一种有效压缩频繁模式挖掘的算法

童咏昕 马世龙 李 钰

(北京航空航天大学 软件学院, 北京 100191) (北京航空航天大学 计算机学院, 北京 100191)

摘 要: 频繁模式挖掘的研究最近致力于在一个合理的容错范围内寻找有代表性的模式来压缩庞大的挖掘结果集. 一种新型启发式算法 AMSA(Approximating Mining based Simulated Annealing)被提出, 其采用了模拟退火思想来保证有效性和压缩的质量. 依据 FIMI(Frequent Itemset Mining Implementations Repository)提供的公用数据集进行的实验结果也证明了这一结论. 通过与 FPclose算法和 RPglobal算法分别进行了性能的比较, AMSA 挖掘的结果集规模小于 FPclose算法和 RPglobal算法得到的结果集规模, 特别是当支持度阈值很低时, RPglobal不可在合理时间内产生结果集, AMSA 却可在合理时间内得出较精准的结果集.

关 键 词: 数据挖掘; 模拟退火; 启发式方法

中图分类号: TP 311. 13

文献标识码: A 文 章 编 号: 1001-5965(2009) 05-0640-04

Effective algorithm for mining compressed frequent patterns

Tong Yongxin

(School of Software, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Ma Shilong Li Yu

(School of Computer Science and Technology, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract Researches of frequent-pattern mining have recently focused on discovering representative patterns to compress a large of results within a reasonable tolerance bound. A novel heuristic algorithm, approximating mining based simulated annealing (AMSA), was proposed. The algorithm uses a method based simulated-annealing to improve efficiency and quality of the compression. Our experimental studies demonstrate the algorithm is efficient and high quality on a common dataset supported by frequent itemset mining implementations repository (FIMI). The mining result of AMSA is smaller than mining results of FPclose and RPglobal by performance study. Especially, if min_sup threshold is low, RPglobal fails to generate any result within reasonable time range, while AMSA generates a concise and succinct mining result.

Key words data mining; simulated annealing; heuristic method

在过去的十几年对数据挖掘研究的过程中, 频繁模式挖掘一直扮演着一个极其重要的角色. 但由于众所周知的“向下封闭特性”导致了挖掘出的频繁模式数量会呈指数爆炸规模[1-2], 如何从完全挖掘结果集中选择一个规模较小的、用户最感兴趣的模式集合始终是该领域的研究热点与新挑战.

针对上述问题, 一种通用的解决方法是压缩全部结果集, 并从中选出一些有代表性的模式来代替全部结果集. 目前压缩频繁模式的研究主要分为无损压缩与有损压缩两类方式. 所谓无损压缩是指挖掘闭频繁项集[3-5], 而所谓有损压缩方式较多, 如挖掘极大频繁模式[6]、集成模式、基于集合覆盖的近似挖掘[7-8]等等. 该类方法得到的结果集与完全结果集相比有部分的信息损失, 但均在一个合理的范围之内. 其中基于集合覆盖思

收稿日期: 2008-08-10

基金项目: 国家 973计划资助项目 (2005CB321902)

作者简介: 童咏昕 (1982-), 男, 北京人, 硕士生, yxtong@nlsde.buaa.edu.cn.

想的近似挖掘方法成为挖掘出核心模式的主要方法之一. 该类算法是将挖掘的完全结果集作为被覆盖集合, 通过聚类技术产生有代表性的模式集合作为覆盖集, 从而将挖掘压缩模式的问题转化为 NP难问题, 再采用基于贪心的近似算法来达到寻找压缩模式的目的[8].

通过上述分析可知压缩模式集合的性能主要依赖于近似算法的效率和准确度, 但传统的基于贪心思想的近似算法很容易陷入贪心思想造成的局部最优解中, 而丢失了全局最优的信息. 一种基于模拟退火思想的近似挖掘算法 AMSA (Approximating Mining based Simulated Annealing)被提出来解决该问题. 该算法可以很大程度上避免陷入局部最优的可能性, 并达到全局最优. 此外, 在该算法中通过调整相关参数影响算法的优化效果. 因此, 当产生的频繁模式过多时, 可以通过调整参数既使结果在容错区间内, 并保持较高执行效率.

