二分类图上的非冗余协同图模式挖掘算法

王章辉 赵宇海 王国仁 李 源
(东北大学信息科学与工程学院 沈阳 110819)

摘 要 图模式广泛应用于构建高效图分类模型的特征空间识别. 协同图模式是一种内部节点高度相关的图结构,与普通图模式相比,协同图模式具有更高的区分能力,从而更加适用于分类模型的特征选择. 文中研究了从二分类图中挖掘非冗余协同图模式的问题,通过限制协同图模式的区分能力远远高于其所有子图模式的非冗余性质,大幅度减少了挖掘结果的数量,同时保留了具有强区分能力的协同图模式. 由于协同图模式理论上必须检测其所有子图是否满足约束条件,挖掘它们非常具有计算挑战性. 基于非冗余协同图模式的多种特性,提出相对应的削减规则;通过对区分能力的边界估计,提出两个快速检测非冗余协同图模式方法,在此基础上给出了一种高效的深度优先挖掘算法 GINS. 大量真实与合成数据集上的实验结果表明,GINS 算法明显优于其他两个代表性算法,作为图分类模型的分类特征时,非冗余协同图模式获得了较高的分类精度.

关键词 二分类图;非冗余;协同图模式;图分类;图挖掘;子图模式;分类器
中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2015.01434

Mining Non-Redundant Synergy Graph Patterns from Two Classes of Graphs

WANG Zhang-Hui ZHAO Yu-Hai WANG Guo-Ren LI Yuan
(*College of Information Science and Engineering, Northeastern University, Shenyang* 110819)

Abstract Graph patterns are widely used to define the feature space for building an efficient graph classification model. Synergy graph patterns refer to those graphs, where the relationships among the nodes are highly inseparable. Compared with the general graph patterns, synergy graph patterns which have much higher discriminative powers are more suitable as the classification features. This paper investigates the problem of mining non-redundant synergy graph patterns from two classes of graphs. By guaranteeing the property that the discriminative powers of synergy graph patterns are much higher than all their subgraphs, mining non-redundant synergy graph patterns can dramatically reduce the number of results and still capture the strong discriminative powers synergy graph patterns. However, finding all non-redundant synergy graph patterns is computationally challenging because all their subgraphs should theoretically be checked. Also, through studying the properties of non-redundant synergy graph patterns, the corresponding pruning techniques are proposed. Moreover, two fast synergy graph pattern detection methods are proposed based on the bound estimation of the discriminative powers. Based on those techniques, an efficient depth-first algorithm GINS is presented for this mining problem. Extensive experiments are conducted on a series of real-life and synthetic datasets. The results show that

收稿日期:2014-02-06;最终修改稿收到日期:2015-01-21. 本课题得到国家自然科学基金(61272182,61100028,61073063,61173030,61332014)、国家"八六三"高技术研究发展计划项目基金(2012AA011004)、国家杰出青年科学基金项目(61025007)、新世纪优秀人才支持计划(NCET-11-0085)、中央高校基本科研业务费(N130504001)资助.**王章辉**,男,1985 年生,博士研究生,主要研究方向为数据挖掘、生物信息. E-mail: wzh_neu@163.com. **赵宇海**(通信作者),男,1975 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为数据库、数据挖掘、生物信息. E-mail: zhaoyuhai@ise.neu.edu.cn. **王国仁**,男,1966 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为不确定数据管理、XML 数据管理、查询处理与优化、并行数据库系统、生物信息. **李 源**,男,1986 年生,博士研究生,主要研究方向为数据挖掘、生物信息.

GINS is more efficient than two representative competitors. Besides, when the non-redundant synergy graph patterns are considered, the classification accuracy is improved much.

Keywords　two classes of graphs; non-redundant; synergy graph patterns; graph classification; graph mining; subgraph pattern; classifier

1　引　言

图结构作为一种通用的表示不同对象之间复杂关系的数据结构，已广泛应用于多种跨学科领域，如生物信息学[1-2]、化学信息学[3-4]、药物信息学[5]等. 在这些科学应用中，图数据中隐含的模式信息可以用来帮助构建分类模型和理解分析这些复杂的数据结构. 例如，在化合物数据分析中，图模式可以帮助研究人员揭示化合物中具有化学毒性的结构部分[6]. 从图数据中进行图模式挖掘具有重要的研究意义和应用价值.

已经得到广泛研究的图模式挖掘主要包括频繁子图挖掘[7-8]和显著子图挖掘[9-11]. 频繁子图挖掘算法基于用户指定的支持度阈值和相对应的数据库，可以获得所有满足支持度阈值的频繁子图模式，但这些挖掘算法往往产生大量的甚至达到指数级的频繁子图，不但降低了挖掘算法的效率，而且难以对大量的挖掘结果进行深入分析和利用. 为了减少挖掘结果的数量，显著子图挖掘算法采取对模式搜索空间采样或者近似过滤的方法，仅关注一少部分满足确定显著性定义的图模式，这些方法虽然可以大规模减少输出模式的数量，但容易丢掉具有丰富研究意义的图模式.

图分类研究中，生物和医学数据被广泛作为分析对象[9-10,12-14]. 在这些应用领域中，图数据的所属类别很多情况下只有两类. 例如，在基于调控网络的致病基因识别中，生物专家通常关心的问题是，某些基因的交互作用是否与特定疾病的发生有关；在药物分子设计中，医疗专家的主要兴趣在于，某种药物分子结构是否对特定的疾病具有显著疗效. 在实际研究中，针对诸如此类的研究，大多通过收集两种不同类别的数据，即患者组(case)和对照组(control)数据，来进行对比分析. 如果将上例中的患者组标记为“＋”类，对照组标记为“－”类，我们称此类数据为二分类数据. 特别的，如上例所述的许多生物和医疗数据可以建模为图数据，我们称其为二分类图数据，其中的患者组称为正类图集，对照组称为负类图集. 二分类图是图分类研究中使用最广泛的基本数据类型之一，文献[9-10,12-14]等诸多研究工作都是基于该类数据展开的. 为了便于叙述，本文的研究工作也以该类数据为例展开. 在第 5 节中，进一步讨论了如何将本文的研究工作扩展至多类图数据的问题.

图分类研究工作中，区分子图挖掘[12-14]可以获得大量具有区分能力的图模式，帮助用户构建图数据分类模型. 现有的区分子图挖掘算法大多只关心图模式的区分能力，并以此作为唯一的选择标准，导致结果中存在一定的冗余模式. 其主要原因在于，忽略了图模式本身的结构信息(如协同交互等).

在生物信息学与生物医学研究中，协同交互现象大量存在[15-17]. 为了直观地理解这种现象，图 1 给出了一种在野生型酵母基因中存在的简单协同交互现象. 如图所示，两条不同的路径 $A\rightarrow B\rightarrow C$ 与 $X\rightarrow Y\rightarrow Z$($A,B,C$ 等可表示为不同的基因或者蛋白质)作用于相同的生物过程，影响野生型酵母细胞的存亡. 由于两条路径的并行结构，对任意单路径的基因扰乱(扰乱 A 或 Y)，都不会导致酵母细胞的死亡. 只有当同时对两条路径进行基因扰乱时(扰乱 A 与 Y)，才会导致酵母细胞死亡. 本例中的 A 和 Y 相对于酵母细胞类别(存活或死亡)就存在着协同交互关系，即二者通过紧密的联合作用影响酵母细胞的存活或死亡. 在实际应用中，基因或蛋白交互等生物调控网络通常用图结构数据来表示. 因此，发现图 1 所示的协同交互关系可以转化为图数据中的协同图模式挖掘问题.

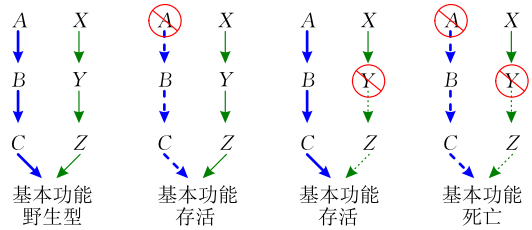

图 1　酵母基因中的协同交互示例

协同图模式从模式结构本身出发，要求其所有的子图模式都不具有超过其本身的表现性能. 协同的含义，从图的角度通俗来讲，是指只有当某些特定的图节点组合在一起，形成的图模式才具有更高的性能. 任意减少其中的节点或边，都将降低图模式本

身的表现性能.我们称这样的图模式为协同图模式.挖掘协同图模式可以帮助理解和解释图数据中潜在的性质.如前所述,在癌症致病基因识别研究中,可以根据患者组和对照组的基因交互数据,构建正类图集和负类图集,形成二分类图数据库.通过对二分类图数据进行协同图模式挖掘,找到与特定疾病相关的致病基因组,并从这些基因之间相互作用的角度解释疾病发生的原因.协同图模式可以为疾病的诊断、预防和治疗提供新的研究视角.因此,协同图模式挖掘具有重要的理论研究意义和实际应用价值.

