Application of machine learning on soil classification based on CPTU data

J. Liu
NGI, Oslo, Norway, Jingdong69.liu@gmail.no

G. Sauvin[1], S. Yang[2]
NGI, Oslo, Norway, Guillaume.Sauvin@ngi.no[1], Shaoli.Yang@ngi.no [2],

ABSTRACT: Cone Penetration Testing (CPTU) is one of the most popular soil investigation methods offshore, both for offshore oil and gas industry and offshore windfarm industry. A CPTU test is a continuous recording of the subsurface at the test location. It contains the vertical variations of the mechanical characteristics of the subsoil. These variations in turn indicate changes in geological layers and their properties. During a test the cone tip resistance, the friction along the rod and excess pore water pressure can be obtained. A considerable amount of CPTU data is available from industry and research projects so far. Soil type, such as sand, silt and clay, can be classified by CPTU data. Normally this is done based on soil behavior type (SBT). In addition to SBT method, machine learning is used for soil classification in this study. A big dataset, combining CPTU and corresponding lab data, is used for training purpose and application of the method in soil classification is discussed. Synthetic classes are used in this study and ML technique is used to mimic the Ic and SBT classification based on CPTU. Grain-size distribution as well as Attemberg`s limits datasets are used to further refine the classification according to Unified Soil Classification System (USCS).

Keywords: Cone Penetration Testing (CPTU); Soil Classification; Machine Learning; Soil Behavior Type (SBT)

1. Introduction

Machine learning (ML) has been used frequently in Geotechnical engineering field, e.g., in liquefaction assessment (Garcia, et al., 2012, [1], Ardakani and Kohestani, 2015, [2]), in predicting spatial soil type distribution (Ghaderi, et al., 2018, [3]), in site characterization (Tsiaousi, et al., 2018, [4]), in pile capacity evaluation (Goh, 1995, [5], Maizir et al. 2015, [6], Mazaher, A., 2016, [7]), in settlement prediction of foundations (Shahin, et al., 2002, [8], Alkroosh and Nikraz, 2011, [9]) and in slope stability analysis (Cho, 2009, [10], Peng, et al., 2014, [11]). ML has also been used for soil classification, e.g., Bhargavi and Jyothi, 2011, [12], used data mining technics to classify agriculture soil types, and it was indicated that Fuzzy classification rules is a good method for the soil classification.

CPTU is a widely used in situ test method to get variation of soil properties and it is one of the most efficient and repeatable method for site investigation. A significant amount of CPTU data is available and can be used for ML analysis (Alkroosh and Nikraz, 2011, [9], Ardakani and Kohestani, 2015, [2], Ghaderi et al., 2018, [3]). Classical CPTU based soil classification uses empirical soil classification charts by Robertson et al., 1986, [13], Robertson, 1990, [14] and Robertson, 2016, [15]). Bhattacharya and Solomatine, 2006, [16] applied decision trees, Artificial Neural Network (ANN) and Support Vector Machines (SVM) techniques to CPT data for soil classification with a prediction accuracy of about 83%.

In this study, we use a CPTU database for soil classification purpose. The SBT index Ic is directly calculated from CPTU data and used for model testing. Synthetic classes are used and ML technique is used to mimic the Ic and SBT classification based on CPTU. Ku et al., 2010, [17] concluded that SBT index, Ic, either in terms of the RW formulation (Robertson and Wride,1998, [18]) or the JD formulation (Jefferies and Davies, 1993, [19]) is an effective parameter for soil classification. Ic reflects the mechanical behavior of soils and can distinguish sand like soils from clay like soils without knowledge of particle size distribution and plasticity. In terms of the RW formulation, the most suitable cut-off Ic value for distinguishing sand-like soils from clay-like soils is 2.67; in terms of the JD formulation, the most suitable cut-off Ic value is 2.58. These cut-off values were established based on statistical analyses of the data set derived and employed in the study. However, research with additional high quality data to further validate these cut-off values is desirable. In this regard, collection of soils that can expand the database is especially important

In this study, we also use the Unified Soil Classification System (USCS) for further refining the soil classification based on CPTU data. We apply Random Forest Classifier (RFC) technique and find this to be a good method for soil classification.

