算法设计与分析期末作业

选题 A：利用所学知识研究最优属性约简问题

标记数据可以用一个信息表描述，即$S = (U, C, D)$。在一个信息表中，寻找该信息表的一个最优属性约简问题可以形式化定义如下：

输入：一个信息表$S = (U, C, D)$

输出：一个最优属性约简$B \subseteq C$

约束条件：$POS_B(D) = POS_C(D)$

最优化目标： $\min |B|$

表 1 一个简单医疗信息表

	a_1	a_2	a_3	a_4	a_5	a_6	d
Patient	Headache	Temperature	Lymphocyte	Leukocyte	Eosinophil	Heartbeat	Flu
x_1	Yes	High	High	High	High	Normal	Yes
x_2	Yes	High	Normal	High	High	Abnormal	Yes
x_3	Yes	High	High	High	Normal	Abnormal	Yes
x_4	No	High	Normal	Normal	Normal	Normal	No
x_5	Yes	Normal	Normal	Low	High	Abnormal	No
x_6	Yes	Normal	Low	High	Normal	Abnormal	No
x_7	Yes	Low	Low	High	Normal	Normal	Yes

例如表 1 为一个简单的医疗信息表，其中$U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$表明有 7 个样本，$C = \{a_1, a_2, a_3, a_4, a_5, a_6\}$表明该信息表有 6 个属性，$D = \{d\}$表示该信息表具有一个标记信息同时该标记存在两种状态，即 Yes 或 No。

函数$POS_B(D)$表示在属性集B下能够确定标记的样本集，其形式化定义如下：

$$POS_B(D) = \{x \in U \mid |[x]_B/d| = 1\},$$

其中$[x]_B$表示在属性集B下元素$x \in U$的等价类，$[x]_B/d$表示$[x]_B$关于标记属性d的等价划分（或理解为关于标记属性分类的类别数）。例如，对于属性集合$B = \{a_1, a_2\}$，函数$POS_B(D)$的计算过程如下：$[x_1]_B = [x_2]_B = [x_3]_B = \{x_1, x_2, x_3\}$、$[x_4]_B = \{x_4\}$、$[x_5]_B = [x_6]_B = \{x_5, x_6\}$和$[x_7]_B = \{x_7\}$。由于$|[x_1]_B/d| = \left|\{\{x_1, x_2, x_3\}\}\right| = 1$，则$x_1, x_2, x_3 \in POS_B(D)$；由于$|[x_4]_B/d| = \left|\{\{x_4\}\}\right| = 1$，则$x_4 \in POS_B(D)$；由于$|[x_5]_B/d| = \left|\{\{x_5, x_6\}\}\right| = 1$，则$x_5, x_6 \in POS_B(D)$；由于

$|[x_7]_B/d| = \left|\{\{x_7\}\}\right| = 1$，则$x_7 \in POS_B(D)$；因此，对于属性集合$B = \{a_1, a_2\}$有$POS_{\{a_1,a_2\}}(D) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$。

研究内容如下：

1. 设计高效的计算函数$POS_B(D)$算法，并分析其最坏时间复杂度。输入测试数据以及任意一个属性子集对应下标，输出$POS_B(D)$包含样本对应下标。例如，输入表 1 中数据表（用矩阵表示）和属性子集下标：1 2，则输出$POS_B(D)$包含样本对应下标为： 1 2 3 4 5 6 7
2. 利用所学知识设计至少两种方法求解最优属性约简，并理论分析其最坏时间复杂度以及通过实验分析所提出算法的效率。输入测试数据，输出最优属性约简的下标。
3. 撰写实验报告，不少于 15 页。至少分为以下五个大部分（子节标题自己命名）：1. 计算函数$POS_B(D)$的算法及分析、2. 基于 XXX 的最优属性约简算法及分析、3. 基于 XXX 的最优属性约简算法及分析、4. 仿真实验分析和 5. 代码。字体格式：字号五号、中文字体宋体、英文字体 Times New Roman 和行距固定 20 磅。

选题 B：利用所学知识研究超图的最小顶点覆盖问题

超图理论由法国数学家 C. Berge 创立，他系统的建立了超图理论。在过去的几十年中，超图理论应用于很多实际问题中，例如特征追踪、场景配准、图像聚类、关联规则挖掘、图像分割和特征选择等。超图是图论中简单图的泛化形式，二者最大的差别在于简单图中每一条边最多可以连接两个顶点，而在超图中边可以连接两个以上的顶点。在超图理论中，将这种能够连接两个以上顶点的边称为超边。现实生活中存在很多中顶点规模较多和连接形式丰富的超图结构。下面介绍一个超图的例子。

例 1 对于一次大型的人工智能领域的国际学术会议，其中包含$k \ge 1$个场报告：$e_1, e_2, \ldots, e_k$，令V为到现场参加该次会议的学者构成的集合。假设每一场报告至少有一个学者参加，则可以构建一个如下超图：

