◎博士论坛◎

主动学习算法综述

刘 康,钱 旭,王自强
LIU Kang, QIAN Xu, WANG Ziqiang
中国矿业大学(北京) 机电与信息工程学院,北京 100083
College of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing 100083, China

LIU Kang, QIAN Xu, WANG Ziqiang. Survey on active learning algorithms. Computer Engineering and Applications, 2012, 48(34):1-4.

Abstract: As a method of constructing an effective training set, the goal of active learning algorithm is to find informative sample which can enhance the classification results of the model during the iteration, thereby reducing the size of the training set and improving the efficiency of the model within the limited time and resources. At present, active learning has become a hot issue in the field of pattern recognition, machine learning and data mining. The fundamental ideas, some latest research results and algorithm analysis of active learning are introduced. Some problems for further research are presented and analyzed.

Key words: active learning; pattern recognition; machine learning

摘 要: 主动学习算法作为构造有效训练集的方法,其目标是通过迭代抽样,寻找有利于提升分类效果的样本,进而减少分类训练集的大小,在有限的时间和资源的前提下,提高分类算法的效率。主动学习已成为模式识别、机器学习和数据挖掘领域的研究热点问题。介绍了主动学习的基本思想,一些最新研究成果及其算法分析,并提出和分析了有待进一步研究的问题。

关键词: 主动学习;模式识别;机器学习
文献标识码: A **中图分类号:** TP181 doi:10.3778/j.issn.1002-8331.1205-0149

1 引言

监督学习模型,例如:支持向量机(SVMs)[1]或者神经网络[2],广泛应用于分类问题[3]。所有分类模型都需使用标记样本训练,并且分类模型的分类效果依赖于标记样本的质量。因此,训练样本需完整地表示所含类别的统计属性。然而,获取训练样本不仅费时、费力,而且训练集包含大量的冗余样本。为了尽可能地减小训练集及标注成本,在机器学习领域中,提出主动学习方法,优化分类模型。

主动学习算法可以由以下五个组件进行建模[4]:

$$A=(C,L,S,Q,U)$$

其中 C 为一个或一组分类器; L 为一组已标注的训练样本集; Q 为查询函数,用于在未标注的样本中查询信息量大的样本; U 为整个未标注样本集; S 为督导者,可以对未标注样本进行标注。主动学习算法主要分为两阶段:第一阶段为初始化阶段,随机从未

基金项目: 国家自然科学基金项目(No.70701013);中国博士后科学基金项目(No.2011M500035);高等学校博士学科点专项科研基金(No.20110023110002)。

作者简介: 刘康(1987—),男,博士研究生,主要研究方向:机器学习与模式识别;钱旭(1962—),男,博士,教授,博士生导师,主要研究方向:机器学习与信息融合;王自强(1973—),男,博士研究生,主要研究方向:机器学习与模式识别。
E-mail:liukang1112@gmail.com

收稿日期: 2012-05-18 **修回日期:** 2012-07-30 **文章编号:** 1002-8331(2012)34-0001-04
CNKI **出版日期:** 2012-08-30 http://www.cnki.net/kcms/detail/11.2127.TP.20120830.1719.006.html

标注样本中选取小部分，由督导者标注，作为训练集建立初始分类器模型；第二阶段为循环查询阶段，S 从未标注样本集 U 中，按照某种查询标准 Q，选取一定的未标注样本进行标注，并加到训练样本集 L 中，重新训练分类器，直至达到训练停止标准为止。

主动学习算法是一个迭代的过程，分类器使用迭代时反馈的样本进行训练，不断提升分类效率。目前，常用于分类问题的主动学习算法有三种形式[5]：(1)基于委员会的启发式方法(QBC)；(2)基于边缘的启发式方法(MS)；(3)基于后验概率的启发式方法(PP)。

2 主动学习算法

2.1 基于委员会的主动学习算法

主动学习方法选择一定数量的分类模型，构成分类委员会。利用初始训练集训练委员会中的每个模型，并将训练完成的模型用于分类未标记样本池中的样本。该方法选择所有模型分类结果中最不一致的样本。

2.1.1 熵值装袋查询(EQB)

为了减少计算复杂度及假设空间的搜索时间，提出基于装袋的熵查询方法构造委员会[6]。将原始训练集划分为 k 个训练集，然后，每个训练集被用于训练模型，并对未标记样本池进行预测，对每个样本 $x_i \in U$ 都有 k 个标签。该方法使用熵值度量预测标签的信息量，选择具有最大熵值的样本。具体形式如下所示：

$$x^{\mathrm{EQB}} = \arg\max_{x_i \in U} \left\{ \frac{H^{\mathrm{BAG}}(x_i)}{\log(N_i)} \right\} \tag{1}$$

