机器学习在肿瘤早期诊断与预后预测中的应用*

施 维 薛 均 潘璀然 任元凯 倪正杰 张远鹏 王 理 吴辉群 蒋 葵 董建成

（南通大学医学院医学信息学系 南通 226001）

〔摘要〕 简单介绍机器学习法，综述机器学习在肿瘤早期诊断与预后预测中的应用，重点阐述支持向量机、人工神经网络和深度学习 3 种机器学习方法在肿瘤诊断与预测中的良好表现。

〔关键词〕 机器学习；肿瘤；诊断；预测

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2016.11.003

Applications of the Machine Learning in Early Diagnosis and Prognostic Prediction of Tumors *SHI Wei, XUE Jun, PAN Cui－ran, REN Yuan－kai, NI Zheng－jie, ZHANG Yuan－peng, WANG Li, WU Hui－qun, JIANG Kui, DONG Jian－cheng, Department of Medical Informatics, Medical School of Nantong University, Nantong 226001, China*

〔**Abstract**〕 The paper briefly introduces the machine learning methods, summarizes the machine learning in early diagnosis and prognostic prediction of tumors, and focuses on good performance in diagnosis and prediction of tumors of the 3 machine learning methods including Support Vector Machine (SVM), Artificial Neural Nets (ANN) and deep learning.

〔**Keywords**〕 Machine learing; Tumor; Diagnosis; Prediction

1 引言

肿瘤（Tumor）是机体在各种致癌因素作用下，局部组织的某一个细胞在基因水平上失去对其生长的正常调控，导致其克隆性异常增生而形成的新生物（Newgrowth）。肿瘤组织无论在细胞形态还是组织结构上，都与其发源的正常组织有不同程度的差异，这种差异称为肿瘤细胞的异型性。良性肿瘤细胞的异型性小，一般与其来源的正常细胞相似，但恶性肿瘤具有高度的异形性。为了在肿瘤引起的症状发生之前就能发现肿瘤的（良恶性）类型，研究人员采用了多种生物医学（技术）方法，如血芯片检测、纳米检测、TCM 等，期待实现肿瘤的早期诊断以及治疗预后预测[1]。但是由于这些方法早期筛查的敏感度低，难以区分出良性与恶性肿瘤。尽管基因签名（Gene Signature）可以显著提高对肿瘤患者的预后预测，但这种方法在临床上的应用并不被看好[2]。近年来，随着计算机能力的提升和相关医

〔修回日期〕 2016－10－17

〔作者简介〕 施维，硕士研究生；通讯作者：董建成。

〔基金项目〕 国家自然科学基金（项目编号：81271668）；江苏省高校自然科学研究项目（项目编号：14KJB310014）；南通市市级科技计划拟资助项目（项目编号：MS12015112）；南通大学医学院教学研究课题（项目编号：Y2014－03）。

疗大数据的发展，越来越多的医生开始使用机器学习帮助诊断，研究表明机器学习法在辅助诊断方向的应用已使得肿瘤诊断与预测的准确率在过去的几年中提高了 15% ~20%[3]。本文对机器学习法在肿瘤早期诊断与预后预测中的应用进行综述。

2 机器学习法概述

2.1 机器学习的主要方法

机器学习是人工智能领域中最能体现智能的一个分支，目标是赋予机器一种新的能力。大多机器学习模型都要经过两个过程：一是从已有的数据集中获得未知的依赖关系；二是用学习到的依赖关系对新数据集进行分析处理。在机器学习领域，主要有 3 类不同的学习方法：有监督学习、无监督学习和半监督学习。有监督学习过程通过学习已标注的数据集的特征和结果建立模型，利用训练完的模型对未知数据进行分类、回归等工作，常见的如决策树（Decision Tree，DT）、支持向量机（Support Vector Machine，SVM）和人工神经网络（Artificial Neural Nets，ANN）等。目前，在肿瘤诊断和预测研究中大量使用的也就是以上 3 种有监督学习方法。无监督学习主要包括降维和聚类，通过对无标注样本信息进行学习，找到高维输入数据的低维结构或将样本划分为不同的簇（Cluster）。在实际的机器学习应用中，如基因序列比对，很容易找到海量无标注数据，由于这些数据的人工标注需要消耗大量资源，因此产生了未标注数据远多于标注数据的情况。半监督学习将未标注数据和已标注数据一起训练学习，解决了有监督学习模型泛化能力不强和无监督学习不精确的问题[4]。

