# Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils

**P.Bhargavi,M.Sc.,M.Tech.**
Department of CSE
Madanapalli Institue of Tecnology and Science,Madanapalli.

**Dr.S.Jyothi,M.Sc.,M.S,Ph.d.**
Head,Department of Computer Science
Sri Padmavathi Mahila Viswa Vidyalayam,Tirupati

## Summary
The advances in computing and information storage have provided vast amounts of data. The challenge has been to extract knowledge from this raw data that has lead to new methods and techniques such as data mining that can bridge the knowledge gap. This research aimed to assess these new data mining techniques 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 Department of Soil Sciences and Agricultural Chemistry, S V Agricultural College, Tirupati, The database contains measurements of soil profile data from various locations of Chandragiri Mandal, Chittoor District. The research establishes whether Soils are Classified Using various data mining techniques. In addition, comparison was made between Naive bayes classification and analyse the most effective technique. The outcome of the research may have many benefits, t o agriculture, soil management and environmental.

***Key words:***
*Naïve bayes, soil profiles, Bayesian Statistics, Soil Database, Classification.*

## 1. Introduction

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. For example, 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" [2]. 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 patterns exist across various soils profiles characterized at various research sites. The data set has been assembled from soil surveys at various agriculture areas located in chandragiri mandal, Chittoor district Andhra Pradesh, 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 soils studies which have been conducted by the Department of Soil Science and Agricultural Chemistry, S V Agricultural College, Tirupati, 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. It is envisaged that the information gained from this research will contribute to the improvement and maintenance of soils and the agricultural environment of Soil Science. The research has a number of potential benefits to the Soil Science. However, the analysis and interpretation of a large data set is problematic. This paper outlines research which may establish if new data mining techniques will improve the effectiveness and accuracy of the Classification of large soil data sets. The classification of such soil data sets is difficult given the complex relationships between large numbers of variables collected for each geographical location. The use of standard statistical analysis techniques is both time consuming and expensive. If alternative techniques can be found to improve this process, an improvement in the classification of soils may result. The overall aim of the research was to classify the soils using Navie Bayes classification technique based on texture of soil profiles

## 2. Classification in Data Mining

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

Manuscript received August 5, 2009
Manuscript revised August 20, 2009

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 section 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. This result was originally proved by Cox [Cox, 1946] and has been reformulated for an AI audience [Heckerman, 1990]. 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. And one must often search large spaces, though most any technique will have to do this and the joint probability provides a good local evaluation function. 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. Any ambiguities in choosing a prior are generally not serious, since the various possible convenient priors usually do not disagree strongly within the regions of interest. Naive Bayes analysis is not limited to what is traditionally considered “statistical” data, but can be applied to any space of models about how the world might be. 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. A natural way to do this is to say that there are a certain number of classes, that a random patient has a certain probability to come from each of them, and that the patients are distributed independently [8] – once we know all about the underlying classes then learning about one patient doesn’t help us learn what any other patient will be like.

**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 five classes of BN classifiers are: Naïve-Bayes, Tree augmented Naïve-Bayes (TANs), Bayesian network augmented Naïve-Bayes (BANs), Bayesian multi-nets and general Bayesian networks (GBNs).Unlike other classifiers the Naïve-Bayes has been used as an effective classifier for many years.

### 2.1 Naive Bayes classifier

**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 articular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be 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 these 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 work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

In spite 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 [7]. 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. However, its naivety may result in poor performance in domains with strong dependencies among attributes. In this paper, 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.

Equations

If a displayed equation needs a number, place it flush with the right margin of the column (e.g., see Eq. 1).

$$y_i(N) = \sum_{n=0}^{m-1} w_n(N) b_n(N) = \sum_{n=0}^{m-1} \overline{b_n^*(N) r_i(N)} \cdot b_n(N) \quad (1)$$

### 3. 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. 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 most common engineering classification system for soils in North America is the Unified Soil Classification System (USCS). 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.

Other engineering soil classification systems in the States include the AASHTO Soil Classification System and the Modified Burmister(See biographical sketch of Prof. Donald M. Burmister).

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 than is provided by the USCS code.

For soil resources, experience has shown that a natural system approach to classification, i.e. grouping soils by their intrinsic property (soil morphology), behavior, or genesis, results in classes that can be interpreted for many diverse uses. Differing concepts of pedogenesis, and differences in the significance of morphological features to various land uses can affect the classification approach. Despite these differences, in a well-constructed system, classification criteria group similar concepts so that interpretations do not vary widely. This is in contrast to a technical system approach to soil classification, where soils are grouped according to their fitness for a specific use and their edaphic characteristics.

