An Optimized Feature Selection and Classification of Soil Parameters for Crop Yield Prediction

S.Manimekalai, Ph.D., Research Scholar, PG and Research Department of Computer Science, Chikkanna Government Arts College, Tirupur, India, manimekalaipari@gmail.com

K.Nandhini, Assistant Professor, PG and Research Department of Computer Science, Chikkanna Government Arts College, Tirupur, India, krishnandhini@yahoo.com

Abstract—In India, agriculture is generally considered as an essential contributor to the economic growth. Mainly, the agricultural activities depend on the seasonal changes, biological and economic causes. The agricultural yield prediction is a challenging and desirable process for each country. In modern world, cultivators are struggling to produce the yield since unpredictable climatic changes and drastically reduce in water resources. Crop yield prediction is achieved by using historical data about crop yield, weather and soil. Many classification algorithms have been proposed to predict the crop yield precisely. This article presents a wheat yield prediction based on Artificial Neural Network (ANN) classification with Feature Selection (FS) using Particle Swarm Optimization (PSO). Initially, different soil parameters are collected and the most relevant features are selected by using PSO. Then, the selected features are classified by ANN model to obtain the trained model. By using the trained model, the quality of crop yield is predicted and compared the performance of the proposed system. Finally, the experimental results show that the proposed ANN classification with FS achieves better performance than the other techniques.

Keywords—Crop yield prediction, Soil characteristics, ANN, Feature selection, PSO.

I. INTRODUCTION

Agriculture plays a significant role in most of the countries like India where the demand of food has been increased due to an unconditional growth of population. In earliest years, people cultivate the crops in their own land and also natural crops are harvested which have been used by different people, animals and birds [1] since such crops can provide healthy life. In modern years, the agriculture is slowly degrading due to the innovative technologies. Also, people don't have the knowledge about the cultivation of the crops in a right time and at a suitable place. By using the modern techniques, the seasonal climatic conditions are also being changed against basic properties such as soil, water and air which provide to insecurity of food. There is no proper solution and technologies by analyzing the issues and problems like weather, temperature, etc., for avoiding the agricultural degradation.

There are different ways for increasing the economical growth in the agricultural field. There are several ways to increase the crop yield and the quality of the crops. Different data mining algorithms like decision trees, naive bayes, etc., are also used for predicting the crop yield production efficiently [2-3]. Data mining refers the process of finding correlations or patterns in huge relational databases. The patterns, associations or relationships among all this data can provide information which can be converted into knowledge about historical patterns and future trends [4]. For instance, information about crop production can support the cultivators in order to identify the crop losses and prevent it in future. Accurate information about history of crop yield is an essential to make decisions related to agricultural risk management.

Hence in this paper, wheat yield prediction using feature selection is proposed by considering different soil parameters. Initially, different soil parameters such as phosphorus (P), potassium (K), sulphur (S), calcium (Ca), magnesium (Mg), zinc (Zn), copper (Cu), iron (Fe) and manganese (Mn) are collected. Then, the PSO algorithm is used to select the most relevant parameters to train the classification model. For achieving classification, ANN is

 DOI : 10.18231/2454-9150.2018.0666

proposed that helps us to predict the crop yield quality efficiently.

The rest of the article is structured as follows: Section 2 presents the previous researches related to crop yield prediction. Section 3 explains the proposed crop yield prediction using different classifiers. Section 4 illustrates the performance effectiveness of the proposed algorithms and Section 5 concludes the research work.

II. LITERATURE SURVEY

Agricultural crop yield prediction [5] was proposed by using Artificial Neural Network (ANN). In this approach, different soil and environmental parameters were considered for predicting the suitable crop. Here, the prediction was achieved by using feed forward back propagation ANN. ANN with zero, one and two hidden layers were considered. Moreover, optimum numbers of hidden layers and each unit in hidden layers were obtained based on the Mean Square Error (MSE) computation. However, ANN can work only with the numerical information and the performance was depending on the user's ability.

Ensemble machine learning model [6] was proposed for crop yield prediction. In this model, AdaBoost algorithm was ensemble with Support Vector Machine (SVM) and Naive Bayes as AdaSVM and AdaNaive. This ensemble model was used for predicting the crop production over a given time period. Initially, historical crop production data and climatic data were collected and combined together. Then, the model was built by classifying the number of input data based on the SVM and naive bayes algorithms. This process was improved by using this ensemble classification model. However, this model requires an improvement on prediction performance.

