高技术通讯 2011 年 第 21 卷 第 12 期:1240 ~ 1245

doi:10.3772/j.issn.1002-0470.2011.12.004

用于网络入侵检测的 VFSA-C4.5 特征选择算法①

李 超②*** 李文法③** 段淼毅*,**

(*北京航空航天大学计算机学院 北京 100191)

(**北京交通大学计算技术研究所 北京 100029)

摘 要 提出了一种新的用于网络入侵检测的特征选择算法——VFSA-C4.5 算法。该算法采用快速模拟退火(VFSA)搜索策略对特征子集空间进行随机搜索,然后利用提供的数据在 C4.5 决策树上的分类错误率作为特征子集的评估标准来为入侵检测获取最优特征子集。在著名的 KDD1999 入侵检测数据集上进行了大量的实验,结果表明该算法相对于其它一些应用于入侵检测的特征选择算法,在保证较高检测率的前提下,可有效地降低误报率、入侵检测的计算复杂度和提高检测速度,能更适用于现实高速网络应用环境。

关键词 网络入侵检测,特征选择,快速模拟退火(VFSA),决策树

0 引 言

网络入侵检测系统是网络安全防御体系的一个重要组成部分[1]。入侵检测系统通过对网络和主机上某些关键信息进行收集分析,检测其中是否有违反安全策略的事件或攻击事件发生,并对检测到的事件发出警报。入侵检测系统处理的大量数据中含有众多相关与冗余信息,影响了入侵检测的准确率和速度,因而对这些数据进行修剪与剔除显得尤为重要。特征选择在这方面扮演着十分重要的角色。特征选择针对数据流的高维特征空间存在大量的相关与冗余特征的特性,在此高维空间上应用搜索算法来寻找最优的特征子集,剔除那些冗余的特征。在得到最重要和必要的特征形成向量后再进行训练和检测,这不仅可以降低入侵检测系统的计算复杂性,提高检测速度,而且可以获得很好的检测效果。

特征选择包括搜索策略与评估函数两个重要部分。在搜索策略方面,针对大数据集,文献[2]提出了诸如正向搜索、反向搜索、顺序搜索等启发式搜索策略,文献[3]提出了比启发式搜索更有优势的的随机搜索策略,如遗传算法。在大规模数据集上的特征选择,这些搜索策略的计算资源耗用大,收敛速度慢,并且在很多情况下得到的是局部最优解。在评估标准方面,文献[4]介绍了一种基于支持向量机(support vector machine,SVM)的特征选择算法,依据该算法可以选出那些分类信息明确的特征。文献[5]介绍了一种可以自动计算属性间相互关系的算法。文献[6]使用递归属性排除及线性向量机作属性评估。但是这些算法在特征子集的评估速度上有待提高。文献[7-9]分别提出了基于遗传算法(genetic algorithm,GA)和 SVM 的特征选择算法(GA-SVM)、基于相关性选择(correlation feature selection,CFS)和 SVM 的特征选择算法(CFS-SVM)、基于主成分分析(principal component analysis,PCA)和 C4.5 决策树的特征选择算法(PCA-C4.5),但在文献[7-9]中,选择特征化费了较多的时间,应用于入侵检测时效果并不令人十分满意。针对上述缺点,本文提出了一种快速模拟退火(very fast simulated annealing,VFSA)和 C4.5 决策树相结合的特征选择算法(VFSA-C4.5)。实验结果表明 VFSA-C4.5 算法相对于其它一些应用于入侵检测的特征选择算法,在保证较高检测率的前提下,有效地降低了误报率,减少了入侵检测的计算复杂度,提高了检测速度,能更适用于现实高速网络应用环境。

① 863 计划(2007AA01Z416)和 973 计划(2007CB311100)资助项目。

② 男,1976 年生,博士,高级工程师;研究方向:网络安全,入侵检测,数据挖掘;E-mail:super_1cm@hotmail.com

③ 通讯作者,E-mail:liliyuanzhu@sina.com

(收稿日期:2011-03-18)

