入侵检测系统中的相反性综合降维模型*

张常有[1,2]，曹元大[2]，王玉梅[1]，于 炯[2]
（1. 石家庄铁道学院计算机与信息工程分院，河北 石家庄 050043；
2. 北京理工大学计算机科学技术学院//智能信息技术北京市重点实验室，北京 100081）

摘 要：为了提高入侵检测系统的性能，提出了一种综合降维方法。首先，借用类比推理方法，为两个多维向量建立相似距离算法。然后，基于人工免疫系统和遗传算法设计了一种对正常行为样本集合和异常行为样本集合的优化算法。最后，对采集到的网络行为样本，分别计算与优化的两个行为样本集合的相似度。把这两个相似度作为纵坐标和横坐标，行为样本被映射成二维坐标平面上的点。系统根据点的位置，判定行为是否异常。

关键词：入侵检测；综合；降维；相似度；人工免疫
中图分类号：TP393 **文献标识码**：A **文章编号**：0529－6579（2009）01－0133－05

A Synthetic Dimension Reduction in Intrusion Detection System

ZHANG Changyou[1,2], *CAO Yuanda*[2], *WANG Yumei*[1], *YU Jiong*[2]
（1. School of Computer & Information, Shijiazhuang Railway Institute, Shijiazhuang 050043, China;
2. Beijing Laboratory of Intelligent Information Technology, School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China）

Abstract: In order to improve the performance of IDS (Intrusion Detection System), a synthetic dimension reduction method is proposed in this paper. First of all, a similarity distance algrithm between two vectors based on analogy resoning is difined. Then, an optimization method based on Artificial Immune System (AIS) and Genetic Algorithm (GA) is used to meliorate the normal-behavior-set and abnormal-behavior-set. Finaly, a new behavior sample is sniffered from network. The distances between this new behavior sample and each of the two meliorated sets are calculated. Using these two distances as ordinate and abscissa, this new behavior sample is mapped into a point in a two-dimensional coordinates plane from a multi-dimensional vector space. According to the location of this point, a behavior can be determined whether it is an intrusion or not.

Key words: intrusion detection system; synthetic; dimension reduction; artificial immune

入侵检测系统（IDS，Intrusion Detection System）的目标是通过收集和分析系统信息，进而监控、探测、标识对网络和计算机系统的有害行为和有害企图。这样，IDS能辨别系统的状态是“正常”，还是“异常”[1]。所以，一个IDS被定义为警戒系统。它自动探测主机或网络中的恶意活动[2]。当系统发现对主机或系统的有害行为时，就产生一个警戒信号向系统中的安全设备报警。入侵检测系统分为两类：异常检测和误用检测[3]。对于一个入侵检测系统，正确性和实时性是两个重要因素。当前网络发展的高速化、复杂化等特性对入侵检测系统的数据处理性能提出了新的挑战。因为当网络速度超过了数据处理速度时，入侵检测分析的速度也必须相应加快，需要改进传统的分析方法。

* 收稿日期：2008－09－18
基金项目：国家自然科学基金资助项目（60563002）；北京市自然科学基金资助项目（4082027）
作者简介：张常有（1970年生），男，副教授；E-mail：zhangchangyou@tsinghua.org.cn

解决这一问题有两种基本思路：①提高入侵检测系统的处理能力，包括数据处理的能力和数据采集能力。②采用新的算法或预处理，降低数据处理的难度。

依照第2种思路，针对网络行为模式的正常样本集合和异常样本集合，降低网络行为向量的维度，从而提高数据处理效率。流形学习（Manifold learning）是一种通过从高维数据中发现低维结构的方法，来简化高维数据。算法目标是将一套给定的高维数据点映射到替代的低维空间[4]。Animesh Patcha[5]提出了一个称为SCAN（Stochastic Clustering Algorithm for Network Anomaly Detection）的异常检测方案。该算法有能力高精度检测入侵行为，甚至使用不完整的审计数据。面向网络环境，很多研究者提出了一些新的入侵检测方法[6-10]。

此外，考虑到训练数据的局限性，用遗传算法和免疫算法相结合，对正常行为样本集合和异常行为样本集合作优化处理。对新采集的网络行为数据，分别计算其到正常行为样本集合和异常样本行为集合的距离，并视为纵、横坐标。这样，行为样本被映射为二维空间的点。依据点的位置，系统判断该行为的入侵概率。降维处理有效提高了入侵检测的实时处理效率。

