Karar Ağaçları Öğrenme Yöntemleri ile Letter Verisi Sınıflandırılması

Decision Trees Learning Methods with Classification of Letter Data

Ahmet ERKOÇ
Elektrik-Elektronik Mühendisliği
KTO Karatay Üniversitesi
Konya, Türkiye
ahmeterkoc@gmail.com

***Abstract* — In this study, we compared the decision tree classification methods on the Letter dataset formed by David J. Slate from the data in the UCI site.**

Rep Tree, Random Forest and J49 algorithms were used for classification and the results were statistically compared.

***Keywords* — Letter Data, Decision Tree Learning, Rep Tree, Random Forest, J49**

I. INTRODUCTION

The concept of classification is simply to distribute the data in the various classes in the data set to the class. Various algorithms try to find out the distribution of these distributions from the given training clusters and try to classify distributions correctly in the direction of this test data. We call the values of these classes (labels). In this way we determine the class of data during training and testing.

The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. We typically train on the first 16000 items and then use the resulting model to predict the letter category for the remaining 4000.

Letter data are classified in twenty-six uppercase letters [A, B, C, ... ,Z] from the English alphabet, respectively. In Table 1, the value ranges of Letter data attributes are given.

TABLO I. LETTER DATASET ATTRIBUTE AND RANGE VALUES LETTER

No	Letter dataset attribute and range values		
	Quality	***Min.***	***Max.***
1	X-Box	0	15
2	Y-Box	0	15
3	Width of Box	0	15
4	Height of Box	0	15
5	Onpix Total	0	15
6	X-BarMean	0	15
7	Y-BarMean	0	15
8	X – Variance	0	15
9	Y – Variance	0	15
10	XY Correlation	0	15
11	X^2Y Mean	0	15

No	Letter dataset attribute and range values		
	Quality	***Min.***	***Max.***
12	XY^2 Mean	0	15
13	X-EGE	0	15
14	XEGVY	0	15
15	Y-EGE	0	15
16	YEGVX	0	15

II. CLASSIFICATION ALGORITHMS USED IN THE EXPERIMENTS

In this study, it is determined to use three different learning algorithms by the software called WEKA workbench. These Algorithms are shown in Table 2.

Table 2. Classifiers Used In the Experiments

No	Classifiers	Learning Type
1	Rep Tree	Decision Tree Learning
2	Random Forest	Decision Tree Learning
3	J49	Decision Tree Learning

Decision tree is a tree formed data structure that verifies divide and rule approach. Decision tree is used for supervised learning. It is a tree structured model in which the local region is found recursively, with a set of division in a few steps.

A. RapTree

REP Tree algorithm is based on the principle of calculating the information gain with entropy and reducing the error arising from variance. This method is firstly suggested by Quinlan.

With the help of this method, complexity of decision tree model is decreased by "reduced error pruning method" and the error rising from variance is reduced.

REP Tree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees.

That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree.

Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once.

Missing values are dealt with using C4.5's method of using fractional instances. The example of REP Tree algorithm is applied on UCI repository and the confusion matrix is generated for class gender having six possible values.

B. Random Forest

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.

In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

Random forests is an ensemble learning algorithm. The basic premise of the algorithm is that building a small decision-tree with few features is a computationally cheap process. If we can build many small, weak decision trees in parallel, we can then combine the trees to form a single, strong learner by averaging or taking the majority vote. In practice, random forests are often found to be the most accurate learning algorithms to date.

The same random forest algorithm or the random forest classifier can use for both classification and the regression task.

Random forest classifier will handle the missing values. When we have more trees in the forest, random forest classifier won't over fit the model. Can model the random forest classifier for categorical values also.

C. J48

The C4.5 algorithm for building decision trees is implemented in Weka as a classifier called J48.

This algorithm is an extension of ID3 algorithm and possibly creates a small tree. It uses a divide and conquers approach to growing decision trees that was leaded by Hunt and his co-workers.

The run-time complexity of the algorithm matches to the tree depth, which cannot be greater than the number of attributes. Tree depth is linked to tree size, and thereby to the number of examples. So, the size of C4.5 trees increases linearly with the number of examples. C4.5 rules slow for large and noisy datasets Space complexity is very large as we have to store the values repeatedly in arrays.

III. RESULTS AND CONCLUSION

In this study, decision trees the analysis of the Letter data is done. Data mining, confidential, important, is a data analysis technique that reveals useful information that is unknown beforehand. It method, unlike conventional analysis techniques, analysis can be done with non-numerical data, and hidden patterns can be uncovered.

Data mining is random in the data set instead of choosing a sample, it is hard to use the entire it is separated from the very analysis technique. Decision trees are not just data mining is one of the analytical techniques used in the type of classification. Decision trees it is quite easy to interpret the results produced because of its visual nature.

Many algorithms belonging to the decision trees are tried and the Random Forest algorithm is found as the most successful algorithm with the number of 19280 correctly classified samples. This algorithm, which makes the most accurate classification, produced a total of 15 classes. As a result of these analyzes, it can be predicted which English letter will come in return.

Classifier:	Rep Tree
Test Option:	Cross-Validation-10
Correctly Classified Instances:	%84.2

Classifier:	Random Forest
Test Option:	Cross-Validation-10
Correctly Classified Instances:	%96.4

Classifier:	J48
Test Option:	Cross-Validation-10
Correctly Classified Instances:	%87.9

The comparison of decision trees methods on Letter data was done using Rep Tree, Random Forest, J48 (C4.5). These methods are used in many classification fields today and it is seen that successful results are obtained.

It seems that the choice of ANN gives easy, fast and consistent results. These methods are often preferred in research. When we look at the results we have received, the most successful classification methods are Random Forest with 96.4%. Other methods and success rates are given in the table.

KAYNAKLAR

[1] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[3] Murugappan, M., Ramachandran, N., Sazali, Y. (2010). Classification of Human Emotion from EEG Using Discrete Wavelet Transform. Journal of Biomedical Science and Engineering, 3(04), 390.

[4] https://www.academia.edu/2541077/Rastlant%C4%B1sal_Karar_A%C4%9Fa%C3%A7lar%C4%B1yla_Nesne_ve_Renk_Da%C4%9F%C4%B1l%C4%B1m%C4%B1na_G%C3%B6re_Sahne_S%C4%B1n%C4%B1fland%C4%B1r%C4%B1lmas%C4%B1

[5] https://www.journalagent.com/iuyd/pdfs/IUYD-43531-RESEARCH_ARTICLE-AKCETIN.pdf

[6] http://dergipark.gov.tr/download/article-file/30544

[7] http://ijiset.com/vol2/v2s2/IJISET_V2_I2_63.pdf

[8] http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf

[9] http://www.socsci.ru.nl/idak/teaching/batheses/bachelor_thesis_patrick_ozer.pdf

[10] http://shodhganga.inflibnet.ac.in/bitstream/10603/21206/14/14_chapter%205.pdf

[11] https://www.academia.edu/4375403/Decision_Tree_Analysis_on_J48_Algorithm_for_Data_Mining