

基于免疫的 RBF 网络在线学习算法研究

潘磊, 王伟智

(福州大学自动化研究所, 福建 福州 350002)

摘要: 提出一种新的基于免疫的 RBF 网络在线算法. 首先是融入增加结点策略和减少结点策略, 把传统的免疫 RBF 网络算法改进成在线学习算法. 其次是改进了权值学习算法, 径向基函数相当于这一类的概率密度, 隐层到输出层权值相当于这一类的值. 用这种方法权值不需要训练. 试验结果表明, 该方法效果理想、速度快, 识别率高.

关键词: 免疫; RBF; 在线学习; RAN 算法

中图分类号: TP183

文献标识码: A

A new on-line training algorithm of RBF network based on the immunity

PAN Lei WANG Wei-zhi

(Institute of Automation, Fuzhou University, Fuzhou, Fujian 350002, China)

Abstract Put forward a new network of RBF of based on the immunity which is a on-line algorithm. The first is adding strategy and remarkable degree based on the output of radial basis function which is a pruning strategy. The second is improval of the weight value study algorithm. radial basis function is equal to probability of this type, weight is equal to value of this type. Weight doesn't need to train. Experiment results show that this method is ideal, speed is quick, and detection rate is high.

Keywords immunity, RBF, on-line training, RAN algorithm

径向基神经网络现有的各种训练算法一般都是采用所谓离线的方式. 如果对象特性没有取遍所有最有代表的特征样本, 径向基网络离线训练方式就无法进行有效地逼近. 基于这种认识, 我们进一步研究了径向基网络的在线学习算法. 本文参考了其它在线径向基神经网络增加结点和减少结点的策略, 把 RAN 算法(径向基神经网络的一种在线学习算法)中的增加结点策略中的“新性”条件和一种基于径向基函数输出值显著程度的减少结点策略结合进传统的免疫 RBF 网络算法, 从而把传统的免疫 RBF 网络算法改进成一种在线学习算法.

径向基神经网络学习算法包括两部分: 一是隐层径向基函数的中心和宽度的确定; 二是隐层到输出层权值的确定. 第一个是学习训练算法的关键. 第二个相对第一个次要些, 却占用不少时间. 笔者在这个方面把径向基函数和隐层到输出层权值与概率中的数学期望公式进行了类比, 径向基函数相当于这一类的概率密度, 隐层到输出层权值相当于这一类的值. 这样, 不但使整个网络更具有很好理解的意义, 还使权值确定和调整更容易了.

1 相关技术

1.1 免疫系统基本理论和相关的学习算法

1) 免疫的定义. 免疫系统是机体对外来的抗原(病毒)进行识别, 通过产生免疫应答(比如产生 B 细胞(抗体的一种))将其清除, 以维持机体内环境相对稳定的一种生理反应.

2) 免疫系统的特点是: ①识别自己与非己; ②联想记忆; ③学习和优化; ④自组织调节.

收稿日期: 2008-06-12

作者简介: 潘磊(1983-), 男, 硕士研究生; 通讯联系人: 王伟智, 研究员.

基金项目: 国家自然科学基金资助项目(60675058); 福建省自然科学基金资助项目(A0610013, A0710008)

1.2 径向神经网络基本理论

径向神经网络 (radial basis function, 简称 RBF) 有很深的数学理论基础, 它能以任意精度逼近任意连续函数. 径向神经网络只有一个隐层, 其激励函数为径向基函数. 该函数一般为高斯函数:

$$\text{radbas}(x) = \exp\left\{ - 0.5 \left\| x - x_c \right\|^2 / \sigma^2 \right\}$$

其中: x_c 为径向基函数的中心; σ 为径向基函数的半径. 初始权值向量 W 都为 1.

