Mushroom Diagnosis Assistance System Based on Machine Learning by Using Mobile Devices

Intisar Shadeed Al-Mejibli
University of
Information Technology and Communications
IRAQ

Dhafar Hamed Abd
Al-Maaref University College
IRAQ

dhafar.dhafar@gmail.com

Received : 14/9/2017
Revised : 8/10/2017
Accepted : 11/10/2017

Abstract: Picking the wild mushrooms from the wild and forests for food purpose or for fun has become a public issue within the last years because many types of mushrooms are poisonous. Proper determination of mushrooms is one of the key safety issues in picking activities of it, which is widely spread, in countries. This contribution proposes a novel approach to support determination of the mushrooms through using a proposed system with mobile devices. Part of the proposed system is a mobile application that easily used by a user - mushroom picker. Hence, the mushroom type determination process can be performed at any location based on specific attributes of it. The mushroom type determination application runs on Android devices that are widely spread and inexpensive enough to enable wide exploitation by users.

This paper developed Mushroom Diagnosis Assistance System (MDAS) that can be used on a mobile phone. Two classifiers are used which are Naive Bays and Decision Tree to classify the mushroom types. The proposed approach selects the most effective of the already known mushroom attributes, and then specify the mushroom type. The use of specific features in mushroom determination process achieved very accurate results.

Keywords: Android, Decision Tree, Mushroom Determination, Mobile Device, Naive Bays.

1. Introduction

Mushrooms are one of the fungi types and currently gain super food status because they have some of the most potent nutrients on the plant. Generally, mushrooms have significant advantages such as destroying cancer cells, eradicating viruses, boosting human immune system, managing weight, improving nutrition and promoting better diet quality, and etc [1][2]. In addition to, their environmental benefits such as cleaning polluted soil and being a potential source of fuel. Hence, generally, mushroom consumption has been increased in many countries, and specifically, the consumption of edible wild mushrooms has become more popular among people [3].

Wild edible mushroom are consumed because they are source of nutrition and they are occasionally source of medicine. The people, who are practicing folk medicine, are consumed mushroom which known as medicinal mushrooms [4]. From the other hand, psychedelic mushrooms are sometimes consumed for purposes of recreational or entheogenic and not commonly used as food, because they have strong psychological effects [3]. Therefore, the value of wild edible fungi is of similar value with many vegetables and in some cases have a higher food value. Thus, the nutritional value of wild edible fungi should not be underestimated [3] [5] [6].

103
In spite of mushroom various benefits, must take into consideration that not all of them are edible because out of 10,000 species of mushrooms, 50 to 100 types are toxic [1] known as inedible. The inherent risks of eating the wild inedible mushrooms are variety and great, with effects ranging from simple symptoms such as a mild stomachache and simple indigestion to life threatening conditions and death in some cases. In the field of botany, no general rule exists for the identification of these dangerous and even life-threatening fungi. There are many wild edible mushrooms such as champignon and others are inedible like death cap which is the most dangerous type as it causes death in more than 50 percent of the incidents [7].

Currently, mushroom determination process is performed by robot and automated sorting machine in food industry. In these cases, the used technique is developed for a specific task and it cannot be used in more general applications. This attributed to constrain these simple techniques to limited features such as color, making them not proper for general applications [8].

The recent updates in communication, mobile devices and technology facilitate developing a mushroom determination system using specific characteristics selected from mushrooms features. In fact, combining a mobile phone with the mushroom determination system has many advantages. For example, mobile phones are of daily use and easy to be understood by anyone. Further, they are powerful and quickest means of using technology. Moreover, with the internet enabled mobile phone, it becomes easily benefit of an internet server. As a result with phone, internet and a determination system, any person could easily know whether a mushroom is edible or inedible [9].

Such system based on analyzing and investigating the mushroom features in order to classify the mushroom depending on their well known features (such as color histogram, hog, and odor). The main goal of this research is to design and implement Mushroom Diagnosis Assistance System (MDAS) which is offers an application running on a mobile Android device. This application provides the user with ability of identifying some mushroom attributes directly in the nature. The analyzing is done on basis of the user interaction with designed and implemented mobile application. Then the mushroom is determined and the results are presented as a final decision which is the mushroom edible or not. In addition the proposal offers many other services like providing the used with information about the all known types of mushrooms.

The rest of this paper is organized as follows: Background is described in section 2. Section3 presents research objectives. Dataset description details in section 4. Section 5 explains proposed system. Analyzing results detail in section 6. Section 7 includes the conclusion.

