基于信息源聚类的最大熵加权信任分析算法

侯 森,罗兴国,宋 克
(中国人民解放军信息工程大学信息技术研究所,河南郑州 450002)

摘 要: 在信息网络中,不同的信息源以不同的可信性和准确性提供了各式各样的信息.为了预测这些信息反映事实的真实度,学者们提出了一些信任分析算法来迭代地计算信息源的信任度及其提供事实的准确度.然而这些算法往往忽略了信息源和事实描述对象之间的相关性.本文作者提出了一种基于信息源聚类的最大熵加权信任分析算法,该算法使我们能够在进行信任分析时有效地融合诸如描述对象属性、信息源关联性等信息.实验证明该算法能够明显的提高分析性能.

关键词: 信息网络;最大熵;信任分析;聚类
中图分类号: TP14 **文献标识码:** A **文章编号:** 0372-2112 (2015)05-0993-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.05.024

A Maximum Entropy Weighted Trust-Analysis Algorithm Based on Sources Clustering

HOU Sen, LUO Xing-guo, SONG Ke
(*National Digital Switching System Engineering and Technology Research Center, Zhengzhou, Henan* 450002, *China*)

Abstract: In information network, different sources publish facts with different degrees of credibility and accuracy. To predict the truth values of the facts, several fact-finder algorithms are suggested which iteratively compute the trustworthiness of an information sourceand the accuracy of the facts it provides. However they ignore a great deal of relevant background and contextual information. In this paper, we proposed a novel maximum entropy weighted method to processing trust analysis, allowing us to elegantly incorporate knowledge such as the attributes of the objects and the implications of the sources. Experiments demonstrate that our algorithmsignificantly improves performance over existing.

Key words: information gain; maximum entropy; trust analysis; clustering

1 引言

信息网络(Information Network)是一种用以描述网络实体复杂关系的抽象概念[1].在信息网络中存在着若干可信度未知的信息源,它们以不同程度的可靠性和准确性为我们提供了大量的事实,而我们可以通过各种方法预测事实的真实度,从而发现真理性知识.信息网络中的信任分析(Trust Analysis)正在被越来越多的学者和研究人员认识和重视.文献[2]中定义信任是一种衡量不确定性的测度标准.事实发现(Fact-finder)则是代表目前信任分析最新研究水平的一种算法,它能够有效地预测事实的可信任程度.事实发现依赖于一种基于图的拓扑分析方法,其中节点代表着信息源或者事实等信息网络实体,而边则代表着实体间的抽象关系.在从大量低质量输入数据和观察结果中提取有用信息的过程中,即使在信息源可靠程度未知的情况下,事实发现算法也能够表现出很高的效率.

近年来国内外学者做了不少关于信息网络中信任分析和事实发现的工作.早期比较有影响力的研究有PageRank[3]算法和Authority-Hub[4]算法,PageRank和Authority-Hub都可以用来对Web站点进行分级排名.Trust-finder算法[5]是第一个无监督信任分析方法,该算法由Yin等人于2008年发表,该算法可以进行异构网络的事实发现.Trust-finder算法认为真实事实趋于相近,而错误事实往往彼此背离.同时Yin等人认为,如果一个信息源一直提供高准确度的事实,则它在很大概率上会继续提供真实事实.基于上述假设,Trust-finder算法迭代的计算所有信息源的信任度和事实的置信度.Alban Galland、Jeff Pasternack等人陆续提出了很多新的算法,在信任分析过程中引入了更多的参数,主要有:Cosine算法、2Estimate算法、3Estimate算法[6]、Average.Log算法[7]、Investment算法以及PooledInvestment算法[8,9]等.

收稿日期:2013-05-10;修回日期:2014-06-16;责任编辑:蓝红杰
基金项目:国家863高技术研究发展计划(No.2009AA012201);上海市科委重大科技攻关项目(No.08dz501600)

然而上述算法都没有考虑到信息源和事实描述对象之间的关系,本文作者将提出一种新的基于信息源聚类的最大熵加权信任分析算法,该算法使我们能够更准确的评估信息源在不同知识领域对事实真实性的影响.我们第一次将信息增益的概念引入到信任分析过程.通过最大熵加权,我们能够能精确的融合描述对象属性和事实关联度等特性.

2 事实发现算法

我们首先介绍一下事实发现算法的基本设定以及在众多事实发现算法中使用的迭代逻辑基础.假定存在大量相互冲突的事实,而这些事实是由大量不同的信息源提供,那么我们该如何从中发现关于描述对象的真实的事实.Trust-finder 算法是第一个无监督事实发现算法,它是 2008 年由 Yin 等人提出.它基于信息源-事实信息网络架构,给出了迭代的计算信息源可信度和事实置信度的计算方法.

图 1 是一个典型的"信息源-事实"信息网络架构示意图.假设有信息源集合 S,集合中的每一个元素都提供了一组事实.信息源 s 的事实集合用 F_s 来表示,事实空间为 $F=\bigcup_{s\in S}F_S$.每一个事实属于一个事实集合 $M_f\subseteq F$,该集合中的元素彼此互斥.

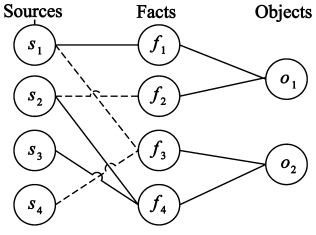

图1 信息源、事实信息网络结构

2.1 基本定义

定义 1 信息源的可信度(Trustworthiness of source,表示符号 $T(s)$),即信息源 s 提供的所有事实的置信程度.

定义 2 事实的置信度(Confidence of fact,表示符号 $C(f)$),即所有提供了事实 f 的信息源的可信程度.

定义 3 事实之间的蕴含度(Implications of facts,表示符号 $\mathrm{imp}(f_1\rightarrow f_2)$,表示事实 f_1 对事实 f_2 的影响程度.

假设在每一个互斥集合 M_f 中只有一个正确的事实,事实发现的目的就是从集合中预测出正确事实来.为此,事实发现算法迭代地计算信息源的可信度和事实的置信度.第 i 次迭代后,事实 f 的置信度 $C^i(f)$ 由所有提供该事实的信息源上一次迭代中的可信度 $T^{i-1}(S_f)$ 表示,信息源 s 的可信度 $T^i(s)$ 则由它本身提供的所有事实当次迭代中得到的置信度 $C^i(F_s)$ 来表示.由于真实的事实比错误事实更具有一致性,因而我们最终能够从错误事实中辨别出正确事实来.

2.2 计算模型

2.2.1 基本算法模型

首先,根据上述定义我们可以得到事实发现算法的基本形式:

$$C(f)=\sum_{s\in S_f}T(s) \tag{1}$$

$$T(s)=\frac{\sum_{f\in F_s}C(f)}{|F_s|} \tag{2}$$

基本算法根据已知的信息网络实体的真实程度相加求平均来估计信息源和事实的信任水平.PageRank 算法和 Authority-Hub 算法都是基于这种相加平均的逻辑.PageRank 算法可以看做是一种基于声誉的推荐系统,Authority-Hub 算法则采用了一种基于 Authority 和 Hub 的二分法.基本算法并没有考虑信息源和事实之间的关联关系.

2.2.2 Trust-finder 算法

第一个无监督事实发现算法是 2008 年由 Yin 等人提出 Trust-finder 算法,作者研究了如何从冲突信息中寻找真实事实的方法.Trust-finder 算法的输入是不同站点提供的大量的事实,不同于 PageRank 算法和 Authority-Hub 算法,作者假定这些事实通常是相互冲突的,而且信息源和事实均可以是异构.Trust-finder 算法通过信息源和事实之间的迭代关系计算两者的信任水平,继而从中辨别出真实事实.

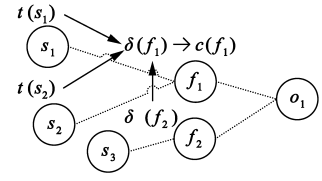

图2 加入事实之间关联性的事实置信度示意图

Trust-finder 算法采用了与式(2)相同的方式来计算信息源的可信度.而对于事实的置信度,作者认为不仅是由提供它的信息源决定,而且还和与它描述同一对象的其他事实之间有着关联关系,即描述同一对象的事实之间有着相互印证或否定的关系.因而我们可以根据图 2 重写事实置信度的计算公式如下

$$C(f)=1-\prod_{s_i\in S_f}(1-T(s_i))+\rho\sum_{f'\in O_f,f'\neq f}\delta(f')\cdot\mathrm{imp}(f'-f) \tag{3}$$

其中 O_f 表示与事实 f 描述同一对象的事实集合.

近年来不少学者做了大量关于信息网络中信任分析的研究,提出了许多事实发现的改进算法,这些算法大都基于与 Trust-finder 算法相同的网络结构,只是在信息源可信度和事实置信度的计算方式上有所不同.

2.3 现有算法的不足之处

上述事实发现算法都是基于如下假设,即所有信息源在不同描述领域的重要性都相同.而且上述所有的事实发现算法都假定对于描述对象的单个属性有且只有唯一的正确事实.然而这些假设具有如下不足:

·领域专家问题:不同的信息源可能在某些领域具有较高的可信度,但是在其他方面并不擅长.比较常见的例子是一个电影评论网站可能在描述一部电影时具有良好表现,但是在描述科学类、时政类话题时并不具有很高的可信度.

·一致性学习问题:即假定描述某个对象的特定属性的事实都是同一类型的.然而在真实的信息网络中,描述同一对象的事实可以是多种形式的,信息源能够以不同的概率提供不同类型的事实,描述同一对象的相同属性.比如电影评论网站描述一部电影好坏的方式可以是一篇文字形式的影评,也可以是数字形式的评分,或者是一段视频类型的访谈.如何从多种类型的事实中一致的学习到关于对象的相关知识将是会一个非常实用的研究.

·多值真相问题:即假定对于某个描述对象的特定属性有且只有唯一的真实事实.但是往往很多事实都只是在一定程度上反应了真相,真实的事实并不具有一个确定的值.而且有时甚至在互斥的事实集合中没有一个能够正确反映对象的真实属性.

通过对事实发现算法模型的分析,我们引入信息熵的概念利用信息源聚类改进传统算法.

3 基于知识领域的信息源聚类

根据"信息源-事实"信息网络架构,我们发现传统的事实发现算法都没有有效地利用描述对象的相关知识.描述对象可以是各种类型的事物,可以是人、物,甚至是某种抽象的关系,相似的描述对象可以被划分到相同的分组,相关联的分组又可以构成特定的"知识领域"."知识领域"反映了从信息源到描述对象的关联性.如果能够根据信息源各自擅长的知识领域的不同而将其划分到不同的聚类,则我们有可能得到一种更符合实际而且更准确高效的学习方法.我们将在相同知识领域表现优异的信息源集合称为"类",将若干信息源划分到不同类的过程叫做"信息源聚类".

3.1 知识领域

首先我们将图 1 中的网络拓扑扩展到一种如图 3 所示的更为复杂的形式.

图 3 中,我们根据信息源和事实描述的对象将整个网络分为了若干知识领域.每个知识领域可以涵盖一个或者几个对象及其相关属性.假设共有 k 个对象 $\{o_1, o_2, \cdots, o_k\}$,则我们可以用 $g_c = \{g_{c1}, g_{c2}, \cdots, g_{ck}\}$ 来表示一个知识领域,其中 g_{ck} 表示该知识领域包含对象 o_k 的概率.

我们在扩展的拓扑结构中还考虑了信息源和事实的多样性.即我们假设信息源可以以不同的概率做出关于同一对象的不同类型的事实论断,而事实通过概率形式不同程度地反映出对象属性.这样的假设更贴近于真实网络,同时通过对不同类型的事实赋予概率的形式能够有效解决一致性学习问题,而概率概念本身也有助于处理多值真相问题.这样的结构较好地模拟了真实网络中不同 Web 网站在不同领域各有专攻的情况.

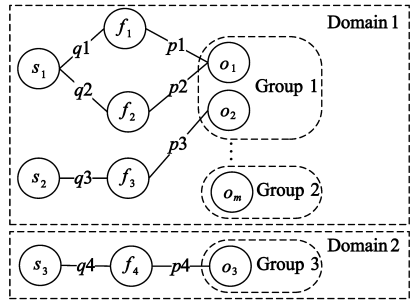

图3　知识领域和对象分组

3.2 基于知识领域的信息源聚类方法

在同一领域中,不仅内部成员比外部实体对领域知识的事实论断更具有可信性,而且组内成员也更加关注于领域内知识的论断,具有更高的相似度,因而我们能够用聚类的方法来进行信息源的划分.

假设存在信息源集合 $S = \{s_1, s_2, \cdots, s_i\}$,事实集合 $F = \{f_1, f_2, \cdots, f_j\}$(其中 $F = \bigcup_{s \in S} F_s$),以及对象集合 $O = \{o_1, o_2, \cdots, o_k\}$.我们定义事实概率矩阵 $\boldsymbol{Q}_{i \times j}$ 和对象概率矩阵 $\boldsymbol{P}_{j \times k}$,分别满足等式 $\boldsymbol{S}^{\mathrm{T}} = \boldsymbol{QF}^{\mathrm{T}}$ 和 $\boldsymbol{F}^{\mathrm{T}} = \boldsymbol{PO}^{\mathrm{T}}$.由于 $\boldsymbol{S}^{\mathrm{T}} = \boldsymbol{QF}^{\mathrm{T}} = \boldsymbol{QPO}^{\mathrm{T}}$,我们可以得到从信息源到描述对象的转移概率矩阵,其中 t_{ik} 表示 s_i 描述了对象 o_k 某一属性的概率,可知元素 $t_{ik} = \sum_{v=1}^{j} q_{iv} p_{vk}$.我们可以用 $t_i = \{t_{i1}, t_{i2}, \cdots, t_{ik}\}$ 表示信息源 s_i 在对象空间上的投影.

我们用模糊 C 聚类的方法实现知识领域.我们用聚类中心矢量 $\boldsymbol{C} = \{c_1, c_2, \cdots, c_c\}$ 代表 c 个知识领域,同时定义 $\boldsymbol{C}^{\mathrm{T}} = \boldsymbol{GO}^{\mathrm{T}}$.

$\boldsymbol{G}_{c \times k} = \begin{bmatrix} g_{11} & \cdots & g_{1k} \\ \vdots & \ddots & \vdots \\ g_{c1} & \cdots & g_{ck} \end{bmatrix} = (g_{ck})$,其中 g_{ck} 表示聚类中心 c_c 在维度 o_k 上的值.

然后我们定义隶属度矩阵 $\boldsymbol{U}_{i \times c}$,满足 $\boldsymbol{S}^{\mathrm{T}} = \boldsymbol{UC}^{\mathrm{T}}$,其中元素 $u_{i \times c}$ 表示第 i 个信息源属于第 c 个领域聚类的隶属度.

$\boldsymbol{U}_{i\times c}=\begin{bmatrix} u_{11} & \cdots & u_{1c} \\ \vdots & \ddots & \vdots \\ u_{i1} & \cdots & u_{ic} \end{bmatrix}=(u_{ic})$,约束条件 $\sum_{n=1}^{c} u_{in}=1,\ \forall i, 0\leqslant u_{in}\leqslant 1$.

我们计算信息源 s_i 和领域聚类中心 g_c 的距离如下:

$$D_{ic}=\sqrt[p]{\sum_{r=1}^{k}(\,|t_{ir}-g_{cr}|\,)^{p}} \tag{4}$$

上式中 p 代表着距离参数.当 $p=1$ 时式(4)为汉明距;当 $p=2$ 式(4)为欧氏距离.聚类的目标是找到最优的聚类矢量 $\boldsymbol{G}_{\text{opt}}$,使得总体距离最小.聚类目标函数如式(5):

$$\min D^2=\sum_{m=1}^{i}\sum_{n=1}^{c} u_{mn}\Big(\sum_{r=1}^{k}(\,|t_{mr}-g_{nr}|\,)^2\Big) \tag{5}$$

3.3 最大熵加权法的推导

本节我们引入信息熵的概念到事实发现过程中.根据隶属度矩阵 $\boldsymbol{U}_{i\times c}$ 的定义,隶属度可以看做是第 i 个信息源属于第 c 个知识领域的概率.从而我们可以定义信息源 s_i 关于知识领域 g_c 的信息熵:$H_{ic}=-u_{ic}(\ln u_{ic})$.

信息熵反映了信息源对于不同的知识领域包涵的信息量,信息熵越大,信息源越能提供该知识领域内对象的有价值的事实论断,因而更具可信度.引入信息熵优化能够更合理的划分知识领域,判别不同信息源最擅长的知识领域,定量的分析信息源在每个知识领域蕴含的信息量,从而有效地利用好信息源和描述对象之间的相关性.

H 是矩阵 $\boldsymbol{U}_{i\times c}$ 的总体信息熵:

$$H=-\sum_{m=1}^{i}\sum_{n=1}^{c}[\,u_{mn}(\ln u_{mn})].$$

根据最大熵原则可以得到如下公式:

$$\max(H)=-\sum_{m=1}^{i}\sum_{n=1}^{c}[\,u_{mn}(\ln u_{mn})] \tag{6}$$

为了便于计算,我们将式(6)转换成下列形式(7):

$$\min(-H)=\sum_{m=1}^{i}\sum_{n=1}^{c}[\,u_{mn}(\ln u_{mn})] \tag{7}$$

其中 $\sum_{n=1}^{c} u_{in}=1,\ \forall i, 0\leqslant u_{in}\leqslant 1$.

为了到达最小化式(5)中的距离和式(7)中的信息熵的目的,我们提出了一个双目标优化问题,目标函数如下:

$$\min\Big[\sum_{m=1}^{i}\sum_{n=1}^{c} u_{mn}\Big(\sum_{r=1}^{k}(\,|t_{mr}-g_{nr}|\,)^2\Big)+\frac{1}{\theta}\sum_{m=1}^{i}\sum_{n=1}^{c}[\,u_{mn}(\ln u_{mn})]\Big] \tag{8}$$

约束条件:

$$\sum_{n=1}^{c} u_{in}=1,\ \forall i, 0\leqslant u_{in}\leqslant 1. \tag{9}$$

目标函数中的常量 θ 是为了保持总体距离 D 和信息熵 H 之间的平衡.

我们使用拉格朗日乘子法,为约束 $\sum_{n=1}^{c} u_{in}=1$ 引入拉格朗日乘子 λ,则目标函数可以写成:

$$L(u_{ic},g_{ck},\lambda)=\sum_{m=1}^{i}\sum_{n=1}^{c} u_{mn}\Big(\sum_{r=1}^{k}(\,|t_{mr}-g_{nr}|\,)^2\Big)+\frac{1}{\theta}\sum_{m=1}^{i}\sum_{n=1}^{c}[\,u_{mn}(\ln u_{mn})]+\lambda\Big|\sum_{n=1}^{c} u_{in}-1\Big| \tag{10}$$

通过对目标函数 $L(u_{ic},g_{ck},\lambda)$ 求解,我们可以得到唯一的最优解 u_{rc} 和 g_{ck}:

$$u_{rc}=\frac{\exp\Big(-\theta\sum_{r=1}^{k}(\,|t_{ir}-g_{cr}|\,)^2\Big)}{\sum_{n=1}^{c}\exp\Big(-\theta\sum_{r=1}^{k}(\,|t_{ir}-g_{nr}|\,)^2\Big)} \tag{11}$$

$$g_{ck}=\frac{\sum_{m=1}^{i} u_{mc}t_{ir}}{\sum_{m=1}^{i} u_{mc}} \tag{12}$$

我们根据上一节对象概率矩阵 $\boldsymbol{P}_{j\times k}$ 和聚类中心矩阵 $\boldsymbol{G}_{c\times k}$ 的定义可知:$\boldsymbol{F}^{\mathrm{T}}=\boldsymbol{P}\boldsymbol{O}^{\mathrm{T}}$ 以及 $\boldsymbol{C}^{\mathrm{T}}=\boldsymbol{G}\boldsymbol{O}^{\mathrm{T}}$.假设 $\boldsymbol{P}_{j\times k}$ 可逆,则可推出 $\boldsymbol{C}^{\mathrm{T}}=\boldsymbol{G}\boldsymbol{P}^{-1}\boldsymbol{F}^{\mathrm{T}}$.

然后根据隶属度函数的定义 $\boldsymbol{S}^{\mathrm{T}}=\boldsymbol{U}\boldsymbol{C}^{\mathrm{T}}$,我们最终可以得到从信息源到事实的加权矩阵 $\boldsymbol{W}_{i\times j}=\boldsymbol{U}\boldsymbol{G}\boldsymbol{P}^{-1}$,从而在每次迭代后我们可以给每个事实根据知识领域加上一个相应的权重 w_{ij}.

$$\begin{aligned}\boldsymbol{W}_{i\times j}&=\boldsymbol{U}\boldsymbol{G}\boldsymbol{P}^{-1}\\&=\begin{bmatrix} u_{11} & \cdots & u_{1k} \\ \vdots & \ddots & \vdots \\ u_{i1} & \cdots & u_{ik} \end{bmatrix}\cdot\begin{bmatrix} g_{11} & \cdots & g_{1k} \\ \vdots & \ddots & \vdots \\ g_{i1} & \cdots & g_{ik} \end{bmatrix}\cdot\begin{bmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{j1} & \cdots & p_{jk} \end{bmatrix}^{-1}\\&=(w_{ij})\end{aligned} \tag{13}$$

4 基于信息源聚类的最大熵加权事实发现算法

如上所述,我们最终推导出了基于信息源聚类的最大熵加权法权值.将权值带入到事实发现算法的迭代过程中,我们将可以得到一个全新的事实发现算法:基于信息源聚类的最大熵加权事实发现算法.本文中,我们给出两个迭代算法的具体流程.

4.1 Clustering based Maximum Entropy Weighted Sum(CMEWS)算法

我们重写迭代式(1)可以得到:

$$c'(f_j)=\sum_{s_i\in S_j} w_{ij}t'(s_i) \tag{14}$$

结合式(2),我们可以得到第一个算法:CMEWS 算

法.如同其他迭代方法一样我们需要一个初始化状态.我们假设初始状态下所有信息源具有统一的可信度 t_0,t_0可以设置在一个估计的平均值范围内,比如可以设置其为 0.9.我们还需要知道事实概率矩阵 $\boldsymbol{Q}_{i\times j}$和对象概率矩阵 $\boldsymbol{P}_{j\times k}$,另外我们随机的设定 c 个聚类中心点作为初始化的知识领域,随着迭代的进行,我们将得到最优的聚类划分以及相应的权值.

算法开始时,我们对于信息源和事实的信任程度的认识处于比较模糊的水平,通过迭代,我们逐步得到了对信息源可信度和事实置信度的定量的了解,当算法达到稳定时迭代停止.

4.2 Clustering based Maximum Entropy Weighted Trust-finder(CMEWT)算法

我们重写式(3):

$$c'(f_j) = 1 - \prod_{s_i\in S_j}(1 - w_{ij}t'(s_i)) + \rho \sum_{o(f')\in O_j}\delta(f') \cdot \mathrm{imp}(f' \rightarrow f) \tag{15}$$

根据文献[5],上式中$(1 - w_{ij}t'(s_i))$通常较小,乘积形式往往会导致下溢.为了方便计算,我们采用了对数形式的可信度,定义事实置信值如下形式:

$$\tau(s_i) = -\ln(1 - w_{ij}t'(s_i)) \tag{16}$$

$$\delta(f_j) = -\ln(1 - c'(f_j)) = -\ln\Big(\prod_{s_i\in S_j}(1 - w_{ij}t'(s_i))\Big) \tag{17}$$

进而根据式(15),当考虑相似事实之间的关联关系后:

$$\delta^*(f) = \delta(f) + \rho \sum_{o(f')\in O_j}\delta(f') \cdot \mathrm{imp}(f' \rightarrow f) \tag{18}$$

最终,我们可以得到下式形式的事实置信度:

$$c(f_j) = \frac{1}{1 + \mathrm{e}^{-\gamma\delta^*(f)}} \tag{19}$$

其中 ρ 和 γ 分别代表关联事实的影响和延迟因子.

根据式(19)和式(2),我们可以得到第二个算法:CMEWT 算法,迭代过程同上.

5 实验分析

本节我们将通过实验验证算法的性能.我们首先描述一下实验数据集.我们在两类数据集上进行了准确性和算法效率测试:一是高度可配置的合成数据集,另一个是真实网络数据集.我们把我们的算法和 Sum、TrustFinder、Average.Log、PoolInvestment 以及 3-Estimate 五种标准算法进行了综合比较,实验结果证明了我们算法的有效性.

我们固定迭代次数为 20 次.作为评价标准,我们采用了全局预测准确率,即准确判定的事实占总事实数的比率.第二个评价标准是完整性,是指正确分类的信息源占总体的比例.这个参数主要是反应算法聚类的能力.第三个标准是收敛速度,主要是指算法达到稳定状态所需要的迭代次数,次数越少,算法效率越高.

合成数据集

首先我们测试了一个中等规模的可配置的人工数据集.我们用 Matlab 7 搭建了信息源-事实网络仿真环境,数据集基于类似图 3 的网络结构.我们设定存在 20 个信源,每个信息源以随机概率提供 10 个不同类型的事实论断,描述不同对象的属性,对象的总数为 16 个.为了方便计算,我们假定对象属性为布尔类型,参考值均为 TRUE.而事实论断值介于(-1,1)之间.如果事实论断为 1 表示该事实支持对象属性为 TRUE;如果事实论断为 -1 表示该事实不支持对象属性为 TRUE;否则我们认为事实部分支持或者部分反对对象属性为真.迭代计算得到的事实置信度则表示该事实可以信赖的程度,如果事实置信度与我们的预设相一致,则我们认为算法正确.