2 RBF 在线学习算法的基本原理

2.1 原有免疫算法用于学习 RBF 隐层中心^[1]

根据抗原 (待分类的数据) - 抗体 (中心) 的亲合力, 选出一部分最好抗体. 抗体克隆出若干个体, 并使其发生变异, 从而形成下一代群体. 群体通过克隆抑制 (去除中心记忆集相似度过大的数据中心)、免疫抑制 (去除中心合并集相似度过大的数据中心) 选出一些最好个体加入中心记忆集, 并用记忆集中的一些个体替换抗体. 用随机产生数据中心替换亲和力较低的数据中心 (体现免疫的自组织特性). 详细流程见文献 [1].

2.2 增删策略的引入及免疫算法的修改

通常的免疫 RBF 网络都是把 de Castro 提出的人工免疫网络 (aNet) 和 RBF 结合. de Castro 提出的人工免疫网络把克隆选择算法和免疫网络理论结合在一起^[2, 3]. 笔者认为, 免疫网络理论具有基本网络结构方面的功能, 比如说免疫网络理论中的免疫抑制相当于基本网络的删除策略. 克隆选择算法也存在一些不足, 比如步骤多和一些参数需要人工定义, 带来操作上的一些不便. 如果已经有很多的经验参数会比较好, 反之经验或是训练实例不丰富的时候就不好使用.

本研究把人工免疫网络中的免疫网络理论去掉, 并且把克隆选择算法中的克隆抑制去掉, 换成一种 RBF 网络的删除策略. 删除策略是一种简易有效的方法, 较之前者免疫网络理论更加易于操作, 克服了计算步骤多的缺点.

改进后的算法相当于修改后的克隆选择算法加增删策略. 克隆选择算法和遗传算法的两点区别是: 遗传算法更多的是强调全局搜索, 而忽略局部搜索; 克隆选择算法则更多的两者兼顾. 在遗传算法中, 变异是次要算子, 而在克隆选择算法中是主要算子.

以下介绍具体的增删策略.

1) 隐层结点的增加策略. 借用 RAN 算法^[4]训练样本的“新颖性”条件. 该方法当新样本满足“新性”条件时分配一个新节点.

新性条件为: ①距离准则: 当前样本输入距离最近的数据中心超过某一定值 ϵ ; ②误差准则: 网络的输出与样本输出的偏差大于某一定值 ϵ_{min} . 新性条件两个同时满足则分配新的隐节点. 这两个“新性”条件在免疫算法中具体体现为两个公式:

条件 1: $A(k, G^l) < \delta_{in}$;

条件 2 $I_k < \epsilon_{min}$.

式中: $A(k, G^l) = \max A(k, G^l)$, 即为 G 中与 $x(k)$ 亲合度最大的记忆单元 G^l ; δ_{in} 为记忆阈值; ϵ_{min} 为免疫力阈值. 条件 1 成立, 表示现有免疫记忆单元不能充分识别新侵入的抗原; 条件 2 成立, 则意味着目前的 B 细胞集不能很好地清除新侵入抗原. 若条件 1 和条件 2 都成立, 说明满足“新性”条件, 分配一个新节点.

2) 隐层结点的减少策略. 径向神经网络的隐层结点可以看成是每一类的模式, 也就是有几类就有几个隐层结点. 隐层结点径向基函数选取的是高斯核函数, 它形式上和正态分布函数一样, 不妨把径向基函数的输出值看成是匹配这个隐层结点 (即这一类) 的概率 (或叫可能性).

如果一个隐层结点的径向基函数输出值在一段时间之内都十分不显著, 说明这个隐层结点对结果的贡献不大, 是冗余的, 应当考虑删除. 这个时间长度可以用学习样本数据的个数来衡量, 比如连续 N_d 个输入, 显著程度可以用是否低于某个阈值来衡量.

