A Study on WEKA Tool for Data Preprocessing, Classification and Clustering

Swasti Singhal, Monika Jena

Abstract— The basic principles of data mining is to analyze the data from different angle, categorize it and finally to summarize it. In today's world data mining have increasingly become very interesting and popular in terms of all application. The need for data mining is that we have too much data, too much technology but don't have useful information. Data mining software allows user to analyze data. This paper introduces the key principle of data pre-processing, classification, clustering and introduction of WEKA tool. Weka is a data mining tool. In this paper we are describing the steps of how to use WEKA tool for these technologies. It provides the facility to classify the data through various algorithms.

Keywords: Data mining; data preprocessing, classification, cluster analysis, Weka tool etc.

I. INTRODUCTION

Firstly, why data mining? As we know that there is a explosive growth of data from terabytes to petabytes. Major problem was the availability of data. We are drowning in data but starving for knowledge. Companies invested in building data warehouses that contain millions of records and attributes but they are not getting the ROI (return on investment). They cannot produce sufficient output due to lack of knowledge, lack of staffs and appropriate tools. Data mining is the process of automatic classification of cases based on data patterns obtained from a dataset. A number of algorithms have been developed and implemented to extract information and discover knowledge patterns that may be useful for decision support. Data mining also known as KDD(knowledge discovery in databases).data preprocessing , pattern recognition, clustering [11], classification[12] are the popular technologies in data mining .In this paper , we will discuss detailed about the data preprocessing that comes in the database server module after that we will discuss the database module in which we will defined two methods: classification and clustering by the data mining tool i.e. Weka (Waikato Environment for knowledge analysis) tool 3.6.8.

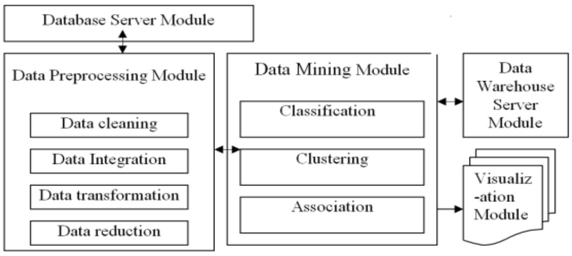


Fig1.

Manuscript received May, 2013.

Swasti Singhal , Computer Science and Engineering,, Amity University, Noida, INDIA.

Monika Jena, Prof Computer Science and Engineering,, Amity University, Noida, INDIA.

II. BACKGROUND

The aim of this section is to give the brief insight into the data preprocessing and Weka background. This chapter is organized as follows: section 2.1 gives the overview of data preprocessing and 2.2 give the introduction of Weka 3-6-8 interface.

A. DATA PREPROCESSING

(I) Why preprocess the data? Because data in the real world is dirty, incomplete and noisy. Incomplete in lacking attributes values and lacking attributes of interest or containing only aggregate value noisy in terms of containing errors or outliers and inconsistent containing discrepancies in names or codes. Now the question arises why is the data dirty? Because incomplete data may come from "not applicable" data value when data has to be collected and the major issue is a different consideration between the times when the data was analyzed and human hardware and software issues are common.

Noisy data may come from the when a human enters the wrong value at the time of data entry as Nobody is perfect. Errors in transmission of data and instruments that collect the faulty data. Inconsistent data may come from the different data sources. Duplicates records also need data cleaning.

(II) Why data preprocessing is important?

Data is not clean, Duplicity of data and the no quality data and the most important is no quality result so data preprocessing is important. Quality decisions must be based on the quality data. Data warehouse needs consistent integration of quality data. By the processing of data, data quality can be measures in term of accuracy, completeness, consistency, timeliness, believability, interpretability.

