A Hierarchical-cluster-based Method for Identifying the Outliers from the Test Data of Solid Mechanics Provided for 2018 TC304 Student Contest

GAO Jingze[1] , WANG Guoliang[2] , DENG Renming[3] , DENG Laiming[4]

[1]*School of Civil Engineering, Harbin Institute of Technology. Email:2638606109@qq.com*
[2]*School of Civil Engineering, Harbin Institute of Technology. Email:3466524823@qq.com*
[3]*School of Civil Engineering, Harbin Institute of Technology. Email:673395360@qq.com*
[4]*School of Civil Engineering, Harbin Institute of Technology. Email:2409729176@qq.com*

Abstract: A hierarchical-cluster-based method is proposed for the identification of the possible outliers in the available experimental data. This method can divide the data of interest into different categories according to their similarity. Then the "abnormal" data that show different characteristics and form a category along can be identified. The presented method is applied to pick out the possible outliers from the test data of solid mechanics provided for 2018 TC304 student contest. To provide confidential result, the available and widely accepted correlation models among different pair of the given five variables including are referred. The "abnormal" data is identified for each pair of variables. To sum up the final results, four values corresponding to three different variables are predicted as outliers.

Keywords: hierarchical cluster, correlation, outlier identification

1. Introduction

There are often outliers in the experimental data. Wrong data records, instrument failure and many reasons can cause the outliers. Therefore, it is necessary to establish a method which can identify the outliers in the experimental data.

Therefore, an outliers identificaion model for soil mechanics experimental data involved in TC304 student contest was established.

This model adopts the hierarchical cluster as the basic method, using existing correlations for more effective variable selection. In this model, the detection model and validation model were established to identify the outliers in the contest more accurately.

Hierarchical cluster is an effective method to identify the outliers. Through the similarity measurement, the similarity between multidimensional data points with multiple variables in the contest can be calculated. These multidimensional data points are sorted from high to low according to the similarity degree. Based on the similarity degree, the multidimensional data points are connected step by step, then they are divided into several classes. On this basis, it is only necessary to introduce a rule to identify which one of the experimental data is outlier.

The variables for hierarchical cluster are the variables not only involved in the contest but conformed to the existing correlations.

In this paper, we proposed a detection model which combining some existing pairwise correlations between liquidity index, vertical effective stress, preconsolidation stress, remolded undrained shear strength, undrained shear strength with hierarchical cluster. We can pick out the suspicious experimental data from the results of this model. Then, all of these suspicious data were tested by validation model combining some existing multivariate correlations with hierarchical cluster. Finally, we can find out the outliers in the contest by synthesizing the results of these two models.

2. Detection Model

A detection model was built to pick out the suspicious data in the contest. It combines hierarchical cluster with existing pairwise correlations. Suspicious data points can be found by analyzing the similarity between data points formed by two variables which satisfying the existing pairwise correlations.

For convenience, we have numbered the data points in the contest. The result is shown in Table 1.

Table 1. Data points in the contest problem

Serial number	Liquidity index	Vertical effective stress(kPa)	Preconsolidation stress(kPa)	Remolded undrained shear strength(kPa)	Undrained shear strength(kPa)
1	0.98	3.7	13.87	0.88	5.95
2	1.31	7.4	12.95	0.59	4.29
3	1.78	13.87	9.25	0.39	4.07
4	1.51	17.57	17.57	0.39	5
5	1.31	21.27	45.12	0.39	5.95
6	1.34	24.05	21.27	0.59	6.43
7	1.63	27.75	24.05	0.39	7.62
8	1.42	31.45	24.97	0.68	16.74
9	2.52	35.14	29.6	0.68	7.86
10	1.27	39.77	29.6	0.78	12.38
11	1.21	44.39	30.52	0.88	13.1
12	1.38	49.02	36.07	0.98	13.81
13	1.45	51.79	55.49	1.18	17.38
14	1.51	58.27	60.12	1.37	13.1
15	1.22	61.97	48.09	0.98	18.57
16	1.18	66.59	72.14	0.88	17.14
17	0.93	71.21	97.11	1.18	26.19

2.1 Hierarchical cluster principle

The main purpose of hierarchical cluster is to divide data points into many categories. Data points in the same class have strong similarity, while data points in different classes have less similarity.

