Mesopotamian journal of Big Data
Vol. **(2021)**, 2021, **pp**. 14–20
DOI: https://doi.org/10.58496/MJBD/2021/003 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

Parallel Generalized Hebbian Algorithm for Large Scale Data Analytics

Mohanad G. Yaseen,*[1], , Mohammad Naeemullah[2] , , Ibarhim Adeb Mansoor[3],

[1] *Department of Computer, College of Education, AL-Iraqia University, Iraq*

[2] *Department of Computer Science Maulana Azad College. Rauza Bagh Aurangabad. Maharashtra, India*

[3] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia*

ARTICLE INFO

Article History

Received 14 Jan 2021
Accepted 20 Feb 2021
Published 22 Mar. 2021

Keywords

Big data

Parallel Processing
GHA

Machine Learning

ABSTRACT

In order to store and analyse large amounts of data on a parallel cluster, Big Data Systems such as Hadoop and DBMSs require a complex configuration and tuning procedure. This is mostly the result of static partitioning occurring whenever data sets are imported into the file system or transferred into it. Following that, parallel processing is carried out in a distributed fashion, with the objective of achieving balanced parallel execution among nodes. The system is notoriously difficult to configure, particularly in the areas of node synchronisation, data redistribution, and distributed caching in main memory. The extended Hebbian algorithm, abbreviated as GHA, is a linear feedforward neural network model for unsupervised learning that finds the majority of its applications in principle components analysis. Sanger's rule is another name for the GHA that may be found in the academic literature. Its formulation and stability, with the additional feature that it may be used to networks that have more than one output. A unique hardware architecture for principal component analysis is presented here in the form of a paper. The Generalized Hebbian Algorithm (GHA) was chosen as the foundation for the design because to the fact that it is both straightforward and efficient. The architecture may be broken down into three distinct parts: the memory unit, the weight vector updating unit, and the primary computing unit. Within the weight vector updating unit, the computation of various synaptic weight vectors uses the same circuit in order to cut down on the area expenses. This is done in order to save space. The GHA architecture incorporates a versatile multi-computer framework that is based on mpi. Therefore, GHA may be efficiently executed on platforms that utilise either sequential processing or parallel processing. When the data set is studied for a short period of time or when a dynamic number of virtual processors is selected at runtime, we predict that our architecture will be able to profit from parallel processing on the cloud. In this research, a parallel implementation of a variety of machine learning algorithms that are built on top of the MapReduce paradigm is presented with the purpose of improving processing speed and saving time.

1. INTRODUCTION

The term "big data"[1] refers to a relatively recent development that has garnered a lot of attention in recent years. It refers to the process of quickly capturing, discovering, and analyzing enormous volumes of data that come from a variety of complicated sources and types[2]. In its most general meaning, the term "big data" refers not only to the data in question but also to the methodology behind the processing and examination of that data. Extraction of principal components is a powerful statistical method that may be used for feature extraction, data reduction, and signal processing, among other applications. It is possible to iteratively extract primary eigenvectors in the real domain by utilizing the Generalized Hebbian Algorithm (GHA) that was developed by Sanger (1992)[3]. In some circumstances, such as the processing of the signal from a sensor array, we come across complicated data. In this study, the Complex-valued Generalized Hebbian Algorithm (CGHA), which was developed by [4], is given. It has been demonstrated that CGHA converges. A single-layer linear neural network may be used to implement CGHA, just as it was possible to do with GHA. The estimate of the Direction of Arrival (DOA) is used to show how CGHA may be used to the processing of signals coming from sensor arrays. Because of the exponential growth in the quantity of data that has been saved over the past several years, big data processing technologies are being created at a breakneck pace.

*Corresponding author. Email: maymy832410@gmail.com

Apache Spark is a parallel computing software framework that was built on the basis of Hadoop. Apache Spark[4] is excellent for large-scale iterative computing in a cloud computing software framework since it was developed on the basis of Hadoop. As a result of this, an analytical system is sought as a solution for the problem of managing applications that require a lot of data. With the advent of big data platforms, this should become obsolete. A hybrid infrastructure, which combines desktop grids and cloud computing, is endowed with a fair amount of cost and a large number of resources, and it is capable of giving the finest solutions as big data analytics. Hadoop[5] is a revolutionary solution that is specialised for in-memory processing, can carry out parallel processing on enormous datasets, and can also readily bond and interrelate with contemporary big data warehouses in environments with distributed storage. A MapReduce framework and a distributed file system are the two distinct components that are maintained by a Hadoop system. The process of using Hadoop systems for analytics begins with the loading of data files onto a distributed file system, followed by the execution of MapReduce calculations in parallel on the data. Recent developments have increased the necessity for the deployment of scalable machine learning algorithms on exceedingly large data sets.

