一种基于模糊聚类的模糊本体生成方法*)

强 宇[1,2] 刘宗田[1] 李 旭[1] 周 文[1] 陈慧琼[1]

(上海大学计算机学院 上海 200072)[1] (蚌埠坦克学院 安徽蚌埠 233013)[2]

摘 要 本文研究了一种从模糊背景生成模糊本体的方法。模糊本体由以下几部分组成,分别是:模糊形式概念分析、模糊概念聚类及模糊本体生成。首先,模糊形式概念分析将模糊逻辑嵌入形式概念分析以构成模糊概念格。其次,模糊概念聚类从模糊概念格构造概念层次。最后,模糊本体生成部分从概念层次生成模糊本体。

关键词 形式概念分析,模糊逻辑,概念聚类,本体生成

A Fuzzy Ontology Generation Method Based on Fuzzy Cluster

QIANG Yu[1,2] LIU Zong-Tian[1] LI Xu[1] ZHOU Wen[1] CHEN Hui-Qiong[1]

(School of Computer Engineering and Science, Shanghai University, Shanghai 200072)[1] (Bengbu Tank College, Bengbu 233013)[2]

Abstract In this paper, we research a method for generating the fuzzy ontology from fuzzy context. Fuzzy ontology is constructed by several parts: fuzzy formal concept analysis, fuzzy concept cluster and fuzzy ontology generation. First, Fuzzy formal concept analysis introduces the fuzzy logic into formal concept analysis to construct fuzzy concept lattice. Second, Fuzzy concept cluster constructs concept hierarchy from fuzzy concept lattice. And the last, Fuzzy ontology is generated from concept hierarchy.

Keywords Formal concept analysis, Fuzzy logic, Fuzzy cluster, Ontology generation

1 引言

本体是将域概念化成人类可理解、机器又可读的形式,此形式由实体、属性、关系及公理组成[1]。本体采用类表示概念,并且支持类间的分类和非分类关系。在很多应用领域中,由经典本体支持的形式化概念不足以表示不确定信息。例如从科学出版物提取的关键词可用以推论相应的研究领域,但同等对待所有的关键词不合适,因为某些词比另外一些更有意义。而且也很难判断一个文本是否绝对属于一个领域。

通常的解决方案是将模糊逻辑引入本体以处理不确定信息。隶属度用以评估一个概念层次上概念间的相似性。从预先定义的概念层次手工生成模糊本体很难,常需要专家解释,因此自动生成很有益。此文提出的方法可自动生成不确定信息的模糊本体。与现有的模糊本体生成技术比,此法可自动构造本体类的概念层次结构。

2 预备知识

从概念层自动生成本体须有很多相关知识,像自然语言处理(NLP),关联规则生成[2],统计模型[3],聚类[4]。对于本体学习,聚类是最有效的技术。COBWEB是很有效的技术,可用于生成本体的概念表示和关系[5]。

FCA是数据分析和知识表示的形式化技术,对概念聚类有效。但多数概念格在生成概念时具有巨大的时空复杂性,故很需要简化格。在冰山格中[6],关联规则用于聚类格上的概念,概念分层[7]用于生成概念层次。

Pollant [8]提出了一种模糊背景,将模糊逻辑与FCA结合起来。采用语言变量(与模糊集有关的语言项),以表示背景的不确定性。但语言变量的定义往往需要人类的解释。在实际应用中,信息多是模糊、不确定的,从模糊背景构造的格比经典格有更广阔的实用背景。

3 概念层次生成框架

概念层次的自动生成法如图1所示。

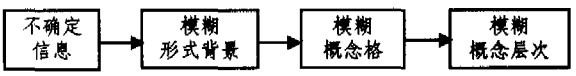

图1 模糊概念层次的生成

其中模糊概念层次包括几个部分,分别是:不确定信息、模糊形式概念分析、模糊概念格、模糊聚类技术、模糊概念层次。

模糊形式概念分析包含从不确定数据的数据库构造模糊形式背景、从模糊形式背景构造模糊概念格。

模糊概念聚类是采用一定的聚类技术、依据格结点间的相似度参数对模糊概念格上的概念进行聚类,生成模糊概念聚类集合,进而生成模糊概念层次。

4 模糊形式概念分析

为了做模糊形式概念分析,可将模糊逻辑植入形式概念分析,以表示模糊信息。

定义1 模糊背景是一个三元组 $K=(O,D,I)$,其中 O 是对象集,D 是属性集,I 是域 $G\times M$,每对关系 $(o,d)\in I$ 有隶属度值 $I(o,d)$,值落在[0,1]中。

