2015 南京理工大学大学生数学建模竞赛

承　诺　书

我们仔细阅读了中国大学生数学建模竞赛的竞赛规则.

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们授权全国大学生数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

我们参赛选择的题号是（从 A/B 中选择一项填写）：A

我们的参赛报名号为（报名编号）：3

所属学院（请填写完整的全名）：自动化学院

参赛队员（打印并签名）：1. 展慧馨

2. 黄丽丽

3. 程光冉

日期：2016 年 5 月 24 日

评阅编号（由组委会评阅前进行编号）：

e 时代的信息安全系统

摘要

本文对信息系统安全程度问题进行了深入的讨论，建立了数学模型。首先，对于典型的信息系统——不同类型的网站建立了基于 R 型聚类分析法和主成分分析法下的多指标系统安全综合模型，接着基于上述讨论，选出最重要的评价指标，并建立信息系统安全程度综合得分模型和等级评价制度。经过我们查阅数据所选择的 14 种网站来进行安全度排名和等级评价，最后，将近期手机系统数据代入之前的模型，模型检验用 t 参数检验法，证明所建立模型的正确性，依据现代模式识别的理论做了模型推广。

针对问题一，我们将资产集、业务集、拓扑结构、脆弱性集作为衡量信息系统安全程度的四大指标高层，每个高层下又分设共七个小指标。利用相似性度量和聚类分析法将七个指标分成四个类，依据聚类图舍弃与其他指标相关系数较小且对于信息系统安全程度影响最低的指标。接着利用主成分分析法求解剩余六个指标的主成分方程，挑选贡献率最高的四个成分，进一步验证之前舍弃对信息安全系统影响程度最低的指标的合理性，并将生育成分的贡献率作为权重，得到信息系统安全程度综合得分模型和等级评价制度。

针对问题二，利用问题一建立的模型，根据聚类分析法结果可初步得出独立性最高、对综合得分影响最高的一组重要的信息安全度量指标，再根据主成分分析中得出的贡献率高的成分的特征向量可初步得出另外一组重要的信息安全度量指标，综合两组可得一个重要的信息安全度量指标，也就是单个泄露信息漏洞可能泄露的信息量。

针对问题三，我们将手机热度类型对应指标 1，手机全球排名对应指标 2，2015 年手机出售台数在全球手机出售台数所占比例对应指标 3，手机系统对应指标 4，软件平均漏洞数对应指标 5。对于手机模型，为了分析简化，我们将上述的指标 5、6、7 合并为一个指标即指标 5。指标 1 手机热度类型排名和指标 2 全球排名近似相等。然后将组内三人手机相应指标对应的数据代入上面的模型，得到手机安全度的排名。

关键词：相似性度量　聚类分析法　主层次分析法　综合得分模型　等级评价

一、问题重述

1.1 背景资料与条件

斯诺登事件为我们敲响了信息安全的警钟，也让我们更进一步认识到当前网络信息安全所面临形势的严峻性。保障我国网络信息安全，是当前面临的重要问题。信息安全度量是业界公认的一个难题，信息安全度量一般需要回答两个问题：信息系统安全不安全？信息系统的安全程度是多少？

1.2 需要解决的问题

（1）基于“2015年信息安全事件汇总报告”（http://mt.sohu.com/20160113/n434399073.shtml）以及其它网络数据，建立一个计算信息系统（孤立隔离，或广泛互联的系统）的安全程度的数学模型。

（2）选取一个重要的信息安全度量指标，说明选取的理由，并给出计算该指标的数学模型。

（1）利用上述两个模型，具体对你们小组成员所持有的手机信息系统进行研究，给出计算结果，并进行简单分析与比较。

二、问题分析

2.1 问题的重要性分析

计算机和网络技术在当今社会迅猛发展并且得到广泛应用，使得各行各业对信息系统的依赖日益加深，信息技术几乎渗透到了社会生活的方方面面。信息系统及其所承载信息的安全问题日益突出，信息和服务在保密性、完整性、可用性等方面出现问题，都会给组织机构带来很大的负面影响。由于互联网的存在，组织机构的信息系统不可避免的与外部系统有着复杂的联系，因此信息安全和保密尤其重要。保障我国网络信息安全，是当前面临的重要问题。信息安全度量是业界公认的一个难题，信息安全度量一般包括两个方面：网络信息系统的安全与否以及它的安全程度。因此，我们我们需要通过建立模型来确定网络信息系统是否安全以及它的安全程度。

2.2 问题的思路分析

2.2.1 问题一的分析

基于“2015 年信息安全事件汇总报告”和调查的网络数据，我们选择一个广泛互联的信息系统，通过聚类分析法建立了一个安全评价模型。该模型有四个大指标评价标准：资产集、业务集、拓扑结构、脆弱性集。每个大指标包含由统计数据或者相关文献得来的小指标。通过打分制度建立综合打分模型。

2.2.2 问题二的分析

问题二要求我们选取一个重要的信息安全度量指标。我们通过主层次分析法剔除一些相关性较小的指标，并选取相关性最大的安全指标即最重要的安全度量指标。通过 MATLAB 编写程序来计算系统的安全度。

2.2.3 问题三的分析

问题三要求我们利用小组成员的手机信息系统对上述两个模型进行验证。

三、问题假设

1. 本论文所建立的模型是计算广泛互联的信息系统的安全程度的数学模型。
2. 假设忽略环境因素对信息系统安全程度的影响。

3. 假设忽略硬件如传输线路等对信息系统安全程度的影响。
4. 假设所测信息系统采用相同的服务器和防火墙。
5. 假设所查数据都是真实准确的。

四、符号说明

4.1 名词的解释

（1）资产集：为攻防场景中有价值的资源的集合，分为点击量和全球网站排名；
（2）业务集：为攻防场景中所有业务的集合，业务反映了网络的具体功能，即网站类型；
（3）拓扑结构：网站中各个站点相互连接的形式；
（4）脆弱性集：为攻防场景中可被攻击序列利用的缺陷的集合，又称为漏洞集；
（5）NC 规则网络：网络中每个节点仅与该节点周围的若干个节点相连的网络；
（6）ER 随机网络：网络中以概率 p 连接 N 个节点中的每一对节点的网络；
（7）WS 小世界网络：网络中含有 n 个节点、每个节点有 k 个邻居、以概率 p 随机化重连边的网络；
（8）BA 无标度网：网络含有 n 个节点、每次加入 m 条边的网络；

4.2 符号说明

符号	符号的含义
X_1	某信息系统对应行业的热度总积分
X_2	三个月内某信息系统的平均排名
X_3	三个月内某信息系统的日均点击量
X_4	某信息系统对应拓扑结构的传播阈值
X_5	某信息系统的漏洞总数
X_6	某信息系统的泄露信息漏洞总数
X_7	某信息系统的单个泄露信息漏洞的泄露信息量
a_{ij}	第 i 个评价对象的第 j 个指标的取值
r_{jk}	第 j 个评价对象的第 k 个指标的相关系数
Z	某信息系统在主成分综合评价模型中的综合得分
b_{ij}	a_{ij}对应的标准化指标
r_{ij}	第 i 个指标与第 j 个指标的相关系数