2.2 数据集划分

在实际运用机器学习法时，还有一个关键问题就是数据的预处理，数据的噪声、缺失、异常和重复等现象都会影响机器学习的效果。此外，机器学习方法在数据特征维度低时工作效果更佳[5]。同时，为了使机器学习方法构建的分类模型能够获得可靠的结果，训练数据集和测试数据集的数据量应该足够大且相互独立。已经被广泛使用的划分标注数据集的方法有 Holdout 方法、交叉验证法等。如 5 折交叉验证，将标注数据集分为 5 份，其中将 4 份做训练，1 份做测试，经过 5 次训练后，标注数据集中每个样本都经过了训练和测试步骤，精度为所有不同验证周期的平均值。

2.3 使用科学的评价指标

当运用机器学习方法构建了分类模型后，使用科学的评价指标对模型进行评估也是重要环节。准确率（Accuracy）、灵敏度（Sensitivity）、特异度（Specificity）以及曲线下面积（Area Under the Curve，AUC）通常被用来对模型的分类效果进行评价，一个分类模型的 AUC 值通常在 0.5 ~1 之间，越大的 AUC 值代表分类模型具备越好的性能[5]。

3 机器学习在肿瘤诊断与预后预测中的应用

3.1 概述

如今高危生活习惯、暴露在环境中的致癌物、年龄、体重、家族史等因素在肿瘤诊断与预测中也起到了重要作用，然而目前临床医生还不能充分利用上述指标参数为肿瘤诊断与预测提供决策支持。过去临床医生用于肿瘤诊断的依据包括组织（病理）学数据、临床数据以及以人群为基础的统计数据[6]，随着分子生物学的快速发展，基因组、蛋白组等分子生物信息的获取使得分子标志物、细胞参数以及基因的表达成为非常有影响力的诊断指标[7]。研究者使用机器学习可以发现高通量测序技术产生的大量基因数据中的关系和特征，临床医生使用机器学习也可以分析复杂数据，从而对肿瘤诊断与预测做出科学的判断。

3.2 支持向量机

3.2.1 工作步骤 支持向量机是肿瘤诊断与预测中广泛使用的一种机器学习方法，其目的在于寻找一个超平面，将训练的标注数据集中的数据分开，且超平面与类域边界的垂直距离最大[8]。支持向量机首先将数据的初始向量映射到高维空

