Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

在线集成学习

赵惟肖　硕士研究生

2019年06月30日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

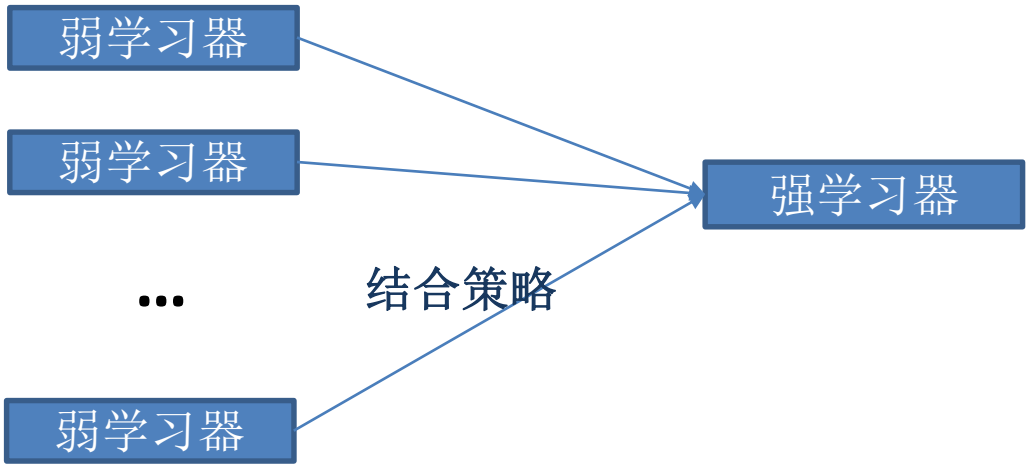
背景简介

背景简介

- 预期收获
  - 1. 了解集成学习和在线学习基本思想
  - 2. 回顾离线bagging（装袋）和离线boosting（提升）的基本原理
  - 3. 理解在线bagging和在线boosting的算法原理

背景简介

- 集成学习
  - 对训练集数据，训练若干个弱学习器，通过一定的结合策略，最终形成一个强学习器，以“博采众长”。
  - Bagging系列算法：随机森林
  - Boosting系列算法：AdaBoost、GBDT



背景简介

- 在线学习
  - 批量学习（Batch）：在训练模型时，一次性的把所有样本全部输入，有存储要求，“填鸭式”。
  - 在线学习（Online）：每输入一个样本，计算一次误差并调整参数，适用于实时产生数据的web网站，“水流式”。
  - 在工业界，参与训练的数据量很容易过TB，对资源的压力很大。Batch->Online能带来明显的经济效益。

