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.

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.

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].
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. The 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-Medoids, 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

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

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].

Initialization

Collection of data

Data pre-processing

Feature Selection

Learning

Training model

Test Model

Algorithms

Testing

Test data

O/P

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.

Table 2: List of some Open Source and Commercial Tools

<table>
<thead>
<tr>
<th>Open source</th>
<th>Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEKA</td>
<td>SPSS</td>
</tr>
<tr>
<td>ORANGE</td>
<td>IBM Intelligent Miner</td>
</tr>
<tr>
<td>Rapid Miner</td>
<td>Microsoft SQL Server</td>
</tr>
<tr>
<td>DataMelt</td>
<td>Oracle Data Mining (from Oracle 10g)</td>
</tr>
<tr>
<td>Apache Mahout</td>
<td>Angoss Knowledge STUDIO</td>
</tr>
<tr>
<td>ELKI</td>
<td>SAS Enterprise Miner</td>
</tr>
<tr>
<td>Knime</td>
<td>KXEN</td>
</tr>
<tr>
<td>MOA</td>
<td>KEEL</td>
</tr>
<tr>
<td>Rattle or R</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Software tool</th>
<th>Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEKA</td>
<td>Machine Learning based</td>
<td>It is java based open source data mining tool, based on various data mining and machine learning algorithms.</td>
</tr>
<tr>
<td>ORANGE</td>
<td>Machine Learning, Data mining, data visualization</td>
<td>It is open source data mining tool, based on Visual programming, visualization, large tool box, plate form independent, interaction and data analysis.</td>
</tr>
<tr>
<td>Rapid Miner</td>
<td>Statistical analysis, predictive analysis</td>
<td>It is an important predictive analytic platform. It is user friendly, rich library of data science and has many machine learning algorithms.</td>
</tr>
<tr>
<td>DataMelt</td>
<td>Statistical analysis , numeric and symbolic computations, scientific visualisation</td>
<td>It is based on cluster analysis, linear regression, neural networks, curve fitting, fuzzy system, analytic calculations &amp; interactive visualisations.</td>
</tr>
<tr>
<td>Apache Mahout</td>
<td>Machine Learning based</td>
<td>It is a library of machine learning algorithms which is used in clustering, classification &amp; frequent pattern data mining. It is also used in a distributed mode with integration of Hadoop.</td>
</tr>
<tr>
<td>ELKI</td>
<td>cluster analysis and outlier detection</td>
<td>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.</td>
</tr>
<tr>
<td>KNIME</td>
<td>Enterprise reporting, Business Intelligence</td>
<td>It is based on java and built upon Eclipse. It is scalable, highly extensive, and capable of data visualization and import –export workflow.</td>
</tr>
<tr>
<td>MOA</td>
<td>Machine Learning based</td>
<td>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.</td>
</tr>
<tr>
<td>KEEL</td>
<td>Machine Learning based</td>
<td>It is java based tool and licensed under GPLv3 and based on clustering, classification, and association. It has a very user friendly GUI.</td>
</tr>
<tr>
<td>Rattle or R</td>
<td>Statistical computing</td>
<td>It is a statistical programming language and R program can run on Mac OS Linux, and Windows, it is based on statistics, clustering, modelling and visualisation.</td>
</tr>
</tbody>
</table>

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


DOI: 10.9790/0661-1904035661 www.iosrjournals.org