机器学习综述

孔欣然
（江苏省新海高级中学，江苏连云港，222002）

摘要：机器学习作为人工智能领域一个关键的研究内容，主要研究如何在积累经验的过程中不断提高计算机程序的性能。通过机器学习，计算机具备了智能手段，并在很多领域都得到了广泛应用。本文详细介绍了机器学习的重要概念，系统地阐述了机器学习的两种典型算法，介绍了算法的原理、流程及利弊，同时也介绍了机器学习在日常生活和生产中的几种典型应用。

关键词：机器学习；人工智能；KNN算法；K-means算法

DOI:10.16589/j.cnki.cn11-3571/tn.2019.24.034

0 引言

我们身处计算机网络时代，数据传播速度加快，信息量增大。在这种趋势的引导下，大数据成为了当今研究的热门主题。而正是由于机器拥有强大的运算大数据能力，使得机器学习成为分析大数据、挖掘潜在规律的主要方式。目前，机器学习已在无人驾驶、智能机器人、专家系统等各个领域产生了广泛而深远的应用[1-5]。

1 机器学习中的重要概念

■ 1.1 训练集与测试集

训练集[6]：在已知数据中选取的用来模拟曲线的数据。

测试集[6]：在已知数据中用来测试模拟曲线精确度的数据。

为了检验模拟曲线的精确度，在实际操作过程中，我们经常按照一定的比例（如 8:2,7:3）把获得的数据划分为训练集和测试集。

这样做的原理是，当我们拟合模型时需要完全依靠训练集里的数据完成拟合。尽管对训练集数据来说，该模型是比较精确无误的，但我们并不能保证当它应用在其他数据时，还保持着较高贴合度。所以需要用测试集来验证模型的精确度。

显然，将一部分数据固定分为训练集，另一部分为测试集，仅验证一次也有可能会出现模型精确度有偏差的情况。因此，为了减少数据划分给模型带来的影响，在实际应用中，我们通常采用 s 交叉验证法[6]。

s 交叉验证法：我们先将数据分为 s 等份，留存第一份测试数据，其余 s-1 份作为训练数据进行训练和评估。第一次，我们用第 1 份做测试，第二次用第 2 份，第 n 次用第 n 份（$1 \leqslant n \leqslant S$）做测试。就这样进行 s 次，从中挑出拟合度最好，精确度最高的模型作为预测模型（注：s 的选择要满足训练集样本数量占总体数量一半以上）。

