

# 基于记忆的 SVM 相关反馈算法

孙树亮 林雪云

(福建师范大学福清分校 福建 350300)

**摘要** 支持向量机(SVM)方法并不假设样本的分布条件,它基于结构风险最小化原则,对小样本情况下的学习问题给出最优解,并且在样本趋于无穷时能保持良好的一致收敛性。在 SVM 的基础上提出的 MSVM 方法,通过记忆功能,用历次反馈的累积样本代替一次反馈样本,从而增加了学习样本数量,减小了查准率的振荡,提高了检索精度;同时为了减轻用户负担,提出了记忆性标注。实验证明,MSVM 方法可以避免因训练样本集过小而出现的局部最小化的问题,能较为准确地分类图像库中的图像,同时有效地减轻了用户的负担。

**关键词** 支持向量机,反馈,记忆性标注,累积样本

**中图分类号** TP391.4 **文献标识码** A

## Relevance Feedback Algorithm Based on Memory Support Vector Machines

SUN Shu-liang LIN Xue-yun

(Dept. of Mathematics & Computer Science, Fuqing Branch of Fujian Normal University, Fujian 350300, China)

**Abstract** Support vector machine(SVM) is based on the minimum of structure risk and used for small samples in machine learning. Memory support vector machine(MSVM) feedback is based on SVM and used cumulation samples replacing feedback samples by memory. It reduces the risk of recall vibration. MSVM feedback also proposes memory label which is used for lightening user's burden. MSVM feedback is proved its superiority by relevant experiments.

**Keywords** Support vector machine, Feedback, Memory label, Cumulation sample

### 1 引言

20 世纪 90 年代中期,在文本检索领域中提出的相关反馈技术(RF)被引入到基于内容图像检索(CBIR)领域。相关反馈技术通过把人的参与引入到信息检索过程中,从而把检索模式从一次进行变成交互式地多次进行,并成为提高检索性能的有效方法。在相关反馈的交互过程中,只要求用户根据需求对系统当前的检索结果给出是否相关或者相关程度如何的判断,然后系统根据用户的反馈进行学习来给出更好的检索结果。

从 20 世纪 90 年代 RF 被引入 CBIR 至今,已取得了许多研究成果<sup>[1-3]</sup>。支持向量机(SVM)作为一种最有效的分类技术之一<sup>[4]</sup>,被广泛地应用到 RF 中。

### 2 支持向量机(SVM)

SVM 建立在统计学习理论的基础上,基于结构风险最小化原则,不需要特定问题的先验知识,在有限的训练样本情况下,可以很好地控制学习机器的推广能力,因而在图像检索中可以有效地改善检索结果。在每次反馈中对用户标记的正例和反例样本进行学习,建立 SVM 分类器作为模型,并根据该模型进行检索。

设有非线性映射  $\varphi: R^d \rightarrow H$ ,将输入空间的样本映射到高维的特征空间  $H$  中。那么在变换后空间内的分类超平面方程为:  $\varphi(x) \cdot w + b = 0$ ,类似于线性可分情况,将其转化为如

下的优化问题:在约束条件  $\sum_{i=1}^n y_i \alpha_i = 0$  和  $\alpha_i \geq 0, i = 1, 2, \dots, n$  下,对  $\alpha_i$  求解下列函数的最大值:

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \langle \varphi(x_i), \varphi(x_j) \rangle \quad (1)$$

令核函数  $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$ ,此时目标函数变为:

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (2)$$

而相应的分类函数也变为:

$$f(x) = \text{sign}(\sum_{i=1}^n \alpha_i y_i K(x_i \cdot x) + b)$$

判别函数可看成特征空间中某点到超平面距离的一种代数度量,去掉  $\text{sign}(g)$ ,转化形式后的分类函数为:

$$f(x) = \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \quad (3)$$

传统的基于 SVM 的 RF 算法将 RF 看成二值分类问题<sup>[5]</sup>。SVM 反馈的优点表现在以下几个方面:

1) SVM 是基于小样本学习的理论,能够寻找小样本情况下学习问题的最优解,不需要利用样本趋向于无穷大的渐进性条件。因而,即使是小样本数据,同样能得到较好的效果。