1 网络行为的相似距离

1.1 行为建模

网络行为的相关度较高的属性主要有：服务类型（srvType），源地址（srcIP），源端口（srvPort），目的地址（dstIP），目的端口（dstPort），时延（dur），源端发送字节数（srcBytes），目的端发送字节数（dstBytes），状态（flag）等。因此，每一个网络行为向量可用如下9维（或多于9维）的向量表示：

$$X = [\text{srvType}, \text{srcIP}, \text{srvPort}, \text{dstIP}, \text{dstPort}, \text{dur}, \text{srcBytes}, \text{dstBytes}, \text{flag}]^T$$

注意到，从数据类型上看，向量X的分量有两类：①字符型。其匹配计算就是严格的相等与否。这类分量适合于上文所述的类比相似度算法。本模型中的字符型分量有服务类型（srvType）、源地址（srcIP）、源端口（srvPort）、目的地址（dstIP）、目的端口（dstPort）、状态（flag）等。②数值型。这类数据的取值是一个能用大小度量的数。他们之间的差别能够用差额来度量。对于这类分量，如直接使用式（1）计算，结果不理想。本模型中这类数据有时延（dur），源端发送字节数（srcBytes），目的端发送字节数（dstBytes）等。

对于数值型分量，必须预先处理，使其适合相反性综合距离模型。具体的离散化方法，可参考文献[11]。离散化以后的数值型分量转化为字符型分量，向量X可整体用于相似距离计算。

1.2 综合距离模型

本文中，距离用向量之间的相似度来表示。相似度算法采用类比推理的相反性综合模型。该模型同时考虑了相同分量和相异分量对相似度结果的不同贡献。行为向量$X = [x_1, x_2, \cdots, x_n]$与两个行为样本集合之间的相似距离作为入侵检测的基础。两个网络行为向量之间的相似度代表了他们的差异程度。为了方便阐述，我们给出如下定义，

（1）定义1 （行为向量之间的相似度）：设X与Y表示任意两个行为向量，它们之间的相似度按式（1）计算。

$$\text{Sim}(X,Y) = \frac{f(X \cap Y)}{f(X \cap Y) + \alpha \cdot f(X - Y)}, (\alpha \geqslant 0) \tag{1}$$

式（1）中，$f(X \cap Y)$表示两者之间的相同分量对相似度的贡献，$f(X - Y)$表示两者之间的相异分量对相似度的贡献，$\alpha \geqslant 0$表示相异分量的贡献系数，其值不小于0。明显，$\text{Sim}(X,Y)$是一个0到1之间的数。

（2）定义2 （行为向量与行为集合之间的相似度）：设有行为集合A，则X与A之间的相似度为

$$\text{Sim}(X,A) = \max\{\text{Sim}(X,A_j), A_j \in A, j = 0,1,\cdots,m\} \tag{2}$$

式（2）中，$\text{Sim}(X,A_j)$为行为X与集合A中的元素A_j之间的相似度。最终取最大相似度作为相似结果。

1.3 基于人工免疫的样本集合优化

人工免疫系统模仿自然免疫系统，提供了一种解决潜在问题的神奇途径。免疫网络的数学框架由Jerne在20世纪70年代提出。随后的研究者[12-13]随又进一步从不同的侧面提出了新的AIS理论，完善了其模型、算法和应用。

考虑到训练数据集可能存在的片面性，采用人工免疫方法与遗传算法相结合，优化异常行为样本集合AI_0，优化过程如图1。

图1主要阐明了异常行为样本库AI的生成优化过程。首先，采用数据挖掘方法生成初始集合AI_0，可以根据经验知识加以补充。然后用遗传算子对它们进行变异和增殖，生成一个更大的候选样本集合AI'_0。对个体进行亲和度测定，计算与初始

