IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 4, Ver. III (Jul.-Aug. 2017), PP 56-61
www.iosrjournals.org

Comparative Analysis of Data Mining Techniques, Tools and Machine Learning Algorithms for Efficient Data Analytics

[*]Keshav Singh Rawat
Central University of Himachal Pradesh, Dharamsala
Corresponding Author: Keshav Singh Rawat

Abstract: *Data mining is more demanding now due to large amount of data are created by web, social media like- twitter, face book, web, and other sources. Data mining is Extracting knowledge from raw data available on large data sets on computer. Extracting knowledge from large data set requires decision making algorithms, machine learning is a process to classification in data mining the data. Many free and open source data mining tools are available on World Wide Web and they are performing classification process using various techniques. This paper analyse various free and open source data mining tools like Weka, R, etc. Our aim to find most accurate tool and technique of classification process. Comparative analysis indicates that we can achieve best result using various combinations of tools and classification technique*.

Keywords: *Data mining, Data classification, free and open source, Machine learning.*

Date of Submission: 14-07-2017 Date of acceptance: 24-07-2017

I. Introduction

In the present computerized world, we are encompassed with huge information that is rapidly growing each day. We have data in large amount but unavailable to get knowledge from that data. The knowledge discovery is very important to achieve benefits from such data. Knowledge discovery process is described in figure 1.

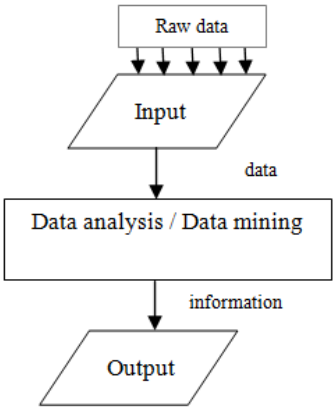

Figure 1: Knowledge discovery

The field of data mining is an emerging research area with important applications in Engineering, Science, Medicine, Business and Education. The size of data base in educational application is large where the number of records in a data set can vary from some thousand to thousand of millions. The size of data is accumulated from different fields exponentially increasing. Data mining has been used different methods at the intersection of Machine Learning, Artificial Intelligence, Statistics and Database Systems [1]. The overall aim of the data mining process is to extract information from huge datasets and transform it into understandable structure for further use.

Data is increasing rapidly day to day due to rapid development of information technology and its usage by the public. Useful information can be obtained when these raw data's are studied properly. Data mining is a process that turns raw data into useful information. The unprocessed data's need to be collected, stored and maintained properly before they are applied to any of the mining techniques. These mining techniques automatically searches large store of data and discover patterns. Data mining depends on effective data collection and warehousing as well as computer processing. The data mining process is described in figure 2. Some of the application of data mining are (i) Spatial mining (ii) Multimedia data mining (iii) Text Mining (iv) Web data mining [24].

