Journal of Computer Applications
计算机应用,2013,33(5):1357 - 1361
ISSN 1001-9081
CODEN JYIIDU
2013-05-01
http://www.joca.cn

文章编号:1001 - 9081(2013)05 - 1357 - 05
doi:10.3724/SP.J.1087.2013.01357

基于量子进化算法的网络入侵检测特征选择

张宗飞*
(台州职业技术学院 计算机工程系,浙江 台州 318000)
(*通信作者电子邮箱 zfzhang2005@126.com)

摘　要:针对当前网络入侵检测中普遍存在检测速度较慢的缺陷,提出了一种新的网络入侵检测特征选择方法。该方法将量子进化算法应用于网络入侵检测的特征选择,从网络连接的原始特征属性中选出一组有效的特征用于入侵检测,以提高检测效率。首先以增强寻优性能为目标改进了量子进化算法,基于特征属性的 Fisher 比构造了特征子集的评价函数,然后按照量子进化算法的流程设计了网络入侵检测特征选择算法。通过 KDD99 样本数据集的实验,表明算法是有效的,既保证了入侵检测的分类性能,也提高了入侵检测的效率。

关键词:网络入侵检测;特征选择;量子进化算法;Fisher 比

中图分类号: TP393.08; TP18　　**文献标志码**:A

Feature selection for network intrusion detection based on quantum evolutionary algorithm

ZHANG Zongfei*
(*Department of Computer Engineering, Taizhou Vocational and Technical College, Taizhou Zhejiang* 318000, *China*)

Abstract: Concerning the disadvantages of slow detection speed in current network intrusion detection, a new feature selection method of network intrusion detection was put forward. The method applied Quantum Evolutionary Algorithm (QEA) to feature selection of network intrusion detection, extracted an optimal subset used in intrusion detection from the original feature set in network connections, so as to get better detection efficiency. First, QEA was improved in order to make its searching performance better, and the criterion function of feature subset was constructed based on the Fisher ratio of feature attributes. Then, the feature selection algorithm of network intrusion detection was designed according to QEA flow. Last, experiments were carried out using the sample data from KDD99. The experimental results show that the proposed algorithm is effective, and it can not only ensure the classification performance of intrusion detection but also improve the detection efficiency.

Key words: network intrusion detection; feature selection; Quantum Evolutionary Algorithm (QEA); Fisher ratio

0　引言

随着网络规模的扩大和网络应用的普及,针对网络的攻击事件越来越多,这些攻击轻则使人们无法正常使用网络,重则带来严重的后果,网络安全已经成为一个全球性关注的问题。网络入侵检测是基于主动防御策略的新一代网络安全技术,它通过分析网络连接中的特征属性来检测是否存在入侵行为,是目前网络安全领域中的重要技术。然而随着网络速度的提升,网络入侵检测面临的一个突出问题是无法实时处理网络传输中海量的数据包。相关研究表明,网络入侵检测时提取和分析处理过多的网络连接特征信息是导致检测速度下降的主要原因之一。据此,一些学者通过特征选择来降低网络连接的特征维数,以解决网络入侵检测中低检测速度的问题,取得了一些成果[1-2]。近年来国内的一些研究者将网络入侵检测的特征选择转化为优化问题,采用优化算法来获得特征子集后用于入侵检测,如文献[3-4]分别使用遗传算法和模拟退火算法来选择特征子集;文献[5-6]使用粒子群算法来获取特征子集,这些特征子集与原始特征集合相比,能够提高入侵检测的效率。然而由于网络连接数据结构比较复杂,特征维数较高,从网络连接的原始特征集合中选择有效的特征子集是具有 NP 难度的组合优化问题,致使这些优化算法的寻优性能受到限制,难以得到有效的特征子集,因此其综合检测性能仍然有待提高。

量子进化算法(Quantum Evolutionary Algorithm,QEA)是量子计算理论与进化计算原理相结合的产物,是一种新发展起来的智能优化算法[7]。QEA 独特的编码和进化机制,使其在求解组合优化问题中显现出优越的性能。本文使用 QEA 来选择网络连接中的特征属性,利用 QEA 的全局寻优能力对特征空间进行全面搜索,剔除无关的、冗余的特征属性,获得最优特征子集用于入侵检测,以提高检测效率。

1　特征选择

特征选择问题的数学模型:对于给定的一个特征集合 $\boldsymbol{F}=\{f_1,f_2,\cdots,f_n\}$,$n$ 是特征集的大小,它的一个特征子集可以用一个二进制向量表示:$\boldsymbol{S}=\{s_1,s_2,\cdots,s_n\}(s_i\in\{0,1\},i=1,2,\cdots,n)$,$s_i=1$ 表示第 i 个特征 f_i 被选择;反之,表示第 i 个特征 f_i 不被选择。特征选择需要解决两个关键问题,即设计搜索策略和评价函数,作用是生成和评价特征子集[8]。

