第37卷 第14期 计 算 机 工 程 2011年7月
Vol.37 No.14 Computer Engineering July 2011

·安全技术· 文章编号：1000—3428(2011)14—0103—03 文献标识码：A 中图分类号：TP311.52

基于GQPSO算法的网络入侵特征选择方法

牟 琦，毕孝儒，厍向阳

(西安科技大学计算机学院，西安 710054)

摘 要：高维网络数据中的无关属性和冗余属性容易使分类算法的网络入侵检测速度变慢、检测率降低。为此，提出一种基于遗传量子粒子群优化(GQPSO)算法的网络入侵特征选择方法，该方法将遗传算法中的选择变异策略与QPSO有机结合形成GQPSO算法，并以网络数据属性之间的归一化互信息量作为该算法适应度函数，指导其对网络数据的属性约简，实现网络入侵特征子集的优化选择。在KDDCUP1999数据集上进行仿真实验，结果表明，与QPSO算法、PSO算法相比，该方法能更有效地精简网络数据特征，提高分类算法的网络入侵检测速度及检测率。

关键词：GQPSO算法；归一化互信息；适应度函数；特征选择；网络入侵检测

Feature Selection Method for Network Intrusion Based on GQPSO Algorithm

MU Qi, BI Xiao-ru, SHE Xiang-yang

(School of Computer, Xi'an University of Science and Technology, Xi'an 710054, China)

【Abstract】Aiming at problem that independent and redundant attributes of high dimensional network data cause classification algorithms' slow detection speed and low detection rate in network intrusion detection, a feature selection approach for network intrusion based on Genetic Quantum Particle Swarm Optimization(GQPSO) algorithm is proposed. The approach organically combines selection and variation of genetic algorithm with QPSO to form GQPSO algorithm, and normalizes mutual information between attributes of network data is defined as the algorithm's fitness function, which guides its reduction of network data attributes to realize optimal selection of network intrusion feature sub-set. Simulation experiment is done in KDDCUP1999. Result shows that compared with QPSO and PSO algorithms, the approach is more effective for feature selection of network data and improvement of network intrusion detection speed and detection rate of classification algorithms.

【Key words】Genetic Quantum Particle Swarm Optimization(GQPSO) algorithm; normalized mutual information; fitness function; feature selection; network intrusion detection

DOI: 10.3969/j.issn.1000-3428.2011.14.033

1 概述

在网络入侵检测中，由于高维数据中无关和冗余属性的存在，致使分类检测算法的检测速度慢、检测率不高，限制了其在实际中的应用，因此如何对网络数据属性进行选择，获取优化特征子集，提高分类算法的检测性能成为当前网络入侵检测研究热点，文献[1-2]将粒子群优化(Particale Swarm Optimization, PSO)算法应用于网络入侵检测数据的特征选择。文献[3]将免疫思想与粒子群优化(Immune and Particale Swarm Optimization, IPSO)算法相结合，应用于入侵检测数据的属性选择中，该方法在一定程度上保持了粒子多样性，提高了PSO算法的收敛精度。文献[4]提出基于距离准则的适应度函数，并用其指导PSO对网络数据的特征优化选择，较好地提高了分类算法质量。文献[5]提出基于量子粒子群优化(Quantum Particale Swarm Optimization, QPSO)的网络入侵特征选择方法，以QPSO算法较好的全局寻优能力实现了网络入侵特征子集的优化选择。但以上特征选择算法均存在迭代后期收敛速度慢、且容易陷入局部最优的缺陷。

本文提出一种基于遗传量子粒子群优化算法(Genetic Quantum Particale Swarm Optimization, GQPSO)的网络入侵特征选择方法，该方法将遗传算法的选择和变异思想与QPSO相结合，并在新的适应度函数指导下，实现网络入侵特征的优化选择。

2 基于GQPSO的网络入侵特征选择算法

2.1 遗传选择、变异操作

针对QPSO算法在迭代后期收敛速度慢和易陷入局部最优的不足，本文将遗传算法中的选择变异操作引入QPSO算法。其基本思想是，将种群中每个粒子的适应度值与当前种群平均适应度值 AF 进行比较，大于 AF 的粒子予以保留，对于适应度值小于 AF 的粒子，对其每一位以概率 P_m 进行取反变异，即如 $rand() > AF$ ，则保持不变，否则对该位进行取反操作，以提高整个种群的收敛速度，避免QPSO算法陷入局部最优值。

2.2 适应度函数的构造

在智能优化特征选择算法中，适应度评价函数的选择至关重要。本文采用基于粗糙集联合信息熵的归一化互信息的评测方法。其核心思想是，选择一个属性子集，属性各自与类属性有较大的关联但几乎没有内部关联，以此达到消除无关属性和冗余属性的目的。设2个属性 X 和 Y ，则其相关性

