International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 01 | Jan -2017 www.irjet.net p-ISSN: 2395-0072

An Empirical Study on Mushroom Disease Diagnosis: A Data Mining Approach

Dr. Dilip Roy Chowdhury[1], Subhashis Ojha[2]

[1]*Assistant Professor, Dept. of Computer Science & Application, University of North Bengal, West Bengal, India*

[2]*Assistant Teacher, Mathurapur B.S.S. High School, Malda, West Bengal, India*

Abstract - *In this paper, a data mining application is introduced for selecting highly effective factors or symptom of different disease diagnosis in mushroom yield. The study also focuses on several factors causing a specific mushroom disease. Highly potential symptoms among several factors were focused out for better management in this regard. That's why data mining techniques are being used for ranking among symptoms. This paper focuses on identifying specific diseases among several diseases using a data mining classification based approaches. Real data has been taken from mushroom farm and thereafter purification of potential factors is done through data mining approaches. The classification technique and disease prediction of mushroom dataset were prepared using Naïve Bayes, SMO and RIDOR algorithms. A statistical comparison has been produced in order to find the best symptoms needed for mushroom disease diagnosis. Besides this, it searches the best performing classification algorithm among all.*

Key Words: Classification Algorithm, Data Mining, Mushroom, Ripple Down Rules (RIDOR), Sequential Minimal Optimization (SMO).

1. INTRODUCTION

Mushroom is one type of fungus type plant containing no chlorophyll. There is around 45000 type fungus available in the world. Among them, around 2000 fungus are edible vegetable food. Mushroom can be edible and non-edible. Cultivating mushroom in scientific ways reduces the probability to occur poison in mushroom yield. In our country, four kinds of mushroom are available namely Button Mushroom, Oyster Mushroom, Paddy Straw Mushroom, Milky Mushroom. Like all other crops, a huge number of biotic and abiotic agents or factors affect mushroom yields. Among the biotic agents, fungi, bacteria, viruses, nematodes, insects and mites are responsible for damaging in mushrooms directly or indirectly. A number of harmful fungi are seen in compost and casing soil during the cultivation of mushroom. Many of these work as competitor molds thereby adversely affecting spawn run. On the other hand the fruit body at various stages of crop growth is hampered by others, thus producing distinct disease symptoms. If this continues, total crop failures occur based on the stage of infection, quality of compost and environmental conditions. At any phase of growth an undesirable growth or development of certain molds can occur and can adversely affect the final mushroom yield [1]. Finding the diseases and its reason in context to mushroom is very much necessary where there is a scarcity of domain expert on this field.

Data mining is a process to get the meaningful data from the large data scattered in large data repository. By using different tools and techniques, right information is provided for data mining process. It is popularly known as information or knowledge discovery, which is one of the recent trends found to be useful in several complex fields. It is the process of evaluating data from different outlooks and summarizing it into useful information that can be used to identify the symptom of different diseases in mushroom. Using data mining mushroom data set can be analyzed. Data mining allows users to investigate data from many different dimensions or angles, categorize it, and summarize the relationships identified [2]. Technically, data mining is used for set up relationship among several fields in dataset. The objective of application of data mining techniques on mushroom dataset is to analyze such data and to find relative importance of different disease parameters for differential diagnosis in mushroom. A specific disease diagnosis is not a result of one deciding factor, in addition it heavily hinges on multiple symptoms.

This paper identifies the symptoms associated with mushroom disease that help to the farmers/growers to improve the quality of production as well as to cope up the huge losses for damaging. This study reveals the accuracy of some classification techniques has been measured and relative importance has been found among several disease symptoms of mushroom. The main objective of this proposed work is to find out differential diseases diagnosis and identification of the highly importance attribute for disease diagnosis and also to find the optimal classification mechanism.

