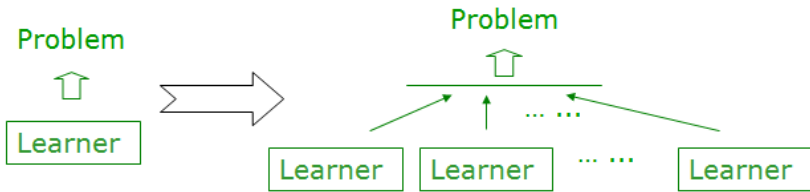


八、集成学习

集成学习

Ensemble Learning (集成学习):

Using multiple learners to solve the problem



Demonstrated great performance in real practice

- ❑ KDDCup'07: 1st place for "... Decision Forests and ..."
- ❑ KDDCup'08: 1st place of Challenge1 for a method using Bagging; 1st place of Challenge2 for "... Using an Ensemble Method "
- ❑ KDDCup'09: 1st place of Fast Track for "Ensemble ... "; 2nd place of Fast Track for "... bagging ... boosting tree models ..."; 1st place of Slow Track for "Boosting ... "; 2nd place of Slow Track for "Stochastic Gradient Boosting"
- ❑ KDDCup'10: 1st place for "... Classifier ensembling"; 2nd place for "... Gradient Boosting machines ... "
- ❑ KDDCup'11: 1st place of Track 1 for "A Linear Ensemble ... "; 2nd place of Track 1 for "Collaborative filtering Ensemble", 1st place of Track 2 for "Ensemble ..."; 2nd place of Track 2 for "Linear combination of ..."

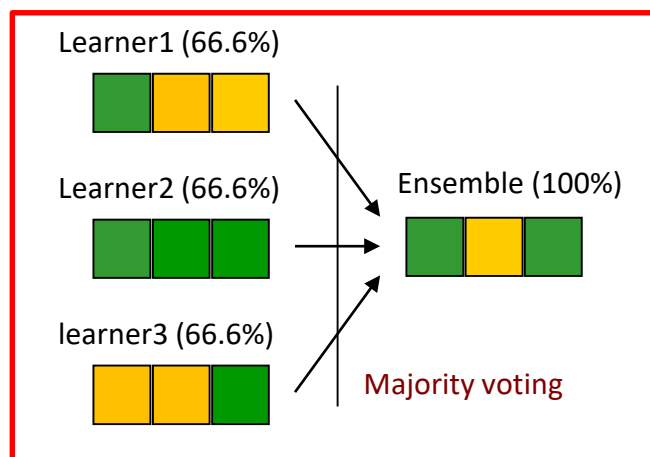
- ❑ KDDCup'12: 1st place of Track 1 for "Combining... Additive Forest..."; 1st place of Track 2 for "A Two-stage Ensemble of..."
- ❑ KDDCup'13: 1st place of Track 1 for "Weighted Average Ensemble"; 2nd place of Track 1 for "Gradient Boosting Machine"; 1st place of Track 2 for "Ensemble the Predictions"
- ❑ KDDCup'14: 1st place for "ensemble of GBM, ExtraTrees, Random Forest..." and "the weighted average"; 2nd place for "use both R and Python GBMs"; 3rd place for "gradient boosting machines... random forests" and "the weighted average of..."
- ❑ KDDCup'15: 1st place for "Three-Stage Ensemble and Feature Engineering for MOOC Dropout Prediction"
- ❑ KDDCup'16: 1st place for "Gradient Boosting Decision Tree"; 2nd place for "Ensemble of Different Models for Final Prediction"
- ❑ KDDCup'17: 1st and 2nd place of Task 1 for "XGBoost"; 1st place of Task 2 for "XGBoost", 2nd place of Task 2 for "Weighted Average of Multiple Models"
- ❑ KDDCup'18: 1st place for "Gradient Boosting"; 2nd place for "Two-stage stacking"; 3rd place for "Weighted Average of Multiple Models"

During the past decade, almost all winners of KDDCup, Netflix competition, Kaggle competitions, etc., utilized ensemble techniques in their solutions

