B V RAJU COLLEGE

VISHNUPUR

BHIMAVARAM

FACULTY RESEARCH PUBLICATIONS

ACADAMIC YEAR 2019-20

INDEX

DIMENSIONALITY AND KNOWLEDGE REDUCTION IN MASSIVE PATIENT DATASETS USING ICA AND HDFS

K. B. V. Brahma Rao[1], P. Suresh Varma[2], R Krishnam Raju Indukuri[3], M. V. Rama Sundari[4]

[1]*Department of MCA B. V. Raju College Vishnu Campus Bhimavaram A. P, India*
[2]*Department of CSE Adikavi Nannaya University Rajamahendravaram A. P, India*
[3]*Department of MCA B. V. Raju College Vishnu Campus Bhimavaram A. P, India*
[4]*Department of CSE, SITE, Tadepalligudem, A. P, India*
[1]brahmaraokbv@gmail.com
[2]vermaps@yahoo.com
[3]Irk.bvrice@gmail.com
[4]mvramasundari@yahoo.co.in

Abstract

The central idea of Independent Component Analysis (ICA) is to reduce the dimensionality of dataset consisting of a large number of interrelated variables, while retaining the variation present in the dataset as much as possible. In this paper we use, ICA and HDFS based algorithm for massive patient datasets to achieve lossless data reduction and to acquire required knowledge. The experimental results demonstrate that the proposed ICA technique efficiently processes massive datasets, eliminates irrelevant data, reduces storage space for the data and also the computation time. The results of ICA are compared with those of the Principal Component Analysis (PCA) and normal method where the accuracy of ICA is found to be good with respect to the Receiver Operating Characteristic (ROC) curve technique, also known as Area Under the Curve (AUC) or Area Under Receiver Operating Curve (AUROC).

Keywords: *Big Data Analytics, Independent Component Analysis, HDFS Principal Component Analysis, Knowledge Reduction*

1. Introduction

The data is growing day by day in industry such as social websites, healthcare scientific areas etc., for the last ten years makes it difficult to store, manage and analyzing it either to make decisions or knowledge reduction. In order to deal with the data detonation and knowledge reduction, we develop a parallel large-scale knowledge reduction method using independent component analysis to acquire the required knowledge for analysis.

Patient informatics data needs to be store in an efficient manner and include a lot of attributes (Variables). Most of the tools crash when large data is stored in it. The proposed technique removes some needless data from a dataset by conserving its properties. The experimental results demonstrate the proposed technique that can efficiently process massive datasets with highly speedup the classification of data and mainly reduce the storage requirements. In all the experiments the introduced method based on independent component analysis is compared with normal method and principal component analysis. This is found to be better than above specified methods.

ICA test is used to select or extract relevant and specific information about attributes (variables) in large dataset. ICA looks for independent factors while PCA looks for uncorrelated factors. If the variables are independent, it means they are not dependent on other variables. If two variables are uncorrelated, it means there is no true relation between them.

Python is an object-oriented, interpreted and high-level programming language with dynamic semantics. Python is having high-level data structures which are combined with dynamic typing and dynamic binding to make a very attractive Rapid Application Development. This is also used for scripting or glue language to connect existing components together. Python is a simple and easy to learn syntax and reduces cost of program maintenance. Python consists modules and packages which encourages program modularity and reuse of code. The Python interpreter and its library are available in source and binary form with freeware for major platforms and can be freely distributed. There is no more compilation step and the edit-test-debug cycle is incredibly fast. It is easy to debug Python programs and a bug or bad input will never cause a segmentation fault. The debugger is written in Python itself, testifying to Python's contemplative power.

Discovering the dependency attributes is an important issue in data analysis. It is also required to find the partial dependency attributes because while some of the attributes are removed, we may loss the required data.

To find the Knowledge has become a new challenge using big data. The independent component analysis has been successfully used in data mining. Generally MapReduce technique has been using for big data analysis in the recent times. But it requires lot of resources and is somewhat difficult to analyse the data. If the data is structural, the ICA technique is very much useful to analysis and acquires the required knowledge. Large amounts of data are producing daily from various sources using sensors and computer related devices in different formats by industries and scientific community. The size of the data may be zetta or yotta byte. This data is processed by different applications and is used to convert the core data into the same format. To process this much of data, Google developed a software frame work is known as MapReduce. But Python is providing a rich set of packages to analyse the data.

A parallel method is improving the performance of data mining for the effective computation of approximation. The parallel method makes our approach more ideal for executing large scale data. Mining the big data and knowledge discovery is a new challenge in the current days because the volume of data growing is at an unmanageable rate. The Python packages are acknowledged much responsiveness from both scientific community and industry for its applicability in big data analysis.

