主动学习在未标记数据挖掘中的应用

蒋炎岩*

(南京大学 计算机科学与技术系, 南京 210046)

Active Learning: An Effective Approach of Mining Unlabeled Data

Yanyan Jiang*

(Department of Computer Science and Technology, Nanjing University, Nanjing 210046, China)

Abstract: Statistical learning technologies are becoming popular and widely applied in most recent years. To better exploit the information of plenty unlabeled data, active learning, which dynamically acquires label from human oracle, is proposed. In this paper, we surveyed the scenario, theory, and algorithms of active learning. Four categories of active learning algorithms are studied: maximizing informativeness, minimizing expected error, minimizing version space and their hybrids. Both theory analysis and empirical study substantiate the effectiveness of active learning algorithms, as well as open problems and research insights are presented.

Key words: active learning; unlabeled data; data mining; machine learning

摘 要: 近年来,统计学习技术已得到了广泛的应用.为了更好地利用现实问题中大量的未标记实例,主动学习算法实现了动态的实例选择-标记过程.其能够根据部分已知的标记,主动地挑选未标记实例获取标记,并由此获得比静态随机采样监督学习更高的预测准确率.本文对主动学习领域进行了调研,对主动学习的原理和算法进行了较为深入的探讨,研究了最大化信息量,最小化期望误差,最小化解释空间三类算法以及将它们相互组合得到的综合算法.文章还对这几类算法中有代表性的四种进行了实验,籍此证实了主动学习算法的有效性,并根据实验结果对主动学习领域的未来做了展望.

关键词: 主动学习;无标记数据;数据挖掘;机器学习

中图法分类号: TP301 **文献标识码**: A

1 引言

近年来,统计学习(statistical learning)技术已在科学,金融,工业等领域发挥了重要的作用,并逐渐开始在许多问题的求解中取代传统的统计回归方法[1].早期的统计学习技术主要围绕监督学习(supervised learning)展开,学习算法根据已知的样本特征和标记求解模型的参数.在带标记实例容易获取的情形下,监督学习得到了广泛的应用.然而,监督学习通常需要数百或数千个带标记的实例才能得到较为精确的结果[2],因此在诸如语音识别,信息抽取,模式分类等标记获取较为困难的场景下,对监督学习的改进算法应运而生.本文主要研究一类有效利用未标记数据的统计学习方法:主动学习(active learning)方法[2-3].

在标记难以自动获取的情形下,通常需要由领域专家(human oracle)进行人工标记.让专家对大量实例进行

* 作者简介: 蒋炎岩(MG1133013),2011 年获南京大学计算机科学学士学位,现于南京大学计算机软件研究所攻读博士学位.

标记的过程是极不经济的,因此我们期望在标记尽可能少实例的同时达到较高的预测准确率.从直觉上来看,随机地选择实例给专家标记并非最佳策略.主动学习算法正是通过有选择性地从大量未标记实例中挑选"最有价值"的那些实例向专家提问,从而提高预测准确率.通常,主动学习算法被实现为一个与领域专家多阶段交互的程序,在向领域专家提问的过程中求解模型参数.一个典型的主动学习的过程如图 1 所示.

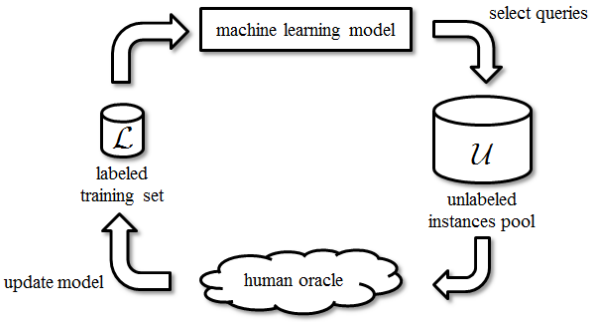

Fig.1 Active learning process

图 1 主动学习过程

主动学习方法现已得到了较为广泛的应用.Google, IBM, Microsoft 等企业已将主动学习技术应用到商业项目中[2].大量文献中的实例研究和理论分析也表明,主动学习技术能以较少的实例数获得较高的预测准确率.本文旨在对主动学习领域进行较为系统的归纳和总结.论文组织如下:第二部分主要介绍主动学习的概念和应用场景,并从理论的角度探索了主动学习的有效性;第三部分结合我们对主动学习算法的理解,对最大化信息量,最小化期望误差,最小化解释空间和综合算法四大类算法进行了归纳和总结;第四部分研究了四种主动学习算法在实际数据集中的表现,以及它们与监督学习算法的对比.最后,结合我们的理论知识和实验结果,我们对实验中观察到的现象做了一定的总结,并对主动学习领域的未来做了展望.

