Mushroom Classification By Family

Lauren Brakke
lab84717@uga.edu
Institute for Artificial Intelligence, University of Georgia
Alyssa Joaquin
agj04689@uga.edu
Institute for Artificial Intelligence, University of Georgia
Drew Becker
ajb6717@uda.edu
Institute for Artificial Intelligence, University of Georgia

Abstract

Binary classification of mushroom edibility using data mining techniques has been thoroughly explored. Expanding on prior research on mushroom classification, we utilize the dataset of mushroom identification attributes to evaluate different multinomial classification methods for the mushroom class *family*. We perform attribute selection based on two metrics, information gain and gain ratio, to create ranked groups of the top ten and top five for comparison of model results, in addition to one control group of all unranked attributes. Using WEKA, we built models using the control and experimental groups and the following classifiers: random forest, J48, JRIP, Naive Bayes, and nearest-neighbor. With all attributes available, random forest, J48, JRIP, and NN performed exceptionally well above 99% accuracy, with Naive Bayes performing around the 92 percentile. For both information gain ranked groups, all classifiers with the exception of Naive Bayes retained accuracy above 90% for top five, and above 99% for top 10. For the gain ratio groups, the top 10 group with the exception of Naive Bayes performed around the 99% mark. When reduced to the top five, each learner's accuracy decreased significantly by at least 20%. Overall, reducing from top ten to top five attributes had a much larger effect on accuracy for the gain ratio group, and Naive Bayes had consistently lower accuracy for all experimental groups. Using the top attributes for information gain achieves the most stable results, but may be subject to overfitting. The t-test with the smallest score, gain ratio gain ratio top five attributes and all attributes, indicates that their difference is significant.

1. Introduction

Prior research by Wagner, Heider, and Hattab focuses on classifying mushrooms as poisonous or edible. However, due to mushroom similarity, "a binary classification cannot be reliable" (Wagner et al., 2021). Wagner et al. suggest that their work can be extended beyond binary classification. Building off this research, our project objective is to use this mushroom data for multinomial classifications of the class *family*.

Framing our objective in question form, we ask: what is the most effective method for classifying different families of mushrooms? Comparing several classification methods: random forest, J48, JRIP, Naive Bayes, and nearest-neighbor, and experimental groups: four groups describing the top 10 and top 5 attributes ranked by information gain and gain ratio, which ones perform the best relative to our evaluation criteria?

Using accuracy as our performance metric, Naive Bayes can be immediately ruled out, as it performs consistently worse than all other classifiers across each experimental group. When attributes are selected by information gain rank in groups of ten and five, random forest, J48, JRIP, and nearest-neighbor all score around 99% accuracy, dropping to around 92.5% when reduced to five attributes. For attribute selection by gain ratio, the top 10 also achieve about 99%

accuracy, but experience a greater decrease when reduced to top five attributes, dropping to a range of 66-72% accuracy. Using the top ranked attributes for information gain produces more stable results, with consistently high accuracy for random forest, J48, JRIP, and nearest-neighbor. Looking at t-tests, GainRatio top 5 attributes and all the attributes have the smallest t-score, indicating that the difference associated with these values is significant.

To divide the work, we assigned 1-2 methods per person: Drew implemented Naive Bayes and led the data preprocessing efforts, Alyssa implemented random forest and nearest-neighbor, and Lauren implemented JRIP and J48. The t-tests and two-tailed t-tests were divided evenly, with each group member calculating 4-5 tests. The final report was divided evenly into parts for each member to handle but still written collaboratively: Alyssa wrote the introduction and literature review, Lauren wrote the abstract, summarization of the dataset, and future work sections, and Drew handled the data pre-processing, summary of model process, and analysis of results. Each group member also played a role in the editing process.

2. Literature Review

In this section, we discuss related research efforts and explain why we chose to focus on the classification of mushrooms by family. Wagner et al. predicted the edibility of mushrooms by classifying them as either poisonous or nonpoisonous. They used two datasets for their classification; their primary dataset was based on a textbook for mushroom identification, containing 173 species and 23 families of mushrooms (Wagner et al., 2021). The secondary dataset served as pilot data for classification tasks, composed of hypothetical mushroom entries structurally equivalent to that of the primary dataset (Wagner et al., 2021). They evaluated the classification performance of several different machine learning algorithms: naive Bayes, logistic regression, and linear discriminant analysis, and random forests (Wagner et al., 2021). Their findings concluded that random forest provided the best classification, achieving perfect results in both accuracy and F2 score (Wagner et al., 2021). Furthermore, Wagner et al. discussed a potential extension of their work, stating that their classification task can be extended to classifying a certain family or species of mushroom. We will build on the work of Wagner et al. by expanding their binary task to a multiclass task. Therefore, our work will be set apart in that instead of classifying mushrooms into two groups- poisonous and not poisonous- we will classify them into 23 groups based on their family.

