运用核Fisher鉴别分析和MPM分类器的入侵检测

陈振国，李冬艳

(华北科技学院计算机系 河北 三河 065201)

【摘要】为了提高分类器的正确率和减少训练时间，将特征提取技术与分类算法结合，提出了一种基于核Fisher鉴别分析和最小极大概率机算法的入侵检测算法。利用核Fisher鉴别分析技术提取关键特征，运用最小极大概率机对提取特征后的数据进行分类，采用离线数据集KDDCUP99进行实验。实验结果表明，该算法是可行和有效的，使分类性能和训练时间都得到了提高。

关 键 词 数据分类； 入侵检测； 核Fisher鉴别分析； 最小极大概率机； 网络安全
中图分类号 TP393 文献标识码 A

Intrusion Detection Based on Kernel Fisher Discriminant Analysis and Minimax Probability Machine Classifier

CHEN Zhen-guo，LI Dong-yan

(Department of Computer, North China Institute of Science and Technology Sanhe Hebei 065201)

Abstract To improve the performance of Minimax Probability Machine (MPM) in the detection rate and the training time, Intrusion Detection Based on Kernel Fisher Discriminant Analysis and Minimax Probability Machine Classifier (KFDA-MPM) algorithm is proposed which combines the feature extraction technology and classification algorithm. In this method, the KFDA is used to extract the optimal feature set and then the MPM is adopted to classify the optimization data. Results of the experiment using the Knowledge Discovery and Data Mining Cup 1999 (KDDCUP99) datasets indicate the effectiveness of the algorithm.

Key words data classification; intrusion detection; kernel Fisher discriminant analysis; minimax probability machines; network security

在网络飞速发展的今天，如何有效地发现新的入侵行为，对于保证系统和网络资源的安全非常重要。入侵检测作为网络安全措施的一个重要环节，也越来越受到人们的关注。

入侵检测可以看作是一个模式分类和识别问题，即通过对网络数据的分析，判断网络用户的状态，从而确定网络行为是否正常。近年来，为了获得更好的检测效果和提高检测系统的实时性，研究人员将智能学习算法和统计学理论(如免疫算法[1]、神经网络[2]、支持向量机[3-5]和分类器融合方法[6]等)应用于入侵检测，使检测的效果和检测速度都得到了改善和提高。但随着网络数据规模的不断增大，仅仅依靠分类算法已不能满足对入侵检测性能提升的需要。基于上述原因，本文选择对网络数据进行预先处理，利用特征分析技术提取入侵数据的关键特征，通过降低数据集的维数降低数据的复杂性，然后再使用分类算法对降维后的数据集进行训练和检测。实验结果表明该方法提高了对入侵的检测率，降低了分类器的训练时间。

本文选用核Fisher鉴别分析和最小极大概率机(Minimax Probability Machines)算法[7]进行入侵数据集合关键特征的提取和分类，并提出了一种基于核Fisher鉴别分析(Kernel Fisher Discriminant Analysis，KFDA)和最小极大概率机(Minimax Probability Machines，MPMs)的入侵检测算法。

1 基于KFDA-MPM的入侵检测

1.1 基于KFDA的特征抽取

设$X=\{x_i\}, i=1,2,\cdots,N$为训练样本集。$x_i^j \in X$ 代表类别 j 的第 i 个样本，$j=1,2,\cdots,K$；每个类别含有 N_j 个样本。经过非线性映射 ϕ，在特征空间H上应用Fisher鉴别准则函数：

$$J(\boldsymbol{w})=(\boldsymbol{w}^{\mathrm{T}}S_b^{\phi}\boldsymbol{w})/(\boldsymbol{w}^{\mathrm{T}}S_w^{\phi}\boldsymbol{w}) \quad (1)$$

式中

收稿时间：2007 – 06 – 10
作者简介：陈振国(1976 –)，男，硕士，讲师，主要从事智能信息处理、网络安全、数据挖掘等方面的研究.

