ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 1,Jan. 2018,pp. 75－84
DOI:10.16337/j.1004-9037.2018.01.009

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

非均衡 IPTV 数据集下的用户报障预测

吴志峰[1]　黄若尘[1]　魏昕[1]　黄荣谔[2]　周　亮[1]

(1.南京邮电大学通信与信息工程学院，南京,210003；2.中国电信江苏分公司，南京,210017)

摘　要：针对传统算法在非均衡交互式网络电视(Internet protocol television,IPTV)数据集下用户报障预测效果不理想的问题,本文将影响网络服务质量(Quality of service,QoS)的传统网络参数和主观反映用户体验质量(Quality of experience,QoE)的 MOS 评分结合来预测用户是否报障。本文在已有的 ODR-BSMOTE-SVM 算法基础上,针对过采样算法产生噪声以及核参数没有进行优化的缺陷,提出了一种改进型算法。该改进算法首先采用欠采样、过采样算法及数据清洗算法对原始非均衡数据进行处理,然后通过自适应变核参数寻找近似最优值,最终实现提升分类效果。实验结果表明,较传统标准支持向量机(Support vector machine，SVM)算法和 ODR-BSMOTE-SVM 算法,本文算法能获得更佳的预测效果。

关键词：非均衡数据;服务质量;数据清洗;支持向量机

中图分类号：TP181　　**文献标志码**:A

Prediction for User′s Complaint in Imbalanced IPTV Dataset

Wu Zhifeng[1], Huang Ruochen[1], Wei Xin[1], Huang Rongxu[2], Zhou Liang[1]

(1. College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China; 2. China Telecommunications Jiangsu Branch, Nanjing, 210017, China)

Abstract: In the imbalanced internet protocol television (IPTV) dataset, the traditional algorithm performs not well in terms of predicting the user′s complaint. For this problem, this paper combines traditional network parameters that influence the network quality of service (QoS) with MOS score that objectively reflects the quality of experience (QoE) to predict user′s complaint. And then we propose an improved algorithm based on the existing ODR-BSMOTE-SVM algorithm for the defects that the over-sampling algorithm will produce noise and there is not any optimization for kernel parameters. In the improved algorithm, under-sampling algorithm, over-sampling algorithm and data cleaning algorithm are firstly used to process the original imbalanced dataset. Then, through searching for the approximate optimal value by adaptive variable kernel parameters, the classification effect is ultimately improved. Experimental results show that the improved algorithm performs better than the traditional standard support vector machine (SVM) and the ODR-BSMOTE-SVM algorithm in predicting user′s complaint.

Key words: imbalanced data set; quality of service; data cleaning; support vector machine

---

基金项目:国家重点基础研究发展计划(2013CB329005)资助项目;国家自然科学基金(61322104，61571240)资助项目;江苏高校优势学科建设工程资助项目;江苏省自然科学基金(BK20161517)资助项目;江苏省青蓝工程资助项目。
收稿日期:2016-03-03;**修订日期**:2016-10-24

引 言

伴随着多媒体通信技术的迅猛发展,以宽带互联网为基础的交互式网络电视(Internet protocol television, IPTV)极大便利了普通居民在家中享受交互式、个性化、自由定制的视频服务与增值应用服务[1]。在视频传输过程中,传统的网络服务质量(Quality of service, QoS),如带宽、丢包、延迟和抖动等,在一定程度上影响用户的观看体验[2-4],但这些网络参数无法精准描述 IPTV 业务的传输状况能否真实满足观看者的需求,同样也不能准确预测用户有无投诉报障[5-6]。而精确分类预测报障用户,便于提前改善服务,提高用户满意程度,增加用户黏度。

如今 IPTV 服务提供商不断将关注点从网络 QoS 转变为用户体验质量(Quality of experience, QoE)[7]。正是因为用户体验是一种主观的情感变化和个体化的感受,它的诸多因素很难量化和具体化。在文献[8]中阐述了量化 QoE 的方法有两类别法和 MOS 评分法[9],然而从大量已知的 IPTV 用户报障反馈中不难看出,高 MOS 值并不能很好地降低用户报障率[10]。

不仅如此,在 IPTV 技术日益成熟的趋势下,报障的用户占整体用户的比例也日益下降。因此,用户数据将不可避免地成为非均衡数据集,且非均衡比例将持续增大。如今,如何从庞大的非均衡的数据集中成功对少数类进行分类也成为不少学者研究的热点[11]。为了解决该问题,文献[12]从数据层面上阐述了众多关于过采样和欠采样的方法,如合成少数类过采样(Synthetic minority oversampling technique,SMOTE)、Borderline-SMOTE(BSMOTE)等算法和随机欠采样、Informed 欠采样等欠采样算法。文献[13]也提出了利用泰森多边形来改善非均衡数据集的分类性能。

在分类预测算法方面,传统的算法和大数据下的机器学习算法也在不断更新[14]以适应日益复杂的应用场景。其中由 Vapnik 等人提出的支持向量机(Support vector machine, SVM)借助最优化方法来解决机器学习中不少"维数灾难"和"过学习"等困难。文献[15]提出的基于 ODR 和 BSMOTE 结合的 SVM 分类算法能够在非均衡数据下取得不错的效果。

本文针对传统单一 QoS 或 QoE 在预测用户报障时存在的固有缺陷,将二者结合起来进行预测,可以显著提升预测分类性能。分类结果可用于 IPTV 服务提供商提前改善服务质量的重要参考指标。

