Data Mining Techniques In Agriculture Prediction Of Soil Fertility

K.Samundeeswari and Dr.K.Srinivasan

Abstract:

Data Mining is emerging research field in Agriculture crop yield analysis.The techniques of data mining are extremely popular in the area of agriculture. In this paper focus on Data Mining techniques in agricultural field. Different Data Mining techniques are in use, such as K-Means,K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for very recentapplications of Data Mining techniques in agriculture field. The productive capacity of a soil depends on soil fertility. Achieving and maintaining appropriate levels of soil fertility, is of utmost importance if agricultural land is to remain capable of nourishing crop production. This paper focus on the problem of predicting soil fertility andSteps for building a predictive model of soil fertility. Soil fertilityis a very important agricultural problem that remains to be solved based on the available data. The problem of soil fertilitycan be solved by employing Data Mining techniques. This paper aims at predicting soil fertility by using Different types of Data Mining techniques.

Keywords : Agriculture,Artificial Neural Networks ,Classification,Data Mining, K-Means, K-Nearest Neighbor, Support Vector Machines,Soil fertility, Yield Prediction.

1. INTRODUCTION

Data Mining is a very crucial research domain in recent research world. The techniques are useful to elicit significant and utilizable knowledge which can be perceived by many individuals. Data mining programs consists of diverse methodologies which are predominantly produced and used by commercial enterprises and biomedical researchers. These techniques are well disposed towards their respective knowledge domain. The use of standard statisticalanalysis techniques is both time consuming and expensive. Efficient techniques can be developed and tailored for solving complex soil data sets using data mining to improve the effectiveness and accuracy of the Classification of large soil data sets [1].

A soil test is the analysis of a soil sample to determine nutrient content, composition and other characteristics. Tests are usually performed to measure fertility and indicate deficiencies that need to be remedied [2]. The soil testing laboratories are provided with suitable technical literature on various aspects of soil testing, including testing methods and formulations of fertilizer recommendations. It helps farmers to decide the extent of fertilizer and farm yard manure to be applied at various stages of the growth cycle of the crop.

A soil test is the analysis of a soil sample to determine nutrient content, composition and other characteristics. Tests are usually performed to measurefertility and indicate deficiencies that need to be remedied[4].Soil fertility is a crucial attribute which is considered for land evaluation, also achieving and maintaining necessary levels of fertility isimportant for nurturing crop production, hence this paper includes steps for building an efficient and accurate predictive model of soil fertility with the help of data mining techniques.The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

2.DATA MINING TECHNIQUES

Data mining techniques are mainly divided in two groups, classification and clustering techniques [8]. Classificationtechniques are designed for classifying unknown samples using information provided by a set of classified samples.This set is usually referred to as a training set as it is used to train the classification technique how to perform itsclassification. Generally, Neural Networks and Support Vector Machines, these two classificationtechniques learn from training set

K.Samundeeswari, Guest Lecturer,
Department of Computer Science, Govt. Arts College for Women, Krishnagiri - 635 001,Tamil Nadu,India
E-mail: samun.arun@gmail.com

Dr.K.Srinivasan, *Assistant Professor & Head, Department of Computer Science, Periyar University Constituent , College of Arts & Science, Pennagaram, Dharmapuri – 636803,Tamil Nadu, E-mail: vasanmsc23@yahoo.co.in*

how to classify unknown samples. Another classification technique, K- Nearest Neighbor [10], does not have any learning phase, because it uses the training set every time a classification must be performed. A training set is known, and it is used to classify samples of unknown classification. The basic assumption in the K-Nearest Neighbor algorithm is that similar samples should have similar classification. The parameter K shows the number of similar known samples used for assigning a classification to an unknown sample. The K-NearestNeighbor uses the information in the training set, but it does not extract any rule for classifying the other.

