基于历史分类加权和分级竞争采样的多视角主动学习

姚拓中,安 鹏,宋加涛
(宁波工程学院电子与信息工程学院,浙江宁波 315016)

摘 要: 多视角主动学习是一种相比于传统主动学习能够取得更大程度版本空间缩减的技术,已被应用于多种类型的大数据分析中.本文针对现有的多视角主动学习算法在分类假设生成和采样策略中存在的不足分别提出了相应的改进方案.本文将 Boosting 思想应用到多视角主动学习框架中,通过将历史上各次查询得到的分类假设进行加权式投票来实现每次查询后分类假设的强化;与此同时,还提出了一种自适应的分级竞争采样策略,当分类争议样本规模较大时通过无监督谱聚类获得上述样本的空间分布描述,并在各个聚类中结合样本的分类不确定度和冗余度信息通过二次规划求解以获得可靠的批处理采样.为了证明上述改进的有效性,本文将多视角主动学习应用到图像分类领域中,并通过基于不同图像特征的视角来分别生成相应的分类假设.实验表明,本文提出的两点改进策略不仅均有助于提升多视角主动学习的性能,而且基于上述不同视角随机组合的多视角主动学习方法相比于经典的单视角主动学习算法能够更快地实现收敛并达到较高的场景分类准确性.

关键词: 多视角主动学习;分类器集成强化;分级竞争采样;图像分类
中图分类号: TN911.73 **文献标识码:** A **文章编号:** 0372-2112 (2017)01-0046-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.01.007

Multi-View Active Learning Based on Weighted Hypothesis Boosting and Hierarchical Competition Sampling

YAO Tuo-zhong, AN Peng, SONG Jia-tao
(*School of Electronics and Information Engineering, Ningbo University of Technology, Ningbo, Zhejiang* 315016, *China*)

Abstract: Multi-view active learning is a technique which can realize more significant reduction on version space than traditional active learning and has been used in large-scale data analysis. This paper proposes two improvements in both hypothesis generation and sampling strategy. We integrate boosting-like idea into the active learning framework which uses the weighted voting of all hypothetic outputs from the past queries. Furthermore, a novel adaptive hierarchical competition sampling is presented. In this sampling strategy, if the number of the contention samples is large, an unsupervised spectral clustering is activated to obtain the coarse distribution of these contention samples in the feature space and then both the classification uncertainty and redundancy measures are considered in each cluster to query the unlabeled samples in batch mode by solving quadratic programming. We apply multi-view active learning in image classification in order to prove the effectiveness of the improvements and different image features are used as views to generate the corresponding hypothesis. The experiments demonstrate that our two proposals can both efficiently improve the performance of the multi-view active learning and the random combination of these views can also obtain faster convergence and better classification accuracy than state-of-the-art single-view active learning algorithms.

Key words: multi-view active learning; weighted hypothesis boosting; hierarchical competition sampling; image classification

---

收稿日期:2015-05-20;修回日期:2016-04-28;责任编辑:李勇锋
基金项目:浙江省自然科学基金(No. LY15F020011, No. LQ15F020004);浙江省公益技术应用研究计划(No. 2016C33255);宁波市自然科学基金(No. 2015A610132)

1 引言

主动学习理论最早由 Simon[1] 提出,是一种能够从大量的未标记样本中挑选出一部分具有较高信息量且对分类器性能提升有帮助的样本进行人工标记的技术,它改变了传统只对已标记样本进行被动学习的形式,能够有效降低特征空间中的样本复杂度. 根据 PAC 学习理论,在理想情况下为了获取期望分类误差小于 ε 的分类器,主动学习的样本复杂度为 $O(\log(1/\varepsilon))$,相比传统被动学习的样本复杂度 $O(1/\varepsilon)$ 可以获得指数形式地减少,因而尤其适合于大数据的分析. 主动学习理论在近十几年里得到了不断完善和发展,并于近些年已开始在图像检索、人脸识别、行为分析,目标跟踪和场景重建等诸多领域体现出广阔的应用潜力.

2 相关研究

选择性采样(Selective Sampling)是主动学习算法的关键. 根据采样方式的不同,主动学习方法大致可分为以下两大类:基于“池”的方法和基于“流”的方法[2,3]. 基于“流”的方法由于不能对未标注样本进行逐一比较,需要根据样本的评价指标人工设定相应的阈值,因而限制了其应用和发展. 基于“池”的方法则是目前最流行的主动学习方法,其根据产生的分类假设数目不同,同样可分为基于单一假设方法[4,5] 和基于委员会的方法[6,7] 两大类. 后者由两个或两个以上的分类假设组成一个委员会并对那些能够最大程度缩减版本空间的样本进行采样,通常具有比前者更好的采样性能. 然而,后者存在的一大问题是委员会中的假设构成相对单一以及多样性存在不足. 值得关注的另一种基于委员会的方法是以 Co-Testing 为代表的[8] 多视角主动学习方法. 该类方法的优势在于结合多个对于样本标签相互条件独立的视角可以获得最大化改进其中一个视角的样本查询性能,并且在采样时能够查询那些至少造成一个视角所对应的分类假设错误的样本. 从理论上,与传统基于委员会的采样方法相比其能够实现更大程度的版本空间缩减.

Muslea 等在 Co-Training 算法[9] 的基础上于 2000 年提出了第一种多视角主动学习算法 Co-Testing,其成为了之后多视角主动学习的基础. 最近十余年里,多视角主动学习获得了快速发展:在 2002 年,Muslea 等通过结合 Co-Testing 和基于多视角学习的 Co-EM 算法提出了 Co-EMT 算法[10]. 该算法通过 Co-Testing 算法对未标注样本进行采样的同时利用多视角学习来改进生成的分类假设,并被成功应用到文本分类中;此后,Mulsea 等提出了 Aggressive Co-Testing 算法,其在能够较好地学习样本类别的强视角基础上,利用对样本描述不具有高区分度的弱视角来同步改进采样和分类假设生成[11];Dima 等将包括 Co-Testing 在内的多种主动学习技术应用到移动机器人的野外自主导航中,通过野外复杂场景感知的实验对比进一步验证了多视角主动学习能够在相同的查询次数中更好地提升自身的分类性能[12];Wei 等则提出了一种基于协同正则化的多视角主动学习框架,通过结合多视角一致性和局部邻近假设实现采样的优化[13].

