BEST: International Journal of Humanities, Arts, Medicine and Sciences (BEST: IJHAMS)
ISSN (P): 2348-0521, ISSN (E): 2454-4728
Vol. 5, Issue 08, Aug 2017, 75-80
© BEST Journals

AUTOMATED SOIL TESTING PROCESS USING COMBINED MINING PATTERNS

T. MATHAVI PARVATHI

Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

ABSTRACT

Our proposed work presents an automated online system, which uses data mining techniques called combined mining patterns to predict the category of the analyzed soil datasets and provide suggestions of the crops which can be cultivated for better yield. The basic problem of predicting the crop yield is formalized by the classification rule, where Naive Bayes and K-Nearest Neighbor algorithms are used. The soil testing starts with the collection of a soil sample which is being collected from the agricultural field. The first and foremost principle of the soil testing process is that an agriculture field can be sampled in such a way that chemical analysis of the soil sample will accurately reflect the agricultural field’s true nutrient status.

KEYWORDS: Combined Mining, Soil Data Mining & Pattern

INTRODUCTION

The fertility and expected growth potential of soil can be determined by performing the soil testing. It is important to determine the nutrient, contaminants and other characteristics such as acidity and PH level of soil. In order to improve the effectiveness and accuracy of classification of large soil data sets, we need to solve complex soil data sets using data mining techniques. The soil testing laboratories are provided with advanced literature on various aspects of soil testing, including testing methods and formulations of fertilizer recommendations. Experienced soil testing persons help farmers to decide the extent of fertilizer and farm yard manure to be applied at various stages of the growth cycle of the crop.

This paper improves the effectiveness and accuracy of the Classification of large soil datasets and provides suggestion of the crops which can be cultivated for better yield. In particular, this research work aims to compare the performance of the data mining algorithms such as Naive Bayes and K-Nearest Neighbor algorithms, which is used for classification. Each and every country trying to invest more money in the agricultural research, because hunger is forcing people to cultivate short duration crops to increase the productivity. The plant growth depends on multiple factors such as soil type, crop type, and weather. Due to a lack of plant growth information and expert advice, most of the farmers fail to get a good yield.

Data Mining is an important domain in which various researches are going on for the past three decades. Data mining plays a major role in analyzing the data sets and predicts the abnormal behavior of any kind of disease, soil categories, etc. From the recent research many data mining algorithms have been proposed to improve the efficiency of data searching, data preprocessing, data cleaning, data aggregation, data classification, feature extraction, normalization, etc.

MATERIALS AND METHODS

We have developed an automated system for soil classification and determine the nutrient, contaminants and other characteristics such as acidity and PH level of soil. After obtaining the above features from the soil 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 data set used, was collected from one of the soil testing laboratories in India. This approach has very sound practical knowledge of soil testing. The outcome of this research will reduce the time consuming process of the soil testing laboratories.

Data Set Collection

The data-set is collected from surveys which are being carried out regularly in Tanjore District. The data-set are acquired by field sampling and then the collected samples are sent for chemical and physical analysis at the soil testing laboratories which are located nearby. It contains information about a number of soil samples taken from various regions of Tanjore district. Dataset has approximately 9 attributes and a total 1988 instances of soil samples. Table1 describes data collected for each soil sample.

Table 1: Attributes Description

Field	Description
Ph	pH value of soil
EC	Electrical conductivity, decisiemen per meter
OC	Organic Carbon, %
P	Phosphorous, ppm
K	Potassium, ppm
Fe	Iron, ppm
Zn	Zinc, ppm
Mn	Manganese, ppm
Cu	Copper, ppm

Automated System

Classification of soil is essential for the identification of soil properties such as the nutrient, contaminants and other characteristics such as acidity and PH level of soil. Automated system can be a very powerful in identifying the soils categories quickly and efficiently. Olden methodologies which are used for classification technique consumes a lot of time and hence it is not a reliable automated system for soil classification.

We propose an automated system for soil classification and determine the nutrient, contaminants and other characteristics such as acidity and PH level of soil. The implementation of an automated system is very difficult without data mining techniques such as classification and combined mining. Other than these data mining techniques, various rules for soil classification were collected from soil testing laboratories. Soil samples were obtained with the help of this automated system.

