PCA Based on Mutual Information for Acoustic Environment Classification

Xueli FAN, Haihong FENG, Meng YUAN
Shanghai Acoustics Laboratory, Chinese Academy of Sciences, Shanghai, China
fanxueli@mail.ioa.ac.cn

Abstract

Principal Component Analysis (PCA) is a common method for feature selection. In order to enhance the effect of selection, a Principal Component Analysis based on Mutual Information (PCAMI) algorithm is proposed. PCAMI introduces the category information, and uses the sum of mutual information matrix between features under different acoustic environments instead of covariance matrix. The eigenvectors of the matrix represent the transformation coefficients. The eigenvalues of the matrix are used to calculate the cumulative contribution rate to determine the number of dimension. The experiment on acoustic environment classification shows that PCAMI has better dimensionality reduction results and higher classification accuracy using neuron network than PCA.

1. Introduction

Signal processing technology pays more and more attention to the different acoustic environment (i.e. sound scene), which sets different parameters or chooses variety strategies for different sound scene[1-4]. Automatic identification of current acoustic environment is important and has developed greatly in many fields, such as the automatic program selection for digital hearing aids[5-7], which can identify the users' environment and adjust signal processing strategies or parameters automatically without switching by hand.

Acoustic environment classification as a common application of pattern recognition[8] has two main functional blocks: feature selection and classifier. The first one selects the most characterizing features and the second one identifies which class the features belong to. In many cases, feature selection is a critical problem to recognize the acoustic environment. On the one hand, high dimension of feature will lead to "curse of dimensionality" problem[9]: approaches that are suitable for a low pattern dimensionality may become unworkable for large dimension because of unrealistic needs of computation and data. On the other hand, redundancy among features will bring in false information, increase computational complexity and cause over-fitting of learning methods.

Principal Component Analysis (PCA)[10] is widely used for feature selection. PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs). PCs are uncorrelated and ordered with variance contribution of data so that the first few retain most of the variation present in all of the original variables.

However, for the classification problems, PCA do not use the category information, which means in a dataset PCA obtains the same results no matter how many classes there are or no matter how to classify the data. The direction of maximum variance is got by treating all data as one class, in which the difference among classes probably is not the largest.

Shannon's information theory[11] is more appropriate for classification, including information entropy, mutual information, conditional mutual information and so on, which provides a suitable formalism for quantifying the uncertainty of the event or arbitrary relationship between variables. Consequently, the idea introducing information theory to feature selection field has become a hotspot in recent years [12-14].

This paper brings mutual information into PCA for acoustic environment classification. Instead of maximizing variance, the algorithm proposed calculates the mutual information between features under each class and sums them up, which takes class into account and considers the redundancy in features. It has been proved that this method yields more appropriate PCs than PCA, and has better effect on dimension reduction. Using Neural Network as classifier in acoustic environment classification experiment, accuracy of correct classified is higher with our algorithm than PCA.

In the following section the shortcoming of PCA is analyzed and an improved version of PCA is proposed.

In section3, the proposed algorithm is applied to acoustic environment classification problem to show the effectiveness of the proposed method. And finally conclusions follow in section 4.

2. Formatting your paper

In this section we will present PCA briefly and illustrate the limitation. Then a Principal Component Analysis based on Mutual Information (PCAMI) algorithm for feature selection is proposed.

2.1 PCA

Suppose that x is a vector of p random variables, and PCA is interested in the variances of the p random variables and the structure of the covariance or correlations between the p variables. To get the PCs, the first step is to look for a linear function $\alpha_1'x$ of the elements of x having maximum variance, where α_1 is a vector of p constants $\alpha_{11}, \alpha_{12}, \cdots, \alpha_{1p}$. Next, look for a linear function $\alpha_2'x$, uncorrelated with $\alpha_1'x$ having maximum variance, and so on, so that at the k-th stage a linear function $\alpha_k'x$ is found that has maximum variance subject to being uncorrelated with $\alpha_1'x$, $\alpha_2'x, \cdots, \alpha_{k-1}'x$. The k-th derived variable $\alpha_k'x$ is the k-th PC. Consider the vector of random variables x has a known covariance matrix Σ, which is the matrix whose (i, j)-th element is the covariance between the i-th and j-th elements of x when $i \neq j$, and the variance of the j-th element of x when $i = j$. It turns out that for k =1, 2, $\cdots$, p, the k-th PC is given by $z_k = \alpha_k'x$ where α_k is an eigenvector of Σ corresponding to its k-th largest eigenvalue λ. Furthermore, if α_k is chosen to have unit length (α_k' α_k = 1), then var(z_k) = λ , where var(z_k) denotes the variance of z_k.

