DOI: 10.3969/j.issn.1001-8972.2010.14.107

机器学习理论研究

郭亚宁 冯莎莎
山东大学威海分校数学与统计学院 264209

摘 要

文章开始介绍了机器学习的几个典型算法，并对当前的热点统计学习理论基本概念及发展前景进行了分析，在文章最后介绍了机器学习理论与其他相关领域的关系。

关键词

机器学习；统计学习理论；数据挖掘；模式识别

Abstract

The article is beginning with several typical algorithm of Machine Learning. And then we analyzed the basic concepts and prospects of the current hot spot called Statistical Learning Theory. We introduced the relationship between the Machine Learning Theory and the other relative field.

Key words

machine learning; Statistical Learning Theory; data mining; pattern recognition

引言

机器学习的核心是学习。学习是人类特有的一项能力，如何让机器像人类一样，能够通过外界环境的影响来改善自己的性能，是机器学习领域研究的重点。什么是学习？不同领域给出了不同的观点，如今在机器学习领域影响最大的是 H. Simon 的观点：学习是这样的改进，系统经过这样的改进后，在完成同样的工作时能完成的更好。

机器学习的过程是一个从未知到已知的过程。如果一台机器拥有这样的程序，随着机器解决问题的增多，在该程序的作用下，机器性能或解决问题的能力增强，我们就说这台机器拥有学习能力。机器解决问题能力的增强主要表现在：初始状态下，对于问题 Q，机器给出结果 A，该机器在解决问题{Q1,Q2,…,Q m}后，再次遇到问题 Q 时给出结果 A1，而结果 A1 比结果 A 更精确，我们就说机器解决问题的能力得到了增强。

1 机器学习的主要算法

随着机器学习理论研究的逐渐深入，它的应用也日益广泛，许多优秀的算法应运而生。算法可以分为基于符号的和基于非符号的两类。前者包含机械式学习、归纳学习、基于解释的学习等，后者包括基于遗传算法的学习和基于神经网络的学习等，以下来介绍几种典型的机器学习算法。

1.1 机械式学习

1.1.1 主要思想。机械式学习又称为死记硬背式学习，是最原始的学习算法。顾名思义，机械式学习即为对每次输入的信息及解决的问题存入知识库，当再次遇到该问题时，直接查询知识库，得到该问题的解决办法。该问题的符号表示如下：待解决问题为{y1,y2,…,yn}，在输入信息{x1,x2,…xm}后，该问题得到了解决，于是将记录对{{x1,x2,…,xm},{y1,y2,…,yn}}存入知识库，以后当遇到问题{y1,y2,…,yn}时，查询知识库，取出{x1,x2,…,xm}作为对问题{y1,y2,…,yn}的解答。

1.1.2 算法评价。能实现机械式学习算法的系统只需具备两种基本技能：记忆与检索。此外，存储的合理安排、信息的合理结合以及检索最优方向的控制也是系统应该考虑的问题。该算法简单、容易实现、计算快速，但是由于系统不具备归纳推理的功能，对每个不同的问题，即使是类似的问题，也需要知识库中有不同的记录对，因此占用的大量的存储空间，这是典型的以空间换时间的算法。

1.2归纳学习

1.2.1 算法简述。归纳学习算法是研究最广泛的基于符号的学习算法。对于给定的信息，通过归纳推理，结合知识库里的信息，得出想要的结论。归纳学习的过程是由特殊实例推导出一般情况的过程，这样就使类似的问题可以利用同样的方法求解。

