Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 160 (2019) 690–695

Procedia Computer Science

www.elsevier.com/locate/procedia

The 3rd International Workshop on Healthcare Interoperability and Pervasive Intelligent Systems (HiPIS 2019)
November 4-7, 2019, Coimbra, Portugal

Benchmarking Business Analytics Techniques in Big Data

Catia Oliveira, Tiago Guimarães, Filipe Portela*, Manuel Santos

Algoritmi Research Centre, University of Minho, Guimarães, Braga, Portugal

Abstract

Technological developments and the growing dependence of organizations and society in the world of the internet led to the growth and variety of data. This growth and variety have become a challenge to the traditional techniques of Business Analytics. In this project, we conducted a benchmarking process that aimed to assess the performance of some Data Mining tools, like RapidMiner, in Big Data environment. Firstly, was analyzed a study where a group of Data Mining tools are evaluated and determined what is the best Data Mining tool, according to the evaluation criteria. After that, the best two tools considered in the study are analyzed regarding their ability to analyze data in a Big Data environment. Finally, studies were carried out on the evaluations of the RapidMiner and KNIME tools for their performance in the Big Data environment.

Keywords: Big Data, Analytics, Data Mining, Benchmarking.

1. Introduction

Nowadays, there is even more data cross on the internet every second than were stored in the entire internet just twenty years ago. That increase came with challenges like process them in. The Big Data offers solutions for the analysis of large sets of data, such as real-time data analysis. However, through data mining, predictive analytics, text mining, among others, is possible, more accurate insights and timely because only the relevant data are processed to achieve greater precision [1].

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: cfp@dsi.uminho.pt

10.1016/j.procs.2019.11.026

Organizations obtain great benefits in the analysis and extraction of information from data [2] because organizations depend on such data to succeed and gain competitive advantage. Thus, it is vital for organizations to give meaning to these data through your analysis in a timely manner, in order to help in the decision making of the organization. But do analysis in Big Data is not always easy, so in this project are identified some challenges in making Data Mining with Big Data. This paper is divided into seven sections: Introduction, Background, Challenges in Mining Big Data, Evaluation of Data Mining Tools, Benchmarking, Discussion and Conclusion and Future Work.

2. Background

2.1. Big Data

The term of big data is often used to describe enormous datasets, but big data is not only that. The size itself does not define big data. Does not exist only one definition for big data, but according to Garg et al. (2016) [1] can be define like a collection of complex and large data sets whose size is beyond ability of traditional data processing applications and other relational database management tools to process, manage and capture the whole data within the desired span of time. Big data in addition to referencing the volume also refers to data variety, and velocity called as the three V's of big data. Volume refers to the size of data that is generated from a variety of sources. Variety, according to Gandomi & Haider (2015) [3], refers to the structural heterogeneity in a dataset, that is that the dataset can be structured, semi-structured and unstructured. Structure data means that the data have a defined format in a way to be easy to analyze. Unstructured data type has difficulties in analyzing like video, images and audio due to its structure. Semi-structure data are not in accordance with strict standards, instead of table data can be stored in XML (Extensible Markup Language) format. Velocity refers to how fast the data is generated and processed [4]; this data comes from mobile users, social media, internet users, etc. [1]. The data is who comes from a sensor is constantly moving to the database store, so the traditional systems do not be able to deal with data in constant movement [5].

Nowadays, to gain a competitive advantage in an organization, it is necessary to analyze large data to obtain crucial information for decision making. In addition, it is necessary to predict what the customer will need in the future. Therefore, analyzing big data, besides being challenging, is very important.

2.2. Data Mining

Data Mining can be defined as the process of discovering knowledge from large datasets. This process of discovering knowledge is done through some techniques and used to help the process of decision making. It is possible to analyze the customer behavior, in respect of some product, by analyzing historical data, and then make decisions about the future of some products [6]. This term is also referred to Knowledge Discovery in Databases (KDD) process, this process has various steps such as data selection, data pre-processing, data transformation, etc. and refers to the process of discovering useful knowledge from data, such as associations, summaries, outliers, etc. [7].

