Intelligent Chemical Fertilizer Recommendation System for Rice Fields

Shiva Hassanjani Roshan¹, Seyyed Javad Kazemitabar² & Ghorban Kheradmandian³

¹Master of Telecommunications, Babol Noshirvani University of Technology, Iran. Email: hasanjani.eleceng@gmail.com
²Assistant professor of Electrical Engineering, Babol Noshirvani University of Technology, Iran. Email: j.kazemitabar@nit.ac.ir
³Tosan Intelligent Data Miners, Iran. Email: kheradmand@aut.ac.ir

DOI: http://doi.org/10.38177/ajast.2021.5318

Copyright: © 2021 Shiva Hassanjani Roshan et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Article Received: 28 June 2021  Article Accepted: 29 August 2021  Article Published: 30 September 2021

ABSTRACT

In this paper, a recommendation system for supplementary chemical fertilizers of rice fields is proposed using data mining methods. Traditionally, an expert determines the necessary amount of chemical fertilizer for each field after testing the amount of existing organic materials in the soil. The recommendation provided by the expert is a combination of agricultural science and region-specific conditions. In this paper, thru recognizing the existing pattern in recommendations proposed by two groups of experts for the agricultural lands in Mazandaran Province in recent years, a predictive model is proposed. Different artificial intelligence techniques are compared with each other and the best one among them is introduced.

Keywords: Chemical fertilizer, Rice field, Mazandaran, Support vector machine, Random forest, xgboost, Genetic algorithm.

1. Introduction

Agriculture industry highly depends on good quality raw agricultural supplies. Therefore harvesting acceptable quality products is one of the most important concerns in the agriculture industry. Crops and vegetables being the main resource of a major part of such industries, take an important role in the economy associated with it. Plants need rich soil in addition to water for their growth and vitality. Richness is supplied by chemical and non-chemical fertilizers. Fertilization is one of the most important factors that determine the crop yield in agriculture. However, more fertilizer does not always mean more crops. Either improper use or unreasonable proportion of fertilizer could result in the reduction of crop yield. As a matter of fact, excessive consumption of chemical fertilizers with heavy elements such as cadmium, lead, and zinc has inflicted much damage to the environment as well as risking health in people consuming the affected products. The proper way is to use fertilizer according to the demand of the crop and the measurement of soil nutrients. Thus, it is necessary to evaluate the fertilizer of soil and study the relation between the fertilizer and crop yield.

2. Related Work

Recently, scientists have recommended machine learning based methods for predicting the proper amount of fertilizer used in the fields. Generally, two groups of factors influence plant growth which are genetic and environmental factors. Genetic factors are related to gene structure of the plant. Temperature, humidity, radiant energy, and so on are some of the environmental factors which could affect plant growth. Some of the environmental factors such as light might not be that controllable but there are still other factors that can be controlled with correct management.

As one of the most crucial environmental factors, soil moisture causes the absorption of natural nutritional materials by the plant. If soil moisture becomes more or less than its defined allowed limit, the process of plant growth will be problematic (Ntukamazina 2017, Chang 2017). In (Chatterjee 2018), Artificial Neural Network
(ANN) is recommended for predicting the amount of soil moisture. Authors use Modified Flower Pollination Algorithm (MFPA) for training the artificial neural network and the result of the experiment was compared with ANN methods based on Particle Swarm Optimization algorithm (PSO) and Cuckoo Search algorithm (CS). Ultimately, NN-MFPA was introduced as the superior model with respect to other models. In another study (Gago 2015), the use of autonomous vehicles was suggested for water consumption management in order to implement a stable agricultural system. This system supervises the amount of consumed water using temporal and spatial sensing tools which are equipped with remote measurement modules. Gill et al. proposed Support Vector Machine (SVM) for predicting the amount of soil moisture in (Gill 2006). Repetition is the foundation of learning in neural networks. Backpropagation method is one of the methods for repetition in the neural network.

