分类号 密级

UDC

学 位 论 文

大规模图数据中统计显著子图挖掘算法研究与实现

|  |  |  |  |
| --- | --- | --- | --- |
| 作者姓名： | 黄海 | | |
| 指导教师： | 赵宇海 教授 | | |
|  | 东北大学计算机科学与工程学院 | | |
| 申请学位级别： | 硕士 | 学科类别： | 工学 |
| 学科专业名称： | 计算机软件与理论 | | |
| 论文提交日期： | 2016年12月 | 论文答辩日期： | 2016年12月 |
| 学位授予日期： | 2017年1月 | 答辩委员会主席： |  |
| 评阅人： |  | | |

东 北 大 学

2016年12月

##### A Thesis in Computer Software and Theory

**The Research and Implementation on Techniques of Statistical Significant Subgraph Mining from Massive Graph Data**

By Huang Hai

Supervisor: Professor Zhao Yuhai

**Northeastern University**

**Decemeber 2016**

独创性声明

本人声明，所呈交的学位论文是在导师的指导下完成的。论文中取得的研究成果除加以标注和致谢的地方外，不包含其他人己经发表或撰写过的研究成果，也不包括本人为获得其他学位而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：

日 期：

学位论文版权使用授权书

本学位论文作者和指导教师完全了解东北大学有关保留、使用学位论文的规定：即学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人同意东北大学可以将学位论文的全部或部分内容编入有关数据库进行检索、交流。

作者和导师同意网上交流的时间为作者获得学位后：

半年 □ 一年□ 一年半□ 两年□

学位论文作者签名： 导师签名：

签字日期： 签字日期：

摘要

近年来，图模式广泛被用在化合物，社交网络，蛋白质交互网络等数据的建模中。统计显著子图是指在分类图数据中与类标签具有显著相关关系的子图结构。统计显著子图的研究具有很重要的现实意义。过去以频繁度为显著性度量的频繁子图挖掘不足以度量子图的显著性。统计显著性为计算机学科与生物，化学等应用学科架起了沟通的桥梁。然而，在挖掘统计显著子图过程中因为子图数目巨大，在假设检验的框架下会产生大量的假阳性结果。而且，在如此多子图中进行搜索面临着巨大的计算量。

针对以上问题，本文利用一种新的算法框架，来控制挖掘结果中假阳性结果的比例，同时提高算法的计算效率。改进方法主要包括以下两点：（1）采用基于置换检验的多重假设检验校正过程，来控制挖掘结果中假阳性个体的数量。利用p值过滤原理提出了p值下界索引和精确p值索引两种技术对置换检验算法进行改进。减少了该过程的冗余计算，提高算法的运行效率。（2）利用群智能优化算法加快子图搜索的过程。将蚁群优化算法首次用于统计显著子图挖掘过程中。蚁群优化算法用于统计显著子图挖掘使得算法在保证挖掘结果的有效性的前提下提高了算法的时间效率。并通过优化蚁群算法来提高挖掘结果的多样性。

通过在实际数据集和人工合成数据集上的实验，本文所提出算法相比同类算法在时间效率上提高一个数量级。与此同时，挖掘结果的经验族错误率和作为分类特征进行图分类的准确率保持了与同类算法相近的效果。该实验结果表明了算法的高效性和有效性。

**关键词：**多重假设检验校正；统计显著子图挖掘；蚁群优化算法；族错误率；置换检验

Abstract

In recent years, graph pattern has been widely used in the modeling of compounds, social networks, protein interaction networks and other data. A statistically significant subgraph is a subgraph structure that has a significant correlation with the class label of the classified graph data. The study of statistically significant subgraphs is of great practical significance. The frequent subgraph mining is not enough to measure the significance of subgraph. Statistical significance set up a bridge of communication for the computer science and biology, chemistry and other applied subjects. However, a large number of false positive results are generated in the framework of hypothesis testing because of the large number of subgraphs in the mining of statistically significant subgraphs. Moreover, searching in so many subgraphs faces a huge amount of computation.

To solve these problems, this thesis uses a new algorithm framework to control the proportion of false positive results in mining results, and improve the efficiency of the algorithm. The improvement methods mainly include the following two points: (1) The multiple hypothesis testing correction process based on permutation test is used to control the number of false positive individuals in mining results. By using the p - value filtering principle, it propose two methods, p - value lower bound index and exact p - value index, to improve the permutation test algorithm. Reduced the redundant calculation of this process, improve the operation efficiency of the algorithm. (2) Using the group intelligent optimization algorithm to accelerate the process of subgraph search. The ant colony optimization algorithm was first used in the process of mining statistically significant subgraphs. The ant colony optimization algorithm is applied to the statistical significant subgraph mining to improve the time efficiency of the algorithm under the premise of guaranteeing the validity of the mining result. And optimize the ant colony algorithm to improve the diversity of mining results.  
 Through the experiment on the real data set and the synthetic data set, the algorithm proposed in this thesis improves the time efficiency by one order of magnitude compared with the similar algorithm. At the same time, the empirical error rate of the mining results and the accuracy rate of classification use these results as a classification feature keep similar to the similar algorithm. The experimental results show that the algorithm is efficient and effective.

**Keywords:** Correction of multiple hypothesis test;Significant subgraph mining; Ant colony optimization algorithms;FWER;permutation test

目录

[独创性声明](#_Toc3514)

[摘要](#_Toc20672)

[Abstract](#_Toc18976)

[第 1 章 绪论](#_Toc21591)

[1.1 研究背景](#_Toc22946)

[1.2 研究目的与意义](#_Toc2334)

[1.3 研究内容和主要贡献](#_Toc31492)

[1.4 论文组织结构](#_Toc10484)

[第 2 章 相关研究工作](#_Toc17962)

[2.1 子图挖掘](#_Toc9129)

[2.1.1 频繁子图挖掘](#_Toc9754)

[2.1.2 区分子图挖掘](#_Toc18360)

[2.1.3 频繁子图与区分子图关系](#_Toc5953)

[2.2 多重假设检验校正](#_Toc16365)

[2.2.1 邦弗朗尼原理](#_Toc8894)

[2.2.2 多重假设检验错误测度](#_Toc15293)

[2.2.3 多重假设检验校正方法](#_Toc25757)

[2.3 蚁群优化算法](#_Toc9681)

[2.4 本章小结](#_Toc10349)

[第 3 章 基本概念和问题定义](#_Toc8416)

[3.1 基本概念](#_Toc32224)

[3.1.1 费舍尔精确检验](#_Toc32328)

[3.1.2 基于置换检验的FWER控制过程](#_Toc14214)

[3.1.3 统计显著子图](#_Toc6745)

[3.2 问题定义](#_Toc19565)

[3.3 本章小结](#_Toc21497)

[第 4 章 基于*p*值过滤和改进蚁群算法的统计显著子图挖掘](#_Toc7079)

[4.1 算法概述](#_Toc17050)

[4.2 控制FWER的统计显著性阈值校正](#_Toc18685)

[4.2.1 基于*p*值下界的过滤原理](#_Toc9357)

[4.2.2 *p*值索引](#_Toc25701)

[4.2.3 算法WEST-FALLYOUNG++](#_Toc4405)

[4.3 基于蚁群优化的统计显著子图挖掘算法](#_Toc23023)

[4.3.1 优化模型](#_Toc5555)

[4.3.2 基于蚁群的子图搜索框架](#_Toc22080)

[4.3.3 基于多样性的蚁群搜索优化](#_Toc22522)

[4.3.4 算法示例](#_Toc27615)

[4.4 本章小结](#_Toc31592)

[第 5 章 实验测试及结果分析](#_Toc26581)

[5.1 实验环境与数据集](#_Toc7007)

[5.2 算法性能评价](#_Toc30841)

[5.3 子图挖掘结果显著性评价](#_Toc8022)

[5.3.1 挖掘结果统计效力的比较研究](#_Toc18982)

[5.3.2 挖掘结果用于图分类的实验研究](#_Toc6713)

[5.4 本章小结](#_Toc25656)

[第 6 章 结束语](#_Toc6201)

[参考文献](#_Toc13930)

[致谢](#_Toc7336)

[攻读硕士期间发表论文和参与项目](#_Toc6443)

# 绪论

近年来随着科学技术水平的不断发展和人们生活水平的提升，一些新的技术得到了广泛的应用和推广，大量的数据用图模式来建模和描述。图模式广泛运用于化合物，社交网络，蛋白质交互网络等数据的建模。在这些研究中，经常需要对图模式进行分类。图数据的分类是图数据挖掘的重要研究领域之一。而恰当的选取子图作为分类特征是获得高分类性能的前提。因此，需要一种方法选取出最具有显著性的子图作为分类特征，同时将与类标签相关性较小的子图排除在图的分类特征之外，即将假阳性子图结构的比例控制在一定范围之内，以提高统计效力。本文主要研究的主要问题是在大规模图数据中，如何挖掘统计显著子结构，并控制挖掘结果中的假阳性个体。

## 研究背景

在生物信息学，计算化学，医学信息和社交网络等科学研究领域的应用中，产生了大量的用图模式建模的数据。为了进一步对这些数据进行分析与研究，通常需要从中提取特征子图，作为原始图模式的代表。子图挖掘与相对比较成熟的序列数据和项集数据等相比图的数据量大。例如，在实际应用中有大量的构建自动分类模型的需求用来预测大量的图模型属于哪一个类别。在实际应用中，化合物分子通常被用图模式来存储，化学家们通常需要预测哪些化合物具有活性，哪些化合物不具有活性。要构建高性能的图分类器，一个重要的步骤是要提取特征子图。对于从图数据中提取特征子图已经有了一系列的研究，如：一部分学者致力于频繁子图挖掘算法，提出了一系列的频繁子图挖掘算法。Akihiro等人在2002年首先将A-prior的思想用于频繁子图挖掘的算法中，此后出现了一系列基于该算法思想的频繁子图挖掘算法。主要包括AGM算法[1]，FSG算法[2]等。后来韩家炜等人利用FP-growth的思想应用于频繁子图的挖掘过程中，促进了图挖掘的进一步发展。但是怎么对边扩展有一定的区别。此外还有其他的子图挖掘算法，如Wang等人于2005年提出了利用索引来挖掘频繁子图的算法GraphMiner[6]；2007年Zhu提出了利用用户约束条件来挖掘频繁子图的gPrune算法[7]等。随着研究的逐渐深入，研究人员发现利用频繁子图作为图的显著子结构对于某些场合并不是很适合。于是，一些研究者们提出了区分子图模式挖掘。Yan等人在2009年提出了LEAP算法[8]利用子图在正负样本中支持度的比较来挖掘区分子图。Ning等人在2010年提出GAIA算法[9]，该算法利用边的正负样本比例作为适应度函数，利用演化计算的思想挖掘区分子图，是一种高效的用来挖掘区分子图的算法。但是区分子图的定义不够科学，难以量化子图与类别的相关关系。于是研究者们提出了一种新的显著性子图的定义，即计算子图的*p*值。如：GraphSig[10]，GraphRank[11]等算法就是利用子图在原始图数据集中的分布估计子图的统计显著性。

然而上述研究有两个主要问题：

1. 频繁子图挖掘以频繁度作为子图显著性度量标准不足以度量子图的重要性。以图1.1所示图数据集为例。图中包含正负两种类别的图模式。图中子图A-B的支持度为8，图*G*1到*G*8均为其超图。然而，很显然其难以作为图分类的特征。而子图B-C的支持度仅为3，但是其在正样本集中出现频率为0.75,而在负样本集中出现频率为0。因此，其可以作为图分类的特征之一。因此，频繁度大的子图不一定具有较高的区分能力。

正负样本

图1.1 具有类标签的图数据集

Fig. 1.1 The graph dataset with class label

1. 子图模式数目巨大。计算每一个子图的显著性作为一次假设检验过程，如此多的假设检验过程会产生大量的假阳性结果。若子图数目为*m*,显著性阈值设置为*α，*则其中假阳性个体的数目大约为*m×α*。假定所有子图之间相互独立，则至少有一个假阳性结果的概率为1-(1-*α*)*m*。当*m*特别大的情况下，若选择较大的*α*将会产生大量的假阳性结果。而若选择较小的*α，*将会使得控制过程过于严格。在多重假设检验中如何选择一个恰当的统计显著性阈值是一个非常重要的问题。

## 研究目的与意义

图是一种现实生活中最普遍的数据模型。显著性子图挖掘是对以图模式存储的数据进行进一步处理的基础。这项研究成果可以广泛用于各个领域。例如，在药物研发过程中，化学家们在努力寻找化合物中对疾病最为有效的子结构，如果能够准确高效地找出真正对于疾病有显著性效果的子结构并且避免大量的假阳性结果将会节约大量的后续试验的时间和资源。在社会网络中，如果能够准确地发现具有显著性的群体对于广告服务提供，流行性疾病控制等方面均有重要贡献。应用前景方面，控制显著性子图挖掘中的假阳率的应用前景是巨大的，也有着很高的经济效益，广泛地应用于药物开发、基因数据分析，蛋白质交互网络等方面。

在传统的数据分类中，一个很重要的问题是如何挖掘出具有高区分能力的模式作为数据的分类依据。对数据进行分类，从而根据分类结果来提供给用户潜在的有用信息。同样，在图数据库中，该问题同样十分重要。相比传统的序列，项集或者结构化数据库，在图数据库中挖掘区分模式将会是更加复杂的。

当前已经有了一系列的进行统计显著子图挖掘的算法。然而，目前的做法存在两个严重的问题。（1）算法效率低下。图数据集中子图的数目与图数据集中样本的个数是指数相关的。目前已经有研究表明，在图，序列，项集中进行模式挖掘都是NP难问题。而且因为图模式的复杂性导致对其子图空间进行搜索将会是非常低效的。在大数据环境下，这样的算法性能难以满足实际应用的需求。提出新的算法以提高该过程时间性能是一个亟待解决的问题。（2）过去的研究大多数集中在如何提高算法的效率上。许多算法研究并未考虑到在大数据环境下进行多重假设检验校正会得到大量的假阳性结果。如果能够减少挖掘结果中假阳性个体的数目将会为后续的检测节约大量的人力，物力，财力。也能够尽可能多的得到真正与类标签显著相关的子图模式。

