文章编号:0438-0479(2002)06-0687-05

一种新的支持向量机增量学习算法

曾文华[1],马 健[2]
(1. 厦门大学计算机科学系,福建 厦门 361005; 2. 杭州电子工业学院计算机分院,浙江 杭州 310037)

摘要: 提出一种新的支持向量机增量学习算法. 分析了新样本加入训练集后,支持向量集的变化情况. 基于分析结论提出新的学习算法. 算法舍弃对最终结论无用的样本,使得学习对象的知识得到了积累. 实验结果表明本算法在保证分类准确度的同时,在增量学习问题上比传统的支持向量机有效.

关键词: 支持向量机;增量学习;学习算法
中图分类号: TP 181 **文献标识码**:A

增量学习技术作为一种智能知识发现技术,已经得到了广泛的研究. 它与传统的学习技术相比,优越性在于它不仅可以舍弃无用样本并减小训练集,而且可以充分利用学习的历史结果,使学习具有了延续性. 很多学者基于传统的学习方法提出了新的增量学习算法[1,2]. 但是由于传统学习算法不能保证很好的泛化能力,常常陷于对问题的过学习和局部最小等现象,因而基于传统学习方法的增量学习算法通常得不到原问题较好的结果.

支持向量机(Support Vector Machine —SVM)是一种新的机器学习技术,由 Vapnik 和他的同事于 1995 年提出[3]. 基于统计学习理论的坚实基础,SVM 有着很强的学习能力和较好的泛化性能. SVM 学习采用优化方法得到的结果是全局最优解,不会产生传统方法中的过学习和局部最小等问题. SVM 学习结果为支持向量集,通常是学习样本集的一小部分,支持向量集充分体现了整个样本集的属性. 本文基于 SVM 寻优问题的 KKT 条件和样本之间的关系,分析了样本增加后支持向量集的变化情况,基于分析结论提出了一种新的 SVM 增量学习算法.

1 支持向量机理论

1.1 支持向量机

给定分类问题,其样本集为$\{X_i, y_i\}$, $i = 1, \cdots, l$, $X_i \in R^n$, $y_i \in \{\pm 1\}$. 支持向量机即为线性分类器,通过构造最优分类面,使得类别间的分类间隔最大[4]. 较大的分类间隔意味着分类器具有较好的泛化能力.

设超平面方程为:

$$wX + b = 0 \tag{1}$$

则决策函数为:

$$f(x) = \mathrm{sgn}(wX + b) \tag{2}$$

求最大分类间隔的超平面问题可以转化为如下二次寻优问题:

$$\begin{aligned} &\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i \\ &\text{s.t.} \quad y_i (wX_i + b) \geq 1 - \xi_i \\ &X_i \in R^n, \ y_i \in \{\pm 1\}, \ i = 1, \cdots, l \end{aligned} \tag{3}$$

其对偶问题为:

$$\begin{aligned} &\max \sum_{i=1}^{l} \alpha_i - \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j X_i X_j \\ &\text{s.t.} \quad 0 \leq \alpha_i \leq C \\ &\sum_{i=1}^{l} \alpha_i y_i = 0, \ i = 1, \cdots, l \end{aligned} \tag{4}$$

α为 Lagrange 乘子. 由上述问题得到最优解 α,则 SVM 的分类函数为:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i (X \cdot X_i) + b \right) \tag{5}$$

如果分类问题是非线性的,则采用一种称为核函数(Kernel Function)的方法,使输入空间映射到高维的核函数特征空间,将非线性问题转化为该空间中的线性分类问题. 根据泛函理论,特征空间中对偶问题和得到的决策函数中的点积可以由输入空间中的核函数来替换. 此时对偶问题为:

收稿日期:2002-04-25
基金项目:工业控制技术国家重点开放实验室资助项目(K01001)
作者简介:曾文华(1964 -),男,博士,教授.

$$\max \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j X_i X_j \tag{6}$$

决策函数为:

$$f(X) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(X_i, X) + b \right) \tag{7}$$

其中 $K(\ )$ 为核函数.

1.2 KKT 条件

对偶问题的最优解 $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_l]$ 使得每个样本 X 满足优化问题的 KKT 条件[5]:

$$\alpha_i = 0 \Rightarrow y_i f(X_i) \geq 1 \tag{8}$$

$$0 < \alpha_i < C \Rightarrow y_i f(X_i) = 1 \tag{9}$$

$$\alpha_i = C \Rightarrow y_i f(X_i) \leq 1 \tag{10}$$

其中非零的 α_i 为支持向量. 考虑函数系 $f(X) = h$, 可知 $f(X) = 0$ 为分类面, $f(X) = \pm 1$ 为分类间隔的边界, 其上的样本即为支持向量.

定理1 对样本集进行训练得到SVM分类器, α 为Lagrange乘子. $\alpha = 0$ 对应的样本分布在分类器分类间隔之外, $0 \leq \alpha \leq C$ 的样本位于分类间隔之上, $\alpha = C$ 位于分类间隔之内[5]. 也即是:

$$\alpha_i = 0 \Rightarrow | f(X_i) | \geq 1 \tag{11}$$

$$0 < \alpha_i < C \Rightarrow | f(X_i) | = 1 \tag{12}$$

$$\alpha_i = C \Rightarrow | f(X_i) | \leq 1 \tag{13}$$

2 增量学习后支持向量集变化分析

当核函数类型及其参数确定后, 支持向量集可以完全描述整个样本集的分类特征, 支持向量集和训练样本集之间的等价关系可以得到证明. 通常支持向量集只是样本集的一小部分, 因此通过对支持向量集的研究增量学习是可行且有效的. 如果新增样本带有原样本集不包含的分类信息, 则学习后支持向量集必然发生变化以体现新信息的加入. 如果考虑新增样本对支持向量集的影响, 则应该考虑以下问题:

1) 什么样的新增样本使得支持向量集发生变化?

2) 这种变化是怎样的?

3) 新增样本集是如何组成的?

给定原 SVM 分类器:

$$f(X) = \sum_{i=1}^{l} \alpha_i y_i K(X_i, X) + b \tag{14}$$

$y = \operatorname{sgn}(f(X))$ 和新的样本集 $A = \{X_i, y_i\}$, $i = l+1, \cdots, l_n$. 由于支持向量集由二次寻优问题得到, 因此如果样本集的改变意味着新增样本集中含有违背 KKT 条件的样本, 并且这些样本中有些转化为支持向量, 新增样本加入时其 Lagrange 乘子 $\alpha = 0$. 根据式(8), 如果其满足 KKT 条件, 则 $y_i f(X_i) \geq 1$; 否则, $y_i f(X_i) < 1$, $i = 1, \cdots, l$. 可得:

$$-1 < y_i f(X_i) < 1 \tag{15}$$

$$y_i f(X_i) < -1 \tag{16}$$

式(15) 表明样本可能位于原来分类间隔中; 式(16) 表明样本可能位于另一类的分类区域内, 如图 1 所示. 因此可以得到如下结论:

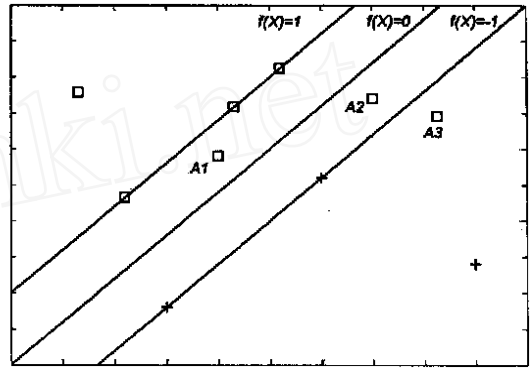

图 1 新增样本与 KKT 条件的关系

方形标记代表‘y = +1’, 十字标记代表‘y = -1’. A1, A2 和 A3 是新增样本

Fig. 1 The relation between new samples and KKT condition

定理2 $f(X)$ 为 SVM 分类决策函数, $\{X_i, y_i\}$ 为新增样本. 满足 KKT 条件的新增样本将不会改变支持向量集, 违背 KKT 条件的新增样本将使支持向量集发生变化. 违背 KKT 条件的样本可以分为三类:

1) 位于分类间隔中, 与本类在分类边界同侧, 可以被原分类器正确分类的样本, 满足 $0 \leq y_i f(X_i) < 1$;

2) 位于分类间隔中, 与本类在分类边界异侧, 被原分类器分类错误的样本, 满足 $-1 \leq y_i f(X_i) \leq 0$;

3) 位于分类间隔外, 与本类在分类间隔异侧, 被原分类器分类错误的样本, 满足 $y_i f(X_i) < -1$.

定理 2 表明, 对新增样本再学习得到新的 SVM 分类器时, KKT 条件比分类函数的分类判断更合理, 分类错误是样本违反 KKT 条件的特定情况. 定理 2 告诉我们, 只有违背 KKT 条件的样本才会影响增量学习后的支持向量集. 因此新增样本可以分为两部

分:违背 KKT 条件的样本和满足 KKT 条件的样本,而后者由于其包含的信息已经为原分类器所反映,因此可以不被学习.

定理 3　新增样本违背 KKT 条件,则原样本集中非支持向量可能转化为支持向量[5].

定理 3 的证明可以由图 2 所示的特例来说明. 原来的支持向量集由 $S1$～$S5$ 组成,$A1$～$A3$ 为新增样本. 对新样本训练过以后由图 2 可以很直观的判断:$S1$,$A3$ 和原来的样本 $N1$ 组成新的支持向量集. 因此增量学习中,只考虑新增样本和原来的支持向量集的做法虽然符合定理 2,但是定理 3 表明这样可能会丢失原来样本集中的信息,而且也不能十分有效的舍弃无用样本.

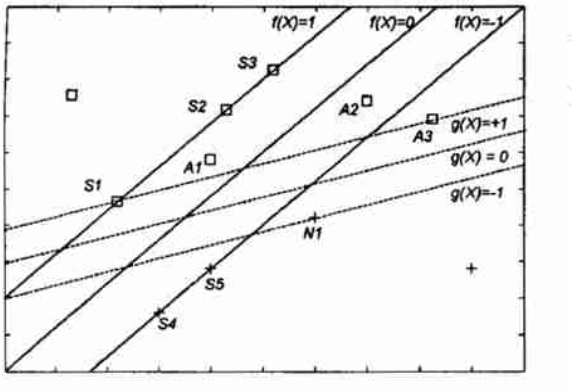

图 2　对新增样本训练后支持向量集可能发生的变化　$S1$～$S5$ 为原支持向量集,$A1$～$A3$ 为新增样本集. 训练后的支持向量集由 $S1$,$A3$ 和 $N1$ 组成,其中 $N1$ 为原样本集中的样本

Fig. 2　The possible change of SV set after training on incremental set

3　一种新的 SVM 增量学习算法

实际上,我们不是把新增样本加入原来的样本集,而是将它们合在一起,两者的地位是等同的. 尽管通常情况下,原来的样本集比新增样本集大. 因此如果由新增样本集训练得到一个新的 SVM,则原来样本集可以看作新 SVM 的新增样本,进而其中也可能存在违背新 SVM 的 KKT 条件的样本,而这些样本可能成为增量学习后的支持向量.

3.1　增量学习方案

基于以上分析,我们提出一种新的增量学习方案. 首先检验是否存在新增样本违背原来的 KKT 条件,如果存在,则由新增样本得到一个新的 SVM. 检验是否存在原来的样本违背新的 KKT 条件,如果仍然存在,则由两个 SVM 的支持向量集加上两次违背条件的样本,构成新的训练集. 最后由新的训练集得到最终的 SVM.

本方案中考虑的不仅是分类错误,而是样本是否违背 SVM 的 KKT 条件,原来样本集合中新的样本集的信息都被获取. 而不违背 KKT 条件的样本被舍弃,不作考虑,而可能有价值的样本得到保留,用来进行新的训练.