In (Gill 2006), the disadvantages of neural networks based on backpropagation algorithm are examined and compared with the suggested SVM model. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were studied for this task. The acquired results showed that SVM is a better choice for predicting the amount of soil moisture compared to the neural network. Esmaeelnejad et al. suggested a neural network based model for predicting the amount of soil moisture in Gilan, in northern Iran (Esmaeelnejad 2015). The proposed neural network of these researchers included two hidden layers for predicting the amount of soil moisture. The amount of phosphorus in soil is another environmental factor for soil fertility. Considering the importance of this matter, the artificial neural network was implemented and experimented for estimating the amount of phosphorous in Kuhin drainage basin in Qazvin Province, Iran in (Keshavarzi 2015). In the study, 85 samples were gathered from 1000 hectares of the selected location for experimenting the contents of the phosphorous and other corresponding features were predicted by the Digital Elevation Model (DEM). 80 percent of the gathered samples were used for the learning section and 20 percent of them were used for testing. The parameters of RMSE and the coefficient of determination ($R^2$) were considered for the evaluation of the proposed model in (Keshavarzi 2015). Measuring the coefficient of determination shows the amount of trust in the prediction model. Through the learning process of the ANN model, the learning rate was measured by fitness functions. Finally, a network with the least error rate was selected. The final results showed the accuracy of the proposed model and this model could determine the amount of phosphorous in the soil using topographic features of the soil. Project (Samadianfard 2018) was conducted in the town of Adana in Turkey for the purpose of monthly estimation of the temperature of the soil and in the depths of 5, 10, 20, 50, and 100 centimeters from the ground. The researchers have introduced a hybrid multilayered perceptron neural network based on the firefly optimization algorithm. The values of RMSE, MAE, MAPE, and MBE were considered for the evaluation of the proposed model and Taylor diagram was used to determine the amount of similarity between the predicted soil moisture and the amount of observed soil moisture.

3. Materials And Method

A. Study Site Description and Soil Sampling

This study was conducted with the purpose of examining the amount of physical and chemical elements on advising chemical fertilizers in the rice fields of Mazandaran, northern Iran (Fig. 1). The area of the field, the type of rice, the soil texture, the saturated mud acidity (pH), electrical conductivity (EH), saturated precipitation (SP),
organic carbon (OC), total neutralizing value (TNV), the organic material percentage (OM), total nitrogen percentage (Ntotal), absorbable phosphorous (Pava), absorbable potassium (Kava), oxide potential (Eh), zinc (Zn), sand percentage, silt percentage, and clay percentage were considered as the input characteristics. Urea, TSP, SOP, ZnSO₄ are the advised fertilizers for the rice field. The dataset used in this research was gathered by Babol pedology lab and rice research institute of Iran located in Amol. This dataset is the result of 161 physical and chemical experiments of the soil and existing nutritional elements in it.

Fig.1. Location of the conducted study

B. Preliminaries

The dataset used in this study, had 161 records. Each record would reflect the following fields for a specific land: pH, E.C, SP, O.C TNV, O.M., Ntotal, Pava, Kava, Zn, Eh, clay, silt and sand. There were nearly 300 missing fields out of the total of 2737. Missing values of dataset records are either due to lost information or information that are not presented. This is a common problem in real world data that researchers have to cope with. The loss of data can heavily affect the modelings done in the data mining process or might even lead to a wrong conclusion and decision. Some of the techniques for missing entries are as follows:

1. Data removal
2. Replacing the respective value with a random value
3. Replacing the lost value with mean, median, or the mode of that column of features
4. Using centrality properties (e.g. median, mean, and so on) of the classified data
5. Predicting the lost data using the method of prediction and the most probable value

In the current study, the random forest algorithm is used to predict the lost data.

C. Overview of Artificial Intelligence Techniques

1) Random Forest Algorithm

Random forest algorithm was introduced in 2001 for the first time by Breiman. Random forest is a supervised classification algorithm which consists of a series of decision trees. Decision tree algorithm is a form of classification (Dasheng 2016 & Niuniu 2015) and one of the most common data mining techniques. Due to its simplicity and efficiency it is extensively used in different applications and problems of machine learning. In the decision tree, by following a series of questions related to the data characteristics and examining the current data for making a decision, its class or category is determined. CART is a binary tree algorithm in the decision tree algorithm. Random forest is a series of CART trees and it is expressed in four steps (Li 2017):

Step.1: \( k \) subsets of training samples \((D_1, D_2, \cdots, D_k)\) among all existing samples in the training section \((D)\) is selected by the Bootstrap sampling method. Ultimately, \( k \) decision trees will be created.

