ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.7, July 2007, pp.1639−1651
DOI: 10.1360/jos181639

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于特征选择的轻量级入侵检测系统*

陈 友[1,2+], 程学旗[1], 李 洋[1,2], 戴 磊[1,2]

[1](中国科学院 计算技术研究所,北京 100080)

[2](中国科学院 研究生院,北京 100049)

Lightweight Intrusion Detection System Based on Feature Selection

CHEN You[1,2+], CHENG Xue-Qi[1], LI Yang[1,2], DAI Lei[1,2]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: Phn: +86-10-62600949, E-mail: chenyou04@mails.gucas.ac.cn

Chen Y, Cheng XQ, Li Y, Dai L. Lightweight intrusion detection system based on feature selection. *Journal of Software*, 2007,18(7):1639−1651. http://www.jos.org.cn/1000-9825/18/1639.htm

Abstract: The intrusion detection system based on feature selection deals with huge amount of data which contains redundant and noisy features causing slow training and testing process, high resource consumption as well as poor detection rate. Feature selection, therefore, is an important issue in intrusion detection and it can delete redundant and noisy features. In order to improve performances of intrusion detection system in terms of detection speed and detection rate, a survey of intrusion detection system based on feature selection is necessary. This paper introduces the concepts and algorithms of feature selection, surveys the existing lightweight intrusion detection systems based on feature selection algorithms, groups and compares different systems in three broad categories: filter, wrapper, and hybrid. This paper concludes the survey by identifying trends of feature selection research and development in intrusion detection system. Feature selection is not only useful for intrusion detection system, but also helpful to provide a new research direction for intrusion detection system.

Key words: feature selection; lightweight intrusion detection system; filter; wrapper; hybrid

摘 要: 基于特征选择的入侵检测系统处理的数据含有大量的冗余与噪音特征,使得系统耗用的计算资源很大,导致系统训练时间长、实时性差,检测效果不好.特征选择算法能够很好地消除冗余和噪音特征,为了提高入侵检测系统的检测速度和效果,对基于特征选择的入侵检测系统进行研究是必要的.综述了这一领域的研究进展,从过滤器、封装器、混合器 3 种模式对基于特征选择的轻量级入侵检测系统进行分类比较,分析和总结各种系统的优缺点以及它们各自适用的条件,最后指出入侵检测领域特征选择的发展趋势.特征选择不仅可以提升入侵检测系统的性能,而且使得对入侵检测的研究向特征提取算法的方向转移.

关键词: 特征选择;轻量级入侵检测系统;过滤器;封装器;混合器

中图法分类号: TP393 文献标识码: A

* Supported by the National Basic Research Program of China under Grant No.2004CB318109 (国家重点基础研究发展计划(973)); the National Information Security 242 Project of China under Grant No.2005C39 (国家 242 信息安全计划)

Received 2006-10-08; Accepted 2007-03-26

入侵检测是一种通过收集和分析被保护系统信息,从而发现入侵的技术.它的主要功能是对网络和计算机系统进行实时监控,发现和识别系统中的入侵行为或企图,给出入侵警报.可将入侵检测看作是区别系统状态是“正常”还是“异常”的二分类问题[1].对入侵检测系统的要求首先是正确性,其次是实时性.只有检测速度快,才能及时处理网络中传输的海量数据,不会因为速度慢而丢失信息、造成漏警,更能及时采取措施,将入侵带来的损失降到最低.随着网络的高速提升,入侵检测系统面临的一个主要问题是检测速度低、负荷大,来不及处理网络中传输的海量数据,并且这个问题变得越来越严重.检测速度已成为入侵检测系统实时性要求的一个重要指标,如何在保证检测正确性的前提下开发出检测速度快的轻量级入侵检测系统,成为当前研究的热点.很多研究者通过特征选择来解决这个问题,提取和处理的特征数目过多是导致速度下降的主要原因之一.特征和分类器性能之间并不存在线性关系,当特征数量超过一定限度时,会导致分类器性能变差.实际上,有些特征没有包含或者包含极少的系统状态信息,它们对检测结果几乎没有影响.所以,特征选择——去除冗余特征,保留能够反映系统状态的重要特征是提高检测速度的一种有效方法.在尽量不降低分类精度的前提下降低特征空间的维数,就是特征选择,即依据一定的评价函数从原始特征集中选择与输出结果有关的或重要的特征子集.特征选择有两个关键的问题,即选择合适的评价函数和与之相适应的搜索方法.

1 特征选择的数学模型及一般化过程

给定一个特征子集 $F=\{f_1,f_2,\ldots,f_N\}$,N 是特征集的大小.一个特征子集可以用一个二进制向量表示:$S=\{s_1,s_2,\ldots,s_N\}$,$s_i\in\{0,1\}$,$i=1,2,\ldots,N$,$s_i=1$ 表示第 i 个特征 f_i 被选择,反之,对第 i 个特征 f_i 不作选择.将评价函数在给定的特征子集 S 上所具有的最大性能 $G(S)$作为目标函数值,则特征选择问题转化为下列优化问题:

$$\max_S G(S).$$

特征选择一般经过 4 个阶段:特征子集产生、特征子集评估、评估停止、结果验证.其中,特征子集产生与子集评估两个阶段是最主要的,即搜索策略和评估函数.

1.1 特征子集产生

特征子集产生包含两个部分:特征空间搜索方向与搜索策略.搜索方向主要有 4 种.(1) 正向选择.开始时不含任何特征,然后每次增加一个.(2) 反向消除.开始时包含了所有的特征,然后每次减少一个.(3) 双向搜索.由正向选择和反向消除相结合.(4) 随机产生.随机地产生特征子集,主要用于不确定搜索.搜索策略主要包括 3 种:穷举搜索[2]、启发式搜索[3]和不确定搜索[4].穷举搜索是搜索所有可能的特征子集,这种搜索策略一定可以发现最优的特征子集,但搜索空间大,当特征数较多时是无法实现的;启发式搜索按照一定的启发式规则搜索特征子集,这种搜索策略搜索空间比较小,可能丢失最优子集.不确定搜索实际上是一种对上述两种搜索的平衡方法,比较典型的不确定搜索有遗传算法[5].

1.2 特征子集评估

特征选择可以看作是一个优化问题,其关键是建立一种评估标准来区分哪些特征组合有助于分类,哪些特征组合存在冗余性、部分或者完全无关.不同的评估函数可能会给出不同的结果.根据评估函数与分类器的关系,特征选择方法分成过滤器模式[6,7]和封装器模式[7−9]两种.其中,过滤器的评估函数与分类器无关;而封装器采用分类器的分类错误率或正确率作为评价函数,其中,过滤器的评价函数又可以细分为距离测度、信息测度、相关性测度和一致性测度[3,10−12].

