ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.5, May 2009, pp.1301−1312
doi: 10.3724/SP.J.1001.2009.03274

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向草图检索的小样本增量有偏学习算法*

梁 爽[1,2], 孙正兴[1,2+]

[1](南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

[2](南京大学 计算机科学与技术系,江苏 南京 210093)

Small Sample Incremental Biased Learning Algorithm for Sketch Retrieval

LIANG Shuang[1,2], SUN Zheng-Xing[1,2+]

[1](State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

[2](Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

+ Corresponding author: E-mail: szx@nju.edu.cn, http://cs.nju.edu.cn/szx/

Liang S, Sun ZX. Small sample incremental biased learning algorithm for sketch retrieval. *Journal of Software*, 2009,20(5):1301−1312. http://www.jos.org.cn/1000-9825/3274.htm

Abstract: This paper proposes an algorithm named Small Sample Incremental Biased Learning Algorithm to solve three difficulties of relevance feedback in sketch retrieval, including small sample issue, asymmetry of training data and real-time requirement. The algorithm combines active learning, biased classification and incremental learning to model the small sample biased learning problem in relevance feedback process. Active learning employs uncertainty sampling to choose the best labeling samples, so that the generalization ability of classifier is maximized with the limited training data; Biased classification constructs hyperspheres to treat positive and negative data differently, which distinguishes the user's target class accurately; Newly labeled samples in each feedback loop are used to train the classifier incrementally to reduce the training time. Incremental learning also collects training data to further alleviate the small sample problem. Experimental results show that this algorithm improves the performance of sketch retrieval. And it can be well extended to other retrieval domains like CBIR (content based image retrieval), 3D retrieval, and so on.

Key words: sketch retrieval; relevance feedback; small sample incremental biased learning; active learning; biased classification; incremental learning

摘 要: 为了解决草图检索相关反馈中小样本训练、数据不对称及实时性要求这 3 个难点问题,提出了一种小样本增量有偏学习算法.该算法将主动式学习、有偏分类和增量学习结合起来,对相关反馈过程中的小样本有偏学习问题进行建模.其中,主动式学习通过不确定性采样,选择最佳的用户标注样本,实现有限训练样本条件下分类器泛化能力的最大化;有偏分类通过构造超球面区别对待正例和反例,准确挖掘用户目标类别;每次反馈循环中新加入的

* Supported by the National Natural Science Foundation of China under Grant Nos.60721002, 60373065, 69903006 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z334 (国家高技术研究发展计划(863)); the Program for New Century Excellent Talents in University of China under Grant No.NCET-04-0460 (新世纪优秀人才资助计划)

Received 2007-03-06; Accepted 2008-01-29

样本则用于分类器的增量学习,在减少分类器训练时间的同时积累样本信息,进一步缓解小样本问题.实验结果表明,该算法可以有效地改善草图检索性能,也适用于图像检索和三维模型检索等应用领域.

关键词: 草图检索;相关反馈;小样本增量有偏学习;主动式学习;有偏分类;增量学习

中图法分类号: TP181 **文献标识码:** A

手绘草图以其交互灵活性和创新性成为新一代的数字媒体数据类型.近年来,PDA,TabletPC 等电子手写输入设备的快速普及进一步推动了手绘草图及其应用领域的发展[1],草图检索(sketch retrieval)也随之成为学术界和工业界的一个热点研究课题[2].

与 CBIR(content based image retrieval)相比,草图与图像在内容信息上有 3 个方面的区别:首先,图像的基本组成单元为像素,而草图则由笔划或图元组成,包含更多语义信息;其次,图像以点阵形式存储,而草图一般以结构化的形式存储,即存储其笔划或图元信息,所以草图需要不同于图像的处理技术;最后,草图关注图形的结构信息,缺少图像中相对丰富的颜色、纹理等特征,这导致草图和图像的内容有不同的侧重点.因此,CBIR 忽略草图的结构化特征及其笔划所表达的语义信息,无法完整捕捉草图检索的查询概念.目前,采用符合草图特性的内容表示实现草图检索成为该领域研究的主流:Leung[3]利用草图的多形式表示来消除用户绘制方式对草图表示的影响;Fonseca[4]利用笔划的层次结构、形状信息和高维度的索引来检索复杂的图形.我们在前期工作[5,6]中对草图检索的内容表示问题进行了研究,采用草图的全局几何特征和图元之间的拓扑结构关系表征草图的内容,同时引入相关反馈技术优化查询结果.

即使采用最优的特征提取和相似度匹配算法,已有的草图检索方法也不能完全达到用户满意的效果,即系统能够返回给用户的相关草图的数量有限.这是因为正如许多其他信息检索问题一样,草图检索面临两个主要的挑战:其一是人与系统在理解上的语义鸿沟.计算机系统只能提取形状、曲率、笔速、空间关系等草图底层的视觉特征,但用户的需求和理解是高层的概念,他们根据头脑中的直觉印象理解草图,无法获取准确的底层视觉特征.这就产生了人与计算机在理解上的差异,导致计算机无法准确地捕捉用户头脑中的检索意图;草图检索的第 2 个难点是用户对草图内容的动态理解.对同一幅草图而言,不同的用户对其有不同的理解,即使同一个用户在不同的时期也会产生不同的理解.因此,检索系统几乎无法找到能够适合所有情况的特征提取或相似度匹配算法.为了减小人与人之间和人与计算机之间在理解上的差异并进一步提高检索的性能,草图检索系统必须在检索过程中不断地对查询和相似度匹配进行优化.这要求检索系统提供自适应的在线学习或分类,通过交互实时地捕捉用户的主观检索意图,并在用户的主观理解和计算机感观的底层视觉特征之间建立映射,增近人和计算机之间的距离.

为了更加准确地挖掘用户的查询兴趣,本文提出了小样本增量有偏学习(small sample incremental biased learning)算法,并将其应用于草图检索中的相关反馈.实验结果表明,该算法是有效的.

本文第 1 节介绍相关工作的研究现状,并分析存在的问题.第 2 节描述草图检索中小样本增量有偏学习算法的总体流程.第 3 节和第 4 节具体阐述小样本增量有偏学习算法中的关键技术.第 5 节将给出实验设计及其结果分析.第 6 节总结全文并讨论未来的研究方向.

