



论 文

面向临床心电图分析的深度学习算法

金林鹏^{①②}, 董军^{①*}^① 中国科学院苏州纳米技术与纳米仿生研究所, 苏州 215123^② 中国科学院大学, 北京 100049

* 通信作者. E-mail: jdong2010@sinano.ac.cn

收稿日期: 2014-03-19; 接受日期: 2014-08-11; 网络出版日期: 2015-01-21

摘要 作为模式识别研究的典型应用, 计算机辅助心电图分析在可穿戴终端及面向基层的“云”服务平台具有重要价值. 本文首先说明了面向临床的心电图分类模型的复杂性, 接着从非线性函数拟合能力角度分析了现有的特征提取算法和分类算法, 并采用深度学习构造心电图分类模型. 针对多导联心电图这种特殊的二维结构, 提出导联卷积神经网络, 并利用“平移起始点”和“加噪”增加训练样本数. 通过 15 万多条记录的测试数据, 取得了准确率为 83.66% 和 AUC 为 0.9086 的成绩. 最后我们还移植分类模型到移动终端设备, 其实时分析结果满足应用需求.

关键词 深度学习 心电图 (ECG) 非线性拟合 卷积神经网络 数据分布与分类

1 引言

心电图 (electrocardiogram, ECG) 是记录人体心脏电活动的可视时间序列, 它已经在临床上广泛用于心脏相关疾病检查, 并且形成了比较完善的 ECG 判断标准. 由于 ECG 往往会包含各种干扰信号, 比如呼吸运动引起的频率低于 0.5 Hz 的低频波, 工频导致的 50, 60 Hz 高频波, 身体肌肉活动产生的约 33 Hz 肌电杂波等等, 临床医生首先要根据常识忽略干扰厉害的波形, 然后判断 ECG.

计算机辅助 ECG 分析方法^[1]在数十年间得到了长足发展, 现有心电设备已经广泛采用了其中的一些内容, 包括 ECG 主波 (R 波) 位置提取、各波起止位置确定等. 如果 ECG 的各个波形特征能够准确提取, 那么按照疾病分类规则就可做出有效判断. 然而, 即使是目前研究最成熟的 R 波提取方法^[2], 其性能在标准数据库 MIT-BIH^[3]上测试很好, 而在实际临床数据集上则下降颇多; 至于 P 波、T 波等提取算法在 MIT-BIH 上的准确率就低, 更不用说临床应用了.

为此, 各种数理特征如小波特征^[4]、高阶统计量^[5]、功率谱特征^[6]、Lyapunov 系数^[7]、Hermit 系数^[8]、Shannon 熵^[9]、Hermite 多项式系数^[10]、线性预测误差特征^[11]等一一被提出, 它们配合时域特征、波形特征, 以及主成分分析 (principal component analysis, PCA)、线性判别分析 (linear discriminant analysis, LDA)、独立成分分析 (independent components analysis, ICA) 等统计分析方法, 被共同用于 ECG 分类识别中. 不可否认这些方法在给定数据集上的有效性, 但是从已发表的文献上看, 不少方法只是在标准数据库或其一个子集上得出的结论 (即使是整个 MIT-BIH 数据集, 其实也只有 47 个病人), 算法的泛化能力很难得到保障. 还有一个很显著的心搏分类结论就是, 病人内

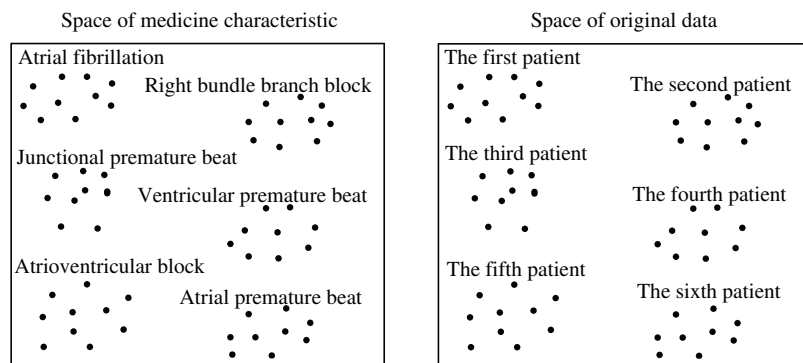


图 1 不同数据空间中的 ECG 分布情况

Figure 1 Distribution of ECG in different data spaces.

(训练集和测试集均包含相同个体的 ECG 样本) 的正异常识别率要远高于病人间的 (训练集和测试集的 ECG 样本所属个体没有交集), 而病人间的某种疾病识别率要介于病人内和病人间的正异常识别率 [12].

2 ECG 数据分布研究

我们结合医生诊断规则和数理统计方法, 对 ECG 潜在的数据分布进行分析.

为了模拟医生诊断时思维过程, 我们先从某种疾病分类入手. 医生在区分正异常时, 其实也是不断地与记忆中疾病模板进行匹配, 如果都不是, 那就判为正常. 对于某种特定疾病, 医学专家有一套诊断标准 (疑难杂症除外), 如 ECG “右束支阻滞” 的诊断规则为:

- (1) P 波出现在每个 QRS 波前;
- (2) PR 间期: 0.12~0.20 s;
- (3) QRS 波: >0.12 s;
- (4) V1 或 V2 导联呈 rSR', RSR' 和 Rsr';
- (5) V5, V6 和 I 导联呈粗钝的 S 波.

假设 x 是任意一个人的 ECG 数据, 经过特征提取函数得到各种医学特征 $f_m(x)$, 即诊断规则中的各项指标; 很显然, 同一种疾病的 ECG 数据点在医学特征空间是聚在一起的, 即满足模式识别中的聚类假设. 图 1 左图粗略地给出了医学特征空间中 ECG 数据分布情况, 黑点为异常 ECG, 空白区域即为正常 ECG. 当然, 这里只是示意各类 ECG 数据的相对分布位置, 同时, 个别数据点还会交错在一起 (对应疑难杂症).

由此可知, 在医学特征空间中不需要很复杂的分类器就能满足要求 (支持向量机 SVM 就能很好地处理这种分类任务), 前提是 $f_m(x)$ 不但能处理各种噪声干扰, 还必须能应付不同人之间由于生理差异导致的数据多样性. 显然, $f_m(x)$ 是一个高度复杂的非线性函数. 成功设计满足临床应用要求 (病人间正异常诊断、特定疾病诊断) 的 ECG 自动分析算法的关键是医学特征提取函数 $f_m(x)$ 的构造.

下面再分析一下病人内基于心搏正异常分类的 ECG 数据分布情况. 为了从 MIT-BIH 或其他数据库上获取 ECG 心搏样本, 大部分文献的做法是先用主波提取算法找到 R 波位置, 然后以 R 为中心向前向后取点组成一个心搏样本. 接着以随机方式、奇偶数方式等划分数据集; 或取一条记录的前

几分钟心搏为训练集/校验集, 剩余部分为测试集. 然而, 不论采用何种方式, 一个不变的原则就是训练集和测试集都包含同一个人的心搏样本, 只不过是时刻不同而已.

ECG 本身就可以做身份识别, Wubbeler 等^[13]的工作表明, 以 R 波峰值点为中心分别在其前后提取 100 ms 长度的波形数据作为特征, 然后计算各个体之间特征的欧氏距离, 并结合阈值法, 就可以实现很高效的身份识别算法 (以 PTB 数据库^[3]中 74 个个体 ECG 数据为实验样本, 识别正确率最高达到 99%). 据此可知, 同一个人的 ECG 原始数据倾向于聚在一起. 图 1 右部分粗略地给出了原始数据空间中 ECG 数据分布情况, 同一个人的 ECG 数据聚在一起, 而不同人的数据分布在不同的区域.

对照图 1 左右两部分我们会发现两者惊人的相似, 唯一的区别就是, 一个是在医学特征空间, 而另一个是在原始数据空间.

不过很巧合的是“病人内 ECG 分类”可以绕过对 $f_m(x)$ 的构造, 在原始数据空间中 ECG 心搏就具有很好的可分性^[13], 接下来的任务就是在保持数据分布不变的情况下, 尽量减少噪声和样本维数 (因为输入维数越少, 又具有足够的可分性, 我们的分类器就能做得很好, 在第 3 部分会详细说明这点), 于是各种数理特征如小波系数^[4]、识别率高的医学特征如 RR 间期、流形算法、PCA/LDA/ICA 等方法各显神通, 在 MIT-BIH 或其他数据库上都取得了很高的识别率, 个别文献在给定的数据集上达到 100% 也就可以理解了, 但这么好的结果也就仅限于病人内 ECG 分类 (由图 1 右部分可知, 病人间 ECG 数据分布复杂到正常样本不聚在一块, 异常样本也不聚在一块).

