进化算法及其在入侵检测中的应用*)

郑洪英[1] 廖晓峰[1] 倪 霖[2] 肖 迪[2]
(重庆大学计算机学院 重庆 400030)[1] (重庆大学机械学院 重庆 400030)[2]

摘 要 进化算法模拟自然进化过程,从随机产生的一群个体出发,采用"适者生存"的进化机制,最后收敛到最优解。针对复杂问题,进化算法有很强的搜索能力和最优化性能。而入侵检测问题可以转化成数据的最优分类问题,因此引入模拟退火搜索算法来实现聚类结果的整个优化过程。算法最后使用 KDD Cup 1999 数据集,并在 MATLAB6.5 中进行了仿真实验,检测效果说明了这种方法的可行性和有效性。

关键词 进化算法,入侵检测,最优化,聚类

Evolutionary Algorithm and its Applications in Intrusion Detection

ZHENG Hong-Ying[1] LIAO Xiao-Feng[1] NI Lin[2] XIAO Di[2]
(College of Computer Science and Engineering, Chongqing University, Chongqing 400030)[1]
(College of Mechanic Engineering, Chongqing University, Chongqing 400030)[2]

Abstract Evolutionary algorithm (EA) is an effective algorithm which simulates the natural evolution (survival of the fittest), begins with a population of random individuals, and converges to the fittest individual representing the optimum solutions. EA has powerful search and optimization performance in a complex problem. Intrusion detection is actually extract abnormal data from normal data, so we can transform the problem of intrusion detection into optimization problem. In this paper, the simulated annealing algorithm is used to optimize the clustering results. The experiment with KDD cup 1999 data sets using matlab 6.5 tool shows that the method is feasible and effective.

Keywords Evolutionary algorithm, Intrusion detection, Optimization, Clustering

1 引言

随着计算机技术和通信技术的发展,由入侵而造成的损失以及和计算机相关的犯罪也急剧增加。因此,网络安全成为人们关注的焦点,即如何确保系统按照预期目标正常、稳定地运行。入侵检测系统(IDS)[1]是从计算机或网络中抽取信息,检测来自于系统外部的入侵者的出现和内部人员对系统的误用。

入侵检测实际上是把异常数据从正常数据中抽取出来,从而识别入侵。因此入侵检测问题可以转化成数据的最优分类问题,也就是找到正确分类数据的最优解。进化算法模拟自然进化过程,从随机产生的一群个体出发,采用"适者生存"的进化机制,最后收敛到最优解。针对复杂问题,进化算法有很强的搜索能力和最优化性能。

在入侵检测研究中对网络数据进行分类是基于以下两个基本假设:

① 用户和程序行为是可见的;

② 正常行为与入侵行为本质上是可区分的。

目前,入侵检测主要分为误用检测(misuse detection)和异常检测(anomaly detection)两类方法。由于误用检测不能对未知入侵和稍微变化的入侵数据进行有效的模式匹配,因此造成低检测率和高误报率。未知的和全新入侵行为的检测主要由异常检测来实现,这也是入侵检测研究领域的焦点。

2 进化算法

进化算法在求解问题时是从多个解开始的,然后通过一定的法则进行逐步迭代以产生新的解。这多个解的集合称为一个种群(population,或称为群体等),记为 $p(t)$,这里 t 表示迭代次数。一般地,$p(t)$中元素的个数在进化过程中是不变的,个数称为群体规模,常记为 N。$p(t)$中的元素称为个体(individual)或染色体(chromosome),记为 $x_1(t), x_2(t), \cdots\cdots$,等。在进化中,要选择当前解进行交配以产生新解。进化算法的流程如图 1 所示。

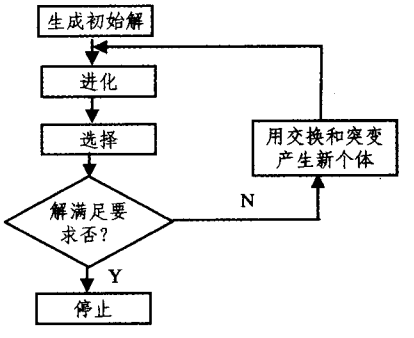

图 1 进化算法工作流程

进化算法常常需要将问题的解进行编码,即通过变换将 X 映射到另一空间 X_g(称为基因空间)。这个变换 $F: X \to X_g$ 要求是可逆的,F^{-1} 称为解码变换。通常,X_g 中的点是字符串(如位串或向量等)的形式,假设 $X_g = B^l = \{0,1\}^l$,即 X_g 是长度为 l 的二进制串全体,则一个长度为 l 的二进制串称为一个染色体。染色体的每一位称为基因(gene),基因的取值称为等位基因(allele),基因所在染色体中的位置称为基因位