1 相关工作

相关反馈是通过引入用户主观评价来在线提高检索性能的最自然的策略.在反馈过程中,用户针对当前的查询结果给出相关/正例或不相关/反例的反馈信息,系统通过分析处理用户标记的反馈信息来在线优化搜索策略,以得到更好的结果集,进而实现个性化的、结合用户主观性认知的检索.相关反馈的概念最初在文本检索中提出,在 CBIR 领域也有大量研究.但图像的特征类别较多,颜色、纹理等物理信息丰富,而草图的内容表示只关注形状和结构信息,这增加了用户兴趣挖掘的难度和复杂性.仅采用形状信息表征用户的检索意图,具有更大的模糊性和多变性,导致系统难以在用户和底层特征之间建立准确的映射.因此,草图检索中需要更加有效的相关反馈机制来完成检索任务.

相关反馈技术已由最初的启发式权重调整发展到近期的在线学习机制[7].基于启发式的方法考虑不同特征之间的相关性,不断地对查询进行优化或调整特征加权的方案,例如查询点位移和权重调整策略[8]、自组织映射(self-organizing map,简称 SOM)[9]以及 Boosting 技术[10]等.这些方法速度快且鲁棒性好,但由于其前提是能够找到合适的参数,启发式方法的最优性难以得到保证.近年来,研究者从优化学习的角度考虑相关反馈,将其看作一个在线的学习分类问题,并采用流行的机器学习方法来实现相关反馈.例如,MacArthus 等人[11]和 Wu 等人[12]分别将决策树和最近邻分类器引入相关反馈.Cox 等人[13]则采用贝叶斯分类器对数据库中的图像进行分类,并开发了相应的 CBIR 系统 PicHunter.Li 等人[14]采用基于线性规划(linear programming,简称 LP)的相关反馈挖掘用户兴趣,并在反馈的同时进行特征选择.一些研究者[15]将支撑向量机(support vector machine,简称 SVM)引入相关反馈,采用 SVM 区分相关\不相关的结果集或学习特征的权重.SVM 具有小样本学习和泛化能力好的特点,成为实现相关反馈的最有前景的技术之一[16].

但由于相关反馈特有的应用场景,其中存在 3 个主要的难点问题.首先是相关反馈的小样本问题.在相关反馈环境中,用户所能标记的样本的数量(一般每次小于 20 个)有限.在这种小样本的训练环境下,一些学习算法的稳定性得不到保证,以至于学习结果无法提供有意义的分类标准,导致分类器泛化能力不高,分类性能差.其次是正反例数据的不对称性.一些机器学习方法将相关反馈看作严格的二值分类问题,同样对待正例和反例[17].有理由假设正例在特征空间中以某种线性或非线性的方式聚类,但反例却不一定可以聚类,因为它们来自不同的类别,且其有限数量不能代表所有反例在特征空间的真正分布.把所有的反例全部分为一类,可能会破坏系统的鲁棒性并降低检索的性能,这在相关反馈的小样本训练环境中体现得更加明显.最后,由于用户与系统进行实时的交互,反馈算法必须保证检索系统的实时性.也就是说,反馈算法必须足够快,并避免复杂的计算.

本文针对相关反馈的上述 3 个难点,提出采用小样本增量有偏学习的相关反馈算法,并将其引入草图检索.该算法能够实时地捕捉用户的查询兴趣,逼近用户的检索意图,在内容匹配的基础上改善检索的性能,进而拉近人的理解与计算机特征处理之间的距离.下文将对小样本增量有偏学习算法的原理和具体设计进行详细介绍.

2 草图检索中的小样本增量有偏学习算法设计

本节首先介绍草图检索的总体结构,再给出小样本增量有偏学习算法的基本思想和算法流程.

2.1 草图检索系统

图 1 给出了草图检索的结构图,其中包括相似度匹配和相关反馈两个主要模块.相似度匹配通过计算查询草图与数据库中草图的相似度,返回初始的检索结果集.在用户提交查询后,草图检索系统分别对输入的查询草图和数据库中的草图进行特征抽取.特征提取是草图相似度匹配的关键,它直接影响匹配效果,也为后续相关反馈建立基础.草图的特征分为全局特征和结构特征两类:全局特征从整体上反映草图概念,接近人类的视觉感知;结构特征则将草图看作一组构成单元,反映草图内在的结构特性.我们抽取 4 种全局形状描述子和基本图元(直线段、弧线段和椭圆)之间的 8 种空间拓扑关系[6],这些特征具有平移、旋转和缩放等几何不变性,且可以组成一个 12 维的特征向量,用于计算草图的相似度.这就将传统检索中的图匹配问题简化为特征向量之间的相似度计算,假设数据库中共收集了 m 幅草图,则计算所有草图相似度的时间和空间复杂度均为 $O(m)$,从而在极大程度上降低了计算的复杂度,有效地提高了检索的效率.

获得初始候选结果集后,草图检索系统通过调用相关反馈模块引入用户的主观评价.用户首先对候选集进行判定,如果返回的结果令用户满意,则检索成功;否则,检索系统调用小样本增量有偏学习算法进行相关反馈,优化系统性能.主动式学习首先通过采样选择标注样本集合,并返回给用户要求其标记.用户标记的结果被看作训练样本,用来训练有偏分类器.分类器通过学习用户的标注信息,挖掘相关和不相关结果的特性,从而判断用户关心的草图在特征空间中分布的区域.训练后的分类器将数据库中的所有草图分为相关和不相关两部分,并将相关性较高的草图作为优化结果返回给用户.整个反馈过程不断循环,每次新标记的样本将用于增量训练分类器,并重新对特征空间进行划分.随着反馈次数的增多,检索结果不断改进,直至检索到用户满意的草图为止.

