# 进化图数据挖掘算法及其在信息学科建设中的应用

## Evolutionary Graphical Data Mining Algorithm and its Application in the Academic Discipline Building of Information and Computing Science

常新功 （山西财经大学信息管理学院 太原 030006）

**摘要**：为了克服当前主流的图数据挖掘算法常采用的贪婪式查找带来的易于陷入局部最优这一问题，提出了一种基于进化算法的图数据挖掘算法，以增强算法的全局查找能力。定义了基于图的交叉和变异算子。考虑到进化算法局部搜索能力弱的特点，本文在变异算子的设计中还融入了爬山算法的思想，以进一步提高解的质量。应用该算法于信息与计算科学的学科建设之中。首先，将网上收集的有关各个院校信息学科专业建设的信息转化为图表示方式，然后用该算法从这些图数据中挖掘出信科专业建设的典型模式。该模式为我校的信息科学专业建设和教学改革提供了理论依据，具有重要的参考价值。

**关键词**：进化算法 图数据挖掘 最小描述长度 教育 学科建设

信息与计算科学专业是教育部于1998年新颁布的数学类专业。该专业涵盖数学、计算科学、计算机科学、信息科学以及控制科学等多个学科，且处在快速的发展之中[1,2]。兴办时间短、涵盖面广和快速发展是该专业的三个特点，也是造成该专业建设困难的原因所在。本文作者作为山西财经大学信息与计算科学专业的教研室主任参加了2004年山西省教委对我校信科专业的评估工作，在评估过程中感觉到广泛地借鉴和吸取其它院校的办学经验是办好信科专业的关键所在，同时也感觉到充分利用先进的数据挖掘技术于专业建设之中可以起到事半功倍的作用。

图数据挖掘是从以图表示的数据中自动提取新的、有用的模式的过程。Subdue[3,4]是当前主流的图数据挖掘算法之一，它以一个大的单个图或由多个图组成的图数据集为输入，以MDL为模式度量，以基于贪心算法的柱状查找为查找方式，输出压缩率最高的子图结构。本文在其基础上提出了基于进化算法（EA，evolutionary algorithms）的图数据挖掘算法，较好地克服了柱状查找易于陷入局部极值的问题，并将其应用于我校的信息学科建设之中，收到了良好的效果。

## 1 图数据挖掘的基本概念

图具有非常强的建模能力，它以结点表示对象，以边表示对象间的关系。结点和边带有标签（label）以作区分。图1是表示三个分子结构的一个样本图形数据集。其中，椭圆表示结点，数字为结点的编号，字符C，S，N，O为结点的标签，bond，dblbond为边的标签。

### 1.1 子结构及其实例

重复出现的数据片段容易引起人们的关注，而这些数据片段的公共结构正是我们要寻找的模式。一个模式如能在越多的场合被用来解释数据，这个模式就越有用。图数据中所有彼此同构的连通子图的结构称为一个子结构（substructure）。而这些连通子图称为该子结构的实例（instances）。表1中给出了图1中的五个子结构及其实例。其中g（n1，n2，n3，…）表示图g中由结点n1，n2，n3，…构成的一个子图。子结构和实例的关系就像是类和对象、型和值之间的关系。子结构即模式，寻找有意义的子结构并应用于实践正是本文的目的所在。

### 1.2 子结构的评价与MDL

本文和Subdue均以MDL（minimum description length，最小描述长度）[5]作为一个子结构的评判准则。MDL基于如下思想：数据中的任何规则性的东西可以被用来压缩该数据。其中，规则性的东西可被视为一种子概念，在用了该子概念之后，对数据的描述所需的符号就比不用该子概念对数据的描述所需的符号

少。子结构就是一种子概念，其实例越多，规则性越强。设 G 代表一个图，S 是它的一个子结构，在用了子结构 S 之后 G 的描述长度定义为 DL(S) +DL(G|S)，其中 DL(S)是子结构 S 的描述长度，即编码 S 所需的二进制位数。DL(S)的求法参见[4]。

