ISSN:2320-0790

An International Journal of Advanced Computer Technology

Integrating Ontology with Data mining with a Case of Mushroom Analysis

Dr.V.Maniraj[1], J.Nithya.[2]

[1]Associate Professor, Dept of Computer Science, A.V.V.M Sri Pushpam College,Poondi

[2]M.Phil, A.V.V.M Sri Pushpam College,Poondi

Abstract: Data mining is extract the knowledge/information from a large amount of data which stores in multiple heterogeneous databases. Data mining provides various techniques. Here using machine learning algorithm such as Pre-process, Classifier, Cluster and association rule for help to improve the Predictive accuracy. Here presently the real data in the Explorer window for Exploring a data and applied these techniques to real mushroom data for predicting a edible mushrooms. They are described in terms of physical characteristics. The mushroom dataset can be based it's attributes. All of the attribute values were nominal. Each field has a set of letter as possible values. It consists of a set of records with 22 attributes and a class label, Poisonous (or) Edible. The classifier could very well be a safe way to determine which Mushrooms I can and I can't Eat. Then considering clustering algorithm with some enhancements to aid in the process of identification of Mushroom characteristics. To applied a classify techniques for predicting a correctly classified and incorrectly classified instances. Association rule mainly used for finding the best rule. In this rule is used to making decision for identify the mushroom is edible or poisonous.

1.1. INTRODUCTION

A Mushroom Database is analyzed. Two data sets were used, one consisting of all records and the other consisting of a subset of records. In addition, an application was developed to demonstrate a technique for creating a human-machine interactive, web-enabled client-side text - based classification tool. In this process, Interactive Visual System by a human whereas the Mushroom Database application developed in this study requires human interaction during the classification process. However, the actual final prediction is made based on machine learning. This study focuses on the use of data mining techniques to analyze a previously obtained data set. To this end, the study will use a nominal data set, the Mushroom Database, and the data mining tool Weka. Various data mining algorithms are used against the Mushroom Database. The data set corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota families [1]. Each species was identified as belonging to one of three classes, definitely edible, definitely poisonous, or unknown edibility and thus, not recommended. Because of the potential dangers associated with unknown edibility, this class was combined with the poisonous one, resulting in a bigger 'definitely poisonous' class. The Mushroom Database contains 8124 instances and 23 attributes.

2.1. BACKGROUND

Data mining is a technique that discovers previously unknown relationships in data. Is the practice of automatically searching large stores of data from discover patterns and trends that go beyond simple analysis. Data mining technique uses sophisticated mathematical algorithms to segment the data and to predict the future events based on past events. Data mining is also known as Knowledge Discovery in Data. The key properties of data mining are Automatic discovery of patterns, Prediction of outcomes , Creation of actionable information , Focus on large data sets and databases. It can answer questions that cannot be addressed through simple query and reporting techniques. The goal is the extraction of patterns and knowledge from huge amount of data, not the extraction of data itself. In data mining Pre-processing is should must needed to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise, error and those with missing data. Several key concepts, algorithms and techniques, will be discussed because they are used in this study. Finally, the confusion matrix will be discussed because it is very essential key concept utilized in the analyses accomplished in this study.

2.1 ALGORITHMS AND TECHNIQUES :

2.1.1 WEKA:

WEKA is a data mining tool developed by the University of Waikato in New Zealand. It implements data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and it also has some application to real-world data mining problems. It is

a collection of machine learning algorithms for data mining tasks.

The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, classify for claasification,cluster for groups, association rules; it also includes a visualization tools for graphical representation. The new machine learning schemes can also be inculded with this package. WEKA is open source software issued under the GNU General Public License.

The goal of this tool is to help you to learn WEKA Explorer Data Mining Process. This will guide you step by step through the analysis of a simple problem using WEKA Explorer pre-processing, classification, clustering, association, attribute selection for selecting attributes, and visualization tools.

At the end of each problem there is a representation of the results with explanations. Each part is concluded with the exercise for individual practice. By the time you reach the end of this tutorial, you will be able to analyse your data with WEKA Explorer using various learning schemes and interpret received results.

Before starting WEKA, you should be familiar with data mining algorithms such as C4.5 (C5), ID3, K-means, and Apriority. All working files are provided. For better performance, the archive of all files used in this paper and it can be downloaded or copied from CD to your hard drive as well as a printable version of the lessons.

