基于神经网络的安全风险概率预测模型*)

刘 芳[1] 蔡志平[1] 肖 侬[1] 王志英[1] 陈 勇[2]
(国防科技大学计算机学院 长沙 410073)[1] (都柏林大学圣三一学院计算机系 爱尔兰都柏林)[2]

摘 要 网络安全风险概率预测对分布式网络环境及其内在的不确定事件进行动态分析和评价,是构建网络安全保障体系的重要环节。深入研究网络态势感知中的特征提取、聚类分析、相似性度量和预测方法,提出了一种基于神经网络的安全风险概率预测模型。采用入侵检测数据进行了实例验证,仿真实验结果验证了风险预测方法的可行性与有效性。

关键词 网络安全,风险概率,预测模型,神经网络

Risk Probability Estimating Model Based on Neural Networks

LIU Fang[1] CAI Zhi-ping[1] XIAO Nong[1] WANG Zhi-ying[1] CHEN Yong[2]
(School of Computer, National University of Defense Technology, Changsha 410073, China)[1]
(Distributed Systems Group, Department of Computer Science, Trinity College, Dublin, Ireland)[2]

Abstract Risk probability estimating analyzes the distributed network and assesses inherently uncertain events and circumstances dynamically. And the estimating probability of security risk is an absolutely necessary step of developing network security protection system. The problems of feature extraction, cluster analysis, similarity measurement and estimation methods in network situation awareness were addressed. A risk probability assessment formula was proposed, and an estimating model adopting the neural networks was presented. An experiment based on DARPA intrusion detection evaluation data was given to support the suggested approach and demonstrated the feasibility and suitability for use.

Keywords Network security, Risk probability, Estimating model, Neural networks

1 引言

风险分析和评估是辨别各种系统的脆弱性及其对系统构成威胁的过程。风险评价对信息系统的需求分析和设计具有重要的量化参考价值。只有在安全评估中正确、全面地了解信息系统所面临的安全风险,才能在信息安全措施的选择、信息安全保障体系的建设等问题中做出合理的决策[1,2]。

在通常的安全风险分析中,定性预测方法包括德尔菲法、比较法、分解-综合比较法等,信息安全研究领域常常采用的"专家调查法"就是德尔菲法的实际应用。定性预测方法的优点主要是简便易行,具有很好的实用性。但由于预测主要依靠专家,因此归根到底仍属专家们的集体主观判断,不能真正客观地做出预测;而且因为定性预测的方法征询意见的时间较长,对于需要快速判断的安全风险问题(例如在分布式网络环境中)就不太适合使用。

风险概率定量分析方法主要有 Markov 预测法、最大熵法、统计参数解析法等[3-5]。其中,Markov 预测法适用于随机波动性较大的数据列的预测,常常被用来预测风险概率未来变动趋势,但无法准确预测风险概率的取值。风险概率估计方法还有二阶矩法、随机模拟法、非参数估计方法、统计样条法、极端事件风险估计法等。针对信息系统安全复杂多变的特点,还需要有更加灵活实用的安全风险概率预测方法。

安全风险评价是信息系统安全评估的前提和基础,但由于无法准确预测未知的风险事件,导致信息系统安全风险评价中往往采用专家或评估者主观赋值的方法,严重地影响了安全风险评价的客观性与准确性。如何充分考虑信息系统过去的风险信息,基于历史数据对信息系统的安全风险概率进行客观的预测,是本文的研究重点。

针对分布式网络环境中信息交互的安全风险,本文的第 2 节提出了一种风险概率评价模型;第 3 节对特征提取、聚类、预测等关键技术进行了研究,讨论了如何采用神经网络技术解决网络态势感知中的风险预测问题,给出了一种基于神经网络的风险概率预测模型;第 4 节采用 DARPA 的入侵检测评估数据进行了仿真实验,验证了风险概率预测方法的可行性与有效性。最后是文章的总结。

2 安全风险概率预测模型

2.1 安全风险概率预测的基本思想

信息系统之间或信息系统与人之间进行交互的行为,常常可以用一些特征来描述。本文假设在网络环境中,任何信息交互行为(如正常访问、入侵等)都可以表示为特征向量,不同的向量元素可以表示不同的信息交互特征,向量元素之间

*)基金项目:国家自然科学基金(No. 90104025, No. 60603062),国家重点基础研究发展计划(973)(No. 2007CB310901),湖南省自然科学基金(No. 06JJ3035)。**刘 芳** 博士,研究方向为信息安全和体系结构;**蔡志平** 博士,研究方向为网络安全和网络测量;**肖 侬** 研究员,博士生导师,研究方向为体系结构和大规模网络存储;**王志英** 教授,博士生导师,研究方向为先进计算机体系结构、信息系统安全、微处理器设计技术;**陈 勇** 博士,研究方向为信息安全。

是相互可比较的[6]。因此提取的特征向量必须尽可能精确和全面地描述分布式网络环境中信息系统的交互行为。

信息系统中第 $i(i\in N)$次信息交互的安全风险概率可以表示为[6]：

$$P^i=F(X^i)+Z^i \tag{1}$$

$$P^i=F(x_1^i,x_2^i,\cdots,x_n^i)+Z^i \tag{2}$$

其中 P^i 是信息系统的第 i 次信息交互的安全风险概率，X^i 是表征第 i 次信息交互的 n 维特征向量，$x_j^i(j=1,2,\cdots,n)$是向量 X^i 中的第 j 个分量元素，组成信息交互的各个特征分量。特征向量 X^i 描述了与第 i 次信息交互相关的历史记录，它们的值来源于观察或收集的数据。Z^i 是随机扰动因素，通常在分析中可以将其假设为 0[6]。F 是将信息交互的特征映射到安全风险概率的函数，它的具体形式可能是已知或未知的，这取决于不同的实际应用。

当映射 F 已知(可能是线性或非线性的)时，预测 P 的值并不困难。但实际情况中，映射 F 常常是未知的。在这种情形下，F 就像一个“黑盒”，输入当前数据和历史数据，输出风险概率值。因此历史数据对于预测信息系统的信息交互的安全风险概率值非常重要。然而，即使对于过去的信息交互，常常也只能通过信息交互的结果来分析每次信息交互是正常行为还是攻击，与信息交互相关的安全风险概率还是未知的。如果能够得到每次信息交互的历史特征向量所对应的风险概率值，就有可能通过分析确定映射 F 的具体形式，并对新的信息交互的风险概率进行预测[7]。

2.2 安全风险概率的度量

通常，风险概率的定义[8]如下：

定义 1 风险概率是风险发生可能性的百分比表示，是一种主观判断；

定义 2 风险概率是对安全风险语句中风险条件部分描述的事件状态实际发生的可能性的衡量。

本文借鉴风险概率的定义，对分布式网络环境信息交互的安全风险概率进行度量。将信息交互行为用 n 维特征向量 X 表示，则每次信息交互行为对应到 n 维向量空间 R^n 中的一个点。信息交互的特征向量在 n 维向量空间中的分布情况具有以下的特性：相似的特征点密集，不相似的特征点相距较远，即特征点按类群聚[9]，聚集成多个不同的簇(其中簇(cluster)是指一个数据对象的集合，具有同类对象之间具有相似性、而不同类对象之间具有相异性的特点)，且特征点的安全风险与它所属的簇以及簇中的其他点有关。基于式(1)，就可以对第 i 次信息交互 $X^i(i\in N)$所属簇的安全风险概率 $P_r(X^i)$进行近似地度量[6,7]：

$$P_r(X^i)\approx\frac{U(N_r(X^i))}{\| N_r(X^i)\|} \tag{3}$$

其中 $N_r(X^i)$是 n 维向量空间中点 X^i 的一个半径为 r 的邻域，也即 $N_r(X^i)$：$\{Y|\rho(X^i,Y)\leqslant r\}$，$\rho(\cdot)$为特征空间中向量距离的求解函数。$\| N_r(X^i)\|$是 $N_r(X^i)$域中所有点的个数，$U(N_r(X^i))$是该域中所有非正常点的个数(即安全风险发生的次数)。由式(3)可知，信息交互 X^i 所属簇的安全风险概率 $P_r(X^i)$，也即是该簇中所有点的平均风险概率值。

2.3 安全风险概率预测模型

在被评估信息系统的历史交互数据所对应的 n 维向量空间中，当一个簇被确定后，根据历史数据可以得到簇中非正常点的个数，从而可以计算簇中所有特征点的平均风险概率值[7]，所以首先需要对被评估信息系统的历史交互数据进行聚类。由于信息系统的信息交互特征向量没有任何样本分类的先验知识，所以首先需要根据各个特征向量的特征相似程度对历史交互数据进行聚类，聚类结果可以反映特征点在 n 维特征空间中的分布情况，每个信息交互所对应的特征点将位于特征向量空间中的某一个簇内。

设在 n 维特征向量空间 R^n 中，共有 l 个簇，即 $C_n=\{I_1,I_2,\cdots,I_l\}$。其中 $I_j(j=1,\cdots,l)$是特征向量空间中的一个簇，簇 I_j 中有 k 个特征向量，即 $I_j=\{X_j^1,X_j^2,\cdots,X_j^k\}$，$X_j^i(i=1,2,\cdots,k)$是簇 I_j 中的某个特征向量。簇 I_j 中包含的多个特征点$\{X_j^1,X_j^2,\cdots,X_j^k\}$，有的对应正常的信息交互，有的对应非正常的信息交互。根据式(3)，本文将簇中非正常特征点的个数与簇中特征点总数之比，定义为簇的平均风险概率，也就是簇 $I_j(j=1,\cdots,l)$中所有点的平均向量 $\overline{X}_j$ 所对应的风险概率值 $P(\overline{X}_j)$。

通过度量簇 I_j 中每个特征向量 $X_j^i(i=1,2,\cdots,k)$与平均向量 $\overline{X}_j$ 之间的相似程度，可以对簇 I_j 中每个特征点对应的风险概率值进行计算。特征点 X_j^i 的风险概率 $P(X_j^i)$可以用如下函数近似地表示：

$$P(X_j^i)\approx s(X_j^i,\overline{X}_j)\times P(\overline{X}_j) \tag{4}$$

其中，$s(X_j^i,\overline{X}_j)$是表征两个特征向量 X_j^i 和 $\overline{X}_j$ 之间相似程度的相似系数。

因此可以基于信息交互的历史数据，计算特征向量空间中的每个特征向量的风险概率值。这样得到的风险概率预测模型对应于式(1)中的映射 F，由此可以对新的信息交互的安全风险概率进行预测。

综上所述，分布式网络环境中的信息交互的安全风险概率预测模型主要由三部分组成[6,7]：

· 历史数据的预处理

(1)从历史交互数据中抽象出特征属性，生成 n 维的特征向量集合 $C^n\subset R^n$；

(2)基于特征向量，对历史交互数据进行聚类，得到 l 个簇，即 $C_n=\{I_1,I_2,\cdots,I_l\}$。其中 $I_j(j=1,\cdots,l)$是特征向量空间中的一个簇，簇 I_j 中有 k 个特征向量，即 $I_j=\{X_j^1,X_j^2,\cdots,X_j^k\}$，$X_j^i(i=1,2,\cdots,k)$是簇 I_j 中的某个特征向量。

· 预测模型的构造

基于聚类结果，对于每个簇，例如 $I_j=\{X_j^1,X_j^2,\cdots,X_j^k\}$，计算簇中的每一个特征向量的风险概率值；并对特征向量空间中的每个簇依次进行下面的操作：

(1)计算簇 I_j 的平均风险概率值 $P(\overline{X}_j)$：

$$P(\overline{X}_j)=\frac{\sum_{i=1}^{k}E(X_j^i)}{|I_j|} \tag{5}$$

其中 $|I_j|$是簇 I_j 中元素的个数。如果簇 I_j 中第 i 个特征点 X_j^i 所表征的信息交互是非正常的，则 $E(X_j^i)=1$，反之，$E(X_j^i)=0$。

(2)计算簇 I_j 中所有特征点的平均特征向量 $\overline{X}_j$；

(3)计算簇 I_j 中每一个特征向量 $X_j^i(i=1,2,\cdots,k)$与平均特征向量之间的相似系数 $s(X_j^i,\overline{X}_j)$；

(4)计算簇 I_j 中每个特征向量 X_j^i 的风险概率近似值 $P(X_j^i)$：

$$P(X_j^i)\approx s(X_j^i,\overline{X}_j)\times P(\overline{X}_j) \tag{6}$$

通过上述步骤，就可以得到特征向量空间 C^n 中每个特征向量所对应的风险概率。

· 风险概率的预测

对于分布式网络环境中新的信息交互，按照如下步骤，计算对应的风险概率近似值：

(1)从新的信息交互中提取特征值，生成一个新的特征向量 X^p；

(2)对新的特征向量进行聚类分析；

(3)经过聚类，X^p 一般会聚类到特征向量空间中的某一个簇中，设 $X^p \in I_k$；计算 X^p 和簇平均向量 $\overline{X}_k$ 之间的相似系数 $s(X^p, \overline{X}_k)$；

(4)新的特征向量 X^p 的风险概率 $P(X^p)$ 可以按下式计算：

$$P(X^p) \approx s(X^p, \overline{X}_k) \times P(\overline{X}_k) \tag{7}$$

其中 $P(\overline{X}_k)$ 是簇 I_k 的平均风险值。

需要指出的是，基于该预测模型计算得到的风险概率值，是对信息交互的风险概率的近似度量；计算得到的安全风险概率值是一个整体的概念，它不是个体敏感(Case Sensitive)的，它表征的是具有相同特征的一类信息交互行为的风险概率值。

3 安全风险概率预测的关键技术

由于分布式网络环境中信息系统和安全风险的复杂性和多变性，需要对安全风险概率预测模型中的关键技术进行深入研究。

3.1 特征属性的提取

分布式网络环境中，信息系统的交互数据是错综复杂的，因此在从大量数据中发现有意义的知识和规律之前必须对其进行一系列的预处理。

安全风险概率预测中处理的是多指标的问题，即每次信息交互用多个指标(特征分量)来描述。由于指标太多，使得分析的复杂性增加，甚至会使具有显著差别的重要特征指标在总体特征中占的比重变小，因此选取过多的特征反而会增加分析的负担并使分析结果变差。由于在实际工作中，指标间经常具备一定的相关性，故人们希望用较少的指标代替原来较多的指标，但依然能反映原有的大部分信息。因此本文采用主成分分析 PCA(Principal Component Analysis)技术，从特征指标中提取主成分，降低特征向量空间的维数，可以有效提高风险概率预测的效率和准确性。

