基于特征选择和多分类支持向量机的异常检测

张晓惠，林柏钢
（福州大学 数学与计算机科学学院，福建 福州 350108）

摘 要：现有大部分的异常检测系统都是把数据分成正常和异常两类，这样可能会丢失重要信息。特征选择的目的是减少异常检测冗余特征的同时，高度保持和原始特征的一致性。实现了特征选择和多分类支持向量机的异常检测技术。采取粗糙集、SVDF、LGP、MARS 相结合的特征选择方法。同时利用多分类支持向量机把数据分成五类。通过实验分析，表明 DoS 攻击相对于其他 3 种攻击的漏报率是最高的。

关键词：异常检测；粗糙集；支持向量机；多类分类；特征选择
中图分类号：TP309.5 **文献标识码**：B **文章编号**：1000-436X(2009)10A-0068-06

Anomaly detection based on feature selection and multi-class support vector machines

ZHANG Xiao-hui, LIN Bo-gang
(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract: The most of Intrusion detection systems divided data into two classes, which were normal and abnormal, so that it might lose some important information. The goal of feature selection was to decrease the redundant features for anomaly detection, and maintain the same high accuracy as the original features. It proposed an anomaly intrusion detection technique based on feature selection and multi-class support vector machines(SVM). The feature selection method merged RS, SVDF, LGP and MARS. Then, data was divided into five classes by the multi-class SVM. The experimental results demonstrate that the false positive rate of DoS is the highest one among four methods.

Key words: anomaly detection; rough set; support vector machine; multi-class; attribute selection

1 引言

入侵检测技术作为防火墙之后的第二道安全闸门，在不影响网络性能的情况下，防止或减轻来自网络攻击的威胁。从检测技术上来看[1]，入侵检测可以分为误用检测和异常检测。误用检测[2]是指运用已知的攻击方法，按照事先定义好的入侵模式，来判断这些入侵行为是否曾经出现，以此来判断是否出现攻击。而异常检测将偏离正常用户行为视为入侵嫌疑。由于异常检测可以检测到针对系统的未知攻击，于是成为入侵检测技术研究的重点。目前，针对异常检测主要采用机器学习方法，包括神经网络、遗传算法、贝叶斯模型、隐马尔可夫模型、数据挖掘、支持向量机、免疫学等[1]。

特征选择是在给定的原始数据集中选择其中重要的数据特征，减少数据维数，并且同时保留分类信息。文献[3]提出入侵检测的特征选择算法，说明经过特征选择的入侵检测系统可以显著缩短其检测时间。支持向量机（SVM, support vector machine）[4~6]是建立在统计学习理论基础上的一种机

收稿日期：2009-08-18
基金项目：福建省科技厅专项项目资助（2007F5071）
Foundation Item: The Sience and Technology Development Project of Fujian (2007F5071)

器学习方法，能较好地解决小样本学习问题，同时具有很好的泛化能力。SVM 用于入侵检测系统[7]，实验证明能比较好的检测入侵行为。由于 SVM 本身算法的限制，在训练样本规模比较大时，会占用较大空间的内存，而且耗时长。目前有关于特征选择和支持向量机在入侵检测方面的研究，但是大多是采用支持向量机的二分类方法，即把数据分为正常行为和入侵行为。分类粒度过粗，可能会丢失其中的一些信息，比如哪类攻击产生的虚警率最高等。

针对上述方法的某些不足，本文实现了一种基于特征选择和多分类支持向量机异常检测方法，该异常检测的主要机理，采用粗糙集、SVDF（support vector decision function）、LGP（linear genetic programming）和 MARS（multivariate adaptive regression splines）技术相结合的方法提取数据特征，然后采用多分类支持向量机算法判断入侵行为。

2　特征选择基本算法

2.1　决策表离散化

用粗糙集理论进行属性约简的时候，要求决策表中的值用离散数据表达。如果某些属性值域为连续值，则必须在算法运行前进行离散化。常用的离散化算法有 Naïve Scaler 算法、Semi Naïve Scaler 算法、等距离划分法、等频率划分法、Boolean 逻辑和Rough集理论相结合的离散化算法及基于属性重要性的离散化算法等。本文采用 Naïve Scaler 算法对原始数据集进行离散化，具体算法如下[8]。

对每一属性 $a \in C$，进行下面的过程：

1) 根据 $a(x)$ 的值，从小到大排列各个记录 $x \in U$；

2) 从上到下扫描，设 x_i 和 x_j 代表 2 个相邻的记录，若 $a(x_i) = a(x_j)$，则继续扫描；否则，若 $d(x_i) = d(x_j)$，也就是说决策相同，继续扫描；否则，得到一个断点 $c, c = (a(x_i) + a(x_j))/2$。也就是说在属性值和决策值都不相同的情况下得到一个断 c。

2.2　基于粗糙集理论的属性约简

粗糙集理论[9]（RST, rough set theory）是 Pawlak 于 1982 年提出的一种新的处理模糊和不确定性知识的数学理论和工具，属性约简是粗糙集理论的核心问题之一。

给定一个四元组的信息系统 $S=(U,C\cup D,V,f)$，其中 U 是给定网络连接的数据集，为一个非空的有限集合，C 是从网络连接中抽取的 41 个特征集，$D=\{d\}$ 为决策属性集。$V=\cup V_r$ 是特征的取值范围构成的集合，其中 V_r 是特征 r 的值域。决策属性 D 的取值范围为{1，2，3，4，5}，其中 1 表示正常的网络连接，2 表示 DoS 攻击，3 表示 U2R 攻击，4 表示 R2L 攻击，5 表示 Probe 攻击。$f: U \times R \to V$ 是信息函数，它指定 U 中每一个对象各个特征的取值。$H(D|C)$ 为集合 D 相对应集合 C 的条件信息熵，$Core_D(C)$ 表示 C 中所有 D 不可省略关系的集合，$SGF(a,C,D)$ 为属性 $a(a \in C)$ 相对于特征集合 C 对于决策属性集和 D 依赖性的重要度[10]。

本文采用基于条件信息熵的特征约简算法，算法描述如下[8]：

输入：网络特征信息表 $S=(U,C \cup D,V,f)$

输入：表 S 的一个特征约简 B

步骤 1：计算 $H(D|C)$

步骤 2：令 $Core_D(C)=\varnothing$；

{（for $\forall a \in C$）计算 $SGF(a,C,D)$；

if $SGF(a,C,D)$ >0 then $Core_D(C) = Core_D(C) \cup \{a\}$ }

步骤 3：if $H(D|Core_D(C)) = H(D|C)$ then

{ $B = Core_D(C)$ 为最小约简，转步骤 4}

else

{ 令 $B = Core_D(C)$；

while $H(D|B) != H(D|C)$ do

{for $\forall a_i \in C-B$计算 $H(D|B\cup\{a_i\})$

$a_j = \min\{a_i \mid H(D|B\cup\{a_i\})\}$；

$B = B \cup \{a_j\}$；

计算 $H(D|B)$；

}

}

步骤 4：输出特征约简 B

2.3　几种其他的特征选择算法

文献[11]对 3 种算法数据特征提取算法进行了研究，用 KDD CUP 99 数据进行了实验，这 3 种算法分别是：SVDF（support vector decision function）、LGP（linear genetic programming）和 MARS（multivariate adaptive regression splines）。实验表

明，存在属性子集对决策是很重要的。

3 支持向量机算法

3.1 支持向量机的基本原理[12]

支持向量机（SVM）经常用于各种分类和预测，它能使错误的检测率减小到最小。支持向量机最初是用于处理量分类问题。基本原理是用一个由一定数量的支持向量决定的超平面来分类数据。支持向量就是一个训练数据的子集，该子集通常被用于定义二类数据的边界。在无法用支持向量机分离二类问题的情况下，它就用核函数将输入数据映射到高维数据空间，然后在高维数据空间解决这个分类问题。在高维特征空间中，可以用线性超平面来分离数据集。

数学上，线性边界可以表示为

$$\omega^T x + b = 0 \tag{1}$$

用训练数据来估计一个函数 $f: R^n \to \{\pm 1\}$。用 $x \in A, y = 1$ 表示 A 类点，用 $x \in B, y = -1$ 表示 B 类点，$(x_i, y_i) \in R^n \times \{\pm 1\}$。如果训练数据是线性可分的，那么就存在一对 $(\omega, b) \in R^n \times R$ 使得：

$$\omega^{\mathrm{T}} x + b \geqslant 1 \quad (x \in A) \tag{2}$$

$$\omega^{\mathrm{T}} x + b \leqslant -1 \quad (x \in B) \tag{3}$$

其决策函数是：

$$f_{\omega,b}(x) = sign(\omega^{\mathrm{T}} x + b) \tag{4}$$

其中，ω 为权重向量；b 为偏离值。不等式约束式（2）和式（3）可合并成：

$$y(\omega^{\mathrm{T}} x + b) \geqslant 1 \quad x \in (A \cup B) \tag{5}$$

此时分类间隔为 $2/\|\omega\|$，使间隔最大等价于使 $\|\omega\|^2$ 最小。满足条件（1）且使 $\|\omega\|^2/2$ 最小的分类面就叫作最优分类面。转化为优化问题就是：

$$\min \Phi(\omega) = \|\omega\|^2 / 2 \tag{6}$$

约束条件是 $y(\omega^{\mathrm{T}} x + b) \geqslant 1$ (7)

3.2 多分类支持向量机

上面介绍的是支持向量机的二分类问题，然后在很多情况下，要将样本数据集分为多类。例如在入侵检测系统中，样本集可以细分为正常类、DoS、U2R、R2L、Probe。这就要对支持向量机进行改进，下面介绍几种方法可以实现多分类支持向量机[13]。

1) “一对一”方式：若将数据集分为 k 类，将 k 类分类分解成 $k(k-1)$个二分类，每个二分类只将一类与另一类数据分开，这 $k(k-1)$个二类分类的判决函数组合起来可以形成 k 类分类的判决函数。当一个样本数据输入时，分别用这 $k(k-1)$个二类分类的判决函数对其进行判断，每次将样本判为某一类，最终得到判入次数最多的类为样本的最后结果。

2) “一对多”方式：若将数据集分成 k 类，将 k 类分类分解成 k 个二类分类，每个二类分类包含所有样本，第 i 个二类分类将第 i 类和其他类分开，这 k 个二类分类的判决函数组合起来就可以形成 k 类分类的判决函数。当一个样本数据输入时，分别用这 k 个二类分类的判决函数进行判断，如果只有第 i 个输出是属于第 i 类，而其他的判决函数都输出为其他类，则判断该样本属于第 i 类，否则不处理。

4 异常检测模型

对入侵检测系统的异常检测模块而言，它接收来自前一模块的数据结果，进行异常分析。本文的异常检测模块接收到数据后，首先，进行数据预处理，对数据进行离散化；然后，对数据进行特征选择；最后，对选择后的数据结果进行数值化和归一化处理后，进行 SVM 检测。异常检测模型如图 1 所示。

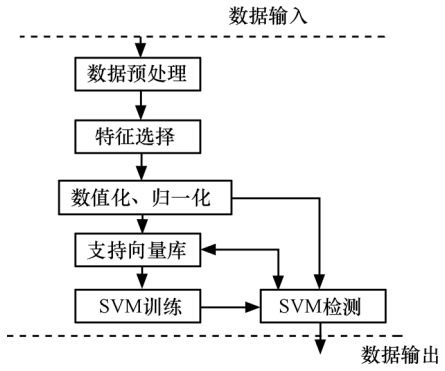


图 1 异常检测模型

5 实验结果与分析

本文实验中采用来自 1999 年 DARPA 为 KDD 竞赛提供的一个异常检测的标准数据（http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html）。该数据集中大约有 500 万条训练数据和 30 万条测试数据。测试数据包含了 38 种不同攻

