International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET207247

Identification of Poison using C4.5 Algorithm

Lai Lai Yee[1], Myo Ma Ma[2]
[1]Faulty of Information Science, University of Computer Studies (Pyay), Pyay, Bago, Myanmar
[2]Faulty of Computer Science/University of Computer Studies (Mandalay), Mandalay, Myanmar

ABSTRACT

Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses or other information repositories. This can be viewed as a result of the natural evolution of information technology. The key point is that data mining is the application of these and other AI and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional. This paper is classification system for Toxicology using C4.5. Firstly, the input data are randomly partitioned into two independent data, a training data and a test data. And then two third of the data are allocated to the training data and the remaining one third is allocated to the test data. Final step is C4.5 Algorithm Process, the training data is used to derive C4.5 algorithm. Classification Process, test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable the rules can be applied to the classification of new data.

Keywords : Classification, Data Mining, Toxicology, C4.5

I. INTRODUCTION

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. A knowledge discovery process include data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining system can be classified according to the kinds of databases mined, the kinds of knowledge mined, the techniques used or the applications. Data mining also called Knowledge-Discovery Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volume of data for patterns using tools such as classification, association rule mining, clustering, etc.

Data mining is iterative and interactive processes that explore and analyse voluminous data in order to discover valid, novel and meaningful patterns, associations or rules, using computationally efficient techniques. It is related to the sub area of statistics called exploratory data analysis, which has similar goals and relied on statistical measures and also closely related to the sub areas of artificial intelligence called knowledge discovery and machine learning. Data mining has been attracted huge attention in numerous research communities due to its wide applicability in many areas such retail industry, financial forecast, and decision support and intrusion detection. Data mining methods include associations, clustering, classification and prediction.

One of the most important fields of the data-mining domain is the association mining.

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

There are four main objectives in this paper. The first is to study classification theory under data mining system in details. The second is to know how decision trees were constructed by C4.5 algorithm. The third is to know C4.5 (based ID3) algorithm is applied to the mining of very large real-world databases. The last is to apply many kinds of poison using computerized system.

II. IDENTIFICATION OF POISON

This paper claims that classification of toxicology data. The goal of this research paper is to classify poison data by using decision tree classification method and to assess its accuracy by holdout method. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Many classification and prediction methods have been proposed by researchers in machine learning, expert systems, statistics and neurobiology. Data classification is a two-step process. These two-steps are model construction and model usage. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction training set. The model is represented as classification rules, decision trees, or mathematical formulae. Model usage is used for classifying future or unknown objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over fitting will occur.

(1) Classification of Data Mining

Data mining is an interdisciplinary field, the confluence of a set of disciplines including database systems, statistics, machine learning visualization and information science. Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, business, bioinformatics, or psychology.

(2) Poison

Poison deals with the science that embodies the knowledge of the sources, characters and properties of poisons, the symptoms they produce the nature of their fatal effects and the remedial measures that should be taken to combat their actions or effects. It is difficult to give an exact definition of the term “toxic”, for substances which are harmless to the body in certain conditions may become dangerous in other conditions. For instance, the salts of potassium are not only poisonous in small doses, but are essential for the maintenance of a health condition of the body. In large quantities however they act as acute poisons, capable of destroying life. Broadly speaking, a poison may be defined as any substance administered in whatever way produces ill-health, disease or death. It may be of synthetic, mineral, animal or vegetable origin and may be administrated by mouth, injection and inhalation or through skin or mucous membrane contact. This also covers the poisonous, which are not often used criminally, except during warfare.

(3) Comparing Classification Methods

Classification and prediction methods can be compared and evaluated according to the following criteria.

(i) Predictive accuracy: This refers to the ability of the model to correctly predict the class label of new or previously unseen data.
(ii) Speed: This refers to the computation costs involved in generating and using the model.
(iii) Robustness: This is the ability of the model to make correct predictions given noisy data or data with missing values.
(iv) Scalability: This refers to the ability to construct the model efficiently given large amounts of data.
(v) Interpretability: This refers to the level of understanding and insight that is provided by the model.

(4) Other Classification Methods

Classification is a preliminary data analysis step for examining a set of cases to see if they can be grouped based on 'similarity' to each other. Data analysis methods vary on the way how they detect patterns. The ultimate reason for doing classification is to increase understanding of the domain or to improve predictions compared to unclassified data.

The decision tree method like the nearest neighbours method, exploits clustering regularities for construction of decision-tree representation. It shows implicitly which variables are more significant with respect to classification decisions. The decision tree learning method requires the data to be expressed in the form of classified examples.

The method of Memory Based Reasoning also called the nearest neighbour method finds the closet part analysis of the present situation and chooses the same solution such systems demonstrate good results in vastly diverse problems. Such systems do not create any models or rules summarizing the previous experience.

