《智能信息处理》课程作业

模糊形式背景概念约简的方法研究

王云涛

作业	分数[20]
得分	

2020 年 11 月 11 日

模糊形式背景概念约简的方法研究

王云涛

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘 要 形式概念分析(Formal Concept Analysis，FCA)[1]是由 R. Wille 于 1982 年提出一种从形式背景进行数据分析和规则提取的强有力工具，在信息检索、数据挖掘、软件工程等领域上的应用十分广泛。概念格是形式概念分析理论中的核心数据结构，在信息检索、知识发现等方面得到了广泛的应用。概念格约简是知识发现领域的一个重要课题。在实际生活中，大多数概念都是模糊的，即模糊形式背景。对于引入模糊后的概念格，概念数量相对的更加庞大，但是其中的许多概念对知识发现意义并不大，而且存在着许多彼此相近的概念，这在某种程度上阻碍了知识的发掘，于是概念格的约简方法变得十分重要，以此使得概念数量减少，简化知识表示。本文通过聚类约简方法进行基于模糊概念格的概念约简。

关键词 形式概念分析；模糊形式背景；概念格；模糊概念格；聚类约简；
中图法分类号 TP311.20 DOI 号 10.3969/j.issn.1001-3695.2014.01.030

Research on the method of fuzzy formal context concept reduction

Wang Yuntao

(Computer science and technology，Dalian maritime university, LiaoningDalian,116026,China)

Abstract *Abstract Formal concept analysis is proposed in 1982 by R. Wille a powerful tool for data analysis and rule extraction from the background of the application in information retrieval, data mining, software engineering and other fields is very extensive.Concept lattices are the core data structures in the theory of formal concept analysis. It is put into wide use in information retrieval, knowledge discovery, and so on. Reduction of concept lattice is an important subject in the field of knowledge discovery. In real life, most of the concepts are fuzzy, that is, the fuzzy formal context.For the introduction of fuzzy concept lattice, number of concepts relative to the more large, but where many of the concepts of knowledge found little importance, and there are many ideas similar to each other, which in some extent hindered the mined knowledge, so concept lattice reduction methods become very important, so that the concept of quantity reduction, simplifying knowledge representation.In this paper, the concept of fuzzy concept lattice based on fuzzy concept lattice is reduced by the method of clustering reduction.

Key words Formal concept analysis; Fuzzy formal context; Concept lattice; Fuzzy concept lattice; Clustering reduction;

1 引言

形式概念分析(Formal Concept Analysis，FCA)[1]是由 R. Wille 于 1982 年提出一种从形式背景进行数据分析和规则提取的强有力工具，在信息检索、数据挖掘、软件工程等领域上的应用十分广泛。形式概念分析是在数学的基础上建立的，对组成本

体的概念、属性以及二者的关系等用形式化的语境进行表述，然后根据语境，构造出概念格。概念格是知识刻画方式的一种，人们总是希望可以从它中等到更多可使用的知识，因此实现概念格约简的重要性就体现出来了，它的出现使得形式概念背景中的知识更容易被发现，也更加简洁。在此，采用聚类约简的思想，结合模糊形式背景对格进行概念约简。

2 形式概念分析

2.1 形式背景和概念格[2]

定义 1 设 U 是对象的集合，M 是属性的集合，I 是 U 与 M 间的关系，则称三元组 K=(U,M,I)为一个形式背景。(u.m)∈ I 表示对象 u 具有属性 m。背景可以用一个对象集合、属性集合以及他们之间的二元关系建立起来的表格来表示，它的每行表示某一对象，每列则表示某一属性。

定义 2 设 K=(U,M,I)为形式背景，对象集合 U={u1,u2,..,um}，属性集合 M={m1,m2,..,mn}。

设 m∈ M，具有属性 m 的对象集合。定义

g(m)={u∈ U/(u,m)∈ I}

设 u∈ U，具有属性 m 的对象集合。定义

f(u)={m∈ M/(u,m)∈ I}

定义 3 设 K=(U,M,I)为形式背景，若 A⊆U，B⊆M，那么存在 f(A)={m∈ M/ $\forall$ u∈ A,(u,m)∈ I}和 g(B)={u∈ U/ $\forall$ m∈ B,(u,m)∈ I}

如果集合 A,B 满足 f(A)=B,g(B)=A。则称由(A,B)二元组被称为概念。A 是概念(A,B)的外延，B 是概念(A,B)的内涵。

定义 4 设 K=(U,M,I)为形式背景，存在两个概念(A1,B1)和(A2,B2)满足

$$(A_1, B_1) \le (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_1 \supseteq B_2)$$

