一种基于PCA技术的入侵检测特征提取方法*

A Feature Extracting Method for Intrusion Detection Based on PCA

钟淑瑛，李陶深，张 敏

Zhong Shuying，Li Taoshen，Zhang Min

（广西大学计算机与电子信息学院，广西南宁 530004）

（School of Comp.，Elec. and Info.，Guangxi Univ.，Nanning，Guangxi，530004，China）

摘要：为了提取入侵检测数据信息的特征和提高入侵检测系统的处理效率，提出一种基于PCA（Principal Components Analysis）技术的入侵检测特征提取方法，并利用Matlab统计工具箱对该方法进行仿真实验。实验结果表明，利用该方法所提取的特征足以代表入侵检测数据的主要信息，根据特征提取结果所进行的数据压缩处理是可行的。

关键词：入侵检测 PCA 特征提取 Matlab

中图法分类号：TP393.03 **文献标识码**：A **文章编号**：1002-7378(2005)04-0244-03

Abstract：In order to extract the feature of intrusion detection data and increase the efficiency of intrusion detection system, a feature extracting method for intrusion based on PCA is revealed. The relevant simulation experiment using the Matlab statistics toolbox is carried out. The experimental results show that the extracting features in this method are enough to represent the main information of intrusion detection data, and the performing data compression in response to extracting feature information is feasible.

Key words：intrusion detection，PCA，feature extracting，Matlab

随着Internet的快速扩展，人们需要处理的数据信息不再局限于主机系统，而是囊括了大批量的网络数据信息。当处理系统无法负荷海量数据时，人们一方面努力提高系统处理效率，另一方面则是采用相关技术对海量数据进行特征提取，以达到压缩数据的目的。

本文提出一种基于的主成分分析（Principal Components Analysis，简称PCA）技术，针对入侵检测数据的特征[1]提取方法，并利用Matlab的统计工具箱对该方法进行仿真实验。实验结果表明，利用该方法提取的特征囊括了入侵检测数据的主要信息，根据特征提取结果所进行的数据压缩处理是可行的。

收稿日期：2005-06-24

作者简介：钟淑瑛（1979-），女，广西南宁人，硕士，助教，主要从事分布式入侵检测系统研究。

* 广西科技攻关项目（桂科攻0385001）和广西留学回国人员科学基金项目（桂科回0342001）联合资助。

1 基于PCA技术的入侵检测特征提取方法

1.1 入侵检测流程

在入侵检测系统能够对信息流做出判别之前，必须经过如图1所示的几道处理流程。

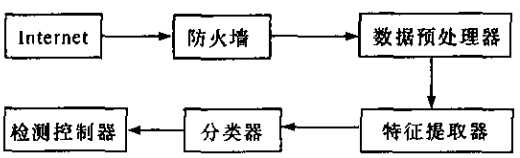

图1 入侵检测系统信息流程

从图1可以发现，特征提取器的地位举足轻重，它和分类器连接在一起，是沟通数据预处理器和检测控制器的桥梁。本文提出的特征提取方法，从网络安全的角度出发，将特征数据源定位在网络数据上。

1.2 基于PCA技术的特征提取方法

PCA技术可以将数据从高维数据空间变换到低维特征空间，因而可以用于数据的特征提取及压缩等方面。近年来，PCA分析技术已经由传统的针

对数据进行线性空间压缩，转向基于神经网络方法的非线性组合模式，对数据进行非线性空间压缩[2]。从数学的角度来看，其根本思想在于降维，而降维是从简化方差和协方差的结构考虑。

PCA 技术的具体操作就是丢弃作为原变量的次要成分存在的线性组合，保留最重要的线性组合。如果选择保留所有的线性组合（线性组合的数目不会明显的增大数据矩阵的秩），那么次要成分则代表线性组合的有用信息。第1个主要成分为指向输入模式分类的最多变量方向的向量，第2个成分则为垂直于第1个向量的向量，指向第2个多变量的方向，以此类推[1]。以上问题的解决可以简化为关于模式相关矩阵 C 的特征值或特征向量问题。主要成分的次序则由相应特征值的取值决定。

