Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Reducing Dimensionality for Cloud Infrastructure with Kernel PCA

Swati*, Prof. Suresh Chand Gupta
Department of Computer Science and Engineering
Panipat Institute of Engineering and Technology, KUK
Samalkha, Haryana, India

Abstract— As we know data generated by various social activities is increasing day by day, analysis of such type of data is seems to be impossible but with cloud computing which works as analytic service provider ,it is going to be quite easy but before this user need to develop their applications to target cloud platform means an effort should be needed so that there is minimum cost for transforming data to clouds[1] So in this paper our goal is to reduce dimensions such that there is little variance or no variance so that we can minimize the cost of uploading data by removing irrelevant features.

Keywords— big data, cloud computing,dimensionalty reduction method, pca, kpca

I. INTRODUCTION

Actually big data is nothing, it means only the large amount of data that could not be processed by basic methods and tools .the main challenges that big data faces lying in searching ,sharing, analysis and visualization[1]. The biggest problem lie in data analysis because of scalability of data .so we need better analytics technique to deal with big data. But for analytics technique we need better computing infrastructure ,expensive software[2]. Cloud computing provide us not only better computing infrastructure but also analytic services like AaaS(Analytics as a service), MaaS(Model as a service). even than cloud computing faces many challenges like(even than we need human analyst for some application ,high estimated cost) and also we know big data contain most of information which are either not relevant or redundant so why we waste money to upload such kind of data. Infact user need to develop their application to target cloud platform so there is minimum cost.

Data excavating that is one of the most vital pace in vision invention procedure for data can be requested quickly on multimedia and hardware period to enhance worth of continuing data resources and can be consolidated alongside data that is held online so we can say it is extra than just data scrutiny but beforehand data is undeviatingly given to data excavating method to remove meaningful data a little feature reductions methods are requested out of that dimensional reduction is the method of changing elevated dimensional to low dimensional data

Although We have countless dimensional reduction method linear and non linear but there continue assorted methods to do so, PCA is by distant the most accepted (unsupervised) linear technique. Therefore, in our analogy, we merely contain PCA as a benchmark. in this paper we expresses the PCA computation merely in words of spot produce and consecutive exploits the kernel(non linear technique) mislead to implicitly compute the Elevated dimensional mapping. The choice of kernels is crucial:disparate kernels yield melodramatically disparate embeddings.

II. PREVIOUS WORK

CL Philip Chen et al., 2014[1], It is already true that Big Data has drawn huge attention from researchers in information sciences, policy and decision makers in governments and enterprises. However, there are so much potential and highly useful values hidden in the huge volume of data. A new scientific paradigm is born as data intensive scientific discovery (DISD), also known as Big Data problems. A large number of fields and sectors, ranging from economic and business activities to public administration, from national security to scientific researches in many areas, involve with Big Data problems. On the one hand, Big Data is extremely valuable to produce productivity in businesses and evolutionary breakthroughs in scientific disciplines, which give us a lot of opportunities to make great progresses in many fields. There is no doubt that the future competitions in business productivity and technologies will surely converge into the Big Data explorations.

Marcos D. Assun ,Rodrigo N. Calheiro, Silvia Bianchi, Marco A. S. Netto, Rajkumar Buyya,2014[2],Big Data analytics is still a challenging and time demanding task that requires expensive software, large computational infrastructure Cloud computing helps in alleviating these problems by providing resources on-demand with costs proportional to the actual usage. Furthermore, it enables infrastructures to be scaled up and down rapidly, adapting the system to the actual demand .Although Cloud infrastructure offers such elastic capacity to supply computational resources on demand, the area of Cloud-supported analytics is still in its early days.

Bouzas, D. et al, 2014 [3], In this paper, they counsel a novel algorithm for dimensionality reduction that uses as a criterion the public data (MI) amid the transformed data and their corresponding class labels. The MI is a influential criterion that can be utilized as a proxy to the Bayes error rate. Furthermore, present quadratic nonparametric implementations of MI are computationally effectual and do not need each prior assumptions concerning the class densities. They display that the quadratic nonparametric MI can be formulated as a kernel
goal in the graph embedding framework. Moreover, they counsel its linear equivalent as a novel linear dimensionality reduction algorithm.

