本文引用格式：凌宇,杜玉晓,李向欢.基于 F-Score 特征选择的癫痫脑电信号识别方法[J].自动化与信息工程,2023,44(5):58-62;73.

LING Yu, DU Yuxiao, LI Xianghuan. Epileptic EEG signal recognition method based on F-Score feature selection[J]. Automation & Information Engineering, 2023,44(5):58-62;73.

基于 F-Score 特征选择的癫痫脑电信号识别方法*

凌宇　杜玉晓　李向欢

（广东工业大学，广东 广州 510006）

摘要：随着癫痫脑电信号自动检测算法研究地不断深入，需要处理的特征维度也不断增加，且冗余特征增大了算法的复杂度，导致算法性能下降。为此，提出一种基于 F-Score 特征选择的癫痫脑电信号识别方法。首先，从原始癫痫脑电信号数据集中提取特征，并计算每个特征的 F-Score 统计值；然后，根据分类模型的分类准确率，通过序列前向搜索方法，选择最优特征集；最后，利用支持向量机和逻辑回归分类模型进行实验，并与传统的特征降维方法 PCA 进行对比。实验结果表明，本文方法可有效降低特征矩阵的维数，提高算法运算效率。

关键词：F-Score；PCA；特征提取；特征选择；癫痫脑电信号识别

中图分类号：R742.1　　**文献标志码：**A　　**文章编号：**1674-2605(2023)05-0009-06

DOI：10.3969/j.issn.1674-2605.2023.05.009

Epileptic EEG Signal Recognition Method Based on F-Score Feature Selection

LING Yu　DU Yuxiao　LI Xianghuan

(Guangdong University of Technology, Guangzhou 510006, China)

Abstract: With the continuous deepening of research on automatic detection algorithms for epileptic EEG signals, the number of feature dimensions to be processed continues to increase, and redundant features increase the complexity of the algorithm, leading to a decrease in algorithm performance. To this end, a method for epileptic EEG signal recognition based on F-Score feature selection is proposed. Firstly, extract features from the original epileptic EEG signal dataset and calculate the F-Score statistical value for each feature; Then, based on the classification accuracy of the classification model, the optimal feature set is selected through a sequence forward search method; Finally, experiments were conducted using support vector machines and logistic regression classification models, and compared with the traditional feature dimensionality reduction method PCA. The experimental results show that the proposed method can effectively reduce the dimensionality of the feature matrix and improve the computational efficiency of the algorithm.

Keywords: F-Score; PCA; feature extraction; feature selection; epileptic EEG signal recognition

0　引言

目前，癫痫的临床诊断主要以脑电图（electro-encephalogram, EEG）为依据。随着计算机技术的飞速发展，人们开始利用计算机处理癫痫脑电信号。计算机处理癫痫脑电信号的基本原理是提取癫痫脑电信号的特征并进行分类[1]，应用较多的分类方法是机器学习算法。在机器学习算法中，理论上认为特征越多，分类性能就越好。然而，大量特征可能存在冗余，降低分类模型的准确率。机器学习算法的基础是特征选择，从原始数据特征集中筛选出最优特征子集，可降低特征矩阵的维度[2]，提高算法的运算效率。目前，常用的特征选择方法可分为过滤式和封装式[3]。其中，过滤式方法通过设置阈值对特征评价进行筛选；封装式方法通过机器学习算法来寻找特征评价。常用的特

* 基金项目：国家自然科学基金（61976059，61640213）

征评价标准有相关系数[4]和互信息[5]。文献[6]先利用极限学习机对非线性特征进行评价，再利用多目标演化算法来筛选最优子集。

为全面反映癫痫脑电信号，需要从原始脑电信号中提取多个维度的特征，包括时域、频域、时频域和非线性特征[7]，导致原始癫痫脑电信号特征集中有许多冗余特征。为此，本文提出一种基于 F-Score 特征选择的癫痫脑电信号识别方法。首先，利用 F-Score 对原始脑电信号的特征进行评价；然后，采用序列前向搜索方法，以分类模型的分类准确率为反馈来寻找最优的特征子集。

