

文章编号:1001-9081(2010)10-2834-04

基于最小二乘模糊单类支持向量机的网络故障检测

张立,孟相如,张亚普

(空军工程大学 电讯工程学院,西安 710077)

(abel0000@126.com)

摘要:针对基于单类支持向量机的网络故障异常检测存在的训练速度慢和检测精度低等问题,提出一种最小二乘模糊单类支持向量机(LSFOC-SVM)。该方法采用最小二乘损失函数和等式化约束改进标准单类支持向量机的训练算法,将二次规划转化为解线性方程组,降低了计算代价;并通过构造基于特征空间距离的模糊隶属度函数和优化选择告警阈值,适当扩大了故障预警范围,提高了故障检测率。与同类方法相比,该方法在保证检测效果的同时大幅度地提升了训练效率。应用测试结果表明该方法是可行的。

关键词:网络故障检测;支持向量机;单类分类;最小二乘;模糊隶属度

中图分类号: TP393.08; TP18 **文献标志码:** A

Network fault detection based on fuzzy one class SVM with least squares and equality constraints

ZHANG Li, MENG Xiang-ru, ZHANG Ya-pu

(Institute of Telecommunication Engineering, Air Force Engineering University, Xi'an Shaanxi 710077, China)

Abstract: A new classifier named Least Squares Fuzzy One Class Support Vector Machine (LSFOC-SVM) was proposed to enhance the efficiency and effect of one class support vector machine applied to network fault abnormal detection. The proposed LSFOC-SVM not only reduced the high computational cost by learning with the least squares and equality constraint which obtain a set of linear equations instead of quadratic programming, but also enhanced the fault detection rate by extending the fault alarm area properly with fuzzy membership based on distance in feature space and appropriate alarm threshold. The comparative study results indicate LSFOC-SVM can improve the training efficiency greatly without affecting the diagnosis accuracy. And application tests verify the feasibility of this method.

Key words: network fault detection; Support Vector Machine (SVM); one class classification; least squares; fuzzy membership

0 引言

网络故障检测是根据目标网络运行状态判断其是否发生故障的过程,其实质是模式识别的单值问题。作为故障管理的关键环节,网络故障检测的效果直接影响到网络安全与性能^[1]。目前,随着知识工程、专家系统、神经网络和支持向量机(Support Vector Machine, SVM)^[2]等技术在领域的成功应用,以智能技术为核心的检测方法成为业界的研究热点。然而,以上方法的检测模式以误用检测为主^[3-4],由于网络故障类型多、新故障层出不穷、样本获取难度大,使得误用检测难以掌握充分的故障知识,从而影响检测精度。异常检测方法可以避免高代价的故障样本采集,且对未知故障具有良好的检测效果。在 SVM 研究领域, Scholkopf^[5]提出的 v -单类支持向量机(v -One Class Support Vector Machine, v -OCSVM)根据目标类样本分布划分特征空间区域,对未知样本做出“是”或“非”的判断,可以作为异常检测的实现基础。但 v -OCSVM 本身的缺陷制约了其在网络故障检测中的应用。一方面, v -OCSVM 的训练是二次规划过程,计算代价高,求解速度慢;另一方面,由于可能存在的噪声或离群点的影响,使得目标样本区域过大而导致误检率较高^[6]。

本文采用最小二乘损失函数和等式化约束^[7]改进

v -OCSVM 的训练算法,并在预测阶段引入基于特征空间距离的模糊判别方法,提出了最小二乘模糊单类支持向量机(Least Squares Fuzzy One Class Support Vector Machine, LSFOC-SVM)。LSFOC-SVM 的模型可以通过求解线性方程组得到,同时构造模糊隶属度和优化选择告警阈值,适当扩大了故障预警范围。相比同类方法,该方法在保证检测效果的同时大幅度地提升了训练效率。实验结果表明,该方法能够快速准确地检测故障,可以满足网络故障检测的应用要求。