本文的实验数据来源于江苏省电信用户的 IPTV 观看数据,其中包含着少部分的用户报障信息。纵观全省 IPTV 数据,用户报障所占比例不大,故实验数据源是一个典型的非均衡数据集。同时,本文在深入理解传统非均衡数据处理与支持向量机等相关理论的基础上,引入了 ODR-BSMOTE-TOMEK 和自适应变 SVM 核参数相结合的集成算法(OBT-Adaptive-SVM)。该方法在均衡用户报障与非报障数据基础之上,重点清除了人工生成样本点在 SVM 分类边界上难以分类的杂质点。在运用 SVM 分类用户报障的同时自适应调整核参数 sigma 以寻找最佳的分类效果。本文提出的 OBT-Adaptive-SVM 算法比传统的分类算法在预测精度上有显著提高,因而成功地应用于预测 IPTV 用户报障与否的实例中。

1 系统描述

本文系统由清洗及特征提取、数据均衡与建模预测 3 个模块构成,如图 1 所示。

清洗及特征提取模块负责清洗原始数据、特征提取等功能;数据均衡模块负责均衡清洗后的原始数据集;建模预测模块拟采用支持向量机 SVM 完成对均衡后的数据进行预测分类。

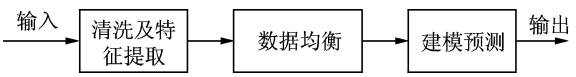


图 1 系统架构

Fig. 1 System architecture

2 理论分析

2.1 支持向量机

作为一种二值分类模型,SVM 可以利用核技巧将输入空间转换到高维特征空间,使之在本质上具有间隔最大特性的非线性分类器。在间隔最大化策略下,SVM 可以等效为求解凸二次规划的最优化算法。算法描述如下。

在处理非线性回归的问题中,输入训练数据集为

$$T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)\}$$

其中 $\boldsymbol{x}_i\in\boldsymbol{X}=\boldsymbol{R}^n$,$y_i\in Y=\{-1,+1\}$,$i=1,2,\cdots,N$,$\boldsymbol{x}_i$ 为第 i 个特征向量,也称为实例,y_i 为 $\boldsymbol{x}_i$ 的类标记,当 $y_i=+1$ 时,称 $\boldsymbol{x}_i$ 为正例;当 $y_i=-1$ 时,称 $\boldsymbol{x}_i$ 为负例。

输出:分类决策函数 $f(x)$

Step 1 选取适当的核函数 $K(x,z)$ 和适当的参数 C,构造并求解最优化问题

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,y_j)-\sum_{i=1}^{N}\alpha_i \tag{1}$$

$$\text{s.t}\quad\sum_{i=1}^{N}\alpha_i y_j=0 \tag{2}$$

$$0\leqslant\alpha_i\leqslant C,i=1,2,\cdots,N \tag{3}$$

求得最优解 $\boldsymbol{\alpha}^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^{\mathrm{T}}$,其中 α_i 为拉格朗日乘子。

Step 2 选择 $\boldsymbol{\alpha}^*$ 的一个正分量 $0<\alpha_j^*<C$,计算分离超平面方程的截距 b

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i K(\boldsymbol{x}_i\cdot\boldsymbol{x}_j) \tag{4}$$

Step 3 构造决策函数

$$f(x)=\text{sign}\Big(\sum_{i=1}^{N}\alpha_i^* y_j K(\boldsymbol{x}\cdot\boldsymbol{x}_i)+b^*\Big) \tag{5}$$

当 $K(x,z)$ 是正定核函数时,问题 (1~3) 是凸二次规划问题,理论上存在解。其中,常用核函数有高斯径向基核函数、多项式核函数、sigmoid 核函数及字符串核函数等,本文选用高斯径向基核函数 $K(x,z)=\exp(-\|\boldsymbol{x}-\boldsymbol{z}\|^2/2\sigma^2)$ 作为核函数,因为它在许多典型应用中具有良好的效果[16-17]。对应的支持向量机是高斯径向基函数 (Radial basis function) 分类器,在此情形下,分类决策函数为

$$f(x)=\text{sign}\Big(\sum_{i=1}^{N_s}\alpha_i^* y_j\exp\Big(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}\Big)+b^*\Big) \tag{6}$$

2.2 BSMOTE 过采样算法与 ODR 欠采样算法

为了减少非均衡数据集对分类的不利影响,通常可以使用过采样算法与欠采样算法处理非均衡数据集。本文采用已有的边界人工少数类过采样算法(Borderline synthetic minority oversampling technique, BSMOTE)和逐级优化递减欠采样算法(Optimization of decreasing reduction, ODR)。其中 BSMOTE 算法在克服传统 SMOTE 算法固有盲目性缺陷的基础之上,仅复制加大分类边界少数样本点的数量,通过提高少数类样本点中对 SVM 算法分类贡献率大的样本点的使用率,进而改善 SVM 分类器的性能。

因为在多数类样本中存在噪声样本和大量冗余样本,它们不仅会降低分类效率,还会严重干扰 SVM 分类器决策过程。而 ODR 算法是通过 KNN 算法判断多数类样本对邻域内样本分类影响的好坏程度、优先清除对分类效果有负面影响的样本,然后再删除影响不大的样本。

