Bagging & Boosting

DCS310

Sun Yat-sen University

Outline

- Introduction to Ensembles
- Bagging
- Boosting

Motivation of Ensembles

- It is difficult to learn a *strong classifier* that can always classify instances correctly

- But it is easy to learn a lot of *'weak' classifiers*

  A weak classifier may not perform well on the whole dataset, but may perform well on parts

- If weak classifiers perform well at different parts of input space, it is possible to *obtain a strong classifier by combining these weak classifiers*

- Two questions

  1) How to produce these weak classifiers?

  2) How to combine the weak classifiers?

Two Combining Methods

1) Committee

- Unweighted average or majority vote

2) Weighted average

- Give better classifiers bigger weighting

For example, consider a two-class classification problem {-1, 1}

Two basic classifiers: $\hat{y}_1 = f_1(\boldsymbol{x})$ $\hat{y}_2 = f_2(\boldsymbol{x})$

Final classifiers: $\hat{y}_e = sign\big(\alpha_1 f_1(\boldsymbol{x}) + \alpha_2 f_2(\boldsymbol{x})\big)$

Remark: The weak/basic classifiers could be any kinds, *e.g.* decision trees, SVM, neural networks, logistic regression *etc.*

Ensemble Clustering

① Toward Multi-Diversified Ensemble Clustering of High-Dimensional Data: From Subspaces to Metrics and Beyond, IEEE TCYB, 2021.

② Enhanced Ensemble Clustering via Fast Propagation of Cluster-Wise Similarities, IEEE TSMC-A, 2021.

③ Ultra-Scalable Spectral Clustering and Ensemble Clustering, IEEE TKDE, 2020.

④ Locally Weighted Ensemble Clustering, IEEE TCYB, 2018.

⑤ Robust Ensemble Clustering Using Probability Trajectories, IEEE TKDE, 2016.

Outline

- Introduction to Ensemble
- Bagging
- Boosting

Creating Weak Classifiers

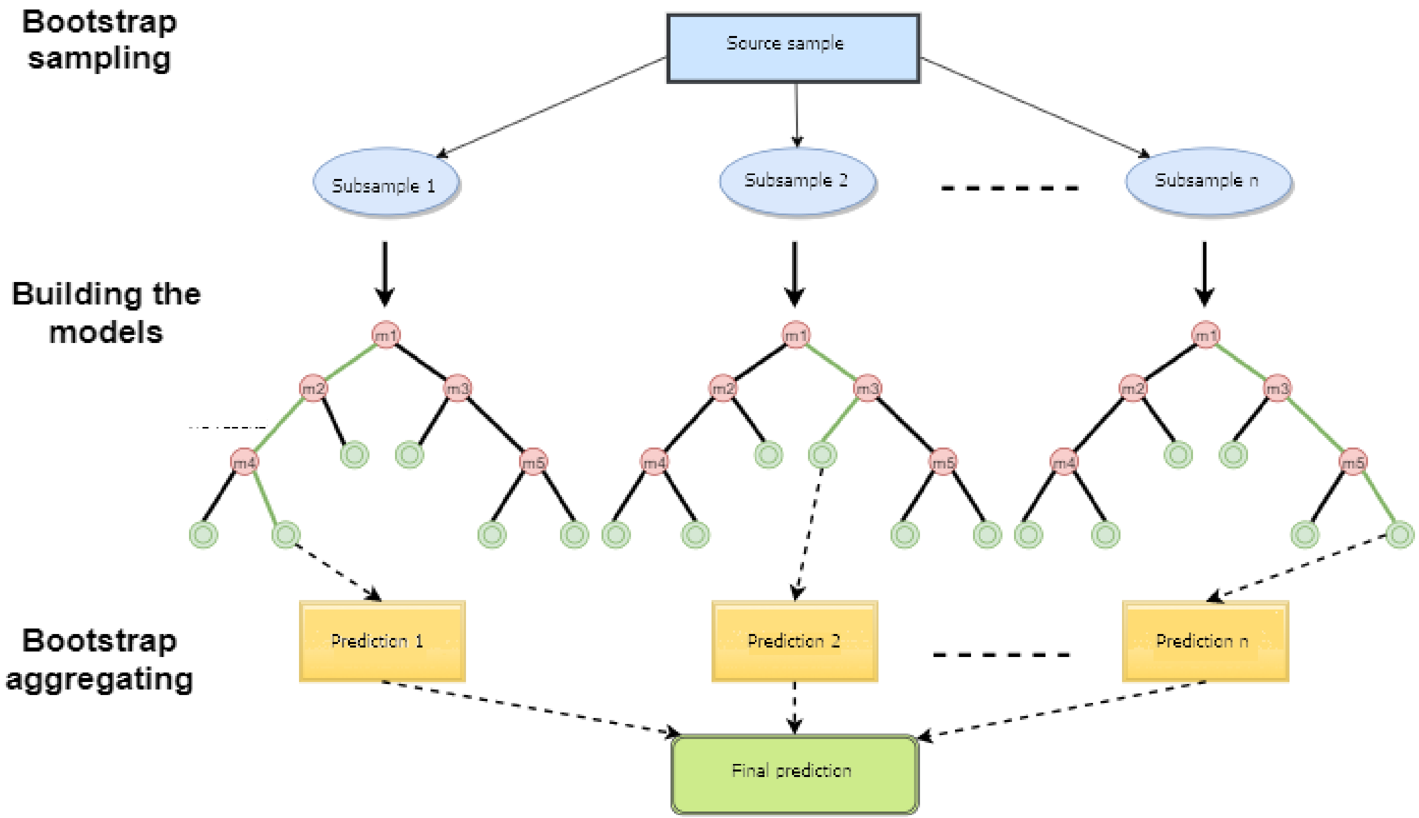
- We don't know how to create classifiers that perform well at parts of the input spaces