2 基于特征选择的轻量级入侵检测系统分类

特征选择有两种模式:过滤器和封装器[13,14].过滤器利用数据本身的特性作为特征子集的评价指标,而封装器利用机器学习算法的正确率作为特征子集的评价指标.一般来说,过滤器的特征选择速度比较快,选择的结果与采用的学习算法没有关系,选择效果比较差;封装器特征选择速度慢,需要交叉认证和大量的计算资源,选择结果依赖于采用的分类算法,选择效果一般较好.为了解决两种特征选择模式各自存在的问题,发挥它们的优

势,很多研究者提出了混合器模式.本文主要从这 3 种特征选择模式出发,在每一种模式中介绍几种典型的特征选择算法,然后通过实验比较基于这些特征选择算法的入侵检测模型的性能,总结和分析各种入侵检测模型的优缺点和适用条件.

2.1 基于过滤器模式的轻量级入侵检测系统

本节主要介绍两种过滤器模式的特征选择算法:相关性[12]和信息增益[15].基于过滤器的入侵检测系统流程如图 1 所示,图中出现的变量定义为 $D(F_0,F_1,\ldots,F_{N-1})$,具有特征数量为 N 的数据集;S_0,特征搜索空间的初始子集;S,生成的特征子集;γ,评价函数值;δ,评估停止条件;M,与分类器无关的评价函数;S_{best},最优特征子集;γ_{best},最优评价函数值;C,分类器;TrD,训练数据集;TeD,测试数据集.

图中出现的函数定义为 $Generate(D)$,根据数据集 D 生成一个特征子集 S;$Eval(S,D,M)$,根据数据集 D、评价函数 M,对特征子集 S 进行评估,返回γ;$Build(TrD,S_{best})$,通过 TrD 和最优特征子集 S_{best},建立分类器 C;$Test(TeD,C)$,通过测试集 TeD 检测分类器 C 的性能.

通过特征选择,找到 S_{best} 之后,在 S_{best} 和 TrD 上建立分类器 C,这样建立的分类器耗用计算资源少,性能优于在全部特征上建立的分类器.

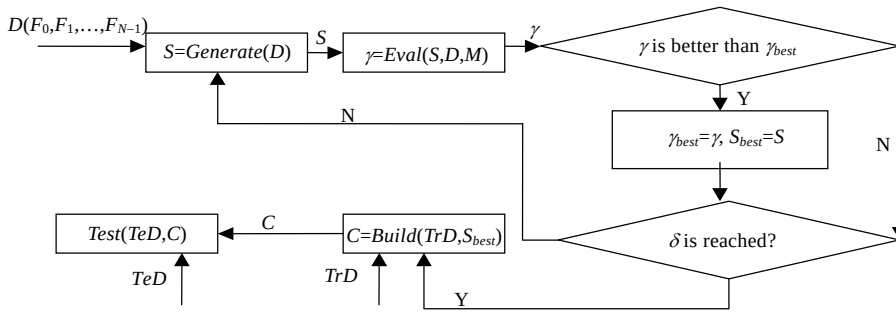

Fig.1 Flow chart of intrusion detection system based on filter

图 1 基于过滤器模式的入侵检测系统流程

2.1.1 相关性特征选择(correlation feature selection,简称 CFS)

相关性[12]用于特征选择的理论依据是:在原特征集中增加或删除相关特征,不影响该特征集的分类能力.为了证明这个理论依据,引入下面的定义和定理.

定义 **1**. 设 R_n 为 n 维特征空间;x_n 为原特征空间 R_n 中的模式向量; $F^n=\{x_j^n\}_{j=1}^N$ 为 m 个模式类别的总样本数,且有 $\sum_{i=1}^{m}N_i=N$,其中,N_i 为第 i 类模式的样本数.定义 Fisher 比准则为

$$J_n(\partial^n)=\frac{\partial^{n^T}S_b^n\partial^n}{\partial^{n^T}S_i^n\partial^n} \tag{1}$$

式中,S_b^n,S_i^n 分别为模式类间和类内散步矩阵,二者均为非负定矩阵;∂^n 是最优的分类鉴别矢量,使式(1)取得极大值;$J_n(\partial^n)$则反映了最优的分类鉴别矢量的分类能力.

定义 **2**. 对于 R_n 空间的样本集 $F^n=\{x_j^n\}_{j=1}^N$,存在阵列 $H^n=(x_1^n,x_2^n,\ldots,x_N^n)^T=(Y_1^N,Y_2^N,\ldots,Y_n^N)$,其中,$Y_i^N$ 是所有样本的第 i 个特征组成的 N 维矢量.若 $Y_1^N,Y_2^N,\ldots,Y_n^N$ 线性相关,则称模式间特征相关;若 $Y_1^N,Y_2^N,\ldots,Y_n^N$ 线性无关,则称模式间特征无关.

定理 **1**. 若在原特征空间 R_n 中增加一个相关特征构成特征空间 R_{n+1},样本阵列 H^{n+1} 中的列矢量 d_1 和特征空间 R_n 中的列矢量 $Y_1^N,Y_2^N,\ldots,Y_n^N$ 线性相关,则有 $J_n(\partial^n)=J_{n+1}(\partial^{n+1})$.

证明:由 d_1 与 $Y_1^N,Y_2^N,\ldots,Y_n^N$ 线性相关,根据矩阵的性质,将 H^{n+1} 进行初等变换,使 d_1 变为零矢量,即存在矩阵 $Q_{(n+1)\times(n+1)}$使得式(2)成立

$$H^{n+1}=(Y_1^N,Y_2^N,...,Y_n^N,0)Q \tag{2}$$

即

$$(x_1^{n+1},x_2^{n+1},...,x_N^{n+1})=Q^T\left(\overline{x_1^{n+1}},\overline{x_2^{n+1}},...,\overline{x_N^{n+1}}\right) \tag{3}$$

式中, $\overline{x_i^{n+1}}=(x_i^{n^T},0)^T(i=1,2,...,N)$.根据非负定矩阵的秩分解定理,特征空间 R_n 中的 S_b^n,S_i^n 的秩分解为

$$S_i^n=\sum_{i=1}^{l}\beta_i^n\beta_i^{n^T},\ S_b^n=\sum_{i=1}^{l}\lambda_i\beta_i^n\beta_i^{n^T} \tag{4}$$

式中, $l=Rank(S_i^n)$.所以,可得到特征空间 R_{n+1} 中 S_b^{n+1},S_i^{n+1} 的秩分解为

$$S_i^{n+1}=\sum_{i=1}^{l}Q^T\beta_i^{n+1}\beta_i^{(n+1)^T}Q,\ S_b^{n+1}=\sum_{i=1}^{l}\lambda_iQ^T\beta_i^{n+1}\beta_i^{(n+1)^T}Q \tag{5}$$

式中, $\beta_i^{n+1}=(\beta_i^{n^T},0)^T$.由式(1)、式(5)可得 $J_n(\partial^n)=J_{n+1}(\partial^{n+1})$. □

定理 1 表明,若在原特征空间中加入新的相关特征,则新旧特征空间中最优分类鉴别矢量的分类能力保持不变.也就是说,若在原特征空间中删除相关特征,则两特征空间的最优分类鉴别矢量的分类能力相等.这也就是用相关分析进行特征选择的理论依据.