在国内,近些年里多视角主动学习的相关研究也开始兴起:Cheng 等结合 Co-Testing 和 SVM,提出了 Co-SVM 算法并在实验中取得了比单视角主动学习更优的图像检索效果[14]; Wang 等提出了一种结合半监督学习的多视角主动学习方法以获取更好的样本复杂度下降[15],并从理论上深入分析了多视角主动学习在非理想条件下的样本复杂度,证明了在无边界 Tsybakov 噪声条件下单视角主动学习的样本复杂度最多仅能实现多项式形式的降低,而多视角主动学习的样本复杂度则可以实现指数形式的下降[16]; Zhang 等通过结合视角内和视角间的不确定度有效改进了多类别图像的分类性能[17];Yang 等则提出了一种基于批处理模式的多视角主动学习方法,通过三种不同特征作为视角来实现图像检索中高置信度未标注图像的采样[18].

然而,现有的绝大多数多视角主动学习方法在采样策略上均存在以下两个明显缺点:(1)每次查询时仅对一个样本进行采样,不具有批处理采样的能力,而批处理采样对于提高主动学习的收敛性起到重要的作用,并且目前已经在很多单视角主动学习中得到应用,而在多视角主动学习领域则应用很少[14,18];(2)仅根据样本的分类置信度来挑选样本,并未考虑未标注样本在特征空间中的分布结构,从而容易导致采样偏置的发生. 此外,相对于现有的研究基本聚焦于采样策略,而对分类假设生成策略的研究则非常少,而分类假设生成的可靠性决定了样本分类标记的准确性,该领域的研究也同样非常具有价值.

3 算法概述

如何生成可靠的分类假设以及高效的选择性采样是多视角主动学习最为关键的技术难题. 为此,本文在经典 Co-Testing 算法的基础上提出了一种新的多视角主动学习方法,简称 $BSVM_{AL}^{MV+QP}$. 从算法设计思路上,$BSVM_{AL}^{MV+QP}$ 算法类似于 Co-EMT,在分类假设生成和采样中分别提出了如下的改进策略:

(1)将 Boosting 思想应用到多视角主动学习的整体框架中,通过类似于递归式弱分类器强化的方式将历史上各次查询得到的分类假设进行加权式投票,从而实现每次查询后分类假设的强化;

(2)提出了一种自适应的分级竞争采样策略. 当分类争议样本数目较大时,通过无监督谱聚类获得上述样本在特征空间中的初步结构分布描述,并在各个聚类中有效结合样本的分类不确定度和冗余度信息来实现高效的批处理采样.

4 多视角主动学习算法 $\mathrm{BSVM}_{\mathrm{AL}}^{\mathrm{MV+QP}}$

4.1 分类假设生成策略

如果主动学习在每次查询中都能够采样到足够数目改进分类假设的高信息量样本,那么随着查询次数的增加,被错误分类的未标注样本数目将不断减少,而这点与 Boosting 技术在弱分类器强化程中的情形非常类似. 为此,本项目拟将经典的 Adaboost 算法思想应用在多视角主动学习框架中,从而逐步强化每次查询后生成的分类假设.

图 1 给出了基于 Adaboost 架构的分类假设强化的基本流程. 本项目以基本分类器为 SVM 的多视角分类器取代了传统 Adaboost 算法中的单视角弱分类器,各次查询后的多视角分类器可以看作 Adaboost 递归过程中的一个个弱分类器. 多视角分类器的分类假设 $h^i(x)$ 由 $v_1,v_2,\cdots,v_n$ 上述 n 个权重为 ω_i 的 SVM 基本分类器输出的分类假设通过权重 $\omega_1,\omega_2,\cdots,\omega_n$ 线性加权获得. 与传统 Query by Boosting 等算法仅针对基本分类器本身进行强化不同,在每次样本查询时,本算法将对各个已标注样本的分类权重进行更新,并将历史上各次递归查询时输出的分类假设通过线性加权的方式得到当前递归查询时的分类假设 $H^i(x)$.

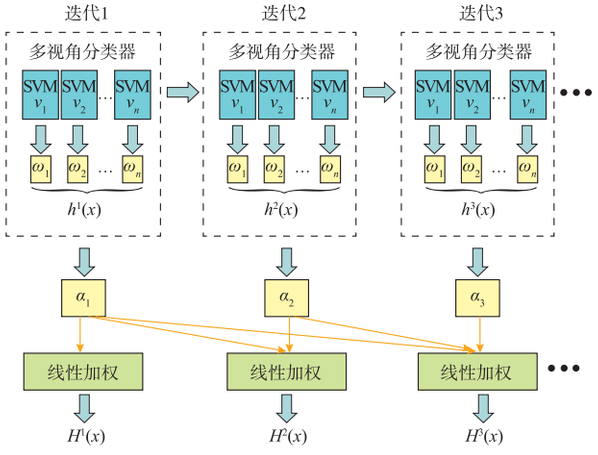


图1 基于Adaboost架构的分类假设强化

通过 Adaboost 思想实现分类假设强化的核心步骤如下:

(a)在第 $t=1,2,\cdots,T$ 次递归时,借鉴文献[19]中基于加权投票的思想得到样本 x_j 基于多个视角加权的初始分类假设:

$$h^t(x_j)=\sum_{f\in\{f^t_1,f^t_2,\cdots,f^t_n\}}\omega^t_i f^t_i(x_j) \tag{1}$$