## 研究内容和主要贡献

本文针对大规模图数据集上的统计显著子图进行了深入研究。引入了置换检验技术用来控制挖掘结果的族错误率。该方法在统计学领域有着充分的理论保证。然而其计算耗时较大，如何把他应用于实际问题中需要提高其计算效率。本文提出了两种*p*值索引技术，极大的减少了算法运行过程中的重复计算。对于计算子图的统计显著性这一计算量很大的过程，可以大幅度减少计算时间。考虑到图数据结构的复杂性和子图空间的指数级爆炸增长，本文通过对支持度与统计显著性的关系制定了相应的削减规则，从而极大的减小搜索空间；在削减过程中利用统计显著性的下界与支持度的单调关系，减少了大量的不可能成为统计显著子图的分支。同时引入了蚁群优化算法用于子图空间的挖掘，从而极大的提高了算法的时间性能。在实验中，本文分别采用了经验族错误率和支持向量机分类算法，对挖掘结果的统计效力进行了实验验证，从而保证了算法的高效性的同时，也确保了算法的有效性。

本文的主要贡献在于：

1. 提出了大规模图数据集中挖掘统计显著子图的优化过程的数学模型。可以充分利用现有的组合优化求解算法来进行求解。
2. 在多重假设检验校正的过程中，对现有的FastWY算法进行了改进。通过对该算法的分析与研究，发现了其中大量的重复计算。针对其中的大量计算冗余，本文提出了两种*p*值索引技术，从而减少了重复计算。利用p值下界过滤原理，使用支持度阈值过滤不可能成为统计显著子图的个体，从而减小搜索空间。
3. 采用了基于群智能优化的算法对子图空间进行搜索。提出了ACOMiner作为多重假设检验校正过程中所采用的子图挖掘算法。将蚁群优化算法运用于子图挖掘过程中，充分利用了子图优化过程与组合优化问题的相似性来求得统计显著子图。过去的统计显著子图挖掘算法中，尚未发现利用群智能算法的相关研究。
4. 通过实验验证，对比了不同参数下算法的时间，空间效率。对比了本文所提出算法与现有算法的时间空间性能。通过对比几种算法挖掘结果的经验族错误率以及将所得到的特征子图作为特征向量利用支持向量机训练分类器的分类效果进行了比较。证明了本文的算法在将算法的时间效率提高了近一个数量级的同时，其挖掘结果依然具有较强的统计效力。

## 论文组织结构

本文后序章节采用如下的组织结构方式：

1. 为“绪论”部分。该部分首先对文章的研究背景进行了简单的描述。介绍了研究目的。
2. 为本文的“相关工作”部分。本章主要介绍了子图模式挖掘的概念，包括频繁子图挖掘，区分子图挖掘以及统计显著子图挖掘。介绍了这些子图特征提取方法的研究现状和意义，引出了如何高效地进行统计显著子图挖掘这一主要研究问题。然后介绍了多重假设检验校正方法的研究现状。以及对蚁群优化算法目前研究情况做了简单的介绍。
3. 为“基本概念和问题定义”部分。本章主要介绍了在本算法中提到的基本概念，主要包括费舍尔精确检验和置换检验的概念。然后用形式化的语言对本文要解决问题进行了定义。
4. 为“改进蚁群算法用于统计显著子图挖掘”部分。本章对算法的框架进行了详细的介绍。首先介绍了算法的整体流程，包括多重假设检验校正过程和子图挖掘过程。然后分别介绍了多重假设检验校正的主要过程和蚁群优化的数学模型以及具体的挖掘过程。其中详细分析了统计显著性与子图支持度的关系。将该性质利用在了挖掘过程中作为削减策略。同时在本章中提出了一系列的策略，通过在挖掘过程中调节参数控制蚁群优化方向和对挖掘结果进行后处理两类方法来保证挖掘结果的具有较少的信息冗余和较高的覆盖度。
5. 为“实验测试和结果分析”部分。本章通过在三种真实数据集，包括大规模数据和较小规模数据以及平衡数据和不平衡数据集上进行了实验。也利用了人工合成的大规模图数据进行了实验验证。通过大量实验证明了算法的准确性和高效性。通过对不同参数下算法的运行效率得到了不同参数对于算法效率的影响。通过与其他算法的时间空间效率比较验证了算法性能的提升。保证了挖掘结果的统计效力。并保证了分类的准确率。
6. 为“结束语”部分。对本文进行了总结与展望。

# 相关研究工作

在图数据集上的模式挖掘中，特征模式的挖掘具有重大的意义，包括频繁子图挖掘，区分子图挖掘等，是图分类中必要的技术之一。本文利用蚁群优化算法在大规模图数据中进行统计显著子图的挖掘。在本章将重点介绍特征模式相关知识、研究现状，多重假设检验校正的方法，最后介绍蚁群优化算法的研究现状。

## 子图挖掘

在数据挖掘领域图挖掘是一个重要的研究领域。这个领域的研究主要集中在从图数据集中挖掘特征子图。研究者们提出了一系列的特征子图度量标准，从频繁度，到区分度再到统计显著性。这些研究都有两个主要问题：（1）生成候选子图的有效机制；（2）怎么更好的处理候选子图，从候选子图中获得目标子模式。接下来详细介绍了这些方面的研究现状以及方法。

### 频繁子图挖掘

数据挖掘的最基础的任务是从数据中提取统计显著的知识。结构化数据和非结构化数据都能很自然的用图模式来表示。例如，在蛋白质交互网络中，蛋白质网络可以用图的形式来保存。用顶点表示基因，用有向边或者无向边表示物理交互或者功能关联。频繁子图挖掘是图挖掘中很重要的一个领域。频繁子图挖掘的目标是为了从频繁子图中提取所有的频繁子图。频繁子图就是指在一个图数据集中出现次数超过一定阈值的子图。频繁子图挖掘是一个很活跃的研究领域，同时也具有重要意义。频繁子图挖掘的算法在化学，生物，网络等方面都有着广泛的应用。

**定义2.1(子图同构)**对于一个图*G*(*V*1,*E*1,*LV*1,*LE*1,*Ψ*1)和子图*g*(*V2*,*E2*,*LV2*,*LE2*,*Ψ2*)。如果(1)且，,Ψ2(*v*)=Ψ1(*v*),;(2)且*,*Ψ2(*u,v*)=Ψ1(*u,v*);则*g*是*G*的子图,*G*是*g*的超图。

**定义2.2(频繁子图)**在一个图事务数据库中，一个子图*g*在数据库中其超图的个数为*sup(g)*。预先设定一个频繁度阈值σ，若，则图*g*为频繁子图。

频繁子图

图2.1 频繁子图

Fig. 2.1 Frequent subgraph

在下图中，*G*1,*G*2,*G*3为一个图事务数据库，根据上述子图同构的定义可得，是*G*1,*G*2,*G*3的子图，因此其支持度为3；*g*2是*G*1的子图，因此其支持度为1。若设支持度阈值为2，则*g1*是频繁子图，而*g*2不是频繁子图。

频繁子图挖掘通常利用两种技术(1)基于Apriori的方法；(2)基于模式增长的方法。这两种算法都与他们在关联规则挖掘中的对应算法相似。基于Apriori的算法利用广度优先搜索策略来遍历给定数据库的子图空间。因此，这种算法在得到*k*边子图是否为频繁子图时，必须先考虑全部*k-1*边子图。基于模式增长的算法被称为FP-growth算法[12]。这类型算法利用深度优先搜索策略来探索子图空间,主要代表为gSpan算法和FFSM算法。对于已经发现的子图*g*，算法递归的拓展子图，直到所有的基于子图*g*的频繁子图全部找到。

大多数的频繁子图挖掘算法采用迭代式的挖掘策略。每一轮迭代可以分为两个阶段。一是生成候选子图。二是计算子图的支持度，并与支持度阈值进行比较。所以频繁子图挖掘的算法研究通常集中在这两个方面。研究人员采用了一系列的策略来提升这两个方面的时间效率。然而由于子图同构检测很难提升，因此如何高效的生成候选子图成为了研究人员们的主要研究方向。如何系统地生成候选子图而尽可能减少冗余是一个关键的点。许多频繁子图挖掘的算法的特点在于所采用的候选子图扩展的方法不同。常用的方法主要有如下几种：

1. 基于层次连接的策略：一个*k+1*边的候选子图是通过合并两个共享*k-1*边子图的*k*边子图来获得。公共的*k-1*边子图被称为这两个*k*边子图的图核心。这个策略的关键点在于一个*k*边子图最多可以有*k*个*k-1*边子图可以共享*k-1*边。连接操作可能会产生众多冗余候选。
2. 基于最右路径扩展的策略：这种策略是一种最普遍的候选子图生成策略。他先从频繁*k*边子树通过添加树最右路径上的顶点从而生成生成*k+1*边子树。一个有向无环图利用最右路径扩展枚举可以生成一棵以空结点为根的枚举树，每一个结点为一个树模式。一个结点S拥有指向另一个结点T的边，则T是S最右路径扩展。所有的一边子树是根节点的最右扩展。因此，所有的子树模式可以通过深度优先搜索或者广度优先搜索的方式来进行遍历。
3. 基于等价类的扩展：这种策略的基本思想是一个*k+1*个顶点的子树是通过连接两个*k*子树生成的。这两个*k*子树必须在同一个等价类里边。一个等价类包含类前缀和一系列的成员。每一个类成员用(*l,p*)对来表示。*l*是第*k*个顶点的标签，*p*是第*k*个顶点双亲结点的深度优先搜索位置。
4. 右-左树连接：该策略利用最右叶结点和最左叶结点按照广度优先搜索的方式生成候选子图。给定两棵树*s,t*，如果Right(*s*)=Left(*t*)，他们的右-左树连接被定义如下：。lml(*t*)代表树*t*的最左叶子，Right(*t*)代表通过移除得到的最右树。rml(*t*)和Left(*t*)相反。

在频繁子图挖掘算法中最耗时的部分为候选子图生成和支持度计算。后者又是最耗费时间的部分。在频繁子图领域，研究者们致力于发现有意义的频繁子图集合，从而减少结果集合的大小。比如说极大频繁子图，闭频繁子图，近似频繁子图以及区分子图。为了将频繁子图利用于分类器并取得更好的分类结果，同时将特征选择的技术融入到子图挖掘的过程中去，一些研究者们将目光转向了区分子图挖掘过程中。

### 区分子图挖掘

区分子图是指那些在分类图数据集中，在某一类数据中是频繁的，而在另一类数据集中是不频繁的。区分子图是一些从选定的图集合中选定一些特定的特征子结构。这些特征子结构可以被用来识别图的类别，即做图分类。因此区分子图在结构化数据集中有着广泛的用途。其中有两个方面是目前最为火热的研究方向。一是蛋白质活性识别和功能预测。二是化合物分子活性预测。

下面给出区分子图的形式化定义：

**定义2.3(区分子图)**对于一个分类图事务数据库*G*,其中正样本集为*G*+，负样本集为*G-*。若子图*g*在*G*+中的支持度远大于(或者远小于)在*G*-中的支持度，那么子图*g*为区分子图。

在区分子图挖掘的过程中通常需要设定一个频繁度计算公式*d(G)*，通常有两种区分度的定义方式：

1. 
2. 

区分子图

图2.2 区分子图

Fig. 2.2 Discriminative subgraph

在图2.2中有一个二类图数据库，在正样本集中图*g*的支持集为{*G*1,*G*2,*G*3}，在负样本集中的支持集为。于是图*g*在正样本集中的支持度为3，在负样本集中的支持度为0，因此按照第一种区分度的定义方式计算得到*d(g)*=3。如果按第二种定义方式，由于支持度较小那一类支持度为0不能直接计算，可以采用平滑的方法，如将正样本集中的支持度和负样本集中的支持度同时加1，得到其区分度为4。

一种直观的挖掘区分子图的算法是首先枚举各个类中的频繁子图，然后在所有的频繁子图中选择那些在另一类中不频繁的子图集合。这种暴力枚举和选择策略保证了可以生成所有的区分子图。然而，这个过程需要生成大量的候选子图并且要在选择过程中计算频繁度，这里边包含了子图同构过程。而判断子图同构问题是一个NP完全的问题。这两个限制导致这种直接的算法不能处理大规模实际图数据集。另外，好多图在某一类中是频繁的，在另一类中也是频繁的。因此，直接暴力搜索的算法很难具有较高的效率。

为了克服这些缺陷，研究者们提出了三个直接挖掘区分子图的算法。

LEAP是最初的区分子图挖掘算法。该算法通过分支限界的技术具有最大区分度的子图模式。该算法认为结构相似的子图倾向于具有相似的区分度。同时，这个算法也运用了频繁度递减的挖掘策略。该策略利用了区分度与子图频繁度之间的关系。

GraphSig在挖掘区分子图的过程中也运用到了频繁子图挖掘的方法。从每一个顶点开始启动随机游走，将图转化为特征向量。然后，该算法将图划分为若干个小的组合，这样每一个组内的子图含有相似的特征向量。然后在每一组内挖掘频繁子图，并将频繁度阈值设定为一个较大的值。这是因为，同一组内的子图具有相似的特征向量，因此同组内的子图会拥有很高的频繁度。利用较高的频繁度阈值来挖掘频繁子图避免了大量的枚举无效子图，这使得graphSig可以用到相对较大的数据集上。

CORK算法[13]利用子图之间的一致性来度量子图模式的区分度。该算法在理论上可以得到近似最优解。给定一个子图集合，一致性指任意两个图模式不能利用该集合来区分的总数。

