

第七章 随机方法 Stochastic Methods

本章目录



- 1 随机搜索
- 2 Boltzmann学习
- 3 Boltzmann网络和图模型
- 4 进化方法 – 遗传算法
- 5 遗传规划



中山大學

7.1 内容介绍

- 对于高维和复杂的模型，由于经常出现许多局部极值，这时必须利用各种处理技巧。本章将主要研究两大类通用随机搜索方法。
- 其一，以Boltzmann学习机作为范例，是一种来自物理学的概念和技术；它已形成高度发展和严格的理论，并且在模式识别中取得很多成功，因而将花主要的篇幅讲述；
- 其二，以遗传算法为范例，源自生物学的若干概念，特别是有关进化的数学理论，它更具启发性和灵活性，当计算资源充足时，不失是一个很吸引人的方法。

3



中山大學

7.2 随机搜索

- 假定给定多个变量 s_i 其中每个变量的数值都取两个离散值之一。为简单起见，记它们为 ± 1 。优化问题是这样描述的：确定 N 个 s_i 的合适取值，使下述代价函数或能量函数最小：

其中的权值 w_{ij} 是对称的，取值可正可负，可以令到自身的反馈为0。

4



中山大學

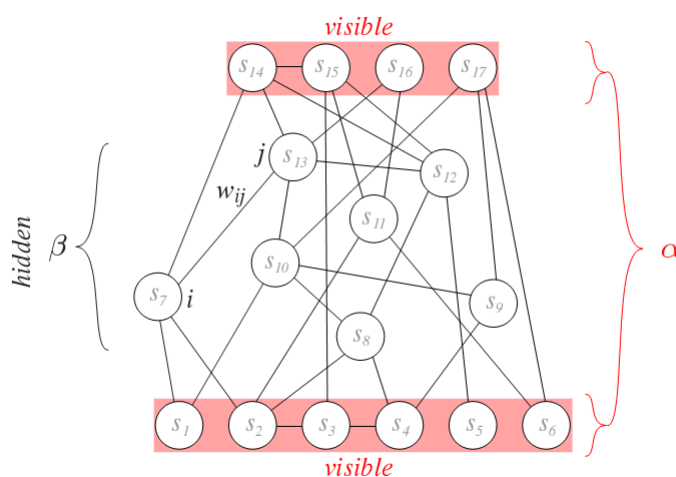
7.2 随机搜索

- 想象一下该网络代表N个物理磁体，每个磁体的北极要么指上 ($s_i = +1$)，要么指向下 ($s_i = -1$)。 W_{ij} 是描述磁体间的物理分离度的函数。对每对磁体间存在一个交互作用的能量，即：
- 优化的任务就是在由这些磁体组成的集团的所有构型当中寻找到最稳定的构型，也就是对应于最低能量的那个构型，即使下面的能量函数最小：

5



中山大學



6

7.2.1 模拟退火 (Simulated Annealing) 中山大學

- **模拟退火算法**：基本思想是把某类优化问题的求解过程与统计热力学中的热平衡问题进行对比，试图通过模拟高温物体退火过程的方法，来找到优化问题的全局最优或近似全局最优解。

7

启发：

- 一个物体（例如金属）的退火过程大体上是这样的：
 - 首先对该物体加热（熔化），那么物体内的原子就可高速自由运行，处于较高的能量状态。但是作为一个实际的物理系统，原子的运行总是最低的能态。
 - 一开始温度较高时，高温使系统具有较高的内能，而随着温度的下降，原子越来越趋向于低能态，最后整个物体形成最低能量的基态。

8



中山大学

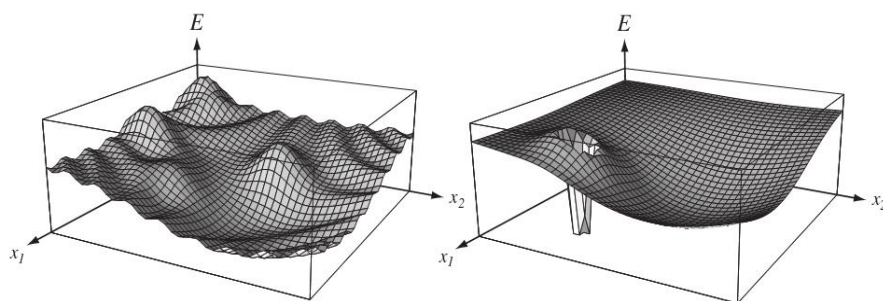


图 7-2 左边的能量函数或能量地形很适合用模拟退火之类的优化求解方法。这类方法利用了随机性,在一控制参数(或温度)的控制下能避免陷入局部极小因而能发现全局最小点,就好像有一个球一边震动,一边在该地形曲面上滚动一样。右边的病态的“高尔夫球场”类型,则很不适合模拟退火求解,因为其能量最小点的区域太小了,而且被一些局部能量高峰阻隔。这种构型空间的问题,我们还将图 7-6 中做更清晰的解释