协同图模式的表现性能根据实际应用可以有不同的表现形式,对于图分类应用,可以选择图模式的区分能力作为其表现性能的度量标准.区分能力的度量可以有多种方式,本文采用图模式的置信度作为区分能力的度量标准.由于协同图模式的特殊限制,使得传统的频繁子图挖掘和区分子图挖掘框架不再适用于协同图模式挖掘.本文首次提出非冗余协同图模式挖掘问题,为了进一步减少协同图模式的冗余结果,我们只关注那些比它们所有子图模式具有更高区分能力的协同模式作为挖掘结果,提出一种非冗余协同图模式挖掘算法 GINS(Graph MIning for Non-Redundant Synergy Patterns),用来实现二分类图上的协同图模式挖掘.由于协同图模式的特殊限制,理论上对于任意拥有 k 条边的一个图模式,必须检测 2^k 数量的子图模式才能确定其是否是协同图模式,这是一项十分浪费时间的工作.同时还需要面对的是置信度度量不具有反单调性质和如何使用挖掘出来的结果集构建一个高效的分类器问题,这使得提出的挖掘算法面对诸多挑战.

本文研究了从二分类图中挖掘非冗余协同图模式的问题,通过限制协同图模式的区分能力远远高于其所有子图模式的非冗余性质,大幅度减少了挖掘结果的数量,同时保留了具有强区分能力的协同图模式;基于非冗余协同图模式的多种特性,提出相对应的模式扩展削减规则,通过对区分能力的边界估计,提出两种快速检测非冗余协同图模式方法,并给出高效的深度优先挖掘算法 GINS;最后选择 top-k 个非冗余协同图模式构建图分类模型,大量实验证明了算法的高效性和有效性.

本文第 2 节介绍文章的相关工作;第 3 节给出问题的形式化描述;第 4 节介绍和分析提出的挖掘算法;第 5 节介绍算法的多类扩展;第 6 节给出实验结果与分析;第 7 节总结本文工作.

2 相关工作

近年来,图模式挖掘问题的相关研究一直备受众多研究者的关注.在数据挖掘领域,频繁子图挖掘问题早已吸引了众多研究者的目光,同时也产生了大量优秀的研究成果.文献[7]和文献[8]介绍了两种十分具有代表性的频繁子图挖掘算法.最近几年的研究热点已经从频繁图模式挖掘转移到显著图模式挖掘和区分图模式挖掘问题上,并取得了大量的研究成果.文献[9,11]采取对图模式搜索空间近似过滤和采样的方法,快速得到一部分显著图模式,从而大大减少了模式结果数量.文献[12-14]采用启发式的搜索策略,快速得到一组具有区分能力的图模式结果集.

文献[18]介绍了从频繁子图挖掘区分图模式的方法,虽然这种方法可以获得所有的区分图模式,但其十分浪费时间.文献[19]采用一定数量的对应组度量图模式的区分能力,可以获得理论上的最优结果.文献[10]使用相对高支持度阈值从小规模数据组中挖掘区分图模式.文献[9]基于结构近似导致区分能力近似的假设,大幅度削减模式搜索空间,快速得到挖掘结果.以上所有区分图模式挖掘算法都未考虑到挖掘结果出现冗余信息的情况,虽然非冗余的模式挖掘已经广泛应用于二进制数据和序列数据中[20-21],用来减少结果数量.到目前为止,挖掘非冗余区分图模式尚未得到广泛研究.

3 问题描述

本节首先介绍一些基础概念,接下来形式化描述二分类图上的非冗余协同图模式挖掘问题.

3.1 基础概念

本文主要考虑简单的连通无向标签图,通过简单修改,本文提出的算法也适用于其他类型的图数据.若无特殊说明,本文中所涉及的图均指简单的连通无向标签图.一个标签图 G 可以定义为一个四元组 $G=(V,E,\Sigma,F)$,其中,V 是顶点集合,$E\subseteq V\times V$ 是边集合,Σ 是标签集合,$F:V\cup E\rightarrow\Sigma$ 是一个函数,用来对图中顶点和边分配相应的标签.此外,一个图还可以属于唯一一个具体的类别.图 2 给出了一个二分类图数据库 D,其中 G_1,G_2,G_3 和 G_4 属于正类图集合,用 G^+ 表示;G_5,G_6,G_7 和 G_8 属于负类图集合,用 G^- 表示.如果图数据库 D 中所有图只属于正

类或者负类两个类别，我们称这样的图数据库 D 为二分类图数据库.

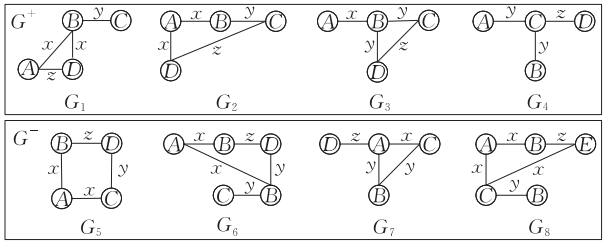

图 2　二分类图数据库 D

定义 1.　图同构. 给定两个图 $G_1=(V,E,\Sigma,F)$ 和 $G_2=(V',E',\Sigma',F')$，图同构是一个双射函数 $f:V\leftrightarrow V'$ 满足如下条件：

(1) $\forall u\in V, F(u)=F'(f(u))$；

(2) $\forall e_1=(u,v)\in E, e_2=(f(u),f(v))\in E'$ 并且 $F(e_1)=F'(e_2)$；

(3) $\forall e_2=(u,v)\in E', e_1=(f^{-1}(u),f^{-1}(v))\in E$ 并且 $F(e_1)=F'(e_2)$.

则称这两个图同构.

定义 2.　子图同构. 给定两个图 $G_1=(V,E,\Sigma,F)$ 和 $G_2=(V',E',\Sigma',F')$，如果存在一个映射函数 $f:V\to V'$ 和 G_2 的一个子图 S，满足 f 是从 G_1 到 S 的同构，则称 f 是一个从 G_1 到 G_2 的子图同构，也称 G_1 子图同构于 G_2.

如果存在一个从 G_1 到 G_2 的子图同构，称 G_1 是 G_2 的子图，G_2 是 G_1 的超图或者 G_2 支持 G_1，表示为 $G_1\subseteq G_2$. 如果 $G_1\subseteq G_2$ 且 $G_1\neq G_2$，称 G_1 是 G_2 的真子图，G_2 是 G_1 的真超图.

定义 3.　支持度. 给定一个二分类图数据库 $D=G^+ +G^- =\{G_1,G_1,\cdots,G_n\}$ 和一个图模式 g，支持图模式 g 的集合记作 $D_g=\{G_i\mid g\subseteq G_i, G_i\in D\}$. 图模式 g 的支持度为 $|D_g|$，表示为 $supp(g)$；图模式 g 的频繁度为 $|D_g|/|D|$，表示为 $freq(g)$.

例 1.　对于图 2 给出的二分类图数据库 D，支持图 3 中图模式 g 的图集合为 $\{G_1,G_2,G_3,G_6\}$，相对应的支持度 $supp(g)=4$，频繁度 $freq(g)=0.5$.

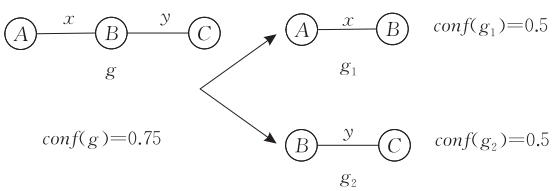

图 3　协同图模式例子

对于一个给定的二分类图数据库 D，$supp(g,G^+)$ 表示图模式 g 在正类图集合中的支持度；$supp(g,G^-)$ 表示图模式 g 在负类图集合中的支持度；$supp(g)$ 表示图模式 g 在整个二分类图数据库 D 上总的支持度，显然 $supp(g)=supp(g,G^+)+supp(g,G^-)$. 图模式 g 的区分能力可以用其在二分类图数据库中的置信度来表示，记作 $conf(g)$，如式(1)所示：

$$conf(g)=\frac{\max\{supp(g,G^+),supp(g,G^-)\}}{supp(g)} \quad (1)$$

图模式 g 的区分能力其实就是其在正类或者负类图集合中支持度的最大值与相对于图数据库总的支持度的比值. 置信度值在二分类图数据库上是一个介于 0.5 到 1 之间的一个实数，取值等于 0.5 表示没有任何的区分能力，取值越大，表明区分能力越强，其最大值可取为 1.

