第12章：演化学习

胡成玉 |  **中国地质大学（武汉）**

机器学习与数据挖掘团队

huchengyu@cug.edu.cn

http://grzy.cug.edu.cn/huchengyu/

本章内容

一、演化学习基础知识

二、遗传算法

三、演化神经网络

四、演化学习问题与挑战

一、演化学习基础知识

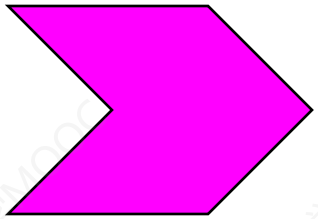
定义：演化学习基于演化算法提供的优化工具设计机器学习算法。

演化算法

- 遗传算法
- 演化规划
- ……

机器学习

- 神经网络
- 强化学习
- ……



演化学习

- 演化特征学习
- 演化监督学习
- 演化非监督学习
- 演化强化学习
- ……

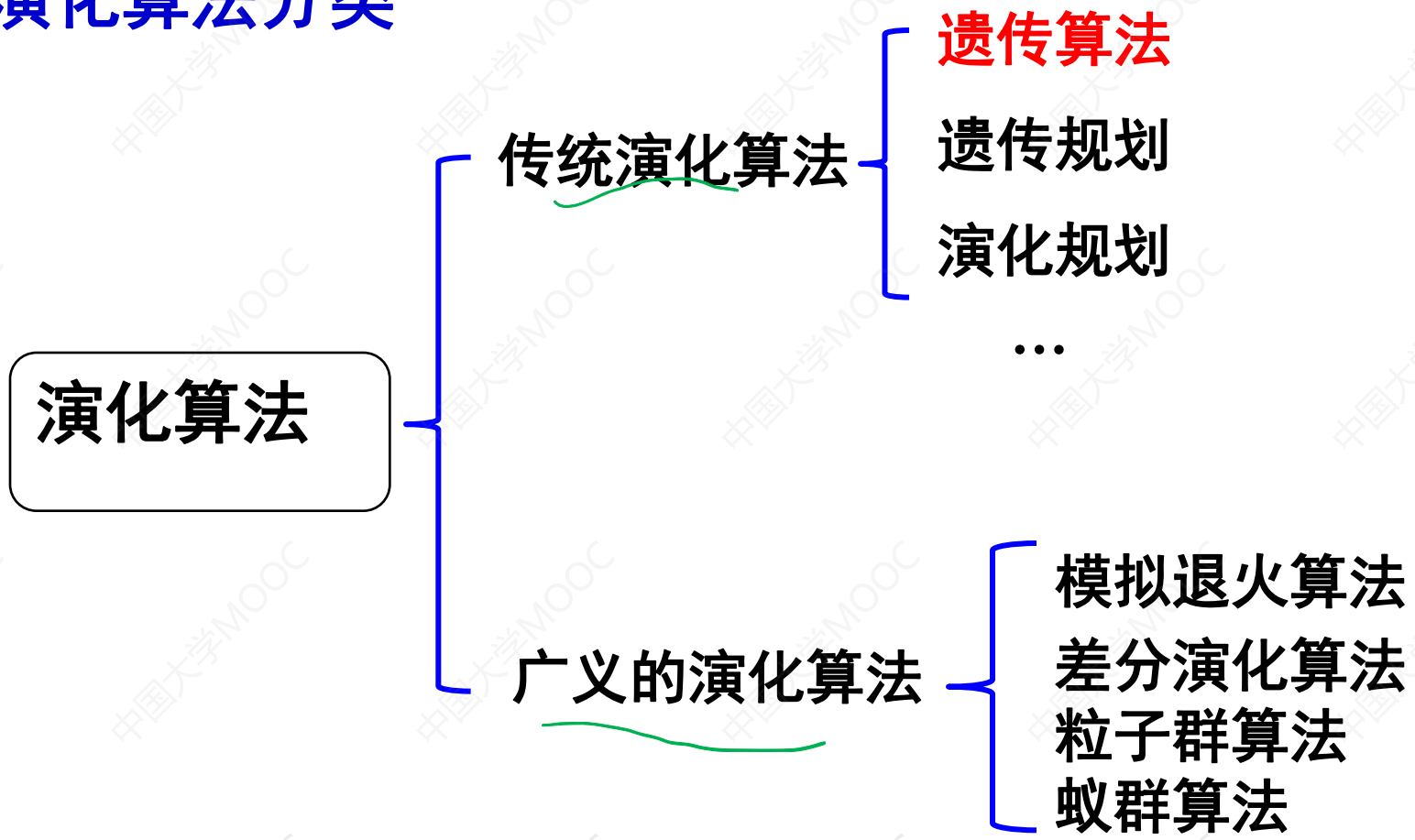
一、演化学习基础知识

演化算法：或称“进化算法”，它 是一个“算法簇”，其灵感都来自于大自然的生物进化。演化算法有很多的版本，比如，有不同的遗传基因表达方式，不同的交叉和变异算子，以及不同的再生和选择方法，与传统的优化算法相比，**演化算法的特点**在于：

