

基于样本分布不平衡的近似支持向量机^{*}

陶晓燕^{1,2} 姬红兵¹ 马志强²

(西安电子科技大学工程学院 西安 710071)¹ (空军工程大学电讯工程学院 西安 710077)²

摘要 针对标准的近似支持向量机(PSVM)没有考虑样本分布不平衡的问题,提出了一种新的 PSVM 算法—BPSVM;根据训练样本数量的不平衡对正负样本集分别分配了不同的惩罚因子,并将原始优化问题中的惩罚因子由数值变更为一个对角阵,最后推导出了线性和非线性 BPSVM 的决策函数。实验结果表明,BPSVM 的性能优于 PSVM,与 SVM 方法相比效率更高。

关键词 近似支持向量机,非平衡分布,平衡近似支持向量机

Proximal Support Vector Machines for Samples with Unbalanced Distribution

TAO Xiao-Yan^{1,2} JI Hong-Bing¹ MA Zhi-Qiang²

(School of Electronic Engineering, Xidian University, Xi'an 710071)¹

(The Telecommunication Engineering Institute, Air Force Engineering University, Xi'an 710077)²

Abstract For the problem of unbalanced data classification is not discussed in the standard Proximal Support Vector Machines, a new PSVM algorithm is presented, namely BPSVM. The different penalty factors are assigned to the positive and negative training sets according to the unbalanced population, and the penalty values are transformed into a diagonal matrix. Finally the decision functions for the linear and nonlinear PSVM are obtained. The experimental results show that the generalization of BPSVM could be better than PSVM, and comparable to SVM with higher efficiency.

Keywords Proximal SVM (PSVM), Unbalanced distribution, Balanced PSVM (BPSVM)

1 引言

支持向量机^[1] (Support Vector Machines, SVM)是基于统计学习理论中的结构风险最小原则提出来的,它利用再生核以及内积定理,有效地避免了经典学习中的过学习、维数灾难、局部极小等非常棘手的问题,在小样本条件下,仍然具有良好的泛化能力。SVM 已成为模式识别、函数回归等领域一个最有力的处理方法^[2,3]。

然而,标准 SVM 需要求解复杂的二次规划问题,训练时间较长。近年来,人们针对 SVM 提出了许多改进方法^[4]。Platt^[5]提出了 SMO (Sequential Minimal Optimization) 算法,其基本思想是循环迭代,将原来问题分解为一系列的子问题,按照某种策略,通过反复求解子问题,最终使结果收敛到原问题的最优解。SMO 解决了占用内存的缺点,但这种方法是一种盲目的“爬山法”,需要有好的启发式策略。为了克服这个难题,很多学者提出了多变量迭代更新方法,它不像 SMO 一次迭代只更新工作集的系数,而是更新所有训练集的系数,这

种方法原理简单,容易实现,但更新率的值难以确定。最近,为了加速 SVM 的学习速度,一些学者提出了简化标准 SVM 优化问题的思路,出现了很多变异 SVM 算法。如 Suykens^[6]等人提出的 LS-SVM (Least Square Support Vector Machines), Mangasarian^[7]等人提出的广义支持向量机 (Generalized Support Vector Machines)。这些方法应用都是直接求解非奇异的线性方程组,从而大幅度提高了学习速度。尤其是 Fung 与 Mangasarian^[8,9]提出的近似支持向量机 (Proximal Support Vector Machines),更引起了人们的极大兴趣。但是,上述 PSVM 对正负样本误差的惩罚是一样的,然而实际情况下,训练样本的分布往往是不平衡的,取同样的误差惩罚就会影响分类结果^[10]。针对标准的 SVM 和 ν -SVM 的样本不平衡问题, H. G. Chew^[11,12]等人提出了根据两类样本的数量,分别对其分配不同的惩罚因子,但优化问题的求解没有变。笔者采用文^[11]中的思想定义惩罚因子,针对 PSVM 的特点,在优化问题中引入了一个新的对角阵,进而推导出了适用于非平衡分布数据分类的线性和非线性 PSVM 的决策函