例 2.　对于图 2 给出的二分类图数据库 D，可以计算图 3 中图模式 g 在该数据库上的置信度值. 其中 $supp(g,G^+)=3$，$supp(g,G^-)=1$，所以置信度值 $conf(g)=3/(3+1)=0.75$.

对于二分类图数据库 D 中任何图模式，都存在一个与之对应的置信度值来表示它的区分能力，我们就可以根据置信度度量标准定义协同图模式.

定义 4.　协同图模式. 给定一个二分类图数据库 $D=G^+ +G^- =\{G_1,G_1,\cdots,G_n\}$ 和一个图模式 g，图模式 g 被称为协同图模式当且仅当式(2)成立，

$$conf(g)-\max_{\forall g'\subset g}(conf(g'))>0 \quad (2)$$

其中 g' 表示图模式 g 的任意一个真子图.

例 3.　对于图 2 给出的二分类图数据库 D，图 3 中图模式 g 就是一个协同图模式，因为它所有的真子图的置信度值都不大于它本身的置信度值.

3.2　问题定义

本文采用置信度值作为度量图模式的区分能力，在进行问题定义之前，我们首先引出强区分图模式和弱区分图模式的定义，接下来定义非冗余协同模式，最后给出本文的全局目标.

定义 5.　强区分模式. 给定一个二分类图数据库 $D=G^+ +G^- =\{G_1,G_1,\cdots,G_n\}$，一个图模式 g 和一个用户指定强区分阈值 β，$0.5<\beta\leq 1$，如果 $conf(g)\geq\beta$，就称图模式 g 为一个强区分模式.

强区分模式表示该模式具有较强的区分能力，用户可以根据指定的阈值选择自己需要的结果.

定义 6.　弱区分模式. 给定一个二分类图数据库 $D=G^+ +G^- =\{G_1,G_1,\cdots,G_n\}$，一个图模式 g 和一个用户指定弱区分阈值 α，$0.5<\alpha<\beta$，如果 $conf(g)\leq\alpha$，就称图模式 g 为一个弱区分模式.

增大 β 的取值，可以得到拥有更高区分能力的图模式，减小 α 的取值，可以获得拥有更弱区分能力的图模式，本文要求 α 的取值小于 β 的取值.

定义 7. 非冗余协同图模式. 给定一个二分类图数据库 $D=G^+ + G^- = \{G_1, G_1, \cdots, G_n\}$，一个图模式 g，当以下 3 条标准满足时，该图模式就是一个非冗余协同图模式.

(1) $freq(g) \geq \delta$，其中 δ 是用户给定的频繁度支持阈值，$0 < \delta \leq 1$；

(2) 图模式 g 是一个强区分模式；

(3) 图模式 g 是一个协同图模式并且它的所有真子图都是弱区分模式.

二分类图上的非冗余协同图模式挖掘问题可以描述如下：给定一个二分类图数据库 $D=G^+ + G^- = \{G_1, G_1, \cdots, G_n\}$，3 个用户输入参数，频繁度支持阈值 δ、弱区分阈值 α 和强区分阈值 β，任务就是从二分类图数据库 D 中挖掘所有满足定义的非冗余协同图模式.

4 GINS 算法

本节介绍如何从一个二分类图数据库中快速挖掘非冗余协同图模式算法. 首先简单了解一下著名的深度优先搜索(Depth First Search，DFS)编码搜索树框架[8]，本文采用此编码搜索方式挖掘所有非冗余协同图模式；接下来介绍非冗余协同图模式的性质和与之对应的削减规则，用来削减 DFS 搜索树上不满足约束条件的分支；同时使用置信度值边界估计进一步加速搜索过程的完成. 论文提出两种协同图模式检查策略，加快对非冗余协同图模式的判定工作. 最后，综合以上技术，提出高效的挖掘算法 GINS.

4.1 DFS 编码搜索树

所有的分支界限搜索算法都需要采用一种具体的搜索框架，用来确保所有的目标模式都可以被找到. 在搜索框架的基础上，可以应用各种削减策略加速搜索过程的完成. DFS 编码树搜索框架是一种高效的并得到广泛应用的搜索框架. 在 DFS 编码搜索树中，可以使用最小 DFS 编码实现图同构问题的检测. 在 DFS 编码搜索树中，如果两个图 g_1 和 g_2 图同构，那么它们一定具有相同的最小 DFS 编码. 根据这条优秀性质，可以把图挖掘问题转化成为满足一定约束条件的序列挖掘问题，这样可以大大降低图挖掘问题的复杂度. 本文采用这种著名的搜索框架，实现非冗余协同图模式挖掘. GINS 算法根据用户给定的频繁度支持阈值 δ，按照深度优先搜索方法构建 DFS 编码树. 图 4 给出基于图 2 二分类图数据库 D 满足频繁度支持阈值 $\delta=0.25$ 的 DFS 搜索树，图 4 DFS 搜索树中包含所有满足频繁度支持阈值的图模式. 在图 4 中，搜索树上每一个节点代表一个图模式，节点上方出现的一组数字表示该节点被搜索的顺序编号和支持该节点所表示的图模式的图编号，即冒号之前的数字表示被搜索的顺序编号，本文用该顺序编号表示其代表的图模式；圆括号内的数字表示图 2 二分类数据库 D 中支持该节点所代表图模式的图编号.

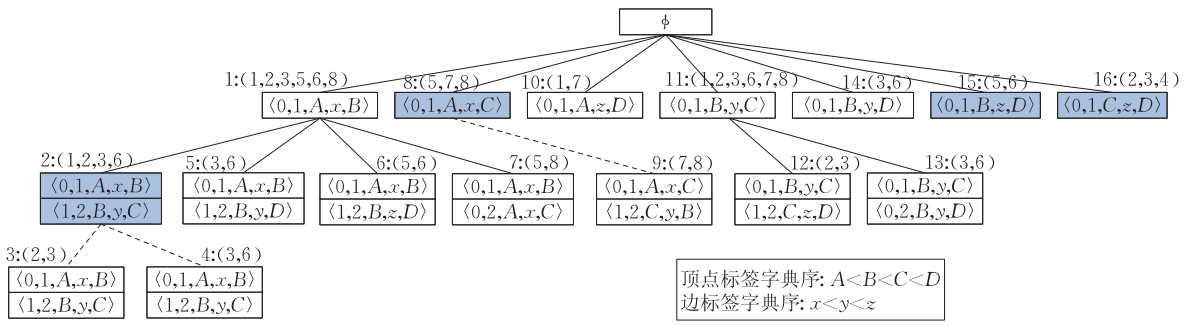

图 4 图 2 中二分类数据库 D 的 DFS 编码搜索树

例 4. 对于图 4 中节点编号为 8 的图模式 $\langle 0,1,A,x,C\rangle$，圆括号内的数字(5,7,8)表示在图 2 的二分类数据库 D 中，图 G_5、G_7 和 G_8 支持节点编号为 8 的图模式. 在图 4 中，灰颜色的节点代表满足用户给定的频繁度支持阈值 $\delta=0.25$、弱区分阈值 $\alpha=0.6$ 和强区分阈值 $\beta=0.75$ 约束的非冗余协同图模式结果.

4.2 非冗余协同图模式性质和削减规则

本小结主要介绍非冗余协同图模式的一些重要性质，根据这些重要性质给出相应的削减规则削减图模式搜索空间. 结合置信度值的上界估计，共同加速图模式搜索过程的完成.

性质 1.　对于一个给定的图模式 g，其满足非冗余协同图模式定义，则它所有的超图模式 $p(g\subset p)$ 必然不满足非冗余协同图模式定义.

证明.　由于图模式 g 是一个非冗余协同图模式，其置信度值 $conf(g)\geq\beta>\alpha$. 对于它所有的超图模式 p 都存在也必然存在一个真子图 g 不是一个弱区分模式，不满足定义 7 中第 3 条标准，则图模式 g 的所有超图 p 模式都不是非冗余协同图模式. 证毕.

