ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 1, January 2018

ENHANCED DECISION TREE CLASSIFICATION WITH THE ENSEMBLE BAGGING MOTHED FOR SOIL DATA CLASSIFICATION IN AND AROUND PERAMBALUR DISTRICT

Dr.S.Sivakumar[1], K.M.Samathuvam[2]
[1]Head & Assistant Professor, Dept. of Computer Science, Thanthai Hans Roever College, Perambalur.
[2]Research Scholar, Dept of Computer Science, Thanthai Hans Roever College, Perambalur

Abstract

Data Mining Techniques have led over various methods to gain knowledge from vast amount of data. There are the various research tools available for the large amount of data. Data mining has more technique to analyze the large amount of data like various classification algorithms. We have studied various data mining research tools available for analyzing the large amount of data. We have studied the WEKA data mining tools and various data mining classification algorithms like Bayesian classification Algorithm, Rule based classification and classification by Decision tree. This dissertation aim to study various research tools available for the comparison for large amount of data. They are applied to soil science database and found out relationship between them. A large data set of soil database is extracted from the Department of Soil Science, Thanthai Hans Roever Krishi Vishwavidyalaya, Perambalur, and Tamilnadu, India. The database contains measurement of soil profile data from locations of Perambalur districts. The data belongs to the hills area of Veppanthattai, Perambalur and Alathur taluk. Comparison was made between Decisions Naïve Base Classification and most effective techniques. The database is the measurement of soil profile data from location of Perambalur districts, Tamil Nadu, in India. The outcome of the research may have many benefits, to agriculture, soil management and environmental. Data mining software application includes various methodologies that have been developed by both commercial and research centers. These techniques have the Accuracy and Error Measures are important parameter to improve the efficiency of classification algorithms. Both the classifier and predictor have their own measures to improve the efficacy. Been used for industrial, commercial and, scientific purpose.

KEYWORDS: Data mining, Classification, Soil Classification,Decision Tree Indusion, Ensemble Classifier, Bagging Method.

1.Introduction

A.Classification in Determing

Classification is a data mining technique used to predict group membership for data instances. Classification maps data into predefined groups or classes. Pattern recognition forms the basis of classification, where an input pattern is classified into one of the several classes based on similarity (or proximity) to the predefined classes.The following are some of the Classification and Predication Mining approaches to mining the data from the large dataset in efficient manner.

- ✓ Classification by Decision Tree Induction
- ✓ Bayesian Classification
- ✓ Rule Based Classification
- ✓ Classification by backpropagation
- ✓ Support Vector Machines
- ✓ Associative Classification: classification by Association Rule Analysis
- ✓ Lazy Learners (or Learning from Your Neighbors)
- ✓ Genetic Algorithms
- ✓ Rough Set Approach
- ✓ Fuzzy Set Approach

- ✓ Prediction

B. Ensemble Methods in Classification

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produce more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods. In the popular Netflix Competition, the winner used an ensemble method to implement a powerful collaborative filtering algorithm. Another example is KDD 2009 where the winner also used ensemble methods. You can also find winners who used these methods in Kaggle competitions, for example here is the interview with the winner of crowd Flower competition.

It is important that we understand a few terminologies before we continue with this article. Throughout the article I used the term "model" to describe the output of the algorithm that trained with data. This model is then used for making predictions.

This algorithm can be any machine learning algorithm such as logistic regression, decision tree, etc. These models, when used as inputs of ensemble methods, are called "base models".

Voting and Averaging Based Ensemble Methods

Voting and averaging are two of the easiest ensemble methods. They are both easy to understand and implement. Voting is used for classification and averaging is used for regression.

In both methods, the first step is to create multiple classification/regression models using some training dataset. Each base model can be created using different splits of the same training dataset and same algorithm, or using the same dataset with different algorithms, or any other method.

II RELATED WORK

The Western Australian Department of Agriculture (AGRIC) conducted a large scale soil mapping project in the south west of the state in the mid-1980s. This soil mapping project was conducted with the support of the National Soil Conservation Program (NSCP), National Landcare Program (LCP) and Natural Heritage Trust.

The Purdie soil classification is the basis of Australian soil classification standards which have subsequently been adopted as the official system (Isbell, 1996). The use of soil classification maps have been shown to play a substantial role in agricultural production, salt control, large scale land management and land improvement.

According to Schoknecht, Tille, Purdie (2004, p. 14) "the soil groups of Western Australia are classified into 60 main groups; this provides a standard way of giving common names to the main soils of the state".