$$S_b^{\phi}=\sum_{j=1}^{K}(N_j/N)(m_j^{\phi}-m^{\phi})(m_j^{\phi}-m^{\phi})^{\mathrm{T}}$$

$$S_w^{\phi}=(1/N)\sum_{j=1}^{K}\sum_{i=1}^{N_j}(\phi(x_i^j)-m_j^{\phi})(\phi(x_i^j)-m_j^{\phi})^{\mathrm{T}}$$

$$m^{\phi}=(1/N)\sum_{i=1}^{N}\phi(x_i)$$

$$m_j^{\phi}=(1/N_j)\sum_{i=1}^{N_j}\phi(x_i^j)$$

根据再生核理论，H空间中的任何解 $\boldsymbol{w}$ 都是由H空间中的样本所张成，即：

$$\boldsymbol{w}=\sum_{i=1}^{N}\boldsymbol{\alpha}_i\phi(x_i) \tag{2}$$

将式(2)代入式(1)得到:

$$J(\boldsymbol{\alpha})=\frac{\boldsymbol{\alpha}^{\mathrm{T}}K_b\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}K_w\boldsymbol{\alpha}} \tag{3}$$

式中

$$K_b=\sum_{j=1}^{K}(N_j/N)(M_j-M)(M_j-M)^{\mathrm{T}}$$

$$K_w=(1/N)\sum_{j=1}^{K}\sum_{i=1}^{N_j}(\eta_i^j-M_j)(\eta_i^j-M_j)^{\mathrm{T}}$$

$$M_j=((1/N_j)\sum_{i=1}^{N_j}k(x_1,x_i),\cdots,(1/N_j)\sum_{i=1}^{N_j}k(x_N,x_i))^{\mathrm{T}}$$

$$M=((1/N)\sum_{i=1}^{N}k(x_1,x_i),\cdots,(1/N)\sum_{i=1}^{N}k(x_N,x_i))^{\mathrm{T}}$$

$$\eta_i^j=(k(x_1,x_i^j),\cdots,k(x_N,x_i^j))^{\mathrm{T}}$$

$$k(x_i,x_j)=\phi(x_i)^{\mathrm{T}}\phi(x_j)$$

由此，在 K_w 非奇异的情况下，最大化式(3)的最优解向量 $\boldsymbol{\alpha}$ 为广义特征方程 $K_b\boldsymbol{\alpha}=\lambda K_w\boldsymbol{\alpha}$ 的前m个最大特征值所对应的特征向量，m=min(K–1,N)，则在H空间中，任何样本 $\phi(x)$ 在特征向量$\boldsymbol{w}$上的投影为：

$$\boldsymbol{w}\cdot\phi(\boldsymbol{x})=\sum_{i=1}^{N}\boldsymbol{\alpha}_i k(x_i,\boldsymbol{x}) \tag{4}$$

2.2　最小极大概率机

在一个二类分类问题中，设$\boldsymbol{x}$和$\boldsymbol{y}$代表随机矢量，则对应的均值矢量与协方差矩阵可分别表示为$\boldsymbol{x}$～$\left(\bar{x},\sum_x\right)$和 $\boldsymbol{y}$～$\left(\bar{y},\sum_y\right)$，其中“～”表示随机变量具有特定的均值与协方差矩阵，但其分布不是无约束的，且 $\boldsymbol{x},\bar{x},\boldsymbol{y},\bar{y}\in R^n$；$\sum_x,\sum_y\in R^{n\times n}$。

需要确定一个超平面 $\boldsymbol{a}^{\mathrm{T}}z=b$，其中$\boldsymbol{a},z\in R^n$，且 $b\in R$，使得它在特定均值与协方差矩阵分布的条件下获得两类最大分类概率P，并可归结为如下问题：

$$\max_{\boldsymbol{\alpha},\boldsymbol{a},b}\boldsymbol{\alpha}\quad \text{s.t}\ \ \inf P\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}\geq b\}\geq\boldsymbol{\alpha} \tag{5}$$

$$\inf P\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{y}\leq b\}\geq\boldsymbol{\alpha}$$

