

基于独立成分分析和流形学习的眼电伪差去除

高军峰, 郑崇勋, 王沛

(西安交通大学生命科学与技术学院, 710049, 西安)

摘要: 针对眼电伪差严重干扰脑电 (EEG) 信号的理解和分析的问题, 提出了一种新的方法用于实时地去除脑电中的眼电伪差. 该方法使用独立成分分析 (ICA) 分解 EEG 信号, 提取独立成分的地形图和功率谱作为特征, 并采用基于模板的 Isomap 算法降低特征的维数. 将新的特征样本送到分类器中以识别眼电伪差独立分量, 几个典型分类器的分类结果显示, 基于模板的 Isomap 算法结合使用最近邻算法进行分类时, 识别伪差的正确率最高. 实验结果表明, 提出的方法在有效去除眼电伪差的同时, 很好地保留了大脑神经信号, 也证明了新的 Isomap 算法用于眼电伪差特征的降维的有效性.

关键词: 流形学习; Isomap 算法; 脑电; 独立成分分析; 主成分分析

中图分类号: R318 **文献标志码:** A **文章编号:** 0253-987X(2010)02-0113-06

Real-Time Removal of Ocular Artifacts from EEG Signals Using ICA and Manifold Algorithm

GAO Junfeng, ZHENG Chongxun, WANG Pei

(School of Life Sciences and Technology, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Aiming at the problem that frequent occurrences of ocular artifacts seriously interfere with the electroencephalogram (EEG) interpretation and analysis, a novel technique to eliminate ocular artifacts from EEG signals in real-time is proposed. The independent component analysis (ICA) is employed to decompose EEG signals, and these independent components features of topography and power spectral density are extracted. Specifically, a template-based isometric mapping (Isomap) algorithm is adopted to reduce the feature dimensionality. The low-dimensional feature samples are fed to a classifier to identify ocular artifacts components. The classification performances of several typical classifiers show that the template-based Isomap algorithm with the nearest neighbor classifier performs best. The experimental results demonstrate the efficiency for removing ocular artifacts with little distortion of underlying brain signals.

Keywords: manifold learning; Isomap algorithm; electroencephalogram; dependent component analysis; principal component analysis

脑电 (EEG) 信号频繁地受到眼电伪差 (OA) 的干扰. 眨眼伪差和眼动伪差是眼电伪差的主要来源, 它们的幅度可以是自发脑电的几倍甚至几十倍, 一些文献已经证实它们有比较明显的低频特性和脑电地形图特征^[1]. 由于眼电伪差严重影响了脑电信号的分析和理解, 因此许多研究人员提出了各种伪差

去除的方法. 最简单的一种方法是通过观察眼电信号去除被伪差干扰的信号部分, 但是此方法对于污染严重的信号无能为力. 另外一个广泛使用的方法是基于时域或者频域的回归方法. 该方法主要基于离线分析, 而且它往往需要一个参考通道, 而眼电 (EOG) 参考通道通常不可避免地包含了脑电神经

信号,所以该方法也不是理想的 OA 去除方法。

最近一些作者证明了利用独立成分分析(ICA)方法去除 OA 的有效性^[1-2]。在 ICA 方法中,观察时间序列 $\mathbf{X}(t)=[x_1(t), x_2(t), \dots, x_K(t)]^T$, $t=1, 2, \dots, N$, 其中 N 是抽样点数, K 是感知器个数。对于 ICA 方法而言, $\mathbf{X}(t)$ 被看作是未知的 K 个独立源 $\mathbf{S}(t)=[s_1(t), s_2(t), \dots, s_K(t)]^T$ 的线性混合 $\mathbf{X}(t)=\mathbf{A}\mathbf{S}(t)$ 。ICA 方法试图通过引入解混矩阵 $\mathbf{W}$ 来计算未知的 $\mathbf{S}(t)$ 信号源。理想情况下, $\mathbf{W}$ 是混合矩阵 $\mathbf{A}$ 的逆,所以 $\mathbf{W}^{-1}$ 也被称为混合矩阵。