1.2.2 执行过程。先介绍两个概念：

<1>示例空间：系统中的训练集全体。

<2>规则空间：训练集中全体实例潜在的规则全体。

归纳学习的过程就是示例空间与规则空间的相互利用与反馈。1974 年，Simon 和 Lea 提出了双空间模型，形象的对这一执行过程进行了描述，如图 1 所示：

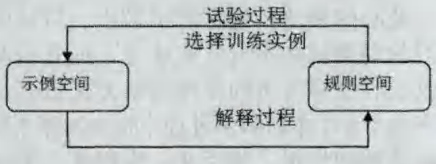

图 1

1.2.3 算法评价。归纳学习算法简单，节省存储空间，在一段时间内得到了广泛的应用。在应用过程中，该算法逐渐显现出它的缺点：

<1>归纳结论是通过对大量的实例分析得出的，这就要求在结论的得出要有大量实例作支撑，而这在许多领域都是无法满足的。

<2>归纳结论是由不完全训练集得出的，因而其正确性无法保证，只能使结论以一定概率成立。

<3>该算法通过对实例的分析与对比得出结论，对于信息的重要性与相关关系无法辨别。

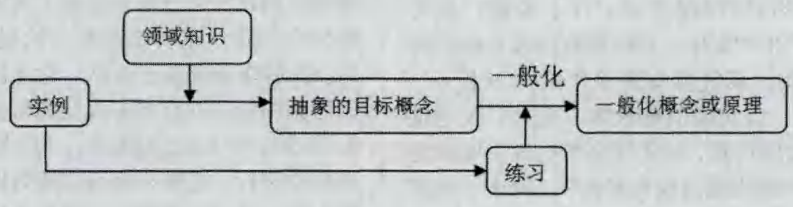

图 2

1.3基于解释的学习

1.3.1 算法思想。该算法的实现要求已知完整的领域知识与对应的一个实例，通过对该实例利用已知知识进行分析来完成对目标概念的学习，然后通过后继的不断练习，得到目标概念的一般化描述 。

1.3.2 执行过程。该学习算法的执行过程如图2所示。

1.3.3 算法评价。得到一个领域完善的知识往往是比较困难的，这就对该算法提出了更高的要求。为解决知识不完善领域的问题，有以下两个研究方向[2]：

<1>改进该算法使其在不完善的领域理论中依然有效。

<2>扩充该领域的知识使其拥有更强的解释能力。

通常情况下，第二种改进方法更重要些。

1.4基于神经网络的学习

1.4.1 算法简介。神经网络是由许多类似神经元的节点和它们之间带权的连接组成的复杂网络结构，是为模仿人类大脑的复杂神经结构而建立起来的抽象数据模型，希望相似的拓扑结构可以使机器像人脑一样进行数据的分析、存储与使用 。

1.4.2 算法思想。神经网络学习的过程就是不断修正连接权的过程。在网络的使用过程中，对于特定的输入模式，神经网络通过前向计算，产生一个输出模式，并得到节点代表的逻辑概念，通过对输出信号的比较与分析可以得到特定解。在整个过程中，神经元之间具有一定的冗余性，且允许输入模式偏离学习样本，因此神经网络的计算行为具有良好的并行分布、容错和抗噪能力[3]。

1.4.3 算法评价。神经网络学习算法是一种仿真算法，拥有良好的认识模拟能力和有高度的并行分布式处理能力。但神经网络模型及其参数设置难以确定，需要长时间的试验摸索过程，并且，对于最后得到的神经网络，其反映的知识往往难以让人理解。为解决这些问题，构造神经网络集成并从神经网络或神经网络集成中抽取规则成为当前研究的热点。

1.5基于遗传算法的学习

1.5.1 算法思想。遗传算法以自然进化和遗传学为基础，通过模拟自然界中生物的繁殖与进化过程，使训练结果逐渐优化。与遗传过程类似，在学习过程中，通过选择最好结果并使其组合产生下一代，使“优秀的遗传因子”逐代积累，最后得到最优的解。

1.5.2 算法评价。遗传算法解决了神经网络学习中的一个缺点，它不需要知道原始信息而只需知道学习的目的即可进行，具有很强的并行计算能力和适应能力。此外，遗传算法采取的随机搜索方法提高了该学习算法对全局搜索的能力。

遗传算法的缺点主要体现在三个方面：无法确定最终解的全局最优性；无法控制遗传过程中变异的方向；无法有效的确定进化终止条件。基于这三个缺点，有人提出了遗传算法与其他学习算法的结合，优点互补已达到更好的效果 。

2 统计学习理论

2.1基本概念

鉴于在实际问题中样本数总是有限的，这就导致传统的学习方法得不到尽如人意的结果。而统计学习理论正是研究小样本统计估计和预测的理论，它的核心思想是通过控制学习机器的容量实现对推广能力的控制[4]。统计学习理论的研究方向主要包括以下四个方面[5]：

