DOI:10.11918/j.issn.0367-6234.201803092

多源非平衡交通检测数据的异常识别方法

邢 雪[1,2], 于德新[2,3], 周户星[2,3], 田秀娟[2]

(1.吉林化工学院 信息与控制工程学院,吉林 吉林 132022;2.吉林大学 交通学院,长春 130022;
3.吉林省智能交通工程研究中心,长春 13002)

摘 要: 为保证交通检测数据的准确性并服务于实时的交通状态判别和预测,交通大数据采用多种检测源数据协同处理并利用机器学习的方法进行异常识别. 异常检测数据的识别主要基于机器学习中 AdaBoost 方法实现. 在算法的训练过程中,为消除单一检测源数据的离群现象,训练数据选取同一路段上多种检测源提供的数据集. 在算法的决策过程中,通过代价敏感方法的优势来改进 AdaBoost 的决策. 实验结果表明:基于非均衡特性改进的 AdaBoost 模型迫使分类器更加关注了待识别的异常样本,增强了 AdaBoost 决策过程中训练决策树规则的代表性,提高了异常类样本的分类准确率. 高速公路实例检测数据集验证了改进算法与相关经典算法的检测准确度、误检率、误警率等指标,其中改进模型与原模型相比,准确率提高了 5.547%,误检率减低了 6.792%. 多种算法的 ROC 曲线对比表明改进的 AdaBoost 方法筛选交通检测样本的可靠度更高,可有效调整由非平衡数据导致的分类误差.

关键词: AdaBoost;数据异常识别;多源交通数据;非平衡检测数据;机器学习

中图分类号: U491.1 **文献标志码**: A **文章编号**: 0367-6234(2019)09-0165-06

A method of abnormal data recognition of multi-source traffic with non-equilibrium feature

XING Xue[1,2], YU Dexin[2,3], ZHOU Huxing[2,3], TIAN Xiujuan[2]

(1. College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, Jilin, China;
2. Transportation College, Jilin University, Changchun 132002, China;
3. Jilin Engineering Research Center for Intelligent Transportation System, Changchun 132002, China)

Abstract: The identification and prediction of real-time traffic conditions rely on data processing. Abnormal data recognition in traffic big data uses machine learning methods with multi-source traffic to ensure the accuracy of traffic detection data. The recognition of anomaly detection data is based on AdaBoost method in machine learning. To eliminate the outlier phenomenon of the single detection source data, the training dataset of the training process selected datasets provided by multiple detection sources on the same road section. The cost-sensitive method optimizes the decision-making process of the improved algorithm. Experimental results show that the improved AdaBoost model forced the classifier to pay more attention to abnormal class samples, which enhanced the representation of training decision tree rules in the AdaBoost and improved the classification accuracy of abnormal samples. The highway test dataset verified the detection accuracy, false detection rate, false alarm rate, and other indicators of the improved algorithm and related classical algorithms. The accuracy rate of the improved algorithm was increased by 5.547%, and the false detection rate was reduced by 6.792%. The comparison of ROC curves shows that the improved AdaBoost method is more reliable in identifying abnormal samples of traffic detection and can effectively adjust the classification error caused by non-equilibrium data.

Keywords: AdaBoost; abnormal data recognition; multi-source traffic data; non-equilibrium detection data; machine learning

交通状态信息的采集可以通过磁频、波频、视频和安装在车内 GPS 等移动定位设备等技术来完成,另外基于 RFID 技术和手机信令技术也可以起到补充采集作用,因而交通领域产生了大量时空数据集. 为进行高效的交通状态识别[1-2]和预测[3-4],需要掌握精准的交通实时数据,获取交通感知数据集的过程中会出现离群数据[5-6],即所得交通信息数据里存在部分数据,与其他数据相比较明显不一致. 离群数据产生的主要原因:1)采集周期较短;2)采集设备不够完善;3)检测数据的传输错误;4)环境因素突变造成. 在交通状态判别流程若忽略离群数据

收稿日期: 2018-03-28
基金项目: 国家科技支撑计划(2014BAG03B03)
作者简介: 邢 雪(1983—),女,博士研究生;
于德新(1972—),男,教授,博士生导师
通信作者: 邢 雪, patricia_xx@126.com