2 主动学习的原理

2.1 主动学习的概念

主动学习的思想方法来源于统计学中的最佳实验设计[4].在实证性学科的验证性实验中,通过合理选择实验参数能够有效地缩短实验周期.类似地,假设在统计学习算法的运行过程中,已标记实例集为$\mathcal{L}$,未标记实例集为$\mathcal{U}$,则主动学习的任务就是设计对实例"价值"的评估函数f,其能够不断在未标记的实例中选出价值最大的实例$u^* = \underset{u\in\mathcal{U}}{\operatorname{argmax}} f(u,\mathcal{L},\mathcal{U})$进行标记,或根据$f$的数值在数据流中选择性地丢弃没有价值的实例.通过启发式地选择被标记的实例,在相同数量标记实例的情形下,主动学习通常能够获得比传统监督学习(即随机选取实例进行标记)更高的预测准确率.

2.2 主动学习算法的应用场景

主动学习算法的应用场景决定了未标记实例集$\mathcal{U}$的特性.据此,主动学习算法可以分为 membership query 算法[5],stream-based 算法[6]和 pool-based 算法[7].

能够在输入空间中任意选择实例进行标记的算法称为 membership query 算法.这类算法的优点是能够在整个输入空间中寻找对学习结果最有利的实例,因此具有最佳的灵活性,实现也最为简便.然而,由于 membership query 算法可能产生大量合法但无意义的输入,此类算法不适合领域专家人工给出标记的应用场景.

与 membership query 相对的是 stream-based 算法.Stream-based 算法适用于未标记实例可以无限获取的情形.其不断通过采样或生成得到新的无标记实例,并根据$\mathcal{L}$的信息决定丢弃新的实例,或是将生成的实例交给领域专家标记.此类算法在语音识别,信息获取等领域得到了广泛的应用.

Pool-based 算法的应用场景则是上述两种算法的折中,适用于未标记集合$\mathcal{U}$为确定有穷集的情形,即算法需

要在一些给定的实例中进行挑选.特别的,当未标记集合是整个输入空间时,pool-based 算法退化成为 membership query 算法,而基于数值的 pool-based 算法也可以通过设置阈值的方式修改为 stream-based 算法.因此,pool-based 算法是主动学习领域中研究最广泛的一类算法.

2.3 主动学习的有效性

主动学习在早期主要是经验主义的启发式算法.随着研究的深入,其理论本质也被不断发掘,目前的理论研究主要基于样本复杂度展开[2].考虑如下例的统计模型中,所有实例均是一维直线上均匀分布的数据点,数据的标记完全由参数θ确定:

$$g(x,\theta)=\begin{cases}1 & x>\theta\\ -1 & \text{otherwise.}\end{cases}$$

根据 PAC 学习模型[8],统计学习算法为了达到对参数θ的ε准确率估计,需要获得$O(1/\varepsilon)$个随机的样本.考虑 membership query 主动学习算法,算法每次会根据已经得到的结果选择确定度最低的实例进行标记,即对处在不确定区间中位点的实例进行标记.该二分查找过程能够通过$O(\log(1/\varepsilon))$次询问达到ε准确率的估计,显著优于传统统计学习算法.

上述示例从直观上揭示了主动学习在特定模型上的有效性.文献[9-10]从理论的角度研究了主动学习的有效性,其结论归纳如下:在最坏情况下,主动学习达到ε准确率的样本数将达到$O(1/\varepsilon)$,但在实际问题中,主动学习算法几乎总是能够在渐近的意义下优于随机取样.

除理论分析外,实例研究也为主动学习的有效性提供了有力的证据.文献[11]显示使用主动学习进行自然语言处理的研究者的完成度明显高于其他方法.此外,文献[2]对实例研究的总结也显示,绝大部分已发表的文献均显示主动学习在众多实际问题中具有良好的表现.

