UTM Computing Proceedings
Innovation in Computing Technology and Applications
Volume: 2| Year: 2017 |ISBN: 978-967-0194-95-0

Spark Machine Learning for Decision Tree Classification

Norhayati Saad [1], Shafaatunnur Hassan [2]

[1] UTM Big Data Centre, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

[2] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

[1] norhayati82@live.utm.my, [2] shafaatunnur@utm.my

Abstract. Nowadays, due to the powerful of the volume of data, most researchers undergoing obstacles in handling and recovering the required data. The intent of this research is to explore and learn a deep understanding of Apache Spark Machine Learning. This research concentrates on classification problems of Handwritten Digits MNIST dataset using Decision Tree and Random Forest Classifier. Therefore, this research has conducted and evaluate several experiments using training and test MNIST dataset using the classifying algorithm. The finding of this study can be used for further understanding of Apache Spark and Classification Problems.

Keywords: Apache Spark, Classification Problems, Decision Tree, Random Forest Classifier, MNIST

1.0 Introduction

As the exponential growth of data volume and the multitude of sources have created new technical and application changes. For example, data generated has been estimated at 2.5 Exabytes of data per day [3]. That value of data is enormous and are not being used.

To deal with this classification problems, Apache Spark which uses Central Processing Unit (CPU) as a processor are use to overcome the problem. Spark divide and runs in parallel to solve the problem faster. Sparks MLlib, the ML library consists of popular learning algorithms like classification, which is beneficial when dealing with classification problems [2].

For experimental purpose, classic MNIST handwritten digit recognition dataset from the MNIST Database of handwritten digits is used for training and testing the ML algorithm to handle classification problems [3]. Moreover, Decision tree and Random Forest Classifier will be the chosen ML techniques to be used as it has been widely used to handle ML classification problems.

The main objective of the project is to explore basics understanding of Apache Spark and implementing Apache Spark ML technique on Classification Problems. The other objective is tounderstand the challenges or new knowledge gain from using Apache Spark. Furthermore, this research only focus on Apache Spark platform, Classification Problem, Decision Tree, Random Forest Classifier and MNIST Dataset.

This paper is organized as follows: In Section 1, present the introduction of this paper. Then, Section 2 explains some related works. While in Section 3, provide the research methodology that is being used in this research. Next, discuss the experimental results in Section 4. In Section 5, present the discussion of the results that are obtained in the experiments and finally, Section 6 provides the conclusion for this research.

2.0 Related Work

From the past decade, with the development of multimedia technology and Internet, image classification problem has received considerable attention [4]. Numbers of data in diverse classes are unbalanced for some classification problems [5]. Therefore, with the help of ML, technologies could exploit unlabeled cases. Hence, enhance classification accuracy and attract more attentions in the image classification field [1]. Classification is a supervised ML technique in which the data is separated into training and testing sets [7].

Furthermore, apache Spark is a popular open-source platform for large-scale data processing that is well-suited for iterative ML tasks [11]. Spark provides primitives for in- memory cluster computing that enables it to query data much faster compared to Hadoop which is well suited for large-scale ML purposes. This tool is specialized at making data analysis faster [7]. Apache Spark also provides high level APIs in Java, Scala, and Python and supports a rich set of high-level tools including Spark SQL for SQL and structured data processing, Spark MLib for ML, GraphX for graph processing and Spark streaming. Spark can run a job 100 times faster than Hadoop.

Moreover, a decision tree consists of a tree structural model and a set of decision nodes and leaves [8]. It is one of ML methods use statistical learning to identify boundaries. Decision trees look at one variable at a time and are a reasonably accessible though rudimentary ML method. Training data is used to train the decision tree model by running each data point through branches for making predictions. The tree maps each training data point per fectly to which the accurate object is in. Traditional decision tree learning algorithms aim to maximize the classification accuracy [8].

In addition, the random forests algorithm is a ML technique that is increasingly being used for image classification and the creation of continuous variables such as percent tree cover and forest biomass. Random forests Classifier or random decision forests are an ensemble learning method for classification, regression and other tasks [10]. Random forests, like decision trees, can be used to solve classification and regression problems [11].

The MNIST dataset for this research contains binary images of handwritten digits that has a training set of 60,000 examples and a test set of 10,000 examples [3]. This dataset is a good database as it only need minimal efforts on preprocessing and formatting. Each one of the digitized numerals are represented with a 20 x 20 sized pixel box while preserving their aspect ratio. Although originally binary images, the standard dataset is now greyscale due to resizing with interpolation [14]. There are many places online to get the other MNIST dataset, such as Kaggle Website that can gives convenient access to the data.

3.0 Methodology

This section presents the methodlogy adopted in this research. There are three phases to complete. Figure 1 illustrates the research steps involved in this research which includes phase 1 of background study, phase 2 of define the objectives for this reseach and finally phase 3 of documentation.