因此病人内高效的识别算法到了临床实际应用, 分类性能会明显下降. 至于病人间的疾病分类, 在医学特征空间只需要考虑一个类, 而正异常分类要考虑多个类, 并且还有许多处于交错区域的疑难杂症, 某种程度上病人间的疾病分类要比正异常分类要容易.

3 模式分类与函数拟合

给定数据集 $X = \{x \in R^d\}$, $Y = \{y \in R\}$, 模式分类就是要找一个函数 f , 使得不同类别数据点的函数值 $f(x)$ 落入不同的区间. 当然, 不同的学习模型会有不同的表述, 但本质上均是拟合函数 f 的过程. 下面从非线性函数拟合角度来分析常见的模式分类算法.

假设要研究的类别为 2 个, 记为 w_1 和 w_2 , 各类的先验概率为 $p(w_i)$, 类条件密度为 $p(x|w_i)$, 则最小错误率贝叶斯 (Bayes) 模型的决策如下: 如果 $p(x|w_1)p(w_1) > p(x|w_2)p(w_2)$, 则 $x \in w_1$; 反之 $x \in w_2$, 即贝叶斯模型的决策函数 $g(x)$ 为

$$g(x) = p(x|w_1)p(w_1) - p(x|w_2)p(w_2).$$

若 $g(x) \in [0, +\infty)$, 则 $x \in w_1$, 否则 $x \in w_2$. 理论基础很完备, 但在复杂的实际应用中, 知道类条件密度 $p(x|w_i)$ 同知道函数 $g(x)$ 一样困难.

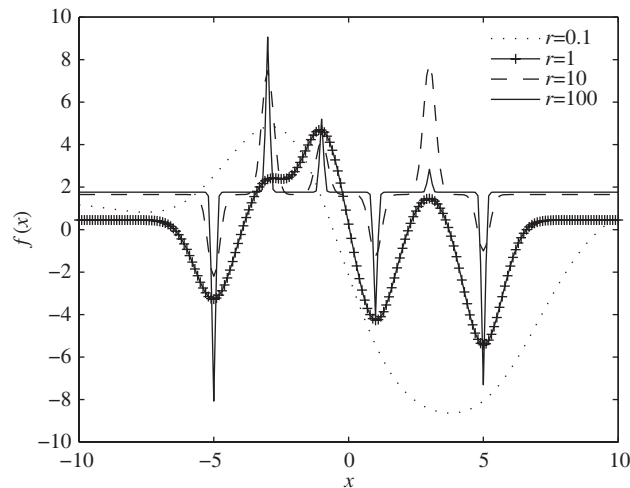
高斯混合模型 (Gaussian mixture model, GMM) 是一个典型的贝叶斯决策模型, 它假定类条件密度服从正态分布, 基于最小错误率的决策函数 $g(x)$ 如下:

$$g(x) = \operatorname{argmax}_i \{w_i N_d(\mu_i, \sigma_i)\},$$

其中 c 为类别数, w_i 是每类的先验概率. 显然, 决策函数 $g(x)$ 是由 c 个正态分布以组合方式进行拟合.

支持向量机 (support vector machine, SVM) 的决策函数 $g(x)$ 如下:

$$g(x) = \left(\sum_{i=1}^n a_i^* y_i K(x_i, x) \right) + b^*, \quad (1)$$

图 2 不同 r 拟合出不同函数 $f(x)$ Figure 2 Different r corresponding to different function $f(x)$

其中 $K(x_i, x)$ 为核函数, n 是支持向量个数 (小于等于训练样本数). 显然 $g(x)$ 由 n 个核函数通过线性组合的方式拟合而成, 以高斯核函数为例, 通过调整 r 和松弛因子 C (其中 r 起的作用更大), 式 (1) 中 a_i^* , b^* 和 n 随之改变, 从而拟合出不同的 $g(x)$, 至于是否与真实决策函数相符, 只有通过测试校验集来确认; 但不论怎么调整, 函数拟合受限于训练样本数.

图 2 给出了基于高斯核函数的 SVM 拟合一维函数的几种情况, 需要说明的是, 一旦给定支持向量 (或训练样本) 的个数, 不论怎么调整系数, SVM 所能拟合的函数族就完全确定了, 对于更复杂的非线性函数, 当前 SVM 就无能为力了. 从这个角度来说, 有差异的训练样本数越多, SVM 的非线性拟合能力就越强. 不过从 SVM 训练角度来说, 我们希望训练样本足够并且尽可能少, 虽然有大规模数据训练算法, 但若训练规模达到百万级别, 实际计算求解还是会有很大问题. 另外, 拟合高维非线性函数比低维非线性函数显然要复杂、困难很多. 因此, 在给定的训练集下, 我们总是先通过特征提取函数 $f(x)$ 从原始数据中提取特征 (其实就是降维过程), 然后再交由 SVM 进行函数 $g(x)$ 拟合, 这样做的另一个好处就是有效降低了核函数的计算复杂度 (实际上, 这里用复合函数 $g(f(x))$ 拟合最终的决策函数, 弥补 SVM 拟合高维非线性函数能力的不足).

高斯过程 (Gaussian process, GP) 的决策函数 $g(x)$ 如下:

$$\begin{cases} g(x) = \sum_{i=1}^n C(x, x_i) A_i, \\ A = K^{-1} Y = [A_1, A_2, \dots, A_n]^T, \end{cases}$$

其中 $C(x, x_i)$ 为协方差函数 (比如平方指数协方差函数), A 是通过训练所确定的一个常数矩阵, 决策函数 $g(x)$ 由 n 个协方差函数通过线性组合的方式拟合合成, 函数拟合受限于训练样本数 n .

神经网络与其说是“黑匣子”, 还不如说是通用的非线性函数逼近器. GMM 可以看作是 2 层的径向基网络, SVM 可看作是网络结构和权值自适应调整的 3 层径向基网络, GP 可看作类似于 3 层径向基网络 (激励函数为协方差函数, 隐层单元数为训练样本数). 这些模型都是不超过 3 层的浅层神经网络.

虽然 3 层神经网络理论上可以逼近任意函数, 但其代价是隐层单元数 (比如计算量和参数个数)

可能变得非常大, 由此会带来一系列计算求解问题. 实际上, SVM 将浅层神经网络做到了极致, 通过核方法, SVM 同时解决了网络结构调整和参数优化问题, 缺点是非线性函数拟合能力受限和超大规模训练集计算求解问题. 当然, 在不是很复杂的模式分类问题如病人内 ECG 分类问题上, SVM 就可以做得很好.

神经网络除了通过“纵向增加隐层神经元的方式”(浅层架构) 来提高非线性函数逼近能力外, 还可以通过“横向增加层及对应神经元的方式”(深层架构)^[14] 来达到目的. 比起浅层架构, 深层架构只需要少得多的神经元就可拟合同程度的非线性函数, 但参数优化问题制约着其实际应用. 不过, 2006 年 Hinton 等^[15] 在深度信念网 (deep belief networks, DBNs) 上做的工作改变了这种状况. 一方面, 通过人为增加层及对应神经元, 神经网络的非线性函数逼近能力不断得到提高; 另一方面, 通过无监督和有监督相结合的方式训练, 每层的参数得到有效调整. 因此, 在复杂的应用领域如声音、图像, 深层架构 (深度学习) 获得了巨大的成功. 当然, 对于一些简单的应用领域 (决策函数为线性函数或简单的非线性函数), 非要用深度学习去做 (即用复杂非线性函数去拟合决策函数), 效果反而不好.

至于近邻法、决策树、隐马尔科夫 (Markov) 模型 (主要通过状态转移矩阵和混淆概率矩阵进行预测分类) 等算法的非线性函数逼近能力不是很强, 这里就不具体展开说明了.

值得要说明的是, 对于复杂的模式识别领域, 通过组合多个分类器使得最终拟合的决策函数足够复杂也是很有用的方法, 其原理同 SVM 组合多个核函数类似. 文献 [16] 用卷积神经网络 (convolutional neural network, CNN) 提取特征, 之后交由 SVM 进行分类, 获得了比两者都要好的结果; 从非线性函数逼近角度就不难理解其本质原因了.

4 特征提取与函数拟合

设计面向临床应用的 ECG 分类算法实际上就是要拟合一个非线性决策函数 $F(x) = g(f(x))$, 其中 $f(x)$ 为特征提取算法所拟合的函数 (可以是医学或数理特征提取函数, 还可以是混合特征提取函数), $g(x)$ 为分类器所拟合的函数. 由第 2 部分分析可知, 病人内 ECG 分类算法最终拟合的决策函数 $F(x)$ 是高度复杂的非线性函数.

