计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
45(Suppl.): 300~304, 2008

# 主动学习研究综述

龙 军 殷建平 祝 恩 赵文涛
(国防科学技术大学计算机学院 长沙 410073)
(jdragon_nudt@hotmail.com)

# A Survey of Active Learning

Long Jun, Yin Jianping, Zhu En, and Zhao Wentao
(*School of Computer Science, National University of Defense Technology, Changsha* 410073)

**Abstract** Recently, active learning has become a hotspot of the research in machine learning. The technology efficiently reduces the sample complex through actively selecting the instance to learn. Introduced in this paper is the latest research progress of active learning, including sample complex, sample selection methods and real world application. Finally, the open problems remained in active learning are pointed out.

**Key words** active learning; instance selection

摘 要 近年来,主动学习成为机器学习领域的研究热点.这一技术通过主动选择要学习的样例从而有效地降低学习算法的样本复杂度.介绍当前主动学习的研究进展,包括主动学习的样本复杂度,样例选择算法和实际应用,最后指出主动学习领域中还保留的开放问题

关键词 主动学习;样例选择

中图法分类号 TP18

传统监督学习问题中,学习算法以外界给定的已标注样例集作为训练集进行训练,从中归纳出模型.而在很多的现实应用中,对样例集进行标注代价昂贵、枯燥乏味或是异常困难,而获取未被标注的样例则相对容易.例如,针对基因序列测试,要标注一段基因序列需要进行代价昂贵的实验,相反,获取基因片段代价则相对小得多.

面对这种情况,传统的监督学习方法(即被动学习)构建正确率满足要求的分类器将十分困难.因此,主动学习方法被提出以有效地处理这类问题.在主动学习中,学习器能够主动地选择包含信息量大的未标注样例并将其交由专家进行标注,然后置入训练集进行训练,从而在训练集较小的情况下获得较高的分类正确率,这样可以有效地降低构建高性能分类器的代价.

主动学习方法一般可以分为两个部分:学习引擎和选择引擎.学习引擎负责维护一个基准分类器,并使用监督学习算法对系统提供的已标注样例集进行学习从而使该分类器的性能提高;而选择引擎负责运行样例选择算法选择一个未标注的样例并将其交由人类专家进行标注,再将标注后的样例加入到已标注样例集中.学习引擎和选择引擎交替工作,经过多次循环,基准分类器的性能逐渐提高,当满足预设条件时,整个过程终止.

主动学习在降低样本复杂度方面比传统被动学习具有优势,近年来取得较大的发展,但就现状而言,仍然存在很多值得深入研究的问题

## 1 样本复杂度

首先关注的是在理论上,主动学习能在多大程度上降低样本复杂度.当前已经有了很多研究,但

收稿日期:2007-07-10
基金项目:国家自然科学基金项目(60603015,60603062);湖南省自然科学基金项目(06jj30035)

总的来说, 这方面的研究仍然还没有达到成熟.

根据 PAC 理论, 为获取期望错误率小于 ε 的分类器, 传统监督学习算法的样本复杂度为 $O\left(\frac{1}{\varepsilon}\ln\left(\frac{1}{\varepsilon}\right)\right)$, 主动学习需要获得比这更低的样本复杂度才有实际意义.

当前已获得的成果包括以下几类: ①样例集中无噪声的问题; ②样例集中存在噪声的问题; ③主动学习在何种情况下不能有效降低样本复杂度的问题.

对问题①, Cohn 等人[1]给出了一种最简单的情况, 即确定单位线段上的 0-1 分界线问题. 他指出, 为了使期望错误率小于 ε, 在采用二分搜索并且样例均匀分布的情况下, 主动学习算法的样本复杂度为 $O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$.

Fruend 等人[2]对更复杂的情况作了分析. 他将问题抽象如下: 假设空间中包含的都是齐次线性分界面, 同时样例均匀分布于 $R^d$ 上的单位球面, 而且假设空间中存在完美分类器(即与所有样例一致). Fruend 等人[2]证明, QBC 算法能够只需要 $O\left(d\log\frac{d}{\varepsilon}\right)$ 次标注就可以学习到一个错误率小于 ε 的分类器, 而对应的学习线性分界面的监督学习算法则需要 $\Omega(d/\varepsilon)$ 的样本复杂度. Fruend 等人的这一结果是历年来能取得的最好的样本复杂度结果.

