International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 07 | July 2018 www.irjet.net p-ISSN: 2395-0072

A Data Mining Based On Ensemble Classifier Classification Approach for Edible Mushroom Identification

Muhammad Husaini[1]

[1]*School of Computer Sciences, University of Science Malaysia, Pulau Pinang, Malaysia*

Abstract - *Mushrooms constitutes alternative source of food that have been weighed as ingredient of gourmet cuisine across the globe. Although among 45000 mushroom species available around the world, only few of them are edible. In this case, data mining classification approach is introduced in order to classify between edible and poisonous mushroom. To classify the mushroom dataset, four distinctive algorithm were compared namely Naïve Bayes, Ripple-Down Rule (RIDOR), and Bayes Net. Three different ensemble classifier such as Boosting, Begging, and Stacking were used to investigate the performance of the algorithms. From the result, it is inferred that Naïve Bayes and Bayes Net that used Begging method has obtained minimal classification accuracy with 95.687 % 96.187 % respectively. The maximal classification accuracy obtained for training and testing in this study was 100% which shows the result obtained were promising.*

Key Words: Boosting, Begging, Ensemble Classifier, Data Mining, Mushroom, Classification Algorithm, Ripple Down Rules (RIDOR).

1. INTRODUCTION

In the earliest history, people believe that mushroom is a treasure that provide good health and strength to human body. However, mushroom become so popular nowadays because they have a lot of important nutrition like niacin, riboflavin, selenium, potassium, and vitamin D as a prevention of hypertension, Alzheimer, Parkinson, and high risk of stroke [1]. According to [2], among 45000 species of mushrooms exist around the world, they can be identified as edible, definitely poisonous, or unknown and only 2000 species are edible. However, there are not an easy way to distinct between edible and poisonous mushroom. Most of the poisonous mushroom can show up as edible mushrooms due to size and colour [3].

In the beginning of research, manual assessment for mushroom classification was used to identify the characteristic of the mushroom [4], [5]. However with the emerging of technology, the modern technique tool like data mining that has been applied in this area of study [6],[7]. As mention by [2],[8], and [9], data mining algorithm like Support Vector Machine (SVM), Decision Tree (C4.5), Naïve Bayes and RIDOR algorithms have been used in their research for mushroom classification. Moreover, these classification algorithm for a single classifier is limited based their performance. Therefore new approach called ensemble classifier like boosting, staking, voting, and etc. could be the solution. In general, this method could provide better classification accuracy than a single predictor can do due to combination of classifier by merge the results of base classifiers.

This paper analyse the classification technique association with type of mushroom to help public to identify the edible mushroom correctly. This study of analysis include by revealing the accuracy of some classification techniques that has been measured and other relative importance of classifier algorithm like precision, time, mean absolute error and etc. The main target of this research work is to compare the classification result with previous work in order to find the best classification technique. Section 1 provides an overview of research. Section 2 provides a brief review of related work on mushroom dataset used. Section 3 describes methodology. In Section 4 result of classification techniques on mushroom data. Section 5 concludes the paper with future perspective.

2. LITERATURE REVIEW

Based on the current research [9] regarding on mushroom dataset from UCL, they found out that decision tree algorithm and support vector machine algorithm produce 100% of correctly classify instances from the databased for both method they are using which are evaluate of training data and 10 fold cross-validation. However decision tree was consider better algorithm than SVM in term of processing time. Negative side of this research show that Naïve Bayes was the worst classifier with the lowest percentage accuracy among those model. In accordance with [9], another study [10] using the same method on classify the mushroom dataset with different classifier models also found that Naïve Bayes was the poorest classifier with the lowest accuracy and precision after ZeroR compare to other. Also Verma and Dutta [11] present an approach for classifying the different types of mushrooms, which are edible or non- edible based on Naïve Bayes, ANN, and ANFIS classifier algorithm. They compare the result in reference to the highest accuracy. Result reveal that ANFIS outperform other classifier with 99.88%. The poorest classifier still belongs to Naïve Bayes algorithm.