- 在线学习

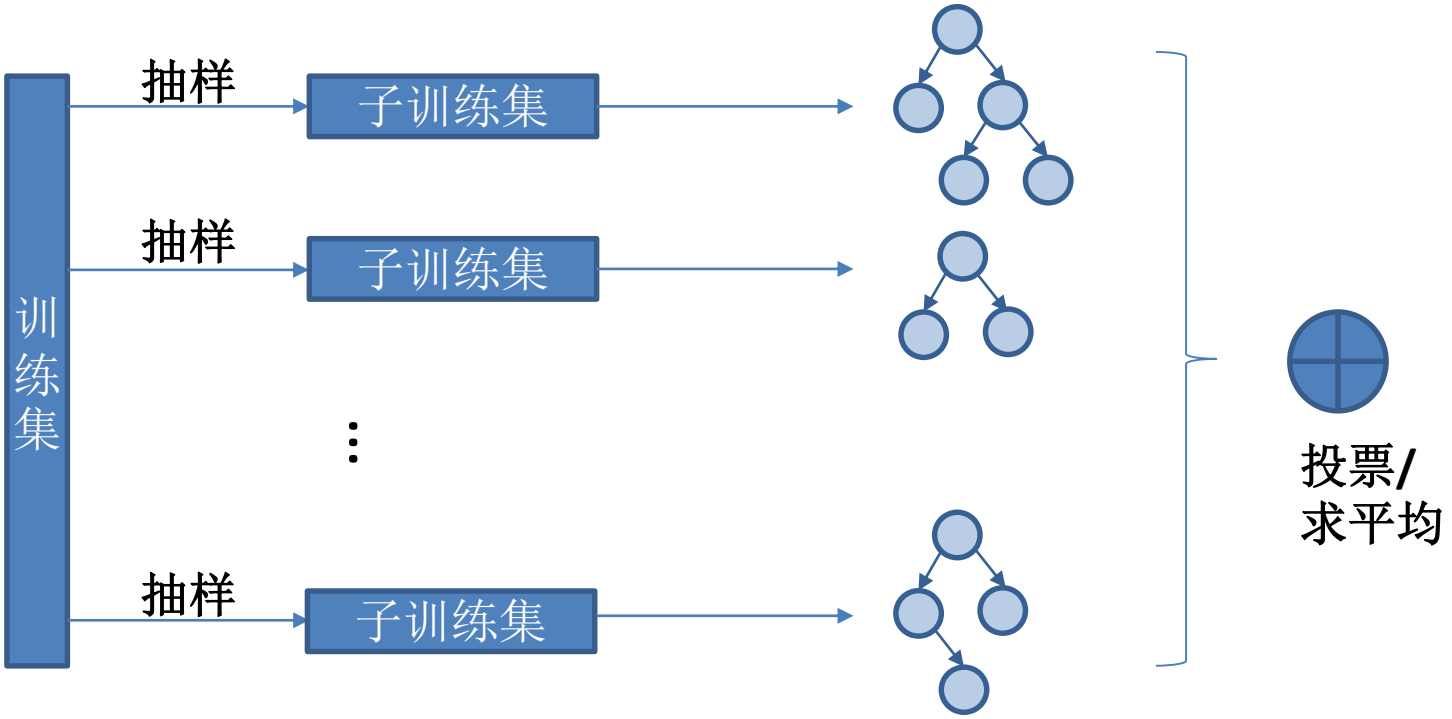
在线学习

- 基于决策树：ID5R、ITI
- 基于SVM：I-SVM
- 基于优化算法：OGD、FOBOS、FTRL
- 基于神经网络：ENN
- 基于无监督：SOINN
- 基于集成：Online-Bagging、Online-Boosting

