APPLICATION OF BAYESIAN APPROACH TO DECISION TREE ALGORITHM FOR CLASSIFICATION OF SOIL TYPES

A Zakiuddin Ahmed
Research Scholar, Department of Computer Science, Jamal Mohamed College (Autonomous) Tiruchirappalli, Affiliated to Bharathidasan University, Tiruchirappalli, India

Dr. T. Abdul Razak
Department of Computer Science, Jamal Mohamed College (Autonomous) Tiruchirappalli, Affiliated to Bharathidasan University, Tiruchirappalli, India

ABSTRACT

This paper details the application of a decision tree algorithm for classification of soil types. The productivity of agriculture depends on environmental conditions and soil types. Soil dataset of particular area is downloaded from the Kaggle website for the purpose of finding the classification of soil types and based on the types of soil predicted the agriculturist will sow the seed. In this paper Soil types are classified by applying Bayesian approach to Decision Tree algorithm with Bayesian model is used for finding the classification of soil types. The idea behind is rather simple but powerful. The proposed algorithm offers some unique features not to be found in any other tree inducers while at the same time it can produce better results for many difficult problems. Experimental results are presented which illustrate the performance of generating best decision tree for classifying soil type from the given soil dataset. The Algorithm of Bayesian approach to Decision Tree helps to classify the soil types more accurately than the existing Algorithms KNN, SVM and Decision Tree selected for this research paper.

Keywords: Decision Tree, Bayesian Model, Machine Learning, Soil Dataset


http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=8

1. INTRODUCTION

Machine Learning Technique is the application and also the essential part of Artificial Intelligence (AI). Machine learning automatically learns from the observations or data
Application of Bayesian Approach to Decision Tree Algorithm for Classification of Soil Types

without being explicitly programmed. AI refers to the simulation of human intelligence in machines that are programmed to think like human and imitate their action. The term AI may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem solving. The process of learning starts with the observations or data from the given datasets. Data preprocessing techniques are applied to datasets.

Machine Learning is a method of data analysis that automates the analytical model building; computer need not have to be explicitly programmed.

Decision Tree belongs to supervised machine learning algorithm family. Decision Tree is a very simple model and is widely used in business solutions to support decision making process and risk analysis. Decision Trees are the trees that are used to solve the problem by using tree representation of given datasets. Each internal node in a decision tree corresponds to an attribute and each leaf node corresponds to the class label. The soil datasets collected from a particular region from the Kaggle website is given as input to the WEKA tool and tested with the accuracy and all with the existing algorithms with the WEKA tool. The program written in the python, Java or any other languages for the newly designed algorithms can be taken for analysis. The results of existing algorithms KKN, SVM and Decision Tree; and the proposed algorithm are compared to show that our proposed algorithm produces accurate result than the existing ones.

Soil data testing is often performed by commercial labs that offer several of tests, targeting group of compounds and minerals. The advantage associated with the soil test 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.

In this research work, we have used machine learning tool WEKA for the purpose of knowing the results of existing algorithms such as K-Nearest Neighbor, Support Vector Machine and Decision Tree. A new algorithm for the proposed method has been devised and the algorithm is implemented using Python. The results obtained for the existing algorithms and the results obtained from the proposed algorithm are furnished in the Table 2. Both the values of existing and proposed algorithms are compared visually using the bar chart in figure-2.

2. RELATED WORKS

Yethiraj. N.G et al., [1] studied the relevance of data mining tools and techniques and found out that there are several algorithms and techniques being applied in agricultural discipline. Some of the techniques, such as, ANN, ID3, the k-means, and the KNN and Support Vector Machines were applied in the field of Agriculture. Data mining application in agriculture is a relatively new approach for forecasting or predicting of animals, and agricultural management.

Barghavi.P et al., [2], examined that the data mining tools and techniques could be applied to characterize soil data and found that data mining depends on the amount of data used in the course of action. An increase in dataset size improves accuracy in the result, which may also improve the verification of suitable patterns compared to well-known statistical analysis.

Ramar.K et al., [3] says Data mining classification techniques applied to soil database can be successful in determine meaningful relationships from the data. Genetic Algorithm (GA) is an effective tool to use in data mining and pattern recognition. However, GA has problems with premature convergence which hold back diversity in the population and prevent exploration of the whole search space.
Anju Rathee et al [4] explained various algorithms such as Iterative Dichotomiser 3 (ID3), C4.5 Classification and Regression Tree (CART) and compared their performance and results and their evaluation is done by the existing datasets. The performance of different decision tree algorithms are investigated on the basis of their assurance and time seized for visualize the tree. Among all the three algorithms discussed in this paper, it has been observed that C4.5 is the best algorithm among all since it gives the preferred precision and proficiency over the other algorithms.