五、模型的建立与求解

5.1 数据的采集和处理

5.1.1 数据的来源说明

本文中的数据来源于 http://www.alexa.cn/和 http://top.chinaz.com/hangyetop/index.html?qq-pf-to=pcqq.discussion 和 2015 年互联网安全报告。

5.1.2 数据预处理

通过查阅文献[1]可知四种拓扑类型 BA 无标度网、WS 小世界网络、NC 网络、ER 随机网络的阈值$\lambda = \langle k\rangle / \langle k^2\rangle$,$\langle k\rangle$表示各节点的平均度，$\langle k^2\rangle$表示网络的平均均方度，反映网络的异质化程度。根据 NC 网络定义、ER 随机网络模拟算法、WS 小世界网络模拟算法和 BA 无标度网络模拟算法，生成四种典型的网络，参数如表 5-1，λ 越大，安全度越高。

表 5-1 仿真网络参数

参数类型	NC 网络	ER 随机网络	WS 小世界网络	BA 无标度网络
平均度<k>	6	6.014	6	6.004
平均均方度< k^2>	36	42.432	37.07	86.756
传播阈值λ	0.16666	0.14173	0.16186	0.06921

5.2 问题一

影响一个网络信息系统安全的因素有很多，在系统分析和评估过程中，为避免遗漏重要因素，往往在一开始选取指标时，考虑尽可能多的相关因素，然后利用变量聚类法，研究变量之间的相似关系，按变量的相似关系把变量聚合成若干类，进而找出影响系统的主要因素。从而将问题简单化。

5.2.1 模型的建立

由以上的分析，我们将影响系统安全的因素分为四大类：①资产集即系统有价值的资源。可认为是某一信息系统拥有或控制的、能够带来经济利益的全部资产。②业务集即网络类型和它提供的服务，通过以往案例可知电商网站和金融网站等涉及到金钱交易和有大量用户数据资料的网站容易被黑客攻击。③拓扑结构即网站中各站点相互连接的方式，不同的拓扑结构阈值不同，被攻击的可能性也有所区别。根据拓扑结构可测算一个信息系统的稳定性。④脆弱性即信息系统存在的漏洞，漏洞越多越容易被攻击。由此可得各指标分类如下：

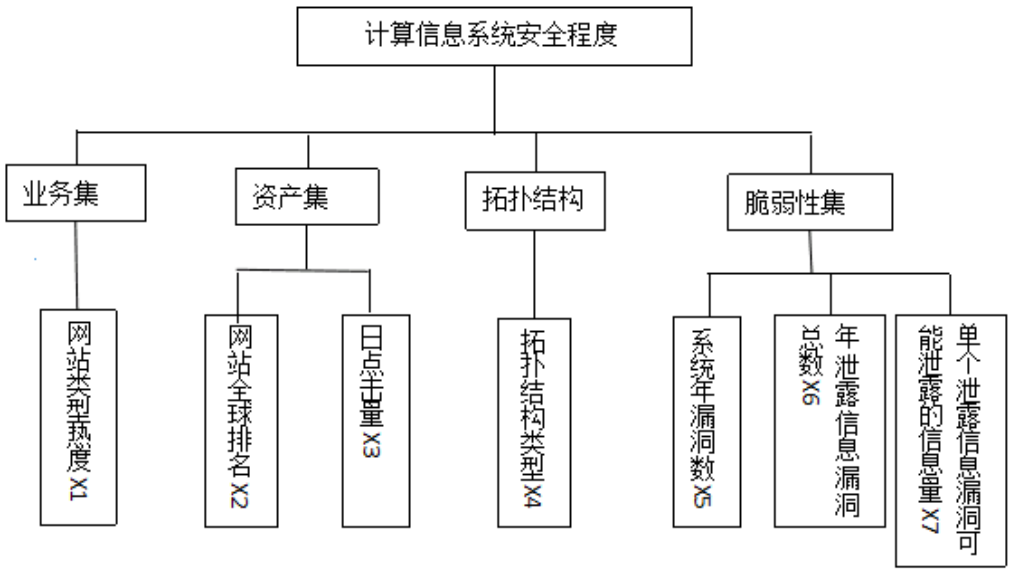


图 1　各指标关系图

5.2.1.1 资产集指标的确定

根据（4.1）中对此指标的解释，资产集是一个系统有价值的资源，我们将其划分为网站日均点击量和全球网站排名。网站的点击量和排名越高，代表网站越有价值。通过 ALEXA 全球网站排名，可以查出一个网站近三个月的平均日点击量和全球网站排名。详细数据见表 5-2。

5.2.1.2 业务集指标的确定

业务集是一个网站的类型及它提供的服务，由常识可知越热门的网站越容易被攻击。通过调查数据，我们得到了近一年 14 个类型不同的网站的类型热度。详细数据见表 5-2。

5.2.1.3 拓扑结构指标的确定

拓扑结构指标是指网站中各站点相互连接的方式，由文献[1]可知，不同的拓扑结构阈值不同，阈值λ 越大，安全风险越不容易爆发，系统越安全。通过数据预处理，我们得到四种拓扑结构的阈值。

5.2.1.4 脆弱性集指标的确定

脆弱性集是指网站被攻击的可能性。我们将其具体划分为三个指标：系统年漏洞总数、泄露用户信息的漏洞数和单个泄露信息漏洞可能泄露的信息量。详细数据见表 5-2。

5.2.1.5 整体指标的确定

由以上的分析，我们将影响信息系统安全的指标分为 7 个，分别是：指标 1-网站类型热度 x_1，指标 2-网站排名 x_2，指标 3-网站日点击量 x_3，指标 4-网站结构稳定性 x_4,指标 5-网站年漏洞总数 x_5，指标 6-网站年泄露信息漏洞总数 x_6，指标 7-网站单个泄露信息漏洞可能泄露的信息量 x_7。根据查阅的数据，我们得到的 14 个信息系统作为评价对象的各指标值如下:

表 5-2 信息系统对应指标参数

序号	信息系统名称	网站类型	网站类型热度	平均排名（3 个月）
1	百度	综合类	14739059	4
2	淘宝	电子商务类	3842948	12
3	新浪	综合类	14739059	9187
4	世纪佳缘	生活服务类	15204799	4615
5	东方财富网	财经类	3982408	757
6	爱奇艺	视频类	13297141	232
7	网易	综合类	14739059	103
8	中国光大银行	银行类	6034680	70965
9	中国电信	通信类	6034680	2479
10	中国南京网	政府类	4318928	46856
11	起点中文网	小说类	13297141	2232
12	极客工坊	科技类	6083860	48403
13	慕课网	教育类	9710674	4749
14	好大夫在线	医疗类	1924897	57854
序号	**信息系统名称**	**日均点击量（3 个月）**	**拓扑结构类型**	**拓补结构稳定性λ**
1	百度	268044000	BA 无标度网络	0.06976
2	淘宝	690360000	WS 小世界网络	0.16216
3	新浪	444000	NC 网络	0.16666
4	世纪佳缘	8325000	WS 小世界网络	0.16216
5	东方财富网	19476000	NC 网络	0.16666
6	爱奇艺	3699000	WS 小世界网络	0.16216

