基于免疫原理的入侵检测算法研究①

Research on Intrusion Detection Algorithm Based on Immune System

裴 凯 王卫红 （浙江工业大学 软件学院 浙江杭州 310023）

摘 要： 将生物免疫思想引入入侵检测系统中，提出了一种新的基于免疫的入侵检测模型；分析了网络中数据的特性，引入马氏距离，从而将字符串比较转化为数值计算以此解决了匹配的效率问题。该算法充分利用了网络中的数据包数量，考虑了各部之间的关系。实验结果表明，本文算法可得到很好检测效果。

关键词： 网络安全 入侵检测 免疫系统 免疫因子 马氏距离

1 引言

随着计算机技术的发展，诸如未授权使用、拒绝服务等网络安全威胁也在增加。计算机病毒的数量已经从其 1986 年出现第一个增加到了如今的超过 73000 个。

现在出现了许多灵感来源于生物原理的计算机技术。生物的免疫系统因为拥有处理海量信息的能力而成为人们很感兴趣的一个课题[1,2]。尤其是免疫系统能够以并行的、分布式的模式来进行许多复杂的计算。就像神经系统一样，免疫系统能够学习新的信息，回顾以前学习过的信息，并且能在分布式的环境下进行模式识别[3]。

计算机网络的安全可以和生物的免疫系统作类比。在网络中，由于计算机组件故障或者入侵行为（包括内部的和外部的）而有可能出现威胁和危险（包括影响私有性，完整性以及可用性）。采用免疫的原理来解决计算机的安全问题[4-6]，这个思想在 1994 年就提出了。New Mexico 大学的 Forrest 和他的小组已经为此进行了长时间的研究并且建立一个人工免疫模型[7,8]。Dasgupta 和 Fabio 也提出了一些方法来解决这些问题，比如正选择方法和负选择方法。

在此类方法中，系统模拟免疫系统识别自身的分子和细胞（称为“self”）以及外来的分子和细胞（称为“nonself”）[9-11]，其主要缺点就是检测子的产生过程是随机的，所以新形成的检测子很有可能错误地识别自身细胞和外来细胞。为了尽可能的减少这种错误，检测子必须经过一个成熟过程，此过程的计算量比较大，效率低下。为了解决此类问题，许多研究者提出了种种改进办法，但效果始终不太理想。

对此，我们提出一种改进的免疫模型。本文的主要工作是通过对数据的预处理，使之形成向量，引入马氏距离算法，从而将串比较转化为数值计算来提高效率。实验的结果表明，此方法具有很好的检测效果。

2 基于马氏距离的入侵检测系统

在统计学上，马氏距离是由印度人 P. C. Mahalanobis 于 1936 年提出的[12]。它是基于不同变量之间的相关性来进行识别和分析的。在判断一个未知样本与已知样本之间的相似性方面，马氏距离是一个有效的方法。马氏距离与我们熟知的欧氏距离不同，它考虑了数据集之间的相关性和分散性信息。马氏距离广泛应用于聚类分析和其他分类技术中。马氏距离也经常用于检测外来点集，尤其是在线性退化模型中。

马氏距离已经用于人脸识别及字体辨认方面[13]。马氏距离区别于其他距离判别方法的最重要的一个参数就是协方差矩阵[14]，它能更好“反应”系统内部之间的“特性”，这对入侵检测中，正常行为与入侵行为的区分是相当有利的。因此我们采用马氏距离来反应这种“相关性”。

2.1 免疫模型

在我们的实验系统中，模拟免疫系统分为训练区

① 基金项目：国家基金（60473024）

和检测区。在训练区中，根据已有的数据，通过计算，得出能更好反应正常与异常的不同特性的检测子；在检测区中，样本数据与经过训练的检测子进行比较，得出判断结果。

通常的基于免疫原理的入侵检测模型都是模拟生物免疫，生成的检测子是不能与自身细胞结构相匹配的。在我们的实验系统中，为了与马氏距离相结合，作了一定的改进。检测子分为两类，一类是由正常行为来训练，一类是由入侵行为来训练。最后用于测试的样本要分别与这两类检测子进行匹配，比较与两者的特性谁更接近，才能做出最后的判断。这样做的好处在于，避免自己定义阈值来判断相似的程度，只是把训练检测子的计算量转移到了测试匹配过程中。