定义2 给定一个模糊形式背景 $K(O,D,I)$ 和阈值 Φ_d,

*)本文受国家自然科学基金(60275022)和上海市高等学校青年发展基金(03AQ99)资助。强 宇 博士生,讲师,研究方向为人工智能,数据挖掘。刘宗田 博导,教授,研究方向为人工智能,软件工程。李 旭 硕士生,研究方向为数据挖掘。周 文 博士生,研究方向:人工智能。

对 $\forall O_1 \subseteq O, f(O_1)=\{d\in D \mid \forall o\in O, I(o,d)\geqslant \Phi_d\}$，

对 $\forall D_1 \subseteq D, g(D_1)=\{o\in O \mid \forall d\in D_1, I(o,d)\geqslant \Phi_d\}$；

带隶属度的模糊背景的模糊概念是二元对 (O_i,D_i)，$O_i\subseteq O, D_i\subseteq D, f(O_i)=D_i, g(D_i)=O_i$。

定义 3 设 $(O_1,D_1),(O_2,D_2)$ 是模糊形式背景 (O,D,I) 的两个模糊概念，当且仅当 $O_1\subseteq O_2$，则有 $D_2\subseteq D_1$，则 (O_1,D_1) 是 (O_2,D_2) 的子概念，(O_2,D_2) 是 (O_1,D_1) 的超概念。

定义 4 带模糊隶属度阈值的模糊形式背景的模糊概念格是 K 的所有模糊概念的集合。

定义 5 模糊形式概念 (O_1,D_1) 和其子概念 (O_2,D_2) 的相似度可定义成 $E_{\text{概念相似度}}(c_1,c_2)=\dfrac{\varphi(O_2)}{\varphi(O_1)}$

$$\varphi(O_i)=\sum_{o\in O_i}\mu(o,D_i),\ \mu(o,D_i)=\min_{d\in D_i} I(o,d)$$

模糊形式背景可表示成表 1，采用阈值 Φ_d，可消除低隶属度值。基于对象的成员隶属度可定义成 $\varphi(O_1)=\min\limits_{d\in D_1} I(o,d)$，其中 $o\in O, I(o,d)$ 是 I 中定义的对象 O 和属性 d 间的隶属度值[9]。因为对象与概念的关系是对象和属性的关系的交集，故据模糊理论[9]，隶属度交集取最小值。

表 1 是模糊背景。

图 2 是应用渐进生成法得到的模糊概念格，图 2 中包含模糊形式概念的隶属度值及模糊形式概念结点间的相似度参数。

表 1 处理后的模糊背景

o_i	d_1 咳嗽	d_2 咳嗽	d_3 咳嗽	d_4 头疼	d_5 头疼	d_6 血压	d_7 血压
1	0.8	0.2	0.9	0.1	0.8	0.2	0.0
2	1.0	0.0	0.0	1.0	0.6	0.4	0.0
3	1.0	0.0	0.1	0.9	0.9	0.1	0.0
4	0.3	0.7	0.7	0.3	0.0	0.6	0.4
5	0.6	0.4	0.7	0.3	0.0	0.8	0.2
Φ_{di}	0.74	0.26	0.48	0.52	0.46	0.42	0.12

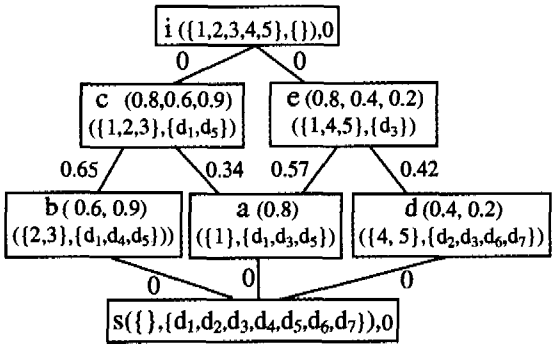

图 2 模糊概念格

5 模糊概念聚类技术

在传统概念格中，形式概念是用数学方法生成的，关于属性值差异很小的对象可分类成一个形式概念，在较高的层次上可聚类成一个概念，基于此法，可将此概念聚类成一个概念。基于此法，可用模糊概念聚类技术把概念聚成多个聚类，概念聚类可以有如下性质：

