机器学习算法比较

郭 成

（武汉轻工大学 数学与计算机学院，湖北 武汉 430023）

摘 要：如今，机械学习在数据挖掘、图像处理、自然语言处理以及生物特征识别等领域的应用已十分广泛。在机器学习中有一种“无免费午餐（NFL）”的定理，它指出没有任何一个算法可适用于每个问题，尤其是与监督学习相关的。因此，应尝试多种不同的算法来解决问题，同时还要使用“测试集”对不同算法进行评估，并选出最优者。笔者基于机器学习的发展，对几种常见算法优劣进行了研究分析，并讨论了其发展前景。

关键词：机器学习；监督学习；算法比较

中图分类号：TP333.35 **文献标识码：**A **文章编号：**1003-9767（2019）05-049-02

Algorithm Comparison of Machine Learning

Guo Cheng

(School of Mathematic and Computer Science, Wuhan Polytechnic University, Wuhan Hubei 430023, China)

Abstract: Nowadays, machine learning has been widely used in data mining, image processing, natural language processing and biometric recognition. There is a theorem of "no free lunch (NFL)" in machine learning, which points out that no algorithm can be applied to every problem, especially in relation to supervised learning. Therefore, we should try many different algorithms to solve the problem. At the same time, we should use "test set" to evaluate different algorithms and select the best one. Based on the development of machine learning, the advantages and disadvantages of several common algorithms are studied and analyzed, and their development prospects are discussed.

Key words: machine learning; supervised learning; algorithm comparision

0 引言

机器学习算法是一种可从数据中学习、从经验中提升自己而不需要人类干预的算法。学习内容可能是一个从输入映射到输出的函数、无标记数据中的隐含结构或者是“基于实例的学习（instance-based learning）”，通过比较新的实例和存储在内存中的训练数据，给新的实例赋予一个类别标记。“基于实例的学习”不会在这些具体的实例上创造一层抽象。

1 机器学习分类

机器学习算法按人工干预程度划分，大致可分为三种类型：强化学习、无监督学习、监督学习。

强化学习是一种智能体通过与环境进行交互获得的奖赏指导行为，它以“试错”的方式进行学习，其目标是使智能体获得最大的奖赏[1]。无监督学习只有输入变量没有输出变量，这种学习问题使用无标记的训练数据来对数据中隐含的结构进行建模。监督学习即使用标记的训练数据来学习从输入变量（X）到输出变量（Y）的映射函数。

本文旨在探讨监督学习算法的优劣，因此对无监督学习与强化学习不再赘述，以下将重点分析监督学习几种算法的介绍与比较[2]。

监督学习问题可以分为两类。第一，分类问题，在监督学习中，输入变量 X 可以是连续的，也可以是离散的；当输出变量 Y 为有限个离散值时，预测问题便成为分类问题。第二，回归问题，其主要作用是预测自变量和因变量二者的关系，自变量和因变量的关系为：输出变量的值会随着输入变量值的变化而发生变化。回归模型正是表示从输入变量到输出变量之间映射的函数[3]。

2 K 近邻算法（K-NN）

K-NN 算法是最简单的分类算法，算法原理是通过计算待分类样本和训练样本之间的差异性，按照由小到大的排序对差异进行排序，再选出前面 K 个差异最小的类别，并统计

作者简介：郭成（1994—），男，湖北孝感人，硕士研究生在读。研究方向：图像处理与模式识别。

在 K 个类中出现次数最多的类，这一类别即为最相似的类，最终将待分类样本分到最相似的训练样本的类中 [4]，这一机制与投票（Vote）类似。使用 K-NN 算法时，精确度高，但对异常值不敏感，且无数据输入假定，局限性在于 K-NN 算法是基于实例的学习，使用算法时，使用的训练样本数据必须尽可能接近实际数据。K-NN 算法必须保存全部数据集，如果训练数据集很大，必须有足够的存储空间。因为必须对数据集中的每个数据计算距离，所以需要耗费很长时间。此外，K-NN 算法无法给出数据的基础结构信息，因此，平均实例样本和典型实例样本所具备的特征就无从得知。

