Management Science and Engineering 管理科学与工程, 2012, 1, 24-28
http://dx.doi.org/10.12677/mse.2012.13004 Published Online September 2012 (http://www.hanspub.org/journal/mse.html)

Classification Rules Based on Discrete Morse Theory

Yucai Zhan, Xiyu Liu

School of Management science and Engineering, Shandong Normal University, Ji'nan
Email: 191806347@qq.com

Received: Mar. 26th, 2012; revised: May 9th, 2012; accepted: Jun. 10th, 2012

Abstract: With the emergence and development of discrete Morse theory, it has been widely applied, such as Topology, Computer Graphics and geometric modeling. Classification mining is the process of learning through the training sample data set to construct classification rules, and is an important aspect of data mining, knowledge discovery. The essence of the classification mining is to get high accuracy, interesting and easy to understand classification rules. In this paper, discrete Morse Theory is used to construct classifier, Purpose is to elect the useful information which people interested in from large amounts of data. First we summarizes the relevant theoretical knowledge about data mining and discrete Morse theory, describes the relationship between the Hasse diagram, discrete gradient vector field and discrete Morse function, and describes the algorithm to build a discrete gradient vector field and discrete Morse function. Finally, for the problem of classification mining, we construct the simplicial complex about the classification rules, use the discrete Morse theory to solve the problem of classification rules, and show the feasibility and efficiency of the method through the example.

Keywords: Discrete Morse Function; Discrete Gradient Vector Field; Classification Rules

基于离散 Morse 方法的分类规则研究

战玉彩，刘希玉

山东师范大学管理科学与工程学院，济南
Email: 191806347@qq.com

收稿日期：2012 年 3 月 6 日；修回日期：2012 年 5 月 9 日；录用日期：2012 年 6 月 10 日

摘　要：随着离散 Morse 方法的出现和发展，其应用也越来越广泛，主要领域有拓扑学、计算机图形学和几何建模等。分类规则挖掘则是通过对训练样本数据集的学习构造分类规则的过程，是数据挖掘、知识发现的一个重要方面。分类规则挖掘的实质是希望得到高准确性、有趣的和易于理解的分类规则。本文利用离散 Morse 方法构造分类器，从大量数据中选出人们感兴趣的有用信息。首先综述了数据挖掘和离散 Morse 方法的相关理论知识，描述了 Hasse 图、离散梯度向量域和离散 Morse 函数三者之间的关系，并介绍了构建离散梯度向量域和离散 Morse 函数的算法。最后针对分类的挖掘问题，构造了关于分类规则的单纯复形，并利用离散 Morse 方法分析解决了关于分类规则的问题，并通过例证表明了该方法的可行性和高效性。

关键词：离散 Morse 函数；离散梯度向量域；分类规则

1. 引言

数据挖掘是知识发现的一个步骤，是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取出隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[1]。分类规则挖掘是数据挖掘领域中最重要的研究领域之一，同时，也是其它诸如

人工智能、模式识别、人工神经网络等学科的重要研究内容，并且有丰富的结果和广泛的应用，因此对分类规则挖掘的研究是很有必要的[2]。

Morse 理论最初是由 Marton Morse 在其论文中提出的，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，是一种非常有用的优化工具。随后，Forman 在 Morse 理论的基础上进行了进一步的研究，并将离散结构引入 Morse 理论中，形成了离散 Morse 方法本文又将离散 Morse 方法延伸到狭义离散 Morse 方法，并以此来解决分类规则挖掘问题。

2. 相关理论

根据参考文献[3-5]，下面给出几个关于离散 Morse 方法的概念描述：

定义 1 单元： p 维单元 $\alpha^{(p)}$是一个同胚于开放的 p 维球体：$\left\{x \in R^p : \|x\| < 1\right\}$。当单元的维度 p 很明显的时候，我们可以用 α 来代替 $\alpha^{(p)}$。

定义 2 CW-复形： 零维单元 K^0 是一个点，沿着它的边界连接一维单元线段，得到 K^1，沿着 K^1 的边界连接二维单元面(如三角形)，得到 K^2，沿着 K^1 边界连接三位单元体(如正方体)，得到 K^3，依次类推下去，就可得到每一个 n 维单元 K^n。这些单元组成的离散集合就建立了一个 CW-复形。如果单元的个数是有限的，则我们就说 CW-复形是有限的。

