# Classification Algorithm for Edible Mushroom Identification

Agung Wibowo
Computer Science
STMIK Nusa Mandiri Sukabumi
Sukabumi, Indonesia
agung.awo@nusamandiri.ac.id

Yuri Rahayu
Computerized Accounting
AMIK BSI Sukabumi
Sukabumi, Indonesia
yuri.yru@bsi.ac.id

Andi Riyanto
Computerized Accounting
AMIK BSI Sukabumi
Sukabumi, Indonesia
andi.iio@bsi.ac.id

Taufik Hidayatulloh
Computer Science
AMIK BSI Jakarta
Jakarta, Indonesia
taufik.tho@bsi.ac.id

***Abstract*— Indonesia has 13% species of mushroom in the world but there is a very limited study on determining edible or poisonous mushroom. Classification process of poisonous mushroom or not will be easily conducted by learning machine using mining data as one of the ways to extract computer assisted knowledge. Currently, there are three comparisons of the best classification algorithms in data mining, namely: Decision Tree (C4.5), NaïveBayes and Support Vector Machine (SVM). The study method used is experiment with assisted tool of WEKA that has been testing in the comparison of the three algorithms. To conduct the testing, it is used the mushroom data of Agaricus and Lepiota family. The mushroom data were taken from The Audubon Society Field Guide to North American Mushrooms, in UCI machine learning repository. Results of the testing indicate that the C4.5 algorithm has the same accuracy level to the SVM by 100% however, from the speed aspect, process of the C4.5 algorithm is faster than the SVM.**

***Keywords*— Classification algorithm; Data mining; c4.5;Naive Bayes; Mushroom; Support Vector Machine.***

## I. INTRODUCTION

Mushroom is fleshy and edible fruit bodies of several species of fungi members of Basidiomycetes that usually grow in ground surface or substrate of other plants such as straw and wood [1] Indonesia is categorized as one of the agricultural countries and known as the warehouse of prominent mushroom in the world [2]. The number of species of mushroom that has been known until now is less than 69.000 out of the estimation of 1.500.000 species in the world and in Indonesia, there are less than 200.000 species [3]. This million species of mushroom, in general, can be divided into two types, namely edible and poisonous mushrooms. The Family of Agaricus and Lepiota wildly live in the open spaces; both with various shapes, colors, and characteristics which are not known by many people are poisonous. The Family of poisonous Agaricus and Lepiota can cause illness for one who consumes and also can cause death. The Family of Agaricus and Lepiota that are living wildly can be consumed and even used as medicines [4].

In Indonesia, until now (in 2017) based on the google search engine, it is found out two research publications in the field of edible and poisonous mushroom identification and classification. The mushroom identification method that has been done used Naïve Bayes and Voting Feature Interval(VFI5) algorithms with the prediction accuracy of 99,552% and 84.53% [4],[5]. Naïve Bayes is one of classification algorithm and classification is one of the major studies in data mining [6].

According to [7], he expressed that out of the three popular classification algorithms, these algorithms (C4.5, Naive Bayes, and SVM have been conducted the comparison test for intrusion detection [9]. The C4.5 Algorithm is the easiest algorithm to be implemented in the programming [7], and [8] said that c4.5 is the classification algorithm with the highest accuracy with the lowest level of error. This three are the best classification algorithm because of including in the top tenth the most influencing algorithm in data mining [7],[10]. From previous researchers, it can be seen a limited method in identifying the types of mushroom, one of the identification methods used is VF15 so that it is necessary for easier and more accurate alternative ways to identify a mushroom, namely algorithm comparison, and data mining.

The popular application of data testing for data mining in Indonesia is Rapid Miner or WEKA (Waikato Environment for Knowledge Analysis). This research is conducted by using an experimental method and assisted tool WEKA. This tool has a small size with faster loading time, with a simple interface and do not require many sources of data in the data processing compared to Rapid Miner. This research aims to find a feasible classification algorithm for edible mushroom

978-1-5386-0954-5/18/$31.00 ©2018 IEEE

with the best accuracy and easy to implement. Classification Algorithm

### A. C4.5 Algorithm

The C4.5 algorithm is the algorithm used to form decision tree. This algorithm is the very strong and popular classification and prediction methods. The decision tree method changes considerable facts into decision tree that represents rules. A decision tree model consists of a series of rules to divide the numbers of heterogeneous populations into smaller ones, more homogenous by considering the purpose of variables. The purposes of variables commonly are grouped definitely and the decision tree model more direct to probability calculation for each record on the categories or to classify record by grouping into one class [11]. There are two variables used to determine the root of a decision tree, namely entropy and gain values.

