文章编号 1004-924X(2015)增-0453-06

基于主成分分析进行特征融合的心拍分类

王 迪[1],司玉娟[1,2*],刘 通[1],郭永丛[1],刘书文[3]

(1.吉林大学 通信工程学院,吉林 长春 130012;
2.吉林大学 珠海学院,广东 珠海 519041;
3.北京大学 软件与微电子学院,北京 100871)

摘要:针对心拍分类中单一特征分类准确率低而组合特征存在特征冗余和维度过高的问题,提出了一种基于主成分分析(PCA)进行特征融合的心拍分类算法。该算法在对单一特征进行规范化后,将多种特征组合为一个具有大量信息的高维特征;之后引入 PCA 算法去除冗余成分,得到低维融合特征;最后,利用支持向量机(SVM)为分类器完成心拍分类。以心拍时域特征、离散小波特征、离散傅里叶特征为例,在 MIT-BIH 数据库上进行了实验。实验结果表明,融合特征维度为 100 时,分类准确率可达 97.389%。与单一特征相比,融合特征提高了分类准确率;与组合特征相比,融合特征降低了特征维度。

关 键 词:心拍分类;分类准确率;特征融合;主成分分析; MIT-BIH 数据库

中图分类号:R540.41;TP391.4 **文献标识码**:A **doi**:10.3788/OPE.20152313.0453

Heart beat classification based on feature fusion by principle component analysis

WANG Di[1], SI Yu-juan[1,2*], LIU Tong[1], GUO Yong-cong[1], LIU Shu-wen[3]

(1. *College of Communication Engineering, Jilin University, Changchun* 130012, *China*;
2. *Zhuhai College of Jilin University, Zhuhai* 519041, *China*;
3. *College of Software & Microelectronic, Peking University, Beijing* 100871, *China*)
* *Corresponding author, E-mail: siyj@jlu.edu.cn*

Abstract: For lower accuracy by a single feature and too high dimensionality by a combined feature in heartbeat classification, a new heartbeat classification algorithm was proposed based on the feature fusion by using Principle Component Analysis(PCA). With proposed algorithm, each single feature was normalized, then all kinds of features were combined together to a new one with high dimensionality and more information. In order to reduce dimensionality of combined features, the PCA was employed to remove redundant components. Finally, the Support Vector Machine(SVM) was used as a classifier to classify different heartbeats. By taking time domain feature, Discrete Wavelet Trans form (DWT) feature and Discrete Fourier Transform(DFT) feature as examples, the experiments were performed in MIT-BIH database. This study was compared with the new feature mentioned above,

收稿日期:2015-04-15;**修订日期**:2015-05-07.

基金项目:吉林省重点科技攻关项目(No. 20150204039GX);吉林省长春市重大科技攻关专项资助项目(No. 14KG064);广东省科技计划资助项目(No. 2013B010101020)

three single features and combined feature without Principle Component Analysis. The experiments results indicate that the accuracy of the new features is 97.389% when its dimension is 100. The new fusion feature has lower dimension than combined feature and higher classification accuracy than single feature.

Key words: heart beat classification; classification accuracy; feature fusion; Principle Component Analysis(PCA); MIT-BIH database

1 引　言

心电图是心脏活动信号的记录。其中，心拍的类别是诊断心脏疾病的重要依据。然而，人工对心拍进行分类费时费力，因此，对 ECG 信号进行自动分类的研究越来越受到重视。

心拍特征是进行心拍分类的主要依据，提高心拍分类准确率是进行心拍分类的主要目的，当前提高分类准确率的主要方法是寻找适合进行心拍分类的单一特征，如 Philip 利用心拍的时域形态学特征实现心拍的分类[1]，或者组合使用不同类别心拍特征，如 Hari 通过使用神经网络、心电时域与离散小波域特征建立分类模型实现心拍分类[2]。但是这些方法都存在缺陷：单一特征含有的分类信息较少，通常很难得到理想的分类准确率，组合特征虽然具有足够的分类信息，但其特征维度会随特征种类增加而快速升高，同时存在特征冗余，这样会限制特征种类的选择，不利于分类。基于这些缺陷，本文提出了基于主成分分析(PCA)进行多特征融合心拍分类算法。主成分分析是一种多元统计分析方法。将原变量通过一系列线性组合构成新变量，使这些新变量在彼此互不相关的前提下尽可能多地体现原变量的信息，普遍应用于去除相关特征、提取和数据压缩中[3]，如应用主成分分析进行光全散射特征波长的选择[4]，基于主成分分析进行唇部轮廓建模[5]，使用主成分分析进行烟雾检测[6]。本文利用主成分分析充分考虑特征组合后各特征项之间的相关，将原始组合特征转换为较低维的特征，新的融合特征保留了原特征的主成分，且相互之间互不相关，去除了特征冗余，进而获得组合特征最佳描述特征。实验结果表明，该算法得到的融合特征能有效降低组合特征维度，提高心拍分类准确率。

