Machine Learning in Soil Classification

B. Bhattacharya
UNESCO-IHE Institute for Water Education
P.O. Box 3015, 2601 DA Delft,
The Netherlands
E-mail: b.bhattacharya@unesco-ihe.org

D. P. Solomatine
UNESCO-IHE Institute for Water Education
P.O. Box 3015, 2601 DA Delft,
The Netherlands
E-mail: d.solomatine@unesco-ihe.org

Abstract– **In a number of engineering problems, e.g. in geotechnics, petroleum engineering, etc., intervals of measured series data (signals) are to be attributed a class maintaining the constraint of contiguity and standard classification methods could be inadequate. Classification in this case needs involvement of an expert who observes the magnitude and trends of the signals in addition to any a priori information that might be available. In this paper an approach for automating this classification procedure is presented. Firstly, a segmentation algorithm is applied to segment the measured signals. Secondly, the salient features of these segments are extracted using boundary energy method. Based on the measured data and extracted features classifiers to assign classes to the segments are built; they employ Decision Trees, ANNs and Support Vector Machines. The methodology was tested for classifying sub-surface soil using measured data from Cone Penetration Testing and satisfactory results were obtained.**

I. Introduction

In a number of engineering problems there is a necessity to classify contiguous intervals (segments) of series data (signals). Series data has an additional index variable (distance or time) associated with each data value. Standard classification algorithms in these situations are often inadequate due to the additional contiguity constraint. Examples from the following domains can be mentioned: classification of sub-soil layers using Cone Penetration Testing [2] [7], well-log analysis in petroleum engineering [9], palaeoecology [4], etc. In these cases measurements are taken from a vertical bore or with a test apparatus which is pushed down the earth and it is required that the stratigraphical information is preserved in the classification. The problem is solved in two phases: firstly, a segmentation algorithm is used to cluster contiguous blocks of instances and secondly, these segments are classified by domain-experts.

We investigated the problem with a specific interest of automating classification of soil layers from measured data. In civil engineering it is a prerequisite to know the soil classes up to some depths prior to any construction. The direct method to identify the soil classes by drilling boreholes and testing soil samples is very expensive. A cheaper alternative is the so-called Cone Penetration Testing (CPT) which is one of the most popular soil investigation methods [2]. In CPT, a metallic cone is pushed into the soil and an indication of the in-situ soil strength is obtained by measuring the force needed to let it advance at a constant rate. A CPT recording is a quasi-continuous picture of the subsurface at the test location. It contains the vertical variations of the mechanical characteristics of the subsoil. These variations in turn indicate variations in geological layers and their properties. During a test, two primary signals are recorded: 1) the cone tip resistance stress (q_c), 2) the frictional stress (f_s) which is used to derive the more widely used friction ratio $R_f = f_s * 100/q_c$. Additionally, information is available from borehole drilling in the proximity of CPTs typically with the frequency of 1 borehole for 10 CPTs. Observing the variations of q_c and R_f (Fig.1) and using the nearby borehole information, an expert firstly segments the logs i.e., finds boundaries of layers (class boundaries), and secondly, using the domain knowledge assigns a soil class C_i to each segment (where $i = 1, 2, \ldots, I$ and I = number of classes).

In practice a manual segmentation and classification procedure is followed. This procedure requires expertise, and is expensive, time consuming, subjective and not completely reproducible. The challenge is to automate this procedure. In order to achieve this a new algorithm called

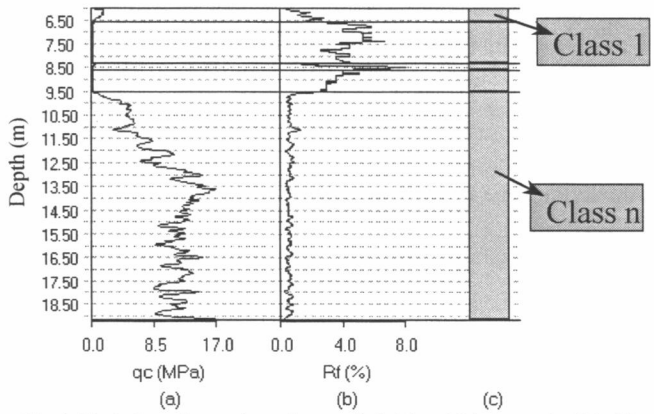


Fig. 1. Variation of cone tip resistance (q_c) (a) and friction ratio (R_f) (b) along depth of a Cone Penetration Testing which is used to segment the logs and assign classes to the segments (c).

