Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Mining a Tea Insect Pest Database

Jyotshna Solanki (Student)
PDM College of Engineering & Technology
Bahadurgarh (Haryana), India

Prof. (Dr.) Yusuf Mulge (Principal)
PDM College of Engineering &Technology
Bahadurgarh (Haryana), India

Abstract: ***Data mining techniques are widely acceptable in different application areas. These application areas are different based on dataset specification as well as the criticality of the application area. In this research, mining approach is defined for tea insect pest database analysis. This dataset is having the feature set of tea gardens and it is required to identify the pest over the tea leafs. In this present work, a weighted decision approach is defined to predict the disease. The presented work is defined in two main stages. In first stage, the attribute analysis based weight assignment approach is defined. The weightage will be assigned under entropy value computation. In second stage, this weighted dataset will be taken by naive bays algorithm and the classification will be implemented. This classification stage will separate the disease instances under different classes. The analysis of work will be done under recognition rate parameters. Data mining is a process of extracting/equating the meaningful data from the large database by using different tools and techniques. Data mining is the process of evaluating data from different outlooks and summarizing it into useful information.***

Keywords: Data Mining, Naïve Bayes algorithm, Techniques, etc.

I. INTRODUCTION

Data mining is a process of extracting/equating the meaningful data from the large database by using different tools and techniques. Data mining techniques are widely acceptable in different application areas. These application areas are different based on dataset specification as well as the criticality of the application area. In this research, mining approach is defined for tea insect pest database analysis. This dataset is having the feature set of tea gardens and it is required to identify the pest over the tea leafs. In this present work, a weighted decision approach is defined to predict the disease. The presented work is defined in two main stages. In first stage, the attribute analysis based weight assignment approach is defined. The weightage will be assigned under entropy value computation. In second stage, this weighted dataset will be taken by naive bays algorithm and the classification will be implemented. This classification stage will separate the disease instances under different classes. The analysis of work will be done under recognition rate parameters

***Naïve Bayes Algorithm :Bayesian Classification*:**

Bayesian classifiers are statistical classifier. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. In theory Bayesian classifiers have the minimum error rate in comparison to all other classifiers. In practice this is not always the case, owing to inaccuracies in the assumption made for its use, such as class conditional independence and the lack of available probability data.

Bayes theorem

According to Bayes' theorem, the probability that we want to compute P (H|X) can be expressed in terms of probabilities P (H), P (X|H), and P(X) as

P (H|X) = P (X|H) P (H)/P(X)

And these probabilities may be estimated from the given data.

Naïve Bayesian classification

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.

Bayesian belief network

Bayesian belief networks specify joint conditional probabilities distributions. They allow class conditional independencies to be defined between subsets of variables. They provide a graphical model of causal relationships, on which learning can be performed. Trained Bayesian belief networks can be used for classification.

A belief network is defined by two components - a directed acyclic graph and a set of conditional probability table. Each node in the directed acyclic graph represents a random variable. The variables may be discrete or continuous valued. They may correspond to actual attributes given in the data or to "hidden variable" believed to form a relationship.

Each arc represents a probabilistic dependence. If an arc is drawn from a node y to node z then y is a parent or immediate predecessor of z. each variable is conditionally independent of its no descendants in the graph, given its parents.

A belief network has one conditional probability table (CPT) for each variable. The CPT for a variable Y specifies the conditional distribution P (Y| Parents (Y)), where parents(Y) are the parents of Y.

Naïve Bayes Tree

NB Tree appears to be a viable approach to inducing classifiers, where:

- Many attributes are relevant for classification and
- Attributes are not necessarily independent
- Database is large
- Interpretability of classifier is important

II. LITEREATURE SURVEY AND METHODOLOGY FOLLOWED

Data miningis one of the important and young feature in the field of computer science. It generally refers to or means the extraction of hidden information from large amount of databases. Data mining helps organizations to discover and identify the hidden patterns in databases. This chapter includes details like classification and its techniques. Classificationis the most commonly used data mining technique. It generally refers to mapping of data item into predefined groups and classes. This chapter also involves discussion about lazy learners and eager learners. Data classification is a two-step process which involves two basic steps like model construction and model usage. Description of the survey done is described in detail.

3.1 Data Mining in Education

Data miningis one of the important and young feature in the field of computer science. It generally refers to or means the extraction of hidden information from large amount of databases. Data mining helps organizations to discover and identify the hidden patterns in databases. The extracted patterns are then made in use to build data mining models and hence can be used to predict performance and behavior with high accuracy. It involves the use of sophisticated data analysis tool, which helps to discover unknown valid patterns and relationships in large databases. Tools can be statistical models, mathematical algorithm and machine learning methods. Data mining is used in several areas of interest. Several techniques are used to mine data like that of decision trees, genetic algorithm, artificial neural network, visualization and nearest neighbor method. These methods are used to reveal hidden information from large databases in many organizations. Mining of data helps to solve problems, providing up advantageous results. There are different alternative names for data mining such as that of knowledge Discovery in Databases (KDD), Knowledge Extraction, Data Archeology, Data Dredging, Information Harvesting, Business Intelligence etc.

3.2 Model Construction and Model Usage

Data classification is a two step process which involves two basic steps like model construction and model usage. In the model construction, each tuple is assumed to belong to a predefined class which is determined by one of the attribute, generally referred to as class label attribute. Data tuples are termed as samples, objects or examples. These tuples build up a training data set. Model representation is done in the form of classification rules, decision tree or mathematical formulae.

