

改进的 FUP 算法在五金产品质量分析系统中的应用

李松生¹, 赵燕伟², 顾熙仁²

(1. 上海大学 机电工程与自动化学院, 上海 200072; 2. 浙江工业大学 特种装备制造与先进加工技术教育部重点实验室, 杭州 310032)

摘要: 以阀门作为五金产品质量数据分析的实例, 对阀门产品质量数据进行分析, 主要是分析阀门产品的缺陷数据。针对传统的增量式关联规则算法 FUP 没有考虑到数据的时间属性, 在 FUP 算法的基础上提出一种改进算法, 并且将改进的 FUP 算法运用到产品质量分析系统中。通过实验结果对比发现, 使用了改进算法以后, 原来的许多规则已经不在生成的规则列表中出现, 而一些新的规则被生成了。

关键词: 计算机应用; 关联规则; 缺陷数据; FUP 算法; 质量分析

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5497(2012)Sup. 1-0251-04

Application of improved FUP algorithm on hardware product quality analysis system

LI Song-sheng¹, ZHAO Yan-wei², GU Xi-ren²

(1. School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200072, China; 2. Key Laboratory of Special Purpose Equipment and Advanced Manufacturing Technology, Ministry of Education, Zhejiang University of Technology, Hangzhou 310032, China)

Abstract: The valve was used as an example for hardware product quality analysis, and the defect data of valve was mainly analyzed. Traditional incremental updating algorithm FUP for association rules does not take the time properties of data into account. Therefore an improved FUP algorithm was proposed, and it was used in the analysis of product quality. By comparison of experimental results, many of the original rules do not generate in the list of rules, while the new rules are generated after using the improved algorithm.

Key words: computer application; association rules; defect data; FUP algorithm; quality analysis

0 引言

数据挖掘就是从大量数据中提取有用信息的过程。通常, 数据挖掘任务可以分为描述型任务和预测型任务^[1]。预测型任务就是根据其他属性的值预测特定属性的值, 如回归、分类、离群点检测, 其中分类分析就是通过分析示例数据库中的数据, 为每个类别做出准确的描述, 或建立分析

模型, 或挖掘出分类规则, 然后用这个分类模型或规则对数据库中的其他记录进行分类, 分类算法有多种, 本文主要描述基于关联规则的分类算法。

关联规则的经典算法是 Apriori 算法^[2-3], 它是一种最有影响的挖掘关联规则频繁项集的算法, 采用 Apriori 算法对生产过程中的历史缺陷数据进行分析。文献[4]提出一种分布式关联规则挖掘算法, 该改进的 Apriori 算法通过对频繁

收稿日期: 2012-02-09.

基金项目: 国家自然科学基金项目(60970021).

作者简介: 李松生(1959-), 男, 高级工程师, 博士。研究方向: 机械设计及理论。E-mail: lisongsheng111@sina.com

项集阈值的设置,减少中间候选项集的数量,降低算法复杂度,提高算法执行效率,能解决传统关联分类算法中存在的冗余和冲突规则问题;文献[5]提出优先考虑短规则分类的关联分类算法。Cheung^[6]提出了一种增量式更新关联规则的方法 FUP,使数据库发生变化后,对已经得到的关联规则进行维护,其后在 FUP 的基础上,又提出了 FUP2 算法^[7],从而不仅可以处理数据的增加,而且还可以处理数据的删除和修改。

本文主要以五金产品的阀门为例,从解决质量管理的实际角度出发,提出了改进的 FUP 算法,并将其运用到产品质量分析系统中,对企业的业务数据(特别是质量数据)进行分析。

1 基于关联规则方法的五金产品品质数据分析

1.1 增量关联规则算法 FUP

由于交易数据库中的数据是不断增加的,这必然会引起关联规则发生变化,如果采用 Apriori 算法对更新后的数据库进行挖掘,由于数据库规模会越来越大,这样 Apriori 算法的挖掘效率会变得越来越低^[8]。增量式更新算法就是针对这个问题提出的。设原有交易数据库中的数据集记为 DB(也称为旧数据集),新增加数据集记为 db(也称为新数据集),则当前事务数据库为(DB+db)。

采用 Apriori 算法获得数据集 DB 的频繁项目集是 L(DB),则 FUP 算法的基本思想是:

(1) 利用 Apriori 算法生成新事务数据集 db 的频繁项目集 L(db),比较 L(db) 和数据集 DB 的频繁项目集 L(DB),找出其相同部分,将相同部分放入当前事务数据库(DB+db)的频繁项目集中。

(2) 对于 $t \in L(DB) - L(db)$ 的频繁项目集,如果 $t \in L(DB)$ 且 $t \notin L(db)$,则扫描 db 得到 t 在 db 中的支持度 support_d,再根据 DB 中已经求得的支持度 support_D,求出 t 在(DB+db)中的支持度为

$$\text{support}_{UD} = \frac{\text{support}_d + \text{support}_D}{d + D} \quad (1)$$

如果 $\text{support}_{UD} \geq s \times (D + d)$,则把 t 放入当前事务数据库(DB+db)的频繁项目集中,否则 t 不是频繁项目集。

(3) 对于 $t \in (L(DB) - L(db))$ 的频繁项目

集,如果 $t \in L(db)$ 且 $t \notin L(DB)$,则扫描 DB 得到 t 在 DB 中的支持度 support_D,再根据 db 中已经求得的支持度 support_d,求出 t 在(DB+db)中的支持度为

$$\text{support}_{UD} = \frac{\text{support}_d + \text{support}_D}{d + D} \quad (2)$$

如果 $\text{support}_{UD} \geq s \times (D + d)$,则把 t 放入当前事务数据库(DB+db)的频繁项目集中,否则 t 不是频繁项目集。

其中利用 Apriori 算法生成新事务数据集 DB 的频繁项目集 L(DB)的流程图如图 1。

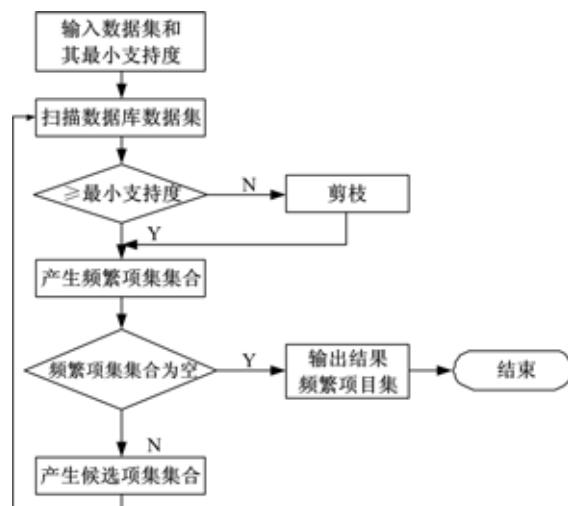


图 1 Apriori 算法流程图

Fig. 1 Flow chart of Apriori algorithm

1.2 FUP 算法的改进

由于缺陷数据的积累过程是一个比较长的过程,而在过程中,一些数据会因为某种分析因素的变化而渐渐地失去其分析价值。

因此,在分析缺陷数据时必须考虑数据的时间属性。对于不同时间的数据,其用于分析的价值是不同的。对五金缺陷数据分析来说,近期的数据相对于前面的数据更具有分析价值,而很早以前的数据基本上就没有太大的参考价值了。

针对以上分析,在 FUP 算法的基础上提出一种改进算法以解决在五金产品缺陷数据挖掘中遇到的问题。

在原先的 FUP 算法框架中,加入参数 s_1 ($0 \leq s_1 \leq s$),其含义是某项目集(这个项目集可以在旧数据集里面也可以在新数据集里面)在新数据集 db 中的最小支持度。