2.3 基函数和权值

本文的基函数及权值和一般的 RBF 网络不同. 一般的 RBF 网络隐层函数是:

$$\text{radbas}(x) = \exp\left\{ - 0.5 \left\| x - x_c \right\|^2 / \sigma^2 \right\}$$

本文提出用高斯分布函数来作为径向基函数:

$$\text{radbas}(x) = 1/(2\pi\sigma)^{d/2} \exp\left\{-0.5 \frac{\|x - x_c\|^2}{\sigma^2}\right\}$$

每一个隐层结点代表一类模式, 而结点径向基函数代表属于这一类模式的概率密度函数. 该隐层结点到输出端的权值代表这一类模式的值. 所以, 本文的基函数和权值较一般 RBF 网络更具有物理现实意义. 而最后的输出值等于每一层高斯分布函数输出值乘以这一层相应的权值的总和, 即 $\sum \omega \cdot \text{radbas}(x)$. 这个有点类似于数学期望公式. 但隐层结点要取遍样本空间的所有关键样本, 所以刚开始效果会差一点, 越往后效果越好. 为此, 要引入免疫算法来配合 RBF. 免疫算法具有联想记忆和自组织调节的特点, 相当于在目前最优点处附近搜索更优的点. 免疫算法用于径向基神经网络训练中心. 输出端用公式 $(\sum \omega \cdot \text{radbas}(x)) / (\sum \text{radbas}(x))$.

本文的权值学习不同于一般的权值学习. 一般的权值学习基于 delta 学习原则, 权值为一个最小二乘解, 而本权值在系统初始化的时候就设定为这一类的值. 增加隐层结点的时候权值也就学习好了, 即为这个增加结点这一类的值. 一般的 RBF 网络训练分两步, 先选中心和再调权值. 本文把后一步省去, 所以权值学习的时间等于增加结点的时间, 比其它方法权值学习大大缩短时间.

3 RBF 在线学习算法的实现

设在每个递推步 k ($k=1, 2, \dots, t, \dots$) 输入样本 $k[x(k), y(k)]$, 具体学习算法如下:

1) 系统初始化. 以一个确定的数量进行随机初始化. 刚开始用随机方法确定隐层中心, 隐层到输出层的权值用 $y(k)$ 来表示.

2) 检验是否加入新隐层节点 (B 细胞). 输入新样本 k , 判断是否满足“新性”条件, 如果满足, 算法执行步骤 3), 诱导 B 细胞克隆扩增产生新的 B 细胞 (隐层节点).

3) 新隐层节点 (B 细胞) 的在线生成^[5].

① 选择和 $x(k)$ 亲和力最大的数据中心, 复制 N_c 个, 产生数据中心的复制集 L .

② 应用下面公式对 P 个相同的数据中心进行变异处理, 形成变异集 L' , α_k 为变异率, 反比于亲和力. 这一过程实际上是在有最大亲和力的数据中心附近进行搜索以得到更有亲和力的数据中心.

$$C'_k = C_k - \alpha_k(C_k - X_i) \quad (k=1, 2, \dots, p)$$

③ 计算变异集 L' 中与 $X(k)$ 的亲和力, 选出亲和力最大的 N_m 个数据中心生成数据中心集 L'' , 同时产生新增候选集 B , 新增隐层节点转化为在候选集 B 中寻找某最优隐层节点 B_h (数据中心为 C_h , 宽度为 σ_h), 将其加入到原有 RBF 网后, 对样本 $[x(k), y(k)]$ 的逼近效果最好, 即

$$F_k(B_h) = \max F_k(B_b) \quad (b \text{ 从 } 1 \text{ 到 } N_m)$$

$$F_k(B_b) = 1 \left\{ 1 + \left| y(k) - \sum_{j=1}^{p-1} w_j \phi_j(x(k)) - w_b \phi_b(x(k)) \right| \right\} \quad (b \text{ 从 } 1 \text{ 到 } N_m)$$

考虑到记忆集合实际是在数据中心附近搜寻而产生, 因此, $\forall b, 1 \leq b \leq N_m$, 令 W_b 为原有 RBF 网中数据中心隐层节点到输出层节点之间的连接权值.

4) 检验是否删除冗余隐层节点. 如果连续多个样本都无法激活某隐层节点, N_d 为隐层节点 i 连续未被输入样本激活的次数, 则隐层节点 i 从 RBF 网中删除. 这一过程实际上是模拟了免疫应答中低激励值 B 细胞的凋亡. 为防止误删有用的隐层节点, N_d 应取较大的值.

