IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Comparative Analysis of Data Mining With Big Data Using WEKA Software Tool

Srinivasa Rao Putta

Dept. of Computer Science, Rayalaseema University, Kurnool, India

Corresponding Author: srinivasa.putta@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i6.713715 | Available online at: www.ijcseonline.org

Accepted: 10/Jun/2019, Published: 30/Jun/2019

***Abstract*—** Big data has become more popular as people and organizations realize the importance and the value that the data has in formulating important information. As the data continue to increase, some challenges arise on the methods or techniques that are needed to be used in extracting meaningful information from the big data. Increase in data has led the researchers to make expansions on the existing data mining techniques to help with adapting to the evolving nature of big data thus leading to the development of new analytical techniques. Research has led to the development of various data mining techniques used on big data. It is, therefore, necessary to evaluate and compare different data mining techniques for big data.

***Keywords*—**Data Mining, Big Data, WEKA Software Tool

I. Introduction

Data mining refers to the process of analyzing various datasets that are provided to discover new patterns of unsuspected relationships. Different techniques are used in data mining where they are used in processing and analysis of various types of data patterns with the most popular tasks being classification, clustering, association rules mining, and summarization. Big data is a term that describes the collection of large data sets which can't be processed through the use of a traditional database system and its techniques [3]. It thus requires new techniques and approaches to store, process, analyze and visualize the data in a meaningful timeframe to find consistent patterns. Some of the techniques that are used in data mining include association, classification, clustering, finding sequential patterns and decision trees.

As organizations continue gathering more data in their data centers, it is necessary for them to find ways to get meaningful patterns from the data. These patterns provide information which is necessary to help the organization in making informed decisions regarding strategies to handle the market, competitors and organization processes. The organizations therefore need to use data mining techniques to formulate meaningful information from big data. This document conducts a comparative analysis of data mining techniques on big data with the use of Weka tool.

This paper starts with brief introduction, literature review, and discuss the methodology adopted and finally reviewed the results with suggestions for further research.

II. Related Work

This section evaluates different data mining techniques that are conducted on big data. There are several data inning techniques that may be applied to big data. These techniques may be classified into two subsections which are unsupervised and supervised learning. The supervised learning techniques incorporate a technique which is built before analyzing the algorithm being applied then to the analysis to help in estimating the parameters to the model [1]. Such techniques include decision tree, classification, neural networks, and Bayesian classification. Unsupervised learning does not require prior creation of a model with the algorithm being applied directly to the datasets to help in getting the results with a model being formulated based on the results being generated. A good example is clustering.

A decision tree uses a structure that depicts a flowchart with nodes where each of the nodes represents the value of an attribute while each branch is denoting the test results and tree leaves denoting various classes. The trees may be easily converted into classification trees. They are advantageous in that they are simple to understand and can give accurate and reliable results even when the data is minimal or constrained. It contains several algorithms such as C4.5, ID3, J48, CART, NB Tree, and REP Tree [4]. Bayesian classification uses a statistical classifier and can be used in the prediction of class membership probabilities.

Clustering technique groups data into classes for comparing the objects which are within the cluster and decide on their similarity index. Objects that ate within a similar cluster should have a high similarity index and are different from objects within another cluster which should also have a high similarity index within the cluster. Examples include partitioning methods, hierarchical methods, density based methods and grid-based methods.

Association rule mining discovers the association or correlation of relationships among a set of data items. This technique is mostly used when finding an everyday item that is set among various large data sets. Examples of algorithms include DDA, CDA, and Apriori. Prediction technique is mostly used in the identification of a relationship between the dependent and independent variables in a data set [5]. It may even be used in the prediction of values that are missing in a specific data set. An example of a predictive technique is a regression. Sequential patterns aim at discovering similar patterns in a data transaction in data transactions over a specified period. The patterns that are found are used in ore analysis.

III. METHODOLOGY

The methodology section explains the methods that are used in making a comparative analysis on data mining with big data using Weka tool. The section explains the data used in making the comparative analysis. The section also explains the features of Weka tool and how the tool will be used to conduct the analysis. The section finally explains how the results are interpreted from the analysis.

The methodology used in making a comparative analysis of the data mining techniques is the use of a database that contains dummy information on students from a high school for the past five years [2]. Some of the attributes in the tables that are in the database include roll number, enrollment number, final result, branch, and total marks. The tool that is used in the WEKA data mining tool. The WEKA, data mining tool, has four main modules, these are the explorer, experimenter, knowledge flow, and the simple CLI. The experimenter explores the data provided to it.

