集成学习SVM在图像检索中的应用

梁竞敏
LIANG Jing-min

广东女子职业技术学院 艺术设计与信息技术系,广州 511450
Department of Arts Design and Information Technology,Guangdong Women's Polytechnic College,Guangzhou 511450,China
E-mail:gzmliang@126.com

LIANG Jing-min.Application of integration learning SVM in image retrieval.Computer Engineering and Applications, 2009,45(18):182-184.

Abstract: An image retrieval method combining SVM and Adaboost algorithm is proposed.The proposed approach selects the most informative samples in database to train SVM,it can reduce the feedback rounds and the number of samples effectively,and it can use both advantages to improve the accuracy of image retrieval.At last Adaboost method is proposed to integrate studying with SVM,and it improves the image retrieval performance.The experiments show that the method work well solves the small sample size problem and it can improve the retrieval efficiency and performance consistently under the condition of limited training samples.

Key words: content-based image retrieval;support vector machines;integration learning;relevance feedback;Adaboost algorithm

摘 要:提出一种基于SVM和Adaboost集成学习相结合的相关反馈算法。在相关反馈过程中选择最具信息的样本训练支持向量机,可以有效减少相关反馈的次数和所需学习样本的数量,通过两者的互补来有效地提高图像检索的精度。最后提出Adaboost算法对SVM分类器进行加权投票,这样进一步提高了图像检索的性能。实验表明,该方法较好地解决了图像检索中的小样本选择问题,能够显著提高图像检索的效率和性能。

关键词:基于内容的图像检索;支持向量机;集成学习;相关反馈;Adaboost算法
DOI:10.3778/j.issn.1002-8331.2009.18.054 **文章编号:**1002-8331(2009)18-0182-03 **文献标识码:**A **中图分类号:**TP391

1 引言

目前检索系统采用的描述方法主要是基于低层视觉特征的方法,这些方法有两个主要的问题:一是存在高层概念和低层特征间的缺口[1]。在许多情况下,从高层概念到低层特征的映射无法由用户完成;二是人类视觉感知具有主观性。不同的人对相同的视觉场景,会有不同的感受。甚至同一个人在不同的场合下,对同一个视觉场景也可能有不同的视觉感受。这说明:期望通过一次搜索找到所需要的图像在具体应用中是不现实的[2]。此外,对系统检索结果的评价也具有主观性,所以检索过程需要用户的参与[3]。

当前相关反馈已成为图像检索技术中研究的热点,并且是图像检索中最具挑战性的研究方向[4]。相关反馈是图像检索中重要的一个环节,是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,哪些是"错误"的检索结果,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[5]。

目前将相关反馈问题看作一个二分类问题是一种较为普遍的学习模式。这种方法的基本思路是在检索过程中根据用户对图像的标注相关或不相关,动态学习一种分类器,用于将图像集分为相关和不相关两类,并将相关部分作为结果展现给用户。具体构造何种分类器取决于人们对先验信息的了解程度,不同的目标图像类分布的假定导致了不同的分类算法,较常用的是SVM分类器[6]。

提出一种基于SVM和Adaboost集成学习相结合的相关反馈算法。在相关反馈过程中采用主动学习方法选择最具信息的样本训练支持向量机,这样有效地减少相关反馈的次数和所需学习样本的数量,提高图像检索的精度;最后提出Adaboost算法对SVM分类器进行加权投票。实验表明,该方法具有较好的检索性能。

2 支持向量机

支持向量机最初是为解决两分类问题的。分类其实就是对样本相似性及相似程度的判断。为了说明支持向量机的原理,先从线性可分说起,然后再推广到线性不可分的情况。

2.1 支持向量机

设线性可分样本集为$\{(x_i,y_i)\}$,$i=1,2,\cdots,l$,$\boldsymbol{x}_i\in R^n$,$y_i\in$

基金项目:广东省科技计划项目工业攻关项目资助课题(No.2007B010200036)。
作者简介:梁竞敏(1974-),男,讲师,主要研究方向:计算机应用、企业信息化管理、图像检索与模式识别。
收稿日期:2008-09-19 **修回日期:**2008-12-15

{-1,+1},是类别符号。n 维空间中线性判别函数的一般形式为 $g(x)=\boldsymbol{w}\cdot\boldsymbol{x}+b$,其中 $\boldsymbol{x}$ 是输入向量,$\boldsymbol{w}$ 是权值向量,b 是阈值,分类线方程为 $\boldsymbol{w}\cdot\boldsymbol{x}+b=0$。将判别函数进行归一化,使两类所有样本都满足 $|g(\boldsymbol{x})|\geqslant 1$,也就是使得离分类面最近的样本的 $|g(\boldsymbol{x})|=1$,此时分类间隔等于 $2/\|\boldsymbol{w}\|$,因此使间隔最大等价于使 $\|\boldsymbol{w}\|$ 或 $\|\boldsymbol{w}\|^2$ 最小。要求分类线对所有样本正确分类,就是要求它满足:

$$y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]-1\geqslant 0,i=1,2,\cdots,l \tag{1}$$