我们首先分析一下医学特征提取函数 $f_m(x)$. 假设每种特征提取算法准确性为 98%(目前这样的水平无法达到), 如果一种算法使用 20 种特征, 那么这 20 种特征全部都准确的概率只有 66.76%. 在这么低准确性的特征数据上, 提高算法最后的分类准确性, 是很困难的. 下面从非线性函数拟合角度来分析目前常用的特征提取方法.

PCA 是以方差大目标的线性降维算法, LDA 是考虑了类别信息的线性降维算法, 而 ICA 则是实现盲源分离, 三者均是线性模型. 因此, 在复杂的应用领域, 核函数方法被引入到这些算法中, 发展了核 PCA、核 LDA 和核 ICA. 不过, 所有基于核函数的算法在非线性函数拟合方面都存在着和 SVM 一样的缺陷.

小波变换在很多领域都取得了很不错的效果, 这里关注的是其在特征提取方面的表现, 其表达式为

$$WT_f(a, \tau) = \frac{1}{\sqrt{a}} \int_0^R f(t) \psi \left(\frac{t - \tau}{a} \right). \quad (2)$$

其中 $\psi_{a,\tau}(t)$ 为小波基函数, 具有尺度 a 和平移 τ 两个参数. 利用基函数的多尺度特性将信号在不同尺度下展开并提取有用信息; 在此基础上还可利用 Lyapunov 系数来描述信号的局部奇异性. 由式

(2) 可知, 小波变换只是简单的积分变换 (提取主要信息而丢弃细节信息), 其非线性函数拟合能力非常有限.

Shannon 熵、Hermite 系数、功率谱等其他数理特征也都是通过简单的数学公式计算所得 (这里就不展开说明了, 可参看相关文献), 其非线性函数拟合能力同小波变换一样, 都是非常有限的.

IsoMap (isometric feature mapping)、LLE (locally linear embedding) 等流形算法的非线性函数拟合能力并不是很强, 因为这些算法的本质是将数据从高维降到低维时, 尽量保持原近邻数据的某种关系. 降维过程必然伴随着部分数据信息的丢失, 噪声的存在还可能会严重扰乱数据本来结构中的邻域关系. 此外, 相关实验表明, 当训练样本规模达到百万级别时, 确定数据点领域关系的计算量是普通计算机不可接受的.

不过, 像核方法和流形算法所遇到的大规模样本计算问题, 通过样本空间划分是可以解决的, 但这其实可归为多分类器组合领域 (不同区域对应一个子函数, 整个样本空间所拟合的函数就是组合所有子函数). 适当的多分类器组合总是很有效的方法, 这在第 3 部分已有说明.

由以上分析可知, 现有的各种数理特征提取方法的非线性函数拟合能力都是非常有限的. 此外, 任何特征提取方法在将原始的高维数据降为低维数据时, 必然伴随着部分信息的丢失. 我们希望丢失冗余信息 (比如噪声), 但实际情况是无法确认到底多少维才能满足要求; 在某个数据集上对最终分类没有贡献的特征, 在另一个数据集上可能就有贡献. 另外, 将特征提取过程和分类过程孤立开来的缺陷是, 一旦提取的特征不能完全反应样本内在属性, 后续分类器再怎么设计, 也无法拟合出真实的决策函数. 再看深度学习^[15], 无监督学习就是特征提取过程, 而有监督学习就是要拟合最终的决策函数. 显然, 深度学习将特征提取过程和分类过程融合在一起, 避免了特征提取过程出现的部分偏差, 而这也是深度学习成功的原因之一.

5 面向临床应用的 ECG 分类方案研究

显然, 可临床应用的 ECG 分类算法应该是面向病人间的, 部分导联脱落、部分 QRS 波群不明显、噪声干扰都是普遍存在的, 目标测试集是开放的, 也即意味着最终要拟合的决策函数是一个高度复杂的非线性函数. Martinez 等^[17]首次将深度学习应用到生理信号的分类识别当中, 取得了 70%~75% 准确率. 鉴于理论分析和已有工作, 我们采用深层神经网络, 并且不做任何特征提取过程, 直接用原始数据拟合最终的 ECG 决策函数.

深层神经网络随着层数和神经元数增加, 其非线性函数拟合能力就会增强, 但待调整参数也会成倍增加, 也即意味着网络模型所能拟合的函数个数增加 (一组参数对应一个函数). 参数调整过程实质上是在函数空间中找一个最优的函数, 这比传统最优化问题 (在搜索空间找一个最优的点) 要困难的多. 当训练样本不多时, 满足条件的函数太多, 而 (back propagation, BP) 算法只可能找到有限的几个, 因此没有特别的正则化方法 (经典表述是 “引入一些先验知识来限制学习过程”); 就优化而言, 就是限制待调整参数的取值范围, 减少搜索空间, BP 训练必然失败.

DBNs 的无监督预训练可以看作是正则化方法, 通过预训练将参数确定在某个最优值的附近, 以限制其任意取值; 稀疏编码^[18]则是典型的参数正则化方法, 目的是迫使大部分参数为零; CNN 其实也是正则化方法, 卷积层可看作是只有局部连接权值而其他连接权值均为 0 的全连接层, 权值共享进一步限制参数的取值 (我们称为 “卷积正则化”). 深层神经网络要想用 BP 算法成功训练, 正则化方法是必须的, 而且越多越好. 当然, 如果优化算法足够优秀, 那就另当别论了.

常规 ECG 共有 12 个导联的数据, 其中 8 个是正交的, 其余 4 个可从已有的 8 个导出. 每个导联

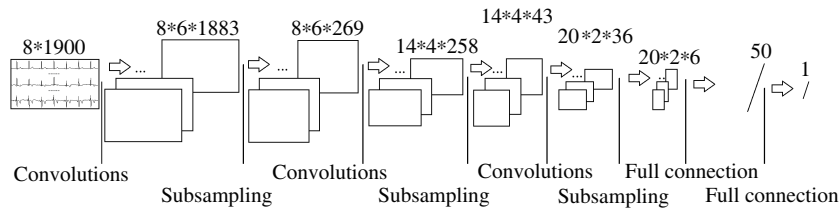


图 3 CNN 结构示意图

Figure 3 Structural representation of CNN

通常采集 10 s 的 ECG 记录数据, 大概包含 12~18 次搏动, 经 500 Hz 采样后, 每个导联的心电信号对应一个维数为 5000 的向量, 8 个导联的 ECG 记录数据共为 40000 维. 我们将“卷积正则化”加入到深层神经网络中, 即可解决普通计算机由于权值过多导致内存溢出问题, 又利用了 BP 训练. 另外, 现有的 CNN 结构^[19] 主要面向二维图像, 虽然多导联 ECG 与二维图像相似, 但导联间数据相关性与导联内数据相关性程度不一样. 为此, 我们基于多导联 ECG 这种特殊的二维结构, 加入新的正则化方法(我们称之为“导联卷积正则化”), 进一步限制参数的取值. 图 3 是一种典型的卷积神经网络.

5.1 卷积神经网络

CNN 是模拟视觉识别图像的过程而创造的, 其网络结构专门针对二维图像设计(典型的 2 维卷积操作), 并首先成功用于手写体识别.

在图 3 中, 我们把 8 导联心电图看作是二维图像, 输入层维数是 8×1900 ; 卷积层 A 共有 8 个特征面, 每个特征面使用一个 3×18 卷积核(记为 $[3, 18]$, 表示一个输入元素个数固定, 与时间和初值等都无关的单值函数), 该层输出 8 个 6×1883 (其中 $6 = 8 - 3 + 1$, $1883 = 1900 - 18 + 1$) 大小的特征面; 取样层 A 采取 1×7 取样核(记为 $[1, 7]$), 产生 8 个 6×269 (其中 $6 = 6 / 1$, $269 = 1883 / 7$) 的特征面; 卷积层 B 采用 3×12 的核, 产生 14 个 4×258 的特征面; 取样层 B 采用 1×6 的核, 产生 14 个 4×43 的特征面; 卷积层 C 采用 3×8 的卷积核, 产生 20 个 2×36 的特征面; 取样层 C 同样采用 1×6 的核, 产生 20 个 2×6 的特征面(实际上是 240 个点); 最后两层是使用多层感知器进行分类输出.

