第16卷 第2期
2006年2月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 2
Feb. 2006

图像检索中的主动学习及其可测量性

凌俊斌,庄卫华,刘鲁西
(河海大学 计算机及信息工程学院,江苏 南京 210098)

摘 要:主动学习对于复杂、主观、使用少量训练实例的图像检索查询具有非常有效的作用。在图像检索中应用主动学习与支持向量机相结合的方法进行相关反馈,通过两者的互补来有效地提高图像检索的精度。对比了推理算法、简单主动算法以及角度多样性算法3种主动学习算法,并研究了最好的样本选择策略。还讨论了主动学习中概念复杂度的可测量性,并对未来的研究方向提出了建议。相信随着这些可测量性问题被重点提出,主动学习的成果可以被广泛应用。

关键词:图像检索; 支持向量机; 主动学习
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1005-3751(2006)02-0132-03

Active Learning and Its Scalability for Image Retrieval

LING Jun-bin, ZHUANG Wei-hua, LIU Lu-xi
(Computer and Information Engineering College of Hohai University, Nanjing 210098, China)

Abstract: Active learning plays a vital role in image retrieval with a little amount of training instances. That's because we can combine active learning with support vector machines to improve the accuracy of image retrieval. This paper compares three kinds of active learning algorithms: speculative algorithm, simple active algorithm and angle-diversity algorithm, and develops the best strategy of sample selection. Besides, this paper discusses the two scalability issues of active learning: scalability in dataset size, and scalability in concept complexity, and provides many suggestions of the directions of the study. With the development of the issues of scalability problem, the production of active learning can benefit those areas definitely.

Key words: image retrieval; SVMs; active learning

0 引 言

在图像检索中应用主动学习与支持向量机(Support Vector Machines)相结合的方法来进行相关反馈,通过两者的互补能有效地提高图像检索的精度。主动学习可以找到最具有信息的样本来引发用户反馈以达到目标查询概念被快速学习的目的。该方法将数据映射到高维空间来增加线性学习器的计算能力,为支持向量机提供了一个重要的构成模块,SVMs的一个重要特征就是在一定的程度上使逼近理论的问题与学习理论的问题相互独立。

文中给出了3种主动学习策略并比较它们的性能,讨论了主动学习在概念复杂度中的可量测性。在此,考查了概念的缺乏性、多样性和孤立性对学习性能的影响。最后提出了主动学习研究未来的可能发展方向。

1 主动学习策略

主动学习在学习过程中可以根据学习进程,选择最有利于分类器性能的样本来进一步训练分类器,它能有效地减少评价样本的数量。

主动学习从形式上是一个循环反复的过程。首先,候选样本集中所有的样本都未带类别标注,根据先验知识或者随机地从候选样本集中选择少量样本并标注它们的类别,构造初始训练样本集,确保初始训练样本集中至少包含有一个正例样本和一个负例样本。利用初始训练样本集中这些带类别标注的样本训练一个分类器,在该分类器下,采用某种采样算法,从候选样本集中选择最有利分类器性能的样本,标注类别并加入到训练样本集中,重新训练分类器,再次选择最有利分类器性能的样本。重复以上过程,直到候选样本集为空或达到某种指标[1]。

支持向量机是在高维特征空间使用线性函数假设空间的学习系统,它由一个来自最优化理论的学习算法训练,具有使用核、没有局部最小,以及通过间隔或者是维数无关的量(比如支持向量个数)来控制容量的特点[2]。支持向量机在使训练样本分类误差最小化的前提下,尽量提高分类器的泛化推广能力,从实施的角度看,训练支持向量机等价于解一个线性约束的二次规划问题,使得分隔特征空间中两类模式点的两个超平面之间的距离最大,而且它能保证得到的解为全局最优点,具有较好的泛化和推广

收稿日期:2005-05-24
作者简介:凌俊斌(1981—),男,江苏南通人,硕士研究生,研究方向为多媒体技术;庄卫华,副教授,硕士生导师,主要研究方向为计算机应用技术。

