第37卷 第16期 计 算 机 工 程 2011年8月
Vol.37 No.16 Computer Engineering August 2011

·软件技术与数据库· 文章编号：1000—3428(2011)16—0045—03 文献标识码：A 中图分类号：TP391

基于广义离散 Morse 理论的强关联规则挖掘

刘 俊，刘希玉

(山东师范大学管理与经济学院，济南 250014)

摘 要：针对强关联规则的挖掘问题，提出构造事务数据库的单元复形，利用广义离散 Morse 理论发现强关联规则的方法。在基本的离散 Morse 理论和关联规则的基础上延伸得到广义离散 Morse 理论和强关联规则的定义，通过在事务数据库的单元复形上定义离散 Morse 函数挖掘强关联规则，例证表明该方法的可行性和高效性。

关键词：离散 Morse 理论；拓扑；广义离散 Morse 函数；广义离散梯度；强关联规则

Strong-association Rules Mining Based on Generalized Discrete Morse Theory

LIU Jun, LIU Xi-yu

(School of Management and Economics, Shandong Normal University, Jinan 250014, China)

【Abstract】For the problem of strong-association rules mining, a method is proposed which constructs a cell complex on transaction database and uses generalized discrete Morse theory to find the strong-association rule. It gets the definition of generalized discrete Morse theory and strong-association rule by extending the basic discrete Morse theory and association rule, mining the strong-association rule by defining discrete Morse theory on cell complex of transaction database. Example verifies the feasibility and efficiency of the method.

【Key words】discrete Morse theory; topology; generalized discrete Morse function; generalized discrete gradient; strong-association rule

DOI: 10.3969/j.issn.1000-3428.2011.16.015

1 概述

Morse 理论最初的应用主要是通过分析黎曼流形上 Morse 函数的临界点来研究流形的拓扑结构，在几何体和拓扑空间之间起到非常重要的作用。随后，在其基础上又进行了进一步的研究，将 Morse 理论引入到离散结构中，形成了离散 Morse 理论，使得 Morse 理论的应用更加广泛。

本文定义一种广义离散 Morse 理论，用来挖掘强关联规则。

2 相关概念

在介绍离散 Morse 理论之前，先了解一些涉及到的相关概念。

(1)单元(cell)：一个维度为 p 的单元 $\alpha^{(p)}$固定同胚于一个开放的 p 维球体，即：

$\{x\in R^p:|x|<1\}$

(2)单元复形(cell-complex)：一个单元复形由一系列 p 维单元黏结而成。从一个 0 维单元(顶点)K^0 开始，将 1 维单元(线段)沿边界黏结到 K^0 上，可得到单元复形 K^1，将 2 维单元(面)沿边界黏结到单元复形 K^1 上得到单元复形 K^2。依此类推，可得到单元复形 K^n。

(3)超图[1]：一个超图是一个序偶(N,L)，其中，N 为节点；L 为族，且 L 的每个元素都是节点的族，称为超链接。

(4)超树[2]：如果在一个有向超图中，每个节点最多为一个超链接的源节点，且它不含有任何环，则该超图称为超树。

(5)正规部分[1]：一个超图(N,L)的正规部分是简单图(N,R)的连通部分，其中，R 是(N,L)的正规超链接的集合。

3 离散 Morse 理论

离散 Morse 理论的主要目的是建立单元复形的离散 Morse 函数或离散梯度向量域，通过对离散 Morse 函数或离散梯度向量域的研究得到单元复形的拓扑信息和属性。

3.1 离散 Morse 函数

定义在给定的单元复形上的离散 Morse 函数是一个实值函数，它随着维度不断增加。其定义如下：

定义 1(离散 Morse 函数) 对一个单元复形 K 的每个单元都映射一个实数的函数 f: $K\to R$，如果对每一个 p 维单元 $\alpha^{(p)}\in K$，它都满足：

$$\#\{\tau^{(p+1)}\ \mathrm{f}\ \alpha^{(p)}:f(\tau)\leqslant f(\alpha)\}\leqslant 1$$

$$\text{and}\ \#\{\nu^{(p-1)}\ \mathrm{p}\ \alpha^{(p)}:f(\nu)\geqslant f(\alpha)\}\leqslant 1$$

