使用定性属性的数据库关联规则的增量挖掘①

冯永华, 王晓峰

(上海海事大学 信息工程学院, 上海 201306)

摘 要: 目前数据库关联规则的增量挖掘作为数据挖掘的一个重要的领域, 已经广泛应用于教育, 医疗, 卫生等领域, 因此它成为了当今数据挖掘中最活跃, 最重要的一个分支领域. 数据库中的数据存在大量未知的数据以及不可知的数据变化. 若采用 Apriori 算法进行计算, 一方面很难取得较好的结果, 另一方面支持度的变化对结果的影响很大, 无法确定支持度的变化, 因此借助属性论中定性属性的机理以及属性计算网络的边界学习算法, 结合 IUBM 算法提出了一种基于定性属性的关联规则的增量挖掘算法. 比如在以分数划线招生制度下, 定性基准的一分之差, 可能完全改变一个学生的一生的命运. 通过实验表明, 该算法在处理大规模数据的增量式关联规则的挖掘中减少了冗余规则的产生, 同时挖掘效率得到了很大的提升. 对于诸如预测大学生就业的情况及招聘企业对于应届生学习情况的了解等应用十分有意义.

关键词: 数据库; 定性属性; 关联规则; 增量挖掘; 边界学习算法

Database Incremental Association Rules Mining Based on the Qualitative Properties

FENG Yong-Hua, WANG Xiao-Feng

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: Nowadays, the present incremental database association rule mining is an important area as data mining, has been widely used in education, medical, health and other fields, so it has become the most active data min are change and unknown. If use the Apriori algorithm to calculate, on the one hand it is difficult to achieve good results, on the other hand, a great impact on support changes we can not determine the support change. So with the mechanism of qualitative attribute theory and attribute computing network boundary learning algorithm with IUBM algorithm, we propose an algorithm for mining association rules based on incremental qualitative attribute. For example, in order to score crossed the enrollment system, qualitative datum point, it may completely change the life of a student's life. With the experiments show that, this algorithm reduces the redundant rules generated in the incremental mining association rules in large-scale data processing, at the same time the mining efficiency has been greatly improved. Apply the researches to prediction of College Students' employment for graduates to understand the learning situation of application is very meaningful.

Key words: database; qualitative attributes; association rules; the incremental mining; boundary learning algorithm

1 引言

从当今社会来看, 海量的数据时代已经来临, 尤其是在互联网, 电信, 金融等领域, 而关联规则是数据挖掘问题中一个重要的研究内容, 广泛用于互联网, 电信, 金融等领域. 但在实际应用中, 数据库里的数据不是一成不变的, 是随着时间的改变而发生变化的, 如文献[1]提出的动态数据集中关联规则挖掘的进化论方法, 文献[2] 提出的数据集中模糊关联规则的挖掘及文献[3]提出的多目标优化遗传算法的关联规则挖掘以及生产企业从生产现场采集的数据, 商场的商品交易数据等, 因此关联规则的增量挖掘问题在数据挖掘领域十分重要, 对于关联规则的增量挖掘问题的研究非常有意义.

国内外对于关联规则的增量挖掘有很多不同的研

① 收稿时间:2015-01-12;收到修改稿时间:2015-04-07

究, 大多数的关联规则的增量式更新算法都是以 Aprior 算法为核心进行改变或者优化, 包括冯玉才等提出的 IUA 算法和 PIUA[4]算法, 但是由于每种算法都无法避免多次扫描数据库, 在时间和效率等各方面存在着诸多的缺点.

针对关联规则的增量问题, 本文采用属性计算网络的边界学习算法和 IUBM 算法的融合, 提出了一种新的基于定性属性的关联规则的增量式更新算法, 并将该算法和 FUP 算法应用于生物信息的数据关联规则挖掘, 对两种实验结果进行了分析对比.

2 关联规则的更新

关联规则: 假设 I = {i1, i2, i3, … , im }是 m 个不同项目的集合, 事物数据库为 D, X=>Y 的蕴含式, 其中 X∈I, Y ∈I, 并且 X∩Y= Ø.

支持度(support)S: 设关联规则 X ⇒Y 在事物集中成立, 把 D 中包含 X ∪Y 的事物占整个数据集 D 的百分比;

置信度(conference)C: D 中包含 X 同时也包含 Y 的事务占事务 X 的百分比;

用户可以指定最小支持度阈值和最小置信度阈值;

项目的集合称为项目集, 如果项集满足最小支持度, 则称为频繁项目集, 反之称为非频繁项目集.

