WEKA Tool With Different Data Mining Technique

1 *Monika Lambsonge   2*Girish Talmale 3* Poonam Prasad

1*.2* G.H Raisoni College Of Engineering, 3* Senior Scientist CSIR-NEERI
Computer Science And Engineering                     Analytical Instrumentation Department
Nagpur                                                                   Nagpur

Abstract:
Now a day’s huge data is a very important issue today. Everyday users suffer major problems like “How to analyze large data or how to predict a large amount of data?” here WEKA tool gives a solution to these types of problems, we can easily solve the big data issue. Weka is an extreme engineering technology that helps to analyze big data and sort out the solution by using different algorithms. analysis of big data plays an important role for a better solution.

Index term: Data mining techniques, Weka tools, Weka Explore interface, classification, clustering

I. Introduction
In this article, a vast data is collected daily. analyze that large data and abstract meaningful information from that data. peoples perform many types of work in the world like industrial, banking, medical, business in the company, shop, mall, etc. so they have a lot of data but nobody have a time to sort important data manually there is a need to extract our important data in easy way. We find out the effective solution that is WEKA tool. We can easily sort our important data daily and make our work easy that helps to predict large data and give a better solution by using different data mining algorithms.

II. About Weka:

Waikato Environment for Knowledge Analysis shortly denoted as WEKA developed in 1993 from the Waikato University in New Zeland, it is free open-source software fully implemented in JAVA, easily available on the internet. WEKA is mainly created to analyze several amount of big databases like industrial, banking, medical, school, colleges, etc. as well as support several data mining tasks like pre-processing, clustering, classification, regression, visualization. WEKA helps to support big data with the help of different algorithms and find out the differences between two similar variables. Graphical method support to imagine the future selection tasks.

WEKA is a open source tool developed by the Waikato university from New Zealand. This is a free software under the GNU license [6].

- WEKA is fully created in the JAVA and easily runs on a new platform.
- A extensive compilation of data pre-processing and unloading techniques.
- Ease to understand graphical user interface model.

WEKA support to collect several data like clustering, classification and feature selection algorithms.
III. Classification method of weka [1]:

1. Clustering [7]:
   Clustering is a technique of segmentation a set of data that builds a valuable cluster of objects which have similar characteristics. Clustering defines classes and puts objects in each class. For example, in the prediction of obnoxious gases by using clustering we can sort out that the list of obnoxious gas which has the same risk factor.

2. Classification [1][8]:
   Classification is the method to identify the objects in a group to target classes. In the classification method mainly used mathematical techniques such as linear programming, decision trees, statistics, and neural network. By using that we can learn how to classify the group of objects.

   - **A decision tree:**
     In a decision tree, there are some nodes like internal, branches, leaf. The internal node is used to refer a test on attribute[15]. Branch node is used to refer the result of a test & leaf node is used to hold a class label. The upmost node is a root node in the tree.

   - **Rule induction:**
     Rule induction is a data mining technique of deducing if-then rules from a data set. These symbolic decision rules explain an inherent relationship between the attributes & class labels in the data set. But, in general, these rules hold & are deduced from real-life experience based on our everyday observation.

   - **Neural Network:**
     Neural networks and data mining. An artificial neural network, often just called a neural network is a mathematical model inspired by biological neural networks [14]. A neural network consists of an interconnected group of artificial neurons and it processes information using a connectionist approach to computation.
3. Prediction [1]:
Prediction is the method to identify the relations between two variables.

4. Association [1]:
Association is the method to find out the relation between different modules that are present in the same database.

IV. USER INTERFACE

When we open the WEKA tool, there are five applications in WEKA GUI chooser. Our main work in explorer, because explorer is a user interface, we can upload our dataset and explore it by using different algorithms. The selection of algorithm apply according to our dataset for getting the better result.

1. PREPROCESS
In pre-processor, there has a facility for upload data from a database, comma-separated values (CSV) file, etc. and pre-processing this data using filtering algorithms use for filtering the database according to their specific attribute and instances.

   - FILTER
     There are many filter present in Weka used for clean up the unwanted dataset like attributes and instances which can be easily removed by the use of filter.
     In filter two different packages i.e supervised and unsupervised.

   - SUPERVISED
     Contains a supervised filter that is the filter that takes class distribution into account must utilize for eg. the weka.filters.supervised filter interface.
     
     Attribute:-column-wise filtering.
     Instances: - row-wise filtering.

   - UNSUPERVISED
     Contains unsupervised filters that are filter work without taking any class distribution method into account, must utilize the weka.filters.unsupervised filter interface.
     
     Attribute:-filter column-wise.
     Instances:-filter row-wise.

2. CLASSIFIER
In classifier, there has a facility to apply classification and regression algorithms for eg. predictive model, visualize erroneous prediction, receiver operating characteristic curves, visualization etc.

3. ASSOCIATE
In associate, there has a facility to provide association rule learners to identify the interrelationship between the attribute in data.

4. CLUSTER
In a cluster there has a facility of clustering technique for eg. simple K-means.
5. SELECT ATTRIBUTE
In select attribute, there has a facility to provide the most predictive attribute.

6. VISUALIZE
In visualize there has a facility to show scatter plot matrix, and analyze it using several selection operators.