概念聚类有层次关系，关系可从模糊概念导出，由概念聚类表示的概念可以是其他概念聚类的子概念或超概念。可用近似置信阈值确定两概念是否近似。

阈值 Ls 采用 0.6，应用聚类技术得到的模糊概念层次如图 3 所示。

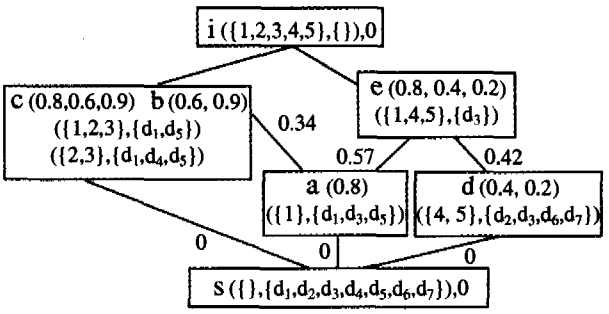

图 3 模糊概念层次

模糊聚类的算法伪码是：

保存模糊格中的所有边，放入边集。边包括结点 c_i, c_j 和结点间的相似度信息。

其中 c_i 是 c_j 的父结点。

循环查找边集中的每条边，如边相连的两点的相似度值大于阈值，则两点放入一个聚类集，子结点吸收父结点的内涵，保存子结点内涵。

连到 c_i 父结点的边更新为连到新结点，

连到 c_j 的子结点的边更新为连到新结点。

直到边集中无边可消去。

6 聚类有效性评价

本体是基于概念层次构造的，故本体质量多依赖于概念层次的质量。概念层次是基于聚类技术得到的，容错法[10]通常用于评估聚类的有效性。另外还可以度量平均无插值精确性以评估概念层次的可恢复性。

通过定义聚类的有效性可以对模糊聚类作出有效评价；

聚类 C 的有效性 V 定义为：

$V(c)=1-r(c)$ 其中 $r(c)$ 是聚类容错性。

$$r(C)=\sum_{d_i\in D}\sum_{i=1}^{n}\sum_{j=1}^{n} p(c_i)\,p(c_j)\,D^{d_i}(c_i,c_j)$$

其中 D 是聚类 C 中的属性集，$P(c_i)$ 是聚类 C 中 c_i 的概率，$D^{d_i}(c_i,c_j)$ 是属性 d_i 上 c_i 和 c_j 的距离。

$D^{d_i}(c_i,c_j)=|I(c_i,d_i)-I(c_j,d_i)|$，$I(c_i,i)$ 和 $I(c_j,j)$ 是属性 d_i 上对象 c_i 和 c_j 的隶属度值。

V 值越大，说明聚类有效性越好。

7 模糊本体生成

应用模糊聚类技术导出的模糊概念层次可以用于模糊本体的构造。模糊概念格中的每个模糊概念包含外延和内涵信息，模糊概念层次体现了模糊概念之间的分类关系；而本体描述的是概念和概念之间的关系，包含 5 个基本的建模元语，分别是：类、关系、函数、公理和实例。所以，从模糊概念格到模糊本体的构造，需要将两者的内容做映射，具体的映射方法如下所示：