1 最小二乘模糊单类支持向量机

LSFOC-SVM 以 v -OCSVM 为基础,融合最小二乘损失、等式化约束与模糊判别思想,分别对训练和测试环节提出改进,包括基于最小二乘单类 SVM 训练算法和基于特征空间距离的模糊判别。

1.1 v -OCSVM 与异常检测

对训练数据集 $x_i \in \mathbf{R}^n (i = 1, \dots, l)$, v -OCSVM 的基本思想是在高维特征空间中寻找一个最优超平面 $(w, \phi(x))$, 使之以尽可能大的距离 ρ 将尽量多的样本从原点分开。其形式化描述如下:

$$\min_{w, \rho, \xi} \frac{1}{2} \|w\|^2 + \frac{1}{vl} \sum_{i=1}^l \xi_i - \rho \quad (1)$$

收稿日期:2010-04-07;修回日期:2010-05-20。

基金项目:陕西省自然科学基金资助项目(SJ08F14);空军工程大学电讯工程学院研究生创新基金项目。

作者简介:张立(1981-),男,湖北武汉人,博士研究生,主要研究方向:网络故障诊断;孟相如(1963-),男,陕西蓝田人,教授,博士生导师,主要研究方向:宽带通信网络;张亚普(1985-),男,河北石家庄人,硕士研究生,主要研究方向:网络故障诊断。

$$\begin{aligned} \text{s.t.} \quad & (\mathbf{w} \cdot \phi(\mathbf{x}_i)) \geq \rho - \xi_i \\ & \xi_i \geq 0, i = 1, \dots, l \end{aligned}$$

其中: $\phi(\cdot)$ 是核空间映射函数, \mathbf{w} 为权向量, $\xi_i \in R$ 是为适应离群点引入的松弛变量, 可调参数 $v \in (0, 1)$ 控制着总样本数中错误样本比率的上界。式(1)的 Lagrange 对偶为:

$$\min_{\alpha} \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (2)$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq \frac{1}{vl}, i = 1, \dots, l$$

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

其中: $\alpha_i (i = 1, 2, \dots, l)$ 为 Lagrange 乘子, 核函数 $K(\mathbf{x}_i, \mathbf{x}_j) = [\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)]$ 。由式(2)解出 α_i , 得到决策函数:

$$f(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^l \alpha_i K(\mathbf{x}, \mathbf{x}_i) - \rho \right) \quad (3)$$

v -OCSVM 通过训练目标样本在特征空间中建立自我区域。基于否定选择思想, 待测数据若处在该区域内, 则认为是自我的同类样本, 否则判别为非我的异类样本。基于正常数据的网络故障异常检测具有样本丰富、实现简单、对未知故障适应性强等优势。但应用于实际, 仍需提高 v -OCSVM 的训练速度和判别准确率。

1.2 最小二乘单类 SVM 训练算法

在大样本情况下, 式(2)的凸二次规划问题计算代价极高, 这成为制约其应用的主要问题。为了提高训练速度, 本文引入最小二乘损失函数和等式化约束的方法改进 v -OCSVM, 得到式(4)所示的优化问题:

$$\begin{aligned} \min_{\mathbf{w}, \rho, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{vl} \sum_{i=1}^l \xi_i^2 - \rho \\ \text{s.t.} \quad & (\mathbf{w} \cdot \phi(\mathbf{x}_i)) = \rho - \xi_i, i = 1, 2, \dots, l \\ & \text{用 Lagrange 方法求解上述优化问题:} \\ L(\mathbf{w}, \rho, e, \alpha) = \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{vl} \sum_{i=1}^l \xi_i^2 - \rho \\ & \sum_{i=1}^l \alpha_i [\mathbf{w} \cdot \phi(\mathbf{x}_i) - \rho + \xi_i] \end{aligned} \quad (4)$$

其中 $\alpha_i \in R$ 为拉格朗日乘子, 由极值条件可得:

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}} = 0 \rightarrow \mathbf{w} = \sum_{i=1}^l \alpha_i \phi(\mathbf{x}_i) \\ \frac{\partial L}{\partial \rho} = 0 \rightarrow \sum_{i=1}^l \alpha_i = 1 \\ \frac{\partial L}{\partial \xi_i} = 0 \rightarrow \xi_i = \frac{vl}{2} \alpha_i \\ \frac{\partial L}{\partial \alpha_i} = 0 \rightarrow \mathbf{w} \cdot \phi(\mathbf{x}_i) - \rho + \xi_i = 0 \end{cases} \quad (5)$$

上式可化为求解下面的线性方程组:

$$\begin{bmatrix} \mathbf{G} + (vl/2)\mathbf{I} & -\mathbf{e} \\ \mathbf{e}^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \rho \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (6)$$

其中: \mathbf{G} 为元素是 $G_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ 的 Gram 矩阵, \mathbf{I} 为单位矩阵, 向量 $\alpha = [\alpha_1, \dots, \alpha_l]^T$, $\mathbf{e} = [1, \dots, 1]^T$ 。以上的改进将约束条件由不等式变为等式, 一次损失函数换成二次损失函数, 其分类的几何意义是对目标样本用超平面最小二乘逼近, 使得与原点的平均距离最大。这种转化只需求解线性方程组而无需求解二次规划, 能大幅度提高计算速度, 节省训练时间。

1.3 基于特征空间距离的模糊判别

设任意待测样本 \mathbf{x}_j 非线性映射到特征空间为 $\phi(\mathbf{x}_j)$, 则 $\phi(\mathbf{x}_j)$ 与原点的距离 γ_j 可表示为:

$$\gamma_j = \langle \mathbf{w}, \phi(\mathbf{x}_j) \rangle \quad (7)$$

其中 $\langle \cdot, \cdot \rangle$ 为特征空间点积。将式(6)的第一子式代入(8)中, 得:

$$\gamma_j = \left\langle \sum_{i=1}^l \alpha_i \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \right\rangle = \sum_{i=1}^l \alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \sum_{i=1}^l \alpha_i K(\mathbf{x}_i, \mathbf{x}_j) \quad (8)$$

根据式(9), v -OCSVM 的原始判别规则为: $\gamma_j > \rho$ 为目标样本, $\gamma_j = \rho$ 为边界样本, $\gamma_j < \rho$ 为异类样本。经过 1.2 节最小二乘改进后的 v -OCSVM, 依据目标类样本误差平方的总体信息划分特征空间, 这种情况下, 可能存在的噪声或离群点极易导致目标类别区域的扩张。若继续采用基于式(9)的硬划分规则会降低异类样本的识别率。从故障止损的角度考虑, 正常样本的错判尚可通后续的分类分析纠正, 而故障样本的漏判则有可能引起难以估量的损失。因此, 本文对分类边界采用模糊描述, 基于特征空间距离构造待测样本属于异类的模糊隶属度函数:

$$\mu_{-}(\mathbf{x}_j) = \begin{cases} 1, & \gamma_j \leq \rho \\ \left(1 + \left|\frac{\gamma_j - \rho}{\|\mathbf{w}\|}\right|^2\right)^{-1}, & \gamma_j > \rho \end{cases} \quad (9)$$

式(10)将原始判定的异类样本和边界样本确定判别为异类样本(隶属度为1); 对于原始判定的目标样本, 随着特征空间中 γ_j 与 ρ 相对距离的增大, \mathbf{x}_j 属于异类的可能性以广义钟型隶属度函数逐渐衰减。

根据式(10)的模糊隶属度函数, 建立如下模糊判别规则。
输入: 待测样本 \mathbf{x}_j , 训练使用的核函数 $K(\cdot, \cdot)$, 训练所得的最优 Lagrange 乘子向量 α^* 、距离 ρ^* 以及给定的告警阈值 θ ;
输出: \mathbf{x}_j 所属类别。

