INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

Application of machine learning techniques in estimation of crop yield

Keerti Mulgund
keerticm76@gmail.com
Smt Kamala and Sri Venkappa M. Agadi College of Engineering and Technology, Lakshmeshwar, Karnataka

Rupa S. G.
rupasg1998@gmail.com
Smt Kamala and Sri Venkappa M. Agadi College of Engineering and Technology, Lakshmeshwar, Karnataka

ABSTRACT

It is a process in which we can know what happened in the past. And we know that past is the best predictor of the future. In this research paper, we apply descriptive analytics in the agriculture production domain for sugarcane crop to find efficient crop yield estimation. In this paper, we have three datasets like Soil dataset, Rainfall dataset, and Yield dataset. And we make a combined dataset and on this combined dataset we apply several supervised techniques to find the actual estimated cost and the accuracy of several techniques. In this paper, three supervised techniques are used like as K-Nearest Neighbor, Support Vector Machine, and Least Squared Support Vector Machine. It is a comparative study which tells the accuracy of training proposed model and error rate. The accuracy of the training model should be higher and the error rate should be minimum. And the proposed model is able to give the actual cost of estimated crop yield and it is a label like as LOW, MID, and HIGH.

Keywords— *Crop yield estimation, Support Vectors, Least square Support Vector Machine, Data analytics, Agriculture analytics*

1. INTRODUCTION

Agriculture is one of the important industrial sectors in India and the country's economy is highly dependent on it for rural sustainability. Due to some factors like climate changes, unpredicted rainfall, a decrease of water level, use of pesticides excessively etc. The level of agriculture in India is decreased. To know the level of production we performed descriptive analytics on the agriculture data. The main objective of this research work is to provide a methodology so that it can perform descriptive analytics on crop yield production in an effective manner. Although some studies revealed statistical information about agriculture in India, few studies have investigated crop prediction based on the historical climatic and production data. ANNs accept been acclimated for assorted purposes including classification, clustering, agent quantization, arrangement association, action approximation, forecasting, ascendancy applications and optimization. Using ANN predictions accept been acclimated for the banking industry and altitude prediction. In this work, an ANN is used to predict crop yields based on the data provided by the Telangana State in India. During the review of the several research papers. We found that there are several models exist like as- Principal component regression, Partial least squares, Adaptive forecasts, ARIMA model etc. But the similarity between these models that either they are based on regression or classification. Now we are developing a system which is supervised based model. And it will work as a mixed approach it means classification technique as well as regression technique.

The paper [1] proposes various classification methods to classify the liver disease data set. The paper emphasizes the need for accuracy because it depends on the dataset and the learning algorithm. Classification algorithms such as Naïve Bayes, ANN, ZeroR and VFI were used to classify these diseases and compare the effectiveness, correction rate among them. The performance of the models were compared with accuracy and computational time. It was concluded that all the classifiers except naive Bayes showed improved predictive performance. Multilayer Perceptron shows the highest accuracy among the proposed algorithms.

The paper [2] tries to solve the problem of food insecurity in Egypt. It proposes a framework which would predict the production, and import for that particular year. It uses Artificial Neural Networks along with Multi-layer perceptron in WEKA to build the prediction. At the end of the process, we would be able to visualize the amount of production import, need and availability. Therefore it would help to make decisions on whether food has to be further imported or not.

The soil datasets in the paper [3] are analyzed and a category is predicted. From the predicted soil category the crop yield is identified as a Classification rule. Naïve Bayes and KNN algorithms are used for crop yield prediction. The future work stated is to create efficient models using various classification techniques such as support vector machine, principal component analysis.

2. LITERATURE SURVEY

The paper [4] shows the importance of crop selection and the factors deciding the crop selection like production rate, market price and government policies are discussed. This paper proposes a Crop Selection Method (CSM) which solves the crop selection problem and improves the net yield rate of the crop. It suggests a series of the crop to be selected over a season considering factors like weather, soil type, water density, crop type. The predicted value of influential parameters determines the accuracy of CSM. Hence there is a need to include a prediction method with improved accuracy and performance.

The soil datasets in the paper [3] are analyzed and a category is predicted. From the predicted soil category the crop yield is identified as a Classification rule. Naïve Bayes and KNN algorithms are used for crop yield prediction. The future work stated is to create efficient models using various classification techniques such as support vector machine, principal component analysis.

