一种 CDRWPCA 网络故障特征提取算法

杨　婷[1]，　孟相如[1]，　温祥西[1,2]，　刘青原[1]
(1.空军工程大学信息与导航学院,陕西西安,710077;2.空军工程大学空管与领航学院,陕西西安,710051)

摘要　针对主元成分分析(PCA)在网络故障特征提取过程中可能丢失分类信息的问题,提出了一种中心距离比值加权主元成分分析(CDRWPCA)算法。算法计算样本每维特征的中心距离比值来衡量特征间的差异,并根据特征差异构造权重因子,对更具有鉴别性的特征赋予更大的权重,得到加权数据集;然后对加权数据集运用 PCA 进行特征提取后将提取后的数据集送入支持向量机(SVM)验证算法的有效性。算法相比较与 PCA 算法增加了时间复杂度,但相对于 PCA 算法本身的时间复杂度,增加不多。在网络故障诊断中的实验结果表明算法能在提取特征维数更少的情况下,提高了故障识别率。

关键词　特征提取;主元成分分析;中心距离比值
DOI　10.3969/j.issn.1009-3516.2013.06.016
中图分类号　TP393　　**文献标志码**　A　　**文章编号**　1009-3516(2013)06-0068-05

A Center Distance Ration Weighted Principal Component Analysis Algorithm for Network Fault Feature Extraction

YANG Ting[1], MENG Xiang-ru[1], WEN Xiang-xi[1,2], LIU Qing-yuan[1]
(1. Information and Navigation College, Air Force Engineering University, Xi′an 710077, China;
2. Air Traffic Control and Navigation College, Air Force Engineering University, Xi′an 710051, China)

Abstract: In view of the problem that the useful classification information in principal component analysis (PCA) may be lost in the process of network fault feature extraction, a new method named center distance ration weighted principal component analysis (CDRWPCA). According to sample category information, the center distance ratio of the difference between characteristics is measured by using this algorithm. By doing so, the weight is designed based on the feature discrimination. Then the weighted datasets are used for PCA feature extraction. Finally, the extracted datasets are sent to support vector machines (SVM) so as to verify the effectiveness of the algorithm. Experiments on network fault diagnosis demonstrate that the the proposed algorithm can improve the compression ratio and the final fault recognition rate.

Key words: feature extraction; principal component analysis; center distance ratio

由于网络故障的复杂性和传播性,连锁产生的故障关联症状往往使得采样数据中包含大量冗余,这些冗余特征的存在不仅会增加计算的复杂度,而且其中的冲突信息会干扰学习算法的最终决策。因此对特征参数进行分析,提取出最能描述网络故障的特征参数是非常有意义的。

特征提取是通过映射的方法将特征投影到低维空间,进而达到去除冗余特征的目的。特征提取的

收稿日期:2013-04-15
基金项目:国家自然科学基金资助项目(61201209)
作者简介:杨　婷(1989－),女,湖北应城人,硕士生,主要从事网络故障诊断研究.
E-mail:ytyhhxs@126.com

方法分为两类,即线性方法和非线性方法,典型的线性提取方法有主元成分分析(Principal Component Analysis,PCA)[1]、Fisher 判别分析(Fisher Discrimination Analysis,FDA)[2]等,典型的非线性提取方法有核主成分分析(Kernel Principal Component Analysis,KPCA)[3]和核 Fisher 判别分析(Kernel Fisher Discrimination Analysis,KFDA)[4]等。其中 FDA、KFDA 等方法是一种有监督的特征提取方法,但往往存在类内散度奇异值问题,普适性较差。PCA、KPCA 等方法是一种无监督的特征提取方法,仅仅在数据变化最大的几个方向进行特征提取,在提取过程中可能丢失对分类有用甚至是关键的鉴别信息。

为此,本文考虑将样本的分类信息融入 PCA 特征提取中,提出了一种有监督的特征提取方法——中心距离比值加权主元成分分析(Center Distance Ration Weighted Principal Component Analysis,CDRWPCA)。根据中心距离比值构造权重因子,对更具有鉴别性的特征赋予更大的权重,采用 PCA 对加权数据进行特征提取。