- 具有高鲁棒性和广泛适应性；
- 具有自组织、自适应、自学习的特性；
- 本质并行性；
- 能够不受问题性质的限制，有效处理传统优化算法难以解决的复杂问题。

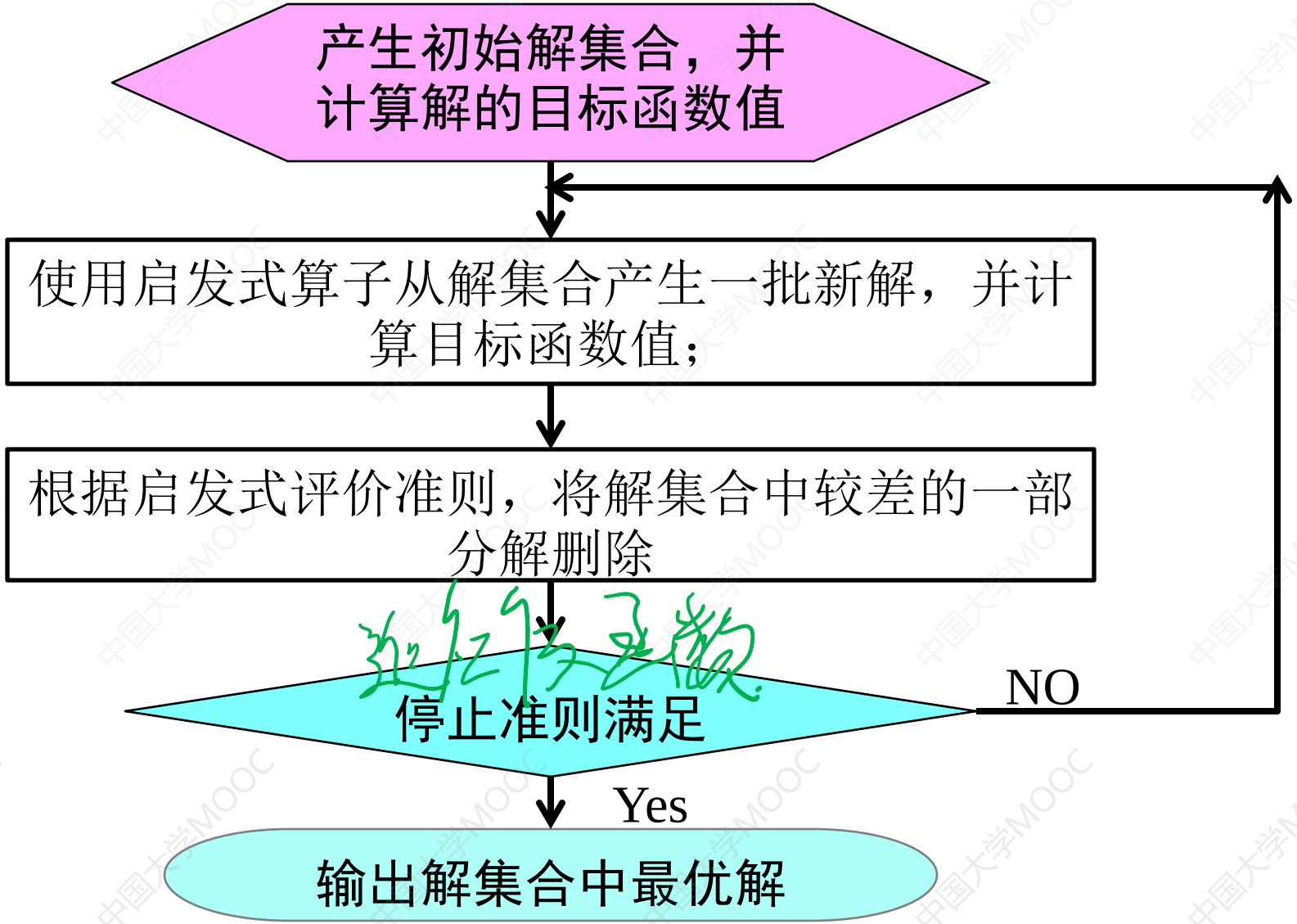
一、演化学习基础知识

演化算法分类



一、演化学习基础知识

演化算法的通用流程



二、遗传算法

定义：遗传算法是模拟生物在自然环境中的遗传和演化过程而形成的一种自适应全局优化概率搜索算法。

历史：