Step.2: Among \( N \) indices of the classification tree node, \( m \) characteristics are randomly selected and the best characteristic is selected among \( M \) indices according to the least node purity principle. The trees will grow that way.

Step.3: This step is the repetition of step 2. As a result, \( k \) decision trees are created.

Step.4: \( k \) decision trees which have grown quite well will form the combinatorial random classifier forest. The sample positioned at the last class of the random forest will wait for the majority vote as shown below:

\[
f(x_t) = \text{majority vote}\{h_t(x)\} \ (i = 1, 2, \ldots, k) \tag{1}
\]

The block diagram of the random forest algorithm is shown in Fig. 2.

Figure 2: Random forest algorithm

2) Support Vector Machine Algorithm

Support vector machines (SVMs) which were proposed by Vapnik (Bierman 2001) are one of the supervised machine learning algorithms that can be used for both classification and regression problems. Support vector machine is a border which separates the classes of data in the best way possible. SVM problem is a constrained convex quadratic optimization problem.

The fitness function is one of the most essential factors in SVM and its main goal is to use the function \( f(x) = wx + b \) in proportion to the data \( \{x_i, y_i\}, x_i \in R^n, y_i \in R^n \ (i = 1, 2, \ldots, n) \) (Fuxue 2010). According to the SVM theory, the fitness function is defined in relation (2).

\[
f(x) = wx + b = \sum_{i=1}^{\infty} (a_i - a_i^*) \cdot k(x, x_i) + b \tag{2}
\]

In relation (2), \( a_i \) and \( a_i^* \) are support vector machines and only a small portion of their values are non-zero. \( a_i, a_i^* \), and \( b \) can be acquired through quadratic optimization problems.

ISSN: 2456-883X

www.ajast.net
3) XGBoost Algorithm

XGBoost is a combinatorial optimization based on Gradient Boosting Decision Tree (GBDT). The main idea of the boosting algorithm comes from the fact that the performance of many decision trees is better than a single one. In fact, a single decision may not present an acceptable result. But when these trees are used in large numbers, they show better performance. This leads to an enhancement of data processing speed and more accurate results.

The fitness function of XGBoost is defined in (3):

$$F_{Obj}(\theta) = L(\theta) + \Omega(\theta) \quad (3)$$

Where:

$$L(\theta) = I(y_i, \hat{y}_i) \quad , \Omega(\theta) = \gamma T + \frac{1}{2} \lambda \| \omega \|^2$$

As seen in (3), the fitness function is composed of two sections. $L(\theta)$ is a convex cost function which shows the difference between $y_i$ and $\hat{y}_i$. $\Omega(\theta)$ is related to the adjustment section. $T$ is the number of the leaves of the tree and $\gamma$ is the learning rate. $\gamma$ has a value between $[0, 1]$.

![Fig.2.](image)

XGBoost causes the $\frac{1}{2} \lambda \| \omega \|^2$ section to increase in comparison with GBDT algorithm. $\lambda$ is the adjustment parameter and $\omega$ shows the leaf weight. Increasing this item can lead to an increase in the capability of the model (Jarraya 2014).

4) Genetic Algorithm

In the field of evolutionary modeling in biological systems in the 1960s, some works were done by a biologist named Faster. The genetic algorithm, however, was proposed by John Holland, computer scientist in the University of Michigan, for engineering applications and in its modern form in 1975. Later, this method gained its popularity through the efforts of Goldberg in 1989 and has a respectable position among other methods today due to its abilities (Sivanandam 2008). This method is based on the theory of gradual evolution and fundamental ideas of Darwin. In this algorithm, the fit members of a generation are bred for creating a generation of potentially fitter members. The genetic algorithm has two general models which are continuous and binary models. The difference between the genetic algorithm and the classic optimization methods are two general points:

- In genetic algorithm, a population consisting of points are created in each iteration instead of a single point.
Unlike conventional methods, random selections are used instead of deterministic selections while selecting the next population in the calculations.

