基于高阶逻辑的复杂结构数据半监督聚类

李琳娜[1] 陈海蕊[2] 王映龙[3]
(中国科学技术信息研究所 北京 100038)[1] (濮阳职业技术学院 濮阳 457000)[2]
(江西农业大学软件学院 南昌 330045)[3]

摘 要 半监督聚类近年来成为了机器学习和数据挖掘领域的研究热点。目前存在的半监督聚类方法都采用属性-值的知识表示方式。但属性-值语言在表示复杂结构数据时存在很多弊端,而基于高阶逻辑的知识表示语言 Escher 能较好地表示复杂结构数据。在 Escher 的知识表示方式下,首先当先验知识是实例之间的约束信息时,提出了搜索 K-Means 算法的 K 个初始质心的方法;其次,对先验知识不完全、能够发现的初始质心的个数 r 小于 K 的情况,提出了搜索其余的 $K-r$ 个初始质心的算法 MSS-KMeans 和 SMSS-KMeans;最后在复杂结构数据集上,验证了所提算法的可行性。最终的实验结果表明,基于高阶逻辑知识表示方式的半监督聚类方法要优于基于属性-值语言的半监督聚类方法。

关键词 高阶逻辑,模板,半监督聚类,复杂结构数据,先验知识
中图法分类号 TP391 **文献标识码** A

Semi-supervised Clustering of Complex Structured Data Based on Higher-order Logic

LI Lin-na[1] CHEN Hai-rui[2] WANG Ying-long[3]
(Information Institute of Science and Technology Chinese, Beijing 100038, China)[1]
(Puyang Vocational and Technical Colledge, Puyang 457000, China)[2]
(College of Software, Jiangxi Agriculture University, Nanchang 330045, China)[3]

Abstract Semi-supervised clustering algorithms have recently received a significant amount of attention in the machine learning and data mining of communities. All of current algorithms use attribute-value languages to represent knowledge. Attribute-value languages have inherent drawback for represent complex structured data. However, knowledge representation language Escher based on higher-order logic, can represent complex structured data. With Escher as knowledge representation formalism, firstly, when prior knowledge is pairwise constraints between instances, the method of initializing the K-Means algorithm was proposed; secondly, when the number r of cluster centers which can be initialized with incomplete knowledge is less than K, the algorithms MSS-KMeans and SMSS-KMeans were proposed to initialize the rest $K-r$ cluster centers. Finally, the empirical study carried out on datasets of complex structure data showed the feasibility of the presented algorithms. The final experimental results demonstrate the comparability between the presented algorithms and the known algorithms based on attribute-value language.

Keywords Higher-order logic, Template, Semi-supervised clustering, Complex structure data, Prior knowledge

聚类通常被认为是一种无监督的数据分析方法[1],然而在实际问题中,可以很容易地获得有限的样本先验信息,如样本的成对限制信息。这些知识不仅仅可以进行聚类结果的外延性验证,还可以用来"引导"或"调整"聚类过程,从而改善无监督聚类算法的性能[2],这样的聚类方法称为半监督聚类。半监督聚类已成为数据挖掘和机器学习领域的研究热点。

根据使用先验信息方法的不同,已有的半监督聚类算法被分成两大类[3]:一类是基于测度的方法,这类方法首先训练相似性测度,以满足类属或限制信息,然后使用基于测度的聚类算法进行聚类;另一类是基于约束的方法,该方法对先验信息的使用又分为使聚类的目标函数满足已知的先验信息,让先验信息限制聚类过程中实例的划分和利用已知的先验信息初始化聚类参数。本文所提的半监督聚类算法属于后者,算法既利用先验信息初始化聚类参数,又用其引导聚类过程。

2001年,Wagstaff 等提出了基本的利用先验知识限制聚类过程的算法 COP-KMEANS[4],他们根据给定的实例对象之间的约束信息使聚类结果满足给定的限制。

随后 Basu 等人提出了算法 Seeded- KMeans 和 Constrained- KMeans,这两个算法都是利用给定的一些实例的类属信息先验知识。前者将根据先验知识得到每个簇的实例成员的平均值,作为簇的初始质心,然后运行基本的 K-Means 算法,该方法要求先验信息中要有关于每个簇一些成员的类

