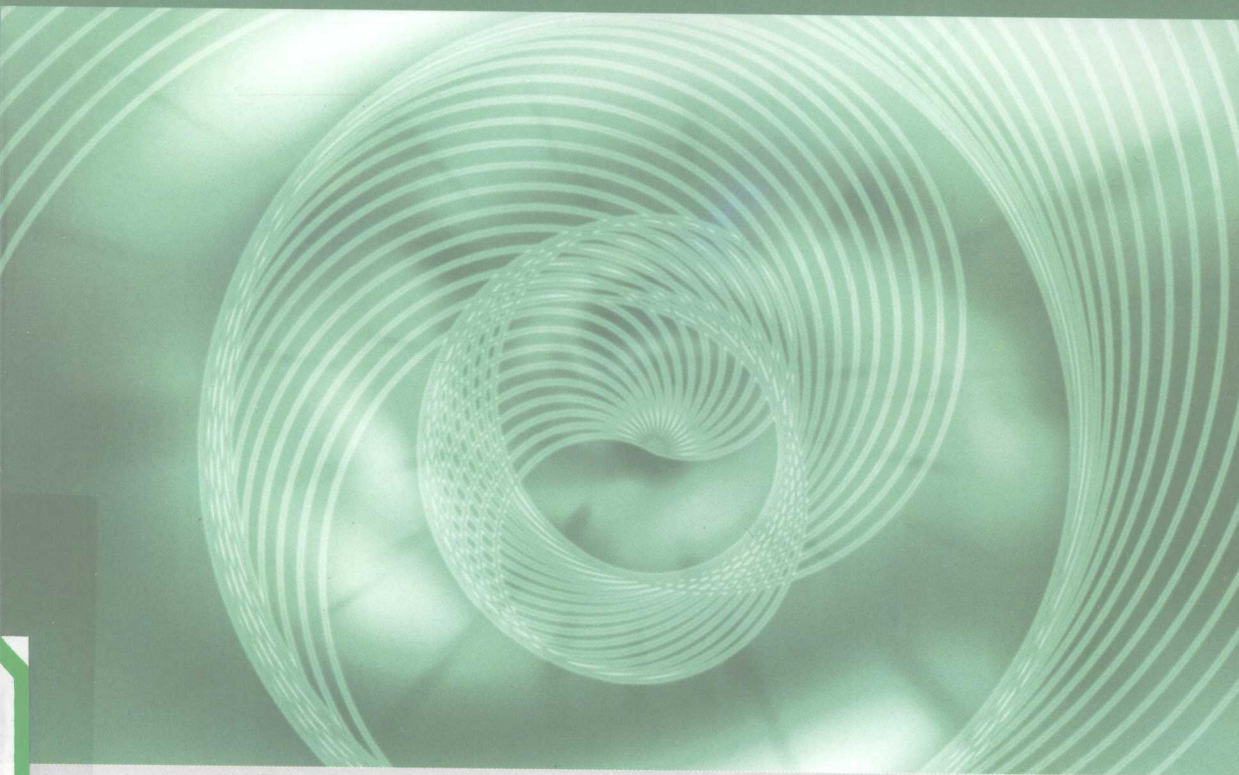
基于混合进化的子结构发现

常新功 著



国防工业出版社
National Defense Industry Press

基于混合进化的子结构发现

常新功 著

国防工业出版社

·北京·

内 容 简 介

本书介绍了使用进化算法进行图学习的一些概念、思想、方法和技术。全书共分7章，其中前3章为基础篇，介绍了图学习的基本概念、基本思想、发生发展历程、应用领域和典型的图学习算法Subdue系统，另外还介绍了进化算法的基本理论、基本思想、典型范式、一般框架、各个组成要素、典型实例和一个基于进化规划的子结构发现算法EPSD。第4章～第6章为算法设计篇，分别介绍了基于混合进化、基于回溯机制、基于带全部实例的个体表示和基于个体协同的四种混合进化子结构发现算法。第7章为应用篇，介绍了子结构发现算法在学科建设、区域经济研究、地震数据分析和反恐数据分析中的四个典型应用。附录中还给出了本书用到的多个图数据集。

