浙江科技学院学报，第 17 卷第 1 期，2005 年 3 月
Journal of Zhejiang University of Science and Technology
Vol. 17 No. 1, Mar. 2005

一种适于心电数据的可分性评价准则研究

葛丁飞[1]，李时辉[2]，瞿　晓[1]
（1. 浙江科技学院 信息与电气工程学系，浙江 杭州 310023；2. 义乌工商职业技术学院 计算机工程系，浙江 义乌 322000）

摘　要：在心电特征提取中，现存的基于多种可分性准则和 K-L 变换等技术在特征提取的有效性等方面都有各自的优缺点，都不能保证取得满意的结果。这就需要有一个评价准则来衡量最终特征的有效性和类间的可分性。为此，本文针对心电数据提出了一种基于标准差和欧氏中心距的可分性评价准则，并应用于 MIT-BIH 数据库中心律失常数据的特征提取、特征有效性的评测和决策树的设计。实验结果表明，这是一种有效和实用的可分性评测准则。

关键词：可分性准则；特征提取；心律失常；标准差
中图分类号：R540.41；TN911.7　　**文献标识码**：A　　**文章编号**：1671-8798(2005)01-0009-04

Research on a separability measurement criterion suitable for cardiac data

GE Ding-fei[1], LE Shi-hui[2], QU Xiao[3]
(1. Department of Information and Electrionic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China; 2. Department of Computer Engineering, Yiwu Industrial and Commercial College, Yiwu 322000, China)

Abstract: Existing feature extraction techniques for cardiac signals are based on various separability criterion and like K-L transformation ect. Most of them have some advantages and disadvantages and are not always satisfactory for effective feature extraction. There is a need to develop a criterion to measure the performance of cardiac features and the separability between different classes. A separability measurement criterion suitable for cardiac data was presented, which was based on standard deviation and Euclidean center distance. The criterion was applied to cardiac arrhythmia data obtained from MIT-BIH database in order to extract features, measure the performance of features and build the decision tree for multiclass classification. The experimental results show that it is an effective and practical separability measurement criterion.

Key words: separability criterion; feature extraction; cardiac arrhythmia; standard deviation

心律失常的自动检测和分类在临床上有着重要　　的应用，其关键是如何提取有效的心电特征，国际

收稿日期：2004-12-01
基金项目：浙江省自然科学基金项目(Y104284)
作者简介：葛丁飞(1965—　)，男，浙江东阳人，工程师，主要从事生物信号的研究与教学工作。

上众多专家都对其进行了广泛而深入的研究[1~4]。目前，在二次心电特征提取中，常用的方法是基于多种可分性准则和K-L等变换的特征值排序法，由于这些可分性准则在特征提取的有效性等方面都有各自的优缺点，都不能保证取得满意的结果[5~8]。比如：基于类内类间距离的可分性准则虽然便于分析和判断，但是，不能包含概率分布和各类交叠的信息，因此，有时不能取得最终满意的结果[5]。这就需要有一个准则来衡量最终特征的有效性。再者，在特征值排序中，如何选取选前 k 个特征向量对所提特征的可分性有着重要的影响。另外，在多类决策树的设计中，关键问题之一是类别分组，一种常用的方法是依据类间的可分性程度进行分组[6]。这也需要一个定量的准则来衡量最终特征的有效性和类间的可分性。

作为衡量特征的有效性和类间的可分性评价准则，需要具有计算简单、容易实现以及与分类结果保持一致的特性。而现有的这些可分性准则包括基于类内类间距离的可分性准则大多难以满足这些要求。比如：基于熵函数的可分性准则往往会遇到对概率分布作出估计困难的问题；基于概率分布的可分性准则，如Bhattacharyya距离和散度等，虽然不需要对概率分布作出估计，但是，有时会遇到奇异或近奇异协方差矩阵求逆困难的问题，其中Bhattacharyya距离已经成为一种流行的可分性评价准则[8]。

本文结合欧氏距离和概率分布信息提出了一种基于中心距和标准差的可分性评价准则。并应用于MIT-BIH数据库中的心电数据的特征提取、特征有效性的评测和决策树的设计。结果表明，该准则是一种计算简单、实现容易和有效的可分性评价准则。

