《智能信息处理》课程作业

基于属性聚类的形式概念分析中的概念格简化

田应彪

作业	分数[20]
得分	

2020 年 11 月 13 日

基于属性聚类的形式概念分析中的概念格简化

田应彪

（大连海事大学 信息科学与技术学院 大连 116026）

摘 要 在形式概念分析(FCA)中，概念格图形化地描述了信息系统的对象和属性之间的潜在关系。概念格的关键复杂性问题之一在于提取有用信息。在庞大的上下文中，属性的无组织性通常不会在 FCA 中产生信息格。此外，在更大的多值上下文中理解属性和对象之间的集体关系更加复杂。在本文中，介绍了一种新的方法来推导一个更小的和有意义的概念格，从中可以推断出概念的摘录。在现有的基于属性的概念格约简方法中，大多数情况下要么减少属性大小，要么减少上下文大小。本文使用的方法包括使用属性的结构相似性和差异性将属性组织成簇，这通常被称为属性聚类，以产生派生的上下文。

关键词 形式概念分析,概念格, 属性聚类

Simplification of concept lattice in formal concept analysis based on attribute clustering

Yingbiao Tian

(Department of Information Science Technology, Dalian Maritime University, Dalian 116026)

Abstract: In formal concept analysis (FCA), concept lattices graphically describe the potential relationships between objects and attributes of an information system. One of the key complexity problems of concept lattices is to extract useful information. In large contexts, the disorganization of attributes usually does not produce information lattices in the FCA. In addition, understanding the collective relationship between attributes and objects in a larger multi-valued context is more complex. In this paper, a new approach is introduced to derive a smaller and meaningful concept lattice from which abstractions of concepts can be infered. In existing attribute-based conceptual lattice reduction methods, in most cases either the attribute size is reduced or the context size is reduced. The approach used in this article includes the use of structural similarities and differences of attributes to group attributes, often referred to as attribute clustering, to generate derived contexts.

Key words Formal concept analysis, concept lattice, attribute clustering

1 引 言

形式概念分析(FCA)已成为数据分析和知识处理的重要工具。它创建了一个概念的层次顺序，每个概念由两个项组成——程

度和意图。集合的这一水平射序构成一个偏序集(部分有序集)，可以用格图表示。FCA侧重于从任何信息系统获得两种形式的产出。第一个是概念格，它是由某些对象和属性簇组成的偏序集。第二个是一组称为属性含义的公式，它描述信息系统中存在的属性

FCA 也有一些不足之处。FCA 经常在巨大的背景下产生大量的形式概念,这在 2006 年的 ICFCA 上被首次指出是一个开放的问题。此外，属性的独立处理产生更大更复杂的概念格。并且用户很难从这样的格子中探索真正相关的方面。因此，概念格形成后，首要任务是确定一个最小概念格，它可以避免冗余，同时保持结构一致性。为此，我们的目标是根据属性的结构相似性对其进行分组，然后修改简化形式的上下文。由于属性本质上是相互关联的，很明显，属性的分组(聚类)在知识提取过程中可以发挥有益的作用。

聚类是一种常见的方法，通过这种方法，数据集被划分成相似项目的组。聚类分析已经被用作在知识挖掘的各个领域提取数据的有效工具。它是一种无监督的模式分组，从观察、数据收集和属性到集合(组)中获得。通常，这种类型的聚类是在对象组上执行的。然而，为了达到我们的目的，我们根据上下文的属性进行聚类。标准的聚类方法将聚类限制为互斥和穷举，这意味着集合中的每一项都包含在一个聚类中。研究人员也在 FCA 环境中有效地处理了聚类技术。Elloumi 等人(2004)在模糊环境下使用模糊聚类分析中的聚类技术，使用关联规则来减少上下文大小，使得结果上下文保留关联规则。Kumar 和 Srinivas 使用模糊 k-均值聚类技术来减小概念格的大小。Martin 等人(2013)提出了聚类方法来度量相同上下文的不同模糊环境下模糊概念格的变化。