2.2 Limitation of PCA

PCA considers all data in one class and finds the direction in which the distribution of the data is most disperse, that is, the variance of the data is largest, and does not use the category information. Supposing two datasets with same data but different division, the PCs got from PCA are identical. These results are obvious not appropriate. The method using covariance does not consider the relationship between the feature and class or the features within one class, which are important to identify the most characterizing features.

Assume that a two-dimension dataset X consists of 200 data in two classes and each class has 100 data of Gaussian distribution, whose average value and covariance matrix are [1,3]， [1,0;0,50] and [5,3]， [1,0;0,50] respectively. After PCA, the contribution rate of PC1 is 91.2%, that is, the rate of largest eigenvalue λ of covariance matrix Σ of X to the sum of two eigenvalues is 91.2%. Source data and PC1 are shown in Figure 1 and the length of PC1 is 15 times than original to show it clearly.

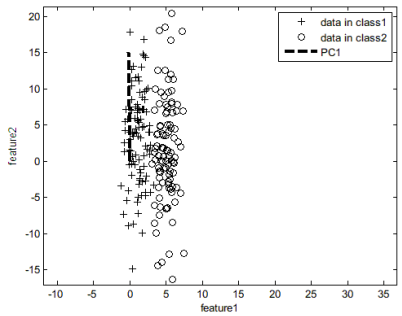

Figure 1. Direction of the PC1 of PCA from Gaussian data

It can be seen from the figure 1 that the variance of data in the direction of vertical axis (feature1) is larger than that of horizontal axis (feature2), which the PC1 from PCA conforms to. However, we cannot separate the data from class1 and the data from class2 in the direction of PC1 due to the data mapped to the direction of PC1 are totally confused. However, if the same dataset is divided into two new classes: the data of feature2 > 0 as class1 and the data of feature2 ≤ 0 as class2, we also get the same PC1 by PCA. In this case, we can identify these two classes in the direction of PC1. So, the class information is important to get the suitable PCs for classification and unfortunately PCA does not consider it. On the other side, variance is not the unique and apposite rule to get the suitable PCs.

2.3 PCA based on mutual information for Feature Selection

Mutual information measures a general dependence between two variables. The measurement is not just appropriate for the random variables of linear correlation but also for the random variables of nonlinear correlation. If $p(f_1)$， $p(f_2)$ are the marginal probability distribution of random variable F_1 and F_2 respectively, and $p(f_1, f_2)$ is the joint probability distribution, then the mutual information $I(F_1; F_2)$ of them is:

$$I(F_1;F_2) = \Sigma_{f_1} \Sigma_{f_2} p(f_1, f_2) \log \frac{p(f_1, f_2)}{p(f_1)p(f_2)}. \qquad (1)$$

It is symmetric in F_1 and F_2 and always nonnegative. The mutual information is equal to zero if and only if

F_1 is independent on F_2. So we can calculate the mutual information between features to evaluate the relationship between them. Considering the importance of categories, conditional mutual information $I(F_1; F_2|C)$ is used to show the relationship between features under each class.

A new algorithm of Principle Component Analysis based on Mutual Information (PCAMI) for feature selection is proposed for improving the performance of PCA. PCAMI introduces the class factor and mutual information to PCA, which uses the sum of conditional mutual information matrix among the features when class is given instead of covariance matrix in PCA. The PCs are obtained from the equation below:

$$B'\Psi_{I_F} B = \Lambda, \tag{2}$$

Where B is the transformation matrix of PCAMI, column vector β_k is the k-th eigenvector of Ψ_{IF}, and Λ is the diagonal matrix of which the diagonal elements are the eigenvalues of Ψ_{IF} , p is the original dimension of dataset, and Ψ_{IF} is the sum of conditional mutual information matrix:

$$\Psi_{I_F} = \sum_{c=1}^{l} MI(F \mid C). \tag{3}$$

Here l is the total number of categories, $MI(F|C)$ is the conditional mutual information matrix whose (i,j)-th element is the conditional mutual information $I(F_i; F_j|C)$ between the i-th and j-th feature when $i \neq j$, and the conditional entropy $H(F_j|C)$ of the j-th feature when $i = j$.

