《智能信息处理》课程作业

基于监督学习算法关联性的形式概念分析

王树声

作业	分数[20]
得分	

2021 年 11 月 28 日

基于监督学习算法关联性的形式概念分析

王树声

(大连海事大学 理学院, 辽宁省大连市 中国 116026)

摘 要 形式概念分析，也被称为概念格，由 *Wille R* 于 1982 年首先提出，它是应用数学的一个新分支，提供了一种支持数据分析的有效工具。概念格通过 *Hasse* 图生动和简洁地体现了这些概念之间的泛化和特化关系，被认为是进行数据分析的有力工具。概念格可以用于许多机器学习的任务，同时监督学习作为机器学习的重要一部分，各种经典的监督学习方法之间也存在着灵魂与思想的相似。本文通过形式概念分析的理论方法，研究监督学习方法之间的关联性。首先简要介绍了形式概念分析的基本概念、过程和基本方法；接着对监督学习方法进行了特征提取，构成形式背景，构建概念格；最后从概念格中提取了相关规则。

关键词 形式概念分析；概念格；*Hasse* 图；监督学习

Formal Concept Analysis Based on Relevance of Supervised Learning Algorithm

Wang Shusheng

(Department of Mathematics, Dalian Maritime University, Dalian, 116026)

Abstract Formal concept analysis, also known as concept lattice, was first proposed by Wille R in 1982. It is a new branch of Applied Mathematics and provides an effective tool to support data analysis. Concept lattice vividly and concisely reflects the generalization and specialization relationship between these concepts through Hasse diagram, which is considered to be a powerful tool for data analysis. Concept lattice can be used in many machine learning tasks. At the same time, supervised learning is an important part of machine learning. There are also similarities between soul and thought among various classical supervised learning algorithms. This paper studies the correlation between supervised learning algorithms through the theoretical method of formal concept analysis. Firstly, the basic concepts, processes and methods of formal concept analysis are briefly introduced. Then, the supervised learning algorithms are comprehensively compared, the features are extracted, the formal background is formed, and the concept lattice is constructed. Finally, the correlation between supervised learning algorithms is concluded.

Key words Formal concept analysis, Concept lattice, Hasse diagram, Supervised learning

1 引 言

概念格，也称为*Cralois*格，又叫做形式概念分析[1]，是由*Wille R*于 1982年提出的，其基本思想是基于对象与属性之间的关系，根据这一关系来建立一种概念层次结构，其中每个概念都是对象与属性的统一体。另外，概念格通过*Hasse*图生动和简洁地体现了这些概念之间的泛化和特化关系。因此，概念格被认为是进行数据分析的有力工具。从数据集中（概念格中称为形式背景）中生成概念格的过程实质上是一种概念聚类过程；然而，概念格可以用于许多机器学习的任务。目前，已经有了一些建造概念格的算法，并且概念格在信息检索、数字图书馆、软件工程和知识发现等方面得到应用。

目前，在自然语言学[2]（计算语言学）中，如果可以通过汉语词法、句法、语义等知识库的学习发现其中隐含的规律，将有助于计算机对现代汉语的理解。而形式概念分析上的规则提取可以有效提取知识库中的隐藏的规则，因此利用形式概念分析对语言知识进行研究是一个很有意义的课题。在知识发现领域，概念格可以从关系数据中构造出来，然后从概念格上可以提取各种类型的知识，如蕴含规则、关联规则、分类规则等等；在软件工程领域，概念格可以从类库的规范说明上构造，从而对类库结构的可视化以及类库的重构和优化提供支持；在知识工程领域，概念格可以用于知识库的重新结构化；在信息检索方面，概念格可以实现对信息的有机组织并过虑掉无用的信息。而且，有人指出概念格将会在生物和生命科学领域有重大应用。

监督学习方法[3]是机器学习中重要一部分。监督学习可以认为是学习一个模型，使它能对给定的输入预测相应的输出。其词汇中包含适用问题、模型类型、学习策略、损失函数等重要信息，本文通过分析监督学习方法的隐含的语义信息来构建其形式背景并建立对应的概念格。

2 形式背景

形式概念分析首先要建立形式背景。形式背景被定义为一个三元组，公式为$F=(T,A,R)$，其中 T 为所有对象集合，A 为所有属性的集合，$R\subseteq T\times A$ 为 T 和 A 中元素之间的二元关系集合[4]。该三元组可以表示为二维表。在下面表 1 所示的形式背景中，关于对象集合 $T=\{t_1,t_2,t_3,t_4\}$，属性集合 $A=\{a_1,a_2,a_3,a_4,a_5\}$，二元关系 R 为确定性关系。实际上，形式背景一般都不是直接存在的，需要从数据源中提取，从而就需要对数据源进行分析，采取不同的策略和算法来提取形式背景。

