

主成分分析法在软件静态测试中的研究与应用

余为峰, 黄松

(解放军理工大学 指挥自动化学院 军用软件测评中心, 江苏 南京 210007)

摘要:随着信息化程度的不断提高以及人们对软件需求的扩大,软件的复杂性也已经远远地超出了以前的水平,大大地增加了软件设计和开发的难度。以软件复杂性为出发点,介绍了主成分分析法(PCA)的基本思想、原理和主要作用,分析了主成分分析法在软件静态测试中的应用价值与可行性,最后通过一个具体的软件进行了详细的算例分析,获得了较好的效果,帮助软件开发人员和测试人员在静态分析中识别复杂性和风险性比较高的函数和模块起到了很好的作用。

关键词:主成分;主成分分析法;静态分析

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2011)06-0073-04

Research and Application of Principal Component Analysis to Software Static Testing

YU Wei-feng, HUANG Song

(Software Test and Evaluation Center for Military Training, Institute of Command Automation,
PLA University of Science & Technology, Nanjing 210007, China)

Abstract: As the degree of informationization is enhancing continually and people need more and more demands, software also becomes more complicated than before, thus, it consumedly increases the difficulty in the design and development of software. Firstly, it introduces the basic idea, principle and function of principal component analysis method, and then analyses application value and feasibility of principal component analysis method to software static testing. Finally, it performs a detailed case in terms of the method, gets the good result, the method helps the developers and testers achieve some ability to identify some function and module with high complexity and risk in the software static analysis.

Key words: principal component; principal component analysis; static analysis

0 引言

软件是一种特殊的产品,其开发与生产有着自身的规律和特殊性。与其它工程项目相比,软件项目具有更多的不确定性和复杂性。其中代码级设计与分析的不完整性与复杂性是造成软件外在复杂性的的重要因素。

随着软件外包业的发展,它有别于软件产品开发,客户对于产品的要求不再局限于系统是否能够正确运行,而是在设计、代码的品质上也有了更多的要求。有的客户甚至会在产品交付后先来检查代码品质,只要是代码不符合要求就会被拒绝。

但在项目的实际执行中,面对客户的这些要求,常常遇到诸如编写的代码不符合规范、编码效率低、代码层次结构较复杂、代码错误多、代码难以维护等现象,

从而影响到项目的时程和交付的品质,影响到客户的满意度和对专业程度的质疑。

主成分分析法是一种多元统计分析技术,广泛地应用于经济、社会、科教、环保等领域中众多对象的评价和排序^[1]。其中在软件风险度和评价领域,文献[2]曾使用主成分分析法对面向过程开发的软件产品进行了风险评估,并取得了较好的效果。文中利用主成分分析法,通过分析软件产品复杂性等度量指标对软件产品和代码进行评估,采用的是一种客观的评估方法。该方法能在软件开发与静态测试过程中帮助开发者或测试者识别软件的高风险、高复杂性函数和模块,从而有效地提高软件的质量。

1 主成分分析法的主要思想和主要原理

1.1 主成分分析法的主要思想

主成分分析法是一种将多维因子纳入同一系统中进行定量化研究且理论较完善的多元统计分析方法,在解决很多实际问题时取得了较好的效果^[3]。在实证

收稿日期:2010-11-10;修回日期:2011-02-17

基金项目:国家863计划(2009AA01Z402)

作者简介:余为峰(1985-),男,硕士研究生,研究方向为软件测试;
黄松,教授,博士,硕士生导师,研究方向为系统仿真、软件测试。

问题研究中,为了全面、系统地分析问题,必须考虑众多影响因素。这些涉及的因素一般称为指标,在多元统计分析中也称为变量。因为每个变量都在一定程度上反映了所研究问题的某些信息,并且指标之间彼此有一定的相关性,因而所得的统计数据反映的信息在一定程度上有重叠。在用统计方法研究多变量问题时,变量太多会增加计算量和增加分析问题的复杂性,人们希望在进行定量分析的过程中,涉及的变量较少,得到的信息量较多。主成分分析法正是适应这一要求产生的,是解决这类问题的理想工具。

1.2 主成分分析法的主要原理

主成份分析法是通过研究指标体系的内在结构关系,从而将多个指标转化为互不相关的、包含原来指标大部分信息的少数几个指标,即主成份^[4]。它是一种降维的统计方法,它借助于一个正交变换,将其分量相关的原随机向量转化成其分量不相关的新随机向量,这在代数上表现为将原随机向量的协方差阵变换成对角形阵,在几何上表现为将原坐标系变换成新的正交坐标系,使之指向样本点散布最开的 p 个正交方向,然后对多维变量系统进行降维处理,使之能以一个较高的精度转换成低维变量系统,再通过构造适当的价值函数,进一步把低维系统转化成一维系统。