All elements in Ψ_{IF}, both conditional mutual information and conditional entropy, are real numbers and nonnegative according to their definition. Ψ_{IF} is also a symmetric matrix because the (i,j)-th element $I(F_i; F_j|C)$ of conditional mutual information $MI(F|C)$ is equal to the (j,i)-th element $I(F_j; F_i|C)$ and the sum of $MI(F|C)$ is symmetric too. As a result, the eigenvalues of Ψ_{IF} are real, which means Λ is real diagonal matrix, and the matrix B whose columns are eigenvectors of Ψ_{IF} is orthogonal matrix.

According to (2), k-th PC z_k can be represented as $\beta_k'x$, where x denotes the original vector of p random variables and β_k is a vector of p constants $\beta_{k1}, \beta_{k2}, \cdots, \beta_{kp}$. Up to p PCs could be found, but it is hoped, in general, that most of the variation in x will be accounted for by m PCs, where $m \ll p$. The reduction in dimensionality is achieved by accessing the cumulative contribution of PCs. First get all the eigenvalues μ of Ψ_{IF} and arrange them from large to small. It is similar to PCA that the larger eigenvalue μ is, the more important PC corresponded to is. For example, the eigenvector β_1 corresponding to the largest eigenvalue μ_1 multiplied by x is PC1, and the eigenvector corresponding to the k-th largest eigenvalue β_k multiplied by x is k-th PC. Second calculate the contribution rate θ of each PC, which is defined as the rate of eigenvalue corresponding to each PC to the sum of all eigenvalues:

$$\theta_k = \frac{\mu_k}{\sum_{k=1}^{p} \mu_k}. \tag{4}$$

Then obtain the cumulative contribution δ, which is defined as the sum of contribution rate. For example, the k-th cumulative contribution δ_k is the sum of the first k cumulative contributions:

$$\delta_k = \sum_{i=1}^{k} \theta_i. \tag{5}$$

In general, we select m-dimension whose cumulative contribution reaches to the range of 85% to 95% to reduce the dimensionality while retaining as much as possible of the information.

3. Experimental results

In this section we will present some experimental results using the algorithm proposed in section 2 for several datasets.

3.1 Gaussian distribution data

Section 2.2 indicates the limitation of PCA, and to compare with it, we will examine PCAMI using the same dataset. Here rewrite the dataset: the two-dimension dataset X consists of 200 data in two classes and each class has 100 data of Gaussian distribution, whose average value and covariance matrix are [1,3]，[1,0;0,50] and [5,3]，[1,0;0,50] respectively.

The PC1 of PCA and PCAMI are plotted in Figure 2, and the same as Figure 1, the length of PCA1 is 15 times than original to show it clearly. The cumulative contributions of the PC1 are 91.2% and 89.7% respectively.

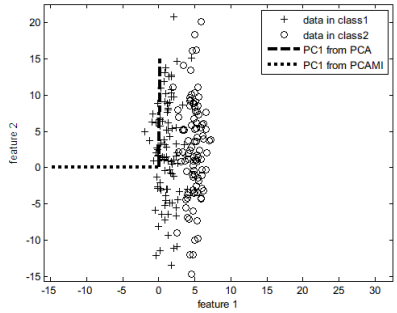

Figure 2. PC1 from PCA and PCAMI of Gaussian data

From Figure 2, it can be seen that the direction of PC1 getting from PCAMI is almost parallel to feature1-axis, which is vertical to the direction of PC1 from PCA. In the direction of former, if we map all the data just as the data are mapped to feature1-axis, the data of two classes are not confused and can be distinguished well, which shows that PCAMI has better performance than PCA.