本书可供所有从事机器学习和数据挖掘的专业技术人员阅读和使用，也可供管理科学和系统工程专业的读者学习参考。

图书在版编目(CIP)数据

基于混合进化的子结构发现/常新功著. —北京：国防工业出版社，2009.1

ISBN 978-7-118-06066-9

Ⅰ.基… Ⅱ.常… Ⅲ.子结构法 Ⅳ.0241.82

中国版本图书馆CIP数据核字(2008)第190976号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路23号 邮政编码100048)

天利华印刷装订有限公司印刷

新华书店经售

*

开本 710×960 1/16 **印张** 13½ **字数** 236千字

2009年1月第1版第1次印刷 **印数** 1—4000册 **定价** 25.00元

(本书如有印装错误，我社负责调换)

国防书店：(010)68428422 发行邮购：(010)68414474

发行传真：(010)68411535 发行业务：(010)68472764

前　言

关系无处不在。近年来，学习数据中各种关系的图学习和图数据挖掘得到了广泛的关注并成为数据挖掘与机器学习领域的一个重要分支。图是通用的数据表示形式，与其他表示形式相比，图数据含有更为丰富、复杂的信息，更能反映问题的本质，同时也会导致更大、更复杂的假设空间，从而向数据挖掘与机器学习领域提出了新的挑战。本书将在众多复杂问题求解过程中表现突出的进化算法与简洁高效的局部搜索算法相混合，形成混合进化算法，并应用于图学习和图数据挖掘的子结构发现之中，取得了较同类算法更好的实验结果。

本书的前3章为基础篇，分别介绍了图学习和进化算法的基本概念、基本思想和基本理论以及相应的应用领域和具有代表性的算法实例等内容。

本书的第4章～第6章为算法设计篇，其主要内容有：

1. 依据混合进化算法理论提出了混合进化子结构发现算法HEASD。在HEASD中，给出了基于图的染色体表示和遗传算子，并将爬山算法的思想融于交叉和变异算子的设计之中，实验结果表明了该算法的有效性。同时本书还提出了一种新的子结构扩展方法——单标签扩展，并对其正确性和有效性进行了理论证明和实验验证。

2. 子图同构问题是图学习和图数据挖掘的瓶颈问题，是造成问题复杂的根源所在。其表现之一就是它造成了进化的单向性，从而导致了查找的不完全性。为此提出了基于带回溯个体的混合进化子结构发现算法HEASDBT，将回溯机制融入到了进化过程之中，可以对假设空间中的某些关键区域进行密集搜索，实验结果表明了该算法的有效性。

3. 实例丢失现象是图学习中广泛存在的造成解质量降低的一个重要原因。为此提出了两个算法HEASDFI和HEASDCI，前者采取“预防”的策略，尽量避免实例的丢失；后者则采取“治疗”的办法，重新找回丢失的实例。实验结果表明

了以上两种算法的有效性。在 HEASDCI 中，还提出了一个新的遗传算子——个体协同算子，使多个代表同一子结构的不同个体可以对同一目标进行协同查找，以提高解的质量。由于个体协同算子需要进行频繁的图同构操作，而图同构操作虽然不像子图同构那样已被证明是 NP 完全问题，但目前还没有多项式级的算法存在，为此本书提出了一个时间复杂度为多项式级的近似图同构算法以提高个体协同算子的执行效率。

本书的第 7 章为应用篇，详细介绍了子结构发现算法在学科建设、区域经济研究、地震数据分析和反恐数据分析中的四个典型应用。

本书内容主要取自于作者 2005 年—2008 年在天津大学攻读博士学位期间的研究成果，期间得到了寇纪淞教授和李敏强教授的精心指导和指引启迪，在此向两位导师表示衷心的感谢。

本书的编辑和出版还得到了刘炯编辑的指导和帮助，在此对其热心、负责、干练的工作作风和辛勤的工作表示由衷的敬意和感谢。

进化算法和图学习都属于目前国际上科学研究的前沿领域，处于快速的发展之中，无论是在理论还是应用方面都存在着大量的问题尚待进一步深入研究。由于作者学识水平和可获得资料的限制，本书中不妥之处在所难免，敬请同行专家和诸位读者批评指正，也希望能够和各位同好在学术上进行多方面的合作。