2. Interpretation of CPTU data

Typical raw CPTU data includes the cone tip resistance, q_c, the sleeve friction, f_s and pore water pressure, u_2. Interpretation of raw CPTU data (u_2, f_s and q_c) has traditionally been done by derivation of parameters like friction ratio (R_f), normalized cone tip resistance (Q_t), and normalized friction ratio (F_r) and their subsequent use in existing charts or classifications.

The corrected cone tip resistance, q_t:

$$q_t = q_c + u_2(1-\alpha) \quad (1)$$

where α is the net area ratio

The friction ratio (R_f) represents the ratio between f_s and q_c:

$$R_f = \frac{f_s}{q_t} \times 100\% \quad (2)$$

With u_2 the pore pressure measured between the cone tip and the friction sleeve and the net area ratio determined by the characteristics of the used cone. Stress-normalized equivalents of the variables q_t and R_f should be used to account for the in-situ vertical stresses: the normalized cone tip resistance, Qt:

$$Q_t = \frac{q_t - \sigma_{v0}}{\sigma_{v0}'} \quad (3)$$

and the normalized friction ratio, F_r:

$$F_r = \frac{f_s}{q_t - \sigma_{v0}} \times 100\% \quad (4)$$

where σ_{v0} is the total overburden pressure, and σ_{v0}'' the effective vertical stress.

Pore pressure ratio, B_q, is defined as:

$$B_q = \frac{u_2 - u_0}{q_t - \sigma_{v0}} \times 100\% \quad (5)$$

Where u_0 is equilibrium pore pressure.

Jefferies and Davies (1993, [19]) introduced the SBT index Ic to represent the radius of the concentric circles in the classification diagram of Robertson (1990, [14]).

Jefferies and Davies (1993, [19]) proposed an expression for Ic:

$$Ic = \sqrt{[3.0 - log_{10}\, Q_t\,(1 - B_q)]^2 + [1.5 + 1.3 log_{10}\,(F_r)]^2} \quad (6)$$

Robertson and Wride (RW) (1998) proposed an other expression for Ic:

$$Ic = \sqrt{[3.47 - log_{10} Q_{tn})]^2 + [1.22 + log_{10}\, F_r]^2} \quad (7)$$

$$Q_{tn} = \frac{(q_t - \sigma_{v0})/\sigma_{atm}}{(\sigma_{v0}'/\sigma_{atm})^n} \quad (8)$$

$$n = 0.381 Ic + 0.05 \frac{\sigma_{v0}'}{\sigma_{atm}} - 0.15,\ n \le 1.0 \quad (9)$$

Where σ_{atm} is the reference pressure.

The Ic variable captures only the soil type from the raw CPTU data, and carries little or no information on the in situ soil state (consolidation, cementation, or sensitivity). In contrast, the 2D classification charts (Roberson et al., 1986, [13], and Robertson 1990, [14]) do include such additional soil state information.

3. Machine learning method

Artificial Intelligence (AI) means making or automating the human tasks with help of Algorithms and programming. In recent 20 years, AI algorithms have been proved promising in many fields, including geotechnical engineering. Machine learning (ML) is a broad subfield of AI that uses multivariate, nonlinear, nonparametric regression or classification algorithms and techniques to learn from existing data and develop predictive models. ML can be very useful for solving problems where deterministic solutions are not available or are very expensive in terms of computational cost, but for which there is significant data available.

Random Forest (RF) is one of the many machine learning algorithms used for supervised learning, this means for learning from labelled data and making predictions based on the learned patterns. RF can be used for both classification and regression tasks. Random Forests (RFs) are ensemble-based decision trees and were developed to overcome the shortcomings of traditional decision trees. In RF, like other ensemble learning techniques, the performance of a number of weak learners is boosted via a voting scheme. The main hallmarks of random forest include; 1) bootstrap sampling – randomly selecting number of samples with replacement, 2) random feature selection – randomly selecting only a small number of features in the split of each node, 3) full depth decision tree growing, and 4) Out-of-bag error estimation – calculating error on the samples which were not selected during bootstrap sampling (Jiang et al., 2009, [20]).

4. Testing of Machine learning method

In this study, we used JD formulation of Ic (Jefferies and Davies, 1993, [19]). Figure 1 shows the commonly used soil behaviour type chart.

