IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Parallel Computing Approaches for Dimensionality Reduction in the High-Dimensional Data

Siddheshwar V. Patil[1*], Dinesh B. Kulkarni[2]

[1] Department of Computer Science and Engineering, Walchand College of Engineering, Sangli, India
[2] Department of Information Technology, Walchand College of Engineering, Sangli, India

**Corresponding Author: siddheshwar.patil@walchandsangli.ac.in, Tel.: +91-9975646421*

DOI: https://doi.org/10.26438/ijcse/v7i5.17501755 | Available online at: www.ijcseonline.org

Accepted: 26/May/2019, Published: 31/May/2019

Abstract**—** The machine learning, as well as data mining techniques, deals with huge datasets. The numbers of dimensions (many features or instances) for these datasets are very large, which reduces performance (accuracy) of classification. The high dimensionality data models generally involve enormous data to be modeled and visualized for knowledge extraction which may require feature selection, classification, and prediction. Because of the high dimensionality of the datasets, it often consists of many redundant and irrelevant features which will grow the classification complexity while degrade the learning algorithm performance. Recent research focuses on improving accuracy by the way of dimension reduction techniques resulting in reducing computing time. So, it leads researchers to easily opt for parallel computing on high-performance computing (HPC) infrastructure. Parallel computing on multi-core and many-core architectures has evidenced to be important when searching for high-performance solutions. The general purpose graphics processing unit (GPGPU) has gained a very important place in the field of high-performance computing as a result of its low cost and massively data processing power. Also, parallel processing techniques achieve better speedup and scaleup. The popular dimensionality reduction methods are reviewed in this paper. These methods are Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Random Projection (RP), Auto-Encoder (AE), Multidimensional scaling (MDS), Non-negative Matrix Factorization (NMF), Locally Linear Embedding (LLE), Extreme Learning Machine (ELM) and Isometric Feature Mapping (Isomap). The objective of this paper is to present parallel computing approaches on multi-core and many-core architectures for solving dimensionality reduction problems in high dimensionality data.

Keywords**—** High-performance computing, Parallel computing, Dimensionality reduction, Classification, High-dimensionality data, Graphics processing unit

I. INTRODUCTION

Since the massive data volumes are generated at an extraordinary rate from different sources (e.g. marketing, financial, health, medical, social networks, government agencies etc), the big data growing faster in both the dimensionality and size [1], [2]. The survey says that, in libSVM database, in year 1990s, maximum dimensionality of dimensionality stood almost 62,000, then again increased in first decade of current century which is reached to 16 million. It is further increased in the current decade to almost 29 million. So, on this increased data, the present learning algorithms do not always operate in an appropriately suitable way. This becomes challenge for researchers to process extremely high dimensionality dataset. Still, Computational analysis of such high-dimensional information is a difficult problem; prod the development of various high-performance computational techniques. An important and rising category of strategies for handling such information is dimensionality reduction. The performance of the machine learning or any data mining algorithm depends on features (attributes) selected from dataset. High dimensional dataset reduces performance of the training algorithm since learning algorithm attempt to use all the features. Dimensionality reduction method is employed as the first stage to investigate and reduce massive information set. The prime objective is to find significant features and eliminate unnecessary from the high dimensional dataset. i.e. to decrease dimensions of the dataset, increase the accuracy of classification; decrease the computational cost. The prediction models are normally used for various intelligent and expert system applications including neural networks, multi-agency, knowledge discovery and management, data mining, text mining, multimedia mining, and fuzzy, genetic algorithms. So, the applications having maximum accuracy and less computational time are today's need. Due to the computational complexity (time and space), it is great mathematical challenge for traditional classification methods which works on high dimensionality. For such numerous computational problems, algorithms based on CPU are insufficient to provide a result in a required time span. Additionally, such cases become larger that CPU algorithms based on multi-cores will also inadequate to solve it. So, finally, there is a need of many-core architectures (GPGPUs) to solve such problems [3]. These requirements can be found in science and technology; natural sciences, physics, biology, chemistry, information technologies, geospatial information systems (GIS), structural mechanics, and computer science (CS) problems. Today, many of such problems are solved by enormous parallel processors. Also, for a program running on multiple parallel processors to predict the hypothetical maximum speed up, Amdahl's law is used.

