基于支持向量数据描述的高速公路事件检测

赵晓芳,刘智勇
(五邑大学 信息学院,广东 江门 529020)

摘 要:提出了一种基于支持向量数据描述(SVDD)算法的快速事件检测方法。该算法把有事件样本和无事件样本分别用全体样本优化的SVDD算法进行优化。但每次只对那些对超球体边界有影响的数据进行优化。该方法既保留了全体样本优化SVDD算法的优点,又达到加快训练速度的目的。采用I-880数据库中实际交通的历史数据进行实验,并与全体样本优化SVDD实验结果相比较。实验证明该分类方法能够获得较高检测率和较低的误报率,且需要较短的训练时间,表明了所给方法的有效性和可行性。

关键词:SVDD;事件检测;分类
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)12-0248-03

Freeway Traffic Incident Detection Based on Support Vector Data Description

ZHAO Xiao-fang, LIU Zhi-yong
(School of Information, Wuyi University, Jiangmen 529020, China)

Abstract: The problem of freeway incident detection is researched by using support vector data description (SVDD). This method optimiza the incident samples and free-flow samples separately, but it only optimizes the dates which have influence to hyper-sphere one time. So it not only holds the advantages of SVDD, and but also gets obtained high training speed. Using the I-880 database which has actual traffic history data to carry on the experiment, and compares with the classic SVDD, confirmed this classified method can obtain a higher detection rate and lower false alarm rate. So the results of simulation experiments show that the proposed method is effective and feasible.

Key words: support vector data description; incident detection; classification

0 引 言

随着经济的快速发展,车辆猛增,高速公路出现了事故频频、交通拥挤、交通堵塞、环境污染等一系列问题,给人们的生活带来了巨大的不便。日前,人们对交通事件自动检测系统已作了深入的研究,有关交通事件自动检测算法研究已取得一些进展。事件算法发展至今,可大致分为模式识别算法、统计预测算法、突变理论算法和神经网络算法等[1~3]。

文中提出了一种基于支持向量数据描述(SVDD)的高速公路事件检测方法。SVDD是Tax等人提出的解决单值分类问题的支持向量机方法[4]。与有监督的SVM不同在于,SVM需要两类样本来寻找区分两类数据的最优超平面,而SVDD只需要一类样本即可,其目标在于寻找包含该类样本的最优超球体,将该样本与其他类样本区分开来。但SVDD在数据量比较大时,核矩阵的比较大,优化速度很慢。在高速公路事件检测应用中,把有事件样本和无事件样本分别用SVDD算法进行优化,但每次只对那些对超球体边界有影响的数据进行优化,该方法既保留了SVDD算法的优点,又达到加快训练速度的目的。

收稿日期:2008-03-17
基金项目:广东省自然科学基金项目(06029813);广东省高等学校自然科学重点研究项目(05z025)
作者简介:赵晓芳(1982-),女,河南襄城人,硕士研究生,主要研究方向为智能交通控制、模式识别与智能系统;刘智勇,博士,教授,主要研究方向为智能交通控制、模式识别与智能系统。

1 算法介绍

1.1 SVDD算法简介

SVDD的主要思想是通过计算高维空间中包含目标类映射数据的最小超球体边界来对该组数据进行描述,从而实现将目标类与非目标类分开。在数学上SVDD是一个寻找最小超球体使之尽可能含一类目标数据的优化问题。设样本集为:$\{x_i, i = 1,\cdots,l\}$,x_i

$\in R^d$。设法找一个以 a 为中心,以 R 为半径的能够包含所有样本点的最小球体。优化问题为:

$$\min F(R,a,\xi_i) = R^2 + C\sum_{i=1}^{l}\xi_i \tag{1}$$

约束为

$$(\phi(x_i)-a)(\phi(x_i)-a)^{\mathrm{T}} \leqslant R^2+\xi_i, i=1,\cdots,l \tag{2}$$

$$\xi_i \geqslant 0, i=1,\cdots,l \tag{3}$$

将上面的优化问题(式(1))变成其对偶形式为:

$$\max\sum_{i=1}^{l}\alpha_i K(x_i,x_i) - \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i,x_j) \tag{4}$$

约束为

$$\sum_{i=1}^{l}\alpha_i = 1 \tag{5}$$

$$0 \leqslant \alpha_i \leqslant C, i=1,\cdots,l \tag{6}$$

解优化问题可以得到 α_i 的值,不为零的 α_i 所对应的样本被称为支持向量。根据 KKT 条件,对应于 $0<\alpha_i<C$ 的样本满足

$$R^2 - (K(x_i,x_i) - 2\sum_{j=1}^{l}\alpha_j K(x_j,x_i) + a^2) = 0 \tag{7}$$

其中,$a=\sum_{i=1}^{l}\alpha_i\phi(x_i)$。因此,用任意一个支持向量根据式(7)可求出 R 的值。对于需要检测的样本 z,设

$$f(z) = (\phi(z)-a)(\phi(z)-a)^{\mathrm{T}} = K(z,z) - 2\sum_{i=1}^{l}\alpha_i K(z,x_i) + \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i,x_j) \tag{8}$$

