مجلة العلوم البحثة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.lyISSN 2521-9200
Received 13/01/2019 Revised 22/04/2019 Published online 01/05/2019

Data Mining: WEKA Software (an Overview)

*Fatma A. Ibrahim, Omar A. Shiba
Computer Department, Faculty of Science, Sebha University, Libya
*Corresponding Author: fat.ibrahim1@sebhau.edu.ly

Abstract Data mining (also known as knowledge discovery from databases) is the process of extraction of hidden, previously unknown and potentially useful information from database. In today's world data mining have progressively become interesting and popular in terms of all application. The data mining requires huge and small amount of data sets for extraction of knowledge from it. The main aim of data mining software is to allow user to examine data. This paper is a review paper that introduces some topics related to data mining steps and also describing the steps of how to use WEKA tool for various technologies & different facility to classify the data through various algorithms.
Keywords: Data mining, Data pre-processing, KDD, Machine learning, WEKA tool.

تنقيب البيانات: برمجيات WEKA (نظرة عامة)

*فاطمة إبراهيم و عمر شيبه
قسم الحاسوب– كلية العلوم– جامعة سبها، ليبيا
*للمراسلة: fat.ibrahim1@sebhau.edu.ly

الملخص تنقيب البيانات(يُعرف أيضاً باكتشاف المعرفة من قواعد البيانات) هو عملية استخراج المعلومات المخفية وغير المعروفة مسبقًا والتي قد تكون مفيدة من قاعدة البيانات، ويتطلب تنقيب البيانات كمية صغيرة وكبيرة من مجموعة البيانات لاستخراج المعرفة منها. الهدف الرئيسي من برمجيات تنقيب البيانات هو السماح للمستخدم بفحص البيانات. هذه الورقة عبارة عن استعراض لبعض المواضيع المتعلقة بخطوات تنقيب البيانات وأيضاً تصف خطوات كيفية استخدام أداة WEKA للعديد من التقنيات المختلفة لتصنيف البيانات من خلال خوارزميات مختلفة.
الكلمات المفتاحية: تنقيب البيانات، المعالجة المسبقة للبيانات، KDD، التعليم الآلي، أداة WEKA.

I. Introduction

Data mining is one of the areas of computer science, it is refers to extracting useful information from big amount of data[1], it has been defined as the extraction of implicit, previously unknown, and possible be useful information from databases/data warehouses. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form, which is easily perceptual to humans[2,3,4]. Data mining attempts to formulate analyse and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from unstructured data. It uses mathematical analysis to derive patterns and trends that exist in data. Popularly known as Knowledge Discovery in Databases (KDD) [3], and the purpose of the Knowledge Discovery Process is *"a decision support process in which we search for patterns of information in data"*. This is done by automatically searching one database or multiple databases to find important patterns in the data[5]. KDD consists of the following serial steps [6,7]:

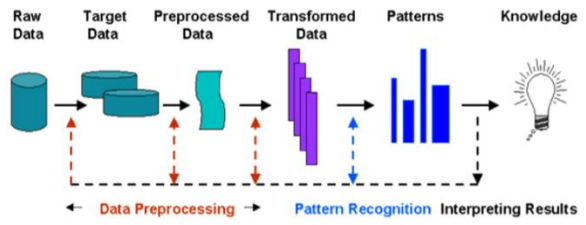

Figure 1: The Knowledge Discovery in a Database Process[1]

1. *Data Selection:* Compiles heterogeneous and unrelated data from different sources, which requires the process of selecting important and relevant data to become the target data to be processed and extract knowledge from them.
2. *Data Pre-processing:* Performs basic operations to remove the noise data, attempt to find missing data or to put a strategy to process it, detect/remove extreme values, and resolve discrepancies between data.
3. *Data Transformation:* The data is transformed and put into appropriate formats for mining by performing some methods as association and clustering.
4. *Data Mining:* This step involves application of knowledge discovery algorithms to the cleaned, transformed data in order to extract meaningful patterns from the data. Data mining algorithms include *classification, clustering, regression, etc.*
5. *Pattern Evaluation:* There are various types of information need different type of representation, in this step presentation of mined patterns in understandable and useful form where evaluation of the outcomes is prepared with statistical justification and significance testing.
6. *Knowledge Representation*: In this step the knowledge discovered is represented in the

correct form using different visualization techniques.

Moreover, data mining tools predict future trends and patterns, behaviours, help to make knowledge driven decisions[4,8], there are many other tools used in data mining, such as Rapid Miner, Orange, Tanagra, etc, which are not good enough compared with WEKA, for more details see [8,9,10].

