COMPREHENSIVE ANALYSIS OF DATA MINING CLASSIFIERS USING WEKA

Hemlata
Department of Computer Science & Applications
Maharshi Dayanand University
Rohtak, Haryana, India
ORCID: 0000-0002-6105-7399

Abstract: Data Mining or knowledge extraction from a large amount of data i.e. Big Data is a crucial and important task now a days. Data Mining and its applications are the most promising and rapidly emerging technologies. A number of Open Source Big Data Mining tools are available. Users or researchers must have the knowledge of the characteristics, advantages, capabilities of the tools. This paper gives an experimental evaluation of the algorithms of WEKA. The classification algorithms are analysed on the basis of accuracy and precision by taking the real dataset. The paper presents the comprehensive evaluation of different classifiers of WEKA. It will help the future researchers or data analysing business organisation to select the best available classifier while using WEKA.

Keywords: Data Mining, Big Data, Classifiers, Big Data Mining Tools, Weka

I. INTRODUCTION

There is a rapid development in Information Technology and Data Science industry in last few decades. As the size of data increases its retrieval is becoming complex. The extraction of meaningful data from the huge amount of data is a challenging task. The process of discovering knowledge or meaning from the unstructured and unformatted data is known as Data Mining.

Many Big Data Mining Tools are available but to select the best is very difficult. This paper gives a comprehensive analysis of the most important Big Data Mining Tool - WEKA. It presents the experimental evaluation of the data mining classifiers available in WEKA. The analysis is done by implementing the classification algorithm on real dataset and comparing the results. The comparison and analysis is done on the basis of precision, accuracy and confusion matrix.

II. DATA MINING

Data Mining means the practice of examining large pre-existing databases in order to generate new information. It is the procedure of pattern discovery from huge datasets. It is an essential process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science [1]. Simply, the process of extracting meaningful and usable data from larger set of any raw data is known as data mining. It indicates the procedure of analysing data patterns in large batches of data using one or more software [2]. Data mining is also known as Knowledge Discovery in Data (KDD). It is the process of collecting, searching through, and analyzing a large amount of data in a database, as to discover patterns or relationships. The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases is data mining [3]. It is the nontrivial extraction of implicit, previously known, and potentially useful information from data. Data Mining is the science of extracting useful information from large data sets or databases [4]. It uses computational techniques from statistics, machine learning, and pattern recognition. Data mining is the process of applying computer-based methodology for knowledge discovery from data [5]. The two components are:-

- Knowledge Discovery
  Concrete information gleaned from known data. Data you may not have known, but which is supported by recorded facts [6].
- Knowledge Prediction
  Uses known data to forecast future trends, events, etc. (i.e.: Stock market predictions).

Data mining is highly useful in the following domains [7] –

- Market Analysis and Management.
- Corporate Analysis & Risk Management.
- Fraud Detection

A typical data-mining system consists of a data-mining engine and a repository that contains the data-mining artifacts, such as the models, created in the process. The process of data mining takes data as the input and gives knowledge as the output [8]. It is graphically represented in Figure 1.
A. Techniques of Data Mining
Various techniques of Data Mining are:

1. Clustering
Clustering is the method of grouping similar objects. A cluster of similar data objects can be treated as one group. In cluster analysis, we first partition the set of data into groups based on similarity and then assign the labels to the groups. Clustering analysis can be applied in diverse fields such as data analysis, image processing, market research and pattern recognition [9]. It helps the marketing executives in discovering different groups in their customer base on the basis of customer’s purchasing patterns. Cluster analysis is a tool to gain insight into the distribution of data to observe characteristics of each cluster [10].

2. Association Rules Discovery
The process of identifying the similar patterns which frequently go together is known as association rules. The data objects with associated or similar characteristics are grouped under associated rules. Associations are used in retail sales to identify patterns that are frequently purchased together [11]. This process refers to the process of uncovering the relationship among data and determining association rules [12]. For example, a retailer generates an association rule that shows that 70% of the time milk is sold with bread and only 30% of the times biscuits are sold with bread.