1) 每一个到现场参加会议的学者被看作是一个顶点，既V为顶点集；

2) 每一场报告被看作是一个超边，其中超边e_i连接的顶点含义为所有参加报告e_i的学者构成集合。

例 1 中构建了一个学者参加学术会议报告的超图结构，其中一个学者可以参加多个不同的主题的报告，且某个主题的报告可以有多个人参加。该类型超图结构可以形式化定义为如下：

定义 1 设一个有限集合$V = \{v_1, v_2, \ldots, v_n\}$，集合$V$的子集簇$E = \{e_1, e_2, \ldots, e_k\}$，其中对于任意一个$e_i$满足$e_i \subseteq \mathrm{V}$，若子集簇$E$为集合$V$上的超图，则$E$满足下列两个条件：

1) $e_i \neq \emptyset$，其中$i = 1,2,\dots,k$；

2) $\cup_{i=1}^{k} e_i = V$。

若集合V和其子集簇E满足定义 1 中的两个条件，则可以用一个二元组$H = (V, E)$表示为一个超图，其中顶点集为$V = \{v_1, v_2, \dots, v_n\}$，超边集为$E = \{e_1, e_2, \dots, e_k\}$。

在一个超图$H = (V, E)$中，其中$V = \{v_1, v_2, \dots, v_n\}$和$E = \{e_1, e_2, \dots, e_k\}$，则有：

1) 在超图H中超边$e \subseteq E$连接的顶点集表示为$V(e)$，连接顶点$v \in V$的超边集表示为$E(v)$；

2) 若超某个超边$e \in E$满足$|e| = 1$，则表明构成了一个环；

3) 若两个顶点v_x和v_y满足$\exists e \in E \wedge \{v_x, v_y\} \subseteq V(e)$，则表明顶点$v_x$和$v_y$连接；

4) 若存在一个超边$\{v_i\} \in E$，则表明顶点v_i与它自身连接；

5) 超图H中若任意两个超边$e_x, e_y \in E$满足$V(e_x) \cap V(e_y) \neq \emptyset$，则表明超边$e_x$和$e_y$相交，否者这两个超边不相交。

6) 某个顶点$v \in V$的度表示连接该点的超边数目$d(v) = |E(v)|$，而构成环的顶点的度为 2；

与经典图论类似，在超图理论中有诱导子超图和部分超图的概念，定义如下。

定义 2 在一个超图$H = (V, E)$中，其中$E = \cup \{e_i | i \in I\}$，$I$为超边的标记集，有：

1) 对于非空顶点子集$V' \subseteq V$，超图H的一个诱导子超图为$H' = (V', E')$，其中$E' = \cup_{i \in I} \{e_i' | e_i' = V(e_i) \cap V' \wedge e_i' \neq \emptyset\}$；

2) 对于非空子集$J \subseteq I$产生超边子集$E' = \{e_j | j \in J\}$，超图H的一个部分超图为$H' = (V', E')$，其中$\cup_{j \in J} e_j = V'$。

图 1 是一个简单的超图，以该简单超图为例，简单分析了上述相关概念。

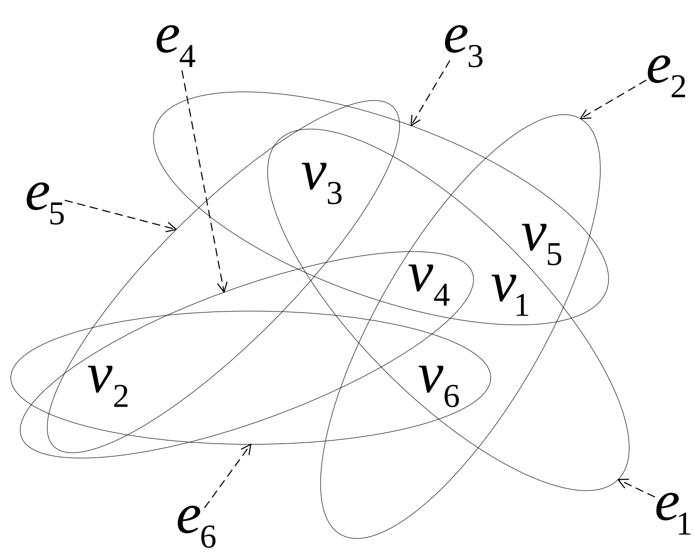

图 1 一个简单的超图

例 1 对于图 1 中超图，有：

1) 该超图可以表达为$H=(V,E)$，其中顶点集为：$V=\{v_1,v_2,v_3,v_5,v_5,v_6\}$和超边集合为：$E=\big\{\{v_1,v_3,v_4,v_6\},\{v_1,v_4,v_5,v_6\},\{v_1,v_3,v_4,v_5\},\{v_2,v_3\},\{v_2,v_4\},\{v_2,v_6\}\big\}$；

2) 顶点v_2连接的超边为$E(v_2)=\{e_4,e_5,e_6\}$，超边e_1连接的顶点集为$V(e_1)=\{v_1,v_3,v_4,v_6\}$；

3) 超图中一共有六个顶点，它们的度分别为$d(v_1)=3$、$d(v_2)=3$、$d(v_3)=3$、$d(v_4)=4$、$d(v_5)=2$和$d(v_6)=3$；

