ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.11, November 2008, pp.2814−2821
DOI: 10.3724/SP.J.1001.2008.02814

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种半监督*K*均值多关系数据聚类算法*

高 滢[1,2+], 刘大有[1,2], 齐 红[1,2], 刘 赫[1,2]

[1](吉林大学 计算机科学与技术学院,吉林 长春 130012)

[2](吉林大学 符号计算与知识工程教育部重点实验室,吉林 长春 130012)

Semi-Supervised *K*-Means Clustering Algorithm for Multi-Type Relational Data

GAO Ying[1,2+], LIU Da-You[1,2], QI Hong[1,2], LIU He[1,2]

[1](College of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](Key Laboratory of Symbolic Computation and Knowledge Engineering for the Ministry of Education, Jilin University, Changchun 130012, China)

+ Corresponding author: E-mail: gying@jlu.edu.cn

Gao Y, Liu DY, Qi H, Liu H. Semi-Supervised *K*-means clustering algorithm for multi-type relational data. *Journal of Software*, 2008,19(11):2814−2821. http://www.jos.org.cn/1000-9825/19/2814.htm

Abstract: A semi-supervised *K*-means clustering algorithm for multi-type relational data is proposed, which extends traditional *K*-means clustering by new methods of selecting initial clusters and similarity measures, so that it can semi-supervise cluster multi-type relational data. In order to achieve high performance, in the algorithm, besides attribute information, both labeled data and relationship information are employed. Experimental results on Movie database show the effectiveness of this method.

Key words: data mining; semi-supervised learning; clustering algorithm; multi-type relational data; *K*-means clustering

摘 要: 提出了一种半监督*K*均值多关系数据聚类算法.该算法在*K*均值聚类算法的基础上扩展了其初始类簇的选择方法和对象相似性度量方法,以用于多关系数据的半监督学习.为了获取高性能,该算法在聚类过程中充分利用了标记数据、对象属性及各种关系信息.多关系数据库 Movie 上的实验结果验证了该算法的有效性.

关键词: 数据挖掘;半监督学习;聚类算法;多关系数据;*K*均值聚类

中图法分类号: TP181 **文献标识码**: A

传统的数据挖掘任务,例如关联规则挖掘、市场购物篮分析、聚类分析等通常假定数据由同种类型、相互独立的实体构成,且实体间互不关联.而现实世界中的许多数据却是多关系的,即数据由多种不同类型的实体组成,各类实体属性不尽相同,且实体间通过多种关系相互关联[1].多关系数据在生物信息学、Web 导航、社会网、

* Supported by the National Natural Science Foundation of China under Grant Nos.60496321, 60773099, 60573073 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant Nos.2006AA10Z245, 2006AA10A309 (国家高技术研究发展计划(863)); the Science and Technology Development Plan of Jilin Province of China under Grant No.20030523 (吉林省科技发展计划); the European Commission under Grant No.TH/Asia Link/010 (111084) (欧盟项目)

Received 2008-02-19; Accepted 2008-08-26

知识获取与利用、地理信息系统和自然语言理解等领域广泛存在[2,3].忽略数据的多关系特性,而仍然使用传统的数据挖掘方法对数据进行挖掘将得到不准确的结论.因此,为了提高挖掘结果的准确性,数据的多关系特性必须进行适当的处理.

在数据挖掘任务中,由于大量标记数据的获取需要昂贵的代价,实际数据往往由大量无标记数据和少量标记数据组成.半监督学习是处理此类数据的一种学习方法,近年来受到众多研究者的关注[4].目前,许多研究者已提出多种半监督学习方法.从学习方式上看,现有半监督学习方法可分为半监督分类和半监督聚类两大类[5].半监督分类是在有监督分类的基础上,通过无标记数据指导分类过程,以提高分类的准确性;半监督聚类是在无监督聚类的基础上,通过标记数据指导聚类过程,以提高聚类质量.然而,无论哪类学习方法,针对多关系数据的研究都很少.

