ISBN 978-981-14-1455-8
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
WCSE_2019_SPRING
Yangon, Myanmar, February 27-March 1, 2019, pp. 148-153
doi:10.18178/wcse.2019.03.025

Predictive Analytics on High-Dimensional Big Data using Principal Component Regression (PCR)

Kyi Lai Lai Khine[1] and Thi Thi Soe Nyunt [2 +]

[1] Cloud Computing Lab, University of Computer Studies, Yangon, Myanmar
[2] Faculty of Computer Science, University of Computer Studies, Yangon, Myanmar

Abstract. Nowadays, the increasing volume, complexity of formats and delivery speed of "Big Data" from diverse application domains have exceeded the capabilities of traditional data management tools and technologies. There is a need to re-design classical data analysis methods and algorithms to be adaptable in parallel and distributed architecture which can work well with the vast amounts of data not only in size of samples but also in number of dimensions. Moreover, high-dimensional big datasets have experienced many issues and challenges to handle huge collection of wide (dimensions) and tall (samples) data nature extracting useful value from it. Principal Component Analysis (PCA) is an important machine learning algorithm in dimensionality reduction for highly correlated large-scale data. In this system, we will apply PCA as selecting regressors for multiple linear regression model we called Principal Component Regression (PCR) for high-dimensional big data analytics with the aim to select effective and efficient features or dimensions. Additionally, we will develop the parallel and distributed version of PCA as preliminary machine learning approach for multiple linear regression model implemented on two widely-used scalable and distributed platforms such as Disk-Based MapReduce and Memory-Based Spark solving the scalability issue of big data. Large-scale OpenStreetMap (OSM) data which can provide as reality fulfillment to GIS market and spatial world will be applied for experimentation of the system.

Keywords: Big Data, High-Dimensional, Principal Component Analysis, Multiple Linear Regression, Principal Component Regression, MapReduce, Spark, OpenStreetMap (OSM)

1. Introduction

Today, technology fast changing era, the terms "Big Data" and "Predictive Data Analytics" have been popular in statistics, data mining and machine learning environments. Generally, classical data analysis techniques always shoot a question about how to deal with extremely large datasets [1]. We have also discovered that traditional data analysis techniques are developed on small and moderate data or datasets. Thus, analyzing high-dimensional data and then predicting some insights from it become more and more difficult in big data era. We call "High-Dimensional" data in which one observation or record contains dimension in thousands or millions or more while only tens or hundreds of observations. While exponential increase in the size of data caused by a large number of dimensions, "Curse of Dimensionality" has become a big issue to address. Principal Component Analysis (PCA), a dimension reduction technique, has become a remedy to "Curse of Dimensionality" [2]. Sai Kiranmayee Samudrala, Jaroslaw Zola and et al. [2] explained existing traditional dimensionality reduction methods and techniques have encountered scalability issue because they don't scale well in datasets containing thousands of data points with millions of dimensions called high-dimensional data. They proposed a parallel framework for dimensionality reduction in large-scale datasets using distributed memory machines with MPI. Typically, there are two common distributed processing platforms, MapReduce and Spark for massive data analysis. MapReduce enables high-

[+] Thi Thi Soe Nyunt. Tel.: + (95-9-5321948); fax: +(013-610-633).
E-mail address: (thithisn@gmail.com).