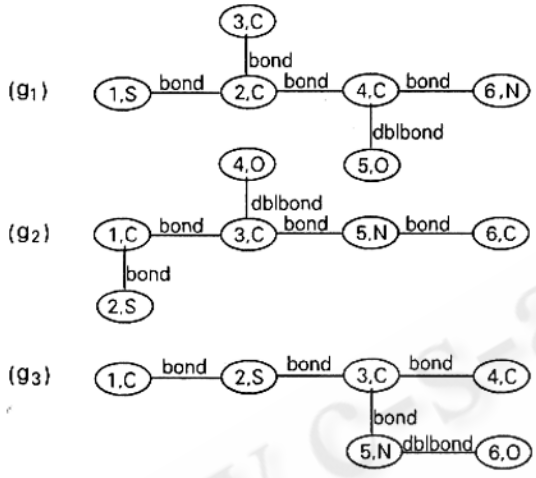


图 1　一个样本图数据集

表 1　图 1 中的几个子结构及其实例

| 编号 | 子结构 | 实例 |
|---|---|---|
| $S_1$ | O | $g_1(5)$, $g_2(4)$, $g_3(6)$ |
| $S_2$ | C dblbonc O | $g_1(4,5)$, $g_2(3,4)$ |
| $S_3$ | N O | $g_3(5,6)$ |
| $S_4$ | C bond C bond N | $g_1(2,4,6)$, $g_2(1,3,5)$, $g_3(4,3,5)$ |
| $S_5$ | O dblbond C bond C bond N | $g_1(2,4,5,6)$, $g_2(1,3,4,5)$ |

DL(G|S)表示在将 G 中 S 的所有实例分别换为一个代表该子结构的新的结点后，对所得图进行编码所需的二进制位数。子结构 S 对 G 的压缩率定义为 DL(G)/(DL(S) +DL(G|S))，其中 DL(G)为图 G 未被压缩时的描述长度。这样压缩率可以作为一个子结构 S 的度量，用了子结构 S 之后 G 的描述长度越小，S 对 G 的压缩率越大，子结构 S 的规则性就越强。实践和理论证明，基于 MDL 准则的方法能够找到更为关键、更有意义的子结构[4]。本文就是要在众多的代表不同院校办学信息的图数据中找出描述长度最小，即压缩率最大的子结构作为办学的典型模式以作参考。

## 2　基于 EA 的图数据挖掘算法

EA 是一种基于自然选择和遗传变异等生物进化机制的全局性随机搜索算法[6]。它从代表侯选解集的初始种群开始，迭代地进行选择、交叉、变异等遗传操作，最终发现最优解或近似最优解。进化算法除具有普适性、可伸缩性、并行性和鲁棒性等特点外，还具有群体搜索的特点，即它是从一群而不是单一的初始点开始沿多条路径搜索，这使其可以有效的跳出局部极值点。本文正是利用这一点在子结构发现算法中引入进化机制以提高解的质量。

### 2.1　染色体的表示

一个子结构就是一个染色体。它由定义该子结构的图(包括结点数组、边数组等内容)以及指向该子结构的实例的指针数组、目前找到的实例个数、该子结构的压缩率等项组成。

### 2.2　种群的初始化

首先生成所有的象表 1 中 $S_1$ 那样的单点结构，然后再对所有的单点结构进行扩展，扩展的办法是对每个单点结构的所有实例增加一条相邻的边，形成一个单边子图。将所有这些子图按是否同构分类就形成了所有的单边结构。例如，单点结构 $S_1$ 通过其实例 $g_1(5)$，$g_2(4)$，$g_3(6)$ 可分别扩展为 $g_1(5,4)$，$g_2(3,4)$，$g_3(5,6)$，其中前二个对应 $S_2$，最后一个对应 $S_3$。在对所有单点结构扩展后生成的子结构中随机选择 popsize(种群大小)个构成初始种群。

### 2.3　适应值、选择和精英保留

一个子结构 S 的适应值定义为压缩率 DL(G)/(DL(S) +DL(G|S))。选择采取联赛选择，它较之轮盘赌选择可以保持更高的种群多样性。本文还采用精英保留策略，当前代的最佳个体不经选择直接进入下一代，以避免由进化机制的随机性导致的已找到的最优解的丢失。

### 2.4　交叉

设 Sub1 和 Sub2 为两个子结构，其交叉过程为：首先在 Sub1 的所有实例中任选一个实例，记作 ins1，在 Sub2 的实例中找出所有与 ins1 相重叠(有公共结点)的所有实例并任选其一作为 ins2，如没有相重叠的返

回空值;其次将 ins1 和 ins2 合并成一个新的子图并转成一个子结构,记作 NewSub;然后开始收集 NewSub 的实例,将 Sub1 的任一实例和 Sub2 的任一实例合并成一个新的子图,在得到的这些子图中所有与 NewSub 同构的子图即为其实例;最后返回 NewSub。例如,表1中 $S_2$ 和 $S_4$ 交叉,在 $S_2$ 的实例中任选一个实例,假设为 $g_1(4,5)$,在 $S_4$ 中和 $g_1(4,5)$ 重叠的实例为 $g_1(2,4,6)$;$g_1(4,5)$ 和 $g_1(2,4,6)$ 合并为 $g_1(2,4,5,6)$,转成子结构 $S_5$;将 $S_2$ 的任一实例和 $S_4$ 的任一实例合并,所得结果中和 $S_5$ 同构的有 $g_1(2,4,5,6)$, $g_2(1,3,4,5)$,它们即为 $S_5$ 的实例,$S_5$ 作为结果返回。