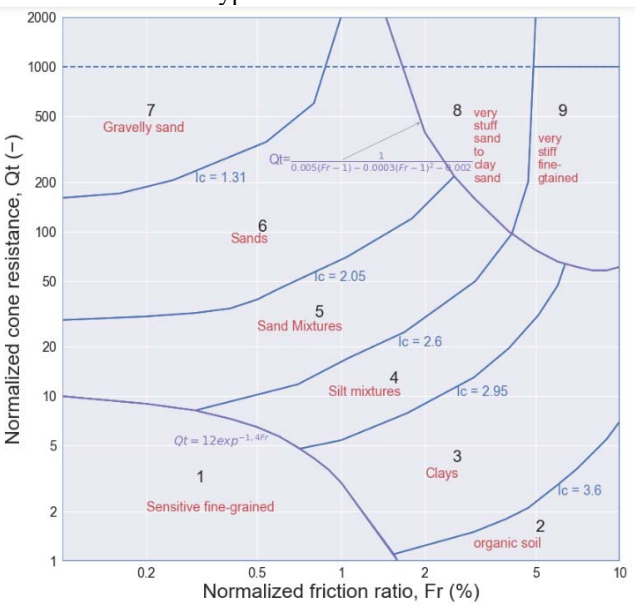

Figure 1. Normalized CPT soil behavior type chart.

Table 1 shows the proposed description and corresponding I_c values for soil classification (Robertson et al., 1986, [13] and Robertson 1990, [14]).

Table 1. Soil Behaviour Type

Proposed description	SBT_n	SBT	I_c
Sensitive fine-grained	1	1	N/A
Clay - organic soil	2	2	> 3.6
Clays: clay to silty clay	3	3	2.95-3.60
Silt mixtures: clayey silt & silty clay	4	4&5	2.60-2.95
Sand mixtures: silty sad to sandy silt	5	6 & 7	2.05-2.60
Sands: clean sands to silty sands	6	8	1.31-2.05
Dense sand to gravelly sand	7	9 & 10	< 1.31
Stiff sand to clays sand	8	12	N/A
Stiff fine-grained	9	11	N/A

To determine if any of the soil layers contains "sensitive clays and silts" from zone 1 or "stiff soil" from zones 8 and 9, the following rule can be used: (see Figure 1):

1) The soil layers belong to zone 1, if
$$Qt < 12e^{-1.4Fr}, \quad (10)$$
2) The soil layers belong to zone 8 and 9, if
$$Qt > \frac{1}{0.005(Fr-1)-0.0003(Fr-1)^2-0.002}, \quad (11)$$

3) Zones 8 and 9 are separated by line I_c =2.6

Our database includes 357 CPT tests sampled at 0.5 meter. There are 12921 CPT data points in total to be used in the application of machine learning methods (Figure 3). Nine types of soils can be classified according to Ic values proposed by JD and 67.6% of the soil in the database belong to group 3 (clay to silty clay). Table 2 presents the database distribution of the different soil behavior types, while Table 3 gives a list of statistical summary of mean, standard deviation, median, 25% and 75% percentile, minimum and maximum values of input features Rf, Qt, Fr and Bq. Figure 2 shows data in the train datasets and Figure 3 shows data in the test datasets. Data that was predicted in a wrong category was shown in Figure 3 too. It should be noted that synthetic classes are used in this study and ML technique is used to mimic the SBT classification.

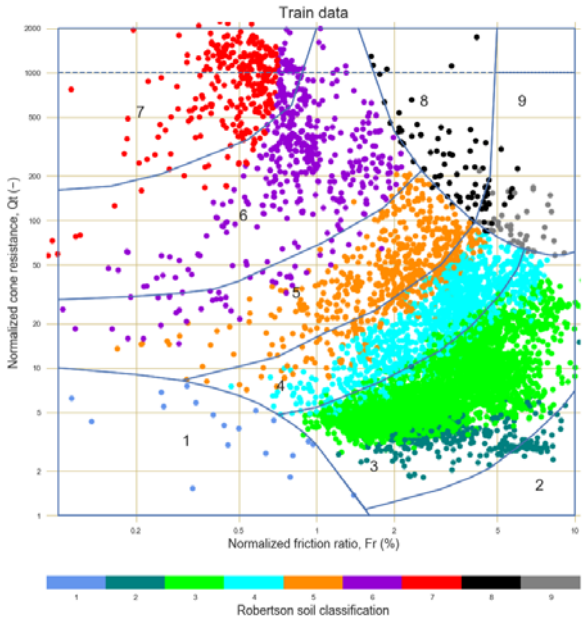

Figure 2. Soil types based on the available train dataset, boundary lines are based on Robertson

There are various techniques that help improving the performance of imbalanced datasets. Under-sampling and over-sampling are two of them. Under-sampling removes samples randomly from over-represented classes, and it is useful for large dataset. Over-sampling adds more samples from under-represented classes and is useful for a relatively small dataset. SMOTE (Synthetic Minority Over-sampling Technique) is an over-sampling method which creates synthetic samples of the minority class. In this study, we use the Python package Imblearn to perform over-sampling for the minority classes.

