第35卷 第23期 计 算 机 工 程 2009年12月
Vol.35 No.23 Computer Engineering December 2009
·安全技术· 文章编号：1000—3428(2009)23—0118—02 文献标识码：A 中图分类号：N945.15

基于KCCA优化的网络入侵检测算法

钱鹏江[1]，王士同[1,2]，徐 华[1,2]，颜惠琴[3]
(1. 江南大学信息工程学院，无锡 214122；2. 南京理工大学，南京 210094；3. 无锡职业技术学院，无锡 214121)

摘 要：入侵检测系统所得原始特征通常是高维度的，这些高维度特征带来了较大的计算开销。针对该问题，采用核典型相关分析方法进行原始特征的二次提取，得到简约而重要的二次特征。在该二次特征的基础上运用二叉树多分类支持向量机法判别待测网络状态所属类别。仿真实验证明，该算法在不显著损失检测准确度的情况下可提升系统实时性，从而达到优化目标。

关键词：入侵检测系统；核典型相关分析；二次特征；二叉树支持向量机

Network Intrusion Detection Algorithm Based on KCCA Optimization

QIAN Peng-jiang[1], WANG Shi-tong[1,2], XU Hua[1,2], YAN Hui-qin[3]
(1. School of Information Technology, Jiangnan University, Wuxi 214122; 2. Nanjing University of Science and Technology, Nanjing 210094; 3. Wuxi Institute of Technology, Wuxi 214121)

【Abstract】In Intrusion Detection System(IDS), the original features normally lead to considerable computational complexity because of their high dimensions. In this paper, reduced and important further features are obtained by introducing Kernel Canonical Correlation Analysis(KCCA), and the binary tree based multi-class classification SVM is used to complete the classification task by these further features. This algorithm is devoted to increase the real-time performance as much as possible under the condition of not clearly losing classification accuracy. Simulation experiment confirms the above advantages.

【Key words】Intrusion Detection System(IDS); Kernel Canonical Correlation Analysis(KCCA); further features; binary tree based SVM

入侵检测系统(Intrusion Detection System, IDS)从网络关键链路收集信息进行分析、判断和决策，其所得原始信息通常高达几十维甚至上百维，直接基于此原始信息的检测判别伴随的是可观的计算开销，因此，能高效精简或提炼原始信息的方法不可或缺。核典型相关分析(Kernel Canonical Correlation Analysis, KCCA)方法就是其中之一。本文使用KCCA把IDS收集的大量高维原始特征降维成少量精辟的二次特征，然后以此二次特征为依据使用二叉树多分类支持向量机技术完成入侵检测任务。

1 核典型相关分析

设有输入空间样本集 $S_x=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_\ell\}$ 及对应的输出空间样本集 $S_y=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_\ell\}$，$\boldsymbol{x}_i\in R^N,\boldsymbol{y}_i\in R^N$；$\boldsymbol{\Phi}_x$ 和 $\boldsymbol{\Phi}_y$ 分别为它们到各自 Hilbert 空间的核映射；$\boldsymbol{w}_{\boldsymbol{\Phi}_x}$ 和 $\boldsymbol{w}_{\boldsymbol{\Phi}_y}$ 分别为 Hilbert 输入和输出空间的投影方向(projection direction)[1]。

由文献[1-2]知，关于 S_x 和 S_y 的 KCCA 可表示成如下最优化问题：

$$\max_{\alpha,\beta}\ \boldsymbol{\alpha}'\boldsymbol{K}_x\boldsymbol{K}_y\boldsymbol{\beta}$$
$$\text{s.t.}\quad \boldsymbol{\alpha}'\boldsymbol{K}_x\boldsymbol{K}_x\boldsymbol{\alpha}+\kappa\boldsymbol{\alpha}'\boldsymbol{K}_x\boldsymbol{\alpha}=1,\ \boldsymbol{\beta}'\boldsymbol{K}_y\boldsymbol{K}_y\boldsymbol{\beta}+\kappa\boldsymbol{\beta}'\boldsymbol{K}_y\boldsymbol{\beta}=1 \tag{1}$$

其中，κ 是泛化系数[2-3]；$\boldsymbol{K}_x,\boldsymbol{K}_y$ 分别对应 Hilbert 输入和输出空间核矩阵。

由拉格郎日法得式(1)的最优解：

$$\boldsymbol{\beta}=\frac{(\boldsymbol{K}_y+\kappa\boldsymbol{I})^{-1}\boldsymbol{K}_x\boldsymbol{\alpha}}{\lambda} \tag{2}$$
$$(\boldsymbol{K}_x+\kappa\boldsymbol{I})^{-1}\boldsymbol{K}_y(\boldsymbol{K}_y+\kappa\boldsymbol{I})^{-1}\boldsymbol{K}_x\boldsymbol{\alpha}=\lambda^2\boldsymbol{\alpha} \tag{3}$$