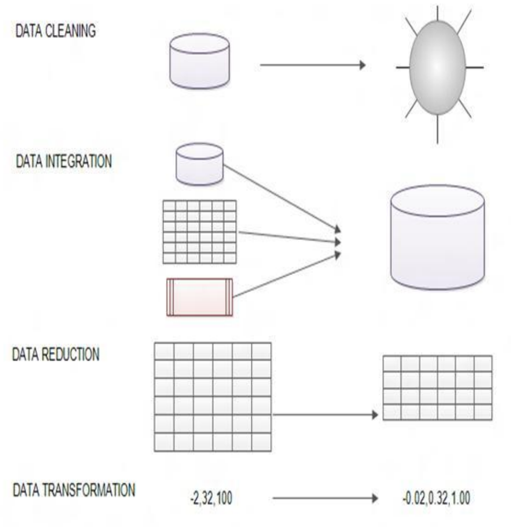


Fig2

Data pre-processing is the important step in data mining process. The question is why the data needs to be cleaned before it is processed? Sometimes data is collected in an ad hoc manner. This happens because usually the data that will be used has not been good, this includes:

Incomplete: shortage of attribute values or certain other attributes.

Noisy: contains error or outlier values that deviate from the expected.

Inconsistent: discrepancies in the use of the code or name.

Here good data quality based on the good decisions and data warehouse integration requires consistent data quality.

Data entry mistakes can occur and/or the data may have missing or unknown entries. During the data cleaning and preprocessing stage noise is removed from the data. Outliers and anomalies in the data can pose special problems for the data analyst during the data cleaning process. Raw data can be available but it is necessary to make it useful. Data collection is the important aspect in every field but if the information is irrelevant then it will be the huge problem. Problem like missing values, impossible data combination, out of range values. Such problem can produce misleading result. Misleading results and this problem can give the poor quality of data. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set. Kotsiantiset al presents a well-known algorithm for each step of data pre-processing. [13]

As seen in the Fig1 it database server module shows the steps of data pre- processing technologies or method used in data preprocessing includes data cleaning fig2 to eliminate the incorrect values and to check the inconsistency of the data. Second step is data integration it is to combining the data from the databases, files and data cubes etc. Data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system.

Data transformation can be divided into two steps:

Data mapping maps data elements from the source data system to the destination data system and captures any transformation that must occur code generation that creates the actual transformation program. Next is data reduction which Elaborating the data into a form that is smaller in size but still produces the same analytical results. last is data discretization which is Part of the data reduction but has its own importance, especially for numerical data in statistics and machine learning, discretization refers to the process of converting or partitioning continuous attributes, features or variables to discredited or nominal attributes/features/ variables/ intervals. This can be useful when creating probability mass functions – formally, in density estimation. It is a form of binning, as in making a histogram.

When the data is good? It can be checked in terms of the: **Accuracy, Completeness and Consistency**, **Timeliness**, **and Value added Interpretability**, **Accessibility**, **Contextual**, **and Representational.**

B. WEKA INTERFACE:

The Weka or woodhen (Gallirallus australis) is an endemic bird of New Zealand. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand [18]. The Weka suite contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. It provides many different algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent.

As shown in fig3 the GUI Chooser consists of four buttons:

Explorer: An environment for exploring data with WEKA.

Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.

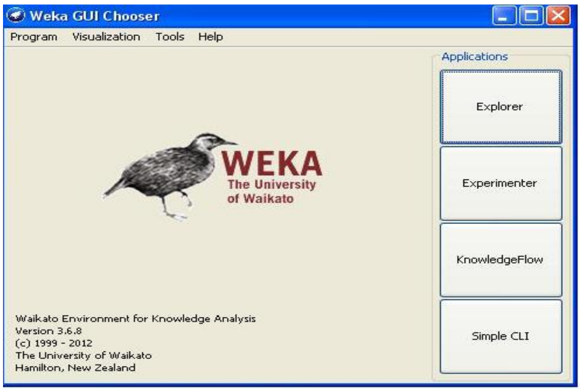
Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag and drop interface. One advantage is that it supports incremental learning.

Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

This Java-based version (Weka 3) is used in many different application areas, in particular for educational purposes and research. There are various advantages of Weka:

1. It is freely available under the GNU General Public License.

2. It is portable, since it is fully implemented in the Java programming language and thus runs on almost any architecture.