<1> 经验风险最小化准则下统计学习一致性的条件；

<2> 在这些条件下关于统计学习方法推广性的界的结论；

<3> 在这些界的基础上建立的小样本归纳推理准则；

<4> 实现新的准则的实际方法(算法)。

其中，最具有指导意义的理论成果是推广性的界。下面介绍一个核心概念VC维。

2.1.1 VC维。VC维是这样定义的[6]：一个数据集中包含N个点，这N个点可以用2^N中方法区分正例和负例，即确定了 种不同的学习问题。如果对于这些问题中的任何一个，都存在假设类Π中的一个假设h将正例与负例分开，这时我们称Π散列了N个点。可以被Π散列的点的最大数目成Π的VC维，记作VC(Π)。VC维度量的是机器的学习能力。

2.1.2 推广性的界。推广性的界研究的是在各种函数集中，经验风险与实际风险的之间的关系。Vapnik和Chervonenkis经研究后得出如下结论：对预测函数集中的任意函数（包括使经验风险最小的函数），经验风险$R_{emp}(\omega)$与实际风险$R(\omega)$至少以$1-\eta$的概率满足 ：

$$R(\omega) \le R_{emp}(\omega) + \sqrt{\frac{\left| h(\ln\frac{2n}{h}+1) - \ln(\frac{\eta}{4}) \right|}{n}}$$

有公式可见，学习系统的实际风险受VC维h和样本数n的影响。要想降低风险以便更好的推广样本，除了使经验风险尽可能小外，还应该尽量降低VC维 。

2.2发展前景

以上的分析都是理论性的，对于VC维，除特殊函数集外，其值是难以计算的，因此，如何建立VC维的计算算法是当前研究的热点问题。另外就是推广性的界中界的计算方法，如何能得到范围更小的界是今后的研究方向之一。另外，抛开VC维，如何选择更好的反映机器学习能力的参数也是未来的一个发展目标。

3 机器学习与相关领域的关系

机器学习是牵涉面很广的学科，它的理论来源于许多学科，成果的应用也十分广泛。本文就机器学习与模式识别、数据挖掘等领域的联系以及它的一些应用做下简单介绍。

3.1机器学习与模式识别

机器学习是人工智能领域所要研究的核心问题之一，它的理论成果已经应用到了人工智能的各个分支中。简单地讲，机器学习使机器的性能得到提高，模式识别研究的问题是如何将不同的事物划分为不同的类别。机器学习算法在模式识别系统中的应用会使模式识别系统具备更强的分类能力。目前，在模式识别领域得以应用的机器学习算法主要有遗传算法、神经网络、支持向量机、k—近邻法等。模式识别过程如图3所示[7]。

3.2机器学习与数据挖掘

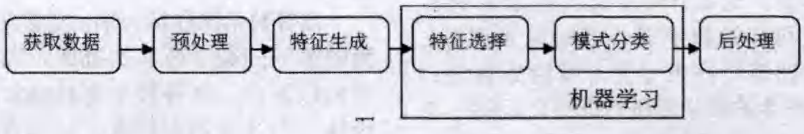

图3

下转第214页

学生英语听力能力不过关。实行双语教学也并非每个学生都可以轻松接受。对于普通院校来说，英语四、六级通过率本身就不高，实行双语教学对不同水平的学生存在一定的难度。在我校，虽然学校采取一定措施鼓励学生参加英语四、六级考试，但是我校对学生通过国家英语四、六级考试并没有硬性的规定。这使得学生缺少通过英语四、六级考试的压力，此外，不同专业学生英语考级的通过率各不相同。对于环境工程专业学生来讲，大部分学生在大三都能通过英语四、六级考试（2006 级学生在大三时英语四级通过率达到 91.6%），但是仍然存在一部分学生连国家英语四级考试都没有通过。而且尽管有些大学生英语四、六级都通过了，但英语实际应用能力，特别是听说能力较差。

因此，《环境科学概论》双语教学实施过程中，建议对上双语课学生的英语水平先进行测试，按不同英语水平进行分班教学，对于英语听说水平较弱学生采用中文上课，对英语水平较高的学生采用英语教学，以达到更好的教学效果 。

总之，双语教学在我国毕竟是新生事物，还有待更多的教育者进行实践和探讨。实施双语教学是实现高等教育国际化，面向现代化、面向未来、面向世界的复合型人才的有效途径。在普通高校实施《环境科学概论》双语教学过程中，建立一整套规范完善的双语教学模式，从原版教材的引进、双语教师的培养，到教学方法的更新完善，还有待于教育者更多的教学研究与实践，以期获得较丰富系统的经验理论。

参考文献

[1]孙凤娟. 法学专业双语教学的理论和实践探索[J]. 科技信息.2008, 14: 341-342.