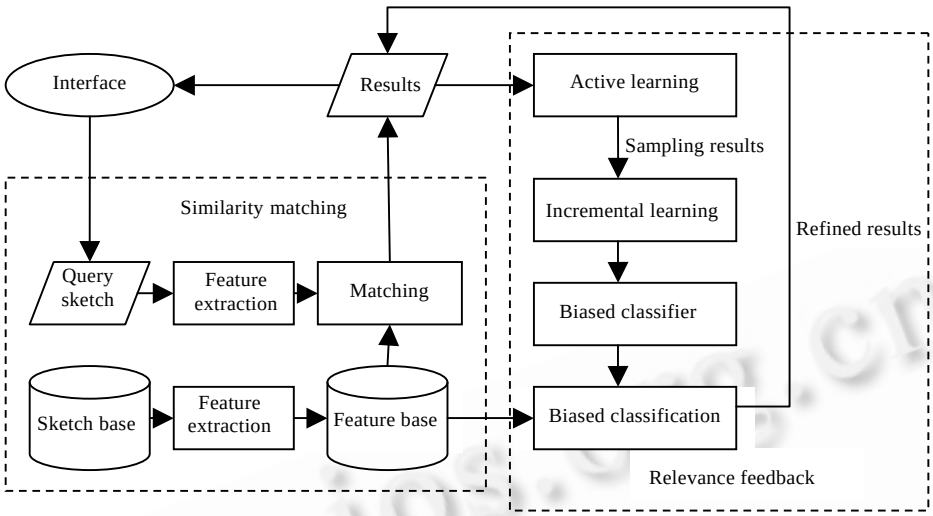

Fig.1 Structural framework of sketch retrieval

图 1 草图检索结构图

2.2 小样本增量有偏学习

考虑到相关反馈的应用特性,可将相关反馈看作小样本有偏分类/学习问题.其中,用户只能提供有限的训练样本,且只对若干类对象中的 1 类感兴趣.针对相关反馈中小样本训练、数据不对称和实时性要求等难点,我们结合采用主动式学习、有偏分类与增量学习的思想,提出了面向草图检索的小样本增量有偏学习算法,以增强分类器的稳定性并提高检索的性能.

小样本增量有偏学习算法包括主动式学习、有偏分类和增量学习 3 个部分:(1) 主动式学习根据分类器的需求选择"最富信息"的训练样本,回显给用户并要求其标记,进而在小样本的训练环境下最大化分类器的泛化能力;(2) 有偏分类机制区别对待相关和不相关的样本,通过描述正例在样本空间中的聚类区域直观地将相关样本与其他类别的草图进行区分,更准确地捕捉用户的目标空间;(3) 增量学习采用用户在每轮相关反馈中新标记的样本增量训练分类器,通过积累样本增强分类器的稳定性,缓解小样本问题.同时,增量学习减少了重新训练分类器的时间,也解决了实时性要求.下面给出小样本增量有偏学习的完整算法流程.

算法 1. 小样本增量有偏学习算法.

假设草图的特征表示为 $\boldsymbol{x}=\{x_1,x_2,\ldots,x_m\}$,其中,$\{x_1,\ldots,x_m\}\in\Re^m$ 为草图的 m 维特征属性,则草图检索中小样本增量有偏学习算法的具体流程如下:

第 1 步.给定一个查询,首先进行草图相似度匹配,按照相似度(similarity)递减的顺序,将初步检索结果 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_k\}$返回给用户.

第 2 步.如果用户检索到目标草图,即候选结果集中包含用户的目标草图,则检索过程结束;否则,接下一步.

第 3 步.用户对回显的草图数据进行标记,若 $\boldsymbol{x}_i$ 相关,则其标记 $y_i=1$;否则 $y_i=-1$,得到训练数据集 $\boldsymbol{TS}=(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_k,y_k)$.

第 4 步.如果是第 1 次反馈,则采用用户标记的样本集 ***TS*** 在线生成有偏分类器 *BC*;否则,采用 ***TS*** 对 *BC* 进行增量学习.

第 5 步.采用有偏分类器 *BC* 对数据库中的草图进行区分,得到相关和不相关的草图集.并计算草图的相关性,将前 k 个最相关的草图作为优化结果集返回给用户.

第 6 步.如果结果集包含用户的目标样本,则检索结束;否则,接下一步.

第 7 步.采用主动式学习算法对草图数据库进行采样.计算草图样本的信息量,将前 k 个"最富信息"的样本

返回给用户,以进行下一轮反馈.

第 8 步.转第 3 步.

为了体现系统的有效性并提高检索的效率,小样本增量有偏学习算法的回显过程分为两个阶段.首先,检索系统将最相关的检索结果返回给用户,使用户总是先得到最匹配的草图检索结果,这体现了草图检索系统的有效性;接下来,如果用户对检索结果不满意,则进行反馈.系统采用主动式学习对草图数据库进行采样,将“最富信息”的样本回显给用户并要其标记,从而在小样本的训练环境下最大化有偏分类器的泛化能力,并减轻用户的标记负担,这就提高了小样本增量有偏算法的学习效率.同时,该算法采用每轮反馈新标记的样本增量训练有偏分类器,既积累了训练样本信息,又节省了重新训练分类器的时间,能够满足相关反馈实时性的要求.

此外,因为小样本增量有偏学习算法独立于检索内容的特征表示,所以该算法具有很强的扩展性.只要检索对象的内容可以表征为特征向量的形式,即可使用小样本增量有偏学习算法进行相关反馈来提高检索的性能.因此,该算法同样可以应用到其他检索领域,例如图像检索、三维模型检索等.

3 相关反馈中的小样本学习问题

小样本问题是影响相关反馈的主要难点之一.在相关反馈的标记过程中,用户希望标记的样本数量尽可能地少,标记数据与整个数据库中的未标记数据相比往往是微乎其微的,这导致系统无法采集到足够的样本对分类器进行训练.在小样本的训练环境下,仅采用有监督学习的分类器往往无法取得很好的泛化能力.如何利用有限的样本条件来提高分类器的学习性能也就成为一个热点的研究问题.在本文中,我们采用主动式学习解决小样本问题,通过选择“最富信息”的样本降低对反馈算法样本数量的要求.

3.1 主动式学习

通常来讲,检索系统必须满足两个要求,即检索系统必须准确地捕捉用户的查询概念,并在有限的小样本训练条件下快速地获取用户的查询意图.这是因为在与系统的实时交互过程中,用户希望在尽可能少的标记负担和反馈次数下,尽快寻找到目标草图.在这种小样本学习的环境下,训练样本的好坏将严重影响分类器的学习性能和泛化能力.因此,不同的样本选择策略将产生不同的学习结果.如何有效地对样本空间进行采样,并选择最优的回显样本成为相关反馈的关键.

检索系统自主选择回显样本的过程,可以看作系统的主动式学习(active learning)[18].主动式学习是分类器根据现有的数据分布情况,主动查询特定样本类别,以取得最优学习效果的采样过程.换句话说,也就是学习分类器自主地确定最优的训练样本的过程.下面给出主动式学习的定义.

定义 1(主动式学习). 假设数据集合 D 由标记样本集 L 和未标记样本集 U 组成,样本的特征空间为 X^n,则主动式学习 l 是一个二元组(C,S),即

$$l=(C,S),\text{其中,}\begin{cases}C: X^n \to \{-1,+1\}\\ u = S(U,C,L), u \subset U\end{cases}.$$

$C:X^n\to\{-1,+1\}$为由标记数据 L 训练所得的分类器;采样算法 S 决定了样本的回显策略,它根据当前的标记数据 L 选择未标记数据 U 中最优的回显样本集 u,并向用户查询其标记.

