基于图的多关系数据挖掘理论研究与方法*)

王映龙[1,2] 宋泽锋[1] 陈 卓[1]

(北京科技大学信息工程学院 北京 100083)[1] (江西农业大学计算机与信息工程学院 南昌 330045)[2]

摘 要 在过去的几年,结构化数据挖掘的需求日渐兴起,图是计算机学科和离散数学中最好的结构数据研究之一,基于图的数据挖掘已越来越广泛。本文介绍了基于图的数据挖掘的理论基础及其研究方法。

关键词 图,树,路径,结构化数据,数据挖掘

Graph-based Multi-relations Data Mining Fundamental Research and Application

WANG Ying-long[1,2] SONG Ze-feng[1] CHEN Zhuo[1]

(School of Information Engineering, Beijing University of Science and Technology, Beijing 100083, China)[1]

(School of Computer and Information Engineering, Jiangxi Agriculture University, Nanchang 330045, China)[2]

Abstract The need for mining structured data has increased in the past few years. One of the best studied data structures in computer science and discrete mathematics are graphs. It can therefore be no surprise that graph based data mining has become quite popular in the last few years. This article introduces the theoretical basis of graph based data mining and surveys the state of the art of graph-based data mining and brief descriptions of some representative approaches.

Keywords Graph, Tree, Path, Structured data, Data mining

1 引言

对于现实世界中普遍存在的关系数据库和复杂的结构化对象,传统的基于属性——值的数据挖掘方法存在明显的缺陷,各种数据的多关系数据挖掘研究具有重大的理论意义与应用价值。图拓扑结构是数学中最根本的结构,它与逻辑语言有非常紧密的联系,预计图的挖掘将对数据挖掘和机器学习的新发展作出更多的贡献。此外,图的挖掘有很深的实际应用潜力,因为图结构数据广泛存在于各行各业的实用领域(包括生物、化学、材料学和通信网络)中。

2 基于图的数据挖掘理论基础

2.1 图的分类

基于图的数据挖掘这一研究领域的历史不长,但它的理论基础并不局限在某一理论。这是因为在数学里对图的研究有悠久的历史。在这个部分,回顾基于图的数据挖掘方法的五个理论基础。它们是子图分类、子图同构、图的不变式、挖掘措施和解决方法。子图分为一般子图、诱导子图、连通子图、有序子结构、无序子结构、路径等几种不同的类(如图 1),并且基于图的数据挖掘方法主要取决于它的目标类。子图同构在基于图的数据挖掘里是子结构匹配和计算的理论基础。图的不变式在一些方法里对有效减少目标子图的搜索空间提供了一个重要数学标准。再者,挖掘措施定义将被挖掘的模式特征和常规数据挖掘的相似性。在本文里,理论依据解释为唯一不受指引没有标签但有/无环和平行的边的图,由于篇幅有限,这里只介绍有或无循环边和平行边的无向图的理论基础。但几乎相同讨论了有向图或者被标记图的应用。在基于图的数据挖掘里使用的大多搜索算法都来自人工智能,但一些额外搜索算法在数学里使用。

