International journal of Master of Engineering Research and Technology | Vol-2, Issue-6, 2015
ISSN (ONLINE) 2394-6172

Analysis of Data Mining Tool: Weka

Harsha Jain
M.Tech Scholar
Amity university, Gwalior, (India)
Harshajain249@gmail.com

Dheeraj Pal
Asst. Professor
Amity university, Gwalior, (India)
Spal.dheeraj@gmail.com

Abstract
In this paper we are describing the mining of data it can be easier with the help of tool weka, as now a days there is lot of data which cannot be mine easily so there is a tool which help in mining the data and in this tool there is a lot of features we can directly apply them all in our data and get the proper and efficient results. This paper describes the basic principle of mining the data that is data sources, classifiers, clustering, evaluation association rules, data pre-processing, & its tools. Basically the paper is going to present the full description of the features of weka tool or in other words we can understand the use of tool weka.

Keywords: Data sources, data mining, weka tool, classifiers, filters, association rules, clustering, data pre-processing.

INTRODUCTION

Data mining is a term which is very popular now a days as the collection of data increased day by day so, mining of data it means collection of important data from huge data is a big task, we can understand it by an example like mining of gold, mining of coalmining of diamonds etc. in that all materials are necessary but still we mine for better and best. Same as in data mining from lot of data we have to find out the useful or an important data. So it can be easier with the help of tool in this we have provided a lot of or number of algorithms that can be implemented for finding or extract the useful information. As in other words data mining also known as knowledge Discovery database (KDD), the overall goal of mining is to extract information and convert it in to an understandable form for future use. Data mining involves six common classes of task classification, clustering, regression, data pre-processing etc. which are going to discuss in detail in other portion of this paper, weka tool have many versions we are here discussing weka version 3-7-12 (Waikato Environment For knowledge Analysis tool). Firstly, we are going to discuss about the data sources that in which format the file is loaded. After that how the data is processed in the tool and the other features such as association rules, filters, data sinks, classification, clustering, evaluation visualization etc. In classifiers there is some functions such as trees, rules etc. In filters there are two types such as supervised & unsupervised. Weka tool had four applications by which we can see the data results in different forms such as Explore, Experiment, knowledge flow environment, and Command line interface (CLI) we discuss all this features in detail. These all are the weka tool applications first of all we discuss about the processing of data means why the data process is necessary after we apply the mining. Data processing how the data process how we can get the data after mining the data.

DATA PROCESSING- Today's real world databases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple heterogeneous sources. Low quality data will lead to low quality mining results. How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results? There are number of data pre-processing techniques by applying all these techniques we can process the data in a perfect manner. Techniques which are used through which the data is processed or go through are as follows Data cleaning, Data integration, Data transformation, Data reduction, Data discretization .Data cleaning can be applied to remove noise and correct inconsistency in data. Data integration merges the data from multiple sources into coherent data store, such as data warehouse. Data transformation such as normalization may be applied. For example normalization may improve the accuracy and efficiency of mining the algorithms involving distance measurement. Data reduction can reduce the data size by aggregating, and generalized. Now the question arise why should we process the data? Imagine if you are manager in a company, and suddenly you have to inspect the database of company while, identifying all the attributes dimensions to be included in your analysis such as sold items, purchase goods, cost of items, customer information, date of items etc. Then you noticed that certain of attributes and tuples not recorded yet. This is called inconsistency of data where the information is missing. Incomplete, noisy, and inconsistent data are commonplace properties of large real world databases and data warehouses. Incomplete data can occur from number of reasons .By the processing of data. Data quality can be measure in terms of accuracy, completeness, timeliness, believability, interpretability. Now if the data is processed than why should we clean the data? Sometimes the data is incomplete, inconsistent and incompleteness of data i.e. shortage of attributes. Data integration can be defined as yet some attributes representing a given concept may have different names in different databases causing inconsistencies and redundancy. For example the attribute for customer identification maybe referred to as customer id in one data store and customer id in another. Typically data cleaning, data integration are performed as a pre processing step. When preparing the data for a data warehouse. Additional data