3 主动学习的算法

主动学习算法设计的要素在于如何构造恰当的实例评估函数,使其能根据已标记实例集$\mathcal{L}$中的信息从未标记实例集$\mathcal{U}$中选择一个最佳的实例x^*,对其标记后能够最大程度地提高模型的预测准确率.这一基本框架适用于所有三种类型的主动学习算法[2].然而由于$\mathcal{U}$中实例标记的不确定性,寻找恰当实例的过程也没有绝对的黄金准则.通常,主动学习算法采用启发式的贪心策略,即每次从未标记实例中选择某一方面属性最大(或最小)的实例进行标记.在这一部分中,我们将对主流的主动学习算法按其所属类型的不同分别展开讨论.

3.1 选择最具信息量的实例

最大化信息量算法的基本思想非常直观:我们总是选择那些会对当前模型带来较大潜在改变(即蕴含较多信息)的未标记实例进行标记.

一类最大化信息量的算法是选择那些"最不能确定"的实例,称为 uncertainty sampling 算法[7].其基本原理在于绝大多数统计学习算法得到模型后,在给出实例分类的同时也能够给出一个对该分类结果的置信度.从直观上看,对高置信度的实例进行标记的结果将是基本确定的,即不会对模型产生任何影响;反之,选择那些置信度较低的实例进行标记,应当能够有效地提高模型的辨别能力.

根据具体的统计学习算法能够设计出不同的 uncertainty sampling 主动学习算法.例如基于概率的二元分类算法可以选取后验概率最接近$1/2$的实例进行标记[7]:

$$x^*=\underset{x\in\mathcal{U}}{\operatorname{argmax}}\left|\Pr[x=1|\mathcal{L}]-\frac{1}{2}\right|$$

上述算法可以推广到多元分类算法.由于一个实例有多种可能的标记,我们可以引入实例被标为每一种标记的概率来度量实例的不确定性.Keast confident 算法以每一个实例可能性最大的标记作为其代表标记,并从中选择不确定性最高的:

$$x^*_{LC}=\underset{x\in\mathcal{U}}{\operatorname{argmax}}\left(1-\Pr\left[\underset{y}{\operatorname{argmax}}\Pr[y|x]\middle|\mathcal{L}\cup x\right]\right)$$

作为 least confident 算法的扩充,margin sampling 算法则以每一个实例最大和次大的标记置信度的差作为度量的标准:

$$x_M^* = \underset{x \in \mathcal{U}}{\operatorname{argmin}} \left(\Pr[\hat{y_1}|x] - \Pr[\hat{y_2}|x] \right)$$

如果考虑所有标记的可能,则可以使用信息熵算法:

$$x_H^* = \underset{x \in \mathcal{U}}{\operatorname{argmax}} - \sum_i \Pr[y_i|x] \log \Pr[y_i|x]$$

文献[12-13]中对上述算法进行了大量的实例研究.研究结果表明,上述主动学习算法在实际应用中均优于随机选择实例进行标记.

此外,通过发掘统计学习算法的本质并适当地定义未标记实例的置信度,uncertainty sampling 思想还可以向非概率的统计学习模型推广.文献[14-15]分别提出了基于决策树和支持向量机的 uncertainty sampling 算法.

另一类最大化信息量的算法是最大化模型改变量的期望.这类算法的原理是寻找那些对模型潜在影响最大的未标记实例进行标记.令$\nabla \ell(\mathcal{L})$表示模型关于参数和已标记实例的梯度,我们最小化模型改变量的期望[16]:

$$x^* = \underset{x \in \mathcal{U}}{\operatorname{argmax}} \sum_i \Pr[y_i|x] \left\| \nabla \ell(\mathcal{L} \cup x) \right\|$$

在实例研究[13]中,最大化期望模型改变算法也有较好的预测准确率.但这类算法的最大弱点在于其效率较低,还易受到原始数据缩放的影响,因此没有得到广泛的应用.