1 特征提取与特征选择算法

1.1 PCA 特征降维

主分量分析（principal component analysis, PCA）是一种常用的数据降维方法[8]，它将原始数据集中的多维特征映射到低维空间，从而减少数据的维度。PCA 可以减少计算量，提高算法的运算效率，消除噪声，提高模型的泛化能力；但可能丢失重要的特征信息，影响算法的准确率。

1.2 F-Score 特征选择

F-score 是一种通过计算类内间距来衡量特征分类能力的方法，由 HUANG 等[9]于 2006 年提出，可有效实现特征选择。假设X_k（$k=1,2,3,\cdots,N$）为特征集中的所有特征，n^+、n^-为特征样本的正负实例数量，则特征集中的第i个特征的 F-Score 计算公式[10]为

$$F_i=\frac{\left(\bar{x}_l^{+}-\bar{x}_l\right)^2+\left(\bar{x}_l^{-}-\bar{x}_l\right)^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}\left(x_{k,i}^{+}+\bar{x}_i^{+}\right)^2+\frac{1}{n^--1}\sum_{k=1}^{n^-}\left(x_{k,i}^{-}+\bar{x}_i^{-}\right)^2} \tag{1}$$

式中：$\bar{x}_l$为该特征样本在特征集中的平均值，$\bar{x}_l^{+}$、$\bar{x}_l^{-}$分别为该特征样本在正负样本上的平均值，$x_{k,i}^{+}$、$x_{k,i}^{-}$分别为第 k 个正负类样本在第 i 个特征上的值。

在进行特征选择时，采用序列前向搜索方法来搜索最优特征子集，即依据一定的评价标准，从原始特征集中选择可以使评价标准最好的特征加入到最优特征子集 S 中。

1.3 癫痫脑电信号特征提取

特征提取是癫痫脑电信号分类算法的重要步骤，提取特征的好坏直接影响癫痫脑电信号分类模型的分类效果[11]。本文提取的癫痫脑电信号特征包括峰度和偏态、Hjorth 参数、功率谱密度、香农熵、谱熵、近似熵等。

1） 峰度和偏态[12]。峰度（Kurtosis, K）用来表示采样点分布的陡峭程度，在波形图上描述波峰尖锐程度的特征。偏态（Skewness, S）用来表示信号的偏斜程度，与正态分布进行比较，往左偏是左偏态，往右偏是右偏态。峰度和偏态的计算公式分别为

$$K=\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^4/SD^4 \tag{2}$$

$$S=\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^3/SD^3 \tag{3}$$

式中：SD为信号标准差，$\bar{X}$为信号均值。

2） Hjorth 参数[13]包括活动性（Activity）、移动性（Mobility）和复杂性（Complexity）3 个参数，计算公式为

$$A_{\text{activity}}=SD^2 \tag{4}$$

$$M_{\text{mobility}}=\frac{SD'}{SD} \tag{5}$$

$$C_{\text{complexity}}=\sqrt{\frac{SD''}{SD'}\Big/\frac{SD'}{SD}} \tag{6}$$

式中：SD'为一阶差分信号的标准差，SD''为二阶差分信号的标准差。

3） 功率谱密度（power spectral density, PSD）[14]用来表示信号的能量特征，计算公式为

$$PSD=\frac{\sum_{i=1}^{N}X_i}{f_s} \tag{7}$$

式中：X_i为癫痫脑电信号，f_s为脑电信号的频率。

4） 香农熵（ShEn）[15]又被称为信息熵，可以对

信号中包含的信息量进行度量，而信息量大小可以表示信息的不确定程度，计算公式为

$$ShEn = -\sum_f P_h(X)\ln P_h(X) \tag{8}$$

式中：P_h为概率密度函数的周期图估计。

5） 谱熵（SpEn）[16]用于量化信号的规律性和顺序性，计算公式为

$$SpEn = -\frac{1}{\log N_f}\sum_f P_f(X)\ln P_f(X) \tag{9}$$

式中：N_f为频率分量个数，P_f为概率密度函数。

6） 近似熵（ApEn）[17]用来表示脑电信号的复杂性和规律性，计算公式为

$$ApEn = \Phi^{d_E-1}(r) - \Phi^{d_E}(r) \tag{10}$$

式中：$\Phi^{d_E}(r)$为相邻两点之间的相似度，r为相似距离。与其他特征相比，近似熵具有抗干扰性强、数据量依赖较小等特点。

1.4 分类模型

癫痫脑电信号分类算法的分类模型采用支持向量机（support vector machine, SVM）和逻辑回归（logistic regression, LR）机器学习分类模型。