9



中山大学

7.2.2 Boltzmann(玻尔兹曼)因子

- 物理学中有一个关键结论: 系统位于一个特定(离散)构型 γ 具有能量 E_γ 的概率由

给出, 其中 $Z(T)$ 是个归一化函数。式中分子部分称为“Boltzmann因子”, 分母部分是所谓的“配分函数”, 它保证上式确是个真正概率。

$$Z(T) = \sum_{\gamma'} e^{-E_{\gamma'}/T}$$

- 大致来说, 高温给了系统更多的能量, 使出现高能量构型的概率增大。这也定性地解释了Boltzmann因子中概率对 T 的相依关系:

在高温时, 所有构型的概率分布大致平均, 而在低温时, 系统则集中分布在具有最低能量的构型周围。

10



中山大學

模拟退火算法

首先将网络随机初始化，并设定一个高的初始“温度” $T(1)$ ，然后随机选择一个节点 i ，假定其现在的状态是 $s_i = +1$ ，计算在这种构型下系统总能量 E_a ，接着再计算如果改变到候选状态，即 $s_i = -1$ 时，对应的系统总能量 E_b 。如果候选状态能量 $E_b < E_a$ ，则接受这次状态改变，反而则按照如下概率接受这个状态改变：

11



中山大學

Hint:

值得指出的是，温度函数 $T(k)$ （这里 k 是迭代的次数）被称为冷却进度或退火进度。 $T(1)$ 应该足够高，以使得全部构型有大致一样的概率。温度应该十分缓慢地逐渐下降，系统能够到达状态空间的任何区域，同时又避免陷入不希望的局部极小处。

一种典型的退火进度利用了公式 $T(k+1) = cT(k)$ ，其中 $0 < c < 1$ 。若不考虑计算资源，初始值都宜设高一些。实际中， $0.8 < c < 0.99$ 之间的 c 可以工作得很好。

12



中山大學

模拟退火算法可以分解为解空间、目标函数和初始解三部分。

模拟退火的基本思想：

- (1) 初始化：初始温度 T (充分大)，初始解状态 S (是算法迭代的起点)，每个 T 值的迭代次数 L
- (2) 对 $k=1, \dots, L$ 做第(3)至第6步：
- (3) 产生新解 S'
- (4) 计算增量 $\Delta t' = C(S') - C(S)$ ，其中 $C(S)$ 为评价函数
- (5) 若 $\Delta t' < 0$ 则接受 S' 作为新的当前解，否则以概率 $\exp(-\Delta t'/T)$ 接受 S' 作为新的当前解。
- (6) 如果满足终止条件则输出当前解作为最优解，结束程序。
终止条件通常取为连续若干个新解都没有被接受时终止算法。
- (7) T 逐渐减少，且 $T > 0$ ，然后转第2步。

13



中山大學

算法 1 (随机模拟退火)

