Forest Cover Type Prediction with Machine Learning with R and Weka

Perera PLM[1]
Jayakody JRKC[2]

ABSTRACT

Data Mining has especially become popular in the fields of forensic science, fraud analysis and healthcare, since it reduces costs in time and money. Classification techniques can be a significant factor for determining the missing class attributes of a given test data. In this paper, Kaggle community (data scientists' community) membership was taken to predict the class type of the random forest test data set. Relevant train data and test data were given in the site. Many models were developed with the training data and ran it against with the test data to decide the missing classtypes of the test data. Varies data mining techniques such as decision trees, Random forest, nearest neighbor and Neural network was used to predict the missing forest type. Different data visualization techniques were followed to identify the dataset to enhance the model after tuning it with the best selected feature sets. R and Weka tools were used extensively for the analysis.

According to the research, prediction accuracy was increased in the following order such as Decision tree, Random forest,Nearest-Neighbour and Neural Network.R with Decision tree was given the least accuracy (0.35490) where as Weka with Neural Network was given the highest accuracy (0.72789). Weka tool algorithms were given good prediction rates compares to R tool. Feature selection and feature extraction was improved the prediction accuracy with different ML techniques.

KEYWORDS: Decision Tree, K-nearest neighbor, Neural Network, R, Random Forest, Weka

INTRODUCTION

During the last few years Data Mining has become more and more popular. With the information age, the digital revolution made it necessary to use some heuristics to analyze the large amount of data that has become available.

Data Mining is a process that consists of applying data, analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (models) over the data" (Fayyad, 1996). Another, sort of pseudo definition is; "The induction of understandable models and patterns from databases". (Rokach, 2007) In other words, we initially have a large (possibly infinite) collection of possible models (patterns) and (finite) data. Data Mining should result in those models that describe the data best, the models that fit (part of the data). Classification are used for the kind of Data Mining problem which are concerned with prediction. Using examples of cases, it is possible to construct a model that is able to predict the class of new examples using the attributes of those examples.

LITERATURE REVIEW

In order to analyze the data set, R and Weka tools were extensively used. As classification algorithm, Decision trees, Random Forest Nearest Neighbor and Neural network techniques were used.

Decision tree learning is a method commonly used in data mining (Rokach, 2008). The goal is to create a model that predicts the value of a target variable based on several input variables.

[1]Undergraduate, Department of Computing and Information Systems, Faculty of Applied Sciences, Wayamba University of Sri Lanka

[2]Lecturer, Department of Computing and Information Systems, Faculty of Applied Sciences, Wayamba University of Sri Lanka

Table 1: Main Attribute List of Forest Cover Type

Data Field	Description
Elevation	Elevation in meters
Aspect	Aspect in degrees azimuth
Slope	Slope in degrees
Horizontal_Distance_To_Hydrology	Horizontal distance to nearest surface water features
Vertical_Distance_To_Hydrology	Vertical Distance to nearest surface water features
Horizontal_Distance_To_Roadways	Horizontal Distance to nearest roadway
Soil_Type	(40 binary columns, 0 = absence or 1 = presence) - Soil
Wilderness_Area	(4 binary columns, 0 = absence or 1 = presence) -
Cover_Type	(7 types, integers 1 to 7) - Forest Cover Type designation
Horizontal_Distance_To_Fire_Points	Horizontal Distance to nearest wildfire ignition points

A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. Each element of the domain of the classification is called a class. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes (Rokach, 2008).

Random forest which has the base in decision trees. Random forests are an ensemble learning method for classification and regression which operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. (Quinlan, 1986). The general techniques of bootstrap aggregating or bagging were applied in the training algorithm for random forests.

The k-Nearest Neighbors algorithm (k-NN) is a non-parametric method used for classification and regression (Altman, 1992). In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. It is expected that features are selected carefully to extract relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction is performed on raw data prior to applying k-NN algorithm on the transformed data in feature space.

As the last ML technique neural network algorithm such as multilayer perception were used for further experiment. A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs (Rosenblatt, 1961). A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable (Rosenblatt, 1961). Multilayer perceptron neuron uses a nonlinear activation function which was developed to model the frequency of action potentials, or firing of biological neurons in the brain.