通过相关性进行特征选择,最典型的是 Pearson 相关性[12],它可以计算一个特征子集的相关度.Pearson 相关性计算公式为

$$Merit_s=\frac{k\bar{r}_{cf}}{\sqrt{k+k(r-1)\bar{r}_{ff}}} \tag{6}$$

其中,s 是含有 k 个特征的特征子集,$Merit_s$ 是对特征子集 s 相关度的一个评估结果. $\bar{r}_{cf}$ 是类与特征之间的平均相关度, $\bar{r}_{ff}$ 是特征与特征之间的平均相关度.当类与特征之间相关度越高,特征与特征之间的相关度越小时,特征子集 s 分类效果越好.

2.1.2 信息增益(information gain,简称 IG)

1948 年,Shannon 提出并发展了信息论,研究以数学的方法来度量信息,提出了信息增益等基本概念,并得到广泛的应用.信息增益又称为互信息.样本中属性的信息增益越大,其包含的信息量也越大.也就是说,在特征选择时,应计算各个属性的信息增益.具有最高信息增益值的属性是给定集合中具有最高区分度的属性.属性 A 的信息增益[15]定义为

$$Gain(A)=I(s_1,s_2,\ldots,s_m)-E(A) \tag{7}$$

其中,$I(s_1,s_2,\ldots,s_m)$由样本的熵确定,其定义为

$$I(s_1,s_2,...,s_m)=-\sum_{i=1}^{m}P(C_i)\log_2P(C_i) \tag{8}$$

其中,$P(C_i)$是任意样本属于 C_i 的概率,$P(C_i)=s_i/s$;m 表示样本类别数;s_i 是属于类 C_i 的样本数;s 是总的样本数.式(7)中,$E(A)$定义为

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+s_{2j}+...+s_{mj}}{s}I(s_{1j}+s_{2j}+...+s_{mj}) \tag{9}$$

设属性 A 具有 v 个不同的值$\{a_1,a_2,\ldots,a_v\}$,可以用属性 A 将 s 划分为 v 个子集$\{s_1,s_2,\ldots,s_v\}$,其中,s_j 包含 s 中 A 值为 a_j 的那些记录.项$(s_{1j}+s_{2j}+\ldots+s_{mj})/s$ 是第 j 个子集的权值,等于子集 $A=a_j$ 中的样本个数除以 s 中样本总数.s_{ij} 是子集 s_j 中类 C_i 的样本数,且式(9)中有

$$I(s_{1j}+s_{2j}+...+s_{mj})=-\sum_{i=1}^{m}P_{ij}\log_2P_{ij} \tag{10}$$

其中,$P_{ij}=s_{ij}/s_j$ 是 s_j 中样本属于类 C_i 的概率.式(7)中的 $Gain(A)$是从特征 A 上获得该划分的信息增益,它表示具有最高信息增益的特征是给定记录集中具有区分度的特征.

2.2 基于封装器模式的轻量级入侵检测系统

基于封装器模式的入侵检测系统详细流程如图 2 所示.与过滤器相比,封装器更有利于提高分类器的性能,但是会耗用更多的计算资源与存储资源.与过滤器模式的流程图相比,封装器在评估特征子集时采用了与分类器相关的评价函数 A,而不是无关评价函数 M.

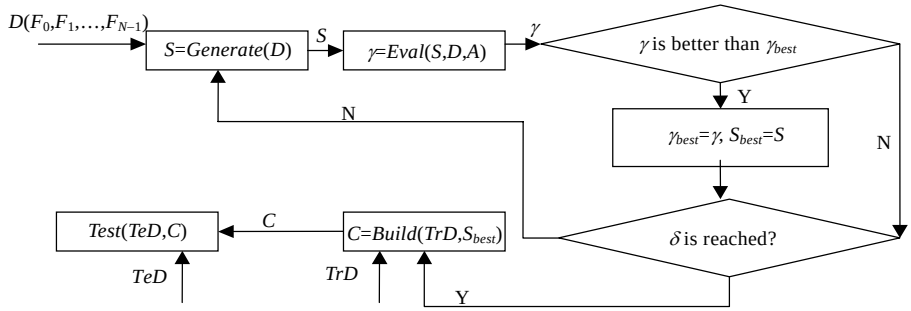

Fig.2 Flow chart of intrusion detection system based on wrapper

图 2 基于封装器模式的入侵检测系统流程

2.2.1 支持向量机(support vector machine,简称 SVM)

支持向量机是由 Vapnik 博士提出的基于统计学习理论的一种新的模式识别技术.其用于特征选择的主要思想[16−19]是:把已知一组 N 个 d 维的独立同分布的训练样本 $X=\{(x_i,y_i)|x_i\in R^d,y_i\in\{-1,1\},i=1,2,\ldots,N\}$通过非线性变换 $h(\cdot)$映射到一个高维特征空间 f.在此高维特征空间中,构造最优性决策函数 $y(x)=\mathrm{sgn}(\omega\cdot h(x)+b)$.$\omega$是分类超平面的系数向量,$b$ 是分类阈值,应用 Lagrange 乘子法,ω可以表示为

$$\omega=\sum_{i=1}^{N}\alpha_i y_i h(x_i) \tag{11}$$

α_i是 Lagrange 乘子.如果 $y(X)$是正值,则 X 属于正值的类;如果 $y(X)$是负值,则 X 属于负值的类.$y(X)$的值取决于ω_i和 X:如果ω_i是一个很大的正值,那么第 i 个特征对正值类的贡献大;如果ω_i是一个很大的负值,则第 i 个特征对负值类的贡献大.如果ω_i 的值在零值左右偏移,则第 i 个特征不具有很好的分类能力.一个对于特征的排序可以通过支持向量机的函数完成.引入原样本空间核函数:

$$k(x_i,x_j)=h(x_i)\cdot h(x_j) \tag{12}$$

则将高维空间的点积运算 $h(x_i)\cdot h(x_j)$转变成原空间的核函数,避免了指定非线性变换形式和真正实施复杂的非线性映射.

2.2.2 神经网络(neural network,简称 NN)

神经网络[20]由于具有良好的非线性映射能力和对任意函数的准确逼近能力,在特征选择领域得到了广泛的应用.特别是当可以获得的信息仅仅由训练数据来提供时,神经网络能够更为有效地进行特征选择.在特征选择领域应用得比较普遍的是 BP(back propagation)[21]神经网络,这里具体介绍一个 3 层前馈型的神经网络:网络输入层、网络隐含层和网络输出层.把向后搜索策略用于基于神经网络的特征选择算法中,向后搜索先用所有的特征作为网络输入,训练网络达到要求精度,比较删去每个特征后网络在训练集上的精度变化,然后选择精度下降最小的特征进行评价.使用验证集上的精度变化作为评价标准,如果删除此特征后,验证集上的精度下降在允许范围内,则删除此特征;否则,保留它.同样的方法应用于删除隐节点,直到网络中所有的输入特征和隐节点都能删除为止.$(x_1,x_2,\ldots,x_N)$为网络的输入层,N 为原始特征集合中元素个数;网络隐含层的神经元个数为 n_h;最后一层网络输出层具有 n_L 个节点,这一参数主要是由实际问题中类别总数来确定,在我们的实验中该值为 2.网络输出层和隐含层的传输函数均为对数-S 形函数,即 $f(x)=\dfrac{1}{1+\mathrm{e}^{-x}}$.首先利用 BP 算法进行网络学习,使用了如式(13)所示的扩展互熵误差函数来训练网络,其中互熵误差函数(the cross-entropy error function)E_0 为标准误差项,