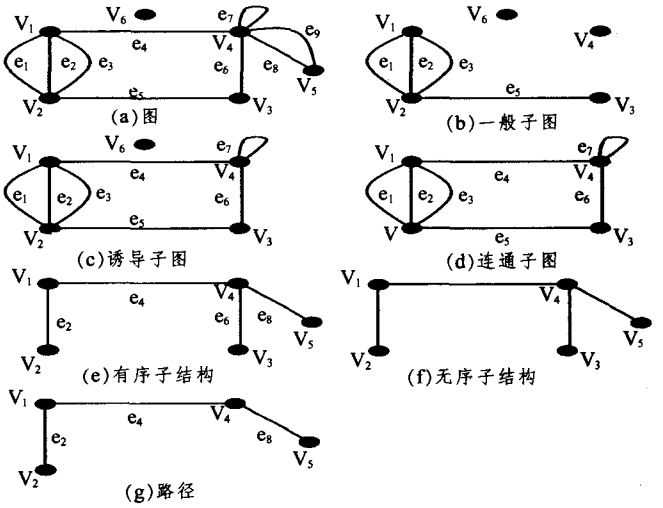

图 1 图的分类

2.2 子图同构

给出图 $Gx(Vx,Ex,fx)$ 和图 $Gy(Vy,Ey,fy)$,"子图同构"就是找出子图 $Gsx(Vsx,Esx,fx)$,$Gsy(Vsy,Esy,fy)$ 及边集 Vsx 和边集 Vsy 双向映射关系 gxy,使得子图 Gsx 和 Gsy 是同一子图,也就是 $fx(exh)=(vxi;vxj)\in Esx$ 当且仅当 $fy(eyh)=(vyi;vyj)\in Esy$,这里 $vyi=gxy(vxi)$ and $vyj=gxy(vxj)$。存在 gxy 确保子图 Gsx 和 Gsy 之间的拓朴结构相同。如在图 1 中的(b)、(d),它们共同以顶点$\{v_1,v_2,v_3\}$和边$\{e_1,e_2,e_3,e_4\}$构成,由 $vi=gbd(vi)$,$i=1,2,3$ 的双向映射形成子图。这个映射就是图(b)和图(d)之间的同构。

*)国家自然科学基金(60675030)资助;教育部科技重点项目(教技司[2000]175)资助。**王映龙** 副教授,博士研究生,研究方向为数据挖掘;**宋泽锋** 博士研究生,研究方向为数据挖掘;**陈 卓** 博士研究生,研究方向为粗糙集理论及其应用、数据挖掘。

在基于图的数据挖掘里，子图同构被进一步扩展到多图范围。给定图集$\{Gk(Vk,Ek,fk)\mid k=1,\cdots,n\}$，找出子图$Gs(Vs,Es,fs)$、子图集合$\{Gsk(Vsk,Esk,fsk)\mid k=1,\cdots,n\}$和图$Gs$与每一$Gsk$(所有$k=1,\cdots,n$)的双向映射$fs$。满足条件存在$Gs(Vs,Es)$是被给图集的公共子图。这个子图同构定义提供了被给定图的拓扑结构相同的匹配和计数的基础。子图同构问题(也就是决定两图是否有拓扑结构相同的问题)计算复杂度还不知道。它是一个NP完全或多项式，所有去分类的尝试至今不能解决。另一方面，子图同构(也就是判断一个图是否是另一个图的子图问题)被认为是NP完全问题。

2.3 图的不变量

图的不变量是对图的拓扑结构特征化的度量。如果两个图拓扑结构相同，也就是说它们同构，它们就有相同的图的不变量，但反之不成立。如图的顶点数、每个顶点的度数、连接顶点的边数、环数等等都是图的不变量。同构图的这些不变量总是相同的，但根据这些给定的图的不变式不能推出图的同构。图的不变量不能等同图的同构和计数，但是用它可以去减少解决图同构的搜索空间。如果两子图的不变式中的任何一个不相同，这两个子图肯定是不同构的。

过去几十年里，离散数学就研究了用图的不变量去解决图的同构问题。最有代表性的是NAUTY算法，被认为是解决图同构问题的最好算法[1]。它主要是利用图中每一顶点的图的不变量，如顶点的度数和邻接顶点度数的个数，来显著地减少搜索空间。在两图间有不相同的图的不变量的映射在图的同构问题中绝对不存在，因在图的拓扑结构里这样的顶点落在两图的不相应的位置。基于图的不变量的分而治之的方法在大多实际问题中显著地增强了计算效率。

图结构的更直接的表示和处理是用代数框架，如邻接矩阵[2]。第i行和第i列对应第i个顶点，矩阵中的i,j元素就是连接顶点的边$\{f(eh)=(vi,vj)\}$的集合。严格地说，因为它的元素不是一真正的数值，它不是一个矩阵。在这里我们为了方便称它为矩阵。若两顶点之间不存在边，则对应的元素为空或0，如图1(a)的邻接矩阵表示如下：

	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	$\{e_2,e_3,e_4\}$	0	$\{e_1\}$	0	0
v_2	$\{e_2,e_3,e_4\}$	0	$\{e_3\}$	0	0	0
v_3	0	$\{e_5\}$	0	$\{e_6\}$	0	0
v_4	$\{e_1\}$	0	$\{e_6\}$	$\{e_7\}$	$\{e_8,e_9\}$	0
v_5	0	0	0	$\{e_8,e_9\}$	0	0
v_6	0	0	0	0	0	0