Feixas, M. et al, in 2014 [4], Data Theory (IT) instruments, extensively utilized in countless logical fields such as engineering, physics, genetics, neuroscience, and countless others, are additionally functional transversal instruments in picture processing.

Izquierdo-Verdiguier, E. et al, 2014 [5], This paper presents a novel semisupervised kernel partial least squares (KPLS) algorithm for nonlinear feature extraction to tackle both landcover association and biophysical parameter retrieval problems. The counseled method finds protrusions of the early input data that align alongside the target variable (labels) and incorporates the wealth of unlabeled data to deal alongside low-sized or underrepresented data sets. The method relies on joining two kernel functions: the average radial-basis-function kernel established on labeled data and a generative, i.e., probabilistic, kernel undeviatingly learned by clustering the data countless periods and at disparate scales across the data manifold.

Jaime Zabalza et al., 2014 [6] In this paper as a widely used approach for feature extraction and data reduction, Principal Components Analysis (PCA) suffers from high computational cost, large memory requirement and low efficacy in dealing with large dimensional datasets such as Hyper spectral Imaging (HSI). To this end, a novel Folded-PCA is proposed, in which the spectral vector is folded into a matrix to allow the covariance matrix to be determined in a more efficient way.

L.J.P. van der Maaten,, E.O. Postma, H.J. van den Herik,2008[7], In recent years, a variety of nonlinear dimensionality reduction techniques have been proposed, many of which rely on the evaluation of local properties of the data. The paper presents a review and systematic comparison of these techniques. The performances of the techniques are investigated on artificial and natural tasks. The results of the experiments reveal that nonlinear techniques perform well on selected artificial tasks, but do not outperform the traditional PCA on real-world tasks. The paper explains these results by identifying weaknesses of current nonlinear techniques, and suggests how the performance of nonlinear dimensionality reduction techniques may be improved

III. PREVIOUS WORK

Advances in data collection and storage skills across the past decades have managed to an data overload in most areas. Most Researchers working in areas as varied as engineering, economics, and customer deals, need larger and larger simulations and observations were daily basis. Elevated dimensional datasets, in difference alongside tinier, extra established data sets have -been extensively studied.

In countless Contraption Discovering setbacks the reduction of the dimensionality of a input Matrix is an vital pace beforehand each scrunity of the data can be performed. The finished criterion for cutting the dimension is the desire to uphold most of the relevant data of the early data according to a little optimality criteria. Assorted methods such as PCA have been extensively utilized for cutting data. Main constituent scrunity (PCA), additionally recognized as Karhunen Loeve development, is a classical reduction and Feature Selection Methods, PCA finds a linear mapping amid the early features set to a lower dimensional feature set. The early pace of PCA is to change the matrix into vectors of elevated dimension and next compute the covariance matrix in high-dimensional vector space. As the Covariance matrix cannot be assessed for colossal size Matrices and the relatively small number of training samples. Furthermore, computing the Classification based on PCA is very time consuming. in this research work we have developed kernal based high dimensional image feature reduction method such that only relevant informations is preserved and there is little variance via mapping of high dimensional image into lower dimensional image.

IV. RESEARCH OBJECTIVE AND METHODOLOGY

Because kernal pca is an umbrella term actually to define various dimensionality reduction method have many application For every application we need to only modify kernal function acc to their need.and also we have several kenal function also inverse multiquardic ,integral operator kernel function,Gaussian kernel function. So we have used kernal pca because we have to update only kernal function and also complexity is not changed because we need to update only kernal function.here we calculate inner dot product by using gaussian distribution function to form kernal matrix.

in our work we have modified our kernal in such a way that it can minimise the dimension of high dimensional image by using mri scan and after that we can upload our reduced image to public or private cloud.