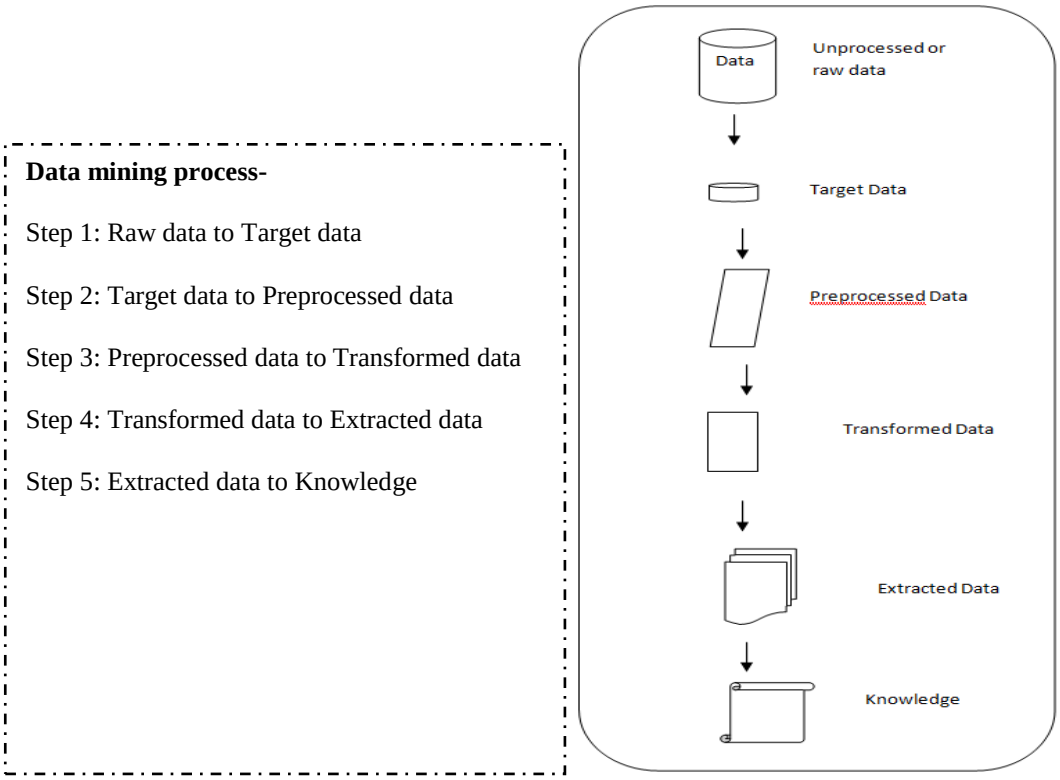

Figure 2: Data mining steps

The remaining paper is prepared as follows: Section 2 describes the data mining techniques. Section 3 introduced machine learning process. Section 4 describes some popular open source data mining tools. Section 5 contains the comparative analysis of different data mining tools. Section 6 summarises the Conclusion and the tail of this paper contains references.

II. Data Mining Techniques

Data mining is an Extraction of use full, analytical data from large databases. It plays out an Identification and assessment of unknown patterns in database. It is capable innovation with awesome potential to help associations to find and create data from their data sets. he various types of data for mining are (i) Relational Databases (ii) Flat files (iii) Transaction Databases (iv) Multimedia Databases (v) Spatial Data bases (vi) Time-Series Databases (vii) Data Warehouses(viii) World Wide Web. Data mining instruments anticipate future patterns and practices [2]. Information mining can be led on any sort of information as long as the information are significant for an objective application. There are four basic approaches of data mining.

Classification - It is a data mining technique that assigns data in a collection to target classes. It is utilized to characterize diverse data in various classes. Classification is like bunching in a way that it likewise fragments data records into various portions called classes. It also require algorithms for new data classification. There are many common algorithms to do classification (i) Decision trees, (ii) Artificial Neural Networks, (iii) Rules Induction, (iv) Bayesian classifiers, (v) Clustering, (vi) SVM etc.

Regression- It is the technique to identify and analyses the relation between two or more independent and dependent variables. Independent and response variables are used in order to analyze the data set.

Clustering- It is defined as unsupervised classification of data items in clustering given data set is divided into two groups call clusters so that similar data come together in one cluster. Clustering is one of the important method of data mining as we have large amount of data set available on well as well as data repository basically clustering is used to divide the data in two similar objects that is cluster on group and the objects or data that are dissimilar are located in different cluster or group. The various clustering methods are used to find useful and different classes of data pattern that are classified according to (i) input type (ii) criteria of clustering which define the similarity between the data objects (iii) concepts like numerical data, fuzzy system etc. Clustering is widely used as one of the important steps in the exploratory data analysis [25]. Some clustering algorithms are –

Partition Based- K means, K-Medoils, K-modes, PAM, CLARANS, CLARA, FCM.

Hierarchical Based – CURE, BIRCH, ROCK,Echidna, Chameleon.