击，总的可以分为 4 种类型的攻击：Probe（扫描与探测）、DoS（拒绝服务攻击）、U2R（对本地超级用户的非法访问）和 R2L（未经授权的远程访问）。

在数据集中一个连接记录即为一个完整的会话，每条连接记录由四类属性集组成：基本属性集、内容属性集、基于时间的属性集、基于主机的属性集，总共有 41 个属性，其中包含有 9 个离散属性和 32 个连续属性。这 9 个离散属性是：protocol-type、service、flag、land、logged-in、root-shell、su-attempted、is-hot-login 和 is-guest-login。41 个属性的具体说明如表 1 所示。

表 1　　网络数据特征

标号	网络数据特征	标号	网络数据特征
1	duration	22	is_guest_login
2	protocol_type	23	count
3	service	24	srv_count
4	flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	land	28	srv_rerror_rate
8	wrong_fragment	29	same_srv_rate
9	urgent	30	diff_srv_rate
10	hot	31	srv_diff_host_rate
11	num_failed_login	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	num_compromised	34	dst_host_same_srv_rate
14	root_shell	35	dst_host_diff_srv_rate
15	su_attempted	36	dst_host_same_srv_port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_rerror_rate
21	is_host_login		

由于原始数据集过于庞大，在本实验中选择小部分的数据集作为实验数据。随机抽取 17 071 条记录，其中正常记录 8 833 条、DoS 有 3 926 条、U2R 有 520 条、R2L 有 1 126 条、Probe 有 2 666 条，也就是说正常数据和入侵数据大约各占 50%。在随机抽取的 17 071 条记录中，选择其中的 70%作为训练数据，剩下的 30%作为测试数据。

5.1　特征选择结果

对离散化的测试数据用粗糙集进行属性约简，得到 26 个属性集，它们分别是：3、4、5、6、7、10、13、14、16、23、24、25、27、28、29、30、31、32、33、34、35、36、37、38、39、40，记为特征选择方法 1。其中 3、4、5、24、32、33 这 6 个属性最为重要，若只选用 6 个重要属性，记为特征选择方法 2。

文献[14]对 SVDF、LGP、MARS 和粗糙集算法进行了比较，分别选择 6 个属性（粗糙集算法用提取到的 6 个重要的属性）。用 SVDF 算法选择特征，得到的属性子集是：2、4、5、23、24、33；用 LGP 算法得到的属性子集是：3、5、12、27、31、35；用 MARS 算法选择到的属性子集是：5、24、27、33、34、35。实验证明，分别用这 4 个属性子集来判断入侵，都能取得较好的效果。4 个属性子集如图 2 所示。

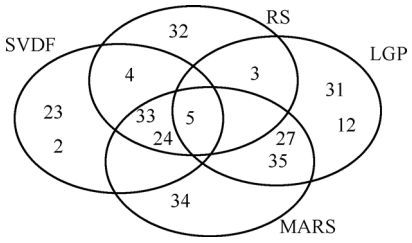


图 2　4 个属性子集比较

由图 2 可知，这 4 个属性子集提取出来的特征共有 13 个。可以认为这 13 个属性对于决策是重要的。这 13 个属性分别是：2、3、4、5、12、23、24、27、31、32、33、34、35。4 种算法结合的特征选择算法记为特征选择方法 3。

5.2　属性的数值化

经过特征选择后的属性集减小了原始数据集的维数。特征选择后，本文采用 SVM 多类分类算法判断是否入侵。由于 SVM 只能处理数值化的数据，而我们的数据样本中存在一些属性是字符串型的，所以必须对特征属性进行预处理。比如 protocol_type 属性的取值有：tcp、icmp、udp 等，把 tcp 对应为整型 1、icmp 为 2、udp 为 3、其他协议类型为 4。service 属性的取值有 auth、bgp、courier、csnet_ns、ctf、daytime、discard、domain、domain_u、echo、eco_i、ecr_i、efs、exec、finger、ftp、ftp_data、gopher、hostnames、http、http_443、imap4、IRC、iso_tsap、klogin、kshell、ldap、link、login、mtp、