- Thus, we instead seek to create classifiers that are as diverse as possible, that is, encouraging their predictions *to be uncorrelated*. For example,

1) Creating subsets of the training dataset by bootstrapping
   - Randomly draw N' samples from the N-sample training dataset *with replacement* 随机抽取K个子集
   - Repeat the above procedure K times, producing the subsets $S_1, S_2, \cdots, S_K$

2) Training a decision tree on each of the subset S_k

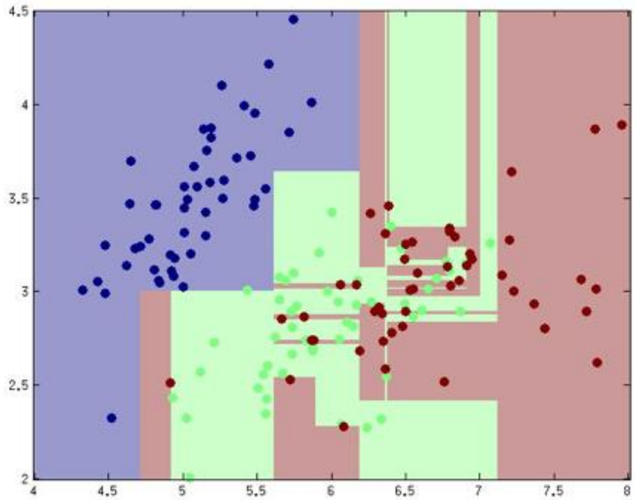
3) Combine K decision trees into one by majority voting



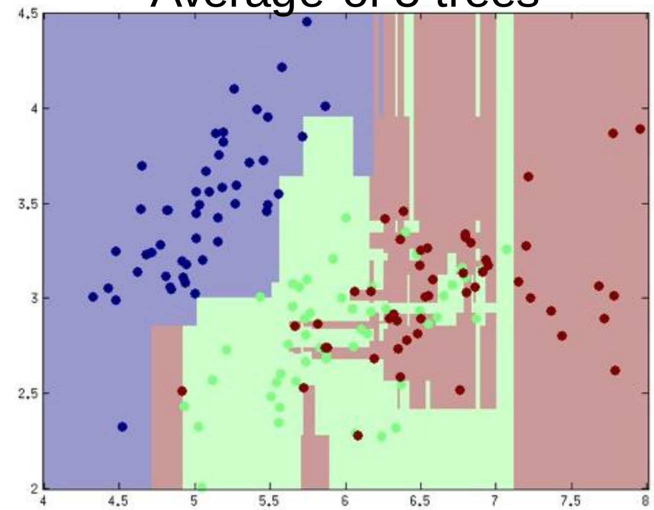
At testing, pass test data through all K classifiers, and the majority voting result is used as the final prediction