的存在,将导致无意义的离群数据和交通事件的重要隐含信息混在一起. 为保障交通事件的评判精度性和预测交通态势的时效性,有效地通过多维数据特征在数据集中剥离出离群数据,已成为交通信息处理中所面临的基本问题. 对于交通检测动态数据的识别,文献[7-8]利用阈值法对交通中异常数据进行筛选和识别,文献[9-10]则针对交通管理系统中出现的缺失数据提出组合参数识别法. 最近几年不少国内学者将动态交通数据中的离群数据分为错误数据和不精确数据,并运用交通流理论建立数据的判别规则. 文献[11]利用粗糙集-模糊识别技术对交通数据预处理后进行状态识别. 文献[12-13]以多源检测数据预处理方法为基础实现交通评价. 文献[14-15]提出灰色理论和近邻聚类方法处理具有异常数据的交通流数据. 本文针对交通数据检测中无意义的离群数据,结合先进的机器学习方法,在准确度高、运算速度快的迭代分类算法 AdaBoost 的基础上,利用代价敏感方法的优势,提出了一种基于 AdaBoost 优化决策的筛选离群样本的方法.

1 检测离群数据的决策树构建

1.1 AdaBoost 的基本理论

AdaBoost 分类器是机器学习中比较流行的分类算法[16],在给定特征空间 X 和分类标识 $y \in \{+1, -1\}$, AdaBoost 的核心思想是针对同一个训练集训练不同的弱分类器 $h_t(x)$, 其中 $x \in X$, 然后组合这些弱分类器形成强分类器 $H(x)$, 即

$$H(x) = \mathrm{sign}(f(x)) = \mathrm{sign}\Big(\sum_{t=1}^{T} \alpha_t h_t(x)\Big). \quad (1)$$

从一个包含 n 个元素的训练集 $\{x_i\}$ 开始,对每个元素分类. 通过每轮弱分类器 $h_t(x)$ 的分类结果为训练集元素分配权重 $D_t(x)$. 每轮学习根据分类和权重选择最优的弱分类器 $h_t(x)$, 一旦弱分类器选定即可获取通过分类标识 y_t 与分类器结果 h_t 确定本轮的优度系数 a_t, 同时根据系数 a_t 更新的权值分配 $D_t(x)$. 最后通过不断迭代训练 T 次之后完成强分类器 $H(x)$ 分类过程.

1.2 适应交通数据特征的模型

模型需要整个数据集被分为两个部分,一个训练集和一组测试集,前者用于构建模型,后者用于测试模型的检测能力. 首先将选定数据集随机分成训练集和测试集,并对训练集数据进行多次迭代分类;之后不断利用训练集的分类结果对训练集元素进行权重的变化;更新迭代分类中的权重系数,在有限次的训练后完成强分类的组合,其中在本研究中弱分类器选用决策树分类器.

在交通检测数据集中,每个交通检测点数据可获取大量感知数据即由多种检测源的数据组成,假设有 n 个数据源,每个数据源均通过多个交通参数对检测对象进行描述,则每个时段均能得到一组多源感知数据. 为分析道路截面检测器采集的数据方便,提取数据集中常用的 3 种检测来源(感应线圈数据、地磁数据、卡口数据)的数据,提取其中交通流参数即流量、速度以及占有率进行数据分析,为异常识别数据某采集时间某检测器的交通参数,例如异常识别数据采集时间 t_i 感应线圈得到交通流量 q_{Ci}、地点平均速度 v_{Ci}、时间占有率 o_{Ci}, 则另需提取空间相关的其他检测器数据. 根据上述分析交通特征选择属性分别为交通数据采集时间 t_i, 感应线圈得到交通流量 q_{Ci}, 感应线圈得到地点平均速度 v_{Ci}, 感应线圈得到时间占有率 o_{Ci}, 地磁得到交通流量 q_{Ui}, 地磁得到地点平均速度 v_{Ui}, 地磁得到时间占有率 o_{Ui}, 卡口数据交通流量 q_{Ki}, 卡口数据地点平均速度 v_{Ki}, 卡口数据时间占有率 o_{Ki}, 交通数据质量标志 y_i, $i = 1,2,\cdots,n$, 其中 y_i 取值属于 $\{+1, -1\}$, 表示由检测数据集评判的数据信息为正常数据或是离群数据的判决标签.

