基于动态确定度传播的选择性采样*

张晓宇

(中国科学技术信息研究所 战略研究中心, 北京 100038)

摘　要: 传统主动学习中采用的批量采样模式忽略了样本之间的相互关系,因而会不可避免地引入冗余。针对上述问题,提出了一种动态批量采样模式,采取"逐一标注,批量训练"的流程,综合利用当前分类模型和先前标注样本对后续采样进行动态指导;在此基础上,进一步提出了基于动态确定度传播的选择性采样算法,有效地提高了所选取样本的信息量。实验结果证明,基于动态确定度传播的选择性采样算法能够显著改进分类结果。

关键词: 相关反馈; 半监督学习; 主动学习; 多视角学习; 选择性采样

中图分类号: TP37　　**文献标志码**: A　　**文章编号**: 1001-3695(2012)05-1929-05

doi:10.3969/j.issn.1001-3695.2012.05.088

Effective selective sampling with dynamic certainty propagation

ZHANG Xiao-yu

(*Research Center for Strategic Science & Technology Issues*, *Institute of Scientific & Technical Information of China*, *Beijing* 100038, *China*)

Abstract: In traditional active learning, selective sampling was performed in batch mode, which neglected examples' correlation and thus inevitably brought in redundancy. This paper presented a dynamic batch sampling mode, using both the existing classification boundary and the previously labeled examples as guidance for further selection. Then it proposed a dynamic certainty propagation(DCP) algorithm for informative example selection. Experimental results demonstrate the effectiveness of selective sampling with DCP algorithm.

Key words: relevance feedback; semi-supervised learning; active learning; multi-view learning; selective sampling

在基于内容的图像检索(content-based image retrieval, CBIR)[1~3]中,相关反馈[4,5]是一种通过人机交互获取用户检索需求的有效手段。其大致流程为:由用户对一些图像进行标注,利用标注信息对检索模型进行重新训练以提高检索性能。一方面,由于语义鸿沟的存在,要想从图像底层特征准确地刻画用户需求,需要用户标注尽可能多的图像作为训练样本;另一方面,标注过程费时费力,从用户体验出发自然希望标注尽可能少的图像。为了解决上述矛盾,需要充分利用好图像库中海量的未标注图像和有限的已标注图像。从机器学习的角度而言,相关反馈过程本质上是一个分类问题,即基于已标注图像训练集,将图像库中的图像分为与查询相关和不相关两类。因此,可以将机器学习中的算法引入到相关反馈过程中,以改进其性能。

1　相关工作

主动学习[6,7]是针对已标注样本少、未标注样本多且易于获得的情况而提出的一种机器学习方法。其主要思想是:仅仅选取最有信息的样本进行标注,使得根据这些样本的标注信息所训练出来的新模型的分类性能得到尽可能大的提升。将主动学习方法应用到相关反馈过程中,由系统有针对性地主动选取对模型改进帮助最大的图像让用户进行标注,从而可以凭借有限的图像标注量获得尽可能大的检索性能提升。此类方法中具有代表性的包括 SVM_{Active}[8]、SSAIR[9]、Co-SVM[10]等。在主动学习算法中,核心问题是最有信息样本的选取方法,即选择性采样方法。在 SVM_{Active} 中,距离当前分类面最近的一批样本即为最有信息的样本。假设 f 是当前的 SVM 分类器,$|f(x)|$ 可以看成是一个样本 x 距离分类面的距离,则在 SVM_{Active} 中具有最小 $|f(x)|$ 值的一批样本作为最有信息的样本选取出来进行标注。

多视角学习[11~13]是另一种针对已标注样本少、未标注样本多且易于获得的情况而提出的机器学习方法。其主要思想是:从不同的视角去描述样本,并根据不同视角间的差异性获取额外信息,从而提高分类性能。由于解决的问题与主动学习相同,因此研究工作将多视角学习融合到主动学习的框架内,以进一步改进分类效果。在 SSAIR 和 Co-SVM 两种方法中,都综合利用了多视角学习和主动学习这两种算法。其大体思路是:首先根据不同的图像子特征,利用训练样本训练出相应的子分类模型;然后将不同子分类模型间分歧最大的样本选取出来作为最有信息的样本,并进行标注。假设 f_i 是基于不同视角训练获得的子分类模型,m 是不同视角的数目,则一个样本 x 在不同视角下分类结果的差异可以用以下公式表示:

