《智能信息处理》课程作业

# 形式概念分析研究和数据挖掘应用

刘海鹏

| 作业 | 分数[20] |
| --- | --- |
| 得分 | |

2020 年 11 月 11 日

# 形式概念分析研究和数据挖掘应用

刘海鹏

(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：近些年来，信息的数量不断扩大，我们收集信息的能力已经提高了很多。因特网的流行给我们带来了很多有用的数据信息，同时也给我们带来了数据巨大而造成的麻烦，我们需要将这些数据转换成我们需要的信息以及知识。数据库知识发现是指从非常大的数据量中提取有效的，新颖的，潜在有用的，可信的，并且可以被人理解的模型。概念格是近年来数据分析快速发展的有力工具，用来发现数据中隐藏的知识模式。因此，研究概念格的基本理论并将其应用于知识发现具有重要意义。

**关键词** 形式概念分析；数据挖掘；知识发现；概念格

**中图法分类号** TP311.20 **DOI 号** 10.3969/j.issn.1001-3695.2014.01.030

## Research on formal concept analysis and Data mining applications

Liu Haipeng

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

**Abstract** ： *In recent years, the amount of information has been expanding, and our ability to collect information has improved a lot. The popularity of the Internet has brought us a lot of useful data information, but also brought us huge data and caused trouble, we need to convert these data to the information we need and knowledge. Database knowledge discovery refers to the extraction of effective, novel, potentially useful, trustworthy, and understandable models from very large amounts of data. Concept lattice is a powerful tool for the rapid development of data analysis in recent years, which is used to discover the hidden knowledge patterns in data. Therefore, it is of great significance to study the basic theory of concept lattice and apply it to knowledge discovery.

**Key words:** Formal concept analysis,Data mining, Concept lattice.

## 1 引言

需求是科技发展的原始动力。八十年代,人工智能投资研究性项目遭受挫折,科研转入实际应用时提出了知识发现(KDD)。它是一个新兴的,面向商业应用的研究。一份最近的 GartenrGroup 报告列举了五项在今后 3 一 5 年内对工业将产生重要影响的关键技术,其中 KDD 和人工智能排名第一,并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位。Gartner 的 HPC 研究表明:“随着数据捕获、传输和存储技术的快速发展,大型系统用户将更多的需要采用新技术来挖掘市场以外的价值,采用更为广阔的并行处理系统来创建新的商业增长点”。可以看出 KDD 完全是与应用紧密相关的、快速发展的新兴学科。

Friedman 在 1997 提出主要有三个基础技术激发了知识发现的开发、应用和研究的兴趣,它们是:一、海量的数据搜集。二、高性能多处理器计算机。三、数据挖掘算法。目前,商业数据仓库和计算机自动收集的数据记录导致超大规模数据库的出现、高性能计算和并行体系结构的发展、对巨量数据的快速访问以及对此类数据应用精深的统计方法计算能力的提高等原因,说明以上三种基础技术已发展成熟,因此知识发现开始走向商业应用。

KDD 是一个广阔的研究领域,是多种方法的集成,包括统计学、数据库技术、人工智能、机器学习、模式识别、机器发现、不确定性建模、数据可视化、高性能计算、最优化、管理信息系统和基于

知识的系统等。它是一个含有多个步骤的高度交互和迭代的过程。而数据挖掘算法是用来从数据中发现模式、聚类和模型,这些模式最终将被归纳为用户可理解和可视化的形式。数据挖掘是 KDD 全过程的一个非常重要和关键的步骤,是 KDD 目前的研究热点。

## 2 相关概念与技术

定义 1.1:概念 C 的广义子概念 desc(C)定义如下:

(1)C 的直接子概念是 C 的广义子概念，即 C1<C=> C1∈desc(C)。

(2)C 的直接子概念的广义子概念包含于 C 的广义子概念中，即(C1<C)∧(C2 desc(C1))=>{C2∈dese(C)}。

假设分类算法的输入向量集合，即训练集 S={vi | vi=vi1,vi2…, ci}，其中 ci 是类别标识。不考虑分类属性 C，训练集对应于形式背景 T=(O, D, R)，S 中的每个向量 vi 就是 T 中的一个对象，(vi1, vi2…)是其对应的属性值。