主动式学习的目的是获取最大的信息增益或在最大程度上减小决策过程中的不确定性,从而在最小的训练样本规模下获得最高的学习性能.其出发点是系统可以通过合理地选择标注样本来减轻对训练数据的要求.在特殊情况下,其计算复杂度也会降低.例如,某些 NP 完全的学习问题可能在多项式复杂度内解决.在相关反馈中,系统首先调用采样算法 S,主动选择一定数量的回显样本,并要求用户标记.反馈系统再使用新的标记数据对分类器 C 进行更新,从而不断优化分类器的泛化能力.

主动式学习和传统的被动式方法的本质区别在于选择回显标记样本的策略不同,即系统获取训练数据的方式不同,如图 2 所示.在传统的被动式相关反馈学习算法中,检索系统根据用户的要求,按照相关性的高低返回候选对象,无法自主地选择回显的结果样本,而仅仅是对用户的需求作出反应.这种被动式的样本选择策略是固

定不变的,系统总是返回给用户最想得到的结果,使用户得到最相关的候选结果集,但学习分类器却只能被动地接受训练数据.虽然用户在每次反馈中得到了最好的结果,但这些标记样本对分类器来讲并不一定是最优的训练数据集.在主动式学习中,为了取得更好的分类性能和检索效果,分类器被赋予自主选择和收集数据样本的能力.在反馈中,分类器或检索系统可以根据自身的需要向用户发问.也就是说,分类器可以返回最想要用户标记的对象,从而要求用户为其提供最具价值的样本信息,减少用户不必要的标记负担.

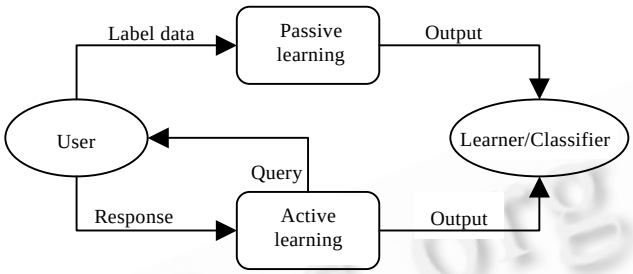

Fig.2 Difference between active and passive learning

图 2 主动式学习与被动式学习的区别

3.2 选择性采样算法

更确切地说,在相关反馈中遇到的主动式学习形式可称作选择性采样(selective sampling).实际上,在机器学习中,不同样本对分类器的作用是不一样的.这种作用称为样本的“信息量”,样本含有的信息量越大,对分类结果的确定越重要.为了最大化相关反馈中有限样本的信息量,选择性采样通过评估未标记的样本数据库,主动选择“最富信息”的样本作为回显样本,进而减轻小样本训练环境对分类器学习的约束.

选择性采样的相关工作主要可以分为两类:基于委员会的投票选择(query by committee,简称 QBC)和不确定性采样(uncertainty sampling).QBC 的方法首先由 Seung 等人[19]提出,该方法训练一系列分类器委员,并选择分类器之间判别争议最大的样本作为下一个查询样本.Freund 等人[20]给出了 QBC 方法的理论证明,即在一定的假设下,有效地选择一些值得标记的数据,可以在标记少量样本的条件下达到标记所有数据的训练效果;选择性采样的第 2 类方法是不确定性采样.Lewis 和 Gale[21]给出了不确定性采样的主要思想,即只采用一个分类器来对样本进行判别,并同时给出对所有未标记样本判别的不确定性.确定性最低的样本,也就是分类器最不确定的样本将被选择作为下一个返回的待标记样本.Lewis 等人指出,不确定性采样可以大幅度地减小训练数据的规模,能够有效地应用于小样本的训练环境.近来,Tong 和 Chang[18]在 CBIR 中使用了 SVM 不确定性采样的方法,取得了很好的应用效果.

在不确定性采样中,检索系统将不确定性最大的样本作为“最富信息”的样本返回给用户进行标记,从而尽可能快速地排除分类器学习的不确定性因素,提高分类器的泛化能力.直观认为,“最富信息”的样本是距分类边界最近的样本.这是因为为了提高分类器在决策上的可信度,我们希望分类器的分类间隔尽可能地小.而在分类边界上或其附近的样本点是最难被分类的模式,分类器对其判定最不可靠.对这些样本点的标注可以有效地提高分类器决策的可信度,快速地缩小分类器不确定的区域范围,从而提高其泛化能力.因此,这些距离分类边界最近的样本点也就是求解分类任务中的“最富信息”的样本.将它们回显给用户,并要求他/她进行正确的标记,可以在最大程度上对分类器进行优化.

此外,增量学习也是缓解小样本学习的有效方法,同时也节省了重新训练分类器的时间,加快了相关反馈的响应速度.在每轮相关反馈过程中,用户新标记的样本被用于增量训练分类器.随着相关反馈次数的增加,训练样本的信息量会随之增大,分类器的性能也越来越稳定.

4 相关反馈中的不对称问题

相关反馈中正反例数据的不对称性要求检索系统分别对待相关和不相关的样本.本文采用有偏分类对相

关反馈进行建模,更加有效地估计正、反例样本在特征空间中的分布.考虑到 SVM 泛化能力好的特点,我们选择有偏 SVM(biased SVM,简称 BSVM)作为有偏分类的分类器,并进行不确定性采样.

4.1 有偏分类

传统的相关反馈采用二值分类的方法,通过建立超平面对样本空间进行划分.但二值分类同等对待正例和反例,无法处理数据不均衡的现象,导致系统的鲁棒性低,无法满足相关反馈中的有偏分类需求,因而不适合解决相关反馈问题.为了区分用户感兴趣的草图类别,本文将草图检索中的相关反馈过程看作一个在线的有偏分类[17],其中用户只关心若干类数据中的 1 类,或只对其中 1 个类别比较偏心.下面给出有偏分类的定义:

定义 2(有偏分类). 设样本空间Ω由类别为$\{\omega_1,\omega_2,\ldots,\omega_n\}$的 n 类样本组成,x 表示Ω中的一个样本,则有偏分类是指找到一个判别函数 $BC(x)$,$\forall x\in\Omega$,使得

$$BC(x)=\begin{cases}1, & \text{if } x\in\omega_i\\ -1, & \text{if } x\notin\omega_i\end{cases},$$

其中,ω_i 是用户感兴趣的类别.

