Classification of Mushrooms using Supervised Learning Models

Balika J. Chelliah
Assistant Professor(S.G), Department of Computer Science, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India-600089

S. Kalaiarasi
Assistant Professor(O.G), Department of Computer Science, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India-600089

Apoorva Anand
Undergraduate, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India-600089

Janakiram G
Undergraduate, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India-600089

Bhaghi Rathi
Undergraduate, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India-600089

Nakul K. Warrier
Undergraduate, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India-600089

Abstract – Mushroom hunting is becoming increasingly popular as a leisure activity, it is paramount that we have some way of classifying them as poisonous or non-poisonous. Using supervised machine learning models on the dataset that UCI makes available of various characteristics of mushrooms, we can get a prediction system that can classify mushrooms. The inspiration of this project is to understand which machine learning models work best on the dataset and which features are most indicative of poisonous mushrooms.

Index Terms – learning, Classification, Support vector machines, Logistic regression, Gaussian naive bayes, Random forest classifier, Decision trees

1. INTRODUCTION

Machine Learning is a hypnotizing point, it empowers machines to make sense of how to accomplish errands that irrefutably required a man to do. While machine learning used to require the most serious supercomputers just a few years earlier, the section of circulated registering, more affordable CPUs, and much better estimations allowed Machine Learning to end up open more broadly.

Hunting of different species of mushrooms is acknowledging new tops in noticeable quality, it is essential that we have some technique for gathering them as poisonous and non-poisonous. Using coordinated machine learning models, we can get a genuinely correct gauge system that can classify mushrooms. The inspiration for this endeavor is to fathom which machine learning models work best on the dataset, and which features are most unique of harmful mushrooms.

The dataset consolidates depictions of theoretical illustrations contrasting with 23 sorts of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each specie is recognized as undeniably satisfactory, unquestionably unsafe, or of darken edibility and not recommended.

This last class was joined with the unsafe one. The Guide clearly communicates that there is no direct control for choosing the edibility of a mushroom; no administer like "gifts three, let it be" for Poisonous Oak and Ivy.

In general, various machine learning techniques are used to analyze the features obtained from the datasets. So, in another way we can say that a machine learning is a system or model that takes the documents, analyze the input, and learns automatically with the involvement of any human. They have the ability to understand how to adjust the methods of learning according to some actions.

2. RELATED WORK

In order to identify and classify mushrooms as poisonous and non-poisonous, we need to understand the fundamentals of each component that are going to be used. To give the reader a wider scope of this domain, we have a referred several papers which address the same problem.

1. Concept acquisition through representational adjustment

- J Schlimmer

The is a thesis promotes the hypothesis that the necessary abstractions can be learned. The specific task studied is inducing a concept description from examples. A model is presented that relies on a weighted, symbolic description of concepts. The model extends previous work by allowing for noisy examples, unknown values, and concept change over time. Key results illustrate that the model should scale-up to larger tasks than those studied and have a number of potential applications.

2. Trading of simplicity and coverage in incremental concept learning

- W Iba, J Wogulis, P Langley

The paper presents an incremental learning method called HILLARY that addresses several of the more difficult aspects of learning from examples. Specifically, HILLARY employs 'hill climbing' to incrementally learn disjunctive concepts from noisy data in either a relational or at tribute-value representation. This paper discusses HILLARY's learning algorithm, tradeoff, and evaluation function, and we present empirical studies of the system's learning behavior on both natural and artificial domains. This paper shows that HILLARY's performance will deteriorate linearly with the amount of noise, independent of the memory limitations.

3. Extraction of logical rules from training data using back propagation networks

- W Duch, R Adamczak, K Grabczewski

The paper gives an overview of techniques developed to redress the full potential of trained artificial neural networks (ANNs). Essentially, this paper focuses on mechanisms, procedures, and algorithms designed to insert knowledge into ANNs, extract rules from trained ANNs, and utilise ANNs to refine existing rule bases. This paper also introduces a new taxonomy for classifying the various techniques and delineates criteria for evaluating their efficiency.

4. Extraction of crisp logical rules using constrained backpropagation networks

- Wlodzislaw Duch , Rafal Adamczak, Krzysztof Grabczewski, Masumi Ishikawa, Hiroki Ueda

The paper compares neural networks and back propagation algorithms. By adding regularization terms to the error function, these two algorithms introduce barriers to the structure of the network. Networks with a very less number of connections are created, leading to a small number of crisp logical rules. These two algorithms work on the mushroom classification data to output the optimal logical response. One, by extraction of logical rules from the sample data to acquire knowledge is an important and difficult problem in computational intelligence. Two, neural networks, in particular multi-layered perceptrons, are useful classifiers that can learn arbitrary vector mappings from the input to the output space and successfully use this mapping in novel situations.

3. PROPOSED MODELLING

The proposed model is to find the solution for the above-mentioned problems.

- Our aim is to recreate all the major machine learning algorithms to compare them for inconsistencies, speed and accuracy.
- By looking at all models, we hope to achieve superior knowledge on which learning models fit the problem to be solved.

