Design and Development of Efficient Method for Soil Classification

S.SathyaPriya[1] S.Suganthi[2]
[1]Assistant Professor [2]M. Phil. Researcher Scholar
[1,2]Department of Computer Science
[1,2]Sankara College of Science and Commerce, Coimbatore, India

Abstract— Data mining applications have been developed by both commercial and research Centers. These techniques have been used for industrial, commercial and scientific purposes. For example, data mining techniques has been used to analyze large data sets and establish useful classification and patterns in the data sets. "Agricultural and biological research studies have used various techniques of data analysis including, decision trees, statistical machine learning and other analysis methods".
Key words: Soil Data, Soil Classification

I. INTRODUCTION

Soil Classification is the most important one. It influences many other properties and significance of land use and management. The Soil texture is an important property for agriculture soil classification. It influences fertility, drainage, water holding capacity, aeration, tillage, and strength of soils.

This research proposed a new classification method and developed a GUI for soil classification. Multi class support vector machine is used to classification. The soil data used in this research consists of 111 instances with 8 attributes like (i.e., 'Clay', 'Clayey Peat', 'Clayey Sand', 'Humus Clay', 'Peat', 'Sandy Clay', and 'Silty Sand'). The texture of the Soil data is varied from sand to silty clay loam where as in sub-surface horizons it varied from sand to clay.

II. CLASSIFICATION TECHNIQUES

Classification consists of assigning a class label to a set of unclassified cases.

- Supervised Classification: The set of possible classes is known in advance.
- Unsupervised Classification: Set of possible classes is not known. After classification we can try to assign a name to that class. Unsupervised classification is called clustering.

A. Supervised Classification

- The input data, also called the training set, consists of multiple records each having multiple attributes or features.
- Each record is tagged with a class label.
- The objective of classification is to analyze the input data and to develop an accurate description or model for each class using the features present in the data.

This model is used to classify test data for which the class descriptions are not known.(2)

1) Bayesian Classifier

- The classes are mutually exclusive and exhaustive.
- The attributes are independent given the class. Called "Naïve" classifier because of these assumptions. Empirically proven to be useful. Scales very well.

Bayesian classifier is defined by a set C of classes and a set A of attributes. A generic class belonging to C is denoted by c_j and a generic attribute belonging to A as A_i.(3)

2) K Nearest Neighbors Algorithm

The closest neighbor (NN) rule distinguishes the classification of unknown data point on the basis of its closest neighbor whose class is already known. M. Cover and P. E. Hart purpose k nearest neighbour (KNN) in which nearest neighbor is computed on the basis of estimation of k that indicates how many nearest neighbors are to be considered to characterize class of a sample data point. It makes utilization of the more than one closest neighbor to determine the class in which the given data point belongs to and and consequently it is called as KNN. These data samples are needed to be in the memory at the run time and hence they are referred to as memory-based technique (4).

- K ← number of nearest neighbors
- For each object Xin the test set do
- Calculate the distance D(X,Y) between X and every object Y in the training set
- Neighborhood ← the k neighbors in the training set closest to X
- X.class ← SelectClass (neighborhood)
- End for

3) Decision Tree Induction

Decision tree classification is the learning of decision trees from class labelled training tuples. A decision tree is a flowchart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

a) Advantages

Amongst other data mining methods, decision trees have various advantages. Decision trees are simple to understand and interpret. They require little data and are able to handle both numerical and categorical data. It is possible to validate model using statistical tests. They are robust in nature, therefore, they perform well even if its assumptions are somewhat violated by the true model from which the data were generated. Decision trees perform well with large data in a short time. Large amounts of data can be analyzed using personal computers in a time short enough to enable stakeholders to take decisions based on its analysis.

b) Limitations

The problem of learning an optimal decision tree is known to be NP-complete. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. Decision -tree learners create over-complex trees that do not generalize the data well. This is called over fitting. Mechanisms such as pruning are necessary to avoid this problem. [1]

III. PROPOSED SOIL CLASSIFICATION TECHNIQUE

Indian Standard Classification System (ISC) was adopted by Bureau of Indian Standards is in many respect similar to the Unified Soil Classification (USC) system.