著 者

目　录

第1章 绪 论

本书关注两方面的内容：图学习和进化算法。图是建模复杂结构和复杂交互的利器，然而由于图表示的灵活性，从图数据中进行学习或数据挖掘是困难的：一方面图学习的假设空间(子图结构集合)非常巨大，另一方面在图数据中学习时常常要面对图同构、子图同构等目前还没有多项式时间算法的问题。为此将擅长解决复杂问题的进化算法引入图学习，以期发现更优的解。本章 1.1 节介绍图学习的重要性和复杂性；1.2 节简要介绍图学习的典型应用领域；1.3 节详细阐述本书的研究目标、图学习中的核心任务——子结构发现的研究现状；1.4 节为全书的内容安排；1.5 节对本章进行小结。

1.1 图学习的目的

科学技术的飞速发展使当今社会呈现出信息化、网络化和全球化的特点，随之而来的海量数据对人类现有的数据处理能力提出了新的挑战，“数据爆炸”、“数据的丰富，知识的匮乏”正是对这一现象的真实写照。然而更为复杂的是，除了海量性之外，数据本身往往还呈现出结构化的特点，即数据内部或数据之间存在着形形色色、错综复杂、常常较数据本身更能反映问题实质的联系，这又进一步加大了数据的自动处理和分析的难度，同时也使得许多数据挖掘结果不能尽如人意。“数据的丰富和复杂，高质量知识的匮乏”呼唤更为强有力的数据挖掘工具。近年来，从结构数据中进行挖掘和学习引起了众多研究者的关注，并逐渐成为数据挖掘和机器学习领域的一个主流分支。在第五届图挖掘和图学习国际研讨会(MLG'07)的主页上这样写到：“数据挖掘和机器学习正在经历一场结构化革命。数十年来人们一直关注独立同分布数据，然而现在许多研究者开始或正在研究建立在更为复杂的数据表示形式上的问题……”[1]

结构数据既可以表现为像在蛋白结构预测和自然语言分析中所呈现出的序

列或树等简单的结构形式，也可以表现为在论文引用图和 World Wide Web 中所呈现出的复杂的图结构形式。对于现实生活中的任一事物，从内外两个方面来看，对内它可以被描述为一个反映其内部各组成成分之间交互关系的网络；对外它可以被描述为更大网络中的一个组成成分，因此从本质上讲结构数据可以归纳为以下两种类型：

(1) 数据之间是相互独立的，但数据本身由一系列相互关联的实体组成(内部结构)，如分子；

(2) 多个数据之间互相联系形成了更为复杂的结构(外部结构)，如 World Wide Web 中的网页和社会网络中的个体。

图是一种通用的结构表示形式，它也是数学、计算机科学、工程学科等许多自然科学学科中最为常用的数据表示工具之一，它以结点代表问题领域中的对象或对象的属性，用边代表对象之间或对象与属性之间的关系，很好地建模了以上两种结构类型。

从算法的角度讲，图是最一般的数据结构，其他常用的线性或非线性数据结构都是图的特例。例如，对于一个标量，可以用仅含一个结点的图来表示，结点的标签(附于图中结点或边的字符、字符串或数值，以反映结点类型或结点之间的关系类型的不同)即为该标量值；对于一个 n 维向量，可以用一个含有 n 个结点，n-1 条边的图来表示，每个结点代表一个分量，结点标签为各个分量的取值，而边则反映了分量之间的相邻关系，任意两个相邻的分量之间均连接一条边；对于一个字符串，可以用结点表示字符，用边表示字符之间的相邻关系；而对于常见的非线性结构——树来说，它本身就是图的特殊情形：树是不含回路的图。

既然图是最通用的数据表示形式，几乎可以表示一切数据，那么如何从大量甚至海量的图数据中发现有用的知识呢？很自然地，希望如图 1-1 上部所示的模式来解决这个问题，即输入图数据，计算机进行处理，然后得到有用的知识。事实上，当欲解决问题的处理过程很清楚时，或者说该问题已良好建模时，可通过直接编程来解决。例如，解线性方程组、求一个图中任意两结点之间的距离、求一个图的最小生成树等都属于这样的问题。但是对于未良好建模的问题，例如，识别手写字符、按是否会拖欠贷款或是否会购买本公司的产品进行客户分类、在图数据集中识别典型子结构、图的分类聚类等，对于这些问题人们事先并不知道如何由给定的输入计算出期望的输出，或者说即使知道，由于其计算代价太高，也不便实行。虽然这些问题不能用传统的编程途径来解决，但可以采用另外一种策略，即利用计算机强大的计算能力，从样例(即已知的相