name、netbios_dgm、netbios_ns、netbios_ssn、netstat、nnsp、nntp、ntp_u、other、pop_2、pop_3、printer、private、remote_job、rje、shell、smtp、sql_net、ssh、sunrpc、supdup、systat、telnet、time、uucp、uucp_path、vmnet、whois、Z39_50，它们依次对应 1 到 59。flag 属性取值有：OTH、REJ、RSTO、RSTOSO、RSTR、S0、S1、S2、S3、SF、SH，依次对应 1 到 11。

5.3　SVM 分类结果

我们知道 SVM 在输入样本集规模比较大的时候会占用较大的内存空间，并且耗时比较长，这使得 SVM 在实际应用中受到了限制。在入侵检测系统中如果直接采用 41 维属性作为 SVM 输入，则消耗的时间、空间资源过于庞大，不适合用在实时入侵检测系统中。运用上述介绍的 4 种特征选择算法，分别选择 6 个属性，4 种特征选择算法是：RS、SVDF、LGP、MARS。然后采用 SVM 二类分类算法。例如检测正常类的实验中，把数据分成正常和异常；在检测 DoS 攻击的实验中，把数据分成 DoS 攻击和非 DoS 攻击。实验表明，上述 4 种算法选择的属性对决策的影响很大。实验结果如表 2 所示。

表 2　　4 种特征选择算法的检测精度

类型	MARS/%	SVDF/%	LGP/%	RS/%
正常	84.9	80.83	94.16	89.84
DoS	99.77	99.71	99.8	99.34
Probe	99.5	99.12	100	99.63
U2R	100	100	60	100
R2L	98.57	99.31	100	100

由表 2 知，分别用 4 种算法选择出来的 6 个属性，对入侵检测的决策都起着重要作用。于是有了上述提出的特征选择方法 3。

如果采用特征选择方法 2 选择特征，然后采用 SVM 分类算法，得到较好的效果，但是它耗时长；如果采用特征选择方法 2，虽然在训练时间和检测时间上得到了提高，但是它检测的准确率有明显的下降；若采用特征选择方法 3，不仅减少了训练时间和检测时间，而且在检测效果近乎与方法 1 相同。若采用传统的 SVM 二类分类算法，也就是把数据分成正常和异常两大类，在检测精度上比 SVM 多类分类效果好，实验结果如表 3 所示。若采用 SVM 多类分类算法，把数据分成正常、DoS、U2R、R2L 和 Probe 五类，实验结果如表 4 所示。

表 3　　3 种不同选择算法的检测精度（分两类）

特征选择方法	训练时间/s	检测时间/s	检测精度/%
方法 1	1.8	0.8	99.84
方法 2	0.9	0.4	91.72
方法 3	1.2	0.5	99.56

表 4　　3 种不同选择算法的检测精度（分五类）

特征选择方法	训练时间/s	检测时间/s	检测精度/%
方法 1	8.3	3.9	92.62
方法 2	4.6	2.5	86.72
方法 3	5.8	3.1	92.04

采用 SVM 多类分类算法相对于 SVM 二分类算法的检测时间是比较长的，这是由 SVM 多类分类算法本身的限制。由表 4 知，方法 3 在训练时间和检测时间均减小的情况下，检测精度没有明显降低。用特征选择方法 3 和多分类 SVM 相结合的异常检测模块，系统总的误报率和漏报率分别是 3.11%和 9.86%。本文提出的入侵检测系统的误报率较低，得到了较好的效果，而系统的漏报率相对而言比较高。各类的具体检测结果如表 5 所示。

表 5　　5 种类别的检测精度

实际数据包类型	检测结果类型				
	正常	DoS	U2R	R2L	Probe
正常	96.89%	0.16%	0.28%	1.99%	0.68%
DoS	15.33%	82.53%	0.18%	1.71%	0.25%
U2R	9.62%	0	63.46%	26.92%	0
R2L	8.08%	0.36%	0.62%	90.94%	0
Probe	2.59%	1.35%	0.04%	0	96.02%