近年来,基于 ICA 方法的一些自动化的伪差去除技术被陆续报道。这些方法抽取独立成分的一些典型特征进行 OA 分量的自动化辨识^[3-4],但是在很多情况下,特征集合中的很多特征都具有冗余性^[5],比如 Halder^[3]提取的功率谱和地形图特征。如果能发现那些实际控制特征样本的低维特征,无疑对分类更有利。本文使用了一种非线性的降维算法——流形学习(ML)^[6]来揭示高维特征空间的低维流形结构,并借此降低特征维数。值得一提的是,目前 ML 主要用在数据可视化和离线的降维分析。即使少数作者把流形算法用在了模式识别邻域,流形学习也只是应用在批处理的方式上,而不能用来处理在线的流数据。传统的方式往往使用了太多的样本构建流形,因此速度因素影响了流形算法的在线推广。本文选择了 Isomap^[6]算法并且提出了一个基于模板的 Isomap 的使用方式,利用该改进方式,Isomap 算法可以实现在线降维,结合 ICA 方法,能更有效地去除 OA。

1 方法

1.1 传统 Isomap 算法

系统复杂的特征不仅增加分类算法的处理时间,而且会降低分类准确率,目前已有许多降维算法用于解决此问题。主成分分析(PCA)是使用最广泛的技术之一,可是此技术最大的缺点就是假设特征样本必须存在于一个线性子空间。多维尺度变换(MDS)是另外一个广泛使用的降维方法,它在降维的同时保持了输入特征空间的欧式距离不变。

图 1a 显示了一个被称为“swiss roll”的曲面,尽管曲线上的点都位于 3 维空间,但是这个形状可以被 2 个变量参数化。在数学上,我们称在该曲面中存在一个二维的流形结构,能揭示这个流形结构的算法称为流形算法,它被看作是 PCA 算法的非线性对比算法和对 MDS 算法的扩展。本文采用了基于全

局的 Isomap 方法,它能在特征样本数据比较有限的情况下仍有比较好的稳定性。Isomap 算法构建了基于测地线距离(在流形结构中的点间距离)的嵌入型^[6]。在图 1a 中,一条曲线和一条虚线连接两个特征点,前者代表两点在曲线上的实际距离,反映了两个点的内在相似性,后者代表欧式距离,显然它不能反映点间的内在相似性^[7]。在此种情况下,如果要求保持特征点的相似性不变而进行特征空间的降维,PCA 和 MDS 算法无法完成这一任务,而 Isomap 算法可以有效地解决此问题。令 $\mathbf{X}=[x_1, x_2, \dots, x_n]^T$ ($x_i \in \mathbf{R}^D$) 代表 D 维输入空间的所有特征向量, n 是特征点个数。令 $[y_1, y_2, \dots, y_n]^T$ ($y_i \in \mathbf{R}^d$, $d < D$) 代表由 ML 算法计算得到的流形结构,即降维空间的特征向量。ML 学习算法的步骤表述如下。

(1)构建邻域图。使用 p -Isomap 选择某个点的 p 个邻域点,即如果某个点 i 是另外一个点 j 的 p 个最近的点之一,就认为此两点直接相连,并计算两点的欧式距离 $d_x(i, j)$; 否则认为此两点断开, $d_x(i, j) = \infty$ 。依此,可以得到一个基于 p -最近邻算法的邻域图 G 。

(2)流形结构中任意两点测地线距离 $d_M(i, j)$ 的估计。基于邻域图 G ,首先初始化并令 $d_G(i, j) = d_x(i, j)$,然后用最短路径算法(Floyd-Warshall 算法)计算最短路径,即

$$d_G(i, j) = \min\{d_G(i, j), d_G(i, k) + d_G(k, j)\} \quad (1)$$

最后的 $d_G(i, j)$ 即为测地线距离 $d_M(i, j)$ 的估计。存储这些距离到矩阵 $\mathbf{C} \in \mathbf{R}^{n \times n}$ 中,将该矩阵中心均值化,即

$$\mathbf{B} = -\frac{1}{2}\mathbf{HCH} \quad (2)$$

$$\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{\Gamma}\mathbf{\Gamma}^T \quad (3)$$