我们利用 PCA 技术进行主成分分析的主要步骤是先由原始数据矩阵（设为 n 行 m 列）求出相关矩阵 V（m 阶方阵），再求出 V 的 m 个特征值（按由大至小的顺序排好）：$\lambda_1,\lambda_2,\cdots,\lambda_m$，以及相应的已正交标准化的特征向量 $\nu_1,\nu_2,\cdots,\nu_m$。其中，ν_1 是这样一个向量，以它的各个分量作为系数，求出的各变量的线性组合，就是第1主成分。同样地，以 ν_2 为系数，求出各变量的线性组合，就得到第2主成分。通常取 2～3 个主成分已经足够能够包含或者代表原有数据的全部信息。将原始数据矩阵中的 n 行信息数值代入上述各主成分的线性组合公式，就可以得到各行信息的主成分分值（score）。将每行个体的前2～3个主成分分值在二维或三维空间中点成散点图。从散点图就可以很清楚地看出每行信息的地位和各条信息之间的关系。我们设计的基于PCA 技术的特征提取方法的实现算法如下。

步骤1：设1行 m 列的矩阵 O_i 表示1条入侵检测原始数据，入侵检测数据集中共有 k 条数据，$i=1,2,\cdots,k$。根据原始矩阵 O_i 求出相关矩阵 X_i，其中 X_i 为 m 阶方阵。初始化，置 $i=1$。

步骤2：求出 X_i 的 m 个特征值（按由大至小的顺序）：egenvaluei1，egenvaluei2，…，egenvalueim，以及相应的正交标准化特征向量 $\upsilon_{i1},\upsilon_{i2},\cdots,\upsilon_{im}$，其中 $\upsilon_{i1},\upsilon_{i2},\cdots,\upsilon_{im}$ 均为 n 行列向量。

步骤3：以 υ_{i1} 的各个分量作为系数，求出的各变量的线性组合，得到第一主成分。以 υ_{i2} 的各个分量作为系数，求出各变量的线性组合，得到第二主成分。以此类推，通过 υ_{im} 得到第 m 主成分。

步骤4：置 $i=i+1$，重复步骤1、步骤2、步骤3，当 $i=k$ 时，算法结束。

2 仿真实验结果与分析

2.1 基于MATLAB 的PCA 技术实现

将入侵检测的原始数据矩阵中的 n 行信息数值代入主成分线性组合公式，就可以得到各行信息的主成分分值。本实验中基于MATLAB 实现PCA 的程序的一般语句格式为：

[pc，scores，variance，T2]=princomp(X)

其中，右端的输入数值是数据矩阵，左端输出的 pc 为 m 阶方阵，每列是相应的特征向量数值，也即是每个主成分在各原始变量上的系数。第1列是相应于最大的特征值的特征向量，因此它就是第1主成分在各变量上的系数；第2列是第2主成分的系数，以此类推。scores 是一个 n 行 m 列的矩阵，第1行代表第一条信息的 m 个主成分分值，第2行代表第2条信息的 m 个主成分分值，以此类推。variance 为1维 m 元列向量，它代表每条信息到样本中心的距离，数值越大表示此点离样本中心越远。

2.2 结果与分析

基于 PCA 分析的仿真实验所采用的数据集为DARPA[3]的KDD 数据集[4]。为了方便阐述，我们用IDSdata 来表示DARPA 的KDD 数据集。根据上述的程序 princomp 左端的前3个输出量，分别给出相关 PCA 分析结果显示图。

从图2可以看出，IDSdata 数据集中38个变量项中，只有src _bytes，dst _bytes，logged _in，count，srv _count，same _srv _rate，dst _host _srv _count，dst _host _same _srv _rate 这8个变量项在权值上发生了变化。发生权值变化的变量项不在数据压缩的范围之内。

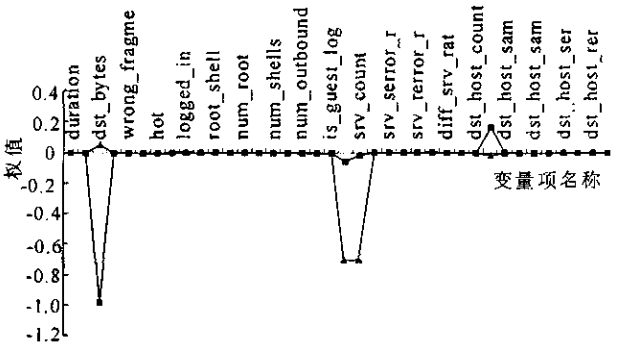

图2 IDSdata 数据集的第1、2主成分分布

—■—：第1主成分；—▲—：第2主成分。

从图3所示的box 图中可以看到IDSdata 的大致情况是：count 和srv _count 以及dst _bytes 这3个变量项的数值变化要比其他项的大得多，说明以上3

