摘 要

网络流量的特性分析一直是通信网络性能分析的一个极其重要的问题。对网络流量的理解对解决许多网络方面的问题很重要，诸如流量预测，异常检测和容量规划等。但是网络流量呈现高维多变的特性，目前还没有一种很好的方法进行研究。

PCA（主成分分析法）是一种最常用的分析高维对象的方法，成功应用于图像识别、神经网络等领域。由于网络流量同样存在高维问题，因此，Anukool Lakhina 等人提出引入 PCA 算法进行分析。

实验在一组真实的网络流量数据上进行。采用 PCA 算法对其进行处理后，我发现整个流量呈现低维特性，即，一个网络有近百条流量，这些流量能够用一部分独立的变量较为精确的近似。

接下来，我还研究了如何运用 PCA 将流量的内在特征分解成三部分：共同的周期性的趋势，短暂生存的脉冲型的流量和背景噪声。这三类特征对于研究每条流量的结构提供了一个更直观的工具，其中脉冲型的流量有助于检测网络流量的突发点。然后，我还初步研究了 PCA 的这一分解在时域上的稳定性，即采用 PCA 对前一时段流量的分析结果对于分析后一时段的流量是否有效。

总的来说，根据仿真结果来看，将 PCA 算法引入流量分析领域还是可行的。

【关键词】：PCA，网络流量分析，主成分，特征流

Abstract

The characteristic analysis of network traffic is an important issue in communication network. A thorough understanding of network traffic is essential for addressing a wide variety of network problems, including traffic forecasting, anomaly detection, capacity planning. However, at present, it's difficult in analyzing the network traffic because of its high dimensional multivariate structure.

PCA (Principal Component Analysis) is a most commonly used technique to analyze high dimensional structures. It has been successfully used in the field including image recognition, neural network. So it's an attempt to apply PCA in network traffic analysis.

I apply PCA on a real network traffic data. Then I find that the set of flows has small intrinsic dimension. In fact, even in a network with nearly a hundred flows, these flows can be accurately modeled using a small number of independent components.

Then, I show how to use PCA to decompose the structure of flows into three constituents: common periodic trends, short-lived bursts, and noise. This provides a tool in understanding the character of flow. Furthermore, I explore the extent to which this decomposition varies over time. That is, whether the original decomposition is still useful in the subsequent decomposition.

Token together, simulation experiments confirm that applying PCA in network traffic analysis is feasible.

第一章 引言

1.1 概述

网络流量分析一直是通信网络性能分析的一个极其重要的问题。但是目前，网络流量分析的许多工作都是集中在孤立地研究网络中单条链路的流量上。然而，当今网络研究者面临的更重要的问题是需要同时对网络所有链路的流量进行建模和分析，包括流量工程，流量矩阵估算，异常检测，攻击检测，流量预测和容量计划等。

然而，同时对网络所有链路的流量进行分析是一个难题，因为仅对一条链路的流量建模都很复杂。因此，全网络流量分析仍是一个复杂而重要的挑战。

解决全网络流量分析的一个办法是必须认识到网络中不同链路的流量不是相互独立的，实际上是由一系列 OD 流和一个路由矩阵决定的。一个 OD 流是流量的集合，这些流量从一个入口点进入网络，从一个出口点离开。这些由路由决定的点到点流量的重合就产生了所有链路的流量。因此，代替研究所有链路的流量，一个更直接和根本的方法是研究网络的 OD 流。

然而，尽管在概念上 OD 流比链路流量更接近网络的特征，研究它们也遇到同样的问题。最主要的问题是 OD 流呈现高维多变特性。这样，最主要的问题就是高维问题。

通常，当要分析高维对象时，一个常用而有效的方法就是通过低维近似它的主要特性。因为，通常由于高维导致的复杂性可能由一小部分独立的变量控制，因此能够很好地由这一小部分近似。维数分析和维数

降低技术就是找到这些变量，从而更好地理解原对象。

而最常用的分析高维对象的方法是 PCA 算法（主成分分析法，也叫 Karhunen-Loeve 变换或特征值分析法）。给定一个高维对象和相关的坐标空间，PCA 能够找到一个适合给定对象降维的新的坐标空间。一旦将这个对象放入这个新空间，只用一部分坐标表示这个对象就可以达到误差最小。当一个高维对象能以这种方式用一部分维数近似时，我们称这小部分维数是这个对象的内在维数。

1.2 论文主要工作

Anukool Lakhina，Mark Crovella，Christophe Diot 等人[1,6,7]提出将 PCA 算法用于网络流量分析领域。因此，在这篇论文中，我将采用 PCA 算法来研究采样到的某个真实网络流量的内在维数和结构，而不涉及具体的采样方法。尽管我研究的这个网络有近百条流量，但是我们将看到：在很长一段时间里，它的结构也能很好地用小维数来近似。事实上，我们将发现：只要采用 3 到 7 维就能较好地近似网络中的所有流量。

引入特征流的概念能够更好地研究这一低维特性。一个特征流捕获了网络中所有流量的某一特性，每条流量都能表示成特征流的加权，加权系数表示每个特征流的重要性。特征流的这一重要特性使我接下来研究它的性质。研究发现，特征流可分为三类：

(1). 确定型的特征流，即反映流量中可预测的、周期的趋势的特征流。

(2). 脉冲型的特征流，即反映流量中一些偶然爆发的、生存时间很短的脉冲形式的流量。

(3). 噪声型的特征流，即反映了流量中类似高斯噪声的、相对稳定的流量。

这样，通过把流量分解成这些特征流的组合，我们就能够比较直观地看出各流量的内在结构，同时，也能够更好的了解整个网络的行为。

事实上，将特征流按照这种方式进行分类，我们将发现，我们能够获得整个网络流量的重要信息。首先，我们发现每条流量都能由一小部分特征流近似。这样，每条流量都有自己相应的某些特征。其次，这些特征可以以一种可预知的方式变化。特别的，我们发现网络中有的流量呈现可预见的周期性变化的趋势；有的流量是由脉冲型的和噪声型的特征流组成；还有的流量仅仅是一些脉冲型的流量或者仅仅是一些噪声型的流量。因此，特征流的这一分类方法为我们重构和认识整个网络流量提供了一个有用的工具，而其中脉冲型的流量有助于检测网络流量的突发点。然而，这篇论文对特征流的研究还有待于进一步的深入。

论文还对 PCA 算法对网络流量的分解在时域上的相对稳定性进行了研究。实验发现，PCA 对上一时段流量的分析结果对下一时段流量的分析也很有帮助。

最后，从一个更宽的角度来说，这篇论文是对网络流量进行维数分析。这篇论文研究的网络流量的内在维数和结构也许会对研究网络流量的其他行为有所帮助，这也是这篇论文的后续工作。

这篇论文的结构如下。第二章中我将简单描述网络流量分析的现状。第三章中我将详细介绍 PCA 算法的相关知识，通过一个简单的例子和相应的图形进行介绍。第四章将介绍在网络流量分析领域运用 PCA 算法的详细步骤。在第五章中，我将给出在一组真实的网络流量数据上运用 PCA 算法进行处理后的仿真结果及一些相关的结论。最后在第六章中总结论文的主要工作和一些后续工作。

第二章　网络流量分析

2.1 概述

网络流量监测是网络管理和系统管理的一个重要组成部分，网络流量数据为网络的运行和维护提供了重要信息。这些数据对网络的资源分布、容量规划、服务质量分析、错误监测与隔离、安全管理都十分重要。网管人员可以利用它们来监控网络的数据流量，分析网络的使用情况及性能，尽早发现网络的瓶颈，便于调整网络的路由，合理分配网络流量，保证网络高效、稳定、可靠地运行。

传统语音业务是时分复用方式的工作机制。因此，一般说来业务流量比较平稳，突发性的流量很少，而且业务流量一般是对称的，即流入流量跟流出流量大致平衡。而目前高速发展的IP数据业务却和传统的语音业务有着很大的不同，IP数据业务是统计复用方式的工作机制，所以业务流量不稳定，突发性大，而且流出流量和流入流量一般来说不平衡，差异性比较大。

近年来，随着Internet呈爆炸性地增长，人们经常会遇到网络拥塞和服务质量低等一系列问题，加强网络管理和改善网络的运行已成为当务之急。而了解网络行为，更多地知道网络流量情况和尽量多的测量信息，无异为重要的手段之一，因此网络流量的测量与分析一直为人们所关注。

国外最早的网络测量始于70年代初[19]，逐渐成熟于80年代，90 年代已渐成体系。在网络测量的方法、工具以及网络基础设施框架和流量的测量模型等方面都做了探索和改进。而我国网络的发展起步较晚，90年测量模型等方面都做了探索和改进。而我国网络的发展起步较晚，90年

代初才引入Internet，大规模的快速发展于90 年代末，近来随着Internet网络的发展，我国已成为世界上Internet用户第二的国家。网络流量的成倍增加，同样需要解决流量的监测、预测和网络规划的问题。我国的一些大的ISP和网络规划及运营者也在进行网络流量测量、网络行为、性能分析这方面的工作，正逐步缩小和国外的差距。

2.2 网络流量分析的目的

一、网络规划

要更好地管理网络和改善网络的运行，网络管理者需要知道其网络的流量情况和尽量多的测量信息。比如，网络管理者可以通过一些工具察看某一网段的负荷情况，等到某一网段的负荷过重时，就可以决定是否在这一网段采用更高速的介质或将这一网段划为两个网段。

但是，网络管理者往往不知道哪些应用导致了流量的增长。如果流量的增长只是因为更多的人在WWW上浏览，那么在这一网段增加代理服务器比把这一网段一分为二更好一些；另一个方面是对网络流量进行长期的监测，通过分析历史趋势，更好地规划网络。

二、基于流量的计费

现在ISP对网络用户提供服务绝大多数还是采用固定租费的形式。这对一般用户和ISP求说，都不是一个好的选择。采取这一形式的很大原因就是网络提供者不能够统计全部用户的准确流量情况。这就需要有方便的手段对用户的流量进行监测。

三、网络应用状况监测与分析

了解网络的应用状况，对研究者和网络提供者都很重要。通过网络应用监测，可以了解网络上各种协议的使用情况(如WWW，pop3，ftp，rtp

等)，以及网络应用的使用情况，研究者可以据此
研究新的协议与应用，网络提供者也可以据此更好地规划网络。

四、网络用户行为监测与分析

这对于网络提供者来说非常重要，通过监测访问网络的用户的行为，可以了解到：

1. 某一段时间有多少用户在访问我的网络
2. 访问我的网络最多的用户是哪些
3. 这些用户停留了多长时间
4. 他们来自什么地方
5. 他们到过我的网络的哪些部分

通过这些信息，网络提供者可以更好地为用户提供服务，从而也获得更大的收益。

2.3 网络流量的应用

一、网络流量可用于校园之中，如学生的宿舍网络异常，网管人员将锁定其IP，并且予以警告，如规劝不听，将予以处分，以免造成网络异常，防止学校服务器挂点。

二、有名木马后门程式，从网络流量也可以发现其异常状况，进而防止被侵入。

三、IP重复进入，可能是骇客侵入的征兆，也可从网络流量得知。

四、伺服器网络管理，限定进入IP最高数目，防止网络拥挤，如：网络游戏每个伺服器都有设最高使用者进入人数，像“天堂”这个游戏，每个伺服器的最高使用者到五千人进入。

2.4 网络流量的测量

Internet流量数据有三种形式：被动数据（指定链路数据）、主动数据（端至端数据）和BGP路由数据，由此涉及到两种测量方法：被动测量方法和主动测量方法[15]。

一、主动测量

主动测量是指在网络上布置测试平台，主动发送测量的流量，从A到B，获得两端点之间的测量结果信息，如发送ICMP包或UDP包等。

主动意味着测量过程中产生新的网络流量。这些流量也许是为了引起网络部件的特殊响应（如：traceroute），也许是为了查看网络为流量提供服务类型的性能（如：treno）。主动测量给网络增加了潜在的荷载负担，特别是如果没有仔细设计使得该方法产生的流量数最小，那么附加的流量会扰乱网络，歪曲分析结果。如：为了测量在IP网络云中瓶颈链路的带宽，定期地向测试路径发送巨大的TCP流量，那么由此产生的附加流量可能会产生Heisenberg效应，而拥塞通过网络云到达这点的路径，并且测量的吞吐量低于瓶颈链路的带宽。

另外，主动测量至少需要多个网络部件某种形式的参与。如：ping命令用于估计主机A到主机B的RTT，需要主机B响应ICMP ECHO请求信息。有几种形式的合作已经广泛应用在Internet上，如：响应ICMP请求和匿名FTP服务器允许主机A和服务器之间进行吞吐量测量，可以将这种合作定义为被动合作。另一种合作方式是主动合作，如果要测量A至B路由的对称性，从B到A和从A到B同样需要进行路由测量，需要B也要同样主动参加测量。