II. Data Per-processing

In real world, data are aggregated from different sources to create a certain databases/data warehouses. Thus, data can be dirty, noisy incomplete or inconsistent, to processing such as these issues should be performing preprocessing before data mining steps are initiated to obtain good data quality, which lead to high-quality mining results. Preprocessing includes major tasks the following[11,12,13]:

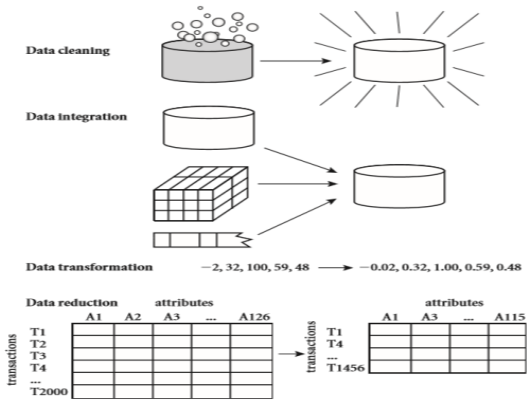

Figure 2: Forms of data pre-processing

1. **Data cleaning:** It fills missing data and remove noise data, and finds the outliers and correct inconsistencies in the data.
2. **Data integration:** It combines data relevant from different sources of data in a single data store, such as a data warehouse.
3. **Data transformation:** It converts a set of data values into a common format for processing, as a normalization application.
4. **Data reduction:** It reduces the volume of data, but still produces the same analytical results.

III. WEKA Tool

WEKA (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms (ID3, KNN, ...) for solving real - world data mining problems written in Java, and runs on almost any platform. It funded by the University of Waikato, New Zealand in 1993. It is free software and available from http://www.cs.waikato.ac.nz/ml/weka. The WEKA workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality[2,4,8].

Advantages of WEKA include:

1. Free availability under the GNU General Public License.
2. Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
3. A comprehensive collection of data preprocessing and modelling techniques.
4. Ease of use, because of graphical user interfaces.

WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection [2].

A. WEKA GUI Chooser

There are four graphical user interfaces in WEKA, as shown in figure 3.

Figure 3: WEKA GUI Chooser

1. **Explorer:** An environment for exploring data with WEKA, and gives access to all of its facilities using menu selection and form filling.
2. **Experimenter:** It is environment for performing experiments and statistical tests between learning schemes, and evaluating machine learning algorithms.
3. **Knowledge Flow:** The environment supports some functional from the Explorer, and but with a drag-and-drop interface. You can design configurations to handle streaming data.
4. **Workbench:** A unified graphical user interface that combines interfaces (Explorer, Experimenter, Knowledge Flow) into one application, allowing to user of specify which applications and additions will show, In addition to with settings relating to them.
5. **Simple CLI:** Command-line interface is simple interface for writing commands of WEKA for operating systems that do not provide their own command line interfaces.

B. Classification with WEKA

Explorer GUI contains taps for preprocess, classify, cluster, associate, select attributes, visualize, shown in figure 4. Classification is a process of classify each item in a dataset to any belonging to set of predefined labelled as classes or groups, by examining the features for item.

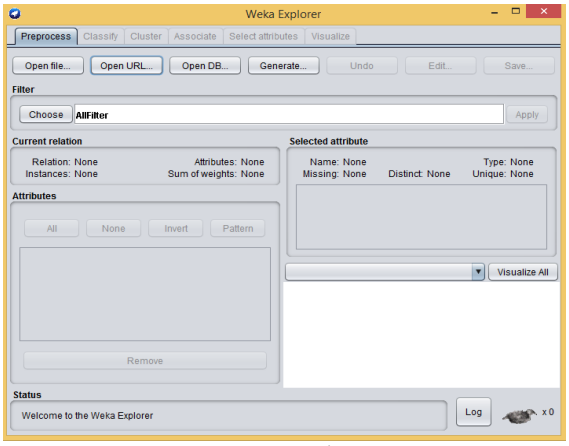

Figure 4: Explorer GUI

Classification in WEKA is executes a step by step process. First, is data loading from various sources, include files, URLs and databases by tap preprocess in Explorer GUI. WEKA has of capacity to read different formats (.txt, .csv, .arff), such as in ".csv" format from real world Excel datasheet, once data loaded into WEKA, the data set automatically saved into ARFF (Attribute Relation File Format) format. Then, follow these steps:

1. Preprocessing the dataset
2. Choose classify and apply algorithm
3. Generate model
4. Analysis the result or output

The classification process is selected by click on the tap "Classify" in Explorer GUI, as shows figure 5. Classify includes many learning algorithms such as Decision Tree, K-nearest nighbour, Naive Bayes, Support Vector Machine, as following:

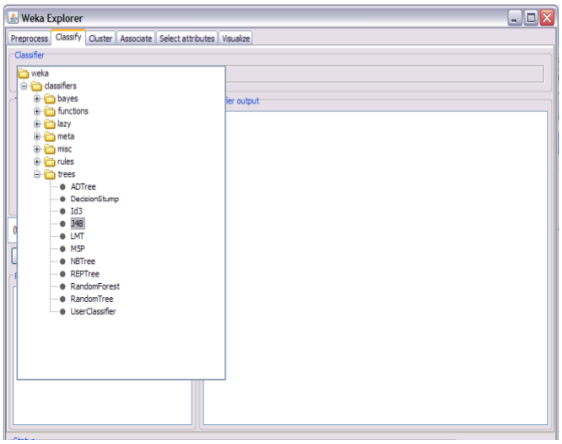

Figure 5: learning algorithms in Classify

Decision Tree: is most commonly used in WEKA algorithms. In this algorithm, it is used divide and conquer technique as basic learning strategy, it goals to create a model to predict the value of the target variable based on the input values. The structure of decision tree includes a root node, branches, and leaf nodes. Each internal node indicates a test on an attribute, each branch indicates the outcome of a test, and each leaf node keeps a class label. There are various decision tree algorithms are used in classification such as ID3,AD Tree, J48. Figures 6,7 show an example for representation of classifier tree J48 (Weather) output, and tree view visualization.

```
J48 pruned tree
------------------

outlook = sunny
|   humidity = high: no (3.0)
|   humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :  5

Size of the tree :  8
```

Figure 6: Output after building and testing the classifier: decision tree

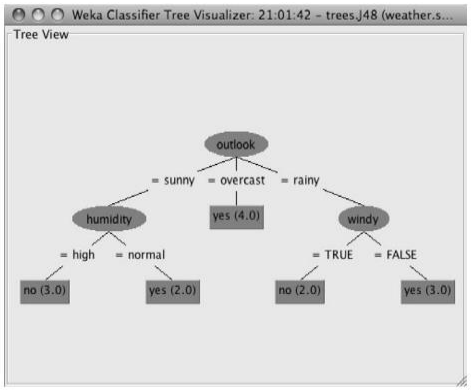

Figure 7: The decision tree building

K-Nearest Neighbor (K-NN): is a pattern based learning or lazy learning, it is compare every attribute of every case in the set of similar cases with every identical attribute of the input case, and classifies new cases based on a similarity measure. K-NN algorithm is consists of steps following:

1. Calculating the distance of input case x_j from all training cases x_k by equation for Euclidean distance:

$$dis(x_j,x_k) = \sum_i \sqrt{(x_{j,i} - x_{k,i})^2}$$

2. Order training cases based on the distance and selection of K-nearest neighbor.
3. Using the class which owns the majority between the k-nearest neighbors (this method considers the class as the class of input case which is observed more than all the other classes between the K-nearest neighbors).

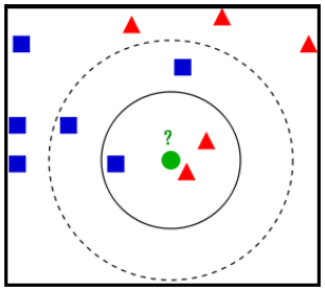

Figure 8: Example of k- nearest neighbor

In figure 8, the test case (circle) should be classified either into square class or into triangle class. If k = 3 (as solid line circle) test case (circle) is classified into triangle class because there are 2 triangle cases and only 1 square case in the inner circle. If k = 5 (as dashed line circle) test case (circle) is classified into square class because there are 3 square cases and only 2 triangle cases in the inner circle.

Naive Bayes: this is a classifier based on Bayes' theorem, which is used to predict class labels by the estimates of the probability masses are used as input for a Naive Bayes classifier. It is computes the conditional probabilities of the different classes given the values of attributes, and then selects the class with the highest conditional probability. Equation of Bayes' theorem is:

$$P(C_x|X) = \frac{P(X|C_x)P(C_x)}{P(X)}$$

$$P(C_x|X) = P(C_x|X_1) \times P(C_x|X_2) \times \ldots \times P(C_x|X_n) \times P(C_x)$$

Where:

P(Cx|X) is a posterior probability of class (C_x,target) given predictor (x,attributes).