这三种算法，都要比暴力算法在性能上有明显提升。然而这些算法对于大规模图数据依然是不够有效的。于是，研究者们提出了利用演化计算的方法来进行区分子图挖掘。如GAIA算法。GAIA算法利用一种新型的子图编码方法，使得算法可以任意遍历子图模式并且利用一种启发式挖掘的方法遍历子图空间。在这种方法中心的候选是通过扩展已有的候选子图来获得的。并且，基于演化计算的方法，如遗传算法，可以使具有较低区分度的子图更加容易被淘汰，减少了子图的搜索空间。利用这种方法，GAIA可以更快地获得区分子图模式。另外，这种方法具有天然的并行性，以此来进一步提升挖掘的效率。

区分子图挖掘提出了一个朴素的思想用来度量子图的显著性，研究人员们多从提升效率减少重复挖掘的角度来改进算法，然而这样的算法对于子图显著性的度量不够科学系统，也不够另人信服。于是研究者们开始从假设检验的角度来进一步挖掘真正具有统计意义的模式。

在研究过程中，研究者们对于区分子图的统计效力缺乏很严密的数学推理。于是，研究者们开始从统计显著性的角度来衡量子图的统计效力。这种方式采用了假设检验的思路框架作为基础。首先，设定了原假设。在这里，因为需要证据证明子图与类标签具有显著相关关系，因此首先假定原假设为子图与类标签没有相关关系，备择假设为子图与类标签具有显著相关关系。

形式化的描述统计子图挖掘的问题如下。假设给定两个图集合*D*和*D’*，其中图集合*D*的大小，且图集合*D*大小，为了不失一般性，本文规定。对于每一个子图，。首先定义一个零假设为子图*H*是否出现与*G*的类别相互独立。本文的任务是寻找有足够证据拒绝原假设的子图*H*，子图*H*被认为是统计显著的子图。

统计显著子图挖掘是大规模图数据库中一个越来越重要的问题。现在已经有的频繁子图挖掘算法由于其难以度量较低支持度的子图的统计显著性，因此存在严重的问题。如果把频繁子图中的频繁度阈值降到很低，则会有大量的挖掘结果需要进行后续处理，而如果阈值较高的话，又会有大量的显著结果丢失。统计显著子图挖掘开启了一个新的图模式挖掘研究方向。在化合物分子，蛋白质网络等方面展现了充分的潜力来识别重要的子结构。

### 频繁子图与区分子图关系

区分子图相对于频繁子图进一步挖掘了子图与类标签之间的相关关系，本文所研究的统计显著子图是区分子图挖掘领域一个新的研究方向。统计显著子图挖掘利用统计学中假设检验的方法，可以进一步挖掘出具有较小支持度的区分模式。从统计学的角度精确地衡量了子图模式与目标变量之间的关系。使得对于子图重要性程度有了更加科学合理的解释。统计显著子图，区分子图，频繁子图三者之间有着密切的联系。他们三者之间的关系如图2.3。

三者关系

图2.3 三者关系

Fig. 2.3 The Relationship between frequent subgraph,discriminative subgraph and significant statistical subgraph

从图中可以看出，当挖掘频繁子图的时候，如果把支持度阈值设定到很低的水平，统计显著子图将是频繁子图的一个子集，而如果仅仅挖掘频繁子图，将会有大量的结果集与类标签并没有相关关系，因此以此作为特征对图数据库进行分类很难取得良好的效果。区分子图是一种朴素的对子图与类标签之间关系的描述方法，他只是从直观上对模式在正负样本集中的支持度差异进行了比较，缺乏深刻的理论依据支撑。难以对挖掘结果进行科学的描述，因此其挖掘结果难以保证较高的统计质量。而统计显著子图利用成熟的统计检验手段，进一步将子图与类标签之间的关系用统计规则描述，用统计显著性值来描述子图的统计显著性情况，可以从理论上表明挖掘结果的可靠程度。因此在后续的对特征子图的利用过程中可以取得较好的效果。统计显著子图是一种更为科学有效的特征子图挖掘方法。

## 多重假设检验校正

近年来，随着科学技术的发展，对于高维数据研究越来越多，多重假设检验的问题受到了越来越多人的关注。以统计显著子图挖掘为例，对于一个图数据库，其子图规模是非常巨大的，即使只考虑待检验子图，这个数量也是很大的。因此，如何控制犯第一类错误的概率，逐渐成为一个严峻的问题。

### 邦弗朗尼原理

假设人们拥有一些数据，并预期从该数据中找到某些特定类型的事件。即使得到的数据是完全随机的，这些事件还是有一定的可能性会出现。然而，这些事件出现只是因为随机性，并不是人们希望的具有一定因果意义的。随着人们获得的数据量逐渐增加，这种情况出现的次数也会大幅度的增加。这些随机出现的模式往往会有一些不同寻常的规律，这些特征看上去会有很重要的意义，但是实际上他只是随机出现的一些特征，并没有别的原因。从这个角度看，这些数据的出现不具有统计效力。随着如今数据的规模不断增大，在数据挖掘领域，如果依然沿用传统的*p*值来衡量数据的统计显著性就会产生大量的假阳性结果。下面给出邦弗朗尼原理的一个例子：

假设对于一个图数据集*D*，若设统计显著性阈值*ɑ*=0.05，挖掘得到10000个统计显著子图。则每一个子图犯第一类错误的概率为0.05，这就使得挖掘结果中假阳性个体期望为500个。这部分结果是否出现于图的类标签是相互独立的，他们成为挖掘结果只是因为随机性。因此，如果进行后续检验，将会耗费大量的人力物力，或者具有一定的破坏性。其成本可能是难以承受的。而且，实际中数据量的增大，假阳性的个体会进一步的增加，为了解决这种问题引入了多重假设检验校正的方法。

下面给出邦弗朗尼原理的一个阐述：在数据随机性假设的基础上，可以计算得到期望发生事件发生次数的期望值。如果该值明显高于希望找到的真正实例的数目，那么可以预计，得到的几乎任何事件都是没有意义的，只是统计上由于随机性出现的假象，而非所需寻找事件的证据[14]。

针对上述问题，统计学家进行了一系列的研究来避免大量假阳性个体的出现。

### 多重假设检验错误测度

假设检验问题是统计推断的一个重要的方面。单个假设检验的理论在最近几年已经有了比较完善的理论基础。然而，在多重假设检验方面，研究还处于比较初步的阶段。人们还没有找到一个完美的方法。随着科学的迅猛发展和技术的不断进步，在互联网时代，大数据成为了人们耳熟能祥的词汇。然而，这对于统计工作者来说却是一个空前巨大的挑战。多重假设检验的理论是大规模数据分析的主要理论基础。例如，对于大规模图数据中的子图挖掘，同时识别两个类别中成千上万的子图模式是否与类标签有显著性关联便涉及到多重假设检验问题。下面，重点介绍几种多重假设检验校正问题中的错误测度[15]：

(1)FWER(Family Wise Error Rate)：族错误率

族错误率指挖掘结果中至少有一个假阳性结果的概率。从族错误率定义可以看出，族错误率对于错误的控制很严格，是一种非常保守的错误测度。但是在特定的应用背景下，该测度还是十分有效的。在药物临床试验中，要对很多种药物效果进行检验，主要希望知道哪一种药物最有效，并不希望找出所有有效的药物种类(相对于对照组)。在这种需求下，虽然控制FWER会丢掉一些有效的处理方式，即真正显著的假设检验有较大的概率没有被发现，然而对于所关心的问题该度量方式是有效的。

在单个假设检验的情况下，定义了*p*值，在多重假设检验中引入了类似的值的定义。将多重假设检验中的*p*值称作调整的*p*值：

 (2.3)

其中，*Err*表示设定的错误评价指标，如FWER，FDR等。这是对应于某一个单一假设检验*Hi*的调整以后的*p*值为*Hi*刚好被拒绝是检验所犯第一类错误的总体水平。若设定*Err*为FWER，则调整*p*值为：

 (2.4)

对于给定的总体检验水平，如果，则拒绝*Hi*。

然而，在实际过程中，很难找到这样的*p*值阈值，通常做法是，得到该值的一个下界。

(2)FDR(False Discovery Rate)错误发现率

由于FWER错误测度太过保守，在某些问题中可能是不太合适的。于是Benjamini和Hochberg提出了另一种较为开放的错误测度FDR。错误发现率FDR用公式2.5计算得到：

 (2.5)

在上述FDR定义下，指示函数*I*(*R*>0)主要有两个用途：一是添加该指示函数以后，使得*R=0*时有意义；二是当*R=0*时，*V=0*检验所犯的第一类错误等于0。因此该定义是合理的。当所有原假设均为真，则；否则。

1. pFDR(Positive False Discovery Rate)阳性错误发现率

Storey针对微阵列实验数据分析问题提出了另一种错误度量pFDR。阳性假发现率pFDR是指在假设显著存在的条件下，被错误拒绝的个数与总拒绝假设个数的比值的期望。即：

 (2.6)

考虑到pFDR与FDR之间的关系：

 (2.7)

由此可以得出二者仅仅相差一个概率*Pr*(*R*>0)。对于微阵列数据的假设检验问题，通常情况下显著性检验是真实存在的。因此在*Pr*(*R*>0)这个概率条件下定义错误度量是一种合理的方式。而且，当检验个数很多时，，此时pFDR与FDR的值几乎相等。

### 多重假设检验校正方法

下面介绍了几种目前常用的多重假设检验校正方法。假设检验校正方法主要可以分成三类：直接调整；基于置换检验的方法和保留方法。

1. 直接调整方法：

邦弗朗尼校正是一种常用的多重假设检验校正方法[16]。他属于直接校正的方法。邦弗朗尼校正的目标是为了控制族错误率。为了将族错误率控制到小于*ɑ*，邦弗朗尼校正用*ɑ*除以检测个数，得到校正以后的显著性阈值。假设*Nt*是执行的多重假设检验个数，所以*p*值小于的假设被认为是统计显著的。

Benjamini and Hochberg’s方法[17]是一种用来控制FDR的直接调整方法。若*H1*，*H2*...*H3*是*n*个假设检验并将其按对应*p*值升序排列。得到对应的*p*值序列为*p*1，*p*2...*p*3。如果将FDR控制在*ɑ*，这种方法寻找一个最大的*i*使得，然后所有*p*值小于*pi*的假设被认为是统计显著的。

1. 基于置换检验的方法[18-20]：

该方法是本文中主要采用的多重假设检验校正的方法。基于置换检验的假设检验校正方法将类标签随机洗牌，然后重新计算假设检验的*p*值。随机洗牌过程打破了模式和类标签之间的关系，因此重新计算得到的*p*值是零假设的一个近似分布。该方法目标是将族错误率控制在*ɑ*内。其过程主要包括4步：

1. 随机排列类标签*N*次。
2. 利用随机排列的类标签计算所有模式的*p*值。
3. 得到每一个排列下的最小值，，...。
4. 取，...的*ɑ*分位数作为校正后的统计显著阈值。

基于置换检验的多重假设检验校正方法考虑到了模式之间的交互作用，因此该校正方法能够得到更精确的截断*p*值。然而，基于置换检验的多重假设检验校正方法又是非常好使的算法。

1. 基于保留的方法[21]：

基于保留的评估方法是有Webb提出来的。他是为了克服前两者的缺陷：一是直接调整过于严格；二是基于排列的校正方法过于耗时、他提出了一种方法将数据集分成两部分，一部分为探索数据集，另一部分为验证数据集。首先从探索性数据集中挖掘显著模式。在该结果中值小于*ɑ*的模式在验证集中进行进一步验证。为了将族错误率控制在*ɑ*内，在验证集上用邦弗朗尼校正来调整*p*值。此时的校正因子为探索集上*p*值小于显著性阈值的模式。因此，待检验模式的数目会有大幅度的减小。因此，该方法试图发现那些有适当的*p*值的模式。类似的，可以用来控制FDR。

基于保留的方法比基于置换检验的校正方法节省更多的时间。然而，这种方法的效果与数据集的划分方式密切相关。如果一个模式恰好只落入了探索集或者验证集，那么这样的模式就不会被发现了。因此，这种方法一方面使得噪声规则不容易被识别为显著的，同时另一方面也导致许多显著的关联规则没有被识别到。

## 蚁群优化算法

蚂蚁是一种简单的生物，但是大量的蚂蚁同时工作会完成大量的复杂的工作。单个蚂蚁可能会毫无目的地游荡直到他精疲力尽。平均每只蚂蚁在他们的大脑里只有10000个神经元。这并不能让他们完成很复杂的任务。但是蚂蚁们通常数以百万计组合成群。1000000只蚂蚁集合在一起，神经元数目可以达到100亿。这几乎可以达到一个普通人的神经元数目了。为什么这样微小的一种生物可以在地球上存活如此之久，科学家们将这归功于他们的适应性和社会组织能力。借鉴了蚂蚁的这种群体活动的方式，科学家们研究了模拟蚁群行为的优化算法[22]。下面对该类算法做简单介绍。

|  |
| --- |
| **算法2.1** *ACO*蚁群优化算法  **输入：**图结构*G*  **输出：**TSP最短路径 |
| 1. **function** AS(*G*): 2. **for** each edge: //对每一条边初始化信息素 3. set initial pheromone value *t0* 4. **end for** 5. **while** not stop: //开始迭代直到达到终止条件 6. **for** each ant *k*: //每一只蚂蚁寻找一条路径 7. randomly choose an initial city 8. **for** *i=1* to *n*: //要经过所有*n*个城市 9. Choose next city with probability 10. **end for** 11. **end for** 12. Compute the length *Ck* of the tour constrcted by the *k-*th ant 13. **for** each edge: //释放信息素 14. Update the pheromone value 15. **end for** 16. **end while** 17. **end function** |

蚁群算法第一次是由Colorni等于1991年发表[23]。该算法在旅行售货商问题上得到了很好的效果。在蚁群算法中，模拟每一只蚂蚁从一座城市旅行到另一座城市，并且模拟了蚂蚁在经过的路径上释放信息素的情形。信息素在释放的同时还要挥发。一只蚂蚁从当前城市移动到另一座城市时是根据相邻城市信息素比例来确定移动的概率。同时蚂蚁也被假想拥有对于该问题的一些相关知识来帮助他们在旅行过程中做出决策。蚁群优化算法的伪代码如算法2.1。

