Machine Learning Methods for the Prediction of Paddy Productivity in Nepal

Chaturbhujj Bhatt, Subarna Shakya and Tej Bahadur Shahi*

Central Department of CSIT, Institute of Science and Technology, Tribhuvan University, Kathmandu, 44600, Nepal

*Corresponding author. E-mail: tejshahi@cdcsit.edu.np

ABSTRACT

Machine Learning techniques have got a rich focus on agriculture management systems due to its significant improvement in classification algorithms. The agricultural data is difficult to study because they consist of different attributes such as geographic locations, soil types, and seasonal conditions. It is challenging to identify the most important attribute that affects the prediction of agriculture yields such as paddy productions. This study is mainly focused on the prediction of paddy productivity of a particular geographic location (Kanchanpur District) which is also categorized as a super zone for paddy cultivation by the Nepal Government. This study aims to collect the agriculture data using manual questionnaire designed with the help of agriculture experts and measure the performance of four machine learning algorithms, namely, Support Vector Machine, Neural Network, Naïve Bayes and Decision Tree for the prediction of paddy productivity (low, medium and high). From the result analysis, it was seen that Decision Tree (SimpleCart) was able to classify 80.19% of the data correctly which was better than SVM, Naïve Bayes and Neural Network in comparison to results of evaluation metrics.

Keywords: Paddy Productivity, Feature Selection, Machine Learning, Classifications

INTRODUCTION

It is important to farmers that they should get information on new paddy types when they are assessing whether to accept these new varieties for farming or not. The information about the productivity of such paddy varieties can be used as part of a farmer's decision-making process so that they can make an optimal decision on choosing paddy varieties for plantations that expect the highest paddy yields. It is common that changing to new paddy variety will get into greater productivity with or without little change in farm resources and outlays. Therefore, it is very important that the information about paddy varieties and their paddy productivity estimation must be reliable and accurate as much as possible. To solve these issues, data mining, and machine learning techniques can be useful where data mining is the extraction of the hidden patterns from large databases to help firms that focus on the most significant information in their agriculture data. Also, machine learning task predicts future trends and patterns which allows businesses to make proactive, knowledge-driven decisions (Han, Kamber, & Pei, 2012). Since the

Article history:
Received 8 March 2020; Received in revised from 6 October 2020;
Accepted 3 November 2020; Available online 11 November 2020

data mining can be used for the automated of data, it made the investigation of past procedures provided by such as decision support systems.(Nirkhi, 2010) The decision support system can help the decision-maker by providing the right information at the right time if this information is managed by an intelligent system (Revathi, Revathi, & Hemalatha, 2011).

It is challenging and complex to study agriculture data from a research perspective is to identify the key attributes that determine the paddy productivity in the complex farming situations or environments which consists of many features such as geographic location, soil types, and seasonal conditions, etc. The key attributes include nutrition and soil type, grain yield and quality, sowing and harvest dates, tolerance to environmental stresses, measurement inaccuracy, sampling discrepancy, outdated data sources and other errors (Kumar, Singh, Kumar, & Singh, 2015). The productivity of the agriculture system has been decreasing for a decade because of varieties of climate changes. It also needs a focused study to determine the relationship between paddy productivity and climate factors and uncertainty. The traditional system had some drawbacks. Irrelevance of the delivered information, the inability of the system to cover all farmers, lack of avenues to improve performance, unaccountability regarding advice given by the system, etc. are major problems (Salzberg, 1994).

This research aims to prepare the paddy data and design a system, for paddy productivity and suggest the best model for paddy productivity prediction among four machine learning models in a special super paddy zone located in Kanchanpur District, Nepal. The result from these experiments can be used as a reference for other locations as it shares the same farming situations.

RELATED WORK

Classification or prediction is the most widely used data mining task. Classification algorithms are supervised methods that uncover the hidden relationship between the target class and the independent variables (Salzberg, 1994). Supervised learning algorithms allow labels to be assigned to the observations so that new data can be classified based on training data(Han et al., 2012; Kumar et al., 2015). Examples of classification tasks are image and pattern recognition, medical diagnosis, loan approval, detecting faults or financial trends(Salzberg, 1994; Wu et al., 2008).

In recent years, Machine learning has attracted researchers from the interdisciplinary field to investigate how computers can learn or improve their performance based on data. When we have labeled data for training, we can use supervised machine learning methods, which are much efficient in prediction tasks as well as classification tasks (Keerthi, Shevade, Bhattacharyya, & Murthy, 2001).