每个输入特征面(输入层相当于 1 个特征面)通过各自不同的卷积核映射成多个输出特征面, 而组成同一个输出特征面的每个神经元与每个输入特征面的相同局部区域相连, 且权值共享, 但不同输出特征面的神经元权值不共享; 然后每个输出特征面通过缩放操作(间隔取最大值或间隔取平均值)减少尺寸大小. 显然, 若输出特征面过少, 则不利于神经网络学习, 因为其他有利于学习的特征可能被忽略了; 但输出特征面过多, 待调整参数以及网络前向计算时间均成倍增加, 不利用 BP 训练. 因此, 选择合适的特征面数就显得极其重要. 在实际应用中, 一般可先尝试较少个数的特征面, 然后逐步增加个数, 并观察所得训练模型的性能, 直到找到一个前向计算时间和分类性能都不错模型.

作为深层神经网络的一种, CNN 可采用经典的训练方法: 从低层到高层, 每一层的卷积核(权值)先通过受限玻尔兹曼(Boltzmann)机^[15]或降噪编码机^[20]进行无监督学习以得到比较好的初值, 然后再交由监督学习继续调整, 以获得更优值. 不过当标注样本过多时, 无监督学习所起作用会弱化掉^[20], 此时直接采用监督学习, 每一层的卷积核(权值)也能得到非常有效的学习.

在监督学习过程中, 每隔一定周期, 就用当前训练模型去测试校验集, 将测试结果最好的模型保存下来, 而不是等训练结束后再进行保存. 这是因为使训练集误差小的决策函数不一定是真实所需判别函数(经典的“过拟合”问题), 而只有准确预测未见样本的决策函数才是需要的.

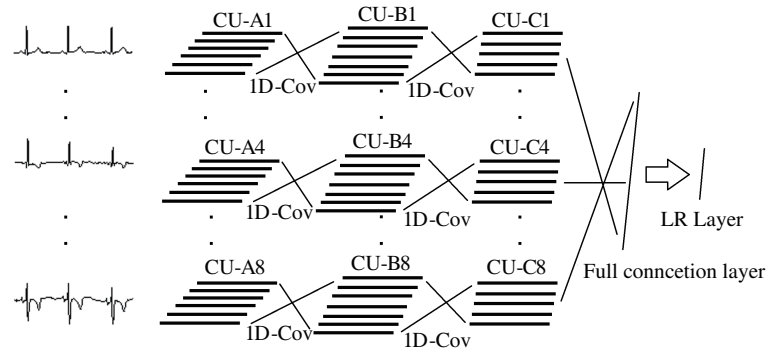


图 4 LCNN 结构示意图

Figure 4 Structural representation of LCNN

5.2 基于“导联卷积正则化”的深层神经网络

“卷积正则化”一般对行方向和列方向数据均进行卷积融合计算, 对图像而言是非常合情合理的; 但若原封不动的应用到多导联心电图数据, 那么不同导联 (列方向) 同一时间段内 (行方向) 的心电图数据就会有卷积计算. 由于导联内数据是时间相关的, 采用卷积计算是合情合理的; 但导联间数据是独立的, 若要进行卷积计算, 最好的做法是不同导联所有不同组合情况 (即两导联间的所有组合情况进行卷积计算, 三导联间的所有组合...) 都这么做. 然而就现有的“卷积正则化”而言, 不论如何设计卷积核, 都只能实现部分导联的卷积融合计算.

相同导联的 ECG 数据具有相似性属性, 不同导联的 ECG 数据则不具有该属性. 既然卷积计算有缺陷, 那么我们干脆将不同导联间的卷积计算去掉. 卷积核采用 $1 \times X$ 就可实现这个目的, 但在同一个特征面里, 不同导联数据共享同一个卷积核, 从数字信号处理角度来看, 不同导联数据就被强制使用同一个滤波器 (卷积计算可看作是滤波过程). 然而, 不同导联都具有各自内在的属性, 理应都有各自最佳的滤波器, “卷积正则化”的强制共享显然是一个不合理的设定. 我们希望通过样本训练, 每个导联都能找到最适合自己的滤波器. 据此, 我们设计了基于“导联卷积正则化”的深层神经网络 (称之为导联卷积神经网络 LCNN), 其结构如图 4 所示. 卷积单元 CU 包括多个卷积/抽样特征面, 1D-Cov 表示 1 维卷积计算, 每个导联均有 3 个卷积单元, 且不同导联间的卷积单元互不相干. 若有 8 个导联, 则一共有 24 个卷积单元, 而传统的 CNN 就只有 3 个卷积单元. 每个导联的心电图数据依次通过最适合自己的 3 个卷积单元, 之后汇总所有导联的信息做最后的分类, 这是合情合理的设定.

图 4 中 LR 层就是传统 Logistic Regression 层, 仅限于二分类问题; 相比 Softmax Regression (SR) 层, 其优势是保证神经网络误差函数值不变的情况下, 有效减少待调整参数的个数; 但在某些情况下可能会运算溢出 (出现 NaN), 导致无法进一步 BP 训练, 此时可考虑采用 SR 层.

为了更清晰地描述 LCNN 前向计算过程, 我们给出其计算公式. 设输入样本为 $x = [x_1, x_2, \dots, x_8]$, 其中 $x_i (1 \leq i \leq 8)$ 为第 i 通道的数据, f_{cov} 为卷积函数, f_{sub} 为抽样函数, 则有

$$\begin{cases} f(x) = g_E \left(g_D \left(\bigcup_{i=1}^8 g_{C_i} (g_{B_i} (g_{A_i} (x_i))) \right) \right), \\ g_D(x) = \phi(Wx + b), \\ g_E(x) = \frac{1}{1 + e^{-(Wx+b)}}, \end{cases}$$

其中 g_D 是全连接层计算函数, g_E 是 LR 层计算函数, $\phi(\cdot)$ 是激励函数. $g_{A_i}, g_{B_i}, g_{C_i}$ 为卷积单元计算函数, 表达式均为 $f_{\text{sub}}(f_{\text{cov}}(x))$, 唯一区别是计算权值不同, 两函数的具体计算公式如下:

$$\begin{cases} f_{\text{cov}}(v) = \bigcup_{j,x} v_{ij}^x = \bigcup_{j,x} \left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} w_{ij,(i-1)m}^p v_{(i-1)m}^{(x+p)} \right), \\ f_{\text{sub}}(v) = \bigcup_{x=1, Q_i+1, \dots}^m v_{im}^x = \bigcup_{x=1, Q_i+1, \dots}^m \phi \left(\max \left(\bigcup_{q=0, 1, \dots, Q_i-1} v_{(i-1)m}^{(x+q)} \right) \right). \end{cases}$$

这里设第 i 层的卷积核和取样核分别为 $1 \times P_i$ 和 $1 \times Q_i$, v_{ij}^x 表示“第 i 层第 j 个特征面”位置是 x 的神经元输出值, $w_{ij,(i-1)m}^p$ 表示“第 i 层第 j 个特征面”指向“第 $i-1$ 层第 m 个特征面”的核权值, b_{ij} 表示第 i 层第 j 个特征面的偏置.

LCNN 可采用和 CNN 相同的训练算法, 不过 LCNN 的无监督学习更高效. 当我们采用降噪编码机等进行无监督学习, 每个导联可单独训练, 这样就使降噪编码机的输入维数大大减少, 同时还具有天然的并行计算特性, 非常有利于缩短训练时间.

5.3 增加样本和不平衡数据集处理

BP 算法是基于梯度下降法而设计的神经网络优化算法, 如果函数空间中只有一个最优函数, 那么 BP 算法就能做得很出色. 除了正则化方法以外, 增加训练样本也可使满足条件的函数 (由神经网络拟合出来) 个数减少, 从而有利于拟合出真实的决策函数. 实际上, 训练样本数的增加还能弱化“正则化方法”带来的作用, 比如在用 CAE 训练 MNIST 手写数字数据库时, 若通过扭曲变换方式等来增加实际训练样本数, 可弱化无监督预训练所带来的作用 [20].

然而, 无限制的增加训练样本是不现实的; 同时, 对于心电图数据的收集, 正常样本总是大量存在, 而某种疾病的样本数就会相对少、互相也不一样. 也就是说, 不同类别心电图数据的不平衡性是普遍存在的.

为此, 我们充分利用 ECG 记录数据的周期特性, 通过平移起始点, 将一条 ECG 记录所有可能的情况都包含进去. 这样做不仅增加了训练样本数, 还增强了分类算法的鲁棒性, 因为给定一条好多个心搏的 ECG 记录, 其起始点可以在一个心搏内的任何位置, 而不同起始点的 ECG 记录在数据空间却是完全不同的点. 似乎将所有 ECG 记录的起始点都选在 R 波位置是一种选择, 可实际情况是临床应用的 ECG 数据复杂多样性, QRS 波群不明显、部分导联脱落但医生还能给出诊断结论的数据等比比皆是, 而目前成熟的 R 波提取算法都只适用于相对规整的 ECG 数据, 并且这样做还会导致训练样本相对匮乏.

