OUTLIER REMOVAL AND PATTERN REDUCTION TECHNIQUES USING ERROR SIGN TESTING

P.M. Wong[1] and T.D. Gedeon[2]

[1] Centre for Petroleum Engineering
Email: patrick@iantag.petrol.unsw.edu.au
[2] School of Computer Science and Engineering
Email: tom@cse.unsw.edu.au
The University of New South Wales, Sydney, NSW 2052, Australia

Abstract

Most real data sets are noisy in nature. A number of techniques for outlier detection and removal have been proposed recently. This paper reviews the issues on the effects of outliers in training a back-propagation neural network. A new technique to detect and remove the outliers using error sign testing is proposed, and the method is applied to classify a set of wireline log data obtained from an oil well in Australia into different rock types. The results on this example show that the reduced training set improves generalisation on a validation test data set.

Model Assumptions

In this paper, we will assume a multi-layer feed-forward network trained using back-propagation, and will use the general expression "neural network" to mean such a network. All connections are from units in one level to units in the next level, with no lateral, backward or multi-layer connections. Each unit is connected to each unit in the preceding layer by a simple weighted link. The network is trained using a training set of input patterns with desired outputs, using the back-propagation of error measures. The network is tested using a validation set of patterns which are never seen by the network during training and thus can provide a good measure of the generalisation capabilities of the network. The separation of the total set of patterns into training and test sets is generally at random to avoid introducing experimenter bias.

By back-propagation we mean the general concept of developing the error gradient with respect to the weights, and not restricted to the original gradient descent method. In the examples we use here, we have used the basic sigmoid logistic activation function, $y = (1 + e^{-x})^{-1}$, though again this is not germane to the substance of our results.

Introduction

Neural networks are increasingly popular for solving pattern recognition, classification and mapping problems. One of the important features of such networks is their ability to learn and generalise from examples. Whilst neural network methods are capable of solving complex problems, three major research areas have techniques which being actively developed to compensate for the deficiencies of back-propagation networks. These areas include the optimisation of network topology (Gedeon and Harris, 1992; Bose and Garga, 1993), the development of fast learning algorithms (Barnard, 1992; Scalero and Tepedelenlioglu, 1992; Zhang and Wang, 1992) and the design of the training patterns structure (Wann et. al., 1990; Mehrotra et. al., 1991; Gedeon and Bowden, 1992; Slade and Gedeon, 1993). This paper addresses the last research area which attempts to use appropriate training patterns in order to reduce the amount of training time and improve generalisation ability.

Recent work has addressed the issue of using boundary samples in training the network. Wann et. al. (1990) used Hamming distances to define the boundary samples and also showed that using boundary samples were sufficient to achieve good performance in generalisation. Mehrotra et. al. (1991) provided some guidelines on the number of boundary samples which should be used in successful classification. These two papers, however, did not discuss the effects of the quality of the data in training the network. Most real data sets are not only complex and highly non-linear, but also noisy in nature. Noisy data containing outliers can generally have two effects on training neural networks. The first effect is to slow down the learning of the good data, and the second is to overfit the outliers.

Gedeon and Bowden (1992) used a new technique in simplifying the error surface in the pattern space by eliminating half of the given training samples. The method was based on the errors associated with each pattern at 1000 epochs. The errors were sorted in ascending order and every second sample of the sorted patterns then selected, and a half-sized training set was obtained. The new training set was then used for a further training stage. The technique was demonstrated by an example study. The results showed that training using the reduced training set, in some cases performed better (i.e. produced better generalisation) on the validation set compared to training on the whole training set. Strictly speaking, the method used in that study was not an outlier removal technique since only half of the noisy data (as well as the good data) were eliminated from the original training data set.

Slade and Gedeon (1993) reviewed a number of outlier detection and removal techniques, including Absolute Criterion, Least Median Squares and Least Trimmed Squares methods. They also proposed a technique based on the error distribution of the training patterns at intervals of 50 epochs. All of these techniques were based on analysis of the error values and their distribution over the training patterns at a specific epoch, and the reduction of patterns was then performed.

A new outlier detection and removal technique based on the change of error during training, will be described in the next section. Unlike Gedeon and Bowden (1992) and Slade and Gedeon (1993), the proposed technique examines the "quality" of each pattern in the training set within a certain number of epochs before outliers start to affect the performance of the learning process. An example study using wireline logs and core data from an oil well in a real reservoir will be used to demonstrate the proposed method.