Example: Bagged Decision Trees

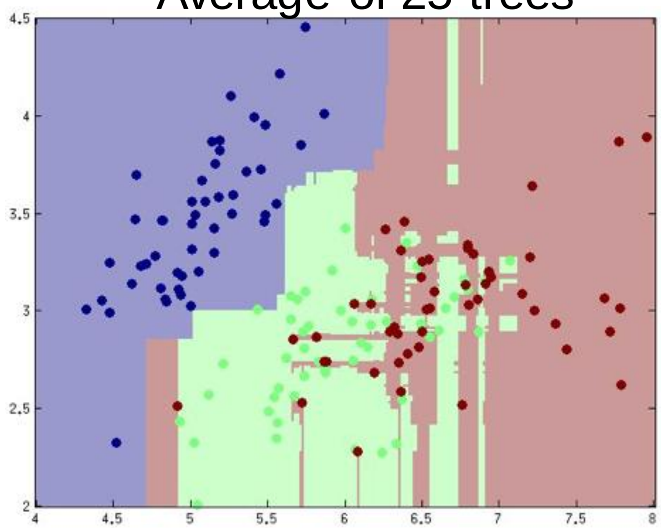
Full dataset



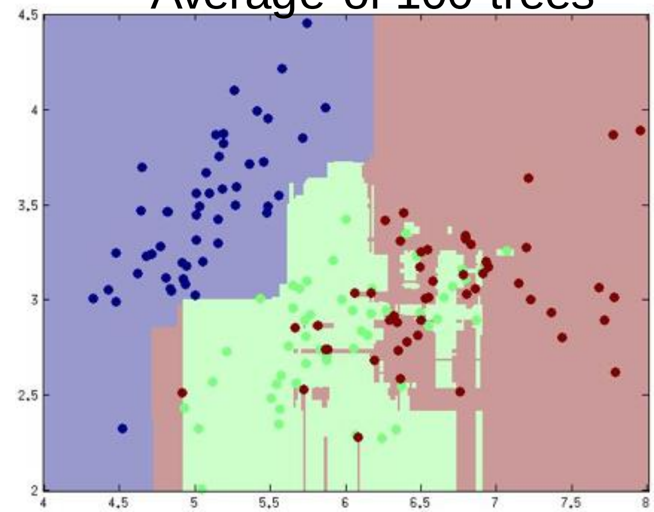
Average of 5 trees



Average of 25 trees



Average of 100 trees



Random Forest

- What happens if the classifiers obtained above are *still strongly correlated* ? 随机选的子集之间可能出现线性相关

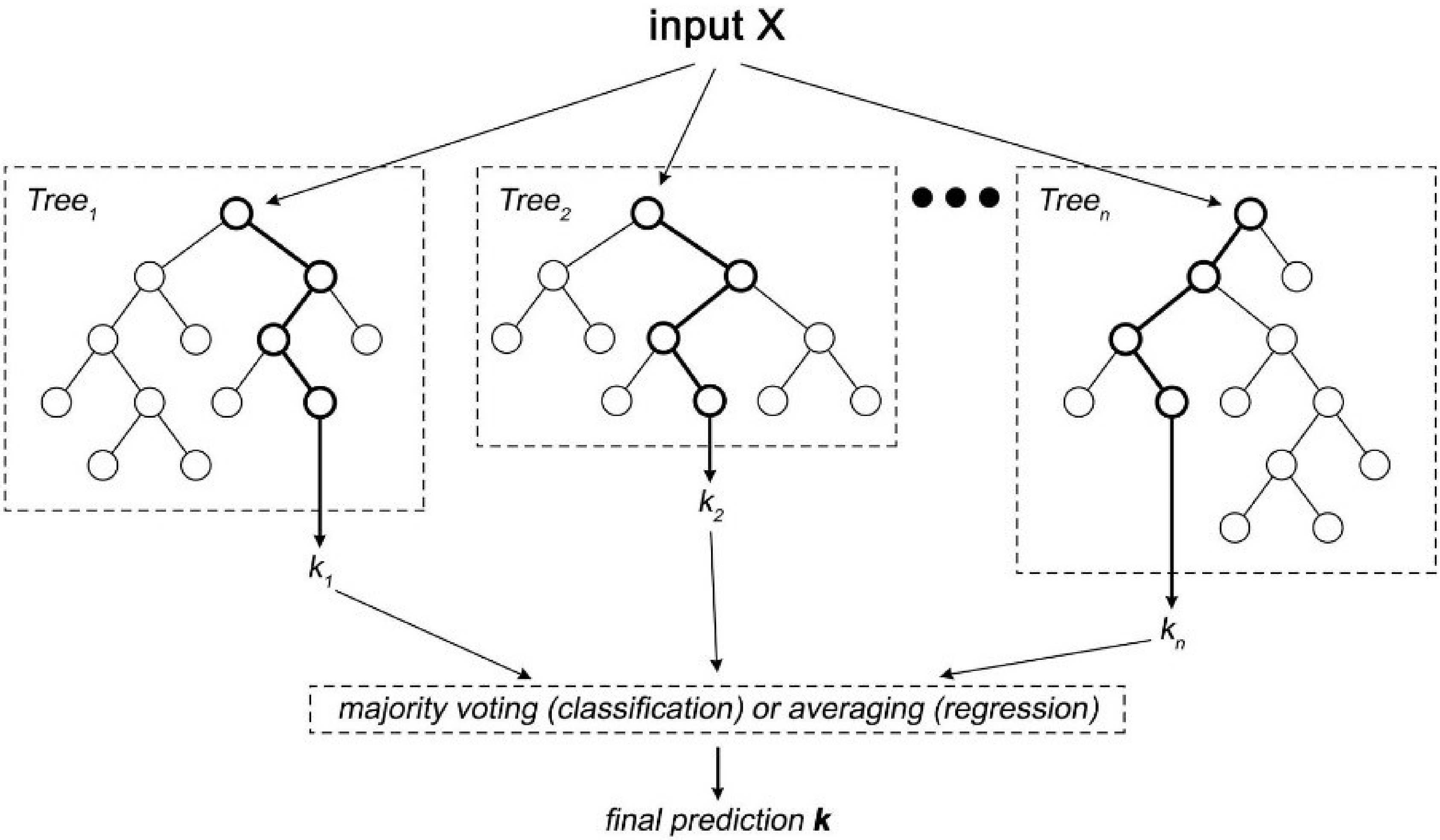
  ⇒ Combining them by majority vote doesn't help much on the performance improvement