那么函数 f 是一个定义在单元复形 K 上的离散 Morse 函数[3]。

从离散 Morse 函数的定义中可以看出，函数 f 最多分配一个比 $f(\alpha)$大的数给比单元 α 维度低的单元，同时也最多分配一个比 $f(\alpha)$小的数给比单元 α 维度高的单元。

定义 2(临界单元[4]) 函数 f: $M\to R$ 是一个离散 Morse 函数，一个 p 维单元 $\alpha^{(p)}$是一个临界单元，如果它满足下面的条件：

$$\#\{\tau^{(p+1)}\ \mathrm{f}\ \alpha^{(p)}:f(\tau)\leqslant f(\alpha)\}=0$$

$$\#\{\nu^{(p-1)}\ \mathrm{p}\ \alpha^{(p)}:f(\nu)\geqslant f(\alpha)\}=0$$

3.2 离散梯度向量域

虽然定义在单元复形上的离散 Morse 函数反映了单元复形的一些属性，但是有关单元复形的拓扑信息更多地存在于

基金项目：国家自然科学基金资助项目(60873058)；山东省自然科学基金资助项目(Z2007G03)

作者简介：刘 俊(1986－)，女，硕士研究生，主研方向：Morse 理论，数据挖掘；刘希玉，教授、博士

收稿日期：2011-01-28 **E-mail：**liujun271@163.com

单元复形的离散梯度向量域中。

组合向量域[1]V 定义在单元复形 K 上，是相关单元 $\{\alpha^{(p)}\ \mathrm{p}\ \beta^{(p+1)}\}$ 的一个不相交的集合。集合中的单元满足：

$$\{\alpha^{(p)}\ \mathrm{p}\ \beta^{(p+1)}\}\in V \Rightarrow V(\alpha)=\beta \text{ and } V(\beta)=0$$

利用从单元 α 指向单元 β 的箭头表示这个配对，如果单元 α 不属于任何配对，则有 $V(\alpha)=0$。

V-路径[5]是单元 $\alpha_0^{(p)},\beta_0^{(p+1)},\alpha_1^{(p)},\beta_1^{(p+1)},\cdots,\alpha_r^{(p)},\beta_r^{(p+1)}$的一个交替序列，这些单元满足：

$$V(\alpha_i^{(p)})=\beta_i^{(p+1)} \text{ and } \beta_i^{(p+1)}\ \mathrm{f}\ \alpha_{i+1}{}^{(p)}\neq\alpha_i^{(p)}$$

如果 $r\geqslant 1$ 且 $\alpha_{r+1}=\alpha_0$，就说一个 V-路径是非平凡的且闭合的。

在讨论了组合向量域和 V-路径后，可以得到离散梯度向量域的定义。

定义 3(离散梯度向量域) 一个离散梯度向量域是一个带有非平凡的闭合 V-路径的组合向量域。

从离散梯度向量域的角度定义临界单元：如果一个单元 α 不和任何一个其他的单元成对，则单元 α 是临界单元，即：$V(\alpha)=0$。

3.3 离散 Morse 函数与离散梯度向量域的关系

文献[1]证明了对于每一个离散 Morse 函数 f，总存在一个和 f 具有相同临界单元的离散梯度向量域 V。同样，对于每一个离散梯度向量域 V，总存在一个和 V 具有相同临界单元的离散 Morse 函数 f。离散 Morse 函数和离散梯度向量域在反映单元复形的拓扑结构方面是互相对应的。

4 算法的构建

在了解了离散 Morse 理论后，可以知道，离散 Morse 理论和离散梯度向量域都是反映单元复形拓扑结构的重要理论，因此，如何构建离散 Morse 函数和离散梯度向量域就成为一个重要的问题。