1)给每个模糊概念层次中的概念节点一个标识，每个标识名对应模糊本体中的一个类名；模糊概念之间的层次关系对应本体中相应类之间的分类关系；

2)本体中每个类对应的属性由模糊概念层次中相应模糊概念内涵对应的模糊语言变量值表示，属性的值对应形式背景中模糊隶属度值；

3)本体中类的实例即模糊形式背景的对象。

由上述三个步骤生成的本体具一致性，故不需要一致性检查；本体中类的对象即实例的属性值是用模糊值表示，体现了现实性；但是，本体中表示的关系相对比较单一，现实中存

在的大量非分类关系如何得到，须通过专家参与，人为加入已有的本体原型，并扩展，才得到比较完整的本体模型。

整个从模糊形式背景到本体模型的生成过程可以用图 4 表示。

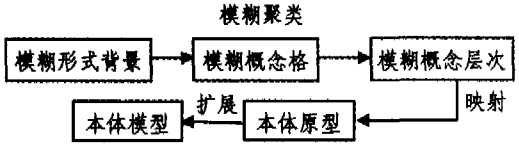

图 4　模糊本体生成过程

下面以图 4 的模糊概念层次举例说明模糊本体的生成。以此为例，映射到本体可得 A、B、C、D 四个类（见图 5），以及类之间的分类关系（见图 5）。

表 2　本体原型中类间的关系

本体	概念聚类层次中的概念
A	c(0.8,0.6,0.9)b(0.6, 0.9) ({1,2,3},{d_1,d_5})
B	e(0.8, 0.4, 0.2) ({1,4,5},{d_3})
C	a(0.8) ({1},{d_1,d_3,d_5})
D	d(0.4, 0.2) ({4, 5},{d_2,d_3,d_6,d_7})

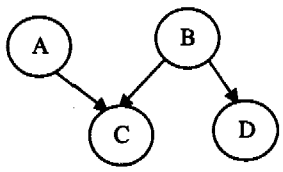

图 5　本体原型中类间关系

例如对于类 A，具有属性咳嗽及血压；实例为 1，2，3；实例 1 的属性咳嗽的值为经常（可信度为 0.8），属性血压的值为高（可信度为 0.8）。

同理，可以得到 B，C，D 等类的属性及实例，在实际应用中还可以通过其它方式加入概念间的非分类关系，最后得到完整的本体模型。

结束语　本文研究了基于模糊背景生成模糊本体。提出了一种基于模糊聚类技术生成模糊本体的方法。定义了反映模糊聚类有效性的度量参数。未来的研究方向还包括基于模糊聚类的模糊本体生成算法研究，与 cobweb 的对比实验等。

参 考 文 献

1　Guarino N, Giaretta P. Ontologies and Knowledge Bases: Towards a Terminological clarification: Toward, 1995
2　Maedche A, Staab S. Ontology Learning for the Semantic Web. IEEE Intelligent systems, Special Issue on the Semantic Web, 2001, 6(2)
3　Faatz A, Steinmetz R. Ontology enrichment with texts from WWW. In: Proc. of Semantic Web Mining second Workshop at ECML/PKDD-2002, Finland, 2002
4　Bisson G, Nedellec C. Designing Clustering Methods for Ontology Building: The Mo'K Workbench. In: Staab S, Maedche A, Nedellec C, WiemerHasting P, eds. Proc. of the Workshop on Ontology Learning, 14th European Conf. on Artificial Intelligence, ECAI'00, Germany, 2000
5　Clerkin P, Cunningham P, Hayes C. Ontology Discovery for the Semantic Web Using Hierarchical Clustering. In: Proc. of Workshop at ECML/PKDD, 2001, Germany, 2001
6　Stumme G, Taouil R, Bastide Y, Pasquier N, Lakhan L. Computing iceberg concept lattice with Titanic. Journal on Knowledge and Data Engineering, 2002, 42(2)
7　Vogt F, Wille R. TOSCANA: a Graphical Tool for Analyzing and Exploring Data. In: Tamassia R, Tollis I G, eds. GraphDrawing'94, Heidelberg, 1995. 226～233
8　Pollandt S. Fuzzy-Begriffe: Formale Begriffsanalyse unscharfer Daten. Springer Verlag, Berlin-Heidelberg, 1996
9　Zadeh L A. Fuzzy Sets. Journal 1996. of Information and Control, 1965, 8: 338～353
10　Chu W, Chiang K. Abstraction of High Level Concepts from Numerical Values in Databases. 1994. 133～144

（上接第 144 页）

人智能行为变量，$\theta=(\theta_1,\theta_2,\cdots,\theta_n)\in\Theta$，$\Theta$ 为 n 维实空间 R^n 的开集，这里用 $\Theta(\theta_1,\theta_2,\cdots,\theta_n)$ 表示该结构对应的条件概率分布在流形 S 上的坐标参数，由于其结构实现的复杂性，要对其进行结构分解，分解为多个子系统模型，这些子系统模型协同完成复杂体系结构模型具有的功能。用 $S_1,S_2,\cdots,S_m$，$(1<m\leqslant n)$ 表示分解后的子系统所对应的子流形，证明可分解性定理即等价于证明 $S_1S_2\cdots S_m$ 微分同胚于 S，下面用数学归纳法证明：