基本概念

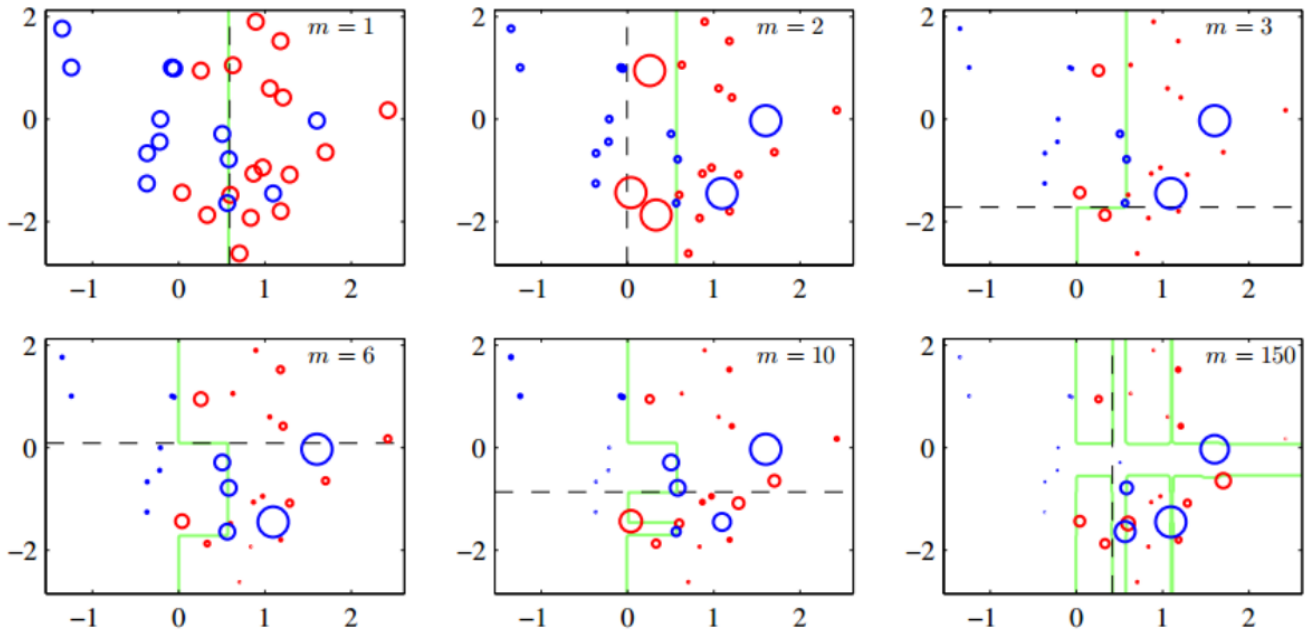
基本概念

- 基本概念——随机森林（Bagging）
  - 用随机的方式建立一个“森林”，“森林”包括很多决策树，对决策树的结果进行投票。（并型结构）



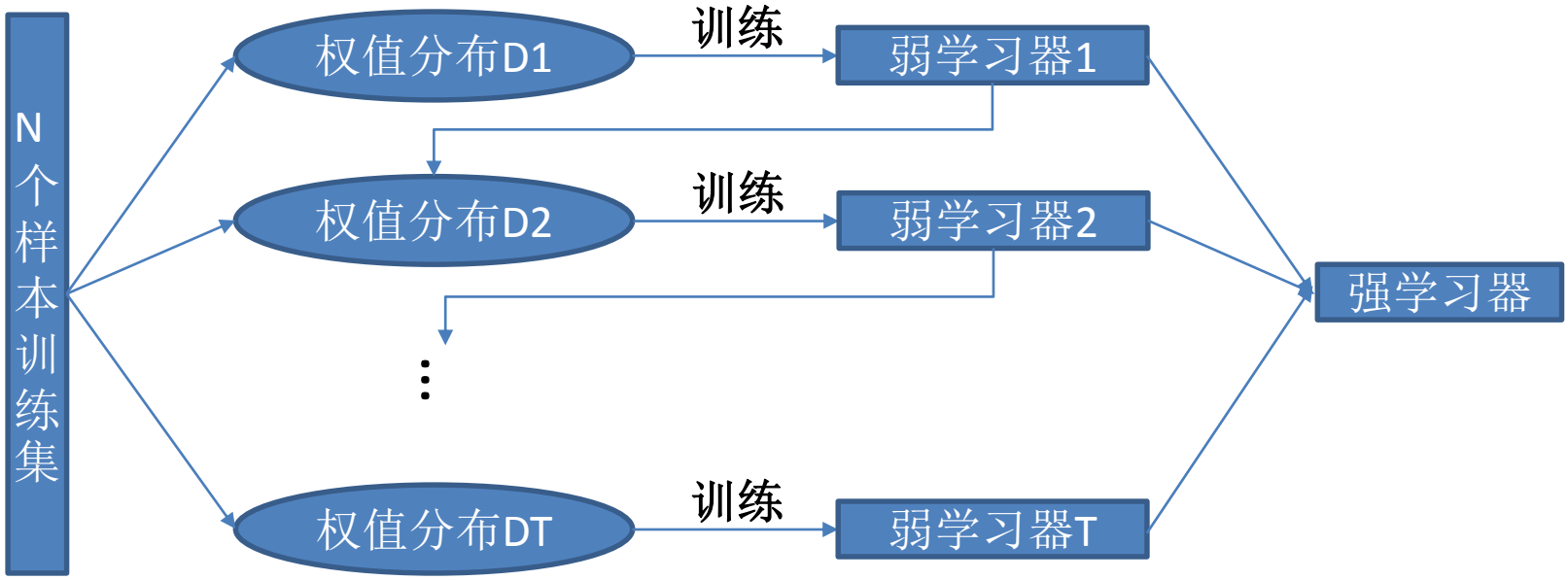
基本概念

- 基本概念——AdaBoost（Boosting）
  - 初始化：如果有N个样本，则赋予相同的权值：1/N
  - 迭代中：如果某个样本被准确分类，降低权值；被错误分类，提高权值。并结合新权值训练此次迭代的弱分类器。
  - 输出：弱分类器组合的强分类器。