其中 $H^{\mathrm{BAG}}(x_i)$ 表示熵值，N_i 表示预测样本 x_i 的类别的数量，$1 \le N_i \le N$。

当所有的分类器对样本的预测值一致时，委员会返回熵值为0，将此样本加入训练集并不会提高分类器的效果。相反，如果预测值最不一致，其返回的熵值最大，将此样本加入训练集能显著提高分类器的效果。

2.1.2 自适应不一致最大化(AMD)

针对高维数据，需要将特征空间划分为一定数量的子集，并使用该子集构造委员会[7-8]。给定输入空间，根据数据子集 x^v，将 d 维输入划分为不相交的 V 个部分，其中 $\bigcap_{v=1}^{V} x^v = x$。选择样本的准则如下所示：

$$x^{\mathrm{AMD}} = \arg\max_{x_i \in U} \left\{ H^{\mathrm{MV}}(x_i) \right\} \tag{2}$$

对于特定的子集，H^{MV} 表示分类器预测值的熵。

$$p^{\mathrm{MV}}\left(y_i^* = w \middle| x_i^v\right) = \frac{\sum_{v=1}^{V} \boldsymbol{W}^{\varepsilon-1}(v,w)\delta\left(y_{i,v}^*, w\right)}{\sum_{v=1}^{V}\sum_{j=1}^{N_i} \boldsymbol{W}^{\varepsilon-1}(v,w)} \tag{3}$$

其中 $\boldsymbol{W}^{\varepsilon-1}$ 是 $N \times V$ 维权重矩阵，在每次迭代中，$\boldsymbol{W}^{\varepsilon-1}$ 根据第 $\varepsilon-1$ 次迭代抽样获取样本的真实标签值实时更新。

$$\boldsymbol{W}^{\varepsilon}(v,w) = \boldsymbol{W}^{\varepsilon-1}(v,w) + \boldsymbol{\delta}\left(y_{i,v}, w\right), \forall i \in S \tag{4}$$

权重矩阵加权每个视角的确信度，从而预测标签。

2.2 基于边缘的主动学习算法

假设考虑二分类问题，样本点 x_i 到分类超平面的距离由下述公式给出：

$$f(x_i) = \sum_{j=1}^{n} \alpha_j y_j K(x_j, x_i) + b \tag{5}$$

$K(x_j, x_i)$ 表示核函数，定义了候选样本点 x_i 与支持向量 x_j 之间的相似度。支持向量机的训练集是由稀疏矩阵表示，所以被选择的样本的权重值为非零值(即 $\alpha_i > 0$)。换言之，样本点越靠近当前分类模型的边缘，该样本越有可能成为支持向量。

2.2.1 边缘抽样

基于边缘查询的方法主要用于支持向量机模型的主动学习中，数据点距分类超平面间的距离，即决策函数的绝对值，能够直观地估计出未标记样本的确定性程度[9-10]。

考虑一个未标记样本池，在多分类问题中，采用一对一的策略，边缘抽样的启发式方法最小化式(6)得到样本：

$$x^{\mathrm{MS}} = \arg\min_{x_i \in U} \left\{ \min_{w} \left| f(x_i, w) \right| \right\} \tag{6}$$

其中 $f(x_i, w)$ 表示样本点到分类超平面的距离。距分类界面越近的样本，分类模型对其确信度越低，对分类界面而言，该样本具有大量的分类信息。

2.2.2 基于多层次不确定性抽样

基于边缘抽样的方法扩展到多分类问题时，支持向量机还可以考虑不同类别的距离差异[11-12]。提出了多层次的不确定性抽样，具体形式如下所示：

$$x^{\mathrm{MCLU}} = \arg\min_{x_i \in U} \left\{ \max_{w \in N} f(x_i, w) - \max_{w \in N \setminus w^+} f(x_i, w) \right\} \tag{7}$$

其中 w^+ 表示最大的确信度。针对式子(7)可得出结论：x^{MCLU} 值越大，对应样本属于某类别的确信度越大，相反，确信度越小。

2.2.3 基于空间重构的抽样

在支持向量机的分类模型中，不仅可以使用点

到超平面的距离度量不确定性,也可以使用支持向量系数作为度量准则,该准则将多分类问题转化为二分类问题。在重构空间中,支持向量系数对应的样本点用于训练第二个支持向量机模型 $f^{\mathrm{SSC}}(x)$,该模型用于分类 $\alpha>0$ 与 $\alpha=0$。将模型 $f^{\mathrm{SSC}}(x)$ 用于候选样本集中,选出的样本就成为支持向量[13]:

$$x^{\mathrm{SSC}}=\arg\min_{x_i\in U} f^{\mathrm{SSC}}(x_i)>0 \tag{8}$$

只要在候选样本集中选出支持向量集,那么,第二个分类器可以从中随机选择样本进行训练。

2.3　基于后验概率的主动学习算法

在后验概率主动学习算法中,后验概率反映出样本类别的确信度。该算法根据预测所得样本后验概率值的大小,对候选样本集进行排序。通过分析后验概率的变化或每个候选样本的每类分布情况,确定出不确定区域,并从中选择样本,构成训练集。

