GC-BES:一种新的基于嵌入集的图分类方法

王桂娟[1,2] 印 鉴[1] 詹卫许[3]
(中山大学信息科学与技术学院 广州 510275)[1] (华南师范大学计算机学院 广州 510631)[2]
(南方电网信息中心 广州 510000)[3]

摘 要 已提出很多图分类方法。这些方法在挖掘频繁子图时,只考虑了子图的结构信息,没有考虑子图的嵌入信息。实际上,有些频繁子图挖掘算法在计算子图的支持度时,可以获得嵌入信息。在L-CCAM子图编码的基础上,提出了一种基于嵌入集的图分类方法。该方法采用基于类别信息的特征子图选择策略,充分利用嵌入集,在频繁子图挖掘过程中直接选择特征子图。通过实验表明,该方法是有效的、可行的。

关键词 频繁子图,图分类,图挖掘,特征选择
中图法分类号 TP311.13 **文献标识码** A

GC-BES:A Novel Graph Classification Approach Based on Embedding Sets

WANG Gui-juan[1,2] YIN Jian[1] ZHAN Wei-xu[3]
(College of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275, China)[1]
(School of Computer, South China Normal University, Guangzhou 510631, China)[2]
(Information Center, South Power Grid, Guangzhou 510000, China)[3]

Abstract Many graph classification approaches have been proposed. These approaches only look at the structural information of the pattern, but do not take advantage of the embedding information during mining frequent subgraph. In facts, in some efficient subgraph mining algorithms, the embedding information of a pattern can be maintained. A graph classification approach was presented. Based on L-CCAM coding, it uses a feature subgraph selection strategy based on label information to select the feature subgraph, while making full use of embedding sets to directly generate feature subgraph in mining frequent subgraph. Experiment results show that it is effective and feasible.

Keywords Frequent subgraph, Graph classification, Graph mining, Feature selection

图,作为一种通用的数据集结构,应用在许多科学领域。可以用图来表示化合物结构、3-D蛋白质结构、程序调用图、社会网络等数据对象之间的各种复杂关系。人们迫切需要设计图分类模型来对图数据进行分类,如生物学家可以通过图分类来预测蛋白质具有某种功能;化学家通过图分类可以预测药物有毒;程序员通过识别出程序流图中的特征子图,可以找到程序中的Bug[1]。

在图数据分类中,最直接的方法[2,3]是:首先,利用频繁子图挖掘算法[4-8]在训练图集中挖掘出频繁子图,然后利用某些特征选择策略,从频繁子图中选出特征子图,构造特征向量,最后用分类模型进行分类。该方法的主要缺点是利用挖掘算法产生的频繁子图数目巨大,甚至达到指数级。大量的频繁子图既阻碍了特征子图的选择,又大大地降低了挖掘算法的效率。

为了解决上述瓶颈问题,最近的研究从挖掘全部频繁子图转向挖掘有意义的、重要的、有分辨力的子图,即直接从训练图集中挖掘出分辨子图,采用分类模型进行分类。Leap算法[9]采用该方法的图分类算法,直接从训练图集中挖掘出最佳的子图模式;文献[10]采用基于模型的搜索树,直接挖掘出分辨子图;算法gPLS[11]采用有偏最小平方回归直接挖掘特征子图;graphSig[12]采用在每个结点进行随机游动,把图转变成向量,再把图划分成组,使得在相同组中的图具有相似向量的方法,来挖掘特征子图。

这些图分类方法在挖掘频繁子图的过程中,只考虑了子图的结构信息,没有考虑子图的嵌入信息。但在诸如FFSM[7]子图挖掘算法中,图模式的嵌入信息是可以获得的。本文提出了一种基于嵌入集的图分类方法GC-BES(Graph Classification Based on Embedding Set),其在采用带有图标记的L-CCAM编码的基础上,利用FFSM算法[6]的连接和扩展技术,在挖掘频繁子图的过程中利用嵌入集,采用基于类别信息的特征子图选择策略[13],直接生成特征子图,最后用SVM分类器进行分类。本文的主要贡献如下:

1)在CCAM[1]编码的基础上,提出了带有图标记的L-CCAM编码;

到稿日期:2011-07-06 返修日期:2011-09-29
王桂娟(1973-),女,博士生,讲师,主要研究领域为图数据挖掘、图数据管理,E-mail:wgjgood_ok@126.com;印 鉴(1968-),男,博士,博士生导师,主要研究领域为数据库、数据仓库、人工智能、数据挖掘;詹卫许(1974-),男,博士,工程师,主要研究领域为无线通信、数据库、数据仓库的设计与实现、数据挖掘。