步骤1 将 $\mathbf{x}_j, K(\cdot, \cdot), \alpha^*$ 代入式(9)计算 γ_j ;
步骤2 比较 γ_j 与 ρ 的相对大小, 根据式(10)确定 \mathbf{x}_j 属于非目标类的模糊隶属度 $\mu_{-}(\mathbf{x}_j)$;
步骤3 若 $\mu_{-}(\mathbf{x}_j) < \theta$, 则 \mathbf{x}_j 属于目标类别; 否则, \mathbf{x}_j 属于异常类别。

对 LSFOC-SVM 而言, 属于 $(0, 1]$ 的告警阈值 θ 实际上决定着故障预警范围和检测效果。当 $\theta = 1$ 时, 模糊判别退化为原始的硬划分判别; 随着 θ 的减小, 特征空间中故障判定范围增大, 漏报率会相应减小, 但同时查全率(目标样本的检出率)也会减小。故判决前应通过实验优化选取适当的 θ 值, 获得查全率和故障漏报率的折中, 以提高检测器的总体性能。

2 基于 LSFOC-SVM 的网络故障检测

LSFOC-SVM 算法设计的网络故障异常检测过程如图1所示。学习阶段, LSFOC-SVM 算法训练程序提取历史样本库中积累的网络正常状态数据, 训练后将得到的模型存储于 SV 库; 检测阶段, 部署于目标网络的状态监视程序提取当前网络的特征数据, 预处理后输入检测模块。LSFOC-SVM 算法的模糊判别程序利用 SV 库提供的正常状态模式对其进行否定检测。考查待测样本对模式的符合程度, 若大于告警阈值, 则判定目标网络存在故障, 网络告警并触发故障类别分析程序, 输出诊断操作建议; 若小于告警阈值, 则认为网络工作正常, 继续网络状态监视。

3 实验与分析

3.1 数值实验

为测试算法有效性及比较各算法性能, 本文采用 DARPA 入侵检测评估数据集^[8], 以真实环境下的多种网络攻击模拟网络故障, 比较标准 SVM, v -OCSVM 和 LSFOC-SVM 的检测性能, 选取样本集结构如表1所示。

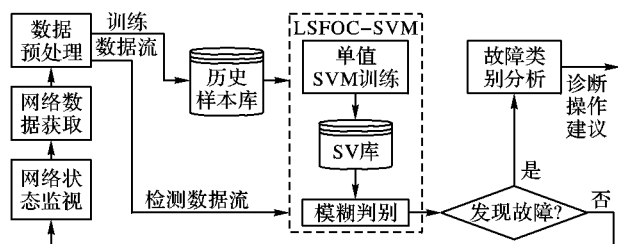
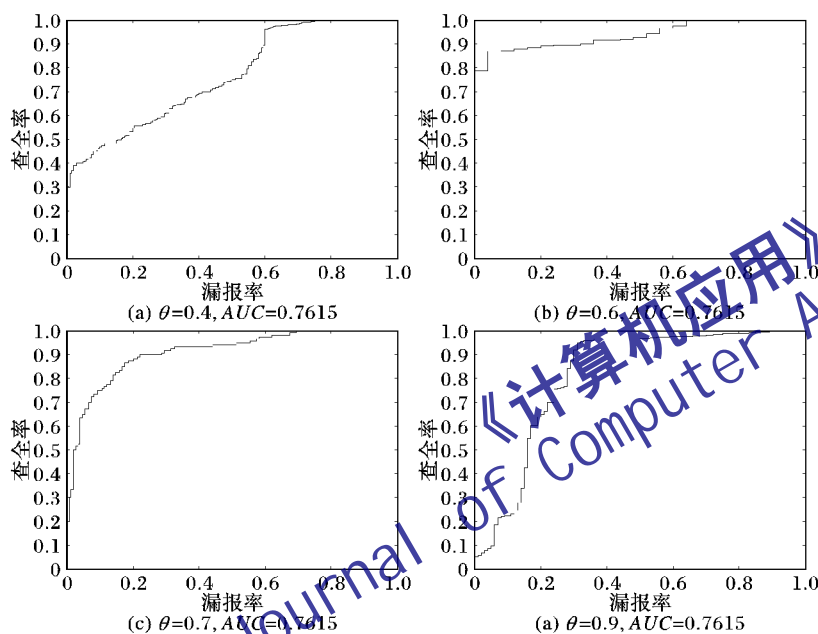


