SOIL CLASSIFICATION AND CROP SUGGESTION USING MACHINE LEARNING

Shravani V[1], Uday Kiran S[2], Yashaswini J S[3], and Priyanka D[4]

[1-4]*Undergraduate students, Department of Computer Science, Sapthagiri College of Engineering*
[4]***Mr. Shankar Rana**, Assistant prof., Department of Computer Science, Sapthagiri college of Engineering*

Abstract— *Agriculture is one of the most important components of our society. Soil is a critical factor for a successful agriculture. The composition of soil differs from soil to soil. The Growth of Crops is affected by these chemical features of soil. Choosing the right type of crops for that particular type of soil is also important.*

Machine Learning techniques can be used to classify the soil series data. The results of such classification can further be combined with crop dataset to predict the crops that are suitable for the soil series of a particular region and its climatic conditions.

Soil dataset and crop dataset are used. The datasets comprise of chemical attributes and geographical attributes of soil and crops. Algorithms like SVM and Ensembling technique can be used to classify the soil series data and predict the suitable crops.

Index Terms— **Soil, Crop, Agriculture, crop recommendation, soil classification, machine learning.**

I. INTRODUCTIÓN

A Agriculture is a very essential part of our society. Agriculture is a source of livelihood in most parts of the world. Agricultural produce is of great importance. But in recent years, the agricultural produce is gradually decreasing. Soil plays a crucial role in agriculture. Soil consists of nutrients, that are used by the plants to grow. There are different types of soils available and each having different properties. Crop's productivity is mainly based on the type of soil. The possible way to improve productivity is that we choose a right crop for the right land type. This can be done by first analysing the soil then classifying it into different soil groups. Based on these soil groups and the geographical conditions, one can decide which crop is best suited and is beneficial. The traditional methods are Costly, long process and also time consuming. Hence, there is a need for new technologies and methods to enhance the existing system in order to get faster and better results.

Machine learning is one of the budding technologies in the field of agriculture. Machine Learning can be used to improve the productivity and quality of the crops in the agricultural sector. It can be used to find patterns among the agricultural data and classify it into a more meaningful data. This data can be used for further processes. Machine learning techniques usually follows the following procedure: collecting data, processing the data, training-testing of data samples. The algorithm such as SVM can be used for classification of soil and crop prediction based on previous patterns followed and the type of the soil. The project requires the following datasets: soil dataset with several chemical properties has its features and crop dataset with geographical attributes as its features.

The project aims at creating a model that efficiently classifies the soil instances and map the soil type to the crop data to get better prediction with higher accuracies. Soil prediction involves types of crop classifications and geographical attributes. It also aims at creating a system that processes the real-time soil data to predict the crops with higher accuracy. The model involves two phases: training phase and testing phase. Two datasets are used: Soil dataset and crop dataset. The predicted and actual classes are compared and the list with correct classes is obtained.

II. LITERATURE SURVEY

2.1: Soil Classification using Machine Learning Methods and Crop Suggestion Based on Soil Series

This project creates a model that can predict soil series with land type and according to prediction it can suggest suitable crops. It makes use machine learning algorithms such as weighted K-Nearest Neighbor (KNN), Bagged Trees, and Gaussian kernel-based Support Vector Machines (SVM) to classify the soil series. The Soil classification philosophies used, follows the existence knowledge and practical circumstances. On the land surfaces of earth, classification of soil creates a link between soil samples and various kinds of natural entity. Based on these classifications and the mapped data, the suitable crops were suggested for a particular region.

System overview

The system uses the soil series data obtained by Soil Resources Development Institute (SRDI) of Bangladesh. The group of soils which is formed from the same kind of parent materials and remains under the similar

conditions of drainage, vegetation time and climate forms the soil series. It also has the same patterns of soil horizons with differentiating properties. Each soil series were named based on its locality. The main purpose of this system is to create a suitable model for classifying various kinds of soil series data along with suitable crops suggestion for certain regions of Bangladesh. In this paper, they have worked on 9 soil series datasets obtained from six upazillas of Khulna district, Bangladesh. The dataset considered had 383 samples with 11 classes. The crop database was created by considering the Upazilla codes, map units and class labels. The method involved two phases: training phase and testing phase. Two datasets were used: Soil dataset and crop dataset. Soil dataset contains class labelled chemical features of soil. The system follows an architecture as seen in Fig2.1.1. The machine learning methods were used to find the soil class. Three different methods used were: weighted K-NN, Gaussian Kernel based SVM, and Bagged Tree.

