机器学习算法系列（17）：非平衡数据处理

一、Introduction

常用的分类算法一般假设不同类的比例是均衡的，现实生活中经常遇到不平衡的数据集，比如广告点击预测（点击转化率一般都很小）、商品推荐（推荐的商品被购买的比例很低）、信用卡欺诈检测等等。

对于不平衡数据集，一般的分类算法都倾向于将样本划分到多数类，体现在模型整体的准确率很高。

但对于极不均衡的分类问题，比如仅有1%的人是坏人，99%的人是好人，最简单的分类模型就是将所有人都划分为好人，模型都能得到99%的准确率，显然这样的模型并没有提供任何的信息。

在类别不平衡的情况下，对模型使用F值或者AUC值是更好的选择。

处理不平衡数据，可以从两方面考虑：一是改变数据分布，从数据层面使得类别更为平衡；

二是改变分类算法，在传统分类算法的基础上对不同类别采取不同的加权方式，使得模型更看重少数类。

本部分对数据层面的一些方法做一个介绍，改变数据分布的方法主要是重采样：

- 1）过采样：增加少数类样本的数量
- 2）欠采样：减少多数类样本的数量
- 3）综合采样：将过采样和欠采样结合

二、过采样

2.1 随机过采样

采样算法通过某一种策略改变样本的类别分布，以达到将不平衡分布的样本转化为相对平衡分布的样本的目的，而随机采样是采样算法中最简单也最直观易懂的一种方法。

随机过抽样是增加少数类样本数量，可以事先设置多数类与少数类最终的数量比例，在保留多数

类样本不变的情况下，根据比例随机复制少数类样本，在使用的过程中为了保证所有的少数类样本信息都会被包含，可以先完全复制一份全量的少数类样本，再随机复制少数样本使得满足数量比例，具体步骤如下：

1.首先在少数类S_{min}集合中随机选中一些少数类样本
2.然后通过复制所选样本生成样本集合E
3.将它们添加到S_{min}中来扩大原始数据集从而得到新的少数类集合$S_{min-new}$

S_{min}中的总样本数增加了 |E|个新样本，且$S_{min-new}$ 的类分布均衡度进行了相应的调整，如此操作可以改变类分布平衡度从而达到所需水平。

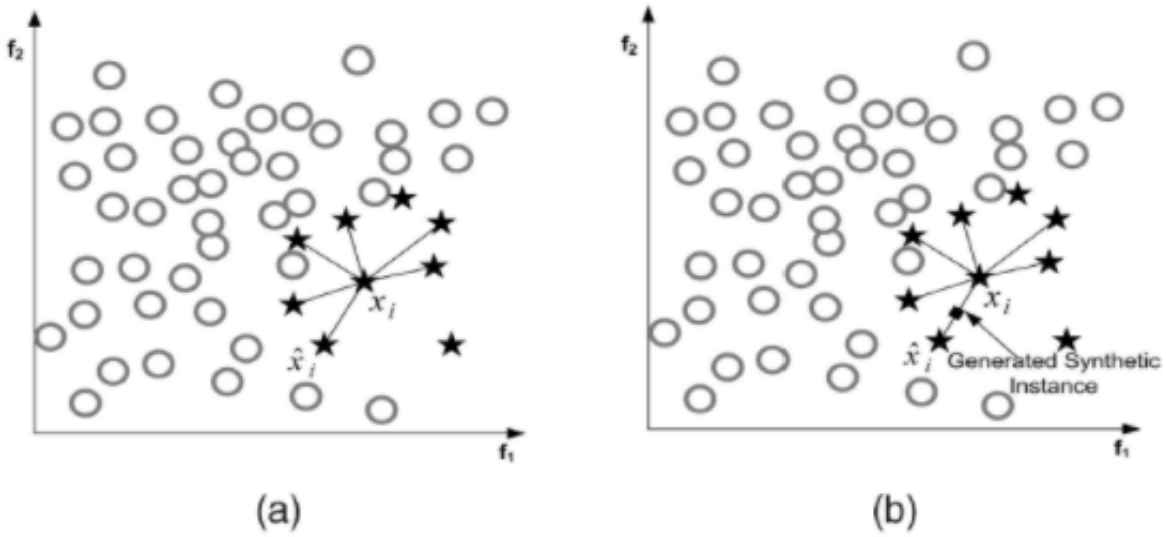
重复样本过多，容易造成分类器的过拟合

2.2 SMOTE算法(Synthetic Minority Oversampling Technique)

在合成抽样技术方面，Chawla NY等人提出的SMOTE过抽样技术是基于随机过采样算法的一种改进方案，由于随机过采样简单复制样本的策略来增加少数类样本，这样容易产生模型过拟合的问题，即使模型学习到的信息过于特别（Specific）而不够泛化(General)。

