多分类学习

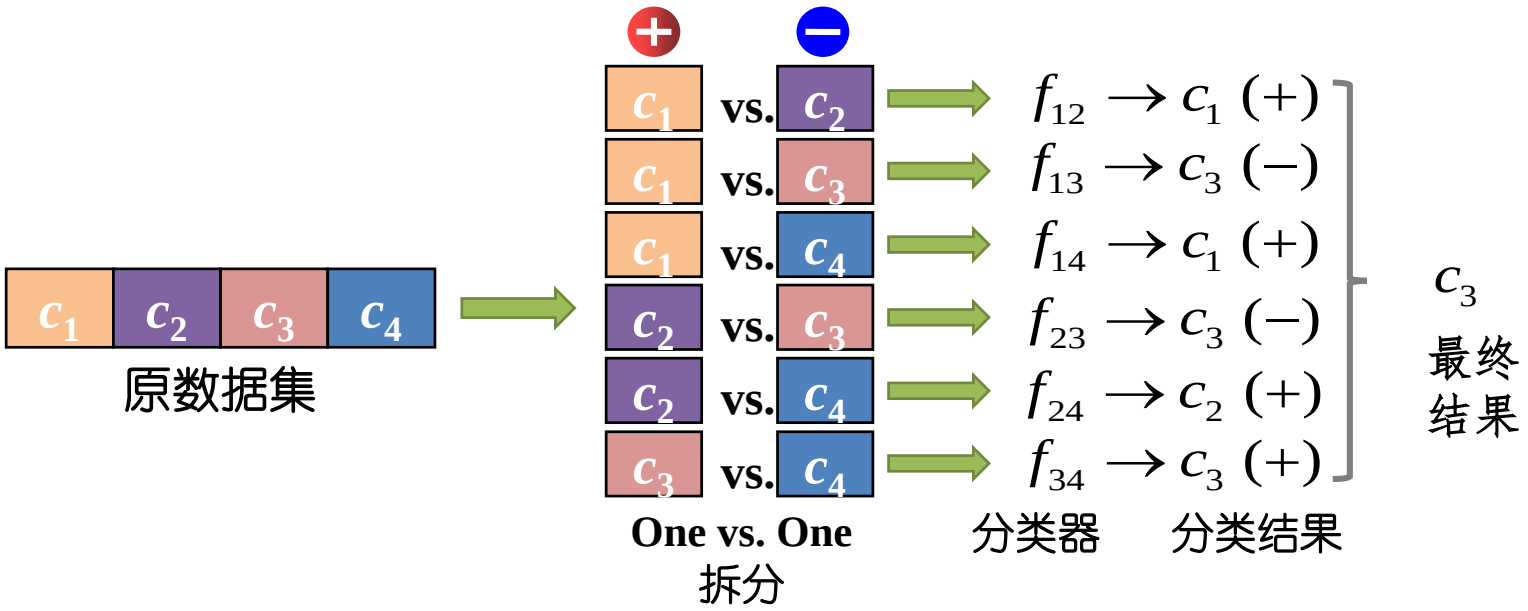
多分类学习

- 多分类学习方法
  - 二分类学习方法推广到多类
  - 利用二分类学习器解决多分类问题（常用）
    - 对问题进行拆分，为拆出的每个二分类任务训练一个分类器
    - 对于每个分类器的预测结果进行集成以获得最终的多分类结果

- 拆分策略
  - 一对一（One vs. One, OvO）
  - 一对其余（One vs. Rest, OvR）
  - 多对多（Many vs. Many, MvM）