METHODOLOGY

Random forest cover type dataset was provided by Jock A. Blackard and Colorado State University. Test dataset, sample submission dataset and a training dataset were provided as .csv files. The

study area of the dataset includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. Each observation is a 30m x 30m patch. It is requested to predict an integer classification for the forest cover type. The seven types are 1 - Spruce/Fir, 2 - Lodgepole Pine 3 - Ponderosa Pine, 4 - Cottonwood/Willow, 5 – Aspen, 6 - Douglas-fir and 7 – Krummholz. The training set (15120 observations) contains both features and the Cover_Type. Only the features are contained in the test set. Therefore with the help of the training test it is requested to predict the cover type for every row in the test set (565892 observations). After a deeper examination, it was found that the data set contains 41 attribute set [Table 1]. As tools R and Weka were used and as Machine Learning technologies, decision trees, Random Forest, KNN and Neural Network technologies were used with different feature set to find a class type for each test set.

DATA COLLECTION AND ANALYSIS

Decision trees, random forest, nearest neighbor and neural network classifier methods were used to predict the class type with R and Weka tools. R was used with decision trees and random forest whereas Weka was used with decision trees, random forest, nearest neighbor and neural network methods.

R with Decision Trees

R and decision trees was used with R-Studio to load the test data and the training data. In order to run the training test, the library rpart named as Recursive Partitioning and Regression Trees and the CART decision tree algorithm were used. Since rpart comes with base R, it was imported before use it. In rpart [Figure 1] formula was used to predict the type with decision trees.

```
fit<-
rpart(as.factor(Cover_Type)~AttributeList(x1,x2,
x3,x4,x5.....xN,     data = train, method=class))
```

Figure 1: Decision Tree Code in R to Generate Class Type with All Variables

After execution of the algorithm the accuracy was reported as 0.35490 which is bit low. Then with R it was experimented with the variables to get an idea about different fields. [Figure 2]

```
Summary (trains Elevation)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1863    2376    2752    2749    3104    3849
```

Figure 2: Summary Statistics of Elevation Variable

The results from the decision trees were graphed to further analyze the dataset. The rattle, rpart and RcolorBrewer packages were downloaded and installed in order to generate the graphs. Finally, decision trees were reviewed with the help of the packages as given below [Figure 3] to further analyze the dataset.

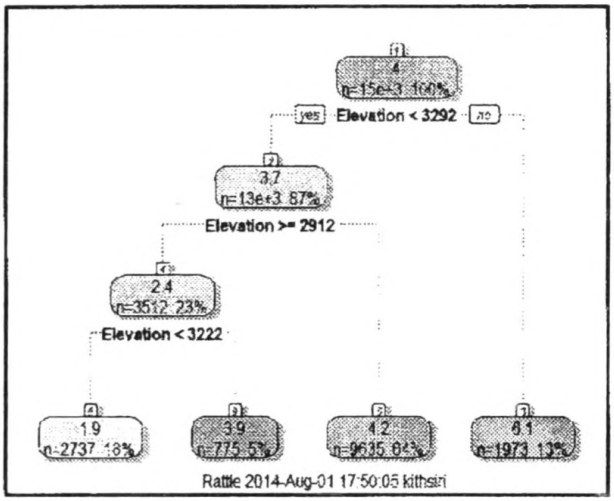

Figure 3: Decision Tree Hierarchy Generation for Elevation Variable

R with Random Forest

In the next step R and random forest was used. As it is explained previously, it was an improvement over decision trees. In this method Random Forest package was installed with R. Before it was executed with random forest, the number of seeds were set to 50. Then with the [Figure 4] it was executed the train test with R.

```
fit<- randomforest(as.factor(Cover_Type) ~
AttributeList(x1,x2,x3,x4....xN,     data = train,
controls= cforest_unbiased (ntree=2000,
mtry=6))
```

Figure 4: Code of Random Forest Usage to Predict the Class Type

Finally, R tool with the random forest accuracy was increased to 0.62859.