表 1 形式背景的示例

	a_1	a_2	a_3	a_4	a_5
t_1	1	1	1	1	0
t_2	1	1	1	0	1
t_3	0	1	1	0	0
t_4	0	0	1	1	1

针对常见的监督学习方法，提取相关形式背景，为了便于说明，监督学习方法的关系背景如表 2，其对应关系如下：1 代表感知机，2 代表 k 近邻法，3 代表朴素贝叶斯，4 代表决策树，5 代表逻辑斯谛回归与最大熵模型，6 代表支持向量机，7 代表提升方法，8 代表隐马尔可夫模型，9 代表条件随机场；*a* 是该方法是否适用于分类问题，*b* 是该方法的模型类别是否是判别模型，*c* 是该方法是否使用线性模型，*d* 是该方法是否可以使用概率模型，*e* 是该方法可以使用非概率模型，*f* 是该方法的学习策略是否使用极大似然估计，*g* 是该方法学习的损失函数是否为对数似然函数，*h* 是该方法是否简单。其中数值 1 代表具备该属性，0 代表不具备。

表 2 监督学习方法关系-形式背景

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
1	1	1	1	0	1	0	0	1
2	1	1	0	0	1	0	0	1
3	1	0	0	1	0	1	1	1
4	1	1	1	1	1	1	1	1
5	1	1	0	1	1	1	0	0
6	1	1	0	0	1	0	0	0
7	1	1	0	0	1	0	0	0
8	0	0	0	1	0	1	1	0
9	0	1	1	1	1	1	1	0

3 概念格

建格的过程实际上是概念类聚的过程[1]。因此，在概念格中，建格算法具有很重要的地位对于同一批数据，所生成的格是唯一的，即不受数据或属性排列次序的影响，这也是概念格的优点之一。概念

格的建格算法可以分为两类：批处理算法和增量算法。概念格可以添加背景知识，甚至可以只用背景知识建造。

批处理算法根据其构造格的不同方式，可分为 3 类，即从顶向下算法、自底而上算法、枚举算法。从顶向下算法首先构造格的最上层节点，再逐渐往下。自底而上算法则相反，首先构造底部的节点，再向上扩展。枚举算法则是按照一定顺序枚举格的所有节点，然后再生成 *Hasse* 图，即各节点之间的关系。增量算法和批处理算法不同，增量算法的思想都是大同小异的——基本思想都是将当前要插入的对象和格中所有的概念交，根据交的结果采取不同的行动[5]。主要区别在于连接边的方法。

本文将根据监督学习方法关系的简单形式背景产生对应概念格。首先，根据所给形式背景约减生成单值形式背景，再确定单值形式背景中的父子关系，根据父子继承关系绘制 *Hasse* 图，最后补充各形式概念的上确界和下确界，形成概念格。

3.1 约简形式背景

形式背景的约减包括聚类（行约减）和关联（列约减）。通过表 2 可看出，6 与 7 是一组有相同属性的行，故将其合并；*d* 与 *f* 是一组有相同对象的列，故将其合并。最后得到约减后的形式背景如表 3。

表 3 监督学习方法关系-约减形式背景

	d，*f*	*a*	*b*	*c*	*e*	*g*	*h*
6,7	0	1	1	0	1	0	0
1	0	1	1	1	1	0	1
2	0	1	1	0	1	0	1
3	1	1	0	0	0	1	1
4	1	1	1	1	1	1	1
5	1	1	1	0	1	0	0
8	1	0	0	0	0	1	0
9	1	0	1	1	1	1	0

3.2 生成单值形式背景

单值的形式背景即根据前一步约减后的形式背景，把值为“1”的位置改为“×”，去掉其他位置的“0”以表示该形式对象有此属性。最后得出结果见表 4。

表 4 监督学习方法关系-单值形式背景

	d，*f*	*a*	*b*	*c*	*e*	*g*	*h*
6,7		x	x		x		
1		x	x	x	x		x
2		x	x		x		x
3	x	x				x	x
4	x	x	x	x	x	x	x
5	x	x	x		x		
8	x					x	
9	x		x	x	x	x	

3.3 确定父子关系

父子关系也称基于属性个数的排序。在获取到的单值形式背景的基础上做顺序的调整，找到属性继承的父子关系，例如 3 可由对 8 的全部属性继承的基础上添加自身属性 *a* 得到。通常情况下，为方便查找，从上倒到下按属性的多少进行排列。表 5 所示为基于属性个数的排序。

