Regular paper | Journal of the KIECS. pp. 1269-1276, vol. 18, no. 6, Dec. 31. 2023, t. 122, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2023.18.6.1269

그래프 분류 기반 특징 선택을 활용한 작물 수확량 예측

옴마킨* · 이성근*†

Crop Yield Estimation Utilizing Feature Selection Based on Graph Classification

Ohnmar Khin* · Sung-Keun Lee*†

요 약

작물 수확량 예측은 토양, 비, 기후, 대기 및 이들의 관계와 같은 다양한 측면으로 인해 다국적 식사와 강력한 수요에 필수적이며, 기후 변화는 농업 생산량에 영향을 미친다. 본 연구에서는 온도, 강수량, 습도 등의 데이터 세트를 운영한다. 현재 연구는 농부와 농업인을 지원하기 위해 다양한 분류기를 사용한 기능 선택에 중점을 두고 있다. 특징 선택 접근법을 활용한 작물 수확량 추정은 96% 정확도를 나타내었다. 특징 선택은 기계학습 모델의 성능에 영향을 미친다. 현재 그래프 분류기의 성능은 81.5%를 나타내며, 특징 선택이 없는 Random Forest 회귀 분석은 78%의 정확도를 나타냈다. 또한, 특징 선택이 없는 의사결정 트리 회귀 분석은 67%의 정확도를 유지하였다. 본 논문은 제시된 10가지 알고리즘을 대상으로 특징 선택 중요성에 대한 실험 결과를 나타내었다. 이러한 결과는 작물 분류 연구에 적합한 모델을 선택하는 데 도움이 될 것으로 기대된다.

ABSTRACT

Crop estimation is essential for the multinational meal and powerful demand due to its numerous aspects like soil, rain, climate, atmosphere, and their relations. The consequence of climate shift impacts the farming yield products. We operate the dataset with temperature, rainfall, humidity, etc. The current research focuses on feature selection with multifarious classifiers to assist farmers and agriculturalists. The crop yield estimation utilizing the feature selection approach is 96% accuracy. Feature selection affects a machine learning model's performance. Additionally, the performance of the current graph classifier accepts 81.5%. Eventually, the random forest regressor without feature selections owns 78% accuracy and the decision tree regressor without feature selections retains 67% accuracy. Our research merit is to reveal the experimental results of with and without feature selection significance for the proposed ten algorithms. These findings support learners and students in choosing the appropriate models for crop classification studies.

Keywords

Crop Yield Prediction, Feature selection, Graph classification, Ensemble algorithms, Regression models
작물 수확량 예측, 특징 선택, 그래프 분류, 앙상블 알고리즘, 회귀 모델

* 순천대학교 인공지능공학부(sklee@scnu.ac.kr)
† 교신저자 : 순천대학교 인공지능공학부
· 접 수 일 : 2023. 10. 13
· 수정완료일 : 2023. 11. 12
· 게재확정일 : 2023. 12. 27
· Received : Oct. 13, 2023, Revised : Nov. 12, 2023, Accepted : Dec. 27, 2023
· Corresponding Author : Sung-Keun Lee
Dept. of Artificial Intelligence Eng., Sunchon National University,
Email : sklee@scnu.ac.kr

Ⅰ. Introduction

Today in agriculture, crops are grown at much higher precision to enable farmers to treat plants and animals almost individually, which increases the effectiveness of farmers′ findings significantly. This indicates to even estimate crop yields and assessing crop quality for individual plants. Feature selection is a significant act for a machine learning classifier that involves choosing a subset of the most proper attributes from a dataset to build a predictive model. There are five approaches for feature selection.

Machine learning is well provided when it comes to investigating data about ground situations, including water status, temperature, and chemical makeup, all of which retain an effect on crop development and livestock well-being. Today′s purpose is to attempt new findings for crop improvement. We select the features utilizing the proposed five methods which are feature selection with correlation, uni-variate feature selection, recursive feature elimination(RFE), recursive feature elimination with cross-validation(RFECV), and tree-based feature selection. We utilize ten algorithms to train models, predict, and compare.

Graph data science is good and adequate to provide random forest natively which are great models. But, the other models, bagging classifier, extra trees classifier, linear discriminant analysis classifier, quadratic discriminant analysis, k neighbors classifier, and decision tree classifier are quite good with feature selection. Finally, evaluate the data performance without feature selection methods and run the proposed model accuracy with curves and diagrams. The pretty smooth curves and classifier metrics are illustrated.

The current investigation intends to support a better and more effective crop estimation for farming. The aim of the paper is:

(i) To classify the harvest estimation under the environmental conditions.