In this case, clustering techniques can be used to split a set of unknown samples into clusters. One of the most used clustering techniques is the K-Means algorithm [5]. Given a set of data with unknown classification, the aim is to find a partition of the set in which similar data are grouped in the same cluster. The parameter K plays an important role as it specifies the number of clusters in which the data must be partitioned. The idea behind the K-Means algorithm is, given a certain partition of the data in K clusters, the centers of the clusters can be computed as the means of all samples belonging to clusters. The center of the cluster can be considered as the representative of the cluster, because the center is quite close to all samples in the cluster, and therefore it is similar to all of them. There are some is advantages in using K-Means method. One of the disadvantages could be the choice of the parameter K. Another issue that needs attention is the computational cost of the algorithm. There are other Data Mining techniques statistical based techniques, such as Principle Component Analysis (PCA) , Regression Model and Biclustering Techniques [10] have some applications in agriculture or agricultural - related fields.

3.DATA MINING IN AGRICULTURE

Data mining in agriculture is a very current research topic. It consists in the application of data mining techniques to agriculture. Current technologies are nowadays able to provide a lot of information on agricultural-related activities, which can then be an examined in order to find important data. A related, but not equivalent term is precision agriculture.

3.1Sorting apples by watercores

Before going to market, apples are examined and the ones showing some faults are removed. However, there are also invisible faults that can spoil the apple flavor and look. An example of invisible defect is the watercore.

This is an interior apple disorder that can affect the longevity of the fruit. Apples with slight or mild watercores are pleasant, but apples with medium to harsh degree of watercore cannot be stored for any length of time. Moreover, a few fruits with serious watercore could spoil a whole batch of apples. For this reason, a computational system is under study which takes X-raypictures of the fruit while they run on conveyor belts, and which is also able to inspect (by data mining techniques) the taken pictures and evaluate the probability that the fruit contains watercores.[6]

3.2Optimizing pesticide use by data mining

Current studies by agriculture researchers in Pakistan (one of the top four cotton producers of the world) showed that effort of cotton crop yield maximize through pro-pesticide state policies have led to a dangerously high insecticide use. These studies have reported a negative association between insecticide use and crop yield in Pakistan. Hence immoderate use (or abuse) of pesticides is harming the farmers with adverse financial, environmental and social impacts. By data mining the cotton Pest Scouting information along with the meteorological recordings it was shown that how pesticide uses can be optimized (reduced). Clustering of data revealed interesting patterns of farmer practices along with insecticide use dynamics and hence help identify the reasons for this insecticide abuse.

METHODOLOGY	APPLICATIONS
K - means	Forecasts of pollution in atmosphere classifying soil in combination with GPS
K – Nearest Neighbor	Simulating daily precipitations and other weather variable
Support vector Machine	Analysis of different possible change of the weather scenario
Decision Tree Analysis	Predication soil dept
Unsupervised clustering	Generate cluster and determine any existence of pattern
WEKA Tool	Classification system for sorting and grading Mushrooms

3.3Explaining pesticide abuse by data mining

To observe cotton growth, various government departments and agencies in Pakistan have been recording pest scouting, agriculture and metrological information for decades. Bristly estimates of just the cotton pest scouting data recorded stands at around 1.5 million records, and growing. The initial agro-met data recorded has never been digitized, integrated or standardized to give a complete image, and hence cannot support conclusion making, thus requiring an Agriculture Data Warehouse[7]. Creating a novel Pilot Agriculture addition Data Warehouse come behind by analysis through querying and data mining some fascinating discoveries were made, such as pesticides sprayed at the wrong time, wrong pesticides used for the right reasons and secular relationship between pesticide usage and day of the week.

4. APPLICATION OF DATA MINING TECHNIQUES/ALGORITHMS IN AGRICULTURE

There are number of studies which have been carried out on the application of data mining techniques for agricultural data sets. Naive Bayes Data Mining Technique is used to classify soils that analyze large soil profile experimental datasets. [4] Decision tree algorithm in data mining is used for predicting soil fertility. By using clustering techniques (Based on Partitioning Algorithms and Hierarchical algorithms) writer inspect the current usage and details of agriculture land disappeared in the past seven years. The overall aim of the study was to determine the land utilization for agriculture and non agriculture areas for the past ten years.

