IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Volume-6, Issue-1 E-ISSN: 2347-2693

Survey on Classification Techniques for Soil Data Prediction to Better Yielding of Crops

S.Manimekalai[1*], K.Nandhini[2]

[1*]Department of Computer Science, Chikkanna Government Arts College Tirupur, Tamilnadu, India
[2]Department of Computer Science, Chikkanna Government Arts College Tirupur, Tamilnadu, India

*Corresponding Author: manimekalaipari@gmail.com

Available online at: www.ijcseonline.org

Received: 06/Jan/2018, Revised: 10/Jan/2018, Accepted: 25/Jan/2018, Published: 31/Jan/2018

Abstract- Yield prediction is a significant contribution for agriculture data mining to the proper choice of crops for sowing. This makes the difficulty of predicting the yielding of crops a remarkable challenge. Earlier yield prediction was performed by considering the farmer's experience on a selected field and crop. The main thing of the crop yielding is soil. This work presents the use of classification techniques to predict the soil datasets. The predicted results will express the yielding of crops. The issue of predicting the soil data is recognized as data mining technique. The soil is classified by using these techniques Naive Bayes, Decision Tree, fuzzy and neural network are used. The set of rules JRip is applied and validated on this paper using weka tool.

Keywords: Data mining, Fuzzy, neural network, decision tree, soil dataset.

I. INTRODUCTION

Indian economy is highly depending on agriculture. Agriculture is the main source of income for most of the population. So farmers are regularly interested about yield prediction. Many factors are important like soil, weather, rain, fertilizers and pesticides are used to increase the crop production. In agriculture field Data mining plays a main role in crop yielding. There is a need to transform the large data into technologies and make them available to the farmers. It is can be very useful for farmers to take efficient and effective decision. Soil is one of the parameter which is used to increase crop production is considered.

Data mining is the process to find interesting knowledge from large amounts of data [1]. The aim of the data mining process is to extract knowledge from an existing data set and transform it into a human understandable formation for advance use. It is the process of analyzing data from different view and encapsulates it into useful information. There is no restriction to the type of data that can be analyzed by data mining. It analyzes data hold in a relational database, a data warehouse, a web server log or a simple text file. Analysis of data in successful way requires understanding of appropriate techniques of data mining. This paper is to gives the details about different data mining techniques in view of agriculture domain for soil classification. Generally Data mining tasks can be Classified into two categories: Descriptive data mining and Predictive data mining.

Descriptive data mining is to identify the general properties of the data in the database. Predictive data mining is used to predict explicit values based on patterns decided from known results. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Predictive data mining approach is used to predict future crop, weather forecasting, pesticides and fertilizers to be used, revenue to be generated and so forth with appreciate to agricultural data mining.

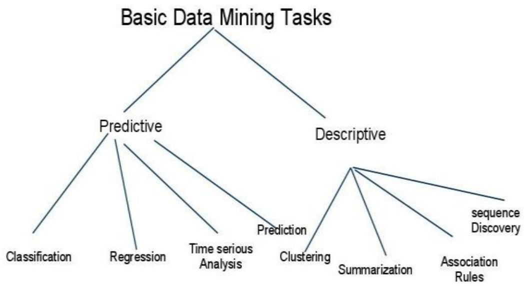


Figure 1.1: Data mining Techniques

The data mining techniques include Classification, Clustering, Association rules and Regression. The different data mining techniques are used for solving different agricultural problem. The different data mining techniques are shown in Figure 1.1.

1.1 *Soil classification*

Plantation of crop according to the soil type is very important in the success of crop. It influences many other properties and importance of land use and management. The Soil texture is main attribute for agriculture soil classification. It effects fertility, aeration, water holding capacity, drainage, tillage, and strength of soils [2]. There are many characteristics that decide the nature of the soil ex: pH (power of Hydrogen) value, moisture, Ec (Electrical Conductivity), ESP (Exchangeable Sodium Percentage), etc The pH value is used to decide whether the soil is acidic, the Ec value is used to decide whether the soil is saline (salt content) and the ESP value is used to decide if the soil is alkali.

Field	***Description***
Ph	Ph value of soil
Ec	Electrical conductivity
Oc	Organic carbon
Zn	Zinc
Cu	Copper
Fe	Iron
P	Phosphorous
Mn	Manganese
K	Potassium

Table 1.1.1: Soil Attributes

The soil samples that belong to either acidic or saline or alkali are supposed to be problematic soils as they are not conducive for crop growth. The other soil samples are non-problematic as they are conducive for crop growth. To analyze the soil type of a geographical area, soil samples are collected then the samples are classified into different types [6]. Using data mining techniques one can efficiently classify the soil samples into different categories. Different types of classifications algorithms are used to soil dataset to predict its fertility.