跟踪和可视化Internet拓扑结构是主动测量最主要的应用，CAIDA

国际组织最近开发的skitter动态测量工具可用于动态发现和绘制全球Internet拓扑。同时主动测量技术可以探测网络的特定现象，如发现许多Internet端至端的延迟分布具有重尾特征。Internet的健壮性和可靠性很大程度取决于ISP网络有效可靠的路由，Internet路由行为的分析直接影响下一代网络硬件、软件和操作政策。主动测量还有其它的应用领域：评估IP地址空间的利用率，路由的不对称性和不稳定性，按网络地址前缀长度的流量分布，BGP路由表的空间使用效率，单播和组播路由不一致的程度等。

总之，主动测量的优点是灵活、方便。它是端到端之间的测量，可得到端到端之间的网络性能信息。它的不足：由于需要向网络发送流量，会增加网络负担，对网络性能产生影响。大量的流量可能会在瓶颈处产生拥塞，从而使测量值偏离实际值，有系统误差，即Heisenberg效应，Heisenberg测不准原理。

二、被动测量

被动测量是在网络中的一点收集流量信息，如使用路由器或交互机收集数据或者一个独立的设备被动地监测网络链路的流量。被动测量可以完全取消附加流量和Heisenberg效应，这些优点使人们更愿意使用被动测量技术。有些测度使用被动测量获得相当困难：如决定分组所经过的路由。但被动测量的优点使得决定测量之前应该首先考虑被动测量。如果关心的不是完整的Internet路由，而是AS之间的路由，那么能监测两个对等BGP之间的流量，因为流量中包含全部的AS之间的路由信息。被动测量技术遇到的另一个重要问题是目前提出的要求确保隐私和安全问题。

网络流量是采用大小不一的报文传送，收集到的数据可以进行各种

流量分析，如：流量中各种应用的成分、报文的长度分布、报文到达时间、性能和路径长度等，对这些流量行为的了解能帮助设计下一代互联网设备和体系结构。

网络管理员最感兴趣的被动测量流量是流量的流矩阵，即：有多少流量从一个网络流向另一个网络的表格，这个信息能有助于优化设计决定。不同的流量粒度矩阵有不同的用处，AS粒度流量矩阵有助于优化拓扑结构；一个公司或大学网络管理者为了了解各部门之间流量交换的情况，可以建立系或工作组粒度的流矩阵；国家粒度的流量矩阵有助于了解各国的开放策略和国际商业前景，美国是世界Internet流量主要中转国，71%的其它国家之间的国际流量经过美国。

同时，被动测量还有许多其它应用，包括：识别、刻画和跟踪网页缓冲和代理的优化配置；网络体系结构的安全危害；拥塞控制算法的有效性；流量增长是由于增加了用户还是每个用户流量的增加；流行协议和应用使用的变化；新的技术和协议（如：组播和Ipv6）的渗透力和影响。以上的被动测量应用是Internet流量行为研究的主要内容。

总之，被动测量的优点是：一般不会增加额外的网络流量。但是，被动测量主要用于单点监测，难以进行端至端的行为分析，如路由分析、链接等。改进的方法是：能否将被动测量方法也应用于端至端的性能分析，避免主动测量产生的误差。其中关键的技术是分组标识算法的研究，即对经过这两个监测点的报文进行识别。

有时为了能够从被动收集的数据中提取某些参数可能需要借助于主动测量。另外，被动测量是应该有尽可能低的丢失率，否则测量的数据将难以进行精确估计。

2.5 网络流量的分析

网络流量的分析在网络行为学中起着一个衔接的作用，主要利用网络流量测量部分收集到的各种流量信息，通过运用一些分析和建模方法对其进行分析，以期发现流量的特性，对网络性能做出客观的评价，并以此作为对网络进行控制和优化的依据。流量分析使得人们能够识别网络中现存的问题，并能够找出问题产生的原因。另一方面，使得人们能够识别出将来会发生的一些潜在问题，对网络的性能做出预测，使得网络管理员能够提前查出并解决产生问题的因素，避免网络故障的发生。

流量的分析方法主要有两种，一种是基于测量的流量分析方法，一种是基于模拟仿真的流量分析方法[20]。

一、基于测量的分析

基于测量的分析是对网络数据进行实时测量，然后再对测量的数据进行分析。按照处理时间的不同可以分为在线分析和离线分析两种。

在线分析是从一个局部、详细的角度对流量进行实时分析。它根据客户提出的分析要求进行分析，这里的分析要求可以有范围要求（可以是通过两点之间，某个子网内部以及不同子网之间）、时间要求（可以是某一时刻，也可以是某一时间段）、业务类型要求（Internet 上现有的各种业务，比如 Telnet 服务、FTP 服务等）、分析内容要求（延迟变化、吞吐量变化、丢包率变化、流量变化等）以及显示要求（刷新频率及各种显示方式，如表格方式、折线方式、柱状方式等），最终产生分析结果。

离线分析主要是使用数学分析工具和数学模型等技术和方法对网络流量进行更进一步的分析。数学建模主要分为网络建模（网络设备、通

信链路等）和流量建模。正确的模型可以使得我们能够通过模拟来研究各种模型参数对网络性能的影响，进而提高对网络属性和行为的理解。好的网络模型都是依赖于可参数化的流量模型，这些参数都是从网络测量中得来的。

针对网络流量的特点，研究人员常使用 ARMA(Auto Regressive Moving Average)、网络流量的“分组火车”模型、基于用户行为的模型、小波基模型等[20]。

二、基于模拟仿真的分析

模拟网络行为是指模拟网络流量再实际网络中传输、交换和复用的过程。网络仿真获取的网络特性参数包括网络全局性能统计量、网络节点的性能统计量、网络链路的流量和延迟等，由此既可以获取某些业务层的统计数据，也可以得到协议内部的某些特殊的参数的统计结果。

网络仿真技术有两个显著的特点：首先，网络仿真能够为网络的规划设计提供可靠的定量依据。其次，网络仿真能够验证实际方案或比较多个不同的设计方案。

目前世界上的网络仿真软件可以分为高端和低端两类产品。高端产品一般具有复杂的建模机制、比较完备的模型库、完善的外部接口、强大的功能并能够得到比较可靠的仿真结果。其主流产品基本上都来自美国公司，例如 MIL3 公司的 OPNET、CACI 公司的 COMNET、UC BERKELEY NS 等。低端产品一般只有简单的建模机制、较小的模型库、简单的外部接口，功能单一并且仿真结果的可靠性较差。比较知名的产品也大都来自美国。

第三章　主成分分析法

3.1 概述

3.1.1 什么是 PCA

什么是 PCA? PCA 是 Principal Component Analysis 的缩写，意思是主成分分析。PCA 是一种很有用的统计学的方法，通过突出数据间的相同点和不同点来识别数据内在的特征，并重新表示数据。

PCA 还是一个数学过程，它把一组相关的变量转化成一组数量较少的不相关的变量，转化后的这组不相关的变量就叫做主成分（PC）。其中，第一个主成分包含了原数据尽可能多的信息，接下来的每个主成分都包含了原数据剩余信息中的最大量。并且转化后的每个不相关的变量都能用原变量线性表示，它们可由原数据的协方差矩阵或相关矩阵的特征向量得到。因此，PCA 就是通过提取最少数量的主成分来达到降低维数的目的，而提取的这些主成分能最大限度的包含原数据的特征和最小限度的损失原数据的信息量。

3.1.2 PCA 的优点

我们说 PCA 的目的主要有以下几个方面：

(1). 概述变量间的关系。

(2). 可将原来的变量转换成新的没有相关的变量。

(3). 可用来简化多变量的维数，即降低变量个数，但亦会丧失部分信息。

(4). 可解决回归分析里共线性问题。

(5). 可以用来作为一组变量的综合指标，如物价指数等。

在这篇论文中，我将采用 PCA 主要是为了发现和降低数据的维数，发现高维数据背后更丰富的内涵。因为多维空间通常都很难通过视图表示，最主要的研究手段就是降维，然后分析这些合成的结果，它们反映了原对象多方面的特性。通过二到三个变量组成的视图可以以最小的信息损失量来概括原对象多方面的性质。因为高维空间很难以视图形式表示，因此 PCA 主要用于将高维降至二到三维。

PCA 中主成分的提取原理可以用图 3.1 表示。

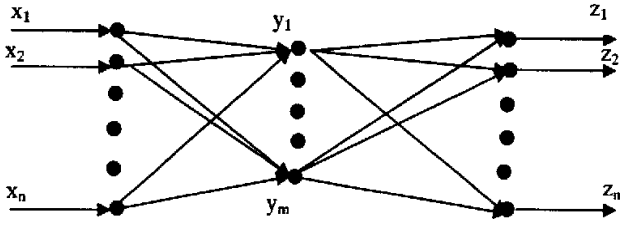

图 3.1 主成分提取原理

图 3.1 所要表达的意思是变量$\{z_1, z_2, \ldots, z_n\}$通过一组不相关的变量$\{y_1, y_2, \ldots, y_m\}$（$m < n$）可以近似地表示原变量$\{x_1, x_2, \ldots, x_n\}$，并且这种表示关系所引起的信息损失量能够保证很小[16]。

PCA 的另一个优点是一旦你找到原数据的内在特征，并对数据进行压缩，例如通过减少数据的维数进行压缩，这样的压缩导致的信息损失量很小。

因此，概括地说，PCA 是一种降低维数和多元分析的技术。它主要应用于数据压缩、图像处理、可视化、数据挖掘分析、模式识别和时间序

列的预测等领域。PCA 的广泛应用主要由于以下三个很重要的特性：

第一，当把一组高维向量压缩为一组低维向量并用这组低维向量重建原向量时，其均方误差表现为理想的线性特性。

第二，这一模型的参数能从原数据直接计算得到，例如将采样数据的协方差矩阵对角化。

第三，给定参数，压缩和解压缩都很容易操作，它们仅需要一些矩阵的运算。

当然，我们在运用 PCA 的时候，要注意如下事项：

1. 当变量间共线性低则无须利用 PCA 作简化。

2. PCA 可以使用协方差矩阵，也可以采用相关矩阵分析，但是当变量间的单位不同或差异较大时，应使用相关矩阵进行分析较佳。

3. 简化多变量的个数时，最重要的是考查丧失部分信息后信息损失的问题。

其中简化多变量的个数可以遵循以下几种方法：

1. 解释的信息比例

如果只取最大的 m 个主成分代替原有的 p 个变量，则这 m 个主成分解释的信息比例为：

$$R=\frac{\lambda_1+\lambda_2+\cdots+\lambda_m}{\lambda_1+\lambda_2+\cdots+\lambda_p}$$

一般以能解释原有变量信息达 70% 以上为原则。

2. 陡坡图(Scree plot)

由特征值对特征值的总和（即特征值的个数）画图，找到开始平坦的点，即为所求个数。

3. 还有其他多种统计上正式分析，但没有标准制式规定的分析统计

量。

这一章将着重介绍 PCA 算法的基本思路，理论推导等相关知识。

3.2 主成分提取的基本思路

设原测量数据有 p 个变量。主成分分析的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样本点集合的重心重合，新坐标系的选取的第一坐标轴与数据变异的最大方向对应，新坐标系的第二坐标轴与第一坐标轴标准正交，并且对应于数据变异的第二大方向……依次类推。这些新坐标轴分别被称为第一主轴、第二主轴……依次类推。如果经舍弃少量的信息后，由主轴 $P_1, P_2, \ldots, P_m$ 构成的子空间能够十分有效的表示原数据的变异情况，则原来的 p 维空间就被降至 m 维，这个新生成的 m 维子空间被称为 m 维主超平面。当 $m=2$ 时，就称其为主平面。可以用原样本点集合在主超平面上的投影来近似地表示原样本点集合。