对应大量蚂蚁的集体行为主要有以下三个基本特征[24]：

1. 正反馈
2. 负反馈
3. 启发式信息：在蚁群算法中构造启发式规则,帮助蚁群搜索过程找到可行解以及尽可能找到最优解。

蚁群算法是一种近似算法。一般用他来解决至今没有找到精确的有效算法的问题的。有效算法一般指多项式复杂度的算法。以旅行售货商问题为例，它要求一个完全图的最小哈密尔顿回路。这个问题属于NP完全问题[25]。也就是说，当问题规模稍微大一点，现有的精确算法的运算量就会急剧增加。因此就产生了各种近似算法，以解的质量来换取效率，寻求满意解而不是最优解，蚁群算法就是其中的一种。蚁群算法对于NP完全问题的近似解可以取得较为满意的解。

## 本章小结

本章主要对于论文相关的其他研究做了简要的介绍。首先介绍了子图挖掘的两个主要方向，并同时介绍了他们之间的关系。接下来，介绍了多重假设检验校正的错误度量标准以及校正的方法。最后，介绍了蚁群优化算法的基本原理和算法主要流程，并对其适用范围进行了分析。

# 基本概念和问题定义

本章详细介绍了第四章算法所涉及到的基本概念，包括费舍尔精确检验，基于置换检验的FWER控制过程以及统计显著子图的相关概念。为第四章做好准备。并给出了问题的定义，明确了研究目标。

## 基本概念

### 费舍尔精确检验

费舍尔精确检验[26]是一种用来分析列联表统计显著性的检验方法。在使用中通常将其用于较小的样本规模。但是实际上它可以用于任意大小的数据集。这是以其发明者命名的一种精确检验方法。之所以被称之为精确检验是因为这种方法可以精确的计算出偏离零假设的显著性。而许多非精确检验则需要样本量趋向于无穷大时，其极限才是显著性的精确值。因此在大样本下可以使用卡方检验代替费舍尔精确检验。

在统计学中，列联表是一种通过矩阵形式展示多个变量频率分布的表格。在商业智能，工程和科学研究，调查研究中有广泛的应用。利用列联表可以发现两个变量之间的交互关系。假定有A和B两个变量，每一个变量有两种取值0和1，列联表可以用如表3.1的方式表示，两个变量的频率分布：

表3.1 描述一对变量A,B相关关系的列联表

Table 3.1 The contingency of variation A and B

|  |  |  |  |
| --- | --- | --- | --- |
| A  B | 0 | 1 | total |
| 0 |  |  | =+ |
| 1 |  |  | =+ |
| total | =+ | =+ | =+(+) |

在表3.1中，*x*00表示属性A和B同时为0的样本个数；*x*01表示属性A取1，属性B取0个样本个数；*x*10表示属性A取0，属性B取1的样本个数；*x*11表示属性A和属性B同时取1的样本个数；*B*0，*B*1分别表示属性B取0和取1的样本个数；*A*0，*A*1分别表示属性A取0，和取1的样本个数；*n*表示总的样本个数。

表3.1显示了属性A和属性B的分布情况。为了得到属性A和属性B的相关关系的显著性可以通过一系列的统计检验方法获得：如皮尔逊卡方检验，G检验，费舍尔精确检验，以及Barnard’s 检验。如果在不同列，两种个体的比例有着显著的区别，那么就说，这两个变量之间存在着关联性。接下来，主要介绍费舍尔精确检验的方法。

以上述列联表为例，观察到表中值的概率可以利用超几何分布计算：

 (3.1)

在式4.2中，当*x00*确定时，其他三个值也可以相应确定。于是可以得到：

*q(x00)=p* (3.2)

假设A属性的取值与B属性的取值是无关的。观察值的显著性可以通过计算当原假设正确时与观察值同样极端或者更极端的情况下的概率和来得到。为了计算观察到的数据的显著性，必须计算这些所有可能情况的J概率值，并将其相加。可以形式化的得到对于特定观测值的显著性：

 (3.3)

*PL*是观察到更小值的概率和*PR*是观察到更大值的概率。*PL*和*PR*被用作单尾检验的*p*值，而定义为双尾检验的*p*值。如果*p*值小于给定显著性阈值*ɑ*，则拒绝原假设，即A属性的取值与B属性的取值是相关的。否则接受原假设，即两个属性之间是无关的。由此得到*ɑ*是犯第一类错误的概率。

### 基于置换检验的FWER控制过程

置换检验[27]因其对总体分布自由，因此具有广泛的应用。通过对样本类标签顺序上的置换，切断类标签与样本之间的关系，对其多次采样，计算其统计检验量。置换检验是一种基于重采样技术的统计推断手段。通过大量的实验计算可以近似构造出原始分布，从而估计出结果的族错误率，进而对挖掘结果中的族错误率进行控制。利用置换检验的方法可以对多重假设检验的族错误率进行控制。该方法随机排列记录的类标签，在每一个排列下重新计算子图的*p*值。随机排列的过程打破了子图和类标签之间的关联，因此重新计算得到的*p*值是对零假设的一个近似。即在子图跟类标签相互独立的情况下，子图*p*值的分布。

置换检验示意图

图3.1 置换检验过程图示

Fig. 3.1 The example of permutation test process

为了将族错误率控制在小于*ɑ*，随机生成了*N*个排列。在这*N*个排列中是不应该有统计显著模式的。因此，如果在这些排列下发现了统计显著子图，那么这些结果就是假阳性结果。于是需要确定一个*p*值阈值，使得在该阈值下所有排列中发现统计显著性子图的比例小于。为了找到该截断*p*值，在每一个排列下获得最小的*p*值，并且将其按照由小到大的顺序排列。其第*ɑ*分位数的值被作为截断阈值。于是认为在原始图数据库中，如果其*p*值小于该阈值，则认为其是统计显著的。

图3.1展示了置换检验的整个过程。图中每一行代表一次排列，每一列对应一个模式，即一次假设检验。每一个模式在对应的排列下都能通过费舍尔精确检验得到一个统计显著性。于是表中*pi*j代表在第*i*个排列下，第*j*次假设检验得到的子图*p*值。对应每一次排列得到了其最小统计显著性*p*值。一共得到*n*个最小显著性*p*值。于是将这些*p*值按递增排序，取其*ɑ*分位数显著性作为校正后的最优显著性阈值。以此来保证挖掘结果整体的族错误率。

基于置换检验的多重假设检验校正方法考虑到了子图模式之间的交互作用，通过过采样的方法近似的描绘出假设的原始分布，因此相比直接调整的方式可以得到更为精确的最优*p*值阈值。置换检验作为一种成熟的统计技术，在理论上有了充分的保证。然而，考虑到这种算法计算过程中相当于*n*次挖掘统计显著子图，而挖掘过程又是一种非常耗时的操作。因此，这是一种时间复杂度很高的方法，于是本文采用了一系列的技术来减少其时间消耗。

### 统计显著子图

在实际应用中图数据集通常包括两类。统计显著子图是指在某一类图数据中统计明显多于另一类图数据的子图。该明显程度利用统计学中的假设检验过程来定义。统计显著子图挖掘过程就是多个假设检验过程。首先原假设为该模式的发生与否与类标签是相互独立的，而备择假设就是该模式出现与否与类标签相关。然后，对每一个子图进行假设检验。通过费舍尔精确检验得到每一个子图的显著性*p*值。此时，需要一个统计显著性阈值作为是否具有统计显著性的判断标准。本文中研究的问题需要考虑到整个结果集的显著程度，于是需要将结果集的族错误率(FWER)控制在预设的范围之内。这就是一个多重假设检验校正的问题。这就是本文所要挖掘的子图模式。

表3.2 假设检验结果类别

Table 3.2 The result of hypothesis test

|  |  |  |
| --- | --- | --- |
|  | 备择假设为真 | 零假设为真 |
| 被认为是统计显著的 | TP | FP |
| 被认为不是统计显著的 | FN | TN |

表3.2展示了在假设检验过程中可能会出现的四种情况。TP表示在备择假设为真的情况下拒绝了原假设，即认为该假设是显著的。TN表示了在零假设为真的情况下接受了零假设。其中FP发生即犯了第一类错误。FP占所有假设检验的比例就是犯第一类错误的概率。该概率在统计意义上等于假设检验过程中设定的统计显著阈值*p*值。在进行统计显著子图挖掘的过程中算法要保证FP样本的个数要尽可能的少，同时TP的个数要尽可能多。这样在挖掘过程中就需要在敏感度和准确度之间做一个权衡。如何确定一个恰当的统计显著性阈值时本文主要研究的问题之一。

在对统计显著子图进行挖掘的过程中，主要有两大难点需要被解决：一是如何确定区分子图是否显著的统计显著性阈值；二是在子图挖掘过程中如果将所有子图都枚举出来然后进行假设检验将会是一个非常耗时的过程，大量的子图模式将成为算法的主要瓶颈，如何避免枚举出大量的候选结果，以此来加速算法的执行效率也是一个主要解决的问题。本文主要针对这两个问题设计了一系列的算法来保证算法的高效性和有效性。

## 问题定义

本文的主要研究内容是从分类的图事务数据库中挖掘对图的类别具有显著意义的子图，同时将挖掘结果中不具有显著意义的子图数目控制在给定的阈值范围内。本文所研究的对象是一个多重假设检验的问题。待检验的个体为图数据库中的所有子图集合。原假设为该子图与原图的类标签是无关的，备择假设为该子图与原图的类标签是相关的。当子图的*p*值小于一定值时，拒绝原假设，接受备择假设，即该子图与原图的类标签是显著相关的。

对于子图*p*值的计算采用费舍尔精确检验来计算其*p*值。对于某一个子图*s*的分布，可以用如表3.3的列联表表示。

表3.3 用列联表表示子图分布

Table 3.3 The contingency of the distribution of subgraph

|  |  |  |  |
| --- | --- | --- | --- |
|  | *s*出现 | *s*未出现 | *total* |
| *G+* | *x* | *n-x* | *n* |
| *G-* | *x’* | *n'-x'* | *n’* |
| *total* | *x+x'* | *n-x+n’-x’* | *n+n’* |

表3.3中对应的符号含义为：*x*表示子图*s*在正样本集中出现的频数，*x’*表示子图*s*在负样本集中出现的频数。*n*表示正样本集的数目，*n’*表示负样本集的数目。由上图中展示的子图*s*的分布情况，利用费舍尔精确检验值*p*计算公式(2)可以得到该子图的*p*值。因为子图个数巨大，于是需要考虑多重假设检验校正的问题，本文采用了对挖掘结果的族错误率进行控制的方法，来降低挖掘结果中假阳性个体的比例。综上所述，得到问题的定义如下：

**问题定义**：给定含有*n*个图的图事务数据库*G*，给定FWER阈值*ɑ*，该图数据集中统计显著子图为满足如下条件的图*g*：

（1）挖掘结果子图集合为*G*res；

（2）对于*G*res中的子图，其中至少有一个是与类标签无关的概率小于*ɑ*，即();

（3）考虑到挖掘结果的多样性，对于结果集中任意两个子图在给定相似度阈值*δ*时，*sim(s1,s2)*>*δ*。

由以上定义可以得知，在统计显著子图挖掘过程中，不是将子图看作单个的个体，而是将所有待检验子图看作一个整体。

## 本章小结

本章介绍了算法中所涉及到的基本概念以及问题定义。首先本章介绍了，对于子图统计显著性检验的费舍尔精确检验，引出了子图*p*值的概念。然后介绍了置换检验的概念，并介绍了怎么利用置换检验来控制挖掘结果的族错误率。接下来介绍了统计显著子图的相关概念。最后给出了本文所研究问题的形式化定义，明确了对该问题研究的重点。

# 基于*p*值过滤和改进蚁群算法的统计显著子图挖掘

在本章中介绍了利用改进的蚁群算法从大规模图数据中挖掘统计显著子图的算法框架。4.1节介绍了算法的总体结构，然后详细讨论各部分的具体内容。4.2节介绍了基于多重假设检验过程中族错误率的控制方法，对已有的方法的计算效率进行了改进。4.4节提出了针对于图数据库的蚁群优化搜索方法，通过演化计算的方法加快对子图的搜索过程。4.5节对本章内容进行了总结。

## 算法概述

本文提出了基于p值过滤和改进蚁群算法的统计显著子图挖掘算法PABSMiner。算法主要包括两部分。第一部分为多重假设检验校正部分。该部分中利用用户给定的族错误率阈值进一步确定子图显著性*p*值阈值，并且利用支持度与统计显著性下界的关系得到能够达到该显著性阈值的最小支持度。该部分提出了利用两种*p*值索引对算法运行效率进行改进的方法。这两种*p*值索引包括对已知正负样本支持度的精确*p*值索引和已知总体支持度对于*p*值下界的索引。第二部分为基于蚁群优化的子图挖掘算法。该部分利用第一部分得到的统计显著性阈值以及支持度阈值，利用演化计算的方法，对整个子图空间进行搜索。这样的搜索过程可以利用群体的行为指导搜索过程进行，大量的削减搜索空间。算法的主要输入为图事务数据库和用户自定义的族错误率阈值。输出结果为该事务数据库中的统计显著子图集，该结果集可以保证其中出现第一类错误的概率小于输入的族错误率阈值。其流程如算法4.1。

