doi: 10.3969/j.issn.1006-1576.2011.04.028

代价敏感支持向量机的投影次梯度求解方法

梁万路
（解放军炮兵学院 5 系 43 队，合肥 230031）

摘要：针对传统的分类算法以及精度作为评价指标不能够满足现实分类问题的需要，将代价敏感方法引入支持向量机中，提出一种新的学习算法 CSSVM，并得到了类似于 Pegasos 的投影次梯度求解方法，用于大规模数据的处理。Pegasos 的步骤包括初始化、迭代、确定梯度下降的步长、确定梯度下降方向、更新、投影和结束。实验结果表明，该算法能有效提高识别率和识别精度，具有一定的竞争力。

关键词：不均衡数据；代价敏感；支持向量机；大规模数据
中图分类号：TP311.13　　**文献标志码**：A

Projection Sub-Gradient Solving Method for Cost-Sensitive SVM

Liang Wanlu
(No. 43 Team, No. 5 Department, Artillery Academy of PLA, Hefei 230031, China)

Abstract: Aiming at the traditional method and its precision which used as evaluation index can not meet the requirements of practical classification. Introduce cost sensitive method into SVM, put forward a new learning algorithm CSSVM (cost-sensitive SVM), and acquire projection sub-gradient solving method which is similar as Pegasos to deal with large scale data. The Pegasos process includes initialization, iteration, ascertaining step lengths and direction of sub-gradient descent, update, projection and the end. The test results show that this algorithm can effectively improve identifying rate and identifying precision and it is competitive.

Keywords: class-imbalance data; cost-sensitive; SVM; large scale data

0 引言

分类问题是机器学习和数据挖掘等领域的重要研究内容。现有一些成熟的分类方法，如支持向量机（SVM）[1-2]是由 Vapnik 所领导的贝尔实验室在 1963 年提出的一种非常有潜力的分类技术，主要应用于模式识别领域[3]，在均衡数据的分类中取得了良好的分类效果。然而许多实际的应用领域中都存在不均衡的数据集[4]，例如故障检测、医疗诊断、信息检索、文本分类等。在这种情况下，少数类样本的识别率往往更重要。传统的分类算法以及精度作为评价指标不能够满足现实分类问题的需要，因此针对不均衡数据分类的代价敏感方法的研究越来越引起人们的关注。目前，已有的代价敏感分类算法设计主要是通过直接对损失函数进行加权[5-8]，进而设计出代价敏感损失函数及其相应算法。笔者通过在 SVM 的设计中集成样本的不同误差分类代价，对 Hinge 损失函数进行加权，提出一种新的代价敏感支持向量机，得到了类似于 pegasos[9]的代价敏感支持向量机的投影次梯度求解方法，并通过在大规模数据库上的实验验证了算法的有效性。

1 代价敏感支持向量机

假设训练样本集 $S=\{(x_i,y_i)\}_{i=1}^m, x_i\in R^n, y_i\in\{+1,-1\}$，优化问题可表示为正则化项加 Hinge 损失的形式，如式 (1)：

$$\min_w \frac{\lambda}{2}\|w\|^2+\frac{1}{m}\sum_{(x,y)\in S}\max\{0,1-y(\langle w,x\rangle+b)\} \tag{1}$$

其中正则化参数 λ 为一个预先确定的正实数，在模型复杂度和经验风险之间的寻求最佳折衷，以期获得更好的泛化能力。b 为偏置项，在不均衡数据分类中扮演着重要的角色[9]。

SVM 在不均衡数据分类中存在不足，代价敏感方法是解决不均衡数据分类的重要手段。目前，构造代价敏感分类算法最常用的方法就是直接对损失函数加权，笔者将这种方法引入支持向量机中，可得代价敏感支持向量机（CSSVM）：

$$\min_w \frac{\lambda}{2}\|w\|^2+\frac{1}{m}c(y_i)\sum_{(x,y)\in S}\max\{0,1-y(\langle w,x\rangle+b)\} \tag{2}$$

收稿日期：2010-11-29；修回日期：2011-12-17
基金项目：国家自然科学基金项目“统计学习理论与算法研究”（60575001）和“基于损失函数的统计机器学习算法及其应用研究”（60975040）
作者简介：梁万路（1987—），男，安徽人，硕士，从事模式识别与人工智能研究。

其中，$c(y_i)$代表标签为y_i类样本的误分代价。为简化符号，在二分类问题中记正样本的误分代价$c(y=1)=c_1$，负样本的误分代价$c(y=-1)=c_2$。

2 代价敏感支持向量机的投影次梯度求解

以色列学者 Shai 在文献[9]中提出求解支持向量机的原问题的投影次梯度方法，称为 pegasos。这种算法交替地进行随机梯度下降和投影。算法为达到精度ε所必须的迭代次数为$\tilde{O}(1/\varepsilon)$，而传统的支持向量机优化算法的迭代次数为$\Omega(1/\varepsilon^2)$。算法的运行时间为$\tilde{O}(d/\lambda\varepsilon)$，其中，$d$ 是样本的维数，λ是正则化参数。既然算法的运行时间不直接依赖于训练样本集的大小，所以算法适合解大规模数据集的优化问题。

