基于 em 算法且能以概率 1 全局收敛的混合学习算法*

王士同

(华东船舶工业学院计算机系 镇江 212003)

摘要 文章指出了随机神经网络 em 学习算法仍然存在着收敛于局部极小值之缺陷.针对三层随机感知机,文章将 em 学习算法与 Solis 和 Wets 的随机优化算法结合起来,提出了三层随机感知机的混合型新学习算法 HRem.文章从理论的角度证明了混合型新学习算法 HRem 能以概率 1 全局收敛于随机感知机的基于 Kullback-Leibler 差异度量的最小值.这一理论结果对 em 学习算法的深入研究有重要意义.

关键词 随机神经网络,em 学习算法,随机优化算法.

中图法分类号 TP18

最近,S. Amari[1,2]教授提出了基于信息几何的随机神经网络的 em 学习算法.国外学者[3]及国内学者张建、史忠植[4]较全面地探讨了多层随机神经网络的 em 算法理论框架.em 学习算法是基于统计模型上微分几何结构分析,通过 e-投影和 m-投影两种对偶投影理论发展起来的减小 Kullback-Leibler 差异的随机神经网络学习算法理论.这一算法理论是当前发展人工神经网络学习理论的重要框架.

em 算法的基本思想是把神经网络模型理解为一个统计模型 $P(y|x,\Theta(W))$,它表示神经网络在输入为 x 时,输出为 y 的概率函数.其中 W 是神经网络的连接权值和阈值,Θ 是统计模型的参数.由上述所有概率分布组成的集合 $S=\{P(y|x,\Theta)\}$,从微分流形的意义上看,S 形成一个流形,Θ 作为坐标被用来识别统计模型.而由神经网络实现的统计模型为 $M=\{P(y|x,\Theta(W))\}$是 S 的子流形,用 W 来进行识别.神经网络的学习问题被表示为在训练样本集 $Y=\{(x^{(t)},y^{(t)}),t=1,2,\ldots,T\}$下,子流形 M 上参数 W 的估计.这样,可用信息几何的结果.$P(y|x,\Theta)$可由指数坐标(e-坐标)θ 进行识别.同时还可编码成混合簇表示,用混合坐标(m-坐标)η 进行识别.训练数据集 $Y=\{(x^{(t)},y^{(t)}),t=1,2,\ldots,T\}$看作是非完整数据,对它的信息量估计形成一个数据子流形 D.学习算法的目标就是寻找 M 上的点 P 和 D 上的点 Q 来减少 Kullback-Leibler 距离

$$K(M,D)=\min_{P\in M,\theta\in D}K(P,Q);$$

$K(P,Q)$是 Kullback-Leibler 差异度量.[1,4]学习算法 em 通过 M 到 D 的 e-投影,实现在神经网络已知的情形下的非完整数据的最好估计;通过 D 到 M 的 m-投影实现神经网络最佳参数的选择.两个投影都是减少上式的 Kullback-Leibler 差异度量 $K(P,Q)$.

和常见的 BP 等算法一样,学习算法 em 的一个致命缺陷是不能保证收敛于 Kullback-Leibler 差异度量的最小值,即经常收敛于局部极小.本文试图结合 Solis 和 Wets 的随机优化算法,提出基于 em 学习算法的神经网络混合型新学习算法 HRem.本文的理论分析将证明:新的混合型学习算法 HRem 能以概率 1 全局收敛于 Kullback-Leibler 差异度量的最小值.这一理论结果对于发展和研究 em 学习算法理论具有重要的学术意义.

1 三层随机感知机及其 em 学习算法

三层随感知机的结构如图 1 所示.它是包括一个隐层的前馈随机网络,有一个输出单元,输入 $x\in R^n$,是 n 维的向量.$z=(z_i),i=1,2,\ldots,m$,是 m 个隐单元的输出,z_i 取值 0 或 1.当输入为 x 时,Z_i 取值 0 或 1 的概率为

$$P(z_i|x)=f(Z_i,w_i x)=)f(x_i,\sum_{j=1}^{n}w_{ij}x_j+w_{i0})$$

其中 $w_i=w_{i1},w_{i2},\ldots,w_{in}$,是由输入到第 i 个隐单元的连接权向量,w_{i0}为阈值,f 为 Sigmoid 函数:

---

* 本文研究得到国家自然科学基金和江苏省跨世纪学术带头人基金资助.作者王士同,1964 年生,博士后,教授,主要研究领域为人工智能,神经网络,模糊数学.