2 概念

2.1 概念和概念层次的基本概念

以表格的形式收集数据是研究FCA的第一步，数据表被称为(正式)上下文。十字和空格表示对象和属性之间存在或不存在关系。设 g 是对象的集合，m 是属性的集合，依赖关系。FCA 已经建立，在任何知识发现系统中都有广泛的应用。其应用领域包括人工智能、决策系统、基因表达数据分析、信息检索、本体设计、故障诊断、软件代码分析、专家系统等。

$I \subseteq G \times M$ 是 g 和 m 之间的关联关系。然后，三元组(g, m, I)被称为上下文，由 k 表示。拥有属性 m ∈ M 的对象 $g \in G$ 由$(g, m) \in I$ 或 gIm 给出，称为对象 g 具有属性 m。对于 $A \subseteq G$ 和 $B \subseteq M$，我们定义:

$$A' = \{m \in M \mid g\, I\, m, \forall g \in A\}$$

$$B' = \{g \in G \mid g\, I\, m, \forall m \in B\}$$

如果 A ⊆ G 和 B ⊆ M，这样对于上下文(g, m, I)，$A' = B$ 和$B' = A$。那么这一对(A，B)就被称为概念，而 A，B 被分别说成是概念(A，B)的范围和意图。因此，意图和范围是一个概念的身份。属于概念的所有对象的集合构成了范围；而意图构成了由对象共享的所有属性的集合。

2.2 多值上下文和概念扩展

本质上价值众多的特征或属性，如重量、大小、分数、性别等。区分现实生活中的几个对象，这种属性被称为多值属性。FCA 中的多值上下文表示方案可以有效地处理具有一般属性的上下文。与单值上下文不同，概念格不能直接从 MV 上下文中导出。因此，使用将 MV 上下文转换为单值上下文转换过程被称为概念缩放，这是用户指定的转换过程.

定义 1a(连接) 设$(L, \leq)$是一个偏序集，设 S 是它的子集$(S \subseteq L)$。S 的上界是元素 x ∈ S，使得对于所有 $s \in S$ 有 $s \leq x$。对偶地，S 的下界是元素 y ∈ S，使得对于所有 s ∈ S，有 $y \leq s$。在 S 的所有上界的集合中的最小元素被称为S的上确界或最小上界，并且被对偶地表示为∨S。 S 下界中最大的元素称为 S 的下确界或最大下界，用∧S 表示。如果 $S = \{x, y\}$，我们简单地用 $x \vee y$ 代替∨S，用 x ∧ y 代替∧S。上确界和下确界也分别称为连接和相遇。

定义 1b(格) 如果 l 中存在$\forall a,b \in L, a \in b$ 和 $a \wedge b$，则称偏序集(l，≤)为格。换句

话说，偏序集 l 的任意两个元素都存在并满足运算。如果 l 的每个子集都有上确界和上确界，则称 l 为完全格。

定义 2(语境) 三元组 $K = (G, m, I)$称为形式上下文，如果 g 和 m 分别是非空的对象和属性集，并且 $I \subseteq G \times M$ 是 g 和 m 之间的关联(二元)关系

定义 3(多值上下文) 多值(MV)上下文 (G, M, W, I)由对象 G，属性 M，属性值 W 的集合以及 G 和 M,W 之间的三元关系/组成。换句话说，$I \subseteq G \times M \times W$，其中 $(g, m, W) \in I$ 和 $(g, m, v) \in I$ 意味着 w = v。通过 $(g, m, w) \in I$，我们的意思是‘对于对象 g，属性 m 拥有值 w’。如果 W 包含 n 个元素，那么四元组 $(G, M, W) \in I$ 称为 n 值上下文。每个 MV 属性都是 $m: G \to W$ 这样 $m(g) = w$ 的部分映射。

定义 4(概念) 假设 (g, m, I)是一个形式背景，那么对于任何一个 $A \subseteq G$ 和 $B \subseteq M$ 来说，如果 $a \uparrow = b$ 和 $b \uparrow = a$，则这一对 (a, b)被称为形式概念。集合 a 和 b 分别被称为范围和意图。a'和 b'是概念形成算子。上下文 (G, M, I)的所有概念(A，B)的集合形成一个完整的格，用 $B(G, M, I)$或 B(K)表示。