4. IMPLEMENTATION

The implementation basically consists of these following modules:

- Linear Regression

In measurements, coordinate backslide is a straight approach for exhibiting the association between a scalar ward variable y and no less than one illustrative factors (or free factors) showed X. In straight backslide, the associations are shown using direct pointer works whose dark model parameters are assessed from the data.

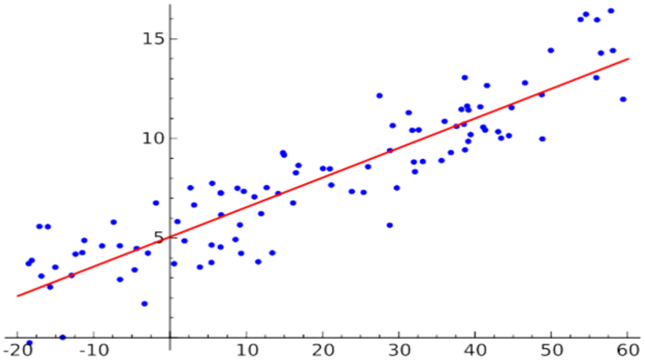

Fig.1 An example of a Best-Fit Line

- Logistic Regression

In insights, strategic relapse, or logit relapse, or logit demonstrate is a relapse show where the needy varIn bits of knowledge, key backslide, or logit backslide, or logit exhibit is a backslide indicate where the poor variable (DV) is full scale.

The double strategic model is utilized to check the likelihood of a coordinated reaction in light of no short of what one marker (or free) factors (highlights).iable (DV) is all out.

The binary logistic model is used to check the probability of a matched response in light of no less than one marker (or free) factors (features).

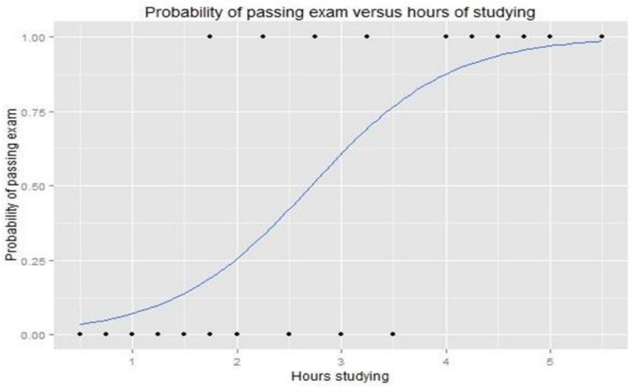

Fig.2 Demonstration of Logistic Regression in Action

- Gaussian Naive-Bayes

In machine learning, credulous Bayes classifiers are a gathering of direct probabilistic classifiers in perspective of applying Bayes' hypothesis with strong (innocent) self-sufficiency doubts between the features. Credulous Bayes classifiers are significantly flexible, requiring different parameters coordinate in the amount of variables (features/pointers) in a learning issue.

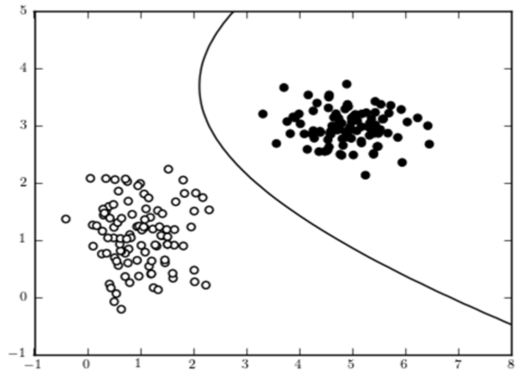

Fig.3 Simple Gaussian Naive Bayes Classification

- Random Forest Classifier

Irregular woods or arbitrary choice woodlands are an outfit learning methodology for request, backslide and distinctive assignments, that work by building countless trees at planning time and yielding the class that is the technique for the classes (game plan) or mean desire (backslide) of the individual trees.

Woodlands of trees part with calculated hyperplanes, if arbitrarily constrained to be unstable to simply picked feature estimations, can get accuracy as they create without torment from overtraining.

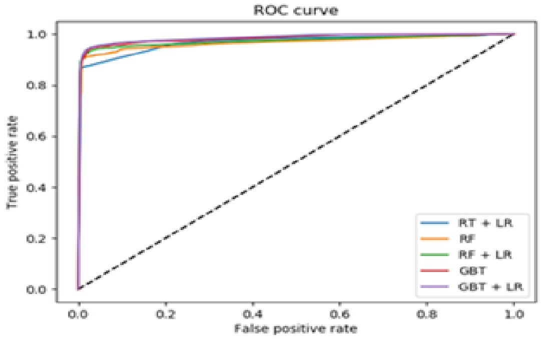

Fig.4 Receiver Operating Characteristic curve for random forest classifier

- Support Vector Machine

In machine learning, bolster vector machines are coordinated learning models with related learning counts that separate data used for course of action and backslide examination.