式中: $\mathbf{\Gamma}$ 是 $n \times 1$ 的单位向量。

(3)对矩阵 $\mathbf{B}$ 进行如下的谱分解

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{Y}\mathbf{Y}^T \quad (4)$$

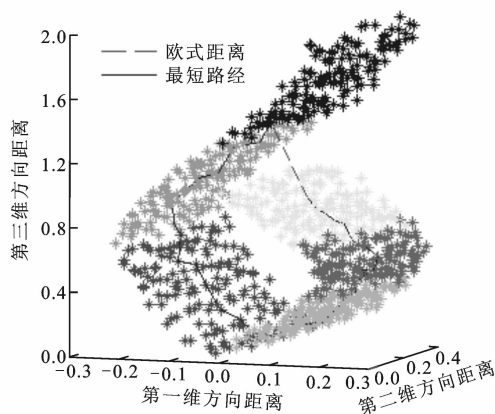
式中: $\mathbf{\Lambda}$ 是对角矩阵。考虑到在 $\mathbf{\Lambda}$ 中可能有负的对角元素值,因此由以下步骤计算 $\mathbf{Y}$ 。

(4)令 $[\mathbf{\Lambda}_+]_{ij} = \max\{\mathbf{\Lambda}_{ij}, 0\}$, 所以

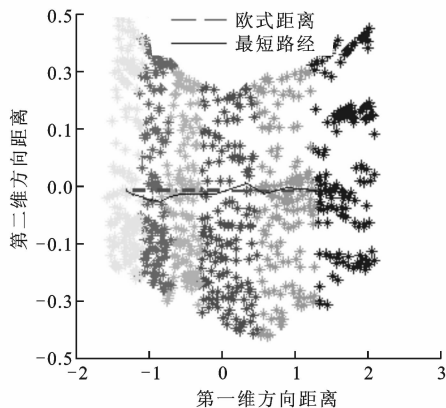
$$\mathbf{Y} = \mathbf{U}\mathbf{\Lambda}_+^{1/2} \quad (5)$$

最后,令 $\mathbf{Y}=[\mathbf{Y}]_{n \times d}$, 即选择 $\mathbf{Y}$ 的前 d 列更新 $\mathbf{Y}$, 结果即为矩阵 $[y_1, y_2, \dots, y_n]^T$ 。

对于图 1a 的曲面,经过以上算法计算后的流



(a)“swiss roll”曲面



(b)“swiss roll”曲面经过 ML 算法计算后的表示

图1 “swiss roll”曲面以及在该曲面应用 ML 算法得到的二维流形结构

形结构如图 1b. 很明显,最短路径距离较好地近似表示了测地线距离(相对值),因此通过以上算法可以很好地保持高维空间的特征点间的相似性.

1.2 FT-Isomap 算法

考虑到本文研究实际是基于有导师的机器学习问题,即在分类 OA 之前,可以获得大量有类标示的眼电伪差和非眼电伪差的特征样本(过程见第 2 节). 基于这样的两个集合,改变传统的 Isomap 算法的使用方式,提出一个基于模板的新的使用方式. 以下将新方法称为 FT-Isomap 算法. 该算法步骤描述如下:

(1)分别选取 m 个眼电伪差和 m 个非眼电伪差作为特征向量模板;

(2)选取另外的 n 个测试样本,采用 Isomap 算法,建立 $2m+n$ 个特征向量样本的流形结构;

(3)应用已训练好的分类器识别 OA. 参数 m 、 n 和其他 Isomap 算法中的参数取值需要通过交叉验证的训练过程获得,具体训练和测试方法将在 2.3

节详述.

2 实验

2.1 EEG 数据采集

实验对象为 16 个健康受试者(8 名男性,8 名女性,平均年龄 22 岁),他们被要求做 2 min 的实验. 电极按照 10-20 国际标准放置在 Fp1、Fp2、F3、Fz、F4、T7、C3、Cz、C4、T8、P3、Pz、P4、O1、Oz、O2 的 16 个通道上,按照标准采集垂直和水平眼电信号. 使用 Neuroscan 放大器,带通滤波为 0.1~45 Hz,抽样率为 500 Hz,使用右耳参考电极. 在实验过程中,两个箭头随机地以 3 s 的间隔显示在屏幕的两边,受试者被要求随着箭头出现而左右移动眼球并眨眼,随后 EEGLAB^[8] 将 EEG 数据分为 640 个数据集,每个数据集持续 3 s.