These machine learning techniques are implemented in a distributed file system environment via the Hadoop MapReduce framework. When compared to hand-tuned implementations of individual learning algorithms, a machine learning method offers great performance, the ability to parallelize on a cluster of computers, and scalability over very big datasets. The use of high-level programming languages to script declarative machine learning algorithms frees the user from having to worry about low-level implementation and performance tweaking duties. An application of machine learning algorithms was suggested as a means of increasing the parallel speed at which multicore computers operate and the effectiveness of the use of time. In this work, it is hypothesised that these techniques may be used for single-core computers by leveraging Hadoop.

A recently developed enhanced EMD time-frequency analysis approach was proposed by Huang et al [5]. EMD was developed for the purpose of analysing nonlinear and nonstationary signals, as opposed to other methods of analysis such as Fourier transforms and wavelet decomposition. This is due to the fact that feature extraction through EMD is solely based on the properties extracted from the data, without taking into consideration the concept of stationarity. Any complex dataset may be broken down into a limited and frequently manageable number of intrinsic mode functions using the EMD approach (IMFs). The International Monetary Fund is comprehensive, responsive, and, to a considerable extent, orthogonal. The Hilbert transformation allows for the instantaneous frequency of a nonlinear and nonstationary signal to be calculated. This is possible due to the fact that IMFs may be regarded as single components. The EMD approach has been utilised in a variety of different applications, ranging from the study of rainfall to the identification of problems with robotic roller bearings. From its initial use in pattern recognition to domains such as failure detection of mechanical components, the support vector machine (SVM) has been getting a growing amount of attention in a variety of disciplines recently. In 1995, Corinna and Vapnik put up their proposal for the current iteration of the SVM.

The structural risk minimization (SRM) concept underpins the support vector machine (SVM), which stands in contrast to the more popular empirical risk minimization (ERM) premise. Because it is based on the SRM principle, the SVM often has superior generalisation performance compared to more traditional neural network approaches. The process of training a support vector machine is analogous to solving a linearly restricted quadratic programming problem. The training time scale, which is midway between quadratic and cubic scales with regard to the number of training samples, is one of the drawbacks, though. A parallel SVM[6] method that is built on the Spark platform has been offered as a solution to the issues of sluggish optimization and excessive memory consumption that arise while training the SVM algorithm with huge amounts of data. During the iterative processing, the MapReduce computing model must do many read and write operations to the disc, which slows down the learning rate. Scalability, dependability, and the capacity to evenly divide work are just some of the benefits that come with using Spark, a framework for distributed clusters that is built on memory. It is a form of parallel computing that can process large amounts of data quickly and effectively. The parallel SVM and the Spark computing platform are given a comprehensive examination in this paper. An effective parallel SVM technique is built here, and it's all thanks to data partitioning. The findings of the experiments indicate that the algorithm has the potential to hasten the convergence of the model, to increase the classification efficiency, and to be appropriate for the processing of huge amounts of data.

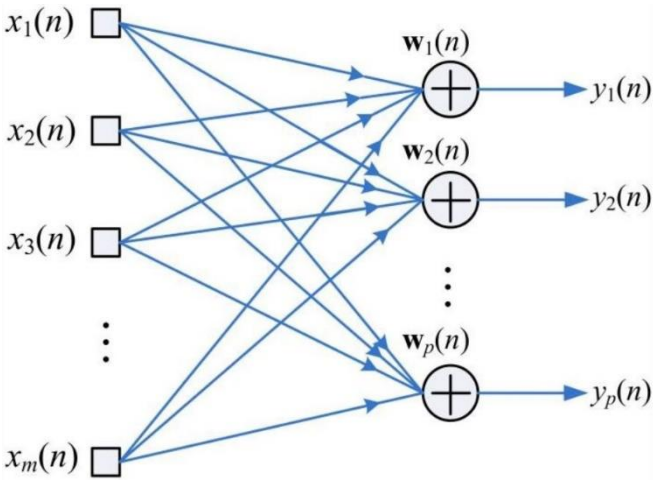

Fig.1. The neural model for the GHA.

