DOI: 10.17667/riim.2020.1/3.

Classification of Mushroom Data Set by Ensemble Methods

Şahin YILDIRIM
Mechatronic Engineering Department
Faculty of Engineering, Erciyes University
Turkey, Kayseri
sahiny@erciyes.edu.tr

Mehmet Safa BİNGÖL
Mechatronic Engineering Department
Faculty of Engineering, Erciyes University
Turkey, Kayseri
msbingol@erciyes.edu.tr

***Abstract*—Due to disease of mushrooms, it is very important to classify mushrooms for predicting the best quality mushrooms. Classification algorithms, in the most general sense, attempt to estimate which class an object belongs to. Ensemble methods are used to increase the classification success rate. Ensemble methods simultaneously use two or more classification algorithms. There are many methods to analyse the main parts of mushrooms. For above mentioned descriptions; in this simulation study five types of classification algorithm are employed to predict the structure of mushrooms. Mushroom dataset is used to predict the classes of mushrooms. The results are improved compared to other classification methods (logistic regression, naive bayes classifier, k-nearest neighbor, support vector machines, random forest, neural networks) that these methods will be used to predict exact mushrooms features and classifications in real time approaches.**

Keywords—classification; ensemble methods; machine learning; mushroom dataset.

I. INTRODUCTION

Recently, many different algorithms are used for classification. Logistic regression [1], naive bayes classifier [2], and support vector machines [3] are algorithms that classify based on statistical methods. K-nearest neighbour [4] algorithm uses the data directly for classification, without building a model first. Decision tree [5], repeatedly splits the data set resulting in a tree-like structure. The main purpose in the classification process is to try to determine which class an object belongs to. Choosing the classification method that is suitable for the problem is important for increasing the accuracy rate.

Ensemble methods aim to increase the accuracy rate by using more than one classifier for the classification problem. Here, the classifiers used can be the same type or different types. The main purpose of methods is to increase the accuracy rate over the best individual classifier [6, 7, 8].

When using the Ensemble method, the result should be better than a single classifier, otherwise there is no point in using the ensemble method. Because the calculation cost of the ensemble method is higher than a single classifier [9]. General structure of ensemble learning approach is given in Figure 1.

Ensemble learning has been used in many different applications such as classification of birdsong [10], credit risk evaluation [11], to improve deep learning performance [12], a telemedicine tool framework for lung sounds classification [13].

Approximately 14,000 species of mushroom are known in the world. 2000 species are reported to be edible and, among these edible mushrooms, about 200 are wild species [14].

Recently, many researchers have conducted research projects on mushrooms to classify poisonous mushroom and edible mushroom species using different classification techniques on the mushroom dataset [15, 16, 17].

In this study, the samples in mushroom dataset, which has 2 classes, are classified as poisonous and edible. 5 different ensemble methods are used for this classification process. Results are given in tabular form and evaluated for accuracy and time.

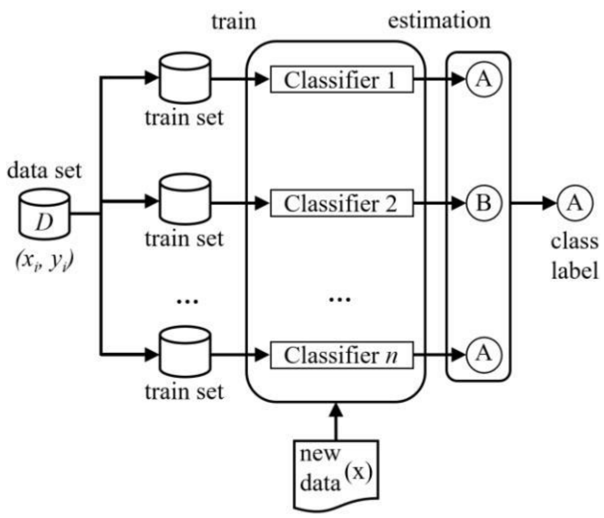

Fig. 1: The flow chart of ensemble learning method

II. MATERIALS AND METHOD

In this study, 5 different ensemble classifiers (Subspace Discriminant [18], RUSBoosted Trees [19], Subspace KNN [20], Bagged Trees [21], Boosted Trees [22]) were tested on mushroom dataset.