(ii) To create a way of maximum yield gains for farmers.

(iii) To develop a unique crop prediction for valuable model accuracy.

Ⅱ. Related Works

A large number of earlier research is expressed in the literature section. They proposes Spatial Temporal Federated Learning(STFL) for Graph Neural Networks (GNNs)[1]. FL environment provides data privacy while performing a useful model conception. Experimentation outcomes of the ISRUC S3 dataset demonstrate the effectiveness of STFL on graph forecasting studies. Weather change is a new challenge for food security and financial planning[2]. In this research, they propose a novel graph-based recurrent neural network for crop forecasting, comprising geographical and temporal facts and additional growth to predict power. It′s the first geographical validity in crop yield forecasting at the county rank nation. Experiments illustrate deep learning strategies on different metrics and test the significance of geo-spatial and material facts. FedGraphNN is created on a suitable graph FL and includes a vast amount of datasets from various fields[3]. Federated GNNs act more harmful in most datasets with a non-IID split than centralized GNNs. The GNN model achieves the finest outcome in the centralized environment. FedGraphNN technique is computationally facilitated and secured to large-scale graph datasets. The present paper proposes a special 3-tiered taxonomy of the FedGNNs publications to supply the idea of how GNNs perform in Federated Learning[4]. How GNN training is accomplished under various FL method architectures and capacities of graph data overlapping across data silos, and how GNN collection is executed under different FL surroundings. Now, numerous machine learning and deep learning techniques test the trained model to use the data assembled and kept in centralized data storage[5]. The effects of the tested algorithms indicate that federated averaging by ResNet-16 regression algorithm with Adam optimizer yields outcomes that are compared with centralized learning algorithms for yield forecast in a federated environment. To forecasts more satisfactory crop yield, hybrid MLR (Multiple Linear Regression) and ANN(Artificial Neural Network) are presented [6]. The computational period for hybrid MLR-ANN

and traditional ANN is computed. The outcomes display that the given hybrid MLR-ANN algorithms give more satisfactory accurateness than the traditional ANN. A crop yield prediction accuracy is useful for more prominent foodstuffs production[7]. This paper presents a predictive model of Pakistan wheat production using 12 algorithms by dividing data samples into three sets which reproduces on other yields and areas. The crop classification using remote sensing images that is required to use the temporal, high spatial, and spectral resolutions of images[8][12]. They created a combination of optimal feature selection (OFSM) with hybrid convolutional neural network-random forest (CNN-RF) networks and a novel crop classification method for many remote sensing images to balance classification accuracy and processing time. Indicating that the Conv1D-RF technique supplies a useful methodology for time sequence presentation. Crop data is gathered and feature selection is executed utilizing the Relief model[9]. Features are extracted using the LDA (Linear Discriminant Analysis) algorithm. Particle Swarm Optimization-Support Vector Machine (PSO-SVM), KNN, and Random Forest are operated for classification. By using these tools, the farmer makes knowledgeable judgments about which vegetables to plant on his farms. The special farm reduces the labor and improves products[10][11]. The purpose of this research is to assist a particular growth yield and gain increased products at lower prices. It also supports the full-price prediction required for a farm. The current research primarily contains the following three sections. In the materials and methods section, graph classification, classifiers with feature selection, and regressors without feature selection are stated. We express the large number of classifier comparisons in the results section. The results and future performance are in the conclusion part.

III. Materials and Methods

3.1 Materials

The crop dataset is in CSV format. There are 2200 rows and 8 columns of attributes in the dataset. The feature names in the data as: Index(['Nitrogen', 'Phosphorous', 'Potassium', 'Temperature', 'Humidity', 'pH', 'Rainfall']). Crops are class label. We replacement the missing data with mean and median of the whole column. A very important plot to visualize the violin and swarm for all the features combinations. Eight features of the crop data are plotted. For the feature comparison, pair grid is utilized. We definitely use this discovery for feature selection.

3.2 Proposed Graph Classification

The graph is powerful to display crops and interactions. It is a homogeneous graph. In graph classification, the input is a graph, and the goal is to train a classifier that predicts the class of the graph accurately and gains better accuracy. The overall processes of GNN are expressed below.

We are splitting the data into separate spaces and then classifying the data accordingly to their labels. All this data is downloaded from Kaggle.

Nitrogen, pH, Potassium, Phosphorous, Humidity, Temperature, Rainfall and Crops are within this database. These attributes are bringing in to the nodes. It is a homogeneous network where nodes are crops and edges between nodes are attributes. Bringing in the edges and consequently it is generating the graph. Each node of crops has a set of features that are illustrated in Figure 1. Generally, the graph does not have any fixed order.