1　PCA 基本原理及分析

PCA[5]是通过剔除严格线性相关或相关性较强的自变量的信息,达到降维目的的一种分析方法。

设随机变量 $\boldsymbol{X}=[x_1,x_2]^{\mathrm{T}}$,其观测值见图 1 中圆圈。从图 1 中可看出,$x_1$ 和 x_2 之间并不存在严格的线性相关关系。但如果将坐标轴旋转得到相互正交的新坐标轴(t_1,t_2),则不难看出,变化主要体现在 t_1 上。分别称 t_1 和 t_2 为原变量 x_1 和 x_2 的第 1 主元和第 2 主元。如果 t_1 所提供的信息占系统总信息的绝大部分,则可以仅考虑 t_1 而忽略 t_2,从而使变量由 2 个变成 1 个,达到降维的目的。

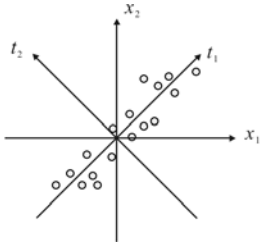


图 1　主元分析示意图

Fig. 1　Schematic diagram of PCA

为了定量描述主元所提供信息量的相对大小,定义如下的方差贡献率 δ_i 和主元的累计贡献率 η_a:

$$\begin{cases}\delta_i = \lambda_i / \sum_{j=1}^{N}\lambda_j \\ \eta_a = \sum_{i=1}^{a}\delta_i = \sum_{i=1}^{a}\lambda_i / \sum_{i=1}^{N}\lambda_i\end{cases} \tag{1}$$

累计贡献率 η_a 用来衡量前 a 个主元所含的信息量占总信息量的份额。一般情况下,可取前 a 个主元,使其累计方差贡献率 $\eta_a \geqslant 85\%$。PCA 提取主元的具体步骤描述如下:

STEP 1　对自变量 $\vec{\boldsymbol{x}}=[x_1,x_2,\cdots,x_D]^{\mathrm{T}}$ 进行 N 次观察并去除均值,得到其测量数据矩阵 $\boldsymbol{X}=[\vec{\boldsymbol{x}}_i,\vec{\boldsymbol{x}}_{i+1},\cdots,\vec{\boldsymbol{x}}_N]\in\boldsymbol{R}^{N\times D}$;

STEP 2　计算 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{\psi}_x = E(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) = \frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$;

STEP 3　求 ψ_x 的特征根 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_a \geqslant 0$ 及其相应的单位正交特征向量 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_a$;

STEP 4　计算得分矩阵 $\boldsymbol{T}=(t_i)=<x_i,u_i>=\boldsymbol{u}_i^{\mathrm{T}}x$;

STEP 5　按式(1)计算主元 t_i 的方差贡献率 δ_i 和前 a 个主元的累计方差贡献率 η_a,根据预先设定的累计方差贡献率 η_a(例如 85%),确定主元个数。

在图 2 中由于两类样本的分布在竖直方向上变化比较大(变化区间为 0 ～ 30),提取得到的第 1 主元方向实际包含的分类信息很小,而变化较小的水平方向(变化区间为 0 ～ 3),提取得到的第 2 主元方向几乎包含了所有的分类信息。图 3 可以清楚地看到,第 1 主元方向上的映射无法将两类样本区分开,而在特征值较小的第 2 主元方向的映射却可以完全将两类样本分开。显然 PCA 提取效果较好的条件是:在分类方向上数据变化较大,而在实际情况中往往不满足这一点。因此,考虑赋予对分类更重要的特征更大的权重,扩大分类方向上的数据变化。

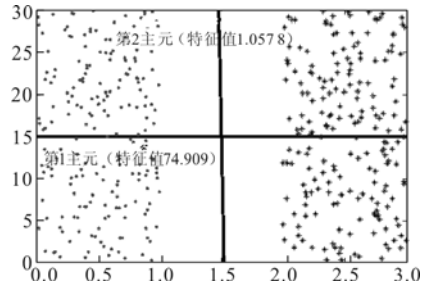
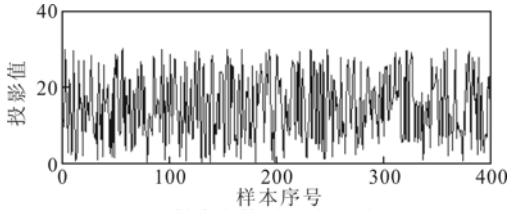
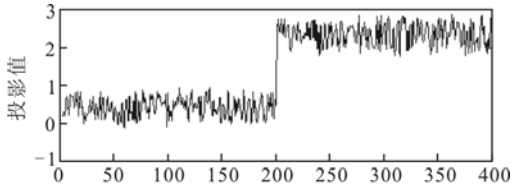


