ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

SPECIFIC CROP FERTILITY ON SOIL DATASET USING DATAMINING TECHNIQUES

R.Shobana& D. Saranya

Department of computer science and application, D.K.M College for women, Vellore, India.

shobanavasu.mca@gmail.com

Abstract

Agriculture is a backbone of Indian economy and hence the aim of this paper to review research on various factors, properties and components of soil, and find which type of soil is most suitable for particular nutrients like Nitrogen, Phosphorus, Potassium, calcium, zinc and focuses on specific crop on various soil and also hybrid of soil components on specific crop. This research aims at analysis of soil dataset using decision tree algorithms in data mining. Different decision tree algorithms are applied to soil dataset to predict its fertility. This paper focuses on classification of soil fertility to predict crop production with the help of J48 algorithm, Random Forest, Simple Cart, NB Tree, tools and techniques.

Keywords

Crop yield, Soil data, Specific crop, various soil datasets, Classification algorithm

I. Introduction

Enhancement of food technology is today's need. India is living in th e era of huge population wherein the ratio and proportion of food and humans has no toning, resulting in high rates of inflation. Agriculture is totally dependent on the soil quality but as time passes more and more agricultural production results in the loss of nutrients present in the soil. We require identifying techniques that will slow down this elimination of nutrients and also will return the required nutrients with the soil, so that we keep getting high quality and good quantity crop productions.

Agriculture is a backbone of Indian economy and hence the aim of this paper to review research on various factors, properties and components of soil, and find which type of soil is most suitable for particular nutrients like Nitrogen, Phosphorus, soil and also hybrid of soil components on specific crop. This research aims at analysis of soil dataset using decision tree algorithms in data mining. Different decision tree algorithms are

applied to soil dataset to predict its fertility. This paper focuses on classification of soil fertility to predict crop production with the help of J48 algorithm, tools and techniques.

Keywords: Crop yield, Soil data, Specific crop, various soil datasets, Classification algorithm

II. Purpose

The work in this thesis highlights an important model. Data Mining has two primary Models: Descriptive Data Mining Model and predictive Data Mining Model. Descriptive mining models describe or summarize the general characteristics or behaviour of the data in the Soil database. Predictive models perform inference on the current data in order to make the prediction. Both of them are fundamentals to understand Soil behaviours. In general , in materials informatics, Data mining can be used in the following task

(i) Association analysis

Association analysis is good at discovering patterns, and can be usedtodevelopheuristicrulesforSoil behaviourbasedonlargedatasets

(ii) Classifier/Predict modeling

Some machine learning algorithms can be used for Soil class prediction and Soil classification models such as Support Vector Regression (SVR) and Neural Network (NN), can be used to build up the Predict models. These

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

models can be used to predict crystal structure or composite properties from hybrid Soildata.

(iii) Cluster analysis

As an exploratory data analysis tool, it can sort different Soil or properties into groups in such a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. And, cluster analysis can be integrated with high-throughput experimentation for rapidly screening combinatorial data

(iv)OutlierAnalysis

Inpropertiesanalysisorcombinatorialexperiments,outlieranalysisis used to identify anomalies, especially to assess the uncertainty and accuracy of results, and distinguish between true discoveries and false-positiveresults.

III. Dataset Collection

Dataset required for this research contains various attributes and the 10 attributes of soil samples with its description are listed

Table 1: Attribute Description

Attribute	Description
Ph	pH value of Soil
EC	Electrical Conductivity
OC	Organic Carbon, %
P	Phosphorous, ppm
K	Potassium, ppm
Fe	Iron, ppm
Zn	Zinc, ppm
Mn	Manganese, ppm
Cu	Copper, ppm
label	Soil fertility class(very low, low, moderate, moderately high, high,

IV. Data MiningTechnique

Classificationandpredictionisoneofthecoretasksof DataMining.Aclassificationtechniqueis a systematic approach to building classification models from training and testing data sets.Several classification models such as Decision Tree Classifier, Rule-Based Classifier, Neural Network Classifier, naive Bayesian Classifier, Neuro-Fuzzy classifier, Support VectorMachinesand etc. are reported in literature. Each technique employs a learning algorithm to identify a model that best fits the relationships between the attribute set and class label of the input data.The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of data set that has never seen before. Therefore, the key objective of the data mining algorithm is to build models with good generalization capacity.