The paper is organized as follows. Literature of the review is described in section II. Research methodology is discussed in section III. Results and Discussion is mentioned in section IV. Finally the conclusion in section V.

II. REVIEW LITERATURE

The author V.Bhuyar et al represented the classification of soil fertility rate using J48, Naïve Bayes, and Random forest algorithm in the paper [1]. The author concludes that J48 algorithm gives better result than other algorithms. The J48 set of rules facilitates the farmer and decision makers to perceive the soil fertility rate and at the assist of nutrients observed in the soil sample exclusive fertilizers may be recommended [2].

The author B. Murugesakumar et al integrates the work of different authors in one place. That is useful for researchers to gain information of current framework of data mining techniques and applications to classify soil fertility [6].

Author R.Ramesh Vamanan et al aimed to evaluate the various classification techniques of data mining and apply them to a soil science database to establish if meaningful relationships can be found. The application of data mining techniques has been organised for Tamil Nadu soil data sets. This paper compares the different classifiers and the outcome of the 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 [7].

The Author V.Rajeswari et al Narrate the comparative analysis of three algorithms like Naïve Bayes, JRip and J48 is projected. Finally the author concludes that JRip classification algorithm gives better result of this dataset and is correctly classified into maximum number of instances comparing with the other two. JRip can be recommended to predict soil types [10].

The author Sofianita et al demonstrated the application of SOM in soil classification to identify the type of soil. The results are compared with *k*-means algorithm. The proposed algorithm has shown its ability in classifying the soil with a 91.8% of accuracy [11].

III. RESEARCH METHODOLOGY

Classification of soil is very challenging to study because depending upon the fertility category of the soil, the field experts can take a decision which type of crops to be cultivated in a particular soil and also determine the type of fertilizers to be used for the same. Popular classification techniques include decision trees, neural networks, KNN, and Naïve Bayesian classifier, Fuzzy classification etc.

Naive Bayes

A Naive Bayes classifier is a easy probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. Depending on the best nature of the probability model, naive Bayes classifiers may be skilled very efficiently in a supervised learning setting. A bonus of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters essential for classification.

Fuzzy Classification

Fuzzy classification rules are widely taken into consideration as a nicely applicable illustration of classification knowledge with uncertainty and they permit readable and interpretable rule bases. The maximum vital task inside the design of fuzzy classification systems is to discover a set of fuzzy rules from training data with uncertainty to address a specific classification problem.

Fuzzy logic works with membership values in a way that mimics Boolean logic.

To this end, replacements for basic operators AND, OR, NOT must be available. A usual substitution is called the *Zadeh operators*:

Boolean	**Fuzzy**
AND(x,y)	MIN(x,y)
OR(x,y)	MAX(x,y)
NOT(x)	1 – x

Table 2.2 Basic operators

For true/1 and false/0, the fuzzy expressions produce the equal end result because Boolean expressions.

Artificial Neural Networks

(ANN) is systems inspired by the research on human brain. Artificial Neural Networks (ANN) networks in which each node represents a neuron and each link represents the way two neurons interact. Each neuron performs very simple tasks, while the network representing of the work of all its neurons is able to perform the more complex task. Artificial neural network is a new techniques used in flood forecast. The advantage of ANN approach in modelling the rain fall and run off relationship over the conventional techniques flood forecast [9]. Neural network has several advantages over conventional method in computing [8].

Neural network models can be viewed as simple mathematical models defining a function

f: X → Y or a distribution over X or X and Y. Sometimes models are intimately associated with a particular learning rule. A common use of the word "ANN model" is honesty the definition of a *class* of such capabilities (in which members of the class are received via varying parameters, connection weights, or specifics of the architecture including the variety of neurons or their connectivity).

Genetic Algorithm

Genetic algorithm are generally used to generate excellent solutions to optimization and

Search troubles by using relying on bio-stimulated operators inclusive of mutation, crossover and selection. Genetic algorithm calls for:

1. A genetic representation of the solution area.
2. A fitness feature to assess the solution area.

Classification by Decision Tree

Decision tree induction is the learning of decision trees from class-labeled training tuples. Decision tree algorithms are as follows.

J48 (C4.5)

J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is a program that creates a decision tree primarily based on a set of categorized input data. This decision tree can then be tested in opposition to unseen test data to quantify how well it generalizes. This algorithm was developed by Ross Quinlan. It is an extension of Quinlan's earlier ID3 algorithm. C4.5 uses ID3 algorithm that accounts for continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. The decision trees generated by using C4.5 may be used for classification, and because of this, C4.5 is often called as a statistical classifier.