The train and test partitions are created. Thus, split training to 75% and testing to 25%.

A popular task is node classification where the representation of each node predicts a certain class. The graph classification aggregates over all node representations of a graph and compares them with other graphs. Passing between the nodes is a mathematical function in Equation (1), in which 'f' upgrades the receiver node utilizing the messages from the adjacent sender node. c_{11}, c_{12}, c_{13}, and c_{14} are the adjacent matrices. x_1, x_2, x_3, and x_4 are all nodes, and W is the weight of neighbors. The graph's structure fixes it. Depending on the final task, the weight of each feature of the nodes is different.

$$f(x_1) = (c_{11}x_1W+c_{12}x_2W+c_{13}x_3W+c_{14}x_4W) \quad (1)$$

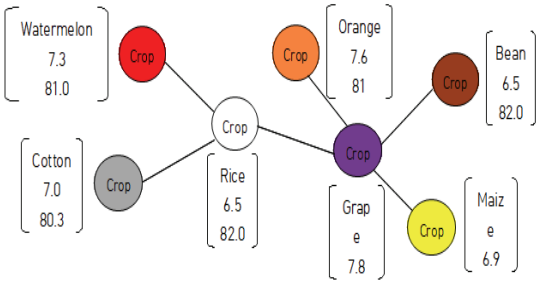

Fig. 1 Graph network for related Crops

A graph's edges are summarized using a table which is an adjacent matrix. The nodes connect to themselves and stack all node features of a graph in a matrix like in Figure 2. And three main elements including the learnable weights, 'W' with all these dimensions. The importance weights for each neighbor are calculated by altering the adjacency matrix, (n x n) consequently and these functions are the same. With this, a highly optimized matrix multiplication library to execute (f) for all the nodes at one time. They are fixed by the graph's structure in Figure 2. In Equation (2), weights of neighbors, 'W' are fixed of the graph. 'A' refers to adjacency matrix and 'X' is for all nodes.

$$f(x) = \sigma(AXW) \qquad (2)$$

We add deep learning algorithms and train our model by node embedding, hidden layers, and 100 Epochs. In the test function, predict node class, extract the class label with the highest probability, checked how many values were predicted correctly, and create a precision percentage using a sum of correct predictions divided by the total number of nodes. Finally, evaluate the accuracy of the graph's predictions on a crop dataset using the test function, and we get pretty good results with 81.5% accuracy which is illustrated in the graph of Figure 3. In Figure 3, x indicates Epochs and y indicates accuracy. The unit of Epochs is the number of seconds and accuracy is a rational number.

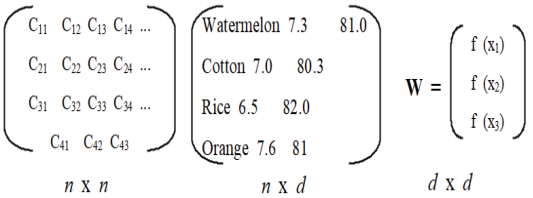

Fig. 2 All node features of a graph in a matrix.

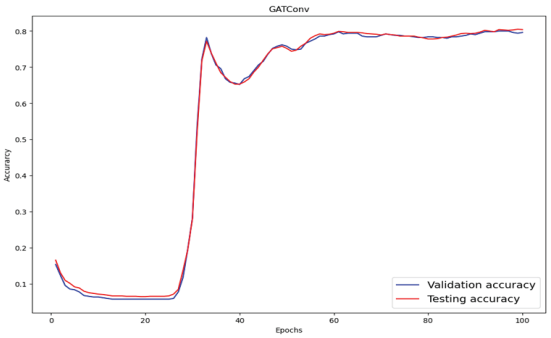

Fig. 3 Evaluation of graph classification

3.3 Feature Selection

The goal is to get a pretty good accuracy based on feature selection in prediction[13]. Consequently, we extract the features using the proposed five methods.

(i) Using "feature selection with correlation analysis" involves identifying the features that are approvingly associated with the target variable. Nitrogen, phosphorus, and Potassium are correlated with each other. Accordingly, they are selected. Linear Discriminant Analysis discovers an accuracy of 94% and there are a few wrong predictions in the confusion matrix.

(ii) "Univariate feature selection" involves selecting the features that have the strongest relationship with the target variable. This can be done using statistical tests such as chi-squared tests. The best three features of Nitrogen, Phosphorous, and Potassium are utilized to classify.

(iii) "Recursive feature elimination (RFE)" is a famous feature selection process that acts on the recursive extracted features and creates a model by the remained features until the optimal subset of

features is identified. The best features of RFE are the Index(['Temperature', 'Humidity', 'pH', 'Rainfall'], dtype = 'object').

