Rating Prediction using Deep Learning and Spark

Monika Mishra, Mingoo Kang+, Jongwook Woo*
Computer Information Systems Department
California State University Los Angeles, California, United States
+Hanshin University, Korea
[e-mail: mmishra2@calstatela.edu, kangmg@hs.ac.kr, jwoo5@calstatela.edu]
*Corresponding author: Jongwook Woo

Abstract

Big Data architecture has been presented to afford the massive data set for storing and computation. And, deep learning has been popular for machine learning community as known that it has high accuracy for the massive data set. The aim of this paper is to build the models with Deep Learning and Big Data platform, Spark. With the massive data set of Amazon customer reviews, we develop the models in Amazon AWS Cloud services in order to predict the users' ratings for the items at Amazon. And, we present a comparative conclusion in terms of the accuracy as well as the performance with the Deep Learning architecture with Spark ML and the traditional Big Data architecture, Spark ML alone.

Keywords: Big Data, Deep Learning, Spark, Analytics Zoo, Amazon EMR, machine learning, recommendation

1. Introduction

Amazon has been leading the e-commerce markets and expands to be a leader at the IT industry providing many services in the cloud computing. Besides the companies like Amazon, many e-commerce has the disadvantage as online shoppers do not have the ability to physically inspect or try on the items being considered for purchase [1-3]. In the predictive analysis, the goal of creating a recommendation system is to recommend one or more "items" to "users" of the system, which should become helpful for the customer who cannot access the products physically. This model can learn about a user's preferences through observations made on how they rate products. Based on those observations, it recommends new items to the users when requested. It also predicts ratings for different products. These predictions are of great significance for the business.

These days, the data set becoms too large to store and process, which initiates Big Data. Big Data is defined as non-expensive frameworks, mostly on distributed parallel computing systems, which can store a large-scale data and process it in parallel. A large-scale massive data set means a data of giga-bytes or more, which cannot be processed or stored using traditional computing systems [4]. Hadoop and Spark are the popular Big Data platforms and lately NoSQL DB and search engine such as Elasticsearch are regarded as Big Data frameworks.

Spark has been adopted to Big Data platform as an efficient in-memory distributed computing engine. Furthermore, its Machine Leaning library, Spark ML, is popular for predicting the trends of massive data set.

Deep Learning has received highlights past several years, mostly after Google shares its TensorFlow library and NVidia's GPU become non-expensive as multi-core parallel computing processor in a single chip.

There has been many approaches to integrate distributed systems and multi-core GPU systems, such as, DeepLearning Pipeline for Apache Spark by Databricks, TensorFlowOnSpark by Yahoo, BigDL/Analytics Zoo by Intel, DL4J by Skymind, Distributed DeepLearning with Keras & Spark by Elephas. This paper adopts Analytics

This study was supported by the AWS Cloud Credits for Research program.

Zoo and Amazon EMR to execute the models using Spark ML and Analytics Zoo.

The paper has the section 2 as Related Work. The section 3 presents the background with dataset. The section 4 illustrate the Big Data Deep Learning architecture of our work. The sections 5 and 6 are the experimental methods and the conclusion respectively.

2. Related Work

Bhavesh [5] classify Amazon product review to positive and negative in the traditional systems. He concentrated on just one product category – *baby*. He also performed sentimental analysis for one of the baby products.

Max [6] performed descriptive analysis using *Sparklyr* platform but not predictive analysis.

Monika et al [7] present descriptive and predictive analysis using Big Data on Cloud Computing by building recommendation models.

Our paper presents predictive analysis to predict users' ratings using Big Data Spark ML. Furthemore, predictive analysis is done even by adopting the integrated systems of Deep Learning and Spark to compare the performance and accuracy.

3. Backgrounds

The big data analysis and prediction is mostly based on Hadoop and Spark. Big Data is defined as non-expensive frameworks, mostly on distributed parallel computing systems, which can store a large-scale data and process it in parallel [6]. Hadoop & Spark clusters have mostly selected as the solutions for Big Data analysis and prediction in the world.

In the paper, the data set is uploaded to Hadoop Distributed File systems and Amazon AWS S3 and transformed to be analyzed using Spark and Analytics Zoo. The data set is acquired from the *amazonaws* site [10]. The data has details about the products reviewed on Amazon site between 2005 and 2015 in the United States. The Amazon product review dataset contains 15 attributes and has about 6.93 million records. The total file size is 3.63 GB, which is big so that the traditional systems cannot afford or takes serious time to compute the prediction.