间，然后用超平面将数据点分成两类，其工作步骤，见图 1。

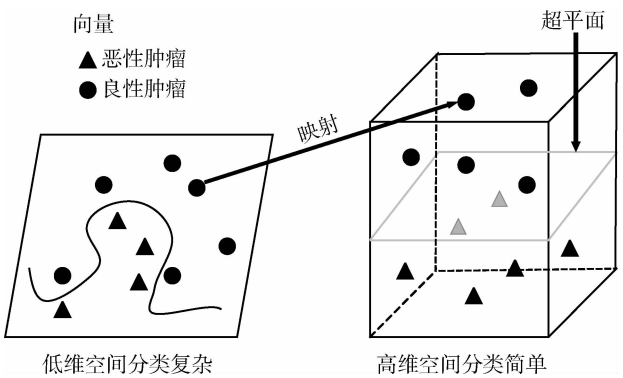

图 1 支持向量机的工作步骤

3.2.2 具体应用 表 1 展示了本节提到模型的详细参数。从表中可见研究者大多选择了以乳腺癌为主题的数据集来训练和测试模型；验证方法多以交叉验证为主；以 SVM 构建的分类模型的准确率最高为 99.4%。Huang 等利用网格法搜索最优核函数构建 SVM，选用包含 5 种病毒的数据集训练模型，最终用模型分析乳腺癌和纤维腺瘤数据，结果显示 HSV－1 和 HHV－8 的特征组合使模型得到最高的分类准确率[9]。Stoean 等不仅仅满足于模型的高准确率，通过结合 SVM 和进化算法（Evolutionary Algorithms）以及使用协同进化框架（Cooperative Coevolution）设计模型，试图获得模型进行决策的逻辑过程中的关键属性和参数值，以使临床医生可以将注意力放在那些关键的指示属性以及阈值范围上，结果显示，“团块厚度”、“细胞形状均匀性”、“边缘粘附”、“裸核”和“有丝分裂”这 5 个特征的组合对乳腺癌的分类支持度最高[10]。Kim 等利用 SVM 建立了一个乳腺癌 5 年内复发预测模型，通过与人工神经网络以及 Cox 比例风险回归模型进行对比发现，该乳腺癌复发预测模型的准确率最高[11]。Chen 等利用基于粗糙集的特征选取方法，采用分层采样的验证方法构建模型，也获得了很高的准确率[12]。

表 1 支持向量机方法在肿瘤诊断与预后预测中的应用研究

作者	肿瘤类型	病例数	数据种类	准确率	验证方法
Chen et al	乳腺癌	683	细胞核特征数据	0.994(AUC)	分层采样
Stoean et al	乳腺癌	699	临床数据、细胞核特征数据	—	随机采样
Huang et al	乳腺癌	80	DNA 病毒数据	86%	5 折交叉验证
Kim et al	乳腺癌	679	临床数据	0.85(AUC)	随机采样

3.3 人工神经网络

3.3.1 网络模型 人工神经网络从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。训练集数据进入网络后，每个维度的属性参数经过加权、求和等数学过程计算后，得到输出结果。反向传播（Back Propagation，BP）算法提出后，其强大的非线性映射和泛化能力解决了许多非线性问题，BP 神经网络在临床辅助决策、影像学处理和波形分析等方向都获得了较好的应用效果[13]。BP 神经网络的计算过程由正向传递和反向传播两部分构成。输入信息经隐藏层处理后到达输出层，若输出结果与预期值有差别，则函数会计算误差并反向传播到隐藏层中的每个神经元，修改权值等参数。人工神经网络模型，见图 2。