7	网易	115350000	BA 无标度网络	0.06976
8	中国光大银行	738000	ER 随机网络	0.14173
9	中国电信	5310000	WS 小世界网络	0.16216
10	中国南京网	27285	BA 无标度网络	0.06976
11	起点中文网	13398000	NC 网络	0.16666
12	极客工坊	198000	ER 随机网络	0.14173
13	慕课网	4644000	BA 无标度网络	0.06976
14	好大夫在线	108000	WS 小世界网络	0.16216
序号	**信息系统名称**	**年漏洞总数**	**年泄露信息漏洞总数**	**单个泄露信息漏洞可能泄露的信息量**
1	百度	2330	180	291
2	淘宝	2330	180	291
3	新浪	2330	180	291
4	世纪佳缘	2330	180	291
5	东方财富网	543	41	267
6	爱奇艺	2330	180	291
7	网易	2330	180	291
8	中国光大银行	543	41	267
9	中国电信	549	35	563
10	中国南京网	2330	180	291
11	起点中文网	2330	180	291
12	极客工坊	2330	180	291
13	慕课网	1169	22	112
14	好大夫在线	383	25	961

5.2.3 模型的求解

5.2.3.1 指标的相似性度量

要用数量化的方法对信息系统的安全评价指标进行分类，就必须用数量化的方法描述指标之间的相似程度。本模型中采用以上 7 个指标x_1~x_7对安全程度进行刻画。在对各项指标进行聚类分析时，首先要确定指标的相似性度量，这里采用相关系数来度量。记指标 x_j 的取值（x_{1j}，x_{2j}, …，x_{nj}）$^T \in R^n$ (j=1,2, ...，m)，则可用两指标 x_j 与 x_k 的样本相关系数 r_{jk} 作为它们的相似性度量，r_{jk} 求解过程如下:

$$r_{jk}=\frac{\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)\left(x_{ik}-\overline{x_k}\right)}{\left[\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)^2\sum_{i=1}^{n}\left(x_{ij}-\overline{x_k}\right)^2\right]^{\frac{1}{2}}}$$

通过 MATLAB 求解的相关系数矩阵如下：

表 5-3　7 个指标间的相关系数

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1.0000	-0.5552	-0.1216	-0.1785	0.6041	0.5514	-0.4300
x_2	-0.5552	1.0000	-0.3107	-0.0082	-0.3492	-0.2839	0.3606
x_3	-0.1216	-0.3107	1.0000	-0.0348	0.2974	-0.3034	-0.1162
x_4	-0.1785	-0.0082	-0.0348	1.0000	-0.2183	-0.0994	0.3346
x_5	0.6041	-0.3492	0.2974	-0.2183	1.0000	0.9714	-0.4765
x_6	0.5514	-0.2839	0.3034	-0.0994	0.9714	1.0000	-0.3711
x_7	-0.4300	0.3606	-0.1162	0.3346	-0.4765	-0.3711	1.0000

r_{jk} 越接近 1，x_j 与 x_k 越相关；r_{jk} 越接近 0，x_j 与 x_k 相似性越弱。

由相关系数矩阵可以看出某些指标间确实存在很强的相关性，比如指标 x_5，指标 x_6，指标 x_1，指标 x_5 或指标 x_3，指标 x_5。

5.2.3.2 指标聚类

指标聚类的基本思想是：我们所研究的指标之间存在程度不同的相似性（亲疏关系——以样品间距离衡量）。于是根据不同信息系统的多个观测指标，具体找出一些能够度量指标之间相似程度的统计量，以这些统计量为划分类型的依据。把一些相似程度较大的指标聚合为一类，把另外一些彼此之间相似程度较大的或指标又聚合为另一类，直到把所有的指标聚合完毕。R 型聚类分析是对变量进行分类处理，适用于此次模型，首先利用 MATLAB 程序中的 pdist 函数用欧式距离法根据各指标间的相关系数导出各指标间的距离。

各指标间的距离 $d(x_j, x_k)$ 求解如下：

$$d_{(j,k)}=\left[\sum_{p=1}^{7}\left|x_p-y_p\right|^2\right]^{\frac{1}{2}}$$

求解的各指标间的距离如下：

表 5-4　7 个指标间的距离

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	\	1.5552	1.1216	1.1785	0.3959	0.4486	1.4300
x_2	1.5552	\	1.3107	1.0082	1.3492	1.2839	0.6394
x_3	1.1216	1.3107	\	1.0348	0.7026	0.6966	1.1162
x_4	1.1785	1.0082	1.0348	\	1.2183	1.0994	0.6654
x_5	0.3959	1.3492	0.7026	1.2183	\	0.0286	1.4765
x_6	0.4486	1.2839	0.6966	1.0994	0.0286	\	1.3711
x_7	1.4300	0.6394	1.1162	0.6654	1.4765	1.3711	\

然后用类平均法分类，利用 MATLAB 程序中 linkage 函数进行聚类，分类完成后再用 dendrogram 函数绘制如下聚类图。（横坐标为 x_1~x_7 这七个指标，纵坐标为平台高度）