2. LITERATURE REVIEW

Data mining applications are being recently used in agricultural research and many activities can be observed in

this field. The data mining techniques used in agriculture for prediction of problems, disease detections, optimizing the pesticide and so on. It is capable to provide a lot of information on agricultural-related activities, which can then be analyzed in order to find important information and to collect relevant information. The data mining techniques are used for disease detection, pattern recognition by using multiple application in this study[2]. In an another study, Ahsan Abdullah, Stephen Brobst, Ijaz Pervaizhave shown that how data mining integrated agricultural data including pest scouting, pesticide usage and meteorological recordings is useful for optimization (and reduction) of pesticide usage[3]. Jyotshna Solanki, Prof. (Dr.) Yusuf Mulge discussed about the techniques used in agriculture for exacting the meaningful data for relevant information. Cunningham, S. J., and Holmes worked on a classification system capable of sorting mushrooms into quality grades and achieving an accuracy similar to that attained by human inspectors [4]. In another study, D Ramesh, B Vishnu Vardhan describe certain Data Mining techniques adopted in order to estimate crop yield analysis with existing data [5].

As per as this study is concern, it is found that, till now work has been done on the mushroom to find out the mushroom which is edible or not. No such cases found which works on the disease of the mushroom cultivation and their management. This study actually focuses on this area to diagnosis the mushroom disease using classification techniques of Data Mining.

3. DATA MINING CLASSIFICATION ALGORITHMS

3.1. NAÏVE BAYES

Naïve Bayesian classification is a supervised learning process. In addition, it is a statistical method for classification purposes. This classification is based on Bayesian theorem. Bayes Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved, in this sense, is considered "naïve" [6]. It is looked attractive on when the dimensionality of the supplied inputs is high. The process of maximum likelihood is being used for parameter estimation in naïve Bayesian models. Let D be a training set of tuples and their associated class labels. The representation of every tuple is being done by an *n*-dimensional attribute vector, X= (X_1,X_2,X_3,....,X_n). Let there are m Classes C= (C_1,C_2,C_3,.....,C_m). Under given dataset, the Naïve Bayesian classifiers will predict that given a tuple *X* belongs to the class having the highest posterior probability, conditioned on *X*. That is, the na¨ıve Bayesian classifier predicts that tuple ***X*** belongs to the class C_i if and only if,

$P(C_i / X) > P(C_j / X)$ for $1 <= j <= m , j \neq i$

Thus, we maximize $P(C_i / X)$. The class C_i for which $P(C_i / X)$ is maximized is called the maximum posteriori hypothesis[6]. By Bayes' theorem

$P(C_i / X) = P(X / C_i) P(C_i) / P(X)$.

Only $P(X / C_i) P(C_i)$ is maximized since P(X) is constant. Under given data sets with many attributes, it would be extremely computationally expensive to evaluate $P(X / C_i)$. For reduction of computation, the naïve assumption of class-conditional independence is made.

$$P(X / C_i) = \prod_{k=1}^{n} P(x_k / C_i)$$

$$= P(x_1/C_i) \times P(x_2/C_i) \times P(x_3/C_i) \times \ldots\ldots\ldots P(x_n / C_i)$$

Bayesian classifiers have the minimum error rate in comparison to all other classifiers [6].

3.2. RIPPLE -DOWN RULE LEARNER (RIDOR)

Ripple-Down Rules devised by Prof. Paul Compton is a strategy of building system incrementally while they are already in use [7]. When a system is not able to handle a case or situation appropriately, there is a need to alter the system in such a way that the previously acquired knowledge of the system is not degraded, just revised where necessary. The alteration is made simply and rapidly and the difficulty of making a change should not increase as the system develops. RDR can be categorized as a type of apprentice learning [7]. The following formulation of a ripple-down rules is shown below:

if *condition* **then** *conclusion* **[because** *case***] except**

if ....

else if ...

If a rule fires, but it generates an incorrect conclusion, RDR has a fascination to add an except branch. Besides this, one condition is also kept in knowledge that if a rule fails to fire when it should, an else branch is added. In RDR, the appropriate condition is chosen in such a way that the new rule to be inserted is as general as possible for diagnosing correctly the new case. An RDR keeps track those cases which make a new rule create.