关联规则的挖掘问题归结为以下两个问题:

①找出所有事物中的频繁项目集;

②利用频繁项目集生成关联规则;

本文的研究重点是在第一个问题上. 设原数据库为 DB, 原支持度为 S, |DB|为 DB 中事物总数; 新增加的数据为 db, 新的支持度为 S' , |db|为 db 中的事物中总数.

关联规则的更新问题就是找出 |DB+db|在新的最小支持度下的所有频繁项目集 L, 并生成相应的关联规则.

前文提到的关联规则的增量挖掘中, 许多专家提出了多种不同的算法, 比如朱玉全等提出的 FIUA 算法, 虽然此算法解决了支持度变化时关联规则的变化, 但是没有数据的增加. D.W.Cheung 提出的 FUP[5]算法和 FUP2[6]算法, 但是此算法需要多次扫描数据库, 耗时方面和内存开销巨大.

本论文的目的在于考虑到数据库里的数据有增加并且最小支持度发生变化时, 关联规则的更新问题. 针对这一问题, 可以利用 FP-growth[7]算法在重新计算一遍, 但是时间和空间开销巨大, 而且没有利用已有的结果, 浪费了前面所做的工作. 为了充分利用已知的结果, 本文应用属性论中计算网络的边界学习算法与 IUBM 算法进行融合, 提出了一种新的基于定性属性的关联规则增量挖掘算法.

3 定性属性的描述

定性属性是属性论中一个非常重要的定义, 在现实生活中许多数据往往是连续型的, 并且时间是其中一个隐含的参数, 而决策者一般都是用语言评判等级, 所以在挖掘之前必须将定性属性转换为定量属性并再次对定量属性进行离散化.

应用定性属性的关联规则的增量挖掘分为以下三种情况:

第一: 在定性基准不变的情况下, 数据库里的数据有了增加, 此时的关联规则的变化情况;

第二: 在定性基准变化和数据库的数据都发生变化时, 关联规则的变化;

第三: 在定性基准发生变化而数据库里的数据不发生变化时关联规则的变化;

由于文章篇幅有限, 本文针对第一种情况从属性论定性属性的角度出发结合 IUBM 算法对原数据库和新增数据库进行关联规则的增量的挖掘, 在以后的学习当中会逐个实现上述第二和第三种情况.

4 FUP 算法思想

首先扫描 db,产生 1-频繁项目集 L1, 对于任意项 x;

若 x 在 db 中是频繁项目集, 在 DB 中也是频繁项目集, 则将 x 加入 LDB+db 中;

若 x 在 db 中是非频繁项目集, 在 DB 中是频繁项目集, 计算 x 的支持度

$$\text{Support(x)}=\frac{\left|\mathrm{x}\right|_{db}+\left|x\right|_{DB}}{\left\|db\right|+\left|DB\right\|} \quad (1)$$

|x|DB 为 x 在 DB 中的数, |x|db 为 x 在 db 中的数, 如果 Support(x)>=minsupport, 则把 x 加入到 L1 中;

若 x 在 db 中是频繁项目集, 在 DB 中是非频繁

项目集, 扫描 DB 得到 x 在 DB 中的支持度, 再由式(1)计算 x 的支持度, 如果 Support (x)>= minsupport, 则把 x 加入到 L1.

对于 LDB-1 和 Ldb-1, 令集合 M=(LDB-1 -Ldb-1), 若 M 不为空, 对于 M 中的每个元素 X,按式(2)计算支持度

$$\text{Support}(x) = \frac{|x|_{DB}}{\|db| + |DB\|} \quad (2)$$

如果 Support(x)>=minsupport, 则把 x 加入到 L1;

由得到的 Lk 构造 db 中的 Lk+1,对于每个 K+1 项集 x:

若 x 在 DB 中是频繁项目集, 且在 db 中也是频繁项目集, 则把 x 加入到 Lk+1 中;

若 x 在 DB 中是频繁项目集,在 db 中是非频繁项目集, 扫描 db 得到 x 在 db 中的支持度, 再由式(1)计算 x 的支持度, 如果 Support(x)>=minsupport, 则把 x 加入到 Lk+1;

若 x 在 DB 中是非频繁项目集 ,在 db 中是频繁项目集, 扫描 DB 得到 x 在 DB 中的支持度, 再由式(1)计算 x 的支持度, 如果 Support(x)>=minsupport, 则把 x 加入到 Lk+1;

若 x 在 DB 中是非频繁项目集 在 db 中是非频繁项目集, 则舍弃 x.