The ICA algorithm assumes that the given variables are linear mixtures of some unknown variables. The latent variables are not dependent on other variables and they are also called the independent components of the observed data. The main advantage of ICA technique in data analysis is that it does not need any initial or supplementary information concerning data. The following figures show the difference between PCA and ICA.

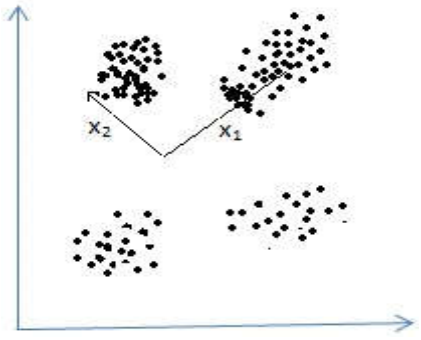

Figure: 1 Original Data Space x

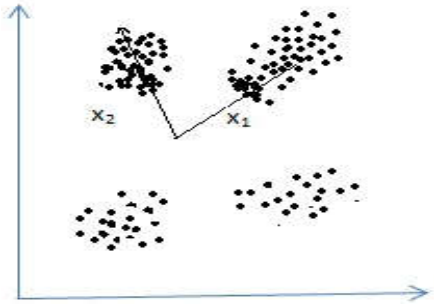

Figure: 2 Original Data Space x

The equation of PCA is x = WX.

Where i. x is the observations

ii. W is the mixing matrix

iii. X is the source or independent components

Now we have to find an un-mixing matrix i.e. the components become as independent as possible. The most common method to measure independence of components is Non-Gaussianity. As per the central limit theorem, distribution of the sum of independent components tends to be normally distributed is also known as Gaussian.

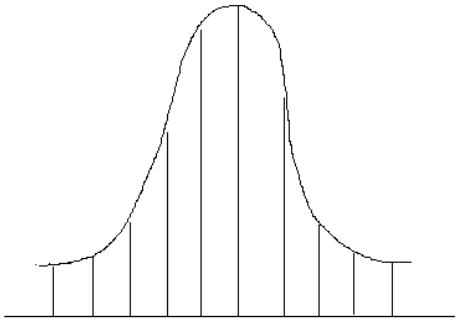

Figure 3. Central Limit Theorem

The common measure of shape is called the kurtosis. The mean and standard deviation have the same units as the original data and the variance has the square of those units. Though the kurtosis is similar to skewness has no more units. It is a pure number like z-score. Now we can look for the transformations that maximize the kurtosis of each component of the independent components. The third order moment of the distribution is called Kurtosis. Non-gaussian distribution will maximize the kurtosis. The following figure non-gaussian distribution (ICA) which in turn makes the components independent.

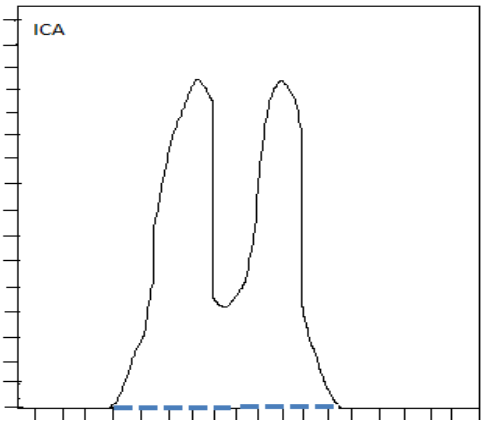

Figure 4. Non-gaussian Distribution

ICA has also provided the necessary formalism and thoughts for the development of some propositional machine learning systems. ICA technique has also been used for knowledge representation, dealing with imperfect data, reducing knowledge representation, data mining and for analyzing attribute dependencies. This technique has found many applications such as medical data, security analysis, power system, image processing, voice recognition and finance. This technique is also one of the

research areas that have successfully used for knowledge discovery or Data Mining in datasets.

The Hadoop Distribution File System (HDFS) is the primary data storage system used by Hadoop applications and its architecture consists NameNode and DataNode to implement distributed file system. The distributed file system provides high-performance access to data across highly scalable Hadoop clusters. The output of ICA algorithm is passed to HDFS and it executes using MapReduce.

The MapReduce is a programming model that consists two functions: Map and Reduce. This model operates on the big data sets with a parallel, distributed algorithm on a cluster. A map function performs filtering and sorting, whereas reduce function performs a summary operation.

To expand the application of ICA in the field of big data mining and deal with massive data sets is combined with HDFS to operate the big data sets in parallel and distributed to achieve dimensionality reduction and acquire required knowledge. And also eliminating redundancy, less memory space for storing data and reduce the execution time.

2. Related Work

Chaman Lal Sabharwal [1] used PCA based algorithms in two diverse genres, Qualitative Spatial Reasoning (QSR) to achieve data reduction along with improved regression analysis respectively. M. Usman Ali [2] proposed Principal Component Analysis (PCA) and Factor Analysis are used for dimensionality reduction of Bioinformatics data. These techniques were applied on Leukaemia data set and the number of attributes was reduced from to.