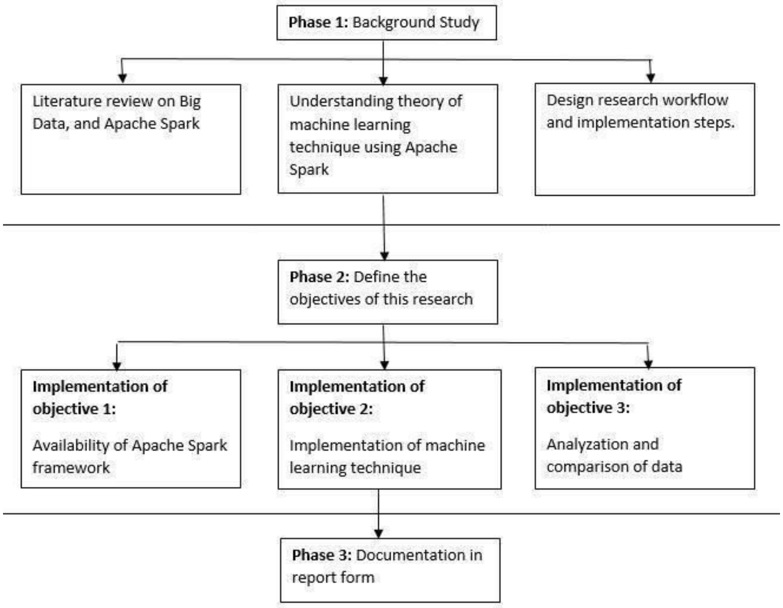
Phase 1: Background Study
Literature review on Big Data, and Apache Spark
Understanding theory of machine learning technique using Apache Spark
Design research workflow and implementation steps.
Phase 2: Define the objectives of this research
Implementation of objective 1: Availability of Apache Spark framework
Implementation of objective 2: Implementation of machine learning technique
Implementation of objective 3: Analyzation and comparison of data
Phase 3: Documentation in report form

Figure 1. Research workflow

4.0 Experimental Results

In this research, several experiment are made using two sets of machine learning techniques Decision Tree and Random Forest Classifier using MNIST dataset.

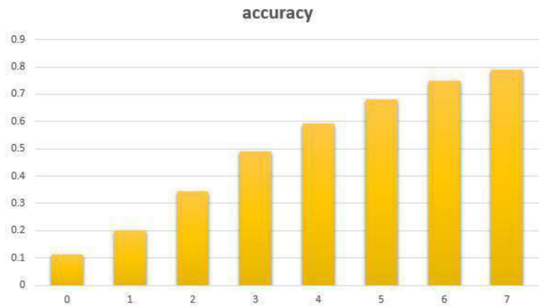

Figure 2. Result of ‘maxDepth’ experiment

Figure 2 shows the the accuracy results of the learned tree using maxDepth. The results took around 1.14 minutes to be completed since it is training numerous trees which get deeper and deeper. It can be concluding from the figure that the accuracy increases and achieving more powerful classifiers with the deeper, larger trees. However, deeper trees indeed are more powerful but they are not always better than the other.

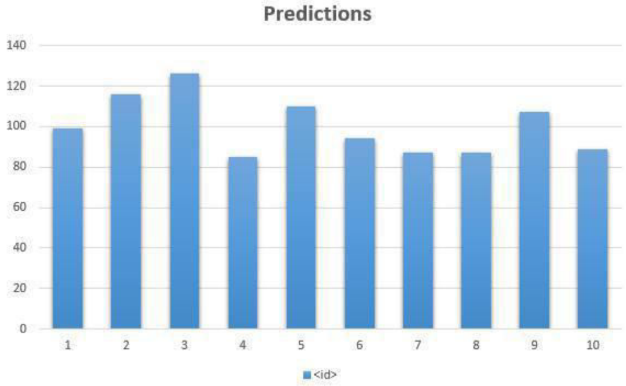

Figure 3. Random Forest Classifier prediction results

Figure 3 shows the predictions results of Random Forest Classifier . As shown from the results, many labels and predictions are matched, however some are not. More tuning can done to improved the predictions accuracy deeply.

5.0 Discussion

One of the main challenge in performing this research is implementing the theory of Apache Spark machine learning techniques into one of the classification problems. This research intends to explore and deep understanding Apache Spark ML

Based on the results that are obtained from this experiment, a proper tuning to the parameters based on held-out data and more testing must be done. Furthermore, more limitations and challenges need to be improved and upgrade for a much better framework environment in Apache Spark.

6.0 Conclusion

The principle objective is to explore and have a deep understanding in Apache Spark by handling classification problems using MNIST database. This experiment result can guide other researchers or data scientist beginner to understand how Apache Spark and encourage more interaction to learning and implement Apache Spark for BDA. In future, different machine learning techniques need to be experiment using Apache Spark to get more significant result and understanding learning. Therefore, further research needs to be done for the bigger improvement in BDA.

References

1. Smith, Jinhua Liu, W. Y. a. C. S., Hualong Yu. Combining Active Learning and Semi-Supervised Learning Based on Extreme Learning Machine for Multiclass Image Classification. Springer International Publishing Switzerland, 2015.
2. Lekha R. Nair, S. D. S., Sujala D. Shetty. Applying spark based machine learning model on streaming big data for health status prediction. Computers Electrical Engineering, 2017.
3. LeCun, Y. The MNIST Database of handwritten digits.
4. Chang, C.-C. and Lin, C.-J. LIBSVM : A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2013.