

基于核主成分分析的异常轨迹检测方法

鲍苏宁*, 张磊, 杨光

(中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116)

(*通信作者电子邮箱 baosuning@126.com)

摘要:针对现有算法不能有效应用于多因素轨迹异常检测的问题,提出基于核主成分分析(KPCA)的异常轨迹检测方法。首先,为了改善轨迹特征提取的效果,采用KPCA对轨迹数据进行空间转换,将非线性空间转换到高维线性空间;其次,为了提高异常检测的准确率,采用一类支持向量机对轨迹特征数据进行无监督学习和预测;最终检测出具有异常行为的轨迹。采用大西洋飓风数据对算法进行测试,实验结果表明,该算法能够有效提取出轨迹特征,并且与同类算法相比,该算法在多因素轨迹异常检测方面具有更好的检测效果。

关键词:异常轨迹检测;核主成分分析;高维特征空间;一类支持向量机

中图分类号: TP312 **文献标志码:** A

Trajectory outlier detection method based on kernel principal component analysis

BAO Suning*, ZHANG Lei, YANG Guang

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou Jiangsu 221116, China)

Abstract: In view of the fact that the existing algorithms cannot effectively be applied to multi-factor trajectory outlier detection, this paper proposed a new method named TOD-KPCA (Trajectory Outlier Detection method based on Kernel Principal Component Analysis). Firstly, in order to enhance the effect of trajectory feature extraction, the method used KPCA to do the space transformation for trajectories and converted nonlinear space to a high dimension linear space. Furthermore, in order to improve the accuracy of outlier detection, the method used one-class Support Vector Machine (SVM) to do unsupervised learning and prediction with trajectory feature data. Finally, the method detected those trajectories with abnormal behavior. The proposed algorithm was tested on the Atlantic hurricane data. The experimental results show that the proposed algorithm can effectively extract trajectory features, and compared with the same algorithm, the proposed algorithm has better detection results in terms of multi-factor trajectory outlier detection.

Key words: TRAjectory Outlier Detection (TRAOD); Kernel Principal Component Analysis (KPCA); high-dimensional feature space; One-Class Support Vector Machine (One-Class SVM)

0 引言

随着全球定位系统(Global Positioning System, GPS)、射频识别(Radio Frequency Identification, RFID)等定位设备的应用普及,轨迹数据呈现爆炸式增长^[1]。通过对轨迹数据集进行异常检测,可以检测出行进路线异常的飓风等自然灾害,根据位置传感器返回的轨迹数据集可以检测出出现故障的传感器。Lee等^[2]提出了基于TRAOD(Trajectory Outlier Detection)算法的异常轨迹检测框架;刘良旭等^[3]使用基于R-Tree的高效异常轨迹检测算法,根据轨迹间的距离特征矩阵来计算轨迹之间的距离以确定其是否匹配;Xiong等^[4]采用分层概率模型来检测一些局部正常而整体表现异常的轨迹;文献[5]采用基于马尔可夫假设的密度估计方法,对每条轨迹赋予一个概率值,并通过概率阈值来判定异常轨迹。现有轨迹异常检测方法都只是在原始轨迹数据空间进行轨迹特征的提取和处理,并没有考虑到轨迹内部各个特征之间的联系,提取效果不好,并且很难应用到多因素异常轨迹检测中。

针对轨迹异常检测中在原始轨迹空间特征提取效果差以

及不能有效应用于多因素轨迹异常检测的问题,提出基于核主成分分析(Kernel Principal Component Analysis, KPCA)的异常轨迹检测方法,将KPCA和一类支持向量机(One-Class Support Vector Machine, One-Class SVM)相结合。采用KPCA将轨迹数据从非线性的原始空间映射到高维线性空间,提高轨迹特征提取的有效性,并在此基础上,采用一类支持向量机进行无监督学习和预测。此外,该方法能够在不作任何修改的情况下有效适应多因素轨迹数据的异常检测。

1 基于KPCA的轨迹空间转换方法

基于KPCA的轨迹空间转换方法的主要思想是最大化类间散布,最小化类内散布,通过核矩阵分解保留代表轨迹特征的最优主分量。图1显示了基于KPCA的轨迹空间转换方法框图。