JRip

This algorithm implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which changed into proposed by William W. Cohen as an optimized model of IREP. The algorithm is briefly defined as follows:

Initialize RS = {}, and for each class from the much less prevalent one to the greater common one, DO:

1. Building stage:

Repeat 1.1 and 1.2 until the description length (DL) of the ruleset and examples is 64 bits greater than the smallest DL met to date, or there are no positive examples, or the error rate >= 50%.

1.1. Grow phase:

Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: p(log(p/t)-log(P/T)).

1.2. Prune phase:

Incrementally prune each rule and allow the pruning of any final sequences of the antecedents; The pruning metric is (p-n)/(p+n) -- but it's actually 2p/(p+n) -1, so in this implementation we simply use p/(p+n) (actually (p+1)/(p+n+2), thus if p+n is 0, it's 0.5).

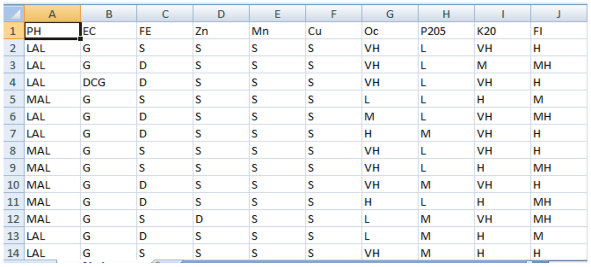
2. Optimization stage:

after generating the initial ruleset {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/(P+N).Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the ruleset. After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete the rules from the ruleset that would increase the DL of the whole ruleset if it were in it. And add resultant ruleset to RS.

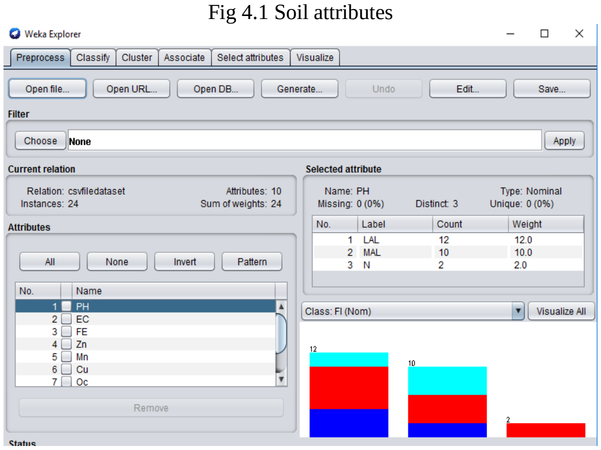
IV. RESULTS AND DISCUSSION

In this work, the data is collected from the sample agricultural soil dataset. There are taken 24 instances data which contains the attributes such as PH, EC,FE,ZN,MN,CU,OC,P205,K20,FI. This data set organized in Excel Sheet which has been saves as type is CSV extension it shown in Fig 5.1.



	A	B	C	D	E	F	G	H	I	J
1	PH	EC	FE	Zn	Mn	Cu	Oc	P205	K20	FI
2	LAL	G	S	S	S	S	VH	L	VH	H
3	LAL	G	D	S	S	S	VH	L	M	MH
4	LAL	DCG	D	S	S	S	VH	L	VH	H
5	MAL	G	S	S	S	S	L	L	H	M
6	LAL	G	D	S	S	S	M	L	VH	MH
7	LAL	G	D	S	S	S	H	M	VH	H
8	MAL	G	S	S	S	S	VH	L	VH	H
9	MAL	G	S	S	S	S	VH	L	H	MH
10	MAL	G	D	S	S	S	VH	M	VH	H
11	MAL	G	D	S	S	S	H	L	H	MH
12	MAL	G	S	D	S	S	L	M	VH	MH
13	LAL	G	D	S	S	S	L	M	H	M
14	LAL	G	S	S	S	S	VH	M	H	H

Fig 4.1 Soil attributes



None

Current relation

Relation: csvfiledataset Attributes: 10
Instances: 24 Sum of weights: 24

Selected attribute

Name: PH Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	LAL	12	12.0
2	MAL	10	10.0
3	N	2	2.0