3 决策树

K 近邻算法可完成很多分类任务，但最大的缺点是其数据形式难以理解，数据的内在含义无法给出，而这两点恰好是决策树的主要优势。决策树算法是从数据的特征出发，将特征作为基础，划分不同的类别 [5]。

决策树的主要优势在于计算输出的结果易理解，且复杂度低，如果中间值缺失，对结果影响也较小，决策树算法也可处理不相关特征数据。但是，当决策树的复杂度较大时，可能会造成过拟合问题。此时，可通过裁剪决策树的办法，降低决策树的复杂度，提高决策树的泛化能力。如果决策树的某一叶子结点只能增加很少信息，可将该节点删掉，将其并入到相邻的结点中，从而降低决策树的复杂度，消除过拟合问题。

4 朴素贝叶斯

朴素贝叶斯与决策树最大的不同在于前者是给出最大可能性结果的猜想和概率，后者是“ 武断 ”的给定唯一分类结果。

用朴素贝叶斯算法进行分类，主要是基于贝叶斯定理与特征条件独立假设。如果给定训练数据集，首先基于特征条件独立假设学习输入 / 输出的联合概率分布，在此模型基础上，针对给定的输入 x，利用贝叶斯定理计算后验概率最大的输出 y。

朴素贝叶斯法实际上学习了生成数据的机制，所以从属性上来看，应将其归于生成模型。条件独立假设的意思即在类确定的条件下，用于分类的特征都是条件成立的。正因为有了这一假设，朴素贝叶斯法得以简单化，但其不足在于，分类准确率不高，且分类性能不够好 [6]。

5 Logistic 回归

Logistic 回归是一种简单的分类算法，主要思想是利用已有数据，对分类边界线建立回归方程，以此进行分类。

Logistic 回归的目的是找到非线性函数 sigmoid 的最佳拟合参数，从而相对准确的预测分类结果。为了找出最佳的函数拟合参数，最常用的优化算法为梯度下降法。随机梯度下降法是一种在线学习算法，它不需要重新读取整个数据集进行批处理运算，而是在新数据到来时自动完成迭代，并更新拟合参数，也正是因此计算损耗较低。

总的来说，Logistic 回归算法具有计算代价低，易于理解和实现等优点。但是，Logistic 回归算法易出现欠拟合和分类精度不高等现象。

6 结　语

收敛性是评价一个优化算法的重要标准，也就是说在优化后，算法中的参数是否达到稳定值。在实际中，当参数值接近稳定时，仍然会出现小的周期性波动，究其原因，是数据集并非线性可分，这些点在每次迭代时会引发系数的剧烈改变，造成周期性的波动。显然人们希望算法能避免来回波动，从而收敛到某个值，并且收敛速度要足够快。目前机器学习发展势头迅猛，配合大规模数据标注的出现，深度神经网络也走出了此前的困境，在学术界和工业界都取得了很大进步，如 AlphaGo 大败围棋顶级选手，而在此背景中，监督学习发挥着不可或缺的作用。通过监督学习预训练模型，可很好地解决深度模型的过拟合问题，从而在实际应用中能得到模型更贴合预计的反馈。笔者相信，在今后，以监督学习为代表的机器学习将会大放异彩，发挥越来越重要的作用。

参考文献

[1] 李航 . 统计学习方法 [M]. 北京 : 清华大学出版社 ,2017:362.

[2]Michalski R S,Carbonell J G,Mitchell T M.Machine Learning[M].Berlin:Springer-Verlag,1994:243.

[3]Ruppert,David.The Elements of Statistical Learning:Data Mining,Inference,and Prediction[J].Journal of the American Statistical Association,2004,99(466):567-567.

[4]Liu B.Web Data Mining:Exploring Hyperlinks,Contents and Usage Data[M].Berlin:Springer-Verlag,2006:361.

[5]Collins M,Schapire R E,Singer Y.Logistic Regression,AdaBoost and Bergman Distances[J].Machine Learning,2002,48(1-3):253-285.

[6] 孙亮 , 黄倩 . 实用机器学习 [M]. 北京 : 人民邮电出版社 , 2017:364.