定义 3 Hasse 图： Hasse 图是通过结点的维数来确定结点的队列安排的。

单纯复形 K 的 Hasse 图是满足以下条件：

1) H 的每个结点代表 H 的一个单元。

2) H 结点间的链接代表 K 的相关单元。

3) 每个链接的根结点是单纯复形 K 中的最高维单元。

定义 4 离散 Morse 方法： 函数 f：$K \to R$ 是单纯复形 K 上每个单元到全体实数的一个映射，对任意单元 $\sigma^{(p)} \in k$，如果满足以下条件：

$$\#\left\{\tau^{(p+1)} \succ a^{(p)} : f(\tau) \le f(a)\right\} \le 1$$

且

$$\#\left\{v^{(p-1)} \prec a^{(p)} : f(v) \ge f(a)\right\} \le 1$$

则函数 f 就叫做单纯复形 K 上的离散 Morse 函数。

定义 5 组合梯度向量域： 单纯复形 K 上的组合梯度向量域是相关单元 $\left\{a^{(p)} \prec \beta^{(p+1)}\right\}$ 的一个不相交的集合。我们可以把离散梯度向量域定义成函数 V：$K \to K \cup \{0\}$，且满足

$$\left\{a^{(p)} \prec \beta^{(p+1)}\right\} \in V \Rightarrow V_{(a)} = \beta$$

且

$$V(\beta) = 0 \text{。}$$

从单元 α 到单元 β 画一个箭头来表示这个配对。如果单元 σ 不属于任何配对，则 $V(\sigma) = 0$。

定义 6 V-路径： 单元 $a_0^{(p)}$，$\beta_r^{(p+1)}$，$a_1^{(p)}$，$\beta_1^{(p+1)}$，……，$\alpha_r^{(p)}$，$\beta_r^{(p+1)}$ 满足

$$V\left(\alpha_i^{(p)}\right) = \beta_i^{(p+1)}$$

且

$$\beta_i^{(p+1)} \succ \alpha_{i+1}^{(p)} \ne a_i^{(p)}$$

则这样的单元构成的交替序列叫做 V-路径。

定义 7 离散梯度向量域： 如果当 $r \ge 1$ 时满足 $a_r + 1 = a_0$，则 V-路径是非平凡且闭合的。不存在非平凡且闭合的 V-路径的组合向量域叫做离散梯度向量域。

定义 8 广义的离散 Morse 函数： 为一个单纯复形 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 $a^{(p)} \in K$，它都满足：

$$\left\{\tau^{(p+1)} \succ a^{(p)} : f(\tau) \ge f(a)\right\}$$

且

$$\left\{V^{(p-1)} \succ a^{(p)} : f(v) \le f(a)\right\}$$

那么函数 f 是一个定义在单纯复形 K 上的广义离散 Morse 函数。

定义 9 广义离散梯度向量域： 在一个单纯复形中，如果存在单元 $\alpha \prec \beta$，有 $f(a) \ge f(\beta)$，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。

3. 算法步骤

在参考文献[6]中给出了构建离散 Morse 函数和构建离散梯度向量域的方法，本文将给出详细的构建离散 Morse 函数和构建离散梯度向量域的算法步骤。

3.1. 超树和超图

超图是一个配对(N, L)，其中 N 表示图中所有的结点的集合，L 表示跟 N 相关的所有关联。L 中的元素称为超链。规则超链指的是只跟两不同结点相关的链接。只与一个结点相关的链接是环。而非规则超链是指环和同时跟三个或以上的结点相关的链接。

如果一个简单有向超图的每个结点至多是一个超链的根，并且不包含任何超环。则称其为超树。

T. Lewiner 在文献[7]中已证明：超树 HF 的任一规则部分 R 要满足下面的条件：

1) HF 的规则部分 R 是简单树。

2) R 中至多有一个结点是环或者是非规则部分的超链接。

对于超树，如果它的规则成分中不存在与根结点相关的环或不规则超链，则称之为临界部分。

3.2. Hasse 图、离散 Morse 函数及离散 Morse 向量域之间的关系

对于每个离散 Morse 函数，总存在一个离散梯度向量域同其有相同临界单元，而对每个离散梯度向量域，也存在一个有相同临界单元的离散 Morse 函数。故构建离散梯度向量域也就是构造相应的离散 Morse 函数。构造离散 Morse 函数和离散梯度向量域需要用到 Hasse 编码，事实上，离散梯度向量域的配对跟 Hasse 图中的配对是一一对应的。离散梯度向量域可以看作是从 Hasse 图中不同层次的超图上提取的超树。