图 2　PCA 主元分类方向

Fig. 2　Classification direction of PCA



(a) 样本在第1主元投影值



(b) 样本在第2主元投影值

图 3 PCA 主元方向映射情况

Fig. 3 Principal direction map of PCA

2 CDRWPCA 算法

2.1 算法描述

FDA 提取的特征最具有判别性,最有利于分类。FDA[6] 的基本思想是:选择使得 Fisher 准则函数达到极值的向量,使得样本在该方向投影后,类内散度最小,类间散度最大。受 FDA 的启发,考虑运用样本特征的类内散度和类间散度来衡量样本特征对分类贡献度的大小。与 Fisher 准则不同的是采用样本每维特征的中心距离比值[7] 来计算该维特征的差异信息。特征差异越大的特征越具有鉴别性,对特征鉴别性更大的特征赋予更大的权值。

假设 C 类模式的训练样本为 $X^i=\{x_n \mid x_n \in \mathbf{R}^{N\times D}, y_n=i, n=1,2,\cdots,N_i, i=1,2,\cdots,C\}$, $N_1+N_2+\cdots+N_c=N$ (2)

式中:N_i 为第 i 类样本的样本个数;i 为样本类别;D 为样本维数;N 为所有样本总数。

则第 i 类样本第 k 维特征的中心为:

$$m_k^i=\frac{1}{N_i}\sum_{n=1}^{N_i}x_{n,k},k=1,2,\cdots,D,i=1,2,\cdots,C \quad (3)$$

第 $i(i=1,2,\cdots,C,)$ 类样本第 k 维特征到其自身中心的距离(自中心距)为:

$$S_{w,k}^i=\sum_{n=1}^{N_i}(x_{n,k}^i-m_k^i)^2,k=1,2,\cdots,D \quad (4)$$

第 $i(i=1,2,\cdots,C)$ 类样本第 k 维特征到第 $j(j\neq i$ 且 $j=1,2,\cdots,C)$ 类样本第 k 维特征的距离(互中心距)为:

$$S_{b,k}^{i,j}=\sum_{n=1}^{N_i}(x_{n,k}^i-m_k^j)^2,k=1,2,\cdots,D \quad (5)$$

第 k 维特征中心距离比值为:

$$p_k=\frac{\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}S_{b,k}^{i,j}}{\sum_{i=1}^{c}S_{w,k}^i},k=1,2,\cdots,D \quad (6)$$

p_k 为每维特征互中心距与自中心距的比值,自中心距反映样本第 k 维特征类内分布情况,互中心距反映样本第 k 维特征的类间分布情况,类内散度越小,类间散度越大,越有利于分类。因此以 p_k 的大小来衡量样本特征间的差异性。p_k 越大,则第 k 维特征对分类的贡献率越大。

为了保证权重因子为正,且随着 p_k 的增加而增加,定义权重矩阵为 $\mathbf{A}=(a_{ij})_{D\times D}$,每个 a_{ij} 表示为:

$$a_{ij}=\begin{cases}\exp(\eta p_i), i=j\\ 0, i\neq j\end{cases} \quad (7)$$

式中 η 为系数因子。$\mathbf{A}$ 为一个正定的对角矩阵,反映了相应的特征对分类贡献率的大小。

为了消除量纲的影响,保证 $0\leqslant a_{ii}\leqslant 1$, $\sum_{i=1}^{D}a_{ii}=1$,根据式(8) 对 $\mathbf{A}$ 进行归一化:

$$a_{ii}^{'}=a_{ii}/\mathrm{trace}(\mathbf{A}) \quad (8)$$

CDRWPCA 算法特征提取过程如下:

首先,训练样本数据归一化消除量纲。得到的数据集用 $\boldsymbol{X}=[\vec{x}_i,\vec{x}_{i+1}\cdots,\vec{x}_N]\in\mathbf{R}^{N\times D}$ 表示;

其次,根据式(4)~式(6)计算数据 X 每维特征中心距离比值,由式(7)~式(8)得到归一化后的权重矩阵。对相应特征进行加权得到新的数据集 $\boldsymbol{X}_{new}$,其中 $\boldsymbol{X}_{new}=\mathbf{A}\boldsymbol{X}$,并将其去中心化;