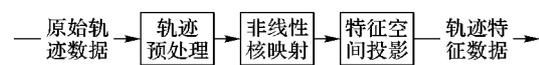


图1 轨迹特征提取框架

收稿日期: 2014-01-15; 修回日期: 2014-03-06。 基金项目: 中央高校基本科研业务费专项资金资助项目(2013XK10); 教育部博士点基金资助项目(20110095110010); 江苏省自然科学基金资助项目(BK20130208)。

作者简介: 鲍苏宁(1991-), 男, 江苏泰州人, 硕士研究生, 主要研究方向: 移动对象轨迹数据挖掘; 张磊(1977-), 男, 江苏沛县人, 副教授, 博士, 主要研究方向: 移动对象轨迹数据挖掘; 杨光(1988-), 女, 山东济宁人, 硕士研究生, 主要研究方向: 移动对象轨迹数据挖掘。

1.1 轨迹预处理

给定一条轨迹的 k 个样本点,如下表示这些点的集合

$$\mathbf{v} = [x_1, \dots, x_k, y_1, \dots, y_k, v_1, \dots, v_k, d_1, \dots, d_k]^T \quad (1)$$

这里 x_k, y_k, v_k, d_k 分别代表轨迹上第 k 个点的经度、纬度、速度和方向。采用合成少数类过抽样 (Synthetic Minority Over-sampling Technique, SMOTE) 算法^[6]对所有轨迹进行过采样,使所有轨迹具有相同的长度。

采用离差标准化 (Min-max normalization, Min-max)^[7]方法将轨迹特征数据映射到 0 和 1 之间,去除数据的单位限制,将其转化为无量纲的纯数值,便于不同单位或量级的指标进行比较。

连接标准化后的特征值:

$$\mathbf{v} = [x_1', \dots, x_L', y_1', \dots, y_L', v_1', \dots, v_L', d_1', \dots, d_L']^T \quad (2)$$

1.2 非线性核映射

利用核方法^[8]可将轨迹数据映射到高维特征空间^[9]中,能够获得比在原始空间中更好的特征属性。

向量集合 $\{\mathbf{v}_n\}_{n=1}^N$ 表示标准化的轨迹特征数据的集合,其中 \mathbf{v}_n 代表第 n 条轨迹特征数据的行向量。对特征向量集合中的每个向量,使用一个合适的非线性变换 $\Phi(\mathbf{v}_i): \mathbf{R}^m \rightarrow F$, $\mathbf{v}_i \rightarrow \mathbf{v}'_i$, 将特征向量 \mathbf{v}_i 映射到高维特征空间 F 中。在 KPCA 中,非线性变换 $\Phi(\mathbf{v}_i)$ 对应不同的核函数 $k(\mathbf{v}, \mathbf{v}')$ 。 F 空间中的协方差矩阵:

$$\Sigma = \frac{1}{N} \sum_{n=1}^N \Phi(\mathbf{v}_n) \Phi(\mathbf{v}_n)^T \quad (3)$$

求解协方差矩阵 Σ 的特征值及对应的特征向量:

$$\lambda \mathbf{u} = \Sigma \mathbf{u} \quad (4)$$

协方差矩阵 Σ 的第 i 个特征向量 \mathbf{u}_i 可通过 $\Phi(\mathbf{v}_1), \dots, \Phi(\mathbf{v}_N)$ 的线性组合来表示 ($\alpha_n^{(i)}$ 为系数):

$$\mathbf{u}_i = \sum_{n=1}^N \alpha_n^{(i)} \Phi(\mathbf{v}_n) \quad (5)$$

将式(3)和式(5)代入式(4):

$$\lambda_i \sum_{n=1}^N \alpha_n^{(i)} \Phi(\mathbf{v}_n) = \frac{1}{N} \sum_{n=1}^N \Phi(\mathbf{v}_n) \Phi(\mathbf{v}_n)^T \sum_{m=1}^N \alpha_m^{(i)} \Phi(\mathbf{v}_m) \quad (6)$$

将高维特征空间中的内积运算用核函数代替,即

$$k(\mathbf{v}, \mathbf{v}') = \Phi(\mathbf{v})^T \Phi(\mathbf{v}') \quad (7)$$

如果在式(6)的两边同时左乘 $\Phi(\mathbf{v}')^T$ 得:

$$\lambda_i \sum_{n=1}^N \alpha_n^{(i)} k(\mathbf{v}', \mathbf{v}_n) = \frac{1}{N} \sum_{n=1}^N k(\mathbf{v}', \mathbf{v}_n) \sum_{m=1}^N \alpha_m^{(i)} k(\mathbf{v}_n, \mathbf{v}_m) \quad (8)$$