1　相关概念

所谓频繁模式挖掘 (或称作频繁项集挖掘) 定义为如下形式. 令 I是一个项的集合, 记作 $\{o_1, o_2, \cdots, o_d\}$, 称 I的一个非空子集为一个项集. 此外, 对于一个给定的事务数据库 D, 它是一个项集的集合, 记作 $\{t_1, t_2, \cdots, t_n\}$, 且 $t_i \subseteq I$. 对于任意一个项集 α, 将包含 α的事务的集合记作 $D_\alpha = \{i \mid \alpha \subseteq t_i$ 且 $t_i \in D\}$. 再定义一个项集 α的势为 α中包含的项的数量, 例如, $|\alpha| = \{o_i \mid o_i \in \alpha\}$.

定义 **1**　对于一个事务数据集 D, 如果 $|D_\alpha|/|D| \geqslant \sigma$, 其中 $|D_\alpha|/|D|$称为 α的支持度, 记作 $s(\alpha)$, 而 σ是用户给定的最小支持度阈值 ($0 \leqslant \sigma \leqslant 1$), 则称 α是频繁项集.

定义 **2**　对于一个项集 α, 如果它是闭频繁项集, 当且仅当, 不存在 α的超集 β, 使 β与 α有相同的支持度.

定义 **3**　存在一种距离的度量函数 $Dis: P \times P \rightarrow [0, 1]$, 该函数对于一个给定的模式集合 S中任意两个模式 $p_1, p_2 \in P$, 存在一种映射关系, 使得函数值在 $[0, 1]$. 在本文中采用一种 Jaccard距离来度量两个模式之间的相似程度:

$$D(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|}$$

定义 **4**　设存在模式 P和模式 P', 如果 $P \subseteq P'$, 且 $D(P_1, P_2) < \delta$则称模式 P被模式 P'“ δ覆盖了”. 其中 δ为用户给定的调节集合覆盖程度的参数.

定义 **5**　假设有一个如果存在模式集合, 如果一个有代表性的模式 P_r, 对于模式集合中每个元素 P 都被 P_r, δ覆盖, 则称这个模式的集合形成了一个 δ——聚类.

2　问题形式化与 AMSA算法

2.1　问题形式化

定义 **6**　给定一个事务数据库, 最小支持度阈值 M 和聚类质量测度 δ(即一个容错范围), 则挖掘有代表性模式集合问题被转换为寻找用最少的代表性模式覆盖从数据库中挖掘出的完全结果集的问题.

定理 **1**[8]　寻找最少个数的代表性模式问题是 NP难的, 并可以规约为一个最小集合覆盖的问题.

证明　此处证明略, 详情见文献 [8].

由于挖掘压缩频繁模式的问题可转化为寻找最小数量有代表性模式的最小集合覆盖问题, 所以压缩频繁模式结果集的质量依赖于对该 NP难问题的近似程度. 而当前对该问题的求解算法是基于贪心方式的最小集合覆盖思想[8], 但是其易于陷入局部最优解, 最终可能导致较低的近似率.

而随机的 Local Search策略采用随机游走的思想避免了陷入局部最优解的困境, 从而在对 NPC问题的近似求解中成为了高效的算法优化策略, 模拟退火就是其中一种经典的随机 Local Search策略. 一种新型的基于模拟退火的最小集合覆盖算法被提出 SCSA (Set Covering based Simulated Annealing), 进而产生了 AMSA算法. 下面先简要回顾模拟退火的基本思想, 随后详细阐述 AMSA算法.

模拟退火算法是 1983年 Kirkpatrick将 Metropolis的模拟退火思想引入组合优化领域形成的一种随机优化方法. 该算法模拟高温金属降温的热力学过程, 在优化过程中先随机地选择一个初始状态并考察该状态的目标函数值. 其后对当前状态进行一次随机扰动, 再计算新状态下的目标函数值, 并以概率 1接受好结果, 以某种概率 P接受较差结果作为当前值, 直到系统冷却. 此外, 由于该算法会以某种概率接受较差的结果值, 因此可以跳出局部最优解的约束.

2.2　AMSA算法

AMSA 算法由两个部分组成: 第 1部分为 AMSA主算法, 第 2部分为 SCSA算法, 其是一种基于模拟退火思想求解最小集合覆盖的算法.

算法 **1**　基于模拟退火的近似频繁模式挖掘算法 (AMSA)

输入: 由最小支持度阈值 $\hat{M}$ 产生的频繁模式完全集 F; 最小支持度阈值 M; 对于聚类的容错值 δ

输出: 有代表性模式的集合 R.