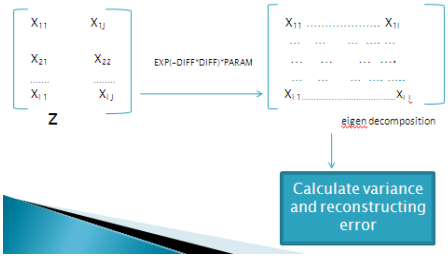

Fig. 1 flow chart

A. Kernel PCA:

Kernel PCA (KPCA) is the reformulation of established linear PCA in a high-dimensional space that is crafted employing a kernel function. In present years, the reformulation of linear methods employing the 'kernel trick' has managed to the proposition of prosperous methods such as kernel ridge regression and Prop Vector Machines. Kernel PCA computes the main eigenvectors of the kernel matrix, rather than those of the covariance matrix. The reformulation of PCA in kernel space is frank, as a kernel matrix is comparable to the in product of the datapoints in the elevated dimensional space that is crafted employing the kernel function. The request of PCA in the kernel space provides Kernel PCA the property of constructing nonlinear mappings.

B. Kernel Functions and Kernel Tricks:

The basic idea to deal with linearly inseparable data is to project it onto a higher dimensional space where it becomes linearly separable. Let us call this nonlinear mapping function ϕ so that the mapping of a sample x can be written as $x \rightarrow \phi(x)$, which is called "kernel function." Now, the term "kernel" describes a function that calculates the dot product of the images of the samples x under ϕ.

$$\kappa(xi,xj)=\phi(xi)T\phi(xj)$$

A (Gaussian) radial basis function (RBF) kernel can be used to map the data onto infinite dimensions. There are other popular kernels, e.g., polynomial kernels or sigmoid kernels, but the focus of this article will be on the RBF kernel, which is defined as

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp\left(- \frac{\|\boldsymbol{x_i} - \boldsymbol{x}_j\|_2^2}{2\sigma^2} \right)$$

where xi and xi are a pair of feature vectors and $\|x_i - x_j\|^2$ is the squared L2-norm (i.e., squared Euclidean distance) $\Sigma_k(x_{ik} - x_{jk})^2$ between the two sample vectors. Roughly speaking, the term "kernel" can be interpreted as a "similarity function" between a pair of samples. The minus sign inverts the distance measure into a similarity score, and due to the exponential term the resulting similarity score will range between 1 (for exactly similar samples) and 0 (for very dissimilar samples).

C. Steps of kernel PCA:

[in this we are given image as data,and sigma,numev,outlier all are given as a user input]

1. initialise k= zero[n,n] matrix and param=0.5/sigma*sigma
2. .repeat 3 and 4 for i=1 to n
3. repeat 4 for j=1 to n
4. create kernal matrix k[I,j] for each I and j by applying kernal function
5. krow=sum(k,1)/n[this will calculate sum over column]
6. .ksum=sum(krow)/n
7. repeat for each entry
8. for i=1 to n
9. for j=1 to n
10. k[I,j]=k[I,j]-krow[i]-krow[j]+ksum
11. [alpha,lambda]=eigs(K,numev,'lm',opts)[where alpha will be matrix of n*numev and lambda will be diagonal matrix of numev*numev]
12. resvar = (trace(K)-trace(lambda)) [where trace is sum of diagonal elements]
13. Calculate reconstruction error by recerr function

    we have deviced this algorithm for high dimensional images.

V. RESULTS AND ANALYSIS

We are employing matlab and we work in this paper alongside pca and kpca and will display aftermath by contrasting pca and kpca whereas pca utilized for linear makeover and kpca for non linear transformation.

A. PCA:

In pursuing figure we have seized a two dimensional data and made a plot employing MATLAB, the data consists of Elevated variance.