到稿日期:2008-10-21 返修日期:2009-02-27 本文受国家自然科学基金(60875029)资助。
李琳娜(1981-),女,博士,主要研究方向为机器学习等,E-mail:today8181@163.com;陈海蕊(1979-),女,讲师,主要研究方向为数据挖掘等;王映龙(1973-),男,博士,主要研究方向为图挖掘等。

属信息。后者在根据前者的方法得到 K 个初始质心后，在聚类过程中保持这些已知类属信息的实例的类标号不变[5,6]。

2004 年，Basu 根据实例对象之间约束信息这样的先验知识，得到一些相互不相交的实例的集合，然后将每个集合看作一个簇，用这些集合成员的平均值作为相应的簇的初始质心。当集合的个数 r 大于预得到的簇个数 K 时，选择最大的 K 个集合；当集合的个数 r 小于 K 时，随机选择 $K-r$ 个不能属于这些集合的实例作为其余的初始质心[7,8]。

2007 年，Wang Chao 等人针对当给定的初始种子的个数 r 小于预得到的簇个数 K 时的问题，提出了两种寻找其余的 $K-r$ 初始质心的方法 FS-KMeans 和 SS-KMeans。前者选择 $K-r$ 个距离已知的 r 个质心最远的实例作为其余的初始质心；后者根据已知的 r 个初始质心对数据集进行 K-Means 聚类，然后从聚类结果中随机选择一个簇对其执行二分 K 均值划分。算法将其划分为两个类，将每个类成员的平均值作为其初始质心[9]。

这些技术都是采用属性-值的知识表示方式，不同之处仅仅在距离计算方法，属性的选择或属性的加权方式上，没有针对具体的数据集或任务考虑知识表示方式对聚类结果的影响。采用属性-值的知识表示方式虽然可以得到效率较高的算法，但对于复杂结构数据的学习任务，具有以下缺点：

1) 属性-值语言是基于命题逻辑的，是一种表达能力比较弱的形式化语言，不能描述具有复杂结构的数据。

2) 由于数据通常具有很多特征，如何选择对归纳学习最有利的特征是一个很困难的问题，很难有一个对所有的应用领域都适用的特征选择标准。

3) 不能够描述属性值之间的实质关系。

为了解决属性-值的知识表述方式以上几个方面的弊端，C. Giraud-Carrier，J. W. Lloyd 等人提出了类型化的高阶逻辑的知识表示方式。这方面最经典的知识表示系统是 Escher 语言[10]，与基于命题逻辑的知识表示方式相比，具有如下特征：

1) 以更加直接、自然、简洁的方式表达具有复杂结构的数据。

2) 支持各种数据类型，如集合、多集及图等任意复杂的类型，能描述复杂结构的数据。

3) 实例空间中的每一个实例都用一个封闭的项进行描述，将实例的所有信息集中在一个位置，有利于学习过程中使用这些信息。

本文采用 Escher 的知识表示方式，研究了复杂结构数据的半监督聚类问题。首先对先验知识是实例之间的约束信息时，提出了搜索 K-Means 算法的 K 个初始质心的方法。其次，对先验知识不完全、能够发现的初始质心的个数 r 小于 K 的情况，提出了搜索其余的 $K-r$ 个初始质心的算法 MSS-KMeans 和 SMSS-KMeans。最后在复杂结构数据集上，验证了所提算法的可行性。最终的实验结果表明基于高阶逻辑知识表示方式的半监督方法与基于属性-值语言的方法具有可比性。

1 Escher 介绍及其表示方式下实例之间的距离

1.1 Escher 介绍

Escher 是一个基于高级逻辑的知识表示语言，能够表达如下类型：

· 整数、浮点数、字符串、字符集合、布尔型。这一组类型称为基本类型，是表达数据的项的基本构造块。在 Escher 中，Int 表示整数类型，Float 表示浮点类型，Char 表示字符串类型，String 表示字符串集合的类型，Bool 表示布尔类型。