2 主成分分析法在软件静态分析中的可行性分析

软件静态分析是一种不通过执行程序而进行测试的技术^[5],它主要可以通过代码检查和静态结构分析两种途径来实现^[6]。当对两个文本在不同的语义级别上执行时,例如一个程序针对其规格文档,可以使用静态分析技术对程序的完整性和正确性进行评价^[7]。代码审查作为一种有效方法,主要是对代码与设计的一致性、代码对标准的遵循性、代码逻辑表达的正确性以及代码结构的合理性进行检查,其主要目的是发现模块内部的错误。静态结构分析是一种机械化的特征分析方法,其关键功能是检查软件结构的表示和描述是否一致。客观的软件静态测试与分析方法是基于软件度量的,常常需要借助软件工具进行,一般包括代码分析、控制流分析、数据流分析、接口分析、表达式分析,这些度量数据本身并不能反映软件的质量,必须对这些度量数据进行分析处理,从中提取有用的信息,对影响质量的程度给予评价。

在进行基于度量的软件静态测试和分析时,可利用软件度量工具收集一些程序中关于各类函数的原始数据。这时,由于采用了不同的度量指标,各个函数的指标之间必然包含着关于软件产品的重叠信息。为使评估指标体系能够准确、科学地反映被评估对象的本

质属性,按照建立评估指标体系的原则和程序,形成评估指标体系后还应对指标进行筛选、简化,使其成为最佳指标集^[8]。因此需要对这些度量指标进行降维,提取其主要度量指标,利用主要度量指标的信息来对软件的相关函数或者模块进行处理,识别出高复杂性、高风险的函数类,便于开发者和测试者进行决策。而主成分分析法正是基于这样一种实际问题,利用降维的思想,把多指标转化为少数几个综合指标,从而降低分析的难度。

3 主成分分析法的主要步骤

主成分分析法的主要步骤如下^[9]:

设通过软件度量工具获得一个 $n \times p$ 阶的度量数据矩阵,其中 p 是选取的指标数, n 是针对指标 p 收集的样本数目。

步骤 1 构造样本阵,对样本阵元进行如下标准化变换:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, 2, \dots, n; j = 1, 2, \dots, p; \text{其中 } \bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n}, s_j^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}, \text{得标准化阵 } Z.$$

步骤 2 对标准化阵 Z 求相关系数矩阵: $R =$

$$[r_{ij}]_{p \times p} = \frac{Z^T Z}{n-1}. \text{其中, } r_{ij} = \frac{\sum_{k=1}^n z_{ik} z_{kj}}{n-1}, i, j = 1, 2, \dots, p.$$

步骤 3 解样本相关矩阵 R 的特征方程 $|R - \lambda I_p| = 0$ 得 p 个特征根,并对 λ 的各个值进行由大到小的排序。按 $\frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^p \lambda_j} \geq 0.85$ 确定 m 的值,使信息的

利用率(贡献率)达到 85% 以上,对每个 $\lambda_j, j = 1, 2, \dots, m$, 解方程组 $Rb = \lambda_j b$ 得单位特征向量 $b_j^0 = b_j / \|b_j\|$ 。

步骤 4 将标准化后的指标变量转换为主成分 $u_j = z_i^T b_j^0, j = 1, 2, \dots, m, i = 1, 2, \dots, n$, 得主成分决策矩阵为 $U = (U_1, U_2, \dots, U_m)$, 其中 $U_i = (u_{1i}, u_{2i}, \dots, u_{ni})^T, j = 1, 2, \dots, m, U_1$ 称为第一主成分, U_2 称为第二主成分, \dots, U_p 称为第 p 主成分。

步骤 5 对 m 个主成分进行综合评价:

对 m 个主成分进行加权求和,即得最终评价价值。最后根据主成分的分量的值的情况来确定静态代码中复杂性和风险性比较高的函数并进行相应的分析和决策。

4 主成分分析法的应用算例分析与研究

文中针对软件代码层上的静态分析,依据 Logscope^[10] 和 MASIA^[11] 给出的度量元指标体系,提出了某

编辑软件程序的总共包含 11 个元素的度量元指标集,具体指标如表 1 所示。

表 1 某编辑软件程序的度量元指标集

序号	度量元	度量元含义
1	NI	程序中操作符总数
2	VOCAB	程序中的词汇量
3	LENGTH	程序的长度
4	VG	圈复杂度
5	LEVEL	层次总数
6	INDCALLS	间接调用数
7	HIER-CPX	层次复杂度
8	STRU-CPX	结构复杂度
9	NBCALLING	调用的个数

下面选择某系统工具对该软件程序进行度量所获得的数据为例进行说明。表 2 给出的是该程序的原始数据,其中 $n = 13, p = 9$ 。针对原始数据,现在对软件进行复杂性等特性的分析。