图1 基于LSFOC-SVM的网络故障异常检测过程

由于考虑到实际中未知故障的出现频度,在测试集中加

表1 实验样本集结构

算法	训练样本集	测试样本集
标准 SVM	正常样本 7000 条; 攻击样本 5000 条 (DoS 攻击样本 3500 条, Probing 攻击样本 1000 条, R2L 攻击样本 500 条)	正常样本 5000 条; 攻击样本 3000 条 (DoS 攻击样本 1500 条, Probing 攻击样本 1160 条, R2L 攻击样本 400 条, U2R 攻击样本 40 条)
v -OCSVM	正常样本 12000 条	
LSFOC-SVM	正常样本 12000 条	

图2 θ 取不同值对LSFOC-SVM的性能影响

当 θ 较小时(图2(a)),总体检测效果不好($AUC = 0.7615$),这是因为故障判定范围过大,导致漏报率较低时,查全率也较低;当 θ 较大时(图2(d)),故障的判定范围缩小,接近于原始硬划分,总体检测效果一般($AUC = 0.8147$),且查全率较高时漏报率也较高,这体现出离群点和单类模型不完备的故障知识对检测结果的影响;当 θ 介于0.6和0.7之间时, AUC 在0.9以上,检测器有较高的准确性,且曲线随查全率陡峭上升,即检测器在查全率较高时获得了较低的漏报率。通过以上比较,选定性能较优的 $\theta_1 = 0.6$ 和 $\theta_2 = 0.7$ 作为后续实验的告警阈值。

4)对各SVM检测器设置优化参数,训练样本集,得到检测模型;

5)用得到的模型分别对测试集进行检测,结果如表2所示(其中正常样本共5000条,故障样本共3000条)。

从训练时间来看,LSFOC-SVM明显少于其他两种算法,分别只有 v -OCSVM训练时间的1/4,标准SVM训练时间的1/5。说明最小二乘改进和等式化约束减小了计算代价,极大地提高了训练速度。从对未知故障(U2R)的检测来看,标准SVM仅识别出22条(55%),而LSFOC-SVM与 v -OCSVM分别检出了38条(95%)和33条(82.5%),检测率都达到了

入了少量未在训练集中出现的U2R类型样本,模拟未知故障。实验环境是:Windows XP系统,CPU 1.4 GHz,内存512 MB,Matlab 6.5。训练和预测步骤如下:

1)对实验样本集进行数值化和归一化,并采用文献[9]的方法约简属性得到9维数据;

2)将预处理后的数据输入各SVM检测器,通过交叉验证和网络搜索法确定RBF核参数;

3)在训练集上,采用ROC曲线^[10]分析告警阈值 θ 对LSFOC-SVM检测效果的影响,结果如图2所示。

80%以上,说明标准SVM误用检测模型依赖于故障学习,两种基于单类SVM的异常检测方法对未知故障显示出良好的适应性,特别是LSFOC-SVM由于使用了模糊检测,故障识别率更高;从告警效果来看, v -OCSVM的虚警率与漏报率比较平均,但都超过了20%,表明其检测性能较差;标准SVM有较低的虚警率,但漏报率受到未知故障的影响超过了10%;而LSFOC-SVM通过模糊判别和优化告警阈值,获得了最低的漏报率(仅1.27%),更加接近网络故障检测最大限度预警故障隐患的实际要求。虽然同时虚警率有所增长,但可以通过后续的故障类别分析纠正;从总的检测精度来看,两个告警阈值下的LSFOC-SVM的检测精度都在86%以上,略低于标准SVM的检测精度,明显优于 v -OCSVM。说明了LSFOC-SVM的改进能够大幅度提高单类支持向量机的检测精度,达到接近标准SVM的检测水平。