1.1　搜索策略的设计

目前,特征选择的基本搜索策略有三种:穷举搜索、随机搜索和启发式搜索,三种搜索策略各有优缺点,在实际应用中经常组合使用[8]。QEA 作为一种概率随机搜索算法,由于在计算过程中采用了启发式搜索策略,使得搜索空间远远小于 $O(2^n)$,在特征选择中具有独特的优势,然而 QEA 在求解复

收稿日期:2012-11-23;**修回日期**:2013-01-07。　**基金项目**:浙江省教育厅科研项目(Y201225119)。
作者简介:张宗飞(1971 -),男,浙江台州人,讲师,硕士,CCF 会员,主要研究方向:进化算法、网络安全。

杂优化问题时,其优化性能还不够理想,因此对于本文的特征选择问题,需要先对 QEA 进行改进,然后使用改进的 QEA 进行求解。

种群进化是 QEA 寻优的驱动机制,通常使用式(1)的量子旋转门通过式(2)更新量子位来实现进化操作,更新过程中 θ_i 的值通过事先定义的旋转角度表获得[7],表中给出的角度值是离散的且固定的,对染色体量子位的更新操作限制在几个固定状态。

$$\boldsymbol{U}(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{1}$$

其中 θ 为旋转角度。

$$\begin{bmatrix} \alpha_i^{t+1} \\ \beta_i^{t+1} \end{bmatrix} = \boldsymbol{U}(\theta_i) \begin{bmatrix} \alpha_i^t \\ \beta_i^t \end{bmatrix} = \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix} \begin{bmatrix} \alpha_i^t \\ \beta_i^t \end{bmatrix} \tag{2}$$

其中:$[\alpha_i^t, \beta_i^t]^{\mathrm{T}}$ 为更新染色体第 i 个量子位的概率幅;$[\alpha_i^{t+1}, \beta_i^{t+1}]^{\mathrm{T}}$ 为更新后染色体第 i 个量子位的概率幅;$\theta_i = s(\alpha_i\beta_i) \cdot \Delta\theta_i$,$\Delta\theta_i$、$s(\alpha_i\beta_i)$ 分别表示旋转角的大小和方向。

在复杂优化问题中,这种有限状态的更新操作的寻优能力有限,难以搜索到全局最优解,而且查找 θ_i 需要经过复杂的多路条件判断,影响了算法的效率。因此,改进 θ_i 的计算方法是提高 QEA 寻优性能的关键,也是目前 QEA 改进的主要方向[9-10]。本文的改进策略设计过程如下:

1)提出量子位相位角的角距离概念。

对于组合优化问题,通常将量子位的概率幅简化为用实数表示,因此本文用一对满足归一化条件的实数 (α,β) 表示一个量子位的概率幅,此时量子位可以映射为二维实空间上的单位向量,并且第二、三、四象限上的量子位均可以投影到第一象限上,据此给出量子位相位角和角距离的定义。

定义 1 假设给定的量子位 $|\varphi\rangle$ 的概率幅为 $[\alpha,\beta]^{\mathrm{T}}$($\alpha$,$\beta$ 为实数),称 $\arctan(|\beta|/|\alpha|)$ 为 $|\varphi\rangle$ 的相位角,记为 $\omega_{|\varphi\rangle}$($\omega_{|\varphi\rangle} \in [0, \pi/2]$)。

$$\omega_{|\varphi\rangle} = \begin{cases} \arctan(|\beta|/|\alpha|), & |\alpha| \neq 0 \\ \pi/2, & |\alpha| = 0 \end{cases} \tag{3}$$

由定义 1 可知 $|0\rangle$ 的相位角值为 0,$|1\rangle$ 的相位角值为 $\pi/2$,即 $\omega_{|0\rangle} = 0$,$\omega_{|1\rangle} = \pi/2$。

定义 2 假设量子位 $|\varphi_1\rangle$ 的相位角为 $\omega_{|\varphi_1\rangle}$,$|\varphi_2\rangle$ 的相位角为 $\omega_{|\varphi_2\rangle}$,称 $\omega_{|\varphi_2\rangle} - \omega_{|\varphi_1\rangle}$ 为 $|\varphi_1\rangle$ 到 $|\varphi_2\rangle$ 的角距离,记为 $\Delta\theta_{|\varphi_1\rangle \to |\varphi_2\rangle}$。

