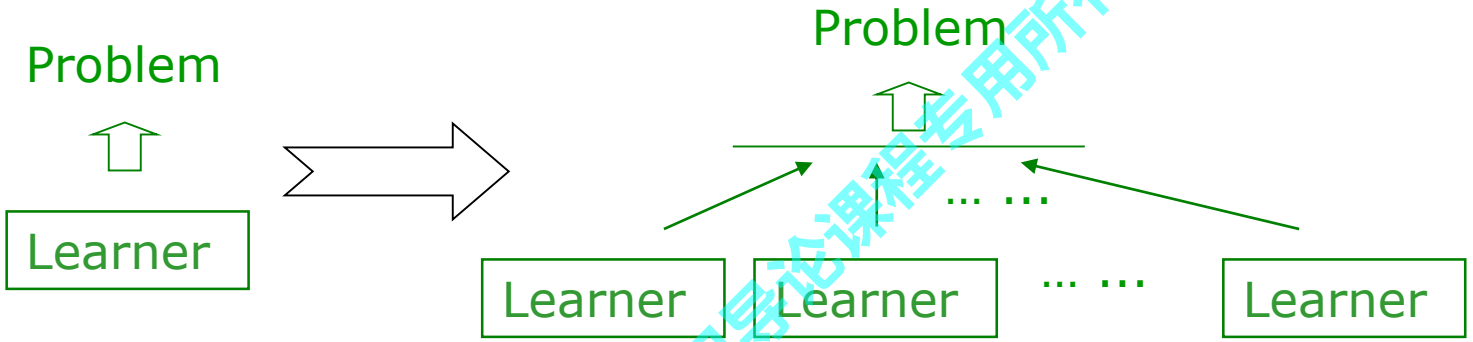
八、集成学习

主讲教师：周志华

集成学习 (Ensemble learning)

集成学习通过构建并结合多个学习器来完成学习任务



- 同质(homogeneous)集成：集成中只包含同种类型的“个体学习器”
  相应的学习算法称为“基学习算法” (base learning algorithm)
  个体学习器亦称“基学习器” (base learner)
- 异质(heterogeneous)集成：个体学习器由不同的学习算法生成
  不存在“基学习算法”

Why Ensemble?

集成的泛化性能通常显著优于单个学习器的泛化性能

一个观察：

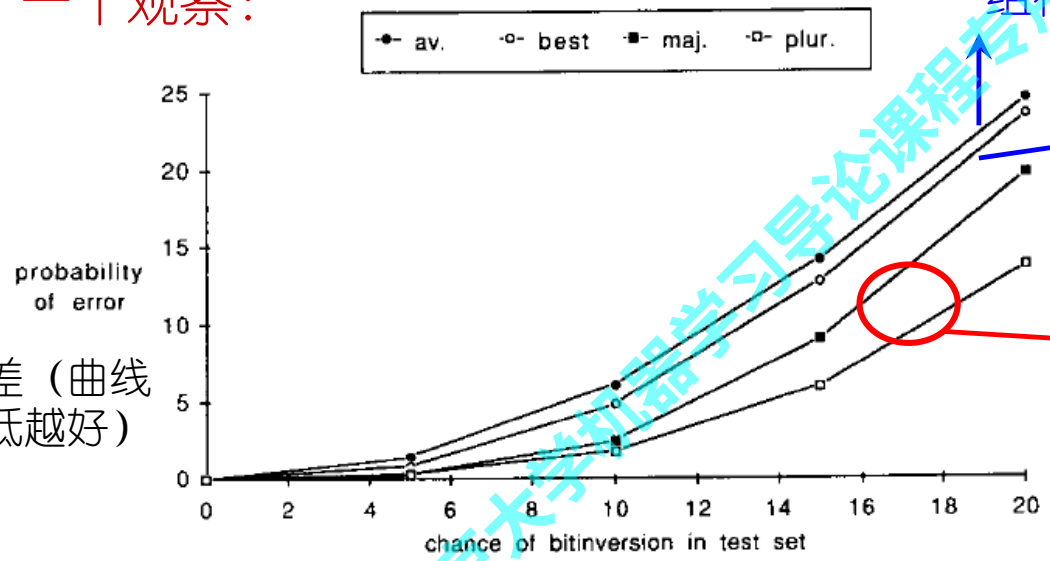


Fig. 4. Performance versus noise level in the test set is shown for individual and for consensus decisions. Data displayed shows the average and the best network, as well as collective decisions using majority and plurality for seven networks trained on individual training sets.

