FEATURE SELECTION AND GROUPING OF CULTIVATION ENVIRONMENT DATA TO EXTRACT HIGH/LOW YIELD INHIBITION FACTOR OF SOYBEANS

Katsuhiro Nagata[1], Midori Namba[1], Seiichi Ozawa[1], Yuya Chonan[2], Satoshi Hayashi[2], Takuji Nakamura[2], Hiroyuki Tsuji[2], Noriyuki Murakami[2], Ryo Nishide[3] and Takenao Ohkawa[1]

[1]Kobe University, Japan

[2] Hokkaido Agricultural Research Center, NARO, Japan

[3] The Center for Data Science Education and Research, Shiga University, Japan

nagata@cs25.scitec.kobe-u.ac.jp, namba@cs25.scitec.kobe-u.ac.jp, ozawasei@kobe-u.ac.jp, ohkawa@kobe-u.ac.jp, y.chonan@affrc.go.jp, hayashis@affrc.go.jp, takuwan@affrc.go.jp, tuzihiro@affrc.go.jp, noriyuki@affrc.go.jp, ryo-nishide@biwako.shiga-u.ac.jp

ABSTRACT

This work aims to extract high or low yield factors by analyzing soybean cultivation data and its cultivation environment. In this study, methods for using soybean data and investigating the cause of yield affected by the cultivation environment are proposed. As the soybean is affected by surrounding environment at each growing stage, the cultivation environment is also examined at each corresponding stage by dividing environment data. Qualitative values are generated at each stage, and the plants' condition for each stage is simply expressed because the slight changes of the environment do not affect much the growth of soybean. Then, similar cultivation environments at each growing stage are grouped by clustering. In the grouping, features of cultivation environment are selected to eliminate groups that have the few number of soybean fields whose environments are similar, and features with low possibility to be classified as either the high and low yield factors are removed. Thus, a distinctive group that is inclined toward either high yield or low yield of soybean cultivation has been identified. The cultivation environment that may affect the yield of soybeans has been revealed.

Keywords*:* soybean, feature selection, clustering.

1. INTRODUCTION

In recent years, there is a high demand in agricultural efficiency due to the aging society and a lack of successor in Japan. To solve this problem, data mining using past data is increasingly used (Harel, D. et al, 2014). Not many methods have been established to improve the amount of yield and quality of soybeans (Japan Agricultural Development and Extension Association (JADEA), 2012), which is the target of our work. Previous studies used LCM (Uno, T. et. Al, 2004), or tried to discover a frequent pattern from soybean cultivation data (Umejima, K. et. al, 2016, Namba, M. et. al, 2017). However, these studies did not investigate the dual effects of cultivation environments at multiple stages.

The yield is considered to be affected by the dual effects of cultivation environments at multiple stages. Therefore, the authors have studied a method using decision tree (Buchanan, G. , B. et al, 1978). Decision tree is easy to interpret visibly and can represent the dual effects of cultivation environments at multiple stages. However, decision tree makes factors more complicated to grasp because of its hierarchical structure. In this paper, we propose a method using clustering and feature selection. We

do not intend to deepen the tree by using clustering, Ward's hierarchical method (Ward, Jr. , H. , J. , 1963), and feature selection.

2. METHODOLOGY

2.1 Consideration of soybean growth by stage division and qualitative value conversion

It is impractical to manage the daily growth of a crops because it burdens the farmers. In addition, since soybeans are known to be affected by the surrounding environment depending on the stage, the influence is to be examined by dividing the cultivation environment data by each stage or the corresponding period. In this study, we focus on the fact that the growth of soybean depends on the temperature before the flowering stage and the number of days thereafter (National Agriculture and Food Research Organization (NARO), 2003). The growth of soybean is divided into 8 stages by accumulated temperature before flowering beginning, and by 10 days after flowering beginning into 5 days. The accumulated temperature is a value obtained by adding the mean temperature every day from the sowing date, and the next stage begins from the next day when the value exceeds a reference value. Then, let the typical value for every cultivation environment element in each stage be a value in the stage (Figure 1).