本文研究了用于处理多关系数据的半监督聚类方法.在传统 K 均值聚类算法的基础上,提出了半监督 K 均值多关系数据聚类算法(semi-supervised K-means clustering algorithm for multi-type relational data,简称 SKMMR).该算法在聚类过程中充分利用标记数据、对象属性信息和各种关系信息,以提高聚类准确性.算法具有如下特征:

- 深入分析对象中所包含的各种关系,包括类内关系、类间关系、显式关系、隐式关系;
- 通过新的初始类簇选择方法扩展传统 K 均值聚类算法,使用标记数据指导初始类簇的选择过程;
- 通过新的相似性度量方法扩展传统 K 均值聚类算法,使对象间的相似度不仅包括对象属性信息,还包括各种关系信息.

本文第 1 节给出半监督多关系数据聚类的问题定义及符号表示.第 2 节介绍半监督 K 均值多关系数据聚类算法.第 3 节给出多关系数据集 Movie 上的实验结果.第 4 节总结全文.

1 半监督多关系数据聚类的问题定义

1.1 问题定义及符号表示

在半监督学习[6]和多关系数据聚类[7]的问题定义的基础上,本文给出半监督多关系数据聚类的问题定义.

在半监督多关系数据聚类问题中,已知多种类型的对象及其间关系的集合,其中一些为标记对象,其余为未标记对象.令 X 表示对象集合,X^L 表示 X 中的标记对象,R 表示 X 中对象间关系,则:

$X=\{X_1,\ldots,X_m\}$,且 $X_1=\{x_{11},x_{12},\ldots,x_{1n_1}\},\ldots,X_m=\{x_{m1},x_{m2},\ldots,x_{mn_m}\}$,表示 X 中存在 m 种不同类型的对象,且每类对象的数目为 $n_i(1\le i\le m)$;每类对象 X_i 有其自身的属性,记为 $X_i.A$.

X 中存在部分标记对象,记为 $X^L=\{X_1^L,X_2^L,\ldots,X_m^L\}$,其中,$X_i^L=\{(x_{i1},y_{i1}),(x_{i2},y_{i2}),\ldots,(x_{il_i},y_{il_i})\}$ 为 X_i 中的标记对象,这里,l_i 是 X_i 中标记对象数.

X 中对象间的关系 $R=\{R_{intra},R_{inter}\}$,其中,$R_{intra}=\{R_1,\ldots,R_m\}$表示类内关系,即同种类型对象间的关系,$R_{inter}$ 表示类间关系,即不同类型对象间的关系.在所有关系中,有些关系能够通过数据库的关系模式直接获得,称为显式关系;有些关系需要对数据库对象深入分析才能得到,称为隐式关系.

给定上述 X,X^L 和 R,在不违反 X^L 中给定标记的情况下,将各类对象 X_i 自动划分成 K_i 个组的过程称为半监督多关系数据聚类.

1.2 举 例

图 1 是一个多关系数据的例子,该例是 UCI Movie 数据集的一部分.图 1 中共含有 4 种类型的对象,即 X={{Movie},{Actor},{Director/Producer},{Studio}};每种类型对象均有各自的属性,如 title,category,locale 等是 Movie 对象的属性,name,gender 等是 Actor 对象的属性;每种类型的对象中都有部分标记对象,在图 1 中标记对象未被标出;对象间的关系分为类内关系和类间关系,如 Actor 和 Director/Producer 内的 co-work 关系是类内关系,has Actor,directedBy 等是类间关系;在所有关系中,普通箭头表示显式关系,虚箭头表示隐式关系.

