一类目标函数的逆向构造 [1]

贾 颖 * 杜利民 侯自强

(中国科学院声学研究所交互信息系统实验室 北京 100080)

摘 要 面向解决真实世界问题的神经应用需求，本文提出了一种构造目标函数的逆向方法，即将目标函数的构造任务转化为误差信号的设计．应用这一方法，我们构造出了一类的目标函数，它不仅可以解除均方误差（ MSE ）函数的假饱和状态，从而缩短了网络的训练时间，而且能够克服相对熵函数带来的过度适应性问题，从而提高了网络的泛化能力．

关键词 目标函数，逆向构造，误差信号， MSE, 相对熵

中图号 TN911.7, TN-052

1 引 言

从 80 年代起，沉寂已久的神经网络研究在世界范围内得到了复兴，在模型、算法、应用上都取得了一系列实质性的重大进展，目前，在语音处理、图像处理、计算机视觉、模式识别这些应用领域中，神经网络的应用正在从简单问题上的成功范例向解决真实世界问题的大规模系统转变．这种应用趋势就要求神经网络的学习算法具有学习时间短和泛化能力强的特点．

收敛速度慢和训练时间长是现有的误差反向传播 (BP) 算法的致命弱点，尤其是对于训练那些解决真实世界问题的大规模网络．造成收敛速度慢的一个主要原因是输出层节点的假饱和现象．尽管相对熵准则能够有效地克服假饱和现象，但是由于在接近最优解时，相对熵给出误差信号仍然很大，使得网络过度适应于训练数据，从而降低了在测试数据上的泛化能力．因此当神经网络面向解决真实世界问题时，构造更加有效的目标函数的需求就显得愈加迫切．

本文提出了一种构造目标函数的逆向方法，即根据误差信号来构造目标函数，而误差信号的设计直接体现了我们希望的信号形式．利用这一方法，我们构造出了一类目标函数，它不仅可以克服均方误差 (MSE) 函数的假饱和现象，从而缩短了网络的训练时间，而且能够克服相对熵函数带来的过度适应性问题，从而提高了网络的泛化能力．在我们进行的基于混合 HMM/ANN(Hidden Markov Model/Artificial Neural Networks) 模型的汉语连续语音识别系统的研究中，使用这类目标函数训练神经贝叶斯概率估计器 (一个 125-438-23 的 MLP(Multi-Layered Perceputions)) 的实验结果表明：这类目标函数不仅概念清晰，计算简单，而且效果显著，优于 MSE 和相对熵函数，给出了较理想的收敛过程．

2 均方误差函数和相对熵函数

考虑一个具有 $L(L \geq 2)$ 层的全连接前向网络，设其网络节点的作用函数为 sigmoid，则它的训练可以归结为求解下面的一个非线性无约束优化问题：

$$W^* = \arg \min_{W \in R^N} E(W, X), \tag{1}$$

[1] 1998-07-24 收到， 1999-02-15 定稿

* 目前在英特尔中国研究中心从事语音识别研究工作． Ying.Jia©intel.com

式中 W 是网络中所有的连接权参数，即 $W=(W^1,W^2,\cdots,W^L)$， W^l 是第 l 层连接权向量， X 是训练样本集， $E(W,X)$ 是目标函数 (也称代价函数或误差函数)， W^* 是全局最优解.

误差反向传播 (EBP) 算法是 1985 年由 PDP 研究小组提出，用于解决多层网络的训练问题，它将输出层的误差信号 δ^L 反向逐层传播，并且根据当前层的误差信号 δ^l 来更新 W^l，即

$$w_{i,j}^l(t+1)=w_{i,j}^l(t)+\eta\delta_j^l(t)y_i^{l-1}(t). \tag{2}$$

设训练集由 N 个样本对组成，即 $X=\{(x(n),t(n)),\quad n=1,2,\cdots,N\}$，则目标函数的一般形式为

$$E(W,X)=\frac{1}{N}\sum_{n=1}^{N}e^2(n), \tag{3}$$

则各层节点的误差信号定义为

$$\delta_j^l(n)=-\frac{\partial e^2(n)}{\partial y_j^l(n)}\cdot\frac{\partial y_j^l(n)}{\partial x_j^l(n)}=-\frac{\partial e^2(n)}{\partial y_j^l(n)}\cdot\varphi'(x_j^l(n)). \tag{4}$$

