文章编号:1671-8879(2013)06-0084-04

交通事件检测的加权支持向量机算法

王武功,马荣国
(长安大学 公路学院,陕西 西安 710064)

摘 要:针对交通事件数据样本少,检测效率低的问题,将加权支持向量机引入到交通事件检测中,采用样本重要度加权法提高算法的检测率,根据识别误差确定样本重要度权值,建立了交通事件检测的样本重要度加权法支持向量机算法,最后应用实测数据对标准支持向量机算法、样本重要度加权法、样本数目加权法 3 种算法的检测效果进行测试。研究结果表明:样本数目加权法算法能够根据样本的好坏自适应确定样本重要度权值,提高了算法的鲁棒性;当负正样本比率减少时,3 种算法的检测效果均变差,而对于同样的样本,标准支持向量机的检测率最低,样本重要度加权法的效果最好,加权算法的选择要依据样本的数量、分布不平衡以及识别目标而定;在交通事件检测中,为了提高检测率,选择样本重要度加权效果最好,在不同的样本不平衡率下,检测效果是不同的,不平衡率越严重,检测效果越差。

关键词:交通工程;交通事件检测;不平衡样本;加权支持向量机;样本重要度
中图分类号:U491 **文献标志码**:A

Weighed support vector machine for traffic incident detection

WANG Wu-gong, MA Rong-guo
(School of Highway, Chang'an University, Xi'an 710064, Shaanxi, China)

Abstract: A weighed support vector machine(SVM), based on the importance of the samples, was proposed to solve the problem of low detection caused by imablanced samples. The weight of the samples was determined by their discriminate errors. The measured data was used to test the performance of the proposed algorithm. The basic SVM, the weighed SVM based on the sample number and the weighed SVM based on sample importance were tested under different imbalanced samples. The results show that the algorithm can determine the weight of the sample according the sample, which can improve the robust of the algorithm; the bigger the imbalance of the samples, the lower the detection ratio of all three algorithms. With the same samples, the ratio of detection of the basic SVM is the lowest and the weighed SVM based on sample importance is the highest; in traffic incident detecting, in order to improve detecting rates, the weighted SVM based on sample important is the best choice; under different unbalanced sample rates, detection effects are different; the higher the unbalanced rate, the worse the detection effect. 1 tab, 1 fig, 9 refs.

Key words: traffic engineering; automatic incident detection; imbalanced sample; weighed support vector machine; importance

收稿日期:2012-12-10
基金项目:国家自然科学基金项目(51208053)
作者简介:王武功(1972-),男,陕西华县人,工学博士研究生,E-mail:wwg@chd.edu.cn。

0 引　言

交通事件检测一直是智能交通研究的热点之一。现有的交通事件检测算法大致可分为 3 类：模式识别法、统计预测法、智能算法。模式识别法、统计预测法都需要根据经验确定参数，算法的移植性较差。当前广泛使用的算法是智能算法，主要有神经网络法、遗传算法、支持向量机法、小波算法等[1-3]。神经网络法具有良好的非线性逼近能力和自适应自学习的功能，但是存在需要确定网络结构，要求的训练样本量大等问题[3]。支持向量机(SVM)形式上类似于多层前向网络，对小样本、非线性、高维问题具有良好的泛化能力，而且，可自动解决网络结构问题，因此，近年来，SVM 已成为机器学习领域的研究热点，已成功应用于文本分类、手写识别、图像分类、生物信息学等领域。同时，支持向量机算法在事件检测中的应用也得到广泛研究。有关学者通过对核函数形式及不同的输入特征对检测性能的影响的研究，发现高斯径向基函数具有最高的识别率，输入特征选用占有率与流率组合的检测性能最好[4-5]。Chen 等针对支持向量机算法参数选择的问题，将 BOOSTS 分类器与支持向量机相结合，采用多个支持向量机集成算法进行交通事件检测[6]。以上的算法，都是基于标准支持向量机进行的。标准支持向量机是在假设类分布平衡，样本数据大致相当的前提下使用时，具有较高的精度，但是，对于不平衡数据，标准的支持向量机的性能大大下降。而针对样本不平衡问题，采用惩罚因子来平衡两类样本的数目差别[7]，将模糊数学与支持向量机相结合，根据不同输入样本对分类贡献的不同，赋以相应的隶属度，从而提高分类性能[8]。且有学者将加权支持向量机成功应用于文本分类中[9]。

