浙江科技学院学报,第 19 卷第 2 期,2007 年 6 月
Journal of Zhejiang University of Science and Technology
Vol. 19 No. 2, June 2007

# Method of Feature Extraction Suitable for Hyperdimensional Time Series Data

LOU Tian-liang[1], JIANG Hui-zhong[2]
(1. Department of Computer Engineering, Yiwu Industrial & Commercial College, Yiwu Zhejiang 322000, China;
2. Chinese-German School, Zhejiang University of Science and Technology, Hangzhou 310023, China)

**Abstract**: The increasing hyperdimensional data are acquired from multichannel sensors due to contain more significant information, but the amount of data becomes very large. Extracting significant features from these data is essential for processing and transmission. Optimal discrimination plane (ODP) technique was developed to reduce the redundant data. The features were extracted from the data using the ODP based on Fisher's criterion method. The patterns were projected onto the two orthogonal vectors that built up the ODP, and two-dimensional feature vectors were attained and utilized as features to represent the patterns. Electrocardiogram signals are applied to the analysis as an example in this study. A quadratic discriminant function based classifier and a threshold vector based classifier were employed to measure the performance of the extracted features, respectively. The results show that the proposed ODP is an effective and feasible technique to extract the features from the hyperdimensional time series data.

**Key words**: hyperdimensional time series data; multichannel sensors; feature extraction; classification
**CLC number**: R540.41; TN911.7 **Document code**: A **Article ID**: 1671-8798(2007)02-0089-04

# 一种适于高维时间序列的特征提取方法

楼天良[1],蒋惠忠[2]
(1. 义乌工商职业技术学院 计算机工程系,浙江 义乌 322000; 2. 浙江科技学院 中德学院,杭州 310023)

**摘 要**:为了获得更多的信息,越来越多的数据利用多路传感器进行采集,由此产生了大量的超高维时间序列。特征的提取在处理和传输这些数据中起到至关重要的作用。为此,提出一种最优鉴别平面(ODP)技术以消除数据冗余。该平面由两个在 Fisher 准则基础上建立起来的相互垂直的矢量组成,将模式样本投影到 ODP 上可得到二维特征矢量。为了衡量特征的有效性,分别用二次判别函数分类器和阈值矢量分类器对特征进行分类测试。同时,以心电信号为例对 ODP 方法进行测试,结果表明,该方法应用于超高维数据的特征提取是行之有效的。

**关键词**:高维时间序列;多路传感器;特征提取;分类
**中图分类号**:R540.41; TN911.7 **文献标识码**:A **文章编号**:1671-8798(2007)02-0089-04

**Received date**: 2007-04-20

**Biography**: LOU Tian-liang, male, born in 1965 in Dongyang, Zhejiang, associate professor, specializes in computer information processing technology.

The dimension of sensed data becomes higher and higher because of higher spectral resolution, increasing number of satellite or sensors, and continuously observed multichannel electrocardiogram (ECG) data in intensive care unit (ICU)[1,2]. In order to efficiently obtain necessary information from hyperdimensional data or to transmit the data through a communication channel, the quantity of the data must be reduced. This can be achieved by extracting significant features.

The purpose of the work is to develop a technique to extract features from these hyperdimensional time-series data. The basic idea of the method is that each sampling point in a segment of the signals is weighted and fused according to the weighting factors. The procedures of the proposed method in the paper include redundancy removed by Principle Component Analysis (PCA), data normalized by whitening transform, feature extraction by optimal discrimination panel (ODP) approach developed by Fisher's criterion method, and classification based on the threshold vector method and quadratic discriminant function (QDF)[3]. Whitening transform could make the within-class dispersion spheric. Finally, two-dimensional features are extracted to represent the patterns in the research.

Due to the large number of patients in intensive care unit (ICU) and the need for continuous observation, numerous methods for cardiac arrhythmias classification have been proposed such as Lyapunnov transform[1], nonlinear method[4], AR analysis[5], parametric methods ect[6]. However, these methods seem to lose some of classification information[1]. Thus, the multichannel ECG data including normal sinus rhythm (NSR) and premature ventricular contraction (PVC) were used to analyze and test the ODP method in this study.