根据式(8)可计算核矩阵 \mathbf{K} 中的元素:

$$K_{mn} = k(\mathbf{v}_m, \mathbf{v}_n) \quad (9)$$

为使均值 $\frac{1}{N} \sum_{i=1}^N \Phi(\mathbf{v}_i) = 0$, 需要中心化高维特征空间中的核矩阵 \mathbf{K} :

$$\mathbf{KL} = \mathbf{K} - \mathbf{1}_N \mathbf{K} - \mathbf{K} \mathbf{1}_N + \mathbf{1}_N \mathbf{K} \mathbf{1}_N \quad (10)$$

其中 $\mathbf{1}_N$ 代表 $N \times N$ 的矩阵,并且矩阵中的每个元素值都是 $1/N$ 。

1.3 特征空间投影

为了达到降维并提取主分量的目的,还需要对核矩阵 \mathbf{KL} 进行投影计算,运用豪斯荷尔德变换以及 QR 分解法^[10]求解

核矩阵 \mathbf{KL} 的特征值和特征向量,并在给定提取率 P 的约束下计算特征空间的投影,即主分量。

2 基于 KPCA 的异常轨迹检测方法

基于 KPCA 的异常轨迹检测 (Trajectory Outlier Detection method based on KPCA, TOD-KPCA) 方法,在 KPCA 的基础上采用一类支持向量机^{[5]13-15}对轨迹特征进行训练和预测,能够达到有效的异常检测效果。支持向量机 (Support Vector Machine, SVM) 已被广泛高效地运用到高维数据问题的处理中^[11]。一类支持向量机在处理数据聚类方面有着坚实的数学基础,适合于轨迹分析和异常检测^[12]。

2.1 核函数的选择

由于 KPCA 在高维特征空间提取的轨迹特征数据已经从原始轨迹数据转换为普通形式的特征数据值,所以在使用一类支持向量机进行无监督学习时,可以直接应用标准的高斯核函数:

$$k(x, y) = e^{-\|x-y\|^2/\sigma} \quad (11)$$

2.2 TOD-KPCA 算法的整体步骤

TOD-KPCA 算法描述如下:

输入: 轨迹样本集合 Trajectory $\text{trajs} = \{T_i\}_{i=1}^N$;

输出: 轨迹标签集合 $\text{labels} = \{\text{label}_i\}_{i=1}^N$ 。

1) 对轨迹样本进行 SMOTE 过采样^[6]。

2) 采用 Min-max 方法^[7]标准化每条轨迹。

3) 获取轨迹样本 (N 条轨迹,每条轨迹具有 m 个原始属性值) 矩阵:

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \dots & t_{1m} \\ t_{21} & t_{22} & \dots & t_{2m} \\ \vdots & \vdots & & \vdots \\ t_{N1} & t_{N2} & \dots & t_{Nm} \end{pmatrix}$$

4) 计算核矩阵 \mathbf{K} 。其中,可通过调节核函数的宽度 ω 来获取不同的核矩阵。

5) 中心化核矩阵 \mathbf{K} , 得到核矩阵 \mathbf{KL} 。

6) 计算核矩阵 \mathbf{KL} 的特征值 e_1, e_2, \dots, e_N 及对应的特征向量 $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N$ 。

7) 对特征值 e_1, e_2, \dots, e_N 进行降序排序。

8) 正交化特征向量,得到正交化后的特征向量 $\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_N$ 。

9) 累加每个特征属性的贡献率 p_1, p_2, \dots, p_n , 根据给定的提取率 P , 如果 $P_t \geq P (P_t = p_1 + \dots + p_t)$, 则提取前 t 个主分量 $\alpha_1, \alpha_2, \dots, \alpha_t$ 。

10) 计算特征空间上的投影 $\mathbf{Y} = \mathbf{KL} \cdot \boldsymbol{\alpha}$, 其中, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_t)$ 。

11) 通过 10) 得到新的轨迹数据集 $\chi = \{\chi_i\}_{i=1}^N$, 其中 χ_i 代表一条轨迹的特征向量, $\chi_i = \{\alpha_1, \alpha_2, \dots, \alpha_t\}$, 其中 α_t 代表轨迹 χ_i 的第 t 个主分量。