SSAIR:
$$f(x) = \sum_{i=1}^{m} f_i(x) \tag{1}$$

Co-SVM:
$$f(x) = \sum_{i=1}^{m} \mathrm{sgn}[f_i(x)] \tag{2}$$

收稿日期: 2011-08-16; **修回日期**: 2011-09-30　　**基金项目**: 中央级公益性科研院所基本科研业务费专项资金资助项目(ZD2011-7-3);中国科学技术信息研究所科研项目预研资金资助项目(YY-201114);国家自然科学基金资助项目(60475010)

作者简介: 张晓宇(1983-),男,江苏镇江人,助理研究员,博士,主要研究方向为模式识别与智能系统、科技情报与知识管理(zhangxy@istic.ac.cn).

类似地,将具有最小$|f(x)|$值的一批样本作为最有信息的样本选取出来进行标注。

2 动态批量选择性采样

由前文可见,无论是单纯的主动学习(如 SVM_{Active}),还是融合了多视角学习的主动学习(如 SSAIR 和 Co-SVM),本质上采用的都是批量选择性采样模式,即:事先定义一种度量(如$|f(x)|$)来量化未标注样本在当前分类模型下预测结果的确定度;然后批量地选择确定度最小的 k 个样本进行标注;最后根据这 k 个样本的标注信息对分类模型进行训练更新。这一过程可以概括为批量标注、批量训练。

在文献[14]中,作者深入分析了分类问题中未标注样本的价值所在,并在此基础上提出了选择性采样的两个重要准则:a)低确定度准则,即选取的样本应具有较低的确定度;b)低冗余度准则,即选取的样本不应与先前已选取的样本之间存在过多的信息冗余。

不难发现,批量选择性采样模式虽然遵循了低确定度准则,却忽略了低冗余度准则。可以借助一个简单的二维特征空间内的分类问题对其进行解释。如图 1 所示,在批量规模 $k=2$ 的情况下,按照批量选择性采样的方法,样本 A 和 B 将被选取作为最有信息样本,因为它们是距离当前分类面最近的两个样本,从而具有最低的确定度。但是,当样本 A 被标注之后,样本 C 却显得比 B 更加有信息。这是因为样本 B 和 A 在特征空间上极为相似(即样本 B 与 A 直接存在着很高的信息冗余),当近邻样本 A 被标注之后,样本 B 的确定度实际上被大大地提高了,因而此时如果再选取样本 B 进行标注,其所能提供的额外信息量非常有限。在批量选择性采样中,由于样本是以批量的方式同时选取并标注的,因此同一批被选取的样本之间的相互关系被忽略了,进而影响了所选取样本的信息量。

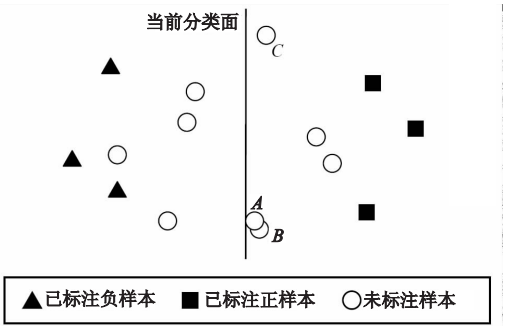


图1 低冗余度采样准则示例

逐一选择性采样的方式可以解决批量选择性采样的不足,即:每次只选取距离当前分类面最近的一个样本进行标注,然后立即通过训练进行分类模型的更新,并根据更新后的分类面选取下一个样本,如此循环。逐一选择性采样可以看成是批量选择性采样在批量规模 $k=1$ 时的一个特例,简而言之也就是“逐一标注,逐一训练”。由于每选出一个样本进行标注后,都会马上对分类模型进行更新,从而在选取样本过程中使得当前分类面不断向最优分类面逼近,同时先前标注的样本信息也可以被充分利用以指导下一个样本的选取。因此,在标注样本数量相同的情况下,逐一选择性采样选取的样本要优于批量选择性采样,从而最终获得的分类模型也更优。但是,由于在逐一选择性采样中,每选择一个样本就要进行一次模型更新,导致其计算复杂度非常高,因而在实际的图像检索系统中使用逐一选择性采样策略进行相关反馈是不切实际的。

