International Journal of Control and Automation
Vol. 11, No. 10 (2018), pp.107-118
http//dx.doi.org/10.14257/ijca.2018.11.10.10

An Efficient Analysis Model for Growth Environment Information System using Multi Regression and Modified K-means Algorithms

Se-Hoon Jung[1] and Chun-Bo Sim[2]

[1]School of Major Connection (Bigdata Convergence), Youngsan University, Yangsan, Korea
[2]School of Information Communication & Multimedia Engineering, Sunchon National University, Suncheon, Korea
shjung@ysu.ac.kr, cbsim@sunchon.ac.kr

Abstract

This paper set out to propose an analysis model to make various uses of Big Data to be measured through the expansion of IoT agricultural technology and crop cultivation. Certain elements of crop cultivation environments were extracted by designing a model for the extraction to combine principal component analysis and K-means algorithm and an analysis model to predict and assess crop quality and yield according to the extracted elements. The data of mushroom cultivation was used in the experiment. In the extraction of elements according to cultivation environments, elements were distinguished that had big impacts on the mushroom cultivation environment. The error range was within approximately 5% between the quality of mushrooms that had been harvested and the prediction value of mushroom quality based in the analysis model, which indicates that the analysis model produced similar results to the actual predictions.

Keywords: *K-means, PCA, Multi Regression Analysis, Big Data, Mushroom*

1. Introduction

Big Data, which is produced across various fields today, is reproduced as new processed data to serve the goal of usability[1-2]. Reproduced data mainly include the goal of finding meanings and patterns hidden in them based on classification and analysis. There has been an ongoing increase of Big Data-based system applications across a range of fields in society, and the trend has been increasing in the field of agriculture as well. The utilization of Big Data has increased around the world to analyze cultivation and growth environments in the field of agriculture. But there are limitations to the creation of efficient agricultural environments due to the insufficient sharing of different cultivation methods and environments among different crops and species. It is particularly essential to analyze the related information of cultivation items that require a sensitive growth environment and need enough knowledge of their cultivation and constant management. Today, smart farms are built with sustainable, automated cultivation facilities to provide the observation and analysis data of growth and development environments by adopting IoT sensors[3]. This study, thus, set out to propose an analysis system for growth environments to provide information about the optimal state of an environment and reduce the risks to hinder growth by analyzing and guessing before the data of growth environments. The study proposed an analysis model for crop growth environments by applying multiple regression analysis and an altered K-means algorithm[4-6] so that it would be possible to increase high quality farming yields by processing various data

Received (July 17, 2018), Review Result (September 10, 2018), Accepted (September 29, 2018)

ISSN: 2005-4297 IJCA

including old cultivation environments information and analysis data that offered Big Data with a single consistent system and identifying an efficient method of agricultural cultivation. The proposed model consisted of two major parts: the first one involved the application of altered K-means algorithm to the entire data of crop cultivation environments to analyze the elements of crop growth environments and identify the elements affecting the crop cultivation environments. The elements of cultivation environments were extracted where the principal components were applied up to 80% for each element. The other part involved the application of multiple regression analysis to analyze the identified elements to influence the growth environments and predict the crop yields according to the data of cultivation environments.

2. Related Work

2.1. Regression Analysis

Regression analysis is a statistical method to explain causal relations in nature or in society with explanatory variables to influence and response variables to be influenced. A regression model expresses response variables with the function of explanatory variables, and an estimated regression model is used to predict the values of response variables with those of explanatory variables. Binomial types expressed in Boolean values are used for response variables in regression analysis. When there are three values of response variables or more, multinomial and continuous types are used. Regression analysis, in general, on the premise of linear relations between independent and dependent variables. There are interactive effects in such linear relations just like the increasing values of independent variables will lead to the certain increase or decrease of dependent variables between weight and height, for instance. Formula (1) shows a linear functional formula to present relations between correlated independent and dependent variables. Multiple regression analysis has the same basic concept as simple linear regression analysis, but it uses two independent variables or more. Predictive abilities can be increased by using many different independent variables. This model was used to match linear relations between Y Group of quantitative dependent variables and X Group of independent variables.

$$\hat{y} = (\bar{y} - b_1\bar{x}) + (\sum(x-\bar{x})(y-\bar{y})/\sum(x-\bar{x})^2) \tag{1}$$