*)国家 863 高技术研究发展计划项目、国家科技支撑计划项目(2006BAH02A09)资助。郑洪英 博士研究生,研究方向为信息安全;廖晓峰 博导,主要领域是信息安全、非线性科学;倪 霖 副教授,研究方向是信息系统;肖 迪 博士,主要领域是信息安全。

(locus)。

进化算法尽管是一种搜索寻优的方法，但是它和传统的方法有很大的不同，它不要求所研究的问题是连续、可导的，但是却可以很快地得出所要求的最优解。具有：①有指导搜索、②自适应搜索、③渐进式寻优、④并行式搜索、⑤黑箱式结构、⑥全局最优解、⑦通用性强等特点。

3 进化算法在入侵检测中的应用

3.1 进化算法的引入

在近期入侵检测系统的研究过程中，人们提出了一些新的入侵检测技术，例如免疫系统[2~4]、神经网络[5~7]、基于代理(Agent)的检测等。而目前最常见的入侵检测系统是专家系统[8]，一般由一系列规则组成，这些规则是将专门领域专家的经验和知识进行抽象而形成的。专家系统能够将大量的经验和知识集中到同一系统中，识别当前网络是否符合误用入侵行为的特征。专家系统必须通过不断升级来保证其有效性。也有将数据挖掘引入入侵检测研究，但进行检测时仍存在两方面的困难，即对于未在训练集出现的和全新的入侵行为的检测，则出现较低的检测率和较高的误报率；另一方面，由于需要对大量已标识的历史数据进行训练学习，因此造成了对训练集数据的强烈依赖性。

因此可以利用进化算法的优化能力开展基于进化算法的网络入侵检测研究，以提高检测算法对未知入侵检测的有效性为目标，从检测率和误报率两个重要指标出发进行网络入侵检测。例如，Liu Yongguo 等[12]提出了一种基于遗传聚类的网络入侵检测方法。

3.2 基于进化算法的网络入侵检测模型

入侵检测的核心问题在于对安全审计数据进行分析，分析正常数据和入侵数据的分布特征，以检测其中是否包含入侵或异常行为的迹象，数据分析模块是入侵检测系统的核心。基于异常的入侵检测系统的数据分析模块主要涉及两个问题：(1)如何建立计算机系统或网络的正常行为模型即检测模型；(2)如何以检测模型作为检测入侵的依据，来确定待检行为是否为入侵行为。

本文采用聚类分析和进化算法相结合的思想，首先对数据分布进行分析，通过得到一个近似最优的聚类结果来建立一个较准确的检测模型。这个模型的建立过程实际上是不断对聚类结果使用进化算法进行优化的过程，利用进化算法对聚类准则函数进行优化，从而得到一个近似的最优解，这样也就对数据进行了较准确的划分，把正常数据和入侵数据自动划分到不同的类；接下来再对各个类加标签，是“正常类”、“入侵类”还是“怀疑类”，然后给出各个类的定义，以此来建立检测模型；最后在这个检测模型的基础上进行入侵检测并给出进行入侵检测的方法。该检测模型所用训练数据集易于从实际运行环境中获得，且具有一定的自动化、自我学习能力和较好的检测性能。入侵检测过程分成三个步骤，即训练、测试和评估，这个过程可以使用图 2 来描述。