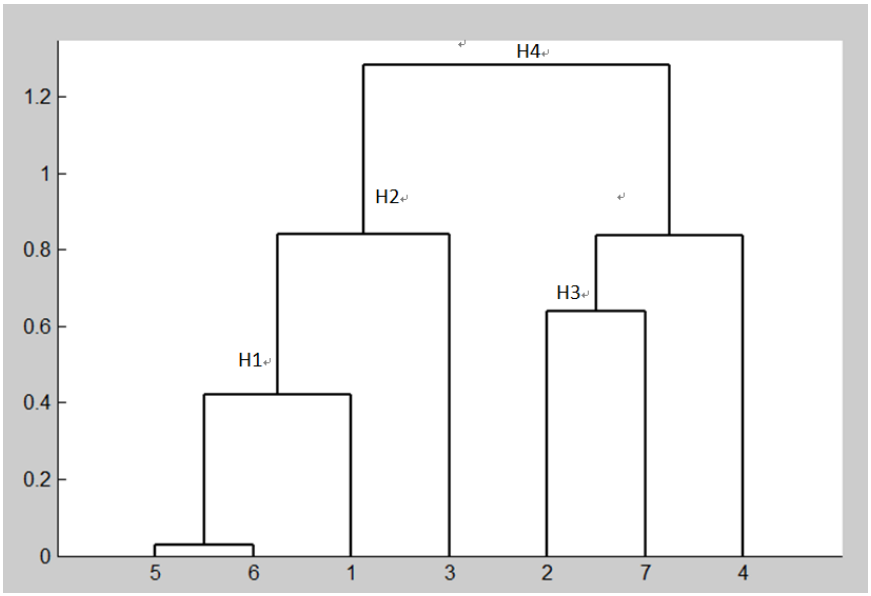


图 2　七个指标的聚类图

第一步，所有的元素分为一类 H_4｛x_1，x_2，x_3，x_4，x_5，x_6，x_7｝，每一类的平台高度为 0；

第二步，取新类的平台高度为 0.4，把 x_1，x_5，x_6，合成一个新类 H_1；

第三步，取新类的平台高度为 0.6，把 x_2，x_7 合成一个新类 H_3；

第四步，取新类的平台高度为 0.8，把 H_1，x_3 合成一个新类 H_2；

第五步，取新类的平台高度为 1.2，把 H_2，H_3，x_4 合成一个新类 H_4

H_4 就包含了所有指标 x_1~ x_7，根据不同的平台高度绘制了如下二叉树。

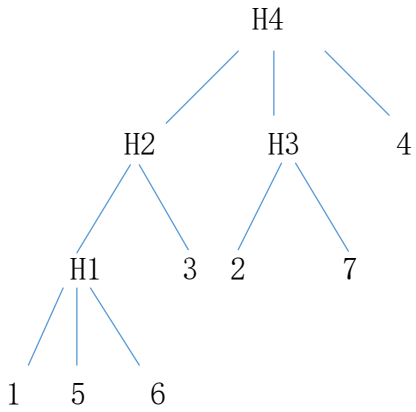


图 3　七个指标的二叉图

在聚类图和二叉树中可看出，第一类G_1包括指标x_4，第二类G_2包括指标x_2、x_7，第三类G_3包括指标x_3，第四类G_4包括指标x_1、x_5、x_6。如果将 7 个指标分为 4 类，从平台最低的G_4类剔除指标 6（年泄露信息漏洞总数），这样就从 7 个指标中选定了 6 个相关性比较高的分析指标：网站类型热度 x_1，网站排名 x_2，网站三月平均日点击量 x_3，网站结构稳定性 x_4，网站年漏洞总数 x_5，网站单个泄露信息漏洞可能泄露的信息量 x_7。

5.2.3.3　信息系统安全程度的综合得分计算

主成分分析的主要目的是将我们手中相关性很高的变量转化成彼此相互独立或不相关的变量，从而去解释原来资料中的大部分变异。通常是选出比原始变

量个数少，能解释大部分资料中的变异的几个新变量，即所谓主成分，进而以解释资料的综合性指标。针对本问题的特点，正好适合用主成分分析进行聚类分析后的 6 个指标，从而计算信息系统安全程度的综合得分。

（1）对表 5-2 数据进行标准化处理

进行主成分分析的指标变量有 6 个，分别为 x_1，x_2，x_3，x_4，x_5， x_7，共有 14 个信息系统评价对象，第 i 个评价对象的第 j 个指标的取值为 a_{ij}。将各指标值 a_{ij} 转换成标准化指标值 b_{ij}，有

$$b_{ij}=\frac{a_{ij}-\mu_j}{s_j}, i=1,2,\cdots,14;\ j=1,2,\cdots,6$$

其中：

$$\mu_j=\frac{1}{n}\sum_{i=1}^{n}a_{ij}, s_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(a_{ij}-\mu_j\right)^2}, j=1,2,\cdots,m$$

即第 j 个指标的样本均值和样本标准差。对应地，我们称

$$x_j^{'}=\frac{x_j-\mu_j}{s_j}, j=1,2,\cdots,6$$

为标准化标准变量。

通过 MATLAB 得到的标准化矩阵如下：

表 5-5　14 个评价对象的 6 个指标的标准化矩阵

	指标x_1	指标 x_2	指标 x_3	指标 x_4	指标 x_5	指标x_7
评价对象 1	1.1212	-0.6877	0.9839	-1.4971	0.7041	-0.2555
评价对象 2	-1.0604	-0.6874	3.2021	0.6628	0.7041	-0.2555
评价对象 3	1.1212	-0.3318	-0.4217	0.7680	0.7041	-0.2555
评价对象 4	1.2144	-0.5090	-0.3803	0.6628	0.7041	-0.2555
评价对象 5	-1.0325	-0.6585	-0.3217	0.7680	-1.3774	-0.3755
评价对象 6	0.8325	-0.6789	-0.4046	0.6628	0.7041	-0.2555
评价对象 7	1.1212	-0.6839	0.1819	-1.4971	0.7041	-0.2555
评价对象 8	-0.6216	2.0628	-0.4201	0.1852	-1.3774	-0.3755
评价对象 9	-0.6216	-0.5918	-0.3961	0.6628	-1.3704	1.1051
评价对象 10	-0.9651	1.1283	-0.4238	-1.4971	0.7041	-0.2555
评价对象 11	0.8325	-0.6014	-0.3536	0.7680	0.7041	-0.2555
评价对象 12	-0.6117	1.1883	-0.4230	0.1852	0.7041	-0.2555
评价对象 13	0.1144	-0.5038	-0.3996	-1.4971	-0.6482	-1.1508
评价对象 14	-1.4444	1.5546	-0.4234	0.6628	-1.5638	3.0960

（2）计算相关系数矩阵 R=（r_{ij})$_{mxn}$，r_{ij}是第 i 个指标与第 j 个指标的相关系数，计算过程如下：

$$r_{ij}=\frac{\sum_{k=1}^{14}b_{ki}\cdot b_{ki}}{n-1},i,j=1,2,\cdots,6$$

通过得到的相关系数矩阵如下：

表 5-6　6 个指标间的相关系数

	x_1	x_2	x_3	x_4	x_5	x_7
x_1	1.0000	-0.5552	-0.1216	-0.1785	0.6041	-0.4300
x_2	-0.5552	1.0000	-0.3107	-0.0082	-0.3492	0.3606
x_3	-0.1216	-0.3107	1.0000	-0.0348	0.2974	-0.1162
x_4	-0.1785	-0.0082	-0.0348	1.0000	-0.2183	0.3346
x_5	0.6041	-0.3492	0.2974	-0.2183	1.0000	-0.4765
x_7	-0.4300	0.3606	-0.1162	0.3346	-0.4765	1.0000

（3）计算特征值和特征向量。用 MATLAB 程序计算相关系数矩阵 R 的特征值λ，$\lambda_1\geqslant\lambda_2\geqslant\cdots\lambda_5\geqslant\lambda_7\geqslant0$，及对应的特征向量 u_{11}，u_{12}，…，u_{76}，u_{77} 由特征向量组成 6 个新的指标变量如下:

$y_1=-0.4979x_1'-0.2373x_2'+0.4275x_3'-0.1051x_4'+0.2852x_5'+0.6486x_7'$；
$y_2=0.4389x_1'-0.3484x_2'-0.3274x_3'-0.5911x_4'-0.2192x_5'+0.4258x_7'$；
$y_3=-0.1871x_1'+0.7461x_2'-0.4837x_3'-0.0436x_4'+0.0734x_5'+0.4087x_7'$
$y_4=0.2317x_1'+0.4782x_2'+0.6640x_3'+0.2568x_4'-0.4533x_5'-0.0730x_7'$；
$y_5=-0.5066x_1'+0.0452x_2'-0.0582x_3'-0.7203x_4'+0.0364x_5'-0.4667x_7'$；
$y_6=0.4627x_1'+0.1869x_2'+0.1787x_3'+0.2300x_4'+0.8104x_5'-0.0882x_7'$。

其中：y_1是第一主成分，y_2 是第二主成分，…，y_6 是第六主成分。

（4）选择 4 个主成分

计算 6 个主成分对应特征值 λ_j 的信息贡献率和累计贡献率。称

$$c_j=\frac{\lambda_j}{\sum_{k=1}^{6}\lambda_k},j=1,2,\cdots,6$$

为主成分y_i的信息贡献率，同时，有

$$\alpha_p=\frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{6}\lambda_k}$$