若 $f(z)\leqslant R^2$,则 z 被判决为目标类,否则 z 被判决为非目标类。$K(x_i,x_j)$ 一般取径向基核函数,即 $K(x_i,x_j)=\exp(-\|x_i-x_j\|^2/\sigma^2)$

1.2 快速算法的步骤

数据量比较大时,SVDD 算法要计算的核函数矩阵很大,所以优化速度很慢,在不降低检测率的情况了,用快速的优化方法来加快训练速度[5]。该方法每次优化所求的核矩阵都很小,所以优化速度很快。在训练过程中,随着训练样本的增多,后来的样本大多只需要做简单的数学判断而不参与优化,所以样本越多,这种算法的优势越能体现出来,后面的试验可以证明。

具体的步骤如下:

1)选择部分样本按 SVDD 方法进行优化,得到支持向量 $\boldsymbol{X}_N$ 和球心 a_N 以及半径 R_N。

2) 读入新样本 X_{N+1},计算

$$\lambda_{N+1} = \frac{d(x_{N+1},a_N)}{R_N} \tag{9}$$

$$d(x_i,x_j) = \sqrt{k(x_i,x_i) - 2k(x_i,x_j) + k(x_j,x_j)} \tag{10}$$

3) 若 $\lambda_{N+1}\geqslant C_0$,则该新样本对超球体边界有影响,用 $\boldsymbol{X}_N$ 和当前新样本 x_{N+1} 组成新的训练样本进行 1) 操作;若 $\lambda_{N+1}<C_0$,则新样本对边界无影响,读入新样本,转向 2) 操作。其中 $C_0\in(0,1)$。

当数据量很大时,也可在读入 m 组样本后再逐个判断对边界的影响,找到对边界有影响的 m_0 个样本和支持向量组成新的训练样本,进行优化处理。

1.3 算法比较

文献[6]提出了基于支持向量数据描述的数据约简的模式识别方法,采用网上数据[7]进行试验,该方法首先对数据进行约简,再进行优化,如果约简后的数据量仍然很大就也很难达到预期的目的。为了便于比较,同样采用文献[7]中的同一种数据 Banana Data Set,选取前十组训练和测试数据分别进行分类实验获得识别率和操作时间,采用 MATLAB 编程,时间用 cputime 计时函数进行统计,直接采用标准二次规划函数 quadprog 进行优化。在此 C 值都取 1,$\sigma^2=0.3$ 时,采用 SVDD 算法得到的最好效果的识别率和操作时间分别为 87.69% 和 25.3s;采用文献[6]中的方法,数据约简的过程每组平均需要的优化时间为 544.3s,进行分类试验的最好结果的识别率和操作时间为 88.33% 和 20.55s;采用文中快速算法,得到的最好结果为 87.86% 和 24.03s。而采用 SVM 对前十组 Banana Data Set 原始数据进行操作的平均每组却需要 666.5s,但其识别率为 88.32%。

比较上述结果可知,先进行数据约简再优化的方法和 SVM 的操作时间都比较长,但识别率相对比较高,文中所述的快速算法和文献[6]的方法达到了相近的检测率,但所用的时间不足其 1/20。其每次需要优化的数据个数只是 20 多个,所以速度很快,数据量越大,这种效果越明显。采用 SVDD 所用的时间和检测率与文中所述方法所得的结果没有很大差别,下面的试验可以证明当数据量增大时,这种差别是很大的。

2 在高速公路事件检测中的应用

2.1 输入参数的选择

数据来自高速公路路段 I-880 实地线圈数据集和事件数据集[8],为了使该算法具有较强的实用性,对流量采用像素级的参数融合,使其能够体现交通流从拥挤开始到结束的特征。所以这里选取以下交通参数的归一化值作为输入:占有率、速度、占有率/速度和占有率/流量。

2.2 事件自动检测算法的仿真与实现

选取某检测点的 659 组训练样本和 70 组测试样本,其中测试样本中 28 组为有事件发生时的数据,用检测率和误报率来评价事件检测算法,用 MATLAB 中的 tic 和 toc 来记录整个过程需要的时间。取 $C=$

0.9，C_0 =0.95，初始的正负样本集的大小均为 50 组。为了便于比较，采用同样的训练样本和测试样本分别对 SVDD 算法和文中算法进行仿真，试验结果见表 1。

表 1 试验结果的比较

算法	σ^2 值	0.0005	0.001	0.008	0.01	0.05	0.1	0.3	0.6	1
SVDD	检测率(%)	92.86	75.00	57.14	50.00	39.29	39.29	39.29	-	-
	误报率(%)	62.86	57.14	18.57	22.86	38.86	32.86	32.86	-	-
	操作时间(s)	110.4	104.1	135.3	321.9	186.9	177.9	156.2	-	-
文中算法	检测率(%)	-	-	-	42.86	71.43	96.43	96.43	96.43	1
	误报率(%)	-	-	-	22.86	12.86	5.71	2.86	4.29	12.86
	操作时间(s)	-	-	-	42.39	24.23	21.48	18.4	19.1	19.2