如目标函数 $E(W,X)$ 为均方误差函数 (MSE)，即

$$e^2(n)=\frac{1}{2}\sum_{m=0}^{M_L}(t_m(n)-y_m^L)^2; \tag{5}$$

此时，误差信号就为

$$\delta_j^l=\begin{cases}y_j^L(1-y_j^L)(t_j-y_j^L),\\ y_j^l\cdot(1-y_j^l)\cdot\sum\limits_{i=0}^{M_l}\delta_i^{l+1}w_{ji},\end{cases} \tag{6}$$

从上式我们可以看到，当输出层节点处于假饱和状态，即期望的输出信号 $t_j=0$, 而节点的实际输出 $y_j^L\to 1.0$，或 $t_j=1.0$，而实际输出 $y_j^L\to 0$ 时， MSE 提供的误差信号 $\delta_j^L\to 0$，如图 1 中曲线 1 所示， δ^L 经反向传播使得各层的 $\delta_j^L\to 0$. 因此一旦在输出层出现假饱和现象， MSE 就很难或要经历漫长调节时间才能解除这种状态. 在大规模网络训练中，因为节点的扇入数目很大，节点经常进入饱和状态，此时若采用 MSE 准则，则势必导致训练过程十分漫长或不收敛.

在真实世界问题中，能够用于训练的样本数据通常是有限的和被噪声污染了的，在这种情况下，用具有明确概率意义的目标函数来度量网络的收敛程度就具有更为广阔的应用空间，相对熵准则 (Hopfield(1987),Baum, Wilczek (1988)) 就是这样的一类目标函数

$$E(W,X)=\frac{1}{N}\sum_{k=0}^{N}\left[\sum_{n=1}^{M_L}t_k\ln\frac{t_k}{y_k^L}+(1-t_k)\ln\frac{1-t_k}{1-y_k^L}\right]. \tag{7}$$

对于相对熵函数，可以导出其误差信号为

$$\delta_j^l = \begin{cases} t_j - y_j^L, & l = L; \\ y_j^l(1-y_j^l)\sum_{i=0}^{M_l} \delta_i^{l+1} w_{j,i}^l, & l \neq L. \end{cases} \tag{8}$$

由 (8) 式我们可以看到，当输出层节点处于假饱和状态时，$|\delta_j^L| = 1$，如图 1 中曲线 2 所示．因此相对熵目标函数可以使假饱和状态迅速解除，加快了网络的收敛速度，但是当实际输出值 y_j^L 接近期望输出 t_j 时，相对熵给出的误差信号仍然很大，这就使得训练过度适应于训练数据，而不能渐进地收敛于局部最小点，因此为了提高网络的泛化能力，十分有必要构造性能更加优越的目标函数．

3 用逆向法构造出的一类目标函数

从上面对均方目标函数和相对熵函数的讨论中可以看到，构造新的目标函数形式是十分必要的但又是困难的．同时我们还看到，虽然直接构造目标函数存在很大的困难，但是从目标函数的一阶导数中定义出的误差信号直接控制了网络的训练过程． BP 算法就是按照输出节点的误差信号来调节连接权参数的，因此输出层节点的误差信号的形式决定了学习曲线．从这个意义上讲，我们完全可以根据对学习曲线的要求来设计误差信号．一旦确定了误差信号，目标函数也就隐含地确定了，我们称这种构造目标函数的方法为目标函数的逆向构造法．下面我们就给出一类利用逆向法构造出的目标函数．

对 MSE 和 CE 误差信号的分析使我们意识到，理想的误差信号应该是：当输出层节点出现假饱和时，误差信号能够提供较大的值，正如相对熵准则；当实际输出接近期望的输出值时，误差信号又能较小，使得网络能够渐进地收敛于局部极小点，而不至于使训练过度，即误差信号具有图 1 中曲线 3(点线) 的特点．曲线 3 使我们想到可以将误差信号表示成