Waikato Environment for Knowledge Analysis
Version 3.6.8
(c) 1999 - 2012
The University of Waikato
Hamilton, New Zealand

Fig3

3. It is a huge collection of data preprocessing and modeling techniques.

4. It is easy to use due to its graphical user interface.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

C. Data preprocessing steps in Weka:

Firstly, Run Weka software, launch the explorer window and select the “Preprocess” tab. Then Open the iris data-set, and enter what information do you have about the data set (e.g. number of instances, attributes and classes)? What type of attributes does this data-set contain (nominal or numeric)? What are the classes in this data-set?

Which attribute has the greatest standard deviation? What does this tell you about that attribute? After entered the data set under “Filter” choose the “Standardize” filter and apply it to all attributes. What does it do? How does it affect the attributes’ statistics? Click “Undo” to understanding the data and now apply the “Normalize” filter and apply it to all the attributes. What does it do? How does it affect the attributes’ statistics? How does it differ from “Standardize”? Click “Undo” again to return the data to its original state. At the bottom right of the window there should be a graph which visualizes the data-set, making sure “Class: class (Nom)” is selected in the drop-down box click “Visualize All”. What can you interpret from these graphs? Which attribute(s) discriminate best between the classes in the data-set? How do the “Standardize” and “Normalize” filter affects these graphs? Under “Filter” choose the “Attribute Selection” filter. What does it do? Are the attributes it selects the same as the ones you chose as discriminatory above? How does its behavior change as you alter its parameters?

III. CLASSIFICATION IN WEKA

Classification is the process of finding a set of models that describe and distinguish data classes and concepts, for the purpose of being able to use the model to predict the class whose label is unknown.

Classification is a two step process, first, it build classification model using training data. Every object of the dataset must be pre-classified i.e. its class label must be known; second the model generated in the preceding step is tested by assigning class labels to data objects in a test data set. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The model is represented as classification rules, decision trees, or mathematical formulae. Second step is model usage. It is for classifying future or unknown objects. It estimates accuracy of the model. The known label of test sample is compared with the classified result from the model.Model construction describe a set of predetermines classes. 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. If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known. We assume [3] that after data preparation, we have a data set where each record has attributes X1,X2, X3 upto Xn, and Y. Our Goal is to learn a function f:(X1,…,Xn)→Y, then use this function to predict y for a given input record (x1,…,xn).In this Classification: Y is a discrete attribute, called the class label whether Prediction: Y is a continuous attribute. Classification Called supervised learning, because true labels (Y- values) are known for the initially provided data .Some application involve credit approval, target marketing, medical diagnosis, fraud detection..

A. STEPS INVOLVE IN WEKA

Basically there are four steps involved in weka for classification.

- Preparing the data
- Choose classify and apply algorithm
- Generate trees
- Analysis the result or output

Firstly, Prepare the data, load the data and the data should be in .arff format. After loaded the data choose classify then choose classification algorithm and generate the trees.

IV. CLUSTERING IN WEKA

This pattern divides the records in database into different groups. In the same group, the groups have the similar properties. Between groups the differences should be as bigger as possible, and in the same group, the differences should be as smaller as possible. There is no predefined class that’s why its comes under the unsupervised learning .some examples of cluster applications are seen as in marketing, land use , insurance , earthquake studies and in city planning Methods [5] involve in cluster analysis are portioning methods, hierarchical Methods, density-Based Methods, grid-Based Methods, model-Based Methods, clustering high-dimensional data, constraint-based clustering [8] , and Outlier analysis..Many algorithms exist for clustering. Following figures showing three major clustering methods and their approach for clustering.

(I) K-means Clustering

The term "k-means" was first used by James Mac Queen in 1967 [14], though the idea goes back to 1957 [15]. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published until 1982. K-means is a widely used partitioned clustering method in the industries. The K-means algorithm is the most commonly used partitioned clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time.