1） SVM 模型在解决高维问题以及小样本的分类问题中应用广泛，其可将线性不可分的问题映射到高维空间，并在高维空间中寻求一个超平面，将低维线性不可分问题转换为高维空间的线性可分问题[18]。但在高维空间中如何计算内积成为一个难题，为此 SVM 引入核函数，将在高维空间中的内积运算转化为低维空间中的输入核函数运算。本文采用的核函数为径向基函数，即

$$K(x_i, x_j) = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right), g > 0 \tag{11}$$

2） LR 模型用来预测输出变量的取值或者事件发生的概率[19]，通常用来解决分类问题，其表示简单、准确率高，模型中使用正则化项能够避免过拟合。LR 的表达式模型是一个线性函数，其输入特征与对应的权重向量相乘，再使用 logistic sigmoid 函数将结果映射到 0~1 之间的概率空间。

1.5 算法步骤

假设从原始癫痫脑电信号中提取的特征集为$F = (f_1, f_2, \ldots, f_i)$输出的最优特征子集为 S，基于 F−Score 的特征选择算法流程如图 1 所示。

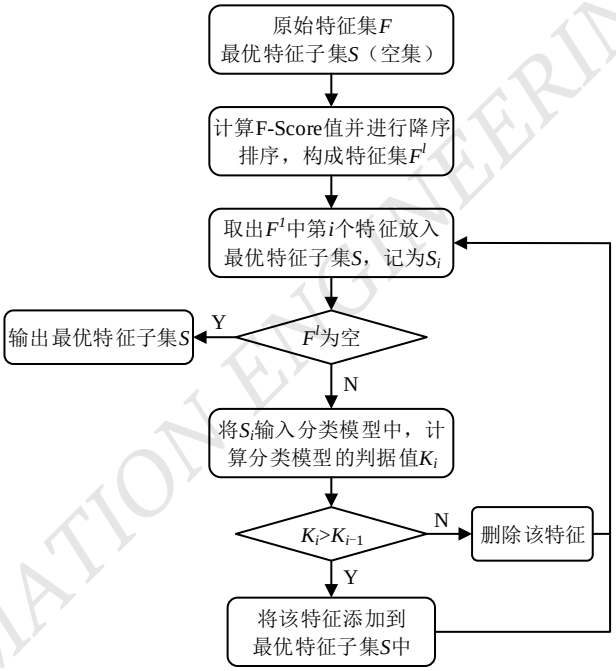

图 1 基于 F-Score 的特征选择算法流程图

基于 F-Score 的特征选择算法具体步骤为：

1） 对特征集 F 中的每个特征进行基于 F-Score 算法的特性评价，计算每个特征的 F-Score 值；

2） 将每个特征的 F-Score 值降序排序，重新构建特征集 F^1；

3） 每次从特征集 F^1 中取出 F-Score 值最大的特征放入最优特征子集 S，如果特征集 F^1 为空，算法结束，否则继续执行下一步；

4） 将特征子集 S 输入到分类模型中进行分类，以分类模型的 K 为判据；假设当前的特征集为 S_i，分类模型的判据值为 K_i，从 F^1 中取出当前 F-Score 值最大的特征加入 S_i 中，记为 S_{i+1}，同样计算 S_{i+1} 的判据值 K_{i+1}；

5） 比较 K_i 与 K_{i+1}，如果 $K_{i+1} \le K_i$，表示这个特征对分类效果起不到正向促进作用，将这个特征从 S 中去除，并返回步骤 3）；如果 $K_{i+1} > K_i$，表示这个特