Improved crop yield prediction [7] was proposed by using neural network. In this system, initially different data related to the crop yield like types of crops, crop diseases, soil pH level, etc., including environmental parameters were gathered. Then, the gathered data was normalized and classified by using neural network. The prediction of crop yield was performed on an adaptive cluster approach over dynamically updated historical crop data and improved the decision making process. However, the performance efficiency was not analyzed precisely.

Crop recommendation system [8] was proposed for improving the crop yield by using machine learning technique. In this system, a voting model was used for constructing an efficient and accurate prediction model. Initially, different parameters related to crop, soil and environment were collected by soil testing lab dataset. Then, the obtained data was given to the recommendation system which has ensemble model with majority voting technique using SVM and ANN as learners for recommending a crop for specific parameter. However, this system utilizes only few numbers of attributes.

Analysis of soil behavior and prediction of crop yield [9] was proposed by using data mining approach. In this analysis, the problem of crop yield prediction was formulated as a classification rule where naive bayes and K-Nearest Neighbor (KNN) algorithms were used. Here, the soil dataset was collected from the soil testing laboratory Jabalpur, Mathya Pradesh. The prediction accuracy was computed by evaluating the datasets. The classification algorithms were applied on the soil dataset to predict the crop yield effectively. However, an automatic prediction system was required to improve the crop yield prediction.

Bitter melon crop yield prediction [10] was proposed by using machine learning algorithm. In this technique, bitter guard leaves dataset was collected from Ampalaya farms. The collected leaves were classified using convolutional neural network into good or bad via their description. Training of data was achieved through the abilities of Keras, Tensor Flow and Python worked together. However, processing time of this system was high.

III. PROPOSED METHODOLOGY

In this section, the proposed crop yield prediction using different classifiers are explained in detail. Initially, different soil parameters such as P, K, S, Ca, Mg, Zn, Cu, Fe and Mn are gathered. The most relevant features are selected based on the PSO algorithm. Then, the dataset with selected features is split into training and testing dataset. The training dataset is given to the ANN classifier as input

 DOI : 10.18231/2454-9150.2018.0666

to learn the soil parameters. Finally the testing dataset is used to predict the crop yield based on the learned soil parameters. The overall process of the proposed system is shown in Fig. 1.

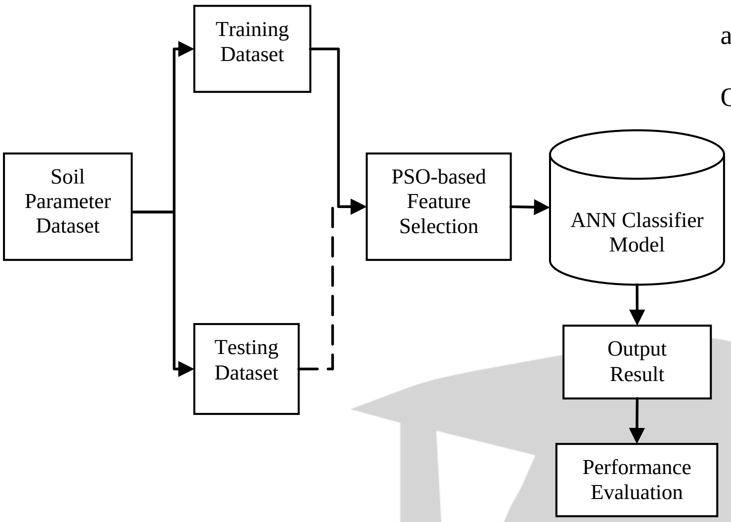

Fig. 1. Process of Proposed Crop Yield Prediction

A. *Feature Selection based on PSO*

The most optimal features (parameters) are selected based on the PSO algorithm. Consider each parameter as a particle and each particle are represented as $\{P_1, P_2, \dots, P_n\}$. The optimal parameter with optimal fitness function $f(P)$ is selected by optimizing the particle sequences where $f(P)$ consists of high prediction accuracy.