因此,本文提出了动态批量选择性采样模式,其主要思想是:每次选取一个最有信息的样本进行标注,并根据该样本的标注信息指导下一个样本的选取,如此循环;当一定数目的样本标注完成后,通过一次训练对分类模型进行更新。可见,动态批量选择性采样所使用的是“逐一标注,批量训练”的过程,它介于“批量标注,批量训练”与“逐一标注,逐一训练”之间。在动态批量选择性采样中,样本的选取不再仅仅取决于当前分类面,同时还在新标注样本信息的指导下动态地进行调整,从而在有效利用样本之间相互关系的同时也有效地弥补了当前分类模型的不足。而在执行效率方面,由于动态批量选择性采样的训练过程仍然是批量的,即在一定批量的样本全部完成标注之后才进行一次训练,其计算复杂度与批量选择性采样相比并没有显著变化。从这个意义上说,动态批量选择性采样很好地融合了批量选择性采样和逐一选择性采样两种模式的特点,从而有效地兼顾了性能和效率。三种选择性采样策略的相互关系如图 2 所示。

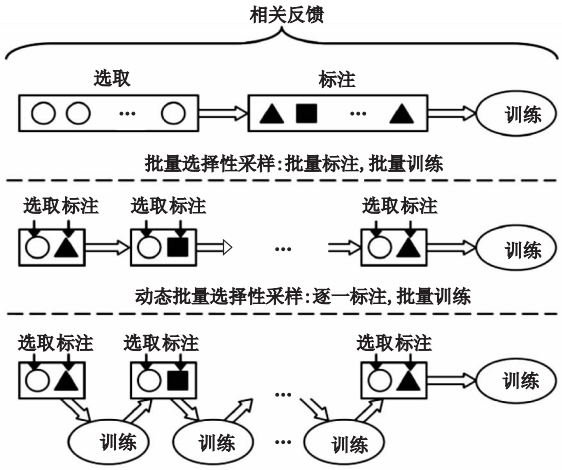


图2 三种选择性采样模式比较

3 动态确定度传播

为了更好地兼顾低确定度和低冗余度这两条准则,本文在动态批量选择性采样模式的框架内提出了动态确定度传播(dynamic certainty propagation, DCP)算法。其主要思想是:在每轮相关反馈之前,根据各未标注样本到分类面的距离赋予相应的初始确定度;相关反馈过程中,每次选取确定度最低的一个样本进行标注,每当一个新的样本被标注之后,其确定度将发生改变,并且根据样本之间的相互关系进行传播,以改变其他未标注样本的确定度,重复这一过程,直到一定批量的样本被标注;最后,利用标注样本进行分类器的重新训练。通过考虑未标注样本之间的相互关系,在保持低确定度的同时有效地降低了所选取样本的冗余度,从而使得最终获取的样本具有更高的信息量。

3.1 算法流程

用 $X=\{x_1, x_2, \cdots, x_n\}=U\cup L$ 表示整个样本集合。其中,U 和 L 分别表示未标注样本集和已标注样本集。对于每

一个样本 $x_i \in X$ ($1 \leqslant i \leqslant n$),$y_i \in \{0, 1\}$ 对应于它的类标。如果 $x_i \in L$,则 y_i 已知;如果 $x_i \in U$,则 y_i 未知,需要通过 $f(x_i)$ 去预测 y_i,其中 f 是当前 X 上的分类器。

