问题背景:

解决分类问题是机器学习的其中重要一个分支。要实现分类首先需要获得相应数据进行学习，而现实生活中我们得到的数据通常是不平衡的（例如银行需要对于异常交易数据进行分类，而学习的数据中大部分是正常的交易数据，异常交易数据比较占少数），这就会使分类的结果产生影响。为此，本论文所提出的MTGP-SMOTE方法在数据层面运用合成少数过采样技术（SMOTE）方法对原始数据进行过采样（oversampling），而后借助遗传编程中的多树遗传编程（Multitree Genetic Programming）策略，将单个树的运行结果作为一个独立的填补数据，从而实现多树对总体数据的填补。

概念理解：  
Majority class (*Maj*): 原始数据样本中较多的一类数据（例如银行交易数据中的正常交易数据）

Minority class (*Min)*: 原始数据样本中较少的一类数据（例如银行交易数据中的异常交易数据）

SMOTE算法：合成少数过采样技术

算法流程：

1. 确定临近数量
2. 选择近邻样本
3. 合成新样本

遗传编程（GP）：一种模仿自然界中进化过程的一种算法，其中包含

1. 初始化（Initialization）

常用的方法有：Grow，Full，Ramped half-and-half（本论文在初始化时所用的方法）

1. 选择（Selection）

从当前一代中选择某些个体作为下一代的父母的过程，本论文用到的选择方法是精英策略（elite strategy）

1. 适应度（Fitness）

衡量遗传过程的好坏，对后续交叉、变异等操作起到决定性影响

1. 交叉（Crossover）

从两个随机选取的父代个体中产生两个子代个体进行互换

1. 变异（Mutation）

类似于产生新的基因，利于形成群体的多样性，使算法有机会从局部最优解中解脱出来。

1. 复制（Replication）

父代个体在未经任何变化的条件下从当前代复制到新一代（本论文中使用到的算子没有包括复制操作）

以上操作是否发生完全取决于该个体的适应度，即个体的适应度越大，被选中的概率越大。

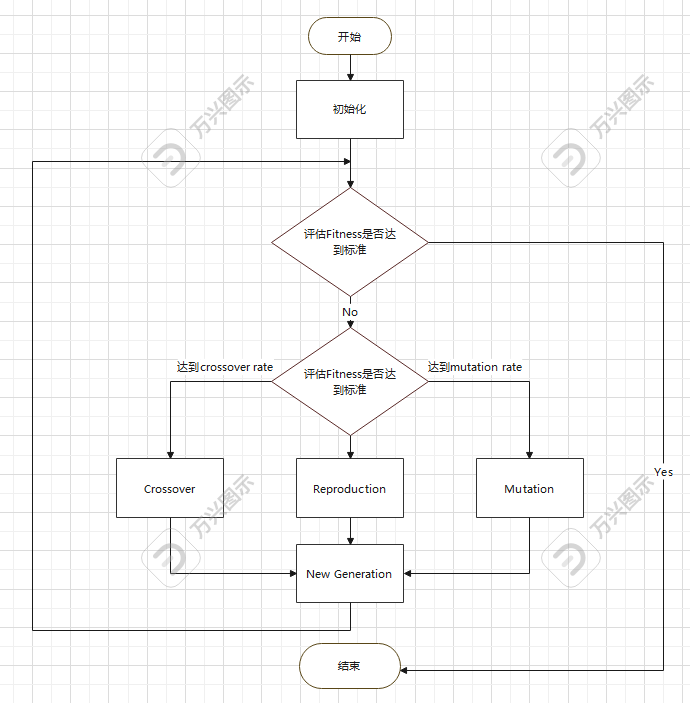
MTGP-SMOTE的具体流程

1. 确定需要填补的少数类的个体数量（即为树的数量）
2. 确定Terminal set和Function set，定义树中运算的基本结构
3. 获得目标评估对（Mait，Mint）用于指导个体的生成。不同的树要分配不同的（Mait，Mint）作为后续评估目标生成的标准
4. 适应度评估

利用距离测量和角度测量两种方法进行评估，评价个体的生成情况，决定后续是否进行交叉或变异操作

文本

AI 生成的内容可能不正确。



一些问题

1. 论文所用到了变异算子和交叉算子，为什么不使用复制算子（replication），复制算子的使用对整个系统会有什么影响？