2. **Background**

This section presents some principles to facilitate reading the other sections of this research. It describes the classification concept, Naive Bayes and Decision tree.

2.1. **Classification**

Classification process is predicting categorical class labels which are discrete or unordered. Training set and the values (class labels) are used in constructing a model to classify the attribute and then use it in classifying new data. The task of data classification includes two-step process. The first process is learning step representing by constructing the classification model. The second process is a classification (testing) step and in this step represents the constructed model which is used given data to predict class labels for them.

The researchers have been proposed several classification methods in machine learning, pattern determination, and statistics. For example the techniques that can be used for classification are Naive Bays, and decision tree [10].

2.2. **Naive Bayes**

The Naive Bayes algorithm is known with its simplicity. It is regarded as a probabilistic classifier which counts the frequency and combinations of values in a given data set to compute a set of probabilities. The algorithm based on Bayes theorem and assumes the class variable value results from the independence of all attributes [11] [12].

The Naive Bayesian algorithm offers an adaptable method for dealing with whatever number of features or classes on probability theory, and it frequently used with supervised learning. It has the ability to that examines all its training input and it is known as the asymptotically fastest learning algorithm. Although, the model nature is characterized with simplicity, this algorithm demonstrates efficient performance in a very wide variety of problems. Further, small amounts of improper data, or “noise,” do not affect the results by much [13].

Many studies have found that the performance of simple Bayesian classifier is comparable to the performance of decision tree and selected neural network classifiers [10][14].
2.3. Decision tree

Decision tree is one of the common methods which are used for inductive inference. The induction of decision tree is the decision tree learning from training tuples of class-labeled. The structure of decision tree is like tree, where test on an attribute is represented by internal node (non leaf node), outcome of the test is denoted by branch, and class label is held by leaf node (or terminal node). The root node is the topmost node in a tree. [10][15]

The algorithm consists of the following process: it used all tuples in the training set, selects the most attribute that produces maximum information for classification, and for this attribute a test node is generated. Then, top down induction is performed in which the current set of tuples is divided by decision trees based on their values of the current test attribute. The process of classifier generation stops, if one of the following conditions is met: first all tuples in a subset related to the same class, or it would not be useful to precede with further separation into other subset. The algorithm of decision tree usually uses “information gain” measures which is an entropy-based measure. This measurement is used as a heuristic for selecting the features that will best split the training data into independent classes. [10][13].

3. Research Objectives

The main goal of this research is determining if mushrooms are poisonous or edible. This could be achieved by designing and implementing a Mushroom Diagnosis Assistance System (MDAS) that is able to analyze the input data and conclude precise decision that lead to mushroom diagnosis. Additionally, this research aims to reduce the gap of knowledge between different individuals in mushroom diagnosis. The following details the specific objectives of this research:

- Identify the target and the effective attributes (variables).
- Preprocess mushroom data so that it is suitable for training.
- Design and implement an application running on a mobile device that is performing the mushroom determination task.
- Maintain an updated database in an authorized institute that includes known species of mushrooms and newly discovered ones which are recently recorded.
- Use the system after testing and validating its performance.

4. Dataset Description

The used datasets is obtained from UCI (Unique Client Identifier) Repository of Machine Learning Database [16]. The dataset has includes description of the hypothetical sample which is corresponding to the 23 species of the gilled mushroom in the Leptota and Agarics Family. Each one of those species is identified as the definitely edible, definitely poisonous, unknown edibility, or is not recommended at all. The latter class has been combined with a poisonous and edible based on 22 physical attributes as recorded in [17].

Data distribution:
The dataset was distributed into two different classes:
- Class 1 = Edible with the number of 4208 instances (51.28%)
- Class 2 = Poison with the number of 3916 instances (48.2%)

There are total of 8124 instances that were captured.

Setting a Target
the main target of this experiment is determining if mushroom is edible (Y) or poisonous (N)

Normalizing Attributes
before any data train, the data normalization process is required. This to ensure all data value must be in string so it can be train.

The data from the UCI repository came encoded as a .txt file with 22 feature values encoded as single ASCII characters and a label of either “Y” for edible or “N” for poisonous on each line of the file. Full explanations of the encodings are available at the repository at [17]. The whole data file contains 8,124 instances of mushrooms. Roughly half are classified as poisonous and half as edible. Mushrooms that were classified as poisonous do include some of unknown toxicity and have been classified as poisonous on the side of caution. [17].