$$fitness(P) = max(prediction\ accuracy) \qquad (1)$$

After number of iterations, the best optimal parameter P with the optimal fitness function $f(P)$ is selected. The location of i^{th} particle at time period T and d-dimensional search space is $l_i^t = (l_{i1}^t, l_{i2}^t, \dots, l_{id}^t)^T$ and its velocity is $v_i^t = (v_{i1}^t, v_{i2}^t, \dots, v_{id}^t)^T$. The best previous location of the particle is denoted as $l_{ib} = (l_{ib1}^t, l_{ib2}^t, \dots, l_{ibd}^t)^T$ and the best global location of the entire particle is denoted as l_g. Hence to obtain the best optimal parameter, the location and velocity of each particle is updated in each iteration process (t) as follows:

$$v_{id}^{t+1} = \omega v_{id}^t + c_1 rand_1 (l_{ibd}^t - l_{id}^t) + c_2 rand_2 (l_{gd}^{t-1} - l_{id}^t) \qquad (2)$$

$$l_{id}^{t+1} = l_{id}^t + v_{id}^{t+1} \qquad (3)$$

In equation (2), c_1 and c_2 are referred as learning factors and ω is the inertia coefficient. Also, $rand_1$ and $rand_2$ are denoted as the random numbers which are uniformly distributed.

Algorithm for PSO-based Feature Selection

Input: Soil parameters such as P, K, S, Ca, Mg, Zn, Cu, Fe and Mn

Output: Most optimal parameter

1. Begin
2. Initialize n number of particles (parameter)
3. Initialize its initial location l_i^t and velocity v_i^t
4. $While(t! = t_{max})$
5. for each particle do
6. Compute the fitness function by using (1)
7. Find $pbest$ of each particle
8. Obtain the $gbest$ among $pbest$
9. Select any two random variables as $rand_1$ and $rand_2$
10. Update particle location and velocity by using (2) & (3)
11. End for
12. End $While$
13. Obtain the optimal parameter P
14. End

B. *ANN-based Classification*

The selected parameters are given to train the ANN classifier. ANN has three layers namely input, hidden and output layer. The probabilities are denoted as $f(x) = x$ are given to the input layer of neurons. The hidden layer of ANN is defined as tan-sigmoid transfer function.

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \qquad (4)$$

Each input has its own weight values as $w_1, w_2, \dots, w_n$ and the weighted sum of the inputs is done by the adder function as follows,

$$u = \sum_{i=1}^{n} w_i x_i \qquad (5)$$

The output layer of ANN is described by the following equation.

$$y = f(\sum_{i=1}^{n} w_i x_i + b_i) \qquad (6)$$

In the equation (6), y is the output neuron value i.e., low crop yield or high crop yield; $f(x)$ is the transfer function, w_i refers the weight values, x_i denotes input data values and b_i refers to the bias value. Based on the output neuron values, the relationship between crop yield quality and the

 DOI : 10.18231/2454-9150.2018.0666

considered parameters is learned which generates the prediction models. The basic structure of ANN is shown in Fig. 2.

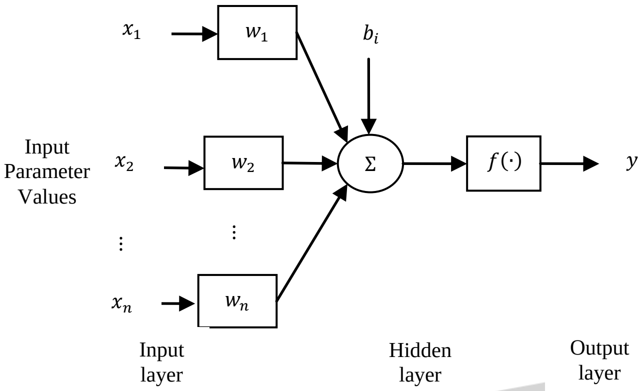

Fig. 2. Basic Structural Design of ANN

The parameter values for ANN are given in Table 1.

TABLE 1. SOIL PARAMETERS FOR ANN

Parameters	Values
Input layer neurons	70
Hidden layer neurons	50
Output layer neurons	1
Learning rate	0.01
Transfer function	Tan-Sigmoid
Maximum number of iteration	100

During training, learning rate is used to control the weight and bias value changes in each iteration process i.e., each updation of weight and bias values. By configuring the parameters mentioned in Table 1, training dataset can be trained using ANN. Based on this trained model, the crop yield quality i.e., low or high is predicted efficiently.