III. RESULTS AND DISCUSSION

This paper shows that identification of poison data. The aim of this research is to classify poison data by using decision tree classification method and to assess its accuracy by holdout method. There are four main objectives in this paper. The first is to study classification theory under data mining system in details. The second is to know how decision trees were constructed by C4.5 algorithm. The third is to know C4.5 algorithm is applied to the mining of very large real-world databases. The last is to apply many kinds of poison using computerized system.

There are two processes. Input data are randomly partitioned into two independent data, a training data and a test data. Typically two third of the data are allocated to the training data and the remaining one third is allocated to the test data. C4.5 algorithm process, the training data is used to derive C4.5 algorithm. Classification Process, test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable the rules can be applied to the classification of new data.

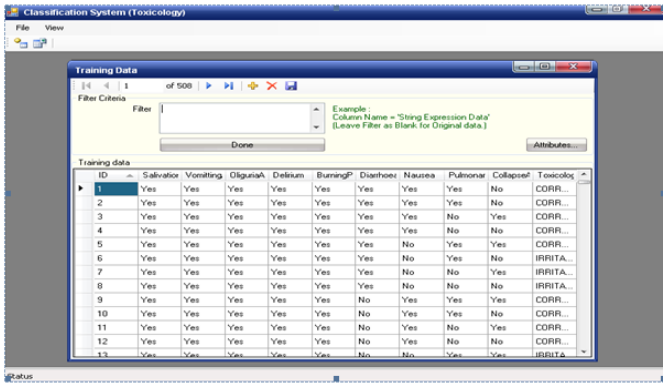
Finally, complete content and organizational editing before formatting.

(1) Database Design and Implementation

Database Design for training sample data and testing sample data.

Column Name	Data Type	Field Size	Required	All Zero Length	Primary Key	Description
ID	Number	Long Integer	Yes	No	Yes	Auto Number Field
Result	Text	10	No	Yes	No	Testing or Pre-defined value
Salivation	Text	10	No	Yes	No	Testing or Pre-defined value
Vomiting and Purging	Text	5	No	Yes	No	"--" must be used instead of empty
Oliguria, anuria, ablumin and casts in urine	Text	5	No	Yes	No	"--" must be used instead of empty
Delirium	Text	5	No	Yes	No	"--" must be used instead of empty
Burning pain in the gullet, mouth, throat and stomach	Text	5	No	Yes	No	"--" must be used instead of empty
Diarrhoea	Text	5	No	Yes	No	"--" must be used instead

This is the multi-documents interface application and the following is the application main user interface screenshot of "Identification of poison Using C4.5 Algorithm" system. In this form, there are four main menus: File and View as shown in Figure(1).



Figure(1): Application Main Form

This system supports users in identifying by providing suggestion according to the user's input. It also allows the users to retrieve the information about poison. It can easily modify data from the database. It gives some ideas based on the concepts of identifying systems

The limitation and further extension of this system are: it is implemented only by using C4.5 algorithm, it does not compare with other classification methods such as Naive Bayesian Classifier (NBC) and Bayesian Belief Network (BBN). The future work will extend decision tree induction (C4.5 algorithm) to work on the other data sets. We can plan to test the poisons dataset by using other classifiers such as Naive Bayesian and K-Nearest Neighbor. In classification problems, it is commonly assumed that all objects are uniquely classifiable, i.e., that each training sample can belong to only one class. In addition to efficiency, classification algorithms can then be compared according to their accuracy. Accuracy is measured using a test set of objects for which the class labels are known. Accuracy is estimated as the number of correct class predictions, divided by the total number of test samples.

IV. CONCLUSION

With the recent emergence of the field of data mining, there is a great need for algorithms for building classifiers that can handle very large databases. This system implements the user to view the knowledge of data mining. This system has described the feature subset selection problem in supervised learning, which involves identifying the relevant or useful features in a dataset and giving only that subset to the learning algorithm. For the evaluation, the user used hold-out method as accuracy estimation technique. It shows how a decision tree is constructed by C4.5 algorithm. By studying poisons, the user can prevent risk factors and maintain healthy living. This system demonstrates efficiency and effectiveness in dealing with poison for identification.

V. REFERENCES

[1]. N. J. MODI "Medical Jurisprudence and Toxicology".

[2]. J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001.

[3]. http:// en. Wedipedia. Org.
[4]. http:// www. Decisiontrees. Net
[5]. Pang_Ning, Tan Michael Steinbach and Vipin Kumar "Introduction to Data Mining".
[6]. http:// eruditionhome.com/datamining

Cite this article as :

Sh Lai Lai Yee, Myo Ma Ma, "Identification of Poison using C4.5 Algorithm", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 2, pp. 218-222, March-April 2020. Available at doi : https://doi.org/10.32628/IJSRSET207247
Journal URL : http://ijsrset.com/IJSRSET207247