2.3.1　Kullback-Leibler最大化

该算法通过分析样本最大化后验概率的变化值。使用Kullback-Leibler值度量增加样本前后的分布差异[14]。该方法选择最大后验概率所对应的样本,并且计算KL值。待选样本需满足下式:

$$x^{\mathrm{KL-max}}=\arg\min_{x_i\in U}\left\{\sum_{w\in U}\frac{1}{(u-1)}\times KL\left(p^{+}(w|x)\middle\|p(w|x)\right)p\left(y^{*}=w|x_i\right)\right\} \tag{9}$$

其中对于类别 w,$p^{+}(w|x)$ 指在增加后的训练集 $X^{+}=X\cup\left(x_i,y_i^{*}\right)$ 上的后验概率。目前,针对该方法,提出集成级联的方式对先前所选的样本进行加权,并且样本对当前分类器而言不相关。级联的方式可以缩减分类模型数量,提高分类效率。

2.3.2　Breaking Ties算法

类似于EQB策略,该方法估计候选样本池中每个样本点的后验概率。目前已有模型能够估计后验概率值,例如:人工神经网络,最大似然估计分类器等等。对于支持向量机的决策函数输出,使用sigmoid函数估计样本的概率[15]:

$$p\left(y_i^{*}=w|x_i\right)=\frac{1}{1+\exp\left\{Af(x_i,w)+B\right\}} \tag{10}$$

其中 A,B 都是估计值。只有获得后验概率值,才能评估未标记样本池中的不确定区域。

对于二分类问题而言,Breaking Ties算法专注于后验验概率最小差异的样本。在多类别的问题中,当两个最大的概率值差值很小时,分类器的确定度达到最小[16]。具体形式如式子(11):

$$x^{\mathrm{BT}}=\arg\min_{x_i\in U}\left\{\max_{w\in N}p\left(y_i^{*}=w|x_i\right)-\max_{w\in N\backslash w^{+}}p\left(y_i^{*}=w|x_i\right)\right\} \tag{11}$$

通过上述形式可知,在支持向量机的模型下,BT算法和多层不确定性抽样算法(MCLU)的形式非常类似。

3　主动学习算法分析

主动学习作为一种新的机器学习方法,其主要目标是有效地发现训练数据集中高信息量的样本,并高效地训练模型。与传统的监督方法相比,主动学习具有如下优点:能够很好地处理较大的训练数据集,从中选择有辨别能力的样本点,减少训练数据的数量,减少人工标注成本。委员会查询(QBC),边缘查询(MS),后验概率查询(PP)作为主动学习算法的典型代表,近年来有效地推动了主动学习算法的迅速发展,目前已成为机器学习、模式识别和数据挖掘研究领域中最前瞻和热点的研究方向之一。下面简要论述这些算法之间的联系和区别,以便于读者在这些基本算法的基础上进行扩展和应用。

在QBC算法中,使用标记样本训练多个参数不同的假设模型,并用于预测未标记的样本。因此,QBC算法需要训练一定数量的分类器,在实际应用中,其计算复杂度相当大。为了约束计算量,使用EQB方法简化计算。针对高维数据的情形,AMD算法能够将特征空间划分为子空间,它是EQB算法的变形,不同的分类方法将相同的样本分类在不同的区域中,在计算过程中避免了维数灾难的问题。该算法优点:分类器可以使用多种分类模型以及组合模式,如:神经网络,贝叶斯法则等等。

对于边缘的启发式方法而言,主要针对支持向量机的情形。根据分类模型计算出样本到分类界面的距离选择样本。在MS算法中,仅仅选择距离分类界面最近的样本加入训练集,它是最简单的边缘抽样的方法。而在MCLU算法中,与MS不同之处在于:选择离分类界面最远的两个最可能的样本的距离差值作为评判标准。在混合类别区域中,MCLU能够选择最不确信度的样本,而MS的效果不佳。

在某些情形下,MS和MCLU都会选出冗余的样本,引入多样性准则,剔除相似的样本,减少迭代的次数。常用的多样性准则采用样本间相似度,即样本间的相似度越高,说明样本所反映的数据特点越