Research regarding to ensemble classifiers has two major issues, which are combining of training methods of the base classifiers and combining methods for the decisions making

of the base classifiers [12].Ensemble classifier generation methods using homogeneous base classifiers can be broadly classified into five groups that are based on manage the training parameters, manage the error function, manage the feature space, manage the output labels, and manage the training patterns. All these methods aim to accomplish diversity among the base classifiers.

According to Sun et al. [13], they have been categorize top three most popular ensemble classifier methods which are bagging, boosting, and random subspace ensembles. They found out that this method depends on type of base classifier, parameter and setting for each individual classifier. Until this time, these approaches have been proven to be quite flexible in a broad area of research such as sentiment classification, face recognition, and etc. [14], [15] .In contrast with one single classifier, an ensemble classifier has excellences to manipulate a classification task which is difficult for traditional methods, to attain higher prediction accuracy [16].

3. METHODOLOGY

3.1. Dataset

The dataset for the mushroom was acquired from the UCI repository. This dataset contains samples of mushrooms from the Agaricus and Lepiota Family and then they are classified as definitely edible, definitely poisonous or of unknown edibility and not recommended. This mushroom dataset (Table 1) contain 8124 number of instances with 22 number of attributes. There are 2 class labels where definitely edible become one class label as 'e' and definitely poisonous or of unknown edibility and not recommended form one class label as 'p'. The dataset has very even class distribution with 51.8% are edible and 48.2% are poisonous.

Table -1: Mushroom Attribute Dataset

Attribute Description	Attribute Type
class	Nominal
cap-shape	Nominal
cap-surface	Nominal
cap-color	Nominal
bruises	Nominal
odor	Nominal
gill-attachment	Nominal
gill-spacing	Nominal
gill-size	Nominal
gill-color	Nominal
stalk-shape	Nominal
stalk-root	Nominal
stalk-surface-above-ring	Nominal
stalk-surface-below-ring	Nominal
stalk-color-above-ring	Nominal
stalk-color-below-ring	Nominal
veil-type	Nominal
veil-color	Nominal
ring-number	Nominal
ring-type	Nominal
spore-print-color	Nominal
population	Nominal
habitat	Nominal

3.2. Splitting dataset for training and testing

Mushroom dataset had been split into training and testing where 90% of mushroom dataset was used for training and the remaining used for testing purpose. Only training dataset will undergo several process of analysis where testing dataset was preserved after the process in order to check either classifiers algorithm overfit or not.

3.3. K-fold cross validation

In this work, K-fold cross validation has been used as the training procedure, where the training dataset was randomly divided into 'K' where K=10 set. The classifier was trained with sub-sample, i.e., 90% of the overall feature set and the training classifier is validated with the remaining 10% dataset. The cross-validation process is then repeated for all the K sets. The K classification result from the folds, then be average to estimate the mean classification accuracy. In this study, the feature set of mushroom training dataset were split into a training set and testing set using K-fold cross validation scheme.

3.4. Classifiers Algorithm

Classification is one of the most crucial steps in any pattern recognition algorithm. Commonly, classification algorithm can be categorize into supervise and unsupervised learning. In supervise learning output data is provided with labels while for unsupervised learning there is no prior information about output labels. Pervious researchers [2][9][10] found that Naïve Bayes and RIDOR algorithm not really perform well to classify the dataset. For that reason, Boosting, Bagging, and Stacking has been used in this research. Later performance of each classifiers algorithm was compared.

4. RESULT

4.1. Performance of the Mushroom Dataset Using ZeroR Algorithm

Table 2 shows the performance for the ZeroR classifier technique on the mushroom dataset based on different evaluation on training dataset like Correctly Classified Instances, Incorrectly Classified Instances, Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, and Root relative squared error. Based on the result, it can be seen that, the dataset has been distribution evenly with 51.8% are edible and 48.2% are poisonous.

Table -2: Result for ZeroR

Evaluation on Training Dataset	Algorithm
	ZeroR
Correctly Classified Instances	51.7565%
Incorrectly Classified Instances	48.2435 %
Kappa statistic	0
Mean absolute error	0.4994
Root mean squared error	0.4997
Relative absolute error	100 %
Root relative squared error	100 %

4.2 Performance of the Training Mushroom dataset using AdaBoost

From Table 3, it can be observed that the performance of mushroom dataset using AdaBoost based on Naïve Bayes, RIDOR, and Bayes Net algorithm almost comparable. All of the algorithm had 100% correctly classify and 0 % incorrectly classify. However regarding on Mean absolute error, Root mean squared error, Relative absolute error, and Root relative squared error RIDOR algorithm perform better compare to Naïve Bayes and Bayer Net.