Hierarchical cluster can be divided into agglomerative hierarchical cluster and divisive hierarchical cluster. At the beginning of agglomerative hierarchical cluster, all of the data points are separated into one class. Then, the distance between the class and the class is calculated, and the two classed with the smallest distance are merged until all data points are combined into one class. In this paper, we choose agglomerative hierarchical cluster as the basic method we use.

Also, we choose Mahalanobis distance as the similarity measurement. Compared with the Euclidean distance, Mahalanobis distance has many advantages. Such as it can eliminate the influence of variable correlations and it is scale independent. Its definition is shown in equation 1.

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \Sigma^{-1}(\vec{x}-\vec{y})} \qquad (1)$$

where $\vec{x}$ and $\vec{y}$ is two samples in the sample population, Σ is the covariance matrix for the sample population.

When calculating the distance between classes, we use single linkage method. The definition of the distance between class m and class n is shown in equation 2.

$$D(m,n) = \min_{\overline{x_i \in m, y_j \in n}} \{d(\overrightarrow{x_i},\overrightarrow{y_j})\} \qquad (2)$$

2.2 Definition and detection algorithm of abnormal data points

In limited experimental data, normal data points which far away from other data points may be mistaken for outliers. In order to avoid the case above, we need to give an accurate definition to outliers.

In this paper, we defined those multidimensional data points which do not conform to existing correlations as outliers. It indicates that the outliers are not only far from other data points, but also do not conform to existing correlations. So, our variable selection for detection model must work based on existing correlations. As mentioned above, Mahalanobis distance can deal with variables that are related well.

The detection algorithm must serve to find out the outliers defined above. Therefore, we adopt a new algorithm to identify the outliers. First, selecting variables based on the existing correlations. Second, using mahalanobis distance to calculate the similarity between data points. Third, using the single linkage method to clustering. Fourth, removing a set of clustering which at the top of the clustering tree (Liang 2012). If the distance of this pair of classes more than twice the average of the distance between all of the pair of classes and only a single data point is released, marking this single data point as an abnormal data point. Fifth, repeating step 4 until the single data point no longer appears.

2.3 Existing pairwise correlations selected for modelling

The selection of existing pairwise correlations must meet two requirements. First, these existing pairwise correlations must cover all of the variables involved in the contest as far as possible. Second, these existing pairwise correlations must allow us to infer which variable of the abnormal data point is the outlier from the detection results.

Therefore, we selected four existing pairwise correlations for modeling, they are shown in equation 3-6. (Wood 1990, Wroth and Wood 1978, Ladd and Foott 1974, Ching and Phoon 2012)

$$\sigma_v' = \sigma_{v,L}' \,\mathrm{R}\exp\{[k-\log(R)]\times LI\} \tag{3}$$

$$S_u^{re} = S_{u,L}^{re}\,\mathrm{R}\exp[\log(R)\times -LI] \tag{4}$$

$$\ln(\frac{S_u}{P_a}) = a\ln(\frac{\sigma_p'}{P_a}) + b \tag{5}$$

$$\ln(\frac{S_u}{\sigma_v'}) = a\ln(OCR) + b \tag{6}$$

where $\sigma_{v,L}'$, $S_{u,L}^{re}$, R , k , a and b are the model constant.

2.4 Outliers analysis of the data in the contest

2.4.1 Detection model 1

Based on equation 3, we choose LI and $\ln(\sigma_v')$ to establish the detection model 1.

The clustering process is shown in Table 2. In the table, the clustering order is from top to bottom. And the clustering tree is shown in Figure 1.

We can get the average of the distance between all of the pair of classes by calculating the average value of the third column in Table 2. In detection model 1, it is 1.9802.