从该算法流程可以看出，第1，2步主要是利用FastWY(一种最新的基于置换检验的多重假设检验校正方法)的改进算法[28]WYplus进行假设检验校正，获得校正后的统计显著性阈值标准，然后在第二步利用*p*值与支持度之间的关系，得到最小支持度阈值，用来在第三步挖掘过程中进行削减。对于支持度小于该值的子图，将不可能成为统计显著子图，于是在蚁群搜索的过程中，如果某一个子图的支持度小于该值，其子图将不再作为候选模式进行搜索。这将大大减少第三步中的搜索空间，提升挖掘速度。

|  |
| --- |
| **算法4.1** PABSMiner(基于p值过滤和改进蚁群算法的统计显著子图挖掘)  **输入**：图事务数据集*G*；族错误率阈值*α*  **输出：**图数据集*G*中的统计显著子图*Gres* |
| //利用改进的westfall-younglight算法，在给定的族错误率阈值下获得相应的统计显著性*p*值阈值*p*  1*.p*=WYplus(*α,G*)  //利用支持度与统计显著性阈值下界的关系，求出满足统计显著性条件的子图的最小支持度  2*.σ*=minSupp(*p*)  //利用蚁群优化算法，在图数据集中挖掘显著性值小于步骤1中所得到的阈值的子图  3*.Gres*=ACOMiner(*G,σ,p*) |

## 控制FWER的统计显著性阈值校正

在挖掘统计显著子图过程中，需要对每个子图进行统计假设检验，这其实是一个多重假设检验校正的过程。如果以传统假设检验中常用的0.05或者0.01作为显著性阈值，由于每一个子图都有一定的概率犯第一类错误，那么会导致整个结果集中有大量的假阳性结果。考虑到后续的实验以及继续分析的耗费，大量假阳性结果将会浪费大量的人力物力进行排除。如果利用直接调整的假阳性校正方法，如邦弗朗尼校正法，因为没有考虑到子图与子图之间的交互关系，会导致对于显著性的要求过于严格，导致丢失大量与图类别有显著相关关系的结果。因此，研究者们对如何对显著性阈值进行校正提出了一系列的方法。其中置换检验作为一种经典的假设检验校正方法，能够较为准确的估计数据的零分布，从而确定出恰当的显著性阈值。但是由于需要生成很大数量的排列，并在每次对类标签重排后都需要获得该次重排列下的最小*p*值。这样的计算量很难在大数据环境下进行统计显著性的校正。但是经过分析，这些方法依然存在一些冗余的计算。本文将对我们对现有算法的改进进行详细介绍。

### 基于*p*值下界的过滤原理

在本文中主要利用费舍尔精确检验确定子图的统计显著*p*值。费舍尔精确检验是计算一个22列联表的*p*值的最流行的方法之一。结合表3.1分析，在一个数据集中，费舍尔精确检验假设对于一个子图其支持度*x+x’*，以及正样本数*n*，和负样本数*n’*是确定的。因此根据费舍尔精确检验*p*值计算公式可以得知，该子图的*p*值仅与其正样本集中的支持度或者负样本集中的支持度相关。*p*值的自由度仅为1。在这样的假设下，可以认为子图*si*与类标签*Y*之间是否独立的数据模型可以用超几何分布来描述。在子图支持度，正样本数，负样本数固定的情况下，该子图出现数据集中分布的概率公式如式4.1[29]：

 (4.1)

其中，，，。如果在一个数据集中得到子图在正样本集中的支持度为，然后就能得到该子图的单尾显著性值：

 (4.2)

因为，从公式中可以得到，，*k*只有有限的取值范围。同样，一个子图的显著性*p*值只与*γ*有关，因此其*p*值同样仅有有限的取值范围。于是，当一个子图的支持度已知的情况下，可以得到该子图的最小可达*p*值。即，在当前数据集中，该子图所能达到的最小显著性程度被定义如4.3：

 (4.3)

因为，在一个数据集中正样本集和负样本集中的样本个数都是确定的，于是*Ψ(x+x’,n,n’)*可以写成*Ψ(x+x’)*。于是对于一个支持度为*x+x’*的子图，其能达到的最小*p*值为。因此，假如确定了子图的统计显著性阈值*p*，就可以通过支持度获得待检验子图的集合。即，如果该子图的支持度所能达到的最小*p*值大于显著性阈值，则该子图不可能成为统计显著子图。定义待检验子图为满足如式4.4所示条件的图模式：

 (4.4)

在上式中，*δ*为校正后的显著性阈值，I(*δ*)即为在该显著性阈值下待检验子图的集合。不在该集合中的子图将不可能成为统计显著子图。

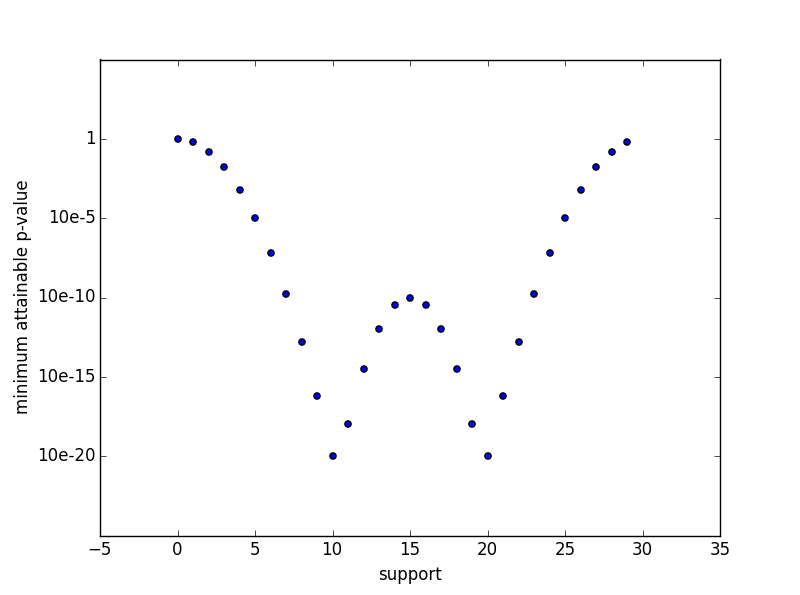


图4.1：整个数据集样本个数为30，正样本为10，对应每一个支持度的最小可达*p*值

Fig. 4.1 The number of sample is 30,the number of positive sample is 10,the relationship between support and minimum attainable p-value

在上述关于子图支持度与最小可达p值的关系中，两者之间的关系并不是单调关系，因此依然很难利用支持度对统计显著子图进行削减。因为在挖掘过程中，子图模式总是从小到大进行扩展，相应的其支持度总是由大到小变化。如果能够得到支持度阈值与统计显著子图之间存在单调关系，则当子图支持度下降到一定程度以后，该子图及其所有子图将不再有足够大的支持度成为统计显著子图。即，其支持度对应的统计显著性下界依然大于需要的统计显著性阈值。研究者们通过计算得出了支持度与最小可达*p*值与支持度之间的关系如图4.1。从图中可以看出，支持度与最小可达*p*值在整体上并不存在单调关系，整体曲线关于最大支持度的1/2对称。在左半部分，当支持度小于正支持度的时候呈单调递减，当支持度大于正支持度时单调上升。

为了能够利用子图的支持度削减统计显著子图的搜索空间，需要将最小可达性阈值与支持度的关系进行改造，使之具有单调关系。为了达到这个要求，研究者们利用一个更低的下界代替精确的最小可达*p*值下界：

 (4.5)

根据该*p*值下界可以得到新的待检验子图集合。由于，因此必然有。这也保证了，该转换不会导致显著子图结果的丢失。由于的单调性，可以得到由支持度表示的待检验子图集合：。在这里，。这种形式的变换，对于后续的算法设计有着重要的作用。因为可以从该式中发现，如果能够得到的值，就可以将统计显著子图挖掘转化为频繁子图挖掘算法。频繁子图挖掘已经在数据挖掘领域进行了广泛而又深入的研究。在基于置换检验的一些已有算法中，如FastWY，WestFall-Young-Light中，该性质有着重要的意义：如果中的所有模式的*p*值都能求出，并且其中最小的*p*值，那么不在待检验子图集合中的子图的*p*值将不可能小于。因此所有子图的最小*p*值，搜索过程只需要在待检验子图中进行即可。

### *p*值索引

在计算p值过程总，子图的统计显著性只与其正负样本支持度相关。如果两个不同的子图具有相同的正负样本支持度，则其具有相同的统计显著性。一个子图的最小可达*p*值仅与其在整体中的支持度相关[30]。根据以上性质，本文提出了两种*p*值索引的定义。

|  |
| --- |
| **算法4.2** 精确 *p*值索引算法  **输入：**精确*p*值索引表*T*，正样本集支持度*x+*，负样本集支持度*x-*  **输出：**精确p值δ |
| 1. **function** get\_exactp*(x+,x-)* 2. **if** *T(x+,x-)==0*//如果索引项不存在，则计算该索引项对应p值 3. *T(x+,x-)=p(x+,x-)* 4. **return** *T(x+,x-)* 5. **else** //索引项存在，直接返回对应p值 6. **return** *T(x+,x-)* 7. **end if** 8. **end function** |

**定义4.1**：**(精确*p*值索引)**在一个图数据库中，定义一个二维表，表中每一个元素为对应行和列作为正负样本支持度的子图的*p*值。

对于给定的子图*s*，假设其正样本集支持度为*x*+，负样本集支持度为*x*-，则可以从精确*p*值索引中快速得到对应的元素。对于一个给定的图数据库，正负样本集规模分别为*n*1，*n*2对于该数据库可以建立如图4.1所示的索引，以快速的获得特定子图的*p*值。对于每一个子图*s*，若已知其正负样本支持度，可以直接从表中取得对应的*p*值，这样，所有的*p*值最多将只需算一次，于是减少了大量的重复计算。

精确p值索引

图4.2 精确*p*值索引示例

Fig. 4.2 An example of exact p-value index

**定义4.2：(*p*值下界索引)**对于一个图数据库，定义一个数组，数组长度为图数据库中图模式个数。数组中每一个元素为对应支持度的*p*值。

在挖掘过程中，有时候需要通过支持度阈值下界作为削减规则。若该支持度下*p*值下界已小于目标*p*值，则该子图分支将被削减。对于一个给定的图数据库，利用一个数组预先存储不同支持度下的*p*值下界，如图4.2。

p

图4.3 *p*值下界索引示例

Fig. 4.3 An example of lower bound p-value index

因此，判断一个子图分支是否需要继续挖掘，只需要利用支持度，从*p*值下界索引中获取对应的*p*值下界，根据其值判断是否削减。其具体计算过程见算法4.2以及算法4.3。

|  |
| --- |
| **算法4.3** *p*值下界索引算法  **输入：***p*值下界索引表*I*，子图支持度*σ*  **输出：**该支持度下*p*值下界 |
| 1. **function** get\_lowerboundp*(I,σ)* 2. **if** *I(σ)*==0//如果索引项不存在，则计算该索引项对应的p值下界 3. *I(σ)=Ψ(σ)* 4. **return** *I(σ)* 5. **else** //如果索引项存在，直接返回对应p值下界 6. **return***I(σ)* 7. **end function** |

### 算法WEST-FALLYOUNG++

在本节中，将主要介绍如何利用上节中得到的支持度与显著性阈值的关系对图集进行挖掘，以获得在指定族错误率下的校正显著性阈值*δ\**以及要达到该显著性阈值所需的最小支持度阈值*σ*。

在4.2.1中讨论了子图统计显著性与子图支持度的关系。可以得到如下性质：

**性质4.1**：假设校正后的显著性阈值为*δ*，且，则可以利用支持度*s>σ*的子图的*p*值来精确估计挖掘结果的族错误率(FWER())。

**证明**：因为*s>σ*，则子图。设，若，则。利用的定义可以得出，。于是，。

利用这个性质，可以确保算法可以利用到所有有效的信息来计算。

首先，预先计算出类标签的*J*个排列。然后，对*k*=1初始化，和。在这里，初始化，于是相对应的待检验子图的支持度范围为。初始支持度阈值。在这里从*k*=1开始计算，跳过了*k*=0。因为，当支持度为1的时候就已经将所有的模式作为待检验子图加入显著子图候选集中。接下来，对子图进行枚举。假设在枚举过程中发现一个子图*s*，然后进行如下步骤：

1. 计算得到子图*s*的支持度*xi*，并检查该子图是否为待检验子图，如果不是则继续进行挖掘。
2. 如果该子图支持度在待检验子图支持度范围内，则对一个超几何随机变量预先计算该支持度下所有可能的*p*值。该值的取值决定于*x*，*x’*和*xi*。
3. 对于全部*J*个排列计算该子图在正样本集中的支持度，并且计算得到对应的*p*值。对于每一个排列，如果得到的*p*值小于当前最小*p*值，则将当前最小*p*值更新为该值。
4. 利用更新后的最小*p*值估计当前族错误率。如果族错误率太大，逐渐增加支持度下界*k*，这样就会减少显著性阈值，并且缩小的范围。直到族错误率小于，继续进行枚举。
5. 继续向下挖掘子图，直到没有重新启动子图挖掘过程。

该算法将一直进行下去，直到收敛到一个确定的支持度阈值，并且得到了。可以确定最优的显著性阈值一定位于内，并且尽管没有获得所有*J*次采样的最小*p*值，但获得了所有小于的最小*p*值采样。因为，所以通过这些样本可以获得最优的统计显著性阈值。同时，可以大幅度的减少计算量。该算法的伪代码，如算法4.2所示。

在该算法中，需要计算大量的*p*值。而经过分析发现，对于子图*s*1，*s*2，如果他们有相同的正负样本集支持度，那么他们具有相同的显著性*p*值。如果*s*1，*s*2具有相同的支持度，则他们具有相同的显著性阈值下界。于是，由于大量的子图模式具有相同的正负支持度，或者支持度，为了减少重复计算本文提出了利用*p*值索引来减少计算量。对于两种不同的*p*值，本章利用上文引入的两个*p*值索引进行快速计算*p*值。在算法4.4中利用了和两个函数分别用来快速计算显著性阈值下界和精确显著性值。其流程见算法4.3。