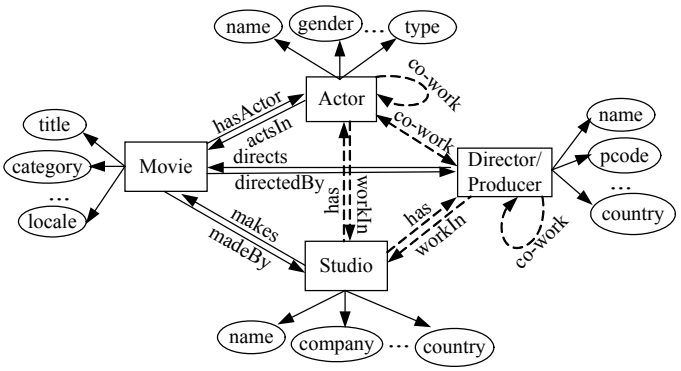


Fig.1 An example of multi-type relational data

图 1 多关系数据示例

2 半监督*K*均值多关系数据聚类算法

K 均值聚类[8]是将数据集自动划分为 *K* 个组的一种常用方法.在选择 *K* 个初始聚类中心以后,*K* 均值聚类通过迭代执行如下步骤来实现:

(1) 将每个对象 x_i 分配到与之最近的聚类中心;

(2) 根据当前属于每个聚类中心的对象,更新各聚类中心 c_j.

在为对象分配新聚类中心时,若所有对象所属的新聚类中心均与原来一致,则算法因收敛而结束.

本文在 *K* 均值聚类算法的基础上,通过扩展初始聚类中心(初始类簇)的选择方法及相似性度量方法,使之用于多关系数据的半监督学习.扩展之后的新算法称为半监督 *K* 均值多关系数据聚类算法(SKMMR).

2.1 SKMMR

SKMMR 通过标记数据指导初始类簇的选择,通过充分利用对象属性及各种关系信息,重新定义多关系数据中对象间相似度,改进了原始 *K* 均值聚类算法,使之用于多关系数据的半监督学习,其过程描述见算法 1.

算法 1. *SKMMR*(X,X^L,R.C).

//输入:m 类数据对象 $X=\{X_1,\ldots,X_m\}$,X 中的标记对象 $X^L=\{X_1^L,X_2^L,\ldots,X_m^L\}$, 对象间关系 $R=\{R_{intra},R_{inter}\}$.

//输出:聚类结果 $C=\{C_1,C_2,\ldots,C_m\}$,其中,$C_i=\{c_{i1},c_{i2},\ldots,c_{iK_i}\}$,$1\le i\le m$.

```
BEGIN
1. for i=1 to m do
2.   {C_i←SelectInitialClusters(X_i, X_i^L);
3.    do {
4.       for each x_ij∈X_i
5.          if (x_ij∈ X_i^L) AssignToLabeledCluster(x_ij,C_i);
6.          else AssignToClosestCluster(x_ij,C_i);
7.       UpdateClusters(C_i);
8.    } until converges;
9.   }
END
```

算法 1 的第 2 行是从数据 X_i 中选择 K_i 个初始类簇的过程.该过程是在标记数据 X_i^L 的指导下完成的,详细描述见算法 2.算法 1 的第 5 行和第 6 行,首先判断对象 x_{ij} 是否为标记对象,若是,则根据其标记 y_{ij},将其直接分配到 y_{ij} 所对应的类簇;若对象 x_{ij} 为未标记对象,则将该对象分配到与之最近的类簇.该过程使用了新的相似性度量

方法,其具体描述见第 2.2 节.算法 1 的第 7 行,更新各类簇.由于算法经过迭代后,对象所属的类簇可能会发生变化,该步记录此时每个类簇所包含的对象情况.

算法 2. *SelectInitialClusters*(X_i, X_i^L.C_i).

//输入:第 i 类对象 $X_i=\{x_{i1},x_{i2},...,x_{in_i}\}$, X_i 中标记对象 $X_i^L=\{(x_{i1},y_{i1}),(x_{i2},y_{i2}),...,(x_{il_i},y_{il_i})\}$.