选择最具信息量实例的一类算法通常能够充分利用学习算法的特性,大多拥有简明,高效等特点,因此被广泛使用[2].同时,这类算法也面临一些固有的缺陷.首先,由于总是有偏好地启发式挑选具有特定性质的实例进行标记,在特定数据集上的预测准确率会大幅下降;此外,不符合原始分布的 outlier 有可能会对这类算法的预测模型造成较大的影响.

3.2 选择最小期望误差的实例

最大化信息量算法的本质是根据当前已标记实例的信息贪心地选择最有信息量的实例,而选择最小期望误差[17](expected error reduction)算法则从全局的角度进行考虑,总是标记那些具有最小期望误差(风险)的实例:

$$x^* = \underset{x \in \mathcal{U}}{\operatorname{argmax}} \sum_i \Pr[y_i|x] \left(\sum_{u=1}^{U} \left(1 - \Pr[\hat{y}|x^{(u)} \wedge (\mathcal{L} \cup x)] \right) \right)$$

这类算法的基本流程是计算每一个实例加入标记集合后整体误差的期望.这类算法在实例研究中体现出较高的准确率和稳定性.然而,由于其时间开销较大,即便采用朴素的统计回归模型也需要$O(|\mathcal{U}||\mathcal{L}|)$次计算,这类算法并未得到十分广泛的应用.虽然在特定的问题上可以较为高效地应用此类算法,但整体而言此类算法的执行效率仍比 uncertainty sampling 等算法要低得多.

3.3 选择最小化解释空间的实例

对于一个已标记的实例集$\mathcal{L}$,与其一致的所有统计学习模型称为$\mathcal{L}$的解释空间(version space).解释空间越大,就意味着我们有越多的模型可以选择.当解释空间只有一个点时,统计学习的模型也可以唯一确定.因此,另一类主动学习算法总是选择那些能够最大程度缩小解释空间的实例进行标记.

这一思想引出了一类重要的算法:query-by-committee(QBC)[18].其基本思想是用不同的统计学习算法在同一已标记的实例集$\mathcal{L}$上进行训练,得到一个模型的集合$\mathcal{C} = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$.我们用$\mathcal{C}$中的每一个模型对$\mathcal{U}$中的实例进行标记,然后选择那些不同模型预测结果最不一致的实例进行标记.因此要实现 QBC 算法,就必须:

1. 选择适当的算法构造出模型集合$\mathcal{C}$,
2. 量化$\mathcal{C}$中模型对实例预测的不一致性.

如果将统计学习算法看作是在解释空间中求解一个最佳的模型,则$\mathcal{C}$的挑选就是通过启发式手段缩小可行解释空间的策略.文献[16,18-19]等提出了大量启发式的算法并做了相应的理论分析.

在选择实例的阶段,我们同样可以采用信息熵算法,最大化不同模型给出标记集合的信息熵[20]:

$$x_{VE}^{*}=\underset{x\in\mathcal{U}}{\operatorname{argmax}}-\sum_{i}\frac{V(y_i)}{C}\log\frac{V(y_i)}{C}$$

尽管学界没有对$\mathcal{C}$大小的选取达成共识,但实例研究表明,即便在$|\mathcal{C}|$很小(2 或 3)的情形下,QBC 算法也能获得较高的准确率[2,15].QBC 算法简单易行,实例研究又说明了其高效性,因此 QBC 算法在主动学习领域也得到了广泛的应用.

最小化解释空间这一思想也可以直接加以利用.通常精确地计算出最小化解释空间的实例是非常困难的,因此我们一般使用启发式算法或近似算法.文献[15]中给出了一种基于支持向量机的启发式最小化解释空间算法:对于每一个未标记实例,我们都计算出将它标记为-1 时的分平面距离m^-和将它标记为 1 时的分平面距离m^+,从最小化解释空间的意义上,我们选择到二者到分平面相对距离最大的实例进行标记:

$$x^{*}=\underset{x\in\mathcal{U}}{\operatorname{argmax}}\min\left\{\frac{m^-}{m^+},\frac{m^+}{m^-}\right\}$$

3.4 综合算法

上述三种类型的主动学习算法的思想是相互独立的,而且,不同算法在不同应用场景下的准确率往往相差很大.因此,一个自然的思路是将不同类型的算法组合起来.文献[21]提供了一种聚合多种算法的思路:在有多个主动学习算法的情形下,引入算法的评分机制,根据主动学习过程的推进,预测正确的那些算法将会得到奖赏,而预测错误的那些则会获得惩罚,籍此选出对当前实例集最佳的主动学习算法.