Our models for predictive analysis is to implement the rating prediction in both Spark ML and Analytics Zoo. Each metric of the models is evaluated in terms of accuracy, MAE (*Mean Absolute Error)* and performance, *Time*.

4. Big Data Deep Learning Architecture

Spark is in-memory distributed computing engine with linear scalibilty and it has been popular as integrated to Big Data plaforms such as Hadoop and NoSQL DB. As Deep Learning grows exponentially on single chip with multi-core computing power, it has had many different architectures to integrate Spark and Deep Learning: DeepLearning Pipeline for Apache Spark by Databricks, TensorFlowOnSpark by Yahoo, BigDL/Analytics Zoo by Intel, DL4J by Skymind, Distributed DeepLearning with Keras & Spark by Elephas.

Fig. 1 shows the BigDL architecture by Intel, which is integrated into Analytics Zoo. In the paper, Analytics Zoo is adopted as a experimental platform using AWS EMR cloud service. Analytics Zoo supports Deep Learning models, Keras and Tensor Flow as well as BigDL and can run on Hadoop/Spark cluster.

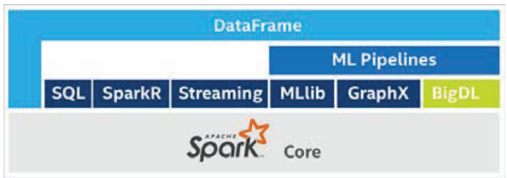

Fig. 1. BigDL Architecture [8]

5. Experimental System and Results

5.1 Cloud Computing Systems

We used storage and big data services offered by Amazon AWS cloud. We extensively worked on Hadoop and Spark system components in AWS EMR. The EMR instances of the Hadoop/Spark cluster details are given below:

EMR Instances: r3.2xlarge
Number of Nodes: 3
Memory size: 183 GB (= 61 GB x 3)
CPU: 8 vCPU

CPU speed: 3.1 GHz
Storage: 960 GB (= 2 x 160GB x 3)

5.2 Prediction Programming

For prediction, we adopt Python 2.7 programming language with the machine / deep leaning libraries in Spark ML (Machine Learning) and Analytics Zoo.

The features selected to build models are [user, item] and the label is [rating] in 1 to 5. In Spark ML, we have used ALS (Alternating Least Squares) algorithm to build the recommender. Using Spark ML pipeline, we've defined parameters and used *pipeline fit* method to train the model. Then we evaluate the model by computing MAE in the rating prediction for each product by the user.

In Analytics Zoo, a neural network recommendation system, Neural Collaborative Filtering(NCF) with explict feedback, is used as a Recommender API in Analytics. Besides, optimizer of BigDL is adopted to train the model. NCF leverages a multi-layer perceptrons to learn the user/item interaction function, at the mean time, NCF can express and generalize matrix factorization and the users can build a NCF with or without matrix factorization [7].

5.3 Experimental Results

In Spark ML, the dataset is split into 80:20 ratios for training and testing. Initially, we train the model with *CrossValidation* method with 8 folds. Then, we used *TrainValidation* method. Parameters are applied to train the ALS models as follows:

Rank: [1, 5]
Maximum Iteration: [5, 10]
Regularization Parameters: [0.3, 0.1, 0.01]
Alpha: [2.0, 3.0]

The pipeline is used as an estimator to train the model. The models are then evaluated using *RegressionEvaluator* for the Mean Absolute Error (MAE).

In Analytics Zoo on the same Hadoop Spark cluster, the data set is also randomly split into 80:20 ratios for training and testing. NCF neural network algorithms is adopted with *Optimizer* for training a model. Then, several batch sizes are given in the range of 110K to 300K for more about 6.5 million records with 10 epochs.

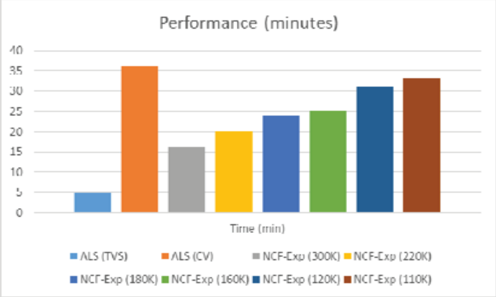

Fig. 2. Performance

Fig. 2 shows the experimental result of the computing times when evaluating the models with the training and test data set. ALS modes in PySpark take 5 and 36 minutes for TVS (*TrainValidationSplit)* and CV (*CrossValidation)* respectively. For Analytics Zoo, it takes 16 to 36 minutes as decreasing the batch size from 300,000 to 110,000.

