DOI: https://doi.org/10.24113/ijoscience.v4i8.152

A Review on Soil Property Detection using Machine Learning Approach

Smriti Singhatiya
P.G. Student
Department of Computer Science and Engineering
MPCT
Gwalior, India
singhatiyasmriti@yahoo.com

Dr. Shivnath Ghosh
Associate Professor
Department of Computer Science and Engineering
MPCT
Gwalior, India

Abstract: **The agricultural sector is the backbone of the Indian economy. Although focused on industrialization, agriculture remains an important sector of the Indian economy, both in terms of contribution to gross domestic product (GDP) and jobs for millions of people across the country. One of the key factor for productive agriculture is soil. The purpose of the work is to predict the type of terrain using data mining classification methods. Agricultural properties and soil ownership play a crucial role in agricultural decision-making. This research sought to evaluate various mining association techniques and apply them to a soil database to determine if significant relationships could be created. Performance prediction is one of the applications that uses the concept of data mining to increase crop productivity. This makes the problem of crop productive performance is an interesting challenge. An earlier performance prediction was made taking into account the cultivator's experience with a particular crop and culture. This work introduces a system that uses data mining techniques to predict the category of analyzed soil datasets.**

Keywords: Yield Prediction, Data Mining, Soil Analysis, Machine Learning.

I. INTRODUCTION

Horticulture depends to a large extent on the quality of the soil, but in the long run, the increase in agricultural production will result in loss of soil. It is necessary to recognize the methods that repress this elimination of supplements and, moreover, restore the necessary supplements in the soil, thus continuing to receive high quality and large-scale plant productions [1]. In horticulture, soil welfare means that soil can force physical, composite and organic practices for reliable profitability with high yields. The great nature of the soil guarantees us the maintenance and the arrival of water and supplements, the improvement and the continuous development of the roots, while maintaining the biotic state that gives normal results and combats rotting [2].

Soil is very important for plant life. It consists of solids (minerals and organic matter), liquids (water and solutes) and gases (mainly oxygen and carbon dioxide) and contains living organisms. All these elements provide their physical and chemical properties. To maintain fertility, achieve better yield and protect the environment, it is necessary to nurture the soil properly. On the other hand, soil tests are essential to manage it properly. A soil test is the study of a soil sample to discover an additional substance, its composition and various attributes. As a general rule, soil tests are performed to determine the wealth and indicate the gaps to be corrected [3]. The analysis of soil nutrients is very useful for the farmer in determining the type of yield to be grown in a particular soil condition.

Soil fertility, which refers to the intrinsic capacity of the soil to provide essential nutrients to plants in sufficient and adequate proportions for optimal growth, is one of the key elements for determining soil productivity. The management of Indian soil fertility requires sustainable high-level production to produce adequate food for the growing population. Good soil fertility management requires careful identification of the limits of current nutritional deficiencies and monitoring of changes in soil fertility to predict their shortage. These gaps must be mitigated by sound and best practices in terms of nutrients, water, plants and energy for soil management, in order to maintain food production at a reasonable level to ensure high productivity at the same time. future. Therefore, managing soil fertility at optimal levels is one of the key factors for achieving high and sustainable productivity.

Soil fertility is one of the most important factors in crop control. The macronutrients (N, P, K) and the micronutrients (Zn, Cu, Fe and Mn) are important soil components that control their fertility. The characterization of soil in relation to the assessment of the fertility status of an area or region is important in the context of sustainable agricultural production. Due to the unbalanced and inadequate use of fertilizers and the low efficiency of other inputs, the efficiency of the reaction (production) of chemical fertilizers has significantly decreased in recent years with intensive cultivation. Fluctuations in nutrient intake are a natural phenomenon and some of them may be sufficient while others are inadequate.

In this work, data mining classification methods are used to study soil nutrients. Data mining involves extracting information from a data set and transforming it into a structure that is understandable for future use. Various data collection methods are available for the field research sector. Classification is one of the data mining techniques that

automatically create a model of classes from a set of records that contains class labels. Popular classification techniques include decision trees, neural networks, k-nearest neighbour, SVM, and Naïve Bayesian classifier etc.

II. DATA MINING APPROACH

Neural Network

Neural networks are a category of nonlinear flexible models that may adaptively recognize models from data. In theory, it's been shown that neural networks will learn from experience in an acceptable range of nonlinear processing units and might estimate any complicated useful relationship with great accuracy. empirically, several made applications have established their role in pattern recognition and prediction. though many types of neural network models are planned, the foremost common model for time series prediction is that the direct acting network model. Figure 1 shows a typical three-layer feedforward model used for prediction purposes. The input nodes are the previous delayed observations, whereas the output provides the forecast for the future value.