A STUDY OF SOIL CLASSIFICATION

The classification of soil was considered critical to research because depending upon the soil categories; we have to define the suggestion of the crops which can be cultivated for better yield. In earlier days, the domain knowledge expert determines which crops can be cultivated in that particular soil and which fertilizers should be used for the best yield. The following section describes Naive Bayes, K-Nearest Neighbor algorithm briefly.

Naive Bayes

Naive Bayes classifier is also called as a simple probabilistic classifier. Naive Bayes classifier works based on applying Bayes' theorem with strong (naive) independence assumptions.

Naive Bayes classifier is an efficient classification algorithm for two-class and multi-class classification problems. Two-class is also called as binary classification problems. The technique involved behind this algorithm is very simple to understand when it is described using binary or categorical input values by selecting the most probable hypothesis given in the sample data that we have to use as our prior knowledge to understand about the problem. Bayes' Theorem provides an efficient way to calculate the probability of a hypothesis as per our prior knowledge. Bayes' Theorem is also called as naive Bayes or idiot Bayes because the calculations of the probabilities for each hypothesis are simplified to make their calculation almost similar. Rather than calculating the values of each and every attribute value P(d1, d2, d3|h), they are simply assumed to be conditionally independent and then given the expected target value and calculated as P(d1|h) * P(d2|H) and so on.

There is a very strong assumption that is most unlikely in real data sets, i.e. that the attributes do not interact with each other. Nevertheless, the approach performs surprisingly well on data sets where this assumption does not hold.

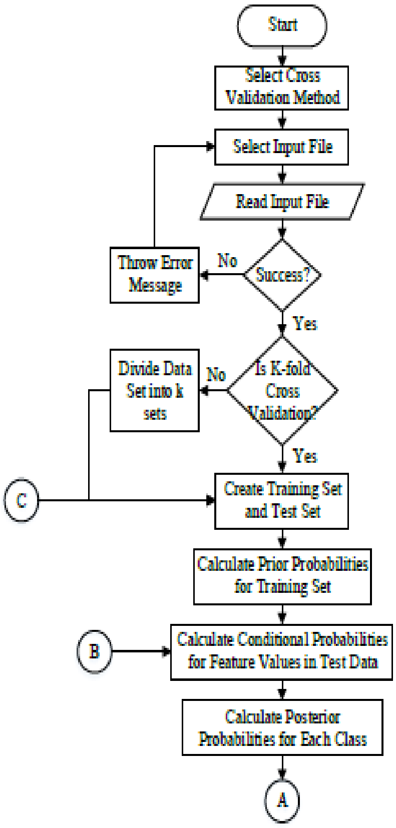
Start
Select Cross Validation Method
Select Input File
Read Input File
Throw Error Message
No
Success?
Yes
Divide Data Set into k sets
No
Is K-fold Cross Validation?
Yes
C
Create Training Set and Test Set
Calculate Prior Probabilities for Training Set
B
Calculate Conditional Probabilities for Feature Values in Test Data
Calculate Posterior Probabilities for Each Class
A

Figure1: Process Flow for Soil Test Report

K-Nearest Neighbor

K Nearest Neighbor is also called as k-NN algorithm; it is a non parametric simple lazy machine learning algorithm. The algorithm procedure is very simple to understand, but it works incredibly well, when it is implemented in practice. Also, it is surprisingly versatile and its applications range from vision to proteins to computational geometry to graphs and so on. K Nearest Neighbor (k-NN) algorithm is also a lazy algorithm. What does it means is that it does not use any sort of training data points to do any generalization technique. In other words, the training phase is very minimal to perform the generalization technique and hence the training phase is pretty fast. Lack of generalization technique means that, K Nearest Neighbor (k-NN) algorithm keeps all the training data sets. More exactly, all the training data sets are needed during the initial testing phase. This is in contrast to other techniques like Support Vector Machine (SVM) where you can discard all non support vectors without any problem.