对问题②, Kaariainen[3]证明当噪声率为 η 时, 样本复杂度的一般下界为 $\Omega\left(\frac{\eta^2}{\varepsilon^2}\right)$. 更进一步, Kaariainen 假定当假设空间中性能最好的分类器的错误率至多为 $\beta>0$ 时, 为使错误率小于 ε, 主动学习的样本复杂度下界为 $\Omega\left(\frac{\beta^2}{\varepsilon^2}\log(1/\delta)\right)$, 其中 δ 为可靠性参数(表示学习以 1− δ 的概率能够成功).

Balcan 等人[4]则提出一种主动学习算法, 该算法在未标注样例集包含噪声的前提下, 针对 Fruend 等人[2]所描述的问题情况, 可以获得样本复杂度为 $O\left(d^2\log\frac{1}{\varepsilon}\right)$ 的一般上界.

对问题③, Dasgupta[5]证明, 当假设空间包含非齐次线性分界面, 则存在一些目标假设, 使得无论怎样使用主动学习算法, 达到小于 ε 的错误率的一般下界为 $\Omega\left(\frac{1}{\varepsilon}\right)$. 同时, Dasgupta 等人[6]还证明, 在样例均匀分布时, 无论用什么样的主动学习算法, 标准感知机都需要 $\Omega\left(\frac{1}{\varepsilon^2}\right)$ 次标注使错误率小于 ε.

## 2　样例选择算法

根据获得未标注样例的方式不同, 可以将主动学习算法分为两种类型: 基于流的和基于池的. 基于流(stream-based)的主动学习[2,7-8]中, 未标注的样例按先后顺序逐个提交给选择引擎, 由选择引擎决定是否标注当前提交的样例, 如果不标注, 则将其丢弃. 而基于池(pool-based)的主动学习[9-11]中则维护有一个未标注样例的集合, 由选择引擎在该集合中选择当前要标注的样例. 以下分别介绍.

### 2.1　基于池的样例选择算法

基于池的样例选择算法是当前研究得最为充分的. 按照选择的标准不同可以分为以下几类:

基于不确定度缩减的方法、基于版本空间缩减的方法、基于未来泛化错误率缩减的方法和其他方法. 分别介绍如下:

1) 基于不确定度缩减的方法

这类方法选择那些当前基准分类器最不能确定其分类的样例进行标注. 这种方法以信息熵作为衡量样例所含信息量大小的度量, 而信息熵最大的样例正是当前分类器最不能确定其分类的样例. 从几何角度看, 这种方法优先选择靠近分类边界的样例, 所以又可以称为最近边界方法. 这种方法可以应用于任何形式的基准学习器, 如 logistic regression[9]、隐马尔可夫模型[12]、支撑向量机[13-14]以及归纳逻辑编程[15]等. 它在大多数问题上能取得比随机选择更好的性能, 但有可能采集到孤立点.

2) 基于版本空间缩减的方法

这类方法选择那些训练后能够最大程度缩减版本空间的样例进行标注. 在二值分类问题中, 这类方法选择的样例总是差不多平分版本空间, 其思想来源于二分搜索. 这包括 QBC[16], SG-net[1], QBag[17], QBoost[17]和 Active Decorate[18]等.

QBC 算法[2]从版本空间中随机选择若干假设构成一个委员会, 然后选择委员会中的假设预测分歧最大的样例进行标注. 评价分歧度有如下标准: 投票熵[7]、Jensen-Shannon 分歧度[19]、Kullback-Leibler 分歧度[20]等. Fruend 等人[2]给出了 QBC 方法的严格理论证明. 为了优化委员会的构成, 增强其多样性, QBag[17], QBoost[17]和 Active Decorate[18]算法分别采用 Bagging, AdaBoost 和 Decorate 等成熟的分类器集成算法从版本空间中产生委员会.