基本概念

- **基本概念——AdaBoost（Boosting）**
  - **下一个分类器重点关注之前分类器预测不够准的样本（串行结构）**



可参考《提升方法（Boosting Methods）-刘晓双-2017-09-10 19_00_00》

算法原理

算法原理

P	满足集成学习算法在大数据背景下的快速训练
C	大量有标签样本，“水流式”输入模式
D	学习新知识和不遗忘旧知识之间的平衡
L	IEEE A+类会议

T	对Bagging和Boosting进行在线化
I	单个样本（X，y）
P	Online Bagging 1.根据泊松分布设定随机数 2.进行随机数次数的在线决策树求解 Online Boosting 1、2步同Online Bagging 3.对样本权重进行更新
O	对单个样本更新后的Online Bagging/Online Boosting模型

算法原理

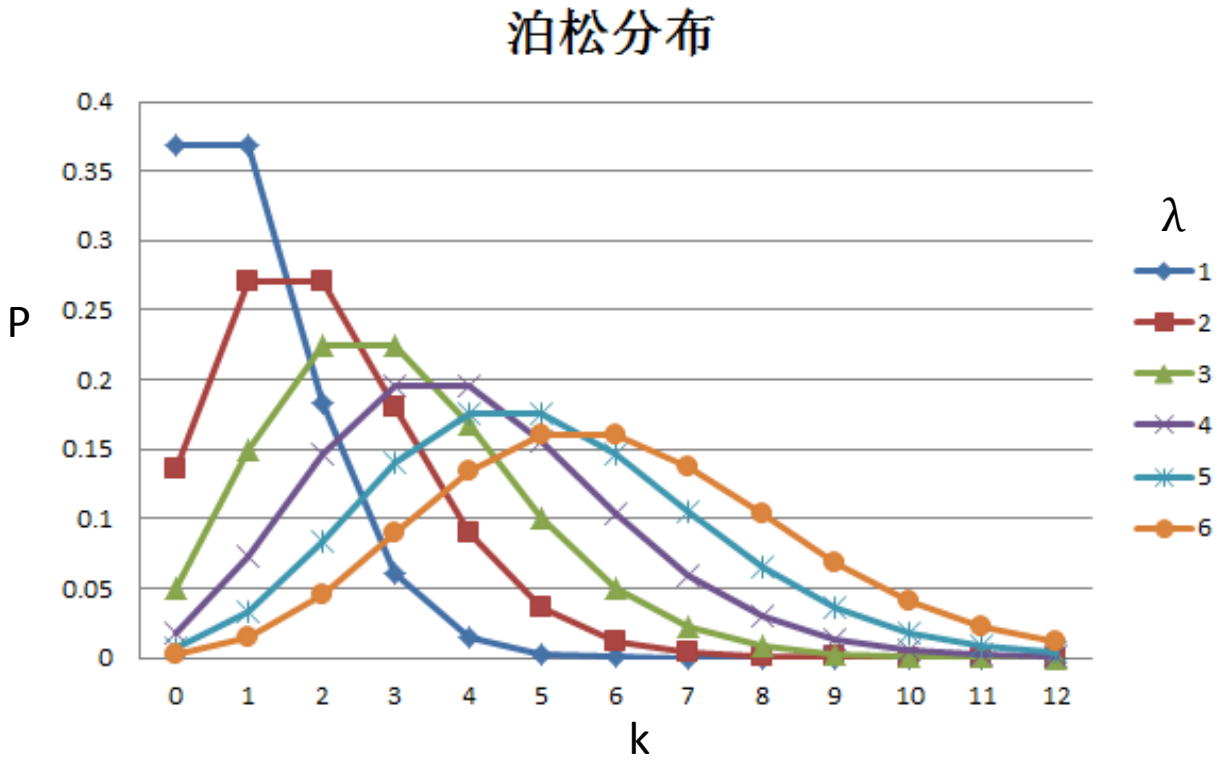
- Online Bagging
  - 泊松分布：在二项分布的伯努利试验中，若试验次数N很大，二项分布的概率p很小，且乘积λ = N*p比较适中，则事件出现的次数的概率可以用泊松分布来逼近。
  - 二项分布$P(k) = \binom{N}{k}\left(\frac{\lambda}{N}\right)^k\left(1-\frac{\lambda}{N}\right)^{N-k}$
  - 当$N \to \infty$,可转化为Poisson(λ)分布：

