
Validating a Neural Network Application - The Case of Financial Diagnosis

Per Egil Pedersen

Department of Economics and Business Administration
Agder College
4890 Grimstad, Norway
Per.Pedersen@hia.no

Abstract

It has been argued that neural network applications should be benchmarked using several data sets of realistic and real problems, and competing algorithms (Prechelt, 1995). However, if applying a neural network model to a particular real problem is in focus, validation should be considered as a suitability evaluation in which several bases of evaluation are combined in a composite judgement. In this paper, five bases of such evaluation are introduced and applied to the validation of a neural network model of financial diagnosis.

1 INTRODUCTION

To improve the validation of neural network (NN) algorithms, it has been suggested that multiple data sets from realistic and real problems should be used. If the data sets are produced with careful consideration of the threats to their validity and the proposition tested is NN algorithms performing better than other algorithms, this seems a reasonable strategy. As a general principle for evaluating NN applications to a particular real problem, it is somewhat limited. Such evaluations are more likely to focus on the *suitability* of an algorithm to a particular problem. Suitability evaluations start with propositions about the new and competing models, they are based upon theory of the application area, they require that the validity of the data sets can be demonstrated, they are less concerned with the generality of the model, and they investigate multiple aspects of performance.

Many of the improvements suggested for validating NN algorithms (Prechelt, 1995) are relevant to applications of these models to real problems, but the five bases listed above make suitability evaluations different from traditional validation. We elaborate on these five bases in section 2. In section 3, a multilayered perceptron (MLP) model of financial diagnosis is evaluated using the five bases of evaluation. In the final section, we discuss the general applicability of the five evaluation bases in light of the conclusions that could be drawn from applying them to the MLP model of financial diagnosis.

2. BASES OF EVALUATION

The first basis of evaluation is the *propositions* made about NN models in a particular

application area. Superior performance to other models when evaluated by a simple measure such as mean square error is only one of several possible propositions, and suitability evaluations may focus on some or, preferably, all propositions. The second basis of evaluation is *theory* of the application area. Usually, existing models are founded on empirical or theoretical arguments. Such arguments are used to select relevant features of the problem and relevant responses or outcomes. Furthermore, a theory also suggests a particular relationship between relevant features and responses or outcomes. Thus, NN models should not only produce the right outcomes; the outcomes should also be produced in the right way. The third basis of evaluation is the *validity of the data sets* used in the application. Both internal and external validity aspects are of relevance. Recommended experimental, measurement and sampling procedures should be followed and reported in NN applications as in other model applications, so that the validity of the data sets can be evaluated. The fourth basis for evaluating the suitability of the NN model is how the propositions of the NN algorithm is evaluated *statistically*. The traditional proposition is that of superior performance. It has been argued that this proposition is evaluated poorly (Prechelt, 1995). Even though the use of several data sets from multiple problems is irrelevant to tests of suitability to a particular problem, using multiple data sets and proper statistical validation against competing models are highly relevant. The final basis of evaluation is the models' *behaviour and representations*. As an extension of the argument that model outcomes must be produced in the right way, analysis of the models' behaviour under varying conditions can be performed. In addition, analysis of the models' representations may be performed by comparing the representations of the model to the relevant "representations" proposed by theory. When all the five bases of evaluation are used, a composite judgement of the suitability of an NN model to a particular real problem can be made. As an example of how these evaluation bases can be used, we apply them to the validation of an NN model of financial diagnosis.

3 EVALUATING A MODEL OF FINANCIAL DIAGNOSIS

Financial diagnosis is the classification task performed when a subject makes a judgement of the financial situation of the firm based upon information from the financial statement (Methlie, 1987). This task is performed in several contexts, such as bankruptcy prediction, going concern judgement and loan decision contexts. The diagnosis task constitutes an important basis for prediction and judgement in all these contexts and has many similarities across the different contexts.

3.1 PROPOSITIONS

Several authors have proposed that NN models outperform traditional statistical techniques in predicting the environmental outcome of these contexts (Tam & Kiang, 1992; Wilson & Sharda, 1994). Some researchers have argued that the superior performance of these models can be explained by their ability to represent intermediate features or abstractions of relevance to the task (Raghupathi et al., 1991; Srivastava, 1993). Some researchers even have proposed that these abstractions resemble those used by skilled diagnosticians (Berry & Trigueiros, 1993; Singleton & Surkan, 1995).

