XIII ICSMFE, 1994, New Delhi, India / XIII CIMSTF, 1994, New Delhi, Inde

DATA ANALYSIS AND MODELING USING AN ARTIFICIAL NEURAL NETWORK

ANALYSE ET MODELISATION DE DONNEES PAR UN RESEAU DE NEURONES ARTIFICIELS

Girish Agrawal[1] *J. David Frost*[2] *Jean-Lou A. Chameau*[3]

[1]Visiting Assistant Professor, School of Civil Engineering, Purdue University, West Lafayette, Indiana, U.S.A.
[2]Associate Professor, School of Civil Engineering, Georgia Institute of Technology, Atlanta, Georgia, U.S.A.
[3]Professor and Director, School of Civil Engineering, Georgia Institute of Technology, Atlanta, Georgia, U.S.A.

SYNOPSIS: The results presented are from the authors' ongoing research into applications of artificial neural networks and fuzzy theory to geotechnical engineering. The focus of this paper is on the use of artificial, or more properly, computational neural networks as tools—from a practical engineering perspective—for modeling and data analysis. Neural networks can be used to estimate functions from sample data as can statistical techniques. The major difference is that statistical approaches require guessing as to the functional dependency of outputs on inputs whereas neural systems do not require articulation of any such physical or mathematical model. Neural networks can recognize ill-defined patterns without an explicit set of rules; they adaptively infer it from sample data. Published data is used to demonstrate the comparative ease with which a neural network can be trained to perform a variety of tasks involving modeling and prediction.

INTRODUCTION

Traditional modeling and analysis teaches that natural phenomena can be teased apart to reveal underlying principles, and that those principles explain the behavior of everything from individual particles to large masses of soil. This reductionist approach is not the only valid way of investigating the world. One does not always have the luxury of an available model which accurately reflects real phenomena; this is especially true when the material is natural soils.

A substantial part of geotechnical design relies upon observation of the behavior of geotechnical structures under similar conditions. Geotechnical engineers apply experience and judgement in their evaluations in an attempt to characterize the parameters of interest for the entire soil mass under consideration by extrapolating from a few observations or from laboratory testing. The inherent empiricism in many design procedures needs to be clearly appreciated.

Additionally, despite the great improvement in techniques for modeling the behavior of soils, there are other difficulties. One of the greatest lies in the inherent variability of geotechnical data itself, which is large even in nominally uniform soil masses. This variation causes a scatter in the results which is difficult to correct. Applying too many refinements and corrections only serves to make analysis complicated and may be of doubtful value.

Due to the complexity involved, probabilistic modeling requires significant data manipulation and placement of restrictive constraints on the type and distribution of data—in order to meet the statistical assumptions of the model being employed. Parameter free modeling may work better but the standard procedures are difficult to use, time consuming and cumbersome, and do not always result in tractable problems.

It is reasonable to state that expert geotechnical engineers learn to detect patterns of behavior and match them with behavior they have seen earlier. The same is true in using statistical correlations and/or models to characterize the relationship of one or more parameter of interest to that of others. This is essentially a problem of pattern recognition where a pattern can be thought of as an example or model, *i.e.*, something which can be copied.

The task may then be stated as an attempt to develop a set of automated procedures which perform the following two-phase task:

(a) Training or learning phase — learn to classify correctly when presented with a series of data sets with known classifications; and
(b) Generalization or prediction phase — when presented with a previously unseen set of data, will indicate the class to which the new data set most likely belongs.

In other words, the system should have the ability to recognize a pattern or patterns—based solely on given data (observational or measurements). In mathematical terms, this entails the learning of a mapping between an input space and an output space, such that the mapping generalizes properly when applied to new inputs.

Artificial neural networks and fuzzy sets

An artificial neural network consists of a group of identical neurons—processing elements, loosely based on biological neurons, which receive an input and produce an output using a non-linear transfer function—arranged in interconnected layers. Among the many possible neural architectures and training methods described in the literature, layered feed-forward networks with supervised learning using backpropagation have been selected as those best suited to data analysis tasks of the type most common in geotechnics (Agrawal, 1992). In error backpropagation, learning is accomplished by an iterative procedure where the network steps through weight space to minimize a given error function. Such a net can learn a function given example data from specific occurrences. The standard backpropagation algorithm has been modified and implemented in a general computing environment. *GIRJAAL*, the neural-fuzzy software package developed for this work, is used in all cases presented in the subsequent. The focus of this paper is on applications and thus no further details of network architecture and training methods are given. Detailed explanations of the methodology used and the operational algorithm can be found in Agrawal (1992).

