模糊聚类算法在软件度量数据分析中的应用

杜星海,侯 红
(西北大学 软件工程研究所,陕西 西安 710069)

摘 要:为了提高软件质量,控制和改进软件开发过程,需要有效地度量软件开发过程和分析其过程各个阶段收集的度量数据。文中将模糊聚类算法应用到软件度量的数据分析中。先给出了数据挖掘相关知识和理论,再介绍了该算法在软件度量数据分析中应用的实验研究。由于较快地发现有严重缺陷的模块,进而提高了软件测试效率。

关键词:软件度量;软件测试;数据挖掘;模糊聚类分析
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1005-3751(2005)12-0132-03

Software Measurement Data Analysis with Fuzzy Clustering Algorithm

DU Xing-hai, HOU Hong
(Institute of Software Engineering, Northwest University, Xi'an 710069, China)

Abstract: In order to improve software quality, control and better software process, should have the measurement of software development and analyses the data that are collected at various phases of the software development process. This paper introduces fuzzy clustering algorithm into analyzing software metrics data. The knowledge of data mining is briefly described and the framework of its application is discussed, the emphasis is placed on the experiment of analyzing software metrics data. It not only improves the level of software development and test of software, but also provides software quality management a scientific foundation.

Key words: software measurement; software testing; data mining; fuzzy clustering analysis

0 引 言

软件质量是与软件产品能满足规定的或隐含需求的能力有关的特征或特征的全体,也是软件开发过程中所使用的各种开发技术和验证方法的最终体现。软件度量是一种软件质量管理重要手段之一,是不断将实际质量和预期质量相对比的过程。它应用在软件开发生命周期的各个过程中,通过对软件开发过程进行有效度量及早发现软件中缺陷多的模块,可以有效地控制花在这些模块上额外开发和维护的工作量。

数据挖掘是使用模式识别技术、统计和数学技术,自动地在大量数据中寻找预测性的信息来支持自动决策[1]。模糊聚类是一种无监督学习数据挖掘技术之一。本文讨论C-均值的模糊聚类算法在软件度量数据分析中的应用。根据度量值把软件模块聚类成几个簇。具有相似软件度量的缺陷的模块,被分在一个簇中;没有缺陷的模块被分在另一个簇中。当对所有模块应用模糊聚类算法后,可以较快地发现其中存在严重缺陷的模块,进而提高了软件测试效率。

收稿日期:2005-03-04
作者简介:杜星海(1978—),男,陕西咸阳人,硕士研究生,主要研究领域为软件度量。

1 度量数据的分析

1.1 度量数据

软件度量数据是在软件开发过程中进行度量时所收集的。一个良好的数据应具有正确性、准确性、一致性,同时与表示数据所需要的精度、相关联的特定活动或持续时间等有关。数据收集通常采用人工收集和自动收集两种方式,人工收集方式收集数据时容易发生数据的偏差、误差、遗漏等情况,而自动收集方式因采用某种成熟的算法收集数据精确度相对高。

分析度量数据时,通常采用统计技术或数学知识描述属性值的分布情况和多个属性之间的关系,然后基于这些模式或关系对软件度量的属性做出判断。统计技术有简单技术和高级技术之分,简单技术包括盒形图、三点图、控制图等;高级技术包括分类树、变换、多元数据分析、多准则决策支持等。除此,用于度量数据分析的技术还有数据挖掘技术。目前,主成分分析、神经网络分析和决策树等一些数据挖掘技术在软件度量中的应用已经研究。文中主要详细介绍模糊聚类算法在软件度量数据分析中的应用。

1.2 模糊聚类分析

一般来说,任何形式的数据分析是一种尝试回答一些具体的问题。聚类分析是将数据对象集成若干个簇,簇内的数据对象之间具有较大的相似度,而不同簇的两个数据

对象具有较小的相似度。传统意义上,聚类分析应用了典型的组理论。一个数据对象必须属于且只属于一个簇。模糊聚类允许簇相互重叠。一个数据对象可以属于几个簇,但它隶属各个簇的隶属度的总和应是 $\sum_i u_i = 1$。

模糊聚类分析的基本理论是模糊 C - 均值聚类算法[2]。函数通过迭代运算使得值达到最小。

$$J(f) = \sum_{x\in X}\sum_{k\in K} f^m(x)(k)\cdot d^2(x,k)$$

式中,f 是一个模糊区,$f(x)(k)$ 是第 k 簇中的模式 x,m 是模糊指数,$d(x,k)$ 是模式 x 和第 k 个簇中心之间的距离。FCM 是一个簇中心优化和簇成员优化相互交替的迭代优化算法,FCM 需要一个模式集(表示为向量)、一个距离值(主要采用欧氏距离)和一个该模式集应分的簇数。开始随机选取几个簇中心,计算模糊区 f。之后,簇中心不断根据 f 优化成 J,f 不断根据簇中心优化成 J,反复迭代到 f 的值低于一个极小值,或迭代次数超过已知最大值停止。

