Advances in Applied Mathematics 应用数学进展, 2022, 11(6), 3496-3501
Published Online June 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.116372

Hans汉斯

机器学习在医疗数据发展中的应用思考

杨兴俊[1]，杨兴华[2]

[1]上海工程技术大学，上海
[2]云南财经大学，云南 昆明

收稿日期：2022年5月15日；录用日期：2022年6月3日；发布日期：2022年6月17日

摘　要

随着互联网技术的发展，大数据依然成为当今医疗行业的战略资源。本文基于机器学习算法K近邻、支持向量机、Catboost对公开数据威斯康星州乳腺癌诊断数据进行建模实验，从实践角度深入了解机器学习算法在健康医疗数据中的应用，为当前的医学瓶颈提供新的思路。

关键词

机器学习，SVM，Catboost，K近邻

Reflections on the Application of Machine Learning in the Development of Medical Data

Xingjun Yang[1], Xinghua Yang[2]

[1]Shanghai University of Engineering Science, Shanghai
[2]Yunnan University of Finance and Economics, Kunming Yunnan

Received: May 15th, 2022; accepted: Jun. 3rd, 2022; published: Jun. 17th, 2022

Abstract

With the development of Internet technology, big data is still a strategic resource for today's medical industry. Based on the machine learning algorithm K-nearest neighbor, support vector machine, and Catboost, this paper conducts modeling experiments on the public data of Wisconsin breast cancer diagnosis data, and deeply understands the application of machine learning algorithms in health and medical data from a practical point of view, and provides new insights for the current medical bottleneck.

文章引用: 杨兴俊, 杨兴华. 机器学习在医疗数据发展中的应用思考[J]. 应用数学进展, 2022, 11(6): 3496-3501.
DOI: 10.12677/aam.2022.116372

Keywords

Machine Learning, SVM, Catboost, K-Nearest Neighbor

1. 引言

随着计算机技术的发展，大数据的获取、储存、分析成为可能。在医疗行业大数据的发展为医生提供了决策的依据，特别在癌症的治疗方面取得巨大成功。当然这主要得益于大数据的获取更便捷以及后期的数据挖掘技术的发展使其成为医疗的决策依据。医院通过对癌症病人的病灶分析，建立自己的医疗数据库，其储存的数据包括病灶的大小，体积、癌症诊断结果，即癌症是良性还是恶性等相关特征。随着数据库的积累，从数据中挖掘宝贵的信息便是当下机器学习和深度学习的主要任务。计算机从海量大数据中学习相应的模型，在可接受的泛化能力下，当医院有新的病人时可以通过计算机算法辅助诊断，减少病人的就诊效率以及医生的工作量。

2. 研究现状

互联网技术的快速发展为当今信息化的社会带来许多宝贵的资源，不仅给我们带来了经济效益，同时也带来了新时代的数字资源。大数据资源依然成为当今能和自然资源、人力资源相抗衡的人类不可或缺的战略资源。大数据的发展使其与医疗行业的信息技术紧密联系在一起。每年不同的医疗机构产生的数据量达到 TB 级别甚至达到 PB 级别。在如此大量的数字资源背景下，机器学习算法、深度学习算法的普及为智慧医疗的发展奠定坚实的基础，使其患者在求医的过程中省去许多繁琐不必要的程序，医疗机构也能在基础医疗信息电子化以及医疗信息共享化下提高医疗资源的使用效率，提高疾病诊断的准确性。通过机器学习算法对健康医疗大数据的挖掘和分析为当前的医学瓶颈提供新的思路，医疗大数据的发展将改变传统模式下的医疗模式。计算机辅助诊断为医生提供相应的决策依据，不仅尊重患者的价值观、个体性差异和需求、保持医疗服务的连续性和可及性，全面提高医疗质量。对此国内外科技巨头以及医疗机构、学者都在积极开展机器学习和病理诊断的联合研究。

2.1. 国外研究现状

国外学者和互联网科技公司从不同角度对机器学习在医疗数据中的应用进行研究。Lin C (2014)等利用 NoSQL 技术提出了病人主使的数据结构，可以克服不同医疗机构之间记录数据模式之间的差异[1]。Zhang (2019)等人利用支持向量机对纽约市诊所获取的电子病例中提取相应的医疗诊断数据进行癌症数据分类[2]。Weng (2017)等人将随机森林、逻辑回归、梯度增强机、神经网络等 4 种机器学习算法应用于英国家庭无心血管疾病患者的常规临床数据预测[3]。Ho (2012)等人通过神经网络模型对肝癌患者切除手术后 1 年、3 年、5 年无疾病生存史的 80%患者数据进行预测，另外 20%作为验证组，采用逻辑回归进行预测，结果显示神经网络模型的性能较高[4]。另外科技公司，比如 Predilytics 利用大数据和机器学习使得整个行业被数字化，发挥出大数据的价值。其次，谷歌公司通过大数据预测出美国甲型 H1N1 流感的传播情况，虽然预测的结果依然存在偏差，但这是大数据成功应用的一个典范。同时，2019 年全球人工

