JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MUSHROOM CLASSIFICATION: A COMPARISON OF CLASSIFICATION ALGORITHMS USING MACHINE LEARNING TECHNIQUES

Ms D Usha Rani
Department of Computer Science,
TBAK College for Women,
Kilakarai, Tamilnadu, India

Abstract : Mushrooms are used in the medical field to cure diseases like anemia, increase body immunity, diabetes, and cancer treatment. Some of the mushroom varieties are edible, but others are highly poisonous. The identification of mushrooms whether edible or poisonous is a difficult process because of the large number of mushrooms have similar characteristics. The principle of this paper is to classify the mushrooms by using Machine learning classification algorithms through a data mining tool. For the best classification, five classification algorithms were compared. The results showed that JRip classification is the best based on the difference between before and after applying data reduction in time taken.

Keywords: Data Mining, Classification, Data Pre-processing, Data Reduction, CFS subset evaluation, JRip, Decision Table, SGD, SMO, Logistic Regression.

I.INTRODUCTION

Data mining is also usually mentioned as knowledge discovery from Data (KDD). The aim of data mining is to mine useful and relevant information from huge databases or data warehouses. Knowledge discovery is a collaborative process, comprising of developing an understanding of the application domain, choosing and making a data set, pre-processing, data transformation.
Mushrooms are live in several habitats like above the ground, on the ground, or even on the plants such as deceased wood. It contains high protein, vitamins, minerals, and antioxidants. In the science field, it is one type of fungus. Mushrooms are the most sustainably produced food is not only taste but also have an excessive nutritional value. Mushrooms are also said to be a mediator to fight cancer cells and kill some types of viruses that carry infectious diseases such as polysaccharides, glycoproteins, and proteoglycans.

II. LITERATURE REVIEW

Narumol Chumuang, et al., in their paper compared seven classification algorithms for getting the highest accuracy rate. The algorithms are Naïve Bayes Multinomial Text,Naïve Bayes Updateable,Naïve Bayes,SGD Text,LEL,K-NN and Stacking. Finally, the K-NN classification algorithm shows a 100% accuracy rate.

Agung Wibowo,et al., in their paper they compared three classification algorithms for getting the highest accuracy rate. The algorithms are c 4.5, Naïve Bayes and SVM. Finally, c 4.5 classification algorithm shows 100% accuracy rate.

Kanchi Tank ., in their paper based on supervised learning algorithms such as Support Vector Machines(SVM), K-Nearest Neighbour (KNN), Naïve Bayes,Decision Trees (DT), Random Forest (RF) and Logistic Regression (LR). The result shows that Random Forest and K-Nearest Neighbour gave the highest test accuracy of 92.21% and 92.04%.

Shuhaida Ismail., in their paper presented the methodology and results for mushroom classification experiment based on their behavioral features such as characteristics, population, and habitat. Three classification algorithms were experimented with and compared. The algorithms are Decision Trees, K-Nearest Neighbour, and Naive Bayes. Principal Component Analysis (PCA) is identifying which attributes are important to classify the mushrooms. The decision Tree algorithm shows a 100% accuracy rate.

III. PROPOSED METHODS

Data Mining tools and Machine Learning techniques are used to convert raw data into some actionable, meaningful information. Three important phases are involved which are Data Pre-processing, Data Reduction and classification.

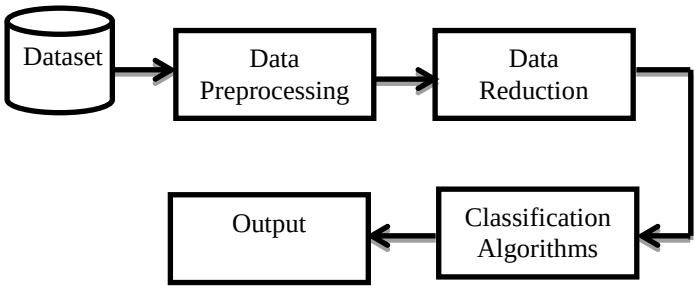

Figure 1: Methodology for mushroom classification

A. Data set

An openly available dataset is downloaded from the UCI Machine Learning Repository. It contains 23 attributes with 8124 instances of mushrooms.