满足上述条件(1),并且使 $\|\boldsymbol{w}\|^2$ 最小的分类面就叫做最优分类面,定义 N 个 Lagrange 算子 $\alpha_i,i=1,\cdots,N$。求解该二次优化问题,可以得到最优分类面,其中 $\boldsymbol{w}=\sum_{i=1}^{N}\alpha_i y_i\boldsymbol{x}_i$,$\boldsymbol{x}_i$ 是位于分类间隔面上的样本,称为支持向量[6]。分类函数为:

$$f(\boldsymbol{x})=sign\left(\sum_i \alpha_i y_i\boldsymbol{x}_i\cdot\boldsymbol{x}+b\right) \tag{2}$$

最优分类面是在线性可分的前提下讨论的,如果某些训练样本不能满足式(1)的条件,可以在条件中增加一个松弛项 $\xi_i\geqslant 0$ 和惩罚系数 C,使得式(1)变成:

$$y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]-1+\xi_i\geqslant 0,i=1,2,\cdots,l \tag{3}$$

求广义的最优分类面最大限度将样本分开,同时使分类间隔最大的问题可以表示为下面的二次规划问题:

$$\text{Min.}\phi(\boldsymbol{w},b)=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}\xi_i \tag{4}$$

约束性条件变为式(3),其中 ξ_i 可看作训练样本关于分离超平面的偏差,$\xi_i=0$ 时问题变为线性可分情形,C 为惩罚系数,它实际上起控制对错分样本惩罚程度的作用。求解这一优化问题的方法与求解最优分类面时的方法相同,一样转化为一个二次函数极值问题,其结果与可分情况下得到的几乎完全相同,约束条件变为:

$$0\leqslant\alpha_i\leqslant C,i=1,2,\cdots,l \tag{5}$$

可以看出,由于 C 的引入,α_i 的值受到限制,C 越小,α_i 的值也就越小。

对于非线性问题,SVM 通过选择适当的非线性变换,将输入空间 X 中的训练样本映射到某个高维特征空间 F,使得在目标高维空间中这些样本线性可分。根据泛函的有关理论,若核函数 $K(\boldsymbol{x},\boldsymbol{x}_i)$ 满足 Mercer 条件[7],它就对应某一变换空间中的内积 $<\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x})>$,函数 $\phi:X\to F$ 是一个从非线性输入空间 X 到高维特征空间 F 的映射,所以求映射 $\phi:X\to F$ 只要知道如何由输入 $\boldsymbol{x}$、$\boldsymbol{x}_i$ 计算内积 $<\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x})>$ 即可,由:

$$K(\boldsymbol{x}_i,\boldsymbol{x})=\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}) \tag{6}$$

将式(2)改写,即可得到对应高维空间的分类函数为:

$$f(x)=sign\left[\sum_{i=1}^{n}\alpha_i y_i K(\boldsymbol{x}_i\cdot\boldsymbol{x})+b\right] \tag{7}$$

这样,在高维空间的内积运算,转化为低维输入空间中的一个简单函数计算。

2.2 SVM 学习方法

目前 SVM 训练算法速度都比较慢,这是因为训练样本的数量决定了二次规划问题目标函数中矩阵的维数,使得求解规划问题的速度与维数呈指数增长。为了提高训练速度,减少学习样本数,缩短训练时间,将主动学习引入 SVM 中。主动学习与传统的被动学习的本质区别在于,它可以在候选的样本集中,主动选择对于当前分类器不确定度最大的新样本进行训练,来进一步设计新的分类器,从而使得可以用尽可能少的标注样本数来实现尽可能高的分类精度。

SVM 主动学习机包括两个独立的部分 (f,q),f 是一个 SVM 分类器,$f:X\to\{-1,1\}$ 使用训练样本集进行学习,q 是一个查询函数,根据训练样本集,决定下一步应从候选集 U 中选择哪一个样本进行标注。

SVM 主动学习机由查询函数 q 采取某种查询策略,从未标注的候选样本集 U 中选择下一个应标注的样本。根据泛函原理可知,对于线性可分问题,分类间隔中的样本对分类器的影响较大。因此,本文中 q 采用的查询策略为:每次选择离分类面最近的一个样本作为新样本进行训练。采用这种策略,每次选择进行学习的样本都是不确定性最大的样本,它对分类器的影响也最大,候选样本集中剩下的样本对分类器的影响逐步减弱。这种策略充分体现了 SVM 的本质,即分类器仅与支持向量有关,与其他向量无关,下面给出本文采用的 SVM 主动学习算法:

输入:未带类别标注的候选样本集 U,每次从 U 中采样个数为 1。

输出:分类器 f,预标注样本。

(1)从候选样本集 U 中选择 i 个样本并正确标注其类别,构造初始训练样本集 T,使 T 中至少包含一个输出 y 为 1 和一个输出 y 为-1 的样本;

(2)根据训练集 T 构造 SVM 分类器 f;

(3)对 U 中所有样本使用 f,标注为 (x,y),其中 y 为分类器 f 给向量 $\boldsymbol{x}$ 预先打上的标注;

(4)从样本集 U 中选择一个离分类边界最近的未标注样本 (x,y);

(5)将该样本正确标注后加入训练集 T 中(y 为 x 的正确标注);

(6)若检测精度达到某一设定值,算法终止,返回 f;否则重复第(2)步。

3 基于集成学习 SVM 的相关反馈

3.1 Adaboost 算法

Adaboost 算法的思想[8]是:给定一弱学习算法和一训练集 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,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,$y_i\in\{+1,-1\}$。初始化时,Adaboost 为训练集指定每个训练样本的权重都为 $\frac{1}{n}$。接着,调用弱学习算法进行 T 次迭代,每次迭代后,按照训练结果更新训练集上的分布,对于训练失败的训练样本赋予较大的权重,使得下一次迭代更加关注这些训练样本,从而得到一个预测函数序列 $h_1,h_2,\cdots,h_t$,每个预测函数 h_t 也赋予一个权重,预测效果好的,相应的权重越大。T 次迭代之后,在分类问题中最终的预测函数 H 采用带权重的投票法产生。Adaboost 算法[8]描述如下:

输入:$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,其中 $x_i\in X,y_i\in Y=\{+1,-1\}$。

初始化:$D_1(i)=\frac{1}{n}$　　//表示第 1 次迭代,训练样本权重为 $\frac{1}{n}$

for $t=1$ to T　　//T 为迭代次数

在 D_t 下训练:得到弱的假设:$h_t:X\to\{+1,-1\}$;　　//得到第 t 次预测函数

计算 h_t 的错误率:$E_t=\sum_{h_t(x_i)\neq y_i}D_t(i)$　　//错分样本

计算分类假设 h_t 权值:α_t　　//h_t 的权重

更新权值:$D_{t+1}=D_t\times F(E_t)$

输出最后的分类函数:$H(x)=sign(\sum_{i=1}^{T}\alpha_i h_t(x))$

3.2 相关反馈

为了缩小低层特征和高层语义之间的语义鸿沟,采用人机交互的相关反馈方法,首先用户对查询结果进行评价,标记出相关的图像,由于 SVM 分类精度不高等问题,通过 Adaboost 算法法对 SVM 分类器进行加权投票, 最后将图像按相似度的大小返回给用户。

3.2.1 基于 Adaboost 算法集成 SVM 的分类算法

(1)初始化样本权重:$D_1(i)=\frac{1}{n}$　//表示第 1 次迭代,训练样本权重为$\frac{1}{n}$

(2)对于 t=1 to T 进行迭代　//T 为迭代次数

在 $D_t(i)$下训练,使用 SVM 分类器训练得到弱分类假设:

$h_t:X\to\{+1,-1\}$;　//得到第 t 次预测函数

①计算 h_t 的错误率:$E_t=\sum_{h_t(x_i)\neq y_i}D_t(i)$　//错分样本

②计算分类假设 h_t 权值:$\alpha_t=\frac{1}{2}\ln\left(\frac{1-E_t}{E_t}\right)$　//h_t 的权重

③更新权值:$D_{t+1}(i)=\frac{D_t(i)}{Z_t}\times F(E_t)=\frac{D_t(i)}{Z_t}\times\begin{cases}e^{-\alpha_t} & \text{if } h_t(x_i)=Y_i\\ e^{\alpha_t} & \text{if } h_t(x_i)\neq Y_i\end{cases}$

//Z_t 为归一化因子

(3)最后输出分类函数:$H(x)=\sum_{t=1}^{T}\alpha_t h_t(x)$

3.2.2 相关反馈算法

首先定义一些标记:n:系统的检索规模,即界面显示的图像数目;I_p:正例集合;I_n:反例集合;I_p^1:当前反馈中的正例集合;I_n^1:当前反馈中的反例集合;β:衰减系数;(x_i,y_i):训练样本。其中 x_i 为图像,y_i 为对应的标注;$w(i)$:图像库中每幅图像对应的权值。详细的学习和检索算法如下:

(1)系统初始化。所有图像权值 $w(i)$=0,正例集合 I_p 和反例集合 I_n 为空集。在通过低层特征相似度计算的初始检索结果中选择 n 幅图像提供给用户[2]。