本文选择 BSVM 作为有偏分类的分类器.通常,基于 SVM 的相关反馈技术采用经典的二值 SVM[18]或单类 SVM(one SVM,简称 1SVM)[22]分类器.然而,经典 SVM 将相关反馈看作是严格的二值分类,不考虑正反例数量不对称的情况;单类 SVM 将相关反馈看作估计正例样本在特征空间中的分布情况,在一定程度上解决了有偏分类问题.它采用相关的样本训练 1SVM 分类器,在特征空间中寻找一个最优的超球面,使其包含尽可能多的正例.该超球面所包含的区域即为用户相关的目标草图在特征空间中的分布范围.但由于 1SVM 不使用反例信息,在反例数量比例很大时反馈结果不够理想.为了解决相关反馈中的有偏分类问题,并结合使用正、反例信息,我们选择 BSVM[23]作为有偏分类的分类器.BSVM 将经典 SVM 与 1SVM 相结合,其基本策略是通过一对超球面来描述数据,其中,内超球面是包含正例的最小超球面,而外超球面则是不包含反例的最大超球面.换句话说,内超球面包含尽可能多的正例,而外超球面则将尽可能多的反例排斥在外.因此,接下来的目标就是寻找最优超球面,使其在包含正例的同时,将反例排斥在外.

图 3 给出了二维输入空间中 BSVM 超球面的示意图,其中,实心圆和十字分别表示相关和不相关的草图,而虚线表示的超球面则是我们待寻找的决策超球面.

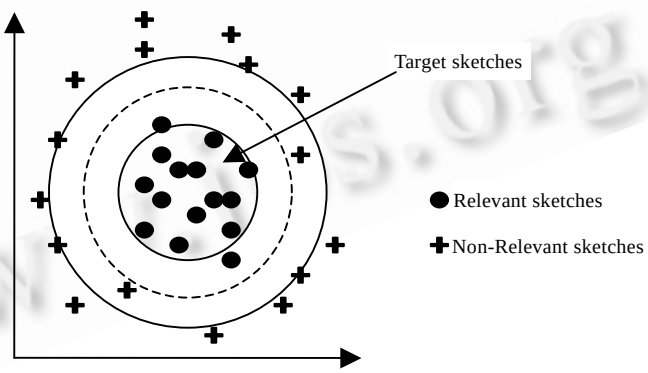

Fig.3 BSVM hypersphere in 2D space

图 3 二维空间中 BSVM 的超球面

4.2 BSVM不确定性采样

本文将 BSVM 作为有偏分类器,对数据库中的未标记草图进行不确定性采样,降低分类器对训练数据数量的要求,其算法流程如下:

算法 2. BSVM 不确定性采样算法.

假设数据集合 D 由标记样本集 L 和未标记样本集 U 组成,样本的特征空间为 X^n.C:$X^n\to\{-1,+1\}$为 BSVM

有偏分类器,$label:X\to\{-1,+1\}$为用户标记函数,则 BSVM 不确定性采样的过程如下:

第 1 步. 根据当前的已标注数据集 L,计算不确定性最大的未标记样本集,$u=Uncertainty\ Sampling(U,C,L)$.

第 2 步. 将 u 返回给用户.

第 3 步. 对 u 中每个样本 x:

要求用户提供标记,$l=label(x)$,

更新已标记的样本集,$L=L\cup\{(x,l)\}$.

第 4 步. 更新未标记样本集,$U=U-u$.

第 5 步. 更新 BSVM 分类器,$C=BSVM(L)$.

BSVM 的最优超球面可以通过最优问题求解.设训练数据集为

$$(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)\in\Re^m\times Y,Y=\{-1,+1\} \tag{1}$$

其中,n 为训练数据的个数,m 为特征空间的维数,y 为训练数据的标记.在 BSVM 不确定性采样过程中,各样本点到 BSVM 超球面的距离可计算如下:

$$d(\boldsymbol{x})=|R^2-\|\Phi(\boldsymbol{x})-\boldsymbol{c}\|^2| \tag{2}$$

其中,$\Phi(\boldsymbol{x}_i)$为映射函数,$\boldsymbol{c}$ 和 R 分别为最优超球面的球心和半径.样本点到决策超球面的距离表征了样本的可信度:该距离越大,样本分类的可信度越高;反之,则可信度越低,不确定性也就越高.因此,与决策超球面距离最近的样本即为“最富信息”的样本.根据样本点到分类边界的距离,我们即可按照下面的公式计算每个样本点所蕴含的信息量:

$$information(\boldsymbol{x})=\begin{cases}0, & \text{if } d(\boldsymbol{x})\ge\omega\\ 1-\dfrac{d(\boldsymbol{x})}{\omega}, & \text{if } d(\boldsymbol{x})<\omega\end{cases} \tag{3}$$

其中,阈值ω可根据经验设定.如果样本到分类边界的距离 $d(\boldsymbol{x})$超过阈值,则信息量置 0;否则,进行归一化.这种归一化方法直观、有效,可以在不改变样本信息量顺序的前提下将其归一化到区间[0,1]上.本文选择信息量较高的草图作为回显标注集,并按照信息量递减的顺序返回给用户.

用户进行标注后,系统采用用户的标注信息更新训练 BSVM 分类器,并区分相关和不相关的样本.样本的相关性可以根据 BSVM 的目标函数进行计算如下[23]:

$$\begin{aligned}f(\boldsymbol{x})&=R^2-\|\Phi(\boldsymbol{x})-\boldsymbol{c}\|^2\\&=R^2-\|\Phi(\boldsymbol{x})-\sum_{i=1}^{n}\alpha_i y_i\Phi(\boldsymbol{x}_i)\|^2\\&=-k(\boldsymbol{x},\boldsymbol{x})+2\sum_{i=1}^{n}\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x})+R^2-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\end{aligned} \tag{4}$$

其中,$k(\boldsymbol{x},\boldsymbol{x})$为 BSVM 核函数,参数$\alpha_i$为 Lagrange 乘子,可通过二次规划确定其最优值.根据目标函数的符号即可对相关和不相关的草图进行分类:若 $f(\boldsymbol{x})>0$,则待分类的草图是相关的;否则,该草图不相关.此外,$f(\boldsymbol{x})$中的 R 和 $\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 对所有待分类的草图都是相同的,在比较草图的相关性大小时可以消去.因此,本文采用简化的评价函数(公式(5))来表示草图的相关性:

$$relevance(\boldsymbol{x})=-k(\boldsymbol{x},\boldsymbol{x})+2\sum_{i=1}^{n}\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x}) \tag{5}$$

由于相关性高的草图更有可能成为用户感兴趣的目标草图,所以我们计算数据库中所有草图的相关性,按其高低进行排序,将相关性较高的草图作为优化的检索结果集返回给用户.