II. Dimensionality reduction

The objective of any machine learning is to learn methods of input data mapping to desired outputs. If input data consist of noise and irrelevant attributes, classification of machine learning techniques will get affected. So, feature extraction and dimension reduction techniques are necessary to perform preprocessing on training data to achieve good results. Principal Component Analysis (PCA) [4], Non-negative Matrix Factorization (NMF) [5], Linear Discriminant Analysis (LDA) [4], Random Projection (RP) [6], AE [7][8][9], Multi-dimensional Scaling (MDS) [10] are some of the techniques. For dimensionality reduction, there exists unsupervised nonlinear Multidimensional Scaling based techniques like Locally Linear Embedding [11], Isomap [12], Extreme Learning Machine (ELM) [13] which achieved remarkable results for data. PCA algorithm [4] shows the mapping of original data model to diverse degrees. Then it learns the principal components (linearly un-correlated features) which define variance term of data. Dimensionality reduction is done by showing input given data through principal component subsets which will describe best variance of data. Attributes having minimum involvement to variances is measured as less descriptive. These attributes will be removed. Linear Discriminant Analysis (LDA) technique searches the vectors on underlying space. They must be discriminate among different classes. Furthermore, Linear Discriminant Analysis produces major mean differences among desired classes. It generates linear combination, if distinct features are available specific to subject. For samples from various classes, there can be two defined parameters in which first is within-class scatter matrix and next is between-class scatter matrix. Objective function is to decrease within-class distance measure whereas improve between-class measure. NMF [5] divides the input data to two parts. One is positive basis matrix while another is positive coefficient matrix. From data, the commonly occurring parts are learned by Positive basis matrix. To indicate degree in which commonly occurring parts reconstruct data, positive coefficient matrix is used. NMF accomplishes reduction of dimensions by selecting data along positive basis matrix. NMF retains data parts that are commonly occurring parts whereas deletes the parts of data that are rarely occurring parts. Another computationally efficient technique for dimensionality reduction is Random Projection (RP) [6] which preserves the Euclidean distance in the data points to reduced dimension space. Random Projection is attained by selecting data along sparse random matrix or orthogonal random matrix. The dimensionality reduction techniques based on neural network is Auto Encoding (AE) [7][8][9]. It has equal output and input data. When there are more input layer neurons than hidden layer neurons then Auto Encoding performs dimensionality reduction. Linear Auto Encoding is similar to PCA which learns variance information [4]. Nonlinear Auto-Encoder learns non-linear features. Multidimensional scaling (MDS) achieves a lower dimensional representation of data by keeping constant distances between data points [10]. The method is known as distance method. Distance is used as similarity or dissimilarity measure. Another nonlinear technique is LLE which calculates low-dimensional neighborhood while maintaining collection of high-dimensionality inputs. LLE removes requirement to evaluate pairwise distances between broadly separated data points. From the local linear fits, LLE recovers global nonlinear structure [11]. The nonlinear technique for dimensionality reduction is Isomap [12] takes geodesic interpoint distances as an alternative for Euclidean distances (EDs). Shortest paths with curved surface of manifold are obtained by using Geodesic distances. For complex natural observations, Isomap discovers nonlinear degrees of freedom. Extreme Learning Machine (ELM) based dimensionality reduction is efficient learning method for regression and classification [13]. It has number of dimensionality mapping functions such as Gaussian, sigmoid, multi-quadratic, Fourier series, wavelet etc. It has power to handle both large and small datasets in efficient way. For high-dimensional datasets (dimensions greater than 10), before using K-nearest neighbors algorithm (k-NN) usually, dimension reduction is performed to prevent from getting poor results [14].

III. Mathematical model

Relevant For dimension reduction, consider following observations.

Input vectors: $\boldsymbol{x}$
Dimension size: M
Number of observations: N
n^{th} observation: x_n
The entire dataset of observations: X
Since X is $M * N$ matrix, where complete observations will be taken as columns.

