IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP4308

Training Distribution Strategies for Optimizing Neural Network Classification Models

Steven Walczak
University of Colorado at Denver, College of Business, swalczak@carbon.cudenver.edu

Irena Yegorova
City University of New York, iyegor72@yahoo.com

Bruce H. Andrews
University of Southern Maine, bandrews@usm.maine.edu

ABSTRACT

Neural networks have been repeatedly shown to outperform traditional statistical modeling techniques for both discriminant analysis and forecasting. While questions regarding the effects of architecture, input variable selection, learning algorithm, and size of training sets on the neural network model's performance have been addressed, little attention has been focused on distribution effects of training and out-of-sample populations on neural network performance. This article examines the effect of changing the population distribution within training sets, in particular for a credit risk assessment problem.

INTRODUCTION

As information availability continues to grow (e.g., through the World-Wide-Web), the complexity of business decision making increases proportionally (Walczak 2001b). Decision support systems, data mining tools, and artificial intelligence programs attempt to facilitate business decision making. Neural networks, a nonparametric modeling technique, have been shown to work well for many types of business problems (Li 1994, Patuwo et al. 1993, Widrow et al. 1994, Zahedi 1996). Many researchers have demonstrated empirically that neural network models outperform the more traditional statistical models including regression, logit, decision trees, and discriminant analysis (Bansal et al. 1993, Patuwo et al. 1993, Piramuthu 1994).

What factors enable the nonparametric neural network models to outperform the traditional parametric statistical methods? All parametric statistical methods, including regression and Bayesian classification, necessitate that the population distribution or variable distributions adhere to pre-defined characteristics such as a multivariate normal distribution (Klecka 1980). When variable distributions are unknown as frequently happens in business problem solving (e.g., bankruptcy prediction and investment risk analysis), then the more traditional methods, including Bayesian classification, cannot be applied accurately (Patuwo et al. 1993). Nonparametric approaches, such as neural networks, are needed to determine group conditional distribution functions when *a priori* distributions are unknown (McLachlan 1992).

An unanswered question in the field of neural networks is the effect of unequal population distributions and their maintenance as a representative sample or alteration as a stratified sample in the training group used to build the neural network classification model. A heuristic that is normally followed by neural network researchers is to include the greatest amount of data possible in the training samples (Hu et al. 1999, Patuwo et al. 1993, Smith 1993, Zahedi 1996), which necessarily forces a representative training sample that maintains distribution differences.

Another potential problem with the use of stratified training sets is that the effect of unequal distributions is greatest when the overall population has very few elements (e.g., a 90/10 distribution between two categories over 100 samples leaves only 10 samples of the smaller category to be divided across the training and test sets) and in practice many interesting business problems have limited data sets (Smith 1993). Berardi and Zhang (1999) specifically state that small group classification with neural networks is particularly sensitive to sampling variations. Although recent evidence suggests that neural network training may be optimized with very small data sets (Walczak 2001a), most researchers feel more confident if larger training sets can be instantiated. A common method for maximizing the size of the training set when small real-world data populations exist is to utilize either bootstrapping or jackknifing (Efron 1982). The jackknife process, which is a specialization of the bootstrap method, creates N different training sets of size N-1, with each data sample being used as the holdout test sample a single time. The aggregation of the N test results effectively approximates the results of an overall model (Efron 1982). Unfortunately, the use of the bootstrap or jackknife methodologies necessarily creates a representative training sample that closely emulates the data distribution inequalities found in the population.

This article examines the effect of using stratified training samples when data samples have an unequal distribution for a two-group classification problem in the domain of credit scoring for bank loans. A modified bootstrap process is created to maintain predefined distributions within the training sets. The results empirically indicate that equal distributions of each category within the training set produces the optimal generalization capabilities of neural network classification models, while representative training samples (especially when group membership probabilities are widely disparate) will produce sub-optimal results.

BACKGROUND ON NEURAL NETWORK BUSINESS CLASSIFICATION MODELS

As previously stated, neural networks are widely used for solving business classification problems. Two of the more common applications of neural network classification models in the financial domain (Zahedi 1996) are for bankruptcy prediction (Fletcher & Goss 1993, Raghupathi 1996, Sharda & Wilson 1996, Wilson & Sharda 1994) and credit/loan appraisal (Piramuthu et al 1994, West 2000). West (2000) indicates that a lender using a neural network credit scoring system was able to achieve a 10 percent improvement in accuracy over their previous system.