其中，$\boldsymbol{I}$ 为 $\ell\times\ell$ 单位矩阵。

这样，给定任意 $\tilde{\boldsymbol{x}}$，它在 $\boldsymbol{w}_{\boldsymbol{\Phi}_x}$ 上的投影为[1]

$$P_x(\tilde{\boldsymbol{x}})=\boldsymbol{\Phi}'_x(\tilde{\boldsymbol{x}})\boldsymbol{w}_{\boldsymbol{\Phi}(x)}=\boldsymbol{K}_{\tilde{x}}\boldsymbol{\alpha} \tag{4}$$

其中，$\boldsymbol{K}_{\tilde{x}}=[k_x(\tilde{\boldsymbol{x}},\boldsymbol{x}_1),k_x(\tilde{\boldsymbol{x}},\boldsymbol{x}_2),\cdots,k_x(\tilde{\boldsymbol{x}},\boldsymbol{x}_\ell)]'$。

同理，给定任意 $\tilde{\boldsymbol{y}}$，它在 $\boldsymbol{w}_{\boldsymbol{\Phi}_y}$ 上的投影为

$$P_y(\tilde{\boldsymbol{y}})=\boldsymbol{\Phi}'_y(\tilde{\boldsymbol{y}})\boldsymbol{w}_{\boldsymbol{\Phi}(y)}=\boldsymbol{K}_{\tilde{y}}\boldsymbol{\beta} \tag{5}$$

其中，$\boldsymbol{K}_{\tilde{y}}=[k_y(\tilde{\boldsymbol{y}},\boldsymbol{y}_1),k_y(\tilde{\boldsymbol{y}},\boldsymbol{y}_2),\cdots,k_y(\tilde{\boldsymbol{y}},\boldsymbol{y}_\ell)]'$。

注意，典型相关分析要求输入与输出空间的样本均值为零，可以分别在 $\boldsymbol{K}_x$ 和 $\boldsymbol{K}_y$ 上进行中心化处理即可。

2 基于KCCA优化的网络入侵检测算法

本文采用二叉树多分类 SVM 法作为系统的分类识别技术。其具体原理参见文献[4]，本文不作介绍。

上文已经说明了在高维度原始特征空间进行二次特征提取的必要性。给出具体实现方法如下：

(1)把各种网络状态定义为输出空间 Y，采集所得的大纬度信息定义为输入特征空间 X。现从 2 个空间各取若干对样本构建训练样本集 S_{train} 和测试样本集 S_{test}。

基金项目：国家自然科学基金资助项目(60773206, 60704047)；国家“863”计划基金资助项目(2006AA10Z248)；江南大学青年基金资助项目(105000-52210752)

作者简介：钱鹏江(1979－)，男，博士研究生，主研方向：模式识别，网络安全；王士同，教授、博士生导师；徐 华，博士研究生；颜惠琴，硕士

收稿日期：2009-07-06 **E-mail：**qianpjiang@126.com

(2)设 S_{train} 容量为 ℓ，经式(2)和式(3)得$(\lambda_i,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i)$, i=1,2,…, ℓ。由于 KCCA 法特征值等价于典型相关程度[1-2]，只取特征值为正实数的情况，得$(\lambda_j,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)$, j=1,2,…,C。且 $\lambda_1>\lambda_2>\cdots>\lambda_C$。这里 $\boldsymbol{\alpha}_j$ 和 $\boldsymbol{\beta}_j$ 即输入和输出空间中的 C 个典型相关向量。

(3)将训练集 S_{train} 的原始 N 维输入特征 $\boldsymbol{x}_i$, i=1,2,…, ℓ，经式(4)降为仅 C 维的二次特征，且 C 远远小于 N，再加上 y_i，构成二次特征空间训练样本集 F_{train}。同理得二次特征空间训练样本集 F_{test}。

(4)以 F_{train} 为训练样本，采用二叉树多分类 SVM 法构造多个分类器，以判别 F_{test} 中每个待测样本所属的类别。

注意：核矩阵进行如下中心化处理，达到样本集零均值化[3]：

$$\hat{\boldsymbol{K}}=\boldsymbol{K}-\frac{1}{\ell}\boldsymbol{j}\boldsymbol{j}'\boldsymbol{K}-\frac{1}{\ell}\boldsymbol{K}\boldsymbol{j}\boldsymbol{j}'+\frac{1}{\ell^2}(\boldsymbol{j}'\boldsymbol{K}\boldsymbol{j})\boldsymbol{j}\boldsymbol{j}' \tag{6}$$

其中，$\boldsymbol{j}$ 是 $\ell\times1$ 元素全为 1 的列向量。

3 算法的性能分析

IDS 很关注系统的实时响应时间，N 维的原始特征降成 C 维的二次特征必然带来系统计算复杂度的下降。但必须指出 KCCA 的降维过程也带来了额外的计算开销，该开销对系统的时间影响也是不容忽视的。