应用 FCM 算法时,先假设数据集被分成不同的簇数,以不同簇数聚类数据集之后对比其各自聚类质量,把聚类的质量最好的簇数作为所需的簇数。度量聚类的质量方法有划分系数、紧密性、分离性等。

2　模糊聚类算法在软件度量数据分析的应用实验

文中研究的对象是一个商业医学照相软件系统。这个软件几乎有近似 4000000 行的源代码,分为 4500 模块。模块中,58% 是 Pascal 编写的,29% 是 Fortran 编写的,7% 是汇编编写的。本研究的对象是 Fortran 和 Pascal 语言编写的 390 个模块。数据库设计为 390 个记录,每个记录有 12 个项,前十一个项是不同的软件度量值,第十二个项是模块修改的次数。数据库的数据是在度量这 390 个模块过程中收集的。

应用在软件开发的软件度量有:

1 源代码行数

2 可执行代码行数(非注释和空行)

3 字符的总数量

4 注释行数

5 注释字符的数量

6 代码字符数量

7 Halstead 的 n (被定义为操作符和操作数的总数)

8 Halstead 的 nh (n 的近似值)

假设数据集分簇个数范围为 2 到 10。对数据集分别以簇数为 2,3…聚类,聚类后计算各自的划分系数[3]、CS 指数[4]、分离性[5]、SSE(the average sum of squared error)等,如表 1。划分系数和 CS 指数越大或分离性和平均 SSE 值越小说明聚类质量越好。在表中,簇数是 7 时,聚类的质量最好。因此设定数据集应聚类的合适簇数是 7 个。

当数据集被聚类成 7 个后,对这几个簇分别计算各自的最小值、最大值、平均值、中值、样本标准偏差和记录个数,也包括每个簇的质心的变化,如表 2。

表 1　不同簇数的聚类有效性度量

簇数	划分系数	CS 指数	分离性	平均 SSE * 10^3
2	0.8832	0.0064	9.5549	3.3596
3	0.7275	0.0001	20.7707	2.7435
4	0.6529	0.0003	20.8820	2.7124
5	0.6031	0.0000	16.5433	2.5807
6	0.5145	0.0005	29.1250	2.5836
7	0.4865	0.0006	23.9647	2.5753
8	0.4262	0.0001	41.6580	2.5878
9	0.4060	0.0001	40.0919	2.5762
10	0.4014	0.0009	20.0038	2.5821

表 2　不同簇的属性

簇	最小值	最大值	平均值	中值	样本标准偏差	总数	质心变化
1	0	27	4.176471	2	4.676488	102	4.7474
2	0	47	21.25	16.5	12.94065	20	21.1544
3	8	41	19.31818	14	12.12391	22	16.5572
4	0	19	2.261682	1	3.10002	107	3.1178
5	14	98	36.75	32.5	21.99638	12	29.0585
6	0	25	5.325561	4	4.888021	86	6.1586
7	1	46	10.02439	7	9.379466	41	9.9213

排列每一维的簇质心变化,如表 3,会发现簇质心和簇的变化数的顺序是一样的,所有的度量值变化和质心之间是一种单调的关系[6]。

表 3　簇质心的属性的排列顺序

属性	簇的排列顺序
源代码行	4,1,6,7,3,2,5
可执行行数	4,1,6,7,3,2,5
总字符	4,1,6,7,3,2,5
注释行	4,1,6,7,3,2,5
注释字符	4,1,6,7,3,2,5
代码字符	4,1,6,7,3,2,5
Halstead N	4,1,6,7,3,2,5
Halstead Nh	4,1,6,7,3,2,5
输出:变化质心	4,1,6,7,3,2,5

对每个簇计算所有度量的最小值、最大值、平均值、中值、标准偏差,如表 4。

表 4　不同簇的度量

簇	最小值	最大值	平均值	中值	标准偏差
(a)簇 1					
代码行	22	130	71.696078	71	22.405705
可执行行	18	86	53.852941	56	16.402365
字符	511	2570	1561.598	1580	450.37072
注释	4	48	14.921569	13	7.7621655
注释字符	58	1646	592.41176	585	388.42057
代码字符	349	1424	814.12745	793	24022533
N	67	357	151.55882	149.5	48.276754
Nh	110	390	199.28039	190.15	59.222974
(b) 簇 2					
代码行	239	468	364.25	375	60.476942
可执行行	183	403	302.3	306	58.529435
字符	6013	10526	7875.65	7835.5	1155.2359
注释	26	122	72.4	69.5	21.933872
注释字符	916	4006	2138.05	2004	949.07836
代码字符	4467	7151	5466.1	5463.5	681.45285
N	646	1430	997.65	637.5	208.01576
Nh	645.5	1488.6	1079.99	1054.3	190.60697

(续表4)