cleaning can be performed to detect and remove redundancies that may have resulted from data integration. Data integration is a term covering several areas such as Data warehousing, Data migration, Entrprise/application/information integration. In data transformation the data are consolidated in to form appropriate for mining. Data transformation involves following techniques such as smoothing which help to remove the noise from data, Normalization it is used for where the attribute data are scaled. Data aggregation where summary or aggregation attributes are applied, generalization where low level or primitive data are replaced by higher level concept through the use of concept hierarchies. Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. There are a number of strategies for data reduction. These include data aggregation (e.g building a data cube), attribute subset selection (eg. removing irrelevant attributes through correlation analysis) dimensiality reduction (e.g, using encoding schemes minimum length encoding or wavelets), and numerosity reduction (e.g clusters and parametric model).

WEKA INTERFACE-WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. WEKA supports many data mining tasks such as data preprocessing, classification, clustering, regression and feature selection to name a few. The workflow of WEKA would be as follows first we enter the data than data is pre-processed than mining the data and final knowledge data. WEKA was written in java code, it is a free software which is easily available and it is also a platform independent. weka that contains a collection of visualization tools and algorithms for data analysis and predictive modeling together with graphical user interfaces (GUI) for easy access to this functionality.

GUI chooser consists of four buttons: Explore Experiment, Knowledge flow environment and simple CLI.
Explore: An environment for exploring data. It supports data-pre- processing, attribute, selection, learning and visualization.
Experimenter: An environment for performing experiments and conducting statistical tests between machine learning algorithms.
Knowledge Flow: It is similar to Explorer but has a drag-and-drop interface. It gives a visual design of the KDD process.
Simple CLI: Provides a simple command-line interface for executing 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 java programming language and thus run on almost any architecture.
3) It is a huge collection of data pre processing and modeling techniques.
4) It is easy to use due to its graphical interface weka supports several data mining task data preprocessing, clustering, classification, regression, visualization and feature selection.

DATA PRE-PROCESSING STEPS IN WEKA

How to load a file in weka. In addition to a native data ARFF data file format, weka has capability to read in ".csv" format files. This is fortunate since many databases or spreadsheet applications can save or export data into flat files in this format. As can be seen in the sample data file, the first row contains the attribute names (separated by commas) followed by each data row with attribute values listed in the same order (also separated by commas).In fact once loaded into weka, the data set can be saved into ARFF format. Weka performs a series of operations using weka attributes and discretization filters and then performs association rule mining on the resulting data set. While all of these operations can be performed from the command line, we use the GUI interface for WEKA Explorer. Initially (in the Preprocess tab) click "open" and navigate to the directory containing the data file (.csv or .arff).In this case we will open the data file. Since the data is not in ARFF format, a dialog box will prompt you to open this. You can click on 'USE CONVERTER' and click ok in the next dialog box. Once the data is loaded weka will recognize the attributes and during the scan of data will compute some basic statistics on each attribute. The left panel shows the list of recognized attributes, while the top panels indicate the list of base relation and the current working relation. Clicking on any attribute will shows the basic statistics on that attribute. What type of attribute thus this data-set contains (nominal or numeric?) What are the classes in this data set? Which has the greatest standard deviation? What does it's telling 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.

CLASSIFIACTION IN WEKA

Classification (also known as classification trees or decision trees) is a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. In Classification, training examples are used to learn a model

that can classify the data samples in to known classes. The classification involves following steps:

a. Create training data set.
b. Identify class attribute and classes.
c. Identify useful attributes for classification (Relevance analysis).
d. Learn a model using training examples in Training set.
e. Use the model to classify the unknown data samples.

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 that after data preparation, we have a data set where each record has attributes X1, X2, X3 up to Xn, and Y. Our Goal is Y, and then uses this function to learn a function f :(X1… Xn) 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. 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.