3 改进的 ODR-BSMOTE-SVM 算法

当前处理非均衡数据集下的 SVM 通常包含两种主要的方法：一是数据样本的改善；二是算法层面的改进[18]。传统 ODR-BSMOTE-SVM 算法的缺点是没有重新审视 BSMOTE 算法所生成人工样本点的质量好坏。因为其中有些人工样本点可能会干扰 SVM 分类器决策的过程，同时原算法中应用 SVM 的分类结果往往不是全局最优解。故本文提出如下改进以克服传统算法的固有缺陷。

3.1 ODR-BSMOTE-TOMEK 数据均衡算法

在 BSMOTE 算法作用下增加边界的少数类样本点，虽然这样能通过增大支持向量的个数从而提升 SVM 的分类性能，但是不可避免地又会在 BSMOTE 人工生成的样本点过程中重新生成一些 SVM 难以分辨的杂质点，这反过来造成 SVM 分类性能下降。本文在非均衡数据经过 ODR-BSMOTE 算法之后，引入了数据清洗 TOMEK 过程。算法描述如下：

Step 1 随机从样本集合 S 中抽取出样本点 $\boldsymbol{x}_i \in S$。在样本集 S 中寻找与 $\boldsymbol{x}_i$ 最近邻的点 $\boldsymbol{x}_j \in S$。

Step 2 样本集 S 中寻找与 $\boldsymbol{x}_j$ 最近邻的点 $\boldsymbol{x}_k \in S$。

Step 3 判断 $\boldsymbol{x}_i == \boldsymbol{x}_k$ 是否成立，若成立则跳转 Step 4，否则 $\boldsymbol{x}_i = \boldsymbol{x}_j$，$\boldsymbol{x}_j = \boldsymbol{x}_k$，然后跳转 Step 2。

Step 4 判断 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_k$ 的类别是否一致，若一致，则将这两个点保存至新的样本集 S_{new}，然后从样本集 S 中删除这两点。若不一致，则直接从样本集 S 中删除这两点。

Step 5 判断样本集 S 中的个数是否为大于 0 的偶数，若为偶数则重复 Step 1，否则结束退出。

最终输出的样本集 S_{new} 则是改进 SVM 算法的输入数据集的来源。

3.2 改进的自适应变核参数 SVM 算法

相对于核函数，真正决定 SVM 性能的因素其实是核参数。目前国内外也有很多成熟的算法来寻找合适的核参数值[19-20]。但是本文的数据源来自全省用户数据，体量较大，从算法的执行效率、时间成本与系统负载均衡等多方面角度考虑，最终本文在算法层面上的改进是在深入研究传统标准 SVM 的基础之上，采用改进的自适应变核参数 SVM 算法（Adaptive-SVM）寻找最优值，算法流程如图 2 所示。图 2 中，在初始化高斯核参数 sigma 的同时也设定好 sigma 的最大上限值 max_sigma。出于减轻程序运行负荷的考虑，算法优先使用粗步长，如 0.1，先将 sigma 值从初始值按粗步长逐渐增大，然后计算粗步长下各个 sigma 值所对应的预测分类效果，同时仅存储当前最佳分类效果所对应的参数 sigma 值作为最佳局部点。当 sigma 值大于 max_sigma 值时，算法停止粗步长搜索。

紧接着算法根据粗步长搜索得到的最佳局部点，将参数 sigma 值自动定位到该最佳局部点的左侧附近，在限定好细步长变化范围后，改用细步长改变参数 sigma 值。本文选取的细步长与粗步长比例取 1∶10，计算细步长下各个 sigma 值所对应的预测分类效果，同时存储当前最佳分类效果所对应的参数 sigma 值作为全局最佳点。

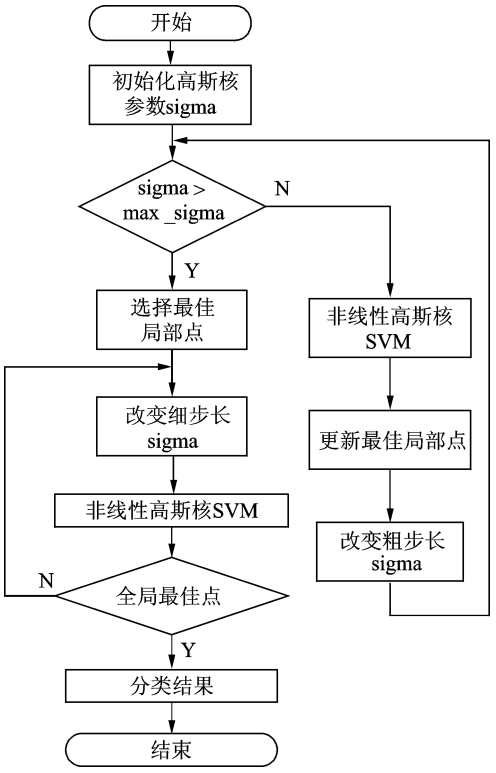


图 2 自适应变核参数 SVM 算法流程图

Fig. 2 Flow chart of SVM algorithm with adaptive variable kernel parameters

因为该自适应变核参数 SVM 算法所改变的步长不是连续型数值,而是间隔较小的离散值,故无法收敛于真正最大值。因此该算法在实际运行中仅需在离散值中找到最合适的结果。这种做法的优势在于用较小分类精度损失为代价,换取 SVM 算法在大数据下训练与测试时运行效率的大幅提高,把在步长足够小情况下所计算出的局部最优值近似为最优值,并在算法结束前同时输出最佳的分类效果。