performance with two main operations, “Map” and “Reduce” to be processed in parallel [3]. Spark, massive data processing framework, extends MapReduce consisting primitives for data sharing, named Resilient Distributed Datasets (RDDs). Computing PCA of a large-scale matrix Y of size N ×D (N rows and D columns), it can be obtained “d” principal components ($d \leq D$) that explains the most variance (information) of the data in matrix Y [4]. In big data era, it can be realized that traditional statistical, data mining and machine learning techniques are responsible to fulfil two essential properties such as scalability and flexibility adaptable with parallel and distributed processing. Moreover, classical dimensionality reduction machine learning algorithms were designed to work well with small-scale data [5,6]. Such algorithms are needed to transform into parallel, distributed and scalable version adaptable on scalable and distributed platforms. PCA will face scalability problem when it is applied to large-scale data although it is assumed a well-known dimensionality reduction model in data analysis [7]. Therefore, PCA is needed to develop on MapReduce and Spark to facilitate large-scale data analysis as they are increasing interest in distributed platforms to process big data [8,9]. Classical PCA techniques may also become a big problem for extremely high-dimensional setting. Young Kyung Lee and et al. [9] expressed that applying the standard and classical principal component analysis technique may become a big problem when it fails to offer consistent results in very high-dimensional setting. They proposed some modifications for standard PCA that can be worked well in high-dimensional data.

In fact, the multiple linear regression model is used to explain the linear relationship between one dependent variable and two or more independent variables in predictive data analysis. In spite of being successful in many applications, however, regression analysis can face serious difficulties when “Multicollinearity” exists among the independent variables (“regressors”). In multiple linear regression model, regressors are sometimes highly correlated with one another. Moreover, a dataset consists of extremely high dimensions (multiple regressors) for multiple linear regression analysis may degrade the predictive power of the model [10]. To handle the problem, variables or regressors selection techniques have become the essential matter improving the predictive power of regression model. PCA is also one of the remedial techniques which is mostly used for reducing the multiple dimensions associated to multiple linear regression which create new variables called the principal component (PCs) that are orthogonal and uncorrelated to each other. Zebin Wu, Yonglong Li, and et al. [11] discussed how to develop a parallel and distributed implementation of PCA on cloud computing architectures for dimensionality reduction of large-scale high-dimensional hyperspectral image dataset. PCA is implemented using Apache Hadoop MapReduce and Apache Spark for in-memory computing. Furthermore, vast amounts of raw, unstructured and spatially attributed data are continuously generated and available from OpenStreetMap (OSM). PCA on MapReduce and Spark will be developed as preliminary machine learning approach for predictive high-dimensional OSM data analysis in this system. Tonglin Zhang and Baijian Yang [12] also described about principal component analysis, a dimension reduction technique, has many issues and challenges in solving high-dimensional big matrix data. Moreover, they presented that large-scale standardized matrix computation in PCA. While PCA is being considered to be implemented on distributed platforms, Mohammad Hoque, Md. Rezaul Raju and et al. pointed that matrix-matrix multiplication is essential in solving many areas of science and engineering. Parallel matrix-matrix multiplication is still remaining as a research problem in distributed environments with higher number of processors.

The paper is organized as follows. Section 2 presents the background theory of principal component analysis and brief explanation about two parallel frameworks. OpenStreetMap (OSM) data for spatial data analysis and its interesting application areas are then expressed in Section 3. Our main experimentation of the PCA algorithm on two distributed platforms and respective experimental result explanations in detail are presented in Section 4. Finally, some discussions and conclusion are given in Section 5.

2. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) can explain a large number of input variables by a small number of principal component results with the purpose of interpreting and understanding easily. Therefore, PCA is often applied the first step of data analysis, which may be followed by linear regression, multiple linear regression, cluster analysis, image analysis, and many others. PCA also transforms a set of uncorrelated

linear combinations, called principal components from the observations of variables. In computing PCA, input matrix Y of size N ×D (N rows and D columns) can be obtained "d" principal components ($d \leq D$) that explains the most variance (information) of the data in this matrix Y. We have investigated that many classical PCA approaches are developed on small and moderate data or datasets. According to storage and computational barriers, PCA will be implemented on two popular distributed platforms, MapReduce and Spark to facilitate large-scale predictive data analysis. They are increasing interest in distributed platforms to process big data [11,12]. The parallel processing paradigm MapReduce enables high-performance processing among clusters of commodity computers. There are two main functions or operations, "Map" and "Reduce" in distributed manner for a job or a task to be processed in parallel. The input datasets can be organized as key/value pairs, and the "Map" function divides the main task into several subtasks to be operated in parallel manner and produces a set of intermediate pairs or results after processing upon these key/value pairs. The "Reduce" function takes the responsibility of processing all intermediate values concerned with intermediate key and then collecting all intermediate results for the main task [13]. Spark is newly developed for large-scale data processing framework which can implement in-memory cluster computing. It extends the MapReduce consisting primitives for data sharing, named Resilient Distributed Datasets (RDDs), and offers an API based on coarse-grained transformations for data recovery efficiently using lineage. Spark outperforms 100 times faster than MapReduce in memory, or 10 times faster on disk by providing advanced Directed Acyclic Graph (DAG) execution engine.