Other projects also utilize the mushroom data for various data mining research projects. In “Test-cost sensitive naive Bayes classification”, Chai et al. illustrate how to obtain a test-cost sensitive naive Bayes classifier and compare its performance to several other classifiers, using several datasets including the mushroom dataset. Dawood et al. use an Artificial Neural Network (ANN) to train and test the mushroom dataset for prediction in “Artificial Neural Network for Mushroom Prediction”. They use a JNN tool for training and validation, identify key attributes of the dataset, and achieve a 100% accuracy of prediction (Dawood et. al, 2020). Similarly, in “Mushroom Classification Using ANN and ANFIS Algorithm”, Verma and Dutta use Artificial Neural Network and Adaptive Neuro Fuzzy inference system to implement different classification techniques for classification of edible or non-edible.

In another paper, “Behavioral Features for Mushroom Classification”, Ismail et al. use the Principle Component Analysis (PCA) algorithm for feature selection for classification of poisonous or non-poisonous, using J48 decision trees. They rank all mushroom features, with

odor named as their top ranked feature (Ismail et al., 2018). Chumuang et al. present their results of classification using physical features illustrating that k-NN shows 100% accuracy in "Mushroom Classification by Physical Characteristics by Technique of *k*-Nearest Neighbor". In "Classification Algorithm for Edible Mushroom Identification", Wibowo et al. use Weka to test the comparison of three classification algorithms: Decision Tree (C4.5), Naive Bayes, and Support Vector Machine (SVM). Their results demonstrate that SVM and C4.5 have the same accuracy level (100%) with Bayes falling behind at 95.8887%,, but C4.5 is faster than SVM (Wibow et. al, 2018).

In "Edibility Detection of Mushroom Detection Using Ensemble Methods", Pinky et al. use ensemble methods bagging, boosting, and random forest to classify mushrooms as edible or poisonous. Their findings illustrate that random forest and dissimilarity measure-based bagging have the highest accuracy, but random forest is faster (Pinky et al., 2019). Our literature review is not entirely comprehensive of the robust data mining research that utilizes the mushroom dataset, but it illustrates that its prior usage primarily revolves around classification as poisonous or edible. Our project differentiates from most research by analyzing data mining techniques for the classification of mushrooms based on their family.

3. Summarization of the Data Set

The Mushroom Data Set includes a primary and secondary dataset. The primary dataset includes attributes based on mushroom identification for 173 species from 23 families (Wagner et al., 2021). The primary dataset has 24 columns, 3 classes - *family*, *species*, *class* (edible or poisonous), and 21 variables describing various numeric and nominal attributes. Numeric variables include quantitative ranges for *cap-diameter*, *stem-heigh*t, and *stem-width*. Nominal variables are represented as single letters from a set or binary values for each variable including attributes *cap-shape*, *cap-color*, *does-bruise-or-bleed*, *has-ring*, *gill-attachment*, *gill-spacing*, *gill-color*, *stem-surface*, *stem-color*, *veil-type*, *veil-color*, *ring-type*, *spore-print-color*, *habit*, and *season*. More detailed descriptions of each variable can be found below in Table 1.

The secondary dataset is a simulated dataset based on a textbook identification guide, including 353 hypothetical mushroom entries per species, resulting in 61,069 total hypothetical mushrooms (Wagner et al., 2021). The primary challenge with this dataset is that although it was generated from an identification guide that included *family* and *species* values for each instance, the original generated dataset is unlabelled with *family* or *species* values. However, because the simulated dataset is sorted, and the original authors published a specific report of how they generated the dataset (noted above), we were able to create a simple python program to add the *family* and *species* labels from the primary dataset as a variable to the data. We also include summary statistics of all numeric variables in Table 2. Although our data is not balanced on the class values for family, each family has a large number of instances and is substantial for our model.

