Engineering, 2013, 5, 268-271
http://dx.doi.org/10.4236/eng.2013.510B056 Published Online October 2013 (http://www.scirp.org/journal/eng)

Feature Extraction by Multi-Scale Principal Component Analysis and Classification in Spectral Domain

Shengkun Xie[1*], Anna T. Lawnizak[2], Pietro Lio'[3], Sridhar Krishnan[4]
[1]Global Management Studies, Ryerson University, Toronto, Canada
[2]Department of Mathematics and Statistics, University of Guelph, Guelph, Canada
[3]Computer Laboratory, University of Cambridge, Cambridge, UK
[4]Electrical and Computer Engineering, Ryerson University, Toronto, Canada
Email: [*]shengkun.xie@ryerson.ca

Received June 2013

ABSTRACT

Feature extraction of signals plays an important role in classification problems because of data dimension reduction property and potential improvement of a classification accuracy rate. Principal component analysis (PCA), wavelets transform or Fourier transform methods are often used for feature extraction. In this paper, we propose a multi-scale PCA, which combines discrete wavelet transform, and PCA for feature extraction of signals in both the spatial and temporal domains. Our study shows that the multi-scale PCA combined with the proposed new classification methods leads to high classification accuracy for the considered signals.

Keywords: Multi-Scale Principal Component Analysis; Discrete Wavelet Transform; Feature Extraction; Signal Classification; Empirical Classification

1. Introduction

The performance of a method used to recover the deterministic pattern is often impacted by stochastic correlations among noises of signals. As a common multivariate statistical method, the principal component analysis (PCA) [1] is often used for data dimension reduction and feature extraction of signals. The data features can be extracted by mapping signals onto a feature subspace that is spanned by only the first few principal components. However, the classical PCA may not perform well when it is applied to temporally correlated data or non-stationary data. When PCA is applied to these types of data, two following common problems are encountered. The first one is that PCA of measurements of stochastic processes is a single scale modeling approach, which means that local measurements do not vary with different underlying frequencies. Data from complex systems are often multi-scale and non-stationary in nature [2], therefore the conventional PCA often is not suitable for analyzing these typesof data. The second problem is that most of the data coming from complex systems are often temporally correlated.

Since PCA is based on the analysis of the data variance-covariance matrix, to avoid the potential undesirable effects on the outcomes of PCA caused by autocorrelation of the data, one may improve the data analysis by applying instead PCA in wavelet domain of the data, *i.e.* to the transformed data obtained by taking discrete wavelet transform (DWT), or wavelet packet transform, or stationary wavelet transform of the original data. The reason is that in wavelet domain the data have good decorrelation and localization properties [3]. The encountered problems of the conventional PCA when applying to auto-correlated or non-stationary data can be resolved by combining wavelet transforms with PCA because the wavelet coefficients of the data at each wavelet scale are approximately stationary and temporally uncorrelated ([4,5]). Additionally, due to the fact that in wavelet domain the significant features of the data can be extracted by a set of large values of wavelet and scaling coefficients and the fact that PCA can explain the large variations through few principal components, the combination of DWT with PCA can extract spatial and temporal data features simultaneously. In this paper, we illustrate the usefulness of DWT and PCA to the classification problem of a set of EEG data. We propose two methods, namely confidence interval classification and empirical classification, to classify the extracted features in the spectral domain.

2. Methods

2.1. Multi-Scale Principal Component Analysis

In the multi-scale PCA approach, first, signals are orga-

[*]Corresponding author.

nized into a data matrix, denoted by $\mathbf{\Phi}$, and DWT is then applied to $\mathbf{\Phi}$ (to each signal). After taking the DWT of Φ, PCA is applied at each level of the wavelet coefficients matrix. This procedure eliminates the principal components loading and their scores [1] that correspond to small eigenvalues and it reconstructs the wavelet coefficients by using the selected significant components and their associated scores at each level. The reconstructed signals are then obtained by taking PCA of wavelet approximation coefficients plus principal components of the wavelet detail coefficients at each level. Therefore, the wavelet coefficient matrix may be reconstructed by using selected wavelet coefficients and the final extracted data matrix can be obtained by taking inverse discrete wavelet transform (IDWT).