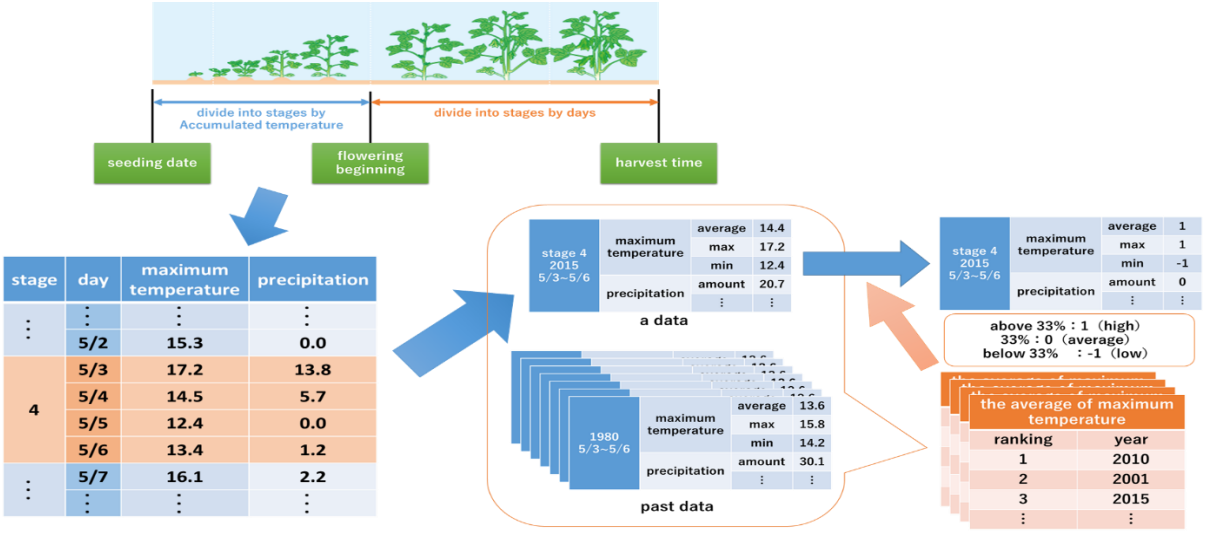

Figure 1. Dividing stage and making qualitative value (Syngenta Japan, "bits of knowledge about soybeans", https://www.syngenta.co.jp/cp/columns/hatasaku_blank_01?TB_iframe=true&width=740&height=800, Accessed 26 April 2019 (in Japanese))

Soybeans and other crops are not affected by small environmental changes. Therefore, the change of the environment is roughly grasped by converting the representative value of each element acquired for each stage to a qualitative value. As shown in Figure 1, each element is ranked by comparing its value with values in the same period in the last 30 years. According to the ranking, the elements are classified by the following three conditions; ranked in the top 33% as "high", bottom 33% as "low", and the others as "average". These three conditions are represented as {1, -1, 0} respectively.

2.2 Creation of field groups with similar cultivation environment by clustering using feature selection

There are many attributes of the data after being qualitatively valued. As the value varies depending on the field, too many complex yielding factors will be obtained if this data is used. Complex factors are difficult to use in actual soybean cultivation and are not useful. Therefore, we make groups of fields with similar cultivation environment by using clustering (Ward, Jr. , H. , J. , 1963), which enables to

classify given data without external criteria. Ward's method is used in this work. The clustering was performed between fields and not within the fields. For example, clustering of Figure 2 shows the degrees of similarity are calculated for seven fields, and clusters are formed in terms of similarity. As a result, a total of five clusters are created.

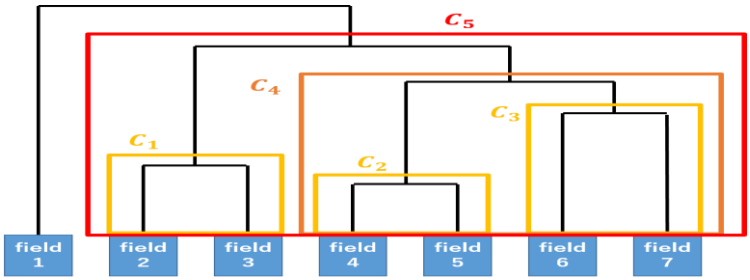

Figure 2. Example of clustering

Among the clusters created in this way, a cluster with a small number of fields is not useful because there are few fields that apply to the high or low yield factors in the end, and the factor will be unreliable. So, in this research, we will try not to create such a cluster so that the results are not influenced by feature selection. We select the elements used to create clusters with more than the number of differences between high and low yielding fields (we defined this number as 'dif') and remove elements that have never been used. Thereafter, clustering is performed again using only the selected element.