K. Keerthi Vasan [3] discussed on the efficiency of PCA for intrusion detection and determine its Reduction Ratio (RR), ideal number of Principal Components needed for intrusion detection and the impact of noisy data on PCA. H. Telgaonkar Archana [4] discussed well known techniques of Dimensionality Reduction namely Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Performance analysis is carried out on high dimensional data set UMIST, COIL and YALE which consists of images of objects and human faces. Classify the objects using KNN classifier and naive bayes classifier to compare performance of these techniques. Difference between supervised and unsupervised learning is also inferred using these results.

Nandakishore Kambhatla [5] developed a local linear approach to dimension reduction that provides accurate representations and is fast to compute. K. Keerthi Vasan [6] focused on the efficiency of PCA for intrusion detection and determine its Reduction Ratio (RR), ideal number of Principal Components needed for intrusion detection and the impact of noisy data on PCA.

Laurens van der Maaten [7] presented a review and systematic comparison of PCA and classical scaling techniques. The performances of the nonlinear techniques are investigated on artificial and natural tasks. The results of the experiments reveal that nonlinear techniques perform well on selected artificial tasks, but that this strong performance does not necessarily extend to real-world tasks. Alireza Sarveniazi [8] reviewed dimensionality reduction methods in detail the last and most new versions that extensively developed in the past decade.

G.N.Ramadevi [9] presented the study of the dimension reduction techniques and its applications in real life. The PCA (Principal Component Analysis) is one of the dimensionality reduction techniques of feature reduction algorithm to reduce the dimensionality of the dataset without losing the data. PCA can be applied on data before clustering will results more accurate and reduce the time substantially. PCA is used for data visualization and noise reduction. Steven H. Berguin [10] proposed a new method for dimensionality reduction is presented that scales as plog(p), where p is the number of design variables. It works by taking advantage of adjoin design

methods to compute the covariance matrix of the gradient. This information is then used with principal component analysis to develop a linear transformation that allows an aerodynamic optimization problem to be reformulated in an equivalent coordinate system of lower dimensionality.

Khaled Labib [11] presented a method for detecting Denial-of-Service attacks and Network Probe attacks using Principal Component Analysis as a multivariate statistical tool. They discussed the nature of these attacks, introduced Principal Component Analysis and merits of using it for detecting intrusions. Jianqing Fan [12] presented an overview of methodological and theoretical developments of PCA over the last decade, with focus on its applications to big data analytics. They discussed relationship between PCA and factor analysis as well as its applications to large covariance estimation and multiple testing.

Jiaying Weng [13] presented an overview of some classic and modern dimension reduction methods, followed by a discussion of how to use the transformed variables in the context of analyzing survey data. Lan Fu [14] demonstrated the Discrimination Analysis of Multivariate Statistical Analysis, Linear Dimensionality Reduction and Nonlinear Dimensionality Reduction Method under the circumstances of the wide range of applications of high-dimensional data.

Zebin Wu [15] developed a parallel and distributed implementation of a widely used technique for hyperspectral dimensionality reduction: principal component analysis (PCA), based on cloud computing architectures. M. Song [16] discussed the effect of applying dimensionality reduction (preprocessing) techniques on the performance of trace clustering. They used three popular feature transformation techniques; singular value decomposition (SVD), random projection (RP), and Principal Components Analysis (PCA), and the state-of-the art trace clustering in process mining.

Tonglin Zhang [17] presented a new PCA approach, where the basic idea is to use the technique of scanning data by rows. They demonstrated that their PCA approach can be applied even if the size of data is higher than the memory size of the computer. Sabharwal [18] presented PCA based algorithms in two diverse genres, qualitative spatial reasoning (QSR) to achieve lossless data reduction and health informatics to achieve data reduction along with improved regression analysis respectively.

Prof Rasendu Mishra [19] presented a survey of various dimensionality reduction techniques for reducing features sets in order to group documents effectively with less computational processing and time. Jan Kalina [20] discussed the challenges and principles of Big Data analysis in biomedicine.

Sonam Malik [21] proposed K-PCA based techniques that can reduce the dimensionality of data for aiding both understanding and classification. Gordana Ivosev [22] described Principal Component Variable Grouping (PCVG), an unsupervised, intuitive method that assigns a large number of variables to a smaller number of groups that can be more readily visualized and understood.

Fasong Wang [23] discussed the data mining problem with ICA. The data model of under-complete ICA in data mining is given and then gives the most popular ICA algorithm-Natural Gradient Algorithm (NGA). Several applications of data mining with ICA is considered, such as latent variable decompositions, multivariate time series analysis and prediction, text document data analysis, extracting hidden signals in satellite images, weather data mining and so on. Tonglin Zhang [24] proposed a PCA approach without the computation of principal components. This approach provides an exact solution to PCA for regression.