Grid Based- STING, CLIQUE, Wave Cluster, Optigrid.

Density Based-DBSCAN, OPTICS, DBCLASD, DENCLUE.

Model Based-COBWEB, EM, CLASSIT,SOMs.

Association Rule Learning- It extract interesting correlations, frequent patterns, associations among sets of items in the transaction databases and data repositories. Association rules are applied in various areas such as communication networks, risk management, inventory control, etc. Some recognize algorithm of association rules are apriori, Eclat and FP growth[24].

(i) Apriori algorithm is the best method to mine association rules. It use BFS(breadth first search) strategy based on divide and conquer method.

(ii) Eclat algorithm is a DFS (depth first search) algorithm using set intersection method.

(iii) FP growth algorithm (frequent pattern growth) uses an extended FP tree structure to store the database in a compressed form this algorithm also adopt a divide and conquer method to decompose both the mining task and the databases.

Anomaly or Outlier Detection- There are various issue exist in mining information in substantial dataset such information repetition, the estimations of traits is not particular, information is not finished and anomaly or Outlier. Anomaly detection is also known as outlier detection. It is used to find date items, which do not match with items in the data set, it has three categories unsupervised, semi-supervised and supervisor anomaly detection. This technique can be used in a variety of areas like- (i) fraud detection (ii) system health monitoring, (iii) intrusion detection (iv) fault detection, (v) eco-system disturbances (vi) event detection used in sensor networks. Some efficient data analytics algorithm [3] for data mining are given in table 1.

Table1: List of Data Mining Algorithms

Data mining technique	Algorithm	Reference
Classification	TLAESA	[4]
	SLIQ	[5]
	FastNN	[6]
	SFFS	[7]
	GPU- based SVM	[8]
Clustering	DBSCAN	[9]
	RKM	[10]
	TKM	[11]
	BIRCH	[12]
	Incremental DBSCAN	[13]
Association rules	FP-tree	[14]
	CLOSET	[15]
	CHARM	[16]
	MAFIA	[17]
	FAST	[18]
Sequential patterns	SPADE	[19]
	CloSpan	[20]
	Prefix Span	[21]
	SPAM	[22]
	ISE	[23]

III. Machine Learning

Machine learning is a process of designing algorithms that can help to make production on data. machine learning can be supervised and unsupervised, machine learning process includes data initialisation, learning model and testing models. the machine learning process is described in figure 3. In the data initialization, collected data is pre-processed to removal of noise then feature can be selected for better output. Learning model contain algorithms and tested data where testing model test data set for efficiency and provide expected output[28].

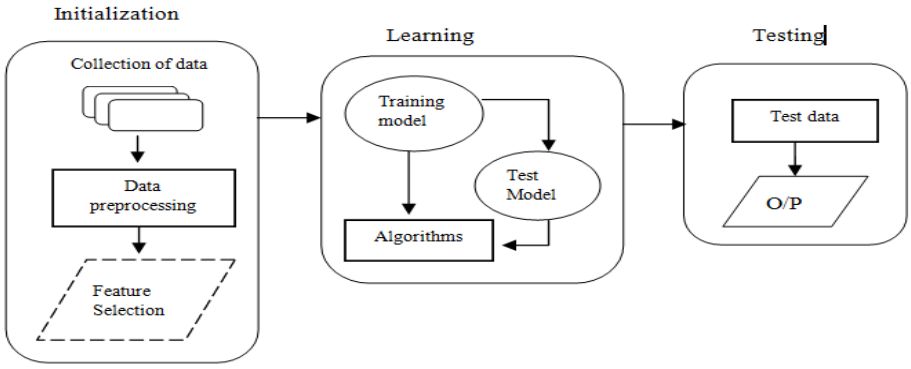

Figure 3: Machine learning process

Machine learning is useful to improve various applications like – (i) image classification (ii) Face recognition (iii) Speech recognition (iv) Signal denoising (v) Genetics (vi) Anti spam (vii) Weather forecasting.