为主成分 y_1，y_2，…，y_6 的累计贡献率。这里我们取 α_p 接近于 0.90 时，选择前 4 个指标变量 y_1，y_2, y_3，y_4 作为 4 个主成分，代替原来 6 个指标变量，从而可对 4 个主成分进行综合分析。通过 MATLAB 程序得到的 6 个主成分贡献率如下：

表 5-7　6 个主成分贡献率

序号	特征值	贡献率%	累计贡献率%
1	2.5411	42.3523	42.3523
2	1.1504	19.1732	61.5255
3	1.0236	17.0607	78.5862
4	0.5896	9.8263	88.4125
5	0.5327	8.8777	97.2902
6	0.1626	2.7098	100

（5）通过选择的 4 个主成分计算信息系统安全程度综合评价值和评价等级
分别以 4 个主成分的贡献率为权值，构建主成分综合评价模型即

$$Z=42.3523y_1+19.1732y_2+17.0607y_3+9.8263y_4$$

我们利用专家打分的方法，确定 5 个等级以比较准确地判断信息系统安全程度的具体大小。这 5 个等级从低到高依次是：“信息系统安全程度”程度差，“信息系统安全程度”程度合格，“信息系统安全程度”程度一般，“信息系统安全程度”程度较好，“信息系统安全程度”程度很好。“信息系统安全程度”Z 值对应等级确定方法如表所示：

表 5-8　评价等级表

Z 的数值	(-∞,-1.2)	[-1.2,-0.6)	[-0.6,0)	[0,0.6)	[0.6,+ ∞]
对应等级	信息系统安全程度较差	信息系统安全程度合格	信息系统安全程度一般	信息系统安全程度较好	信息系统安全程度很好

我们对表 5-2 中的 14 个评价对象进行综合评分和等级评价，结果如下：

表 5-8　14 个评价对象的信息系统安全程度得分

排名	信息系统名称	信息系统序号	信息系统安全程度综合得分	信息系统安全程度等级评价
1	世纪佳缘	4	0.689529	很好
2	淘宝	2	0.684725	很好
3	起点中文网	11	0.635432	很好
4	爱奇艺	6	0.633619	很好
5	新浪	3	0.629885	很好
6	百度	1	0.622425	很好
7	网易	7	0.505922	较好
8	慕课网	13	-0.161591	一般
9	极客工坊	12	-0.25882	一般
10	东方财富网	5	-0.387788	一般
11	中国电信	9	-0.483452	一般
12	中国南京网	10	-0.553937	一般

13	中国光大银行	8	-0.1058359	合格
14	好大夫在线	14	-0.149759	较差

由生活经验可知，世纪佳缘、淘宝等网站涉及用户的个人信息、财产信息较多，受到不良攻击的可能性也就最大，故网站安全建设必定较为严密，从而所以信息安全程度高。这近一步的验证了我们的模型的合理性。

5.3 问题二

从 5.2.3.2 中剔除指标x_6后剩下的 6 个指标选取一个重要的信息安全度量指标，我们可结合 5.2.3.2 中的聚类图和 5.2.3.3 中的贡献率表格选取出所需要的信息安全度量指标。

（1）由 5.2.3.2 选出信息安全度量指标

由图 2 和图 3 可知，指标x_4（拓补结构稳定性）和指标x_7（单个泄露信息漏洞可能泄露的信息量）和其他指标的相关性较小，独立性较高，故其对信息系统安全程度的影响也较高。所以由 5.2.3.2 我们初步选定指标x_4和指标x_7作为重要信息安全度量指标。

（2）由 5.2.3.3 选出信息安全度量指标

由我们经主成分选择后的 4 个主成分的特征向量如下:

y_1=−0.4979x_1'−0.2373x_2'+0.4275x_3'−0.1051x_4'+0.2852x_5'+0.6486x_7'；
y_2=0.4389x_1'−0.3484x_2'−0.3274x_3'−0.5911x_4'−0.2192x_5'+0.4258x_7'；
y_3=−0.1871x_1'+0.7461x_2'−0.4837x_3'−0.0436x_4'+0.0734x_5'+0.4087x_7'；
y_4=0.2317x_1'+0.4782x_2'+0.6640x_3'+0.2568x_4'−0.4533x_5'−0.0730x_7'。

通过分析，可看出第一个主成分主要反映指标x_1，x_5，x_7，第二个主成分主要反映指标x_4，第三个主成分主要反映指标x_2，第四个主成分主要反映指标x_3。

由表 5-7 可知第一个成分所占故由 5.2.3.3 选取的重要信息安全度量指标初步选定为x_1，x_5，x_7。

（3）选定一个重要的信息安全度量指标

综合（1）和（2），我们选定指标x_7（即单个泄露信息漏洞可能泄露的信息量）作为一个重要的信息安全度量指标。

5.4 问题三

根据上面建立的模型，我们将手机按手机类型热度对应指标C_1，手机全球排名对应指标C_2，2015 年手机出售台数在全球手机出售台数所占比例对应指标C_3，手机系统对应指标C_4，软件平均漏洞数对应指标C_5。对于手机模型，为了分析简化，我们将上述的指标C_5、C_6、C_7合并为一个指标即指标C_5。指标 C_1手机类型热度排名和指标C_2全球排名近似相等。我们小组三人使用的分别是华为、小米、vivo 手机，对应的指标及数据如表 5-9。其中指标C_5的计算根据调查【2】2015 年 APP 九大行业榜单产品,可得到各类型 APP 平均漏洞数（具体数据见附录）然后用手机内各 APP 类型的个数乘以相应的比例即可得指标C_5对应值，进而得到本小组各成员手机信息系统的信息安全程度总得分和安全程度如下：

5-9 小组成员手机对应指标

手机型号	手机热度类型C_1	手机全球排名C_2	销售量比例C_3	手机系统C_4	软件平均漏洞数C_5	对应评分	安全程度等级
华为	3	3	8.40%	安卓	2.65	1.311	很好
小米	4	4	5.60%	安卓	2.26	0.7439	很好
vivo	9	9	3.30%	安卓	4.45	-2.0549	较差

由上表可知，本小组各成员安全性：华为>小米>vivo. 由手机排名可知，安全性主要由软件漏洞数决定，和 5.2、5.3 的模型相符，也符合实际。说明上面建立的模型正确。