A. *Mushroom dataset*

In this study, the mushroom dataset were used [23]. The mushroom data set has 8124 sample. Each sample has specific 23 features. In the dataset, mushrooms are divided into poisonous and edible classes. This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms [23]. Mushroom parts are given in Figure 2.

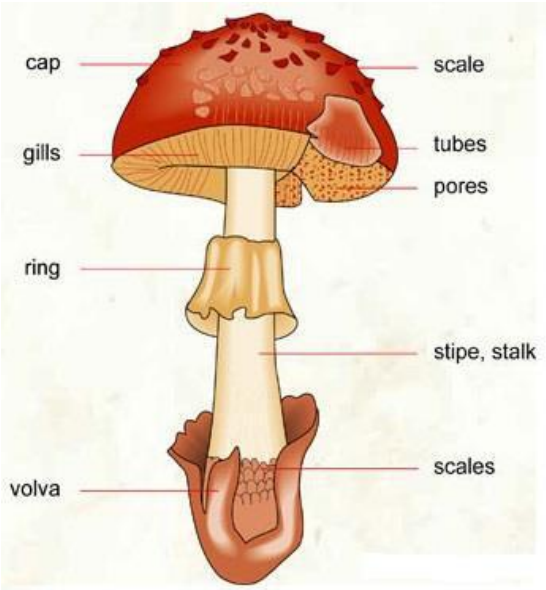

Fig. 2. Mushroom parts [24].

The features of the mushroom dataset are given as following [15]:

1) Attribute Information:(classes: edible, poisonous)
2) cap-shape: conical, convex, knobbed, bell, flat, sunken.
3) cap-surface: grooves, smooth, fibrous, scaly.
4) cap-color: buff, gray, brown, green, pink, cinnamon, purple, white, red, yellow.
5) bruises: bruises, no.
6) odor: anise, creosote, almond, foul, musty, fishy, none, spicy, pungent,
7) gill-attachment:free, descending, attached, notched.
8) gill-spacing: crowded, distant, close.
9) gill-size: narrow, broad.
10) gill-color: brown, buff, black, gray, green, purple, orange, red, white, pink, yellow, chocolate.
11) stalk-shape: tapering, enlarging.
12) stalk-root: rhizomorphs, cup, bulbous, equal, rooted, club.
13) stalk-surface-above-ring: silky, scaly, smooth, fibrous.
14) stalk-surface-below-ring: silky, scaly, smooth, fibrous.
15) stalk-color-above-ring: pink, buff, gray, orange, red, yellow, cinnamon, white, brown.
16) stalk-color-below-ring: pink, buff, gray, orange, red, yellow, cinnamon, white, brown.
17) veil-type: universal, partial.
18) veil-color: yellow, orange, brown, white.
19) ring-number: two, one, none.
20) ring-type: sheathing, flaring, large, none, evanescent, pendant, zone, cobwebby.
21) spore-print-color: orange, black, yellow, brown, buff, purple, green, white, chocolate.
22) population: abundant, clustered, numerous, scattered, several, solitary.
23) habitat: waste, meadows, paths, grasses, urban, woods, leaves.

Edible and poisonous mushroom population type percentage is given in Figure 3.

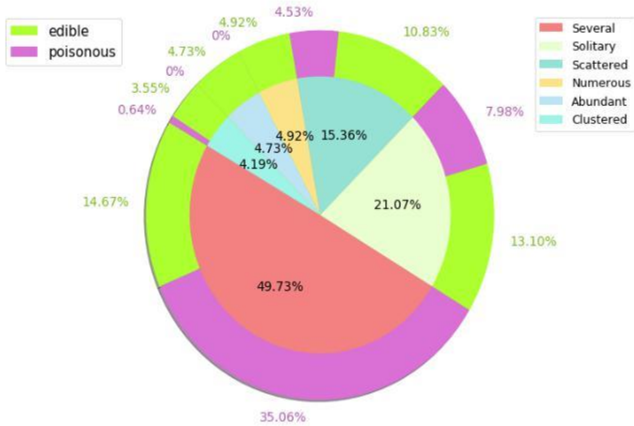

Fig. 3. Representation with flow chart edible and poisonous mushroom population type percentage [25].