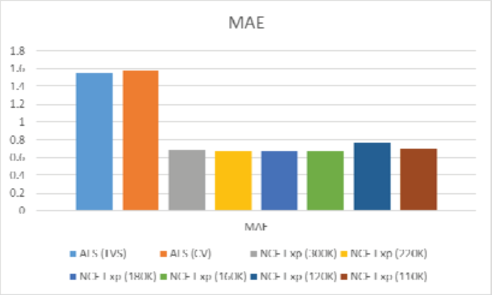

Fig. 3. Mean Absolute Error (MAE)

Fig. 3 shows the MAEs as a measurement to evaluate the accuracy of the models. ALS modes in PySpark has 1.55 and 1.574 for TVS and CV respectively. For Analytics Zoo, the MAEs varies from 0.693 to 0.7036 while decreasing the batch size from 300,000 to 110,000. When the batch size is 220,000, MAE becomes the minimum value 0.6652. In summary, integrating deep learning with Spark has 55% less MAE than Spark alone. The computing times in Analytics Zoo are similar to build an ALS model with Cross Validation in Spark ML.

6. Conclusion

In the paper, Hadoop Big Data systems and Amazon S3 are adopted to store and analyze for over 5GB amazon data set, which is linearly scalable when the data set grows further.

Analytics Zoo is a platform to integrate the scalable Spark computing engine and deep learning algorithms, which can obtain the strength of distributed computing systems and multi-core parallel computing systems.

The paper presents the experimental result to compare the Spark ML and Analytics Zoo. In order to measure and compare the performance and accuracy of the systems, recommendation model is implemented to predict ratings as predictive analysis.

We observe that the Deep Learning Spark models present 55% more accurate predictions than Spark ML alone while the computing time is similar, especially with Cross Validation modeling. Besides, it illustrated that the integrating Spark cluster and multi-core deep learning can be obtained as a distributed and single-chip parallel computing systems.

References

[1] A. S. Jain, “Top 10 Benefits of Online Shopping (and 10 Disadvantages),” *ToughNickel*, 2018. [Online]. Available: https://toughnickel.com/frugal-living/Online-shopping-sites-benefits.

[2] C. Morah, “Shopping Online: Convenience, Bargains And A Few Scams,” *Investopedia*, 2018. [Online]. Available: https://www.investopedia.com/articles/pf/08/buy-sell-online.asp.

[3] D. L. Montaldo, “When It Is Best to Shop Online and When It Is Not?,” *The Balance*, 2019. [Online]. Available: https://www.thebalance.com/the-pros-and-cons-of-online-shopping-939775.

[4] Woo, Jongwook, & Xu, Yuhang (July 18-21, 2011), Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing, The 2011 international Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), Las Vegas.

[5] B. Patel, “Predicting Amazon product reviews' ratings,” *Towards Data Science*, 26-Apr-2017. [Online]. Available: https://towardsdatascience.com/predicting-sentiment-of-amazon-product-reviews-6370f466fa73.

[6] M. Woolf, “Playing with 80 Million Amazon Product Review Ratings Using Apache Spark,” *minimaxir*, 02-Jan-2017. [Online]. Available: https://minimaxir.com/2017/01/amazon-spark/.

[7] Intel Analytics Zoo, https://software.intel.com/en-us/ai/analytics-zoo

[8] BigDL: Distributed Deep Learning on Apache Spark, https://software.intel.com/en-us/articles/bigdl-distributed-deep-learning-on-apache-spark

[9] Monika Mishra, Jaydeep Chopde, Maitri Shah, Pankti Parikh, Rakshith Chandan Babu, Jongwook Woo, "Big Data Predictive Analysis of Amazon Product Review", KSII The 14th Asia Pacific International Conference on Information Science and Technology (APIC-IST) 2019, pp141-147, ISSN 2093-0542

[10] S3.amazonaws.com (n.d), Amazon Reviews Multilingual Dataset from https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_US_v1_00.tsv.gz