3.3 OBT-Adaptive-SVM 集成算法

以往在处理非均衡数据下的分类时,单一的数据层面上的改进或算法层面上的改进在一定程度上可以提高分类精度,但是效果不够理想[21-22]。故本文将改进的 ODR-BSMOTE-TOMEK 数据均衡算法和改进的自适应变核参数 SVM 算法有机地结合成适用于用户报障预测的 OBT-Adaptive-SVM 集成算法,算法流程如图 3 所示。算法的基本思想是:首先设参数 α 是需要删除的多数类样本个数与多数类和少数类样本之间的差值的比值。首先确定一个合适的 α,然后利用 ODR 和 BSMOTE 算法,按照预定值分别减少多数类样本,增加少数类样本,再将处理后的训练集经过 3.1 节所设计的 TOMEK 数据清洗技术,删除那些 BSMOTE 产生的和现有数据构成的对 SVM 分类性能不良影响的噪声点,从而将处理后的训练集利用 SVM 进行分类。最后,根据 3.2 节所提出的算法自适应调整 SVM 核参数 sigma 使之达到最佳的预测分类效果。

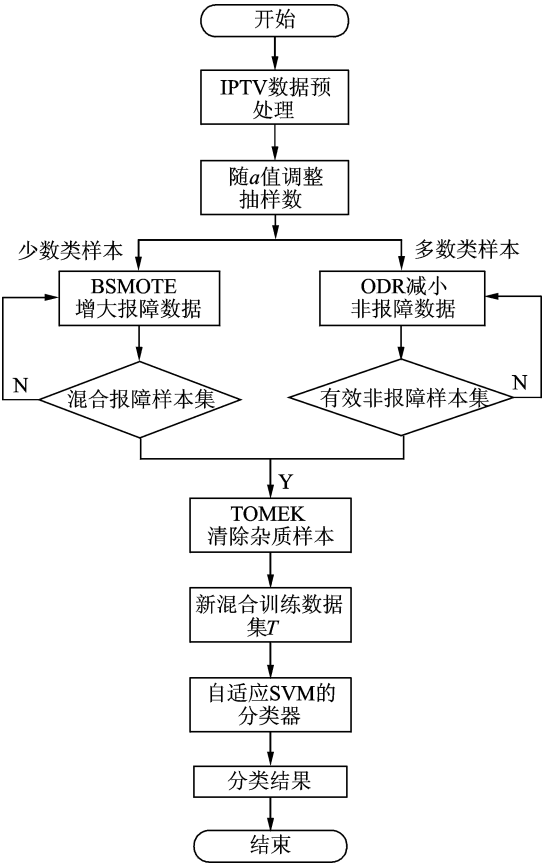


图 3 OBT-Adaptive-SVM 集成算法流程图

Fig. 3 Flow chart of OBT-Adaptive-SVM algorithm

4 实验及数据分析

4.1 实验数据及预处理

本文的原始数据源自于江苏电信全省 IPTV 用户家中机顶盒上传的数据,且已分类出报障用户和非报障用户,接着进行如下处理。

(1) 特征参数的选取

原始数据集含有 24 个字段,但是很多字段是非数值型,无法参与 SVM 计算。因此,本文最终选择 10 个字段作为 SVM 算法的特征参数,最后 1 个特征 CLASS 是区分该用户是否报障,用"+1"表示报障用户,"−1" 表示非报障用户,如表 1 所示。

(2) 相同用户数据合并取平均

本文选用的原始数据集含有 4 723 101 条

表 1 机顶盒上传数据各字段含义

Tab. 1 Meaning of each set-top box's field

字段	含义
SEVERITY	告警等级
ALARM_NUM	告警数
LOSSRATE	丢包率
DOWN_BANDWIDTH	机顶盒下行带宽
MEDIARATE	码率
MDI_DF	网络抖动时延
MDIMLR	每秒丢失媒体数据包数量
VSTQ	当前网络状况
MOS_VALUE	MOS 值
CPU_USAGE	CPU 使用率
CLASS	报障与否标识符

IPTV 用户记录数据，其中报障记录有 48 172 条，占 1.02%，非报障记录有 4 674 929 条，占98.98%，即报障记录与非报障记录的比例约为 1∶97。现将相同用户所有 IPTV 观看记录进行求平均值作为该用户的记录。经过处理后的数据总量为 439 050 条，其中报障用户有 4 871 个，占 1.11%，非报障用户有 434 179 个，占98.89%，即报障用户与非报障用户的比例约为 1∶89，由此可知，数据集的不平衡程度还是相当大，若不经过数据层面算法的处理，很难对用户报障与否进行很准确的预测。

4.2 不平衡数据集的评价标准

在评测非均衡样本集的分类器性能时，传统的性能评估指标是从整体分类情况角度看待多数类和少数类的准确率，有时并不适用于非均衡数据集。测试集中分类样本集的混淆矩阵如表 2 所示。正因为如此，越来越多的学者在研究非均衡数据分类时采用如下评判标准[23]，本文也采用这些评价标准。