原样本点集合在主超平面的第 h 主轴上的投影构成综合变量 $v_h \in R^m$，称为第 h 主成分，$h=1,2,\ldots,m$。若以方差 $Var(v_h)$ 表示第 h 主成分 v_h 所携带的变异信息，则主成分分析的结果是

$$Var(v_1) \geq Var(v_2) \geq \cdots \geq Var(v_m) \geq 0$$

3.3 主成分提取的理论推导

记 X 是一个有 n 个样本点和 p 个变量的数据矩阵。

$$X=(x_{ij})_{n\times p}=[x(1),\cdots,x(p)]$$

为推导方便，且不失一般性，设X是经标准化的（即$E(x_j)=0$，$Var(x_j)=1$）。现要求一个综合变量t_1，t_1是$x(1),\cdots,x(p)$的线性组合，即

$$t_1=Xp_1,\quad \|p_1\|=1$$

要使得t_1能携带最多的原变异信息，即要求t_1的方差取到最大值，t_1的方差为

$$Var(t_1)=\frac{1}{n}\|t_1\|^2=\frac{1}{n}p_1'X'Xp_1=p_1'Vp_1$$

这里，记$V=\frac{1}{n}X'X$是X的协方差矩阵。当X中的变量均是标准化变量时，V就是X的相关系数矩阵。

把上面的问题写成数学表达式，即求优化问题

$$\max_{\|p_1\|=1} p_1'Vp_1 \qquad (1)$$

采用拉格朗日算法求解，记λ_1是拉格朗日系数，令

$$L=p_1'Vp_1-\lambda_1(p_1'p_1-1)$$

对L分别求p_1和λ_1的偏导，并令其为零，有

$$\frac{\partial L}{\partial p_1}=2Vp_1-2\lambda_1p_1=0 \qquad (2)$$

$$\frac{\partial L}{\partial \lambda_1}=-(p_1'p_1-1)=0$$

得

$$Vp_1=\lambda_1p_1$$

由此可知，p_1是V的一个标准化向量，它所对应的特征值是λ_1。而根据目标函数式(1)及式(2)，有

$$Var(t_1) = p_1' V p_1 = p_1' (\lambda_1 p_1) = \lambda_1 p_1' p_1 = \lambda_1$$

所以，欲使t_1的方差达到最大值，p_1所对应的特征根λ_1必定要取最大；换言之，即要求p_1是矩阵V的最大特征根λ_1所对应的标准化特征向量。这里，p_1被称为第一主轴，$t_1 = Xp_1$被称为第一主成分。

类似的，我们可以求第二主轴p_2，p_2与p_1标准正交（$p_2' p_1 = 0$，$\|p_2\|^2 = 1$），第二主成分$t_2 = Xp_2$是携带变异信息第二大的成分，并且$Var(t_2)$仅次于$Var(t_1)$。t_2的方差为：

$$Var(t_2) = \frac{1}{n}\|t_2\|^2 = \frac{1}{n} p_2' X' X p_2 = p_2' V p_2$$

同样，写成求优化问题

$$\max_{\|p_2\|=1} p_2' V p_2$$

$$p_2' p_1 = 0, \quad p_2' p_2 = 1$$

定义拉格朗日函数为

$$L = p_2' V p_2 - \lambda_2 (p_2' p_2 - 1)$$

对L分别求p_2和λ_2的偏导，并令其为零，得

$$V p_2 = \lambda_2 p_2$$

$$p_2' p_2 = 1$$

p_2是矩阵V的标准化特征向量，它所对应的特征根是λ_2，并且

$$\lambda_2 = p_2' V p_2 = Var(t_2)$$

由于有约束 $p_2' p_1 = 0$，因此 λ_2 只能是矩阵 V 的第二大特征值，p_2 是对应于 V 第二大特征值的标准化特征向量。

依次类推，可求得 X 的第 h 主轴 p_h，它是协方差矩阵 V 的第 h 个特征值 λ_h 对应的标准化特征向量。而第 h 主成分 t_h 为

$$t_h = Xp_h$$

$$Var(t_h) = \frac{1}{n} p_h' X' X p_h = \lambda_h$$

由此有，$Var(t_1) \geq Var(t_2) \geq \cdots \geq Var(t_m)$。所以，用数据变异大小来反映数据中的信息，则第一主成分 t_1 携带的信息量最大，t_2 次之……如果抽取了 m 个主成分，这 m 个主成分所携带的信息量总和为

$$\sum_{h=1}^{m} Var(t_h) = \sum_{h=1}^{m} \lambda_h$$

综上所述，X 的第 h 主轴 p_h，它是协方差矩阵 V 的第 h 个特征值 λ_h 所对应的标准化特征向量，又称为负载向量(loading vector)。而第 h 主成分 t_h 为原样本点集合在主超平面的第 h 主轴上的投影构成综合变量，又称为主元、主元得分向量
(Score vector)并且有：

$$Var(t_1) \geq Var(t_2) \geq \cdots \geq Var(t_m)$$

定义如下变量：

PCA分析负载阵为 P，$P = [p_1, p_2, \cdots, p_m]$

PCA分析主成分得分矩阵为T，$T=[t_1,t_2,\cdots,t_m]$

则有，$T=XP$

或$t_h=Xp_h$

另外，可以严格证明，主成分t_i的样本均值等于零，样本方差等于X阵协方差阵V的第i大特征根，并且主成分之间以及主成分所对应的负载向量p_i之间都是正交的。

3.4 PCA算法分析

主成分分析的方法即为找出原有变量的线性组合并使其变异数最大。

1. 算出协方差矩阵S，或相关矩阵R

2. 求S或R的特征值及单位特征向量

3. 将特征值依大小顺序排列，设分别为$\lambda_1,\lambda_2,\ldots,\lambda_p$

4. 求出对应的单位特征向量$v_1,v_2,\ldots,v_p$，其中$v_i'v_i=1$，且$v_i'v_j=0$，则

$$\begin{aligned}
y_1 &= v_1'\cdot x = v_{11}x_1+v_{12}x_2+\cdots+v_{1p}x_p\\
y_2 &= v_2'\cdot x = v_{21}x_1+v_{22}x_2+\cdots+v_{2p}x_p\\
&\vdots \qquad\qquad\qquad \vdots\\
y_p &= v_p'\cdot x = v_{p1}x_1+v_{p2}x_2+\cdots+v_{pp}x_p
\end{aligned}$$

分别称为第一主成分，第二主成分，…，第p主成分。接下来将详

细描述 PCA 算法的各个步骤。

3.4.1 获取数据

为了便于下面的说明，构造下面两个变量[2]：

$$x = (2.5 \quad 0.5 \quad 2.2 \quad 1.9 \quad 3.1 \quad 2.3 \quad 2 \quad 1 \quad 1.5 \quad 1.1)^T$$

$$y = (2.4 \quad 0.7 \quad 2.9 \quad 2.2 \quad 3.0 \quad 2.7 \quad 1.6 \quad 1.1 \quad 1.6 \quad 0.9)^T$$

它们构成矩阵 $d = (x \quad y)$。

这是一组二维数据，之所以选择它作为例子，是为了能够采用图形来说明 PCA 算法每一步的结果。这组数据的图形如图 3.2 所示。

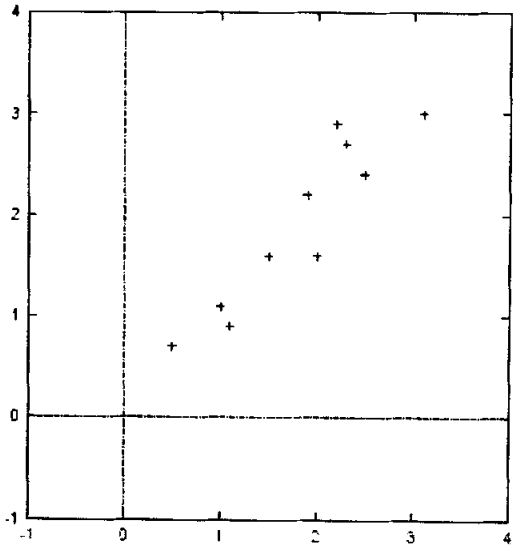

图 3.2 PCA 算法示例图形

3.4.2 数据处理

为了使 PCA 能更有效的工作，必须首先对数据矩阵进行处理，所谓的处理也就是将代表数据各维数的变量减去其对应的均值，即使各变量均值为零。这样在以下步骤的坐标变换中，坐标空间的原点位置将保持不变。

因此在我们的这个示例中，也就是将变量 x 减去其均值 $\bar{x}$，变量 y 减去其均值 $\bar{y}$，处理后的数据如下所示：

$$D=\begin{pmatrix}0.69 & -1.31 & 0.39 & 0.09 & 1.29 & 0.49 & 0.19 & -0.81 & -0.31 & -0.71\\ 0.49 & -1.21 & 0.99 & 0.29 & 1.09 & 0.79 & -0.31 & -0.81 & -0.31 & -1.01\end{pmatrix}^{T}$$

其图形如图 3.3 所示。

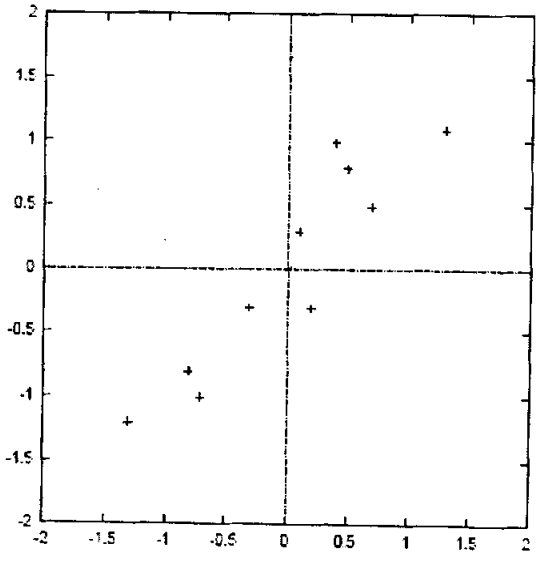

图 3.3 数据处理后的图形

3.4.3 计算协方差矩阵

就是计算代表原数据各维数的变量间的协方差组成的矩阵，即：

$$C=\begin{pmatrix}\mathrm{cov}(x,x) & \mathrm{cov}(x,y)\\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y)\end{pmatrix}$$

我们的数据是二维的，因此其协方差矩阵将是 2x2 维的。其协方差矩阵如下：

$$\mathrm{cov}=\begin{pmatrix}0.616555556 & 0.615444444\\ 0.615444444 & 0.716555556\end{pmatrix}$$

从上面的协方差矩阵可以看出，非对角线上的元素都是正数，这说明变量 x 和 y 是同增长或同减小的。

3.4.4 计算特征值和特征向量

因为协方差矩阵是方阵，因此我们可以计算它的特征值和特征向量。这一点很重要，因为我们可以从中获取反映原数据矩阵的重要的信息。以下是计算得到的特征值和特征向量：

$$\lambda=\begin{pmatrix}0.0490833989\\ 1.28402771\end{pmatrix}$$

$$v=\begin{pmatrix}-0.735178656 & -0.677873399\\ 0.677873399 & -0.735178656\end{pmatrix}$$

注意到，这些特征向量都是单位向量，即它们的模为 1。这一点在 PCA 算法中很重要，许多数学工具也能直接得到单位长度的特征向量的结果。

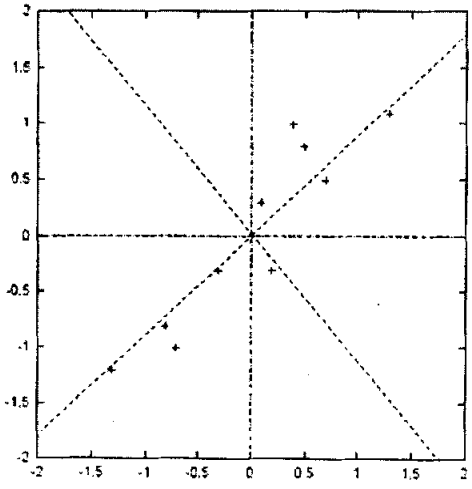

图 3.4 处理后的数据和其特征向量

这些特征向量说明什么呢？让我们来看看图 3.4。

从图 3.4 中我们可以看出特征向量所反映出来的原数据矩阵内在的明显特征。从协方差矩阵可以看出，变量 x 和 y 是同增长或同减小的。在图 3.4 中，我以虚线的形式画出了特征向量的方向。这两个特征向量是相互正交的，且都经过原点，更重要的一点是，它们提供了反映原数据矩阵内在特征的重要信息。我们看到其中一个特征向量穿过原数据的各个点的附近，就像画了一条最匹配的直线，即原数据各个点都分布在这个特征向量的两侧。这个特征向量告诉我们原数据矩阵的两个变量间的关系。另一个特征向量给我们提供的原数据矩阵内在特征的信息量较第一个少。

因此，通过计算协方差矩阵的特征向量，我们就能提取原数据矩阵的内在特征。

3.4.5 选择主成分

这一步是数据压缩和降维的所在。如果你仔细观察了上一步骤得到的特征向量和特征值，你就会发现这些特征值差别很大。事实上，对应的特征值大的特征向量也就是原数据矩阵的主成分。在我们的例子中，对应的特征值较大的那个特征向量也就是穿过原数据点的那条直线的方向。