显然,角距离是一个矢量,包含大小和符号,大小表示偏转幅度,符号表示偏转方向,+ 表示逆时针方向,- 表示顺时针方向。由定义 2 可得两个特殊的角距离如式(4)、(5)所示:

$$\Delta\theta_{|\varphi\rangle \to |0\rangle} = \omega_{|0\rangle} - \omega_{|\varphi\rangle} = -\arctan(|\beta|/|\alpha|) \tag{4}$$

$$\Delta\theta_{|\varphi\rangle \to |1\rangle} = \omega_{|1\rangle} - \omega_{|\varphi\rangle} = \pi/2 - \arctan(|\beta|/|\alpha|) \tag{5}$$

2)使用角距离计算旋转角。

分析式(2)的量子位更新过程可以发现,在量子旋转门的作用下量子位的概率幅发生了变化,从而改变了量子位取 0 与取 1 的概率,其实质就是使量子位发生偏转,据此提出一种基于角距离的动态调整 θ_i 策略。旋转角 θ_i 按照式(6)、(7)计算。

$$\theta_i = (1 - f_x/f_b)\,\Delta\theta_{|\varphi_i\rangle \to} \tag{6}$$

其中:f_x 为更新个体的适应度值,f_b 为当前最优个体的适应度值。

$$\Delta\theta_{|\varphi_i\rangle \to} = \begin{cases} \Delta\theta_{|\varphi_i\rangle \to |0\rangle} = -\arctan(|\beta_i|/|\alpha_i|), \\ \qquad (f_b \geqslant f_x) \wedge (b_i = 0) \wedge (x_i = 1) \\ \Delta\theta_{|\varphi_i\rangle \to |1\rangle} = \pi/2 - \arctan(|\beta_i|/|\alpha_i|), \\ \qquad (f_b \geqslant f_x) \wedge (b_i = 1) \wedge (x_i = 0) \\ 0, \quad \text{其他} \end{cases} \tag{7}$$

其中:x_i 为更新个体第 i 个量子位的观测值,b_i 为当前最优个体第 i 个量子位的观测值。

通过式(6)、(7)计算旋转角 θ_i 时,得到 $\Delta\theta_{|\varphi_i\rangle \to}$ 的符号能够使当前染色体的所有量子位向着最优个体对应量子位的基态偏转,保证了算法的进化方向;而 θ_i 的大小能够根据当前待更新量子位的状态和染色体的适应度值自适应地计算合适的角度值,从而保证了算法的进化性能。可见由式(6)、(7)计算旋转角 θ_i 的过程是动态的和连续的,对问题解空间的搜索更加全面和精细,而且计算简单,容易理解。

3)使用 H 门修正量子位概率幅。

分析式(6)、(7)的调整策略,可以得知这种调整方式仍然存在局限性,当个体与当前最优个体非常远时,f_x/f_b 的值将非常小,此时调整幅度就非常大,有可能会使量子位的概率幅迅速趋向于 0 或 1,导致观测值过早收敛,尤其在进化早期,大部分个体适应度很低,过大幅度的调整会使这些个体过早收敛,陷入早熟。为了克服这一缺陷,采用 H 门对更新后的量子位概率幅进行修正[11],修正策略为设定一个极小阈值 $0 < \varepsilon \ll 1$,当更新后量子位的 $|\alpha|^2$ 或 $|\beta|^2$ 小于阈值 ε 时,采用式(8)对该量子位进行修正。

$$[\alpha_i^{t+1}, \beta_i^{t+1}]^{\mathrm{T}} = \begin{cases} [\sqrt{\varepsilon}, \sqrt{1-\varepsilon}]^{\mathrm{T}}, |\alpha_i^{t+1}|^2 \leqslant \varepsilon \wedge |\beta_i^{t+1}|^2 \geqslant 1-\varepsilon \\ [\sqrt{1-\varepsilon}, \sqrt{\varepsilon}]^{\mathrm{T}}, |\beta_i^{t+1}|^2 \leqslant \varepsilon \wedge |\alpha_i^{t+1}|^2 \geqslant 1-\varepsilon \\ [\alpha_i^{t+1}, \beta_i^{t+1}]^{\mathrm{T}}, \quad \text{其他} \end{cases} \tag{8}$$

从式(8)可以看出,对量子位的修正操作是在量子位的 $|\alpha|^2$ 或 $|\beta|^2$ 的值过于接近 0 或 1 时对其进行修正,从而增强观测种群的多样性,避免算法陷入早熟收敛。