## 1 Method

### 1.1 Procedures of the Feature Extraction

The PCA is used to reduce the redundancy of time-series data, white transform is applied to the data in order to normalize them into spheric distribution. The features are extracted by ODP approach that is a linear technique and involved two projecting vectors based on Fisher's criterion. The ODP approach also finds another projecting vector that is orthogonal to the Fisher's vector.

1.1.1 Redundancy of the data reduced by PCA Calculate the within-class scatter matrix $S_w$ of the classes and its eigenvalues and eigenvectors. In order to select the $d$ eigenvectors corresponding to the $d$ largest eigenvalues of $S_w$, the separability criterion based on standard deviation and Euclidean center distance (SDECD) is introduced. The SDECD can be expressed as[3,7]

$$J=\frac{\sqrt{\sum_{i=1}^{d}(\mu_{1i}-\mu_{2i})^2}}{3\left(\frac{1}{d}\sum_{i=1}^{d}\sigma_{1i}+\frac{1}{d}\sum_{i=1}^{d}\sigma_{2i}\right)} \tag{1}$$

where $\sigma_{1i}$ and $\sigma_{2i}$ represent the standard deviations of individual variables of each class, respectively. $\mu_1=[\mu_{11},\mu_{12},\ldots\ \mu_{1d}]^T$ and $\mu_2=[\mu_{21},\mu_{22},\ldots\mu_{2d}]^T$ are the expected vectors for the classes, respectively. The criterion to select the $d$ eigenvectors is to make the $J\geqslant 1.00$, which is calculated based on the reduced data. The sample vector $x_d$'s are generated by projecting each pattern onto these chosen eigenvectors.

1.1.2 The Data Normalized by Whitening Transform After redundancy of the data is reduced by PCA, the within-class dispersion of each class is normalized to spheric distribution by whitening transform. The within-class scatter matrix of the reduced data ($S_{wr}$) is computed in this research. Suppose the eigenvalues and eigenvectors of $S_{wr}$ are expressed by $\lambda_k$ and $p_k$ ($k=1,2,3\ldots d$), respectively, $p_k$ is a $d$-dimensional column vector, then the white transform is given by[8]

$$y_d=P^T x_d \tag{2}$$

where $P=(p_1/\sqrt{\lambda_1},p_2/\sqrt{\lambda_2},\ldots\ p_d/\sqrt{\lambda_d})$ is $d_x\ d$ matrix, $y_d$ is a $d$-dimensional column vector.

1.1.3 Feature Extraction After the data is normalized, the ODP method is applied to the data to extract the two-dimensional features. The feature vector $Z$'s are obtained by projecting $y_d$'s onto the two projecting vectors that are orthogonal between them, and used to represent the ECG segments. Suppose $S_{wn}$ denotes the within-class scatter matrix of the normalized data $y_d$'s. First projecting vector

$v_1$ called Fisher's vector is[3]

$$v_1 = S_{wn}^{-1}(m_1 - m_2) \tag{3}$$

Then another projecting vector $v_2$ can be found according to the ODP. The $v_2$ also maximizes the Fisher criterion and orthogonalizes with $v_1$, which can be wrote as

$$v_2 = \left[ S_{wn}^{-1} - \frac{m_1 - m_2)^{-T}(S_{wn}^{-1})^2(m_1 - m_2)}{(m_1 - m_2)^T(S_{wn}^{-1})^3(m_1 - m_2)} * (S_{wn}^{-1})^2 \right](m_1 - m_2) \tag{4}$$

where $m_1$ and $m_2$ are the expected vectors for the classes, which are computed based on the normalized data, respectively.