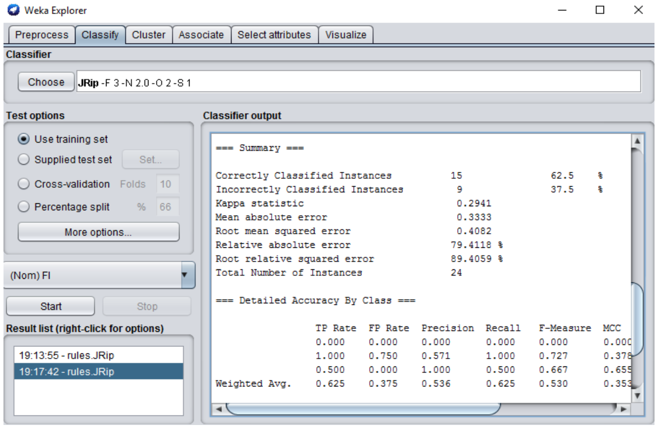
Attributes

No.		Name
1	□	PH
2	□	EC
3	□	FE
4	□	Zn
5	□	Mn
6	□	Cu
7	□	Oc

Class: FI (Nom) Visualize All

Fig 4.3 Soil attributes in weka

This work is predicted the soil classification using JRip algorithm. This paper is implemented by using weka tool. In weka tool first open dataset file it shown in fig 5.2 and choose JRip algorithm then get the result. It is based on the training data set it concludes that weighted average of True Positive Rate of JRip classifier is 0.625, False Positive Rate is 0.375 it shown in Fig 5.2.



JRip -F 3 -N 2.0 -O 2 -S 1

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) FI

19:13:55 - rules.JRip
19:17:42 - rules.JRip

```
=== Summary ===

Correctly Classified Instances          15               62.5    %
Incorrectly Classified Instances         9               37.5    %
Kappa statistic                          0.2941
Mean absolute error                      0.3333
Root mean squared error                  0.4082
Relative absolute error                 79.4118 %
Root relative squared error             89.4059 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.000    0.000    0.000      0.000    0.000      0.000
                 1.000    0.750    0.571      1.000    0.727      0.378
                 0.500    0.000    1.000      0.500    0.667      0.655
Weighted Avg.    0.625    0.375    0.536      0.625    0.530      0.353
```

Fig 4.2 JRip classifier result

The numbers of correctly classified instances are 62.5% and incorrectly classified instances are 37.5%. JRip algorithm can be recommended to predict soil types in agriculture data mining.

V. CONCLUSION

Soil prediction plays an important role in crop yielding in agriculture. This paper focalized on soil type. The numerous techniques and its strategies are used within the surveyed papers. The techniques like Naïve Bayes, Genetic algorithm, neural network, and J48 are basically used for higher classifications. The set of rules JRip is applied and validated on this paper using weka tool. It is located that 67 % of correctly classified instances of soil utilized in crop yielding. All those soil parameter values are taken into consideration earlier than cropping.

REFERENCES

[1]. Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann, 2006.

[2]. Bhuyar V. Comparative analysis of classification techniques on soil data to predict fertility rate for Aurangabad District. International Journal of Emerging Trends and Technology in Computer Science. 2014 Mar-Apr; 3(2):200–3.

[3]. Beniwal S., Arora J., (2012). Classification and Feature Selection techniques in data mining. International Journal of Engineering Research and Technology (IJERT).

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

[5]. N. Hemageetha., G.M. Nasira., Classification of Soil type in Salem District Using J48 Algorithm. IJCTA, 9(40), 2016.

[6]. B. Murugesakuma., Dr. K.Anandakumar., Dr. A.Bharathi., "Survey on Soil Classification Methods Using Data Mining Techniques". International Journal of Current Trends in Engineering & Research (IJCTER) e-ISSN 2455–1392 Volume 2 Issue 7, July 2016.

[7]. R. Vamanan & K. Ramar, (2011), "Classification of Agricultural Land Soils A Data Mining Approach", International Journal on Computer Science and Engineering, ISSN: 0975-3397, Vol. 3.

[8]. AR. PonPeriasamy, E. Thenmozhi., "A Brief survey of Data Mining Techniques Applied to Agricultural Data" International Journal of Computer Sciences and Engineering Volume-5, Issue-4 E-ISSN: 2347-2693

[9]. Ramesh Babu Palepu., Rajesh Reddy Muley. "An Analysis of Agricultural Soils by using Data Mining Techniques". International Journal of Engineering Science Computing 2017.

[10]. Veenadhari S, Misra B, Singh CD. Data mining techniques for predicting crop productivity—A review article. In: IJCST. 2011; 2(1).

[11]. V.Rajeswari. K.Arunesh., "Analysing Soil Data Mining Classification Techniques". Indian Journal of Science and Technology, Vol 9(19), May 2016.

[12]. Sofianita., Jamian., "Soil Classification: An application of Self Organising Map and K-Means" 978-1-4244-8136-1/10/$26.00_c 2010 IEEE