As it is shown in Table 2, Figure 1 and our detection algorithm, we can find out the points which serial number is 1 and 9 are the abnormal data point. They are shown in a scatter diagram, Figure 2. Obviously, they are different from other data points.

2.4.2 Detection model 2

Based on equation 4, we choose LI and $\ln(S_u^{re})$ to establish the detection model 2.

The clustering process is shown in Table 3. In the table, the clustering order is from top to bottom. And the clustering tree is shown in Figure 3.

We can get the average of the distance between all of the pair of classes by calculating the average value of the third column in Table 3. In detection model 2, it is 1.8243.

As it is shown in Table 3, Figure 3 and our detection algorithm, we can find out the point which serial number is 9 is the abnormal data point. It is shown in a scatter diagram, Figure 4. Obviously, it is different from other data points.

2.4.3 Detection model 3

Based on equation 5, we choose $\ln(\frac{S_u}{P_a})$ and $\ln(\frac{\sigma_p'}{P_a})$ to establish the detection model 3.

The clustering process is shown in Table 4. In the table, the clustering order is from top to bottom. And the clustering tree is shown in Figure 5.

We can get the average of the distance between all of the pair of classes by calculating the average value of the third column in Table 4. In detection model 3, it is 2.0976.

As it is shown in Table 4, Figure 5 and our detection algorithm, we can find out the points which serial number is 5 and 8 are the abnormal data point. They are shown in a scatter diagram, Figure 6. Obviously, they are different from other data points.

2.4.4 Detection model 4

Table 2. Clustering process for detection model 1

Steps	Serial number of class m	Serial number of class n	Distance
1	15	16	0.5273
2	5	6	0.5572
3	12	13	0.6065
4	14	20	0.7013
5	10	11	0.7395
6	21	22	1.0540
7	8	19	1.1683
8	18	23	1.2527
9	24	25	1.6263
10	4	7	1.6491
11	26	27	1.7265
12	17	28	2.3757
13	3	29	2.7037
14	2	30	3.3293
15	1	31	**4.3759**
16	9	32	**7.2897**
	Average distance		1.9802

Table 3. Clustering process for detection model 2

Steps	Serial number of class m	Serial number of class n	Distance
1	2	6	0.2137
2	11	16	0.2472
3	4	7	0.8564
4	15	19	0.9268
5	10	21	1.0126
6	3	20	1.1662
7	12	22	1.3234
8	5	23	1.4051
9	8	24	1.4338
10	18	26	1.4449
11	13	14	1.7706
12	1	27	1.8215
13	17	29	1.9251
14	28	30	2.0861
15	25	31	3.5682
16	9	32	**7.9871**
	Average distance		1.8243

Table 4. Clustering process for detection model 3

Steps	Serial number of class m	Serial number of class n	Distance
1	10	11	0.3607
2	12	18	0.9755
3	6	7	1.0055
4	14	16	1.3262
5	9	20	1.4086
6	2	4	1.4989
7	22	23	1.5093
8	13	15	1.5944
9	19	25	1.6323
10	1	24	2.1083
11	17	21	2.1725
12	3	27	2.1772
13	26	28	2.2482
14	29	30	3.3229
15	8	31	**4.5175**
16	5	32	**5.7039**
	Average distance		2.0976

Note: The origin data points in the contest are from class 1 to 17. When they synthesize new classes, the new classes have a serial number which from 18 to 32.

Table 5. Clustering process for detection model 4

Steps	Serial number of class m	Serial number of class n	Distance
1	12	15	0.3078
2	3	11	0.3080
3	6	7	0.4509
4	10	19	0.5265
5	18	21	0.7914
6	4	20	0.8667
7	14	16	0.9975
8	13	23	1.0212
9	9	25	1.3178
10	24	26	1.4677
11	17	27	1.8554
12	22	28	1.9726
13	2	29	2.9587
14	8	30	**4.6907**
15	5	31	**6.2815**
16	1	32	**7.4782**
	Average distance		2.0808

Note: The data which is underlined indicates that this data is more than twice the average distance.

