An Image Compression Algorithm Based on PCA and JL

Zhaodi Xiao[1, 2]
[1]School of Sciences, South China University of Technology, Guangzhou, China
[2]Foshan Power Supply Bureau, Guangdong Power Grid Corp, Foshan, China
Email: tibusong@foxmail.com

***Abstract*—Principal Components Analysis (PCA) is one of the most frequently used dimensionality reduction methods. PCA is suitable in time-critical case (i.e., when distance calculations involving only a few dimensions can be afforded) [1]. When it comes to image compression, PCA has its significant advantages: good performance in removing of correlations, and high compression ratio. Johnson-Lindenstrauss Lemma is a probability method leading to a deterministic statement of dimensionality reduction. This paper proposes a image compression algorithm: PCA for image compression based on improved Johnson-Lindenstrauss Lemma.**

***Index Terms*—PCA, johnson-lindenstrauss lemma, image compression, euclidean distances preserving, dimensionality reduction**

I. Introduction

With the development of science and technology, now we are living in an era of explosion of information, including a large number of images. The objective of image compression is to reduce irrelevance and redundancy of the image data to an acceptable level [2] in order to be store or transmit data in an efficient form [3].

In this paper, first of all, we revisit PCA and improved Johnson-Lindenstrauss Lemma respectively. After that, we will give a statement of the proposed dimensionality reduction method: a PCA method based on improved Johnson-Lindenstrauss Lemma. This method results in a minimum-k dimensional data set V of the original data set with a relatively high probability, with only $(1\pm\varepsilon)$ Euclidean Distance distortion.

II. An Improved Johnson-Lindenstrauss Lemma

The goal of Johnson Lindenstrauss Lemma is to, for a point set P in R^p, find a point set Q in R^k ($k < p$) and a mapping from P to Q [4]. Johnson Lindenstrauss Lemma has been applied to image processing. Since the features of images are represented as high dimensional vectors. With dimension reduction techniques, we can compress the vectors while the similarity between any two vectors is preserved. So we can carry out image analysis in lower space.

Manuscript received February 10, 2013; revised April 16, 2013.

The Johnson-Lindenstrauss Lemma is a fundamental result in dimensionality reduction that states that any m points in high-dimensional space can be mapped to a much lower dimension $k \geq \mathrm{O}(\frac{\log \mathrm{m}}{\varepsilon^2})$, without distorting pair wise distances between any two points by more than a factor of $(1+\varepsilon)$. In fact, such a mapping can be found in randomized polynomial time by projecting the high-dimensional points onto k -dimensional linear subspaces.

JAVIER ROJO AND TUAN S. NGUYEN (2010) [5] worked directly with the distributions of random distance rather than resorting to the moment generating function technique, an improvement on the lower bound for k is obtained.

Since the conclusion for lower bound for JL Lemma using L_2-L_1 norm in (JAVIER ROJO AND TUAN S. NGUYEN, 2010) (Improving Johnson-Lindenstrauss Lemma) only works for random matrix with i.i.d. entries drawn from Guassian distribution and one of the Achlioptas distributions (eq. q=1,2,3), this paper only discusses the case where L_2-L_2 norm is used.

A. Lower Bound for JL Lemma using L_2-L_2 norm

L_2-L_2 norm means that both of the distance measures of the original space and the target space are L_2-norm.

Let k be an even integer, and $0<\varepsilon<1$. Let $\lambda_1=k(1+\varepsilon)/2$ and $d=k/2$. Then the decreasing function in k is:

$$g(k,\varepsilon)=e^{-\lambda_1}\frac{\lambda_1^{d-1}}{(d-1)!}$$

For any $0<\varepsilon<1$ and integer n, let k be the smallest even integer satisfying $(\frac{1+\varepsilon}{\varepsilon})$ $g(k,\varepsilon)\leq\frac{1}{n^2}$. Then for any set V of n points in R^p, there is a linear map f: $R^p \to R^k$ such that for any $u,v\in V$

doi: 10.12720/joig.1.3.114-116

$$P[(1-\varepsilon)\|u-v\|^2 \le \|f(u)-f(v)\|^2 \le (1+\varepsilon)\|u-v\|^2] \ge 1-\frac{1}{n^2}$$

The lower bound of k can be obtained by finding the smallest even integer satisfying $(\frac{1+\varepsilon}{\varepsilon})\ g(k,\varepsilon) \le \frac{1}{n^2}$.

III. Principal Components Analysis (PCA)

Principal Components Analysis(PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal [6].

Assume that the original dataset X contains p dimensions and n observations and it is required to reduce the dimensionality into k dimensional subspace Y. This map is given by

$$Y=E^T X$$

where $E_{p\times k}$ is the projection matrix containing k eigenvectors corresponding to the k greatest eigenvalues.

IV. An Image Compression Algorithm: PCA based on Johnson-Lindenstrauss Lemma (JL-PCA)

When we apply PCA to reduce the dimensionality of the original high-dimensional data, it lacks a standard or a measure in terms of the projected dimensions (number of features of images). Johnson Lindenstrauss Lemma has its inherent advantages in Euclidean distances preserving. From the philosophic and dialectic point of view, an elaborately formed matrix is a special form of a randomly chosen transformation matrix which is the motivation of using Johnson-Lindenstrauss Lemma in the process of PCA. This paper proposed a similarity preserving algorithm for image compression, a PCA method based on Johnson-Lindenstrauss Lemma. Given the range of toleration of distortion of Euclidean distances after dimensionality reduction, we can find out the lower bound of the projected dimension, so that we can guarantee pairwise Euclidean distances preserving at a given probability. Images compressed by this algorithm are easier to be analyzed, especially for image search. For example, its computational cost will be reduced and similarity will be preserved when we carry out image search.