表 2 测试集中分类样本集的混淆矩阵

Tab. 2 Classification confusion matrix of sample sets in test sets

分类	预测为不报障	预测为报障
用户不报障(多数类)	TN	FP
用户报障(少数类)	FN	TP

(1)训练集总体准确率：算法分类器在训练集中正确预测出用户报障和不报障的总个数与训练集总数的比值。

(2)测试集用户报障召回率 Recall_Min，表达式为

$$\text{Recall\_Min}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

(3)测试集用户报障查准率 Precision_Min，表达式为

$$\text{Precision\_Min}=\frac{\text{TP}}{\text{FP}+\text{TP}}$$

(4)测试集用户不报障召回率 Recall_Maj，表达式为

$$\text{Recall\_Maj}=\frac{\text{TN}}{\text{TN}+\text{FP}}$$

(5)用户报障分类精度和用户不报障分类精度的测试集几何平均值 G_mean[24]，表达式为

$$G\_\text{mean}=\sqrt{\text{Recall\_Min}*\text{Recall\_Maj}}$$

G_mean 是保持用户报障、用户不报障分类精度平衡的情况下最大化它们的精度，也就是说只有当 Recall_Min 和 Recall_Maj 同时都最高的时候，G_mean 的值才最大。

(6)测试集用户报障的 F_measure，表达式为

$$F\_\text{measure}=\frac{2*\text{Recall\_Min}*\text{Precision\_Min}}{\text{Recall\_Min}+\text{Precision\_Min}}$$

F_measure 指标是一种综合考虑查全率和查准率的分类评价指标，能综合体现出分类器对用户报障和用户不报障的分类效果，但更加侧重于用户报障样本的分类效果。

因为报障用户与非报障用户的比例约为 1∶89，所以本文在测试数据集中取得报障与非用户报障比例也是 1∶89。

4.3 实验结果与分析

为了验证本文提出的 OBT-Adaptive-SVM 算法的性能，根据表 3 中的数据集进行 3 次实验，分别计算传统标准 SVM、ODR-BSMOTE-SVM 算法和 OBT-Adaptive-SVM 算法的最终分类情况。算法中涉及到 k-NN 算法时 k 一律取值为 5，同时为了更直观地表示最终结果，SVM 的核函数一律选择高斯径向基函数，惩罚因子 $C=1\,000$，核参数 sigma 从 0.1 开始并以 0.1 为步长递增到 2.0，以搜索最佳分类的大致位置。核宽度 sigma 之所以最大取 2.0，这是因为算法分类性能在核参数达到特定值后，受 sigma 的

影响急剧减小。而 OBT-Adaptive-SVM 算法在大致搜索到最优局部点后,在其附近以 0.01 为步长递增寻找近似全局最优点。

表 3 数据集的基本信息表

Tab. 3 Basic information table of datasets

数据集类型	SVM	ODR-BSMOTE-SVM	OBT-Adaptive-SVM
训练集报障样本数	4 000	10 000	10 000
训练集非报障样本数	4 000	10 000	10 000
测试集报障样本数	800	1 000	1 000
测试集非报障样本数	71 200	89 000	89 000

此外为了公平比较算法性能,ODR-BSMOTE-SVM 算法与 OBT-Adaptive-SVM 算法中的某些参数设置相同,在 ODR 算法中均固定删除系数 α 为 0.3,取 BSMOTE 算法中的 $k=5$ 近邻,然后 5 近邻中选择 $s=3$ 的随机少数点。3 个算法的分类结果具体如下:

(1)采用表 3 中的数据,对 SVM 算法进行 Matlab 仿真,得到结果如图 4,5 所示。从图 4,5 中可以看出,不对原始数据进行任何数据层面算法的处理,SVM 算法得到的最佳点将会在核参数 sigma 为0.3 附近。此时的报障与不报障的召回率在 65%左右,但此时的 G_mean 和 F_measure 的值普遍都很低,均在 0.1 以下。因此,该算法分类效果并不太明显。

(2)采用表 3 中的数据,对 ODR-BSMOTE-SVM 算法进行 Matlab 仿真,得到结果如图 6,7 所示。由图 6,7 可以看出,数据集在经过 ODR-BSMOTE 算法处理后,再经过 SVM 得到分类结果明显优于不做任何数据均衡处理的标准 SVM 算法,且高斯核宽度 sigma 在 0.2 以前可以获得不错的 G_mean 和 F_measure。因此,该算法的分类效果较标准 SVM 算法有所提高。

(3)采用表 3 中的数据,对 OBT-Adaptive-SVM 算法进行 Matlab 仿真,当核参数 sigma=0.1 时得到结果如图 8,9 所示。由图 8,9 大致可以看出,数据集在经过 ODR-BSMOTE-TOMEK 算法处理后的分类结果明显优于 ODR-BSMOTE-SVM 算法,且高斯核宽度 sigma 在 0.2 以前可以获得非常好的 G_mean和 F_measure。

接下来,再经过改进的自适应变核参数 SVM 算法处理后,核参数自动调整到最优局部点的左侧附近,即起始点 0.01,这是受到核参数 sigma 必须大于 0 的要求,然后算法以步长为 0.01 递增改变核参数,最终得到细步长下的分类结果,如图 10,11 所示。