[Hansen & Salamon, TPAMI90]

一组神经网络的平均性能

选择最优神经网络

两种简单的神经网络集成方法

误差（曲线越低越好）

如何得到好的集成？

	测试例1	测试例2	测试例3
h_1	√	√	×
h_2	×	√	√
h_3	√	×	√
集成	√	√	√

(a) 集成提升性能

	测试例1	测试例2	测试例3
h_1	√	√	×
h_2	√	√	×
h_3	√	√	×
集成	√	√	×

(b) 集成不起作用

	测试例1	测试例2	测试例3
h_1	√	×	×
h_2	×	√	×
h_3	×	×	√
集成	×	×	×

(c) 集成起负作用

令个体学习器“好而不同”

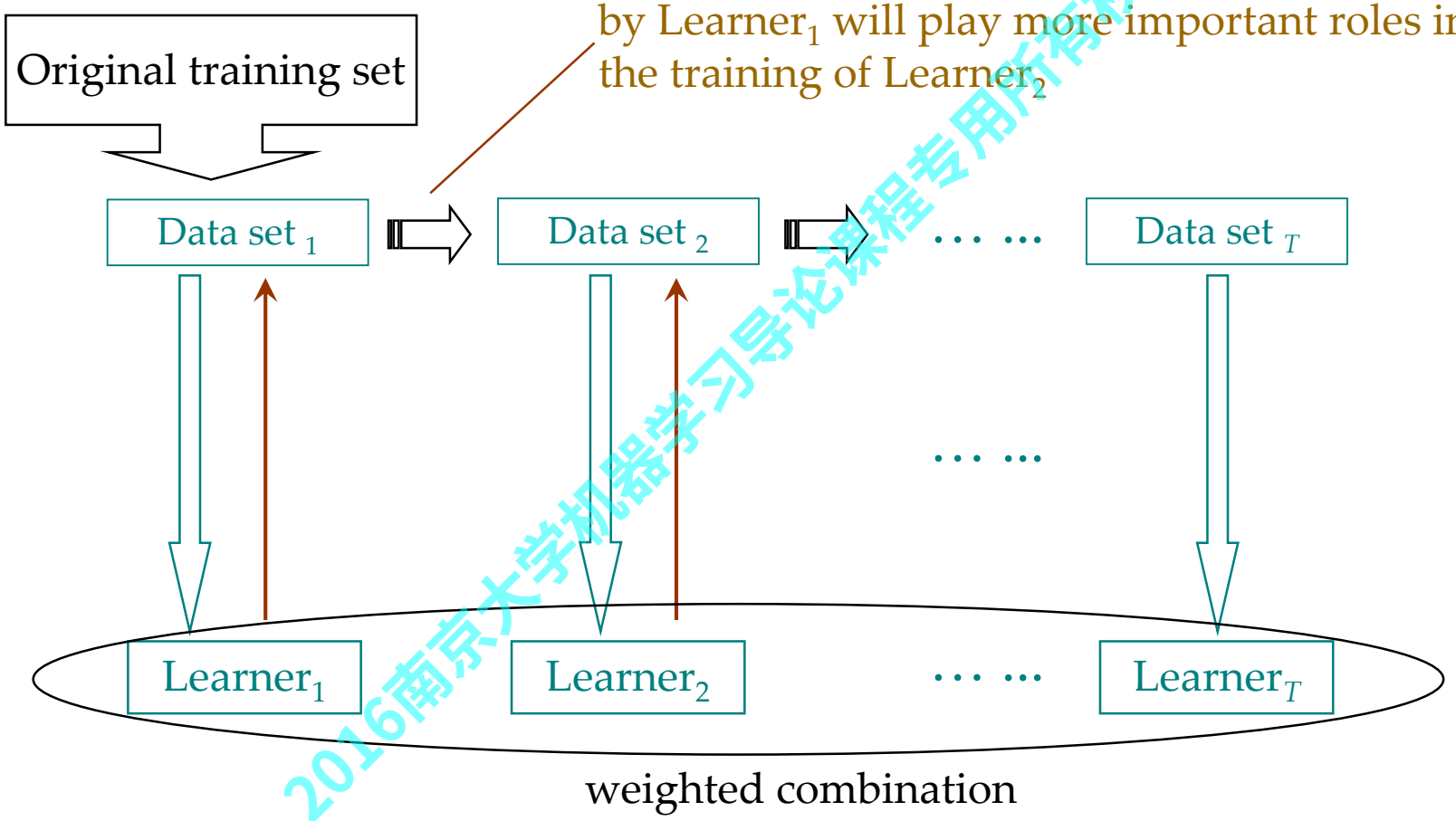