Conventions: a) Bold and lowercase letter ($\boldsymbol{x}$): vector
b) Non-bold and uppercase letter(X): matrices

The dimensionality reduction purpose is to get a new mapping χ of a reduced dimension m such that maximum information will be taken from the main observations set $\boldsymbol{x}$.
It can have conversion operation which maps the original vectors to new vectors. They are feature vectors.

$$\chi = Conversion(\boldsymbol{x})$$

Finally, projection of x_n is written as χ_n.
There is a method to recover a value which is close to original vector,

$$\boldsymbol{x'} = Recover(\chi) \text{ such that } \boldsymbol{x'} = \boldsymbol{x}$$

IV. LITERATURE REVIEW

E. Martel, E., R. Lazcano et al. [1] proposed a dimensionality reduction by Principal Component Analysis to raise the performance and effectiveness of wide hyperspectral image algorithms. They have shown the execution of the PCA on high-performance infrastructures such as Graphics Processing Units (GPUs) from NVIDIA and Kalray many-core. They have shown full utilization of the HPC infrastructure which will help to reduces the time needed for processing of a given hyperspectral images. Furthermore, the experimental results got on different datasets of hyperspectral images are compared to results obtained using PCA algorithm based field programmable gate array (FPGA). According to their results, dimensionality reduction by Principal Component Analysis on HPC devices outperforms the FPGA based method.

S. Ramirez-Gallego, I. Lastra et al. [2] have presented a method based on information theory which gives maximum relevance and minimum redundancy for dimensionality reduction. It gives high accuracy, so called to be the most relevant method for dimensionality reduction but having computationally expensive as it affects many features. Authors present fast-mRMR that minimizes this computational load. Along the fast-mRMR, authors showcased implementations of the same algorithm on variety of platforms, specifically, sequential execution on CPU, parallel computing on GPU (graphics processing unit), and distributed computing using apache spark. The results shows that parallel implementation (GPU) are best than sequential (CPU) in cases when there are large number of records (>100000 approximately). Results also shows that spark is the better than CPU version for high number of features (> 2000 roughly).

H. Kvinge, E. Farnell, M. Kirby, and C. Peterson [3] explained that the dimensionality-reduction methods are useful for extracting meaningful information in high dimensional datasets. They have used Secant-Avoidance Projection algorithm for data reduction. The objective is to avoid secant directions and also make a provision that dissimilar data points should not be mapped in reduced space. Such algorithms are based on constructive proof from Whitney's embedding theory which is a mathematical framework from differential topology. From the set of points, computing all (unit) secants is computationally expensive. So, GPUs are used to accomplish parallel forms of these algorithms. They have presented a data-reduction algorithm in polynomial-time which generates a significant representation in low-dimension of a data set.

J. Chen, Z. Tang [15] presents new method on the apache spark for big data called as parallel random forest. This algorithm combines task and data parallel techniques to get the optimization. In the data parallel technique, to minimize cost of data communication, the vertical data partitioning is used. Data multiplexing process is used to allow training data to be used further then reduce features from data. In task parallel technique, a twofold parallel technique is used in the training of the random forest, and a task directed acyclic graph is generated according to training of parallel random forest and dependency of Resilient Distributed Dataset (RDD) objects. To increase the accuracy for big, high-dimensional data, feature selection method is applied for training and a weighted voting in testing process earlier to parallelization. Result shows benefit of parallel random forest algorithm than Spark MLlib based algorithms.

Results are also good for classification accuracy, scalability and performance.

Result shows that, the performance of eigen decomposition based on GPU is 315 times more than conservative CPU dependent technique. Also, experiments on Hyperspectral Remote Sensing dataset (HSI) classification shows consistency of results using parallel code when compares with sequential approach. Authors suggest future work shall be on developing a dimension reduction toolbox using parallel methods for a unified HSI.

Y. Wang, A. Shrivastava et al. [16] presented a randomized algorithm for similarity search on CPU and GPU for datasets having ultra-high dimensions.

Authors evaluated the experiments on real-world benchmark high dimensional datasets such as malicious URL, social networks, text, click-through prediction, etc. The experiments focus on the difficulties deals with high dimensionality datasets consist of millions of features. Authors say that the current experimentations fail on the scale or much slower than the proposed algorithm.