2.2 ICA 算法和特征提取

ICA 算法将采集信号盲分解为若干个独立成分^[9],使用 EEGLAB 默认的 Extended Infomax algorithm^[3] 来分解 EEG. 一些作者已经指出眼电伪差成分(OAC)具有典型的功率谱和地形图,EEG 专家通过观察这些特征来识别眼电伪差^[1]. 为了在线识别 OAC,提取 IC 特征,采用 Bartlett 算法计算功率谱,对功率谱曲线从 1~25 Hz 每隔 0.5 Hz 抽取一个值,这样得到的 49 个值作为功率谱特征. 混合矩阵 W^{-1} 的一列即为对应的 IC 地形图特征,最后的特征向量即为功率谱和地形图特征的合并,特征向量的维数是 67 维. 另外,所有的数据集送交一个神经生理专家,由他根据一些典型的特征(时域波形和功率谱等)挑选出典型的 OAC. 观察图 2 中的分量的特征,将 IC1 和 IC2 选为 OAC. 最后,将 1006 OACs 和 1006 non-OACs 选为两个原始的特征向量集合.

2.3 训练和参数选择

当 Isomap 算法用于分类问题时,尤其适合与 k 最近邻算法(KNN)结合^[10],称为 FT-Isomap + KNN 综合方法. 设降维参数为 d ,Isomap 算法中用于构建邻域图的近邻参数为 p ,通过以下训练过程可以得到最佳参数的取值:

(1)从两个向量集合中随机抽取 m 个眼电伪差和非眼电伪差的特征向量样本,并将集合中剩下的所有样本分为 n 个样本为一组的若干测试组;

(2)对 $2m+n$ 个样本运行 FT-Isomap 算法;

(3)使用 KNN 算法,在降维空间对 n 个测试样本分类,算法中的参数 k 设置为 9;

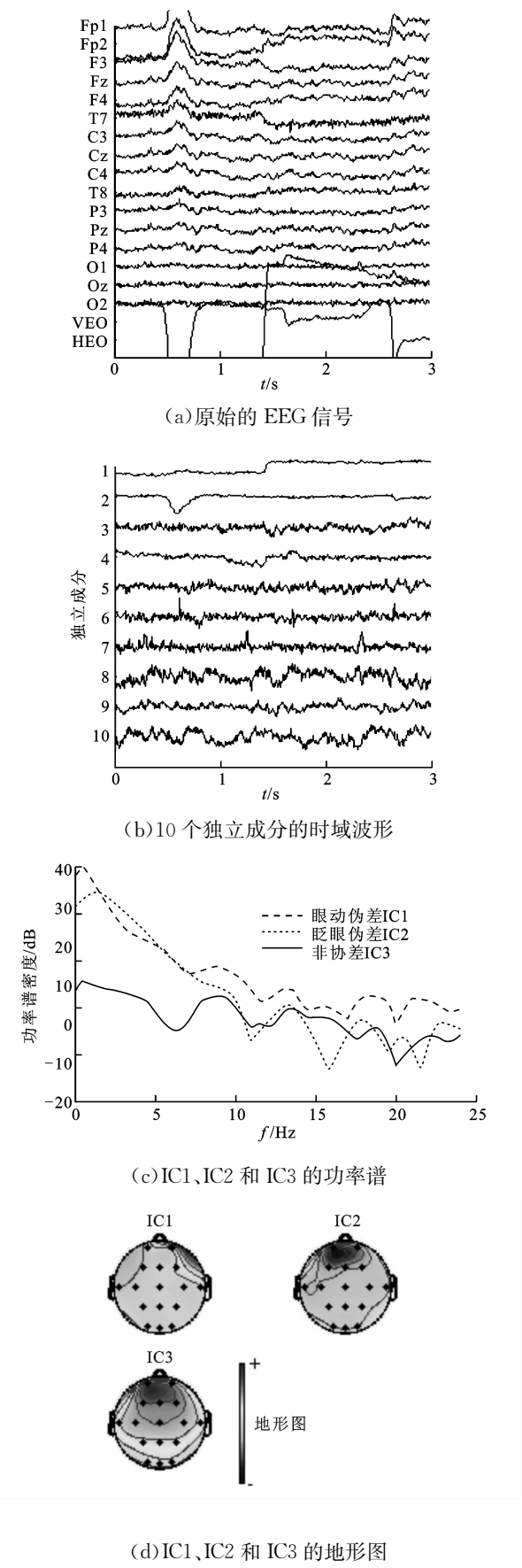


图 2 EEG 信号和 ICA 分解的独立成分的各种特征

(4)选择另外的测试组,重复以上的(2)、(3)步骤,然后对所有测试组的准确率进行平均;