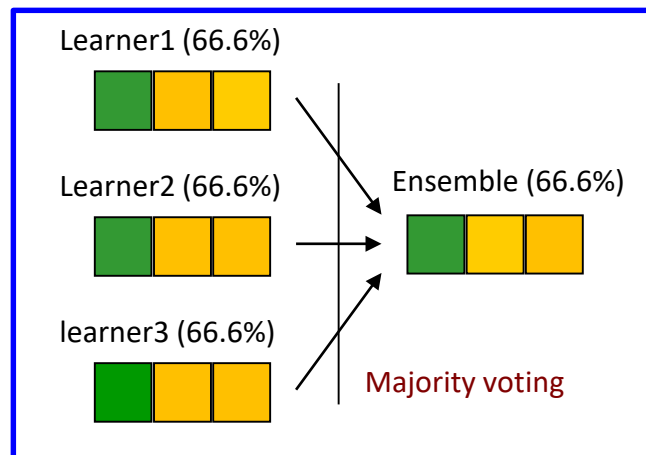
To win? Ensemble !

如何得到好的集成？

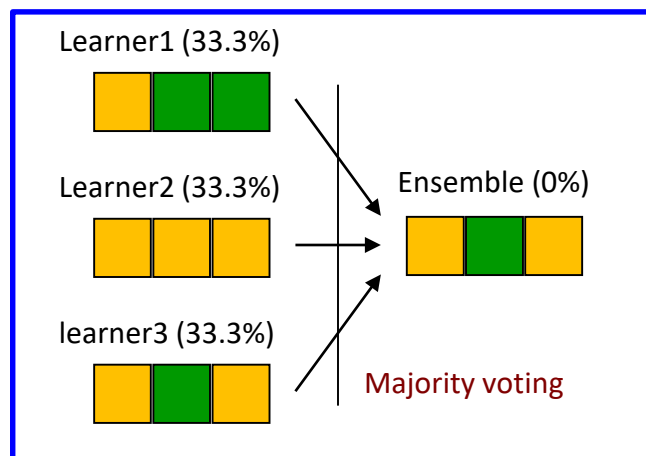
Some intuitions:



Ensemble really helps



Individuals must be different

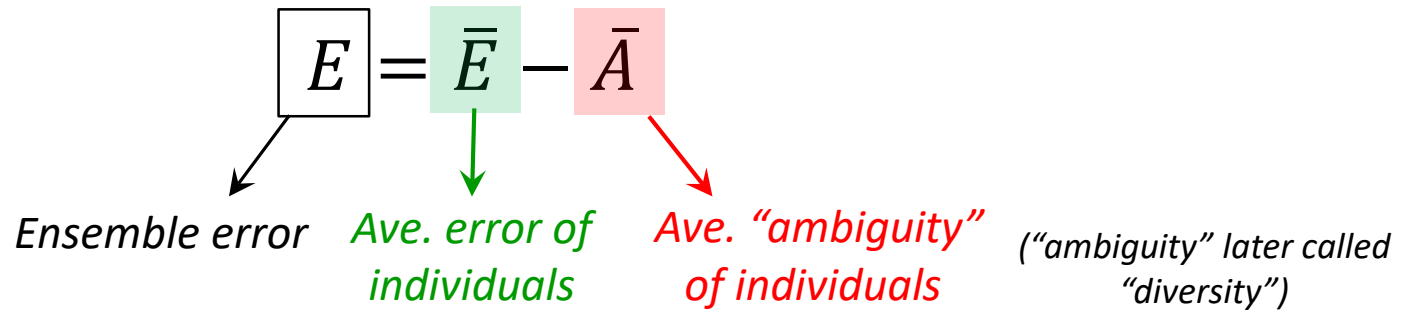


Individuals must be not-bad

令个体学习器 “好而不同”

“多样性” (diversity) 是关键

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

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


The diagram illustrates the error-ambiguity decomposition. It shows the equation $E = \bar{E} - \bar{A}$ where E is in a black box, \bar{E} is in a green box, and \bar{A} is in a red box. Arrows point from each term to its description: a black arrow from E to "Ensemble error", a green arrow from \bar{E} to "Ave. error of individuals", and a red arrow from \bar{A} to "Ave. 'ambiguity' of individuals". A note in parentheses next to the red arrow states: ("ambiguity" later called "diversity").