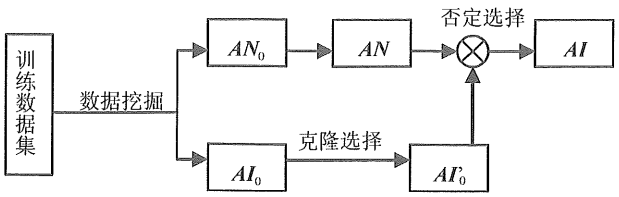

图 1　行为样本集合的产生和优化过程

Fig. 1　The processes of sample set creating and optimizing

AI_0 的相似度，筛选出优秀样本；再进行否定选择，即删除其中与 AN 中相等（或非常相近）的元素。最后产生优化过的异常行为样本集合（AI）。优化过程分为两步：

（1）克隆选择。

克隆选择算法的目的是扩大异常样本量，或者优化抗体在该样本空间的分布特性。这些分布特性包括样本的密度、样本分布的均匀度等。本节采用的克隆选择算法以 AI_0 为原始参数，采用多点交叉，随机变异的方法，扩大样本空间，提高这些样本在该空间分布的均匀度。扩大异常行为样本空间的大小和优化样本在该空间分布的均匀度有利于降低 IDS 的漏报率。

（2）否定选择。

否定选择的目的是保护自体细胞不受到误损。也就是说，AI 中不能存在与 AN 中相同或相近的行为向量。否定选择的算法与上节中的克隆选择类似，要分别计算 AI_0'中元素与集合 AN 的相似度，排除其相似度为 1 和非常接近 1 的向量，避免误报。

正常行为样本集合 AN 采用类似的步骤优化处理。

2　综合距离模型

2.1　数学模型

本文将网络行为抽象为一个 n 维向量，如 $X = [x_1, x_2, \cdots, x_n]$。其中 x_i 为该向量的一个分量，表示行为的一个侧面。这个 n 维向量称为行为空间的一个点。全部网络行为集合构成了行为曲面。具有不同属性的行为集合的全部，将构成不同的曲面。在入侵检测系统中，我们关心异常行为集合和正常行为集合。为了画图方便，不失一般性，设正常行为集合和异常行为集合分别在三维空间构成“异常平面”和“正常平面”，如图 2 示。其中，P 和 Q 分别为两个行为向量所代表的空间点。

图 2（a）中，$|AC|$ 为 P 点到“异常平面”的距离；$|BD|$ 为 Q 点到“异常平面”的距离；

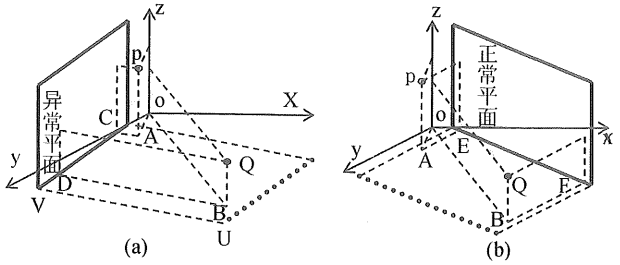

图 2　行为向量的简单距离模型

Fig. 2　The simple distance model of behavior vector

$|UV|$ 为阈值。

$\because |AC| < |UV|$，且 $|BD| < |UV|$

$\therefore P$、Q 均为异常

设 $fp(x)$ 为向量 x 的异常概率函数，则有 $fp(P) = |UV| - |AC|$ 和 $fp(Q) = |UV| - |BD|$ 分别代表点 P 和 Q 的异常概率。

又$\because |AC| < |BD|$

$\therefore fp(P) > fp(Q)$

即，P 点异常概率大于 Q 点异常概率。

再看图 2（b），$|AE|$ 为 P 点到“正常平面”的距离；$|BF|$ 为 Q 点到“正常平面”的距离；有，$fp(P) = |AE|$，$fp(Q) = |BF|$。

又$\because |AE| < |BF|$

$\therefore fp(P) < fp(Q)$。

即，P 点异常概率小于 Q 点异常概率。

两个图中得到了相矛盾的结论。为了达到判断结果的一致性，令

$$fp(P) = \frac{|AE|}{|AE| + |AC|}$$

$$fp(Q) = \frac{|BF|}{|BF| + |BD|}$$

2.2　综合距离模型

综合考虑“正常平面”和“异常平面”的距离，如图 3 所示。

2.3　综合降维模型

根据分析，定义综合降维模型如下。

如果设 IDS 系统的正常行为样本集合为 AN，则把行为 X 与 AN 的相似度 S_{AN}^{X} 称为 X 行为的正常距离。如式（3）计算，

$$S_{AN}^{X} = \mathrm{Sim}(X, AN) = \max\{\mathrm{Sim}(X, AN_j),\ AN_j \in AN,\ j = 0, 1, \cdots, m\} \quad (3)$$