Data Mining requires an algorithm or method to analyze data. The objective of data mining process is to build an efficient predictive or descriptive model of large data that be able to explain it. The predictive model, as the name says, is for to predict unknown or future values of variables of interest. The descriptive model finds a pattern in data. That's also important to optimize the model parameters for successful applications of any data mining approach [8].

There are many data mining techniques, but the principal techniques for finding patterns are regression, classification, clustering and association. Classification classifies data into different classes, and the goal is to create a set of classification rules that will answer a question or predict a behavior [10]. Regression is used to predict a range of numeric values, in a set of data and see the relationship between two variables [6]. Clustering is a technique that is grouping the data based on their characteristics, aggregating them according to their similarities [6].

There are 2 categories of data mining models [11]:

- Unsupervised Model- This category focus on finding patterns/clusters in the given data.
- Supervised Model- This category deal with training the system with historical data and make predictions.

3. Challenges in Mining Big Data

Analyze a large amount of data can be challenging because many tools are not able to deal with such data. Social Networking like Facebook, Twitter, Instagram, etc., are generated huge amounts of data per day at very high speed. This data needs to be processed in real-time and can be used to predict stock market behavior, for example. For that,

we need tools that deal with this type of data [11]. Therefore, we cannot store, manage and analyze big data with traditional methodologies or data mining software tools. So Big Data Mining, according to Jaseena and David (2014), is the capability of extracting useful information from large datasets or streams of data which that due to volume, variety and velocity it was not possible before. Dealing with Big Data itself is challenging, your mining continues to be challenging, if not even more. Jaseena and David (2014) considered as challenges of Big Data Mining heterogeneity, scale, timeliness, complexity and privacy.

3.1. Heterogeneity, Scale and Complexity

As stated in point 2.1, Big data is defined by your variety; they can be structured, semi-structured and unstructured. Therefore, the presence of different rules or patterns in data become challenging for analyzing. Convert unstructured data to structured, for analyzing is a challenging for Big Data Mining [12].

Incomplete data can also be a challenge because it creates uncertainties. In Data Mining, there is a way to deal with incomplete data, such as ignore the missing values or data imputation. In data imputation, the gaps are fulfilled with the goal to produce a better model them the original one (with original data) [12].

As has been said, traditional tools are not able to handle this growing data, and real-time processing, so alone dealing with the volume of data is already a challenge. Therefore, organization, data analysis, retrieval and modeling are also challenging due to scalability and complexity of data that needs to be analyzed [12].

3.2. Timeliness and Privacy

Analyzing a large amount of data can take a long time because of its size and complexity. However, there are situations where results are needed at the moment. Due to the importance of data to make a decision the processing/mining task must be finished within a certain period, otherwise, the results become less valuable or even worthless [12] [13]. The privacy is also a challenge because the tools used for analysis, stores, manages, etc., utilizes data from a lot of sources. This leads to a risk of exposure of the data, making it vulnerable. So, the analysts must deal with the data carefully [14].

4. Evaluation of Data Mining Tools

Ventura et al. [15], have done a study about open-source Data Mining tools. This study only addresses the data mining tools not include de big data area. Thus, at the end of this study, the best data mining tools considered in the study are analyzed in the big data area, investigating what each of these tools has to offer for the big data area.

The study evaluates 19 open-source data mining tools and provides an extensive study based on a wide set of features that any tool should satisfy. The evaluation is carried out by two methodologies, the first one is based on scores provided by experts for a subjective judgement and the second one performs an objective analysis about witch feature are satisfy by each tool [15].

4.1. Open-source Data Mining Tools

The list of open-source data mining tools was taken from the KDnuggets (KDnuggets is a leading site on Business Analytics, Big Data, Data Mining (DM), Data Science, and Machine Learning). Table 1 contains the name and a short description of the tools considered [15].

Table 1. Tools considered for benchmarking.

Data Mining Tools	Description
ADaM Algorithm Development and Mining	This is a data mining toolkit and provides a suite of tools for each of the basic data mining processes.
ADAMS (Advanced Data Mining And Machine learning Systems)	ADAMS is a flexible workflow engine aimed at quickly building workflows *. Each step of the knowledge discovery process is described by a graphical user interface.