最后,对数据集 $\mathbf{X}_{new}$ 进行 PCA 特征提取。

用 PCA 算法计算协方差矩阵的时间复杂度为 $O(nD)$,对 $D\times D$ 协方差矩阵进行特征分析的时间复杂度为 $O(D^3)$,因此 PCA 算法的时间复杂度为 $O(nD)+O(D^3)$。CDRWPCA 算法比 PCA 算法增加了中心距离比值加权的步骤,式(3)的时间复杂度为 $O(N)$,式(4)的时间复杂度为 $O(D^2)$,式(5)的时间复杂度为 $O\left(\frac{C(C-1)}{2}D^2\right)$,式(6)的时间复杂度为 $O(D)$,所以 CDRWPCA 相比较与 PCA 算法增加的时间复杂度为 $O(N)+O(D^2)+O(C(C-1)/2D^2)+O(D)$,相比较与 PCA 本身的时间的复杂度 $O(nD)+O(D^3)$,增加不多。

2.2 算法验证

为了验证算法的有效性,采用 benchmarks[8] 数据集,实验数据结构见表 1。

表 1 实验数据结构

Tab. 1 The structure of the test samples

数据集	个数×维数	数据集	个数×维数
diabetis	768×8	image	2 310×18
flare	1 066×9	splice	3 175×60
german	1 000×20	twonorm	7 400×20
heart	270×13	thyroid	215×5

支持向量机(Support Vector Machines,SVM)运用原始 PCA 进行特征提取后再送入 SVM 分类和 CDRWPCA 特征提取后再送 SVM 分类的结果

进行了比较。PCA 和 CDRWPCA 主元个数的选取采用累积贡献率大于 90%的前几个主元,SVM 核函数为高斯径向基 $\mathbf{K}(x,y)=\exp(-\beta\|x-y\|^2)$,参数由 5 折交叉验证得到,系数因子 η 由经验值得到,结果见表 2。

从表 2 可以看出,PCA 特征提取的方法达到了降维的目的,尤其在 flare、german、image 数据集上降维效果较好。但是一些数据集经 PCA 特征提取后送入 SVM 分类的识别率有所降低,如 diabetis、german、image、splice 数据集。说明这些数据集不满足 PCA 提取效果较好的条件,即在分类方向上数据变化较大,此时用 PCA 进行特征提取会丢失一些分类信息。CDRWPCA 相比于 PCA 可以达到同等甚至更优的降维效果,尤其在 flare、heart 数据集上,CDRWPCA 的降维效果得到了明显改善。CDRWPCA 特征提取后的分类准确率要高于 PCA 特征提取后的分类准确率,说明运用加权的办法,将分类信息融入特征提取过程中的方法是有效的。

表 2　实验结果对比表

Tab. 2　The contrast of the experimental results

数据集	SVM		PCA-SVM		CDRWPCA-SVM	
	识别率/%	特征维数	识别率/%	特征维数	识别率/%	特征维数
diabetis	78.25	8	77.47	6	78.26	6
flare	67.54	9	67.54	4	67.64	1
german	76.70	20	76.00	13	77.00	13
heart	84.81	13	84.81	8	86.30	1
image	98.05	18	97.10	5	97.53	5
splice	81.33	60	79.33	46	86.67	42
thyroid	98.14	5	98.60	4	98.60	4
twonorm	97.00	20	97.25	17	97.25	17

图 4 为 heart 数据集系数因子 η 对于 CDRWPCA-SVM 特征压缩率和识别率的影响。从图中的柱状图看出,系数因子 η 值越大,提取以后得到的特征维数越少。从 SVM、PCA-SVM 和 CDRWPCA-SVM 的线状图走势看出:当系数因子 η 接近 0 时,CDRWPCA-SVM 的识别效果等同于 SVM 和 PCA-SVM 的识别效果;当系数因子在某一个合适的范围内时,CDRWPCA-SVM 的识别效果要优于 SVM 和 PCA-SVM 的识别效果。对于 heart 数据集而言,系数因子 $\eta\eta$ 的取值在[0.017 5,0.031]内时,CDRWPCA-SVM 的识别率高于 PCA-SVM 和 SVM,当 η 超过了一定范围,CDRWPCA-SVM 的识别率变得不理想。综合考虑特征压缩率和最后对分类的影响,选择合适的系数因子 η 在 CDRWPCA 特征提取中至关重要。