表 5 中的列标题代表实际数据包类型，而行标题代表检测结果类型。例如：0.16%代表正常数据被检测为 DoS 攻击的概率是 0.16%。由表 5 可知，DoS 攻击产生的漏报率最高。

6　结束语

本文结合 4 种特征选择算法，对 KDD CUP 99 数据进行特征选择，并采用多分类支持向量机对数据进行学习、训练、分类。实验证明，该方法取得了较好的效果，并且误报率比较低。在漏报率上，DoS 的漏报率是最高的。经过深入分析，有些指标的颗粒度还可以细化。下一步工作的重点是降低漏报率，提高异常检测的准确度。

参考文献：

[1] WANG Y. A Hybrid Intrusion Detection[D]. The Degree of Doctor of Philosophy Lowa State University, 2004.

[2] ILGUN K, KEMMERER R A, PORRAS P A. State transition analysis: a rule-based intrusion detection approach[J]. IEEE Transaction on Software Engineering, 1995, 21(3): 181-199.

[3] ANDREW H S. Identify important features for intrusion detection using support vector machines and neural networks[A]. IEEE Proceedings of the 2003 Symposium on Application and the Internet[C]. 2003.209-217.

[4] VAPNIK V. Statistical Learning Theory[M]. New York: Springer, 1995.

[5] CORTS C, VAPNIK V. Support vector networks[J] .Machine Learning, 1995, 20: 273 - 297.

[6] CRISTIANINI N, SHAWE-TAYLOR J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods[M]. Cambridge: Cambridge University Press, 2000.

[7] 柏海滨, 李俊. 基于支持向量机的入侵检测系统的研究[J]. 计算机技术与发展, 2008, 18(4): 137-143.
BAI H B, LI J. Research of intrusion detection system based on support vector machine[J].Computer Technology and Development,2008, 18(4): 137-143.

[8] 张义荣, 鲜明, 肖顺平等. 一种基于粗糙集属性约简的支持向量异常入侵检测方法[J]. 计算机科学, 2006, 33(6): 64-68.
ZHANG Y R, XIAN M, XIAO S P, *et al*. An anomaly intrusion detection technique of support vector machine based on rough set attribute reduction[J].Computer Science,2006, 33(6): 64-68.

[9] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11 (5):341-356.

[10] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.
WANG G Y. Rough Set Theory and Knowledge Acquisition[M]. X'an:Xi'an Jiaotong University Press, 2001.

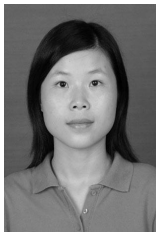
[11] SUNG A H, MUKKAMALA S. The Feature Selection and Intrusion Detection Problems[M]. Springer Verlag Lecture Notes Computer Science 3321, 2004. 468-482.

[12] 肖海军, 王小非, 洪帆等. 基于特征选择和支持向量机的异常检测[J]. 华中科技大学学报, 2008, 36(3): 99-102.
XIAO H J, WANG X F, HONG F, *et al*. Attribute selection-based and support vector machine for anomaly detection[J]. J Huazhong Univ of Sci&Tech,2008, 36(3): 99-102.

[13] 徐勋华, 王继成. 支撑向量机的多类分类方法[J]. 微电子学与计算机, 2004, 21(10): 149-152.
XU X H, WANG J C. Support vector machine for multi-class classification[J]. Microelectronics and Computer,2004, 21(10): 149-152.

[14] ZAINA A, MAAROF M A, SHAMSUDDIN S M. Feature selection using rough set in intrusion detection[A]. TENCON 2006[C]. 2006.

作者简介：



张晓惠（1984-），女，福建连江人，福州大学硕士生，主要研究方向为网络安全与智能技术。

林柏钢（1953-），男，福建福州人，福州大学教授、博士生导师，主要研究方向为数据通信与网络、编码与密码、信息安全。