Before presenting the application examples, an additional comment is appropriate. In classification/recognition tasks, separation into distinct classes is not always possible, or indeed desirable. Available tools however, such as the standard classification methods of discriminant analysis and logistic regression, do not allow for classification with incomplete matches, *i.e.*, a given object is not allowed to belong to more than one class and is classified as a member of a particular group with an associated probability of misclassification. There is no concept of gradual change of class boundaries; where an object may share the properties of more than one class. Further, it is necessary to account for the ambiguity (as distinct from randomness) inherent in geotechnical data analysis. Probabilistic methods deal only with the uncertainty reflected by randomness. Both of the preceding shortcomings are properly handled by the use of fuzzy set theory. Kosko (1992) has shown that fuzzy sets and neural networks share the same state space so that a neural network automatically implements fuzzy classification.

APPLICATION EXAMPLES

The two examples, chosen to demonstrate the ability of the procedures to handle different problems, are:

(a) parameter estimation — interpreting the network output in terms of continuous variables wherein drained strength parameters are estimated without the need for the user to articulate a model; and
(b) modeling — grain-size distribution curves are modeled by showing that the network can estimate D_{50} based on given values of D_{10}, D_{90}, and % fines content in the sample.

Model free estimation of drained strength parameters

The available input data consists of 22 sets of results from laboratory testing of field compacted samples of silty clay and are taken from Altschaeffl *et al.* (1987). Each data-vector consists of four measurements—dry density, effective stress strength angle and intercept, and water content. The problem can be defined as follows:

- Develop a model to estimate drained strength parameters, ϕ' and c', from measured dry density (γ_d) and moisture content (W_c).

It is important to recognize some of the contrasts between a neural network based procedure and regression correlations to model ϕ' and c'. The two strength parameters describe different aspects of the same soil property, and hence should be seen as two aspects of the same pattern. In other words, a particular pair of the input parameters, γ_d and W_c, should produce a particular ϕ' - c' regression pair, but a separate statistical correlation is needed to model each of them. On the other hand, a single neural network is able to model both ϕ' and c'. Additionally, developing a good regression correlation involves trying out various parameter combinations until a good fit is achieved. This is a trial and error procedure which can become very time consuming. Further, assumptions have to be made concerning the interrelationships between various parameters and the distributions of these parameters. No such assumptions are required for developing a neural network based model. Selecting groups of parameters to be used is also not needed with a network and this in turn obviates the need for any preprocessing of the data. The data is presented, as is, to the network which automatically learns to model the input-output relationship (if any) by adapting its weights and biases. Once training is complete, the network can be used to estimate the drained strength parameters for a given pair of dry density and moisture content measurements.

Training is done using a 2-5-2 network (a three layer network with 5 hidden nodes and two each of input and output nodes). The available data is partitioned into odd and even numbered sets—based on the numbering in the available data tabulation. Two cases are considered: Case 1 — all odd numbered samples are used for training and the even numbered data are used to test the predictive ability of the trained network; and Case 2 — the network is trained using even numbered data and tested using odd numbered data. In