Some of the interesting aspects of previous research are that three of the five cited studies use a population of paired samples with representative training sets, so that the failed and non-failed groups or default and full repayment groups have equal probability of occurrence. Fletcher and Goss (1993) use a sample size of 36 firms and rely on an N-fold cross validation to handle the small amount of data, which in practice is the same as a bootstrap. Small data sets are a common problem in developing business classification models. Piramuthu et al. (1994) use two different data sets, each with equal probabilities of membership, that have 36 and 100 samples respectively for loan default classification (two groups) and credit risk classification (five groups) and a 10-fold cross validation (bootstrap) is used to overcome the small sample sizes.

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

Wilson and Sharda (1994) performed some early experiments in the domain of bankruptcy classification that examined the effect of different distributions between two groups on test sets that also had varying distributions. They used three combinations of training and test sets: 50/50, 80/20, and 90/10 distribution probabilities with the second number representing bankrupt firms. Their preliminary findings indicated that classification results for the nine different neural network models was best when the distribution of the training set matched the distribution of the test set. Hence, a representative training sample that preserves the distribution inequalities of the population produces the optimal performance. Hu et al. (1996) follow the representative training sample philosophy to construct neural network models that classify Sino-foreign joint ventures as unsuccessful or successful, with the population having a 90/10 probability distribution. The initial results for the Sino-foreign performance classification problem were sub-optimal.

Later results on bankruptcy problem that utilizes three different group membership probabilities (Sharda & Wilson 1996), indicated that the stratified 50/50 training group (representative for the 50/50 test set only) outperformed all other representative or stratified training set neural network models on all combinations of test set distributions (50/50, 80/20, and 90/10). Unfortunately, the other bankruptcy and credit scoring neural network research forces a 50/50 representative distribution and test set by limiting the population and hence the sensitivity of the smaller group in the real-world to the training set distribution cannot be effectively measured (Berardi & Zhang 1999).

METHOD

Before presenting the methodology used to investigate the effect of representative versus stratified training sets for classification problems that have unequal population distributions, the impetus for considering a non-representative training set is examined.

Need for Stratification in the Training Set

Whenever a classification problem has equal probability of membership in each of its categories, then the issue of representative versus stratified training sets is eliminated. However, when unequal probabilities of group membership exist, a classification model maximizes its generalization performance by weighting predictions accordingly (Klecka 1980, McLachlan 1992). This means that if a two-group classification problem has a probability of membership in the first group of 80 percent, then it should be 80 percent likely that any unclassified sample belongs to the first group. Alternatively, an 80 percent classification accuracy may be achieved by placing all new observations into the first group, regardless of actual group membership.

Significant inequality within group distributions may cause certain neural network and statistical models to maximize their performance by effectively eliminating membership in the smaller group. As an example, a logistic regression model was constructed for the previously mentioned Sino-foreign joint venture (Hu et al. 1996). This logistic regression model achieved a classification performance of almost 91 percent, which was over 2 percent above the closest neural network model, by classifying all of the joint ventures as not-so-successful (group 2). The distribution between the not-so-successful group and the successful group was 90.84/9.16 for both the training and evaluation samples. The disparity of the exclusion effect just discussed increases as the probability of group membership in the smaller group approaches zero.

Problem and Data Description

The classification problem used to investigate the effect of stratified versus representative training samples is a loan default/repayment problem. The data set is the same as used by Yegorova et al. (2000) and is acquired from the files of a regional economic development lender whose role, among other things, is to provide financing to small companies that are expected to promote job growth and contribute to the local economy. A cross-sectional review of the industries involved reveals a variety of businesses including woodworking, paper, boating, and equipment manufacturing. The sample used in this paper is limited to loans extended to small, expanding manufacturing businesses, since this category has the largest percentage of loans and also includes a larger proportion of loan defaults. The lender's terminated loan portfolio includes 102 loans made to expanding manufacturing companies. Terminated loans are defined here as loans that are either paid off by the borrower or are in default. Loans that were in non-accrual status as of the sample date, but not charged-off by the lender, were excluded from the sample. This elimination process and incomplete data resulted in only 61 loans with 15 defaults in the final sample. The sample data have a 75/25 distribution for the paid off and defaulted loans made by the lender.