4.1. Mushroom Attributes

The mushroom 22 attributes in the dataset are represented in table 1. These attributes represent the well known mushroom attributes. Hence, a string of size 22 bytes (22 characters) is required to describe each mushroom. All the mushroom description strings are of the same size because each attribute has only one possibility [17].

There are 8124 instances in data set. These instances are distributed as 4208 edible mushrooms and 3916 poisonous mushrooms. Stalk-root attribute, which is numbered 11, misses 2480 attributes. The sign ‘?’ is used to note the
missing attributes in dataset and same sign is used in string describing of the mushroom when the attribute is missing [17]. A binary vector (0-1) can be used to describe the mushroom instead of using string because the data of mushroom is nominal.

### Table 1. Attributes of Mushroom Description in the Dataset

<table>
<thead>
<tr>
<th>Attribute no.</th>
<th>Attribute name</th>
<th>Possibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cap-shape</td>
<td>bell (b), conical (c), convex (x), flat (f), knobbed (k), sunken (s)</td>
</tr>
<tr>
<td>2</td>
<td>cap-surface</td>
<td>fibrous (f), grooves (g), scaly (y), smooth (s)</td>
</tr>
<tr>
<td>3</td>
<td>cap-color</td>
<td>brown (n), buff (b), cinnamon (c), gray (g), green (r), pink (p), purple (u), red (e), white (w), yellow (y)</td>
</tr>
<tr>
<td>4</td>
<td>bruises ?</td>
<td>bruises (i), no (f)</td>
</tr>
<tr>
<td>5</td>
<td>odor</td>
<td>almond (a), anise (l), creosote (c), fishy (y), foul (f), musty (m), none (n), pungent (p), spicy (s)</td>
</tr>
<tr>
<td>6</td>
<td>gill-attachment</td>
<td>attached (a), descending (d), free (f), notched (n)</td>
</tr>
<tr>
<td>7</td>
<td>gill-spacing</td>
<td>close (c), crowded (w), distant (d)</td>
</tr>
<tr>
<td>8</td>
<td>gill-size</td>
<td>broad (b), narrow (n)</td>
</tr>
<tr>
<td>9</td>
<td>gill-color</td>
<td>black (k), brown (n), buff (b), chocolate (h), gray (g), green(r), orange (o), pink (p), purple (u), red (e), white (w), yellow (y)</td>
</tr>
<tr>
<td>10</td>
<td>stalk-shape</td>
<td>enlarging (e), tapering (t)</td>
</tr>
<tr>
<td>11</td>
<td>stalk-root</td>
<td>bulbous (b), club (c), cup (u), equal (e), rhizomorphs (z), rooted (r), missing (?)</td>
</tr>
<tr>
<td>12</td>
<td>stalk-surface above-ring</td>
<td>fibrous (f), scaly (y), silky (k), smooth (s)</td>
</tr>
<tr>
<td>13</td>
<td>stalk-surface below-ring</td>
<td>fibrous (f), scaly (y), silky (k), smooth (s)</td>
</tr>
<tr>
<td>14</td>
<td>stalk-color above-ring</td>
<td>brown (n), buff (b), cinnamon (c), gray (g), orange (o), pink (p), red (e), white (w), yellow (y)</td>
</tr>
<tr>
<td>15</td>
<td>stalk-color below-ring</td>
<td>brown (n), buff (b), cinnamon (c), gray (g), orange (o), pink (p), red (e), white (w), yellow (y)</td>
</tr>
<tr>
<td>16</td>
<td>veil-type</td>
<td>partial (p), universal (u)</td>
</tr>
<tr>
<td>17</td>
<td>veil-color</td>
<td>brown (n), orange (o), white (w), yellow (y)</td>
</tr>
<tr>
<td>18</td>
<td>ring-number</td>
<td>none (n), one (o), two (t)</td>
</tr>
<tr>
<td>19</td>
<td>ring-type</td>
<td>cobwebby (c), evanescent (e), flaring (f), large (l), none (n), pendant (p), sheathing (s), zone (z)</td>
</tr>
<tr>
<td>20</td>
<td>spore-print-color</td>
<td>black (k), brown (n), buff (b), chocolate (h), green (r), orange (o), purple (u), white (w), yellow (y)</td>
</tr>
<tr>
<td>21</td>
<td>population</td>
<td>abundant (a), clustered (c), numerous (n), scattered (s), several (v), solitary (y)</td>
</tr>
<tr>
<td>22</td>
<td>habitat</td>
<td>grasses (g), leaves (l), meadows (m), paths (p), urban (u), waste (w), woods (d)</td>
</tr>
</tbody>
</table>
When attribute includes value 1 is used and 0 used when the attribute does not include values. The attributes have a definite length because there are only certain possibilities for describing the attribute, i.e. for attribute 21, population, there are six options, abundant (a), clustered (c), numerous (n), scattered (s), several (v), solitary (y). The feature scattered (s) can be representing as the binary string 000100 [17].