3.2 Acoustic environment dataset for Feature Selection

The common sound datasets consist of NOIZEUS，NOIS-EX-92 and so on are used and divided into five categories: clean speech, speech in babble noise, speech in nonbabble noise, noisy and music. 500 sound files are selected randomly, and there are 100 files in each class. The common 28 features are extracted as the inputs of PCA and PCAMI for feature selection, including the first four cepstrum coefficients and their variance and first difference, zero-crossing rate and its variance, bandwidth[15], three envelope modulations, spectrum entropy and its variance, center of gravity and its fluctuation strength and its maximum, four features on fundamental frequency[16] and V2W[17].

The contribution rate and cumulative contribution of each PC from PCA and PCAMI are shown in the list below.

TABLE 1. Eigenvalues μ and cumulative contribution from PCA and PCAMI

	PCA		PCAMI	
	μ	δ (%)	μ	δ (%)
PC1	12.78	45.64	16.92	61.53
PC2	5.72	66.05	5.69	82.23
PC3	2.73	75.80	1.17	86.50
PC4	1.18	80.02	0.95	89.94
PC5	0.98	83.51	0.60	92.13
PC6	0.65	85.82	0.46	93.80
PC7	0.63	88.06	0.32	94.98
PC8	0.52	89.90	0.28	95.99
PC9	0.49	91.65	0.15	96.52
PC10	0.46	93.31	0.12	96.98
PC11	0.32	94.46	0.11	97.37
PC12	0.32	95.60	0.10	97.74
PC13	0.26	96.54	0.09	98.07
PC14	0.16	97.13	0.08	98.36
PC15	0.14	97.64	0.07	98.61
PC16	0.12	98.07	0.06	98.84
PC17	0.11	98.48	0.05	99.03
PC18	0.09	98.79	0.05	99.22
PC19	0.08	99.08	0.04	99.37
PC20	0.06	99.28	0.04	99.53
PC21	0.04	99.44	0.03	99.65
PC22	0.04	99.58	0.02	99.72
PC23	0.03	99.71	0.02	99.80
PC24	0.03	99.80	0.02	99.87
PC25	0.02	99.88	0.02	99.94
PC26	0.02	99.95	0.01	99.97
PC27	0.01	100.00	0.01	99.99
PC28	0.00	100.00	0.00	100.00

It can be seen that the cumulative contribution of PCAMI is higher than PCA's when PCs are of same dimension. If 85% is the threshold of choosing the number of PCs, PCA needs 6 PCs while PCAMI needs 3. When 90% is the threshold, PCA needs 9 PCs and PCAMI just needs 6. So the effect on dimensionality reduction of PCAMI is better than that of PCA.

The next experiment will check whether the PCs of PCAMI after dimensionality reduction have better classification accuracy compared with PCA.

3.3 Acoustic environment dataset for Classification

In this experiment, PCs from PCA and PCAMI in experiment B are used as the new features and the inputs of classifier which is neural network. The dimensions of PCs selected with the cumulative contribution beyond 85%, 90% and 95% are used to compare PCAMI with PCA on the classification accuracy.

Table 1 shows the number of PCs needed from PCAMI and PCA respectively when the cumulative contribution is limited to 85%, 90% and 95%. The PCs are used as inputs of the classifier-Neural Network (NN) to examine the classification accuracy rate. A three-layer feedforward backpropagation neural network (also called multilayer perception or MLP) is implemented. The nodes in the hidden layer use a logsig activation function, while a linear transfer function is used in the output layer. The weights of

each node are adjusted using a gradient descent algorithm to minimize the mean squared error (MSE) between the output of the network for a certain training data set and the desired output. The network is trained using the Levenberg-Marquardt backpropagation algorithm. The sound files are divided into two parts randomly: 60% of all files for training the classifier and 40% for testing respectively. The results of acoustic environment classification are shown in the list below.

TABLE 2. Classification accuracy

dim of PCA	accuracy (%)	dim of PCAMI	accuracy (%)
6	79.5	3	79.5
9	89.5	6	90.5
12	92.5	8	93.0

The list above shows that the classified accuracies are almost the same in horizontal. According to Table 1, the cumulative contribution of the first 6 PCs from PCA is 85.82%, and the cumulative contribution of the first 3 PCs from PCAMI is 86.50%. According to Table 2, they lead to the same accuracy. Besides, we can find the similar states for the dimension of 9 and 12 of PCA. So the rule on which we decide the dimensionality of PCs is appropriate for PCAMI. The cumulative contribution using eigenvalues of conditional mutual information can measure the importance of PCs quantitatively. On the other hand, to get the same classification accuracy, PCAMI needs lower dimensionality than PCA, that is, PCAMI can reach to higher accuracy of same dimension than PCA.