征可以提高分类效果，将这个特征保留在 S 中，并返回步骤 3）；

6） 直到遍历特征集 F^1 的所有特征，生成的特征集 S 即为最优特征子集。

2 实验结果及分析

本文实验仿真采用 MATLAB 实现。实验对比 PCA 和 F-Score 2 种特征选择算法在 SVM 和 LR 2 种分类模型上的性能。

2.1 实验数据

本文采用的 EEG 数据集来自伯恩大学的 Bonn 数据集。Bonn 数据集中包含 Set A、Set B、Set C、Set D、Set E 5 组数据，选取 Set A（正常脑电信号）和 Set E（癫痫脑电信号）2 组进行分类实验。Set A 和 Set E 脑电信号波形图如图 2 所示。

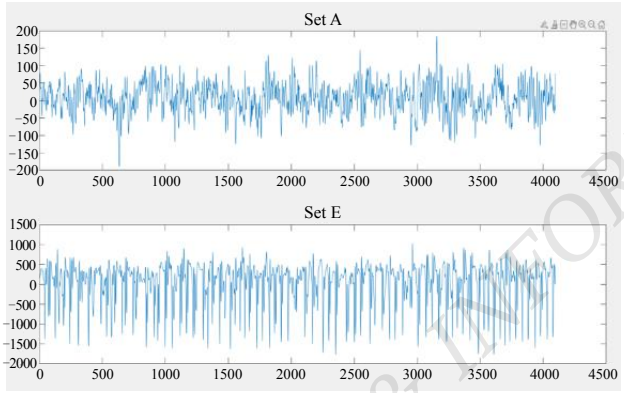

图 2 Set A 和 Set E 的脑电信号波形图

实验前，将每个 EEG 信号分成 4 个相等的部分，获得 400 个标准的 EEG 样本和 400 个癫痫发作样本，每个样本长度为 1 024。

2.2 实验结果分析

本文对比经过 PCA 和 F-Score 特征选择后的特征集，分别在 SVM 模型和 LR 模型的分类效果，实验流程如图 3 所示，特征选择的结果如表 1 所示。

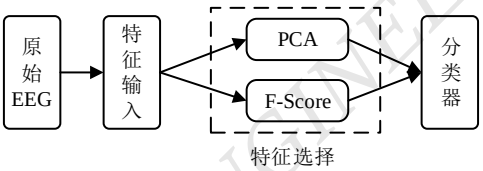

图 3 实验流程图

表 1 特征选择结果

	特征维数
原始特征	31
PCA	18
F-Score	15

本文选取准确率（Accuracy）、精确率（Precision）、特异性（Specificity）和敏感度（Sensitivity）4 个指标对分类模型进行评估。其中，准确率是模型正确预测的样本数量与总样本数量之比；精确率衡量模型在预测为正类的样本中的准确性；特异性衡量模型对于实际为负类的样本的预测能力；敏感度衡量模型对于实际为正类的样本的预测能力。分类效果如表 2 和表 3 所示。

表 2 SVM 模型分类效果

	准确率（Accuracy）	精确率（Precision）	特异性（Specificity）	敏感度（Sensitivity）
原始特征	85.31%	84.21%	87.12%	87.12%
PCA	84.3%	86.6%	89.56%	85.4%
F-Score	91.2%	90.23%	91.11%	89.99%

表 3 LR 模型分类效果

	准确率（Accuracy）	精确率（Precision）	特异性（Specificity）	敏感度（Sensitivity）
原始特征	83.31%	85.21%	82.42%	81.45%
PCA	84.2%	86.3%	86.4%	84.8%
F-Score	90.72%	89.74%	88.21%	86.79%

由表 2 和表 3 可以看出：原始特征经过特征选择后，分类模型的分类效果有一定提升，且 F-Score 特征选择算法的分类效果比 PCA 特征降维的效果更好。

原始癫痫脑电信号特征集为 31 维，经 F-Score 特征选择算法得到的最优特征子集为 15 维；经 PCA 特征降维后特征为 18 维，表明经过 F-Score 特征选择算法处理过后可有效降低特征集维度，减少分类模型计算的复杂度。

