



基于边界样本欠取样支持向量机的电信用户欠费分类算法

李创创, 卢光跃, 王航龙

(西安邮电大学无线网络安全技术国家工程实验室, 陕西 西安 710121)

摘要: 电信用户欠费预测是一个不平衡数据集分类问题。针对传统支持向量机 (SVM) 对不平衡数据集中少数类检测精度低的问题, 基于分类平面由边界样本的位置决定, 提出了一种通过删除部分多数类边界样本的方法来改善传统 SVM 算法的不足, 将该算法和其他几种算法在电信数据和多个不平衡 UCI 数据集上的实验结果进行对比, 验证所提算法对少数类的检测精度和总体评价指标都有所提高。

关键词: 欠费; 不平衡; SVM; 边界; 欠取样

中图分类号: TP181

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017208

SVM classifier for telecom user arrears based on boundary samples-based under-sampling approaches

LI Chuangchuang, LU Guangyue, WANG Hanglong

National Engineering Laboratory for Wireless Security, Xi'an University of
Posts and Telecommunications, Xi'an 710121, China

Abstract: Telecom users' arrears forecasting is a classification problem of unbalanced data set. To deal with the problem that the traditional SVM on the unbalanced data set had a low detection accuracy of minority class, a novel method was proposed. Based on the fact that the position of classification plane was determined by the boundary samples, the proposed method was implemented via removing some of samples closed to the classification plane to avoid the deficiency of the traditional SVM algorithm. Finally, the proposed method was compared with other approaches on unbalanced data sets. The simulation results show that the proposed method can not only increase the detection accuracy of minority but also improve the overall classification performance.

Key words: arrear, unbalance, support vector machine, boundary, under-sampling

1 引言

近年来, 随着电信运营商之间的竞争日益激烈, 我国电信运营商在推出大量新业务的同时也降低了用户的开户门槛, 使得拖欠、拒交话费用户的比例不断增加^[1], 对电信企业正常运营造成

很大影响。为了减少坏账, 增加公司利润, 越来越多的运营商通过建立用户欠费预测模型, 提前预测可能出现欠费的用户并对该类用户采取相应的措施。支持向量机 (support vector machine, SVM) 是以统计学为基础的一种经典机器学习方法^[2], 它在解决高维非线性数据集中的分类问题



中表现出了良好的性能^[3], 因此被越来越多的学者所关注, 现已广泛应用于许多领域的分类问题中。

传统 SVM 分类算法在均衡数据集上可以得到良好的分类效果, 当数据集不均衡时, 分类平面会偏向于少数类^[4-8], 导致少数类的漏检率增大。在生活领域中人们需要面对很多非平衡数据集上的分类问题, 例如电信欠费用户预测。由于电信用户中欠费用户数量远远少于正常用户, 并且对少数类的检测精度的要求要高于多数类, 因此必须改进 SVM 来提高对少数类的检测精度。

针对传统 SVM 固有的缺点, 已提出了两类改进算法。一类是基于算法层面的, 通过改进算法模型来使其适应不均衡数据集, 比如模糊支持向量机^[9] (fuzzy support vector machine, FSVM) 就是根据每一个样本离样本中心的距离赋予每一个样本不同的权值来改善 SVM 分类平面偏移问题。另一类为数据层面上的, 其通过对训练集的数据进行处理, 使其多数类样本和少数类样本保持基本均衡, 最具代表性的算法有随机过取样 (over SVM) 算法和随机欠取样 (under SVM) 算法。随机过取样算法采用随机复制训练集中少数类样本, 使新的训练集数据集中两类样本保持平衡^[10]。然而, 增多样本不仅使计算复杂度增高, 并使分类间隔减小, 容易产生过拟合。而随机欠取样算法采用随机去除训练集中的多数类部分样本使多数类和少数类样本数相等^[11]。由于欠取样算法只是选取原来多数类样本的一个随机子集, 并不能够代表原多数类样本的全部信息, 当选择的子集远离边分类边界, 会导致 SVM 分类平面过偏移, 使得多数类检测精度损失过大。