The proposed system based on these attributes in taking the decision for example:

**Mushroom is most likely poisonous if:**
- Spore print color is green

**While**

**Mushroom is most likely edible if:**
- Odor is almond, anise, or no smell at all

5. **Proposed System**

Proposed system consists of unified database, web application (server) and mobile phone application (client). Figure 1, represents the three components of proposed system. Unified database is maintained by an authorized institute such as hospital or college of Agriculture and it keeps updated information about all the known and newly discovered mushroom spices to a unified accessible database. Web application and Mobile phone application have common services, which are Mushroom Determination, Adding New Mushroom Spices, and Exploring Mushroom Spices. Although these two applications have a common services, these services are different in sort of their performance because the web application is designed for specialists and Mobile application is designed for users who are interested in Mushroom picking. The web application has more privileges in accessing and modifying the unified database.

In addition, web application offers some services for specialists only such as reply to user question and check the new discovered mushroom data before adding it to unified database.

The following explains the three common services:
1- **Mushroom Determination** service by which the user may easily know the type of suspicious mushrooms (edible or not) based on providing the application with some of mushroom attributes. If the suspicious mushroom is unknown the mobile application offers the user to send request to server side including mushroom attributes such as image which is sent after compression.

2- **Adding New Mushroom Spices** service, this service allow user to add a newly discovered mushroom spices to a unified database. This service requests the user to specify some of the discovered mushroom features like (color, description, image (sent after compression), and location) to send them to unified database. The newly discovered mushroom will be added to the unified database after the authorization of specialist.

3- **Exploring Mushroom Spices** service: this service allows the user to explore information about mushrooms, their features, images, and etc.

Further, the mobile phone application offers more services such as Remote Mushroom Determination. This services facilitates connect the user with an authorized expert from a local institution such as hospital or agricultural college, in order to provide the user with assistance in some critical situation. For example when Mushroom Determination service fails in its task or when emergency situations occurred Mushroom determination process includes two phases training phase and testing phase. Figure 2 shows the methodology of the proposal in training phase. Training phase uses two classifiers (Nave Bayse and Decision Tree) for comparison purposes. Test phase base on Decision Tree.

![Diagram](image)

**Figure 2. The Methodology of the Proposal in Training Phase**

5.1. **Mushroom Determination**

Five of the features described in previous section are used to specify the mushroom class (edible or not) after applying many experiments on the available dataset the proposal find out that the features shown in table 2 are enough to know the mushroom class. These five features are odor, spore-print-color, habitat, gill-size and cap-color.

5.2. **Features suggested by fast decision tree algorithm**

Five features are the most effective features in final decision. Figure 3, shows the steps of mushroom determination process based on the most dominant features [16].
Figure 3. Classification Process of Mushroom Determination

Table 2. The Most Dominant Attributes of Mushroom

<table>
<thead>
<tr>
<th>Attribute no.</th>
<th>Attribute name</th>
<th>Possibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>odor</td>
<td>almond (a), anise (l), creosote (c), fishy (y), foul (f), musty (m), none (n), pungent (p), spicy (s)</td>
</tr>
<tr>
<td>2</td>
<td>spore-print-color</td>
<td>black (k), brown (n), buff (b), chocolate (h), green (r), orange (o), purple (u), white (w), yellow (y)</td>
</tr>
<tr>
<td>3</td>
<td>habitat</td>
<td>grasses (g), leaves (l), meadows (m), paths (p), urban (u), waste (w), woods (d)</td>
</tr>
<tr>
<td>4</td>
<td>gill-size</td>
<td>broad (b), narrow (n)</td>
</tr>
<tr>
<td>5</td>
<td>cap-color</td>
<td>brown (n), buff (b), cinnamon (c), gray (g), green (r), pink (p), purple (u), red (e), white (w), yellow (y)</td>
</tr>
</tbody>
</table>
6. Analysis Results
Two classifiers have been applied which are Naive Bays and Decision Tree in determination task. The two classifiers applied with features selection and without features selection. Table 3 shows the two classifiers performance without features selection regarding different parameters.

Table 4 presents the accuracy by class with features selection. Table 5 shows the two classifiers performance with features selection regarding different parameters. In general, the results show that the Decision Tree classifier achieved the best performance.