Weighted KNN: KNN uses all training data, since weighting is used. The nearest neighbours are given more weightage than the farther ones. The distance is calculated and the majority of such classification is taken as the final classification. The obtained accuracy of the classification using KNN was 92.93 %

Gaussian Kernel based SVM: SVM separates the objects of classes into different decision planes. A decision boundary separates the objects of one class from the object of another class. Support vectors are the data points which are nearest to the hyper-plane. Kernel function separates the non linear data by transforming the inputs to higher dimensional space. In this project they have used gaussian kernel function. The accuracy obtained using SVM was 94.95%

Bagged Trees: Bagging generates a set of models that classifies the random samples of data and predicts the classes. When given a new instance, the predictions of these models are aggregated to find the final prediction of class. The accuracy obtained using this algorithm was 90.91%.

After classifying the soil series, the crops, that are suitable for that series for the given map unit of corresponding upazilla were suggested. The results inferred that K-NN and Bagged tree showed comparative accuracy, but SVM outperforms the other two algorithms and produces an accuracy of 94.95%. The system was able to achieve an average accuracy of 92.93%.

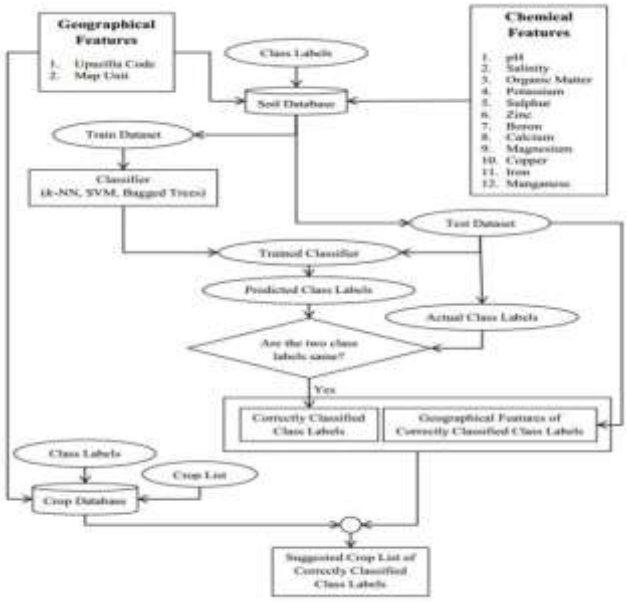

Fig2.1.1: The system architecture of Soil Classification and Crop Suggestion system.

2.2: Big Data Analytics for Crop Prediction Mode Using Optimization Technique

Big data analysis is used in discover novel solutions which means analyzing bulky data set ,so which place a decision making in agricultural field. In this project they are predicting two classes that is good yield and bad yield based on soil and environmental features like average temperature, average humidity, total rain fall and production yield hybrid classifier model is used in optimizing the features and it is also divided into three phases that is viz pre-processing, feature selection and SVM_GWO that is grey wolf optimizer along with Support Vector Machine(SVM) classification is used to improve the accuracy, precision, recall and F-measure. In this the result shows that SVM_GWO approach is better as compared to typical SVMs classification algorithm.

The proposed methodology of GWO (Grey Wolf Optimization over SVM (Support Vector Machine) classifier. The proposed approach is divided into three phase.

Data Pre-Processing: Data pre-processing is deemed to be the main step in the data mining method and machine learning projects. In effective data collection can subsequently lead to improper combination, weak control and incorrect values . It moreover provides misleading results if the identification of the data is not carried out clearly. Therefore, it is essentially important to carry over quality and representation of data at the initial stage **Feature Selection:** The algorithm of feature selections

used to select a subset of the features that are considered be important so that the redundant data can be removed. In addition, it also eliminates the irrelevant and noisy features of the data in order to make more simple and accurate. It also later helps in saving memory and storage. Feature selection algorithms are categories viz in two categories subset selection methods and feature ranking methods. The selection methods of Subset essentially find the features that are required for finest subset.