在新一轮的相关反馈过程中,系统采用用户新标记的样本对 BSVM 分类器进行训练,以降低分类器学习的复杂度.BSVM 增量学习的思想是采用支撑向量(support vector,简称 SV)代替整体数据集进行训练,在保证分类器性能的同时减少训练样本的数量.假设原始训练样本和新加入的样本集合分别为 IS 和 INS,则 BSVM 增量学习可以表示为

$$BSVM=Incremental\_Learning[Sub(IS)\cup INS] \tag{6}$$

其中,$Sub(IS)=IS_{SV}$,表示 IS 中的支撑向量集合.

5 实验设计及分析

为了验证方法的有效性,我们在 Intel P4 PC(2.0G Hz CPU,512MB 内存),Microsoft Windows XP 的环境下设计了原型系统进行实验.

5.1 实验设置

本文共收集了 55 个草图类别,其中包含 30 类工程零件草图和 25 类电器元件草图作为实验数据,如图 4 所示.草图数据分别由 10 个不同的用户绘制,每个用户对每个类别各绘制 2 幅草图,所以每个草图类别共包含 20 幅草图,数据库中共计存有 1 100 幅草图.

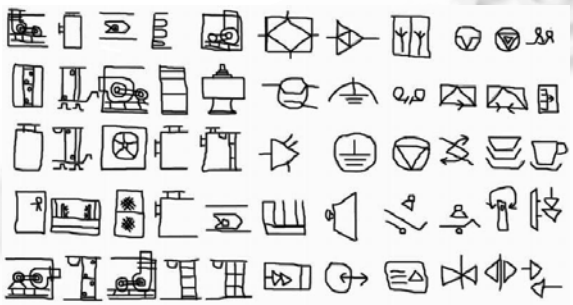

Fig.4 55 sketch data classes for sketch retrieval

图 4 草图检索的 55 个草图数据类

本文采用查准率图来衡量检索的性能,以曲线的形式给出查准率和候选草图结果集的大小之间的关系.查准率是候选集中相关结果的数目与整个候选集中结果数目的比值,其计算如下:

$$precision=\frac{|relevant\cap candidates|}{|candidates|} \tag{7}$$

我们进行多次检索查询,并计算最后的平均查准率值.显然,在查准率图中,曲线越高,检索效果越好,因为在返回相同个数的候选草图的情况下,曲线越高代表相应的查准率越高;反之亦然.

5.2 结果分析

为了验证草图检索小样本增量有偏学习算法的效率和有效性,本文分 3 部分进行实验:首先验证小样本增量有偏学习算法主动式学习的效果,并说明其在相关反馈的小样本训练环境下的有效性;其次,给出小样本增量有偏学习算法与二值分类器的反馈效果比较,以说明小样本增量有偏学习算法中 BSVM 有偏分类机制的有效性;最后验证小样本增量有偏学习算法的效率和实时性.本文使用 LIBSVM 算法库[24]对相关反馈实验进行开发.

5.2.1 小样本增量有偏学习的小样本训练有效性

本文将小样本增量有偏学习算法与无采样机制的 BSVM 相关反馈进行效果比较,以验证小样本增量有偏算法中主动式学习的有效性.为确保各种方法实验结果之间的可比性,所有算法均选择相同的参数和相同的 SVM 核函数.通常情况下,高斯核函数比其他类型的核函数具有更好的效果,因此选择高斯核函数作为 SVM 的核函数.这样,小样本增量有偏学习与 BSVM 相关反馈方法的根本区别仅在于选择回显样本的策略不同.此外,Li 等人[14]的研究表明,基于线性规划(linear programming,简称 LP)的相关反馈可以在小样本学习的环境下取得较好的结果.因此,也将小样本增量学习算法与 LP 相关反馈进行对比,以验证其小样本训练的有效性.为方便起见,在图中采用 SSIB(small sample incremental biased)表示小样本增量有偏学习算法.

图 5 给出了小样本增量有偏学习算法与传统被动式相关反馈的性能曲线.图中的 4 条曲线分别代表小样本

增量有偏学习算法和 BSVM 相关反馈(分别标记 20 和 40 个样本)、LP 相关反馈经过 3 次反馈后的平均查准率.从中可以看出,在同样回显样本数量的情况下,小样本增量有偏学习的性能曲线高于无采样机制的 BSVM 和 LP 相关反馈.这说明小样本增量有偏学习算法由于提供了主动式学习机制,其效果优于普通的被动式相关反馈方法.在无采样机制的相关反馈中,检索系统只返回给用户最相关的样本.这种被动的学习方式不考虑回显样本的信息量,不具备主动选择回显样本的能力,因此无法最优化分类器的性能.而小样本增量有偏学习算法对有偏分类器进行选择性采样,将“最富信息”的样本回显给用户,可以在小样本的训练环境下使分类器达到最优的泛化能力,从而在最大程度上提高草图检索的性能.此外,注意到小样本增量有偏学习算法在 20 个回显样本下与 BSVM 相关反馈在 40 个回显样本下的性能曲线非常接近,这说明在小样本增量有偏学习算法中,用户每次只要标注 20 个样本即可达到在 BSVM 相关反馈中每轮标注 40 个样本的反馈效果.因此,小样本增量有偏学习算法的主动式学习机制使得草图检索系统可以在少量用户标记的情况下取得很高的反馈性能,且所需标注的样本数量远少于传统的被动式相关反馈所需要的训练样本量(前者几乎为后者的一半),在很大程度上减小了用户标记的负担.

图 6 是小样本增量有偏学习算法和 BSVM 相关反馈、LP 相关反馈的查准率随标记样本个数的变化曲线.从图中可以看出,随着标记样本个数的增加,3 条曲线均逐渐升高,即相关反馈方法的性能均逐步提高.这是因为在相关反馈小样本训练的环境下,样本的数量对分类器的训练效果影响很大.随着样本个数的增多,分类器的学习效果越来越稳定,其泛化能力也将逐渐提高.此外,小样本增量有偏学习算法的性能曲线高于 BSVM 和 LP 相关反馈,这说明在同样数量的标记样本的情况下,小样本增量有偏学习算法的泛化能力要好于采用其他机器学习方法的相关反馈.因此,采用小样本增量有偏学习算法中的主动式学习机制可以有效地解决相关反馈中小样本训练的问题.