现实各类机器学习、数据挖掘应用中，广泛使用集成学习技术

想获胜，用集成

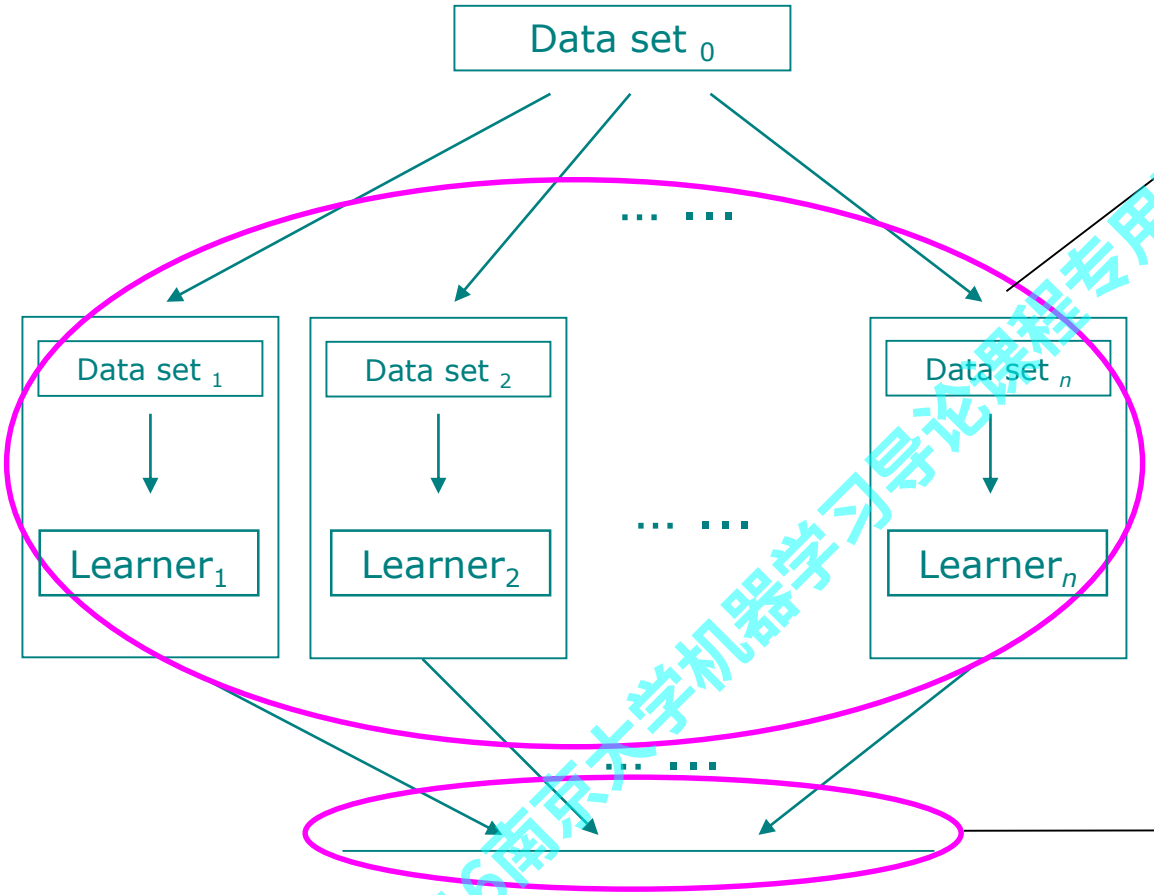
很多成功的集成学习方法

- 序列化方法
  - **AdaBoost** [Freund & Schapire, JCSS97]
  - GradientBoost [Friedman, AnnStat01]
  - LPBoost [Demiriz, Bennett, Shawe-Taylor, MLJ06]
  - … …
- 并行化方法
  - **Bagging** [Breiman, MLJ96]
  - Random Forest [Breiman, MLJ01]
  - Random Subspace [Ho, TPAMI98]
  - … …