Table 1: Mushroom Dataset (Attributes and Features)

S. No	Name of the Attributes	Name of the Features
	Cap Shape	Bell, Conical, Convex, Flat, Knobbed, Sunken
	Cap Surface	Fibrous, Grooves, Scaly, Smooth
	Cap Color	Brown, Buff, Cinnamon, Gray, Green, Pink, Purple, Red, White, Yellow
	Bruises	Yes, No
	Odor	Almond, Anise, Creosote, Fishy, Foul, Musty, None, Pungent, Spicy
	Gill Attachment	Attached, Descending, Free, Not attached
	Gill Spacing	Close, Crowded, Distant
	Gill Size	Broad, Narrow
	Gill Color	Black, Brown, Buff, Chocolate, Gray, Green, Orange, Pink, Purple, Red, White, Yellow
	Stalk Shape	Enlarging, Tapering
	Stalk Root	Bulbous, Club, Cup, Equal, Rhizomorphs, Rooted, Missing
	Stalk Surface above Ring	Fibrous, Scaly, Silky, Smooth
	Stalk Surface below Ring	Fibrous, Scaly, Silky, Smooth
	Stalk Color above Ring	Brown, Buff, Cinnamon, Gray, Orange, Pink, Red, White, Yellow
	Stalk Color below Ring	Brown, Buff, Cinnamon, Gray, Orange, Pink, Red, White, Yellow
	Veil Type	Partial, Universal
	Veil Color	Brown, Orange, White, Yellow
	Ring Number	None, One, Two
	Ring Type	Cobwebby, Evanescent, Flaring, Large, None, Pendant, Sheathing, Zone
	Spore Print Color	Black, Brown, Buff, Chocolate, Green, Orange Purple, White, Yellow
	Population	Abundant, Clustered, Numerous, Scattered, Several, Solitary
	Habitat	Grasses, Leaves, Meadows, Paths, Urban, Waste, Woods
	Class	Non-toxic, Toxic

B. Data Pre-processing

Data cleaning is one of the pre-processing steps that is used to prepare raw data for analysis by removing unwanted data or bad data and filling in the null values. ignore the row, Fill the missing value manually, Global constant value, Central tendency for the attribute, attribute mean or median value, most probable value is the steps to fill the missing values. Any one of the steps is enough to remove unwanted data. Smoothing by bin means, bin medians, bin boundaries, Regression, and Outlier analysis are the steps to remove the noisy data. In this dataset stalk- root has 31% of missing values. Maximum missing percentage value will affect the performance. so applied unsupervised Replace missing values to this attribute. After applying this filter, it will show 0% of the missing value.

C. Data Reduction

Data reduction is a process that reduced the original data set that is much smaller in volume and also maintains the integrity of the original data. It does not affect the result obtained from data mining before data reduction and after data reduction is the same. Data reduction increases the efficiency of data mining.

Correlation-based Feature Selection (CFS) is one of the data reduction techniques. It is used to reduce the number of variables or attributes for analysis in data set by extracting needed attributes from a large pool. It is used to improve the performance of an algorithm and improve visualization. After applying this feature selection, it will select Odor, Gill-Spacing, Stalk-surface-above-ring, and Veil-color attributes.