In the following, the properties of the evolutionary genetic algorithm are explained:

**Population:** One of the important genetic methods is focusing on a population of chromosomes instead of a single chromosome as a point in the search space. A set of chromosomes in the search space is called a population. According to this algorithm, each generation includes chromosomes which have more features than the population of the previous generation. Each population or generation of chromosomes has a size which is known as the population size.

**Search Space:** Every point in the search space shows an acceptable solution. The purpose of searching the search space is finding a point or points as the best solutions of the problem.

**Fitness Value:** The value acquired from each generation through the fitness function determines the fitness of the problem.

**Chromosome:** In the genetic algorithm, a chromosome is a series of parameters in a way that defines a proposed solution for the problem which the genetic algorithm is trying to solve.

**Continuous Genetic Algorithm:** In the continuous genetic algorithm, the goal is to solve optimization problems in which an optimal solution for the problem variables is searched. In this algorithm, the problem variables are shown with a chromosome which is an array of different values to be optimized (Haupt 2004).

For example, if a chromosome has $N_{\text{var}}$ variables, a $N_{\text{var}}$-dimensional optimization problem is considered whose parameters are shown with $P_1, P_2, \ldots, P_{N_{\text{var}}}$. Therefore, each chromosome is determined with a $1 \times N_{\text{var}}$ array. This means:

$$\text{Chromosome} = [P_2, P_2, \ldots, P_{N_{\text{var}}}]$$

(4)

In this case, the values of the variables are in floating point values. A chromosome has a fitness which is acquired through the evaluation of the cost function $f$ with variables $P_1, P_2, \ldots, P_{N_{\text{var}}}$. Meaning:

$$\text{Cost} = f(\text{Chromosome}) = f(P_1, P_2, \ldots, P_{N_{\text{var}}})$$

(5)

**Initial Population:** In order to start the continuous genetic algorithm, an initial population is generated from $N_{\text{pop}}$ chromosomes.

The matrix showing the population is a $N_{\text{pop}} \times N_{\text{var}}$ matrix in which each line represents a population of each variable.

$$\text{Pop} = rand(N_{\text{pop}}, N_{\text{var}})$$

(6)

All defined variables are generated in the interval of $[0, 1]$ and are normalized by a uniform random number generator. If the range of values be in the interval of $[P_{\text{Lo}}, P_{\text{Hi}}]$, the unnormalized values are acquired through relation (7):

$$P = \frac{P_{\text{Lo}} \times (P_{\text{Hi}} - P_{\text{Lo}}) + P_{\text{Lo}}}{2} + \frac{P_{\text{Hi}} - P_{\text{Lo}}}{2}$$

(7)
\[ p = (p_{Hi} - p_{Lo})p_{\text{norm}} + p_{Lo} \]  

(7)

In which \( p_{Hi} \) and \( p_{Lo} \) are the largest and the smallest numbers in the actual range of the variables respectively. Also, \( p_{\text{norm}} \) is the normalized value of the parameter. Then, each chromosome is evaluated with the cost function and the evaluation results of each one of them is determined. These cost values are sorted from the smallest to the largest value so that they could be used in the next steps of the algorithm.

**Natural Selection:** Through the natural selection of each chromosome, the fitter ones remain to be transferred to the next generation and is used for the rebreeding process in the next generation. Chromosomes having less fitness value are gradually eliminated from the population of that generation. Only \( N_{\text{keep}} \) chromosomes from the population who have the highest amount of fitness remain for mating.

**Parent Selection:** \( N_{\text{keep}} \) chromosomes with highest fitness value will breed among the pairs of parents that are randomly selected. Each selected pair will produce two children who have their parents’ features. The parents also remain as a part of the next generation. Now, the method and the criteria for selecting the parents are considered. According to Darwin’s evolutionary theory, the best of each generation remain for producing future generations. Various methods exist for the selection process. Roulette wheel selection, rank selection, and steady state selection can be mentioned as examples.

**Elitism:** Elitism is a method in which the prototype of the best chromosomes are directly carried over to the next generation for creating new generations.

**Mating:** After randomly selecting the generators, the children are generated from some of their combinations which is called a crossover process. This operator works on a pair of chromosomes.