所有图的不变量都能从这种图的直接表达方式中提取出来。缺点是在内存消耗和处理时间上有很大的复杂性。

一个最平常的且非常重要的图的不变量是“规范标签”和“规范形式”[2,3]。一个图能用多种形式表示，所有的邻接矩阵来自表示同一图的行和列的不同排列的邻接矩阵。在不同的图的表示中，这种不确定性都是共有的，这个诱导了图同构问题的搜索空间的混合增长。规范标签是一个最有效的补救。规范标签有不同的定义，但它必须唯一地表示一个图。例如，$n\times n$的邻接矩阵能按矩阵元素的下列排列而形成代码标签化：$x1,1x1,2x2,2x1,3x2,3x3,3\cdots xn-2,xn-1,nxn,n$，因为无向图的矩阵是对角线对称，这里仅用右上角的列，相似的代码也适合有向图。我们能按字母顺序最小代码唯一地定义规范标签和矩阵对应的规范标签的邻接矩阵的规范表。规范标签和规范表显著地缩小了图的表示不定性和搜索空间。

最近一个新的研究方向是用图的不变量去构建一个特征化图的高维特征空间。如果图能转换成特征向量，那么各种不同的机器学习、数据挖掘和统计方法都能运用。这个最新出现的方法，收集了图G的很多图的不变量，形成一个由图不变量组成的特征向量XG。当图G非常复杂时，XG要近似估计图的拓扑的维数是非常大，影响了许多挖掘方法的计算问题。为了减少这些问题，新的方法引入了核函数$K(XGx,XGy)$和$\phi:XG\to H$的映射由内值$<\phi(XGx);\phi(XGy)>$表示K。K表示图Gx和图Gy的相似程度，H是Hilbert空间。由于内值$<\phi(XGx);\phi(XGy)>$不能直接由计算H里的向量导出，这种方法能避免一些策略性的问题。而且因为相似度K是标量值，稀少分配问题也消失了。这个方法的缺点是这些核不能有效地计算出来，在最近的一些研究中提出了一些合适的可供选择的方法和相似措施。

2.4 挖掘度量

用于挖掘图的子结构的度量和平常数据挖掘相似，所选择的度量依赖于挖掘方法的目标和约束。基于图挖掘最常用的“支持度”的定义和菜篮子分析一样[4]。给定图数据集D，子图Gs的支持度$\sup(Gs)$定义为

$\sup(Gs)=D$图集中子图Gs个数/D图集中图的总数

如果图Gsy是图Gsx子图，这个度量反映单调性$\sup(Gsx)\leqslant\sup(Gsy)$。在一些挖掘方法中，只要给定一“最小支持度”值minsup，子图{Gs}支持度就大于最小支持度minsup。

那些常用在机器学习领域的其它挖掘度量也应用在基于图的数据挖掘方法中，尤其如信息熵、信息增益、吉尼指数和最小描述长度(MDL)[5,6]。

3 图的挖掘方法

在基于图的数据挖掘中，上面提到的子图同构问题必须用有效的搜索来解决。粗略地说，当前使用的有四种搜索方法，根据搜索的完全性分为：启发式搜索方法和完全搜索方法；根据子图同构匹配的观点分为：直接和间接的匹配方法。间接匹配方法不能解决子图同构问题，但在一些相似性的措施下可解决子图相似问题。

3.1 基于贪心搜索的方法

Subdue系统和GBI使用的是贪心搜索方法，并且在图的匹配问题上使用的是不精确的图的匹配算法，以避免图的同构问题的高度复杂性。它们都从图中抽取出了公共子结构，但它们进行的是不完全搜索，有可能导致最佳的子结构丢失。