The experimenter allows creation, analysis, modifications and running of extensive experiments. The knowledge flow has a similar function to explorer but supports incremental learning. Simple CLI provides the functionality of the whole system through the command line. After the analysis was conducted, it's possible to see the accuracy. The results are then summarized to the advantages and disadvantages of each technique (Garg & K. Sharma, 2013, p. 4).

Results and Discussion

This section evaluates the results from the comparative analysis of big data using the Weka tool. The results from the analysis are then interpreted and simplified into merits and demerits of each data mining algorithm. The algorithms are grouped into classification algorithms, decision tree algorithms, and clustering algorithms. The results are explained below:

Classification algorithms (Decision tree &naïve Bayesian)

Merits

- Fast is classifying unknown records.
- Ability to handle both discrete and continuous data.
- Doesn't require a normalized database.
- Works well with numeric data.

Demerits

- Sensitive to small changes in data.
- Prone to errors when too many classes are used.
- Inability to predict the value of a continuous class attribute.
- Decision tree algorithms.

Decision tree algorithms (C4.5, ID3, CART)

Merits

1. Able to use continuous data helps in improving computational efficiency.
2. Able to handle data that has missing values.
3. Ability to find leaf nodes that enable the test data to be pruned thus the reduction of the number of tests.
4. Has the ability to handle both nominal and numeric values.

Demerits.

1. It requires that the target attributes only to have discrete values.
2. The data may be over-classified or over fitted if only a sample is tested.
3. May lead to having an unstable decision tree.

4. Only one variable is split.

Clustering algorithms (K-Means & K-Nearest Neighbor)

Merits

- Performs well even with other data missing.
- It is easy to debug and implement.
- Provides accurate results.
- Allows some noise reduction techniques which are used in the improvement of the classifier.
- K-means is fast.
- It has a very efficient algorithm which results even when the data sets are instinct.

Demerits

- Poor runtime.
- Requires high calculation of complexity.
- Does not consider the weight difference.
- Sensitive to irrelevant and redundant features.
- Cannot work with datasets that are not linear
- Cannot handle noisy data outlets.

From the results, it is clear that each algorithm has its own advantages and disadvantages. It is therefore necessary to evaluate the nature of the data being intended to be used in determining to decide on the most appropriate technique to use in mining the data. The strategy will help in getting meaningful information and patterns to organization in a faster way.

IV. Conclusion and Future Scope

Big data is a concept that has recently gained popularity. With organizations producing a lot of big data, they require data mining techniques to get meaningful information from the data. Some of the data mining techniques include clustering, decision trees, prediction techniques, sequential techniques and association rule mining. The Weka tool has been used to make a comparative analysis on big data mining techniques. The outcomes from the tool are then interpreted in the advantages and disadvantages of using various data mining techniques where each technique has its benefits and setbacks.

As big data continues to evolve in the future with more types of data being handled, it is necessary for organizations and researchers to formulate more data mining techniques to be able to evaluate and get meaningful patterns from the data. In the future, it will also be necessary to integrate machine learning algorithms. Machine learning algorithms will help organizations with the preparation of big data and conducting prescriptive analytics to make it easy for data mining on big data. In the future, it will also be necessary for organizations intending to undertake data mining procedures to formulate technology strategies to enable big data strategy to ensure that meaningful patterns and information are extracted from the available big data with the simplest process possible.

References

[1] The Baheti, A., & Toshniwal, D. (2014). Trend Analysis of Time Series Data Using Data Mining Techniques. 2014 IEEE International Congress on Big Data. doi:10.1109/bigdata.congress.2014.69

[2] Garg, S., & K. Sharma, A. (2013). Comparative Analysis of Various Data Mining Techniques on Educational Datasets. International Journal of Computer Applications, 74(5), 1-5. doi:10.5120/12878-9673

[3] Gole, S., & Tidke, B. (2015). A survey of big data in social media using data mining techniques. 2015 International Conference on Advanced Computing and Communication Systems. doi:10.1109/icaccs.2015.7324059

[4] Jamil, J. M., & Shaharanee, I. N. (2014). Comparative analysis of data mining techniques for business data. doi:10.1063/1.4903641

[5] Shobanadevi, A., & Maragatham, G. (2017). Data mining techniques for IoT and big data — A survey. 2017 International Conference on Intelligent Sustainable Systems (ICISS). doi:10.1109/iss1.2017.8389260

Authors Profile

Srinivasa Rao Putta pursed Master of Computer Applications (MCA) from Andhra University and is currently a Research Scholar at Rayalaseema University. He has more than 20 years of experience in Software Develoopemant and Applications.