Predicting soil liquid limit and plasticity index using tree-based learning

Application de l'apprentissage d'arbres à l'estimation de limite de liquidité et d'indice de plasticité des sols

M.C. Piantedosi
Department of Computer Science, University of Massachusetts Lowell, USA. Email: michael@dogsolitude.org
G.R. Livingston
Department of Computer Science, University of Massachusetts Lowell, USA. Email: gary@cs.uml.edu
P.U. Kurup
Department of Civil and Environmental Engineering, University of Massachusetts Lowell, USA. Email: Pradeep_Kurup@uml.edu
E.P. Griffin
Shaw Environmental & Infrastructure, Stoughton, Massachusetts, USA. Email: erin.griffin@shawgrp.com

ABSTRACT

This study documents the successful application of tree learning and data mining techniques to predict liquid limit (*LL*) and plasticity index (*PI*) from in situ cone penetration test (CPT) data. The primary learning methods used were regression trees, linear model trees, and recently popularized ensemble methods that average the results of many differently grown, individual trees. The data used in this study were obtained after a 7.6 magnitude earthquake struck Taiwan on September 1999 in the town of Chi-Chi, and numerous CPTs were conducted in the surrounding area. *LL* and *PI* data were obtained from soil samples taken from adjacent boreholes. Three approaches for predicting *LL* and *PI* were evaluated. The first approach predicted these values from cone resistance (q_c), friction ratio (R_f), total overburden stress (σ_{v0}), equilibrium pore pressure (u_0), and the associated depth of these measurements. The second approach predicted *LL* and *PI* using q_c, R_f, σ_{v0}, u_0, and their depth, along with a binary *plastic identity* feature that indicated whether or not a soil was plastic or non-plastic. The third approach consisted of a two-step prediction process where classification-tree learning methods were used to train models for predicting the plastic identity from q_c, R_f, σ_{v0}, u_0, and their depth. This predicted value was then used, along with q_c, R_f, σ_{v0}, u_0, and their depth, to predict *LL* and *PI*. The third approach was found to give better estimates of *LL* and *PI* than the first method, and the second approach gave the best predictions.

RÉSUMÉ

Cet article montre comment l'apprentissage d'arbres ainsi que certaines techniques de data mining peuvent être appliquées de manière efficace à l'estimation de l'indice de plasticité (*PI*) et de la limite de liquidité (*LL*), à partir de données *in situ* d'essais de pénétration au cône (CPT). Les méthodes d'apprentissage clés qui sont utilisées sont les arbres de régression, les arbres linéaires, ainsi que les méthodes d'ensemble, popularisées récemment, et qui calculent la moyenne des contributions de nombreux arbres individuels. Les données utilisées dans cette étude ont été obtenues après un tremblement de terre de magnitude 7.6, qui a frappé Taiwan au mois de septembre 1999. L'épicentre était situé dans la ville de Chi-Chi, ou de nombreux CPT furent effectués, ainsi qu'aux alentours de Chi-Chi. Les données de *LL* et *PI* ont été obtenues à partir de sismiques de puits adjacentes. Trois approches pour l'estimation de *LL* et *PI* sont évaluées. La première approche estime ces valeurs à partir de la résistance au cône (q_c) du coefficient de friction (R_f), de la contrainte de surcharge totale (σ_{v0}), de la pression interstitielle d'équilibre (u_0), ainsi que de la profondeur associée à ces mesures. La seconde approche estime *LL* et *PI* à partir de q_c, R_f, σ_{v0}, u_0, leur profondeur, ainsi que de *l'identité plastique*, un paramètre binaire indiquant si oui ou non, le sol est plastique. La troisième approche est un processus en deux étapes où: (1) des méthodes d'apprentissage basées sur les arbres de classification sont utilisées pour entrainer les modèles à estimer l'identité plastique, à partir de q_c, R_f, σ_{v0}, u_0, et de la profondeur à laquelle ces valeurs sont mesurées; (2) la valeur ainsi calculée est utilisée avec q_c, R_f, σ_{v0}, u_0, et la profondeur associée, afin d'estimer *LL* et *PI*. La troisième approche donne de meilleures estimations de *LL* et *PI* que la première méthode, et la seconde donne de meilleures estimations que les deux autres.