2. THE HADOOP FRAMEWORK

Hadoop[7] is a software framework that runs applications on inexpensive hardware devices in huge clusters. It is an open source cloud platform software framework. Hadoop is able to achieve this because it offers a set of interfaces and data services for applications that are dependable, robust, and transparent. Hadoop is used to implement the MapReduce algorithm developed by Google. This method is capable of breaking down applications into many smaller units of work, each of which may be carried out or repeated on any cluster node. In addition to this, Hadoop offers a high throughput for both reading and writing data, as well as a distributed file system that can be used to store information on a number of different processing nodes. The use of MapReduce and distributed file systems inside the Hadoop framework allows for automated handling of failing nodes, which contributes to the framework's high degree of fault tolerance. Hadoop's main benefit is not just because it is open source, but also that it was designed to have a high scalability both in terms of data storage and processing. The high level of dependability that distributed processing possesses is ensured by the backup and recovery mechanism of the distributed file system HDFS as well as the job monitoring of Map/Reduce. Installation and operation of the Hadoop framework cluster, which is capable of processing vast amounts of data quickly and effectively, may be carried out on any standard personal computer. The architecture of Hadoop is seen in Figure(2).

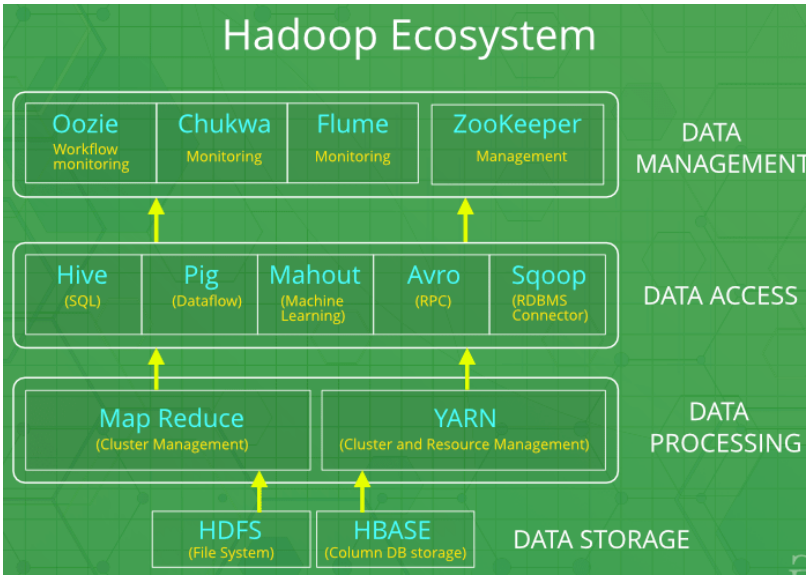

Fig.2. The architecture of Hadoop.

3. GENERALIZED HEBBIAN ALGORITHM

This part presents the Generalized Hebbian Algorithm[8] and demonstrates how the approach may be applied to the rectangular matrix form of singular value decomposition. The section concludes with a brief discussion of its applications. To do an Eigen decomposition, you need to provide an input that is a square matrix that is diagonally symmetric. This means that the cell value at row x, column y must be the same as the cell value at row y, column x. The correlation between data located in one space and other data located in the same space is the type of data that may be expressed by a matrix of this form. For instance, we could want to characterise the frequency with which a given word appears in conjunction with a specific other term. The relationships between the elements in the same area that are revealed by the data are consequently symmetrical; the phrase "word an appears with word b precisely as frequently as the phrase "word b occurs with word a." In singular value decomposition, matrices of input that are rectangular in shape are managed. One example of this is found in the form of ordered word bigrams; to visualise this, picture a matrix in which the rows represent the first word in a bigram and the columns represent the second word. The number of times that word b comes after word an is in no way comparable to the number of times that word an appears after word b. Both of these numbers are significantly different. Rows and columns are two distinct types of spaces; in a bigram, the row represents the space for the first word, while the column represents the space for the second word.

A. Machine Learning

The more all-encompassing field of artificial intelligence may be approached via the lens of machine learning. It comprises generating computations or algorithms that can get to data and learn from it without needing to be modified by a human and without having to obtain access to the data. It then performs an analysis on the data in order to identify the patterns that are included within it. The algorithms, in the long term, are able to make sensible decisions. enough to be able to apply what it has learnt to datasets that it has not previously utilised. Supervised learning, unsupervised learning, and reinforcement learning are the three forms of learning. There is intense rivalry in many fields, notably among shopping malls, to increase the number of customers they serve and, as a result, generate enormous profits. Machine learning is currently being utilised by a variety of stores and markets throughout the world in order to accomplish this objective.