例 5.　对于图 4 DFS 编码搜索树中节点编号为 2 的节点，由于其本身已经是一个非冗余协同图模式，以它为根的所有分支根据性质 1 都可以安全删除，图 4 中用虚线表示删除的分支部分.

对于 DFS 编码搜索树中所有的候选图模式，为了判断其是否是一个非冗余协同图模式，必须验证其是否满足定义 7 的 3 条标准. 接下来，文章引出置信度值的上界估计，可以实现对候选图模式的置信度值在具体计算之前进行预估计，减少了具体计算时间. 随着图模式的增长尽管置信度既不满足单调性质也不满足反单调性质，但图模式的支持度和频繁度满足反单调性质. 在 DFS 编码搜索树中，随着图模式的逐渐扩展，扩展后的图模式的频繁度值越来越小，直到不满足用户指定的频繁度支持阈值后停止扩展. 当算法进行 DFS 编码搜索扩展图模式 g 时，可以根据用户指定的最小频繁度支持阈值估计图模式 g 的所有超图模式置信度的上界. 如果这个上界都无法达到定义 7 的第 2 条标准，那么算法不再需要对图模式 g 进行扩展搜索，也就是说以图模式 g 为根节点的所有分支都可以安全地从搜索空间删除，因为它的所有分支节点都不可能是一个强的区分模式.

性质 2.　对于一个给定的 DFS 编码搜索树中的图模式 g 和用户给定的频繁度支持阈值 δ，它的支持度值和置信度值分别表示为 $supp(g)$ 和 $conf(g)$. 对于其任何超图模式 $p(g\subset p)$，其置信度 $conf(p)$ 值的上界可以通过式(3)进行计算得到.

$$conf(p)_{g\subset p}\leq conf(g)\frac{supp(g)}{\delta\times|D|}\tag{3}$$

证明.　由于 DFS 编码搜索树中的图模式的支持度随着图模式扩展满足反单调性质，则支持度值满足 $supp(g)\geq supp(p)$，$\max\{supp(g,G^+),supp(g,G^-)\}\geq\max\{supp(p,G^+),supp(p,G^-)\}$. 因此，

$$conf(p)_{g\subset p}=\frac{\max\{supp(p,G^+),supp(p,G^-)\}}{supp(p)}\leq\frac{\max\{supp(g,G^+),supp(g,G^-)\}}{supp(p)}.$$

由于满足用户指定的频繁度支持阈值的最小支持度为 $\delta\times|D|$，因此，根据式(1)可得到

$$conf(p)_{g\subset p}\leq\frac{\max\{supp(g,G^+),supp(g,G^-)\}\times supp(g)}{supp(p)\times supp(g)}\leq conf(g)\frac{supp(g)}{\delta\times|D|}.$$

证毕.

性质 2 给出一个 DFS 编码搜索树中图模式扩展的置信度上界估计公式，如果这个上界估计值都无法达到定义 7 的第 2 条标准，对于正在扩展的图模式的所有分支都可以安全删除，可以减少大量运算时间. 根据非冗余协同图模式的定义和性质，提出 3 条削减规则，用来加快搜索过程的完成.

削减规则 1.　给定 DFS 编码搜索树上一个非冗余协同图模式 g，DFS 编码搜索树上以图模式 g 为根的所有分支都可以安全删除.

证明.　由性质 1 可以直接导出. 证毕.

削减规则 2.　给定 DFS 编码搜索树上一个图模式 g，如果 $freq(g)\geq\delta$，并且 $\alpha<conf(g)<\beta$，则 DFS 编码搜索树上以图模式 g 为根的所有分支都可以安全删除.

证明.　由于 $\alpha<conf(g)<\beta$，图模式 g 既不是一个强区分图模式，同时也不是一个弱区分图模式，根据定义 7 的第 3 条标准，图模式 g 在 DFS 编码搜索树中的所有超图模式都不可能是非冗余协同图模式. 证毕.

削减规则 3.　给定 DFS 编码搜索树上一个图模式 g，如果 $freq(g)\geq\delta$，$conf(g)\geq\beta$，并且图模式 g 不是一个非冗余协同图模式，则 DFS 编码搜索树上以图模式 g 为根的所有分支都可以安全删除.

证明.　由于 $conf(g)\geq\beta$，并且图模式 g 不是一个非冗余协同图模式，根据定义 7 的第 3 条标准，图模式 g 在 DFS 编码搜索树中的所有超图模式都不可能是非冗余协同图模式. 证毕.

根据以上提到的非冗余协同图模式的性质和相应的削减规则，接下来简单描述一下一个具体图模式的搜索过程. 给定 DFS 编码搜索树上一个图模式 g，首先需要计算的是图模式 g 的频繁度，如果其频繁度小于用户指定的频繁度支持阈值 δ，根据频繁度的反单调性质，DFS 编码搜索树上以图模式 g 为

根的所有分支都可以安全删除.否则的话,需要进一步计算图模式 g 的置信度值,如果置信度值小于用户指定的弱区分阈值 α,需要进一步在 DFS 编码搜索树上按照深度优先方式搜索图模式 g 的分支;如果置信度值大于 α 但是小于用户指定的强区分阈值 β,根据削减规则 2,DFS 编码搜索树上以图模式 g 为根的所有分支都可以安全删除.如果置信度阈值大于 β,此时,图模式 g 就成为非冗余协同图模式的一个候选,接下来需要检测其所有真子图模式是否都是弱区分模式,也就是看是否满足定义 7 的第 3 条标准.具体的检测方法将在接下来的一节中详细介绍.在确定图模式 g 成为一个非冗余协同图模式之后,根据削减规则 1,DFS 编码搜索树上以图模式 g 为根的所有分支都可以安全删除.搜索过程递归进行,直到找到所有满足要求的非冗余协同图模式,搜索过程结束.

4.3 非冗余协同图模式检测策略

本小节将介绍两个非冗余协同图模式的检测策略,在介绍检测方法之前,给出候选非冗余协同图模式的置信度下界估计.

对于一个给定的候选非冗余协同图模式 g,最直接和明显的检测策略就是计算其所有真子图模式的置信度,判断是否满足弱区分模式的约束条件.但这是一种十分耗费时间的方法,对于任意拥有 k 条边的一个图模式,必须检测完 2^k 数量的子图模式才能确定其是否是一个非冗余协同图模式,大量的子图同构检测使得这种方法的效率极端低下.

由于图模式 g 的频繁度随着图模式的扩展满足反单调性质,也就是说相对于图模式 g 的频繁度值,其真子图模式拥有更高的频繁度值,随着真子图模式边的个数越来越少,其对应的频繁度值越来越大.图模式 g 的真子图模式中拥有最大频繁度的子图模式必然是只有一条边的单边图模式,并且单边图模式的支持度值是很容易计算出来的.因此,可以在进行对候选非冗余协同图模式 g 检测之前,可预先估计一下其所有真子图模式置信度的下界,如果这个下界值都比用户指定的弱区分阈值 α 值大的话,根据定义 7 的第 3 条标准,图模式 g 必然不是一个非冗余协同图模式,根据削减规则 3,DFS 编码搜索树上以图模式 g 为根的所有分支都可以安全删除,从而节省大量的检测时间.

性质 3. 对于一个给定的候选非冗余协同图模式 g,它的支持度值和置信度值分别表示为 $supp(g)$ 和 $conf(g)$.对于其任意真子图模式 $g'(g'\subset g)$,其置信度 $conf(g')$ 值的下界可以通过式(4)进行计算得到.其中 g'' 表示候选非冗余协同图模式 g 的任意拥有单边的图模式.

$$conf(g')_{g'\subset g}\geq\frac{conf(g)\times supp(g)}{\max_{g''\subset g}\{supp(g'')\}} \tag{4}$$

证明. 由于 DFS 编码搜索树中的图模式的支持度随着图模式收缩满足反单调性质,则支持度值满足 $supp(g)\leq supp(g')$,$\max\{supp(g,G^+),supp(g,G^-)\}\leq\max\{supp(g',G^+),supp(g',G^-)\}$.因此,

$$conf(g')_{g'\subset g}=\frac{\max\{supp(g',G^+),supp(g',G^-)\}}{supp(g')}\geq\frac{\max\{supp(g,G^+),supp(g,G^-)\}}{supp(g')}.$$

由于图模式收缩时满足反单调性质,$supp(g')\leq\max_{g''\subset g}\{supp(g'')\}$,根据式(1)可得到

$$conf(g')_{g'\subset g}\geq\frac{\max\{supp(g,G^+),supp(g,G^-)\}\times supp(g)}{supp(g')\times supp(g)}\geq\frac{conf(g)\times supp(g)}{\max_{g''\subset g}\{supp(g'')\}}.$$

证毕.