P(C) is called class previous probability.

P(X|Cx) is the probability which indicate the predictor probability of given class.

P(X) is called previous probability of the predictor.

Example explain implementation of Naive Bayes algorithms: Weather dataset having attributes (Sunny, Overcast, and Rainy). Using training dataset, Naive Bayes algorithms will predict the value as you can play or no.

Table 1: Training dataset of Weather

Weather	Play
Sunny	Yes
Overcast	Yes
Overcast	Yes
Sunny	No
Rainy	No
Sunny	No
Overcast	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Sunny	Yes
Rainy	Yes
Overcast	Yes
Rainy	Yes
Rainy	No
Sunny	Yes
Overcast	Yes

Table 2: Frequency data of Weather

Weather	Yes	No
Sunny	4	2
Overcast	6	0
Rainy	2	3
Total	12	5

Now, we can compute the posterior probability for each class by Naive Bayesian equation. the result of prediction is highest posterior probability for the class.

P(Yes|Sunny) = P(Sunny|Yes) * P(Yes) / P(Sunny)

P(Sunny|Yes) = 4/12 = 0.33, P(Sunny) = 6/17 = 0.35, P (Yes) = 12/17 = 0.70

Now, P(Yes|Sunny)= 0.33*0.70/0.35= 0.66, which has higher probability.

Support Vector Machine (SVM): it is one of the method for data classification by a separating hyper-plane, which can separate the two classes or labels clearly with a maximal separating margin. The margin is the geometrical distance of blank space between the two species. SVM is divides sample points of both labels or class are on different sides of hyper plane. The hyper plane is calculated as:

$$w.x - b = 0$$

Where, w is weight vector, and b is bias of the optimal hyperplane. Figure 9 shows the SVM classifier with hyper plane. Such as training examples labelled either "yes" or "no", a maximum-margin hyper plane is identified which divide the "yes" from the "no" training examples, such as the distance between the hyper plane and the closest examples (the margin) is maximized.

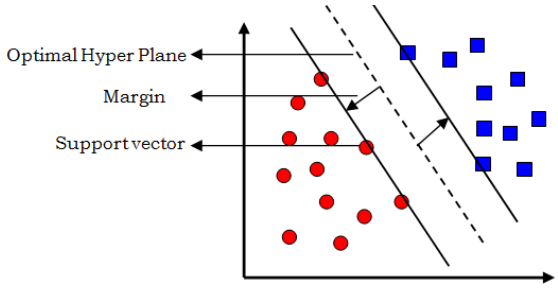

Figure 9: SVM classifier

IV. Conclusion

In this paper we have discussed the concepts and steps of KDD, data mining and WEKA. We have also covered some common classification algorithms used in WEKA tool for data classification such as Decision Tree, K-nearest nighbour, Naive Bayes, Support Vector Machine. We conclude that WEKA software supports many standard data mining tasks, specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

REFERECE

[1]- Y.Li, F.Advisor, T.Beaubouef, *"Data Mining: Concepts, Background and Methods of Integrating Uncertainty in Data Mining"*, CCSC:SC Student E-journal, Vol.3, PP.2-7,2010.

[2]- S.B.Jagtap, Kodge B.G, *"Census Data Mining and Data Analysis using WEKA"*, International Conference in Emerging Trends in Science, Technology and Management-Singapore,2013.

[3]- V.Gupta, Devanand, *"A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues"*, International Journal of Scientific and Engineering Research Volume 4, Issue3, ISSN 2229-5518, March-2013.

[4]- E.G.Kulkarni, Raj B.Kulkarni, *"WEKA Powerful Tool in Data Mining"*, International Journal of Computer Applications, 0975-8887, RTDM 2016.

[5]- O.Sjöblom, "*Data Mining in Promoting Flight Safety*", page no: 85, Ph.D. thesis, 2016.

[6]- S.Mukherjee, R.Shaw, N.Haldar, S.Changdar, *"A Survey of Data Mining Applications and Techniques"*, International Journal of Computer Science and Information Technologies, Vol. 6 (5), ISSN: 0975-9646, 2015.

[7]- S.Patel, H.Patel, *"Survey of Data Mining Techniques Used in Healthcare Domain"*, International Journal of Information Sciences and Techniques, Vol.6, No.1/2, March 2016.