免疫系统可以表示成一个两元组 $\Omega=(I,D)$。其中，I 是检测子的集合，表示为 $I_1,\cdots,I_n$。D 是检测过程，表示为 $D_1,\cdots,D_n$。根据检测的结果 D_i 来来判断样本是否属于正常。

任何入侵检测的方法都不可避免的会产生错误。这些错误可以大致分为两类：

(1) False positive (False alarm)：系统报警，此时并没有产生异常。过高的此类错误会使得入侵检测效率低下。

(2) False negative：系统把一个异常判断为了正常。此类错误的产生将会对系统产生非常危险的后果，因为入侵被视作正常。

一个好的入侵检测方法应该产生较少的 false positive，同时也尽量避免产生 false negative。

2.2 马氏距离训练过程

我们用真实的入侵数据来进行测试。数据来自于 Massachusetts 研究所的 Lincoln 实验室[15]。这些数据是从一个测试网络中取得的，既包含了正常的信息也包含了异常的信息。获取这些数据的目的是为了检验入侵检测系统。这些数据中有一周是完全正常的数据（不含有任何攻击）。这就为训练我们的系统提供了足够的样本。测试数据是由网络通信数据（tcpdump，inside and outside network traffic），审计数据（bsm）和文件系统数据组成的。在我们的实验中，只利用了 outside tcpdump network 这个数据。

设是来自均值向量为 μ、协方差矩阵为 Σ 的总体 G 的样品，则 x 至总体 G 的马氏距离是

$$d(x,G)=[(x-\mu)^{T}\Sigma^{-1}(x-\mu)]^{+}$$

网络中的通信记录通常是以字符串的形式记录下来的，如：

11:58:47.873028 202.102.245.40.netbios-ns > 202.102.245.127.netbios-ns: udp 50

此行数据是从网络界面上流过的数据包里面截取出来的，可以看出此网络行为发生的时间，来源主机与目的主机，访问类型等。通常的协议分析便是针对这种信息包头。但网络中的流量很大，单靠人工分析显得效率及其低下。因此，我们想到把字符串格式转化为数值格式便于计算。在 Dasgupta 等人的文章中，把此种格式进行编码，用二进制数来表示，然后用遗传算法进行训练和改进。但这种编码方式因为是根据个人指定的规则，所以具有很强的随意性。在我们的方法中，用 tcpstat 这个工具来直接获取通信统计。tcpstat 具有强大的功能，灵活的截取策略，可以截取网络中的数据流量，这正是我们所需要的。

为了采用马氏距离，我们必须把数据构成向量形式，我们只利用通信统计中的三组数据来构成向量：用 b 来表示每秒钟的字节数（number of bytes per second）、用 p 来表示每秒钟的数据包数量（number of packets per second）以及用 c 来表示每秒钟的 ICMP 包数量（number of ICMP packets per second），则样本可表述为 $x(b,c,p)^{T}$。这些数据都可以通过 tcpstat 工具直接获得。之所以选择这三类数据作参数而不是其他，是因为它们都是直接反映网络中的流量变化。正常情况下，网络中的流量应该是平稳变化的，一旦出现异常，流量会突然增大或减小。通过观察流量的变化情况，可以初步判断是否出现异常。当然，正常行为的平稳变化可能只会影响数据包的数量而并没有影响 ICMP 的行为，所以我们选择三类参数同时训练，而不是两个或一个参数，也是考虑到尽量避免单一参数带来的误差。为了从实验上证明此理论，我们构造另一个一维样本 $y(b)^{T}$。

1999 年的数据中，有一周是完全正常的数据，不包含任何入侵行为，我们利用其来训练“self”（G_1）；在用于训练“nonself”（G_2）的数据上，为了要尽可能的反应异常行为带来的数据流量上的变化，我们预先挑出各类入侵行为单独训练。在本文的实验中，选取的是 1998 年最初的 20 个攻击数据和 1999 年最初的 100 个

攻击数据。剩下的数据中既包含正常数据，也包含入侵行为，因此用来进行测试。这些攻击的描述见[13]。

在实际问题中，由于数据的局限性，类似协方差矩阵、均值向量此类数据通常都是未知的，所具有的数据资料只是来自两个 p 维总体的样本观测值，称为训练样本，设 $x_1^{(1)}, x_2^{(1)}, \cdots, x_{n1}^{(1)}$ 是来自总体 G_1 的容量为 n_1 训练样本；$x_1^{(2)}, x_2^{(2)}, \cdots, x_{n2}^{(2)}$ 是来自总体 G_2 的容量为 n_2 训练样本，这时可以用训练样本估计均值向量 μ_1, μ_2 及协方差矩阵 Σ。μ_1, μ_2 的估计是各自训练样本的均值向量，即

