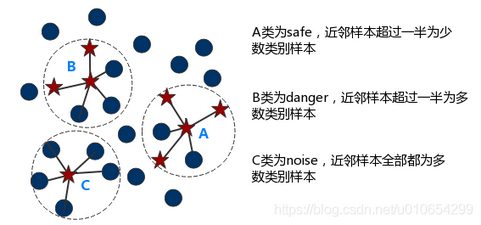
**Borderline SMOTE**算法

**Borderline SMOTE** 是在SMOTE基础上改进的过采样算法, 该算法仅使用边界上的少数类样本来合成新样本并添加到数据集中, 从而改善样本的类别分布。

Borderline smote采样过程是将少数类样本分为3类, 分别是safe、danger和Noise, 具体说明如下, 最后, 仅对danger的少数类样本过采样。(距离计算为K近邻算法)。（k一般取5）。

1. Safe: 样本周围一半以上均为少数类样本， 如图中点A
2. Danger: 样本周围一半以上均为多数类样本, 视为在边界上的样本, 如图中点B
3. Noise; 样本周围均为多数类样本, 视为噪音, 如图中点C.



Borderline-SMOTE又可分为**Borderline-SMOTE1**和Borderline-SMOTE2，Borderline-SMOTE1在对Danger点生成新样本时，在K近邻随机选择少数类样本（与SMOTE相同），Borderline-SMOTE2则是在k近邻中的任意一个样本(不关注样本类别)

本项目使用**Borderline-SMOTE 1,** 即对danger点采用SMOTE算法,

SMOTE算法的基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本，具体如图2所示，算法流程如下。

1. 对于少数类中的每一个danger样本X, 以欧氏距离为标准,计算它到少数类样本集中所有样本的距离, 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，对于每一个danger样本X, 从其k近邻中随机选择N个样本, 假设选择的样本为Xn。
3. 对于每一个随机选出的近邻Xn, 分别与原样本按照如下的公式构建新的样本

Xnew = X+rand(0, 1) \*（Xn-X）