ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.16, No.33, November . 2020
E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

机器学习经典算法及其应用研究综述

徐洪学,孙万有,杜英魁,汪安祺

(沈阳大学 信息工程学院,辽宁 沈阳 110000)

摘要:机器学习是人工智能的一个重要子领域,是现阶段人工智能和数据分析领域的重点研究方向之一,我们有必要对机器学习有个全面而深刻的认识理解。根据训练样本及反馈方式的不同对机器学习算法分为监督学习、无监督学习及强化学习三类,介绍机器学习领域有代表性的若干经典算法及其应用,最后对机器学习算法的发展前景进行展望。

关键词:机器学习;监督学习;无监督学习;强化学习;深度学习

中图分类号:TP181 文献标识码:A

文章编号:1009-3044(2020)33-0017-03

开放科学(资源服务)标识码(OSID):

DOI:10.14004/j.cnki.ckt.2020.3359

Survey on the Classic Machine Learning Algorithms and Their Applications

XU Hong-xue, SUN Wan-you, DU Ying-kui, WANG An-qi

(School of Information Engineering, Shenyang University, Shenyang 110000, China)

Abstract: Machine Learning is an important subfield of Artificial Intelligence and one of the key research directions in the field of Artificial Intelligence and data analysis at this stage. It is necessary for us to have a comprehensive and profound understanding of Machine Learning. According to the different training samples and feedback methods, Machine Learning algorithms are divided into supervised learning, unsupervised learning and reinforcement learning. Some representative classical algorithms in the Machine Learning field and their applications are introduced. Finally, the development prospect of Machine Learning algorithms has prospected.

Key words: Machine Learning; Supervised Learning; Unsupervised Learning; Reinforcement Learning; Deep Learning

1 背景

机器学习领域的著名学者汤姆·米切尔(Tom Mitchell)将机器学习定义为:对于计算机程序有经验E、学习任务T和性能度量P,如果计算机程序针对任务T的性能P随着经验E不断增长,就称这个计算机程序从经验E学习[1]。这个定义比较简单抽象,随着对机器学习的研究越来越深入,我们会发现机器学习的内涵和外延也在不断变化。简言之,机器学习就是用计算机通过算法来学习数据中包含的内在规律和信息,从而获得新的经验和知识,以提高计算机的智能性,使计算机面对问题时能够做出与人类相似的决策[2]。

随着各行各业的发展,数据量增多,对数据处理和分析的效率有了更高的要求,一系列的机器学习算法应运而生。机器学习算法主要是指运用大量的统计学原理来求解最优化问题的步骤和过程。针对各式各样的模型需求,选用适当的机器学习算法可以更高效地解决一些实际问题[3]。

2 机器学习算法的分类

按照现在主流的分类方式,可以根据训练样本及反馈方式的不同,主要将机器学习算法分为监督学习、无监督学习和强化学习3种类型。其中监督学习是机器学习这三个分支中最大和最重要的分支。另外,作为监督学习与无监督学习相结合的半监督学习方法[4],暂不列在本文讨论范围之内。

2.1 监督学习算法(Supervised Algorithms)

在监督学习中,训练集中的样本都是有标签的,使用这些有标签样本进行调整建模,从而使模型产生一个推断功能,能够正确映射出新的未知数据,从而获得新的知识或技能[5]。

根据标签类型的不同,可以将监督学习分为分类问题和回归问题两种。前者预测的是样本类别(离散的),例如给定鸢尾花的花瓣长度、花瓣宽度、花萼长度等信息,然后判断其种类;后者预测的则是样本对应的实数输出(连续的),例如预测某一时期一个地区的降水量。常见的监督学习算法包括决策树、朴素贝叶斯及支持向量机等。

2.2 无监督学习算法(Unsupervised Algorithms)

无监督学习与监督学习相反,训练集的样本是完全没有标签的。无监督学习按照解决的问题不同,可以分为关联分析、聚类问题和维度约减三种。

关联分析是指通过不同样本同时出现的概率,发现样本之

收稿日期:2020-08-10
基金项目:国家自然科学基金资助项目(项目编号:61473195)
作者简介:徐洪学(1962—),男,辽宁大连人,教授,硕士生导师,博士;孙万有(1996—),男,辽宁大连人,硕士;杜英魁(1980—),男,吉林临江人,副教授,硕士生导师,博士;汪安祺(1995—),女,安徽铜陵人,硕士。

间的联系和关系。这被广泛地应用于购物篮分析中。例如,如果发现购买泡面的顾客有百分之八十的概率买啤酒,那么商家就会把啤酒和泡面放在临近的货架上。

聚类问题是指将数据集中的样本分成若干个簇,相同类型的样本被划分为一个簇。聚类问题与分类问题关键的区别就在于训练集样本没有标签,预先不知道类别。

维度约减是指保证数据集不丢失有意义的信息的同时减少数据的维度。利用特征选择方法和特征提取两种方法都可以取得这种效果,前者是指选择原始变量的子集,后者是指将数据由高维度转换到低维度。