(iv) Currently, "Recursive feature elimination with cross-validation and graph classification" is not only finding the most useful features but also searching how numerous features are required for the best accuracy. This data is very easy to classify to make feature selections. Then make the last feature selection method.

(v) In the "Tree-based feature selection and graph classification" strategy, there is a feature important fact. To use this methodology, training data is not the correlated features. Classifier selects at each iteration, thus a series of feature significance lists can vary. Classification higher, the more essential feature is required. To use the feature_importance method, data training is not correlated with features. Consequently, after the best features, we focus on these in Figure 4. In Figure 4, x indicates features' names and y indicates measurement.

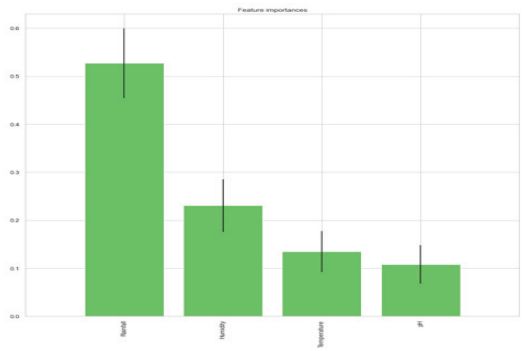

Fig. 4 Plot of the feature importance

IV. Experiment and results

In this experiment, an Ubuntu 18.04.6 LTS Workstation, a Jupyter Notebook, and desktop computers are used. The Jupyter Notebook is an effective product. It investigates the proposed yield dataset with Plotly's library. Later, experiments with the presented distinctive machine learning classifiers and achieve the best accuracies. The accuracies of the above various machine learning classifiers with feature selection are compared to the without feature selection of random forest regressor and the decision tree regressor approaches. These two regressors' accuracies are demonstrated in Table 1. The line chart of prediction results is illustrated in Figure 5.

Table 1. Table of performance evaluation measures

Regressor	Accuracy (Train)	Accuracy (Test)	MAE	RMSE
Random Forest	97%	78%	4	6
Decision Tree	100%	67%	11	19

In Figure 5, x indicates names of algorithms and y indicates the accuracy percentage. The proposed seven algorithms' accuracy is shown in Figure 5. Quadratic Discriminant Analysis accuracy is 92%, Linear Discriminant Analysis Accuracy success is 94%, the Bagging Classifier gets 95%, the extra Trees Classifier has 95%, the K Neighbors Classifier gains 95%, the Random Forest Classifier achieves 95%, and the Decision Tree Classifier consists of 96%. Consequently, the overall average accuracy is 95%. The results are displayed in Table 2. The important step for the proposed classifier is the evaluation stage which promotes the estimation of the difference between the expected and real point. This assumption supports achieving a consistently dedicated classifier for estimation crops.

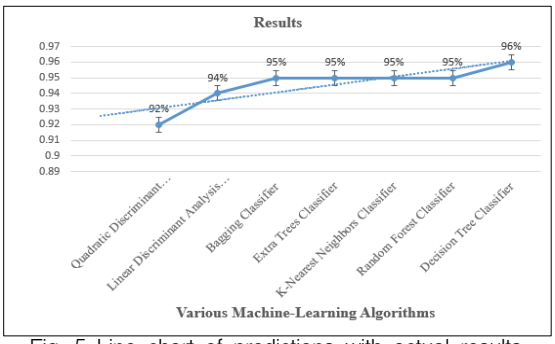

Fig. 5 Line chart of predictions with actual results

Table 1 reveals the diverse results without feature extraction methods. r2_score, accuracy_test, MAE and RMSE for applied regressors are given in below Table 1.

Feature extraction is an essential part of machine learning algorithms. Pre-trained classifiers are experimented on download datasets and run the extracting features first. The correct features are selected and strategy implementation is enhanced. Diverse seven models perform diverse missions. In our research site, RFE and RFECV of five feature selections are used for extraction. In this analysis, a new graph classifier is proposed for crop estimation. The figures and table show the classifier's results for this research, along with their train_accuracies and test_accuracies.

In this study, the seven classifiers are used to test the quite good accuracies with feature selection. Additionally, the accuracies of the random forest regressor and the decision tree regressor without feature selection approaches are compared. The regressor's results are not satisfactory compared to these seven classifiers. Their experimental results details are mentioned in Figure 5 and Table 1. The accuracies vary among these classifiers based on the same input crop dataset. Current research reveals that feature selection approaches do support the best accuracies.