或

$$\max_{\alpha,a,b}\boldsymbol{\alpha}\ \ \text{s.t}\ \ 1-\boldsymbol{\alpha}\geq\sup P\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}\leq b\} \tag{6}$$

$$1-\boldsymbol{\alpha}\geq\sup P\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{y}\geq b\}$$

考虑式(6)中的条件约束，以及Bertsimas与Sethuraman的结果，则有：

$$\sup P\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{y}\geq b\}=\frac{1}{1+d^2}$$

式中　$d^2=\inf_{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{y}\geq b}(\boldsymbol{y}-\bar{y})\sum_y^{-1}(\boldsymbol{y}-\bar{y})$。

把原问题转换为如下的二次凸优化问题：

$$\min\left\|\sum\nolimits_x^{\frac{1}{2}}\boldsymbol{a}\right\|_2+\left\|\sum\nolimits_y^{\frac{1}{2}}\boldsymbol{a}\right\|_2\ \ \text{s.t}\ \ \boldsymbol{a}^{\mathrm{T}}(\bar{x}-\bar{y})=1$$

可以通过多种方式解决这个问题，其在最坏情况下的时间复杂度为$O(n^3)$，与传统的支持矢量机具有相同的时间复杂度。详细内容请参见文献[7]。

2.3　基于**KFDA-MPM**的入侵检测模型

利用KFDA提取关键特征集合，用MPM分类算法对优化后的数据进行分类，可构造出如图1所示的入侵检测模型。

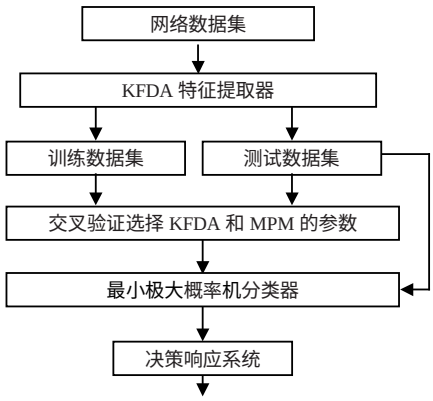

图1　基于KFDA-MPM的入侵检测模型

基于KFDA-MPM的入侵检测分为训练和检测两个阶段。

1) 训练阶段

(1) 根据给定的训练数据集，利用式(3)对基于核函数的Fisher准则进行训练，根据式(2)获取最优投影方向，用式(4)将训练数据映射到该投影方向。

(2) 将映射后的训练数据按照式(8)进行训练，并根据式(6)和式(7)得到 $\boldsymbol{\alpha}$ 和 d^2，并求出均值和协方差。

2) 检测阶段

将给定的测试数据集经过投影映射后进行入侵检测，根据检测的结果，决策响应系统可进行不同

的处理。

3 仿真结果及分析

本文仿真所使用的实验数据取自1999年DARPA的入侵检测评估项目KDDCUP99[8]。其中主要攻击划分为DOS(拒绝服务攻击)、R2L(远端未经授权的访问)、U2R(未经授权提升权限)和Probing(探针)四类[9]。

在样本库的选择上，本文仿真采用KDDCUP99 10%的数据集作为基准数据。在实验中，将该数据集分为Probe数据集、DoS数据集、R2L数据集和U2R数据集四个部分，其中Probe数据集、DoS数据集和R2L数据集分别由900条Normal样本和600条对应的攻击样本组成训练集。由于KDDCUP99训练集中U2R攻击的实例较少，因此U2R数据集由140条Normal样本和80条U2R样本组成训练集，测试集中的样本数量与训练集相同。