0-7803-9048-2/05/$20.00 ©2005 IEEE

CONCC (CONstraint Clustering and Classification) was developed. It can be used in automatic classification with the constraint of contiguity and includes the following two steps:

- *Segmentation*: to find J segments of data from a single series data (e.g. CPT);
- *Classification*: to build a classifier to assign classes to these segments; it is built using measured data and extracted features from segments from a number of series data from a region and trained with classes labelled by experts.

Segmentation of series data has been addressed in [1], and is briefly reported here. This paper presents classification of the found segments using three Machine Learning (ML) methods: Decision Tress (DT), ANNs and Support Vector Machines (SVM). Application of ML in geostatistical problems is quite limited, some applications are reported in [1] [8] [10] [20].

II. Segmentation of Series Data

Segmentation can be defined as the clustering of series data where the constraint of contiguity has to be maintained. If the measured instances are labelled 1, 2,..., N according to depth, and J segments are sought , then J-1 'markers' are needed in some of the N-1 gaps between pairs of neighbouring measurements to produce J segments of contiguous block of data. The number of possible partitions in this case is considerably smaller than the unconstraint clustering case. Accordingly, most of the available algorithms employ the exhaustive or semi-exhaustive search within this reduced search space.

A segment g_j to be identified is defined as:

$$g_j = \{\{x_{1,1},..,x_{1,K}\},....,\{x_{l,1},..,x_{l,K}\}, ...\{x_{nj,1},..,x_{nj,K}\}\}_j \quad (1)$$

where $x_{l,k} \in \Re^n$ and represents the measured data;

$l = 1, 2, ..., n_j$, where n_j is the number of instances in segment g_j;

$k = 1, 2, ..., K$, where K is the number of dimensions (signals);

$j = 1, 2, ..., J$, where J is the number of segments.

After segmentation the classification problem is solved when segments are attributed to classes:

$$g_j \rightarrow C_i \quad (2)$$

Review of methods for segmentation (constraint clustering) is given in [15] and [4]. The reported methods partition a dataset into J groups by minimising the criteria of within-group sum of squared deviations from the segment mean, i.e. dispersion. The reported methods follow these approaches: i) a dynamic programming (computationally demanding) method where the solution is obtained recursively [6]; ii) a split moving window of fixed width that is moved along the data sequence and a marker is placed at points where a substantial change in some statistical criteria is observed [16].

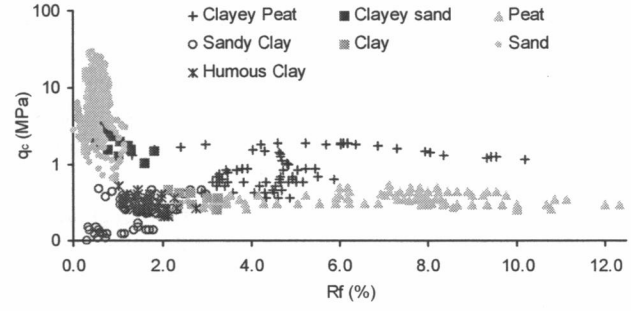
The CONCC algorithm addresses the shortcomings of the existing segmentation algorithms. It uses fuzzy logic to address the imprecision in the measured data and uses a particular threshold-based distance measure between instances and segment centres. Initial tests of the CONCC algorithm were performed and the satisfactory performance was achieved [1]. This paper presents the classification of the segments found by the CONCC algorithm. However, segments found by any other suitable segmentation algorithm, including the manual segmentation procedure of experts, can be classified using the presented approach as well.