2.2. K-means Algorithm

K-means algorithm[7] is a clustering technique to classify input data into k clusters based on unsupervised learning similar to supervised learning. Unlike supervised learning, which updates weight vectors every time a vector is entered, the K-means algorithm updates weight vectors simultaneously after all the input vectors are entered. The criteria of clustering classification are the distance between clusters, dissimilarity among clusters, and minimization of the same cost functions. Similarity between data objects increases within the same clusters. Similarity to data objects in other clusters decreases. The algorithm performs clustering by setting the centroid of each cluster and the sum of squares between data objects and distance as cost functions and minimizing the cost function values to repeat cluster classification of each data object. Intra-cluster distance (IntraCD) is the addition of distance to all the input vectors allocated to the concerned cluster from the centroid of each cluster. Inter-cluster distance (ICD) is the distance of weight vector between two clusters. As seen in formula (2), errors are calculated by adding the sum of IntraCD of all the clusters and subtracting the sum of ICD for all the cluster pairs. β and γ are weighted values.

$$\text{Error} = \beta \sum_{i=0}^{k}(Intra\ CD) - \gamma \sum_{i=0}^{k}(ICD) \tag{2}$$

Macqueen Approach K-means[8] presents a method usually used in researches on the utilization of the K-means clustering algorithm. The user arbitrarily selects K, the initial number of clusters, from the initial objects for clustering. K can be randomly selected from the entire objects. K is assigned to the closest cluster according to the measurement of distance from the cluster centroid to all of the objects. The objects assigned to a cluster will be finally used for reassignment to new centroids. The algorithm will be terminated when centroids are measured under the threshold value set by the user. Kaufman Approach K-means[9] introduces an algorithm to supplement the rising costs of measurement due to the calculation of distance from all of the objects, which had been pointed out as a problem with Macqueen Approach K-means. The algorithm is faster in measurement than Macqueen Approach K-means. This algorithm determines the centroids of all the objects distributed as initial centroids. The distance is measured from an object of the entire group to the initial cluster value. When a selected object has a distance from the initial value over the threshold value set by the user and has a certain number of objects nearby or more, it will be determined as a cluster centroid. The algorithm would repeat the process until the initial K value matches the selected outcome until termination. The approaches of Macqueen Approach K-means and Kaufman Approach K-means to initialization have, however, problems of selecting initial values arbitrarily and choosing initial values according to certain conditions and rules. Max-min Approach K-means[10] selects an object from the entire group and determines it as the first initial value and measures its distance from the other objects. The object with the longest distance from the first initial value is assigned as the second initial value. Distance from the other objects will be calculated based on the first and second initial values. Since two initial values are needed to calculate distance, the sets of distance pairs will be measured to save two measurements. The initial value of distance pairs from the measured set of distance pairs will be defined as the distance measured between the concerned object and two values. The observed value with the maximum distance measurement will be determined as the third initial value, which will have the longest distance from the earlier two initial values assigned. The algorithm will continue until the initial values of K are all met through the repeated process.

2.3. Growth Environment Analysis System

Study [11] proposed to promote the reliability of a farming journal by automatically saving the data of produce conditions and control environments and entering the multimedia data of produce. The farming journal was materialized in a physical layer, which was comprised of soil sensors and internal and external sensors in the cultivation field, a middle layer, which covered the journal's database, video, sensor, and server management, and an application layer, which provided users with GUI. The farming journal was designed to record general works and disease and pest forecasts and check the data inserted in video, voice, text or image.

Study [12] proposed a management and monitoring system for a growth environment to increase a crop yield. The growth monitoring system would check the crop conditions via the sensors and control the environment artificially. Related environment sensors would be necessary for EC, pH, temperature, humidity, intensity of illumination, and CO2. Most of the sensor nodes were organized in a wired fashion, and the system was organized in the RS485 method. When it was organized in a wireless fashion, the Zigbee-based USN technology was applied. The control system covered crop cultivation, environment, nutrient solution and source of light. Data collected from sensor and sink nodes would be sent to the sever of a local gate to monitor the current conditions. Independent gateways were set for sensor and energy monitoring control.