4.1 Data mining methodologies and its use in Agriculture domain Methodology

4.2**The application of k-means algorithm in the field of agriculture:**

The k-means algorithm is used for soil grouping using GPS-based technologies. Classification of plant, soil, and residue regions of scrutiny by color images, grading apples before marketing, Monitoring water quality changes, Detecting weeds in accuracy agriculture, the prediction of wine fermentation problems can be performed by using a k-means approach. Knowing in promote that the wine fermentation process could get stuck or be slow can help the enologist to correct it and protect a good fermentation process [6].

4.3**The k-nearest neighbor application in the field of agriculture:**

The k-nearest algorithm is used in imitating daily weather conditions and other weather variables and Estimating soil water parameters and Climate prediction.

4.4**The applications of neural networks in the field of agriculture:**

The neural network is used in forecasting of flowering and maturity dates of soybean and in forecasting of water resources variables.

4.5**The applications of SVMs in the field of agriculture:**

The implementation of support vector machine is the crop Classification and in the analysis of the climate change scenarios.

5. RESEARCH METHODOLOGY

5.1Dataset collection

Data set required for this analysis. These datasets contain varied attributes and their many values of soil samples taken from literature review. Dataset has ten attributes and a complete 1988 instances of soil samples. Table one shows attribute description. The dataset has 9 attributes.

Table1 describes data collected for each soil sample.

Table 1 : Attribute 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

5.2Automated System

Soil classification system is essential for theidentification of soil properties. Expert system can bea very powerful tool in identifying soils quickly andaccurately .Traditional classification systems includeuse of tables, flow-charts. This type of manualapproach takes a lot of time, hence quick, reliableautomated system for soil classification is needed tomake better utilization of technician's time [9].We propose an automated system that hasbeen developed for classifying soils based onfertility. Being rule-based system, it depends on facts, concepts, theories which are required for theimplementation of this system. Rules for soilclassification were collected from soil testing lab.

The soil sample instances were classified into thefertility class labels as: Very High, High, ModeratelyHigh, Moderate, Low, and Very Low. These classlabels for soil samples were obtained with the help ofthis system and they have been used further forcomparative study of classification algorithms.

In agriculture,a **soil test** commonly refers to the analysis of a soil sample to determine nutrient content, composition, and other characteristics such as the acidity or pH level. A soil test can determine fertility, or the expected growth potential of the soil which indicates nutrient deficiencies, potential toxicities from excessive fertility and inhibitions from the presence of non-essential trace minerals. The test is used to mimic the function of roots to assimilate minerals. The expected rate of growth is modeled by the Law of the Maximum.

Tap water or chemicals can change the composition of the soil, and may need to be tested separately. As soil nutrients vary with depth and soil components change with time, the depth and timing of a sample may also affect results. Composite sampling can be performed by combining soil from several locations prior to analysis. This is a common procedure, but should be used judiciously to avoid skewing results. This procedure must be done so that government sampling requirements are met. A reference map should be created to record the location and quantity of field samples in order to properly interpret test results.

5.3 Storage, handling, and moving:

Soil chemistry changes over time, as biological and chemical processes break down or combine compounds over time. These processes change once the soil is removed from its natural ecosystem (flora and fauna that penetrate the sampled area) and environment (temperature, moisture, and solar light/radiation cycles). As a result, the chemical composition analysis accuracy can be improved if the soil is analyzed soon after its extraction — usually within a relative time period of 24 hours. The chemical changes in the soil can be slowed during storage and transportation by freezing it. Air drying can also preserve the soil sample for many months.

5.4 Soil testing

Soil testing is often performed by commercial labs that offer a variety of tests, targeting groups of compounds and minerals. The advantage associated with local lab is that they are familiar with the chemistry of the soil in the area where the sample was taken. This enables technicians to recommend the tests that are most likely to reveal useful information.

Laboratory tests often check for plant nutrients in three categories:

- Major nutrients: nitrogen (N), phosphorus (P), and potassium (K)
- Secondary nutrients: sulfur, calcium, magnesium
- Minor nutrients: iron, manganese, copper, zinc, boron, molybdenum, chlorine

Do-it-yourself kits usually only test for the three "major nutrients", and for soil acidity or pH level. Do-it-yourself kits are often sold at farming cooperatives, university labs, private labs, and some hardware and gardening stores. Electrical meters that measure pH, water content, and sometimes nutrient content of the soil are also available at many hardware stores. Laboratory tests are more accurate than tests with do-it-yourself kits and electrical meters[8].