方法:

```
1: for each P∈F, support(P)≥M
2: 将 P 插入集合 A
3: for each Q∈F, 且 Q 覆盖 P
4: 将 P 插入集合 Q
5: While A≠0 do
6: R←SCBSA(A, Q)
7: Return R
```

算法 2 基于模拟退火的最小集合覆盖算法(SCSA)

输入: 由最小支持度阈值 $\hat{M}$ 产生的频繁模式完全集; 对频繁完全集进行 δ覆盖的模式的集合 Q.

输出: 最小覆盖集合 C.

方法:

```
1: T←T0
2: C←Q
3: n←n0
4: While(n! = 0)
5: do
6: S←C
7: U←F
8: for k←1 to L
9: 随机选取 Si∈S, U←U-Si
10: If(U 可以覆盖 X) then
11: If(|U|<|C|) then
12: C←U
13: n←n0
14: else
15: 依概率 P=exp{(|C|-|U|)/T}
接受 C←U, n←n0; 若未接受到 n←n-1
16: S←S-Si
17: T←αT
```

上述 AMSA 算法的输入是闭频繁模式的集合, 算法 1的 1~4行都在为基于模拟退火思想的集合覆盖算法准备输入条件. 算法 2中前 3行为确定参数和候选集合, 第 4行是模拟退火算法所要求的连续 n_0 个新解都未被接受时终止算法, 这正是其避免陷入局部最优困境的设计. 第 5~14行表示, 每次该算法都以一定的概率接受当前非最优解, 而迭代的次数是受初始参数 T_0影响的. 当迭代过程结束是, 最后集合 C 中以概率接受的最终结果即为最优结果集.

3 实验结果分析与比较

为比较上述算法, 采用 FMI的"频繁项集挖掘算法标准数据集 (网址为: http //fmi cs helsinki fi/data)[5]"测试 AMSA 算法. 实验模拟采用 Visual C++ 6 0在 CPU 为 Intel(R) Core(TM) Duo T2350 1. 86GHz, 内存为 1. 5GB RAM, 操作系统为 Windows XP的系统上实现了全部算法.

实验的过程主要是用 AMSA 算法与 FPclose 算法[5]和 RPglobal算法[8]进行了挖掘结果数量的比较: 其中 FPclose算法是在给定最小支持度阈值 M 的条件下, 产生出全部频繁闭项集. RPglobal算法的是先用频繁闭项集挖掘算法得出全部闭频繁项集, 然后在产生的闭频繁项集中挖掘出最小支持度阈值为 M 的代表性模式. AMSA 算法的实现过程也是在产生的闭频繁项集中采用 AMSA 算法得出全部有代表性的模式. 实验在 FMI 提供的 accidents, chess, Connect和 pumsb_star4类公共数据集上进行的.

实验结果如图 1~图 4所示. 可以清晰的看出, 由于在采用挖掘压缩的频繁模式, AMSA 算法和 RPglobal算法产生的模式数量远比 FPClose算法要少, 此外, 随着支持度的减少, AMSA 算法产生的模式数量比 RPglobal算法产生的模式数量减少许多. 这充分证明了采用基于模拟退火的全局最优策略在输出结果上比基于贪心思想的近似算

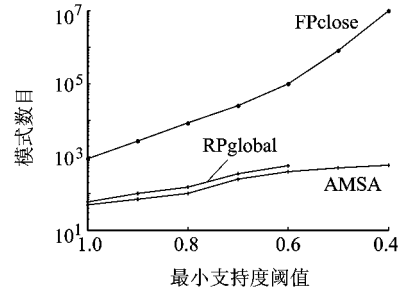

图 1 Accidents数据集输出的结果集

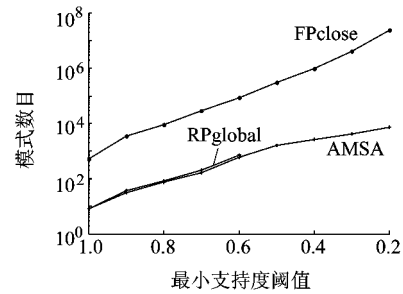

图 2 Chess数据集输出的结果集

法在近似准确度上提高了很多, 这也是全局最优和局部最优的区别所在. 此外, 通过上述 4幅实验比较图, 可以看出当支持度阈值很低时, RPgbbal算法由于候选集中模式数量过多, 而贪心思想又易陷入局部最优的困境中, 因此该算法不能在合理时间范围内得到结果集 (此处的合理时间为 1h). 而 AMSA 通过对初始温度参数的调整, 可以有效地得到全局最优解.