### 1.2 Classification Based on QDF and Threshold Vector

After ODP process, the segments or patterns of the signal represented by $Z = [Z_{v1}, Z_{v1}]$'s are classified using the QDF-Based algorithm and threshold vector $Z_0 = [v_{10}, v_{20}]$. Suppose $m_1'$ and $m_2'$ are the expected vectors of class $\omega_1$ and $\bar{\omega}_2$ in the ODP plane, respectively, $\sum_1$ is the covariance matrix of class $\omega_1$, then the QDF can be described by[3]

$$g(Z) = k^2 - (Z - m_1')\sum_1'^{-1}(Z - m_1')^T \tag{5}$$

where $k$ is an uncertain constant to be selected. The decision-making rule is:

if $g(Z) > 0$, then $Z$ belongs to $\bar{\omega}_2$, otherwise $Z$ belongs to $\omega_1$

A threshold vector $Z_0 = [v_{10}, v_{20}]$ is specified as

$$Z_0 = \frac{N_1 m_1' + N_2 m_2'}{N_1 + N_2} \tag{6}$$

The decision-making rule is:

If $Z_{v1} < v_{10}$ and $Z_{v2} < v_{20}$, then $Z$ belongs to $\omega_1$, otherwise $Z$ belongs to $\bar{\omega}_2$

During the training phase, $m_1'$ and $\sum_1$ are computed using the random selected samples of the class $\omega_1$, the sample mean $m_2'$ is computed using the random selected samples of the class $\bar{\omega}_2$, then the QDF based classifier and the threshold vector $Z_0$ can be determined. During the testing phase, the value $g(Z)$ is computed as the equation of (5) with a pre-selected $k$. The above decision-making rules were used to classify the rest testing data as belonging to a class.

## 2 Experiment and Results

The selected data including NSR and PVC was taken from MIT-BIH arrhythmia database, which was sampled from two sensors with frequency 360 Hz. Four patient's ECGs were selected from the database shown in Table 1. In current study, the sample size of the various segments is 0.9 seconds (325 sample points), which 0.3 seconds before R peak and 0.6 seconds after R peak are picked by Tompkin algorithm[9]. The ECG segments with two channels were concatenated together to form 650-dimensional feature vectors.

**Table 1** Evaluation data from the MIT-BIH arrhythmia database

| Identification Number | Number of NSR ($N_1$) | Number of PVC ($N_2$) |
|---|---|---|
| Record 106 | 1 507 | 520 |
| Record 210 | 2 423 | 194 |
| Record 233 | 2 230 | 831 |
| Record 221 | 2 029 | 394 |

Forty largest eigenvalues and corresponding eigenvectors of the within-class scatter matrix $S_w$ were selected based on the criterion SDECD. Thus, 40-dimensional feature vectors were used to represent the ECG segments after PCA. Two-dimensional feature vectors were extracted to represent the ECG segments by projecting the 40-dimensional feature vectors onto the Fisher's vector and its orthogonal vector. A mapping result of the testing data on ODP is shown in Figure 1.

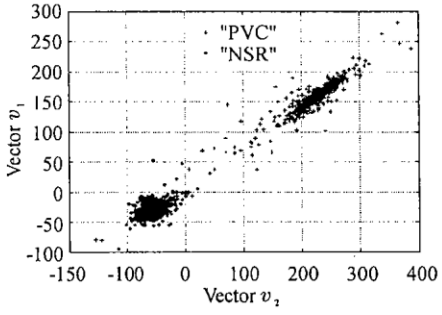


**Figure 1** A mapping result of the testing data on ODP

In training phase, one hundred cases each from the two classes in individual ECG record are random selected for training, and the remaining is used for testing. Table 2 and 3 give the individual ECG record classification results on testing data based on the QDF algorithm and the threshold vector. The universal ECG record classification results are shown in table 4 based on QDF classifier and the training data set containing 400 samples.