设 IDS 系统的异常行为样本集合为 AI，则把行为 X 与 AI 的相似度 S_{AN}^{X} 称为 X 行为的异常度。如式（4）计算，

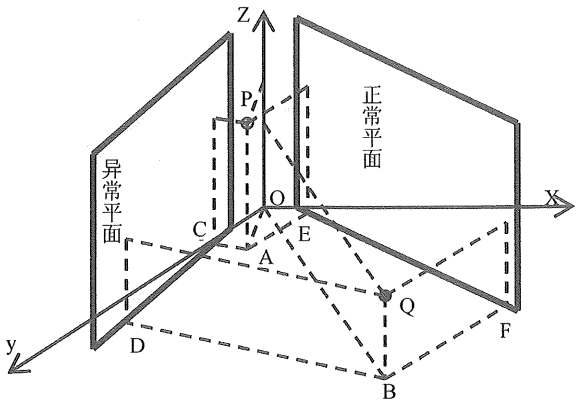

图3 行为向量的综合距离模型

Fig. 3 The synthetic distance model of behavior

$$S_{AN}^{X} = \mathrm{Sim}(X,AI) = \max\{\mathrm{Sim}(X,AI_j), AI_j \in AI, j = 0,1,\cdots,m\} \quad (4)$$

综合考虑行为 X 的正常度和异常度，定义其入侵概率为 $P(X,AN,AI)$。入侵概率的值由式（5）计算。

$$P(X,AN,AI) = \frac{S_{AI}^{X}}{S_{AI}^{X} + \beta \cdot S_{AN}^{X}}, (\beta \geqslant 0) \quad (5)$$

式（5）中，β 是 X 的正常度对入侵概率的贡献系数。β 是一个不小于0的值。

3 行为样本向量降维方法

3.1 行为样本到二维平面的映射

对任意网络行为 X，根据式（3）和式（4）分别能够得到其正常度 S_{AN}^{X} 和异常度 S_{AN}^{X}。若令 $u = S_{AN}^{X}$，$v = S_{AN}^{X}$，则对于本文关心的问题，X 可以映射维一个二元组（u，v）。把 u 视为纵坐标，把 v 视为横坐标，则（u，v）表示为二维平面上的一个点。

考虑到 $u \in [0,1]$，$v \in [0,1]$，网络行为 X 将被映射到坐标系中（0，0）到（1，1）的区域中的一个点。如图4所示。其中，（u_1，v_1）和（u_2，v_2）分别表示行为向量 X_1 和 X_2 在平面上映射得到的两个点。

3.2 行为检测方法

判定行为的入侵性是IDS的根本任务。为了确定一个行为 X 是否为异常行为，需要定义一个阈函数 $u = f(v)$，其对应的曲线在 $v \in [0,1]$ 时，落在（0，0）-（1，1）区域内，如图4中粗实线所示。理想情况下，该域函数曲线把整个空间分成两个区域 D_1 和 D_2。直观上看，D_1 在曲线的左上方，D_2 在曲线的右下方。点（u_1，v_1）落在区域 D_1，（u_2，v_2）落在区域 D_2。

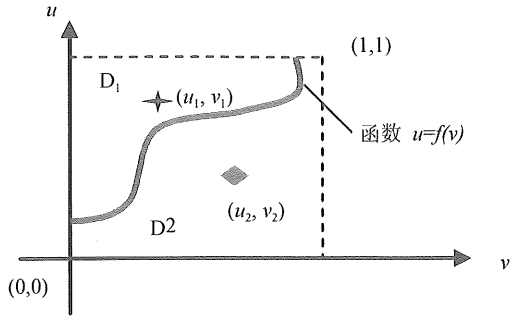

图4 行为向量到平面点的映射模型

Fig. 4 The mapping model from a behavior to a point

设 u_x 表示行为 X 的 u 值，v_x 表示行为 X 的 v 值，这时，

$$u_x < f(v_x)$$

表示点（u_x，v_x）落在 D_2 区，判定 X 为入侵行为。同样，判定 X_2 为异常行为。当系统发现异常，则按照既定的策略报警。

4 结 语

本文针对网络入侵检测系统面临的海量审计数据处理问题，根据关联规则挖掘结果，得到网络行为模式。对网络行为模型，计算其正常性和异常性，映射到平面上的点。从多维降到两维问题，综合考虑两维上的投影，得到入侵与否的一致性评判结果。这种方法能适应并行处理，有利于提高高速分布式网络中的入侵检测的效率。

参考文献：

[1] FORREST S, PERELSON A S, ALLEN L, et al. Self-Nonself discrimination in a computer[C]// Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy. Los Alamitos: IEEE Computer Society Press, 1994: 202 - 212.