在交通事件检测当中，由于事件样本(负样本)的收集相对比较困难，所以，算法学习使用的样本通常都存在严重不平衡。结合已有的研究成果，为了提高交通事件的正确检测率，本文在文献[7]的基础上提出采用加权支持向量机进行事件检测。

加权支持向量机根据样本的数目差别或样本的重要性程度不同，对不同样本采用不同的加权因子，提高对少数样本的识别能力。基于样本数目差别的权值法仅考虑样本数量的不平衡。而样本不平衡的情况，除了数量不平衡，还存在样本分布不平衡。交通流数据就属于这一类。正常运行的交通流参数在随时间变化，可以是稀疏流、拥挤流、阻塞流。交通事件在各种交通流中均可能发生，当交通流是稀疏流或过饱和流时，事件对交通流参数影响不大，难以检测，而误认为是正常交通流。为了提高算法的检测率，降低误报率，就需要加强对难以识别的事件样本加强学习。为此，本文采用基于样本重要性的加权方法建立事件检测的非线性模型，最后，通过仿真试验证明了所提出算法的可行性及有效性。

1 算法介绍

标准的支持向量机是在假设类分布平衡，样本数据大致相当的前提下使用的，具有较高精度，然而对于不平衡数据，标准的支持向量机的性能大大下降。加权支持向量机，根据样本的数目差别或样本的重要性程度不同，通过对不同样本采用不同的加权因子，以解决样本数目或样本的重要性程度的严重不平衡。

1.1 加权支持向量机

对于一组带有类别标记的训练样本集$\{x_i\}$，$x_i\in R^i$，$y_i\in\{+1,-1\}$，$i=1,\cdots,l$，若超平面 $\boldsymbol{wx}+\boldsymbol{b}=0$($\boldsymbol{w}$ 为超平面的法向量；$\boldsymbol{x}$ 为自变量向量；$\boldsymbol{b}$ 为超平面的截距向量)能将样本正确分为两类，则最佳超平面应使两类样本到超平面最小距离之和最大。对于不同的样本，设加权权值为 s_i，则加权支持向量机的优化问题可描述为

$$\min\left(\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}s_i\varepsilon_i\right)$$

$$\text{s.t.}\quad y_i[(\boldsymbol{w}w_i)+\boldsymbol{b}]\geqslant 1-\varepsilon_i\quad(i=1,2,\cdots,l)$$

$$\varepsilon_i\geqslant 0\quad(i=1,2,\cdots,l)$$

式中：y_i 为目标值；ε_i 为误差项；C 为惩罚系数；w_i 为 $y_it(+1,-1)$。

利用 Lagrange 乘子法可以把上述求解最佳超平面的问题转化为对偶问题，即

$$\max\sum_{i=1}^{l}\alpha_i\alpha_jy_iy_jk(x_i,\ x_j)$$

$$\text{s.t.}\quad\sum_{i=1}^{l}\alpha_iy_i=0\quad 0\leqslant\alpha_i\leqslant Cs_i\quad(i=1,2,\cdots,l)$$

相应的决策函数变为

$$f(x)=\text{sign}(\sum_{i=1}^{l}\alpha_iy_ik(x_i,\boldsymbol{x})+\boldsymbol{b})$$

式中：$k(x_i,\boldsymbol{x})$为核函数；α_i 为 Lagrangian 乘子

核函数的选取应使其为特征空间的一个内积。常用的核函数为

线性核函数：$k(x_i,\boldsymbol{x})=x_i\cdot\boldsymbol{x}$

高斯核函数:$k(\boldsymbol{x},x_i)=\mathrm{e}^{-\gamma_i\|\boldsymbol{x}-x_i\|^2}$

Sigmoid 核函数:$k(x,x_i)=1/[1+\mathrm{e}^{(\beta(\boldsymbol{x}\cdot x_i)+b)}]$

式中:β 为函数的参数。

1.2 权值确定方法

1.2.1 样本数目的平衡加权

为了消除样本数目的不平衡对识别精度的影响,对于同类样本采用相同的权值,设正例的类别权值为 s_+,样本数目为 l_+,反例的类别权值为 s_-,样本数目为 l_-,为了使 2 类得到平衡的误差率,s_+ 与 s_- 的比例关系应为:$\dfrac{s_+}{s_-}=\dfrac{l_-}{l_+}$。