$$\lim_{N\to\infty} P(k) = \lim_{N\to\infty} \frac{N!\,\lambda^k}{N^k(N-k)!\,k!}\left(1-\frac{\lambda}{N}\right)^N\left(1-\frac{\lambda}{N}\right)^{-k}$$

$$= \lim_{N\to\infty} \underbrace{\left[\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{k-1}{N}\right)\right]}_{\to 1}\frac{\lambda^k}{k!}\underbrace{\left(1-\frac{\lambda}{N}\right)^N}_{\to \exp(-\lambda)}\underbrace{\left(1-\frac{\lambda}{N}\right)^{-k}}_{\to 1} = \frac{\lambda^k\exp(-\lambda)}{k!}$$

算法原理

- Online Bagging
  - 对在线学习，可以认为数据量$N \to \infty$，则对于每一个新输入的训练样本$\mathrm{d}(X, y)$，通过Poisson(λ)分布得到一个随机数，对样本重复训练随机数次。



算法原理

- Online Bagging
  - 在在线学习时，对于单棵树，转化为批量决策树算法->在线决策树算法问题。
  - 主要思想：每棵树可以在线分裂。
  - 叶子节点分裂：熵或Gini

$$L(R_j) = -\sum_{i=1}^{K} p_i^j \log(p_i^j) \quad or \quad L(R_j) = \sum_{i=1}^{K} p_i^j (1 - p_i^j)$$

  - 每棵树上叶子节点P_j概率 $P_j = [p_1^j, \dots, p_K^j]$

j：第j棵树　i：第i个分类结果　K：总分类数

算法原理

- Online Bagging
  - 可写作左节点和右节点的概率

$$P_{jls} = [p_1^{jls}, \dots, p_K^{jls}] \quad and \quad P_{jrs} = [p_1^{jrs}, \dots, p_K^{jrs}]$$

  - 每棵树上叶子节点P_j分裂必须符合以下两个条件：

1. 落在P_j的样本个数必须大于一个常数（可以人工设定）
2. 叶子节点的Gini必须大于一个常数（可以人工设定）

  - Gini计算公式：

$$\Delta L(R_j, s) = L(R_j) - \frac{|R_{jls}|}{|R_j|} L(R_{jls}) - \frac{|R_{jrs}|}{|R_j|} L(R_{jrs})$$

  - 从而完成整棵树的更新。

l：左节点　r：右节点　s：在线预测中给定的测试集合　|·|：样本数目

算法原理

- Online Bagging

Algorithm 1 On-line Random Forests

Require: Sequential training example $\langle x, y \rangle$
Require: The size of the forest: T
Require: The minimum number of samples: α
Require: The minimum gain: β

1: // For all trees
2: **for** t from 1 to T **do**
3: k ← Poisson(λ)
4: **if** $k > 0$ **then**
5: // Update k times
6: **for** u from 1 to k **do**
7: $j =$ findLeaf(x).
8: updateNode$(j, \langle x, y \rangle)$.
9: **if** $|\mathcal{R}_j| > \alpha$ and $\exists s \in \mathcal{S} : \Delta L(\mathcal{R}_j, s) > \beta$ **then**
10: Find the best test:
$s_j = \arg\max_{s \in \mathcal{S}} \Delta L(\mathcal{R}_j, s)$.
11: createLeftChild$(\mathbf{p}_{jls})$
12: createRightChild$(\mathbf{p}_{jrs})$
13: **end if**
14: **end for**
15: **else**
16: Estimate $OOBE_t \leftarrow$ updateOOBE$(\langle x, y \rangle)$
17: **end if**
18: **end for**
19: Output the forest $\mathcal{F}$.

