## 0 概述

机器学习应用到实际问题中的一般流程为：数据预处理 -> 特征工程 -> 建模 -> 训练 -> 预测。

可以看出，第一个阶段就是数据预处理。数据预处理主要是将收集到的数据进行清洗、格式转换等。因为收集的raw数据往往存在各种各样的问题（缺失值、格式不符合我们的需求等）。

现在业界流行的深度学习模型，号称是“端对端”的建模，可以省略特征工程这一步，但是数据预处理仍然是需要的。

## 第一章 数据格式转换

## 第二章 处理缺失值

## 第三章 类别型变量的独热编码

## 第四章 数据标准化

## 第五章 数据降维

## 第六章 数据不平衡问题处理

### 1、样本不平衡问题简介

“样本不平衡”，更完整的说法是“样本类别不平衡”，是针对分类问题而言的。它是指样本中某些类别的样本数相对其它类别的样本数少很多。

样本不平衡分两种情况：

（1）大样本的不平衡：样本的总数较大。如有1000万样本，其中A类样本只有5万。这种情况下模型很容易忽视A类样本；

（2）小样本不平衡：样本总数较小。如有100个样本，其中10个A类样本，90个B类样本。这种情况下A类样本数（绝对数量）很少，包含的信息太少，无法从中提取模式。

实际中样本不平衡的典型场景包括：CTR预估（点击事件是少数类样本）、异常检测（异常点是少数类）、罕见事件分析（罕见事件是少数类）。

工业上的样本不平衡的极限是30：1。

### 2、欠采样

欠采样，又称下采样，是将多数类样本采少，以跟少数类样本平衡。

1、随机欠采样：在多数类样本中随机删除一些样本。

2、PG欠采样算法：基于k-means或其他原型算法的欠采样算法。设少数类样本数是，多数类样本数是，多数类样本平衡后的样本数是（与接近），则PG算法迭代次，每次从多数类样本中选出

### 3、过采样

过采样，又称上采样，是将少数类样本采多，以跟多数类样本平衡。

(1) 随机复制：就是简单的复制少数类样本。这种方法的缺点是容易造成过拟合，因为少数类样本本来信息就少，随机复制相当于加强了少数类样本的信息，但是没有提供新的信息，所以模型比原来更不易在新的样本上泛化。

(2) smote样本构建算法：与随机复制相反，它构建新的样本。

smote过采样方法的思路很简单：在少数类样本中，任意选取两个样本，然后进行线性插值，产生新的样本。但是这个做法也过于简单，没有考虑到一些信息，比如少数类样本的分布特征，这样在少数类与多数类的分界线附近容易造成误差；

基于简单的smote过采样算法，有很多改进的方法。如borderline-smote算法、kmeans-smote算法等，并且smote算法可以与Boosting算法结合，形成smoteBoost。

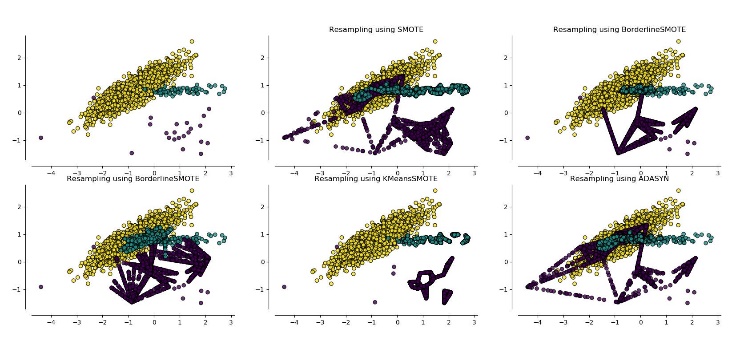


图1 各种smote的改进算法

### 4、cost-sensitive方法：基于模型的方法

cost-sensitive方法是一种通用的方法，它适用于任意模型，基于模型对样本的类别标签进行修改。关于cost-sensitive方法，请参见这篇论文：

《The Foundations of Cost-Sensitive Learning》

cost-sensitive方法的具体实现有meta-cost算法和focal-cost算法等。关于focal-cost，请参见这篇论文：《Focal Loss for Dense Object Detection》。

## 第七章 数据增强