* https://adams.cms.waikato.ac.nz/

Data Mining Tools	Description
AlphaMiner	This tool is a general-purpose data mining system design to facilitate the implementation of data mining processes. Provides the user with a wide range of functionalities to carry out different processes like data access, data manipulation, etc.
Cramer Modelling Segmentation and Rules (CSMR)	CSMR is a data mining suite used for business analytics and provides an integrated environment for predictive modeling, segmentation, data visualization, statistical data analysis and SQL queries.
Databionic ESOM (D.ESOM)	This tool is a suite of programs that perform data mining tasks like clustering, visualization and classification with emergent self-organizing maps.
DataMelt	DataMelt is a software for numeric computation, mathematics, statistics, symbolic calculations, data analysis and data visualization†. With DataMelt, it is possible applying different data mining techniques by means of a graphical user interface.
ELKI (Environment for deveLoping KDD-applications)	ELKI is a data mining software focused in research in algorithms, with an emphasis on unsupervised methods in cluster analysis and outlier detection‡.
The gnome data mine tools (GDataMine)	The GDataMine is a set of open-source data mining programs. It includes algorithms for the association rule mining task, a Bayes classifier and a decision tree classification algorithm.
KELL (Knowledge Extraction based on Evolutionary Learning)	This tool can be used for many different knowledge data discovery tasks. KEEL provides a simple graphical user interface based on data flow to design experiments with different datasets and computational intelligence algorithms in order to assess the behavior of the algorithms §.
KNIME (Konstanz Information Miner)	This tool is designed for discovering the potential hidden in data, mining for fresh insights, or predicting new futures**. The KNIME Analytics Platform incorporates hundreds of processing nodes for data I/O, preprocessing and cleaning, modeling, analysis and data mining.
MiningMart	MiningMart is a graphical tool for processing and transforming data stored in very large databases.
ML-Flex	ML-Flex is an open-source software package designed to enable flexible and efficient processing of disparate data sets for machine learning analysis ††.
Orange	Orange is a machine learning ad data mining software and presents a visual programming front-end for explorative data analysis and visualization.
RapidMiner	This tool provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics.
Rattle (R Analytical Tool Learn Easily)	This tool offers a graphical user interface for DM tasks and is based on the R language.
SPMF (Sequential Pattern Mining)	SPMF is an open-source data mining library focused on pattern mining tasks.
Tanagra	Tanagra proposes several data mining methods such as exploratory data analysis, statistical learning and machine learning.
Vowpal Wabbit (VW)	VW is fast out of the core learning system library and supports a number of machine learning problems and others.
WEKA (Waikato Environment for Knowledge Analysis)	WEKA is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining and visualization‡‡.

4.2. Methodology Used

As has already been said, Ventura et al. [15], used two methodologies a subjective procedure and objective analysis. In the first methodology, four main categories were evaluated: performance, functionality, usability and support of supplementary activities. For each category, criteria were defined, and a set of questions were answered. In the second evaluation procedure, four groups are considered: system requirements, type of approaches, process-dependent features and user interface feature. In this methodology, it is only checked if the features are satisfied by each tool. Here only be the final results of this study will be shown. The details of this study can be seen in the official article "Evaluation and Comparison of Open Source Software Suites for Data Mining and Knowledge Discovery, 2017" written by Abdulrahman et. al [15]. Weka was the reference tool for all other tools to be evaluated. Therefore, the scores were given with reference to this tool. After assigning the scores (1 to 5) and placing the appropriate weights for each criterion, the final result for each category, taking into account each criterion, was as shown in Figure 1. Here, the result is summarizing and is shown in percentage (in per unit basis).