通常，一旦从协方差矩阵得到特征向量，下一步就是按照它们对应的特征值的大小进行排序，这一步正反映了各个成分的重要程度。现在，只要你喜欢，你可以忽略那些重要性小的成分。这样做你确实损失了一些信息，但是只要那些特征值足够小，你损失的信息量也就足够小。假如你删除了一些成分，那么最后的数据矩阵的维数将小于原数据矩阵的维数。也就是说，假如你原来的数据矩阵是n维，这样你就会得到n个特征值和特征向量，然后你仅选择了前p个特征向量，那么最后的数据矩阵就是p维的。

现在我们需要做的就是选择主成分。这就是要求我们从特征向量列表中选择我们需要留下的那些特征向量，并组成一个新的特征向量矩阵。

结合我们的例子，我们只有2个特征向量，因此我们只有2个选择。我们或者可以将这两个特征向量都保留：

$$V=\begin{pmatrix}-0.677873399 & -0.735178656\\ -0.735178656 & 0.677873399\end{pmatrix} \quad (1)$$

或者删除其中重要性小的成分，仅保留其中一个：

$$V=\begin{pmatrix}-0.677873399\\ -0.735178656\end{pmatrix} \quad (2)$$

我们将在下一步骤中看到这两个选择的结果。

3.4.6 导出新数据矩阵

这是 PCA 算法的最后一步，也是最容易的一步。一旦我们选定了主成分并组成新的特征矩阵，只要将其转置并左乘原处理的数据矩阵，即：

$$D_{final} = V^{'}D^{'}$$

这个式子告诉我们什么呢？它告诉我们原数据可以由这些向量表示。我们原来的数据是 2 维的，即 x 和 y，它们由 x 和 y 构成的坐标空间决定。我们可以由任意两个正交的坐标轴构成的空间来表示这组数据。通过 PCA 变换，我们将原数据从由 x 和 y 构成的空间转变成由这些特征向量构成的空间。假如我们为了降维，如删除了一部分特征向量，则重构的数据只决定于剩余的特征向量构成的空间。

回到我们的例子，上一步骤中，我们选择了两个新的特征向量矩阵，下面我将给出如何由这两个特征矩阵重构原数据，以及两者之间的区别。

由矩阵(1)进行转换，我们得到重构的数据如下：

$$D_1 = \begin{pmatrix} -0.827970186 & -0.175115307 \\ 1.77758033 & 0.142857227 \\ -0.992197494 & 0.384374989 \\ -0.274210416 & 0.130417207 \\ -1.67580142 & -0.209498461 \\ -0.912949103 & 0.175282444 \\ 0.0991904375 & -0.349824698 \\ 1.14457216 & 0.0464172582 \\ 0.438046137 & 0.0177646297 \\ 1.22382056 & -0.162675287 \end{pmatrix}$$

其图形如图 3.5 所示。

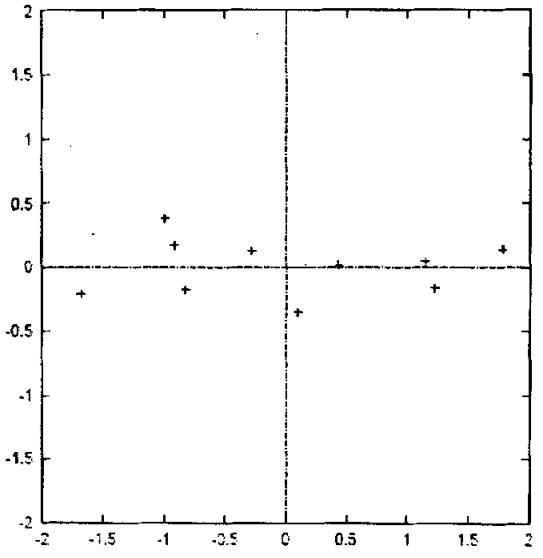

图 3.5 由 2 个特征向量重构的数据

从图 3.5 中可以看出，重构的数据即是原数据，只是进行了坐标轴翻转，两个特征向量变成了原来的 x 和 y 坐标轴。

这一点是很容易理解的，因为在我们采用矩阵(1)进行数据重构的时候，我们并没有删除任何特征向量，因此重构后，也就没有任何信息损失，而仅仅是坐标轴进行了翻转，即两个特征向量变成了原来的 x 和 y 坐标轴。

另一个转换就是采用矩阵(2)，即只保留了第一个特征向量的那个矩阵，那么重构后结果会如何呢？

采用矩阵(2)重构后的数据如下：

$$D_2 = \begin{pmatrix} -0.827970186 \\ 1.77758033 \\ -0.992197494 \\ -0.274210416 \\ -1.67580142 \\ -0.912949103 \\ 0.0991904375 \\ 1.14457216 \\ 0.438046137 \\ 1.22382056 \end{pmatrix}$$

其图形如图 3.6 所示。

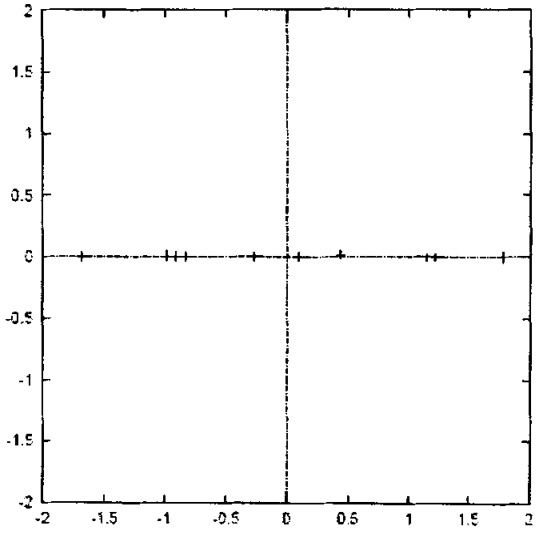

图 3.6 由 1 个特征向量重构的数据

正如我们所预知的那样，它是 1 维的。我们将其和上面由 2 个特征向量重构的数据进行比较，发现它就是上面重构后的数据矩阵的第一行。因此，如果你要用图形表示它，它将是 1 维的，只有 x 轴的坐标。因为我们删除了一个特征向量，因此也删除了一个坐标轴的值。

那么这一步到底有什么用呢？通常，我们通过数据之间的内在特征来表示数据，它们很好的反映了数据之间的关系。这一点很有用，因为这样我们能更好的分类我们的数据。起初，我们通过 x 和 y 坐标来表示数据，这样当然也没问题。但是每一点的 x 和 y 坐标并不能很好的说明这些数据之间的关系。通过转换后，我们就能很容易的看出每一点在新的坐标轴上的位置关系（坐标轴上方或下方）。在使用矩阵(1)，即保留所有特征向量的矩阵进行转换时，我们仅仅将原数据进行了翻转，即只是将原数据从由 x 和 y 构成的空间变换到由特征向量构成的空间。而由矩阵(2)，即只保留了一个特征向量的矩阵进行转换时，我们损失了由被删除的特征向量贡献的信息。

3.5 找回原数据

在使用 PCA 算法进行数据压缩的时候，如何找回原数据显然是最关心的问题。

那么，我们如何找回原数据呢？在进行之前，我们必须明确，在我们的转换过程中，只有将所有的特征向量都包含在特征相量矩阵里，我们才可能精确的得到原数据。如果在转换过程中，我们删除了一部分特征向量，那么找回的数据将会有一定的信息损失。

在上一小节 PCA 算法的步骤中，最后的变换式子如下：

$$D_{final} = V^{'} D^{'}$$

将上式进行变换以得到原数据：

$$D^{'} = \left(V^{'}\right)^{-1} D_{final}$$

然而，当我们将所有的特征向量都包含在特征向量矩阵里时，求特征向量矩阵的逆矩阵就相当于求它的转置矩阵。因为各个特征向量是相互正交的，且模为 1。如此一来，找回原数据的过程得到简化，上面的方程可变为：

$$D' = V^T D_{final}$$

但是，为了得到原来的数据，我们还必须加上进行数据处理时被减去的均值。因此，最后的式子为：

$$d = V^T D_{final} + \bar{d}$$

其中，$\bar{d}$ 是原来数据矩阵中各变量的平均值。

这个式子同样也适用于删除了一些特征向量的情形。因此，即使你删除了一部分特征向量，利用上面的式子也能进行正确的变换。

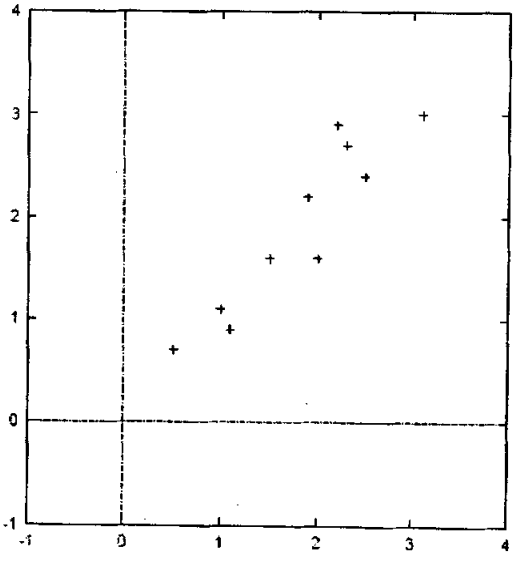

图 3.7 使用 2 个特征向量重构的数据

下面，我将给出由上面两个特征向量矩阵(1)和(2)重构原数据的结果。由矩阵(1)重构的结果如图 3.7 所示。

从图中我们可以发现，这就是原数据的图形。为什么呢？原因很简单，因为我们保留了所有的特征向量，即保留了所有的信息，因此在重构的过程中没有任何信息损失，因此重构后的数据也就是原来的数据。

接着，我将给出删除了一些特征向量后重构的数据的图形，即采用矩阵(2)进行重构，以展示信息如何损失。图 3.8 给出了这一结果。

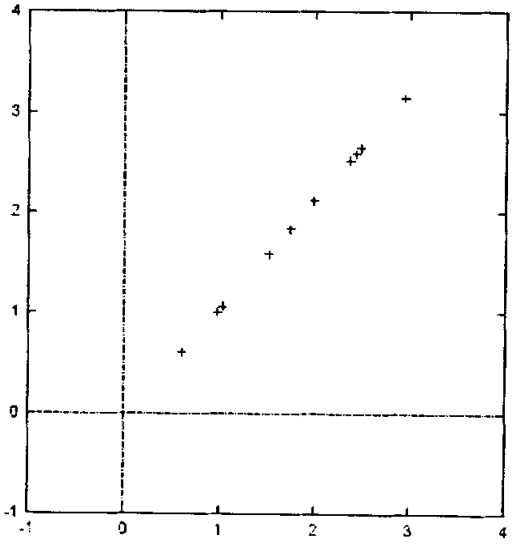

图 3.8 使用 1 个特征向量重构的数据

将其与图 3.7 所示的数据进行比较，你将发现沿着第一个特征向量方向上的信息被保留了，而沿着第二个特征向量方向上的信息被删除了。

第四章　网络流量分析中的 PCA

4.1 概述

在前面一章中已经阐述过，PCA 作为一种降维和多元分析的技术，已经成熟地应用于数据压缩、图像处理、神经网络、可视化、数据挖掘分析、模式识别和时间序列的预测等领域。PCA 的广泛应用主要由于它的三个很重要的特性：

第一，当把一组高维向量压缩为一组低维向量并用这组低维向量重建原向量时，其均方误差表现为理想的线性特性。

第二，这一模型的参数能从原数据直接计算得到，例如将采样数据的协方差距针对角化。

第三，给定参数，压缩和解压缩都很容易操作，它们仅需要一些矩阵的运算。

我们的网络流量也呈现高维多变的特性。因为即使一个中等规模的网络，它的 OD 流也有近百条，而且网络中各流量是相互依赖，即不是相互独立的，而是相互作用的。要研究这样一个网络的流量情况，就应综合地研究这个网络中的所有流量，抽象出来，这就是一个高维对象。既然 PCA 是研究高维对象的一种成熟的技术，为什么我们不能尝试着将其引入流量分析领域呢？

在我们的网络流量分析领域，选用 PCA 的一个原因就是它能将输入的向量空间直接转化成输出空间而不需要在转化的过程中进行人为的控制。比如，许多方法在转换的过程中需要设置一些随机取值的参数，参数取值不同将导致结果或好或坏。但 PCA 却不需要，它只是通过计算输