其中，小组成员的软件统计如下

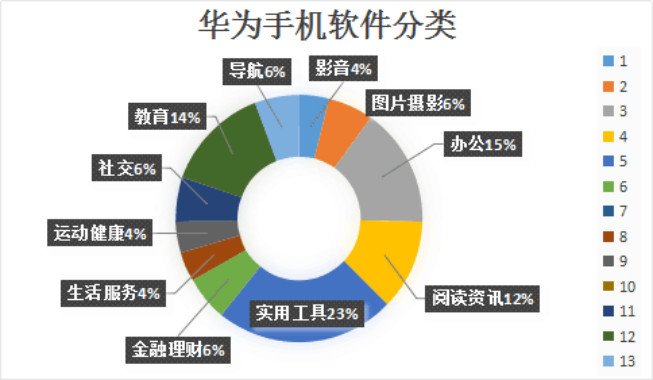


图 4　小组第一个成员手机软件分类图

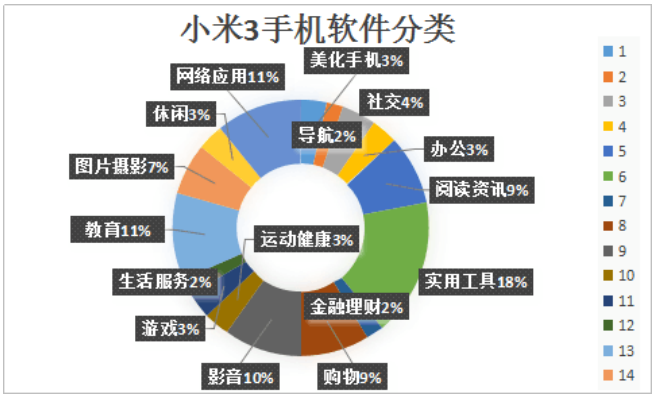


图 5　小组第二个成员手机软件分类图

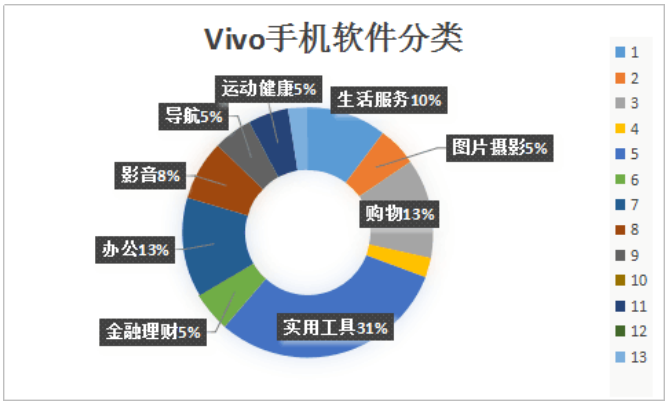


图 6　小组第三个成员手机软件分类图

通过 MATLAB 程序对小组各成员手机系统主成分分析结果如下：

表 5-10　小组各成员手机系统主成分分析

序号	特征值	贡献率	累计贡献率
1	3.731	93.2759	93.2759
2	0.269	6.7241	100
3	0	0	100
4	0	0	100

故仅需要 2 个主成分 y_1，y_2，，标准化变量的前 2 个主成分对应的特征向量(见附录）。从主成分分析系数可以看出，第一主成分主要反映了手机全球销售量比例，且比重为 0.8465，由下表可知，全球销售量比例排名和主层次分析综合排名也相同。

表 5-11　小组各成员手机系统排名情况

序号	手机类型	全球销售量比例	全球销售量比例排名	主层次分析综合评分	主层次分析综合评分排名
1	华为荣耀 6	0.084	1	1.33	1
2	小米 3	0.056	2	0.7439	2
3	VIVO-X6	0.033	3	-2.0549	3

5.5 模型的检验

这里我们采用 t 检验法。

5.5.1 问题一的检验

设总体 $X\sim N\left(\mu,\sigma^2\right)$，其中方差未知，是 $x_1,x_2,...,x_n$ 来自 X 的一个样本，现对均值μ 提出假设，此时正态分布的标准差σ 未知，因而用S^2代替，从而检验统计量服从的分布与 Z 检验统计量服从的分布不同，而且拒绝域的临界点也不一样。这里使用双边检验。本题需检验假设：

H_0：μ =μ_0=a_{ij} ，其中 i=1，2，…，14 ；　j=1，2，…，6。

$$H_1：\mu \neq \mu_0$$

检验的拒绝域为：

$$|t|=\left|\frac{\bar{x}-\mu 0}{\frac{s}{\sqrt{n}}}\right|\geq t_{\frac{\partial}{2}}(n-1)$$

查表得：

$$t_{\frac{\partial}{2}}(n-1)=3.1824$$

将上表中各累计三个月的得分做为样本取值，以排名第 1 的世纪佳缘举例子均值=0.719107，方差=0.05986，t=0.988229<3.1824，故接受 H0。

同上，14 个网站的综合得分（见附录）经计算均合理。

六、模型的评价

6.1 模型的优缺点分析

6.1.1 模型的优点

1、第一问建立模型时考虑多个因素更加合理，然后利用聚分析法和主成分分析法剔除掉相关性较小的指标，计算严谨，且模型的结果与事实吻合度很高。

2、第二问采取综合打分制可以算出各系统的安全得分，简洁明了。

3、第三问采用手机验证，证明模型是合理的。可见，模型适用性很广。

6.1.2 模型的缺点

1、数据选取的较少。

2、主要分析了网站和手机系统，没有考虑其他的信息系统。

七、模型的推广

模式识别是人类的一项基本智能，在日常生活中，人们经常在进行“模式识别”。随着 20 世纪 40 年代计算机的出现以及 50 年代人工智能的兴起，人们当然也希望能用计算机来代替或扩展人类的部分脑力劳动。(计算机)模式识别在 20 世纪 60 年代初迅速发展并成为一门新学科。

模式识别是指对表征事物或现象的各种形式的(数值的、文字的和逻辑关系的)信息进行处理和分析，以对事物或现象进行描述、辨认、分类和解释的过程，是信息科学和人工智能的重要组成部分。

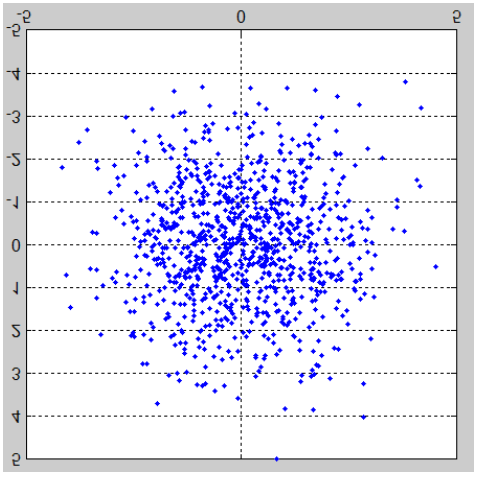
高斯概率密度函数以其数学的灵活性，易处理性和中心极限定理被广泛应用于模式识别领域，中心极限定理表示大量的随机独立变量分布趋于高斯分布函数在实际应用中，只需产生随机数数量足够大。

多维高斯概率密度分布函数如下

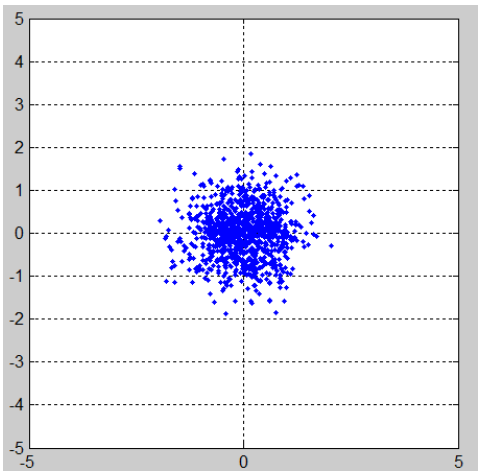
$$p(x)=\frac{1}{(2\pi)^{1/2}|S|^{1/2}}\exp\left(-\frac{1}{2}(x-m)^T S^{-1}(x-m)\right)$$