Based on equation 6, we choose $\ln(\frac{S_u}{\sigma_v^{'}})$ and $\ln(OCR)$ to establish the detection model 4.

The clustering process is shown in Table 5. In the table, the clustering order is from top to bottom. And the clustering tree is shown in Figure 7.

We can get the average of the distance between all of the pair of classes by calculating the average value of the third column in Table 5.

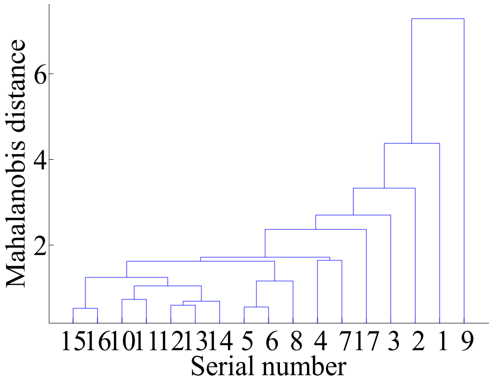

Figure 1. Clustering tree for detection model 1

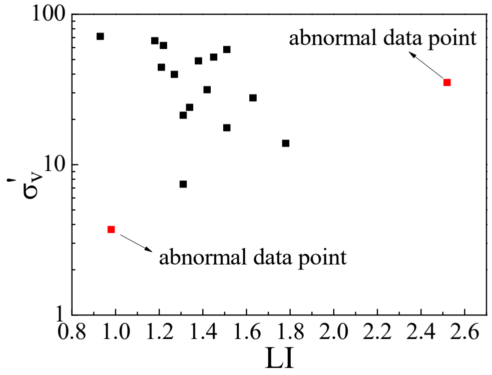

Figure 2. Scatter diagram for detection model 1

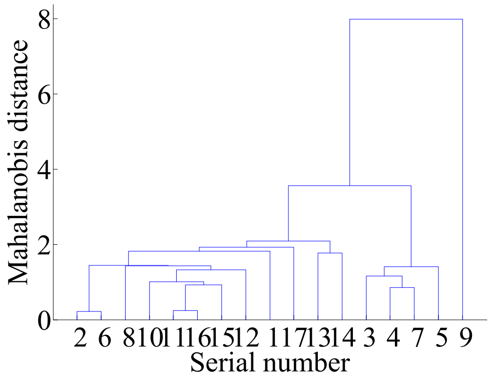

Figure 3. Clustering tree for detection model 2

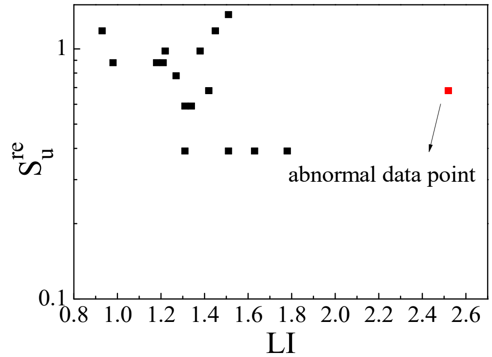

Figure 4. Scatter diagram for detection model 2

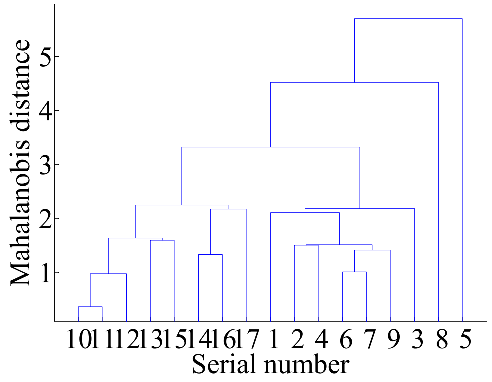

Figure 5. Clustering tree for detection model 3

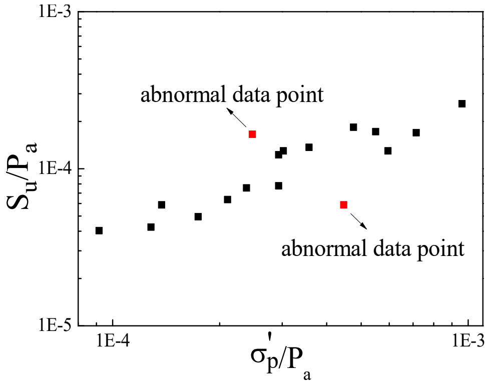

Figure 6. Scatter diagram for detection model 3

In detection model 1, it is 2.0808.