12) 对于数据集 $\chi_i \in \chi$ 中每条轨迹样本 χ_i , 利用所选的决策函数判断 χ_i 是否为异常轨迹, 如果是, 则将轨迹 χ_i 作为一条异常轨迹输出, 并设置相应的轨迹标签 label_i 为 -1。

算法伪码如下:

```

BEGIN:
1) InitializeParameter( $\omega, P$ )
2) labels = Initialize(labels)
3) traj = Preprocess(traj)
4) for  $i = 1$  to number of traj
5)   for  $j = 1$  to number of traj
6)      $K(i, j) = \text{Kernel}(\text{traj}(i, :), \text{traj}(j, :), \omega)$ 
7)    $KL = \text{NormalizeKernel}(K)$ 
8)    $(v, e) = \text{EigenvalueAndVector}(KL)$ 
9)   Sort( $v, e$ )
10)   $(v, e) = \text{Unitization}(v, e)$ 
11)   $se = \text{Sum}(e)$ 
12)   $sr = 0$  //accumulating contribution rate
13)  for  $i = 1$  to number of  $e$ 
14)     $sr = sr + e[i]$ 
15)    if  $sr/se > = P$ 
16)      break;
17)  Projection = Mapping( $KL, i, V$ )
18)  model = Training(Projection)
19)  for  $i = 1$  to number of traj
20)    labels[i] = sgn(traj[i], model)
21)    if (labels[i] == -1)
22)      output as outlier trajectory
23)    else
24)      output as normal trajectory
END
    
```

3 实验及分析

基于 TOD-KPCA 算法,开发了异常轨迹检测系统 Trajectory_Detection。该系统由 VC++2008 开发。实验硬件环境为:CPU 为 Core i3 2.27 GHz,内存为 2 GB,实验机器型号为联想 B460。实验数据集采用飓风数据集^[13]中 1950—2006 年的数据,包括 18944 个点构成的 608 条轨迹。

3.1 参数设置对 TOD-KPCA 算法的影响

方法主要涉及 3 个需要用户预先设定的参数: ω, P, ν 。 ω 表示 KPCA 中核函数的宽度; P 表示提取主分量时设置的提取率; ν 表示异常轨迹所占比例的渐进上限。

提取率 P 的大小将影响算法所提取的主分量的数目。表 1 显示了不同提取率(0.7~0.95)对应的实验结果。

表 1 提取率 P 对实验结果的影响($\omega = 14, \nu = 0.05$)

P	$oNum$	$pNum$	$nOutliner$
0.70	300	3	29
0.75	300	4	30
0.80	300	5	31
0.85	300	6	27
0.90	300	8	24
0.95	300	14	25

表 1 中, $oNum$ 表示原始轨迹数据的维度, $pNum$ 表示提取出的主分量数目, $nOutliner$ 表示检测到的异常轨迹的数目(以下相同)。实验结果表明,随着提取率值的增大,提取的主分量数目也在缓慢增加,说明提取的主分量在提取率增长很大的情况下,依然保持数目变动范围很小,能够充分代表轨迹的特征。

KPCA 中核函数宽度 ω 的取值直接影响算法的降维效果。表 2 显示了不同核函数参数 ω 对应的实验结果。

表 2 参数 ω 对实验结果的影响($P = 0.9, \nu = 0.05$)

ω	$oNum$	$pNum$	$nOutliner$
2	300	63	25
4	300	26	27
6	300	16	28
8	300	13	27
10	300	10	30
12	300	9	29

从表 2 中可以看出,随着核函数参数值的增大,提取的主分量的数目在快速减少,降维效果很明显。

在无监督学习中,参数 ν 用来控制异常轨迹所占比例的渐进上限。表 3 中显示了不同 ν 值(0.02~0.12)对应的实验结果。

表 3 参数 ν 对实验结果的影响($P = 0.9, \omega = 10$)

ν	$oNum$	$pNum$	$nOutliner$	$realRate$
0.02	300	10	13	0.021
0.04	300	10	27	0.044
0.06	300	10	38	0.063
0.08	300	10	52	0.086
0.10	300	10	61	0.100