There are many simulation models developed to study soil dynamics. Among them, most successful models belong to data mining techniques which are often used to study soil characteristics. In paper (Veenadhari, Mishra, & Singh, 2011), authors used a k-means classification technique to classify soils and plants into different categories in combination with GPS11 based technologies. Also, they used SVM to classify the crops into different types.

In Pakistan, the agricultural researcher has conducted a research on the impact of pesticide uses on paddy productivity. It has been reported that there is a negative correlation between pesticide usage and crop yield. Therefore, heavy use of pesticides is harmful to the agriculture system with adverse financial, environmental and social impacts. In this paper, it is also reported that how machine learning techniques can be utilized for agriculture data integration into the system which shows that various features like number of insects, usages of pesticides are key factors to determine the optimal uses of pesticide. Using clustering techniques, the interesting patterns of farmer practices and use of pesticide have been determined(Nithya & Sundaram, 2011).

The authors (Veenadhari et al., 2011), studied the impact of environmental factors on crop yield by Kharif and Rabi Corporation in a state of India. They used decision tree data mining algorithm to analyze the influence of farming situations on soybean and paddy crop productivity. They implemented decision tree algorithm as a web-based expert information system in which an expert system provides advisory services to farmers regarding pests, diseases, and their control measures.

Support vector machine based paddy diseases prediction techniques is proposed. They used an image of paddy leaf as an input and processed further using feature extraction technique. They reported the accuracy of 94.16%, precision of 90.9% and recall 91.6% by optimizing the hyper- parameter(Bashir, Rehman, & Bari, 2019).

METHODOLOGY

Data Preparation

Paddy data were collected from the survey in person among 101 farmers of Bheemdatta Municipality, Ward No. 2, 10, 18, and Krishnapur Municipality, Ward No. 3 of Kanchanpur district, Mahakali Zone. These areas in Kanchanpur district are classified as Paddy Super zone by the Government of Nepal. For the implementation of the super zone program, the Government of Nepal has recently launched Prime Minister Agriculture modernization project(Agriculture Report, 2017, pp. 4-6).

The survey questionnaires were developed by consulting agriculture experts of this project and thirteen questions were formulated in total. These features of a dataset are discussed in Table 1.

Proposed System

The collected data were not in a ready format for further processing. The raw data contains categorical, nominal and numerical values. These values were labeled using an inbuilt encoder available in WEKA tool(Hall et al., 2009). Similarly, the noise in the dataset such as outliers collected in the survey was removed manually.

The machine learning models were built using four algorithms: Support Vector Machine, Decision tree, Naïve Bayes and Multi-layer Perceptron Neural Network which further process the pre-processed data to learn the patterns for future predictions.

The Support Vector Machine in their basic form is a binary classifier that separates the linearly separable data using a hyper-plane which maximizes the distance between supporting hyper-plane (Raorane & Kulkarni, 2012).

The hyper-parameters for the Support Vector Machine were chosen as per the optimal result. The kernel trick was used to make the linearly non-separable data into linearly separable in higher dimension space.

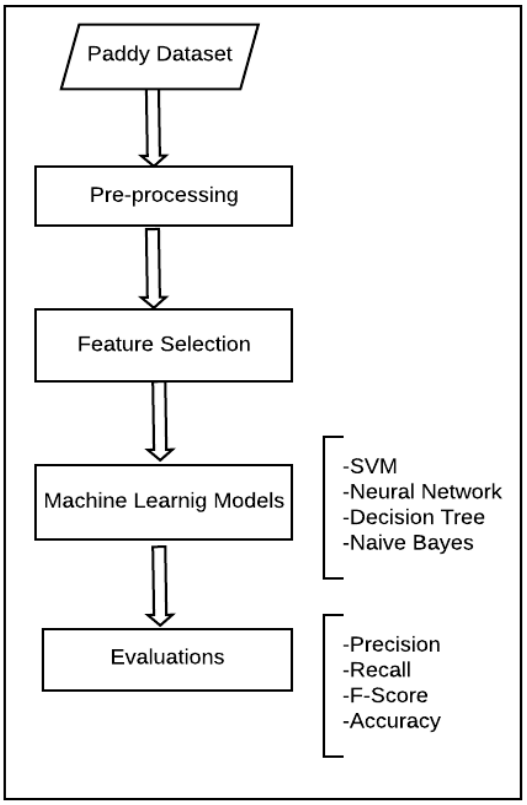

Figure 1 Systematic Diagram of Paddy Prediction System

Multilayer Perceptron Network with three layers was used for building Neural Network. The back-propagation algorithm was used to train the neural network (Nithya & Sundaram, 2011). The SimlpleCart decision tree is used for building a decision tree as provided in WEKA tool. Multinomial Naïve Bayes with optimal hyper-parameter is used to implement the Naïve Bayes classifiers(Singh, Kumar, Gaur, & Tyagi, 2019).