3. Proposed Analysis Model

3.1. Structure of Analysis Model

Figure 1 shows the analysis model diagram for classifying crop cultivation data suggested in the study and identifying independent variables that affect specific dependent variables. There are a variety of factors affecting the crop cultivation environment, and the effects can be different by factor. The models for analyzing such factors and identifying independent variables affecting dependent variables can be divided into three. The density of multi-dimensional factors is analyzed through the transformed K-means algorithm, which is suggested in the study, and independent variables including 80% of main ingredient factors are extracted primarily. As such independent variables have high correlation among variables, they are highly likely to create analysis models and have effect on dependent variables. Moreover, the optimal number of communities can be classified through the transformed K-means algorithm, and independent variables to be applied for the analysis model can be drawn. Independent variables are extracted based on the classified the number of communities, and a predicting model that affects the final dependent variable is created through a multiple regression analysis.

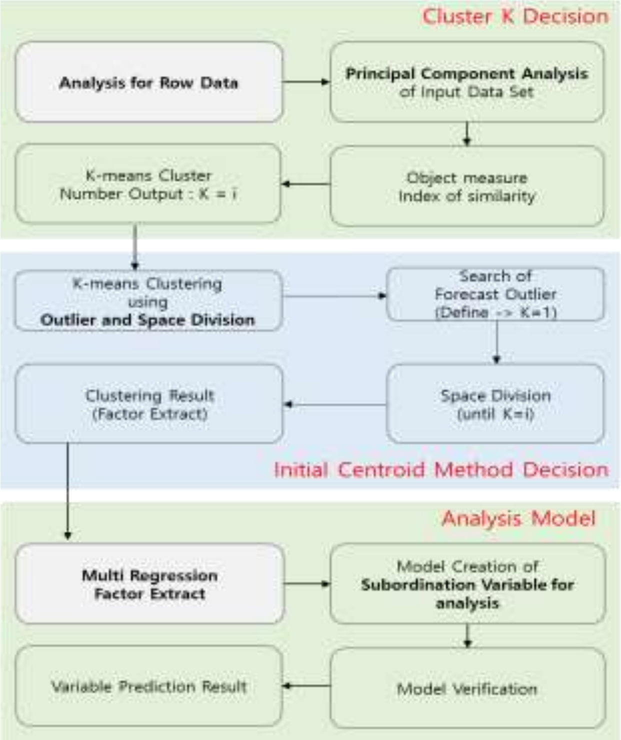
Cluster K Decision
Analysis for Row Data
Principal Component Analysis of Input Data Set
K-means Cluster Number Output : K = i
Object measure Index of similarity
K-means Clustering using Outlier and Space Division
Search of Forecast Outlier (Define -> K=1)
Clustering Result (Factor Extract)
Space Division (until K=i)
Initial Centroid Method Decision
Analysis Model
Multi Regression Factor Extract
Model Creation of Subordination Variable for analysis
Variable Prediction Result
Model Verification

Figure 1. Overall Structure of Analysis Model

3.2. Factor Detection using Modified K-means Algorithm

A model was designed to extract the elements of crop cultivation environments by applying an altered K-means algorithm so that it would be possible to analyze the elements that had huge impacts on crop cultivation according to the elements of cultivation environments. There is a need for a scale to assess whether analysis data are similar or dissimilar in the clustering method not based on the probability model of crop cultivation data. Clustering is usually done with dis-similarity(or distance) rather than similarity. The method of determining K, the number of clusters, is critical for optimal clustering. The present study used K, the optimal number of clusters, to set a temporary scope through the principal component analysis of input crop cultivation data. The scope of principal components to minimize the sum of distance squares within the set scope was appointed as the final number of clusters, K. The multi-dimensional data of crop cultivation data vectors were altered linearly based on the dispersion matrix to analyze the number of clusters in the clustering of input crop cultivation data. The objects of input crop cultivation elements were measured for the principal component direction and element of covariance matrix to set the scope of principal components up to 80% of threshold value. Determined as the final number of clusters was the principal component of the highest dissimilarity among the clusters based on the Euclidean distance formula within the set temporary scope of principal components. Figure 2 presents the clustering extraction algorithm of crop cultivation data based on the application of principal component analysis. And Figure 3 presents to modified K-means logic pseudocode.

① Scatter plots and the number of random data, p in the input data will be checked.

② Principal component analysis will be conducted for the entire input data entities, and principal components will be extracted till the point where a constant value will be maintained to explain the entire data.

③ The central point segmentation method will be applied to C_{kl} , the number of random clusters, and n_k, the number of random central points, based on the principal components that have been extracted through principal component analysis. m_k, the central point of each initial cluster, will be measured with a random cluster index vector.

④ The minimum value of m_k, the central point of each segmented area, will be calculated with A_k, the sum of squared distance to each entity.

