ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(19), DOI: 10.17485/ijst/2016/v9i19/93873, May 2016* ISSN (Online) : 0974-5645

Analysing Soil Data using Data Mining Classification Techniques

V. Rajeswari* and K. Arunesh

Department of Computer Science, Sri SRNM College, Sattur - 626203, Virudhunagar Dist., Tamil Nadu, India; vramanimca@gmail.com, arunesh_naga@yahoo.com

Abstract

Background/Objectives: Soil is an essential key factor of agriculture. The objective of the work is to predict soil type using data mining classification techniques. **Methods/Analysis:** Soil type is predicted using data mining classification techniques such as JRip, J48 and Naive Bayes. These classifier algorithms are applied to extract the knowledge from soil data and two types of soil are considered such as Red and Black. **Findings:** In this paper, Data Mining and agricultural Data Mining are summarized. The JRip model can produce more reliable results of this data and the Kappa Statistics in the forecast were increased. **Application/Improvement:** For solving the issues in Big Data, efficient methods can be created that utilize Data Mining to enhance the exactness of classification of huge soil data sets.

Keywords: Data Mining, Naive Bayes, J48, JRip, Soil Dataset

1. Introduction

Data Mining (DM) becomes popular in the field of agriculture for soil classification, wasteland management and crop and pest management. In[1] assessed the variety of association techniques in DM and applied into the database of soil science to predict the meaningful relationships and provided association rules for different soil types in agriculture. Similarly, agriculture prediction, disease detection and optimizing the pesticides are analyzed with the use of various data mining techniques earlier[2]. In[3] analyzed J48 classification algorithm in high accuracy for predict the soil fertility rate. In[4] investigated the uses of various DM techniques for knowledge discovery in agriculture sector and introduced different exhibits for knowledge discovery in the form of Association Rules, Clustering, Classification and Correlation. In[5] predicted the soil fertility classes using with classification techniques were Naïve Bayes, J48 and K-Nearest Neighbor algorithms. In[6] used Adopted data mining techniques to estimate crop yield analysis. Multiple Linear Regression (MLR) method was used to find the linear relationship between dependent and independent variables. K-Means clustering approach was also use to form four clusters considering Rainfall as key parameter. In[7] analyzed the vegetative factors of landslides in the Shimen reservoir watershed in northern Taiwan. Decision tree, Bayesian Network data mining techniques and the non-linear approaches were implemented. Optimization based Bayesian Network approach was considered as better than non-linear. In[8] analyzed the virtual significance of soil fertility and the crop management factors to predict the maize yields and in determining the yield variability and the gap between farmers. Classification and regression tree analysis was used to predict the result. In[9] investigated two comprehensive methods to calculate the production related yield gap and a soil fertility related nutrient balance. The methodology allows knowledge from micro-scale to higher-scale levels and determines land quality. In[10] predicted soil attributes and analyzed soil data using classification techniques. Soil properties such as pH value, Electrical Conductivity (EC), Potassium, Iron, Copper, etc. were classified using classification algorithms like Naïve Bayes, J48 and JRip. Among the algorithms, J48 was considered as simple classifier and produced better result.

* *Author for correspondence*

2. Agricultural Data Mining

Data Mining is essential to discover the agricultural related knowledge such as soil fertility, yield prediction and soil erosion. Soil prediction helps to for soil remedy and crop management. Classification algorithms involve finding rules that partition the data into disjoint groups. A set of classification rules are generated by such a classification process, which can be used to classify future data[11].

Following section give explanation of classification algorithms such as Naive Bayesian classifier, J48 decision tree classifier and JRip classifier.

2.1 Naive Bayes

A Naive Bayes classifier is one of the classifiers in a family of simple probabilistic classification techniques in machine learning. It is based on the Bayes theorem with independence features. Each class labels are estimated through probability of given instance. It needs only small amount of training data to predict class label necessary for classification[12].

2.2 J48 (C4.5)

The J48 is one of the classification-decision tree algorithm and it slightly modified from C4.5 in Weka. It can select the test as best information gain. This algorithm was proposed by Ross Quinlan. C4.5 is also referred to as a statistical classifier. J48 predicts dependent variable from available data. It builds tree based on attributes values of training data. This classifies data with the help of feature of data instances that said to have information gain. The importance of error tolerance is developed using pruning concept[13,14].