**Mutation:** Using the genetic algorithm will be quite effective for solving optimization problems. But it should be taken into consideration that the high convergence speed of the algorithm is not desirable and it might converge the algorithm to an undesirable region of the solution space. If this solution space is around a global minimum, it will not cause a problem. But if the function under optimization has several local minima in this region, then the algorithm will not be able to find the best solution of the problem and it will finish by reaching a local minimum instead of a global one. In order to prevent the fast convergence problem, the algorithm must be forced to search other regions of the search space with random changes or mutations.

**Algorithm Convergence:** Conventionally, three criteria are considered to be the stopping criteria of the algorithm:

- Algorithm runtime
- The number of produced generations
- Error criterion convergence

**Encoding:** In the genetic algorithm, the encoded form of the parameters or the variables of the problems can be used instead of the parameters and variables themselves. Conventional encoding methods in the genetic algorithm are direct encoding, binary encoding, mutational encoding, value encoding, and tree encoding (Dasheng 2016).
Binary Genetic Algorithm: The most important and basic form of the genetic algorithm is the binary genetic algorithm in which variables are encoded in the binary form. This type of genetic algorithm is also called the discrete genetic algorithm (Haupt 2004) because variables in this algorithm do not have continuous changes and cannot take any value. The set of problem variables whose optimal value has to be found are encoded in the form of binary strings and are concatenated to each other. Like the continuous genetic algorithm, this algorithm starts with defining the parameters of the algorithm and the cost function. The purpose of this algorithm is modifying the output with a favorable method in order to find the appropriate values of input variables. After executing the simulator program, the number which defines the standard deviation or the fitness of that set of information will be assigned to that member of the given population. The binary genetic algorithm starts with an array of $N_{pop} \times N_{bits}$ variable values which are going to be optimized. $N_{bits}$ is the number of bits that each chromosome is going to be encoded with. In this algorithm, each chromosome has $N_{var}$ variables (a $N_{var}$-dimensional optimization problem) with parameters $P_1, P_2, \ldots, P_{N_{var}}$.

Chromosome = $[P_1, P_2, \ldots, P_{N_{var}}]$ (8)

The binary genetic algorithm works in bits. Each variable $x$ has a value which is displayed as a string of bits with the length of $N_{gene}$. All steps of this algorithm is similar to the mentioned steps for the continuous genetic algorithm. But there is a small difference in the mutation in the binary state with the continuous state due to its bit nature. The mutation in the binary state is in this way that each selected mutation variable will become one if its value is zero.

Fig.3. Block diagram of the continuous genetic algorithm
and vice versa so that the fast convergence of the algorithm which impedes acquiring the global minimum is prevented. The necessary number of bits for the mutation is acquired through (9) (Haupt 2004):

\[
\text{Mutations} = \mu \left( N_{\text{pop}} - 1 \right) \times N_{\text{bits}} \quad (9)
\]

The block diagram of the binary genetic algorithm is shown in Figure 4.

![Block diagram of the binary genetic algorithm](image)

**Fig.4.** Block diagram of the binary genetic algorithm

4. Results and Discussions

As mentioned, the dataset selected for this study was not complete and missing data were reported. The existence of missing data is considered a crucial problem and hinders the progress of the research. In this study, the random forest method is used to tackle this problem. This way, the missing data was recovered. Actually, the goal was to recommend the proper amount of supplementary urea, TSP, SOP, and ZnSO4 for the rice field. In order to predict these values, random forest, XGBoost, and SVM algorithms based on the genetic algorithm were implemented. The experiments above was executed and repeated in the R for each one of the advised fertilizers separately. The results of the mentioned programs are reported in the tables (1-4).