改进的 FUP 算法的流程图如图 2 所示。

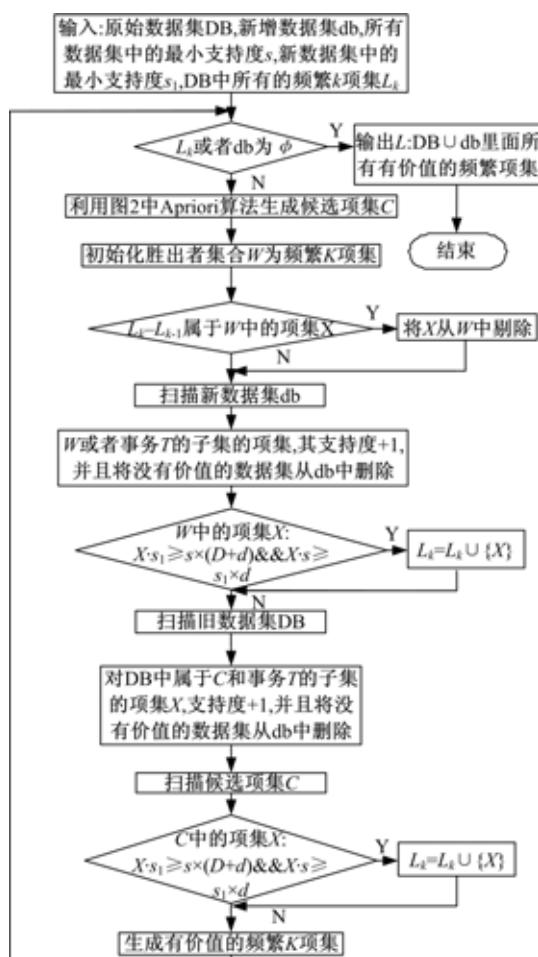


图 2 改进的 FUP 算法流程图

Fig. 2 Improved algorithm flow chart of FUP

2 改进的 FUP 算法在阀门产品质量数据分析中的应用

将改进的 FUP 算法应用到阀门产品质量数据分析系统中,应用改进 FUP 算法得到部分规则如表 1 所示。应用 Apriori 算法得到部分规则如表 2 所示。

从表 2 可以找出最具有实际意义的规则是 1、5、7、9 这些规则。从规则 1 可以看出阀体尺寸不标准这一缺陷问题主要出现在编号为“20101”的工序,而使用的原材料是“CU1021”。通过工序的编号,系统就可以找到这个工序、使用的加工机器编号以及操作工。

从表 1 可以看出,使用了改进算法后,原先的许多规则已经不在生成的规则列表中出现,而一些新的规则被生成了。这是因为许多数据条目尽管在旧数据集合里是频繁项目。而且在总的数据

集合里也是频繁项目,但是在新的数据集合里面却不是频繁项目了,所以这些规则变得“过时”而被改进算法去除;另外一些数据条目虽然在旧数据集合里不是频繁项目,但是在新的数据集合和总的数据集合里是频繁项目的,因此一些新的规则被生成。

表 1 使用改进算法关联分析得到的部分规则

Table 1 Association rules with FUP algorithm

序号	规则描述	支持度/%	置信度/%
1	工序信息编号 = “20502” & 零件原材料型号 = “CU1021” & 零件型号 = “FG-A0305” → 缺陷 = “阀盖处外部泄漏”	9.3	100
2	零件型号 = “FG-A0305” → 缺陷 = “阀盖处外部泄漏” & 产品型号 = “QF-A01123”	9.3	100
3	工序信息编号 = “20506” & 零件原材料型号 = “CU1023” & 零件型号 = “FT-A0302” & 产品型号 = “QF-A01121” → 缺陷 = “阀体尺寸不符合标准”	6.7	90
4	工序信息编号 = “20506” & 零件原材料型号 = “CU1023” & 零件型号 = “FT-A0302” → 缺陷 = “阀体尺寸不符合标准” & 产品型号 = “QF-A01121”	6.7	90
5	工序信息编号 = “20503” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0301” & 产品型号 = “QF-A01122” → 缺陷 = “阀体有裂纹”	8.1	100
6	工序信息编号 = “20503” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0301” → 缺陷 = “阀体有裂纹” & 产品型号 = “QF-A01122”	8.1	85

如规则 1 在原先的规则列表中没有,是从新增的数据集合中生成的。从规则 1 可以看出阀盖处外部泄漏这一缺陷问题主要出现在编号为“20502”的工序,而使用的原材料是“CU1021”。通过工序的编号,系统就可以找到这个工序、使用的机器编号以及操作工,说明有可能对应的机器有故障或老化,或者是技术工人的操作不规范,导致缺陷的产生。那么就需要检修该机器或加强员工的培训。