3.3. SEQUENTIAL MINIMAL OPTIMIZATION (SMO)

It is the extension of Support Vector Machines (SVM) is a method for the classification of both linear and nonlinear data [6]. Extensions of basic SVM algorithm such as the sequential Minimal optimization developed by John C. Platt of Microsoft research, which is utilized in this paper, implemented in WEKA can be used to train SVM faster, better group samples clusters. The goal of SVM is to search

the linear optimal separating hyperplane (i.e. " decision boundary") that can separate two classes with the largest distance (i.e. "gap" or "margin") within a border line (support vectors) [8]. If data are linearly inseparable, original input data is transformed into a higher dimensional space with the help of a nonlinear mapping constructed through mathematical projection ("kernel trick") where separating decision surface is found. After that, a linear separating hyperplane is searched in new space. The maximal marginal hyperplane found in the new space corresponds to a nonlinear separating hypersurface in the original space [6]. The training time of SVM might be slow but because of their ability to model complex nonlinear decision boundaries it is accurate to learn both simple and high complex classification models, and avoids over fitting by using complex mathematical principles.

Sequential Minimal Optimization (SMO) algorithm which is the new efficient technique for training SVMs. SMO breaks the very large quadratic programming (QP) optimization problem occurred in SVM training into a sequence of minimal possible QP problems involving only two variables, and each of these problems is solved analytically. SMO heuristically selects a pair of variables for each problem and optimizes them. This procedure repeats until all the patterns satisfy the optimality conditions [9]. The SMO algorithm needs less amount of memory, thus very large SVM training problem can accommodate in the memory of a personal computer, as a result, large matrix computation is avoided [10]. SMO algorithm selects two Lagrange multipliers α_1 and α_2 and optimizes the objective value for both these α's. Finally it adjusts the b parameter based on the new α's. This process is repeated until the α's converge [11]. SMO update two Lagrange multipliers as a SMO Step as shown below:

Given two examples E1 and E2:

$$\alpha_2^{new} = \alpha_2 + \frac{y_2(E_2 - E_1)}{\eta}$$

Where $\eta = K(\vec{x}_1, \vec{x}_1) + K(\vec{x}_2, \vec{x}_2) - 2K(\vec{x}_1, \vec{x}_2)$

Clips the value at the end of the segment:

$$\alpha_2^{new,clipped} = \begin{cases} H & if \quad \alpha_2^{new} \geq H; \\ \alpha_2^{new} & if \ L < \alpha_2^{new} < H; \\ L & if \quad \alpha_2^{new} \leq L; \end{cases}$$

if $y_1 = y_2$ then:

$L = \max(0, \alpha_2 + \alpha_1 - C)$

$H = \min(C, \alpha_2 + \alpha_1)$

Otherwise:

$L = \max(0, \alpha_2 - \alpha_1)$

$H = \min(C, C + \alpha_2 - \alpha_1)$

$$\alpha_1^{new} = \alpha_1 + s(\alpha_2 - \alpha_2^{new,clipped})$$

Where $s = y_1 y_2$

Major components of SMO are an analytical method to solve for two Lagrange multipliers.

4. PROPOSED METHODOLOGY

4.1. EXPERIMENTAL PROGRESSION

A survey cum experimental methodology is being put to some diagnosis purposes. Through extensive finding of the literature and discussion with experts as well as mushroom growers about differential diseases occurred in mushroom, a number of factors/symptoms/attributes considered for diagnosing mushroom diseases are identified. These influencing attributes are being categorized as input variables. For this purpose, real data is collected from mushroom farm. These collected data is then sorted out using manual techniques. Then data is converted into an appropriate format required by the processing. After that, features and parameters selection (i.e. attributes selection) is identified. The above mentioned process is shown below:

Step1: Generation of Mushroom dataset from real field as well as mushroom growers.
Step2: Possible values and relevant variables are considered.
Step3: Input data is to be converted into appropriate file format.
Step4: Produced file is fed into the Evaluator
Step5: Recognition of High influential Variables is performed.
Step6: Classification Algorithms are being applied and comparison of output is taken.