总体而言,本文提出的LSFOC-SVM能够快速准确地检测故障,比标准SVM和 v -OCSVM更加适合网络故障检测。

3.2 应用测试

3.1节的数值实验结果说明了LSFOC-SVM与同类检测算法相比在训练时间和故障检测率上具有一定优势,下面在实际网络中测试基于LSFOC-SVM的网络故障异常检测效果,实验网拓扑如图3所示。

分别提取根节点路由器Router1与下级交换机相连的两个以太网口的部分MIB变量作为样本特征,获得特征结构如表3所示的样本数据。

积累网络正常状态下的样本数据,待样本库增长趋于停止时,将得到的318组样本输入LSFOC-SVM(RBF核参数 $\sigma = 0.5, v = 0.05$)训练提取网络正常状态模型(训练时间仅为0.029 s)。人为设置链路断开、端口配置错误、网络设备停机及网络拥塞等故障场景,使用LSFOC-SVM进行异常检测,告警阈值 θ 设为0.6,检测结果如表4所示。

可以看出,本文提出的LSFOC-SVM网络故障异常检测方法在实际网络环境中检测率达到90%以上,且对于各种类型的故障无一漏报,显示了良好的检测性能和实用性。同时,其极快的训练速度能够支持网络故障的在线检测,为网络故障检测提供了一条有效的途径。

表 2 故障检测结果

算法	训练时间/s	正常样本检出条数	虚警率/%	故障样本检出条数	漏报率/%	检测精度/%
标准 SVM	82.32	4 738	5.24	2 648	11.73	92.32
<i>v</i> -OCSVM	62.89	3 884	22.32	2 227	25.77	76.39
LSFOC-SVM	$\theta = 0.7$ $\theta = 0.6$	4 179	16.42	2 749	8.37	86.60
		4 118	17.64	2 962	1.27	88.85

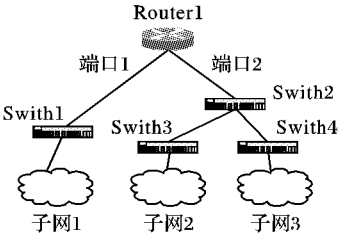


图 3 实验网络拓扑

表 3 样本特征结构

编号	特征名称	数值范围
1	链路协议状态	1-up 2-down
2	端口管理状态	1-up 2-down
3	端口当前状态	1-up 2-down
4	输入丢包率	0 ~ 100%
5	输出丢包率	0 ~ 100%
6	CRC 错误率	0 ~ 100%
7	平均输入流量	0 ~ 100%
8	平均输出流量	0 ~ 100%

表 4 检测结果

检测集	检测条数	检出条数	检测率/%
正常样本	168	152	90.5
故障样本	127	127	100.0
总计	295	279	94.6

4 结语

本文提出的 LSFOC-SVM 融合了 *v*-OCSVM 的异常检测、ISSVM 的最小二乘和等式化约束以及基于特征空间距离的模糊检测等思想。与同类检测算法相比,具有训练快速、判别准确、漏报率低及对未知故障适应性强等特点,可以满足网络故障检测的实际需要,对事故的预防起到积极的作用。然而,LSFOC-SVM 求得的 α 虽然满足极值条件,但不能保证是全局

最优解,这体现在其检测精度略低于标准 SVM 以及较高的虚警率上。进一步提高 LSFOC-SVM 的泛化性能,使之减少对后续识别环节的依赖,是下一步的研究方向。

参考文献:

[1] HAJJI H. Statistical analysis of network traffic for adaptive faults detection [J]. IEEE Transactions on Neural Network, 2005, 16(5): 1053 – 1063.