Grew Wolf Optimization Based Feature Selection: GWO select optimal feature subset for classification, in this approach they used optimization for given relative weight to features according to its relative information and reducing training error. So grey wolf optimization provides effective relation between features and get effective information from features by this information reduce the support vectors in SVM which reduce the training error of classifier. So it's direct impact on testing error of classification and increases the accuracy, precision and recall.

Machine Learning Classifier: A feature selection algorithm selects a subset of vital features and removes superfluous, unrelated and raucous features for simpler and more precise data illustration.

Proposed Algorithms: Proposed Algorithm of SVM_GWO . The objective of this study is to increase the accuracy of prediction model by using different parameters for future precision agriculture. The data set has been taken from Food agriculture organization and applied to the processing model through map reduce. To process the big data, map reduce is used to minimize the time of execution and further classification is done. Feature selection and extraction are important steps in classification systems.

This paper presents a hybrid model i.e. SVM_GWO that uses a combinational approach for improving the classification accuracy, recall, precision, f-measure by selecting the optimal parameters settings in SVM. In this classification we have extract the feature vector with minimum error and converge and then SVM_GWO is developed for selecting the optimal SVM parameters. Result show that the proposed approach is better than the typical SVM classification algorithm with classification accuracy 77.09%, precision 75.38%, recall 74.189% and f measure 73.15%.

2.3: Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique

The ensembling technique is used to build a model that combines the predictions of multiple machine learning models together to recommend the right crop based on the soil specific type and characteristics with high accuracy. The independent base learners used in the ensemble model are Random Forest, Naive Bayes, and Linear SVM. Each classifier provides its own set of class labels with an acceptable accuracy. The class labels of individual base learners are combined using the majority voting technique. The crop recommendation system classifies the input soil dataset into the recommendable crop type, Kharif and Rabi. The dataset comprises of the soil specific physical and chemical characteristics in addition to the climatic conditions such as average rainfall and the surface temperature samples. The average classification accuracy obtained by combining the independent base learners is 99.91%.

A brief step by step procedure of designing the crop recommendation system is explained as follows:

Step 1: Input

The input dataset is a comma separated values file containing the soil dataset, which has to be subjected to pre-processing.

Step 2: Pre-processing of input data

Input dataset is subject to various pre-processing techniques such as filling of missing values, encoding of categorical data and scaling of values in the appropriate range

Step 3: Splitting into training and testing dataset

The pre-processed dataset is then split into training and testing dataset based on the specified split ratio. The split ratio considered in the proposed work is 75:25, which means 75% of the dataset is used for the training the ensemble model and the rest 25% is used as test dataset.

Step 4: Building individual classifiers on the training dataset The training dataset is fed to each of the independent base learners and the individual classifiers are built using the training dataset.

Step 5: Testing the data on each of the classifiers The testing dataset is applied on each of the classifiers, and the individual class labels are obtained.

Step 6: Ensembling the individual classifier output using Majority Voting Technique.

Proposed Algorithm: Ensemble Framework, Random Forest, Naive Bayes, Linear SVM, Majority Voting.

The dataset considered for usage in the given proposed work is a soil dataset primarily comprising of soil physical and chemical properties, along with the climatic details. An open source dataset is obtained from the data repository site of the Government of India, data.gov.in

The dataset size is 5MB containing 9000 rows and 15 attributes that are of prime importance. The crops considered are Cotton, Sugarcane, Rice, Wheat. The dataset attributes that are of prime importance are, Soil Type, pH value of the soil, NPK content of the soil, Porosity of the soil, Average rainfall, Surface temperature, Sowing season .

A crop recommendation system has been designed that takes into consideration the soil dataset with respect to the four crops Rice, Cotton, Sugarcane, Wheat. The soil dataset is first preprocessed and then the ensembling technique performs a critical function in the classification of the four crops. The individual base learners used in the ensemble model are Random Forest, Naive Bayes, and Linear SVM. Majority Voting Technique has been used as the combination method to provide the best accuracy.