表 5 监督学习方法关系-基于属性个数的排序

	d，*f*	*a*	*b*	*c*	*e*	*g*	*h*
8	x					x	
6,7		x	x		x		
2		x	x		x		x
3	x	x				x	x
5	x	x	x		x		
1		x	x	x	x		x
9	x		x	x	x	x	
4	x	x	x	x	x	x	x

3.4 绘制 *Hasse* 图

Hasse 图也称哈斯图[6]，在数学分支序理论中，是用来表示有限偏序集的一种数学图表，它是一种图形形式的对偏序集的传递简约。

具体的说，对于偏序集合 $(S;\leq)$，把 S 的每个元素表示为平面上的顶点，并绘制从 x 到 y 向上的线段或弧线，只要 y 覆盖 x（就是说，只要 $x<y$ 并且没有 z 使得 $x<z<y$）。这些弧线可以相互交叉但不能触及任何非其端点的顶点。带有标注的顶点的这种图唯一确定这个集合的偏序。

Hasse 图的作图法为：以“圆”表示元素；若 $x<y$，则 y 在 x 的上层；若 y 覆盖 x，则连线；不可比的元素在同层。应用 *Hasse* 图表示各结点所组成的偏序集及节点间的关系，由上到下表示的即为两节点间的父子关系，根据表 5 所绘 *Hasse* 图如图 1 所示。

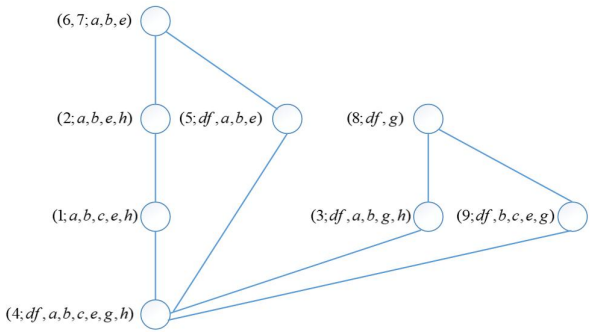

图 1 *Hasse* 图

3.5 生成概念格

针对表 5 的简单形式背景，采用手工方式生成概念格。图 1 已经给出 *Hasse* 图，即已得出概念间的偏序关系，只需补出上下确界即可得到概念格。

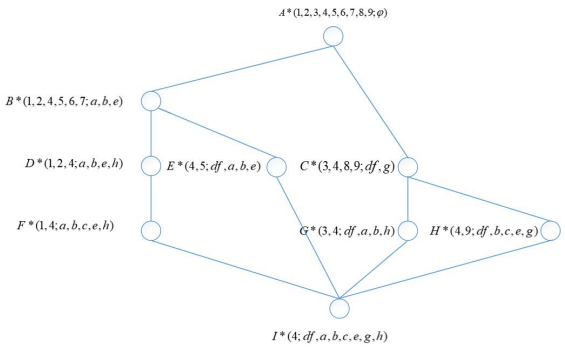

图 2 概念格

4 从概念格中提取关联规则

4.1 关联规则概述

关联规则[7]是很有价值的一类规律，它指的是一个形如 *ATB* 的表达式，其中 *A* 和 *B* 是属性集合，其直观含义是：具有属性 *A* 的行也可能具备属性 *B*。已有不少学者讨论了从概念格上提取规则或函数依赖的问题，比如采用“标记法”从格中提取规则的 *Rulelearner* 算法[8]、使用改进的 *Bordat* 建格算法作为基础的 *CLACF* 集成算法[9]。和其它分类器相比，概念格上提取的规则具有相当或更好的分类效果。

4.2 关联规则的提取方法及相关解释

下面我们引用一种在建好的概念格上生成规则的算法[10]。基本思想为针对格中每个结点来生成无冗余的所有规则，主要的依据是其双亲结点。

定理 1 如果格中结点 $H=(X;X')$ 只有一个双亲结点 $M=(Y;Y')$，则 H 所产生的规则前件只能为单个描述符，且存在 $p\in\{X',-Y'\}$，都有一条无冗余规则 $p\to X'-p$。

定理 2 如果格中结点 $H=(X;X')$ 具有 n 个双亲结点 $M_1(Y_1;Y_1')$，$M_2(Y_2;Y_2')$，…，$M_n(Y_n;Y_n')$，则对于任意一个描述符 $p\in\{X'-(Y_1'\cup Y_2'\cup\cdots\cup Y_n'\}$，都存在一条规则 $p\to X'-p$。