■ 1.2 欠拟合与过拟合

在机器学习得出训练模型时，我们经常会遇到两种结果，一种是欠拟合[6]，一种是过拟合[6]。

欠拟合：欠拟合是指在训练数据和预测结果时，模型精确度均不高的情况。如图 1，该曲线未经过大部分数据且偏离较大，与数据匹配度较低，这直接导致在测试时表现不佳。

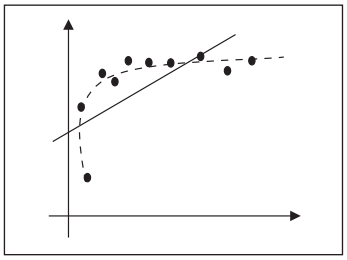

图 1　欠拟合示意图

产生原因：模型未能准确地学习到数据的主要特征。

解决策略：我们可以尝试对算法进行适当的调整，如使算法复杂化（例如在线性模型中添加二次项、三次项等）来解决欠拟合问题。

过拟合：顾名思义，指的是模型出现拟合过度的情况。过拟合表现为模型在训练数据中表现良好，在预测时却表现较差，如图 2 所示。

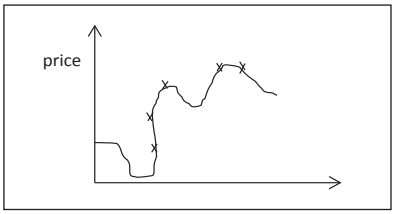

图 2　过拟合示意图

产生原因：这是源于该模型过度学习训练集中数据的细节，而这种随机波动并不适用于新数据，即模型缺乏普适性，所以模型在预测时表现较差。

解决策略：可以通过扩大训练集数据容量的手段，降低噪声对模型的干扰，以达到使模型学习到更多数据关键特征的目的。

■ 1.3 分类与回归

监督学习在生活中应用广泛。总的来说，监督学习可分为分类和回归[6]。

回归问题（连续变量预测）：输入一个新数据，训练模型就会预测其输出值（实数）。回归任务属于定量输出。

分类问题（离散变量预测）：输入一个数据，训练模型将推断出它的类别。分类任务属于定性输出。

比如，天气预报预测明天的温度，就是一个典型的回归任务；而预测明天是否下雨，是阴还是晴，则是一个典型的分类任务。

■ 1.4 监督学习与非监督学习

机器学习的核心是机器使用算法分析海量数据，通过学习数据，挖掘数据中存在的潜在联系，并训练出一个有效的模型，将其应用于决定或预测。目前，机器学习分为监督学习和非监督学习这两种基本类型[6]。

监督学习：在监督学习中，数据已被标记。换句话说，我们已经确定了分类标准。计算机使用训练模型来识别每种标记类型的新样本，确定该变量的输出值。此时，我们得到的结果可能是个连续值（即回归任务），也可能是一个离散值（即分类任务）。

非监督学习：与监督学习相反，在非监督学习中，数据没有被标记。换句话说，我们还不清楚数据的分类标准。通常情况下，当面对众多的数据却缺乏了解时，我们可以使用非监督学习，利用机器先将这些陌生的数据挖掘出某些潜在的共性，并对它们聚类。因此，非监督学习中机器划分的类，是我们之前未知的。比方说，我们使用机器给大量未知特征的图片聚类，观察机器得出的结果我们才发现，在两个类簇中，一类图片都是人像，另一类则都是风景画。

2 机器学习中的典型算法

■ 2.1 KNN 算法（k 近邻法）

1968 年 Cove、Hart 提出 KNN 算法，目前 KNN 算法[1]已广泛应用于监督学习。

2.1.1 算法原理

该方法的原理是：我们选取一个 k 值，输入样本，系统筛选出该样本在已知数据中最相似的 k 个数据，这 k 个数据中从属于不同类，哪个类别的占比最多，那么该样本也属于这个类别。在 KNN 算法中，已知数据成为空间里的特征向量，对应空间中的一个点，通常我们用空间中两个实例点的距离来反映两者相似度。

2.1.2 三大基本要素

k 值的选择，度量距离和分类决策规则是 KNN 算法的基础

（1）K 值的选择：理论上来说，k 值的选取可能会直接影响输出结果。而且 K 值较小，模型会比较复杂，容易导致过拟合情况的出现；K 值较大，可以减少误差，但又会使模型的复杂度过低。所以，在实际应用中，我们常先取一个较小的值为 k，在用交叉验证法找到最优的 K 值。

(2) 度量距离：设两个 n 维变量 $a(X_{11},X_{12},\cdots,X_{1n})$ 与 $b(X_{21},X_{22},\cdots,X_{2n})$ 间的距 $d=\sqrt[p]{\sum_{k=1}^{n}|x_1-y_1|}$，其中 p 是一个变参数。

特别地，当 p=1 时，就是曼哈顿距离，公式为：

$$d=\sum_{k=1}^{n}\left|x_{1k}-x_{2k}\right|$$

当 p=2 时，就是欧氏距离，公式为：

$$d=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2}$$

当 $p\to\infty$ 时，它是所有坐标距离的最大值，也就是切比雪夫距离，公式是：

$$d=\max\left(\left|x_{1k}-x_{2k}\right|\right)$$

不同距离公式的选取也会对结果造成影响。如图 3 所示，我们知道当采用欧几里得距离（即 p=2）时，到原点距离为 1 的点组成的图形是一个半径为 1 的圆；若采用曼哈顿距离（p=1）时，组成的图形就是一个对角线长为 2 的菱形；而当采用切比雪夫距离（$p\to\infty$ 时），所得图形就为边长为二的正方形了。