NAI BAYESIANCLASSIFIER

Naive Bayesian classifier is a statistical classifier that can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. It is fast and incremental that can deal with discrete and continuous attributes and has excellent performance in real-life problems. The naive Bayesian classifier, or simple Bayesian classifier generally used for classification or prediction task. As it is simple, robust and generality, this procedure has been deployed for various applications such as Materials damage detection Agricultural land soils classification, Network intrusion detection, Machine learning applications. Therefore, the application of this method is extended to classification of Soil data sets and to reduce the computational cost of classification of Soil for dataselection.Bayesian formula can be written as

$$p(\underline{D1,D2\ldots.Dn})p(ci)$$

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

$$P(Ci/D1,D2,\ldots Dn)=\frac{ci}{p(D1,D2\ldots\ldots Dn)}$$

Random forest

The Random Forests algorithm is able to classify large amounts of data with accuracy/ It helps for classification and regression that construct a number of decision tree at training time and output the class that is the mode of the classes output by individual trees.Random Forests are the combination of tree predictors where each tree depends on the values of the random vector sampled individually with the same distribution for all trees. The basic principle is that group of "weak learners" can come together to form a "strong learner".Random Forests are a wonderful toolformaking predictions do not overfit because of the law of large numbers.

Random Forests grows many classification trees.

1. If the number of cases in the training set is N, sample N cases at random- but with replacement, from the original data. This sample will be the training set for growing the tree.

2. If there are M input variable, a number M is specified that the each node, m variables are selected at random out of M and the best split of m from M is used to split the node. The value of m will be taken as constant.

3. Each tree is grown to be the largest extent possible. There is no pruning.

v. Comparison Of Decision Tree Algorithms For Soil Fertility Prediction

Soil fertility is considered to be one of the critical attributes for deciding specific crop pattern by using hybrid of soil in particular area. In this section, results of various decision tree algorithms on dataset are shown. Based on these, the best classifier is selected and further used for tuning its performance

J48 (C4.5)

J48 is an open source java implementation of the C4.5 algorithm in the weka data mining tool. C4.5 is the program that creates a decision tree based on a set of labeled input data.

The decision trees generated byC4.5 can be used for classification, and for this reason, C4.5 is referred to as a statistical classifier. [7] proposed a secure hash message authentication code. A secure hash message authentication code to avoid certificate revocation list checking is proposed for vehicular ad hoc networks (VANETs). The group signature scheme is widely used in VANETs for secure communication, the existing systems based on group signature scheme provides verification delay in certificate revocation list checking. In order to overcome this delay this paper uses a Hash message authentication code (HMAC). It is used to avoid time consuming CRL checking and it also ensures the integrity of messages. The Hash message authentication code and digital signature algorithm are used to make it more secure . In this scheme the group private keys are distributed by the roadside units (RSUs) and it also manages the vehicles in a localized manner. Finally, cooperative message authentication is used among entities, in which each vehicle only needs to verify a small number of messages, thus greatly alleviating the authentication burden.

NB Tree

This algorithm is used for generating a decision tree with navie Bayes classifiers at the leaves

SimpleCart

It is a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric. It is used to implement minimal cost complexity pruning.

In this paper, Decision tree techniques are used to compare on the basis of accuracy and Error Rate. Tenfold cross validation was used in the experiment.

This paper showed that J48 (C4.5) model turned out to be the best classifier for soil samples. Comparison of suitable and unsuitable with their accuracy are shown using different Classifiers.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

Classifier	NB Tree	SimpleCart	J48
Correctly Classsified Instances	1700	1824	1827
Incorrectly Classified Instances	288	164	161
Accuracy(%)	85.51	91.75	91.90

Table 2: Comparison of different classifiers

Tuning Performance OfJ48 Algorithm

Accuracy of J48 algorithm for predicting soil fertility was highest, hence it was used as a base learner. Now , the aim was to increase its accuracy with the help of some others meta-techniques like attribute selection and boosting with the help of Weka.

With attribute selection

Attribute selection reduces dataset size by removing irrelevant/ redundant attributes. It finds minimum set of attributes such that resulting probability distribution of data classes is as close as possible of original distribution. Attribute evaluator method-Classification subset evaluator was used, Which evaluates the worth subset of attributes by considering the individual predictive ability of each attribute. Following are the results using Attribute Selected Classifier with the learner as J48

Correctly Identified Instances	1935	94.23%
Incorrectly Identified Instances	143	7.8%

Table 3: Using Attributes Selected Classifier with J48 as Base Learner

Using Attribute Selected Classifier with J48 as Base Learner
It can be clearly seen that the accuracy has been increased from 91.90 to 93.20 after application of attribute selection technique.