⑤ B_k will be calculated which is the minimum average distance between the central point of a random cluster and the entities included in an external cluster.

⑥ S(k), which is the maximum cluster dissimilarity based on a difference betweenA_k , the degree of separation based on the average distance between the entities included in different clusters and all the other entities, and B_k, the degree of cohesion based on the average distance between an entity within a cluster and one in an external cluster, will be treated C_k, which represents the number of clusters, K.

⑦ Set C_k , the number of input vector data.

⑧ All of vector data will be measured for mean μ , standard deviation σ , and standard normal distribution $\emptyset_{\mu,\sigma^2}(x_k, y_k)$.

⑨ Select the objects distributed in the space fractile and $\mathrm{P}(\bar{X} \geq x_k, y_k) = \pm 0.95$ by using μ , the mean of vector data, and σ , standard deviation.

⑨-① When there is one object distributed, assign the object to m_1 , the centroid of C_1 , the first cluster.

⑨-② When there are two objects or more distributed, assign the object whose two vectors record the biggest length first to m, the centroid of C_1, the first cluster.

⑩ Measure the distance(A_i) between the remaining objects(x_i) and , the centroid of C_1 , the first cluster and assign the object whose distance measurement is the biggest to m_2 , the centroid of C_2 , the second cluster.

$$C_2(m_2) = a_i \leftarrow \max_{1\le i\le n}(A_k) = \max_{1\le i\le n}\|x_i - m_1\| = \|\alpha_i - m_1\|$$

⑪ Distance to the remaining objects(x_j) will be measured in the following formulas after the exclusion of m_1 and m_2 , the centroids of C_1 and C_2 excluded in Stage ⑩, respectively:

⑪-① Measure the distance pairs to m_1 and m_2, the centroids of clusters, and then the maximum value of distance pairs.

$$CA_i = \max \|x_j - m_1\|, \|x_j - m_1\|, \qquad 1 \le j \le n$$

⑪-② Of the maximum measurements, assign the object with the maximum value to m_3, the centroid of C_3, the third cluster.

$$C_3(m_3) = x_i \leftarrow \max_{1\le i\le n}(CA_i) = CA_1$$

⑫ Repeat Stage ③ with K, the number of clusters set by the user in advance. When the conditions of the initial value of K are met, stop Stage ⑩ and perform clustering for the other objects.

⑬ The remaining objects form a cluster in a direction that is the closest to the area selected based on the initial values. Then the user will appropriate the sum for the centroid of another cluster and repeat the Stages ⑩~⑫ until there is no more travel of each cluster centroid.

Figure 2. Modified K-means Algorithm

```
BEGIN
FUNCTION main()
    array VectorData(),int centroid()

IF learn THEN
    FUNCTION vectorData()
    FOR cnt = 0 to k=1 DO
      FOR cnt = 0 to PCA(Max) DO
         FOR cnt = 0 to clustering k DO
             value = n(object) * p(variable)
         END
         int threshold()
      END
    END
    int k-value()
    int centroid()

ELSE IF classification THEN
   FUNCTION centroid()
    FOR cnt = 0 to k=1 DO
      FOR cnt = 0 to centroid point DO
         FOR cnt = 0 to clustering k DO
             Distance[][] = centroid * point
         END
         factor = Kvalues(clustering)
      END
    END
    sequence factor extract()
    sequence Kvalues_arr_load()
    sequence Euclidean distance()

ELSE
END
```

Figure 3. Modified K-means Logic Pseudocode

3.3. Design of Prediction Model using Multi Regression

The multiple regression analysis technique was applied to make numerical predictions of influences on the final crop yield and quality based on the elements of crop cultivation environments extracted with an altered K-means algorithm. Of data prediction techniques, multiple regression analysis was chosen which was good at analyzing relations among independent variables and predicting influences on dependent variables. The crop quality scores, which were a dependent variable, were favorable for the conditions of multiple regression analysis since they were known beforehand for each crop and quantitative. A variety of crop growth environment elements were applied including temperature, humidity, intensity of illumination, CO2, cultivation site, and outside temperature. The farmers graded the crops according to these elements with their subjective numbers of quality. Formula (3) shows the basic model of multiple linear regression analysis with the k number of independent variables to identify the effects of growth environment elements on quality scores and their elements. The regression coefficient of crop cultivation environment elements, which were independent variables, was β_0; the measurement error of crop quality ε_i; which was a dependent variable, was β_0; and the estimation of regression model was to find the regression coefficient.