$$\delta_j^L = (2t_j - 1)^{n-1}(t_j - y_j^L)^n. \tag{9}$$

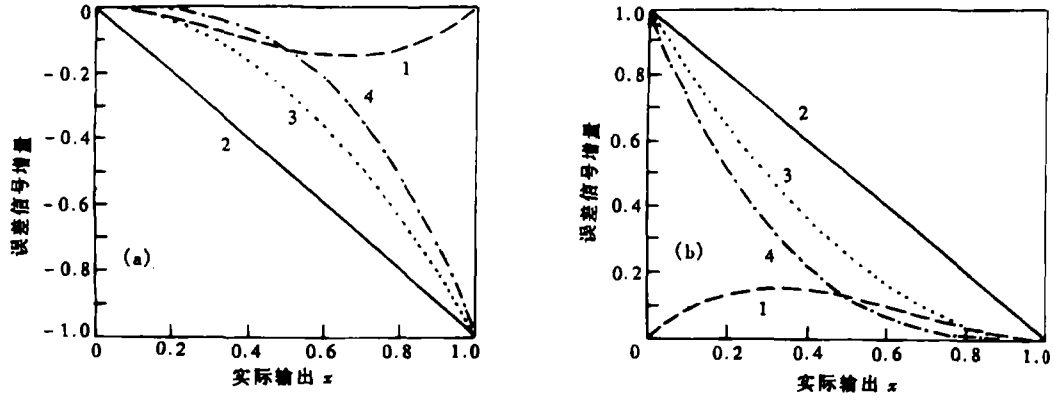

图 1　(a) 理想输出 $t_j = 0$　(b) 理想输出 $t_j = 1.0$
曲线 1 对应于 MSE 的误差信号，曲线 2 对应于 CE 的误差信号
曲线 3 对应于用逆向法构造出的目标函数 ($n = 2$) 的误差信号，
曲线 4 对应于用逆向法构造出的目标函数 ($n = 3$) 的误差信号

图 1 中曲线 3(点线) 和 4(点划线) 分别是 δ_j^L 在 n=2 和 n=3 时的情形．与图 1 中曲线 1(MSE，虚线) 和 2(CE，实线) 的比较，可以看到，在输出节点出现假饱和的情况下，$|\delta_j^L|$

最大，而当输出值接近期望值时，$|\delta_j^L|$ 又能渐进于 0, 所以它实现了我们所期望的形式. 当 n=1 时，(9) 式的误差信号正是相对熵函数的 δ_j^L . 与 (9) 式相对应的目标函数就为

$$e(n) = -\int \sum_{j=0}^{M_L} \delta_j^L \mathrm{d}x_j^L = -\sum_{j=0}^{M_L} \int_{0+}^{1^-} \frac{(2t_j-1)^{n-1}(t_j-y_j^L)^n}{y_j^L(1-y_j^L)} \mathrm{d}y_j^L. \tag{10}$$

事实上，(10) 式中构造出的目标函数具有贝叶斯概率的估计特性，即

性质 1 (10) 式中构造的目标函数的最小化解 $y_j^L(x_n)$ 是贝叶斯概率 $p(c_j|x_n)$ 的一致估计.

证明 对 (10) 式取数学期望，得

$$E\left\{-\sum_{j=1}^{M_L}\int \frac{(2t_j-1)^{n-1}(t_j-y_j^L)^n}{y_j^L(1-y_j^L)}\mathrm{d}y_j^L\right\}$$
$$= \int \sum_{j=1}^{M_L}\left\{p(c_j|x_n)\int \frac{(y_j^L)^n}{y_j^L(1-y_j^L)}\mathrm{d}y_j^L - (1-p(c_j|x_n))\int \frac{(1-y_j^L)^{n-1}}{y_j^L}\mathrm{d}y_j^L\right\}\cdot y_j^L(x_n)\mathrm{d}x_n,$$

又因为 $0 \le p(c_j|x_n) \le 1$, 因此最大似然解 $y_j^L(x_n)$ 就为

$$y_j^L(x_n) = p^{1/n}(c_j|x_n)/[(1-p(c_j|x_n))^{1/n} + p^{1/n}(c_j|x_n)].$$

当 n=1 时， $y_j^L(x_n) = p(c_j|x_n)$; 当 $n \ge 2$ 时， $y_j^L(x_n) \approx p(c_j|x_n)$. 所以 n: $\infty \to 1$ 时，有 $y_j^L(x_n) \to p(c_j|x_n)$, 即 $y_j^L(x_n)$ 是 $p(c_j|x_n)$ 的一致估计. 证毕

性质 2 当 $n = 1$ 时， (10) 式中的误差函数就是相对熵目标函数。

上面构造出的这类目标函数不仅计算量没有任何增加，而且可以有效地抑制假饱和现象和提高网络的泛化能力，用这类目标函数训练得到的网络的输出是贝叶斯概率的一致估计，所以在解决真实世界问题的应用中具有极大的潜力. 从目标函数的构造实例中我们可以看到，用逆向法构造目标函数可以直接面向网络训练过程的控制和优化，从而极大地开拓了目标函数构造的新途径。

4 实验结果

语音识别一直是神经网络最为活跃的应用领域之一，从 60 年代早期 Widrow 的实验，到 80 年代在小词汇孤立词识别中取得的巨大成功，这些成功的应用范例始终激励着人们去探索神经网络技术在大词汇，非特定人连续语音识别中的应用途径，其中基于混合 HMM/ANN 模型的连续语音识别系统是目前最具有生命力的应用方案. 在混合模型中，神经网络被用作贝叶斯概率估计器，向所有 HMM 基元模型的状态提供局部区分性观测概率. 以汉语识别为例，如果以声韵母为建模基元且不考虑基元在连续语流中的变体，共有 61 个基元模型，再假设声母基元用 3 个状态的 HMM , 韵母用 4 个状态的 HMM 表示，则要求神经贝叶斯概率估计器具有 218 个输出节点，为 218 个基元状态提供观测概率. 连续 HMM 的观测向量一般至少有 25 维分量，所以就网络规模而言，混合 HMM/ANN 模型中的神经贝叶斯概率估计器的确是一个面向真实世界应用的大规模网络，对这种规模的网络进行有效地训练也的确是一个非常具有挑战性的课题.

