Agricultural Journal 6 (3): 82-86, 2011
ISSN: 1816-9155


# Classification of Agricultural Land Soils: A Data Mining Approach

[1]V. Ramesh and [2]K. Ramar
[1]Department of CSA, SCSVMV University, Kanchipuram, India
[2]Sri Vidya College of Engineering and Technology, Virudhunagar, Tamil Nadu, India

**Abstract:** The problem of the knowledge acquisition and efficient knowledge exploitation is very popular also in agriculture area. One of the methods for knowledge acquisition from the existing agricultural databases is the methods of classification. In agricultural decision making process, weather and soil characteristics are play an important role. This research aimed to assess the various classification techniques of data mining and apply them to a soil science database to establish if meaningful relationships can be found. A large data set of soil database is extracted from the Soil Science and Agricultural Department, Kanchipuram and National Informatics Centre, Tamil Nadu. The application of data mining techniques has never been conducted for Tamil Nadu soil data sets. The research compares the different classifiers and the outcome of this research could improve the management and systems of soil uses throughout a large number of fields that include agriculture, horticulture, environmental and land use management.

**Key words:** Data mining, soil profiles, agriculture, classification techniques, classifiers, horticulture

## INTRODUCTION

Data mining 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 natural trees, statistical machine learning and other analysis methods (Cunningham and Holmes, 1999). The analysis of agricultural data sets with various data mining techniques may yield outcomes useful to researchers in the Agricultural field. Data Mining software applications includes various methodologies that have been developed by both commercial and research centers. These techniques have been used for industrial, commercial and scientific purposes. Agricultural and biological research studies have used various techniques of data analysis including natural trees, statistical machine learning and other analysis tools. This research determined whether data mining techniques could also be used to classify soils that analyze large soil profile experimental datasets. The research aimed to establish if data mining techniques can be used to analyze different classification methods by determining whether meaningful pattern exists across various soil profiles characterized at various research sites. The data set has been assembled from soil surveys at various agricultural areas located in Kanchipuram district, Tamil Nadu, India. The research has utilized existing data collected from seven commonly occurring soil types in order to classify soils and correlations between a numbers of soil properties. The soil studies which have been conducted by the Soil Science and Agricultural Department, Kanchipuram provide a vast amount of information on the classification of soil profiles and chemical characteristics. The analysis of these agricultural data sets with various data mining techniques may yield outcomes useful to researchers in the Soil Sciences and Agricultural Chemistry. This research has a number of potential benefits to the soil science. The overall aim of the research is compare the different classifiers and the outcome of this research may have many benefits to agriculture, soil management and environment.

**Review of literature:** A number of studies have been carried out on the application of data mining techniques for agricultural data sets. For example, the k-means method is used to perform forecasts of the pollution in the atmosphere (Jorquera *et al*., 2001), the k-nearest neighbor is applied for simulating daily precipitations and other weather variables (Rajagopalan and Lall, 1999) and the different possible changes of the weather scenarios are analyzed using SVMs (Tripathi *et al*., 2006).

Data mining techniques are often used to study soil characteristics. As an example, the k-means approach is used for classifying soils in combination with GPS based techniques (Verheyen *et al*., 2001). The research conducted by Ibrahim was to apply unsupervised clustering to analyze the generated clusters and determine

**Corresponding Author:** V. Ramesh, Department of CSA, SCSVMV University, Kanchipuram, India

whether there are any significant patterns. Decision tree analysis method has been used in the prediction of natural datasets in agriculture and was found to be useful in prediction of soil depth for a dataset.

In another study, WEKA was used to develop a classification system for the sorting and grading of mushrooms. The system developed a classification system that could sort mushrooms into grades and attained a level of accuracy equal to or greater than the human inspectors. The process involved the pre-processing of the data not just cleaning the data but also creating a test dataset in conjunction with agricultural researchers.

## MATERIALS AND METHOD

**Soil classification:** The classification of the soil was considered critical to the study because the soil types must be the same in all locations across the study are for the results to be accurate. 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. Engineers, typically Geotechnical engineers, classify soils according to their engineering properties as they relate to use for foundation support or building material. Modern engineering classification systems are designed to allow an easy transition from field observations to basic predictions of soil engineering properties and behaviors.