根据上述交通数据属性,确定交通检测数据特征空间 $\boldsymbol{X}$ 的特征数为 10,给定自变量数据集合 $\boldsymbol{X}$ 和数据变量 $\boldsymbol{Y}$ 的矩阵分别描述为

$$\boldsymbol{X} = [x_1 \quad x_2 \quad \cdots \quad x_{10}] = \begin{bmatrix} t_1 & q_{C1} & v_{C1} & o_{C1} & q_{U1} & v_{U1} & o_{U1} & q_{K1} & v_{K1} & o_{K1} \\ t_2 & q_{C2} & v_{C2} & o_{C2} & q_{U2} & v_{U2} & o_{U2} & q_{K2} & v_{K2} & o_{K2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_n & q_{Cn} & v_{Cn} & o_{Cn} & q_{Un} & v_{Un} & o_{Un} & q_{Kn} & v_{Kn} & o_{Kn} \end{bmatrix}, \quad (2)$$

$$\boldsymbol{Y} = [y_1 \quad y_2 \quad \cdots \quad y_n]^{\mathrm{T}}. \quad (3)$$

式中: X_i 为一组数据单元, n 为选择输入样本的样本数目; $y_i \in \{+1, -1\}$ 为对应数据异常识别的结果.

2 交通数据检测改进模型

2.1 实时交通检测数据中离群数据特征

道路交通检测器获取的数据包含交通数据采集时间、检测器所属类型、流量、地点平均速度、时间占有率等数据属性. 以下 3 种情况表现为实时道路数据的离群表象:1)道路交通状态检测获取的数据值与实际道路交通状态值偏离较大;2)获取的道路交通状态数据为错误数据,超出了道路交通状态值的合理范围或违背了道路交通的相关规律;3)道路交通状态为异常状态,导致数据出现偏离常规数据趋

势. 本文分析了提取到的山东高速公路 2014 年 11 月 5 日在同一截面 165 个时间点的 3 类交通检测器的数据，图 1 为对相同路段的多源参数（以流量、地点平均速度、时间占有率三参数确定数据点位置）综合散点图. 分析数据可知，一方面交通状态的异常性体现为多检测源的同时或相邻时段内的同步异常；另一方面通过数据对比发现存在不符合交通状态的离群样本存在，且数目明显少于其他类样本的数目（非平衡数据）[17]. 针对交通数据多源同步特性，通过多源交通数据特征剥离出离群数据而不影响交通异常状态数据分析，从而有效保障交通事件的评判度和预测交通态势的效率.

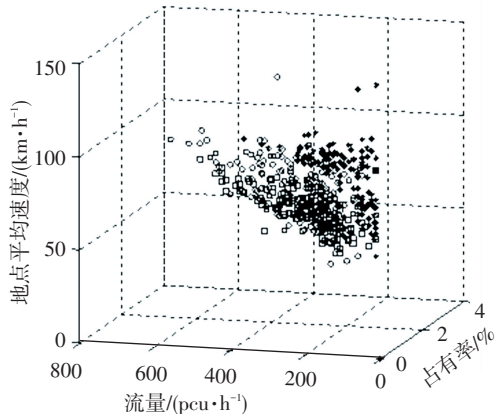


图 1　对相同路段检测时间序列的多源综合散点图

Fig.1　Scatterplot of multi-source parameters data in the same section

2.2　针对交通数据异常识别的改进 AdaBoost 模型

交通数据中非平衡数据的识别具有现实意义，而数据稀缺和极端值可导致 AdaBoost 分类方法性能下降，基于此问题本文提出通过在弱分类器中侧重少数类样本赋予更大的权重，避免由于原训练集中的少数类数据量较少，导致训练得到的决策树规则没有代表性的缺点，迫使分类器更加关注少数类样本，提高少数类样本的分类准确率，从而能够很好地解决非平衡数据集分类问题，这样就迫使最终强分类器对少数类样本具有更高的准确率. 由于 AdaBoost 算法中指数误差界没有任何直接依赖分类，所以之后的文献[18-19]主要针对分类非对称（class-conditional）直接修改权重更新规则. 更新规则是在错误的结果约束极小化过程中，这些变化是真正影响的理论属性而 AdaBoost 演算本身无法保证. 针对非平衡数据特性提高分类代价敏感度，描述为

$$J(f)=E\left([y=1]\,\mathrm{e}^{-C_{\mathrm{P}}f(x_i)}+[y=-1]\,\mathrm{e}^{-C_{\mathrm{N}}f(x_i)}\right),$$

$$f(x)=\frac{1}{C_{\mathrm{P}}+C_{\mathrm{N}}}\log\frac{C_{\mathrm{P}}P(y=1\mid x)}{C_{\mathrm{N}}P(y=-1\mid x)}.$$

式中 C_{P} 和 C_{N} 表示对于正类和负类错误分类的代价.

