基于 LBSVM 机器学习的相关反馈图像检索

欧阳军林,刘建勋,曹步清
OUYANG Jun-lin,LIU Jian-xun,CAO Bu-qing

湖南科技大学 计算机科学与工程学院,湖南 湘潭 411201
Department of Compute Engineer,Hunan University of Science and Technology,Xiangtan,Hunan 411201,China
E-mail:yangjunlin0732@163.com

OUYANG Jun-lin,LIU Jian-xun,CAO Bu-qing.Relevance feedback image retrieval based on LBSVM.Computer Engineering and Applications,2009,45(2):112-115.

Abstract: Relevance feedback technology based on machine learning becomes focus in image retrieval.In relevance feedback based on SVM,sample is lack and unbalance,precise feedback is low.For these problem,in this paper,a new relevance feedback method based on machine learning is presented,which combines on Boosting and SVM.It improves image retrieval speed and accuracy.A new feedback method based on long machine learning is presented.Experiments show that the proposed system is not only efficient but also effective.

Key words: relevance feedback;machine learning;Boosting;image retrieval

摘 要 基于机器学习的相关反馈技术是基于内容的图像检索研究的热点。由于基于 SVM 的相关反馈技术存在样本数量少,样本正负比例不平衡,反馈准确率低等问题,文中先对 Boosting 方法进行改进,提出了用先验知识的 Boosting 方法与 SVM 结合的短期机器学习相关反馈方法(BSVM),在此基础上为进一步提高系统反馈速度与准确率,通过保存训练好的分类器和它对应的样本,提出了基于长期机器学习的相关反馈方法(LBSVM)。文中提出的两种方法与其它方法进行了比较实验,结果表明,该方法优于其它方法。

关键词 相关反馈;机器学习;Boosting 方法;图像检索
DOI:10.3778/j.issn.1002-8331.2009.02.032 **文章编号**:1002-8331(2009)02-0112-04 **文献标识码**:A **中图分类号**:TP391

1 引言

当前相关反馈方法已成为图像检索技术中研究的热点,提高图像检索性能的强大工具。相关反馈方法的基本思路是:在查询的过程中允许用户对检索结果进行评价和标记,指出结果中哪些"正确"的检索结果(与查询图像相关),哪些是"错误"的检索(与查询图像不相关),然后将用户标记的相关信息作为训练样本反馈给系统进行学习,指导下一轮的检索,使得检索结果更符合用户的需要。相关反馈技术发展至今大约 10 余年,许多学者提出了各种各样的相关反馈方法。从相关反馈原理大致可以分为以下 4 类:查询向量转移的相关反馈[1],权值调整的相关反馈[2],基于统计学习理论的相关反馈[3],基于机器学习的相关反馈[4-7]。其中基于机器学习的相关反馈方法是一种比较新颖的反馈方法,许多学者将相关反馈看作模式识别中的有监督学习或分类问题,主要采用的机器学习理论有:支持向量机[4,7](SVM)、决策树[5]、神经网络[6]、基于内核的学习等。文献[4,7]采用 SVM 作为图像检索的相关反馈方法,用户对第一次得到的检索结果进行标记,把标记的相关和不相关图像作为 SVM 的训练样本,由训练样本得到分类器对整个图像库进行分类,根据 SVM 分类函数值的大小把相关图像顺序反馈给用户,实验证明该方法具有较好反馈检索效果。

然而这种方法对某一种类型的图像库具有较好性能,对大型图像库或者复杂图像库就存在一些问题。首先,用户标记的样本数量有限,其次,用户标记的相关和不相关样本可能不平衡,最后,单一的 SVM 分类并不是非常准确。而 Boosting 方法能把弱学习算法提升为强学习算法,因此在文献[7]SVM 相关反馈的基础上,用改进的结合先验知识的 Boosting 方法增强 SVM 的分类性能,提出了两种新颖的相关反馈方法:Boosting 与 SVM 结合的短期学习相关反馈方法(BSVM),Boosting 与 SVM 结合的长期学习相关反馈方法(LBSVM)。前一种短期学习相关反馈方法是在 SVM 相关反馈方法上的改进,主要用于解决 SVM 相关反馈方法中样本数量少,样本不平衡的问题,同时提高 SVM 分类准确性;而长期学习相关反馈方法是短期学习相关反馈的进一步改进,主要作用有两个:提高反馈的速度,其次提高反馈的准确率。图 1 是整个相关反馈检索系统结构框图。