These propositions represent the first basis for evaluating an NN application to financial diagnosis, but their form requires a change from a predictive to a behavioural perspective. The first proposition can be validated by comparing model performance to the performance of competing algorithms. The second proposition can be validated by testing the performance differences between models with and without the ability to develop internal representations. The third proposition is much more difficult to validate, but an evaluation can be performed by comparing the internal representations of NN models to cognitively relevant representations suggested by theory of financial

diagnosis. A suitability evaluation of an NN model of financial diagnosis should include the evaluation of all three propositions, and consequently, it goes beyond *traditional* validation.

3.2 FINANCIAL DIAGNOSIS THEORY

Traditional studies in accounting and finance take a *predictive* approach to financial diagnosis, and apply a variety of statistical models to predict the environmental outcome in different contexts of the task (see e.g. Altman et al., 1981). Behavioural studies take one of two different approaches. *Judgement modelling* studies model diagnosticians' *judgements* of the outcomes in different task contexts with methods similar to those of the predictive approach (e.g. Libby, 1975). Cognitive studies are either experimental or descriptive in orientation. *Experimental cognitive* studies apply cognitive theory to predict the effect on diagnosticians' judgements of manipulating variables such as information content and form (e.g. Iselin, 1993). *Descriptive cognitive* studies follow the information processing theory tradition of Newell and Simon (1972) and apply protocol analysis to the verbal utterances of human diagnosticians during diagnosis (e.g. Biggs et al., 1993).

Theories of all these approaches contribute in different ways to how NN applications to financial diagnosis should be evaluated. Predictive studies identify the relevant diagnostic features of the task, and both judgement modelling and descriptive cognitive studies identify relevant responses or outcomes. All approaches suggest specific models of the financial diagnosis task. Most of these are formal models that can be used as benchmark models. All approaches also contribute to the identification of diagnostic knowledge and intermediate abstractions relevant to the task. Typically, these abstractions are formalised in diagnostic concepts the most frequently applied being leverage, profitability, liquidity and financing (Pedersen, 1995). These concepts correspond well to the "underlying dimensions" found in studies applying principal components analysis to financial statement information (e.g. Gombola & Ketz, 1983).

The theory of financial diagnosis guides suitability evaluation in several ways. Diagnostic features of the theory should also be diagnostic in an NN model. Intermediate abstractions represented in NN models of financial diagnosis should somehow resemble diagnostic concepts of financial diagnosis theory, and the cognitive relevance of behavioural models can be evaluated by analysis of internal representations.

Financial diagnosis theory also suggests a model including intermediate abstractions of diagnostic relevance between the features and responses of the task. A model introducing a set of represented abstractions intermediating original features and diagnostic response is very similar to the simplest multilayered NN models including one layer of hidden units. Even though many different NN models can be applied, and these can represent different *types* of intermediate abstractions, a multilayered perceptron (MLP) including one layer of hidden units is suggested here as a model of the financial diagnosis task.

3.3 RESEARCH DESIGN

To provide a valid data set of financial diagnoses, a controlled experiment was set up. A random sample of 75 firms was selected from an established Norwegian small firms register. Financial statement information from two consecutive years was collected, and balance statements, income statements, funds flow statements and selected ratios were presented in a booklet to 108 subjects. Of the subjects, 98 had prior auditing experience and were familiar with the financial diagnosis task. The rest of the subjects had other accounting experience. Thus, the subjects were considered skilled diagnosticians. To

control for individual variation in diagnostic behaviour, all subjects received a booklet of three financial statements. The distribution procedure was randomised, so that on an average 4.3 skilled subjects diagnosed each firm. The subjects used a response form to characterise the profitability, financing, liquidity, leverage and general financial situation of each firm on predefined 5-point Likert scales. A composite judge measure of the subjects' evaluation of the general financial situation was calculated as the average value of their response on the general financial situation indicators. Composite judge measures have been recommended in financial diagnosis studies to reduce individual variation and improve diagnostic predictions (Libby, 1981). The composite judge measure also transformed the response of the classification task into an approximately interval scaled variable with good distribution properties without changing the classificatory character of the task itself.