对每一个未标注样本,用确定值来衡量其预测类标的确信程度,确定值的符号("+"或"-")表示相应的预测类标,确定值的绝对值被称之为确定度。对于一个未标注样本而言,其确定度越低,该样本就越有信息。用 C 表示未标注样本的确定值。初始情况下,每个未标注样本 $x_i \in U$ 的确定值由当前分类器决定:

$$C^{(0)}(x_i) = f(x_i) \tag{3}$$

采用逐一标注的方法,每次仅仅选取一个具有最低确定度的未标注样本 $x_l \in U$ 进行标注:

$$l = \mathrm{argmin}_i \mid C^{(r-1)}(x_i) \mid \tag{4}$$

其中:r 表示一轮相关反馈中第 r 个样本的选取。

在动态确定度传播算法中,未标注样本的确定值并不是一成不变的,而是随着样本的标注过程动态变化。它不仅取决于当前分类器,同时也受先前标注样本的影响。

在样本 x_l 被选取并标注之后,首先改变其确定值:

$$C^{(r)}(x_l) = \begin{cases} M & y_l > 0 \\ -M & y_l < 0 \end{cases} \tag{5}$$

其中:M 是一个预先设定的正常数来表示已标注样本的确定度。

x_l 确定值的改变将传播到其他的未标注样本,并对其确定值产生相应的影响。这种影响的程度取决于各样本与 x_l 之间的在特征空间的相似度。为了更好地描述样本之间的相互关系,本文构建了一个图模型 $G = (V, E)$,其中节点集合 V 即为样本集 X,而边集合 E 表示样本之间的关系,其权重由一个 $n \times n$ 的矩阵 $\boldsymbol{W}$ 决定。$\boldsymbol{W}$ 中的元素 w_{ij} 是用热核(heat kernel)[15] 表示的对应样本 x_i 和 x_j($1 \leqslant i, j \leqslant n$)之间的关系:

$$w_{ij} = \exp\left(-\frac{\| x_i - x_j \|^2}{t} \right) \tag{6}$$

其中:t 是一个反映样本之间影响强度的参数。显然,在特征空间内距离较近的样本之间的边权重较高,而随着样本之间距离的增大,其对应的权重将呈指数级衰减。基于 $\boldsymbol{W}$,将 x_l 确定值的改变传播到其他的未标注样本:

$$\Delta C^{(r)}(x_i) = w_{il} \Delta C^{(r)}(x_l) \tag{7}$$

其中:

$$\Delta C^{(r)}(x_i) = C^{(r)}(x_i) - C^{(r-1)}(x_i) \tag{8}$$

可见,x_l 仅仅会对近邻的其他样本产生比较显著的影响,而对于较远的样本,其影响可以忽略不计。

因此,对于每一个未标注样本,其新的确定值可以表示为

$$C^{(r)}(x_i) = C^{(r-1)}(x_i) + w_{il}[C^{(r)}(x_l) - C^{(r-1)}(x_l)] \tag{9}$$

使用式(9),可以在每次样本标注之后动态地对其他样本的确定值进行更新。

下面总结了动态确定度传播的算法步骤。

输入:当前分类器 f,已标注样本集 L,未标注样本集 U,样本相互关系矩阵 $\boldsymbol{W}$,在每轮相关反馈中需要标注的样本数目 k。

a)用 f 对 U 进行分类,初始化未标注样本的确定值:

$$C^{(0)}(x_i) = f(x_i), x_i \in U$$

b)对 $r = 1, 2, \cdots, k$,重复执行下述操作:

(a)选择具有最小确定度的一个样本 x_l 进行标注:

$$l = \mathrm{argmin}_i \mid C^{(r-1)}(x_i) \mid, x_i \in U$$

(b)从 U 中删除 x_l,并将 x_l 加入 L:

$$U = U - \{x_l\}, L = L \cup \{x_l\}$$

(c)改变 x_l 的确定值:

$$C^{(r)}(x_l) = \begin{cases} M & y_l > 0 \\ -M & y_l < 0 \end{cases}$$

(d)将 x_l 确定值的改变传播到其他的未标注样本:

$$C^{(r)}(x_i) = C^{(r-1)}(x_i) + w_{il}[C^{(r)}(x_l) - C^{(r-1)}(x_l)], x_i \in U$$

c)利用 L 对 f 进行更新。

3.2　参数赋值

在动态确定度传播算法中,总共包含两个参数,即 t 和 M,其中,t 是控制传播范围的参数,M 表示已标注样本的确定度。为了保证动态确定度传播算法具有更好的通用性,本文并不经验性地对其赋予定值,而是根据数据自身的分布自适应地取值。

