Mushroom Evaluation

Chong Yu
John Jay College of Criminal Justice
Chong.yu@jjay.cuny.edu

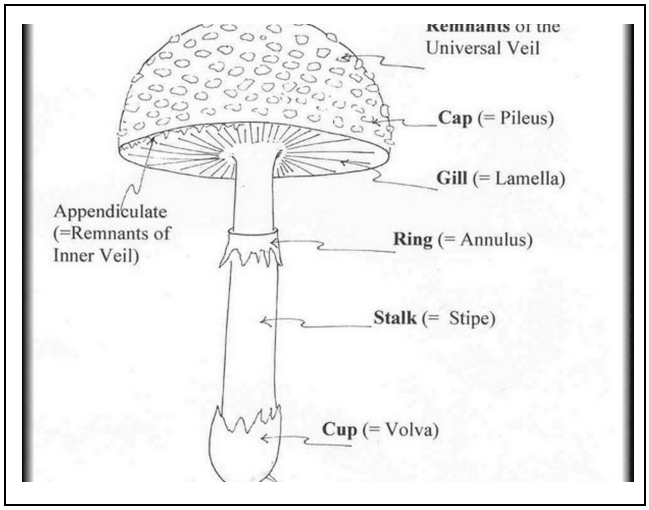
Hunter Johnson, Mathematics & Computer Science Department
John Jay College of Criminal Justice

ABSTRACT

When choosing an algorithm, there should be a consideration with the data that will be evaluated because of the accuracy, speed, precision and recall may fluctuate. In this paper, I choose to compare a dataset of hypothetical samples. From the creators to describe the dataset ": *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 (1981). Although this dataset was originally contributed to the UCI Machine Learning repository nearly 30 years ago, mushroom hunting (otherwise known as "shrooming") is enjoying new peaks in popularity*." I selected eight features that would help improve classification and applied them to four machine learning models: Logistic Regression, Decision Tree, Random Forest Classifier and Support Vector Machine. The main goal is to determine whether or not a mushroom would be fatal or nonfatal. Considering this dataset problem in regarding life threatening situations the ideal goal is to achieve the least error because the factor of death is a tragic consequence.

1 INTRODUCTION

Mushrooms are often marketed as a "meat replacer" due to their protein content and fleshy texture. Mushrooms are a good source of the B vitamins riboflavin, pantothenic acid, and niacin. Some countries treat mushrooms as a kind of high nutrition food. However, not all mushrooms are edible. It can be fatal to eat a mushroom without prior knowledge of its species. Various classification algorithms will be used to form the best model to predict whether new emerging mushrooms are edible based on the detected data of the mushrooms. The main goal is to determine whether or not a mushroom would be fatal or nonfatal. Considering this dataset problem in regarding life threatening situations the ideal goal is to achieve the least error because the factor of death is a tragic consequence.



Above, we have various features of the mushrooms such as cap-shape, cap-surface , cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring ,veil-type, veil-color, ring-number, ring-type, spore-print-color, population and habitat.

Furthermore, it is an opportunity to compare the classifiers and also understand how they operate. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the credibility of a mushroom; no rule like "leaflets three, let it be" for Poisonous Oak and Ivy.

In this paper, we will discuss Logistic Regression, Decision Tree, Random Forest Classifier, and Support Vector Machine ensemble learning technique. We will then look at implementations of each and show a bit of the final product of each algorithm. This will be followed by graphical analysis of the algorithm. And finally, the algorithm will be rated and judged accordingly.

1. Logistic Regression
2. Decision Tree
3. Random Forest Trees
4. SVM

To check for prediction. Out of these SVM's performance was poorer as compared to the others. And RandomForest's accuracy was 100%. Next, the feature importances was used using RFC to see what features are impacting the classification the most. Then, I again make the prediction using the important features, using the same ALGOS mentioned above. And find the performance of SVM to have improved. Then, I again make predictions on the Test data. And this time, the number of edible mushrooms have increased by a fraction.

2 DATA

Data Set Characteristics	Multivariate
Attribute Characteristics:	Categorical
Associated Tasks:	Classification
Number of Instances:	8124
Number of Attributes:	22
Missing Values?	Yes
Area:	Life
Date Donated	1987-04-27
Number of Web Hits:	528720

This was the approach towards this dataset. Each row represented a mushroom, and it was labeled as e = edible or p = poisonous.

```
class,cap-shape,cap-surface,cap-color,bruises,odor,gil
above-ring,stalk-color-below-ring,veil-type,veil-color
p,f,s,e,f,s,f,c,n,b,t,?,215,s,k,p,w,p,w,38,o,e,w,v,l
p,x,y,w,t,p,f,c,n,k,e,e,94,s,s,w,w,p,w,43,o,p,n,v,u
e,f,s,n,t,n,f,c,b,e,e,?,244,s,s,e,e,p,w,37,t,e,w,c,w
p,x,s,e,f,y,f,c,n,b,t,?,286,k,s,p,w,p,w,82,o,e,w,v,l
```

The data contained 8,124 rows and 23 columns. We notice that there are 25 features in all in the training data. Out of which 23 are categorical and 2 are continuous features. There are no missing values. So I need to impute the categorical data. Since, there is a lot of ordinality in the categorical data. So, I use Label Encoder to encode them. Now, I check for any outliers in the data. It seems that there aren't any in this dataset. To improve our performance on the test dataset, we split our training dataset into cross validation dataset. Now by starting for predictions.