This work is capable of calculating the k-NN graph roughly on the web-spam dataset (1.2 billion) in only 10 seconds. For the webspam data, calculating a whole k-NN graph in maximum 10 seconds will require minimum 20 teraflops.
Authors provided the CPU and GPU implementations for their algorithms.

Ramirez-Gallego et al. [17] said that the dimensionality reduction methods can be parallelized in apache spark which is a big data platform so as to accelerate the performance efficiency and accuracy. They proposed the information theory based implementation of a feature selection on a distributed architecture. The experimental work on real-world datasets showcase that the distributed framework is efficient to deal with ultrahigh-dimensional datasets also with a large sample size datasets. This algorithm beats the sequential algorithms in many cases they used. Authors suggested possible future research on high-speed data streams which focuses on developing a novel approach based on information theory. Analyzing the impact of approximate selection on high-dimensionality dataset by fast way which doesn't suffer accuracy. Another one is developing automatic system which filters most appropriate set of attributes (features), thus removing the requirement to provide the number of features (attributes) at every execution.

R. Jin, G. Chen et al. [18] presented a parallel version of an optimized iterative semantic compression method. This algorithm is particularly for data having large scale. Due to the enormous growth of data in size as well as in dimensions, data reduction has great research value and practical significance leads to use the compression technology. Considering the drawbacks of the semantic compression algorithm, Authors proposed a new method of bidirectional order selection based on interval partitioning. Authors proposed a parallel optimization iterative semantic compression algorithm on GPU as well as on spark to increase the speed of the algorithm. A number of experiments are performed on the large datasets, which also shows the efficiency of their proposed algorithm.

Current high-resolution Mass Spectrometry (MS) tools are generating thousands of spectra in only one systems experiment. In each spectrum, it contains thousand numbers of peaks, out of it; less number of peaks only helps to deduct the peptides. Therefore, preprocessing of such data to detect noisy and irrelevant peaks is ongoing research area. Many sequential noises reducing algorithms are unfeasible because of its high time-complexity.

M. Awan, F. Saeed [19] presented a dimensionality reduction algorithm based on GPU for MS2 spectra. It uses a new Binary Spectra as well as Quantized Indexed Spectra (QIS) data structure optimizes memory as well as GPU computational operations. The data structures used in the work helps to communicate within CPU and GPU by least possible data and enables for storing and processing complex data structure of 3-D to array structure of 1-D by keeping Mass Spectrometry integrity. Based on the size of input data, the proposed algorithm also considers sufficient memory of GPUs as well as switches among in-core as well as out-of-core modes. Result shows speed-up of almost 385X over the sequential algorithm. It shows the processing of million spectra in 32 seconds only.

K. Siddique, Z. Akhtar et al. [20] proposed parallel computing framework on bulk synchronous for big data applications. The work uses Apache Hama, a platform for distributed computing. It uses parallel bulk synchronous processing. Evaluation exposes the Hama`s potential and efficiency of problems arises in big data. Specifically, authors presented an evaluation of Hama's graph model. Authors also presented an apache giraph model using page rank method. Result describes apache Hama platform is superior to other one, giraph for parameters like computational speed and scalability. Still, the many issues continues to prevent the complete power of Hama which can be available for large-scale. The authors describe these challenges and solutions to solve them and focus further research possibilities.

T. Mingjie, Y. Yu et al. [21] proposed a parallel approach on high dimensional data for skyline query processing. For the multidimensional data points set available, the skyline query retrieves points which do not come under any other points from set. Because of pervasive usage of skyline queries, there are many challenges that are still not addressed. Authors introduced a novel efficient approach for running of skyline queries on a large-scale data. In this work, the data is firstly partitioned with z-order curve. To reduce dimensions of intermediate data, it uses a novel data partitioning. Secondly, each computation node partitions input data points into distinct sets. Further it implements parallel skyline computation to produce skyline candidates. At last, indexes are built and use the efficient method to combine the produced skyline candidates. Several implementations show that skyline code gains the performance improvement far better as compared to existing approaches.