$$Y_{i=score} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i, \tag{3}$$

4. Results

4.1. Factor Detection using Modified K-means Algorithm

A total of 9,414 pieces of raw data of mushroom cultivation measured for a week were used to assess the proposed analysis system for crop cultivation environments. There was a total of 13 elements of mushroom crop cultivation environments, and the mushroom cultivation data was used according to temperature, humidity, intensity of illumination, outside humidity, and CO2. The overall attributes were compared with the final quality element items, and the analysis results show that six attributes, temperature, humidity, intensity of illumination, CO2, outside temperature, and odor, had close connections with crop cultivation environments. A proposition based on this was a clustering method through principal component analysis and K-means analysis. The method was then applied for performance evaluation. The temporary range of six principal components was set with covariance matrix in a principal component analysis. The measurements of dissimilarity indicate that four principal components recorded 0.41, which suggests that the four elements had the biggest impact on the quality scores of mushroom cultivations. Figure 2 presents the analysis results of the 13 influential factors of mushroom cultivation environments, which were analyzed with an altered K-means algorithm, by the clusters. The final elements include the components whose cluster density was high based on the results of six principal components in a principal component analysis.

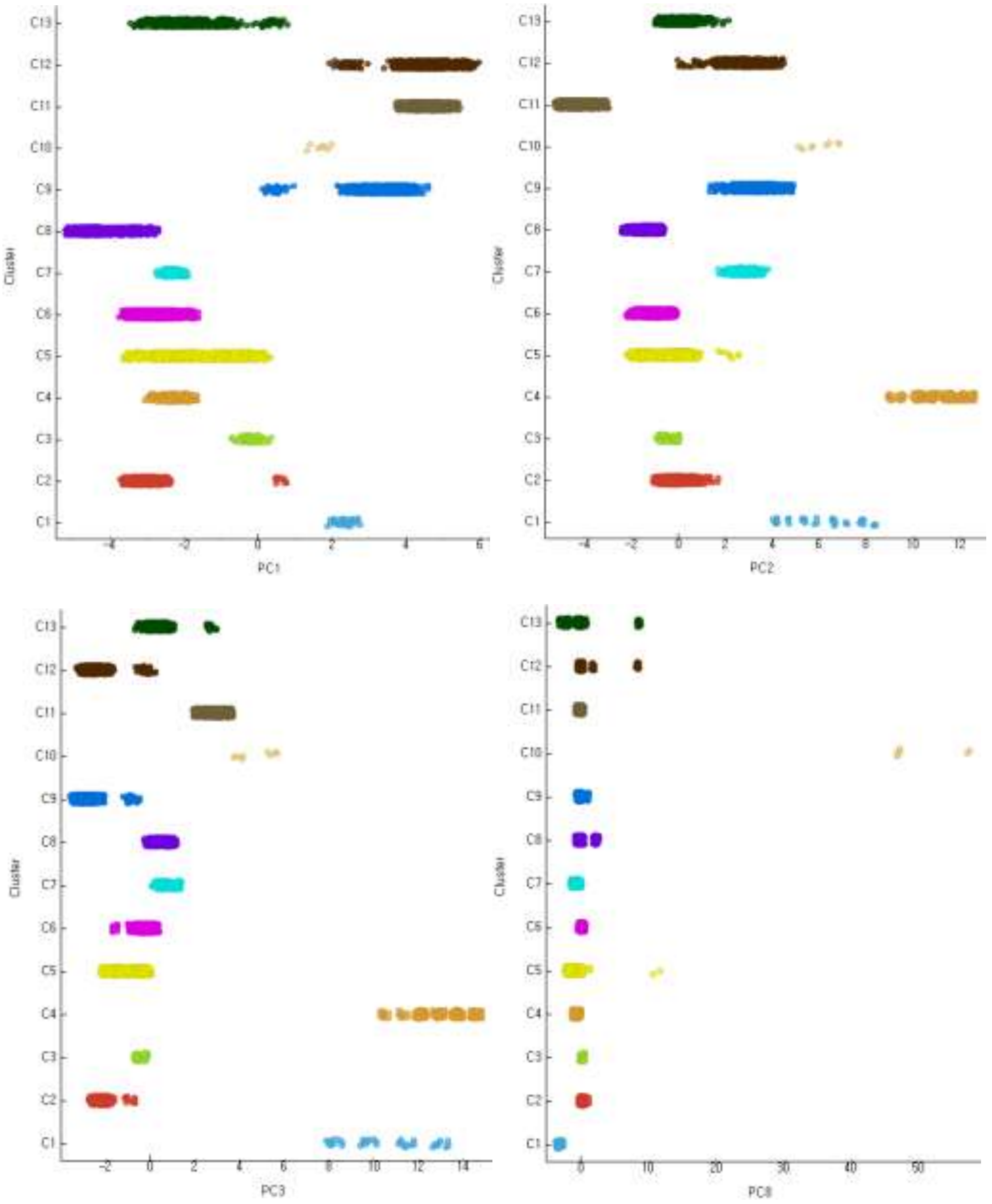

Figure 2. The Analysis Results of the 13 Influential Factors of Mushroom Cultivation Environments, which were Analyzed with an Altered K-means Algorithm, by the Clusters

Multiple regression analysis was conducted to analyze and predict a model of numerical quality evaluation according to the mushroom crop yields based on the four elements of mushroom cultivation environments. Table 1 shows the regression coefficient, standard error, P value, and sum of squares of each of these four elements. Regression coefficient was measured with temperature, humidity, illumination, and CO2. The mushroom quality scores were in the range of 0~100 points based on the old mushroom cultivation data. Formula (4) presents a formula to measure evaluation scores according to the data of four cultivation environments. Table 2 shows the deviation in quality scores between the seven pieces of data of 100 predictions data based on learning results and actual data.

Table 1. Multi Regression Analysis Result(Mushroom)

Variable	Coefficient	Std. Error	*P-Value*	*Sum of Squares*
Intercept	64.88592442	0.440996191	0	39712.41255