由于不同起始点的心电图数据对应到数据空间中的不同点, 这样就避免了传统重采用技术“简单复制”或“启发式构造”的一些缺陷.

加噪处理或对原始数据做某种扭曲变换, 是神经网络训练中为增加样本而普遍采用的技术. 实际上, 只要对类别的最终判断无影响, 不论怎么设计变换都可以. 然而, 对于给定的数据集, 若某种变换方法迎合了测试集中的内容, 那么它将会比其他方法有效, 而当测试集无穷大时, 也就不会有什么优势了. 针对心电图数据, 我们叠加正常情况下不影响医生做出判断的噪声, 包括白噪声、低频基线漂移噪声和高频 (工频) 噪声, 具体过程如下:

- ① 每个导联分别叠加幅度在 $[0, 0.1]$ 随机变化的白噪声;
- ② 每个导联分别叠加幅度在 $[0, 0.1]$ 随机变化, 相位随机, 频率在 $[0, 0.2]$ 随机变化的低频噪声;

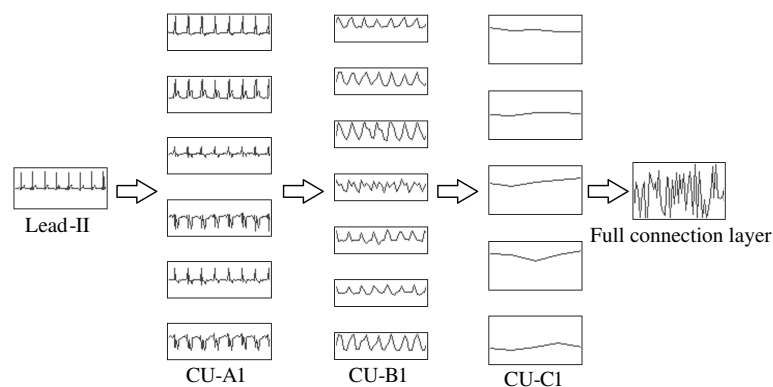


图 5 基于深度神经网络的 ECG 特征
Figure 5 ECG features based on deep neural network

- ③ 每个导联分别叠加幅度在 $[0, 0.15]$ 随机变化, 相位随机, 频率在 $[45, 90]$ 随机变化的高频噪声;
- ④ 每个导联分别减去叠加噪声后的该导联平均值.

经过以上处理, 原始 ECG 在训练过程中, 每次训练叠加的噪声将不相同, 从而增加了训练样本数目.

不论通过“平移起始点”, 还是通过“加噪”来增加训练样本, 都可看作是正则化方法, 进一步限制参数的取值.

5.4 对深度神经网络自动提取特征的探讨

相比传统的模式分类算法, 深度神经网络实际上包含了自动特征提取过程, 每一层对应一组特征输出, 最后一层给出分类结果.

为了直观地显示深度神经网络所提取的特征, 以第 6 部分最终训练所得的 LCNN 作为特征提取模型, 用编号为 4176 记录为例, 画出各层所提取的特征 (以时间序列方式对待). 图 5 仅以 II 导联为例, 给出其经过卷积单元 A1, B1, C1 所提取的特征, 最后一个则是汇总所有导联信息, 经过全连接层所提取的特征.

由图中可以看出, 各层所提取的特征以增强的方式从不同角度表现原始信号, 并且随着层数的增加, 其表现形式越来越抽象. 很显然, 这些特征并没有临床诊断的物理意义, 仅仅是数理值.

从模式分类角度来说, 满足 Fisher 判别准则的特征最有利于分类. 由图 1 可知, 医学特征很好地满足了判别准则, 但问题是现有人工特征提取方法都不太准确, 每引入一个特征, 便会累积误差, 从而无法准确拟合决策函数.

再看 LCNN, 最后层 (LR 层) 其实是一个线性分类器, 要提升分类性能, 则其输入 (即全连接层所提取的特征) 就要尽量满足 Fisher 判别准则. 监督学习是以提高分类性能为目标, 为此, 在学习过程中, 全连接层所提取的特征就会尽量满足判别准则, 而前几层就会配合全连接层对自身参数做调整. 整个过程是自动完成的, 避免了中间特征的引入导致误差累积, 只要网络参数得到有效调整, 就能够准确拟合决策函数.

虽然不同特征提取方法的具体过程不同, 但殊途同归, 只要提取出的特征有利于分类即可. 一方面, LCNN 通过输入原始数据而避免了中间误差的引入; 另一方面, 通过正则化方法, 网络参数得到有

效调整, 从而使全连接层提取的特征尽量满足 Fisher 判别准则, 最终有利于分类.

6 数值实验

6.1 实验平台与评价指标

实验平台为 Core i7-3.4 GHz, 8 GB 内存的 64 位 Windows 7 操作系统, 深层神经网络代码基于 Python 平台的 theano 工具集编写, 其中 theano 版本为 0.6rc. 我们使用以下几个性能指标来衡量分类模型的性能:

- 特异性 Sp: 真正的正常 ECG 有多少比例没有被误判, 也称为正常 ECG 检测灵敏度.
- 灵敏度 Se: 真正的异常 ECG 有多少比例能被正确检测出来, 也称为异常 ECG 检测灵敏度.
- 准确率 Acc: 正确分类的 ECG 比例.
- -P: 所有被分为正常的 ECG 中真正为正常的比例, 也称为正常 ECG 检测精度.
- AUC: ROC(受试者工作特征曲线) 曲线下的面积.
- TPR: ROC 曲线纵轴, 正异常分类中为特异性 Sp.
- FPR: ROC 曲线横轴, 正异常分类中为 (1- 灵敏度 Se).

使用 AUC 评价一个分类器性能的一个粗略的建议如下所示^[21]:

- 0.90–1.00=excellent classification;
- 0.80–0.90=good classification;
- 0.70–0.80=fair classification;
- 0.60–0.70=poor classification;
- 0.50–0.60=failure.

为简单起见, 本文所有数值实验均只采用监督学习, 没用无监督预训练过程; 另外为了避免过早地收敛到局部极值, 优化算法采用 Vogl 等^[22]提出的惯性量和变步长 BP 算法, 学习步长随着迭代次数变化而变化. 设定最大训练周期为 500, 初始步长为 0.02, 步长衰减率第 2, 3 周期为 0.9405, 其余为 0.99, 训练中以使得校验集准确率最高为最后训练结果.

6.2 实验数据集

① MIT-BIH 数据集

MIT-BIH 数据集共有 48 条 ECG 记录, 每条记录时长 30 分钟, 采样率为 360 Hz. 每条 ECG 记录由 2 个导联组成, 不过每条记录的导联并不完全相同, 仅 40 条记录均有 II 导联和 V1 导联. 很多文献不考虑导联一致性, 仅按通道处理数据, 显然不符合实际临床应用.

② 基于记录标注的 CCDD 数据集^{[23]1)}

我们课题组为算法测试开发的中国心血管疾病数据库网站上所有数据均来自临床诊断的 ECG, 从记录号 944 开始的数据均是按记录标注. 由于来源于真实的临床环境, 部分导联脱落、QRS 波形异常、无效记录等情况均是存在的. 具体实验时, 我们取 “data944–25693” 中一半多数据做训练, 剩余数据作为 “小规模测试集”; 再取 “data25694–179130” 作为大规模测试集, 具体细节见附录 A.

1) <http://58.210.56.164/ccdd/>.

表 1 HeartBeat-CNN 在 MITBIH 上的实验结果
Table 1 Experimental result of HeartBeat-CNN in MITBIH database

	Classification result		Specificity	Sensitivity	Accuracy
	Normal	Abnormal			
	0	1	99.00%	98.51%	98.89%
0	64617	650			
1	309	20434			

CCDD 数据集是分批的, 每批数据按如下方法处理: 每 560 条记录为一个包, 第 1 个包正异常样本数均为 280, 之后将正异常样本均匀分配到剩余包中. “data944-25693” 批次中, 校验样本共有 560 个, 训练样本共有 $22 \times 560 = 12320$ 个, 测试样本有 $21 \times 560 = 11760$ 个.