Then analysis of attributes and implementation is performed using the data mining tool. After implementation, results are generated and analyzed. Stepwise description of methodology used is depicted above. This study considers sixteen different diseases namely Green Mould, False Truffle, Brown Plaster, Inky Caps, Cinnamon Mould, Wet Bubble, Dry Bubble, Cobweb, Olive Green Mould, Yellow Mould, Sepedonium Yellow Mould, Lipstick Mould, Lilliputia Mould, Pink Mould, Oedocephalum Mould, White Plaster Mould. Each object has 18 attributes or symptoms (factors). Here, 110 data records of mushroom are being used for analyzing for decision making.

4.2. THE DATASET

The dataset for the Mushroom has been accumulated from expert as well as farmer. The Mushroom dataset consists of 110 instances and 19 attributes including classification attribute. Table1. shows the attribute data set for the study.

Using cross-validation parameter the evaluation has been performed to observe the performance of the three classification algorithms. It has been done to confirm the best technique for the Mushroom Diseases Diagnosis with the help of performance factors or criteria.

This study has been focused on such criteria's like Accuracy, Attribute Selected Classifier, Kappa Statistic, Confusion Matrix, True Positive(TP) Rate, False Positive(FP) Rate, Precision, Recall, F-Measure, Receiver Operating Characteristics(ROC) Area, and Error rate such as Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Root Relative Squared Error.

Table -1: Attribute Data Set

Attribute Description	Attribute Type	Number of Attribute Values
Effect_Of_Damage	Nominal	16
Effect_on_Cap	Nominal	6
Mycelium_or_Patch_Color	Nominal	11
Compost_And_Casing_Soil_Color	Nominal	11
Odour	Nominal	6
Habitat	Nominal	10
Liquid_Exudate	Nominal	3
Use_damp_wood	Nominal	2
erheated_compost	Nominal	3
Chicken_Or_Horse_Manure_Mixer_In_Compost	Nominal	2
Excessive_Ammonia_Or_Nitrogen_In_Compost	Nominal	2
Peat_In_Casing	Nominal	2
Improper_Phase_II_Composting	Nominal	2
Dry_and_leathery	Nominal	2
High_Humidity	Nominal	2
Lack_Of_Ventilation	Nominal	2
Mycelium_tends_to_be_thicker_ropes	Nominal	2
Patches_on_compost_during_cool_down	Nominal	2
Disease_differential	Nominal	16

In this paper, selection of high influential attributes by using select attributes facility using WEKA, a Data Mining Tool is performed. For attribute evaluation, Information Gain attribute evaluators are used.

5. STATISTICAL ANALYSIS FOR DECISION MAKING

The study has been used 10 fold Cross validation techniques for classifying the entire data set. The Weka classifier takes 110 labeled data. Then it produces 10 equal sized sets. Each set is divided into two groups: 90 labeled data are used for training and 10 labeled data are used for testing. It also produces a classifier with an algorithm from 90 labeled data and applies that on the 10 testing data for set 1 and does the same thing for set 2 to 10 and produces 9 more classifiers. Afterwards, it averages the performance of the 10 classifiers produced from 10 equal sized (90 training and 10 testing) sets.

Kappa statistics plays a significant role in terms of classification in mushroom dataset. It is a chance-corrected measure of agreement between the classifications and the true classes of the entire data set. Kappa actually calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. When a value is greater than 0, that means the classifier is doing better than chance, which happens in all the three classifier that the study have been used.

Another important protagonist works here in this study is Mean absolute error (MAE). It measures the average magnitude of the errors in a set of forecasts, without considering their direction thereof. It measures accuracy for continuous variables. It is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. Root mean squared error (RMSE) measures the average magnitude of the error by using quadratic scoring rule. Here the difference between forecast and corresponding observed values are each squared and then averaged over the sample.