Error Sign Testing

During the training phase each input vector of the training set is presented to the network and an output vector is obtained. An error vector is then calculated by taking the difference between the desired and the output vectors. The magnitude of the error vector can be used as a measure of how easily the input pattern is learnt in a given epoch, the lower the value the easier to learn. For successful learning of a good pattern, the error magnitude for a particular pattern in the n^{th} epoch of batch training, say E_n, should be smaller than the previous one, E_{n-1}. Therefore, counting the number of negative signs of the expression E_n-E_{n-1} for K epochs can be used to define a *good* pattern. In order words, a pattern with a large percentage of negative signs will be a good pattern. Similarly, a small percentage of negative error signs (or large percentage of positive error signs) will be a noisy pattern or outlier. This method of detecting outliers is the Error Sign Testing (EST) method.

The value of K can be determined by the root-mean-square-error (rmse) of all the training patterns presented to the network. In most applications, the rmse starts from a high value and drops very quickly in a small number of epochs. In this example study, K was determined at an epoch when oscillation of rmse began, say 200. This is due to the presence of the outliers, and hence the learning of good patterns slows down and the network starts to overfit the outliers.

After K epochs, the percentage of negative signs of each pattern is calculated. In this study, a pattern which has less than 10% of the negative error signs is defined as an outlier. The outliers are then removed and further training is performed using a reduced pattern size. A validation data set is also used to test the performance of the trained network, but is not used for training, nor are any outliers removed from this test set.

Example Study

The technique we proposed for pattern reduction was used on a data set from the wireline logs and core data in an oil well located at the North West Shelf, Australia. The data set consisted of 231 points along the well with 4 wireline log measurements. The log variables in this well included deep resistivity (RLLD), bulk density (RHOB), sonic travel time (DT) and gamma ray (GR). The rock type of each of the points was identified through the careful examination of the core sample taken at that location. In this study, five dominant rock types were found and were then named as rock types 1 to 5 for simplicity purposes. All the log measurements were normalised in the range of 0 to 1. A cross-plot of RHOB versus DT is shown in Figure 1. The core sampling process is not usually done on every well and the predictions of rock types must then rely on the available log data. Neural network methods are increasingly popular in this area, and some recent applications in well log analysis can be referred to Rogers et. al. (1992) and Wong et. al. (1994).

Figure 1. Relationships of RHOB and DT for different rock types

In this study, the four log measurements and the corresponding rock types were used as the input patterns and the desired output patterns respectively. The validation data set was constructed using every third point along the well and hence the size of this data set was 77 patterns (i.e. one-third of the original). The initial training set was then formed by the remaining 154 patterns.

The network architecture used has four input units corresponding to each of the log variables, and four units in the hidden layer. Each rock type was represented by one output unit, and hence five output units were used. The classification was considered to be correct if the outputs of the network were within 0.1 of the desired value of 0 or 1. Bias nodes were also included in the hidden and output layers. The learning rate and the momentum were both set to 0.1. The network configuration used produced acceptable predictions of the rock type. By acceptable, we meant the results from neural network produced a consistent improvement compared to those obtained from the standard statistical techniques (see later sections).

Experiments and Results

The log measurements usually contain noisy data, especially in a lithologically complex reservoir. However, the outliers are not easily recognised in most of the cases (refer to Figure 1). Note that this does not include potentially unbounded outliers, which are clearly markedly different to the normal data and are excluded using standard methods. This exclusion of markedly different outliers is necessary for our method, as it is specialised to the location of any remaining outliers, and the sum of squares error term is not sufficiently sensitive to compensate for very extreme values which could pull the network towards them and still reduce the overall squared error. In this study, the proposed error sign testing method was used to detect and remove any remaining outliers. Four experiments with different training sets were trained and the performance of each set (i.e. classification accuracy) was evaluated using the validation test data. The structure of each training set is described as follow:

(1) whole data set - this data set consisted of the original 154 data which was used as a control experiment.

(2) half of the whole data set - this data set was selected based on the percentage of negative error signs during training using each of the 154 original patterns. The patterns were then sorted in ascending order and the new training set was formed using every second sorted pattern. The aim of this training set was to remove half of the good and noisy data.

(3) clean data set - the outliers, defined by the EST method, were removed from the original data set. After doing this, the data set was then considered to be "clean".

(4) half of clean data set using error sign testing method - this data set was formed using the clean data set in (3) followed by the half reduction method using the EST method as in (2). This experiment was designed to further simplify the set of patterns in (3).

Each of the above experiments was performed with 10 different sets of initial weights, and was trained for 10,000 epochs. In each run, the same initial weights are used for training the original data set (1), and the corresponding sets of reduced patterns (2), (3) and (4). This was done in order to minimise the effects of the initial random functionality of the network unit weights. The validation set was also used to record the highest clasification accuracy (in percentage) during the training phase.

The results of the comparison study are tabulated in Table 1. Note that the classification accuracy shown are the maximum values of the number of runs done and can be from different runs. The average training times (measured in number of epochs) required to achieve the maximum accuracy in each experiment are also shown. The success of the method is shown by the high classification accuracy on the validation set. Statistical evaluation of the same data using discriminant analysis was also done on the whole training set. The classification accuracy was only 57% on the validation data. Therefore, the results using neural network methods showed better performance in generalisation.