Ensemble error Ave. error of individuals Ave. "ambiguity" of individuals ("ambiguity" later called "diversity")

The more **accurate** and **diverse** the individual learners,
the better the ensemble

However,

- the “ambiguity” does not have an operable definition
- The error-ambiguity decomposition is derivable only for regression setting with squared loss

很多成功的集成学习方法

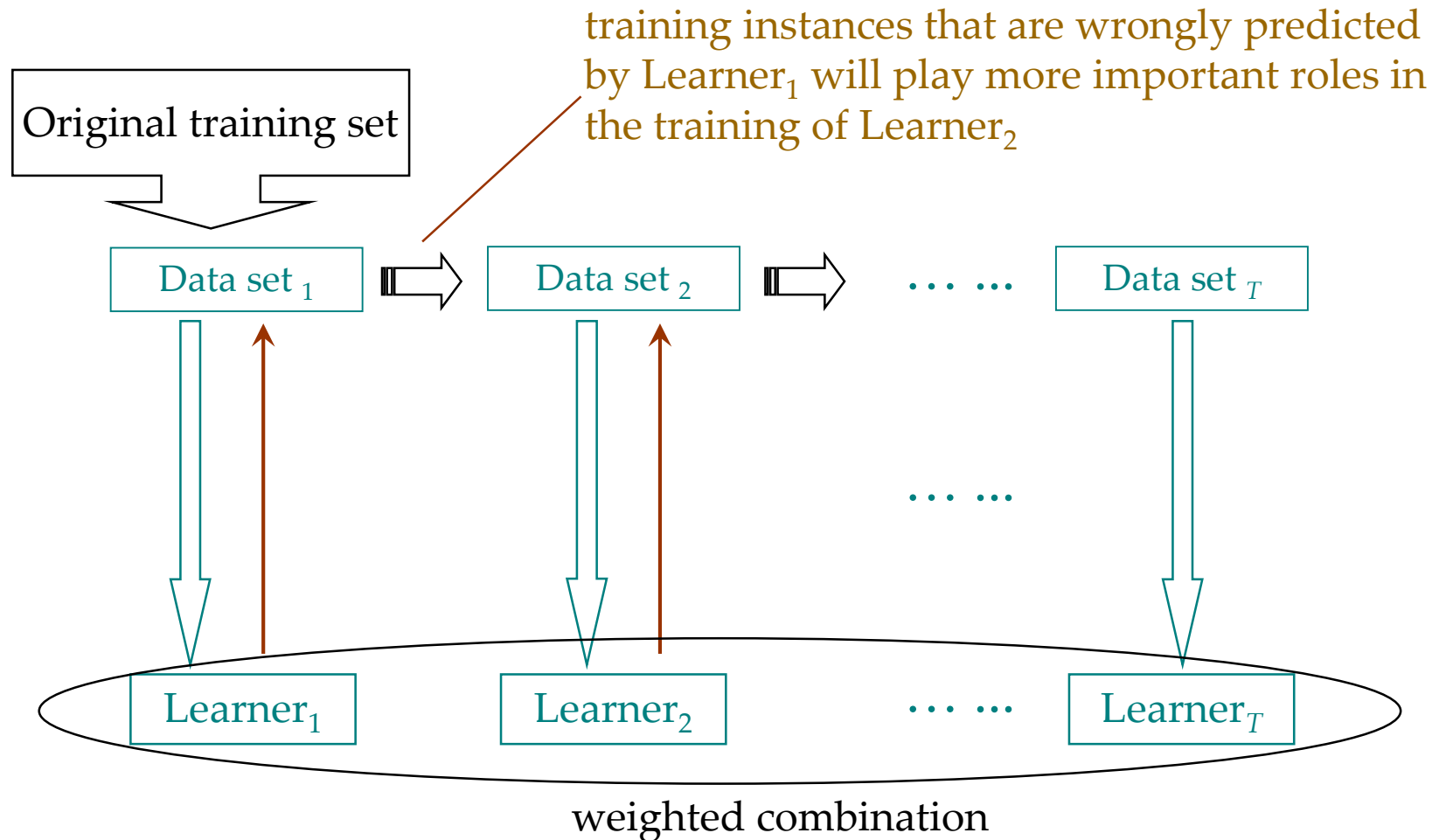
■ 序列化方法

- **AdaBoost** [Freund & Schapire, JCSS97]
- GradientBoost [Friedman, AnnStat01]
- LPBoost [Demiriz, Bennett, Shawe-Taylor, MLJ06]
-