Table -3: Performance of the Training Mushroom dataset based on AdaBoost Algorithm

Evaluation on Training Dataset	Algorithm		
	Naïve Bayes	RIDOR	Bayer Net
Correctly Classified Instances	100 %	100 %	100 %
Incorrectly Classified Instances	0 %	0 %	0 %
Kappa statistic	1	1	1
Mean absolute error	0.0001	0	0
Root mean squared error	0.003	0	0.0001
Relative absolute error	0.0153 %	0 %	0.0003%
Root relative squared error	0.6001 %	0 %	0.0128%

4.3. Performance of the Training Mushroom dataset using Begging

Table 4 shows the performance of the Mushroom dataset based on Begging classification method. It is clear that RIDOR algorithm perform really well by having 100% correctly classify, 0% incorrectly classify and 1 for Kappa statistic. With the Kappa statistic greater than 0 means RIDOR algorithm doing better than chance. Both classifier, Naïve Bayes and Bayer Net performance are comparable to each other with only slight difference between them. However, according to correctly classified instances and Kappa statistic, Bayes Net perform better than Naïve Bayes.

Table -4: Performance of the Mushroom dataset based on Begging

Evaluation on Training Dataset	Algorithm		
	Naïve Bayes	RIDOR	Bayes Net
Correctly Classified Instances	95.687 %	100 %	96.187 %
Incorrectly Classified Instances	4.313 %	0 %	3.813 %
Kappa statistic	0.9134	1	0.9235
Mean absolute error	0.0424	0.0001	0.0382
Root mean squared error	0.176	0.003	0.1643
Relative absolute error	8.4858 %	0.0175 %	7.6572%
Root relative squared error	35.2153 %	0.592 %	32.8803%

4.4. Performance of the Training Mushroom dataset using Stacking

Meta classifier: Decision Tree (J48)

Table 5 discuss the performance of the Mushroom dataset based on Stacking method that used Decision Tree (J48) as a Meta classifier. Table 5 shows the evident that RIDOR algorithm is the best algorithm for stacking method compare

to Naïve Bayes and Bayes Net since it has 100% accuracy with value of 1 for Kappa statistic. RIDOR algorithm also has the lowest value for Mean absolute error, Root mean squared error, Relative absolute error, and Root relative squared error compare to Naïve Bayes and Bayes Net. Both Naïve Bayes and Bayes Net are comparable since there are not much different result between them.

Table -5: Performance of the Mushroom dataset based on Stacking

Evaluation on Training Dataset	Algorithm		
	Naïve Bayes	RIDOR	Bayes Net
Correctly Classified Instances	98.1498 %	100 %	98.3123 %
Incorrectly Classified Instances	1.8502 %	0 %	1.6877 %
Kappa statistic	0.963	1	0.9662
Mean absolute error	0.0365	0	0.0333
Root mean squared error	0.1347	0.0001	0.1288
Relative absolute error	7.3154 %	0.0028 %	6.6698 %
Root relative squared error	26.9634 %	0.0127 %	25.7707 %

4.5. Overall Performance Using Test Dataset

The overall performance of the ensemble classifier using testing dataset shown in Table 6. Both Correctly Classified Instances and Incorrectly Classified Instances show good result since most of the percentage accuracy appear more than 95%. While Mean absolute error also display excellent result with most of them almost to zero or equal to zero. Last but not least, only begging that used Naïve Bayes algorithm and Stacking that used Bayes Net algorithm have the lowest value which are 0.9839. However the rest of algorithm have maximum value of 1. Therefore, this indicate that there is no contradiction between training and testing performance of the classifier algorithms.