簇	最小值	最大值	平均值	中值	标准偏差
(c)簇3					
代码行	209	471	287.63636	277.5	65.173428
可执行行	170	345	232.04545	220	53.301012
字符	4761	11794	6204.4091	5926	1483.5273
注释	13	102	59.95545	62	19.998755
注释字符	293	5578	2037	1808.5	1097.4097
代码字符	2762	5542	3792.6364	3878.5	782.7984
N	472	1180	704.18182	673.5	167.69606
Nh	561.3	1063	791.07273	794.95	103.97012
(d)簇4					
代码行	3	74	33.317757	30	16.547016
可执行行	2	59	23.317757	22	10.874584
字符	59	1770	676.92523	641	392.1887
注释	0	27	6.2523364	6	4.7426079
注释字符	0	1411	253.14019	195	264.06658
代码字符	30	970	336.5514	311	176.76369
N	3	160	60.308411	57	32.676635
Nh	2	221.5	92.461682	91.4	45.478983
(e)簇5					
代码行	511	944	645.75	627.5	116.65031
可执行行	401	692	513.75	493.5	88.810447
字符	9731	21266	14129.333	14687.5	3200.3
注释	68	194	115	97.5	97.5
注释字符	1792	9946	5113.0833	4974	2703.2842
代码字符	5251	10394	8139.1667	7878	1381.5789
N	943	2083	1468.3333	1479.5	305.45892
Nh	1034.8	1777.3	1343.175	1344.75	222.50694
(f)簇6					
代码行	62	545	129.61628	127	51.313837
可执行行	47	538	104.59302	102	52.8100065
字符	611	5584	2738.2209	2728.5	761.08
注释	0	67	19.639535	17	10.643875
注释字符	0	3645	1050.1395	966.5	665.65213
代码字符	240	3356	1472	1422	384.08829
N	36	445	261.25581	262	62.416285
Nh	66.6	553.1	315.78256	306.85	75.413404
(g)簇7					
代码行	111	261	186.53659	180	40.638096
可执行行	86	233	152.2439	149	35.998459
字符	2508	5883	4119.561	3922	945.11433
注释	4	86	32.146341	28	19.459652
注释字符	48	2977	1302.9512	1167	779.84986
代码字符	1290	4482	2592.9024	2484	618.44638
N	267	680	460.4878	445	108.02503
Nh	287.5	791.4	515.61707	513.7	98.394789

将表4和表1对比,发现每个簇除了有较大的度量值,还有较小的度量值。在簇4,1,6,7中,值偏高的度量个数分别是0,2,1,0;而在簇3,2和5中,值偏高的个数分别是5,3,5。因此度量值偏小的有4个簇,度量值偏高的有3个簇。与簇质心的排列顺序是4,1,6,7,3,2,5相比,可以反映值偏低的4个簇和值偏高的3个簇之间有明显区别。值偏高的簇的每一个模块的变化数都比值偏低的簇大得多,而软件度量值变化和软件模块的修改次数有线性关系,有严重缺陷的模块就可能分到这些值偏高的簇中。在风险模块分析中,这些值偏高的模块就需要引起测试重视,额外的工作量将花在这些模块上,也就不仅提高软件测试工作效率,还保证了软件质量。

3 总 结

讨论了模糊聚类方法在软件度量数据分析中的应用。期望这能从收集的软件度量数据中发现一些隐含的、以前未知的可理解的信息,为软件质量评估、软件开发过程管理和控制、资源合理组织和分配以及制定切实可行的软件开发计划、降低成本和获得高质量软件提供了科学理论依据。

参考文献:

[1] 朱三元.软件质量及评价技术[M].北京:清华大学出版社,2001.

[2] 李雄飞,李 军.数据挖掘与知识发现[M].北京:高等教育出版社,2003.

[3] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. New York: Plenum Press, 1981.

[4] Chen M S, Han J, Yu P S. Data mining: an overview from a database perspective[J]. IEEE Trans Knowledge Data Eng, 1996, 8(6): 866-883.

[5] Fayyad U M. Data mining and knowledge discovery: making sense out of data[J]. IEEE Expert, 1996, 11: 20-25.

[6] Lind R K. An experimental study of software metrics and their relationship to software errors[D]. Milwaukee: University of Wisconsin, 1986.

(上接第131页)

的灵活性优点,又能达到网络换代的自然过渡。种种迹象表明,MPLS将成为继IP技术后的新一代主流网络技术,为人们带来高性能、更如意的网络服务。

参考文献:

[1] 石晶林,丁 炜.MPLS宽带网络互联技术[M].北京:人民邮电出版社,2001.

[2] 华为-3com公司技术研发小组.MPLS技术白皮书[Z].华为-3com公司,1999.

[3] Rekhter B D Y.多协议标签交换技术与应用[M].罗志祥,朱志时等译.北京:机械工业出版社,2001.

[4] 广东省电信有限公司技术研发组.MPLS技术简介[EB/OL]. http://www.gdtel.com.cn/jsjl/jszt/2002-05-23/15.shtml,2002.

[5] Zielke G. Multi-Protocol Label Switching: A New Strategy for the NetworkBackbone[EB/OL]. http://www.infotel-systems.com,2000.

[6] Leon-Garcia A, Widjaja I.通信网——基本概念与主体结构[M].乐正友,杨为理等译.北京:清华大学出版社,2003.