参数 t 控制着新标注的样本对于其周围样本的影响程度,以保证只有近邻样本才会被其显著地影响。本文采用最近邻距离来自适应地刻画一个样本集空间分布的紧密程度。对于一个样本 x_i,计算其到最近邻之间的距离,记为 $d_{nn}(x_i)$,然后对所有最近邻距离构成的整个集合 $\{d_{nn}(x_i)\}(x_i \in X)$,计算其算术平均值为

$$d = \frac{1}{n} \sum_{i=1}^{n} d_{nn}(x_i) \tag{10}$$

最终参数 t 赋值为

$$t = 2d^2 \tag{11}$$

参数 M 描述的是已标注样本的确定度,由于样本已标注,因而其确定度 M 应该相对较高。首先,计算所有未标注样本确定度 $\{\mid C^{(0)}(x_i) \mid\}(x_i \in U)$ 的均值和方差为

$$\mu_C = \frac{1}{\mid U \mid} \sum_{x_i \in U} \mid C^{(0)}(x_i) \mid \tag{12}$$

$$\sigma_C^2 = \frac{1}{\mid U \mid} \sum_{x_i \in U} [\mid C^{(0)}(x_i) \mid - \mu_C]^2 \tag{13}$$

其中:$\mid U \mid$ 表示未标注样本集合 U 的大小(即未标注样本的数目)。然后,对 M 进行如下赋值:

$$M = \mu_C + 3\sigma_C \tag{14}$$

这样赋值,一方面保证了 M 足够大,从而可以显示出对于已标注样本足够高的确定度;另一方面也确保其不过分地大,仍然与大多数未标注样本的确定度具有可比性。

值得注意的是,上述参数赋值方法并不是唯一的,也可以采用其他方法对参数进行有效赋值。在实际使用中发现,算法的性能对于参数一定程度上的变化并不敏感。

3.3　算法讨论

现在来解释一下动态确定度传播算法是如何有效遵循低冗余度准则的。仍然以图 1 所示的二维特征空间上的分类问题作为示例。设选择性采样之前,样本 A、B 和 C 的确定值分别为 $C(x_A)$、$C(x_B)$ 和 $C(x_C)$。显然,初始状态下它们满足如下关系:

$$\mid C(x_A) \mid < \mid C(x_B) \mid < \mid C(x_C) \mid \tag{15}$$

依据低确定度原则,首先选择样本 A 进行标注。随后,对样本 A、B 和 C 的确定值进行更新(假设样本 A 被标注为正样本):

$$C'(x_A) = M$$
$$C'(x_B) = C(x_B) + w_{BA}[C'(x_A) - C(x_A)] = C(x_B) + w_{BA}[M - C(x_A)]$$
$$C'(x_C) = C(x_C) + w_{CA}[C'(x_A) - C(x_A)] = C(x_C) + w_{CA}[M - C(x_A)] \quad (16)$$

由于样本 B 与 A 之间的距离远小于样本 C 与 A 之间的距离,因而 w_{BA} 远大于 w_{CA},因此在这种情况下

$$|C'(x_B)| < |C'(x_C)| \quad (17)$$

将不再成立,样本 C 将具有更小的确定度,从而选择样本 C 作为下一个最有信息样本进行标注。

4 图像检索系统

在被动学习的 CBIR 系统中,相关反馈是直接在返回的结果图像上进行的,即用户如果对检索结果不满意,将直接在结果图像上进行标注。由于图像检索系统所返回的图像都是在分类器的角度而言最为确信的正类图像,即分类面正方向一侧距离分类面最远的一些样本,因此这些图像如果作为相关反馈中的待标注样本显然缺乏信息量,即使标注以后对图像检索结果的改进也将非常有限。