1 常用的基于Bhattacharyya距离的可分性评价准则

一种常用的可分性评价准则Bhattacharyya距离是均值向量和协方差矩阵的函数，它属于参数方法的范畴[8]。其计算公式是：

$$J_B=\frac{1}{8}(\mu_1-\mu_2)^T\left[\frac{\sum_1+\sum_2}{2}\right]^{-1}(\mu_1-\mu_2)+\frac{1}{2}\ln\frac{\left|\frac{1}{2}(\sum_1+\sum_2)\right|}{[|\sum_1||\sum_2|]^{1/2}} \tag{1}$$

这里 $\mu_1=[\mu_{11},\mu_{12},\mu_{13}\cdots\mu_{1d}]^T$、$\mu_2=[\mu_{21},\mu_{22},\mu_{23}\cdots\mu_{2d}]^T$ 分别是 $d\times1$ 类均值向量，$\sum_1$、$\sum_2$ 分别是 $d\times d$ 协方差矩阵，d 为特征矢量的维数。其中，第一项反映了由于均值向量不同引起的类分布之间的差异；第二项反映了由于协方差矩阵不同引起的类分布之间的差异。

2 基于欧氏中心距和标准差的可分性评价准则

公式(1)中，$\sum_1$、$\sum_2$ 是奇异或近奇异矩阵时，J_B 会遇到计算和实现上的困难，如矩阵的求逆和 $|\sum|$ 非常近似于零等等，虽然理论上可以通过求矩阵的特征值和特征向量的方法予以解决，但是对于一个巨大的矩阵而言，其计算往往是极为复杂的，而且实现也是非常困难的[8]。为此，本文引入基于欧氏中心距和标准差的可分性评价准则 J，其计算公式如下：

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

其中，σ_{1ii}、σ_{2ii}，$i=1,2,3\cdots d$，分别是特征向量各分量的标准差，分子为两类之间的欧氏中心距。在理论上，它描述了两类之间的空间距离，其值越大，越有利于分类；分母中的每项为特征向量各分量标准差的平均值，它包含了概率分布的信息，其值越小，样本的收敛性越好。根据正态分布函数性质，样本分布在离中心点3倍标准差距离外的概率小于0.3%。J 值越大，可分性程度越高。当 J 值大于1时，我们假设两类样本之间可分性达99.00%。

3 实验与结果

3.1 基于欧氏中心距和标准差的可分性评价准则 J 在特征提取中的应用

本实验中采用MIT-BIH数据库中双导联心率失常信号，包括心室性心动过速(VT)、心室纤维性颤动(VF)各300个样本。利用一个基于360 Hz带通滤波器对信号进行预处理，其上下边带截止频率是1 Hz和50 Hz。数据窗口为325个采样点(0.9 s)，每一个样本的650个采样点(325×2)构成了一个650维的样本特征失量。再利用K-L变换进行二次特征提取，用以下两种方法选取K-L变换的特征向量：①选前 k 个最大特征值的特征向量，使得样本

在前 k 个轴上的能量占整个能量的 99%以上；②利用基于欧氏中心距和标准差的可分性评价准则 J，选前 k 个最大的特征值的特征向量，使得 $J\geqslant 1$ 或 J 值没有明显增加为止。利用线性分类器进行分类。VT 和 VF 的实验结果如表 1 所示。

表 1　不同 k 值的实验结果

k 值		40	50	60	70	100	200	300	350	400
k 个轴所占能量%		98.571	99.230	99.604	99.740	99.999	99.999	99.999	99.999	99.999
J 值		0.1132	0.1654	0.1856	0.2145	0.2649	0.5198	0.7787	0.9082	1.038
分类精度%	VT	69.6	70.9	72.0	73.5	74.8	82.8	88.0	94.4	98.4
	VF	59.1	62.9	64.5	65.1	67.6	80.4	89.2	95.2	99.4

3.2　基于欧氏中心距和标准差的可分性评价准则 J 在多类决策树设计中的应用

利用 MAR 模型对 MIT-BIH 中双导联心率失常信号进行建模，然后提取特征，包括心房早期收缩(APC)、心室早期收缩(PVC)、心室性心动过速(VT)、心室纤维性颤动(VF)、室上性心动过速(SVT)信号各 300 个样本，数据窗口为 325 个采样点(0.9 s)，建模时的信号频率为 250 Hz。引用最小误差平方 SSE 准则确定模型阶次为 4，MAR 系数作为特征，其特征向量维数为 16。再利用 K-L 变换对 MAR 系数进行二次特征提取，得到二次特征 K-LMAR 系数，其特征向量的维数为 10。在决策树的设计中，依据各类 ECG 特征值之间的 J 进行分组，J 值小的被分为同一组，决策树的决策过程如图 1 所示，其中的每一步都利用线性分类器进行分类。基于 MAR 系数和 K-LMAR 系数的类间 J 值如表 2 和表 3 所示。分类的结果如表 4 所示。