CLUSTERING IN WEKA

This pattern divides the record in data base in to different groups. In the same group, the groups have the similar properties. Between groups the difference should be as bigger as possible and in the same group the difference should be as smaller as possible. There is no pre-defined class that's why it comes under the unsupervised learning. Some example of cluster application is seen as in marketing, land use, insurance, earth quake studies and in city planning. Method involves in cluster analysis are portioning methods, hierarchical methods, density based methods, clustering high dimensional data, constraint based clustering, and outliner analysis. Many algorithms exist for clustering following figure showing the three major clustering methods and their approach for clustering.

K-MEANS CLUSTERING

The term "k-means" was first used by James Mac Queen in 1967. The standard algorithm was first proposed by Stuart Loyd in 1957. K-means is a simple technique for clustering analysis. Its aim is to find the best division of n entities into k groups (called clusters), so that total distance between the group's members and corresponding centroid, irrespective of the group is minimized. This algorithm randomly selects K number of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and cluster mean. Then it computes new mean for each cluster. This process iterates until the criterion function converges.

Characteristics of K-means clustering This algorithm attempts to determine K partitions that minimize the squared error functions. It is relatively scalable and efficient in processing large data sets because the computational complexity of the algorithm is O (nkt)

Where N=total number of objects, K=number of clusters, T=number of iterations

This method often terminates at local optimum.

Hierarchical - clustering: Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. 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.

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 cluster neighborhood for a given radius (Eps) has to contain at least a minimum number of instance (Min pts) .One of the most well known density based clustering algorithms is the DBSCAN, DBSCAN separates data base in to three classes:

1. Core points: These are points that are at the interior of a cl User.
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.

CONCLUSION

In this paper, firstly we describe the data mining how the data process and why it is essential to process the data, brief description of the data processing, than the steps of processing the data in weka tool 3-7-12 than, classification of weka it means how to create a tree, graph how we can perform the experiment by this tool where all the clustering and algorithms are available. So in this paper the use of tool weka is described and by which we can use the tool, without any doubt weka tool is very important in the field of data mining as it performs many experiment through which we can analysis our result and in future also this tool can be used in research field, as weka tool had so many versions it means there is a wide use of this tool and the future research is also can be done by with the help of this tool.

REFERENCES

[1] A Study on WEKA Tool for Data Preprocessing, Classification and Clustering by "Swasti Singhal", "Monika Jena" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013.

[2] K-Means Clustering in Spatial Data Mining using Weka Interface by "Ritu Sharma(Sachdeva) M.Tech Student Department of Computer Science Jamia Hamdard University, Anita Rani Lecturer DAV Centenary College. International Conference on Advances in Communication and Computing Technologies (ICACACT) 2012 Proceedings published by International Journal of Computer Applications.

[3] Comparison of major clustering algorithm with tool weka" 2014 international conference on advance in ict for emerging regions.

[4] Utility of Association Rule Mining: a Case Study using Weka Tool by "A.lekha, Dr.C.V SriKrishna and Dr.Viji vinod Emerging Trends in VLSI, Embedded System, Nano Electronics and Telecommunication System (ICEVENT), 2013 International Conference on.

[5] An overview of big data opportunities, applications and toolsby"Benjelloun,FtaimaZahra,Lahcen,AyoubAit,Belfkih,Samir,"International journal of Intelligent system and computer vision(ISCV)2015.

[6] Predictive analytic using data mining technique by"Gulati, Hina", computing for sustainable global development (Indicator).2015 2nd International conferences.

[7] Comparison of Classification Techniques for predicting the performance of Students Academic Environment by "M. Mayilvaganan, D. Kalpanadevi", 2014 International Conference on Communication and Network Technologies (ICCNT).

[8] Multi-Level Counter Propagation Network For diabetes classification by"Velu.C.M, kashwan, K.R, Signal processing image processing pattern recognition (ICSIPR), 2013 international conference.

[9] Predicting disease by using data mining technique based on health care information by "Chan-chine-chung, wang, huang, granual computing (Grc), 2012 International conference.