本文提出一种基于边界样本的欠取样 SVM (boundary samples-based under-sampling support vector machine, B-SVM) 算法, 该算法首先对不平衡数据集进行仿真得到初步的分类平面, 计算多数类样本到平面的距离, 根据计算结果删除部分多数类边界样本, 从而使分类平面向多数类方

向偏移。实验中将本文所提算法应用于某省的电信数据和标准的 UCI 数据集^[12]中, 并与已有算法进行了对比, 实验结果表明本文算法对少数类的分类精度有了明显的改善。

2 SVM 理论

支持向量机的基本想法是求解几何间隔最大的分类平面^[13]。以二分类为例, 设给定的训练集样本为代表 $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, $\mathbf{x}_i \in R^l$ 代表 l 维样本, $y_i \in \{+1, -1\}$, $(i=1, 2, \dots, n)$ 代表样本类别。为了能更好地将样本分为两类, 首先将样本通过映射函数 $\phi(\mathbf{x})$ 映射到高维空间, 在高维映射空间中寻找分类平面 $\mathbf{w} \cdot \phi(\mathbf{x}) + b = 0$; 然后通过优化分类平面的权向量 $\mathbf{w}$ 和截距 b , 寻找最大分类间隔平面, 即优化如下目标函数:

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

$$\text{s.t. } y_i(\mathbf{w} \cdot \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad i=1, 2, \dots, n \quad (2)$$

$$\xi_i \geq 0, \quad i=1, 2, \dots, n$$

其中, C 为惩罚因子, 用于平衡结构风险和经验风险, ξ_i 为松弛变量。

上述问题是一个凸二次规划问题, 利用拉格朗日乘子法可以将其转化为如下对偶问题:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) - \sum_{i=1}^n \alpha_i \quad (3)$$

$$\text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0 \quad (4)$$

$$0 \leq \alpha_i \leq C, \quad i=1, 2, \dots, n$$

其中, α_i 是拉格朗日乘子, 必须满足 Karush-Kuhn-Tucker (KKT) 条件:

$$\alpha_i (y_i (\mathbf{w} \cdot \phi(\mathbf{x}_i) + b) - 1 + \xi_i) = 0, \quad i=1, 2, \dots, n \quad (5)$$

$$(C - \alpha_i) \xi_i = 0, \quad i=1, 2, \dots, n \quad (6)$$

求解式 (3) 和式 (4) 得:

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \phi(\mathbf{x}_i) \quad (7)$$

最终得到分类决策函数为:

$$f(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^n \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \right) \quad (8)$$

其中, $K(\mathbf{x}, \mathbf{x}_i)$ 为核函数, $K(\mathbf{x}, \mathbf{x}_i) = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}_i)$ 。

本文使用的是高斯核函数 $K(\mathbf{x}, \mathbf{x}_i) = e^{-\sigma \|\mathbf{x} - \mathbf{x}_i\|^2}$, σ 为径向基核函数自带的参数, 决定了样本映射到特征空间的分布。

图1是测试数据集不平衡时SVM分类性能的示例。选用高斯函数生成了一个多数类与少数类之比为50:1的二维不平衡数据集, 十字型的为多数类样本; 图1中以等高线的形式给出了分类平面, 分类平面是等高线为零的线。从图1中可以看到, 分类超平面向着少数类偏移, 部分少数类样本被错分。可见, 当数据集不平衡时, 多数类样本在投影空间的密度远大于少数类; SVM为了使总体正确率最大化, 分类平面会向少数类方向偏移, 导致少数类的检测精度低。