SMOTE的主要思想是利用特征空间中现存少数类样本之间的相似性来建立人工数据，特别是，对于子集$S_{min} \subset S$，对于每一个样本$x_i \subset S_{min}$使用K-近邻法，其中K-近邻被定义为考虑S_{min}中的K个元素本身与x_i的欧氏距离在n维特征空间X中表现为最小幅度值的样本。由于不是简单地复制少数类样本，因此可以在一定程度上避免分类器的过度拟合，实践证明此方法可以提高分类器的性能。但是由于对每个少数类样本都生成新样本，因此容易发生生成样本重叠（overlapping）的问题。算法流程如下：

- 1）对于少数类中的每一个样本(x_i)，以欧氏距离为标准计算它到少数类样本集S_{min}中所有样本的距离，得到K近邻；
- 2）根据样本不平衡比例设置一个采样比例以确定采样倍率N，对于每一个少数类样本x_i，从其K近邻中随机选择若干个样本， 假设选择的近邻为$\tilde{x}$；
- 3）对于每一个随机选出的近邻$\tilde{x}$，分别与原样本按照如下的公式构建新的样本:
  $x_{new} = x + rand\ (0, 1) \times (\tilde{x} - x)$

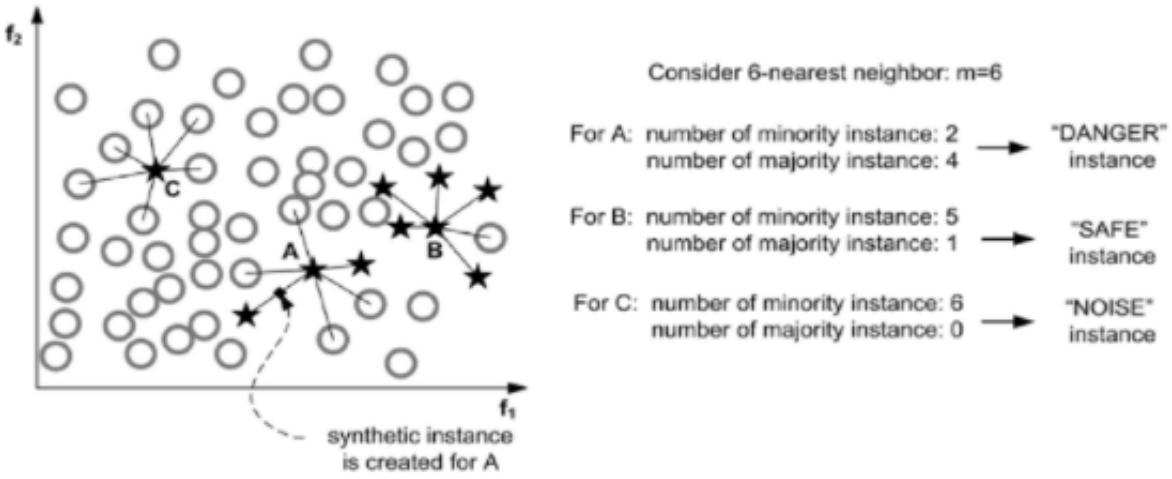


2.3 Borderline-SMOTE算法

原始的SMOTE算法对所有的少数类样本都是一视同仁的，但实际建模过程中发现那些处于边界位置的样本更容易被错分，因此利用边界位置的样本信息产生新样本可以给模型带来更大的提升。Borderline-SMOTE便是将原始SMOTE算法和边界信息算法结合的算法。算法流程如下:

1.首先，对于每个$x_i \subset S_{min}$确定一系列K-近邻样本集，称该数据集为S_{i-kNN}，且$S_{i-kNN} \subset S$;
2.然后，对每个样本x_i，判断出最近邻样本集中属于多数类样本的个数，即：$|S_{i-kNN} \cap S_{maj}|$;
3.最后，选择满足下面不等式的$x_i: \frac{k}{2} < |S_{i-kNN} \cap S_{maj}| < k$,将其加入危险集$DANGER$，