定义 5(概念层次) 设(G，M，I)为所有概念的集合。对于任意两个概念 $(A1, B1)$，$(A2, B2) \in B$，子超概念关系‘≤’定义如下：$(A1, B1) \leq (A2, B2) \Leftrightarrow a1 \subseteq a2$ *或* $B1 \supseteq B2$)。较低的概念被称为次概念，而较高的概念被称为超概念。

定义 6(格同态) 设 L2 、L1 是两个格。如果‘f’保持连接并满足运算，即$\forall\ a,b \in L\ f(a \wedge b) = f(a) \wedge f(b)$和 $f(a \vee b) = f(a) \vee f(b)$，则映射 $f: L1 \to L2$ 称为格同态。

定义 7(概念包含) 对于来自不同概念格的任何两个概念 l1、l2，如果 l1 的范围和意图分别包含在 l2 的范围和意图中，我们说 l1 包含在 l2 中，并且我们表示 $l1 \subseteq l2$。

定义 8(集群语境) 设 m 是所有属性的集合，设 π(m)表示集合 m 的所有分区的集合。对于任何分区 $n \in \pi(m)$，设 $B(G, n, j)$表示对应于上下文 (g, n, j)的概念格，其中 $J \subseteq G \times N$，使得对于任何对象 $g \in G$ 和 $n \in n$，$(g, n \in) j$ 当且仅当存在属性 $m \in m$，也可以表示为 gJN'。我们称之为语境 (G, N, J)，这样就形成了一个集群语境。

2.3 FCA 中简化技术的背景

自然界中可用的巨大背景通常会产生复杂的概念格，在不丢失相关信息的情况下，很难理解概念格的大小和潜在关系。因为概念格的关键问题通常被认为是其在大小、结构层次、底层信息等方面的复杂性。目前已经提出了几种具有不同特征的概念格约简方法。

在其中一些方法中，冗余信息从概念格中移除。一般来说，约简方法侧重于找到能够保持原始格结构不变的相关对象或属性集。Ganter 和 Wille 通过移除可约对象和属性获得了澄清的上下文，并且所得到的概念格保持了与原始概念格的同构。面向粗糙集的方法将模糊聚类分析扩展到基于决策的上下文，从任何信息系统中分解冗余知识。Medina(2012)研究了形式概念格、面向对象概念格和面向属性概念格三种框架中的属性约简。不考虑框架，已经发现属性可以被分为三个层次的必要性，并且在任何层次的属性约简都是相同的。另一类约简方法通过适用的原则或标准选择形式概念、对象或属性来工作。一些学者提出了一种方法，该方法使用某些约束来减少概念的数量，这些约束是从属性依赖公式(ADFs)导出的，这些公式是与形式上下文一起额外输入的。与给定的一组模数转换器兼容的一组概念被简化为重要的概念。

2.4 FCA 中简化技术的背景

属性聚类是一种对相互关联的属性进行分组的方法。一个集群内的属性之间的相关性更高，而不同集群的属性之间的相关性较低。使用属性聚类可以最小化数据挖掘算法的搜索维度。每个簇由一个唯一的质心组成，它具有簇内属性的更多共同性质。研究人员大多预先定义所需的集群数量。Jain，Murty 和 Flynn (1999)已经介绍了使用统计模式识别技术的聚类技术的概述。根据它们，

聚类过程的基本步骤是:(1)数据表示，(2)相似性度量，(3)分组，(4)聚类表示。此外，他们还确定了一些领域，如信息检索和字符识别、图像分割和数据挖掘，作为聚类技术的适用领域。聚类技术可以用于对象或属性集或形式概念。聚类技术已经在具有不完整上下文的 FCA 环境中采用。

现今有几种可用的聚类技术。其中 k-means，hierarchical，DB Scan，OPTICS 和 STING 是一些流行的聚类技术。在几乎所有的聚类技术中都明确或隐含地使用了邻近性度量。由于比较由上述技术产生的聚类的质量在 FCA 的研究中是一项困难的任务，所以有时这些聚类技术可能是不合适的。此外，开发一种合适的比较方法来确定集群的质量仍然是一项艰巨的任务，因为集群没有展示对象/属性之间的全部关系。为了避免这一缺点，我们提出了基于众所周知的相似性度量概念的属性聚类。此外，这种聚类技术不需要任何验证，因为验证度量本身就是相似性度量。在数据挖掘过程中，数据的接近度通常使用相似性/距离相关的度量来度量。与此相一致，Jaccard 相似性系数和距离度量被广泛使用在二进制、分类、序数或混合属性的情况下，欧几里德或相关距离度量不能用于度量属性之间的距离。因此，距离度量被属性/对象之间的相异度的概念修正(韩和 Kamber 2006)。通常，距离测量是在对象集上执行的，而聚类技术只在一组对象。但是，为了实现我们的目标，我们在属性集上使用这些度量。