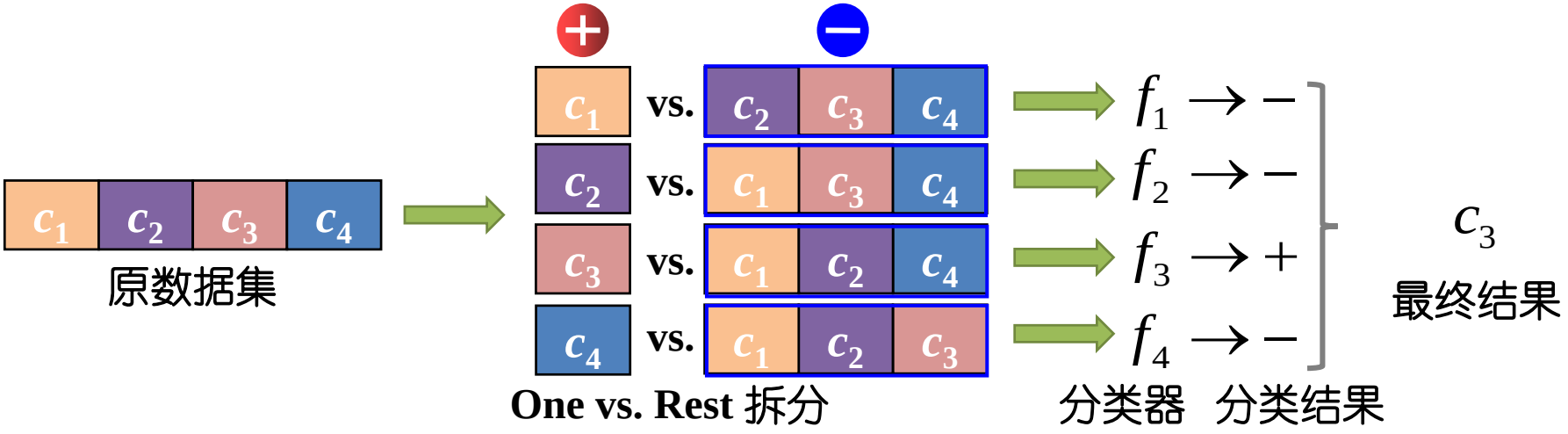
多分类学习 - 一对一

- 拆分阶段
  - N个类别两两配对
    - N(N-1)/2 个二类任务
  - 各个二类任务学习分类器
    - N(N-1)/2 个二类分类器
- 测试阶段
  - 新样本提交给所有分类器预测
    - N(N-1)/2 个分类结果
  - 投票产生最终分类结果
    - 被预测最多的类别为最终类别



多分类学习 – 一对其余

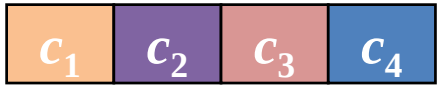
- 任务拆分
  - 某一类作为正例，其他反例
    - N 个二类任务
  - 各个二类任务学习分类器
    - N 个二类分类器
- 测试阶段
  - 新样本提交给所有分类器预测
    - N 个分类结果
  - 比较各分类器预测置信度
    - 置信度最大类别作为最终类别