则称(A1,B1)为(A2,B2)的子概念，(A2,B2)为(A1,B1)的超概念。由形式背景(U,M,I)中所有概念根据它们之间的层次关系有序组成的集合，常称为(U,M,I)的概念格。

2.2 模糊形式概念[3]

用来表示模糊性概念的集合叫做模糊集合。一般的集合指的是：具有某些属性的事物（对象）的整体。描述这些事物的属性的概念是清晰的，明确的，所以每个事物（对象）对于集合中属性的隶属关系也是明确的，隶属度也是固定取值[0，1]。但是在人们的现实生活中还有着许多模糊的，不清楚的概念，例如成绩好、很晚、好热、有钱等，这些概念所描述的事物不能单纯的用"是"或"否"来说明。模糊集合指的是：具有某些模糊概念所描述的属性的对象的全体。它与一般集合的差别在于：它描述这些事物的属性的概念本身是不清晰的、不明确的，所以每个事物（对象）对于集合中属性的隶属关系也是明确的。基于这一思想，Zadeh 将特征函数的取值范围由 $\{0,1\}$ 推广到连续闭区间 $[0,1]$，建立“隶属函数”的概念，并以此为基础给出了“模糊集合”的定义，进而对经典集合进行了推广。

定义 5 对于论域 U，设 A 是 U 到 $[0,1]$ 上的一个映射，即

$A: U \to [0,1], x \mapsto A(x)$。

则称 A 是论域 U 的模糊子集，或者论域 U 上的模糊集合，简称为模糊集。函数 $A(*)$ 称为模糊集 A 的隶属函数，$A(x)$ 称为 x 相应于模糊集 A 的隶属度。

定义 6 一个模糊形式背景是一个三元组，其中 U 为所有对象的集合，A 为所有属性的 $K = (U, A, \tilde{I})$ 集合，$\tilde{I}$ 是一个定义在域 $U \times A$ 上的模糊关系，每个关系中的元素 $< x, a >$ 都有一个隶属度 $I(x, a)$，简记为 $I_x(a)$，$0 \le I(x, a) \le 1$。

可知，$I_x(a_i)$ 反映对象 x 对属性 a_i 的隶属程度。

定义 7 给定一个模糊形式背景 $K = (U, A, \tilde{I})$ 和阈值 T，定义

$$X' = \{ a \in A \mid \forall x (x \in X \wedge I(x, a) \ge T) \}$$ ；

$B'=\{x\in U \mid \forall a(a\in B\wedge I(x,a)\geq T)\}$。

其中对象集合的子集 $X\subseteq U$，属性集合的子集 $B\subseteq A$。一个具有阈值 T 的概念定义为 $C=(X,B)$，其中 $X\subseteq U$，$B\subseteq A$，$X'=B$，$B'=X$，称 X 为模糊形式概念 C 的外延，B 为模糊形式概念 C 的内涵。

由定义 7 可知，模糊概念的对象与属性之间的关系用隶属度表示。同理，外延和概念之间的关系可以看成是外延和概念的各个属性之间的关系，外延与属性的隶属度应该是外延中的每个对象与这个属性隶属度的最小值。由此，引出如下定义。

定义 8 在定义 6 和 7 的基础上，外延 X 与属性集 A 中各个属性之间的关系为：

$I_X(a_i)=\min I(x,a_i),\forall x\in X$。

我们记 $I_X(a_i)$ 为外延 X 对于属性 a_i 的隶属程度。

根据 Zadeh 表示法，模糊形式概念 $C=(X,B)$ 的内涵可以表示为：

$$B=\sum_{i=1}^{|A|} I_X(a_i)/a_i$$

3 聚类约简

聚类是数据挖掘中研究分类问题的一种多元统计方法，主要用于发现、挖掘不同类别的数据以及识别数据中特定的分布和模式。聚类实质上是一个过程，在这个过程中将给定的大量数据进行划分，分成不同的类（簇）的；通过划分得到的结果的要求是：同一个类（簇）中的对象的相似度较大，而不同类（簇）间的对象的相似度较小。一般情况下，聚类被看作是一个无监督的学习过程，它不需要预先规定数据所属的类别，或者给出训练样例来指明数据具有的关系。聚类分析被应用于很多方面，例如：在生物学上，聚类能用于推导植物和动物的分类，对基因进行分类，获得对种群中固有结构的认识；聚类在地球观测数据库中相似地区的确定，汽车保险单持有者的分组；聚类对 Web 上的文档进行分类，以发现信息。

