

挖掘完全频繁项集的蚁群算法

黄 红 星

(福建农林大学 计算机与信息学院,福建 福州 350002)

摘 要: 关联规则是数据挖掘发现的重要知识,完全频繁项集的发现是挖掘关联规则的关键步骤.蚁群算法是一种元启发式算法,已经有效应用于许多组合优化问题.因此,提出一种新的应用蚁群算法挖掘完全频繁项集的方法.对比实验表明,该算法是智能高效的.

关键词: 数据挖掘;关联规则;完全频繁项集;蚁群算法

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2014)12—0144—04

Ants Algorithm for Mining Frequent Itemsets in Larger Database

HUANG Hong-xing

(College of Computer and Information, Fujian Agriculture and Forestry University, Fuzhou 350002, China)

Abstract: In the field of data mining, an important issue for association rules generation is frequent itemsets discovery, which is the key factor in implementing association rule mining. The ant colony optimization is one of the meta-heuristics for combinatorial optimization problems. Therefore, a novel approach of applying ant colony optimization for mining frequent itemsets is proposed. Compared experiments show that this algorithm is intelligent and effective.

Key words: data mining; association rule; frequent itemsets; ant colony optimization

1 引言

面对海量数据的不确定性、无尺度性、非线性、非平稳、非高斯、非结构或半结构化等,传统的数据分析方法受到了极大的限制与挑战.《Science》杂志刊登《数据处理》专题指出^[1],到2007年,人类已储存295 EB的信息.如果所有数据都储存在CD光盘上,堆起来的高度可直抵月球.因此,数据的搜集、维护和使用已成为科学研究的主要方面,对许多学科而言,海量数据意味着更严峻的挑战,更好地组织和利用这些数据会有助于我们抓住巨大的机遇.数据挖掘主要发现的关联分析(关联规则和相关关系)、分类和预测、聚类分析、离群点分析(异常检测)和演变分析(序列模式和时间序列)等重要模式和知识^[2],其本质从数学上可看作复杂的最优化问题.而生物启发式计算由于高效的优化性能,无需问题特

殊信息等优点,已广泛应用于工程优化、计算机科学、组合优化、优化调度等领域,并且在先进重大装备制造、新能源与新材料、航空航天、电子信息、生物医药、农林渔业、经济管理等重点领域得到初步应用^[3].因此,利用生物启发式计算进行高效的海量数据挖掘成为一个有重大应用潜力的研究课题.

挖掘完全频繁项集是发现关联规则的核心步骤,也是许多数据挖掘任务中的关键问题,如分类、聚类、Web挖掘等,而数据挖掘中关联规则经典类Apriori算法^[4]在挖掘大型数据库时存在着性能瓶颈.蚁群在觅食过程中总能找到蚁巢和食物源之间的最短路径.受其启发,M. Dorigo等提出了一种新型的模拟生物智能的优化算法—蚁群优化算法^[5],该算法在解决复杂优化问题中表现出良好的性能.目前,应用蚁群算法于数据挖掘大量的研究主要还集中在解决聚类^[6]、分类^[7]等,或者直接挖掘关联规

收稿日期: 2014—03—23; **修回日期:** 2014—05—09

基金项目: 福建省教育厅科技项目(K80MKTY38A);福建农林大学青年科技项目(2011xjj19)

则. 例如屠莉^[8]利用蚂蚁系统(AS)在由数据集中属性和属性值对应超顶点和子顶点而构成的无向图挖掘频繁项集,与经典的 Apriori 和 FP-growth 算法比较,该算法可以较快挖掘出关联规则,具有较好的规则质量. 但是该算法挖掘的只是单维布尔型规则,所使用的数据集也较小,并且蚂蚁系统收敛速度慢,易出现停滞现象等. 沈国强^[9]等人给出一种基于遗传算法和蚂蚁算法相结合的多维关联规则挖掘算法,克服了遗传算法局部搜索能力弱和蚂蚁算法搜索速度慢的缺陷,在对具有稀疏特性的多维关联规则的挖掘中体现了良好的性能. R. J. Kuo^[10-11]等人应用蚁群系统(ACS)以及与聚类分析相结合挖掘多维约束的关联规则,并应用于大型数据库的关联规则挖掘,取得较满意的结果. 但是挖掘应用更广泛的完全频繁项集还很少,笔者曾利用蚁群算法挖掘最大频繁项集^[12],但是完全频繁项集挖掘与旅行商问题也是不同的,旅行商问题的解与节点之间的连接及其顺序有关,而完全频繁项集挖掘与节点之间的连接及其顺序无关,所以传统的蚁群算法模型并不适合解决完全频繁项集挖掘;另外一方面,最大频繁项集只是完全完全频繁项集的子集,问题规模相对较小,消耗空间和时间复杂度相对较低.