IV. Data Mining Tools

The development of data mining algorithms needs data mining. There are many tools available either free and open source or commercial tools. The criteria used to classify data mining to depend on the different determining task methods visualizations, data structure interaction, import export option, and license policies [24]. The most frequently used data mining tools are given in table 2.

Table2: List of some Open Source and Commercial Tools

Open source	*Commercial*
WEKA	SPSS
ORANGE	IBM Intelligent Miner
Rapid Miner	Microsoft SQL Server
DataMelt	Oracle Data Mining (from Oracle 10g)
Apache Mahout	Angoss Knowledge STUDIO
ELKI	SAS Enterprise Miner
Knime	KXEN
MOA	
KEEL	
Rattle or R	

V. Comparative Analysis Of Data Mining Tools

In this section, The best available data mining tools [2][24][26][27] were taken and comparative study was made by considering parameter licence, technical specification and features and describing in table 3.

Table 3: Features of most common used data mining tools.

Software tool	Type	Features
WEKA	Machine Learning based	It is java based open source data mining tool, based on various data mining and machine learning algorithms.
ORANGE	Machine Learning, Data mining, data visualization	It is open source data mining tool, based on Visual programming, visualization, large tool box, plate form independent, interaction and data analysis.
Rapid Miner	Statistical analysis, predictive analysis	It is an important predictive analytic platform. It is user friendly, rich library of data science and has many machine learning algorithms.
DataMelt	Statistical analysis , numeric and symbolic computations, scientific visualisation	It is based on cluster analysis, linear regression, neural networks, curve fitting, fuzzy system, analytic calculations & interactive visualisations.
Apache Mahout	Machine Learning based	It is a library of machine learning algorithms which is used in clustering, classification & frequent pattern data mining. It is also used in a distributed mode with integration of Hadoop.
ELKI	cluster analysis and outlier detection	It is java based open source data mining tool and licensed under AGPLv3.It focuses on outlier detection and cluster analysis with a compilation of several algorithms from both these domains.
KNIME	Enterprise reporting, Business Intelligence	It is based on java and built upon Eclipse. It is scalable, highly extensive, and capable of data visualization and import –export workflow.
MOA	Machine Learning based	It can handle large volumes of real time data streams at a very high speed and can be used through GUI, command line, or Java API.
KEEL	Machine Learning based	It is java based tool and licensed under GPLV3and based on clustering, classification, and association. It has a very user friendly GUI.
Rattle or R	Statistical computing	It is a statistical programming language and R program c an run on Mac OS Linux, and Windows, it is based on statistics, clustering, modelling and visualisation.

VI. Conclusion And Future Scope

In this paper, we gives a brief introduction to data mining and various data mining techniques like classification, clustering, association rule, regression, anomaly detection and their features. A comparative analysis of various data mining tools is also discussed in this paper. This paper shows various data mining technique can be applied for given domain and which data mining tool will help to analyze data by applying efficient algorithms. The best technique and tool of data mining can be determined by comparing the classification method and total accuracy. The discussion may be used to develop a new or modified mining algorithm or software tool that achieved best results.

References

[1] Sumit Garg,Arvind K. Sharma , "Comparative Analysis of Data Mining Techniques on Educa-tional Dataset" ,International Journal of Computer Applications (0975 – 8887),Volume 74– No.5, July 2013

[2] Rangra et al., International Journal of Advanced Research in Computer Science and Software Engineering 4(6), June - 2014, pp. 216-223

[3] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, "Big data analytics: a survey", Journal of Big Data , springer open ,2015

[4] Micó L, Oncina J, Carrasco RC. A fast branch and bound nearest neighbour classifier in metric spaces. Pattern Recogn Lett. 1996;17(7):731–9.