Soil testing is used to facilitate fertilizer composition and dosage selection for land employed in both agricultural and horticultural industries.Prepaid mail-in kits for soil and ground water testing are available to facilitate the packaging and delivery of samples to a laboratory. Similarly, in 2004, laboratories began providing fertilizer recommendations along with the soil composition report.

Lab tests are more accurate, though both types are useful. In addition, lab tests frequently include professional interpretation of results and recommendations. Always refer to all proviso statements included in a lab report as they may outline any anomalies, exceptions, and shortcomings in the sampling and/or analytical process/results.Some laboratories analyze for all 13 mineral nutrients and a dozen non-essential, potentially toxic minerals utilizing the "universal soil extractant" (ammonium bicarbonateDTPA)

5.5Soil contaminants:

Common mineral soil contaminants include arsenic, barium, cadmium, copper, mercury, lead, and zinc.Lead is a particularly dangerous soil component.

6. SOIL CLASSIFICATION

Soil classification deals with the systematiccategorization of soils based on distinguishing Characteristics as well as criteria that dictate choices inuse. Soil classification is a dynamic subject, from thestructure of the system itself, to the definitions of classes and finally in the application in the field. Soilclassification can be approached from the perspective ofsoil as a material and soil as a resource.The most common engineering classificationsystem for soils is the Unified Soil Classification System(USCS) [6].

The USCS has three major classificationgroups:
(1) Coarse-grained soils (e.g. sands and gravels);
(2) Fine-grained soils (e.g. silts and clays);
(3) Highly

Organic soils (referred to as "peat"). The USCS furthersubdivides the three major soil classes for clarification. Afull geotechnical engineering soil description will alsoinclude other properties of the soil including color, in-situmoisture content, in-situ strength, and somewhat moredetail about the material properties of the soil that isprovided by the USCS code.

The soils are classified into different orders, suborders,great groups, sub-groups, families and finally intoseries as per USDA Soil Taxonomy as in [8]. The solidphase of soil can be divided into mineral matter andorganic matter. The mineral particles can be furthersubdivided into classes based on size. The classification of soil particles according to size are Sand, Silt, Clay. Theproposition of Sand, Silt, and Clay present in soil determinesits texture.

6.1Soil Data

The soil data used in this paper consists of 111 instanceswith 8 attributes like (i.e., Depth, Sand, Silt, Clay,Sandbysilt, Sandbyclay, Sandbysiltclay, TextureClass). Thetexture of the Soil data is varied from sand to silty clayloam where as in sub-surface horizons it varied from sandto clay as in [2]. Table2.Shows the different soil attribute.

Table 2. Soil Attribute

SYMBOL	DESCRIPTION
S	**Sand**
Sicl	**SiltyClay 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**

7. A COMPARATIVE STUDY OF SOILCLASSIFICATION

The classification of soil was considered criticalto study because depending upon the fertility class ofthe soil the domain knowledge expert's determineswhich crops should be taken on that particular soiland which fertilizers should be used for the same.The following section describes Naive Bayes, J48,JRip algorithms briefly.

7.1Naive Bayes

A naive Bayes classifier is a simpleprobabilistic classifier based on applying Bayes'theorem with strong (naive) independenceassumptions. Depending on the precise nature of theprobability model, naive Bayes classifiers can betrained very efficiently in a supervised learningsetting. An advantage of the naive Bayes classifier isthat it only requires a small amount of training data toestimate the parameters (means and variances of thevariables) necessary for classification [5].

7.2 J48 (C4.5)

J48 is an open source Java implementationof the C4.5 algorithm in the Weka data mining tool.C4.5 is a program that creates a decision tree basedon a set of labeled input data. This decision tree canthen be tested against unseen labeled test data toquantify how well it generalizes. This algorithm wasdeveloped by Ross Quinlan. It is an extension ofQuinlan's earlier ID3 algorithm. C4.5 uses ID3algorithm that accounts for continuous attribute valueranges, pruning of decision trees, rule derivation, andso on.The decision trees generated by C4.5 can beused for classification, and for this reason, C4.5 isoften referred to as a statistical classifier [6].