在式(1)中,权重 ω^t_i 代表了视角 i 对于分类的贡献程度,其大小由该视角的分类误差 ε^t_i 所决定;$f^t_i(x_j)$ 代表第 i 个视角 SVM 所输出的满足 Karush-Kuhn-Tucker (KKT)条件最优解的样本 x_j 分类置信度. 在这里,我们定义视角 i 的分类误差 ε^t_i 如下:

$$\varepsilon^t_i=\frac{1}{\left(\sum\limits_{x\in L,y=1}f^t_i(x)-\sum\limits_{x\in L,y=-1}f^t_i(x)\right)} \tag{2}$$

在式(2)中,$\sum\limits_{x\in L,y=1}f^t_i(x)$ 和 $\sum\limits_{x\in L,y=-1}f^t_i(x)$ 分别代表已标记样本集 L 中被视角 i 分类为 $y=1$ 和 $y=-1$ 的样本分类置信度之和. 利用 ε^t_i,我们可根据:$\omega^t_i=\frac{1}{Z^t_1}\ln\left(\frac{1-\varepsilon^t_i}{\varepsilon^t_i}\right)$ 来得到 ω^t_i,Z^t_1 为权重归一化系数. 那么,多视角分类器的分类误差 δ^t 可通过式(3)计算得到:

$$\delta^t=\sum_{j=1}^{N}\beta^t_j\left|h^t(x_j)-y_j\right| \tag{3}$$

(b)经过第 t 次查询后已标注训练集中的样本数目更新为:$J^t=J^{t-1}\cup L^t$. 其中,J^t 代表第 t 次查询时的已标注样本集,L^t 则代表通过采样后新加入的样本集. 由于主动学习具有训练集规模不断增大的特性,因此对于初始训练集通常为小样本分类问题的主动学习而言,需要考虑训练样本集 J 的大小 $|J^t|$ 对于更新多视角分类器权重 α_t 时的影响:

$$\alpha_t=\frac{1}{Z^t_2}\left(\ln\left(\frac{1-\delta^t}{\delta^t}\right)+\lambda|J^t|\right) \tag{4}$$

我们可根据:$\omega^{t+1}_j=\omega^t_j\beta_t^{1-e^t}$ 来更新各个样本的权重. 其中,$\beta_t=\delta_t/(1-\delta_t)$,如果样本 x_j 被正确分类则 $e_j=0$,否则 $e_j=1$.

(c)被查询样本的 x_i 分类假设 $H^t(x)$ 为历史上所有 k 次查询中分类假设的加权和,如式(5)所示:

$$H^t(x)=\begin{cases}1, & \sum\limits_{t=1}^{K}\alpha_t h^t(x)\geqslant\frac{1}{2}\sum\limits_{t=1}^{K}\alpha_t\\ 0, & \text{else}\end{cases} \tag{5}$$

4.2 采样策略

为了在上述不同的样本分布下均能以较高概率查询到高信息量的未标记样本,我们提出了一种基于分级竞争的采样机制,其具体流程如图 2 所示. 当具有分类争议的未标记样本数目超过阈值 T_{num} 时,本文首先通过聚类将特征空间中上述未标记样本集聚合成 K 个不同的类. 接着,根据自顶向下的方式,通过类间和类内采样竞争依次确定各个聚类的采样数目以及每个类内被采样的样本.

4.2.1 类间采样竞争

聚类法由于能够获得样本在特征空间中的基本分

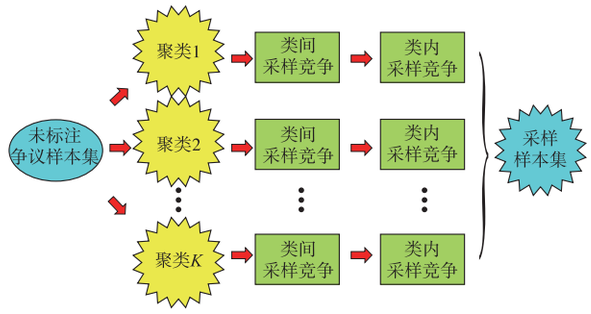


图2　基于分级竞争机制的采样流程

布结构描述,因而对于发掘高信息量的未标记样本具有积极的作用. 在类间采样竞争中,我们通过基于图切割法来最小化目标函数的谱聚类方法对分类争议样本进行准确高效的聚类. 其中,本文使用归一化割(Normalized Cut)[20]作为目标函数进行最小化求解.

不过,归一化割法的时间复杂度和空间复杂度分别为 $O(N^3)$ 和 $O(N^2)$,当分类争议的未标记样本数目较为庞大的情况发生时,求解将变得难以计算. 为此,我们采用了一种快速近似谱聚类方法来降低算法的复杂度,其具体步骤如下所示:(*a*)首先,采用传统的 K 均值聚类将分类争议的未标注样本 $x_1,x_2,\cdots,x_N$ 聚成 K 类,并提取各类的中心 $y_1,y_2,\cdots,y_k$ 作为具有代表性的样本;(*b*)其次,再采用归一化割法将上述类中心 $y_1,y_2,\cdots,y_k$ 聚成 K'类,并将各个分类争议样本 x_i 归类到所对应的类中心 y_i. 在本文中,K 均值聚类的数目 $K=N_{cp}/20$,其中 N_{cp} 为分类争议样本的总数,而谱聚类的数目 $K'=5$. 那么,快速近似谱聚类方法的计算复杂度为 $O(KNT)+O(K^3)$,其中 T 为 K 均值聚类的迭代次数. 不难发现,当 $K \ll N$ 时算法的计算复杂度可以得到显著下降.

