天善智能
TIANSHAN SOFT

评分卡模型中的前沿问题二：非平衡样本的处理

讲师：安迪生

课程简介

- 在信用风控工作中，样本不平衡是常见的问题，即好样本占了绝大多数的比例。此时基于原始数据构建出来的模型容易降低对少类样本的敏感性，不利于模型将好坏样本做区分。因此改善样本的非平衡性是信用风控建模工作中的一部分。

目录

- 过采样与欠采样
- SMOTE算法
- 样本权重法

过采样与欠采样

- 非平衡样本

在分类问题中，经常出现某些类型的样本远少于另外的类型。例如，在信用卡交易中，“盗刷”相比较正常交易，出现的频率低很多。如果我们搜集数据进行盗刷交易的预测，则样本中正常交易的样本量必然远多对于盗刷的样本。

非平衡样本的危害在于，其容易造成分类器在多数类精度较高、少数类的分类精度很低。以最大分类精度为目标，导致算法提高多数样本分类精度而忽略小样本的预测精度。例如，分类器对于正常交易的识别能力很强，很对于盗刷交易则不敏感。

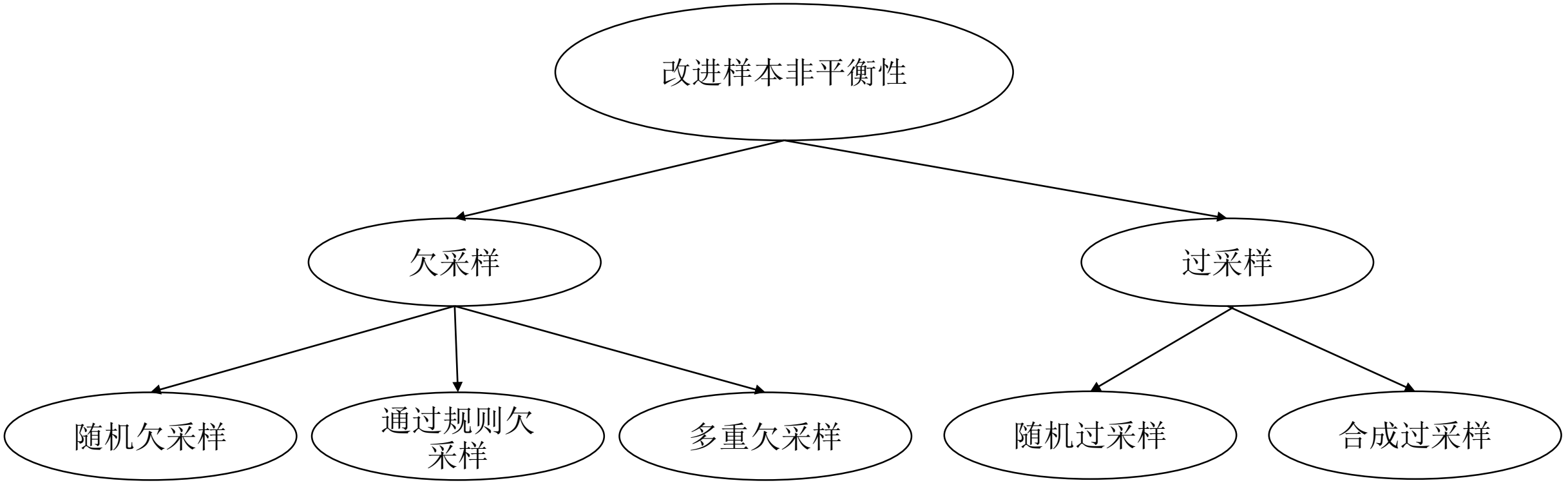
解决非平衡样本对建模的危害由两类方案：

1，改进样本的非平衡性

2，改进损失函数的权重

过采样与欠采样

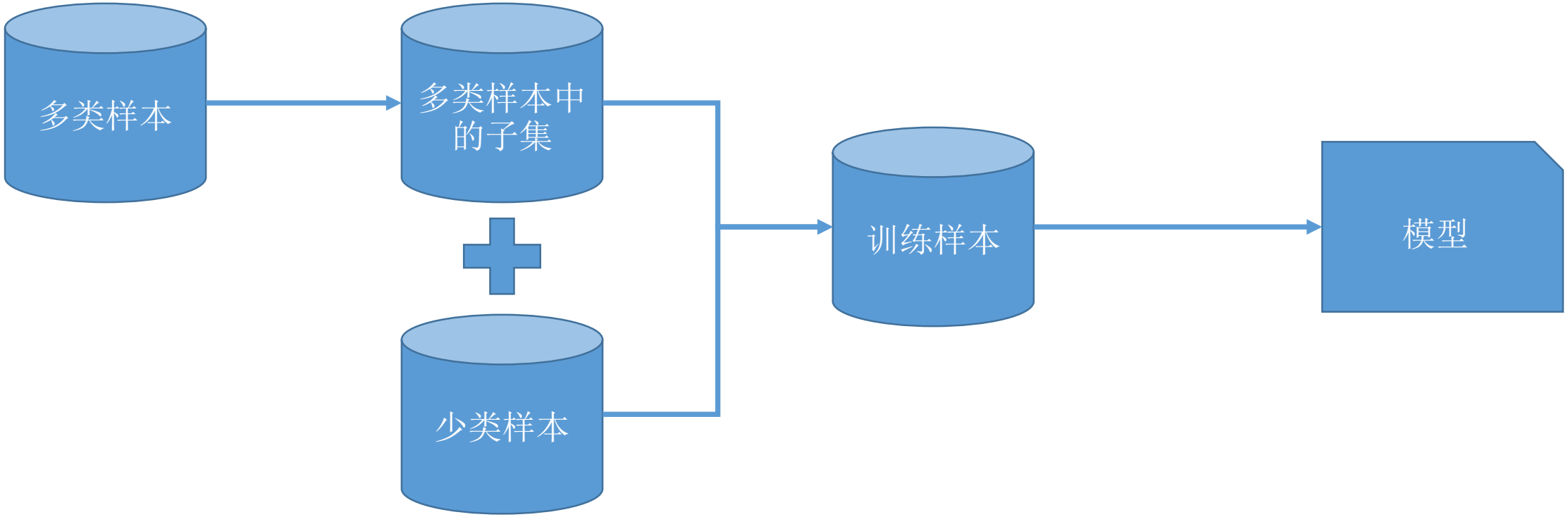
