Soil Classification Using Data Mining Techniques: A Comparative Study

Raghuveer Yadav, Sujal Rani

Abstract: **Soil Classification deals with the systematic categorization of soils based on distinguished characteristics as well as criteria. We developed Data Mining techniques like: GATree, Fuzzy Classification rules and Fuzzy C - Means algorithm for classifying soil texture in agriculture soil data. In this paper, we give a comparative study of developed algorithms. The study is used to compare and analyze the soil data.**

***Keywords*--Data Mining, Soil Database, Classification, Genetic Algorithm, Fuzzy Classification, Fuzzy Clustering.**

I. INTRODUCTION

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” as in [1]. This research determines whether data mining techniques can be used to improve pattern recognition and analysis of large soil profile experimental datasets. Further, the research is aimed to establish if data mining techniques can be used to assist in the classification methods by determining whether meaningful patterns exist across various soils profiles. Various data mining algorithms [10] [11] [12] were applied to the Soil data and were analyzed to determine the best algorithm to classify soil textures.

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.

Data mining algorithms that we have applied to classify agriculture soil are: GATree, Fuzzy classification rules and Fuzzy C-Means algorithm. These techniques were applied on the collected soil data [2] and the achieved performances were compared and analyzed. GATree and Fuzzy Classification rules were used for supervised learning. However, classification based on Fuzzy rules gives much performance than GATree. For Unsupervised learning Fuzzy C-Means algorithm was used for classifying the soil data.

II. SOIL CLASSIFICATION

Soil classification deals with the systematic categorization of soils based on distinguishing characteristics as well as criteria that dictate choices in use. Soil classification is a dynamic subject, from the structure of the system itself, to the definitions of classes, and finally in the application in the field. Soil classification can be approached from the perspective of soil as a material and soil as a resource.

The most common engineering classification system for soils is the Unified Soil Classification System (USCS) [6]. The USCS has three major classification groups: (1) coarse-grained soils (e.g. sands and gravels); (2) fine-grained soils (e.g. silts and clays); and (3) highly organic soils (referred to as "peat"). The USCS further subdivides the three major soil classes for clarification. A full geotechnical engineering soil description will also include other properties of the soil including color, in-situ moisture content, in-situ strength, and somewhat more detail about the material properties of the soil that is provided by the USCS code.

The soils are classified into different orders, sub-orders, great groups, sub-groups, families and finally into series as per USDA Soil Taxonomy as in [8][9]. The solid phase of soil can be divided into mineral matter and organic matter. The mineral particles can be further subdivided into classes based on size. The classification of soil particles according to size are Sand, Silt, Clay. The proposition of Sand, Silt, Clay present in soil determines its texture.

A. Soil Data

The soil data used in this paper consists of 111 instances with 8 attributes like (i.e., Depth, Sand, Silt, Clay, Sandbysilt, Sandbyclay, Sandbysiltclay, TextureClass). 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 as in [2]. Table 1 shows the different soil survey symbols.

TABLE 1
SOIL SURVEY SYMBOLS

S	**Sand**
Sicl	**Silty Clay Loam**
Sic	**Silty Clay**
C	**Clay**
Sl	**Sandy loam**
Cl	**Clay loam**
Sil	**Silty Loam**
L	**Loam**
Ls	**Loamy sand**
Scl	**Sand Clay Loam**
Sc	**Sand Clay**

III. CLASSIFICATION ALGORITHMS

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take a collection of cases as input, each belong to a small number of classes described by a fixed set of attributes, and outputs a classifier that can accurately predict the class to which a new case belongs.

Datasets can have nominal, numeric or mixed attributes and classes. Not all classification algorithms perform well for different types of attributes, classes and for datasets of different sizes. In order to design a generic classification tool, one should consider the behavior of various existing classification algorithms on different datasets. In this paper, the data mining algorithms like GATree, Fuzzy classification rules and Fuzzy C-Means algorithm are analyzed. These are applied to Soil data and then compared for efficient algorithm to classify soil texture**.**

A. GATree

Genetic Algorithms are optimization and Machine Learning algorithms that are based loosely on the process of biological evolution. Interest in genetic algorithms has increased recently in conjunction with an increase in interest in other algorithms based on natural processes, including simulated annealing and Neural Networks. Genetic Algorithms have been widely used in a wide variety of applications including combinational optimization and knowledge discovery. A Genetic Algorithm proceeds by choosing chromosomes to serve as parents and then replacing members of the current population with new chromosomes that are (possibly modified) copies of the parents. The process of reproduction and population replacement goes on until a stopping criterion (the achievement of a performance target or the usage of an allotted amount of CPU time for instance) is met.