2.3 JRip

IREP optimized version is Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen. This algorithm is a propositional guideline learner. J-Rip classifier is one of the decision tree pruning models based on association rules. It is an effective technique to reduce error pruning. In this algorithm, the training data is split into two sets and with the help of pruning operators the error is reduced on both the sets. Finally rules are formed from two sets such as Growing set and Pruning set.

3. Results and Discussion

In this work, we collected the agricultural soil dataset from the soil testing lab., Virudhunagar District. We have taken 110 data which contains the attributes such as Village Name, Soil Type or Color, Soil Texture, PH, EC (Electrical Conductivity), Lime Status, Phosphorous. This system predicted the soil type Red and Black based on the PH and EC value. The PH value of Black soil discovered as greater than 7.7 and Red soil found as less than 7.7. We took three classification algorithms such as JRip, J48, Naive Bayes to predict the soil type Red and Black. While applying three classifier algorithms, JRip considers the entire attribute. But, J48 classifier considers only PH and EC value. Tree is build based on above two attributes. JRip classifier generates the rules efficiently and shows good performance for this soil data set. As comparing these three algorithms JRip resulted in high accuracy. Here, full dataset considered as training set.

Based on the training data set it is concluded that weighted average of True Positive Rate of JRip classifier is 0.982. In the case J48 and Naïve Bayes TP Rate is 0.97 and 0.86 it indicates the low level. So, automatically JRip classifier classified the data set in higher sense. Soil properties differed among sites with Red textured soils and Black textured soils. It since that below 7.0 is acid soil and above 7.0 is alkaline soil. The spectral analysis was sufficiently sensitive to capture the variation in soil fertility between the different soil natures.The soil dataset which contains the attributes like soil type, pH value, etc. are given in Figure 1. This data set organized in Excel Sheet with saves as type is CSV extension.

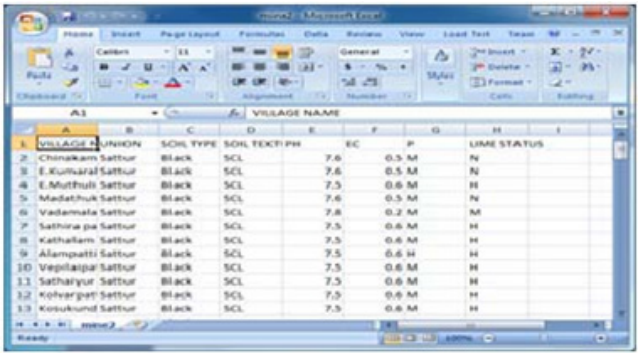

Figure 1. Soil data set.

The number of incorrectly classified instances, error rate of JRip is given in Figure 2.

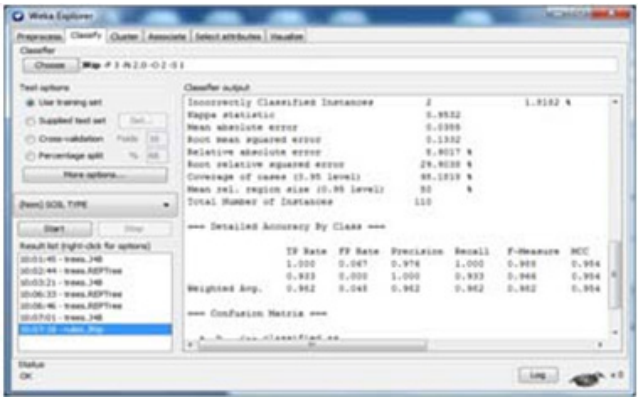

Figure 2. JRip classifier result.

The number of correctly classified instances and incorrectly classified instances are given in Figure 3. Here, JRip classified maximum number of instances.

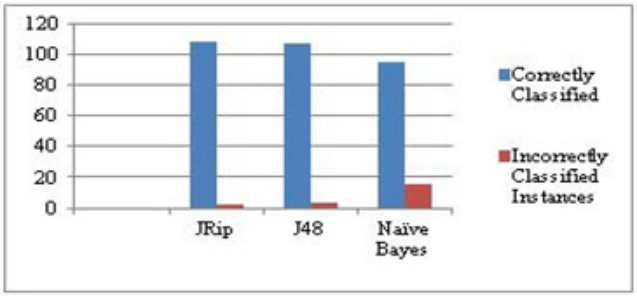

Figure 3. Classifiers error rate.