Evaluation and Validations

The evaluation of the machine learning models was performed using accuracy, precision, recall and F-score (Powers, 2011). The accuracy measure how correctly the classifier is producing result whereas precision and recall are used to make a balanced comparison with respect to the number of misclassified instances.

Table 1 Data Feature Description

S. N	Attributes	Attribute Type	Description
1	Seed_Name	Nominal	Name of Seed used by the farmer for paddy cropping
2	Seed_Source	Nominal	Source type of seed and they are: local, agro-vet, agricultural and mixed
3	Seed_Amount	Numerical	Quantity (Kg) of seed
4	Prod_Area	Numerical	Production Area (in Kattha) used for cropping
5	Soil_Type	Nominal	Nature of Soil on production area and they are: Pango, sandy and both
6	Irrigation_used	Nominal	Source of irrigation used by the farmer during cropping and they are: motor (M), Rain (R), Canal (C), Moror+Rain (MR)
7	Fertilizer_used	Nominal	Types of fertilizer used are: Organic, Chemical, and Both
8	Labour_used	Nominal	Types of labor used are: Human, Machine, Both
9	Disease_Insect	Nominal	Disease or Insect found in crop are: Khaire, Gabero, Gandi, Suke, Paterol, etc.
10	Pestiside_used	Nominal	Types of pesticides used in paddy are: organic, chemical and no
11	Harvest_duration	Nominal	Types of harvest duration of paddy are: Long (above 130 days), Medium (between 110 and 131 days), short (below 110 days)
12	Expenses	Numeric	Total expense (NRs) during cropping
13	Previous_Producton	Nominal	Types are: Low (below 1001 Kg, Medium (between 1000 and 2000 Kg), and High (above 1999 Kg).
14	Current_Production	Nominal	Types are: Low (below 1001 Kg, Medium (between 1000 and 2000 Kg), and High (above 1999 Kg).

The validation of the model is done using 4-fold cross-validation. In 4-fold cross-validation, the initial data are randomly partitioned into 4 mutually exclusive subsets or "folds" i.e. D1, D2, D3, D4 each of approximately equal size. Training and testing is performed 4 times in the ratio of 3:1 which means 3 folds are used as Training and 1 fold is used as Testing.

EXPERIMENTS AND RESULTS

Experiments were run on the different settings of hyper-parameters and the results of each experiment were reported in Table 2 and Figure 1. The results of each experiment are also reported in the form of confusion matrices. The confusion matrix shows the performance of the classifier on each class by displaying all correctly classified items and incorrectly classified items for each class.

The result shows that the decision tree is performing best than the other three models namely Support Vector Machine, Neural Network, and Naïve Bayes. It is due to the limited attribute and the limited number of training data. The second competitive results were obtained with the Naïve Bayes model. Form the experiment, it can be claimed that in such an environment where the number of features and number of training samples is small, the decision tree classification model is better than the Support Vector Machine, Neural Network, and Naïve Bayes.

Table 2 Result of Each Machine Learning Model

Classifiers	Accuracy	Precision	Recall	F-Score
SVM	0.5148	0.490	0.515	0.483
Neural Network	0.6767	0.672	0.673	0.671
Decision Tree	0.8019	0.815	0.802	0.804
Naïve Bayes	0.6237	0.619	0.624	0.615

The confusion matrix is a table for analyzing the result of the classifiers. It deals with how many numbers of samples were correctly classified and how many samples were misclassified.

Table 3 Confusion Matrix for Naïve Bayes

Predicted\ Actual	Low	Medium	High
Low	32	7	0
Medium	15	20	6
High	2	8	11

From Table 3, it is clear that the Naïve Bayes classifier is mostly confused or misclassified the high productivity into medium productivity and medium productivity class into low productivity of paddy class.

Table 4 Confusion Matrix for SVM

Predicted\ Actual	Low	Medium	High
Low	22	15	2
Medium	11	28	2
High	6	13	2

From Table 4, it is clear that the SVM classifier is not able to learn the feature to distinguish high productivity class from other low and medium productivity classes. It has good accuracy in the other two classes in comparison to high productivity classes.