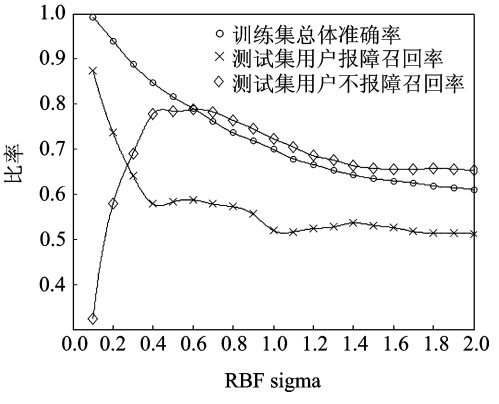


图 4 标准 SVM 的召回率

Fig. 4 Recall rate of standard SVM

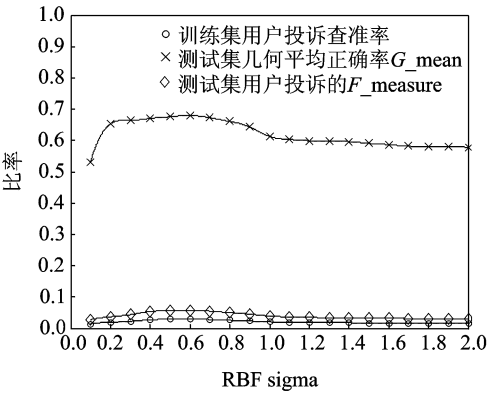


图 5 标准 SVM 算法的 G 和 F 指标

Fig. 5 G_mean and F_measure of standard SVM

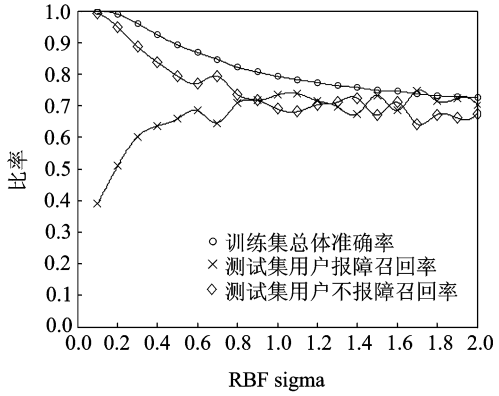


图 6 ODR-BSMOTE-SVM 算法的召回率

Fig. 6 Recall rate of ODR-BSMOTE-SVM

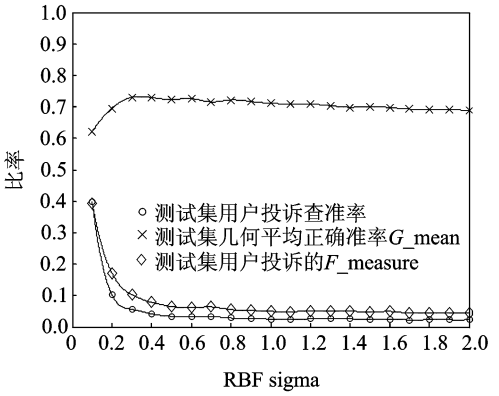


图 7 ODR-BSMOTE-SVM 算法 G 和 F 指标

Fig. 7 G_mean and F_measure of ODR-BSMOTE-SVM

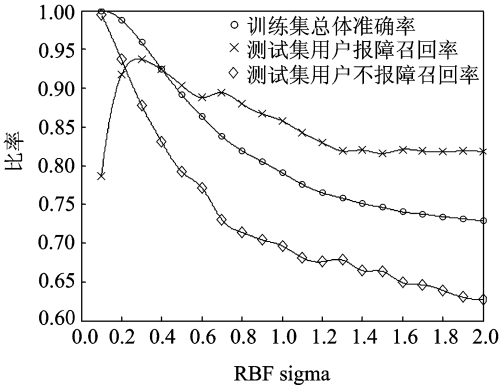


图 8 OBT-Adaptive-SVM 算法的召回率

Fig. 8 Recall rate of OBT-Adaptive-SVM

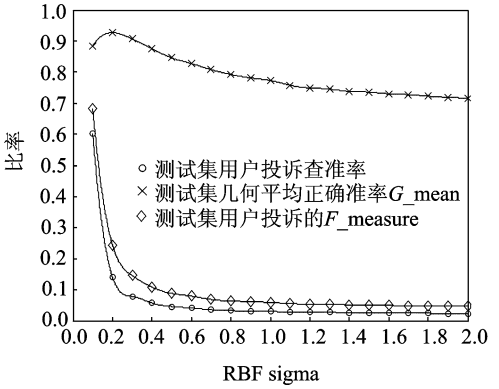


图 9 OBT-Adaptive-SVM 算法 G 和 F 指标

Fig. 9 G_mean and F_measure of OBT-Adaptive-SVM

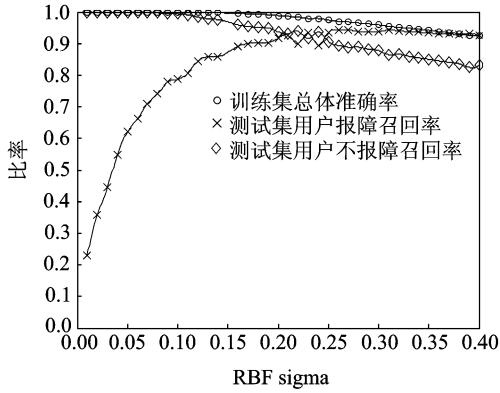


图 10 细步长下 OBT-Adaptive-SVM 算法的召回率

Fig. 10 Recall rate of OBT-Adaptive-SVM under the fine step

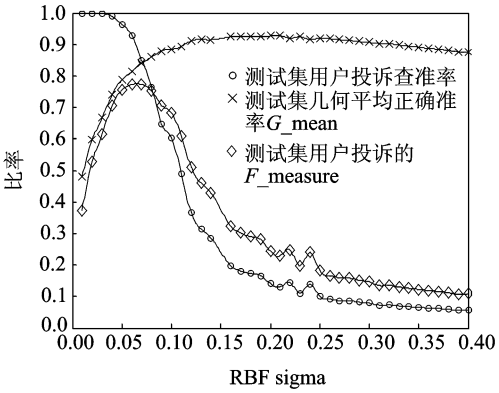


图 11 细步长 OBT-Adaptive-SVM 算法 G 和 F 指标

Fig. 11 G_mean and F_measure of OBT-Adaptive-SVM under the fine step

由图 10,11 可以看出,改进的自适应变核参数 SVM 算法在核参数 sigma=0.21 处得到近似最优分类效果,预测用户报障召回率高达 92.6%,预测用户不报障召回率也高达 93.08%,本文提出的 OBT-

Adaptive-SVM 算法分类效果分别和标准 SVM 算法、ODR-BSMOTE-SVM 算法进行比较,用户报障准确率显著提高。图 12 为本实验中 3 种算法最佳召回率的比较。由图 12 可明显看出,ODR-BSMOTE-TOMEK 算法在经过数据处理后,分类性能较标准 SVM 有所提高。而本文提出的 OBT-Adaptive-SVM 集成算法分类性能相比前两个算法显著地提升,说明本文所提方法在预测 IPTV 用户报障与否的应用中具有可行性。尽管报障用户与非报障用户的比例为 1∶89,但是改变报障用户与非报障用户的比例仅会整体改变这 3 个算法准确率的具体数值,而不改变最终结论。同样,本文在 k-NN 算法中 k 的取值一律为 5,降低了计算复杂度,也不影响最终结论。

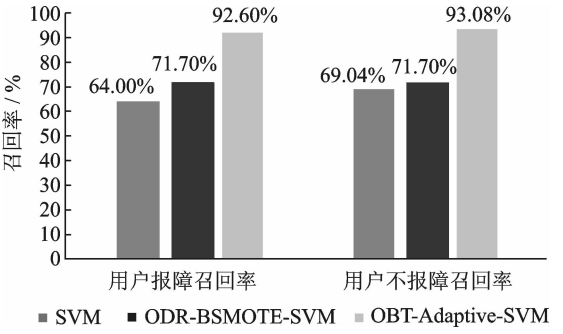


图 12 三种算法的召回率比较

Fig. 12 Comparison of recall rotte of three algorithms

5 结束语