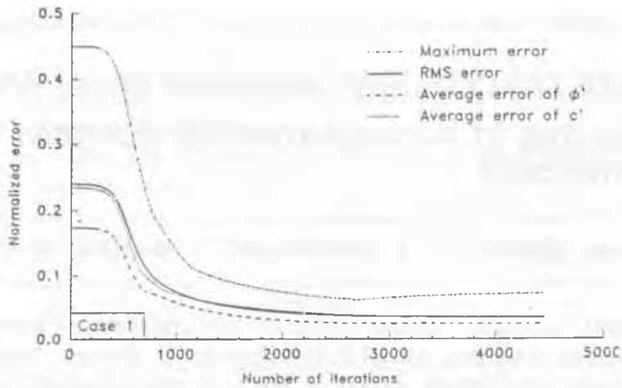

Fig. 1. Modeling ϕ' and c' — error during training

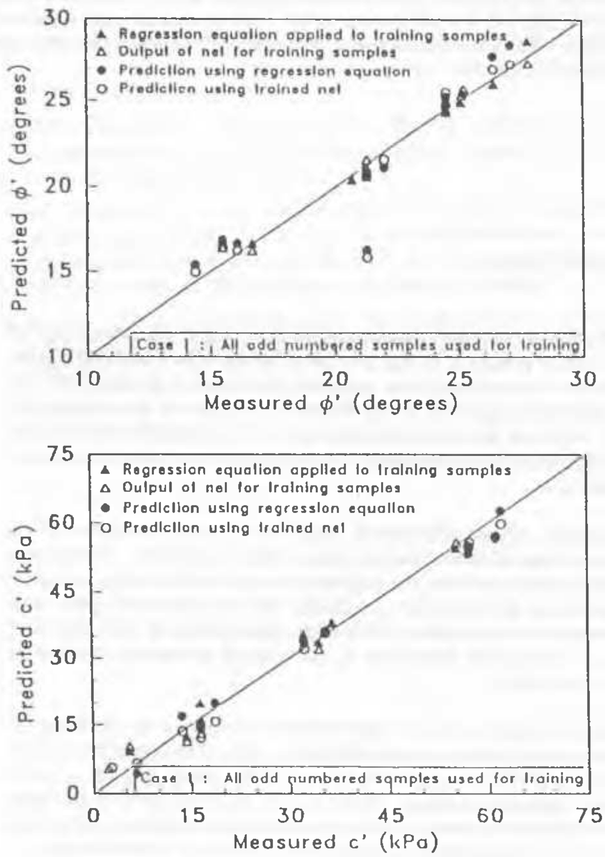

Fig. 2. Modeling ϕ' and c' — results of Case 1

each case, the network is trained using 11 sets of input-output data vectors and the trained network is used to predict ϕ' and c' values for 11 sets of data.

Figure 1 shows a typical error plot during training. The network converges to a minima in less than 1000 iterations. Further iteration does not reduce the error in any significant degree. Similar results are obtained with various combinations of starting network parameters. Figure 2 shows typical scatter plots of predicted *versus* measured values for ϕ' and c'. The rms error for the prediction of the strength intercept is 2.4 kPa and 3.0 kPa for Case 1 and Case 2, respectively. The rms error for the prediction of the effective-stress

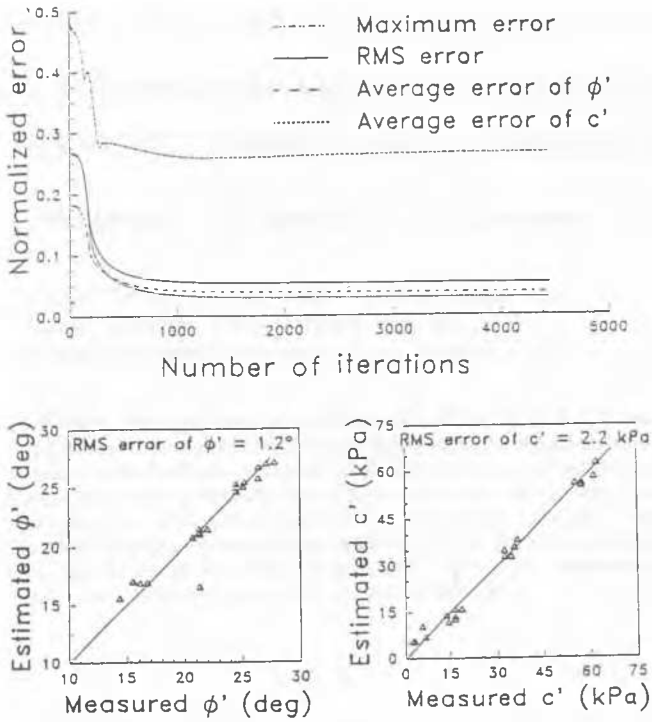

Fig. 3. Modeling ϕ' and c' — results with all data

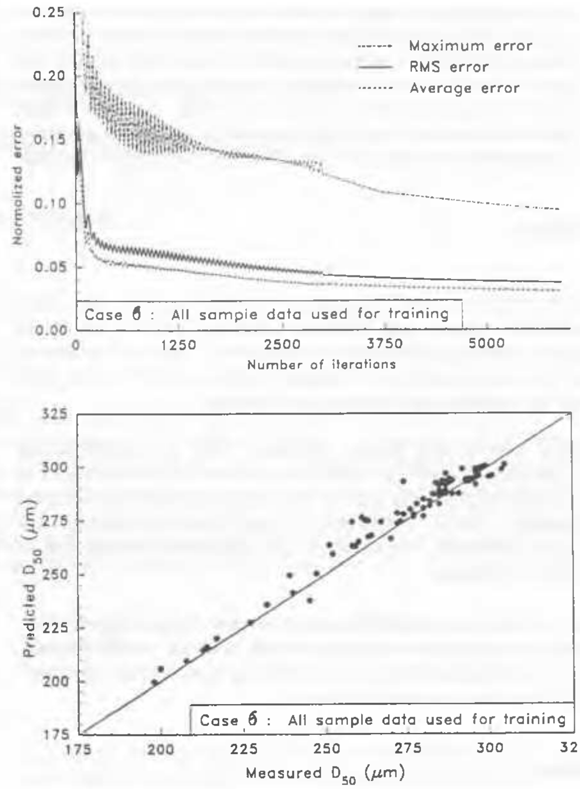

Fig. 4. Estimating D_{50} — results with all data

strength angle, ϕ', seems high (1.74°) in Case 1, but a look at the corresponding scatter plots shows that most of it is contributed by one outlier (Data pair no. 20 in the original set) and if its contribution is neglected, the rms error reduces to approximately 0.65°. That one sample is causing the high error could also be deduced by studying the error tracks as in Figure 1. The maximum error track for ϕ' falls to a low level when no. 20 is not part of the training set (Case 1), whereas the corresponding error track in Case 2 settles down at a relatively higher level. The results of statistical analyses also show that the error in estimating ϕ' for sample no. 20 contributes the most to the error.