图 2 是训练集上均方误差随迭代次数的变化曲线，图 3 是 Validation 集上帧误识别率随迭代次数的变化曲线，这里我们用贝叶斯准则进行分类，即如果输出层中节点 j 的输出值

大于其他所有节点的值，则认为当前的输入向量 x_n 是由状态 j 产生的．可以看到，在帧识别率和 MSE 两方面， CE 和文中提出的误差都能使训练很快地收敛，而均方目标函数却使误差曲线长时间处于一个平坦的区域．

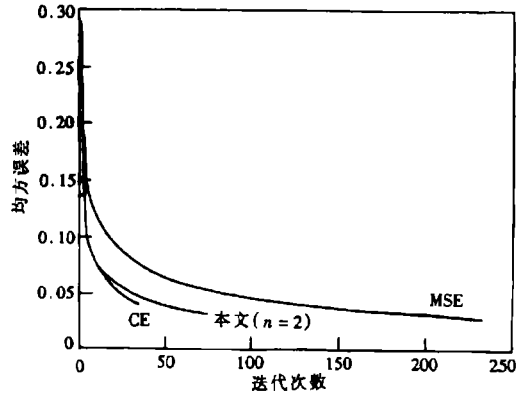

图 2　不同目标函数在训练集上的均方误差曲线

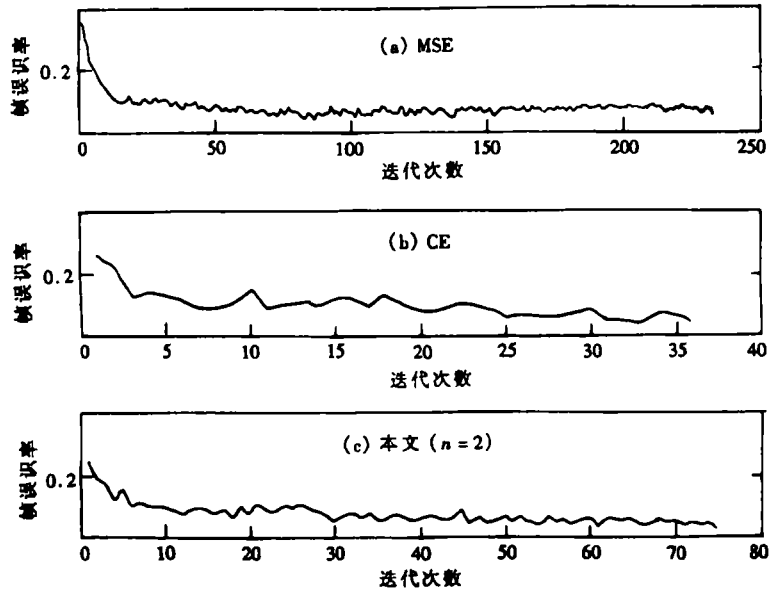

图 3　不同目标函数在验证集上的帧误识率曲线

为了便于比较目标函数 (10) 式与均方目标函数和相对熵函数，这里我们只考虑 22 个声母 (包括零声母) 的混合 HMM/ANN 模型，并且每个声母只用 1 个状态的 HMM 模型表示，其他实验细节参见表 1 ，整个实验是在 Ultra-Sparc-1 工作站上进行，网络的训练数据和性能评价数据是从表 1 中提到的数据库中收集所有 22 个声母和静寂音出现的数据得到，分成 4 组，其中 3 组用于训练， 1 组用于测试．

表 2 是不同目标函数训练得到的网络在训练集， Validation 集和测试集上的帧误识率和总体均方误差，可以看到，用相对熵得到的网络，其泛化性能不如我们构造的目标函数．所以构造的目标函数的确提高了网络的泛化能力．