(5)选取另外的 $2m$ 个特征模板,对应得到测试集合,重复以上 4 个步骤,使用类似于交叉验证(CV)的方法,即可得到当前参数组合(d 、 p 、 m 和 n)下的分类准确率;

(6)选择其他的参数重复以上 5 个步骤.

通过训练和测试可以得到最高的分类准确率和对应的参数组合,此时的分类准确率即为 FT-Isomap+KNN 方法在最佳参数下的最佳准确率.另外,为了和非线性降维方法对比,选取 PCA 作为典型的线性降维方法,选择背向传输神经网络作为分类器.类似地,通过 5 折交叉验证的训练和测试方法,也得到 PCA+NN 在最佳参数下的最高分类准确率.为了与使用降维技术的识别方法相比较,选取了 Fisher 线性分类和神经网络(NN)方法对样本进行非降维地直接分类.

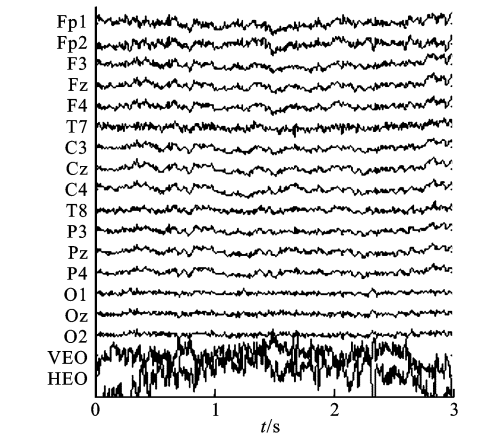
3 结 果

首先给出几种分类方法的最佳参数:①对于 FT-Isomap+KNN 方法, $m=30$, $n=6$, $d=2$, $p=5$;②对于 PCA+NN 方法, $d=5$, 隐层节点数等于 5, 学习率等于 0.2, 输出控制精度 $\epsilon=0.001$;③对于 NN 方法, 隐层节点数等于 5, 学习率等于 0.1, $\epsilon=0.001$.

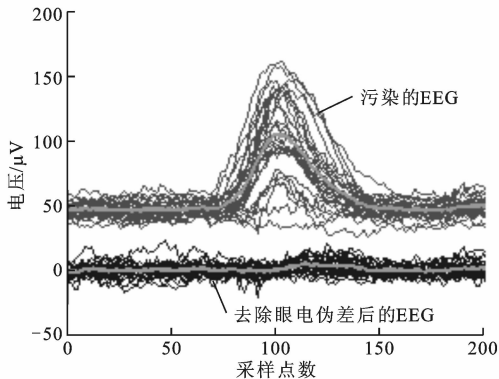
在上述参数下,几种分类方法的最佳分类准确率在表 1 中给出.表 1 显示 FDA 的分类率最差,而 FT-Isomap+KNN 方法的准确率最高,NN 方法达到了 92.42%的平均准确率,这个结果远高于 FDA 方法的 81.77%,说明了特征数据并不处于一个线性的子空间中.我们还发现,PCA+NN 方法的分类准确度低于 NN 方法.这个结果应该是因为 PCA 的假设,即特征数据必须位于线性子空间中,因此使用 PCA 降维后,准确率会下降.