对危险集中的每一个样本点（最容易被错分的样本），采用普通的$SMOTE$算法生成新的少数类样本。



三、欠采样

3.1 随机欠采样

减少多数类样本数量最简单的方法便是随机剔除多数类样本，可以事先设置多数类与少数类最终的数量比例，在保留少数类样本不变的情况下，根据比例随机选择多数类样本。

- 1）首先我们从S_{maj}中随机选取一些多数类样本E
- 2）将这些样本从S_{maj}中移除，就有$|S_{maj-new}| = |S_{maj} - |E|$

优点在于操作简单，只依赖于样本分布，不依赖任何距离信息，属于非启发式方法；缺点在于会丢失一部分多数类样本的信息，无法充分利用已有信息。

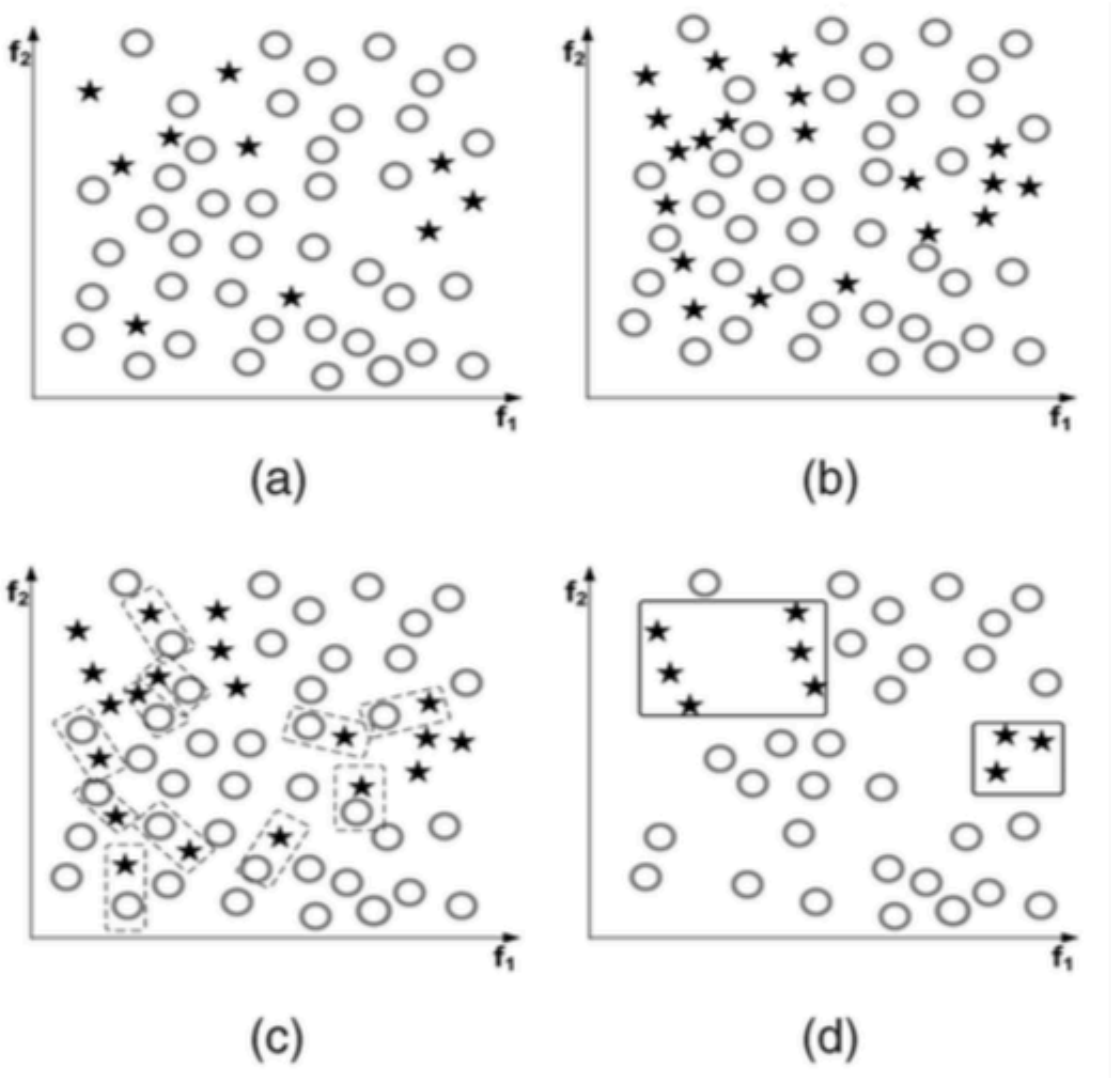
3.2 Tomek Links方法

定义：Tomek links被定义为相反类最近邻样本之间的一对连接。

符号约定：给定一个样本对(x_i, x_j)，其中$x_i \in S_{maj}$，$x_j \in S_{min}$，记$d(x_i, x_j)$是样本x_i和x_j之间的距离