· 数据构造器。对用户定义类型，数据构造器是必需的。0 元的数据构造器通常称为常数。数据构造器在 Escher 用 data 关键字声明。关键字 data 表示一个类型的声明以及该类型相应的数据构造器。每个数据构造器构造了一个以多个常数作为域值的类型，关键字 type 声明一个由其他类型构造的类型。

· 元组。元组本质上是属性-值表达方式中数据表达的基础，其作用显而易见。

· 集合、多元素集合。集合和多元素集合虽然没有元组类型应用范围广，但是集合尤其是多元素集合也是相当有用的数据类型。

· 链表。可以将链表类型转换为集合类型等。

· 树。树定义为 data Tree T= Node T[Tree T]，其中 T 是节点的类型。

· 图。图又分为有向图和无向图。无向图定义为

type Label = Int;

type Graph $v\,w$ = {Label, v}, {(Label→Int, w)}}

其中，v 是顶点所表示的信息的类型，w 是边所表示的信息类型。有向图的定义与其类似。

下面举两个例子展示 Escher 的知识表示方式[11]。

第一个例子是一个比较典型的属性-值描述的数据：根据天气情况判断是否打网球[12]。

首先 Escher 要给出实例的类型定义。在该例中实例 Weather 用元组类型表达，其具体定义如下：

data Outlook=Sunny|Overcast|Rain;

data Temperature=Hot|Mild|Cool;

data Humidity=High|Normal|Low;

data Wind = String|Medium|Weak;

type Weather = (Outlook, Temperature, Humidity, Wind)。

上述声明定义了表示实例 Weather 的元组类型：该元组由 4 个属性组成，其类型分别为 Outlook，Temperature，Humidity，Wind。4 个属性类型由相应的 data 关键字分别定义。

下面给出具体实例的表示。

(Overcast, Hot, High, Weak)：该实例表示阴天、气温较高、空气湿度较大、风力比较弱的情况；

(Sunny, Hot, High, Weak)：该实例表示天气晴朗、气温较高、空气湿度较大、风力比较弱的情况。

1.2 Escher 表示方式下实例之间的距离

在 Escher 的知识表示方式下，一个实例用一个基本项表示，故实例之间的距离是基本项之间的距离[14]。

假设距离函数为 d，则 $d:\beta\times\beta\to R$，R 表示实数集，β 表示一个基本项。函数 d 的定义依赖于给定的函数 ρ。实数值函数 ρ 定义在数据构造器与其自身的积上，且满足以下条件：

1)对每一个类型构造器 $T\in\zeta$，ρ 是与 T 相关的数据构造器上的矩阵。

2)对于至少有一个数据构造器的元数>0 的类型构造

器，ρ 是离散矩阵。

例如，类型构造器 List 有两个数据构造器：(元数>0)和[]，且 $\rho([],:)=1$。相反，Nat 只有一个空的数据构造器，因此无法应用第二个条件。从下面关于函数 d 的定义可以看出，只需要考虑与相同的类型构造器相关的数据构造器 C 和 D 的 ρ 函数值，即 $\rho(C,D)$。

例 1　对类型 1 和 Ω 来说，ρ 可能是离散矩阵。对类型 Nat，Float，可以用 $\rho(n,m)=|n-m|$。对类型构造器，使用定义在数据构造器上的离散矩阵。

定义 1　若 $s,t\in\beta$，递归定义在 β 中的项结构上的函数 $d:\beta\times\beta\to R$ 如下：

1) 如果 $s,t\in\beta\alpha$，这里 $\alpha=T\alpha_1\cdots\alpha_k$，对某个 $T,\alpha_1,\cdots,\alpha_k$，那么如果 $C\neq D, d(s,t)=1$，否则，$d(s,t)=\sum_{i=1}^{n}\frac{1}{2^i}d(s_i,t_i)$。

其中，$s=Cs_1\cdots s_n$，$t=Dt_1\cdots t_m(n\geqslant m)$，$\beta\alpha$ 表示由个体类型为 α 的项所构成的集合。该部分适用于由基本类型或用户定义类型构成的复杂类型的情况，也可以是复杂类型构成更复杂类型的情况。它表示复杂类型之间的距离由它们的子类型之间的距离定义。