无监督学习与人类的学习方式更为相似,被誉为是人工智能最有价值的地方[6]。常见的无监督学习算法包括稀疏自编码、主成分分析及K-means等。

2.3 强化学习算法(Reinforcement Algorithms)

强化学习是从动物行为研究和优化控制两个领域发展而来。强化学习和无监督学习一样都是使用未标记的训练集,其算法基本原理是:环境对Agent(软件智能体)的某个行为策略发出奖赏或惩罚的信号,Agent要使每个离散状态期望的奖赏都最大,从而根据信号来增加或减少以后产生这个行为策略的趋势[7]。

强化学习这一方法背后的数学原理与监督/非监督学习略有差异。监督/非监督学习更多地应用了统计学知识,而强化学习更多地应用了离散数学、随机过程等这些数学方法[8]。常见的强化学习算法包括Q-学习算法、瞬时差分法、自适应启发评价算法等。

3 机器学习经典算法及其应用

机器学习作为一个独立的研究方向已经经过了近四十年的发展,期间经过一代又一代研究人员的努力,诞生了众多经典的机器学习算法,但限于篇幅无法对所有算法一一整理总结,以下只列举了有代表性的一部分经典算法进行描述。

3.1 朴素贝叶斯(Naive Bayesian)

朴素贝叶斯算法基于统计学分类中的贝叶斯定理,将特征条件独立性假设作为前提,是一种常见的有监督学习分类算法。对于给定的一组数据集,朴素贝叶斯算法会求得输入/输出的联合概率分布,然后在统计数据的基础上,依据条件概率公式,计算当前特征的样本属于某个分类的概率,选择概率最大的分类。

在实际情况下,朴素贝叶斯算法的独立假设并不能成立,所以其性能略差于其他一些机器学习算法,但是由于其实现简单、计算复杂度低且对训练集数据量的要求不大,使其在文本分类、网络舆情分析等领域上有着十分广泛的应用[9]。另一方面,由于实际应用中存在各特征相互干涉、训练数据集缺失等的情况,于是又从中优化演变出其他贝叶斯算法[10],以增强其泛化能力。朴素贝叶斯算法的改进及应用如表1所示。

表1 朴素贝叶斯算法的改进及应用

算法名称	朴素贝叶斯分类器	半朴素贝叶斯分类器	贝叶斯网	EM算法
应用场景	特征间相互独立	考虑特征间的依赖性	合理特征间依赖关系	估计残缺的训练数据

3.2 K均值算法(K-Means)

K均值算法是一种常用的聚类算法。其核心思想是把数据集的对象划分为多个聚类,并使数据集中的数据点到其所属聚类的质心的距离平方和最小,考虑到算法应用的场景不同,此处描述的"距离"包括但不限于欧氏距离、曼哈顿距离等。

它的工作流程分为四步:1)在数据集n个对象中任意选取k个对象作为初始的聚类中心;2)计算数据集中其他对象到这些聚类中心的距离,分别将这些对象划分到与其距离最近的聚类中心;3)重新计算聚类均值得到其质心,然后将所得到的质心作为新的聚类中心;4)不断重复第二步和第三步直到标准测度函数最终收敛为止[11]。

K均值算法原理十分简单,需要调节的参数只有一个k,且具有出色的速度和良好的可扩展性,因而K均值算法作为经典的聚类算法,被普遍应用于需要解决此问题的各个领域之中[12]。

3.3 AdaBoost算法(Adaptive Boosting)

Boosting算法是一种可以用来减小监督学习中偏差的机器学习算法,其中AdaBoost算法是最优秀的Boosting算法之一,其核心思想是将分类精度比随机猜测略好的弱分类器提升为高分类精度的强分类器。AdaBoost算法可用于分类和回归,但目前的研究和应用大多集中于分类问题[13]。

为了创建一个强大的分类器,AdaBoost算法使用了多次迭代的方法。它会依据每次训练后的样本集分类是否正确和上次分类的准确率,来确定数据集中样本的权值,将修改过权值的新数据集送给下层分类器进行训练,然后将每次训练得到的分类器相融合,得到的结果是一个比弱分类器更加准确的分类器,并作为最终的决策分类器以实现算法[14]。

相较于其他机器学习算法,AdaBoost算法虽然对异常值和噪声数据比较敏感,但其具有能够显著改善子分类器预测精度、不需要先验知识、理论扎实、克服了过拟合问题等众多优点。这使AdaBoost算法及其演化算法凭借其优秀性能受到不同领域研究人员的关注,目前被广泛应用于机器视觉、计算机安全、计算生物学等诸多领域[15]。

3.4 支持向量机(SVM)

支持向量机是一种对数据进行二元分类的广义线性分类器,其决策边界是对学习样本求解得到的最大边距超平面。基本思想是:找到集合边缘上的若干数据(称为支持向量),用这些点找出一个平面(称为决策面),使得支持向量到该平面的距离最大。

支持向量机的学习策略就是间隔最大化,其算法就是求解凸二次规划的最优化算法,关键在于求得分类间隔最大值的目标解[16]。根据训练数据集是否线性可分的不同,支持向量机分为:线性可分支持向量机、线性支持向量机和非线性支持向量机。