能力[3]。在范围广大的应用中,SVMs 的性能胜过其他大多数的学习系统。

主动学习与 SVMs 一起工作的基本思想是要在支持向量间找出最不确定的未标注的实例向用户进行询问来对超平面进行改进[4]。用 SVMs 实现主动学习,采取何种采样算法是关键,如何选择新的样本进行评价直接关系到整个算法的性能。对于线性可分问题,从接近分类超平面的未标记样本中采样后所得分类器,其分类超平面位置最有可能被改变,而远离分类超平面的样本被采样后对分类器影响不大。因此选择离分类超平面最近的 n 个样本作为新的样本进行评价[4]。将主动学习与 SVMs 结合(记做 SVM_{active})进行图像检索,用三步返回与内容最相关的前 k 张图像,其过程为:

(1)SVM_{active} 将学习目标看作二元输出问题,指出超平面的一边为与查询内容相关,而另一边为无关。

(2)SVM_{active} 通过主动学习快速地获得分类器。主动学习选择最具有信息量的实例来训练分类器,这确保了查询结果以最少的反馈循环而快速收敛。

(3)分类器被最终确定,SVM_{active} 返回与查询结果最相关的前 k 张图像。

下面,给出主动学习的 3 种主要样本选择策略:推理算法、简单主动算法以及角度多样性算法,并对它们的性能进行分析比较。

1.1　推理算法

推理算法的过程是通过递归对用户的反馈进行推理处理来产生样本。虽然推理过程几乎是最佳样品策略,但是需要密集的计算,用它作为其它主动学习策略性能的衡量尺度。

推理算法伪代码如下:

Input: L, U, h; //L 为标注集,U 为未标注集,h 为样本个数

Output: S; //S 为样本

Procedures:

SVM_{Train}(); //SVM 训练算法

f(); //在标注集上通过支持向量机训练得到的分类器

Initialization:

S←0; //将 S 初始化为 0

BEGIN

1. if(h = 0) then return 0;

2. f ← SVM_{Train}(L);

3. for each $X_i \in U$

X_i.distance←|f(X_i)|;

4. X_s ← $\text{argmin}_{x_i \in U}(X_i.\text{distance})$;

5. U ← U − {X_s};

L ← L ∪ {X_s};

6. X_s.label ← +;

7. S ← S ∪ Speculative(L, U, (h − 1) / 2);

8. Xs.label ← −;

9. S ← S ∪ Speculative(L, U, h − (h − 1) / 2);

10. S ← S ∪ {X_s};

11. return S;

END

推理算法从找出最具信息的样本开始(离超平面最近的未标注的实例),然后根据样本的两个可能的标注推理,产生两个样本,一个为正,一个为负。这个算法进行递归的推理来产生输出结果。算法的第 6 和第 8 步分别把样本推理成正和负,再各自递归地调用推理过程选择下一个样本。在 h 个样本产生后,推理算法停止。

1.2　简单主动算法

简单主动算法选择 h 个距离分离超平面最近的未标注的实例来请求用户反馈[5]。在标注过的初始训练样本集 L 的基础上,算法先训练分类器 f,再将分类器 f 应用于未标注的候选样本集 U 来计算每个未标注实例到分离超平面的距离。然后将 h 个未标注实例选为下一组样本来用作训练。简单算法的基本思想是距离超平面最近的这 h 个实例,是在 L 上训练的 f 的最不确定的实例。简单算法是要通过加入最具有信息量的样本来修改 f,进而最大限度地改善不确定性。

1.3　角度多样性算法

被选中的样本需要具有多样性,可以在样本选择中加入多样性度量[5]。角度多样性算法的主要思想是选择距离分类超平面最近的样本集,同时保持多样性,而样本的多样性通过它们之间的角度来进行测量。

角度多样性算法伪代码如下:

Input: L, U, h, λ; //L 为标注集,U 为未标注集,h 为样本个数,λ 为权重参数

Output: S; //S 为样本集

Procedures:

SVM_{Train}(); //SVM 训练算法

f(); //在标注集上通过支持向量机训练得到的分类器

K(); //K 为 SVM 核函数

Initialization:

S←0; //将 S 初始化为 0

BEGIN

1. f ← SVM_{Train}(L);

2. While(|s| < h)

$$X_s \leftarrow \text{argmin}_{x_i \in U}\left(\lambda * |f(X_i)| + (1-\lambda) * \left(\max_{x_j \in S} \frac{K(X_i, X_j)}{\sqrt{K(X_i, X_i)K(X_j, X_j)}}\right)\right)$$

S ← S ∪ {X_s};

3. Return S;

END

假设样本 X_i,其标准向量为 $\Phi(X_i)$,与两个超平面 h_i 和 h_j 之间的对应角度为 x_i 和 x_j,那么利用核 K 表示为:

$$|\cos(\angle(h_i,h_j))| = \frac{|\Phi(x_i)\cdot\Phi(x_j)|}{||\Phi(x_i)||||\Phi(x_j)||} = \frac{K(X_i,X_j)}{\sqrt{K(X_i,X_i)K(X_j,X_j)}}$$

角度多样性算法从最初的超平面开始,用指定的标注集 L 进行训练,然后再对每个没有标注的实例 x_j,计算到分类超平面 h_i 的距离,再将未标注的实例 x_j 和当前样本集 S 之间的角度定义为样本 x_j 到集合 S 中任一实例 x_s 之间的最大的角度,这个角度衡量了实例 x_j 被选为样本后样本集 S 的多样性。

角度多样性算法引入了参数 λ 来平衡两部分:到分类超平面的距离和样本中角度的多样性。加入平衡因子,最终的未标注的实例 x_j 可以写成:

$$\lambda * |f(X_i)| + (1-\lambda) * (\max_{x_j\in s} \frac{K(X_i,X_j)}{\sqrt{K(X_i,X_i)K(X_j,X_j)}})$$

式中,函数 f 计算到超平面的距离,K 是核函数,S 是训练集。之后,算法将选择在 U 中具有最小值的未标注的实例作为样本。算法重复以上的步骤 h 次来获得 h 个样本。根据试验,使用平衡参数 $\lambda = 0.5$ 时,算法可以获得很好的性能 。

2 测量性问题

当处理由大量潜在的查询概念组成的大型数据集合时,主动学习面临两个测量性问题:数据集大小的测量和概念复杂度的测量[6]。数据集大小的测量问题是很明显的,当未标注集 U 很大时,是禁止通过计算来扫描整个未标注集来选择样本的,需要有一个有效的检索方法来选择样本而不涉及整个 U 。在此不作详细叙述。

对概念复杂性问题,文中使用多样性、缺乏性、孤立性 3 种方法来测量。

利用实验来比较 3 种主动学习策略的性能,使用具有 107 个分类、50000 个图像的数据集。

2.1 概念多样性

概念的多样性通过相关图像在输入空间内的分散程度来衡量,多样性也是概念可学习性的一个重要指标。多样性高的概念在整个输入空间都分布着相关图像。例如:概念"花",包含了不同颜色和类型的花,其多样性远胜于"红玫瑰"的概念。

文中的图像数据集中,概念多样性和概念可学习性并没有表现出很强的关系。如图 1 所示,得出的 20 个输出结果与概念多样性的比率关系是很弱的 0.3701(0 表示没有关系,1 表示全关联)。通过仔细检查多样性因素,概念的多样性值被封装在 1.7 到 3.4 这个小的范围内。也就是说,不同概念的多样性差异是很小的。这是因为,当数据维数很高的时候,空间中点的距离趋于接近,其结果就是任意两个图像之间距离的差异是很小的,由此概念之间的多样性差异就很小。

2.2 概念缺乏性

概念的缺乏性用命中率(数据和概念匹配的百分率)来表示。命中率衡量了在一个检索系统中一个概念是怎样被完好描述的。当假定每个关键字等同于一个概念,关键字的命中率即为用该关键字来做标注的图像数目。命中率的高低与数据集的不同有关。一个很具体的概念可能会是很缺乏的(低命中率),因为数据中匹配的实例很少。反之,一个很抽象的概念可能会具有高命中率。