图 3 显示了采用动态确定度传播算法的 CBIR 系统界面。图中区域①为操作区,用户可以进行查询图像的提交、图像检索、图像库的增减等一系列操作。和其他主动学习的 CBIR 系统一样,本文的图像检索系统将相关反馈中待标注的图像与作为检索结果返回的图像分开,分别置于区域②和③:区域②是检索结果区,用来显示图像检索的返回结果(当然,如果用户愿意的话,也可以像使用被动学习的 CBIR 系统一样直接在返回结果上进行标注);区域③是相关反馈区,在这里显示的是最有信息的图像,如果用户对检索结果不满意,只需标注相关反馈区中的少量图像便可以显著地提高图像检索的性能。

图3 图像检索系统用户界面

传统的主动学习 CBIR 系统采用的是批量选择性采样模式,即“批量标注,批量训练”,所以在相关反馈区,待标注的图像是一并显示出来由用户标注的。而本文的 CBIR 系统所采用的是动态批量选择性采样策略,即“逐一标注,批量训练”,因而相应的相关反馈区,待标注的图像是逐一显示的。如图 4 所示,当用户标注完一幅图像之后,系统综合利用当前分类面和新标注图像的信息动态地选取下一幅图像,并显示出来交给用户标注。

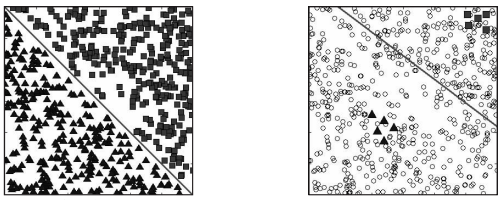
图4 图像检索系统相关反馈流程

5 实验

为了验证本文提出的动态确定度传播算法的有效性,分别在人造数据集和真实世界图像集上对其进行验证。

5.1 人造数据分类

本文构造了一个简单的二类数据集。如图 5(a)所示,整个数据集包括 610 个数据,正类、负类各 305 个,该数据集可以用一个线性分类器完美地将其划分。初始状态下,对于所有数据的真实类标是无法获知的,即所有的数据均是未标注的。从 10 个给定的训练数据出发,可以训练出一个初始的线性分类器,如图 5(b)所示。



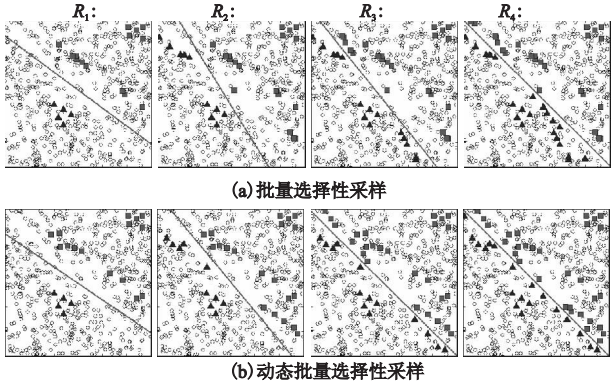
(a) 数据集真实分布 (b) 初始训练数据及相应的分类器

■正样本;▲负样本;○未标注样本;— 分类面

图5 人造数据集

实验将动态确定度传播算法(动态批量选择性采样模式)与传统的批量选择性采样模式进行比较。两种方法均采用线性核的 SVM 分类器,并且在每一轮相关反馈中均选取 10 个样本进行标注。

图 6(a)和(b)分别给出了批量选择性采样和动态批量选择性采样的实验结果。



(a)批量选择性采样

(b)动态批量选择性采样

图6 人造数据集上选择性采样比较

从图中不难发现,在动态批量选择性采样下,分类器能够更快地收敛于最优分类模型。例如,在 3 轮相关反馈(标注了 30 个样本)之后,动态批量选择性采样所对应的分类器就已经几乎达到最优了;而传统的动态批量选择性采样则需要 4 轮相关反馈(标注 40 个样本)才能达到类似的效果。这意味着,在分类器性能相当的条件下,本文方法所需要标注的数据量更小;换句话说,这也表明了本文方法所选取的样本比传统方法

所选取的样本对于分类模型的改进更有价值。

实验随后又在不同的初始训练集上重复进行。结果表明,虽然参与比较的两种选择性采样模式都在一定程度上受到初始训练数据的影响,但是,动态批量选择性采样始终优于传统的批量选择性采样方法,尤其是在初始分类模型性能非常弱的情况下,动态批量选择性采样的优势就更为明显。通过充分挖掘样本之间的相互关系,动态确定度传播算法可以更快地定位到最有信息的样本上,从而有效地提升分类器性能。