智能健康峰会上，IBM Watson 健康事业部副总裁 Alok Gupta 出席并介绍 Watson Health 在整个医疗领域六个关键领域锻造能力，通过将人工智能和大数据分析技术运用于医疗领域，深入洞察医学知识和医学数据，助力于解决健康领域的多重难题，这更进一步说明健康医疗数据作为 21 世纪的战略资源，其价值是巨大的。

2.2. 国内研究现状

我国已有很多人投入到医疗大数据的研究中。高汉松(2013)等构建了一个医疗云数据挖掘平台，融合了云计算和 Hadoop，主要展示不同层次的功能，例如基础层、平台层等等[5]。许德权(2013)等着重阐述了医疗个性化服务与大数据间的关系及应用[6]。我国在现阶段的着重发展方向以及发展领域就是基于医疗大数据的医疗信息化。高校也在积极探索寻求合作，中南大学与移动率先提出了与以往不同的“移动医疗”概念，并将其开发成包含具体应用场景的原型系统，利用物联网技术打破时间空间的局限，实现了医患的远程互动与对接。除此以外，国内的很多中小型企业也投身于开发数据采集挖掘等方面的应用。我国现在较为成功的大型医疗数据项目是位于上海市的医院临床信息共享项目，实现了医院间临床项目信息的实时共享。

3. 机器学习分类算法

基于以上国内外学者将不同机器学习算法应用于临床医疗数据的研究，本文采用 K 近邻、Catboost、支持向量机三种算法对公开数据腺癌临床诊断数据进行预测。

3.1. K 近邻算法

K 近邻(KNN)是通过测量不同特征值之间的距离进行分类。它的基本思路是：如果一个样本在特征空间中的 k 个最相似的样本中的大多数属于某一个类别，则该样本也属于这个类别，其中 k 通常是不大于 20 的整数。KNN 算法中，所选择的邻居都是已经正确分类的对象。该方法在分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。在 KNN 中，通过计算对象间距离来作为各个对象之间的非相似性指标，避免了对象之间的匹配问题，这里的距离一般使用欧氏距离：

$$d(x,y)=\sqrt{\sum_{k=1}^{n}(X_k-Y_k)^2}$$

或者使用曼哈顿距离：

$$d(x,y)=\sqrt{\sum_{k=1}^{n}|X_k-Y_k|}$$

具体的算法步骤为：

1) 计算测试数据与各个训练数据之间的距离；
2) 按照距离的递增关系进行排序；
3) 选取距离最小的 k 个点；
4) 确定前 k 个点所在类别的出现频率；
5) 返回前 k 个点中出现频率最高的类别作为测试数据的预测分类。

3.2. 支持向量机

支持向量机通过“支持向量”学习超平面，用于将数据进行分类。主要包括线性支持向量机、近似线性支持向量机、非线性支持向量机模型，其中，线性支持向量机适用于数据线性可分的情况，当数据

线性不可分的情况，使用核函数将数据映射到高维向量空间中，在这样的高维空间中，可以学习到超平面将数据进行分类。

3.3. Catboost 算法

Catboost 算法是俄罗斯搜索引擎巨头 Yandex 于 2017 年开源的一款 GBDT 计算框架，因能高效处理数据中的类别变量而得名。Catboost 是一种用于实现梯度提升决策树的机器学习方法，主要采用排序提升的方法对抗训练集中的噪声点，从而避免梯度估计的偏差，进而解决预测偏移的问题。

假设前一轮训练得到的强分类器为 $F^{t-1}(x)$，当前损失函数为 $L\left(y,F^{t-1}(x)\right)$，则本轮迭代要拟合的弱分类器为 h^t：

$$h^t=\arg\min_{h\in H}L\left(y,F^{t-1}(x)+h(x)\right)$$

梯度表示为：

$$h^t=\arg\min_{h\in H}E\left(-g^t(x,y)-h(x)\right)^2$$

近似数据表示为：

$$h^t=\arg\min_{h\in H}E\left(-g^t(X_k,Y_k)-h(X_k)\right)^2$$

最终预测偏移的链式传递为：梯度的条件分布和测试数据的分布存在偏移；h^t 的数据近似估计于梯度表达式之间存在偏移；预测偏移会影响 F^t 的泛化能力。

4. 基于机器学习理论的大数据癌症病例诊断应用

4.1. 数据来源

本文采用机器学习数据库 UCI 中的威斯康星州乳腺癌诊断数据[7]，该数据集的样本量后续还会增加，随着沃尔伯格博士报告他的临床病例时，样本会定期送达数据库扩充数据集。该样本目前包括 569 个有效样本，每个样本包括以下几个特征：1) ID 编号；2) 诊断(M = 恶性，B = 良性)；3) 每个细胞核计算十个实值特征，分别为：半径(从中心到周长点的距离平均值)、纹理(灰度值的标准偏差)、周长、面积、平滑度(半径长度的局部变化)、紧凑性、凹度(轮廓凹陷部分的严重程度)、凹点(轮廓凹面部分的数量)、对称性、分形维数。