定义 1.2:如果在 S 中有 vi=vk(i≠k)，称 vi 和 vk 是不可分向量;如果存在一个属性 a，对于 S 中的任意向量 vi∈S，vk∈S，有 ai=ak 称属性 a 是分类无关属性。简单的说，分类无关属性 a 在 S 的所有向量上的取值都相等，或者说在数据库中，存在某一列的值全部相同。

定义 1.3:如果向量集合 S 剔除了不可分向量和分类无关属性，则有其所组成的形式背景定义可为标准分类形式背景。事实上，分类算法在处理训练集时，不可分对象和分类无关属性只会产生冗余的和无效的分类规则，因此从分类的角度，我们可以对形式背景进行纵向和横向的预处理，来生成标准分类形式背景。

定义 1.4:概念 e 覆盖数记为 Cover(C)，定义为集合 S={C}∪dese(C)中对象概念的个数。

## 3 量化的相对约简格的分类规则发现

基于 Galois 概念格的分类系统取得了良好的分类效果，但是算法的时间和空间的效率有待提高。其主要原因之一是 Galois 格的庞大的数据存储量和构造过程的时间花费，它是制约 Galois 格走向应用来处理大数据量的关键问题之一。

算法 Ruleame 是利用 Galois 格来发现分类规则的，而量化的相对约简格的提出有效的降低了大规模概念格的存储复杂度，使得概念格的结构更加简洁明了，知识表示更为丰富。利用量化的相对约简格，更容易实现分类算法，归纳分类规则。本章我们将基于量化的相对约简格来对 Ruleamer 算法进行改造，从时间上和空间上提高算法的效率。然后提出一个新的基于量化的相对约简格的分类规则发现算法。

为了简便起见，下文的量化的相对约简格均指由标准分类形式背景所构造的量化的相对约简格，我们讨论它的性质:

命题 1:量化的相对约简格的空概念的直接超概念外延基数必定是 1。

证明:假设概念 V 是空概念的直接超概念，并且其外延基数|A|=2,不妨设其外延分别是 o1 和 o2，由于 V 已经是空概念的直接超概念，说明在概念格中不存在某个概念，其外延仅仅是 o1，或者是 o2，因此对象 o1 和 o2 在所有的属性上的取值都是相同的。此结论与标准分类形式背景的定义相矛盾。

命题 2:如果形式背景是标准分类形式背景，在量化的相对约简格中，对于每个概念 U=(|A|, B)必有|A|=Cover(U)，Cover(C)就是 C 的外延基数。

证明:由定义 5.4 知道 Cover(C)表示的是概念 C 及其广义子概念中对象概念的个数。而由命题 1 显然可知，在标准分类形式背景中，所有对象概念的外延只包含一个对象。那么如果概念 U 的外延基数是|A|，说明其外延中有|A|个不同的对象，而每个对象都会形成一个对象概念。所以在其本身和广义子概念集合中中必定有|A|个对象概念，即|A|=Cover(U)。

算法在构造好的相对约简格上发现分类规则，首先按照以下步骤来初始化概念格节点:

1)把格中的所有节点都初始化为 active，表示每个节点都是生成分类规则

的候选节点;

2)对应形式背景，将每个对象概念分别标记上它的分类标识;

3)如果一个节点的广义子概念中的对象概念的分类标识都是自则，将此概念节点的分类标识标为 ci 否则，将此概念的分类标识标记为一个特殊的分类标识 mixed。

命题 3:量化的相对约简格中，如果某个概念 U=(|A|, B)的分类标识是 ci，则对于 V=(|Asup|, Bdesc) ∈dese(U)必有 ci 也是 V 的分类标识;如果

U 的分类标识是 mixed,则对于 V=(IAsup▌,Bsup) ∈ sup(U)，mixed 也是 V 的分类标识;显然，如果一个概念存在分类标识是 ci，那么它的广义子概念的分类标识也是 ci;如果一个概念没有分类标识，那么它的广义超概念也没有分类标识。