Table 3. Classifiers Performance without Features Selection

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Classifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive Bays</td>
</tr>
<tr>
<td>Time/Second</td>
<td>0.04</td>
</tr>
<tr>
<td>Correctly Classified Instances</td>
<td>95.8272%</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>4.1728%</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.9162</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.0419</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.1757</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>8.3961%</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>35.159%</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>8124</td>
</tr>
</tbody>
</table>

Table 4. Accuracy of Classifications

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>True Positive</th>
<th>False Positive</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bays</td>
<td>0.984</td>
<td>0.014</td>
<td>0.984</td>
<td>0.984</td>
<td>0.999</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Y</td>
</tr>
</tbody>
</table>
In addition to, all the calculated error measurements were the lowest with Decision Tree.
Hence, in spite that the Decision Tree spent more time in determination process, its performance was better than Naive Bays. However, the difference in time between the two classifiers is approximately less than $(0.01 - 0.0 = 0.01$ second) with feature selection. Waiting 0.01 seconds represents very short time to get such critical decision.

7. **Conclusion**
Many species of mushrooms are poisonous making the task of picking the wild mushrooms from the wild and forests for food purpose or for fun too risky.

The symptoms of eating poisonous mushrooms can vary from slight gastrointestinal discomfort to death. Hence, who wants to pick a mushroom for food must be 100% sure of mushroom species before consumption

This paper developed Mushroom Diagnosis Assistance System (MDAS) consists of three components unified database, web application (server) and mobile phone application (client). MDAS maintains unified updated database about mushroom information under the supervision of the specialists and facilitates mushroom determination process on a mobile phone. This system offers a proper determination of mushrooms to achieve safety in picking mushroom, which is widely spread, in countries.
This proposed determination approach includes two phases. First phase train the machine learning system and second phase used it in determination process. The suggested determination approach specifies the most effective features in determination process, and then finds the final decision. The use of specific features in mushroom determination process achieved very accurate results. Two classifiers are used which are Naive Bays and Decision Tree are applied for validation purpose. Both classifiers applied with and without feature selection. The results show that Decision Tree classifier achieved better results than Naive Bays in term of correctly classified instances, incorrectly classified instances and error measurements. While there is a slightly difference in the required time at the classification process, as the Naive Bays needs slightly less time than Decision Tree.

References


[9] Professor Judy Wajcman, Professor Michael Bittman, Dr Paul Jones, Dr Lynne Johnstone, and Jude Brown, “The Impact of the Mobile Phone on Work/Life Balance”, The AMTA/ARC, Preliminary Report June 2007


نظام مساعد لتشخيص نوع الفطر بالاعتماد على التعليم الآلي باستخدام الأجهزة المحمولة

ظافر حميد عبد
كلية المعارف الجامعة

dhafar.dhafar@gmail.com

انتصار شديد المجيلي
جامعة تكنولوجيا المعلومات والاتصالات

dr.intisar.almejibli@gmail.com

المستخلص:

أصبح اختيار الفطر البري من البرية والغابات لغرض الغذاء أو للمتعة قضية مهتمة في السنوات الأخيرة لأن العديد من أنواع الفطر سامة. التحديد السليم للفطر هو واحد من قضايا السلامة الرئيسية في فعالية انتقاله. التنبؤ على نطاق واسع، في البلدان. تتراوح هذه الورقة نهجا جديدا لدعم تحديد الفطر من خلال استخدام نظام مقترح مع الأجهزة المحمولة. جزء من النظام المقترح هو تطبيق المحول التي تستخدم بسهولة من قبل المستخدم - منخفض ثمانا الفطر. وبالتالي، يمكن إجراء عملية تحديد نوع الفطر في أي مكان بناءً على صفات محددة منه. تطبيق تحديد نوع الفطر يعمل على أجهزة المحول التي تنتشر على نطاق واسع وغير مكلفة بحيث يمكن استغلالها بشكل واسع من قبل المستخدمين.

وضعت هذه الورقة نظام مساعد لتشخيص الفطر (MDAS) الذي يمكن استخدامه على الهواتف المحمولة. يتم استخدام أثني من المصفوفات التي هي تصنف أنواع الفطر، النهج المقترح Decision Tree و Naive Bays يقدم باختبار الصفات الأكثر فعالية من صفات الفطر المعروفة، ومن ثم تحديد نوع الفطر. استخدام ميزات محددة في عملية تحديد نوع الفطر حققت نتائج دقيقة جدا.