IV. RESULTS AND DISCUSSION

In this section, the performance of each classifier in crop yield prediction is evaluated and compared by using MATLAB 2018a. In this experiment, soil parameters are collected for wheat crop. The comparison is made in terms of different metrics such as precision, recall, f-measure and accuracy.

MATLAB is a powerful and flexible tool for data mining. It consists of predefined complex statistical modules. Thus, there is no need of writing an individual code and the code is compact. For testing the complex procedure, it requires less time. User can write his/her own function which he/she can use several times when required for data analysis. It handles very huge database. Moreover, it is open source software system.

A. Precision

Precision is computed based on the prediction at true positive and false positive rates.

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

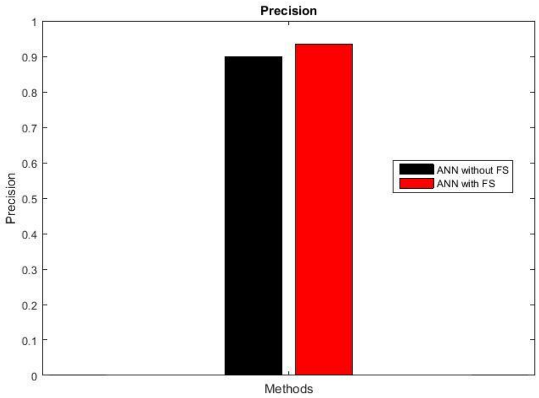

Fig. 3. Comparison of Precision for ANN without FS and ANN with FS

Fig. 3 shows the comparison of precision for ANN with and without FS technique in wheat yield prediction. From the analysis, it is observed that ANN with FS algorithm has high precision than the without FS algorithm.

B. Recall

Recall value is calculated based on the prediction value at true positive and false

negative predictions.

$$Recall = \frac{TP}{TP + False\ Negative\ (FN)}$$

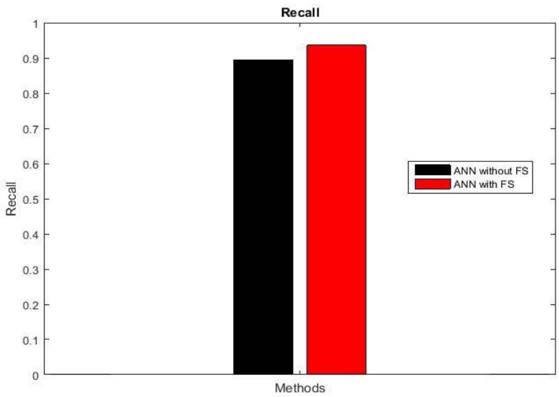

Fig. 4. Comparison of Recall for ANN without FS and ANN with FS

 DOI : 10.18231/2454-9150.2018.0666

Fig. 4 shows the comparison of recall for ANN with and without FS technique in wheat yield prediction. From the analysis, it is observed that ANN with FS algorithm has high recall than the without FS algorithm.

C. F-Measure

F-measure is computed by using both precision and recall as follows:

$$F - measure = 2 \cdot \left(\frac{Precision \cdot Recall}{Precision + Recall}\right)$$

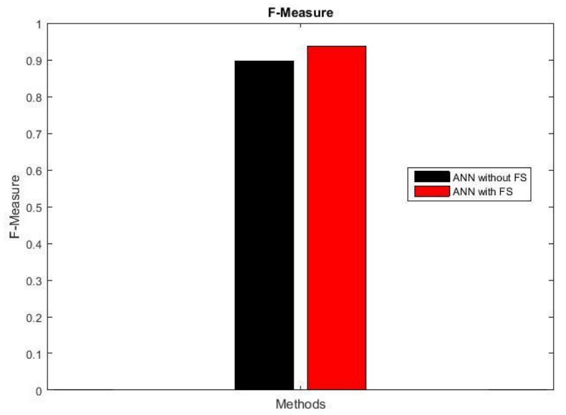

Fig. 5. Comparison of F-Measure for ANN without FS and ANN with FS

Fig. 5 shows the comparison of f-measure for ANN with and without FS technique in wheat yield prediction. From the analysis, it is observed that ANN with FS algorithm has high f-measure than the without FS algorithm.