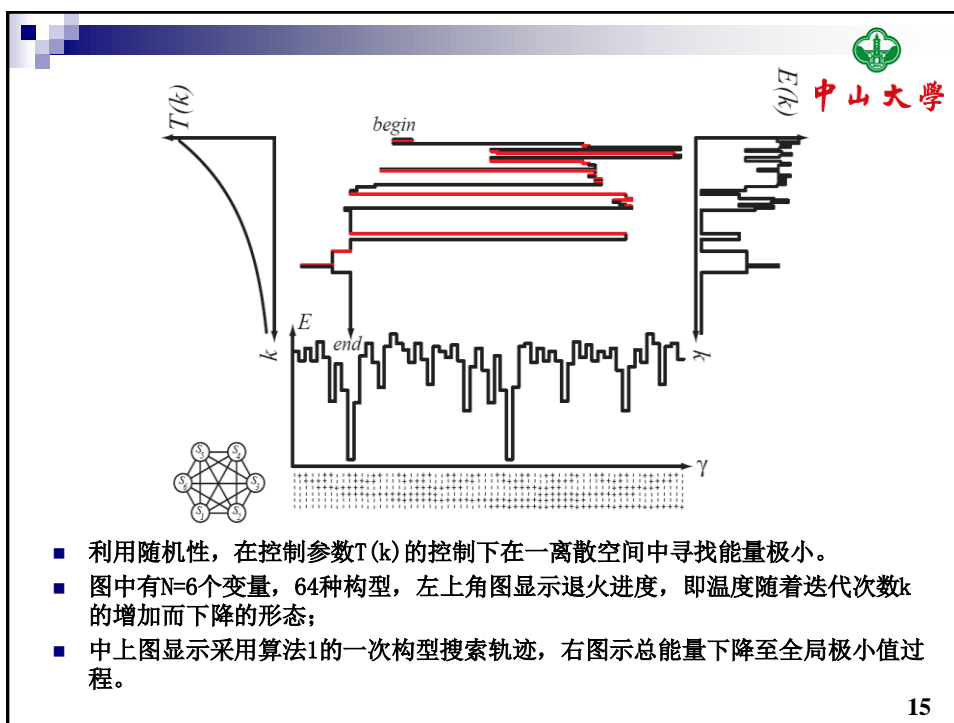
```

1 begin initialize  $T(k), k_{max}, s_i(1), w_i \quad i, j = 1, \dots, N$ 
2    $k \leftarrow 0$ 
3   do  $k \leftarrow k + 1$ 
4     do 随机地选择节点  $i$ ; 假设它的状态为  $s_i$ 
5        $E_u \leftarrow -1/2 \sum_j^N w_{ij} s_i s_j$ 
6        $E_b \leftarrow -E_u$ 
7       if  $E_b < E_u$ 
8         then  $s_i \leftarrow -s_i$ 
9       else if  $e^{-(E_b - E_u)/T(k)} > \text{Rand}[0, 1)$ 
10        then  $s_i \leftarrow -s_i$ 
11   until 所有节点轮询多次
12   until  $k = k_{max}$  或停止准则满足
13   return  $E, s_i, i = 1, \dots, N$ 
14 end

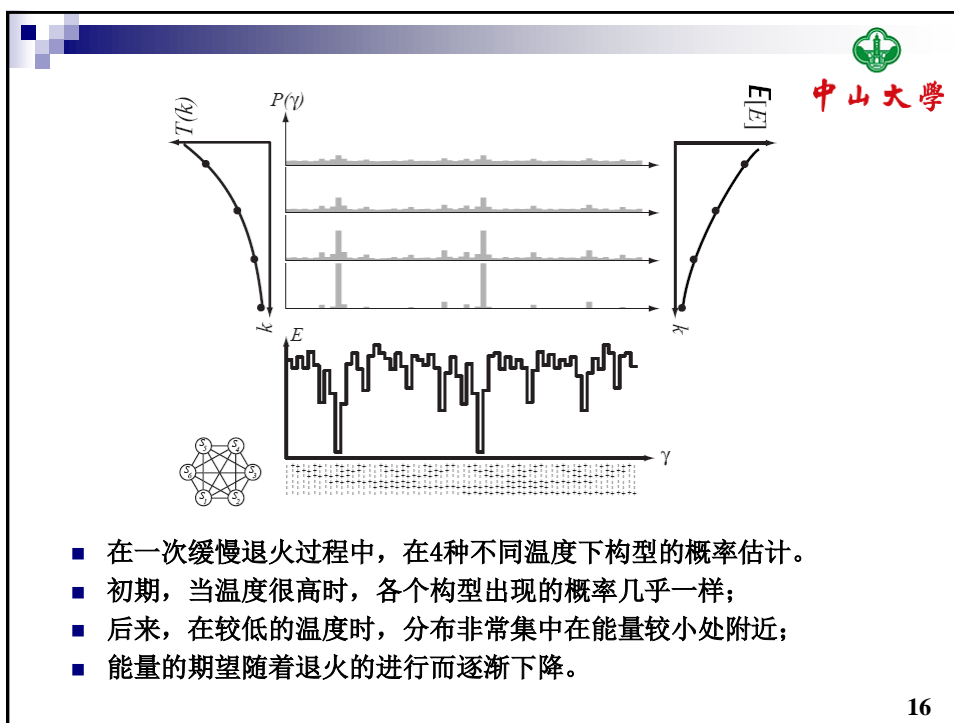
```

几个关键点：起始温度、温度下降的速率、终止温度、终止准则

14



15



16



中山大學

7.2.3 确定性模拟退火

在搜索中允许节点取模拟状态值，在搜索终止时，这些状态值被强制到最优化所需的 ± 1 。要求在高温时很大的朝上的力也不能确保 $s_i = +1$ ，而当温度很低时，很小的外力就可以使 $s_i = +1$ 或 -1 。故若令 $l_i = \sum_j w_{ij} s_j$ 表示施加在节点 i 上的外力，则更新后的状态值成 $s_i = f(l_i, T) = \tanh[l_i / T]$

算法 2 (确定性模拟退火)

