability values with a relative low frequency. This fact can produce overestimation if the predictive algorithm considers the profitability measure of each single member.

Fig. 6 DISTRIBUZIONE DELLA REDDITIVITA'



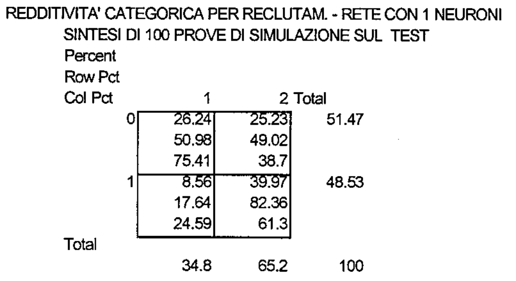
4.2 The selection of the best prospects

The design of a neural network for the selection of the best prospects can be the answer to an extremely clear marketing necessity. Furthermore neural networks are supposed to be better than traditional statistics, when the information bases are poor and *a priori* hypotheses very weak.

The available input variables are, in this case, geographical area, information about the origin of the contact and about subscriptions to Mondadori's magazines. The target variable is a dummy variable, profitable or not profitable.

The best performance has been obtained with a network with a single hidden neuron. The contingency table in figure 7 shows that 61.3% of the truly profitable members are correctly classified by the network: confronting these results with the 48.5% of the marginal distribution makes the advantage evident.

Fig. 7



```
REDDITIVITA' CATEGORICA PER RECLUTAM. - RETE CON 1 NEURONI
        SINTESI DI 100 PROVE DI SIMULAZIONE SUL  TEST
        Percent
        Row Pct
        Col Pct            1          2 Total
                   0|   26.24|   25.23|     51.47
                    |   50.98|   49.02|
                    |   75.41|    38.7|
                   1|    8.56|   39.97|     48.53
                    |   17.64|   82.36|
                    |   24.59|    61.3|
        Total
                        34.8      65.2       100
```

The problem of determining the rule for the selection of the best prospects has also been faced applying standard statistical modelling (multiple linear regression, logistic regression). On the bases of simulation trials we can say that, when available information is poor (lack of strong behavioural evidences), neural networks seem to perform a slightly better than traditional models.

If we go back to consider the problem of the classification of the members, we can notice (figure 4) that a little knowledge of the members' behaviour increases the predictive power of the network: over 80% of the members with a positive predicted profitability, will truly be profitable.

5. Conclusions

On the bases of this experience we can make some considerations on neural network application in database marketing.

First of all we should highlight the importance of both a specific knowledge of the field of application and of data manipulation and screening: an efficient selection of input variables, supported by statistical criteria has the great advantage of making the analysis simpler and more robust.

The definition of an error measurement criterion, as well as a significant number of simulation trials are the necessary steps to validate the goodness of fit and ability to generalise of neural networks.

Having said this, neural networks are versatile and easy-to-use.

The main obstacle lies in interpreting each input variables effect on the target variable's predicted values. This aspect, probably due to the fact that NN are such a recent technology, currently limits their use in decision making.

The %TNN macros are *user friendly* and efficient. At present the SAS System is the only environment that allows us to organise our data bases, develop traditional statistical models and design neural networks. This data warehousing solution allows us to compare the results of each method easily and homogeneously and, hence, choose the best analysis process.

As far as CDE is concerned, the ability to rapidly and accurately predict the profitability of each member opens the door to appealing perspectives in the definition of ever more personalised promotional strategies, with assured gains in terms of corporate profitability.

References

Agresti, A. (1990). Categorical Data Analysis, John Wiley & Sons, New York

Bouroche, J.M. e Saporta, G. (1980). L'analyse des donnėes, C.L.U. Editrice, Napoli.

Jobson, J.D. (1992). Applied Multivariate Data Analysis, Springer-Verlag New York.

Fabbri G,Orsini R. (1993), Reti neurali per le scienze economiche, Franco Muzzio Editore.

Sarle, W.S. (1994). Neural Networks and Statistial Models, SAS Institute Inc., Cary, NC, USA.

Sarle, W.S. (1995). Neural Networks Implementation in SAS Software, SAS Institute Inc., Cary, NC, USA.

SAS Institute Inc., *SAS/STAT User's Guide, Version 6, Fourth Edition,* Cary, NC: SAS Institute Inc., 1989.

Wasserman, P.D. (1989), Neural Computing Theory and Practice, New York: Van Nostrand Reinhold.

Wasserman, P.D. (1993), Advanced Methods in Neural Computing, New York: Van Nostrand Reinhold.

Guido Cuzzocrea: CIDER Univ. Bocconi,
Tel: ++ 39-2-26823641
Fax: ++ 39-2-26823649
E-mail:
GUIDO.CUZZOCREA@UNI-BOCCONI.IT

Alberto Saccardi: CIDER Univ. Bocconi,
Tel. ++ 39-2-58363080
Fax: ++ 39-2-58363087
E-mail:
ALBERTO.SACCARDI@UNI-BOCCONI.IT