在支持向量机算法提出之初就被成功地应用于手写数字的识别上,证明了其算法在理论上具有突出的优势。支持向量机模型与许多机器学习算法能够很好地联合应用,这使得支持向量机有着众多性能更佳的改进模型[17]。在近几年内的发展中,支持向量机在人脸识别、文字识别、图像处理等各方面都得到了广泛应用[18]。

4 总结

经过漫长的发展，经历了萌芽期、停滞期、复兴期、成型期直到现在的蓬勃发展期，最终机器学习算法的研究成果得到广泛应用实属不易。虽然目前有着众多机器学习算法，但没有一种算法能够适用所有问题，针对不同的应用场景，监督学习、无监督学习、强化学习都有各自合适的选择，各种类别的机器学习算法均有擅长的领域和难以克服的缺陷，尽管机器学习的经典算法较为简单，但其是机器学习发展的基础和核心，可将其进行改进和联合使用，以发挥优点弥补不足，促进机器学习能力的提升，希望本文可以给该领域的研究学者提供一些参考和启发。

参考文献：

[1] 赵彰.机器学习研究范式的哲学基础及其可解释性问题[D].上海:上海社会科学院,2018.

[2] 张润,王永滨.机器学习及其算法和发展研究[J].中国传媒大学学报(自然科学版),2016,23(2):10-18,24.

[3] 杨剑锋,乔佩蕊,李永梅,等.机器学习分类问题及算法研究综述[J].统计与决策,2019(6):36-40.

[4] 屠恩美,杨杰.半监督学习理论及其研究进展概述[J].上海交通大学学报,2018,52(10):1280-1291.

[5] 邓建国,张素兰,张继福,等.监督学习中的损失函数及应用研究[J].大数据,2020,6(1):60-80.

[6] 焦李成,杨淑媛,刘芳,等.神经网络七十年:回顾与展望[J].计算机学报,2016,39(8):1697-1716.

[7] 张耀中,胡小方,周跃,等.基于多层忆阻脉冲神经网络的强化学习及应用[J].自动化学报,2019,45(8):1536-1547.

[8] 马骋乾,谢伟,孙伟杰.强化学习研究综述[J].指挥控制与仿真,2018,40(6):68-72.

[9] 李旭然,丁晓红.机器学习的五大类别及其主要算法综述[J].软件导刊,2019,18(7):4-9.

[10] 渠云龙.基于不同场景的贝叶斯分类的改进研究与应用[D].长春:吉林大学,2019.

[11] 柏宇轩.Kmeans应用与特征选择[J].电子技术与软件工程,2018(1):186-187.

[12] 薛琳瑶.K-means算法在地质灾害系统中的应用研究[D].西安:西安工业大学,2018.

[13] Freund Y,Schapire R E.A decision-theoretic generalization of on-line learning and an application to boosting[J].Journal of Computer and System Sciences,1997,55(1):119-139.

[14] 潘志庚,刘荣飞,张明敏.基于模糊综合评价的疲劳驾驶检测算法研究[J].软件学报,2019,30(10):2954-2963.

[15] 贾宏云.基于AdaBoost模型的藏文文本分类研究与实现[D].拉萨:西藏大学,2019.

[16] 林香亮,袁瑞,孙玉秋,等.支持向量机的基本理论和研究进展[J].长江大学学报(自科版),2018,15(17): 6,48-53.

[17] 程建峰,乐俊刘丹.基于SVM算法的用户行为认证方法[J].计算机系统应用,2017,26(11):176-181.

[18] 姜建国,常子敬,吕志强,等.USB HID攻击检测技术研究[J].计算机学报,2019,42(5):1018-1030.

【通联编辑：谢媛媛】

（上接第3页）

[8] 谭宝成,杨成.激光雷达动态障碍物检测[J].西安工业大学学报,2015,35(3):205-209.

[9] 张穗华,骆云志,王钤,等.基于三维激光雷达的障碍物检测方法研究[J].机电产品开发与创新,2016,29(6):14-17.

[10] Asvadi A,Premebida C,Peixoto P,et al.3D Lidar-based static and moving obstacle detection in driving environments:an approach based on voxels and multi-region ground planes[J].Robotics and Autonomous Systems,2016,83:299-311.

[11] 王鑫,吴际,刘超,等.基于LSTM循环神经网络的故障时间序列预测[J].北京航空航天大学学报,2018,44(4):772-784.

[12] 蒋畅江,温登峰,唐贤伦,等.基于改进型轻门控循环单元的语音识别[J].计算机工程与设计,2019,40(11):3265-3268,3356.

[13] 段生月,王长坤,张柳艳.基于正则化GRU模型的洪水预测[J].计算机系统应用,2019,28(5):196-201.

[14] 刘建伟,崔立鹏,刘泽宇,等.正则化稀疏模型[J].计算机学报,2015,38(7):1307-1325.

[15] 张舜,郝泳涛.基于深度学习的障碍物检测研究[J].电脑知识与技术,2019,15(34):185-187,193.

【通联编辑：唐一东】