2 心拍特征选取

为证明基于主成分分析特征融合得到的融合特征能有效解决单一特征分类准确率低，组合特征存在特征冗余和维度较高的问题，选取心拍实验中常用、特征维度较高、且存在较强相关性的3类特征：时域特征、离散小波特征、离散傅里叶特征为例，进行算法实验。

2.1 时域特征

对心电信号进行预处理后，采用 ECGPUWAVE 检测心电信号的分割点，通过获取每个心拍的 P 波起点与 T 波终点，截取出心拍向量，再对心拍进行重采样得到时域特征向量。时域特征向量重采样后维度为300。

2.2 离散小波特征

对心拍样本进行离散小波变换，可以获得能够体现心拍类别间差别的频率成分[7]。

在使用离散小波小波变换对信号进行分析时，对最终的性能及结果会产生重要影响的是母小波的选择及分解尺度的选择。文献[8]中指出，由于2阶 Daubechies 小波(db2)具有很好的平滑特性，所以它更加适合对心电信号的变化进行检测。文献[9-10]表明，在 MIT 数据库 360 HZ 的采样频率下，心拍之间最能体现不同类别间差异的特征频率成分集中在第四尺度。因此，对心拍的特征向量进行离散小波变换后，取第四尺度的概貌系数(21维)和细节系数(21维)共同构成42维的特征向量。

2.3 离散傅里叶特征

对心电时域信号通过离散傅里叶变换(DFT)后进行频域分析是生物医学信号处理的常用方法，可以找出心拍在时域中不太明显而在频域中较明显的特征。

采用离散傅里叶变换对心拍进行处理后，由于频谱图对称，因此取傅里叶频域向量的前150

点作为心拍离散傅里叶变换后频域特征，维度为 150。

3　特征分离度比较

特征选定后，首先通过评价准则衡量特征与类之间的相关性，对特征的分类有效性进行评估。

距离度量，也常被认为是分离度、差异性或辨识能力的度量，通常被用于找到使两类尽可能分类的特征。由于上述心电信号特征为线性不可分数据集，因此在测定距离度量时使用文献[11-12]中提出的基于核空间的距离度量，核函数采用高斯径向基核函数。

将上述 3 种特征单一使用和各种组合使用时的情况考虑在内。单一特征按时域特征、离散小波特征、离散傅里叶特征的顺序表示为单一 1、单一 2、单一 3，组合特征按时域特征＋离散小波特征，时域特征＋离散傅里叶特征，离散小波特征＋离散傅里叶特征，时域特征＋离散小波特征＋离散傅里叶特征的顺序表示为组合 1、组合 2、组合 3、组合 4。

不同特征在进行正常与不正常两类心拍区分时分离度如图 1 所示。

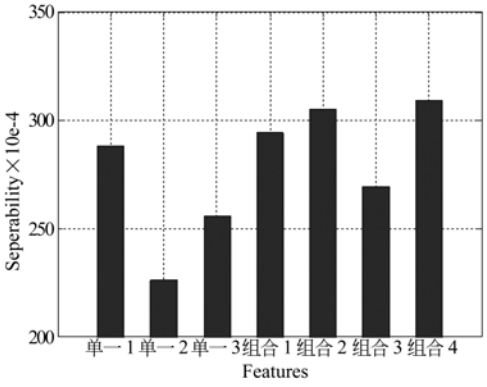


图 1　不同特征组合分离度

Fig. 1　Seperability of different features combined together

由图 1 可以看出，组合特征的分离度普遍比单一特征高，即理论上组合特征在进行分类时会比单一特征可以更好地区分两类心拍。

对组合特征进行主成分分析后，原组合特征与融合特征分离度比较如图 2 所示。融合特征指组合特征经文中算法处理后得到的低维特征。

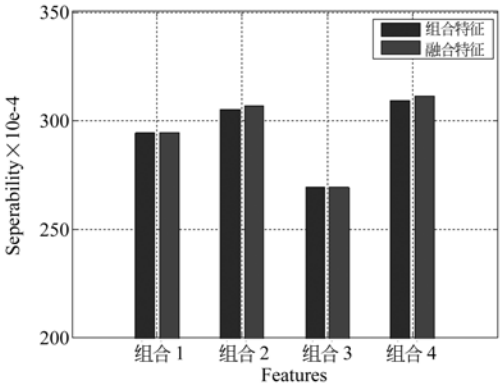


图 2　组合特征与融合特征分离度对比

Fig. 2　Comparison of the separability of different combined features and fusion features