- *Solution:* Introducing extra randomness into the learning process of decision trees

> As building the decision trees, only use subset of the randomly selected attributes

随机选择X中的属性
可以防止子集之间线性相关

- Random forest illustration



Outline

- Introduction to Ensemble
- Bagging
- Boosting

Overview of Boosting Methods

- *Weak classifiers*

Repeat the following steps several times

1) Identifying the examples that are incorrectly classified

2) Re-training the classifier by *giving more weighting to the misclassified examples*

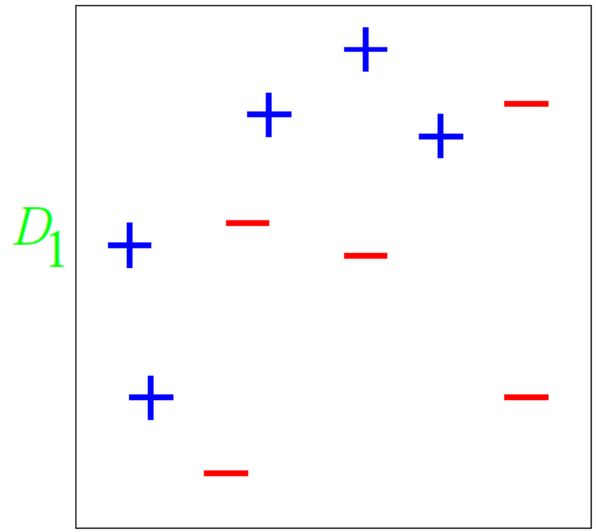
- *Combining*

Combing the prediction results of each classifier by *weighted average*

How to *weight the examples and prediction results* is the key

Adaboost

- Consider a two-class classification problem with 10 training examples

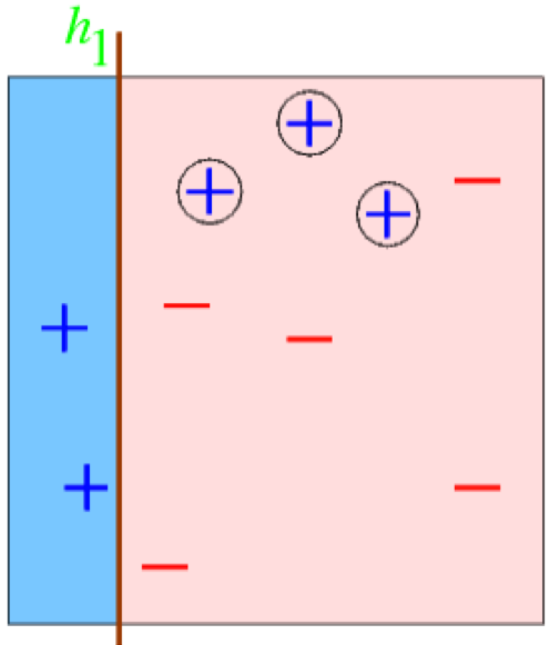


只使用x中一个属性的弱分类器

- For simplicity, only consider the weak classifiers whose decision boundaries are parallel to the axes, that is,