As it is shown in Table 5, Figure 7 and our detection algorithm, we can find out the points which serial number is 1,5 and 8 are the abnormal data point. They are shown in a scatter diagram, Figure 8. Obviously, they are different from other data points.

2.4.5 Analysis of results

We find that detection model 1 and detection model 2 all contain the liquid index and both of their data point which serial number is 9 is abnormal. So, we set the liquid index of the con-

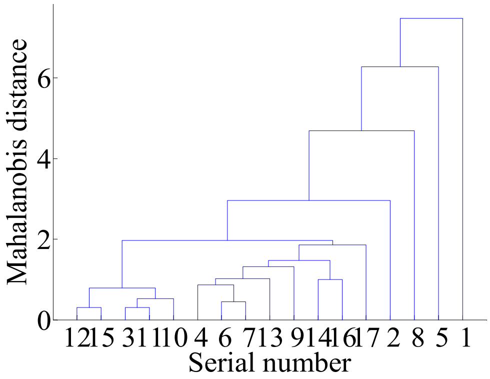

Figure 7. Clustering tree for detection model 4

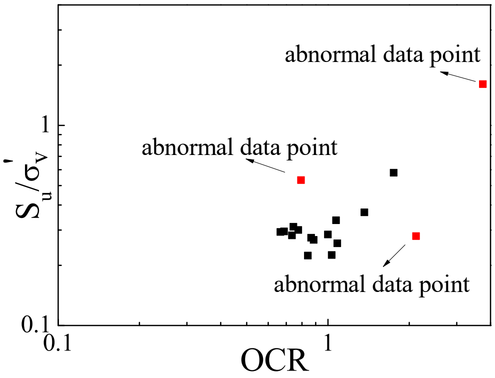

Figure 8. Scatter diagram for detection model 4

Table 6. Suspicious data based on detection models

Serial number	Liquidity index	Vertical effective stress(kPa)	Preconsolidation stress(kPa)	Remolded undrained shear strength(kPa)	Undrained shear strength(kPa)
1	0.98	**3.7**	13.87	0.88	5.95
2	1.31	7.4	12.95	0.59	4.29
3	1.78	13.87	9.25	0.39	4.07
4	1.51	17.57	17.57	0.39	5
5	1.31	21.27	**45.12**	0.39	**5.95**
6	1.34	24.05	21.27	0.59	6.43
7	1.63	27.75	24.05	0.39	7.62
8	1.42	31.45	**24.97**	0.68	**16.74**
9	**2.52**	35.14	29.6	0.68	7.86
10	1.27	39.77	29.6	0.78	12.38
11	1.21	44.39	30.52	0.88	13.1
12	1.38	49.02	36.07	0.98	13.81
13	1.45	51.79	55.49	1.18	17.38
14	1.51	58.27	60.12	1.37	13.1
15	1.22	61.97	48.09	0.98	18.57
16	1.18	66.59	72.14	0.88	17.14
17	0.93	71.21	97.11	1.18	26.19

test data point which serial number is 9 as a suspicious data. In the same way, we can set the vertical effective stress of the contest data which serial number is 1 as a suspicious data.

Also, we find that data points which serial number is 5 and 8 are abnormal by analyzing the result of detection model 3. But we cannot find out which one of preconsolidation stress and undrained shear strength caused these abnormal data points. It will be determined in the validation model. Now, we treat both of preconsolidation stress and undrained shear strength of data point 5 and 8 as suspicious data.