在谱聚类完成后,本文分别定义如下两个类间采样的标准:(1)聚类的样本数目;(2)信息熵. 首先,在样本数目越多的类中,将采样越多数目的样本,即类 C 的采样数目 $Num(C)$ 和该类的样本数目 N_C 成正比:

$$Num(C) \propto N_c \tag{6}$$

其次,采用类内所有样本的信息熵之和来代表类 C 的信息量 $Ent(C)$,如式(7)所示:

$$\begin{aligned} &Ent(C) \propto -\sum_{i=1}^{N_c} P(x_i)\log(P(x_i)) \\ &\text{s.t. } P(x)=\frac{1}{N_C}\sum_{i=1}^{N_c} K(x,x_i) \end{aligned} \tag{7}$$

其中,类 C 的信息熵计算可等价于样本 x 在类 C 中的核密度估计[40]. K 代表帕森窗(Parzen Window),其定义如下:

$$K(x,x_i)=\frac{\exp\left(-\frac{1}{2}(x-x_i)^{\mathrm{T}}\Sigma^{-1}(x-x_i)\right)}{(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}} \tag{8}$$

最终,通过线性加权方式结合上述两个采样标准来获得类 C 的具体采样数目:

$$N_C^S=\frac{N_T}{Z}[\gamma Num(C)+(1+\gamma)Ent(C)],C=1,2,\cdots,K \tag{9}$$

在式(9)中,N_C^S 代表在类 C 中的采样数目,Z 为归一化因子,N_T 为在当前查询中采样的总样本数目,$[\cdot]$代表取整操作.

4.2.2　类内采样竞争

在目前主流 SVM 主动学习的采样方法中,在分类不确定度 F 的基础上结合诸如样本在特征空间中的冗余度 R 这一类方法获得了不少应用并取得了良好的效果[21,22]. 该类方法在进行基于目标函数最小化的采样时,对上述两种度量均采用了简单的加权方式:$\min\{\omega \cdot F+(1-\omega)R\}$. 其中,归一化加权系数 ω 通过人工预先设定,即在每次查询时样本的分类不确定度和冗余度之间具有恒定的权重比. 但事实上,每次查询后上述两者均会发生不同程度的变化,因而恒定的权重无法准确地描述上述两种度量之间的真实关系.

为了改进上述采样方法存在的问题,本项目拟在每次查询时动态地估计样本分类不确定度和冗余度之间的权重以获得最优的采样效果,而 Hoi 等人提出的方法[23]通过二次规划求解较好地平衡了上述两种度量在采样中的重要程度. 本项目拟将该方法扩展到多视角的情况以实现各个聚类内的采样,其可通过如下的最小化方式实现:

$$\begin{aligned} &\min_{P\in R^{n-l}} P^T\tilde{f}_v+\frac{1}{2}p^{\mathrm{T}}K_{u,u}P \\ &\text{s.t. } p^{\mathrm{T}}u=k,0\le p\le 1 \end{aligned} \tag{10}$$

式(10)为求解归一化变量 $p_i\in[0,1]$ 的过程,表示未标注样本被采样的概率. $\hat{f}_v=(|f_v(x_1)|,\cdots,|f_v(x_l)|)^{\mathrm{T}}$,其中 $f_v(x)$ 代表了第 v 个视角输出的样本 x 的分类置信度,$x_1,\cdots,x_l$ 为待查询的未标注样本. $\boldsymbol{u}$ 为单位向量,$k=N_C^S$ 为批处理采样的未标注样本数目.

式(10)的第一部分 $p^{\mathrm{T}}\tilde{f}$ 代表了第 v 个视角中样本的分类不确定度. 通过最小化 $p^{\mathrm{T}}\tilde{f}$,算法倾向于优先采样距离第 v 个视角的分类边界较近的样本. 式(10)的第二部分 $p^{\mathrm{T}}K_{u,u}P$ 则代表了样本之间的冗余度. 通过最小化 $p^{\mathrm{T}}K_{u,u}P$,被采样的样本之间则具有较小的相似度. 由于核函数 $K_{u,u}$ 为凸函数,因此可通过经典的凸二次规划方法对 P 进行求解,从而得到第 v 个视角中未标注样本的采样概率 $\{p_1^v,p_2^v,\cdots,p_l^v\}$.

为了实现各个聚类内的采样,本项目拟借鉴 Co-Testing 算法中具有抗噪能力的 Conservative 采样思想[19],如式(11)所示:

$$N^*\leftarrow\arg\min_{N\subseteq U\wedge|N|=k,v_k,v_l\in V}(\max p_i^{v_k}-\min_i^{v_l}) \tag{11}$$

在式(11)中,$\max p_i^{v_h}$ 和 $\min_i^{v_l}$ 分别为通过二次规划求解得到的样本 i 的最高和最低采样概率,其中 v_h 和 v_l 分别对应于多视角分类误差 ε 最小和最大的视角. 我们将根据($\max p_i^{v_h} - \min p_i^{v_l}$)的值进行升序排列,并最终将差值最小的前 k 个分类争议样本进行采样和标注后加入到已标注样本集 N^* 中.

当训练集中具有分类争议的未标注样本数目 N 过大时,那么方法[23]会因为核函数 $K_{u,u}$ 的复杂度 $O(N^2)$ 过大而变得难以计算,因而其仅适用于小规模的数据集. 为此,本项目拟通过设置阈值的方式实现多视角采样的自适应调整,从而适用于具有大规模未标注样本的数据集:

(a)当分类争议样本的数目未超过 T_n 时,直接采用基于二次规划的多视角采样策略;

(b)当分类争议样本的数目超过 T_n 时,则采用分级采样竞争机制. 二次规划仅需要在各个具有相对较少样本数目的类内进行,这样,每个类内 $K_{u,u}$ 的计算复杂度仅为原先基于全体分类争议样本情况下的 $1/K'^2$.