Given a game plan of getting ready delineations, each set apart as having a place with both of two classes, a SVM planning count makes a model that doles out new cases to one grouping or the other, making it a non-probabilistic twofold straight classifier (regardless of the way that strategies, for instance, Platt scaling exist to use SVM in a probabilistic request setting).

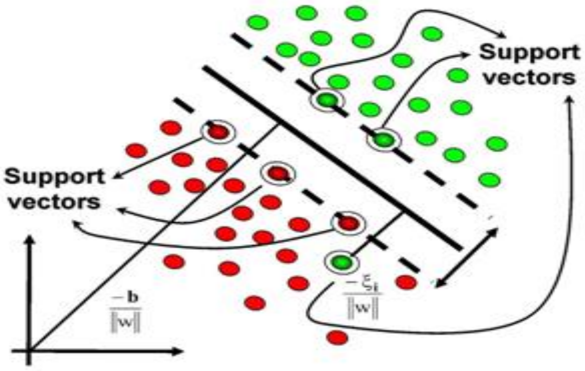

Fig.5 Support vector machine model

- Decision Trees

A tree has various analogies, everything considered, and turns out that it has affected a wide zone of machine getting the hang of, covering both classification and backslide. In decision examination, a decision tree can be used to ostensibly and unequivocally address decisions and essential authority. As the name goes, it uses a tree-like model of decisions. Regardless of the way that a for the most part used instrument in data burrowing for construing a system to accomplish a particular target, its similarly extensively used as a piece of machine acknowledging, which will be the guideline point of convergence of this article.

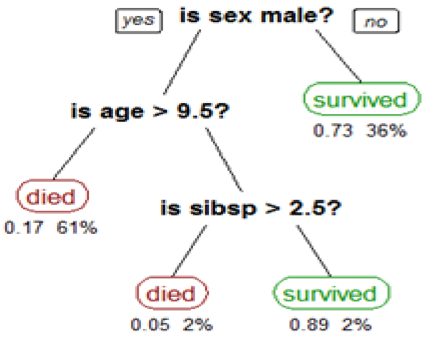

Fig.6 Example of a decision tree model

5. RESULTS AND DISCUSSIONS

After running the various learning models on our dataset to determine poisonousness of mushrooms, we learn that while all the learning models are fairly decent, some exceptionally good, the default decision tree model wins out in accuracy and reliability.

6. CONCLUSION

This comparative study has allowed us to gain greater insight into the inner workings of the major machine learning models and how they play out in real world situations. The conclusion can be scaled and applied to countless other problems of the same nature and happens to be a valuable tool for any data scientist.

REFERENCES

[1] J.R. Sobehart, R.M. Stein, V. Mikityanskaya, and L. Li. Moody's public firm risk model: a hybrid approach to modeling short term default risk. Technical Report, Moody's Investors Service, Global Credit Research. Available electronically at http://www.moodysqra.com/research/crm/53853.asp, 2000.

[2] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: bagging, boosting and variants. Machine Learning, 36:105–142, 1999.

[3] Blake and C.J. Merz. UCI repository of machine learning databases. Technical Report, University of California, Irvine. Available electronically at http://www.ics.uci.edu/~mlearn/MLRepository.html, 2000.

[4] K.P. Burnham and D.R. Anderson. Model Selection and Multimodel Inference: A Practical Information–Theoretic Approach, 2nd. ed., Springer–Verlag, New York, 2002.

[5] C.J. Fombrun, N. Gardberg, and J. Sever. The reputation quotient: a multi–stakeholder measure of corporate reputation. Journal of Brand Management, 7:241–255, 2000.

[6] D.J. Hand. Construction and Assessment of Classification Rules, John Wiley and Sons, Chichester, 1997.

[7] C. Harris–Jones and T.L. Haines. Sample size and misclassification: is more always better? AMSCAT–WP–97–118, AMS Center for Advanced Technologies, 1997.

[8] T.S. Jaakkola and M.I. Jordan. Bayesian logistic regression: a variational approach. Statistics and Computing, 10:25–37, 2000.

[9] Langley, Pat. The changing science of machine learning. Machine Learning, 82:275–279, 2011.

[10] van Iterson, M., van Haagen, H.H.B.M., and Goeman, J.J. Resolving confusion of tongues in statistics and machine learning: A primer for biologists and bioinformaticians. Proteomics, 12:543–549, 2012.

[11] van Rijsbergen, C. J. Information Retrieval. Butterworth, 2nd edition, 1979.

[12] Buehler, Martin, Iagnemma, Karl, and Singh, Sanjiv (eds.). The 2005 DARPA Grand Challenge: The Great Robot Race. Springer, 2007.

[13] Zdziarski, Jonathan A. Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification. No Starch Press, San Francisco, 2005.

[14] Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143:29–36, 1982.

[15] Bennett, James and Lanning, Stan. The Netflix Prize. In Proc. of KDD Cup and Workshop, pp. 3–6, 2007.