SOILS CLASSIFICATION SYSTEM ON THE BASIS OF DMT AND CPT DATA

Michał Kruk, Jarosław Kurek, Piotr Bilski, Oguz Akpolat and Simon Rabarijoely

Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences – SGGW, ul.
Nowoursynowska 159, 02-767, Warsaw, Poland
Michał Kruk michal_kruk@sggw.pl

Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences – SGGW, ul.
Nowoursynowska 159, 02-767, Warsaw, Poland
Jarosław Kurek jaroslaw_kurek@sggw.pl

Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences – SGGW, ul.
Nowoursynowska 159, 02-767, Warsaw, Poland
Piotr Bilski piotr_bilski@sggw.pl

Chemistry Department, Faculty of Science, Mugla University - Kotekli Campus, 48000
Mugla, Turkey
Oguz Akpolat oakpolat@gmail.com

Faculty of Engineering and Environmental Sciences, Warsaw University of Life Sciences – SGGW,
ul. Nowoursynowska 159, 02-767, Warsaw, Poland
Simon Rabarijoely simon_rabarijoely@sggw.pl

Abstract

The paper presents methods of the soils classification on the basis of DMT and CPT data obtained from Warsaw University of Life Sciences campus. The applied methods use the features selection such as Fast Correlation-Based Filter (FCBF), Fisher measure, Correlation-based Feature Selection (CFS) and the automatic classifiers in the form of Support Vector Machine (SVM) and k Nearest Neighbours to estimate the soil category. The starting point are the data from CPT and DMT probes in the form of m x n matrix.

The results of experiments have shown that the proposed numerical features of similar nature are strongly related with the soils types. The features collected from the DMT tests are: the A pressure, required to just begin to, move the membrane ("lift-off"), the B pressure, required to move the center of the membrane 1.1 mm against the soil, a third reading C ("closing pressure") can also optionally be taken by slowly deflating the membrane soon after B is reached, γ [kN/m^3], σ_{vo} [MPa], u [MPa]. These descriptors have been used as the diagnostic features forming inputs to the automatic classifier, which performs the final recognition of soil profiles. The average discrepancy rate between the score of our system and the human expert results, estimated on the basis of 625 measurements, is below 5%.

The proposed method appears to be useful for the automatic classification of the soil profiles. The obtained results have shown that the system is able to recognize seven different soil profiles with good statistical accuracy and good concordance with experts. This result gives good hope to apply the system for supporting and accelerating the geotechnical process of soil profiles measurement. Moreover the system allows to save the time of manual analysis of the data in comparison to the human expert assessment and to accelerate the research in this area.

Keywords: *Soil profiles, Soil classification, Artificial intelligence, SVM, Dilatometer test.*

1. Introduction

Supporting the geotechnical exploration by computer technologies is currently widely used. It may be helpful while obtaining, storing and interpreting measurement data. When the data are digitalized it is easier to process and analyze them using computers algorithms. This significantly shortens the duration of making decision and identification of the soil characteristics. The sophisticated computing algorithms still rely on the measurement devices providing information about the soil parameters at the particular depths (Marchetti 1980). The most traditional and reliable method is drilling boreholes in the ground to get soil samples. They are then analyzed in the laboratory. Although this is the most reliable method which gives results with high precision it has disadvantage – high cost and intrusiveness. The numerous boreholes change the structure of the soil and may damage the ground on which the building foundation is established. Therefore multiple approaches (such as probes and georadars) are introduced to decrease the impact of the boreholes on the collected information.

In the computer-based approach the algorithms analyse collected data automatically, producing the soil profile and its characteristics (such as indexes) as the output. Then they are analyzed by the human expert (geotechnical engineer), who can verify accuracy of the generated profile. The latter may be also verified by comparing it to the results of borehole drilling which is the most reliable method, often used as the reference for verifications.

Such algorithms can be further used to develop the automated soil profile generation system, which could classify the geotechnical layer based on the measured quantities at particular depths. Similar works were done before (Hashash et al. 2004, Shahin et al. 2005), but new approaches must be proposed.
The paper presents the computerised automatic system for soil profile generation. The data are obtained from Warsaw University of Life Sciences (WUoLS) campus, Warsaw, Poland. The data gathered for the computer algorithm are obtained using DMT and CPT probes. It was implemented in Matlab enviroment (Matlab 2014). The research presented here is a continuation of the experiments published before (Rabarijoely et al. 2007) , Rabarijoely and Bilski 2009), Kruk et al. 2014) and Kurek et. al. 2014). The expected result of the research is the automatic system for soil profile generation which accuracy can be similar or better than the human expert. In this paper we combine the feature selection methods with different classifiers and compare them to the boreholes drilling method.