The entropy value is obtained from the formula:

$$Entrophy(S) = \sum_{i=1}^{n} -P_i * log_2 P_i \quad (1)$$

Notes :
S = group of cases
n = number of S partition
p = Proportion of Si on S

The gain value is obtained from the formula:

$$Gain(S, A) = entrophy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * entrophy(S_i) \quad (2)$$

Notes :
S = groups of cases A = Attributes
n = number of partition of A attributes
|Si | = number of case in the i partition
|S| = number of case in the S

Attributes with the highest gain value will be selected into the root from the decision tree. In general, to build the decision tree with the C4.5 algorithm will be in the following processes:

(1) Selecting attribute with the highest gain value as the tree root;
(2) Making the branches for each value;
(3) Dividing the cases in to the branches;
(4) Repeating the process for each branch for all cases in the branch having the same class.

The processes in the decision tree are changing the form of data into the three model that can be represented into the rules [11],[12].

### B. Naive Bayes algorithm

Naïve Bayes is statistic classification model that can be used to predict the probability of membership in a class. Naïve Bayes is based on the Bayes theorem having the similar classification capability with the Decision Tree and neural network [7]. Such Naïve Bayes is the simplification of the Bayes theorem. The following if the formula of Naive Bayes:

$$P(X|H)=P(H|X)P(X) \quad (3)$$

Notes :
X : data with unknown class
H : hypothesis of X data, is the specific class
P(H|X) : Probability of Hhypothesis based on X conditions X (posterior probability)
P(H) : probability of H hypothesis (prior probability)
P(X|H) : X probability based on the conditions of H hypothesis
P(X) : probability of X

### C. Support Vector Machine

Support Vector Machine (SVM) is the learning system which its classification uses hypothesis room in the form of linear function sine a feature space with high dimension. In the SVM concept, it seeks to find out the best separator function (hyperplane) among the limited functions. The best separator of Hyperplane among the two classes can be found out by measuring the hyperplane margin and find out the maximum points. The data in the divider field is called as support vector [5],[12]. In mathematic manner, the main concept of SVM is:

$$\min \frac{1}{2}|w|^2 \quad (4)$$

$$\text{s.t } y_i(x_i.w+b)-1 \geq 0 \quad (5)$$

In which $(x_i.w+b)\geq 1$ for class 1 and $(x_i.w+b)\leq -1$ for class 2, $x_i$ is set data and $y_i$ is the output from the $x_i$ data and w, bare the parameters that are looked for its values. The SVM optimization formulation for the classification cases of two classes is distinguish into linear and non-linear classifications.

## II. Research Methods

This research will be compare the result to determine which algorithm having the best accuracy by using mushroom dataset. The thinking framework is used as a reference in conducting this research and can be seen in Figure 1, as a proof of hypothesis that the C4.5 algorithm is the best algorithm for identification of edible mushroom compared to other classification algorithms that have been tested and/or used (C4.5, Naïve Bayes, VFI5). The dataset used is public data taken from The Audubon Society Field Guide to North American Mushrooms, contributed by Jeff Schlimmer at UCI (https://archive.ics.uci.edu/ml/datasets.html) where there are 8124 data and 22 variables with the nominal data type of 23 species of Agaricus and Lepiota family fungi.

The mushroom classification has 22 attributes that can affect on the classification whether it is edible or poisonous, namely: 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, habitat. The data testing uses machine with atom processor 1.66Ghz, memory of 1 GB, and operation system of Ubuntu 15.10.

## III. Result and Discussion

Results of the testing using evaluate on training data (Table 1) indicate that the C4.5 and SVM(SMO) algorithm shave better classification accuracy value that the Naïve Bayes algorithm. It is required a shorter time for algorithm process than two other algorithms that are tested in this research. The C4.5 algorithm uses shorter time by 0,18 second and faster time of 0,14 second that the SVM(SMO) algorithm and 0.16 second faster than the Naïve Bayes algorithm.

TABLE I. Results of the Testing Using Evaluate on Training Data

| Classifier output | Algorithm | | |
|---|---|---|---|
| | *C4.5(J48)* | *Naive Bayes* | *SVM (SMO)* |
| Correctly Classified Instances | 100% | 95.8887 % | 100% |
| Incorrectly Classified Instances | 0% | 4.1113 % | 0% |
| Kappa statistic | 1 | 0.9175 | 1 |
| Mean absolute error (MEA) | 0 | 0.0405 | 0 |
| Root mean squared error (RMSE) | 0 | 0.1718 | 0 |
| Relative absolute error (RAE) | 0% | 8.1092 % | 0% |
| Root relative squared error (RRSE) | 0% | 34.3742 % | 0% |

Data of testing results of training data is re-tested using 10-fold cross-validation for system testing or evaluation. Evaluation is intended to test the mining data application that we have created to obtain the percentage of system accuracy. 10-fold cross-validation process divides the dataset into ten-part, Nine parts will be used as training data and one part is used as data for testing. There are 10 training and testing processes to obtain the average accuracy. The results of the evaluation system using 10-fold cross-validation can be seen in Table 2. Classification accuracy using Naïve Bayes algorithm has increased classification error of 0.0615% which impact on the percentage of success classification although the process time has increased significantly.