因此,本文给出完全频繁项集的概念和性质,对数据库进行垂直二进制编码,给出计算支持度的方法,通过扫描数据库,构建蚁群路径图,然后利用蚁群的随机搜索和正反馈原理挖掘完全频繁项集. 最后在不同的基准数据集上对算法的性能进行测试.

2 完全频繁项集的概念和性质

设事物数据库 $T = \{t_i \mid 1 \leq i \leq m\}$, 数据集 $I = \{i_j \mid 1 \leq j \leq n\}$, 其中 $t_i \subseteq I$. 任何非空子集 $A \subseteq I$ 称为项集, 事物 t_i 包含项集 A 当且仅当 $A \subseteq t_i$.

定义 1 项集 A 的支持度 $s(A) = |K(A)|/m$, 其中 $K(A) = \{k \mid A \subseteq t_k, 1 \leq k \leq m\}$.

性质 1 $\forall A, B \subseteq I, A \subseteq B$, 则 $K(B) \subseteq K(A)$, $s(B) \leq s(A)$.

定义 2 完全频繁项集 $F(s_{\min}) = \{A \subseteq I \mid s(A) \geq s_{\min}\}$, 其中 $0 < s_{\min} \leq 1$.

性质 2 $\forall A, B \subseteq I, A \subseteq B$, 若 $A \notin F(s_{\min})$, 则 $B \notin F(s_{\min})$.

性质 3 $\forall A, B \subseteq I, A \subseteq B$, 若 $B \in F(s_{\min})$, 则 $A \in F(s_{\min})$.

定义 4 k -项集 $L_k = \{A \subseteq I \mid |K(A)| = k\}$, 完全频繁 k -项集 $F_k(s_{\min}) = \{A \subseteq I \mid A \in$

$F(s_{\min}) \wedge A \in L_k\}$,

3 挖掘完全频繁项集的蚁群算法

3.1 数据库编码和支持的计算

设 $BV_j = (bv_{1j}, bv_{2j}, \dots, bv_{ij}, \dots, bv_{mj})$; 其中 $bv_{ij} = \begin{cases} 1, & \text{若 } I_j \in T_i \\ 0, & \text{否则} \end{cases}$, 项 I_j 支持度计数 $Support(I_j) = \sum_{i=1}^m bv_{ij}$, 从而得到位向量矩阵 $BV = [BV_1, BV_2, \dots, BV_j, \dots, BV_n]$, 即用垂直二进制位向量来表示数据库.

在经过第一步后, 不必再扫描数据库. 项集 $\{I_1, I_2, \dots, I_k\}$ 的支持度计数是项的位向量“与”运算后的向量中“1”的个数, 即 $Support\{I_1, I_2, \dots, I_k\} = \sum_{i=1}^m v_i$, 其中 $v_i = bv_{i1} \wedge bv_{i2} \wedge \dots \wedge bv_{ik}$, $\wedge$ 表示逻辑“与”.

算法 1: 支持的计算

```
Scan the database D;
for(i = 1; i ≤ m; i++)
{
  for(j = 1; j ≤ n; j++)
  {
    if(I_j in the T_i)
    {
      v_ij = 1;
    }
    else
    {
      v_ij = 0;
    }
  }
  for(j = 1; j ≤ n; j++)
  {
    u_{m+1}j = 0;
    for(i = 1; i ≤ m; i++)
    {
      u_{m+1}j = u_{m+1}j + v_ij;
    }
  }
  for(i = 1; i ≤ m; i++)
  {
    u_{i(n+1)} = 0;
    for(j = 1; j ≤ n; j++)
    {
      u_{i(n+1)} = u_{i(n+1)} + v_ij;
    }
  }
}
```

3.2 蚁群路径图的构建

通过扫描数据库, 可以得到每个项与支持它的一个交易列表相对应, 从而将数据库转换为垂直格式. 对与 I_i, I_j 对应的交易列表作“交”运算, 如果交易列表相交所得的结果个数大于给定最小支持度, 则构造一条从 i 到 j 的有向边.