2)频繁子图挖掘过程中,在模式增长的过程中考虑了子图的结构信息,在选择特征子图的过程中则考虑了子图的嵌入信息;

3)采用基于类别信息的特征子图选择策略,在频繁子图挖掘过程中直接生成特征子图;

4)实验表明,本算法是可行的和有效的。

1 预备知识及问题定义

在详细介绍 GC-BES 方法之前,首先给出有关的基本概念和问题定义。

定义 1(标号图) 标号图 G 是一个五元组 $G=(C,V,E,\Sigma,\lambda)$,其中 C 是 G 的类别标记,比如 $C=\{+(正类),-(负类)\}$,V 是顶点集合,E 是边的集合,Σ 为类标记的集合,映射 $\lambda:(V\cup E)\to\Sigma$,即映射 λ 给每个顶点和每条边分配一个标记。每个图都有唯一从 1 开始算起的 ID。在每个图中,每个顶点也都有唯一的从 1 开始算起的 ID。

例如图 1,图数据集中有两个图,它们有各自的 ID,+1 和 −1(其中 +,− 为图的类标记)。每个顶点以(顶点 ID:顶点标记)表示。

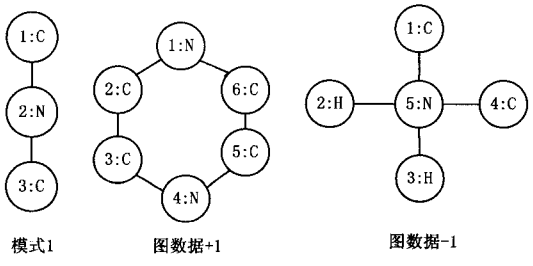

图 1 图和子图模式举例

定义 2(子图同构) 给定两个图 $G=(C,V,E,\Sigma,\lambda)$ 和 $G'=(C,V,E,\Sigma,\lambda)$,一个从 G 到 G' 的子图同构是一个单射函数 $f:V\to V'$,其满足:(1) $\forall u\in V,\lambda(u)=\lambda'(f(u))$;(2) $\forall(u,v)\in E,(f(u),f(v))\in E'$ 且 $\lambda((u,v))=\lambda'((f(u),f(v)))$。

如果存在从 G 到 G' 的子图同构,则称 G 是 G' 的子图,G' 是 G 的超图,记为 $G\subseteq G'$。如果 $G\subseteq G'$ 并且 $G\neq G'$,称 G 是 G' 的真子图。

例如,在图 1 中,模式 1 是图数据+1 和图数据−1 的一个子图。

定义 3(支持度) 给定图数据库 G_s,模式 p 在 G_s 中的支持度定义为

$$SUP(p,G_s)=\frac{\text{包含图模式 } p \text{ 的图的个数}}{G_s \text{ 中图的总个数}}$$

定义 4(频繁子图) 给定图数据库 G_s 和最小支持度阈值 min_sup,如果子图 p 满足 $SUP(p,G_s)\geqslant$min_sup,则称子图 p 为 G_s 中的频繁子图。

定义 5(图分类问题) 输入一个图集合 $G_s=\{(G_i,C_i)\mid i=1,2,\cdots,n\}$,其中 G_i 是一个图数据,C_i 为 G_i 所属的类别。输出一个分类模型,用来判断新图所属类别。

为了便于描述,本文只讨论二元分类问题。对 GC-BES 方法稍加改动,就可以将其用于多元分类问题。

2 L-CCAM 编码

频繁子图挖掘算法中,图模式的规范化编码有最小 DFS 编码[6]和 CAM 编码[7]。文献[1]在 CAM 编码的基础上,提出了 CCAM 编码。在这些编码中的图数据是没有区分类别信息的。本文在 CCAM 编码的基础上,提出了带有图类别信息的 L-CCAM 编码。

定义 6(L-嵌入) 给定一个子图单射 $f:V(g)\to V(g')$,顶点集 $\{f(u)\mid u\in V(g)\}$ 就是 g 在 g' 中的一个 L-嵌入。由于可能存在多个子图单射,故 g 在 g' 可能有多个 L-嵌入。

例如,在图 1 中,模式 1 在图数据+1 中有两个 L-嵌入,分别是{1,2,6}和{3,4,5}(按升序排列),在图数据−1 中有一个 L-嵌入{1,4,5}。那么,模式 1 共有 3 个 L-嵌入编码:〈+1,1,2,6〉,〈+1,3,4,5〉和〈−1,1,4,5〉。