III. Feature Extraction Using Boundary Energy

After the segments are found the following task is to assign classes to the segments. However, for different locations the mapping (2) could be different due to the spatial variability. When data from several test locations are combined an overlap of instances of different classes is usually observed (Fig. 2). The spatial variability has several reasons such as site-specific conditions, location and depth, measuring instruments, etc. To solve the problem it needs to be brought to a higher dimension by bringing in additional features so that a partition of the input space is possible by a classifier for assigning a class to that subset of inputs.

In engineering, experts use subjective criteria, so the automated classification methods should select appropriate features and be compatible with experts. During this research several experts in geology were interviewed and their manual classification procedures were recorded. It was observed that the experts assign high importance to the shape, in addition to the magnitude of the data. This led us to conclude that in automating the classification procedure experts' perception about the shape of signals needed to be parameterised.

Fig. 2. Scatters of the measured data from several CPTs in the cone tip



resistance (q_c) – friction ratio (R_f) space. Overlapping of different classes is discernible.

Shapes of signals can be represented using multi-scale transforms such as Fourier transform, *w*-representation using Marr wavelet and Morlet wavelet, Gabor transform etc. [3]. These representations can be used to derive shape measures using multi-scale energy methods such as boundary energy, multi-scale wavelet energy, etc. In this research the boundary energy has been used in parameterising the shape effects.

A. Boundary energy

Boundary energy is defined as the amount of energy required to modify the shape of a contour to its lowest energy level (a circle), with the same perimeter as the original object. The concept of boundary energy originated from the theory of elasticity and was first applied in biological shape characterisation [19]. Since then it has been widely used as a global shape measure for classification of a variety of shapes. Boundary energy is defined as follows:

$$B_{a,k} = \frac{1}{N}\sum_{n=0}^{N-1} c(a,n)^2 \qquad (3)$$

where $B_{a,k}$ denotes the boundary energy of a signal at scale a along the dimension k, c is the curvature at point n, $n = 1, 2,.., N$; and N is the number of discrete observations. A detailed description on boundary energy can be found in [3].

The multi-scale dimension of boundary energy is brought by successive low-pass filtering of a series data and by computing boundary energy of each of these filtered series data. Gaussian filter is the most common one and the value of 'sigma' in the Gaussian expression is changed gradually from low values to very high values. As a result the curvature of a series data is computed at different sigma values, i.e., at different analysing scales leading to a multi-scale representation of the series data. Such a multi-scale representation of curvature of a series data is called a *curvegram* [3]. Successive low-pass filtering of a series data with varying sigma values lead to a multi-scale characterisation of the energy contained in the series data. With increasing scales the small scale details of a series data vanishes and the most important features become prominent, which can be utilised in subsequent classification. Boundary energy has been successfully applied in biomedical engineering, neuroscience, and in general, in analysing images from a diverse domain (see, e.g., [14]) and has been recognised as an effective tool that can be used as a global shape measure.

It is evident that computation of boundary energy depends upon an accurate estimation of multi-scale curvature of a series data, which for a discrete signal is not an easy task. Commonplace computation techniques of curvature based on finite difference methods can lead to high errors, effectively thwarting the possibility of using boundary energy as a shape measure. In this regard a Fourier based curvature computation technique introduced in [3], which is much more accurate than the traditional curvature computation methods, has been used.

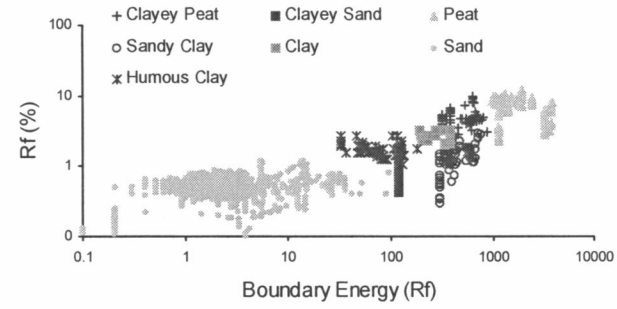


Fig. 3. Scatters of friction ratio (R_f) and the boundary energy of R_f using data from several Cone Penetration Tests. Several clusters of instances corresponding to classes are now more or less disjoint.