基金项目:国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2DD3CB317007);湖南省自然科学基金(the Natural Science Foundation of Hunan Province of China under Grant No.03jjy6024)。
作者简介:欧阳军林(1977-),男,讲师,研究方向:图像处理、图像检索、模式识别;刘建勋(1970-),男,博士,教授,研究方向为图像处理、网格计算;曹步清(1979-),男,讲师,研究方向:图像处理、图像检索、模式识别。

收稿日期:2007-12-28 修回日期:2008-03-21

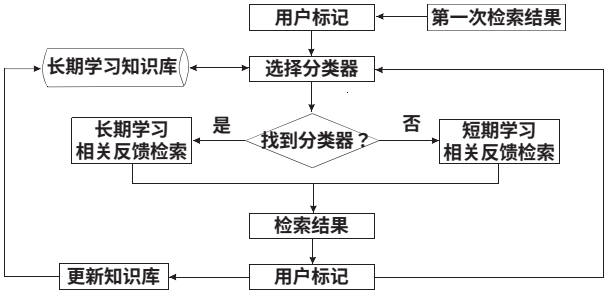

图 1 相关反馈检索系统结构框图

2 短期的基于机器学习的相关反馈方法(BSVM)

Boosting 方法主要的作用是能把弱学习算法提升为强学习方法。在 SVM 相关反馈检索的基础上,为了提高检索准确率,克服 SVM 训练中样本少和样本不平衡问题,采用改进的 Boosting 方法提高了 SVM 反馈的准确率。本章首先简单地介绍 Boosting 方法,接着介绍提出的结合先验知识的 Boosting 方法,最后给出了相应的基于 Boosting 的相关反馈图像检索算法。

2.1 Boosting 方法

Boosting[8]方法的基本思想是:给定一弱学习算法和训练集 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,这里 x_i 为第 i 个训练样本的输入,y_i 为分类问题的类别标志。算法开始给每一个训练样本赋予相等的权值 $1/n$,然后用该学习算法对训练集样本训练 T 轮,每次训练后,对训练失败的训练样本赋予较大的权值,也就是让学习算法在后续的学习中集中对比较难的训练样本进行学习,从而得到一个预测函数序列 $h_1,h_2,\cdots,h_T$,其中每个 h_j 也对应一定的权值,预测效果好的预测函数的权值较大,反之较小。最终的预测函数 H 采用加权投票方式对新样本进行判别。Boosting 算法具体描述如下:

(1) 输入:$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$;$x_i\in X,y_i\in Y=\{+1,-1\}$;初始化:$D_1(i)=1/n$。

(2) For $t=1,\cdots,T$

①在 D_t 下训练,得到预测函数 h_t;

②计算该预测函数的错误率

$$E_t=\sum_{h_t(x_i)\neq y_i}D_t(i) \tag{1}$$

选择 $\alpha_t=\frac{1}{2}\ln\left(\frac{1-E_t}{E_t}\right)$ (2)

③根据错误率更新样本的权值:

当 $h_t(x_i)\neq y_i$ 时:

$$D_{t+1}(i)=D_t(i)*\exp(\alpha_t) \tag{3}$$

当 $h_t(X_i)=Y_i$ 时:

$$D_{t+1}(i)=D_t(i)*\exp(-\alpha_t) \tag{4}$$

(3) 对于未知样本,输出:

$$H(x)=\mathrm{sign}(\sum\alpha_t h_t(x)) \tag{5}$$

2.2 结合先验知识的改进 Boosting 方法

Boosting 方法完全是在没有人参与的情况下进行学习分类,然而,许多实验表明,在有人类先验知识的情况下进行学习,分类的准确性更高,因此采用结合先验知识的 Boosting 方法。为了结合一些先验知识在 Boosting 方法中,必须创建一个规则 π 映射到每个示例 x 属于标记值 $\{+1,-1\}$ 的条件概率 $\pi(y|x)$。对于相关反馈方法中,样本的概率也就是先验知识,可以通过图像检索模块中计算得到的相似度来表示样本先验知识的条件概率,因为用户反馈给系统的样本相似值越大,那么该样本属于这类的概率就越大,因此,本算法中先验知识的条件概率 $\pi(y|x)$ 就是图像检索中的相似度值,它的值在用户第一次检索中就已经算出。在给定样本先验概率的情况下,对上面的 Boosting 方法进行改进,主要从两个方面进行改进:

(1) 为了结合先验知识,添加了两个加权的带有先验知识的新样本。即在训练阶段,对每个训练样本 (x_i,y_i),增加两个新的样本 $(x_i,+1)$,$(x_i,-1)$,两个新样本的初始权重分别为 $\eta\pi_+$ 和 $\eta(1-\pi_+(x_i))$。

(2) Boosting 方法输出函数公式 $H(x)=\mathrm{sign}(\sum\alpha_t h_t(x))$ 的基础上增加一个 0 阶的决策函数 $h_0(x)$,该决策函数的值由样本的先验概率根据如下公式得到。

$$h_0(x)=\sigma^{-1}(\pi_+(x))=\ln\left(\frac{\pi_+(x)}{1-\pi_+(x)}\right) \tag{6}$$

因此 Boosting 方法最终的输出函数修改为:

$$H(x)=\mathrm{sign}\left(\sum_{t=0}^{T}\alpha_t h_t(x)\right) \tag{7}$$

其中 $\alpha_0=\frac{1}{T}\sum_{t=1}^{T}\alpha_t$。

因此通过上面的改进,在相关反馈检索中增加了样本的数量,较好的使正负样本的比例趋于平衡,同时通过 Boosting 方法使 SVM 分类准确率进一步提高,下面给出相应的算法。

2.3 改进的 Boosting 与 SVM 结合相关反馈图像检索算法(BSVM)

基于 BSVM 相关反馈算法的主要思想是:根据用户标记的相关样本和不相关样本,利用 SVM 先对这些小样本进行训练,得到一个预测分类函数,然后利用 Boosting 方法,并调整小样本特征的权值系数,作为新的样本,重复上面的步骤 T 次。最后对新的样本通过上面训练得到的 T 个预测函数进行投票确定最终所属类别,按照所属类别的函数值的大小按降序反馈给用户。BSVM 相关反馈框图如图 2 所示。

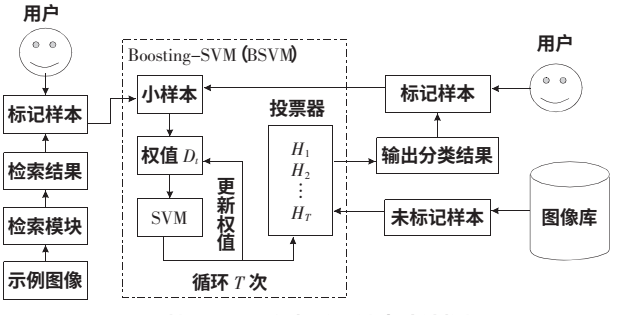

图 2 基于 BSVM 的相关反馈方法结构框图

基于 BSVM 的相关反馈短期学习算法如下:

(1) 用户通过检索模块或者 SVM 反馈模块得到检索结果,并进行标记正例 P^+ 或负例 P^-,把 P^+ 或负例 P^- 送入 BSVM 中。

(2) 根据训练样本对 BSVM 进行训练 T 次。

For $i=1$ to T

①利用支持向量机的分类函数 $f(x)$ 公式 (6) 即得到预测函数 $H_t(i)$。

$$f(x)=\frac{2}{b}\sum_{i=1}\alpha_i y_i k(x_i,x)-k(x,x) \tag{8}$$

②利用前面介绍的 Boosting 方法中式(3),式(4)得到更新的权值 $D_t(i)$。

当 $h_t(x_i)\neq y_i$ 时:

$$D_{t+1}(i)=D_t(i)*\exp(\alpha_t) \tag{9}$$

当 $h_t(X_i)=Y_i$ 时:

$$D_{t+1}(i)=D_t(i)*\exp(-\alpha_t) \tag{10}$$

③根据公式(2)计算参数 α_t

$$\alpha_t=\frac{1}{2}\ln\left(\frac{1-E_t}{E_t}\right) \tag{11}$$

(3)训练完以后,对图像库中的未标记样本根据训练后的投票器进行分类,确定未标记样本的类别是相关还是不相关,并根据公式(7) $H(x)=\sum\alpha_t h_t(x)$,得到相关图像的值按照降序反馈给用户。