---

† https://jwork.org/dmelt/

‡ https://elki-project.io/

§ http://www.keel.es/

** https://www.knime.com/about

†† http://mlflex.sourceforge.net/

‡‡ https://www.cs.waikato.ac.nz/ml/weka/

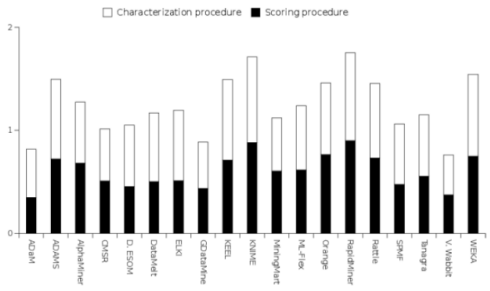

Fig. 1 - Final result of the two methodologies. Taken from: Ventura et al. [15]

As can be seen, the two tools with the best percentage of the two methodologies are RapidMiner and KNIME. After having classified the two best data mining tools, they will be analyzed in relation to the Big Data area.

4.3. RapidMiner and Knime

RapidMiner is a data mining tool and offers a Big Data extension, call Radoop, which allows the analysis of large data in a big data environment. RapidMiner Radoop is a code-free environment for designing advanced analytic processes. Radoop includes more than 60 operators for data transformations as well as advanced and predictive modelling that run on a Hadoop cluster. Supports Cloudera, Hortonworks, MapR, Amazon EMR, Apache, Microsoft Azure HDInsight, supports Hive on Spark and Hive-on-Tez and allows the use of SparkR, PySpark, Pig, and HiveQL scripts. Hadoop is an open-source framework that allows the processing and management of a huge volume of structured and unstructured data across the clusters. Radoop allows performing input and output operations on HDFS (Hadoop Distributed File System) as well on the Hive database. In Hive are analyze large datasets that are stored in HDFS [16]. KNIME is a data mining tool and offers Big Data Extensions that integrate the power of Apache Spark and Apache Hadoop with KNIME. Also allows import/export and read/write HDFS data and perform analytics within Hive and Impala, with Impala, it is possible to query data stored in HDFS in real-time. KNIME is a tool for data analysis, manipulation, visualization, and reporting. Both tools, KNIME and RapidMiner, are a code-free environment graphical user interface.

5. Benchmarking

In this chapter, the two best tools resulting from the previous study, in this case, KNIME and RapidMiner, will be evaluated, and their performance will be compared in different scenarios.

5.1. RapidMiner Radoop vs RapidMiner Analytics

Gaspar et al., (2015) conduct an experiment to compare the performance between Radoop, an extension for the RapidMiner, and RapidAnalytics, an open-source server solution of RapidMiner for data mining and business analytics. This experiment consisted of the execution of several data transformations tasks in the Hadoop cluster (Radoop) and in the RapidAnalytics (in-memory). In this experiment, was created measurements on how Radoop performs and scales with the size of the data set and the number of processes nodes. Was used 4 to 16 nodes in the cluster and experimented with 128MB to 8GB data sets in Radoop and 128MB to 1GB in RapidAnalytics. In the first experiment, increasing the data size, both RapidAnalytics and Radoop scale linearly but Radoop finishes the job much faster, even with 4 processing nodes. But, in small data sets, like 128MB, the run time it is the same in the Hadoop cluster because in each block of 64MB (by default) is processed by only one node. That means that one block of 128MB is processed by two nodes. In the second experiment, data transformation were selected several attributes, filter out some examples and aggregated according to an attribute, then was renamed newly created columns and was saved the result. Radoop performed this task more fast then RapidAnalytics, even in small data sets.

5.2. KNIME – Apache Hive based on MapReduce vs Apache Hive based in Tez

Kotter [17], conduct an experiment to compare the runtime performance of two different Hive execution engines, Hive execution engine based on MapReduce and Tez, by running an SQL query several times for each engine.

Apache Tez is a framework for building high-performance batch and interactive data processing applications. MapReduce, it is a framework for writing applications that process a large amount of structured and unstructured data stored in the HDFS [17]. In this experiment was use Hortonworks Sandbox and the "SQL Generation" to assemble database operations modularly and create complex SQL statements. Apache Tez has better performance then MapReduce; the size of the data set was small. It also compared the performance of the standard data format with the Optimized Row Columnar data format, where each query was executed 15 times. The time execution in MapReduce and Tez with ORC data format remains stable throughout the interactions.

The runtime of each interaction in the Apache Tez varied very little from iteration to iteration. Apache MapReduce shows more irregularities between interactions but not accentuated.