The comparative analysis of classifiers is given in Table 1. Here JRip performed better classification to compare the other algorithms and also Kappa Statistic value becomes nearest 1.00 in JRip algorithm.

Table 1. Comparative analysis of classifiers

Evaluation Criteria (Total number of instances 110)	Correctly Classified Instances	Incorrectly Classified Instances	Prediction Accuracy	Kappa Statistic
JRip	108	2	98.18%	0.9532
J48	107	3	97.27%	0.9305
NaïveBayes	95	15	86.36	0.5926

The JRip algorithm gives the high prediction accuracy is given in Figure 4. The Naive Bayes Algorithm has less accuracy compared than J48 and JRip.

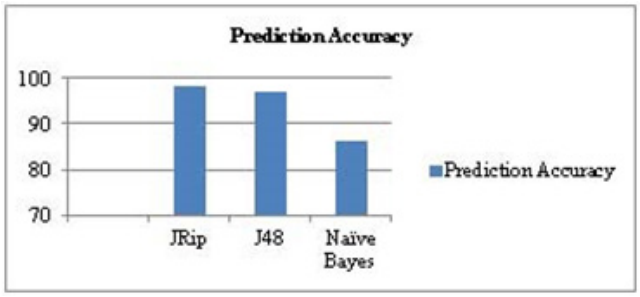

Figure 4. Prediction accuracy for classifiers.

4. Conclusion and Future Work

In this paper, the comparative analysis of three algorithms like Naïve Bayes, JRip and J48 is projected. 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.

5. References

1. Geetha MCS. Implementation of association rule mining for different soil types in agriculture. International Journal of Advanced Research in Computer and Communication Engineering. 2015 Apr; 4(4):520–2.
2. Solanki J, Mulge Y. Different techniques used in data mining in agriculture. International Journal of Advanced Research in Computer Science and Software Engineering. 2015 May; 5(5):1223–7.
3. 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.
4. Fathima NG, Geetha R. Agriculture crop pattern using data mining techniques. International Journal of Advanced Research in Computer Science and Software Engineering. 2014 May; 4(5):781–6.
5. Suman, Naib BB. Soil classification and fertilizer recommendation using WEKA. International Journal of Computer Science and Management Studies. 2013 Jul; 13(5):142–6.
6. Ramesh D, Vardhan VB. Data mining techniques and applications to agricultural yield data. International Journal of Advanced Research in Computer and Communication Engineering. 2013 Sep; 2(9):3477–80.
7. Tsai F, Lai JS, Chen WW, Lin TH. Analysis of topographic and vegetative factors with data mining for landslide verification. Ecological Engineering. 2013 Dec; 61:669–77.
8. Tittonell P, Shephered KD, Vanlauwe B, Giller KE. Unravel-

ling the effects of soil and crop management on maize productivity in small holder agricultural systems of Western Kenya - An application of classification and regression tree analysis. Agriculture, Ecosystems and Environment. 2008 Jan; 123(1-3):137–50.
9. Bindraban PS, Stroorvofel JJ, Jansen DM, Vlaming J, Groot JJR. Land quality indicators for suitable land management: Proposed methods for yield gap and soil nutrient balance. Agriculture, Ecosystems and Environment. 2000; 81:103–12.
10. Gholap J, Lngole A, Gohil J, Shailesh, Attar V. Soil data analysis using classification techniques and soil attribute prediction. 2012 Jun; 9(3):1–4.
11. Anuradha C, Velmurugan T. A comparative analysis on the evaluation of classification algorithms in the prediction of student performance. Indian Journal of Science and Technology. 2015 Jul; 8(15):1–12.
12. Narain B. Study for Data Mining techniques in classification of agricultural land soils. Journal of Advanced Research in Computer Engineering. 2011 Jan-Jun; 5(1):35–7.
13. Venkatesan E, Velmurugan T. Performance analysis of decision tree algorithms for breast cancer classification. Indian Journal of Science and Technology. 2015 Nov; 8(29):1–8.
14. Chandrakar PK, Kumar S, Mukherjee D, Applying classification techniques in Data Mining in agricultural land soil. International Journal of Computer Engineering. 2011 Jul-Dec; (2):89–95.