一致,则需要剔除该样本,反之,相似度越低。可以使用相似系数值来刻画样本点性质的相似性。具体形式如下所示:

$$\cos\theta=\frac{\left|\phi(x_i)\cdot\phi(x_j)\right|}{\left\|\phi(x_i)\right\|\left\|\phi(x_j)\right\|}=\frac{K(x_i,x_j)}{\sqrt{K(x_i,x_i)\cdot K(x_j,x_j)}} \quad (12)$$

其中 $\phi(\cdot)$ 指非线性函数, $K(\cdot,\cdot)$ 指核函数[17]。在核空间中,余弦值无需考虑映射函数的形式[18]。样本间的夹角越小(余弦值越大),说明样本间越相似,反之亦然。

基于概率的启发式方法依赖于样本的后验概率分布形式,所以该方法的计算速度最快。KL方法的不足之处在于:在迭代优化过程中,它每次只能选择一个样本,增加了迭代的次数。此外,如果分类模型不能提供准确的概率评估值,它依赖于之后的优化评估值。而在BT算法中,其思想类似于EQB,在多分类器中,选择样本两个最大概率的差值作为准则。当两个最大的概率很接近时,分类器的分类确性度最低。

所有主动学习算法能够构造分类器期望的训练集,同时通过选择具有判别信息的数据点正确地划分类别边界。训练后的模型具有很强的泛化能力,从而为主动学习的研究提供了很强的实用基础。

4 主动学习算法中有待进一步研究的问题

主动学习已在信息检索、人脸识别和遥感图像分类[19]等诸多领域中得到了成功应用,但是作为一种新的机器学习算法,其在理论和实践应用方面仍有许多问题亟需进一步研究和探讨:

(1)非监督学习算法[20],例如:聚类算法[21],能够不使用专家知识,自主处理训练样本集的信息。将非监督学习算法与主动学习相结合,提出高效的查询方法。

(2)当输入数据的维数很高时,在高维空间进行查询时会面临“维数灾难”问题,因而需要在预处理阶段寻找高效的降维算法[22],减少查询复杂度。

5 结论

主动学习作为一种新的机器学习方法,其主要目标是从大量的数据集中寻找具有高信息量的样本,从而减少训练样本集的数量,减小计算复杂度,提高分类器泛化能力。本文主要介绍了主动学习的基本思想、最新研究进展及其算法分析,并探讨了其在理论和应用中有待进一步研究的问题。

参考文献:

[1] Hastie T, Tibshirani R, Friedman J.The elements of statistical learning: data mining, inference, and prediction, ser.statistics[M].2nd ed.New York: Springer, 2009.

[2] Boser B E, Guyon I M, Vapnik V N.A training algorithm for optimal margin classifiers[C]//Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, 1992: 144-152.

[3] Haykin S.Neural networks and learning machines[M]. 3rd ed.Cambridge, MA: Prentice-Hall, 2008.

[4] Settles B.Active learning literature survey[R].Univ of Wisconsin-Madison, 2011.

[5] Tuia D, Ratle F, Pacifici F, et al.Active learning methods for remote sensing image classification[J].IEEE Trans on Geosci Remote Sens, 2009, 47(7): 2218-2232.

[6] Copa L, Tuia D, Volpi M, et al.Unbiased query-by-bagging active learning for VHR image classification[C]//Proc SPIE Remote Sens Conf, 2010.

[7] Di W, Crawford M.Multi-view adaptive disagreement based active learning for hyperspectral image classification[C]// IEEE International Geoscience and Remote Sensing Symposium, 2010: 1374-1377.

[8] Muslea I.Active learning with multiple views[J].Journal of Artificial Intelligence Research, 2006, 27: 203-233.

[9] Campbell C, Cristianini N, Smola A J.Query learning with large margin classifiers[C]//Proc Int Conf Mach Learn(ICML), 2000: 111-118.

[10] Schohn G, Cohn D.Less is more: Active learning with support vector machines[C]//Proc 17th ICML, 2000: 839-846.

[11] Demir B, Persello C, Bruzzone L.Batch mode active learning methods for the interactive classification of remote sensing images[J].IEEE Trans on Geosci Remote Sens, 2011, 49(3): 1014-1031.

[12] Vlachos A.A stopping criterion for active learning[J]. Comput Speech Lang, 2008, 22(3): 295-312.

[13] Pasolli E, Melgani F, Bazi Y.Support vector machine active learning through significance space construction[J]. IEEE Geosci Remote Sens Lett, 2011, 8(3): 431-435.

[14] Roy N, McCallum A.Toward optimal active learning through sampling estimation of error reduction[C]// Proc Int Conf Mach Learn(ICML), 2001: 441-448.

[15] Platt J.Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods[M]// Advances in Large Margin Classifiers.Cambridge, MA: MIT Press, 1999.

(下转22页)