公式表示：如果不存在任何样本x_k，使得$d(x_i, x_k) < d(x_i, x_j)$，那么样本对$(x_i, x_j)$被称为Tomek Links

使用这种方法，如果两个样本来自Tomek Links，那么他们中的一个样本要么是噪声要么它们都在两类的边界上。所以Tomek Links一般有两种用途：在欠采样中：将Tomek Links中属于是多数类的样本剔除；在数据清洗中，将Tomek Links中的两个样本都剔除。



3.3 NearMiss方法

NearMiss方法是利用距离远近剔除多数类样本的一类方法，实际操作中也是借助KNN，总结起来有以下几类：

- 1）NearMiss-1：在多数类样本中选择与最近的三个少数类样本的平均距离最小的样本
- 2）NearMiss-2：在多数类样本中选择与最远的3个少数类样本的平均距离最小的样本
- 3）NearMiss-3：对于每个少数类样本，选择离它最近的给定数量的多数类样本

NearMiss-1和NearMiss-2方法的描述仅有一字之差，但其含义是完全不同的：NearMiss-1考虑的是与最近的3个少数类样本的平均距离，是局部的；NearMiss-2考虑的是与最远的3个少数类样本的平均距离，是全局的。

NearMiss-1方法得到的多数类样本分布也是”不均衡“的，它倾向于在比较集中的少数类附近找到更多的多数类样本，而在孤立的（或者说是离群的）少数类附近找到更少的多数类样本，原因是NearMiss-1方法考虑的局部性质和平均距离。

NearMiss-3方法则会使得每一个少数类样本附近都有足够多的多数类样本，显然这会使得模型的精确度高、召回率低。

实验结果表明得到NearMiss-2的不均衡分类性能最优。

四、Informed Understanding

Informed欠抽样算法可以解决传统随机欠采样造成的数据信息丢失问题，且表现出较好的不均衡数据分类性能。其中有一些集成（ensemble）的想法，主要有两种方法，分别是EasyEnsemble算法和BalanceCascade算法。

4.1 EasyEnsemble算法

它把数据划分为两部分，分别是多数类样本和少数类样本，对于多数类样本S_{maj}，通过n次有放回抽样生成n份子集，少数类样本S_{min}分别和这n份样本合并训练AdaBoost分类器，这样可以得到n个模型，最终的模型采用加权多数表决的方法，加大分类误差率小的弱分类器的权值，使其在表决中起较大的作用，减小分类误差率小的弱分类器的权值，使其在表决中起较小的作用。这里假设多数类样本为N，少数类样本为P，算法流程如下：

