Project title:

The research and application of fast principle component analysis and singular value decomposition on huge data set

Contents:

Report info:

1. Preface

This project is started in 2009/10/1. Two undergraduate assistants are hired. One is Mr. Wei-Da Lai and the other is Mr. Yuan-Lin Su. Mr. Su is supper in Linux system management. He helped me to purchase our computer equipment and build up our server system. Because the budget of the computer equipment was not fully supported by NSC, we bought a notebook instead of the server. We indeed need a server to accomplish our project. So we cooperated with Prof. Rong-Nian Lai who is an assistant professor in Institute of Traditional Medicine, National Yang-Ming University. Prof. Lai provided us one HP server and a National Highway Institute (NHI) data set that include one million patients record from 1996 to 2007.

Because this project is “The research and application of fast principle component analysis and singular value decomposition on huge data set”, the NHI dataset satisfies the huge-data requirement. This data is the ten years clinical history of one million patients in Taiwan. Without compression the

data, it need about 270 Giga bytes HD capacity for storage. Hence, many statistical analysis methods with PCA, SVD, or eigenvalue computing will stuck for such huge data set. Our SC-SVD method works fine for these applications.

Another application is using our SC-PCA to Steganography problem. We try to find a security way to hide information into images of JPEG format. This is a research about security communication, such that people can hide a message into an image with JPEG format, then he can write an email with an attachment of this embedded image. When this email is transferred by the general email system, people who spy on this mail only see the normal message in the plain text. Nobody knows there is another message embedded in the JPEG image. The assistant Mr. Wei-Da Lai works fine in this problem. We use PCA method to estimate the robust region that message can hide into the JPEG image. Unlike the general steganography problem, our method can resist JPEG compression to 60% compression rate without missing embedded information.

2. Research purposes

The research purpose of this project is to develop a fast algorithm in singular value decomposition. Because singular value decomposition is cost in computation complexity, when data is increasing, the traditional SVD method becomes infeasible. So we are looking for a fast algorithm to make SVD is feasible for large dataset.

In our previous study, when the number of independent vectors is fewer then the square root of number of total vectors, we have a fast singular value decomposition method. If the number of independent vectors is close to the number of total vectors and singular values decay rapidly, we also have a fast approximation SVD algorithm.

If applications satisfy the previous cases, we can use fast SVD algorithm to improve computational cost. There are many applications in this case. We will focus on image steganography problem and NHI data analysis.

In steganography problem, JPEG image format is popular in digital images. However, there are few hiding techniques taking the JPEG image into account. Because JPEG format is a losing compression format, unlike BMP and Tiff format, data embedded in image might be destroyed when image be stored by JPEG compress format. To successful embed message into image which can avoid JPEG compression, we need to find the invariance space of JPEG processing. We will use PCA method to estimate this invariance space, and see what happens when we embed message into the image.

In NHI data analysis, there are many interesting problems for research. We focus on the asthma patients and look for whether there are some drugs rising asthma actively.

3. Literature reviews

For the length limit, we only show the literature review of the Steganography problem here.

Current methods for the embedding of data into the cover image fall into two categories: spatial-based methods (Adelson, 1990; Van Schyndel et al., 1994; Wang et al., 2001) and transform-based methods (Cox et al., 1997; Wolfgang et al., 1999; Xia et al., 1997). Spatial-based methods embed the data into the cover image directly, while transform-based methods embed the data into the cover image by modifying the coefficients of corresponding basis, for example wavelet basis of Fourier basis. Because JPEG image compression is design by the Discrete-Cosine Transform (DCT), we shall embed the message in the DCT domain.

The Watermark and Steganography are similar but in different purpose. The main goal of watermarking is to embed information into image that cannot remove the information from image. The main goal of steganography is to embed information into image that cannot be detected. Hence the recent work that embed message into JPEG image does not consider the compression rate of JPEG image. However, the JPEG image is easily be compressed again such that the embed information will be destroy. The best way is to design a embed method between watermark and steganography, such that the cover image can resist the JPEG compression.