算法 2: 蚁群路径图的构造

```
扫描数据库, 得到每个项 I_i 与支持它的一个交易列表 TID_i;
Support(I_i) = Count(TID_i); // 项 I_i 的支持度是列表 TID_i 的长度
for(i = 1; i ≤ n; i++) // n 是数据库中不同项集数
{
  for(j = 1; j ≤ n; j++)
  {
    {
```

$Support(I_i \Rightarrow I_j) = Count(TID_i \wedge TID_j)$; // 对任意两个交易列表求“交”

if($Support(I_i \Rightarrow I_j) \geq Min_Support$) // 如果结果个数大于最小支持度

从节点 i 到 j 构造一条有向边;

}

3.3 完全频繁项集的挖掘

信息素 $\tau_{ij}(t)$ 表示项 I_i 和 I_j 一起出现的期望度. 启发式信息取为 $Support(I_i, I_j)$, 算法过程中启发式信息不会发生变化. 对于图中没有边的两节点规定 $Support(I_i, I_j) = 0$. 一次循环中蚂蚁从某节点出发, 周游所有连通的路, 并且每个连通的节点只访问一次. 根据路径图的定义和频繁项集的性质, 孤立的节点(项)不可能构成完全频繁项集.

$p_{ij}^a(t)$ 表示 t 时刻蚂蚁 a 从节点 i 转移到节点 j 的状态转移概率:

$$p_{ij}^a(t) = \begin{cases} \frac{[\tau_j(t)]^\alpha [\eta_{ij}(t)]^\beta}{\sum_{l \in allowed_a} [\tau_l(t)]^\alpha [\eta_{il}(t)]^\beta}, & \text{如果 } j \in allowed_a \\ 0, & \text{否则} \end{cases} \quad (1)$$

当蚂蚁从节点 i 到节点 j 后, 节点(项) i 上的局部信息素更新公式如下:

$$\tau_i = (1 - \lambda) \tau_i + \lambda \tau_0, \lambda \in (0, 1) \quad (2)$$

当所有的蚂蚁完成一次循环, 各节点(项)上的全局信息素更新公式如下:

$$\tau_i(t + \Delta t) = (1 - \rho) \tau_i(t) + \rho \Delta \tau_i^{best}(t, t + \Delta t) \quad (3)$$

$$\Delta \tau_i^{best}(t, t + \Delta t) = \begin{cases} \frac{1}{L_{max}}, & \text{如果节点 } i \text{ 包含在最优路径中} \\ 0, & \text{否则} \end{cases} \quad (4)$$

式中 $\lambda, \rho \in [0, 1]$ 表示路径上信息素挥发因子, 初始时刻 $\tau_i^{best}(0) = 0$. L_{max} 为全局长频繁项集长度.

算法 3: 蚁群挖掘频繁项集

// 参数初始化

令时间 $t = 0, \Delta t = 0$, 最大循环次数 NC_{max} ;

初始化信息量 $\tau_i(t) = \tau_0, \Delta \tau_i^{best}(t, t + \Delta t) = 0$;

$\alpha, \beta, \lambda, \rho$ 为常数; 蚂蚁数 Ant_{num} ;

// 生成频繁 K -项集 ($k > 1$)

for($NC = 1; NC < NC_{max}; NC++$) // for1 循环次数控制

{

for($a = 1; a \leq Ant_{num}; a++$) // for2 Ant_{num} 个蚂蚁周游

{ $k = 1$;

将蚂蚁 a 随机置于 L_1 的节点 i 上;

$L_k(a) = \{I_i\}$; // 蚂蚁 a 挖掘的频繁项集;

$tabu_a = \varphi$; // 蚂蚁 a 的禁忌表清空

$p = 0$; // p 为蚂蚁禁忌表指针索引号

$tabu_a(p) = i$; // 将禁忌表指针指向蚂蚁 a 的当前位置 i

$tabu_a = tabu_a \cup i$; // 将蚂蚁 a 的当前位置 i 加入其禁忌表

for($p = 1; p < size(L_1); p++$) // for3 每个蚂蚁需访问所有连通的节点

{

蚂蚁 a 按照式(1) 选择具有最大状态转移概率的节点 j ;

$tabu_a = tabu_a \cup j$; // 将 j 加入到禁忌表 $tabu_k$

$allowed_a = v - tabu_a$; // 更新可行路径节点

if($Support(L_k(a) \cup \{I_j\}) \geq Min_Support$) // 加入新节点后项集是频繁的

$\{L_{k+1}(a) = L_k(a) \cup I_j$;

$tabu_a(p) = j$; // 修改禁忌表指针, 将蚂蚁移到新的节点 j

按照式(2) 更新信息素; // 局部信息素更新

$k = k + 1$;

}

}

// for3 结束

$L(a) = L_1(a) \cup L_2(a) \cup \dots \cup L_{max}(a)$; // $L_{max}(a)$ 为蚂蚁 a 发现的最长频繁项集