Boosting: A flowchart illustration



Bagging



bootstrap a set of learners

generate many data sets from the original data set through bootstrap sampling (random sampling with replacement), then train an individual learner per data set

voting for classification

the output is the class label receiving the most number of votes

averaging for regression

the output is the average output of the individual learners

学习器结合

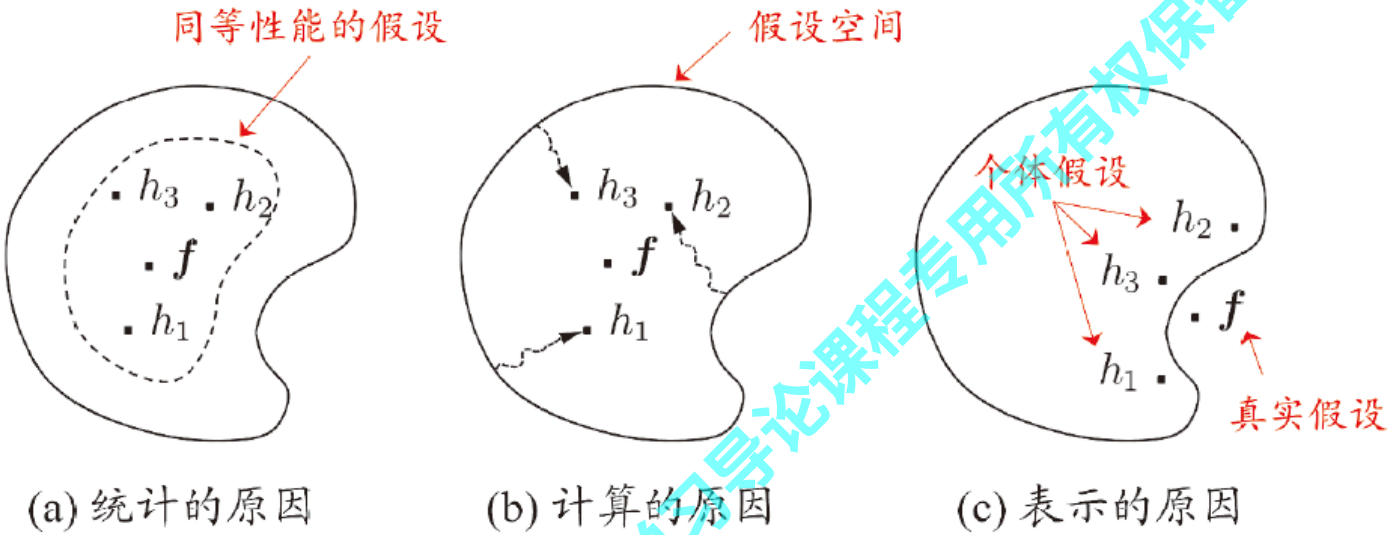


图 8.8 学习器结合可能从三个方面带来好处 [Dietterich, 2000]

常用结合方法：

- 投票法
  - 绝对多数投票法
  - 相对多数投票法
  - 加权投票法
- 平均法
  - 简单平均法
  - 加权平均法
- 学习法

Stacking

输入: 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
初级学习算法 $\mathfrak{L}_1, \mathfrak{L}_2, \ldots, \mathfrak{L}_T$;
次级学习算法 $\mathfrak{L}$.

过程:

```
1: for t = 1, 2, ..., T do
2:     h_t = L_t(D);
3: end for
4: D' = ∅;
5: for i = 1, 2, ..., m do
6:     for t = 1, 2, ..., T do
7:         z_it = h_t(x_i);
8:     end for
9:     D' = D' ∪ ((z_i1, z_i2, ..., z_iT), y_i);
10: end for
11: h' = L(D');
```

使用初级学习算法 $\mathfrak{L}_t$ 产生初级学习器 h_t.

生成次级训练集.