K. Passi, A. Nour and C. Jain [22] explained the importance of dimensional reduction in machine learning techniques. It proposes a technique identify genes for analysis of microarray expression for clinical behavior on cancer datasets. One way to find significant gene by applying a Markov blanket algorithm. Authors compared performance of Markov blanket-based system with various wrapper based dimensionality (genes) reduction approaches on various microarray datasets. The wrapper-based approach depends on memetic algorithms. For Markov blanket, they have used minimum redundant maximum relevance dimensionality reduction optimized with genetic algorithms. They have done performance comparison of the Markov blanket model with the other classifier methods using same set of features. Performance of classification algorithms shows improved accuracy than other methods for cancer microarray datasets.

As the Hyperspectral data usually consists of redundancy which can be eliminated using dimensionality reduction. Z.

Wu, Y. Li [23] presented a technique for dimensionality reduction on cloud computing environments having proficient storage and preprocessing capacity of data. They created a parallel and distributed version of a PCA technique on cloud computing systems. The execution utilizes file system of Hadoop to understand distributed storage and uses Apache Spark. It uses parallel map-reduce model. So, they have shown benefit of both high throughput access and distributed computing abilities of cloud computing environments. They also showed optimization in traditional PCA which is finest for parallel and distributed computing. Further, it is implemented on actual cloud computing architecture. The result for the proposed parallel method on numerous Hyperspectral datasets shows high performance.

V. DATASETS

In the literature review covered in this paper, the experiments are executed using several real-world high dimensional datasets from UCI Machine learning repository and from other sources [17]. The classification performance of dimensionality reduction methods is calculated based on the dataset used for experiments. Few datasets contains small samples but higher dimensions, like brain tumor, colon, leukemia, prostate and lymphomas (DLBCL). Other datasets are epsilon, DNA, ECBDL14, URL; kddb consists of large samples with high dimensions. The Hyperspectral image dataset is available on lesun weebly site. The summary description of some high dimensionality datasets is shown in Table I.

Table 1- Summary of high-dimensional datasets

Data set	# samples	# features	#classes
Colon	62	2000	2
Brain tumor	50	10367	4
Prostate	102	1500	2
Lymphomas	77	5470	2
Epsilon	400000	2000	2
Dna	79739293	200	2
ECBDL14	65003913	630	2
url	1916904	3231961	2
kddb	19264097	29890095	2

VI. CONCLUSION

This paper gives insights of parallel computing techniques in resolving dimensionality reduction problems in high dimensional data. Furthermore, it elaborates important issues of high dimensionality problems. It explains dimensionality reduction model and also describes well-known techniques used for dimensionality reduction. The paper also elaborates parallel processing techniques on GPGPUs and distributed architecture like spark. From the state of art review, it shows high-performance computing approaches are best suitable for solving high dimensional data problems. This paper provides scope of HPC applicable to high dimensional data for new researchers. It motivates them to use the parallel processing techniques and computational power of various multi-core and many-core systems to accelerate the performance for solving high dimensional problems.

REFERENCES

[1] E. Martel, R. Lazcano, J. Lopez, D. Madronal et al., "*Implementation of the Principal Component Analysis onto High-Performance Computer Facilities for Hyperspectral Dimensionality Reduction: Results and Comparisons*", Remote Sens, Vol. **10**, issue. **6**, pp. **864**, **2018**.

[2] S. Ramirez-Gallego, I. Lastra et al.,"*Fast-mrmr: fast minimum redundancy maximum relevance algorithm for high-dimensional big data*", Int. J. Intell. Syst . Vol. **32**, Issue. **2**, pp. **134-152**, **2017**.

[3] H. Kvinge, E. Farnell, M. Kirby, and C. Peterson, "*A GPU-Oriented Algorithm Design for Secant-Based Dimensionality Reduction*", 17th IEEE International Symposium on Parallel and Distributed Computing (ISPDC), Geneva, pp. **69-76**, **2018**.

[4] H. Hotelling, "*Analysis of a complex of statistical variables into principal components*", J. Edu. Psychol., vol. **24**, Issue. **6**, pp. **417-441**, **1933**.