- 改进样本的非平衡性



过采样与欠采样

- 简单欠采样

简单欠采样法是从多类样本中随机抽取一定比例的样本，和少类样本组成新的训练集。该训练集中的两类样本的比例相对较为均衡。



过采样与欠采样

- 简单欠采样（续）

简单欠采样的优点

- 简单易实施

缺点

- 当多类样本分布不均匀时，简单欠采样下的模型的偏差较大
- 容易丢失重要信息
- 当总体样本量不大时，形成的新训练集的样本量更少，从而降低模型的可信度

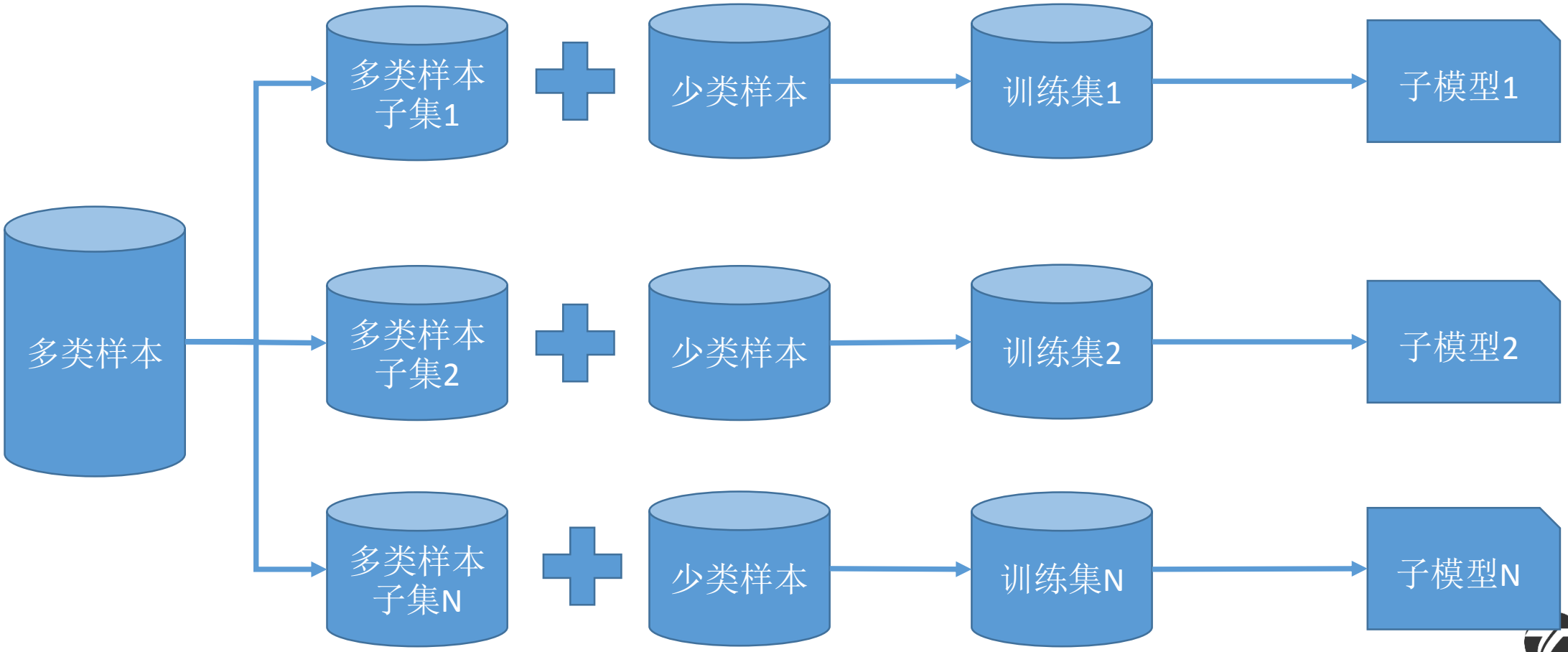
注意事项

- 与训练集不同，测试集中不需要对多类样本进行欠采样。建议用实际的比例进行测试

过采样与欠采样

- 多重欠采样

为了克服简单采样下样本信息容易丢失、模型容易有偏的情形，我们可以采用多重欠采样的方式来构建集成模型。具体流程如下：



过采样与欠采样

- 多重欠采样

多重欠采样的优点

- 保持信息不丢失
- 模型的泛化能力不会减弱，健壮性更强

多重欠采样的缺点

- 在训练阶段增加了复杂度

注意事项

- 训练得到的子模型可以是同质的，也可以是异质的

过采样与欠采样

- 案例