2.3 概念孤立性

孤立性描述了一个概念和其他概念间的分离程度。如果一个概念和其他概念在数据空间中混合,其具有低孤立性。比如关键字"猫科"经常被用于注释虎、狮子和家猫的图片。如果用户以"猫科"查询,将很难区别用户脑中的这个词的意思。而一个较好独立性的搜索字"天安门广场"几乎不会有模糊性,可以确信用户想的是著名的中国北京天安门广场。概念的孤立性度量在 0 和 1 之间,1 表示概念很好地被隔离,0 表示和其他概念重叠。如图 2 所示,在文中试验的图像数据集中,查询性能和孤立性之间的关系是 0.6514。孤立性是一个很好的分类概念复杂性的方法。

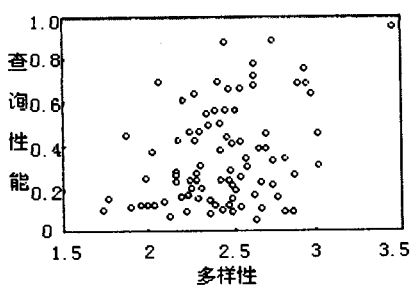

图 1 概念多样性

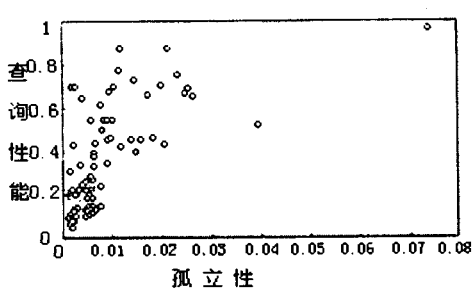

图 2 概念孤立性

3 实验和讨论

在下面的实验中,用检索出与查询结果最相关的前 20 张图像的精度来衡量检索性能,对每个实验进行 20 次,给出其平均精度。对于 50k 的数据集,扫描整个数据集来进行样本选择和检索显得太繁琐。因此,通过索引结构来进行。试验时根据准确率和执行时间作为比较 3 个取样策略的标准,其结果如图 3 所示。

图中表明角度多样性算法在所有的主动学习算法中

(下转第 138 页)

Grid－Peer 之间的资源共享，这类似于银行异地存取。图 3 是一种任务调度模型。在 Grid－Peer 系统中，网格用户可以向任何一个的 Broker 提出资源请求，Broker 接受网格用户请求信息后，查看其所需的资源的信息，把它交给相应的资源管理器处理，资源管理器查询相应的资源组队列，并选择相应的资源分配给任务。如果查不到最合适的资源，可以查找其它的资源组，或向 Broker 返回无合适的资源，Broker 则向其它的 Broker 传送该任务的需求信息，直到找到相应的资源，或者经过一段时间后，以失败结束这次资源查找过程。

4 结 论

文中分析了网格和 P2P 网络的特征，以及二者之间的相似性，对比了其各自的优缺点，分析了网格和 P2P 网络的结合的可能性，再加上 OGSA 的技术特征，提出了具有网格和 P2P 网络技术特征的 P2P－Grid 系统。比较了它与银行运营模式，从而构建出它的物理模型、功能模型以及一种任务调度模型。该系统从整体上说是一个 P2P 网络，它的每个节点是一个局部网格系统，称之为 Super－Peer 或 Grid－Peer。各 Super－Peer 的管理模式和构建规模等都不一定相同。Super－Peer 之间依赖于 P2P 网络技术连接。Super－Peer 的管理工作是由一个或几个服务器和若干 Broker 完成，Super－Peer 之间也通过 Broker 进行交互。虽然文中从 P2P－Grid 系统的物理构建模型、功能模型和任务调度模型的工作原理上做了说明，但是 P2P－Grid 系统的具体实现过程还有很多工作要做。例如：P2P－Grid 系统的资源管理，包括资源的描述、资源的组织与注册、资源的发现与定位、资源调度与任务处理，这些主要依赖设计出有效的 Broker，当然还有安全管理等问题，都有待继续研究。

参考文献：

[1] Foster I，Kesselman C. The Grid：Blueprint for a New Computing Infrastructure[M]. San Fransisco，CA：Morgan Kaufmann，1999.