(4)如果用户觉得满意则停止检索,如果不满意则跳到步骤(1)继续执行。

3 基于长期学习的相关反馈方法(LBSVM)

3.1 长期学习知识库的建立

从上面的相关反馈方法中可以看到,在 SVM 相关反馈的基础上,通过引入 Boosting 方法,提出了 BSVM-RF 方法,该方法克服了样本数量少和样本不平衡问题,同时通过多次训练最后经过投票确定类别,提高了分类准确率,也即增强了 SVM 的分类性能。然而,这种方法还存在的一个问题就是:当用户检索完成以后并没有保留检索中的有用信息,如分类器可以保存留作下次检索用,用户标记属于该分类器的样本等,如果用户没有保存这些有用信息,那么当用户进行下次检索时,必须重新训练,然后进行分类,因此为了提高检索的效率,就可以利用以前训练好的分类器直接进行分类。基于这种思想,提出了基于 BSVM 相关反馈的长期学习算法(LBSVM-RF),该方法把每次 BSVM 训练完成后的 t 个预测函数 h_t 和 t 个权值系数 α_t 保存下来,并把训练中与该分类器相关的图像与不相关的图像在图库中的序号 ID 一起保存,留作下次进行相同类型图像检索时使用。保存 h_t,和 α_t 也就保存以前训练好的分类器,保存训练中相关和不相关图像就是为下次反馈检索中采用该分类器时增加正负样本,提高检索准确率。因此建立如图 3 所示的 BSVM 知识库。

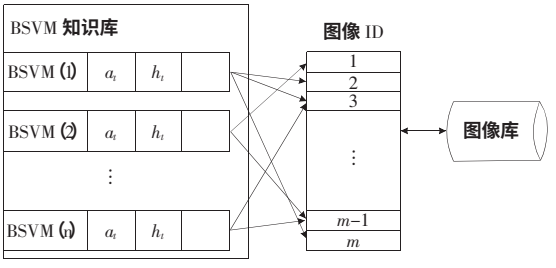

图 3 BSVM 知识库

3.2 基于长期学习的相关反馈方法(LBSVM-RF)

提出的 LBSVM-RF 方法与传统的 SVM-RF 算法相比有如下优点:首先传统 SVM-RF 算法根据用户标记的检索结果,先进行训练,后分类,而本文的方法是利用以前的分类器先分类,后训练,这样的好处是能够利用以前的训练器先进行分类,加快反馈的速度,省去了训练的时间;其次,与传统 SVM-RF 算法相比,提出的方法中长期知识库中积累以前用户反馈的信息,也就是说保存了该分类器的正例样本和负例样本,也即增加了样本数目,因而提高了分类准确度。

上面说明了基于长期学习相关反馈方法的好处,那么存在的问题是,在什么情况下利用长期学习,什么情况利用短期学习?以及在长期学习情况下选用哪个分类器?下面给出了选择分类器进行长期学习必须满足的两个条件。

(1) ①for 对知识库中的每个分类器 i 设置初始变量 BSVM(i)=0

for 对在某次检索中用户标记的每个正例 I_p^1 和负例 I_n^1

if (I_p^1 or I_n^1) 的 ID 在知识库的 BSVM(i) 中,BSVM(i) ++

②在变量 BSVM(i) 中找包含此次训练样本最多的分类器,即 argmax(BSVM(i)) 做为训练分类器。

(2) 假设在某次检索中用户标记的正例样本和负例样本的总数为 m,SVM 知识库中某个分类器保存的图像总数为 n,当满足下面条件时,则选择该分类器进行分类。

$$\frac{m\cap n}{\min(m,n)}\geqslant\alpha \tag{12}$$

其中,分子表示交集,相同的图像数目,α 为一个参数,经实验证明,初始值为 0.3 反馈检索效果最好。

第 1 个条件表示从保存的分类器中,选择一个包含最多相关图像的分类器作为候选分类器;第 2 个条件表示检索中标记图像与分类器中保存的样本之间有相同样本的数量满足一定的比例。当这两个条件同时成立时才选择该分类器,否则采用短期相关反馈学习方法。