Soils are divided into three broad divisions:

1) Coarse grained soils, when 50% or more of the total material by weight is retained on 75 micro IS sieve.
2) For fine grained soils, when more than 50% of the total material passes through 75 micron IS sieve.
3) If the soil is highly organic and contains a large percentage of organic matter and particles of decomposed vegetation, it is kept in a separate category marked as peat (Pt).

In all there are 18 groups of soils: 8 groups of coarse grained, 9 groups of fine grained and one of peat.950

A. Dataset Collection

The dataset involves the major part of the work. It was carried out in Coimbatore district. The primary data for soil survey was carried out by field sampling. Those acquired samples were sent to laboratories for chemical and physical analysis. It contains number of soil samples taken from various regions of the district. Dataset has 9 attributes and a total 2988 instances of soil samples.

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

Table 1: describes data collected for each soil sample.

IV. AUTOMATED SYSTEM

Soil classification is very much necessary for identification of properties of soil. The enhanced expert automated syatem can more efficient in performing the task. The Traditional approach of identifying soil was very slow and not much efficient. The process included tables, flow- chart.

This paper proposes an efficient automated system to enhance the classification process. Based on the fertility of the soil the classification process is undertaken. To perform this task the rules for soil classification was collected from lab. The soil sample instances were classified into: Very High, High, Moderately High, Moderate, Low, and Very Low.

Classifier	Naviye Bayes	KNN	Decision Tree
Correctly Classified Instances	985	2998	3826
Incorrectly Classified Instances	3556	156	265
Accuracy	38.60%	95.68%	96.86%
Mean Absolute Error	1.229	0.0125	0.0577

Table 2: Comparison of Different Classifiers

Actual Value using Soil Testing	Predicted Value	Error
10.3	10.661	0.361
7.7	7.431	-0.269
13.5	4.6	4.653
4.6	9.5	8.478
9.5	1.022	3.035
2.9	0.135	4.915
5.1	4.915	0.185
15.3	15.667	0.367
7	7.402	0.402
18.4	18.743	0.343
4.4	4.388	-0.012

Table 3: Relative Absolute Error

Here the Relative Absolute Error is nearly same for both the prediction algorithms. Even though Least Median Square regression gives better numeric predictions but the time taken to build the model is 67 times that of Linear Regression, hence computational cost used by Linear Regression is much lower than that of least median square technique(5).

REFERENCES

[1] CLASSIFICATION TECHNIQUES ANALYSIS Megha Gupta, Naveen Aggarwal NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, 19 -20 March 2010.
[2] SLIQ: A fast scalable Classifier for Data Mining; Manish Mehta, Rakesh Agarwal and Jorma Rissanen
[3] Robust Bayesian Classifier, Presented by Chandrasekhar Jakkampudi.
[4] A Comparative Study of Classification Techniques in Data Mining Algorithms Sagar S. Nikam,Department of Computer Science, K.K.Wagh College of Agriculture, Nashik, India. Oriental journal of computer science and technology.
[5] Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction, JayGholap, Anurag Ingole, JayeshGohil, ShaileshGargade, Vahida Attar.
[6] A.Kumar&N.Kannathasan, (2011),"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 2012.
[7] L. Armstrong,. Diepeveen& R. Maddern,(2004),"The application of data mining techniques to characterize agricultural soil profiles",
[8] "Methods Manual-Soil Testing in India", Department of Agriculture & Cooperation Ministry of Agriculture Government of India, 2011
[9] "Naïve Bayes",Wikipedia,February2012
[10] "C4.5 (J48)", Wikipedia,February 2012
[11] W. Cohen,(1995),"Fast Effective Rule Induction, in Twelfth International Conference on Machine Learning.
[12] I. Witten & F. Eibe,(2005),"Data Mining: Practical Machine Learning Tools and techniques" 2nd Edition, San Francisco: Morgan Kaufmann,
[13] A. Al-Rawas,S. Al-Alwai, (1998),A. Bisma Y. Al-Alwai,"Soil Classification Decision Support System

Using An Expert System Approach", Engineering Journal of University of Qatar, Vol. 11, p. 103-115.

[14] S. Cunningham& G. Holmes,(1999),"Developing innovative applications in agriculture sing data mining", Department of Computer Science University of Waikato Hamilton, New Zealand, Technical Report.