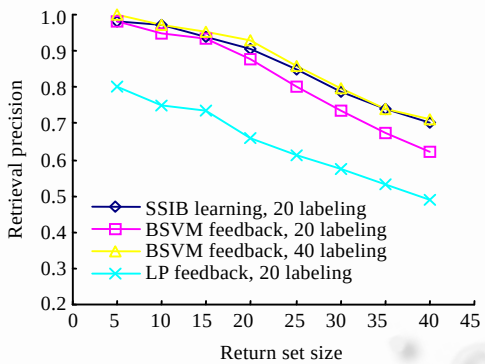

Fig.5 Effectiveness of small sample learning

图 5 小样本学习的有效性

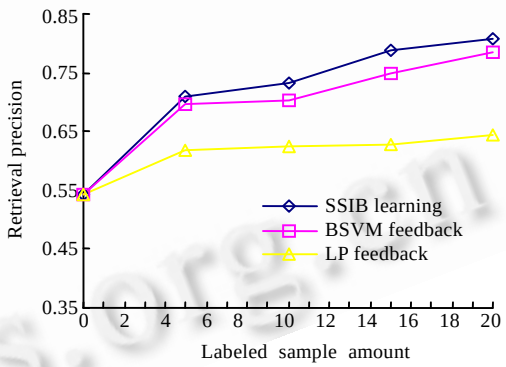

Fig.6 Precision with the labeled sample amount

图 6 查准率随标记样本数量的变化

5.2.2 小样本增量有偏学习的有偏分类有效性

本文将小样本增量有偏学习算法与传统的不具备有偏分类机制的相关反馈技术进行比较,以验证小样本增量有偏学习算法中 BSVM 有偏分类的有效性.我们选择经典 SVM 二值分类、单类 SVM 和 LP 相关反馈与小样本增量有偏学习算法中的 BSVM 有偏分类进行对比.同前,选择相同的 SVM 分类参数和高斯核函数.

图 7 是小样本增量有偏学习算法有偏分类的效果图.图中 4 条曲线分别代表小样本增量有偏学习算法、1SVM 相关反馈、SVM 相关反馈和 LP 相关反馈相对应的查准率曲线.对不同的反馈算法进行比较,可以发现小样本增量有偏学习算法的结果明显优于其他相关反馈效果.这说明,小样本增量有偏学习算法中的有偏分类机制能够更加准确地描述正反例在特征空间中的分布情况,更适合解决相关反馈中的有偏分类问题.因此,采用小样本增量有偏学习算法能够有效地改善草图检索系统的性能.通过不断学习用户的查询习惯和主观查询意图,草图检索系统可以在很大程度上提高检索的精度,满足用户的查询需求.

图 8 绘出了草图检索系统的查准率随反馈次数的变化曲线.从图 8 中可以看出,随着反馈次数的增加,相关

反馈算法均可以提高检索的性能.总体上,小样本增量有偏学习算法的结果曲线最高,这说明该算法比其他相关反馈更能提高检索的精度.而 LP 相关反馈不具备增量训练能力,每轮反馈效果没有较大的提高.此外,由于 SVM 二值分类同等对待正例和反例,无法处理有偏分类问题,只能在反馈后期逐步修正超平面;而 1SVM 相关反馈只利用了相关结果的信息,虽然可以快速地估计相关样本的分布区域,但该方法没有利用不相关的结果信息进一步改善检索精度,导致其反馈效果快速收敛.小样本增量有偏学习算法通过增量训练分类器,逐步提高检索的精度,减小了用户标记负担,加快了相关反馈的响应速度.

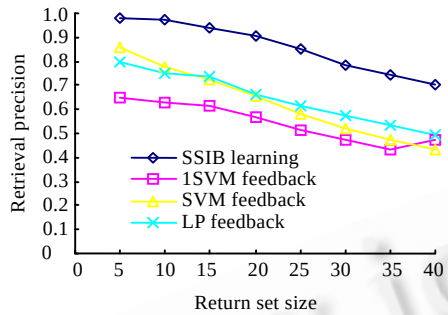

Fig.7 Effectiveness of biased classification

图 7 有偏分类的有效性

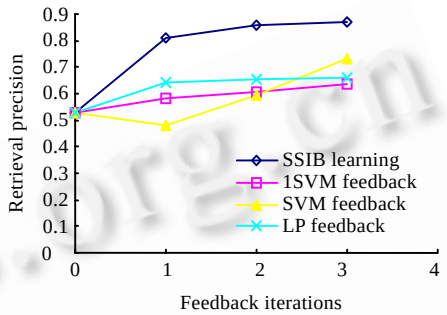

Fig.8 Effectiveness of incremental learning

图 8 增量学习的有效性

5.2.3 小样本增量有偏学习的实时性

此外,响应时间也是评价系统效率的关键因素.在实时交互的草图检索系统中,用户的响应时间越短,草图检索的时延越小,越能满足实时应用需求.采用增量学习的方法,可以在很大程度上减少分类器训练的时间.此外,本文将草图的内容表示成特征向量的形式,从而成功地避免了传统图匹配带来的计算复杂度.在实验中,草图内容匹配的平均响应时间为 0.087s,相关反馈的平均响应时间为 0.112s,均在 1/10 秒左右.因此,本文提出的小样本增量有偏学习算法能够满足草图检索及相关反馈中实时应用的需求.

6 总结与展望

针对相关反馈中小样本训练,正、反例数据不对称以及实时性要求这 3 个难点问题,本文提出了小样本增量有偏学习的相关反馈算法,并将其引入草图检索.该算法具有以下几个方面的特点:为了解决相关反馈中的小样本训练问题,小样本增量有偏学习算法采用主动式学习方法,将“最富信息”的样本回显给用户,进而在小样本训练的条件下最大化分类器的泛化能力;为解决相关反馈中正、反例数据不对称的问题,小样本增量有偏学习算法采用有偏分类机制,通过构造 BSVM 超球面来更准确地捕捉用户的目标草图在特征空间中的区域;为满足相关反馈的实时性要求,采用分类器增量学习的方式,减少了重新训练分类器的时间.此外,通过基于特征向量的内容表示方法来避免传统的图匹配方法所带来的复杂度.实验验证了该算法在提高草图检索性能方面的有效性.