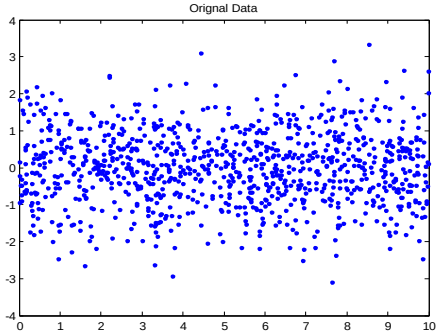

Fig. 2 original data

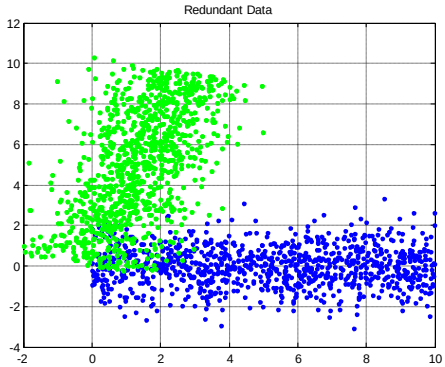

Fig. 3 redundant data after transformation

in figure below we have computed the Principal Component analysis of the given dataset, so that the noise and variance ( ) is reduced significantly.

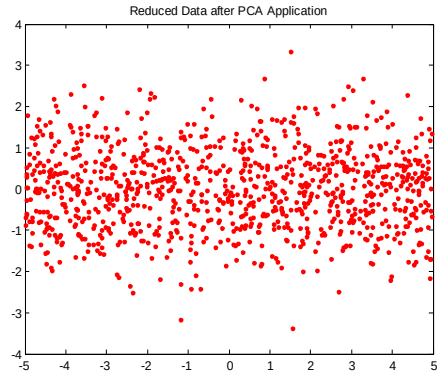

Fig. 4 reduced data after pca application

B. Kernel PCA:

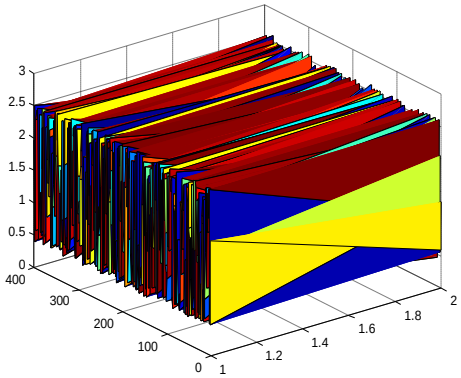

Fig. 5 three dimensional surface image

surf(Z) creates a three-dimensional shaded surface from the z components in matrix Z, using x = 1:n and y = 1:m, where [m,n] = size(Z). The height, Z, is a single-valued function defined over a geometrically rectangular grid. Z specifies the color data, as well as surface height, so color is proportional to surface height.

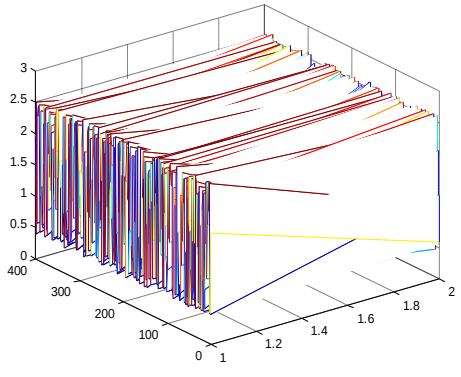

Fig. 6 wire frame structure

Reduced image with kpca and variance shown below:

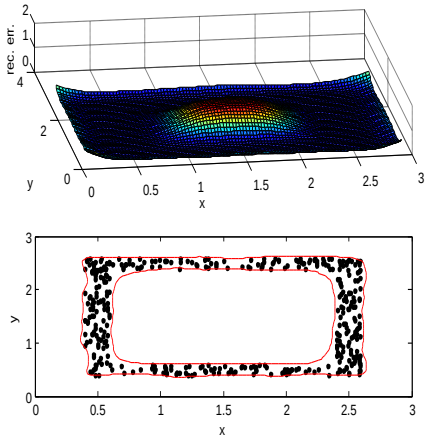

Fig 7 reduced image with kpca

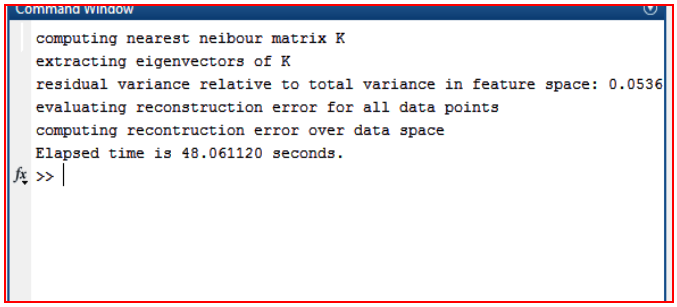

Fig 8 reduced variance

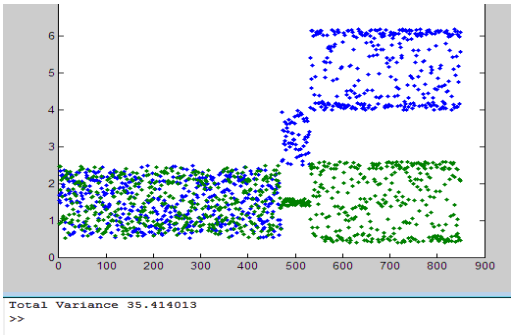

Fig 9 reduced variance with hdpca

In above fig we used high dimensional pca and we can saykernel pca is better.

VI. CONCLUSION AND FUTURE SCOPE

A. Conclusion:

Dimension reduction is the purpose of data to a lesser dimensional space such that uninformative variance in the data is discarded, or such that a subspace in that the data lifetimes is recognized. Dimension reduction has a long record as a method for data visualization, as well as for removing key low dimensional features (for example, the two-dimensional orientation of an object, from its elevated dimensional picture representation). The wanted low dimensional features depend on the task at hand in a little cases. Separately from teaching us concerning the data, dimension cut frequently leads us to larger sketches for inference. The necessity for dimension reduction additionally occurs for supplementary pressing factors. As we know big data have many opportunities and challenges .major challenge is data analysis. cloud computing provide better analytical services .but as we know big data contain most of raw data so before data is uploaded on cloud it need to be refined .so we use popular refining method called kpca As we know kpca is an umbrella, has many application because its kernal function is modified acc to the need of project .we implement kpca on high dimensional image and modified function in a way that high dimensional image map into low dimensional image so that there is no variance or little variance and as we said above we compared both and we see that with hdpca there is variance above 35 percent .so we are very confident about using kernel pca here.

B. Future work

As in my work I implement it only over images so even than it is area of interest that can we reduce all kind of data and 2nd I done only it in simulated way.obviously it is assumption based for real world environment it is still area of interest.

REFERENCES

[1] C.L. Philip Chen, Chun-Yang Zhang "Data intensive application and challanges" (2014).

[2] MarcosD.Assun,RodrigoN,.Calheiro,SilviaBianchi,Marco A. S. Netto, Rajkumar Buyya," big data and cloud (trends and direction)" (2014)

[3] Izquierdo-Verdiguier, Emma, Luis Gómez-Chova, Lorenzo Bruzzone, and Gustavo Camps-Valls. "Semisupervised kernel feature extraction for remote sensing image analysis." Geoscience and Remote Sensing, IEEE Transactions on 52, no. 9 (2014)

[4] Jaime Zabalza, Jinchang Ren, Mingqiang Yang, Yi Zhang, Jun Wang, Stephen Marshall, and Junwei Han. "Novel folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing." ISPRS Journal of Photogrammetry and Remote Sensing 93 (2014)

[5] Bouzas, Dimitrios, Nikolaos Arvanitopoulos, and Anastasios Tefas. "Graph Embedded Nonparametric Mutual Information For Supervised Dimensionality Reduction." (2014).

[6] Feixas, Miquel, Anton Bardera, Jaume Rigau, Qing Xu, and Mateu Sbert. "Information Theory Tools for Image Processing." Synthesis Lectures on Computer Graphics and Animation 6, no. 1 (2014)

[7] L.J.P. van der Maaten, E.O. Postma "review on dimensionality reduction methods" (2014)