关数据)中归纳出输入与输出之间的对应关系，这种解决问题的方法就称为学习。由于在解决这类问题时不可避免地要用到目前最具智能的机器——计算机，因此这门学科又称为机器学习。

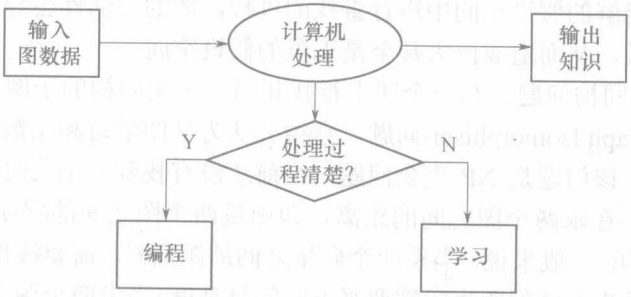


图 1-1　计算机解决问题的模型

学习属于数据驱动(Data Driven)的方法，它完全或主要靠从数据中归纳出特定的模式以解决问题。它与理论驱动(Theory Driven)的方法，如编程，形成鲜明的对比。理论驱动的方法能够精确定义出所需计算机处理的问题的过程，以演绎的方式解决问题。近年来，随着对数据挖掘和机器学习领域的深入研究，学习能够解决的问题的数量和范围呈快速增长和扩大之势，在许多领域，如字符识别、语音识别、人脸识别、文本分类、邮件过滤、基因检测、蛋白质同构体检测、Web 检索、拖欠贷款预测、分子性质确定等，都有着广泛的应用并带来了巨大的科学、经济和社会效益。

本书所关心的问题是从图数据中发现有效的、新颖的、潜在有用的和最终可理解的模式，该过程即称为图学习或图数据挖掘。需要说明的是，虽然相对于机器学习，数据挖掘更强调数据的海量性和算法的可伸缩性，但本书的研究内容则更多地偏向于图学习，对海量性和算法的可伸缩性的研究是今后的研究目标，因此在本书中对图学习或图数据挖掘不作区分，视为同一概念。图学习涉及生物学、化学、药学、机器人学、计算机科学、数学等众多的学科，并已在生物工程、交通运输、通信、计算机辅助设计、社会网络分析、国家安全等领域取得了许多令人振奋的成果。

然而，正是图表示的通用性和灵活性带来了其复杂性。实践和理论表明从图数据中学习是困难的。造成这一困难的原因主要有以下两个方面：

(1) 图表示的灵活性带来的巨大的假设空间。图是一种非常灵活的数据表示形式，结点可以表示任意类型的对象或对象的属性，边可以表示任意类型的对象之间或对象与属性之间的关系，任两个结点之间可以有任意多个任意类型

的关系。这种表示的灵活性使得图具有强大的表达能力，可以对非常宽泛的领域中的许多复杂问题进行很好的建模。但这种灵活性同时也不可避免地带来了复杂性。“Graphs are prisoners of their own flexibility”[2]。由于学习过程就是在包含所有可能解的假设空间中进行查找的过程，图的灵活性会导致候选解的数量成指数增长，从而造成巨大甚至是无限的假设空间。

(2) 子图同构问题。在一个图中查找和另一个图同构的子图的问题称为子图同构(Subgraph Isomorphism)问题。它被公认为是图学习和图数据挖掘领域中的瓶颈问题。该问题是 NP 完全问题，目前还没有找到多项式时间复杂度的算法来解决它。在求两个图之间的距离，即衡量两个图之间的(不)相似性时，常用到子图同构。一般来说，当求两个向量之间的距离时，需要线性时间复杂度，当求两个字符串之间的距离时需要平方时间复杂度，当求两个图之间的距离时，则需要指数时间复杂度[3]。

为了克服以上问题，许多图学习或图数据挖掘算法如 Subdue[4-7]、GBI[8]等算法采取了柱状查找、爬山算法等贪婪式局部查找方法，这类方法在减少查找复杂度的同时也带来了另一个问题，即该类方法容易陷入局部极值，难以得到全局最优或近似全局最优解。

进化算法(Evolutionary Algorithms，EA)提供了另一种解决问题的途径。实践证明进化算法在解决较难的最优化问题时有着良好的效果。具体来说，当一个问题包含较多的变量时，当变量之间存在复杂的关联关系时，当假设空间中存在较多的局部极值点时，用进化算法是一个非常好的选择。