In another study WEKA was used to develop a classification system for the sorting and grading of mushrooms (Cunningham and Holmes, 1999). The system developed a classification system that could sort mushrooms into grades and attained a level of accuracy equal to or greater than the human inspectors.

The data, a total of 68 attributes including photo images, was used by the j4.8 algorithm classifier within WEKA to create a model for the human inspectors and the automated system. The model created using the human rules showed that each inspector used different combinations of attributes when assigning grades to mushrooms (Cunningham and Holmes, 1999). [5] discussed about a method, Sensor network consists of low cost battery powered nodes which is limited in power. Hence power efficient methods are needed for data gathering and aggregation in order to achieve prolonged network life. However, there are several energy efficient routing protocols in the literature; quiet of them are centralized approaches, that is low energy conservation. This paper presents a new energy efficient routing scheme for data gathering that combine the property of minimum spanning tree and shortest path tree-based on routing schemes. The efficient routing approach used here is Localized Power-Efficient Data Aggregation Protocols (L-PEDAPs) which is robust and localized. This is based on powerful localized structure, local minimum spanning tree (LMST). The actual routing tree is constructed over this topology. There is also a solution involved for route maintenance procedures that will be executed when a sensor node fails or a new node is added to the network.

III PROPOSED METHODOLOGY

Existing parameter to analysed physical properties of soil science. In this section we explain the existing parameter to analysed physical properties of soil science.

A. Parameter Used In This Research

1. Correctly Classified Instance

If the problem is a multi-class one (i.e. more than two classes) then AUC is calculated for each class in turn by treating all other classes as the negative class. It is possible to achieve a high AUC on one class while the overall classification accuracy is somewhat lower. Another possibility is due to the precision that is used to output stuff. Classification accuracy is output to four decimal places (and is a percentage between 0 and 100) while AUC is output to three decimal places (and is a number between 0 and 1) [7].

2. Incorrectly Classified Instance

Incorrectly classified instances refer to the case where the instances are used as test data and again are the most important statistics here for our purposes [8].

3. Kappa Statistic

Inter observer variation can be measured in any situation in which two or more independent observers are evaluating the same thing [9].

4 Mean Absolute Errors

The mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes [10].

5. Root Mean Square Error

The root mean square deviation (RMSD) or root mean square error (RMSE) is a frequently-used measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated [11].

6. Root Relative Square Error

The relative absolute error is very similar to the relative squared error in the sense that it is also relative to a simple predictor, which is just the average of the actual values. In this case, though, the error is just the total absolute error instead of the total squared error [12]

B. Data Collection Cleaning and Checking

1. Soil Data Collection

The dataset was collected as part of a survey by soil data of Perambalur district and included a large amount of information from different location within the Perambalur, Alathur and Veppanthattai Taluks. This information was collected from various locations where a pit was dug and samples taken. The samples were then sent for chemical and physical analysis at the agricultural laboratories in Department of Soil Science Thanthai Hans Roever Krishi

Vishwavidyalaya, Valikandapuram, and Perambalur.

2. 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 classification analysis to be conducted on the dataset.

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

3. 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. 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 to be applied to the soil data set to provide relevant outcomes that were required in the research [13].

C. Data Set : 01 – Descriptions

Eucalyptus Soil Conservation

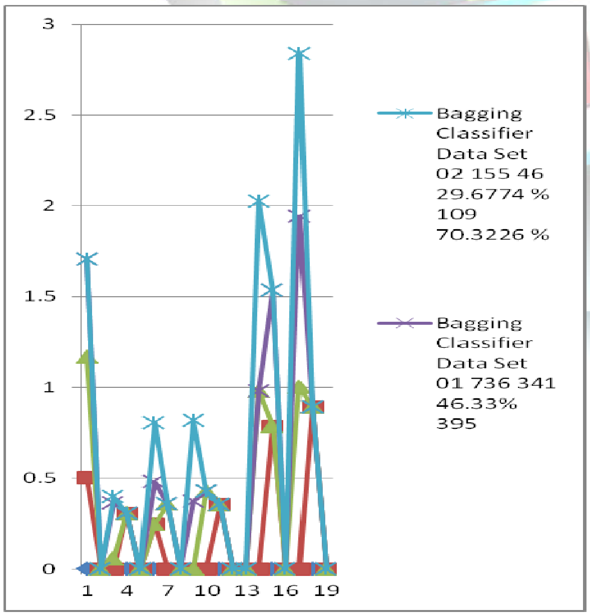
Data source: Roever KVK, Valikandapuram, Veppanthattai Taluk, Perambalur District

The objective was to determine which seed lots in a species are best for soil conservation in seasonally dry hill country. Determination is found by measurement of height, diameter by height, survival, and other contributing factors.