6.3 基于病人内的 ECG 心搏分类

为了和大多数发表的文献做对照, 我们同样用 MIT-BIH 数据集做了基于病人内的 ECG 分类, 采用文献 [4] 的数据处理方法, 从 48 条记录中提取 16 类心搏, 取 N 类为正常, 其他为异常, 做正异常分类 (每条记录的导联不完全相同, 不符合实际临床应用), 采用传统的卷积神经网络 (记为 HeartBeat-CNN), 具体实验步骤如下.

原始数据首先经过 0.5~25 Hz 的带通滤波, 用来滤除高频干扰和基线漂移. 在每个导联, 以 R 波为中心, 取 R 波前 100 个点, 后 199 个点, 包括 R 波在内共 300 个采样点组成一个心搏, 最后提取了 110110 个样本, 取 24100 个样本做训练 (与对照文献保持一致), 剩余样本做测试. 在保证了训练集和测试集一样的前提下, 将训练集做划分, 取 21020 个样本组成真正意义上的训练集, 剩余 3080 个样本组成校验集.

CNN 的三个卷积核分别为 [2,13], [1,10] 和 [1,6]; 三层取样核均为 [1,3], 即在 3 个点中取最大值; 三个特征面数分别为 10, 30 和 40 个, 两个全连接层的神经元数分别为 50 和 2. 训练参数设置如 6.1 小节所述, 采用加权处理不平衡集问题, 正异常权值之比为 3:2, 并加入噪声模块. 训练模型独立运行 10 次, 取准确率最高的作为对比模型.

由表 1 可知, HeartBeat-CNN 在测试集上的准确率达到 98.89%, 而特异性和灵敏度也都很高. 而同样的数据集, 原文献采用小波变换、ICA 和 PCA 进行特征提取, 然后由 SVM 进行分类, 最后用贝叶斯方法进行决策融合, 取得了 98.46% 的准确率 (原文献 16 类, 折合成 2 分类, 准确率再高一些). 再加上 4 个 RR 间期特征才达到 99.32% (折合成 2 分类, 准确率为 99.45%), 当然这个结果和 HeartBeat-CNN 已不具有可比性了.

由于是典型的不平衡集分类问题, 针对同样的数据集, 我们还用当前比较流行的不均衡偏最小二乘分类器 (asymmetric partial least squares classifier, Aplsc)^[24] 做了对比实验, 并尝试不同的主成分个数 (5,10,20,25,30,40,100,200,400,600), 当取 200 时, 分类模型取得 87.03% 的最高准确率.

由表 1 和表 2 可知, 相比 Aplsc 算法, CNN 在特异性、灵敏度、准确率都具有一定的优势. 但这么好的分类结果, 与其说是找到心电图的本质特征, 倒不如说是找到个体的独有特征, 一旦用 MIT-BIH 数据集以外的病人做心搏分类, 结果就会急剧下降.

6.4 面向临床应用的 ECG 分类实验

为了降低计算复杂度, 首先对 ECG 数据进行低通滤波, 然后下采样到 200 Hz, 大大减少了输入数

表 2 Aplsc 在 MITBIH 上的实验结果 (主成分 200 个)
Table 2 Experimental result of Aplsc in MITBIH database (200 principal components)

	Classification result		Specificity	Sensitivity	Accuracy
	Normal	Abnormal			
	0	1			
0	56535	8732	86.62%	88.31%	87.03%
1	2425	18318			

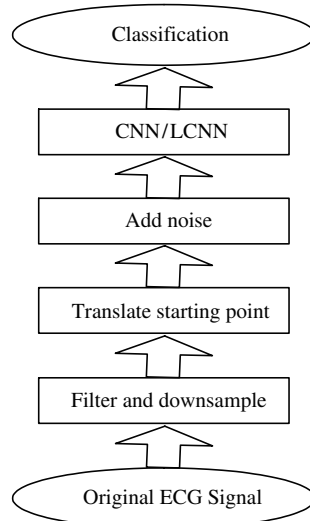


图 6 基于 CNN/LCNN 的 ECG 分类模型
Figure 6 ECG classification algorithm based on CNN/LCNN

据的维数. 按文献 [25] 规定, ECG 机频率响应最低上限为 75 Hz, 根据奈奎斯特 (Nyquist) 采样定律可知 150 Hz 采样频率就可以保留 ECG 中 75 Hz 以下的有用信号, 因此 200 Hz 是可以保留 ECG 中有用信号.

接着跳过开始 25 个采样点, 取中间 1900 个点的数据作为原始数据, 也即 9.5 s 数据. 多导联 ECG 数据有 8 个基本导联, 分别是 II, III, V1, V2, V3, V4, V5, V6 导联, 剩余 4 个导联由 8 个基本导联线性推导而出, 因此我们只保留 8 个导联心电图数据.

图 6 是基于 CNN/LCNN 的 ECG 分类算法框图. 原始 ECG 先经过滤波和下采样, 接着平移起始点和加噪处理, 最后输入 CNN/LCNN 进行分类. 测试阶段没有加噪处理, 并且起始点均为第 1 个采样点. 训练参数设置如 6.1 小节所述, 每个模型独立运行 10 次, 取 AUC 最大 (基于小规模测试集) 的作为对比模型.

为了直观显示各模块对分类性能的贡献, 我们逐步添加不同的模块. 首先用传统 CNN 对原始心电图数据 (输入维数为 8×1900) 做实验, 并分为加噪和不加噪两种情况, 以验证加噪模块对基于病人间的 ECG 记录分类的有效性. 三个卷积核分别尝试为 [3,18], [3,12] 和 [3,8], 三个取样核分别为 [1,7], [1,6] 和 [1,6], 三个特征面数分别为 8, 14 和 20, 两个全连接层的神经元数分别为 50 和 1. 由于没有平移起始点, 我们采用加权方式来处理数据不平衡问题, 正异常权值之比为 1:3, 并记该模型为 CNN[R]. 由表 3 可知, 加噪模块对最终分类模型在特异性和准确率方面提升明显, 但 AUC 和灵敏度方面有所

表 3 CNN[R] 在小规模测试集上的实验结果
Table 3 Experimental result of CNN[R] in small-scale test set

Noise model	AUC	Classification result			Specificity (%)	Sensitivity (%)	Accuracy (%)
		Normal	Abnormal				
No	0.8536	0	6235	2123	74.60	79.57	76.04
		1	695	2707			
Yes	0.8475	0	7147	1211	85.51	68.99	80.73
		1	1055	2347			

下降; 另外, 无论是否加噪, CNN[R] 的分类性能远远达不到临床应用要求.

为此, 我们在“噪声模块”的基础上继续添加“起始点模块”, 并采用“重采样技术”处理不平衡集问题. 这里设定起始点范围为 [1, 200], 即 1 s 数据长度 (可设置更长时间, 因为对于心率小于 60 次/分的 ECG, 1 s 数据还不能包括一个心搏的所有位置); 训练过程中, 每个样本随机选定起始点, 之后选取连续的 1700 个点作为神经网络的输入 (维数为 8×1700), 实际共有 $12320 \times 200 = 2464000$ 个样本参与训练 (训练样本只有 12320 个, 每个样本通过选取起始点拟成 200 个样本).

另外, 为了验证 LCNN 的有效性, 我们将 LCNN 和不同结构 CNN 做对比. 但是不同网络结构不具有可比性, 因此要事先设定一定条件, 比较才有意义. 对于前向神经网络, 每两层之间的神经元不论采用何种连接方式 (如卷积层、抽样层等), 均可看作是全连接层的特例. 因此, 我们设定条件是不论如何调整参数, 必须保证层数及每层神经元数相同; 至于每两层之间的神经元如何连接, 可看作是不同的正则化方法, 在 5.2 小节已有说明.

为了将 LCNN 和 CNN 做比较充分的对比, 我们设定了 4 种结构的 CNN, 分别记为 CNN[A], CNN[B], CNN[C], CNN[D], 再加上 LCNN, 一共 5 种模型, 具体参数设置如下:

- ① LCNN: 三个卷积核分别为 [1,21], [1,13] 和 [1,9], 三个特征面数分别为 6, 7 和 5.
- ② CNN[A]: 三个卷积核分别为 [1,21], [1,13] 和 [1,9], 三个特征面数分别为 6, 7 和 5.
- ③ CNN[B]: 三个卷积核分别为 [3,21], [3,13] 和 [3,9], 三个特征面数分别为 8, 14 和 20.
- ④ CNN[C]: 三个卷积核分别为 [6,21], [2,13] 和 [1,9], 三个特征面数分别为 16, 28 和 20.
- ⑤ CNN[D]: 三个卷积核分别为 [8,21], [1,13] 和 [1,9], 三个特征面数分别为 48, 56 和 40.
- ⑥ 所有模型的三个取样核均分别为 [1,7], [1,6] 和 [1,6].
- ⑦ 所有模型的两个全连接层神经元数均分别为 50 和 1.