3. OpenStreetMap (OSM)

OpenStreetMap (OSM) [14] is a data resource available for open source, rights-free geographic data and information across the world. It is a collaborative mapping project started in 2004 and it was founded by Steve Coast with the purpose of initially focusing on mapping the United Kingdom. The ultimate objective is to create free, editable map of the world and nowadays it becomes the most comprehensive source of volunteered geographic information. The raw, unstructured large-scale OSM XML files uses a topological data format, with four main elements (also known as data primitives): nodes, ways, relations and tags. Many well-known applications and services collaborating with some kinds of geolocation or map-based component using OSM data are as follows: OpenStreetMap-based map for iPhoto for iOS and it has been cited a lot of sources for Apple's custom maps in iOS 6. Interactive data visualization products by Tableau software company has integrated OSM for all their mapping requirements. The professional robot simulator widely used for educational purposes, Webots applies OSM data to create virtual environment for autonomous vehicle simulations [14]. The raw, unstructured XML format large-scale OSM spatial dataset which can provide as reality fulfillment to GIS market and spatial world is applied in this system. Firstly, we have to pre-process this raw state OSM spatial data into suitable format to be processed with MapReduce and Spark platforms in the following figure 1 and the overall system architecture is shown in figure 2.

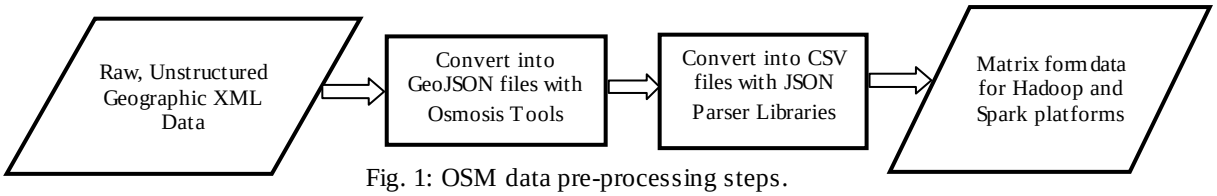

Fig. 1: OSM data pre-processing steps.

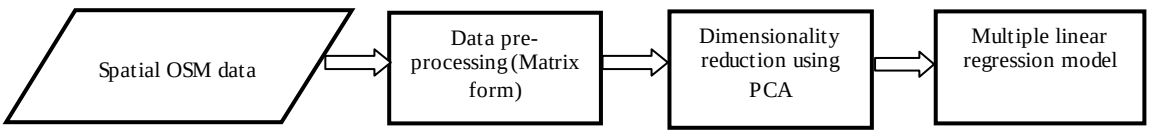

Fig. 2: The overall system architecture.

4. Experimentation

4.1. Experimental Setup

To verify the proposed parallel and distributed implementation of PCA, experiments were performed as follows: we run the experiments on Amazon Elastic Compute Cloud (Amazon EC2) which is a web service that provides resizable computing capacity and EMR (Elastic MapReduce) for creating a cluster of four Amazon EC2 m4. large instances, one for "Master" (Server node) and three "Slave" nodes. The cluster runs Linux Red Hat 4.6.3, and Amazon Hadoop Distribution 2.8.3 and Apache Spark 2.3.0 were installed on this cluster.