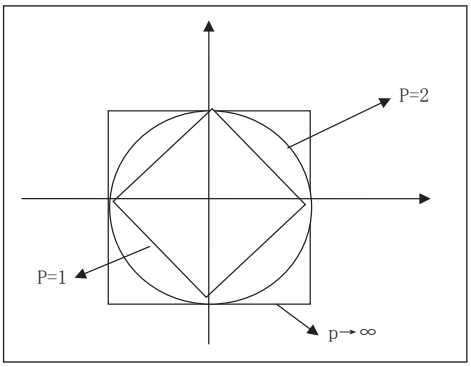

图 3　距离公式的影响示意图

2.1.3 算法流程

Step1. 选择适当的 K 值和度量距离；

Step2. 依次计算样本到各点距离；

Step3. 将距离从小到大排序；

Step4. 选出前 k 个距离最近的点；

Step5. 统计各点所属类别，该样本即属于占比最多的那一类。

2.1.4 KNN 算法优点

KNN 算法自身操作简略，便于理解，易于完成，尤其适合于多分类问题。

■ 2.2 K-means 算法（K 均值算法）

James MacQueen 在 1967 年提出 K-means 算法[2]，目前非监督学习中 K-means 算法已获得普遍的使用。

2.2.1 算法原理

在 K-means 算法中，字母 k 的含义是表示聚类的数量，而单词 means 表示每个类的数学期望。与上文提到的 KNN 算法一样，在 K-means 算法中，样本之间的距离就是样本相似度的直观表达。两个样本离得越近，样本具有的相似度就越高，换句话说，它们就越可能在同一类。在 K-means 算法中，通常采用欧氏距离公式计算距离。

2.2.2 算法流程

Step1. 随机选取 k 个样本，作为初始聚类中心；

Step2. 机器计算样本与选定中心的距离；

Step3. 比较样本与各中心距离，将其归于距离最近的中心所属聚类

Step4. 依次计算并分配各个样本，每个聚类中心和分配给它的样本就构成了一个簇；

Step5. 重新计算每个类别的聚类中心；

Step6. 重复上述过程，直到没有聚类中心再发生变化为止。

3 机器学习在实际生活中的应用

伴随科技的不断进步，机器学习已经成为一个正在蓬勃发展并拥有无限光明未来的热门学科。机器学习，这项能带来巨大生产力的革命性技术，对经济、社会等各个方面将产生革命性的变化。目前，机器学习已与生物、教育、交通等各个领域结合，碰撞出各种智慧的火花。

■ 3.1 推荐系统

将不同用户感兴趣的不同内容精准地进行推荐，这就是推荐系统的目的。目前，推荐系统已广泛地应用到了各大搜索引擎、购物网站以及各类 app。

以网易云音乐为例，它在一众音乐类 app 中脱颖而出，两年半就突破了一亿用户，很大程度上得益于它的推荐系统。网易云利用机器学习，通过分析用户的喜欢与收藏，试听与循环播放，调查用户喜好，从而通过机器推测使用者的基本信息，如性别，年龄等，进行用户画像，并寻找与该用户相似程度较高的用户。举个例子，如果两个人歌单里有一百首歌，其中有九十首歌是重复的，这时，网易云音乐就会将余下十首该用户未接触的歌曲推荐给他。对于不同类别的用户，网易云均能进行实时跟踪，进行每日相关推荐。网易云利用机器学习形成的推荐系统，提高了推荐效率，并成功达到极高的用户满意度。