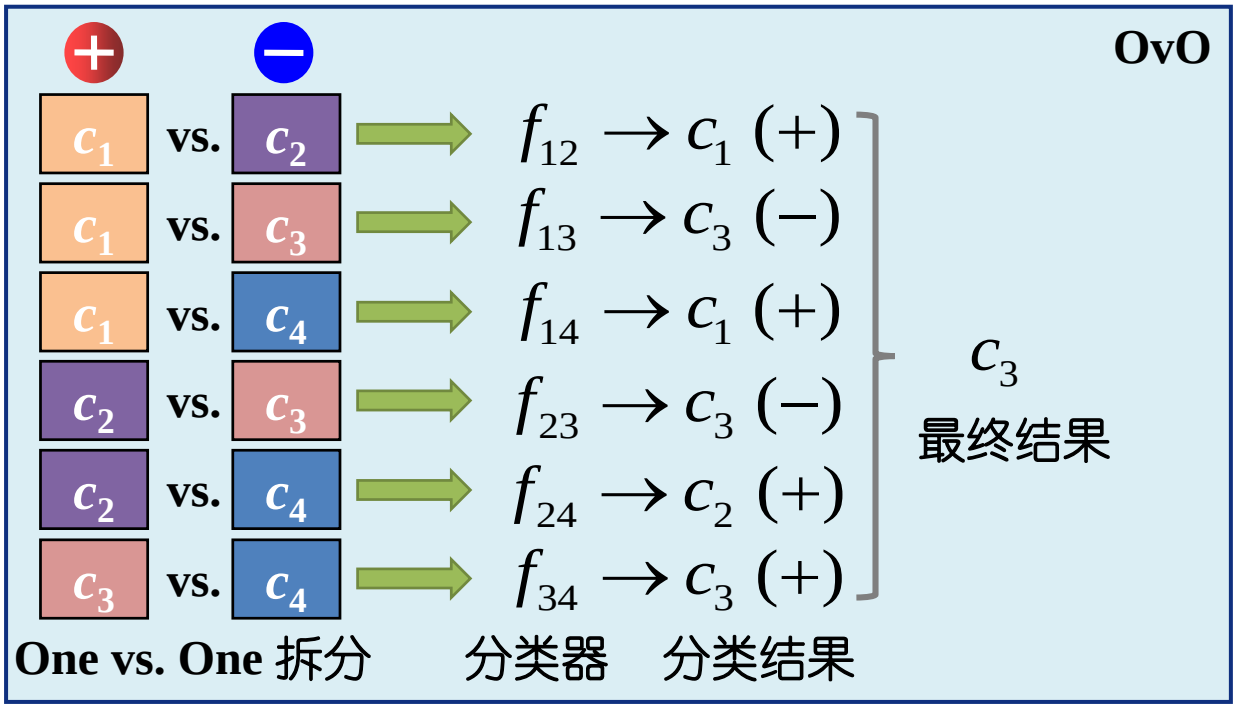


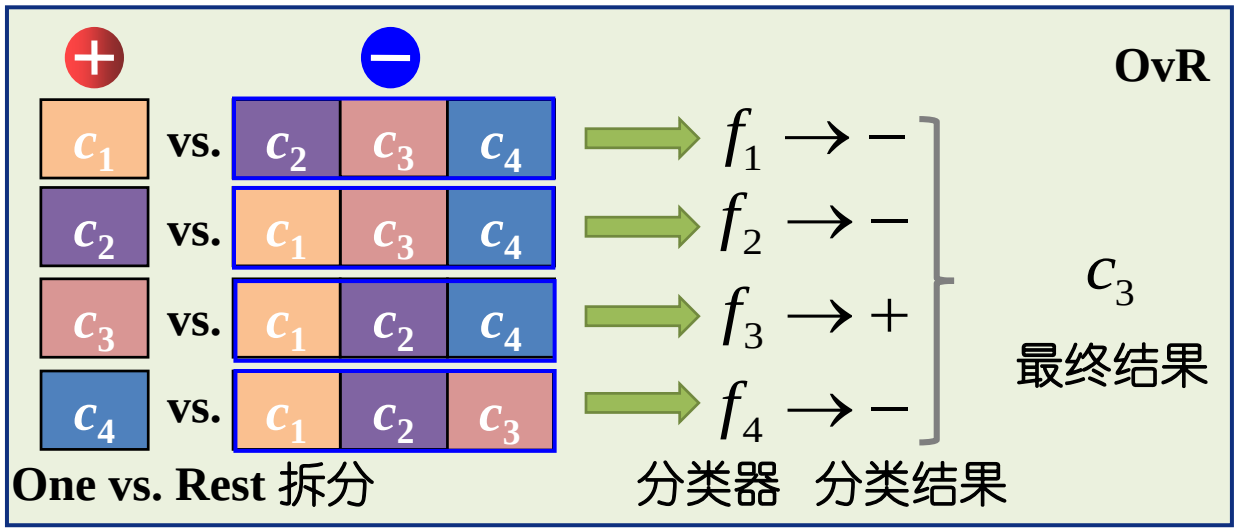
多分类学习 – 两种策略比较

大连理工大学 人工智能学院
School of Artificial Intelligence, Dalian University of Technology

- **OvO的存储开销和测试时间开销通常比OvR大**：OvR只需训练N个分类器，而OvO需训练$N(N-1)/2$个分类器。
- **类别多时，OvO的训练时间开销通常比OvR小**：训练时，OvR的每个分类器均使用全部训练样本，而OvO的每个分类器仅用到两个类样本；
- **预测性能差不多**：至于预测性能，则取决于具体的数据分布，在多数情形下两者差不多。