表 1 不同分类方法的分类准确率

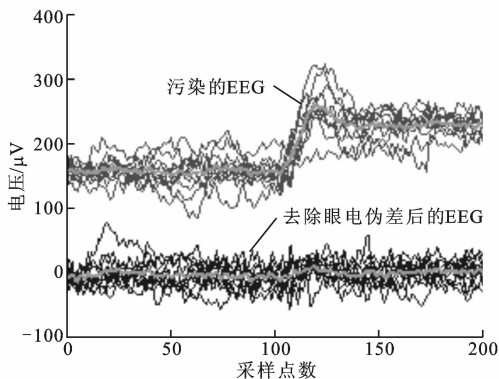
分类方法	分类准确率/%		
	敏感度	特异度	平均
FDA	80.42	83.12	81.77
NN	91.31	93.54	92.42
PCA+NN	90.14	92.86	91.50
FT-Isomap+KNN	99.53	98.05	98.79



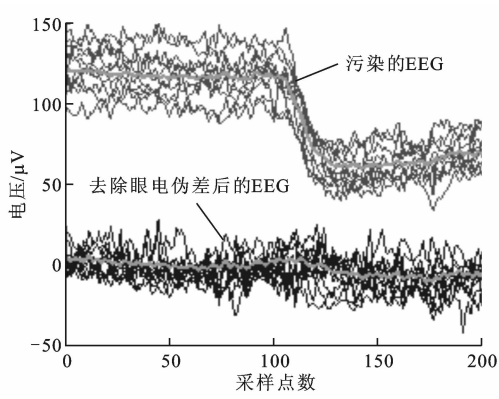
(a) 伪差去除后的 EEG 信号



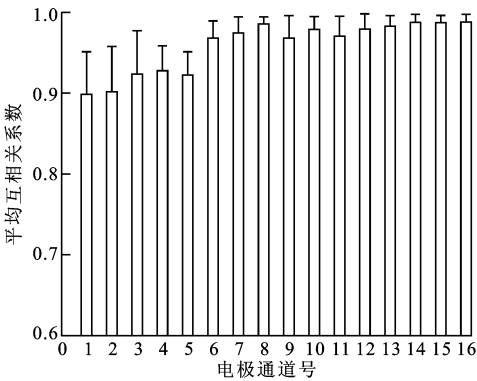
(b) 眨眼伪差干扰去除效果



(c) 眼动伪差干扰(漂移的方向向上)去除效果



(d) 眼动伪差干扰(漂移的方向向下)去除效果



(e) 16 个通道上的平均 CC 值及相应的标准差

图 3 眼电伪差去除效果评价

使用本文提出的方法去除图 1a 中显示的 EEG 信号的眼电伪差, 去除伪差后的 EEG 显示在图 3a 中. 对比图 1a 和图 3a 的 EEG 信号, 可以发现眼电伪差基本已被去除. 为了进一步显示多个数据集的去除伪差的总体效果, 从每个受试者随机选取 8 个数据集, 然后应用本文方法去除伪差(使用了 30 个眼电和 30 个非眼电伪差特征向量作为模板), 最后将伪差去除前后的信号片段显示在图 3b~图 3d, 可以看到, 伪差去除的效果基本满足在线应用的要求.

另外, 使用互相关系数 (CC) 客观评价大脑保持神经信号的能力. 首先观察去除 EEG 信号中被污染的信号片段, 比如对于图 1a 和图 3a, 同时去除前 13 个通道的 200 ms 到 500 ms 的信号片段, 然后在伪差去除前后的 EEG 信号的相同通道上计算相关值. 上述 120 个数据集的 16 个通道的 CC 值显示见图 3e. 在此图中, 我们可以看到前 3 个电极通道上的均值和标准差基本相等, 而且均值低于其他通道, 标准差也大于其他通道. 这与几个通道上受到的眼电污染最严重的实际情况是一致的. 重要的是, 其他通道的 CC 均值都比较高, 标准差也都比较小. 这些结果说明, 本文提出的方法比较好地保持了大脑神经信号.

4 讨 论

FT-Isomap 算法基于传统的 Isomap 算法, 两者的不同之处在于前者是以一些特征模板和需要测试的特征样本构建流形, 这就使得 FT-Isomap+KNN 能够使用在有导师的模式识别领域中.