1.2.2 样本重要性的平衡加权

由于样本分布的不均衡,导致不同样本对分类的贡献大小不同,使得某些样本的识别难度加大,为此,文献[6]提出基于样本重要度的加权法,以解决分布不平衡问题。基于样本重要度的加权法的思想是:样本的识别误差越大,样本识别越困难,应加大其权重。公式为

$$s_{i+1}=s_i+\frac{\alpha(y_i-\hat{y}_i)}{\sum_{i=1}^{l}y_i-\hat{y}_i}$$

样本误差为

$$e_i=y_i-\hat{y}_i$$

式中:α 为比例系数;y_i 为实际值;$\hat{y}_i$ 为标定值;s_i 为第 i 个样本的权重;e_i 为标本误差。

交通事件检测的主要依据是事件发生后,交通流三参数(交通量、车流速度、道路占有率)出现急剧变化,当参数的变化超过某一数值,则认为有事件发生。参数的变化与事件的严重性有关,事件越严重,变化量越大,反之,变化量就小。

目前,事件检测算法的主要问题是检测率较低。原因在于交通流的不稳定,正常的交通流可表现为稀疏流、拥挤流、阻塞流等,可以是不同交通流的交替变化,因而,正常运行的交通流参数也在随时间变化,容易与交通事件流相混淆。所以,部分交通事件对交通流参数影响不大,难以检测,而误认为是正常交通流。为了提高算法的检测率,降低误报率,就需要加强对难以识别的事件样本加强学习。具体到算法中,就是提高对困难样本的权重系数。本文采用的加权 SVM(支持向量机)算法步骤为:

(1)初始化样本权值,所有样本权值相同,$s_i=\dfrac{1}{l}$,应用 SVM 进行学习;

(2)计算样本误差 $e_i=y_i-\hat{y}_i$,修改 $s_{i+1}=s_i+\dfrac{\alpha(y_i-\hat{y}_i)}{\sum_{i=1}^{l}y_i-\hat{y}_i}$,重新学习。直到 $e_i<\delta$,停止学习。δ 为很小的正数。

2 试验结果及分析

2.1 试验环境

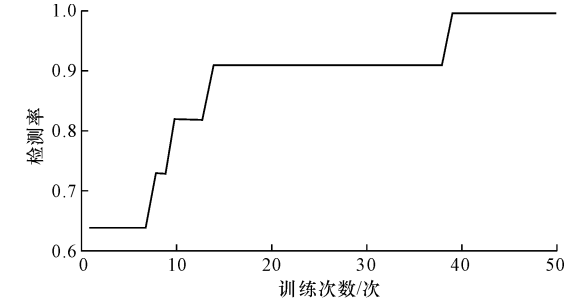
交通事件检测的依据是事件发生时,描述上下游交通状态的参数将发生急剧变化,上游交通的交通密度变大,平均速度变小,而下游交通的交通密度变小,平均速度变大。所以本文选择描述交通流特征的参数为属性变量,即选择上下游交通的占有率、平均速度、交通量为属性变量,即输入信号选取 $t-2$、$t-1$、t 时刻交通密度、平均速度、交通量。

试验数据使用的是 I-880 交通流数据,每 30 s 记录一次交通流数据,选取其中的 400 个样本,前 200 个为学习样本,其中交通事件样本数分别是 26 和 7 个,后 200 个为测试样本,交通事件样本数为 18 个。

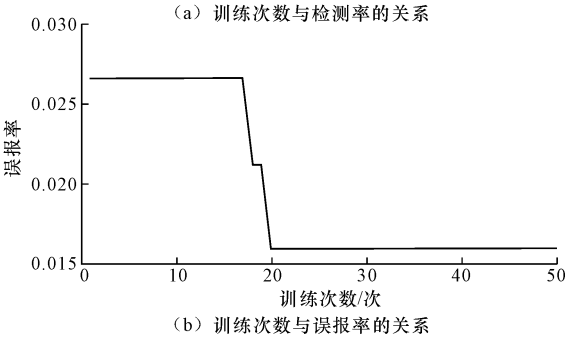
2.2 试验结果及分析

2.2.1 确定样本重要度权值

根据 1.2 节所介绍的方法进行学习,得到样本重要度权值。训练次数与检测结果之间的关系如图 1 所示,可见,随着学习次数的增加,检测率逐渐增大,误报率逐渐变小。



(a) 训练次数与检测率的关系



(b) 训练次数与误报率的关系

图 1 训练次数与检测性能之间的关系

Fig. 1 Relation between training times and performance of detection

2.2.2　检测结果比较

为了验证本文提出方法的可行性，分别用标准 SVM、样本数目加权 SVM 及样本重要度加权 SVM 进行检测，惩罚系数 $C=8$，核函数取高斯核函数 $k(x,x_i)=(-\gamma\|x-x_i\|^2)^{\epsilon}$，参数 $\gamma_i=2$。检测结果如表 1 所示。