具体取值可以由式(14)计算出来,另外两项均为正则化项:

$$E=\frac{E_0}{n_L}+\alpha_1\frac{1}{Pn_h}\sum_{p=1}^{P}\sum_{k=1}^{n_h}f'(a_{kp}^h)+\alpha_2\frac{1}{Pn_L}\sum_{p=1}^{P}\sum_{j=1}^{n_L}f'(a_{jp}^L) \tag{13}$$

$$E_0=-\frac{1}{2P}\left[\sum_{p=1}^{P}\sum_{j=1}^{n_L}(d_{jp}\log O_{jp}^L+(1-d_{jp})\log(1-O_{jp}^L))\right] \tag{14}$$

其中,d_{jp} 为网络输入为第 p 个样本数据时第 j 个神经元处的输出期望值.

从式(14)可以看出,所添加的两项有效地限制了传输函数的导数,使得隐含层和输出层的神经元工作在饱和区域内.利用限制隐含层传输函数导数的正则化神经网络具有很好的泛化能力.在用网络来解决一个问题时,要想使网络获得最低的泛化误差,必须要求隐含层和输出层中各个节点具有不同的敏感度,所以,在误差函数中使用了 α_1,α_2 这两个正则化参数.式(13)中,P 为训练样本的个数, a_{kp}^h , a_{jp}^L 分别是输入为第 p 个数据时隐含层中第 k 个神经元和输出层第 j 个神经元的输入值,其计算公式分别为

$$a_{kp}^h=\sum_{i=0}^{N}\omega_{ik}^o x_{ip}\ ,\ a_{jp}^L=\sum_{i=0}^{n_h}\omega_{ij}^h O_{ip}^h\ .$$

其中, ω_{ik}^o 为网络第 i 个输入与输入层第 k 个神经元之间的权值,x_{ip} 为输入数据样本中第 p 个数据时网络的第 i 个输入值, ω_{ij}^h 为隐含层第 i 个神经元与输出层第 j 个神经元之间的权值, O_{ip}^h 是指网络输入为第 p 个数据时隐含层中第 i 个神经元的输出值,计算公式为 $O_{ip}^h=f(a_{ip}^h)$.式(13)中, $f'(a_{kp}^h)$, $f'(a_{ip}^L)$ 分别是网络输入为第 p 个数据时传输函数在隐含层中第 k 个神经元和输出层中第 j 个神经元处的导数.

2.3 基于混合器模式的轻量级入侵检测系统

基于混合器模式的入侵检测系统详细流程如图 3 所示,它不仅通过与分类器无关的评价函数来评估子集,并且加上了与分类器有关的评价函数来评估.它首先通过与分类器无关的评价函数选出候选的特征集合,然后把这些候选特征集合提交给机器学习算法选出最优的特征子集.

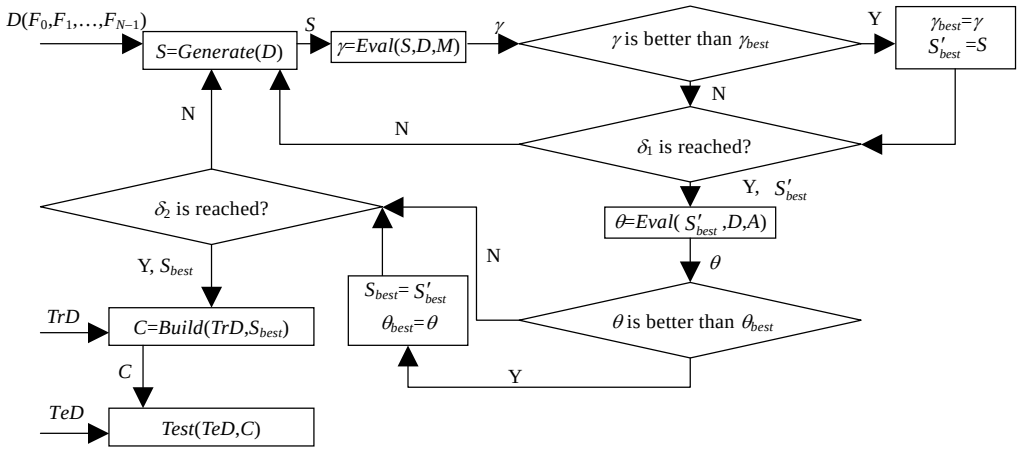

Fig.3 Flow chart of intrusion detection system based on hybrid

图 3 基于混合器模式的入侵检测系统流程

2.3.1 基于 CFS 和 SVM 混合特征选择的入侵检测系统

基于 CFS 和 SVM 的混合器模式[22−24]轻量级入侵检测系统结构如图 4 所示.由图 4 可知,网络数据经过预处理程序清理之后,形成 3 个符合我们实验要求的数据集合:训练集、认证集、测试集.数据集中的特征选择分两个阶段完成:首先由 CFS 选出特征子集,然后把这些候选的特征子集交由 SVM[25]分类器评估.经过两个阶段的处理之后,对特征空间中的特征进行排序和删减,把最终的特征子集提交给 SVM.SVM 依据这些特征和训练

集建立入侵检测模型.最后用测试集测试已建立模型的性能.

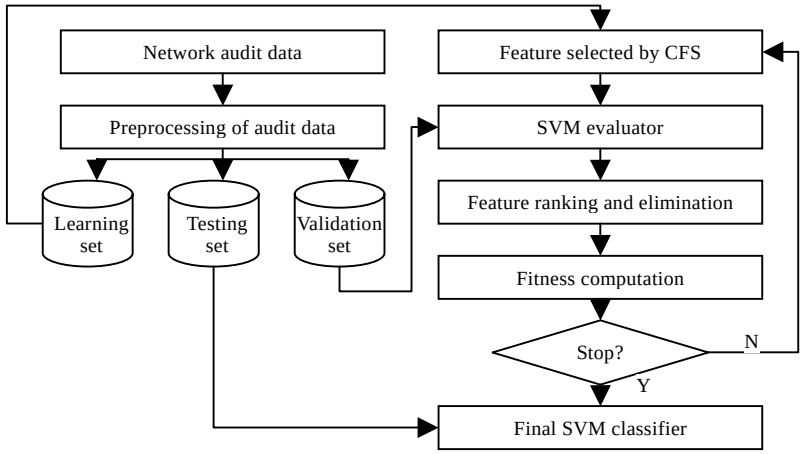

Fig.4 Overall flow of intrusion detection system based on CFS and SVM

图 4 基于 CFS 和 SVM 的混合器模式入侵检测系统结构

2.3.2 基于 IG 和 J4.8 混合特征选择的入侵检测系统

基于 IG 和 J4.8 的混合器模式[13]入侵检测系统的流程如图 5 所示.这个系统主要由两部分组成:特征选择和分类器.特征选择包括两部分:基于信息增益 IG 和决策树 J4.8 的特征子集评估.特征子集的产生是由遗传算法(GA)[5]完成的.利用信息增益对其产生的特征子集评估,选出最优的特征子集 S_{best},然后决策树又对 S_{best} 评估,把评价函数的最好值存放在 θ_{best}.最后当达到最大循环次数或是满足限制条件 δ 时,θ_{best} 对应的那个 S_{best} 作为分类器 J4.8[15]的输入特征子集.