本文方法在数据层面上一方面削弱噪声点和冗余点对报障预测的干扰,另一方面加强少数有效样本点对正确分类的贡献,同时再加入 TOMEK 算法以清除 BSMOTE 算法生成的在 SVM 分类边界上难以区分判断的杂质点。在算法层面上本文方法不仅引入了自适应改变 SVM 核函数的参数 sigma 的算法,还将 OBT 算法和自适应变核参数 SVM 算法二者结合成一种行之有效的集成学习系统。在预测分类报障用户问题上,实验结果显示本文所提的 OBT-Adaptive-SVM 集成算法的预测准确率比传统的 SVM 和 ODR-BSMOTE-SVM 算法都高,在 IPTV 用户预测的应用中取得比传统算法更佳的性能效果。此外,如何同时自适应改变 SVM 的惩罚参数以提高用户报障预测效果将是下一步研究的目标。

参考文献:

[1] 史志明. 网络视频质量评估方法与测试技术研究[D]. 北京:北京邮电大学, 2013.
Shi Zhiming. Research on network video quality assessment method and measure technology[D]. Beijing: Beijing University of Posts and Telecommunications, 2013.

[2] Zhou L, Hu R, Qian Y, et al. Energy-spectrum efficiency tradeoff for video streaming over mobile ad hoc networks[J]. Selected Areas in Communications, IEEE Journal on, 2013, 31(5): 981-991.

[3] Zhou L, Yang Z, Wang H, et al. Impact of execution time on adaptive wireless video scheduling[J]. Selected Areas in Communications, IEEE Journal on, 2014, 32(4): 760-772.

[4] 古强. 直播型 IPTV QoS 若干关键技术的研究[D]. 北京:北京邮电大学, 2010.
Gu Qiang. On some live IPTV QoS key techniques[D]. Beijing: Beijing University of Posts and Telecommunications, 2010.

[5] 李海林,郭崇慧,杨丽彬. 基于时间序列数据挖掘的故障检测方法[J]. 数据采集与处理, 2016, 31(4): 782-790.
Li Hailin, Guo Chonghui, Yang Libin. Fault detection algorithm based on time series data mining[J]. Journal of Data Acquisition and Processing, 2016, 31(4): 782-790.

[6] 周赛赛. IPTV 系统 QoS 关键技术研究与改进[D]. 长沙:中南大学, 2008.
Zhou Saisai. Research and improvement on QoS key technology in IPTV system[D]. Changsha: Central South University, 2008.

[7] 张大陆,祝嘉麒. 网络传输中 IPTV 的 QoE 评估模型的研究[J]. 计算机工程与应用, 2013, 49(20): 71-76,135.
Zhang Dalu, Zhu Jiaqi. QoE evaluation model for IPTV in network transmission[J]. Journal of Computer Engineering and Applications, 2013, 49(20): 71-76,135.