3) 基于泛化误差缩减的方法

这类方法试图选择那些能够使未来泛化误差最大程度减小的样例. 其一般过程为: 首先选择一种损失函数用于估计未来错误率, 然后将未标注样例集中的每一个样例都作为下一个可能的选择, 分别估计其能给基准分类器带来的误差缩减, 选择估计误差缩减最大的那个样例进行标注. 当前针对不同的基准分类器提出相应的算法, 如朴素贝叶斯[21]、贝叶斯网络[22]、最近邻算法[23]等. 这种方法直接针对分类器性能的最终评价指标, 理论上具有很好的效果, 但计算量较大, 同时损失函数的精度对性能的影响也至关重要.

4) 其他方法

包括各种难以归入以上分类的主动学习算法, 包括 COMB[24]、多视图(view)主动学习[11]、预聚类主动学习[25]等.

COMB 算法[24]组合 3 种不同的主动学习器, 迅速切换到当前性能最好的学习器从而使选择样例尽可能高效.

多视图主动学习[11]用于学习问题为多视图学习的情况, 选择那些使不同视图的预测分类不一致的样例进行学习. 其中, 视图是指样例中足够做出分类判断的特征集合. 这种方法对处理高维的主动学习问题非常有效, 但不适用于低维问题.

预聚类主动学习[25]认为基于不确定度缩减的方法会忽略样例的先验分布, 而样例的分布恰恰有可能蕴涵丰富的信息. 因此, 首先运行聚类算法, 然后选择样例时优先选取最靠近分类边界的样例和最能代表聚类的样例(即聚类中心).

### 2.2 基于流的样例选择算法

基于池的算法大多可以通过调整以适应基于流的情况. 但由于基于流的算法不能对未标注样例逐一比较, 需要对样例的相应评价指标设定阈值, 当提交给选择引擎的样例评价指标超过阈值, 则进行标注. 但这种方法需要针对不同的任务调整阈值, 所以难以作为一种成熟的方法投入使用.

QBC 算法[3]也曾用于解决基于流的主动学习问题. 样例以流的形式连续提交给选择引擎, 选择引擎选择那些委员会(此处委员会只由两个成员分类器组成)中的成员分类器预测不一致的样例进行标注.

Saunier 等人[26]试图用基于流的主动学习算法处理道路交叉点的冲撞风险评估问题. 但他指出, 由于面对的问题中请求人类专家对样例进行标注的代价小, 所以可以直接让人类专家标注每一个提交的样例, 然后再对这些样例进行选择性的学习.

不同于 Freund 等人[2]的工作, Cesa-Bianchi 等人[27]针对的问题为: 流中的样例独立同分布(不要求均匀分布)地取自 $R^d$ 上的单位球面, 其标注由一个二值线性概率函数生成且线性系数未知. 他提出的算法维持一个对当前训练集的最小平方估计, 该估计随着算法迭代的进行而逐步更新, 当新样例自流中提交给选择引擎时, 计算对新样例的最小平方估计的边缘值, 当该边缘值小于一个阈值时, 标注该样例, 而该阈值随着算法迭代的进行逐步调整. Cesa-Bianchi 等人[27]证明, 随着提交给选择引擎的样例增多, 该算法需要标注的样例呈对数级地减少.

## 3 应 用

当前, 主动学习已逐步投入具体的应用, 其中包括文档分类及信息提取、图像检索、入侵检测、Web 分析和视频分析等广大领域的实际问题.

1) 文档分类及信息提取

由于文档分类和信息提取任务中对人类专家的依赖很重, 需要枯燥乏味的大量劳动, 所以这一领域较早地刺激了主动学习方法的出现[9-10, 15, 21, 28]. 这一应用涵盖的面也很广, 基于池的主要样本选择算法基本上都涉及到了. 但相关的研究大多集中在 2001 年以前, 近年来较少涉及.