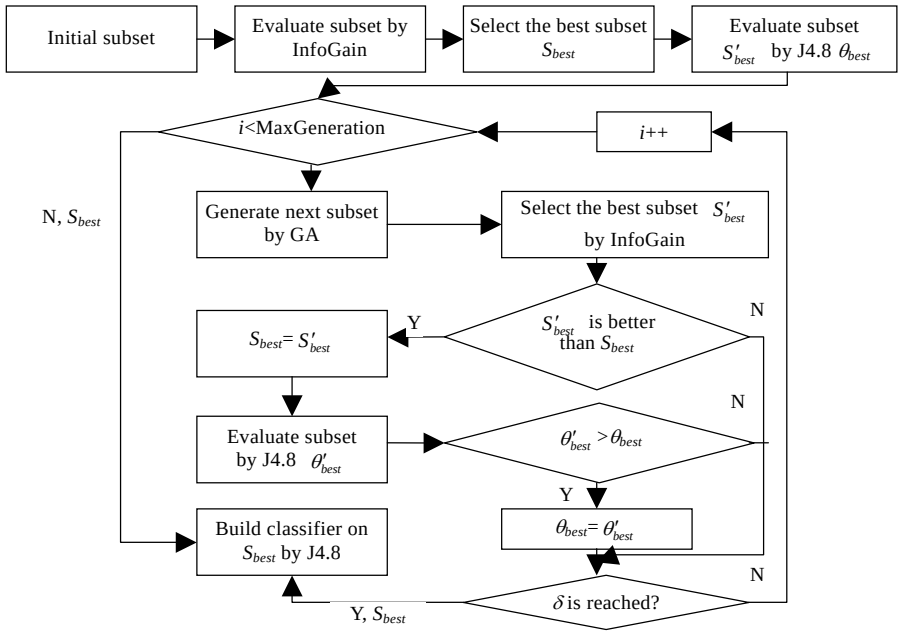

Fig.5 Flow chart of intrusion detection system based on IG and J4.8

图 5 基于 IG 和 J4.8 混合器模式的入侵检测系统流程

3 实验验证及比较分析

本文所有的实验数据都来源于 KDD1999[26],并且在开源工程 WEKA[27]上完成.由于在 KDD 数据集中有 79.2%是 DOS,所以我们下面所有的实验都是针对 DOS 来设计的.我们把 KDD1999 中 DOS 的训练集和测试集合并,并且随机取样部分实例形成 11 个数据集,这些数据集都保持着原样本的分布特性.每个数据集含有 12 350 个实例.我们选择其中一个数据集作为特征选择的训练集,然后把在该数据集上选出的最优特征子集在另外 10 个数据集上训练分类器.为了对基于特征选择的分类器和基于 KDD1999 全部 41 个特征[26]的分类器性能进行比较,我们采用 3 折交叉验证.实验硬件环境是 Intel(R) Pentium(R) processor 1.73GHz,512M 内存.软件环境是 Microsoft Windows XP SP2.

3.1 实验评价标准

基于特征选择的轻量级入侵检测系统的性能评价指标有 4 个:平均建模时间、平均测试时间、检测率、误报率.下面我们给出检测率和误报率的定义:

检测率(true positive rate,简称 TPR)被定义为: $检测率(TPR)=\frac{正确检测出的攻击样本数量}{总的攻击样本数量}$;

误报率(false positive rate,简称 FPR)被定义为: $误报率(FPR)=\frac{被错误判断为攻击的正常样本数量}{总的正常样本数量}$.

3.2 特征选择结果

针对第 2 节描述的全部特征选择算法,我们通过实验让它们在训练集上进行特征选择,特征选择的结果见表 1.表中列出了 6 种特征选择算法在训练集上进行特征选择之后形成的特征子集.其中,filter 型特征选择算法有 2 种:CFS,IG.wrapper 型特征选择算法有 2 种:SVM 和 NN.hybrid 型特征选择算法有 2 种:基于 CFS 与 SVM 的特征选择算法 CFS-SVM 和基于 IG 与 J4.8 的特征选择算法 IG-J48.

特征选择算法选择出的特征均在 KDD1999 全部 41 个特征之中.在特征选择结果栏中,冒号前面的数字表示选择的特征在 KDD1999 的 41 个特征中的序号,冒号后面的字符串是该序号对应的特征名称.如 hybrid 型特征选择算法:CFS-SVM,其特征选择的结果为 5,6,12,23,其中,5 表示 src_bytes,6 表示 dst_bytes,12 表示 logged_in,23 表示 count.

Table 1 Selected feature subsets for different selection algorithms

表 1 各种特征选择算法选择的特征子集

Algorithm	Selected features
CFS	6,12,23,31,37: dst_bytes,logged_in,count, srv_diff_host_rate,dst_host_srv_diff_host_rate
IG	3,5,6,12,23,24,36: service,src_bytes,dst_bytes,logged_in, count,srv_count,dst_host_same_src_port_rate
SVM	5,23,33,34: src_bytes,count,dst_host_srv_count, dst_host_same_srv_rate
NN	6,10,23,37,40: dst_bytes,hot,count, dst_host_srv_diff_host_rate,dst_host_rerror_rate
CFS-SVM	5,6,12,23: src_bytes,dst_bytes,logged_in,count
IG-J48	2,5,23,37: protocol_type,src_bytes,count, dst_host_srv_diff_host_rate

通过表 1 可以看出,在 KDD1999 中有一个非常重要的特征,无论采取何种特征选择方法,该特征总是被选取.它就是 count,表示在过去 2 秒内通过同一节点的与本次连结相同的连结的数目.这说明 count 属性与 DOS 攻击之间存在非常密切的联系.无论从理论上还是在我们后面的实验中都可以证明 count 是一个重要的特征.特征选择的目的就是发现像 count 一样的特征.

3.3 实验验证及结果分析

针对表 1 中每一种特征选择算法选择的特征子集,我们在此基础上建立轻量级入侵检测系统模型,然后把

这些建立的检测模型与那些基于所有 41 个特征建立的模型进行对比,比较它们在平均建模时间、平均检测时间、检测率以及误报率上的性能.我们对各种入侵检测系统的平均建模时间与平均测试时间进行了记录,其结果见表 2.