为了清晰描述改进 AdaBoost 模型，给定 N 为训练集 (X,Y) 中个体数目，其中训练集每个 (x_i,y_i) 的 y_i 表示为

$$y_i=\begin{cases}1, & 1\leqslant i\leqslant m;\\ -1, & m<i\leqslant n.\end{cases}$$

针对交通离群数据的改进 AdaBoost 模型训练过程如下.

步骤 1　对原始训练集上的样本，给定每个分类初始分布为

$$D(i)=\begin{cases}\dfrac{1}{2(n-m)}, & 1\leqslant i\leqslant m;\\ \dfrac{1}{2m}, & m<i\leqslant n.\end{cases}$$

步骤 2　初始化循环轮数 $t=1$.

步骤 3　计算

$$T_{\mathrm{P}}=\sum_{i=1}^{m}D(i),$$

$$T_{\mathrm{N}}=\sum_{i=m+1}^{n}D(i).$$

步骤 4　初始化分类器变量 $f=1$.

步骤 5　在第 f 个弱分类器 $h_f(X)$ 中计算

$$D(i)=\begin{cases}\sum\limits_{i=1}^{m}D(i)\parallel y_i\neq h_f(x_i)\parallel;\\ \sum\limits_{i=m+1}^{n}D(i)\parallel y_i\neq h_f(x_i)\parallel.\end{cases}$$

步骤 6　计算满足等式的 $\alpha_{t,f}$，满足的假设为

$$2C_{\mathrm{P}}B\cosh(C_{\mathrm{P}}\alpha_{t,f})+2C_{\mathrm{N}}D\cosh(C_{\mathrm{N}}\alpha_{t,f})=C_1T_{\mathrm{P}}\mathrm{e}^{-C_{\mathrm{P}}\alpha_{t,f}}+C_2T_{\mathrm{N}}\mathrm{e}^{-C_{\mathrm{N}}\alpha_{t,f}}.$$

步骤 7　计算弱学习器的损失为

$$L_{t,f}=B(\mathrm{e}^{C_{\mathrm{P}}\alpha_{t,f}}-\mathrm{e}^{-C_{\mathrm{P}}\alpha_{t,f}})+T_{\mathrm{P}}\mathrm{e}^{-C_{\mathrm{P}}\alpha_{t,f}}+D(\mathrm{e}^{C_{\mathrm{N}}\alpha_{t,f}}-\mathrm{e}^{-C_{\mathrm{N}}\alpha_{t,f}})+T_{\mathrm{N}}\mathrm{e}^{-C_{\mathrm{N}}\alpha_{t,f}},$$

式中 C_{P} 和 C_{N} 为代价参数.

步骤 8　累计 $f=f+1$，若 $f\leqslant F$，重复步骤 5.

步骤 9　在本轮中比较得到最小损失的弱分类器 $(h_t(X),\alpha_t(X))$ 为 $\underset{f}{\operatorname{argmin}}[L_{t,f}]$.

步骤 10　更新 $D(i)$ 权重为

$$D(i)=\begin{cases}D(i)\mathrm{e}^{-C_{\mathrm{P}}\alpha_t h_t(X_i)}, & 1\leqslant i\leqslant m;\\ D(i)\mathrm{e}^{C_{\mathrm{N}}\alpha_t h_t(X_i)}, & m<i\leqslant n.\end{cases}$$

步骤 11　累计 $t=t+1$，若 $t\leqslant T$，重复步骤 3.

步骤 12　确定的分类器为

$$H(x)=\operatorname{sign}(f(x))=\operatorname{sign}\Big(\sum_{t=1}^{T}\alpha_t h_t(x)\Big),$$

式中 $h_t(x)$ 为弱分类器联合.

评估数据异常识别算法功能性能时，选用检测准确度 D_{acc}、误检率 R_{FP} 和误警率 R_{FN} 指标来衡量，

计算公式分别为

$$D_{acc} = \frac{N_{CN} + N_{CG}}{N_{CN} + N_{CG} + N_{EN} + N_{EG}}, \tag{4}$$

$$R_{FP} = \frac{N_{EN}}{N_{CG} + N_{EN}}, \tag{5}$$

$$R_{FN} = \frac{N_{EG}}{N_{EG} + N_{CN}}. \tag{6}$$

式中 N_{CN} 为检测出的交通离群样本数目; N_{EG} 为未检测出的交通离群样本数目; N_{CG} 为检测出的一般交通样本数目; N_{EN} 为未检测出的一般交通样本数目.