//输出:初始类簇 $C_i=\{c_{i1},c_{i2},...,c_{iK_i}\}$.

BEGIN

1. K←*NumberOfDistinctLabels*(X_i^L);
2. If (K>K_i)
3. Error! The data is not consistent, return.
4. If (K≤K_i)
5. Select K objects with different labels from $\{x_{i1},x_{i2},...,x_{il_i}\}$ and (K_i–K) objects from $\{x_{i(l_i+1)},...,x_{in_i}\}$ as initial clusters.

END

算法 2 是在标记数据的指导下,从 X_i 中选择 K_i 个初始类簇的过程.该过程首先统计 X_i^L 中包含的不同标记的数目 K,若 K>K_i,则按照问题定义,在不违反 X_i^L 的情况下,将 X_i 划分为 K_i 个类簇是不可能的,此时,算法因数据不一致而结束;若 K≤K_i,则从标记数据中随机选择具有不同标记的 K 个数据对象,再从未标记数据中随机选择其余(K_i–K)个对象,从而完成初始类簇的选择.

2.2 相似性度量

计算两个对象之间、对象与类簇之间的相似度在聚类中至关重要.

设 x_{ij},x_{ik} 是 X_i 类的两个对象,考虑到对象具有属性特征、类内和类间关系特征,本文定义 x_{ij} 与 x_{ik} 的相似度为

$$Sim(x_{ij},x_{ik})=\alpha\cdot Sim_A(x_{ij},x_{ik})+\beta\cdot Sim_{intra}(x_{ij},x_{ik})+\gamma\cdot Sim_{inter}(x_{ij},x_{ik}) \tag{1}$$

其中,Sim_A 是属性相似度;Sim_{intra} 是类内关系相似度;Sim_{inter} 是类间关系相似度;α,β和γ是各相似度的权重,满足 α+β+γ=1.α,β和γ的具体取值与数据集相关,可由专家指定或通过实验的方法来确定.

属性可分为数值属性和类别属性.数值属性一般需要标准化处理,即将属性的取值范围映射到同一个区间(如[0,1])内,从而消除数据量纲及属性域的影响.类别属性可分为有序类别属性和无序类别属性.对于有序类别属性,可将其转换为数值属性,从而采用与数值属性相同的方式来处理;对于无序类别属性,仅考虑属性值的异同即可.本文把对象 x_{ij},x_{ik} 的属性相似度 Sim_A 定义为

$$Sim_A(x_{ij},x_{ik})=e^{-\|x_{ij}.A-x_{ik}.A\|} \tag{2}$$

其中,‖ $x_{ij}.A$−$x_{ik}.A$ ‖ 表示对象 x_{ij},x_{ik} 之间的属性距离,令其为对象各属性的差异之和,即

$$\|x_{ij}.A-x_{ik}.A\|=\sum_{r=1}^{p}|x_{ij}.A_r-x_{ik}.A_r|+\lambda\sum_{r=p+1}^{N}\delta(x_{ij}.A_r,x_{ik}.A_r) \tag{3}$$

这里,$x_{ij}.A_r$ 和 $x_{ik}.A_r$ 为对象 x_{ij} 和 x_{ik} 的第 r 个属性;N 是 X_i 类对象的属性总数,并假定从第 1 到第 p(p≤N)个属性为经过标准化的数值属性或有序类别属性,从第(p+1)到第 N 个属性为无序类别属性;函数 $\delta(a,b)$为差异函数,当 a 与 b 相等时,$\delta(a,b)$值为 0,否则为 1;参数 λ用来调节无序类别属性相对于数值属性在属性相似度中所占的比重,λ 取值范围为[0,1].