为了构造类似图 3 的网络结构,我们首先人为地把对象平均分入四个组,同时假定将信源也平均分配到各组;设置同组每个信息源分别提供 8 条事实正确地描述组内的某个对象(事实符号为正),提供 2 条事实错误地描述组外的某一对象(事实符号为负);接下来我们给信息源和事实之间分配随机的概率,即随机产生事实概率矩阵 $\boldsymbol{Q}_{i\times j}$,同时调整同一信源可以有多条事实描述同一对象,这样就模拟了多类型事实的情况;最后我们在生成的数据集上分别运行本文所提出的两个算法和其他基准算法,产生的实验数据与我们预设环境相对比,以此来验证本文算法的有效性.实验结果如图 4 所示.

图 4 所示的是不同算法的全局准确率,图中 x 轴表示迭代次数,y 轴表示全局准确率.绝大多数算法在迭代 10 次之后都达到了收敛.如图 4 所示,由于计算上的复杂性,本文算法收敛速度相对较慢,在 10 次迭代之前算法的全局准确度低于大多数基准算法.但是随着迭代的进行,领域聚类逐渐找到最佳聚类中心,权值逐步优化,因而全局准确率也逐步提高.10 次迭代之后,本

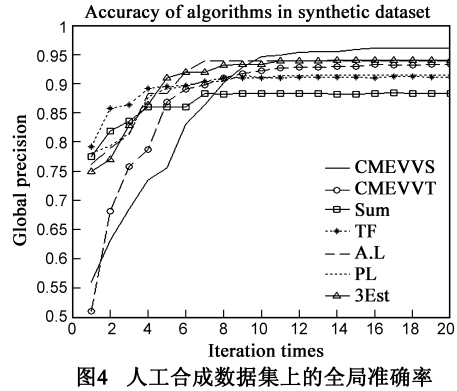

图4　人工合成数据集上的全局准确率

文算法的准确度能够达到甚至超过基准算法.

表 1 所示是 20 次迭代后七种算法的错误率和完整度.表 1 第一行表示算法错误判断的个数,有两种错误,一种是正确事实判为错误,另一种是错误事实判为正确.实验结果表明第二类错误要多于第一类错误.表 1 第二行表示算法的完整度,这一参数表示的是算法基于知识领域聚类的能力.我们的实验环境中,数据集 20 个信息源均匀分布在四个聚类中,实验结果表明聚类后除了 CMEWT 中出现一次误判之外,其余信源都能正确的划分到正确的聚类之中,因而两种算法具有较好地聚类能力.

表 1 算法错误率和完整度

算法名称	CMEWT	CMEWS	Sum	TF	A.L	PL	3Est
错误数	8	13	23	18	12	17	12
完整度	95%	100%	NA	NA	NA	NA	NA

真实世界数据集

第二个数据集来自于在线的常识问卷调查,数据集由文献[6]提供.此处问卷调查包括了 17 个话题,涵盖了从文学、地理到历史知识等 5 个主题.对于每个话题,有 4 到 14 个可能的答案,在这些答案之中只有一个是正确的.一共进行了 601 次问卷,每张问卷有 95 个题目.根据文献[6],在进行了功能依赖性处理后,我们一共得到了 37,170 个有效的事实论述.

该数据集中,我们可以把不同的问卷视为信息源,题目看做是事实,17 个话题可以被看做是对象,而根据不同的主题,我们很自然的可以把对象划分为 5 个知识领域.该数据集做了部分事实的忽略处理,因而得到的总事实数小于 601×95,本文实验中只考虑有效事实部分,即事实总数等于 37,170.

图 5 显示了不同算法在真实世界数据集上的全局准确率.本文的算法依然在 20 次迭代之后能够达到或者优于基准算法的全局准确率.我们注意到该数据集假设每个问题只能有唯一正确的答案,这相对于单值事实,因而在这样的情况下 CMEWS 没有得到太高的性能提升,但是在收敛速度上比 CWMET 稍快.而 CWMET 在 14 次迭代后稍稍优于 3-Estimate 算法和 PoolInvestment 算法.