Sudeep Tanwar [25] used Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) techniques to perform DR over BD. They compared the performance of both techniques in terms of accuracy and mean square error (MSR). Kerstin Bunte [26] reviewed the basic principles of dimensionality

reduction and discussed some of the approaches that were published over the past years from the perspective of their application to big data.

Wei Li [27] presented a method to reduce the microRNA data of esophageal cancer patients by combined PCA and cluster analysis. Dan Feldman [28] presented a practical solution with performance guarantees to the problem of dimensionality reduction for very large scale sparse matrices.

Marco Cavallo [29] proposed a visual interaction framework to improve dimensionality reduction based exploratory data analysis. Aris Kosmpoulos [30] examined the computational feasibility of the most common dimensionality reduction method (Principal Component Analysis) for the chosen problem as well as the computational benefits that it provides for cascade classification and its effect on classification accuracy.

3. Proposed Work

In the review of literature a very little work has been found towards dimensionality and knowledge reduction by removing superfluous data. No attempt is found to design data reduction using independent component analysis technique. This technique is used on linear mixtures of some unknown latent variables. In this paper, we use ICA technique using Python packages and Hadoop open source software. Hadoop is open source software developed by Apache Software Foundation attempted for distributed storage and distributed processing of massive data on computer clusters. The commodity hardware is most sufficient in the clusters for processing big data sets. The structure of processing of big data sets using ICA and Hadoop Distributed File System (HDFS) is described in **[Figure-5]**.

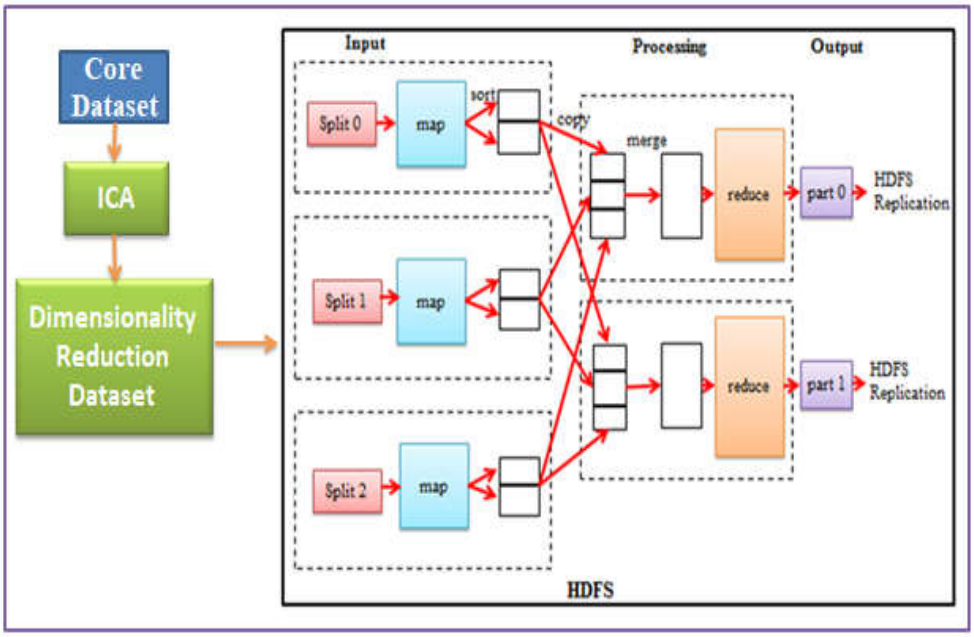

Figure 5. Structure of ICA and HDFS

Independent Component Analysis (ICA) is based on information theory and is also one of the most widely used dimensionality reduction techniques is used for big data sets. Python is providing a rich set of library packages for dimensionality reduction techniques. Hadoop Distributed File System (HDFS) is the very large storage system for dig datasets used by Hadoop applications. HDFS creates multiple replicas of data blocks and assigns them on data nodes, to enable reliable extremely rapid computations. Hadoop consists two most important modules. They are File Storage and Distributed Processing System. The first module is File Storage is also

known as "Hadoop Distributed File System (HDFS)". It is accountable for scalable, reliable, relatively low cost storage. The files are stored across a group of servers in HDFS and data availability is monitoring persistently in a cluster servers. The second module of Hadoop is the parallel data processing system is also known as "MapReduce". The Hadoop distributed file system and the MapReduce framework are successively on the same set of nodes. The Hadoop MapReduce programming allows the execution of Java code and also uses software written in other languages.