在方案中,可能需要对新增样本单独作一次训练,来得到新的 SVM,而整个过程对全部样本只做一次违背 KKT 条件的验证. 如果 SVM 的决策函数为 $f(X_i)$,$y_i \in \{\pm 1\}$ 为类别号,违背 KKT 条件等价于 $y_i f(X_i) < 1$. 接着将给出一种增量学习算法,由于多类问题可以转化为多个二类问题,因此,这里只考虑基本的二类问题.

3.2　增量学习算法

前提:

原样本集 X_0 和由其训练得到的 SVM 分类器 Ω^0,X_0^{SV} 表示 Ω^0 的支持向量集,X_I 是新增样本.

算法:

1) 检验 X_I 中的样本是否违背 Ω^0 的 KKT 条件,如果没有样本违背,则算法停止,Ω^0 为增量学习结果. 否则,根据检验结果,X_I 被分为 X_I^V 和 X_I^S,X_I^V 表示违背 Ω^0 的 KKT 条件的样本集合,X_I^S 表示满足 Ω^0 的 KKT 条件的样本集合.

2) 由 X_I 得到新的 SVM 分类器 Ω^I,X_I^{SV} 表示 Ω^I 的支持向量集.

3) 检验 X_0 中的样本是否违背 Ω^I 的 KKT 条件,如果没有样本违背,则算法停止,Ω^I 为增量学习结果. 否则 X_0 被分为 X_0^V 和 X_0^S 两部分,X_0^V 表示违背 Ω^I 的 KKT 条件的样本集,X_0^S 表示满足 Ω^I 的 KKT 条件的样本集.

4) X_U 表示 $X_0^{SV} \cup X_I^{SV} \cup X_0^V \cup X_I^V$,由 X_U 得到新的 SVM 分类器 Ω,则 Ω 为增量学习结果,$X_H = X_0^S \cup X_I^S$ 作为历史信息将被舍弃.

4　实验和讨论

基于以上研究,我们实现了算法,并对 Heart Scale 数据库[6] 进行了实验,同时讨论了一些相关问题.

4.1 实验和结果

在 LibSVM[6] 函数库的基础上实现了以上算法,实验采用的 Heart Scale 数据库包含了 270 个 13 维向量数据. 样本被标为两类:120 个样本为 + 1 类,其余 150 个为 - 1 类.

实验设计如下,将实验数据分为三组: *H*1, *H*2 和 *H*3,每组分别包含 100,100 和 70 个样本,顺次作为新增样本集进行训练. 在 SVM 训练中,C 取1 000,采用高斯核函数类型. 我们将本算法与传统的 SVM 算法进行了比较,实验结果见表 1 和表 2.

表 1 增量学习算法与传统 SVM 算法在舍弃样本效果的比较(C = 1 000)

Tab. 1 Comparison on effect of discarded samples between our incremental learning result and the traditional algorithm (C = 1 000)

新增样本集	训练集大小		支持向量集大小	
	T	*I*	*T*	*I*
*H*1	100	100	52	52
*H*2	200	107	84	74
*H*3	270	127	101	93

表 2 增量学习算法与传统 SVM 算法学习能力和泛化能力比较

Tab. 2 Comparison on the learning result and predict accuracy

新增样本集	*H*1		*H*2		*H*3	
	T	*I*	*T*	*I*	*T*	*I*
*H*1	100.00	100.00	78.00	78.00	80.00	80.00
*H*2	100.00	99.00	100.00	99.00	81.43	82.86
*H*3	100.00	98.00	100.00	98.00	100.00	98.57

注:表中数据为学习后对数据集的预测准确度(%)

4.2 讨 论

从表 1 和表 2 可知,本增量学习算法能够舍弃无用样本,保留了几乎所有可能变成支持向量的样本,同时保证了分类的精度. 当新增样本被加入后,传统方法的训练集只是原样本和新增样本的合并. 随着样本的增加而增加,而本增量学习算法的训练集并不随着样本的增加而变大,并且可以观察到其变化是随着支持向量集的大小的增加作出相应的变化. 这意味着我们的算法可以不断的发现新的支持向量,被认为是无用样本的舍弃并没有影响学习的效果. 表 2 所示,本算法具有与传统 SVM 算法相同的学习能力和泛化性能.