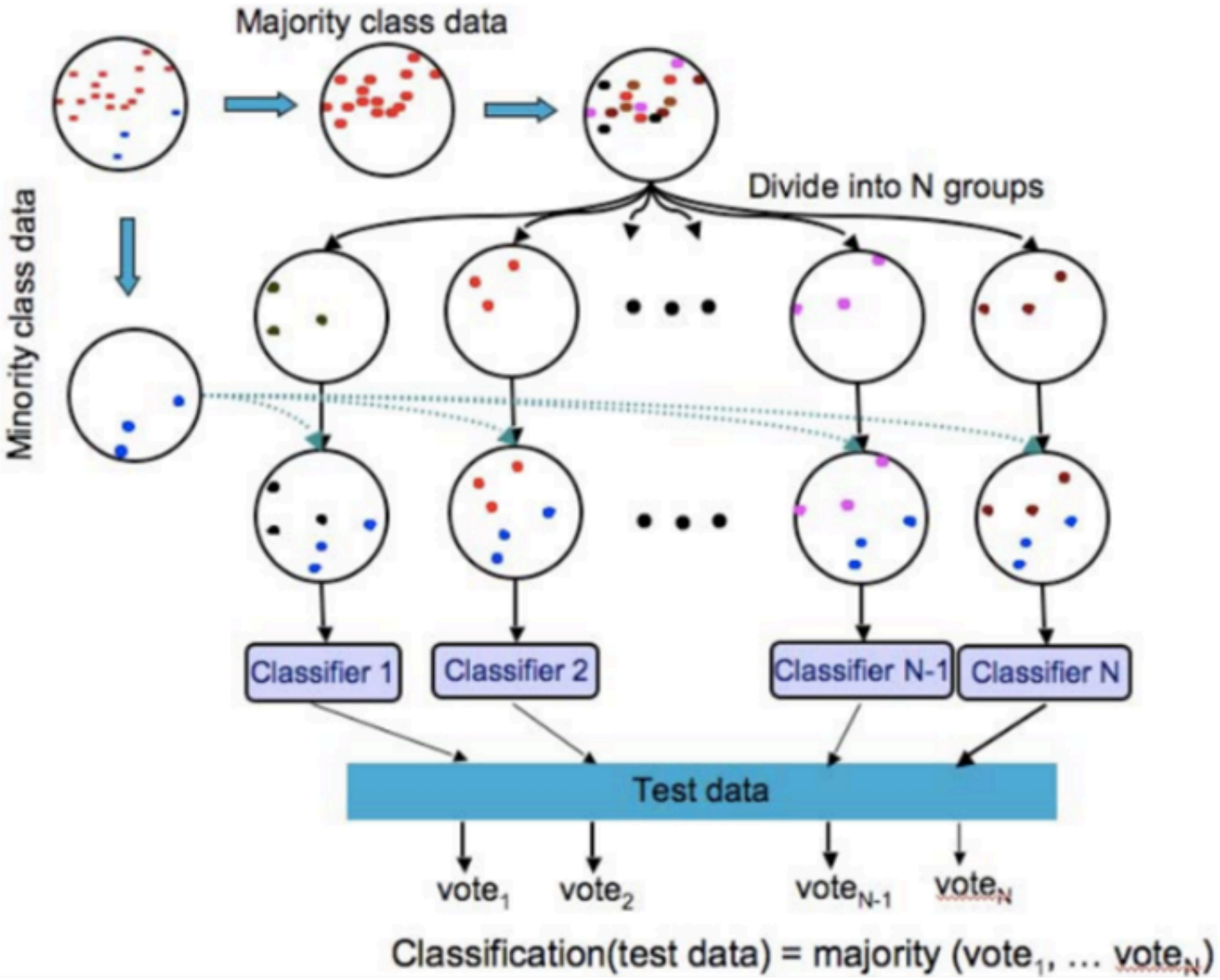
1. 对i从1到T重复以下过程：
   - 从$\mathcal{N}$中随机抽取样本子集$\mathcal{N}_i$，使得$|\mathcal{N}_i| = |\mathcal{P}|$
   - 利用$\mathcal{N}_i$和$\mathcal{P}$训练AdaBoost分类器H_i，阈值设置为θ_i，则

$$H_i(x) = \text{sgn}(\sum_{j=1}^{s_i} \alpha_{i,j} h_{i,j}(x) - \theta_i)$$

2. 最终分类器为

$$H(x) = \text{sgn}(\sum_{i=1}^{T} \sum_{j=1}^{s_i} \alpha_{i,j} h_{i,j}(x) - \sum_{i=1}^{T} \theta_i)$$

EasyEnsemble的想法是多次随机欠抽样，尽可能全面地涵盖所有信息，算法特点是利用boosting减小偏差（Adaboost）、bagging减小方差（集成分类器）。实际应用的时候也可以尝试选用不同的分类器来提高分类的效果。



4.2 BalanceCascade算法

EasyEnsemble算法训练的子过程是独立的，BalanceCascade则是一种级联算法，这种级联的思想在图像识别中用途非常广泛。算法流程如下：

1. 设定训练层级T，$f=\sqrt[T-1]{\frac{|\mathcal{P}|}{|\mathcal{N}|}}$，其中$f$为每一层级的分类器都要达到的FPR（false positive rate）
2. 对i从1到T重复以下过程：
   - 从$\mathcal{N}$中随机抽取样本子集$\mathcal{N}_i$，使得$|\mathcal{N}_i|=|\mathcal{P}|$
   - 利用$\mathcal{N}_i$和$\mathcal{P}$训练AdaBoost分类器H_i，阈值设置为θ_i，则

$$H_i(x)=\text{sgn}(\sum_{j=1}^{s_i}\alpha_{i,j}h_{i,j}(x)-\theta_i)$$

   - 调整阈值θ_i使得分类器H_i的FPR为f
   - 将$\mathcal{N}$中被H_i正确分类的样本剔除
3. 最终分类器为

$$H(x)=\text{sgn}(\sum_{i=1}^{T}\sum_{j=1}^{s_i}\alpha_{i,j}h_{i,j}(x)-\sum_{i=1}^{T}\theta_i)$$