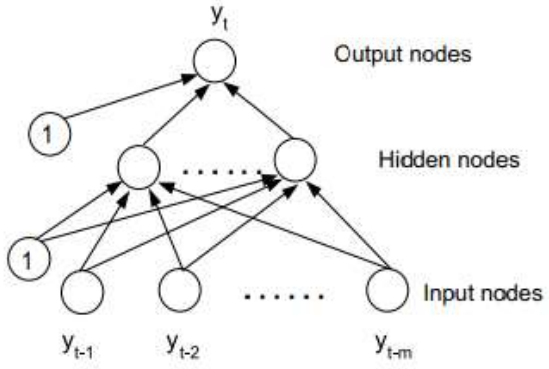

Figure 1: Neural Network

Hidden nodes with appropriate non-linear transfer functions are used to process the information received from the input nodes. The model can be written as:

$$y_t = \alpha_0 + (\sum_{i=1}^{m} \beta_{ij} y_{t-i} + \beta_{0j}) + \varepsilon_t \quad (1)$$

where m is the number of input nodes, n is the number of hidden nodes, f is a sigmoid transfer function, such as logistics:

$$f(x) = \frac{1}{1+\exp(-x)} * \{\alpha_j, j = 0,1,2,\dots\dots,n\} \quad (2)$$

is a vector of weights from the hidden to output nodes and $\{\beta_{ij}$, i=0,1, ... , m; j =1,2,….,n} are the weights of the input to the hidden nodes α0 and β0j the weights of the arcs resulting from the polarization terms whose values are always equal to 1. It should be noted that the linear transfer function it is used in the root node as desired for forecasting problems.

Extreme Learning Machine

The single-hidden layer-feed-forward neural network – also termed as ELM – an learn exactly N different observations on almost all non-linear activation functions with at most N hidden nodes. The main difference between ELM and the traditional formation of an electric network is that the hidden ELM layer does not need a setting that randomly selects hidden level parameters. Input weights, hidden neural bias, and hidden layer output weights are randomly assigned to minimize drive failure. ELM transforms the learning problem into a simple linear system in which the initial weights can be determined analytically. For N arbitrary distinct instances $\{(x_i, y_i), i = 1, 2, \dots, N\}$, where $x_i$and y_iELM with n inputs, m outputs, k hidden neurons, and an activation function g(x) is modelled as:

$$\sum_{i=1}^{n} \beta_i \, g(w_i^T + b_i) = o_i, i = 1,2,\dots\dots\dots N \quad (3)$$

Where w_i and β_i represent the weight vectors connecting the input neurons to an ith hidden neurons to the output neurons, respectively, and bi is a threshold of the ith hidden neurons. The ELM with k = N hidden neurons can reliably approximate these N instances with zero error as

$$\sum_{i=1}^{N} ||o_i - y_i|| = 0 \quad (4)$$

$$\sum_{i=1}^{k} \beta_i g(w_i^T x_i + \beta_i) = y_i \, , i = 1,2,3,\dots\dots\dots N \quad (5)$$

The matrix y is the ELM hidden layer output matrix, in which the i-th column of y is the output of the hidden neuron with respect to the inputs x1, x2. ,,, xN. In the basic ELM, if k << N and Y are a non-square matrix, learning the ELM is equivalent to finding a solution of the least squares β of the linear system Yβ = T.

Decision Tree

The decision model is based on the actual values of the attributes in the data. The decision interval continues until a prediction decision is made for a given record. It has a default destination variable. Decision trees are trained in the data for classification and regression problems. Decision trees are popular in machine learning because they are often quick and precise. It works for categorical and continuous input and output variables. In this technique, the population or sample is divided into two or more homogeneous subpopulations or more based on the most significant fragment in the input variables.

The decision is taken from the tree in the strategic division. This greatly affects the accuracy of the tree. This decision criterion differs for classification and regression trees in Figure 3.4. Decision trees use different algorithms to decide to divide a node into two or more subnodes. The trees break the nodes for all the available variables, then select the division that leads to the most homogeneous sub-nodes. The most popular decision tree algorithms are: Random Forest and least square boosting(LSBoost).

Support Vector Machine (SVM)

This is for classification and regression problems. SVM classifies data into different classes by identifying a hyperplan (line) that separates learning data into classes. The hyperplane's identification, which maximizes the distance between classes, increases the probability of generalizing secret data. SVM offers the best classification performance i.e. the accuracy of the training set. It does not overflow the data.

SVM does not make strong assumptions about the data. Show more efficiency in the correct classification of future data. SVM is classified into two categories, i.e. Linear and non-linear. In a linear approach, training data is represented by a line, i.e. hyperplane, shown separately.