入向量的前几个最大的主成分来达到降维的目的，而不需要对输入空间作转化。因此，输入数据只是在原空间进行分析，转化的结果是确定性的，并且不依赖于初始条件。

接下来，我将详细地阐述 PCA 算法在网络流量分析领域中的应用。

4.2 PCA 在网络流量分析中的应用

将 PCA 算法运用于网络流量数据中，要实现的目的如图 4.1 所示。

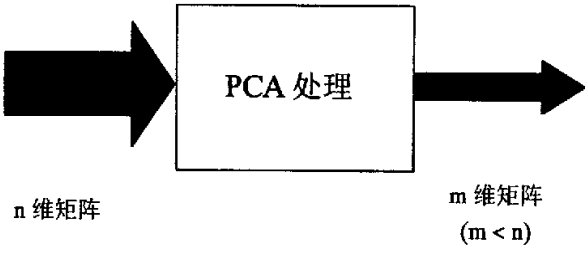

图 4.1 PCA 用于网络流量数据矩阵

在第三章中，我们已经概括地描述了 PCA 算法的一般步骤。由于网络流量分析领域有其自身的特性，因此 PCA 算法也有一些相应的调整，其间也引入了一些网络流量分析领域特有的术语。下面我们将给出具体的步骤。

为了便于下面的讨论，我们将首先引入一些相关的符号变量。让 p 代表我们研究的网络中所有 OD 流的数量，t 代表每条链路采样的点数，在这篇论文中，我们研究的网络有近百条 OD 流，每条 OD 流的数据均持续了相当长一段时间，均采样了数百个点，因此 $t > p$。令 $\boldsymbol{X}$ 为一个 $t \times p$ 的矩阵，它代表一个网络中所有 OD 流的时间序列。这样，矩阵 $\boldsymbol{X}$ 的每列 i

代表第i条OD流的时间序列，而每行j代表网络所有OD流在j时刻的采样值。此外，我们用下标区分矩阵中的每一列，因此第i条OD流用X_i表示。注意到，如此定义的矩阵X的秩最大不超过p。最后，除非特别指出，这篇论文中所有的向量都是列向量。

在数学上，特征提取就是从测量空间R^n到特征空间R^m的映射。映射通常要遵守如下两个准则：

(1). 特征空间必须保留测量空间中的主要信息；

(2). 特征空间的维数应大大低于测量空间的维数。

PCA就是满足上述准则的一种数据压缩方法。

设我们对数据矩阵X进行PCA变换，也就相当于研究矩阵X^TX的特征值问题。其中X^TX是表示各流量的变量间的协方差矩阵。每个主成分v_i就是矩阵X^TX分解后的第i个特征向量，即：

$$X^TXv_i = \lambda_i v_i \qquad i = 1,\ldots,p \tag{1}$$

其中λ_i是v_i对应的特征值。注意到，矩阵X^TX的特征向量是正交的，相应的特征值是非负实数。按惯例，特征向量是单位向量，特征值由大到小排列，即$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$。

由于计算矩阵X的主成分也就相当于计算X^TX的特征向量，因此我们来考虑第一主成分。令p维向量v_1代表矩阵X的第一主成分。正如前所述，第一主轴v_1捕获了原测量数据最大的能量[1]：

$$v_1 = \arg\max_{\|v\|=1} \|Xv\|$$

其中$\|Xv\|$表示原测量数据沿着v轴上的能量。上面的公式也可以写为如下的形式：

$$\begin{aligned} v_1 &= \arg\max_{|v|=1} \|Xv\| \\ &= \arg\max_{v} \frac{\|Xv\|}{v^T v} \\ &= \arg\max_{v} \frac{v^T X^T X v}{v^T v} \end{aligned}$$

依此递推，一旦前$k-1$个主成分确定后，第k个主成分就对应于剩余能量中最大的方向。剩余能量就是原数据的减去前$k-1$个主成分的能量后剩余的能量。这样，我们将第k个主成分v_k表示如下：

$$v_k = \arg\max_{|v|=1} \left\| \left(X - \sum_{i=1}^{k-1} X v_i v_i^T \right) v \right\|$$

相同的原因，由于计算第k个主成分相当于寻找矩阵$X^T X$的第k个特征向量，因此，计算所有的主成分组成的集合$\{v_i\}_{i=1}^{p}$就相当于计算矩阵$X^T X$的所有特征向量。

经过如此计算得到的主成分$v_1, v_2, \ldots, v_p$具有如下几个性质：

(1). 各主成分间互不相关，即对任意的i和j，v_i和v_j的相关系数

$$Corr(v_i, v_j) = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

(2). 组合系数$(a_{i1}, a_{i2}, \ldots, a_{ip})$构成的向量

$$v = a_{i1} v_1 + a_{i2} v_2 + \cdots + a_{ip} v_p$$

也是单位向量。

(3). 各主成分的方差是依次递减的，即

$$Var(v_1) \geq Var(v_2) \geq \cdots \geq Var(v_p)$$

(4). 总方差不增不减，即

$$\begin{aligned} & Var(v_1)+Var(v_2)+\cdots+Var(v_p) \\ =& Var(c_1)+Var(c_2)+\cdots+Var(c_p) \\ =& p \end{aligned}$$

这一性质说明，主成分是原变量的线性组合，是对原变量信息的一种改组，主成分不增加总信息量，也不减少总信息量。

(5). 主成分 $v_1, v_2, \ldots, v_p$ 是矩阵 $X^T X$ 的特征向量，而且，特征值 λ_i 就是第 i 个主成分 v_i 的方差，即

$$Var(v_i) = \lambda_i$$

并且有

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

一旦将原测量数据映射到主成分空间，就很容易分析原数据在某个方向上的变化。考察映射到主成分上的数据，我们看到主轴 i 上的数据分布可以由 Xv_i 给出。因此，对于每个主轴 i，我们可得到：

$$u_i = Xv_i \qquad i = 1, \ldots, p$$

其中 u_i 是 t 维向量，且相互正交。上面的公式说明：以 v_i 为系数，重建了原网络流量数据中所有 OD 流在某个方向上的特征。因此，向量 u_i 捕获了原网络流量数据中所有 OD 流在主轴 i 上的特征信息。因为主轴是按照能量贡献率的大小排列的，因此向量 u_1 捕获了原网络流量数据中所有

OD 流最大的特征信息，u_2 提取了次大的特征信息，依此类推。正因为 $\{u_i\}_{i=1}^{p}$ 能捕获原网络流量数据中所有 OD 流中共同的时变趋势，我们将其定义为原数据矩阵 X 的特征流。

主成分集合 $\{v_i\}_{i=1}^{p}$ 依照列的次序构成主成分矩阵 V，其为 $p \times p$ 型。同样，我们也能够以 u_i 为列构成 $t \times p$ 型矩阵 U。结合 V 和 U，我们可以得到每条 OD 流 X_i：

$$X_i = U\left(V^T\right)_i \qquad i = 1,\ldots,p$$

其中 X_i 是第 i 条 OD 流的时间序列，$\left(V^T\right)_i$ 是矩阵 V 的第 i 行。上面的公式表明：每条 OD 流 X_i 都可以表示成特征流 u_i 的线性组合，组合系数为 $\left(V^T\right)_i$。

图 4.2 和图 4.3 显示的是分解后的某个特征流 u_i 和它对应的主成分 v_i 的典型例子。从图中可以看出，特征流 u_i 捕获了网络流量中的某一时变特性，而主成分 v_i 则确定了这一时变所发生的方向。

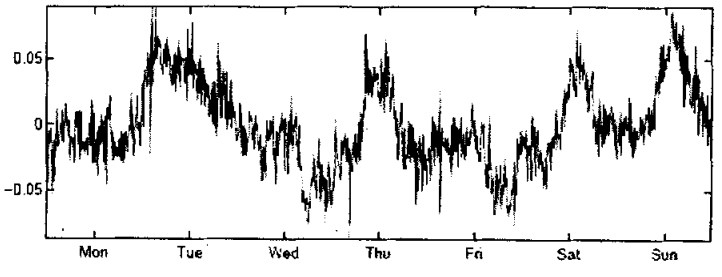

图 4.2 分解后的一个特征流

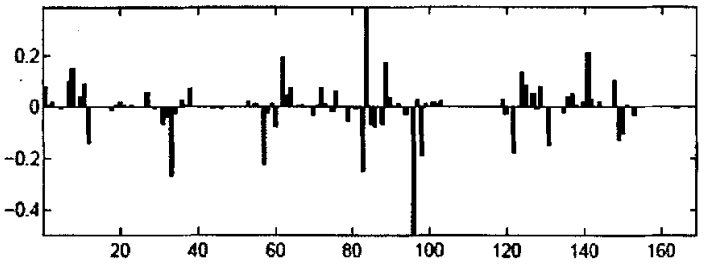

图 4.3 特征流相应的主成分

我们再来考察集合$\{\lambda_i\}_{i=1}^{p}$的性质。λ_i是特征值，同时也隐含了相应的主成分的能量信息：

$$\|Xv_i\| = v_i^T X^T X v_i = \lambda_i v_i^T v_i = \lambda_i$$

其中，第二个等式从方程(1)中导出，最后一个等式遵循v_i是单位向量的事实。因此，λ_i非常有用，它是关系降维的重要因素之一，并且可以很容易的通过图形的方式表示。特别地，我们发现，如果在p个特征值中只有r个λ_i是不可忽略的，那么这意味着原数据矩阵X是依赖于原p维空间的r维子空间的。由此，我们可以将原数据矩阵X近似为：

$$X' \approx \sum_{i=1}^{r} u_i v_i^T$$

其中$r < p$，是原数据矩阵X的有效内在维数。

以上我详细地阐述了 PCA 算法用于网络流量分析领域的具体步骤。需要指出的是，这一想法是 Anukool Lakhina，Mark Crovella，Christophe Diot 等人提出的[1,6,7]。在下一章中，我将给出这一算法用于一组真实的流量数据的结果，并给出采用 Matlab 工具运行的结果。

在进行下一章前，我将利用一小节的篇幅描述一个很重要的概念：OD 流。

4.3 OD 流

研究全网络流量时，我们必须认识到，一个网络中不同链路的流量不是相互独立的，而这个网络的流量事实上是由一个 OD 流集合和一个路由矩阵决定的。

那么，什么是 OD 流呢？所谓 OD(Origin-Destination)流，就是一些流量的集合，这些流量从网络的一个节点（称入口点）进入网络，从一个节点（称出口点）离开网络。这些由路由决定的点到点之间的流量的重叠，就构成了整个网络中所有链路的流量。因此，研究网络流量的一个更直接、更基本的方法不是研究所以链路的流量，而是研究网络中的 OD 流集合。

我们平时接触最多的是链路流，比较容易获取的数据也是链路流的数据。那么，链路流和 OD 流之间的关系如何呢？它们之间通过路由矩阵 A 关联。矩阵 A 的大小是链路流×OD 流，其中，如果 OD 流 j 经过链路 i，则 $A_{ij}=1$，否则 $A_{ij}=0$。因此，OD 流向量(x)和链路向量(y)之间存在如下关系：

$$y = Ax$$

流量工程就是调整路由矩阵 A 的过程，也就是给定一些 OD 流 x，在一定的程度上影响链路流 y。因此，精确的流量工程和链路容量规划等领域都依赖于更好的理解 OD 流的特性。

一个典型的具有n个节点的网络，将包含n^2个OD流。因此，即使在一个中等规模的网络中，如具有数十个节点的网络，将包含上百甚至上千个OD流。这意味着向量y是属于一个高维空间的。一系列的OD流组成的矩阵X将是一个高维多变的时间序列矩阵。

第五章　分析网络流量数据

5.1 概述

这篇论文中采用PCA算法对OD流的分析是基于一个真实的网络的数据。这组数据是实验室从中国最大的 IP 网络的骨干链路上获取的MRTG(Multi-Router Traffic Grapher)数据。由于保密的原因，在此不便公布这个网络的真实名字。MRTG 是一种基于 SNMP 的监控网络链路流量负载的工具，它能记录一个路由器各个接口流入和流出的流量信息。从MRTG数据中，以1800秒（即半小时）的间隔滤出连续10天的记录，即从2001年4月28日00:00到2001年5月7日24:00。这些是链路流的数据，然后实验室利用路由矩阵 A 将其转换为OD流矩阵。因此，这个网络共有63条OD流，而每条OD流均采样了481个点，因此PCA算法研究的数据矩阵 X 是一个 481×63 的矩阵。这是一个高维的数据矩阵，在采用PCA算法进行处理前，我们只能用图形孤立的显示每条OD流的状况。在前面的章节已经强调过每条OD流的流量不是相互独立的，因此这种图形显示并没有多大的意义。

采用PCA算法分析这组链路流量，将得到一个 63×63 的主成分矩阵 V 和一个 481×63 的特征流矩阵 U。这一处理是在Matlab中完成的。下面，我将给出PCA算法作用于网络流量数据的仿真结果，并进行详细的讨论。