当 $m=2$ 时，S_1、S_2 为 S 的一个结构分解，设 (S_1,Ψ_1) 为 r 维微分流形，$1\leqslant r<n$，流形上的点的坐标参数是 $\Theta(\theta_1,\theta_2,\cdots,\theta_r)$，$(S_2,\Psi_2)$ 为 $n-r$ 维微分流形，流形上的点的坐标参数是 $\Theta(\theta_{r+1},\theta_{r+2},\cdots,\theta_n)$，$\Psi_1$、$\Psi_2$ 分别为 S_1 和 S_2 的坐标卡集。构造积拓扑空间 S_1S_2，并定义微分构造：

$$\Psi'=\{(U_\alpha V_\beta,\ f_\alpha g_\beta)\mid(U_\alpha,f_\alpha)\in\Psi_1,(V_\beta,g_\beta)\in\Psi_2\}\tag{13}$$

对任意 $(u,v)\in U_\alpha V_\beta$，有

$$(f_\alpha g_\beta)(u,v)=_{def}(f_\alpha(u),g_\beta(v))\tag{14}$$

因为

$$S_1=\bigcup_\alpha U_\alpha,S_2=\bigcup_\beta V_\beta\tag{15}$$

所以

$$S_1S_2=\bigcup_\alpha U_\alpha\bigcup_\beta V_\beta=\bigcup_{\alpha,\beta}U_\alpha V_\beta\tag{16}$$

又

$$\begin{aligned}(f_\alpha g_\beta)(u,v)&=(f_\alpha(u),g_\beta(v))\\&=_{def}(x,y)\in(f_\alpha(U_\alpha),g_\beta(V_\beta))\end{aligned}\tag{17}$$

且

$$\begin{aligned}(f_\alpha^{-1}g_\beta^{-1})(x,y)&=(f_\alpha^{-1}(x),g_\beta^{-1}(y))\\&=(f_\alpha^{-1}\circ f_\alpha(u),g_\beta^{-1}\circ g_\beta(v))=(u,v)\end{aligned}\tag{18}$$

所以

$$(f_\alpha g_\beta)^{-1}=f_\alpha^{-1}g_\beta^{-1}\tag{19}$$

即

$$(f_\alpha g_\beta)^{-1}(x,y)=(f_\alpha^{-1}(x),g_\beta^{-1}(y))\tag{20}$$

$$\begin{aligned}f_\alpha g_\beta:U_\alpha V_\beta&\to(f_\alpha g_\beta)(U_\alpha V_\beta)\\&=f_\alpha(U_\alpha)g_\beta(V_\beta)\subset R^rR^{n-r}=R^n\end{aligned}\tag{21}$$

是同胚映射。

设 $m=k$ 时，S_1 分解为 $(S_1^1,S_1^2,\cdots,S_1^r)$，$S_2$ 分解为 $(S_2^{r+1},S_2^{r+2},\cdots,S_2^k)$，$S_1$ 和 $S_1^1S_1^2S_1^r$ 是同胚的，S_2 和 $S_2^{r+1}S_2^{r+2}S_2^k$ 是同胚的，S_1S_2 和 S 依然是同胚的。

当 $m=k+1$ 时，无论是对 S_1 或 S_2 进行分解，S_1S_2 和 S 均还是同胚的。证毕。

结论　本文证明了智能机器人复杂体系结构的可分解性，这为智能机器人体系结构模型的层次化、模块化实现奠定了理论基础。证明中所采用的微分流形理论，不仅可以用于分析智能机器人体系结构的可分解性，而且可以用于分析智能机器人的信息、控制和问题求解能力的分布模式、内在机理、学习机制和整体特性，从而在智能机器人的结构与功能之间建立拓扑对应关系，更好地指导智能机器人体系结构的设计工作。

参 考 文 献

1　蒋心松. 机器人学导论[M]. 沈阳：辽宁科学技术出版社，1994
2　Hecht-Nielsen R. Kolmogorov's mapping neural network existence theorem [A]. In: Proc. of the International Conference on Neural Networks, volume 3 [C]. New York, 1987. 11～14
3　罗四维. 大规模人工神经网络理论基础[M]. 北京：清华大学出版社，2004
4　Amari S. Differential Geometrical Methods in Statistics [M]. Berlin: Springer-Verlag, 1985