我们将训练集中的好样本随机分成了10等份， 每一份都和坏样本组成新的训练集，训练出10个逻辑回归模型。

在测试集上，每一个逻辑回归模型都对测试样本进行预测，最终对概率进行加权融合。融合后的概率的AUC值从75%增加到77%。

过采样与欠采样

- 简单过采样

与简单欠采样相对应的是简单过采样，即对少类样本按比例进行重复抽样。简单过采样可以改善非平衡性但是没有生成新的样本，并且会放大少类样本中的噪声（即错误数据）。

优点:

- 简单易实施

缺点：

- 容易放大少类样本中的噪声

注意事项

- 测试集中不需要进行过采样
- 有些模型（例如决策树）在构建时不需要对少类样本进行过采样，可以将过采样后的比例看成权重带入模型中

SMOTE算法

- 合成过采样：SMOTE算法

合成少数类的过采样技术，基于简单过采样的一种改进方案。由于随机过采样采取简单复制样本的策略增加少数类样本，容易产生过拟合，使得模型不够 泛化。

smote基本思想：对少数类样本分析，并根据少数类样本人工合成新样本添加到数据集中。

1、对于少数类中每个样本，得到K近邻（K自定义）。

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。对每个少数类样本x_0，从其k近邻中随机选取若干个样本