A trial version of Weka package can be downloaded from the University of Waikato website at http://www.cs.waikato.ac.nz/~ml/weka/index.html.

3.1 LITERATURE REVIEW

3.1.1. BASIC CONCEPT OF DATA MINING :

Data Mining, it is an Extraction of hidden, predictive information from large databases[5].It is also called as Knowledge Discovery from Databases (KDD).It perform an Identification and evaluation of hidden patterns in database. It is powerful technology with great potential to help organizations to locate and generate information from their data warehouses. Data mining tools predict future trends and behaviors.

It help organization to make proactive knowledge driven decisions, they prepare databases for identifying hidden patterns and also automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends.

To mine such type of data there are number of data mining tools are available. As a result it has become rather difficult for an unknown user to select the best possible data mining tool for his work. This paper presents an overview of data mining with the steps included in mining data and the different data mining methods and it also provides the reader the comparisons study of various freely available data mining tools such as WEKA tool, RapidMiner tool and NetTool Spider for web mining available today with their own strengths and weaknesses.

3.1.2. Data Mining

Data mining is also known as refers to extracting or “mining" knowledge from large amounts of data. It is also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns such as association rules[6]. It applies many older computational techniques from statistics, information retrieval, machine learning and pattern recognition. Following are the data mining steps:

- **Data Cleaning**:In this step, data that contain corrupted,errorneous or empty records are removed.
- **Data Integration**:In this step to proceed with data mining, data need to be collected and integrated into a single format structure. However, different sources of data usually do not contain uniform structures and interpretations of data; therefore integration into a single format needs to take place.
- **Data Selection**: definitely not all of the data collected are needed though. Data selection is mainly allows for choosing only such data that are relevant to the task to be performed.
- **Data Transformation**:The data that have passed the cleaning step are still not ready for data mining purposes, for they still need to be transformed into format accepted by the data mining algorithm.
- **Data Mining**:In this step,different kinds of algorithms may be applied on the data in order to discover potential knowledge hidden within the data.
- **Pattern Evaluation**:The importance of results provided by data mining needs to be evaluated, for not all of the findings may be of interest to the inquiry. Redundant patterns are therefore removed.
- **Knowledge Presentation**:Results that appear to be the most important undergo transformation and visualization in order to be provided in the most understandable form.
-

3.1.3 Data Mining Methods

- **Classification:** Supervised Learning. The classes are known

- **Clustering:** Unsupervised Learning. The classes are unknown
- **Association Rule Mining:** Identifying the hidden, previously unknown relation between the entities.
- **Temporal mining**: Use with temporal data, modeling temporal events, time series, pattern detection, sequences and temporal association rules are some tasks.
- **Time Series Analysis:** Describe the trend, nature and behavior of time series data. Predict the future trend and behavior of the data.
- **Web Mining:** Mining web data; Web content mining, Web structure mining and Web usage mining.
- Spatial Mining: Use with GIS for mining knowledge from spatial database. Spatial classification and clustering and rule generation are some task under this mining.

3.1.4 Categories of Mining

The two categories of data mining are Descriptive mining and Prescriptive mining. Summarizing or characterizing the universal properties of data in data repository is known as Descriptive mining. Prescriptive mining is to perform inference on existing data, to make predictions based on the past data.

Association rule mining, classification and clustering are some of the data mining techniques.

The data can be classified into different categories based on the mining techniques that are applied in Data mining. Some of them are (a) Relational data, (b) Transactional data, (c) Spatial data, (d) Temporal and time-series data and (e) World Wide Web data[9].

Recently the Chinese government gave great importance to the culture industry for economic growth. To analyse the factors about recognition, satisfaction and participation of residents on cultural activities, Apriori association rule mining algorithm was applied on a survey data.

The mining results revealed that income, occupation and educational background as the main factors of culture industry.

Based on the results, suggestions were given to make decision support to improve the living standard and education background of residents to improve the participation in cultural activities [30].