表 2 利用 Apriori 算法关联分析得到的部分规则
Table 2 Association rules with Apriori algorithm

序号	规则描述	支持度/%	置信度/%
1	工序信息编号 = “20101” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0301” & 产 品型号 = “QF-A01121” → 缺陷 = “阀体尺寸不符合标准”	4.3	100
2	工序信息编号 = “20101” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0301” → 产 品型号 = “QF-A01121” & 缺陷 = “阀体尺寸不符合标准”	4.3	100
3	工序信息编号 = “20101” & 零件型号 = “FT-A0301” → 零件原材料型号 = “CU1021” & 产 品型号 = “QF-A01121” & 缺陷 = “阀体尺寸不符合标准”	4.3	85
4	零件型号 = “FT-A0301” & 缺陷 = “阀体尺寸不符合标准” → 工序信息编号 = “20101” & 零件原材料型号 = “CU1021” & 产品型号 = “QF-A01121”	4.3	100
5	工序信息编号 = “20101” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0302” & 产 品型号 = “QF-A01121” → 缺陷 = “阀体尺寸不符合标准”	8.7	85
6	工序信息编号 = “20101” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0302” → 产 品型号 = “QF-A01121” & 缺陷 = “阀体尺寸不符合标准”	8.7	66.7
7	工序信息编号 = “20103” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0301” & 产 品型号 = “QF-A01121” → 缺陷 = “阀体有裂纹”	5.1	100
8	工序信息编号 = “20103” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0301” → 产 品型号 = “QF-A01121” & 缺陷 = “阀体有裂纹”	5.1	100
9	工序信息编号 = “20118” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0302” & 产 品型号 = “JZF-A01131” → 缺陷 = “密封面不平”	6.1	100
10	工序信息编号 = “20101” & 零件原材料型号 = “CU1021” & 零件型号 = “FT-A0302” → 产 品型号 = “JZF-A01131” & 缺陷 = “密封面不平”	6.1	80

3 结束语

五金产品实际生产中,由于分析的历史数据是一个经常扩充的数据集合,扫描原始的数据库的代价非常昂贵。增量式更新算法能充分利用已挖掘出的知识来提高挖掘效率。本文对增量式更新算法 FUP 提出改进,考虑时间属性,去掉渐渐变得没有分析价值的数据。与实验结果对比,改进的 FUP 算法去掉了在旧数据集中是频繁项目而在新数据集中是非频繁项目的规则,取得了较好的性能。

参考文献:

- [1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases [C]// Proc of the 1993 ACM SIGMOD Int'l Conf on Management of Data (SIGMOD'93), 1993: 207-216.
- [2] Agrawal R, Srikant R. Fast algorithms for mining association rules [C]// Proc of the 20th Intel'l Conf on Very Large Data Bases (VLDB'94), 1994: 487-499.
- [3] 蒋盛益,李霞,郑琪. 数据挖掘原理与实践 [M]. 北京:电子工业出版社,2011.

[4] 郭鸿,黄桂敏,周娅. 基于 Kademlia 的下关联规则挖掘算法研究 [J]. 计算机工程与设计, 2011, 32(1): 221-223.

Guo Hong, Huang Gui-min, Zhou Ya. Study of association rules discovered algorithm based on Kademlia [J]. Computer Engineering and Design, 2011, 32(1): 221-223.

[5] 武建华,沈钧毅,王元元. 一种改进的关联分类算法 [J]. 计算机工程, 2009, 35(9): 63-65.

Wu Jian-hua, Shen Jun-yi, Wang Yuan-yuan. A improved algorithm on association and taxonomy [J]. Computer Engineering, 2009, 35(9): 63-65.

[6] Cheung D W, Han J, Ng V, et al. Maintenance of discovered association rules in large databases: an incremental updating technique [C]// Proc 1996 Int'l Conf on Data Engineering, New Orleans, Louisiana, 1996.

[7] Cheung D W, Lee S D. A general incremental technique for updating discovered association rules [C]// Proc 1997 Int'l Conf on Databases Systems for Advanced Applications, Melbourne, Australia, 1997.

[8] 郭炜星. 数据挖掘分类算法研究 [D]. 杭州:浙江大学, 2008.

Guo Wei-xing. Study of data discovered association algorithm [D]. Hangzhou: Zhejiang University, 2008.