基于盒式图的数据过滤与回归分析算法

杜庆峰,李 岩

(同济大学 软件学院,上海 200331)

摘 要:讨论了软件度量的数据过滤和回归分析问题,提出了一种用盒式图进行数据过滤,再用回归分析得出线性回归直线的算法。

关键词:软件度量;数据清洗;回归分析;盒式图

中图分类号:TP311.5 **文献标识码**:A **文章编号**:1674-7720(2010)13-0001-02

Data filtering and regression analysis algorithm based on box plot

DU Qing Feng, LI Yan

(School of Software Engineering, Tongji University, Shanghai 200331, China)

Abstract: This paper discusses the problem of data filtering and regression analysis of software metrics and puts forward a new algorithm that using box plot to filter data, and get linear regression straight line by regression analysis.

Key words: software metrics; data filtering; regression analysis; box plot

软件度量是对软件开发项目、过程及其产品进行数据定义、收集以及分析的持续性定量化过程,目的在于对此加以理解、预测、评估、控制和改善,从而保证软件开发中的高效率、低成本、高质量[1]。但是,得到正确的度量只是测量程序的一部分。软件质量是与所收集和分析的数据质量密切相关的,数据清洗过程的目的就是要解决“脏数据”的问题。数据清洗是指去除或修补源数据中的不完整、不一致、含噪声的数据。在源数据中,可能由于疏忽、懒惰,甚至为了保密使系统设计人员无法得到某些数据项的数据[2]。根据决策系统中“garbage in garbage out”(如果输入的分析数据是垃圾则输出的分析结果也将是垃圾)原理,必须处理这些噪声数据。去掉噪声平滑数据的技术主要有分箱(binning)、聚类(clustering)、回归(regression)等[3]。本文在回归分析的基础上,加入了盒形图进行数据过滤,从而得出一条线性回归直线,使模式或者关系变得更加明显,从而用这些模式和关系对测量的属性作出判断。

1 盒形图和回归分析简介

1.1 盒形图

该方法可以描述数据集取值范围的情况,展示数据主要聚集的区域,发现离群数据可能的位置,以便于对离群数据进行处理。盒形图显示一个变量的信息,如对相同CMM等级的不同项目完成每个FP的工作量分析,根据中位数 m、上四分位数 u、下四分位数 l、盒长 d、和尾(tail)来分析。

中位数是在数据集中排列居中的项。也就是说,如果中位数取值为 m,则数据集中有一半的值大于 m,一半的值小于 m。将所有数值按大小顺序排列并分成四等份,处于三个分割点位置的得分就是四分位数。最小的四分位数称为下四分位数 l,所有数值中,有四分之一小于下四分位数,四分之三大于下四分位数。中点位置的四分位数就是中位数。最大的四分位数称为上四分位数 u,所有数值中,有四分之三小于上四分位数,四分之一大于上四分位数。也有叫第25百分位数、第75百分位数的。将上四分位数和下四分位数的距离定义为盒长 d,因此,$d=u-l$。接下来定义分布的尾(tail)。理论上,上尾值点为 $u+1.5d$,下尾值为 $u-1.5d$,这些值必须进行舍位处理,以接近真实数据,位于上尾和下尾之外的值称为离群值。

1.2 回归分析方法

回归分析方法是研究要素之间具体数量关系的强有力的工具,运用这种方法能够建立反映要素之间具体的数量关系的数学模型,即回归模型。线性回归技术的基础就是散点图。将每个属性对表示为一个数据点 (x,y),

然后用回归技术计算出能够最好地拟合这些点的直线。目标是将属性 y(因变量)根据属性 x(自变量)表示为等式:$y=a+bx$。

线性回归的理论是从每个点垂直向上或向下画一条线段到趋势直线,表示从数据点到趋势直线的垂直距离。在某种意义上,这些线段的长度表示数据和直线的差异,且这种差异应尽可能地小。因此,"最佳拟合"的直线式是指使该距离最小的直线。

在数学上要计算"最佳拟合"直线的斜率 b 和截距 a 是很简单的。每个点的差异称为残差,生成线性回归直线的公式是残差的平方和达到最小。可以将每个数据点的残差表示为:

$$r_i=y_i-a-bx_i$$

最小化残差平方和得到以下关于 a、b 的等式:

$$b=\sum(x_i-m_x)(y_i-m_y)/\sum(x_i-m_x)^2 \quad (1)$$

$$a=m_y-bm_x \quad (2)$$

m_x 是 x_i 的平均值,m_y 是 y_i 的平均值[4]。

2 算法实现

在进行数据清洗时,由于数据是无序输入的,所以先对其排序,再用盒形图法行数据清洗。以下是伪代码:

```
void BubbleSort(double m,double q,int n)  //先对输入
                                  //的数据进行冒泡排序,并相应修改
        //第二组数据的顺序,以保证它们之间的对应关系
{     for(int i=0;i<n;i++)
        for(int j=n-1;j>i;j--)
        {
         输入数据的排序
         修改第二组数据
        }
}
  void box(double *m,double *q,int &n) //盒形法筛选
//掉离群项目工作量数据,n 为输入数据个数,m、q 为指针
{
   double a,b,c,top,bottom,l;   //上分位 a,中位数 b,
                                //下分位 c
   if(n%2==0)   //计算出 3 个四分位数
   {
      b=(*(m+n/2)+*(m+n/2-1))/2;   //数据个数为
                  //偶数时,中位数取中间两数的平均值
      a=*(m+n/4);
      c=*(m+3*n/4);     }
}
else
{    b=*(m+n/2);
     a=*(m+n/4);
     c=*(m+3*n/4);     }
   l=c-a;  top=c+1.5*l;bottom=c-1.5*l;   //计算出盒
                                //长,上尾数,下尾数
   if(bottom<0)  bottom=m;   //并进行必要的舍位处理
    int  j=n;
    for(int  i=0;i<j;i++)   //判断是否为离群值,
    {
       if(*(m+i)>top || *(m+i)<bottom)
         如有,将其从数组中剔去
    }
}
```

接下来要对筛选出来的数据进行回归分析,从而得到一个数据模型。

```
void  regress(double*  m,double*  q,int  n)   //对数组
          //m 和数据 q 的数据用线性回归法进行拟合
          //并用一条直线表示出它们之间的对应关系
{   double  average_m,average_q,total_m,total_q,L_mq,
L_mm;
    double  a,b;  //拟合直线 y=a+bx 的 2 个待定系数
    for(int  i=0;i<n;i++)
    {
      //计算两组数据的和 total_m 和 total_q
    }
    average_m=total_m/n;  //求的第一组数据的平均值
    average_q=total_q/n;  //求的第二组数据的平均值
    for(int  j=0;j<n;j++)
    {
     利用公式(1)计算两组数据 m,q 它们所有数据偏
                          离程度的对应相乘之和 L_mq
    }
    for(int  k=0;k<n;k++)
    {
     计算第一组数据 m,它的所有数据偏离
     程度的平方和 L_mm
    }
    b=L_mq/L_mm;   //计算出拟合直线的待定系数
                                  //b 的拟合值
    a=average_q-b*average_m;   //利用公式(2)算出参
                                  //数 a
}
```

从而得到一条线性直线,算法结束。

3 算法在实验数据上的实现

从 SSMBSS(上海软件度量基准体系)中选取了一组数据(见表 1),首先将其用散点图列出来(见图 1),然后用盒形图进行数据清洗(见图 2),最后用回归分析得出拟合直线(见图 3)。

综上所述,对于软件度量过程中出现的数据冗余和失真的情况,可以通过数据过滤和回归分析进行处理,除去那些离群的数据,并得出相应的拟合直线,这样就可以分析出数据的规律,保证软件的质量,提高效率。

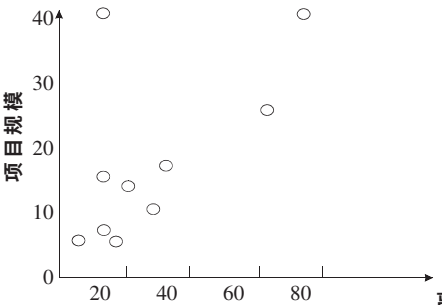

图 1 散点图

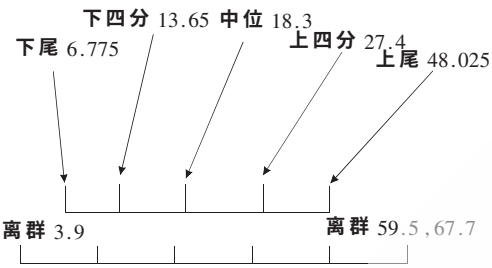

图 2 盒形图分析结果

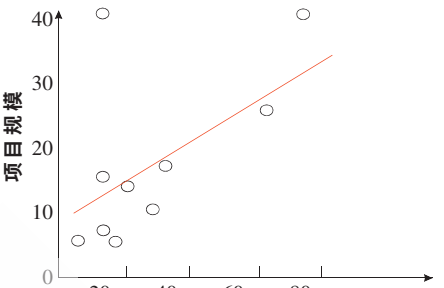

图 3 拟合直线

表 1

项目工作量/月	项目规模/代码行
16.7	6 050
22.6	8 363
32.2	13 334
3.9	5 942
17.3	3 315
67.7	38 988
10.1	38 614
19.3	12 762
10.6	13 510
59.5	26 500

参考文献

[1] FENTONenton N E, PFLEEGER S L. Softwaremetrics: a rigorous&practical apporach[M](第 2 版). 北京: 清华大学出版社, 2003.

[2] 郭志懋,周傲英.数据质量和数据清洗研究综述.软件学报[J],2002(11).

[3] 王石,李玉忱,刘乃雨,等.在属性级别上处理噪声数据的数据清洗算法.计算机工程[J],2005(5).

[4] 徐建华.现代地理学中的数学方法.北京:高等教育出版社,2002.

(收稿日期:2009-03-15)

作者简介:

杜庆峰,男,1968 年生,副教授,主要研究方向:软件项目管理与质量控制,软件测试。