2) SVM 的方法便于寻找最优结构。构造 SVM 的过程实际上是求解一个凸二次优化问题,理论上可以得到全局最优解,不存在局部最优的问题。

3) 引入核函数可以将非线性问题转化为线性问题,并且有效地解决了“维数灾难”。

4) SVM 训练时间比较稳定。

到稿日期:2010-04-29 返修日期:2010-08-29

孙树亮(1982-),男,博士,主要研究方向为模式识别与智能系统、信息处理, E-mail: tjussl\_07@126.com; 林雪云(1976-),女,硕士,主要研究方向为数据库与信息处理。

### 3 基于记忆的 SVM 反馈算法的基本思想

尽管 SVM 反馈方法可在小样本条件下得到较满意的检索结果,但是对于用户需要标记的样本集的大小仍有一个下限。为了改善反馈中固有的小样本问题,即反馈中固有的正负样本不对称问题,本文提出了基于记忆的 SVM 相关反馈算法(memory support vector machine,MSVM)。本算法用历次反馈的累积样本集代替一次反馈样本作为学习样本集,可以提高训练样本集的数量,同时能够避免 SVM 反馈因训练样本的较大变化而导致检索结果正确率的振荡。同时,为了降低用户标注样本的疲劳性和人为误差,采用记忆性标注的方式,有效地提高了系统的检索性能。

### 4 基于记忆的 SVM 相关反馈算法

MSVM 方法的反馈过程与 SVM 方法的反馈过程相比,在以下两方面做了改进:

(1)训练样本集不仅仅是每次检索结果,而且是包含历次检索结果的积累决策样本集。

初始时,正例集合  $I_p$ 、反例集合  $I_n$  为空集。第  $l$  次反馈结果中正例集合  $I_p^l$ 、反例集合  $I_n^l$  用来更新正例集合  $I_p$  和反例集合  $I_n$ :

$$I_p = (I_p \cup I_p^l) - I_n^l, I_n = (I_n \cup I_n^l) - I_p^l \quad (4)$$

这样就可以防止样本由于不同的反馈次数而造成同时属于正例样本集与反例样本集这种情况。

(2)考虑到用户对反馈样本标注的疲劳性和不同时间对同一个样本的喜好程度的差异,提出系统记忆性标注的方法。

设第  $l$  次反馈结果中正例集合为  $I_p^l$  和反例集合为  $I_n^l$ ,前  $l-1$  次反馈结果中正例集合为  $I_p$  和反例集合为  $I_n$ ,系统返回图片的数量为  $n$ ,则称

$$S_M^l = (I_p^l \cup I_n^l) \cap (I_p \cup I_n) \quad (5)$$

为记忆样本集,即出现在本次检索结果中且之前已被用户标注过的样本集。这些检索结果在实际显示时由系统根据之前的标注方案自动显示其标注结果。而用户本次实际需要标注的样本数  $u^l$  是:

$$u^l = |I_p^l \cup I_n^l - I_p \cup I_n| \quad (6)$$

这样随着反馈次数的增加,需要标注的样本数越来越少,从而可以大大降低用户的疲劳度。

MSVM 反馈算法描述如下。

(1)根据相似度进行检索,在初始检索结果中选择  $n$  幅图像提供给用户。初始时,所有图像权值  $w(i)=0$ ,记忆样本集  $S_M$  为空集, $u^0=0$ ,正例集合  $I_p$  和反例集合  $I_n$  为空集。

(2)第  $l$  次记忆样本集  $S_M^l$  及需要标注的图像数  $u^l$  为:

$$S_M^l = (I_p^l \cup I_n^l) \cap (I_p \cup I_n)$$

$$u^l = |I_p^l \cup I_n^l - I_p \cup I_n|$$

(3)得到第  $l$  次反馈结果中正例集合  $I_p^l$  和反例集合  $I_n^l$ ,用来更新正例集合  $I_p$  和反例集合  $I_n$ :

$$I_p = (I_p \cup I_p^l) - I_n^l, I_n = (I_n \cup I_n^l) - I_p^l$$

(4)分类器训练样本集  $(x_i, y_i)$ :

$$x_i \in I_p \cup I_n, y_i = \begin{cases} +1, & \text{for } x_i \in I_p \\ -1, & \text{for } x_i \in I_n \end{cases} \quad (7)$$