Training Set	Average Training Time (epochs)	Classification Accuracy (%)
(1) Whole Data Set	3500	66
(2) Half of Whole Data Set	3400	64
(3) Clean Data Set	2900	66
(4) Half of Clean Data Set	1400	68

Table 1: Performance of Different Training Sets on the Validation Data.

The results on this example also showed that a larger set of training patterns did not necessarily improve the generalisation ability of the trained network. Pattern reduction techniques aim to simplify the error surface of the pattern space, however simple elimination of half of original training patterns did not show significant improvement. This is most likely due to the presence of half of the original outliers in the remaining training data set, and these outliers may affect the process of error surface simplification. When the outliers were defined and removed from the training set, the same classification accuracy was obtained compared to the whole data set. Hence, the removal of outliers did not seem to improve generalisation, however the clean data set took less training time to achieve the same results as obtained from the original data set. Therefore removing outliers in the training set does reduce the amount of training time.

The clean data set was also reduced to half its size and the results showed better performance in generalisation. This improvement was probably due to the simplification of the error surface in the pattern space. Figure 2 shows the outliers identified by the EST method in the run with maximum classification accuracy. In this case, 12 outliers (out of 154 data) were found in rock types "2" and "3", and they were coded as "2x" and "3x". As displayed in Figure 2, the outliers defined tend to lie within the clusters of other rock types, and therefore including these data in the training set will significantly increase the training time, and overfitting of these outliers will also occur.

Figure 1. Outliers (coded as "2x" and "3x") identified by the Error Sign Testing Method

Finally, it is important to mention that the neural network method in this study performed better than the standard statistical approach. However, the classification accuracy obtained from discriminant analysis can be used to provide a minimum baseline level of accuracy for the neural network method to achieve.

Conclusions

A new technique, called the error sign testing method, can be used to detect and remove outliers in a noisy training data set. The method was applied to a real data set obtained from an oil well. From a series of experiments, the following conclusions can be drawn:

(1) Removal of outliers in the training set does not necessarily improve performance in generalisation over a noisy validation set. Performance is of course much better on a clean validation set.

(2) Pattern reduction of a clean or noise-free training set may improve generalisation, due to the simplification of error surface in the pattern space.

(3) Pattern reduction also has an advantage in speeding up the training time, as the number of epochs decreases, as well as the time gained by reducing the number of patterns in the training set, and hence the length of each epoch.

References

Barnard, E., 1992, "Optimisation of Training Neural Nets", *IEEE Trans. Neural Networks*, 3, p. 232-240.

Bose, N.K., and Garga, A.K., 1993, "Neural Network Design using Voronoi Diagrams", *IEEE Trans. Neural Networks*, 4, p. 778-787.

Gedeon, T.D., and Bowden, T.G., 1992, "Heuristic Pattern Reduction", *IJCNN*, Beijing, p. 449-453.

Gedeon, T.D., and Harris, D., 1992, "Hidden Units in a Plateau", *Proc. 1st Int. Joint Conf. on Intelligent Systems*, p. 391-395.

Mehrotra, K.G., Mohan, C.K., and Ranka, S., 1991, "Bounds on the Number of Samples Needed for Neural Learning", *IEEE Trans. Neural Networks*, 2, p. 548-558.

Rogers, S.J., Fang, J.H., Karr, C.L., and Stanley, D.A., 1992, "Determination of Lithology from Well Logs using a Neural Network", *AAPG Bulletin*, 76, p. 731-739.

Scalero, R.S., and Tepedelenlioglu, N., 1992, "A Fast New Algorithm for Training Feedforword Neural Networks", *IEEE Trans. Signal Processing*, 40, p. 202-210.

Slade, P., and Gedeon, T.D., 1993, "Bimodal Distribution Removal", in Mira, J, Cabestany, J and Prieto, A, *New Trends in Neural Computation*, pp. 249-254, Springer Verlag, Lecture Notes in Computer Science, vol. 686.

Wann, M., Hediger, T., and Greenbaun, N.N., 1990, "The Influence of Training Sets on Generalisation in Feed-Forward Neural Networks", *IJCNN*, Part 3, San Diego, p. 137-142.

Wong, P.M., Gedeon, T.D., and Taggart, I.J., 1994, "An Improved Technique in Porosity Prediction: A Neural Network Approach", *IEEE Trans. Geoscience and Remote Sensing*, (in press).

Zhang, Y.X., and Wang, D.X., 1992, "Fast Learning in a Backpropagation algorithm with a Sine-type Thresholding Function", *Applied Optics*, 31, p. 2414-2416.