5 FUP 算法分析

FUP 算法主要是利用旧频繁项目集对 DB+db 中的频繁项集是否为最终的频繁项目集进行分析, 从而得到项目的频繁项目集. 该算法在新增加的数据集与原数据集相差不大的情况下具有较高的效率, 但是在现实生活的大数据集中, 新增加的数据集与原数据集存在一定的差异, 并且数据量非常大而且不断的发生变化, 从而对某些重大的决策产生很大的影响.

6 基于定性属性的关联规则的增量式更新算法思想

本算法的核心思想是前期利用属性论中的边界学习算法将数据进行离散化和分组, 在与定性基准一一对应, 减少了一一判断每个数据的时间, 然后利用 IUBM 算法挖掘出新的关联规则.

定义 1. 将数据库中的事物分成若干部分, 若一个项目集为频繁项, 则它至少在一个部分中是频繁的.

定义 2. 若一个项目集是非频繁项目集, 则所有包含该项目集的超集也一定是非频繁项目集.

6.1 边界学习算法

过程如下:

其中 R 为实数集, J 为样本序列数, K 为属性量的序列数. 其中 αk, βk 为属性的定性基准.

步骤 1: 输入样本集 X={Z(k)j | Z(k)∈R, j= 1,2,3…; k=1,2,3,. }

步骤 2: 预处理样本集. 原来的学习输出结果只有两种, 即属于定性映射[αk, βk], 结果为 1, 不属于则结果为 0.

改进边界学习法加入了度的概念, 使用转换程度函数(3)对输出的结果进行处理, 即输出结果属于[0,1],则样本转化为输入样本集 X={ Z_j^k | Z_j^k R, j=1,2,3...; k=1,2,3,... } 和输出样本集 Y={ $R_j^{(k)} \left| R_j^{(k)} \right.$? R, k 1,2,3...; j = 1,2,3... }.

$$\eta(Z_i^k) = \frac{1}{1+\exp((-\frac{Z_i^k - \varepsilon}{\delta_i}))} \quad (3)$$

步骤 3: 对输出结果采用快速排序法进行排序并对输入样本集进行排序 并限制 αk 永远比 βk 小, 从而减少了调节次数, 调节次数 I 初值为 0, 提取靠0且大于0和靠近 1 且小于 1 的两个样本作为初始基准[αk ,βk].

步骤 4:

(1)学习 α: 若有样本对应的结果大于 0, 则逆序选择一个正例样本组 Zik 转向步骤 5, 否则转步骤 6.

(2)学习 β: 若有样本对应的结果大于 0, 则顺序选择一个正例样本组 Zik 转向步骤 5, 否则转步骤 6.

步骤 5: $Z_i^k < \alpha^{(k)}$ $\alpha^{(k)} = Z_j^{(k)}$ 若 $Z_j^{(k)} > \beta^{(k)}$ 且令 $\beta^{(k)} = Z_j^{(k)}$, I=I+1, 转步骤 4.

步骤 6: 输出调节次数 I 及定性基准, 结束.

边界学习算法求出了数据库数据的定性属性的边界. 将数据库中的数据分割成为了多个分区, 并将各个分区进行编码, 即数据库中的数据属于该分区对应的定性属性.

6.2 IUBM 算法

再加入新的数据后, 支持度会有四种变化:

I. 在 DB 和 db 中都是频繁项目集;

II. 在 DB 中是频繁项目集, 但是在 db 中不是频繁项目集;

III. 在 DB 中不是频繁项目集, 但 在 db 中是频繁项目集;

IV. 在 DB 和 db 中全不是频繁项目 集;

针对这四种支持度的变化, IUBM 算法的算法过程如下:

首先采集应用边界学习算法得到的分组数据, 关联规则的挖掘实际可以描述为频繁项目集的挖掘, 所以在应用边界学习算法得到的分组数据可以从数据的大小上直观的分辨出频繁项目集所在的分组区间, 则将相应的区 间数据应用 IUBM 算法, 从而大大减少了许多没有实际价值的数据量.

输入: 原数据库 DB, DB 中的所有频繁项目集 LDB, 新增数据 db, 原支持度 S, 变换后的支持度 S'

输出: DB+db 的频繁 k-项目集

若 S< S' ,则在 DB 中删除支持度小于新的支持度的项集, 更新 DB 中的频繁项目集, 得到新的频繁项目集, 若 S>S', 则用 ABM 算法更新 DB 中的频繁项目集, 得到 DB 中的新的频繁项目集.

同时扫描 db,构造矩阵 bv, 用 ABM 找出所有的 k-项目集.

将 DB 中的频繁项目集和 db 中的频 繁项目集进行比较.