6. Conclusion and Future Work

Based on these studies, it is possible to conclude that both tools have a good performance in analyzing data in a Big Data environment. Apache MapReduce is inefficient for applications that need to constantly reuse the same dataset [18], this may have been a cause for poor performance compared to Apache Tez; however Tez represents an alternative to MapReduce with the fast response time. Radoop showed a good performance with the increasing of data size and the number of nodes, except with small datasets, even with 4-8 processing nodes, here the single machine had a better performance. The development of this project helped to realize the challenges that Big Data brings to the area of Data Mining. The research and evaluation performed on different tools helped to understand what criteria should be considered when selecting a Data Mining tool. In addition, how RapidMiner and KNIME adapted to the Big Data area in such a way that it can be implemented. Both tools offer quite similar methods and techniques. However, more benchmarking is needed to understand which of these two tools brings more benefits in Big Data analysis in the same environment with the same conditions.

Acknowledges

This work has been supported by national funds through FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2019and Deus ex Machina (DEM): Symbiotic technology for societal efficiency gains - NORTE-01-0145-FEDER-000026.

References

[1] Garg, N., Singla, S., & Jangra, S. (2016). Challenges and Techniques for Testing of Big Data. Procedia Computer Science, 85, 940–948.
[2] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
[4] Gupta, S., & Chaudhari, M. S. (2015). Big Data Issues and Challenges. International Journal on Recent and Innovation Trends in Computing and Communication, 3(2), 62–67. https://doi.org/10.1109/HICSS.2013.645
[5] Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and good practices. In 2013 6th International Conference on Contemporary Computing, IC3 2013. https://doi.org/10.1109/IC3.2013.6612229
[6] Sharma, A., & Kaur, B. (2017). A Research Review On Comparative Analysis Of Data Mining Tools , Techniques And Parameters, 8(7), 523
[7] Gaur, N., & Baj, J. (2018). DATA MINING TOOLS- A COMPARATIVE STUDY, XII, 1–5.
[8] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & Coello Coello, C. A. (2014). A Survey of Multiobjective Evolutionary Algorithms for Data Mining : Part I, 18(1), 4–19.
[9] Jothi, N., Rashid, N. A. A., & Husain, W. (2015). Data Mining in Healthcare – A Review. Procedia - Procedia Computer Science, 72, 306
[10] Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V, & Rong, X. (2015). Data Mining for the Internet of Things : Literature Review and Challenges, 2015(i). https://doi.org/10.1155/2015/431047
[11] Hashmi, A. S., & Ahmad, T. (2016). Big Data Mining : Tools & Algorithms, 4(1), 36–40
[12] Jaseena, K. U., & David, J. M. (2014). Issues, Challenges, and Solutions: Big Data Mining, 131–140
[13] Atanassov, A., & Al-Barznji, K. (2017). A SURVEY OF BIG DATA MINING : CHALLENGES AND TECHNIQUES A SURVEY OF BIG DATA MINING : CHALLENGES AND TECHNIQUES, (December)
[14] Jothi, B., Amudha, S., & J, J. (2018). Research Challenges in Mining of Big Data: A survey, 118(20), 241–247
[15] Ventura, S., Altahi, A. H., Luna, J. M., & Vallejo, M. . (2017). Evaluation and Comparison of Open Source Software Suites for Data Mining and Knowledge Discovery (September). https://doi.org/10.1002/widm.1204
[16] Beckmann, M., Nelson, F. F., Pires de Lima, B. S. L., & Costa, M. A. (2014). A User Interface for Big Data with RapidMiner A User Interface for Big Data with RapidMiner, (August). https://doi.org/10.13140/2.1.5152.4488
[17] Koetter, T. (2015). Hive execution engine comparison with the KNIME Analytics Platform. [Blog] KNIME Blog [Accessed 24 Out. 2018]
[18] Torres, H., Portela, F., & Santos, M. F. (2018). An Overview of Big Data Architectures in Healthcare An overview of Big Data architectures in healthcare, (May). https://doi.org/10.1007/978-3-319-77700-9