另一种增加主动学习算法稳定性的策略是在选择实例的过程中引入对模型"代表性"的因素[13].对于任意基础主动学习算法的评估函数f(例如 uncertainly sample 算法),修正优化目标函数为:

$$x^{*}=\underset{x\in\mathcal{U}}{\operatorname{argmax}}\,f(x,\mathcal{L},\mathcal{U})\times\left(\frac{1}{U}\sum_{u=1}^{U}\operatorname{sim}(x,x^{(u)})\right)^{\beta}$$

式中第二项乘子为每一个实例相对于其他实例的相似性.这一算法在实际应用中能有效地增强传统主动学习算法对 outlier 的抵抗能力.

综合算法从本质而言是一种多目标优化.多目标优化中的多个目标(如信息量,代表性等)同时达到最优的可能性是极小的,因此在各个目标之间权衡即成为了主动学习领域研究的热点之一,例如文献[22]提出了同时考虑信息量和代表性的多目标优化算法.

4 实例研究

4.1 实验环境

为了公平地对比不同类型算法的优劣,我们固定统计学习算法为支持向量机,并在实验中统一采用 LIBSVM[23]运行库,所有实验均采用默认参数在 LIBSVM 的默认流程下运行.我们选定 Digit1 和 USPS 两个数据集[1]进行对比实验.实验的软件环境为 Ubuntu Linux 下的 Python 2.6, GNU Octave 3.2.3 和 LIBSVM3.1.我们共选用了五种不同类型的算法作为对比,现描述如下.

4.1.1 随机采样(RND)

我们将随机采样(RND)算法的输出结果作为实验的基准,严格意义上来说,它并不属于主动学习算法.实验过程中,RND 算法每次从未标记的实例集$\mathcal{U}$中随机地选择一个x进行标记,然后对集合$\mathcal{L}\cup x$调用 LIBSVM 库,并对学习模型的精确度进行测量.

4.1.2 Uncertainty Sampling(US)

我们选取了一种典型的支持向量机算法[15]作为 uncertainty sampling 的代表:总是选取距离分平面最近的

[1] 由于 MVS 和 QUIRE 算法的运行时间较长,完成 text 数据集实验需要超过 100 小时,故没有在 text 数据集上完成实验.

实例作为不确定性最大的实例进行标记.在若干篇文献中都对它进行了研究.文献[7]中的研究表明,它不仅是一种典型的 uncertainty sampling 算法,还是一种最小化解释空间的启发式算法.由于其原理简单,时间复杂度低等特性,在现实中得到了广泛的应用.在完成这一实验的过程中,我们修改了 LIBSVM 的实现,使其在输出每一个实例预测标记的同时还能输出每一个实例距离分平面的距离.

4.1.3 Query-By-Committee(QBC)

我们选择了文献[18]中的 query-by-bagging 算法作为 query-by-committee 的代表.在我们的算法实现中,committee 大小$|\mathcal{C}|=3$,$\mathcal{C}$中的每一个成员都采用相同的支持向量机算法,但输入的标记实例不同.在主动学习的每一个阶段中,我们对$\mathcal{L}$进行三次随机采样,每次随机选出$2|\mathcal{L}|/3$个元素,将三次采样得到的实例送给$\mathcal{C}$中的三个 committee 分别学习,并对$\mathcal{U}$中所有实例投票.算法总是选择投票结果分歧最大的实例进行标记,当有多个满足条件的实例时,则随机选择一个.

4.1.4 Minimizing Version Space(MVS)

我们还实现了文献[15]中最小化解释空间的 min-max 启发式算法:

$$x^* = \underset{x\in\mathcal{U}}{\operatorname{argmax}} \min\left\{\frac{m^-}{m^+}, \frac{m^+}{m^-}\right\}$$

由于计算m^+和m^-需要重新运行支持向量机算法,其时间效率较低.因此,我们对其进行了一定程度的近似:我们不再寻找全局最优解,而是从$\mathcal{U}$中随机地采样 20 个实例$\mathcal{U}'$作为$\mathcal{U}$的代表,并从$\mathcal{U}'$中选择数值最大的实例进行标记,即$x^* = \underset{x\in\mathcal{U}'\subset U}{\operatorname{argmax}} \min\left\{m^-/m^+, m^+/m^-\right\}$.这一算法将以较大概率返回排名前 5%以内的实例.