3.3. 离散梯度向量域的构建算法

在单纯复形 K 上构建生成树 ST。为生成树上的所有的结点和链接进行配对，并且确保不在生成树上的结点和链接都处于临界状态。步骤如下：

1) 选取根结点。

2) 在生成树确定一个规则部分 R。

3) 如果 R 为临界部分，则可设任一结点为根结点。

4) 如果 R 不是临界部分，把叶子结点与其唯一的链接配对。

5) 将 R 中未配对的结点和链接放到集合 $R_i=\left(i=1,2,3\cdots\right)$ 中。

6) 重复 4)和 5)，直到所有链接配对完成。

7) 最后一个结点(未配对)设置为根结点。

生成树的结点表示顶点，生成树的链接表示边。用结点到边的箭头来表示这些配对。由于不存在非平凡闭合的 V-路径，所以构建的向量域是离散的。

3.4. 离散 Morse 函数的构建

在单纯复形 K 上构建离散 Morse 函数的步骤如下：

1) 在单纯复形 K 上构建超树 HF。

2) 在 HF 上选取一个规则部分 R，确定规则部分的方法跟构建生成树的方法一致。

3) 如果 R 是孤立的和超树 HF 的补集不相关，则为超树 HF 的结点和链接赋值。

4) 将 R 的根结点及其相关联的链接赋值为 C。

5) 将其它结点赋值为该结点与根结点的距离加 C。

6) 两结点间的链接赋值为该两结点中较大者的值。

7) 如果单纯复形 K 中存在环或者非规则超链接，则为它们赋值为 C。

8) 如果 R 与超树 HF 的补集相关，则为超树 HF 补集的结点和链接进行赋值。

9) 重复以上步骤 5)、6)和 7)，直到所有结点和链接被赋值成功。

10) 得到单纯复形 K 的离散 Morse 函数 f。

4. 利用离散 Morse 方法构造分类模型

数据挖掘是从大型数据库的数据中提取出人们感兴趣的、隐含的、事先未知的潜在有用信息的知识。分类规则挖掘则是通过对训练样本数据集的学习构造分类规则的过程，是数据挖掘、知识发现的一个重要方面。分类规则挖掘的实质是希望得到高准确性、有趣的和易于理解的分类规则。目前常用的分类规则挖掘方法有决策树方法、统计方法、神经网络方法、粗糙集方法和遗传算法等。而分类规则也可以利用离散 Morse 方法通过以下方法来进行挖掘。

数据分类是通过挖掘已有的数据训练集，集中同一类数据对象的共同特征，提取分类规则，对整个数据集进行合理分类的过程[8]。分类的目的是能根据已

经分类的数据构造出一个分类模型，即分类器。

构造一个分类器，需要有一个样本数据集作为输入。该数据集由一组数据库中记录构成，每条记录与一个特定的类别相对应。通常这些训练样本是之前的一些经验数据。

给定一个有 k 个类训练数据集 D，下面利用离散 Morse 方法构造一个分类器，来对数据集进行分类。但是得到的结果仍然可能会有相当数量的分类规则是特定用户不感兴趣的。所以仍需根据用户所需进一步对挖掘过程产生的不感兴趣的分类规则进行调整。这可以利用文献[9]中的思路设定支持度度量来实现。当支持度都大于或等于用户事先设定的最小支持度阈值的分类规则才被认为是有趣的。

