一种改善非平衡分布数据 SVM分类能力的新策略①

岑 涌 罗林开
(厦门大学信息科学与技术学院模式识别与智能系统研究所 厦门 361005)

摘 要 支持向量机利用接近边界的少数向量来构造一个最优分类面。但是若两分类问题中的样本呈现非平衡分布时,即两类样本数目相差很大时,分类能力就会有所下降。提出分别使用重复数量少的一类样本、选择数量多的类样本以及引入类惩罚因子的三个方法来改善分类能力。实验表明,三种方法对不同类型数据集合,一定程度上都改善了支持向量的分类能力。

关键词 支持向量机 非平衡分布 惩罚因子
中图分类号 TP274

A Novel Strategy for Improving the Performance of SVM Classification for Unbalance Distribution Data

Cen Yong Luo Linkai
(Institute of Pattern Recognition and Intelligent System, School of Information Science and Technology, Xiamen University, Xiamen 361005)

Abstract Support vector machine constructs an optimal hyper-plane utilizing a small set of vectors near boundary. However, when the two-class problem samples are imbalanced distribution, SVM has a poor performance. This article presents repeat training minority class samples, selects training majority class samples and introduces punish parameter three methods. Computational results indicate that it improves the capability of SVM classification for the unbalanced samples of different styles datasets.

Key words SVM, unbalance distribution, introduce punish parameter
Class number TP274

1 引言

支持向量机(Support Vector Machine, SVM)是在统计学习理论的基础上发展起来的一种新的机器学习方法,它基于结构风险最小化原则(Structure Risk Minimum, SRM),能有效地解决学习问题,具有良好的推广性和较好的分类精确性。在许多数据挖掘领域应用 SVM取得了不错的效果,然而在一些二分类问题的实际应用中发现在样本量有限且两类样本数量不均衡的情况下,算法的分类能力就会有所降低。为了解决此问题,文献[1]提出的 DFP-SVM算法是通过将约束性数学规划问题转换为无约束性规划问题来改善两类样本数量十分不均衡问题时的分类能力;文献[2]提出了一种通过改善该矩阵来调整类边界的方法来处理数据非均衡分布情况下的支持向量机的分类能力。本文则尝试通过改变训练样本集合中两类样本间的构成来改善 SVM对非平衡分布数据的分类能力。

2 支持向量机概述

支持向量机产生的二元分类器,称为最佳分类超平面,通过非线性映射将输入向量映射到高维特征空间中。SVM使用基于支持向量非线性可分类边界构造线性模型来评估决策函数。如果数据是线性可分的,SVM训练线性机最优超平面来无差错的分离数据并且使得超平面和最近的数据点之间的距离达到最大。那些决定最优超平面最近的训练点被称为支持向量,所有其它的训练样本都与决定的两类边界无关。在大多数数据非线性可分的时候,SVM使用非线性机找到超平面和最小化训练误差[3]。

定义带有标签的训练样本[x_i, y_i],输入向量

① 收到本文时间:2006年 1月 5日

$x_i \in R^n$, 类的值 $y_i \in \{-1, 1\}$, $i=1, 2, \ldots, l$. 考虑线性可分的情形, 决策规则的最优超平面分离二元决策类, 可以由一些支持向量给出:

$$Y = \mathrm{sign}\left(\sum_{i=1}^{N} y_{\alpha_i}(x \cdot x_i) + b\right) \quad (1)$$

其中 Y是结果, y_i是训练样本 x_i得到的类值, $(\cdot)$表示为内积。向量 $x=(x_1, \ldots, x_n)$为输入, 向量 x_i, $i=1, \ldots, l$为支持向量。在式(1)中, b和 α_i为决定超平面的参数。

考虑线性不可分的情形, 在高维情形下的式(1)的表达式为:

$$Y = \mathrm{sign}\left(\sum_{i=1}^{N} y_{\alpha_i} K(x, x_i) + b\right) \quad (2)$$

函数 $K(x, x_i)$定义各种不同非线性决策平面输入空间生成内积的核函数。

常用的核函数有以下几类:

多项式核函数: $K(x, x_i) = (x \cdot x_i + 1)^d$, 其中 d为多项式的度。径向基核函数: $K(x, x_i) = \exp(-1/\delta^2 (x - x_i)^2)$, 其中 δ为径向基函数核的宽度。两层神经网络的核函数: $K(x, x_i) = S[(x \cdot x_i)] = 1/[1 + \exp\{v(x \cdot x_i) - c\}]$, 其中 v和 c为Sigmoid函数 $S[(x \cdot x_i)]$满足不等式 $c \geq v$的参数。

3 处理策略

由于两类样本在分布上的非平衡性, 使得在支持向量机的训练过程中, 两类样本对决策超平面的构成的过程中影响力度会有所不同, 导致最终的决策函数, 即分类超平面将有偏差, 使得分类能力有所降低。因此, 本文考虑能否通过改变样本在支持向量机训练中的构成来改变非平衡分布带来的不良影响。通常情况下我们将两类样本构成比例为7:3及其以上时, 认定为非平衡分布。首先, 对于在样本集合中数量上占大多数的一类的样本我们称其为强势类样本, 相反, 对于在数量上占少数的则称为弱势类样本。

方案Ⅰ:

对于训练集中的弱势类样本, 在进行训练的时候进行多次选取训练, 即重复训练其中的一部分或者全部的弱势类样本, 直到数量上与强势类样本类达到一定的平衡分布。

方案Ⅱ:

对于训练集中的强势类样本, 在进行训练的时候相对的采取只选取其中和弱类样本数相当的一部分进行训练的方法。

方案Ⅲ:

很多数据集中两类样本数目呈现非平衡分布, 但是一般的支持向量机中只有一个惩罚参数 C, 而且对两类样本的惩罚力度是相同的, 这样对样本的非平衡分布问题无能为力。所以考虑引入两个惩罚权重因子, 分别对两类样本错划程度进行分开惩罚, 进而来达到控制样本非平衡分布的影响。

$$\min_{\omega, b, \xi} \frac{1}{2}\|\omega\|^2 + C\left(\mu_+ \sum_{i_+=1}^{l} \xi_{i_+} + \mu_- \sum_{i_-=1}^{l} \xi_{i_-}\right) \quad (3)$$

$$s.t \quad y_i((\omega \cdot x_i) + b) + \xi_i \geq 1, \quad i=1, \ldots, l \quad (4)$$

$$\mu_+ + \mu_- = 1 \quad (5)$$

$$0 \leq \mu_-, \mu_+ \leq 1$$

公式(3)中的 C和 ξ_i为惩罚参数和松弛变量。μ_+为1类样本惩罚权重因子, μ_-为2类惩罚权重因子, 满足(5)式中的约束条件。ξ_{i_+}表示1类样本, ξ_{i_-}表示2类样本。由式(3)可知, 对某类样本的误判惩罚力度最终由惩罚参数 C和惩罚权重 μ_+、μ_-的乘积共同决定, 使得原先的算法能够通过调整两个惩罚权重因子来改善在非平衡分布数据集上的分类能力。

4 实验与分析

4.1 实验数据集

本文选用了两个不同的数据集, 分别来源于Libsvm网站[4]上关于 International Joint Conference on Neural Networks (IJCNN) 的数据和国际机器学习标准数据库 UCI[5]中的 German Credit(银行信用分析)数据集。

对于数据集 IJCNN是取其中给出的一部分, 样本属性数为22, 其中二元离散属性数为10. 本文从中随机选取了200个正类样本和800个负类样本。对于 German Credit数据集, 是由300个正类样本和700个负类样本组成, 样本属性数为24, 其中二元离散属性数为9. 从组成的两个数据集上的样本可知, 居于强类地位的都是负类样本, 居于弱类地位的则都是正类样本。针对两个数据集都只选取其中每一类50%的样本做为训练集, 从剩余的样本中, 选择相等数目的两类样本做为共同的测试集。经过变化后的各个方案实际的样本组成如表1所示:

表1 各种方案下实际数据样本的组成

数据集		训练样本集		测试样本集	
		正类样本数	负类样本数	正类样本数	负类样本数
IJCNN	C-SVM	100	400	100	100
	方案Ⅰ	400	400		
	方案Ⅱ	100	100		
	方案Ⅲ	100	400		
German Credit	C-SVM	150	350	150	150
	方案Ⅰ	350	350		
	方案Ⅱ	150	150		
	方案Ⅲ	150	350		

