

模糊神经网络学习样本的选取与网络扩展能力研究^{*}

Study on Pattern Selection and Generalization Ability for Fuzzy Neural Networks

许少华¹ 梁久祯¹ 何新贵²(大庆石油学院 安达151400)¹(北京系统工程研究所 北京100101)²

Abstract This paper deals with lower learning convergence and generalization of neural networks caused by shortage of pattern set. Considering contradiction in a pattern set that may be brought by rough selection, we propose a strategy in which some pattern's regimentation are remedied according to their fuzzy value. Experiments illustrate that this method is suitable to the case that some patterns are wrongly divided. On the other hand, the uncompleted pattern input vectors are a key factor leading to lower generalization of neural networks. In this paper we also consider the influence of adding new input heft on neural networks. Test results proved the validity of this work.

Keywords Neural networks, Fuzzy, Learning, Pattern, Generalization, Logging explanation

1. 引言

目前,有关神经网络的研究日趋走向成熟阶段,早在1989年,仅含一个隐层的BP网络就被证明可以逼近任意的连续函数^[1-3],从而揭示了神经网络强大的计算能力。另一方面,神经网络方法用于问题求解无需建模,给实际问题的求解过程带来很大的方便。对于网络本身的建立,已有许多学习算法,目前使用比较多的是BP学习算法或其变形的形式。但是,网络在学习阶段普遍存在局部极值和学习速度较慢等问题^[4]。针对这两个问题,自反传播学习算法发明以来,许许多多的

学者尝试用各种方法解决,如出现了快速BP算法、模拟退火算法、遗传算法等。这些算法在求解极值问题上有了很大的改进,同时也有各自的缺陷。如快速BP算法避免不了局部极值问题;模拟退火算法虽然一概率1收敛于全局极值,但一般学习速度很慢;遗传算法受编码长度的限制,对于规模较大或精度较高的问题难以应付。

实际问题的求解除了上述问题之外,在很大程度上网络的收敛速度和精度与学习使用的样本集的特点有关。由于以往的学者在采用实验样本时常以“经典”问题作为样本集,如XOR问题、6奇偶数问题等,所使

^{*}黑龙江省自然科学基金资助课题。

素空间上较容易地发现。

·得到复杂的解:通过胚胎生成,简单的基因型可以映射成复杂的表现型。

·重复性:胚胎生成在基因型中使用可重复性结构、如子程序等,可提高算法生成可行解的能力。

·适应性:在进化过程中保证遗传操作产生有效解、满足约束条件,修正失效的功能。

结论 目前研究者们使用的胚胎生成主要有三种类型:外部的、显式的、隐式的,这三种类型各有优缺点,在解的质量、基因型的大小及执行效率上也有所不同,针对具体问题,应该设计有效合理的胚胎生成。生物学家一直提倡胚胎生成的思想,进化计算的研究者也看到了它的潜在优势,尤其在计算机进化设计的领域中研究者们越来越对其感兴趣。

· 94 ·

参考文献

1. Told S, Latham W. Evolutionary Art and Computers. Academic Press, 1992
2. Bentley P J, Wakefield J P. Generic Evolutionary Design. In: Chawdhry P K, Roy R, Pant, R K, eds. Soft Computing in Engineering Design and manufacturing. Springer Verlag, Part 6, 289~298
3. Koza J, et al. Genetic Programming II. San Francisco, CA: Morgan Kaufmann, 1999
4. Grizou P. Cellular Encoding of Genetic Neural Networks. [Technical report 92-21] Laboratoire de l'Informatique du Parallélisme, Ecole Normale Supérieure de Lyon, May 1992
5. de Garis H. An Artificial Brain. New Generation Computing v12, 2, Springer-Verlag, 1994
6. Bentley P J. Evolutionary Design By Computers. Morgan Kaufmann Pub. 1999