4)使用染色体交叉操作增强种群活力。

QEA 的缺陷之一是在进化后期有时会陷入停滞状态,导致算法的寻优效率降低,采用染色体交叉操作对此进行改进,基本思想是当算法出现停滞时,通过交叉给种群个体施以较小的扰动,改变染色体部分量子位的状态,从而增强种群活力,维持算法继续进化搜索。算法停滞的判断准则是:给定一个较小阈值 $0 < \delta \ll 1$,若连续 5 代进化相邻两代种群的平均适应度值的变化均小于 δ,则认为算法已经出现了停滞,应该采取交叉操作。染色体交叉的步骤是:

①对种群中的染色体进行随机排序;

②随机确定 $m(0 < m \leqslant 5)$ 个量子位作为交叉点;

③对交叉点的量子位概率幅循环移位 n(n 为染色体个数)次。

1.2 评价函数的设计

特征选择根据特征子集评价准则与后续分类器的关系分为过滤式(Fiter)和封装式(Wrapper)两种[8]。考虑到网络入侵检测中数据量大,数据特征维数较高,因此采用基于距离度量的过滤式评价准则,使用 Fisher 比来设计特征子集的评价函数,主要设计思想是分类能力强的特征表现为类内距离尽可能小,类间距离尽可能大。

对于 d 维特征空间中由 n 个样本构成的数据集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$，$x_i(i=1,2,\cdots,n)\in\mathbf{R}^d$，划分为 c 个类:$C_1,C_2,\cdots,C_c$,每一类中包含 n_i 个样本，$\sum_{i=1}^{c}n_i=n$。

定义 3　第 k 维特征在样本集上的类间离散度 $S_{\mathrm{b}}^{(k)}$ 与类内离散度 $S_{\mathrm{w}}^{(k)}$ 的比值:$S_{\mathrm{b}}^{(k)}/S_{\mathrm{w}}^{(k)}$ 称为该特征的 Fisher 比。

$$S_{\mathrm{b}}^{(k)}=\sum_{j=1}^{c}\frac{n_j}{n}(m_j^{(k)}-m^{(k)})^2 \tag{9}$$

$$S_{\mathrm{w}}^{(k)}=\sum_{j=1}^{c}\left(\frac{1}{n_j}\sum_{x\in C_j}(x^{(k)}-m_j^{(k)})^2\right) \tag{10}$$

其中:$x^{(k)}$ 表示样本 x 的第 k 维特征值,$m_j^{(k)}$ 和 $m^{(k)}$ 分别表示第 j 类样本和所有样本的第 k 维特征的均值;$S_{\mathrm{b}}^{(k)}$ 表示不同类样本间的距离,$S_{\mathrm{w}}^{(k)}$ 表示同类样本间的距离。

定义 4　根据定义 3,由 k 个特征构成的特征子集在样本集上的类间离散度 $S_{\mathrm{b}}^{(Rk)}$ 与类内离散度 $S_{\mathrm{w}}^{(Rk)}$ 的比值:$S_{\mathrm{b}}^{(Rk)}/S_{\mathrm{w}}^{(Rk)}$ 称为该特征子集的 Fisher 比。

$$S_{\mathrm{b}}^{(Rk)}=\sum_{i=1}^{k}\sum_{j=1}^{c}\frac{n_j}{n}\left(m_j^{(i)}-m^{(i)}\right)^2 \tag{11}$$

$$S_{\mathrm{w}}^{(Rk)}=\sum_{i=1}^{k}\sum_{j=1}^{c}\left(\frac{1}{n_j}\sum_{x\in C_j}\left(x^{(i)}-m_j^{(i)}\right)^2\right) \tag{12}$$

为了简化计算,本文将网络入侵检测数据样本分为 2 类:正常数据类和入侵数据类,分别称为正类样本和负类样本,此时问题简化为二分类问题。对于上述的样本数据集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$,将 $\boldsymbol{X}$ 中正类样本集记为 $\boldsymbol{X}_1$,负类样本集记为 $\boldsymbol{X}_2$,n_1 为正类样本数,n_2 为负类样本数,根据定义 4 可得:

$$S_{\mathrm{b}}^{(Rk)}=\sum_{i=1}^{k}\left(\frac{n_1}{n}(m_1^{(i)}-m^{(i)})^2+\frac{n_2}{n}(m_2^{(i)}-m^{(i)})^2\right) \tag{13}$$

$$S_{\mathrm{w}}^{(Rk)}=\sum_{i=1}^{k}\left(\frac{1}{n_1}\sum_{x\in X_1}(x^{(i)}-m_1^{(i)})^2+\frac{1}{n_2}\sum_{x\in X_2}(x^{(i)}-m_2^{(i)})^2\right) \tag{14}$$