Sim_{intra} 是类内关系相似度.如果两个对象存在类内关系,则 Sim_{intra} 值为 1,否则为 0,即

$$Sim_{intra}(x_{ij},x_{ik})=\begin{cases}1, & intra\text{-}relationship(x_{ij},x_{ik})\\ 0, & \neg intra\text{-}relationship(x_{ij},x_{ik})\end{cases} \tag{4}$$

intra-relationship(x_{ij},x_{ik})用来判断对象 x_{ij},x_{ik} 是否存在类内关系.

Sim_{inter} 是类间关系相似度.任一对象可能与其他多个类的对象存在类间关系,设 X_{inter} 为与 X_i 内对象 x_{ij},x_{ik}

存在类间关系的类的集合,则 x_{ij},x_{ik} 的类间关系相似度应与 X_{inter} 中元素数目成正比.对于任意元素 $X_p \in X_{inter}$ $(1\le p\le m$,且 $p\ne i)$,x_{ij},x_{ik} 的类间关系相似度与 X_p 内同时和两对象有关系的对象数占和其中一个对象具有关系的对象数的比例相关,且与该比例成正比.为了保证类间关系相似度的取值范围为[0,1],本文定义其为

$$Sim_{inter}(x_{ij},x_{ik})=1-\mathrm{e}^{-\sum_{X_p\in X_{inter}}\frac{|inter\text{-}relationship(x_{ij},X_p)\cap inter\text{-}relationship(x_{ik},X_p)|}{|inter\text{-}relationship(x_{ij},X_p)\cup inter\text{-}relationship(x_{ik},X_p)|}} \tag{5}$$

这里,$inter\text{-}relationship(x_{ij},X_p)$表示类 X_p 内与对象 x_{ij} 具有类间关系的对象集合,| |表示集合的势.

对于对象与类簇的相似度,本文定义其为该对象与属于该类簇的所有对象的平均相似度,即

$$Sim(x_{ij},c_{ip})=\underset{x_{iq}\in c_{ip}}{avg}\,(Sim(x_{ij},x_{iq})) \tag{6}$$

3 实 验

3.1 实验数据及其关系分析

本文在 UCI 的多关系数据库 Movie[9]上进行实验.Movie 数据库的关系模式如图 2 所示.从关系模式中能够直接获得显式关系,如{Movie}与{Director/Producer}的类间关系,{Movie}与{Studio}的类间关系.通过关系表 Cast 能够获得{Movie}与{Actor}的类间关系;通过分析数据库对象,可以进一步获得隐式关系,如{Actor},{Studio}与{Director/Producer}两两之间的关系、{Actor}和{Director/Producer}的类内关系.分析关系模式及数据库内容之后,所得关系如图 3 所示.由此,Movie 数据库中所含对象及关系数见表 1.