Temperature	0.006936071	0.003341822	0.500967715	20.25626918
Humidity	0.040577390	0.017977714	0.693805183	200.6268514
Illumination	0.008187306	0.175713039	0.426978991	78030.39735
Co2	-0.007495560	0.007627452	0.21489621	0.178319922

$$Y_{i=score} = 64.88592442 + 0.006936071 * x_{i(Temp.)} + 0.004057739 * x_{i(Humi.)} \\ +0.0.008187306 * x_{i(Illu.)} + 0.007495560 * x_{i(co2)} + \varepsilon_i, \quad (4)$$

Table 2. Mushroom Data Quality Measurement and Prediction

Number	Score	Est. Score	Coefficient			
			Temp.	Humi.	Ill.	Co2
1	82	84.82849	24.66	92.09	618	2
2	81	84.81085	24.66	91.84	618	1
3	80	84.81531	24.66	91.95	618	1
		…………………………………..				
98	81	80.23244	19.89	51.1	262	4
99	79	79.74646	18.88	55.79	183	1
100	74	79.5912	17.79	52.15	183	1

The quality error was set at 10.96. The experiment results show an error of approximately 5% (1~10 points) between the actual and predicted measurements of cultivated mushroom quality. There were, in particular, big differences in mushroom quality according to temperature and humidity changes. When there were severe temperature changes, there was a difference of about 5~7 points between actual and predicted quality scores. The optimal temperature for a mushroom cultivation environment was predicted to be approximately 20~24 degrees. In addition, there was a smaller deviation in mushroom quality in a cultivation environment of 70~95 humidity than in one of less than 70 humidity.

5. Discussion

This study proposed a model for the analysis of crop cultivation environments to analyze the elements of crop cultivation environments and the effects of big influences of such environments on final crop yield and quality. A model was built with regression analysis, which was capable of analyzing the categories of large-volume crop environment data and the data sets of certain crops, and Big Data analysis, which identified crop elements based on an altered K-means algorithm. An algorithm was established to extract elements that would have a lot of influence on the concerned crop from a huge pool of crop cultivation elements by applying principal component analysis and K-means algorithm. The experiment results were used to apply the elements of mushroom cultivation environments and extract six principal components of close connections with mushroom quality among 13 elements. The K-means algorithm was applied to extract four elements that had big effects on the cultivation environment. In addition, a multiple regression analysis-based prediction model was applied to compare and analyze actual and predicted mushroom quality. The experiment results show that there was a difference of about 5% (1~10 points) and that there were differences in crop quality scores according to certain elements.