4) 由于对于任意一个$e\ \in E$均有$|V(e_1)|>1$，则该超图中不存在环；

5) 由于$\{v_1,v_6\}\subseteq V(e_1)=\{v_1,v_3,v_4,v_6\}$，则顶点$v_1$和$v_6$连接；

6) 对于超边e_1和e_2有$V(e_1)\cap V(e_2)=\{v_1,v_4,v_6\}$，因此超边$e_1$和$e_2$相交，另外对于超边$e_3$和$e_6$有$V(e_3)\cap V(e_6)=\emptyset$，则超边$e_3$和$e_6$相互独立不相交；

7) 对于顶点子集$V'=\{v_3,v_4,v_5,v_6\}$，则由其产生超图的诱导子图为$H'=(V',E')$，其中超边集为$e_1'=V(e_1)\cap V'=\{v_3,v_4,v_6\}$、$e_2'=V(e_2)\cap V'=\{v_4,v_5,v_6\}$、$e_3'=V(e_3)\cap V'=\{v_3,v_4,v_5\}$、$e_4'=V(e_4)\cap V'=\{v_3\}$、$e_5'=V(e_5)\cap V'=\{v_4\}$和$e_6'=V(e_6)\cap V'=\{v_6\}$；

8) 若超图H中的超边标记集为$I=\{1,2,3,4,5,6\}$，对于由标记子集$J=\{1,2,3\}$产生的部分超图为$H'=(V',E')$，其中超边集$E'=\big\{\{v_1,v_3,v_4,v_6\},\{v_1,v_4,v_5,v_6\},\{v_1,v_3,v_4,v_5\}\big\}$，顶点集$V'=V(e_1)\cup V(e_2)\cup V(e_3)=\{v_1,v_3,v_4,v_5,v_6\}$。

图论中一个经典的问题就是最小顶点覆盖问题，类似的超图中最小顶点覆盖问题可以定义如下：

定义 3 在一个超图$H=(V,E)$中，一个非空顶点子集$K\ \subseteq V$产生的诱导子超图为$H'=(V',E')$，若$|E'|=|E|$，则称顶点子集V'为超图 H 的一个顶点覆盖。

定义 4 在一个超图$H=(V,E)$中，最小顶点覆盖和极小顶点覆盖定义如下：

1) 一个在顶点覆盖K中，若删除任意一个顶点v而顶点子集$K-\{v\}$不是一个顶点覆盖，则称顶点子集K为一个极小顶点覆盖；

2) 顶点规模最小的极小顶点覆盖称为最小顶点覆盖；

超图理论中最小顶点覆盖问题实际上和经典图论中的相同，在现实生活中会遇到很多该类型问题，如例 1 中一次大型的人工智能领域的国际学术，由于时间的限制，不同主题的报告将按照并行的方式在不同会场同时进行。对于某个学者而言，他某个时间点只能依据自身兴趣爱好选择相应的主题参加报告，而对于大型的国际学术会议而言，该学者只能参加少量的主题报告。如果该学者需要了解本次国际学术会议每一个主题会议的报告情况来撰写一份参会总结，则他需要与参加其他主题报告的其他学者交流。由于参加会议人数众多，他如何只与少量的学者交流，以最小的代价获取本次会议所有主题的报告情况？

上述问题实际上就是超图中最小顶点覆盖问题，它可以形式化描述如下：

该次大型的人工智能领域的国际学术参会情况的结构定义为超图$H=(V,E)$，其中顶点集$V=\{v_1,v_2,\dots,v_n\}$表示为n个学者构成的集合，超边集$E=\{e_1,e_2,\dots,e_k\}$表示为有k个不同的主题报告。学者v_1参加的主题报告为$E_1\subset E$，如何从v_2至v_n中选择较少的学者交流而获得本次会议所有主题的报告情况。

下面将上述问题转化为超图中最小顶点覆盖问题：

超图$H=(V,E)$的部分超图为$H'=(V',E')$，其中$E'=\{e_1,e_2,\dots,e_k\}-E_1$和$V'=\cup_{e'\in E'}V(e')$。部分超图$H'=(V',E')$的一个最小顶点覆盖就是上述问题的解。

研究内容如下：

1. 利用所学知识设计至少两种方法求解超图的最小顶点覆盖，并理论分析其最坏时间复杂度以及通过实验分析所提出算法的效率。输入测试数据，输出最小顶点覆盖对应顶点集下标。

   例如：输入：1 3 4 6
   1 4 5 6
   1 3 4 5
   2 3
   2 4
   2 6

   输出：1 2

2. 撰写实验报告，不少于 15 页。至少分为以下五个大部分（子节标题自己命名）：1.超图最小顶覆盖问题的描述、2. 基于 XXX 的超图最小顶点覆盖算法及分析、3. 基于 XXX 的超图最小顶点覆盖算法及分析、4. 仿真实验分析和 5. 代码。字体格式：字号五号、中文字体宋体、英文字体 Times New Roman 和行距固定 20 磅。