淘宝的推荐系统也同样利用了机器学习。淘宝用户的足迹、收藏、购物车和购买记录，都是淘宝用来计算客户购买某种商品的可能的重要基础。进而筛选其中可能性高的进行推荐，并赠送优惠券。淘宝使用基于机器学习的智能化推荐系统，不仅缓解了商品屯积量过大的问题，节约了成本，避免造成资源浪费，同时，它也精确瞄准了每一个客户的喜好，从而达到利益最大化。

■ 3.2 模式识别

（1）图像识别

图像识别指的是机器对图像进行处理、分析、记忆，从而识别各种对象的技术。比如，机器识别图片上的动物是猫是狗。面对海量的图片，人类通过机器学习进行自动识别，这不仅节省了大量时间、劳动力和成本，同时，这也提高了工作效率和准确性。

随着科学技术的不断进步，计算机学习早已不再局限于固定的模板，而是能灵活地识别出多变的手写字体和人脸等。基于机器学习的图像识别技术已成功取得了从“哪一种”精确到了“哪一个”的重大突破。目前，图像识别已在生物医学、遥感技术、通讯领域和军事刑侦等多个领域有了广泛的应用。

（2）语音识别

语音识别指的是在进行了学习和理解后，机器把语音信息转成文字信息，或者理解用户的语音命令并执行操作，如 iphone 手机的 siri, 用户可以通过声控的方式，来打开手机上的应用程序，或者搜索附近电影院、宾馆等天气情况信息，也可以了解明天的天气情况，甚至可以与 siri 进行生动对话。iPhone 手机使用基于机器学习的语音合成和识别技术，一方面，给用户带来了新鲜体验，另一方面语音命令减少了文字输入的麻烦，使操作简单、便捷、迅速。

■ 3.3 垃圾邮件过滤系统

机器通过学习实现划分收到的邮件，识别垃圾邮件甚至是带有病毒的有害邮件，进行拦截和删除，保护电脑安全。

4 总结与展望

随着谷歌（Google）旗下 DeepMind 公司开发的围棋机器人 AlphaGo 一次次战胜人类顶尖棋手，作为人工智能最有前景的一个重要分支，机器学习近年来得到人们前所未有的重视。本文介绍了机器学习有关的基本概念，以及 KNN 算法和 kmeans 算法。KNN 算法、K-means 算法虽然有效，但同时也存在弊端，希望在不久的将来，机器学习能更大程度地改善我们的生活。

参考文献

* [1] 周昀锴 . 机器学习及其相关算法简介 [J]. 科技传播 .2019, 6 : 153-154.

（下转第 38 页）

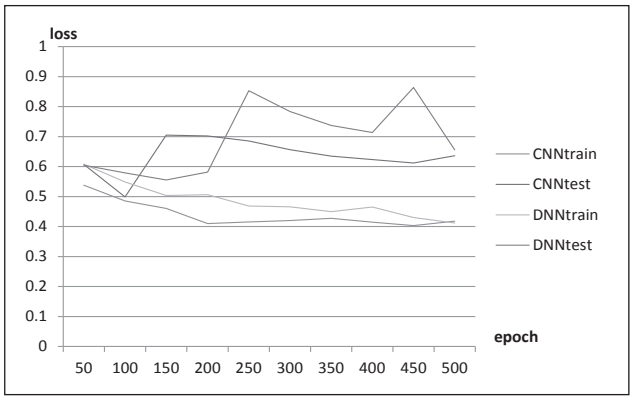

图 2　二次实验数据图

4 总结与展望

本文针对环境污染问题，利用深度学习去进行图像识别，从而将识别出的垃圾进行分类。获得较好的实验效果，且实验数据显示卷积神经网络的准确率要高于深度神经网络。

未来将根据不同城市的垃圾分类方法不同，数据集的分类还需更加详细，种类应更加全面，同时针对物品的不同部位属于不同类型的垃圾，还需要增加预识别功能，确定要不要对投放进来的散装垃圾进行拆分，实现垃圾一体化。