3. Regression
Regression methods are used to predict the value of the response variable from one or more predictor variables where the variables are numeric[13]. Different forms of Regression are Linear, Multiple, Weighted, Polynomial, Nonparametric, Robust.

4. Sequential Pattern Discovery
Sequential pattern discovery or Sequential pattern Mining is specialized data mining task to analyse sequential data and to discover sequential patterns. It is the process of discovering interesting sub-sequences in a set of sequences. The criteria of sub sequence can be its frequency of occurrence, length or profit. As the data can be encoded in sequences of symbols, sequential pattern mining can be applied in many fields such as e-learning, market basket analysis, bioinformatics, and many more[14]. It has been used to find sub-sequences that appear in a sequence database i.e. that are common to several sequences. These sub-sequences are called frequent sequential patterns.

The patterns which occur very frequently in the data are called frequent patterns [15]. Various kinds of frequent patterns –
- Frequent Item Set – The item sets which occur frequently together are called Frequent Item Set. For example, bread and butter.
- Frequent Subsequence – A sequence of patterns which appear frequently such as purchasing a smartphone is followed by a memory card.
- Frequent Sub Structure – Substructure means different structural forms, such as graphs or trees. The occurrence of these sub-structures combined with item-sets or sub-sequences is known as Frequent Sub Structure.

5. Collaborative Filtering
Collaborative Filtering is the process of predicting the behaviour of a user on the basis of the behaviour of many other users with similar taste. It filters the interests of the users and predicts on the collaborative (collective) result. In other words, collaborative filtering is the process of filtering
for information or patterns using various collaboration techniques [16]. Collaborative filtering can be used in very large database i.e. Big Data. It can be applied in many different types of data such as sensing and monitoring data, financial data, electronic data etc.

Collaborative filtering usually requires three inputs [17]:
- Users' active participation
- Method to represent users' interests
- Algorithms

6. Classification

Classification is the most important data mining technique which is based on machine learning. Classification is used to group the data item set in one of the predefined set of classes or groups. Classification method uses the mathematical and statistical techniques such as decision trees, linear programming, and neural network. In classification, a model is built that can learn how to classify the data items into groups [17]. Classification is a data mining function that assigns items in a collection to target categories or classes. The main objective of classification is to accurately predict the target class for each data [18]. For example, a classification model identifies loan applicants as low, medium, or high credit risks. Various methods used for data mining classification are:-
- Decision tree:
  Decision trees are powerful and popular tools for classification and prediction. Decision tree is a predictive model that can be viewed as a tree. The goal of decision tree is to create a model that estimates the value of a target variable based on several input variables [19]. The nodes represent the input variables; whereas the edges to children represent the possible values of that input variable. Decision tree is a classifier model in the form of a tree structure having two types of nodes [20]:
  - A leaf node – represents the value of the target attribute.
  - A decision node – represents some tests to be carried on a single attribute value, with one branch and sub-tree for each possible outcome of the test.
- Rule-based methods:
  Rule-based classifiers categorize the data with the use of a collection of “if . . . then . . .” rules. The predecessor of the rule is an expression made of the combination of attributes. The resultant rule is a positive or negative classification. For building a rule-based classifier we have to derive rules directly from the dataset [21]. Most important advantage of rule-based classifiers is that they are extremely expressive because they are symbolic. Rule-based classifiers or decision trees are easy to interpret, easy to generate and they can classify new instances efficiently.
- Memory based learning:
  Memory-based learning method is a collection of learning algorithms which compare the objects of new problem with the objects of training. It does not perform explicit generalization [22]. It can also be referred to as instance-based learning method because it builds hypotheses directly from the training objects which mean that the hypothesis complexity can grow with the growth of data. The ability to adapt to its model to previously unseen data is the main advantage of memory-based learning over other methods of machine learning [23]. One example of memory based learning algorithm is the k-nearest neighbour algorithm.
- Neural networks:
  Artificial neural networks (ANNs) are based on biological neural networks i.e. the nervous system of our brain. They are used to approximate functions that depend on a large number of inputs [24]. Artificial neural networks are the systems of interconnected “neurons”. These neurons calculate values from inputs. These neural networks are competent in machine learning and pattern recognition [25]. An artificial neural network functions by creating connections between many elements each element represent a single neuron of a brain. These neurons are constructed and simulated by a computer system. Every neuron takes many signals as input and gives a single signal as output. This output is then sent to the next neuron. The neurons are strongly interconnected and organized into different layers. The input layer receives the input and the output layer produces the final output [26]. In general one or more hidden layers are sandwiched in between the two. This layer system makes it impossible to forecast or know the exact flow of data.
  - Bayesian network:
    The Naive Bayes Classifier technique is based on Bayesian theorem of probability. It is particularly used when the dimensionality of the inputs is high. The Bayesian Network Classifier is efficient in computing the most possible output based on the input. In this classifier it is possible to add new raw data at runtime and get a better probabilistic classifier [27]. When the class variable is given, Naive Bayes classifier considers that the presence of a particular attribute is not related to the presence of any other attribute. Bayesian networks are directed acyclic graphs (DAG) whose nodes represent random variables. The nodes can be any observable quantities, variables, unknown parameters or hypotheses [28]. Edges are the conditional dependencies. Nodes which are not connected represent the independent variables. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node.
  - Support vector machines:
    Support Vector Machines are learning models which are supervised and have learning algorithms to analyze data and patterns. SVMs are also known as support vector networks. The basic SVM is a binary linear classifier which accepts input data set and gives two possible outcomes [29]. Well-founded theoretical approach, quality of generalization and ease of training are the basis of SVM. The support vector machine is an algorithm to learn and classify rules of data. The SVM can be used to learn polynomial, radial basis function (RBF) and multi-layer perceptron (MLP) classifiers [30].