[2] KEMMERER R A, VIGNA G. Intrusion detection: a brief history and overview [J]. Computer, 2002, 35 (4): 27 - 30.

[3] 蒋建春，马恒太，任党恩，等. 网络安全入侵检测：研究综述[J]. 软件学报，2000，11(11):1460 - 1407. JIANG Jianchun, MA Hengtai, REN Dangen, et al. A survey of intrusion detection research on network security [J]. Journal of Software, 2000, 11(11):1460 - 1466.

[4] SEUNG H S, LEE D D. The manifold ways of perception [J]. Sience,2000, 22:2268 - 2269.

[5] PATCHA P, PARK J. Network anomaly detection with incomplete audit data [J]. Computer Networks, 2007, 51 (5): 3935 - 3955.

（下转第137页）

poral logic: Mathematical foundations and computational aspects [M]. Oxford :Oxford University Press, 1994.

[3] LAMBEK J. The mathematics of sentence structure [J]. Amer Math Monthly , 1958,65:154 - 170.

[4] BARENDEGT H P. The lambda calculus its syntax and semantics [J]. Studies in Logic and The foundations of Mathematics, 2004,103(6).

[5] http://www. cosoft. sysu. edu. cn/TempDB/index. htm, [2008].

[6] CURRY H B. Some logic aspects of grammatical structure [C]// JAKOBSON R. Stucture of language and its mathematical aspects. AMS, Providence, RI.

[7] HOWARD W. The formulae-as-types notion of construction [M]. Manuscript, Published in Seldin and Hindley , 1980.

[8] RAO J H, KÜNGAS P. Logic-based web services composition from service description to process model [C]//In 2nd Intl. Conference on Web Services,2004.

[9] UUSTALU T, VENE V, The essence of dataflow programming [C]// In 3rd Asian Symp. on Programming Languages and Systems, APLAS'05, Number 3780 in Lect. Notes Comp Sci, 2005.

[10] HAJEK P. Metamathematics of fuzzy logic [C]//Trends in Logic 4. Kluwer, 1998.

（上接第136页）

[6] FUGATE M, GATTIKER J R. Anomaly detection enhanced classification in computer intrusion detection [C] // LNCS 2388. Berlin, Heidelberg: Springer-Verlag, 2002: 186 - 197.

[7] KIM D, PARK J. Network-based intrusion detection with support vector machines [C] // LNCS 2662. Berlin, Heidelberg: Springer-Verlag, 2003: 747 - 756.

[8] PARK J, SHAZZAD K, KIM D. Toward modeling lightweight intrusion detection system through correlation-based hybrid featureselection [C] // FENG D, LIN D, YUNG M. Proceedings of the CISC. Heidelberg: Springer-Verlag, 2005: 279 - 289.

[9] TAYLOR C, ALVES-FOSS J. NATE: Network analysis of anomalous traffic events, a low-cost approach [C] //Proceedings of the 2001 Workshop on New Security Paradigms. New Mexico: ACM, 2001: 89 - 96.

[10] HORNG S, FAN P, CHOU Y, et al. A feasible intrusion detector for recognizing IIS attacks based on neural networks [J]. Computers & Security, 2008, 27 (3 - 4): 84 - 100.

[11] ZHANG Changyou, CAO Yuanda, Yang Minghua, et al. The immune recognition method based on analogy reasoning in IDS [J]. Wuhai University Journal of Natural Sciences, 2006, 11 (6): 1839 - 1843.

[12] 焦李成，杜海峰. 人工免疫系统进展与展望 [J]. 电子学报, 2003, 31 (10): 1540 - 1548.
JIAO Licheng, DU Haifeng. Development and prospect of the artificial immune system [J]. Acta Electronica Sinica, 2003, 31 (10): 1540 - 1548.

[13] 肖人彬，王磊. 人工免疫系统：原理、模型、分析及展望 [J]. 计算机学报, 2002, 25 (12): 1281 - 1293.
XIAO Renbin, WANG Lei. Artificial immune system: principle, models, analysis and perspectives [J]. Chinese Journal of Computers, 2002, 25 (12): 1281 - 1293.