[8]- S.Usharani, K.Kungumaraj, *"A Survey on Data Mining with Big data- Applications, Techniques, Tools, Challenges and Visualization"*, International Journal of Advanced Research in Computer and Communication Engineering, Vol.4, Issue 12, ISSN 2278-102, December 2015.

[9]- V. Sridevi, A. Kanagaraj, *"A Survey of Data Mining Techniques and Tools"*, International Journal of Advanced Research in Computer and Communication Engineering, Vol.6, Issue 9,ISSN 2278-1021, September 2017.

[10]- S.Hussain, *"Survey on Current Trends and Techniques of Data Mining Research"*, London Journal of Research in Computer Science and Technology, Vol.17, Issue 1, 2017.

[11]- K.Kushwaha, P.Mishra, *"A Survey on Data Mining using Machine Learning Techniques"*, International Journal of Advanced Research in Computer and Communication Engineering, Vol.5, Issue 9, ISSN: 2278-1021, September 2016.

[12]- J.Han, M.Kamber, J.Pei,"*Data Mining Concepts and Techniques*", 3rd,Elsevier, ISBN: 978-0-12-381479-1, 2011, page no: 83,84.

[13]- S.Singhal, M.Jena, *"A Study on WEKA Tool for Data Preprocessing, Classification and Clustering"*, International Journal of Innovative Technology and Exploring Engineering, Vol.2, Issue-6, ISSN 2278-3075, May 2013.

[14]- Omar A. A. Shiba, *"Comparing Case Slicing Technique With Selected Decision Tree Classification Algorithms"*, KUTPM Journal, Vol.2, No.2, p.16, 2004.

[15]- Omar A. A. Shiba, *"CST New Slicing Techniquesto Improve Classification Accuracy"*, In the Proceedings of The International Arab Conference On Information Technology, Benghazi-Libya (ICT2010), 14-16 Dec 2010.

[16]- D.Baumgartner, G.Serpen, *"Large Experiment and Evaluation Tool for WEKA Classifiers"*, International Conference on Data Mining, July 13-16, 2009, Las Vegas, USA.

[17]- H.Sahu, S.Shrma, S.Gondhalakar, *"A Brief Overview on Data Mining Survey"*, International Journal of Computer Technology and Electronics Engineering, Vol.1, Issue 3, ISSN 2249-6343, 2011.

[18]- R.Deepa, S.Vaishnavi, *"A Survey on Data Mining Methods and its Applications"*, International Journal of Advance Engineering and Research Development, Vol.5, Issue 01, ISSN 2348-4470, January 2018.

[19]- M.Kuhkan, *"A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm"*, International Journal of Computer Engineering and Information Technology, Vol.8, No.6, ISSN 2412-8856, June 2016.

[20]- K.Khamar, *"Short Text Classification Using KNN Based on Distance Function"*, International Journal of Advanced Research in Computer and Communication Engineering, Vol.2, Issue 4, ISSN: 2278-1021, April 2013.

[21]- O.Georgina, A.Isah, J.Alhasan, *"Analytical Study of Some Selected Classification Algorithms in WEKA Using Real Crime Data"*, International Journal of Advanced Research in Artificial Intelligence, Vol.4, No.12, 2015.

[22]- A.Tate, B.Rajpurohit, J.Pawar, U.Gavhane, G.B.Deshmukh, *"Comparative Analysis of Classification Algorithms Used for Disease Prediction in Data Mining"*, International Journal of Engineering and Techniques, Vol.2, Issue 6, Nov-Dec 2016.

[23]- R.P.Aharwal, *"Evaluation of Various Classification Techniques of WEKA Using Different Datasets"*, International Journal Of Advance Research And Innovative Ideas In Education, Vol-2 Issue-2, ISSN: 2395-4396, 2016.

[24]- D.Kaur,R.Bedi,S.K.Gupta, *"Review of Decision Tree Data Mining Algorithms: ID3 and C4.5"*, Proceedings of International Conference on Information Technology and Computer Science, ISBN:9788193137307, July 11-12, 2015.

[25]- S.Aruna**,** L.V.Nandakishore, *"Application of Gist SVM in Cancer Detection"***,** Annals. Computer Science Series. 9th Tome 2nd Fasc. – 2011.

[26]- Yukai Yao et al,"*K-SVM: An Effective SVM Algorithm Based on K-means Clustering*", Journal of Computers, Vol.8, No.10, October 2013.