

基于深度学习的疾病诊断 *

陆家发

张国明

陈安琪

(江苏省卫生统计信息中心
南京 210008)

(1 江苏省卫生统计信息中心 南京 210008
2 南京大学计算机科学与技术系 南京 210023)

(江苏健康职业学院
南京 210029)

〔摘要〕 以 UCI 心脏病数据集为试验测试数据, 利用机器学习中的深度学习技术进行疾病诊断分类, 分别与随机森林、支持向量机以及神经网络分类进行对比, 指出深度学习技术对医疗大数据的挖掘具有巨大潜力和重要的应用价值。

〔关键词〕 数据挖掘; 深度学习; 疾病诊断

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.04.010

Disease Diagnosis Based on Deep Learning LU Jia-fa, Health Statistics and Information Center of Jiangsu Province, Nanjing 210008, China; ZHANG Guo-ming, 1 Health Statistics and Information Center of Jiangsu Province, Nanjing 210008, 2 Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China; CHEN An-qi, Jiangsu Jiankang Vocational College, Nanjing 210029, China

〔Abstract〕 The paper takes heart disease dataset of UCI as experimental data, carries out disease diagnosis and classification based on the deep learning technology of machine learning, compares with Random Forest (RF), Support Vector Machine (SVM) and Neural Network (NN) classification, and indicates that the Deep Learning (DL) technology is of great potential and important value for exploration of medical big data.

〔Keywords〕 Data mining; Deep learning; Disease diagnosis

1 引言

目前, 我国处于医疗信息化的飞速发展时期, 全国各地正在建设以电子病历为核心的区域健康信息平台, 医疗数据量呈井喷式增长, 其中包括患者

的诊疗数据、费用数据、就诊时间序列数据以及相关管理数据等, 医疗卫生数据的获取日益便捷。但隐含在海量医疗数据中的有效信息并没有得到充分挖掘和有效利用。同时, 现阶段医疗卫生事业也存在很多问题, 如医生超负荷工作、医患关系紧张、医疗费用高昂等。医院作为主体, 需要降低医疗成本, 主要是最小化诊断测试的成本, 同时提高疾病诊断的效率和正确性。当前, 医疗诊断主要依靠医生的专业知识和丰富的诊断经验进行判别并给出治疗方案, 而没有利用丰富的历史医疗数据。如何从海量的医疗数据中发现隐含的有用信息, 从而对医疗决策提供诊断辅助并提高医疗服务水平是解决问

〔修回日期〕 2017-03-10

〔作者简介〕 陆家发, 助理工程师。

〔基金项目〕 江苏省卫生和计划生育委员会科研课题“基于区域健康信息平台的医疗大数据挖掘平台研究与设计”(项目编号: X201601)。

题的关键所在。

数据挖掘技术可以从海量数据中学习特定的模式,从而提取更多的有用信息,如神经网络、决策树、最大似然以及基于统计知识的逻辑回归和支持向量机等^[1-2]。以上方法都可以看成是带有 1 层隐层节点或没有隐层节点的浅层机器学习模型^[3],其局限性在于有限样本和计算单元情况下对复杂函数的表示能力有限,针对复杂分类问题其泛化能力不强^[4]。近年来,以深度学习为代表的机器学习在医疗领域的应用越来越多,在语音、图像识别等方面都取得了很好的效果^[5],但在医疗数据挖掘中还没有成熟的应用。鉴于此,本研究将深度学习技术应用于疾病诊断,探索深度学习在医疗大数据挖掘中的可行性。