表 1 训练细节

数据库	来源	一个人的 4 遍发音，1267 个汉语单音节字，人工标注出声韵母所在的位置		
	特征向量	12 阶 LPC 系数 +12 阶倒谱系数 +1 个归一化能量		
网络结构 (3 层)		节点数目	作用函数	内容
	输入层	25×5		$X_{n-4}^{n-1}+x_n+X_{n+1}^{n+4}$
	隐层	500–1000	Tanh	
	输出层	22	Sigmoid	22 个声母 (含零声母)
训练算法	目标函数	MSE, CE, (10) 式 ($n=2$)		
	初始化	在 [0,1] 之间均匀分布，Oja 内部表示		
	搜索方向	梯度下降		
	学习率	沿下降方向作一维近似搜索		
	样本加载	页内随机加载		
	权值更新	On-line		
	终止准则	依据网络在验证集上的泛化性能		

表 2 帧误识率 / 均方误差

	训练集	验证集	测试集
MLP1(MSE)	0.00362/0.02979	0.12261/0.09988	0.11429/0.097409
MLP2(CE)	0.00603/0.03034	0.11256/0.09784	0.11172/0.096049
MLP3(HighCE)	0.01227/0.03617	0.12311/0.098338	0.11136/0.096669

5 结 论

本文提出了一种构造目标函数的逆向方法，即根据误差信号来构造目标函数。利用这一方法，构造出了一类目标函数，它不仅可以克服均方误差 (MSE) 函数的假饱和现象，而且能够克服相对熵函数带来的过度适应性问题。在算法实现上具有概念清晰，计算简单的特点。还证明了用这类目标函数训练得到的网络的输出是贝叶斯概率的一致估计，所以这类目标函数在解决真实世界问题的应用中具有极大的潜力。

参 考 文 献

[1] Richard M D, Lippmann R P. Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Computation, 1991, 3: 461–483.

[2] Gish H. A probabilistic approach to the understanding and training of neural network classifiers, Proc. IEEE Intl Conf. on ASSP, New Mexico, USA: 1990, 1361–1364

[3] Haykin S. Neural Netwoks: Comprehensive Foundation. New York: Macmillan College Publishing Company, 1994, 696.

[4] Karayiannis N B. Accelerating the training of feedforward neural networks using generalized Hebbian rules for initializing the internal representations. IEEE Trans. on Neural Networks, 1996, 7(2): 419–425.

[5] Morgan N, Bourlard H. Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach. IEEE Signal Processing Magazine, 1995, 12(3): 25–42.

[6] 杜利民，侯自强. 自动语音识别中的人工神经网络方法. 物理学进展，1996, (9).

[7]　焦李成. 神经网络系统理论，西安：西安电子科技大学出版社，1992, 26-41.

INVERSE CONSTRUCTING OF A SET OF OBJECTIVE FUNCTIONS

Jia Ying　Du Limin　Hou Zijiang

(*Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100080)

Abstract　To meet the requirements with large-scale neural networks for real-world applications, an inverse way of constructing objective functions was proposed in this paper, which translates the task of constructing objective functions into the design of error signals. Followed this way, a set of objective functions has been given as examples to eliminate the false saturation in Mean Squared Error (MSE) and overspecialization in Cross Entropy (CE). The verification of its power was also made by the comparison with MSE and CE in the tasks of estimating the scaled likelihood for the Hidden Markov Models' states in the Hybrid HMM/ANN models, and showed consistent advantages with the theoretical expectations.

Key words　Objective functions, Inverse constructing, Error signal, MSE, Cross entropy

贾　颖：　男，1971 年生，博士生，主要从事汉语连续语音识别和神经网络方面的研究工作.

杜利民：　男，1958 年生，博士，研究员、博士生导师、IEEE 高级会员、中国电子学会理事. 一直从事语音信号与信息处理技术的研究. 先后开展矢量量化、人工神经网络、小波变换、隐马尔柯夫模型等数字信号处理方法研究和信号自适应增强、微弱信号自适应检测、声门关闭时刻检测、鼻腔小舌启闭时刻检测、语音声学特征分析、语音合成、语音信号压缩、高保真音乐信号压缩等应用研究. 在国际国内学术会议和期刊上发表论文 40 余篇. 95 年入选中科院“百人计划”，领导交互信息系统实验室，系统开展汉语语音交互信息系统的关键算法和系统技术的研究.

侯自强：　男，1936 年生，研究员、博士生导师、中国图像图形学会理事长、中国声学学会副理事长、中国电子学会高级会员、理事、中国科健集团董事长. 研究兴趣包括数字信号处理、人工神经网络、小波变换、分维变换等理论方法和声呐最佳时空处理、信道分配、高分辨率时延估计、自适应波束形成、MRI 血流成像、消费电子学、HFC MAN 技术、语音识别和理解、多媒体智能信息技术等工程技术.