Table -6: Overall Performance Using Test Dataset

	Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error
AdaBoost	Naïve Bayes	100 %	0 %	1	0.0034
	RIDOR	100 %	0 %	1	0
	Bayer Net	100 %	0 %	1	0.0008
Begging	Naïve Bayes	99.2 %	0.8 %	0.9839	0.0065
	RIDOR	100 %	0 %	1	0.0008
	Bayes Net	100 %	0 %	1	0.005
Stacking	Naïve Bayes	100 %	0 %	1	0
	RIDOR	100 %	0 %	1	0
	Bayes Net	99.2 %	0.8 %	0.9839	0.0.243

5. CONCLUSIONS

In this study, it had been noted that AdaBoost classification technique perform the best compare to Begging and Stacking method. It also found that Naïve Bayes, RIDOR, and Bayer Net algorithm in AdaBoost perform extremely well for both training and testing result shows 100% accuracy. These algorithm also give best outcome for Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, and Root relative squared error. By comparing with previous method [2][9][10], this technique especially for RIDOR algorithm improve really well.

For the near future, research on image processing in order to identify the edible and non-edible mushrooms can be investigated. Also different optimization like Ant colony and Fish swarm optimization can be used.

REFERENCES

[1] M. E. Valverde, T. Hernández-pérez, and O. Paredes-lópez, "Review Article Edible Mushrooms : Improving Human Health and Promoting Edible Mushrooms : Improving Human Health and Promoting," no. January, 2015.

[2] D. R. Chowdhury and S. Ojha, "An Empirical Study on Mushroom Disease Diagnosis : A Data Mining Approach," 2017.

[3] C. FM, "Amanita phalloides in Victoria," p. 849–850., 1993.

[4] A. van der Neut, "The development of a set of characteristics for D. U. S. tests of cultivated mushroom varieties," First Int. Semin. Mushroom Sci., pp. 153–160, 1991.

[5] G. Polder and G. W. A. M. Van Der Heijden, "IDENTIFICATION OF MUSHROOM CULTIVARS USING IMAGE ANALYSIS," vol. 35, no. 1, pp. 347–350, 1992.

[6] A. T. C. C. Salvador, M. R. Martins, H. Vicente, J. Neves, and J. M. Arteiro, "Modelling molecular and

inorganic data of Amanita ponderosa mushrooms using artificial neural networks," Agrofor. Syst., vol. 87, no. 2, pp. 295–302, 2013.

[7] H. V. A. Teresa Caldeira, J. M. Arteiro, J. C. Roseiro, and J. Neves, "An artificial intelligence approach to Bacillus amyloliquefaciens CCMI 1051 cultures: Application to the production of anti-fungal compounds," Bioresour. Technol., vol. 102, no. 2, pp. 1496–1502, 2011.

[8] A. A. R. Khan, S. S. Nisha, and M. M. Sathik, "CLUSTERING TECHNIQUES FOR MUSHROOM DATASET," no. June, pp. 1121–1125, 2018.

[9] A. Wibowo, Y. Rahayu, and A. Riyanto, "Classification Algorithm for Edible Mushroom Identification," pp. 250–253, 2018.

[10] SUNITA BENIWAL and BISHAN DAS, "MUSHROOM CLASSIFICATION USING DATA MINING TECHNIQUES," Int. J. Pharma Bio Sci., vol. 6, no. 1, pp. 1170–1176, 2015.

[11] S. K. Verma and M. Dutta, "Mushroom Classification Using ANN and ANFIS Algorithm," vol. 8, no. 1, pp. 94–100, 2018.

[12] A. Rahman and B. Verma, "Knowledge-Based Systems Ensemble classifier generation using non-uniform layered clustering and Genetic Algorithm," Knowledge-Based Syst., vol. 43, pp. 30–42, 2013.

[13] D. Z. S. Sun, and C. Zhang, "An experimental evaluation of ensemble methods for EEG signal classification," Pattern Recognit. Lett., vol. 28, no. 14, pp. 2157–2163, 2007.

[14] K. Tumer and N. C. Oza, "Classifier ensembles: Select real-world applications," Inf. Fusion, vol. 9, no. 1, pp. 4–20, 2008.

[15] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Inf. Sci. (Ny)., vol. 181, no. 6, pp. 1138–1152, 2011.

[16] G. Wang, C. Zhang, and J. Zhuang, "An Application of Classifier Combination Methods in Hand Gesture Recognition," vol. 2012, pp. 1–18, 2012.