5.2 仿真结果及讨论

正如前面的章节说明的那样，这篇论文的基础就是采用PCA算法将

一个网络中的所有 OD 流分解成相应的主成分和特征流。在这一章中，我将给出这一处理的结果。首先，我们将看到一部分特征流就足够很好的重建 OD 流了——也就是说，OD 流呈现低维性质。然后，我将研究每条 OD 流的组成，也就是说，每条 OD 流究竟是由那些特征流组成的。这一问题又引出对特征流的考察上来。最后，我又探讨了 PCA 对 OD 流的分解在时域上的相对稳定性问题。

5.2.1 OD 流的低维特性及重构

PCA 算法运用于网络流量数据的主要作用就是降维。降维也就是要删除一部分维数，即删除一部分信息。在前面的章节中已经描述过，特征值 λ 的一个重要性质就是反映了对应的特征向量，即主成分 v 所携带的信息，即

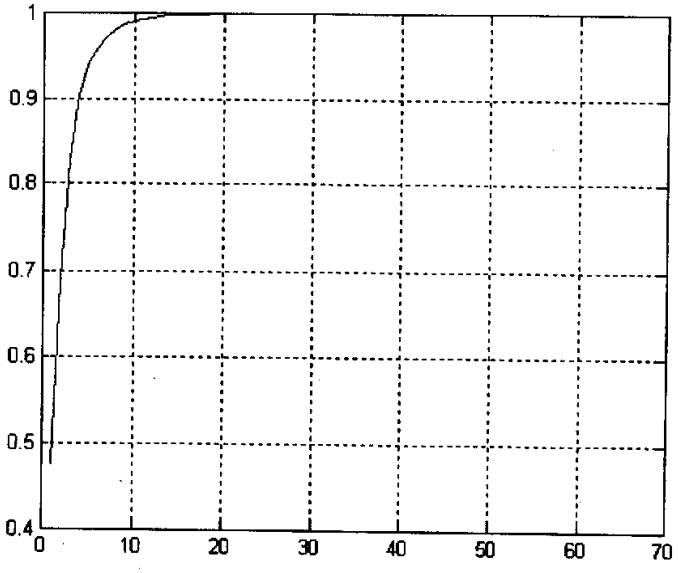

图 5.1 主成分的累积贡献率

$$Var(v_i) = \lambda_i$$

因此，我们可以通过 λ 的比重来直观的显示每个主成分的信息贡献率。图 5.1 显示的是主成分的累积信息贡献率。

图 5.1 显示的结果是，仅仅前几个主成分就贡献了流量的大部分信息。我们看到，曲线的坡度很大，这表明一部分主成分，在 3 到 7 维之间，贡献了大部分的流量信息。从另一个角度来说，这一结果也显示了整个 OD 流的时间序列的内在有效维数是 3 到 7 维——大大低于 OD 流的数量。

为了更形象的说明 OD 流的低维性质，我将给出低维重构 OD 流的图例。我仅仅通过采用几个主成分来达到重构 OD 流的目的。这一重构是根

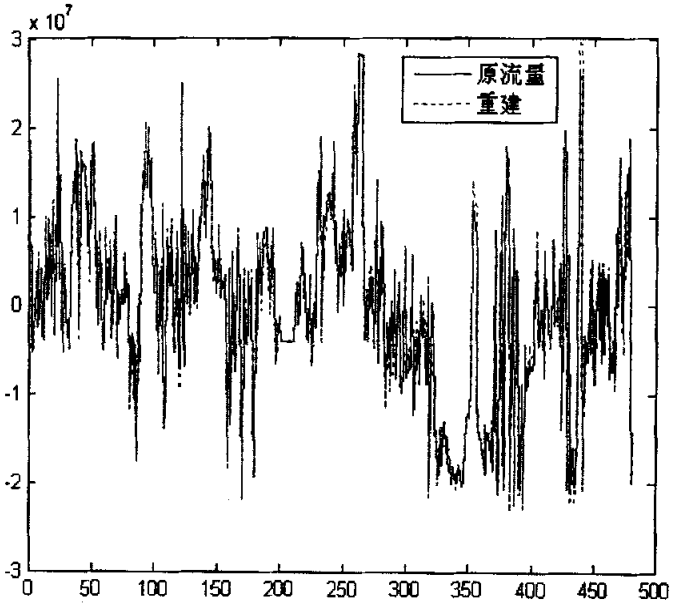

图 5.2(a) 采用 5 个主成分重建的 OD 流 1

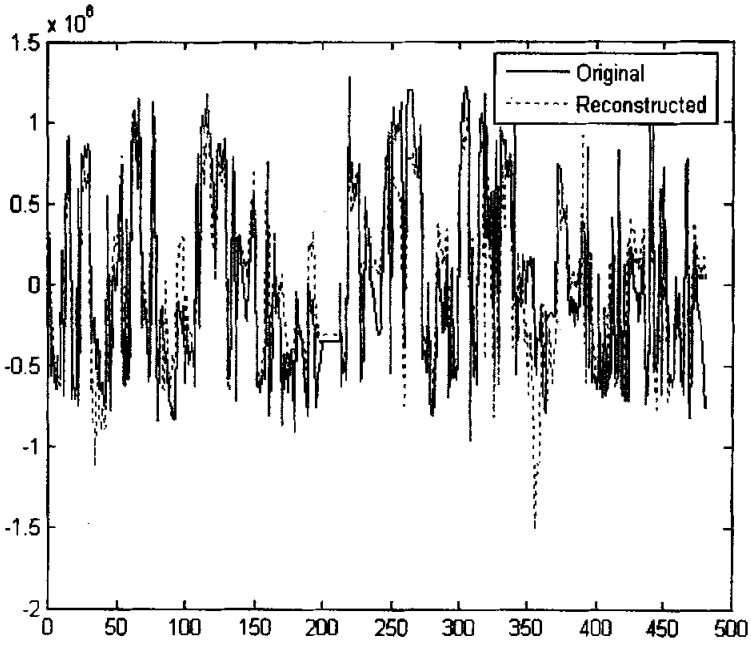

图 5.2(b) 采用 8 个主成分重建的 OD 流 33

据上一章中的公式进行的：

$$X' \approx \sum_i u_i v_i^T$$

我从所有的 OD 流中选取了两条进行重构，结果如图 5.2 所示。图 5.2 表明，即使从原数据中删除了近 60 维数，还是能够很好的重构这些 OD 流。不过，仔细观察图 5.2(a)和(b)可以发现，OD 流 1 采用 5 个主成分重构的误差比 OD 流 33 采用 8 个主成分重构的误差要小。为什么会产生这样的结果呢？因为 OD 流 1 中的特征分布比较集中，而 OD 流 33 的特征分布相对比较均匀，因此 OD 流 33 中 8 个主成分所占的信息比重也赶不上 OD 流 1 中 5 个主成分所占的信息比重。这也暗示着各条 OD 流间的差别还是很大的。这将在下一小节具体介绍。

那么，OD 流的这一低维性质是如何产生的呢？至少有两方面的原因导致这一低维性质。首先，在原数据中，如果各维数间的差异非常大，那么数据就可能存在低维性质。这也就是说，原数据中一部分维数的方差占据了很大的比重。其次，如果这些时间序列在维数上有共同的趋势或特征，那么它们将可能存在低维特性——也就是说，如果维数间存在不可忽视的相关性，那么就存在这一低维性质。

5.2.2 OD 流的结构

下面我们来研究每条 OD 流具体由哪些特征流组成的，也就是研究上一章的公式：

$$X_i = U\left(V^T\right)_i \qquad i = 1,\ldots,p$$

前面提到，这一公式表明：每条 OD 流 X_i 都可以表示成特征流的线性组合，组合系数即为 $\left(V^T\right)_i$。也就是说：主成分矩阵 V 的第 i 行即表明各个特征流对第 i 条 OD 流的贡献率。这一点很重要，这样我们就可以通过矩阵 V 的每行来研究 OD 流的组成——每条 OD 流是由哪些特征流组成的，两条 OD 流可以通过转化成特征流的表示找出相似点和不同点。

观察矩阵 V 的每行，我们将发现在特征流如何构成 OD 流方面的一些结果。首先，我们将看到每个 OD 流都仅仅由一小部分特征流组成。

我们来考察矩阵 V 的任意一行。我们关心的是这一行中到底有多少个值是显著非零的。因此，我们可以设置一个门限值来计算这一行中有多少个值的绝对值大于此门限值。理论上，通常此门限值取为 $1/\sqrt{p}$，其中 p 是代表我们研究的网络中所有OD流的数量。将矩阵 V 的每一行减去

这一门限值$1/\sqrt{p}$。注意，在运算前，保证矩阵V的每一列是单位向量。图5.3就是主成分矩阵V的每个元素与门限值$1/\sqrt{p}$比较的结果。

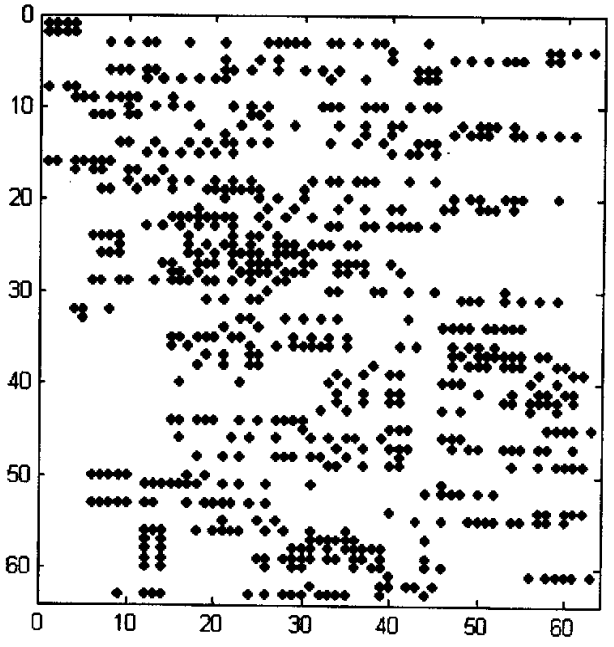

图5.3 主成分矩阵V各元素与门限值的比较

图5.3中，圆点是代表矩阵V中绝对值大于此门限值的元素。我们看到，处理后矩阵V的大多数行都在10个点左右。对于OD流来说，这意味着任意一条OD流都由10个左右的特征流组成，通常更少。这一结果也意味着我们可以认为每条OD流都有相应的一部分特征。这样，我们可以通过这些特征来研究每条OD流的性质。

其次，我们来研究OD流间的差异。强调一点，图5.3中，特征流的索引按照对应的特征值的大小排序。我们看到，图5.3中，有的OD流主要由前几个特征流组成，有的OD流主要由中间部分的特征流组成，而有的OD

流却主要由后面部分的特征流组成。将图5.3矩阵各行的顺序调整一下，这一结论将看得更清楚。处理后的结果如图5.4所示。

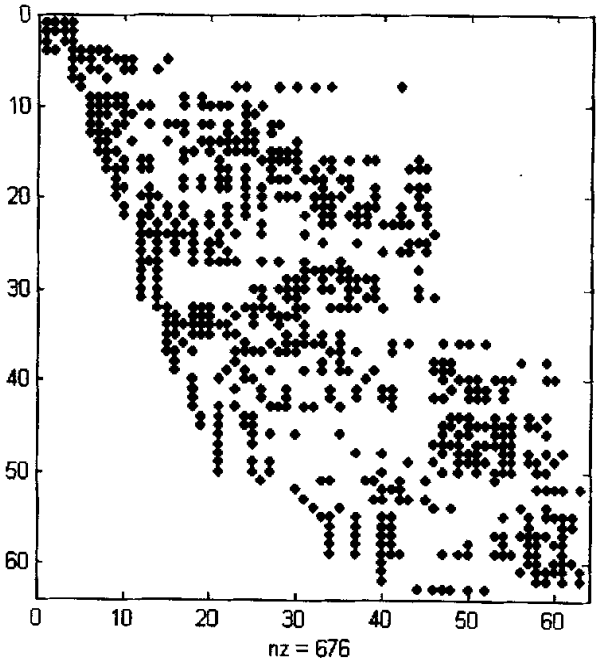

图5.4 处理后的结果

那么，这一现象到底说明什么呢？OD流和特征流之间到底存在什么关系呢？在第一章中，我曾经提到，特征流可以分为三类：确定型的特征流、脉冲型的特征流和噪声型的特征流。那么图5.4中，各OD流的组成有如此分明的关系，这是否同特征流的这一分类有关呢？要回答这一问题，让我们接下来看下一小节的讨论。