4.2 方法应用

本文采用的支持向量机算法工具为 SVM—Light软件[6]，验证方法采用 5折交叉验证方法。在支持向量机中的核函数的选取时，几个方案中都采用径向基核函数。原因是考虑到径向基核非线性的样本映射到高维空间中时，不同于线性核，而是能够处理分类和属性之间非线性的情形。此外，线性核可以认为是径向基核一种特殊情形——线性核伴随惩罚参数能和径向基核在一些参数组合(C，δ)下有同样的表现效果。另外，考虑影响模型选择复杂度的超平面参数个数，多项式核比径向基核多很多超平面参数。Sigmoid核和径向基核在一些特定参数时有同样的效果，更重要的是 Sigmoid核在某些参数情形下是不合理的[7,8]。

在实验过程中，对于径向基核函数参数 δ，另其始终为数据集中样本属性数的倒数：δ=1/K(K为属性数)。同时设置精度要求为默认的 0.001。特别地，对于引入针对类惩罚权重因子的方案，我们设定两个因子的构成原则为：μ_-，μ_+之比为样本数目比值的反比。

4.3 实验结果和分析

Vapnik等人的研究中发现，核函数参数和误差惩罚参数 C对学习机器的性能至关重要。理论上太小的参数 C的值会引起训练数据的欠拟合，因为放置在训练数据上的权重过小而导致测试集上过大的均方误差。相反，如果参数 C的值太大，SVM会引起训练数据的过拟合，意味着$\frac{1}{2}\|\omega\|^2$项将失去意义，目标也将回到只是最小化经验风险而已。

在本文中对每一个数据集根据交叉验证和逐步迭代的方法找到相应的最优的参数 C的值，然后进行相应的分类效果评估，来达到避免由于 C过大或过小所引起的消极影响。

4.4 分类效果评估

预测的准确率(Accuracy)是评估分类方法的主要指标，定义对样本有：

$$\text{准确率}=\frac{\text{正确分类的实例数}}{\text{总实例数}}\times 100\% \quad (6)$$

为了考虑单独类别的准确率，我们使用灵敏度(Sensitivity)，特效性(Specificity)以及评价标准来作为准确率的度量。这些度量定义分别为：FN是指将正类样本错分为负类样本数目，FP指将负类样本错分为正类样本数目。同时标记正类样本正确分类数和负类样本正确分类数分别为 TP和 TN。

由此可得：$\text{Sensitivity}=TP/(TP+FN)$ (7)

$\text{Specificity}=TN/(FP+TN)$ (8)

表 2中给出了两个数据集上在 4种方案下的总体准确率，表 3则分别给出了 4种方案在两个数据集上的灵敏度和特效性的情况。

表 2 几种方案实验总体准确率结果

	准确率(%)	C—SVM	方案Ⅰ	方案Ⅱ	方案Ⅲ
IJCNN	训练集	95.5%	98.5%	97.0%	97.5%
	测试集	90.0%	94.0%	94.0%	93.5%
German Credit	训练集	79.33%	81.33%	84.67%	83.33%
	测试集	73.67%	73.67%	78.33%	76.67%

表 3 IJCNN和 German数据集上的灵敏度和特效性分析

数据集		C—SVM	方案Ⅰ	方案Ⅱ	方案Ⅲ
IJCNN训练集	Sen.	0.85	0.9775	0.96	0.92
	Spec.	0.9825	0.9925	0.98	0.99
测试集	Sen.	0.84	0.93	0.92	0.94
	Spec.	0.96	0.95	0.95	0.93
German训练集	Sen.	0.7	0.796	0.86	0.787
	Spec	0.834	0.843	0.84	0.854
测试集	Sen.	0.707	0.72	0.787	0.753
	Spec.	0.77	0.753	0.78	0.78

4.5 结果分析

对于方案Ⅰ(重复训练弱类样本的方法)：在两个数据集上对处于数量上少数的正类样本点进行了重复训练，在 IJCNN数据集上使得分类能力有了明显的提高，但是在 German Credit数据集上效果不是很明显。