参考文献

* [1] 新华网 . 我国生活垃圾年产量已超过 4 亿吨 [DB/OL]. www.xinhuanet.com//energy/2017-09/18/c_1121678247.htm.2017-09-18.
* [2] 新华网 . 习近平对垃圾分类做出重要指示 [DB/OL].www.xinhuanet.com//2019-06/03/c_1124577181.htm,2019-06-03.
* [3] 新华网 . 垃圾分类有法可依 [DB/OL]http://www.xinhuanet.com/2019-04/24/c_1124408373.htm.2019-04-24.
* [4] 尹宝才 , 王文通 , 王立春 . 深度学习研究综述 [J]. 北京工业大学学报 , 2015, 41(01) : 48-59.
* [5] 李彦冬 , 郝宗波 , 雷航 . 卷积神经网络研究综述 [J]. 计算机应用 , 2016, 36(09) : 2508-2515+2565.

（上接第 6 页）

表1　视觉定位结果与机器人末端执行机构实际位置对比 单位：mm

序号	视觉定位结果	机器人末端位置
1	(444.8112,132.8965，-25.0719)	(443.9,133.7，-25.3)
2	(412.3321,197.5294，-25.1482)	(412.6,196.4，-24.3)
3	(457.5165,203.0563，-25.3341)	(456.2,202.8，-26.1)
4	(464.3551,259.3223，-25.1236)	(463.8,258.1，-25.2)
5	(528.1085,186.9865，-25.0125)	(526.9,186.3，-25.1)

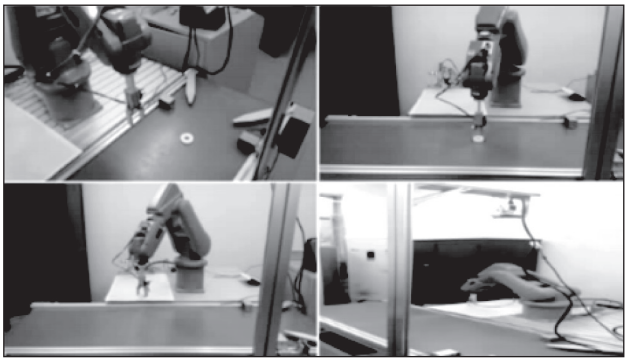

图 6　工件抓取实验

6 结束语

本文是根据工业机器人在生产线上的实际工作运用的一个全新课题，传统机电一体化、控制类系统已经不能满足现代化生产的需求，视觉定位抓取很好的解决了机器人不能识别工件的颜色、形状、大小等问题，有效的提高工件识别与抓取精度。但目前我们的国产视觉还处于起步阶段，受外界光源、环境因素的影响比较大，因此，我们更好抓紧时间完善视觉系统，为我国的工业产业转型升级做出更大贡献。

参考文献

* [1] 郭彤颖，安冬 . 机器人系统设计及应用 [M]. 北京 : 化学工业出版社 , 2016.
* [2] 刘小波 . 工业机器人技术基础 [M]. 北京 : 机械工业出版社，2016.
* [3] 许龙 . 基于机器视觉的 SMT 芯片检测方法研究［D］. 西安 : 西安电子科技大学 , 2014.

（上接第 84 页）

* [2] 刘思宏 , 余飞 . 基于大数据下 K-means 聚类算法的在线学习行为路径应用研究 [J]. 兰州文理学院学报 (自然科学), 2019, 33(01) : 70-74.
* [3]. 盘点谷歌内部如何使用深度学习技术 [J]. 信息与电脑 (理论版), 2017(16) : 9.
* [4]. 梁迎丽 , 梁英豪 . 人工智能时代的智慧学习 : 原理、进展与趋势 [J/OL]. 中国电化教育 , 2019(02) : 16-21[2019-01-30]
* [5] 张耀铭 , 张路曦 . 人工智能 : 人类命运的天使抑或魔鬼——兼论新技术与青年发展 [J]. 中国青年社会科学 , 2019, 38(01) : 1-23.
* [6] 周志华 . 机器学习 [M]. 清华大学出版社 .2016.1