本文通讯联系人:王士同,镇江 212003,华东船舶工业学院计算机系

本文 1997-05-04 收到原稿,1997-06-16 收到修改稿

$$f(x,w)=\frac{\exp\{w\cdot x\}}{1+\exp\{w\}}$$

输出单元，接收隐单元的值 z，输出为随机值 y，y 取值 0 或 1．输出 y 的概率依赖于隐单元的信号 z

$$P(y|z)=f(y,v\cdot z),\quad v\cdot z=\sum_{i=1}^{m}v_iz_i+v_0$$

$v=(v_1,v_2,\ldots,v_m)$，是从隐单元到输出单元的连接权值，v_0 是输出单元的阈值．

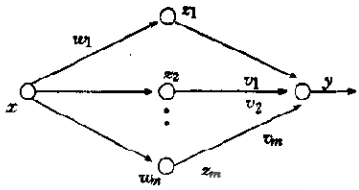

图1　三层随机感知机

于是，三层随机感知机可作为下列概率分布函数．

$$P(y|z|x;W)=P(y|z;x;W)P(z|x;W)=P(y|z;x;W)\prod_{i=1}^{m}P(z_i|x;W)$$

其中 W 是 $w_1,w_2,\ldots,w_n,v$ 的全部．

在训练样本集 $Y=\{(x^{(t)},y^{(t)}),t=1,2,\ldots,T\}$ 的情形下，对第 t 个训练样本，它所确定的数据子流形为

$$D_t=\{\bar{\eta}|\bar{\eta}_{1k,t}=y^{(t)}a_k,\bar{\eta}_{2k,t}=x^{(t)}a_k\}.$$

其中 $\sum_k a_k=1$，这里 $\bar{\eta}_{1k,t},\bar{\eta}_{2k,t},a_k$ 的含义见文献[4]．

下面给出三层随机感知机算法[4]，其中 $M^*,M^{\exists *},D^*,D^{\exists *}$ 的含义见文献[4]．

begin

(1)[初始化]，设神经网络的初始连接权值 w_i^0 和 v^0，计算

$$\theta_{10,t}^{\exists 0}=\sum_k k\cdot v^0$$

$$\theta_{20,t}^{\exists 0}=\sum_k[w_k^0+\log(1+\exp\{k\cdot v^0\})]$$

$$\theta_{0k,t}^{\exists 0}=x^{(t)}k\cdot v^0-y^{(t)}[w_k^0+\log(1-\exp\{k\cdot v^0\})]$$

$$\theta_{0k,t}^{\exists 0}=[x^{(1)}-x^{(t)}][w_k^0+\log(1+\exp\{k\cdot v^0\})]$$

$$\theta_{0k,1}^{\exists 0}=\frac{1}{x^{(1)}}[w_k^0+\log(1+\exp\{k\cdot v^0\})]$$

从而得到 M^* 上的点 $P_0^{\exists *}$；

(2)循环执行下面的(3)(4)两步，直到某终止条件满足．令 s 表示循环次数标，记 $s=0,1,2\ldots$

(3)[e-投影步骤]e-投影 $P_s^{\exists *}$ 到 D^*，得到投影点 $Q_{sH}^{\exists *}\in D^*$，投影点 $\eta^{\exists}$ 坐标为

$$\theta_{10,t}^{\exists S+1}=y^{(t)}\qquad\theta_{20,t}^{\exists S+1}=x^{(t)}\qquad\theta_{1k,t}^{\exists S+1}=0\qquad\theta_{2k,t}^{\exists S+1}=0$$