KCCA 分 2 个阶段进行：离线训练时期和在线识别阶段。前者计算 C 个典型相关向量及构造二次特征样本。但该阶段是预先进行的，对入侵检测的实时性影响可以不计；后者则对系统的实时性影响至关重要。待测样本首先需要和 ℓ 个训练样本按式(4)进行原始特征核映射，耗时 $O(\ell)$，然后与 C 个典型相关向量点积得到 C 维的二次特征，耗时 $O(C)$，因此每个待测样本由 KCCA 增加的额外实时复杂度为 $O(\ell+C)$。

4 仿真实验

本实验数据来自 KDD 1999 IDS 数据库，选取其中的 back, neptune, normal, smurf 这 4 种网络状态的若干记录构建了实验用数据集，该数据集原始特征为 41 维度。

系统平台：1.66 GHz 酷睿 2 CPU，3 GB DDR 667 MHz 内存，Windows XP，Matlab 7.1 等。

实验中二叉树多分类 SVM 和 KCCA 输入空间使用的核函数均为高斯径向核：

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\mu\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2) \tag{7}$$

KCCA 输出空间的元素仅是一标量，因此，采用相对简单的多项式核：

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=(<\boldsymbol{x}_i,\boldsymbol{x}_j>+b)^d \tag{8}$$

实验主要参数设置：KCCA 中式(7) μ = 0.010 1，式(8)b=1, d=2；典型相关向量个数 C=5；二叉树多分类 SVM 原始特征空间式(7) μ = 0.000 05，二次特征空间 μ = 0.004 3。

以不同训练样本容量测试了算法优化前和优化后 IDS 系统的性能指标，如表 1 所示。

随训练样本容量改变的 4 种时耗(优化前判别时耗、优化后判别时耗、KCCA 额外时耗和 KCCA 优化算法总时耗)性能差异如图 1 所示。

由表 1 和图 1 可知，当训练样本容量适度(本实验中约小于 400)，网络信息经 KCCA 由 41 维降为 5 维后识别准确度虽略有下降，但总体响应速度得到了大幅的提升(判别时耗大幅减少)。如训练样本容量为 280 时，准确度下降 2.66%，响应速度却提升 28.41%。但也发现 KCCA 额外时耗正比于训练样本的容量，当样本容量增大到一定程度时，KCCA 降维的额外开销反而会制约系统的整体响应时间。因此，该算法的关键是寻找训练样本容量的合适区间，以达到系统在不显著损失识别准确度的同时又能得到大幅的实时响应速度提升。

表 1　KCCA 优化前与优化后系统性能对照表

样本容量/条		未优化		KCCA 优化后		
训练集	测试集	判别时耗/s	准确度/(%)	总时耗/s		准确度/(%)
				KCCA 额外时耗	判别时耗	
40×4	1 200	6.39	99.00	3.23	0.72	94.17
50×4	1 200	7.64	99.08	4.17	0.73	94.25
70×4	1 200	9.89	99.08	6.22	0.86	96.42
110×4	1 200	10.18	99.08	10.75	1.46	97.08
120×4	1 200	11.38	99.08	11.95	1.59	97.08

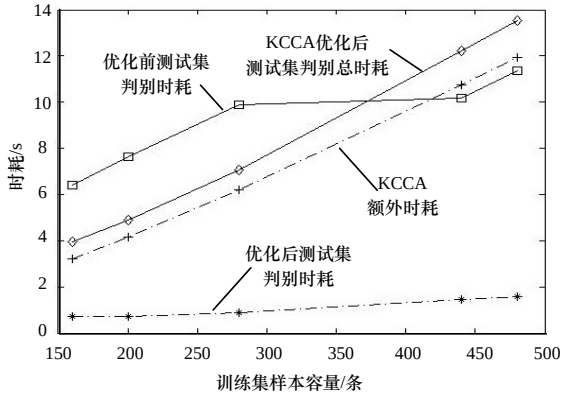

图 1　4 种时耗随训练样本容量变化曲线图

5 结束语

本文设计了一种基于 KCCA 优化的网络入侵检测算法，并通过仿真实验对该算法的优势进行了验证。实验结果表明，选定合适的训练样本容量，该算法能使 IDS 系统在不显著降低识别准确度的情况下较大幅度地提升实时响应速度，这对于检测精度有一定要求但实时性第一的入侵检测环境是很实用的。

参考文献

[1] Pan J J, Kwok J T, Yang Qiang, et al. Multidimensional Vector Regression for Accurate and Low-cost Location Estimation in Pervasive Computing[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(9): 1181-1193.

[2] Hardon D, Szedmak S, Shawe-Talyor J. Canonical Correlation Analysis: An Overview with Application to Learning Methods[J]. Neural Computation, 2004, 16(12): 2639-2664.

[3] Shawe-Talyor J, Cristianini N. Kernel Methods for Pattern Analysis[M]. Cambridge, UK: Cambridge University Press, 2004.

[4] Fei Ben, Liu Jinbai. Binary Tree of SVM: A New Fast Multiclass Training and Classification Algorithm[J]. IEEE Transactions on Neural Networks, 2006, 17(3): 696-704.

编辑　任吉慧