B. *Ensemble Learning*

The general working principle of ensemble learning can be outlined in the following:

Suppose that the training data set consists of m samples D={(x_1,y_1), (x_2,y_2), ..., (x_m,y_m)}, that there are *k* class labels $y_i \in Y = \{1,...,k\}$, that the classification algorithm is represented by *L*, and that the community size is set to *n* [26].

Ensemble learning step:

Step 1: D data set is used directly (Voting) or creates *n* new D_i data sets from D data set (Bagging, Boosting).

DOI: 10.17667/riim.2020.1/3.

Table 1. The results of five types of methods

Number	Method	Accuracy (%)	Training Time (sec)	Prediction Speed (obs/sec)
1	Boosted Trees	89.6	28.13	74000
2	Bagged Trees	100	27.69	14000
3	Subspace Discriminant	99.2	30.43	4900
4	Subspace KNN	99.9	33.20	1600
5	RUSBoosted Trees	89.6	30.16	69000

Step 2: The following operation is repeated n times; different data sets are trained by the same algorithm $C_i = L$ (D_i) or the same data set are trained by different learning algorithms $C_i = L_i$ (D).

Step 3: compare the decisions of the test set with classifiers

Step 4: Output from each classifier for a new instance x, $y_i = C_i$ (x)

Step 5: The results of n classifiers $\{C_1, C_2, ..., C_n\}$ are combined. The general formula of Ensemble learning is given in equation 1.

$$C^*(x) = \underset{(y \in Y)}{argmax} \sum_{i:C_i(x)=y}^{n} 1 \qquad (1)$$

III. RESULTS

The mushroom data set was used to train 5 different ensemble classificaiton methods. Accuracy, training time and prediction time of the trained ensemble methods was examined. The results are given in Table 1.

The training time of Bagged Trees is the shortest and the training time of Subspace KNN is the longest. The prediction speed of Boosted Trees is the fastest and the prediction speed of Subspace KNN is the slowest. The accuracy rate of Bagged Trees is the highest and the accuracy rate of Boosted Trees and RUSBoosted Trees are the lowest.

Five types of ensemble methods were employed to classification mushroom dataset. On the other hands, the classification results showed that the Bagged Trees ensemble method has good performance to classification mushroom dataset. Because, the accuracy rate of Bagged Trees is the highest and the training time of Bagged Trees is the shortest.

IV. CONCLUSIONS AND DISCUSSIONS

In classification problems, ensemble learning based on combining more than one classifier to improve classification performance of only one classifier has been proposed. Ensemble classifiers can mitigate some of the mistakes made by individual classifiers so that the performance of an ensemble classifier is probably better than the performance of the best single classifier.

This paper has presented an invastigation regarding to ensemble methods for classifiying mushroom dataset. These approaches were improved that this kind of methods will be employed to other types of plants, in real time labrotory experimental works.

References

[1] S. Dreiseitl, L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," Journal of biomedical informatics, 2002, 35(5-6), 352-359.

[2] X. Feng, S. Li, C. Yuan, P. Zeng, Y. Sun, "Prediction of slope stability using naive Bayes classifier," KSCE Journal of Civil Engineering, 2018, 22(3), 941-950.

[3] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, A. S. Hegde, P. L. Kubben, "Performance evaluation of DWT based sigmoid entropy in time and frequency domains for automated detection of epileptic seizures using SVM classifier," Computers in biology and medicine, 2019, 110, 127-143.

[4] A. Bablani, D. R. Edla, S. Dodia, "Classification of EEG data using k-nearest neighbor approach for concealed information test," Procedia computer science, 2018, 143, 242-249.

[5] E. Gokgoz, A. Subasi, “Comparison of decision tree algorithms for EMG signal classification using DWT,” Biomedical Signal Processing and Control, 2015, 18, 138-144.

[6] T. G. Dietterich, “Ensemble methods in machine learning,” in International workshop on multiple classifier systems, Springer, June 2000, p. 1-15.