给定概率代价函数 F_{PC} 和标准期望代价 E_c 的定义,概率代价函数主要采用检测样本的先验概率、检测样本数和未检测出的样本数来联合定义,描述为

$$F_{PC} = \frac{p(+)N_{EG}}{p(+)N_{EG} + p(-)N_{CG}},$$

式中 $p(+)$ 和 $p(-)$ 为检测出交通离群样本和检测出的一般交通样本的先验概率.

标准期望代价 E_c 表示为

$$E_c = N_{CG} \cdot F_{PC} + N_{EG}.$$

3 实验结果与分析

3.1 实验数据采集

为了检验改进 AdaBoost 模型的实际应用性能,首先对提出的模型和相关经典算法在随概率代价函数变化下各个指标进行对比,指标包括检测准确度、误检率、误警率和标准期望代价等. 研究选取了山东高速公路检测数据集中 2014 年 11 月 5 日 13 个监测点的感应线圈数据、地磁数据和卡口数字化处理后数据进行检测器数据异常识别,采集数据集的特征描述见表 1.

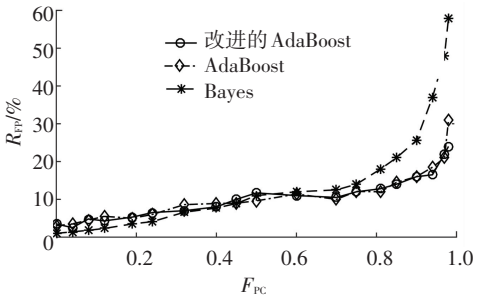
表 1 不平衡采集数据集的特征描述

Tab.1 Properties description of the non-equilibrium datasets

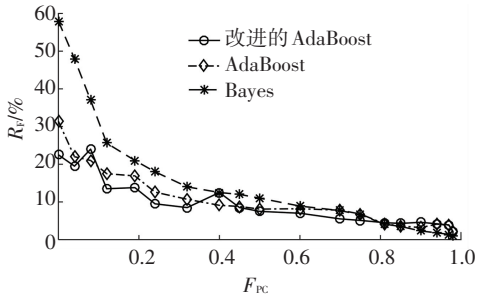
数据集	数据属性	样本数	离群样本数	正常样本数	原始不平衡率/%
济北站数据集	线圈数据	4 824	182	4 642	3.921
	地磁数据	6 399	443	5 956	7.452
	卡口数据	10 870	659	9 619	6.851
高唐站数据集	线圈数据	5 013	198	4 762	3.949
	地磁数据	6 512	531	5 972	8.154
	卡口数据	11 194	625	10 478	5.583