算法的输入:由标准分类形式背景构造的量化的相对约简格和类别集合 C;

算法的输出:谓词描述的分类规则;

```
GenClassRules(L,C)
1 初始化所有格节点
2 while L 中还有节点是 aetive do
3 {
    u <- Cover(w)最大且分类标识不是 mixed 的节点 w
    MakeRulesFromNode(L, u)
6   foreach v∈{u}∪dese(C)且 v 是对象概念 do
7       { foreach w ∈{v}∪super(v) do
8           { cover(w)<-cover(w)-1
9           if cover(w)<=0 then 标记 w 为 inactive }
10      标记 v 为 inactive
11}
MakeRulesFromNod(L,u)
1 anter <- {u}∪super(u)中的属性节点的内涵并集
2 label <-的分类标识
3 生成规则 anter=>label
```

算法 1.1

以上算法与 Ruleamer 相比较有以下的特性和优点:

1)对于量化的相对约简格中的概念 U=(A，B=(v1，v2…vk))，除了全概念和空概念，如果 B≠∅，则这个概念是属性概念，并且如果|B|二 k，表示这个属性概念可以由属性 v1，v2，…vk 分别确定。无需花费时间专门判断某个概念是否是属性概念以及一个概念是有哪些属性概念共同确定的。

2)由命题 1 知道，如果某个概念的外延基数是 1，那么它必定是对象概念;除了全概念和空概念，如果 B≠∅表明这个概念是由其它概念相交生成的内部概念。因此无需花费时间专门判断某个概念是否是对象概念。

3)由命题 2 可知，在相对约简格中，Cover(v)就是概念 v 的外延基数，我们不需要单独计算每个概念的覆盖数。

4)量化的相对约简格，以外延基数来表示概念外延，以相对约简内涵表示概念内涵，较低了概念格的空间复杂度，提高了算法的健壮性。

分析 Ruleame 算法可以知道，主算法的时间复杂度大约是 O(|L|3)内，其中|L|是 Galois 格 L 中的概念个数，但是在初始化阶段，求解每个概念 v 的 Cover(v)和标识每个概念的类别时，它们的时间复杂度却分别为 O（|L|2)。综上所述，基于量化的相对约简格的分类规则发现算法，可以从时间和空间上提高算法效率，来解决大数据量问题。

然而，算法 1.1 和 Ruleamer 算法所产生的分类规则是冗余的。通过仔细分析算法 5.1 我们发现，在 GenClassRules 中的第六至第十步，分别来计算 Cover 值、标记 inactive 是没有必要的,分析原因:

假设存在一个分类标识 ci，在概念格中，我们所关心的是所有分类标识为 ci 的节点，以及虽然分类标识不是 ci 但对分类标识为 ci 的节点有影响的节点。

定义 1.5:如果节点 v 的分类标识是 ci，并且其所有的广义超概念的分类标识都是 mixed，我们称 v 是分类标识 ci 的关节点。

考虑分类标识 c;的关节点 v，其上的广义超概念虽然分类标识不是 ci，但是对分类有影响，其下的广义子概念的分类标识必定都是 ci。因此以这个节点为分界线，它的广义超概念和广义子概念对于分类都是有影响的节点。算法只要寻找到每个分类标识的关节点，由此关节点就可以获得分类属性。

命题 4:对于任意两个分类标识 ci，cj∈C，ci≠cj 设它们的关节点分别是 vi 和 vj，那么有 dese(vi) ∩desc(vj)= ∅。

证明:两个不同的分类标号 ci 和 cj 的关节点的广义子概念是不相交的，假如相交，那么表示有某个对象的分类标识既为 ci 又为 cj，这是不可能的。

命题 5:对于任意一个相对约简概念 U=(A，B=(v1，v2，…vk))，如果它的分类标识是 ci，必有分类规则 vi=>ci 其中 vi ∈B。它是一个前件为一个属性的最简分类规则。