主成分的概念首先由 Karl Parson 在 1901 年引进，不过当时只对非随机变量来讨论。1933 年 Hotelling 将这个概念推广到随机向量[10]。主成分分析过程实质上是对原坐标系进行平移和旋转变换，使得新坐标系的原点与数据群点的重心重合，新坐标系的第一个轴与数据变化的最大方向对应，新坐标系的第二个轴与第一个轴标准正交，并且对应于数据变化的第二大方向……以此类推。因此经过舍弃少量信息后，新坐标系的前 $m(m<n)$ 个方向能够有效地表示原来数据的变化情况，则原来的 n 维特征向量空间就被降至 m 维[11]。

具体步骤如下：

设有 N 个样本，每个样本有 n 个属性，将原始数据写成矩阵 $X=(x_{ij})_{N\times n}$。

(1)原始数据标准化，仍记为 X；

(2)建立变量的相关系数阵：$R=(r_{ij})_{n\times n}=X'X$；

(3)求 R 的特征根 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ 及相应的单位特征向量：

$$a_1=\begin{bmatrix}\alpha_{11}\\ \alpha_{21}\\ \vdots\\ \alpha_{n1}\end{bmatrix}, a_2=\begin{bmatrix}\alpha_{12}\\ \alpha_{22}\\ \vdots\\ \alpha_{n2}\end{bmatrix}, \cdots, a_n=\begin{bmatrix}\alpha_{1n}\\ \alpha_{2n}\\ \vdots\\ \alpha_{m}\end{bmatrix} \tag{8}$$

(4)写出主成分：

$$F_i=\alpha_{1i}X_1+\alpha_{2i}X_2+\cdots+\alpha_{ni}X_n \quad (i=1,\cdots,n) \tag{9}$$

其中的 $m(<n)$ 个主成分作为新的数据向量，来取代原始数据：

$$X_{PCA}=[F_1,F_2,\cdots,F_m]=X[\alpha_1,\alpha_2,\cdots,\alpha_m] \tag{10}$$

(5)通过主成分分析方法，得出的各个主成分的贡献率为

$$\gamma_i=\lambda_i/\sum_{j=1}^{n}\lambda_j \quad (i=1,\cdots,n) \tag{11}$$

$\gamma_i(i=1,\cdots,n)$ 为各个属性之间的相关度。相关度越大，表示该属性成分综合原数据的能力越强，一般取较大的前 $m(<n)$ 个 γ_i，并且 m 的取值满足 $\sum_{i=1}^{m}\lambda_i/\sum_{j=1}^{n}\lambda_j=\beta$(一般情况 β 取值在[0.8,0.9]为宜)。

经过属性特征提取后，可以最大程度地去掉不重要的和冗余的属性特征。接下来，根据提取出的主成分对信息系统中的信息交互特征进行聚类和建模预测。

3.2 聚类算法

聚类分析是把一个给定的数据对象集合分成不同的簇(cluster)的方法。作为一种没有预先指定类别的无监督分类法，聚类分析常常作为一个独立的分析工具，用于了解数据的分布，或进行数据预处理[10]。

一个好的聚类方法能产生高质量的聚类结果。这些聚类的簇要具备两个特点，即簇内高相似性和簇间低相似性[12]。聚类结果的好坏取决于该聚类方法采用的相似性评估方法以及该方法的具体实现[13]。聚类分析有许多具体的算法。基于分布式网络环境中信息交互的复杂性和动态性，本文采用自组织映射 SOM(Self-Organizing Map)神经网络[14]，对被评估信息系统的历史交互数据进行聚类。

1981 年 Kohonen 提出了自组织特征映射的概念。Kohonen 认为一个神经网络接受外界输入模式时，将会分为不同的对应区域，各区域对输入模式具有不同的响应特征，而且这个过程是自动完成的，其特点与人脑的自组织特性相类似。

具体的自组织特征映射算法[14]可以归纳如下：

(1)初始化。输入单元与输出单元连接的所有权值可以随机地选取某一个较小的值。各个输出单元 j 的邻接输出单元 NE_j 的选取，$NE_j(t)$ 表示在时刻 t 时输出单元 j 的邻接单元的集合，它是随时间的增长而不断缩小的。