Amir Ahmad et al [5] presented a method called random projection random discretization ensembles which fundamentally makes the gatherings of multivariate decision trees with the help of univariate decision tree algorithm. Random Discretization is proposed which basically creates random discredited features. It has been accomplished that Random Projection Random Discretization Ensembles is used basically in small ensembles and it is also robust to the noisy data.

V. Rajeswari et al [6] experimented with classification algorithms such as JRip, J48 and Naive Bayes to predict the soil types from the given soil dataset. These algorithms were applied to extract the knowledge from soil data and two types of soil are considered such as Black and Red soil. Based on the soil type predicted using the above said classification algorithms, the agriculturist will cultivate the crop.

3. PROPOSED MODEL
In the proposed model, an algorithm is designed using Bayesian approach to Decision Tree which will classify the soil types more perfectly than already existing algorithms for the given soil dataset of a particular region. The proposed algorithm for classification of soil type using Bayesian approach to Decision Tree is designed and it will be implemented by writing code using python language. The results acquired from the proposed algorithm will be then compared with the results obtained for existing algorithms such as KNN, SVM and Decision Tree algorithms with the help of WEKA tool.

A Decision Tree Classifier operates by breaking down a dataset into smaller and smaller subsets based on different criteria. The different sorting criteria will be used to divide the dataset, with the number of examples getting smaller with every division. Decision Tree is referred to as a predictive modeling technique from the subfield of machine learning. It uses the divide and conquers method for splitting the data according to attribute values. Decision Tree analysis is utilized for predicting soil taxonomy from seismic data. Decision Tree processing is a technique that includes the recursive partitioning of data into progressively homogeneous subsets. After the partitioning is over, the subsets will be called as nodes. The label of the majority class is allocated to each terminal node.

Decision Tree Algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. The information gain is based on the decrease in entropy after a dataset is split on an attribute.

3.1. Bayesian Model
A Bayesian model is a statistical model where we use probability to represent all ambiguity within the model, both the uncertainty regarding the input and output to the model. Applying Bayes rule to the unknown variables of a data modeling problem is called Bayesian modeling.

Bayesian statistics is used for organizing an optional method for developing predictive connections between climate and soil types. Mapping soil properties within this probability model is a method by which values of a given soil resources at each location are extracted
Application of Bayesian Approach to Decision Tree Algorithm for Classification of Soil Types

from a set of specified soil resources groups. The decision tree algorithm is enhanced by using Bayesian model to improve the accuracy of soil classification.

Deduction or induction inference methods used by Machine Learning algorithm can make the computer to be trained from existing data or theories. To provide as training data, only small amount of samples are needed to form knowledge base by Machine Learning algorithm. It is easier than the direct extraction from the higher expert. This algorithm derives the rules and measures to connect the two parts and predictable output is related with set of examples. In this analysis part, the decision tree and Bayesian modeling are used and it needs training to retrieve the information. Usually soil sampling is measured as proficient technique to get such kind of results. Without any personal constraints and non-financial background, making an effort to collecting samples, are easier said than done.

3.2. Proposed Flow Diagram for Bayesian approach to Decision Tree.
The flow diagram for the proposed model is shown below.

- Soil dataset downloaded from Kaggle is given as input to the proposed algorithm.
- Preprocesses the soil dataset and features are extracted
- Training and test datasets are defined as usual
- Now the prepared dataset above is given as input to the DTBM
- Form leaf and branch required for an Decision Tree
- Now the Bayesian Model is introduced with the Decision Tree and which will form the final Decision Tree.

Figure 1 Proposed Flow Diagram
3.3. PROPOSED ALGORITHM

The proposed algorithm shows both the Decision Tree and the Bayesian Model in the machine learning mechanism. For this research paper soil dataset is downloaded from Kaggle and it has been used as input to the existing algorithms and as well as in the proposed algorithm. Initially, 80% of the training sample is given and trained for the prediction and 20% of the testing data are used for the process of soil prediction. The classification process has been done with the help of the combination of both the Bayesian and decision tree model in the machine learning algorithm. The classes are defined to predict the attribute value for the soil prediction.

The steps for soil classification using Bayesian approach to decision tree is given below:

Step 1: The input is given as dataset and with the help of the dataset pre-processing will be done.