基金项目：陕西省自然科学基金资助项目(2009JM7007)

作者简介：牟 琦(1974－)，女，副教授，主研方向：网络安全，网络集成，数据库技术；毕孝儒，硕士研究生；厍向阳，副教授、博士后

收稿日期：2011-03-02 **E-mail：**bi_xiao_ru@sina.com

强弱可用归一化互信息来度量：

$$SU(X,Y)=\frac{2\cdot I(X;Y)}{H(X)+H(Y)} \tag{1}$$

其中，$I(X;Y)=H(X)+H(Y)-H(X,Y)$ 是属性 X 和 Y 的互信息；$H(X),H(Y)$ 是熵函数，以每个属性值的概率为基础，其定义如下：

$$H(X)=-\sum_{i=1}^{n}p(a_i)\mathrm{lb}p(a_i) \tag{2}$$

$H(X,Y)$ 是 X 和 Y 的联合熵，由 X 和 Y 的所有组合值的联合概率计算出来，其定义如下：

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(a_i,b_j)\mathrm{lb}p(a_i,b_j) \tag{3}$$

则基于归一化互信息的特征选择决定一个属性子集的优良与否可用下式度量：

$$\frac{\sum_j SU(X_j,C)}{\sqrt{\sum_i\sum_j SU(X_i,X_j)}} \tag{4}$$

其中，C 是类属性；i 和 j 包括属性子集中的所有属性，式(4)也就是 GQPSO 特征选择算法的适应度函数，很明显其值越大，粒子的适应度越高。

2.3 粒子编码方法

属性选择就是在保证原有数据集分类能力不变的条件下，删除其属性集中的无关属性和冗余属性，以实现特征的优化选择。因此，可以把入侵检测数据集每一个属性定义为粒子的一维离散二进制变量，M 个属性构成粒子的 M 维离散二进制空间。

对于每一个粒子，如果第 i 位为 1，表示第 i 个属性被选中，反之表示该属性没被选中。因此，每一个粒子代表了一个不同的属性子集，也就是一个候选解。比如，粒子 $i=1010100$，那么表明属性 2,4,6,7 被剔除，属性 1,3,5 被选中，因此优化后的属性子集为 {1,3,5}。

2.4 粒子更新

在粒子进行迭代更新时，为了确保各个粒子能有效地进行自身和社会学习，并避免过早地收敛从而陷入局部最优值，所以各个粒子按如下方式进行更新操作：

$$\begin{cases}x_i^{t+1}=p+\beta\cdot|Mbest-x_i^t|\cdot\ln(1/u) & \text{if } k\geqslant 0.5\\ x_i^{t+1}=p-\beta\cdot|Mbest-x_i^t|\cdot\ln(1/u) & \text{if } k<0.5\end{cases} \tag{5}$$

其中，x_i^t 表示粒子 i 在第 t 时刻的位置；β 是收缩扩张因子，它用来控制算法的收敛速度，可在运行中动态调节；μ,k 为 [0,1] 之间产生的随机数。同时，为了保证所有粒子向最优粒子靠拢，将 p 定义如下：

$$p=(c_1\cdot pbest+c_2\cdot gbest)/(c_1+c_2) \tag{6}$$

其中，c_1,c_2 为 [0,1] 之间的随机数；p 表示在全局最优值与局部最优值之间的一个随机值。$Mbest$ 为整个粒子群的中心位置，按下式确定：

$$Mbest=\frac{1}{m}\sum_{i=1}^{m}pbest_i=\left(\frac{1}{m}\sum_{i=1}^{m}pbest_{i1},\frac{1}{m}\sum_{i=1}^{m}pbest_{i2},\mathrm{L},\frac{1}{m}\sum_{i=1}^{m}pbest_{id}\right) \tag{7}$$

其中，m 为种群粒子个数；d 为粒子的维数。

2.5 算法描述

输入 网络数据训练样本，最大迭代次数 T

输出 网络数据的最优特征子集

Step1 随机初始化粒子种群，初始化粒子的个体最好位置 $pbest$、群体的全局极值 $gbest$；

Step2 根据式(4)评价每个粒子的适应度值；

Step3 对每个粒子，将其适应度值与其经历过的最好位置 $pbest$ 进行比较，如果优于 $pbest$，则将其作为当前的最好位置 $pbest$；

Step4 对每个粒子，将其适应度值与群体所经历过的最好位置 $gbest$ 进行比较，如果优于 $gbest$，则将其作为群体最优位置；

Step5 根据式(5)~式(7)更新粒子位置；

Step6 计算当前种群平均适应度值 AF，对适应度值大于 AF 的粒子予以保留，对于小于 AF 的粒子进行变异操作；

Step7 若迭代次数等于 T，转 Step8，否则转 Step2；

Step8 把群体最优位置转化为对应的特征子集。

2.6 算法时间复杂度分析

从 GQPSO 算法的描述可以看出，算法的计算时间主要花费在迭代过程中。设种群规模为 P，粒子维数为 D，迭代次数为 K，每次迭代都要经过适应度计算、个体和群体最优状态选择和粒子位置更新等步骤。其时间复杂度分析如下：