在线应用时, 本文方法作为在线去伪差系统的一个滤波器, 工作在 3 s 的 EEG 信号上. 对于一个在线系统, 比如 BCI 系统, 一次输出的抽样块是 100 个抽样点, 伪差去除以后, 时间窗移动 100 个样本点

继续下一次伪差去除. 100 个抽样点在本文的抽样率下处理时间是 200 ms, 采用 ICA 和 FT-Isomap 算法去除眼电伪差的运行时间大致是 70 ms 和 60 ms, 可见时间上可以满足在线应用的要求. 另外, 由于本文提出的方法是对 EEG 做分段处理的, 因此伪差处理后 EEG 在分段处信号是不连续的, 在做 EEG 的其他分析时, 也必须进行分段处理^[11].

参考文献:

- [1] JUNG T P, MAKEIG S, WESTERFIELD M, et al. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects[J]. *Clinical Neurophysiology*, 2000, 111(10): 1745-1758.
- [2] LEE T W. Independent component analysis: theory and applications[M]. Hingham, MA, USA: Kluwer Academic Publishers, 1998.
- [3] HALDER S, BENSCH M, MELLINGER J, et al. Online artifact removal for brain-computer interfaces using support vector machines and blind source separation[EB/OL]. <http://downloads.hindawi.com/journals/cin/2007/082069.pdf>. 2009-02-22.
- [4] SCHLOGL A, KEINRATH C, ZIMMERMANN D, et al. A fully automated correction method of EOG artifacts in EEG recordings[J]. *Clinical Neurophysiology*, 2007, 118(1): 98-104.
- [5] CAYTON L. Algorithm for manifold learning [M]. San Diego, USA: University of California, 2005.
- [6] TENENBAUM J B, SILVA V D, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. *Science*, 2000, 290(5500): 2319-2323.
- [7] SEUNG H S, DANIEL D L. The manifold ways of perception[J]. *Science*, 2000, 290(5500): 2268-2269.
- [8] DEIORME A, MAKEIG S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis[J]. *Journal of Neuroscience Methods*, 2004, 134(1): 9-21.
- [9] COMON P. Independent component analysis, a new concept? [J]. *Signal Processing*, 1994, 36(3): 287-314.
- [10] VIACHOS M, DOMENICONI C, GUNOPILOS D, et al. Non-linear dimensionality reduction techniques for classification and visualization[C]// *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM, 2002: 645-651.
- [11] SHAO S Y, SHEN K Q, ONG C J, et al. Automatic

EEG artifact removal: a weighted support vector machine approach with error correction[J]. *IEEE Trans Biomed Eng*, 2009, 56(2): 336-344.

[本刊相关文献链接]

- 心算任务中与自主神经有关的头皮脑电活动研究, 2009, 43(10): 124-128.
- 基于多导脑电复杂性测度的脑疲劳分析, 2008, 42(12): 1555-1559.
- 基于 Fisher 判据时频分析的运动相关脑电特征选择及优化, 2008, 42(8): 1026-1030.
- 基于近似熵的认知能力对事件相关电位的影响研究, 2008, 42(10): 1300-1303.
- 基于脑电功率谱-连续隐马尔科夫链的精神疲劳分级模型, 2007, 41(12): 1474-1478.
- 基于样本熵的注意力相关脑电特征信息提取与分类, 2007, 41(10): 1237-1241.
- 基于功率谱分析的头皮电位和心率变异性关系研究, 2007, 41(8): 991-994.
- 基于多通道线性描述参数的生理性精神疲劳分析, 2007, 41(6): 737-740.
- 基于多导脑电特征的生理性精神疲劳分析, 2007, 41(2): 250-254.
- 基于可生长自组织映射的意识任务分类, 2006, 40(10): 1152-1156.
- 探测深度对心脏光学标测中荧光信号的影响, 2006, 40(10): 1157-1160.
- 一种新的注意力相关脑电分类算法设计, 2005, 39(10): 1162-1164.
- 基于多通道脑电特征运动意识任务的分类, 2005, 39(8): 904-907.
- 基于非线性参数的意识任务分类, 2005, 39(8): 900-903.
- 实时脑电信号眼电伪差去除方法的研究, 2004, 38(12): 1306-1309.
- 利用双谱分析的癫痫脑电特征研究, 2004, 38(10): 1097-1100.
- 快速多变量自回归模型的意识任务的特征提取与分类, 2003, 37(8): 861-864.
- 基于多通道时频相干的诱发电位单次提取, 2003, 37(6): 638-641.

(编辑 杜秀杰)