2) 如果 $s,t\in\beta\alpha$，这里 $\alpha=\beta\to\gamma$，对某 β,γ，那么

$$d(s,t)=\frac{\Sigma_{B_\beta}d(V(sr),V(tr))}{1+\Sigma_{B_\beta}d(V(sr),V(tr))}$$

其中，$\beta\to\gamma$ 表示集合或多集类型，其中的每个元素类型为 β,γ 通常是自然数类型，$V(sr)$ 表示将 s 作用于 r 得到的值。

3) 如果 $s,t\in\beta\alpha$，这里 $\alpha=\alpha_1\times\cdots\times\alpha_k$，对某个 $\alpha_1,\cdots,\alpha_k$，则

$$d(s,t)=\frac{1}{n}\sum_{i=1}^{n}d(s_i,t_i)$$

其中，$s=(s_1,\cdots,s_n)$，$t=(t_1,\cdots,t_m)$。该情况主要适用于元组类型。

4) 如果不存在 $\alpha\in\upsilon^*$ 使得 $s,t\in\beta\alpha$，那么 $d(s,t)=1$。该情况表示如果两个个体的类型不匹配，则它们的距离为 1。

2　基于 Escher 的半监督聚类算法 MSS-KMeans 和 SMSS-Kmeans

由于非领域专家也能给出一些实例对象之间的约束信息，故这样的先验知识比实例的类标记信息更容易获得。同时一些实例的类标记信息可以转化成实例对象之间的约束信息，故实例对象之间的约束信息比实例的类标记信息更一般化。在本文中，先验知识是实例对象之间的约束信息，这些知识用来初始化 K-means 聚类算法的初始质心及引导聚类过程。

本文的 K-means 聚类算法不同于传统的该方法。在将每个点指派到最近的质心后，重新计算每个簇的质心时，不是计算簇中包含的所有实例的平均值，而是选择簇中相对中心的实例作为簇的质心。相对中心的实例是与其它实例的距离的平方和(SSE，Sum of the Squared Error)最小的实例。该方法相似于通常使用的 k-medoid 算法中寻找质心的思想。

实例对象之间的约束信息包括：两个实例对象必须属于同一个类(must-link)和两个实例对象一定不属于同一个类(cannot-link)。下面介绍如何根据先验信息初始化 K 个初始质心。

首先采用如下方式将这些信息转化为图。将每个实例对象看作一个数据点，若两个实例对象必须属于同一个类，则相应的数据点之间有边相连。若这些先验知识是完全的，整个图应该是由一些相互不相连的连通子图构成的，每一个连通的子图就是一个预先已知一些成员的簇。然后对每一个这样的簇，在该簇的已知成员中选择使得距所有已知不能属于该簇的实例的 SSE 值最大的实例作为该簇的初始质心。即若根据先验知识得到某一个簇 M，其有 p 个必属于它的实例，$M=\{x_i\mid 1\leqslant i\leqslant p\}$。集合 N 是根据先验知识获得的 q 个不能属于簇 M 的实例，$N=\{y_j\mid 1\leqslant j\leqslant q\}$。若 x_i 使得 $\operatorname{argmax}\sum_{1\leqslant j\leqslant q}|y_i-x_i|$ 成立，则选择实例 x_i 为 M 的初始质心。

从上面的步骤中得到的关于每一个簇中的实例的信息，在聚类的整个过程中保持不变。在聚类的过程中，若两个在先验知识中约束为不能属于同一个簇的实例分配到同一簇的时候，随机选择其中的一个，将其划分到别的簇中。

在很多实际应用领域，经常存在着不完全的背景知识，即利用先验的背景知识能够初始化的初始质心的个数 r 少于预得到的簇个数 K。针对这个问题，提出两个初始化其余的 $K-r$ 个初始质心的方法，分别为：