1. 遗传算法最早由美国密西根大学的 J. Holland 教授提出，起源于20世纪60年代对自然和人工自适应系统的研究。
2. 70年代，De Jong 基于遗传算法的思想在计算机上进行了大量的纯数值函数优化计算实验。
3. 80年代，由Goldberg进行归纳总结，形成了遗传算法的基本框架。

二、遗传算法

遗传算法的基本思想

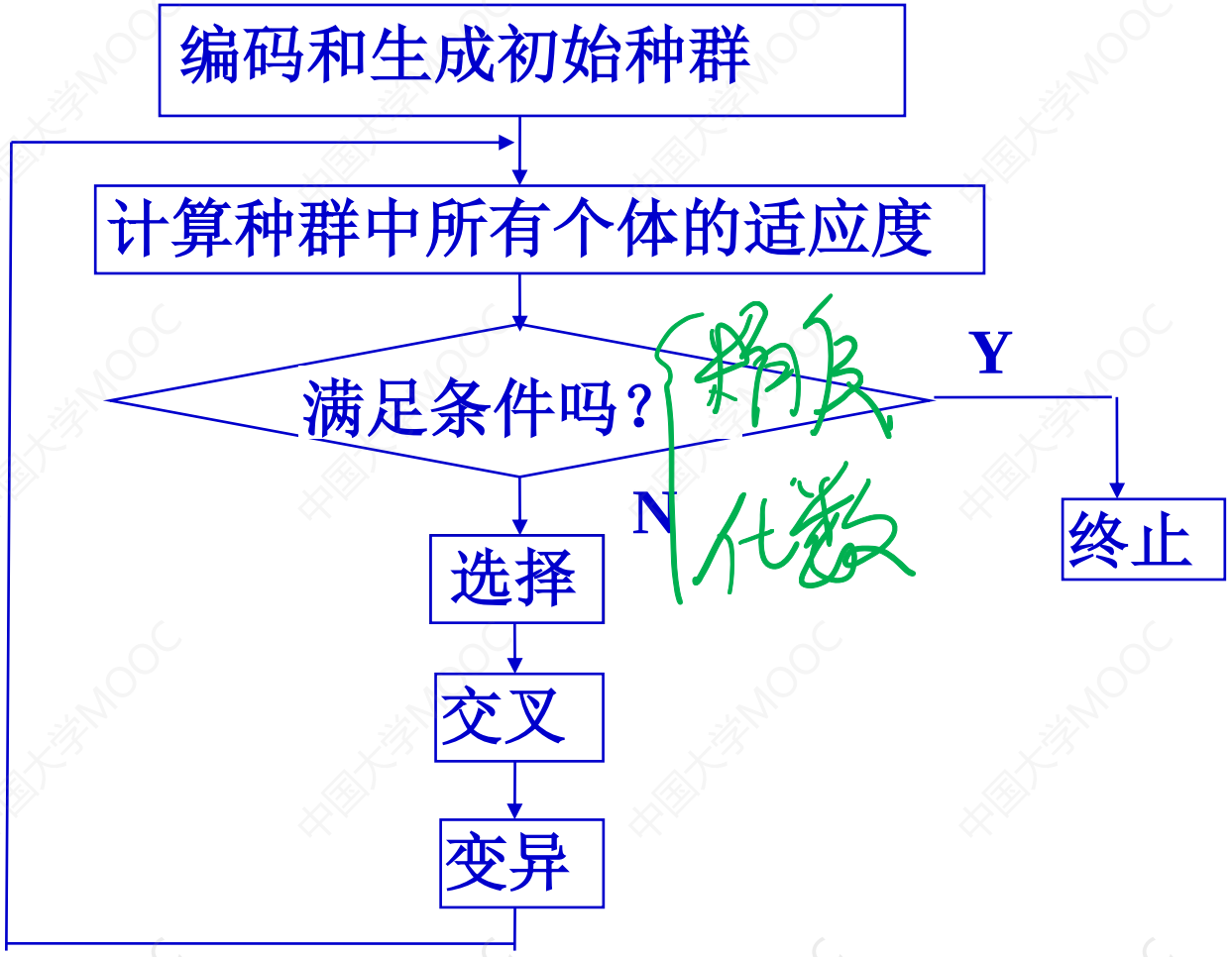
从初始种群出发，采用优胜劣汰、适者生存的自然法则选择个体，并通过杂交、变异来产生新一代种群，如此逐代演化，直到满足目标为止。