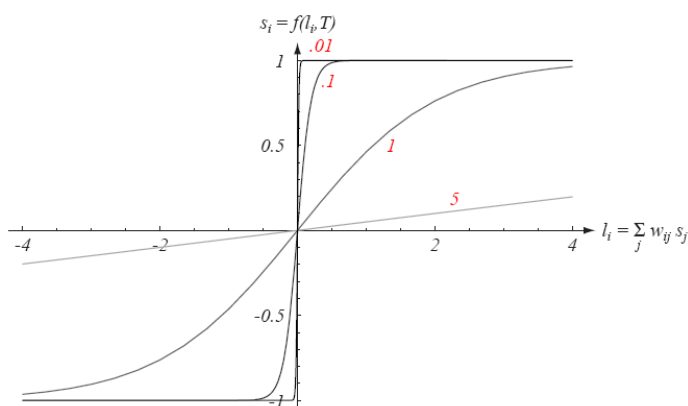
```

1 begin initialize  $T(k), w_{ij}, s_i(1), i, j = 1, \dots, N$ 
2    $k \leftarrow 0$ 
3   do  $k \leftarrow k + 1$ 
4     随机地选择节点  $i$ 
5      $l_i \leftarrow \sum_j w_{ij} s_j$ 
6      $s_i \leftarrow f(l_i, T(k))$ 
7   until  $k = k_{max}$  或收敛准则满足
8   return  $E, s_i, i = 1, \dots, N$ 
9 end
```

17

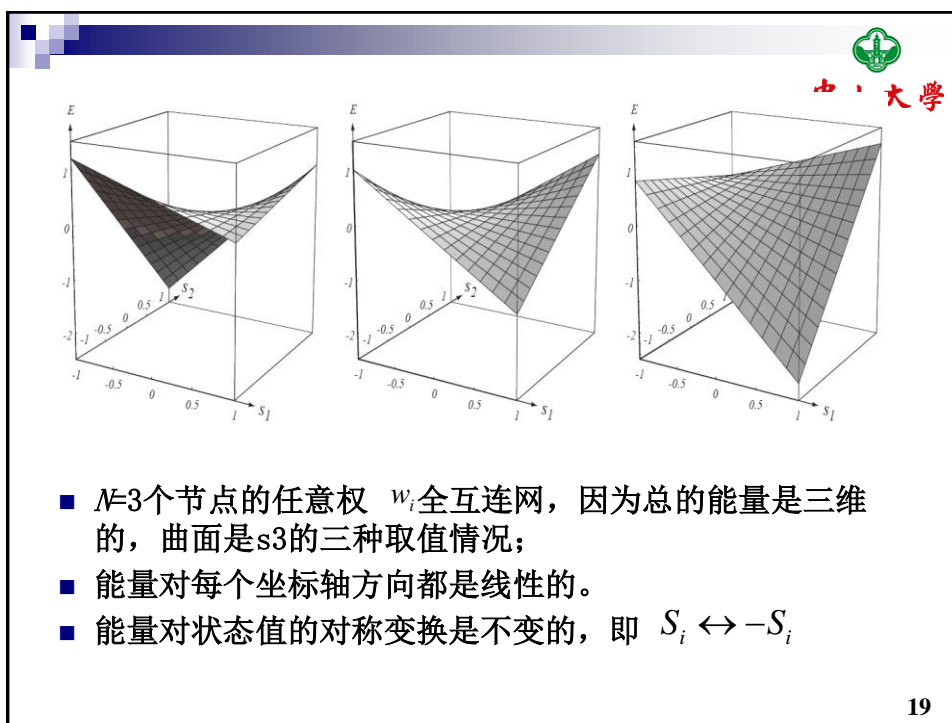


中山大學

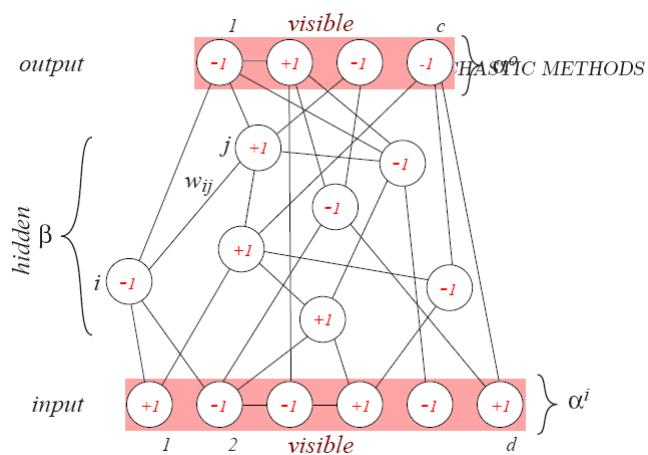


- 在确定性退火中，每个节点的状态可在范围 $-1 \leq s_i \leq 1$ 内连续取值
- l_i 表示与节点 i 相连的所有其他节点的外力。

18



7.3 Boltzmann学习



- 当图7-1那样的网络用于学习时，区别两类可见单元是很重要的。



中山大學

7.3.1 可见状态的随机Boltzmann学习

- 假定全部可见单元的概率分布已知为 $Q(a)$ ，现在要求实际经由随机仿真获得的对于给定样本集合的概率分布 $P(a)$ 与已知的 $Q(a)$ 一致。
- 可见构型的概率等于所有可能的隐状态构型的求和：

$$\begin{aligned}
 P(\alpha) &= \sum_{\beta} P(\alpha, \beta) \\
 &= \frac{\sum_{\beta} e^{-E_{\alpha\beta}/T}}{Z}
 \end{aligned}$$

其中， $E_{\alpha\beta}$ 是对应可见单元和隐单元构型的系统能量。 Z 是系统总的配分函数。

21



中山大學

7.3.1 可见状态的随机Boltzmann学习

- 对实际分布和期望分布差异的一个自然度量是相对熵，Kullback-Leibler距离或Kullback-Leibler散度，即：

$$D_{KL}(Q(\alpha), P(\alpha)) = \sum_{\alpha} Q(\alpha) \log \frac{Q(\alpha)}{P(\alpha)}.$$

D_{KL} 非负只有 Q, P 相等时才为0.