Each subject also indicated the most important cues of the financial statements. Of the 108 subjects, 105 indicated the use of one or more cues. Of these, 83.8 % indicated that cue values of two consecutive years were used, suggesting that the majority of the subjects were sensitive to correlated features. 55.4 % of all cues indicated were cues representing financial ratios. The main reason for the subjects using these ratios was probably that the ratios were almost firm size independent. By selecting the 16 most frequently indicated financial ratios with cue values from two consecutive years as preliminary input variables in the MLP model, the most important features were represented by variables with size independence and good distribution properties.

Even though recommended experimental, measurement and sampling procedures were followed, evaluating the validity of the applied procedures is left to the reader. Naturally, such evaluations can only take place if these procedures are reported.

3.4 STATISTICAL VALIDATION OF PROPOSITIONS

Since 75 data sets of financial diagnoses were too few cases to properly set the weight values of an MLP model with 32 input units and one response unit, we applied sensitivity analysis similar to that of Moody and Utans (1995) to constrain the number of input variables to 12. These input variables represented values from two consecutive years of the six financial cues indicated by the diagnosticians as the six most important.

The MLP model was set up following the original principles of Rumelhart et al. (1986). Asymmetric sigmoid output functions were used. Inputs and output were scaled to the [0,1] scale, but no other representational transformations were made. The original epoch-based learning of Rumelhart et al. (1986) was applied, setting η of the hidden layer to 0.5 and η of the output layer to 0.4. A small momentum term α of 0.1 was used. Initial weights were randomly selected from a uniform distribution in the range [-0.2, 0.2] for the hidden layer and in the range [-0.7, 0.7] for the output layer. The recommended N-fold cross validated mean square error (MSE) was used in all performance evaluations (White, 1990). To control model complexity, MSE was computed for models including 0, 2, 4, 6, 8, 10, 12 and 14 hidden units. This procedure was similar to a constructive algorithm selecting the best model based upon N-fold cross validated performance measures.

Six *benchmark models* were developed applying recommended procedures of the accounting and finance literature (Libby, 1975). Two models used OLS regression on a selected set of principal components of the original financial cues to avoid multicollinearity problems. Two benchmark models were produced using OLS regression on the original 32 variables selected by the subjects, and two benchmark models were produced using the same method on the 12 input variables of the MLP model. Performance results of the best of these benchmarks are shown in table 1.

Table 1. Performance results of the best benchmark models.

Benchmark / Results	MSE	Corr. with target	Corr. with distance
9-factor regression	0.232	0.059	0.252
12-variables regression	0.221	0.169	0.149
12-var. stepwise regression	0.212	0.177	0.128

Table 1 shows the cross validated mean square errors (MSE) and two measures illustrating the distribution of the models' error terms. Both the correlations of the error terms with the target output values and the correlations of the error terms with distance from mean target values indicate an unfavourable skewness in the distributions of the error terms.

A selection of the performance results of the MLP model is shown in table 2. The number of hidden units is indicated in the model name. Cross validated MSE is shown for every 5000 learning iterations. As expected, MSE reached a flat minimum during learning, and increased with learning beyond this minimum due to overtraining. Even though performance was best for the model with four hidden units, the larger difference in MSE was between the models with hidden units and the model without such units.

Simple t-tests of the differences between means were used to evaluate the propositions made in section 3.1. A t-test of the difference in MSE between the MLP model with four hidden units stopped at the optimal number of training iterations and the best benchmark model showed that the difference was significant and in favour of the MLP model at $\alpha = 0.05$ ($t=1.96$, $d.f.=74$). This test indicated that the MLP model was superior in modelling the diagnostic response of our subjects. The correlation of the MLP model's error terms with target values was 0.048, and the correlation of the error terms with distance from mean targets was -0.055. These measures indicated that the error terms had a uniform distribution over the range of target values, and the unfortunate distribution of errors found in the benchmark models was not found in the MLP model. Thus, we can conclude that the superiority proposition of section 3.1 holds when evaluated both by MSE and by the properties of the error term distributions.

Table 2. Selected cross validated MSEs for the MLP model

Model: / Iterations:	5000	10000	15000	20000	25000	30000
HID0	0.182	0.182	0.185	0.186	0.186	0.187
HID2	0.232	0.185	0.171	0.159	0.158	0.156
HID4	0.175	0.160	0.151	0.147	0.147	0.145
HID6	0.180	0.174	0.163	0.157	0.156	0.159

A t-test of the difference in MSE between the best NN models with and without hidden units showed a significant difference in favour of the MLP model at $\alpha = 0.05$ ($t=2.27$, $d.f.=74$). Since the two models were developed with comparable parameter settings, this test supported the proposition of section 3.1 that the reason for the improved performance of the more complex models is their ability to represent and utilise intermediate abstractions of relevance to diagnosis. In our suitability evaluation, this formal test contributes positively to a conclusion that the MLP model is a suitable model of financial diagnosis.