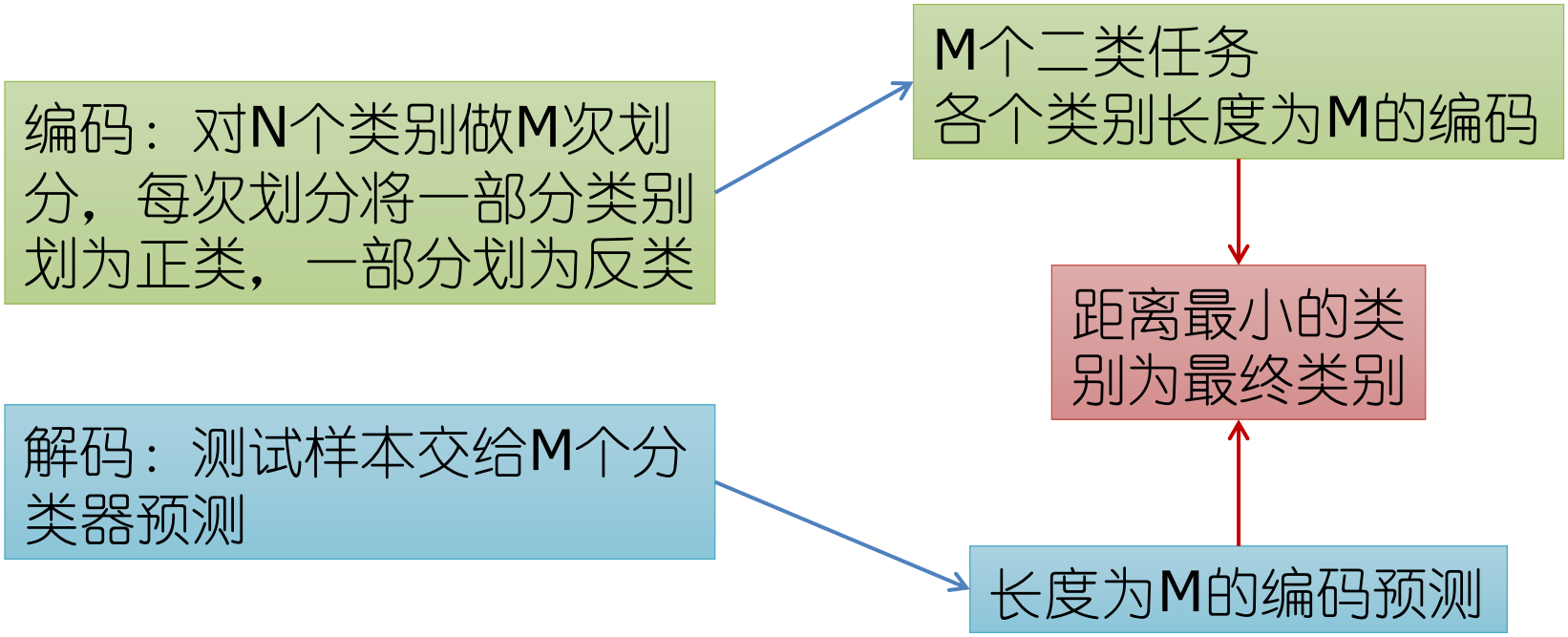




多分类学习 - 多对多

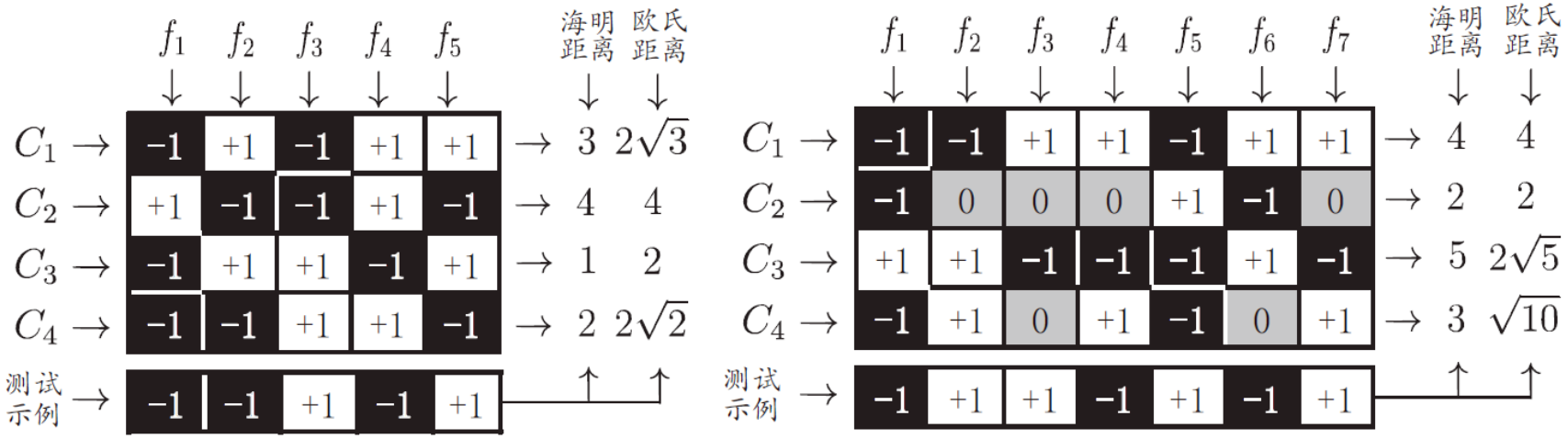
- 多对多（Many vs Many, MvM）
  - 若干类作为正类，若干类作为反类

- 纠错输出码（Error Correcting Output Code, ECOC）



