文章编号: 1001- 2486(2003) 06- 0016- 05

网络入侵检测系统中的数据缩减技术*

邹　涛, 孙宏伟, 田新广, 张尔扬
(国防科技大学电子科学与工程学院, 湖南 长沙　410073)

摘　要: 在进行事件分析之前, 网络入侵检测系统首先要面对数据缩减的问题。以 ANIDS 为背景, 分析了两种重要的数据缩减技术: 相关特征子集选择和特征再构造。提出了一种基于 Wrapper 方法的最优特征子集选取算法 SRRW。在考虑学习算法偏置的情况下, 通过识别强相关特征并引入约束, 能够更快地搜索并获得最优的相关特征子集。从特征再构造角度出发实现数据缩减, 并通过因子负荷量矩阵分析了原始特征之间的相关性。

关键词: 网络入侵检测; 数据缩减; 相关特征选取; 主成分分析
中图分类号: TP391　　文献标识码: A

Data Reduction in Network Based on the Intrusion Detection System

ZOU Tao, SUN Hong-wei, TIAN Xin-guang, ZHANG Er-yang
(College of Electronic Science and Engineering, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: NIDSs deal with the problem of data reduction before analyzing the events. Two important measures used in ANIDS are proposed: FSS and new feature construction. A novel algorithm named SRRW is put forward first, which can produce OFS by recognizing all strongly relevant features and restrict them in searching process. A feature construction method is used to get the OFS. The correlations between the original features can be analyzed by factor loading matrix.

Key words: NIDS; data reduction; relevant feature selection; PCA

分布式入侵检测系统是入侵检测技术发展的一个重要趋势。为了实现入侵的协同检测, 系统需要对数据进行存储[1]、分析及各组件之间的共享。如何有效地实现数据缩减, 在不丢失有用信息的前提下尽可能地压缩数据, 以减轻系统对传输、存储及分析的要求是 NIDS 需要研究的一个重要问题。本文以智能网络入侵检测系统(ANIDS)为背景, 分析其中所采用的相关特征子集选择和特征再构造两种数据缩减技术。

1　最优特征子集选择

1.1　概 述

最优特征子集选择(OFSS)直接影响到模式识别或者分类问题中学习器的性能。无关特征只会带来目标假设空间不必要的增大, 进而导致学习时间的增加。根据 PAC 计算学习理论, 学习样本复杂度满足以下公式:

$$m \geqslant \frac{1}{\varepsilon}[4\mathrm{lb}(2/\delta) + 8VC(H)\mathrm{lb}(13/\varepsilon)] \tag{1}$$

其中, m 为训练样本数目, H 为学习的假设空间, $VC(H)$为 H 的 VC 维[2], $VC(H)$随$|H|$单调递增。(1)式表明, 训练样例个数 m 要大于等于公式右侧取值时, 才能以 $1-\delta$ 的概率学习得到错误率不大于ε的目标概念。而假设空间 H 直接与特征的维数有关, 特征维数越大, 样本复杂度也越大。因此, OFSS 能降低样本复杂度, 有效提高机器学习的学习准确率[3,4]。另外, 在归纳学习之前做 OFSS, 还能使所获得的规则集更为简洁, 便于理解。

*　收稿日期: 2003- 05- 13
作者简介: 邹涛(1974—), 男, 博士生。

Terran Lane 为了减小入侵检测中对用户轮廓的存储要求,研究了实例选取和聚类等数据缩减技术[5,6]。S. Mukkamala 等利用 SVM 及 ANN 提出一种 PFRM(Performance-based Feature Ranking Method)算法来实现重要特征的鉴别和选取[7]。文献[8]采用相同数据对 PFRM 算法和基于遗传算法(GA)的 Wrapper 算法进行了分析和对比。

1.2　相关工作

特征子集选取算法根据其目标函数是否与学习算法有关可以分为两类。如果特征的选取与学习算法无关,则称为 Filter 方法;否则称为 Wrapper 方法。基于 Filter 和 Wrapper 的 OFSS 方法比较如图 1 所示。

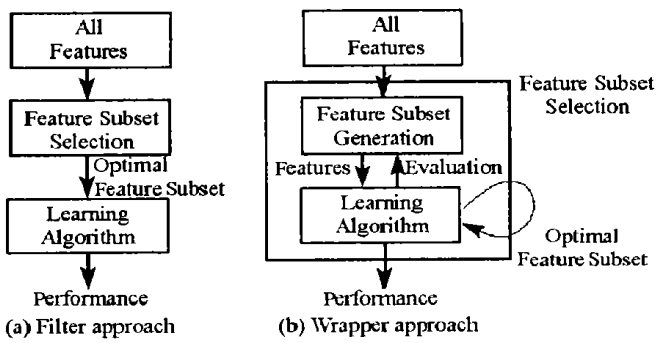

图 1　Filter 与 Wrapper 方法比较

Fig. 1　Filter approach and Wrapper approach

基于 Filter 的 OFSS 方法有 Relief 算法[9],以及利用互信息、信息增益决定相关特征的方法。由于 Relief 算法在计算某个特征的权值时独立于其他的特征,因此对于相关性很强的特征集会造成特征冗余[10]。信息增益算法通过计算每个特征的信息增益,选择信息增益值最大的特征组成最优特征子集。

Wrapper 方法一般采用学习算法的分类准确率来实现特征子集的评价。其计算效率虽然比 Filter 方法低,但是 OFSS 的搜索过程考虑到了用来构造分类器的学习算法的偏置,因此其选择得到的最优特征子集一般都要优于 Filter 方法。本文采用 Wrapper 方法,提出了一种基于强相关约束的新的 OFSS 算法。

1.3　基于强相关约束的 **Wrapper** 方法

首先定义训练样例集合 X 上特征的相关性。

定义　一个特征 f_k 为训练样例集合 $X\{f_1, f_2, f_3, \cdots, f_m, C\}$ 上的容错门限为 α 的强相关特征,当且仅当,$\exists (x_i \in X, x_j \in X)$,满足目标类别 $c(x_i) \neq c(x_j)$,且:

$$\begin{cases} f_l(x_i) \neq f_l(x_j), & l = k \\ f_l(x_i) = f_l(x_j), & \text{其他} \end{cases} \tag{2}$$

$$count(x_z) \geqslant \alpha, \quad z = i, j \tag{3}$$

其中,$f_l(x_i)$ 为样例 x_i 第 l 个特征的取值,$count$ 为 X 中满足上述条件的样例对数目。

OFSS 过程可以看成是一个最优化搜索问题[11]。如果特征的个数为 N,则有 2^N 种可能。常用的序列搜索算法包括:前向搜索 SFS、后向搜索 SBE、pq 序列搜索等。而遗传算法(GA)是解决最优化问题的有效手段。因此,我们在基于强相关约束的 Wrapper 方法(SRRW)中加入了遗传搜索,如图 2 所示。输入训练数据,首先对所有特征逐一做相关性测试,求取强相关特征子集。设总特征个数为 m,则搜索空间为 2^m;而若扫描得到 ml 个强相关特征,则 GA 的搜索空间将减小为原来的 $1/2^{ml}$。将 F 及约束条件送给 GA,做带约束的遗传搜索。GA 得到的特征子集送给学习算法,通过机器学习获得输出规则,并向 GA 返回分类准确率及特征个数以计算适应度函数。

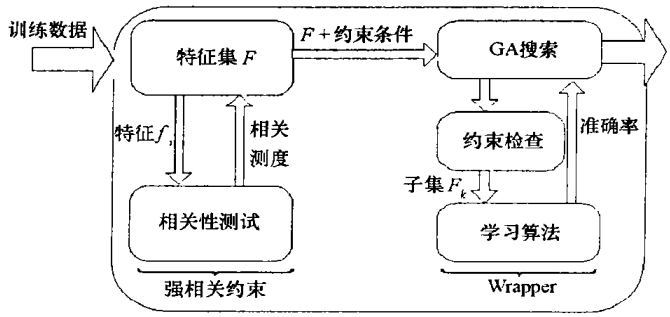

图 2 基于强相关约束的 Wrapper 方法 SRRW
Fig. 2 SRRW algorithm

SRRW 定义为一个三元组: $SRRW = (\alpha, RGA, MLA)$, 其中:

$$RGA = (R, C, E, P_0, M, \Phi, \Gamma, \Psi, T), \quad E = (L_{FS}, A_{ML}), \quad A_{ML} = (\beta)$$

其中, RGA: 带约束的遗传算法; β: 学习算法偏置; A_{ML}: 准确率; MLA: 学习算法; R: 约束算子; C: 编码方法; E: 适应度评价函数; P_0: 初始种群; M: 种群规模; Φ: 选择算子; Γ: 交叉算子; Ψ: 变异算子; L_{FS}: 编码为 1 的特征个数; T: 终止条件。

适应度函数的选择是 SRRW 算法的关键。基于式(1)的分析以及统计学习理论中关于推广性的界的结论, 经验风险 $R_{emp}(D_n)$ 和期望风险 $R(D_n)$ 之间以至少 $1-\eta$ 的概率满足

$$R(D_n) \leqslant R_{emp}(D_n) + \sqrt{\frac{VC(H)\{\ln[2n/VC(H)]+1\} - \ln(\eta/4)}{n}} \tag{4}$$

及期望风险最小化的考虑, SRRW 算法中的适应度函数采用了 $\min\_(L_{FS}, R_{emp}(D_n))$ 偏置。式(4)中 $VC(H)$ 为定义在实例空间 D_n 上的 H 的 VC 维、n 为训练集中样例的个数。

SRRW 算法不但通过强相关特征选取能够有效避免信息增益 OFSS 算法中对强相关特征的可能遗漏性问题, 而且由于在 RGA 中引入了强相关性约束, 减小了搜索空间。由于 Wrapper 方法每一轮循环都需要学习一次, 因此, 搜索空间的减小实际上是带来了学习次数的减少。而与 Filter 方法相比, Wapper 方法还避免了 OFS 可能与学习算法偏置不一致的问题。

2 特征再构造

数据缩减中 OFSS 技术的另外一种扩展是采用特征之间的组合来构造新的 OFS。主成分分析 PCA 即是通过对原特征集做 PCA 变换[12], 得到各特征互相独立的新的特征集; 取其中贡献率最大的部分特征构成 OFS。利用因子负荷量矩阵, 还可实现特征之间的相关性分析。

设有 n 个样本 $x^0_1, x^0_2, \cdots, x^0_n$, x^0_{ik} 表示样本 k 的第 i 个特征取值, 将 PCA_OFSS 应用于网络连接数据 $X\{f_1, f_2, f_3, \cdots, f_m\}$ 的步骤为:

(1)按公式(5)归一化样本:

$$x_{ik} = \frac{x^0_{ik} - \frac{1}{n}\sum_{k=1}^{n} x^0_{ik}}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(x^0_{ik} - \frac{1}{n}\sum_{k=1}^{n} x^0_{ik}\right)^2}}, \quad i = 1, 2, \cdots, m; k = 1, 2, \cdots, n \tag{5}$$

(2) 求 X 的特征相关矩阵, 并做奇异值 SVD 分解, 计算主成分;

(3) 计算各主成分方差 var、总方差 $total\_var$ 及分数方差 $frac\_var$;

(4) 选取满足 $frac\_var > min\_frac$ 的主成分, 构成新的 OFS;

(5) 计算因子负荷量矩阵, 矩阵元素为 $a_{ij} = \sqrt{\lambda_i} l_{ij}$, 其中 l_{ij} 为特征值 λ_i 对应特征矢量中的第 j 个元素;

(6) 利用因子负荷量矩阵做聚类分析, 得到原始特征之间的相关性。

3　实验数据及结果

实验采用 MIT 实验室发布的 DARPAR 98 入侵检测标准数据集。该数据集中包含正常网络连接及 land 攻击、neptune 攻击、猜测口令、端口扫描等 22 种攻击方式。

表 1　SRRW 算法参数表

Tab. 1　Parameter table of SRRW

参数	α	β	C	m	n	P_0	M	Φ	Γ	φ	T
取值描述	1	k-DNF	0, 1	41	50 000	random +	40	RWS	SP	RM	50 代
		MDL				restriction			0.65	0.015	

表 1 为SRRW 算法参数表。表 2 为OFSS结果。41 个特征中符号特征占 3 个, 数值特征 38 个。α = 1 时的强相关特征搜索结果为f_5、f_8。

表 2　OFSS结果

Tab. 2　Result of OFSS

	特征数	训练准确率	数据大小	缩减比例	预测错误率
全特征集	41	100%	100%	——	7.59% / 12.62%
Wrapper	23	100%	59.1%	40.9%	4.13% / 7.58%
SRRW	15	100%	41.2%	58.8%	3.66% / 5.81%
SRRW_OFS	$\{f_1, f_3, f_4, f_5, f_7, f_8, f_9, f_{28}, f_{31}, f_{33}, f_{35}, f_{37}, f_{39}, f_{40}, f_{41}\}$				

其中, Wrapper 是仅仅采用训练准确率作为适应度函数时得到的结果。预测错误率包含了分别在两组测试集上的测试性能。从实验结果可以看出, GA 的适应度函数在考虑了算法的期望风险最小化后得到的 SRRW 算法具有更好的数据缩减和预测分类能力。