4.2. 构建模型

本文通过 R 语言 mlr3verse 包对 569 个有效样本进行分析，其中 80%用于训练，20%用于测试。本次实验以癌症的诊断结果作为因变量，其余变量作为自变量并对其进行标准化，分别建立 K 近邻算法、随机森林算法、支持向量机、Catboost 算法，然后对多种分类算法进行基准测试，采取 10 折交叉验证，对各种模型进行评估，挑选出泛化能力较好的模型调出最好的参数对实验样本进行预测。

4.3. 研究结果

本文基于 R 语言最新机器学习库 mlr3verse 包，通过随机森林算法、支持向量机、catboost 算法构建癌症诊断预测模型。其模型评估结果如表 1 所示：

在基准测试中挑选性能最优的模型对数据进行学习，由上表可知：本次实验最终采用 Catboost 算法构建癌症诊断模型，其预测结果如表 2 所示：Truth 代表的是原始癌症的诊断结果，Response 代表的是 Catboost 模型预测的诊断结果。

Table 1. Performance comparison of various algorithms
表 1. 各类算法性能比较

Learner	Accuracy	AUC	Precision	recall	iters
ksvm	0.9665	0.9953	0.9751	0.9722	10
kknn	0.9648	0.9854	0.9556	0.9916	10
Catboost	0.9666	0.9908	0.9587	0.9889	10

Table 2. Confusion matrix table
表 2. 混淆矩阵表

Response	Truth	
	B	M
B	71	1
M	1	40

如上表所示：经过 Catboost 预测乳腺癌诊断结果发现，有一列病例是良性，但模型错误的分类为恶性，一例癌症诊断结果是恶性，诊断成良性，但总体来说，经过 10 折交叉验证，模型的性能较佳。当然这不是必然的结果，计算机的诊断仅仅作为辅助，最终的决定权依然在于医疗机构。经过上面的实验，大数据作为 21 世纪的战略资源，合理应用医疗大数据，尊重患者个人隐私的同时提高了医疗诊断效率。

5. 应用思考

通过机器学习算法，当医疗机构在诊断癌症时，通过计算机辅助，不仅可以提高医疗机构的诊断效率，同时还能为医疗机构提供相应的决策依据。当然随着传感器的发展使得大数据的收集成为可能，一方面在互联网背景下，数据的增长规模不可同日而语，另一方面随着机器学习，深度学习的快速发展，人类将迈入人工智能的社会，机器人诊断，机器人开处方必将成为可能。现在设想一下如果人类通过计算机在成千上万的医疗数据中不断学习，甚至随着时间的发展数据还能呈指数式增长，计算机的学习能力也能在其中不断更新迭代，那么机器人医生的出现将不在是梦。然而，目前就我国而言，医疗数据的获取成为制约我国医疗数据发展的壁垒，未来我国医疗大数据如何应用于解决人类面临的问题，关键在于全国各大医疗机构的数据公开，倘若医疗数据可以全面公开，那这样的数据必将成为研究的重点，因此大数据的公开透明是医疗进步的关键一环。其次人工智能人才的培养，国家应该大量建立人工智能实验室，为人才的培养奠定良好的研究环境。

参考文献

[1] Lin, C., Huang, L., Chou, S., *et al*. (2014) Temporal Event Tracing on Big Healthcare Data Analytics. *Proceedings of* 2014 *IEEE International Congress on Big Data*, Anchorage, 27 June 2014-2 July 2014, 281-287. https://doi.org/10.1109/BigData.Congress.2014.48

[2] Zhang, X., Xiao, J. and Gu, F. (2019) Applying Support Vector Machine to Electronic Health Records for Cancer Classification. *Proceedings of* 2019 *Spring Simulation Conference* (*SpringSim*), Tucson, 29 April 2019-2 May 2019, 1-9. https://doi.org/10.23919/SpringSim.2019.8732906

[3] Weng, S.F., Reps, J., Kai, J., *et al*. (2017) Can Machine Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data. *PLOS ONE*, **12**, e0174944. https://doi.org/10.1371/journal.pone.0174944

[4] Ho, W.-H., Lee, K.-T., Chen, H.-Y., *et al*. (2012) Disease-Free Survival after Hepatic Resection in Hapatocellular Carcinoma Patients: A Prediction Approach Using Artificial Neural Network. *PLOS ONE*, **7**, e29179. https://doi.org/10.1371/journal.pone.0029179

[5] 高汉松，肖凌，许德玮，等. 基于云计算的医疗大数据挖掘平台[J]. 医学卫生信息管理杂志, 2013, 34(5): 7-12.

[6] 许德泉，杨慧清. 大数据在医疗个性化服务中的应用[J]. 中国卫生信息管理杂志, 2013, 10(4): 301-304.

[7] Wolberg, W., Street, W. and Mangasarian, O. (1995) Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository.