超图在数据挖掘领域中的几个应用

崔 阳 杨炳儒

(北京科技大学信息工程学院 北京 100083)

摘 要 数据挖掘技术的进一步发展同新理论和新方法的应用密切相关。超图以图论和集合论为基础,近年来在数据挖掘领域超图理论已经得到运用。首先概述了超图的基本概念,然后重点介绍结合了超图理论的新的关联规则挖掘算法 Maradbcm,以及超图在聚类、空间数据挖掘方面的运用情况。

关键词 超图,数据挖掘,Maradbcm 算法,空间数据挖掘

中图法分类号 TP311.13 **文献标识码** A

Application of Hypergraph in Data Mining

CUI Yang YANG Bing-ru

(Dept. of Computer Science and Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract The development of data mining is tightly related to the application of new theories and methods. Definition of hypergraph is based on Graph Theory and Set Theory, and has been used in data mining recent years. Some important concepts of hypergraph were given at first. Then the application on Maradbcm algorithm, clustering and spatial data mining was introduced.

Keywords Hypergraph, Data mining, Maradbcm algorithm, Spatial data mining

数据挖掘技术的目的在于发现数据中有用的模式,并使用该模式帮助解释当前的行为或预测未来的结果,以人们容易理解的形式提供有用的决策信息。随着数据挖掘技术的不断发展,一些新技术与方法逐渐被应用到数据挖掘的研究中,从而为一些新老问题的解决提供了新思路。

超图理论由于较抽象和复杂,起初的研究和应用进展都较缓慢。但近年来,超图日益为研究者所关注,有关的研究课题不断增多,并已在包括数据挖掘等诸多领域中得到了应用。

1 超图的概念

C. Berge 于 1970 年第一次提出超图,并对超图理论进行了系统阐述[1]。超图的理论基础是图论和集合论。具有共同属性特征的对象属于一个集合,不同的抽象层次可归属于集合的集合;如此构成以集合的包含关系为基础的结构,这种结构可用超图来表示。

超图的相关定义如下:

定义 1 设 $X=\{x_1,x_2,\cdots,x_n\}$ 是 n 个结点的有限集。X 的某一子集被称为 E_i,若 $\bigcup_{i=1}^{m}E_i=X$,则称 $P=\{E_1,E_2,\cdots,E_m\}$ 为边集,称 $H=\{X,P\}$ 为超图。在超图的图形表示中,将属于同一条边的点由一条闭曲线包围在一起,这条闭曲线称为超图的一条超边。

定义 2 在 P 中任取子集 $F\subset P$,用 F 做边集构成超图 $H_F=(X_F,F)$,其中 $X_F=\bigcup_{E_i\in F}E_i$,则称 H_F 为超图 H 的部分超图。

若在超图的边集中定义了方向,则超图为有向超图。有向超图是超图概念的扩展。其定义为:

定义 3 设 $H=(X,P)$ 为超图,若 $P=\{E_1,E_2,\cdots,E_m\}$ 为有向边集,则称 H 为有向超图。

此外,有向超图的一些相关定义还包括:

定义 4 与同一条边关联的端点称为邻接;若两个有向超边有一公共的顶点,则称此两条有向超边为邻接。

定义 5 设在超图 $H=\{X,P\}$ 里存在一个点边集序列 $\mu=(x_1E_1x_2E_2\cdots E_{q-1}x_qE_qx_{q+1})$,其中 x_i 是 H 的相异节点,E_i 是 H 的相异边,且 $x_k,x_{k+1}\in E_k,(k=1,2,\cdots,q)$。序列 μ 中共含 q 个相异边,若 $x_{q+1}\neq x_1$,则序列 μ 是 H 的一条链,其长为 q;$x_{q+1}=x_1$,则序列 μ 是 H 的一个圈。

定义 6 若有向超图 H 上有链起于顶点 A,止于顶点 B,便记作 $A\equiv B$,此时称 A 到 B 是可达的。

定义 7 若在集合 X 或 P 上定义权函数,则超图 $H=(X,P)$ 称为点或边权超图。

有关超图的表示分为无向超图和有向超图两种情况。在无向超图中,若 $|E_i|>2$,则画成一曲线包围 E_i 中的所有点;若 $|E_i|=2$,则画成一曲线连接两点;若 $|E_i|=1$ 则为自环。有向超图的一种表示法是超边、端点和引线表示法[2],即在 E_i 的每一端引出一根线,称为引线。若 $E_i\cap E_j\neq\Phi$,则在其交集中的端点用引线联接。

2 超图在数据挖掘领域的典型应用

2.1 关联规则挖掘

到稿日期:2009-07-20 返修日期:2009-09-30 本文受国家自然科学基金(60675030)资助。

崔 阳(1979—),男,博士生,主要研究方向为知识发现,E-mail:cuiyang14@163.com;杨炳儒(1943—),教授,博士生导师,主要研究方向为知识工程与知识发现。

关联规则是数据挖掘领域最为成熟和最为活跃的研究课题，目前对其的研究成果已不胜枚举。有关关联规则挖掘的算法，最著名的仍当属 Agrawal 与 Srikant 在 1994 年提出的 Apriori 算法[3]。该算法使用一种称为逐层搜索的迭代方法搜索频繁 n-项集，并在搜索过程中利用 Apriori 性质来压缩搜索空间，提高频繁项集逐层产生的效率。

Apriori 算法的两个主要缺点是效率较低和无法分析稀有信息。而 Apriori 算法之后出现的绝大多数关联规则挖掘算法，都是在其基础上进行改进的，因此也无法从根本上克服这两个缺点。北京科技大学知识工程研究所近年来提出了 KDD 中的双库协同机制[4,5]，并以此为基础提出了新的关联规则挖掘算法 Maradbcm。

Maradbcm 算法与各种 Apriori 算法虽然在本质上都基于统计方法，但存在显著区别，主要表现之一便是二者基于的学术思想不同。Apriori 算法是基于组合论的数据库全局搜索，而 Maradbcm 算法则以内在机理研究为基础，基于知识短缺进行定向挖掘。

知识短缺是指在现有的知识库中没有重复的和没有冗余的知识。冗余的知识指知识库中存在多余的知识或者存在多余的约束条件。一般认为，如果两条规则链中第一条规则的条件相同，且最后一条规则的结论等价，则称此两条规则链存在冗余。

为了发现知识短缺，Maradbcm 算法采用了有向超图来表示知识库中的知识[6]，这是一种新的知识表示方法。在这种知识表示方法中，有向超图 H 为两元组 (V,E)，其中图的顶点 V 是知识库中的知识素结点的集合，E 是知识库中规则对应的有向边。若一条规则 $r_i=p_1\wedge p_2\wedge\cdots\wedge p_k\rightarrow p_j$，则有向边 $e_i=\langle(p_1,p_2,\cdots,p_k),p_j\rangle$ 是一个有序偶。有向边的第一个元素是 V 的一个子集，与规则的前件相对应；有向边的第二个元素是 V 的一个元素，与规则的后件相对应。

Maradbcm 算法中引入了知识素结点和知识合结点的概念。知识素结点是指对应在知识库中的数据库中每个属性程度词，并将知识素结点的集合定义为 $E=\{e_1,e_2,\cdots,e_n\}$。知识合结点是指不含否定联结词的合式公式 $\theta_0e_1\theta_1e_2\cdots\theta_{m-1}e_m\theta_m$，其中 $e_i\in E,i=1,2,\cdots,m;\theta_i\in J,i=0,1,\cdots,m$。这里 J 是由符号“$\wedge$”，“$\vee$”，“(”，“)”4 个符号及其任意组合而形成的集合。对于关联规则而言，只有“$\wedge$”合取形式，即 $e_1\wedge e_2\wedge\cdots\wedge e_m$ 的形式。

有向超图 H 的邻接矩阵 $A(H)$ 表示如下：

$$a_{ij}=\begin{cases}1, & \text{当}\begin{cases}\langle p_i,p_j\rangle\in E,i,j=1,2,\cdots,n\\ \langle p_i,p_j\rangle\in E,j=1,2,\cdots,n;i=n+1\end{cases}\\ 0, & \text{其它}\end{cases}$$

式中，p_i 为知识库中知识合结点，$A(H)$ 可以直接由知识库得到。

有向超图 H 的可达矩阵 $P(H)$ 表示如下：

$$a_{ij}=\begin{cases}1,\text{当 } A \text{ 可达 } b\\ 0,\text{其它}\end{cases}$$

式中，A 为知识库中知识素结点或合结点，b 为知识库中的知识素结点。A 可达 b 表示为 $A\equiv b$。可达的概念如定义 6 所述。该矩阵的列是固定的，即为知识素结点的个数；而行由知识素结点和出现在知识库中的合结点组成。对应结点 A 的一行表示该结点的可达情况，对应素结点 b 的一列表示其它结点可达 b 结点的情况。

通过有向超图的邻接矩阵 $A(H)$，计算有向超图的可达矩阵 $P(H)$，得到的 $P(H)$ 中 0 元素就是短缺的知识。算法的简略步骤为：

假设素结点的个数为 m，知识库中出现的知识合结点个数为 n，则首先设 $P(H)=0$。

(1)读取知识库中的一条知识 $r:a\rightarrow b$(设知识结点 a 对应的行为第 i 行，b 所对应的列为 j 列)

(2)for $r_:=1$ to n do

(3) if $P(r,j)=1$ then

(4) for $s_:=1$ to m do

(5) $P(i,s):=P(i,s)\vee P(j,s)$

(6)判断知识库中是否还有知识未读取，若有则读取下一条知识，转(2)；若没有则转(7)；

(7)结束

其中“$\vee$”表示布尔和。

在得到短缺知识集后，还要对短缺知识进行剪枝、定向搜索、评价等，最终将获取的知识呈现给用户。利用有向超图进行定向挖掘的 Maradbcm 算法在关联规则挖掘方面具有较大的优越性。

2.2 聚类

聚类分析是数据挖掘中广为研究的方法之一。当前各种聚类算法的适应性、效率与最优解仍须进一步研究。基于超图的聚类算法很多。有一种基于超图模式的聚类方法是将高维空间中的原始数据及其相互关系映射到带权超图 $H=(V,E)$[7]。其中 V 表示数据项点集，E 表示连接相关数据项子集的超边集。设 w_j 为对应于 E 中每一条超边 $e_j\in E$ 的权重，用以衡量超边连接的多个相关数据项之间的相关程度。同时给出评价函数 $f=\sum w_j$。在数据集的超图确定后，选用一种图形分割最优算法，对超图进行分割。分割依据是令评价函数 $f=\sum w_j$ 的值最小，即令各子分割之间相关性之和最小。该方法在数据点多、聚类类别多和数据项维数高的情况下效率要好于传统聚类算法。

基于超图的聚类的应用很多。例如，在虚拟环境中利用基于超图的聚类进行遍历预测[8]，用超图模型对无尺度网络进行聚类[9]。近年来还有人提出了利用模糊超图模型进行聚类[10]等。

基于超图的聚类在 Web 挖掘方面也有应用前景。在一些比较重视对用户浏览习惯进行分析的网站，如电子商务中的 B2B 网站等，可以利用超图对一个 Web 网站的所有页面进行聚类，通过分析用户的浏览记录来挖掘用户的浏览习惯。首先建立频繁网页集，用以保存多数用户在一次网站访问中共同浏览的页面。这些页面的内容高度相关。之后建立网页超图，Web 站点中的页面即为超图的顶点，而频繁网页集则作为超边[11]。

为进一步描述网页与不同的频繁网页集之间联系的密切程度，可以挖掘出各频繁网页集中的所有关联规则，并在此基础上计算出每一个频繁网页集中各关联规则可信度的平均值，作为网页与该频繁网页集联系紧密程度的度量。每个超边的权取该边所对应的频繁网页集中所有关联规则可信度的平均值。对页面的聚类同样通过超图分割完成，分割原则是被切割的超边权值和尽可能小，以保证相互关联较小的网页

在不同的子图，而关联密切的网页在同一子图。在将网页聚类的基础上，可以根据用户的浏览事务对用户聚类，以便针对不同类型的用户给出偏重有所不同的页面显示。依据的原则是用户事务和网页聚类的相似度。

2.3 空间数据挖掘

空间数据是一类具有多维特征，即时间维、空间维以及众多的属性维的数据，描述了复杂系统的状态、系统的性质、系统的空间分布和系统的发展演化。空间数据挖掘指的是从空间数据库保存的海量空间数据中抽取出空间对象之间的相互关系及反映其演化规律的知识。

超图理论在空间数据挖掘中也具有应用价值。将有向超图同面向对象技术相结合所形成的用于表示模式矢量的结构以及子模式之间关系的图形称为超图模型[12]。超图模型支持类、对象、属性和联系等概念。其中属性包括类的性质和对象的性质；联系包括类与类(对象与对象)之间的连接和关系。超图模型可以表达模式的复杂结构和关系，并将模式同面向对象的相关概念联系在一起，其中包括类、继承、聚合、概括(支持超类)、多层次关系(超类间的联系)以及复合类等。

之所以将超图引入空间数据挖掘，是因为空间数据蕴含有复杂的规律和知识，而超图模型能够用图的方式抽象地描述这种复杂性，从而将空间数据的组织结构予以简化；同时用超图模型还可进行空间数据的可视化表示，便于计算机实现[13]。

超图模型中顶点表示模式矢量的组成(属性、拓扑结构、子模式)，不同的顶点可表示不同抽象层次的模式，顶点之间的有向弧段代表对象类之间的关系。在超图模型中，可以用闭合曲线和赋值的连线表示空间数据的空间维；用点和闭合曲线的赋值以及赋值曲线来表示空间数据的属性维；用属性值的变化以及连线值的变化来表示空间数据的时间维。超图模型还可以表示空间对象之间的层次关系，以及关系的强弱。

空间知识发现的常见的知识类型，如关联规则、聚类和分类规则等，也可利用各种表示逻辑关系的图来表示知识之间的层次和相互依赖。由于超图模型综合了超图和有向图的优点，因此可以将关联规则可视化表示，图中节点表示数据的项，边表示关联关系，规则的可信度和支持度可用不同的颜色和数值来表示。有向图适合于数据和规则数目较少情况下的知识的可视化。

结束语 超图基于图论和集合论，已在数据挖掘研究中得到了运用。超图模型能够利用图的逻辑结构，有效组织和传递数据集的结构、关系和含义，实现关联规则、聚类和分类等知识的获取和表示。超图(特别是有向超图)是 Maradbcm 这一全新关联规则挖掘算法的重要理论基础之一。随着研究的深入，超图理论还将在数据挖掘各研究领域中起到更大的作用。

参 考 文 献

[1] Berge C. Graph and Hypergraph[M]. Amstezdam: North-Holland, 1973

[2] 黄汝激. 超网络的有向超树分析法[J]. 电子科学学刊, 1987, 19(3): 244-255

[3] Agrawal R, Imielinski T, Swami A. Mining Associations between Sets of Items in Massive Databases[C]// Buneman P, Jajodia S, ed. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data(SIGMOD-93). 1993: 207-216

[4] 杨炳儒, 王建新. KDD 中双库协同机制的研究(Ⅰ)[J]. 中国工程科学, 2002, 5(4): 41-51

[5] 杨炳儒, 王建新, 孙海洪. KDD 中双库协同机制的研究(Ⅱ)[J]. 中国工程科学, 2002, 4(5): 34-43

[6] 杨炳儒, 孙海洪. 基于双库协同机制的挖掘关联规则算法 Maradbcm[J]. 计算机研究与发展, 2002, 39(11): 1447-1455

[7] 张蓉. 数据聚类技术的研究[J]. 计算机工程与应用, 2002, 16: 145-147

[8] Hung Shao-shin, Liu Shing-min. Using hypergraph-based clustering scheme for traversal prediction in virtual environments[C]// 2007 First IEEE Symposium on Computational Intelligence and Data Mining. Honolulu, HI, USA, 2007: 429-436

[9] Zheng Yu, Qian Rong. A Hypergraph model for clustering scale-free network[C]// The 27th Chinese control conference. CCC, 2008: 561-565

[10] Liu Yang, Zhang Ya-jie. A new data clustering method for farmland evaluation based on fuzzy-hypergraph model[J]. Journal of Wuhan University of Technology(Information & Management Engineering), 2007, 29(11): 126-128

[11] 李尊朝. 基于网页超图分割的 WEB 聚类法[J]. 纺织高校基础科学学报, 2003, 16(3): 261-264

[12] 杨炳儒, 张德政. 超图模型：基于超图的设计模式描述和复用实现[J]. 计算机工程与应用, 2001, 13: 46-48

[13] 孙连英, 彭苏萍, 张德政. 基于超图模型的空间数据挖掘[J]. 计算机工程与应用, 2002, 11: 30-32

(上接第 178 页)

[3] 赵越. 一种基于相关性的网络数据流聚类算法[J]. 西南师范大学学报：自然科学版, 2009(12 增刊)

[4] Le A, Al-Shaer E, Boutaba R. On Optimizing Load Balancing of Intrusion Detection and Prevention Systems[C]// INFOCOM Workshops 2008. IEEE, 2008, 4: 1-6

[5] 鲁瑞华, 张为群. 模糊加权中值滤波器[J]. 计算机科学, 2006, 6: 186-188

[6] 孙钦东, 张德运, 高鹏, 等. 并行入侵检测系统的负载均衡算法[J]. 小型微型计算机系统, 2004, 25: 2215-2217

[7] 伍海波, 陶滔, 唐启涛, 等. 基于负载均衡的并行入侵检测系统设计[J]. 微计算机信息, 2009, 25: 88-90

[8] 代伟, 刘敏, 余永武. 基于 Ad Hoc 网络的混合入侵检测算法[J]. 重庆工学院学报：自然科学版, 2008, 22(3): 75-78