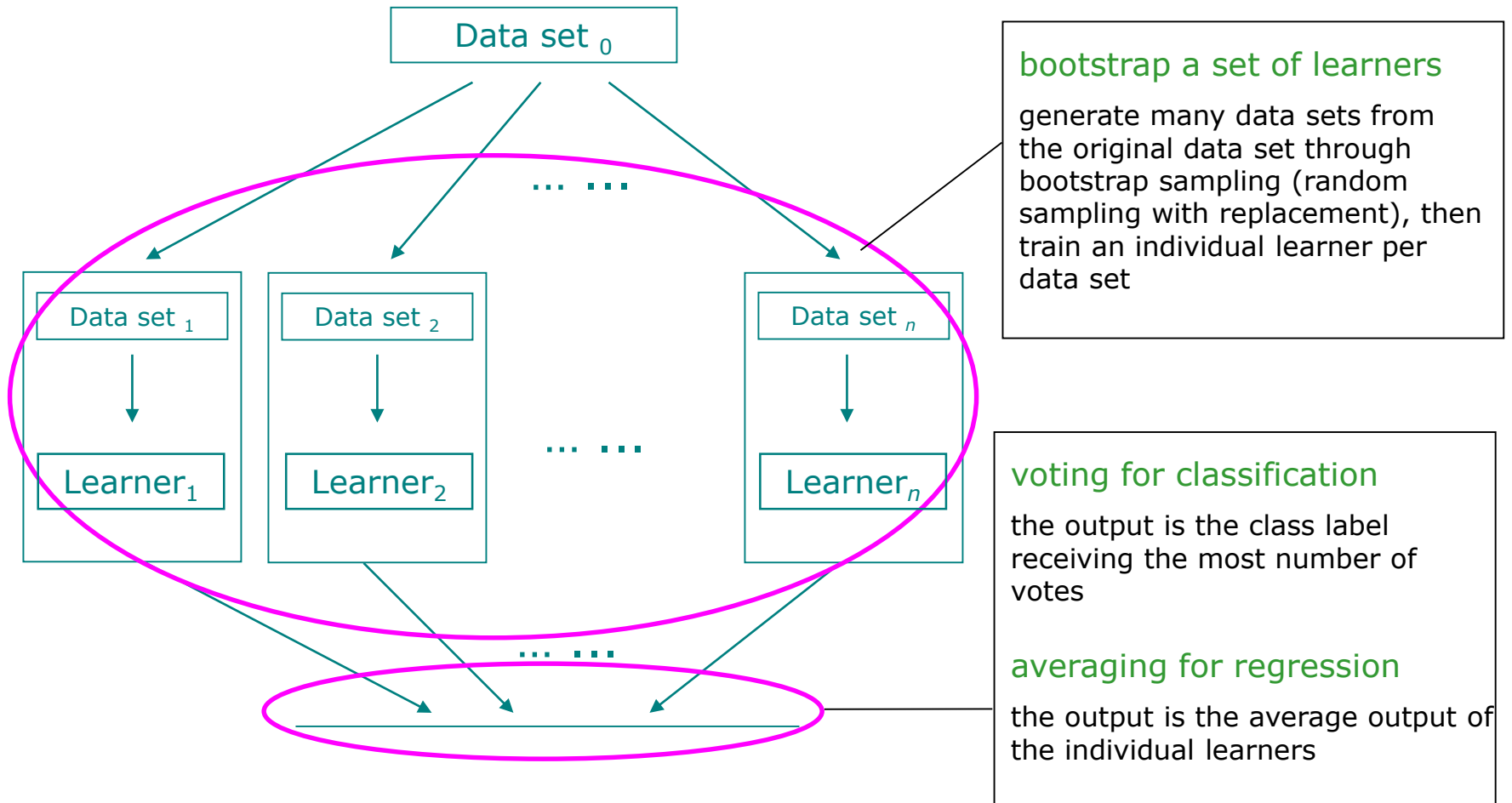
■ 并行化方法

- **Bagging** [Breiman, MLJ96]
- Random Forest [Breiman, MLJ01]
- Random Subspace [Ho, TPAMI98]
-