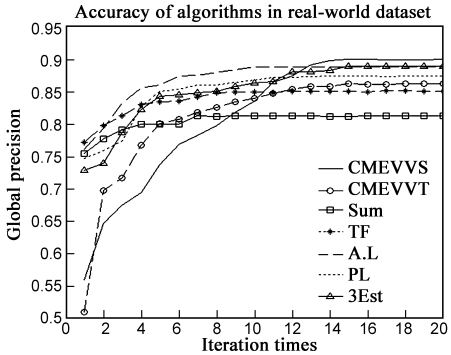

图5 真实世界数据集上的全局准确率

6 结论

本文提出了一种基于信息源聚类的最大熵加权信任分析算法,使我们能够在进行信任分析时有效地融合诸如对象属性以及信息源关联性等信息.传统的事实发现算法都没有有效地利用描述对象的相关知识,我们的方法将与同一描述对象集合强相关的若干信息源聚类于同一知识领域,聚类中心被视为知识领域专家,从而有效的利用了信息源和描述对象之间的相关性;然后根据信息源在各个知识领域的影响程度动态地改变事实置信度中的信息源权值,最终迭代地计算信息源和事实的信任水平.作者第一次将信息熵的概念带入到信任分析中,作者还考虑了信息源和事实的不确定性,从而能够有效地处理一致性学习和多值事实问题.实验结果证明该算法能够有效地提升信任分析能力.但是该算法还存在收敛速度慢和容易陷入局部极值等问题,这将是我们下一步的研究重点.

参考文献

[1] Han Jia-wei. Mining heterogeneous information networks by exploring the power of links[A]. Proceedings of the 20th international conference on Algorithmic learning theory[C]. Berlin Heidelberg: Springer-Verlag, 2009. 3 – 4.

[2] Sun Yan, et al. Information theoretic framework of trust modeling and evaluation for adhoc networks[J]. IEEE Journal on Selected Areas in Communications, 2006, 24(2): 305 – 317.

[3] Sergey Brin, Lawrence Page. The anatomy of alarge-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(7): 107 – 117.

[4] Jon M, Kleinberg. Authoritative sources in a hyperlinked environment[A]. Proceedings of the ACM-SIAM Symposium on DiscreteAlgorithms [C]. San Francisco, California, USA: SIAM, 1998. 668 – 677.

[5] Yin X, Yu P, Han Jia-wei. Truth discovery with multiple conflicting information providers on the web[J]. IEEE Transaction on Knowledge and Data Engineering, 2008, 20(6): 796 – 808.

[6] Alban Galland, Serge Abiteboulетal. Corroborating information from disagreeing views[A]. Proceedings of the 3rd ACM International Conference on Web Search and Data Mining[C]. New York City, USA: ACM, 2010. 131 – 140.

[7] 杨静,赵家石,张健沛.一种面向高维数据挖掘的隐私保护方法[J].电子学报,2013,41(11):2187 – 2192.
Yang Jing, Zhao Jia-shi, Zhang Jian-pei. A privacy preservation method for high dimensional data mining[J]. Acta Electronica Sinica, 2013, 41(11): 2187 – 2192. (in Chinese)

[8] J Pasternack, D Roth. Knowing what to believe (when you already know something)[A]. Proceedings of the 23rd International Conference on Computational Linguistics[C]. Beijing: Coling, 2010. 877 – 885

[9] L Blanco, V Crescenzi, et al. Probabilistic models to reconcile complex data from inaccurate data sources[A]. Proceedings of the 22nd International Conference on Advanced Information Systems Engineering[C]. Hammamet, Tunisia: Springer, 2010. 83 – 97.

[10] 李光,王亚东.一种改进的基于奇异值分解的隐私保持分类挖掘方法[J].电子学报,2012,40(4):739 – 744.
Li Guang, Wang Ya-dong. An improved privacy-preserving classification mining method based on singular value decomposition[J]. Acta Electronica Sinica, 2012, 40(4): 739 – 744. (in Chinese)

作者简介

侯　森　男,1979 年生于河南郑州.解放军信息工程大学信息技术研究所博士研究生.研究方向为数据挖掘、通信与信息系统.
E-mail: scuths@hotmail.com

罗兴国　男,1952 年生于重庆.教授,博士生导师.解放军信息工程大学信息技术研究所总工程师,研究方向为数字通信、移动通信、高性能计算和数据挖掘.

宋　克　男,1976 年生,硕士,解放军信息工程大学信息技术研究所讲师,研究方向为网络体系结构、片上系统、集成电路设计.