1)依次选择 $K-r$ 个距离所有的已发现质心的 SSE 值最大的样例作为初始质心，即 MSS-KMeans 算法(Maximal-SSE-Seeded-KMeans)；

2)首先根据已发现的 r 个初始质心采用 K-Means 聚类算法进行聚类，然后从该聚类结果中选择这样的簇，使得所有样例距其质心的 SSE 值最大。对其进行二分 K-均值算法，将其划分为两个簇，将每个簇中相对中心的点分别作为其初始质心。重复前面的过程，直到获得 K 个初始质心。该方法称为 SMSS-KMeans 算法(Splitting-Maximal-SSE-Seeded-KMea-ns)。

3　实验

所有的实验运行平台是具有以下参数的 PC：

- 赛扬 2.3GHz CPU
- 2GMB RAM
- Linux

3.1　数据集介绍

实验使用两个比较经典的数据集(数据集 1 和数据集 2)及一个人工构造的数据集(数据集 3)。

数据集 1 选自 UC 的 Audiology 数据集，该数据集中共有 226 个实例、26 个类。对该数据集分别采用属性-值的知识表述方式和 Escher 的知识表述方式。在 Escher 的表述方式下，每个实例用 list 类型表示。

数据集 2 是蛋白质的合成，包括 DNA 分子的转录以及随后的 mRNA 的翻译过程。mRNA 包含两个区域，除了包含蛋白质氨基酸序列的遗传信息外，还有一个非编码区，不构成蛋白质的编码。这些非编码区域包含有一些连续的部分，称为信号结构，可以将信号结构，其二级结构视为一棵树。根据信号结构的生物功能，可将它们分成不同的信号结构类[15]。属于同一个信号结构类的信号结构的生物功能在信号结构组合的时候具有共同的特征，属于同一个信号结构类的信号结构不必相同但必须相似。数据集 2 是由 66 个信号结构构成的数据集，它们属于 5 个信号结构类[16]。在 Escher 的表述方式下，信号结构的数据类型是有向图。

数据集 3 是元学习处理模型选择问题，是归纳从任务到学习器的映射[17]。目前很多的元学习采取首先抽取任务的特征，然后归纳特征到合适的学习模型的映射的学习机制。可将抽取特征的方法分为统计和信息理论特征、标记和基于决策树的特征 3 类。然而，这些方法都是手工或预先计算一些特征。H. Bensusan 等人提出了直接从归纳的决策树中学习，即复杂结构归纳学习应用于元学习问题[18]。首先用一定的决策树学习算法归纳出任务所对应的决策树，将树中每个节点的信息都用一个元组表示。元组包含的具体信息视实际情况而定，可以是该节点所使用的属性、所包含的样例的个数及一些启发式信息。然后将决策树转化为 Escher 所对应的封闭项，用来刻画学习任务的特征。再对这些得到的 Escher 项应用基于高阶逻辑的归纳学习算法加以归纳，得到元理论。收集一些数据集，首先对每个数据集应用决策树学习算法 C4.5，然后根据文献[18]中的方法将每个决策树转化为 Escher 项。每个模型的数据集分别为：神经网络；6 个数据集；决策树，11 个数据集；朴素贝叶斯，8 个数据集；贝叶斯网络，9 个数据集；K-最近邻，10 个数据集；SVM，12 个数据集；FOIL，4 个数据集。

3.2 评价方法

分别采用如下 2 种度量准则对算法的性能进行评价：

1)F-测度

在分类、聚类算法的性能评价中，F-测度是一个常用的评估指标，它由准确率和召回率定义。此处采用如下的成对 F-测度的定义：

$$Precision=\frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsPredictedInSameCluster}$$

$$Precision=\frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsActuallyInSameCluster}$$