由于单边图模式 g'' 的支持度很容易计算得到,在进行候选非冗余协同图模式 g 检测之前,可以使用式(4)对其所有真子图模式的置信度下界进行估计,如果这个下界值都大于用户指定的弱区分阈值 α,根据定义 7 的第 3 条标准,图模式 g 确定不是一个非冗余协同图模式;同时根据削减规则 3,删除不必要的搜索分支.否则的话,需要按照一定的检测策略计算检测图模式 g 的所有真子图模式是否满足定义 7 的第 3 条标准.接下来,我们提出两种检测候选非冗余协同图模式的检测方法.

(1) 自顶而下检测策略(Top-down).对于任意一个拥有 k 条边的候选非冗余协同图模式,自顶而下的检测策略可以通过对候选非冗余协同图模式进行删边操作来完成,通过对候选非冗余协同图模式每次删除一条不同的边,可以获得所有 $k-1$ 条边的真子图模式集合.然后再通过对 $k-1$ 条边的真子图模式进行删一条边操作,同样可以得到所有 $k-2$ 条的真子图模式集合.不断对新产生的真子图模式集合进行删除一条边的操作,最终可以得到该候选协同图模式的所有真子图模式.每当得到一个新的真子图模式,检测其是否是一个弱区分图模式.如果确定其是一个弱区分图模式,继续产生新的真子图模式进行检测.直到确定所有的真子图模式都是弱区

分图模式，则该候选非冗余协同图模式为一个真正的非冗余协同模式. 否则，结束检测过程，确定其不是一个非冗余协同图模式，并根据削减规则 3，以其为根的所有分支都可以安全删除.

(2) 自底而上检测策略(Bottom-up). 对于任意一个拥有 k 条边的候选非冗余协同图模式，自底而上的检测策略采用对候选非冗余协同图模式的 k 条边进行模式扩展方法，产生该候选非冗余协同图模式的所有真子图模式，详细的模式扩展方法可以参考文献[7]. 当扩展产生一个新的真子图模式时，同样需要对其进行检测判断，确定其是否是一个弱区分图模式，接下来的处理过程同自顶而下的检测策略相同，具体削减可以参照上述过程.

算法 1.　GINS 算法.

输入：二分类图数据库 D；3 个参数 δ、α 和 β

输出：所有非冗余协同图模式集合 S

1. 扫描 D 计算所有频繁边.
2. 删除 D 中不频繁的边和顶点.
3. 计算 D 中频繁边的置信度值.
4. 删除置信度值大于 α 小于 β 的频繁边.
5. 添加置信度值大于 β 的边到集合 S.
6. 添加置信度值小于 α 的边到候选集合 S^1.
7. 初始化集合 S^1 中所有边的最小 DFS 编码.
8. FOR S^1 中每个 DFS 编码 s
9. 　调用过程 1 *pattern_Mining*(D,δ,α,β,S,s).
10. 输出非冗余协同图模式集合 S.

最后，综合 DFS 编码搜索框架与非冗余协同图模式相关性质、削减规则和置信度边界估计，给出二分类图上的非冗余协同模式挖掘算法 GINS 如算法 1 所示. 其中算法 1 包含模式挖掘的具体过程，参见过程 1.

过程 1.　*pattern_Mining*(D,δ,α,β,S,s).

输入：二分类图数据库 D；3 个参数 δ、α 和 β；结果集 S；DFS 编码 s

输出：所有非冗余协同图模式集合 S

1. IF s 不是最小 DFS 编码，THEN
2. 　过程 1 结束.
3. IF s 的置信度值小于等于 α，THEN
4. 　IF s 的超图模式置信度上界值大于 β，THEN
5. 　　在 D 中搜索 s 的所有超图模式.
6. 　　扩展 s，满足频繁度阈值的 s 重新调用过程 1.
7. IF s 的置信度值大于 α 并且小于 β，THEN
8. 　过程 1 结束.
9. IF s 的置信度值大于等于 β，THEN
10. 　IF s 的真子图置信度下界值小于等于 α，THEN
11. 　　检测所有的真子图模式是否满足定义 7.
12. 　IF 存在真子图模式不满足定义 7，THEN
13. 　　过程 1 结束.
14. 　ELSE
15. 　　把 s 加入输出结果集合 S.

4.4　GINS 的非冗余性

非冗余协同图模式要求图模式本身是一个强区分模式，其所有真子图都是弱区分模式. 也就是说，一个非冗余协同图模式 g 的任一超图 p 都不可能是非冗余的协同图模式(性质 1). 否则，超图模式 p 必然会存在一个真子图 g 是强区分模式，与非冗余协同图模式的定义矛盾. 图模式 g 被称为非冗余协同图模式的直观意义在于，其本身已经具有很强的区分能力，足以用于区分不同类的图集. 此时，即使其超图模式 p 也具有很强的区分能力，相对于 g 而言也是冗余的，没有必要出现在结果集中. 以图 4 为例，编号为 8 的节点对应的子图模式是一个非冗余协同图模式，以置信度为 1 的水平区分了负类图集中的 G_5，G_7 和 G_8，具有很强的区分能力. 其超图(节点 9)虽然也可以区分负类图集中的 G_7 和 G_8，置信度同样为 1，但相对于节点 8 而言，其超图模式(节点 9)是冗余的.

LEAP[9] 算法仅以区分能力作为目标函数，采用结构近似削减快速地从搜索空间挖掘满足区分能力要求的图模式. 虽然能较快地获得满足区分能力要求的图模式，但不能保证每次得到的区分模式在区分不同类别时不存在冗余信息. 在 LEAP 算法下，图 4 中编号为 8 和 9 的节点所表示的子图模式都可能成为其挖掘结果，因为它们都具有很强的区分能力. 但很明显，按照定义 7，节点 9 对应的子图模式相对于节点 8 来说是冗余的. GraphSig[10] 算法侧重于从大规模图数据中挖掘区分图模式. 但与 LEAP 类似，其仍然将图模式的区分能力作为唯一的度量标准，引导图模式的挖掘，忽视了结果的冗余. 因此，图 4 中编号为 8 和 9 的节点所对应的子图模式也都可能存在于 GraphSig 的挖掘结果中. 也就是说，GraphSig 存在着和 LEAP 算法相似的问题.

本研究在基于 DFS 编码搜索树的查找过程中，通过削减规则保证了结果集中的每个图模式 g 是一个强区分模式，而其任意子模式 g' 均为弱区分模式，避免了结果冗余的现象. 因此，根据定义 7，由 GINS 算法产生的结果是非冗余的.

5　GINS 算法的多类扩展

如前所述，虽然本研究中的非冗余协同图模式

挖掘算法 GINS 是基于二分类图提出的，但在实际应用中，只需对 GINS 进行简单扩展，就可完全适用多类图数据.

给定一个含有 m 个类标签的多分类图数据库 D，分别用 $G^1,G^2,\cdots,G^m$ 表示 m 个不同类标签对应的图集. 根据支持度的定义，图模式 g 在不同类别图集下的支持度分别用符号 $supp(g,G^1),supp(g,G^2),\cdots,supp(g,G^m)$ 来表示. 与二分类图中的情况类似，图模式 g 的区分能力同样可以用其置信度来表示，即选择在不同类别图集下的最大置信度度量图模式 g 区分其他图集的区分能力. 通过对式(1)的简单修改，图模式 g 在多类图数据集上的置信度如式(5)所示：

$$conf(g)=\frac{\max\{supp(g,G^1),supp(g,G^2),\cdots,supp(g,G^m)\}}{supp(g)} \tag{5}$$

图模式 g 在多类图数据库上的置信度取值介于 0 到 1 之间，取值越大，表明该模式的区分能力越强.

给定如式(5)所示的多类图数据中图模式的置信度计算方法后，GINS 算法在多类图数据上的挖掘过程同二分类图数据上的挖掘过程一样. 在基于 DFS 编码搜索树的遍历中，通过式(5)计算置信度的取值范围，选取对应的削减规则缩减搜索空间，避免冗余模式成为候选图模式. 同时，利用提出的两种快速协同图模式检测方法对候选节点进行判断，保证挖掘结果的正确性和非冗余性.