根据具体主成分分析法的算法步骤,利用 MATLAB 计算工具以及相关矩阵解算函数^[12]求得标准化矩阵如表 3 所示。

表 2 该软件程序的原始度量数据

函数名称	NI	VOCAB	LENGTH	VG	LEVEL	INDCALLS	HIER-CPX	STRU-CPX	NBCALLING
Initial_SC	11	41	125	13	2	5	1	1	3
Arith_CCode2	15	51	155	9	2	7	3	2	2
Arith_CCode3	6	12	32	5	1	2	1	1	0
Backup_SC	3	22	27	1	7	1	1	0.5	1
Check_SC_Process	2	23	19	3	1	1	2	0	0
Error_Process	1	29	15	2	1	1	1	0.5	1
Time_Process	1	10	23	1	1	2	1	1	0
Status_Check	12	39	131	11	3	6	3	2	1
Backup_Process	5	15	34	5	1	1	1	1	0
Caculate	2	9	23	1	1	1	3	0.5	1
BCommand_Errors	1	13	51	2	1	1	1	0.5	1
Check_SC	16	56	231	9	3	9	2	1	2
Broadcast_Event	3	8	32	1	1	3	1	0.5	0

表 3 原始数据矩阵的标准化矩阵

0.902894	0.794348	0.816309	1.914395	0.045091	0.698635	-0.70757	0.197958	2.1769
1.625209	1.588697	1.254219	0.975258	0.045091	1.425215	1.592027	1.913591	1.128763
0	-0.79435	-0.54121	0.036121	-0.54109	-0.39124	-0.70757	0.197958	-0.96751
-0.54174	-0.18331	-0.6142	-0.90302	2.976014	-0.75453	-0.70757	-0.65986	0.080626
-0.72232	-0.12221	-0.73097	-0.43345	-0.54109	-0.75453	0.44223	-1.51768	-0.96751
-0.90289	0.244415	-0.78936	-0.66823	-0.54109	-0.75453	-0.70757	-0.65986	0.080626
-0.90289	-0.91656	-0.67258	-0.90302	-0.54109	-0.39124	-0.70757	0.197958	-0.96751
1.083473	0.855452	0.903891	1.444827	0.631276	1.061925	1.592027	1.913591	0.080626
-0.18058	-0.61104	-0.51202	0.036121	-0.54109	-0.75453	-0.70757	0.197958	-0.96751
-0.72232	-0.97766	-0.67258	-0.90302	-0.54109	-0.75453	1.592027	-0.65986	0.080626
-0.90289	-0.73324	-0.26387	-0.66823	-0.54109	-0.75453	-0.70757	-0.65986	0.080626
1.805788	1.894215	2.36359	0.975258	0.631276	2.151795	0.44223	0.197958	1.128763
-0.54174	-1.03876	-0.54121	-0.90302	-0.54109	-0.02795	-0.70757	-0.65986	-0.96751

表 4 样本相关系数矩阵

1	0.873528	0.942776	0.883272	0.246989	0.945769	0.484468	0.7487	0.66245
0.873528	1	0.889404	0.774695	0.331317	0.858337	0.450815	0.559102	0.747193
0.942776	0.889404	1	0.803138	0.240351	0.966427	0.441107	0.619014	0.720457
0.883272	0.774695	0.803138	1	0.112925	0.782963	0.342634	0.69716	0.694089
0.246989	0.331317	0.240351	0.112925	1	0.196574	0.034564	0.11604	0.303262
0.945769	0.858337	0.966427	0.782963	0.196574	1	0.465907	0.681205	0.637071
0.484468	0.450815	0.441107	0.342634	0.034564	0.465907	1	0.480512	0.26266
0.7487	0.559102	0.619014	0.69716	0.11604	0.681205	0.480512	1	0.357338
0.66245	0.747193	0.720457	0.694089	0.303262	0.637071	0.26266	0.357338	1

表 5 主成分值矩阵

函数名	主成分 1	主成分 2	主成分 3
Initial_SC	2.549252	0.858695	-1.75009
Arith_CCode2	3.958135	-0.83528	0.618848
Arith_CCode3	1.115622	-0.59081	-0.39071
Backup_SC	1.135865	2.925198	1.356812
Check_SC_Process	1.811977	-0.37226	0.250815
Error_Process	1.514528	0.24346	-0.5994
Time_Process	1.919019	-0.50349	-0.07752
Status_Check	3.094646	-0.80531	1.190248
Backup_Process	1.248303	-0.53613	-0.37033
Caculate	1.439492	-0.87692	1.098759
BCommand_Errors	1.679253	0.124493	-0.629
Check_SC	4.153477	0.657583	-0.40136
Broadcast_Event	1.891356	-0.28923	-0.29707