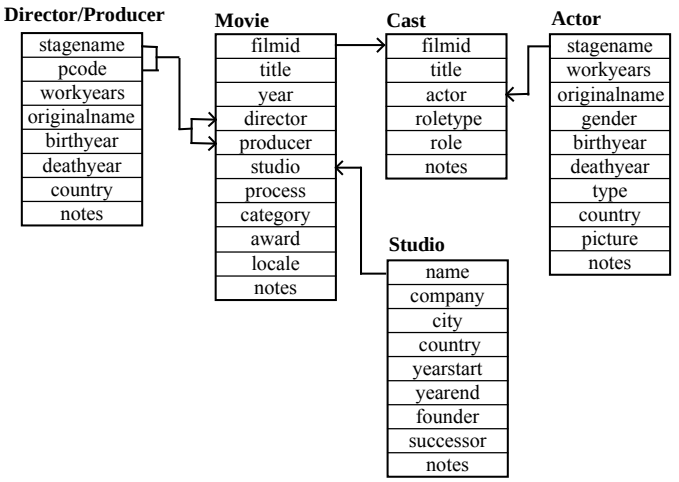


Fig.2 Relational schema of Movie database

图 2 Movie 数据库的关系模式

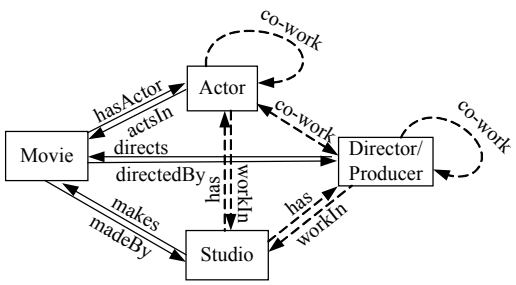


Fig.3 All kinds of relationships in Movie database

图 3 Movie 数据库中的各种关系

Table 1 Number of objects and relationships in Movie database

表1 Movie数据库中的对象及关系数

Type	Name	Number
Objects	Movie	14 067
	Actor	6 822
	Director/Producer	3 311
	Studio	194
Inter-Relationships	Movie-Actor	28 927
	Movie-Director/Producer	15 191
	Movie-Studio	5 892
	Actor-Director/Producer	28 965
	Actor-Studio	9 501
	Director/Producer-Studio	1 981
Intra-Relationships	Actor: co-work	68 621
	Director/Producer: co-work	960

3.2 聚类结果评价

聚类结果常用聚类准确性来评价.Jaccard系数是评价聚类准确性的常用方法.本文使用文献[10]中的Jaccard系数变种方法,令聚类准确性为各类簇内正确聚类的对象对数目之和与各类簇内对象对总数之和的比值.对X_i类对象聚类结果C_i,使用该方法计算的聚类准确性JC为

$$JC=\frac{\sum_{c_{ip}\in C_i} Num((x_{ij},x_{ik})\mid x_{ij}\in c_{ip},x_{ik}\in c_{ip},x_{ij}\neq x_{ik},\ \text{and}\ sameLabel(x_{ij},x_{ik}))}{\sum_{c_{ip}\in C_i} Num((x_{ij},x_{ik})\mid x_{ij}\in c_{ip},x_{ik}\in c_{ip},\ \text{and}\ x_{ij}\neq x_{ik})} \tag{7}$$

其中,函数 $Num()$用于求满足条件的对象对数目,$sameLabel()$用于判断两对象标记是否相同.由于类簇数越大,该方法越趋于获得更高的准确性,在对同种类型对象聚类的多次实验中,本文令其类簇数一致.

另外,本文亦使用类似于文献[11]中的基于熵的方法来评价聚类结果.给定类簇 $c_{ip}\in C_i$,则 c_{ip} 的熵为

$$H(c_{ip})=-\sum_h P_h\log_2 P_h.$$

这里,P_h 为具有 h 标记的对象在类簇 c_{ip} 中所占的比例.整个聚类结果 C_i 的熵为 C_i 所包含的各类簇的熵之和.为了使不同类型对象的聚类结果具有可比性,不同于文献[11]直接将聚类结果的熵作为评价标准,本文使用平均熵来评价聚类结果.聚类结果 C_i 的平均熵为

$$\overline{H}(C_i)=\frac{\sum_{c_{ip}\in C_i} H(c_{ip})}{|C_i|} \tag{8}$$

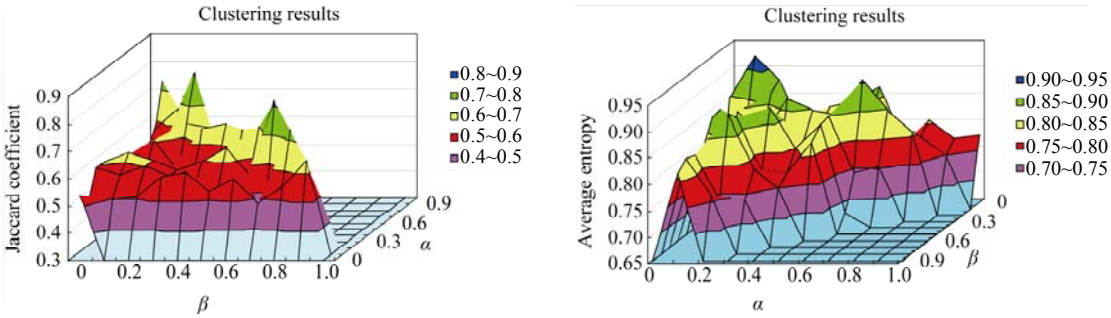
基于平均熵的评价方法,平均熵越小,表明聚类质量越高.