2 浅层机器学习模型

2.1 支持向量机

支持向量机^[6] (Support Vector Machine, SVM) 是 Corinna Cortes 和 Vapnik 等于 1995 年首先提出的,是一种使用核方法的非线性分类器,与神经网络、决策树等方法相比, SVM 在训练样本数量有限的情况下可以取得较好的分类结果。其思想是用核函数映射到特征空间,找到一个分类超平面,所有计算都在原空间中完成,而无须在特征空间进行大规模运算。给定训练集 $S = \{(x^1, y_1), \dots, (x^l, y_l)\} \in R^n \times \{-1, +1\}$, 通过求解其对应的凸优化问题得到决策函数:

$$\max_{\alpha} g(\alpha) = \sum_{i=1}^l \alpha_i y_i k(x^i, x^i) \quad (1)$$

$$0 \leq \alpha_i \leq C, \sum_{i=1}^l \alpha_i y_i = 0$$

式中: α 是 Lagrange 乘子; C 是惩罚系数; $k(\cdot)$ 是满足 Mercer 条件的核函数; ξ_i 是惩罚因子。

在众多核函数中,高斯核函数最为经典,其形式为:

$$k_{\sigma}(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2) \quad (2)$$

式中: $\|x_i - x_j\|^2$ 为欧氏范数, $\sigma \in R^+$ 用于调整核函数的灵活性。当找到式 (1) 的最优解 α_i 时,对样本 x 的分类可通过判定其属于超平面的哪一侧来确定:

$$y = \text{sgn} \left(\sum_{i=1}^l \alpha_i y_i k(x^i, x) + b \right) \quad (3)$$

为处理多类问题,通常采取一对一或一对多的二元分类策略。本试验采用高斯核函数,通过对训练样本进行 boot 采样选择惩罚系数 C 和宽度 σ 。

2.2 随机森林

随机森林 (Random Forests, RF) 是一种以决策树为基本分类器的集成学习模型,由 Leo Breiman 于^[7]2001 年提出,通过组合多种决策树的预测,投票得出最终的预测结果。随机森林是一种非线性建模工具,大量理论和实证研究表明该算法具有较高的正确率,对异常值和噪声具有很好的容忍度,且不易出现过拟合。随机森林分类的基本思想是^[8]: 首先,利用 bootstrap 抽样从原始训练集抽取 k 个样本,且每个样本的容量都与原始训练集相同;其次,对 k 个样本分别建立 k 个决策树模型,得到 k 种分类结果;最后,根据 k 种分类结果对每个记录进行投票,得出最终分类结果。随机森林通过构造不同训练集增加分类模型间的差异,提高组合分类模型的外推预测能力。通过 k 轮训练,得到一个分类模型序列 $\{h_1(X), h_2(X), \dots, H_k(X)\}$,再利用分类模型序列构成一个多分类模型系统,该系统的最终分类结果采用简单多数投票法。最终的分

$$H(X) = \arg \max_Y \sum_{i=1}^k I(h_i(x) = Y) \quad (4)$$

式中: $H(X)$ 表示组合分类模型; h_i 表示单个决策树分类模型; Y 表示输出变量, $I(\cdot)$ 为示性函数。

3 神经网络和深度学习模型

3.1 神经网络

人工神经网络^[9-10] (Artificial Neural Network, ANN) 是参照生物神经网络而发展起来的模拟人脑生物过程的人工智能技术,是由大量人工神经元节点互连形成的非线性系统,能够模拟复杂的非线性函数。人工神经元是神经网络的基本处理单元,每个神经元包含多个输入和 1 个输出。数据通过神经

元之间的连接传输, 神经元与各输入间的连接强度称为权重, 权重决定了输入信号的兴奋或抑制。神经元通过激励函数 (Activation Function) 将所有经过权重调整后的输入相加, 形成一个激励值, 然后根据激励值产生神经元的输出。网络输出依网络的连接方式、权重和激励函数的不同而不同。前馈神经网络是人工神经网络的一种典型结构, 该网络只在训练过程中有反馈信号, 而在分类过程中数据只能向前传送, 直至到达输出层, 层间没有向后的反馈信号。本文使用前馈神经网络进行试验。