With respect to the bases of evaluation, simulation procedures have been reported in some detail. In addition, competing benchmark models and multiple measures of performance were used. Applying cross validated MSE also allowed the use of simple statistical tests to evaluate two of the propositions of section 3.1.

3.5 REPRESENTATIONAL ANALYSIS

To investigate the last proposition of section 3.1, analysis of the connectionist model representations was necessary. Principal components analysis of the 12 input variables of the connectionist model showed that the factors with the highest eigenvalues could be interpreted as representing the diagnostic concepts profitability, leverage, liquidity and financing. Thus, it could be suggested that similar dimensions or "factors" were represented by the four hidden units of the best MLP model.

When using N-fold cross validation, 75 different "versions" of the model with four hidden units were developed. These models had different weight patterns, but by adjusting the weight values for differences in bias weights and turning the signs so that all hidden units were excitatory, the weight patterns could be used in cluster analysis (Hanson and Burr, 1990) to identify groups of hidden units with similar functionality. This analysis revealed that all models had at least two very local hidden units belonging to different clusters and accounting for most of the models' response variation. The rest of the hidden units were more distributed, and only accounted for small variation in model response. Hinton diagrams of the incoming weights of two representative local hidden units are shown in figure 1.

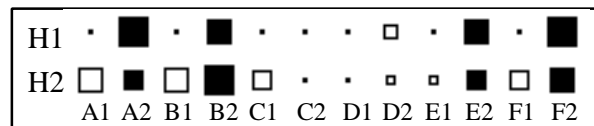


Figure 1. Weight pattern of two local hidden units.

In figure 1, the weights from six financial cues A - F to two local hidden units (H1 and H2) are shown. For each financial cue, two input units represent values of the first and second of two consecutive years. The financial cues are from left to right; operating margin (A), net income/total assets (B), net income/equity (C), average interest rate (D), acid test (E) and equity/total assets (F). These weights indicate no specialisation of each hidden unit on financial cues representing different diagnostic concepts. Instead, it seems that the upper hidden unit (H1) uses the second year values of a broad set of financial cues representing several diagnostic concepts. This hidden unit detects the *level* of the financial cues of the *latest year*, and may be termed "level-oriented". The lower hidden unit (H2) uses both values of a broad set of financial cues, but the two weights from cues of each consecutive year have opposite signs. Consequently, this unit is activated by the *change* in the level of a broad set of financial cues. The hidden unit actually computes the general trend in the financial cues from one year to the next. Consequently, this hidden unit is termed "change-oriented".

To form the final diagnosis of the financial situation, the model seemed to use the two diagnostic concepts "level" and "change" instead of the suggested diagnostic concepts of financial diagnosis theory. This finding is at odds with our expectations and contributes negatively to a suitability conclusion. However, the internal representations may still have cognitive relevance, and may actually suggest that the diagnosticians used heuristics different from what was recommended by theory. With this finding, the analysis also served an exploratory purpose.

4 CONCLUSIONS

We have argued that the validation of an NN application to a real problem should be considered as a suitability evaluation with five bases. We applied these principles to an NN model of financial diagnosis. Using the five bases of evaluation, we concluded that the MLP model was a suitable model of financial diagnosis, but equally important, it was possible for the reader to make an independent judgement of the suitability of the model.

It may be argued that the financial diagnosis task is atypical of NN applications, such as its behavioural character. However, the five bases of evaluation apply to most NN applications to real problems. NN applications are seldom applied without specific propositions about their properties as models. The application area is not likely to be so unexplored that no theories have been suggested. Valid measurement and sampling procedures are equally important in all data collection, irrespective of the model finally selected, and applying proper simulation procedures to secure statistical validity has already been focused by other NN researchers (Prechelt, 1995). Finally, analysis of NN models' internal representations adds a basis of evaluation relevant to all application areas with a suggested theoretical relationship between features and outcomes. The literature on validating NN algorithms have recently stressed the statistical basis of evaluation, whereas we have stressed the *combination* of several bases when evaluating NN applications to real problems. The statistical basis is one of these, but the other bases are of equal importance.