4.2. Experimental Results

In this system, PCA is applied with large-scale OSM spatial data in extracting valuable map information according to specific location is implemented on two distributed platforms. We intend to show that how PCA performs on these two platforms applying large-scale OSM data in extracting important features or dimensions for the regression model. We applied Apache Mahout machine learning library [15] in implementing MapReduce based PCA on top of Hadoop framework and MLlib machine learning library [16] for Spark based PCA implementation. According to the experimentation, we can describe the eigenvalues result of PCA obtained from serial and distributed versions by showing the comparative studies of PCA between the two versions. Top ten eigenvalues obtained from two versions of PCA can be seen in table 1 and the variance explained values of respective principal components are shown in table 2. Additionally, we can summarize that the effects of changing dimensions (increasing in number) can also vary in execution time. In each experiment, we apply the same number of rows or records only changing the number of columns and in the following figure 3, we measure execution time or running time in seconds for various dimensions in OSM spatial dataset. In figure 4, we'd like to show the analysis results of the proposed PCA on two distributed platforms: MapReduce and Spark, namely, PCA_MapReduce and PCA_Spark. The results show that the running time for both PCA algorithms are approximately close for small data size (in data records). For larger data size, however, PCA_Spark can offer faster running time than PCA_MapReduce. That is why; we can conclude that PCA_Spark outperforms better execution time when we scale into larger data size. Unlike PCA_Spark, the running time of PCA_MapReduce arises with a rate when we increase the size of input data records or rows which allows it to scale well.

Table. 1: Top ten eigenvalues obtained from serial and parallel and distributed versions of PCA.

No.	Serial Version	Parallel and Distributed Version
1.	-94.409438	-94.409438
2.	-613.041609	-613.041609
3.	-45.451175	-45.451175
4.	10.280624	10.280624
5.	127.537321	127.537321
6.	72.752940	72.752940
7.	107.646178	107.646178
8.	-68.007273	-68.007273
9.	78.034168	78.034168
10.	89.951034	89.951034

Table. 2: Total variance explained.

Principal Components		Initial Eigenvalues		
		Total	% of Variance	Cumulative % of Variance
	1.	-0.999804	0.090857	0.090857
	2.	-8.990541	0.817017	0.907874
	3.	1.111034	-0.100965	0.806909
	4.	-1.522418	0.1383501	0.945259

	5.	0.074319	-0.006753	0.938505
	6.	-0.598327	0.054373	0.992878
	7.	0.017288	-0.001571	0.991307
	8.	0.310082	-0.028178	0.963128
	9.	0.003321	-3.017159	0.962827
	10.	-0.158343	0.014389	0.977216

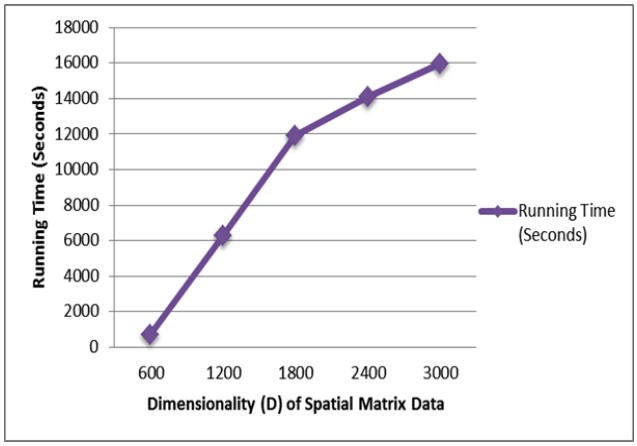

Fig. 3: Execution time upon OSM spatial dataset after varying the number of dimensions in each experiment