[2] Foster I，Ianmitchi A. A Peer－to－Peer Approach to Resource Location in Grid Environments[EB/OL]. http://people.cs.uchicago.edu/~anda/papers/iamnitchi－bookch.pdf. 2002.

[3] Lawniczek B，Majka G，Slowikowski P，et al. Grid Infrastructure Monitoring Service Framework Jiro/JMX Based Implementation[J/OL]. Electr. Notes Theor. Comput. Sci.，http://www.elsevier.nl/locate/entcs/volume82.html 12 pages. 2003.

[4] Ranganathan K，Iamnitchi A，Foster I. Improving Data Availability through Dynamic Model－Driven Replication in Large Peer－to－Peer Communities[A]. CCGrid2002[C]. [s.l.]：IEEE Computer Society，2002.376－381.

[5] Talia D，Trunfio P. Toward a Synergy Between P2P and Grids[J]. IEEE Internet Computing，2003，7(4)：94－96.

（上接第 134 页）

性能最好。角度多样性算法在一些交互中的性能和推理算法的一样，甚至还要好一些，而推理算法被认为是可以达到几乎最佳效果的算法，但是推理算法需要密集的计算，所以就效力和效率而言，角度多样性算法是最理想的选择。

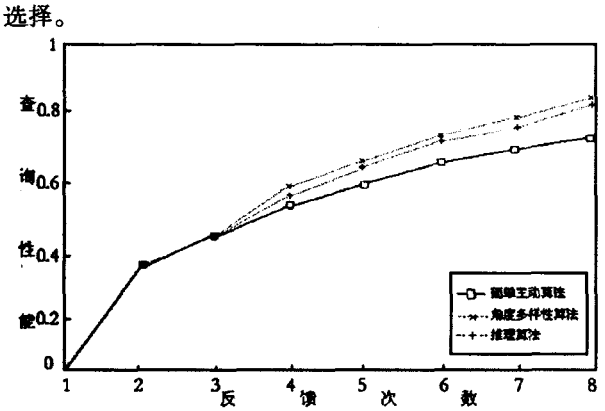

图 3 试验结果

4 结 论

指出了主动学习在学习具有最少信息量的复杂、主观查询概念中的效力，并对比 3 种主动学习算法，研究了最好的样本选择策略。还讨论了主动学习的两个最重要的可测量性问题：概念复杂度和数据集的可测量性问题。当概念在数据集中几乎没有可匹配的实例，并没有很好地和其他概念隔离时，那么概念的可学习性就会受到影响。相信这些可测量性问题应该被重点提出，这样主动学习的成果就可以被广泛应用。

参考文献：

[1] 张健沛，徐 华. 支持向量机(SVM)主动学习方法研究与应用[J]. 计算机应用，2004(1)：1－3.

[2] Cristianini N，Shawe－Taylor J. 支持向量机导论[M]. 李国正，王 猛，曾华军译. 北京：电子工业出版社，2004.

[3] 徐彤阳，姚跃华，朱志勇. 一种基于支持向量机的图像边缘检测方法[J]. 微机发展，2005，15(1)：87－90.

[4] Schohn G，Cohn D. Active learning with support vector machines[A]. In：Proceedings of the Seventeenth International Conference on Machine Learning(ICML－2000)[C]. California：[s.n.]，2000.839 － 846.

[5] Tong S，Chang E. Support vector machine active learning for image retrieval[A]. In：Proceedings of ACM International Conference on Multimedia[C]. Ottawa：[s.n.]，2001.107－118.

[6] Goh K，Li B，Chang E Y. DynDex：A dynamic and non－metric space indexer[A]. In：Proceedings of ACM International Conference on Multimedia[C]. Juan Les Pin，France：[s.n.]，2002.466－475.