Lewis 等人[9]以贝叶斯方法作为基准学习器, 使用基于不确定度缩减的样例选择算法进行文本分类. McCallum 等人[10]将 EM 算法同基于 QBC 方法的主动学习结合. EM 算法能够有效地利用未标注样例中的信息提高基准分类器的分类正确率, 而 QBC 算法能迅速缩减版本空间, 这两种技术的结合经实验证明是相当有效的. Roy 等人[21]使用贝叶斯方法作为基准学习器, 采用基于泛化误差缩减的样本选择算法处理文本分类. 该算法使用损失函数评估每个样例可能导致的未来泛化误差的缩减, 选择能引起未来泛化误差缩减最大的样例进行标注. Tong 等人[28]则使用支撑向量机作为基准学习器, 采用最近边界方法作为样例选择算法.

2) 图像检索

相对于文本, 图像中蕴含的信息量更大. 因此图像检索也是主动学习的一个重要应用领域[29].

Tong 等人[29]使用支撑向量机为基准学习器的主动学习算法来处理图像检索. 该算法采用最近边界方法作为样例选择算法, 同时将图像的颜色、纹理

等提取出来作为部分特征进行学习.

3) 入侵检测

由于入侵检测系统较多地依赖于专家知识和有效的数据收集,所以有研究者采用主动学习算法降低这种依赖性.

Almgren 等人[30]以支撑向量机作为基准学习器,采用最近边界方法作为样例选择算法. Almgren 等人针对 1999 年 KDD 入侵检测数据竞赛的样例集做了测试,实验结果表明,采用主动学习算法可以大幅度降低需要标注的样例集.

其他还有一些针对 Web 信息提取[11]、可视目标检测[31]、信息检索[32]、基因序列分析[33]和道路安全评估[26]的研究.

## 4　开放问题

目前对于以下情况的研究还很少[34]:1)学习器不知道样例如何分布的情况;2)低误差率情况下的均匀或任意分布的有高样本复杂度边界的高效学习算法;3)空间和时间复杂度不能随着可见样例和错误上升而上升的情况;4)针对其他概念类或者一般概念类的学习问题.

对主动学习的现实应用进一步的研究可以包括:1)同具体应用领域的先验知识结合,研究更加高效的主动学习算法,以减少标注样例的代价. 2)结合代价敏感学习及不平衡数据集学习等技术,有效处理实际问题域真正关注的问题,而非仅仅以提高分类正确率为惟一标准.

## 参　考　文　献


[1] D Cohn, Atlas R Ladner. Improving generalization with active learning. Machine Learning, 1994, 5(2): 201—221

[2] Y Freund, H S Seung, E Shamir, *et al*. Selective sampling using the query by committee algorithm. Machine Learning, 1997, 28(2-3): 133—168

[3] M Kaariainen. Active learning in the non-realizable case. In: Proc of the 17th Int'l Conf on Algorithmic Learning Theory. Berlin: Springer, 2006. 63—77

[4] M -F Balcan, A Beygelzimer, J Langford. Agnostic active learning. In: Proc of the 23rd Int'l Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2006

[5] S Dasgupta. Coarse sample complexity bounds for active learning. In: Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2005

[6] S Dasgupta, A T Kalai, C Monteleoni. Analysis of perceptron-based active learning. In: Proc of the 18th Annual Conf on Learning Theory. Berlin: Springer, 2005

[7] I Dagon, S Engelson. Committee-based sampling for training probabilistic classifiers. In: Proc of the 12th Int'l Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1995. 150—157

[8] S Arganmon-Engelson, I Dagon. Committee-based sample selection for probabilistic classifiers. Journal of Artificial Intelligence research, 1999, 11: 335—360

[9] D D Lewis, W A Gail. A sequential algorithm for training text classifiers. In: Proc of the 17th ACM Int'l Conf on Research and Development in Information Retrieval. Berlin: Springer, 1994. 3—12

[10] A K McCallum, K Nigam. Employing EM in pool-based active learning for text classification. In: Proc of the 15th Int'l Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1998

[11] I Muslea, S Minton, C A Knoblock. Active learning with multiple view. Journal of Artificial Intelligence Research, 2006, 27: 203—233