2. The input data

The input data were taken at Warsaw University of Life Sciences (WUoLS) during the expansion of the university campus. Before new buildings could be established, throughout soil investigation had to be performed. Therefore multiple tests were conducted, using the DMT probes and traditional methods, i.e. drilling boreholes. The boreholes test as the most accurate could was treated as the reference method for the proposed approaches. The profiles and probing places are presented in Fig. 1. In this example on the basis of borehole drilling (OW-8, OW-11, OW-6), CPT probing (CPT-4) and DMT probing (DMT-2) the soil profiles were generated by the human expert manually. Our main task in this paper is to describe the automatic method which could do this automatically on the basis CPT and DMT probing (with low intrusiveness) only.

The data for the experiment were gathered from geotechnical investigations by CPT and DMT probes. 19 measurement were made in such way. The probes and the probing are fully described in our previous papers (Rabarijoely et al. 2007) , Rabarijoely and Bilski 2009), Kruk et al. 2014), Kurek et. al. 2014).

In the described experiment 625 records were obtained. They may be presented as 625x8 matrix where particular columns are referenced to:

1. Depth - the depth of measurement
2. γ [kN/m^3] - volumetric weight of soil
3. σ_{vo} [MPa] - total vertical stress
4. u [Mpa] - pore water pressure
5. A pressure - the pressure, required to just begin to, move the membrane ("lift-off")
6. B pressure - the pressure, required to move the center of the membrane 1.1 mm against the soil
7. C pressure - ("closing pressure") can be taken by slowly deflating the membrane soon after B is reached
8. soil class - soil class was obtained by the drilling boreholes which is our referenced method

Only a part of this numerical features will be input to learn and test the classifier. To evaluate and filter which of the numerical data can be useful we used Fisher measure to perform features selection