由定义 3、定义 4 可知,Fisher 比表示特征对分类的贡献度,一个特征子集的 Fisher 比越大,表示该特征子集的分类能力越强。因此,针对网络入侵检测数据集,特征子集的评价函数可设计为:

$$F=\sum_{i=1}^{k}\left(\frac{n_1}{n}(m_1^{(i)}-m^{(i)})^2+\frac{n_2}{n}(m_2^{(i)}-m^{(i)})^2\right)\Big/\sum_{i=1}^{k}\left(\frac{1}{n_1}\sum_{x\in X_1}(x^{(i)}-m_1^{(i)})^2+\frac{1}{n_2}\sum_{x\in X_2}(x^{(i)}-m_2^{(i)})^2\right) \tag{15}$$

2　基于量子进化的网络入侵检测特征选择算法

算法设计的基本思想是以网络连接中的特征属性为原始特征集,使用文中改进的量子进化算法对其逐代优化,从而获得一个最优特征子集。

2.1　算法设计

根据 QEA 的基本流程,设计算法过程如下。

1)种群个体编码及初始化。

根据特征选择的数学模型,一个特征子集表示为一个二进制向量,如果一个网络连接的特征属性数为 m,在 QEA 中一个特征子集就可以用一个长度为 m 的量子染色体来表示,每一个量子位对应一个特征,此时一个量子位携带着 3 个信息:量子位在量子染色体中的位置、量子位的观测值和量子位对应的特征值。

随机产生 n 个式(16)所示的长度为 m 的量子染色体构成初始种群,然后将量子染色体的各量子位初始化为等概率状态,即将量子位概率幅$[\alpha_i,\beta_i]^{\mathrm{T}}$初始化为$(1/\sqrt{2},1/\sqrt{2})$。

$$\boldsymbol{q}=\begin{bmatrix}\alpha_1&\alpha_2&\cdots&\alpha_m\\ \beta_1&\beta_2&\cdots&\beta_m\end{bmatrix} \tag{16}$$

2)种群观测。

种群观测的目的是使量子染色体的各量子位从不确定的概率态坍塌到确定的基态,获得观测态个体,其形式是长度为 m 的二进制串$(x_1x_2\cdots x_m)(x_i\in\{0,1\})$。观测方法是:对于量子染色体的每一个量子位,随机产生$[0,1]$的一个常数 r,若 $r\geqslant|\alpha_i|^2$(α_i 为染色体第 i 个量子位的概率幅),则该量子位的观测值(x_i)取 1,否则取 0。

3)适应度函数设计。

QEA 中使用适应度函数评价个体的优劣,指导个体的进化,个体的适应度函数值越大,表示个体越优秀。本文算法中一个个体就代表一个特征子集,其优劣可以用 2 个指标来衡量:特征子集对训练集样本的分类能力,即特征子集的 Fisher 比;特征子集的特征维数,即特征子集对应个体中“1”的个数。算法要寻找的是分类能力强、特征维数少的特征子集,因此适应度函数可以设计为:

$$f=F/(1+n_p) \tag{17}$$

其中:F 为个体对应的特征子集的 Fisher 比,n_p 为个体中“1”的个数。

4)种群进化。

种群进化通过染色体的更新来实现,按照式(2)、(6)、(7)进行更新计算和式(8)进行修正计算。当符合连续 5 代进化相邻两代种群的平均适应度值的变化均小于交叉阈值 δ 的条件时,按照文中提出的交叉策略进行染色体交叉操作。

5)停止条件设定。

算法的求解目标是获得最优特征子集,因此当搜索到最优特征子集所对应的那个最优个体时,算法就可以停止循环,本文将此最优个体的判定条件设定为:若算法连续 5 代进化中每一代最优个体适应度值与当前保存的最优个体适应度值的差异小于给定的极小阈值 λ 时,则认为当前最优个体就是最优特征子集,算法停止循环。

2.2　算法伪代码

```
//输入:原始特征集,训练样本
//输出:最优特征子集
初始化种群,得到 Q[n,(α,β)_m]
t:=0
for i:=0 to 4                                   //先进化 5 代
  观测 Q[n,(α,β)_m],得到 P[n,{0,1}_m]
  计算 P[n,{0,1}_m]中各个体的适应度及平均适应度,存于
    f[n]与f[i]
  记录当代最优个体及其适应度,存于 b[m]及fb[i]
  if fb[i] > fB  then  fB:=fb[i]; B[m]:=b[m] end if
  更新 Q[n,(α,β)_m]中各个体,得到新一代种群 Q[n,(α,
    β)_m]
  t:=t+1
end for
while not (|fb[4]-fB|<λ ∧|fb[3]-fB|<λ ∧
  |fb[2]-fB|<λ ∧|fb[1]-fB|<λ ∧|fb[0]-fB|<λ)
  if (|f[4]-f[3]|<δ ∧|f[3]-f[2]|<δ ∧
    |f[2]-f[1]|<δ ∧|f[1]-f[0]|<δ) then
    对Q[n,(α,β)_m]中个体进行交叉操作,存于 Q[n,(α,
      β)_m]
```