6 实验结果与分析

本节为了验证所提非冗余协同模式挖掘算法的高效性和有效性，在多个数据集上进行了大量实验. 算法采用基于标准模板库(STL)的 C++编程实现，实验环境为 HP PC 机器，2.33 GHz 双核处理器，2 GB 内存，Windows 7 操作系统.

6.1 实验数据集

实验部分，采用了多个真实数据集并通过组合真实数据集形成合成数据集对所提出的算法进行多角度分析和评价. 第一个数据集为化合物集合数据集(AIDS)，是一个用来测试对艾滋病病毒有无抑制作用的化合物集合，包含 43 905 个化合物，这些化合物根据实验结果对艾滋病的抑制程度可以分为活跃(CA)、中性(CM)和不活跃(CI) 3 种类别. 其中 CA 含有 422 个化合物，CM 含有 1081 个化合物，剩余化合物属于 CI 类别. 该化合物数据集可以从 DTP-NCI/NIH① 获得.

实验同时采用了 PubChem② 数据库上的一系列图结构数据集. PubChem 数据库是一个维护良好的记录生物活动的平台，包括各种分子生物活性检测，抗癌生物检测等记录. 其中每个数据集根据其抗癌检测可以分为活跃和不活跃两类，表 1 对 11 个美国国家癌症研究所(NCI)检测进行简单介绍. 从表 1 可以看出每一种癌症检测活跃化合物的数目都远远小于不活跃化合物的数目，比例大约只占百分之五. 由于实验对比 GraphSig[10] 算法只能处理相同数目的正类集合与负类集合的数据集，为统一起见，所有实验均采用此种等比例的数据集. 实验中对每一个癌症检测数据集都随机选择 1000 个活跃化合物组成二分类数据集的正类数据集合；随机选择 1000 个不活跃化合物组成二分类数据库的负类数据集合，这样就重新组成 11 个规模为 2000 的二分类图数据集，接下来的实验中，采用重新组合的数据集来度量算法的性能.

表 1 抗癌检测数据集

检测 ID	癌症描述	活跃数目	不活跃数目
1	Non-Small Cell Lung	2047	38410
33	Melanoma	1642	38456
41	Prostate	1568	25967
47	Central Nerv Sys	2018	38350
81	Colon	2401	38236
83	Breast	2287	25510
109	Ovarian	2072	38551
123	Leukemia	3123	36741
145	Renal	1948	38157
167	Yeast anticancer	9467	9467
330	Leukemia	2194	38799

6.2 算法的效率

本小节主要从算法运行时间方面考察提出的非冗余协同图模式挖掘算法的效率，两个具有代表性的图模式挖掘算法 GraphSig[10] 和 LEAP[9] 作为与提出的 GINS 算法进行比较的对比算法. 其中 GINS 算法默认采用自顶而下的检测策略记录算法执行时间. 由于 LEAP 算法每次迭代运行只产生一个显著图模式，所以使用其多次迭代产生图模式直到覆盖所有输入图的时间总和作为算法的响应时间. GINS 算法的默认参数设置为 $\delta=0.05$，$\alpha=0.6$ 和 $\beta=0.75$，接下来的实验中当改变其中一个参数时，剩余参数选择默认设置值；同时设置 LEAP 算法的默认

① http://dtp.nci.nih.gov
② http://pubchem.ncbi.nlm.nih.gov

参数值 $\sigma=0.05$ 和 GraphSig 算法的默认参数值 $minSup=0.1\%$ 和 $maxPvalue=0.1$.

首先，使用 AIDS 数据集中所有的 CA 化合物与随机选取与 CA 相等数目的 CM 化合物组成一个新的二分类数据集，接下来的实验采用此数据集来度量 3 个算法的效率. 图 5 显示了在改变频繁度 δ 参数时 3 个算法执行时间的对比，由于 LEAP 算法不受频繁度的约束，算法运行时间仅与参数 σ 相关，所以频繁度的改变对于 LEAP 算法运行时间毫无影响. 从图 5 中还可以看出 GINS 算法受频繁度改变影响较大，随着频繁度值逐渐变大，GINS 算法的运行时间迅速减少，而 GraphSig 算法则表现得不太明显.

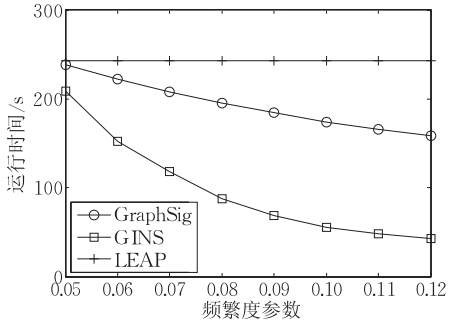

图 5　改变频繁度参数下的运行时间

GINS 算法采用两种检测非冗余协同图模式的策略，一种是自顶而下(Top-down)策略；另外一种是自底而上(Bottom-up)策略，接下来通过实验对比两种策略下 GINS 算法受弱区分参数和强区分参数的影响. 从图 6 中可以看出，随着弱区分参数 α 的逐渐增大，基于两种策略的 GINS 算法的运行时间都逐渐增长，明显可以看出，自顶而下的检测策略优于自底而上的检测策略. 从图 7 中同样可以看出，基于两种策略的 GINS 算法的运行时间都随着强区分参数 β 的增大而迅速减少，相比较来说，自底而上的检测策略随着强区参数 β 的改变表现得更显著一些.

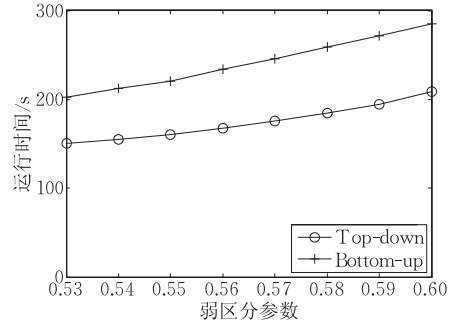

图 6　改变弱区分参数下的运行时间

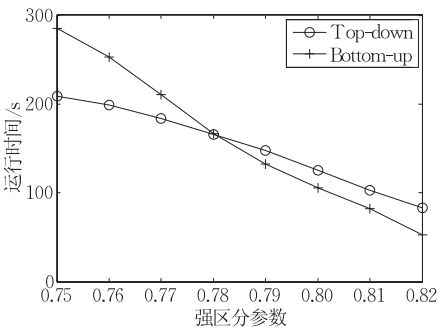

图 7　改变强区分参数下的运行时间

在进行算法的可扩展性分析时，数据集选择从表 1 的 Yeast anticancer 数据集按照等比例的活跃和非活跃数目随机抽取，形成新的测试数据集，供算法进行可扩展性分析. 从图 8 可以看出，随着数据集规模的逐渐变大，3 个算法的运行时间都逐渐增加，说明 3 个算法都明显受数据规模的影响. 其中 LEAP 算法受数据集规模变大的影响更加显著，而 GINS 算法的运行时间同数据集规模的增大接近线性增长，从而可以说明提出的非冗余协同图模式挖掘具有良好的可扩展性.

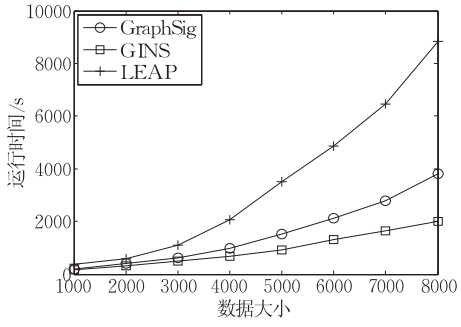

图 8　改变数据集规模下的运行时间

图 9 显示了 3 个算法在表 1 合成的 11 个数据集上的运行效率，从图中可以看出 GINS 算法在 11 个数据集上的运行时间都优于其他两个算法，这得益于 GINS 算法中大量的削减规则和边界估计技术，大大缩短了算法的执行时间.