Subdue系统处理的是连通图[7]，并且这个连通图的顶点和边都要加上标签。Subdue系统寻找的子图结构是能够最大压缩输入图G的子图结构，压缩的方法使用了MDL原则。可以把找到的子图结构看作一个概念。Subdue系统使用的算法是柱型搜索算法。算法开始的时候，有待扩展的子图结构仅仅只是输入图中的一个顶点，然后这个顶点随着算法不断扩展。在每一步扩展中，对每一个子图结构都会计算$I(S)+I((G)S)$[8]，在这个公式中，S是被发现的子图结构，G是输入的图，$I(S)$是必须用于编码被发现的子图结构的bits的数量。在用子图结构S对原图G进行压缩以后，用$I((G)S)$表示必须用于编码被压缩后的图G的bits的数量。如果哪个子图结构的$I(S)+I((G)S)$最小，那么这个子图结构就作为最优子结构保存起来。这个算法是完全贪心的，是没有回溯的。因为柱型搜索的最大柱宽已经预定好了，所以有可能丢失最佳的子结构。Subdue的一个特征是：它的子图匹配算法是不精确匹配算法，所以它允许在子图匹配中出现轻微的图的扭曲。

3.2 归约数据库方法

给定一数据集，用归纳决策树学习、菜篮子分析[9]和ILP

(下转第157页)

后，不用全部从头开始计算两点距离，可以大大减少运行时间。

(3)距离的选取非常重要，实验中我们采用改进的距离公式。可以根据我们挖掘数据的目标，将所感兴趣的离群数据值加权。

实验说明了用本文的算法既保证了得到的结果相当接近于全局最优解，又保证了能非常快速地得到结果。即使对于样本的个数为1010、维数为18这样的高维大样本，用本文的算法也能快速得到比较满意的结果。

结束语 自离群数据的概念提出以来，在数据挖掘领域一直聚集了大量的研究人员从事离群数据挖掘技术的理论和实际应用的研究。本文在分析了传统的离群数据挖掘算法优点和缺点的基础上，针对非线性高维数据集中的离群点侦测问题，提出一种基于局部线性加权的离群点检测算法。该算法利用LLE算法的思想寻找样本数据的内在嵌入分布，并通过距离公式和离群点权值判别式进行权值数据判定，根据权值的大小标识出数据集中的离群点。特别是对于一些线性不可分的数据集，在运用传统离群点检测算法失败的情况下，本文提出的算法仍然能取得良好的离群点检测效果。

参考文献

[1] 夏火松. 数据仓库与数据挖掘技术. 北京：科学出版社，2004

[2] Barnett V , Lewis T. Outliers in Statistical Data[M] . New York : John Wiley and Sons , Inc , 1994

[3] Franco P, Ian S M. Computational Geometry : an Introduction [M] . New York : Springer - Verlag , 1988

[4] Edwin K M, Raymond N T. Algorithms for Mining Distance - Based Outliers in Large Datasets[R]//Proceedings of the 24th International Conference on Very Large Data Bases. New York:Morgan Kaufmann , 1998:392 -403

[5] Markus B M, Peter K H, Raymond N T , et al . LOF : Identifying Density - based Local Outliers[R] //Chen W, Naughton J F , Bernstein P A, eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas , Texas: ACM Press , 2000: 93 - 104

[6] Spiros P, Hiroyuki K, Phillip G B. LOCI : Fast Outlier Detection Using the Local Correlation Integral [R]//Proceedings of the 19th International Conference on Data Engineering. 2003: 315 -326

[7] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications[A]//Haas L M, Tiwary A, eds. Proc. ofthe ACM SIGMOD International Conference on Management of Data[C]. Seattle: ACM Press, 1998: 94-105

[8] Roweis S T , Saul L K. Nonlmensionality Reduction by Locally Linear Embedding [J]. Science , 2000 , 290 (22)

[9] Knorr E, Ng R. Finding Intensional Knowledge of Distance—based Outliers[C]//Scotland: Proc of the 25th VLDB Conference. Edinburgh, 1999: 211-222

[10] E M, Ng R T. Algorithms for Mining Distance—based Oufliers in Large Datasets//Proc. of Int. Conf. Very Large Data—bases (VLDB'98). C J, New York, 1998: 392-403

(上接第153页)

等挖掘方法产生一些归纳规则、一些关系规则和一些模式。这些归纳规则被存于数据库中，根据数据库里先前产生结果用精确表示查询条件而设计的查询语言对数据库进行查询。这个框架适合基于图的数据挖掘。用基于图的数据挖掘方法预先产生子图中的子图和关系，并把它存于归约数据库中，用致力于数据库的查询语言去查询子图和它们之间的关系。这种方法的优点是加快图挖掘的操作，这是因为子图和它们的关系的基本模式已经提前产生了。潜在的缺点是计算量大，提前产生和保存归纳模式的内存大。在一些工作中，这种方法常和下面提到的完全层次搜索方法联合使用。