22



中山大学

学习的过程

- Boltzmann基于相对熵的梯度下降算法。训练模式集确定了 $Q(a)$ ，我们的目的是确定合适的权值，使得在温度 T 上实际分布 $P(a)$ 与已知的 $Q(a)$ 尽可能地接近。
- 可以证明，权值更新公式为：

$$\Delta w_{ij} = \frac{\eta}{T} \left[\underbrace{\mathcal{E}_Q[s_i s_j]_{\alpha \text{ clamped}}}_{\text{learning}} - \underbrace{\mathcal{E}[s_i s_j]_{\text{free}}}_{\text{unlearning}} \right]$$

其中已定义 $\mathcal{E}_Q[s_i s_j]_{\alpha \text{ 冻结}} = \sum_{\alpha \beta} Q(\alpha) P(\beta | \alpha) s_i(\alpha \beta) s_j(\alpha \beta)$

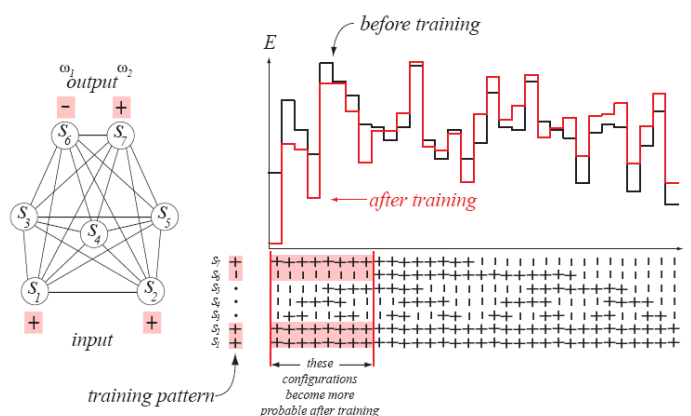
- 同样道理，对输入—输出联想的随机学习，我们也可以得到其权值更新公式为：

$$\Delta w_{ij} = \frac{\eta}{T} \left[\underbrace{\mathcal{E}_Q[s_i s_j]_{\alpha', \alpha'' \text{ 冻结}}}_{\text{学习}} - \underbrace{\mathcal{E}[s_i s_j]_{\alpha' \text{ 冻结}}}_{\text{非学习}} \right]$$

23



中山大学

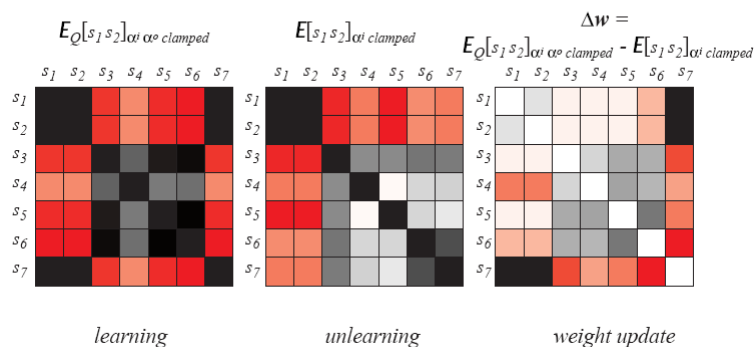


- 一个7-单元互连的Boltzmann学习算法训练为对输入模式 $s_1=+1, s_2=+1$ 赋类别值 ω_2

24



中山大學



■ Boltzmann学习过程

25



中山大學

7.3.2 丢失特征和类别约束


■ 丢失特征:

Boltzmann训练算法的一个关键优点在于不管在学习阶段还是识别阶段，它都能处理丢失特征的情况。训练中如果遇到一个缺损的二值模式，则对应于丢失的那个特征的输入节点的值允许发生改变——也就是说，可以暂时地将它看作是隐节点，而不是箝位输入节点。

■ 模式补足:

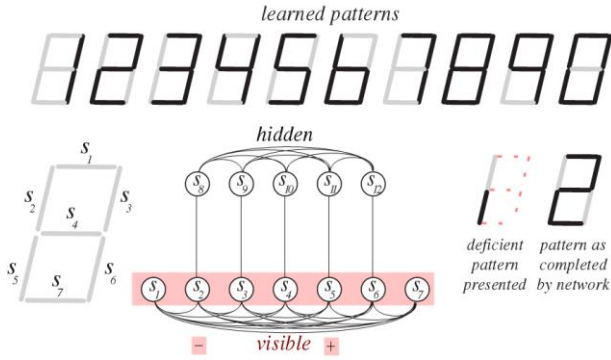
模式补足问题指的是：只给定模式的一部分，要求估计出完整的模式。Boltzmann网络能够用于模式补足，也就是填充缺损模式中的位置特征。

26




中山大學

模式补足示例:



一个具有5个隐单元的12-单元的网络用10个7-段数字模式来训练。

27



中山大學

7.3.3 确定性Boltzmann学习

确定性Boltzmann学习的基本方法就是对状态变量允许模拟取值，到最后自动收敛到问题所需的±1上。明确说，令D表示训练模式x的集合，x中包含模式特征和类别标记，其算法过程如下：

算法 3 { 确定性 Boltzmann 学习算法 }

1 **begin** **initialize** $\mathcal{D}, \eta, T(k), w_{ij} \quad i, j = 1, \dots, N$

2 **do** 随机选择训练模式 \mathbf{x}

3 状态 s_i 随机化

4 退火网络用输入和输出箝位

5 在最后的低 T , 计算 $[s_i, s_j]_{s^0, s^0}$ 箝位

6 状态 s_i 随机化

7 退火网络用输入箝位而输出自由

8 在最后的低 T , 计算 $[s_i, s_j]_{s^0, s^0}$ 箝位

9 $w_{ij} \leftarrow w_{ij} + \eta / T [[s_i, s_j]_{s^0, s^0} \text{ 箝位 } - [s_i, s_j]_{s^0, s^0}]$

10 **until** $k = k_{\max}$ 或收敛准则满足

11 **return** w_{ij}

12 **end**

28

7.3.4 初始化和参数设置



中山大学

- 对于隐单元数，能够表示 n 个不同模式最少的隐单元数也是 $\lceil \log_2 n \rceil$
尽管如此，这个隐节点个数的下界仍然不够紧，因为有可能存在这样的情况，即无法找到合适的权值组合能够惟一的表达各个模式。
- 对于权值，当然可以将所有权值都初始化为0, 但这样作将导致训练过程不必要的慢。
- 我们可任意令一半节点权为正，另一半为负。初始值若限于一定合适的范围，将有利于提高学习速度。

29

7.3.4 初始化和参数设置



中山大学

- 对于初始温度，一般应该服从：

$$T(1) = \frac{\mathcal{E}_+[\Delta E]}{\ln[m_2] - \ln[m_2 R - m_1(1 - R)]}.$$

- 假定 m_1 是本次退火中对应能量下降的状态迁移的数目。 m_2 对应上升。
- 对于学习率 η ，出于稳定性的考虑，要求：

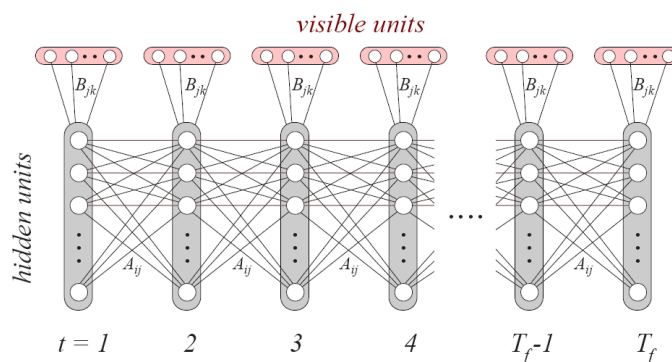
$$\eta \leq T^2 / N$$

30



中山大學

7.4 Boltzmann网络和图示模型

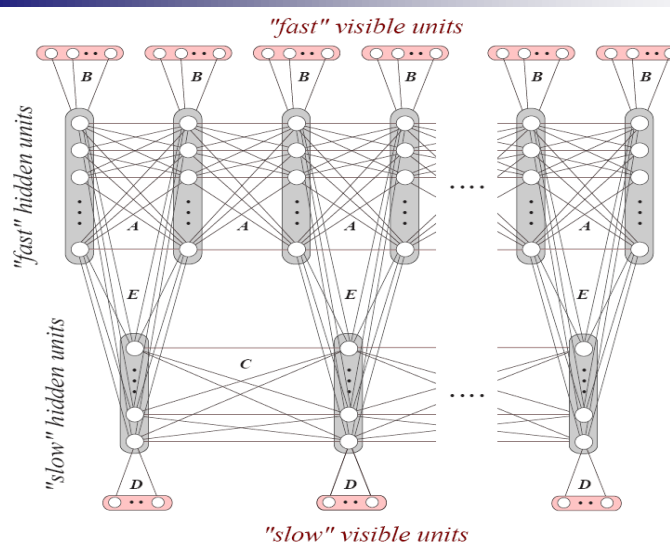


- Hidden Markov模型可以按时间展开为Trellis格型结构

31



中山大學



- 一个Boltzmann拉链由两个Boltzmann链组成，可用于语音识别问题
- 快链学习单个音素结构及转变，慢链学习整个单词或整个短语中更“大”的韵律和重音结构