Table 2 Average building time and testing time for nine different intrusion detection systems

表 **2** 9 种不同的入侵检测模型的平均建模时间与检测时间

System	J48-**ALL**	J48-**IG**	J48-**IG**-**J48**	SVM-**ALL**	SVM-**CFS**	SVM-**SVM**	SVM-**CFS**-**SVM**	NN-**ALL**	NN-**NN**
Building time (s)	1.3	0.13	0.19	77	22	60	35	11468	7800
Testing time (s)	–	–	–	15	5	10	6	22	7

表 2 中除了基于表 1 中 6 种特征选择算法建立的轻量级入侵检测模型之外,还有 3 种检测模型:基于所有 41 个特征建立的决策树分类模型 J48-ALL、基于所有 41 个特征建立的 SVM 模型 SVM-ALL、基于所有 41 个特征建立的 NN 入侵检测模型 NN-ALL.在表的 System 行中,如 J48-**IG**-**J48**,前面的 J48 表示分类器名称,后面的粗体 **IG**-**J48** 表示特征选择算法.我们依据每一种分类器来比较特征选择前后模型的平均建模时间和平均测试时间.基于决策树 J4.8 上建立的入侵检测模型有 3 种:J48-ALL,J48-IG,J48-IG-J48;基于 2 种特征选择算法 IG,IG-J48 建立的检测模型比基于所有 ALL 特征建立的入侵检测模型具有更短的建模时间,特别是基于信息增益 IG 建立的入侵检测模型建模时间最短.在测试时间上,由于基于 J4.8 建立的入侵检测模型检测时间很快,所以就未作统计.基于分类器 SVM 建立的 4 种入侵检测模型涵盖了 filter 型、wrapper 型、hybrid 型特征选择算法,从这 4 种检测模型可以看出,filter 模式 SVM-CFS 具有最短的建模时间与测试时间,wrapper 模式 SVM-SVM 具有最长的建模时间与测试时间,hybrid 模式 SVM-CFS-SVM 居中.针对 3 种分类器:支持向量机 SVM、决策树 J4.8 和神经网络 NN,J4.8 具有更短的建模时间与测试时间,不超过 2 秒,NN 最长,SVM 次之.

在检测率和误报率方面,各种特征选择算法的表现也参差不齐.下面从特征选择算法的有效性验证和各种特征选择算法性能比较两个角度来分析入侵检测模型的检测结果.特征选择算法的有效性验证是指结合特征选择的 IDS(intrusion detection system)与没有结合特征选择的 IDS 在检测率和误报率上的比较,如果基于特征选择的 IDS 具有高检测率、低误报率,则证明特征选择算法是有效的,因为加入特征选择算法之后,IDS 检测性能提高了.在下面所有的实验结果比较图中,曲线名称的第一个连字符前面表示分类器名称,后面表示特征选择算法的名称.

在过滤器模式下,针对分类器 J4.8 构成的入侵检测系统,有基于全部特征的分类器 J48-ALL、基于信息增益的分类器 J48-IG 两种类型.针对分类器 SVM 构成的入侵检测系统,有基于全部特征 SVM-ALL、基于相关性特征选择的分类器 SVM-CFS 两种类型.由图 6 可知,SVM-ALL,SVM-CFS 在 TPR 上比 J48-ALL,J48-IG 平均高出 0.6%;但是在 FPR 上,J4.8 构成的分类器比起 SVM 分类器平均要低 1.6%,这将更有利于系统发现异常情况.从图 6 可知,在以 J4.8 为分类器的入侵检测中,J48-IG 的误报率 FPR 比 J48-ALL 略偏低.以 SVM 为分类器的两种入侵检测模型中,SVM-CFS 在 TPR 上略低于 SVM-ALL,但是相差极小.在 FPR 上,SVM-CFS 具有很小的误报率.由分类器 J4.8 构成的检测模型在 TPR,FPR 上的性能要优于分类器 SVM 构成的模型.从 TPR,FPR 的结果图不难看出,基于各种特征选择算法的 IDS 都具有很高的检测率与很低的误报率,这是因为 KDD1999 数据集相对于其自身而言是一个比较完备的数据集,所以在其上进行 3 折交叉认证,可以获得很好的结果.但是,这些都不影响特征选择算法有效性的验证,因为从图中可以得出,基于特征选择的 IDS 在 TPR 相当的情况下具有更低的 FPR.在同一分类器下,从经过特征选择的分类器和未经过特征选择的分类器性能比较上,可以显而易见地看出特征选择算法的有效性.

图 7 显示了在封装器模式下由分类器 SVM,NN 构成的两类入侵检测系统在 TPR,FPR 上的表现.其中,SVM-SVM 表示分类器用 SVM,特征选择算法也用了 SVM.它是 wrapper 型特征选择方法,其在 TPR,FPR 上的性能优于基于 filter 型特征选择的入侵检测系统 SVM-CFS.针对分类器神经网络构成的入侵检测模型有 2 种:基于所有特征建立的模型 NN-ALL 和基于神经网络 NN 选择算法的 NN-NN.由于神经网络进行特征选择时会耗用大量的选择时间,并且选择后的系统与未选择时相比没有太大的改进,所以在特征选择和轻量级入侵检测

系统方面,神经网络具有一定的局限性.由图可知,SVM-SVM 具有最好的 TPR.NN-ALL,NN-NN 的 TPR 相当.在 FPR 的实验结果中可以看出,NN-ALL,NN-NN 具有很好的性能,SVM-CFS 最差,比它们平均低 1.5%.

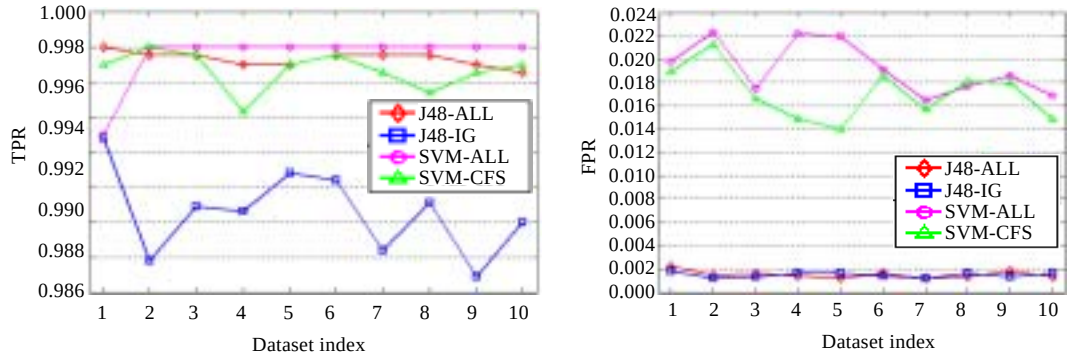

Fig.6 TPR and FPR of each IDS on ten data sets, and the IDS is built using all features or selected features, which are selected by filter feature selection algorithm