定义 7(L-嵌入集) 一个子图模式 p 在一图数据集合 G_s 中的 L-嵌入的集合用 L_p 表示,即 $L_p=\{e\mid e$ 是模式 p 在 G_s 中的嵌入$\}$。

例如图 1,模式 1 在图数据+1 和图数据−1 构成的图数据集中的 L-嵌入集是{〈+1,1,2,6〉,〈+1,3,4,5〉,〈−1,1,4,5〉}。

定义 8(L-邻接矩阵) 给定子图同构映射 f 及其模式 p 的嵌入编码 B,p 的邻接矩阵 M 是一个 $|V|\times|V|$ 矩阵,其中 V 为模式 p 的顶点集。矩阵 M 的元素满足

$$M[i,j]=\begin{cases}B[i]\text{的标记}, & i=j\\ B[i]\text{和 }B[j]\text{之间边的标记}, & i\neq j\end{cases}$$

图 2 表示对应模式 p 的 3 个不同的 L-嵌入编码的 3 个邻接矩阵。

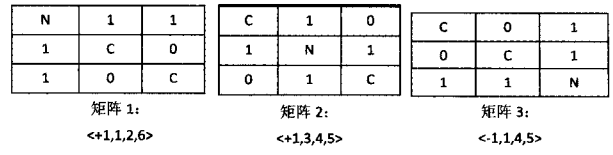

图 2 模式 C-N-C 的 3 个邻接矩阵

定义 9(L-矩阵编码) 一个子图模式 p 的矩阵编码是 p 的邻接矩阵的下三角元素按照行顺序所形成的字符串。

例如,图 2 中矩阵 1,2,3 的矩阵编码分别是 N1C10C,C1N01C 和 C0C11N。

定义 10(L-条件规范化邻接矩阵,L-CCAM 矩阵) 给定一个图数据集合,其中每个图数据都有唯一的 ID,一个子图模式 p 的条件规范化邻接矩阵是模式 p 按字典排序最小嵌入编码所对应的矩阵。

例如图 1,在图数据+1和图数据−1构成的图数据集中,模式C-N-C的条件规范化邻接矩阵就是图 2 中的矩阵 1。

定义 11(L-CCAM 编码) 给定一图数据集合,其中每个图数据有唯一的 ID,子图模式 p 的 L-CCAM 编码就是对应于模式 p 的 L-CCAM 矩阵的矩阵编码。

例如,在图 1 中,给定由图数据+1 和图数据−1 构成的图数据集,模式 1 的 L-CCAM 编码是 N1C10C。给定一个图数据集合,两个同构的子图模式一定具有相同的 L-CCAM 编码(参见文献[1]的证明)。

3 GC-BES 分类方法

3.1 特征子图选择策略

文献[13]提出了基于类别信息的特征子图选择策略:只有那些只在正类或反类中存在的独有频繁子图和在正类中的

支持度与在负类中的支持度相差很大的显著频繁子图才对分类起重要的预测作用。同时，在二元分类问题中，正类的数量通常远远小于反类的数量。例如，文献[2]中 CA 和 CI 的分类问题，正类 CA 的数量只是反类 CI 数量的 1%。为此，本文采用文献[13]的特征选择策略，为了提高效率，只选择正类中的独有频繁子图和显著频繁子图作为特征子图。下面给出相应的独有频繁子图和显著频繁子图的定义。

定义 12（独有的频繁子图） 给定两个图数据库 G_s^1 和 G_s^2，如果一个频繁子图 p 满足：p 在 G_s^1 中出现而不在 G_s^2 中出现，或者 p 在 G_s^1 中不出现而在 G_s^2 中出现，则称频繁子图 p 为 G_s^1 或 G_s^2 的独有频繁子图。

定义 13（显著的频繁子图） 给定两个图数据库 G_s^1、G_s^2 和整数 ∂，如果一个频繁子图 p 的支持度满足

$SUP(p,G_s^1)/SUP(p,G_s^2)\geqslant\partial$ 或 $SUP(p,G_s^1)/SUP(p,G_s^2)\leqslant 1/\partial$

则称频繁子图 p 为显著的频繁子图，∂ 称为显著比。此定义给出了在正类和负类同时具有但支持度差别较大的特征子图所满足的条件。

算法 1 给出了选择正类的独有频繁子图和显著频繁子图的伪代码。

算法 1 FeatureSelect(p,L_p,λ)