数据预处理：这个阶段主要任务是从 KDD Cup1999[9] 中构造训练集和测试集，并分别进行规范化处理。

训练阶段：主要任务是使用具体的聚类分析算法对准备好的数据进行处理和分析，提取出能够精确描述系统行为模式的检测模型。

评估阶段：主要是对使用检测模型进行测试与评估的模块。该模块的主要任务是给出使用检测模型进行入侵检测的方法，并对检测结果进行测试与评估，以得到真正符合要求的模型。

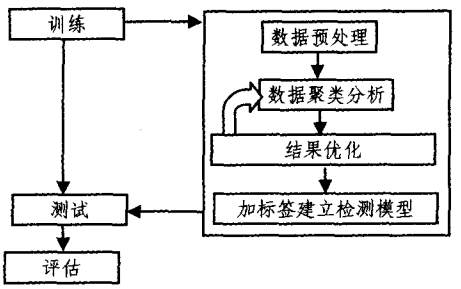

图 2 基于聚类和进化算法的入侵检测模型

4 实验结果

4.1 聚类实现

本文把聚类问题转化成类标识的分配问题，每个类有唯一的一个类标识。然后使用优化算法找到一种最优的分配方案。

已知样本空间 Q 中的 N 个样本 X_i，$i \leqslant N$，对于每一个样本 X_i 需要为它分配一个类标识 c_i，$c_i \in [1 \cdots K]$，因此可以得到一个 N 维的分配向量 $C_i=(c_1, c_2, c_3, \cdots, c_N)$。例如在图 3 中，分配向量 $C_i=(5,3,5,5,2,1,1,3,4,4)$，向量给 10 个样本分配类标识，使得 $c_1=5, c_2=3, c_3=5, c_4=5, c_5=2, c_6=1, c_7=1, c_8=3, c_9=4, c_{10}=4$，从而可以获得 5 个聚类族：属于第一个聚类族的样本分别是 X_6, X_7；第二个聚类族是 X_5；第三个聚类族是 X_2, X_8；第四个聚类族是 X_9, X_{10}；而第五个聚类族是 X_1, X_3, X_4。

$$C_i=(5,\ 3,\ 5,\ 5,\ 2,\ 1,\ 1,\ 3,\ 4,\ 4) \xrightarrow{\text{Assignment}} \begin{cases} CLUSTER_1=(X_6,\ X_7) \\ CLUSTER_2=(X_5) \\ CLUSTER_3=(X_2,\ X_8) \\ CLUSTER_4=(X_9,\ X_{10}) \\ CLUSTER_5=(X_1,\ X_3,\ X_4) \end{cases}$$

图 3 一个分配实例

对于一个分配方案，我们可以使用聚类准则来评价聚类质量。聚类准则选得好，聚类质量就会高。如果聚类质量不满足要求，就要重复执行聚类过程，以便优化聚类结果。在重复优化中，可以改变相似性度量，若有必要还可选用新的准则函数。算法主要使用类内距离和准则函数，如公式(1)。

$$J=\sum_{k=1}^{K}\sum_{\substack{c_i=k \\ c_j=k}} \| x_i-x_j \| \tag{1}$$

因此对于样本空间 Q 中的 N 个样本，算法力求找到一个分配方案 C_i，使得在可能的分配方案中 $C=(C_1, C_2, \cdots, C_{k^n})$，$C_i$ 使得聚类准则函数最小。即

$$Minimum(J(C_i)) \tag{2}$$

$$Subject\ to\ C_i \in C \tag{3}$$

因此引入混沌模拟退火算法就是为了对聚类准则函数进

(下转封四)

(上接第 163 页)

行优化,求出使得聚类准则函数具有最大值的一种类标识分配方案。

4.2 收敛过程

实验过程中的数据集采用 KDD Cup1999 网络数据集,引入模拟退火算法[10,11]实现聚类准则函数的进化过程。模拟退火可以看成是局部搜索的一种变体。在模拟退火中,每当获得当前一个相邻的解,算法根据新的解的目标函数值作出不同的选择。当新的解的目标函数值优于当前解,则选择新的解作为当前解,否则按照一个概率公式决定选择新解作为当前解。图 4 给出了当分类数目为 60、初始温度为 300、结束温度为 1、衰减因子为 0.92、Mapkob 链的长度为 1000 时准则函数的收敛过程。

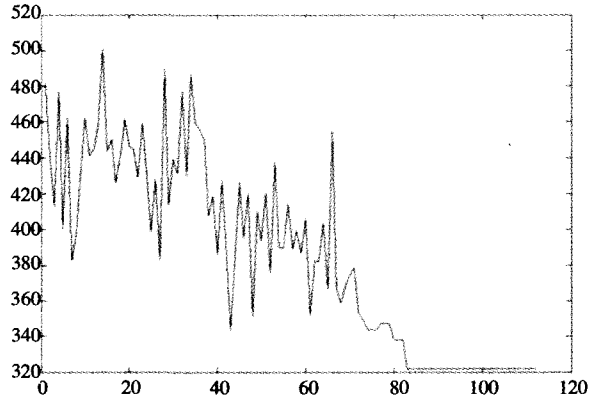

图 4 聚类准则函数的收敛过程

4.3 检测结果

模拟退火算法中冷却进度表不同参数的组合将导致数据聚类结果的差异,从而影响到最终的检测质量。另外,分类个数(label)是影响最终检测质量的重要因素,它的选取主要依据训练集中的数据量多少来确定,过多或过少的分类个数都会导致检测结果的恶化。图 5 给出了 5 个数据集的检测结果。