The result of detection model 4 proves the analysis above.

According to the analysis results, we obtained the suspicious data which underlined in Table 6. They will serve as the basis for the validation model.

3. Validation Model

A validation model was built to verify the result of the detection model. At the same time, it can solve some remaining problems in the detection model. It combines hierarchical cluster with existing multivariate correlations.

Due to the complexity of high-dimensional clustering, we only conduct qualitative analysis in validation model, unlike quantitative analysis in detection model.

3.1 Existing multivariate correlations selected for modelling

In order to solve the remaining problems of detection model and validate the result of detection model effectively, we selected three existing multivariate correlations to establish the validation model, they are shown in equation 7-9. (Ching and Phoon 2012)

$$S_t = LI^{0.413} \times (\frac{S_u^{re}}{P_a})^{-0.947} \times (\frac{\sigma_v^{'}}{P_a})^{0.581} \times 0.564 \qquad (7)$$

$$\frac{S_u}{\sigma_v^{'}} = S_t^{0.144} \times OCR^{0.810} \times 0.206 \qquad (8)$$

$$\frac{S_u}{\sigma_v^{'}} = LI^{-0.258} \times S_t^{0.208} \times OCR^{0.806} \times 0.177 \qquad (9)$$

Equation 7-9 contains all of the variables in the contest. Equation 7 does not contain preconsolidation stress, so it can deal with the remaining problem in the detection model.

3.2 Validation of suspicious data

3.2.1 Validation model 1

Based on equation 7, we choose $\ln(LI)$, $\ln(S_t)$, $\ln(\frac{S_u^{re}}{P_a})$ and $\ln(\frac{\sigma_v^{'}}{P_a})$ to establish the validation model 1.The clustering tree of validation model 1 is shown in Figure 9.

We find that the data point which serial number is 5 is normal by analysing the clustering tree of validation model 1. In this model, we do not set presolidation stress as a variable. It indicates that it is presolidation stress make data point 5 an abnormal data point. Of course, it is undrained shear strength make data point 8 an abnormal data point.

In this way, we have completely solved the remaining problem in detection model.

The following two validation models are used to verify the results obtained.

3.2.2 Validation model 2

Based on equation 8, we choose $\ln(\frac{S_u}{\sigma_v^{'}})$, $\ln(S_t)$ and $\ln(OCR)$ to establish the validation model 2. The clustering tree of validation model 2 is shown in Figure 10.

