

随机森林算法在肌电的重要特征选择中的应用

张洪强¹ 刘光远¹ 赖祥伟²

(西南大学电子信息工程学院 重庆 400715)¹ (西南大学计算机与信息科学学院 重庆 400715)²

摘 要 在肌电信号的情感识别问题中,如何从高维特征中找出起关键作用的特征,一直是情感识别的难题。使用随机森林算法,并依照其对特征的评价准则,来计算肌电信号的 126 个初始特征在不同情感模式分类中的贡献度。依照每个特征的重要程度,优先组合贡献度大的特征并将其用于情感分类。实验数据验证了该方法的有效性。

关键词 情感识别,随机森林,肌电信号,特征选择

中图法分类号 TP391.41 **文献标识码** A

Application of Random Forest Algorithm in Important Feature Selection from EMG Signal

ZHANG Hong-qiang¹ LIU Guang-yuan¹ LAI Xiang-wei²

(School of Electronic and Information Engineering, Southwest University, Chongqing 400715, China)¹

(School of Computer and Information Science, Southwest University, Chongqing 400715, China)²

Abstract It is a hard problem that how to find the effective feature from high-dimensional feature in EMG signal emotion recognition. This paper used random forest algorithm to compute the contribution in different emotion recognition of the 126 EMG signal, depending on the feature evaluation criteria of random forest algorithm, and then preferentially made up the features that have more contribution to emotion recognition, and used then to emotion recognition. Experiments show it is reasonable.

Keywords Emotion recognition, Random forest, EMG signal, Feature selection

1 引言

随着信息技术的高速发展,为了满足人们对和谐人机环境的迫切需求而发展起来的情感计算,已经成为现在人工智能领域不可缺少的重要组成部分。美国麻省理工大学的 Picard 教授于 1995 年在他的技术报告中首次提出了情感计算的概念,随后在此基础上出版了第一部情感计算的专著《Affective Computing》,其中指出了情感计算是有关情感、感情产生以及影响情感方面的计算。让计算机设备也拥有情感,从生理信号中提取情感特征,分析各种生理信号和人类情感的关联,是国际上最近的研究热点。肌电信号(EMG)在所有的生理信号中是最重要的生理信号之一, Picard 等率先从 EMG 中提取了生理信号特征并进行情感识别^[1-6], 研究结果显示:提取的特征可以反映情感状态的不同,如何找到可靠的特征是问题的关键。在从 EMG 中提取特征集时,很难确保提取的特征能包涵足够的信息量以区分各种情感,为了提高情感分类能力,在提取特征集时一般都尽可能包涵所有信息,这就造成了过高的特征维数,给选择出真正起作用的特征增大了难度。考虑到上述问题,本文提出了使用随机森林算法,将每个特征对分类的贡献度量化,依照随机森林算法的评

价准则加以排序^[7,12],这样我们就能了解特征集中哪些是重要特征,这对于如何选择特征分类有很大帮助。

2 EMG 信号特征采集和特征提取

考虑到情感诱发,实验使用不同情感的电影片段,对在校且年龄在 19 到 25 岁间的大一健康新生进行试验,诱发情感,并使用美国 Biopac 公司提供的多导生理记录仪来采集被试的各种情感生理信号。最后采集到的数据是否有效需经过被试的投票结果来决定。对于 EMG 信号的特征提取^[8]:首先计算时域的常用统计量(均值、中值、标准差、最小值、最大值、最小值比率、最大值比率),再对 EMG 信号序列做一次差分和二次差分,然后计算以上的 7 个统计特征,从而得到 21 个时域统计特征。由于肌电信号的非平稳性,后面使用具有正交性、紧支性的 Daubechies5 小波做基函数,对降噪后的 EMG 信号做五层分解,然后对每一层的小波细节系数提取以上的 21 个统计特征,这样总共得到 126 个原始特征。为消除个体差异对情感分类的不利影响,对每个被试的每种情感特征取值都采用该被试在平静状态下的特征取值进行归一化^[9]。