$$\theta_{2k,t}^{\exists S+1}=\sum_{t=1}^{m}\frac{1}{x^{(t)}}[x^{(1)}\exp\{w_k^Sx^{(1)}\}[1+\exp\{w_k^Sx\}]^{x^{(1)}}]\times$$

$$[1+\exp\{\frac{1}{x^{(t)}}(x^{(1)}kv^S-y^{(t)}(w_k^S\log(1+\exp\{k\cdot v^S\})))\}\times$$

$$\frac{1}{x^{(1)}}(w_k^S+\log\{1+\exp\{kv^0e_0\}\})]^{[e_0x^{(1)}]}$$

(4)[m-投影步骤]m-投影 $Q_{sH}^{\exists *}$ 到 M^*，得到投影点 $P_{sH}^{\exists *}\in M^*$，神经网络权值修正的增量形式为

$$W^{\exists S+1}=W^{\exists S}+\varepsilon B[\eta^{\exists S+1}-\eta^{\exists}(\theta^{\exists S})]$$

其中 $B=[\frac{\partial\theta}{\partial w}]$ 是一个矩阵

$$\frac{\partial\theta_{1k}}{\partial v}=k,\quad\frac{\partial\theta_{1k}}{\partial w}=0$$

$$\frac{\partial\theta_{2k}}{\partial v}=\frac{k\cdot\exp\{kv\}e_0}{1+\exp\{kv\}},\quad\frac{\partial\theta_{2k}}{\partial w}=1$$

权值的修正公式为

$$w_i^{S+1}=w_i^S+\varepsilon_i\frac{\partial\theta_{2k}}{\partial w_i}[\eta_{2k}^{\exists S+1}-\sum_{i=1}^{m}\sum_{t=1}^{T}[1+x^{(t)}\exp\{\sum_k[w_k^S+\log(1+\exp\{kv^S\})]\}\cdot(1+\exp\{\sum_k[w_k^S+\log(1+\exp\{kv^S\})]\})^{e_0x^{(t)}}]]$$

$$v^{S+1}=v^S+\varepsilon_i(\eta_{2k}^{\exists S+1}-y^{(t)})+\varepsilon_ik\frac{\exp\{kv^S\}}{1+\exp\{kv^S\}}\times[\eta_{2k}^{\exists S+1}-\sum_{i=1}^{m}\sum_{t=1}^{T}[1+x^{(t)}\exp\{\sum_k[w_k^S+\log(1+\exp\{kv^S\})]\}\times(1+\exp\{\sum_k[w_k^S+\log(1+\exp\{kv^S\})]\})^{e_0x^{(t)}}]]$$

end.

定理 1.[4]对于 em 学习算法，在每一次的 e-投影和 m-投影交替后，恒有

$$K(P_S^*,Q_S^*)\geq K(P_{S+1}^*,Q_S^*)\geq K(P_{S+1}^*,Q_{S+1}^*).$$

这一定理非常重要.它是与 Solis 和 Wets 的随机优化方法进行有效结合的基础,是保证新的混合型学习算法 HRem 以概率 1 全局收敛的前提.

2 Solis 和 Wets 的随机优化算法

Solis 和 Wets 提出的随机优化算法,其优点在于,对于紧致集上的目标函数,它能以概率 1 收敛于目标函数 $K(W)$的全局最小值.

需要指出,em 算法中的 $K(P,Q)$事实上也是 W 的函数,即 $K(P,Q)=K(P(W),Q(W))$,故亦可作为目标函数.

Solis 和 Wets 的随机优化算法 Random Optimization Algorithm 如下所述.

Random-Optimization Algorithm

begin

(1) 选择初始点 W_0(∈搜索空间 W).设 NUM 是迭代总次数;$k=0$;

(2) 生成高斯型随机矢量 ξ_k,若 $W_k+\xi_k\in W$,转(3);否则,转(4);

(3) (3.1) 若 $K(W_k+\xi_k)<K(W_k)$,则

$$W_{k+1}\leftarrow W_k+\xi_k;\quad b_{k+1}\leftarrow 0.4\xi_k+0.2b_k;$$

(3.2) 若 $K(W_k+\xi_k)\geq K(W_k)$,且 $K(W_k-\xi_k)<K(W_k)$,则

$$W_{k+1}\leftarrow W_k-\xi_k;\quad b_{k+1}\leftarrow b_k-0.4\xi_k;$$

(3.3) 否则,$W_{k+1}\leftarrow W_k$,$b_{k+1}\leftarrow 0.5b_k$(b_0 表示 ξ_k 的算术平均值);

(4) 若 $k=NUM$,算法停止.若 $k<NUM$,则 $k\leftarrow k+1$,转(2);

end.