[8] 林闯,胡杰,孔祥震. 用户体验质量(QoE)的模型与评价方法综述[J]. 计算机学报, 2012, 35(1): 1-15.
Lin Chuang, Hu Jie, Kong Xiangzhen. Survey on models and evaluation of quality of experience[J]. Chinese Journal of Computers, 2012, 35(1): 1-15.

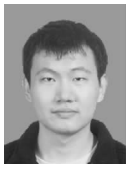
[9] Balachandran A, Sekar V, Akella A, et al. Developing a predictive model of quality of experience for internet video[C]// ACM SIGCOMM Computer Communication Review, 2013, 43(4): 339-350.

[10] Sun S, Wei X, Wang L, et al. Association analysis and prediction for IPTV service data and user's QoE[C]//Wireless Communications & Signal Processing (WCSP), 2015 International Conference on. Nanjing, China: IEEE, 2015: 1-5.

[11] Jeatrakul P, Wong K W. Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm [C]//Neural Networks (IJCNN), The 2012 International Joint Conference on. Brisbane, QLD: IEEE, 2012: 1-8.
[12] He H, Garcia E A. Learning from imbalanced data[J]. Knowledge and Data Engineering, IEEE Transactions on, 2009, 21(9): 1263-1284.
[13] Young W A, Nykl S L, Weckman G R, et al. Using Voronoi diagrams to improve classification performances when modeling imbalanced datasets[J]. Neural Computing and Applications, 2015, 26(5): 1041-1054.
[14] 何清,李宁,罗文娟,等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能, 2014, 27(4): 327-336.
He Qing, Li Ling, Luo Wenjun, et al. A survey of machine learning algorithms for big data[J]. Journal of Pattern Recognition and Artificial Intelligence, 2014, 27(4): 327-336.
[15] 陶新民,童智靖,刘玉,等. 基于 ODR 和 BSMOTE 结合的不均衡数据 SVM 分类算法[J]. 控制与决策, 2011, 26(10): 1535-1541.
Tao Xinmin, Tong Zhijing, Liu Yu, et al. SVM classifier for unbalanced data based on combination of ODR and BSMOTE [J]. Journal of Control and Decision, 2011, 26(10): 1535-1541.
[16] 李忠国,侯杰,王凯,等. 模糊支持向量机在路面识别中的应用[J]. 数据采集与处理, 2014, 29(1): 146-151.
Li Zhongguo, Hou Jie, Wang Kai, et al. Application of fuzzy support vector machine on road type recognition[J]. Journal of Data Acquisition and Processing, 2014, 29(1): 146-151.
[17] 汪友生,胡百乐,张丽杰,等. 基于支持向量机的动脉硬化斑块识别[J]. 数据采集与处理, 2012, 27(3): 283-286.
Wang Yousheng, Hu Baile, Zhang Lijie, et al. Recognition of atherosclerotic plaque based on support vector machine[J]. Journal of Data Acquisition and Processing, 2012, 27(3): 283-286.
[18] 陶新民,郝思媛,张冬雪,等. 不均衡数据分类算法的综述[J]. 重庆邮电大学学报(自然科学版), 2013, 25(1): 101-110,121.
Tao Xinmin, Hao Siyuan, Zhang Dongxue, et al. Overview of classification algorithms for unbalanced data[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2013, 25(1): 101-110,121.
[19] 董国君,哈力木拉提·买买提,余辉. 基于 RBF 核的 SVM 核参数优化算法[J]. 新疆大学学报(自然科学版), 2009, 26(3): 355-358,363.
Dong Guojun, Halmurat Maimait, Yu Hui. Algorithms of optimizing SVM′s kernel parameters with RBF kernel[J]. Journal of Xinjiang University(Natural Science Edition), 2009, 26(3): 355-358,363.
[20] 刘俊芳. 粒子群和人工蜂群的混合优化算法优化 SVM 参数及应用[D]. 太原: 太原理工大学, 2012.
Liu Junfang. A hybrid algorithm of PSO and ABC used to optimize the parameters of SVM and its application [D]. Taiyuan: Taiyuan University of Technology, 2012.
[21] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[M]. Berlin Heidelberg:Springer, 2003: 107-119.
[22] Sun Y, Kamel M S, Wong A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12): 3358-3378.
[23] 林智勇,郝志峰,杨晓伟. 若干评价准则对不平衡数据学习的影响[J]. 华南理工大学学报(自然科学版), 2010, 38(4): 147-155.
Lin Zhiyong, Hao Zhifeng, Yang Xiaowei. Effects of serveral evaluation metrics on imbalanced data learning [J]. Journal of South China University of Technology (Natural Science Edition), 2010, 38(4): 147-155.
[24] Chawla N V, Japkowicz N, Kotcz A. Editorial: Special issue on learning from imbalanced data sets[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 1-6.

作者简介:

吴志峰(1992-),男,硕士研究生,研究方向:大数据、数据挖掘及机器学习,E-mail: njuptwzf@163.com。



黄若尘(1990-),男,博士研究生,研究方向:云计算与大数据。

魏昕(1983-),男,博士,副教授,研究方向:图像处理与人工智能。

黄荣谓(1985-),男,硕士,工程师,研究方向:无线网络与大数据。

周亮(1981-),男,博士,教授,研究方向:多媒体通信与机器学习。

(编辑:张黄群)