Boosting: A flowchart illustration



Bagging



学习器结合

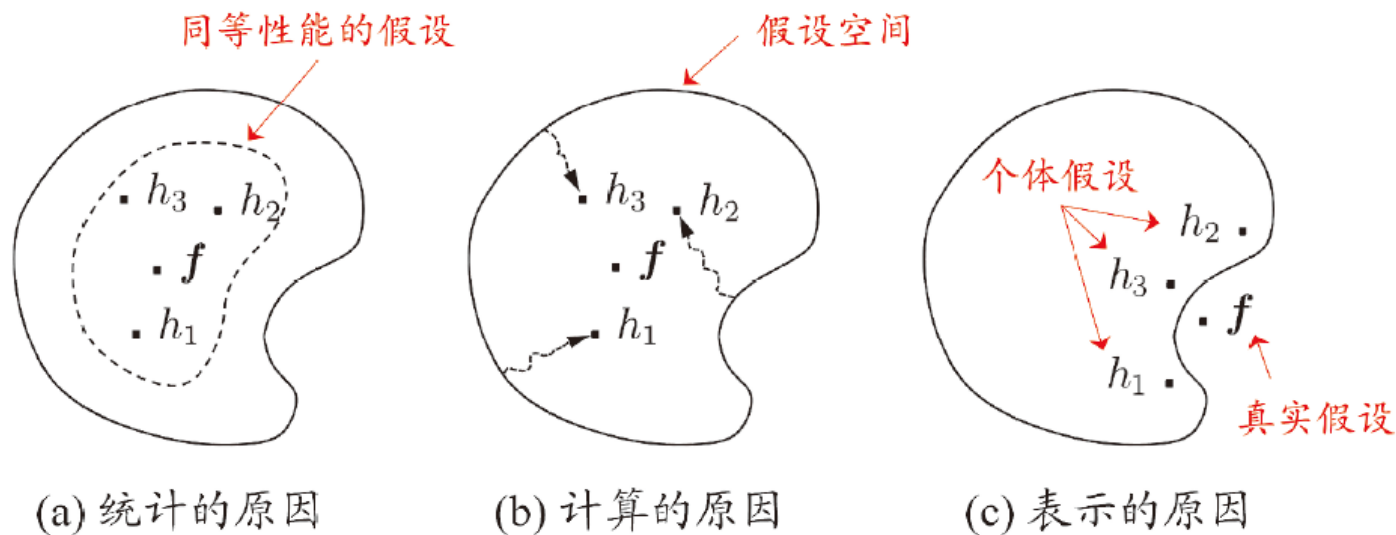


图 8.8 学习器结合可能从三个方面带来好处 [Dietterich, 2000]

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 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$;
初级学习算法 $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_T$;
次级学习算法 \mathcal{L} .

过程:

```
1: for  $t = 1, 2, \dots, T$  do  
2:    $h_t = \mathcal{L}_t(D)$ ;  
3: end for
```

使用初级学习算法 \mathcal{L}_t
产生初级学习器 h_t .

```
4:  $D' = \emptyset$ ;
```

```
5: for  $i = 1, 2, \dots, m$  do  
6:   for  $t = 1, 2, \dots, T$  do  
7:      $z_{it} = h_t(\mathbf{x}_i)$ ;  
8:   end for
```

生成次级训练集.

```
9:    $D' = D' \cup ((z_{i1}, z_{i2}, \dots, z_{iT}), y_i)$ ;  
10: end for
```

```
11:  $h' = \mathcal{L}(D')$ ;
```

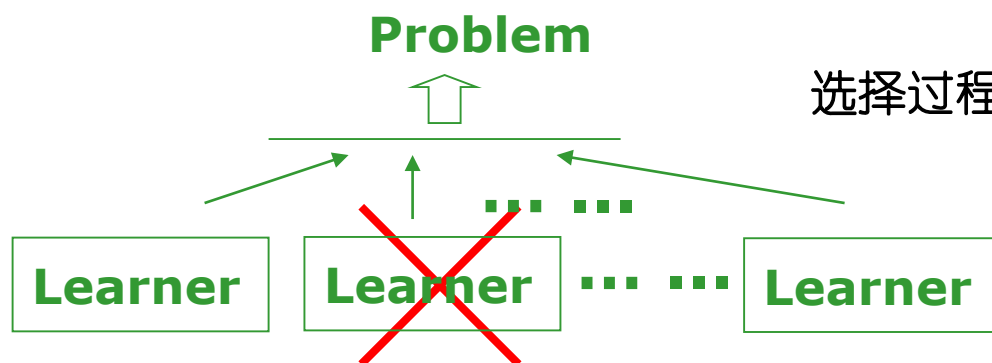
输出: $H(\mathbf{x}) = h'(h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_T(\mathbf{x}))$