BalanceCascade算法得到的是一个级联分类器，将若干个强分类器由简单到复杂排列，只有和少数类样本特征比较接近的才有可能输入到后面的分类器，比如边界点，因此能更充分地利用多数类样本的信息，一定程度上解决随机欠采样的信息丢失问题。

五、综合采样

目前为止我们使用的重采样方法几乎都是只针对某一类样本：对多数类样本欠采样，对少数类样本过采样。也有人提出将欠采样和过采样综合的方法，解决样本类别分布不平衡和过拟合问题，本部分介绍其中的SMOTE+Tomek Links和SMOTE+ENN。

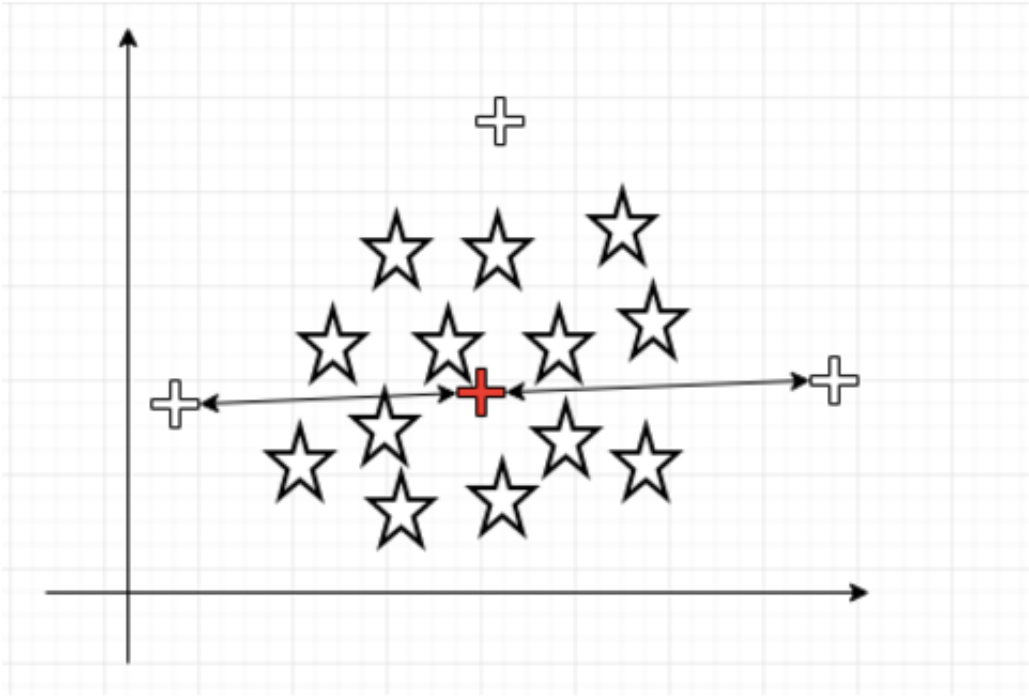
5.1 SMOTE+Tomek Links

SMOTE+Tomek Links方法的算法流程非常简单：
1.利用SMOTE方法生成新的少数类样本，得到扩充后的数据集T
2.剔除T中的Tomek Links对

普通的SMOTE方法生成的少数类样本是通过线性插值得到的，在平衡类别分布的同时也扩张了少数类的样本空间，产生的问题是可能原本属于多数类样本的空间被少数类“入侵”，容易造成模型的过拟合。

Tomek Links对寻找的是那种噪声点或者边界点，可以很好地解决“入侵”的问题，下图红色加号为SMOTE产生的少数类样本，可以看到，红色样本“入侵”到原本属于多数类样本的空间，这种噪声数据

问题可以通过Tomek Links很好地解决。



由于第一步SMOTE方法已经很好地平衡了类别分布，因此在使用Tomek Links对的时候考虑剔除所有的Tomek Links对。

5.2 SMOTE+ENN

SMOTE+ENN方法和SMOTE+Tomek Links方法的想法和过程都是很类似的：

- 1）利用SMOTE方法生成新的少数类样本，得到扩充后的数据集T
- 2）对T中的每一个样本使用KNN（一般K取3）方法预测，若预测结果与实际类别标签不符，则剔除该样本。