} // for2 结束

确定本次循环找到的最优路径 $L_{max} = size(L_{max}(a)), a = 1, 2, \dots, Ant_{num}$; // L_{max} 为最长频繁项集长度

按照式(3) 和(4) 更新信息素; // 全局信息素更新

$t = t + \Delta t$;

} // for1 结束

$L = L(1) \cup L(2) \cup \dots \cup L(M)$; // 重复的集合只计算一次

4 数据实验

实验最佳参数组合为 $\alpha = 1, \beta = 2, \lambda = 0.1, \rho = 0.2, \tau_0 = 1/size(L_1), Ant_{num} = 30, NC_{max} = 300$. 实验数据集分别为 T40, I10, D100K, Mushroom, BMS-Webview-1 和 Retail^[13], 数据集的特征如表 1 所示.

表 1 数据集特征

数据集	项集数	事物数	事物最小长度	事物最大长度	事物平均长度
T40, I10, D100K	942	100 000	4	77	39
Mushroom	119	8 124	23	23	23
BMS-Webview-1	497	59 602	1	267	2
Retail	13 103	41 373	1	52	9

我们对这四个数据集进行测试, 得到在不同数据集和不同支持度阈值下的运行时间如图 1~图 4 所示. 实验结果表明, 同样条件下, 无论事物数大还是小、项集长还是短、数据相关性强还是弱, ACO 算法执行效率比 Apriori 算法平均提高了 20%~40%. 虽然有限迭代的蚁群算法无法找到所有的完全频繁项集, 但是该算法倾向找到长项集和有意义的规则, 这对现实更有意义. 因此, 该算法对于大型数据库中完全频繁项集的挖掘是智能高效的.

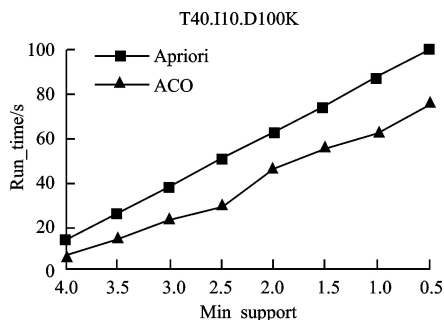


图1 T40. I10. D100k

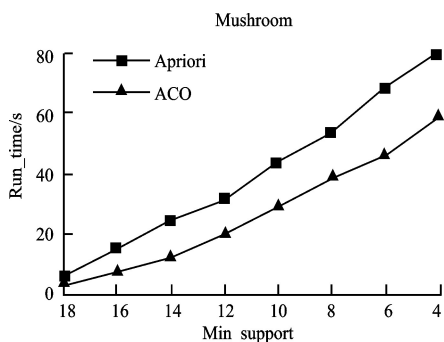


图2 Mushroom

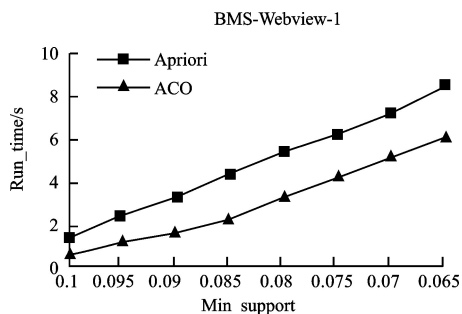


图3 BMS-Webview-1

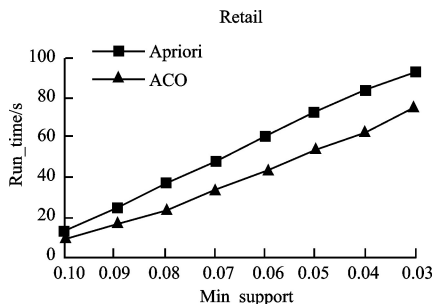


图4 Retail

集,直接在项集关联图上发现完全频繁项集,克服了Apriori算法的弱点.在不同的基准数据集上对算法的性能进行测试表明,新的算法有效提高了挖掘完全频繁项集的智能和性能.