下面的例子给出训练集 D(见表 1)。

1) 根据给定的训练数据集，可构造出数据属性的树形结构(如图 1)。七个叶子结点分别表示“年龄 ≤ 30”，“年龄为 31~40”，“年龄 ＞40”，“非学生”，“学生”，“信誉一般”，“信誉良好”，为方便起见，此处我们分别用“*L*”，“*M*”，“*B*”，“*N*”，“*Y*”，“*F*”，“*N*”来表示它们。中间一层结点表示符合七个叶子结点中的两个条件的购买电脑的情况。最上层结点表示符合七个叶子结点中的三个条件的购买电脑的情况。图中的数字的多少表示顾客购买电脑的可能性大小。
由此可得到一定不购买电脑的类成员：{年龄 ≤ 30，不是学生}属于类{不购买电脑}，{年龄 ＞40，信誉良好}属于类{不购买电脑}。

Table 1. One database about customer in a shopping center (training data set)[10]
表 1. 一个商场顾客数据库(训练样本数据集合)[10]

Rid	Age	student	Credit	Buy Computer
1	≤30	no	fair	no
2	≤30	no	excellent	no
3	30~40	no	fair	yes
4	>40	no	fair	yes
5	>40	yes	fair	yes
6	>40	yes	excellent	no
7	30~40	yes	excellent	yes
8	≤30	no	fair	no
9	≤30	yes	fair	yes
10	>40	yes	fair	yes
11	≤30	yes	excellent	yes
12	30~40	no	excellent	yes
13	30~40	yes	fair	yes
14	>40	no	excellent	no

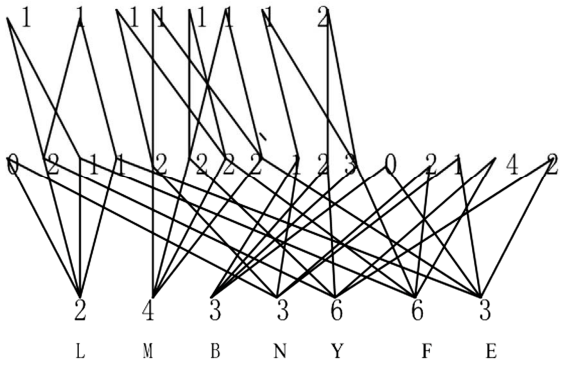

Figure 1. The tree structure of a training data set
图1. 训练数据集的树形结构

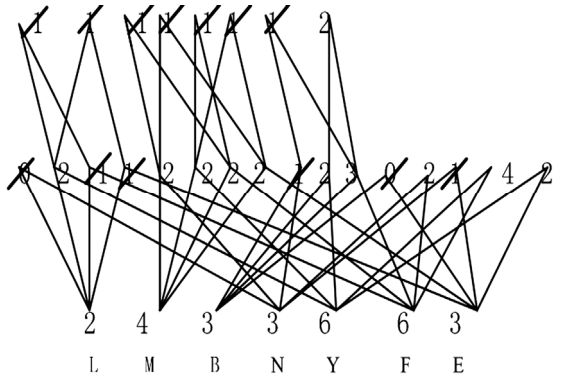

Figure 2. Remove the nodes whose support is less than 2
图 2. 删除支持度 < 2 的结点

2) 本例中，我们取最小支持度为 2，以便使得产生的结果让用户感兴趣。由此可把训练数据集的树形结构的图中支持度小于 2 的结点删除(如图 2)，去掉图中多余的链接之后即可得到 Hasse 图。

3) 由上面得到 Hasse 图即可以构建一个单纯复形 K：一个条件为零维单元(顶点)，两个条件为一维单元(线段)，……N 个条件为 $N-1$ 维单元。

在单纯复形 K 中，最高的维数为 3，所以它是一个平面图。将 Hasse 图中每个结点的支持度与单纯复形中的单元权值一一对应起来，即可得到单元带有数值的单纯复形如下图 3。

4) 下面我们构造一个用离散 Morse 函数：a) 在单纯复形的所有单元中选取权值最大者。b) 对所有单元的权值进行修改：用权值最大者减去该单元的权值得到一个新的权值，把新权值赋给该单元。在此例子中，单纯复形中单元最大权值为 6，用 6 减去各个单元的权值可得到如图 4 的离散 Morse 函数。

5) 构建与此离散 Morse 函数一一对应的离散梯

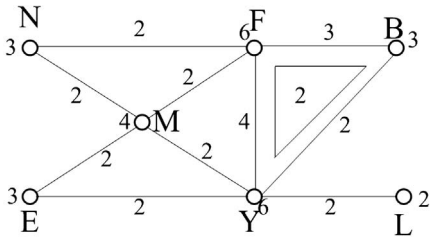

Figure 3. Constructing simplicial complex by the Hasse diagram
图 3. 由 Hasse 图构造的单纯复形