(5)SVM 对训练样本进行训练:

$$f(x) = \sum_i a_i y_i K(x_i, x) + b$$

(6)利用分类函数  $f(x)$  的输出更新每一幅图像的权值:

$$w(i) = (1 - \beta)w(i) + f(x_i) \quad (8)$$

式中, $\beta$  为衰减系数,且  $0 \leq \beta \leq 1$ 。

(7)根据权值对图像库进行排序,返回检索图像。若用户对该次检索结果不满意,则返回到(2),继续学习。若满意,则结束。

### 5 实验结果与分析

在图像库中用 1000 幅图像进行实验,包括山川、人物、动物、花等 10 类图片,每类 100 幅左右。检索特征使用颜色与形状特征, $\rho=1, C=100$ ,返回图像的数量 NRT 分别是 11, 20, 30, 40。

实验结果为相同条件下 5 次实验的平均值。

由图 1 可见:

(1)NRT 等于 11 时,SVM 和 MSVM 出现震荡的现象。这是由于 SVM 采用二分类的方法,每次检索的结果中必然把反例排在结果的后面,因此下一次学习时所用的反例有很大的不同,而有些正例样本与反例在特征上可能是很相似的,因而训练样本的较大变化导致了正例样本的查准率产生较大的振荡。基于 MSVM 方法的检索系统在检索过程中也存在一个性能波动的阶段,经过学习后 MSVM 的检索效果会较 SVM 方法波动更小,而且更快趋于稳定。这是因为 MSVM 方法记忆了历次检索结果,用这样的集合作为训练样本集,能在很大程度上避免这种现象的发生。

(2)由 SVM 和 MSVM 曲线可以看出,查准率随着返回样本数量的增加而提高,这是因为样本数越多,构造的分类器越准确。

(3)在返回相同数量样本的条件下,MSVM 比 SVM 方法有更好的查准率,这是因为 MSVM 具有记忆功能,用历次反馈的累积样本代替一次反馈样本,因此用来构造分类器的样本数量更多,也就具有更好的查准率。

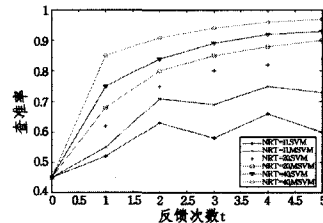


图 1 SVM 和 MSVM 在不同的返回图像数条件下反馈次数与查准率的关系

从图 2 可知,若用户标记的样本数越多,则 SVM 和 MSVM 的泛化能力越强,同时用户标记图像的负担也越大。在返回相同数量的样本条件下,MSVM 方法平均标注样本数更少,可以有效减轻用户的负担。

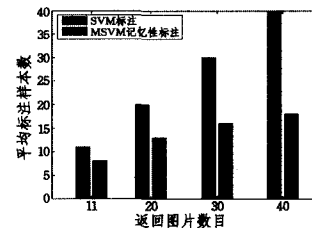


图 2 SVM 和 MSVM 在 3 次反馈条件下平均标注样本数的比较

美国南加州大学的 USI-SIPI image database<sup>[6]</sup>,从 leaf-lesstrees 类中随机取一幅图片进行检索,检索方法分别是 SVM 反馈法和 MSVM 反馈法。反馈 3 次,第一幅是待检索图片,系统返回 11 幅查询图片。检索结果如图 3 和图 4 所示。

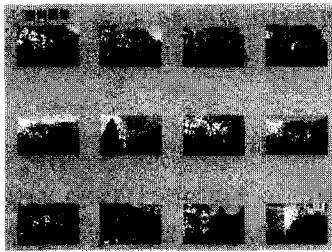


图3 基于 SVM 反馈的检索

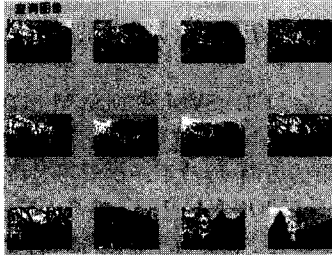


图4 基于 MSVM 反馈的检索

通过对比图3和图4可以看出,经过3次反馈后,MSVM方法比SVM方法有更好的检索效果,图像检索的精确度得到了很大的提高,从而验证了MSVM方法的有效性。

**结束语** SVM方法并不假设样本的分布条件,它基于结构风险最小化原则,对小样本情况下的学习问题给出最优解,并且在样本趋于无穷时可以保持良好的一致收敛性。

MSVM方法是在SVM的基础上,通过记忆功能,用历次反馈的累积样本代替一次反馈样本,从而增加了学习样本数量,减小了查准率的振荡,提高了检索精度。同时为了减轻用户负担,本文提出了记忆性标注。实验证明,MSVM方法可以避免因训练样本集过小而出现的局部最小化的问题,能较为准确地分类图像库中的图像,同时有效地减轻了用户的负担。