输入：子图模式 p 以及相应的嵌入集 L_p，整数 λ

输出：正类的独有频繁子图和显著频繁子图

1)$R^+_{Only}\leftarrow\Phi$，$R^+_{Discri}\leftarrow\Phi$

2)扫描 p 的嵌入集，判断 p 是否在 G_f^-

3)若不存在，则 $R^+_{Only}\leftarrow R^+_{Only}\cup\{p\}$ /*选择正类中独有的频繁子图*/

4)若存在，则计算并判断 $\frac{|L_p^+|}{|G^+|}/\frac{|L_p^-|}{|G^-|}>\lambda$ 是否成立 /*$|L_p^+|$，$|L_p^-|$ 分别为模式 p 在正类图数据和负类图数据中嵌入的个数，$|G^+|$，$|G^-|$ 分别为正类和负类中图数据的个数*/

5)若成立，则 $R^+_{Discri}\leftarrow R^+_{Discri}\cup\{p\}$，/*选择正类中显著的频繁子图*/

6)输出 R^+_{Only}，R^+_{Discri}

L-嵌入集中每个 L-嵌入都有其所在原图数据的类标记，上述算法第 2)、第 3)步，扫描 p 的 L-嵌入集，只需查看有没有带(−)标记的嵌入。如果没有，则表明 p 为正类独有的频繁子图模式；如果扫描 p 的 L-嵌入集，查看到有带(−)标记的嵌入，说明该子图模式 p 在负类图数据中存在。这时，再根据 p 的嵌入集计算 $|L_p^+|$，$|L_p^-|$，然后判断 p 是否是正类中显著频繁子图，如算法的第 5)、第 6)步。在此过程中，只需一次扫描频繁模式 p 的 L-嵌入集，不需要进行多次的子图同构测试，就能选择出特征子图，提高了效率。特征子图的生成可以直接整合到 FFSM 算法中，从而得到算法 2(GBES_FFSM 算法)，其伪代码见 3.2 节。

3.2 GBES_FFSM 算法

本文采用 FFSM[7] 算法来产生候选子图。FFSM 算法的基本思想是使用 FFSM-Join 和 FFSM-Extension 来产生候选子图，并通过相应的剪枝策略来删除重复的候选子图，通过扫描嵌入集计算候选子图的支持度，得到频繁子图。故通过 FFSM 算法进行频繁子图挖掘时，频繁子图的嵌入信息是可以获得的。由于本文采用的编码方法为 L-CCAM 且需要输出 L-嵌入集，因此不能直接采用 FFSM 算法，对其进行修改，得到 GBES_FFSM 算法。GBES_FFSM 算法的伪代码如算法 2 所示，在算法 2 中调用了 FeatureSelect()。

算法 2 GBES_FFSM(U,W,O)

输入：U，次优 L-CCAM 列表，W，频繁连同子图的 L-CCAM 集合，以及 O，频繁子图相应的 L-嵌入集

输出：正类的独有频繁子图和显著频繁子图

1)for X∈Udo

2)if(X is L-CCAM)then

3)W←W∪{X}，C←Φ，$O_c\leftarrow\Phi$

4)for Y∈U do

5)C←C∪FFSM-Join(X,Y) /*通过对 X，Y 进行 Join 操作，获得的候选子图 c*/

6)$O_c\leftarrow O_c\cup$FFSM-Join-Embeddig(X,Y)/* O_c 为 X，Y 进行连接后得到的候选子图 c 的 L-嵌入集*

7)如果候选子图 c 频繁，则 FeatureSelect (c,O_c,λ)

8)end for

9)C←C∪FFSM-Extension(X)，$O_a\leftarrow\Phi$/*通过对 X 进行 Extension 操作，获得的候选子图 a*/

10)$O_a\leftarrow O_a\cup$FFSM-Extension-Embeddig(X)/* O_a 为对 X 进行 Extension 操作后得到的候选子图 a 的 L-嵌入集*/

12)如果候选子图 a 频繁，则 FeatureSelect (a,O_a,λ)

13)GBES-FFSM(C,W,O)