32



中山大學

7.5 进化方法

7.5.1 遗传算法

算法 4 (基本遗传算法)

```

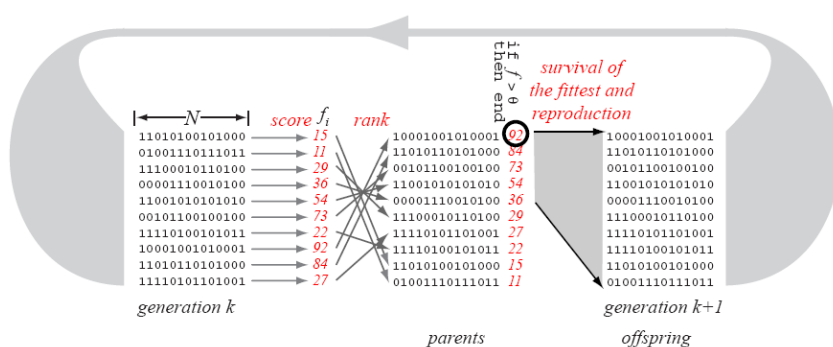
1 begin initialize  $\theta, P_{c1}, P_{mut}, L, N$ -位染色体
2   do 确定每个染色体的适应度,  $f_i, i=1, \dots, L$ 
3     染色体排序
4     do 选择得分最高的两个染色体
5       if  $\text{Rand}[0,1) < P_{c1}$  then 交叉一对随机选择的位
6       else 以概率  $P_{mut}$  改变每一位; 删除父染色体
7     until  $N$  个子代被创建
8   until 任何染色体的得分  $f$  超过  $\theta$ 
9   return 最高适应度的染色体(最佳分类器)
10 end

```

33



中山大學



- 基本遗传算法是一个随机迭代搜索算法。在第 k 代的种群中存在 L 个分类器个体，其中每一个都是用一个长度为 N 的二进制位串表示，成为染色体 (Chromosome)

34



中山大學

遗传算子

在基本遗传算法中，每个分类器的表达是一个二进制的位串，称为染色体。

复制：染色体被原样复制一遍，不发生改变；

交叉：把两条染色体混合或配对的过程，得到两条新的染色体。在染色体上随机确定一个位置并开断，将A染色体的第一部分和B染色体第二部分连接，另一半也是如此；

变异：允许每个位以一个很小的概率改变自身，比如从0变成1，或者相反。

染色体表达

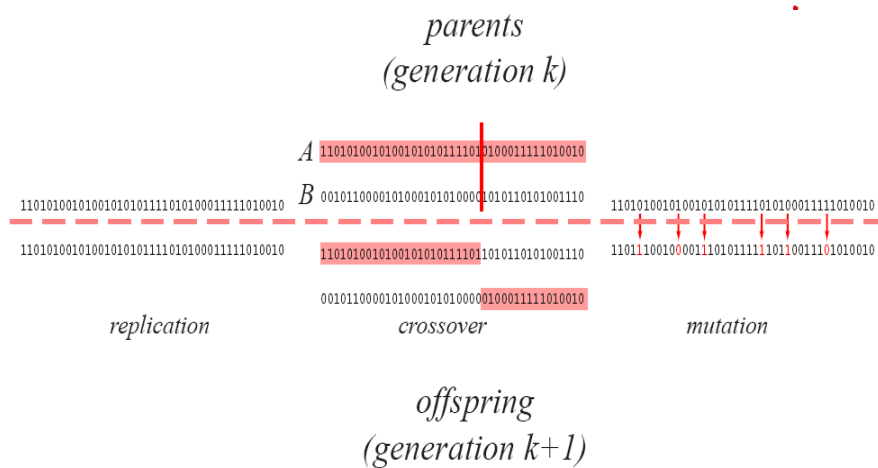
最早期最简单的是令染色体中各个位表示一个具有固定权值的两层感知器网络的各个特征；

另一种方法是令染色体的不同片断表示一个具有固定拓扑的多层神经网络的各个权值。类似地，染色体也可以用于表达网络的具体拓扑结构。

35



學



- 由三种基本遗传算法可以将染色体变换成其子代染色体复制，交叉，变异

36

■ 利用遗传算法进行模式识别的一种自然映射方法是：从一个二进制串到一棵二叉分类树。

37

得分

对c类分类问题，通常最简便的做法是进行c次二分法操作，每一次将一个不同类别与其他所有类别区分开。进行分类时，测试模式依次提供给每一个二分法，并进行相应的类别标记。