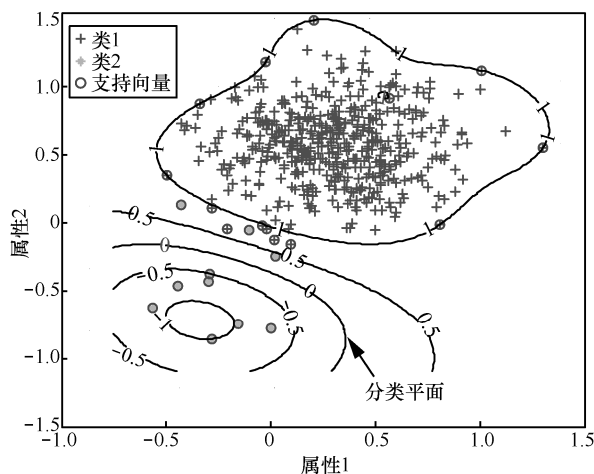


图1 数据样本比例为50:1时的分类平面

3 针对非平衡样本的基于边界样本欠取样 SVM

提高少数类的检测精度, 就必须使分类平面向多数类方向偏移, 传统的欠取样算法没有考虑到样本在投影空间的分布特性, 导致在测试集中分类效果差, 因此必须充分考虑样本的分布特性。

由于在投影空间中边界样本占该类样本的比例都很小, 删除掉部分多数类边界样本使分类平面偏移对多数类分类正确率影响很小, 因此可以删除部分多数类边界样本来使超平面向多数类方向偏移。

本文首先在原始数据集上运用传统 SVM 得到边界样本, 再计算边界样本距分类平面的距离, 根据计算结果, 有选择地删除部分多数类边界样本。边界样本跟拉格朗日乘子 α_i 和核宽度 σ 的值有密切的关系。根据式(3)和式(4)可求解得到 α_i , α_i 的值有3种可能, 即:

- $\alpha_i = 0$, 根据式(6)可得 $\xi_i = 0$, 则样本 $\mathbf{x}_i$ 被正确分类, 并远离分类平面;
- $0 < \alpha_i < C$, 根据式(5)得 $y_i(\mathbf{w} \cdot \phi(\mathbf{x}_i) + b) - 1 - \xi_i = 0$, 根据式(6)得 $\xi_i = 0$, $y_i(\mathbf{w} \cdot \phi(\mathbf{x}_i) + b) = 1$, 则样本 $\mathbf{x}_i$ 是标准的支持向量, 恰好落在最大间隔平面上;
- $\alpha_i = C$, 根据式(5)和式(6)得 $y_i(\mathbf{w} \cdot \phi(\mathbf{x}_i) + b) - 1 - \xi_i = 0$ 和 $\xi_i \geq 0$ 。如果 $0 \leq \xi_i < 1$, 样本 $\mathbf{x}_i$ 被正确分类, 并且处于分类平面和最大间隔平面之间。如果 $\xi_i \geq 1$, 样本 $\mathbf{x}_i$ 被错分。

由上可知, $\alpha_i > 0$ 对应的样本为支持向量, 与分类平面距离较近, 决定了分类平面的位置和分类间隔的大小。

一旦分类平面得到后, 可以计算每一个样本与分类平面的距离。 γ_i 表示样本 $\mathbf{x}_i$ 与分类平面的函数间隔, 也表示该点离分类平面的“距离”:

$$\gamma_i = y_i(\mathbf{w} \cdot \phi(\mathbf{x}_i) + b) = y_i \left(\sum_{j=1}^J \alpha_j^* y_j K(\mathbf{x}_j, \mathbf{x}_i) + b \right) \quad (9)$$

其中, J 为总体的支持向量数, α_j^* ($\alpha_j^* > 0$) 为对应支持向量的拉格朗日乘子。

对于多数类样本而言, 当 $\gamma_i \leq 0$ 时表示样本落入少数类一侧, 该样本被错分, $\gamma_i > 1$ 样本在多数类一侧, 该样本分类正确并远离分类平面, $0 < \gamma_i \leq 1$ 表示样本在分类超平面和最大间隔平面



之间,对应的样本为边界样本,可以根据 γ_i 的值删除掉部分多数类样本。

然而对于某些稀疏数据集(如电信数据集),存在很多属性值为零的情况,在投影空间中边界样本密度较大,在这种情况下删除边界多数类样本会导致多数类样本的检测精度损失过大。径向基核函数的参数 σ 决定着边界样本密度。 σ 越小边界样本越多, σ 越大边界样本越少。当 σ 很小时,由于多数类边界很多,删除全部多数类边界样本会使分类平面向多数类过偏移,导致整体分类性能下降。所以在删除边界样本时,为了不使分类平面过偏移,边界样本的数量不宜大于该类的5%。