2.3 Extraction of high/low yield factors by decision trees

The data after clustering is only considered as influential at each stage. Since the crop grows with time, its yield is considered to be affected by the dual effects of cultivation environments at multiple stages. So, it is necessary to take into consideration the influence between stages in the whole period, not only at each stage. Thus, we use the decision tree (Buchanan, G. , B. et al, 1978) to find the factor from the information of fields to the clusters. For example, Figure 3 shows that fields, which do not belong to cluster 1 in stage 3 and belong to cluster 2 in stage 10, become high yielding fields. When the condition belonging to cluster 1 and cluster 2 in Table 1, a high yield factor "The average of maximum temperature is high in the stage 3 and the average of precipitation is low in the stage 10." can be finally obtained.

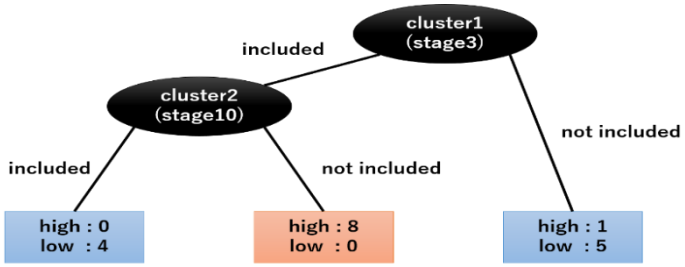

Figure 3. Sample of decision tree

Table 1. Example of the condition belonging to clusters

Cluster	The condition belonging to clusters
Cluster1	The average of maximum temperature is high in stage 3
Cluster2	The average of precipitation is low in stage 10

3. EXPERIMENT

In order to verify the effects of clustering, experiments have been conducted for the method using "Without clustering method" described in Section 2.3 and the proposed method called as "With clustering method" here in comparison. The outline of the field data are shown in Table 2. The total of 42 types of data are used in six fields from the soybean cultivation areas in Hokkaido. The yield of each field is ranked, and the experiments have been performed with the top 21 fields as high-yielding and the bottom 21 fields as low-yielding fields. We acquired the cultivation environment for every field from NARO mesh agriculture meteorological data (Table 3). In this experiment, dif in Section 2.3 is defined as

$$dif = 6$$

The Ward method is used for clustering, and the C5.0 algorithm is used for the decision tree.

Table 2. Field data

The number of fields	42
Location	Iwamizawa, Naganuma, Hitsujigaoka, Memuro, Bibai, Asahikawa
Year	2004 – 2017
Seeding date	12 May – 6 June
Cultivar	Yukihomare, Toyomusume, Toyoharuka, Yukishizuka

Table 3. environmental data acquired from NARO

The elements	Mean temperature	Maximum temperature	Minimum temperature	The amount of solar radiation	Precipitation	Sunshine duration
Location	Iwamizawa, Naganuma, Hitsujigaoka, Memuro, Bibai, Asahikawa					
Year	2004 – 2017					
The period of data acquisition	1 May – 30 November					
Qualitative value	Average, Maximum, Minimum, Range	Average, Maximum, Minimum	Average, Maximum, Minimum	Sum, Maximum, Minimum, Range	Sum, Maximum, Minimum	Sum, Maximum, Minimum, Range

We finally discuss the results of experimental methods with the experts (i.e. the cultivators of soybean in Hokkaido) to examine whether the results are matched with their experience.

4. RESULTS

The results obtained by "With clustering method" and "Without clustering method" are summarized in Figures 4 and 5. In "With clustering method", we obtained the following:

- If the range of the mean temperature in stage 3 is high and the total value of the solar radiation amount is below the average, then the yield is high.
- A field that does not fulfill all the conditions shown in Figure 5 has low yields.

In "Without clustering method", we obtained the following:

- If the total value of the amount of solar radiation in stage 9 is low, the yield will be low.

- Even if the above conditions are not met, low yields may occur when the maximum of sunshine duration in stage 8 is high.
- If the amount of solar radiation in stage 9 is above the average, and the maximum of the sunshine hours in stage 8 is below the average, the yield is low.