### 2.5 变异

设 Sub 为一个子结构,首先对 Sub 的所有实例以所有可能的加一边的方式进行扩展;然后对所得的子图按同构关系进行归类,不同的类对应不同的新的子结构;最后在这些新的子结构中任选 n 个,其中适应值最大的,作为结果返回。例如对表1中 $S_4$ 变异,扩展其实例 $g_1(2,4,6)$ 可得 $g_1(2,4,5,6)$, $g_1(1,2,4,6)$, $g_1(2,3,4,6)$,扩展 $g_2(1,3,5)$ 可得 $g_2(1,2,3,5)$, $g_2(1,3,4,5)$, $g_2(1,3,5,6)$,扩展 $g_3(4,3,5)$ 可得 $g_3(2,3,4,5)$, $g_3(3,4,5,6)$。将这些实例按同构关系可分为 5 组,分别对应 CCNO, SCCN, CCCN, CCNC, SCNC五个子结构。设 n = 2,任选的两个子结构为 CCNO,CCCN,其中适应值较大的 CCNO 即 $S_5$ 作为结果返回。

上述的 n 是一个正整数,当 n =1 时变异操作是纯随机性的变异,n 越大贪婪性越强。这种机制将爬山算法的思想引入到了 EA 中,以缓解 EA 局部搜索能力不强这一问题。它对算法的运行效率和运行结果都有较大的促进和提高。

### 2.6 基于 EA 的子结构发现算法

以下为基于 EA 的子结构发现算法,其中 popsize 为种群规模,limit 为最大迭代次数,pc 为交叉概率,pm 为变异概率。函数 flip(x) 先随机产生一个介于 0 和 1 之间的随机数,然后拿这个数和 x 比较,如该数小于 x 则返回真,否则返回假。

```
输入 popsize, limit, pc, pm,n
初始化种群并计算每个染色体的适应值
generation = 0
while (generation < limit)
    i = 0
    将上一代种群中的最优个体保存为当前种群中的第 0 个个体
    i++
    do {
      在上一代种群中随机选择两个个体 chrom1 和 chrom2
      if (flip(pc))
        chrom1 = chrom1 和 chrom2 的交叉结果
      if (flip(pm))
        chrom1 = chrom1 变异后所得的结果
      将 chrom1 保存为当前种群中的第 i 个个体
    } while (i < popsize)
    返回最优个体
```

## 3 挖掘典型的专业办学模式

### 3.1 数据的收集和表示

本文从各个院校的主页收集该校信科专业培养目标、培养要求、主干课程、实践教学、毕业去向等信息,并将其转为图表示,其中一个院校是一个星状结构。目前我们共收集了 126 所院校的信科专业建设的数据,示例见图2。

### 3.2 调整子结构评价方法以偏置查找

我校是财经类院校,因而更偏重于参考和采纳财经类院校的办学经验,但是其它院校也有许多可借鉴之处。为此,我们采用加权的办法,财经类院校权重为 1.5,重点院校权重为 1.2,其它院校权重为 1,以此来区分不同院校所具有的参考价值,从而偏置算法的查找,挖掘出更有意义的模式。这是对原算法的一个改进,在原算法中一个子结构 S 的压缩率为 DL(G)/(DL(S)+DL(G|S)),本文将其乘了一个因子 α,改为 αDL(G)/(DL(S)+DL(G|S))。其中 α 按如下算法计算:

```
α = 0
for each instance ins of S
  α = α + w(ins)
α = α/N
```

其中 w(ins) 为子结构 S 的实例 ins 所属院校的权重,N 为子结构 S 的实例个数。这样当 S 的实例多数属于财经类或重点院校时,其 α 值就大,从而压缩率就大,表示该子结构较优。另外,对于不同的权重设置,