文献[1]给出了构建离散 Morse 函数和离散梯度向量域的思路，下面分别给出其具体算法的流程。

4.1 离散 Morse 函数的构建

在该部分涉及到生成树的概念：如果 $N'\subset N, L'\subset L$，则得到图(N,L)的子图；如果 $N'=N$，则得到图(N,L)的生成子图；如果生成子图(N',L')中不包含环，则得到一棵生成树。

构建离散 Morse 函数算法的流程如下：

(1)在有限单元复形 K 上构造一棵生成树 T。

(2)在 K 中构造 T 的补图 G。

(3)G 的节点：对应于 K 的一个三角形或者是 K 的一条边界边(即该条边在 K 中，但是不在 T 中)。

(4)G 的边：如果 G 的 2 个节点对应的是 2 个有共享边的三角形，或者对应的是一个三角形和 K 的边界边，则这 2 个节点之间存在一条链接。

(5)为生成树 T 分配值。

(6)将生成树的根节点和与其相关联的链接初始化为 c。

(7)为 T 的其余节点分配值，其值为该节点到根节点的距离+c。

(8)为 T 的每个链接分配值，其值为与该链接相关联的 2 个节点中值较大的那个节点的值。

(9)直到 T 的每个节点和链接都分配到值。

(10)为 K 中生成树 T 的补图 G 分配值(G 中值的分配从度为 1 的节点开始，节点的度表示该节点所对应的三角形中未分配值的边的条数)。

(11)将 G 中度为 1 的节点初始化为 T 中的最大值+1。

(12)为该节点对应的三角形中的自由边(未分配值的边)分配相同的值。

(13)重置 G 中节点的度。

(14)返回步骤(11)执行，直到 G 中所有节点都已完成。

(15)得到单元复形 K 的离散 Morse 函数 f。

4.2 离散梯度向量域的构建

构建离散梯度向量域算法的流程如下：

(1)在单元复形 K 上构造超树 HF。

(2)对于 HF 的正规部分 R，选择一个节点作为 R 的根。

(3)如果正规部分 R 是一个临界部分，则 R 的任何节点都可以作为根节点。

(4)否则，如果正规部分 R 不是一个临界部分，而恰有一个节点是一个环或非正规超链接的源节点，则该节点作为根节点。

(5)将 R 中的叶节点和与其相关联的唯一链接进行配对。

(6)R 中未配对的节点记为 $R^{(1)}$，并在 $R^{(1)}$上执行第(5)步。

(7)反复执行第(6)步，直到除根节点外 R 的所有节点都已配对。

(8)处理根节点。

(9)如果正规部分 R 是一个临界部分，则该根节点不配对，成为临界点。

(10)否则，如果正规部分 R 不是一个临界部分，则该根节点和与它相关联的环或非正规超链接配对。

(11)处理非正规部分。

5 强关联规则挖掘

对离散 Morse 理论进行扩展得到广义离散 Morse 理论，并用来挖掘强关联规则。

定义 4(广义离散 Morse 函数) 为一个单元复形 K 的每个单元都映射一个实数函数 f: $K\to R$，如果对每一个 p 维单元 $\alpha^{(p)}\in K$，它都满足：

$$\{\tau^{(p+1)}\ \mathrm{f}\ \alpha^{(p)}: f(\tau)\geqslant f(\alpha)\} \text{ and } \{\nu^{(p-1)}\ \mathrm{p}\ \alpha^{(p)}: f(\nu)\leqslant f(\alpha)\}$$

那么函数 f 是一个定义在单元复形 K 上的广义离散 Morse 函数。

定义 5(广义离散梯度) 在一个单元复形中，如果存在单元 $\alpha\ \mathrm{p}\ \beta$，有 $f(\alpha)\geqslant f(\beta)$，则在 α、β 之间画一个箭头，其中，α 是箭头的尾；β 是箭头的头。这样形成的梯度域称为广义离散梯度。单元 α 可以是多个箭头的尾。