由于在KDDCUP99数据集所提供的数据中不仅包含数值数据，同时还有非数值数据，因此本文在进行检测前，对非数值数据进行量化。同时对数据集中的连续属性采用基于动态层次聚类的连续属性离散化算法[10]做离散化处理。

本文中选用SVM、MPM和KFDA-MPM三种方法在相同的数据集合上进行仿真实验，经过100次重复实验的平均结果如表1所示。

表1 KFDA-MPM与SVM、MPM性能比较

数据集	算法	正确检测率/(%)	训练时间/ms
Probe	SVM	79.1	4.318
	MPM	79.7	4.234
	KFDA-MPM	94.6	0.540
DoS	SVM	82.6	4.237
	MPM	82.4	4.359
	KFDA-MPM	96.5	0.612
U2R	SVM	83.4	3.672
	MPM	83.7	3.608
	KFDA-MPM	94.7	0.410
R2L	SVM	85.3	4.453
	MPM	85.5	4.389
	KFDA-MPM	91.9	0.548

根据三种算法在正确检测率与训练时间上的比较可以看出，采用KFDA-MPM算法进行入侵检测明显优于SVM和MPM算法，在检测率方面取得了良好的效果，并且训练时间也得到了改进。

MPM算法的时间复杂度与支持矢量机算法一样都与问题的规模相关，问题规模的大小，对MPM算法的性能有很大的影响。本文选用核Fisher鉴别分析技术提取数据的关键特征，减少了数据的维数，从而降低了问题的规模，提高了MPM算法的性能。但当问题规模特别庞大的情况时，本文算法对性能提升不明显。

4 结 束 语

本文采用MPM算法构造分类器，引入了核Fisher鉴别分析算法(KFDA)用于提取关键特征，并提出了基于KFDA-MPM的入侵检测算法。试验结果表明，该算法能够提高入侵的检测率，同时降低了分类器的训练时间，取得了理想的效果。下一步的研究工作将考虑采用分布式MPM算法进行网络入侵检测。

参 考 文 献

[1] HOFMEYR S A. The implications of immunology for secure systems design[J]. Computers & Security, 2004, 23 (6): 453-455.
[2] CANNADY J. Artificial neural networks for misuse detection[C]//In: Proceedings 1998 National Information Systems Security Conf (NISSC 98). Arlington: [s.n.], 1998.
[3] FUGATE M, GATTIKER J R. Computer intrusion detection with classification and anomaly detection using SVMs[J]. Int J Pattern Recognition Artif Intell, 2003, 17(3): 441-458.
[4] 饶 鲜, 董春曦, 杨绍全. 基于支持向量机的入侵检测系统[J]. 软件学报, 2003, 14(4): 798-803.
[5] 李 辉, 管晓宏, 咎 鑫, 等. 基于支持向量机的网络入侵检测[J]. 计算机研究与发展, 2003, 40(6): 799-807.
[6] GIACINTO G, ROLI F, DIDACI L. Fusion of multiple classifiers for intrusion detection in computer networks[J]. Pattern Recognition Letters, 2003, 24(12): 1795-1803.
[7] LANCKRIET G R G, GHAOUI L E, BHATTACHARYYA C, et al. Minimax probability machine[C]//Proceedings of Advances in Neural Information Processing Systems. Berkeley: Department of EECS University of California, 2002.
[8] RICHARD L, JOSHUA W, DAVID J, et al. The 1999 DARPA off-line intrusion detection evaluation[J]. Computer Networks, 2000, 34(4): 579-595.
[9] MAHONEY M V, CHAN P K. An analysis of the 1999 DARPA/Lincoln laboratory evaluation data for network anomaly detection[C]//In: Proceedings of the Sixth International Symposium on Recent Advances in Intrusion Detection, RAID 2003. Pittsburgh: Springer-Verlag, 2003.
[10] 苗夺谦. Rough Set理论中的连续属性的离散化方法[J]. 自动化学报, 2001, 27(3): 296-302.

编 辑 熊思亮