Of the five bases, two should be given special attention. First, applying and reporting valid data collection procedures seem to have been given little attention in NN applications. NN researchers seldom report the procedures applied in data collection whereas traditional studies pay considerable attention to the validity of data sets. Second, early NN applications focused on analysis of NN models' internal representations. Specific analysis techniques were developed (Hanson & Burr, 1990), and applied (Gorman & Sejnowski, 1988) to these representations. Our view is that these techniques should be applied and developed further so that NN model evaluations are not left to statistical validation alone.

If applying the five bases, the evaluation is given a form closer to traditional model evaluations within the problem areas, and NN models will more likely be considered natural alternative models in the application areas¹.

¹ The data set used in this study is available upon request to Per.Pedersen@hia.no

References

- Altman, E.I., Avery, R.B., Eisenbeis, R.A. & Sinkey, J.F. (1981). *Application of classification techniques in business, banking and finance*. Greenwich, CT: JAI Press.
- Berry, R., & Trigueiros, D. (1993). Applying neural networks to the extraction of knowledge from accounting reports: A classification study. In Trippi, R.R. & Turban, E. (eds.), *Neural networks in finance and investing* (p. 103-123). Chicago, IL: Probus Publishing.
- Biggs, S.F., Selfridge, M., & Krupka, G.R. (1993). A computational model of auditor knowledge and reasoning processes in the going-concern judgement. *Auditing: A Journal of Practice & Theory*, 12(supplement), 82-99.
- Gombola, M.J., & Ketz, J.E. (1983). Financial ratio patterns in retail and manufacturing organizations. *Financial Management*, 12(2), 45-56.
- Gorman, R.P., & Sejnowski, T.J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1, 75-89.
- Hanson, S.J., & Burr, D.J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences*, 13, 471-518.
- Iselin, E.R. (1993). The effects of the information and data properties of financial ratios and statements on managerial decision quality. *Journal of Business Finance & Accounting*, 20, 249-266.
- Libby, R. (1975). Accounting ratios and the prediction of failure: Some behavioral evidence. *Journal of Accounting Research*, spring 1975, 150-161.
- Libby R. (1981). *Accounting and human information processing: Theory and applications*. Englewood Cliffs, NJ: Prentice Hall.
- Methlie, L.B. (1987). On knowledge based decision support systems for financial diagnosis. In Holsapple, C.W., & Whinston, A.B. (Eds.), *Decision support systems: Theory and application*. Berlin: Springer Verlag.
- Moody, J.E., & Utans, J. (1995). Architecture selection strategies for neural networks: Application to corporate bond rating prediction. In Refenes, A.P. (Ed.), *Neural networks in the capital markets* (pp. 277-300). New York: Wiley.
- Newell, A., & Simon, H.A. (1972) *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Pedersen, P.E. (1995). Connectionist models of financial diagnosis. Unpublished doctoral dissertation, Norwegian School of Economics and Business Administration, Bergen, Norway.
- Prechelt, L. (1995). Some notes on neural learning algorithm benchmarking. *Neurocomputing*, 9, 343-347.
- Raghupathi, W., Schkade, L.L., & Raju, B.S. (1991). A neural network approach to bankruptcy prediction. Reprinted in Trippi, R.R., & Turban, E. (Eds.), *Neural networks in finance and investing* (pp. 141-158). Chicago, IL: Probus Publishing.
- Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In Rumelhart, D.E., & McClelland, J.L. (Eds.), *Parallel distributed processing; Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.
- Singleton, J.C., & Surkan, A.J. (1995). Bond rating with neural networks. In Refenes, A.P. (Ed.), *Neural networks in the capital markets* (301-307). New York: Wiley.
- Srinivasan, V., & Kim, Y.H. (1987). Credit granting: A comparative analysis of classification procedures. *The Journal of Finance*, 42, 665-683.
- Srivastava, R.P. (1992). Automating judgmental decisions using neural networks: A model for processing business loan applications. *Proceedings of the 20th Annual ACM Computer Science Conference* (p. 351-357), NY: Association for Computing Machinery.
- Tam, K.Y., & Kiang, M.Y. (1992). Managerial applications of neural networks: The

- case of bank failure predictions. *Management Science*, 38, 926-947.
- White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3, 535-549.
- Wilson, R., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems*, 11, 545-557.