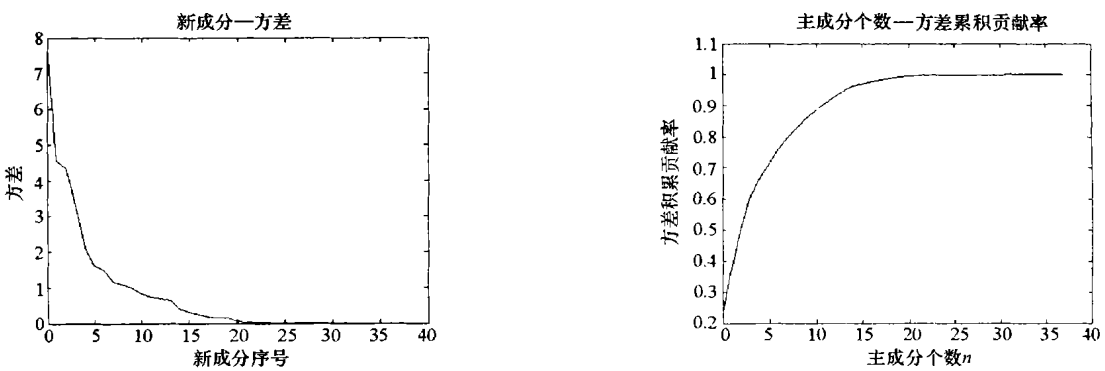

图 3　(a) 新成分 — 方差图; (b) 方差累积贡献率图

Fig. 3　(a) New_components — variance; (b) Fractional variance contribution curve

PCA 实验数据: n = 2000; m = 38。由图 3(b) 可见, 只需要 20 个主成分即可使方差累积贡献率达到 99.3633%。按公式计算 20 个主成分的因子负荷量矩阵, 并根据此矩阵用一阶 Minkowski 距离度量的重心层次聚类方法(非一致系数门限取为 0.9) 分析原始特征之间的相关性, 得到表 3。

表3 特征相关聚类表

Tab.3 Result of the features clustering

簇	1	2	3	4	5	6	7	8	9	10	11	12	13	14
i of	1.27	4,12,15,17	20,21	22,23	24,25	6,10,11	8	7	3	29	5	28	2	16
f_i	32	18,26,30,31	33	35,36	37,38	13,14		19	9			34		

4 结束语

ANIDS中的网络数据缩减手段有很多种。本文仅针对OFSS和PCA做出分析,提出了一种新的基于强相关约束的Wrapper方法——SRRW。从实验结果可以看出,SRRW算法不但能够有效地实现数据缩减,同时还能够有助于减小模型的期望风险。与从特征再构造的角度实现数据缩减的PCA相比,SRRW保留了特征的原始物理意义,使得由后续的机器学习算法得到的规则集合更易于理解。PCA方法也具有较好的数据缩减性能,在方差累积贡献率达到99.3633%时的数据缩减比例接近50%;并且可以通过对因子负荷量矩阵的分析获得原始特征之间的相关性,可作为聚类等无监督学习算法的前期处理。

参考文献:

[1] Tim B. Multisensor Data Fusion for Next Generation Distributed Intrusion Detection Systems[C]. In Proceedings 1999 IRIS National Symposium on Sensor and Data Fusion, May 1999.

[2] Blumer A, Littlestone N. Learning Faster than Promised by the VapnikChervonenkis Dimension [J]. Discrete Applied Mathematics, 1989, 24: 47-53.

[3] Langley P. Elements of Machine Learning [M]. Morgan Kaufmann, Palo Alto, CA., 1995.

[4] 边肇祺. 张学工, 等. 模式识别[M]. 北京: 清华大学出版社, 1999.

[5] Lane T, Brodley C E. Temporal Sequence Learning and Data Reduction for Anomaly Detection [C]. ACM Transactions on Information and System Security, 1999, 2(3): 295-331.

[6] Lane T, Brodley C E. Data Reduction Techniques for Instance-based Learning from Human/Computer Interface Data[C]. Proceedings of the Seventeenth International Conference on Machine Learning, 2000: 519-526.

[7] Mukkamala Srinivas, Sung Andrew, Abraham Ajith. Identifying Key Variables for Intrusion Detection Using Soft Computing Paradigms [C]. St. Louis, MO, USA: The IEEE International Conference on Fuzzy Systems FUZZ-IEEE' 03, 2003.

[8] 邹涛, 孙宏伟, 等. 入侵检测系统中两种审计数据缩减技术的比较与分析[J]. 计算机应用, 2003, 23(7).

[9] Kira K, Rendell L A. The Feature Selection Problem: Traditional Methods and a New Algorithm[C]. Proceedings of the 10th National Conference on Artificial Intelligence, CA, July 1992.

[10] Bradley P S. Mathematical Programming Approaches to Machine Learning and Data Mining[D]. University of Wisconsin, Madison, 1998.

[11] Langley P. Selection of Relevant Features in Machine Learning[C]. In Proceedings of the AAAI Fall Symposium on Relevance, AAAI Press, 1994.

[12] 夏绍玮, 杨家本, 杨振斌. 系统工程概论[M]. 北京: 清华大学出版社, 1995.