3.2 实验数据分析

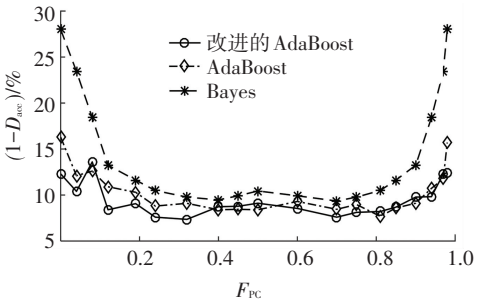
本实验针对检测准确度、误检率、误警率和标准期望代价指标分析不同方法在高速公路检测数据集的异常识别效果. 比较济北站数据集和高塘站数据集的实验结果,图 2 为基于高速公路检测数据集(济北站)在改进 AdaBoost 方法、AdaBoost 方法和 Bayes 方法中检测指标对比图,图 3 为基于高速公路检测数据集(高唐站)在改进 AdaBoost 方法、AdaBoost 方法和 Bayes 方法中检测指标对比图. 图 2、3 中均以概率代价函数 F_{PC} 为横坐标比较各检测指标,其中图 2(a)、3(a)表示误检率 R_{FP} 随概率代价函数的变化曲线,图 2(b)、3(b)表示误警率 R_{FN} 随概率代价函数的变化曲线,图 2(c)、3(c)表示分类错误率 $1-D_{acc}$ 随概率代价函数的变化曲线,图 2(d)、3(d)表示标准期望代价 E_c 随概率代价函数的变化曲线. 本实验使用不同方式的训练构造决策规则,对相同数据集分类中,不同算法效果相差明显;而对于不通数据集综合,相同算法的特征可以延续. 首先对相同数据集分析,在图 2 对比 Bayes 方法、AdaBoost 方法和改进的 AdaBoost 方法体现的 R_{FP}、R_{FN} 指标曲线中, AdaBoost 方法和改进的 AdaBoost 方法性能接近并明显优于 Bayes 方法,而 3 种方法在图 2(d)的 E_c 中差别不大. 同理在图 3 中对比 Bayes 方法、AdaBoost 方法和改进的 AdaBoost 方法体现的 R_{FP}、R_{FN} 指标曲线中,在图 3(b)、3(c)中改进的 AdaBoost 方法略优于 AdaBoost 方法,并明显优于 Bayes 方法,而 3 种方法在图 3(d)的 E_c 中差别不明显. 另一方面综合两个组数据指标曲线可以发现,对图 2 和图 3 实例验证的检测数据集,随着数据集规模的加大,改进的 AdaBoost 方法在 $1-D_{acc}$、R_{FP} 两个指标上优于 AdaBoost 方法,前者比后者平均低 5.547% 和 6.792%. 其原因是分类样本例的比例不均衡,AdaBoost 侧重考虑非均衡的数据特性,被错误分类的离群数据降低了检出率,而改进的 AdaBoost 算法整体的误检率降低,充分体现了算法中引入代价参数针对性地提高了检测准确性. 另外研究通过 ROC 曲线表征比较各类方法的检测性能,如图 4 采用 ROC 曲线全面评价各类识别方法的优劣,以误检率 R_{FP} 为横轴, 以检出率($1-R_{FN}$)为纵轴, 从曲线变化可以看出改进的 AdaBoost 方法明显优于其他算法,ROC 曲线比较更为直观全面.



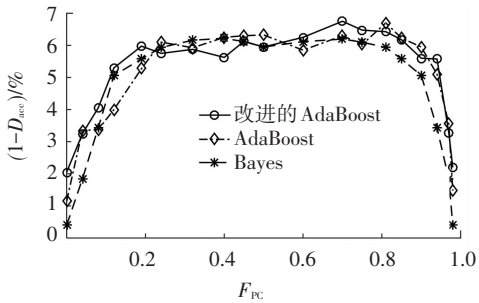
(a)误检率与概率代价关系曲线



(b)误警率与概率代价关系曲线



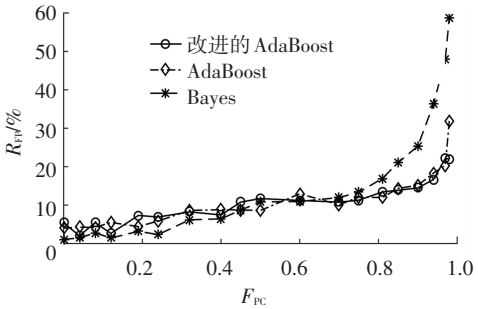
(c)分类错误率与概率代价关系曲线



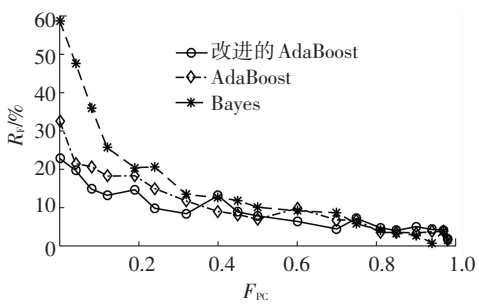
(d)标准期望代价与概率代价关系曲线

图 2　基于高速公路检测数据集(济北站)检测指标对比

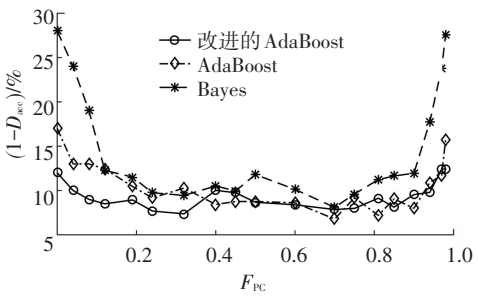
Fig.2　Detection index comparisons of highway detection dataset (Jibei Station)