Consider the problem of separating the set of training vectors belonging to two distinct classes,

G = {(x_i; y_i); i = 1; 2;:::;N} with a

hyperplane $w^T*(x) + b = 0$ (xi is the ith input vector, yi ∈ {−1; 1} is known binary target), the original SVM classi1er satisfies the following conditions:

$$w^T*_\phi(x_i) + b \geq 1 \text{ if } y_i = 1$$
$$w^T*_\phi(x_i) + b \leq 1 \text{ if } y_i = -1 \quad (6)$$

where ϕ: $R^n \rightarrow R^m$ is the feature map mapping the input space to a usually high dimensional feature space where the data points become linearly separable.

The distance of a point xi from the hyperplane is

$$d(x_i, w, b) = \frac{|w^T * \emptyset(x_i) + b|}{|w^2|} \quad (7)$$

The margin is 2/|w| according to its definition. Hence, we can find the hyperplane that optimally separates the data by solving the optimization problem:

$$min\emptyset(w) = \frac{1}{2}|w|^2 \quad (8)$$

For the inseparable linear problem, we first assign the data to another large space *H* using a non-linear mapping, which we call Φ. So we use the linear model to achieve classification in new space *H*. Through defined "kernel function" *k*, is converted as follows:

$$max \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j k(\vec{x_i} * \vec{x_j}) \quad (9)$$

$$s.t. \sum_{i=1}^{l} a_i y_i = 0 \qquad 0 \leq a_i \leq C, \quad (10)$$

i=1,2,……l

And corresponding classification decision function is converted as follows:

$$f(x) = sign\,[\sum_{i=1}^{l} a_i y_i k(\vec{x_i} * \vec{x}) + b] \quad (11)$$

The selection of kernel function aims to take the place of inner product of basic function. The kernel function investigates the non-separable problems as follows:

$$k(x_i x_j) = \exp(-\gamma||x_i - x_j||) \quad (12)$$

k-Nearest Neighbor (kNN)

kNN is used for both classification and regression problems, in Figure 1.5. It is one of the simplest classification algorithms. Determine the parameter k which is number of nearest neighbors. When there is new data point to classify, then its k nearest neighbors is find out from the training data. The distance is calculated using one of the measure from Euclidean distance, Minkowski distance, Mahalanobis distance. The larger is k, the better is classification.

J48 (C4.5)

J48 is one of the classification algorithms and has been slightly modified by C4.5 in Weka. You can choose the test as the best information gain. This algorithm was proposed by Ross Quinlan. C4.5 is also called statistical classifier. J48 predicts the dependent variables of the available data. Constructs a tree based on the values of the attributes of the training data. This classifies the data using the functionality of the data instances that are said to have a gain of information. The importance of fault tolerance is developed using a pruning concept.

III. RELATED WORK

In the field of agriculture, several jobs have been done using data mining techniques. According to the research article [7], the researcher accepts the challenge of transforming raw data into useful information. A large amount of data is collected and stored for analysis with sensors and a GPS. This information is used with the neural network for the prediction of grain yield.

Dr. Ramesh [11] has proposed data models that provide great accuracy and generality in terms of performance prediction capabilities. The work to predict yield is done by analyzing the annual precipitation, using KNN, ANN and SVM for this purpose. Data mining techniques are also used in the field of land.

Shweta Taneja [8] worked to identify useful relationships between different soil types. Here, the grouping technique with the WEKA tool has been implemented to create clusters of soil based on their salinity.

P.Bhargavi [5] proposed data mining techniques, when applied to an agricultural soil profile, may improve the verification of valid soil profile classification. The researcher used Naive Bayes classification technique for the classification of the soil. Many researchers have suggested various yield prediction and classification techniques for crops or soil, but a little prominence is given to do it by analysing nutrients and micronutrients content in soil.

To the best of our knowledge Gideon O Adeoye [12] to his knowledge, focused on the physical factors of the soil, the nutrient content available in the soil and the content of the ear corn leaves. Regression equations were obtained for each of the soil factors and the plant to predict yield, which predicts yield with the levels of each factor in all soils, all other constant factors.

The nutrient content of the leaves on yield was also studied by D. Almaliotis [13]. Referring to the work discussed above, this manuscript presents the analysis and classification of soils in terms of nutrients and micronutrients, and to predict crops.

Moni Paul et al. [14] describes a system that uses data mining techniques to predict the category of soil data records

analyzed. The category of crop yields expected show. The crop yield prediction problem is formalized as a classification rule, in which the nearest Bayes and K-Nearest Neighbour methods are used.