以上参数设置保证了 5 种模型的网络层数及每层神经元数均相同.

我们首先将 CNN[R] 加噪模型和 CNN[B] 模型做对比, 两者网络结构相似, 且均包含加噪模块, 区别是一个采用“平移起始点 + 重采样”训练方式, 另一个采用“加权”训练方式. 由表 4 可知, CNN[B] 模型无论是特异性、灵敏度、准确率, 还是 AUC 指标, 均超越 CNN[R] 加噪模型, 最终分类性能提升明显.

虽然 CNN[B] 模型的各项性能指标都相当不错, 但这毕竟只是“小规模测试集”的实验结果 (测试样本仅为 11760 个), 与实际临床应用 (测试样本为无穷大) 还存在相当大的距离.

为此, 我们将 CNN[B] 模型在“大规模测试集”做了分类实验 (测试样本为 151274 个), 也做了 CNN[A] 模型、CNN[C] 模型、CNN[D] 模型, 以及 LCNN 模型的分类实验. 我们分别从计算时间 (分类一条记录所需的时间) 和临床应用指标来对比以上 5 种模型的性能, 具体实验结果见表 5.

表 4 不同分类模型在小规模测试集上的实验结果

Table 4 Experimental result of different classification model in small-scale test set

Model	AUC	Classification result			Specificity (%)	Sensitivity (%)	Accuracy (%)
		Normal	Abnormal				
		0	1				
CNN[R]	0.8475	0	7147	1211	85.51	68.99	80.73
		1	1055	2347			
CNN[B]	0.8948	0	7487	871	89.58	73.46	84.91
		1	903	2499			

表 5 不同分类模型在大规模测试集上的性能指标

Table 5 Performance index of different classification model in large-scale test set

Model	Time (ms)	Acc (%)	AUC	FPR=1%		FPR=5%		FPR=10%		-P=95%		-P=90%	
				TPR	-P	TPR	-P	TPR	-P	FPR	TPR	FPR	TPR
LCNN	2.82	83.66	0.9086	16.0	95.4	53.4	93.2	72.6	90.3	1.81	26.7	10.6	73.9
CNN[A]	2.80	82.90	0.9013	13.8	94.7	49.5	92.7	70.1	90.0	0.37	5.50	10.0	70.1
CNN[B]	6.60	82.90	0.8977	13.9	94.7	47.5	92.4	68.1	89.8	0.41	6.04	9.59	67.0
CNN[C]	9.32	81.75	0.9001	15.9	95.3	50.0	92.8	69.3	89.9	1.78	26.2	9.86	69.0
CNN[D]	11.9	83.46	0.9025	16.6	95.5	51.5	93.0	70.6	90.1	1.50	22.1	10.2	71.0

由表 5 可知, 相比传统的 CNN, 本文专门针对多导联 ECG 而设计的 LCNN, 无论在计算时间, 还是临床应用指标, 都具有一定的优势. 而基于 CNN 设计的 4 种模型中, CNN[A] 模型的不同导联数据在同一特征面强制使用相同卷积核 (滤波器), CNN[B] 和 CNN[C] 模型的部分导联数据存在卷积融合计算, CNN[D] 模型则一次性对所有导联数据进行卷积融合计算. 相比而言, CNN[D] 模型的各种临床应用指标都相当不错, 但其计算时间也是最耗时的. 实验结果同 5.2 小节的理论分析结论不谋而合.

为了进一步验证所提心电图分类模型的有效性, 我们将其与传统人工特征提取的分类方法做对比. 文献 [26] 提出的 ECG-MTHC 模型融合了四个分类模块, 包括 RR 间期正常分析、QRS 波群相似性分析、基于数值特征和 QRS 波群形态特征的统计分类器 (SVM)、心电图典型特征 (P 波、T 波、R 波、PR 间期等) 分析, 最终给出一条记录的分类结果.

在表 6 中, ECG-MTHC 模型实际仅给出 14 多万条记录的分类结果, 剩余 1 万多条记录由于中间特征提取出错而无法给出. 由此也反映了 LCNN 模型的健壮性, 更能适应实际临床环境.

为了参照对比, 我们给出了 LCNN 模型在相同测试集的实验结果, 此外也给出全部测试集的结果. 由表 6 可知, LCNN 模型在特异性和准确率方面明显高于 ECG-MTHC 模型, 不过灵敏度要差一些. 灵敏度之所以差, 一方面是因为训练集中正常样本是异常样本的 2 倍多; 另一方面, 考虑到 CCDD 数据库 ECG 记录的长度, 我们设定起始点范围为 [1, 200] (基于 200 Hz 采样率), 这意味着心率小于 60 次/分的记录的起始点不能完全覆盖. 通过采集更符合要求的的心电图数据作为训练样本, 并融入基于医生规则的分类模型如 RR 间期等, 可非常有效地提高灵敏度, 而这也是我们下一步要开展的研究工作.

实际上, 现有心电图临床应用的一些方法准确率普遍不是很高. 2012 年 Hakacova 等 [27] 对市场心电图机的自动诊断结果进行统计分析, 总共统计了 500 例心电图, 发现 Philips medical 准确率只

表 6 不同分类模型在小规模测试集上的实验结果
Table 6 Experimental result of different classification model in large-scale test set

Model	Classification result		Specificity (%)	Sensitivity (%)	Accuracy (%)
	Normal	Abnormal			
	0	1			
ECG-MTHC	0	42767	54.62	95.13	72.49
	1	3011			
LCNN	0	65574	83.75	83.67	83.72
	1	10091			
LCNN	0	71378	83.84	83.43	83.66
	1	10961			

有 80%, 而 Draeger medical systems 只有 75%, 同时该报告也给出 3 名普通医生的心电图判读准确率平均为 85%. 本文所提的心电图分类模型在 15 万多条记录上取得了 83.66% 的准确率, 具有显然的参考和应用价值.

7 结论

本文对面向临床应用的 ECG 分类算法进行了深入的研究, 并专门针对多导联 ECG 分类问题设计了导联卷积神经网络, 而“平移起始点”和“加噪”两个策略使得 BP 算法更加有效的寻找最优参数. 通过测试 15 多万条心电图数据, 我们的模型取得了准确率为 83.66%, AUC 为 0.9086 (excellent classification) 的结果. 据此, 我们同上海宏康信息技术有限公司合作, 将训练所得的分类模型移植到 Android 移动平台, 实现了心率判断、早搏等七种病症的识别功能.

网络结构参数设定和优化、起始点范围选择, 以及分类器融合途径等都有待进一步研究. 我们将不断以既有思想和方法为基础^[28], 并吸纳新的理论和技术, 为计算机辅助 ECG 分析方法的广泛的应用不断探索和努力.

致谢 非常感谢朱洪海博士所做的基础工作, 以及张高登硕士在采集和处理心电图数据方面所做的工作.

参考文献

1 Clifford G D, Azuaje F, McSharpy P E. Advanced Methods and Tools for ECG Data Analysis. London: Artech House, 2006

2 Zhu H H, Dong J. An R-peak detection method based on peaks of Shannon energy envelope. Biomed Signal Process Control, 2013, 5: 466-474

3 Goldberger A L, Amaral L A N, Glass L, et al. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation, 2000, 101: 215-220

4 Ye C, Kumar B V, Coimbra M T. Heartbeat classification using morphological and dynamic features of ECG signals. IEEE Trans Biomed Eng, 2012, 59: 2930-2941

5 Kutlu Y, Kuntalp D. A multi-stage automatic arrhythmia recognition and classification system. Comput Biol Med, 2011, 41: 37-45