定义 6(强关联规则) 关联规则[6]是形如 $A\Rightarrow B$ 的蕴含式，其中，$A\subset I, B\subset I$，$I=\{i_1,i_2,\cdots,i_m\}$ 是项的集合，并且 $A\cap B=\varnothing$。规则 $A\Rightarrow B$ 在事务集 D 中成立，具有支持度 s，其中，s 是 D 中事务包含 $A\cup B$ 的百分比；规则 $A\Rightarrow B$ 在事务集 D 中具有置信度 c，如果 D 中包含 A 的事务同时也包含 B 的百分比是 c，即：

$$support(A\Rightarrow B)=P(A\cup B)$$

$$confidence(A\Rightarrow B)=P(B/A)$$

定义置信度为 100%的关联规则为强关联规则。

对于一个给定的事务数据库，可以利用广义离散 Morse 理论通过以下方法挖掘出其中存在的强强关联规则，这里以含有 5 个项(I_1、I_2、I_3、I_4、I_5)的事务数据库为例，步骤如下：

(1)根据给定的事务数据库确定项的个数，建立项集的树形图。

对于含有 5 个项的事务数据库，对项进行排列组合得到频繁 1-项集、频繁 2-项集、频繁 3-项集、频繁 4-项集、频繁 5-项集，构造它的树形图如图 1 所示。

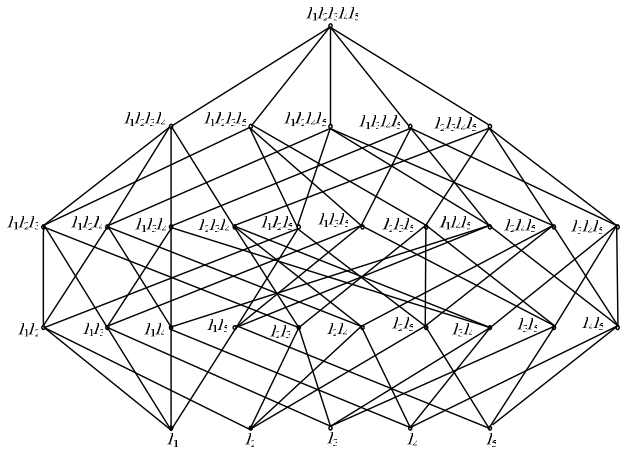

图 1　含有 5 个项的树形图

(2)扫描数据库，计算每个项集的计数。根据如果一个集合不能通过测试，则它的所有超集也都不能通过相同的测试这一性质，对于包含了支持度计数小于最小支持度项集的那些集合，它们的支持度就可以不用再计算，这样就大幅提高了效率。

该例中首先计算频繁 1-项集的支持度计数，去掉支持度计数小于 2 的频繁 1-项集，那么同时可以去掉含有支持度计数小于 2 的频繁 1-项集的频繁 2-项集，只需要再计算剩余频繁 2-项集的支持度计数即可。同样的频繁 3-项集、频繁 4-项集、频繁 5-项集的支持度计算也是如此，反映在树形图中即如图 2 所示。

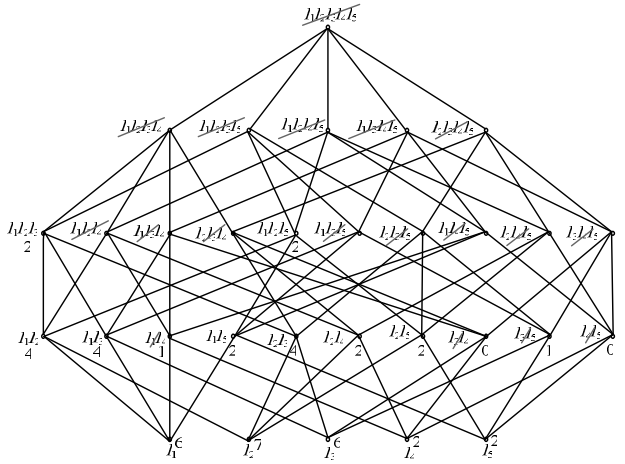

图 2　支持度计数的计算示意图

(3)根据树形图可以得到一个单元复形 K：其中频繁 1-项集为 0-维单元(顶点)，频繁 2-项集为 1-维单元(线段)，……，频繁 n-项集为(n-1)-维单元。各个频繁项集的支持度计数分别作为各维单元的权重。

