CONCEPTUAL CLUSTERING FOR THE GEOTECHNICAL DATA ANALYSIS

Piotr Bilski and Simon Rabarijoely

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

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

Abstract

The paper presents the application of the conceptual clustering algorithm for the geotechnical data analysis. First, the problem of the geotechnical layers identification for the construction engineering purposes is introduced. Then, the idea of the computer-based data analysis and its advantage over the traditional methods is explained. The proposed algorithm of conceptual clustering, adjusted for the continuous data processing is presented in detail. Its general steps, used data structures and the tuning procedure of the approach parameters to obtain the optimal results are explained. After presenting the investigation site and the available data sets (obtained from the geotechnical probes used in the in situ tests) for experiments, the results are discussed and conclusions drawn. The comparison between the reference profiles and the ones obtained thanks to clustering are provided, showing advantages of the proposed approach. The possible applications of the method to create automated soil profiles and calculate geotechnical indexes are provided as well.

Keywords: unsupervised learning, geotechnical investigation, soil profiles

1. Introduction

Practical aspects of the civil engineering include multiple applications regarding construction of buildings and structures important for the environment and human society (such as dams, bridges, etc.). The most important operation during the construction process is the design of the foundations, based on which the remainder of the structure is established. To propose the optimal design, the detailed knowledge about the structure of the soil below is required. During recent years, multiple invasive and non-invasive methods of determining the soil structure in the ground were proposed. The former (including drilling boreholes and geotechnical probes) are currently established and widely used during civil engineering operations (Marchetti, 2010), while the latter (such as radars) are in the experimental stage (Huisman, 2003). The boreholes give the most accurate information about the soil characteristics, but require time-consuming and expensive laboratory tests on the soil extracted from the ground. On the other hand, the measurement data acquisition by probes (such as Cone Penetration Test – CPT or Dilatometer of Marchetti Test - DMT) is executed much faster (up to twenty minutes) in situ. Therefore they are used as the support for the boreholes investigation, minimizing its usage and making the procedure less expensive.

To interpret the measurement results from probes, human expert knowledge supported by nomograms (charts connecting measured quantities with soil categories – Figure 1) is required. Unfortunately, the content of the diagrams depends on the location where the measurements are taken. Currently used charts were created for northern America and western Europe, which makes their application problematic. First, the classification accuracy may be lower because the particular soil type reacts to the probe differently, depending on its geological history (for example, whether it was subjected to the glacier or not). Second, the nomograms provide only the general information about the soil type, according to the Eurocode 7 and 8 classification rules (Frank, 2008). They disregard the internal inconsistencies in each soil type, which is caused by the overlapping of neighbouring categories. This means that inside the particular soil type multiple subcategories may exist, characterized by different physical traits. For these reasons, the application of computer data processing methods is justified and
can lead to the new information about the geotechnical structure, important in some applications (such as complex structures).

![Figure 1: The example of the chart for the CPT probe with two collected parameters parameters on the axes: cone resistance and friction ratio (Robertson, 1983)](image)

The paper presents the implementation of the conceptual clustering algorithm to the analysis of geotechnical data collected by the probes. This is a new version of the algorithm, originally proposed to work with discrete data (Fisher, 1987). Its advantage over other approaches (such as graph (Bilski, 2009), nearest neighbour (Frost, 2000) or c-means clustering (Goktepe, 2005), applied to the described domain, is the knowledge representation in the form of complexes, easily converted to rules. This allows for creating the artificial intelligence identification module with the In Section 2, the process of measurement data acquisition and forming the sets for processing is presented. Section 3 introduces the algorithm, its steps, operations and used data structures. In Section 4 the location for the experiments is described. Section 5 contains experimental results and discussion. In Section 6 conclusions and future prospects are considered.