用的样本自身的独立性较强。实际中大量的问题样本之间的相关性较大,有的问题由于样本没有经过精心的筛选,致使在同一个训练集中若干个样本出现矛盾现象,即在具有相近输入向量的两个或多个样本的输出相异,这种现象会严重影响网络的学习,本文重点考虑样本的选取问题,并用实验验证样本处理的效果。另一方面,在样本输入向量中对样本类型影响的因素考虑不全是降低网络泛化能力的一个关键因素。文中也考虑加入新的输入分量对网络的泛化能力的影响。

2. 模糊神经网络及其学习算法

为确定起见,我们考虑一种正则化模糊神经网络(NFNN),用 μ_{A_k} 表示输入变量 x_i 的第 k 个节点的隶属函数,并假定 x_i 具有 n_i 个用于模糊划分的项节点,即对应于输入变量 x_i 的模糊隶属函数个数为 n_i ,则一般的NFNN结构如图1所示。

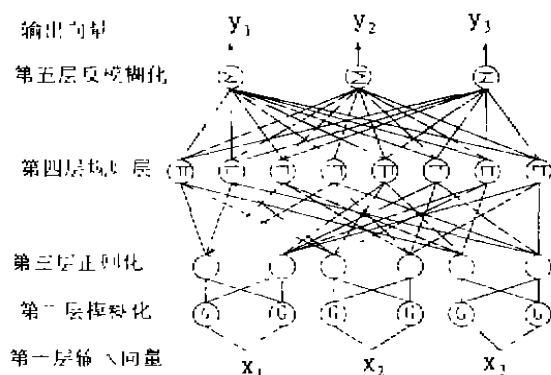


图1 正则化模糊神经网络结构图

第一层:输入层,该层的输入向量可以是精确的数值向量,也可以是模糊量。

第二层:模糊化层。该层使用高斯函数作为隶属函数,第一层的 x_i 与该层相应的第 j 个节点的输出为:

$$\mu_{A_{kj}}(x_i) = \exp\left(-\frac{(x_i - m_{kj})^2}{\sigma_{kj}}\right) \quad (1)$$

式中, m_{kj} 和 σ_{kj} 分别表示 A_{kj} 的均值和方差,令 $N = \sum_{i=1}^n n_i$,而 $j=1, 2, \dots, n_i$,该层有 N 个节点。

第三层:正则化层,该层对第一层的输出进行正则化操作,即

$$\mu'_{A_{kj}}(x_i) = \frac{\mu_{A_{kj}}(x_i)}{\sum_{j=1}^{n_i} \mu_{A_{kj}}(x_i)} \quad (2)$$

式中, $\mu'_{A_{kj}}$ 表示 $\mu_{A_{kj}}$ 的正则化输出。该层节点数与第二层节点数相同,即 N 个节点。

第四层:规则层。该层将前件(正规化节点)和结论节点(输出节点)连接起来,连接的规则是:每个规则节点仅与属于同一个输入分量的一个规则节点发生连接。因此,NFNN网络的初始化结构中共有 $K = \prod_{i=1}^n n_i$ 个规则节点,第 k 个规则节点的输出为:

$$z_k = \prod_{i=1}^n \mu'_{A_{ki}}(x_i) \quad (3)$$

其中, $s=s(k, i)$,且 $1 \leq s \leq n_i$,由下式确定:

$$1) s(k, i) = (k + \sum_{j=1}^{i-1} n_j' - \sum_{j=1}^{i-1} s(k, j) n_j^{i-1}) / n_i^{i-1} \mod n_i \quad (4)$$

2) 若 $s(k, i) = 0$,置 $s(k, i) = n_i$ 。

第五层:反模糊化层。所有第四层的规则节点都与该层输出节点连接。该层完成中心平均反模糊化操作。设该层有 P 个节点,第 j 个输出分量为:

$$y_j = \sum_{k=1}^K b_{kj} z_k = \sum_{k=1}^K b_{kj} \prod_{i=1}^n \mu'_{A_{ki}}(x_i) \quad (5)$$

用梯度下降法学习调整NFNN的参数,极小化误差函数为:

$$E = \frac{1}{2} \|D - Y\|^2 \quad (6)$$

式中, Y 是NFNN的实际输出向量, D 为期望输出向量。

若 w_{ij} 是待调参数,学习规则为:

$$\begin{aligned} w_{ij}(t+1) &= w_{ij}(t) - \eta \frac{\partial E}{\partial w_{ij}} - \alpha \Delta w_{ij}(t) \\ \Delta w_{ij}(t) &= w_{ij}(t) - w_{ij}(t-1) \end{aligned} \quad (7)$$

式中, η 为学习速度; α 为惯性系数。

3. 学习样本的选取方法

样本集的分布特征对网络的训练有很大的影响,分布均匀的样本集在网络训练时收敛速度较快,网络的泛化能力较好。相反,分布不均匀或异类样本在样本空间中距离较近的情况就不利于网络的学习。如下图所示:在图2中均匀分布的9个两类样本只需用四个超平面就可以分割,而在图3中同样是9个两类样本,由于分布不均则至少需要6个超平面来分割。

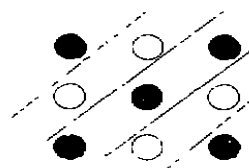


图2

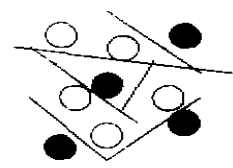


图3

实际中,许多异类样本在样本空间中距离较近的情况是由白噪引起的,有些是样本本身的选取发生了错误。因此,这些样本应处理掉,否则将影响网络的学习。

习和泛化能力。有关去噪的方法很多,我们给出一种按模糊隶属函数值相近样本向量类别矫正策略,该方法对于样本错误划分的情况很有效。

设 A 为所考虑原始样本集,其中样本个数为 M , B 为处理后的样本集, A 中有部分样本被错误划分,下面是按模糊隶属函数值相近的样本向量类别矫正方法。

1) 统计样本中心向量矩阵 $m = (m_{ij})_{n \times n}$ 和方差向量矩阵 $\sigma = (\sigma_{ij})_{n \times n}$, 并按(1)式计算 $\mu_{ij}(x)$, $i = 1, 2, \dots, n; j = 1, 2, \dots, n$ 。

2) 随机从 A 中抽取一个样本 x , 计算其他样本 $z \in A \setminus \{x\}$ 与 x 的模糊距离 $d = \sum_{i,j} |\mu_{ij}(z) - \mu_{ij}(x)|$, 若 $d < T$ (阈值) 则将 z 放入 B 集合, 重复此过程直到遍历所有 $z \in A \setminus \{x\}$, 将 x 放入 B 集合。

3) 统计 B 集合中样本的类型, 设属于 a 类的样本个数最多。

4) 将 B 集合中样本的类型改写为 a 类, 置 $A = A \setminus B$ 。

5) 若 A 空结束, 否则返回2)。

注意: 阈值 T 的选取随问题的特点而发生变化, 若取得太小则起不到样本类别矫正的目的, 网络的学习不会有太大的改进; 取得太大, 虽然网络学习很快, 但却容易发生类别过矫正错误。

4. 网络的泛化能力

神经网络的泛化能力问题一直受到其应用领域的重视。网络的逼近能力解决了计算能力问题, 而网络的泛化能力则标志着网络的实用能力问题。网络的泛化能力与网络的结构和样本集的选取有关。在网络的结构一定的情形下, 样本集的选取严重影响着网络的泛化能力。

均匀分布的样本集具有较好的代表性, 具有噪声的样本集, 即使学习过程是收敛的, 但由于噪声的干扰, 映射关系 $F: X \rightarrow Y$ (X 为样本输入, Y 为样本输出) 较无噪声的复杂, 网络的泛化能力会明显下降。