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n_1} x_i^{(1)} = \bar{x}^{(1)}, \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n_2} x_i^{(2)} = \bar{x}^{(2)}$$

两个训练样本的协方差矩阵各为

$$S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} (x_i^{(1)} - \bar{x}^{(1)})(x_i^{(1)} - \bar{x}^{(1)})^T$$

$$S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} (x_i^{(2)} - \bar{x}^{(2)})(x_i^{(2)} - \bar{x}^{(2)})^T$$

当 $\Sigma_1 = \Sigma_2 = \Sigma$ 时，Σ 的一个联合估计是

$$S = \hat{\Sigma} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

这样，定义线性判别函数 $W_1(x)$，$W_2(x)$ 及 $W(x)$ 的估计各为

$$\begin{cases} \hat{W}_1(x) = \hat{a}_1^T x + \hat{b}_1, \text{其中 } \hat{a}_1 = S^{-1}\bar{x}^{(1)}, \hat{b}_1 = -\frac{1}{2} \\ (\bar{x}^{(1)})^T S^{-1}\bar{x}^{(1)} \\ \hat{W}_2(x) = \hat{a}_2^T x + \hat{b}_2, \text{其中 } \hat{a}_2 = S^{-1}\bar{x}^{(2)}, \hat{b}_2 = -\frac{1}{2} \\ (\bar{x}^{(2)})^T S^{-1}\bar{x}^{(2)} \\ \hat{W}(x) = \hat{a}^T(x - \bar{x}), \text{其中 } \hat{a} = S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}), \bar{x} = \frac{1}{2} \\ (\bar{x}^{(1)} + \bar{x}^{(2)}) \end{cases}$$

这样，我们就可以做出判断：

$$\begin{cases} x \in G_1, \text{若 } \hat{W}_1(x) \geq \hat{W}_2(x) \\ x \in G_2, \text{若 } \hat{W}_1(x) < \hat{W}_2(x) \end{cases}$$

$\hat{W}_1(x) \geq \hat{W}_2(x)$，说明待检测数据的“特性”比较接近正常的 G_1，则可判为自我（self）；相反，$\hat{W}_1(x) < \hat{W}_2(x)$，则说明待检测数据更接近异常的 G_2，则可判为非我（nonself）。

3 实验与结果分析

我们基于免疫思想，利用马氏距离来处理这些数据。实验与免疫系统的映射关系如表 1 所示：

表 1　免疫系统与实验系统映射关系

免疫系统	self	nonself	淋巴细胞	外来细胞
实验系统	正常数据	异常数据	用于训练的异常数据	待检测的数据

实际数据中，难免会出现个别样本的“特性”与总体“特性”相差较大的情况，例如，本属于正常样本的数据，其协方差矩阵更接近于异常样本。为尽量减小此类样本对总体的影响，我们可以采取回代法来剔除此类样本数据。设 G_1, G_2 为两个总体，$x_1^{(1)}, x_2^{(1)}, \cdots, x_{n1}^{(1)}$ 与 $x_1^{(2)}, x_2^{(2)}, \cdots, x_{n2}^{(2)}$，是分别来自 G_1 和 G_2 的训练样本，其容量分别是 n_1 和 n_2。以全体训练样本作为 $n_1 + n_2$ 个新样品，逐个代入已建立的判别准则中判别其归属，这个过程称为回判。如果将本属于 G_1 的样品 $x_i^{(1)}$ 误判为 G_2，则将 $x_i^{(1)}$ 从 G_1 中剔除，剩下的数据重新进行训练。如果将本属于 G_2 的样品 $x_i^{(2)}$ 误判为 G_1，则将 $x_i^{(2)}$ 从 G_2 中剔除，剩下的数据也重新进行训练。由此得到的两个样本集更能准确反应各部的特征。

我们把马氏距离计算的结果与其他基于免疫思想的入侵检测方法得出的结果进行比较，具体效果表 2：

表 2　实验结果对比

检测技术	检测结果（产生 1% False alarm）
正选择方法	96.4%
负选择方法	87.5%
马氏距离方法（一维）	57.9%
马氏距离方法（三维）	96.8%