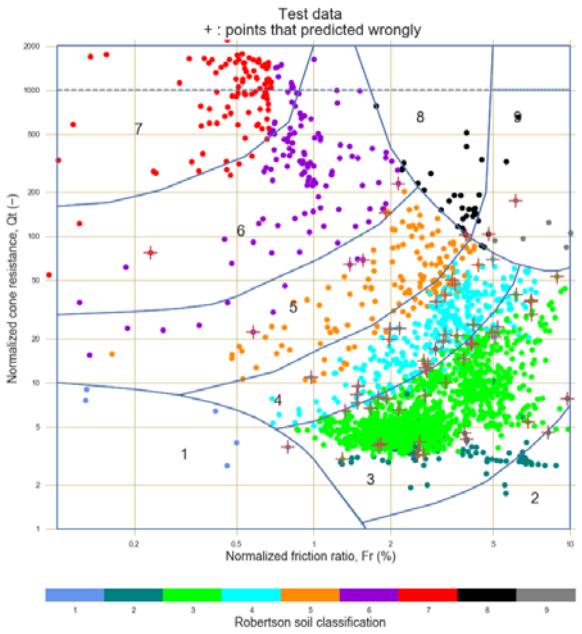

Figure 3. Soil types based on the available test dataset, boundary lines are based on Robertson

Table 2. Different Soil Behaviors types in the datasets

Soil Behaviours types	Number of samples	Percentage of samples	samples for train data	samples for test data
1	34	0.3%	28	6
2	480	3.7%	393	87
3	8734	67.6%	6992	1742
4	1558	12.1%	1240	318
5	801	6.2%	651	150
6	710	5.5%	569	141
7	426	3.3%	326	100
8	101	0.8%	73	28
9	77	0.6%	64	13
Sum	12921	100%	10336	2585

Table 3. Statistical description of samples

Feature	Number of _samples	Mean	Standard Deviation	Min
Rf	12921	2.1	1.47	0.0
Bq	12921	0.33	0.28	-3
QT	12921	65.3	222	0.5
Fr	12921	2.8	1.62	0.0

Feature	25% percentiles	median	75% percentiles	Max
Rf	1.24	1.6	2.93	21.8
Bq	0.047	0.38	0.59	0.99
QT	4.38	5.8	19.7	2942
Fr	1.90	2.4	3.44	22.3

We train our model following this procedures:

(a) Python Scikit-learn package is used to train the model and predict the result using the random forests method (Buitinck, et al., [22]).

(b) The available dataset is randomly split in two parts: training dataset (80% of the whole dataset) and test dataset (20% of the whole dataset). The training dataset is used to train the model and the test dataset is used to evaluate the prediction accuracy.

(c) The derived CPT parameters 'R_f', 'Q_t', 'F_r','B_q' are input data, and soil behaviour types, SBT classes(1-9) are target and output.

(d) The random forest parameters such as the numbers [1,5,10,30,50,100] of trees in the forest, the maximum depth [1,2,5,10,15,20,30, 40,50,100] of the tree and the minimum number [1,2,3,4,5,6,7]of samples required to be at a leaf node are tested to fit the model.

(e) If we set the number of trees in the forest as 30~40, the maximum depth of one tree as 20~30, accuracy of the total testing samples is $\frac{2525}{2585} * 100\% = 97.7\ \%$ (Table 4). Since the soil type 3 is the majority type, types 2 and 4 have lower accuracy.

(f) If we randomly reduce the number of samples of the soil type 3 from 6992 to 1240, keeping the same numbers for the remaining samples, the accuracy for test data changes from 97.7 % to 95.6%=($\frac{2475}{2585} * 100\%$). Accuracy for the majority type is reduced, however the types 2 and 4 show better accuracy (Table 5)

(g) From Figure 4, Q_t is the most important input parameter for the model with 45.45% importance, followed by Fr with 23.47% importance, Bq with 18.12% importance and Rf with 12.95% importance to soil classification.