需要指出的是,文中所提出的算法独立于检索内容的特征表示,只要检索对象的内容可以表征成特征向量的形式,即可使用小样本增量有偏学习算法对其检索结果进行优化.因此,本文的方法同样适用于图像检索和三维模型检索等其他应用领域.然而,用户和计算机之间的映射是一个十分复杂的问题,如何利用相关反馈来进一步缩小用户与计算机之间的语义鸿沟,如利用用户建模技术建立长期反馈机制,仍是有待于深入研究的课题.

References:

[1] Sun ZX, Feng GH, Zhou RH. Techniques for sketch-based user interface: Review and research. Journal of Computer-Aided Design & Computer Graphics, 2005,17(9):1891−1899 (in Chinese with English abstract).

[2] Kamel I, Barbará D. Retrieving electronic ink by content. In: Nwosu K, Bobbie P, Orji C, eds. Proc. of the IW-MMDBMS’96. Washington: IEEE Press, 1996. 54−61.

[3] Leung H. Representations, feature extraction, matching and relevance feedback for sketch retrieval [Ph.D. Thesis]. Pittsburgh: Carnegie Mellon University, 2003.

[4] Fonseca MJ. Sketch-Based retrieval in large sets of drawings [Ph.D. Thesis]. Lisbon: University Technique Lisbon, 2004.

[5] Liang S, Sun ZX, Li B. Sketch retrieval based on spatial relations. In: Sarfraz M, Wang YS, Banissi E, eds. IEEE Proc. of the Int'l Conf. on Computer Graphics, Imaging and Visualization: New Trends. Washington: IEEE Press, 2005. 24−29.

[6] Liang S, Sun ZX. BSVM-Based relevance feedback for sketch retrieval. Journal of Computer Aided Design & Computer Graphics, 2006,18(11):1753−1757 (in Chinese with English abstract).

[7] Huang TS, Zhou XS. Image retrieval with relevance feedback: From heuristic weight adjustment to optimal learning methods. In: Mercer B, eds. IEEE Proc. of the Int'l Conf. on Image Processing. Piscataway: IEEE Press, 2001. 2−5.

[8] Rui Y, Huang TS, Mehrotra S. Content-Based image retrieval with relevance feedback in MARS. In: Torwick I, ed. Proc. of the IEEE Int'l Conf. on Image Processing. Washington: IEEE Press, 1997. 815−818.

[9] Laaksonen J, Koskela M, Laakso S, Oja E. PicSOM: Self-Organizing maps for content-based image retrieval. Pattern Recognition Letters, 2000,21(13-14):1199−1207.

[10] Tieu K, Viola P. Boosting image retrieval. Int'l Journal of Computer Vision, 2004,56(1-2):17−36.

[11] MacArthus SD, Brodley CE, Shyu C. Relevance feedback decision trees in content-based image retrieval. In: Werner B, ed. Proc. of the IEEE Workshop on Content-Based Access to Image and Video Libraries. Washington: IEEE Press, 2000. 68−72.

[12] Wu P, Manjunath BS. Adaptive nearest neighbor search for relevance feedback in large image databases. In: Georganas N, Popescu-Zeletin R, eds. Proc. of the 9th ACM Int'l Conf. on Multimedia. New York: ACM Press, 2001. 89−97.

[13] Cox IJ, Miller ML, Minka TP, Papathomas TV, Yianilos PN. The Bayesian image retrieval system, PicHunter: Theory, implementation and psychophysical experiments. IEEE Trans. on Image Processing, 2000,9(1):20−37.

[14] Li B, Sun ZX, Liang S, Zhang YY, Bo Y. Relevance feedback for sketch retrieval based on linear programming classification. In: Zhuang YT, Yang SQ, Rui Y, He QM, eds. Proc. of the Pacific-Rim Conf. on Multimedia 2006. LNCS 4261, Berlin, Heidelberg: Springer-Verlag, 2006. 201−210.

[15] Hong P, Tian Q, Huang TS. Incorporate support vector machines to content-based image retrieval with relevant feedback. In: Ward RK, ed. Proc. of the IEEE Int'l Conf. on Image Processing. Piscataway: IEEE Press, 2000. 750−753.

[16] Zhou XS, Huang TS. Exploring the nature and variants of relevance feedback. In: Kak A, Smith J, eds. Proc. of the IEEE Workshop on Content-based Access of Image and Video Libraries. Washington: IEEE Press, 2001. 94−101.

[17] Huang TS, Zhou XS, Nakazato M, Wu Y, Cohen I. Leaning in content-based image retrieval. In: Elman J, Sur M, Weng J, eds. Proc. of the 2nd Int'l Conf. on Development and Learning. Washington: IEEE Press, 2002. 155−164.

[18] Tong S, Chang E. Support vector machine active learning for image retrieval. In: Georganas N, Popescu-Zeletin R, eds. Proc. of the 9th ACM Int'l Conf. on Multimedia. New York: ACM Press, 2001. 107−118.

[19] Seung HS, Opper M, Sompolinsky H. Query by committee. In: Haussler D, ed. Proc. of the 5th Annual ACM Workshop Computational Learning Theory. New York: ACM Press, 1992. 287−294.

[20] Freud Y, Seung HS, Shamir E, Tishby N. Information, prediction, and query by committee. In: Hanson S, Cowan J, Lee Giles C, eds. Advances in Neural Information Processing Systems 5. San Francisco: Morgan Kaufmann Publishers, 1992. 483−490.

[21] Lewis DD, Gale WA. A sequential algorithm for training text classifiers. In: Croft W B, van Rijsbergen CJ, eds. Proc. of the 17th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: Springer-Verlag, 1994. 3−12.

[22] Chen Y, Zhou XS, Huang TS. One-Class SVM for learning in image retrieval. In: Mercer B, eds. Proc. of the IEEE Int'l Conf. on Image Processing. Piscataway: IEEE Press, 2001. 34−37.

[23] Chan CH. Using biased support vector machine in image retrieval with self-organizing map [MS. Thesis]. Hong Kong: The Chinese University of Hong Kong, 2004.

[24] Chang CC, Lin CJ. LIBSVM: A library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/ libsvm

附中文参考文献:

[1] 孙正兴,冯桂焕,周若鸿.基于手绘草图的人机交互技术研究进展.计算机辅助设计与图形学学报,2005,17(9):1891−1899.

[6] 梁爽,孙正兴.面向草图检索的相关反馈方法.计算机辅助设计与图形学学报,2006,18(11):1753−1757.

梁爽(1983－),女,辽宁沈阳人,博士,主要研究领域为多媒体信息检索,智能人机交互.

孙正兴(1964－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为多媒体计算,计算机视觉,智能人机交互.