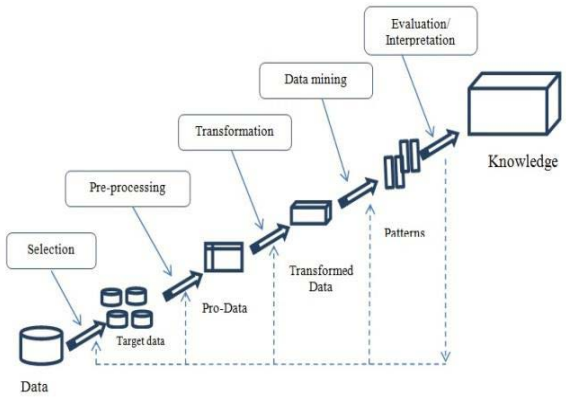


Fig. 1. KDD process

In sports management, Association rule mining algorithm was applied for a case study on Indian Cricket team; especially mining relationship on the team's performance data in one day international (ODI) matches.

This analysis used in determining the factors associated with the match outcome so as to enable the team to frame match winning strategies.

In recent years, Evolutionary Algorithm has been broadly accepted in many systematic areas and it derives mechanisms of biotic progression and applies them in problem solving.

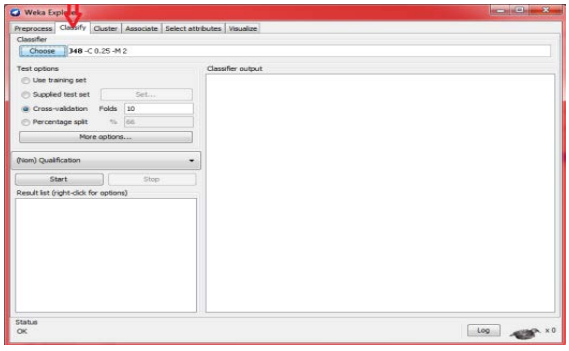
The algorithm also applied in the field of Tax inspection excavation, traffic management and network analysis.

4.1 FOCUS OF STUDY

4.1.1 CLASSIFY DATA

4.1.1.1 Building "Classifiers"

Classifiers in WEKA for predicting nominal or numeric quantities. There are multiple machine learning schemes available in WEKA include decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptron's, logistic regression, and bayes' nets. Once you have to load your data set , all the tabs are available to you. Click on the 'Classify' tab.



'Classify' window displays on the screen. Now you can start data analyzing process using the provided algorithms. In this paper you will analyze the data with C4.5 algorithm using J48, WEKA's implementation of decision tree learner[8].

Choosing a Classifier:

Click on 'Choose' button in the 'Classifier' box just below the tabs and select C4.5

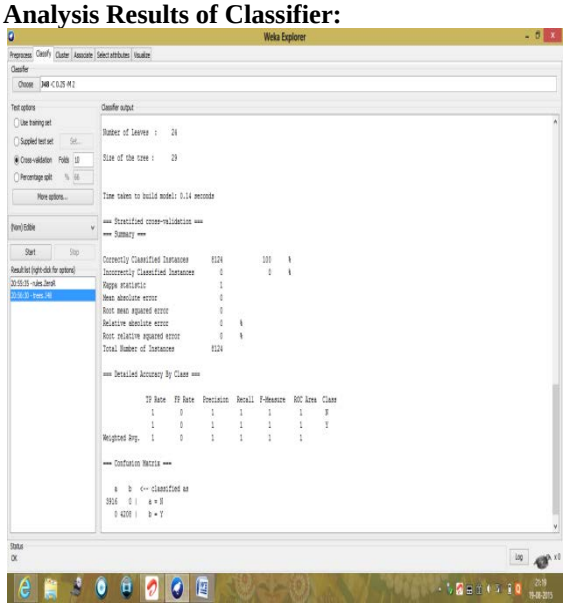
Classifier WEKA → Classifiers→ Trees → J48

Setting Test Options:

Before you can run the classification algorithm, you need to set test options. Set test options in the 'Test options' box. The test options that available to you are:

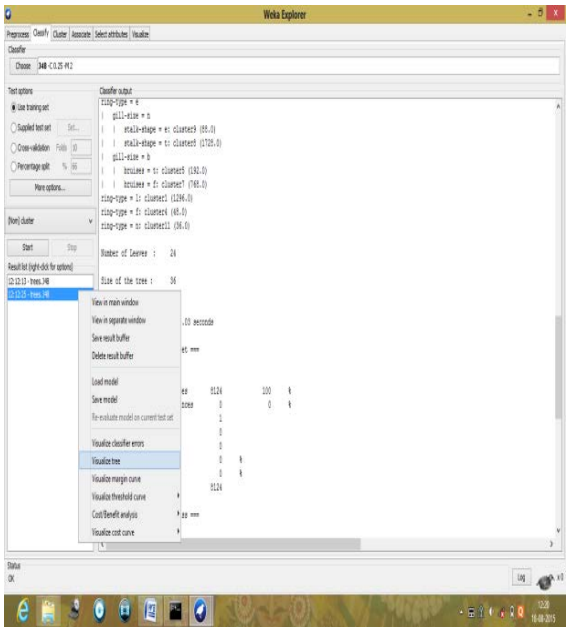
1. **Use training set** - Evaluates the classifier and well it predicts the class of the instances it was trained on.
2. **Supplied test set** - Evaluates the classifier on how well it predicts the class of a set of instances loaded from a file. Clicking on the 'Set…' button brings up a dialog allowing you to choose the file to test on.
3. **Cross-validation** - Evaluates the classifier by cross-validation, using the number of folds that are entered in the 'Folds' text field.
4. **Percentage split** - Evaluates the classifier on how well it predicts a certain percentage of the data, which is held out for testing. The amount of data held out depends on the value entered in the '%' field.

Analysis Results of Classifier:

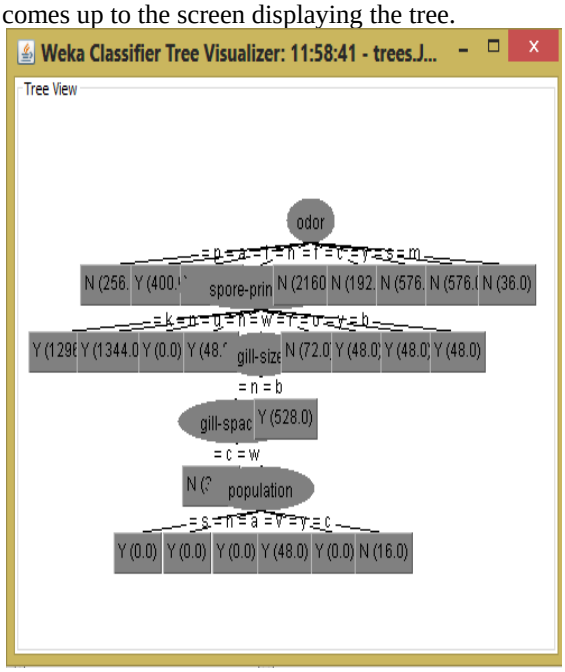


WEKA lets you to see a graphical representation of the classification tree. Right-click on the entry in 'Result list' for which you would like to visualize a tree.

It invokes a menu containing the following items:



Select the item 'Visualize tree'; a new window comes up to the screen displaying the tree.



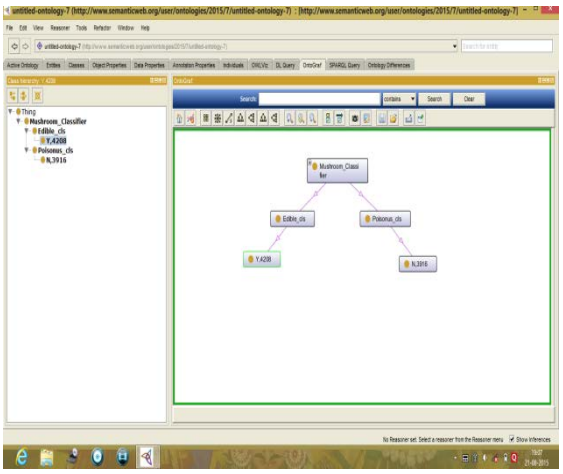
WEKA also lets you to visualize classification errors. Right-click on the entry in 'Result list' again and select 'Visualize classifier errors' from the menu.

5.1 Results of Study:

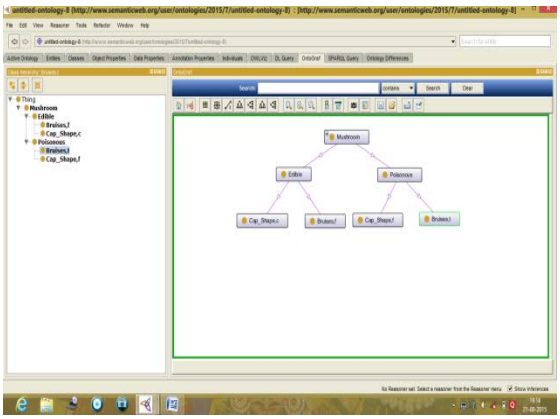
Results are derived from weka tools using Machine Learning algorithm in clusters, select attributes and Association.