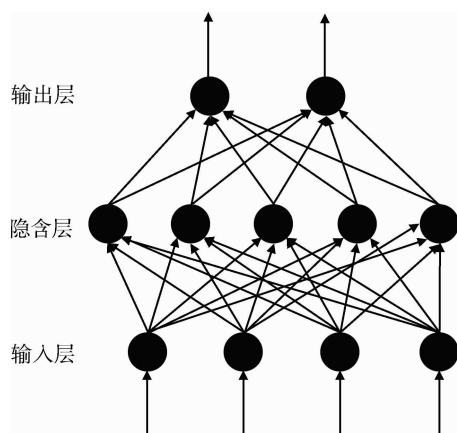


图 1 3 层前馈神经网络模型

图 1 为一个 3 层的前馈神经网络, 其中第 1 层为输入层, 第 2 层为隐含层, 第 3 层为输出层。假设输入层、隐含层和输出层的神经元个数分别为 I 、 J 和 K , 输入为 $(x_0, x_1, x_2, \dots, x_{i-1})$, 隐含层输出为 $(h_0, h_1, h_2, \dots, h_{j-1})$, 输出层为 $(y_0, y_1, y_2, \dots, y_{k-1})$, 输入层神经元 i 到隐含层神经元 j 的权值为 v_{ij} , 隐含层神经元 j 到输出层神经元 k 的权值为 w_{jk} , θ_j 和 θ_k 分别表示隐含层与输出层神经元的阈值, 设激活函数为 f 。该神经网络隐含层各神经元的输出值为:

$$h_j = f\left(\sum_{i=0}^{i-1} v_{ij} x_i - \theta_j\right) \quad (5)$$

输出层各神经元的输出值为:

$$y_k = f\left(\sum_{j=0}^{j-1} w_{jk} h_j - \theta_k\right) \quad (6)$$

若激活函数 f 选用线性函数, 那么神经网络的输出 y_k 将是输入 x_i 的线性函数。因此, 若要做高次函数的逼近则需选用适当的非线性函数作为激活函数, 如 S 函数和双曲函数。

3.2 深度学习

深度学习^[3,11] (Deep Learning, DL) 算法的概念源于人工神经网络的研究, 是对神经网络的发展, 与神经网络具有相似的分层结构。相比于传统神经网络, 深度学习有了重大的改进, 其通过构建具有多个隐层的非线性网络结构来实现复杂的函数逼近, 通过海量的训练数据学习更抽象的深层次特征来提升分类或预测的准确性, 在训练难度上 (如梯度弥散问题) 通过逐层训练来有效降低。而且深度学习可以自动学习特征, 而不必请专家手工构造特征, 极大推进智能化、自动化。通过逐层的特征学习, 将样本在原空间转化到新的特征空间, 得到更易于分类或预测结果的输出。通过海量数据的不断学习, 将低层次的特征进行抽取或组合, 形成更加抽象的高层次特征, 发现数据的内在规律, 更能够刻画数据的本质和丰富的内在信息。用深度学习进行医疗疾病诊断具有如下优势: 充分挖掘海量医疗数据中隐含的深层关系, 通过对深度网络的学习获得具有高判别能力的更抽象、更高层的特征表示, 提高疾病诊断的正确率。

4 试验验证

4.1 数据源

使用公开 UCI 的心脏病数据集 (270 个样本) 作为测试数据, 对 4 种不同分类算法 (支持向量机、随机森林、神经网络和深度学习) 进行对比分析。数据集包括与心脏病相关的 13 个属性描述, 如年龄、性别、心绞痛类型、静息血压、胆固醇、心电图、空腹血糖、最高心率值、是否运动引起的心绞痛、缺陷类型等。本试验采用 4 种算法对 UCI 数据集中的心脏病数据进行测试。

4.2 试验对比

选取 70% 样本作为训练样本, 30% 样本作为测试样本, 采用 boot (有放回抽样) 采样做 20 次试验, 计算灵敏度 (Accuracy)、特异性 (Sensitivity) 和正确率 (Specificity) 的平均值。