4.1.5 Querying Informative and Representative Examples(QUIRE)

最后,我们选择了 QUIRE 作为综合算法的代表,来自于文献[22].QUIRE 的运行机制与传统的主动学习算法十分类似:其对于每一个未标记实例都计算一个评估函数,并从中选取最佳的实例进行标记.QUIRE 算法的主要特点是它并非单方面启发式地选择不确定性最大或解释空间最小的实例.来自文献中的分析表明,其采用的评估函数

$$f(x,\mathcal{L},\mathcal{U}) = \min_{y_u\in\{-1,+1\}^{|\mathcal{U}|}} \max_{y_s\in\{-1,+1\}} y^T\left([\kappa(x_i,x_j)]_{n\times n}+\lambda I\right)^{-1} y$$

能够近似地写成两项之和,其中第一项代表了根据已标记实例推算x标记的置信度,即信息量;第二项则代表了x与所有未标记数据的契合程度,即代表性.这说明 QUIRE 算法是一种兼顾了信息量和代表性的综合算法.我们使用了作者提供的代码,并且在兼容的环境下完成了实验.

4.2 实验结果

我们在 Digit1 和 USPS 两个数据集上运行了 US, QBC, MVS, QUIRE 和 RND 五种算法,每一种算法都会在一个数据集上用不同的初始数据(10 个已标记的实例)运行多次,测试结果如图 2 所示.图中横坐标为标记实例的个数,纵坐标则为模型在整个数据集上的预测准确率.

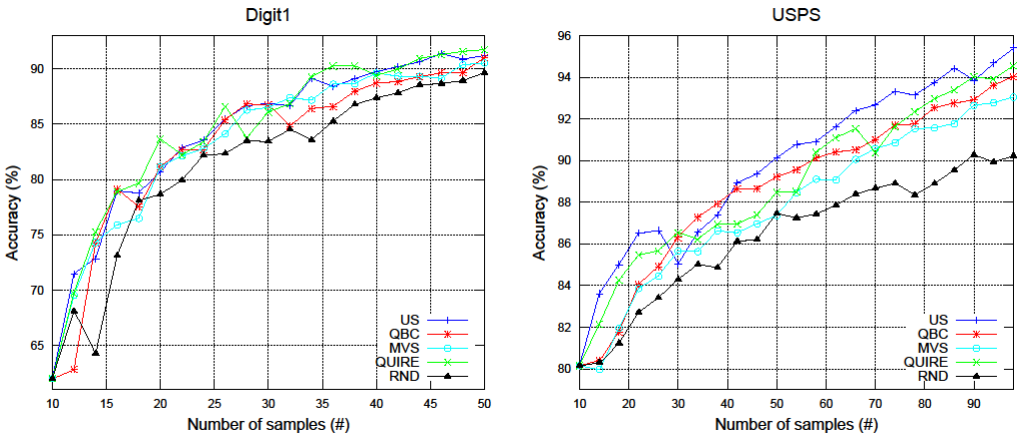

Fig.2 Empirical study results
图 2 实验结果

我们统计了各个算法的平均准确率,如表 1 所示.表中方框表示数据集中准确率最高的数值,下划线则表示数据集中准确率次高的数值.

	US	QBC	MVS	QUIRE	RND
Accuracy(Digit1)	86.585	85.691	85.557	87.219	83.397
Accuracy(USPS)	90.985	89.811	88.913	90.175	87.402

Table.1 Average accuracy
表 1 平均准确率

为了验证相关结果的置信度,我们对实验数据进行了 Welch t-检验,对比在不同数据集上各个算法的预测准确率是否有显著差异,结果如表 2.表中数字为 t 值,其中粗体加框的数值表示置信度超过 95%,加框的数值表示置信度超过 90%,有下划线数字则表示置信度超过 80%.