Table -2: Comparative Study among the Algorithms

Evaluation on Training Data Set	SMO	Naïve Bayes Simple	RIDOR
Correctly Classified Instances	110 (100%)	**110 (100%)**	98 (89.0909 %)
Incorrectly Classified Instances	0 (0%)	**0 (0%)**	12 (10.9091 %)
Kappa statistic	1	**1**	0.8837
Mean absolute error	0.1094	**0.0002**	0.0136
Root mean squared error	0.2284	**0.0012**	0.1168
Relative absolute error	93.3603%	**0.1278 %**	11.6397 %
Root relative squared error	94.3904%	**0.4881 %**	48.2497 %
Total Number of Instances	110	**110**	110

Finally, the square root of the average is taken into consideration. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. Both MAE and RMSE are used to diagnose the variation in the errors in a set of forecasts. Logically RMSE will always be larger or equal to the MAE. A comparative study among the classification algorithms used here has been shown in Table 2.

According the comparative study it is found that the NaiveBayesSimple Classifier have been most effective for classifying and decision making.

The Naïve Bayes, SMO and RIDOR classifiers are compared with accuracy measures that are depicted in below Table 3. Here, Initially the Naïve Bayes and SMO classification algorithm have high accuracy to compare with RIDOR algorithms for Mushroom Diseases Diagnosis dataset.

Table -3: Accuracy Measure for Classification

Mining Classifiers	Accuracy (%)
SMO	100
Naïve Bayes	100
RIDOR	89.0909

But, after analyzing elaborately, this study finds that more filtration is required for choosing best algorithm among three classification algorithm. So, the TP Rate, FP Rate, Precision, Recall, F-Measure, ROC area values have been verified.

In detailed Accuracy by Class, both Naïve Bayes and SMO classification algorithm has also same value without comparison of error rate measure. Besides this, at the time of comparing different type of error rates namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) in the mushroom dataset, then among two algorithm consisting of same accuracy it has been investigated that Naïve Bayes Classification algorithm has become finally more preferable because of less prone to error among all different type of error rates during this experimental analysis. Naïve Bayes Classification algorithm performs best with accuracy 100% with less error rate.

Table -4: Error Rate Measure for Best Classification

Name of Classification Algorithm	MAE	RMSE	RAE	RRSE
SMO	0.1094	0.2284	93.3603	94.3904
Naïve Bayes	0.0002	0.0012	0.1278	0.4881
RIDOR	0.0136	0.1168	11.6397	48.2497

6. CONCLUSION AND FUTURE SCOPES

The study has found that, with ever changing and growing agricultural knowledge and increasing spread of diseases, the mushroom farmers are sometimes confused and overloaded with information during decision making. It is then suggested to consult the specialists to resolve confusion. But, however, this facility of consulting specialists might not always be available or not timely available [12]. This is certainly a serious problem of mushroom agricultural domain. For this, huge damage is seen during any type of agricultural cultivation. Moreover, while farmer cultivate mushroom, same situation might be happened. As a result, at any phase of undesirable growth of mushroom could be seen and thus it causes adversely affect the final mushroom yield.

In this paper different classification algorithms such as the Naïve Bayes, SMO and RIDOR are experimentally evaluated using Mushroom Diseases Diagnosis dataset. Till now, no work has been found for diagnosing mushroom diseases. So, this field has challenging scope for future research work and directions, which may be extended further. Here, the classification algorithms are used on Mushroom Diseases Diagnosis dataset with the help of the data mining tool as WEKA. It is deduced that Naïve Bayes Classification algorithm provides better results for Mushroom Diseases Diagnosis dataset when compared with other classification algorithms.

In near future there is defiantly scope for using more different classifiers with more consistent dataset to get more accurate values for decision making. Finally, applications of advance data mining technology can enhance disease diagnosis and treatment to avoid undesirable loss during mushroom cultivation.

ACKNOWLEDGEMENT

I am obliged to Dr. Dilip Roy Chowdhury who had inspired me to create this research work. Evaluating my interest and urge he encouraged and helped me to complete this experimental work.

.