5.2.3 特征流的研究

特征流的研究在理解OD流的性质上起着非常重要的作用。现在我们

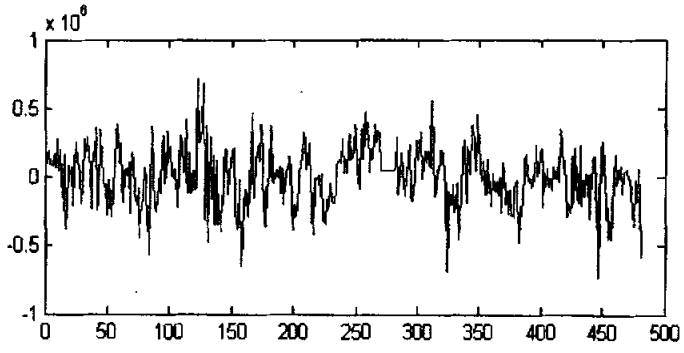

图 5.5(c) 噪声型特征流

图 5.5(a)所示的特征流呈现很强的周期性质，它反映网络每天的行为。正因为这一性质，故称它为确定型的特征流。

图 5.5(b)所示的特征流呈现一些剧烈的，生存时间很短的脉冲。这些脉冲型的特征流呈现的脉冲的幅值已经大于特征流均值 4 到 5 个标准方差。这些脉冲清晰的捕获了网络中流量激增或骤减的时刻。

图 5.5(c)所示的特征流呈现平稳的，类似高斯噪声的性质。这些噪声型的特征流反映了网络的一般行为。通常大多数的特征流都属于这一类别。

根据上面的分析，我们可以看出，特征流按照这种分类方式表现出截然不同的三个方面的特性。然而，需要说明一点，目前特征流的这一分类方法没有太多理论上的依据，也就是说，没有一个明确的划分标准来准确判断每一个特征流究竟属于哪一类特征流。目前仅仅是从特征流的时间序列的图形中判别其归属于哪一类别。如脉冲型的特征流，它表现为存在至少一个尖脉冲，并且脉冲的幅值大于特征流均值 4 到 5 个标准方差；噪声型的特征流类似于高斯白噪声，幅值变化范围不大；确定型的特征流呈现一定的周期趋势。

转向特征流，研究这三类常见的特征流，并描述如何通过研究这三种类型的特征流来理解上一小节中链路流的组成同特征流的分类之间的关系。

首先，我们来认识这三种类型的特征流。来观察所有的特征流，我们将发现，所有的特征流几乎都分属于三种不同的类别。各个类别中有代表性的特征流如图 5.5 所示。

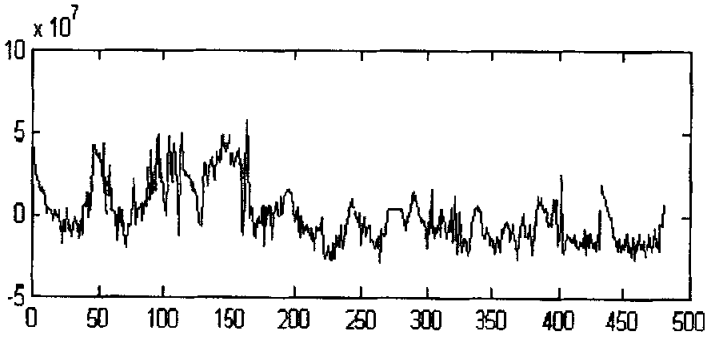

图 5.5(a) 确定型特征流

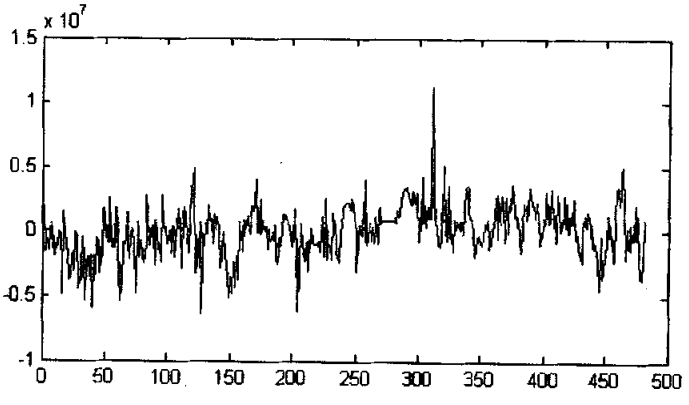

图 5.5(b) 脉冲型特征流

因此正如上面所说，并不是任何一个特征流都非得按照这种分类方法确定无疑地归为其中的一类。但是，这种分类方法事实上还是十分有效的，在实验过程中，将看到几乎所有的特征流都可以很容易地按这种分类方法归为其中的一类。这样一来，结合上一小节中 OD 流的组成和特征流的关系，我们将能够更好的理解 OD 流的性质。

采用这一分类方法的一个明显的好处是它能将任何 OD 流分解成这三类不同特性的特征流。也就是说，我们能够按照这三种类型的特征流：确定型的特征流、脉冲型的特征流和噪声型的特征流重构每条 OD 流。当我们如此处理后，每一种类型的特征流分别捕获了 OD 流中不同的特征：它的确定性的趋势，它的呈现脉冲尖峰的时刻以及它的相对稳定的随机组成成分。通过这一分类，我们从大容量的流量中提取的脉冲型的和噪声型的信息十分清晰，而且，从背景噪声中提取的脉冲型的信息也十分清晰。注意到，我们仅仅通过变换和简单的特征流分类方法，而没有采用建模就达到了这一目的。

既然这三类特征流如此重要，那么我们来研究 OD 流分解成这三类特征流后，它们之间的相互关系如何。首先，我们来观察当特征流按照它们的重要性（如对应的特征值大小）排列的时候，不同类型的特征流排列的位置有什么不同。实验结果表明，确定型的特征流往往出现在前几个特征流中，余下的特征流就分别属于脉冲型的特征流和噪声型的特征流。具体结果见表 5.1。

表 5.1 显示，确定型的特征流往往出现在前面几个，因为它携带的原流量的信息量最大。而脉冲型的特征流和噪声型的特征流携带的原流量的信息量较小，因此出现在其后，然而，这组数据的处理结果中，脉冲型的特征流和噪声型的特征流的分布界限不是很明了。至于是何种原

因，还有待进一步研究。

特征流类型	分布（特征流索引）
确定型特征流	1，2，4
脉冲型特征流	3，15，21，22，23，24，27，28，31，33，34，38，45，46，47，48，49，50，51，55，56，57，58，60，61
噪声型特征流	5，6，7，8，9，10，11，12，13，14，16，17，18，19，20，25，26，29，30，32，35，36，37，39，40，41，42，43，44，52，53，54，59，62，63

表 5.1 各类型的特征流的分布

从上面不同特征流的归属，我们得知原流量中最重要的信息来源于由周期趋势导致的不平稳的流量。去除这些周期趋势，激增或骤减的呈现脉冲形式的流量占据了次重要的地位。最后，对流量最不重要的贡献来自于噪声。这一结论从表 5.2 中以数据的形式更直观的表示。表 5.2 显示了每种类型的特征流的能量贡献率。

特征流类型	能量贡献率
确定型特征流	76.64%
脉冲型特征流	14.02%
噪声型特征流	9.34%

表 5.2 各类型的特征流的贡献率

将特征流按照这种方法进行分类的另一个好处是我们能将脉冲型的

流量从大容量的流量和背景噪声中提取出来。这一点是非常有用的，因为，脉冲型的流量就是流量激增或骤减的时刻，流量的激增或骤减往往预示着一些异常情况的产生，而这很可能同网络容量方面的蓄意攻击有关。图 5.6 显示的是某条 OD 流和与其对应的合成的脉冲型特征流。

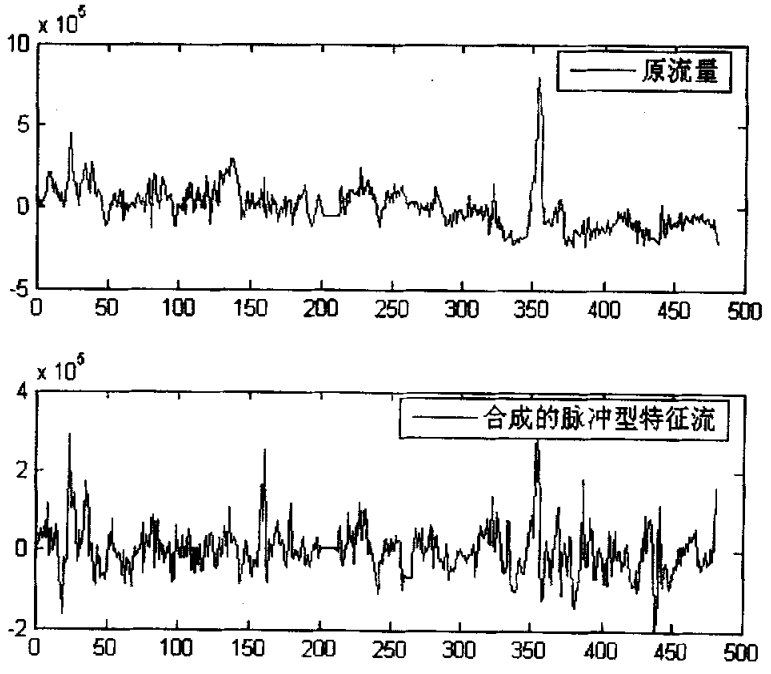

图 5.6 OD 流和对应的合成后的脉冲型特征流

在图 5.6 中，我利用表 5.1 中属于脉冲型特征流的u_i和如下公式：

$$X' \sim \sum_i u_i v_i^T$$

合成了原 OD 流中的脉冲型的成分。让我们将两者进行对照，原流量中第 353 点处的流量激增从原时间序列中就可以直接观察到，而合成的脉冲型成分在这一点处也显示了尖脉冲，两者吻合。而合成的脉冲型成分中

第24点和第160点显示的两个幅度稍小的尖脉冲就不容易从对应的链路流量的时间序列中直接观察得到，而这两处往往也意味着流量方面的异常。经过验证，这两处也存在不同程度的流量激增。

因此，脉冲型的特征流对于网络流量突发点的检测很有帮助，能够对网络容量方面的异常攻击锁定检测范围，及时发现并阻止这一类型的攻击。至于确定是何种类型的攻击，还需要分析包头信息，这一点不在这篇论文的讨论范围中。

5.2.4 OD流结构的相对稳定性

前面的讨论表明，PCA算法在研究OD流的内在结构方面很有成效。但是，在许多实际的应用场合中，研究下一时段网络流量的趋势也很重要。

这一小节讨论的内容是将OD流分解成各个主成分对应的特征流对于分析下一时段的流量，或者说不是这一时段采用PCA算法进行处理的输入时是否同样有效。

假设给定时间段$[t_0, t_1)$上的OD流，采用PCA算法对其进行处理后得到主成分集合$\{v_i\}$。接下来，在某一时刻t_2，$t_2 > t_1$，采用刚刚得到的主成分集合$\{v_i\}$将这一时段新的OD流分解成特征流。那么，这一分解是否仍然可行呢？也就是说这一分解是否仍然保持了低维特性？由于高维空间的限制，不能将这个问题回答得很深入，下面我将给出一些初步的实验结论。

通常检验OD流的低维特性的一个途径就是计算采用一部分维数近似原流量时引起的误差的大小。采用两段时间上连续的OD流X_1和X_2，

我将对数据 X_1 采用 PCA 算法进行处理并得到其相应的主成分集合 $\{v_i\}$。接着，利用主成分集合 $\{v_i\}$ 得到数据 X_1 的特征流，再利用 $\{v_i\}$ 以同样的方式得到数据 X_2 的伪特征流。之所以称其为伪特征流，是因为它是采用数据 X_1 的主成分集合 $\{v_i\}$ 对数据 X_2 进行处理后得到的，而不是对数据 X_2 直接运用 PCA 算法进行处理得到的，但是它们可能包含了数据 X_2 的真正的特征流的性质。然后，我利用这些特征流来重建数据 X_1，得到重建后的数据，将其记为 X_1'。再利用这些伪特征流来重建数据 X_2，得到重建后的数据，将其记为 X_2'。

