文章编号: 1673-5196(2007)04-0090-03

C-SVM 在不同类别样本数目不均衡下的优化

张秋余[1], 赵付清[1], 王 静[1], 余冬梅[1], 李建建[1], 张润花[2]

(1. 兰州理工大学 计算机与通信学院, 甘肃 兰州 730050; 2. 甘肃建筑职业技术学院, 甘肃 兰州 730050)

摘要: 在解决故障检测等分类问题时, 若不同类别样本数目相差很大, C-SVM 训练的分类错误总偏向于样本数较少的类别, 因而影响了分类的精确性. 为提高精确性, 提出一种优化算法, 在训练过程中针对不同类样本, 采用不同的权值来优化训练过程, 按正负类样本在总样本中所占的比例, 加大样本数较少的类别权值, 降低样本数较大的类别权值来实现两类样本间的均衡. 实验结果表明, 该方法对两类样本数目相差很大的问题有效.

关键词: C-SVM; 不均衡样本数; 参数优化; 加权

中图分类号: TP391 **文献标识码**: A

Optimization of C-SVM in case of samples with unequal numbers in their different varieties

ZHANG Qiu-yu[1], ZHAO Fu-qing[1], WANG Jing[1], YU Dong-mei[1]
LI Jian-jian[1], ZHANG Run-hua[2]

(1. College of Computer and Communication, Lanzhou Univ. of Tech., Lanzhou 730050, China; 2. Gansu Construction Vocational Technical College, Lanzhou 730050, China)

Abstract: In solving the problem of trouble-locating in the case of samples with great difference in their number for their different varieties, the training with C-SVM was undesirably under bias towards those varieties with fewer samples, so that the training accuracy was unsatisfactory. In order to improve its accuracy, an optimization algorithm was proposed based on taking different weights for different classes in the process of training. According to the proportion of positive and negative samples in the total samples, the weight for the minor variety with fewer numbers of samples was increased and the other decreased, so that the balance between two samples varieties was realized. It was showed by experiments that the proposed approach could improve the accuracy of classification.

Key words: C-SVM; unequal sample numbers; parameter optimization; weighting

支持向量(SVM, support vector machine)[1~3]是 Cortes & Vapnik 在 1995 年首先提出的机器学习方法[4,5]. 该方法是建立在有限样本尤其是小样本情况下、基于统计学习理论的 VC 维理论和结构风险最小化原则基础之上, 而不是基于传统的经验风险最小化原则建立起来的. 对于非线性问题, 首先可以通过巧妙地构造一个适当的内积函数, 将输入空间中的非线性问题转化为某个高维空间中的线性问题, 然后在这个新空间中求取最优线性分类面.

在 C-SVM 算法中, 对于不同类别数目差异很大的情况, 分类效果并不是很好. 为了提高算法的分类精确性, 本文对不同类别的样本采用不同的权值, 最后通过实验证明该算法的有效性.

1 支持向量机的基本原理

SVM 是从线性可分的最优分类面发展而来的, 基本原理如图 1 所示. 在图中, 实心点和空心点代表两类样本, H 为分类面, H_1、H_2 分别为各类中离分类线最近的样本且平行于分类线的直线, 它们之间的距离叫做分类间隔(margin). 所谓最优分类线就是要求分类线不但能将两类正确分开(训练错误率为 0), 而且要使分类间隔最大. 分类线方程为 $x \cdot w + b = 0$, 可以对它进行归一化, 使得对线性可

收稿日期: 2006-11-09
基金项目: 甘肃省科技攻关项目(2GS047-A52-002-03)
作者简介: 张秋余(1966-), 男, 河北辛集人, 副研究员.

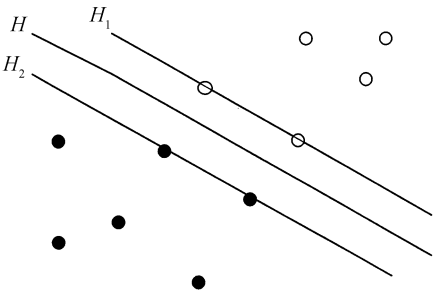

图 1　最优分类面

Fig.1　Optimal separation plane

分的样本集(x_i,y_i),$i=1,\cdots,n$,$x\in R^d$,$y\in\{+1,-1\}$满足

$$y_i[(w\cdot x_i)+b]-1\geqslant 0 \quad (i=1,\cdots,n) \tag{1}$$

此时分类间隔等于$2/\|w\|$,使间隔最大等价于使$\|w\|^2$最小.满足式(1)且使$\frac{1}{2}\|w\|^2$最小的分类面就叫做最优分类面[6],H_1、H_2上的训练样本点就叫做支持向量.上面的方法是在保证训练样本全部被正确分类,即经验风险R_{emp}为0的前提下,通过最大化分类间隔来获得最好的推广性能.如果希望在经验风险和推广性能之间求得某种均衡,可以通过引入正的松弛因子ξ_i来允许错分样本的存在.这时,原问题变为