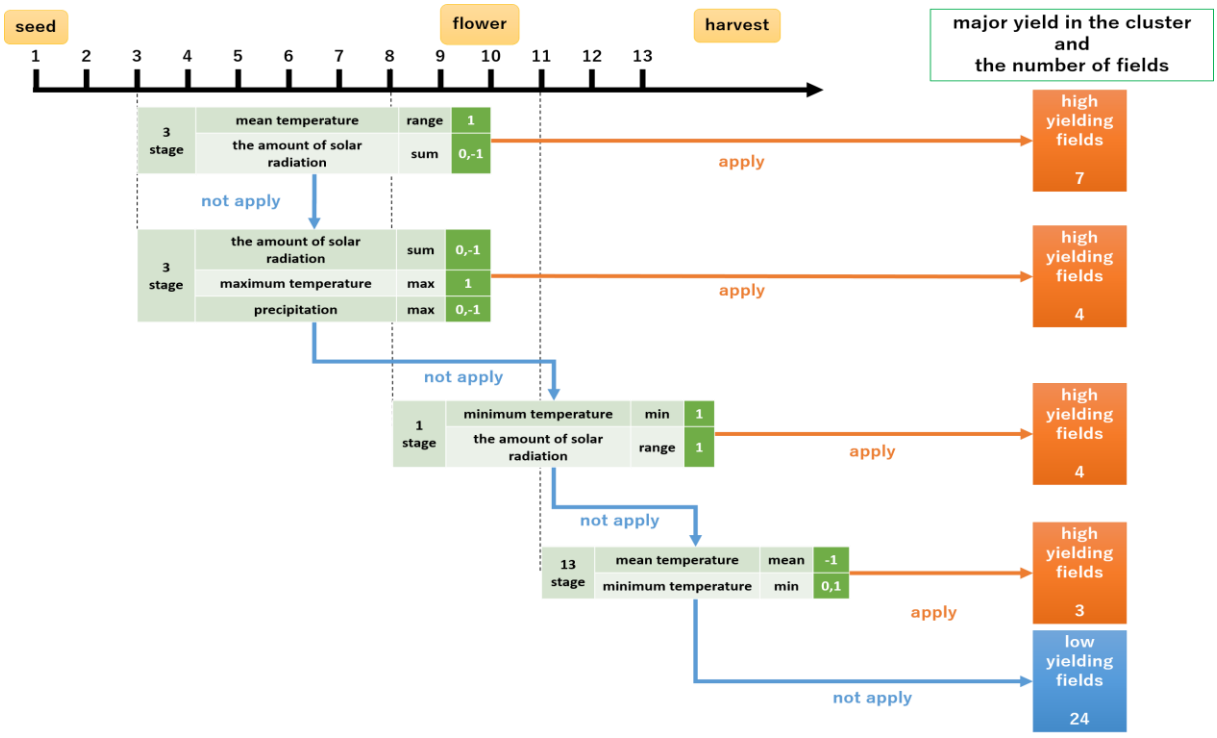

Figure 4. Decision tree of "With clustering method"

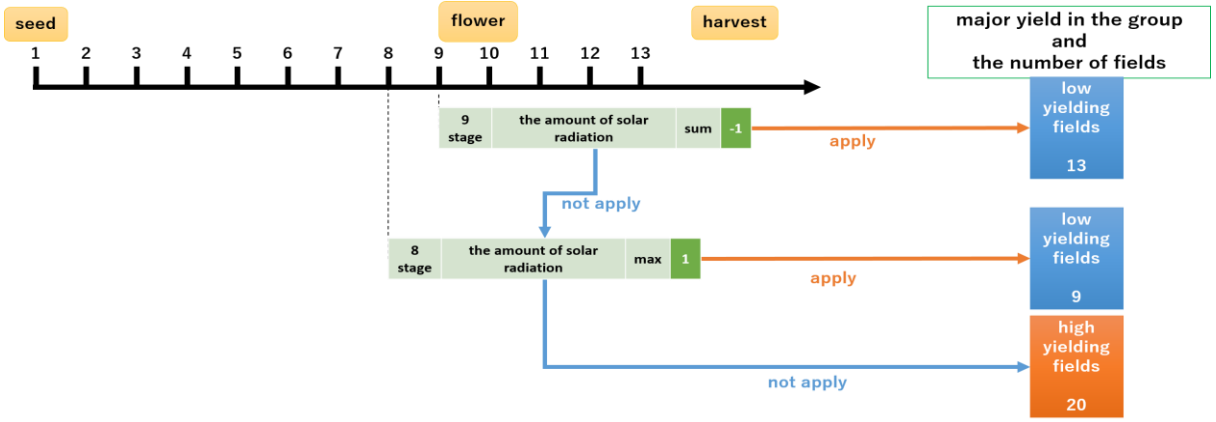

Figure 5. Decision tree of "Without clustering method"