由图 2 可以看出各种组合特征与其相应的融合特征相比，分离度变化很小，这说明 PCA 降维有效保留了原特征中的分类信息。

通过分离度的比较分析，可以看出，基于主成分分析进行的特征融合，能够在增加组合特征种类，提高分类准确率的同时，降低特征维度，同时有效保留原特征中的分类有效信息。

4　基于主成分分析进行特征融合的心拍分类

1）输入 N 个心拍的 3 种不同类别特征：$\boldsymbol{a}_1$，$\boldsymbol{b}_1$，$\boldsymbol{c}_1$，$\boldsymbol{a}_2$，$\boldsymbol{b}_2$，$\boldsymbol{c}_2$，… $\boldsymbol{a}_n$，$\boldsymbol{b}_n$，$\boldsymbol{c}_n$。采用多特征组合，得到特征向量矩阵 $\boldsymbol{X}$：

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{a}_1,\boldsymbol{b}_1,\boldsymbol{c}_1\\ \boldsymbol{a}_2,\boldsymbol{b}_2,\boldsymbol{c}_2\\ \vdots,\vdots,\vdots\\ \boldsymbol{a}_n,\boldsymbol{b}_n,\boldsymbol{c}_n\end{bmatrix}, \tag{1}$$

2）对矩阵 $\boldsymbol{X}$ 按列根据式(2)标准化处理得到标准化矩阵 $\boldsymbol{X}^*$。

$$x^*=\frac{x-x_{\min}}{x_{\max}-x_{\min}}, \tag{2}$$

式中：$x_{\min}$ 表示每一列的最小值，$x_{\max}$ 表示每一列的最大值，x 为原矩阵的元素，x^* 为标准化矩阵中元素。

3）根据式(3)建立自相关矩阵。

$$\boldsymbol{R}=\boldsymbol{X}^{*\mathrm{T}}\boldsymbol{X}^*/(\mathrm{N}-1), \tag{3}$$

式中：$\boldsymbol{X}^*$ 为归一化标准处理后的矩阵。

4）根据相关矩阵求得其特征值和特征向量。得到自相关矩阵的特征值 $\lambda_1\geq\lambda_2\geq\cdots\lambda_m$ 及各特征值对应的特征向量 $\boldsymbol{u}_1,\boldsymbol{u}_2\cdots\boldsymbol{u}_m$。

5)确定主成分个数:方差贡献率及累计方差贡献率分别为:

$$\eta_i = \lambda_i / \sum_i^m \lambda_i, \tag{4}$$

$$\eta_\Sigma(p) = \sum_i^p \eta_i. \tag{5}$$

选取主成分的个数时需要依据累计方差贡献率,通常情况下,如果累计方差贡献率为 75%～95%[13-15],则其对应的前 p 个主成分包含原始变量的绝大部分信息,主成分个数就是 p 个。

p 个主成分对应的特征向量为 $\mathbf{U}_{mp} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_p]$,则 N 个样本的 p 个主成分构成的矩阵为:

$$\mathbf{Z}_{Np} = \mathbf{X}^* \mathbf{U}_{mp}, \tag{6}$$

7)从降维后特征向量矩阵 $\mathbf{Z}_{Np}$ 随机选取一定个数的心拍作为训练集,利用 SVM 训练得到训练模型 Model,将剩余的所有心拍作为测试集。

8)将测试集中所有心拍通过分类器模型 Model 进行预测,得到心拍预测标签 Label。

所有特征参与特征组合,分类信息量随之增加,维度迅速上升,但此时并非所有属性对分类决策都有贡献。PCA 算法将组合特征向量映射到 p 维空间上,通过重新构造出的 p 维特征表征原有特征,剔除噪声和特征冗余。PCA 算法不是简单地将心拍特征向量去除若干个,而是减少了表达原组合特征矩阵的正交基向量个数,丢弃了其他维度的信息。尽管降维中必然伴随着信息的丢失,但 PCA 算法基于方差最大化原则,将原来组合特征通过少数几个特征来表示。这几个特征与原特征相比,维度大幅下降,同时还集中了原始特征的大部分信息。算法通过支持向量机训练分类得到测试集标签后,与原测试集标签比较,通过分类准确率衡量算法有效性。