[5] D. Lee, H. Seung, "*Learning the parts of objects by non-negative matrix factorization*", Nature, vol. **401**, pp. **788-791**, **1999**.

[6] D. Achlioptas, "*Database-friendly random projections*", Proc. 20th Symp. Principles Database Syst., pp. **274-281** , **2001**.

[7] Y. Bengio, "*Learning deep architectures for AI*", Found. Trends Mach. Learn., vol. **2**, no. **1**, pp. **1-127**, **2009**.

[8] Y. Bengio, A. Courville, P. Vincent, "*Representation learning: A review and new perspectives*", IEEE Trans. Pattern Anal. Mach. Intell., vol. **35**, issue. **8**, pp. **1798-1828**, **2013**.

[9] M. Chen, Z. Xu et al., "*Marginalized denoising autoencoders for domain adaptation*", Proc. 29th Int. Conf. Mach. Learn., pp. **767-774**, **2012**.

[10] T. Cox, M. Cox, "*Multidimensional Scaling*", Handbook of Data Visualization. Springer Handbooks Comp.Statistics. Springer, Berlin, Heidelberg, pp. **316-341**, **2008**.

[11] S. Roweis, L. Saul, "*Nonlinear Dimensionality Reduction by Locally Linear Embedding*", Science, vol. **290**, pp. **2323-2326**, **2000**.

[12] J. Tenenbaum, V. de Silva, J. Langford, "*A Global Geometric Framework for Non- linear Dimensionality Reduction*", Science, vol. **290**, pp. **2319-2323**, **2000**.

[13] L. Kasun, Y. Yang, G. Huang and Z. Zhang, "*Dimension Reduction With Extreme Learning Machine*", IEEE Transactions on Image Processing, vol. **25**, no. **8**, pp. **3906-3918**, **2016**.

[14] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, "When is nearest neighbor" meaningful?. Database Theory - ICDT99, **217{235}**, **1999**.

[15] J. Chen et al., "*A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment*", IEEE Transactions on Parallel and Distributed Systems, vol. **28**,issue. **4**, pp. **919-933**, **2017**.

[16] Y. Wang, A. Shrivastava, J. Wang, and J. Ryu, "*Randomized Algorithms Accelerated over CPU-GPU for Ultra-High Dimensional Similarity Search*", ACM Proceedings of International Conference on Management of Data. pp. **889-903**, **2018**.

[17] S. Ramirez-Gallego et al., "*An Information Theory-Based Feature Selection Framework for Big Data Under Apache Spark*", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. **48**, Issue. **9**, pp. **1441-1453**, **2018**.

[18] R. Jin, G. Chen, Anthony K. H. Tung, Lidan Shou, and Beng Chin Ooi, "*An Optimized Iterative Semantic Compression Algorithm And Parallel Processing for Large Scale Data",* KSII

Transactions on Internet and Information Systems. Vol. **12**, issue. **6**, pp. **2761- 2781**, **2018**.

[19] M. Awan and F. Saeed, "*An Out-of-Core GPU based Dimensionality Reduction Algorithm for Big Mass Spectrometry Data and Its Application in Bottom-up Proteomics*", ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, New York, pp. **550-555**, **2017**.

[20] K. Siddique, Z. Akhtar et al., "*Apache Hama: An Emerging Bulk Synchronous Parallel Computing Framework for Big Data Applications*", IEEE Access. 4, pp. **8879-8887**, **2016**.

[21] T. Mingjie, Y. Yu et al., "*Efficient Parallel Skyline Query Processing for High-Dimensional Data*", IEEE Transactions on Knowledge and Data Engineering, Vol. **30**, issue. **10**, **2018**.

[22] K. Passi, A. Nour and C. K. Jain, "*Markov blanket: Efficient strategy for feature subset selection method for high dimensional microarray cancer datasets*", IEEE International Conference on Bioinformatics and Biomedicine, USA, 2017.

[23] Z. Wu, Y. Li et al., "*Parallel and Distributed Dimensionality Reduction of Hyperspectral Data on Cloud Computing Architectures*", IEEE Journal of Selected Topics, vol. **9**, issue. **6**, pp. **2270-2278**, **2016**.