Natural system approaches to soil classification, such as the French Soil Reference System (Referential pédologique français) are based on presumed soil genesis. Systems have developed, such as USDA soil taxonomy and the World Reference Base for Soil Resources, which use taxonomic criteria involving soil morphology and laboratory tests to inform and refine hierarchical classes.

Another approach is numerical classification, also called ordination, where soil individuals are grouped by multivariate statistical methods such as cluster analysis. This produces natural groupings without requiring any inference about soil genesis.

In soil survey, as practiced in the United States, soil classification usually means criteria based on soil morphology in addition to characteristics developed during soil formation. Criteria are designed to guide choices in land use and soil management. As indicated, this is a hierarchical system that is a hybrid of both *natural* and objective criteria. USDA soil taxonomy provides the core criteria for differentiating soil map units. This is a substantial revision of the 1938 USDA soil taxonomy which was a strictly natural system. Soil taxonomy based soil map units are additionally sorted into classes based on technical classification systems. Land Capability Classes, hydric soil, and prime farmland are some examples.

In addition to scientific soil classification systems, there are also vernacular soil classification systems. Folk taxonomies have been used for millennia, while scientifically based systems are relatively recent developments[4].

**3.1 The Unified Soil Classification System** (or **USCS**) is a soil classification system used in engineering and geology disciplines to describe the texture and grain size of a soil. The classification system can be applied to most unconsolidated materials[5], and is represented by a two-letter symbol. Each letter is described below

| First and/or second letters | | Second letter | |
|---|---|---|---|
| **Letter** | **Definition** | **Letter** | **Definition** |
| **G** | gravel | **P** | Poorly graded (uniform particle sizes) |
| **S** | sand | **W** | well graded (diversified particle sizes) |
| **M** | silt | **H** | high plasticity |
| **C** | clay | **L** | low plasticity |
| **O** | organic | | |

**Table 1**

**3.2 The Classification of Soils in Chandragiri Mandal: 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, physio-chemical and chemical properties of soils determined were used as criteria for classifying soils. The soils of chandragiri Mandal were classified into different orders, sub-orders, great groups, sub-groups, families and finally into series as per USDA Soil Taxonomy[11]. The texture of the surface of chandragiri mandal varied from sand to silty clay loam where as in sub-surface horizons it varied from sand to clay. According to soil survey Manual [12] the symbols used are shown in table 2.

| | |
|---|---|
| **C** | **Clay** |
| **Cl** | **Clay loam** |
| **L** | **Loam** |
| **S** | **Sand** |
| **Sl** | **Sandy loam** |
| **Scl** | **Sandy clay loam** |
| **Sc** | **Sandy clay** |
| **Ls** | **Loamy sand** |

**Table 2**

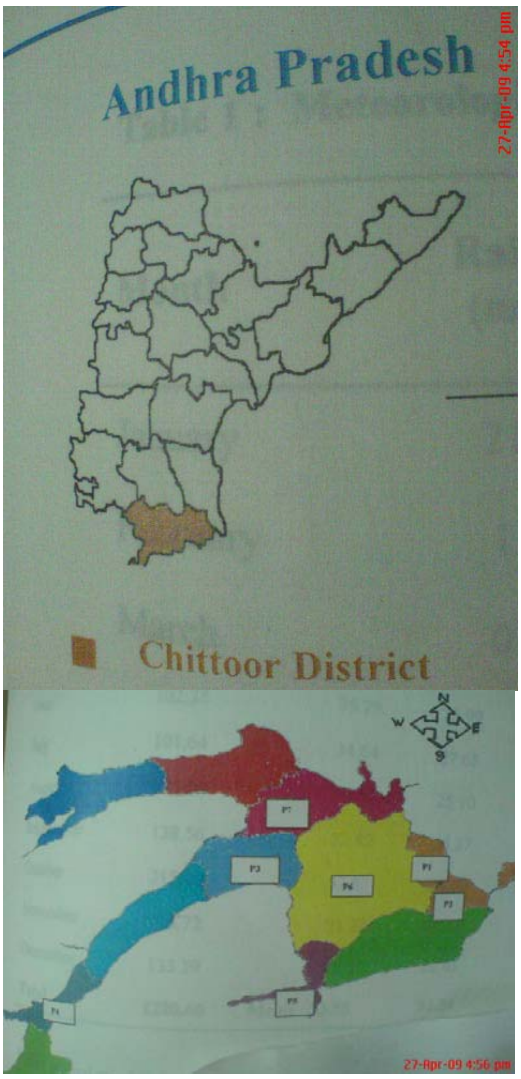


**Figure 1: Characteristic of soils of Chandragiri Mandal,Chittoor District,Andhra Pradesh**

## 4 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:

### 4.1 Data collection cleaning and checking

Relevant data was selected from a subset of the Soil science database.