For a small number and less spatially correlated signals, the de-noising by dimension reduction method maynot be an appropriate choice as dimension reduction may cause severe loss of information of the signals. In this paper, we propose a method for de-nosing small number of signals. It is based on retaining only significantly large values of PC scores of wavelet coefficients. The proposed method re-calculates the PC scores at each level of wavelet details. Consider the following level j dependent regression model:

$$\boldsymbol{L}_j = \hat{\boldsymbol{L}}_j + \boldsymbol{e}_{\boldsymbol{L}} \quad (1)$$

where $\boldsymbol{L}_j$ are the PC scores at level j of wavelet details, $\hat{\boldsymbol{L}}_j$ is the estimate of $\boldsymbol{L}_j, j = 1,\ldots,L$, and $\boldsymbol{e}_{\boldsymbol{L}}$ are the residuals. To obtain $\hat{\boldsymbol{L}}_j$, one can apply either hard or soft thresholding method with the universal threshold $\lambda\varphi 2 \quad {}_i\sqrt{logN}$ to $\boldsymbol{L}_j$ [6] for $1 \le i \le p$, where N is a length of the original p signals and σ_i is the standard deviation of the PC scores of the variable i at level 1 of wavelet decomposition. Denote the matrix of PC scores of wavelet detail coefficients at the level 1 by $D_1 = l_{ik}^1$ for $1 \le i \le p$ and $1 \le k \le N_1$,where l_{ik}^1 is the cell of the matrix D_1 and N_1 is thelength of wavelet detail coefficients at the level 1. The estimate of σ_i can be obtained by solving the eigen value problem of D_1; but the most commonly used estimator of σ_i is the median absolute deviation (MAD) estimator basedon D_1, that is a robust estimator defined as:

$$\hat{\sigma}_i = \frac{median\{|l_{i,1}^1 - \overline{l}_i^1|, |l_{i,2}^1 - \overline{l}_i^1|, \ldots, |l_{i,N_2}^1 - \overline{l}_i^1|\}}{0.6745} \quad (2)$$

where $\overline{l}_i^1$, for $1 \le i \le p$, is the average of the PC scores of wavelet detail coefficients of variable i at the level 1. The wavelet level dependent thresholding method applied to each set of PC scores gives a new estimate of PC scores matrix at each level j.

2.2. Classification in Spectral Domain

The idea behind de-noising by the multi-scale PCA is to process the original signals, so that, they become more deterministic in the feature space. Each extracted signal will be a superposition of a set of PC score functions. Thus, taking FFT of the signals obtained by applying the multi-scale PCA produces features that behave more deterministically in Fourier frequency domain than in the original time domain. In order to classify signals in spectral domain, we propose both confidence interval classification and empirical classification methods.

2.2.1. Confidence Interval Classification Method

The confidence interval classification (CIC) method can be described as follows. Suppose that, the training signals are divided into K groups and that each group is labeled by l, where $1 \le l \le K$. By taking the FFT of the data we transform the data of each group from time domain into spectral domain. For each Fourier frequencyw_i, where $1 \le i \le n$, we calculate the average $P_l(w_i)$ of the Fourier power spectra of the transformed data of each group l. Next, for each average value $P_l(w_i)$ we construct a 95% level confidence interval (CI) based on the approximate normal distribution for each group l of the training data set. In the case of the presented work, the number of the considered frequencies n = 1000, as for each group the energy of these 1000 points contains more than 95% of the total energy of the data of the considered group. Next, for each test signal and for each group l of the training data set we test, for each frequency w_i, if the Fourier power spectrum $P_l(w_i)$ of the test signal is within the 95% CI of the average of the Fourier power spectra $P_l(w_i)$ of the group l of the training data set. For test signal and each training data group l we denote by n_l, a number of Fourier frequencies for which the Fourier power spectrum $P_l(w_i)$ falls within the 95% CI of $P_l(w_i)$. Finally, we calculate for each test signal the ratio of n_l to n for each training data set l and denote itby C_l. We classify the test signal into a group l if C_l is the maximum value of $\{C_1, C_2 \cdots C_K\}$.

2.2.2. Empirical Classification

The empirical classification (EC) method can be described as follows. Suppose that, signals are divided into K groups and that each group is labeled by l, where $1 \le l \le K$. Here, each group l consists of two subsets, *i.e.* it consists of n_1^l test signals and n_2^l training signals. Within each data group l we denote each test signal by j and each training signal by k, where $1 \le j \le n_1^l$ and where $1 \le k \le n_2^l$. We denote for each Fourier frequency w_i, $i = 1 \ldots n$, the Fourier power spectrum of the jth test-signal from group l by the function $P_{lj}(w_i)$, and the Fourier power spectrum of the kth training signal from group l by $G_{lk}(w_i)$, respectively. The proposed empirical classification procedure steps are as follows. 1) Calculate for given l, j, k and i the ratios