Keywords: Regression trees, model trees, bagging, soil property prediction, plasticity index, liquid limit

1 INTRODUCTION

Liquid limit (*LL*) and plasticity index (*PI*) are important indices required for soil classification. A soil's *LL* and *PI*, along with its natural water content, can give a qualitative idea about its compressibility characteristics and shear strength. The traditional procedures for determining these two indices are based on laboratory tests (ASTM D4318) of disturbed soil specimens obtained from boreholes. These methods can be time-consuming and expensive. The potential use of a well-known in situ test method, the cone penetration test (CPT), for determining *LL* and *PI* has been discussed by Cetin and Ozan (2009). The feasibility of using the CPT for estimating *LL* is perhaps due to its similarity (though distant) to the Swedish fall cone test, which is a standard laboratory method used widely in Europe and Asia for determining *LL* (BS 1377 1990). This study focuses on the application of tree learning and data mining techniques to predict *LL* and *PI* from CPT data. The CPT essentially consists of pushing an instrumented cylindrical probe, known as the cone penetrometer, into the soil at a rate of 2 cm/s. An electronic depth encoder measures the penetration depth. The cylindrical probe has a conical tip with apex angle equal to 60º. The device is equipped with a load cell at the tip to measure the cone or tip resistance (q_c), which is the force offered by the soil to the tip during intrusion divided by the projected cone area. The projected cone area of the standard cone penetrometer is 10 cm^2. The cone penetrometer is also equipped with a friction sleeve (150 cm^2 surface area) and a load cell to measure the sleeve friction (f_s), which is the shear stress between the surrounding soil and the shaft of the probe. The term friction ratio (R_f), often used in CPT data interpretation, is the ratio of the sleeve friction to the cone

Proceedings of the 17th International Conference on Soil Mechanics and Geotechnical Engineering
M. Hamza et al. (Eds.)

doi:10.3233/978-1-60750-031-5-1020

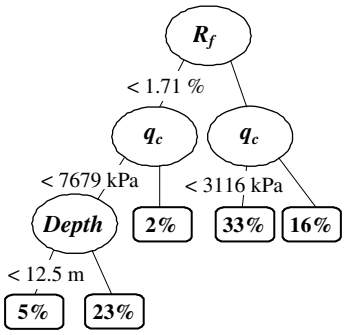

Figure 1. An example of a regression tree for predicting liquid limit

resistance, expressed as a percentage. The data acquired during a CPT is plotted on a computer screen in real time (as q_c and f_s vs. depth), and may be used for soil classification and estimating various soil properties (Lunne et al. 1997).

In this study three approaches were evaluated for predicting *LL* and *PI* from CPT data. The first approach essentially tried to predict these values directly from cone resistance (q_c), friction ratio (R_f), total overburden stress (σ_{v0}), equilibrium pore pressure (u_0), and the depth associated with these measurements. In the second approach, a binary *plastic identity* feature that differentiates plastic soils from non-plastic soils is added to the values listed in the preceding sentence. The third approach consists of a two-step prediction process where classification-tree learning methods were used to train models for predicting the values of plastic identity from q_c, R_f, σ_{v0}, u_0, and depth. This predicted value was then used, along with q_c, R_f, σ_{v0}, u_0, and depth, to predict *LL* and *PI*.

2 METHODS

2.1 Machine Learning

Machine learning methods provide powerful data mining tools not only for discovering new knowledge but also for fusing non-commensurate features, such as those used in this study, into predictive models. Machine learning has been successfully used in a variety of applications, ranging from biomedical, such as proteomics (Shaughnessy et al. 2008), to geotechnical engineering (Livingston et al. 2008).

2.1.1 Tree-based learning

Tree-based learning is one of the earliest developed and best understood machine-learning methods (Witten and Frank 2005). Tree-based learning methods generate models in which the internal nodes form a tree-like structure. The internal nodes represent tests on a feature's values, and the leaf nodes represent classes. The tree is used to classify a case by traversing from the *root node* at the top to a *leaf node* at the bottom. The values of the case's features are used to determine the path taken to reach a leaf node. In a decision tree, one of the simplest tree types, the class represented by a leaf node reached via this traversal would be the class predicted for the case.