- 6 Übeyli E D. Combining recurrent neural networks with eigenvector methods for classification of ECG beats. *Digit Signal Process*, 2009, 19: 320–329
- 7 Übeyli E D. Recurrent neural networks with composite features for detection of electrocardiographic changes in partial epileptic patients. *Comput Biol Med*, 2008, 38: 401–410
- 8 Lei W K, Li B N, Dong M C, et al. An application of morphological feature extraction and support vector machines in computerized ECG interpretation. In: *Proceedings of 6th Mexican International Conference on Artificial Intelligence*, Aguascalientes, 2007. 82–90
- 9 Li X H, Shu L, Hu H L. Kernel-based nonlinear dimensionality reduction for electrocardiogram recognition. *Neural Comput Appl*, 2009, 18: 1013–1020
- 10 Doquire G, de Lannoy D, Francois D, et al. Feature Selection for Interpatient Supervised Heart Beat Classification. In: *Proceedings of International Conference on Bio-inspired Systems and Signal Processing*, New York, 2011. 67–73
- 11 Martis R J, Chakraborty C, Ray A K. A two-stage mechanism for registration and classification of ECG using Gaussian mixture model. *Pattern Recogn*, 2009, 11: 2979–2988
- 12 de Chazal P, O'Dwyer M, Reilly R B. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Trans Biomed Eng*, 2004, 51: 1196–1206
- 13 Wubbeler G, Stavridis M, Kreiseler D, et al. Verification of humans using the electrocardiogram. *Pattern Recogn Lett*, 2007, 28: 1172–1175
- 14 Sutskever I, Hinton G E. Deep narrow sigmoid belief networks are universal approximators. *Neural Comput*, 2008, 20: 2629–2636
- 15 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Comput*, 2006, 18: 1527–1554
- 16 Lauer F, Suen C Y, Bloch G. A trainable feature extractor for handwritten digit recognition. *Pattern Recogn*, 2007, 40: 1816–1824
- 17 Martinez H P, Bengio Y, Yannakakis G N. Learning deep physiological models of affect. *IEEE Comput Intell Mag*, 2013, 8: 20–33
- 18 Lee H, Battle A, Raina R, et al. Efficient sparse coding algorithms. In: *Proceedings of NIPS*, 2006. 801–808
- 19 LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. In: *Proceedings of the IEEE*, New York, 1998. 2278–2324
- 20 Masci J, Meier U, Ciresan D, et al. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In: *Proceedings of the 21th International Conference on Artificial Neural Networks*. Berlin: Springer, 2011. 52–59
- 21 Gorunescu F. *Data Mining Concepts Techniques and Models*. Berlin: Springer, 2011. 185–317
- 22 Vogl T P, Mangis J K, Rigler A K, et al. Accelerating the convergence of the back-propagation method. *Biol Cyber*, 1988, 59: 257–263
- 23 Zhang J W, Liu X, Dong J. CCDD: An enhanced standard ECG database with its management & annotation tools. *Int J Artif Intell Tool*, 2012, 21: 1–26
- 24 Qu H N, Li G Z, Xu W S. An asymmetric classifier based on partial least squares. *Pattern Recogn*, 2010, 43: 3448–3457
- 25 China Food and Drug Administration. YY 1139–2000 single and multichannel electrocardiograph Version 1.0. 2000 [国家食品药品监督管理局. YY 1139–2000 单道和多道心电图机版本 1.0. 2000]
- 26 Wang L P. Study on approach of ECG classification with domain knowledge. Dissertation for Ph.D. Degree. Shanghai: East China Normal University, 2013 [王丽苹. 融合领域知识的心电图分类方法研究. 博士学位论文. 上海: 华东师范大学, 2013]
- 27 Hakacova N, Trägårdh-Johansson E, Wagner G S, et al. Computer-based rhythm diagnosis and its possible influence on nonexpert electrocardiogram readers. *J Electrocardiol*, 2012, 45: 18–22
- 28 Dong J, Zhang J W, Zhu H H, et al. Wearable ECG monitors and its remote diagnosis service platform. *IEEE Intell Syst*, 2012, 27: 36–43

附录 A

中国心血管疾病数据库网站上所有数据均来自实际环境, 因而存在部分无效数据. 我们将降维为 200 Hz 后数据个数不足 1925 以及标注为 “0x00” (无效) 或无标注的记录丢掉; 然后取标注为 “0x0101” 和 “0x020101” 的记录为正常, 其他都为异常. 这样处理后各批数据集的正异常分布情况如表 A1 所示.

小规模数据集均来自 “data944–25693”. 考虑批组整除原因, 我们去掉了 29 条记录, 其编号如下: 1173, 1257, 1464,

表 A1 数据分布情况
Table A1 Data distribution

Dataset	Normal	Abnormal	Total
data944-25693	17467	7202	24669
data25694-37082	4911	6352	11263
data37083-72607	25020	10249	35269
data72608-95829	16210	6508	22718
data95830-119551	10351	12948	23299
data119552-141104	9703	11529	21232
data141105-160913	9713	9831	19544
data160914-175871	6944	7781	14725
data175872-179130	2289	935	3224

1549, 1593, 1791, 1817, 2078, 2193, 2235, 2265, 2289, 2313, 2338, 2958, 2970, 3023, 3130, 3305, 3396, 3557, 3676, 3893, 4296, 4305, 4585, 25607, 25692-25693, 这样小规模数据集的共有 24640 个样本.

训练集的记录编号为 4520-4584, 4586-4613, 4615-4652, 4654-4761, 4763-4967, 4969-4972, 4975-5146, 5148, 5151-5279, 5281-5300, 5302-5348, 5350-5540, 5542-5568, 5570-5713, 5715-5777, 5779, 5781-5792, 5794-5974, 5976-6074, 6076-6118, 6120-6127, 6129-6134, 6136-6206, 6208-6281, 6283-6441, 6443-6502, 6504-6538, 6540-6654, 6656-6997, 6999-7005, 7007-7012, 7014-7060, 7062-7451, 7453-7506, 7508-7531, 7533-7594, 7596-7642, 7644-7732, 7734-7739, 7741-7829, 7831-7887, 7889-7939, 7941-7946, 7948-7956, 7980-8064, 8066-8108, 16556-17045, 17047-17128, 17407-17422, 17424-17454, 17456-17928, 17930-17933, 17935-17955, 17957-18093, 18095-18258, 18260-18441, 18443-18538, 18540-18562, 18565-18642, 18644-18814, 18816-18817, 18819-18984, 18986-19327, 19329-19439, 19441-19647, 19649-19657, 19659-19852, 19854-20173, 20175-20700, 20702-20881, 20883-21252, 21254-21301, 21303-21586, 21588-21815, 21817-21865, 21867-21889, 21891-21986, 21988-22149, 22151-22226, 22228-22462, 22464-22623, 22625-22793, 22795-22935, 22937-23032, 23034-23038, 23040-23043, 23045-23067, 23069-23268, 23270-23587, 23589-23611, 23613-23973, 23975-24108, 24110-24646, 24648, 24650-24775, 24777-24820, 24822-24962, 24964-25201, 25203-25282, 25284-25301, 25303-25330, 25332-25457, 25459-25495, 25497-25606, 25608-25691, 共有 12320 个样本.

校验集的记录编号为 4176-4278, 4362-4413, 4415-4519, 7957-7979, 17129-17389, 17391-17406, 共有 560 个样本. 小规模测试集由 “data944-25693” 中的剩余样本组成, 共有 11760 个样本. 大规模测试集由除 “data944-25693” 之外的所有其他数据集组成, 如表 A1 所示, 共有 151274 个样本.

Deep learning research on clinical electrocardiogram analysis

JIN LinPeng^{1,2} & DONG Jun^{1*}

¹ *SuZhou Institute of Nano-tech and Nano-bionics, Chinese Academy of Sciences, Suzhou 215123, China;*

² *University of Chinese Academy of Sciences, Beijing 100049, China*

*E-mail: jdong2010@sinano.ac.cn

Abstract As one of the classical applications of pattern recognition research, electrocardiogram (ECG) classification has important application values for wearable ECG devices and “cloud” service platforms. In this paper, first of all, the complexity of the ECG classification model for clinical application is illustrated. Consequently, the approximation ability of a nonlinear function in an existing feature extraction and classification algorithm is analyzed, and deep learning is employed for ECG classification. Then, lead convolutional neural networks (LCNN) is presented considering the special two-dimensional structure of multi-lead ECG, in which “translating starting

point” and “adding noise” are two of the main strategies to increase the training sample. Tests conducted using more than 150,000 ECG records show that the proposed method has an accuracy of 83.66% and 0.9086 AUC. Finally, the classification model is implemented on a mobile terminal, where its real-time analysis performance is shown to meet application requirements.

Keywords deep learning, electrocardiogram (ECG), nonlinear approximation, convolutional neural networks, data distribution and classification



JIN LinPeng was born in 1984. He received an M.S. degree in computer application technology from Ningbo University, Ningbo City, in 2010. He is currently a Ph.D. candidate at the Suzhou Institute of Nano-tech and Nano-bionics, Chinese Academy of Sciences. His research interests include pattern recognition and artificial intelligence.



DONG Jun was born in 1964. He received a Ph.D. degree from Zhejiang University in 1997. He is currently a “Hundred Talent Program” Professor at the Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Sciences. His research interests include computer-aided cardiovascular diseases analysis and calligraphy creation simulation. Dr. Dong is an adjunct professor at Xiamen University.