目前，已有的 pegasos 算法解决的是不含偏置项 b 的支持向量机形式：

$$\min_w \frac{\lambda}{2}\|w\|^2+\frac{1}{m}\sum_{(x,y)\in S}\max\{0,1-y\langle w,x\rangle\} \tag{3}$$

Pegasos 算法步骤为：

第 1 步：初始化。选取任一向量w_1，使其范数不大于$1/\sqrt{\lambda}$，即$\|w_1\|\le 1/\sqrt{\lambda}$。

第 2 步：在算法的第$t(t=1,2\cdots T)$次迭代，T是算法要执行的迭代次数。首先选择样本数为k的子集$A_t\subseteq S$，并用如下所示近似目标函数代替最初的目标函数。

$$\min_w \frac{\lambda}{2}\|w\|^2+\frac{1}{k}\sum_{(x,y)\in A_t}\max\{0,1-y\langle w,x\rangle\}$$

第 3 步：确定梯度下降的步长（学习率）$\eta_t=\frac{1}{\lambda t}$。

第 4 步：确定梯度下降方向。首先确定A_t的一个子集A_t^+，所包含的样本为使用w_t判断当前损失为非零的，即$A_t^+=\{(x_i,y_i)\in A_t: y_i\langle w_t,x_i\rangle<1\}$。对近似目标函数在$w_t$处求偏导为：$\nabla_t=\lambda w_t-\frac{1}{|A_t|}\sum_{(x_i,y_i)\in A_t^+}y_i x_i$。

第 5 步：更新。

$$w_{t+\frac{1}{2}}=w_w-\eta_t\nabla_t=(1-\eta_t\lambda)w_t+\frac{\eta_t}{k}\sum_{(x_i,y_i)\in A_t^+}y_i x_i$$

第 6 步：投影。$w_{t+1}=\min\{1,\frac{1/\sqrt{\lambda}}{\|w_{t+\frac{1}{2}}\|}\}w_{t+\frac{1}{2}}$。

第 7 步：结束。输出w_{t+1}。

如果在每轮迭代中选用的样本个数都为 m 个，即$A_t=S$，算法就成为次梯度投影算法，如果任意选择一个随机样本迭代，算法成为随机次梯度方法。

由于代价敏感支持向量机的优化形式相对于标准支持向量机不同，所以在A_t^+的范围，近似目标函数处的次梯度以及投影区域等方面发生了变化，同时，在代价敏感分类中偏置项 b 的作用非常关键，而现有的方法解决的是不含偏置项 b 的支持向量机形式，这里也给出处理方法。下面将算法的改进之处进行详细说明如下：

1) A_t^+的范围变换为：

$$A_t^+=\{(x_i,y_i)\in A_t: y_i(\langle w_t\cdot x_i\rangle+b)<1\}$$

2) 近似目标函数发生变换为：

$$\min_w \frac{\lambda}{2}\|w\|^2+\frac{1}{k}\sum_{(x_i,y_i)\in A_t^+}c(y_i)\max(0,1-y_i(\langle w,x_i\rangle+b))$$

3) 近似目标函数在w_t处的次梯度为$\nabla_t=\lambda w_t-\frac{1}{|A_t^+|}\sum_{(x_i,y_i)\in A_t^+}c(y_i)y_i x_i$。

4) 加权损失的投影区域为$w\le\sqrt{\frac{c(y_i)}{\lambda}}$。由强凸对偶定理[10]分析可得：

$$\lambda\|w^*\|^2=\sum_{i=1}^m\alpha_i-\frac{1}{m}c(y_i)\sum_{i=1}^m(1-y_i\langle w\cdot x_i\rangle),\alpha_i\in[0,\frac{1}{m}*c(y_i)]$$

$$\lambda\|w^*\|^2\le\sum_i^m\alpha_i\le m*\frac{1}{m}*c(y_i)=c(y_i)$$

$$w\le\sqrt{\frac{c(y_i)}{\lambda}}$$

5) 偏置项 b 的处理：

求解在 b 处的次梯度为：

$\nabla_b=-c(y_i)\frac{1}{|A_t|}\sum_{(x,y)\in A_t^+}y$，初始化$b_0=0$，更新$b_t=b_t+\frac{\eta_t}{|A_t|}c(x_i,y_i)\sum_{(x_i,y_i)\in A_t^+}y_i$，偏置项 b 不需要投影。

3 实验结果及分析

为了验证求解代价敏感支持向量机原问题的

Pegasos 算法，参数 λ 的取值在 $\{\lambda \mid 2\times10^{-8} \le \lambda \le 2\times10^{8}\}$ 之间，对大规模数据库的分类效果进行交叉验证。实验用大规模数据库的详细信息如表 1。