5 实验结果与分析

在本实验中,我们分别采用了在图像检索领域中常用的 COREL 彩色图像集[24]和 13 Natural Scene Categories 黑白图像集[25]来分别评价本文提出的分类假设生成和采样策略的效果. 在 COREL 图像集中,本文分别提取颜色和纹理两种简单特征来对图像进行描述. 由于从 13 Natural Scene Categories 图像集中采集的训练样本均为灰度图像,因此我们除了用同样方式提取纹理特征向量之外,还统计每幅图像灰度的均值和方差. 在单视角主动学习时,我们将两种特征对应的特征向量连接成一个特征向量以表示当前图像的特征. 在多视角主动学习时,则分别将两种特征分别作为不同的视角.

5.1 算法比较和参数设置

在实验中,我们将本文算法 $BSVM_{AL}^{MV+QP}$ 与几种经典的 SVM 主动学习算法进行比较. 我们参与比较的算法如下所示:(1)采用随机的方式进行采样的 SVM 主动学习,简称 SVM_{AL}^{Rand};(2)将基于样本到分类超平面的距离作为采样标准的 SVM 主动学习[25],简称 SVM_{AL}^{DIST};(3)在样本到分类超平面距离基础上结合角度多样性作为采样标准的 SVM 主动学习[22],简称 $SVM_{AL}^{DIST+AngDIV}$. 其中,距离和角度多样性之间的权重在主动学习前通过交叉验证方式获得;(4)在样本不确定度和冗余度基础上采用基于二次规划采样的 SVM 主动学习[23],简称 SVM_{AL}^{QP}. 此外,我们还将本文算法根据不同模块组合来获得相应的算法,以此来评价各个模块在算法中起到的作用:(5)将 $BSVM_{AL}^{MV+QP}$ 变成单视角的模式,简称 $BSVM_{AL}^{SV+QP}$. 该算法是为了比较多视角主动学习与单视角主动学习之间的差异;(6)$BSVM_{AL}^{MV+QP}$ 没有结合 Boosting 技术改进生成的分类假设,简称 SVM_{AL}^{MV+QP}. 该算法是为了评价 Boosting 对于主动学习的影响.

我们将把 $BSVM_{AL}^{MV+QP}$ 与上述几种 SVM 主动学习算法进行比较,并考虑以下两个因素来对算法进行评价:(1)Label Size:在图像检索中的初始训练样本集的大小;(2)Batch Size:每次主动学习查询中被选择进行人工标记的未标记样本数目;本文通过调整不同的 Label Size 和 Batch Size 来评估对算法的影响程度. 本文采用平均精度(Average Precision, AP)和平均精度均值(Mean of Average Precision, MAP)作为衡量各个算法的标准. 此外,所有 SVM 主动学习算法中的核函数均采用卡方核(Chi-Square Kernel)以实现直方图交叉操作,并以"One-Against-All"形式训练多类别主动学习分类器以获得场景的类别估计.

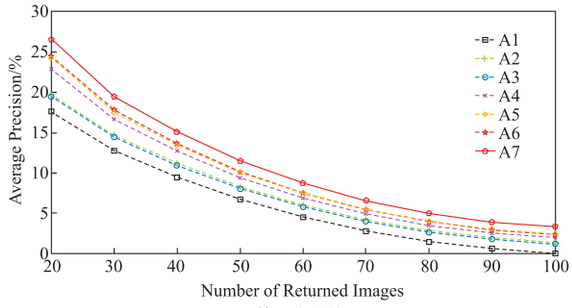
5.2 固定 Label Size 和 Batch Size 分析

在第一个实验中,我们分别固定 Label Size = 50 和 Batch Size = 50,来模拟一个 8 次查询的主动学习过程. 图 3 给出了图像检索时前四次相关反馈的 AP 以描述主动学习的分类性能变化. 其中,A1 – A6 分别依次代表 SVM_{AL}^{Rand}, SVM_{AL}^{DIST}, $SVM_{AL}^{DIST+AngDIV}$, SVM_{AL}^{QP}, $BSVM_{AL}^{SV+QP}$ 和 SVM_{AL}^{MV+QP} 上述六种主动学习方法(用虚线表示),A7 则为本文提了出的方法 $BSVM_{AL}^{MV+QP}$(用实线表示).

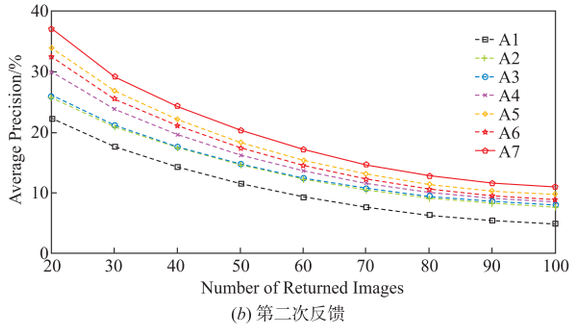
在图 3 给出的前四次相关反馈中我们可以定性地观察到:(1)和 SVM_{AL}^{Rand} 相比,其余算法在不同返回图像数目的情况下均具有更高的 AP,可见上述各种主动学习的采样策略均具有比随机采样更好的性能,并且随着反馈次数的增加 SVM_{AL}^{Rand} 与其它算法的 AP 差异越来越明显;(2)四种基于二次规划采样的方法 SVM_{AL}^{QP}, $BSVM_{AL}^{SV+QP}$, SVM_{AL}^{MV+QP} 和 $BSVM_{AL}^{MV+QP}$ 的 AP 均要比其余三种算法更高,这说明了相比之下基于二次规划的采样策略具有更好的性能,并且本文方法 $BSVM_{AL}^{MV+QP}$ 在所有基于二次规划采样的算法中均具有最好的表现;(3)$BSVM_{AL}^{MV+QP}$ 具有比 $BSVM_{AL}^{SV+QP}$ 更高的 AP,证明了多视角主动学习在实际图像检索应用中相对于单视角主动学习的优势;(4)SVM_{AL}^{MV+QP} 的 AP 从第二次反馈后开始比 $BSVM_{AL}^{MV+QP}$ 更低,可见 Boosting 起到了改进主动学习分类性能的作用.