Step 2: The feature extraction process will be performed, which will reduce the dimensionality of the data given as input. Then certain datasets are trained and tested.

Step 3: Using the tested dataset, the decision tree with the Bayesian model are combined for classification and performed.

Step 4: Finally, tree formation has been done and soil type is predicted by using the performance metrics such as precision, recall, F1-Score and accuracy.

4. RESULTS AND DISCUSSION

The proposed algorithm Bayesian approach to Decision tree is implemented using the Python language and the results are furnished in the Table 1 given below. The proposed algorithm is used by measures such as Recall, F-measures, Precision and accuracy. These four metrics are connected with mining classification. These four metrics are necessary to mark the judgment of the classification task.

The Precision and Recall are two extremely important model evaluation metrics. Precision refers to the percentage of our results which are relevant and the Recall refers to the percentage of total relevant results correctly classified by our proposed algorithm. The F-Score or F1-Score conveys balance between the Recall and Precision.

![Graphical representation of proposed](http://www.iaeme.com/IJARET/index.asp 812)

**Figure 2** Graphical representation of proposed

The performance results of the proposed Decision Tree with Bayesian Method algorithm is furnished in the above Table 1. And the results of Table 1 are visually shown in the above Figure 1 using the bar chart. There are six classes which are classified in the proposed algorithm namely class 0, class 1, class 2, class 3, class 4 and class 5. The overall accuracy is
Application of Bayesian Approach to Decision Tree Algorithm for Classification of Soil Types

85% predicted from the proposed Decision Tree with Bayesian model. The weighted average of precision, recall, and F1-Score are 0.87, 0.86 and 0.8541 respectively got from this proposed algorithm.

Table 1 Performance analysis of proposed technique

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.91</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>1</td>
<td>0.83</td>
<td>0.86</td>
<td>0.74</td>
</tr>
<tr>
<td>2</td>
<td>0.72</td>
<td>0.77</td>
<td>0.57</td>
</tr>
<tr>
<td>3</td>
<td>0.81</td>
<td>0.85</td>
<td>0.83</td>
</tr>
<tr>
<td>4</td>
<td>0.78</td>
<td>0.58</td>
<td>0.67</td>
</tr>
<tr>
<td>5</td>
<td>0.80</td>
<td>0.74</td>
<td>0.91</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>0.85</td>
</tr>
<tr>
<td>Weighted average</td>
<td>0.87</td>
<td>0.86</td>
<td>0.8541</td>
</tr>
</tbody>
</table>

4.1. Comparative Analysis of Existing Techniques with Proposed Model

Table 2 Comparisons of the existing techniques with proposed model

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>0.81</td>
<td>0.83</td>
<td>0.84</td>
<td>0.8356872</td>
</tr>
<tr>
<td>SVM</td>
<td>0.81</td>
<td>0.82</td>
<td>0.81</td>
<td>0.802529</td>
</tr>
<tr>
<td>DECISION TREE</td>
<td>0.79</td>
<td>0.81</td>
<td>0.7805</td>
<td>0.794765</td>
</tr>
<tr>
<td>Proposed Algorithm</td>
<td>0.87</td>
<td>0.85</td>
<td>0.86</td>
<td>0.854867</td>
</tr>
</tbody>
</table>

Figure 3 Graphical representation of comparison of the existing algorithms KNN, SVM and Decision Tree with proposed algorithm

The comparison of the existing algorithms such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Decision Tree (DT) classification algorithms with the proposed algorithm Decision Tree with Bayesian Method (DTBM) is shown in the Table 2. The same is depicted with help of Bar-chart in the Figure 2. Here the proposed algorithm has a higher precision, recall and F1-score and accuracy.
5. CONCLUSION

Soil type classification for the given datasets of a particular region with more parameters helped to achieve the classification tasks successfully and accurately. Professional methods are adapted to perform complex soil data sets utilizing the data mining techniques to improve the accuracy and efficiency of the classification of enormous soil datasets. In this research work, we have discussed the Decision Tree with Bayesian Model in soil prediction and soil classification in Machine Learning Techniques. Here, analysis of soil classification using different algorithms such as K-Nearest Neighbor, Support Vector Machine, and Decision Tree and Proposed Bayesian approach to Decision Tree Algorithm has been compared. In this paper, we have presented a comparative analysis of the various classification algorithms with proposed algorithm. Finally, the proposed Bayesian approach to Decision Tree Algorithm achieves better results compared to the other three existing algorithms such as K-Nearest Neighbor, Support Vector Machine and Decision Tree for soil type classification.

REFERENCES