When the input data set provided to an algorithm is too large, then definitely it will take more time to be processed the data set and it is suspected to be redundant (e.g. The same measurement in both feet and meters) then the large input data set will be transformed into a reduced representation set of features (also named features in vector representation). Transforming the huge input data set into the set of features is called as feature extraction. If the multiple features extracted from the given data set are carefully chosen it is expected that the features set will extract the corresponding information from the input data set in order to perform the desired task using this reduced representation instead of the full size input, either it can be tabular column, graphical notation such as directed or undirected graph, etc. Feature extraction is performed on raw input data set before applying k-NN algorithm on the transformed data in feature space.

Algorithm

```
INPUT: Target data sets X(1,…..k)
OUTPUT: Labels(Y)
Step 1: k-Nearest Neighbor
Step 2: Classify(X,Y,x) //X-training data; Y: class labels of X, x:unknown samples
Step 3: for i = 1 to m do
Step 4: Compute distance d (Xi, x)
Step 5: end for
Step 6: Compute set I containing indices for the k smallest distance d (Xi, x).
Step 7: Return majority label for {Yi, Where I € I}
```

RESULTS AND DISCUSSIONS

The purpose of the study is to examine the most effective techniques to predict the category of the analyzed soil data-sets and provide suggestion of the crops which can be cultivated for better yield. Few of techniques are discussed here. Naive Bayes is very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold and K Nearest Neighbor method is used to classify the soil testing attributes and categories the different kind of soil for better yield of crops.

CONCLUSIONS

In this research work, combined mining and pattern recognition techniques for soil data mining studied. The major aim of this research work is to come out of the techniques being used in the agricultural soil science and its allied area. A comparison of different data mining classification algorithm could produce an efficient solution for soil classification. Better understanding of soils could improve productivity in farming, maintain biodiversity, etc.

REFERENCES

1. ”Fuzzy Logic”. Stanford Encyclopedia of Philosophy. Stanford University. 2006-07-23. Retrieved 2008-09-29.
2. Shalvi D and De Claris N, Unsupervised neural network approach to medical data mining techniques, in Proceedings of IEEE International Joint Conference on Neural Networks, (Alaska), pp. 171-176, May 1998.
3. Crozier, C.R., B. Walls, D.H. Hardy, and J.S. Barnes 2004. “Response of cotton to P and K Soil Fertility Gradients in North Carolina”, Journal of Cotton Science8:130-141(http://journal.cotton.org).
4. Alahakoon D., Halgamuge S.K., and Srinivasan B, Dynamic self organizing maps with controlled growth for knowledge discovery, IEEE Transactions on Neural Networks, vol. 11, pp. 601-614, 2000.
5. Tisdale, S.L., W.L. Nelson, J.D. Beaton, and J.L. Havlin. 1993. “Soil Fertility and Fertilizers”,5th ed. Prentice Hall, NJ.
6. Banerjee M, Mitra S, and Pal S.K, Rough fuzzy MLP: Knowledge encoding and classification, IEEE Transactions on Neural Networks, vol. 9, pp. 1203-1216, 1998.
7. Subba Rao. Indian Agriculture past Laurels and Future Challenges, Indian Agriculture: Current Status, Prospects and Challenges. Convention of Indian Agricultural Universities Association, 27:58-77, December 2002.
8. Lee R. S. T. and Liu J. N. K., Tropical cyclone identification and tracking system using integrated neural oscillatory leastic graph matching and hybrid RBF network track mining techniques, IEEE Transactions on Neural Networks, vol. 11, pp. 680-689, 2000.
9. M. Kumari & S. Godara, (2011), “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction”, International Journal of Computer Science and Technology,, Vol. 2, and ISSN: 0976
10. “Soil test”, Wikipedia, February 2012.
11. Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. and Teller E... Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics, 21(6):1087-1092, 1953.
12. P. Gruhn, F. Goletti, & M. Edelman, (2000) “Integrated Nutrient Management, Soil Fertility, and Sustainable Agriculture: Current Issues and Future Challenges”, International Food Policy Research Institute, N.W. Washington, D.C. U.S.A.; Technical Report.