B. Supervised and Unsupervised Learning

It is a method for imparting intelligence upon computers; the method is trained using data that has been labelled in accordance with a certain output. Unsupervised learning is another approach that may be used for the same purpose; however, in this method, we work with data that has not been labelled.

C. Classification

The process of classification is an example of supervised learning. This method requires labelled data in order to function properly, and it involves classifying the information into a variety of categories.

D. Regression

Through the use of regression, we are able to forecast the continuous result by basing our projections on one or more of the anticipated variables. In this, we employed a variety of methodologies, such as linear regression, logistic regression, polynomial, and others.

E. Clustering

Clustering is a form of learning that is done without supervision. During the clustering process, the unlabeled data are organized into categories. When we do this, we cluster the data points that are quite similar to one another.

F. Association rule learning

It is learning with no adult supervision. In order for it to be more productive, it examines each data item to determine whether or not it is dependent on other data items and then maps those dependencies. It makes an effort to unearth a few peculiar connections or links between the many components that make up the dataset. The if and then rules make use of it. For instance, when customers buy a phone at that particular point in time, the buyer also buys the back cover for the phone. The market basket analysis is one of the most important methods that large corporations use to discover connections between different kinds of products. It gives merchants the ability to identify links between the items that customers buy together as frequently as feasible.

4. PROPOSED METHOD AND RESULTS DISCUSSION

In this part of the article, we will provide an overview of the process of computing the data summaries and propose how we might compute a broad variety of machine learning models by utilising our data summaries. In the following sections, we will show the parallel processing features of our approach and go into further technical depth regarding how our system works. It is common knowledge that our suggested technique cannot be used to concurrently calculate all of the models, especially the iterative models; this is especially true for those models. The three-phase parallel design that we showed served as inspiration for our parallel solution. The entire dataset is loaded into memory as the first phase of the traditional GHA procedure. Please be aware that the data size must be within the limit that can be accommodated by the amount of

RAM that the computer possesses. Memoryup is a performance metric that evaluates the capacity of a parallel clustering method to make effective use of the accessible memory space on each node. It is feasible to compute the memoryup by altering the memory size of each node while maintaining the same dataset and number of nodes. This allows for the computation of the memoryup. The new GHA methodology makes use of the concept of data scanning by rows as its central organising principle. Even when the amount of data surpasses the capacity of the computer's memory, the GHA methodology may still be executed. When processing huge datasets, it is common practise to unnecessarily perform operations on duplicate and non-representative data, which can result in the loss of a sizeable portion of the CPU's time. Eliminating this category of data can frequently lead to a noticeable improvement in the speed at which processing is carried out. The storage and transmission of massive datasets are made much simpler when nonrepresentative data are removed from those datasets. This is an additional advantage of deleting nonrepresentative data from massive datasets. Using a variety of numerical examples, the benefits that the proposed new system offers in terms of computing were analysed. The computation was carried out using a computer equipped with a third-generation Intel Core i7 CPU running at 2.8 GHz and 16 gigabytes of DDR3 memory. The computer language that will be used to implement all of the suggested big data methods The GHA algorithm used C++. The comparison aims to evaluate the speed performance of the algorithm by performing an in-depth analysis of the PGHA's execution time by applying the Radoop technique and Parallel PCA. In the interest of this illustration, let us assume that the level of support fluctuates while the number of computer nodes remains same at 3.

Table I Baseline data sets description

Data set	d	n	Description	Model
CreditCard	30	285K	raise in credit line	NB
YearPredictionMSD	90	515K	rain or not	LR, PCA

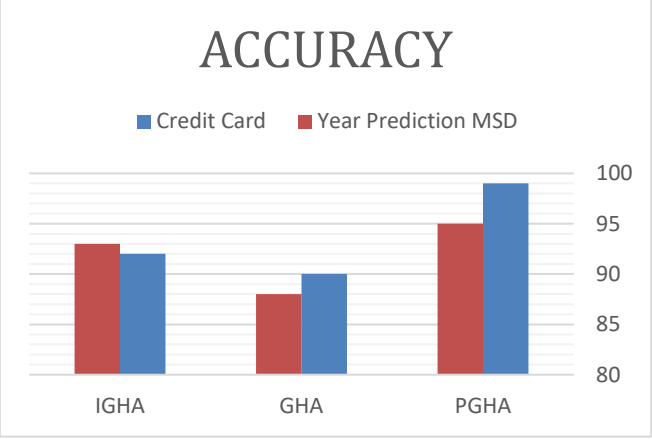

Fig.3. The Accuracy for Parallel GHA.