Table 4. Accuracy for different soil type for the test set before reducing number of majority samples

		Predicted										
	Class	1	2	3	4	5	6	7	8	9	total	accuracy
Actual	1	5	0	0	1	0	0	0	0	0	6	83.3%
	2	0	78	9	0	0	0	0	0	0	87	89.7%
	3	0	5	1724	13	0	0	0	0	0	1742	99.0%
	4	0	0	13	301	3	0	0	0	1	318	94.7%
	5	0	0	0	5	142	3	0	0	0	150	94.7%
	6	0	0	0	0	3	137	0	1	0	141	97.2%
	7	0	0	0	0	0	1	99	0	0	100	99.0%
	8	0	0	0	0	1	0	0	27	0	28	96.4%
	9	0	0	0	0	0	0	0	1	12	13	92.3%
	total	5	83	1746	320	149	141	99	29	13	2585	

Table 5. Accuracy for different soil type for test set after reducing number of majority samples

		Predicted										
	Class	1	2	3	4	5	6	7	8	9	total	accuracy
Actual	1	6	0	0	0	0	0	0	0	0	6	100 %
	2	0	82	5	0	0	0	0	0	0	87	94.3%
	3	0	22	1659	61	0	0	0	0	0	1742	95.2%
	4	0	0	0	314	2	0	0	0	2	318	98.7%
	5	0	0	0	6	140	2	1	1	0	150	93.3%
	6	0	0	0	0	3	138	0	0	0	141	97.9%
	7	0	0	0	0	0	2	98	0	0	100	98.0%
	8	0	0	0	0	1	0	0	27	0	28	96.4%
	9	0	0	0	0	0	0	0	1	12	13	92.3%
	total	6	104	1664	381	146	142	99	29	14	2585	

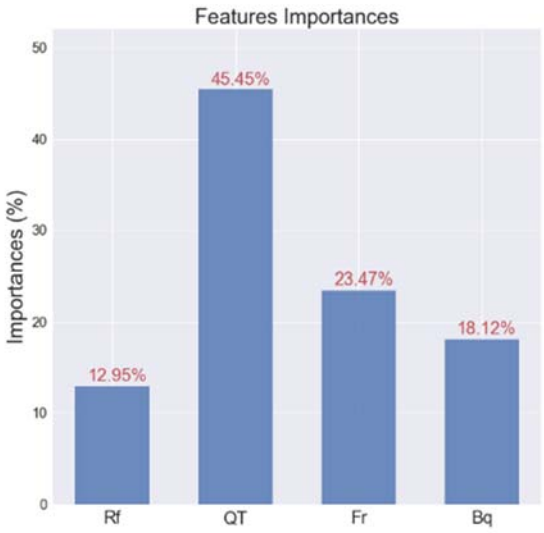

Figure 4. Importance of input parameters

These results show that the random forest method can be used for classification of soils, and that the accuracy can reach about 97% for the test set.

5. Separation of sand-like soils and clay-like soils

In many geotechnical applications, it is important to distinguish between clay-like soils and sand-like soils. Based on USCS classification method, if more than 50% of material is larger than No.200 sieve size (0.075mm), it is coarse grained soils. It is referred to sand-like soils in this study (Figure 5). If 50% or more of material is smaller than No.200 sieve size (0.075mm), it is fine grained soils. It is referred to clay-like soils in this study.

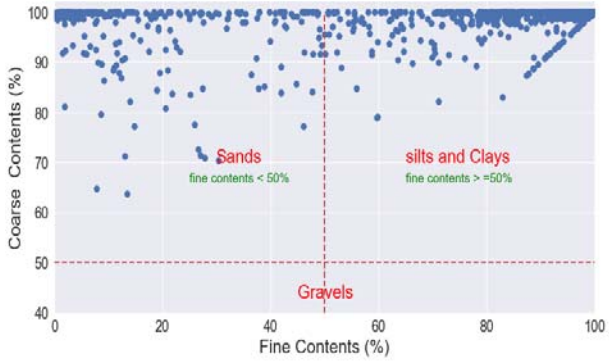

Figure 5. All data with grain size distribution tests

For clay like soils, it can be classified further to different groups based on Atterberg limit tests.

In this study, 1367 grain size distribution data and 1801 Atterberg Limit data with corresponding CPT data are collected from different locations. In total, there are 2792 pairs of CPTU data points and soil types, including 2555 clay like soils and 237 sand like soils.