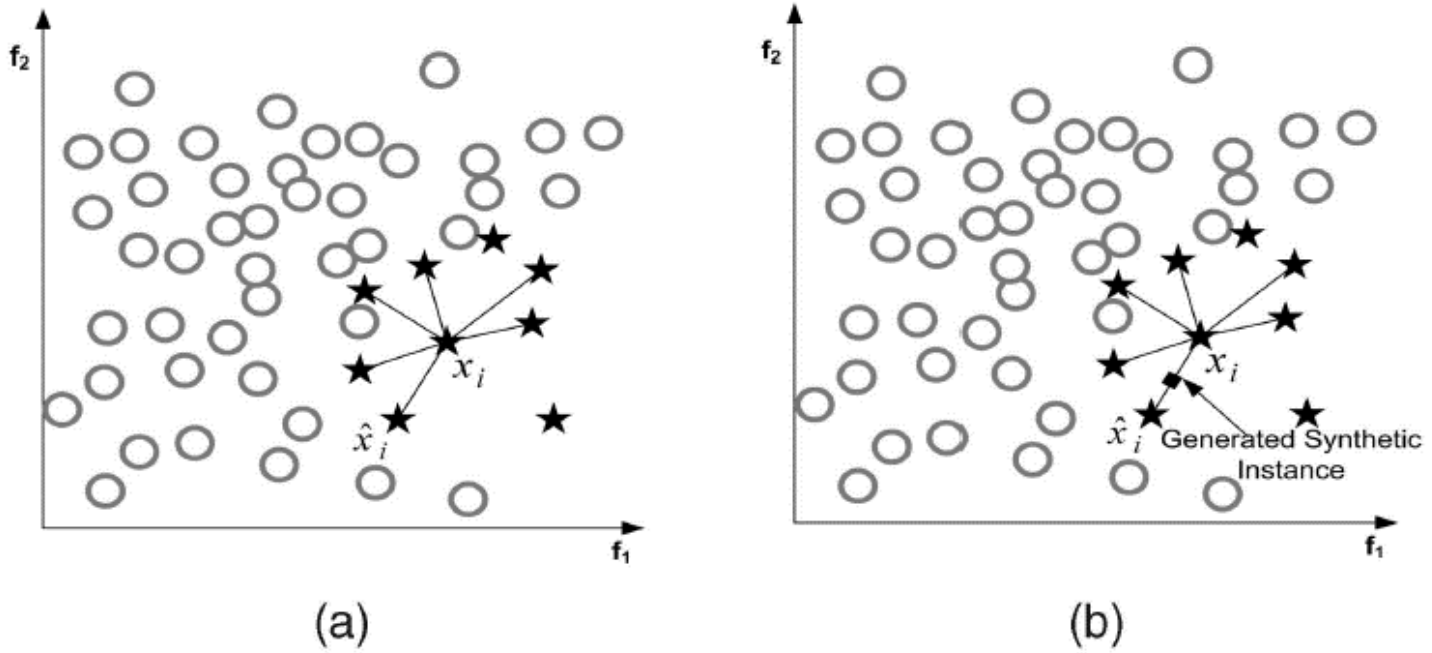
3、对于每个随机选出的近邻样本x_1，分别与原样本按照如下公式构建新样本。

$$x_{new} = x_0 + uniform(0,1) \times (x_0 - x_1)$$

SMOTE算法

- 合成过采样：SMOTE算法（续）

SMOTE算法摒弃了随机过采样复制样本的做法，可以防止随机过采样易过拟合的问题，实践证明此方法可以提高分类器的性能。



SMOTE算法

- 合成过采样：SMOTE算法（续）

SMOTE算法在风控模型中面临的问题：

① 绝大多数场景中数据集是混合型数据，存在非数值特征。对于非数值特征，需要解决2个问题，即如何在寻找k邻近样本中考虑非数值特征，以及在新样本生成中如何生成非数值特征。

② SMOTE会对特征工程会产生影响，例如分箱。分箱前或者分箱后利用SMOTE算法进行样本扩充，会对分箱产生不同的影响

③ 需要解决特征尺度归一化的问题。特征的尺度会影响k邻近样本的寻找。

SMOTE算法

- 解决方案：非数值型特征

对于非数值特征，在寻找K邻近样本的时候，使用其他数值型特征在少类样本上的标准差的中位数作为距离，带入欧式距离的计算中。例：

	F1	F2	F3	F4	F5
样本1	1	2	1	A	C
样本2	3	3	5	A	D
样本3	4	6	3	B	D

当计算样本1和样本2的距离时，特征F4和F5是类别型变量。其中，特征F4在二者上的取值是相同的，不需要计算距离。特征F5也是类别型变量且在算样本1和样本2上的取值不同。这时我们需要计算数值型特征F1、F2和F3的标准差的中位数，即1.63. 于是算样本1和样本2的欧式距离为：