end if
$fb[0]:=fb[1];fb[1]:=fb[2];$
$fb[2]:=fb[3];fb[3]:=fb[4]$
$\overline{f}[0]:=\overline{f}[1];\overline{f}[1]:=\overline{f}[2];$
$\overline{f}[2]:=\overline{f}[3];\overline{f}[3]:=\overline{f}[4]$
观测 $Q[n,(\alpha,\beta)_m]$,得到 $P[n,\{0,1\}_m]$
计算 $P[n,\{0,1\}_m]$ 中各个体的适应度及平均适应度,存于 $f[n]$ 与 $\overline{f}[i]$
记录当代最优个体及其适应度,存于 $b[m]$ 及 $fb[i]$
if $fb[i]>fB$ then $fB:=fb[i];B[m]:=b[m]$ end if
更新 $Q[n,(\alpha,\beta)_m]$ 中各个体,得到新一代种群 $Q[n,(\alpha,\beta)_m]$
$t:=t+1$
end while
解码 $B[m]$,得到最优特征子集并输出

3 实验与结果分析

3.1 实验准备

3.1.1 实验数据

实验使用 KDD99 网络入侵检测数据集[12],首先将 KDD99 中带入侵标记的 10% 训练数据集 kddcup. data-10-percent. gz 和测试数据集 corrected. gz 合并,然后随机选取 14 个子集作为实验的样本数据集,每个子集约 10 000 个实例,其中正常实例与各类型的入侵实例保持原数据集的分布比例,各子集的样本实例分布如表 1 所示。

表 1 实验数据集样本实例分布

数据集	实例总数	正常实例数	入侵实例数
TrD1	10012	1959	8053
TrD2	10009	1957	8052
TrD3	10013	1970	8043
TrD4	10005	1966	8039
TeD1	10021	1961	8060
TeD2	10017	1958	8059
TeD3	10027	1969	8058
TeD4	10025	1971	8054
TeD5	10022	1967	8055
TeD6	10002	1958	8044
TeD7	10009	1962	8047
TeD8	10019	1965	8054
TeD9	10011	1969	8042
TeD10	10027	1959	8068

3.1.2 数据预处理

KDD99 中的数据以网络连接的形式保存,每条数据记录有 41 个特征属性:3 个符号型特征属性,38 个数值型特征属性。实验过程中涉及距离计算,本文采用样本方差来表示距离度量,需要对样本的符号特征值和数值特征值进行处理。

1)符号特征数值化。对于符号特征,无法直接使用方差表达式计算,因此需要将其数值化,方法是采用十进制对符号特征值进行编码,将其映射到数值空间。例如 2 号特征 protocol_type,特征值为:TCP、UDP、ICMP,十进制编码后得:0、1、2,以此转化为数值型特征。

2)数值特征归一化。对于数值特征,样本实例中不同的特征使用不同的度量标准,如果直接使用方差表达式计算距离,就会产生大数吃小数问题,导致某些特征被掩盖,因此需要将其归一化,方法是将原始的各特征值(包括数值化后的符号特征)按下列步骤计算,将其映射到一个标准的特征值空间。

①计算样本集所有样本第 k 维特征的平均绝对偏差 $\overline{m}^{(k)}$:

$$\overline{m}^{(k)}=\frac{1}{n}\sum_{i=1}^{n}(x_i^{(k)}-m^{(k)}) \tag{18}$$

其中:$x_i^{(k)}$ 为样本集中第 i 个样本的第 k 维特征值,$m^{(k)}$ 为样本集中所有样本第 k 维特征的平均值:

$$m^{(k)}=\frac{1}{n}\sum_{i=1}^{n}x_i^{(k)} \tag{19}$$

②计算样本集第 i 个样本第 k 维特征的标准化值 $\overline{x}_i^{(k)}$:

$$\overline{x}_i^{(k)}=(x_i^{(k)}-m^{(k)})/\overline{m}^{(k)} \tag{20}$$

3.1.3 实验环境

实验的硬件环境是 Intel Pentium 4 CPU 2.60 GHz、1 GB内存,操作系统为 Microsoft Windows XP SP2,编程环境为 JDK1.6.0_24 + Eclipse 3.2。