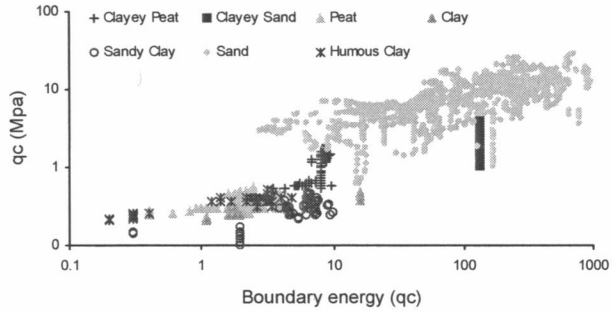


Fig. 4. Scatters of cone tip resistance (q_c) and the boundary energy of q_c using data from several Cone Penetration Tests. Several clusters of instances corresponding to classes are now more or less disjoint.

Fig. 2 shows instances from several segments taken from a number of CPTs ($K = 2$) where segments were labelled by the experts. From the fact that instances overlap it can be concluded that partitioning of the data to build a classifier in this space (q_c–R_f) may not be possible. Fig. 3 shows R_f and the corresponding boundary energy for these segments, whereas Fig. 4 shows q_c and the corresponding boundary energy for these segments. Clusters corresponding to classes are now more or less disjoint and hopefully present a much easier problem for a classifier.

IV. Classification

The overall classification scheme is shown in Fig. 5 with the particular reference to the classification of soil based on CPT data. It can be easily amenable to classification problems of other domains (e.g. petroleum engineering) as well. The training data may be created by segmenting several series data by involving an expert or by using the CONCC algorithm. The series data used in the training is then labelled by experts. Features from the labelled data (during training) or unlabelled data (during operation) are extracted using boundary energy (in the FE unit). Data preparation is carried out in the constructed feature space (in the DP unit). The classifier (CA) consists of two units. The pre-classification unit (PC) classifies each instance during the testing and operational use. The compaction unit

determines a single class (called a 'compact class') for a segment.

Classifiers are trained using DT, ANN and SVM to learn the following mapping from the labelled training data:

$$\{x_{l,1}, B_{l,1}, x_{l,2}, B_{l,2}, \ldots, x_{l,K}, B_{l,K}\} \rightarrow C_i \quad (4)$$

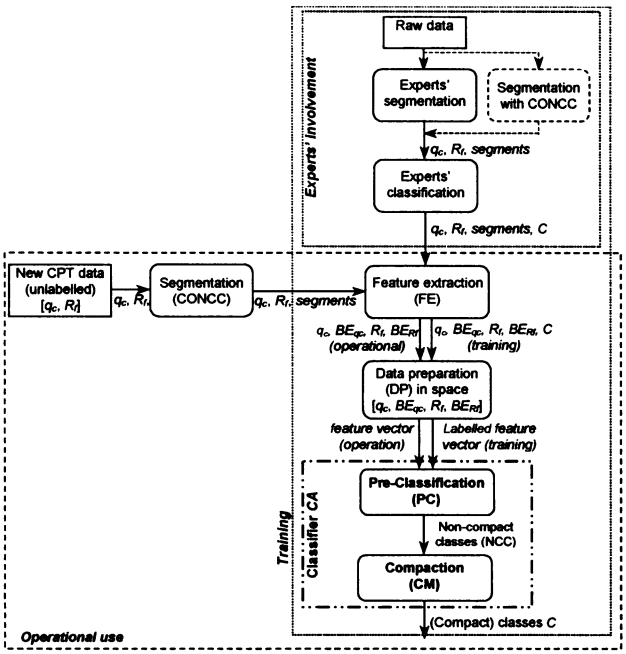


Fig. 5. The classification scheme during the training and operational phase. During training experts are involved in preparing the training data. Once the classifier CA is trained it replaces experts.

where l = index of instances, l = 1, 2,..., n_j and n_j is the number of instances in g_j;

K = number of signals (dimension);

$B_{l,k}$ = Boundary energy at point l in dimension k for input variable $x_{l,k}$.