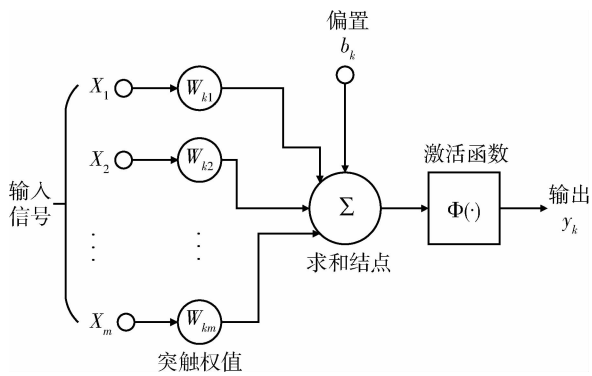

图 2 人工神经网络模型

3.3.2 具体应用 表 2 展示了本节提到模型的详

细参数。从表中可见研究者已开始选用跨平台数据集训练人工神经网络模型。Turgay 等运用 ANN 构建的预测模型来区分恶性和良性乳腺肿瘤[14]。输入层数据的维度为 36，包含由放射科医师参与决定的乳房影像属性参数；隐藏层设置 1 000 个节点。为了防止过拟合的情况，作者还采用了提前停止（Early Stopping，ES）算法。该模型与其他 ANN 构建模型不同之处在于其使用了大量有肿瘤标记的乳房影像数据，但由于缺少对人口学数据的预处理，作者未能将筛查和诊断数据集分开输入预测模型。Chen 等采用 ANN 对非小细胞型肺癌术后患者的存活情况进行预测[15]。作者在数据预处理的特征选取阶段选择了与存活情况最相关的 LCK 和 ERBB2 两个基因，以及 4 种临床指标对 ANN 进行训练，预测结果运用 Kaplan – Meier 生存函数进行评估。该模型的优势在于整合了多类别数据并进行了交叉预测，但是模型只能对非小细胞型肺癌的患者数据进行预测，因此可拓展性较弱。Alkim 等建立了强化学习的参数自适应的学习矢量化（Learning Vector Quantization）人工神经网络模型[16]。该模型不仅大大降低了参数调整的时间，而且因为其自适应的结构，还能对不同类型的疾病预测进行快速调整。同时，作者还提到了模型强化机制：在权值向量和输入向量之间加入偏置值，根据权值向量在每轮学习中的竞争次数将偏置值加入距离计算中。试验选用两种不同疾病的数据集进行预测，都获得了较高的准确率。

表 2 人工神经网络方法在肿瘤诊断与预后预测中的应用研究

作者	肿瘤类型	病例数	数据种类	准确率	验证方法
Turgay et al.	乳腺癌	62 219	乳房影像数据、人口学数据以及临床和细胞核数据	0.965（AUC）	10 折交叉验证
Chen et al.	肺癌	440	临床数据、基因数据	83.5%	跨数据集验证
Alkim et al.	乳腺癌和甲状腺疾病	914	细胞数据、临床数据	99.5%	—

3.4 深度学习

3.4.1 *内涵* 其概念来源于人工神经网络的研究，仍旧采用神经网络的分层结构，不同的是深度学习的“逐层初始化”训练机制能够训练更多的隐藏层[17]。肿瘤细胞的基因表达水平与正常细胞有差异，因此在近些年的肿瘤分类研究中，多采用基因芯片技术获取成千上万个基因的表达值。深度学习为解决这类问题提供了技术支持，被广泛应用的模型有卷积神经网络模型[18]、深度信念网络模型[19]以及堆栈自编码网络模型[20]等。因为肿瘤与细胞内的多种基因突变及调控异常有关，因此这些异常会在基因表达数据上有所反映。但是由于基因表达数据的维度高、样本量低、全局特征缺乏等问题，不经过预处理的数据会限制深度学习中卷积和池化等技术的分析应用。

3.4.2 *具体应用* 表 3 展示了本节提到模型的详细参数。从表中可见由深度学习方法构建的模型所获得的平均准确率要略低于 SVM 和 ANN。Fakoor 等使用深度神经网络模型之前，在特征选取部分运用主成分分析方法将高维的基因数据降维，然后使用无监督的稀疏自编码神经网络对数据集进行特征选取[21]。在分类器学习阶段，模型对 13 个肿瘤基因数据集进行训练测试，结果发现分类效果比未使用无监督方法的基准分类器好。Hua 等考虑到深度信念网络的表示层和隐藏层之间的权重是无向的，因此深度信念网络可以在收敛效率和分类准确性上有更好的表现[22]。首先用卷积神经网络对肺结节影像进行特征提取，然后用限制玻尔兹曼机预训练深度信念网络，最后对肺结节进行分类，解决了长期存在的恶性肿瘤或没有实际计算形态和结构特征的良性肺结节性质的分类的基本特征提取问题。Tomczak 等运用分类受限玻尔兹曼机（classRBM）对乳腺癌患者术后 10 年内的复发情况进行预测[23]。从实验结果中还发现，“肿瘤分期超过 50 mm”在乳腺癌复发的影响力较弱。

表 3 深度学习方法在肿瘤诊断与预测中的应用研究

作者	肿瘤类型	病例数	数据种类	准确率（%）	验证方法
Fakoor et al.	结肠癌、腺瘤等	2 057	基因数据	97.5	10 折交叉验证
Hua et al.	肺癌	1 010	肺部影像数据	82.2	留一交叉验证
Tomczak	乳腺癌	949	临床数据	73.8	—

