ISSN 2229-3531 (Print)

JUCIT Vol. 7(3), 31-33 (2016). Periodicity-2-Monthly

(Online ISSN 2455-9997

JOURNAL OF ULTRA COMPUTER & INFORMATION TECHNOLOGY
An International Open Free Access Peer Reviewed Research Journal of Computer Science Engineering & Information Technology
website:- www.compitjournal.org
Estd. 2010

Comparative Study of Open Source Data Mining Software for Big Data

ATIF AZIZ[1] RAJEEV ARYA[2] and SANA SHAFIQUE[3]

[1]Research Scholar NIMS University Jaipur Rajasthan India
[2]Director, Truba Institute of Engineering and Information Technology
[3]Research Scholar Bhagwant University, Ajmer Rajasthan India
Email of Corresponding Author:- atifaziz82@rediffmail.com
http://dx.doi.org/10.22147/jucit/070301

Acceptance Date 8th November, 2016, Online Publication Date 2nd Dec. 2016

Abstract

The immergence of big data has shaped innumerable possibilities in data analysis. The study focuses on the three open source data mining software namely:- Weka, Rapid Miner and Tanagra. The market is flooded with open source software but the above mentioned software are chosen as these are widely used within the data mining community. The paper compares the software on the basis of features that are available in the software.

Key words: Big Data, Data mining software , Weka, RapidMiner, Tanagra

Introduction

In the recent years with the advancement and use of technology in every walk of life, the data is being stored everywhere whether it may be hospitals, super markets, banking sector, service industries or production industries, all these create huge amount of data.In 2007 more data was created than the storage capacity could handle[1]. 2.5 quintillion bytes of data are generated everyday[3]. Big data can be best described using volume, variety and velocity[4]. The different aspects indicates the diversity of contemporary data. The data is created in a very short time frame, this can be described by high velocity of big data[2] and extraction of useful information is of prime importance.

Industries have huge amount data most of it in the raw form or in semi-structured or unstructured format but they don't know how to identify meaningful patterns. Analysis task should be able to present opportunity to produce valuable insights for business, government, science and everyday life.

Machine learning emphasis on the automated learning of machines and ability to handle larger data sets. There are so many machine learning approaches like decision tree, neural networks, genetic algorithm etc. On the basis of training machine learning can be classified into supervised learning, unsupervised learning and semi supervised learning. The quality and quantity of training is the main drawback of supervised learning algorithms. When the training data is biased or insufficient the supervised learning algorithm may fail[5]. The lack of prior knowledge leads to incoherent features which may not correlate with human judgement, despite of all this unsupervised learning still gives us knowledge about data[5]. The commonly used SSL algorithms include self-training, generative models, co-training, graph based methods and multi view learning. SSL can overcome the drawbacks of unsupervised learning by associating some prior knowledge to the unsupervised models[5]. To find out the valuable information from the data which is available various

open source data mining software are available. For this reason there is a need for comparing and contrasting the capabilities of the software.

Criteria for Comparing data Mining software :

The various features which have been used for comparing and contrasting of software are:- User groups, Data structures, Tasks And Methods, Interaction and Visualization, Import and export of data and models, Platforms, Licenses, Functionality Aspect

SNo.	General Characteristics	Weka	Rapid Miner	Tanagra
1	Activity:	high	high	high
2	License :	GPL	GPL	other
3	Language:	Java	Java	C++
	GUI/command Line	Both	GUI	Both
	Data Source Characteristics			
4	JBDC	Yes	Yes	No
5	ARFF	Yes	Yes	Yes
6	CSV	Yes	Yes	No
7	Excel	No	yes	Yes
8	Data Size	Medium	Medium	Medium
	Functionality			
9	Association	Yes	Yes	Yes
10	Evaluation	Yes	Yes	Yes
11	Data Visualization	High	High	Low
12	Model Visualization	High	High	Low
	Usability Aspect			
13	Human Interaction	Manual	Manual	Manual
14	Interoperability	self	self	self
15	Extensibility	Excellent	Excellent	simple
	Documentation			
16	Learning Document	Poor	Poor	Poor
17	Use	limited	limited	Limited
18	Support Services	Low	Medium	Low
	Data Import			
19	Specific input format files	Arff,.libsvm)	arff, .xrff	CSV
20	Excel/spreadsheet	Yes	No	Yes
	Association rules			
21	Apriori	Yes	Yes	Yes
22	FPGrowth	Yes	Yes	No
23	Eclat	Yes	No	No

Conclusion

The study presented an overview of the recent study on three open source data mining software namely: Weka, Tanagra and RapidMiner. Although many different open source software are developed and tested in this field but the above software are chosen as these are widely used by academicians, students and researchers. The above table indicates that Weka and Tanagra are more friendly in importing data from excel and spread sheets where most of the data is stored nowadays as compared to RapidMiner. Model Visualization characteristics are good in Weka and RapidMineras Compared to Tanagra. Overall Weka has got superior features as compared to RapidMiner and Tanagra.

References

1. Jamiy, F. E., Daif, A., Azouazi, M. & Marzak, A., The potential and challenges of Big data - Recommendation systems next level application. *International Journal of Computer Science Issues,* 11(5) (2014).
2. Anon., *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. s.l.:EMC Education Services (2015).
3. Anon., n.d. [Online]
Available at: http:/ //www-01.ibm.com/software/data/bigdata [date visited : 10/11/2016]
4. Dobbs, R., Manyika, J., Roxburgh, C. & Lund, S., *Big data: The next frontier for innovation, competition, and productivity,* s.l.: McKinsey Global Institute (2011).
5. Madhoushi, Z., Hamdan, A. R. & Zainudin, S., *Sentiment Analysis Techniques in Recent Works.* London, UK, s.n. (2015).