2. **Principles of geotechnical investigation**

The description of the data acquisition methodology is important to understand the origin and structure of the data sets, processed by the algorithms. The generic scheme of the expert system is in Fig. 2. First, the data sets are acquired by the selected hardware. In this paper two sources of data are considered: CPT and DMT probes. Both devices are slowly pushed into the ground with the constant speed (usually 2cm/s) and provide the measurements at the particular depth interval (every 20 cm). In the CPT investigation two parameters are measured: the resistance of the cone inserted into the ground – \( q_c \) and the resistance of the friction along the sleeve of the cone - \( f_s \). In the DMT investigation three parameters are collected, although in the presented research only two are used: pressure A (taken at the first contact of the soil with the probe’s membrane, and B (measured after pushing this membrane with the air transported by the internal pneumatic tubing). Multiple extensions of the basic probes (SDMT or CPTU) can be used, providing more measured parameters, determining the size of the created data set.

![Figure 2: Structure of the data processing methodology for the geotechnical data analysis](image)
From the data acquisition process the data sets are obtained, having the form of the $n \times m$ table, where $n$ is the number of depths, at which the measurements are taken (i.e. the number of examples $e_i$) and $m$ is the number of measured parameters $p$ (in this work, only two). Additional columns are added to indicate the index of the example and the depth. The example of the data set from the DMT investigation is in Table 1. Only the two latter columns are processed by the intelligent algorithm, but the depth is also used during the soil profile generation.

<table>
<thead>
<tr>
<th>Ident.</th>
<th>Depth [m]</th>
<th>Pressure A [bar]</th>
<th>Pressure B [bar]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.6</td>
<td>0.20</td>
<td>1.10</td>
</tr>
<tr>
<td>2</td>
<td>1.8</td>
<td>0.20</td>
<td>1.10</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
<td>0.24</td>
<td>1.01</td>
</tr>
<tr>
<td>4</td>
<td>2.2</td>
<td>0.35</td>
<td>1.13</td>
</tr>
</tbody>
</table>

The measured parameters have the form as in Figure 3. It is assumed that subsequent soil categories are distinguishable based on their values. The task may be difficult because of slight changes between readings at every depth. The second reason is the overlapping neighbouring soil types, which means there are no abrupt changes in values, they change gradually. Additionally, the probe may encounter the rock in the soil, which is indicated by the high spike. Multiple abrupt changes in the waveform obtained from measurements may be eliminated by the denoising operation, but this may also delete important variability from data, therefore it should be used carefully.

The second step in the methodology from Figure 2 is the clustering, i.e. the process of unsupervised learning aimed at finding the dependencies and similarities in data to group them into categories. Multiple algorithms can be applied here, differing in the form of stored knowledge and the definition of similarity between the examples. The knowledge extracted from data can be used in three ways:

- To automatically generate the soil profile, i.e. the diagram connecting illustrating the location of particular soil categories at particular depths. This is the fundamental application of measurement data analysis, useful for both construction and science purposes.

![Figure 3: Measured parameters in the DMT (a) and CPT (b) investigation](image-url)
To provide data for the calculation of geotechnical indexes, such material index or pore pressure index. They are important for the construction of the building structure foundation. To store knowledge about the characteristics of the subsequent categories and use it to identify categories in other locations.

The latter application is the third step in the scheme from Figure 2 and is potentially the most challenging, as the same soil categories in different locations may be represented by different values of measured parameters. Therefore knowledge acquired from clustering can be applied to the soil identification only within the limited scope. This paper focuses on the knowledge extraction from the geotechnical data to create profiles.

3. Conceptual clustering algorithm

The conceptual clustering algorithm is presented in Figure 3. It processes the data set $D$ to produce a set of $k$ categories $C$. The scheme operates in the loop, creating complexes covering examples from $D$. The idea is to create $k$ complexes covering exclusively examples from $D$ and having the optimal structure. All examples belonging to the category covered by the particular complex should be as similar to each other as possible and maximally dissimilar to any other example from $D$. The complex is a set of conditions the example from $D$ must meet to be covered by it. Because the processed data is continuous, the conditions have the inequality or interval form. The first operation (select_seed) is executed before the main loop and consists in randomly selecting $k$ examples that will be the centres (seeds) $s$ of the groups covered by the constructed complexes. The latter are built during the specialize_complex operation, i.e. specializing the initially universal complex to cover the centre $s$ and other examples from the set $D$, but not any example already covered by other complex. The result of the operation is the set of complexes $C$ and the set of examples covered by them $T$. To ensure that, the ensure_separation procedure is executed, which modifies every complex so it does not cover any examples covered by other complexes. This concludes the categories generation. However, the clusters may not be optimal, because they were created around the centres selected randomly. Therefore each cluster must be checked according to the criterion of the optimal distance between all members of the group and its centre (upgrade_seeds procedure). For the newly selected centres the procedures of the complex specialization and separation are repeated during the next iteration of the main loop. The process is repeated until the stopping criterion is not met (i.e. the quality of clusters does not improve – see Section 3.2).