$$\min_{w,b,\xi}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i \tag{2}$$

$$\text{s.t.}\quad y_i[(w\cdot x_i)+b]-1+\xi_i\geqslant 0 \tag{3}$$

$$\xi_i\geqslant 0 \quad (i=1,\cdots,n) \tag{4}$$

式中:w是超平面的法向量;b是超平面偏值;C为惩罚参数,取值小表示对经验误差的惩罚小,学习机器的复杂度小而经验风险值较大,这样就是"欠学习",而C取大值则为"过学习";ξ_i是松弛因子,表示x_i到超平面$(w\cdot x_i)+b=0$的距离;$(w\cdot x_i)$表示向量w和x_i的内积.利用Lagrange优化方法可以将上述最优分类面问题转化为其对偶问题:

$$\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_iy_j\alpha_i\alpha_j(x_i\cdot x_j)-\sum_{j=1}^{n}\alpha_j \tag{5}$$

$$\text{s.t.}\quad \sum_{i=1}^{n}\alpha_iy_i=0$$

$$(0\leqslant\alpha_i\leqslant C;i=1,2,\cdots,n) \tag{6}$$

式中:α_i为拉格朗日乘子.在最优化求解α_i的过程中,只有部分$\alpha_i\neq 0$对应的训练样本称为支持向量.如果$0<\alpha_i<C$称对应的支持向量为非边界支持向量;$\alpha_i=C$,则称对应的支持向量为边界支持向量,实际上它是分错的训练样本点.

此外,SVM通过一个非线性映射φ将输入空间变换到高维空间来求解最优分类面,在特征空间上所涉及的有关φ的计算及判决函数都以φ的内积形式出现[7,8],并引入核函数$K(x,x_i)=\langle\varphi(x)\varphi(x_i)\rangle$来代替$\varphi$的点积运算.由此得到的分类函数为

$$f(x)=\text{sign}\left(\sum_{i=1}^{n}y_i\alpha_iK(x_i,x_j)+b\right) \tag{7}$$

用N_{BSV+}、N_{BSV-}、N_{SV+}、N_{SV-}分别表示正类的边界支持向量数、负类的边界支持向量数、正类总的支持向量数、负类总的支持向量数,$N+$和$N-$分别表示正、负类别样本数目.从文献[9]知:

$$\frac{N_{BSV+}}{N+}\leqslant\frac{A}{C\cdot N+}\leqslant\frac{N_{SV+}}{N+} \tag{8}$$

$$\frac{N_{BSV-}}{N-}\leqslant\frac{A}{C\cdot N-}\leqslant\frac{N_{SV-}}{N-} \tag{9}$$

从式(8,9)可以看出:如果两类别样本数目不等,即$N+\neq N-$,则对于样本数目大的类别,其错误分类率小;而对于样本数目小的类别,其错误分类率大.也就是说,对正类点集和负类点集应用相同的惩罚参数C,则意味着哪一类样本点的个数多,就更看重哪类点.例如,正样本数目$N+$大时,$\frac{A}{C\cdot N+}$小,也即$\frac{N_{BSV+}}{N+}$小,错误分类率上界小,所以错误分类率小.然而希望对正类点和负类点的惩罚是相同的.

根据以上分析可以看出,C-SVM算法不适于训练集中各类别样本数目相差很大的分类问题.

2　针对不同类样本的不同加权优化算法

针对C-SVM算法的上述缺陷,提出一种加权支持向量机算法,对正类点集和负类点集应用不同的惩罚参数.为此,对适当选定的参数C,令

$$C+=\frac{N-}{N++N-}C,\qquad C-=\frac{N+}{N++N-}C \tag{10}$$

式中:$N+$,$N-$分别是正类训练点和负类训练点的个数;$C+$是对正类点集的惩罚参数,$C-$是对负类点集的惩罚参数.其原始优化问题描述为

$$\min_{w,b,\xi}\ \frac{1}{2}\|w\|^2+\sum_{y_i=1}C+\xi_i+\sum_{y_i=-1}C-\xi_i \tag{11}$$

$$\text{s.t.}\quad y_i[(w\cdot x_i)+b]-1+\xi_i\geqslant 0 \tag{12}$$

$$\xi_i\geqslant 0 \quad (i=1,\cdots,n) \tag{13}$$

对偶问题为

$$\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_iy_j\alpha_i\alpha_jK(x_i\cdot x_j)-\sum_{j=1}^{n}\alpha_j \tag{14}$$

$$\text{s.t.}\quad \sum_{i=1}^{n}y_i\alpha_i=0 \tag{15}$$

$$0 \leqslant \alpha_i \leqslant C_+, y_i = +1 \tag{16}$$

$$0 \leqslant \alpha_i \leqslant C_-, y_i = -1 \tag{17}$$

每次对训练集进行训练时，都采用这种方法选取惩罚参数 C_+ 和 C_-，决策函数不变.