The classifier (Eq. 4) learns to classify each instance. Finally the mapping (2) is ensured by compacting the classes of instances within a segment. This is required due to the reason that frequently within a series data there could be instances (measurements) that obviously belong to the class C_1 but are within the segment corresponding to another class C_2 and should be attributed to the class C_2. Such points (vectors) could be the result of noise or instrumentation errors. Such points may also appear, say in a CPT, due to the small inclusions of another soil class in the otherwise predominantly homogeneous soil layer. The heterogeneity of the natural environment (*e.g.*, soil) is the prime reason behind the presence of these mischievous measurements. Such points will be called *aliens*. For measurements in natural environments such as CPTs or well-logs of petroleum engineering the aliens are observed mostly in the proximity of another segment (i.e. another class). However, aliens can also be observed due to various other reasons such as instrument errors, etc. Experts ignore these aliens and pick up the general pattern of the segment. The compaction algorithm assigns weights to the classes determined by the PC unit. The weights (w_i) are computed as points on a Gaussian curve with zero mean and standard deviation = 2 (Fig. 6) and the compact class ($\overline{C}_p$) is determined by:

$$\overline{C}_p = w_i C_p / \sum_{i=1}^{n_j} w_i \quad (5)$$

where n_j = number of instances in a segment and C_p is the class for the ith instance.

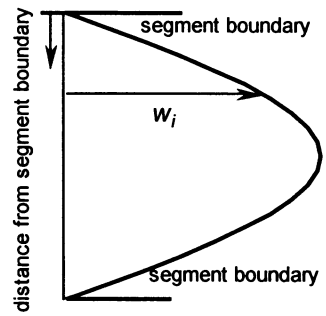


Fig. 6. The variation of weight w_i of an instance in relation to its position within a segment. The compaction algorithm of the classifier multiplies the class of each instance by w_i in order to determine a single class of a segment.

V. Application to Geotechnics

The above methodology was applied to classify soil on the basis of CPT data. The data that has been used to build the classifier is taken from the CPTs conducted at Nesselande, a residential zone under development near Rotterdam, The Netherlands. During 1996 to 2000 CPTs were conducted at 565 locations in an area of 2.5x3.5 km of Nesselande. At about 60 locations boreholes were drilled as borehole information is vital in determining the soil characteristics from CPTs. The Nesselande area is underlain by extensive peat and soft clay deposits. The Late Pleistocene and Holocene (sand) deposits predominantly make up the upper 10 m of soil in the Nesselande area. A detailed geological description of the area can be found in [17]. The study undertaken by the municipality of Rotterdam aims at finding the thickness of the soft sediments overlying (Pleistocene) sands, the presence and geometry of the sand bodies in the soft Holocene deposits, the hydrological contact of these sand-bodies with the (Pleistocene) sand, existence of peat within a few meters below the ground surface, engineering properties of the soil layers present in the subsurface, and their spatial distribution. Municipality of Rotterdam has the task of determining the soil classes of the subsurface of Nesselende using CPTs and borehole information.

Obtaining a proper dataset having more or less equal representation of all the classes as well as all the geographical regions of the site was a big problem. This is because the manual classification procedure of experts is time consuming and costly. Therefore, building a classifier using a minimum amount of data was considered. In consultation with the experts in total seven CPTs of the area were chosen. Four CPTs were considered for training and three CPTs for testing. Due to the scarcity of data no cross-validation dataset was chosen. Each CPT of the testing group is statistically comparable to a CPT of the training group. The number of instances in training and testing were 5150 and 2830 respectively whereas the number of segments in training and testing were 72 and 46 respectively.

A. The classification task

The requirement of details to be found in soil classes of an area varies as per the requirement of the end users. Often it may be enough to know the primary soil classes (sand, peat and clay for this area). Some other times the presence of the secondary soil classes also need to be determined. The determination of just the sandy or non-sandy soil types poses an important geotechnical task. This may be due to the reason that the end user wants to model the spatial extent of the sand layers, or for the purpose of determining the initial settlements from the proposed constructions, or to study the drainage characteristics to assess the migration of contaminants.

Based on the above discussions the following three classification problems were contemplated:

a) binary classification, where the task of the classifier is to determine whether the soil is sandy or not;
b) three-class classification, where the classifier has to identify the primary soil class only (i.e. sand, clay or peat);
c) seven-class classification, where the classifier has to determine the appropriate class from the set of seven classes observed in this area.