参考文献:

- [1] Science Staff . Dealing with data: challenges and opportunities[J]. Science, 2011, 331(6018): 692-693.
- [2] Chen M S, Han J, Yu P S. Data Mining: An overview from database perspective [J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8 (6): 866-883
- [3] Jiawei Han, Micheline Kamber. 著, 范明, 译. 数据挖掘: 概念与技术[M]. 北京: 机械工业出版社, 2006.
- [4] R Agrawal, R Srikant . Fast algorithms for mining association rules [C] // Proceeding of the 20th international Conference on very large database, Morgan Kaufman: Pub Inc, 1994: 487-499.
- [5] M Dorigo, V Maniezzo, A Coloni. The ant system: optimization by a colony of cooperating agents[J]. IEEE Trans On System Man and Cybernetics -Part B, 1996, 26(1): 29-41
- [6] Lumer E, Faieta B. Diversity and adaptation in populations of clustering ants[C] // Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to Animats. Cambridge, MA: MIT Press. 1994: 501-508
- [7] Parepinelli R S, Lopes H S, Freitas A. An ant colony algorithm for classification rule discovery[C] // Data Mining Heuristic Approach. [S. l.]: Idea Group publishing, 2002.
- [8] 屠莉, 陈峻. 挖掘关联规则的蚁群算法[J]. 南京邮电大学学报: 自然科学版, 2006, 26(5): 36-39.
- [9] 沈国强, 覃征. 一种新的多维关联规则挖掘算法[J]. 小型微型计算机系统, 2006, 27(2): 291-294.
- [10] R J Kuo, C W Shih. Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan [J]. Computers and Mathematics with Applications, 2007(54): 1303-1318.
- [11] R J Kuo, S Y Lina, C W Shih. Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan [J]. Expert Systems with Applications, 2007(33): 794-808.

(下转第 151 页)

5 结束语

本文给出了蚂蚁挖掘完全频繁项集的算法, 算法只需要扫描一次数据库, 不需要产生大量的候选

实验. 每类图像本文的查准率均高出基于累加直方图的检索(CH)许多. 本文在相似度计算前缩小了检索的范围,减少了与样本图像特征差异较大的图像产生的干扰,因而能够在查准率上提高了很多.

6 结束语

在基于内容的图像检索中,一般是直接用样本图像的特征与图像库图像的特征进行相似性度计算,忽略了图像库不同类别图像之间的排斥性,相似图像之间的吸引力. 本文通过改进 K-means 算法的初始类心选取方法对图像库进行聚类,提高了图像检索效率. 由于颜色特征作为单一的特征,不能够完全准确的描述图像内容,如何用其他特征描述图像并进行检索,是下一步工作,以进一步提高图像检索的准确率.

参考文献:

- [1] Velpkamp R C. Content-based image retrieval system: a survey[R]. Utrecht: Utrecht University, Mirela Tanase Department of Computing Science, 2001.
 - [2] Rui Y, Huang TS, Chang S-F. Image retrieval: current techniques, promising directions and open issues[J]. Journal of Visual Communication and Image Representation, 1999, 10(3): 39-62.
 - [3] Huang Yuan-yuan, He Yun-feng. An image retrieval method based on color features[J]. Journal of Chinese Computer Systems, 2007, 28(7): 1277-1281.
 - [4] 吕明磊, 刘冬梅, 曾智勇. 基于改进 K-means 算法的图像检索方法[J]. 计算机应用, 2013, 33(S1): 195-198.
 - [5] 田玉敏, 林高全. 基于颜色特征的彩色图像检索方法[J]. 西安电子科技大学学报: 自然科学版, 2002, 29(1): 127-131.
 - [6] 宋琳琳, 王相海. 一种基于 HVS 加权颜色特征的图像检索算法[J]. 微电子学与计算机, 2008, 25(6): 60-62.
 - [7] 张鑫, 温显斌, 孟庆霞. 基于颜色特征的图像检索方法研究[J]. 计算机科学, 2012, 39(11): 243-245.
 - [8] 常哲, 侯榆青, 李明俐, 等. 综合颜色和纹理特征的图像检索[J]. 微型计算机系统, 2011, 32(1): 161-164.
- 作者简介:**
马娟娟 女, (1990-), 硕士研究生. 研究方向为图像检索、软件工程. E-mail: majuanjuan129@163.com.
张伟 男, (1956-), 副教授. 研究方向为模式识别、软件工程.
李朝锋 男, (1971-), 教授. 研究方向为图像及视频质量评价, 人工智能, 模式识别.
杨弘 男, (1966-), 研究员. 研究方向为渔业经济与信息技术.
-
- (上接第 147 页)
- [12] 黄红星, 景林, 黄习培. 挖掘最大频繁项集的朴素蚁群优化算法[J]. 计算机工程与设计, 2011, 32(1): 301-304.
 - [13] Frequent Itemset Mining Implementations Repository [EB/OL]. [2014-03-23]. <http://fimi.cs.helsinki.fi/>, 2004.
- 作者简介:**
黄红星 男, (1979-), 硕士, 讲师. 研究方向为优化算法及其应用. E-mail: qingtiandaxia@126.com.