基于上面的命题和讨论，我们提出了新的基于量化的相对约简格的分类规则发现算法:

```
NewGenClassRules(L, C)
1 为每个概念节点注明分类标识
2  foreaeh  c∈C do
3   u<-分类标识为 c 的节点中 Cover(v)值最
```

大的 v

4 MakeRulesFromNode(L，u，c)

MakeRulesFromNode(L，u，c)

1 Bdesc<-{u}∪dese(u)的所有内涵的并集

2 生成规则 b=>e，其中 b∈Bdesc

3 if u 有两个以上的直接超概念 then

4 Bsuper 介 super(u)中的属性概念内涵的并集的集合

5 生成规则 B=>e，其中 B∈Bsuper

算法 NewGenelassRules 的时间复杂度。O(|C|)，C 是分类标识的集合，其中在第一步初始化时候的时间复杂度是 O(|L|2)，算法可以在初始化的时候，记住分类标识 c 的最大覆盖值 Cover(v)的位置来提高性能;由节点生成分类规则的算法 MakeRulesFromNode 和算法 5.1 中的相同。比较算法 5.1 和算法 5.2，后者在时间性能上有较大提高，并且可以生成前件为一个属性的最简分类规则。

## 4 小结

本章利用量化的相对约简格的优良性能，对 Rulearer 算法进行了精心的改进，同时利用相对约简格提出一种新的分类规则发现算法。基于量化的相对约简格对 Rulearer 算法进行改造后，无需在算法中去专门判断一个概念是否是属性概念，无需专门判断某个概念是否是对象概念，无需专门去计算一个概念的覆盖数，因此在时间复杂度上较 Rulearer 算法有很大的提高，同时由于概念格是量化的和相对约简的，从外延和内涵的角度较低了存储量，在空间复杂度上也有很大提高。关节点是量化的相对约简格中与某个分类标识有密切联系的一个概念，关节点的广义超概念的类别标识是不明确的，关节点本身和关节点的广义子概念的类别标识都是相同的，基于我们提出的关节点的定义，在量化的相对约简格上可以实现新的分类规则发现算法，算法可以分别由关节点的广义超概念、关节点和其广义子概念来生成分类规则，此算法可以生成前件为一个属性的最简分类规则。我们以随机产生的标准分类形式背景作为输入实现了此分类算法，如上所述，与 Rulearer 算法相比较在时间性能、空间性能、发现规则的简化程度上都有明显的改进和提高。最后我们比较了基于概念格的分类和决策树分类方面的异同。

## 参 考 文 献

[1]徐伟华,杨蕾,张晓燕.模糊三支形式概念分析与概念认知学习[J].西北大学学报(自然科学版),2020,50(04):516-528.

[2]陈希邦. 基于粒计算与形式概念分析的面向对象（属性）概念的粒描述[D].江西理工大学,2020.

[3]刘萍,彭小芳.基于形式概念分析的词汇相似度计算[J].数据分析与知识发现,2020,4(05):66-74.

[4]李金海,李玉斐,米允龙,吴伟志.多粒度形式概念分析的介粒度标记方法[J].计算机研究与发展,2020,57(02):447-458.

[5]折延宏,胡梦婷,贺晓丽,曾望林.两种多粒度形式概念分析模型的比较研究[J].计算机工程与应用,2020,56(10):51-55.

[6]曾望林. 基于属性粒化的多尺度形式概念分析研究[D].西安石油大学,2019.

[7]张榕宁.国外网络信息检索研究现状[J]. 图书馆论坛,2006,(8):188—190.

[8]杨青,王瑞菊. 浅析网络住处检索中的问题与对策[J].图书馆学研究,2004,(6):82—83.

[9]李爱红.网络搜索引擎的比较研究[J].中国信息导报.1999(1):25 一 26.

[10]徐亨南.人工智能与智能信息检索.信息检索(江西图书馆学刊),2005,35(l):53 一 54.

[11]金海,袁平鹏.语义网数据管理技术及应用[M].北京：科学出版社,2010.

[12]黄果,周竹荣,周亭.基于语义网的信息检索研究.西南大学学报(自然科学版),2007,29(l):77 一 80.