Classifiers using DT, ANN and SVM were built for the above-mentioned three classification tasks. The experiments were conducted with WEKA [18] for DT, Neurosolutions and NeuralMachine for MLP ANN, and WEKA and RHUL [13] for SVM.

B. Results and discussions

1) Binary classification: The results obtained with all the three methods were close to each other (Table 1). The instances with the sandy soil type were better classified by DT and ANN, whereas, the instances of the non-sandy soil type were better classified by SVM (with WEKA). It was noticed that almost all erroneously classified instances were located near the segment boundaries. Measurements near a segment boundary are often noisy and it can be concluded accordingly that the performance of the classifiers was excellent.

Finally, the segment classes were determined using the compaction algorithm. It was observed that all the segments were correctly classified by the three methods, i.e. a classification accuracy of 100% was reached.

2) Three-class classification: The three-class classification problem is comparatively more difficult than the binary classification problem mainly due to the large overlap of the instances of the clayey soil with that of the other two classes. The performance of the classifiers were comparable, however, the SVM-based classifier (with WEKA) gave slightly better results (Table 2), with ANN and DT following closely. All the methods provided more accurate classification of the instances from the sandy soil. For the clayey soil the SVM-based classifier was better than the others. All the methods provided poor results for the peaty soil. For each classifier if the correct class was not predicted then the predicted class was the geologically neighbouring class.

TABLE 1. CLASSIFICATION ACCURACY OF THE BINARY CLASSIFIERS (ON THE TEST DATASET)

Soil class	% of correctly classified instances			% of correctly classified segments		
	DT	ANN	SVM	DT	ANN	SVM
Sandy soil	99.3	100.0	96.6	100	100	100
Non-sandy soil	96.3	96.4	99.5	100	100	100
Total	97.6	98.0	97.8	100	100	100

TABLE 2. CLASSIFICATION ACCURACY OF THE THREE-CLASS CLASSIFIERS (ON THE TEST DATASET)

Soil class	% of correctly classified instances			% of correctly classified segments		
	DT	ANN	SVM	DT	ANN	SVM
Sand	100.0	99.8	98.5	100	100	100
Clay	82.9	96.5	98.5	90.9	90.9	100
Peat	58.7	60.5	58.7	66.7	66.7	66.7
Total	85.6	91.0	90.7	82.6	82.6	87.0

TABLE 3. CLASSIFICATION ACCURACY OF THE SEVEN-CLASS CLASSIFIERS (ON THE TEST DATASET)

Soil class	% of correctly classified instances			% of correctly classified segments		
	DT	ANN	SVM	DT	ANN	SVM
Silty sand	99.2	98.7	85.1	100	100	100
Clayey sand	83.3	70.8	54.2	100	100	100
Sandy clay	85.4	85.4	69.4	75	75	50
Clay	75.0	77.5	65.0	100	100	100
Humus clay	81.4	93.5	77.0	100	100	80
Clayey peat	65.1	64.6	49.7	57.1	42.9	28.6
Peat	74.1	84.0	65.4	100	100	50
Total	86.5	89.4	74.4	82.6	78.3	60.9

3) Seven-class classification: The results of the classifiers for the seven-class classification problem are shown in Table 3. The classifiers were very accurate for the silty sand and clayey sand. They were moderately accurate

for the sandy clay, clay and humous clay, but were poor for the clayey peat. The ANN model was much better than the others in classifying the instances of the humous clay and peat. The sandy segments were correctly classified in all occasions, the clayey segments were classified with the accuracy of 3 out of 4 segments. The DT gave the best results for the segments of the clayey peat, still the accuracy was not high. The SVM-based classifiers gave poor results; these results are, however, preliminary since no optimisation of the built SVMs (of regularisation constants and the kernels) was undertaken. For each classifier if the correct class was not predicted then the predicted class was the geologically neighbouring class.