5 结论和进一步的研究

我们讨论了样本与 KKT 条件之间的关系,基于讨论结果,分析了新增样本加入后,可能引起的支持向量集的变化,进而提出了一种新的 SVM 增量学习算法. 实验证明,本算法可以在新增样本增加后有效压缩样本集的大小,舍弃无用样本,同时分类能力和泛化性能不受影响.

对于某种问题,经过几次增量学习后,新增样本含有的有用信息将越来越少,这时采用增量学习的开销会收效甚微. 如何提取少量信息将是一个需要进一步研究的问题. 在本算法中,新样本被先后顺次训练,而当训练集为一个大型的样本库时,可以将其分为几个合适的小部分,然后采用对各部分并行学习方法. 因此本算法在并行学习上的应用也是今后进一步研究的方向.

参考文献:

[1] Ratsaby J. Incremental learning with sample queries[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8):883 - 888.

[2] Wang E H C, Kuh A. A smart algorithm for incremental learning[J]. International Joint Conference on Neural Networks, 1992, 3:121 - 126.

[3] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.

[4] Christopher J, Burges C. A Tutorial on Support Vector Machines for Pattern Recognition[M]. Boston: Kluwer Academic Publishers, 1998.

[5] 周伟达,张 莉,焦李成. 支撑矢量机推广能力分析[J]. 电子学报,2001,29(5):590 - 594.

[6] Chang Chih-chung, Lin Chih-jen. LIBSVM: a Library for Support Vector Machines. Available: http://citeseer.nj.nec.com/chang01libsvm.html, 2001-09-07.

A Novel Approach to Incremental SVM Learning Algorithm

ZENG Wen-hua[1], MA Jian[2]

(1. Department of Computer Science, Xiamen University, Xiamen 361005, China;
2. School of Computer Science, Hangzhou Institute of Electronics
Engineering, Hangzhou 310037, China)

Abstract: This paper presents a novel approach to incremental support vector machine (SVM) learning algorithm. It analyses the possible change of support vector set after new samples are added to training set. Based on the analysis result, a novel algorithm is presented. In this algorithm useless sample is discarded and knowledge is accumulated. The experiment result shows that this algorithm is more effective than traditional SVM while the classification precision is also guaranteed.

Key words: support vector machine; incremental learning; learning algorithm

2001 年度我校获部省市级科技奖项目情况

序号	项目名称	获奖类别	获奖等级
1	过渡金属电极体系的表面增强拉曼光谱研究和应用	中国高校科学技术奖	1 等
2	新手性配体、新手性催化剂的分子设计与在不对称催化中的应用	中国高校科学技术奖	2 等
3	金属元素及复杂形态分析联用新技术与新方法	福建省科技进步奖	1 等
4	基于膜分离过程 6-APA 生产技术	福建省科技进步奖	2 等
5	随机度量理论及其应用	福建省科技进步奖	2 等
6	Hammock 的分解和 Nazarova-Roiter, Zavadskij 算法	福建省科技进步奖	2 等
7	微污染水源水生物接触氧化—气浮工艺制水技术研究	福建省科技进步奖	3 等
8	环境与生命物质流动体系的电致化学发光研究	福建省科技进步奖	3 等
9	铁蛋白铁核结构与新功能研究	福建省科技进步奖	3 等
10	铜基甲醇合成催化剂各组分的协同催化作用机理研究	福建省科技进步奖	3 等
11	HIV 基因工程重组抗原及第三代 HIV1 + 2 抗体 ELISA 试剂盒的研制	厦门市科技进步奖	重大贡献
12	嵌入式 TCP/IP 芯片	厦门市科技进步奖	2 等
13	大黄鱼养殖病害防治技术研究	厦门市科技进步奖	2 等
14	微污染水源水生物接触氧化—气浮工艺制水技术研究	厦门市科技进步奖	3 等
15	8 种棕榈植物耐盐性的筛选试验	厦门市科技进步奖	3 等