(1)设特征子集数目 $m(1<m<D)$，则计算属性之间归一化互信息时间复杂度为 $O(f(m))=O(m)$，因此计算所有粒子适应度值的时间复杂度为：$O(P\cdot f(m))=O(P\cdot m)$。

(2)粒子群体最优位置选择的时间复杂度为 $O(P)$。

(3)粒子个体最优位置选择的时间复杂度为 $O(P)$。

(4)粒子位置更新的时间复杂度为 $O(m\cdot P^2)$。

(5)粒子变异操作平均时间复杂度为 $O(m\cdot P/2)$。

因此，GQPSO 算法时间复杂度为：

$$T(P)=2\cdot K\cdot O(P)+K\cdot O(m\cdot P^2)+K\cdot O(m\cdot P/2)\approx K\cdot O(m\cdot P^2)$$

3 网络入侵特征选择实验与分析

3.1 实验数据集与参数设置

实验采用 KDDCUP1999[6]作为训练集与测试集的选取来源，数据中攻击类型分为以下 4 种：DoS，Probe，R2L，U2R。在对选取的训练和测试样本进行预处理的基础上，形成实验数据集如表 1 所示。

表 1 实验样本集的组成

攻击类型	训练样本集			测试样本集		
	样本数目	正常样本	攻击样本	样本数目	正常样本	攻击样本
DoS	950	600	350	850	550	300
Probe	800	550	250	650	450	200
R2L	600	450	150	500	350	150
U2R	200	150	50	150	100	50

实验以支持向量机(Support Vector Machine, SVM)[7]作为分类检测算法，在此基础上采用 Matlab7.0 实现本文算法。其中，SVM 的核函数采用 RBF 函数，核参数 g=0.125 和惩罚系数 C=8 采用交叉验证参数寻优方法获取。

3.2 实验结果和比较分析

在实验数据集上，应用本文提出的 GQPSO 特征选择算法，所产生的特征子集如表 2 所示。

表 2 采用 GQPSO 特征选择算法后的特征子集

攻击类型	特征子集
DoS	protocol_type,dst_bytes,service,logged_in,is_hot_login,is_guest_login,srv_dif_host_rate ,srv_serror_rate, srv_rerror_rate, diff_srv_rate
Probe	protocol_type,src_bytes,hot,logged_in,is_hot_login,srv_serror_rate,dst_host_count
R2L	duration,service,dst_bytes,count,.dst_host_count,dst_host_srv_count
U2R	protocol_type,service,num_file_creations,dst_host_count,dst_host_srv_rate

为了验证本算法的有效性，实验分别采用未进行特征选择样本集和特征选择后的样本集对 SVM 训练，并获取检测结果如表 3、表 4 所示。

表 3 未进行特征选择的 SVM 检测结果

攻击类型	属性数目	训练时间/s	检测时间/s	检测率/(%)	误报率/(%)
DoS	41	7.875	4.840	98.11	0.10
Probe	41	1.015	0.610	51.83	0.20
R2L	41	0.656	0.359	76.22	0.00
U2R	41	0.001	0.020	68.28	1.56

表 4 采用 GQPSO 特征选择算法后的 SVM 检测结果

攻击类型	属性数目	训练时间/s	检测时间/s	检测率/(%)	误报率/(%)
DoS	10	0.9010	0.5020	99.98	0.00
Probe	7	0.1411	0.0470	91.77	0.10
R2L	6	0.3280	0.1420	98.26	0.00
U2R	5	0.0001	0.0161	100.00	0.30

由表 3、表 4 可以看出，经 GQPSO 特征选择后，SVM 分类算法在 4 类攻击上的训练和检测时间大为减少，特别是在DoS攻击检测中，SVM训练和检测时间分别减少了88.56%和 89.67%；同时，SVM 算法在 4 类攻击上的检测率有明显提高，而且保持了很低的误报率，尤其在 Probe、R2L、U2R 3 种攻击检测中，SVM 检测率分别提高了 77.06%、28.92%和 46.46%，且误报率保持在 0.20%左右。