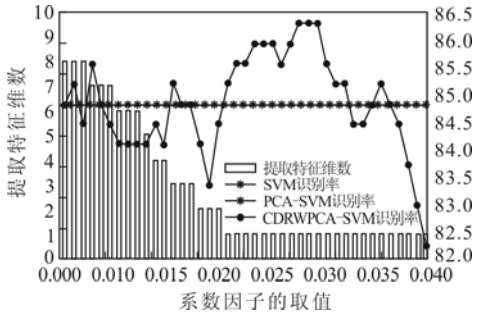


图 4　系数因子 η 对最终结果的影响

Fig. 4　The impact on final results of factor η

3　运用实例

网络中的各种攻击事件和病毒会导致网络中出现大量的“软故障”,因此采用美国国防部高级研究计划局(DARPA)提供的 Kdd'99[9] 数据集进行测试。该数据集包括 4 类网络攻击(分别是 DOS、R2L、U2R 和 PROBE)和一类正常数据集,其中 DOS、R2L、U2R 和 PROBE 依次对应故障 1、故障 2、故障 3 和故障 4。

对收集到的信息进行数值化和归一化处理,将各属性值的范围锁定在[0,1]。随机采样选择 1 200 组样本作为训练集,另取 1 920 组样本作为测试集,样本集结构见表 3。

表 3　实验样本集结构

Tab. 3　The structure of the test samples

类别	训练样本集	测试样本集
正常样本	500×41	800×41
故障 1	300×41	480×41
故障 2	100×41	160×41
故障 3	270×41	450×41
故障 4	30×41	30×41
总计	1 200×41	1 920×41

为了直观的看出不同样本的可分情况,随机选取 5 类训练样本集各 30 个,经 PCA 和 CDRWPCA 提取的 5 类样本的前 2 个主元的二维特征投影见图 5。从图 5 看出 PCA 和 CDRWPCA 取前二维主元,均能在一定程度上将 5 类样本区分开,但是 PCA 方法较难将故障 1 和正常样本区分开,而 CDRWPCA 算法明显提高了故障 1 和正常样本的区分度。

将实验样本集数据预处理后,对比 SVM、PCA-SVM 和 CDRWPCA-SVM 的识别效果,多分类的过程采用文献[10]提出的二叉树的方式,PCA 和 CDRWPCA 主元个数的选取采用累积贡献率大于 90%的前几个主元,选择高斯径向基作为 SVM 的核函数,核带宽为 1,惩罚因子为 2,系数因子 η 的选择综合考虑分类精度和特征压缩率,本文根据经验

选择 0.000 1,实验结果见表 4。

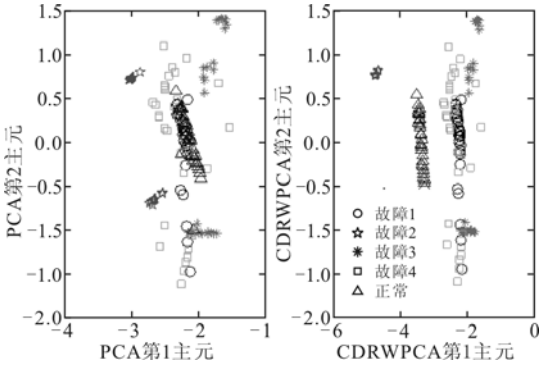


图 5 提取的二维特征平面图

Fig. 5 Plan of two dimensions

表 4 实验结果对比表

Tab. 4 The contrast of the experimental results

数据集	SVM		PCA-SVM		CDRWPCA-SVM	
	识别率/%	特征维数	识别率/%	特征维数	识别率/%	特征维数
故障 1	100.00	41	100.00	9	100.00	7
故障 2	100.00	41	100.00	9	100.00	7
故障 3	87.33	41	86.22	9	88.44	7
故障 4	100.00	41	100.00	9	100.00	7
正常	100.00	41	100.00	9	100.00	7
总体	97.03	41	96.77	9	97.29	7