Decision trees are learned by recursively selecting the features that sort the data into non-overlapping subsets where total polarization of classes between subsets is the best case. As the splitting continues, the features chosen for splitting the data are arranged into a tree. The first feature chosen for splitting forms the root of the tree. Each subset created in this splitting is again split by another feature, which is attached as a child node of the first feature and so on. This recursive splitting continues until a subset's members have the same class or the subset's size is less than a user-supplied threshold.

2.1.2 Numeric Prediction Trees

Unlike a discrete classification problem where unknown classes are predicted, numeric prediction is the process of predicting unknown numeric values. *Regression trees* and *model trees* are two well understood numeric prediction techniques which are based on decision tree learning techniques. The only difference in applying a numeric tree vs. a decision tree is that when a leaf is reached, in a regression tree a numeric value is assigned, and in a model tree a linear expression is evaluated.

Training a regression tree is similar to training a decision tree, but the feature chosen for a given node is the feature that minimizes numeric error at that node. The splitting ends when no cases that reach the current node vary significantly. Finally, a leaf node is attached to the end of each branch, which contains the average target value of all cases that reach that leaf.

Training a model tree is similar to training a regression tree. The main difference is that a standard linear regression model is calculated for each node. Finally, the linear model for each internal node is combined with each of its children in turn until all models encountered down a particular branch are aggregated at the leaves of the tree.

Figure presents a sample regression tree for predicting *LL*. For example, assume a case with q_c of 6000kPa and R_f of 1.5% at a depth of 20m. The evaluation is started at the root node where the value of R_f is checked. Since the R_f for this case is 1.5% and is less than 1.71%, the left branch is taken. Next, the model checks if q_c is less than 7679kPa. This is true, so the left branch is taken at this node as well. Finally the depth is checked vs. 12.5m. Since the depth for the case is 20m, which is greater than 12.5m, the right branch is taken and the model assigns a *LL* value of 23% to this case.

Figure presents a sample model tree for predicting *PI*. Again, assume a test case with q_c of 6000kPa and R_f of 1.5% at a depth of 20m. First, q_c is checked against 3040kPa. Since it's greater, the right branch is taken and R_f is checked against 0.699%. Since this is also greater the right branch is taken again. The *PI* is now computed by substituting the case values for depth, σ_{v0}, q_c, and R_f into the linear expression contained in the right-most leaf node.

2.1.3 Bagging

Ensemble machine learning methods, such as bagging and boosting of tree learning methods, are significant recent developments in machine learning. Generally, an ensemble method learns a number of slightly different models from the same training data, which are then combined to form a more robust aggregate model. Contemporary studies have shown that these developments allow tree-based learning methods to not only overcome some of the problems that plagued them in the early- and mid-1990's, but also compete with or outperform most other methods on complex problems (Caruana and Niculescu-Mizil 2006).

Because slight changes to the training set can produce very different trees, the standard top-down tree induction process is inherently unstable. Bagging (Witten and Frank 2005) attempts to mitigate this instability by repeatedly inducing trees using altered versions of the original training set, and then combining the predictions made by the trees. This is accomplished by