14)end if

15)end for

算法中第 4)、5)步，对具有公共核的频繁子图 X，Y 进行 FFSM 算法中的 Join 操作，获得候选子图 c，$O_c\leftarrow O_c\cup FFSM\text{-}Join\text{-}Embeddig(X,Y)$ 表示对 X，Y 进行连接后得到的候选子图 c 的 L-嵌入集 O_c；第 7)步通过扫描 O_c，计算 C 的嵌入个数，判断 c 是否是频繁的，如果是频繁子图，则调用 FeatureSelect (c,O_c,λ) 选择出特征子图；第 9)、10)步，对频繁子图 X 进行 Extension 操作，获得候选子图 a，$O_a\leftarrow O_a\cup FFSM\text{-}Extension\text{-}Embeddig(X)$ 表示对 X 进行 Extension 操作后得到的候选子图 a 的 L-嵌入集；第 11)步通过扫描 O_a，计算 a 的嵌入个数，判断 a 是否频繁，如果是频繁子图，则调用 FeatureSelect (a,O_a,λ)，选择出特征子图。

算法 3 给出了完整的 GC-BES 算法。

算法 3 GC-BES(G_s^+，G_s^-，∂)

输入：正类图数据集合 G_s^+，负类图数据集合 G_s^-，最小支持度阈值 ∂ $(0<\partial<1)$

输出：正类的独有频繁子图和显著频繁子图

1)S←{频繁的顶点和频繁的边}，L←{频繁边的嵌入集}

2)P←{频繁的边}

3)GBES-FFSM(P,S,L)

4 实验分析

本文所有的算法都在 STL 库支持下用 C++实现，用带有○3 选项的 g++编译，用于实验的计算机具有 2.0 GHzCPU 和 2GHz 内存，运行于 WindowsXP 操作系统。本文采用 LIBSVM[14] 进行分类，取参数 $k=9$。为了与文献[2]中的分类方法比较，本文使用了与文献[2]相同的两个实验数据集：一个是 PTC 数据集(http://www.informatik.uni-freiburg.de/ml/ptc/)。对于该数据集，有 4 个二元分类问题，分别是 MM、FM、MR 和 FR；另一个是 NCI-HIV 化合物数据集(http://dtp.nci.gov/docs/aids_data.html)。该数据集中的所有化合物都被分成 3 类：CA(非常活跃)、CM(中等活跃)和 CI(不活跃)。本文也考虑下面的 3 个二元分类问题：1) CA/CM；2) (CA+CM)/CI；3)CA/CI。表 1 给出了这两个数据集的重要特征。

表 1　数据集 PTC 以及 NIC-HIV 的重要特征

	PTC	NIC-HIV	Class distribution	
# compounds	417	42682	PTC	
Avg. # vertices	25	46	MM	38.3%
Avg. # edges	26	48	FM	40.9%
# vertex labels	24	85	MR	44.2%
# edge lables	4	3	FR	34.4%
Max. # vertices	106	438	NIC-HIV	
Min. # vertices	2	2	CA/CM	28.1%
			(CA+CM)/CI	3.5%
			CA/CI	1%

文献[2]用 ROC 曲线度量分类方法的性能，为了能与文献[2]中分类方法进行比较，本文也用 ROC 曲线度量 G-BES 分类方法的性能。ROC 曲线在 x 轴上显示被错误分类的负类的比率，在 y 轴上显示被正确分类的正类的比率。一个分类器对应的 ROC 曲线面积越大，说明该分类器的分类性能越好。

图 3 给出了 GC-BES 分类方法对应的 ROC 曲线面积。为了与文献[2]中的分类方法比较，图 2 也显示了在文献[2]中给出的 ROC 曲线面积。在图 2 中，Topo 和 Geom 是文献[2]中的分类方法。Topo 利用 FSG[5] 频繁子图挖掘算法挖掘出频繁拓扑子结构来建立分类模型，而 Geom 则是利用 FSG[5] 频繁子图挖掘算法挖掘频繁几何子结构来建立分类模型。在图 2 中，给出分类方法在各种参数下的最好结果。通过比较 ROC 曲线面积可以看出，在 FM，FR 分类问题上，GC-BES 的分类性能要差于文献[2]中的分类方法，但在其他的 5 个分类问题上，GC-BES 的分类性能要明显优于文献[2]中的分类方法。