现阶段，聚类方法主要分为 5 种：划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法。下面对这几种聚类方法进行简单介绍：大部分的划分方法和层次聚类方法是基于距离的；但同时层次聚类方法也基于密度或连通性的；与上两种方式不同的是基于密度的方法与距离无关，不基于各种各样的距离进行聚类操作，而是基于密度的进行的；基于网格的方法与其他任何的聚类方法都可以进行集成。

4 基于模糊概念格的聚类约简

模糊概念格上的聚类约简算法的基本思想是：找到概念格中有两个概念相似时，确定这些概念共同的上确界和下确界，然后将上下确界以及上下确界之间的所有格节点聚在一起，成为一个新的概念。下面我们详细说明这种方法，并给出实例。

设有模糊概念格 $(L,\leq)$，$\forall x,y\in L$，记

$C=\{z \mid \text{glb}\{x,y\}\leq z\leq \text{lub}\{x,y\}\}$。

即 C 为 $\{x,y\}$ 上确界和下确界及其之间的所有概念节点所构成的概念集合，glb{x,y}为下确界，lub{x,y}为上确界，则 $(C,\leq)$ 是 $(L,\leq)$ 的子格。

定义 9 （聚类定义）设有模糊概念格 $(L,\leq)$ 和模糊概念相似度阈值 δ，若 $\exists x,y\in L$，且 x 和 y 的相似度满足 $E(x,y)\geq\delta$，令

$A=\{z \mid \text{glb}\{x,y\}\leq z\leq \text{lub}\{x,y\}\}$。

则将 A 中所有概念聚在一起形成一个聚类概念 A'。

显然，集合 A 是 L 的子集，A 和 L 上的偏序关系 $\leq$ 构成的集合 $(A,\leq)$ 是 $(L,\leq)$ 的子格，我们把 $(A,\leq)$ 叫做聚类子格。

注：$E(*,*)$ 是相似度函数，$0\leq E(*,*)\leq 1$，

例 $E(x, y)$ 是指模糊概念格中 x 和 y 的相似度。本文不给出相似度函数的具体表达式，对于任何相似度函数计算得出相似的两个相似点，进行聚类约简的方式本文完全使用，而在不同的形式背景下所使用的相似度函数是不同的。

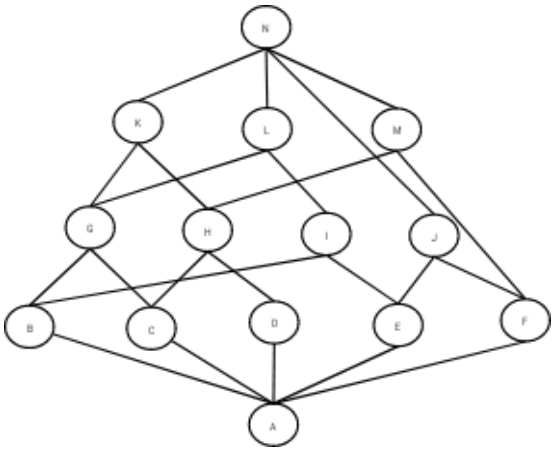

图 4.1 聚类前概念格

以图 4.1 所示的初始概念格为例，若 $E(B, D) \geq \delta$，则聚类步骤如下：(本节主要研究对模糊概念格的聚类约简[4]，在此省略对模糊形式概念的外延和内涵的描述，仅以符号表示节点。)

(1) 求 B 和 D 的上下确界：$\text{glb}\{B, D\} = A$，并且 $\text{lub}\{B, D\} = K$。

(2) 求 A 和 K 之间的所有元素构成的集合：$A_1 = \{A, B, C, D, G, H, K\}$，$A_1$ 是一个聚类子格。

(3) 将 A_1 中的所有元素合并为一个聚类概念 A_1'。

同理，若 $E(F, M) \geq \delta$，则聚类步骤如下：

(1) 求 F 和 M 的上下确界：$\text{glb}\{F, M\} = F$，并且 $\text{lub}\{F, M\} = M$。

(2) 求 F 和 M 之间的所有元素构成的集合：$A_2 = \{F, M\}$，A_2 是一个聚类子格。

(3) 将 A_2 中的所有元素合并为一个聚类概念 A_2'。

根据聚类定义聚类之后的概念聚类集如图 4.2 所示。

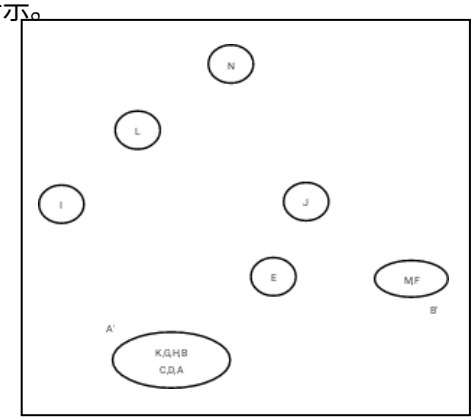

图 4.2 概念聚类集

模糊概念格是一个具有概念层次结构的完备格，但是在聚类的过程中，消除了聚类子格和其他格节点之间的层次关系，需要通过层次化方法使其恢复原有的格结构的层次关系，使其与原有的概念格之间是同构的，并且依然具有概念格的优秀的性质。