The accuracy obtained using the ensembling technique is 99.91%. Hence, the proposed work provides a helping hand to the farmer in the accurate selection of the crop for cultivation. This creates an exponential gain in the crop productivity which in turn boosts the economy of the country.

2.4: Smart Farming Prediction Using Machine Learning

Agriculture is one of the major game changer and a major revenue producing sector in India. Different seasons, market and Biological Patterns influence the crop production ,but because of changes in these patterns result in an excellent loss to farmers .This factors can be minimized by using a suitable approach related to the knowledge of soil types ,pressure ,suitable weather, crop type. whereas, weather and crop types and be predicated using useful dataset that can aid to farmers by predicting the maximized profitable crops to grow. These paper mainly focus on the algorithms used to predict crop yield ,crop cost prediction. With the help of all these features smart farming can be achieved.

The implementation includes the datasets taken from the koggle.com to feed the system with 3000 generic data of agriculture features .These includes temp ,soil quality,etc. To use the predictive system the machine learning algorithms requires two types of data - Trained data, Test data. Trained data is the survey data collected in the period of 12 months, whereas test data is the current survey data. Both these data will be merge together also known as classification techniques (Random forest algorithm will be used).

Research Work: Research work is the first step to gather the data in machine level. For these they have taken only the train datasets and apply the pre –processing on it. It classifies the data into test part and train part. All he features of the required data such as soil type, temp ,humidity etc. is extracted.

Feature Extraction: Feature extraction is required as there is large number of the data to be processed .Feature extraction will take only the necessary data from the test part and train part (Around 25 features).These features will help the farming to boost at all levels .

Classification Technique: Classification technique is the most important part of the process as the implementation of the algorithm occurs here .Random forest algorithm is implemented in the process to give the through results of the datasets. The algorithm takes 20% of the test data (Random data) as the size given to the system and remaining 80% train data is take. After applying the classification techniques we get two results, Algorithm result (accuracy of the datasets) ,Dataset results that will be in the form of a matrix such as truepostive ,true negative etc.). The predicted data can be judge from the matrix itself .In real time the values of the matrix is used to make a prediction of the land to grow the desirable crop in a given features of the month.

They have introduced a Davis Pro2 is an hardware system that will enabled the systems to take periodic data of the fields using its sensors and send the collected data to the cloud storage. Cloud storage will contained all the data of the fields and monitor the changes in the data .All these data will be converted into datasets as required. The Datasets will contained valued of the fields like min temp, max temp, soil types etc Firstly the most important module which we are using is the dataset which is the main component in the machine learning and to find the result .In this we have approximately 3000 data from which we have to find the result. Then after that the clustering came in the story it help us to differentiate between the dataset and group the different data in their respective columns. Then after that the Bayesian network is used to form the Statistical analysis of the attribute in a given dataset. Then after the ANN is used to compares patterns nonlinear effect and underline concept of the relation between them and hence it is a kind of ML technique which has a vast memory. Finally, smart

systems that provide real-time suggestions and make long-term forecasts based on user choices and preferences must be studied and tested.

2.5:Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction

Agricultural research has been profited by technical advances such as automation, data mining. Today ,data mining is used in a vast areas and many off-the-shelf data mining system products and domain specific data mining application soft wares are available, but data mining in agricultural soil datasets is a relatively a young research field. The large amounts of data that are nowadays virtually harvested along with the crops have to be analysed and should be used to their full extent. This research aims at analysis of soil dataset using data mining techniques. It focuses on classification of soil using various algorithms available. Another important purpose is to predict untested attributes using regression technique, and implementation of automated soil sample classification.

In this approach, they have developed an automated system for soil classification based on fertility . After obtaining the fertility class labels with the help of automated system, we carried out a comparative study of various classification techniques with the help of data mining tool known as WEKA. The dataset used, was collected from one of the soil testing laboratories in Pune District (Maharashtra, India).Rest of this paper focuses on the prediction of untested attributes. This research has implemented a very sound practical application of linear regression technique by forecasting an obscure property of the soil test. The outcome of this research will result into substantial diminution in the price of these tests, which will save a lot of efforts and time of Indian soil testing laboratories.