$C \leftarrow \text{conceptual_clustering}(D, k)$

$s \leftarrow \text{select_seed}(k, D)$

while (stopping_criterion_not_met)

$(C, T) \leftarrow \text{specialize_complex}(s, D)$
$(C, T) \leftarrow \text{ensure_separation}(C, D)$
$s \leftarrow \text{upgrade_seed}(s, D)$

end while

while ($D \setminus T \neq \emptyset$)

$s \leftarrow \text{select_seed}(1, D \setminus T)$
$(C, T) \leftarrow \text{add_complex}(s, D, T)$
$(C, T) \leftarrow \text{ensure_separation}(C, D)$

end while

$K \leftarrow \text{generate_clusters}(C, D)$

Figure 4: Conceptual clustering algorithm

It is possible that after the cluster generation, some examples from $D$ are still not covered. For the remaining examples the additional seeds are selected and for them the new complexes created (add_complex and ensure_separation procedures). The first operation is similar to specialize_complex, but it focuses on examples not covered so far. These operations are repeated...
subsequently until no examples from $D$ remain uncovered. This way some number of additional clusters $l$ is created, beyond the initially selected $k$, depending on the ability to cover all data from $D$. After the algorithm completion, the clusters $K$ can be defined as sets of examples covered by the subsequent complexes (procedure generate_clusters).

The algorithm is stochastic, i.e. the obtained result depends on the starting point (the selection of the initial seeds). The general scheme requires explanation of detailed steps, which will be presented in following subsections.

3.1. Data structures

The algorithm uses two structures, storing information about the complexes and clusters. The complex is the set of conditions that must be met to assign the example to the particular category.

Application of the algorithm to the continuous data requires using inequality (1a) and interval (1b) selectors:

$$p_j(e_i) > \theta; \quad p_j(e_i) < \theta \quad (1a)$$

$$p_j(e_i) \in (\theta_1, \theta_2); \quad \theta_2 > \theta_1 \quad (1b)$$

Here the selectors require the example $e_i$ to have the value of the parameter $p_j$ greater or lesser than the predefined threshold $\theta$, or to have its value within the borders defined by $\theta_1$ and $\theta_2$. The complex may contain such conditions for each parameter, with the maximum value of $m$ selectors. The complex is stored in the structure as in Figure 5, where each entry points at the set of selectors (indicating the corresponding parameter) and the selector contains the two- or four-element list, depending on the type of condition: (1a) or (1b). The first and the third value in the list determines the comparison operator: “1” means “greater than”, while “2” stands for “lesser than”, defining the borders of the intervals. Note that the in the presented case the complex stores two selectors maximum, but in another application this number may be greater.

![Figure 5: Structure of complexes](image)

The second structure describes the clusters, storing two types of information: the centre, being the geometric middle point for all examples being the part of the cluster $K_i \in K$ and the list of examples belonging to it. The calculation of the single coordinate (for the selected parameter) in expressed by (2). This structure is used during the profile generation, where three types of information are needed: the index of the example, the depth at which it was acquired (both taken from Table 1) and the number of the category (taken from $K$).

$$p_{ci} = \frac{1}{|K_i|} \sum_{j=1}^{|K_i|} \left\{ p_j(e_i) : e_i \in K_i \right\} \quad (2)$$

3.2. Algorithm operations
The algorithm performs two important operations on data, which require detailed description. The first one is the initial selection of seeds, before the first iteration of the main loop. The total random selection of \( k \) examples is not the optimal choice, as they can be selected close to each other, making the uniform covering of examples from \( D \) impossible. Therefore all examples are first divided into \( k \) intervals, according to their distance from the reference point (here \([0 \ 0]\)). Then, seeds are randomly selected from these groups. The division is based on the dynamic range of the Euclidean distance \( d_E \) (3), from which the length of the interval is calculated (4). The operation is illustrated in Figure 6.