^{*} 基金项目:国防预研基金项目资助(No. 51407030103DZ0117)。陶晓燕 博士研究生,主要研究方向为机器学习,智能信息处理。

提高网络的构造性能。

PONN 网络不同于传统网络的类划分方法,而由高维空间覆盖角度来分析和构造网络,也就克服了传统网络例如 BP 网络,难于训练,中间层神经元难于确定、知识表示缺陷等问题。当然,如何更好地构造 PONN 网络,还有待于利用待识别事物本身特征的空间分布特征作为先验知识,同时加入知识表示的各向异性特征,这一方面的问题值得下一步进行研究。

参考文献

1 曹安照,田丽,陈俊,吕元峰. 神经网络及其研究展望. 自动化与

仪器仪表, 2006, 1:1~3

- Burges C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2 (2):121~167
- Blum A, Rivest R L. Training a 3-node neural network is NP-complete. Neural Networks, 1992, 5:117~127
- 王守觉. 仿生模式识别(拓朴模式识别)——一种模式识别新模型的理论与应用[J]. 电子学报, 2002, 30(10):1417~1420
- Wang Showjue. Priority Ordered Neural Networks with Better Similarity to Human Knowledge Representation [J]. Chinese Journal of Electronics, 1999, 8(1):1~4
- Frey P W, Slate D J. Letter Recognition Using Holland-Style Adaptive Classifiers [J]. Machine Learning, 1991, 6(2):161~182

数,相应的算法称为 BPSVM(Balanced PSVM)。

2 用于非平衡分布样本的 BPSVM

PSVM 是 Mangasarian 等人从最优化理论的角度出发提出的一种变异 SVM,不但提高了标准 SVM 的学习速度,而且还保持了原有较好的泛化能力。如图 1 所示,考虑一个 n -维实空间 R^n 中 m 个点的两类分类问题,用矩阵 $A^{m \times n}$ 表示该训练集,标准 PSVM 等价于求解下面的优化问题:

$$\begin{aligned} \min_{(w, \gamma, y) \in R^{n+1+m}} & \frac{1}{2} \|y\|^2 + \frac{1}{2} (w'w + \gamma^2) \\ \text{s.t. } & D(Aw - e\gamma) + y = e \end{aligned} \quad (1)$$

其中, e 为单位向量, y 代表误差变量, v 是对误差的惩罚因子, D 是 $m \times m$ 的对角阵, 对角线元素取 +1 时, 对应点为正样本集合 A_+ 中的元素, 取 -1 时, 对应负样本集合 A_- 中的元素。这时分类平面不再是边界平面, 而是“近似”平面, 且尽可能地远离异类样本。

2.1 线性 BPSVM

实际中, 经常遇到正负样本分布不平衡的情况, 如果对两

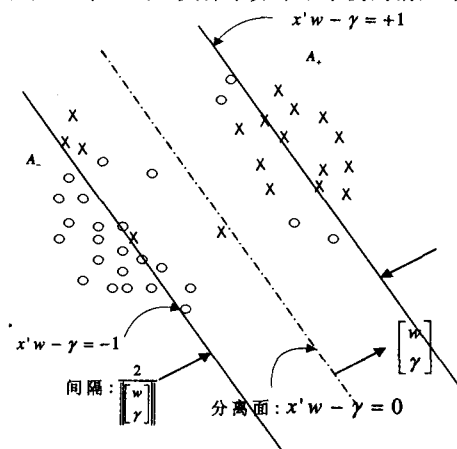


图 1 PSVM 的分类面

类样本的惩罚因子一样的话, 得到的分类面就会偏重于样本数少的一类, 从而影响了正确分类率。为此, 考虑对两类样本分别采用不相等的 v 值, 以产生更准确的分类面。这时式(1)的优化问题变为:

$$\begin{aligned} \min_{(w, \gamma, y) \in R^{n+1+m}} & \frac{1}{2} y'Vy + \frac{1}{2} (w'w + \gamma^2) \\ \text{s.t. } & D(Aw - e\gamma) + y = e \end{aligned} \quad (2)$$

其中, V 是 $m \times m$ 的对角阵, 取值为:

$$\begin{cases} v, & D_{ii} = 1 \\ \frac{\text{num}_+}{\text{num}_-} v, & D_{ii} = -1 \end{cases} \quad (3)$$

num_+ 和 num_- 分别表示正负样本的数目。

利用 KKT 最优充要条件, 式(3)的问题转化为求解下式中拉格朗日梯度为零的方程

$$L(w, \gamma, y, u) = \frac{1}{2} y'Vy + \frac{1}{2} \left\| \begin{bmatrix} w \\ \gamma \end{bmatrix} \right\|^2 - u' (D(Aw - e\gamma) + y - e) \quad (4)$$

其中, $u \in R^m$ 为对应的拉格朗日乘子, 令上式的梯度为零, 得到下面的优化条件:

$$\begin{aligned} w - A'Du &= 0 \\ \gamma + e'Du &= 0 \\ Vy - u &= 0 \\ D(Aw - e\gamma) + y - e &= 0 \end{aligned} \quad (5)$$

化简为:

$$\begin{aligned} w &= A'Du \\ \gamma &= -e'Du \\ y &= V^{-1}u \end{aligned} \quad (6)$$

这里 u 的表达式为:

$$u = (V^{-1} + D(AA' + ee'))^{-1}e = (V^{-1} + HH')^{-1}e \quad (7)$$

其中, H 定义为: $H = D[A - e]$ 。

采用 Sherman-Morrison-Woodbury^[13] 公式简化矩阵求逆, 式(7)可转化为:

$$u = V(I - H(I + H'VH)^{-1}H'V)e \quad (8)$$

最后的分类决策函数可表示为:

$$x'w - \gamma = \begin{cases} > 0 & \text{则 } x \in A_+ \\ < 0 & \text{则 } x \in A_- \\ = 0 & \text{则 } x \in A_+ \text{ 或 } x \in A_- \end{cases} \quad (9)$$

2.2 非线性 BPSVM

对于非线性可分情况, 可以用高斯核 $K(A, A')$ 代替前面的 AA' 推广得到。首先, 用式(6)中的对偶等式 $w = A'Du$ 代替等式约束优化问题中的 w 得到:

$$\begin{aligned} \min_{(w, \gamma, y) \in R^{n+1+m}} & \frac{1}{2} y'Vy + \frac{1}{2} (u'u + \gamma^2) \\ \text{s.t. } & D(AA'Du - e\gamma) + y = e \end{aligned} \quad (10)$$

将核 AA' 用非线性高斯核 $K(A, A')$ 代替, 得到非线性 PSVM 的优化问题:

$$\begin{aligned} \min_{(w, \gamma, y) \in R^{n+1+m}} & \frac{1}{2} y'Vy + \frac{1}{2} (u'u + \gamma^2) \\ \text{s.t. } & D(K(A, A')Du - e\gamma) + y = e \end{aligned} \quad (11)$$

相应的拉格朗日乘子 μ 为:

$$\mu = (V^{-1} + D(KK' + ee'))^{-1}e = (V^{-1} + GG')^{-1}e \quad (12)$$

其中, G 定义为: $G = D[K - e]$ 。

参照文[7]中的推导, 得到非线性 PSVM 的决策函数如下:

$$K(x', A')K(A, A') + e'\mu = \begin{cases} > 0 & \text{则 } x \in A_+ \\ < 0 & \text{则 } x \in A_- \\ = 0 & \text{则 } x \in A_+ \text{ 或 } x \in A_- \end{cases}$$

3 实验结果及分析

分别对人造数据和机器学习领域广泛应用的基准数据进行了实验。为了使训练时间具有可比性, 标准的 SVM 和 PSVM、BPSVM, 均在 Pentium 2.0G, 256M Ram 微机采用 Matlab 6.0 Optimization Toolbox 中的通用程序实现。

3.1 人造数据

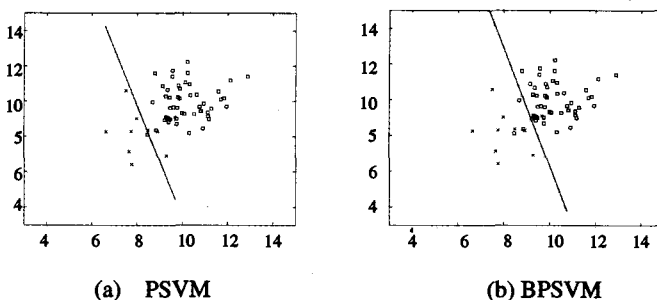


图 2 两类人造数据的分类情况

为了验证新算法的有效性, 我们随机产生了两类高斯分布的样本, 包括 50 个正样本和 10 个负样本, 采用线性分类器。图 2 为实验结果, 可以看出, BPSVM 的分类面较 PSVM

朝数量多的一类移动了,并且提高了分类正确率,PSVM 的为 91.67%,而 BPSVM 的为 93.33%。

3.2 标准数据

通过上面的人造数据实验,可以看出 BPSVM 确实提高了分类器的泛化能力。为了进一步验证新算法的性能,对著名的 UCI^[14] 机器学习数据库中的 6 个基准数据集进行了实验,它们的有关信息如表 1 所示。

表 1 实验数据

数据集	输入维 数	样本 数	类别 数
Breast-cancer-w	9	683	2
Heart	14	294	2
Vote	17	435	2
Iris	4	150	3
Vowel	10	528	11
Wine	13	178	3

表 2 非线性 PSVM 的分类结果

数据集	标准 SVM 正确率(%)	PSVM 正确率(%)	BPSVM 正确率(%)
	时间(s)	时间(s)	时间(s)
Breast-cancer-w	96.56	96.28	97.14
	1687.32	128.5	137
Heart	82	83.01	83.67
	158.39	14.3	13.5
Vote	94.7	94.9	94.49
	405.26	37.5	51.7
Iris	97.73	95.3	98
	56.41	5.14	5.14
Vowel	99.24	98.86	99.25
	3214.18	291.2	381
Wine	97.7	100	98.89
	95.21	8.52	8.94

参数对于支持向量机的分类结果有很大的影响,这里利用网格搜索实现正则参数的选择,其过程为:首先随机选取每个数据集的 50% 作为验证集,然后用验证集中的 90% 训练 PSVM,最后在剩余的 10% 上测试正确率,选定最高识别率所对应的参数为最优的选择。参数的搜索空间确定为: $C, v \in \{10, i = -2, -1, \dots, 5\}$, $\sigma \in \{2^i, i = -4, -3, \dots, 5\}$ 。表 2 列出了非线性 BPSVM、PSVM 和 SVM 三种方法的 10 倍交叉验证(cross-validation)的平均识别率,最高分类结果用黑体标出,其中;多类问题使用了一对多方法。从结果可以看出,尽管 BPSVM 并没有对每一个数据集的正确率都达到最大,但是总体来看,BPSVM 的效果更好一些。同时,相比于标准的 SVM,所需的时间减少了一个数量级。因此,我们可以确切地说所提出的 BPSVM 不但具有很好的泛化能力,同时具有很高的计算效率。

另外,以标准 SVM 的正确率为基准,分别得到了两类和多类数据集的性能提高比例,如表 3 和表 4 所示。这里性能的提高比例定义为:

$$\text{improvement rate} = \frac{\text{correctness}_*}{\text{correctness}_{\text{svm}}} - 1$$