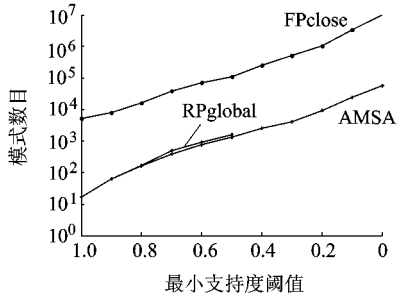

图 3　Connect数据集输出的结果集

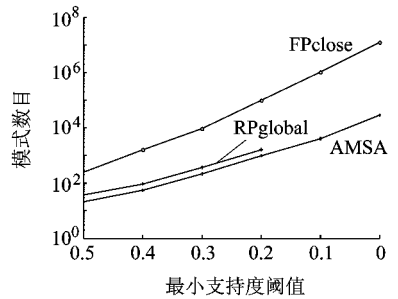

图 4　Pumsb_star数据集输出的结果

4 结　论

针对压缩频繁模式挖掘的问题, 本文提出了一种新型启发式算法——AMSA, 有效地挖掘出一个规模小的有代表性模式集合. 该算法采用了模拟退火思想, 以随机 LocalSearch策略避免了挖掘结果集陷入局部最优的困境. 采用 FMI的公共数据集进行实验比较也充分地证明了此结论.

参考文献 (References)

[1] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Proc of 1994 Int conf on VLDB. Santiago, Chile: VLDB, 1994: 487-499

[2] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[C]// Proc of the 2000 ACM SIGMOD. Dallas, USA: ACM, 2000: 1-12

[3] Pasquier N, Bastide Y, Taouil R, et al. Discovering frequent closed itemsets for association rules[C]// Proc of the 7th ICDT. Jerusalem, Israel: IEEE, 1999: 134-145

[4] Wang J, Han J, Pei J. CLOSET+: Searching for the best strategies for mining frequent closed Itemsets[C]// Proc of the 2003 ACM SIGKDD. Washington DC, USA: ACM, 2003: 236-245

[5] Grahne G, Zhu J. Efficiently using prefix-trees in mining frequent itemsets[C]// Proc of IEEE ICDM Workshop on FMI. Melbourne, FL: IEEE, 2003: 123-132

[6] Bayardo R. Efficiently mining long patterns from databases[C]// Proc of the ACM SIGMOD. Seattle, USA: ACM, 1998: 85-93

[7] Afrati F N, Gionis A, Mannila H. Approximating a collection of frequent sets[C]// Proc of the 2004 ACM SIGKDD. Seattle, USA: ACM, 2004: 12-19

[8] Xin D, Han J, Yan Y, et al. Mining compressed frequent-pattern sets[C]// Proc of the 2005 VLDB. Trondheim, Norway: VLDB, 2005: 709-720

(上接第 **590**页)

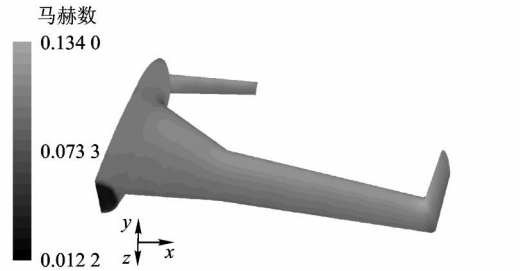

图 11　无人机表面来流速度

5 结　论

以上计算结果及分析表明, 使用计算流体力学软件对飞行器进行低速小尺度条件下的气动力数值模拟, 得到的气动参数可以为飞行器的气动外形设计提供依据和参考. 可以减少研究者重复、低效的劳动, 将更多精力和时间投入到考虑问题的物理本质中去, 因而提高工作效率.

参考文献 (References)

[1] Anderson jr J D. Computational fluid dynamics the basics with applications[M]. New York: McGraw-Hill, 1995

[2] 吴望一. 流体力学 [M]. 北京: 北京大学出版社, 1982
Wu Wangyi. Fluid dynamics[M]. Beijing: Beijing University Press, 1982(in Chinese)

[3] 陈耀松. 力学小议 [J]. 力学与实践, 2001, 23(4): 74-75
Chen Yaosong. Comments on the mechanics development[J]. Mechanics and Engineering, 2001, 23(4): 74-75(in Chinese)

[4] 张锡金. 飞机设计手册第 6册: 气动设计 [M]. 北京: 航空工业出版社, 2002
Zang Xijin. Aircraft design(6): aerodynamic design[M]. Beijing: Aeronautical Industry Press, 2002(in Chinese)