Acknowledgments

This work was supported by Youngsan University Research Fund of 2018. and this paper has been written with the support of Jeollannam-do (2017 R&D supporting program operated by Jeonnam Technopark). and this paper is a revised and expanded version of a paper entitled [A Design of Data Analysis and Prediction System of Acer Mono Sap Based on Artificial Neural Network Using Sensor Data and Unstructured] presented at [7th International Conference on Next Generation Computer and Information Technology, Royton Sapporo, Hokkaido, Japan, August 23 - 25, 2018].

References

[1] C.-W. Kim and S. Park, “Document Clustering Using Semantic Features and Fuzzy Relations”, International Journal of Communication Convergence Engineering, vol. 11, **(2013)**, pp. 179-184.

[2] H. Yun, “Analysis of Similarity of Twitter Topic Categories among Regions”, International Journal of Communication Convergence Engineering, vol. 10, **(2012)**, pp. 27-32.

[3] Z. Hong, Z. Kalbarczyk and R. K. Iyer, “A Data-Driven Approach to Soil Moisture Collection and Prediction”, Smart Computing(SMARTCOMP), 2016 IEEE International Conference on, St. Louis, **(2016)**, pp. 1-6.

[4] S.-H. Jung, J.-C. Kim and C.-B. Sim, “Prediction Data Processing Scheme using an Artificial Neural Network and Data Clustering for Big Data”, International Journal of Electrical and Computer Engineering, vol. 6, **(2016)**, pp. 330-336.

[5] M. G. H. Omran, A. P. Engelbrecht and A. Salman, “An overview on clustering methods”, International Journal of Intelligent Data Analysis, vol. 11, **(2007)**, pp. 583-605.

[6] J.-s. Park, S.-h. Jung, K.-h. Jo, K.-h. Jo and C.-b. Sim, “A Design of Data Analysis and Prediction System of Acer Mono Sap Based on Artificial Neural Network Using Sensor Data and Unstructured Data”, Proceedings of the International Conference on Next Generation Computer and Information Technology(NGCIT), Hokkaido, Japan, **(2018)** August, pp. 96-99.

[7] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm”, International Journal of the Royal Statistical Society. Series C(Applied Statistics), vol. 28, **(1979)**, pp. 100-108.

[8] J. Macqueen, “Some methods for classification and analysis of multivariate observations”, In Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, vol. 1, **(1967)**, pp. 281-297.

[9] K. Zhang, W. Bi, X. Zhang, X. Fu, K. Zhou and L. Zhu, “A New K-means Clustering Algorithm for Point Cloud”, International Journal of Hybrid Information Technology, vol. 8, no. 9, **(2015)**, pp. 157-170.

[10] F. Yuan, Z. H. Meng, H. X. Zhangz and C. R. Dong, “A New Algorithm to Get the Initial Centroids”, Proceeding of the 3rd International Conference on Machine Learning and Cybernetics, **(2004)**, pp. 26-29.

[11] T. Oates Ganesan, “Beyond Object Proposals: Random Crop Pooling for Multi-Label Image Recognition”, Proceedings of the International Joint Conference on Neural Networks(IJCNN), Anchorage, USA, **(2017)** May, pp. 14-19.

[12] Y. W. Lee, J. S. Cho, H. H. Shin, H. Yoe and C. S. Shin “Construction of Farming-diary Management System Using Ubiquitous Technologies”, Proceedings of the International Conference on Korean Internet Information Society, Cheonan, Rep. of Korea, **(2009)** May, pp. 301-305.

[13] D. S. Ko and H. S. Park, “The Study for Design of Growth Environment Monitoring System of Vertical Farm”, Proceedings of the International Conference on Korean Information Technical Society, Chungju, Rep. of Korea, **(2011)** May, pp. 372-375.

Authors

Se-Hoon Jung, he received his BSc and MSc and PhD in Multimedia Engineering from Sunchon National University in 2010 and 2012 and 2017, respectively. Currently, he is an assistant professor with the school of major connection(Bigdata convergence), Youngsan University, South Korea. His research interests include data analysis and data prediction. E-mail : shjung@ysu.ac.kr

Chun-Bo Sim, he received a BSc, MSc, and PhD in computer engineering from Chonbuk National University, South Korea, in 1996, 1998, and 2003, respectively. Currently, he is a professor with the Department of Multimedia Engineering, Sunchon National University, South Korea. His research interests include multimedia databases, ubiquitous computing systems, and big data processing. E-mail : cbsim@sunchon.ac.kr