(2)输入单元接受一组新的输入 $(x_1,x_2,\cdots,x_n)$。

(3)计算所有输入单元与每个输出单元 j 之间的尤克利距离 d_j：

$$d_j=\sum_{i=1}^{n}(x_i(t)-w_{ij}(t))^2 \tag{12}$$

选取一个最小距离的输出单元 j^*。

(4)按下式修改输出单元 j^* 及其邻接单元集合 $NE_j(t)$ 中的输出单元与所有输入单元 i 之间的连接权值：

$$w_{ij}(t+1)=w_{ij}(t)+\eta(t)(x_i(t)-w_{ij}(t)) \tag{13}$$

其中，$j\in NE_j(t)$，$i=1,2,\cdots,n$；$\eta(t)$ 为一增益项，并随时间下降到 0。

(5)如果还有训练样本，转(2)，否则结束。

从上面的映射算法可以看出，SOM 网络的学习规则根据样本的内在联系，能够对样本进行自适应聚类：输出神经元 j

的权值向量 W_j 逐渐向样本集中的某些样本(这些样本总是以该神经元为获胜神经元)靠近;则权值向量集 $\{W_j \mid j=1,2,\cdots,l\}$($l$ 为输出神经元个数)是对样本集中所有样本的描述,而单个权值向量 W_j 可看作是以它为获胜神经元的所有样本的聚类中心,也即是该聚类中所有样本的平均向量。

3.3 相似性度量

相似性度量可以由两个特征点在 n 维特征空间中的距离来测度,度量值是两个向量各相应分量之差的函数。

设特征向量 $X=(x_1,x_2,\cdots,x_n)$ 属于特征向量空间中的一个簇 $I_j(j\in\{1,\cdots,l\})$,则簇 I_j 的平均向量(聚类中心)为 $\overline{X}_j=(\overline{x}_{j1},\overline{x}_{j2},\cdots,\overline{x}_{jn})$。则 X 与 $\overline{X}_j$ 的相似系数 $s(X,\overline{X}_j)$ 可以表示为:

$$s(X,\overline{X}_j)=\frac{\sum_{i=1}^{n}x_i\times\overline{x}_{ji}}{\sqrt{(\sum_{i=1}^{n}x_i^2)(\sum_{i=1}^{n}\overline{x}_{ji}^2)}} \tag{14}$$

本质上来说,式(14)中是两个特征向量 X 和 $\overline{X}_j$ 的欧式(Euclidean)距离相似性测度。两个特征向量越相似,相似系数 $s(X,\overline{X}_j)$ 的值就越大,上限为 1。

3.4 预测方法

预测分析的过程是从过去和现在已知的情况出发,利用一定的方法或技术去探索或模拟不可知的、未出现的或复杂的中间过程,再推断出未来的结果。

基于信息系统中信息交互历史数据的特点,本文采用非线性预测方法中的人工神经网络方法对风险概率进行预测。神经网络通过在一定程度上模拟人脑的结构和功能,已经具备了较强的泛化能力,使得神经网络预测方法适用于多种模式的数据数列的预测,在对分布式网络环境中的信息交互进行风险概率预测分析时具有以下优点:

(1)实现了非线性关系的隐式表达,不需要建立信息系统安全风险的显式关系和数学模型;

(2)容错性好,可以处理数据和信息不全的预测问题,而由于多种原因,历史数据和资料不全的情况在安全风险分析过程中经常遇到。

神经网络的种类很多,其中采用误差反向传播 BP(Back Propagation)算法的多层感知机(Multilayer Perceptron)[14]是应用最广泛的方法之一,这种多层神经网络通常被简称为 BP 网络。它具有良好的非线性映射能力、结构简单、性能良好,与其他传统模型相比,有更好的持久性和预测性。

BP 神经网络是一种有指导的学习方法,其学习过程由正向传播和反向传播组成。在正向传播过程中,当一对学习模式提供给网络后,神经元的激活值,从输入层经各隐含层向输出层传播,在输出层的各神经元获得网络的输入响应。如果输出层得不到期望的输出,则转入反向传播,将误差信号沿原来的连接通道返回,从输出层经隐含层逐层修正各连接权,最后回到输入层,使得误差信号最小。BP 网络间连接权在网络的学习过程中不断得到修正,使输入层与隐含层之间、隐含层与输出层之间的两组权所构成的网络能实现学习样本中输入矢量与输出矢量间特定的映射关系,权的分布体现了各输入分量在输入矢量中所占的特征强度。随着这种误差逆传播修正的不断进行,网络对输入模式响应的正确率也不断上升。

在安全风险概率预测分析中,可以直接使用 BP 网络模型实现系统输入参数与输出参数之间的非线性映射。在网络学习过程中,不断调整网络参数,直到得到满意的输出。基于本文第 2 节中提出的安全风险概率预测模型,利用 BP 神经网络进行风险概率预测分为三大步骤:

• 第一步为训练样本的准备,包括对历史交互数据的预处理(特征提取和聚类)和特征向量风险概率值的计算,得到被评估信息系统的历史交互的特征向量集合以及每个向量所对应的风险概率值,生成 BP 网络的(输入向量,输出响应)训练对的集合 $\{(X^i,P^i)\mid i=1,2,\cdots,k\}$。