使用 R 语言作为试验测试平台，模型训练时通过对训练样本进行 10 次 boot 采样选择参数。深度学习模型包含两层隐含层，神经元数分别为 15 和 13。神经网络为包含 1 个隐含层的前馈神经网络，其神经元数目为 13。试验使用混淆矩阵进行分类结果评价，见表 1。*TP* 表示正样本（有疾病）被正确分类，*TN* 表示负样本（无疾病）被正确分类，*FN* 和 *FP* 表示样本被误分类，如正样本（有疾病）被误分类为健康，负样本（无疾病）被误分类为心脏病。通常用以下 3 个公式来计算灵敏度、特异性和正确率^[12]：

$$Sensitivity = TP / (TP + FN)$$
$$Specificity = TN / (TN + FP)$$
$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$

表 1 混淆矩阵

样本分类	0（无心脏疾病）	1（有心脏疾病）
0（无心脏疾病）	<i>TP</i>	<i>FN</i>
1（有心脏疾病）	<i>FP</i>	<i>TN</i>

支持向量机（SVM）、随机森林（RF）、神经网络（NN）和深度学习（DL）4 种算法对应的灵敏度、特异性和正确率，见表 2。不同分类算法对应的正确率、灵敏度和特异性的曲线，见图 2。由试验结果可知，在分类正确率上，深度学习算法与 SVM 算法基本一致，深度学习的平均正确率略高于 SVM；深度学习的最大值、最小值和平均值明显优于随机森林和传统神经网络算法，特别是对比分类正确率的最小值可以看出传统神经网络最小值为 51.5%，而深度学习为 79.01%，表明深度学习克服了传统神经网络泛化性差的缺点，具有较好的算法稳定性。在灵敏度方面，深度学习明显优于 SVM、随机森林和神经网络。在特异性方面，虽然深度学习算法的最大值低于 SVM，但平均值和最小值表现出较好的分类性能。在较小的 UCI heart 数据集上，通过深度学习对测试数据的不断抽象，获得更好的特征表示，相较于原数据来说，更具有可分性，取得了更好的分类效果。相较于传统的浅层机器学习模型，对大数据的分析是深度学习的优势，

医疗数据量大，来源广泛，更适合对其进行深度学习以得到高层次特征，因此深度学习技术对医疗大数据的分析利用具有重要的意义。

表 2 NN、RF、SVM 和 DL 算法
对应的灵敏度、特异性和正确率

结果		NN (%)	RF (%)	SVM (%)	DL (%)
灵敏度	最大值	85.19	87.65	88.89	88.89
	最小值	51.85	75.31	80.25	79.01
	平均值	79.14	80.8	83.52	83.89
特异性	最大值	90.91	92.31	93.18	94.44
	最小值	63.16	69.39	77.55	79.25
	平均值	81.17	82.67	85.29	86.18
正确率	最大值	90.62	93.75	94.74	89.29
	最小值	40.35	59.52	67.57	73.53
	平均值	77.62	78.45	80.99	80.99