Combining attribute selection and boosting method

Boosting is a machine learning meta algorithm for performing supervised learning. It can boost performance of

weak learner and convert it into a strong learner. It increases the weights of incorrectly identified instances and decreases the weights of correctly identified instances over its limitations.
Adaboost is Weka implementation of boosting method Which is used for boosting a nominal class classifier. The results after the combination of attribute selection and Adaboost with J48 base learner are listed below.

Correctly Identified Instances	1998	97.83%
Incorrectly Identified Instances	87	4.8%

Table 4: Results after using combination of attribute selection and boosting with J48 as base learner

Here, accuracy was enhanced upto 96.83% which makes this predictive method to be more accurate.

CONCLUSION

Datamining is the new research area in agriculture. As agriculture is a soil based industry, there is no way that required yield increases of the major crops can be attained without ensuring that plants have an adequate and balanced supply of nutrients. In this paper different classifier algorithms are used on soil dataset J48 classifier perform better to predict fertility index.

The large amounts of data that are nowadays virtually harvested along with the crops have to be analyzed and should be used to their full extent. Various decision tree algorithms can be used for prediction of soil fertility. My studies showed that J48 gives 91.90% accuracy, hence it can be used as a base learner. With the help of other meta-algorithms like Attribute selection and boosting, J48 gives accuracy of 96.73% which makes a good predictive model.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

References

[1] Kumar A. &Kannathasan N.(2014), " A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining", IJCSI International Journal of Computer Science Issues, Vol. 8,Issue 3,

[2] " Soil test", Wikipedia, February 2016

[3] "C4.5 (J48)", Wikipedia, February 2016

[4] Cohen W. (2010)", Fast Effective Rule Induction", Twelfth International Conference on Machine Learning, 115-123

[5] Gruhn P., Goletti F., and Edelman M.(2010), " Integrated Nutrient Management, Soil Fertility, and Sustainable Agriculture:

Current Issues and Future Challenges", International Food Policy Research Institute 2033 K Street, N.W. Washington, D.C. U.S.A.; Technical Report

[6] Vamanan R. &Ramar K.(2015) " Classification of Agricultural Land Soils A Data Mining Approach", International Journal Of Computer Science and Engineering (ITCSE); ISSN:0975- 3397 Vol.3

[7] Christo Ananth, M.Danya Priyadharshini, "A Secure Hash Message Authentication Code to avoid Certificate Revocation list Checking in Vehicular Adhoc networks", International Journal of Applied Engineering Research (IJAER), Volume 10, Special Issue 2, 2015,(1250-1254)

[8]" CART", Wikipedia, July 2014

[9] Brieman L., Friedman J., Olshen R. and Stone C. (2012) " Classification and Regression trees"'Wadsworth International Group, Belmont, California.

[10] " Soil test", Wikipedia, July 2014.

[11] Compedium on Soil Health, Department of Agriculture &Cooperation , January 2013

[12] N Srihari Rao " Prospective usage of ICT by Farmers for Agriculture", 2014

[13] George bRub, " Data Mining of Agriculture Yield Data: A Comparison of Regression Models"

[14] Jay Gholap "Performance Tuning of J48 Algorithm for Soil Fertility" 2015. Asian Journal of Computer Science and Information Technology 2: 8 (2014) 251- 252

[15] Margaret Dunham, "Data Mining: Concepts and

[16] "Random Forest", Wikipedia, Aug 2014

[17] Kohavi R.(2011), "Scaling Up the Accuracy of Navie-Bayes Classifiers: A Decision tree Hybrid", Second National Conference on Knowledge Discovery and Data Mining , 202-207

ABOUT THE AUTHORS

Prof.SHOBANA. R is an Assistant Professor in the Department Of Computer Science and Applications at D.K.M. College for Women, Vellore. Her special research interests are in Data Mining, Software Engineering, Data Science, Artificial Intelligence, Artificial Neural Network, Cloud computing.

Prof. SARANYA. D is an Assistant Professor in the Department Of Computer Science and Applications at D.K.M. College for Women, Vellore. Her special research interests are in Data Mining, Software Engineering, Cloud computing, Neural Network.