图 6 基于全部特征和过滤器选择特征的入侵检测模型在 10 个数据集上的 TPR,FPR

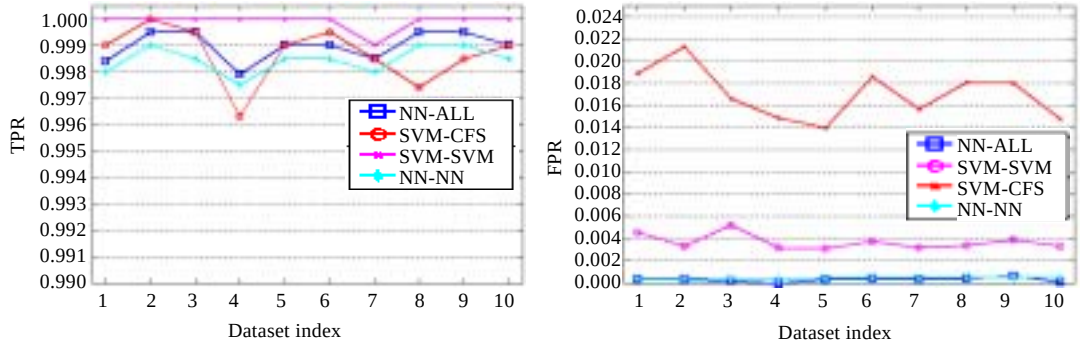

Fig.7 TPR and FPR of each IDS on ten data sets, and the IDS is built using all features or selected features, which are selected by filter feature selection algorithm or wrapper feature selection algorithm

图 7 基于全部特征、过滤器选择特征和封装器选择特征的入侵检测系统在 10 个数据集上的 TPR,FPR

图 8 比较了分类器 J4.8 和 SVM 构成的两类入侵检测系统,在同一分类器下对 filter 模式、wrapper 模式、hybrid 模式 3 种类型的特征选择算法进行了对比.由 J4.8 构成的入侵检测系统,只比较了 filter 模式 J48-IG 和 hybrid 模式 J48-IG-J48.从图 8 中可以看出,hybrid 模式的入侵检测系统 J48-IG-J48 在 TPR 上高于 filter 模式入侵检测系统 J48-IG,平均高出将近 0.9%;在 FPR 上则低于它.针对分类器 SVM 构成的入侵检测系统,有基于 SVM 和 CFS 的 hybrid 模式建立的检测模型 SVM-CFS-SVM,其在 TPR,FPR 上介于基于 filter 模式 SVM-CFS 与基于 wrapper 模式 SVM-SVM 建立的分类器之间.基于 J4.8 分类器构建的 IDS 具有很高的检测率和很低的误报率,这与分类器本身有关,也与样本集和实验方法——3 折交叉验证有关.采用交叉验证相对于独立测试集而言,IDS 具有更好的性能.所以,在 TPR 上,很多 IDS 都接近 100%的检测率.即使如此,我们还是在一个平等的平台下——同样的数据集、同样的实验方案、同样的实验环境、同样的分类器下比较特征选择算法的.在基于 J4.8 构建的分类器下,有两种特征选择算法 IG,IG-J4.8,IG-J4.8 无论是在 TPR,FPR 上均由于 filter 选择算法 IG,与 SVM 建立的 IDS 相比,也具有很大的优势.

从前面的结果来看,混合器模式的入侵检测系统,无论在 TPR,还是在 FPR 上都具有很好的性能,表现突出的是 J48-IG-J48.在同一模式下,以 J4.8 和神经网络为分类器的模型在 FPR 上好于以 SVM 为分类器的模型,但是在 TPR 上却比 SVM 差.选择何种模式、何种特征选择算法、何种分类器,取决于系统对性能的要求,应具体问

题具体分析.如果系统对 TPR,FPR 的要求较高,则 J48-IG-J48 是不错的选择;如果系统对 TPR 要求高,则 SVM-SVM 可以选择,但是它的 FPR 比较差.

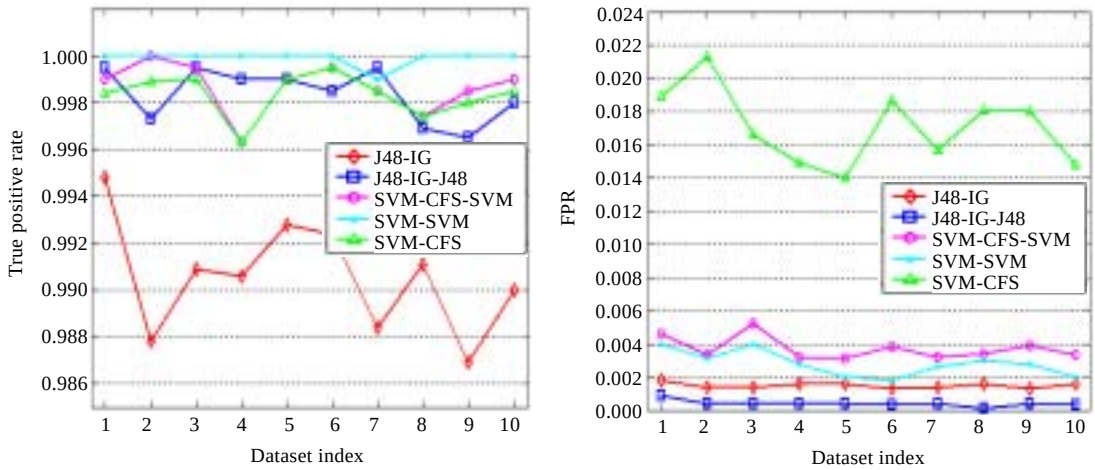

Fig.8 TPR and FPR of each IDS on ten data sets, and the IDS is built using selected features, which are selected by filter, wrapper or hybrid feature selection algorithm

图 8 基于过滤器选择特征、封装器选择特征和混合器选择特征的入侵检测系统在 10 个数据集上的 TPR,FPR

4 总结和未来工作方向

本文根据 3 种特征选择模式来给轻量级入侵检测系统进行分类,对每一种模式下的入侵检测系统进行实验比较和分析,同时还对模式间的入侵检测系统进行实验比较和分析,总结了基于各种模式的入侵检测系统的优缺点和适用条件.轻量级入侵检测系统的关键技术是特征选择,关于特征选择的应用越来越多,而在这些应用中面临的问题主要有两种:数据维数过多、数据集来源受限.目前在入侵检测领域特征选择采用的方法主要有 3 类:过滤器模式、封装器模式和混合器模式[28,29].针对高维数据带来的计算复杂性,过滤器模式能够很好地解决,因为它是通过与分类器无关的评价函数来对特征子集进行评估的.而封装器在计算复杂性上却陷入困境.最近,很多人提出一种混合器模式,它结合过滤器和封装器的优势,不仅继承了过滤器计算速度快的特点,而且继承了封装器分类效果好这一优势.混合器的计算复杂性与过滤器相当,能够很好地处理高维数据,然而在特征选择方面,还需要继续研究去发现更好的搜索策略和评估函数[30]以解决高维数据问题.有限的数据集也给特征选择带来困难.特征选择是一个非常值得研究的领域,它归属于数据挖掘以及其他数据预处理技术[30]的范畴.本文试图研究这一活跃的领域,展示在入侵检测领域用到的一些典型特征选择算法,并且介绍了它们在轻量级入侵检测系统中起到的重要作用和存在的问题.在将来的研究与实践中,我们将结合其他方面的相关技术发展特征选择算法,使它能够更好地为建立正确率高、误报率低、检测速度快的轻量级入侵检测系统服务.