The paper [5] proposes various classification methods to classify the liver disease data set. The paper emphasizes the need for accuracy because it depends on the dataset and the learning algorithm. Classification algorithms such as Naïve Bayes, ANN, ZeroR and VFI were used to classify these diseases and compare the effectiveness, correction rate among them. The performance of the models were compared with accuracy and computational time. It was concluded that all the classifiers except Naive Bayes showed improved predictive performance. Multilayer perceptron shows the highest accuracy among the proposed algorithms.

3. METHODOLOGY

3.1 Dataset collection

The dataset containing the soil specific attributes which are collected from Polytest Laboratories soil testing lab, Pune, Maharashtra, India. In addition, similar sources of general crop data were also used from Marathwada University. The crops considered in our model include groundnut, pulses, cotton, vegetables, banana, paddy, sorghum, sugarcane, coriander. The number of examples of each crop available in the training dataset is shown. The attributes considered where Depth, Texture, Ph, Soil Color, Permeability, Drainage, Water holding and Erosion.

The above-stated parameters of soil play a major role in the crop's ability to remove water and nutrients from the soil. For crop growth to their possible, the soil must provide an acceptable environment for it. The soil is the anchor of the roots. The water holding capacity determines the crop's ability to absorb nutrients and other nutrients that are changed into ions, which is the form that the plant can use. Texture determines how porous the soil is and the comfort of air and water movement which is essential to prevent the plants from becoming waterlogged. The level of acidity or alkalinity (Ph) is a master variable which affects the availability of soil nutrients. The activity of microorganisms present in the soil and also the level of exchangeable aluminium can be affected by PH. The water holding and drainage determine the infiltration of roots. Hence for the following reasons, the above-stated parameters are considered for choosing a crop.

3.2 Crop prediction techniques

MATIS et al. (1985, 1989) proposed an alternative approach to forecast corn and cotton yield that used Markov Chain theory. This method overcomes some of the drawbacks of the regression model. This method, being completely model free, does not require any assumptions about independent and dependent variables. [6].

The technique named CSM Is used to select the sequence of crops over the season. This method may improve the net yield of the crops. This method resolves the selection of crops for the particular season based on the prediction influenced by parameters such as weather, soil type etc. [7].

In SVM model examples are plotted as points in space, in such a way that the examples of the different categories are distinguished by a clear gap that is as wide as possible. And test set examples are also categorized and mapped into same space with the appropriate classification.

Figure 1 shows an overview of the system, which was divided into two parts: collection and prediction. The design principle is to create a framework that allows users to easily configure the system to be site-specific.

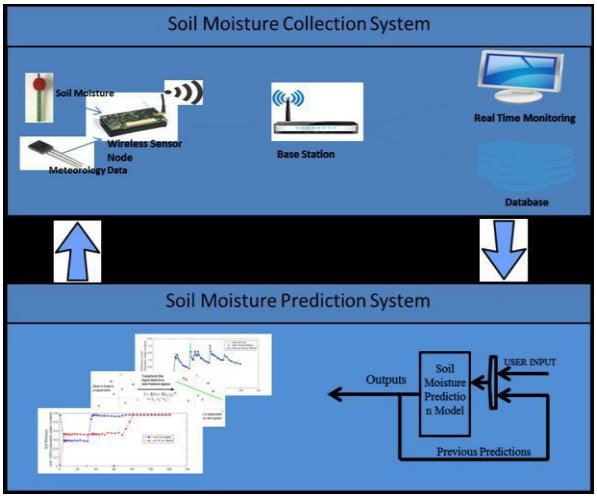

Fig. 1: Dataset collection [18]

In the collection system, a wireless sensor node is prototyped that implemented the proposed framework for sensing soil moisture and other meteorological data. It offered two user-defined variables regulating the level of data granularity and sample intervals. The wireless sensor node can be used for applications such as in-field soil moisture collection and other kinds of remote site data collection since it is specifically designed for applications that require a long lifetime. A prediction system is built on top of the machine learning models to predict soil moisture n days ahead. The models predict the soil moisture value based on meteorological parameters including temperature, humidity, wind speed, solar radiation, precipitation, and soil temperature together with previous days' soil moisture values. The sparse and well-studied machine learning techniques SVM and RVM are applied to the historical data to derive mathematical models. Designed from a Precision Agriculture perspective, the site-specific model is able to incorporate data from other sources at the granularity of one day. The feature of taking user-provided data makes the system more robust by allowing the model to interact with human knowledge or reliable soil moisture data from other sources at a fine granularity. However, the soil and meteorological attributes collected from the hardware devices are the same attributes that are used for deriving prediction models. [9]