基于边界样本的欠取样 SVM (B-SVM) 具体算法步骤如下。

步骤 1 在不均衡数据集上根据式(3)和式(4)解得 $\alpha=(\alpha_1, \alpha_2, \dots, \alpha_n)$, 找出 $0 < \alpha_i$ 的边界样本(支持向量)。

步骤 2 计算边界样本占多数类的样本比例 r , 如果 $r > 0.05$, 则增大 σ 并返回步骤1重新计算 α 。

步骤 3 根据式(9)计算多数类样与分类平面的函数间隔 γ_i 。

步骤 4 删除掉 $\gamma_i < d$ 的多数类样本(d 为人设置的阈值)。

步骤 5 将删除后剩余的多数类样本和少数类样本进行重组构建新的训练集。

步骤 6 根据训练得到新的分类器进行分类。

图2是删除 $\gamma_i < 1$ 的多数类边界样本后分类平面效果,与图1对比,少数类分类全部正确,少数类的分类精度明显提高,这是由于多数类边界样本远离少数类样本,从而使分类平面向多数类方向偏移。

4 实验分析及对比

4.1 不均衡数据分类效果指标

对于电信欠费用户分类问题,关注更多的是

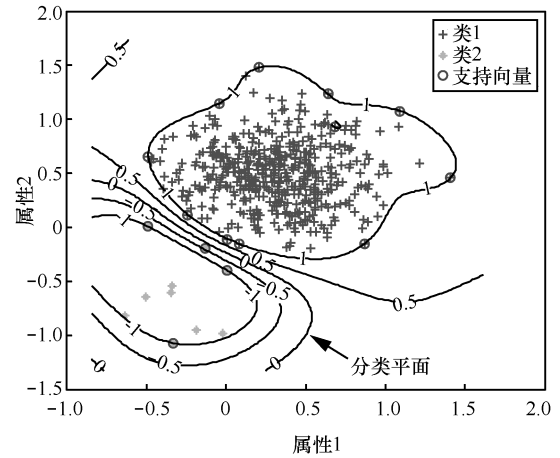


图2 去掉部分多数类边界样本后的分类平面的变化

欠费用户的检测精度,传统 SVM 的评价指标不能很好地反映实际需求。近年来许多学者提出一些用于评价不均衡数据分类效果的指标。最常用的有以下几种(首先定义少数类(正例样本)为 P ,多数类样本(负例样本)为 N)。 FP 是指将多数类错分成少数类的样本总数, FN 是指将少数类错分成多数类总数, TN 和 TP 分别表示正确分类的多数类和少数类样本数。

少数类样本的检测精度为:

$$SE = \frac{TP}{TP + FN} \quad (10)$$

多数类样本的检测精度为:

$$SP = \frac{TN}{TN + FP} \quad (11)$$

总体分类性能评估:

$$GM = \sqrt{SE \times SP} \quad (12)$$

SE 侧重于考量少数类的正确率, SP 侧重于考量多数类的正确率, GM 综合了两个指标体现分类器的整体分类性能。

4.2 实验数据

为了验证本文算法在不均衡数据上的有效性,本文选取了5个不平衡UCI数据集^[13]和某省的电信客户欠费数据^[14],电信数据属性有是否为VIP、付费方式、是否主动离网、是否被动离网和通话时长等92维。其中数值属性可以直接使用,

二值属性、类别属性经过整数编码可以使用。用户属性取值为：公众（0）、集团（1）。付费方式取值为：后付费（0）、预付费（1）。是否离网、是否主动离网、是否被动离网、是否主动停机和是否被动停机取值为：是（1）、否（0）。套餐月费用取值为：46元（0）、66元（1）、96元（2）、126元（3）、156元（4）、186元（5）、226元（6）、286元（7）、386元（8）、586元（9），取欠费金额作为目标属性，整数编码后取值为：未欠费（0）、欠费（1）。处理后的样本部分属性见表1。