该实验中总共有 254 个试验样本,针对每种不同的情感,

到稿日期:2012-04-11 返修日期:2012-07-27 本文受国家自然科学基金(60873143,61103132),教育部科学技术研究重大项目(311032),中央高校基本科研业务费专项资金(XDJK2009B008)资助。

张洪强(1987—),男,硕士生,主要研究方向为智能信息处理与情感计算;刘光远(1961—),男,博士,教授,主要研究方向为计算智能与情感计算;赖祥伟(1978—),男,博士,副教授,主要研究方向为软件测试、情感计算。

情感激发状态不同将导致不同情感的样本数量有所差异。

3 随机森林算法

随机森林算法^[6]是一种组合分类器算法,是由多个决策树 $\{h(x, \theta_i), i=1, 2, \dots, k\}$ 组成的分类器,其中 $\{\theta_i\}$ 是独立同分布的随机向量,由所有决策树综合决定输入向量 x 的最终类标签。每棵决策树的生长依赖于一个独立同分布的随机向量;整体的泛化误差则取决于森林中单棵决策树的分类能力和各分类树之间的相关程度,该算法主要包括两部分:决策树的生长过程和投票过程;随机森林采用 Bagging 方法,应用自助法(bootstrap)^[10]重采样技术生成多个决策树分类器。

3.1 决策树^[11]的生长步骤

1)从容量为 N 的原始训练样本数据中采取放回抽样的方法来随机抽取自助样本集,重复 k (树的数目 n_{tree} 是 k)次形成一个新的训练集 n ,以此生成一个分类器。

2)每个自助样本集生成单棵分类树,该自助样本集是单棵分类树的全部训练数据。设有 M 个输入特征,在树的每个节点处从 M 个特征中随机挑选 $mtry$ 个特征,按照节点不纯度最小的原则从 $mtry$ 个特征中选出一个特征进行分支生长,然后分别递归调用上述过程构造各个分支,直到这棵树能够准确地分类或所有属性都已使用过,在整个森林的生长过程 $mtry$ 将保持恒定。

3)分类树为了达到低偏差和高差异的要求而要充分生长,使每个节点的不纯度达到最小,不进行通常的剪枝操作。随机森林中的每一棵树为二叉树,其生成遵循自上而下的递归分裂原则,即从根节点开始依次对训练集进行划分;在二叉树中根节点包含全部训练数据,按照节点不纯度最小原则,分裂为左节点和右节点。按照同样的规则,节点继续分裂,依次进行直到满足分支停止规则分类树就停止生长。设 $I(n)$ 为节点 n 的不纯度,当节点 n 上的分类数据全部来自于同一类别时,此节点的不纯度 $I(n)=0$;如果分类数据服从均匀分布,则不纯度很大。节点不纯度的度量方法是 Gini 准则,设 $P(w_j)$ 是节点 n 上属于 w_j 类样本个数占训练样本总数的频率,则 Gini 不纯度表示为:

$$I(n) = \sum_{i \neq j} P(w_i) P(w_j) = 1 - \sum_j P^2(w_j) \quad (1)$$

3.2 随机森林投票过程

由生成的多棵分类树组成随机森林,用随机森林分类器对新的数据进行判别与分类,分类结果按树分类器的投票多少而定,也就是对测试样本 X 、预测类 C 、测试样本的最终类投票,投票过程:

$$C_p = \arg \max_c \left(\frac{1}{n_{tree}} \sum_{i=1}^{n_{tree}} I\left(\frac{n_{h_i, c}}{n_{h_i}}\right) \right)$$

式中, n_{tree} 是森林中决策树的数目; $I(*)$ 是示性函数; $n_{h_i, c}$ 是树 h_i 对预测类 C 的分类结果; n_{h_i} 是树 h_i 的叶子节点数。

经投票后,生成混淆表 M ,其是一个 $nc \times nc$ 表,表中的元素 $M(i, j) (i \neq j)$ 表示类型 i 被分类为类型 j 的次数,而仅当 $i = j$ 时, $M(i, j)$ 表示类型 i 被分类正确的个数。

随机森林分类正确率 correct 为:

$$\text{correct} = \frac{\sum_{i=1}^{n_c} M(i, i)}{\sum_{i, j=1}^{n_c} (i, j)} \quad (2)$$