3.2 实验过程与分析

实验按照图 1 所示的流程进行,使用 Java 语言并载入 Weka 工具完成算法代码。

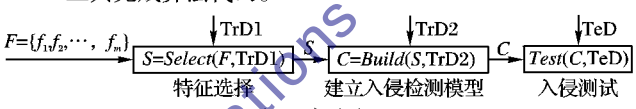


图 1 实验流程

3.2.1 特征选择实验

以 TrD1 为训练样本、样本实例中的 41 维特征属性为原始特征集,按照文中设计的算法进行特征选择实验,算法参数设置为:种群大小为 30,算法停止条件阈值 λ 为 0.005,染色体交叉条件阈值为 0.01,量子位修正阈值 ε 为 0.01。特征选择结果如表 2 所示。

表 2 特征选择结果

解名称	解名称对应的值
最优个体(观测态)	0100110000010000000000100000000011001100000
特征序号	2,5,6,12,23,32,33,36,37
特征名称	protocol_type, src_bytes, dst_bytes, logged_in, count, dst_host_count, dst_host_srv_count, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate

3.2.2 特征子集性能测试实验

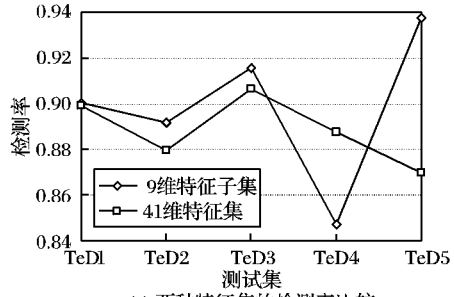
为了验证 3.2.1 节实验中获得的 9 维特征子集的入侵检测性能,首先以 TrD2 为训练样本,使用 9 维特征子集在 Weka 工具中的 J48 分类器上建立入侵检测模型,然后分别在 TeD1 ~ TeD5 上进行入侵检测测试,并采用入侵检测模型建立时间、入侵测试时间、检测率、误报率为评价指标,与基于原始特征集的 41 维特征的实验进行比较。实验结果如表 3 和图 2 所示。

表 3 两种特征集的建模时间和入侵测试时间 s

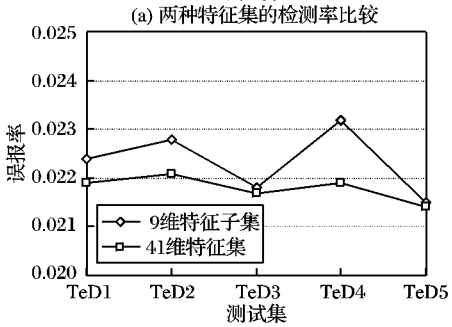
特征集	建模时间	入侵测试时间				
		TeD1	TeD2	TeD3	TeD4	TeD5
9 维特征子集	5.7364	0.1951	0.2107	0.1599	0.1953	0.2160
41 维特征集	14.0822	0.4549	0.3999	0.4500	0.4903	0.4801

表 3 数据表明,基于 9 维特征子集建立的入侵检测模型与基于原始 41 维特征建立的入侵检测模型比较,在建模时间和入侵测试时间上均有明显降低;在检测率和误报率方面,由图 2 可以得出前者的检测率比后者平均提高了 0.996%,但是误报率略差,平均高出后者 0.054%。综合上述 4 个指标

来看,3.2.1 节获得的特征子集的性能是比较好的,它不仅保证了分类性能,而且还减轻了分类器的负担。



(a) 两种特征集的检测率比较

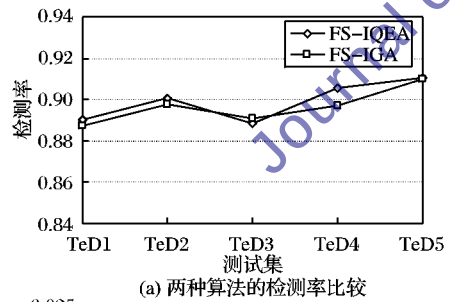


(b) 两种特征集的误报率比较

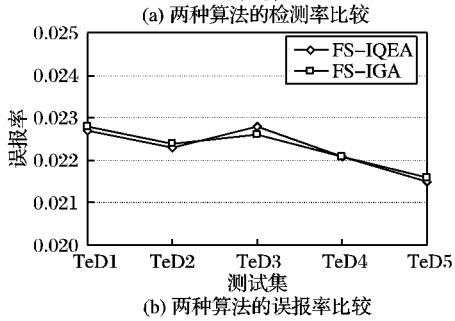
图 2　两种特征集的检测率和误报率

3.2.3　算法有效性验证实验

为了验证本文特征选择算法(FS-IQEA)的有效性,将 FS-IQEA 与文献[3]的基于遗传算法的特征选择算法(FS-IGA)进行比较。实验首先以 TrD3 为训练样本,使用 FS-IQEA 对样本的 41 维特征进行选择,然后使用得到的特征子集,通过训练集 TrD4 在 J48 分类器上建立入侵检测模型,并分别在 TeD6 ~ TeD10 上进行入侵检测测试;接着在相同条件和方法下,进行基于 FS-IGA 的实验。前者算法参数按 3.2.1 节实验设置,后者算法参数按文献[3]设置。实验结果比较如表 4 和图 3 所示。