3.1 Basic Notions

We here introduce about notions of the Independent Component Analysis technique. The equation of ICA is X = AS,

Where

i. X is a random p-vector representing multivariate input measurements.

ii. S is a latent source p-vector whose components are independently distributed random variables.

iii. A is a p x p mixing matrix.

For the given realizations x_1, x_2, …., x_N of X, the goals of ICA are to

i. Estimate A.

ii. Estimate source distributions $S_j \sim f_{Sj}$, j = 1, 2, …., p.

If covariance matrix cov(X) = I and W is orthogonal and then

$$I(WX) = \sum_{j=1}^{p} \left(H(w_j^T)\right) \quad H(X)$$

Hence

$$\min_W I(WX) \leftrightarrow \min_W \{dependence\ between\ w_j^T X\}$$
$$\leftrightarrow \min_W \{the\ sum\ of\ the\ entropies of\ the\ w_j^T X\}$$
$$\leftrightarrow \max_W \{departures\ from\ Gaussianity\ of\ the\ w_j^T X\}$$

Many methods for ICA look for low-entropy or non-gaussian projections. The most common method to measure independence of components is Non-Gaussianity. Finally, the following two concepts are required to compute the independent components.

Negentropy: $J(Y_j) = H(Y_j) - H(Z_j)$, where Z_j is a Gaussian RV with same variable as Y_j. It measures the departures from Gaussianity.

FastICA: This uses simple approximations to negentropy.

$$J(w_j^T X) \approx \left[EG(w_j^T X) - EG(Z_j)\right]^2,$$

and with the data replaced expectations by a sample averages. They use

$$G(y) = \frac{1}{a} \log \cosh y;, 1 \leq a \leq 2$$

Dimensionality reduction algorithm based on ICA and knowledge reduction is mainly including 4 functions: the FastICA function (algorithm 1), the Map function (algorithm 2), the Reduce function (algorithm 3) and main function (algorithm 4).

Algorithm 1: FastICA (nc, rs)

// nc is a number of components (attributes) and rs is a random state

Input: Core Dataset

Output: Dimensionality reduction dataset

1. featurenames = coredataset.columns[:number_columns_required]
2. X = coredatset[featurenames]
3. Y = coredataset.Outcome // These values are either 0 or 1 based on condition
4. Split the data into training and testing (80% / 20 %) using train_test_split()
5. X = ICA.fit_transform(X_train)

// X is dataset feature names and X_train reduction data of core dataset

6. Reduction data is passed to map function.

Algorithm 2: Map (k, v)

Input: Dimensionality reduction dataset from algorithm1.

Output: Produces key and value pairs using predefined condition

1. For a ∈ C do // a is the attribute of dataset
2. Emit key and value pairs (k, v) based on zero and one values of attributes.

Algorithm 3: Reduce (String Cond_Attrib, pairs [$<c_1, n_1>$, $<c_2, n_2>$, …])

Input: Conditional attributes set and its corresponding value list.

Output: Produce different set of attributes using corresponding value.

1. For do

Classify the attributes into groups using the value received from algorithm2.

2. Emit groups of attributes with same value.

Algorithm 4: Main function

Input: Set of decision attributes.

Output: Produces expected output using predefined condition.

1. Result = Ø;
2. Compute conditional and decision attributes H(C | D).
3. Start a job,

{Execute algorithm2 and algorithm3, according to the result compare the value one attribute with remaining attributes and produces required knowledge into Result}

4. Output Result.

In order to deal with the data explosion and knowledge scarcity, we have developed a parallel large-scale knowledge reduction method based on ICA technique for knowledge acquisition using MapReduce for massive patient datasets in this paper. It constructs the parallel algorithm framework model for knowledge reduction using MapReduce, which can be used to compute a reduction for the algorithms based on independent components using logistic regression. The

proposed method enables dimensionality and knowledge reduction algorithm to be applied over massive datasets reduction problem without significant accuracy loss. The experimental results demonstrate that the proposed parallel knowledge reduction method can efficiently process massive datasets using Python packages and Hadoop platform, which highly speed up the grouping process and largely reduce the storage requirements. In all the experiments the proposed method is compared with the normal, PCA and ICA techniques.

Our technique will reduce the utilization of memory and processing time. The superfluous data is removed without important accuracy loss using type of disease. In this paper we have presented theoretical and experimental approach for dimensionality and knowledge reduction from large patient datasets using independent component analysis. The comparison clearly shows that the former method outperforms the latter one.

4. Results and Discussions

In this section, we propose to examine the efficiency of using ICA and MapReduce for big datasets dimensionality and knowledge reduction, as embodied by computing attribute importance and performing parallel search. Section 4.1 describes the datasets used to evaluate the method. Section 4.2 shows the details of hardware and software used in these experiments. Section 4.3 presents and discusses the experimental results of three different algorithms achieved.