The procedure used for classification is as follow:

(a) Random forest classifier (RFC) is applied to fit the classification model.

(b) The dataset is randomly split to two parts (Table 6): training dataset (80%) and test dataset (20%).

(c) The measured CPT data "q_c(kPa)", "u_2(kPa)" and f_s(kPa) are the input parameters in this model and the two classes Sand like soils and Clay like soils as output. Since normalized CPTU data cannot be obtained, the vertical effective stresses is also selected as one input parameters.

(d) Accuracy of 96 % for test data is obtained. But for sand like soil, accuracy is only 65.8%, see 0 for details. This is due to the unbalnced dataset, where clay-like soils are dominating over the sand-like soils. After using synthetic minority over-sampling technique (Nitesh,2002, [21]) to create synthetic (not duplicate) samples of the minority class and balance the minority class to majority class, accuracy of 80.5% can be obtained, see 0 for details.

(e) From Figure 6, the most important input parameters to soil classification are q_c with 46.9% importance and u_2 with 24.5% importance to soil classification. Proportion of various samples

Table 6. Proportion of various samples used in this study

Soil Behaviours types	Number of samples	Percentage of samples	samples for train data	samples for test data
Sand like soils	237	8.49%	196	41
Clay like soils	2555	91.51%	2037	518
summary	2792	100%	2233	559

Table 7. Accuracy for classification of sands and clays for test data before data balance

		Predicted			
	Class	Sand	Clay	Total	Accuracy
Actual	Sand	27	14	41	65,85 %
	Clay	8	510	518	98.46%
	Total	35	524	559	

Table 8. Accuracy for classification of sands and clays for test data after data balance

		Predicted			
	Class	Sand	Clay	Total	Accuracy
Actual	Sand	33	8	41	80,49 %
	Clay	25	493	518	95.17%
	Total	58	501	559	

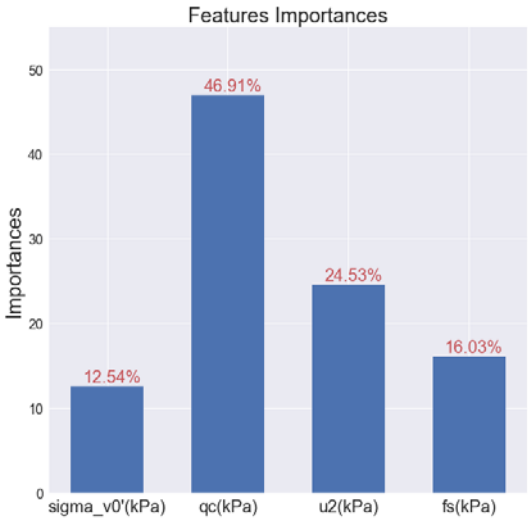

Figure 6. Importance of input parameters

6. *Subcategory of sand-like soils and clay-like soils*

In this section, the same dataset is used for soil classification, but this we subdivided the dataset into 7 soil types (contra 2 in the previous section). In total,1367 grain size distribution data and 1801 Atterberg Limit data with corresponded CPT data. There are 2031 pairs of CPTU data points and soil types, including 1794 clay like soils and 237 sand like soils.

Based on USCS, soils are classified as follows:

Gravels (GW,GP,GM and GC). There is no gravel data and gravel is not discussed in this study.

Sands (SW,SP,SM , SC and Sand with fines between 5% and 12%). For coarse grained sands, Cu and Cc are used to further classify sands to different types.

Cu is uniformity coefficient， $Cu = \frac{D60}{D10}$.

Cc is coefficient of curvature. $Cc = \frac{D30^2}{D60D10}$.

D_{60}, D_{30} and D_{10} are effective size of soil particles corresponding to 60%, 30%, 10% finer in the particle size distribution graph.