从表 3 中可以看出,随着参数 ν 值的增大, $realRate$ (真实异常轨迹所占比例)值也在增大并且与 ν 值非常相近。

3.2 使用 KPCA 前后的实验结果对比

为了验证方法的有效性,分别对基于原始轨迹数据的方法和基于 KPCA 的方法在不同维度的轨迹特征数据上的检测效果进行了比较和分析(图 2~图 4)。参数设置: $\omega = 20, P = 0.9, \nu = 0.05$ 。

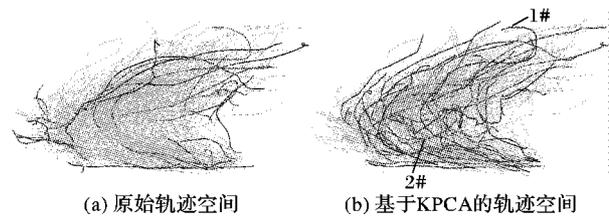


图 2 基于轨迹空间位置实验结果的比较

从图 2(a)中可以看出,基于原始轨迹数据的检测结果中,异常轨迹基本都集中在数据集的边缘,并且都是比较短的轨迹。图 2(b)显示了基于 KPCA 的检测结果,从图中可以看出,检测结果不仅仅局限于轨迹数据集的边缘,还检测出了很多位于数据集内部的异常轨迹(如图 2(b)中的 1#和 2#轨迹片段)。

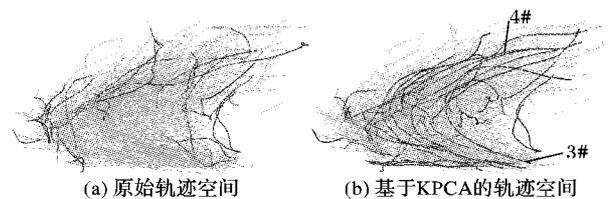


图 3 基于轨迹空间位置及速度实验结果的比较

图 3(a)显示,在基于原始轨迹数据的检测中,虽然加入了轨迹的速度特征,但是检测到的异常轨迹基本还是在轨迹形状上表现异常;图 3(b)显示,在基于 KPCA 的检测结果中,确实能够检测出在形状上表现正常,但是在速度特征上表现

异常的轨迹(如图3(b)中的3#和4#轨迹片段)。

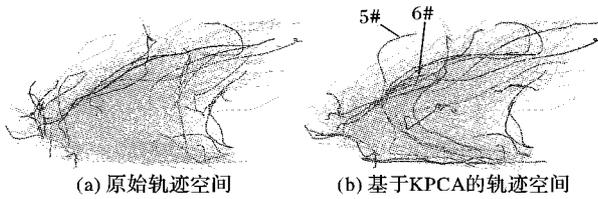


图4 基于轨迹空间位置及方向实验结果的比较

从图4(a)可以看出,在基于原始轨迹数据的检测结果中,加入轨迹的方向特征后,检测的结果并没有多大改变;图4(b)显示,在基于KPCA的检测结果中,能够检测出在方向上表现异常的轨迹(如图4(b)中的5#和6#片段)。

从6张对比图中可以看出,基于KPCA方法的异常检测结果要比基于原始轨迹数据检测的结果有效得多,这是因为原始轨迹空间是一个非线性的数据空间,不能有效地表示出轨迹数据内部的各个特征及其之间的关系,而KPCA具有很强的非线性特征描述能力^[14],通过将原始轨迹数据映射到高维特征空间,将原来非线性的特征数据转变为线性的特征数据,能更好地刻画轨迹数据的特征,从而检测结果也会非常好。

3.3 与同类算法比较

由于以往的异常轨迹检测算法都是基于轨迹空间位置,所以只在轨迹空间位置的基础上和算法TROAD进行比较。图5(a)和图5(b)分别显示了TROAD算法和TOD-KPCA算法在大西洋飓风轨迹数据集(1950—2006年)中检测到的异常轨迹。参数设置: $\omega = 20$, $P = 0.9$, $\nu = 0.05$ 。