4.1. Datasets

We have been studying on the analysis of patient datasets for the last four years. Patient datasets is available to us and the experiment is very meaningful. So we have selected patient big datasets for this experiment. Then, we regard a patient big datasets as a patient knowledge representation system and analysis the specific condition attributes of decision attributes. This experiment can find out the important influence attributes which affect heart attack or kidneys failure. The patient dataset decision table contains 33 attributes and 2 decision attributes. The condition attributes are the influence factors of heart attack or kidneys failure. The decision attributes are heart attack and kidneys failure. The purpose is to remove the irrelevant attributes and confirm the attributes more important. Table 1 shows a decision table of patient heart attack and kidneys failure.

Table 1. A decision table of patient dataset disease information

S. No.	Condition Attributes					Decision Attributes	
	Patient Type	Disease Type	Early Signs	Heart Cough	..	Heart Attack	Kidneys Failure
1	0	0	0	0	..	0	0
2	1	1	1	1	..	1	1
3	0	2	2	0	..	2	2
:	:	:	:	:	:	:	:

Details of Attributes:

1. Patient Type: 0 – In Patient, 1 – Out Patient
2. Disease Type: 0 – Cardiomyopathies, 1 - Coronary Artery, 2 – Diabetes, 3 - Heart Valves, 4 - Heart Defects present at Birth, 5 - High Blood Pressure, 6 - Lung Disease such as Emphysema, 6 - Past Heart Attacks
3. Early Signs: 0 - Chest Discomfort. It's the most common sign of heart danger, 1 - Nausea, Indigestion, Heartburn, or Stomach Pain, 2 - Pain that Spreads to the Arm, 3 - Dizzy or Lightheaded, 4 - Throat or Jaw Pain, 5 - Get Exhausted Easily, 6 – Snoring, 7 – Sweating
4. Heart Cough: 0 – Yes, 1 – No
5. ….
6. Heart Attack: 0 - ST segment elevation myocardial infarction (STEMI), 1 - Non-ST segment elevation myocardial infarction (NSTEMI), 2 - coronary spasm, or unstable angina
7. Kidneys Failure: 0 – Normal, 1 – Partial Failure, 2 – Fully Failed

Though there are many factors that affect the heart attack or kidneys failure, we have selected certain available factors only. Different characteristic attributes have different dimensions. In this paper, we have quantified the attributes first, and then make the data dimensionless which produced Table 1 as the result.

4.2. Hardware and Software used

The experiments have been carried out on six nodes in a cluster. The master node and five compute nodes. Each one of these computer nodes has the following features:

Processors: Intel Core i3 3rd generation or above

Cores: 4 per processor (8 threads)

Network: One Gigabit Ethernet

Hard drive: 1 TB or above

RAM: 4 GB or above

The specific details of the software used are the following:

Python 3.0 or above

MapReduce implementation: Hadoop 2.6.0. MapReduce 1 runtime (Classic).

Cloudera's open-source Apache Hadoop distribution.

Maximum maps tasks: 33.

Maximum reducer tasks: 1.

Operating System: Windows 10 / Ubuntu 15 or above, Version 64 bits

Java SE Development Kit: JDK1.7 or above

4.3 Experimental Analysis

To evaluate the performance of the dimensionality and knowledge reduction algorithm we have considered measurements reduction of data size that effects utilization of memory. A series of experiments are conducted on the big dataset and compared the results among dimensionality and knowledge reduction algorithm using normal, PCA and ICA techniques.

Table 2. Performance metrics of Data Size and Reduction Data size using different dimensionality and knowledge reduction techniques

Data Size in MB	Normal Method	PCA Technique	ICA Technique	Reduction in MB	
				% of Difference w.r.t. Normal Method	% of Difference w.r.t. PCA Technique
70	70	66	65	7.14	1.52
80	80	75	73	8.75	2.67
90	90	87	84	6.67	3.45
100	100	91	88	12	3.3
110	110	94	90	18.18	4.26
120	120	99	94	21.67	5.05
130	130	104	98	24.62	5.77
140	140	108	101	27.86	6.48
150	150	112	104	30.67	7.14
160	160	116	106	33.75	8.62
170	170	119	107	37.06	10.08
180	180	125	109	39.44	12.8
190	190	128	111	41.58	13.28
200	200	131	113	43.5	13.74