### 4.2 Data formatting

The data was formatted into an Excel format from the Access database, based on the ten soil types and relevant related fields. The data was then copied into a single Excel spread sheet. The Excel spread sheet (ESS) 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.

### 4.3 Data coding

The soil data set was then converted into a comma delimited (CSV) format file for the ESS. 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[10] to be applied to the soil data set to provide relevant outcomes that were required in the research.

## 5. 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.

### 5.1 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.

### 5.2 Data mining Results

The benchmark having been established, the data classification was then replicated using WEKA data mining software 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 dataset produced in Excel for the statistical processes were copied and then converted to .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 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 in correctly classified instances and also apply the Naive Bayes to test set and see the correctly and in correctly classified 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 Kappa statistic , Mean absolute error, Root mean squared error , Relative absolute error are less than the remaining Classifiers,like Bayesian classifier,J48.

The time to build the Naive bayes Classifier is less than the remaining Classifier. So, The Naive Bayes Classifier is the efficient classification technique among remaining classification techniques. Normalized Expected Cost of Naïve Bayes is more accurate when compared to Bayesian Network.

## 6. Conclusion

The experiments conducted analyzed a small number of traits contained within the dataset to determine their effectiveness when compared with standard statistical techniques. The agriculture soil profiles that were used in this research were selected for completeness and for ease classification of soils.

The recommendations arising from this research are: 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 amounts of data. Data mining techniques, when applied to an agricultural soil profile, may improve the verification of valid soil profile classification.

## References

[1] A Thesis by D.Basavaraju Characterisation and classification of soils in Chandragiri mandal of Chittoor district,Andhra Pradesh
[2] Cunningham, S. J., and Holmes, G. (1999). Developing
innovative applications in agriculture using data mining. In the Proceedings of the Southeast Asia egional Computer Confederation Conference,1999.
[3] Ibrahim, R. S. (1999). Data Mining of Machine Learning Performance Data. Unpublished Master of Applied Science (Information Technology), Publisher; RMITUniversity Press.
[4] Isbell, R. F. (1996). The Australian Soil Classification.Australian soil and land survey handbook. (Vol. 4).Collingwood, Victoria, Australia: CSIRO Publishing.
[5] Mckenzie, N., and Ryan, P. (1999). Spatial prediction of soil properties using environmental correlation.
[6] Geoderma, 89(1-2), 67-94.Palace, B. (1996). Data Mining: What is Data Mining?Retrieved Aug 30, 2005, fromhttp://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.
[7] Ian h.Written and Eibe Frank.Data Mining Practical Machine Learning Tools and Techniques, Second Edition,Elsevier.
[8] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Elsevier.
[9] Remco R. Bouckaert, Bayesian Network Classifiers in Weka, remco@cs.waikato.ac.nz.
[10] WEKA: Data Mining Software in JAVA:http://www.cs.waikato.ac.nz/ml/weka.
[11] Soil Survey Staff 1998, Keys to soil taxonomy. Eight Edition, Natural Resource Conservation Services, USDA, Blacksburg, Virginia.
[12] Soil Survey Staff 1951, Soil Survey Manual. US Department of Agricultural Hand book No. 18.

**P.Bhargavi**, received the M.Sc(Computers) and M.Tech degrees from Sri Krishnadevaraya .University and Vinayaka missions, India in 1997 and 2005, respectively. She is doing her research in data mining. She is working as an associate professor at Madanapalli Institute of Technology & Science (from 2004) in the Dept. of Computer Science Engineering.

**Dr. S.Jyothi** , received the M.Sc(Maths) degree from Sri Venkateswara University. She received the Dr. degree from Sri Venkateswara University, India. She is working as an associate Professor in the department of computer science, school of mathematical and physical sciences, Sri Padmavathi Mahila Viswa Vidyalayam, Tirupati,India.