距,需要重新对它们进行设计或是修改,以达到规范性和有效性。

5 结束语

主成分分析法是一种降维降难度的多元统计分析方法,在解决实际的问题上已经取得了很好的效果。文中将主成分分析法运用到静态测试及其分析过程中,根据数据处理结果达到了一定效果,但这还只是在该方向的一个初探,如果指标、函数对象的选取以及方法的运用都满足了更高的要求,分析结果也将会更准确、更有针对性。

参考文献:

- [1] 秦寿康. 综合评价原理与应用[M]. 北京:电子工业出版社,2003:132-135.
- [2] Donald E N. Software project risk analysis models with appli-

(上接第 72 页)

进的运动估计算法。改进算法主要采用了更好的提前中止策略,即设置了动态的门限阈值,以及充分利用了视频序列的空间和时间相关性,利用相邻宏块的运动矢量来对块进行运动类型划分,以采用不同的搜索策略对宏块进行起始点预测。试验结果表明,在图像质量稍有提高的情况下,与原来的 MVFAST 算法相比,改进的算法能有效地提高编码速度。

参考文献:

- [1] 毕厚杰. 新一代视频压缩编码标准-H. 264/AVC[M]. 北京:人民邮电出版社,2005.
- [2] 姚晨,沙济彰. 用于块匹配运动估计的 SGDS 算法[J]. 计算机与现代化,2006(1):8-12.
- [3] Koga T, Linuma K, Hirano A, et al. Motion compensated inter-frame coding for video conferencing[C]//in: Proc. Nat. Telecommun. Conf. . New Orleans, LA: [s. n.], 1981.
- [4] Li R, Zeng B, Liou M L. A new three-step search algorithm for block motion estimation[J]. IEEE Trans. on Circuits and Sys-

tem for embedded systems[D]. USA: Wayne State University, 1999.

- [3] 刘小楠,崔巍. 主成分分析法在汾河水水质评价中的应用[J]. 中国给水排水,2009,25(18):105-106.
- [4] 戴毅,霍佳震,张倩. 基于模糊层次综合方法的企业内部风险评价[J]. 同济大学学报(自然科学版),2008,36(6):866-867.
- [5] Hausen H. Comments on practical constraints of software validation techniques[C]// Proceedings of symposium on software validation. [s. l.]: [s. n.], 1984:323-333.
- [6] 柳纯录,黄子河,陈绿萍. 软件评测师教程[M]. 北京:清华大学出版社,2005:160-175.
- [7] Sneed H. Software-testen-state of the art[C]// Software Entwicklungs-System und Werkzeuge, 2 Kolloquium. [s. l.]: [s. n.], 1987:8-10.
- [8] 赵强,康建设,罗武,等. 基于主成分分析法的新装备保障能力评估指标体系[J]. 四川兵工学报,2009,30(8):94-95.
- [9] 王威,张鑫,胡笑涛. 基于主成分分析法的灌区地下水资源承载力评价[J]. 水利与建筑工程学报,2008,8(1):5-6.
- [10] Aggarwal K K, Singh Y. Software Design Metrics for Object-Oriented Software[J]. Journal of Object Technology, 2006, 6(1):121-136.
- [11] Morasca S. Software Measurement: State of the Art and Related Issues[M]. Rovereto, Italy: School of the Italian Group of Informatics Engineering, 1995.
- [12] 苏金明,阮沈勇. MATLAB 实用教程[M]. 第 2 版. 北京:电子工业出版社,2008:133-134.
- [13] Zhu S, Ma K K. A New Diamond Search Algorithm for Fast-Block - matching Motion Estimation[J]. IEEE Trans. Image Processing, 2000, 9: 287-290.
- [14] Zhu Ce, Lin Xiao, Chau Lap Pui. Hexagon-based Search Pattern for Fast Block Motion Estimation[J]. IEEE Trans. Circuits System Video Technology, 2002, 12: 349-355.
- [15] Hosur P I, Ma K K. Motion Vector Field Adaptive Fast Motion Estimation[C]//ICICS'99. Singapore: [s. n.], 1999.
- [16] 哈力旦. 一种改进的运动矢量编码方法[J]. 西安电子科技大学学报,2005(4):60-64.
- [17] 于飞,黄士坦. H. 264 运动估计算法分析[J]. 计算机技术与发展,2009,19(4):115-118.
- [18] 宁矿凤,王小玲. MPEG-4 编码中运动估计和补偿算法研究[J]. 计算机技术与发展,2007,17(6):101-106.
- [19] 李炜,周兵,李波. 运动矢量场自适应搜索算法[J]. 计算机学报,2003,26(2):1-6.
- [20] Hsieh C H, Lu P C, Shyn J S, et al. Motion estimation algorithm using interblock correlation[J]. Electronics Letters, 1990, 26(5):276-277.