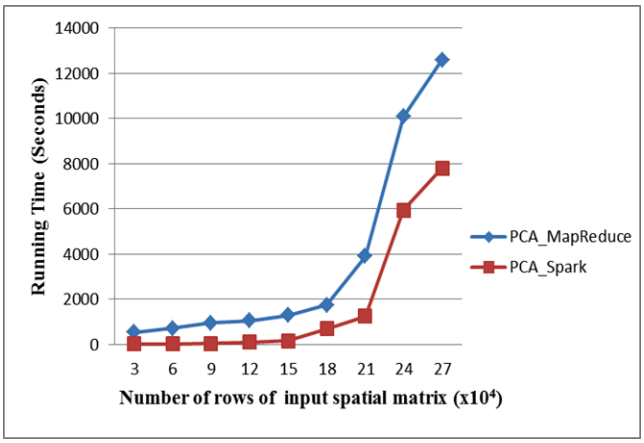

Fig. 4: The PCA on two distributed MapReduce and Spark platforms

5. Conclusion and Discussion

High-dimensional big datasets adversely affect the complexity of data analysis addressing high-dimensionality in big data become very important for constructing efficient statistical, data mining and machine learning models. The benefits that PCA can offer are not only exploring the patterns in data but also compressing the data by reducing the number of dimensions without losing information indeed. Moreover, classical dimensionality reduction machine learning algorithms are needed to redesign distributed and scalable version running on distributed platforms. In this system, PCA is implemented on both MapReduce and Spark platforms as preliminary machine learning approach for future predictive analysis. We intend to show that how PCA performs on these two platforms applying large-scale OSM data in extracting important features or dimensions for the regression model. The experimental results prove that the PCA_Spark outperforms than PCA_MapReduce in execution time. In future works, we will present the detail comparative studies of PCA_MapReduce and PCA_Spark applying diverse high-dimensional datasets in different domains measuring prediction accuracy of the multiple linear regression model.

6. References

[1] Amir, G., Murtaza, H.: Beyond the hype: Big data concepts, methods and analytics. In: 2014 International Journal of Management

[2] Aluru, S., Ganapathysubramanian, B., Zola, J., Samudrala, S.K.: Parallel Framework for Dimensionality Reduction of Large-Scale Datasets. In: Mechanical Engineering Publications, 2015

[3] Deng, S., Wu, W.: Efficient Matrix Multiplication. In Hadoop. In: International Journal of Computer Science and Applications, @ Technomathematics Research Foundation, Vol. 13, No. 1, pp. 93 – 104, 2016

[4] Elgamal, T., Hefeeda, M.: Analysis of PCA Algorithms in Distributed Environments, Qatar Computing Research Institute, April, 2015

[5] Hoque, M., Tymczak, C., Chilakamarri, K., Vrinceanu, D., Raju, M.R.: Parallel Sparse Matrix-Matrix Multiplication: A Scalable Solution with 1-D Algorithm. In: International Journal of Computational Science and Engineering, January, 2015

[6] Inouye, M., Abraham, G.: Fast Principal Component Analysis of Large-Scale Genome-Wide Data, Open-access article, 2014

[7] Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. In: Adaptive data analysis: theory and applications, January, 2016

[8] Kumar, A., Karampatziakis, N., Mineiro, P., Weimer, M., Narayanan, V.K.: Distributed and Scalable PCA in the Cloud, January, 2014

[9] Lee, Y.K., Park, B.U.: Principal Component Analysis in Very High-Dimensional Spaces, 2010

[10] Ul-Saufie, A. Z., A. S. Yahya, N. A. Ramli.: Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, Pulau Pinang. In: International Journal of Environmental Sciences, 2011

[11] Wu, Z., Li, Y., Plaza, A., Li, J.: Parallel and Distributed Dimensionality Reduction of Hyperspectral Data on Cloud Computing Architectures. In: IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, March, 2016

[12] Yang, B., Zhang, T.: Big Data Dimension Reduction using PCA. In: IEEE International Conference on Smart Cloud, 2016

[13] Zhu, J., Ge, Z., Song, Z.: Distributed Parallel PCA for Modeling and Monitoring of Large-scale Plant-wide Processes with Big Data. In: IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS

[14] http://www.openstreetmap.org

[15] https://mahout.apache.org

[16] https://spark.apache.org