输出: $H(\boldsymbol{x}) = h'(h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_T(\boldsymbol{x}))$

图 8.9 Stacking 算法

多样性

“多样性” (diversity) 是集成学习的关键

误差-分歧分解 (error-ambiguity decomposition):

$$E = \overline{E} - \overline{A}$$

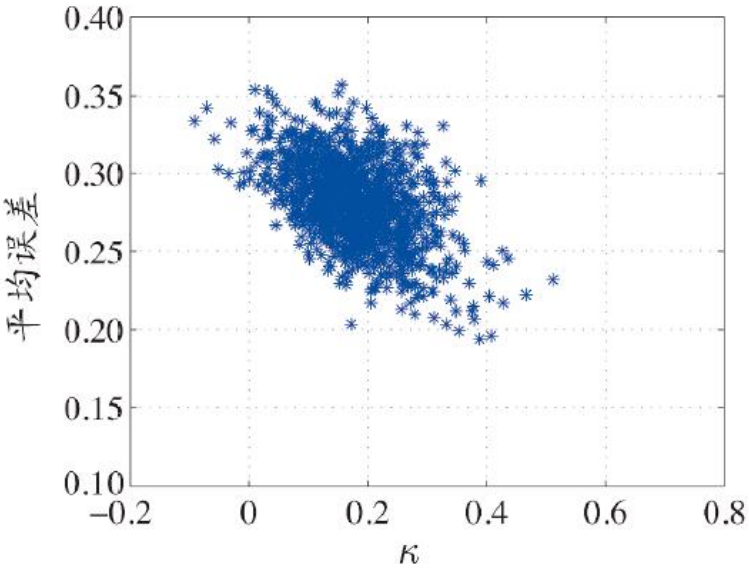
多样性度量

一般通过两分类器的预测结果列联表定义

	$h_i = +1$	$h_i = -1$
$h_j = +1$	a	c
$h_j = -1$	b	d

- 不合度量 (disagreement measure)
- 相关系数 (correlation coefficient)
- Q-统计量 (Q-statistic)
- κ-统计量 (κ-statistic)
- ……

κ-误差图



每一对分类器作为图中的一个点

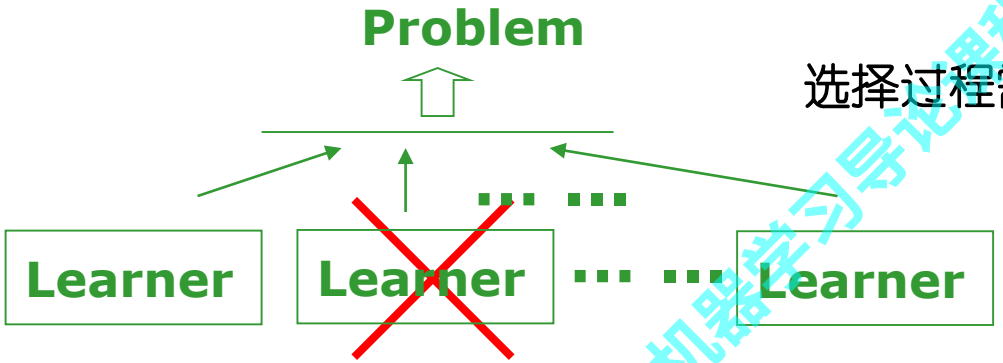
多样性增强常用策略

- 数据样本扰动
  - 例如 Adaboost 使用 重要性采样、Bagging 使用自助采样
  - 注意：对“不稳定基学习器”（如决策树、神经网络等）很有效
    不适用于“稳定基学习器”（如线性分类器、SVM、朴素贝叶斯等）
- 输入属性扰动
  - 例如 随机子空间 (Random Subspace)
- 输出表示扰动
  - 例如 输出标记随机翻转、分类转回归、ECOC
- 算法参数扰动

“越多越好”？

选择性集成 (selective ensemble)：

给定一组个体学习器，从中选择一部分来构建集成，经常会比使用所有个体学习器更好 (更小的存储/时间开销，更强的泛化性能)



选择过程需考虑个体 性能 与 多样性/互补性

仅选出“精度最高的”通常不好！

集成修剪 (ensemble pruning)
[Margineantu & Dietterich, ICML’97]
较早出现，针对序列型集成
减小集成规模、降低泛化性能

选择性集成 [Zhou, et al, AIJ 02] 稍晚，针对并行型集成，MCBTA (Many could be better than all)定理
减小集成规模、增强泛化性能

目前“集成修剪”与“选择性集成”基本被视为同义词

更多关于集成学习的内容，可参考：

Z.-H. Zhou.
Ensemble Methods: Foundations and Algorithms, Boca Raton, FL: Chapman & Hall/CRC, Jun. 2012. (ISBN 978-1-439-830031)

前往……