1　VFSA-C4.5 算法

本文提出的特征选择算法包括搜索策略——快速模拟退火(VFSA)搜索策略和评估标准——C4.5 决策树评估算法。

1.1　特征选择算法的数学模型

给定一个特征子集 $F = \{f_1, f_2, \cdots, f_N\}$,$N$ 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:$S = (s_1, s_2, \cdots, s_N)$,$s_i \in \{0,1\}$,$i = 1,2,\cdots,N$。$s_i = 1$ 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。把 C4.5 决策树在给定的特征子集 S 上所具有的性能 $G(S)$ 作为目标函数值,则特征选择问题转化为优化问题

$$\max_S G(\mathbf{S}) \tag{1}$$

特征选择的求解优化问题 $\max_S G(\mathbf{S})$ 可以通过 VFSA 算法来求解。

1.2　VFSA 搜索策略

模拟退火算法如同随机变异爬山算法、遗传算法一样是属于随机搜索算法[10]。它们对搜索空间进行随机划分,在理论上可以得到全局最优解。模拟退火算法[11]基本过程如下:

第 1 步:初始化初始特征数 M(温度 T)、初始解状态 S(算法迭代的起点)、每个 M 值的迭代次数 K。

第 2 步:对 $n = 1,2,\cdots,K$ 做第 3 步至第 6 步。

第 3 步:产生新解 S_{new}。

第 4 步:计算增量 $\Delta m = f(S_{new}) - f(S)$,其中 $f(S)$ 为评价函数;

第 5 步:若 $\Delta m < 0$ 则接受 S_{new} 作为新的当前解,否则以概率 $\exp(-\Delta m/M)$ 接受 S_{new} 作为新的当前解。

第 6 步:如果满足终止条件则输出当前解作为最优解,结束程序。

第 7 步:M 逐渐减少,且 $M \to 0$,然后转第 2 步。

随机搜索存在的困难是当搜索的空间比较大时,算法复杂性增加,得到最优解的过程复杂化。模拟退火方法在实际使用时,为了提高算法效率,常采用改进的算法。文献[12-14]提到了几种实际使用的模拟退火算法。其中文献[14]中提到的非常快速模拟退火算法(简称 VFSA 算法)最为常用。VFSA 算法的流程与传统模拟退火算法的流程是一样的,只是为了加快算法的执行效率,在模型扰动、接受概率及其退火计划上进行了一些改进[14],而为加快入侵检测系统的速度,较高的执行效率也正是我们需要的。具体改进如下:

模拟退火算法中新模型的产生是对当前模型进行扰动得到的,常规模拟退火算法用的是高斯分布法,而快速模拟退火算法采用依赖于温度的似 Cauchy 分布法即:

$$m_i' = m_i + y_i(B_i - A_i) \tag{2}$$

$$y_i = \mathrm{Tsgn}(u - 0.5)[(1 + 1/T)^{|2u-1|} - 1] \tag{3}$$

式中:m_i 为当前模型中第 i 个变量;u 为[0,1]内均匀分布的随机数;$[A_i, B_i]$ 为 m_i 的取值范围;m_i 为扰动后的模型中第 i 个变量,且 $m_i' \in [A_i, B_i]$。Cauchy 分布法产生新模型的优点是:高温情况下搜索范围大,在低温时搜索仅在当前模型附近。

由广义 Boltzmann-Gibbs 分布,可给出新的接收概率计算公式:

$$P = [1 - (1 - h)\Delta E/T]^{1/(1-h)} \tag{4}$$

式中:$\Delta E = E(m) - E(m_0)$;h 为实数。当 $h \to 1$ 时

$$P = \exp(-\Delta E/T) \tag{5}$$

这是常规模拟退火算法的接受概率公式,为式(4)的特例。$T(k)$ 为迭代 k 次后的数值,表达式为

$$T(k) = T_0 \exp(-ck^{1/N}) \tag{6}$$

式中:T_0 为初始温度;k 为迭代次数;c 为给定常数;N 为待反演参数的个数。式(6)还可写为

$$T(k) = T_0 a^{k1/N} \tag{7}$$

式中,a 为温度衰减率,k 为 $1/N$ 的指数,通常选择 $0.7 \leqslant a \leqslant 1.0$。在实际应用中,可采用0.5或1.0代替式(7)中的 $1/N$。在 VFSA 算法中,最终用式(2)、(3)对当前模型进行扰动产生新模型,新模型按式(4)接收,并按式(7)降温。

VFSA 划分速度与空间降维能力可以通过退火来实现[14]。提高划分速度和降维能力首先在迭代初期每次变异的数目 M 要大,而在迭代后期,当接近满足的评估标准时,变异的特征数目 M 要小,所以我们对 M 进行了补充,给出了 M 的一个计算公式:

$$M = M_{max} \times \min\left[\frac{(I_{max} - i_{current})}{I_{max}}, P_{error}(S)\right] \tag{8}$$