7.3 JRip

This algorithm implements a propositionalrule learner, Repeated Incremental Pruning toProduce Error Reduction (RIPPER), which wasproposed by William W. Cohen as an optimizediversion of IREP.In this paper, three classification techniques(naïve Bayes, J48 (C4.5) and JRip) in data miningwere evaluated and compared on basis of time,accuracy, Error Rate, True Positive Rate and FalsePositive Rate. Tenfold cross-validation was used inthe experiment. Our studies showed that J48 (C4.5)model turned out to be the best classifier for soil samples.

Table-3 Comparison of different classifiers

Classifier	Naïve	Bayes	JRip J48
Correctly Classified Instances	855	1998	2065
Incorrectly Classified Instances	1345	202	135
Accuracy	38.86%	90.81%	93.86%
Mean Absolute Error	0.324	0.0313	0.0283

8. CONCLUSION

Agriculture is the most significant application area particularly in the developing countries like India. Use of information technology in agriculture can hang the scenario of decision making and farmers can yield in better way. For decision making on overall issues related to agriculture field; data mining plays a vital role. In this paper we have discussed about the role of data mining in outlook of agriculture field. We have also discussed several data mining techniques,application of datamining in agriculture and soil containments.

In this paper, we've got suggested an analysis of the soil information using completely different algorithmsand prediction technique. In this paper we have demonstrated acomparative study of varied classification algorithms i.e. Naïve bayes, J48 (C4.5), JRip with the assistance ofdata mining tool .J48 is incredibly easy classifier to form a decision tree.We have demonstrated a comparative study of various classification algorithms i.e. Naïve Bayes, J48 (C4.5), JRip with the help of data mining tool WEKA. J48 is very simple classifier to make a decision tree. In future, we canplan to build Fertilizer Recommendation

System which can be utilized effectively by the Soil Testing Laboratories. This System will recommend appropriate fertilizer for the given soil sample and cropping pattern.

9.REFERENCES

[1]. S.Baskar ,L.Arockiam ,S.Charles,"Applying Data Mining Techniques on Soil Fertility Prediction",International Journal of Computer Applications Technology and Research Volume 2– Issue 6.

[2]. P. Bhargavi1 , Dr. S. Jyothi."Soil Classification Using Data Mining Techniques: A Comparative Study",International Journal of Engineering Trends and Technology- July to Aug Issue 2011.

[3]. Dr.S.HariGanesh , Mrs. Jayasudha,"An Enhanced Technique to Predict the Accuracy of Soil Fertility in Agricultural ining",International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, July 2015.

[4]. Dr. S.Hari Ganesh,Mrs. Jayasudha,"Data Mining Technique to Predict the Accuracy of the Soil Fertility",Dr. S.Hari Ganesh et al, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.7, July- 2015, pg. 330-333.

[5]. 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.

[6]. N.Neelaveni,Ms. S. Rajeswari,"Data Mining in Agriculture- A Survey", International Journal of Modern Computer Science (IJMCS),Volume 4, Issue 4, August, 2016.

[7].V. Rajeswari* and K. Arunesh "Analysing Soil Data using Data Mining Classification Techniques",Indian Journal of Science and Technology, Vol 9(19), DOI: 10.17485/ijst/2016/v9i19/93873, May 2016.

[8].D Ramesh , B Vishnu Vardhan "Data Mining Techniques and Applications to Agricultural Yield Data",International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.

[9].D Ramesh , B Vishnu Vardhan,"Analysis of Crop Yield Prediction Using Data Mining Techniques",IJRET: International Journal of Research in Engineering and Technology.

[10].VelidePhanikumar*and Lakshmi Velide**,"Data mining plays a key role in soil data analysis of Warangal region",International Journal of Scientific and Research Publications, Volume 4, Issue 3, March 2014.