例如随机 1000 个二维数据点，且服从高斯分布 $N(m,S)$，期望 $m=\begin{bmatrix}0\\0\end{bmatrix}$，协方差矩阵为 $S=\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}$，可计算任意分布情况（工况）。

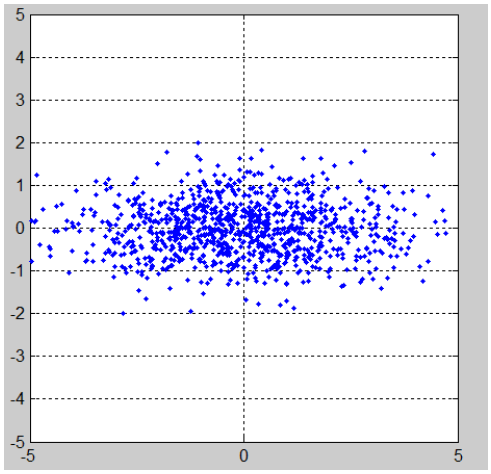
（1） $\sigma_1^2 = \sigma_2^2 = 2$ $\sigma 12 = 0$



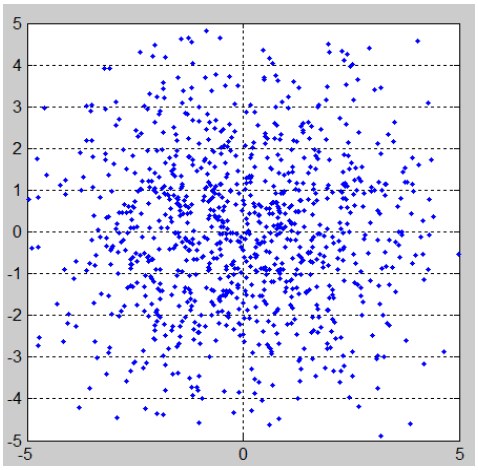
（2） $\sigma_1^2 = \sigma_2^2 = 0.4$ $\sigma 12 = 0$



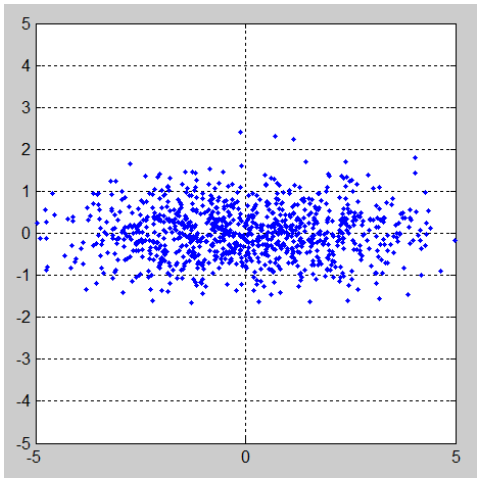
（3） $\sigma_1^2 = 0.4$ $\sigma_2^2 = 4$ $\sigma 12 = 0$



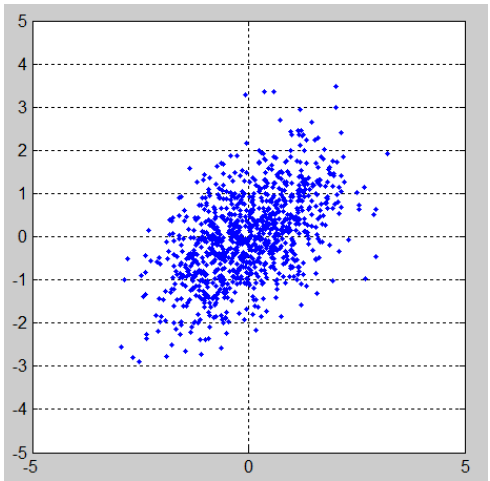
（4） $\sigma_1^2 = \sigma_2^2 = 4$ $\sigma 12 = 0$



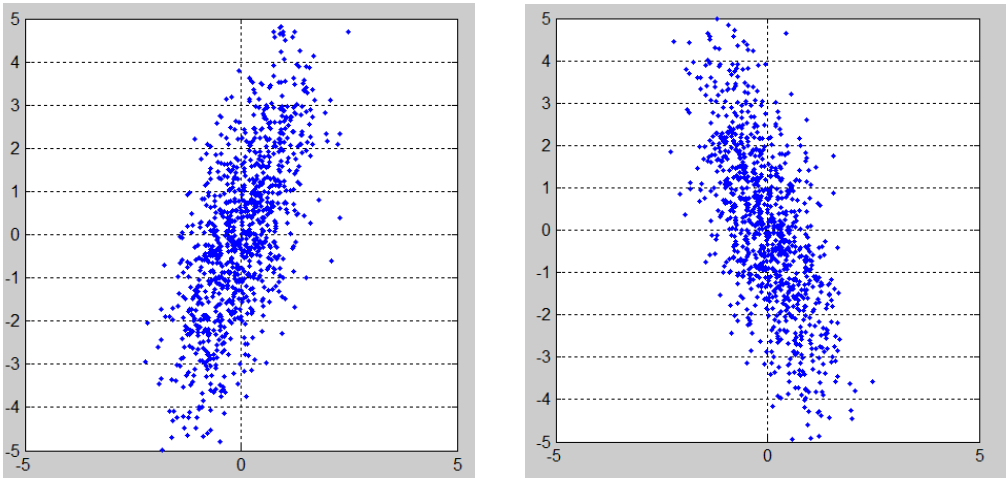
（5） $\sigma_1^2 = 4$ $\sigma_2^2 = 0.4$ $\sigma 12 = 0$



（6） $\sigma_1^2 = \sigma_2^2 = 2$ $\sigma 12 = 1$



（7） $\sigma_1^2 = 0.6$ $\sigma_2^2 = 4.0$ $\sigma 12 = 1.0$ （8） $\sigma_1^2 = 0.6$ $\sigma_2^2 = 4.0$ $\sigma 12 = -1.5$



考虑一个三维空间，分类数为 2，分别为 w_1, w_2，均服从高斯分布 $m_1 = [0,0,0]^T$ 和 $m_2 = [1.2,1.2,1.2]^T$，假设两个分类类别是等概率的，协方差矩阵为:

$$S = \begin{bmatrix} 1.1 & 0.063 & 0.063 \\ 0.063 & 0.74 & 0.063 \\ 0.063 & 0.063 & 0.74 \end{bmatrix}$$