定理 3 如果格中结点 $H=(X;X')$ 具有两个双亲结点 $M_1(Y_1;Y_1')$ 和 $M_2(Y_2;Y_2')$，则 $p_1\in\{Y_1'-Y_1'\cap Y_2'\}$ 和 $p_2\in\{Y_2'-Y_1'\cap Y_2'\}$，都存在一条规则 $p_1p_2\to X'-p_1p_2$，并且前件为两个描述符的规则总数是 $\|Y_1'-Y_1'\cap Y_2'\|*\|Y_2'-Y_1'\cap Y_2'\|$。注意到只有当 $\|Y\|>k$ 时，才可能有前件至多为 k 个描述符的规则。表 6 为根据上述定理所推到出的最简的全局规则示例。

表 6 最简全局规则示例

规则	所属节点范围	相关解释
a,b,e→h	*B*,D**	简单的非概率判别分类模型
a,b,e→df	*B*,E**	学习方法为极大似然估计，同时可以使用概率与非概率的判别分类模型
a,b,e→c,h	*B*,D*,F**	简单的非概率线性判别分类模型
df,g→a,b,h	*C*,G**	损失函数为对数似然函数，学习方法为极大似然估计，简单的概率判别分类模型
df,g→b,c,e	*C*,H**	损失函数为对数似然函数，学习方法为极大似然估计，同时可以使用概率与非概率的判别模型

5 结 论

概率格通过规则提取，重复挖掘数据之间的数学特征，在寻找监督学习分类方法的过程中建造与应用概念层次结构进行方法选取具有很多优势，而概念格的 *Hasse* 图正好体现了一种概念层次结构，反映了学习分类方法之间的共有和私有属性。本文基于若干监督学习方法给出该领域的形式背景，根据监督学习方法与其属性之间的关系，构造决策背景和概念格，进一步刻画了这几种方法属性的关联规则，解释了关联规则的内涵。该工作使我们在进行数据分类时选用算法更有效率更准确。同时对于同一类型方法的不同优化算法也能通过概念格进行分类与聚类。关于概念格的的应用还有许多问题有待研究，例如发展高效的构造概念格及剪枝算法；如何进一步优化算法制定方法，更好地考虑规则提取与属性约简问题。

参 考 文 献

[1] Wille R.Restructuring lattice theory： an approach based on hierarchies of concepts[M]∥Rival I，ed．Ordered Sets．Dordrecht：Reidel，1982：445-470

[2] Song Wei, Zhang Ming. A Concise Course of Semantic Web [M]. Beijing: Higher Education Press, 2004

(宋炜,张铭. 语义网简明教程[M]. 北京:高等教育出版社, 2004)

[3] Li Hang. Statistical Learning Methods [M]. Beijing: Tsinghua University Press, 2012

(李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012)

[4] Yang Fan, Zhai Yanhui, Qu kaishe, Li Deyu. Study on Word Meaning Comprehension Based on Formal Concept Analysis [J]. 2011, 38 (10): 189-191

(杨帆,翟岩慧,曲开社,李德玉. 基于形式概念分析的词义理解研究[J]. 2011, 38(10): 189-191)

[5] Wang Na. Knowledge Acquisition Based on Concept Lattice [J]. Science and technology entrepreneurship, 2010, 6 (4): 118-120

(王娜. 基于概念格的知识获取 [J]. 科技创业, 2010, 6(4): 118-120)

[6] Xie Zhipeng, Liu Zongtian. A Fast Incremental Algorithm for Building Concept Lattice [J]. Chinese Journal of Computers, 2002, 35 (6): 628-639

(谢志鹏,刘宗田. 概念格的快速渐进式构造算法[J].计算机学报, 2002, 35(6): 628-639)

[7] Miao Ru, Shen Xiazhu. Mining Rules from Concept Lattices [J]. Optical disc technology, 2006 (1): 10-11

(苗茹,沈夏炯. 概念格中的规则提取[J].光盘技术,2006(1): 10-11)

[8] Sahami M. Learning classification rules using lattices [A]. In: Lavran N, Wrobel S, eds. Proceedings of ECM L-95 [C]. Grete, Greece, 1995. 343-346

[9] Hu Keyun, Lu Yuchang , Zhou Lizhu, et al.Integrated classification and association rule mining based on concept lattice [A]. Zhong N, Skowron A,eds. Proceeding s o f RSFDGrC99 [C]. Toyko: Springer, 1999. 443-447

[10] Wang Zhihai, Hu Keyun, Hu Xuegang, Liu Zongtian, Zhang Diancheng. General and Incremental Algorithms of Rule Extraction Based on Concept Lattice [J]. Chinese Journal of Computers, 1999, 22 (1): 67-70

(王志海,胡可云,胡学钢,刘宗田,张奠成. 概念格上规则提取的一般算法与渐进式算法[J].计算机学报,1999，22(1): 67-70)