通过集成学习进行知识获取

周志华
(南京大学 计算机软件新技术国家重点实验室,南京 210093)

摘 要:在很多数据资源丰富的应用领域中,机器学习是进行自动或半自动知识获取的有力工具。简要介绍了我们通过集成学习进行知识获取的一些相关工作。

关键词:机器学习;知识获取;集成学习;数据挖掘
中图分类号:TP182 文献标识码:A 文章编号:1673-825X(2008)03-0361-02

Knowledge acquisition via ensemble learning

ZHOU Zhi-hua
(National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210093, P. R. China)

Abstract: In many data rich application domains, machine learning is a powerful tool for automated or semi-automated knowledge acquisition. This extended abstract will briefly introduce some of our works related to knowledge acquisition via ensemble learning.

Key words: machine learning; knowledge acquisition; ensemble learning; data mining

目前被广泛采用的机器学习定义是"利用经验来改善计算机系统自身的性能"。由于"经验"在计算机系统中主要是以数据的形式存在的,因此机器学习需要设法对数据进行分析,而这也正是该领域目前的主流研究内容。值得注意的是,在很多数据资源丰富的应用领域中,机器学习是进行自动或半自动知识获取的有力工具,利用机器学习技术对数据进行分析可以对传统的知识获取技术起到辅助作用。

集成学习(ensemble learning)利用多个学习器来解决问题,可以显著提高学习系统的泛化能力,因此曾被著名学者 T. G. Dietterich 列为机器学习四大研究方向之首。目前已经有很多著名有效的集成学习方法,例如 Boosting,Bagging 等。

由于集成学习利用多个学习器可以获得比仅使用单一学习器更强的泛化能力,因此曾有一些学者试图通过使用大量的个体学习器来获得更好的性能。我们的研究表明,个体学习器未必是越多越好,而是"many could be better than all",对个体学习器中进行选择后再集成可望获得更好的性能,由此,我们提出了选择性集成(selective ensemble)这一范式[1]。

在训练出具有好性能的集成之后,就可以通过 C4.5Rule-PANE 算法[2]来获得泛化能力强、可理解性好的符号规则。大致来说,该算法随机产生输入数据从而利用集成产生一个新数据集,再利用符号规则学习方法对这个新数据集进行学习以产生规则。总的来看,C4.5Rule-PANE 这样的技术涉及到两次学习过程,即先利用原始数据学习得出一个学习器,再利用该学习器产生的数据进行第二次学习。我们对这样的"二次学习"技术奏效的原因进行了研究,分析结果表明,由于实际应用中训练集通常包含噪音、也通常不能完全表达目标分布,因此二次学习技术在真实情况下通常是可以奏效的[3]。生物信息学领域的应用显示出该技术在科学假设生成方面有很好的用途[4]。

成功地通过集成进行知识获取的重要前提是集成有强泛化能力,在有标记数据很多的情况下,利用传统集成学习方法就能产生符合条件的集成,但在有标记数据很少时,如何产生具有强泛化能力的集

收稿日期:2008-03-17 修订日期:2008-3-26
周志华(✉),南京大学 计算机软件新技术国家重点实验室,南京 210093
基金项目:国家自然科学基金重点项目(60635030);国家自然科学基金创新群体项目(60721002)

成就成为一个问题。我们对此进行了研究,提出了使用3个学习器、可以有效地利用未标记数据(unlabeled data)来提高泛化能力的tri-training算法[5],并将其推广,产生了可以使用更多学习器的co-forest算法[6]。

将上述工作结合起来,我们得到一个通过集成学习来进行知识获取的框架:先利用有标记数据训练出一个初始集成,如果有未标记数据可用,就利用未标记数据来改善集成的性能,此后,通过二次学习来获取符号规则形式的知识。

参考文献:

[1] ZHOU Z H, WU J, TANG W. Ensembling neural networks: Many could be better than all. Artificial Intelligence[J]. 2002, 137(1-2): 239-263.

[2] ZHOU Z H, JIANG Y. Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble[J]. IEEE Trans. Information Technology in Biomedicine, 2003, 7(1): 37-42.

[3] ZHOU Z H, JIANG Y. NeC4.5: Neural ensemble based C4.5[J]. IEEE Trans. Knowledge and Data Engineering, 2004, 16(6): 770-773.

[4] JIANG Y, LI M, ZHOU Z H. Generation of comprehensible hypotheses from gene expression data. [EB/OL]. [2008-3-24]. http://lamda.nju.edu.cn/pub.htm.

[5] ZHOU Z H, LI M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[6] LI M, ZHOU Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Transactions on Systems, Man and Cybernetics - Part A, 2007, 37(6): 1088-1098.

作者简介:

周志华(1973-),男,江苏常州人,教授,博士生导师,教育部长江学者特聘教授,主要研究领域为人工智能、机器学习、数据挖掘、模式识别、信息检索等。E-mail: Zhouzh@nju.edu.cn。

(责任编辑:段明琰)

我校郑建宏教授荣获全国"五一"劳动奖章

2008年4月29日,重庆市总工会在渝州宾馆举行庆祝"五一"劳动节颁奖典礼,全市共有19名同志荣获全国"五一"劳动奖章,我校郑建宏教授获此殊荣。郑建宏教授现任我校博士生导师、教育部移动通信工程中心主任、重庆移动通信工程研究中心副主任、重邮信科公司副总经理。从1983年至今,一直致力于教学、科研特别是TD-SCDMA 3G手机研发及产业化工作,长期从事通信工程、信号与信息处理方面的研究与应用开发,具有丰富的通信业知识和实践经验,出版学术专著1部,并在国内外重要学术刊物和学术会议上发表学术论文40余篇,申请国内发明专利32项,授权8项,是我国"863"通信学科领域项目评审专家,国务院政府特殊津贴获得者。

郑建宏教授怀着振兴我国民族移动通信事业的使命意识和美好梦想,从1998年开始从事TD-SCDMA第三代移动通信领域的研究工作,凭着一股"开弓没有回头箭"的大无畏精神,站在TD-SCDMA科技的前沿,成功创建和培养了一支高水平的TD-SCDMA研发队伍,并带领团队一起积极完成了众多的科研任务,通过技术创新,取得了许多突破性的成果,为我国TD-SCDMA被纳入国际第三代移动通信三大主流标准之一作出了突出贡献(该成果荣获2003年国家科技进步二等奖),2001年研制出了世界第一款3G实验样机,2003年研制出世界第一款TD-SCDMA (TSM)手机,2005年成功研制出世界第一款0.13微米工艺的TD-SCDMA手机基带核心芯片(该成果荣获2005年中国高校"十大科技进展"和2006年重庆市技术发明一等奖)