式中 M_{max} 是每次迭代允许变异的最大特征数目,I_{max} 是最大迭代次数,$i_{current}$ 是当前迭代次数,$P_{error}(S)$ 是基于特征串 S 的分类器的所有类的平均分类错误率。在对特征空间进行快速降维的同时,我们希望选出的特征子集空间比较小,这样更利于对要选择的数据进行快速的分类,基于此特征子集

的入侵检测系统就更快速。

1.3 C4.5 决策树算法

C4.5 算法[15]从树的根节点处的所有训练样本开始，选取一个属性来区分这些样本。对属性的每一个值产生一个分支，分支属性值的相应样本子集被移到新生成的子节点上，这个算法递归地应用于每个子节点上，直到节点的所有样本都分区到某个类中，到达决策树的叶节点的每条路径表示一个分类规则。这样自顶向下的决策树的生成算法的关键性决策是对节点属性值的选择。选择不同的属性值会使划分出来的记录子集不同，影响决策树生长的快慢以及决策树结构的好坏，从而导致找到的规则信息的优劣。C4.5 算法的属性选择的基础是基于使生成的决策树中节点所含的信息熵最小。所谓熵在系统学上是表示事物的无序度。不难理解熵越小则记录集合的无序性越小，也就是说记录集合内的属性越有顺序有规律，这也正是我们所追求的目标。集合 S 的熵的计算公式如下：

$$Info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times \log_2\left(\frac{freq(C_j, S)}{|S|}\right) \tag{9}$$

式中 $freq(C_j, S)$ 表示集合 S 中属于类 C_j（k 个可能类中的一个）的样本的数量。$|S|$ 表示集合 S 中样本数量。上面的公式仅仅给出了一个子集的熵的计算，如果我们按照某个属性进行分区后就涉及到若干个子集，我们需要对这些子集进行熵的加权和的计算，公式为

$$Info_f(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times Info(T_i) \tag{10}$$

式中 T 指按属性 f 进行分区后的集合，T_i 指分区后集合中的某一个集合。$|T_i|$ 表示在集合 T_i 中样本数量，$|T|$ 表示集合 T 中样本数量。为了更加明显地比较不同集合的熵的大小，我们计算分区前的集合的熵和分区后的熵的差（我们把这个差叫做增益），增益大的就是我们要选取的节点。公式如下：

$$Gain(f) = Info(S) - Info_f(T) \tag{11}$$

2 基于 VFSA-C4.5 的入侵检测框架

基于 VFSA-C4.5 的网络入侵检测框架见图 1。首先初始化最大迭代次数 I_{max}、每次迭代最大变异数目 M_{max} 及当前迭代次数 i。然后初始化特征子集 S_0，在 S_0 上建立 C4.5 分类器，对分类器的性能进行测试，得出分类器的评估值 $F(S_{best})$，且 $S_{best} = S_0$。进入迭代循环之后，在每一次循环过程中，首先计算本次循环需要变异的特征数目 M，然后基于 M 产生新的特征子集 S，在 S 上建立 C4.5 分类器，测试出分类器的评估值 $F(S)$。把 $F(S)$ 与 $F(S_{best})$ 进行比较，如果 $F(S)$ 小于 $F(S_{best})$ 则 $F(S_{best}) = F(S)$，$S_{best} = S$，否则判断评估标准 δ 是否达到。当基于选择的特征子集的分类器的评估值达到了预先的标准或者最大迭代次数已经达到，就停止特征选择。在 S_{best} 上建立分类器，然后把建立的分类器应用于入侵检测系统中。由于应用特征选择算法之后，特征空间下降，特征空间中相关与杂音特征被剔除，基于选择后特征的入侵检测系统的结构变得简洁清晰，检测速度应该会提高很多，入侵检测系统的检测率也有一定程度的提高。下节将通过实验验证。