	Digit1					USPS				
	US	QBC	MVS	QUIRE	RND	US	QBC	MVS	QUIRE	RND
US	0.00	0.86	0.93	-0.62	**2.14**	0.00	**1.99**	**3.77**	1.63	**6.67**
QBC	-0.86	0.00	0.13	-1.58	1.59	-1.99	0.00	1.61	-0.72	**4.42**
MVS	-0.93	-0.13	0.00	-1.62	1.45	-3.77	-1.61	0.00	-2.75	**3.01**
QUIRE	0.62	1.58	1.62	0.00	**2.67**	-1.63	0.72	**2.75**	0.00	**6.25**
RND	-2.14	-1.59	-1.45	-2.67	0.00	-6.67	-4.42	-3.01	-6.25	0.00

Table.2 t-test results
表 2 t-检验结果

4.3 结果分析

根据 t-检验的结果,我们作出各个算法在不同数据集上比较的偏序,其中实线边(a,b)代表算法a以超过 90%的置信度优于算法b,虚线边(a,b)则代表算法a以 80%至 90%的置信度优于算法b,如图 3 所示.

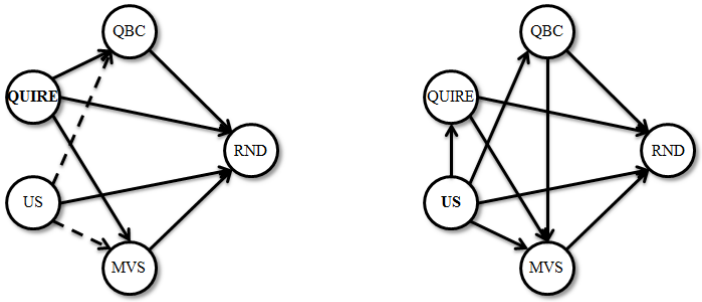

Fig.3 Comparison result between algorithms
图 3 不同算法在数据集上的相对表现

实验结果表明,US, QBC, MVS 和 QUIRE 四种算法均显著优于随机选取实例进行标记的策略,其中 QUIRE 算法在数据集 Digit1 上的效果最好,而 US 在数据集 USPS 上的效果最佳.为了达到相同的预测准确率,RND 需要使用数倍于主动学习算法的实例,这说明主动学习算法确能起到增加预测准确率的作用.

从各个算法两两之间的比较可以发现,对于不同的数据集,算法与算法之间并没有绝对的优劣.这表明了主动学习算法对数据集敏感的特性.虽然主动学习领域中涌现了大量的新算法,并且在大量的数据集上证实了它们的优势,但这仍然不足以证明这些新算法能够完全取代已有的算法.就其根本而言,这些主动学习算法均是为了达到特定目的的启发式方法,而关于主动学习有效性和性能的理论在领域中仍然属于开放问题.

除此之外,时间复杂度也是制约新算法广泛应用的瓶颈.在我们的实现中,US 最为简单,时间复杂度与 LIBSVM 相同,单次运行在 Digit1 和 USPS 数据集上只需若干秒;而 QUIRE 则需要 kernel matrix 作为输入,在相同环境下需要运行则需 3 分钟以上.因此,综合考虑准确率,稳定性,效率等因素,在现实应用中选择适当的算法就变得更加困难.

5 结束语

主动学习算法的基本思想是通过部分带标记实例的信息,反复在大量未标记实例中选取最佳的实例进行标记,以获得较为精确的统计学习模型.本文调研了对最大化信息量,最小化期望误差,最小化解释空间和综合算法四大类主动学习算法,并选择性地进行了实现.理论和实践均表明,主动学习是挖掘未标记数据的有效途径.

如今,为进一步增加主动学习算法的预测准确率,现有的研究已经不仅局限于主动学习算法本身,围绕多个不同的角度展开.例如利用数据集的一些本征特性(如文本或图像),以及与其他统计学习技术的融合,如引入其他挖掘非标记数据的思路(如半监督学习)等算法被相继提出.

尽管主动学习领域的研究不断取得进展,但却没有广泛地在实际中应用.展望未来,我们将其主要原因和挑战总结如下:

从主动学习主体的角度,目前还没有完备的主动学习理论体系,以指导何种主动学习算法适合在怎样的场景下应用.从实验的结果中我们可以看到,主动学习算法对数据集的敏感程度是很高的,在实际问题中滥用主动学习算法甚至可能会降低预测准确率.为主动学习建立完整的理论是一个十分困难的开放问题,目前仅有少部分文献做了一些初步的尝试.