VI. CONCLUSIONS

In this paper a method to classify series data where the constraint of contiguity has to be maintained is presented. Experiments were conducted to classify soil based on CPT data. The main conclusions are:

- due to spatial variability of the measured parameters classification based on the measured parameters was not possible. Additional features were extracted by paramterising experts' perception of the shape of a series data using boundary energy. This novel approach proved to be effective;
- the proposed classification scheme effectively mimics experts' classification procedure and automates the classification task;
- in the case-study of soil classification using CPT data the predictive accuracy of the classifiers on the test set even for the most complex problem was found to be high (83%). When the correct class was not predicted then the predicted class was a geologically neighbouring one. For many practical situations such accuracy of prediction was found to be sufficient by most experts and, if to allow for this error, the accuracy was 100%.

ACKNOWLEDGEMENT

Part of this work was performed in the framework of the projects "Predicting the structure of the subsurface: semi-automatic interpretation of cone penetration testing" and "Data mining, knowledge discovery and data-driven modelling" of the Delft Cluster research programme supported by the Dutch government.

REFERENCES

[1] B. Bhattacharya, and D.P. Solomatine, "An algorithm for clustering and classification of series data with constraint of contiguity", *Proc. 3rd Int. Conf. on Hybrid and Intelligent Systems*, Melbourne, Australia, 2003, pp. 489-498.

[2] A. Coerts, *Analysis of Static Cone Penetration Test Data for Subsurface Modelling - A Methodology* (PhD Thesis), Utrecht University, The Netherlands, 1996.

[3] L.F. Costa, and R.M. Cesar, *Shape Analysis and Classification: Theory and Practice*, Boca Raton, Florida: CRC Press, 2001.

[4] Gordon, A.D. "A survey of constrained classification", *Computational Statistics & Data Analysis*, vol. 21, pp. 17-29, 1996.

[5] S. Haykin, *Neural Networks: A Comprehensive Foundation*, New Jersey: Prentice Hall, 1999.

[6] D.M. Hawkins, and D.F. Merriam, "Optimal zonation of digitized sequential data", *Mathematical Geology*, vol. 5, pp. 389-395, 1973.

[7] G.P. Huijzer, *Quantitative Penetrostratigraphic Classification* (PhD Thesis), Free University of Amsterdam, The Netherlands, 1992.

[8] C.H. Juang, X.H. Huang, R.D. Holtz, and J.W. Chen, "Determining relative density of sands from CPT using fuzzy sets", *J. of Geotechnical Engineering*, vol. 122(1), pp. 1-6, 1996.

[9] M.G. Kerzner, *Image Processing in Well Log Analysis*, Dordrecht, The Netherlands: Reidel Pub., 1986.

[10] J. K. Kumar, M. Konno, and N. Yasuda, "Subsurface soil-geology interpolation using fuzzy neural network", *J. of Geotechnical and Geoenvironmental Engineering*, ASCE, vol. 126(7), pp. 632-639, 2000.

[11] NeuralMachine, *http://www.data-machine.com/*, 28.1.2005.

[12] NeuroSolutions, *http://www.nd.com/*, 28.1.2005.

[13] RHUL, Computer Learning Research Centre, Royal Holloway University of London, (*http://www.clrc.rhul.ac.uk/*, 26/1/2005)

[14] L.J. van Vliet, and P.W. Verbeeck, "Curvature and bending energy in digitised 2D and 3D images", in: K.A. Hogda, B. Braathen and K. Heia (Eds), *Proc. 8th Scandinavian Conf on Image Analysis*, Norway, 1993, vol. 2, pp. 1403-1410.

[15] K. Wagstaff, *Intelligent Clustering with Instance-Level Constraints* (PhD thesis), Cornell University, USA, 2002.

[16] R. Webster, "Optimally partitioning soil transects", *Journal of Soil Science*, vol. 29, pp. 388-402, 1978.

[17] H.J.T. Weerts, *Complex Confining Layers*, Utrecht University, The Netherlands, 1996.

[18] I.H. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 2000.

[19] I.T. Young, and T.W. Calvert, "An analysis technique for biological shape", *Information and Control*, vol. 25, pp 357-370, 1974.

[20] Z. Zhang, and M. T. Tumay, "Statistical to fuzzy approach toward CPT soil classification", *J of Geotechnical and Geoenvironmental Engineering*, vol. 125(3), pp. 179-186, 1999.