It is important to note that eucalypt trial methods changed over time; earlier trials included mostly 15 - 30cm tall seedling grown in peat plots and the later trials have included mostly three replications of eight trees grown. This change may contribute to less significant results.

IV SIMULATION RESULTS

Parameters / Algorithms	Decision Tree/ Data Set 01	Decision Tree/ Data Set 02	Bagging Classifier/ Data Set 01	Bagging Classifier/ Data Set 02
Total Number of Instances	736	155	736	155
Correctly Classified Instances &percentage	366 49.73%	52 33.55%	341 46.33%	46 29.6774 %
Incorrectly Classified Instances &percentage	370 50.27%	103 66.45%	395 53.67%	109 70.322 %
Kappa statistic	0.3059	0.0584	0.3064	0.0307
Mean absolute error	0.2466	0.3563	0.2337	0.3252
Root mean squared error	0.352	0.4272	0.3717	0.4449
Relative absolute error	78.72%	97.91%	74.60%	104.33%
Root relative squared error	88.95%	100.18%	93.92%	89.36%



Conclusion

This problem is solved with the help of data mining classification techniques. When we take large amount of soil data as input and analysed it in weka data mining software then error rate of various classifiers are detected, and we applied the soil data as a input data and analysed to the two data mining classification techniques such as

Decision Table tree and Bagging classifier teased by weka data mining tool. Analyzing to mentioned above data mining classification techniques with soil profile test parameters like Correctly Classified Instances, Incorrectly Classified Instances, Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error. With the help of this method we find when Bagging classification technique is applied to soil data set, the correctly classified instances are more classified. The Kappa statistic , Mean absolute error, Root mean squared error , Relative absolute error are less than the remaining Classifiers, like Bayesian classifier,J48.of soil profile. The time to build the Bagging Classifier is less than the remaining Classifier. So, The Bagging Classifier is the efficient classification technique among remaining classification techniques. Normalized Expected Cost of Bagging is more accurate when compared to Decisions tree classifier. Finally we say Bagging Classificationtechnique is best classification technique.

Future Work

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 n 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] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence. 1996.

[2] Usama Fayyad, Gregory Piatetsky-shapiro,Padhraic Smyth. Knowledge Discovery and Data Mining: Towards a Unifying Framework. 1996. p. 82-88.

[3] Schoknecht, N., Tille, P., and Purdie, B. (2004). Soil- Landscape mapping in south-western Australia (Technical Report). Perth: Department of Agricultural.

[4] Isbell, R. F. (1996). The Australian Soil Classification. Australian soil and land survey handbook. (Vol. 4). Collingwood, Victoria, Australia: CSIRO Publishing.

[5] Christo Ananth, S.Mathu Muhila, N.Priyadharshini, G.Sudha, P.Venkateswari, H.Vishali, "A New Energy Efficient Routing Scheme for Data Gathering ",International Journal Of Advanced Research Trends In Engineering And Technology (IJARTET), Vol. 2, Issue 10, October 2015), pp: 1-4

[6] Ibrahim, R. S. (1999). Data Mining of Machine Learning Performance Data. Unpublished Master of Applied Science (Information Technology), Publisher; RMIT University Press.

[7] Cunningham, S. J., and Holmes, G. (1999). Developing innovative applications in agriculture using data mining. In the Proceedings of the Southeast Asia Regional Computer Confederation Conference,1999..

[8] Mckenzie, N., and Ryan, P. (1999). Spatial prediction of soil properties using environmental correlation. Geoderma, 89(1-2), 67-94.

[9] M. Ankerst, C. Elsen, M. Ester, and H.P. Kriegel. "Visual Classification: An Interactive Approach to Decision Tree Construction". In Proc. 1999 Int. Conf. Knowledge Discovery and DataMining (KDD'99), pp. 392-396, San Diego, CA, Aug. 1999.

[10] Mckenzie N., and Ryan P. (1999). "Spatial Prediction of Soil Properties using Environmental Correlation".

[11] P. Bhargavi et al. "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", IJCSNS International Journal of Computer Science and Network Security, 9(8), August 2009, pp. 117-121.