其中 * 代表 PSVM 或者 BPSVM。结果为正表示性能提高了,负值则表示性能降低了。

表 3 两类数据集上的性能提高比例(%)

数据集	PSVM	BPSVM
Breast-cancer-w	-0.29	0.67
Heart	1.23	2.04
Vote	0.21	-0.22
Average	0.38	0.83

表 4 多类数据集上的性能提高比例(%)

数据集	PSVM	BPSVM
Iris	-2.49	0.28
Vowel	-0.38	0.01
Wine	2.35	1.22
Average	-0.17	0.50

对比两个表的结果,可以看出,相比于标准的 PSVM,新算法在多类数据上识别率提高得要多一些(0.50 - (-0.17) > 0.83 - 0.38),这可能是由于采用一对多多分类策略时,每个子分类器的样本不平衡问题更突出的原因。

结束语 基于 PSVM,本文提出了一种用于非平衡分布数据分类的新算法,即 BPSVM。通过人造和 UCI 标准数据的分析表明,该方法的泛化能力比 PSVM 强,与标准 SVM 相比效率更高,而且,新算法在多类数据集上的分类效果提高得更多。但也应注意到,对于 PSVM 分类算法,本文主要考虑了由正负样本数量不同所引起的不平衡问题,而实际上引起不平衡的因素还有很多,这和数据的分布有着很大的关系,因此,选择什么样的算法来描述数据的分布以及相应的解决办法的研究是下一步工作的方向。

参考文献

- 1 Vapnik V N. The nature of statistical learning theory. New York: Springer, 2000
- 2 Cortes C, Vapnik V. Support vector networks. Machine learning, 1995, 20 (1): 273~297
- 3 Drucker H, Burges C J C, Kaufman L. Support vector regression machines. In: Advances in Neural Information Processing Systems, Cambridge: MIT Press, 1997. 155~161
- 4 许建华,张学工,李衍达. 支持向量机的新发展. 控制与决策, 2004, 19 (5): 481~484
- 5 Platt J. Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods—Support Vector Learning, Cambridge: MIT Press, 1999. 185~208
- 6 Suykens J A K, Vandewalle J. Least squares support vector machines. Neural processing letters, 1999, 9(3): 293~300
- 7 Mangasarian O L. Generalized support vector machines. In: Advances in Large Margin Classifier. Cambridge: MIT Press, 2000. 135~146
- 8 Fung G, Mangasarian O L. Proximal Support Vector Machine Classifiers. Wisconsin: University of Wisconsin, 2001
- 9 Agarwal D K. Shrinkage Estimator generalizations of proximal support vector machines. In: Proc. of the eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada: ACM Press, 2002. 173~182
- 10 Morik K, Brockhausen P, Joachims T. Combining statistical learning with a Knowledge-based approach: a case study in intensive care monitoring. In: Proc. 16th Int. con. on machine learning. San Mateo, Canada: Morgan Kaufman Publishers, 1999. 268~277
- 11 Chew H G, Crisp D J, Bogner R E, et al. Target Detection in Radar Imagery Using Support Vector Machines with Training Size Biasing. In: International Conference on Control, Automation, Robotics and Vision, Singapore, 2000
- 12 Chew H G, Bogner R E, Lim C C. Dual v-Support Vector Machines with Error Rate and Training Size Biasing. In: International Conference on Acoustics, Speech and Signal Processing, Singapore, 2001
- 13 Golub G H, VanLoan C C. Matrix Computations. Baltimore, Maryland: The John Hopkins University Press, 1996
- 14 Murphy M. UCI-benchmark repository of artificial and real data sets. University of California Irvine. [http://www.ics.uci.edu/~mllearn\(1995\)](http://www.ics.uci.edu/~mllearn(1995))