In the current study Mean Absolute Error (MAE) has been used.
Table 1. Mean absolute error for the prediction of urea

| Data mining technique | SVM algorithm based on the genetic algorithm | SVM algorithm | XGBoost algorithm | Random forest algorithm |
|-----------------------|-----------------------------------------------|---------------|------------------|------------------------|
| MAE value             | 0.34                                          | 0.33          | 0.46             | 0.48                   |

Table 2. Mean absolute error for the prediction of ZnSO4

| Data mining technique | SVM algorithm based on the genetic algorithm | SVM algorithm | XGBoost algorithm | Random forest algorithm |
|-----------------------|-----------------------------------------------|---------------|------------------|------------------------|
| MAE value             | 0.35                                          | 0.34          | 0.61             | 0.27                   |

Table 3. Mean absolute error for the prediction of TSP

| Data mining technique | SVM algorithm based on the genetic algorithm | SVM algorithm | XGBoost algorithm | Random forest algorithm |
|-----------------------|-----------------------------------------------|---------------|------------------|------------------------|
| MAE value             | 0.42                                          | 0.39          | 0.3               | 0.34                   |

Table 4. Mean absolute error for the prediction of SOP

| Data mining technique | SVM algorithm based on the genetic algorithm | SVM algorithm | XGBoost algorithm | Random forest algorithm |
|-----------------------|-----------------------------------------------|---------------|------------------|------------------------|
| MAE value             | 0.47                                          | 0.48          | 0.47             | 0.38                   |

The results show that for the prediction of urea, SVM algorithm is suitable. However, for the prediction of ZnSO4 and SOP Random forest algorithm and for the prediction of TSP XGBoost algorithm are suggested.

Acknowledgements

The authors are grateful to Dr. Mirnia, professor of Mazandaran University, who facilitated us with the datasets.
Declarations

Source of Funding

This research did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests Statement

The authors declare no competing financial, professional and personal interests.

Consent for publication

Authors declare that they consented for the publication of this research work.

Data Availability

The data that support the findings of this study are available on request from the corresponding author, J.K. The data are not publicly available due to restrictions.

References

Chang, J (2017), Climate and agriculture: an ecological survey, Routledge.

Chatterjee, S., N. Dey, S. Soumya (2018), Soil moisture quantity prediction using optimized neural supported model for sustainable agricultural.

Dasheng, P. & Q. Wanwen (2016), An improved ID3 decision tree mining algorithm. J. of Huaqiao University.

Esmaeelnejad, L., H. Ramezanpour, J. Seyedmohammadi, & M. Shabanpour (2015), Selection of a suitable model for the prediction of soil water content in north of Iran, Spanish Journal of Agricultural Research.

Fuxue S (2010), SVM in predicting the deformation of deep foundation pit in soft soil area, International Conference on Machine Vision and Human-machine Interface.

Gago, J., C. Douthe, R.E. Coopman, P.P. Gallego, M. Ribas-Carbo, J. Flexas, J. Escalona & H. Medrano (2015), UAVs challenge to assess water stress for sustainable agriculture, Agricultural Water Management 153.

Gill, M. K., T. Asefa, M. W. Kemblowski & M. McKee (2006), Soil moisture prediction using support vector machines, JAWRA Journal of the American Water Resources Association.

Haupt, R. & S. Haupt (2004), Practical to genetic algorithms, Second edition, John Wiley & Sons.

Jarraya, Y., T. Madi, & M. Debbabi (2014), A survey and a layered taxonomy of software-defined networking, IEEE Communications Surveys and Tutorials.

Keshavarzi, A., F. Sarmadian, E. O. El-Sayed, M. Iqbal (2015), A neural network model for estimating soil phosphorus using terrain analysis, The Egyptian Journal of Remote Sensing and Space Sciences.

Li, Y. & H. Wu (2017), Water bloom warning model based on random forest, Artificial Intelligence, Robotics and Human-Computer Interaction, Okinawa, Japan.
Niuniu, X (2015), A survey of decision tree algorithms software guide.

Ntukamazina, N. , R. N. Onwonga & R. Sommer (2017), Effect of excessive and minimal soil moisture stress on agronomic performance of bush and climbing bean (Phaseolus vulgaris L.), Cogent Food and Agriculture 3.1 Applications, Sustainable Computing: Informatics and Systems.

Samadianfard, S., M. A. Ghorbani, & B. Mohammadi (2018), Forecasting soil temperature at multiple-depth with a hybrid artificial neural network model coupled-hybrid firefly optimizer algorithm, Info. Processing in Agriculture.

Sivanandam, S. & S. Deepa (2008), Introduction to Genetic Algorithms, Springer.