表 2　MAR 系数作为特征的类间 J 值

	SVT	APC	PVC	VT	VF
SVT	0	1.6587	1.3775	1.6287	2.8733
APC	1.6587	0	0.9669	1.5397	2.8077
PVC	1.3775	0.9669	0	1.1374	2.2122
VT	1.6287	1.5397	1.1374	0	1.0671
VF	2.8733	2.8077	2.2122	1.0671	0

表 3　K-LMAR 系数作为特征的类间 J 值

	SVT	APC	PVC	VT	VF
SVT	0	1.5056	1.2533	1.5232	2.6022
APC	1.5056	0	0.8182	1.3680	2.3288
PVC	1.2533	0.8182	0	1.0123	1.8658
VT	1.5232	1.3680	1.0123	0	0.9310
VF	2.6022	2.3288	1.8658	0.9310	0

表 4　分类结果

特征组	类别	SVT	APC	PVC	VF	VT
基于 MAR 系数	精度/%	99.0	98.3	99.6	100	100
基于 K-L MAR 系数	精度/%	98.1	97.2	96.3	98.3	97.4

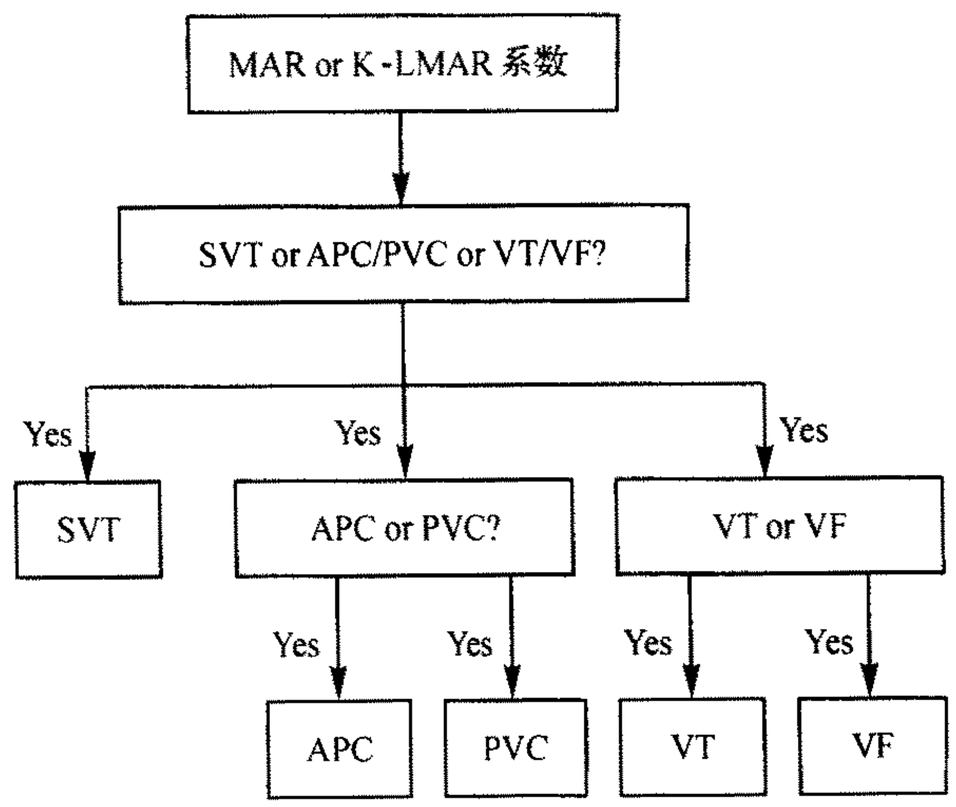

图 1　基于可分性评价准则 J 的心率失常分类决策树

4　讨　论

从表 1 可见，分类精度随着 k 值的增加而提高。因此，基于欧氏中心距和标准差的可分性评价准则不仅具有与分类结果保持一致的特性，而且具有计算简单、实现容易的优点。现有的资料检索结果表明，在 K-L 变换时，选取特征向量常用的准则是取前 k 个最大特征值的特征向量，使得样本在前 k 个轴上的能量占整个能量的 99%以上[6~10]。但是在实验 1 中，这样的准则却不能保证得到满意的结果。从表 1 可知，在 k 取 50 时，即取前 50 个最大特征值的特征向量进行 K-L 变换，样本在前 50 个特征向量轴上的能量已达到整个能量的 99%以上，但是 VT