LBSVM-RF 的基本算法如下:

(1) 根据检索结果,用户标记正例样本 P^+ 或负例样本 P^-。

(2) 根据用户标记的正例样本 P^+ 或负例样本 P^-,利用上面的选择分类器算法寻找以前训练过的最相关的最好的分类器。

①如果找到了分类器 C,则对未知样本直接进行分类,根据分类函数值的大小反馈给用户;同时根据标记的正例样本 P^+ 或负例样本 P^-,结合分类器 C 以前的训练样本,重新的训练 BSVM 分类器,得到新的分类器并保存,留做下次使用。

②如果没有找到合适的分类器,则采用前面介绍的基于 BSVM-RF 方法进行短期相关反馈学习检索。

(3) 根据第(2)步的反馈结果,用户满意则结束,否则,跳转第(1)步执行。

基于长期学习相关反馈检索结构图如前文图 1 所示。

4 实验结果与分析

在 XP 系统下使用 Visual C++6.0 作为开发环境,图像库下载了 J.Z.Wang 的 SIMPLICITY 系统测试库,该库中有 1 000 幅图片 10 类,每类约有 100 幅图片。对传统的 SVM 方法、本章介绍的 BSVM 相关反馈方法以及 LBSVM-RF 进行了实验与对比分析。以检索"花"为例,图 4 给出了图像检索与相关反馈系统的界面。当用户第一次检索完图像后,根据用户标注的相关和不相关图像,作为 SVM 相关反馈方法的输入样本,三种相关反馈方法采用同一种输入样本,后面讨论分析都是以这次检索结果基础上进行。

查准率和查全率是基于内容的图像检索系统性能的两个重要的指标。为确定每一种反馈方法的性能,对 3 种不同的相关反馈方法的实验结果进行了统计分析,统计结果如表 1 所示:表 1 的实验数据表明,在第一次反馈检索过程中,每一种相关反馈方法所需的时间基本上很接近,都需要 1 秒左右。在查准率与查全率方面,提出的两种方法 BSVM 与 LBSVM 比 SVM 相

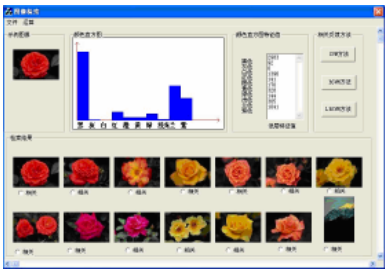

图 4 图像检索与相关反馈检索系统主界面

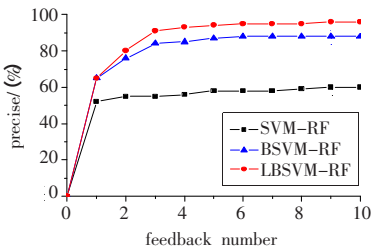

图 5 3 种不同方法 10 次反馈的查准率曲线图

关反馈方法要好。而 BSVM 与 LBSVM 这两种反馈方法几乎相同,主要是系统初期在第一次检索过程中,由于没有保留相关有用的反馈信息,因此长期学习的 LBSVM 反馈方法退化为短期的 BSVM 相关反馈方法。

表 1 系统第 1 次反馈后"花"的实验结果数据

	SVM	BSVM	LBSVM
查询时间/s	1.045	1.157	1.254
查准率/(%)	51	63	63
查全率/(%)	40	57	57

表 2 是采用同样的样本经过第三次反馈后的实验数据。从表中数据可以得知在查询时间方面,LBSVM 是最好的,仅仅需要 0.347 秒,比表 1 中第一次反馈的时间少约 3 倍,这主要的原因就是长期学习 LBBSVM 反馈算法中,它积累了前几次反馈的知识,当用户进行同类别的反馈时,它并不要进行学习训练,直接使用以前训练好的分类器进行分类,因此,它所需要的时间最短。在查准率和查准率方面,LBSVM 反馈方法,显著的提高,查准率达到了 95%,查全率也上升到 76%,主要是采用了 Boosting 方法,其次积累了以前反馈的信息,变相地增加了学习训练的样本,因此准确率也迅速提高。