另外，进化算法的内在并行特点和在进化算法中大量使用的全局搜索算子使其能够成功地跳出局部极值点，取得全局最优或近似全局最优解。然而进化算法在长于全局寻优的同时却拙于细粒度的局部查找，因此将进化算法和擅长局部寻优的爬山算法相结合，取长补短，相辅相成，应用于图学习，取得了较为满意的效果。

本书主要阐述混合进化算法在图学习与图数据挖掘中的核心问题——子结构发现中的应用。

1.2 图学习的应用领域

随着计算机软硬件技术的迅猛发展，人类能够处理规模越来越大和结构越来越复杂的数据，这种计算能力的提高为图学习和图数据挖掘提供了平台基础。另外，“需求是发明之母!”近年来，许多新兴学科的崛起为图学习和图数据挖掘提

供了应用基础。这些新兴学科有：化学信息学(Chemoinformatics)、生物信息学(Bioinformatics)、社会网络分析(Social Network Analysis)、国际互联网(Internet)等。接下来，对图学习和图数据挖掘在这些领域中的应用作一简要介绍。

1. 化学信息学

一直以来，化学家们就用图来表示分子结构，用结点表示不同的化学元素，用边表示一价、二价或三价化合键。化学信息学旨在从分子图结构中预测该化合物是否有毒、作为药品是否有疗效等性质。传统的做法是制药公司为了确定药品的医学性质，不得不对众多的候选药品作昂贵且耗时的科学实验和临床试验。有实例表明，应用图学习和图数据挖掘技术预测分子性质，可以为制药公司节省数百万美元和大量的时间[9]。

2. 生物信息学

分子生物学中包含着大量的结构数据，如 DNA、RNA、蛋白质的分子结构均可用图来表示；分子生物学中还包含着许多形形色色的网络，如蛋白交互网络(Protein-protein Interaction Networks)、新陈代谢网络(Metabolic Networks)、调节网络(Regulatory Networks)、演化发展网络(Phylogenetic Networks)，而网络最自然的表示就是图。生物信息学旨在对以上各种结构或网络的功能进行学习和预测。

目前，最成功的化合物分子功能预测方法即基于对图数据的相似查找。例如，如果要预测一个新的蛋白结构的功能，可以在一个已标注功能的蛋白数据库中进行相似查找，如果该蛋白结构和某个或某些蛋白质的结构很相似，就可以预测这个新的蛋白质拥有和这个或这些蛋白质相同的功能。这种方法的依据来源于一种生物进化的观点，即如果两种蛋白质拥有非常相似的拓扑结构，那么它们很有可能拥有公共的祖先，因而极有可能具有相同的生化功能。这种观点也是“结构决定行为”的体现，它从应用的侧面再一次说明了图学习和图数据挖掘的重要性。

3. 社会网络分析

社会网络具有极其广泛的范畴，电力网络、电话网络、交通运输网络、客户关系网络、产品营销网络、流行病传播网络、计算机病毒传播网络、食物链、同学关系网、朋友关系网等都属于社会网络。可以这么说，世界就是网络的集合。社会网络可定义为基于图表示的异构、多关系数据集合[10]。这种图通常都具有非常大的规模，它用结点来表示网络中的对象，用边(链，link)来表示对象之间的关系或交互，对象和边还可以拥有各自的属性。

顾名思义，社会网络分析即对社会网络的研究，它是一门涉及多个学科的

交叉学科。对它的研究既有科学和社会意义，也有重大的经济意义。一方面，心理学、社会学和人类学学家通过社会网络分析想要揭示复杂的人类社会动力学行为，动物学家通过社会网络分析想要发现动物之间或动物种群之间的交互规则；另一方面，企业和商家想要通过社会网络分析发现潜在的客户和关键客户(key-player，即该客户具有较大的影响力，其行为会影响其他众多客户的购买习惯和行为)，以实现成本更低、收益更大的目标销售。

另外，随着电子通信技术和计算机技术的迅猛发展，电信网络，如手机通信网络和 Internet 的通信日志中含有大量的社会网络数据，这为进行社会网络分析提供了廉价的和易获得的数据来源。

4. Internet、HTML、XML

以超文本(HTML)为结点，超链接(HYPERLINK)为边，Internet 本身就是一个图。事实上，Google 公司在其著名的 PageRank 算法中就是利用这样的图结构来对网站进行排序的。另外，作为数据库领域和企业界的标准，XML 常用来表示半结构数据，其主要结构是树状结构，是一种简单的图。在这样的图数据上，可以执行从简单的网页查询到复杂的网页分类、聚类等一系列图学习和图数据挖掘操作。