• 第二步为神经网络的训练。用第一步得到的训练对集合 $\{(X^i,P^i)\mid i=1,2,\cdots,k\}$,对 BP 神经网络进行训练。

• 第三步是利用训练后的神经网络对安全风险概率进行预测。训练完成后,对于被评估信息系统中新的信息交互,提取特征向量,将新向量作为神经网络的输入,得到的输出即是新的信息交互的风险概率值。

4 仿真实验

为了验证本文提出的安全风险概率预测方法的有效性,我们进行了仿真实验。实验数据均来自美国国防部高级研究计划署(DARPA)提供的入侵检测数据库[1],整个仿真实验均用国际流行和公认最佳的社会科学数据分析软件包 SPSS 13.0 (Statistical Package for the Social Science)和 Matlab6.5 编程实现。

4.1 试验设置

本文利用 DARPA 入侵检测数据库中的数据做仿真实验。DARPA 入侵检测评估项目开始于 1998 年,由麻省理工大学(MIT)的 Lincoln 实验室启动和管理,并由 DARPA 机构和美国空军研究实验室资助。Lincoln 实验室设置了一个模拟的局域网,收集 9 个星期的 Raw TCP DUMP 格式的网络链接数据,其中模拟了政府和空军的 1000 个主机上的 100 个用户的正常通讯,同时包含了 38 种攻击。其中 7 个星期训练数据集包含 500 万个链接数据,2 个星期的测试数据集包含了 200 万个链接数据。每个链接包含大约 100 字节。

38 种攻击被分为以下 4 种主要类型,其中 R2L 和 U2R 攻击所占比例很小。

• 拒绝服务攻击(DoS):例如泛洪攻击等;

• 远程攻击(R2L):从远端机器上的非授权读取,例如基于字典的口令猜测;

• 本地用户非法提升权限的攻击(U2R):非授权地获取管理员权限,例如各种"缓冲区溢出"攻击;

• 网络扫描(Probing):监视和其他探测技术,包括端口扫描和漏洞扫描。

为了从包含大量冗余信息的数据中提取出尽可能多的安全风险信息,抽象出有利于进行判断和比较的特征集合,Wenke Lee 在这方面做了大量的工作,建立了 KDDCUP'99 项目。Lee 从 DARPA1998 数据中抽取 41 维特征,分为基本特征(basic features)、内容特征(content features)、两秒钟内的流量特征(traffic features computed using a two-second window)、主机流量特征(host-based traffic features)。本文对全部数据包中 10% 的数据进行了分析,这部分数据共计 494021 条记录,数据包大小共约 72M。

4.2 主成分分析仿真结果

风险概率预测中,因为历史数据量巨大,且属性众多,所

[1] KDDCUP99, http://kdd.ics.uci.edu/databases/kddcup99/task.html

以首先需要对实例数据进行因子分析。

仿真实验中，采用的数据具体说明如下。

- 数据集合名称：KDD Cup
- 数据库来源：DARPA 入侵检测数据库
- 样本个数：494021
- 样本属性数：41（每个特征向量有 41 个元素，标记为 F1 到 F41）
- 样本属性说明：TCP 链接记录
- 数据集说明：模拟局域网的网络链接数据，仿真实验中表示信息交互数据。

为了验证对数据进行预处理的有效性，本文采用 SPSS 中的主成分分析功能对历史数据进行因子分析。实验中去除了字符变量和几乎全为 0 的变量，对剩下的 36 个变量进行分析。其参数选择为：主成分提取的特征值为大于 0.9、未经旋转的主成分提取结果、要求提取主成分分析矩阵、Varimax 旋转法、主成分相关系数矩阵等。通过主因子分析，我们得到各个特征分量的贡献率（Communalities），特征分量的贡献度越大，表示该属性综合原数据的能力越强。初始的主成分提取结果如表 1 所示。

在表 1 中，

- Eigenvalue：各主成分的特征值；
- Extraction Sums of Squared Loadings：平方和载荷；
- % of Variance：边际贡献，即各主成分的方差占方差总和的百分比；
- Cumulative %：累计贡献，即各主成分方差占总方差百分比的累计百分比。

以累计贡献率的大小决定主成分的选取。计算第 i 个主成分对总方差的贡献率，按贡献率大小的顺序对 $k(k=36)$ 个主成分进行排序，贡献率最大的主成分称为第一主成分，其次称为第二主成分，依此类推。选取主成分的个数取决于主成分的累计方差贡献率，累计贡献率越大，丢失的数据信息就越少，但后续处理计算量大。本文取特征值大于 0.9 的所有主成分。从表 1 中可以看出，取前面 14 个主成分，累计贡献率已经达到 84.415%，即前 14 个主成分已经对大多数数据给出了充分的概括，可以解释总方差的 84.415%。通过主成分分析，减少了 22/36=61%的数据处理量，丢失的信息量也较小。