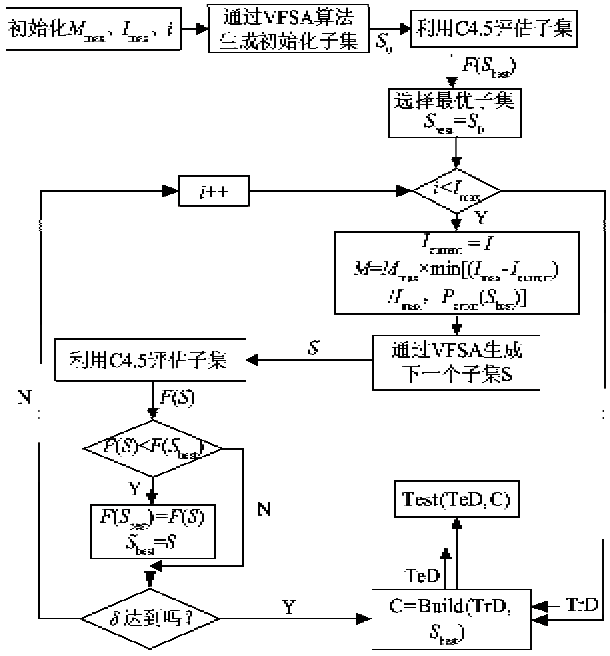

图 1 基于 VFSA-C4.5 的入侵检测框架

3 实验研究

为了验证提出的算法的有效性，一方面，我们测试本文所述特征选择算法用于网络入侵检测的检测效果，主要包括：与相关文献中效果比较好的基于遗传算法和支持向量机的特征选择算法（GA-SVM）、基于相关性和支持向量机的特征选择算法（CFS-SVM）、基于主成分分析和 C4.5 决策树的特征选择算法（PCA-C4.5）的检测率和误报率进行比较；对比使用该算法选择的特征和全部 41 种特征建立检测模型的接收操作特性（ROC）曲线的分值；另一方面，我们从建模时间和检测时间两方面对比该算法

与其它三种算法的检测速度。

3.1　数据集及实验环境

我们采用 KDD1999[16] 作为测试数据集,它是关于入侵检测的一个标准数据集,主要分为训练数据集和测试数据集两部分,该数据集包括大约 490 万条数据记录,每条都是从军方网络环境中模拟攻击所得的原始网络数据中根据设定的 41 个特征提取出来的,它们都是描述网络连接统计信息的特征向量,包含 5 类数据:DoS,Probe,R2L,U2R 四类攻击数据(共包含 24 种攻击类型)以及正常数据。

本文所述的实验都是同一平台下完成,该平台的配置为:Intel processor 1.86GHz,1024MBRAM,Windows XP 操作系统。实验中使用的算法 GA-SVM、CFS-SVM 、PCA-C4.5 和 VFSA-C4.5 均采用机器学习领域著名的开源软件 WEKA[17,18] 进行实验。

3.2　真阳性率和假阳性率的对比

在实验开始之前,我们对 KDD1999 数据集进行了一些预处理工作,以便满足实验的需要,我们将 KDDCup1999 的数据集进行了提取。我们数据集中随机提取了 1986 条正常数据(统一标记为"0")和 2056 条攻击数据(包括上述 4 类攻击,统一标记为"1"),其中攻击数据占整个数据集的1.2%。我们采用了 10 折交叉验证(ten fold cross validation)的方法,重复实验 10 次,取真阳性率(true positive rate,TPR)和假阳性率(false positive rate,FPR)的平均值对几种算法进行了对比。其中,TPR 定义为正确检测出的攻击样本的数量/总的攻击样本的数量,FPR 定义为错误判为攻击的正常样本数量/总的正常样本的数量。

在实验中,我们利用 WEKA 软件对 GA-SVM、CFS-SVM 、PCA-C4.5 和 VFSA-C4.5 四种算法挑选不同的参数进行了多次实验,各取其检测效果最好的结果作为比较。实验结果如表 1 所示。从表中我们不难看出,在训练数据充足的情况下,VFSA-C4.5 方法的检测率要优于其它三种方法,而且 FPR 也相对较低。

表 1　四种不同的算法十字交叉对比实验结果

算法	TPR(%)	FPR(%)
GA-SVM	97.6	1.85
CFS-SVM	97.8	2.33
PCA-C4.5	98.8	0.83
VFSA-C4.5	99.5	0.82

同时,为了验证使用 VFSA-C4.5 特征选择算法的入侵检测比没有使用特征选择算法的入侵检测在检测已知攻击和未知攻击上有更高的 TPR,我们也进行了实验,比较结果如图 2 所示。在图 2 中,我们把所有攻击看作一种类型,并建立两种系统:使用所有特征的入侵检测系统和使用 VFSA-C4.5 选择特征的入侵检测系统。针对每种系统,我们使用两种测试集进行测试,一种是已知攻击(取样的训练集中存在的攻击)测试集,一种是未知攻击(取样的训练集中不存在的攻击)测试集。从图 2 可以看到,在检测已知攻击和未知攻击时,与使用所有特征的入侵检测系统相比,使用 VFSA-C4.5 选择特征的入侵检测系统有更高的 ROC 曲线分值,特别是在检测未知攻击时,ROC 曲线分值更高。

