第16卷 第2期
2006年2月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 2
Feb. 2006

软件度量中的数据分析技术

丁剑洁,鱼 滨,侯 红,钟 欣
(西北大学 计算机科学系,陕西 西安 710069)

摘 要:从统计学的角度出发,分析比较了软件度量常用的数据分析技术以及它们的异同,对影响数据分析的因素也作了进一步说明。为软件度量实践中正确地选用数据分析技术提供指导,从而为软件开发的管理决策、项目过程监控提供了客观有效的支持。

关键词:软件度量;统计方法;数据分析;相关性分析;控制图
中图分类号:TP311.52 **文献标识码**:A **文章编号**:1005-3751(2006)02-0191-03

The Data Analysis Techniques in Software Measurement

DING Jian-jie, YU Bin, HOU Hong, ZHONG Xin
(Dept. of Computer Sci., Northwest University, Xi'an 710069, China)

Abstract: From the view of statistical method, the different type of analyzing methods of software measurement is described and compared. Moreover the influence factor of data analyzed is explained in this paper. It helps us to choose right data analysis techniques in practice of software measurement, and support us to monitoring progress and decision making.

Key words: software measurement; statistical method; data analysis; correlation analysis; control chart

0 引 言

软件度量的目的是用软件度量学的方法来科学地评价软件质量,更有力地对软件开发过程进行控制和管理,合理地组织和分配资源,制定切实可行的软件开发计划,降低成本获得高质量软件[1]。要能真正达到软件度量的目的,数据分析是一项重要环节。数据分析在数学、物理学等众多领域中已经成熟应用。但是在软件工程中,软件度量本身是一门重要并有待研究的学科,要实现量化管理,应用哪些数据技术、如何应用是软件工程实践者面临的一项重要课题。

1 软件度量中数据分析的重要地位

软件度量的目的之一就是为软件项目量化管理提供决策支持。在软件工程的活动中,要面临很多决策问题:测试团队想要知道哪种技术能帮助他们在测试中发现更多的错误;开发人员想知道哪种开发技术对目前就要进行的项目是最佳的;维护人员想要知道模块规模和缺陷数之间是否存在一定关系。只有做出正确的决策,才能达到合理地组织和分配资源,制定切实可行的软件开发计划,降低成本获得高质量软件的目的。在做出决策之前,必须实施一系列的度量活动。在度量活动中需要进行一系列数据收集的繁琐工作,然后进行数据分析,如果数据分析做的不好,就会造成大量的资源和已做工作的浪费。所以说,如何选取适当的数据分析技术,进行有效的数据分析,在软件度量中有很重要的地位。

2 常见的数据分析技术

在软件度量中,根据分析目的可以采用多种数据分析方法。简单的例如:计算数据的平均值、中值、标准差、百分比等方法研究数据的性质和特征。其次,还可以使用比较成熟的统计学中的数据分析方法,例如相关性分析,它主要用来揭示事物之间的关联性。因为软件过程是由多个事物参与、多种因素影响的过程,在作决策的时候应多方考虑,这时候可以借助相关性分析来帮助人们得到客观合理的结果。另外,还可以使用各种图表帮助分析问题,它清晰直观,在软件度量中已逐渐被广泛使用。具体请参看表1[2,3]。

2.1 Pearson 积差相关系数

软件工程中经常会面临决策问题,会有诸多因素影响决策,这些因素往往含有一定的关系,此关系可能并不能用某一确定的函数表示出来,需要用相关性分析作定量的研究,以便帮助做出正确的决策。Pearson 积差相关系数是相关性分析里重要的一种,此概念是20世纪初英国统计学家皮尔逊提出的一种计算两个变量线性相关的系数,通常用 r 或 r_{xy} 表示,其作用是考察的两个变量 y 与 x 组成的二维随机向量 (x,y) 的样本相关系数。

收稿日期:2005-05-20
作者简介:丁剑洁(1979—),女,陕西韩城人,硕士研究生,研究方向为软件工程;鱼 滨,副教授,研究方向为软件工程。

表 1 常见数据分析技术描述

数据分析方法		说明	适用范围
相关性分析	Spearman 秩相关系数	研究两个等级变量之间的相关程度，取值范围：[-1,1] 例如，将模块的规模分为五个等级，缺陷数目分为三个等级，给定两组数据来研究两个变量之间的相关性	双变量
	Kandall 和谐系数	研究多个等级变量之间的相关程度，取值范围：[0,1] 例如，一组专家按各自标准分别对一组模块的质量进行评估，由 Kandall 系数可判断出他们做出结论的一致性，也可以科学客观选出质量好的模块和有经验的专家	多变量
	Pearson 积差相关系数（积矩相关系数）	研究两个计量变量之间的相关程度，取值范围：[-1,1] 例如，根据一组模块的千行代码数和缺陷数目来研究两个变量之间是否存在相关关系	双变量
传统的 Shewhart 控制图	折线图	按时间顺序显示一个变量的变化过程	双变量
	散点图	直观描述两个变量之间的关系。	双变量
	直方图	用来显示过程结果在一个连续值域内的分配和分布情况	双变量
	因果图	用来检测和展示问题与它可能原因之间的关系图	多变量
	排列图	用来记录和分析问题或原因有关的信息工具，可识别主要原因，帮助确定先解决哪方面问题	多变量

若对(x,y)作了 n 次观测，得到 n 对数据(x_1,y_1)，…，(x_n,y_n)，则定义 r 为：

$$r=\frac{L_{xy}}{\sqrt{L_{xx}}\sqrt{L_{yy}}}$$

其中

$$L_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

$$L_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2,\bar{x}=\frac{1}{n}\sum_{i=1}x_i,\bar{y}=\frac{1}{n}\sum_{i=1}y_i$$

当 $|r|=1$ 时，x 与 y 存在完全的线性相关关系；$|r|$ 越小，x 与 y 存在线性相关程度越小；$r=0$，可以认为 x 与 y 不相关（不存在线性相关），但不等于 x 与 y 相互独立，x 与 y 之间可能存在其它形式的相关关系。在 $|r|\neq 0$ 时，$r>0$，可认为 x 与 y 正相关，$r<0$，可认为 x 与 y 负相关。

例 1：表 2 中给出的是一组模块的两个变量。一个是规模，用千行代码数（KLOC）来表示，另外还给出了每个模块在单元测试后发现的故障数（FD）。使用 Pearson 积差相关系数分析两个变量之间的相关性。

表 2 模块规模与缺陷数目度量

Module	A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	Q
KLOC(x_i)	10	23	26	31	31	40	47	52	54	67	70	75	83	83	100	110
FD(y_i)	36	22	15	33	15	13	22	16	15	18	10	34	16	18	12	20

根据上面所给公式，计算如下：

$\bar{x}=56.4$，$\bar{y}=19.7$，$L_{xy}=-1140.13$，

$L_{xx}=12697.75$，$L_{yy}=955.4375$，$r=-0.327$

计算得到两个变量的相关系数为 -0.372，即低度的负相关性。所以，可以得出结论，模块、规模和缺陷数目之间并无显著的关系。

另外，Pearson 积差相关系数所适用的变量必须是连续性变量。如果数据属于顺序刻度，则需要用 Spearman 秩相关系数、Kandall 和谐系数等分析方法来分析。

2.2 控制图

控制图是适用于区分异常或特殊原因所引起的波动和过程固有波动的一种工具。控制图在工业领域、统计质量管理中应用很广。在软件度量中，越来越多的人正在应用控制图。

控制图在软件度量中的主要用途是[4,5]：

(1)分析判断软件生产各个活动的稳定性，统计控制状态。

(2)及时发现各项活动中的异常现象和缓慢变异，预防不好的事件发生。

(3)指出各种可解决的问题，以及可能的过程改进。

(4)为评定各项活动的质量和趋势提供依据。

根据数据类型和分析目的有多种不同的控制图类型。这里列举在软件度量中有着重要作用的 X-S 控制图。其中采用了两个重要的变量：控制上线（UCL）和控制下线（LCL）。它一般根据不同的数据来源有不同的计算方法，统计学家已经将各种计算方法造成表格。在这里运用下面的计算方法。绘制控制图的主要步骤如下：

第一步：若数据是按时间收集，将数据按时间顺序在表中填好。

第二步：计算考察数据的均值、标准差、UCL、LCL。其中，UCL=平均值+2×标准差

LCL=平均值-2×标准差

第三步：根据所得数据画出控制图，标出制图人、制图日期等，使其看起来清晰完整。

第四步：分析控制图，确定异常模式或者非随机行为。

Florac 和 Carleton(1999)指出，可以采用如下 4 项检验已确定异常模式或者非随机的行为[4]：

a. 单独的一个点落在两个 3 倍标准差控制线外。

b. 3 个相邻数值中起码有两个落在中心线同侧，并且偏离中心线超过 2 倍标准差。

c. 5 个相邻数值中起码有 4 个落在中心线的同一侧，并且它们偏离中心线都超过 1 个标准差。

d. 起码有 8 个相邻数值落在中心线的同一侧。

根据出现的异常情况，查明原因提出改进方法。

第五步：剔除可疑点，重新绘制控制图，预测活动发展趋势。

例 2：表 3 中客户服务中心统计了十四周的用户提交报告数，用控制图进行分析。

表 3 用户提交报告数

周	1	2	3	4	5	6	7	8	9	10	11	12	13	14
报告数	5	4	8	7	0	12	8	4	5	5	8	9	5	7

计算得到:

均值:6.2,　标准差:2.8,　UCL:11.8,

LCL:0.6,　+1sigma:9,　-1sigma:3.4

利用这些数据,绘制控制图如图1所示。

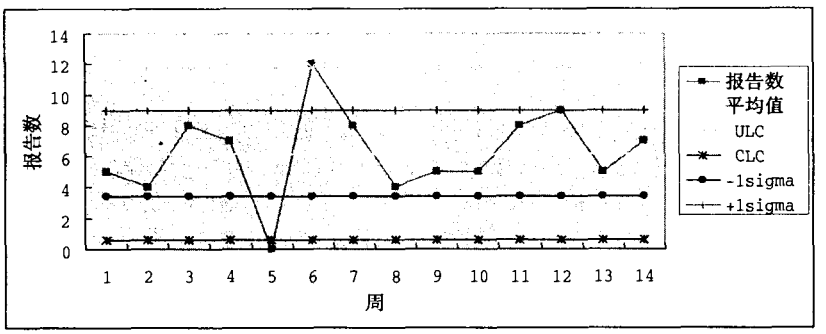

图1　用户提交报告数控制图

从这幅图中,可以看到第5周和第6周的报告数落在了控制上线和控制下线之外,即两倍的标准差之外。由此判断,可能是第5周的报告数出现问题,将其数据累加到第6周,才会出现此情况。若要预测以后提交的报告数,并适当安排时间及工作人员,就必须剔除这两个可疑点的值,重新绘制控制图。

在剔除可疑点时需要注意:

● 确保识别各种正当的特殊原因。如果为了得到"良好的控制图"而剔除可疑点,或者出于任何其他客观原因除去可疑点,这种做法只会妨碍人们做出正确的判断。

● 如果不断地发现和剔除各种特殊原因,那么应该重新确认一下,是否此过程中存在一些极为根本的原因,以至于不适合进行统计分析[4]。

3　影响数据分析的因素

从数据收集到选择分析方法以及做出决策,在这个过程中要受到很多因素的影响,所以应该很谨慎[6]。

3.1　清晰和明白地定义各种度量指标

清晰和明白地定义各种度量指标是有效进行数据分析的前提条件之一。做到这一点,才能使数据具有一致性,保证有可能得到客观公正的结果,否则,可能会起到负面作用。例如,在例2中,有一行数据KLOC表示代码的长度,所以,在采集数据之前,必须明确定义代码行数目的规则,使得数据结果只和所采用规则有关,而和实施对象、实施人员无关。

3.2　软件度量数据的分布

在得到数据集时,应该先判断测量数据的分布情况。例如,如果数据基本符合正态分布,用正态分布的各种指标来分析才是有效的。如果不符合正态分布,可利用变换,例如取对数,再利用正态分布的方法进行分析。

3.3　数据的刻度

数据的刻度分为标称刻度、序号刻度、间隔刻度、比率刻度、绝对刻度。软件测量刻度对于测量数据的运算和统计分析的类型有重要的影响。标称刻度和序号刻度不允许进行平均值、平均偏差的分析,也就是说,对于标称刻度和序号刻度而言,计算其平均值是毫无意义的。对于标称刻度任何有意义的统计分析对于序号刻度而言都是有意义的,对于序号刻度任何有意义的统计分析对于间隔刻度而言都是有意义的,对于间隔刻度任何有意义的统计分析对于比率刻度而言都是有意义的,同样,对于比率刻度任何有意义的统计分析对于绝对刻度而言都是有意义的,请参看表4[2,3]。

4　小　结

数据分析技术是软件度量中的重要环节。文中为软件工程实践者在这一环节提供了一定的帮助和指导,主要论述相关性分析和控制图两类技术在软件度量中的应用。但是,软件度量学是一门重要并且有待研究的学科,软件系统规模增大,结构日趋复杂,数据分析技术也必将随着发展,在更广阔的领域中发挥作用。

表4　测量刻度对统计方法的影响

刻度类型	可定义关系	统计方法举例	可采取检验方法
标称	等于	频率	非参数检验
序号	等于 大于	中值 百分比 Spearman 秩相关系数 Kandall 和谐系数	非参数检验
间隔	等于、大于 测量值间隔的比例	平均数 标准差 Pearson 积矩相关性 多维积矩相关性	非参数检验
比率	等于、大于 测量值间隔的比例、 测量值的比率	几何平均数 变量系数	非参数检验 和参数检验
绝对	等于、大于 测量值间隔的比例、 测量值的比率、测量值	所有统计方法	非参数检验 和参数检验

参考文献:

[1] 刑大红,曹佳冬. 软件度量学综述[J]. 计算机工程与应用,2001,27:17-19.

[2] Fenton N E. Software Metrics: A Rigorous & Practical Approach[M]. 北京:清华大学出版社,2003.

[3] 侯　红,郝克刚,郭小群. 软件测量刻度及其选择方法[J]. 计算机科学,2005(5):216-218.

[4] International Function Point Users Group. IT度量——专家实践[M]. 方德英译. 北京:清华大学出版社,2003. 419-442.

[5] Grant E L, Leavenworth R S. Statistical Quality Control [M]. 北京:清华大学出版社,2001. 117-163.

[6] Cropley D H. Towards for mulating a semiotic theory of measurement information[J]. Measurement, 1998, 24:237-262.