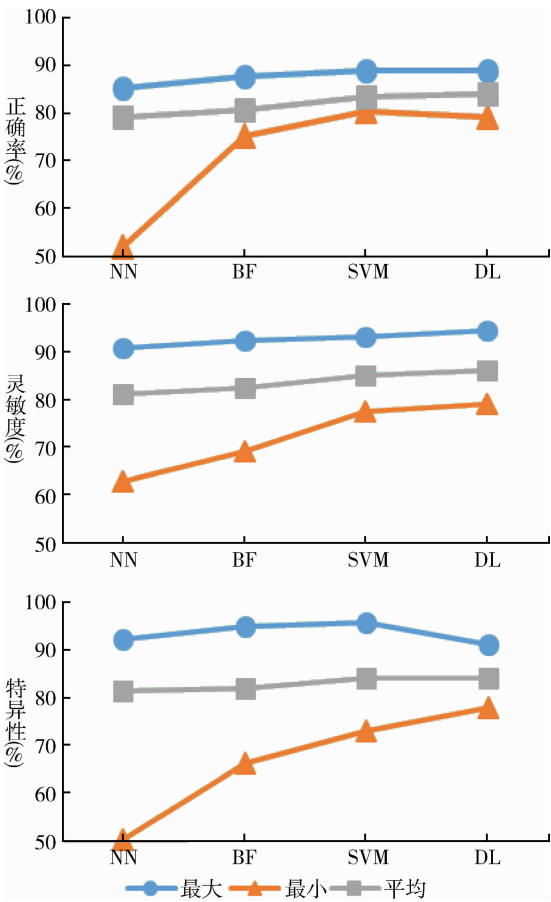


图 2 不同分类算法对应的正确率、灵敏度和特异性曲线

5 结语

针对海量医疗数据中蕴含的大量隐含信息,对其进行有效分析和处理是医疗数据的重要价值所在。设计出区分度高、可判别能力强的特征通常需要耗费大量精力,深度学习可以从海量数据中学习有效特征,随着医疗大数据时代的到来,深度学习将具有巨大的发展潜力。本文利用深度学习技术从数据中提取深度特征,模拟人脑识别过程逐层抽象,结果表明,与 SVM、RF 和 NN 这 3 种浅层网络分类方法相比,深度学习的灵敏度、特异性和正确率都有提高,因此将深度学习应用于医疗诊断具有重要的理论和应用价值。依托各地区的区域健康信息平台,随着医疗数据的不断积累,深度学习模型可以更好地对疾病做出正确诊断,提高已有医疗数据对疾病预测的参考价值,降低诊断成本的同时减少误诊率,提高病人就诊效率和医疗服务水平。

参考文献

- 1 张晔,张晗,尹玢臻,等. 基于支持向量方法机构建急性胰腺炎患者住院天数预测模型 [J]. 医学信息学杂志, 2016, 37 (2): 57 - 60.
- 2 张睿,王觅也,李楠,等. 数据挖掘技术及其在临床恶性肿瘤诊疗中的应用 [J]. 医学信息学杂志, 2015, 36

- (10): 50 - 54.
- 3 孙志军,薛磊,许阳明,等. 深度学习研究综述 [J]. 计算机应用研究, 2012, 29 (8): 2806 - 2810.
- 4 Bengio Y. Learning Deep Architectures for AI [J]. Foundations and Trends in Machine Learning, 2009, 2 (1): 1 - 127.
- 5 余凯,贾磊,陈雨强,等. 深度学习的昨天、今天和明天 [C]. 中国计算机学会人工智能会议, 2013: 1799 - 1804.
- 6 Chang CC, Lin CJ. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology, 2011, (27): 1 - 27.
- 7 Breiman L. Random Forests [J]. Machine Learning, 2001, 45 (1): 5 - 32.
- 8 方匡南,吴见彬,朱建平,等. 随机森林方法研究综述 [J]. 统计与信息论坛, 2011, 26 (3): 32 - 38.
- 9 卢金秋. 数据挖掘中的人工神经网络算法及应用研究 [D]. 杭州: 浙江工业大学, 2006.
- 10 A Brief Introduction to Neural Networks [EB/OL]. [2016 - 05 - 10]. http://www.dkriesel.com/en/science/neural_networks, 2007.
- 11 Hinton G, Osindero S, Teh Y. A Fast Learning Algorithm for Deep Belief Nets [J]. Neural Computation, 2006, 18 (7): 1527 - 1554.
- 12 Milan K, Sunila G. Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction [J]. International Journal of Computer Science and Technology, 2011, 2 (1): 304 - 308.

(上接第 38 页)

5 结语

进入新世纪以来,信息化在各个领域迅速推进,深刻影响着全球经济社会的发展,信息化水平已成为衡量一个国家或地区现代化水平的标志之一。医院信息集成平台是医院信息化建设的核心,对医院的发展发挥重要支撑作用,对提升医院的整

体核心竞争力具有深远意义。

参考文献

- 1 张小亮,景慎旗,朱雨倩,等. 大型医院集成平台建设的实践思考 [J]. 中国数字医学, 2016, 11 (10): 76 - 79.
- 2 张耀海,肖建芳. 医院信息系统集成平台建设与体会 [J]. 医学信息, 2016, (29): 3 - 4.
- 3 陈敏,刘宁. 我国医疗信息化发展方向及热点分析 [J]. 中华医院管理杂志. 2016, 32 (8): 601 - 603.