表 1　初始主成分提取结果

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.022	22.285	22.285	8.022	22.285	22.285
2	4.513	12.536	34.821	4.513	12.536	34.821
3	3.189	8.857	43.678	3.189	8.857	43.678
4	2.955	8.208	51.886	2.955	8.208	51.886
5	1.847	5.129	57.016	1.847	5.129	57.016
6	1.524	4.235	61.250	1.524	4.235	61.250
7	1.191	3.309	64.559	1.191	3.309	64.559
8	1.142	3.172	67.731	1.142	3.172	67.731
9	1.101	3.059	70.790	1.101	3.059	70.790
10	1.012	2.811	73.602	1.012	2.811	73.602
11	1.000	2.778	76.379	1.000	2.778	76.379
12	.994	2.761	79.141	.994	2.761	79.141
13	.966	2.685	81.825	.966	2.685	81.825
14	.932	2.590	84.415	.932	2.590	84.415
15	.867	2.407	86.823			
16	.848	2.356	89.179			
17	.779	2.163	91.342			
18	.727	2.019	93.361			
19	.716	1.988	95.350			
20	.380	1.056	96.406			
21	.364	1.012	97.418			
22	.335	.931	98.348			
23	.161	.447	98.796			
24	.150	.417	99.213			
25	.138	.383	99.595			
26	.046	.129	99.724			
27	.027	.075	99.799			
28	.018	.051	99.850			
29	.017	.048	99.899			
30	.015	.043	99.942			
31	.007	.019	99.961			
32	.006	.016	99.977			
33	.005	.014	99.991			
34	.001	.004	99.995			
35	.001	.003	99.999			
36	.000	.001	100.000			

Extraction Method：Principal Component Analysis.

主成分分析过程中，同时还提取了主成分矩阵，表明了主成分和单个特征分量之间的关系。根据主成分矩阵，对原始数据集 KDD Cup 中的数据进行主成分分析计算，得到新的数据集 KDD Cup(T)，其中每个数据（特征向量）都有 14 个分量元素。

4.3　聚类分析仿真结果

从经过主成分计算的数据集 KDD Cup(T)中随机采样约 5%的数据进行聚类，共计 24626 个样本。

- 数据集合名称：KDD Cup(T)
- 样本个数：24626
- 样本属性数：14
- 样本属性说明：TCP 链接记录
- 数据集说明：该数据集中的数据是信息交互行为的特征向量。

采用 SOM 聚类算法对样本进行聚类。24626 个样本向量通过自适应聚类，分为了 18 个簇，图 1 对聚类数据对 18 个簇中的分布情况给出了直观的描述。

如图 1 所示，第 6 个簇中包含的特征向量最多，其中包含有 39%的 TCP 链接。对聚类结果进行分析，第 6 个簇中并不包含任何攻击链接，在这个簇的网络链接的风险概率值为 0。同样，对所有的簇进行分析，自适应聚类得到的 18 个簇中，只有 8 个簇中包含有非正常链接（攻击），其余的 10 个簇中特征向量都是正常的 TCP 链接。

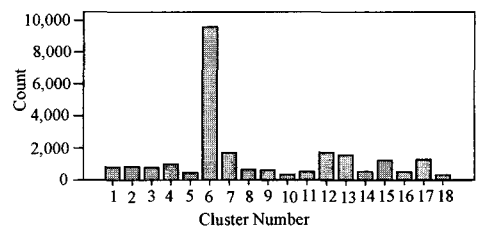

图 1　聚类数据的分布

基于聚类结果，可以计算得到每个样本的风险概率值，生成 24626 个训练对。

4.4　预测分析仿真结果

从经过主成分计算的数据集 KDD Cup(T)中随机采样约 1%的数据作为测试数据，利用聚类分析的结果对预测问题进

行仿真。

- 数据集合名称:KDD Cup (T)
- 训练集样本数:24626
- 测试集样本数:4940
- 条件属性数:14
- 决策属性数:1
- 数据集说明:该数据集中的数据是信息交互行为的特征向量。

(1)本文采用 BP 网络模型进行训练和预测,建立的 BP 模型是输入节点到输出节点的高度非线性映射模型。按照前面的分析,设计 BP 网络为双隐层结构,输入层有 14 个节点(对应于 14 个主成分),输出层 1 个节点,隐含层选取有 18 个节点,即构成 14-18-1 结构的 BP 神经网络;基于经验设定训练速率为 0.9 次/s;允许最大的迭代次数为 10 万次。

对测试数据的风险概率值进行预测的部分结果如表 2 所示,表中的特征向量属于图 1 中的第 12 个簇。从 V1 到 V14 的纵栏,是经过主成分分析,从原始数据的所有 41 个特征中提取出来的 14 个主成分因子。"簇"纵栏表明了特征向量属于哪一个簇。"链接类型"纵栏描述了网络链接的类型。"相似系数"描述了该特征向量与簇平均向量的相似程度,"风险概率"是通过 BP 神经网络计算得到的风险概率值。从表中可以看出,结果表明同种类型的网络链接风险值比较接近,而不同的网络链接的风险值差异较大。