**The USDA soil taxonomy:** The soil taxonomy developed since, the early 1950's is the most comprehensive soil classification system in the world, developed with international cooperation, it is sometimes described as the best system so far. However for use with the soils of the tropics, the system would need continuous improvement.

**The FAO/UNESCO system:** The FAO/UNESCO system was devised more as a tool for the preparation of a small-scale soil map of the world than a comprehensive system of soil classification. The map shows only the presence of major soils being associations of many soils combined in general units. The legend of the soil map of the world lists 106 units classified into 26 groupings. The soil units correspond roughly to great groups from the USDA soil taxonomy while larger main grouping are similar to the USDA soil suborder.

**The french system (ORSTROM/INRA):** The so called French system of classifying soils is based on principles of soil evolution and degree of evolution of soil profiles. It also takes into account humus type, structure and the degree of hydromorphism.

**Classification of soil in Kanchipuram district:** A set of soil properties are diagnostic for differentiation of pedons. The differentiating characters are the soil properties that can be observed in the field or measured in the laboratory or can be inferred in the field. Some diagnostic soil horizons both surface and sub surfaces, soil moisture regimes, soil temperature regimes and physical, physical chemical and chemical properties of soils determined were used as criteria for classifying soils.

According to soil survey manual of Indian Government, the soils of Kanchipuram district are categorizes into 8 classes. The classes generally range from class 1, the best land for agricultural production to class 8, the least productive. In general, class 1 through class 4 are for row production and 5 through 8 are not suitable for row crop production for various reasons. Class 1 is the best land for row crop farming. It is level, well drained, deep, medium textured not subject to erosion or flooding and easily cultivated. Class 1 is just as good but it may have some limitations such as sloping land or slight erosion. Class 3 can still be cultivated but it has some severe limitations.

The land may have moderate slope, erosion or a shallow root zone. Class 4 has severe limitation but can still be cultivated with good management practices. Class 5 is nearly level but has some property which makes it unsuitable for farming. It may be dry, very rocky or most often very wet. This class is quite suitable for pasture, wildlife habitat or forest production. Class 4 is just a more serious version of 5.

It has severe limitations but can be used for the some things. Class 7 has some severe limiting properties. It may be steep or be severely eroded and have deep gullies and it may be very course. This can be turned into pasture but grazing must be controlled. It can also be used as forest or recreation. Class 8 has one or more extreme limitations. It should be left in its natural state for recreation and wild life. It has little agriculture value.

**Classification in data mining:** Techniques used in data mining can be divided in to 2 big groups. The first group contains techniques that are represented by a set of instructions or sub-tasks to carry out in order to perform a certain task. In this view, a technique can be seen as a sort of recipe to follow which must be clear and unambiguous for the executor. If the task is to cook pasta

with tomatoes the recipe may by heat water to the boiling point and then throw the pasta in and check whether the pasta has reached the point of being at dente; drain the pasta and add preheated tomato sauce and cheese. Even a novice chef would be able to achieve the result following this receipt. Moreover, another way to learn how to cook pasta is to use previous cooking experience and try to generalize this experience and find a solution for the current problem.

This is the philosophy, the second group of data mining techniques follows. A technique, in this case does not provide a recipe for performing task but it rather provides the instructions for learning in some way how to perform the task. As a newborn baby learns how to speak by acquiring stimuli from the environment, a computational technique must be taught how to perform its duties. In the case of novice chef, he has all the needed ingredients at the start but he does not know how to obtain the final product. In this case, he does not have the recipe. However, he has the capability of learning from the experience and after a certain number of trials, he will be able to transform the initial ingredients into a delicious tomato pasta dish and be able to write his own recipe.

Number of data mining techniques can be divided in two subgroups as discussed above. For instance, k-nearest neighbor method provides a set of instructions for classification purposes and hence, it belongs to the first group. Neural networks and support vector machines instead follow particular methods for learning how to classify data. The task of supervised classification, i.e., learning to predict class memberships of test cases given labeled training cases is a familiar machine learning problem.