算法4.4利用置换检验获得了最优的统计显著性阈值*δ*\*。该算法生成了类标签的*J*次排列，并且在这*J*次排列下，利用一种子图挖掘算法，分别获得所有子图的最小*p*值。在计算该*p*值过程中，利用了支持度阈值从小依次增大，逐渐逼近最优支持度阈值，并获得最优显著性阈值的方法。同时利用*p*值索引的方法，对于所有需要计算的*p*值均只需计算一次。然后从*J*个最小*p*值中找到其分位数，作为最优的显著性阈值，使得挖掘结果满足预设的族错误率。并且，利用支持度与*p*的关系，可以获得一个子图成为显著子图的最小的支持度阈值。在下文中，将利用最优显著性阈值和最小支持度阈值来挖掘统计显著子图模式。

|  |
| --- |
| **算法4.4** WEST-FALLYOUNG++(WYPlus)算法  **输入：**图事务数据集*G*;族错误率阈值*α*;  **输出：**校正显著性阈值*δ\**;最小支持度阈值*σ*; |
| //主函数   1. **function** WYPlus(*α,G*) 2. **for** *j*=1:*J*: 3. *y(j*)=permute(*y*) //生成j个类标签的排列 4. *pminj*=1 //初始化每一个类标签组合下最小*p*值为1 5. **end for** 6. *σ*=1 //初始化最小支持度阈值为1 7. *δ*=get\_lowerboundp(*σ*) //初始化校正显著性阈值为当前最小支持度下界 8. Adjust\_pvalue(*root,n*) //从根节点开始对数据集进行挖掘 9. return *{pmin}j*的*α*分位数 //获得所有*pmin*中*α*分位数作为*δ\**返回 10. **end function**   //对子图枚举   1. **function** Adjust\_pvalue(*S,x*s) 2. **for** *j*=1:*J*: 3. *x+(j)*=get\_pos\_supp(*s*) //计算子图*s*在正样本集中的支持度 4. *pminj*=min{*pminj*,get\_exactp(*x+(j),x-(j)*)} //更新当前最小*p*值 5. **end for** 6. FWER*(δ)=1/JΣj=1J1*[*pmin(j)≤δ*] //估计当前族错误率 7. **while**(FWER(*δ*)>*α*) //根据族错误率大小调整支持度阈值 8. *σ=σ+1* //增大支持度阈值，减少显著性阈值，减少族错误率 9. *δ*=get\_lowerboundp(*σ*) 10. FWER*(δ)=1/JΣj=1J1*[*pmin(j)≤δ*] 11. **end while** 12. **for***s’* in *Children(s*) //对的子图继续枚举 13. *xs’=supp(s’)* 14. **if** *xs’≥σ* 15. Adjust\_pvalue(*S’,xs’*) 16. **end if** 17. **end for** 18. **end function** |

## 基于蚁群优化的统计显著子图挖掘算法

在本节中，对蚁群优化模型进行了扩展。蚁群优化算法是一种模拟蚂蚁群体智慧来求解组合优化问题的一种方法。子图挖掘也可以看做是一种组合优化问题。一个子图可以看做是若干条边组合起来的。因此，可以利用蚁群优化算法来解决子图挖掘的问题。

利用蚁群优化算法解决子图挖掘问题最将子图挖掘问题转化为一个优化模型。如何对子图进行编码，如何对边组合保证获得的子图为真正的子图，如何保证挖掘结果的多样性都是面临的问题。

### 优化模型

本文的目标是寻找一个子图模式集合，使得这个集合中的所有子图的*p*值小于或者等于最优的统计显著性阈值。于是，这个优化问题可以形式化的描述为如下的数学模型：

已知：*s*是一个子图，并且*f(s)*表示子图*s*的大小,*p(s)*表示子图*s*的显著性；

优化目标：

约束条件：

上述数学模型可以解释为，寻找图的尺寸尽可能大和显著性尽可能小的子图，使得该子图能够满足最优显著性阈值。为什么要求图的尺寸尽可能大呢？因为，对于子图挖掘，一个重要的应用方面是利用挖掘结果构建特征向量，然后对图数据集进行分类。如果只限制子图的显著性，可能会造成特征的冗余或者缺失。如：子图*s*是子图*s’*的超图，假设，如果只限制显著性最小会将*s*,*s’*都作为统计显著子图返回。然而，由于*s*包含*s’*，如果将*s*,*s’*都加入特征集，则造成了信息的冗余;如果只添加则造成了信息的缺失。因此选择了图的尺寸尽可能大，而显著性尽可能小的子图作为特征。下面引出了闭统计显著子图的定义：

**定义4.3**：**(闭统计显著子图)**如果*s’*不存在真超图*s*,使得*s’*，*s*有相同的显著性，则称子图*s’*在数据集中是闭的。如果子图*s’*在数据集中是统计显著的，则称*s’*是闭统计显著子图。

闭统计显著子图的概念刚好与整个优化问题的优化目标一致，接下来，将详细介绍如何利用蚁群优化算法，获得闭统计显著子图。在本节中将介绍如何改进蚁群优化算法用来从图数据集中挖掘闭统计显著子图模式。首先，传统的蚁群优化算法是针对于连通图的。对于子图挖掘算法，由于难以构造连通图，构造了一个单源单汇有向图，于是将子图挖掘问题转化为最短路径问题。

蚁群挖掘实例

图4.4 蚁群优化算法挖掘统计显著子图示例

Fig. 4.4 The example of mining statistical significant subgraph use ACO

在图4.4中，从一个空结点出发，第一层为所有的一边子图，空结点到一边子图存在一条边。第二层为两条边组成的子图。从第一层对应的边到第二层中包含该边的图存在一条边。以此类推。直到某一个图的所有子结点的*p*值均大于当前结点的*p*值，则将该子图作为结果加入候选集。并构建一条指向空结点的路径。于是，问题变成了寻找从源结点到汇结点的权重小于给定显著性阈值的路径。这就是整个问题的数学模型。

### 基于蚁群的子图搜索框架

在本节中，详细介绍了用蚁群优化算法挖掘统计显著子图的整个流程，并给出了算法的伪代码。蚁群优化过程主要包括(1)初始化信息素(2)蚁群多次迭代更新信息素(3)根据信息素含量获得解。接下来详细介绍ACOMiner算法也主要包括这几个步骤。

首先介绍在算法中用到的数据结构：1.边信息素向量*pher*：用来保存每条边上的信息素含量。2.结果集合*resultList*：将满足条件的解加入结果集合。

算法的主要步骤如下：

1. 枚举所有的一边子图。然后初始化边信息素向量。对于每一条边上的初始信息素，按照下面的公式设定：

 (4.6)

1. 在上式中*supp*(*G*-)表示该边在负样本集中的支持度，*supp*(*G*+)表示该边在正样本集中的支持度。
2. 在源结点处释放一定数量的蚁群。对于每一只蚂蚁，利用轮盘赌，根据边信息素向量选择一条边。
3. 对于每一只蚂蚁，列出所有候选的边等待向下挖掘。候选边需要满足以下条件：(1)将该边加入现有子图以后，新扩展得到的子图支持度大于最小支持度阈值；(2)该边需要与现有子图相邻接。例如：新加入的边连接了已有子图中两个不相连接的顶点，或者新加入的边引入了新的顶点。
4. 从候选边里选择一条边扩展当前子图进行扩展。该算法利用轮盘赌方法，根据信息素含量选择恰当的边扩展当前子图。选择各条边的概率可以利用下面的公式进行计算：

 (4.7)

1. 在上式中表示边*e*的信息素含量。表示边*e*的固有信息。即边*e*的初始化信息素含量。
2. 计算当前子图的*p*值。如果，其显著性低于，且其所有子结点的*p*值显著性均大于，则将该子图加入结果列表中。否则，转向第三步，继续进行扩展，直到候选边集合不存在边。
3. 当蚂蚁无法继续扩展时，更新经过路径的信息素。信息素的更新过程利用下述公式：

 (4.8)

上式中C为常数，Q为包含该边的子图的显著性阈值。

1. 当所有蚂蚁结束了一次迭代，每条边上的信息素会被释放掉一部分，用下面的公式模拟信息素释放的过程：

 (4.9)

1. 继续第二步，直到达到预定的迭代次数。

在本算法中，模拟了蚁群的行为用来引导整个子图空间的搜索过程。在搜索过程中也利用了支持度阈值来进行削减。主要用到了Apriori原理的第一条性质来进行削减。于是，整个搜索过程的时间复杂度大大的降低了。整个算法的伪代码如算法4.5。

|  |
| --- |
| **算法4.5** ACOMiner算法  **输入：**统计显著性阈值*δ*，支持度阈值*σ*，图事务数据库*G*  **输出：**统计显著子图模式集*sG* |
| 1. **function** ACOMiner(*δ,σ,G*) 2. EnumerateOneEdgeSubgraph**()**  //枚举所有一边子图 3. InitPher() //初始化边信息素向量 4. *sG*=InitResultList() //初始化结果向量集合 5. **for** *i*=1:*n* //n为迭代次数 6. **for** *j*=1:*m* //m为蚁群种群大小 7. **while**(*candidateList*!=NULL) //扩展子图，直到不能扩展 8. *candidateList*=createCandidateList(); //从空结点开始扩展子图，列出候选边 9. **if** *candidateList*=NULL 10. addSubgraph() //不能继续扩展，将该子图加入结果集合 11. updatePher() //更新信息素 12. **else** 13. ChooseEdge()//选择候选边 14. extendSubgraph() //扩展子图 15. **end if** 16. **end while** 17. **end for** 18. releasePher() //信息素挥发 19. **end for**   **20. return** *sG*  **21. end function** |

### 基于多样性的蚁群搜索优化

因为在统计显著子图挖掘过程中，要用挖掘结果尽可能的覆盖所有的具有显著区分能力的子图，因此如果挖掘结果的冗余程度较大，挖掘结果难以具有很好的覆盖度，结果作为分类标准存在较大的信息冗余。于是，需要在算法中不仅要保证挖掘结果具有较强的统计意义，还需要保证结果有较高的多样性。于是，在本文中，使用了一系列的算法来保证算法挖掘结果的多样性。

首先，在挖掘过程中如果得到两个统计显著子图和。若或者，返回较大子图。这种现象主要出现在一只蚂蚁在寻路过程中，会先后扩展得到两个子图，那么后者将完全包含前者，就将前者从挖掘结果中移除，而将后者加入挖掘结果，从而保证结果的多样性。

其次，通过参数来调节算法挖掘结果的多样性。在算法运行过程中通常倾向于选择信息素含量较大的边进行扩展[31]。如果信息素增长过快，就会使大量的蚂蚁都集中在同一条路径中，导致最后发现的结果都趋于一致，为了避免这种情况的出现，通过调整信息素挥发的速率来调整挖掘结果的多样性。如果信息素挥发程度较大，那么算法的多样性会增强，但是不容易找到显著性最强的子图。然而，如果挥发程度较小，又会过早的收敛到一个局部的最优解，造成了算法的早熟，也导致了结果的多样性较差。于是在挖掘过程中对挖掘结果的统计显著性进行监测。定义蚂蚁相遇的概念如下：如果两只蚂蚁在运行过程中发现了两个相同的子图，那么则称该蚂蚁相遇。在运行过程中，记录某一个子图被挖掘到的次数，若该次数超过一定阈值，则增大信息素挥发程度，以增强挖掘结果的统计显著性。若挖掘过程中得到的子图普遍统计显著性值较大，显著程度较弱，则减小信息素挥发程度，以使得进一步找到统计显著性较强的子图结果。

然后，在挖掘过程中主要调节蚁群路径的变量为边信息素。可以在蚁群寻找路径时加入一定的随机性。即，有一定的概率蚁群不按照信息素浓度划分路径选择概率，而是按照等概率来选择下一步要选择的路径。这样，会进一步增加搜索范围，增加不确定性，以保证挖掘结果的多样性。

最后，对挖掘结果进行一些后处理。对挖掘结果中各个子图所含的边的集合进行比较。若集合相似度较大，则比较两个子图的极大公共子图，若该极大公共子图尺寸超过一定阈值，则取两者中较大的子图作为结果。集合相似度可以利用Jaccard距离等来度量。

结合上面的几个策略，可以保证挖掘结果具有较高的多样性，能够有较强的覆盖能力，尽可能少的信息冗余。以这样的结果作为分类指标，在一定意义上可以提高分类算法的性能。蚁群算法的多样性也是研究人员们重点着力的一个方向。

### 算法示例

为了更直观地描述算法的执行过程，在本节中给出了一个完整的例子来描述利用蚁群进行挖掘的整个过程。

如果在大规模数据集上运行算法，难以直观的理解算法的运行方式。为了便于了解整个挖掘过程，人工制造了一个只有八个图模式的分类图数据集。其中正类样本有四个，负类样本有四个。虽然小样本量难以衡量统计效力。但是，可以一定程度上模拟算法的流程，并在一定程度上展示算法的合理程度。下面，描述了在该数据集上进行统计显著子图挖掘的过程。

示例图数据集

图4.5 蚁群优化算法挖掘统计显著子图示例图数据库

Fig. 4.5 The database of graphs

对图4.5中表示的图数据集进行统计显著子图挖掘：

频繁一边子图

图4.6 频繁一边子图

Fig. 4.6 Frequent One-Edge Subgraph

假设，在统计假设检验校正过程中得到支持度阈值为3，则频繁一边子图如下图4.6所示。其正负样本支持度分别为(4,1)，(3,1)，(2,1)，(2,2)。按照信息素初始化规则：正样本支持度和负样本支持度中的较大值与较小值的比值。如图A-C，其正负样本支持度分别为2,2，其初始化信息素浓度为2/2，即为1；图A-D，其正负样本支持度分别为2,1，于是其初始化信息素浓度为2/1，即为2。根据上述规则，初始化边信息素列表为(4,3,2,1)。接下来开始利用蚂蚁寻找图数据库中的统计显著子图。