表一 二进制数据的示例列联表

	Attribute j		
Attribute i	1	0	Sum
1	q	r	q+r
0	s	t	S+t
	q+s	r+t	p

3 方法

我们提出的方法组织如下:在初始化期间，我们首先获取原始信息系统。在数字数据的情况下，它可以被转换成一个分类(多值)上下文。然后我们确定它的形式概念，从而确定它的概念格。我们使用聚类方法，

在二元属性的情况下，任何两个属性之间的距离可以通过使用列联表来容易地计算。如果一个二进制属性的两个状态同等重要，则该属性是对称的；否则，它是非对称二进制属性。在二进制属性的情况下，距离度量根据它们的对称性而变化。二进制状态 1 和 0 分别表示相应属性的存在或不存在。此外，列联表单元格中给出的基数表示具有所述属性的对象的数量。对象或属性之间的相似性最常用的度量是雅克卡指数或雅克卡相似性系数。二进制数据的示例列联表如表 1 所示。使用表 1 所示的列联表，两个属性/对象之间的 Jaccard 索引可以简单表述为:

$$J(i,j) = \frac{q}{q+r+s} \qquad (1)$$

以类似的方式，两个属性/对象之间的距离使用以下公式进行测量:

$$d(i,j) = \frac{r+s}{q+r+s+t} \qquad (2)$$

显然，距离度量本质上是度量，这两个度量的范围介于 0 和 1 之间。($即 0 \le J(i,\ j),\ d(i,\ j) \le 1$)

即属性聚类过程中的Jaccard指数和距离测度，分别用于确定聚类的质心和对属性进行分组。在构造属性簇之后，我们形成簇上下文(G，N，J)，其中 N∈ π (M)和 π (M)分别表示集合 M 的所有分区的集合，从最初的那个导出。在这个集群上下文中，对象保持不变，而属性是新形成的集群。在群集环境的形成中，我们采用联合原则来确定对象的群集特征的存在与否。换句话说，如果一个对象至少拥有一个集群的属性，那么它就被认为存在于一个特定的集群中。提出的方法的整个过程可以根据以下步骤系统地进行:

第一步:初始化:

(a)输入多值上下文

(b)转换形式上下文

(c)获取形式概念

(d)获取概念格

第二步:属性聚类

a.质心选择

i、计算给定属性的雅克卡相似系数矩阵。

ii、预定义簇的数量，并相应地从具有较高平均雅克卡系数的属性中选择相等数量的质心。

b.属性聚类

i、计算给定属性的距离矩阵。

ii、逐行选择每个非质心属性，并将其与距离矩阵中值最低的质心分组。

第三步:聚类上下文:形成聚类上下文(G，N，J)，其中N∈π(M)和π(M)表示集合M的所有分区的集合。

第四步:概念格:获得格B(G，N，J)。

第五步:映射:使用概念包含图zeta ζ将原始格的概念映射到简化格。

第六步:输出评估:使用概念比较度量，分析从结果格中提取的概念与原始格中的概念。

参考文献

[1]Sumangali K , Kumar C A . Concept Lattice Simplification in Formal Concept Analysis Using Attribute Clustering[J]. Journal of ambient intelligence and humanized computing, 2019, 10(6):2327-2343.

[2]李金海,魏玲,张卓,翟岩慧,张涛,智慧来,米允龙.概念格理论与方法及其研究展望[J].模式识别与人工智能,2020,33(07):619-642.

[3]王霞,谭斯文,李俊余,吴伟志.基于条件属性蕴含的概念格构造及简化[J].南京大学学报(自然科学),2019,55(04):553-563.