3.3 实验结果

为了说明属性、类内及类间关系信息在聚类过程中的作用,本文针对相似度公式(1)中 3 种不同类型的相似度权重参数α,β和γ的多种不同取值进行了实验.图 4(a)、图 4(b)给出了数据集中不含有标记数据,且当α,β和γ(γ=1−α−β)取各种不同值时,分别以公式(7)、公式(8)作为聚类结果评价标准对 Movie 数据集的聚类结果.由图 4 可知,当权重参数设置合理时,同时使用 3 种类型的相似度对数据聚类,聚类结果好于仅使用其中某一种类型的相似度的聚类结果(当α=1,β=0,γ=0 时,是仅使用属性相似度的聚类结果;当α=0,β=1,γ=0 时,是仅使用类内关系相似度的聚类结果;当α=0,β=0,γ=1 时,是仅使用类间关系相似度的聚类结果).另外,由图 4 也可以看出聚类质量较好时的权重参数取值情况.

为了进一步验证本文中的多关系数据对象之间相似性度量的合理性以及算法的有效性,我们将算法 SKMMR 与多关系数据聚类算法 DIVA[11]和 ReCoM[12]进行了实验比较.先将算法 DIVA 和 ReCoM 中的相似性度量方法应用到 SKMMR 中,并与基于本文相似性度量的 SKMMR,对聚类结果比较;又将算法 DIVA 和 ReCoM 对 Movie 数据库的聚类结果与算法 SKMMR 进行比较.为了说明半监督学习特性对聚类结果的影响,本文给出了标记数据占整个数据集不同比例时的聚类结果.实验中,从图 4 聚类结果较好的区域随机选择相似度权重参

数值,图 5(a)、图 5(b)给出了当α=0.6,β=0.1 和γ=0.3 时的实验结果.