多分类学习 – 多对多

- 纠错输出码(Error Correcting Output Code, ECOC)



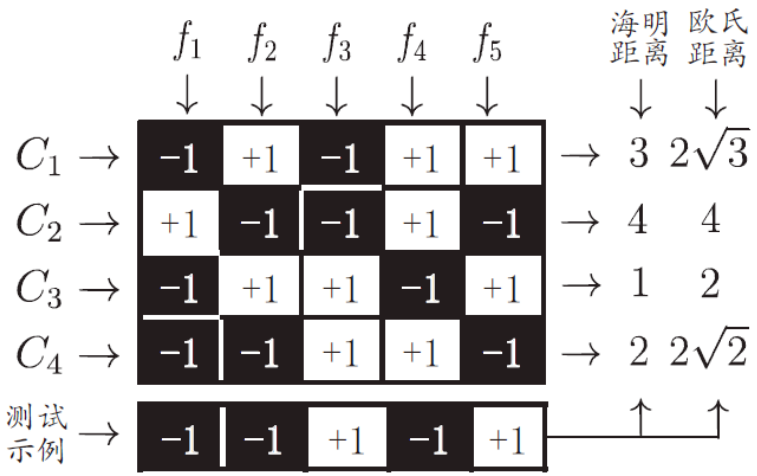
(a) 二元 ECOC 码 [Dietterich and Bakiri,1995]

(b) 三元 ECOC 码 [Allwein et al. 2000]