III. EXPERIMENTAL ANALYSIS

Analysis of Big Data Mining tool Weka is done by following steps:
- Real Dataset Taken
- Selection of Evaluation Technique

A. Data Set Choosen

Real Dataset is taken from online UCI Machine Learning Repository. The dataset is chosen and used for the
performance analysis of various data mining classifier algorithms of WEKA. The details of the dataset are given in table 1:

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>No. of Instances</th>
<th>No. of Attributes</th>
<th>No. of Class Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pima_Diabetes</td>
<td>768</td>
<td>9</td>
<td>2</td>
</tr>
</tbody>
</table>

B. Evaluation Technique
The tool is analyzed on the basis of classification technique of data mining. Various classifiers of WEKA are used as basis of evaluation or analysis. The classifiers of WEKA used in the study are:
- J48
- Decision Table
- Naïve Bayes
- Random Tree
- Attribute Selected Classifier
- Rules ZeroR
- Ibk lazy
- Input Mapped Classifier
- SGD function

C. Experimental Setup
The experiments are conducted on a laptop with Windows 2007 with an Intel Core i3 M350 processor and 4 GB main memory. WEKA 3.8 is freely downloaded from internet as it is open source software.

The tests are conducted on two modes:
- 10-fold cross-validation.
- Evaluation on full training data.

D. Experimental Results
WEKA is an efficient data mining tool with a vast capacity of analysis. It has inbuilt capability of analyzing Big Data using following data mining techniques:
- Classification
- Clustering
- Association

In the present study only classification technique is used for analysis. Classification contains seven different types of classifiers:
- Bayes
- Functions
- Lazy
- Meta
- Misc
- Rules
- Trees

Each and every classifier is considered for analysis on the study. While implementing the classifiers on the real dataset two test modes are considered:
- 10-fold cross validation
- Full training set.

The classifiers are analysed on the basis of Precision. All the nine classifiers, mentioned above in evaluation technique, are implemented using the real dataset pima_diabetes. The precision of the classifiers is represented in chart 1:
From the execution values it can be concluded that SGD Function gives the maximum correct classified instances. RulesZeroR and Input Mapped Classifier give the least correct classified instances. In general all the algorithms produce the correct classified instances in the range 500 to 600 out of 768 instances.

### REFERENCES

9. A. Jović*, K. Bukić* and N. Bogunović, “An overview of free software tools for general data mining”, in the proceedings of 37th International Convention on Information and