[7] A. Subasi, D. H. Dammas, R. D. Alghamdi, R. A. Makawi, E. A. Albiety, T. Brahimi, and A. Sarirete, “Sensor Based Human Activity Recognition Using Adaboost Ensemble Classifier,” Procedia Computer Science, 2018, vol. 140, pp. 104-111.

[8] I. O. Joudeh, A. M. Cretu, R. B. Wallace, R. A. Goubran, A. Alkhalid, M. Allegue-Martinez, and F. Knoefel, “WiFi Channel State Information-Based Recognition of Sitting-Down and Standing-Up Activities,” in International Symposium on Medical Measurements and Applications, IEEE, June 2019, p. 1-6.

[9] C. Kapucu, and M. Çubukçu, “Fotovoltaik Sistemlerde Topluluk Öğrenmesi Temelli Hata Tespiti,” Bilişim Teknolojileri Dergisi, 2019, vol. 12, pp. 83-91.

[10] S. A. Brooker, P. A., Stephens, M. J. Whittingham, S. G. Willis, “Automated detection and classification of birdsong: An ensemble approach,” Ecological Indicators, 2020, 117, 106609.

[11] K. Niu, Z. Zhang, Y. Liu, R. Li, “Resampling Ensemble Model Based on Data Distribution for Imbalanced Credit Risk Evaluation in P2P Lending,” Information Sciences, 2020.

[12] Z. M. Jan, B. Verma, B, “Multiple Strong and Balanced Clusters based Ensemble of Deep Learners,” Pattern Recognition, 2020, 107420.

[13] M. M. Jaber, S. K. Abd, P. M. Shakeel, M. A. Burhanuddin, M. A. Mohammed, S. Yussof, “A telemedicine tool framework for lung sounds classification using ensemble classifier algorithms,” Measurement, 2020, 107883.

[14] P. Kalač, “Chemical composition and nutritional value of European species of wild growing mushrooms: A review,” Food Chemistry, 2009, 113, 9–16.

[15] S. Ismail, A. R. Zainal, A. Mustapha, “Behavioural features for mushroom classification,” In 2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2018 April, 412-415, IEEE.

[16] A. A. R. Khan, S. S. Nisha, M. M. Sathik, “Clustering Techniques For Mushroom Dataset,” 2018, 1121-1125.

[17] S. K. Verma, M. Dutta, “Mushroom classification using ANN and ANFIS algorithm,” IOSR Journal of Engineering (IOSRJEN), 2018, 8(01), 94-100.

[18] A. S. Ashour, Y. Guo, A. R. Hawas, G. Xu, “Ensemble of subspace discriminant classifiers for schistosomal liver fibrosis staging in mice microscopic images,” Health information science and systems, 2018, 6(1), 21.

[19] E. S. Sankari, D. Manimegalai, “Predicting membrane protein types using various decision tree classifiers based on various modes of general PseAAC for imbalanced datasets,” Journal of theoretical biology, 2017, 435, 208-217.

[20] S. Bavkar, B. Iyer, S. Deosarkar, S. “Detection of alcoholism: an EEG hybrid features and ensemble subspace K-NN based approach,” In International Conference on Distributed Computing and Internet Technology, 2019, January, pp. 161-168, Springer, Cham.

[21] G. Martínez-Muñoz, A. Suárez, “Using boosting to prune bagging ensembles,” Pattern Recognition Letters, 2007, 28(1), 156-165.

[22] M. Zięba, S. K. Tomczak, J. M. Tomczak, “Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction,” Expert systems with applications, 2016, 58, 93-101.

[23] (2020) The mushroom classification data setwebsite. [Online].Available: https://www.kaggle.com/uciml/mushroom-classification

[24] (2020) Mushroom Parts [Online].Available: https://infovisual.info/en/biology-vegetal/mushroom

[25] (2020) The mushroom classification website. [Online]. Available: https://www.kaggle.com/mig555/mushroom-classification

[26] P. Yildirim, K. U. Birant, V. Radevski, A. Kut, and D. Birant, “Comparative analysis of ensemble learning methods for signal classification,” in 26th Signal Processing and Communications Applications Conference (SIU), IEEE, May 2018, p. 1-4.