(II) Hierarchical clustering

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram [16]. Two types are there.

Agglomerative (bottom up)

1. Start with 1 point (singleton).
2. Recursively adds two or more appropriate clusters.
3. Stop when k number of clusters is achieved.

Divisive (top down)

1. Start with a big cluster.
2. Recursively divides into smaller clusters.
3. Stop when k number of clusters is achieved.

(III) Density based clustering

Density-based clustering algorithms try to find clusters based on density of data points in a region. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (Min Pts). One of the most well known density-based clustering algorithms is the DBSCAN [7]. DBSCAN separates data points into three classes:

1. Core points: These are points that are at the interior of a cluster.
2. Border points: A border point is a point that is not a core point, but it falls within the neighborhood of a core point.
3. Noise points: A noise point is any point that is not a core point or a border point.

A. STEPS INVOLVE IN WEKA:

Load the data file browsers .arff into WEKA using the same steps we used to load data into the **Preprocess** tab. Take a few minutes to look around the data in this tab. Look at the columns, the attribute data, the distribution of the columns, etc. With this data set, we are looking to create clusters, so instead of clicking on the **Classify** tab, click on the **Cluster** tab. Click **Choose** and select **technique** from the choices that appear.

V. CONCLUSION AND FUTURE WORK

In this paper, firstly we shortly introduce the concepts of Data Mining then data preprocessing and sits steps in Weka and techniques such as classification and clustering by Weka steps. We describe it by the Weka tool. We gave the introduction to Weka 3.6.8. There is no doubt that the data mining is the useful term in the present and future

REFERENCES

1. "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
2. "DATA MINING TECHNIQUES CLASSIFI ATION AND PREDICTION "by Han/Kamber/Pei, Tan/Steinbach/Kumar, and Andrew Moore MirekRiedewald
3. "CLASSIFICATION AND PREDICTION IN A DATA MINING APPLICATION " by SERHAT ÖZEKES and A.YILMAZ ÇAMURCU 2 Istanbul Commerce University, Ragıp Gümüş pala Cad. No: 84 Eminönü 34378, Istanbul – Turkey
4. "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.
5. DATA MINING TECHNOLOGY by Jiawei Han Department of Computer Science University of Illinois at Urbana-Champaign
6. Amazon Elastic Compute Cloud (Amazon EC2),http://aws.amazon.com/ec2/,2009.
7. David Heckerman. Bayesian Network for Data Mining. Data Mining and Knowledge Discovery, 1997:79-119..
8. David Hand, Heikki Mannila and Padhraic Smyth. Principles of Data Mining, the MIT Press, 2001:1-5...
9. A Short Introduction to Data Mining and Its Applications Zhang Haiyang
10. Google Web Applications http://www.google.com/
11. Ritu Chauhan, Harleen Kaur, M.Afshar Alam, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010
12. J.R Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufman, 1993.
13. S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Preprocessing for Supervised Leaning", *International Journal of Computer Science*, 2006, Vol 1 N. 2, pp 111–117.
14. MacQueen J. B., "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.University of California Press. 1967, pp. 281–297.
15. Lloyd, S. P. "Least square quantization in PCM". IEEE Transactions on Information Theory 28, 1982,pp. 129–137.
16. Manish Verma, MaulySrivastava, NehaChack, Atul Kumar Diswar and Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384
17. Timonthy C. Havens. "Clustering in relational data and ontologies" July 2010.
18. Weka: http://www.cs.waikato.ac.nz/~ml/weka/

AUTHORS PROFILE

Swasti Singhal received B.tech degree in Information Technology and pursuing M.Tech in Computer Science and Engineering from Amity University, Noida. Presently she is working in Galgotia's College of engineering, Greater Noida. Her research interest includes Data mining, and software testing.

Monika Jena, Assistant Professor at Amity School of computer science, Noida. She has 12 years of experience in teaching. Her area of interest is in Network Programming and multimedia networks. She has number of publication in reputed journals