A related problem is unsupervised classification where training cases are also unlabeled. Here, one tries to predict all features of new cases the best classification is the least surprised by new cases. This type of classification, related to clustering is often very useful in exploratory data analysis where one has few preconceptions about what structures new data may hold. Bayes theory gives a mathematical calculus of degrees of belief, describing what it means for beliefs to be consistent and how they should change with evidence. This study briefly reviews that theory describes an approach to making it tractable and comments on the resulting trade offs.

In general, a Bayesian agent uses a single real number to describe its degree of belief in each proposition of interest. This assumption, together with some other assumptions about how evidence should affect beliefs, leads to the standard probability axioms. Disadvantages include being forced to be explicit about the space of models one is searching in though this can be good discipline. One must deal with some difficult integrals and sums, although there is a huge literature to help one here. Finally, it is not clear how one can take the computational cost of doing a Bayesian analysis into account without a crippling infinite regress. Some often perceived disadvantages of Bayesian analysis are really not problems in practice. To do a Bayesian analysis of this, we need to make this vague notion more precise, choosing specific mathematical formulas which say how likely any particular combination of evidence would be. Steps for building a Bayesian classifier:

- Collect class exemplars
- Estimate class a priori probabilities
- Estimate class means
- Form covariance matrices, find the inverse and determinant for each
- Form the discriminant function for each class

The motivation behind the development of Bayesian networks has its roots in the regular study of Bayesian probabilistic theory which is a branch of mathematical probability and allows us to model uncertainty about the aim and outcome of interest by combining experimental knowledge and observational evidences. The following chapter will gives us a structure to develop any Bayesian network for any kind of problem. In order to get an entire over view from basic to advanced application by considering an example of type of data or observation and different classification technique which we are dealing with in a project.

The 5 classes of BN classifiers are Naive-Bayes, tree augmented Naive-Bayes, Bayesian network augmented Naive-Bayes, Bayesian multi-nets and general Bayesian Networks. Unlike other classifiers the Naive-Bayes has been used as an effective classifier for many years.

**Naive Bayes classifier:** Naive Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (Naive) independence assumptions. A more descriptive term for the underlying probability model would be independent feature model.

In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to an apple if it is red, round and about 4 in diameter. Even though, these features depend on the existence of the other features, a Naive Bayes classifier considers all of the properties to independently contribute to the probability

that this fruit is an apple. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning, setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood in other words, one can study with the Naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

Inspite of their Naive design and apparently over-simplified assumptions, Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of Naive Bayes classifiers.

An advantage of the Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

The Naive Bayesian classifier is fast and incremental can deal with discrete and continuous attributes has excellent performance in real-life problems and can explain its decisions as the sum of informational gains. In this study, the algorithm of the Naive Bayesian classifier is applied successively enabling it to solve also non-linear problems while retaining all advantages of Naive Bayes. The comparison of performance in various domains confirms the advantages of successive learning and suggests its application to other learning algorithms.

**Data mining process:** The data mining process was conducted in accordance with the results of the statistical analysis. The following steps are a general outline of the procedure that allowed a cluster analysis to be conducted on the dataset.

**Data collection cleaning and checking:** Relevant data was selected from a subset of the soil science database. The soil samples collected from the various regions of Kanchipuram district. Among 2045 soil samples 1500 samples are taken for classification.

**Data formatting:** The data was formatted into an Excel format from the Access database, based on the 10 soil types and relevant related fields. The data was then copies into a single Excel spread sheet. The Excel spread sheet was then formatted to replace any null or missing values in the soil data set to allow coding for the file in the next phase.

**Data coding:** The soil data set was then converted into a comma delimited (CSV) format file for the Excel spread sheet. This file was then saved and opened using a text editor. The text editor was used to format and code the data into the type that will allow the data mining techniques and programs to be applied to it. The coding was formatted so that the input will recognize names of the attributes, the type of value of each attribute and the range of all attributes. Coding was then conducted to allow the machine learning algorithms to be applied to the soil data to provide relevant outcomes that were required in the research.

## RESULTS

The analysis and interpretation of classification is a time consuming process that requires a deep understanding of statistics. The process requires a large amount of time to complete and expert analysis to examine any classification and relationships within the data.