Silts and Clays (ML,CL,OL and MH,CH,OH). For this fine grained or clay like soils, atterberg limits data can be used for further classification (Figure 7)

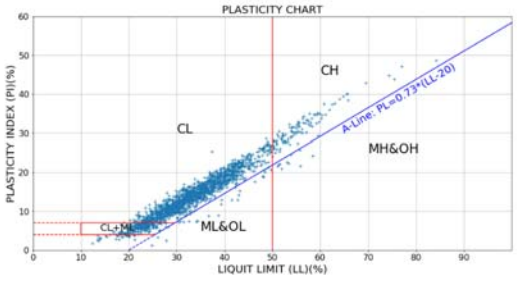

Figure 7. All data with atterberg limit test results

The dataset is classified as in Table 9. Since class MH and SW have smaller number of samples, we combine them with CM and SP respectively, see Table 9. Random forest classifier (RFC) is used to train the model and the dataset are randomly split to two parts: training dataset (80%) and test dataset (20%). The measured CPT data "q_c(kPa)", "u_2(kPa)" and f_s(kPa) as well as the vertical effective stress are selected as input features in the model.

Because the soil type CL is the majority class, the other soil classes have lower accuracy (Table 10). After using synthetic minority over-sampling technique (Nitesh,2002) to create synthetic samples of the minority class and balance the minority class to majority class, we can get better accuracy for the other soil classes, but it remains low (Table 11). One reason could be that the dataset used in this study is too small. It is recommended to test the model on a bigger dataset. The most important input parameter to soil classification is q_c with 33.1% importance (Figure 8).

Table 9. Dataset for classification

Dataset for classification		Combined dataset	
symbol	N_samples	symbol	N_samples
CH	99	CMH	114
CL	1459	CL	1459
CML	171	CML	171
MH	15	ML	50
ML	50	SMC	119
SMC	119	SPW	60
SP	57	S_fine_5_12	58
SW	3	summary	2031
S_fine_5_12	58		
summary	2031		

Table 10. Accuracy for classification of sands and clays before increasing number of minority samples

		Predicted								
	Class	CMH	CL	CML	ML	SMC	SPW	S_fine_5_12	total	accuracy
Actual	CMH	9	10	0	0	0	0	0	19	47,4 %
	CL	4	283	2	0	2	2	0	293	96,6 %
	CML	1	14	24	1	0	0	0	40	60,0 %
	ML	0	4	0	7	1	0	0	12	58,3 %
	SMC	0	1	0	0	13	3	6	23	56,5 %
	SPW	0	0	0	0	2	8	3	13	61,5 %
	S_fine_5_12	0	0	0	0	1	2	4	7	57,1 %
	total	14	312	26	8	19	15	13	407	

Table 11. Accuracy for classification of sands and clays after increasing number of minority samples

		Predicted								
	Class	CMH	CL	CML	ML	SMC	SPW	S_fine_5_12	total	accuracy
Actual	CMH	13	4	0	1	1	0	0	19	68,4 %
	CL	28	200	25	20	11	1	8	293	68,3 %
	CML	0	5	30	4	0	0	1	40	75,0 %
	ML	0	1	2	8	1	0	0	12	66,7 %
	SMC	0	1	1	0	12	3	6	23	52,2 %
	SPW	0	0	0	0	1	9	3	13	69,2 %
	S_fine_5_12	0	0	0	1	1	1	4	7	57,1 %
	total	41	211	58	34	27	14	22	407	

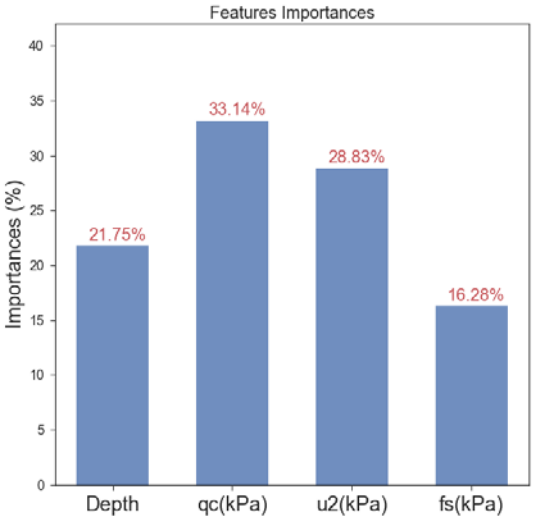

Figure 8. Importance of input parameters

7. Conclusions and discussions

Geotechnical datasets, including in situ CPTU data, grain size distribution, and atterberg limits from laboratory tests is used in this study for soil classification. Random forest method is applied to classify soils. Synthetic classes are used and ML technique is applied to mimic the soil behaviour type index and propose SBT classification based on CPTU. Grain-size distribution as well as Attemberg`s limits datasets are used to further refine the classification according to Unified Soil Classification System (USCS). Accuracy of the random forest method is generally high for CPTU based classification of soils in two main soil types, however it becomes lower for further refining the classification. This is due to small number of samples in the datasets used for further classification. We found data balance technique to be necessary in the analysis considering that one of the soil classes is largely dominated in our dataset.