4. Conclusion

The feature selection plays an important role in classification problems. Because of the existence of irrelevant and redundant attributes, by transforming them to principal components which are irrelevance and has low dimensionality, higher predictive accuracy can be acquired.

The main motivation for this research is to find more effective principal components for acoustic environment classification. This paper introducing conditional mutual information for feature selection presents the algorithm of Principal Component Analysis based on Mutual Information (PCAMI). PCAMI uses the sum of conditional mutual information matrix between features under each class instead of the covariance matrix, which is more appropriate for classification problems. Then the algorithm measures cumulative contribution of principal components to determine the dimensionality.

The performances of PCAMI are tested by three experiments compared with PCA. The results show that PCAMI can get more effective principal components to distinguish the data among classes, obtain larger cumulative contribution when dimension of principal components are same and needs lower dimension when approximate classification accuracies are got than PCA.

5. Acknowledgments

This work is supported by National Natural Science Foundation of China (11104316), and Shanghai Natural Science Foundation (11ZR1446000).

References

[1] B. Ghoraani, and S. Krishnan, "Time—Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals", *IEEE transactions on audio, speech, and language processing,* vol. 19, 2011, pp. 2197-2209.

[2] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification", *ACM Transactions on Speech and Language Processing (TSLP),* vol. 3, 2006, pp. 1-22.

[3] L. Lu, H. J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation", *Speech and Audio Processing, IEEE Transactions on,* vol. 10, 2002, pp. 504-516.

[4] J. H. Choi, and J. H. Chang, "On using acoustic environment classification for statistical model-based speech enhancement", *Speech Communication*, 2011.

[5] C. Molero, N. R. Reyes, P. V. Candeas, and S. M. Bascon, "Low-complexity F 0-based speech/nonspeech discrimination approach for digital hearing aids", *Multimedia Tools and Applications,* vol. 54, 2011, pp. 291-319.

[6] L. Cuadra, R. Gil-Pita, E. Alexandre, and M. Rosa-Zurera, "Joint design of Gaussianized spectrum-based features and least-square linear classifier for automatic acoustic environment classification in hearing aids", *Signal Processing,* vol. 90, 2010, pp. 2628-2632.

[7] E. Alexandre, L. Cuadra, and R. Gil-Pita, "Sound classification in hearing aids by the harmony search algorithm", *Music-Inspired Harmony Search Algorithm*, 2009, pp. 173-188.

[8] C. M. Bishop, and SpringerLink, *Pattern recognition and machine learning* vol. 4: springer New York, 2006.

[9] R. Duda, and P. Hart, *Pattern Classification and Scene Analysis*: New York: Wiley, 1973.

[10] I. T. Jolliffe, and MyiLibrary, *Principal component analysis* vol. 2: Wiley Online Library, 2002.

[11] C. E. Shannon, "A mathematical theory of communication", *ACM SIGMOBILE Mobile Computing and Communications Review,* vol. 5, 2001, pp. 3-55.

[12] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27, 2005, pp. 1226-1238.

[13] L. Yu, and H. Liu, "Efficient feature selection via analysis of relevance and redundancy", *The Journal of Machine Learning Research,* vol. 5, 2004, pp. 1205-1224.
[14] N. Kwak, and C. H. Choi, "Input feature selection for classification problems", *Neural Networks, IEEE Transactions on,* vol. 13, 2002, pp. 143-159.
[15] J. M. Kates, *Digital hearing aids*: Cambridge Univ Press, 2008.
[16] M. C. Büchler, "Algorithms for sound classification in hearing instruments," Swiss Federal Institute of Technology Zurich, 2002.
[17] E. Guaus, and E. Batlle, "A non-linear rhythm-based style classification for broadcast speech-music discrimination", in *116th AES*, Berlin, Germany, 2004.