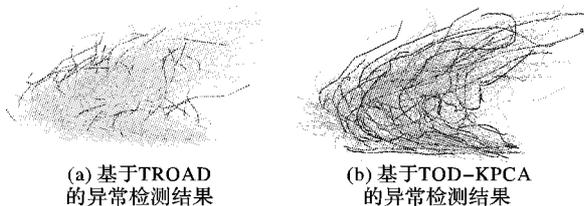


图5 TOD-KPCA与TROAD的实验结果对比

从图5中可以发现:TROAD算法更关注处于数据集边缘的轨迹,并且检测的多为形状上异常的轨迹。而TOD-KPCA算法检测出的异常轨迹,除了那些处于数据集边缘的轨迹外,同时还检测出很多位于数据集内部且表现异常的轨迹以及那些在形状上表现正常但是在其他特征上表现异常的轨迹,这是因为KPCA在非线性数据的特征表现和提取方面具有很强的描述能力,通过在线性的高维特征空间中能够发现轨迹内部的细微差异。

4 结语

本文研究轨迹异常检测,针对多因素轨迹数据的异常检测,提出基于KPCA的异常轨迹检测方法。在对原始轨迹数据进行重新采样和标准化处理后,该方法分为两个阶段:首先利用KPCA方法将原始轨迹数据从原始输入空间映射到高维特征空间,并在高维特征空间中提取出代表轨迹特征的最优主分量,降低轨迹数据的维度同时去除轨迹数据的不平衡性;第二步,使用一类支持向量机对提取的主分量进行无监督学

习得到推理模型,从而推断并检测出异常轨迹。实验表明,TOD-KPCA算法检测到的异常轨迹更具有实际意义,是一种有效的异常轨迹检测算法。

参考文献:

- [1] International Telecommunication Union. World Telecommunication/ICT development report 2010 [EB/OL]. [2013-10-12]. http://www.itu.int/ITU-D/ict/publications/wtdr_10/index.html
- [2] LEE J G, HAN J, LI X. Trajectory outlier detection: A partition-and-detect framework [C]// ICDE2008: Proceedings of the IEEE 24th International Conference on Data Engineering. Piscataway: IEEE, 2008: 140-149.
- [3] LIU L, QIAO S, LIU B, et al. Efficient trajectory outlier detection algorithm based on R-tree [J]. Journal of Software, 2009, 20(9): 2426-2435. (刘良旭, 乔少杰, 刘宾, 等. 基于R-Tree的高效异常轨迹检测算法[J]. 软件学报, 2009, 20(9): 2426-2435.)
- [4] XIONG L, POCZOS B, SCHNEIDER J C, et al. Hierarchical probabilistic models for group anomaly detection [C]// AISTATS2011: Proceedings of the 4th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale: Microtome Publishing, 2011: 789-797.
- [5] OLIVA J B. Anomaly detection and modeling of trajectories [D]. Pittsburgh: Carnegie Mellon University, 2012.
- [6] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 326-331.
- [7] JAIN A, NANDAKUMAR K, ROSS A. Score normalization in multimodal biometric systems [J]. Pattern Recognition, 2005, 38(12): 2270-2285.
- [8] SCHOLKOPF B, SMOLA A, MULLER K R. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation, 1998, 10(5): 1299-1319.
- [9] BISHOP C M. Pattern recognition and machine learning [M]. New York: Springer, 2006.
- [10] HAN X. The convergence of the QL algorithm and QR algorithm with shifts for symmetric tridiagonal matrix [J]. Higher School Journal of Computational Mathematics, 1995, 20(2): 1-5. (韩旭里. 对称三对角矩阵带位移的QL方法和QR方法的收敛性[J]. 高等学校计算数学学报, 1995, 20(2): 1-5)
- [11] JOACHIMS T. Text categorization with support vector machines: Learning with many relevant features [M]. Berlin: Springer, 1998.
- [12] PICIARELLI C, FORESTI G L. Anomalous trajectory detection using support vector machines [C]// AVSS2007: Proceedings of Advanced Video and Signal Based Surveillance. Piscataway: IEEE, 2007: 153-158.
- [13] Tropical Prediction Center. Atlantic tropical storm tracking by year [EB/OL]. [2013-09-15]. <http://weather.unisys.com/hurricane/atlantic/>
- [14] XU J, TAO X. One-class intrusion detection system based on KPCA space-similarity [J]. Journal of Computer Applications, 2009, 29(9): 2460-2460. (徐晶, 陶新民. 基于KPCA空间相似度的—类入侵检测方法[J]. 计算机应用, 2009, 29(9): 2460-2460.)