对于方案Ⅱ(选取部分强类样本的方法)：在两个数据集上，对数量上占多数的负类样本点，只是随机选取数量上和弱类相当的一部分进行训练，在 IJCNN数据集上分类能力的提高不是很明显，但是在 German Credit数据集上变化波动很大，在有的交叉验证集上准确度提高很过，但是在有的上面却没有改善多少，甚至是降低很多。

对于方案Ⅲ(引入类惩罚因子方法)：在两个数据集上，分类能力一定程度上都得到了一些提高，但是一些情况下，分类能力的改善没有方案Ⅰ和方案Ⅱ的明显，但是具有比较稳定的表现，也就是有较好的推广能力。

从数据集特征上分析，German Credit数据集由于每一个样本都是有 24个属性变量(3个连续的、21个陈述性，其中有 9个是二元属性)构成，其中包括有：职业情况(employment status)、个人资料(personal information)、年龄(age)、住房供给情况

(下转第 113页)

的总体设计及其标准化的研究,本文参照了一般的教学流程和教育技术标准中的描述,实现则与网络教育平台所使用的编程语言、系统平台有关。笔者对基于 J2EE平台的多层体系结构的系统平台作了较为详细的论述。最后本文具体描述了学习管理原型系统的实现。

参 考 文 献

[1]沈中南,史元春. 现代远程教育技术规范简介[J]. 计算机工程与应用,2003,39(5):66~69

[2]白凡凡. 计算机管理教学(CMI)的设计和发展[J]. 计算机工程与应用,2003,39(5):220~222

[3]武彩霞. 现代网络教育信息技术标准的应用研究[D]. 成都:西南财经大学,2003

[4] ADLnet SCORM(Sharable Content Object Reference Model) version1.2[EB/OL]. http://www.adlnet.org,2001—10

(上接第 105页)

(housing)和工作(job)等一些社会统计学方面的属性,因此数据集具有分类函数学习中混有比较高的相关信噪比。但是 IJCNN数据集则有比较好的数据质量。

采用重复训练弱类样本方法,在相关数据质量较高的数据集上可以获得比其他三种方案更好的效果,使得改善分类能力的目标得以实现,但是在噪声较多的数据集上此方法增加有用信息的同时也加大了干扰信息,获得更多有用的弱类样本信息的同时也增加了干扰信息,使得分类效果没有得到多少改善;而对于选择强类样本进行训练的方法,在选取到有较大信息量和更具有代表性的数据集时,能使得分类能力明显提高,相反地,若是干扰信息较大的则随机性变化的可能性比较大,此时则会严重影响支持向量机的分类能力。引入惩罚因子的方法在两个数据集上一定程度上改善了分类效果,具有比较好的适用性。

5 结论

实验表明,在数据呈现非均衡分布的情况下,通过改变训练样本集合中样本构成,采用引入类惩罚因子的方法都可以一定程度上改善支持向量机的分类能力,但是重复训练部分少数类样本的方法会使样本数量增加,加大了训练量;选择多数类样本的方法则具有一些随机性和变动性,对样本集本身的质量很敏感;引入惩罚因子的方法,具有更好的稳定性,但是有时候分类效果的提升不是很好。进一步的研究将是尝试借助特征提取方法,建立选取更具有代表性的样本,减少因样本非平衡分布造成的影响,改善支持向量机的分类能力。此外就是利用数据中的降噪等方法来提高数据集的相应质量。在较高质量且样本数量不是很多的情况下,作为一种新的策略,重复少数类样本的方法可以很好地改善支持向量机在数据非平衡分布时的分类能力。

参 考 文 献

[1]孙蕾,周明全,李丙春. 一种非平衡分布数据的支持向量机新算法[J]. 计算机应用,2004,24(12):14~15

[2] Gang Wu, Edward Y. Chang. KBA: kernel Boundary Alignment Considering Imbalanced Data Distribution[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 786~795

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

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[5] http://www.ics.uci.edu/~mlearn/MLRepository.html

[6] http://svmlight.joachims.org/

[7] S. S. Keerthi, C. J. Lin. Asymptotic behaviors of support vector machines with Gaussian kernel[J]. Neural Computation, 2001, 15(7): 1667~1689

[8] H. T. Lin, C. J. Lin. 2003. A study on sigmoid kernels for SVM and the training of non—PSD kernels by SMO—type methods. Technical report. Department of Computer Science and Information Engineering, National Taiwan University