从表中，我们可以看出：

（1）只采用一个参数来训练系统，显然不能更好的反应正常与异常行为各自的特性，参数自身会在一个范围内变化，这种变化会对总体产生一定的影响，因此不适合单独训练。

（2）维数越高，效果越好。三维参数所得到的检测效果显然要好于一维。

(3) 三维参数构成的向量,彼此抵消掉单一参数自身的变化对总体的影响,更能从整体上反应正常和异常行为的特征,因此得到的检测效果要比只采用一维参数要好很多。

(4) 采用三维的马氏距离方法的检测效果要好于其他两种基于免疫原理的检测方法。

4 结论

在本文中,我们尝试利用一种新的方法——马氏距离来分析网络中的数据包数量,以此来检测异常行为。我们选择了真实的数据来测试我们的方法。实验结果表明,马氏距离能够有效的应用于网络安全领域。实验结果还证明了马氏距离在分析网络中的数据包数量方面具有很大的潜力。马氏距离公式中的协方差矩阵,充分考虑了总体分布的相关性和分散性信息,能够更好的反应总体之间的特征,这对于区别正常和异常是相当重要的;同时,在数据选择上,我们利用 tcpstat 这个工具来直接获取与网络流量相关的参数,可直接与马氏距离相结合。马氏距离的样本采用三维向量也是为了更好反应正常与异常行为的特性,避免了单一类型数据所可能造成的偏差;其次,回代法更减小了个别数据对总体的影响。

马氏距离最大的特点就是在计算过程中时刻考虑系统各部之间的相关性信息,因此更能准确反应具有不同性质的数据之间的关系。

参考文献

1 Dasgupta, D. Immunity - based intrusion detection system: A general framework. In: Proc. 22nd Nat. Information Systems Security Conf., Oct. 1999. 147 - 160.

2 S. A. Hofmeyr and S. Forrest. Architecture for an artificial immune system. Evol. Comput, 2000,8(4): 443 - 473.

3 Dipankar D., Fabio G.. An immunity - based technique to characterize intrusion in computer networks. IEEE transaction on evolutionary computation, 2002, 6: 281 - 291.

4 D. Dasgupta. Artificial Immune Systems and Their Applications. New York: Springer - Verlag, 1999.

5 Mobile security agents for network traffic analysis. Proceedings of DARPA Information Survivability Conference and Exposition II (DISCEX - II). Los Alamitos, CA: IEEE Comput. Soc. Press, 1999.

6 J. O. Kephart. A biologically inspired immune system for computers. Proceedings of Artificial Life IV, R. Brooks and P. Maes, Eds. Cambridge, MA: MIT Press, 1994. 130 - 139.

7 J. Kim, P. Bentley. The Human Immune System and Network Intrusion Detection. EUFIT ' 99, September, 1999.

8 Novelty detection in time series data using ideas from immunology. Proc. Int. Conf. Intelligent Systems. 1996. 87 - 92.

9 P. Dhaeseleer, S. Forrest, and P. Helman. An immunological approach to change detection: Algorithms. Proceedings of the 1996 IEEE Symposium on Computer Security and Privacy. Los Alamitos, CA: IEEE Comput. Soc. Press, 1996. 110 - 119.

10 A. Somayaji, S. Hofmeyr, and S. Forrest. Principles of a computer immune system. Proc. 2nd New Security Paradigms Workshop, Sept. 1997. 75 - 82.

11 Leandro N de Castro, Fernando J Von Zben.. Immune and neural network models: theoretical and empirical comparisons. International Journal of Computational Intelligence and Applications, 2001, 1(3): 239 257.

12 Mahalanobis, P. C.. On the Generalized Distance in Statistics. Proceedings of the National Institute of Science, Calctta, 12. 49 - 55.

13 Liu - Hailong, Ding - Xiaoqing. Handwritten Chinese character recognition based on mirror image learning and the compound Mahalanobis function. Journal of Tsinghua University. 2006,46(7):1239 - 1242.

14 梅长林、范金城, 数据分析方法, 北京: 高等教育出版社, 2006:142 - 153.

15 DARPA Intrusion Detection Evaluation (1999). http://www.ll.mit.edu/IST/ideval/index.html.