表 2 系统第 3 次反馈后"花"的实验结果数据

	SVM	BSVM	LBSVM
查询时间/s	1.042	1.253	0.347
查准率/(%)	55	85	90
查全率/(%)	46	66	86

图 5 给出了提出的相关反馈方法和传统的 SVM 相关反馈方法经过 10 次反馈后的查准率性能对比图。从图中可以看出提出的 BSVM 和 LBSVM 相关反馈方法仅仅需要 2~3 次的反馈就可以达到 90%以上的查准率,而传统的 SVM 相关反馈方法经过 10 次反馈后仅能达到 60%的查准率。显然,进一步证明本方法明显的优于传统的 SVM 相关反馈方法。

5 小结

针对传统 SVM 相关反馈方法中样本数量少、正负样本不平衡,SVM 分类的准确性以及系统反馈速度等问题进行了改进,提出两种新颖的基于机器学习的相关反馈方法:结合 Boosting 方法与 SVM 的短期学习相关反馈方法,以及在此基础上提出的长期学习相关反馈方法。提出的短期学习相关反馈方法明显提高了系统反馈的准确率和查全率并为后面提出的长期学习相关反馈方法奠定了基础;而长期学习相关反馈提高了系统的反馈速度,同时也提高了系统反馈的准确率。这两种相关反馈方法相辅相成,实验表明,提出的方法在查准率方面达到 90%以上,而反馈速度是传统 SVM 相关反馈方法的 3 倍。

参考文献:

[1] Rui Y,Huang T,Mehrotra S,et al.A relevance feedback architecture for content based multimedia information systems[C]//Proceedings of IEEE Workshop on Content Based Access of Image and Video Libraries,1997,1:82-89.

[2] Kherfi M L,Ziou D.Image retrieval based on feature weighting and relevance feedback[C]//International Conference on Image Processing,2004,4:689-692.

[3] Rui Y,Huang T S.Optimizing learning in image retrieval[C]//Proceedings of IEEE Conference Computer Vision and Pattern Recognition,South Carolina,2000,2:236-243.

[4] Zhang L,Lin F,Zhang B.Support vector machine for image retrieval[C]//Proceedings of IEEE International Conference on Image Processing,2001,3:721-724.

[5] MacArthur S D,Brodley C E,Shyu C.Relevance feedback decision trees in content-based image retrieval[C]//Proceedings of IEEE Workshop on Content-Based Access to Image and Video Libraries,South Carolina,2000:68-72.

[6] Wood M E J,Campbell N W.Iterative refinement by relevance feedback in content based digital image retrieval[C]//Proceedings of ACM Multimedia,Bristol,U K,1998:13-20.

[7] 许月华,李金龙,陈恩红.一种新的基于 SVM 的相关反馈图像检索算法[J].计算机工程,2004,30(24):116-118.

[8] Freund Y.Boosting a weak learning algorithm by majority[J].Information and Computation,1995,121(2):256-285.

(上接 81 页)

参考文献:

[1] 戴国骏,葛海江,张翔.基于 Cypress PRoC 的多通道无线收发器的设计与实现[J].计算机工程与应用,2007,43(5):137-139.

[2] Davis G G.Cloning the IBM PS/2 series-legal and practical problems[C]//Thirty-Third IEEE Computer Society International Conference on Compcon Spring'88,Digest of Papers,29 Feb-3 March 1988:354-360.

[3] 李秀梅,李学华,陆坤.PS/2 协议的研究及其在单片机系统中的应用[J].微型机与应用,2003,2:22-23.

[4] Cypress Corporation.Wireless USB(TM) 2-Way HID Systems-AN4003.

[5] 李伟光,朱金华,赵博.PC 标准键盘在单片机系统中的应用[J].电测与仪表,2003,40(8):29-31.