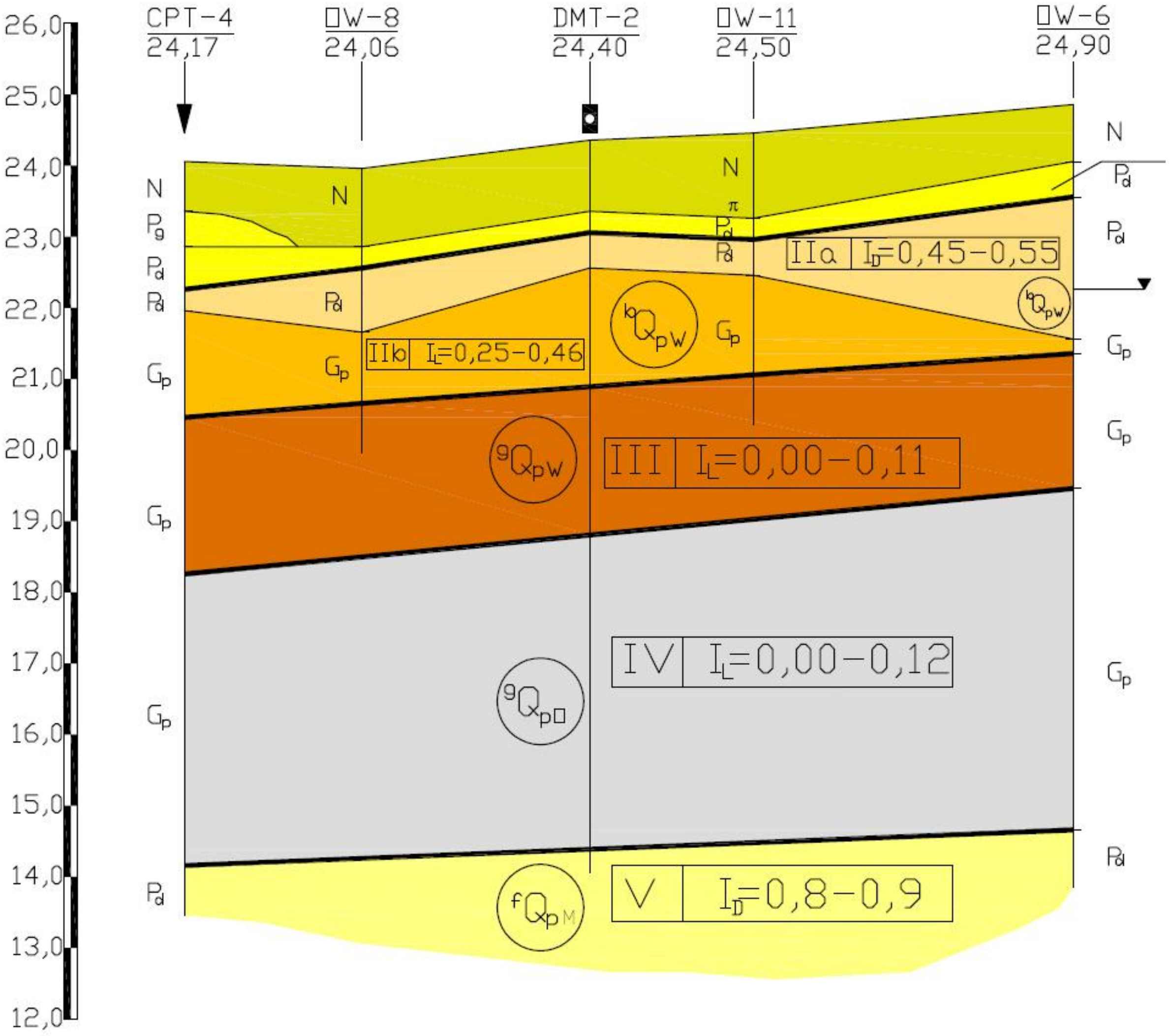


Figure 1: A standard geotechnical cross-section: OW – borehole test (reference method), CPT – cone penetration test, DMT – dilatometer test; (N – fill, Gp – sandy clay, Pd –fine sand, wn – moisture content, ID – relative density, IL –liquidity index)

3. The features selection

The process of feature selection is an important step in developing the efficient procedure of soils classification. Good features should be characterized by the stable values for samples belonging to the same class and at the same time they should differ significantly from different classes (Guyon and Elisseeff 2003) Tan et al. 2006) Liu & Yu 2003) Hall 2000). Thus the main problem in the classification and machine learning is to find out the features of the highest importance for the problem solution. Elimination of features with the weakest class discrimination ability leads to the reduction of the dimensionality of the feature space and improvement of generalization ability of the classifier in the testing mode for the data not taking part in learning.

In our analysis we have used the Fisher measure, depending on the clusterization of the data and description of the clusters using their means and standard deviations. It is evident that the variance of the features describing the cells belonging to the same class should be as small as possible. On the other hand, to distinguish between different classes, the positions of means of feature values for the data belonging to different classes should be separated as much as possible. We have combined both measures together to form the discrimination coefficient $S_{AB}(f)$ defined for the feature f at recognition of two cells belonging to different classes A and B:

$$S_{AB}(f) = \frac{|c_A(f) - c_B(f)|}{\sigma_A(f) + \sigma_B(f)}$$

In this definition c_A and c_B are the mean values of the feature f in the class A and B, respectively. The variables σ_A and σ_B represent the standard deviations determined for both classes.
Fig. 2 illustrates the change of the value of the discrimination coefficient $S_{AB}(f)$ for all extracted features at the recognition between the particular soil type.

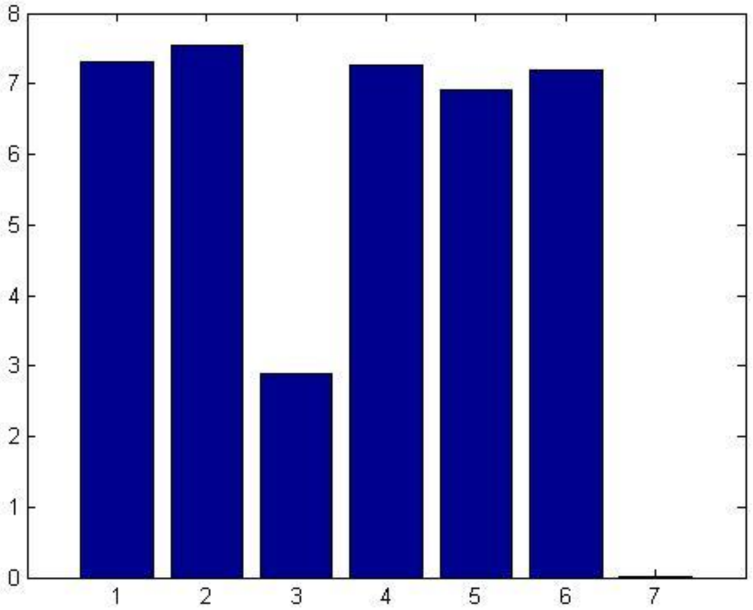


Figure 2 : Values of the discriminating coefficient $S_{AB}(f)$ of all features at the recognition of soil types