5 实　验

实验随机选取 1 200 个心拍样本作为训练集,将剩余的所有样本作为测试集,SVM 分类器选用高斯核函数,在进行单一特征、组合特征与融合特征的分类结果比较时,为了使其具有更强可比性,SVM 的最大迭代次数、阈值设置等都是完全一样的。双特征融合与三特征融合的维度在降维后主成分个数方差贡献率均在选为 97%以上。实验结果用分类准确率来描述分类性能。

文中以上述 3 种特征融合为例先进行特征组合,得到一个新的 492 维(时域特征 300 维+小波域特征 42 维+FFT 变换后频域特征 150 维)特征向量。

根据主成分分析算法进行降维。不同主成分个数对应的方差贡献率如表 1。主成分个数超过 100 后,累计方差贡献率变化很小,方差贡献率均在 97%以上,因此选取降维后的 100 维特征作为新的特征向量。经实验验证,双特征融合时,主成分个数为 100 时方差贡献率也在 97%以上。

表 1　主成分贡献率

Tab. 1　Principle component contribution rate

主成分个数	特征向量方差累积贡献率	主成分个数	特征向量方差累积贡献率
10	83.76%	90	98.69%
20	91.60%	100	98.96%
30	94.26%	110	99.18%
40	95.58%	120	99.36%
50	96.49%	130	99.52%
60	97.23%	140	99.65%
70	97.84%	150	99.75%
80	98.32%	—	—

实验利用心拍特征在 MIT-BIH 数据库上进行正常与不正常心拍的二分类。

首先,测试单一特征的分类效果,其特征维度和分类准确率如表 2。

表 2　单一特征的分类准确率

Tab. 2　Classification accuracy of single features

心拍特征向量	维度	未降维测试集准确率
时域特征	300	96.0833%
DWT 特征	42	95.6963%
DFT 特征	150	96.5000%

其中 DWT 特征表示心拍的离散小波特征,DFT 特征表示离散傅里叶特征。

接着分别测试不同组合特征和融合特征的准确率,降维后的分类结果均是降维至 100 时特征向量得到的分类结果。分类结果如表 3 所示。

从实验结果来看,双特征组合后的分类准确

率普遍比单一特征的分类准确率高，而三特征组合后，分类准确率得到了进一步的提高。这说明，增加组合特征中特征种类，有利于提高分类准确率。将融合特征与组合特征对比，可以看出，融合特征的维度相比组合特征明显减少，但分类准确率并没有因此减少，相反部分融合特征的分类准确率反而高于组合特征的分类准确率。这说明融合后特征保留了原组合特征的分类有效信息，完全可以取代原组合特征作为分类器分类特征。综上，基于主成分分析的特征融合心拍分类算法，有效解决单一特征分类准确率低的问题，同时其通过主成分分析，得到原组合特征最佳描述特征，减少特征冗余，降低特征维度，有效解决多特征组合维度过高，存在特征冗余的问题。

表 3　不同特征组合的分类准确率

Tab. 3　Classification accuracy of different combined features

心拍特征组合	未降维维度	未经 PCA 降维测试集准确率	降维后维度	经 PCA 降维后测试集准确率
时域特征＋DWT 特征	342	96.011 5%	100	95.991 0%
时域特征＋DFT 特征	450	97.149 1%	100	97.327 3%
DWT 特征＋DFT 特征	192	97.293 0%	100	97.272 5%
时域特征＋DWT 特征＋DFT 特征	492	97.334 2%	100	97.389 0%

6　结　论

针对心拍分类中单一特征分类准确率低、多特征维度高，存在特征冗余的问题，本文提出了基于主成分分析进行特征融合的心拍分类算法。以心拍分类的 3 种常用特征：时域特征、离散小波特征、离散傅里叶特征为例，将其融合后使用 SVM 进行分类，与单一特征、组合特征分类准确率进行对比。实验结果证明，融合后的特征能大大降低特征维度，提高分类准确率，有效解决单一特征分类准确率低，组合特征维度过高和存在特征冗余的问题，当融合特征维度为 100 时，分类准确率为 97.389%。

参考文献：

[1] PHILIP C, MARIA O′D, RICHARD B. Automatic classification of heartbeart using ECG morphology and heartbeat interval features[J]. *IEEE Transactions on Biomedical Engineering*, 2004(51): 1196-1206.