3：用possion(λ)确定采样次数k
6：更新k次
7：j第t棵树上叶子节点
8：统计第j个叶子节点的数目，计算Gini
9：判断是否分裂的两个条件
10：在符合条件的叶子节点中，选择一个Gini最大的叶子节点作为分类节点
11：创建左子树
12：创建右子树
16：估计袋外误差（out of bag error）

算法原理

- Online Boosting
  - 回顾Batch AdaBoost：样本被准确分类，降低权值；被错误分类，提高权值。
  - Online AdaBoost：输入新样本$d\langle x, y\rangle$，对每一个弱分类器h_m，初始化$\lambda_m^{\text{corr}} = 1, \lambda_m^{wrong} = 1$，$\lambda_d = 1$，当弱分类器能正确分类该样本时

加大λ_m^{corr}：$\lambda_m^{\text{corr}} \leftarrow \lambda_m^{corr} + \lambda_d$

减小错误率ε_m：$\varepsilon_m \leftarrow \frac{\lambda_m^{wrong}}{\lambda_m^{corr} + \lambda_m^{wrong}}$

减小样本权值λ_d：$\lambda_d \leftarrow \lambda_d\left(\frac{1}{2(1-\varepsilon_m)}\right)$

算法原理

- Online Boosting
  - 当弱分类器错误分类该样本时

加大λ_m^{wrong}：$\lambda_m^{\text{wrong}} \leftarrow \lambda_m^{\text{wrong}} + \lambda_d$

加大错误率ε_m：$\varepsilon_m \leftarrow \frac{\lambda_m^{wrong}}{\lambda_m^{corr}+\lambda_m^{wrong}}$

加大样本权值λ_d：$\lambda_d \leftarrow \lambda_d(\frac{1}{\varepsilon_m})$

算法原理

- Online Boosting

Algorithm: On-line AdaBoost

Require: training example $d\langle \boldsymbol{x}, y\rangle$, $y \in \{-1,1\}$

Require: strong classifier $\boldsymbol{h}$(initialized randomly)

Require: weight $\lambda_m^{corr}, \lambda_m^{wrong}$(initialized with 1), set $\lambda_d = 1$

1: For each base model $h_m \in \boldsymbol{h}$, $m \in \{1,2,\dots,M\}$

2: Set k according to $Poisson(\lambda)$

3: Do k times

4: $h_m = OnlineBase(h_m, d(\boldsymbol{x}, y))$

5: if $y = h_m(\boldsymbol{x})$

6: $\lambda_m^{corr} \leftarrow \lambda_m^{corr} + \lambda_d$ $\varepsilon_m \leftarrow \frac{\lambda_m^{wrong}}{\lambda_m^{corr}+\lambda_m^{wrong}}$ $\lambda_d \leftarrow \lambda_d(\frac{1}{2(1-\varepsilon_m)})$

7: else

8: $\lambda_m^{wrong} \leftarrow \lambda_m^{wrong} + \lambda_d$ $\varepsilon_m \leftarrow \frac{\lambda_m^{wrong}}{\lambda_m^{corr}+\lambda_m^{wrong}}$ $\lambda_d \leftarrow \lambda_d(\frac{1}{\varepsilon_m})$

9: End

10: Get $\boldsymbol{h} = \text{argmax} \sum_{m=1}^{M} \log(\frac{1-\varepsilon_m}{\varepsilon_m}) I(h_m(\boldsymbol{x}) = y)$

优劣分析

优劣分析

- 优势
  - 无需大量存储空间存放训练样本
  - 可处理数据流，做在线训练
- 劣势
  - 精度不如批量集成学习

应用总结

应用总结

- 算法的应用领域
  - 视频跟踪、人脸识别

应用总结

- 算法的应用领域
  - 数据流的分类/回归任务

1.传统行业的应用：

金融银行系统内部的实时计算和决策

原油变化的在线分类/在线空气质量预测/钢炉水温度预报

2.互联网领域应用：

推荐算法

3.物联网邻域应用：

医学诊断、智能家具等传感器数据的实时计算

参考文献

参考文献

[1] Oza, Nikunj C. "Online bagging and boosting." 2005 IEEE international conference on systems, man and cybernetics. Vol. 3. Ieee, 2005.
[2] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof, “On-line Random Forests” in 3rd IEEE ICCV Workshop on On-line Computer Vision, 2009.
[3] Zhang, Kaihua, and Huihui Song. "Real-time visual tracking via online weighted multiple instance learning." Pattern Recognition 46.1 (2013): 397-411.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。