Data from the lender and transformations include 138 variables representing various loan characteristics. Selection of the input variables may have a significant effect on the performance of neural network, as well as statistical, models (Smith 1993, Walczak & Cerpa 1999). The focus of the presented research is to evaluate the effect of stratified training sets and is not concerned with the construction of an optimal loan default evaluation model and as such selects nine variables that are common elements in a number of financial ratios. The nine variables selected for the presented research models are: current assets, liability, current liability, inventory, working capital, equity, sales, cash, and long term debt. These variables should provide a breadth of information regarding the loan recipients and still minimizes the size of the neural network to limit extraneous effects from noise and over-fitting of the data set.

Neural Network Architecture and Training Set Construction

Initially, two different learning algorithms are evaluated, backpropagation (BP) and learning vector quantization (LVQ). Each neural network has the nine input variables (described in the previous section) and two output variables. The two output variables serve as categorical variables for full repayment and default status on the loans. The use of two output variables representing the different classifications is required by the LVQ training method and consequently is also used for the BP training method to eliminate any unforeseen biasing effects from a different architecture. Additionally, the use of the two categorical output variables also eliminates any arbitrary decision regarding the optimum cutoff value for a single valued output to be mapped to the two classification groups.

The size of the networks is minimized to avoid difficulties from over-fitting the data and each architecture has it's quantity of hidden nodes incremented by two until generalization performance starts to decline, indicating over-fitting of the data (Walczak & Cerpa 1999). A subset of the full data set is used to determine the best architecture for each learning algorithm and then these architectures are used exclusively, to again eliminate any bias from using different architectures, to train and test the neural network models anew. The best performing architecture for the BP algorithm is a two-hidden layer architecture with 8 perceptrons in the first hidden layer and 4 perceptrons in the second hidden layer, while the best performing architecture for the LVQ algorithm has a Kohonen layer of 18 elements.

The data set is then divided into training and test sets to build and evaluate the generalization performance of each of the two networks. The first collection of training and test sets is generated using the jackknife methodology (a specialization of the bootstrap method) (Efron 1982), which holds out a single data sample and uses the remaining 60 data samples as the training group. This process is repeated 61 times so that every data sample may serve as the single test case and the neural network is completely re-trained with each of the 60 new training sets to generate an unbiased model. The jackknife method produces a collection of representative training sets that maintain the 75/25 distribution between the two classification groups.

A technique that is similar to *N*-fold cross validation or bootstrapping is developed to create and evaluate different stratified training sets. The "modified bootstrap" is a mixture of the jackknife methodology which guarantees that every member of the population

will be used in a hold-out sample and the bootstrap which enables multiple random samples to be held-out simultaneously, thus creating a smaller training set. The size of the training sets is governed by the quantity of samples from the smallest classification group. As an example, for the loan evaluation data set, a 50/50 stratified training set would only permit 14 or 15 (depending on the sample item to be held out) members of the larger 46 member group. Each member of the smaller group is held out a single time, similar to the jackknife, with training set elements from the larger group randomly selected to satisfy the distribution requirements. This process is repeated until all elements have served as an out-of-sample test item a single time. Due to the reduction in the quantity of the larger group members required for the training set, multiple item tests may be performed on a single neural network model (derived from a single training set), but care must be taken not to duplicate the test evaluation of any population member so as not to introduce any artifacts.

Using the "modified bootstrap" method just described, training sets that satisfy a stratified distribution of 60/40 and 50/50 are instantiated and used in determining the effects of stratification of the training sets. A possible side effect from using the modified bootstrap method is that the size of the training set is constrained by the quantity of samples in the smallest classification group, such as a maximum training set size of 28 to 30 samples for the 50/50 stratified distribution training set. Since fewer members of the known population are present in the training set, a negative generalization bias may ensue (Smith 1993). Results for the three different training set distributions are presented in the next section and even if a training bias is introduced through the modified bootstrap method, the stratified training sets still far outperform the representative training set.