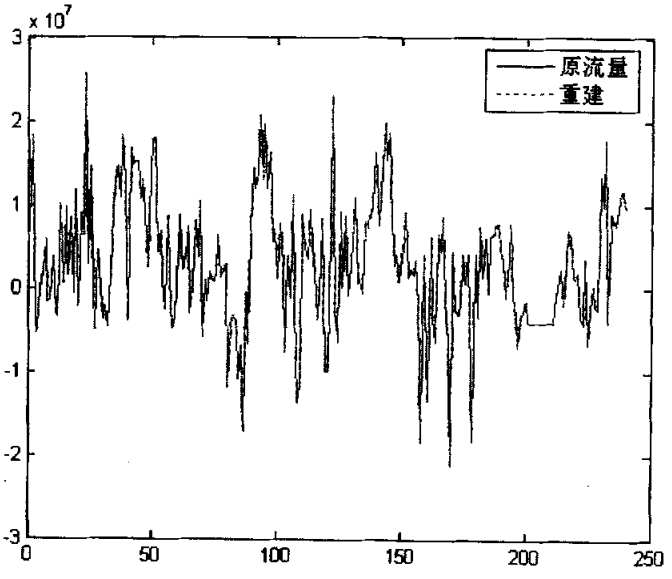

图 5.7(a) 对数据 X_1 进行重构

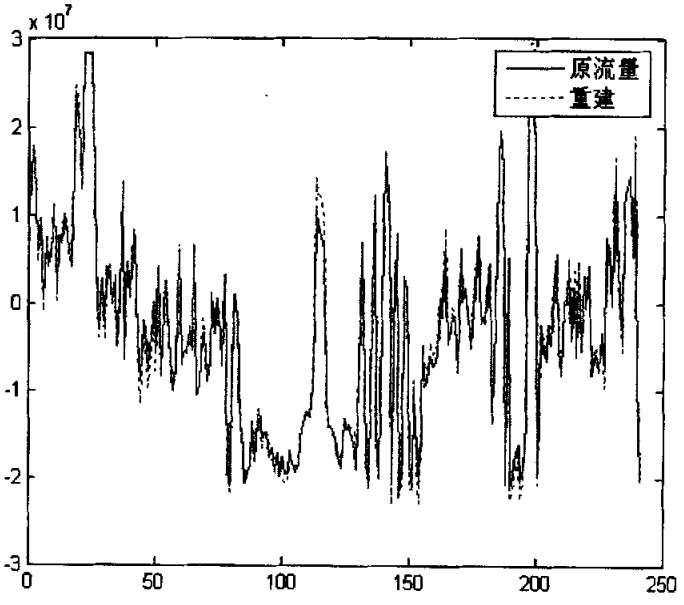

图 5.7(b) 对数据 X_2 进行重构

由于手中的数据资源有限，因此我将原 481 点的数据分成两部分：前 240 点的数据记为 X_1，后 241 点的数据记为 X_2。那么，采用上面的方法分别对数据 X_1 和 X_2 进行重建的结果如图 5.7 所示。

接下来，再计算这两个重构产生的相对误差，分别令其为 R_1 和 R_2：

$$R_1 = \frac{|X_1 - X_1'|}{X_1}$$

$$R_2 = \frac{|X_2 - X_2'|}{X_2}$$

从 5.2.1 小节的讨论得知，重构数据 X_1 产生的误差应该小一些，因为 OD 流的低维性质使其能够由一小部分特征流较为精确的近似。而重构

数据 X_2 产生的误差应该比重构数据 X_1 产生的误差大一些，因为重构采用的主成分集合 $\{v_i\}$ 并不是数据 X_2 真正的主成分集合。但重构数据 X_2 产生的误差究竟比重构数据 X_1 产生的误差大多少呢？

平均地，对于每条 OD 流，我将根据图 5.3，采用 10 个左右的主成分来进行重构。那么，重构数据 X_1 和 X_2 产生的相对误差 R_1 和 R_2 的结果如图 5.8 所示。

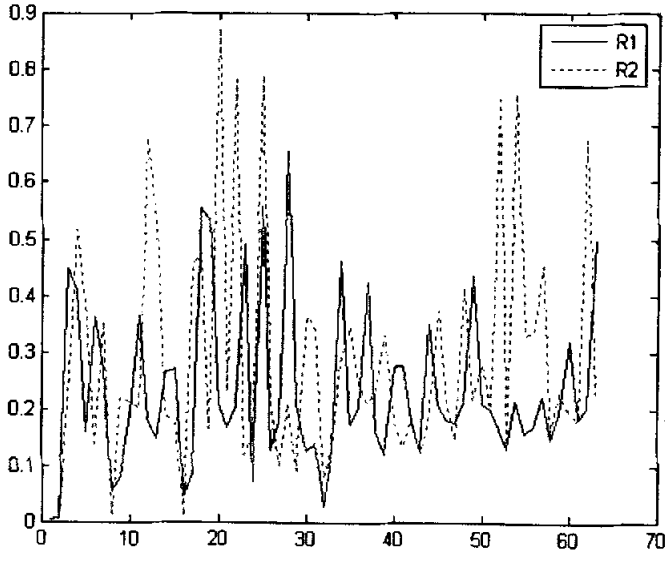

图 5.8 相对误差 R_1 和 R_2

图 5.8 中，蓝色的实线是重构数据 X_1 产生的相对误差 R_1 的值，红色的虚线是采用数据 X_1 的主成分集合 $\{v_i\}$ 重构数据 X_2 产生的相对误差 R_2 的值。由图中得出，相对误差 R_1 的均值为 23.33%，而相对误差 R_2 的均值

为 28.67%。并且总体上重构数据 X_1 产生的相对误差 R_1 比采用数据 X_1 的主成分重构数据 X_2 产生的相对误差 R_2 小。可以看到，这一结果还是可以接受的。

这一结果显示前一时段的流量分解成的特征流对于研究下一时段的流量仍然提供了足够的有用信息，至于这一有效性在时间域上的作用范围有多大，还有待进一步研究，但是上面显示的结果还是较为理想的。

5.3 小结

以上，通过对网络流量数据采用 PCA 算法进行处理后，我研究了 OD 流的低维特性并对其进行了低维重构，通过对特征流的研究来认识 OD 流的内在结构，最后研究了 PCA 算法的处理结果在时域上的稳定性。

因此，从以上结果来看，将 PCA 算法引入网络流量分析领域还是可行的。但是，就目前的研究来看，它只能分析网络流量宏观上的低维特性以及流量的内在结构，对于如何分析诸如延时、丢包数等关系网络性能的参量还有待于进一步研究。

目前，网络流量的分析方法有许多种，其中比较常用的是小波分析法。小波分析法是 80 年代后期发展起来的应用数学分支。它是属于时频分析的一种，而 PCA 算法是一种线性映射的方法，只需在时域上进行变换。它是根据样本点在多维模式空间的位置分布，以样本点在空间中变化最大方向，即方差最大的方向，作为判别矢量来实现数据的特征提取与数据压缩的。因此 PCA 算法相对小波分析法简单。

其次，小波分析法主要用于信号特征的提取，它能够分析出各种不同频率的成分，但是小波分析法一般需要你自己制定用什么样的小波函

数去分解，同时需要指出阶数，这些都需要经验。如果你找对了小波函数和用几阶去逼近，那么你可能能够得到和 PCA 接近的结果。而从 5.2.2 和 5.2.3 小节可以看出，PCA 算法也能够从原流量信息中提取三部分的特征，而这 3 部分特征恰好反映网络流的三种主要特征：规律性的正常行为，异常现象和随机的波动。这一结果只是通过采用 PCA 算法对输入数据在原空间上进行转换，选择保留的维数和设置相应的门限值就达到了，经验的东西少了，确定性增大了。

不过，小波分析法具有多分辨率分析，即多尺度的特点，可以观察网络流量在大时间尺度和小时间尺度下的不同性质。而 PCA 算法就难以达到这个精确程度，并且受线性关系的制约。

第六章　结束语

在这篇论文中，根据 Anukool Lakhina，Mark Crovella，Christophe Diot 等人提出的想法，我采用 PCA 算法分析了一个真实网络的所有流量的时间序列的内在结构。

这篇论文探讨的第一个问题是网络中所有的 OD 流是否存在低维特性，即 OD 流的特性能否被一小部分变量近似捕获。我们知道，即使一个中等规模的网络也存在近百条 OD 流，各条 OD 流的流量间存在错综复杂的关系，因此这个网络的流量矩阵就是一个高维多变的对象。采用主成分分析法，即 PCA 算法对这一流量矩阵进行处理后，我们发现，近百条 OD 流可以用 3－7 个独立的变量精确的描述。

这一低维性质又促使论文继续探讨第二个问题，即能否通过研究这些独立的成分进而研究每条 OD 流的行为以及各 OD 流间的相同点和差异。我们发现，通过将原流量数据投影到各主成分方向，将得到另一组变量，称其为特征流。这些特征流反映了所有 OD 流共同的不同方面的特征，由此能更好的研究 OD 流的结构。进一步的研究发现，所有的 OD 流都表现出三方面的特征：确定的趋势，脉冲和噪声。这也就是特征流的三种分类所表现出来的特性。因此，通过采用 PCA 算法进行处理，我们能够通过图形更直观的研究 OD 流的内在结构，其中脉冲型的特征流有助于检测网络流量的突发点。但是，目前一般通过从特征流的时间序列直接观察来确定其分类，至于明确的分类依据和这三种特征流的分布规律还有待于进一步研究。

接下来，我对 PCA 算法的这一处理在时域上的作用范围进行了研究，即前一时段的流量采用 PCA 算法处理的结果是否能够分析下一时段的流

量特性。实验发现，由此引起的误差还是在可以接受的范围内。因此，PCA 处理保持了一定的时间稳定性。

总的来说，在图像识别、神经网络等高维领域应用成熟的 PCA 算法也适用于网络流量分析领域。

参考文献

[1] Anukool Lakhina, Mark Crovella, Christophe Diot. "Diagnosing Network-Wide Traffic Anomalies". In ACM SIGCOMM, Portland, August 2004.

[2] Lindsay I Smith. "A tutorial on Principal Components Analysis". February 2002.

[3] I.Antoniou, V.V.Ivanov, Valery V.Ivanov, P.V.Zrelov. "Principal Component Analysis of Network Traffic Measurements: the <Caterpillar>-SSA Approach".

[4] Jolliffe I.T. "Principal Component Analysis". Springer-Verlag, 1986.

[5] Jackson J.E. "A User's Guide to Principal Component Analysis". N.Y., 1992: P26~62.

[6] Anukool Lakhina, Mark Crovella, Christophe Diot. "Characterization of Network-Wide Anomalies in Traffic Flows". ICM'04, October 2004.

[7] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft. "Structural Analysis of Network Traffic Flows". In ACM SIGMETRICS, New York, June 2004.

[8] Khaled Labib and V.Rao Vemuri. "Detecting Denial-of-Service and Network Probe Attacks using Principal Component Analysis".

[9] J. E. Jackson and G. S. Mudholkar. "Control Procedures for Residuals Associated with Principal Component Analysis". Technometrics, 1979, P341~349.

[10] Diamantaras K.I, Kung S.Y. "Principal Component Neural Network:

Theory and Applications". New York: John Wiley & Sons, 1996: P44~48.

[11] Xu Lei, Yuille A.L. "Robust principal component analysis by self-organizing rules based on statistical physics approach". IEEE Trans Neural Networks, 1995, 6(1): P131~134.

[12] 孟德顺，《PCA应用中的几个问题》，西北林学院学报，1995。

[13] 程光，龚俭，《大规模高速网络流量测量研究》，计算机工程与应用，2002：17～19。

[14] 王松，夏绍玮，《基于单层网络的自组织的鲁棒主成分分析（PCA）算法》，清华大学学报，1997，37(7)：121～124。

[15] 王松，夏绍玮，《一种鲁棒主成分分析（PCA）算法》，系统工程理论与实践，1998：9～13。

[16] 王松，夏绍玮，《基于误差模型的自适应鲁棒主成分分析》，自动化学报，1999，25(4)：528～531。

[17] 石晶林，郭志刚，曾志民，丁炜，《因特网络流量工程概述》，中国数据通信，2001：9～15。

[18] 刘亮，《基于神经网络的非线性 PCA 方法》，自动化技术与应用，2004，23(5)：8～11。

[19] 吕军，李星，《网络测量分析及研究综述》，计算机工程与应用，2003：19～22。

[20] 李伟，《网络业务流分析的研究与实现》，西安交通大学硕士学位论文，2002。

致 谢

本论文是在导师陈常嘉教授悉心指导下完成的。陈老师深厚的理论基础，渊博的学识，踏实的工作作风，敏捷的思维，严谨的治学态度给我留下了深刻的印象。导师身体力行，谆谆教导，不仅使得学习和研究工作得到顺利的进行，更使我在为人处世上受益匪浅。在此，我向导师陈常嘉教授致以最诚挚的感谢。

感谢在读博士研究生刘紫千，在课题的期间，他给予了我许多指导和帮助。还有很多的研究思路。在这里表示深深的感谢。

感谢北京交通大学通信实验室所以的老师和同学，是他们给我提供了良好的学习，科研环境还有实验条件。

感谢我所有的亲人，在我近20年求学生涯中为我付出的心血，正是他们无私的奉献才让我不断地成长。

最后，衷心感谢在百忙中抽出时间对我论文进行认真评审并提出宝贵意见的各位专家。

攻读研究生期间发表的论文

1．王敏，李纯喜，陈常嘉，《浅谈基于PCA的网络流量分析》，微计算机信息，2006.3：94～95。