(a) Jaccard coefficient of clustering results

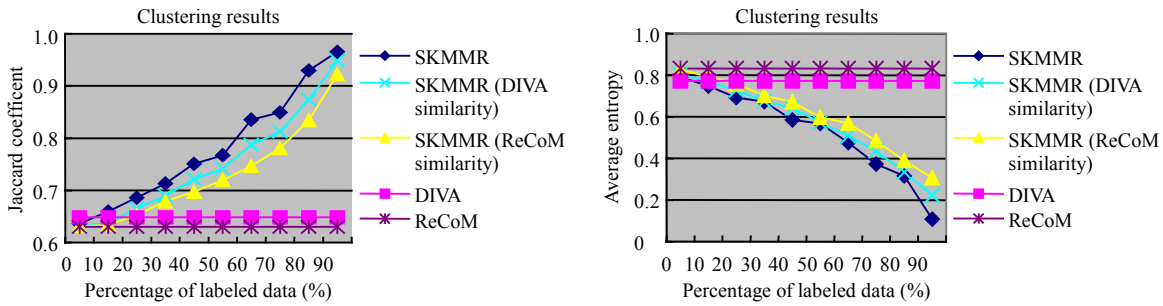
(a) 聚类结果的 Jaccard 系数

(b) Average entropy of clustering results

(b) 聚类结果的平均熵

Fig.4 Clustering results on Movie database with different α, β and γ

图4 α,β和γ取不同值时Movie数据集的聚类结果



(a) Jaccard coefficent of clustering results

(a) 聚类结果的 Jaccard 系数

(b) Average entropy of clustering results

(b) 聚类结果的平均熵

Fig.5 Clustering results with different percentage of labeled data

图5 标记数据占整个数据集不同比例时的聚类结果

由图 5 可知,对于算法 SKMMR,采用本文的多关系数据对象之间的相似性度量方法,聚类结果好于采用算法 DIVA 和 ReCoM 中的相似性度量方法.当标记数据为 0%时,SKMMR 为无监督算法,其聚类结果好于 ReCom,略逊于 DIVA;在给定部分标记数据的情况下,算法 SKMMR 的聚类结果明显好于 DIVA 和 ReCoM 算法;随着标记数据所占比例的增加,半监督算法的聚类效果随之提高.可见,标记数据的使用,有利于提高聚类质量.

4 结 论

为了将半监督学习方法用于多关系数据,本文在原始 K 均值聚类算法的基础上提出了半监督 K 均值多关系数据聚类算法(SKMMR).SKMMR 在深入分析对象及其间关系的基础上,通过扩展 K 均值聚类算法的初始类簇的选择方法及相似性度量方法,将标记数据、对象属性及各种关系信息应用到聚类过程中,以提高聚类质量.在多关系数据库 Movie 上的实验结果验证了该算法的有效性.

References:

[1] Džeroski S. Multi-Relational data mining: An introduction. ACM SIGKDD Explorations Newsletter, 2003,5(1):1−16.

[2] Džeroski S, Lavrač N. Relational Data Mining. Berlin: Springer-Verlag, 2001. 339−364.

[3] Domingos P. Prospects and challenges for multi-relational data mining. ACM SIGKDD Explorations Newsletter, 2003,5(1):80−83.

[4] Bouchachia A. Learning with partly labeled data. Neural Computing and Applications, 2007,16(3):267−293.

[5] Zhu XJ. Semi-Supervised learning literature survey. Technical Report, Computer Sciences TR 1530, University of Wisconsin-Madison, 2007. 1−42.

[6] Chapelle O, Scholkopf B, Zien A. Semi-Supervised Learning. Cambridge: MIT Press, 2006. 3−14.

[7] Long B, Zhang F, Wu XY, Yu PS. Spectral clustering for multi-type relational data. In: Cohen WW, Moore A, eds. Proc. of the 23rd Int'l Conf. on Machine Learning. New York: ACM Press, 2006. 585−592.

[8] Marques de Sá JP, Wrote; Wu YF, Trans. Pattern Recognition Concepts, Methods and Applications. 2nd ed., Beijing: Tsinghua University Press, 2002. 51−74 (in Chinese).

[9] http://archive.ics.uci.edu/ml/datasets.html

[10] Yin XX, Han JW, Yu PS. CrossClus: User-Guided multi-relational clustering. Data Mining Knowledge Discovery, 2007,15(3): 321−348.

[11] Li T, Anand SS. DIVA: A variance-based clustering approach for multi-type relational data. In: Silva MJ, Laender AHF, eds. Proc. of the 16th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2007. 147−156.

[12] Wang JD, Zeng HJ, Chen Z, Lu HJ, Tao L, Ma WY. ReCoM: Reinforcement clustering of multi-type interrelated data objects. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2003. 274−281.

附中文参考文献:

[8] Marques de Sá JP,著;吴逸飞,译.模式识别——原理、方法及应用.北京:清华大学出版社,2002.51−74.

高滢(1978－),女,吉林长春人,博士生,讲师,主要研究领域为数据挖掘,统计关系学习.

齐红(1970－),女,博士,副教授,主要研究领域为数据挖掘,统计关系学习.

刘大有(1942－),男,教授,博士生导师,CCF 高级会员,主要研究领域为知识工程,专家系统与不确定性推理,时空推理,分布式人工智能,多 Agent 和移动 Agent 系统,数据挖掘,多关系数据挖掘,数据结构,计算机算法.

刘赫(1980－),男,博士生,主要研究领域为数据挖掘.