4 结语

分子生物学的快速发展，特别是高通量测序等关键技术的突破，使得肿瘤分类研究从单一维度（形态）向多维度（分子、基因等）进化。本文提到的 3 种机器学习方法（支持向量机、人工神经网络和深度学习）在肿瘤诊断和预测中都获得了较好的表现，SVM 和 ANN 在肿瘤诊断辅助中已经成为研究者构建模型不可或缺的一部分，而深度学习在医学影像分析和跨平台数据处理中崭露头角。以上研究大多不仅仅只为提高模型的准确率，而且考虑了生物医学的意义。从本文提及的研究分析可以发现，将不同的特征选择与分类模型作用于多维整合的异构数据集诊断和肿瘤预测是当下研究者们通用的研究方式。

参考文献

1 Hanahan D, Weinberg R. Hallmarks of Cancer: the next generation [J]. Cell, 2011, 144 (5): 646 - 674.

2 Koscielny S. Why Most Gene Expression Signatures of Tumors Have not Been Useful in the Clinic [J]. Science Translational Medicine, 2010, 14 (2): 305 - 312.

3 Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis [J]. Cancer Informatics, 2007, 2 (1): 59 - 77.

4 刘建伟, 刘媛, 罗雄麟. 半监督学习方法 [J]. 计算机学报, 2015, 38 (8): 1592 - 1617.

5 Sumathi S, Sivanandam SN. Introduction to Data Mining and Its Applications [J]. Decision Support Systems, 2006, 26 (25): 236 - 238.

6 Sun Y, Goodison S, Li J, et al. Improved Breast Cancer Prognosis Through the Combination of Clinical and Genetic Markers [J]. Bioinformatics, 2007, 23 (1): 30 - 37.

7 Ren X, Wang Y, Chen L, et al. ellipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions [J]. Nucleic Acids Research, 2013, 41 (4): 242 - 248.

8 Suykens JAK, Vandewalle J. Least Squares Support Vector Machine Classifiers [J]. Neural Processing Letters, 1999, 9 (3): 293 - 300.

9 Huang CL, Liao HC, Chen MC. Prediction Model Building and Feature Selection with Support Vector Machines in Breast Cancer Diagnosis [J]. Expert Systems with Applications, 2008, 34 (1): 578 - 587.

10 Stoean R, Stoean C. Modeling Medical Decision Making by Support Vector Machines, Explaining by Rules of Evolutionary Algorithms with Feature Selection [J]. Expert System with Applications, 2013, 40 (7): 2677 - 2686.

11 Kim W, Kim KS, Lee JE, et al. Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine [J]. Journal of Breast Cancer, 2012, 15 (2): 230 - 238.

12 Chen HL, Yang B, Liu J, et al. A Support Vector Machine Classifier with Rough set - Based Feature Selection for Breast Cancer Diagnosis [J]. Expert Systems with Applications, 2011, 38 (7): 9014 - 9022.

13 倪然. 人工神经网络联合肿瘤标志对肺癌和大肠癌的预警 [D]. 郑州: 郑州大学, 2009.

14 Turgay AMS, Alagoz O, Chhatwal J, et al. Breast Cancer Risk Estimation with Artificial Nneural Networks Revisited [J]. Cancer, 2010, 116 (14): 3310 - 3321.

15 Chen YC, Ke WC, Chiu HW. Risk Classification of Cancer Survival Using ANN with Gene Expression Data from Multiple Laboratories [J]. Computers in Biology & Medicine, 2014, 48 (9): 1 - 7.

16 AlkiM, E, GüRbüz, E, Kilic L, E. A Fast and Adaptive Automated Disease Diagnosis Method with an Innovative Neural Network Model [J]. Neural Networks the Official Journal of the International Neural Network Society, 2012, 33 (33): 88 - 96.