V. CLASSIFICATION AND CLUSTERING Technique

![Classifier and clustering Algorithm](image)

1. **Classifier Algorithm:**
Classifier algorithm there are four different algorithm techniques.

1.1) **MULTILAYER PERCEPTRON**
Multilayer perceptron [10] is classified as feed forward ANN[15] leave out for the input node. MLP solve a non-linear optimization problem as well as use a back propagation algorithm method which belongs to supervised learning.

1.2) **M5 RULES**
M5 is a plain multilayer perceptron model which belongs to the artificial neural network.

1.3) **REP TREE**
REP tree algorithm is a fast decision tree which is based on the C4.5 algorithm [12][10]. It can able to produce classification and regression algorithms however random feature method generates multiple classifiers as well as random decision tree method generate combines bagging.

1.4) **K-NN**
K nearest neighbors (K-NN) [8] is a simple algorithm. K is used to denote the number of nearest neighbors. this algorithm help to calculate the distance between the query - instances [13] based on Kth minimum distance.

1. **Clustering algorithm:**
In clustering there are three different clustering techniques.

2.1) **FARDEST FIRST**
Farthest first is a fast and greedy algorithm based on clustering. In this algorithm, k points are firstly select as a greedy for selecting the farthest first center and remaining points are cluster whose center is closest. Clustering is the method of a group of similar objects.

2.2) **SIMPLE K-MEANS**
Simple k-means is the method of clustering [11]. Which is used in data mining techniques. Here K is the number of clustering points which is defined center[15] however this algorithm uses the Euclidean distance to measure the distance between instances and clusters [10].

2.3) **SIMPLE EM**
Expectation-Maximization is an iterative aspect method to find MLE or MAP evaluation of data distribution when data is partially missing or hidden. The EM iteration alternates between the performance of one expectation E-step as well as one maximization M-step [9].

IV. Literature survey on WEKA tool:

The author Kulwinder Kaur et al [1] studied the data mining technique with WEKA tool. in this paper author claim an data mining technique has affectively used been in many fields including large data sets such as banking, medical, business, etc. in this paper author represents the several data mining technique in WEKA such as clustering, classification, association, prediction.

The author R.Ramya et al [2] studied the plant disease detection by the used of WEKA tool. in this paper author research on agricultural field. they have represents several different algorithms such as k-means, decision tree, multilayer perceptron, Naive Bayes, SUM, RFA for predicting disease using WEKA tool.
The author Sarangam Kodati et al [3] analysis the heart disease by using the data mining methods in WEKA and orange tool. In this paper, the author represents various data mining methods such as SVM, NB, RF, KNN algorithm for analysis of the heart disease patients.

The author Rausheen Bal et al [4] comparatively studied about metadata with the classifiers algorithm by used the techniques of WEKA tool. In this article, the author represents different classifiers algorithm technique.

The author Dr.Anil Sharma et al [5] having comparatively study about data mining techniques, tools and parameters. In this paper author represents different data mining methods to the analysis of a huge database. These tools greatly support in business database. In this paper authors did experiment with two different database on WEKA tool.

Table 1. Studied research paper on WEKA

<table>
<thead>
<tr>
<th>Sr.no</th>
<th>Title name</th>
<th>Author name</th>
<th>country</th>
<th>year</th>
<th>references</th>
<th>Method used</th>
</tr>
</thead>
</table>


13. V. Mhetre and M. Nagar | Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA | India | 2017 | https://ieeexplore.ieee.org/document/8282735 | Classification, prediction


VII. CONCLUSION:
In this article, we studied the different datasets that are analyzed in WEKA through the use of different data mining techniques and algorithms. WEKA is easily handled all types of huge data like medical, banking, business, etc. that help to predict huge data and sort important information related to your dataset. WEKA included various algorithms which is important to analyze large dataset.

References:
[3] Sarangam Kodati & Dr. R. Vivekanandam,” Analysis of Heart Disease using in Data Mining Tools Orange and Weka”, Global Journals,Volume 18, Issue 1, Version 1.0, Year 2018,India, ISSN: 0975-4172 & Print ISSN: 0975-4350
[5] Dr. Anil Sharma, Balrajpreet Kaur,”a research review on comparative analysis of data mining tools, techniques and parameters”, International Journal of Advanced Research in Computer Science, Volume 8, No. 7, July – August 2017,Punjab,India, ISSN No. 0976-5697
[13] Sarangam Kodati & Dr. R. Vivekanandam, Sri Satya Sai University,” Analysis of Heart Disease using in Data Mining Tools Orange and Weka” Publisher: Global Journals, Online ISSN: 0975-4172 & Print ISSN: 0975-4350
[14] Dr. Anil Sharma, Balrajpreet Kaur, Department of Computer Science and Applications, LPU, Phagwara 144411, Punjab,“ a review research on comparative analysis of data mining tools, techniques and parameters”, ISSN No. 0976-5697
[15] R.Ramya1, Dr.P.Kumar2, D.Mugilan3, M.Babykala4 Assistant Professor1, 3,4, Department of ECE, K.S.Rangasamy College of Technology Professor2, Department of ECE, K.S.Rangasamy College of Technology, Tiruchengode Tamilnadu, India, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, p-ISSN: 2395-0072