从主动学习对象客体的角度看,主动学习适用于标记获取代价较高的情形.通常,实际的主动学习应用都包含与领域专家交互的过程.由主动学习算法的特性可知,其可能会挑选一些”不合情理”的实例交给领域专家标记.此时,领域专家可能以一定几率给出错误的标记,从而对预测准确率产生负面的影响.如何在准确率和健壮性之间进行权衡也是该领域的挑战之一.根据我们的知识,目前还没有文章针对这一方面的内容进行过研究.

综上所述,主动学习算法的研究已经日趋成熟,并逐渐可以走向实用,但领域中仍然包含许多较为本质的开放问题等待解决.

致谢 在此对所有给予我帮助的人致以诚挚的谢意.此外,特别感谢 LIBSVM 作者的开源精神,LIBSVM 的源代码在实验中起到了关键的作用.

References:

[1] Hastie, T. and Tibshirani, R. and Friedman, J. and Franklin, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, 2005.

[2] Settles, B.: Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.

[3] Hanneke, S.: A Bound on the Label Complexity of Agnostic Active Learning. In Proceedings of the 24th international conference on Machine learning, 2007.

[4] Federov, V.: Theory of Optimal Experiments. Academic Press, 1972.

[5] Angluin, D.: Queries and Concept Learning. Machine Learning, 2:319-342, 1988.

[6] Cohn, D. and Atlas, L. and Ladner, R.: Improving Generalization with Active Learning. Machine Learning, 15(2):201-221, 1994.

[7] Lewis, D. and Gale, W.: A sequential Algorithm for Training Text Classifiers. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, ACM/Springer, 1994.

[8] Valiant, L.G.: A theory of the Learnable. Communications of the ACM, 27(11): 1134–1142, 1984.

[9] Dasgupta, S.: Analysis of a Greedy Active Learning Strategy. In Advances in Neural Information Processing Systems (NIPS), 2004.

[10] Balcan, M.F., Hanneke.S and Wortman, J.: The True Sample Complexity of Active Learning. In Proceedings of the Conference on Learning Theory (COLT), 2008.

[11] Tomanek, K. and Olsson, F.: A Web Survey on the Use of Active Learning to Support Annotation of Text Data. In Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing, 2009.

[12] Körner. C. and Wrobel, S.: Multi-class Ensemble-based Active Learning. In Proceedings of the European Conference on Machine Learning (ECML), 2006.

[13] Settles, B. and Craven M.: An Analysis of Active Learning Strategies for Sequence Labeling Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2008.

[14] Lewis, D. and Catlett, J.: Heterogeneous Uncertainty Sampling for Supervised Learning. In Proceedings of the International Conference on Machine Learning (ICML), 1994.

[15] Tong, S. and Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification. In Proceedings of the International Conference on Machine Learning (ICML), 2000.

[16] Abe, N. and Mamitsuka, H.: Query Learning Strategies Using Boosting and Bagging. In Proceedings of the International Conference on Machine Learning (ICML), 1998.

[17] Roy, N. and McCallum, A.: Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In Proceedings of the International Conference on Machine Learning (ICML), 2001.

[18] Melville, P. and Mooney, R.: Diverse Ensembles for Active Learning. In Proceedings of the International Conference on Machine Learning (ICML), 2004.

[19] Dagan, I. and Engelson, S.: Committee-based Sampling for Training Probabilistic Classifiers. In Proceedings of the International Conference on Machine Learning (ICML), 1995.

[20] Settles, B., Craven, M. and Ray, S.: Multiple-instance Active Learning. In Advances in Neural Information Processing Systems (NIPS), 2008.

[21] Baram, Y. and El-Yaniv, R. and Luz, K.: Online choice of active learning algorithms. In Proceedings of the International Conference on Machine Learning (ICML), 2003.

[22] Huang, S.J., Jin, R. and Zhou, Z.H.: Active Learning by Querying Informative and Representative Examples. In Advances in Neural Information Processing Systems (NIPS), 2010.

[23] Chang, C.C. and Lin, C.J.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.