$$\hat{y} = sign(x_1 + b) \text{ or } \hat{y} = sign(x_2 + b)$$

- First iteration

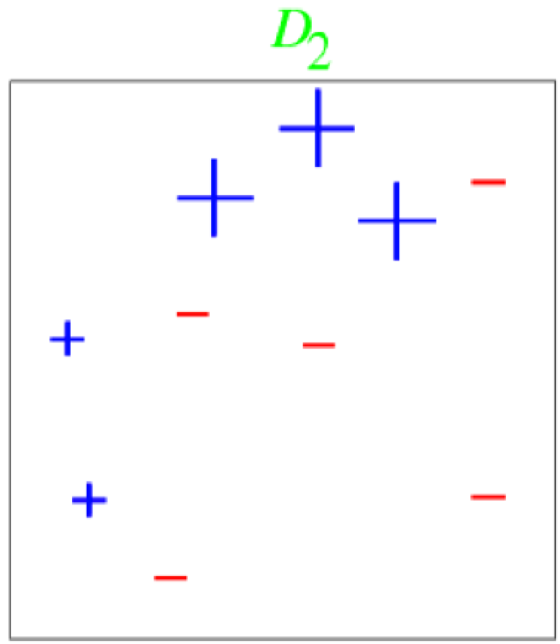


➢ Error rate of the first classifier h_1: $\epsilon_1 = 0.3$

➢ Weighting of the classifier h_1: $\alpha_1 = \frac{1}{2}\ln\left(\frac{1-\epsilon_1}{\epsilon_1}\right) = 0.42$

$$\frac{1-\epsilon_1}{\epsilon_1} = \frac{\text{correct rate}}{\text{error rate}}$$

⇒ The weighting is *positively proportional* to performance of the classifier



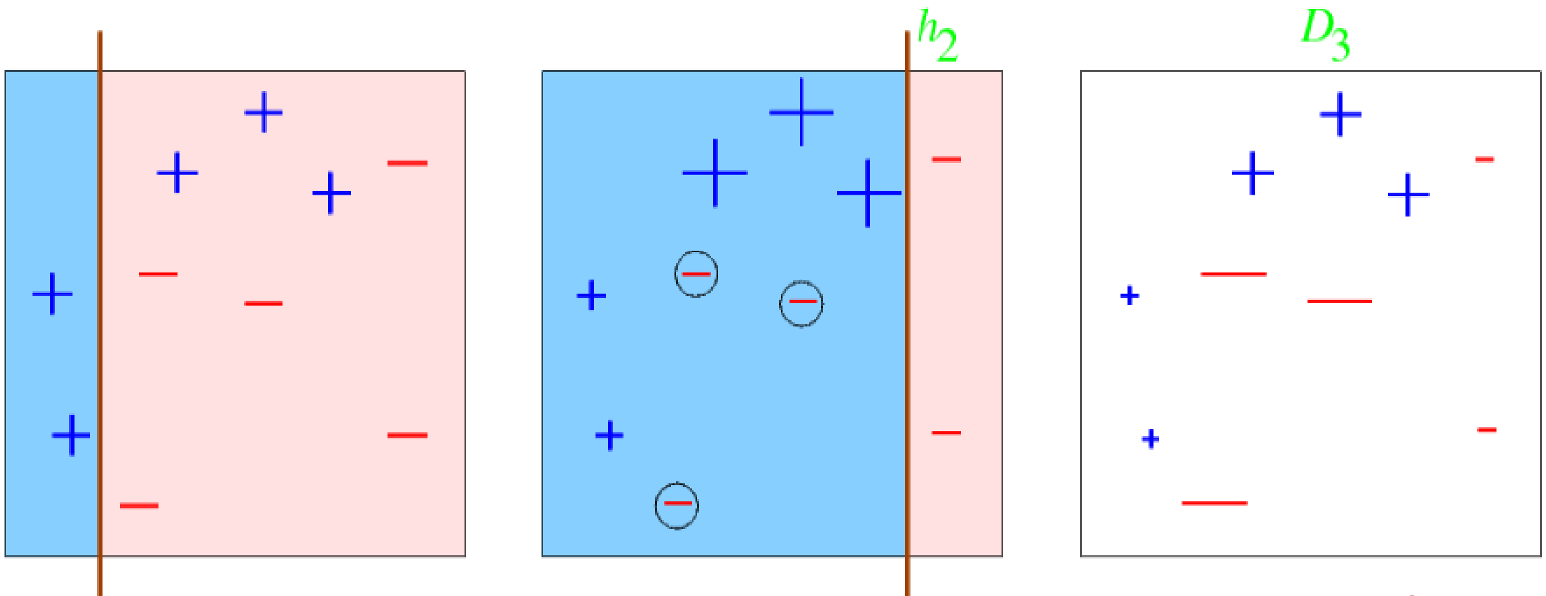
- *Misclassified* examples' weights are amplified by e^{α_1}

对样本进行
权重调整

$$e^{\alpha_1} = \sqrt{\frac{1-\epsilon_1}{\epsilon_1}} = \sqrt{\frac{\text{correct rate}}{\text{error rate}}}$$