表1 电信客户消费部分信息

样本属性	样本1	样本2	样本3	样本4	样本5
是否VIP	0	0	0	0	0
用户属性	0	0	1	0	1
付费方式	0	0	0	0	0
是否离网	0	0	0	0	0
是否主动离网	0	0	0	0	0
是否被动离网	0	0	0	0	0
是否主动停机	0	0	0	0	0
月通话时长/s	0	0	47 431	36 649	68 805
月通话次数/次	0	0	512	51	466
GPRS流量/KB	0	0	43 967	215 791	0
短信发送量/条	0	0	1	238	9
彩信发送量/条	1	2	20	20	23
套餐月费	1	5	2	2	2
通话费用/元	0	0	30.15	2.25	20.85
赠费销账/元	30	0	0	0	0
月初账费/元	42	193	131.85	157.27	150.75
账户余额/元	0	0	169.74	149.33	291.91

6个实验数据集特征信息见表2。

4.3 不同算法的性能比较

为了验证本文所提算法在不均衡数据集上的有效性，应用本文所提算法(B-SVM)和标准SVM算法、模糊SVM算法、随机欠取样SVM算法和随机过取样SVM算法对表1所给的6个不平衡数据集进行了分类实验。在进行数据实验之前，对数据集都进行归一化处理以降低每一个属性值之间的差异。本文使用的仿真工具是LIBSVM和MATLAB。

本文使用五折交叉验证获得最佳的惩罚参数 C 和核宽度 σ 。由于随机欠取样和过取样算法的不稳定性，仿真结果为仿真10次的平均值，见表3。由表3可知，本文所提算法在前4个数据集对少数类的检测精度(SE)和整体分类性能(GM)明显优于其他算法。传统的SVM和FSVM在第5个数据集上的少数类检测精度为0，表明第5个数据集中少数类和多数类在特征空间中分类间隔小，不易分辨。本文通过删除掉与少数类易混淆的多数类使分类平面向多数类偏移，在该数据集上欠采样SVM分类效果更好。

由于电信数据为稀疏数据集，投影空间中的多数类边界样本密度很大，通过调整 σ 的取值范围来使边界样本减少，使算法免于过拟合。本文通过变量 d 来调节删除边界样本的数量。 d 越大，对少数类的检测精度越高，多数类的正确率越低。本文取 $d=1$ ，实际应用中可以根据对少数类的检测精度的要求与对多数类错误的

表2 不平衡数据集描述

数据集	少数类数量	多数类数量	样本总数	不平衡比	类别数	正例类(第几类)
Haberman	81	225	306	26 : 74	2	2
Pima-Indians	268	500	768	35 : 65	2	1
Page-blocks	115	5 358	5 473	2 : 98	5	5
Waveform	1 657	3 343	5 000	35 : 65	2	1
abalone	103	4 074	4 177	2 : 98	29	15
电信数据	250	3 000	3 250	8 : 92	2	1



表 3 仿真结果对比

数据集	评价指标	B-SVM	标准 SVM 算法	模糊 SVM 算法	随机欠取样 SVM 算法	随机过取样 SVM 算法
Haberman	GM	0.68	0.420	0.547	0.592	0.624
	SE	0.55	0.198	0.325	0.625	0.475
	SP	0.84	0.893	0.82	0.56	0.82
Pima-Indians	GM	0.77	0.689	0.703	0.760	0.679
	SE	0.91	0.530	0.54	0.74	0.580
	SP	0.65	0.896	0.915	0.78	0.795
Page-blocks	GM	0.86	0.761	0.702	0.749	0.391
	SE	0.76	0.583	0.5	0.62	0.4
	SP	0.974	0.995	0.985	0.905	0.975
Waveform	GM	0.880	0.877	0.875	0.877	0.792
	SE	0.908	0.827	0.818	0.894	0.654
	SP	0.853	0.928	0.936	0.860	0.96
abalone	GM	0.49	0	0	0.539	0.02
	SE	0.46	0	0	0.58	0.02
	SP	0.523	1	1	0.5	0.997
电信数据	GM	0.931	0.561	0.918	0.871	0.871
	SE	0.92	0.32	0.88	0.88	0.8
	SP	0.942	0.983	0.958	0.862	0.948