$r_{ljk}(w_i) = P_{lj}(w_i)/G_{lk}(w_i)$. 2) Calculate for each Fourier frequency w_i for each test signal j of group l the sample mean of $r_{ljk}(w_i)$, denoted by $\bar{r}_{lj.}(w_i) = \frac{1}{n_2^l}\sum_{k=1}^{n_2^l} r_{ljk}(w_i)$, *i.e.* with respect to all the training signals k of the group l, for $1 \le l \le K$. 3) Calculate the sample variance $S_{lj}^2(\bar{r})$ of $\bar{r}_{lj.}(w_i)$ with respect to w_i. 4) A test signal j is classifiedinto a group l if $S_{lj}^2(\bar{r})$ is the smallest value of $\{S_{1j}^2(\bar{r}), S_{2j}^2(\bar{r}) \cdots S_{kj}^2(\bar{r})\}$.

2.3. Experimental Data

We consider the publicly available data [7] that have five sets, denoted as Set A, B, C, D, and E, respectively. Sets A and B consist of the data segments taken from the surface of EEG recordings of five healthy volunteers. Data in Sets C, D and E come from patients suffering from epilepsy. Set E contains only seizure activity. Each data set (*i.e*., from A to E) contains 100 single-channel EEG signals, each with a total of 4097 sample points. The classification of the normal type and the type with epileptic seizure activities has been widely studied for the considered data sets (*i.e*., Sets A, B, C, D and E) and a high accuracy of these classifications have been achieved. However, in this paper, we focus on the multi-class classification problem.

3. Results

3.1. Two-Class Classification

Before we address the three-class classification problem (*i.e*., the classification problem of the normal, inter-ictal and seizure classes), first, we split randomly the data of each of the Sets A, B, C, D and E, respectively, into the training data set and the test data set of 50 signals each. Next, we study if the test data of each data Set A, B, C, D, and E can be successfully classified as of the type of the respective training data set using the statistical similarity test. To carry out this test, the confidence band of each average of the Fourier power spectra of the pre-processed training data set of each Set A, B, C, D and E is calculated. Next, the computed statistical similarity value of CI test is compared to a pre-defined statistical similarity level to enable a classification decision of each test signal. If the computed statistical similarity value is higher than the pre-defined level, then the pre-processed test signal is classified as of the type of the respective training data. Finally, we count the total number of the correct classifications. **Table 1** shows the results of the accuracy of the two-class classification problem when different levels of the statistical similarity tests are considered for each pre-processed data set. For the statistical similarity level of 0.8 we obtain an accuracy of classification of 50 out of 50 test signals selected from Set A and an accuracy of classification of 49 out of 50 test signals selected from Set E. However, the results in **Table 1** show that the applied CIC method with statistical similarity level of 0.8 does not successfully classify the test data selected from Set B into a group of the respective training data set.

Table 1. The number of correct classifications (displayed in the right columns) out of 50 EEG test signals with respect to different pre-defined statistical similarity levels (listed in the left column) for 5 different data sets (*i.e*., sets A, B, C, D and E) for two-class classification problem.

Statistical similarity level	Set A	Set B	Set C	Set D	Set E
0.95	33	12	34	40	37
0.90	45	19	41	44	46
0.85	48	29	41	45	49
0.80	50	37	41	45	49

3.2. Three-Class Classification

For the three-class classification problem, first, we split randomly the data of each of the Sets A, B, C, D and E, respectively, into the training data set and the test data set of 50 signals each. Instead of only considering the data of Sets A and E separately and ignoring the data of the Sets C, B and D, we combine together data from different sets (e.g., from Set C and Set D). The above described CI based classification method and the proposed EC method need to be modified in order to be applied to the above three-class classification problem. The modifications of the classification methods are needed because we do not classify the test signals into all three possible groups. A test signal from the normal group is classified as either a normal or an inter-Ictal signal and a test signal from the inter-Ictal group is classified as either a normal, or an inter-Ictal, or an Ictal signal. The results of accuracy of the three-classs classification, based on the CIC method and the EC method, are reported in **Table 2** and **Figure 1**. The three-class classification achieves 100% accuracy when the proposed EC method, using only a few PCs (*i.e*., 4 or 5), is applied to the training and test signals. The CIC method used in three-class classification successfully classifies the test data of the inter-Ictal group and the seizure group into the inter-Ictal group and the Ictal group, respectively, but it does not classify successfully test data of the normal group.