- Correctly classified examples' weights are dampened by $e^{-\alpha_1}$

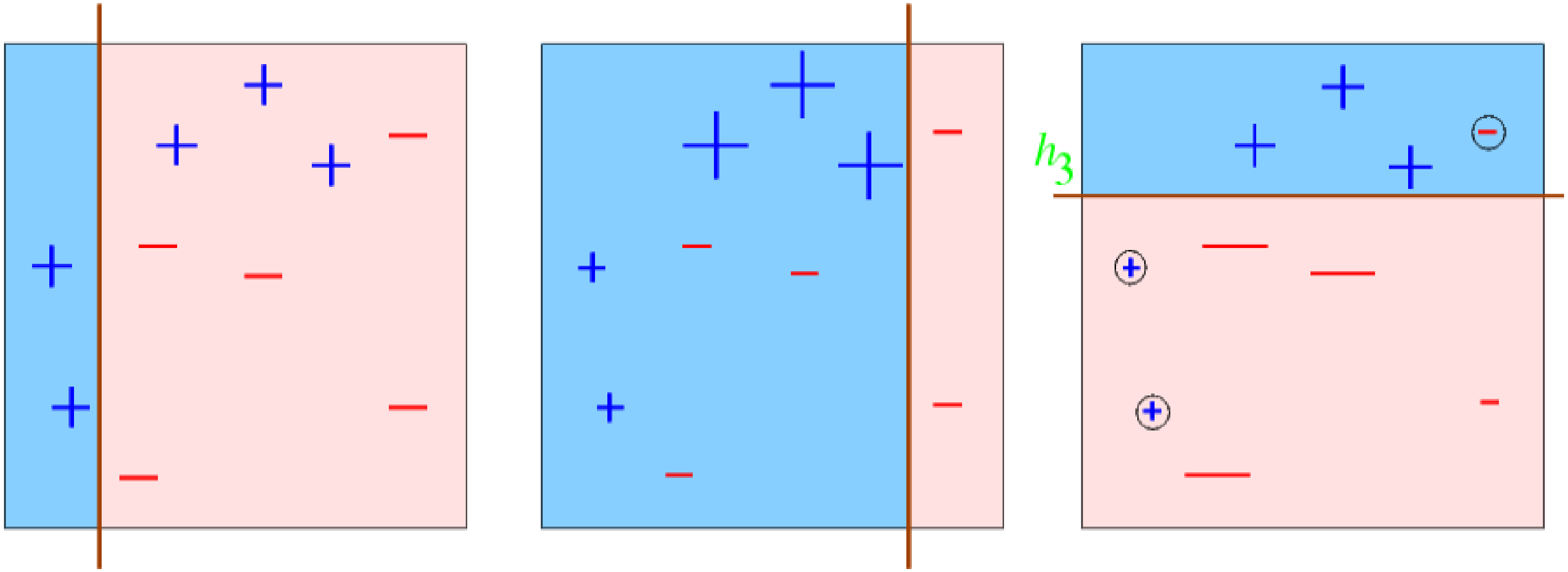
- Second iteration



加权计算的错误率

- Error rate of the second classifier h_2: $\epsilon_2 = 0.21$
- Weighting of the classifier h_2: $\alpha_2 = \frac{1}{2}\ln\left(\frac{1-\epsilon_2}{\epsilon_2}\right) = 0.65$
- Misclassified examples' weights are amplified by e^{α_2}
- Correctly classified examples' weights are dampened by $e^{-\alpha_2}$

- Third iteration



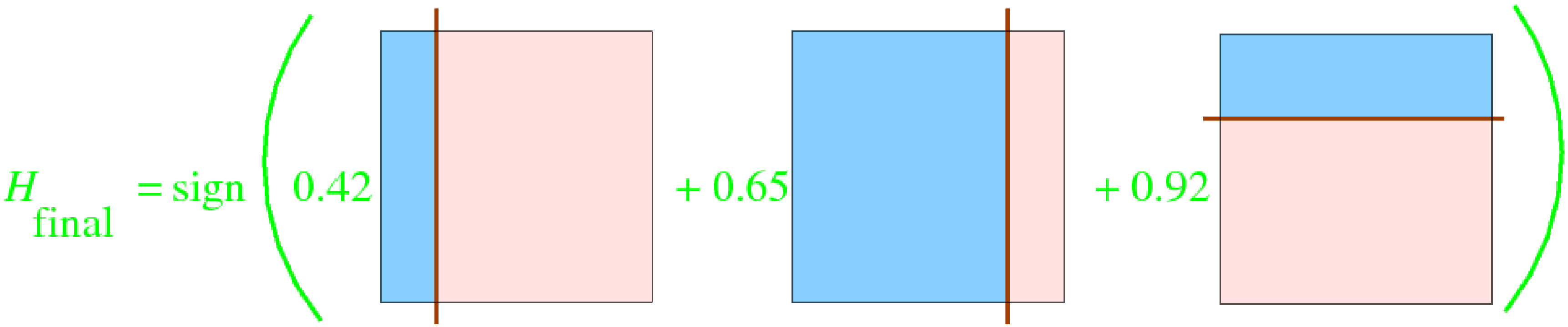
➢ Error rate of the second classifier h_3: $\epsilon_3 = 0.14$