Table 2: Mushroom Dataset (After Applying Data Reduction)

S. No	Attributes
	Odor
	Gill-Spacing
	Stalk-surface-above-ring
	Veil-color

D. Classification Algorithms

i) Decision Table

A Decision Table is a supervised learning method that can be used for both classification and Regression problems, but mostly it is mostly chosen for solving Classification problems. It is a tree-structured classifier, where internal nodes denote the features of a dataset, branches denote the conclusion rules and each leaf node denotes the outcome. It is also called a decision tree, because looks like a tree, it begins with the root node, which enlarges on further branches, and finally constructs a tree-like structure. In a Decision tree, there are two nodes are used, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have numerous branches, whereas Leaf nodes are the output of the decisions and do not contain any extra branches. The decisions or the test are performed on the source of features of the given dataset. It is a graphical representation for getting all the probable solutions to a problem or decision based on specified conditions.

ii) JRip

The RIPPER (repeated incremental pruning to produce error reduction) algorithm is upgraded upon IREP to generate rules that match or exceed the performance of decision trees. Having progressed from numerous iterations of the rule learning algorithm, the RIPPER algorithm has a three-step process. Grow, Prune and Optimize. The first step uses a ‘separate and conquers method to include conditions to a rule until it correctly classifies as a subset of data. Like decision trees, the information gain condition is used to identify the next splitting attribute. When increasing a rule’s specificity no longer decreases entropy, the rule is instantly pruned. Until reaching the stopping condition step one and two are repeated at which point the entire set of rules is optimized using a variety of heuristics.

iii) Logistic Regression

Logistic regression is one of the Supervised Learning Methods in Machine Learning algorithms. It is mostly used to predict the dependent variable using a given set of independent variables. The values may be either Yes or No, 0 or 1, true or False, etc. it provides the probabilistic values between 0 and 1. Linear Regression is used to solve Regression problems, whereas Logistic regression is used to solve the classification problems. The sigmoid function is a mathematical function used to map the predicted values to probabilities. It plots any real value into alternative value within a range from 0 to 1. so it makes a curve like the "S" form. The S-form curve is also called the sigmoid function or the logistic function.

iv) SGD

The word ‘stochastic‘ defines a system or a process that is connected with a random probability. In Stochastic Gradient Descent methods, in each iteration, some of the samples are carefully chosen randomly instead of considering the whole data set. Although using the whole dataset is actually useful for getting to the minima in a less noisy and less random manner, the problem arises when our dataset gets big. It is computationally much less expensive than typical Gradient Descent. Hence, in most situations, SGD is chosen over Batch Gradient Descent for optimizing a learning algorithm.

v) SMO: Sequential Minimal Optimization

SMO breaks the large quadratic programming problem into a sequence of the smallest possible quadratic programming problems. These small quadratic programming problems are solved systematically, which avoids by means of a time-consuming numerical quadratic programming optimization as an inner loop. The total of memory required for SMO is direct in the training set size, which allows SMO to handle very large training sets. Because matrix calculation is omitted, SMO procedures between linear and quadratic in the training set size for several test problems. SMO's calculation time is controlled by SVM calculation; SMO is fastest for linear SVMs and sparse data sets.

IV. RESULT AND DISCUSSION

The classification of this dataset was performed to classify the mushrooms whether edible or poisonous. The results are taken from the confusion matrix.

Table 3: Comparisons of Classification Algorithms

Total Number of Instance:8124				Difference Between Before and After applying Data Reduction
S.No	Classification Algorithms	Time Taken (in Seconds)		
		Before Data Reduction Process	After Data Reduction Process	
	Decision Table	0.52	0.03	0.49
	JRip	0.30	0.17	**0.13**
	Logistic Regression	1.41	0.05	1.36
	SGD	1.06	0.20	0.86
	SMO	0.93	0.27	0.66

Before applying the process of data reduction, the JRip algorithm took less time and after the data reduction process, Decision Table algorithm took less time. The results from the above mentioned table JRip classification algorithm is the best based on the difference between data reduction process in time taken.

Finally, I conclude from Table 1, based on the difference between before and after applying data reduction JRip (0.13) is the best classification algorithm.