IV. PROPOSED METHODOLOGY

This section demonstrates the existing system framework/architecture for soil property assessment based on machine learning techniques as illustrated in figure 2. Flowchart of proposed algorithm:

1. Preprocessing of data is carried out.
2. Then, set features are extracted from data related to the properties of soil such as pH value, calcium content, Nitrogen content, Potassium Content, Phosphorous content, etc.
3. Then on the basis of extracted features are classified in order to analyse Soil Behavior and to predict the crop yielding.

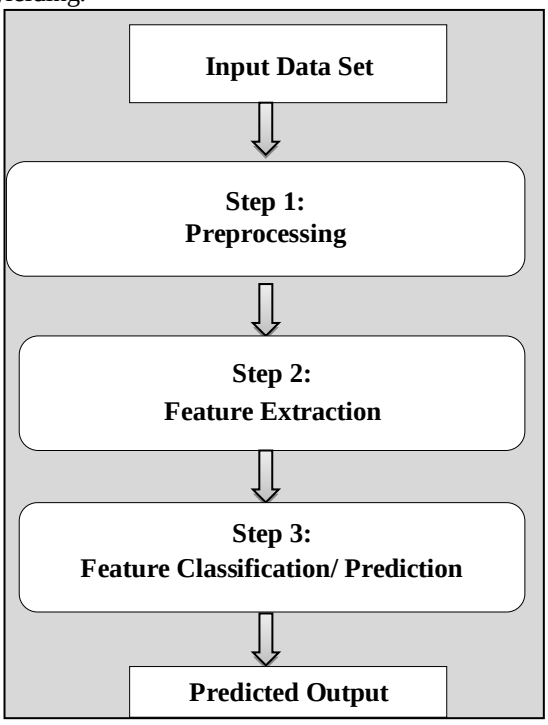

Figure 2: Proposed Model

V. CONCLUSION

In this paper, classification of soil into different categories using soil properties are done by adopting data mining techniques in order to predict the crop yield using available dataset. This study can help the soil analysts and farmers to decide sowing in which land may result in better crop production. The future work may aim to create more efficient models using data mining classification techniques such as support vector machine, neural network, naïve bayes, k-NN, etc. This System will recommend appropriate fertilizer for the given soil sample and cropping pattern.

REFERENCES

[1] Mucherino, P. Papajorgji, P.M. Pardalos, "Data Mining in Agriculture", Springer, 2009.

[2] Mucherino, Petraq Papajorgji, P. M. Pardalos ,"A survey of data mining techniques applied to agriculture", 25 May 2009 Springer-Verlag 2009.

[3] Sally Jo Cunningham and Geoffrey Holmes, "Developing innovative applications in agriculture using data mining", Department of Computer Science, University of Waikato Hamilton, New Zealand.

[4] Cover TM, Hart PE,"K Nearest Neighbor pattern classification", IEEE Trans Info Theory 13(1) : 21-27, 1967.

[5] P.Bhargavi, Dr.S.Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", IJCSNS International Journal of Computer Science and Network Security, Vol.9 No.8, August 2009.

[6] Vishnu Kumar Goyal, "A Comparative Study of Classification Methods in Data Mining using RapidMiner Studio", (IJIRSE) International Journal of Innovative Research in Science & Engineering.

[7] Georg Rub, Rudolf Kruse, Martin Schneider and Peter Wagner, "Data Mining with Neural Networks for Wheat Yield Prediction".

[8] Shweta Taneja, Rashmi Arora, Savneet Kaur, "Mining of Soil Data Using Unsupervised Learning Technique", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 7 No.11, 2012.

[9] M.C.S.Geetha," Implementation of Association Rule Mining for different soil types in Agriculture", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 4, April 2015.

[10] M.Soundarya, R.Balakrishnan," Survey on Classification Techniques in Data mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.

[11] D Ramesh, B Vishnu Vardhan, "Data mining techniqued and applications to agriculture yield data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.

[12] Gideon O Adeoye, Akinola A Agboola, "Critical levels for soil pH, available P, K, Zn and Mn and maize ear-leaf content of P, Cu and Mn in sedimentary soils of SouthWestern Nigeria", Nutrient Cycling in Agroecosystems, Volume 6, Issue 1, pp 65-71, February 1985.

[13] D. Almaliotis, D. Velemis, S. Bladenopoulou, N. Karapetsas, "Appricot yield in relation to leaf nutrient levels in Northern Greece", ISHS Acta Horticulturae 701: XII International Symposium on Apricot Culture and Decline.

[14] Monali Paul, Santosh K. Vishwakarma, Ashok Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach", International Conference on Computational Intelligence and Communication Networks, 2015.