[12] T Shcheffer, S Wrobel. Active learning of partially hidden Markov models. In: Proc of the ECML/PKDD-2001. Berlin: Springer, 2001

[13] G Schohn, D Cohn. Less is more: Active learning with support vector machines. In: Proc of the 17th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufmann, 2000. 839—846

[14] C Campbell, N Cristianini, A Smola. A query learning with large margin classifiers. In: Proc of the 17th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufmann, 2000. 111—118

[15] C Tohompson, M E Califf, R Mooney. Active learning for natural language parsing and information extraction. In: Proc of the 16th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufmann, 1999. 406—414

[16] H S Seung, M Opper, H Sompolinsky. Query by committee. Annual Workshop on Computational Learning Theory, Pittsburgh, Pennsylvania, United States, 1992

[17] N Abe, H Mamitsuka. Query learning strategies using boosting and bagging. In: Proc of the 15th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufmann, 1998. 1—10

[18] P Melville, R J Mooney. Diverse ensembles for active learning. In: Proc of the 21th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufmann, 2004

[19] P Melville, S M Yang, M Saar-Tsechansky, *et al*. Active learning for probability estimation using Jensen-Shannon divergence. The 16th European Conf on Machine Learning, Porto, Portugal, 2005

[20] F Pereira, N Tishby, L Lee. Distributional clustering of English words. In: Proc of the 31st ACL. Morristown, NJ, USA: Association for Computational Linguistics, 1993

[21] N Roy, A McCallum. Toward optimal active learning through sampling estimation of error reduction. In: Proc of the 18th Int'l Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2001. 441—448

[22] S Tong, D Koller. Active learning for parameter estimation in Bayesian networks. In: Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000. 647—653

[23] M Lindenbaum, S Markovitch, D Rusakov. Selective sampling for nearest neighbor classifiers. Machine Learning, 2004, 54(2): 125—152

[24] Y Baram, R El-Yaniv, K Luz. Online choice of active learning algorithm. In: Proc of the 20th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufmann, 2003

[25] H T Nguyen, A Smeulders. Active learning using pre-clustering. The 21st Int'l Conf on Machine Learning, Banff, Canada, 2004

[26] N Saunier, S Midenet, A Grumbach. Stream-based learning through data selection in a road safety application. In: Proc of 16th European Conf of Artificial Intelligence. Amsterdam: IOS Press, 2004. 107—117

[27] N Cesa-Bianchi, A Conconi, C Gentile. Learning probabilistic linear-threshold classifiers via selective sampling. In: Proc of the 16th COLT. Berlin: Spriniger, 2003

[28] S Tong, D Koller. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2001, 2: 45—66

[29] S Tong, E Chang. Support vector machine active learning for image retrieval. In: Proc of the 9th ACM Int'l Multimedia Conf. New York: ACM Press, 2001. 107—119

[30] M Almgren, E Jonsson. Using active learning in intrusion detection. In: Proc of the 17th IEEE Computer Security Foundations Workshop. Washington, DC: IEEE Computer Society Press, 2004

[31] Y Abramson, Y Freund. Active learning for visual object detection. http://www.cs.ucsd.edu/Dienst/UI/2.0/Describe/ncstrl.ucsd_cse/CS2006-0871, 2006

[32] C Zhang, T Chen. An active learning framework for content-based information retrieval. IEEE Trans on Multimedia, 2002, 4(2): 260—268

[33] R Singh, N Palmer, D Gifford, *et al*. Active learning for sampling in time-series experiments with application to gene expression analysis. The 22nd Int'l Conf on Machine Learning, Bonn, Germany, 2005

[34] C Monteleoni. Efficient algorithms for general active learning. The 19th Annual Conf on Learning Theory, Pittsburgh, PA, USA, 2006

龙 军 男, 1978 年生, 博士研究生, 主要研究方向为主动学习、安全

殷建平 男, 1964 年生, 教授, 主要研究方向为计算理论、机器学习、无线传感器网络

祝 恩 男, 1976 年生, 讲师, 主要研究方向为指纹识别

赵文涛 男, 1975 年生, 副教授, 主要研究方向为计算机网络