References

[1] Garcia, S., Ovando-Shelley, E., Gutierrez, J., and Garcia, J. (2012). Liquefaction assessment through machine learning. 15 WCEE, Lisboa 2012.

[2] Ardakani, A. and Kohestani, V.R. (2015). Evaluation of liquefaction potential based on CPT results using C4.5 decision tress. Journal of AI and Data mining, Vol.3, No.1, 2015, 85-92.

[3] Ghaderi, A., Shahri, A.A., and Larsson, S. (2018). An artificial neural network based model to predict spatial soil type distribution using piezocne penetration test data. Bulletin of engineering gology and the environment. Published online: 15 October, 2018.

[4] Tsiaousi, D., Travasarou, T., Drosos, V., Ugalde, J. and Chacko, J. (2018). Machine learning application for site characterization based on CPT data. Geotechnical earthquake engineering and soil dynamics, V GSP 293, 461-472.

[5] Goh A. T. C. 1995b. Empirical design in geotechnics using neural networks. Geotechnique.

[6] Maizir, H., Nurly Gofar and Khairul Anuar Kassim. 2015. Artificial Neural Network Model for Prediction of Bearing Capacity of Driven Pile. Jurnal Teoretis dan Terapan Bidang Rekayasa Sipil.

[7] Mazaher, A. and Mazaher, B. 2016. Determination bearing capacity of driven piles in sandy soils using Artificial Neural Networks (ANNs). IOSR Journal of Mechanical and Civil Engineering

[8] Shahin M. A., Maier H. R. and Jaksa M. B. 2002. Predicting settlement of shallow foundations using neural networks. J. Geotech. & Geoenv. Eng.

[9] Alkroosh I., Nikraz H. 2011. Simulating pile load-settlement behavior from CPT data using intelligent computing. Central European Journal of Engineering.

[10] Cho S.E. 2009. Probabilistic stability analyses of slopes using the ANN-based response surface, Computers and Geotechnics. Vol. 36.
[11] Peng, M., Li, X.Y., Li, D.Q., Jiang, S.H. and Zhang, L.M., 2014. Slope safety evaluation by integrating multi-source monitoring information. Structural Safety.
[12] Bhargavi and Jyothi (2011). Soil classification using data mining techniques: A comparative study. International Journal of engineering trends and technology, July to August Issue 2011: 55-59.
[13] Robertson PK, Campanella RG, Gillespie D, Greig J (1986) Use of piezometer cone data. In-Situ '86 use of in-situ testing in geotechnical engineering, GSP 6, ASCE, Reston, VA, specialty Publication, pp 1263–1280.
[14] Robertson PK (1990) Soil classification using the cone penetration test.Can Geotech J 27(1):151–158
[15] Robertson PK (2016) Cone penetration test (CPT)-based soil behaviour type (SBT) classification system- an update. Can Geotech J 53:1910–1927. https://doi.org/10.1139/cgj-2016-0044
[16] Bhattacharya B, Solomatine, D.P., (2006). Machine learning in soil classification. Neural Networks 19, 186-195
[17] Ku, C.S., Juang, C.H. and Ou C.Y. (2010). Reliability of CPT Ic as an index for mechanical behaviour classification of soils. Geotechnique. 60. No.11, 861-875.
[18] Robertson PK, Wride CE. Evaluating cyclic liquefaction potential using the cone penetration test. Cana-dian Geotechnical Journal. 1998; 35: 442–459.
[19] Jefferies MG, Davies MP. Use of CPTU to estimate equivalent SPT N60. Geotechnical Testing Journal. 1993; 16(4): 458–468.
[20] Jiang, R., W. Tang, X. Wu, and W. Fu (2009). A random forest approach to the detection of epistatic interactions in case-control studies. BMC bioinfor-matics 10 (1), 1.
[21] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer (2002) SMOTE: Synthetic Minority Oversampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321–357, https://arxiv.org/pdf/1106.1813.pdf
[22] Buitinck L., Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt and Gaël Varoquaux (2013). API design for machine learning software: experiences from the scikit-learn project. European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases. 1 September 2013.