5. Discussion

First of all, as the number of nodes of decision tree is fewer in "Without clustering method", it is easy for the "Without clustering method" to interpret. Looking at each factor, the factor of the first condition in stage 9 is matched with the experts' knowledge that high amount of solar radiation around the beginning of flowering leads to high yield. On the other hand, the second factor is not matched with the experts' knowledge.

Next, we analyze the results of "With clustering method". The individual conditions of "With clustering method" are all matched with the experts' opinions. Moreover, the condition in stage 3 is appeared. Usually, this stage is considered to be less related to growth. We may discover a new knowledge that the environment of stage 3 is much important for soybean growth.

6. CONCLUSIONS

In this study, a method is developed to obtain more reliable high or low yield factors by clustering and using feature selection for the factors that influence the yield of soybeans. In order to confirm the effectiveness of “With clustering method”, the results were compared with the practice and knowledge of experts in actual soybean cultivation. The overall result of “With clustering method” is relatively correct, and is useful information for future soybean cultivation.

The points of the study will be summarized and referred for future works. First, soybean growth does not always depend on the elements of Table 3. Other elements of soybean environment must be considered. Second, we experimented with only 42 data of fields. It is necessary to experiment with additional data to get more reliable results. Third, cross validation needs to be conducted to prove the reliability of our method.

ACKNOWLEDGEMENT

This study was partially supported by Ministry of Agriculture, Forestry and Fisheries: Development of diagnostic methods and countermeasure techniques for overcoming high-yield inhibitory factors.

REFERENCES

Harel, K. , Fadida, H. , Slepoy, A. and Shilo, K. (2014) ‘The Effect of Mean Daily Tem-perature and Relative Humidity on Pollen, Fruit Set and Yield of Tomato Grown in Commercial Protected Cultivation’, Agronomy, Vol. 4, Issue. 1, pp. 167-177.

Japan Agricultural Development and Extension Association (JADEA) (2012) ‘Soybean Making for Improvement of Yield and Quality and Stable Production Q & A’, https://www.jadea.org/houkokusho/daizu/documents/daizu-kaitei.pdf, Accessed 26 April 2019 (in Japanese).

Uno, T. , Asao, T. , Uchida, Y. , and Arimura, H. (2004) ‘An Efficient Algorithm for Enumerating Closed Patterns in Transaction Databases’, Lecture Notes in Articial Intelligence, pp. 16-31.

Umejima, K. , Arimitsu, F. , Ozawa, S. , Murakami, N. , Tsuji, H. , and Ohkawa, T. (2016) ‘Optimal Pattern Mining from Time-Series Cultivation Data of Soybeans for Knowledge Discovery’, Proceedings of Joint Workshop on Time Series Analytics and Collaborative Agents Research & Development (in conjunction with the 29th Australasian Joint Conference on Articial Intelligence), pp.19-24.

Namba, M. , Umejima, K. , Nishide, R. , Ohkawa, T. , Ozawa, S. , Murakami, N. and Tsuji, H. (2016) ‘Optimal Pattern Discovery based on Cultivation Data for Elucidation of High Yield Inhibition Factor of Soybean’, Proceedings of the 5th IIAE International Conference on Intelligent Systems and Image Processing, pp.209-216.

Hira, M. , Z. and Gillies, F. , D. (2015), ‘A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data’, Advances in Bioinformatics Volume 2015, Article ID 198363, pp. 13.

National Agriculture and Food Research Organization(NARO) (2003) ‘Cultivation technology after planting utilizing soybean "Yukihomare"’, http://www.naro.affrc.go.jp/project/results/laboratory/harc/2003/cryo03-03.html, Accessed 15 April 2019 (in Japanese).

Ward, Jr. , H. , J. (1963), ‘Hierarchical Grouping to Optimize an Objective Function’, Journal of the American Statistical Association, Vol. 58, pp. 236-244.

Buchanan, G. , B. and Mitchell, M. , T. (1978), ‘Model-directed learning of production rules’, Pattern-Directed Inference Systems, Academic Press, New York.