$$D = \sqrt{(1-3)^2 + (2-3)^2 + (1-5)^2 + 1.63\text{^}2} = 5.26$$

SMOTE算法

- 解决方案：非数值型特征（续）

当生成合成样本时，新样本的数值型变量的取值使用之前介绍的插值公式。而新样本的类别型变量的取值则使用k邻近样本的类别型变量的众数来赋值。例：

	F1	F2	F3	F4	F5
样本1	1	2	1	A	C
样本2	3	3	5	A	D
样本3	4	6	3	B	D

假设样本2和3是样本1的k邻近样本。如果从样本1和样本2中进行新样本合成，新样本在各个变量上的取值为：

$$F1 = 1 + \alpha_1 \times (3 - 1), \alpha_1 \sim uniform(0,1)$$

$$F2 = 2 + \alpha_2 \times (3 - 2), \alpha_2 \sim uniform(0,1)$$

$$F3 = 1 + \alpha_3 \times (5 - 1), \alpha_3 \sim uniform(0,1)$$

$$F4 = majoirty(A, A, B) = A$$

$$F5 = majoirty(C, D, D) = D$$

SMOTE算法

- 样本扩充对IV和WOE的影响

对多类样本（即好样本）的简单随机抽样是不会影响IV和WOE的计算。证明：

	Good	Bad	Good %	Bad %	WOE
Bin 1	G_1	B_1	G_1/G	B_1/B	$\ln(\frac{G_1}{G}/\frac{B_1}{B})$
Bin 2	G_2	B_2	G_2/G	B_2/B	$\ln(\frac{G_2}{G}/\frac{B_2}{B})$
Bin N	G_N	B_N	G_3/G	B_3/B	$\ln(\frac{G_N}{G}/\frac{B_N}{B})$
Total	$G=\sum G_i$	$B=\sum B_i$			

抽样后

	Good	Bad	Good %	Bad %	WOE
Bin 1	$k\times G_1$	B_1	kG_1/kG	B_1/B	$\ln(\frac{G_1}{G}/\frac{B_1}{B})$
Bin 2	$k\times G_2$	B_2	kG_2/kG	B_2/B	$\ln(\frac{G_2}{G}/\frac{B_2}{B})$
Bin N	$k\times G_N$	B_N	kG_3/kG	B_3/B	$\ln(\frac{G_N}{G}/\frac{B_N}{B})$
Total	$G=k\sum G_i$	$B=\sum B_i$			

类似地，可以证明对少类样本进行随机复制，其结果在统计上也不会影响IV和WOE的计算。

但是，SMOTE扩充少类样本时会改变少类样本的分布，因此对IV和WOE的计算产生影响。

样本权重法

- 带权重的损失函数

解决样本非平衡性的另一种方法是将样本的权重带入模型的损失函数中去，从而使得模型在参数估计时提高某类样本的重要性。以逻辑回归模型为例，其损失函数是

$$loss = \sum_{i=1}^{N} \{-y_i \boldsymbol{X_i \beta} + \log(1 + \exp(\boldsymbol{X_i \beta}))\}$$

将样本权重$\{w_i\}$带入loss中，我们有：

$$weighted\ loss = \sum_{i=1}^{N} w_i \{-y_i \boldsymbol{X_i \beta} + \log(1 + \exp(\boldsymbol{X_i \beta}))\}$$

采用梯度下降法求解系数，有：

$$\Delta = \frac{\partial weighted\ loss}{\partial \beta_j} = \sum_{i=1}^{N} (y_i x_{ij} - x_{ij} \frac{\exp(\boldsymbol{x_i \beta})}{1 + \exp(\boldsymbol{x_i \beta})}) w_i$$

$$\beta_j^{k+1} = \beta_j^k - \Delta \times \eta$$

样本权重法

- 案例

在sklearn.linear_model.LogisticRegression中，可以通过参数class_weight来设置样本的权重。其中， class_weight= ‘balanced’ 可以对非平衡样本设置权重从而避免非平衡性的影响。实验结果表明，该方法使得验证集上KS提高了1%。

注：

进行class_weight= ‘balanced’ 的设置时，需要对L1或L2惩罚系数重新调参。

更多商业智能BI和大数据精品视频尽在 www.hellobi.com

BI、商业智能
数据挖掘 大数据
数据分析师
R语言 Python
机器学习
深度学习
人工智能
Hive Hadoop
Tableau
BIEE ETL
数据科学家
PowerBI

天善智能
TIANSHAN SOFT