从表 4 看出 CDRWPCA 比 PCA 降维效果更明显且同时提高了分类精度。从图 6 样本的总体识别率随着主元个数的变化走势中可以看出:当主元个数较少时,特征提取后故障的识别率低于直接用 SVM 进行分类的识别率。说明选择的主元个数过少,使得原始变量的有用信息有所损失。

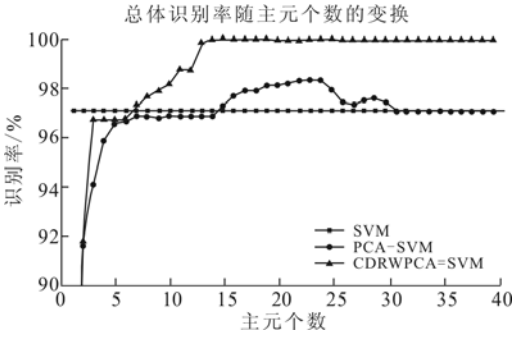


图 6 特征提取分类效果

Fig. 6 The classification results of feature extraction

当主元个数增加到一定值时,特征提取后识别率高于直接用 SVM 进行分类的识别率,说明对原始数据集进行特征提取是有必要的。随着主元个数的增加,CDRWPCA-SVM 比 PCA-SVM 更快的高于 SVM 的识别率,说明 CDRWPCA 选择的前几个主元中由于融合了分类信息,使得提取的主元更有利于分类。随着主元个数的进一步增加,CDRWPCA-SVM 的识别率明显高于 PCA-SVM 和 SVM 的识别率。CDRWPCA 能在提取特征维数比 PCA 更少的情况下,提高故障识别率。

4 结语

为了克服 PCA 在特征提取过程中可能丢失分类信息的缺点,运用加权的办法将样本的分类信息融入特征提取过程中,在 benchmarks 数据集上的实验结果验证了本文算法的有效性。在网络故障特征的提取中运用 CDRWPCA 算法,能在提取更少特征维数的情况下,提高故障识别率。

参考文献(References):

[1] Axel Coussement, Olivier Gicquel, Alessandro Pa-rente. Kernel density weighted principal component analysis of combustion processes[J]. Combustion and flame, 2012, 159(9): 2844-2855.

[2] 周欣,吴瑛. 基于 KPCA 和 LDA 的信号调制识别[J]. 系统工程与电子技术,2011,33(7):1611-1616.
ZHOU Xin, WU Ying. Signal modulation recognition based on KPCA and LDA[J]. Systems engineering and electronics, 2011,33(7):1611-1616. (in Chinese)

[3] Liu Nan, Wang Han. Weighted principal component extraction with genetic algorithms [J]. Applied soft computing, 2012, 12(2):961-974.

[4] Gyeongyong Heo, PaulGader. Robust kernel di-scriminant analysis using fuzzy memberships [J]. Pattern recognition, 2011,44,716-723.

[5] 王桂增,叶昊,主元分析与偏最小二乘[M]. 北京:清华大学出版社,2012.
WANG Guizheng, YE Hao, Principal analysis and partial least-squares [M]. Beijing: Tsinghua university press, 2012. (in Chinese)

[6] Wang Ziqiang, Sun Xia. Multiple kernel local Fisher discriminant analysis for face recognition [J]. Signal processing, 2013, 93(6):1496-1509.

[7] 焦李成,张莉,周伟达. 支撑矢量预选取的中心距离比值法[J]. 电子学报,2001, 29(3):383-386.
JIAO Licheng, ZHANG Li, ZHOU Weida. Pre-extrating support vectors for support vector machine[J]. Acta electronica sinica, 2001, 29(3):383-386. (in Chinese)

[8] Rätsch. Onoda and Müller[EB/OL][2001-01-01]. [2013-4-1]. http://ida. frst. gmd. de/nraetsch/data/benchmar. htm.

[9] Univer University of California trvine. VCI KDD Archive [EB/OL]. [2009-7-25][2013-4-1]. http://kdd. ics. uci. edu/.

[10] 周小平,晏蒲柳,吴静. 基于支持向量机的网络故障在线诊断方法研究[J]. 武汉大学学报:工学版,2006,39(3):102-106.
ZHOU Xiaoping, YAN Puliu, WU Jing. Research on network fault diagnosis method based on support vector machines [J]. Journal of Wuhan university: engineering edition, 2006, 39(3):102-106. (in Chinese)

(编辑:徐楠楠)