5.3 固定 Label Size 分析

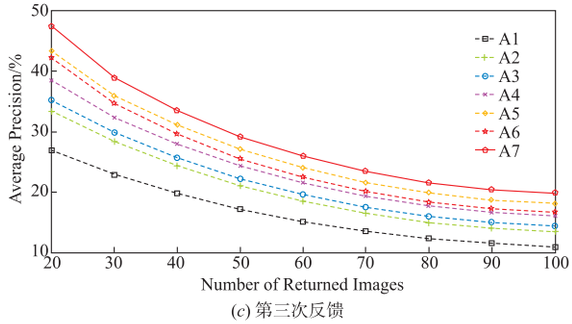
在第二个实验中,我们固定 Batch Size = 50 并通过变化 Label Size 来定量地观察算法的性能变化. 图 4 给出了反馈回的前 20 个图像样本随着 Label Size 的不同取值(Label Size = 10, 20, 30, 40 和 50)而对应的 AP 变化. 其中,A1 – A7 依次代表了 SVM_{AL}^{Rand}, SVM_{AL}^{DIST},



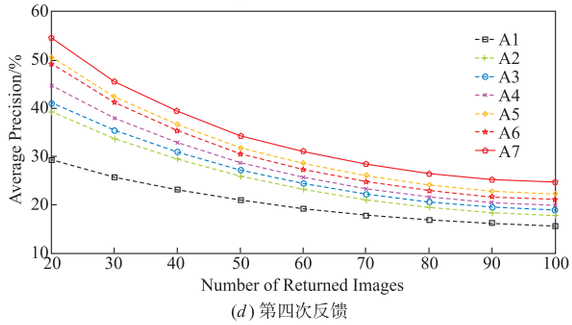
(*a*) 第一次反馈



(*b*) 第二次反馈



(*c*) 第三次反馈



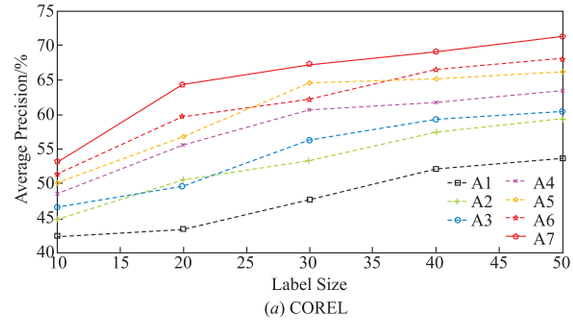
(*d*) 第四次反馈

图3　各种算法的性能比较

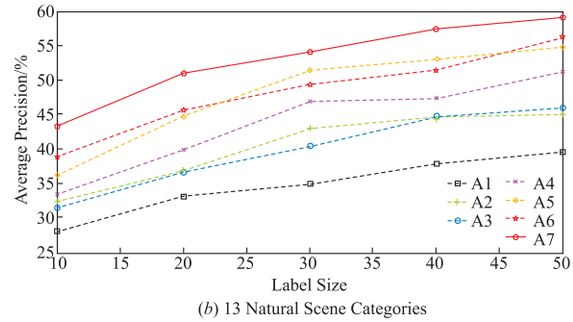
$SVM_{AL}^{DIST+AngDIV}$, SVM_{AL}^{QP}, $BSVM_{AL}^{SV+QP}$, SVM_{AL}^{MV+QP} 和 $BSVM_{AL}^{MV+QP}$ 上述七种主动学习算法.

从图 4 可以看到,和其他方法相比,在不同的 Label Size 情况下本文方法 $BSVM_{AL}^{MV+QP}$ 均具有最高的 AP. 随着 Label Size 的不断增加,$BSVM_{AL}^{MV+QP}$ 算法的 AP 增长均呈现明显的逐步下降趋势,这同时说明了 $BSVM_{AL}^{MV+QP}$ 算法具有较好收敛性能的同时适合处理初始标记样本集较小的情况.

表 1 给出了上述七种主动学习算法基于不同 Label



(*a*) COREL



(*b*) 13 Natural Scene Categories

图4　Label Size变化情况下反馈回的前20个图像样本的AP变化

Size 的 MAP 统计,符号 ± 后面的数值代表了 AP 的标准差. 通过表 1 中可以看到,对于 COREL 和 13 Natural Scene Categories 两个图像集:(1)和 A1,A2 和 A3 上述三种非二次规划采样的算法相比,四种基于二次规划采样的算法 A4,A5 ,A6 和 A7 均具有更高的 MAP,其中 SVM_{AL}^{QP} 比前三种方法中 MAP 最高的 $SVM_{AL}^{DIST+AngDIV}$ 分别高出了 3.56% 和 4.92%,这说明基于二次规划的采样策略能够更为有效地选择具有高信息量的未标记样本来改善分类性能;(2) $BSVM_{AL}^{MV+QP}$ 的 MAP 比 $BSVM_{AL}^{SV+QP}$ 分别高出了 4.42% 和 3.94%,显示了多视角主动学习相对于单视角主动学习的优势;(3) $BSVM_{AL}^{MV+QP}$ 的 MAP 比 SVM_{AL}^{MV+QP} 分别高出了 3.38% 和 4.68%,证明了结合 Boosting 技术有助于强化每次主动学习反馈后的分类效果.