➢ Weighting of the classifier h_3: $\alpha_3 = \frac{1}{2}\ln\left(\frac{1-\epsilon_3}{\epsilon_3}\right) = 0.92$

➢ Stop the iteration

- Final classifier

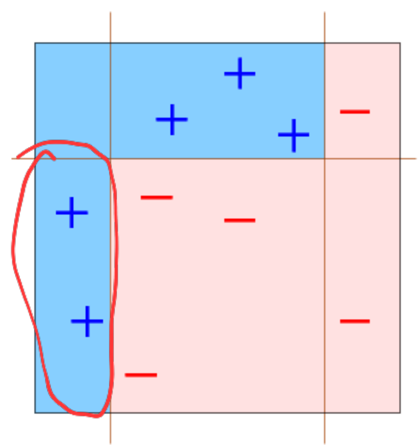
Combining the three classifiers with a linear combination



比较权重之和，进行分类

$0.42+0.65 > 0.92$

=



- Adaboost algorithm

1) Initialize the weight of examples as $\omega_0^{(n)} = \frac{1}{N}$ for $n = 1, \cdots, N$

2) For the k-th iteration, train a classifier $h_k(\boldsymbol{x})$ with the training examples weighted by $w_{k-1}^{(n)}$

3) Evaluate the weighted classification error

$$\epsilon_k = \frac{\sum_{n=1}^{N} \omega_{k-1}^{(n)} I\big(y_i \neq h_k(\boldsymbol{x}_n)\big)}{\sum_{n=1}^{N} \omega_{k-1}^{(n)}}$$

4) Determine the vote stake of the k-th classifier

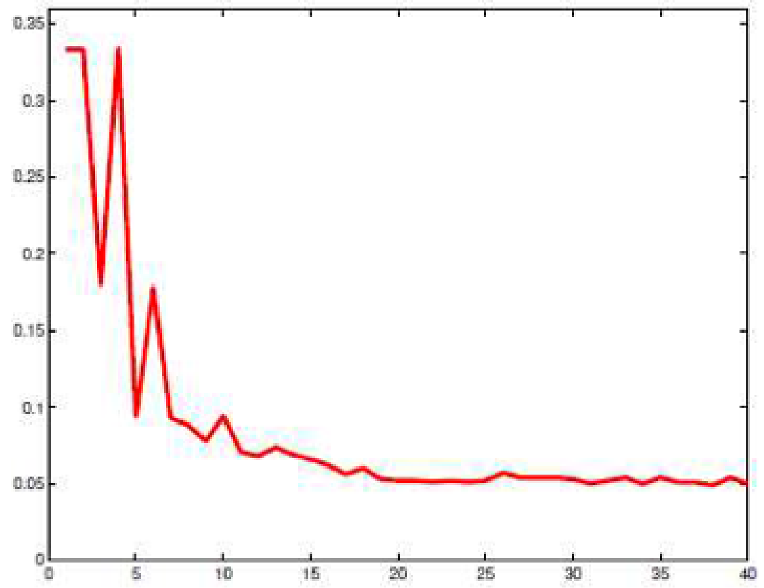
$$\alpha_k = \frac{1}{2} \ln\left(\frac{1-\epsilon_k}{\epsilon_k}\right)$$

5) Update the weights of examples as

判断分类是否正确

$$\omega_k^{(n)} = \omega_{k-1}^{(n)} \exp\{-y_i h_k(\boldsymbol{x}_i) \alpha_k\}$$

- A typical error rate curve as a function of the number of weak classifiers



- Typical weak classifiers
  - Decision trees
  - Logistic regressions
  - Neural networks