[2] HARI M R, ANURAG T, SHAILJA S. ECG signal processing for abnormalities detection using multi-resolution wavelet transform and artificial neural network classifier[J]. *Measurement*, 2013(46): 3238-3246.

[3] 林海明，杜子芳. 主成分分析综合评价应该注意的问题[J]. 统计研究，2013，30(8)：25-31.
LIN H M, DU Z F. Some problems in comprehensive evaluation in the principal component analysis[J]. *Statistical Research*, 2013, 30(8): 25-31. (in Chinese)

[4] 唐红，郑文斌，李宪霞. 主成分分析在光全散射特征波长选择中的应用[J]. 光学 精密工程，2010，18(8)：1691-1698.
TANG H, ZHENG W B, LI X X. Application of principal component analysis to selection of characteristic wavelengths with total light scattering[J]. *Opt. Precision Eng.*, 2010, 18(8): 1691-1698. (in Chinese).

[5] 王丽荣，王建蕾. 基于主成分分析的唇部轮廓建模[J]. 光学 精密工程，2012，20(12)：2768-2772.
WANG L R, WANG J L. Lip contour modeling based on PCA[J]. *Opt. Precision Eng.*, 2012, 20(12): 2768-2772. (in Chinese)

[6] 付小宁，张涛，万里. 基于多光谱分离的烟雾检测[J]. 光学 精密工程，2013，21(11)：2798-2802.
FU X N, ZHANG T, WAN L. Smoke detection based on multispectral separation[J]. *Opt. Precision Eng.*, 2013, 21(11): 2798-2802. (in Chinese)

[7] 范增飞. 基于频域特征的心电图分类研究[D]. 天津：天津大学，2006.

FAN Z F. *ECG classification based on frequency domain features*[D]. Tianjin: Tianjin University, 2006. (in Chinese)

[8] LIU T, SI Y J, WEN D W, *et al.*. Vector quantization for ECG beats classification[C]. *17th IEEE International Conference on Compututational Science and Engineering*, Chengdu, *Computational Science and Engineering*(CSE), 2014: 13-20.

[9] YU S N, CHEN Y H. Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network[J]. *Pattern Recognition Letters*, 2007, 28(10): 1142-1150.

[10] KHORRAMI H, MOAVENIAN M. A comparative study of dwt, cwt and dct transformations in ecg arrhythmias classification[J]. *Expert Systerms with Applications*, 2010, 37(8): 5751-5757.

[11] 蔡哲元, 余建国, 李先鹏, 等. 基于核空间距离测度的特征选择[J]. 模式识别与人工智能, 2010, 23(2): 235-240.
CAI Z Y, YU J G, LI X P, *et al.*. Feature selection algorithm based on kernel distance measure[J]. *Pattern Recognition and Artificial Intelligence*, 2010, 23(2): 235-240. (in Chinese)

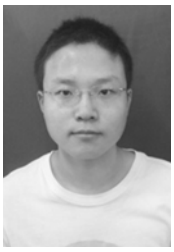
[12] 韩景梅. 支持向量机决策树算法研究及其应用[D]. 上海: 上海交通大学, 2007.
HAN J M. *Research on support vector machine-decision tree arithmetic and its application*[D]. Shanghai: Shanghai Jiao Tong University, 2007. (in Chinese)

[13] WANG D K, LI D W, LIN Y. A new method of face recognition with data field and PCA[C]. 2013 *IEEE International Conference on Granular Computing*(*Grc*), Beijing, 2013: 320-325.

[14] ZHAI M H, SHI F Y, Duncan D. Convariance-Based PCA for multi-size data[C]. 2014 *22nd International Conference on Pattern Recognition*, Stockholm, 2014: 1603-1608.

[15] LIM S T, YAP D F W, MANAP N A. Medical image compression using block-based PCA algorithm [C]. 2014 *International Conference on Computer, Communications, and Control Technology*, Langkawei, Kedah, 2014: 171-175.

作者简介：



王　迪(1991－)，男，河南安阳人，2013年于吉林大学获得学士学位，主要从事信号与信息处理方面的研究。E-mail: 15143086560m0@sina.cn

司玉娟(1963－)，女，吉林长春人，博士，教授，博士生导师，1985 年、1988 年、1996 年于吉林工业大学分别获得学士、硕士、博士学位，主要从事嵌入式系统、生物医学信号的处理与识别等方面的研究。E-mail: siyj@jlu.edu.cn