容忍率来动态调整。

5 结束语

本文针对传统 SVM 对电信欠费用户分类精度低的问题,提出了一种基于边界样本的欠取样 SVM 算法,该算法通过删除部分多数类边界样本使分类超平面向多数类偏移,提高对欠费用户的检测精度。并解决了当边界样本过多时通过调节 σ 的取值范围来解决过偏移的问题。与其他算法在 5 个不平衡 UCI 数据集和电信数据集上的仿真结果对比,表明所提算法能有效提高欠费用户的检测精度。

参考文献:

- [1] 渠瑜. 基于 SVM 的高不平衡分类技术研究及其在电信业的应用[D]. 杭州: 浙江大学, 2010.
- QU Y. Research on SVM-based highly imbalanced classification and its application in telecommunications[D]. Hangzhou:

Zhejiang University, 2010.

- [2] VAPNIK V N. The nature of statistical learning theory[M]. Berlin: Springer, 2000: 138-167.
- [3] JAMIL S, KHAN A. Churn comprehension analysis for telecommunication industry using ALBA[C]/ICET, IEEE, October 18-19, 2016, Islamabad, Pakistan. New Jersey: IEEE Press, 2016.
- [4] RASKUTTI B. Extreme re-balancing for SVMs: a case study[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1):60-69.
- [5] KANG P, CHO S. EUS SVMs: ensemble of under-sampled SVMs for data imbalance problems[M]. Berlin: Springer, 2006: 837-846.
- [6] 陶新民, 张冬雪, 郝思媛, 等. 基于谱聚类下采样失衡数据下 SVM 故障检测[J]. 振动与冲击, 2013, 32(16): 30-36.
- TAO X M, ZHANG D X, HAO S Y, et al. Fault detection based on spectral clustering combined with under-sampling SVM under unbalanced datasets[J]. Journal of Vibration and Shock, 2013, 32(16): 30-36.
- [7] HE H B, EDUARDO A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(8): 1263-1284.
- [8] LIU X Y, ZHOU Z H. Exploratory under-sampling for class-imbalance learning[J]. IEEE Transactions on Systems,

- Man and Cybernetics, 2009, 39(2): 539-550.
- [9] LIN C F, WANG S D. Fuzzy support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2):464-71.
- [10] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 21(9):1263-1284.
- [11] 陶新民, 郝思媛, 张冬雪, 等. 基于样本特性欠取样的不平衡支持向量机[J]. 控制与决策, 2013(7):978-984.
- TAO X M, HAO S Y, ZHANG D X, et al. Support vector machine for unbalanced data based on sample properties under-sampling approaches[J]. Control and Decision, 2013(7): 978-984.
- [12] BATUWITA R, PALADE V. FSVM-CIL: fuzzy support vector machines for class imbalance learning[J]. IEEE Transactions on Fuzzy Systems, 2010, 18(3):558-571.
- [13] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
- LI H. Statistical learning method[M]. Beijing: Tsinghua University Press, 2012.
- [14] 包志强, 崔妍. 电信客户欠费模型评估[J]. 西安邮电大学学报, 2015(4):97-101.
- BAO Z Q, CUI Y. Telecom customer arrearages model evaluation[J]. Journal of Xi'an University of Posts and Telecommunications, 2015(4):97-101.

[作者简介]



李创创 (1991-), 男, 西安邮电大学无线网络安全技术国家工程实验室硕士生, 主要研究方向为数据挖掘。



卢光跃 (1971-), 男, 西安邮电大学无线网络安全技术国家工程实验室教授, 主要研究方向为信号与信息处理、认知无线电和大数据分析。



王航龙 (1989-), 男, 西安邮电大学无线网络安全技术国家工程实验室硕士生, 主要研究方向为数据挖掘。