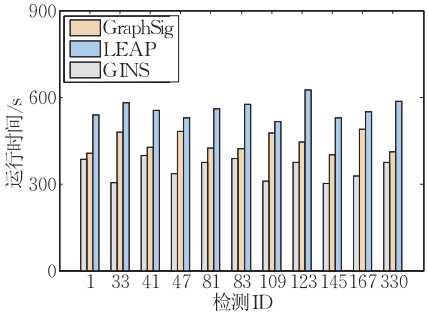

图 9　多个数据集上的运行时间

6.3 算法的有效性

本小节中，我们主要从挖掘算法产生的图模式在分类模型构建的应用角度，对算法进行综合的评价分析，接下来的实验都是在表 1 合成的 11 个数据集上完成的. 为了构建一个好的分类器，把每个合成数据集按照等比划分为训练集和测试集两部分. 分别从 GINS、GraphSig 和 LEAP 这 3 个算法在训练集上运行产生的图模式中选择 top-k 个得分最大的图模式来代表训练集，其中 LEAP 算法为了产生 top-k 个代表训练集的图模式，必须进行算法的多次运行. 在获得 top-k 个代表图模式之后，使用参数 C 介于$[2^{-10},2^{10}]$ 的支持向量机 LIBSVM① 构建分类器，通过构建的分类器在测试数据集上的分类精度对比，证明算法的有效性.

为了避免单次实验的偶然性，所有实验均重复进行 5 次平均计算. 首先考察 3 个算法代表训练集所选择的 top-k 个图模式的平均数目，当测试集中的所有图都至少是一个 top-k 中图模式的超图时，测试集才可以被挖掘到的 top-k 个图模式所代表. 图 10 给出了 3 个算法代表测试集所需要的模式个数的平均值，从中可以看出 GINS 算法需要较少的模式数量就可以代表训练集，这是因为 GINS 所产生的图模式都是非冗余的，而另外两个算法产生的图模式包含一定数量的冗余图模式，必然会需要更多的图模式个数来代表训练集.

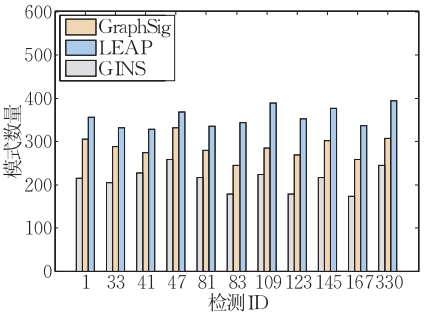

图 10 代表测试集模式的平均数量

实验还观察了 3 个算法选择的 top-k 个模式的平均大小，图 11 给出实验结果，从图中可以看出，GINS 算法平均产生的代表图模式规模更小，这也进一步解释了 GINS 算法比其他两个算法运行更快的原因. 因为目标模式规模小，算法中需要较少的模式扩展，GINS 算法更高效.

为了进一步说明 3 个算法所挖掘图模式的显著性，分别对所选择的 top-k 个代表图模式进行 G-test[9] 平均得分计算，计算结果如图 12 所示. 从图中可以看出 GINS 算法产生的图模式具有更高的显著性.

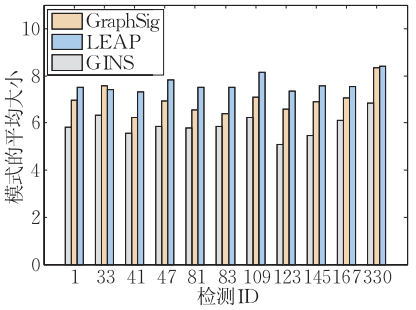

图 11 代表测试集模式的平均大小

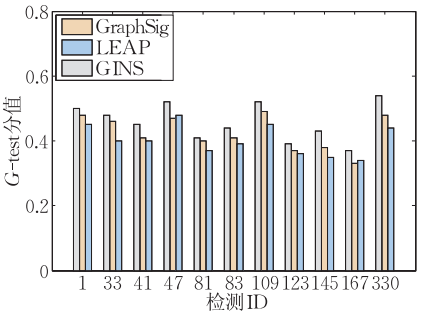

图 12 代表测试集模式的 G-test 平均得分

最后，从分类应用的角度出发，进一步考察 3 个算法的有效性. 基于选择的 top-k 个图模式和支持向量机算法，可以构建相对应的分类器，对数据集中的测试集进行分类评价. 使用接收者操作特征(ROC)曲线下的面积(AUC)[9] 作为分类精度的度量标准. AUC 是一个介于 0 到 1 之间的一个实数，数值越大，说明分类器的分类精度越高，一个好的分类器产生的 AUC 值接近于 1.

表 2 给出使用 3 种算法产生图模式构建分类器的平均分类精度. 从表 2 中可以看出，基于 GINS 算法产生的图模式的平均分类精度明显高于 LEAP 算法，并且在绝大多数数据集上的分类精度高于 GraphSig 算法，通过多次实验分类精度的误差分析可以看出，基于 GINS 算法的分类器更加稳定.

表 2 AUC 平均得分

检测 ID	LEAP	GraphSig	GINS
1	0.79±0.05	0.80±0.03	0.82±0.03
33	0.76±0.03	0.81±0.02	0.82±0.02
41	0.76±0.04	0.76±0.03	0.78±0.03
47	0.77±0.02	**0.80±0.02**	**0.80±0.02**
81	0.77±0.04	0.77±0.02	0.79±0.03
83	0.76±0.04	**0.77±0.02**	**0.77±0.02**
109	0.77±0.02	0.79±0.02	0.80±0.01
123	0.73±0.05	0.74±0.03	0.76±0.02
145	0.78±0.02	0.80±0.03	0.81±0.02
167	0.72±0.03	0.73±0.04	0.75±0.03
330	0.81±0.03	**0.84±0.02**	**0.84±0.02**
平均值	0.76±0.04	0.78±0.03	0.80±0.02

① http://www.csie.ntu.edu.tw/~cjlin/libsvm/

6.4　算法的多类扩展

算法的多类扩展实验中，通过抽取和组合真实数据集合成多个多类数据集. 第 1 个多类数据集为 3 类数据集，通过对化合物集合(AIDS)中的活跃、中性和不活跃三类化合物进行随机抽取，从每一类中随机抽取 422 个化合物组成一个新的数据集，命名为 Class3 数据集. 第 2 个多类数据集为 4 类数据集，通过对表 1 中抗癌检测 ID 为 1，33，41 和 47 数据集的活跃化合物进行随机抽取，从每个数据集的活跃化合物中都随机抽取 500 个化合物组合成一个新的数据集，对该组合数据集命名为 Class4. 同 Class4 数据集的组合方法相同，从表 1 中抗癌检测 ID 为 81，83，109，123 和 145 数据集的活跃化合物中都随机抽取 500 个化合物组合成第 3 个多类数据集，对该数据集命名为 Class5.

由于 LEAP 算法与 GraphSig 算法只能处理两类数据，在处理多类数据时，需要对数据集中多个类别进行两两组合，对每种组合分别执行算法，处理时间为每个组合的执行时间之和. GINS 算法在处理多类数据时默认采用自顶而下的检测策略记录算法的执行时间，参数设置为 $\delta=0.05$，$\alpha=0.5$ 和 $\beta=0.75$，LEAP 算法与 GraphSig 算法的默认参数设置同处理二分类数据时相同.

首先，通过实验分析 GINS 算法在多类数据集上受不同参数的影响. 图 13 给出了 GINS 算法在 3 个多类数据集上受弱区分参数 α 变化的影响情况，从图 13 中可以看出，随着参数 α 的逐渐增大，GINS 算法在多类数据集上的运行时间也逐渐增加，同二分类数据集上的运行时间趋势相同. 同时可以看出，多类数据的类别越多，GINS 算法的运行时间越长，受参数 α 的变化的影响越明显. 图 14 显示了 GINS 算法在 3 个多类数据集上受强区分参数 β 的影响情况，从图 14 中可以看出，随着参数 β 的逐渐增大，GINS 算法的运行时间迅速减少，这一现象也同二分类数据上相同.

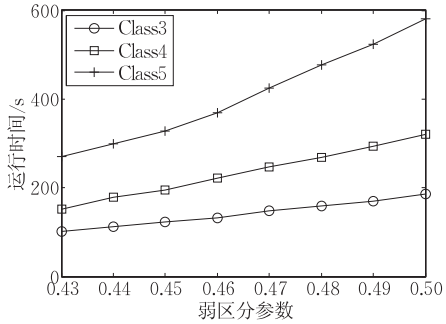

图 13　多类数据改变弱区分参数下的运行时间

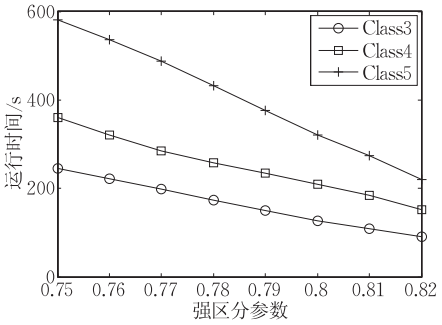

图 14　多类数据改变强区分参数下的运行时间

接下来分析 GINS 算法在多类数据集上的运行效率. 图 15 给出了 3 个算法在 3 个合成的多类数据集上的运行时间，从图中可以看出，GINS 算法在 3 个数据集上都显著比其他两个算法快，这是因为 GINS 算法只需要进行一次执行就可以完成多类数据分类的要求，而 LEAP 算法和 GraphSig 算法在处理多类数据时必须进行多个类别两两组合，分别执行算法，从而导致运行时间显著增加.