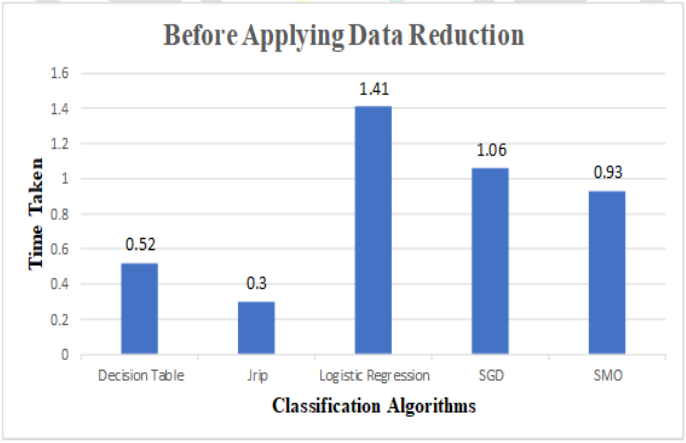

Figure 2: Bar chart representation of classification algorithms (Before Applying Data Reduction)

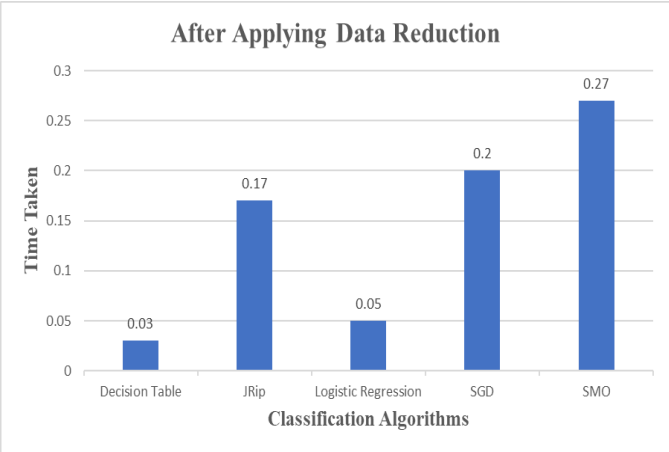

Figure 3: Bar chart representation of classification algorithms (After Applying Data Reduction)

V. CONCLUSION

This paperwork particularly focuses to classify the mushroom whether edible or poisonous because it contains fiber, protein, and antioxidants. The mushroom dataset consists of 23 attributes and 8124 instances. All of them are divided into two classes edible and poisonous. So, the classification of mushrooms is very important. Finally, I conclude JRip is the best algorithm based on processing time for this dataset. In the future, my research work concentrates on developing a mobile application with image processing.

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] F.Y, Osisanwo, Akinsola J.E.T, Awodele O,Hinmikaiye J. O, Olakanmi O, and Akinjobi J."Supervised Machine Learning Algorithms:Classification and Comparison." International Journal of Computer Trends and Technology 48, no. 3 (June 25, 2017): 128–38. https://doi.org/10.14445/22312803/ijctt-v48p126.
[6] Firdous, DrxHina. "Health Benefits Of Mushroom, Uses And Its Side Effects."Lybrate, 2020. https://www.lybrate.com/topic/benefits-of-mushroom-and-its-side-effects (accessed Sep.24, 2020).
[7] E. Yukselturk, S. Ozekes, Y. K. Turel, "Prediction Dropout Student: An Application of Data Mining Methods in An Online Education Program", European Journal of Open, Distance and e-Learning, vol. 17, no. 1, pp.118 – 133, 2014.
[8] X. Y. Ren, Y. Zhao, W. Zeiler, G. Boxem, T. Li, "A Data Mining approach to analyze Occupant Behaviour Motivation", Procedia Engineering, vol 205, pp 2442-2448, 2017.
[9] Zhao, C. and Luan, J. (2006). Data mining: Going beyond traditional statistics. In New Directions for Institutional Research, 131(2) , (pp. 7 16).
[10] L.Arockiam, S.Charles,Arulkumar et.al(2010), "Deriving Association between Urban and Rural Students Programming Skills", International Journal on Computer Science and Engineering Vol. 02, No. 03, pp 687 - 690.