Variable	Type	Description
edibility	Nominal	poisonous=p, edible=e
cap-diameter	Numeric	float number in cm
cap-shape	Nominal	bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
cap-surface	Nominal	fibrous=f,grooves=g,scaly=y,smooth=s

cap-color	Nominal	brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k
does-bruise-or-bleed	Nominal	bruises-or-bleeding=t,no=f
gill-attachment	Nominal	adnate=a, adnexed=x, decurrent=d, free=e, sinuate=s, pores=p, none=f, unknown=?
gill-spacing	Nominal	close=c, distant=d, none=f
gill-color	Nominal	see cap-color + none=f
stem-height	Numeric	float number in cm
stem-width	Numeric	float number in mm
stem-root	Nominal	bulbous=b, swollen=s, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r
stem-surface	Nominal	see cap-surface + none=f
stem-color	Nominal	see cap-color + none=f
veil-type	Nominal	partial=p, universal=u
veil-color	Nominal	see cap-color + none=f
has-ring	Nominal	ring=t, none=f
ring-type	Nominal	cobwebby=c, evanescent=e, flaring=r, grooved=g, large=l, pendant=p, sheathing=s, zone=z, scaly=y, movable=m, none=f, unknown=?
spore-print-color	Nominal	see cap color
habitat	Nominal	grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
season	Nominal	spring=s, summer=u, autumn=a, winter=w
family	Nominal	Amanita, Lepiota, Tricholoma, Wax-Gill, Russula, Pluteus, Entoloma, Bolbitius, Stropharia, Cortinarius, Mushroom, Ink-Cap, Bolete,Paxillus, Chanterelle, Oyster-Mushroom, Crepidotus, Hydnum, Ear-Pick, Bracket-Fungi, Saddle-Cup, Morel, Jelly-Discs

Table 1: All attributes, their type, and description

Variable	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
cap-diameter	0.38	62.34	6.734	5.265
stem-height	0	33.92	6.582	3.37
stem-width	0	103.91	12.149	10.036

Table 2: Summary statistics of of all numeric attributes

4. Data Preprocessing Description

Because our simulated dataset is sorted, our first step for data preprocessing was to apply a randomization filter to the dataset to shuffle all the instances. There is also one attribute, *species*, stored as a *string* type in WEKA, due to a high number of species labels. The *string* type is incompatible with many of the learners in WEKA, and we decided that using *species* to classify the *family* would be unfairly accurate, so we removed the *species* attribute from all the runs in preprocessing. We observed that our dataset has many attributes resulting in very high model accuracy, so we utilized the attribute selector on Weka to narrow down to top five and top ten attributes. We decided to use these two groups of attributes as experimental groups in our comparison of the results of different models. Furthermore, we selected the attributes for each experimental run by using a subset evaluation method that selects attributes that have the highest predictive value and the least redundancy. We made two experimental groups based off of ranker algorithms with two separate criteria, InformationGain and GainRatio. Because the dataset is generated instead of collected, we suspected a level of overfitting in our classifiers. GainRatio is commonly used to combat overfitting, so we chose to use GainRatio as our main attribute

evaluator method. However, we still wanted to compare the results between gain ratio and information gain. The top ten and top five ranked attributes for gain ratio and information gain are summarized in Table 3 and Table 4.

Rank	Attribute	Gain Ratio
1	ring-type	0.4819
2	has-ring	0.4761
3	gill-attachment	0.4089
4	does-bruise-or-bleed	0.3502
5	gill-color	0.3408
6	habitat	0.3381
7	cap-shape	0.3293
8	stem-color	0.2662
9	gill-spacing	0.2367
10	stem-width	0.2111

Table 3: Top 10 Gain Ratio

Rank	Attribute	Information Gain
1	gill-attachment	1.1029
2	stem-width	1.0672
3	gill-color	1.0226
4	cap-shape	0.745
5	stem-color	0.6637
6	cap-surface	0.6056
7	cap-diameter	0.5926
8	stem-height	0.5764
9	ring-type	0.5461
10	cap-color	0.5204

Table 4: Top 10 Information Gain

5. Summarization of Model Development Process

For our models developed in WEKA, we chose the following classifiers: random forest, J48 (based on the C4.5 tree learner), JRIP (based on RIPPER covering algorithm), and nearest-neighbor (k=1). Because our data has 61,069 instances, we chose to split our training and testing data, instead of using 10-fold cross-validation which proved to be too time-consuming and computationally expensive. For each classifier, we ran models using all attributes in the mushroom dataset, the top ten attributes ranked by information gain, the top five attributes ranked by information gain, the top ten attributes ranked by gain ratio. and the top five attributes ranked by gain ratio. For each model, we also left all the hyperparameters at the default values given to them by WEKA for all of the runs. This ensured that equal opportunity for success was given to each learner across all the conditions.