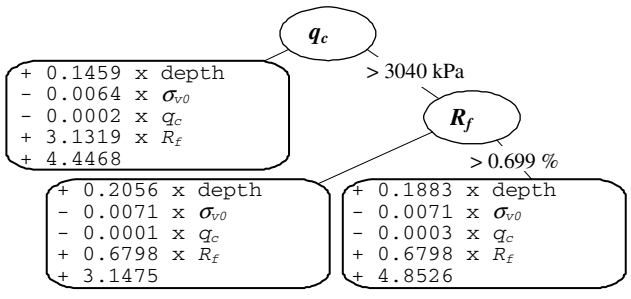

Figure 2. An example of a model tree for predicting plasticity index

resampling with replacement from the training data, where

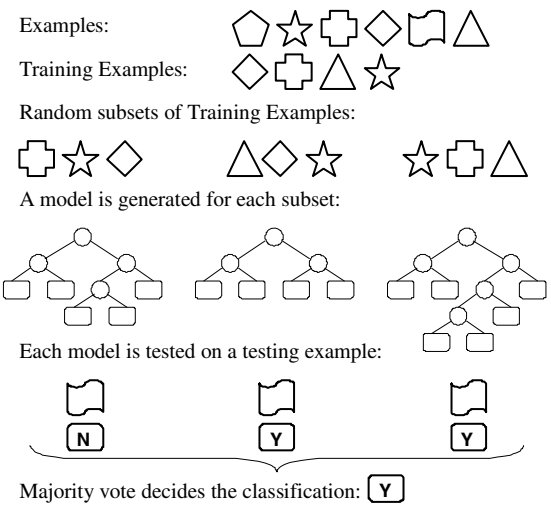

Figure 3. Depiction of the bagging process

some of the original training cases are deleted and other cases are replicated to create a new training set of the same size. A tree is then learned for each new version of the training set. When using the learned trees to make a prediction, a single value is predicted by taking the average of the individual predictions. The bagging process is illustrated in Figure .

2.2 Data

The data used in this study were obtained from tests conducted after a 7.6 magnitude earthquake struck Taiwan on September 2nd, 1999 (Juang 2002). The epicenter was located in the town of Chi-Chi. A total of 139 instances from 12 CPT-SPT pairs were used for training and testing, including five pairs from the Yuanlin site, one pair from the Dachun site, and three pairs each from the Nantou and Wufeng sites. *LL* and *PI* were obtained from soil samples collected during the SPT according to standard ASTM procedures (ASTM D4318). Cone resistance (q_c), friction ratio (R_f), and equilibrium pore pressure (u_0) were obtained from the adjacent CPT sounding, and total overburden stress (σ_{v0}) was calculated based on laboratory determined unit weights. In order to use this data for training and testing, it was necessary for a CPT sounding to be adjacent to, or near, an SPT boring so that the CPT parameters could be correlated to the *LL* and *PI*. Based on reported survey coordinates, most pairs were less than 6m apart, with several being co-located.

2.3 Plastic Identity

Some of the resulting numeric predictions for non-cohesive sandy soils were non-zero, which is incorrect according to civil engineering definition. To remedy this situation, a *plastic identity* feature was added to the data that indicated whether the soil was plastic (fine-grained) if *LL* and *PI* were non-zero, or non-plastic (coarse-grained) otherwise. The addition of this feature not only completely fixed these non-zero results, but also allowed the prediction accuracy to more than double in almost every experiment. Establishing the value of plastic identity requires soil sampling and possibly laboratory testing as well. Therefore, we evaluated the feasibility of first predicting this feature from the other features and then using the predicted values in place of the actual values for plastic identity. The tree classifiers trained for predicting plastic identity were single and bagged decision trees.

2.4 Learning Algorithms Used

The primary prediction methodology used in this study is based upon numeric tree learning. The most effective trees used in this case were regression trees, model trees, and bagged varieties of each. The tree learning methods implemented in the Weka machine learning and data mining package (Witten and Frank 2005) were used in this study: Decision trees were used to predict values for the plastic identity, and both regression trees and model trees were used to predict the values of *LL* and *PI*. Various manual tunings of the trees' confidence factors and numbers of instances per leaf were evaluated, but the gains in accuracy achieved by tuning were insignificant when compared with the default values of 0.25 and 2 respectively.

2.5 Training and Evaluation Procedures

Since these methods are intended for use in the field, one can imagine training the algorithms on data from one set of boreholes and attempting to predict data for new boreholes. Therefore, instead of randomly shuffling around the data instances which each came from a certain depth in a particular borehole, approximately 33 percent (4) of the borehole pairs were held back for testing. Since we only have 12 borehole pairs that don't exhibit all necessary features, a domain expert made these selections, which resulted in 41 data instances for testing, from these 4 borehole pairs, and the other 98 instances, from the remaining 8 boreholes, were used for training.