[2] HABIB T, INGLADA J, MERCIER G, *et al.* Support vector reduction in SVM algorithm for abrupt change detection in remote sensing [J]. IEEE Geoscience and Remote Sensing letters, 2009, 6(3): 606 – 610.

[3] 李千目, 许满武, 张宏, 等. 基于支持向量机的网络应用层故障检测系统[J]. 系统仿真学报, 2006, 18(7): 1806 – 1809.

[4] ZHANG LI, MENG XIANG-RU, ZHOU HUA. Network fault diagnosis using hierarchical SVMs based on kernel method [C]// 2009 Second International Workshop on Knowledge Discovery and Data Mining. Washington, DC: IEEE Computer Society, 2009: 753 – 758.

[5] SCHOLKOPF B, BLATT J C, SHAWE-TAYLOR J, *et al.* Estimating the support of a high-dimension distribution [J]. Neural Computation, 2001, 7(13): 1443 – 1471.

[6] 张新峰, 刘垚巍. 广义超球面 SVM 研究[J]. 计算机研究与发展, 2008, 45(11): 1807 – 1816.

[7] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers [J]. Neural Processing Letters, 1999, 9(3): 293 – 300.

[8] University of California Irvine. UCI KDD Archive [DB/OL]. [2009 – 12 – 16]. <http://kdd.ics.uci.edu/>.

[9] ZHANG LI, MENG XIANG-RU, WU WEI-JIA, *et al.* Network fault feature selection based on adaptive immune clonal selection algorithm [C]// CSO 2009: International Joint Conference on Computation Sciences and Optimization. Washington, DC: IEEE, 2009, 2: 969 – 973.

[10] PIETRASZEK T. On the use of roc analysis for the optimization of abstaining classifiers [J]. Machine Learning, 2007, 68(2): 137 – 169.

(上接第 2824 页)

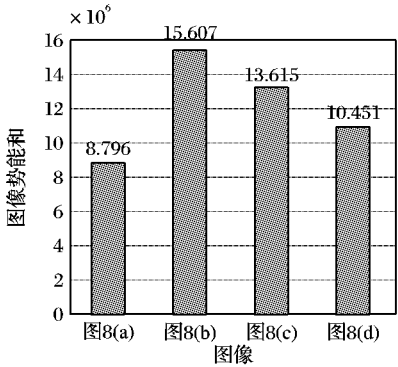


图 9 拉伸图像势能和比较

参考文献:

[1] 任卫军, 褚洪东, 贺显曜. 可变区域的视频图像幅型比非线性缩

放算法[J]. 电视技术, 2008, 32(311): 41 – 42.

[2] 侯发忠, 邹北骥, 李跃强. 基于罗伯茨梯度的彩色图像缩放新法 [J]. 计算机工程与设计, 2009, 30(14): 3367 – 3370.

[3] 李将云, 杨勋年, 汪国昭. 图像缩放的片连续算法[J]. 浙江大学学报: 理学版, 2002, 29(5): 530 – 534.

[4] PRATT W K. Digital image processing [M]. New York, USA: John Wiley & Sons, 2001.

[5] 王森, 杨克俭. 基于双线性插值的图像缩放算法的研究与实现 [J]. 计算机应用, 2008, 28(7): 44 – 45.

[6] 曹凤莲, 沈庆宏, 盛任农, 等. 一款基于新型 Field Programmable Gate Array 芯片的投影仪梯形校正系统研究与实现[J]. 南京大学学报, 2006, 42(4): 362 – 367.

[7] 田敏雄, 沈庆宏, 曹凤莲, 等. 基于图像空间变换和插值运算的投影仪梯形校正法[J]. 电子测量技术, 2007, 30(3): 10 – 12.