4 仿真实例

地下水位仿真实例来自文献 [6]. 实例部分见表 1.

表 1 地下水位及其影响因子监测数据表

Tab 1 Data table of underground water level and the impact factor

序号	河道流量	气温	饱和差	降水量	蒸发量	测量水位
1	0.0752	0.83335	0.85	0.0995	0.8116	0.17685
⋮	⋮	⋮	⋮	⋮	⋮	⋮

通过河道流量、气温、饱和差、降水量、蒸发量 5 个监测因子来估计地下水位. 从资料中归纳出 24 组数据. 1~ 5 组为训练样本. 6~ 24 组为测试样本.

本 RBF 神经网络的输入层 5 个神经元分别对应 5 个监测因子. 随机选取第 1~ 4 个例子做为隐层节点 A~ D. 河道流量等 5 个监测因子作为为径向基函数的中心, 所以径向基函数的中心有 5 个分量. 训练样本中的水位值直接作为隐层到输出层的权值.

用 EXCEL 做仿真实验. 依次输入样本 1, 2, ..., 第 14 个样本与 4 个隐层结点的最大亲和力为 0.679, 小于阈值 0.68, 免疫力为 0.711, 小于阈值 0.72, 满足“新性”条件. 用免疫算法进行新隐层节点的在线生成.

利用新隐层节点 (B 细胞) 的在线生成找出最优的节点, 结果见表 2

表 2 最优的节点
Tab 2 Optimization node

河道流量	气温	饱和差	降水量	蒸发量	权 (水位)
0.0885	0.303	0.52	0.0753	0.2319	0.6201

输入第 14 个样本之前, 隐层的第一个结点 A 已经连续 10 次没有被激活, 也就是说隐层的第一个结点对结果在相当长的一段时间内没有贡献, 应考虑删去. 判断是否激活的方法是输入径向基函数, 值如果小于阈值 ϵ , 说明没有被激活. 阈值 ϵ 取 0.1, 于是用最优的节点代替结点 A.

又用随机算法和经典的 K 均值算法对同一个实例做仿真与本算法做对比, 结果见表 3 可以看出, 本算法比随机算法和 K 均值算法逼近效果要好.

表 3 效果比较

Tab 3 Compare the effect

算法	随机算法	K 均值	online- ai
平均逼近效果	79.18%	81.94%	90.35%

由于初始训练样本较少, 随机算法和 K 均值算法的逼近效果只能达到 80% 左右. online- ai- RBF

算法因为采用了高斯分布函数作为径向基函数, 输出值等于每一类的权值乘以这一层的高斯分布函数的加权和. 所以, 刚开始的时候效果一般, 随着训练的增加, online- ai- RBF 算法有效地去除了冗余的隐层节点, 加进了效果更好的隐层节点, 使得隐层节点更加接近所有的关键样本, 使得输出层的公式更加逼近真实值. 仿真结果表明, online- ai- RBF 算法的平均逼近效果可以达到 90% 以上. 与文献 [5] 的方法比较, 不仅减少了权值的训练, 缩短了时间, 而且在删除策略上更易于操作.

参考文献:

[1] 朱大奇, 史慧. 神经网络原理及应用 [M]. 北京: 科学出版社, 2006: 133- 134.
 [2] de Castro L N, Von Zuben F J. Data mining: a heuristic approach [M]. [s.l]: Idea Group, 2001: 231- 257.
 [3] 左兴权, 李士勇. 人工免疫系统研究的新进展 [J]. 计算机测量与控制, 2002, 10(11): 701- 705.
 [4] 李彬. 径向基函数神经网络的学习算法研究 [D]. 济南: 山东大学, 2005.
 [5] 林金星, 沈炯, 李益国. 基于免疫原理的径向基函数网络在线学习算法及其在热工过程大范围工况建模中的应用 [J]. 中国机电工程学报, 2006, 26(9): 14- 19.
 [6] 飞思科技产品研发中心. 神经网络理论与 MATLAB7 实现 [M]. 北京: 电子工业出版社, 2005: 119.

(责任编辑: 顾泉佩)