遗传算法的特点

1. 遗传算法是从问题解空间多点并行搜索，而非从单个解开始搜索；
2. 遗传算法利用目标函数的适应度这一信息而非利用导数或其它辅助信息来指导搜索；
3. 遗传算法利用选择、交叉、变异等算子而不是利用确定性规则进行操作。

三大算子

二、遗传算法



基本遗传算法的算法流程图

二、遗传算法

遗传算法编码

（1）二进制编码

二进制编码是将原问题的结构变换为染色体的位串结构。在二进制编码中，首先要确定二进制字符串的长度，该长度与变量的定义域和所求问题的计算精度有关。

例如： 假设变量x的定义域为[-2，5]，其要求精度为10E-6，则我们需要将[-2,5]分成7000000个等长小区间，每个小区间用一个二进制位串来表示，于是二进制位串长度至少23位。这是因为：

$$4194304 = 2^{22} < 7000000 < 2^{23} = 8388608$$

二进制编码存在的主要缺点：汉明悬崖。

例如，7和8的二进制数分别为0111和1000，当算法从7改进到8时，就必须改变所有的位。

二、遗传算法

遗传算法编码

（2）实数编码

实数编码是将每个个体的染色体都用某一范围的一个实数（浮点数）来表示，其编码长度等于问题变量的个数。

这种编码方法是将问题的解空间映射到实数空间上，然后在实数空间上进行遗传操作。

实数编码适应于多维、高精度要求的连续函数优化问题。

（3）有序串编码

很多组合优化问题中，目标函数的值不仅与表示解的字符串中各字符的值有关，而且与其所在字符串的位置有关，这时，需要采用独特的有序串编码，比如旅行商优化问题。

二、遗传算法

适应性度量

适应度函数是用于对个体的适应性，进行度量的函数。通常，一个个体的适应度值越大，它被遗传到下一代种群中的概率也就越大。

(1) 常用的适应度函数

原始适应度函数：直接将待求解问题的目标函数f(x)定义为遗传算法的适应度函数。

例如，在求解极值问题 $max_{x\in[a,b]}f(x)$ 时，f(x)即为x的原始适应度函数。

采用原始适应度函数

优点：能够直接反映出待求解问题的最初求解目标

缺点：是有可能出现适应度值为负的情况

二、遗传算法

适应性度量

（2）标准适应度函数：在遗传算法中，一般要求适应度函数值非负，并且，适应度值越大越好，这就往往需要对原始适应度函数进行某种变换，将其转换为标准的度量方式，以满足演化操作的要求，这样所得到的适应度函数被称为标准适应度函数$f_{Normal}(x)$。