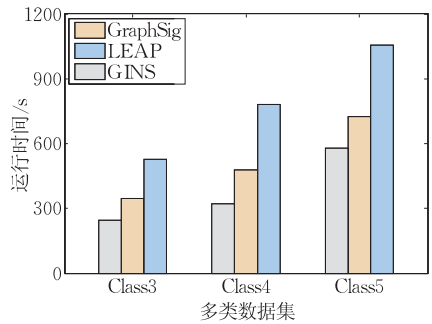

图 15　多类数据集上的运行时间

最后对 3 个算法在多类数据集上产生的图模式所构建的分类器进行分析评价. 同二分类数据的处理方法相同，通过对每个多类数据集等比划分为训练集和测试集两部分，并从每个算法的挖掘结果中找到代表测试集每个类别的不同 top-k 图模式. 不同的是需要对多类数据的每两类组合分别训练一个两类分类器，用每个两类分类器去划分测试样本，根据少数服从多数的原则，确定样本的所属类别. 从表 3 中可以看出，基于 GINS 算法产生的图模式构建的分类器在多类数据集上的平均分类精度显著优于其他两类算法，从而说明 GINS 算法不仅适用于二分类数据，同时也适用于多类数据的分类要求，并具有较高的分类精度.

表 3　多类数据集上的 AUC 平均得分

数据集	LEAP	GraphSig	GINS
Class3	0.76±0.05	0.78±0.03	0.81±0.02
Class4	0.74±0.04	0.75±0.03	0.79±0.03
Class5	0.72±0.05	0.73±0.04	0.78±0.03
平均值	0.74±0.05	0.75±0.03	0.79±0.03

7 结 论

本文提出了一种二分类图上的非冗余协同图模式挖掘算法.通过对非冗余协同图模式性质的研究,给出一系列削减规则加快挖掘过程,基于置信度的边界估计更是进一步减少了挖掘过程中的计算代价;同时提出两种检测非冗余协同图模式策略,加速检测操作的完成.大量实验结果表明,论文提出的GINS挖掘算法同两个代表性算法相比具有较高的执行效率,从分类应用的角度出发,非冗余协同图模式获得较高的分类精度,进一步证明了GINS挖掘算法的有效性.

参 考 文 献

[1] Huan J, Wang W, Bandyopadhyay D, et al. Mining protein family specific residue packing patterns from protein structure graphs//Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology. New York, USA, 2004: 308-315

[2] Sharan R, Suthram S, Kelley R M, et al. Conserved patterns of protein interaction in multiple species. The National Academy of Sciences of the United States of America, 2005, 102(6): 1974-1979

[3] Borgelt C, Berthold M R. Mining molecular fragments: Finding relevant substructures of molecules//Proceedings of the 2002 IEEE International Conference on Data Mining. Maebashi, Japan, 2002: 51-58

[4] Deshpande M, Kuramochi M, Wale N, et al. Frequent substructure-based approaches for classifying chemical compounds. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(8): 1036-1050

[5] Bilgin C, Demir C, Nagi C, et al. Cell-graph mining for breast tissue modeling and classification//Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Lyon, France, 2007: 5311-5314

[6] Helma C, Cramer T, Kramer S, et al. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. Journal of Chemical Information and Computer Sciences, 2004, 44(4): 1402-1411

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

[8] Yan X, Han J. gSpan: Graph-based substructure pattern mining//Proceedings of the 2002 IEEE International Conference on Data Mining. Maebashi, Japan, 2002: 721-724.

[9] Yan X, Cheng H, Han J, et al. Mining significant graph patterns by leap search//Proceedings of the 2008 SIGMOD International Conference on Management of Data. Vancouver, Canada, 2008: 433-444

[10] Ranu S, Singh A K. Graphsig: A scalable approach to mining significant subgraphs in large graph databases//Proceedings of the 2009 IEEE International Conference on Data Engineering. Shanghai, China, 2009: 844-855

[11] Al Hasan M, Zaki M J. Output space sampling for graph patterns. Very Large Database Endowment, 2009, 2(1): 730-741

[12] Jin N, Young C, Wang W. Graph classification based on pattern co-occurrence//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China, 2009: 573-582

[13] Jin N, Young C, Wang W. GAIA: Graph classification using evolutionary computation//Proceedings of the 2010 SIGMOD International Conference on Management of Data. Indianapolis, USA, 2010: 879-890

[14] Jin N, Wang W. LTS: Discriminative subgraph mining by learning from search history//Proceedings of the 2011 IEEE International Conference on Data Engineering. Hannover, Germany, 2011: 207-218

[15] Anastassiou D. Computational analysis of the synergy among multiple interacting genes. Molecular Systems Biology, 2007, 3(1): 1-8

[16] Watkinson J, Wang X, Zheng T, et al. Identification of gene interactions associated with disease from gene expression data using synergy networks. BMC Systems Biology, 2008, 2(1): 10-25

[17] Costanzo M, Giaever G, Nislow C, et al. Experimental approaches to identify genetic networks. Current Opinion in Biotechnology, 2006, 17(5): 472-480

[18] Thoma M, Cheng H, Gretton A, et al. Discriminative frequent subgraph mining with optimality guarantees. Statistical Analysis and Data Mining, 2010, 3(5): 302-318

[19] Thoma M, Cheng H, Gretton A, et al. Near-optimal supervised feature selection among frequent subgraphs//Proceedings of the 2009 SIAM International Conference on Data Mining. Denver, USA, 2009: 1075-1086

[20] Zhang X, Pan F, Wang W, et al. Mining non-redundant high order correlations in binary data. Very Large Database Endowment, 2008, 1(1): 1178-1188

[21] Zhao Y, Wang G, Li Y, et al. Finding novel diagnostic gene patterns based on interesting non-redundant contrast sequence rules//Proceedings of the 2011 IEEE International Conference on Data Mining. Vancouver, Canada, 2011: 972-981

WANG Zhang-Hui, born in 1985, Ph. D. candidate. His major research interests include data mining and bioinformatics.

ZHAO Yu-Hai, born in 1975, Ph.D., assistant professor. His major research interests include database, data mining and bioinformatics.

WANG Guo-Ren, born in 1966, Ph. D., professor. His major research interests include uncertain data management, XML data management, query processing and optimization, parallel database systems and bioinformatics.

LI Yuan, born in 1986, Ph. D. candidate. His major research interests include data mining and bioinformatics.

Background

Graph structure provides a general way to modeling a variety of relationships among different objects, and has been widely used in many scientific domains. With the emergence of abundant scientific graph data, only a small part of the data has a class label. In these scientific applications, graph pattern mining can help build classification models for better predicting unlabeled graphs between different classes and understanding these complex structures.

Building graph classification models being widely used consists of two steps, namely, graph feature generation and classification. First, it can select a set of features by mining frequent or discriminative graph patterns from the graph database with a large number of labeled graphs. Second, building a graph classification model using the set of features selected to predict unlabeled graphs. Many efficient algorithms have been introduced to mine graph patterns, but they always generate too many patterns which are not suited for classification.

This paper first investigates the problem of mining non-redundant synergy graph patterns from two classes of graphs. By guaranteeing the property that the discriminative powers of synergy graph patterns are much higher than all their subgraphs, mining non-redundant synergy graph patterns can dramatically reduce the number of results and still capture the strong discriminative powers synergy graph patterns. With the mined non-redundant synergy graph patterns, a graph classification model can be built. The experimental results verify that the presented algorithm is efficient and effective. Moreover, the mined graph patterns are significant in the classification step.

This research work is supported by the National Natural Science Foundation of China (61272182, 61100028, 61073063, 61173030, 61332014), the National High Technology Research and Development Program (863 Program) of China (2012AA011004), the National Science Fund for Distinguished Young Scholars (61025007), the Program for New Century Excellent Talents Program in University (NCET-11-0085) and the Fundamental Research Funds for the Central Universities (N130504001).