\[
d_k(e_i) = \sqrt{\sum_{j=1}^{m} p_j^2}
\]

\[
l_d = \frac{\max(d_E) - \min(d_E)}{k}
\]

This distance is also used during the cluster generation and determines the index assigned to the cluster. The category with the centre closer to the point \([0 \ 0]\) always obtains lower index. This way the obtained profiles are repeatable.

\[\begin{array}{cccccc}
\text{Interval 1} & \text{Interval 2} & \text{Interval 3} & \ldots & \text{Interval } k \\
\hline
\min(d_E) & \min(d_E)+l_d & \min(d_E)+2l_d & \min(d_E)+3l_d & \max(d_E)
\end{array}\]

Figure 6 : Division of examples from \( D \) for the initial random seed selection

The second operation is the complex specialization (Figure 7), implemented in specialize_complex, add_complex and ensure_separation procedures. It is the core of the proposed method, determining the form of produced complexes. The processed data structures are temporary complexes \( V \) of the same form as in Figure 5. The input parameter is the positive seed \( s_p \), i.e. the example selected from the set \( s \) that should be covered by the resulting complex. The first step is the generation of the universal complex, which covers all examples (\( U \)). Then it is iteratively specialized, i.e. the new complexes \( V \) are created based on the previous ones, with modified selectors such that they cover the positive seed and not negative seeds. For every seed new thresholds \( \theta \) (one for each parameter) are created to exclude subsequent negative seeds from covering. Because after the intersection multiple complexes are created, their number will grow with each negative seed processed. To limit them, the \( r \) parameter is introduced, determining how many temporary complexes from \( V \) are passed to the next iteration. The quality of the complex is introduced in section 3.3. Finally, one complex \( C \) is returned.

\[\begin{align*}
C &\leftarrow \text{complex_specialization}(s_p, s, D) \\
V &\leftarrow U \\
\text{for each } s_n: s \in s; s \neq s_p \\
& \quad V \leftarrow V \cap \{p_j(s_n) \in T \land p_j(s_p) \notin T\} \\
& \quad V \leftarrow \text{select_best}(r, V) \\
\text{end for} \\
C &\leftarrow \text{select_best}(1, V)
\end{align*}\]

Figure 7 : The complex_specialization procedure

During the specialization four cases are possible:

- To the existing complex the new inequality selector (1a) is added, causing the expansion of the middle element in the structure from Figure 5;
- The existing inequality selector is specialized by modifying the value of the threshold: \( \min(\theta_{old}, \theta_{new}) \) for the “lesser than” and \( \max(\theta_{old}, \theta_{new}) \) for the “greater than” condition (see (1a))
The existing inequality selector is changed into the interval one by adding the second condition and creating the interval \((1b)\). If the condition \("\theta_2 > \theta_1"\) is not met, the selector becomes empty, i.e. it does not cover any example from \(D\). This causes elimination of the complex from \(V\).

The existing interval selector is narrowed, obtaining one new border value. As above, if the condition \("\theta_2 > \theta_1"\) is not met, the complex is discarded.

After each modification the set \(T\) of examples covered by the complex is updated. The candidate thresholds are calculated as the middle values of the selected parameter for the positive and negative seed:

\[
\theta = \frac{p_j(s_p) - p_j(s_n)}{2}
\]  

(5)

The procedure from Figure 7 applied to the ensure_separation is slightly different from the introduced scheme. The set of negative seeds is exchanged with the set of examples covered by any complexes. Such examples are excluded from the covering.

3.3. Algorithm parameters

The stochastic nature of the algorithm makes it susceptible to the parameters determined by the operator. The following parameters must be defined:

- The number of clusters to generate, which determines the number of seeds selected in each iteration of the procedure from Figure 4. The experiments show that the reasonable number should be provided by the geotechnical engineer, for instance based on the analysis of the data from boreholes. The resulting number of categories is usually greater than provided to the input of the algorithm because some examples will not be covered by created complexes (for instance the ones related with the rocks).