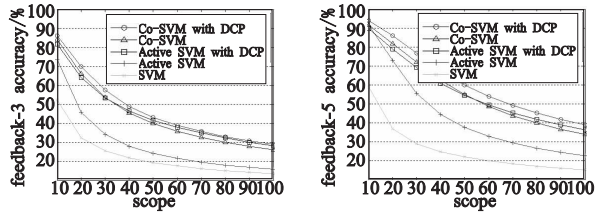
5.2　图像检索

在图像检索中,本文采用了普遍使用的 Coral 图像库[16]作为实验集。在实验集中共有 10 200 幅图像,分为 102 个不同语义类别(如老虎、汽车、人等),每一类包含 100 幅图像。在实验中,每类图像的前 10 幅(共 1 020 幅图像)作为查询图像用来测试图像检索的性能,最终获得的准确率是利用这1 020 幅图像进行查询所得准确率的平均值。

本文使用颜色和纹理特征来描述图像,其中颜色特征由 125 维的颜色直方图和 6 维的颜色矩构成,而纹理特征则利用 3 层的离散小波变换之后在 10 个子带上分别取均值和方差构成一个 20 维的特征。

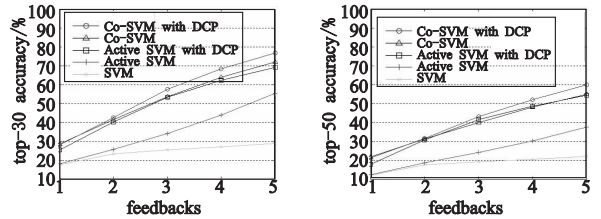
实验将动态确定度传播算法分别应用到 SVM_{Active} 和 Co-SVM 中,并与采用批量选择性采样的原始 SVM_{Active} 和 Co-SVM 进行比较,同时参与比较的还有非主动学习方法。为了公平起见,参与比较的各种方法均使用 RBF 核的 SVM 分类器,相关参数通过交叉验证决定。在每一轮相关反馈中,各种方法均选取相同数目的图像进行标注。

图 7 给出了分别在 3 轮和 5 轮相关反馈之后,返回的前 10 ~ 100个结果的准确率,横坐标 x 表示前 x 个返回结果,纵坐标表示准确率。图 8 给出了前 30 和前 50 个返回结果范围内,准确率随相关反馈轮次变化的趋势,横坐标 x 对应于第 x 次相关反馈,纵坐标表示准确率。



(a) 3轮相关反馈之后的准确率　　(b) 5轮相关反馈之后的准确率

图7　前n(n=10, 20, …, 100)个返回结果的准确率



(a) 前30个返回结果的准确率　　(b) 前50个返回结果的准确率

图8　经过k(k=0, 1, …, 5)轮相关反馈之后的准确率

实验结果分析如下(其中“ > ”表示“优于”):

a)采用了主动学习的相关反馈性能明显优于非主动学习的方法(曲线○、△、□、+ > ×),说明在主动学习中分类模型可以有针对性地对未标注样本进行选取,相对于只能被动接收标注信息的非主动学习方法,其效果更好。

b)在主动学习中,根据所采用的选择性采样模式不同,其最终结果也有明显差别。采用动态批量选择性采样模式的 Co-SVM 和 SVM_{Active} 检索结果好于采用批量选择性采样的原始 Co-SVM 和 SVM_{Active}(曲线○ > △、□ > +),说明在动态批量选择性采样过程中考虑了先后标注的样本之间的关系,有效降低了冗余度。

c)在主动学习中加入多视角学习之后,分类性能进一步提高。融合了多视角学习的主动学习检索结果好于单纯的主动学习(曲线○ > □、△ > +),说明多视角学习可以成为主动学习的有效补充。

d)采用了动态批量选择性采样模式之后,SVM_{Active} 的性能提升(曲线○ > △)比 Co-SVM(曲线□ > +)更为显著,说明在分类模型性能较弱的情况下,动态确定度传播算法的优势更为明显。