(a) 两种算法的检测率比较



(b) 两种算法的误报率比较

图 3　两种算法的检测率和误报率

表 4　两种算法的特征选择时间、建模时间和入侵测试时间　　s

算法	特征选择时间	入侵检测建模时间	入侵测试时间				
			TeD1	TeD2	TeD3	TeD4	TeD5
FS-IQEA	47.3277	5.9903	0.2030	0.1870	0.2197	0.1689	0.2169
FS-IGA	85.0130	6.8015	0.2067	0.1908	0.2090	0.1901	0.2089

实验中,FS-IQEA 选择的特征子集是 10 维,FS-IGA 选择的特征子集是 13 维。从表 4 数据可以看出,算法在选择特征时花费的时间,FS-IQEA 比 FS-IGA 要少 79%;基于 10 维特征建立的入侵检测模型比基于 13 维特征建立的入侵检测模型在建模时间和入侵测试时间上的开销均要略少;图 3 显示前者的检测率也比后者要略好,而两者的误报率相当。综合这些结果可以得出,在网络入侵检测特征选择中,FS-IQEA 的寻优能力和寻优速度均优于 FS-IGA。

4　结语

网络数据包中含有大量无关的、冗余的特征信息,如果使用其全部特征来进行入侵检测,将会降低检测效率,据此本文使用量子进化算法来减少网络连接中的特征维数,提出了基于改进量子进化算法的特征选择算法。在 KDD99 数据集的实验中,算法可以选出 10 维左右的特征子集,与原始特征集比较,使用特征子集在建立入侵检测模型和入侵检测的效率上都有显著提高,并且算法的特征选择效率要优于文献[3]算法。算法的不足之处是选出的特征子集在两次入侵检测实验中的误报率表现都不具优越性,这是今后需要进一步研究之处。

参考文献:

[1] DARTIQUE C, JANG H, ZENG W. A new data-mining based approach for network intrusion detection[C]// Proceedings of the 7th Annual Conference on Communication Networks and Services Research. Washington, DC: IEEE Computer Society, 2009: 372 - 377.

[2] ZAMAN S, KARRAY F. Lightweight IDS based on features selection and IDS classification scheme[C]// Proceedings of the 12th International Conference on Computational Science and Engineering. Washington, DC: IEEE Computer Society, 2009: 365 - 370.

[3] 朱红萍, 巩青歌, 雷战波. 基于遗传算法的入侵检测特征选择[J]. 计算机应用研究, 2012, 29(4): 1417 - 1419.

[4] 李超, 李文法, 段洣毅. 用于网络入侵检测的 VFSA-C4.5 特征选择算法[J]. 高技术通讯, 2011, 21(12): 1420 - 1425.

[5] 郑洪英, 侯梅菊, 王渝. 入侵检测中的快速特征选择方法[J]. 计算机工程, 2010, 36(6): 262 - 264.

[6] 吴春琼. 基于特征选择的网络入侵检测模型[J]. 计算机仿真, 2012, 29(6): 136 - 139.

[7] HAN K H, KIM J H. Quantum-inspired evolutionary algorithm for a class of combinatorial optimization [J]. IEEE Transactions on Evolutionary Computation, 2002, 6(6): 580 - 593.

[8] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2): 162 - 166.

[9] LI Y Y, ZHAO J J, JIAO L C. Quantum-inspired evolutionary multicast algorithm[C]// Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. Piscataway: IEEE Press, 2009: 1496 - 1501.

[10] 张宗飞. 优化网络入侵特征库的量子进化算法[J]. 计算机应用, 2010, 30(8): 2142 - 2145.

[11] HAN K H, KIM J H. Quantum-inspired evolutionary algorithm with a new termination criterion, H_ε gate and two-phase scheme[J]. IEEE Transactions on Evolutionary Computation, 2004, 8(4): 156 - 169.

[12] University of California, Irvine. KDD cup 1999 data[DB/OL]. [211 - 07 - 20]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.