利用以上机制还可以挖掘出不同目的,不同要求的参考模式。

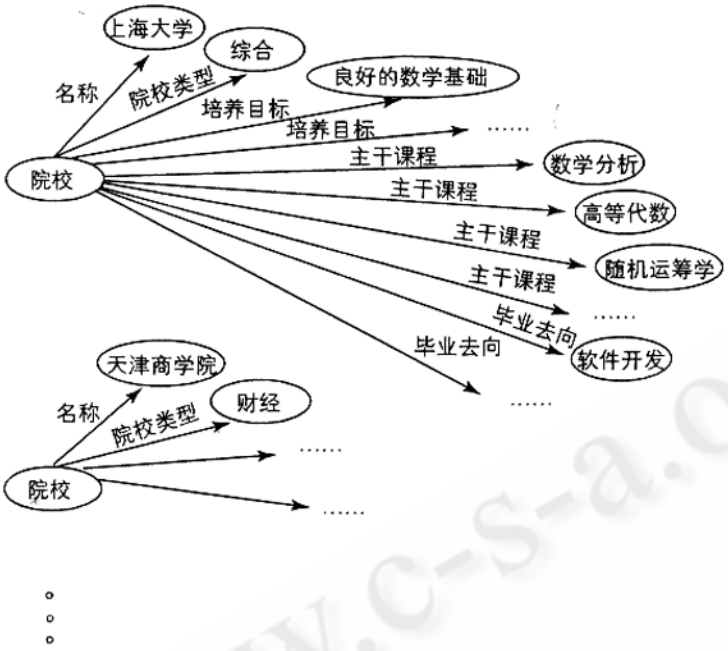


图 2　数据集示例

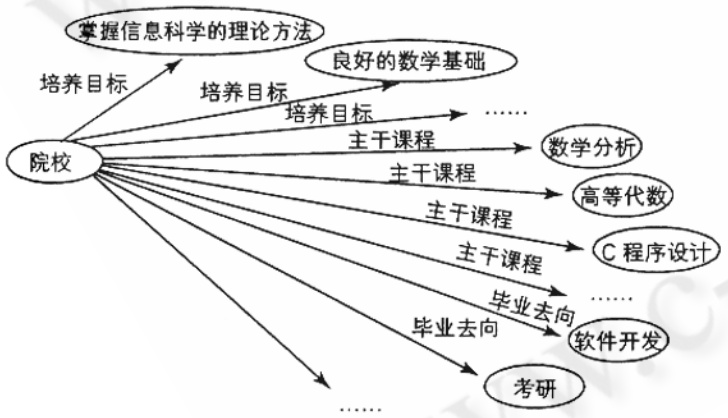


图 3　挖掘结果示例

### 3.3 挖掘的结果及分析应用

本文设置 popsize = 20, limit = 100, pc = 0.2, pm = 0.6, n = 4, 图 3 为挖掘出的结果的一个片段。由图中可以看出,院校名称、院校类型等非规则性的信息并没有出现在结果中,结果中关于培养目标、主干课程、毕业去向等信息反映了国内大多数院校建设信科专业的总的趋势和总的特征,具有非常重要的参考价值。例如,我校的信科专业以前开设线性代数,按照挖掘出的参考模型,我们将其调整为高等代数。看到许多学生毕业后从事软件开发,我们将数据结构与算法这门课程调整为数据结构和算法分析两门课程,为学生今后的开发工作打下坚实的基础。从参考模型中我们看到,有较大比例的本专业学生毕业后选择了考研,为此我们调整课时安排,在第 7 学期为学生空出更多的时间复习,以备考研。

……

## 4 结束语

本文提出了一种基于 EA 的图数据挖掘算法并将其应用到信息学科建设之中。首先将各个院校信息科学专业的办学信息组织成图表示形式,然后用本算法挖掘出兴办信息科学专业的典型模式。这个典型模式具有非常重要的参考价值,它为我校的信息科学专业建设起了积极的作用。应用更多的数据挖掘算法于教育领域是本文作者今后追求的目标及努力的方向。

## 参考文献

1 教育部数学与统计学教学指导委员会数学类教学指导分委员会. 关于《信息与计算科学》专业办学现状与专业建设相关问题的调查报告[J], 大学数学,2003,19(3):1-5.

2 龚日朝,"以特色取胜"建设信息与计算科学专业的新型思路与实践[J],大学数学,2004,20(3):12-15.

3 D. J. Cook and L. B. Holder. Graph-based data mining[J], IEEE Intelligent Systems 15(2):32-41. 2000.

4 I. Jonyer, L. B. Holder, and D. J. Cook. Graph-based hierarchical conceptual clustering[C]. International Journal on Artificial Intelligence Tools, 2001. 10(1-2):107-135.

5 P. Grunwald. A tutorial introduction to the Minimum Description Length Principle[Z]. http://homepages.cwi.nl/pdg/ftp/mdlintro.pdf.

6 李敏强、寇纪淞、林丹、李书全著, 遗传算法的基本理论与应用[M], 北京: 科学出版社. 2002.3.