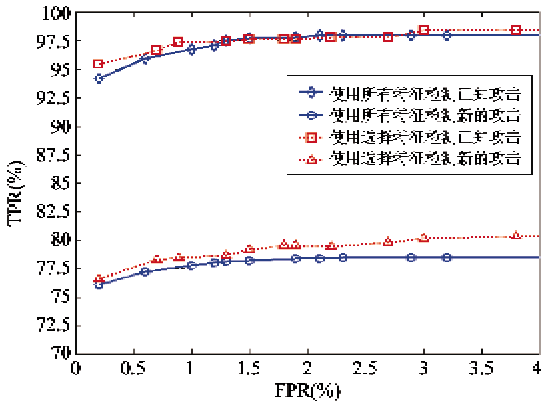

图 2　攻击检测的 ROC 曲线

3.3　检测速度的对比

在训练集上分别使用 GA-SVM、CFS-SVM、PCA-C4.5 和 VFSA-C4.5 四种不同的算法选择最优特征子集,选择的结果如表 2 所示。表中左栏表示应用于入侵检测的四种不同特征选择算法。表中右栏是每种算法选出的最优特征子集。如VFSA-C4.5算

表 2　四种不同的算法选出的最优特征子集

算法	选出的特征
GA-SVM	3,5,12,23,33 : service, src_bytes, logged_in, count, dst_host_srv_count
CFS-SVM	5,6,12,23: src_bytes, dst_bytes, logged_in, count
PCA-C4.5	1,3,5,23,37 : duration, service, src_bytes, count, dst_host_srv_diff_host_rate
VFSA-C4.5	3,5,23,33,34 : service, src_bytes, count, dst_host_srv_count, dst_host_same_srv_rate

法选出的特征子集为 3,5,23,33,34,数字表示该特征在 KDD1999 的 41 个特征中的排序序号。冒号右边的单词是冒号左边的数字对应的特征名称,如 3 对应 service。表中名字的详细解释见 KDD1999[16]。

选出最优特征子集之后,分别在 41 个特征和选出的特征子集上建立入侵检测模型。针对训练数据集,使用不同的阈值,建立多个入侵检测模型,综合评价使用 GA-SVM、CFS-SVM 、PCA-C4.5 和 VFSA-C4.5 四种特征选择算法选择出的特征子集建立的入侵检测模型和使用所有特征子集建立的入侵检测模型在建模时间、检测时间上的不同。入侵检测模型的平均建模时间和检测时间如表 3 所示。从表 3 可以看出:与其它几种特征选择算法相比,使用 VFSA-C4.5 算法建立的入侵检测模型有更少的建模时间和检测时间,检测速度要好于其它算法;与使用所有特征子集的入侵检测模型相比,使用选择特征子集的入侵检测模型有更少的建模时间和检测时间。例如,对于 VFSA-C4.5 入侵检测模型,使用所有特征子集的入侵检测模型平均建模时间和检测时间分别是 78s 和 18s,而使用选择特征子集的入侵检测模型相应的时间分别是 30s 和 6s,分别是使用所有特征子集的入侵检测模型的38.5% 和33.3% 左右。根据建模时间和检测时间的比较结果,可以看到,与 GA-SVM 等相比,我们提出的 VFSA-C4.5 算法消耗更少的计算资源,检测速度更快。同时,使用选择特征子集的模型有更少的建模时间和检测时间也说明特征选择算法有助于提高入侵检测的速度。

表 3 基于四种不同算法的检测模型在所有特征和选择特征上的平均建模时间和检测时间

—		GA-SVM	CFS-SVM	PCA-C4.5	VFSA-C4.5
建模	所有	120	150	85	78
(s)	选出	60	65	33	30
检测	所有	30	45	25	18
(s)	选出	10	15	7	6

4 结 论

现存的对入侵检测的研究主要从两个方面出发:分类器的参数优化和基于数据集的特征选择。本文提出了一种新的应用于入侵检测的特征选择算法——VFSA-C4.5,它主要由搜索策略——快速模拟退火和评估标准——C4.5 决策树两部分组成。快速模拟退火具有全局寻优、降维、速度快的特点,C4.5 决策树在评估时稳定、易于分类,这些都减少了入侵检测的计算复杂度,有助于提高入侵检测的速度。我们在 KDD1999 数据集上进行了一些实验,结果表明,基于 VFSA-C4.5 的入侵检测系统在保证较高检测率的前提下,有效地降低了误报率,它减少了入侵检测的计算复杂度,提高了检测速度,能更适用于现实高速网络应用环境,并已应用于实际的入侵检测系统。