Research Methodology

Dataset Collection: The dataset is part of surveys which are carried out regularly in Pune District. this dataset was collected from a private soil testing lab in Pune. It contains information about number of soil samples.

Automated System: Soil classification system is essential for the identification of soil properties. Expert system can be a very powerful tool in identifying soils quickly and accurately .Traditional classification systems include use of tables, flow-charts. This type of manual approach takes a lot of time, hence quick, reliable automated system for soil classification is needed to make better utilization of technician's time. We propose an automated system that has been developed for classifying soils based on fertility.

A Comparative Study of Soil Classification:

The classification of soil was considered critical to study because depending upon the fertility class of the soil the domain knowledge experts determines which crops should be taken on that particular soil and which fertilizers should be used for the same. The following section describes

Proposed Algorithm: Naive Bayes, J48, JRip algorithms briefly.

Naive Bayes: A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters.

J48 : J48 is an open source Java implementation algorithm in the Weka data mining tool. This algorithm was developed by Ross Quinlan. It is an extension of Quinlan's earlier ID3 algorithm. Continuous attribute value ranges, pruning of decision trees, rule derivation, and so on.

JRip: This algorithm implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP. In this paper, three classification techniques data mining were evaluated and compared on basis of time, accuracy, Error Rate, True Positive Rate and False Positive Rate. Tenfold cross validation was used in the experiment. Our studies showed that J48 model turned out to be the best classifier for soil samples for studying human visual attention over the last few years.

In this paper, They have proposed an analysis of the soil data using different algorithms and prediction technique. In spite the fact that the least median squares regression is known to produce better results than the classical linear regression technique, from the given set of attributes, the most accurately predicted attribute was “P” (Phosphorous content of Algorithm Linear Regression, Least Median Square Regression.

In future, we contrive to build Fertilizer Recommendation System which can be utilized effectively by the Soil Testing Laboratories. This System will recommend appropriate fertilizer for the given soil sample and cropping pattern.

REFERENCES

[1] Gholap, J., Ingole, A., Gohil, J., Gargade, S. and Attar, V., 2012. Soil data analysis using classification techniques and soil attribute prediction. arXiv preprint arXiv:1206.1557.

[2] Sofianita Mutalib, S-N-Fadhlun Jamian, Shuzlina AbdulRahman, Azlinah Mohamed,2010. Soil Classification: An Application of Self Organising Map and k-means [3] Prachi Sharma , Dr. D.V. Padole,2017. Design And Implementation Soil Analyser Using IoT.

[4] S.Pudumalar*, E.Ramanujam*, R.Harine Rajashreeń, C.Kavyań, T.Kiruthikań, J.Nishań, 2016. Crop Recommendation System for Precision Agriculture.

[5] Avinash Kumar, Sobhangi Sarkar and Chittaranjan Pradhan, 2019. Recommendation System for Crop Identification and Pest Control Technique in Agriculture.

[6] Nidhi H Kulkarni , Dr. G N Srinivasan , Dr. B M Sagar, Dr.N K Cauvery,2018. Improving Crop Productivity through A Crop Recommendation System Using Ensembling Technique.

[7] Meiqin Zhang, Shanqin Wang*, Shuo Li, Jing Yi, Peng Fu, 2011. Prediction and Map-making of Soil Organic Matter of Soil Profile Based on Imaging Spectroscopy: A Case in Hubei China.

[8] Parin M. Shah, 2012. Face Detection from Images Using Support Vector Machine.

[9] James O. Hortle, 2018. Alzheimer's Disease and Support Vector Machines: An Introduction to Machine Learning. [10] Heejong Suh, Daehyon Kim* and Changsoo Jang, 2018, Heejong Suh, Daehyon Kim* and Changsoo Jang.

[10] Amos Baranes, Rimona Palas , 2019, Earning Movement Prediction Using Machine Learning-Support Vector Machines (svm).