结束语 进化算法是一种搜索寻优的方法,它不要求所研究的问题是连续、可导的,但是却可以很快地得出所要求的最优解。而入侵检测问题可以转化成数据的最优分类问题,因此,引入模拟退火算法来实现聚类结果的整个优化过程。算法最后使用 KDD Cup 1999 数据集,并在 MATLAB6.5 中进行了仿真实验,检测效果说明了这种方法的可行性和有效性。

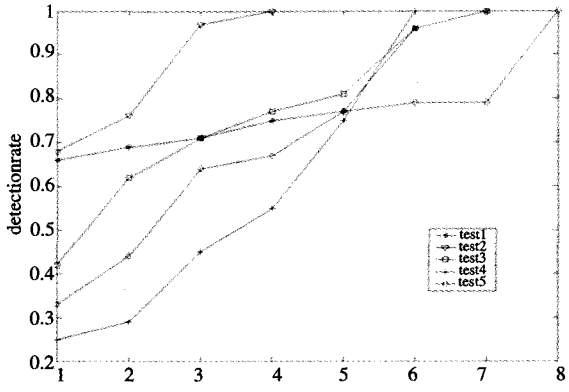

图 5 5 个数据集的检测率

参 考 文 献

1 Axeksson S. Research in intrusion detection: a survey:[Dissertation]. Sweden: Department of Computer Engineering Chalmers University of Technology, 1999

2 Harmer P K, Williams P D, Gunsch G H, et al. An Artificial Immune System Architecture for Computer Security Applications. IEEE Transaction on Evolutionary Computation, 2002,6(3):252～280

3 Hofmeyr S A, Forrest S. Architecture for an artificial immune system. Evolutionary Computation, 1999, 7(1):45～68

4 Dasgupta D. An Immunity-based Technique to Characterize Intrusions in Computer Networks. IEEE Transaction on Evolutionary Computation, 2002,6(3):281～291

5 Lippmann R P, Cunningham R K. Improving intrusion detection performance using keyword selection and neural networks. Computer Networks, 2000, 34: 597～603

6 Markou M, Singh S. Novelty detection: a review-part 2: neural network based approaches. Signal Processing, 2003,83: 2499～2521

7 Joo Daejoon, Hong Taeho, Han Ingoo. The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors. Expert systems with applications, 2003, 25:69～75

8 Sebring M M, Shellhouse E, Hanna M E, et al. Expert systems in intrusion detection: A case study. In:Proceedings of the 11th National Computer Security Conference, Baltimore, Maryland, 1988. 74～81

9 Lincoln Labs. KDD-cup data set. http://kdd.ics.uci.edu/databases/kddcup99.html. 2004-12-2

10 Loukil T, Teghem J, Fortemps P. A multi-objective production scheduling case study solved by simulated annealing. European Journal of Operational Research, 2007,179(3): 709～722

11 El-Bouri A, Azizi N, Zolfaghari S. A comparative study of a new heuristic based on adaptive memory programming and simulated annealing: The case of job shop scheduling. European Journal of Operational Research, 2007,177(3):1894～1910

12 Liu Yongguo, Chen Kefei, Liao Xiaofeng, et al. A genetic clustering method for intrusion detection. Pattern recognition, 2004, 37:927～942

计算机科学

(1974 年 1 月创刊)

第 34 卷第 11 期 (月刊)

2007 年 11 月 25 日出版

国际标准连续出版物号 ISSN 1002－137X

国内统一连续出版物号 CN50—1075/TP

定价:30.00 元 国外定价:5 美元

邮发代号:78—68

发行范围:国内外公开

主管单位:国 家 科 学 技 术 部

主办单位:国家科技部西南信息中心

编辑出版:《计 算 机 科 学》 杂 志 社

重庆市渝北区北部新区洪湖西路 18 号 邮政编码:401121

电话:(023) 63500828 E-mail:jsjkx@swic.ac.cn

网址:www.jsjkx.com

社 长:牟炳林

总 编:彭 丹

主 编:朱宗元

主编助理:徐书令

印 刷 者:重 庆 科 情 印 务 有 限

总发行处:重 庆 市 邮 政 局

订 购 处:全 国 各 地 邮 政 局

国外总发行:中国国际图书贸易总公司(北京 399 信箱)

国外代号:6210—MO