应用于实际的入侵检测系统时,如何将本文所述的算法进一步提高性能,并将其有效地应用于检测 Web 服务器的异常情况,将是我们下一步工作的重点。

参考文献

[1] Zaman S, Karray F. Lightweight IDS based on features selection and IDS classification scheme. In: Proceedings of the 12th IEEE International Conference on Computational Science and Engineering, Vancouver, Canada, 2009. 365-370

[2] Jain A, Zongker D. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, 19(2): 153-158

[3] Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 2000, 33(1): 25-41

[4] Grandvalet Y, Canu S. Adaptive scaling for feature selection in SVMs. *Advances in Neural Information Processing Systems*, 2003, 15: 553-560

[5] Arribas J I, Cid-Sueiro J. A model selection algorithm for a posteriori probability estimation with neural networks. *IEEE Transactions on Neural Networks*, 2005, 16(4): 799-809

[6] Mao K. Feature subset selection for support vector machines through discriminative function pruning analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 2004, 34(1): 60-67

[7] Kim D, Nguyen H N, Ohn S Y. Fusions of GA and SVM for Anomaly Detection in Intrusion Detection System. Advances in Neural Networks. Lecture Notes in Computer Science, New York: Springer-Verlag, 2005, 3498: 415 - 420

[8] Shazzad K M, Jong S P. Optimization of intrusion detection through fast Hybrid feature selection. In: Proceedings of the 6th International Conference on Parallel and Distributed Computing, Applications and Technologies, Washington, DC, USA, 2005. 264-267

[9] Chen Y, Dai L, Li Y, et al. Building efficient intrusion detection model based on principal component analysis and C4.5 algorithm. In: Proceedings of the 9th IEEE International Conference on Advanced Communication Technology, Guilin, China, 2006. 2109-2112

[10] Cherkassky V. The nature of statistical learning theory. *IEEE Transactions on Neural Networks*, 1997, 8(6): 1564-1566

[11] Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by simulated annealing. *Science*, 1983, 220(5):671-679

[12] Nguyen T T, Duong T A. Comparing three improved variants of simulated annealing for optimizing dorm room assignments. In: Proceedings of IEEE International Conference on Computing and Communication Technology, Ho Chi Minh City, Vietnam, 2009. 1-5

[13] Turgut D, Turgut B, Elmasri R, et al. Optimizing clustering algorithm in mobile ad hoc networks using simulated annealing. In: Proceedings of IEEE International Conference on Wireless Communication and Networking, Orlando, USA, 2003. 1492-1497

[14] Vakil-Baghmisheh M T, Navarbaf A. A modified very fast simulated annealing algorithm. In: Proceedings of IEEE International Symposium on Telecommunication, Tehran, Iran, 2008. 61-66

[15] Ruggieri S. Efficient C4.5 [classification algorithm]. *IEEE Transactions on Knowledge and Data Engineering*, 2002,14(2): 438-444

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

[17] Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Computer Communication Review*, 2006,36(5):7-15

[18] 李文法,段洣毅,陈友等. 基于 MRMHC-LSVM 的 IP 流分类. 高技术通讯,2009, 19(6):564-571

VFSA-C4.5 feature selection algorithm for network intrusion detection

Li Chao*,**, Li Wenfa**, Duan Miyi*,**

(*School of Computer Science and Engineering, Beihang University, Beijing 100191)

(**Institute of Computing Technology, Beijing Jiaotong University, Beijing 100029)

Abstract

The VFSA-C4.5 a new feature selection algorithm is proposed to detect network intrusions. The algorithm uses the very fast simulated annealing (VFSA) as the search strategy to specify a candidate subset for evaluation, and then uses the decision tree of C4.5 as the evaluation function to obtain the optimum feature subset for intrusion detection by the data classification error rate. The feasibility of the feature selection algorithm was examined by conducting several experiments on the KDD 1999 intrusion detection dataset. The experimental results show that the VFSA-C4.5 algorithm has higher detection rate and lower false alarm rate compared with other feature selection algorithms for network intrusion detection. Furthermore, the proposed algorithm can reduce computational resources of intrusion detection, improve the detection speed and is more suitable for the real network applications than the traditional ones.

key words: network intrusion detection, feature selection, very fast simulated annealing (VFSA), decision tree