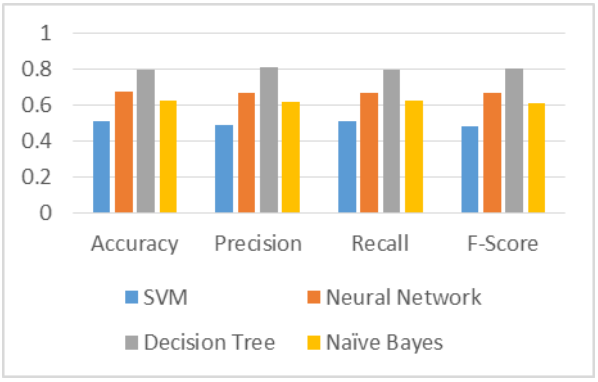

Figure 2 Evaluation Metrics for four Machine learning Models

From table 5, it is clear that the Decision Tree classifier can learn the feature to distinguish three levels of productivity classes with minimum error in comparison to other classifiers. It has good accuracy in medium productivity classes in comparison to the other two classes.

Table 5 Confusion Matrix for Decision Trees

Actual\Predicted	Low	Medium	High
Low	31	8	0
Medium	3	35	3
High	0	6	15

From table 6, it is clear that the Neural Network classifier is not able to learn the feature to distinguish Medium productivity class from low productivity class. It has good accuracy in the Medium productivity class in comparison to the other two classes.

Table 6 Confusion Matrix for Neural Network

Actual\Predicted	Low	Medium	High
Low	30	8	1
Medium	9	27	5
High	2	8	11

CONCLUSION AND RECOMMENDATION

The comparison of classification algorithms is a complex task and it is an open problem. First, the notion of performance can be defined in many ways: accuracy, speed, cost, reliability, etc. Second, an appropriate tool is necessary to quantify this performance. Third, a consistent method must be selected to compare with the measured values. The selection of the best classification algorithm for a given dataset is a very widespread problem. In this sense, it requires to make several methodological choices. So, this research focused on the comparative analysis of decision tree methods for the prediction of paddy productivity.

From the result analysis, it was seen that Decision Tree (SimpleCart) was able to classify 80.19% of the data correctly which was better than SVM, Naïve Bayes and Neural Network in comparison to results of evaluation metrics (Accuracy, Precision, Recall, and F-Measure). The text must be in either English. The submitted paper must be in its final form and of good appearance because it will be printed as it is without any editing. The paper must be typed in one column as it appears in this document.

REFERENCES

Agriculture Report, M. (2017). *Project Document of Prime Minister Agriculture Modernization Project (PM-AMP),*. Retrieved from Kathmandu: https://pmamp.gov.np/en/home/

Bashir, K., Rehman, M., & Bari, M. (2019). Detection and classification of rice diseases: An automated approach using textural features. *Mehran University Research Journal of Engineering and Technology*. ***38*(1)**, 239-250.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*. ***11*(1)**, 10-18.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques* (Vol. 10). Waltham, MA.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural computation*. ***13*(3)**, 637-649.

Kumar, R., Singh, M., Kumar, P., & Singh, J. (2015). *Crop Selection Method to maximize crop yield rate using machine learning technique*. Paper presented at the 2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM), Chennai, India.

Nirkhi, S. (2010). *Potential use of artificial neural network in data mining.* Paper presented at the 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE).

Nithya, A., & Sundaram, V. (2011). Identifying the rice diseases using classification andbiosensor techniques. *International Journal of Advanced Research in Technology*. ***1*(1)**, 76-81.

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*. ***2*(1)**, 37-63.

Raorane, A., & Kulkarni, R. (2012). Data Mining: An effective tool for yield estimation in the agricultural sector. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, ***1*(2)**, 1-4.

Revathi, P., Revathi, R., & Hemalatha, M. (2011). Comparative study of knowledge in Crop diseases using Machine Learning Techniques. *Inter-national Journal of Computer Science and Information Technologies (IJCSIT)*. ***2*(5)**, 2180-2182.

Salzberg, S. L. (1994). *C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993*: Kluwer Academic Publishers.

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). *Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification*. Paper presented at the 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, United Kingdom.

Veenadhari, S., Mishra, B., & Singh, C. (2011). Soybean productivity modelling using decision tree algorithms. *International Journal of Computer Applications*. ***27*****(7)**. 11-15.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*. ***14*****(1)**. 1-37.