①若 X 属于 I 种情况, 则将 x 直接加入新的频繁项目集中;

②若 x 属于 II 种情况, 则

a. 计 算 db 的 支 持 度 Ldb,s=(xi ∧xi +1 ∧...∧xi +k) * t,t = (1,1,1,...,1)Γ

b.如果 LDB+db, s=(LDB,s+Ldb,s) ≥S' *|DB +db|, 则将 x 加入新的频繁项目集中.

c.否则删除 x.

③ 若 x 属于 III 种情况, 则计算

$$L_{DB,S} = (x_i \wedge x_{i+1} \wedge \ldots \wedge x_{i+k}) * t \quad ,$$

t= (1,1,1,...,1)Γ

如果 L_{DB+db}, $s = (L_{DB,s} + L_{db,s}) \geq S' * |DB + db|$ 则将 x 加入到新的频繁项目集中. 否则删除 x 输出新的频繁项目集.

7 基于定性属性的关联规则的增量式更新算法与 FUP 算法的实验对比

采用 MyEclipse 作为开发工具, 开发语言选择 Java, 实现了挖掘算法, 并在操作系统为 Windows XP, CPU 为 Intel Pentium、内存为 2GB、主频为 2.2GHZ 的 PC 机上进行实验, 数据集采用 UCI 机器学习数据库中的 Thyroid Disease Data Set(http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease), 原始数据为抽取 7200 条数据中的 1000 条数据, 新增加的数据为 500,1000,1500, 2000 条记录.

表 1 为数据集首先应用边界学习算法得到的游离甲状腺指数分组的数据

表 1 游离甲状腺指数分组的数据

FTI	数量
43~50	280
50~63	541
63~124	129
124~170	32
170~188	18

从所得到的分组数据中看出, 在 50~63 区间上数量为 541, 数据量为最 大, 所以将 50~63 中的 541 个数据应用 IUBM 算法.

图 1 为 FUP 算法与基于定性属性的关联规则的增量式更新算法耗时对比图.

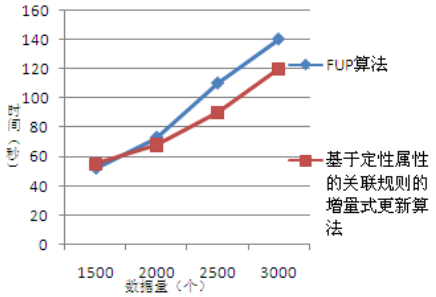

图 1 耗时对比图

在加入数据集为 500 时, 由于数据量很小, 首先采用边界学习算法进行分类的效果不明显, 但当数据增加到 2000 时由图可看出, 时间量减少很大, 从而再一次证明基于定性属性的关联规则的增量式更新算法的高效性.

8 结语

本文在 IUBM 算法的基础上加上了属性论中的定性属性的边界学习算法, 首先将数据进行分类, 并将分类结果再应用 IUBM 算法得到最后的新的频繁项目集, 通过与 FUP 算法的比较, 在时间取得了较好的结果.

该算法也有需要改进的地方, 如利用 IUBM 挖掘时, 支持度的把握需要人为控制, 因此, 算法在自适应方面还需要加强.

参考文献

1 Shenoy PD, Srinivasa KG, Venugopal KR. Evolutionary approach for mining association rules on dynamic database. Proc. of Pacific Asia Conf. on Advances in Knowledge Discovery and Data Mining. Seoul, Korea. 2003. 325–336.

2 de Graaf JM, Koster WA, Witterman JJW. Interesting fuzzy asscciation rules in quantitative database. Proc. of the 5th European Conf. on Principles of Data Mining and Knowledge Discovery. Darmatadt, Gremany. 2001. 140–151.

3 Kaya M, Alhajj R. Muti-objective genetic algorithm based method for mining optimized fuzzy association rules. Proc. fo the 5th Int'l Conf. on Intelligent Optimized Data Engineering and Automated Learing(IDEAL). Execter, England. 2004. 758–768.

4 冯玉才,冯建琳.关联规则的增量式更新算法.软件学报,1998, 9(4): 301–306.

5 Agrawal R, Srikant R. Fast algorithms for mining association rule. Proc. of the 20th International Conference on Very large Data Base (VLDB). Santiago Chile Morgan Kaufmann. 1994. 478–499.

6 Cheung DW, Han J, Ng VT, et al. Maintenance of discovered association rules large databases.An incremental updating technique. Proc 1996 Int Conf Data Enginnering.New Orleans USA: IEEE Computer Society. 1966.106–114.

7 Han J, Pei J, Yin Y. Ming frequent patterns without candidate generation. Proc. of the ACM SIGMOD Intl Conf on Management of Data. Dallas, Texas, United States. 2000. 53–87.