Theories behind the Adaboost

- Define the following *exponential loss*

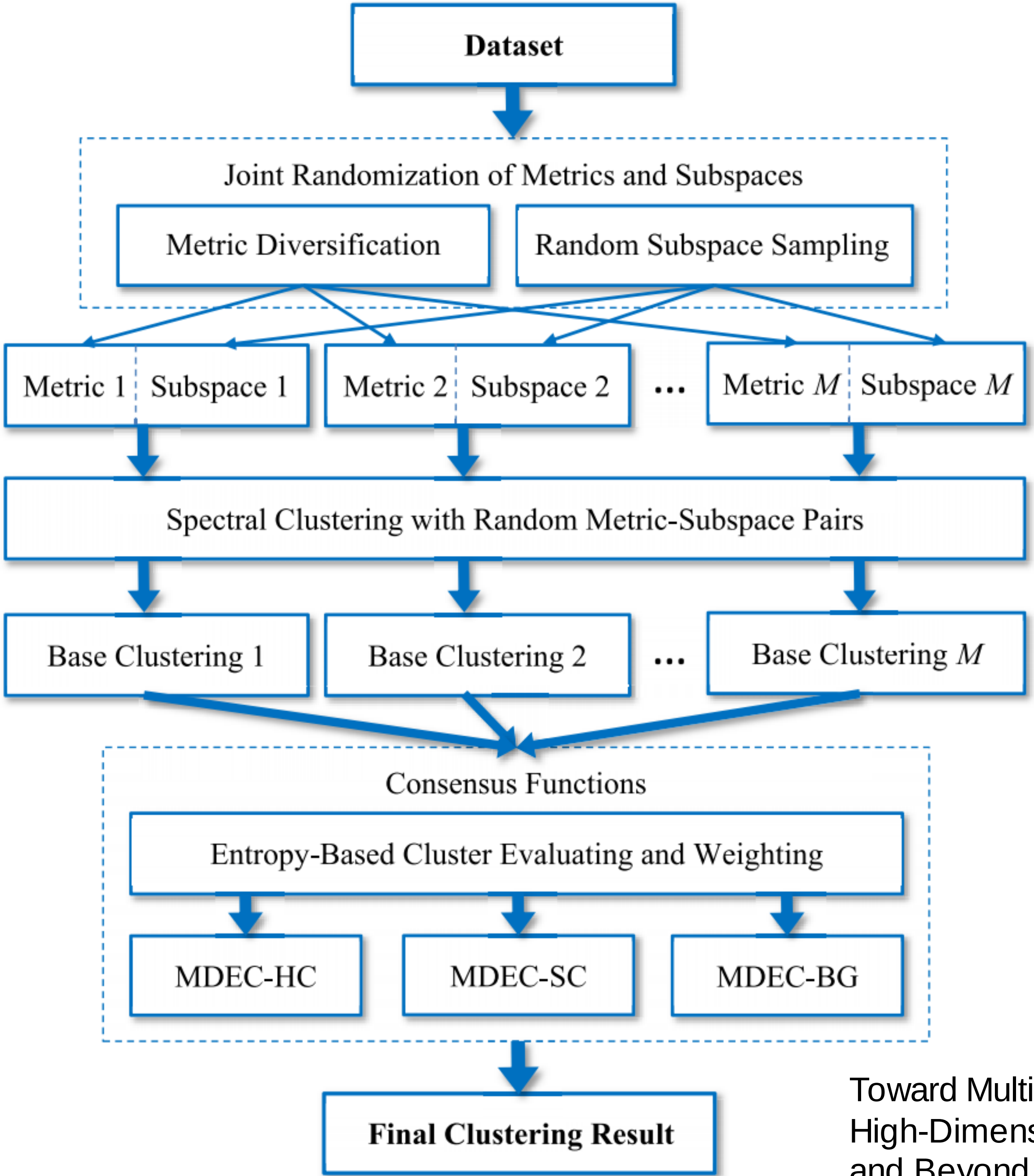
$$\mathcal{L} = \sum_{n=1}^{N} \exp\left\{-y^{(n)} h_{combine}\left(\boldsymbol{x}^{(n)}\right)\right\}$$

where $\boldsymbol{x}^{(n)}$ is the input, $y^{(n)} \in \{-1, 1\}$ is the label; and $h_{combine}(\cdot)$ is the combined classifier

$$h_{combine}(\boldsymbol{x}) = \alpha_1 h_1(\boldsymbol{x}) + \cdots + \alpha_K h_K(\boldsymbol{x})$$

with $h_k(\boldsymbol{x})$ representing the k-th component classifier, *e.g.*, $h_k(\boldsymbol{x}) = sign(\boldsymbol{w}_k^T \boldsymbol{x} + b_k)$

- It can be proved that the Adaboost algorithm is *equivalent to minimize the exponential loss* in a sequential fashion



Toward Multi-Diversified Ensemble Clustering of High-Dimensional Data: From Subspaces to Metrics and Beyond, IEEE TCYB, 2021.