- The stopping criterion of the procedure from Figure 4. Among other possible approaches (such as the scattering), the quality measure for the obtained clustering preferring simple complexes was proposed. It aims at selecting the smallest number of as simple selectors as possible (with minimal number of conditions). The following formula is used:

\[
f_q(C) = \sum_{i=1}^{k} |s \in C_i|
\]  

(6)

The clustering is then the process of minimizing the function (6). If there is no improvement between neighbouring iterations, the main loop is terminated.

- The quality of the candidate complexes \(V\), used to select \(r\) best ones for the next iteration during the complex specialization. The scattering criterion is used here, i.e. the minimal number of examples potentially covered by the candidate complex. Because the continuous data are processed, the theoretical number of examples in the dynamic range is infinite. To avoid this problem, the specified number of examples existing in each variability area is assumed for each measured parameter. This allows for defining the following function (where the symbol \("\succ"\) means the complex \(V\) covers examples \(e\)):

\[
f_q(V) = |e \notin D : V \succ e| - |e \in D : V \succ e|
\]  

(7)

The complexes with the smallest scattering possible are preferred.

4. Experimental results

The algorithm described in section 3 was applied to the data obtained from CPT and DMT probes investigation performed at Warsaw University of Life Sciences (WUoLS) campus. Before the didactic buildings were established, the thorough geotechnical measurements were taken, using boreholes and
probes. Based on the collected data, profiles were generated. They were used in the presented experiments as the reference information, which could be compared to the results of the conceptual clustering. The locations where measurements were taken are in Figure 8, while the example of the obtained profile is in Figure 9.

![Figure 8: Location of the building No. 34 in the WUoLS campus (a) and probings of CPT (▼) and DMT (■) tests](image1)

![Figure 9: Example of the profile generated after collecting data from probes (here locations CPT2 and DMT3 are present) and boreholes (OW28 and OW29)](image2)

The initial analysis of the profiles shows three main soil categories at the depth down to 10 m: sandy clay, brown sandy clay and grey sandy clay. Between the first two the thin layer of the fine sand may be encountered. This way the initial number of clusters to generate was set to 3 and 4, although greater values were tested as well. The depths of the borders between neighbouring layers are different for subsequent locations of probes application sites, which may be the source of inconsistencies between the traditional profile and the one obtained using the conceptual clustering. Note that the profile does not consider intermediate sub-layers between the main layers, which may also be the source of inconsistencies.
In Figure 10 clustering results for the DMT data (from Figure 3a) are presented, depending on the number of expected categories. The dotted lines indicate borders between the main categories as suggested by the profile created using the traditional method. The algorithm in all cases generated more categories than initially expected. This is because the single-example categories (rocks) were also detected. The profile for 3 categories is too generic, as there were problems with distinguishing brown sandy clay from the grey sandy clay. The better results are for 4 and 5 categories, although the grey sandy clay is represented by multiple categories (which means this layer is not uniform). The
Experiments were repeated after denoising the original waveforms with the median filter (of the length 3), leading to the smaller number of categories (as one-category clusters are no longer present). Alternatively, Figure 11 shows results for the CPT data (from Figure 3b). Here there is no fine sand category and sandy clay occupies more space. Increasing the number of initial categories leads to the greater diversity within the main categories (grey sandy clay is decomposed into three subcategories). The denoising procedure has the same impact as in the DMT investigation, eliminating one-example categories and decreasing their overall number.

5. Conclusions

The proposed method is a new approach to the clustering of continuous data. Its main advantage is the form of storing knowledge about the clusters, which can be easily transformed into the premises part of the rule. This way the conceptual clustering may provide the easy and comprehensible knowledge for the rules-based expert system. The main problem is the selection of optimal values for algorithm parameters (such as the number of generated categories or the stopping criterion), determining the clustering result. The reasonable strategy is to start with small values, and, if required, increasing them until the desired result (verified by the geotechnical engineer) is obtained. Also, depending on the algorithm application, various sets of parameters may be useful (for instance, whether the expected result is the general soil profile, or the more detailed diagram of dependencies between soil categories). The open question is the versatility of knowledge extracted from data sets D. To answer it, more experiments with various measurement data from different locations are required.

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

References