- ECOC编码对分类器错误有一定容忍和修正能力，编码越长、纠错能力越强
- 对同等长度的编码，理论上来说，任意两个类别之间的编码距离越远，则纠错能力越强

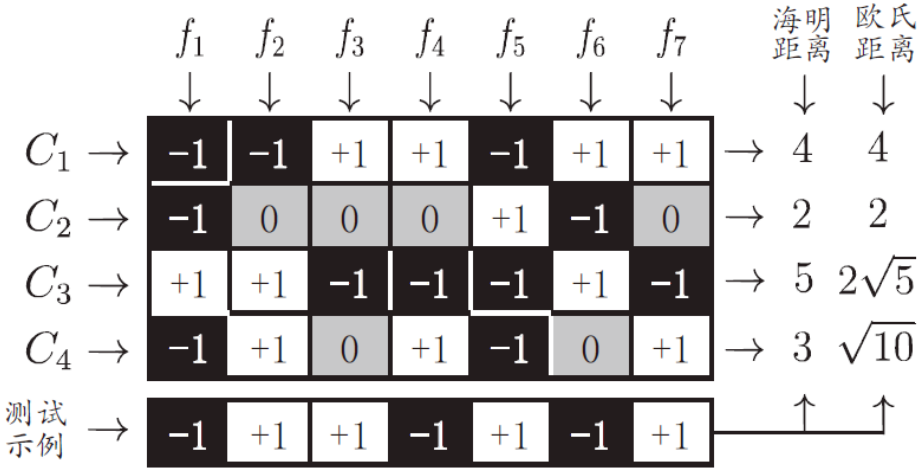
多分类学习 - 多对多

- 纠错输出码(Error Correcting Output Code, ECOC)



(a) 二元 ECOC 码

[Dietterich and Bakiri,1995]



(b) 三元 ECOC 码

[Allwein et al. 2000]