$$成对\ F\text{-}测度=\frac{2\times Precision\times Recall}{Precision+Recall}$$

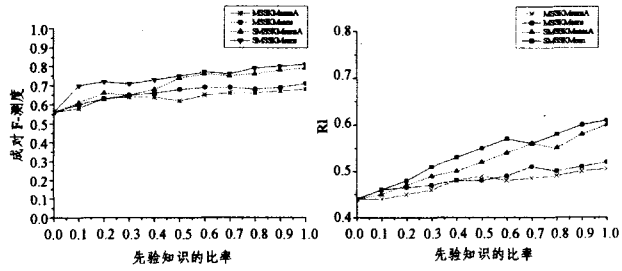
2)Rand Index 评价

最后一个比较常用的聚类评价指标是 Rand Index。对数据集 D，若 P_1 是事实上的 D 的划分或由领域专家给定的 D 的划分，P_2 是实验结果得到的 D 的划分，n 是 D 中所有实例的个数，则 D 中共有 $n\times(n-1)/2$ 个实例对。对 D 中的两个实例 D_i，D_j，A 是 P_1，P_2 同时将 D_i，D_j 分到同一个簇的实例对的个数，B 是 P_1，P_2 同时将 D_i，D_j 分到不同的簇的实例对的个数，则 $RI=\frac{A+B}{n\times(n-1)/2}$。

3.3 实验结果及分析

对每个数据集使用 2-交叉验证的方法生成学习曲线。为了展示先验知识对算法学习效果的影响，取一个数据集的 50%做测试，从剩下的 50%的数据部分中抽取实例对，生成先验知识。若抽取的两个实例对象属于同一个类，相应的约束信息就是它们必须属于同一类。若抽取的两个实例对象不属于同一个类，相应的约束信息就是它们不能属于同一类。图 2 的横轴是根据先验信息能确定的初始质心的个数 r 与预得簇个数 K 的比率，在图中标记为先验知识的比率。

在数据集 1 上，分别采用属性-值和 Escher 的知识表述方式。为了区别算法 MSS-KMeans 和 SMSS-KMeans 在这两种知识表达方式下的结果，在属性-值的表述方式下执行的 MSS-KMeans，SMSS-KMeans 算法分别记为 MSS-KMeansA，SMSS-KMeansA。根据成对 F-测度和 RI 的评价标准，结果分别展示在图 2(a)、(b)中。可以看出，对复杂结构数据，同样的半监督聚类算法在采用属性-值语言表示数据时性能上差于基于 Escher 的表示。

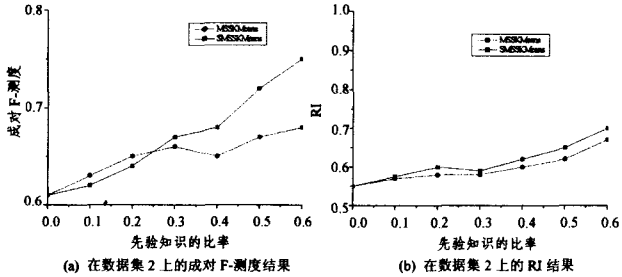


(a) 在数据集 1 上的成对 F-测度结果　(b) 在数据集 1 上的 RI 结果

图 2

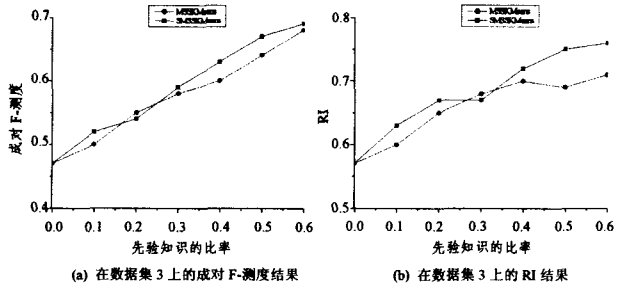
图 3(a)、(b)分别是算法 MSS-KMeans 和 SMSS-KMeans 在数据集 2 上成对 F-测度和 RI 的评价标准下的运行结果。图 4(a)、(b)是在数据集 3 上的运行结果。可以看出，SMSS-KMeans 在性能上要优于 MSS-KMeans，这是由于 SMSS-KMeans 首先根据初始化的质心对数据集进行聚类，然后从聚类结果中选择簇中实例距簇质心的 SSE 值最大的簇。当一个簇可以划分为两个更精确的簇时，其实例距簇质心的 SSE 值通常大于在同一度量标准下的已经很精确的簇的 SSE 值。