6. Analysis of Results

Tables 5-7 display the final results and statistical analysis. Table 5 shows the accuracy of each classifier on the data given each condition. With all 21 attributes available to the learners, many of them, including random forest, J48, NN, and JRIP performed exceptionally well (> 99%). Naive Bayes, while still performing well, scores 92% accuracy, significantly worse than the other learners. When we limit the attributes to the top 10 attributes in Gain Ratio, most of the learners retain their accuracy dropping >1% in most cases. However, Naive Bayes loses nearly 10% accuracy when limited to the top 10 gain ratio attributes. Indeed, when the attributes are limited to the top 5 for gain ratio, there is a steep decrease in accuracy, with random forest and J48 producing 72%, NN classifying at 70%, JRIP at 66% and Naive Bayes at 60%.

In comparison, using the top attributes for Information Gain appears to produce more stable results. Using the top 10 attributes for information gain also produces 99% accuracy for all the learners other than Naive Bayes, which suffers a dropoff to just 82%. Furthermore, if we reduce the attributes to just the top 5 attributes for information gain, the accuracy of most of the learners is reduced only slightly, to 92%, with the exception of Naive Bayes, which is reduced to 71%.

Two tailed t-tests were performed to compare the performance of each learner across all conditions, and one-tailed t-tests were used to compare each experimental group to the control group. Table 6 shows us how all the classifiers in each experimental group compared to the average accuracy for classifiers in the control group. The smallest t score lies between the GainRatio top 5 attributes and all the attributes, indicating that the difference associated with these values is significant. All the other groups performed comparatively with the control group,

Learner	All Attributes	Top 10 Info Gain	Top 5 Info Gain	Top 10 Gain Ratio (without edibility)	Top 5 Gain Ratio
Random Forest	99.947	99.8073	92.5155	98.9597	72.5858
J48	99.7062	99.4558	95.545	99.1331	72.4943
Nearest-Neighbor	99.9181	99.8892	92.949	99.1909	70.7412
JRIP	99.7736	99.2631	92.8864	98.9212	66.1224
Naive Bayes	92.7804	82.7963	71.637	82.9601	60.5019

Table 5: Classifier Accuracy (%)

T-TESTS (one-tailed)	
All/IG10	0.1630491105
ALL/IG5	0.01821628289
ALL/GR10	0.1125209742
ALL/GR5	0.00001060850874

Table 6: One-tailed T-Test Results

Two-Tailed T-Test	
Forest/Tree	0.473578796
Forest/NN	0.6135706462
Forest/Rules	0.3454526532
Forest/Bayes	0.003297473037
Tree/NN	0.2964873717
Tree/Rule	0.202057831
Tree/Bayes	0.005730726249
NN/Rules	0.2605431596
NN/Bayes	0.004731207243
Rules/Bayes	0.01141163197

Table 7: Two-Tailed T-Test Results

7. Discussion of Future Work

Because there is a large amount of current research surrounding the classification of mushroom edibility, future research can aim to classify mushrooms on other attributes, as illustrated in our paper. Another possibility is for future research to aim at extending these data mining techniques to classification of species. There is also the opportunity for the use of collected data for prediction, instead of generated data.

References

Chai, X., Deng, L., Yang, Q., & Ling, C.X. (2004). Test-cost sensitive naive Bayes classification. *Fourth IEEE International Conference on Data Mining (ICDM'04)*, 51-58.

Chumuang, N., Sukkanchana, K., Ketcham, M., Yimyam, W., Chalermdit, J., Wittayakhom, N., & Pramkeaw, P. (2020, November). Mushroom Classification by Physical Characteristics by Technique of k-Nearest Neighbor. In *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1-6). IEEE.

Dawood, K. J., Zaqout, M. H., Salem, R. M., & Abu-Naser, S. S. (2020). Artificial Neural Network for Mushroom Prediction. *International Journal of Academic Information Systems Research (IJAISR)*, *4*(10).

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

Kuo, M. (2022). Key to Major Groups of Mushrooms. mushroomexpert.com. Retrieved October 21, 2022, from https://www.mushroomexpert.com/major_groups.html

Pinky, N. J., Islam, S. M., & Rafia, S. A. (2019). Edibility detection of mushroom using ensemble methods. *International Journal of Image, Graphics and Signal Processing*, *10*(4), 55.

Verma, S. K., & Dutta, M. (2018). Mushroom classification using ANN and ANFIS algorithm. *IOSR Journal of Engineering (IOSRJEN)*, *8*(01), 94-100.

Wibowo, A., Rahayu, Y., Riyanto, A., & Hidayatulloh, T. (2018, March). Classification algorithm for edible mushroom identification. In *2018 International Conference on Information and Communications Technology (ICOIACT)* (pp. 250-253). IEEE.

Wagner, D., Heider, D., & Hattab, G. (2021). Mushroom data creation, curation, and simulation to support classification tasks. *Scientific reports*, *11*(1), 1-12.