对于二元ECOC码：首先进行异或操作

编码值	-1	+1	-1	+1	+1
测试值	-1	-1	+1	-1	+1
异或	0	1	1	1	0

海明距离= 0+1+1+1+0 = 3

对于三元ECOC码，规则同上，但是三元操作不进行异或

海明距离：+1和−1之间的海明距离为1

+1/−1和0之间的海明距离为0.5

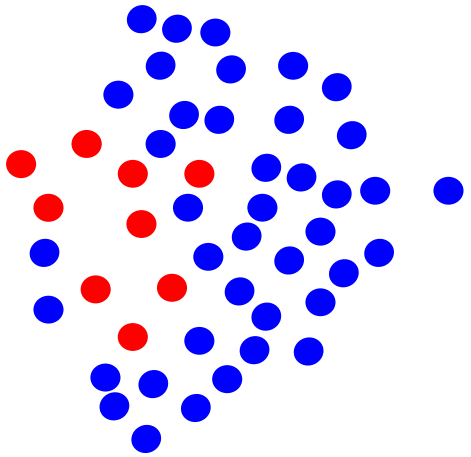
类别不平衡问题

类别不平衡问题

- 类别不平衡（class imbalance）
  - 不同类别训练样例数相差很大情况（正类为小类）

类别平衡正例预测 $\frac{y}{1-y} > 1$

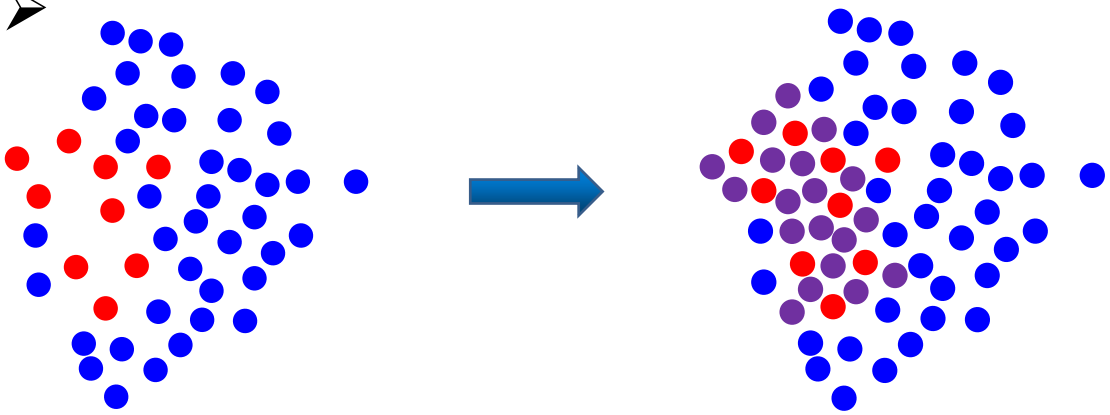
$\frac{y}{1-y} > \frac{m^+}{m^-}$ 正负类比例



- 再缩放
  - 欠采样（undersampling）
    - 去除一些反例使正反例数目接近（EasyEnsemble [Liu et al.,2009]）
  - 过采样（oversampling）
    - 增加一些正例使正反例数目接近（SMOTE [Chawla et al.2002]）

类别不平衡问题

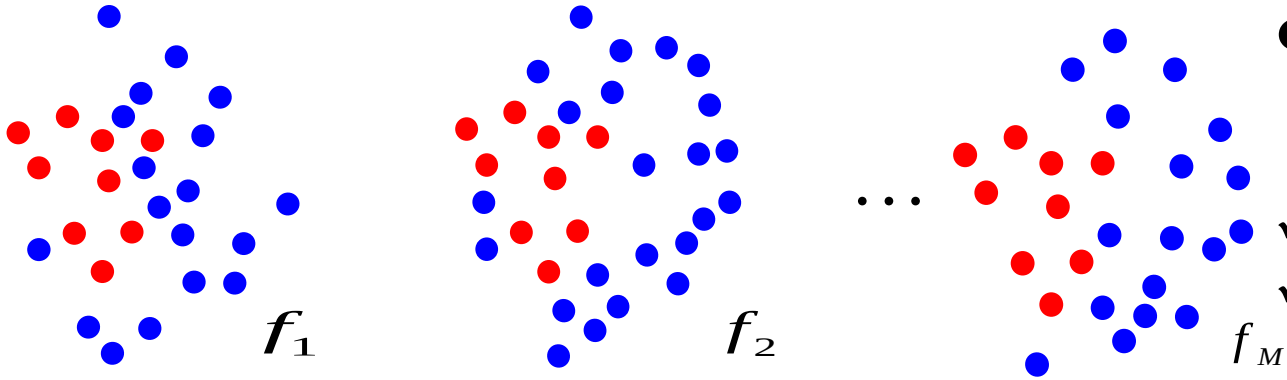
- 过采样 (oversampling):



- ✓ 样本复制
- ✓ 样本插值
- ✓ 样本生成 (GAN)

[1] Chawla N V, Bowyer K W, et al. **SMOTE: Synthetic Minority Over-Sampling Technique**. ***JAIR***, 2002.

- 欠采样 (undersampling)



- 集成学习

$$f = \frac{1}{M}\sum_{m=1}^{M} f_M$$

- ✓ EasyEnsemble[2]
- ✓ BalanceCascade[2]

[2] Xu-Ying Liu, Jianxin Wu, Zhi-Hua Zhou. **Exploratory Undersampling for Class-Imbalance Learning**. ***IEEE TSMCB***, 2009.

优化提要

- 各任务下（回归、分类）各个模型优化的目标
  - 最小二乘法：最小化均方误差
  - 对数几率回归：最大化样本分布似然
  - 线性判别分析：投影空间内最小（大）化类内（间）散度

- 参数的优化方法
  - 最小二乘法：线性代数
  - 对数几率回归：凸优化梯度下降、牛顿法
  - 线性判别分析：矩阵论、广义瑞利商

总结

- 线性回归
  - 最小二乘法（最小化均方误差）
- 二分类任务
  - 对数几率回归
    - 单位阶跃函数、对数几率函数、极大似然法
  - 线性判别分析
    - 最大化广义瑞利商
- 多分类学习
  - 一对一
  - 一对其余
  - 多对多
    - 纠错输出码
- 类别不平衡问题
  - 基本策略：再缩放