Our study shows that the accuracy of classification of the normal group test data depend on the selected feature dimensions for both classification methods. When more features are retained, the classification accuracy of the normal group is decreased regardless which method is used, *i.e*. the CIC method or the EC method (see **Figure 1**). However, for our three-class classification problem, the EC method is more robust than the CIC method in

Table 2. Three-class classification method results for three types of data: normal, inter-Ictal and Ictal.

First 3 PCs						
	CIC			EC		
	Normal	Inter-Ictal	Ictal	Normal	Inter-Ictal	Ictal
Normal	0.55	0.00	N/A	1.00	0.00	N/A
Inter-Ictal	0.45	1.00	0.00	0.00	0.82	0.00
Ictal	N/A	0.00	1.00	N/A	0.18	1.00
First 4 PCs						
	CIC			EC		
	Normal	Inter-ictal	Ictal	Normal	Inter-ictal	Ictal
Normal	0.53	0.00	N/A	1.00	0.00	N/A
Interictal	0.47	1.00	0.00	0.00	0.82	0.00
Ictal	N/A	0.00	1.00	N/A	0.18	1.00
First 5 PCs						
	CIC			EC		
	Normal	Interictal	Ictal	Normal	Inter-Ictal	Ictal
Normal	0.57	0.00	N/A	1.00	0.00	N/A
Inter-Ictal	0.43	1.00	0.00	0.00	0.82	0.00
Ictal	N/A	0.00	1.00	N/A	0.18	1.00

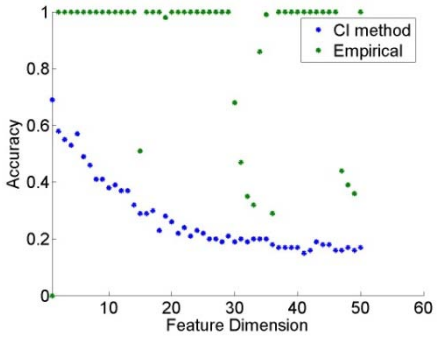

Figure 1. The results show, for the three-class classification method, the classification accuracy of the normal class with respect to different feature dimensions.

classifying the pre-processed & feature extracted test data into the groups of the types of the respective training data sets. The enhanced performance of classification accuracy of our three-class classification for the low feature dimensions implies that feature extraction in the spatial domain greatly improves the classification accuracy of the three-class classification.

4. Conclusion and Future Work

In this paper, we have demonstrated the usefulness of the multi-scale PCA as a new feature extraction method for signals classification problems. The proposed EC method for the three-class classification problem shows an enhanced performance when it is applied to EEG signals in the Fourier frequency domain of the extracted features, obtained by the multi-scale PCA method, and it performs better than the CIC method. Although the discussed feature extraction methods for signals classification were illustrated by applying them to the EEG data, the same methodologies are also applicable to the detection of anomalous events of network traffic.

REFERENCES

[1] I. T. Jolliffe, "Principal Component Analysis," Springer Science+Bussiness Media, Inc., New York, 2004.

[2] M. S. Taqqu, V. Teverovsky and W. Willinger, "Is Network Traffic Self-Similar or Multifractal?" *Fractals*, Vol. 5, 1997, pp. 63-74. http://dx.doi.org/10.1142/S0218348X97000073

[3] B. Vidakovi, "Statistical Modeling by Wavelets," John Wiley & Sons, Inc., Hoboken, 1999. http://dx.doi.org/10.1002/9780470317020

[4] D. Donoho, I. Johnstone, G. Kerkyacharian and D. Picard, "Wavelet Shrinkage: Asymptopia?" *Journal of the Royal Statistical Society*: *Series B*, Vol. 57, 1995, pp. 301-369.

[5] D. Donoho and I. Johnstone, "Minimax Estimation via Wavelet Shrinkage," *Annals of Statistics*, Vol. 26, 1998, pp. 879-921. http://dx.doi.org/10.1214/aos/1024691081

[6] B. Bakshi, "Multiscale Analysis and Modeling Using Wavelets," *Journal of Chemometrics*, Vol. 13, No. 3-4, 1999, pp. 415-434.

[7] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David and C. E. Elger, "Indications of Nonlinear Deterministic and Finite-Dimensional Structures in Time Series of Brain Electrical Activity: Dependence on Recording Region and Brain State," *Physical Review E*, Vol. 64, No. 6, 2001, p. 6190. http://dx.doi.org/10.1103/PhysRevE.64.061907