例如：对极小化问题，其标准适应度函数可定义为

$$f_{nomal}(x)=\begin{cases} f_{\max}(x)-f(x) & \text{当} f(x)<f_{\max}(x) \\ 0 & \text{否则} \end{cases}$$

其中，$f_{\max}(x)$是原始适应函数$f(x)$的一个上界。如果$f_{\max}(x)$未知，则可用当前代或到目前为止各演化代中的$f(x)$的最大值来代替。

二、遗传算法

选择操作

选择操作是指根据选择概率按某种策略从当前种群中挑选出一定数目的个体，使它们能够有更多的机会被遗传到下一代。常用的选择策略：比例选择，排序选择，竞技选择。

比例选择：每个个体被选中的概率与其适应度大小成正比。比如在轮盘赌选择算法中，个体被选中的概率取决于该个体的相对适应度。而相对适应度的定义为：

$$P(x_i)=\frac{f(x_i)}{\sum_{j=1}^{N} f(x_j)}$$

其中，$P(x_i)$是个体x_i的相对适应度，即个体x_i被选中的概率；$f(x_i)$是个体x_i的原始适应度值；分母是种群的累加适应度值

二、遗传算法

选择操作

轮盘赌选择算法的基本思想是：根据每个个体的选择概率P(x_i)将一个圆盘分成N个扇区，其中第i个扇区的中心角为：

$$2\pi \frac{f(x_i)}{\sum_{j=1}^{N} f_i(x_j)} = 2\pi p(x_i)$$

再设立一个移动指针，选择时，假想转动指针，当指针静止时，若它指向第i个扇区，则选择个体i。

从统计角度看，个体的适应度值越大，其对应的扇区的面积越大，被选中的可能性也越大。

二、遗传算法

交叉操作

交叉操作是指按照某种方式对选择的父代个体的染色体的部分基因进行交叉重组，从而形成新的个体。

交叉重组是自然界中生物遗传进化的一个主要环节，也是遗传算法中产生新的个体最重要的方法之一。根据个体编码方法的不同，遗传算法中的交叉操作可分为二进制交叉和实值交叉两种类型。

二进制交叉是指二进制编码情况下所采用的交叉操作，它主要包括单点交叉、两点交叉和均匀交叉等方法。

二、遗传算法

变异操作

变异是指对选中个体的染色体中的某些基因进行变动，以形成新的个体。变异也是生物遗传和自然演化中的一种基本现象，它可增强种群的多样性。遗传算法中的变异操作增加了算法的局部随机搜索能力，从而可以维持种群的多样性。根据个体编码方式的不同，变异操作可分为二进制变异和实值变异两种类型。

（1）二进制变异：该变异方法是先随机地产生一个变异位，然后将该变异位置上的基因值由“0”变为“1”，或由“1”变为“0”，产生一个新的个体。例如：设变异前的个体为A=0 0 1 1 0 1，若随机产生的变异位置是2，则该个体的第2位由“0”变为“1”。变异后的新的个体是A'= 0 1 1 1 0 1 。