表 1　Label Size 变化情况下反馈回的前 20 个图像样本的 MAP 统计

MAP(%)	A1	A2	A3	A4	A5	A6	A7
COREL	47.88 ±1.97	53.10 ±2.16	54.48 ±0.84	58.04 ±1.14	60.58 ±1.00	61.62 ±0.47	65.00 ±1.29
13 Scene	34.66 ±0.93	40.38 ±0.89	39.82 ±0.36	43.74 ±0.20	48.02 ±0.07	48.28 ±0.81	52.96 ±0.69

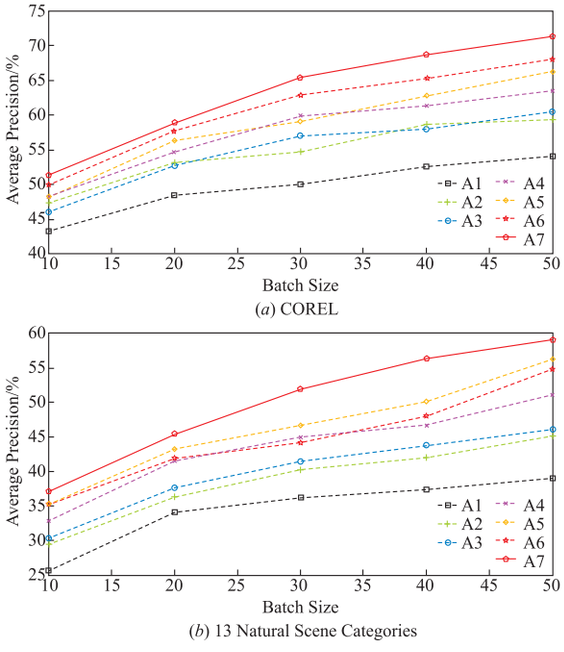
5.4　固定 Batch Size 分析

在第三个实验中,我们固定 Label Size = 50,并通过变化 Batch Size 的大小(Batch Size = 10,20,30,40 和 50)来观察算法的性能变化. 在这里,我们同样采用 AP 和 MAP 作为衡量算法的标准.

图 5 给出了反馈回的前 20 个图像样本随着 Batch

Size 的不同取值而得到的 AP 变化. 根据图 5 可以得到和图 4 相同的结论,即对于两个不同图像集,本文算法 $BSVM_{AL}^{MV+QP}$ 和其余方法相比在不同的 Batch Size 中均具有最高的 AP. 此外,我们还可以观察发现:随着 Batch Size 不断增加,$BSVM_{AL}^{MV+QP}$ 的 AP 增长幅度通常要超过 SVM_{AL}^{DIST} 和 $SVM_{AL}^{DIST+AngDIV}$ 等非二次规划采样的算法. 比如在 COREL 图像集中,当 Batch Size 从 20 增加到 30 时,$BSVM_{AL}^{MV+QP}$ 和 $SVM_{AL}^{DIST+AngDIV}$ 的 AP 增长幅度分别为 5.78% 和 4.32%;当 Batch Size 从 30 增加到 40 时,两者的 AP 增长幅度则变为 2.54% 和 1.71%. 在 13 Natural Scene Categories 图像集中,当 Batch Size 从 20 增加到 30 时,$BSVM_{AL}^{MV+QP}$ 和 $SVM_{AL}^{DIST+AngDIV}$ 的 AP 增长幅度分别为 7.22% 和 6.25%;当 Batch Size 从 30 增加到 40 时,两者的 AP 增长幅度则变为 5.14% 和 4.31%. 上述现象均说明了基于二次规划的采样方法由于能够有效地采样到更多高信息量的样本,因此随着 Batch Size 的增加其 AP 的增幅往往会快于其他的算法.

表 2 给出了上述七种主动学习算法基于不同 Batch Size 的 MAP 统计. 通过表 2 同样可得到和图 5 类似的结论,即对于不同图像集我们看到:A4,A5,A6 和 A7 上述四种基于二次规划采样的算法具有比算法 A1,A2 和 A3 具有更高的 MAP,其中 SVM_{AL}^{QP} 比前三种非基于二次规划采样的算法中 MAP 最高的 $SVM_{AL}^{DIST+AngDIV}$ 分别高出了 2.7% 和 3.64%,证明了基于二次规划的采样策略相对于传统采样方式的有效改进;$BSVM_{AL}^{MV+QP}$ 的 MAP 比 $BSVM_{AL}^{SV+QP}$ 分别高出了 6.58% 和 3.68%,基于聚类的多视角采样具有比单视角采样更好的表现;$BSVM_{AL}^{MV+QP}$ 的 MAP 比 SVM_{AL}^{MV+QP} 分别高出了 2.34% 和 5.12%,同样证明了 Boosting 有助于改进主动学习的分类假设生成.



(a) COREL

(b) 13 Natural Scene Categories

图5 Batch Size变化情况下反馈回的前20个图像样本的AP变化

表 2 Batch Size 变化情况下反馈回的前 20 个图像样本的 MAP 统计

MAP(%)	A1	A2	A3	A4	A5	A6	A7
COREL	49.70 ±1.00	54.68 ±0.06	54.88 ±1.26	57.58 ±0.69	58.54 ±0.43	60.78 ±1.30	63.12 ±0.23
13 Scene	34.42 ±0.70	38.58 ±0.59	39.78 ±0.93	43.42 ±0.07	46.26 ±0.91	44.82 ±0.54	49.94 ±1.35

6 总结

本文在经典多视角学习方法 Co-Testing 的基础上,提出了一种适用于大规模数据集的多视角主动学习算法:$BSVM_{AL}^{MV+QP}$. 该算法同时在分类假设生成和采样中分别提出了相应的改进策略. 在分类假设生成上,在多视角主动学习框架中有效结合了 Boosting 技术思想,通过将历史上各次查询得到的分类假设输出进行加权的方式来实现每次查询后分类假设的强化;在高信息量样本采样上,则提出了一种自适应的分级竞争采样策略,通过无监督谱聚类获得上述样本的空间分布描述,并在各个聚类中结合样本的分类不确定度和冗余度信息通过二次规划求解以获得可靠的批处理采样. 通过基于不同数据集的图像分类实验结果表明,本文提出的分类假设生成和采样策略不仅均有助于改进传统多视角主动学习的性能,而且相比于现有经典的单视角主动学习算法均能够更快地实现收敛并达到较高的分类准确性.