第一只蚂蚁先根据边信息素列表确定要开始扩展的边。通常利用轮盘赌的方式，根据信息素大小按分配选择边的概率。于是信息素较多的结点有更大的概率被选择作为初始边。

假设第一步选择了A-B，于是可以从A-B开始扩展。与A-B相邻的边有B-C，A-C，A-D。将这三条边作为候选边，从这三条边中选一条进行扩展。首先判断扩展以后的支持度，若扩展后的支持度小于预设支持度阈值，则将该边排除于候选边。然后，根据候选边的信息素浓度确定概率大小。

在图4.5所示的数据集中，候选边扩展后频繁的只有边A-B。从现有B-C扩展，有一个位点可以插入边A-B，于是从顶点B处扩展子图，将A-B加入到现有子图中。于是，得到扩展后的子图为图4.7。

挖掘结果示例

图4.7 挖掘结果示例

Fig. 4.7 The result of subgraph mining

然后利用显著性值计算公式，同时能够计算得出该子图的正负样本支持度。根据该子图的正负样本支持度，得到该子图统计显著性。若其显著性小于预设统计显著性，则将其加入候选结果列表。

然后，蚂蚁继续向下扩展。从边集合中寻找与当前子图相邻的边，同上一步确定候选边集合。然后根据边信息素选择要扩展的边。若扩展以后的子图值小于该只蚂蚁该次迭代发现的上一个统计显著子图的值，则用该子图代替前一个子图。若候选边集合为空，则结束迭代。

此时，对该次扩展得到的最后一个统计显著子图中包含的边的信息素进行更新。增加其对应边上的信息素含量。为什么要对最后一个统计显著子图包含的边的信息素进行更新呢？因为，考虑到如果对每一个生成的子图都要对其中所包含边的信息素进行更新，会使得某些边多次增加信息素，使得信息素更新不公平，更偏向于这些边。

上述过程是若干只蚂蚁同时进行的。若所有蚂蚁均走到终点，即无法继续扩展，则考虑信息素释放。然后进行下一次迭代。

直到达到预设迭代次数，算法结束。此时得到统计显著子图集合。

通过上述实例，进一步加深了对算法的了解。

## 本章小结

本章对大规模图数据中统计显著子图挖掘问题进行了深入的研究与分析，提出了一种基于假设检验和演化计算的统计显著子图挖掘算法PABSMiner。该算法首先利用置换检验模拟出数据的零分布，获得了在一定族错误率下统计显著子图的显著性阈值，以及对应的支持度阈值。同时该算法在计算统计显著性的过程中，利用*p*值过滤原理，结合两种*p*值索引来提高算法的运行效率。接下来，设计了基于蚁群算法的挖掘框架，利用群智能算法，加快子图的搜索过程，提高算法的效率。在搜索过程中充分运用蚁群的群体智能，并采用一系列的削减策略如支持度阈值，显著性阈值等等进一步减少了搜索空间，从而提高了算法的执行效率。同时也采用了一系列的措施来增强算法的多样性，提高挖掘结果的覆盖能力，减少信息冗余。最后，通过一个完整的实例，在自己设计的图模式数据集上，描述了算法的整体执行过程，以进一步展示算法的原理。进一步帮助读者理解该算法的原理。

# 实验测试及结果分析

本章主要从算法的效率，子图的显著性和图分类的准确性对提出的基于多重假设检验校正的统计显著子图挖掘算法进行分析和验证，并对比了该算法与过去算法性能的比较。对比其在准确率和时间效率方面的区别。通过对人工合成数据集和真实数据集的实验，证明了用于大规模图数据分类的统计显著子图挖掘算法的高效性和有效性，并且进行了分类准确性的验证。

## 实验环境与数据集

实现语言：Java

CPU：Intel(R)Core(TM)i5-6500 CPU @3.20GHz

内存：16.00GB

操作系统：windows7(64bit)

数据集：在算法的实验测试过程中，使用了多个NCI开源数据集提供的真实数据集，mutag数据集以及若干合成数据集作为实验数据。数据集具体信息如下：

NCI数据集：NCI癌症频谱数据集是一种被广泛用于图分类算法的验证工作中。本文从PubChem数据库中下载了10个NCI数据集。每一个数据集归属于一个抗癌活性预测的生物测定任务。即，如果在一个数据集中一个化合物分子对于对应的癌症具有抵抗活性，将这类分子作为正样本。其余化合物分子样本为负样本。表5.1列出了十个数据集的汇总信息。因为数据集的正负样本不平衡性，大约其中5%为正样本，对于每一个生物检定任务，随机选择与正样本规模相同数量的负样本作为负样本数据。于是获得了十个平衡的数据集，共包含56000个图模式。把这10个数据集合并构成了一个新的NCI-A数据集。与此同时，本文也利用10个数据集中的所有图模式构建了一个新的NCI-B数据集，该数据集是不平衡的。

Mutag数据集：Mutag数据集包含188个化学分子，按照其是否具有诱变性质分为正类和负类。具有诱变性质的认为是正类，没有诱变性质的认为是负类。其中正类有125个，负类有63个。

表5.1 NCI数据集情况汇总

Table5.1 The Description of NCI Datasets

|  |  |  |  |
| --- | --- | --- | --- |
| 生物检定类别ID | 对应癌症类型 | 正样本数 | 负样本数 |
| NCI1 | 肺癌 | 2047 | 38410 |
| NCI33 | 恶性黑色素瘤 | 1642 | 38456 |
| NCI41 | 前列腺癌 | 1568 | 25967 |
| NCI47 | 神经系统瘤 | 2018 | 38350 |
| NCI81 | 结肠癌 | 2401 | 38236 |
| NCI83 | 肺癌 | 2287 | 25510 |
| NCI109 | 卵巢癌 | 2072 | 38551 |
| NCI123 | 白血病 | 3123 | 36741 |
| NCI145 | 肾癌 | 1948 | 38157 |
| NCI167 | 酵母抗癌 | 8894 | 53622 |

合成数据集：本文利用一个合成图数据生成器用来生成100000个无向有标签的连通图。并且随机选择其中的一半作为正样本，另一半作为负样本。设定，顶点的标签一共有20个。在每个图中平均的顶点数和边数分别为20和30。

## 算法性能评价

在本节中，研究了不同数据集以及不同的参数下，算法的时间效率和空间效率。算法的参数主要有族错误率阈值，蚁群算法的参数：信息素权重，信息素挥发率，迭代次数*n*，蚁群种群大小*m*。本节主要针对不同的族错误率阈值下，对多个数据集的运行时间与内存使用峰值进行了比较。

图5.1显示了在不同数据集下，使用不同的族错误率对于运行时间的影响。对于较小的族错误率阈值，由于限制较为严格，可削减搜索空间较大，于是在一系列的削减策略下可以较为明显地降低算法的运行时间。由实验数据也可以得出，算法的运行时间与数据集中图模式个数以及图模式的尺寸都有关系。随着图模式的个数增加，算法的运行时间增长。随着图模式尺寸的增加，算法运行时间增长。

图5.2显示了在不同数据集下，不同的族错误率阈值下，该算法内存使用峰值情况。不同的族错误率阈值会有不同的候选子图以及较小的*p*值索引表，于是占用较小的内存。

图5.1 不同数据集下不同族错误率阈值运行时间对比

Fig. 5.1 The comparision of running time between different datasets with different threshold of FWER

图5.2 不同数据集下不同族错误率阈值内存使用峰值对比

Fig. 5.2 The comparison of maximum memory used between different datasets and threshold of FWER

图5.3显示了在不同的数据集下，三种算法的运行时间效率。可以看出，PABSMiner算法与GraphSig算法基本持平，而大幅度优于传统的FastWY算法。这体现了削减策略和演化计算策略的有效性。

图5.4显示了在不同的数据集下，三种算法的内存使用峰值比较。可以看出，PABSMiner算法在提升算法时间效率的同时，空间使用效率基本与现有算法持平。并且由于减少了搜索空间，避免了生成候选子图过程，节省了较多的空间，在某些数据集下比其他算法更加节省空间。

图5.3 对比GraphSig，FastWY，PABSMiner算法的运行时间

Fig. 5.3 The comparison of run time between different algorithms

图5.4 对比GraphSig，FastWY，PABSMiner算法的内存使用峰值

Fig. 5.4 The comparison of maximum memory used between different algorithms

由于在算法中使用了一系列的削减策略以及索引技术减少了子图搜索空间的范围，同时采用智能优化计算的方法加速了算法的搜索度，因此在时间性能上本算法要优于传统的统计显著子图挖掘算法。在算法中使用到了*p*值索引，索引表的大小与数据集中图模式数目二次相关，同时减少了待选子图的数目，因此内存使用峰值与传统算法基本上持平。

## 子图挖掘结果显著性评价

在本节中从两个角度研究了子图挖掘结果的统计效力。一是从挖掘结果的经验族错误率来比较不同算法之间挖掘结果的统计效力；二是利用挖掘结果构建特征向量，训练图分类器，通过分类器的准确率比较挖掘结果的效力。

### 挖掘结果统计效力的比较研究

对于挖掘结果的族错误率，主要考虑了置换检验的排列次数*J*对于挖掘结果的影响。本文用上述四种数据集进行了测试实验。

(b)Synthetic

(c)NCI-B

(b)NCI-A

(a)mutag

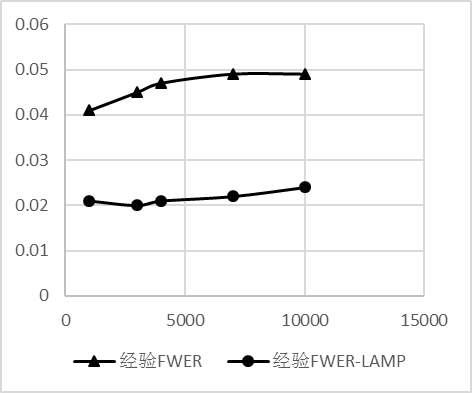
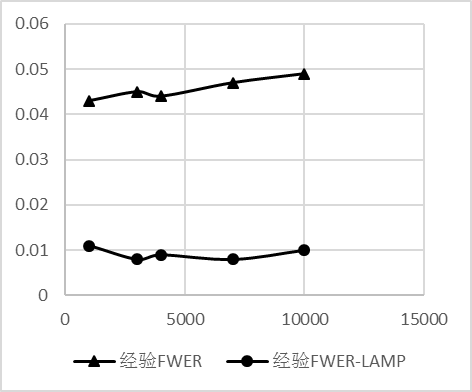
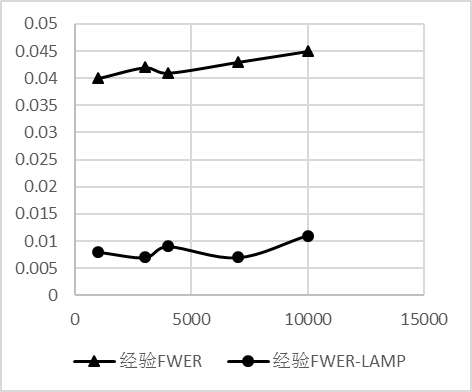
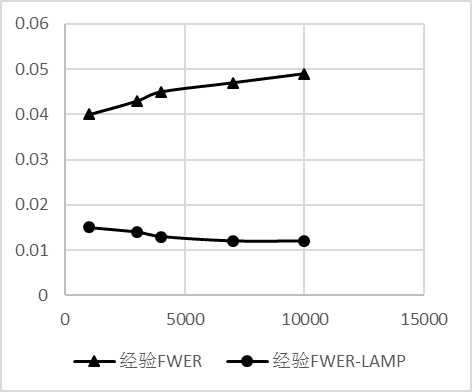


图5.5 不同数据集下两种算法的经验族错误率随着置换次数增加变化趋势

Fig. 5.5 The relationship between empirical FWER and times of permutation

图5.5显示了四种数据集下PABSMiner与LAMP两种算法下挖掘结果经验族错误率对比。由图中可以看出，PABSMiner算法下，设定族错误率阈值为0.05，随着置换检验排列次数的增加，挖掘结果的族错误率将会越来越接近预设阈值。因此*J*越大，对最优显著性阈值的估计将会越准确，得到的挖掘结果的族错误率也会越接近预设值。但是，考虑到*J*的增加会大幅度增加算法的时间花费，于是从实验结果中得到当*J*=10000时，已经可以获得较准确的最优显著性阈值。因此，*J*设定为10000作为最佳置换次数。

进一步，图中描绘出利用LAMP算法获得的校正显著性阈值所得到挖掘结果的族错误率变化趋势。LAMP算法是第一个有效的控制族错误率的算法之一。实验表明LAMP算法是一个很保守的算法。该算法试图把经验族错误率控制在到之间。在实际运用中，即使只是对目标族错误率减半，也会对挖掘结果的统计效力产生巨大的影响。在处理巨大的多重假设检验问题中，主要的目标是寻找潜在的对于目标变量有影响的因子，而如果将族错误率控制的过于严格会使得大量的统计显著结果丢失，造成大量的有效信息的缺失。本文的算法在经验族错误率的稳定性上可以获得较好的效果。

实验中选择经验族错误率作为对比的标准是因为经验族错误率是对于挖掘结果的统计效力的最好代表。如果一个算法设定族错误率阈值为，该阈值对于所有的统计检验过程是最优的。如果经验族错误率低于该阈值，算法会因为过于保守而损失一部分统计效力。因此，经验族错误率越接近，但是不大于，算法的性能越好。

综上所述，该实验表明(1)采用置换检验校正统计显著性阈值对于结果统计效力的提升有重要影响，(2)随着J的增加，当J超过一定值以后其对于结果的统计效力的影响会变的很小。

### 挖掘结果用于图分类的实验研究

统计显著子图挖掘一个重要的应用就是利用挖掘结果中的统计显著子图模式，构建特征向量，对图数据库进行分类。因此，挖掘结果用于图分类的准确率可以很好的显示子图挖掘结果的有效性。将子图挖掘算法用于图分类的算法流程如算法5.1。