二、遗传算法

变异操作

（2）实值变异

◆基于位置的变异方法

该方法是先随机地产生两个变异位置，然后将第二个变异位置上的基因移动到第一个变异位置的前面。

例 设选中的个体向量C=20 16 19 12 21 30，若随机产生的两个变异位置分别时2和4，则变异后的新的个体向量是：

C'= 20 12 16 19 21 30

◆基于次序的变异

该方法是先随机地产生两个变异位置，然后交换这两个变异位置上的基因。

例 设选中的个体向量D=20 12 16 19 21 30，若随机产生的两个变异位置分别时2和4，则变异后的新的个体向量是：

D'= 20 19 16 12 21 30

三、演化神经网络

演化神经网络概述

演化神经网络是基于演化计算和神经网络两大研究方向，将二者有机融合而产生的一种全新神经网络模型。

这种模型把演化计算的自适应机制与神经网络的学习机制有机的结合在一起，有效地克服了传统人工神经网络的很多缺点。

演化神经网络模型的一个主要特点就是它对动态环境的自适应性。这种自适应性过程通过演化的三个等级实现，即连接权值和阈值、网络结构和学习规则的演化。

三、演化神经网络

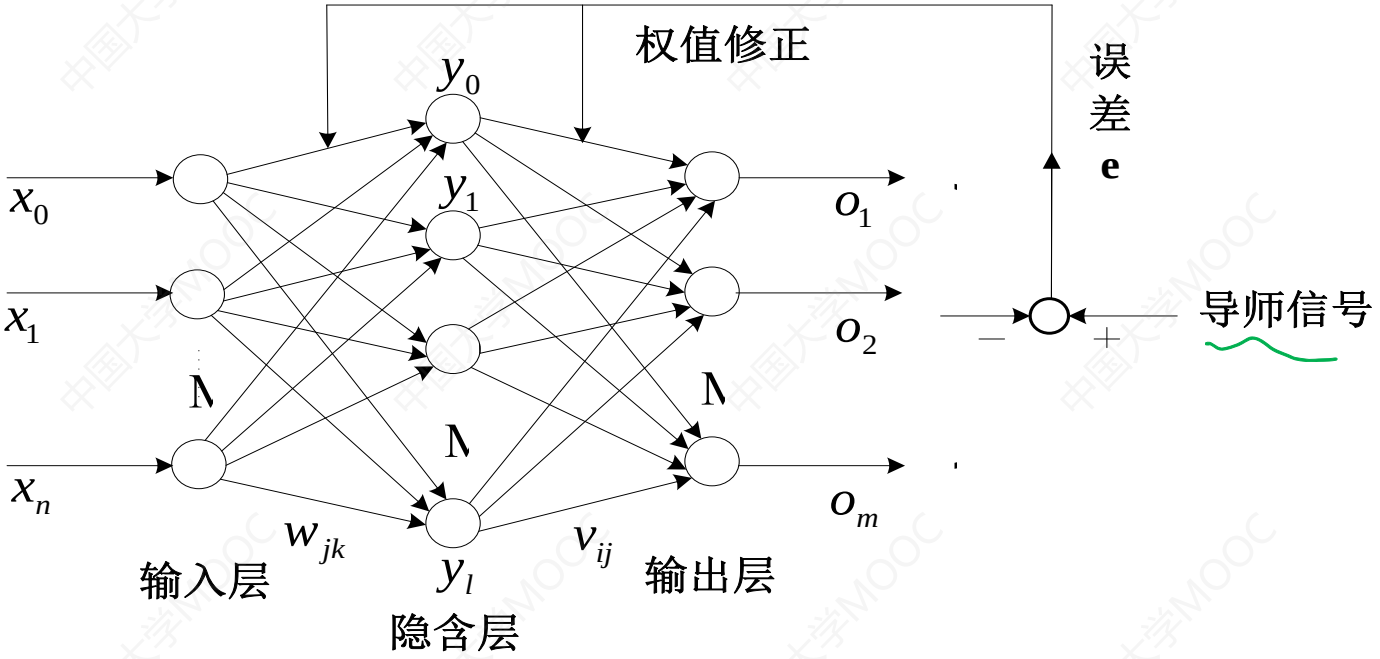
演化神经网络概述

根据演化神经网络实现的三个等级，演化神经网络模型有以下四种不同的类型：

- **初始权值和阈值演化**
- 网络结构演化
- 结构和权值阈值同时演化
- 学习规则演化

三、演化神经网络

初始权值阈值演化---以遗传算法优化BP神经网络的连接权值和阈值为例



BP网络是一类多层的前馈神经网络，目前是人工神经网络中应用最广泛的算法之一，但是存在一些缺陷，比如说学习收敛速度慢、不能保证收敛到全局极小点，网络结构不易确定。

三、演化神经网络

初始权值阈值演化

优化目标：用遗传算法优化BP神经网络的初始权值和阈值，使优化后的BP神经网络具有更好的预测精度。

算法的基本思路：

（1）对神经网络的初始权值和阈值进行编码

（2）然后对种群所有个体进行解码，生成多个神经网络

（3）对每个神经网络进行BP训练，然后以均方根误差作为评价标准，对种群中所有个体进行适应度评价。

（4）进行选择、交叉、变异操作，产生新的种群

（5）判断是否达到停止条件，否则转到（2）运行。

三、演化神经网络

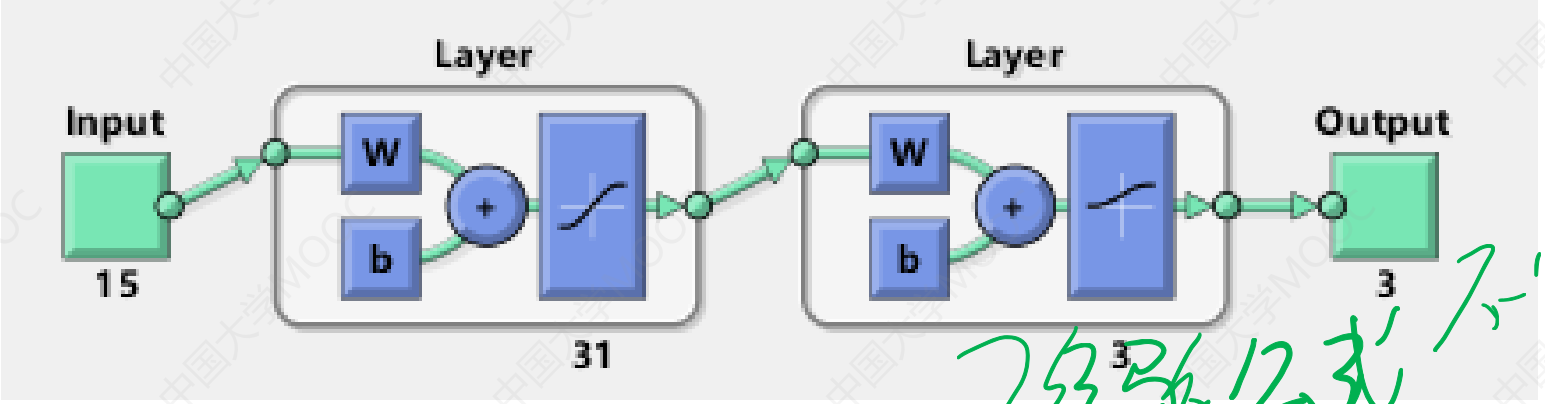