MolFea系统使用的是归约数据库方法，MolFea系统使用的是level-wise版本空间算法[10]。该方法虽然对原图执行了完全的路径搜索，但是它所挖掘的子结构被限制在图中的连续子结构的范围里。

3.3 归纳逻辑编程(ILP)方法

归纳逻辑程序设计(Inductive Logic Programming, ILP)是计算机科学一个重要的分支学科，是归纳学习(Inductive Learning)与逻辑程序设计(Logic Programming)融合的产物。归纳方法是一种重要的机器学习方法，能够在特殊样例上产生泛化模型；逻辑程序设计使用一阶谓词逻辑表示实体间关系并在此基础上作演绎推理。

3.4 数学图论方法

数学图论方法[11]能够从图中抽取出频繁出现的公共子结构，并且能够执行完全搜索。因为数学图论方法要处理图的同构问题，解决图的同构问题的算法是多种多样的。而精确的图的匹配(同构)的难度相当于NP完全问题(NP完全问题是这样的问题：用确定性的算法在多项式时间内无法解决的问题)的难度，这时就要加上一定的限制条件以减少搜索空间，以便在一定程度上回避NP完全问题的难度，从而达到降低时间复杂度的目的。

4 目前图的数据挖掘技术中存在的问题

基于图的频繁子图挖掘的各种算法及其应用，目前在这方面的研究已经取得了一些成绩，并被应用到一些实际的系统中。

虽然目前存在的算法很多，但执行效率很不高，在处理大型数据库时仍然需要大量的时间和空间；求不精确的频繁子图问题在现实生活中广泛存在，但是目前的算法还停留在最初提出的SUBDUE上，有待进一步提高；在large graph型数据库中进行挖掘的算法也比较少，目前主要都是基于transaction型数据库的，因此对large graph型数据库的频繁子图挖掘算法的研究也是一个重要的研究方面；现实生活中有各种各样的图形，如有向图、无向图、加权图、无连通图等，但目前的算法大部分都是针对连通图的挖掘，对加权图等的挖掘算法很少，因此对加权图等的挖掘算法的研究也是一个重要的研究方向等。

参考文献

[1] Mckay B. Nauty users guide (version 1. 5). Technical Report, TR-CS-90-02. Department of Computer Science, Australian National University, 1990

[2] Inokuchi A, Washio T, Motoda H. Complete mining of frequent patterns from graphs: Mining graph data. Machine Learning, 2003,50:321-354

[3] De Raedt L, Kramer S. The levelwise version space algorithm and its application to molecular fragment finding//IJCAI'01: Seventeenth International Joint Conference on Artificial Intelligence, vol2, 2001:853-859

[4] Agrawal R, Srikant R. Fast algorithms for mining association rules//VLDB'94: Twentyth Very Large Dada Base Conference. 1994:487-499

[5] Yoshida K, Motoda H, Indurkhya N. Graph-based induction as a unified learning framework. J of Applied Intel,1994,4:297-328

[6] Cook J, Holder L. Substructure discovery using minimum description length and background knowledge. J Artificial Intel Research, 1994,1:231-255

[7] Cook D J, Holder B. Graph-based Data Mining. IEEE,2000,15 (2):32-41

[8] Cook J, Holder L B, Djoko S. Scalable discovery of informative structural concepts using domain know] edge. IEEE Expert, 1996,11(15)

[9] Agrawal R, Srikant R. Fast algorithms for mining association rules//VLDB'94: Twentyth Very Large Dada Base Conference. 1994:487-499

[10] De Raedt L, Kramer S. The levelwise version space algorithm and its application to molecular fragment finding//IJCAI' OI: Seventeenth International Joint Conference on Artificial Intelligence. vol2. 2001:853-859

[11] Inokuchi A, Washio T, Motoda H. Complete mining of frequent patterns from graphs : Mining graph data. Machine Learning, 2003,50:321-354