Ⅴ. Conclusions

The results show that our proposed classifiers contribute the most acceptable influences. In summary, these machine learning classifiers of feature extraction highlight the critical factors that affect their performance, including the necessity for large amounts of data to improve accuracy. These findings can aid the learners, students, and experimenters in deciding the right models for crop sort studies. In the future, possible regions of investigation consist of combining different new datasets and creating hybrid algorithms.

「순천대학교 교연비 사업에 의하여 연구되었음 / This work was supported by a Research promotion program of SCNU」

References

[1] G. Lou , Y. Liu , T. Zhang, and X. Zheng., "STFL: A Spatial-temporal federated learning framework for graph neural networks," *AAAI Conference on Artificial Intelligence Workshop on Deep Learning on Graphs: Methods and Applications, Vancouver, Canada, 2021.*

[2] J. Fan, J. Bai, Z. Li, A. O. Bobea, and C. P. Gomes, ""A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction," *Proceedings of the AAAI conference on artificial intelligence,* vol. 36, no. 11, 2022, pp. 11873-11881.

[3] C. Yang, H. Xie, L. Sun, L. He, L. Yang, S. S. Yu, Y. Rong, P. Zhao, and J. Huang, "Fedgraphnn: A federated learning benchmark system for graph neural networks," *ICLR 2021 Workshop on Distributed and Private Machine Learning (DPML), Appleton, USA, 2021.*

[4] R. Liu, P. Xing, Z. Deng, A. Li, and C. Guan, "Federated graph neural networks: overview, techniques and challenges," Journal of latex class files, , vol. 14, no. 8, 2021, pp. 1-16.

[5] M. T. K. Makkithaya and N. V. G, "A Federated Learning-Based Crop Yield Prediction for Agricultural Production Risk Management," *2022 IEEE Delhi Section Conference (DELCON),* New Delhi, India, 2022, pp. 1-7.

[6] P. S. M. Gopal and R. Bhargavi, "A novel approach for efficient crop yield prediction," *Computers and Electronics in Agriculture,*" vol. 165, 2019, pp. 1-9.

[7] M. U. Ahmed and I. Hussain, "Prediction of wheat production using machine learning algorithms in northern areas of Pakistan," *Telecommunications policy,* vol. 46, Issue 6, 2022, pp. 1-12.

[8] S. Yang, L. Gu, X. Li, T. Jiang, and R. Ren, "Crop classification method based on optimal feature selection and Hybrid CNN-RF networks for multi-temporal remote sensing imagery," *Remote Sensing,* vol. *12,* no. *19,*

2020, pp. 3119-3225.
[9] S. Gupta, A. Geetha, K. S. Sankaran, and A. S. Zamani, "Machine learning- and feature selection-enabled framework for accurate crop yield prediction," *Journal of Food Quality*, vol. 2022, 2023, pp. 1-7.
[10] S. K. S. Durai and M. D. Shamili, "Smart farming using Machine learning and deep learning techniques," *Decision Analytics Journal*, vol. 2, no. 3, 2022, pp. 1-30.
[11] O. Khin and S. Lee, "Performance Analysis of Deep Reinforcement Learning Algorithms in Agricultural Crop Production," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 18, no. 1, 2023, pp. 99-105.
[12] J. Bong, S. Jeong, S. Jeong, and J. Han, "Study on Image Use for Plant Disease Classification," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 17, no. 2, 2022, pp. 343-350.
[13] O. Khin and S. Lee, " Feature Extraction and Recognition of Myanmar Characters Based on Deep Learning," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 17, no. 5, 2022, pp. 977-984.

저자소개

옴마킨(Ohnmar Khin)

2008년 : 야타나본 대학교 컴퓨터 공학과(공학사)
2011월 : 양곤대학교 컴퓨터공학과 (공학석사)
2021년~ 현재 : 순천대학교 대학원 멀티미디어공학과 박사과정
※ 관심분야 : AI 기반 이미지 프로세싱, 작물 수확량예측 알고리즘, 심층강화학습, 스마트농업

이성근(Sung-Keun Lee)

1985년 고려대학교 전자공학과 졸업(공학사)
1987년 고려대학교 대학원 전자공학과 졸업(공학석사)
1995년 고려대학교 대학원 전자공학과 졸업(공학 박사)
1996년 ~ 1997년 : 삼성전자 네트워크 연구팀
2017년 ~ 2018년 : 미국 조지아텍 ECE 방문교수
1997년 ~ 현재 순천대학교 멀티미디어공학과 교수
※ 관심분야 : 강화학습 기반 QoS 보장 기술, AI 기반 작물 수확량 예측 알고리즘, 멀티미디어 통신