4. Methodology

We collected several natural images from Google search engine by the key words 'natural wall paper'. There are 100 images in our training set. For each image, we compress it by JPEG quality parameter from 100 to 60. Then for each compressed image, we apply 2D DCT on the image and get its DCT base coefficient, say ϕ_j^i, where i is the *i-th* image, and j is the compression quality parameter. Let the 2D DCT of the uncompressed image be ϕ^i. The difference between ϕ_j^i and ϕ^i can be consider the perturbation by JPEG compression, we denote $\delta_j^i = \phi_j^i - \phi^i$. Let matrix A is formed by the collection of δ_j^i, for $i = 1,\cdots,100$ and $j = 60,\cdots,100$. The PCA of matrix A can estimate the main perturbation directions reduced by JPEG compression. If we embed message into the perpendicular direction, the message may preserved by JPEG compress.

The size of A depends on the number of training images and the size of significant DCT coefficient. For example, if we have 100 training images, then there are 4100 columns in matrix A; if the number of significant DCT coefficient is located in the 64-by-64 upper left corner, then there are 4096 rows in matrix A. Hence, we need the fast PCA algorithm to get the principal components.

When the perpendicular space is obtained, we can add the embedded message by the following algorithm. We first define the notation.

Let $R^n = V \oplus W$, $V = span\{\phi_i\}$, where ϕ_i is the principle component derived from the perturbation of JPEG compression and W is the perpendicular space of V in R^n. For every image X, let Y be the 2D DCT coefficient. We extract the most significant n elements as the carrier, denoted by η. Each $\eta \in R^n$ can be represented by $\eta = \phi + \psi$, where $\phi \in V$ and $\psi \in W$.

Our goal is to modify image X such that we can extract some meaningful information from X. We will add a vector $\nu \in R^n$ on η such that $\lfloor S_0^T(\eta + \nu) \rfloor_a = m$ is our message (the binary message in vector type), where S_0^T is the column matrix such that its column span space W and $\lfloor \bullet \rfloor_a \equiv \mathrm{mod}(\bullet / a, 2)$.

For given message m, the main ideas of solve minimize $\|\nu\|$ such that $\lfloor S_0^T(\eta + \nu) \rfloor_a = m$ is the following: (1) minimal $\|\nu\|$ will destroy image less; (2) the mod and floor function will control the robustness of embedded information when the image is compressed again; (3) The projection P_V makes the solution of ν occurs only when $\nu \in W$, that is the additional vector will resist the JPEG compression automatically.

Because $\eta = \phi + \psi$ and $\nu \in W$, let $\psi = \sum c_i s_i$ and $\nu = \sum d_i s_i$, we have $\lfloor S_0^T(\eta + \nu) \rfloor_a = \lfloor c + d \rfloor_a = m$. Let $c_i = n_i * a + b_i$. If $\mathrm{mod}(n_i, 2) = \mathrm{mod}(m_i, 2)$, then $d_i = -b_i$, else $d_i = a - b_i$. Then we solve ν.

5. Experimental results

For each image X, the corresponding S_0 generates the same column space. This property is suitable for commercialization. We only need to compute one S_0 from one image, and then this S_0 can be used by several images. If a pirate

wants to obtain the matrix S_0, he can only obtain $S_0^{'} = QS_0$, where Q is a unitary matrix. With knowing Q, the decode process only obtain the message $\lfloor Q^T(c+d) \rfloor_a$. Hence the pirate cannot reconstruct the correct message m. This steganography algorithm is self-security.

In our experiment, we successfully embedded message into JPEG image, the embedded image can resist JPEG compress from quality 100 to 60 and the decoding shows that there is no error at all. The following two images are the original image (cover image) and embedded image (stego image) respectively.

The original image

The embedded image, the message size is 1560 bits

6. Conclusions

We apply the SC-PCA method for large data analysis. Two fields are applied, one is NHI data analysis and the other is steganography problem. In NHI data analysis we found some drugs used in our NHI system are not suitable for Taiwan National. We will double check the result and publish this discovery.

In the application of steganography problem, we successfully design a method to embed information into JPEG image with both the requirement of steganography and watermark requirements. We will submit this result for publication too.

Reference:

[1] L. M. Marvel, Jr. C. G. Boncelet, and C.T. Retter. “Spread spectrum image steganography”. IEEE Trans. On Image Processing, 8(i):1075 1083, 1999.

[2] J. Zhao and E. Koch. “Embedding robust labels into images for copyright protection”. In Proceedings of International Conference on Intellectual Property Rights for Information, Knowledge and New Techniques, pages 242-251, 1995.