3.3 Learners used in the model

3.3.1 Support Vector machine: Support vector machines (SVM) is a set of supervised learning strategies used for

classification, regression and outlier's discovery. It is a classification technique. Here, we have a tendency to plot every information item as some extent in the n-dimensional house (where n is a variety of options you have) with the worth of every feature being the worth of a selected coordinate. It is a classification technique. During this algorithmic rule, we have a tendency to plot every information item as some extent in the n-dimensional house (where n is a variety of options you have) with the worth of every feature being the worth of a selected coordinate. A Support Vector Machine (SVM) is discriminative classifier correctly bounded by a separating hyperplane. In alternative words, given labelled coaching information (supervised learning), the algorithmic rule outputs associate degree best hyperplane that categorizes new examples. Support vector simple machine (SVM) may be a set of supervised learning strategies used for classification, regression and outlier's uncovering.

3.3.2 NAÏVE Bayes: It is not a single algorithm, but a clan of algorithmic rules. All naive Bayes socio-economic classifiers adopt that the value of a particular feature is independent of the value of any other feature, given the class variable. Naive Thomas Bayes classifier could be a straightforward probabilistic classifier that works supported applying theorem (from Bayesian statistics) with robust naive independence assumptions. It is a classification technique supported Bayes' theorem with associate degree assumption of independence between predictors. In straightforward terms, a Naive Thomas Bayes categorified assumes that the presence of a specific feature in an exceedingly class is dissimilar to the presence of the other feature. As an example, a fruit could also be thought of to be associate degree apple if it's red, round, and concerning a pair of inches in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple. These Learners predict the class label for each of the training data set. The class label that is predicted by the majority of the models is voted through the majority voting technique and the class label of the training data set is decided. From the ensemble models, the rules are generated.

3.3.3 Multi-layer Perceptron (Artificial Neural Network): Multi-Layer Perceptron (MLP) could be a feed-forward neural network with several layers between input and end product layer. Feedforward implies that knowledge period in one direction from the input to the output layer. MLPs square measure typically used for pattern classification, recognition, anticipation and approximation. Associate ANN relies on a set of connected units or nodes referred to as artificial neurons (analogous to biological neurons in the associate animal brain). Every association (synapse) between neurons will transmit an indication from one to a different. The receiving (postsynaptic) somatic cell will method the signal(s) then signal neurons connected to that. Artificial neural networks (ANNs) or connectionist systems square measure computing systems galvanized by the biological neural networks that represent animal brains

In Neural Networks some nodes use a nonlinear activation operate that was developed to model the frequency of action potentials, or firing, of biological neurons. For instance, in image recognition, they could learn to spot pictures that contain cats by analyzing example pictures that are manually labelled as "cat" or "no cat" and exploitation the results to spot cats in different pictures. They are doing this with none a priori information.

3.4 Ensemble Model

3.4.1 AdaBoost: Combining with many other learning algorithms, the meta-algorithm is called a AdaBoost algorithm. This would improve the performance of classification. AdaBoost uses the nested operator and it has a subprocess. The sub-processor is used to generate a better model. The ensemble model creates more than one classifier and generates a better model. The accuracy of classification is expanded by creating more than one classifier by the ensemble model. The ensemble model leads to decision making by combining the results of their classification techniques. By this boosting method, the accuracy of the given algorithm is improved. During the year 1995, Yoav Freund and Robert E. Schapire [9] developed the AdaBoost algorithm. In the training period of AdaBoost algorithm, the input set given is (A_1,B_1),(A_2,B_2),….,(A_m,B_m) where Ai denotes space set A and Bi denotes space set B. It is assumed that B=(-1,+1). The base or weak algorithm are called in AdaBoost for the repeated sets Z=1,….Z. In the training set, AdaBoost pertains to weight. At the beginning point, the weights are evenly distributed, while for the other training the weights are increased to indicate that they are not properly classified.

4. PROPOSED METHODOLOGY

In this paper, forecasting of crop production is done by using the time series data set precisely than the existing models. By using AdaBoost technique, ensemble models such as AdaSVM and AdaNaive are developed. To bring weak learners who are slow in learning, AdaBoost technique helps their understanding. SVM, when joined with AdaBoost (AdaSVM), will make superior classification by giving weak learners with appropriate training. A like method is used for Naive Bayes classifier in which AdaBoost based Naive Bayes (AdaNaive) is used to generate superior classified data.