RESULTS AND DISCUSSION

Those neural networks trained using the BP training algorithm appear to have become trapped in a local minima and produced classification predictions for all members of the population as belonging to the full repayment group. This is similar to the problem encountered by Hu et al.'s (1996) logistic regression model. The BP neural network "learned" to maximize it's performance by classifying all new data samples as belonging to the group that has the highest probability of membership. While this did produce an overall prediction accuracy of 75.41 percent, the fact that no defaulting loan applicants are identified carries a large cost to the lending institution for the classification errors and hence the BP algorithm is not used further (Berardi & Zhang 1999).

The results of the LVQ neural network models for each of the three different training set distributions, one representative and two stratified, is presented in Table 1. It should be noted that because of the jackknife and "modified bootstrap" approaches, the classification accuracy for the LVQ neural networks are for all 61 members of the population and generated from up to 61 different training sets (for the representative 75/25 training group using the jackknife).

The smaller group 2 classifications appears to mirror the probability of membership in the training set until the equally distributed 50/50 stratified training set is used and then it jumps to well over 50 percent classification accuracy. As a further test of this statement, stratified training sets are constructed using the modified bootstrap approach with a group distribution of 65/35. The newly constructed training sets are then used to build neural network classification models that are subsequently used to evaluate only the loan default group 2 test cases. This experiment yields a classification accuracy of the loan default, group 2, members of 37.5 percent.

While the classification accuracy of the smaller loan default group members continues to rise as the probability of group membership approaches equality across the two groups, a corresponding decrease in the classification accuracy of the larger full repayment group members does not occur. This result is unexpected since the much heavier emphasis in the training set for membership in the full repayment group, group 1, should bias the classification results of the associated neural network model accordingly.

Table 1: LVQ neural network classification performance for 3 different training set distributions

Training Set Distribution	Repayment (Group 1) Classifications (N = 46)	Default (Group 2) Classifications (N = 15)	Overall Classification Accuracy
Representative 75/25	33 (71.74 %)	4 (26.67 %)	60.66 %
Stratified 60/40	36 (78.26 %)	6 (40.00 %)	68.85 %
Stratified 50/50	37 (80.43 %)	10 (66.67 %)	77.05 %

To demonstrate, the 61 members of the loan classification population are divided into two distinct groups: one that contains only the 46 members of group 1, the full repayment group, and the other that contains only the 15 members of group 2, the loan default group. This produces two populations that have membership probabilities of 100/0 and 0/100 respectively. A jackknife procedure is used to build LVQ trained neural network models to predict the group membership of these two populations, using the same architecture previously described, with two output categorical variables. The resulting neural network models both produce 100 percent accuracy in classifying all test cases as belonging to the corresponding group. Similar to what happened to the BP neural network mentioned at the beginning of this section, these two monotype populations demonstrate that very large biases (maximum in this case) can produce corresponding probabilistic (certainty) biases in the output of a neural network.

The LVQ neural networks, unlike the BP neural network, are trying to accommodate the presence of two groups in the population. The difficulty arises in that the representative group does not provide enough information for the LVQ neural network to adequately distinguish between the two-group membership criteria. Even though the number of group 2 (loan default) members in the training set stays the same (as in the representative set) in the stratified training sets, the relative importance of the group 2 members increases to 40 and 50 percent of the population, as recognized by the training set. The more balanced representation prevents the larger group from dominating the training and enables the LVQ neural network to more adequately determine the membership criteria for all of the classification groups. This balanced knowledge from the 50/50 stratified training set is what enables the neural network to improve its classification accuracy for both groups in the classification problem.

SUMMARY

The research presented in this article demonstrates that neural network solutions to two-group classification problems with small data sets are maximized when the training sets used to build the neural network classification models are stratified to contain equal membership from each group. This is particularly important for those real-world problems that have unequal membership probabilities. These findings may help explain some of the less than optimal results from previous research (Hu et al. 1996) with neural networks that utilize representative training samples from unequally distributed populations. For the loan repayment classification problem presented in the article, moving from a 75/25 representative training set to a 50/50 stratified training set increased the classification accuracy of the neural network model by over 16 percent.

A modified bootstrap method is described to enable the maximum use of population members in training sets, while still maintaining a stratified balance between the group memberships in the training set. Additional research is needed to extend these results to *N*-group classification problems, where *N* is greater than two, with unequal probabilities of membership in the various groups.

REFERENCES

Full references are available on request from the first author.