Those results can be displayed in tree structure using Protege tool for easy understanding.

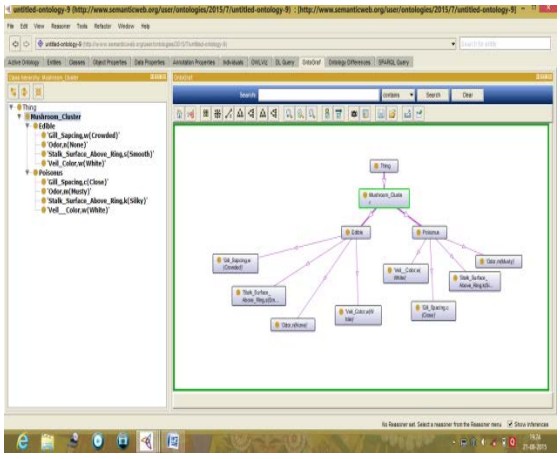
Tree : 1



Tree : 2



Tree : 3



6.1 Descriptions of Mushroom Database Application

6.1.1 DataSet and Method

6.1.1.2 Analyzing the Data and Parsing the Data

The data set that is being used contains 8124 of instances with 22 attributes[10].

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?: bruises=t,no=f
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6. gill-attachment: attached=a,descending=d,free=f,notched=n
7. gill-spacing: close=c,crowded=w,distant=d
8. gill-size: broad=b,narrow=n
9. gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
10. stalk-shape: enlarging=e,tapering=t
11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
12. stalk-surface-above-ring: ibrous=f,scaly=y,silky=k,smooth=s
13. stalk-surface-below-ring: ibrous=f,scaly=y,silky=k,smooth=s
14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
16. veil-type: partial=p,universal=u
17. veil-color: brown=n,orange=o,white=w,yellow=y
18. ring-number: none=n,one=o,two=t
19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

7.1 DISCUSSION AND CONCLUSION

The results obtained based on the decision tree is extraordinary promising. It proves that poisonous mushrooms has their distinct characteristics from edible species which could be used for identification. The tree printed out by the decision tree could be very convenient for determining the poisonous mushroom in a quickly manner. People with even no prior knowledge could easily distinguish mushrooms by simply tracking down the decision tree. The cross-validation is critical to this project, as one can see it improves the accuracy from 96% to 100%. While it's also need to note that the cross-validation implemented in this project is still far from perfect

Here also need to note that, in contrast to KNN, which has to calculate the distances between instances based on 22 attributes, decision tree saves plenty of computational time by using the mutual information.

References :

[1] Mushroom records drawn from the audubon society field guide to north american mushrooms (1981). G. H. Lincoff (pres.), new york: alfred a. Knopf

[2] Artificial intelligence: a modern approach, **3rd edition** (blue) stuart j. Russell and peter norvig. Prentice hall, englewook cliffs, n.j., 2010, 702.

[3] Mushroom poisoning: http://en.wikipedia.org/wiki/mushroom_poisoning

[4] mushroom database: https://courses.cs.washington.edu/courses/cse473/01au/assignments/mushroom-names.txt

[5]—G Effective use of the kdd process and data mining for computer performance professionals — by susan p. Imberman ph.d. College of staten island, city university of new York

[6] —data mining techniques classifi ation and prediction —by han/kamber/pei, tan/steinbach/kumar, and andrew moore mirekriedewald

[7] c—Classification and Prediction in a data mining application — by serhat özekes and a.yilmaz çamurcu 2 istanbul commerce university, ragip gümüş pala cad. No: 84 eminönü 34378, istanbul – turkey

[8] —Survey of classification techniques in data mining — bythair nuphyu e. H. Miller, —a note on reflector arrays (periodical style—accepted for publication),‖ *ieee trans. Antennas propagat.*, to be published.

[9] Data mining technology by jiawei han department of computer science university of illinois at urbana-champaign

[10] Mushroom poisoning: http://en.wikipedia.org/wiki/mushroom_poisoning.