It is easy to observe that our dataset has 5 the most valuable features (1,2,4,5,6), one is mediocre (3) and one is useless (7). The determination of the optimal number of the chosen features is a separate problem. We have solved it by trying different number of the most significant features, testing the trained classifier on the validation data set and choosing the features providing the highest efficiency of recognition.

4. The applied classifiers

In our soil recognition system we have applied and compared two different classifiers: support vector machine (SVM) and k nearest neighbour classifier (KNN).

The SVM is a feed forward network of one hidden layer (the kernel function layer). It is known as an excellent classifier of good generalization ability (Vapnik 1998), Scholkopf 2002). The learning problem of SVM is formulated as the task of separating the learning vectors into two classes of the destination values either $d_i = 1$ (oneclass) or $d_i = -1$ (the opposite class), with the maximal separation margin. The separation margin formed in the learning stage according to the assumed value of the regularization constant C provides some immunity of this classifier to the noise, inevitably contained in the testing data. The great advantage of SVM is the unique formulation of the learning problem leading to the quadratic programming with linear constraints, which is very easy to solve. The SVM of the Gaussian (RBF) kernel has been used in our application. The hyperparameters σ of the Gaussian function and the regularization constant C have been adjusted by repeating the learning experiments for the set of their predefined values and choosing the best one at the validation datasets. The optimal values of these parameters found in these experiment were as follows: $\gamma = 0.5$ and $C = 1000$. To deal with a problem of many classes we have applied multiclass SVM.

The KNN classifier makes decision of class membership of the unknown vector $\boldsymbol{x}$ on the basis of its distances from k known (learning) vectors referred as the prototypes of classes. We assign the unknown vector to the class which appears most frequently in k selected prototypes identified in the previous step (Haykin 1999). Usually the Euclidean distance of $\boldsymbol{x}$ to the prototypes is used. The result depends on the value of k and the best choice of k depends upon the data. Generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct. Proper value of k has been selected by parameter optimization using cross-validation. As a result of such experiments we have found in our case $k = 5$ as the optimal one.

5. The results

The available data set has been split into five exchangeable parts to enable application of the fivefold cross-validation procedure. The class representatives have been split equally into all these parts. Four groups have been combined together and used in learning, while the fifth one used only in testing the trained classifiers. It means that always 500 vectors were used in learning process and 125 vectors took part in testing. The experiments have been repeated five times, exchanging the contents of the four learning and one testing subsets. The misclassification ratio in either learning or testing mode has been calculated as the mean of all five runs.

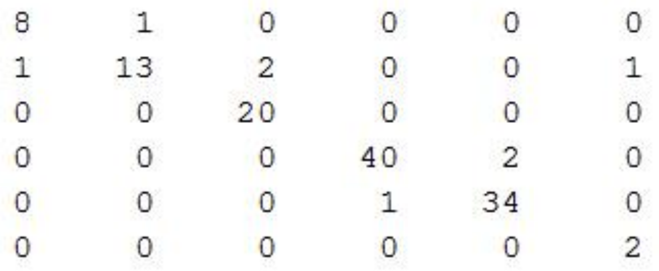
The best results were obtained by SVM classifier with reduced input vector by Fisher method. The comparison of accuracy of SVM and KNN classifiers may be presented in Table 1.

Table 1: Comparison of the accuracy of classification by KNN and SVM classifiers

	KNN	**SVM**
All features	72.8% (91/125)	77.6% (97/125)
4 best features	77.6% (97/125)	82.4% (103/125)
5 best features	88% (110/125)	**93.6% (117/125)**
6 best features	80.8% (101/125)	82.4% (103/125)

The results may be presented in the form of confusion matrix. It illustrates how the cases belonging to all classes of soil have been classified by our system. The columns represent the actual outputs of our system and the rows – the targets. The number in each entry of the 6×6 matrix is the total number of the actually recognized classes (sorts of soil) in testing mode, calculated in all 5 runs of cross-validation experiments. The diagonal entries of this matrix represent the quantity of the properly recognized cases. Each entry outside the diagonal means the number of misclassifications. The entry in the (i,j)th position of the matrix for $i \neq j$ means false assignment of the case of i-th class to the j-th one.