D. Accuracy

It defines the ratio of true positives and true negatives to the sum amount of cases examined. It is measured as,

$$Acc = \frac{TP + True\ Negative\ (TN)}{TP + TN + FP + FN}$$

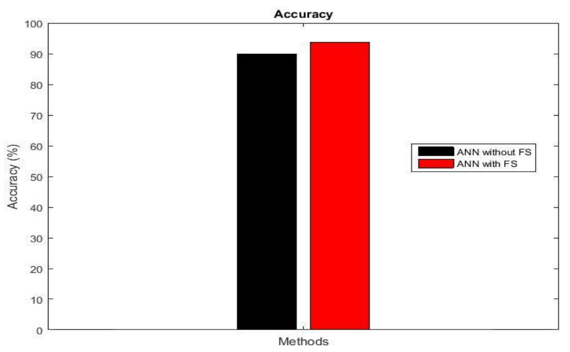

Fig. 6. Comparison of Accuracy for ANN without FS and ANN with FS

Fig. 6 shows the comparison of precision for ANN with and without FS technique in wheat yield prediction. From the analysis, it is observed that ANN with FS algorithm achieves high accuracy than the without FS algorithm.

V. CONCLUSION

In this article, a comparative analysis of crop yield prediction using different classifiers is presented. In this analysis, several soil parameters are gathered and the most optimal parameters are selected based on the PSO algorithm. Once the parameters are selected, they are trained by using ANN classification to predict the crop yield quality. Then, the performance of proposed ANN with feature selection is analyzed and compared with the existing method through experimental results in terms of precision, recall, f-measure and accuracy. From the analysis, it is suggested that the proposed ANN with FS algorithm outperforms than the ANN without FS algorithm.

REFERENCES

[1] P. Priya, U. Muthaiah, and M. Balamurugan, "Predicting yield of the crop using machine learning algorithm," Int. J. Eng. Sci. Res. Technol., vol. 7, no. 4, pp. 1-7, 2018.

[2] D. Ramesh, and B. V. Vardhan, "Analysis of crop yield prediction using data mining techniques," Int. J. Res. Eng. Technol., vol. 4, no. 1, pp. 47-473, 2015.

[3] S. Veenadhari, B. Mishra, and C. D. Singh, "Soybean productivity modelling using decision tree algorithms," Int. J. Computer Appl., vol. 27, no. 7, pp. 11-15, 2011.

[4] R. Sujatha, and P. Isakki, "A study on crop yield forecasting using classification techniques," in IEEE Int. Conf. Computing Technol. Intell. Data Eng., pp. 1-4, 2016.

[5] S. S. Dahikar, and S. V. Rode, "Agricultural crop yield prediction using artificial neural network approach," Int. J. Innov. Res. Electr. Electron. Instrum. Control Eng., vol. 2, no. 1, pp. 683-686, 2014.

[6] N. Balakrishnan, and G. Muthukumarasamy, "Crop production-ensemble machine learning model for prediction," Int. J. Computer Sci. Softw. Eng., vol. 5, no. 7, p. 148, 2016.

[7] P. P. Bhangale, Y. S. Patil, and D. D. Patil, "Improved crop yield prediction using neural network," IJARIIE, vol. 3, no. 2, pp. 3094-3101, 2017.

[8] R. K. Rajak, A. Pawar, M. Pendke, P. Shinde, S. Rathod, and A. Devare, "Crop recommendation system to maximize crop yield using machine learning technique," Int. Res. J. Eng. Technol., vol. 4, no. 12, pp. 950-953, 2017.

[9] D. M. Supriya, "Analysis of soil behaviour and prediction of crop yield using data mining approach," Int. J. Innov. Res. Computer Commun. Eng., vol. 5, no. 5, pp. 9648-9653, 2017.

[10] M. B. Villanueva, and M. L. M. Salenga, "Bitter Melon Crop Yield Prediction using Machine Learning Algorithm," Int. J. Adv. Computer Sci. Appl., vol. 9, no. 3, pp. 1-6, 2011.

 DOI : 10.18231/2454-9150.2018.0666