Attributes

```
class,cap-shape,cap-surface,cap-color,bruises,odor,gill-attachm
ent,gill-spacing,gill-size,gill-color,stalk-shape,stalk-root,st
alk-surface-above-ring,stalk-surface-below-ring,stalk-color-abo
ve-ring,stalk-color-below-ring,veil-type,veil-color,ring-number
,ring-type,spore-print-color,population,habitat
```

We notice that there are 25 features in all in the training data. Out of which 23 are categorical and 2 are continuous features. There are no missing values. So I need to impute the categorical data. Since, there is a lot of ordinality

in the categorical data. So, I use Label Encoder to encode them. Now, I check for any outliers in the data. It seems that there aren't any in this dataset. To improve our performance on the test dataset, we split our training dataset into cross validation dataset.

3 METHODS

1. Logistic Regression
2. Decision Tree
3. Random Forest Trees
4. SVC

Logistic regression is a classification method that uses regression as an internal component. The way it works in two stages: first it computes something closely related to the probabilities of being in each target class and then it labels an example with the highest probability class. Step one is a regression from the predictors to the note quite probabilities. Taking a threshold against not quite probabilities or taking a maximum among several values gives us a class. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

Tree building methods result in a model that can be thought of as a pathwork of constant predictors. The differences among decision tree (DT) methods revolve around how they break down the entire space of the data into smaller and smaller regions and when they stop the break down. Decision Trees are a type of supervised learning used for classification problems. The first step was induction, where the machine was trained on labeled data. The second step was deduction, where the model was tested against unlabeled samples that were held out of the training data set. There are a number of major tree building algorithms. IDS, C4.5, and C5.0 were made by Quinlan. CART was developed independently. In general, the tree building algorithms use the following steps: Evaluate the set of features and splits and pick the best feature and split. Then add a node to the tree that represents the feature-split. For each descendant, work with the matching data and either if the targets are similar enough return the predicted target or if not, return back to step one and repeat.

Random Forest is a type of group training, where a set of classifiers is constructed, then a committee is formed and a vote is taken to select the best classifier. In this case, randomized sets of Decision Trees are created, and one tree is selected per subset. This creates a "forest", which can then be combined into one best tree. If it is decided that a particular variable is not contributing significantly to information gain, the user could choose to remove that variable in order to reduce the dimensionality of the data. The training data set was passed to the model without the class variable at the same time calculating the variable importance.

Support Vector Machine is a kind of supervised learning. It can tackle both linear separable and non-linear separable classification problems. The objective is to locate the widest margin between classes. The training examples closer to the margin are the support vectors, while the ones further from the margin are non-support vectors, and are not part of the model. In essence, we can throw out a lot of our data and just focus on the points that matter. The points that matter are called support vectors. The heart of support vector classifiers (SVC) is to find the support vectors and then do the mathematics necessary to figure out the maximum margin separator between those points. SVCs try to balance off two competing concerns: getting the biggest margin between the example classes and minimizing the number of training errors. We want the biggest margin because under certain assumptions it leads to a good generalization and a good test set error. Also, there are two things that drive the need for more support vectors which are additional complexity in the boundary between the classes and examples that don’t play

nicely with the proposed boundaries. Within the data set, each row is a vector, and the number of columns is the dimensionality. An SVM can use different kernels, such as polynomial, linear, radial and sigmoid. The strengths are a high tolerance to noise data, probabilistic prediction, flexibility in data representations and scalability. The weaknesses are that each kernel type requires a number of parameters, and interpretability of non-linear kernels is a challenge.

4 RESULTS

The first dataset is recorded with alphabetic characters. Each alphabetic character is additionally spoken to as a number, the scope of each element isn't in a fundamental scale. To begin with, I used the most feature importances from model 3 (Random Forest Classifier) and fed it through a sort feature ranking. Then, I use this number to be the range of them, and convert each alphabetic character to a reasonable integer prior from using label encoder. From running each model with cross_val_score. Which is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation we have the results shown below:

Accuracy of Logistic Regression: 89.400226

Accuracy of Decision Tree: 99.601269

Accuracy of Random Forest Classifier: 99.601269

Accuracy of Support Vector Machine: 97.443843

When using the linear classifier such as stochastic gradient descent passing with gradient_of_the_pointwise_error_function we were able to obtain the performance of :

E_in = 0.5210970464135021

5 CONCLUSION

Today, mushrooms can be helpful for antibacterial, anti-inflammatory and antioxidants. While additionally assisting with lessening circulatory strain, moderate glucose, decrease cholesterol, upgrade the insusceptible framework, diminish pressure and help in battling numerous kinds of disease. There are a couple of mushrooms found in the wild that are dangerously poisonous. Some of these look like regular consumable species, however, it very well may be hazardous gathering wild mushrooms without great information for recognizing mushrooms. With the models sharing somewhat similarity in their results This shows there is no single classification algorithm which can give the best prescient model to all datasets. The precision of the prescient model is influenced by the choice of features selected. With this we can conclude that the different classification algorithms are designed to perform better for particular sorts of database. Using the eight features that would help improve classification and applied them to four machine learning models: Logistic Regression, Decision Tree, Random Forest Classifier and Support Vector Machine. The main goal is to determine whether or not a mushroom would be fatal or nonfatal. Considering this dataset problem in regarding life threatening situations the ideal goal is to achieve the least error because the factor of death is a tragic consequence.

6 REFERENCES

[1] Schlimmer,J.S. (1987). Concept Acquisition Through Representational Adjustment (Technical Report 87-19). Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine.

[2] Iba,W., Wogulis,J., & Langley,P. (1988). Trading off Simplicity and Coverage in Incremental Concept Learning. In Proceedings of the 5th International Conference on Machine Learning, 73-79. Ann Arbor, Michigan: Morgan Kaufmann.