1.3 子结构发现的研究现状

图学习和图数据挖掘是从图数据中发现有效的、新颖的、潜在有用的和最终可理解的子图模式的过程。其分支领域有子结构发现、图分类、图聚类、图语法学习等内容。其中子结构发现是图数据挖掘中最基本、最核心的任务，也是图分类、图聚类、图语法学习等其他图学习和图数据挖掘任务的基础。以下对子结构发现的发生、发展历程作一简要介绍。

1.3.1 子结构发现所属的研究领域

在数据挖掘与机器学习领域，对结构数据的研究最早源于 20 世纪 90 年代初诞生的一门新兴的学科——归纳逻辑程序设计(Inductive Logic Programming，ILP)[11-14]。归纳逻辑程序设计是归纳学习(Inductive Learning，IL)和逻辑程序设计(Logic Programming，LP)的交叉学科，因此有这样的公式：ILP=IL∩LP。从归纳学习中 ILP 继承了其目标与方法：即开发新的技术和工具并用归纳学习的方法从示例和经验(背景知识)中挖掘出新的规则和模式；从逻辑程序设计中 ILP 则是借鉴采用了其基于一阶逻辑(First Order Logic)的表示方式。一阶逻辑在为

ILP 提供大量成熟的概念、理论和技术的同时，还为其提供了一致的和极富表达力的表示手段，其一致性体现在示例、背景知识和挖掘到的规则、模式都可以表示为谓词公式，从而可以很自然地将背景知识及中间挖掘结果(也是知识)直接应用于学习和挖掘过程，而知识的应用正是人工智能领域所公认的体现智能特点的关键所在[3]，这使其可以学习和挖掘出更为紧凑、易于理解、更具结构化和智能性的规则与模式。另外，用谓词逻辑表示的规则比命题规则具有更强的表达能力，其内涵更为丰富且易于理解。因此，ILP 学科的诞生克服了当时占主流的基于命题逻辑表示的挖掘方法的两个局限性：表达能力有限与不便于利用背景知识。

关系作为结构的代名词一直以来就是人们关注与研究的焦点，而一种从关系数据库中学习和挖掘规则与模式的想法则使得对结构数据的学习和挖掘得到了更多研究者的认可和关注。关系数据库自 20 世纪 60 年代末诞生以来一直就是数据管理的主要工具与技术。一个关系数据库由多个表构成，表和表之间，表内元组与元组之间，存在着许多一对多和多对多的关系，而关系比数据本身蕴含着更为丰富的内含。再加上随着信息社会的到来和经济的全球化，大量的科学数据和商业数据储藏在众多的关系数据库系统中，这就对 20 世纪末、21 世纪初仍占主流的基于单表的数据挖掘方法从理论和实用的角度提出了挑战。正是在此背景下，一门新的学科——多关系数据挖掘(Multi-Relational Data Mining，MRDM)应运而生，并于 2001 年召开了第一届多关系数据挖掘国际研讨会，旨在从关系数据库中挖掘新的、感兴趣的、潜在有用的模式、规则与知识[15,16]。

鉴于子句逻辑与关系代数的内在关联性，ILP 与 MRDM 有着千丝万缕的联系，它们一起构成了结构数据挖掘的两个主流分支。由于图也是对数据间关系的一种直观自然的表示，其简明的特点使之比基于逻辑的 ILP 有着更为广泛的应用范围，从而使从图数据中学习和挖掘有效的、新颖的、潜在有用的和可理解的图模式的图学习(Graph-based Learning，GL)和图数据挖掘(Graph-based Data Mining，GDM)成为了结构数据挖掘的另一主流分支。

ILP 与 GDM 有许多共性，也有许多不同。其共性体现在：首先两者的目标相同，都是要挖掘出蕴藏在数据中的形形式式的关系，即进行结构数据挖掘；其次两者都具有坚实的数学理论基础，前者基于一阶谓词逻辑，后者基于图论，两者的蓬勃发展都得益于底层的数学基础所提供的概念、理论、方法和成熟的工具；第三，两者的建模能力极强，均属于普适性工具，能够应用于非常广泛的领域。两者的不同之处在于：首先是表示方法不同，前者是基于逻辑的，后