Three sets of experiments were performed. In the first set of experiments, which we refer to as *No Plastic Identity* in our figures, the learning algorithms listed in Section 2.4 were used to learn models from the training data, and then their accuracies were evaluated using the testing data. In the second set of experiments, which we refer to as *Actual Plastic Identity* in our figures, plastic identity, described in Section 2.3, was added to the training and testing data, from which the learning algorithms were used to learn models that were then tested using the modified testing data. In the third set of experiments, which we refer to as *Predicted Plastic Identity* in our figures, additional models were learned from the modified training data *for predicting plastic identity.* These models were then used to predict the values of plastic identity for each of the cases in the original testing set (without the actual plastic identity values), and then the *predicted values* for plastic identity were added to the original testing data. The models learned during the second set of experiments, *learned using actual plastic identify values*, were then evaluated using this new testing dataset, which *contained only predicted plastic identity values*.

3 RESULTS

The results of this study are presented in Figure 4. Without the plastic identity, the models occasionally predicted non-zero values for *LL* and *PI* for non-plastic soils, which would be confusing to a domain expert and are less than ideal. Also, some (not reg. trees) predicted a few small negative values rather than zero and are not shown in Figure 4, but were used to calculate the mean absolute error. These could be automatically rounded up in practice.

When plastic identity was added into the dataset, the accuracy of the learned models improved substantially across all learning methods and most instances of zero *LL* and *PI* were correctly predicted for sandy soils, particularly with regression trees. Plastic identity was predicted with 90 percent accuracy using bagged decision trees, and 82 percent accuracy using only a single decision tree. Finally, adding these predicted plastic identity values into the original dataset and applying numeric prediction generally improved upon the original predictions, but failed to achieve the same accuracy as with the actual plastic identity feature. This is probably primarily due to misclassifications resulting in numeric anomalies.

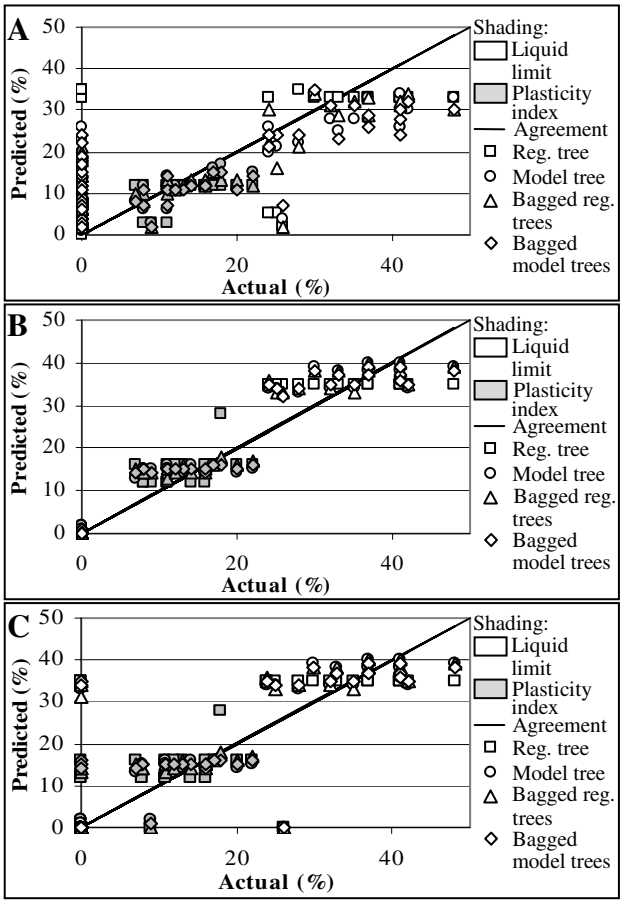

Figure 4. Scatter plots of the predicted values for *LL* and *PI* versus their actual values. A: No plastic identity; B: Actual plastic identity; and C: Predicted plastic identity

4 DISCUSSION

The reason this study focused on tree learning is that the learned trees are not only competitive with other approaches, but they can also provide additional insight into the predicted process beyond that which can be achieved via neural networks, support vector machines, or similar "black-box" techniques.

An example of a problem domain where it would be useful to be able to predict *PI* and *LL* is predicting earthquake induced soil liquefaction (Boulanger and Idriss 2006). *PI* and *LL* are known factors in this process and being able to predict them from datasets, which include only field test information, is very important. *LL* can also be used to get a rough estimate of the compression index C_c from empirical correlations such as: $C_c = 0.009\,(LL{=}10)$.