表 2 预测的部分结果

1	V10	V11	V12	V13	V14	簇	链接类型	相似系数	风险概率
2	−38	250	403	149	−6	2	smurf	0.43375	0.23162982
3	−64	269	1128	400	−22	12	smurf	0.54115	0.288983232
4	−62	253	1129	400	−20	12	smurf	0.5109	0.272829221
5	−84	270	1721	605	−32	12	smurf	0.54715	0.292187333
6	−82	262	1643	576	−37	12	smurf	0.5787	0.309035566
7	−144	268	3474	1208	−80	12	smurf	0.63555	0.339394425
8	−48	231	659	235	−11	12	smurf	0.5164	0.275766314
9	−43	225	521	187	−7	12	smurf	0.4062	0.216917655
10	−82	229	1682	588	−31	12	smurf	0.3573	0.190804229
11	−49	241	647	232	−5	12	neptune	0.13355	0.071317954
12	−65	285	983	349	−17	12	neptune	0.05895	0.031480295
13	−87	307	1552	546	−28	12	neptune	0.0816	0.043575777
14	−36	246	210	81	7	12	ipsweep	0.1824	0.097404678
15	−36	237	205	79	11	12	ipsweep	0.29755	0.158896721
16	−44	244	417	152	7	12	ipsweep	0.28975	0.15473139
17	−70	288	1071	380	5	12	ipsweep	0.30605	0.163435865
18	−36	238	204	79	17	12	ipsweep	0.315	0.168215316
19	−38	245	207	81	17	12	ipsweep	0.2905	0.155131902
20	−61	262	812	288	−9	12	ipsweep	0.2234	0.11929937
21	−92	294	1622	568	−28	12	ipsweep	0.16985	0.090702766
22	−99	303	1801	630	−32	12	ipsweep	0.1728	0.092278116
23	−97	398	1475	522	−24	12	ipsweep	0.1725	0.092117911
24	−151	298	3389	1178	−72	12	ipsweep	0.22855	0.122049557

结束语 安全风险概率是网络安全态势感知中,安全风险评价的关键因素。本文给出了一种基于历史数据的动态的安全风险概率预测模型,该方法基于分布式网络环境的特点,预测模型的构建过程可以离线进行,可以对新的信息交互产生动态的响应,减少安全评估中的主观性。通过仿真实验,我们认为使用该方法进行安全风险概率预测是可行的,但是在实际操作中需要更多地考虑具体的网络状况和应用背景。

参 考 文 献

[1] 陈秀真,郑庆华,管晓宏,等. 层次化网络安全威胁态势量化评估方法. 软件学报,2006,17(4):885-897

[2] 张永铮,方滨兴,迟悦,等. 用于评估网络信息系统的风险传播模型. 软件学报,2007,18(1):137-145

[3] 蒋盛益,李庆华,王卉,等. 一种基于聚类的有指导的入侵监测方法. 小型微型计算机系统,2005,26(6):1042-1045

[4] 陈婷婷,方滨兴,郑军. 基于层次自组织特征映射的网络异常检测系统数据分析器. 计算机应用与软件,2006,23(5):3-8

[5] 王坤,郭云飞. 基于 PCA 的无监督异常检测方法研究. 郑州大学学报,2006,23(5):3-8

[6] Chen Yong, Jensen C, et al. Risk Probability Estimating Based on Clustering[C]// Proc. the 4th IEEE Annual Information Assurance Workshop. West Point, New York, USA, June 2003

[7] Liu Fang, Chen Yong, Dai Kui, et al. Research on Risk Probability Estimating using Fuzzy Clustering for Dynamic Security Assessment[C]// Duentsch Ivo et al. Eds. The 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. LNAI, Regina, Saskatchewan, Canada, September 2005

[8] The American Heritage Dictionary of the English Language. Fourth Edition. Copyright 2000 by Houghton Mifflin Company

[9] 孙即祥. 现代模式识别[M]. 长沙:国防科技大学出版社,2002

[10] Lee C, Landgrebe D A. Analyzing High-dimensional Multispectral Data[J]. IEEE Transactions Geosci. Remote Sensing, 1993, 31(4):792-800

[11] Carreira-Perpinan M A. Continuous Latent Variable Models for Dimensionality Reduction and Sequential Data Reconstruction [D]. Ph. D Thesis. February 2001

[12] Yeung K Y, Ruzzo W L. Principal Component Analysis for Clustering Gene Expression Data. Bioinformatics [M], 2001, 7(9): 763-774

[13] Raychaudhuri S, Stuart J M, Altman R B. Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series[C]// Proc. Pacific Symposium on Biocomputing. 2000

[14] 戴葵. 神经网络实现技术. 长沙:国防科技大学出版社,1998

[15] Liu Fang, Dai Kui, Wang Zhiying. Improving Security Architecture Development Based on Multiple Criteria Decision Making// Chi C-H, Lam K-Y, eds. AWCC 2004 LNCS 3309. November 2004:214-218