表 1　大规模数据库

数据集	正样本个数	负样本个数	维数
astro-ph_29882.dat	7 010	22 872	99 757
astro-ph_test.dat	7 580	24 907	99 757
CCAT_23149.dat	10 786	12 363	47 236
CCAT_test.dat	370 541	410 724	47 236

固定正样本误分代价 c_1 为 1.0，负样本误分代价 c_2 从 0.1 开始，以步长 0.1 递增到 1.0，总体上评价分类性能。当负样本的误分代价递增到 1.0 时，成为标准的支持向量机。本部分实验以正样本的识别率（召回率）：$Recall = TP/(TP+FN)$ 和 $F-measure = 1/(Recall+Precise)$ 为评价指标[11]，其中，$Precise = TP/(TP+FP)$，为正样本的预测精度。F-measure 更接近两者中较小的一个，为两者的调和平均值，所以 F-measure 越大，正样本的识别率和识别精度越高，算法的分类性能越好。实验结果如表 2。

表 2　分类性能

c_2/c_1	Astro-ph		CCAT	
	Recall	F-measure	Recall	F-measure
0.1	0.965 8	0.881 0	0.967 6	0.906 5
0.2	0.939 2	0.920 5	0.951 8	0.918 4
0.3	0.923 4	0.927 0	0.941 2	0.921 6
0.4	0.912 4	0.927 6	0.933 8	0.922 6
0.5	0.904 2	0.926 3	0.923 1	0.922 2
0.6	0.897 0	0.925 7	0.923 1	0.922 1
0.7	0.892 7	0.925 4	0.923 1	0.921 5
0.8	0.887 0	0.924 2	0.916 2	0.920 7
0.9	0.882 0	0.923 1	0.924 0	0.920 0
1.0	0.878 3	0.922 2	0.911 9	0.919 4
win	7	8	9	7

表 2 中，win 表示代价比从 0.1 变化到 0.9 时 Recall 和 F-measure 值大于代价比为 1.0 的次数，即分类效果好于标准支持向量机的次数。

4　结论

实验结果表明，相对于标准的支持向量机，代价敏感支持向量机能提高对少数类样本的识别率和识别精度，具有一定的竞争力。同时，改进后的 pegasos 算法可以很好地用于求解代价敏感支持向量机，在大规模数据的处理中表现出优越的性能。

参考文献：

[1] 张学工. 关于统计学理论与支持向量机[J]. 自动化学报, 2000, 26(1): 23.

[2] Cortes C, Vapnik. V. Support vector networks. Machine Learning, 1995(20): 273-297.

[3] Duda R O,Stork D G,Hart P E. Pattern Classification (2nd), New York: Wiley, 2001.

[4] 杨明, 尹军梅, 吉根林. 不平衡数据分类方法综述[J]. 南京师范大学学报, 2008(2): 35.

[5] Y. Sun, M. S. Kamela, A. K. C. Wong and Wang, Y.Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 2007, 40(12): 3358-3378.

[6] Masnadi-Shirazi, H. and Vasconcelos, N. Asymmetric boosting. Proc. of ICML, 2007.

[7] W. Fan, S. Stolfo, J. Zhang and Chan, P. Adacost:Misclassification cost-sensitive boosting. Proc.of ICML, 1999.

[8] F. R. Bach, D. Heckerman and Horvitz, E. Considering cost asymmetry in learning classifiers. Journal of Machine Learning Research.

[9] Shai Shalev-Shwartz. Yoram Singer. Nathan Srebro, Pegasos: Primal Estimated sub-GrAdient SOlver for SVM.

[10] Hazan, E., Kalai, A., Kale, S., &Agarwal, A. (2006). Logarithmic regret algorithms for online convex optimization. COLT.

[11] D. Lewis, W. Gale. Training text classifiers by uncertainty sampling. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information, New York, 1998: 73-79.

****

（上接第 76 页）

4　结束语

在液压挺柱沉降测试台中采用上下位机通讯，在上位机中采用时序还原和进程优先级设置，保证了测量的实时性，满足了挺柱的使用要求，对间隙等间接测量参数具有指导意义。

参考文献：

[1] 唐大学, 李志明, 孔七一, 等. 发动机的液压挺柱配气机构[J]. 内燃机, 2007, 12(6): 10-14.

[2] 张明. 液压挺柱沉降测试仪的研制[D]. 四川: 四川大学机械工程系, 2006.

[3] 王伟, 徐国华. 多媒体定时器在工业控制中的应用[J]. 微型机与应用, 2001(12): 8210.

[4] 应站煌, 胡建斌, 赵瑞东, 等. 可编程控制器数据采样特殊性问题讨论[J]. 工业控制计算机, 2010, 23(5): 42-44.

[5] 张超, 郑勇, 李健伟, 等. 提取计算机内部高精度时间用于同步测量[J]. 测绘学院学报, 2003, 20(2): 96-99.