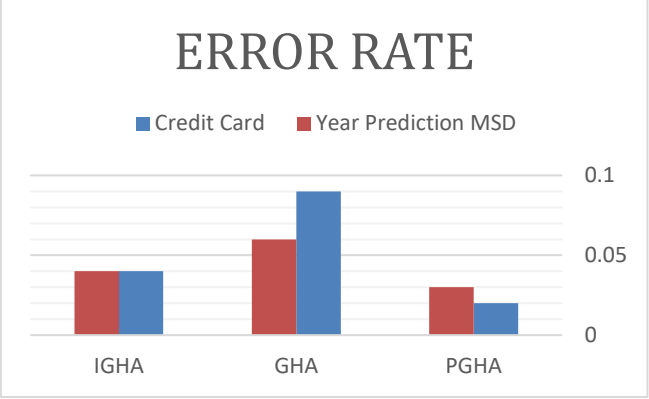

Fig.4. The Error Rate for Parallel GHA.

5. CONCLUSION AND FURTHER WORK

The suggested data processing system that is based on power big data provides a collection of solutions for the Machine Learning, which is the foundation of data mining and analysis. These solutions focus on the fusion of data and the cleaning of data. A unified data file storage format is suggested, as well as a solution for multi-source file formatting and file identification. This is based on the properties of big data. The suggested data cleaning approach makes use of several fundamental cleaning algorithms in order to address the issue of poor data quality that frequently arises in the context of big data. GHA usually contain millions of parameters, and it takes a significant quantity of data in order to adjust all of these factors so that they can accomplish a purpose (for example, classification of images into one of several possible classes). We are able to extract increasingly complicated representations as the size of the network and the data set both continue to expand; however, this comes at the expense of a ridiculously high amount of calculation time. On a CPU with a single core, training certain neural networks can take several weeks. In addition, the size of the neural network may be too large to be contained in a single computer. Because of this, investigating different techniques to parallelize the GHA network became a fascinating challenge. In a setup that is distributed, we may choose to implement Model Parallelism, in which case we would split the model across several computers, or Data Parallelism, and then we would examine the cost of computing as well as the cost of communication.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to express their gratitude to the Universiti Tun Hussein Onn Malaysia, the Informatics Institute for Postgraduate Studies, and the Al Salam University College for their moral support. Please accept my sincere gratitude for the useful recommendations and constructive remarks provided by the anonymous reviewers.

Funding

The authors receive no funding for this work.

REFERENCES

[1] A. K. Tripathi, K. Sharma, M. Bala, A. Kumar, V. G. Menon, and A. K. Bashir, "A parallel military-dog-based algorithm for clustering big data in cognitive industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2134-2142, 2020.

[2] L. Yao and Z. Ge, "Big data quality prediction in the process industry: A distributed parallel modeling framework," *Journal of Process Control*, vol. 68, pp. 1-13, 2018.

[3] I. Gemp, B. McWilliams, C. Vernade, and T. Graepel, "Eigengame unloaded: When playing games is better than optimizing," *arXiv preprint arXiv:2102.04152*, 2021.

[4] A. H. Ali, "A survey on vertical and horizontal scaling platforms for big data analytics," *International Journal of Integrated Engineering*, vol. 11, no. 6, pp. 138-150, 2019.

[5] A. H. Ali and M. Z. Abdullah, "Recent trends in distributed online stream processing platform for big data: Survey," in *2018 1st Annual International Conference on Information and Sciences (AiCIS)*, 2018, pp. 140-145: IEEE.

[6] R. Talib, "How we can use Energy Efficiency built upon the method of K-means clustering to extend the lifetime of WSN," *Al-Salam Journal for Engineering and Technology*, vol. 2, no. 1, pp. 40-45, 2023.

[7] Y. Li and D. Zhang, "Hadoop-Based University Ideological and Political Big Data Platform Design and Behavior Pattern Mining," in *2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI)*, 2020, pp. 47-51: IEEE.

[8] A. H. Ali, M. Aljanabi, and M. A. Ahmed, "Fuzzy generalized Hebbian algorithm for large-scale intrusion detection system," *International Journal of Integrated Engineering*, vol. 12, no. 1, pp. 81-90, 2020.