式中, n_c 是类别的总个数

3.3 特征选择

1)依据 Gini 指数(式(1))的变化程度来反映每个特征的重要程度,其值越大说明特征越重要^[12]。

2)分类准确度降低的程度(式(2))来反映每个特征的重要程度,其值越大说明所选的特征的重要性越高^[7]。

本实验引进随机森林的 Gini 指数和分类误差准确度作为对一个特征贡献度的评价指标。

3.4 随机森林算法的流程

Step1 对于给定的一个生理信号训练样本集,通过 n 次随机的可重复采样,从数据 $(x_1, y_1) \dots (x_n, y_n)$ 出发构建一个步步为营的样本 $(\tilde{x}_1, \tilde{y}_1) \dots (\tilde{x}_n, \tilde{y}_n)$ 。

Step2 基于每一个步步为营样本,构建一棵决策树。

Step3 重复 Step1、Step2,可以得到多棵树。

Step4 让每一棵树都对输入的向量 x_i 进行投票。

Step5 计算所有的投票数,找出其中票数最高的一个就是向量 π_i 的分类标签。

Step6 与正确的分类标签不一样的比例,就是随机森林的错误分类率。

4 实验以及结果分析

依据实验仿真运行结果,当设定生长 1000 棵树,并且设定每个节点的分裂特征数为 11 时,实验效果较好。试验中设定生长 1000 棵树,并且每个节点的分裂特征数为 11,使用肌电信号作为研究对象,其中每类情感都提取了 126 个特征。对高兴与其他情感进行分类,运行随机森林算法得到的实验结果如图 1—图 5 所示。

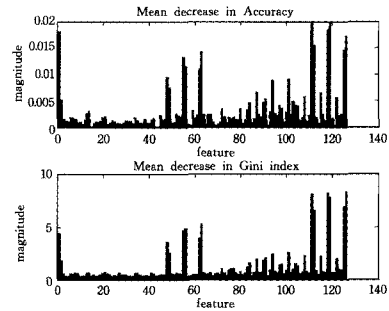


图 1 高兴与愤怒

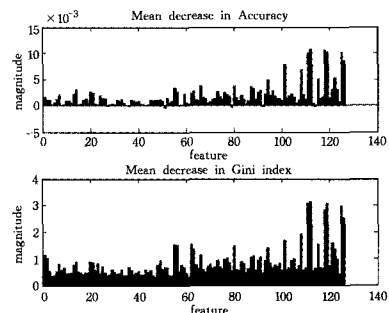


图 2 高兴与悲伤

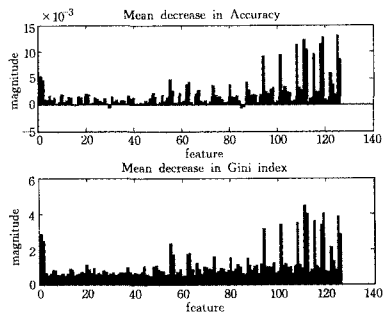


图3 高兴与惊奇

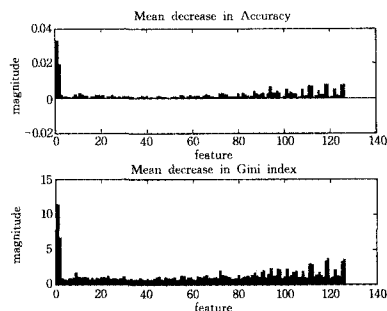


图4 高兴与恐惧

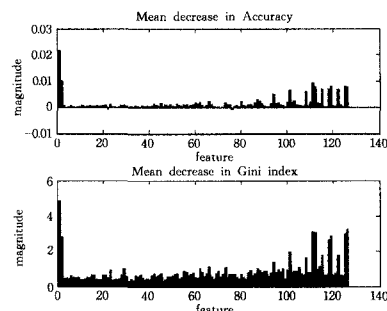


图5 高兴与厌恶

结果分析:

试验中我们使用肌电信号作为研究对象。在对高兴情感与愤怒情感进行二分类识别时,将 217 个高兴样本和 238 个愤怒样本随机打乱,抽取其中的三分之二作为训练集,将剩余的三分之一作为测试集运行随机森林算法得到运行结果。同理将高兴与悲伤(247 个样本)、高兴与惊奇(196 个样本)、高兴与恐惧(240 个样本)、高兴与厌恶(100 个样本)的样本分别组合运行随机森林算法得到结果。对高兴与其他 5 类情感进行二分类实验,得到 5 张结果图,图中横坐标表示特征编号,纵坐标表示 GINI 值和分类误差值,其中这两个值越大说明特征在分类中所起的作用越大,也就是贡献度越大。

由图 1 可以看到,在分类高兴情感与愤怒情感时,1、2、47、38、54、55、101、111、112、118、119、125、126 号特征的贡献度相对较大。

由图 2 可以看到,在分类高兴情感与悲伤情感时,115、118、119、125、126 号特征的贡献度相对较大。

由图 3 可以看到,在分类高兴情感与惊奇情感时,1、2、95、101、108、112、111、115、118、119、125、126 号特征的贡献度相对较大。

由图 4 可以看到,在分类高兴情感与恐惧情感时,1、2 号

特征的贡献度相对较大。

由图 5 可以看到,在分类高兴情感与厌恶情感时,1、2、111、112、118、119、125、126 号特征的贡献度相对较大。

分别用上述选择的重要特征再次使用随机森林算法训练后测试得到的识别率分别为:高兴对愤怒为 73.9%,高兴对悲伤为 71.3%,高兴对惊奇为 72.4%,高兴对恐惧为 77.4%,高兴对厌恶为 76%。由此可以看出肌电信号可以作为情感识别的生理信号,并且可以使用随机森林算法确定其主要特征,对于具有高维特征的肌电信号数据,可以采用随机森林算法对原始特征进行降维,并且能够取得不错的效果。

结束语 随机森林是一种多分类器算法,其对小样本的抗噪声性能不是很理想,这是由于人为采集的实验数据可能有一定的误差,在使用不同样本时会表现出一定的波动,后面我们可以试着从这方面加以改进。

参考文献

- [1] Picard R W, Healey J. Affective Wearable [C]// Proceedings of the First International Symposium on Wearable Computers, 1997. Cambridge, MA, Oct. 1997; 231-240
- [2] Picard R W, Vyzas E, Healey J. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(10): 1175-1191
- [3] Kim K H, Bang S W, Kim S R. Emotion Recognition System using Short-term Monitoring of Physiological Signals[J]. Medical Biology Engine Computer, 2004, 42: 419-427
- [4] Ekman P. An Argument for Basic Emotion [J]. Cognition and Emotion, 1992, 6: 169-200
- [5] Haag A, Goronzy S, Schaich P, et al. Emotion Recognition Using Bio-Sensors: First Step Towards an Automatic System [J]. Affective Dialogue Systems, 2004, 3068: 36-48
- [6] Picard R, Healey J. Digital Processing of Affective Signals [R]. M. I. T Media Laboratory Perceptual Computing Section Technical Report, No. 444. Massachusetts: MIT, 1998
- [7] 洪燕珠. 基于随机森林法的慢性疲劳证候要素特征症状的选择 [J]. 中医杂志, 2010, 51(1): 634-638
- [8] Cheng Ying. The Research of EMG Signal in Emotion Recognition Based on TS and SBS Algorithm [C]// 2010 the 3rd International Conference on Information Sciences and Interaction Sciences. 2010: 363-366
- [9] 温万惠. 情感生理反应样本库的建立与数据相关性分析 [J]. 中国科学 F 辑: 信息科学, 2010, 41(1): 77-89
- [10] Efron B, Tibshirani R J. Bootstrap measures for standar error confidence interval and other measure of statical accuracy [J]. Statistical Science, 1986, 1(1): 54-74
- [11] Breiman L. Randomforests [J]. Machinelearning, 2001, 45(1): 5-32
- [12] Nunes L O. Degradation type classifier for full band speech contaminated with echo, broadband noise, and reverberation [J]. IEEE transactions on audio, speech, and language processing, 2011, 19(8): 2516-2525