（下转第 22 页）

目前信息化已经成为卫生管理与服务各项业务工作的重要支撑[4]。根据服务性质，医学情报所在卫生信息化的过程中主要发挥舆情监测作用，用户可根据个性化定制获得最新资讯。（1）卫生行业舆情监测预警。此模块基于互联网信息采集技术和数据挖掘技术，实时动态监测新闻门户、论坛、博客、微博、微信公众号、APP 新闻客户端等相关互联网站点，实现对海量信息的全方位实时扫描和监测，及时发现如传染病疫情、医疗纠纷、疫苗安全事件等网络突发性事件，将敏感舆情通过 E－mail、短信、电话等方式为管理部门提供预警，对热点信息进行持续跟踪监测，按需求及时完成并推送舆情简报、事件专题报告等，实现对互联网舆情信息的全面掌控与及时预警，为卫生管理部门提供决策支持。（2）医改及基本药物制度舆情监测。医药卫生体制改革和基本药物制度改革是我国卫生改革发展中的重要问题，是卫生政策研究的中心工作。平台开发特定舆情监测模块，即时监测各地区公立医院改革及新农合等医改进展及动态药品价格，实时捕获不同省市级别相关政策法规动向、医院信息化建设程度等，较为系统全面地对海量卫生改革信息进行监测分析。

5 结语

综合性医学信息支撑平台的搭建，运用网络信息平台来突破时间和空间的局限，实现医学情报机构基础工作的全程信息化管理、国内外医学文献资源一站式服务、情报研究及舆情监测全方位服务，充分促进信息、用户与平台资源的共建共享，进而提供更有价值的医药卫生信息服务，同时提升服务质量和效益[5]。平台的搭建不仅促进了医学领域的发展，同时也是医学情报机构实现可持续发展的必经之路。

参考文献

1 陈锐，冯占英，李焱，等. 大数据对生物医学信息服务各环节的影响研究［J］. 图书情报工作，2015，59（9）：68－72.

2 王建文. 基于图书情报系统的知识服务能力优化策略［J］. 科技创新导报，2015，（15）：190.

3 赵莉，王小飞，何福. 基于知识管理的科研情报综合信息系统建设［J］. 情报检索，2013，（6）：82－85.

4 万美. 卫生信息化视角下的医学信息资源建设［J］. 医学信息学杂志，2014，35（4）：77－79.

5 刘海虎. 关于网络环境下提高医学信息服务质量和效益的思考［J］. 健康导报：医学版，2015，20（12）：275.

（上接第 14 页）

17 Rumelhart DE, Hinton GE, Williams R J. Learning Representation by Backpropagating Errors［J］. Nature, 1986, 323（6088）：533－536.

18 Lécun Y, Bottou L, Bengio Y, et al. Gradient－Based Learning Applied to Document Recognition［J］. Proceedings of the IEEE, 1998, 86（11）：2278－2324.

19 Bengio Y. Learning Deep Architectures for AI［J］. Foundations and Trends in Machine Learning, 2009, 2（1）：1－127.

20 Vincent P, Larochelle H, Lajoie I, et al. Stacked Denoising Autoencoders：learning useful representations in a deep network with a local denoising criterion［J］. Journal of Machine Learning Research, 2010, 11（6）：3371－3408.

21 Fakoor R, Ladhak F, Nazi A, et al. Using Deep Learning to Enhance Cancer Diagnosis and Classification［C］. Atlanta：The International Conference on Machine Learning, 2013.

22 Hua KL, Hsu CH, Hidayati SC, et al. Computer－Aided Classification of Lung Nodules on Computed Tomography Images Via Deep learning Technique［J］. Oncotargets & Therapy, 2015,（8）：2015－2022.

23 Tomczak JM. Prediction of Breast Cancer Recurrence Using Classification Restricted Boltzmann Machine with Dropping［EB/OL］.［2015－08－28］. https：//arxiv. org/abs/1308. 6324