**Table 2** Classification accuracy based on QDF

| Classes | ECG Records | 106 | 210 | 233 | 221 |
|---|---|---|---|---|---|
| NSR | Accuracy/% | 99.35 | 97.82 | 97.22 | 99.37 |
| PVC | Accuracy/% | 98.32 | 91.98 | 99.72 | 99.82 |
| | Average/% | 98.84 | 94.90 | 98.47 | 99.60 |

**Table 3** Classification accuracy based on threshold vector

| Classes | ECG Records | 106 | 210 | 233 | 221 |
|---|---|---|---|---|---|
| NSR | Accuracy/% | 99.61 | 98.45 | 99.60 | 99.84 |
| PVC | Accuracy/% | 99.61 | 91.44 | 98.35 | 99.48 |
| | Average/% | 99.61 | 94.95 | 98.98 | 99.66 |

**Table 4** Overall classification accuracy for universal ECG records based on QDF

| Classes | Universal ECG Records | 106, 210, 233, 221 |
|---|---|---|
| NSR | Accuracy/% | 99.01 |
| PVC | Accuracy/% | 95.08 |
| | Average/% | 97.04 |

## 3 Discussion

The objective of this study is to extract the features from hyperdimentional time-series data. A good classification performance with average accuracy of 97.04% has been achieved based on the extracted features and proposed classifiers. In general, class separability not only depends on the class distributions, but also depends on the classifiers to be used. One can sees from Table 2 and 3 that the classification accuracy is almost the same using the two different classifiers. In view of this, the class separability exhibits the less dependency on the classifiers, which does the good features usually hold.

Our experimental results also show the distribution of NSR is closer than that of PVC. So we only used the within-class scatter matrix of class NSR in order to make the distribution of NSR became more circular. One can see from Figure 1 that such a distribution is more suitable for the QDF based classification, too.

The proposed classification results were comparable to some recently published results on arrhythmias classification, for example, classify decimated ECG data including PVC and NSR using artificial neural network, an overall accuracy of 93% was obtained[10]. In addition, we used AR modeling technique to classify the same ECG data shown in Table 1. The model order was 4, and the 4 AR coefficients were used as ECG features to represent ECG segments. The overall accuracy of detecting PVC and NSR is 84.83%.

## 4 Conclusion

It is effective and feasible that proposed ODP method are employed to extract the features from hyperdimentional time-series data.

**References:**

[1] MOHAMED I O, ABOU-ZIED AHMED H M. Study of features based on nonlinear dynamical modeling in ECG arrhythmia detection and classification[J]. IEEE Trans Biomed Eng, 2002,49(10):733-736.

[2] HEON G L, KIYONG N. Cardiovasular disease diagnosis method by emerging patterns[C]//Proceeding of Advanced Data Mining and Application, Xian, China, Springer, 2006:809-818.

[3] 边肇祺,张学工. 模式识别[M]. 北京:清华大学出版社,2002:24-134.

[4] SUN Y, CHAN K L, KRISHNAN S M. Life-threatening ventricular arrhythmia recognition by nonlinear descriptor[J]. Biomedical Engineering Online, 2005,4(1):6.

[5] ARNOLD M, MILTNER W H R, WITTE H. Adaptive AR modeling of non-stationary time series by means of Kalman filtering[J]. IEEE Trans Biomed Eng, 1998,45(5):553-562.

[6] IRENA J, FRANCOIS M, AUDE V. Defibrillation shock success estimation by a set of six parameters derived from the electrocardiogram[J]. Physiol Meas, 2004,25(6):1179-1188.

[7] FUKUNAGA K. Introduction to Statistical Pattern Recognition[M]. New York: Academic Press, 1990: 153-409.

[8] SADAO F, SENYA K. Application of feature extraction scheme to discrimination of Electrocardiofram. T IEE Japan, 2001,121-A(8):725-730.

[9] TOMPKINS W. Biomedical Digital Signal Processing [M]. Englewood Cliffs, New Jersey: Prentice Hall, 1993:246-261.

[10] MELO S L, CALOBA L P, NADAL J. Arrhythmia analysis using artificial neural network and decimated electrocardiograph data[C]//Proceeding of Computers in Cardiology, Piscataway, USA, IEEE, 2000, 27:73-76.