试验结果表明无论 σ^2 取何值，采用文中算法的操作时间都不到 SVDD 的 1/8，且当 σ^2 取 0.3 时，取得了较高的检测率和较低的误报率，表明此算法的可行性和有效性。而采用 SVDD 方法，当 σ^2 取比较小的值时，虽然可以达到不错的检测率，但这种效果是建立在较多的训练样本都作为支持向量的基础上，训练时间比较长，误报率也比较高。

利用文中算法对该检测点每天 14:30～18:30 时间段进行建模，共 24 天的数据，选取前 19 天的数据作为训练样本，采样间隔为 5min，即训练样本为为 931 组，其中有事件样本为 49 组；测试样本为 245 组，其中有事件样本为 20 组。试验结果证明，当取 $C = 1, C_0 = 0.9, \sigma^2 = 0.3$ 时，分时段建模的检测率为 1，误报率为 0，操作时间为 314.8s，达到最好的检测效果。

多次试验证明，多时间段建模的检测结果大多优于不分时间段的整体建模检测结果，但是多时间段建模所需的模型比较多，比较适合于复杂路段，以提高检测效果。

3 结束语

高速公路事件自动检测是高速公路管理的一个重要环节，文中提出把支持向量数据描述应用到高速公路的事件检测上，这种算法不受输入数据顺序的影响，训练速度快，泛化能力强，试验结果表明该算法具有较好的检测结果。另外对一些复杂路段进行分时段建模也是一个值得进一步研究的方向。对于 C_0 的确定可以设为动态值，使其随着样本数目的增加而增大。也可对不同原因引起的交通事件进行细分类，即多类分类识别问题，可采用一对一、一对多方法进行实现。

参考文献：

[1] 徐吉谦. 交通工程总论[M]. 北京：人民交通出版社，1996.

[2] 姜紫峰，刘小坤. 基于神经网络的交通事件检测算法[J]. 西安公路交通大学学报，2000，20(3)：67－69.

[3] 周 伟，罗石贵. 基于模糊综合识别的事件检测算法[J]. 西安公路交通大学学报，2001，21(2)：70－73.

[4] Tax T M J，Duin R P W. Support Vector Domain Description [J]. Pattern Recognition Letters，1999，20(11－13)：1191－1199.

[5] 肖健华. 智能模式与识别方法[M]. 广州：华南理工大学出版社，2006.

[6] 郑晓星，吴今培. 基于支持向量数据描述的数据约简[J]. 计算机应用，2007，20(2)：74－76.

[7] Klaus－Robert M. Intelligent Data Analysis Group[EB/OL]. [1998－08]. http://ida. first. fhg. de/projects/bench/benchmarks. htm.

[8] Petty K. The Freeway Service Patrol Project and the I－880 Database[EB/OL]. [1995－08－14]. http://ipa. eecs. berkeley. edu/pettyk/FSP/.

(上接第 247 页)

也发现系数中的不和谐之处。K_3 和 K_4 是接球球员在 x 和 y 两个方向上的速度，它们对传球的影响应该不分伯仲，但从数据上来看，它们的值有一定的差别。

文中得到的效用函数是在没有对手的情况下得到的，因此还有很大的局限性。下一步所要做的工作首先是要对对手进行建模，而建模所需要的对手数据是很难得到的[9]，可以尝试从 log 文件中分析对手信息，然后进行效用函数的计算。

参考文献：

[1] 张润梅，王 浩，姚宏亮，等. 影响图及其在 Robocup 中的应用[J]. 系统仿真学报，2005，17(1)：134－137.

[2] 姚宏亮. 动态多智能体决策问题研究[D]. 合肥：合肥工业大学，2006：257－289.

[3] 张润梅. 基于影响图的智能 Agent 学习及其在 Robocup 中的应用[D]. 合肥：合肥工业大学，2004：257－289.

[4] Tuyls K，Maes S，Manderick B. Q－learning in Simulated Robotic Soccer Large State Spaces and Incomplete Information [DB/OL]. 2002. http://como. vub. ac. be：8080/Publications/uploads/1/icmla02. ps.

[5] 张晓勇，彭 军. RoboCup 中传球策略的实现[J]. 计算机工程，2004，30(23)：123－124.

[6] 郭 博，程家兴. RoboCup 仿真组的传球策略[J]. 计算机技术与发展，2006，16(2)：129－131.

[7] 于 磊，王 浩，王 骋. RoboCup 中传球策略研究[J]. 计算机工程与应用，2004(28)：59－61.

[8] 张家旺，韩光胜，张 伟. C5.0 算法在 RoboCup 传球训练中的应用研究[J]. 计算机仿真，2006，23(4)：132－134.

[9] Ledezma A，Aler R，Sanchis A，et al. Predicting opponent actions by observation[DB/OL]. 2004. http://www. springerlink. com/content/ec4rq5k3vd278p62/fulltext. pdf.