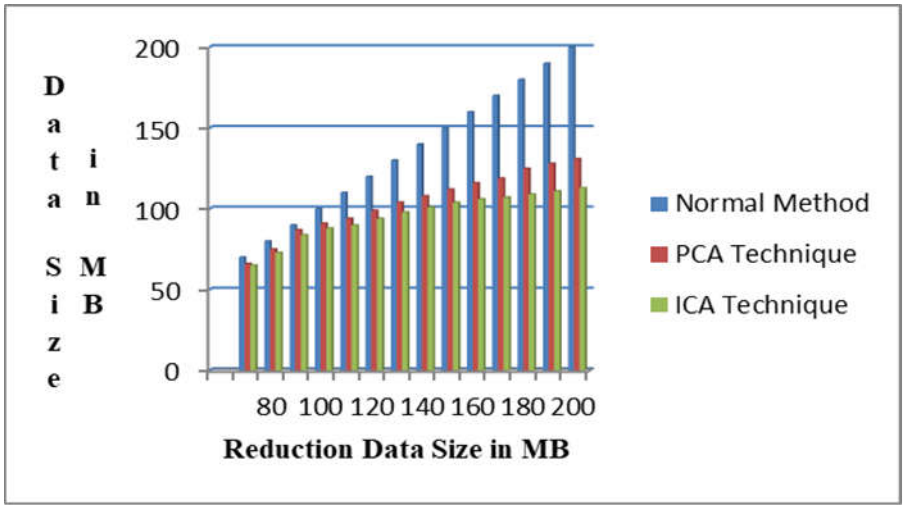

Figure 6. Comparison among three different techniques

The performance metrics of core data size and reduction data size results are interesting when the starting core data size is 70 MB onwards, the corresponding reduction data size is 65 MB that affects utilization of main memory. It shows that dimensionality and knowledge reduction data technique is giving better results than core data. When the data size is increasing, it shows that the dimensionality and knowledge reduction system using the Independent Component Analysis and MapReduce techniques are producing better results rather than Normal method and Principal Component technique.

5. Conclusion

In this paper, we proposed dimensionality and knowledge reduction method using Independent Component Analysis and MapReduce that can handle big datasets. The Hadoop MapReduce is an efficient computational model for distributed parallel processing with big data. The dimensionality and knowledge reduction algorithm is based on independent component analysis is successfully designed and is applied in the experiments. The experimental results demonstrate that the dimensionality and knowledge reduction algorithm using ICA and MapReduce can scale well and efficiently process big datasets using Python and Hadoop. The dimensionality and knowledge reduction algorithm based on independent component analysis can perform better than the knowledge reduction algorithm based on normal method or principal component analysis. Our future research work will focus on applications of the proposed parallel method in dimensionality and knowledge reduction using Independent Component Analysis and Hadoop Distribution File System.

6. Acknowledgements

This research was supported by B.V. Raju Foundation and Sri Vishnu Educational Society. We thank our colleagues from B.V. Raju College who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations of this paper.

We thank Prof. V. Bhaskara Murthy, Department of MCA, B.V. Raju College for assistance with ICA approach for comments that greatly improved the manuscript.

We would also like to show our gratitude to the Dr. Ch.V. Srinivas, Principal, B.V. Raju College for sharing their pearls of wisdom with us during the course of this research, and we thank the anonymous reviewers for their so-called insights.

References

7.1. Journal Article

[1] Chaman Lal Sabharwal, Bushra Anjum, “Data Reduction and Regression Using Principal Component Analysis in Qualitative Spatial Reasoning and Health Informatics”, Scielo., vol. 53, no. 53, (2016), pp. 31-42.

[2] M. Usman Ali, Shahzad Ahmed, Javed Ferzund, “Using PCA and Factor Analysis for Dimensionality Reduction of Bio-informatics Data”, IJACSA., vol. 8, no. 5, (2017), pp. 415-426.

[3] K. Keerthi Vasan, B. Surendiran, “Dimensionality reduction using Principal Component Analysis for network intrusiondetection”, Elsevier., vol. 8, (2016), pp. 510-512.

[4] H.Telgaonkar Archana, Deshmukh Sachin, “Dimensionality Reduction and Classification through PCA and LDA”, International Journal of Computer Applications., vol. 122, no. 17, (2015), pp. 33-37.

[5] Nandakishore Kambhatla, “Dimension Reduction by Local Principal Component Analysis”, ACM., vol. 9, (1997), pp. 1493-1516.

[6] K. Keerthi Vasan, B. Surendiran, “Dimensinality Reduction Using Principal Component Analysis for Network Intrusion Detection”, Science Direct, Elsevier., vol. 8, (2016), pp. 510–512.

[7] Laurens van der Maaten, Eric Postma, Jaap van den Herik, “Dimensionality Reduction: A Comparative Review”, Tilburg centre for Creative Computing, Tilburg University., (2009), pp. 1-35.

[8] Alireza Sarveniazi, “An Actual Survey of Dimensionality Reduction. American Journal of Computational Mathematics”, vol. 4, no. 2, (2014), pp. 55-72.

[9] G.N.Ramadevi, K.Usharani, “Study on Dimensionality Reduction Techniques and Applications”, International Journal Publications of Problems and Applications in Engineering Research (IJPAPER), vol. 4, no. 1, (2013), pp. 136-140.