The results of developing a correlation, for Case 1 data, using linear regression are also shown in Figure 2. The estimated (odd data) and predicted (even data) values for ϕ' and c' are nearly the same as those obtained using a neural network. The results of network training and prediction are marginally superior.

As the final step, the entire set of available data vectors is used for training another 2-5-2 network (Case 3). The starting network parameters are the same as in the preceding cases. The results are presented as Figure 3. The error plot of Figure 3 shows that the network converges to a minima within 500 iterations and the reduction in error with further iterations is marginal. Similar results were obtained with different combinations of starting weights and values for the gain and momentum parameters. This implies that there is a distinct pattern that the network is learning to recognize. Figure 3 also shows the scatter plots of estimated *vs.* measured values for ϕ' and c'. The rms error for the strength intercept is 2.2 kPa. Again, most of the strength angle rms error comes from the single outlier as discussed earlier and if its contribution is neglected, the rms error is less than 0.5°.

Estimating D_{50} based on D_{10}, D_{90} and % fines

Grain size distribution data for material dredged during construction of an underwater berm in the Beaufort Sea was also studied.

Each data-vector has four components (D_{90}, D_{10}, % fines and D_{50}). There is a total of 80 data-vectors available. The sample data is split into two groups for trial runs. Case 4 consists of the odd numbered data (40) being used for training the network and Case 5 uses the remaining 40 for training.

Training is done using a 3-5-1 network (a 3-layer network with 3 input nodes, 5 hidden nodes and 1 output node). In each of Cases 4 and 5, the network is trained using 40 sets of input-output data vectors and the trained network is used to predict D_{50} values for the remaining 40 sets of data.

The networks converge to a stable minima (as evidenced by the average and rms error tracks) in about 1500 iterations in both cases. Further iteration does not reduce the error in any significant degree. The estimated/predicted value of D_{50} has marginal difference from the measured value and, more importantly, the same estimated and predicted final value of D_{50} results for a particular data vector when it is and when it is not part of the training sample, respectively.

The results of these two trial runs provide confidence that the chosen network has been able to learn the pattern/model such that when given the values of D_{90}, D_{10} and % fines for a similar sand, the network can predict D_{50}.

As the final step (Case 6), the entire set of available data vectors is used for training a network. The starting network parameters are the same as those for Cases 4 and 5 above. The error plot of Figure 4 shows that the network starts moving towards a minima almost immediately and stabilizes within 3000 iterations. Iteration is continued until the maximum error falls below the specified normalized error of 0.1 (=15μm) at approximately 5800 iterations. The scatter plot of Figure 4 shows the predicted *vs.* measured values for D_{50}. The root of the mean squared error is 6μm and the average error is less than 5μm.

The purpose of the preceding example is not to argue for the need to be able to predict one grain size parameter based on measurements of others but more to demonstrate the feasibility of training an artificial neural network and then using the trained network to gain the ability to estimate one (or more) related parameters given the measurements of some of the others. Having a trained network available allows one to start with those parameters which are easier to measure and generate the values for those which are not easily measured.

CONCLUSIONS

The procedures that have been demonstrated are very simple to use. They are mathematically rigorous and the results indicate that they can be of enormous use in modeling ill-understood phenomena. They are also easy to incorporate into routine work and if regularly used can result in substantial improvement in modeling and forecasting soil behavior.

In comparison with various classic techniques (such as regression and discriminant analysis), the more powerful neural network formalisms have an advantage in automatically being able to learn and generalize from examples without knowledge of rules. Additionally, in cases where the available data involves vague parameters, the accuracy can significantly exceed that of classical statistical methods.

In addition, it is much more straightforward to develop a trained network than it is to develop an acceptable model using standard methods. Artificial neural networks are a promising technique for developing models since they learn adaptively with minimal external intervention.

REFERENCES

Agrawal, G. (1992). *Geotechnical Data Analysis: Prediction and Modeling Using a Neural-fuzzy Methodology*. Ph.D. dissertation, Purdue University, West Lafayette, Indiana.

Altschaeffl, A.G., Thevanayagam, S. and Agrawal, G. (1987). *Implementation Program to Improve Embankment Design and Performance with Indiana Soils*. FHWA/IN/JHRP-87/6, Joint Highway Research Project, Purdue University, West Lafayette, Indiana.

Kosko, B. (1992). *Neural Networks and Fuzzy Systems*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.