图分类算法主要包括以下几个步骤：

1. 统计显著子图挖掘获得特征子图：在前面的章节详细介绍了统计显著子图挖掘的算法。相当于机器学习中的特征选择问题。
2. 向量化测试集和训练集：假设*g*1…*g*n是从训练集中挖掘得到的显著子图模式。对于测试集和训练集中的一个图模式*G*，可以被表示为一个特征向量。如果，则有，反之则。由此可以将图模式*G*转化为一个特征向量。一旦获得图*G*的向量表示，就可以利用传统的分类器对图进行分类。
3. 训练分类器并对测试集类别进行预测：支持向量机模型将实例表示为空间中的点，然后选择一条最优分割平面使得不同类别的实例被尽可能宽的间隔所分开。对于预测集中的实例，将其置于对于的实例空间中，并根据他们落在分割平面的哪一侧来预测他所属的类别。由于其对复杂建模问题的支持较好，使得其准确性得到了较大的提升。并能够避免其他方法所存在的过拟合问题。支持向量机通过对特定实例的学习准确性和学习能力之间进行平衡，获得了较好的性能，因此有很广泛的应用。在本文的实验中，利用LIBSVM，并使用由作者提供的默认参数作为分类器。LIBSVM是一个集合了支持向量分类，支持向量回归，以及分布估计的一个整合软件包。支持多类分类。

|  |
| --- |
| **算法 5.1** 图分类算法  **输入：**训练集*GTrain*，测试集*GTest*  **输出：**测试集的分类结果*result* |
| 1. **function** GraphClassify(*GTrain,GTest*) 2. //挖掘统计显著子图 3. *sigGraph*=ACOMiner(*α,G*) 4. //向量化测试集和训练集 5. vectorizedGraph(*sigGraph,GTest,GTrain*) 6. //训练分类器 7. svmTrain(*GTrain*) 8. //对测试集进行预测 9. *result*=svmPredict(*GTest*) 10. **return** *result* 11. **end function** |

在这一部分实验中本文通过两个角度对图分类的性能进行比较。一个是通过对分类器的AUC值进行比较；另一个方面是通过对不同算法的分类结果规范化准确率进行比较。因为算法要比较不同统计显著子图挖掘的统计效力，于是统计显著子图挖掘部分采用了三个不同的算法：PABSMiner即本文所提出的算法；FastWY一种最新的基于置换检验的统计显著模式挖掘算法；GraphSig，一种没有考虑多重假设检验校正的传统的统计显著子图模式挖掘算法。对于三种方法采用相同的分类器算法以及参数。已保证结果只与统计子图挖掘结果相关。

图5.6 三种算法挖掘结果用作分类特征AUC对比

Fig. 5.6 The comparison of AUC with different classification feature

图5.6中展示了三种算法对于分类结果AUC的影响。上述实验结果表明，本文提出的算法在大幅度缩短算法时间的情况下可以获得与FastWY相似的AUC值。相比GraphSig,，分类器AUC值有明显提升。

图5.7 三种算法分类结果的正则化准确率对比

Fig. 5.7 The normalized accuracy of different classification feature

图5.7显示了三种算法对于正则化准确率的影响。正则化准确率计算公式如式5.1[33]。

从图中可以得出，FastWY与PABSMiner在分类准确度上基本上持平，效果均好于未考虑多重假设检验问题的GraphSig算法。因此，基于群智能算法的PABSMiner算法在大幅度提升算法运行效率的同时，也保证了挖掘结果统计效力。

 (5.1)

## 本章小结

本章通过实验验证了基于蚁群优化算法的统计显著子图挖掘算法的高效性，验证了基于多重假设检验校正的统计显著子图挖掘算法的有效性，并检验了算法的挖掘结果的可靠性和对于分类的准确性。通过在真实数据集和人工数据集上运行算法所得到的结果，表明将演化计算引入统计显著子图挖掘过程中，可以在提高算法效率的同时保证挖掘结果依然很好。

# 结束语

在图分类研究中，图分类依据从频繁子图到区分子图，然后逐渐发展到统计显著子图。这些方法不断的改进对子图区分能力的评价标准，从而为用户决策提供具有价值的信息。近年来，随着大规模数据的不断发展，复杂数据结构的不断产生，对于图数据挖掘的需求也在日益旺盛，关于图数据挖掘的区分模式的研究热度逐渐增加。然而，由于数据量的急剧增加，图数据结构的复杂性导致现有算法的时间性能不足。其数目庞大的子结构使得其上的挖掘工作变得异常复杂。因此，对于图数据上区分模式挖掘的研究还亟待发展。

本文就大规模图数据中统计显著子图挖掘过程中的时间效率问题和挖掘结果的假阳性比例问题进行了深入的研究。提出了基于蚁群优化算法的统计显著子图挖掘算法。与以往统计显著子图挖掘的算法不同，本文采用了基于演化计算的搜索框架，有效地减少了候选子图生成过多，搜索冗余过的等问题。此外，为了进一步避免挖掘过程中遇到的计算瓶颈，在计算显著性的时候运用了两种p值索引，以及一系列的削减策略，有效的减少了重复计算。

该算法首先利用基于置换检验的多重假设检验校正方法，得到校正后的统计显著阈值，然后利用该统计显著阈值进行挖掘统计显著子图。本文的基本思想包括两个方面：一是利用索引减少重复的*p*值计算时间。在假设检验中，显著性p值的计算是一个相当耗时的步骤，减少这一块的重复计算，可以节省大量的时间。二是不断减少搜索空间，以加快算法的搜索效率。在这方面主要运用了基于蚁群的优化算法，利用群智能算法的优势，通过大规模蚁群对于搜索方向的指引，减少对不必要子图的搜索过程。极大地提高了算法的时间效率。另一方面，还运用了一系列的剪支策略，当搜索过程中满足剪支条件以后，就可以将该分支削减从而极大的节省计算时间，提高算法效率。

最后，根据在真实数据集和人工数据集上的大量实验，通过对不同参数的研究对比以及几种区分子图挖掘算法的对比，证明了提出算法的高效性和有效性。最后我们得出结论，将蚁群优化算法引入统计显著子图挖掘过程中，在极大的提高了计算效率的同时，结果质量几乎保持了同类算法的效果。从而为图数据集上区分序列挖掘的研究提供了一个崭新的视角。

本文提出的算法依据从现实应用中抽象出来的图模型，在具体的应用中可能存在无法直接套用的问题，这就需要根据具体的应用场景来调整算法。同时，本文算法的效率由于计算量的原因也有待提高，所以，今后的研究重点主要有两个方向：一是提出针对大规模数据集的并行挖掘算法，这样可以通过增加机器而提高算法的执行效率。二是研究更加有效的削减策略，以进一步提高算法的效率和准确性。

本文继承了前人的研究基础，取得了一定的研究成果，在一定程度上达到了预期的效果。随着网络数据的飞速发展与革新，各种大规模的以图模式存储的数据引起了研究人员们高度关注。在未来的研究过程中，如何利用新的方法更高效，准确的从这些巨量图模型中挖掘到人们需要的知识结构将成为一个新的研究重点。今后的研究需要进一步的对统计显著子图挖掘以及大规模图数据集上的相关挖掘算法投入更大的精力与物力来把这个领域做得更加深入，透彻，以使其更好的服务于人类社会发展。

参考文献

1. Inokuchi A, Washio T, Motoda H. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data[A].Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery[C]. Springer-Verlag, 2010. 13-23.
2. Kuramochi M,Karypi G.Discovering Frequent Geometric Subgraphs[A]. In Proceedings of the IEEE International Conference on Data Mining[C]. 2002. 258–265.
3. Yan X,Han J.gSpan:graph-based substructure pattern mining[A]. international conference on data mining[C].2002. 721-724.
4. Yan X, Han J.CloseGraph: mining closed frequent graph patterns[A]. knowledge discovery and data mining[C]. 2003. 286-295.
5. Huan J, Wang W, Prins J F, et al. Efficient mining of frequent subgraphs in the presence of isomorphism[A]. international conference on data mining[C]. 2003. 549-552.
6. Wang W, Wang C, Zhu Y, et al. GraphMiner: a structural pattern-mining system for large disk-based graph databases and its applications[A]. international conference on management of data[C]. 2005. 879-881.
7. Zhu F, Yan X, Han J, et al. gPrune: a constraint pushing framework for graph pattern mining[A] Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining[C].Springer-Verlag, 2007. 388-400.
8. Yan X, Cheng H, Han J, et al. Mining significant graph patterns by leap search[A]. international conference on management of data[C].2008. 433-444.
9. Jin N, Young C, Wang W, et al. GAIA: graph classification using evolutionary computation[A]. international conference on management of data[C]. 2010. 879-890.
10. Ranu S, Singh A K. GraphSig: A Scalable Approach to Mining Significant Subgraphs in Large Graph Databases[A]. international conference on data engineering[C]. 2009. 844-855.
11. He H, Singh A K. GraphRank: Statistical Modeling and Mining of Significant Subgraphs in the Feature Space[A]. international conference on data mining[C]. 2006. 885-890.
12. Chen M, Gao X, Li H, et al. An efficient parallel FP-Growth algorithm[A]. cyber-enabled distributed computing and knowledge discovery[C].2009. 283-286.
13. Thoma M, Cheng H, Gretton A, et al. Discriminative frequent subgraph mining with optimality guarantees[J]. Statistical Analysis and Data Mining, 2010, 3(5): 302-318.
14. (美)Anand Rajaraman, Jeffrey David Ullman著.大数据:互联网大规模数据挖掘与分布式处理[M].王斌译.人民邮电出版社, 2012.
15. 裴艳波. 多重假设检验问题中关于三种错误测度-FWER,FDR和pFDR的讨论[D]. 长春：东北师范大学,2005.
16. Ranstam J. Multiple p-values and Bonferroni correction.[J]. Osteoarthritis & Cartilage, 2016, 24(5):763-764.
17. Benjamini Y, Hochberg Y. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing[J]. Journal of the Royal Statistical Society, 1995, 57(57):289-300.
18. Rohart F. Multiple Hypothesis Testing for Variable Selection[J]. Australian & New Zealand Journal of Statistics. 2016, 58(2):245-267.
19. Liu G, Zhang H, Wong L. Controlling false positives in association rule mining[J]. Proceedings of the Vldb Endowment, 2011, 5(2):145-156.
20. Chung E Y, Romano J P. Multivariate and multiple permutation tests[J]. Journal of Econometrics. 2016, 193(1):76-91.
21. Webb G I. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns[J]. Machine Learning. 2008, 71(2):307-323.
22. 曾宇恒, 宋留静, 白嘉豪, et al. 蚁群算法行为属性的改进解决QoS组播路由优化问题[J]., 2015, 0(7): 141-142.
23. Dorigo M, Maniezzo V, Colorni A. Ant system: optimization by a colony of cooperating agents.[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 1996, 26(1):29-41.
24. 王涌鑫, 王希武, 钟启艮. 一种基于搜索多样性的改进蚁群算法研究[J]. 计算机与数字工程, 2013, 41(6):896-898.
25. Traversa F L, Ventra M D. Universal Memcomputing Machines[J]. IEEE Transactions on Neural Networks, 2015, 26(11).
26. Irwin J O. Test of Significance for Differences Between Percentages Based on Small Numbers[J]. Metron, 1935, 12(2).
27. Kim H J, Fay M P, Feuer E J, et al. Permutation test for joinpoint regression with applications to cancer rates[J]. Statistics in Medicine, 2000, 19(3): 335-351.
28. Llinares L, Pez F, Sugiyama M. Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015. 725-734.
29. Sugiyama M, López F L, Kasenburg N, et al. Significant Subgraph Mining with Multiple Testing Correction[C]. Proceedings of the 2015 SIAM International Conference on Data Mining. 2015.
30. Zhang X, Zou F, Wang W. FastANOVA: an Efficient Algorithm for Genome-Wide Association Study.[J]. KDD Proceedings, 2008. 821-829.
31. 嵇可可. 基于动态趋势预测蚁群算法的云计算资源调度优化研究[J]. 科技通报, 2016, 32(1):187-190.
32. Bantis L E, Feng Z. Comparison of two correlated ROC curves at a given specificity or sensitivity level[J]. Statistics in Medicine, 2016.
33. Wang Z, Zhao Y, Wang G. Large-Scale Graph Classification Based on Evolutionary Computation with MapReduce[M]. Web Technologies and Applications. Springer International Publishing, 2015.

致谢

时光匆匆，毕业季又一次来临。两年半的研究生旅途已经接近尾声。回首往事，这两年半来既有机遇也有挑战，有意外也有惊喜。在这里遇到的形形色色的人和事，已经被牢牢的印在脑海中。这段旅程的磨练，让我收获了许多。不论是在学习还是生活都有着巨大的收获。这些收获与这里的众多良师益友的帮助是分不开的。在此谨对这些帮助致以衷心的感谢。

首先，我要感谢我的导师赵宇海教授。在研究生阶段，赵老师在各个方面给予了我细致入微的教导和关怀。赵老师虽然日常工作繁忙，但是仍然保持了高度严谨的治学态度，经常在例会讲座中指导帮助我们，而且用敏锐的学术洞察力为我们的科研指出了明确的方向，帮助我们养成良好的研究习惯。

感谢实验室里的各位老师，感谢你们构筑了实验室里浓厚的学术氛围。感谢印莹，韩东红，王波涛，乔百友等几位老师。

感谢实验室的师兄师姐们，王章辉，李源，甘锐，毕天驰，夏文春，张丽君，姜鹏。感谢他们在学业上给予的帮助和耐心的指导。他们对待学习的认真态度令人敬佩，是我的好榜样。

感谢我的同学们，郎文博，李强，李向南等，感谢他们对我的无私帮助，与他们在一起生活学习感觉非常幸福。

感谢我的父母，感谢父母的养育教导和关心照顾，他们是我前进的灯塔和力量源泉。

最后谨向百忙之中抽出时间评审本论文的专家致以最诚挚的谢意！您辛苦了。

攻读硕士期间发表论文和参与项目

攻读硕士期间参与项目：

1.国家自然科学基金重点项目: 集群环境下基于内存的高性能数据管理与分析(61332006)