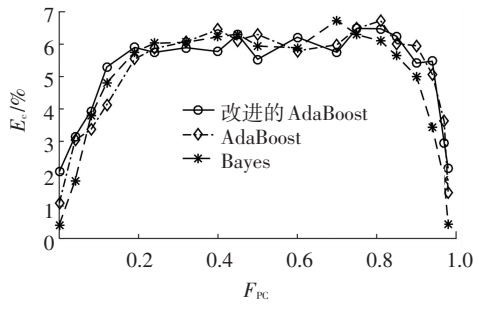
(a)误检率与概率代价关系曲线



(b)误警率与概率代价关系曲线



(c)分类错误率与概率代价关系曲线



(d)标准期望代价与概率代价关系曲线

图 3　基于高速公路检测数据集(高唐站)检测指标对比

Fig.3　Detection index comparisons of highway detection dataset (Gaotang Station)

4　结　论

1)针对交通检测数据中非平衡的离群数据剥离数据集的问题,提出了具有针对性的交通检测数据异常识别模型. 该改进模型保留原始 AdaBoost 算法中训练加权优势;另外模型引入代价敏感方法来强化非平衡特性,改进 AdaBoost 的决策过程. 模型避免了非平衡检测数据导致的分类性能下降问题.

2)实例数据验证模型的 D_{acc}、R_{FP}、R_{FN} 和 E_c 等指标,并利用 ROC 曲线全面评价提出模型的优劣,实验结果表明改进的 AdaBoost 方法在 $1-D_{acc}$、R_{FP} 两个指标上优于 AdaBoost 方法,前者比后者平均低

5.547%和 6.792%,采用改进的 AdaBoost 筛选交通检测样本可提供一个可靠度更高的分类筛选结果,有效调整了非平衡数据导致的分类误差.

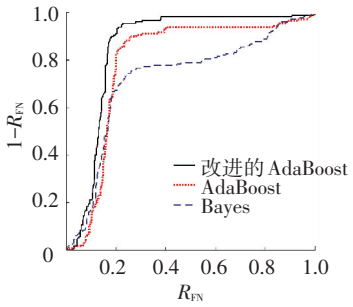


图 4 改进的 AdaBoost 与其他算法的 ROC 曲线图

Fig. 4 Comparison of ROC curves between the improved AdaBoostalgorithm and other algorithms

3)比较改进算法与其他相关模型的适用性,采用分类指标评价本方法的优劣. 实验表明针对交通检测数据集的离群数据提出的改进 AdaBoost 方法可以减少测试误检率. 然而本算法以非均衡的高速公路交通数据样本集为研究的出发点,所以本方法对检测数据集有一定的限制,进一步的研究重点将集中在改善方法的局限性上面.

参考文献

[1] 黄艳国, 许伦辉, 邝先验.基于模糊 C 均值聚类的城市道路交通状态判别[J]. 重庆交通大学学报(自然科学版),2015,34(2),102
HUANG Yanguo, XU Lunhui, KUANG Xianyan. Urban road traffic state identification based on fuzzy C-mean clustering[J]. Journal of Chongqing Jiaotong University (Natural Science), 2015, 34(2): 102

[2] 吴志勇, 丁香乾, 鞠传香.一种基于深度学习的离散化交通状态判别方法[J].交通运输工程与信息学报, 2017,17(5): 129
WU Zhiyong, DING Xiangqian, JU Chuanxiang. A method of discrete traffic state identification based on deep learning[J]. Journal of Transportation Engineering and Information, 2017, 17(5): 129. DOI: 10.16097/j.cnki.1009-6744.2017.05.018

[3] 邴其春, 龚勃文, 杨兆升, 等. 一种组合核相关向量机的短时交通流局域预测方法[J]. 哈尔滨工业大学学报, 2017, 49(3):144
BING Qichun, GONG Bowen, YANG Zhaosheng, et al. A short-term traffic flow local prediction method of combined kernel function relevance vector machine[J]. Journal of Harbin Institute of Technology, 2017, 49(3): 144

[4] 邢雪, 于德新, 田秀娟,等. 基于数据挖掘的高速公路行程时间预测[J]. 华中科技大学学报(自然科学版),2016,44(8): 36
XING Xue, YU Dexin, TIAN Xiujuan, et al. Freeway travel time prediction based on clustering method with data mining[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2016, 44(8): 36. DOI: 10.13245/j.hust.160808