根据图 2 的树形图得到本例的单元复形，此时树形图中只剩余频繁 1-项集、频繁 2-项集、3-项集，因此，得到的单元复形是一个平面图，并将各项集的支持度计数标注于对应的单元上，如图 3 所示。

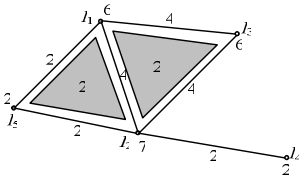

图 3　根据图 2 得到的事务数据库的单元复形

(4)用单元复形中的最大计数减去各单元的计数，得到单元复形的广义离散 Morse 函数，并在该单元复形上构造广义离散梯度。

该例中最大计数为 7，用 7 减去图 3 中标注的各个单元的权重，得到单元复形的广义离散 Morse 函数，如图 4 所示。

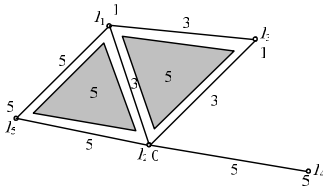

图 4　广义离散 Morse 函数

同时根据广义离散梯度的定义，在图 4 的基础上构造单元复形的离散梯度域，如图 5 所示。

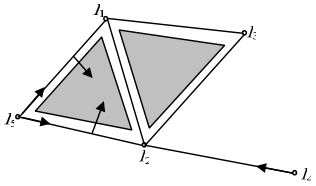

图 5　广义离散梯度

(5)单元复形的广义离散梯度图中的每个箭头代表一个“箭尾单元=>箭头指向单元”的强关联规则。

本例中单元 $I_2I_5 \rightarrow I_1$ 的箭头表示了下面的强关联规则：

$confidence(I_2 \wedge I_5 => I_1)=100\%$

最终得到包含 5 个项的事务数据库中的强关联规则为：

$confidence(I_5 => I_1)=100\%$

$confidence(I_5 => I_2)=100\%$

$confidence(I_4 => I_2)=100\%$

$confidence(I_1 \wedge I_5 => I_2)=100\%$

$confidence(I_2 \wedge I_5 => I_1)=100\%$

6　结束语

关联规则挖掘是一种发现事务之间广泛联系的重要工具，而广义离散 Morse 理论的应用使得对特殊关联规则的挖掘变得更加直观、简单。

广义离散 Morse 理论是对基本离散 Morse 理论的扩展，将其应用于更加广泛的领域是进一步研究该理论的重要途径，这也是下一步要做的工作。

参考文献

[1] Lewiner T. Constructing Discrete Morse Functions[EB/OL]. (2002-06-25). http://www.matmidia.mat.puc-rio.br/tomlew/pdfs/tomlew_msc.pdf.

[2] 张丽娜, 顾耀林. 一种基于离散梯度向量域的可视化应用研究[J]. 计算机工程, 2006, 32(16): 218-220.

[3] Lewiner T, Lopes H, Tavares G. Towards Optimality in Discrete Morse Theory[J]. Experimental Mathematics, 2003, 12(3): 271-286.

[4] 薛红娟, 顾耀林. 拓扑分析在海洋特征提取中的应用[J]. 计算机工程, 2009, 35(3): 263-265.

[5] Forman R. Morse Theory for Cell Complexes[J]. Advances in Mathematics, 1998, 134(1): 90-145.

[6] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 2 版. 范　明, 孟小峰, 译. 北京: 机械工业出版社, 2007.

编辑　顾逸斐