表 1　不同算法的检测结果

Tab. 1　Detection results of different algorithms

负正样本比率/%	方法	检测率/%	误报率/%	平均检测时间/s
13	标准 SVM	84.6	0	3.73
	样本数目加权 SVM	92.3	0.23	3.02
	样本重要度加权 SVM	100	2.07	1.26
3.5	标准 SVM	81.8	0	4.12
	样本数目加权 SVM	90.9	0.45	3.98
	样本重要度加权 SVM	99.8	3.70	1.57

从表 1 可以看出：①对于同质量的样本，3 种算法的检测结果明显不同，标准 SVM 的检测效果最差，加权 SVM 的检测结果要优于标准 SVM，样本重要度加权 SVM 检测结果最好，检测率达到 100%；②加权 SVM 在提高检测率的同时，误报率也变大，说明 2 种加权算法都是以牺牲整体识别误差为代价来提高少数样本的识别能力，加权算法的选择要依据样本的数量、分布不平衡以及识别目标而定，在交通事件检测中，为了提高检测率，选择样本重要度加权效果最好；③在不同的样本不平衡率下，检测效果是不同的，不平衡率越严重，检测效果就差。

3　结　语

(1)根据样本数据中负正样本的数目差别及具体的样本数据分布来平衡两类错分率，甚至减少错分率。

(2)根据学习误差确定出不同类惩罚因子的比例，减少盲目性适合于负正样本数量差异较大的不平衡数据分类。

(3)本文中提出的没有考虑样本个数、样本数据质量及所采用的支持向量机算法对支持向量个数的影响，而实际上，由于支持向量机算法适用于小样本的情况，在样本数不是很大的情况下，支持向量数目对算法的执行影响差别不大。

参考文献：

References：

[1] 樊小红，荆便顺. 基于遗传算法的交通事件检测[J]. 长安大学学报：自然科学版，2005，25(4)：70-72.
FAN Xiao-Hong，JING Bian-Shun. Automatic incident detection based on algorithm[J]. Journal of Chang'an University：Natural Science Edition，2005，25(4)：70-72. (in Chinese)

[2] Jeong Y S，Manoel C N，Myong K J. A wavelet-based freeway incident detection algorithm with adapting threshold parameters [J]. Transportation Research Part C，2011(19)：1-19.

[3] Mahmoud A，Stephan O. Automatic incident detection in vanets：a baysian approach[C]//VTC. Proceedings of the 69th IEEE Vehicular Technology Conference. Barcelon：VTC Spring，2009：1-5.

[4] 陈　斌. 基于支持向量机的高速公路意外事件检测模型[J]. 中国公路学报，2006，19(6)：107-112.
CHEN Bin. Freeway accident detection model based on support vector machine [J]. China Journal of Highway and Transport，2006，19(6)：107-112. (in Chinese)

[5] 陈维荣，关　佩，邹月娴. 基于 SVM 的交通事件检测技术[J]. 西南交通大学学报，2011，46(1)：63-67.
CHEN Wei-rong，GUAN Pei，ZOU Yue-xian. Automatic incident detection technology based on SVM [J]. Journal of Southwest Jiaotong University，2011，46(1)：63-67. (in Chinese)

[6] Chen S Y，Wang W，Henkvan Z. Construct support vector machine ensemble to detect traffic incident[J]. Expert Systems With Applications，2009，36(8)：10976-10986.

[7] 刘万里，刘三阳，薛贞霞. 不平衡支持向量机的平衡方法[J]. 模式识别与人工智能，2008，21(2)：136-141.
LIU Wan-li，LIU San-yang，XUE Zhen-xia. Balanced method for imbalanced support vector machines. [J]. Pattern Recognition and Artificial Intellgece，2008，21(2)：136-141. (in Chinese)

[8] Lin C F，Wang S D. Fuzzy support vector machines [J]. IEEE：Trans on Neural Nerworks，2002，13(2)：464-471.

[9] 赵　晖，荣莉莉. 支持向量机组合分类及其在文本分类中的应用[J]. 小型微型机计算机系统 2005，26(10)：1816-1820.
ZHAO Hui，RONG Li-li. Classification algorithm based on combined support vector machine and its application in text categarization [J]. Mini-micro Systems，2005，26(10)：1816-1820. (in Chinese)