以某型拖拉机的齿轮箱为工程背景，介绍使用基于遗传算法的BP神经网络进行齿轮箱故障的诊断。

样本特征															齿轮状态
0.2286	0.1292	0.072	0.1592	0.1335	0.0733	0.1159	0.094	0.0522	0.1345	0.009	0.126	0.3619	0.069	0.1828	无故障
0.209	0.0947	0.1393	0.1387	0.2558	0.09	0.0771	0.0882	0.0393	0.143	0.0126	0.167	0.245	0.0508	0.1328	无故障
0.0442	0.088	0.1147	0.0563	0.3347	0.115	0.1453	0.0429	0.1818	0.0378	0.0092	0.2251	0.1516	0.0858	0.067	无故障
0.2603	0.1715	0.0702	0.2711	0.1491	0.133	0.0968	0.1911	0.2545	0.0871	0.006	0.1793	0.1002	0.0789	0.0909	齿根裂纹
0.369	0.2222	0.0562	0.5157	0.1872	0.1614	0.1425	0.1506	0.131	0.05	0.0078	0.0348	0.0451	0.0707	0.088	齿根裂纹
0.0359	0.1149	0.123	0.546	0.1977	0.1248	0.0624	0.0832	0.164	0.1002	0.0059	0.1503	0.1837	0.1295	0.07	齿根裂纹
0.1759	0.2347	0.1829	0.1811	0.2922	0.0655	0.0774	0.02273	0.2056	0.0925	0.0078	0.1852	0.3501	0.168	0.2668	断齿
0.0724	0.1909	0.134	0.2409	0.2842	0.045	0.0824	0.1064	0.1909	0.1586	0.0116	0.1698	0.3644	0.2718	0.2494	断齿
0.2634	0.2258	0.1165	0.1154	0.1074	0.0657	0.061	0.2623	0.2588	0.1155	0.005	0.0978	0.1511	0.2273	0.322	断齿