另一方面, 在样本输入向量中对样本类型影响的因素考虑不全是降低网络泛化能力的一个关键因素, 如对于样本集 A_i , x_i 是影响其分类的一个重要因素, 但由于数据采集的原因, A_i 中并没有将 x_i 作为输入。即使网络学习是收敛的, 但其识别率始终维持在一个较低的水平, 在这种情况下, 就要考虑给样本集增加新的影响因素作为输入分量。

5. 实例

油藏水淹层测井解释是注水开发油田非常复杂而又亟需解决的关键问题。由于油田水淹后地下流体性质、岩石物性、地层结构等的复杂变化, 致使测井响应值与水淹状况之间的对应关系十分复杂, 相关的测井

解释的参数较多。由专家经验和测试分析, 我们选用如下有代表性的五参数(厚度、高分辨率深侧向电阻率、高分辨率声波时差、微电位和自然电位)作为输入, 而输出为水淹等级。其中, 水淹等级 2, 3, 4 分别代表弱、中、强水淹。

选用正则化模糊神经网络作为考察对象^[5], 其拓扑结构确定为 5-15-15-243-1 型, 即 5 个输入节点, 15 个模糊化层节点(对应每个输入取 3 个隶属函数), 15 个正则化层节点, 253 个规则层节点和 1 个输出层节点。

网络的学习过程采用单样本循环误差修正方式, 即随机从样本集中抽取一个样本进行网络学习, 直到满足误差精度为止, 再选取下一个样本... 选取 56 个样本数据用于训练。学习参数选取如下: 最大学习次数 30000, 学习速度 0.7, 惯性系数 0.5, 学习精度 0.1 (每个样本实际输出与希望输出的最大偏差)。网络训练次数为每一次包含 26 个样本得到满足精度的修正。学习结果如下: 对原始样本训练, 学习次数 = 1371, 识别率 = 44.6%; 按模糊隶属函数值相近样本向量类别矫正样本后, 学习次数 = 531, 识别率 = 75%。

实际中发现, 在上述例子中另外有两个因素(沉积微相类型、回旋类型)对水淹的类型起很大的影响。下面是加入这两个因素作为新的输入分量之后的学习和识别结果。对原始样本训练, 学习次数 = 1371, 识别率 = 44.6%, 增加影响因素后学习结果, 学习次数 = 2986, 识别率 = 87.3%。

结论 本文重点考虑了由于样本集本身的缺陷而引起的神经网络学习收敛和网络的泛化能力较低问题。实际中大量的问题样本之间的相关性较大, 许多问题由于样本没有经过精心的筛选, 致使在同一个训练集中若干个样本出现矛盾现象, 即存在具有相近输入向量的两个或多个样本的输出相异。针对这一问题, 我们提出了一种按模糊隶属函数值相近样本向量类别矫正策略, 实验证明该方法对于样本错误划分的情况很有效。另一方面, 在样本输入向量中对样本类型影响的因素考虑不全是降低网络泛化能力的一个关键因素。文中也考虑加入新的输入分量对网络的泛化能力的影响, 实验结果充分说明这一处理的有效性。

参考文献

- 1 Hecht-Nielsen R. Theory of the back-propagation neural network. Proc. IJCNN89, 1989, 1: 593~599
- 2 Hornik K. Multi-layer feed forward networks are universal approximators. Neural networks, 1989, 2: 359~366
- 3 Cybenko G. Approximation by superposition of a sigmoidal function. Mathematics of Control, Signals and Systems, 1989, 2: 303~314
- 4 刘耀光, 郑崇勋, 刘明远. 前馈神经网络中的反向传播算法及改进. 进展与展望. 计算机科学, 1996, 23(1): 76~79
- 5 Lianhua H, et al. Evaluation of water flooded interval well logging. ACTA PETROLEI SINICA, 1999, 20(3): 49~55