图 8.9 Stacking 算法

“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

多样性

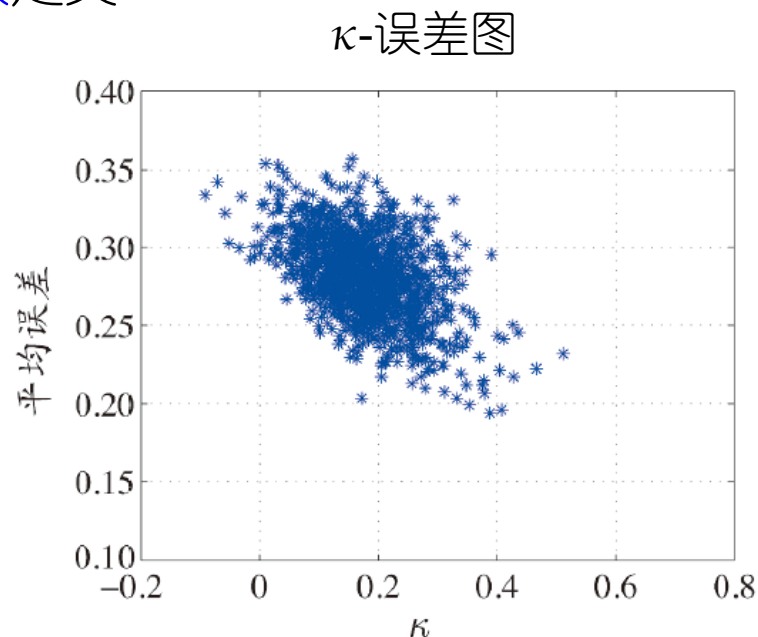
“多样性” (diversity) 是集成学习的关键

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)
-



每一对分类器作为图中的一个点

研究者提出了很多 Diversity measure

Diversity

c represents the total ad. f . The total sum of equal to n .

Table 2 [5] lists definitions of 76 binary similarity and distance measures used over the last century where S and D are similarity and distance measures, respectively.

Table 2 Definitions of Measures for binary data

$S_{JACCARD} = \frac{a}{a+b+c}$ (1)

$S_{SIM} = \frac{2a}{2a+b+c}$ (2)

$S_{CHANCE} = \frac{2a}{2a+b+c}$ (3)

$S_{SP-JACCARD} = \frac{3a}{3a+b+c}$ (4)

$S_{SIMPLE} = \frac{2a}{(a+b)+(a+c)}$ (5)

$S_{SOKAL-MICHENER} = \frac{a}{a+2b+2c}$

$S_{SOKAL-MICHENER} = \frac{a+d}{a+b+c+d}$

$S_{SOKAL-MICHENER} = \frac{2(a+d)}{2a+b+c+2d}$

$S_{HAMMING} = \frac{a+d}{a+2(b+c)+d}$

$S_{FAITH} = \frac{a+0.5d}{a+b+c+d}$

$S_{CHANCE} = \frac{a+d}{a+0.5(b+c)+d}$

$S_{SIMPLE} = a$

$S_{SIMPLE} = a+d$

$S_{SOKAL-MICHENER} = \frac{a}{a+b+c+d}$

$D_{HAMMING} = b+c$

$D_{SIMPLE} = \sqrt{b+c}$

$D_{SQUARED} = \sqrt{(b+c)^2}$

$D_{CHANCE} = (b+c)^{\frac{1}{2}}$

$D_{SIMPLE} = b+c$

$D_{SOKAL-MICHENER} = \frac{b+c}{a+b+c+d}$ (20)

$D_{CHANCE} = b+c$ (21)

$D_{SIMPLE} = (b+c)^{\frac{1}{2}}$ (22)