The classification of Soil data [10] using GATree tool, which is a decision tree builder which is based on Genetic Algorithms (GAs). GATree offers some unique features that are not found in any other tree inducers while at the same time it can produce better results for many difficult problems.

B. Fuzzy Classification Rules

Fuzzy Logic was initiated to represent or manipulate data and information processing, non statistical uncertainties. It is mathematically used to represent uncertainty and vagueness and to provide formalized tools for dealing with the impression intrinsic to many problems. Fuzzy Logic provides an inference morphology that enables approximate human reasoning capabilities that are applied to knowledge based systems. The theory of Fuzzy Logic provides a mathematical strength to capture the uncertainties associated with human cognitive processes, such as thinking and reasoning. The development of fuzzy logic was motivated in large measure by the need for a conceptual frame-work which can address the issue of uncertainty and lexical imprecision.

Fuzzy classification rules are widely considered as a well suited representation of classification knowledge with uncertainty and they allow readable and interpretable rule bases. The most important task in the design of fuzzy classification systems is to find a set of fuzzy rules from training data with uncertainty to deal with a specific classification problem.

We proposed an algorithm that was implemented in programming language ‘C’. It generated Fuzzy rules from training data to deal the soil data classification by first defining the membership functions for the input attributes of the soil data, and then generating the initial fuzzy rules for the training data [11].

C. Fuzzy C – Means Algorithm

In unsupervised classification, class-labels are not known. In our study, Fuzzy C-means algorithm is used to deal Unsupervised data with uncertainty. The goal of Fuzzy C-means algorithm is to group the objects into clusters based only on their observable features such that each cluster contains objects that share some important properties.

Most analytical fuzzy clustering algorithms are based on optimization of the basic c-means objective function, or some modification of it. Fuzzy C-Mean (FCM) clustering is grouped into n clusters with every data point in the dataset belonging to every cluster that will have a high degree of belonging or membership to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belonging or membership to that cluster.

We have applied fuzzy C-means algorithm for Soil data which consists of 10 texture classes. The fuzzy classifiers classify each texture class by clustering them. This is implemented in MATLAB [12].

IV. COMPARATIVE STUDY

This section deals a comparative study of different algorithms which are mentioned in section III. The objective of this study is to know the efficient technique to classify the soil texture.

A. GATree

With the Genetic Algorithms, the binary Decision Tree explains the prediction of the category of Soil data that is built first. To generate the decision tree for the dataset it takes more time and then the rules are classified with that decision tree. The accuracy of Genetic Algorithm is shown in the table 2. Genetic Algorithms are not suitable for managing imprecise or uncertainty in soil data. Genetic based Decision rules are effective for soil data.

TABLE 2
THE ACCURACY OF GENETIC ALGORITHM

Test	Accuracy	Correct Classified
1	0	0/3
2	0.33	1/3
3	0.67	2/3
4	0.67	2/3
5	0.33	1/3
6	1	3/3
7	0.33	1/3
8	0.33	1/3
9	0.33	1/3
10	0.67	2/3

In the above table for the training size of 27 and test size of 3 for different tests the accuracy and the correctly classified instances is shown. The average accuracy for the instances is 0.46.

B. Fuzzy Classification Rules

The proposed Fuzzy Classification algorithm presented in the paper [11] generates rules for the defined membership functions of the input attributes and we have converted each training datum into a fuzzy rule. This algorithm is implemented in C language. Table 3 represents average accuracy rate and average number of rules generated for the existing method [13] and the proposed method [11].

TABLE 3
THE AVERAGE ACCURACY RATE AND AVERAGE NUMBER OF RULES

Method	I	II
Average accuracy rate	96.6%	99%
Average number of rules generated	9.01	9.9

The program that was developed to generate fuzzy classification rules was tested for complexity using the tool SourceMonitor. The metric summary sheet that is generated from the SourceMonitor tool is displayed below for your reference.

C. Fuzzy C-Means Algorithm

Fuzzy C-means clustering is a partitioning method, carried out through an iterative optimization of the objective function. In this method, each feature vector representing the data has a degree of membership into all the clusters and the algorithm works to minimize an objective function. The Fuzzy C-Mean (FCM) clustering is grouped into n clusters with soil data belonging to every cluster that will have a high degree of belonging or membership to that cluster and another data point that lies far away from the center of a cluster that will have a low degree of belonging or membership to that cluster. The average accuracy to classify the clusters is 90.90 [12]. The average number of generated fuzzy rules with this method is zero due to the fact that it is clustering technique.

D. The accuracy of three Algorithms

Table 4 represents the Accuracy and interesting rules generated by GATree for general data and Fuzzy Classification rules for uncertain data. It concludes that for supervised soil data, Fuzzy Classification rules generated more interesting rules than GATree. However, GATree is used for classification of classic soil data and Fuzzy classification rule method is used for classification of Fuzzy soil data.