但考虑到“过拟合”问题。一种避免“过拟合”的方法是在适应度函数中增加对分类器复杂度的征罚项。别外一种方法是运用停止准则。

选择

所谓选择，是指确定在某一代中哪些染色体可以作为父本为下一代提供遗传信息。其主要的选择模式是所谓的“比例适值选择”，即选中某条染色体的概率正比于其适值函数。

该方法的一种小小的修改是令选择概率正比于适应度的某个单调递增函数。如果该函数具有正的二阶导数，那么高适应度染色体被选中的概率就会被增强。

中山大学

38

7.6 遗传规划

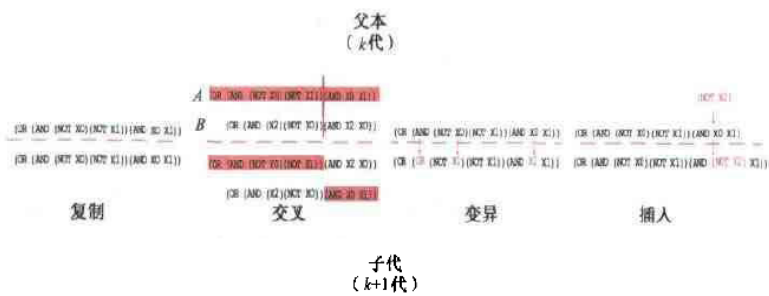
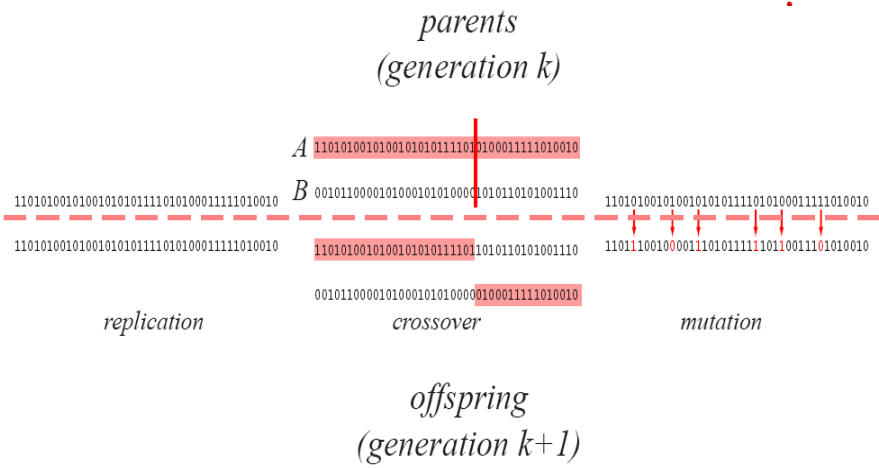


图 7-10 遗传规划的 4 种基本运算,用于将一代的片断变成下一代。复制并不改变片断。交叉是将两个片断混合或交配。其中在片断 A 的某个随机选择的允许位置截断,对片断 B 也这样,然后将 A 的前半部分和 B 的后半部分连接,另一半也如此,这样就得到两个子代片断。在变异中,随机选择的元素以小概率替换另外的元素,但是要替换的元素必须是同一类型。举例来说,数字可以换成数字,单变量运算符可以换成单变量运算符等。对于插入,一个随机选择的元素以小概率更换为相容的片断,以保证文法合法和有意义



- 遗传规划的四种基本运算：复制，交叉，变异，插入



中山大學

The Baldwin effect 鲍德温效应

- 指没有任何基因信息基础的人类行为方式和习惯，经过许多代人的传播，最终进化为具有基因信息基础的行为习惯的现象
- 表明学习会影响进化的速度，个别的学习能增加该物种的进化水准
- 太多的学习和太少的学习一样都将减缓进化速度

41



中山大學

本章小结 (1)

- When a pattern recognition problem involves a model that analytic or gradient descent methods are unlikely to work, we may employ stochastic techniques.
- Simulated annealing , based on physical annealing of metals, consists in randomly perturbing the system, and gradually decreasing the randomness to a low final level, in order to find an optimal solution.

42

本章小结 (2)



中山大學

- Boltzman learning trains the weights in a network so that the probability of a desired final output is increased.
- Some graphical models, such as hidden Markov models and Bayes belief networks, have counterparts in structured Boltzmann networks, and this leads to new applications of Boltzmann learning.

43

本章小结 (3)



中山大學

- Genetic algorithms and genetic programming, based on evolution, perform highly parallel stochastic searches in a space set by the designer.
- Chromosome, a string of bits, used in Genetic algorithms;
- A snippet of computer code used in genetic programming.

44



中山大學

结语

- 随着计算机费用的持续下降，模式分类问题将更多的借助于强大的计算能力，而不是更加精巧和细致的分类器设计来解决。在这种趋势中，进化计算方法是大有前途的方法。（Duda, 2001）