$D_{SIM} = \frac{(b+c)}{4(a+b+c+d)}$ (23)

$D_{CHANCE} = \frac{(b+c)^2}{(a+b+c+d)^2}$ (24)

$D_{SIMPLE} = \frac{n(b+c)-(b-c)^2}{(a+b+c+d)^2}$ (25)

$D_{CHANCE} = \frac{4bc}{(a+b+c+d)^2}$ (26)

$D_{SOKAL-MICHENER} = \frac{b+c}{(2a+b+c)}$ (27)

$D_{SIMPLE} = \frac{b+c}{(2a+b+c)}$ (28)

$D_{CHANCE} = 2 \cdot \frac{1}{\sqrt{1+\frac{a}{b+c}}}$ (29)

$D_{SIMPLE} = \sqrt{1 - \frac{a}{\sqrt{(a+b)(a+c)}}}$ (30)

$S_{SIMPLE} = \frac{a}{\sqrt{(a+b)(a+c)}}$ (31)

$S_{EYRAUD} = \frac{n^2 (na - (a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)}$ (74)

$S_{TARANTULA} = \frac{a}{(a+b)} = \frac{a(c+d)}{c(a+b)}$

$S_{SIMPLE} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} = \frac{a(c+d)}{c(a+b)}$ (76)

$S_{SIMPLE} = \frac{na-bc}{\sqrt{n(a+b)(a+c)}}$ (44)

$S_{SIMPLE} = \frac{a}{\min(a+b, a+c)}$ (45)

$S_{SIMPLE} = \frac{a}{\max(a+b, a+c)}$ (46)

$S_{HAMMING} = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{\max(a+b, a+c)}{2}$ (47)

$S_{FAITH} = \frac{na - \min(a+b, a+c) - (a+b)(a+c)}{2a}$ (48)

$S_{CHANCE} = \frac{a}{(a+b)} + \frac{a}{(a+c)} + \frac{d}{(b+d)}$ (49)

$S_{SIMPLE} = \frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$ (50)

$S_{FAITH} = \chi^2 \text{ where } \chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$ (51)

$S_{SIMPLE} = \frac{1}{n} \cdot \frac{1}{\sqrt{1+\frac{a}{b+c}}}$ (52)

$S_{FAITH} = \left(\frac{P}{n+P} \right)^2 \text{ where } P = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$ (53)

$S_{HAMMING} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$ (54)

$S_{HAMMING} = \cos\left(\frac{\pi(ad-bc)}{n(a+b)(a+c)(b+d)(c+d)}\right)$ (55)

$S_{SIMPLE} = \frac{ad}{(a+b)(a+c)+d}$ (57)

$S_{SIMPLE} = \frac{ad}{(a+b)(a+c)+d}$ (58)

$S_{SIMPLE} = \frac{\sigma - \sigma'}{2n}$ (70)

$S_{SIMPLE} = \frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$ (71)

$S_{SIMPLE} = \frac{\sqrt{ad} + a - (b+c)}{\sqrt{ad} + a + b + c}$ (72)

$S_{SIMPLE} = \frac{ab+bc}{ab+2bc+cd}$ (73)

$S_{SIMPLE} = \frac{n(na - (a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)}$ (74)

$S_{SIMPLE} = \frac{a}{(a+b)} + \frac{d(c+d)}{c(a+b)}$ (75)

$S_{SIMPLE} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} = \frac{a(c+d)}{c(a+b)}$ (76)

The inclusion or exclusion of negative matches, d in the binary similarity measures have been an ongoing issue [9, 12, 15, 16, 17, 18, 26, 27]. The Sokal & Michener, the Roger & Tanimoto, the Faith, the Ochiai II, the Cole, the Gower, Pearson I, and the Stiles etc. are included in the negative match inclusive measures. The Jaccard, the Tanimoto, the Dice & Sorenson, the Kulczynski I, the Sokal & the Tanimoto, the Sokal & the Michener etc. are included in the negative match exclusive measures. The Sokal & Michener, the Roger & Tanimoto, the Faith, the Ochiai II, the Cole, the Gower, Pearson I, and the Stiles etc. are included in the negative match inclusive measures. The Jaccard, the Tanimoto, the Dice & Sorenson, the Kulczynski I, the Sokal & the Tanimoto, the Sokal & the Michener etc. are included in the negative match exclusive measures. The Sokal & Michener, the Roger & Tanimoto, the Faith, the Ochiai II, the Cole, the Gower, Pearson I, and the Stiles etc. are included in the negative match inclusive measures. The Jaccard, the Tanimoto, the Dice & Sorenson, the Kulczynski I, the Sokal & the Tanimoto, the Sokal & the Michener etc. are included in the negative match exclusive measures.