虽然算法 SMSS-KMeans 在性能上要优于算法 MSS-KMeans，但其效率低于后者。二者在性能上的优势差别不是特别悬殊，故都有一定的应用价值。



(a) 在数据集 2 上的成对 F-测度结果　(b) 在数据集 2 上的 RI 结果

图 3



(a) 在数据集 3 上的成对 F-测度结果　(b) 在数据集 3 上的 RI 结果

图 4

结束语　目前存在的半监督聚类算法都以属性-值作为知识表示方式，忽略了属性-值语言在描述复杂结构数据时的弊端。基于高阶逻辑的知识描述语言 Escher 在描述复杂结构数据时能克服属性-值语言的弊端。本文研究了以 Escher 作为知识表示方式的复杂结构数据的半监督聚类问题，主要研究内容为：

在先验知识是实例之间的约束信息的情况下，提出了发

现 K-Means 算法的 K 个初始质心的方法。

当先验知识不完全能够发现的初始质心的个数 r 小于 K 的情况下，提出了搜索其余的 $K-r$ 个初始质心的算法 MSS-KMeans 和 SMSS-KMeans。

在复杂结构数据集上，验证了所提算法的可行性。最终的实验结果表明基于高阶逻辑知识表示方式的半监督方法与基于属性-值语言的方法具有可比性。

基于类型化的高阶逻辑作为复杂结构数据的知识表达方式是一个比较新的研究领域，国内外学术界对这一领域的研究并不多见。随着机器学习与知识发现在复杂结构领域应用深度和广度的拓展，如在计算生物学、医学、病毒营销、反恐、语义 Web、社会网络分析、普适计算等领域的应用，我们相信基于高阶逻辑的复杂结构数据学习必将有着广阔的发展前景。

针对复杂结构数据的半监督聚类，下一步工作的方向包括：

1)进一步从理论上证明、分析我们的实验结果。

2)借鉴一阶逻辑的知识表达方式下计算实例之间距离时谓词加权的思想，改进基于高阶逻辑的知识表达方式下实例之间的距离计算方法。

3)探索基于高阶逻辑复杂结构数据的基于测度的半监督聚类方法。

参 考 文 献

[1] 周涛，张艳宁，袁和金，等. 粗糙核 k-means 聚类算法[J]. 系统仿真学报，2008，20(4)：921-925

[2] Tang W，Xiong H，Zhong S，et al. Enhancing semi-supervised clustering：a feature projection perspective[C]// Proceedings of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York，2007：707-716

[3] 王玲，薄列峰，焦李成. 密度敏感的半监督谱聚类[J]. 软件学报，2007，18(10)：2412-2422

[4] Wagstaff K，Cardie C，Rogers S，et al. Constrained K-means clustering with background knowledge[C]// Proceedings of the 18th Int'l Conf. on Machine Learning. Williamstown，2001：577-584

[5] Basu S，Banerjee A，Mooney R J. Semi-supervised Clustering by Seeding[C]// Proceedings of the 19th Int'l Conf. on Machine Learning. Sydney，2002：7-34

[6] Basu S，Banerjee A，Mooney R J. Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering[C]// Proceedings of the 20th Int'l Conf. on Machine Learning Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining. New Orleans，2003：42-49

[7] Basu S，Banerjee A，Mooney，R J. Active semi-supervision for pairwise constrained clustering[C]// Proceedings of the SIAM Int'l Conf. on Data Mining. Philadelphia，2004：333-344

[8] Bilenko M，Basu S，Mooney R J. Integrating constraints and metric learning in semi-supervised clustering[C]// Proceedings of the 21st Int'l Conf. on Machine Learning. Banff，2004：81-88

[9] Wang C，Chen W，Yin P，et al. Semi-supervised Clustering Using Incomplete Prior Knowledge[C]// Proceedings of the 7th Int'l Conf. on Computational Science. Berlin，2007：192-195

[10] Bowers A F，Giraud-Carrier C G，Lloyd J W. Classification of Individuals with Complex Structure[C]// Proceedings of the 17th Int'l Conf. on Machine Learning. Sydney，2000：81-88