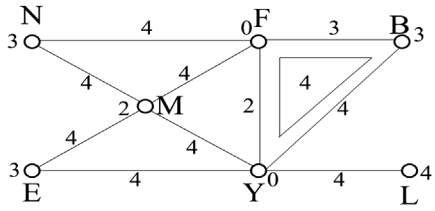

Figure 4. Constructing a discrete Morse function on a simplicial complex
图 4. 在单纯复形上构建离散 Morse 函数

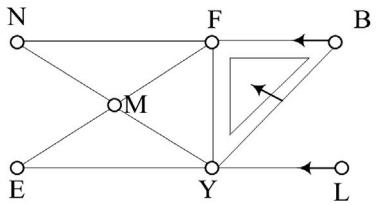

Figure 5. Constructing a Discrete gradient vector field on a simplicial complex
图 5. 在单纯复形上构建离散梯度向量域

度向量域。此处用到广义的离散 Morse 方法的知识，在单元维数较高而离散 Morse 函数值却不高的单元之间画一个箭头，箭头由低维单元指向高维单元，由此构建了广义的离散梯度向量域(如图 5)。

6) 在广义的离散梯度向量域中，箭头表示{合并单元的条件}属于类{购买电脑}，由此可得到购买电脑的类成员。在此处{$L\rightarrow Y$}表示{L, Y}属于{购买电脑}类，即满足条件{年龄 ＜30，学生}属于类{购买电脑}。同理，我们可以得到

{年龄 ＞40，信誉一般}属于类{购买电脑}；

{年龄 ＞ 40，信誉一般，是学生}属于类{购买电脑}。

总上可得到一个分类规则：

{年龄 ＜30，学生}属于类{购买电脑}；

{年龄 ＞40，信誉一般}属于类{购买电脑}；

{年龄 ＞ 40，信誉一般，是学生}属于类{购买电脑}；

{年龄 ≤30，不是学生}属于类{不购买电脑}；

{年龄 ＞40，信誉良好}属于类{不购买电脑}。

5. 结束语

分类是数据分析形式之一，可以用于提取描述重要数据类的模型或预测未来的数据趋势，也使我们日常生活更加快捷、简便。本文研究了构建离散 Morse 函数和离散梯度向量域的方法，并首次把离散 Morse 方法应用到分类规则挖掘上，使得分类规则更加直观、方便。

6. 致谢

首先我要感谢我的导师刘希玉教授，感谢他在我的学习和研究中对我的帮助和指导！感谢国家科学基金(No.61170038)，山东省自然科学基金(No.ZR2011-FM001)对论文的资助和支持！感谢参考文献的作者及其论文给予我的思路和知识资料！感谢我身边的同学和朋友的关心和理解！

参考文献 (References)

[1] U. Fayyad，G. P. Shapiro and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland: AAAI Press, 1996: 82-88.

[2] R. Ayala, L. M. Femandez and J. A. Vilches. Defining discrete Morse functions on infinite surfaces. Universidad de Sevilla, (Spain), 2004: 1-3.

[3] R. Forman. Morse theory for cell complexes. Advances in Mathematics, 1998, 134: 90-145.

[4] 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.

[5] 刘俊, 刘希玉. 基于广义离散 Morse 理论的强关联规则挖掘[J]. 计算机工程, 2011, 37(16): 45-47.

[6] T. Lewiner. Constructing discrete Morse functions. MS Thesis, 2002, 6(6): 33-64.

[7] T. Lewiner，H. Lopes and G. Tavares. Towards optimality in discrete Morse theory. MS Thesis, 2002, 11(2): 1-13.

[8] 张帆. 浅析分类规则挖掘[J]. 科教导刊, 2009, (36).

[9] 蒋良孝, 蔡之华, 刘钊. 一种基于信息增益的分类规则挖掘算法[J]. 中南工业大学学报(自然科学版), 2003, 34(Z1): 69-71.

[10] 韩家炜, 堪博著作; 范明, 孟小峰译. 数据挖掘——概念与技术[M]. 北京: 机械工业出版社, 2007: 3.