6　结束语

在基于内容的图像检索中,相关反馈是获取用户查询意图和偏好的一种重要的人机交互手段。为了使用户仅标注少量的图像便可显著提高图像检索的性能,可以采用主动学习的方法,由系统有针对性地选取最有信息的图像让用户标注。在主动学习中处于核心地位的是选择性采样方法的设计。本文针对传统采样模式的缺陷,提出了一种动态批量选择性采样的模式。该模式采取“逐一标注,批量训练”的流程,综合利用当前分类模型和先前标注样本这两方面的信息对后续样本的选取进行指导,从而有效地克服了批量选择性采样和逐一选择性采样这两种模式的缺陷。基于动态批量选择性采样的模式,提出了动态确定度传播算法,有效地提高了所选取样本的信息量,在保持样本标注量不变的前提下显著改善了图像检索的性能。需要说明的是,本文提出的样本选取方法是一种通用的方法,其应用不仅仅局限于图像检索的相关反馈领域,还可以运用到其他涉及选择性采样的机器学习算法中。

参考文献:

[1] SMEULDERS A W M, WORRING M, SANTINI S, *et al*. Content-based image retrieval at the end of the early years[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1349-1380.

[2] LEW M S, SEBE N, DJERABA C, *et al*. Content-based multimedia information retrieval: state of the art and challenges[J]. ACM Trans on Multimedia Computing, Communications, and Applications, 2006, 2(1): 1-19.

[3] RUI Yong, HUANG T S, CHANG S F. Image retrieval: current techniques, promising directions, and open issues[J]. Journal of Visual Communication and Image Representation, 1999, 10(1): 39-62.

[4] RUI Yong, HUANG T S, ORTEGA M, *et al*. Relevance feedback: a power tool for interactive content-based image retrieval[J]. IEEE Trans on Circuits and Systems for Video Technology, 1998, 8(5): 644-655.

[5] ZHOU X S, HUANG T S. Relevance feedback in image retrieval: a comprehensive review[J]. Multimedia Systems, 2003, 8(6): 536-544.

(下转第 1937 页)

(上接第 1933 页)

[6] COHN D A, GHAHRAMANI Z, JORDAN M I. Active learning with statistical models[J]. Journal of Artificial Intelligence Research, 1996,4(1):129-145.

[7] McCALLUM A, NIGAM K. Employing EM in pool-based active learning for text classification[C]//Proc of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1998: 350-358.

[8] TONG S, CHANG E. Support vector machine active learning for image retrieval[C]//Proc of the 9th ACM International Conference on Multimedia. New York: ACM Press,2001: 107-118.

[9] ZHOU Zhi-hua, CHEN Ke-jia, JIANG Yuan. Exploiting unlabeled data in content-based image retrieval[C]//Proc of the 15th European Conference on Machine Learning. Berlin: Springer, 2004: 525-536.

[10] CHENG Jian, WANG Kong-qiao. Active learning for image retrieval with Co-SVM[J]. Pattern Recognition,2007,40(1): 330-334.

[11] MUSLEA I, MINTON S, KNOBLOCK C A. Active learning with multiple views[J]. Journal of Artificial Intelligence Research, 2006,27(1):203-233.

[12] ZHU Xiao-jin. Semi-supervised learning literature survey, TR1530[R]. Wisconsin: University of Wisconsin-Madison, 2008.

[13] MUSLEA I, MINTON S, KNOBLOCK C A. Selective sampling with redundant views[C]//Proc of the 17th National Conference on Artificial Intelligence. California: AAAI Press, 2000: 621-626.

[14] ZHANG Tong, OLES F J. A probability analysis on the value of unlabeled data for classification problems[C]//Proc of the 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann,2000:1191-1198.

[15] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C]//Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2001: 585-591.

[16] DUYGULU P, BARNARD K, FREITAS J, *et al*. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary [C]//Proc of the 7th European Conference on Computer Vision. London: Springer-Verlag,2002: 97-112.

[17] BRINKER K. Incorporating diversity in active learning with support vector machines[C]//Proc of the 20th International Conference on Machine Learning. California: AAAI Press, 2003: 59-66.