参考文献

[1] H A Simon, G Lea. Problem solving and rule education: a unified view knowledge and organization [J]. Ernuam, 1974, 15(2): 63 – 73.

[2] D Lewis, J Catlett. Heterogeneous uncertainty sampling for supervised learning [A]. Proceedings of the 11th International Conference on Machine Learning [C]. Rutgers University, New Brunswick, NJ, USA, 1994. 148 – 156.

[3] A McCallum, K Nigam. Employing EM in pool-based active learning for text classification [A]. Proceedings of the 15th International Conference on Machine Learning [C]. Madison, Wisconsin, USA, July 24 – 27, 1998. 359 – 367.

[4] D Cohn, L Atlas, R Ladner. Improving generalization with active learning [J]. Machine Learning, 1994, 15(2): 201 – 221.

[5] N Roy, A McCallum. Toward optimal active learning through sampling estimation of error reduction [A]. Proceedings of the 18th International Conference on Machine Learning [C]. Williamstown, MA, USA, 2001. 441 – 448.

[6] Y Freund, H S Seung, E Shamir, N Tishby. Selective sampling using the query by committee algorithm [J]. Machine Learning, 1997, 28(2): 133 – 168.
[7] N Abe, H Mamitsuka. Query learning strategies using boosting and bagging [A]. Proceedings of the 15th International Conference on Machine Learning [C]. Madison, Wisconsin, USA, July 24 – 27, 1998. 1 – 10.
[8] I Muslea, S Minton, C A Knoblock. Selective sampling with redundant views [C]. Proceedings of the 17th National Conference on Artificial Intelligence, Austin, Texas, 2000. 621 – 626.
[9] A Blum, T Mitchell. Combining labeled and unlabeled data with co-training [A]. Proceedings of the Workshop on Computational Learning Theory [C]. Morgan Kaufmann, San Francisco, CA, 1998. 92 – 100.
[10] I Muslea, S Minton, C A Knoblock. Active + semi-supervised learning = robust multi-view learning [A]. The 19th International Conference on Machine Learning [C]. Sydney, Australia, July 8 – 12, 2002. 435 – 442.
[11] I Muslea, S Minton, C A Knoblock. Active learning with strong and weak views: a case study on wrapper induction [C]. Proceedings of the International Joint Conference on Artificial Intelligence [C]. Acapulco, Mexico, 2003. 415 – 420.
[12] C Dima, M Hebert. Active Learning for Outdoor Obstacle Detection [A]. Proceedings of the Robotics Science and Systems [C]. Massachusetts Institute of Technology, Cambridge, Massachusetts, 2005.
[13] D Wei, M M Crawford. Active learning via multi-view and local proximity co-regularization for hyperspectral image classification [J]. IEEE Journal of Selected Topics in Signal Processing, 2011, 5(3): 618 – 628.
[14] J Cheng, K Q Wang. Active learning for image retrieval with Co-SVM [J]. Pattern Recognition, 2006, 40(1): 330 – 334.
[15] W Wang, Z H Zhou. On multi-view active learning and the combination with semi-supervised learning [A]. International Conference on Machine Learning [C]. Helsinki, Finland, 2008. 1152 – 1159.
[16] W Wang, Z H Zhou. Multi-view active learning in the non-realized case [A]. Advances in Neural Information Processing Systems [C]. Vancouver, Canada, 2010. 2388 – 2396.
[17] X Y Zhang, C S Xu, H Q Lu, S D Ma. Multi-view multi-label active learning for image classification [A]. IEEE International Conference on Multimedia and Expo [C]. New York, 2009. 258 – 261.
[18] W H Yang, G Q Liu, L Zhang, E H Chen. Multi-view learning with batch mode active selection for image retrieval [A]. Proceedings of the 21st International Conference on Pattern Recognition [C]. Tsukuba, 2012. 979 – 982.
[19] I Muslea, S Minton, C A Knoblock. Active learning with multiple views [J]. Journal of Artificial Intelligence Research, 2006: 203 – 233.
[20] J B Shi, J Malik. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888 – 905.
[21] S J Huang, R Jin, Z H Zhou. Active learning by querying informative and representative examples [A]. Advances in Neural Information Processing Systems [C]. USA, 2010. 36(10): 1936 – 1949.
[22] C K Dagli, S Rajaram, T S Huang. Leveraging active learning for relevance feedback using an information theoretic diversity measure [A]. ACM Conference on Image and Video Retrieval [C]. Lecture Notes in Computer Science, Tempe, AZ, USA, 2006. 123 – 132.
[23] S C H Hoi, R Jin, J K Zhu, M R Lyu. Semi-supervised SVM batch mode active learning for image retrieval [A]. IEEE Conference on Computer Vision and Pattern Recognition [C]. IEEE, 2008. 1 – 7.
[24] Fei-Fei Li, P Perona. A Bayesian hierarchical model for learning natural scene categories [A]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition [C]. San Diego, CA, June 2005. 524 – 531.
[25] S Tong, E Chang. Support vector machine active learning for image retrieval [A]. Proceedings of 9th ACM Iinternational Cconference on Multimedia [C]. New York, NY, USA, 2001. 107 – 118.

作者简介

姚拓中　男. 1983 年 2 月出生, 浙江宁波人. 2011 年毕业于浙江大学大学信电系, 其后在中科院宁波工业技术研究院做博士后研究, 2014 年进入宁波工程学院电子与信息工程学院, 主要从事计算机视觉和机器学习等方面的研究工作.
E-mail: thomasyao@ zju. edu. cn

安　鹏　男, 1981 年 11 月出生, 山西太原人, 2009 年毕业于清华大学工程物理系, 同年就职于宁波工程学院电子与信息工程学院, 主要从事图像处理和嵌入式系统设计的研究工作.
E-mail: anp04@ 126. com