[2]卢丹怀. 双语教学面临新挑战[J]. 全球教育展望.2001, 10:55-59.

[3]吕良环. 双语教学探析[J]. 全球教育展望.2001, (4): 66-73.

[4]蔺丰奇. 高校实施双语教学过程中存在的问题及对策[J]. 复旦教育论坛. 2003, 1(3): 21-24.

[5]彭小夏. 关于我国高校双语教学的现状问题及对策[J]. 读与写杂志.2008,5 (10):74-75.

[6]曹霞、王建生. 试论高校实施双语教学的挑战与对策[J]. 中国高教研究.2002, 9: 94-95.

[7]胡伟华. 我国高校双语教学现状分析[J]. 西安外国语大学学报.2008,16(4):93-96.

[8]蔡洁. 高校双语教学中的问题与对策[J]. 科技情报开发与经济.2006, 16 (13): 219.

[9]蒋隆敏, 凌智勇. 高校实施双语教学的实践与研究[J]. 江苏高教.2006,: 87-88.

[10]张彤,黄知超,杨连发.双语教学实施方法的几点思考[J]. 北京大学学报.2007, 5: 259-260.

[11]卢丹怀. 双语教育的实质、有效性及不同的教学语言[J]. 全球教育展望. 2004,(2).59-62.

[12]张颖, 单德鑫. 有关环境科学双语教学的思考[J]. 东北农业大学学报(社会科学版).2005, 3 (3): 80-81.

作者简介

黄云凤. 女, 1977 年生于福建惠安. 汉族. 讲师. 博士. 研究方向为环境规划与管理。

上接第 209 页

数据挖掘是从大量的数据中挖掘出隐含的、未知的、用户可能感兴趣的和对决策有潜在价值的知识和规则[8]，是在目前“数据爆炸而信息匮乏”的现实下发展起来的一种技术。与此同时，机器学习也迅猛发展起来。数据挖掘中的许多决策算法都来自于人工智能和机器学习，也有许多应用由机器学习和数据挖掘协同合作，如人工神经网络[9]等。

4 结束语

机器学习最近几年发展迅速，但其毕竟属于新兴领域，发展时间短，技术难题多，一直以来，机器的学习能力都是人工智能领域的“瓶颈”。一方面机器学习领域的发展限制了人工智能领域的发展，另一方面鉴于机器学习与其他领域的密切关系，这就要求研究者们在致力于研究机器学习的同时，可以在其他领域寻找新的学习算法和学习体制，并以此促进机器学习领域的新发展。

参考文献

[1] 杨炳儒. 知识工程与知识发现[M]. 北京：冶金工业出版社.2000.

[2] 闫友彪, 陈元琰. 机器学习的主要策略综述[J]. 计算机应用研究.2004(7):4-10.

[3] 刘琴. 机器学习[J]. 武钢职工大学学报.2001, 13(2):41-44

[4] 杨莉, 赵莉, 曹扬. 基于统计学习理论的信息安全风险管理[J].统计与决策. 2008(16):183-184

[5] Vapnik V.N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag.1995

[6] Ethem Alpaydin. Introduction to Machine Learning[M], 范明, 昝红英, 牛常勇, 译, 北京：机械工业出版社.2009:14-15

[7] 杜明, 周而重. 机器学习在模式识别中的应用研究[J]. 科技信息.2009(9): 37-38.

[8] 侯宇, 田静. 基于决策树方法的数据挖掘分析[J]. 华南金融电脑应用技术. 2009(8):42-43.

[9] 田文英. 机器学习与数据挖掘[J]. 石家庄职业技术学院学报.2004,16(6):30-32.