[10] Steven H. Berguin, Dimitri N. Mavris, “Dimensionality Reduction Using Principal Component Analysis Applied to the Gradient”, AIAA Journal, vol. 53, no. 4, (2015), pp. 1078-1090.

[11] Khaled Labib, V. Rao Vemuri, “An Application of Principal Component Analysis to the Detection and Visualization of Computer Network Attacks”, Annals of Telecommunications, Springer, vol. 61, no. 1, (2004) Jan. pp. 1-14.

[12] Jianqing Fan, Qiang Sun, Wen-Xin Zhou, Ziwei Zhu, “Principal component analysis for big data”, Cornell University, vol. 1, (2018), pp. 1-20.

[13] Jiaying Weng, Derek S. Young, “Some dimension reduction strategies for the analysis of survey data”, Jouranl of Big Data, Springer, (2017), pp. 1-19.

[14] Lan Fu, "The Discriminate Analysis and Dimension Reduction Methods of High Dimension", Open Journal of Social Sciences, Scientific Research, vol. 3, no. 3, (2015), pp. 7-13.
[15] Zebin Wu, Yonglong Li, David E. Goldberg, Jun Li, Fu Xiao, Zhihui Wei, "Parallel and Distributed Dimensionality Reduction of Hyperspectral Data on Cloud Computing Architectures", IEEE, vol. 9, no. 6, (2016), pp. 2270-2278.
[16] M. Song, H. Yang, S.H. Siadat, M. Pechenizkiy, "A comparative study of dimensionality reduction techniques to enhance trace clustering performances", Expert Systems with Applications, Elsevier, vol. 40, no. 9, (2013), pp. 3722-3737.
[17] Tonglin Zhang, Baijian Yang, "Dimension reduction for big data. Statistics and Its Interface", vol. 11, no. 1, (2018), pp. 295-306.
[18] Chaman Lal Sabharwal, Bushra Anjum, "Data Reduction and Regression Using Principal Component Analysis in Qualitative Spatial Reasoning and Health Informatics", Research journal on Computer science and computer engineering with applications, Polibits, vol. 53, no. 53, (2016), pp. 31-42.
[19] Prof Rasendu Mishra, Dr Priti Sajja, "Experimental Survey of Various Dimensionality Reduction Techniques", International Journal of Pure and Applied Mathematics, vol. 119, no. 12, (2018), pp. 12569-12574.
[20] Jan Kalina, "Big Data, Biostatistics and Complexity Reduction", EJBI, (2018), pp. 24-32.
[21] Sonam Malik, Er. Pooja Narula, "Fast Dimensionality Reduction for High Dimensional atasets Supporting Big Data & Cloud Computing", International Journal of Computer Science And Technology (IJCST), vol. 6, no. 3, (2015), pp. 19-24.
[22] Gordana Ivosev, Lyle Burton, Ron Bonner, "Dimensionality Reduction and Visualization in Principal Component Analysis", Analytical Chemistry, vol. 80, no. 13, (2008), pp. 4933-4944.

7.3. Conference Proceedings

[23] Fasong Wang, Hongwei Li, Rui Li, "Data Mining with Independent Component Analysis", Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China, (2006) June 21-23.
[24] Tonglin Zhang, Baijian Yang, "Big Data Dimension Reduction using PCA", Proceedings of IEEE International Conference on Smart Cloud, New York, USA, (2016) November 18-20.
[25] Sudeep Tanwar, Tilak Ramani, Sudhanshu Tyagi, "Dimensionality Reduction Using PCA and SVD in Big Data: A Comparative Case Study", Proceedings of the International Conference on Future Internet Technologies and Trends, (2018) January 21-22.
[26] Kerstin Bunte, John Aldo Lee, "Unsupervised dimensionality reduction: the challenges of big data visualization", Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, (2015) April 22-24.
[27] Wei Li, Zhongyu Su, Wenshuang Li, Tianjia Liu, Pengcheng Guo, "MicroRNA data reduction of esophageal cancer", Proceedings of the First International Conference on Physics, Mathematics and Statistics, Shanghai, China, (2018) May 12-14.
[28] Dan Feldman, Mikhail Volkov, Daniela Rus, "Dimensionality Reduction of Massive Sparse Datasets Using Coresets", Proceedings of the International Conference on Neural Information Processing Systems (NIPS), (2016) December 21-22.
[29] Marco Cavallo, Cagatay Demiralp, "A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration", Proceedings of International Conference on Human Factors in Computing Systems, Montreal, QC, Canada, (2018) April 21-26.
[30] Aris Kosmpoulos, Georgios Paliouras, Ion Androutsopoulos, "The Effect of Dimensionality Reduction on Large Scale Hierarchical Classification", Proceedings of Springer International conference of the CLEF Initiative, (2014) September.