Model usage involves the use of model for classifying the objects that are not known, which helps in predicting up the results. It also helps in estimating the accuracy of the model. The accuracy of the model on a given test is the percentage of test set samples that are correctly classified by the model.

Author[10]made an attempt to use the data mining processes, particularly classification. They worked on enhancing the quality of higher educational system by evaluating the students' performance in courses. CRISP framework use was employed for mining the students' related academic data. Decision tree use was made under classification model, which helped in generation of rules, thus in turn predicting and evaluating the performance. Result prediction was made in the form of grades.

Methodology Followed:

The methodology discussed consists of following steps:-

1. Data collection
2. Preprocessing
3. Data mining
4. Data evaluation
5. Result and analysis

***Data Collection*:**The presented work is to perform the detection of pest disease by for the available dataset. This dataset is collected from UCI repository. The properties of the dataset are listed here under

Data Preprocessing (Preprocessing):Incomplete, noisy and inconsistent data are common properties of any database. Incomplete data can occur because attributes of interest may not be always available. After the data is collected, data preprocessing is done to prepare the final datasheet. Data preprocessing is the activity which removes the irrelevant attributes, removes the inconsistencies and fills the missing values. Data cleaning attempt to fill in missing values, smooth out noise and correct inconsistencies in the data. If there is any irrelevancy in any form then it is removed so that the prediction could be accurate enough. Data preprocessing involves data reduction. In this work, the preprocessing has been accomplished as the selection of related attributes.

Design: Agricultural data processing is one of the major research applications. In this work, the detection of pest disease for tea leaf data is proposed. To perform the accurate detection, a hybrid model is presented in this work. At the earlier stage of this work, the attribute selection will be filtered using weighted decision tree approach.

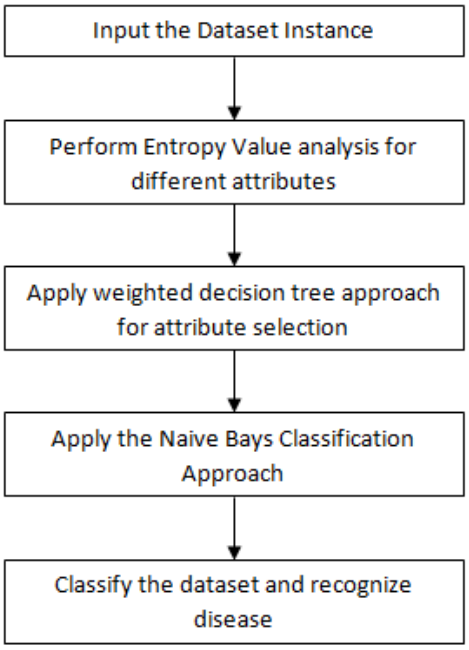

Figure 1 : Flowchart of Work

Decision Tree Approach: Decision Tree is the partition based approach defined to provide the clear region division based on the data point analysis. In this work improved decision tree model is presented for disease classification for crop. The work is defined specifically to identify the pest disease. This work, a weighted analysis model based on decision tree is defined to identify the disease class. The feature analysis is here improved while deciding the decision vector based on specific attribute analysis. This analysis is defined under entropy value. The probabilistic feature analysis is here obtained to derive the first level analysis and later on the entropy weights are applied to obtain the class. The condition aspects are repeated for all the attributes one by one till the clear class identification of disease class or the normal instance is not obtained from the work.

III. CONCLUSION

In this present work, one of the most critical data mining applications is explored for most critical operation. This application area is based on agriculture field. In this work, the pest disease identification is processed over the soyabean dataset. The presented work has improved the decision capability of the nodes by adding the nodes at the attribute level. The effective selection of nodes or attributes is here done based on the cost estimation under entropy value. The estimation is here done to obtain the significant results from the system. The experimentation is here performed on three different sample sets. The results show that the improved model has provided the recognition rate over 95%.

REFRENCES

[1] JiaweiHan ,MichelineKamber, " Data Mining Concepts and Techniques", Morgan Kaufmann publishers, 2001
[2] IAN .H Written and EibeFrank , "Data Mining Practical Machine learning tools and Techniques" , Morgan Kaufmann Publishers is an imprint of Elsever , San Fransisco , 2005
[3] Margret H. Dunham , "Data Mining- Introductory and advanced topics" , published by Pearson Education, 2003 Journals/ conferences
[4] Baird, "The role of academic ability in high level accomplishment and general success". College Board Report No. 82 New York: Research and Development. (1982)
[5] Hilton, T.L. "Persistence in higher education: an empirical study, College Board Report No. 82-5 New York : Research and Development. , (1982)
[6] Ditcher , A and Tetley, " Factors influencing university students academic success: what do students and academic think?" HERDS Annual International conference, Melbourne , 1999
[7] Luan, "Data Mining application in higher education". SPSS Inc, 2004.
[8] Ervina, A , and Md Nor, "Undergraduate students performance : the case of university of Malaya Quality Assurance in Education , 13(4) , 329-343 , 2005.
[9] Golding P. and Mc Namara, "Predicting academic performance in the school of computing and in the school of computing and information technology (SCIT) . 35th ASEE/IEEE frontiers in Education Conference , Indianapolis , 2005
[10] QasemA.Al-Radaideh ,Emad M. Al – Shawakfa and Mustafa L.Al-Najjar, "Mining student Data using Decision Trees" . The 2006 International Arab Conference on Information Technology (ACIT 2006)