In cases where the two binary states are not equally important, such as in the measurement of biomass, the positive matches are usually more significant than the negative matches [1, 6, 10, 26]. Faith included the negative match but only gave the half credits while giving the full credits for the positive matches in eqn (10) [11]. In [4], different weights for positive and negative matches were studied. Weighted similarity measures such as weighted hamming distance or $amod$ [4] are not covered in this paper though.

Historically, all the binary measures observed above have had a meaningful performance in their respective fields. The binary similarity coefficient proposed by Peirce, Yule, and Pearson in 1900s contributes to the evolution of the various correlation based binary similarity measures. The Jaccard coefficient proposed at 1901 is still widely used in the various fields such as ecology and biology. The discussion of inclusion or exclusion of negative matches was actively arisen by Sokal & Sneath in during 1960s and by Goodman & Kruskal in 1970s. In Figure 1, the measures are arranged in historical order.

However, ...

- ❑ [Kuncheva & Whitaker, MLJ 2003]: Empirical study shows that there seems **no clear relation** between many diversity measures and the ensemble performance
- ❑ [Tang, Suganthan, Yao, MLJ 2006]: Exploiting many diversity measures explicitly is **ineffective** in constructing consistently stronger ensembles

There is no well-accepted definition/formulation of diversity

“What is diversity” remains the holy grail problem of ensemble learning

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

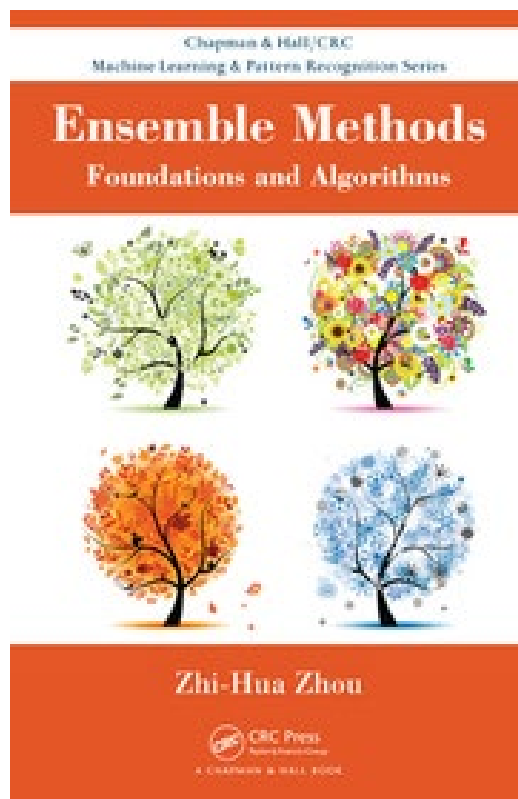
- 例如 **随机子空间** (Random Subspace)

□ 输出表示扰动

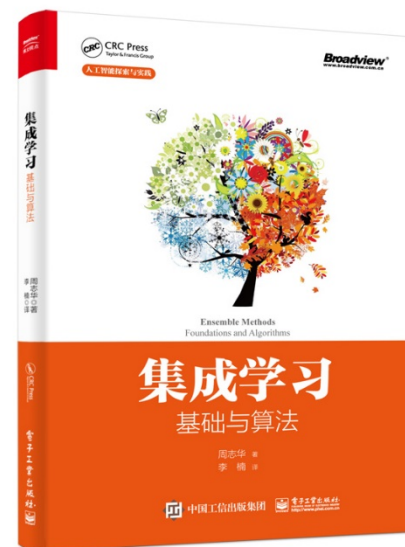
- 例如 输出标记随机翻转、分类转回归、**ECOC**

□ 算法参数扰动

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 978-1-439-83003-1)



前往.....