实验分别采用本文 GQPSO 算法、QPSO 算法和 PSO 算法对表 1 实验样本集进行特征选择，并分别用特征选择后的样本集对 SVM 训练，得到 SVM 检测结果如图 1~图 5 所示。

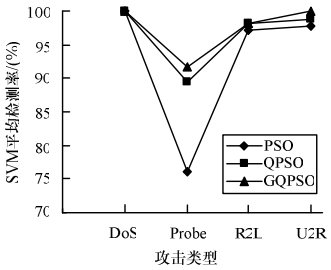

图 1 3 种特征选择算法下 SVM 平均检测率的对比曲线

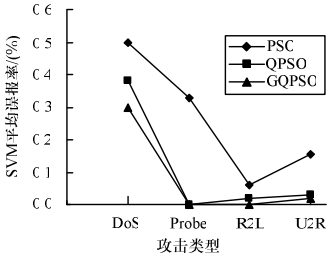

图 2 3 种特征选择算法下 SVM 平均误报率的对比曲线

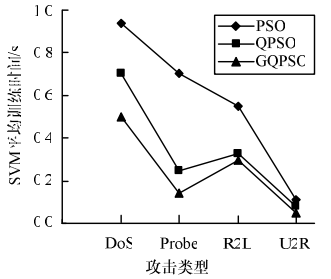

图 3 3 种特征选择算法下 SVM 平均训练时间的对比曲线

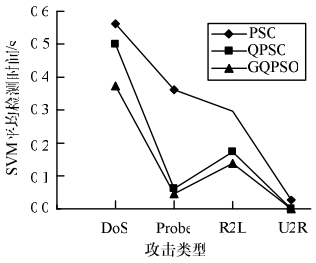

图 4 3 种特征选择算法下 SVM 平均检测时间的对比曲线

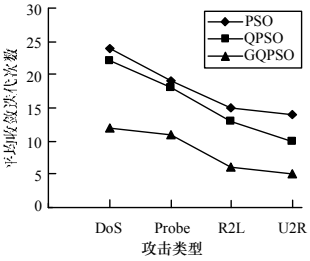

图 5 3 种特征选择算法的平均收敛迭代次数对比曲线

从图 1、图 2 可知，经本文算法进行网络入侵特征选择后，SVM 的检测率、误报率均优于采用 QPSO 和 PSO 特征选择算法，特别是在 Probe 攻击检测中，SVM 平均检测率分别提高 4.28%和 20.75%，平均误报率分别下降 1.25%和 93.33%。由图 3、图 4 可知，经本文算法进行网络入侵特征选择后，SVM 训练时间和检测时间均少于采用 QPSO 和 PSO 特征选择算法，尤其是在 Probe 攻击检测中，SVM 训练时间分别减少 43.33%和 77.92%；SVM 检测时间分别减少 14.28%和 86.84%。由图 5 可知，与 QPSO 和 PSO 特征选择算法比较，本文算法在 4 种攻击检测中，平均收敛迭代次数明显下降，尤其是在 DoS 攻击特征选择中，本文算法的平均收敛迭代次数下降了 13 和 10，表明本文算法有更快的全局收敛速度。

4 结束语

针对高维网络数据中的无关属性和冗余属性致使分类算法的网络入侵检测速度慢，检测率不高的问题，本文提出基于 GQPSO 的网络入侵特征选择方法，该方法涉及一种新的适应度函数，并用其指导 GQPSO 优化算法，以剔除网络数据中的无关属性和冗余属性，实现网络入侵特征的优化选择。实验结果显示，与同类特征选择算法相比，经本文算法进行网络入侵特征选择后，分类算法检测速度更快、检测率更高。

参考文献

[1] Srinoy S. Intrusion Detection Model Based on Particle Swarm Optimization and Support Vector Machine[C]//Proc. of the 2007 IEEE Symposium on Computational Intelligence in Security and Defense Applications. Honolulu, Hawaii, USA: [s. n.], 2007.

[2] 高海华, 杨辉华, 王行愚. 基于 BPSO-SVM 的网络入侵征选择和检测[J]. 计算机工程, 2006, 32(8): 37-39.

[3] 倪 霖, 郑洪英. 基于免疫粒子群算法的特征选择[J]. 计算机应用, 2007, 27(12): 2922-2924.

[4] 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.

[5] 汪世义, 陶 亮, 王华彬. 基于QPSO属性约简在NIDS中的应用研究[J]. 微电子学与计算机, 2010, 27(1): 120-122.

[6] KDD99Cupdataset[EB/OL]. (2010-07-07). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.Html.

[7] Cortes C, Vapnik V. Support Vector Networks[J]. Machine Learning, 1995, 20(3): 273-297.

编辑 陈 文