**Statistical results:** The research activities involved a process to establish if classification could be found in the data. These processes involved the statistical manipulation of the data set in Excel. The aim of the research was to determine if a relationship or correlation can be established with soil data set. The process involved the creation of analysis tools and charting the data so that the classification of soils is displayed and experts can interpret the findings.

**Data mining results:** The WEKA (Waikato Environment for Knowledge Analysis) workbench is an open source collection of state-of-the-art machine learning algorithms and data pre-processing tools. WEKA data mining software is used to determine if any advantage could be gained in both time saving and interpretation of the soil data set. The application of the data to WEKA required that some preprocessing be undertaken. The data set produced in Excel for the statistical processes were copied and then converted into. CSV file format to allow them to be applied to WEKA. The CSV file extension allowed initial analysis to be conducted with later conversion to be taken in to an ARFF WEKA data file for the experimental outcome to be saved. The data mining platform allowed number of data interpretations including classify, cluster and associate routines to be conducted after the pre processing stage. The soil data set did not require any filtering because of the limited amount of missing values and the outcomes required by the researchers. The initial screen provided a set of information that is required by the researchers and took a

Table 1: Experimental results

| Classifiers | Relative absolute error | Mean absolute error | Correct | Kappa statistics |
|---|---|---|---|---|
| Bayes.NaiveBayes | 0.351 | 0.001 | 100.0 | 1.00 |
| Bayes.BayesNet | 10.770 | 0.031 | 92.3 | 0.82 |
| Bayes.NaiveBayes updatable | 0.350 | 0.001 | 100.0 | 1.00 |
| J48 | 23.700 | 0.068 | 92.3 | 0.79 |
| Random forest | 19.690 | 0.051 | 100.0 | 1.00 |

large amount of time to complete with the current statistical methods. The full soil data set was applied to the Naive Bayes to classify the soils and could be established with the model being constructed using a training model to classify the training data set and see the correctly classified instances and also apply the Naive Bayes to test set and see the correctly and in correctly instances. Determine the accuracy when compared with each other.

The results are when Naive Bayes classifier is applied to the soil data set the instances are 100% classified. The other classifiers like Bayes.BayesNet, Bayes.NaiveBayes up datable, trees. J48, trees.RandomForest are also applied to the soil data and results are shown in the Table 1. The Kappa statistic, mean absolute error, root mean squared error, relative absolute error are less than the remaining classifiers like Bayesian classifier.

**Experimental results:** The time to build the Naive Bayes classifier is less than the remaining classifier. Kappa statistics is a measure of degree of nonrandom agreement between observers and/or measurement of a specific categorical variable. The relative absolute error and mean absolute errors of Bayes.Naive Bayes are also minimum with compare to other classifier. So, the Naive Bayes classifier is the efficient classification technique among remaining classification techniques. Normalized expected cost of Naive Bayes is more accurate when compared to Bayesian network.

## CONCLUSION

The experiments conducted analyzed small number of traits contained within the dataset to determine their effectiveness when compared with standard statistical techniques. The agriculture soil profiles that are used in this research were selected for completeness and for ease classification of soils. The recommendations arising from this research implies that data mining techniques may be applied in the field of soil research in the future as they will provide research tools for the comparison of large amount of data. Data mining techniques when applied to an agricultural soil profile may improve the verification of valid soil profile may improve the verification of valid patterns and profile classification when compared to standard statistical analysis techniques.

## REFERENCES

Cunningham, S.J. and G. Holmes, 1999. Developing innovative applications in agriculture using data mining. Proceedings of the Southeast Asia Egional

Jorquera, H., R. Perez, A. Cipriano and G. Acuna, 2001. Short Term Forecasting of Air Pollution Episodes. In: Environmental Modeling, Zannetti, P. (Ed.). WIT Press, UK.

Rajagopalan, B. and U. Lall, 1999. A K-nearest neighbor simulator for daily precipitation and other weather variables. Water Resour. Res., 35: 3089-3101.

Tripathi, S., V.V. Srinivas and R.S. Najundiah, 2006. Downscaling of precipitation for climate change scenarios: A support vector machine approach. J. Hydrol., 330: 621-640.

Verheyen, K., D. Adriaens, M. Hermy and S. Deckers, 2001. High resolution continuous soil classification using morphological soil profile descriptions. Geoderma, 101: 31-48.