TABLE 4
THE ACCURACY AND RULES OF
GATREE AND FUZZY CLASSIFICATION RULES

Method	Average accuracy rate	Average number of rules generated
Genetic Algorithm	46.67%	17
Fuzzy Classification Rules	99%	9.9

Finally, the accuracy rate of Supervised and Unsupervised algorithms are given in table 7. Eventhough accuracy rate of Fuzzy Classification rules is higher than Fuzzy C-means algorithm; Fuzzy C-means algorithm is more suitable algorithm for Unsupervised Fuzzy soil data.

TABLE 5
THE AVERAGE ACCURACY RATES OF
SUPERVISED AND UNSUPERVISED ALGORITHMS

Method	Average accuracy rate
Genetic Algorithm	46.67%
Fuzzy Classification Rules	99%
Fuzzy Clustering	90.90%

V. CONCLUSION

In this paper, we have compared the effectiveness of the classification algorithms Genetic algorithm, Fuzzy classification and Fuzzy clustering. The achieved performances are compared and analyzed on the collected Supervised and Unsupervised Soil data. GATree and Fuzzy Classification rules were used for supervised learning. However, classification based on Fuzzy rules gives much performance than GATree. For Unsupervised learning Fuzzy C-Means algorithm was used for classifying the soil data. This helps one to classify soil texture based on soil properties effectively, which influences fertility, drainage, water holding capacity, aeration, tillage, and bearing strength of soils.

REFERENCES

[1] Cunningham, S. J., and Holmes, G. (1999). Developing innovative applications in agriculture using data mining. In the *Proceedings of the Southeast Asia Regional Computer Confederation Conference*,1999.

[2] A Thesis by D.Basavaraju, Characterisation and classification of soils in Chandragiri mandal of Chittoor district,Andhra Pradesh.

[3] Jiawei Han and Micheline Kamber, Data Mining concepts and Techniques, Elsevier..

[4] Minaei-Bidgoli, B., Punch, W. Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System. Genetic and Evolutionary Computation, Part II. 2003. pp.2252–2263.

[5] Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufman. 1993.

[6] Isbell, R. F. (1996). The Australian Soil Classification.Australian soil and land survey handbook. (Vol. 4).Collingwood, Victoria, Australia: CSIRO Publishing.

[7] Ibrahim, R. S. (1999). Data Mining of Machine Learning Performance Data. Unpublished Master of Applied Science (Information Technology), Publisher; RMITUniversity Press.

[8] Soil Survey Staff 1998, Keys to soil taxonomy. Eight Edition, Natural Resource Conservation Services, USDA, Blacksburg, Virginia.

[9] Soil Survey Staff 1951, Soil Survey Manual. US Department of Agricultural Hand book No. 18.

[10] P. Bhargavi, and Dr. S. Jyothi , Soil Classification Using GATree. International journal of computer science & information Technology (IJCSIT) Vol.2, No.5, October 2010.

[11] P. Bhargavi, and Dr. S. Jyothi , Soil Classification by Generating Fuzzy rules. International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2571-2576.

[12] P. Bhargavi, and Dr. S. Jyothi , Fuzzy C-Means Classifier for Soil Data. International Journal of Computer Applications (0975 – 8887), Volume

6– No.4, September 2010.

[13] H. L. Lin and S. M. Chen, Generating weighted fuzzy rules from training data for handling fuzzy classification problems Proc. 2000 Int. Computer Symp.: Workshop Artificial Intelligence pp. 11 - 18, 2000

Bhargavi Peyakunta is working as Associate Professor in the Department of Computer Science and Engineering, MJR College of Engineering and Technology(Affiliated to JNTU, Anantapur) , Piler.Andhra Pradesh, India **Educational Qualifications:** M.Sc in Computer Science from Sri Krishnadevaraya University, Anantapur and M.Tech degree from Sri Vinayaka missions University, Salem, India. **Teaching & Research Experience:** 13 years of teaching experience & 5 years of research experience. **Current Research Interests:** Data Mining, Fuzzy Systems, Genetic Algorithms and GIS.

Jyothi Singaraju is working as Head (I/C) CSE & IT in Sri Padmavathi Mahila Visvavidyalayam(SPMVV), Tirupati. **Educational Qualifications:** M.Sc in Applied Mathematics from S.V.University, Tirupati , M.S in Software Systems from BITS, Pilani, & Ph.D in Fuzzy Relational Data Bases from S.V.University, Tirupati. **Teaching & Research Experience:** 20 years teaching experience & 24 years research experience. **Current Research Interests:** Fuzzy Systems, Neural Networks, Data Mining, Data Base Management Systems, Genetic Algorithms, Bioinformatics, Bio Metrics and GIS.