5. PERFORMANCE EVALUATION MEASURES

5.1 Accuracy

Accuracy means very nearness to a measured value or the standard set. Accuracy in time series analysis is the value forecasted which is very near to the actual value. The formula for accuracy is A=(TP+TN)/(TP+FP+FN+TN) where the true positive cases are denoted by TP, true negative cases are denoted by TN, FP and FN are denoted for false positive cases and false negative cases respectively.

5.2 Classification Error

The classification Error (E) of any technique „t" are the cases not correctly classified (FP+FN). The formula for calculating classification Error is E_t=(F/N)*100 where t represents the technique, F denotes the number of items classified incorrectly and N reveals the total number of samples.

6. EXPERIMENTAL RESULTS AND DISCUSSION

SVM, AdaSVM, Naive Bayes, AdaNaive are the classification methods used for time series forecast in this paper. Two groups are separated from the data set for training and for testing the algorithms of classification. In order to implement the classification algorithms, the tool used is Rapidminer data analysis. "Read CSV" operator of rapid miner tool is first loaded for Secondary data retained in CSV file. For the classification process, only a subset of data is selected from the loaded data. To select a subset from original data, "Select Attributes" are utilized by the operator. The chosen subset is then subjected to "X-Validation" operator. It develops the classification model which is validated by the test data. AdaBoost based SVM (AdaSVM), SVM, AdaBoost based

Naive Bayes (AdaNaive) and Naive Bayes are implemented for classification by using "X-Validation" operator. The performance of the classification algorithm is evaluated by using the performance operator. Performance evaluation achieved for both the classification algorithms (existing and proposed) are given in Table 1 and table 2.

Table 1: Accuracy of existing and proposed techniques

Crops	Classification Error			
	SVM	AdaSVM	Naïve Bayes	AdaNaive
Rice paddy	90.48	93.72	86.32	96.52
Cotton	87.60	90.56	84.87	93.45
Sugarcane	88.53	91.64	85.60	96.10
Groundnut	89.32	92.75	85.35	95.45
Black Gram	86.70	89.42	82.4	92.6

Table 2: Classification error of existing and proposed techniques

Crops	Classification Error			
	SVM	AdaSVM	Naïve Bayes	AdaNaive
Rice paddy	9.52	6.28	13.68	3.48
Cotton	12.40	9.44	15.13	6.55
Sugarcane	11.47	8.36	14.40	3.90
Groundnut	10.68	7.25	14.65	4.55
Black Gram	13.30	10.58	17.6	7.40

7. CONCLUSION

Our work would help farmers to increase productivity in agriculture, prevent soil degradation in cultivated land, and reduce chemical use in crop production and efficient use of water resources. Our future work is aimed at an improved data set with a large number of attributes and also implements yield prediction.

The system uses supervised and unsupervised Machine learning algorithms and gives the best result based on accuracy. The results of the two algorithms will be compared and the one giving the best and accurate output will be selected. Thus the system will help reduce the difficulties faced by the farmers and stop them from attempting suicides. It will act as a medium to provide the farmers with efficient information required to get high yield and thus maximize profits which in turn will reduce the suicide rates and lessen his difficulties.

8. REFERENCES

[1] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman (2015) , 'Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh' , (SNPD) IEEE/ACIS International Conference.

[2] Aymen E Khedr, Mona Kadry, Ghada Walid (2015), 'Proposed Framework for Implementing Data Mining Techniques to Enhance Decisions in Agriculture Sector Applied Case on Food Security Information Center Ministry of Agriculture, Egypt', International.

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

[4] Anshal Savla, Parul Dhawan, Himtanaya Bhadada, Nivedita Israni, Alisha Mandholia, Sanya Bhardwaj (2015), 'Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture', Innovations in Information, Embedded and Communication systems (ICIIECS).

[5] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman (2015) , 'Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh' , (SNPD) IEEE/ACIS International Conference.

[6] Biom. J. 34 (1992) 4, 501-511 Akademie Verlag Probability Model for Crop Yield Forecasting, R. C. JAIN and RANJANAAG UWALI.A.S.R.I., New Delhi.

[7] [doi 10.1109_ICSTM.2015.7225403] Kumar, Rakesh; Singh, M.P.; Kumar, Prabhat; Singh, J.P. -- [IEEE 2015 International Conference on Smart Technologies and Â Management Â forÂ Computing, Communication.

[8] [doi 10.1109_smartcomp.2016.7501673] Hong, Zhihao; Kalbarczyk, Z.; Iyer, R. K.

[9] Yoav Freund and Robert E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", Journal of Computer and System Sciences, Vol.55, No.1, 1997, pp.119–139.