使用与上面C-SVM 同样的分析方法，可得

$$\frac{N_{BSV_+}}{N_+} \leqslant \frac{A}{C \cdot w_+ \cdot N_+} \leqslant \frac{N_{SV_+}}{N_+} \tag{18}$$

$$\frac{N_{BSV_-}}{N_-} \leqslant \frac{A}{C \cdot w_- \cdot N_-} \leqslant \frac{N_{SV_-}}{N_-} \tag{19}$$

式中：w_+，w_- 分别是对正类训练样本点和负类样本点所采取的权重.

将 $w_+ = \frac{N_-}{N_+ + N_-}$，$w_- = \frac{N_+}{N_+ + N_-}$ 分别带入式(18,19)，得

$$\frac{N_{BSV_+}}{N_+} \leqslant \frac{A}{C \cdot \frac{N_-}{N_+ + N_-} \cdot N_+} \leqslant \frac{N_{SV_+}}{N_+} \tag{20}$$

$$\frac{N_{BSV_-}}{N_-} \leqslant \frac{A}{C \cdot \frac{N_+}{N_+ + N_-} \cdot N_-} \leqslant \frac{N_{SV_-}}{N_-} \tag{21}$$

因为

$$\frac{A}{C \cdot \frac{N_-}{N_+ + N_-} \cdot N_+} = \frac{A}{C \cdot \frac{N_+}{N_+ + N_-} \cdot N_-} \tag{22}$$

即两类的错误分类率上界相同，所以正负两类之间得到了比较平衡的误差率，这就意味着提高了对样本数目较少类的重视程度. 然而这是以降低大类别的分类精度为代价来提高小样本的分类精度.

3 数值实验分析

本文用台湾大学林智仁(Lin Chih Jen)副教授等开发设计的 Libsvm 作为测试工具，选用两组数据，一组是Libsvm 自带的数据，该组数据共有 150 个负类样本，120 个正类样本，每个样本有 13 个特征. 由于要求不同类样本数目差别很大，因此从中选用 150 个负类，并从 120 个正类样本随机抽取 30 个样本作为测试数据. 另一组是IRIS 数据，该组数据共有 150 个样本，分 3 类，每种有 50 个样本，每个样本有 4 个特征，并且第一种与第二、第三种是线性可分的，选第三类中的 50 个样本作为正类，第二类中分别随机抽取 10、20 个作为负类. 因为不同类样本数目相差较大，样本数目又较少，所以训练集与测试集选用同一个数据集. 试验结果如表 1 所示，其中对于未说明的参数均取默认值，+50/−10 表示 50 个正类样本，10 个负类样本.

表 1 数值实验结果

Tab.1 Result of data experiment

训练测试数据集	样本数	核函数	参数 C	测试精度/%	
				不加权	加权
Libsvm 自带数据	+30/−150	径向基函数	0.1	83.3	92.7
			1.0	92.2	100.0
			10.0	97.2	100.0
		多项式核函数	0.1	83.3	91.6
			1.0	86.1	98.3
			10.0	96.1	100.0
Iris 数据	+50/−10	径向基函数	0.01	83.3	95.0
	+50/−20		0.10	77.1	97.1

从表 1 结果可以看出，本文提到的优化策略对于分类精度提高是很明显的，因此这对于提高不同类别样本数相差很大情况下的训练性能是一种可取的优化策略.

4 结论

本文研究了样本数目不均衡情况下C-SVM 的一种分类优化策略，提出了在训练过程中对样本数目少的类别加大权值，对样本数目大的类别降低权值的调节方法. 实验证明，这种方法对于提高 C-SVM 在样本数目不均衡情况下的分类精度具有非常明显的效果，对于故障检测、入侵检测等实际问题具有重要的指导意义.

参考文献：

[1] 边肇祺，张学工. 模式识别 [M]. 第 2 版. 北京：清华大学出版社，1999.

[2] 张学工. 关于统计学习理论和支持向量机 [J]. 自动化学报，2000,26(1):32-41.

[3] BURGES C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998,2(2):121-167.

[4] CORTES C, VAPNIK V. Support vector networks [J]. Machine Learning, 1995,20:273-293.

[5] VAPNIK V. 统计学习理论的本质 [M]. 北京：清华大学出版社，2000.

[6] 邓乃杨，田英杰. 数据挖掘中的新方法——支撑向量机 [M]. 北京：科学出版社，2004.

[7] 齐志泉，田英杰，徐志洁. 支持向量机中的核参数选择问题 [J]. 控制工程，2002,12(4):379-381.

[8] 马永军，李孝忠，王希雷. 基于模糊支持向量机和核方法的目标检测方法研究 [J]. 天津科技大学学报，2005,20(3):29-32.

[9] 刘 爽，贾传荧，陈 鹏. 一种自动选择参数的加权支持向量机算法 [J]. 计算机工程与应用，2006,42(2):64-66.