TABLE II. Results of the Testing of 10-fold Cross-Validation

| Classifier output | Algorithm | | |
|---|---|---|---|
| | *C4.5(J48)* | *Naive Bayes* | *SVM (SMO)* |
| Correctly Classified Instances | 100% | 95.8272 % | 100% |
| Incorrectly Classified Instances | 0% | 4.1728 % | 0% |
| Kappa Statistic | 1 | 0.9162 | 1 |
| Mean Absolute Error (MEA) | 0 | 0.0419 | 0 |
| Root Mean Squared Error (RMSE) | 0 | 0.1757 | 0 |
| Relative Absolute Error (RAE) | 0% | 8.397 % | 0% |
| Root Relative Squared Error (RRSE) | 0% | 35.1617 % | 0% |

Results of the testing in the evaluation process of training data and 10-fold cross-validation of C4.5 and SVM (SMO) algorithm are still in the classification accuracy value by 100%, but in the process time aspect, the C4.5 algorithm is better 6.67 seconds faster than the SVM (SMO) algorithm. Moreover, this algorithm also produces a decision tree and eliminates the number of identification variables from 22 to 5 variables. A C4.5 decision tree for edible mushroom identification has five attributes: Odor, Spore-print-color, Gill-size, Gill-spacing, and Population, with this attributes, can be easily applied to computer programming.

## IV. Conclusion

Comparative classification algorithm testing accuracy in previous data mining has not been done and based on the results of testing of the three best classification algorithms in the data mining. The C4.5 algorithm has the highest accuracy compared to the other two popular classification algorithms, and in terms of processing speed. The decision tree generated by this algorithm can be easily applied to application creation and this algorithm also cuts the number of variables required for identification. For further research, researchers can develop the results of this research into a mobile application equipped with images that make it easier for people to recognize the edible wild mushrooms. Research on the identification of edible mushrooms also can be developed using image processing or compared to other classification algorithm.

## *Acknowledgment*

The authors would like to thank all those who have contributed to this paper. Thanks to The Audubon Society Field Guide to North American Mushrooms which has provided data for this paper.

## *References*

[1] Rial Adity and Setia Hadi Purwono, *Jamur – Info Lengkap dan Kiat Sukses Agribisnis*. Depok, Indonesia/West Java: Agriflo, 2012.

[2] Kristianus Sunarjon Dasa, "Pemanfaatan bagas sebagai campuran media pertumbuhan jamur tiram putih," vol. 11, pp. 195-201, 2011.

[3] Anna Rahkmawati, "Keanekaragaman jamur," Universitas Negeri Yogyakarta, Yogyakarta, Tech. rep. 2010. [Online]. staffnew.uny.ac.id/upload/132296143/pengabdian/ppm-2010-kehati.pdf

[4] Bayu Mahardika Putra, "Klasifikasi Jamur ke Dalam Kelas Dapat Dikonsumsi Atau Beracun Menggunakan Algoritma VFI 5 (Studi Kasus : Famili Agaricus dan Lepiota)," IPB, Bogor, Laporan Akhir 2008.

[5] Galieh Adi and Surya Pradana, "Identifikasi jamur beracun pada jenis jamur famili agaricus dan lepiota berdasarkan klasifikasi," Univeritas

Nusantara PGRI Kediri, Kediri, Laporan Akhir 2016.

[6] Sang Jun Lee and Keng Siau, "A review of data mining techniques," *Industrial Management*

[7] M. Adib Alkaromi, "Komparasi Algoritma Klasifikasi untuk dataset iris dengan rapid miner," *ICTech*, 2015.

[8] Diego R. Amancio et al., "A systematic comparison of supervised classifiers," *CoRR*, vol. abs/1311.0, 2013. [Online]. http://arxiv.org/abs/1311.0202

[9] Dwi Widiastuti, "Analisa Perbandingan Algoritma SVM, Naïve Bayes, dan Decission Tree dalam Mengklasifikasikan Serangan (Attack) pada Sistem Pendeteksi Intrusi," unpublished 2007.

[10] Xindong Wu et al., "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, pp. 1-37, 2008. [Online]. http://dx.doi.org/10.1007/s10115-007-0114-2

[11] Wenefirda Tulit Ina, "Klasifikasi Data Rekam Medis Berdasarkan Kode Penyakit Internasional Menggunakan Algoritma C4.5," *Jurnal Media Elektro*, vol. 1, pp. 105-110, 2013.

[12] Ni Wayan Sumartini Saraswati, "Naïve Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis," *Seminar Nasional Sistem Informasi Indonesia*, pp. 587-591, 2013.