References:

[1] Forres S, Perelson AS, Allen L, Cherukun R. Self-Nonself discrimination in a computer. In: Rushby J, Meadows C, eds. Proc. of the '94 IEEE Symp. on Research in Security and Privacy. Los Alamitos: IEEE Computer Society Press, 1994. 120−128.

[2] Narendra PM, Fukunaga K. A branch and bound algorithm for feature subset selection. IEEE Trans. on Computer, 1977,26(9): 917−922.

[3] Zhou R, Hansen E. Breadth-First heuristic search. Artificial Intelligence, 2006,170(4-5):385−408.

[4] Gheorghies O, Luchian H, Gheorghies A. A study of adaptation and random search in genetic algorithms. In: Proc. of the 2006 IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel. Vancouver: IEEE Computer Society Press, 2006. 2103−2110. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1688566

[5] Bhuyan J. A combination of genetic algorithm and simulated evolution techniques for clustering. In: Proc. of the ACM 23rd Annual Conf. on Computer Science. Nashville: ACM, 1995. 127−134. http://www.informatik.uni-trier.de/~ley/db/conf/acm/csc95.html#Bhuyan95

[6] Liu H, Setiono R. A probabilistic approach to feature selection: A filter solution. In: Proc. of the 13th Int'l Conf. on Machine Learning. 1996. 319−327. http://www.public.asu.edu/~huanliu/publications.html

[7] Das S. Filters, wrappers and a boosting based hybrid for feature selection. In: Brodley C, Danyluk A, eds. Proc of the 8th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001. 74−81.

[8] Yuan H, Tseng SS, Wu GS, Zhang FY. A two-phase feature selection method using both filter and wrapper. In: Proc of the '99 IEEE Int'l Conf. on Systems, Man, and Cybernetics. Piscataway: IEEE Computer Society Press, 1999. 132−136.

[9] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence Journal, 1997,97(1-2):273−324.

[10] Almuallim H, Dietterich TG. Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 1994, 69(1-2):279−305.

[11] Teodoro ML, Phillips GN, Jr Kavraki LE. A dimensionality reduction approach to modeling protein flexibility. In: Proc. of the 6th Annual Int'l Conf. on Computational Biology. Washington: ACM, 2002. 299−308. http://citeseer.ist.psu.edu/teodoro02dimensionality.html

[12] Hall MA. Correlation-Based feature selection for discrete and numeric class machine learning. In: Langley P, *et al.*, eds. Proc. of the 17th Int'l Conf. Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2000. 359−366.

[13] Liu H, Yu L. Towards integrating feature selection algorithms for classification and clustering. IEEE Trans. on Knowledge and Data Engineering, 2005,17(4):491−502.

[14] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 2004, 5(10):1205−1224.

[15] Baglioni M, Furletti B, Turini F. DrC4.5: Improving C4.5 by means of prior knowledge. In: Proc. of the 2005 ACM Symp. on Applied Computing. Santa Fe: ACM, 2005. 474−481. http://www.informatik.uni-trier.de/~ley/db/conf/sac/sac2005.html#BaglioniFT05

[16] Fugate M, Gattiker JR. Anomaly detection enhanced classification in computer intrusion detection. LNCS 2388, Berlin, Heidelberg: Springer-Verlag, 2002. 186−197.

[17] Kim DS, Park JS. Network-Based intrusion detection with support vector machines. LNCS 2662, Berlin, Heidelberg: Springer-Verlag, 2003. 747−756.

[18] Beverly R, Sollins K, Berger A. SVM learning of IP address structure for latency prediction. In: Proc. of the 2006 SIGCOMM Workshop on Mining Network Data. Pisa: ACM, 2006. 1−6. http://www.sigcomm.org/sigcomm2006/papers/minenet-04.pdf

[19] Joachims T. Making large-scale SVM learning practical. In: Schlkopf B, Burges C, Smola A, eds. Advances in Kernel Methods—Support Vector Learning. Cambridge: MIT Press, 1999. 1−13.

[20] Sung AH, Mukkamala S. Identifying important features for intrusion detection using support vector machines and neural networks. In: Proc. of the 2003 Int'l Symp. on Applications and the Internet Technology. Orlando: IEEE Computer Society Press, 2003. 209−216. http://www.computer.org/portal/site/store/menuitem.41cf17dc879177c86ee948ce8bcd45f3/index.jsp?&pName=store_level1&path=store/catalog&file=pr01872.xml&xsl=generic.xsl&

[21] Reed R. Pruning algorithms—A survey. IEEE Trans. on Neural Network, 1993,4(3):740−662.

[22] Park JS, Shazzad KM, Kim DS. Toward modeling lightweight intrusion detection system through correlation-based hybrid feature selection. In: Feng D, Lin D, Yung M, eds. Proc. of the CISC. Heidelberg: Springer-Verlag, 2005. 279−289.

[23] Kim DS, Nguyen HN, Ohn SY, Park JS. Fusions of GA and SVM for anomaly detection in intrusion detection system. In: Advances in Neural Networks. LNCS 3498, New York, Berlin, Heidelberg: Springer-Verlag, 2005. 415−420.

[24] Kompella R, Singh S, Varghese G. On scalable attack detection in the network. In: Proc. of the 4th ACM SIGCOMM Conf. on Internet Measurement. Washington: ACM, 2004. 187−200. http://citeseer.ist.psu.edu/kompella04scalable.html

[25] Rao X, Dong CX, Yang SQ. An intrusion detection system based on support vector machine. Journal of Software, 2003,14(4): 798−803 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/798.htm

[26] KDD cup 1999 data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[27] http://www.cs.waikato.ac.nz/ml/weka/index.html

[28] Chen Y, Li Y, Cheng XQ, Guo L. Survey and taxonomy of feature selection algorithms in intrusion detection system. In: Lipmaa H, Yung M, Lin D, eds. Proc. of the Conf. on Information Security and Cryptology. LNCS 4318, Berlin, Heidelberg: Springer-Verlag, 2006. 153−167.

[29] Chen Y, Dai L, Li Y, Cheng XQ. Building efficient intrusion detection model based on principal component analysis and C4.5 algorithm. In: Proc. of the 9th IEEE Int'l Conf. on Advanced Communication Technology. Korea: IEEE Computer Society Press, 2007. 2109−2112. http://www.icact.org/

[30] Taylor C, Alves-Foss J. NATE: Network analysis of anomalous traffic events, a low-cost approach. In: Proc. of the 2001 Workshop on New Security Paradigms. New Mexico: ACM, 2001. 89−96. http://www.csds.uidaho.edu/papers/Taylor01a.pdf

附中文参考文献:

[25] 饶鲜,董春曦,杨绍全.基于支持向量机的入侵检测系统.软件学报,2003,14(4):798−803. http://www.jos.org.cn/1000-9825/14/798.htm

陈友(1981－),男,安徽安庆人,博士,主要研究领域为网络安全,数据挖掘.

李洋(1978－),男,博士,主要研究领域为入侵检测,网络攻防对抗技术.

程学旗(1971－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为网络信息安全,大规模信息检索与信息挖掘,P2P 计算.

戴磊(1979－),男,博士,主要研究领域为信息安全.