3 结论

本文提出基于 F-Score 特征选择的癫痫脑电信号识别方法，首先，采用原始 EEG 数据集中的 F-Score 统计特性对特征进行评价，并结合序列前向搜索方法搜寻最优特征子集，在搜索过程中采用分类性能评价所选择的特征子集。该特征选择方法能够选择出优化的特征子集，降低数据维数和计算复杂度，进一步提高分类器的性能。

参考文献

[1] YILDIZ A, ZAN H, SAID S. Classification and analysis of epileptic EEG recordings using convolutional neural network and class activation mapping[J]. Biomedical Signal Processing and Control, 2021, 68:102720.

[2] CAI J, LUO J, WANG S, et al. Feature selection in machine learning: A new perspective[J]. Neurocomputing, 2018,300: 70-79.

[3] 计智伟,胡珉,尹建新.特征选择算法综述[J].电子设计工程, 2011,19(9):6.

[4] 周金治,唐肖芳.基于相关系数分析的脑电信号特征选择[J]. 生物医学工程学杂志, 2015,32(4):5.

[5] PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005,27(8):1226-1238.

[6] WANG X, HU T, TANG L. A multiobjective evolutionary nonlinear ensemble learning with evolutionary feature selection for silicon prediction in blast furnace[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021,(99):1-14.

[7] WU M, SUN Y B, WEI Z H, et al. Automatic detection of epileptiform transients in EEG by a two-stage algorithm based on sparse representation[J]. Chinese Journal of Biomedical Engineering, 2009,60:101966.

[8] KE Xi, CHENG Cai. Feature selected based on PCA and optimized LMC[C]//2020 2nd International Conference on Computer Science Communication and Network Security (CSCNS2020)(2020 年第二届计算机科学, 通信和网络安全国际学术会议)论文集, 2020:1-6.

[9] HUANG WEI, YAN HONGMEI, LIU RAN, et al. F-score feature selection based Bayesian reconstruction of visual image from human brain activity[J]. Neurocomputing, 2018,316(17): 202-209.

[10] HYDE, CHARLES E. The Piotroski F-score: evidence from Australia[J]. Accounting and finance,2018,58(2):423-444.

[11] MIROWSKI P, MADHAVAN D, LECUN Y, et al. Classification of patterns of EEG synchronization for seizure prediction [J]. Clinical Neurophysiology, 2009,120(11):1927-1940.

[12] ISLAM K A, TCHESLAVSKI G V. Independent Component Analysis for EOG artifacts minimization of EEG signals using kurtosis as a threshold[C]// International Conference on Electrical Information & Communication Technology. IEEE, 2016.

[13] BO H. EEG analysis based on time domain properties[J]. Electroencephalography & Clinical Neurophysiology, 1970, 29(3):306-310.

[14] BOYLAN G B, RENNIE J M. Automated neonatal seizure detection[J]. Clinical Neurophysiology Official Journal of the International Federation of Clinical Neurophysiology, 2006, 117(7):1412-1413.

[15] GAO W W. Entropy measures for biological signal analyses[J]. Nonlinear dynamics, 2012, 68(3).

[16] MIRZAEI A, AYATOLLAHI A, GIFANI P, et al. Spectral Entropy for Epileptic Seizures Detection[C]// Second International Conference on Computational Intelligence. IEEE, 2010.

[17] KUMAR Y, DEWAL M L, ANAND R S. Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine[J]. Neurocomputing, 2014,133(8): 271-279.

[18] CHEN S, ZHANG X, CHEN L, et al. Automatic Diagnosis of Epileptic Seizure in Electroencephalography Signals Using Nonlinear Dynamics Features[J]. IEEE Access, 2019(99):1.

[19] ROY S, KIRAL-KORNEK I, HARRER S. Deep learning enabled automatic abnormal EEG identification[C]//2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018:2756-2759.

（下转第 73 页）