一共9个训练样本，每个样本为15维特征向量，所以神经网络输入层15个神经元。齿轮有3种故障模式，分别为无故障，齿根裂纹断齿，所以输出层为3个神经元。

三、演化神经网络

初始权值阈值演化-----BP神经网络结构



- 由于样本有15个输入参数，3个输出参数，所以输入层15个节点，隐含层有2*15+1=31个节点，输出层有3个节点。隐含层采用S型正切函数tansig()，输出层神经元输出采用S型对数函数logsig()。
- 权值个数：15*31+31*3=558个，阈值个数：31+3=34个，所以，因此，遗传算法需要优化的参数个数有558+34=592个

史峰等，智能算法30个案例分析第三章，北京航空航天大学出版社，2010

三、演化神经网络

初始权值阈值演化-权值阈值的编码和解码

- 编码。由于神经网络的权值与阈值为[-0.5,0.5]区间的随机数，在本例子中，每个权值与阈值按照10位二进制编码，二进制串的长度为5920。其中前4650位是输入层与隐含层连接权值编码，4651-4960为隐含层阈值编码，4961-5890位是隐含层与输出层连接权值编码，5891-5920位是输出层阈值编码。
- 解码。将每个权值与阈值转化为[-0.5,0.5]区间的实数。
- 精度。权值和阈值的精度可以达到2^{-10}。

输入层与隐含层连接权值	隐含层阈值	隐含层与输出层阈值	输出层阈值
465	31	93	3

三、演化神经网络

初始权值阈值演化-种群初始化

- 设定种群大小为100，每个个体二进制编码长度为5920。

初始权值阈值演化-适应度值计算

- 对每个个体解码，得到592个权值和阈值，每个权值和阈值的区间为[-0.5,0.5]。
- 将权值和阈值赋给BP网络。
- 使用训练样本训练BP网络。
- 使用回代法计算BP网络的均方根误差。
- 适应度函数采用排序的适应度分配函数,对所有网络的均方根误差排序。

三、演化神经网络

初始权值阈值演化---选择算子

- 采用随机遍历抽样算法选择个体，生成100个新个体

初始权值阈值演化---交叉和变异

- 交叉：从种群中随机选择两个个体，利用单点交叉算子产生后代，循环50次后，产生新种群
- 变异：以一定的概率产生变异基因数，用随机方法选出发生变异的基因，如果所选的基因编码为1，则变为0；反之，则变为1。

遗传算法的运行参数

种群大小	遗传代数	个体的二进制位数	交叉概率	变异概率
100	50	10	0.7	0.01

三、演化神经网络

为了判断优化后的初始权值及阈值的好坏，另外再给三组新的数据作为BP网络的测试数据。

测试样本数据

样本特征值															齿轮状态
0.2101	0.095	0.1298	0.1359	0.2601	0.1001	0.0753	0.089	0.0389	0.1451	0.0128	0.159	0.2452	0.0512	0.1319	无故障
0.2593	0.18	0.0711	0.2801	0.1501	0.1298	0.1001	0.1891	0.2531	0.0875	0.0058	0.1803	0.0992	0.0802	0.1002	齿根裂纹
0.2599	0.2235	0.1201	0.0071	0.1102	0.0683	0.0621	0.2597	0.2602	0.1167	0.0048	0.1002	0.1521	0.2281	0.3205	断齿

三、演化神经网络

初始权值阈值演化---结果分析

1、使用随机权值和阈值
测试样本预测结果：

Y1 =

0.9900	0.0354	0.0206
0.0525	0.8644	0.0044
0.0234	0.1734	0.9995

测试样本的仿真误差:0.22349

2、使用优化后的权值和阈值
测试样本预测结果：

Y2 =

0.9791	0.0341	0.0778
0.0340	0.9709	0.0095
0.0222	0.0598	0.9796

测试样本的仿真误差:0.088481

四、演化学习问题与挑战

存在的问题：对演化算法这类随机性启发式优化算法而言，其理论研究不足，比如优化效率高低、与最优解的逼近程度如何、启发式算子效用评估等问题难以有严格的答案，这导致了演化学习也缺乏有效的理论解释。

挑战：近些年，一方面学习模型变得复杂、数据增长迅速、一方面对模型训练时间有严格约束，如何使得演化学习能够进行有效、快速地优化，还有待深入的研究。