The results of this study, which aren't accurate enough to use in practice, may appear to demonstrate no real correlation with the target data to those untrained in machine learning. However, the strength of any machine learning method bears a direct relationship to the quality and quantity of the training data (Witten and Frank 2005). If the algorithms had learned nothing, the results would be extremely random, but that is not the case: the graphs in Figure 4 clearly show a correlation between the predicted and actual values. It is expected that accuracy will improve with a larger database to learn from.

Additionally, the fact that all training data came from the areas around Chi-Chi Taiwan means that the learned models may be site specific and should only be applied in areas with similar geology.

It is also important to stress that the dataset used in this study involved the pairing of nearby CPT and SPT boreholes to obtain full data instances, and that this process is certainly prone to some error. Although the CPT and SPT locations may be adjacent to each other, site heterogeneity is expected to contribute to discrepancies between the CPT data recorded at a certain sounding depth and the *LL* and *PI* of a sample collected from the same depth at a nearby SPT borehole location.

5 CONCLUSIONS

The tree learning methods used in this study were able to estimate *LL* and *PI* with mean absolute errors of approximately 8.7 and 3.9, respectively, using only CPT data. Classification-tree learning methods were able to predict whether a soil is plastic or non-plastic using only CPT data with 90% accuracy, and when these predictions were used, along with the original CPT variables, to predict *LL* and *PI*, the mean absolute errors of the predictions of *LL* and *PI* were 5.3 and 2.6, respectively.

This study has demonstrated new and informative applications of tree learning in the problem domain of soil plasticity and *LL* prediction. Both of these attributes are crucial for characterizing fine-grained soils. While not yet accurate enough to be used in practice, it is expected that accuracy of the predicted *LL* and *PI* will improve with a larger database from test sites around the world.

ACKNOWLEDGEMENTS

Members of the UMass Lowell fall 2006 machine learning class contributed to this study, which was made possible by a 2006-2008 UMass Lowell Healey grant. Kurup and Griffin also appreciate the financial support of the U.S. National Science Foundation under Grant No. CMS-0409594 (cognizant program director is Dr. Richard J. Fragaszy). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the writers and do not necessarily reflect the views of the funding agencies.

REFERENCES

ASTM D4318. 2005. Standard Test Methods for Liquid Limit, Plastic Limit, and Plasticity Index of Soils.

Boulanger R. W. and Idriss, I. M. (2006). Liquefaction Susceptibility Criteria for Silts and Clays. *American Society of Civil Engineers, Journal of Geotechnical and Geoenvironmental Engineering*, 132(11), 1413-1426.

BS 1377. 1990. British Standard Methods of Tests for Soil for Engineering Purposes.

Caruana, R. & Niculescu-Mizil, A. 2006. An empirical comparison of supervised learning algorithms. In ICML '06: *Proceedings of the 23rd international conference on Machine learning*, Pittsburgh, Pennsylvania: 161-168.

Cetin, O. K. & Ozan, C. 2009. CPT-Based Probabilistic Soil Characterization and Classification. *Journal of Geotechnical and Geoenvironmental Engineering*, 135(1), 84-107.

Juang, C. H. (2002). "Soil liquefaction in the 1999 Chi-Chi, Taiwan earthquake, in-situ test data."
http://www.ces.clemson.edu/chichi/TW-LIQ/In-situ-Test.htm (Dec. 3, 2004).

Livingston, G.R., Piantedosi, M., Kurup, P. & Sitharam, T.G. 2008. Using decision-tree learning to assess earthquake-induced liquefaction potential. In *Proceedings of 4th Geotechnical Earthquake Engineering and Soil Dynamics Conference (GEESD-2008),* Sacramento, California.

Lunne, T., Robertson, P. K., and Powell, J. J. M. 1997. Cone Penetration Testing in Geotechnical Practice, Blackie Academic & Professional, London.

Shaughnessy, P., Livingston, G.R. & Graves, M.V. 2008. Towards predicting protein-protein interactions in novel organisms. *Journal of Computational Biology and Drug Design* 1(3): 235-253.

Witten, I.H. & Frank, E. 2005. *Data mining: Practical machine learning tools and techniques, second edition*. Morgan Kaufmann.