[5] 陈淑燕, 王炜, 瞿高峰.服务于智能交通系统的离群交通数据识别[J].东南大学学报(自然科学版), 2008, 38(4): 723
CHEN Shuyan, WANG Wei, QU Gaofeng. Outlier detection in traffic data sets serving for intelligent transportation system[J]. Journal of Southeast University (Natural Science Edition), 2008, 38(4): 723

[6] XING Xue, YU Dexin, ZHANG Wei. Data calibration based on multisensor using classification analysis: a random forests approach [J]. Mathematical Problems in Engineering, 2015, 2015 (708467): 1. DOI:10.1155/2015/708467

[7] NAM D H, DREW D R. Traffic dynamics: method for estimating freeway travel times in real time from flow measurements[J]. Journal of Transportation Engineering, 1996, 3: 185

[8] 陈淑艳,王炜,李文勇.实时交通数据的噪声识别和消噪方法[J]. 东南大学学报(自然科学版),2006,36(2):322
CHEN Shuyan, WANG Wei, LI Wenyong. Noise recognition and noise reduction of real time traffic data[J]. Journal of Southeast University (Natural Science Edition), 2006, 36(2): 322

[9] VANAJAKSHI L, RILETT L R. Loop detector data diagnostics based on conservation of vehicles principle[J]. Transportation Research Record, 2004, 1870: 162. DOI: 10.3141 / 1870-21

[10] SMITH B L, SCHERER W L, CONKLIN J H. Exploring imputation techniques for missing data in transportation management systems [J]. Transportation Research Record: Journal of the Transportation Research Board, 2003, 1836:132. DOI: 10.3141/1836-17

[11] 蒲世林, 李瑞敏, 史其信. 基于粗糙集-模糊识别技术的交通流状态识别算法研究[J].武汉理工大学学报(交通科学与工程版),2010, 34(6):154
PU Shilin, LI Ruimin, SHI Qixin. Study on auto-identification algorithm of traffic flow state based on rough set and fuzzy theory[J]. Journal of Wuhan University of Technology (Transportation Science & Engineering), 2010, 34(6): 154

[12] LIN Dayang, STEVEN B, VARUNRAJ V, et al. Reliability assessment for traffic data[J]. Journal of the Chinese Institute of Engineers, 2012(35): 285. DOI: 10.1080/02533839.2012. 655466

[13] MORGUL E F, OZBAY K, IYER S, et al. Commercial vehicle travel time estimation in urban networks using GPS data from multiple sources [C]// Transportation Research Board 92nd Annual Meeting. Washington DC: Transportation Research Board, 2013: 13

[14] 郭敏,蓝金辉,李娟娟,等. 基于灰色残差 GM(1,N)模型的交通流数据恢复算法[J].交通运输系统工程与信息,2012, 12(1): 42
GUO Min, LAN Jinhui, LI Juanjuan, et al. Traffic flow data recovery algorithm based on gray residual GM(1,N) model[J]. Journal of Transportation Engineering and Information, 2012, 12(1): 42. DOI: 10.16097/j.cnki.1009-6744.2012.01.019

[15] 章渺. 高速公路基本路段实时交通状态识别方法[D].西安:长安大学,2017
ZHANG Miao. Research on method of highway basic section real-time traffic status identification[D]. Xi'an: Chang'an University, 2017

[16] 张亮, 李智星, 王进.基于动态权重的 AdaBoost 算法研究[J]. 计算机应用研究,2017(11) :1
ZHANG Liang, LI Zhixing, WANG Jin. Research on dynamic weights based AdaBoost[J]. Application Research of Computers, 2017(11): 1. DOI: 10.3969/j.issn.1001-3695.2017.11.007

[17] 郭乔进, 李立斌, 李宁.一种用于不平衡数据分类的改进 AdaBoost 算法[J].计算机工程与应用,2008, 44 (21) :217
GUO Qiaojin, LI Libin, LI Ning.Novel modified AdaBoost algorithm for imbalanced data classification[J]. Computer Engineering and Applications, 2008, 44(21): 217

[18] MASNADI S H, VASCONCELOS N. Cost-sensitive boosting[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 33: 294. DOI: 10.1109/tpami.2010.71

[19] SUN Y, KAMEL M, WONG A, et al. Cost-sensitive boosting for classification of imbalance data[J]. Pattern Recognition, 2007, 40: 3358. DOI: 10.1016/j.patcog.2007.04.009

(编辑 魏希柱)