Weka with Random Forest

Weka tool was used to predict the class label in other steps. This tool has many classification algorithms. With the help of different algorithm accuracy was improved by tuning the algorithms after selecting appropriate attribute lists. The environment was setup in eclipse to run with weka and generate a sample answer list for the submission. Random forest with weka was used by keeping only the relevant attribute lists. Attribute list was observed with weka visualization tool to get a general idea of the attributes which is used for the classification.

Load trained data Load test data Load the classifier build
The model for the training data with the classifier for
Each test instances
- For each test instance verify it with the model
- Assign the class type and load into memory
- Until test instances complete

Figure 5: Pseducode to Build the Model in Eclipse with Weka

Eclipse code.[Figure 5] was used to build the model and run the sample. This was caused to improve accuracy up to 0.70594.

Weka with KNN

As the next attempt, nearest neighbor algorithm was used. Test sample was run with weka and eclipse with the IBk algorithm. KNN nearest neighbor was set to two. Dataset attributes was visualized with weka to find most important attribute set Aspect, Slope, Horizontal_Distance_To_ Hydrology, Vertical_Distance_To_Hydrology, Horizontal_Distance_To_Roadways, Hill shade_9am, horizontal_Distance _To_Fire_ Points were identified as main attribute set.KNN was improved the accuracy up to 0.7185.Then at the next step, IBk algorithm with wrapper class in weka were tuned. ClassifierSubsetEval attribute of weka was selected to search for the attributes with BestFirst method. In here classifier was used to estimate the 'merit' of a set of attributes. IBk was selected as the algorithm. Then the accuracy was improved to .71977.

Weka with Neural Network

As the last step the neural network algorithm was used. Multi-layer perception algorithm was selected to run with the default values with weka.jar. ClassifierSubsetEval attribute evaluator was selected to search for the attributes with Best First method. Elevation, Aspect, Slope, Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology, Horizontal_DistanceTo_Roadways, Hillshade_9am, Hillshade_Noon, Hillshade_3pm Horizontal_Distance_To_Fire_Points, Wilderness_Area1 were returned as the best attribute set. Then after tuning accuracy was improved up to 0.72789.

RESULTS AND DISSCUSSION

According to the research, prediction accuracy was increased in the following order [Table 2]. Such as decision tree, Random forest-nearest neighbor and Neural network. Apart from that Weka tool algorithms were given good prediction rates compared to R tool. Feature selection and feature extraction improved the prediction accuracy with both tools and different ML techniques.

Table 2: Forest Cover Type Prediction Accuracy Sheet

Technique	Accuracy
R with DT	.35490
R with Random Forest	.62859
Weka with Random Forest	.70594
Weka with KNN	.70954
Weka with KNN(Wrapper)	.71977
Weka with Neural Network	.72789

In this research, run time factor was the biggest obstacle. When samples were run with the weka, since the data were loaded to its memory, memory overflow happened and weka environment corrupted. Apart from that machine configuration were not sufficient to run these models with the test data to predict the results (RAM was 2GB). When number of iteration was increased upto 5, it took around 4 to 5 days to get the output. Normally, it is ideal to have a RAM size around 10GB. The features were tuned with weka and R and got a good accuracy of predicted results. When the number of folds were increased, time increases exponentially was another obstacle.

CONCLUSION

R and Weka were used extensively to predict the class type of the forest cover. Usage of Random Forest, decision trees, K-nearest neighbor and neural network algorithms are really useful to enhance the predictive accuracy of the results. Based on the above methodology classification accuracy was improved in each step. Since the computation power is less, results were not further enhanced with other types of algorithms. There are around 100 different classification algorithms are available and each can run with different feature engineering variation to increase the prediction accuracy, which can be suggested for extension of this work with future research.

REFERENCES

- Abonyi, J., Feil, B., & Abraham, A. (2005). Computational Intelligence in Data Mining. *Informatica (Slovenia)*, *29*(1), 3-12.
- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.
- Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
- Fayyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Intelligent Systems*, *11*(5), 20-25.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.
- Rokach, L. (2007). *Data mining with decision trees: theory and applications*. World scientific.
- Rosenblatt, F. (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms* (No. VG-1196-G-8).