[5] Mehta M, Agrawal R, Rissanen J. SLIQ: a fast scalable classifier for data mining. In: Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology. 1996. pp 18–32.

[6] Djouadi A, Bouktache E. A fast algorithm for the nearest-neighbor classifier. IEEE Trans Pattern Anal Mach Intel. 1997;19(3):277–82.

[7] Ververidis D, Kotropoulos C. Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition. Signal Process. 2008;88(12):2956–70.

[8] Catanzaro B, Sundaram N, Keutzer K. Fast support vector machine training and classification on graphics processors. In: Proceedings of the International Conference on Machine Learning, 2008. pp 104–111.

[9] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996. pp 226–231.

[10] Ordonez C, Omiecinski E. Efficient disk-based k-means clustering for relational databases. IEEE Trans Knowl Data Eng. 2004;16(8):909–21.

[11] Elkan C. Using the triangle inequality to accelerate k-means. In: Proceedings of the International Conference on Machine Learning, 2003, pp 147–153.

[12] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1996. pp 103–114.

[13] Ester M, Kriegel HP, Sander J, Wimmer M, Xu X. Incremental clustering for mining in a data warehousing environment. In: Proceedings of the International Conference on Very Large Data Bases, 1998. pp 323–333.

[14] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In : Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000. pp. 1–12.

[15] Pei J, Han J, Mao R. CLOSET: an efficient algorithm for mining frequent closed itemsets. In: Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2000. pp 21–30.

[16] Zaki MJ, Hsiao C-J. Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Trans Knowl Data Eng. 2005;17(4):462–78.

[17] Burdick D, Calimlim M, Gehrke J. MAFIA: a maximal frequent itemset algorithm for transactional databases. In: Proceedings of the International Conference on Data Engineering, 2001. pp 443–452.

[18] Chen B, Haas P, Scheuermann P. A new two-phase sampling based algorithm for discovering association rules. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002. pp 462–468.

[19] Zaki MJ. SPADE: an efficient algorithm for mining frequent sequences. Mach Learn. 2001;42(1–2):31–60.

[20] Yan X, Han J, Afshar R. CloSpan: mining closed sequential patterns in large datasets. In: Proceedings of the SIAM International Conference on Data Mining, 2003. pp 166–177.

[21] Pei J, Han J, Asl MB, Pinto H, Chen Q, Dayal U, Hsu MC. PrefixSpan mining sequential patterns efficiently by prefix projected pattern growth. In: Proceedings of the International Conference on Data Engineering, 2001. pp 215–226.

[22] Ayres J, Flannick J, Gehrke J, Yiu T. Sequential PAttern Mining using a bitmap representation. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002. pp 429–435.

[23] Masseglia F, Poncelet P, Teisseire M. Incremental mining of sequential patterns in large databases. Data Knowl Eng. 2003;46(1):97–121.

[24] Kanchan A. Khedika , Mr. L.M.R.J.Lobo,"Data Mining: You've missed it If Not Used ", IOSR Journal of Computer Engineering (IOSR-JCE) ISSN: 2278-0661, ISBN: 2278-8727, PP: 06-14

[25] Garima, H. Gulati and P. K. Singh, "Clustering techniques in data mining: A comparison," *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2015, pp. 410-415.

[26] Mihai ANDRONIE , Daniel CRIȘAN, "Commercially Available Data Mining Tools used in the Economic Environment ", Database Systems Journal vol. I, no. 2/2010 45
http://opensourceforu.com/2017/03/top-10-open-source-data-mining-tools/

[27] J. Vasuki, S. Priyadarshini, " A study of basics of data mining, Machine Learning and Big data", IJIRCCE, Vol. 5, Issue 1, January 2017.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Keshav Singh Rawat. "Comparative Analysis of Data Mining Techniques, Tools and Machine Learning Algorithms for Efficient Data Analytics ." IOSR Journal of Computer Engineering (IOSR-JCE) 19.4 (2017): 56-61.