定义 10 (层次化定义)[5]设集合 L 聚类后的集合为 L'。其中 A 是聚类子格，$A \subseteq L$；A 聚类后的概念节点为 A'，$A' \in L'$。若 $\exists x \in (L - A)$，$\exists y \in A$，满足 $x \leq y$ 或 $y \leq x$，则聚类后将 x 与 A' 连接。

连接的时候会有一些冗余连接，消除冗余连接的方法见如下定义。

定义 11 (去冗余连接)[6]设模糊概念格 $(L, \leq)$ 聚类后的概念集合为 L'，记层次化后的层次关系为 $\leq'$，消除层次结构 $(L', \leq')$ 中的冗余连接方法为：如果存在概念节点 C' 既是概念 C 的子节点，又是概念 C^* 的子节点，而概念 C 是 C^* 的子节点，则删除从概念节点 C' 到概念节点 C^* 的父子连接关系。

图 4.2 所示的概念聚类集，经定义 11 进行层

次化后，产生的新的层次结构如图 4.3 所示。

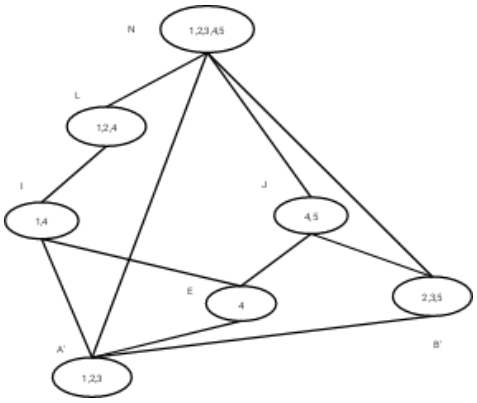

图 4.3 层次化的聚类概念集

图 4.3 所示的层次化概念集中，以聚类概念 B’为例，即 J,N 是 B’的祖先，则删除 B’与 N 之间的连接是冗余连接，将其删除。图 4.3 所示的层次化概念集去冗余连接之后，得到的层次结构如图 4.4 所示。

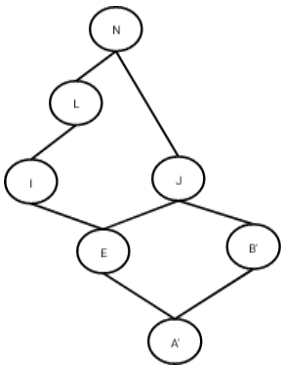

图 4.4 去冗余连接后的层次结构

这样，一个模糊概念格，经上述聚类方法，层次化方法和去冗余之后得到一个约简的层次结构，这一层次结构是一个格，且前后两个格之间是保序的。

结束语

本文就该文所要研究的形式概念分析基础理论的相关概念做了详细的介绍，并且说明了模糊形式概念分析和模糊概念格的相关理论。

此外通过对前人对于聚类约简的理论的运用和方法的调查以及本课题的研究，为概念约简的具体方法进行了说明。通过对概念中对象和属性深入探讨，对约简进行了说明。

参 考 文 献

[1] Ganter B, Stumme G, Wille R. Formal Concept Analysis: Theory and applications - J. UCS Special Issue. Journal of Universal Computer Science, 2004, 10(8): 926-926.

[2] David L D.High Dimensional Data Analysis:The Curses and Blessings of Dimensionality[R].Los Angeles, Mathematics Society Conference: Math Challenges of 21st Century, 2000.

[3] Belohlavek R. Similarity relations in concept lattices. Journal of Logic and Computation, 2000, 10: 823-845.

[4] 黄倩倩. 模糊概念格的聚类约简方法（硕士学位论文）. 大连：大连海事大学，2012.

[5] 李立峰, 王国俊. 一种求概念格属性约简的方法. 计算机工程与应用, 2006, (20): 147-149, 216.

[6] 胡春玉. 模糊形式背景下的概念格属性约简(硕士学位论文). 石家庄: 河北师范大学, 2006.