进而得出三维空间的距离分布

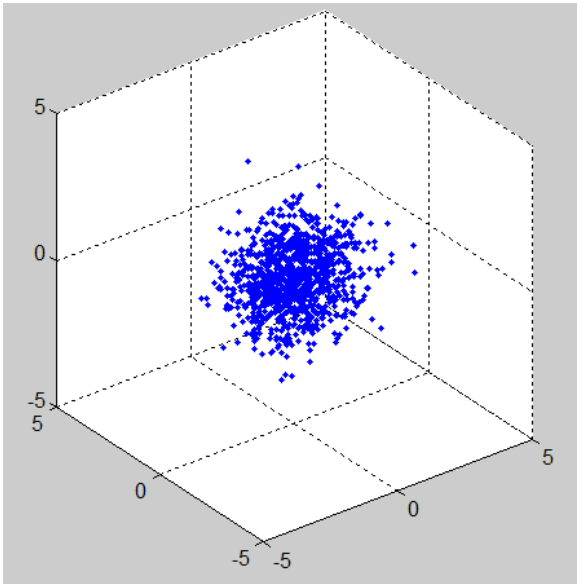


图 7　三维空间距离分布图

由于文章篇幅有限和时间限制，模型的进一步推广会想像多维向量，进而预测 7 维空间。

考虑到很多模式识别案例，就目标进行分类，设定 $x \equiv [x(1), x(2), ..., x(l)]^T \in R^l$，分类数为 c 也就是分为 $w1, w2, ..., wc$ 。

基于上文中我们所做的 R 聚类分析，我们设计一个分类器，考虑一个待分类的目标，任务是将这个目标分为 c 类，分类数 c 事先作为一个先验值，即已知值，

每一个待分类的目标由一组特征值 x（i）表示，i=1,2,3，...，1，即构成一个 l 维特征向量：

$$x=[x(1),x(2),...,x(l)]^T \in R^l$$

假设每一个待分类的目标能够由一个简单的特征向量表示，也就是该组特征向量只属于某一类。分类器如下:

考虑上文三维空间，待分类 $x=\begin{bmatrix}0.34\\0.25\\0.34\end{bmatrix}$，我们用 MATLAB 程序设计了一个欧氏距离分类器，输入为列向量，待分类的数据，输出为属于哪一类的标签，由结果可知，待分类 x 隶属于 w1。分类器见附录。较实用，适用于生活中各个案例。

八、参考文献

[1]李钊.基于复杂网络的复杂信息系统网络拓扑安全性研究.2014（5）
[2]http://www.chinaz.com/news/2016/0301/508971.shtml
[3]黄益民，平玲娣，潘学增.信息安全模型的研究及安全系统方案设计.2001
[4]李鹤田，刘云.信息系统安全风险评估研究综述.2006(1)
[5]李守鹏，孙红波.信息系统安全模型研究.2003(10)
[6]梁艳.智能手机软件漏洞现状.科技创新导报
[7]邓继胜.Web 网站安全胖鱼系统的研究与应用.2008(10)
[8]刘玉玲，冯登国，连一峰.基于时空维度分析的网络安全态势预测方法
[9]刘力维，李建军，陆中胜，谢建春.概率论与数理统计.高等教育出版社
[10]司守奎，孙玺菁. 数学建模算法与应用[M]. 北京:国防工业出版社
[11]姜启源，谢金星，叶俊.数学建模（第四版）.高等教育出版社

附录

1.求解相关系数、指标间距离和 r 型聚类的 MATLAB 程序

```
clc,clear;
a=textread('F:\char1.txt');
corrcoef(a);
d=pdist(a','correlation');
z=linkage(d,'average');
h=dendrogram(z);
set(h,'color','k','linewidth',1.3);
T=cluster(z,'maxclust',4)
for i=1:4
tm=find(T==i)
tm=reshape(tm,1,length(tm))
fprintf('%d categary has %s\n',i,int2str(tm))
```

end

2.t 检验中 14 个网站的综合得分

14739059	4	268044000	0.06976	2330	291
3842948	12	690360000	0.16216	2330	291
14739059	9187	444000	0.16666	2330	291
15204799	4615	8325000	0.16216	2330	291
3982408	757	19476000	0.16666	543	267
13297141	232	3699000	0.16216	2330	291
14739059	103	115350000	0.06976	2330	291
6034680	70965	738000	0.14173	543	267
6034680	2479	5310000	0.16216	549	563
4318928	46856	27285	0.06976	2330	291
13297141	2232	13398000	0.16666	2330	291
6083860	48403	198000	0.14173	2330	291
9710674	4749	4644000	0.06976	1169	112
1924897	57854	108000	0.16216	383	961
14739000	4	268042000	0.06976	2330	291
3843048	12	690360100	0.16216	2330	291
14739049	9176	444700	0.16666	2330	291
15204789	4609	8323000	0.16216	2330	291
3982418	769	19476200	0.16666	543	267
13297151	245	3698900	0.16216	2330	291
14739060	101	115350120	0.06976	2330	291
6034674	70980	738200	0.14173	543	267
6034680	2481	5310000	0.16216	549	563
4318928	46856	27285	0.06976	2330	291
13297141	2241	13398000	0.16666	2330	291
6083860	48390	198000	0.14173	2330	291
9710674	4756	4644000	0.06976	1169	112
1924897	57805	108000	0.16216	383	961
14741059	4	268044100	0.06976	2330	291
3843148	13	690361020	0.16216	2330	291
14738861	9086	443980	0.16666	2330	291
15207599	4730	8324900	0.16216	2330	291
3981408	747	19476020	0.16666	543	267
13296141	262	3698990	0.16216	2330	291
14739659	103	115350000	0.06976	2330	291
6033580	70985	738010	0.14173	543	267
6074681	2469	5310060	0.16216	549	563
4318928	47006	27375	0.06976	2330	291
13287144	2264	13398000	0.16666	2330	291
6083850	49063	198000	0.14173	2330	291
9710669	4751	4643900	0.06976	1169	112

1924889　　57865　108900　　0.16216　383　　961

3.问题三前 2 个主成分对应的特征向量

-0.5176　-0.0408　0.8465　-0.1179

-0.5176　-0.0408　-0.4220　-0.7432

0.4763　-0.7558　0.1986　-0.4030

0.4872　0.6522　0.2568　-0.5209

4. >> clear clc close all

```
warning off
feature jit off
randn('seed',0);
m=[0,0]';
S=[0.6 1;1 4];
N=1000;
X=mvnrnd(m,S,N)';
figure(3),plot(X(1,:),X(2,:),'.');
axis equal
grid on
axis([-5 5 -5 5])
>> clear clc close all
warning off
feature jit off
randn('seed',0);
m=[0,0]';
S=[0.6 -1;-1 4];
N=1000;
X=mvnrnd(m,S,N)';
figure(3),plot(X(1,:),X(2,:),'.');
axis equal
grid on
axis([-5 5 -5 5])
```

```
function [z]=euclidean_classifier(m,X)
%实现欧式距离分类器设计
%输入：
%m:列向量，均值向量，每一列表示待分类数据的均值向量
%X：每一列表示待分类的数据
%输出
%z：输出属于哪一类的标签
[l,c]=size(m);
[l,N]=size(X);
for i=1:N
    for j=1:c
```

```
        de(j)=sqrt((X(:,i)-m(:,j))'*(X(:,i)-m(:,j)));
    end
    [num,z(i)]=min(de);
end
close all
>> warning off
>> feature jit off
>> x=[0.34 0.25 0.34]'
m1=[0,0,0]'
m2=[1.2 1.2 1.2]'
z=euclidean_classifier(m,x)
```