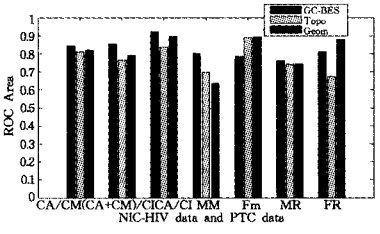

图 3　Topo，Geom 和 GC-BES 所对应的 ROC 曲线面积的比较

本文也在运行时间上与文献[2]进行了比较。本文的运行时间指在训练集中挖掘频繁子图，生成特征向量需要的时间。表 2 显示了在不同的支持度下，Topo，GC-BES 算法对 PTC 以及 NIC-HIV 数据集进行频繁子图挖掘所运行的时间。所有的运行时间是采用 5 折交叉验证取平均的时间。从表 2 可以看出，在 CA/CI，MM，FM，MR，FR 分类问题上，GC-BES 的运行时间要比文献[2]中的 Topo 快 2 倍多；而在 CA/CM 分类问题上，GC-BES 的运行时间要比文献[2]中的 Topo 快 20 倍多；在(CA+CM)/CI 分类问题上，GC-BES 的运行时间要比文献[2]中的 Topo 快 4 倍多。

表 2　在不同的∂下 Topo，GC-BES 算法对 PTC 以及 NIC-HIV 数据集进行频繁子图挖掘所运行的时间

	CA/CM (秒)	(CA+CM)/CI (秒)	CA/CI (秒)	MM (秒)	FM (秒)	MR (秒)	FR (秒)
支持度	10.0/10.0	10.0/5.0	10.0/10.0		3.0/3.0		
Topo	137	1016	392	211	72	66	231
GC-BES	6	249	171	95	35	30	102

GC-BES 的分类性能好主要得益于下面 3 个方面：

1)采用基于区分类别能力的高性能特征选择方法——基于类别信息的特征子图选择策略；

2)在频繁子图挖掘过程中既考虑了子图的结构信息，又考虑了子图的嵌入信息；

3)采用 FFSM[6]——深图优先搜索的频繁子图挖掘算法挖掘频繁子图，而文献[2]采用的 FSG 算法是广度优先搜索算法。

此外，GC-BES 的分类方法也容易被理解和使用。

结束语　本文提出了带有类别标记的 L-CCAM 编码，在此基础上给出了一种基于嵌入集的图分类方法 GC-BES。利用嵌入集，在频繁子图挖掘过程中直接产生正类的独有特征子图和显著频繁子图，然后利用 SVM 进行分类。实验结果显示，在对化合物数据分类时，GC-BES 在分类性能上优于图分类方法[2]。

参 考 文 献

[1] Jin Ning, Young C, Wang Wei. GAIA: Graph Classification Using Evolutionary Computation[C]// SIGMOD10. 2010

[2] Deshpande M, Kuramochi M, Karypis G. Frequent SubsStructure-based Approaches for Classifying Chemical Compounds [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(8): 1036-1050

[3] 刘勇，李建中，朱敬华. 一种新的基于频繁闭显露模式的图分类方法[J]. 计算机研究与发展，2005，44(7)：1169-1176

[4] Inokuchi A, Washio T, Motoda H. An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data[C]// 4th European Conference on Principles of Data Mining and Knowledge Discovery(PKDD 2000). Lyon, France, September 2000

[5] Kuramochi M, Karypis G. Frequent subgraph discovery [C]// Proceedings of the 2001 IEEE International Conference on Data Mining. 2001: 313-320

[6] Yan X, Han J. gSpan: Graph-based substructure pattern mining [C]// Proc. 2002 Int. Conf. on Data Mining (ICDM02). 2002: 721-724

[7] Huan J, Wang W, Prins J. Efficient mining of frequen subgraph in the presence of isomorphism[C]// Proceedings of the 3rd IEEE International Confrence on Data Mining (ICDM). 2003: 549-552

[8] Yan Xi-feng, Han Jia-wei. CloseGraph: Mining Closed Frequent Graph Patters[C]// Proc. of the Ninth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. ACM Press, Aug. 2003: 286-295

[9] Yan Xi-feng, Cheng Hong, Han Jia-wei, et al. Mining Significant Graph Patterns by Leap Search[C]// SIGMOD'08. 2008

[10] Fan Wei, Zhang Kun, Cheng Hong, et al. Direct Mining of Discriminative and Essential Frequent Patterns via Model-based Search Tree[C]// KDD. 2008

[11] Saigon H, Kraemer N, Tsuda K. Partial Least Squares Regression for Graph Mining [C] // Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD2008). 2008: 578-586

[12] Ranu S, Singh A K. GraphSig: A Scalable Approach to Mining Significant Subgraphs in Large Graph Databasws[C]// Proceedings of the 25th International Conference on Data Engineering (ICDE). April 2009

[13] 王桂娟，印鉴，詹卫许. 基于类别信息的特征子图选择策略[J]. 计算机科学，2011，42(8)

[14] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines[EB/OL]. http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001