(2)用户标注与检索目标相关的图像,得到当前反馈结果中正例集合 I_p^1 和反例集合 I_n^1,用来更新集合 I_p 和反例集合 I_n:

$$I_p=(I_p\cup I_p^1)-I_n^1,\ I_n=(I_n\cup I_n^1)-I_p^1 \tag{8}$$

(3)分类器准备训练样本(x_i,y_i),$x_i=I_p\cup I_n$,$y_i=\begin{cases}+1, & \text{当 } x_i\in I_p \text{ 时}\\ -1, & \text{当 } x_i\in I_n \text{ 时}\end{cases}$

(4)利用 3.2.1 节中集成的 SVM 算法构造分类函数:

$$f(x)=\sum_{t=1}^{T}\alpha_t h_t(x) \tag{9}$$

(5)利用分类函数 $f(x)$的输出更新每一幅图像的权值:

$$w(i)=(1-\beta)w(i)+f(x_i),\quad 0\leqslant\beta\leqslant 1 \tag{10}$$

(6)根据权值将整个图像数据库以递减的顺序进行排序,选择前 n 幅图像提供给用户。如果用户已经满意则结束,否则回到步骤(2)。

4 实验结果

在 P4-2.4 G,RAM 1 G 的 PC 机上对一个具有 1 000 幅的图像库进行实验,该图像库分为 10 类,包括花、马、车、恐龙、人、风景等,每类有 100 幅图像。采用查准率(precision)、查全率(recall)和检索速度评价检索算法性能。将传统的 SVM 相关反馈方法(SVM-RF)和提出的方法进行对比实验,采用高斯核函数为支持向量机的核函数,$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\{-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{\sigma^2}\}$,$C$=1 000,$\sigma$=0.5。表 1 为两种方法一次相关反馈后的对比结果,从表 1 中可以看出,采用 Adaboost 算法对 SVM 进行加权投票后,反馈性能得到了很大的提高,查准率从 64%上升到 82%,上升了 28.13%,查全率从 52%上升到 70%,上升了 34.62%,这是因为采用 Adaboost 对 SVM 进行投票后进一步提高了分类准确率。提出方法的反馈时间为 3.866 s,比 SVM 反馈方法多 1.835 s,这是因为采用 Adaboost 方法进行迭代,花费的时间比 SVM 多。

表 1　第 1 次反馈后“花”的实验结果比较

	SVM 相关反馈	提出的方法
反馈时间/s	3.029	4.864
查准率/(%)	64	82
查全率/(%)	52	70

图 1 是 SVM 相关反馈方法经过 1 次相关反馈后返回的 12 幅图像, 图 2 是本文的方法在经过 1 次相关反馈后返回的 12 幅图像,第一幅为检索示例图像,从图 1 和图 2 中可以看出本文的相关反馈方法比 SVM 相关反馈算法较好。

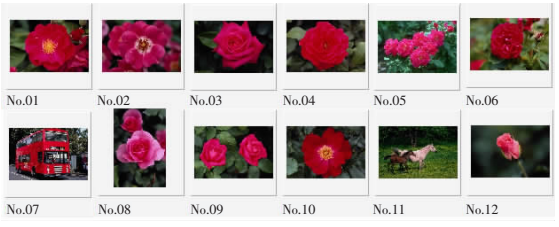

图 1　SVM 方法 1 次相关反馈的检索结果

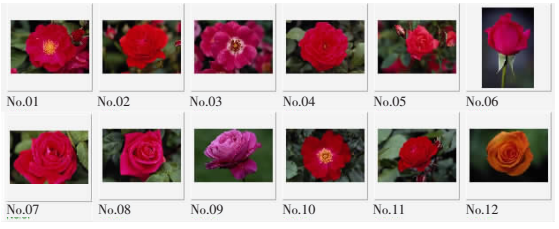

图 2　本文方法 1 次相关反馈的检索结果

5 结论

提出一种 SVM 和 Adaboost 集成学习相结合相关反馈算法。在相关反馈过程中选择最具信息的样本训练支持向量机,最后提出 Adaboost 算法对 SVM 分类器进行加权投票。实验表明,该方法显著提高图像检索的效率和性能。

参考文献:

[1] 官倩宁,田卉.基于 ROI 多特征和相关反馈的图像检索算法[J].计算机科学,2008,35(5):257-260.

[2] 赵海英,卢维娜.基于内容的交互式图像检索方法的实现[J].新疆师范大学学报,2008,27(1):63-66.

[3] Byoung Chul K O,Lee H S,Byun H.Region-based image retrieval system using efficient feature description[C]//Proceedings 15th International Conference on Pattern Recognition,Barcelona,Spain,2000,4:283-286.

(下转 230 页)