[11] Bowers A F，Giraud-Carrier C G，Lloyd J W. A Knowledge Representation Framework for Inductive Leanring[EB/OL]. http：//csl. anu. edu. au/~jwl，2001

[12] Mitchell T. Machine Learning[M]. New York：McGraw-Hill，1997

[13] King R D，Muggleton S H，Srinivasan A，et al. Structure-activity Relationships Derived by Machine Learning：The Use of Atoms and Their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming[C]// Proceedings of the National Academy of Sciences. USA，1996：438-442

[14] Lloyd J W. Logic for Learning：Knowledge Representation，Computation and Learning in Higher-order Logic[M]. Berlin Springer-Verlag，Heidelberg GmbH & Co. KG，2002

[15] Horváth T，Alexin Z，Gyimothy T，et al. Application of Different Learning Methods to Hungarian Part-of-Speech Tagging[C]// Proceedings of the 9th Int'l Workshop on Inductive Logic Programming. London，1999：128-139

[16] Horváth T，Wrobel S，Bohnebeck U. Relational Instance-Based Learning with Lists and Terms[J]. Machine Learning，2001，43(1/2)：53-58

[17] Vilalta R，Drissi Y. A perspective view and survey of meta-learning[J]. Arti. Intell. Rev.，2002，18(2)：77-95

[18] Bensusan H，Giraud-Carrier C，Kennedy C. A Higher-order Approach to Meta-learning[R]. CS-EXT-2000-277. University of Bristol，2000

(上接第 153 页)

[3] Kolahdouzan M R，Shahabi C. Voronoi-Based K Nearest Neighbor Search for Spatial Network Databases[C]// Proc. of 30th Intl. Conf. on Very Large Data Bases. Toronto，Canada，2004

[4] Cho H-J，Chung C-W. An efficient and scalable approach to CNN Queries in a road network[C]// Proc. of 31th Intl. Conf. on Very Large Data Bases. Trondheim，Norway，2005

[5] Saltenis S，Jensen C S，et al. Indexing the Positions of Continuously Moving Objects[C]// Proc. of the 19th SIGMOD Intl. Conf. on Management of Data. Dallas，Texas，USA，2000

[6] Mokbel M F，Xiong Xiaopeng，Aref W G. SINA：Scalable Incremental Processing of Continuous Queries in Spatiotemporal Databases[C]// Proc. of the 23rd SIGMOD Intl. Conf. on Management of Data. Paris，France，2004

[7] Yu Xiaohui，Pu K Q，Koudas N. Mointoring k-Nearest Neighbour Queries over Moving Objects[C]// Proc. of the 21st Intl. Conf. on Data Engineering. Tokyo，Japan，2005

[8] Xiong Xiaopeng，Mokbel M F，Aref W G. SEA-CNN：Scalable Processing of Continuous K-Nearest Neighbor Queries in Spatio-temporal Databases[C]// Proc. of the 21st Intl. Conf. on Data Engineering. Tokyo，Japan，2005

[9] Mouratidis K，Hadjieleftheriou M，Papadias D. Conceptual Partitioning：An Efficient Method for Continuous Nearest Neighbor Monitoring[C]// Proc. of the 2005 Intl. Conf. on Management of Data. Baltimore，Maryland，2005

[10] Mouratidis K，Yiu M L，Papadias D，et al. Continuous Nearest Neighbor Monitoring in Road Networks[C]// Proc. of 32nd Intl. Conf. on Very Large Data Bases. Seoul，Korea，2006

[11] Wang Haojun，Zimmermann R. Location-based Query Processing on Moving Objects in Road Networks[C]// Proc. of 33rd Intl. Conf. on Very Large Data Bases. Vienna，Austria，2007

[12] 陈继东，胡志智，孟小峰，等. 一种基于城市交通网络的移动对象全失态索引[J]. 计算机研究与发展，2007，44(6)：1008-1014

[13] Brinkhoff T. A Framework for Generating Network Based Moving Objects[J]. GeoInformatica，2002，(6)2：153-180