### 参考文献

- [1] Rui Y, Huang T S. Optimizing learning in image retrieval[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Hilton Head Island; IEEE, 2000; 236-243
- [2] 周建新, 高科, 李锦涛, 等. 图像检索中一种有效的 SVM 相关反馈算法[J]. 计算机辅助设计与图形学学报, 2007, 19(4): 535-540
- [3] Zhon X S, Huang T S. Relevance feedback in image retrieval: a comprehensive review[J]. Multimedia Systems Journal, 2003, 8(6): 536-544
- [4] 正涛, 樊孝忠, 郭剑毅. 基于支持向量机的汉语问句分类研究[J]. 华南理工大学学报: 自然科学版, 2005, 33(9): 25-29
- [5] Hong P, Tian Q, Huang T S. Incorporate support vector machine to content-based image retrieval with relevance feedback [C]// Proceedings of International Conference on Image Processing. Vancouver; IEEE, 2000; 750-753
- [6] <http://sipi.usc.edu/services/database/index.html>

(上接第242页)

一确率,如表5所列。

表5 单个特征属性的识别准确率

特征属性	识别率
头部动作	0.571
关注度	0.521
说话语气	0.603
说话时间长度	0.613
交互时机	0.493
上次会话角色类型	0.390
关键词	0.709

从表5可知,不同特征属性对最后的识别率的影响不同,仅用“关键词”特征属性进行测试所得的测试准确率最高,为70.9%;而仅用“上次会话角色类型”特征属性进行测试所得的测试准确率最低,为39%。识别准确率从高到低依次为关键词、说话时间长度、说话语气、头部动作、关注度、交互时机、上次会话角色类型。

### 4.3 不同特征属性组合的识别结果

表6 不同特征属性组合的识别结果

特征属性组合	识别率
{f1, f2}	57.1%
{f3, f4}	71.6%
{f5, f6}	42.9%
{f1, f2, f3, f4}	72.0%
{f1, f2, f5, f6}	48.2%
{f3, f4, f5, f6}	65.6%
{f1, f2, f7}	70.9%
{f3, f4, f7}	79.1%
{f5, f6, f7}	67.4%
{f1, f2, f3, f4, f5, f6}	67.7%
{f1, f2, f3, f4, f5, f6, f7}	80.1%

注: f1: 头部动作, f2: 关注度, f3: 说话语气, f4: 说话时间长度, f5: 交互时机, f6: 上次会话角色类型, f7: 关键词

将不同的特征属性带入朴素贝叶斯模型,对测试集进行测试,得到的实验结果如表6所列。

从表6可以得到如下结论,不同特征属性的组合所得到的识别准确率有一定的差异,而选取7个特征属性的组合所计算出的识别准确率最高,为80.1%,证明所选取特征属性的合理性。**结束语** 本文从获取会议中群体交互语义的角度出发,针对当前智能会议系统只集中于物理交互研究的情况,采用朴素贝叶斯模型,设计并实现了群体交互语义的获取方法。实验结果表明,该方法在群体交互语义的识别方面能够取得不错的效果,识别准确率达到80.1%。但该方法在处理数据上还有诸多不足之处,有待进一步优化,如会话记录中的关键字处理,只是将训练样本集中的关键字数据做简单分类,并没有涵盖全部。总体而言,该方法基本能达到预期的效果,具有一定的适用性和有效性。

### 参考文献

- [1] Yu Zhi-wen, Yu Zhi-yong, Aoyama H, et al. Capture, Recognition, and Visualization of Human Semantic Interactions in Meetings[C]// Pervasive Computing and Communications 2010 (PerCom 2010), 2010; 107-115
- [2] Dey A K, Salber D, Abowd G D, et al. The Conference Assistant: Combining Context-awareness with Wearable Computing [C]// ISWC'99. 1999; 21-28
- [3] Mikic I, Huang K, Trivedi M. Activity Monitoring and Summarization for an Intelligent Meeting Room[C]// IEEE Workshop on Human Motion, 2000; 107-112
- [4] 余意, 易建强, 赵冬斌. 智能空间研究综述[J]. 计算机科学, 2008, 35(8): 1-3
- [5] (美) Tan Pang-ning, Steinbach M, Kumar V. 数据挖掘导论 [M]. 范明, 范宏建, 等译. 北京: 人民邮电出版社, 2006; 141-145