和 VF 分类精度分别为 70.9%和 62.9%，这是不能令人满意的，而此时的 J 值是 0.165 4。但如果利用 J 值作为选取的准则，在 J 值达到 1.038 时，VT 和 VF 分类精度分别为 98.4%和 99.4%，而此时的 k 值应该为 400。也就是说，要达到理想的分类效果需要取前 400 而不是 50 个最大特征值的特征向量进行 K-L 变换。通过 K-L 变换，样本的特征向量维数只能从 650 降到 400。因此，本实验结果还表明，利用 K-L 变换不能始终得到分类和降低维数都令人满意的效果。需要考虑结合其他的手段来降低特征空间的维数，如本实验中可考虑先对信号进行压缩再行 K-L 变换等等。诚然，分类精度不是唯一的考虑因素，有时需要综合考虑，如训练和测试的样本数，训练进程的快慢等问题。

类似地，基于其他多种可分性准则的特征提取也都是采用一定形式的特征排序的方法[5]。如何选取特征向量仍然是个值得继续探讨的问题。

在实验 2 中，由于各类的协方差矩阵都是近奇异矩阵，在利用 Bhattacharyya 距离可分性准则 J_B 时遇到了计算和实现上的困难，而在利用基于欧氏中心距和标准差的可分性评价准则 J 却能有效地解决该问题。在本实验决策树的设计中，依据各类之间的 J 进行分组，J 值小的被分为同一组，决策过程如图 1 所示。实验结果表明，利用基于欧氏中心距和标准差的可分性评价准则 J 来构造决策树方便有效，并获得良好的分类效果。

另外，由分类结果表 4 可见，基于 MAR 系数的分类可取得比基于 K-L MAR 系数的分类稍好的结果，这一结果可从表 2 和表 3 中的 J 值得到一致的解释。因此，J 值可以灵敏有效地反映类间的可分性和特征的有效性。

5 结 论

基于欧氏中心距和标准差的可分性评价准则 J 应用于心电数据特征提取、特征有效性评测和决策树设计可取得良好的结果，J 值可以灵敏地反映类间的可分性和特征的有效性。并能有效地克服现存可分性评价准则实现困难和近奇异协方差矩阵计算困难的问题。

参考文献：

[1] 葛丁飞，李时辉. 基于 ARMA 模型的 ECG 分类和压缩[J]. 浙江科技学院学报，2004，16(1)：7－13.

[2] Chen S W. Two-stage discrimination of cardiac arrhythmias using a total least squares-based Prony modeling algorithm[J]. IEEE Trans Biomed Eng，2000，47：1317－1326.

[3] Poli S，Barbaro V，Bartolini P，*et al*. Prediction of atrial fibrillation from surface ECG：review of methods and algorithms[J]. Ann Ist Super Santita，2003；39(2)：195－203.

[4] Garcia J，Lander P，Sornmo L. Comparative study of local and Karhounen-Loeve based ST-T indexes in recordings from human subjects with induced myocardial ischemia[J]. Comput Biomed Res，1998，31：271－297.

[5] Bian S Q，Zhang X G. Pattern recognition[M]. Beijing：Qinghua University Publication，2002. 176－197，212－223.

[6] Ge D F，Xia S R. Application of AR Model in Telediagnosis of Cardiac Arrhythmias[J]. Chinese Journal of Biomedical Engineering，2004，23(3)：222－229.

[7] Olmos S，Millan M，Garcia J. ECG data compression with the Karhunen-Loeve transformation[J]. Computer in Cardiology，1996，5：253－256.

[8] Fukunaga K. Introduction to statistical pattern recognition[M]. London：Academicc Press Limited，1990. 40－41，98－99，446－450.

[9] Ge D F，Shao Y Q. An algorithm study on telecardiogram diagnosis based on multivariate autoregressive model and two-lead electrocardiogram signals[J]. Space Medicine & Medical Engineering，2004，17(5)：355－359.

[10] Fujimura S，Kiyasu S. Application of feature extraction scheme to the discrimination of electrocardiogram[J]. T IEE Japan，2001，21(8)：725－730.