The lower bound of k for case using L_2-L_1 norm only holds for Gaussian random matrix and Achlioptas-typed random matrix. Obviously, projected matrix made up by selected eigenvectors is not random matrix. As a result, we can not apply conclusions of this case for PCA. Hence, this paper only considers PCA based on Johnson-Lindenstrauss Lemma (JL-PCA) using L_2-L_2 norm to compress images.

First of all, divide an image into segments. Each row acts as an input for ordinary PCA. And we assume that each block has the same components. Secondly, we will find the lower bound of k by applying improved Johnson-Lindenstrauss Lemma using L_2-L_2 norm. After that, we will calculate the eigenvalues of the original matrix $X \in R^{p\times n}$. Thirdly, we will select the top k eigenvalues and find out their corresponding eigenvectors. Then, form transform matrix E^T by combining the selected k eigenvectors. Finally, we can calculate the target matrix by $Y=E^T X$. Then we can get the principal components of the original matrix with only $(1+\varepsilon)$ at a relatively high probability.

Image Compression Algorithm : PCA-JL (L_2-L_2 norm)

Input: original image matrix $X \in R^{p\times n}$

Output: $Y \in R^{k\times n}$

Divide $X \in R^{p\times n}$ into several blocks.

Find out the lower bound of dimensions of projected dataset, k :

$0<\varepsilon<1$, Let k be an even integer, $\lambda_1=k(1+\varepsilon)/2$,

$$d=k/2,\ g(k,\varepsilon)=e^{-\lambda_1}\frac{\lambda_1^{d-1}}{(d-1)!}.$$

Calculate numerically to find k, the smallest even integer satisfying $(\frac{1+\varepsilon}{\varepsilon})\ g(k,\varepsilon) \le \frac{1}{n^2}$.

Calculate eigenvalues of $X \in R^{p\times n}$.

Find the projection matrix $E_{p\times k}$ by selecting k eigenvectors corresponding to the top k eigenvalues.

Calculate $Y=E^T X$.

V. Experiment

It carried out experiment on the platform of Matlab 7.0, using the 64 by 64 Lena image. It divided the original image into 256 blocks of 4 by 4 pixels. Each of the 256 blocks will be treated as an input of the PCA. Hence, each input has 16 dimensions. It set $\varepsilon=0.20$, $n=256$, $\lambda_1=k(1+\varepsilon)/2=0.6k$, $d=0.5k$,

$$g(k,\varepsilon)=e^{-\lambda_1}\frac{\lambda_1^{d-1}}{(d-1)!}=e^{-0.6k}\frac{(0.6k)^{0.5k-1}}{(0.5k-1)!}.$$

Then

$$(\frac{1+\varepsilon}{\varepsilon})\ g(k,\varepsilon)=6e^{-0.6k}\frac{(0.6k)^{0.5k-1}}{(0.5k-1)!}\le\frac{1}{n^2}=\frac{1}{256^2}.$$

After that, we have to find out k numerically.

And also, it carried out experiments on several groups of 64 by 64 pixels images that is cut from 512 by 512 pixels Lena image. It shows that this algorithm does not fully satisfy the improved lower bound of Johnson Lindenstrauss Lema (JAVIER ROJO AND TUAN S. NGUYEN, 2010), that is with a distortion not greater than ε under a probability of $(1-\frac{1}{n^2})$, but with a relatively high probability that is very close to this probability. This provides a way to determine how many principle components to retain when we are applying PCA to compress a group of images under certain distortion of pairwise Euclidean distances.

VI. Conclusions

The images compressed with Euclidean distances preserving characteristics benefit image analysis and image search.

For this algorithm, we have to carry out more experiments to verify it and improve it. And there is a drawback that we have to find out the lower bound of k numerically. These problems are what we are going to solve in the recent future.

It deserves some working to find a precise expression of the probability of this algorithm.

References

[1] S. Deegalla and H. Bostr¨om, "Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification," in *Proc. 5th International Conference on Machine Learning and Applications*, Orlando, FL, 2006.

[2] What is Image Compression? [Online]. Available: http://searchcio-midmarket.techtarget.com/definition/ image-compression

[3] Image Compression. [Online]. Available: http://en.wikipedia.org/wiki/Image_compression

[4] P. K. Agarwal (April 2007). Johnson Lindenstrauss Lemma. [Online]. Available: http://www.cs.duke.edu/courses/spring07/cps296.2/scribe_notes/lecture25.pdf

[5] J. Rojo and T. S. Nguyen, "Improving johnson-lindenstrauss lemma," 2010.

[6] Principal Component Analysis. [Online]. Available: http://en.wikipedia.org/wiki/Principal_component_analysis

Zhaodi Xiao, born in Shaoguan City, China, in 1986. She is a Master candidate on probability and mathematical statistics of South China University of Technology, Guangzhou, China, and is expected to receive her Master's Degree in July 2013. She received her Bachelor's Degree on computer science and technology (bilingual class) from South China University of Technology, Guangzhou, China in July 2009.

Currently, she also serves for Foshan Power Supply Bureau, Guangdong Power Grid Corp. as an IT engineer. She is mainly in charge of the developing and maintenance of information systems. Her research interests focus on machine learning, dimensionality reduction, image processing, software engineering, watermarks, and software automatic testing.