项变量是我们必须考虑的对象之一。

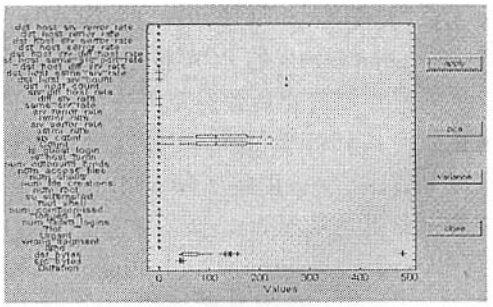

图3 PCA的box图

princomp函数的第2个输出量scores，是IDSdata在主成分所定义的新坐标系中的投影坐标，它与输入信息矩阵的大小相同。将scores的前2～3个主成分分值在二维或三维空间中点成散点图（如图4所示），就可以很清楚地看出每行信息的地位和各条信息之间的关系。从图4可以看出，IDSdata的投影分布较为集中，主成分分析的第1、第2主成分能够充分代表IDSdata的主要内容。

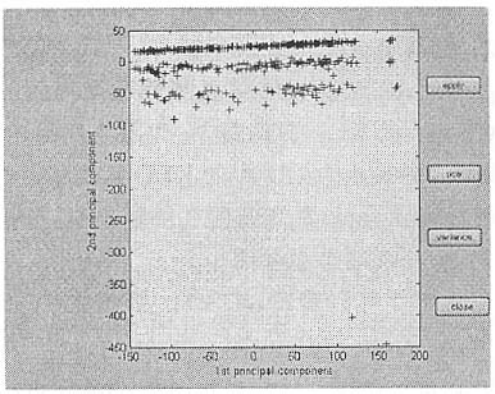

图4 PCA的散点图

princomp函数的第3个输出变量variance，是IDSdata矩阵各列数据相应的方差。各主成分的方差占初始数据方差总和的比例代表该主成分在数据集中的地位，将每个主成分方差占初始数据方差总和的比例在二维空间中形成柱形图，即得主成分的pareto图（见图5）。

从图5中可以看到，第1、2主成分的方差之和所占比例已经占了总成分的98%，第1、2主成分能充分表征入侵检测数据的数据特征，说明本文提出的PCA特征提取方法在解决入侵检测数据特征提取问题上，具备一定的特征提取完备性和充分性。我们可以保留了IDSdata数据集中较大权重值所对应的变量项以及在box图中count和srv_count以及dst_bytes这3个变量项，作为PCA处理后新生成的数据库的主干变量项；在考虑其余变量项的实际意义的基础上，适当删减了影响较小的变量项。这样通过PCA分析可将原来的数据集变量由原来的38个减少为20个，大大减少了后面工作的工作量，为以后的实验提供数据基础。

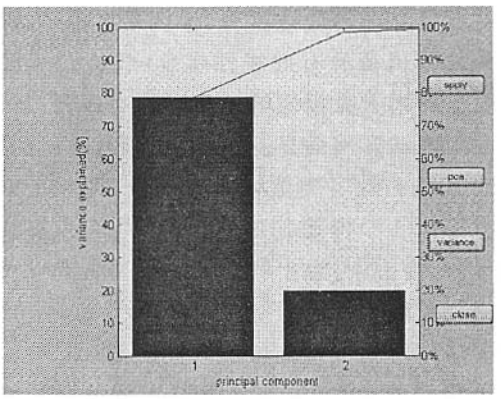

图5 PCA的pareto图

3 结束语

采用PCA技术进行主成分分析可以较直观地观察到入侵检测数据集中为高维特征向量的信息。本文提出的基于PCA技术的入侵检测特征提取方法能够提高入侵检测系统的处理效率。利用MATLAB统计工具箱对该方法进行仿真实验的结果表明，利用该方法所提取的特征足以代表入侵检测数据的主要信息，在此基础上所进行的数据压缩处理是可行的。

参考文献：

[1] Elsevier Computer Science Editorial. Intrusion detection[J]. Information Fusion, 2003, (4): 243-245.

[2] Tzafestas E S, Nikolaidou A, Tzafestas S G. Performance evaluation and dynamic node generation criteria for 'principal component analysis' neural networks [J]. Mathematics and Computers in Simulation, 2000, (51): 145-156.

[3] Kdd_Cup_Dataset[EB/OL]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 2004.

[4] Richard Lippmann, Joshua W Haines, David J Fried, et al. The 1999 DARPA online intrusion detection evaluation[J]. Computer Networks, 2000, (34): 579-595.

（责任编辑：邓大玉）