The confusion matrix of the best results of classification is presented in Fig. 3. It may be easy to observe that the most of mistakes are made in neighbouring classes. It is caused that the neighbouring sort of soils may be mixed or has similar properties.



```
8     1     0     0     0     0
1    13     2     0     0     1
0     0    20     0     0     0
0     0     0    40     2     0
0     0     0     1    34     0
0     0     0     0     0     2
```

Figure 3: The confusion matrix of the best results of classification

6. Conclusion

The paper has presented the research directed to the automatic recognition of soils on the basis of CPT and DMT data gethered from university campus. The proposed approach uses set of numerical features filtered by Fisher measure combined with different solutions of the classifiers. In our solution we compared SVM and KNN classifiers.

The proposed method appears to be useful for the automatic classification of the soil profiles. The results of numerical experiments show that this classification system is able to recognize the sorts of soils from the gathered data with the total accuracy of 93.6%. These results confirm, that an automatic learning based system can reach the efficiency comparable to the geotechnical human expert results.

ACKNOWLEDGEMENTS

This work is supported by the Polish National Centre of Science, Grant No. 2011/03/D/ST8/04309.

References

Forgy E. W. (1965). Cluster analysis of multivariate data: Effciency vs. interpretability of classification. Biometrics, 21:768-769

Guyon, I. & Elisseeff A. (2003), An introduction to variable and feature selection, Journal of Machine Learning Research, 2003, vol. 3, pp. 1158-1182.

Hall M. (2000), Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of the Seventeenth International Conference on Machine Learning Morgan Kaufmann Publishers, San Francisco.

Hashash, Y.M.A., et al. (2004). Numerical implementation of a neural network based material model in finite element analysis, International Journal for Numerical Methods in Engineering, 59, 989-1005.

Haykin S. 1999, Neural Networks, Comprehensive Foundation, Prentice-Hall, Englewood Cliffs, NJ.

Huang A., Mayne P. W. (2008). Geotechnical and Geophysical Site Characterization. Proc. of the 3 rd inter. Conf. on Site characterization, Taipei, Taiwan. Published by: Taylor & Francis Group, London, UK

Kruk, M. et al. (2014), Automated soil profile generation methods on the basis of DMT and CPT data, Proceeding of the International Conference on Artificial Intelligence and Computer Science (AICS 2014), 15 - 16 September 2014, Bandung, INDONESIA

Kurek J. et al. (2014), Automatic estimation of the number of soil profile layers using bayesian information criterion, Proceeding of the International Conference on Artificial Intelligence and Computer Science (AICS 2014), 15 - 16 September 2014, Bandung, INDONESIA

Liu H & Yu L (2003), Feature selection for high-dimensional data: A fast correlation-based based filter solution. Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03), Washington, D.C.

Lunne, T., Robertson, P.K., Powell, J.M. (1997). Cone penetration testing in geotechnical practice. Blackie Academic and Professional, London, England

Marchetti S. (1980). In Situ Tests by Flat Dilatometer. J. Geotech. Eng. Div., ASCE, 106, GT3, 299-321.

Matlab (2014) user manual , MathWorks, Natick.

Młynarek Z. (2007). Site investigation and mapping in urban area. Proc. of the 14th European Conference on Soil Mechanics and Geotechnical Engineering. Madrid, Vol. 1, 175-202.

Rabarijoely S. et al. (2007). The usage of the graph clustering algorithm to the recognition of geotechnical layers. Annals of Warsaw University of Life Sciences – SGGW. Ann. Warsaw Univ. of Life Sciences – SGGW, Land Reclam., No 38, 2007, 57 - 68.

Rabarijoely S. & Bilski P. (2009). Automated soil categorization using CPT and DMT investigations, 2nd International Conference on New Developments In Soil Mechanics and Geotechnical Engineering, 28-30 May 2009, Near East University, Nicosia, North Cyprus

Scholkopf B.& Smola A. (2002). LearningwithKernels, MITPress, Cambridge, MA.

Shahin, et al. (2005). Neural network based stochastic design charts for settlement prediction, Can. Geotech. Jour. (42), 110-120.

Tan P.N et al. (2006) Introduction to data mining, Pearson Education Inc.,Boston.

Vapnik V. (1998), Statistical Learning Theory, Wiley, New York.