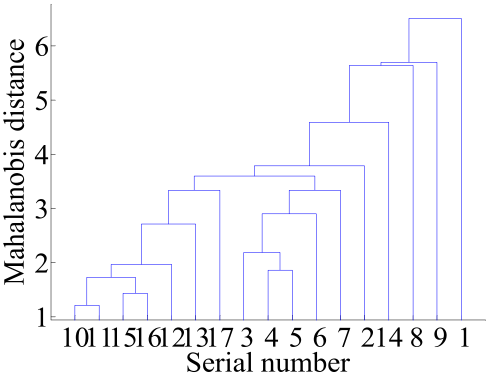

Figure 9. Clustering tree for validation model 1

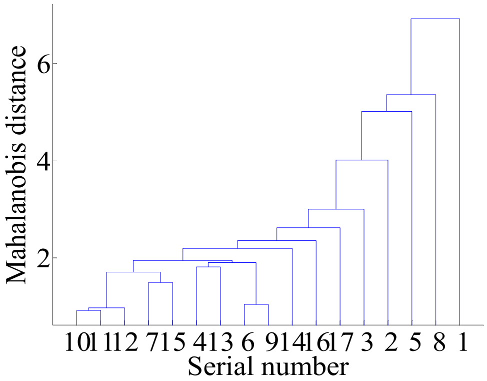

Figure 10. Clustering tree for validation model 2

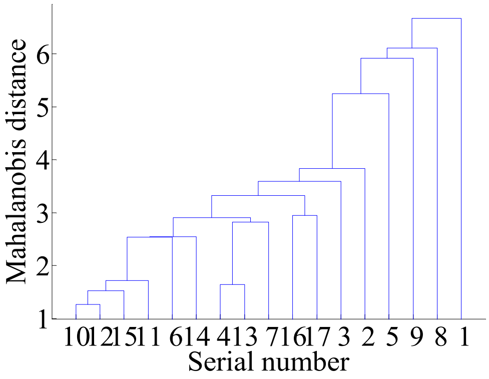

Figure 11. Clustering tree for validation model 3

We find that data point 1,5 and 8 are abnormal by analysing the clustering tree of validation model 2. In this model, we do not set

Table 7. Outliers based on our models

Serial number	Liquidity index	Vertical effective stress(kPa)	Preconsolidation stress(kPa)	Remolded undrained shear strength(kPa)	Undrained shear strength(kPa)
1	0.98	**3.7**	13.87	0.88	5.95
2	1.31	7.4	12.95	0.59	4.29
3	1.78	13.87	9.25	0.39	4.07
4	1.51	17.57	17.57	0.39	5
5	1.31	21.27	**45.12**	0.39	5.95
6	1.34	24.05	21.27	0.59	6.43
7	1.63	27.75	24.05	0.39	7.62
8	1.42	31.45	24.97	0.68	**16.74**
9	**2.52**	35.14	29.6	0.68	7.86
10	1.27	39.77	29.6	0.78	12.38
11	1.21	44.39	30.52	0.88	13.1
12	1.38	49.02	36.07	0.98	13.81
13	1.45	51.79	55.49	1.18	17.38
14	1.51	58.27	60.12	1.37	13.1
15	1.22	61.97	48.09	0.98	18.57
16	1.18	66.59	72.14	0.88	17.14
17	0.93	71.21	97.11	1.18	26.19

liquidity index as a variable. Therefore, this analysis result conforms to our conclusions in detection model and validation model 1 perfectly.

3.2.3 Validation model 3

Based on equation 9, we choose $\ln(LI)$, $\ln(S_t)$, $\ln(\frac{S_u}{\sigma'_v})$ and $\ln(OCR)$ to establish the validation model 3.The clustering tree of validation model 1 is shown in Figure 11.

We find that data point 1,5,8 and 9 are abnormal by analysing the clustering tree of validation model 3. In this model, we set all of the five variables in the contest as variables. Therefore, this analysis result conforms to our conclusions in detection model and validation model 1 perfectly.

4. Conclusion

We finally find the outliers as shown in Table 7 by synthesizing the results of detection model and validation model.

References

Ladd, C.C., and Foott, R. 1974. New design procedure for stability in soft clays. *Journal of the Geotechnical Engineering Division*, ASCE, **100**(7); 763-786

Wood, D.M. 1990. Soil Behaviour and critical state soil mechanics. *Cambridge University Press*, Cambridge, UK.

Wroth, C.P., and Wood, D.M. 1978. The correlation of index properties with some basic engineering properties of soils. *Canadian Geotechnical Journal*, **15**(2): 137-145

J. Ching., K.-K. Phoon. 2012. Modeling parameters of structured clays as a multivariate normal distribution. *Canadian Geotechnical Journal*, **49**: 522-545

LIANG. 2012. A Hierarchical-clustering-based for identifying the first n global isolated points (in Chinese). *Computer Engineering and Application*, **48**(9): 101-103

Acknowledgements

First of all, we would like to thank our advisor, associate professor YU Xiaohui（于晓辉）. Teacher YU has devoted great efforts to the method selection, creative inspiration and wording and format of the article. Teacher YU's guidance and encouragement are the motivation for us to persist.

We would like to thank our soil mechanics teacher QI Jialian（齐加连）. Teacher QI's guidance on soil mechanics has greatly helped us to understand the contest problem.

We would like to thank our senior fellow apprentice HE Zihan（何子涵）. Senior fellow apprentice’s guidance in statistics greatly helped us build the model.