REFERENCES

[1] S.R. Sharma, “Diseases and Competitor Moulids of Mushrooms and their Management”,N.R.C.M. 2007Technical Bulletin, National Research Centre for Mushroom(Indian Council of Agricultural Research).

[2] J. Solanki, Y. Mulge “Different Techniques Used in Data Mining in Agriculture”, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), pp. 1223-1227, Vol. 5, Issue 5, ISSN:2277 128X, 2015.

[3] Abdullah, A., Brobst, S., Pervaiz, I., Umer, M. and Nisar, A. (2004).” Learning Dynamics of Pesticide Abuse through Data Mining”. In Proc. Australasian Workshop on Data Mining and Web Intelligence (DMWI2004), Dunedin, New Zealand. CRPIT, pp. 151-156, Vol. 32. Purvis, M., Ed. ACS, 2004.

[4] Cunningham S. J., and Holmes, “Developing Innovative Applications in Agriculture using Data Mining”. In the Proceedings of the Southeast Asia Regional Computer Confederation Conference, pp. 1-2, 1999.

[5] D Ramesh, B Vishnu Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data", International Journal of Advanced Research in Computer and Communication EngineeringVol. 2, Issue 9, pp. 3477-3480, ISSN (Print) : 2319-5940, ISSN (Online) : 2278-102, 2013.

[6] Jiawei *Han*, Micheline Kamber, Jian Pei., Han –"Data Mining Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, imprint of Elsevier,ISBN 978-0-12-381479-1, 2012.

[7] Paul Compton, Glenn Edwards, Byeong Kang, Leslie Lazarus, Ron Malor, Phil Preston and Ashwin Srinivasan, "Ripple down rules: Turning knowledge acquisition into knowledge maintenance*," Artificial Intelligence in Medicine (Elsevier), Vol. 4, pp. 463-475, 1992.

[8] Richard Enyinnaya, "Predicting Cancer-Related Proteins in Protein-Protein Interaction Networks using Network Approach and SMO-SVM Algorithm", International Journal of Computer Applications, pp. 5 - 9, Vol. 115, No. 3, 2015.

[9] Dmitry Pavlov, Jianchang Mao, Byron Dom,"Scaling-up Support Vector Machines Using Boosting Algorithm", Proceedings. 15th International Conference on Pattern Recognition, Vol. 2, 2000.

[10] John C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", Microsoft Research, Technical Report MSR-TR-98-14, April 21, 1998.

[11] "The Simplified SMO Algorithm", pp. 1-5, CS 229, Autumn 2009.

[12] R.K. Samanta, Dilip Roy Chowdhury and Mridula Chatterjee, A Data Mining Model for Differential Diagnosis of Neonatal Disease, IFRSA's International Journal Of Computing, Vol. 1, Issue 2, pp. 143-150, P-ISSN: 2231-2153, IF: 2.456, e–ISSN 2230-9039 April 2011.

BIOGRAPHIES

Dr. Dilip Roy Chowdhury is working as Asst. Professor in the Department of Computer Science and Application at the University of North Bengal. Before that, he has served Dept. of Computer Science & Application, Gyan Jyoti College, as HOD and other administrative responsibilities along with regular teaching job. Dr. Roy Chowdhury's main research interests lies with the design and implementation of Expert System Development using Artificial Intelligence and Soft Computing techniques. Besides, he is continuing his research in the fields of Artificial Neural Networking, Data Mining, Rough Set Computing, Knowledgebase Design and Information Retrieval. He is associated with various national and international journals of repute as a member of reviewing committee and editorial committee.
Email: diliproychowdhury@gmail.com

Mr. Subhashis Ojha is working at Dept. of Computer Application, Mathurapur B. S. S. High School, Mathurapur, Malda as an Assistant Teacher. Before that he was a faculty at Computer Application Department of Bankura Unnayani Institute of Engineering under West Bengal University of Technology. He has received MCA degree. His research area lies with Expert System, Artificial Intelligence and Soft Computing.
E-mail: subhashis.ojha@gmail.com