当 $NUM=\infty$时,上述算法对于紧致集上的目标函数 $K(W)$能以概率 1 保证收敛于此目标函数的全局最小值.

3 基于 em 算法和随机优化算法的混合新学习算法

现在我们给出三层随机感知机的新的混合型学习算法 HRem.它是 em 学习算法和随机优化方法的有效结合.下文将严格证明对于紧致的权空间 W,HRem 算法能以概率 1 全局收敛于 Kullback-Leibler 距离度量的最小值.

算法 HRem

begin

(1) 置迭代次数 NUM,阈值 ε,ε',$G(<1)$,$k\leftarrow 0$.置初始权矢量 W_0;

(2) 计算 $K(P(W_0),Q(W_0))$.计算方法按 em 算法的步骤而得.令 $K_1\leftarrow K(P(W_0),Q(W_0))$;

(3) $K_2\leftarrow K_1$,$k\leftarrow k+1$;

(4) 运用 em 学习算法修改权 W_k,令 $K\leftarrow K(P(W_k),Q(W_k))$;

(5) $K_1\leftarrow K$;

(6) 若 $K_1<\varepsilon$ 或 $k=NUM$,则结束.神经网络权 W 最终求得;

(7) 否则,如果 $|K_2-K_1|<\varepsilon'$,转(8),否则,转(3);

(8) $k\leftarrow k+1$,$K_2\leftarrow K_1$;

(9) 运用 Solis 和 Wets 的随机优化方法修改权 W_{k-1}得到 W_k,令 $K\leftarrow K(W_k)$;

(10) 若 $K\leq K_2$,转(12),否则,转(11);

(11) 若 $k=NUM$,则结束,要求的网络权得到.否则,转(8);

(12) $K_1\leftarrow K$;

(13) 若 $K_1<\varepsilon$,则结束,要求的网络权得到.否则,转(14);

(14) 若 $|K_2-K_1|>(G\times K_2)\vee\varepsilon$,则若 $k=NUM$,算法结束,否则,转(3);

(15) 若(14)的条件不成立,转(11);

end.

在上述学习算法中,阈值的取值应满足 $\varepsilon'\geq\varepsilon$,$G$ 是一个小于 1 的因子.混合型学习算法 HRem 的设计思想可用图 2形象地刻画出来.在此算法中,当算法迭代次数 k 超过一预定阈值 NUM 时,算法就终止运行.

定理 2. 设 W 是要寻找的合适权 W 的紧致权空间.令 $\hat{W}$是使 $K(P(W),Q(W))$达到全局最小值的某个权 W,即

$$K(P(\hat{W}),Q(\hat{W}))=\min_{W\in W}K(P(W),Q(W))$$

设 $W_\varepsilon\triangleq\{W\mid K(P(W),Q(W))-K(P(\hat{W}),Q(\hat{W}))|<\tilde{\varepsilon},W\in W\}$,并设

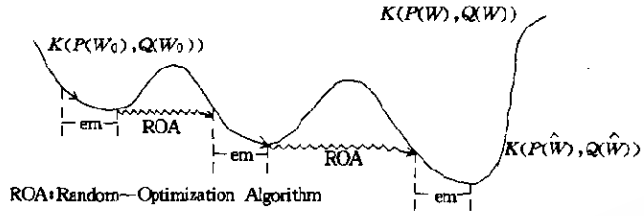

图2

(1) $\forall\ \delta>0$,$U_\delta\cap\mathbf{W}$ 的欧几里德度量是正的,这里 $U_\delta(\hat{W})=\{W\mid\|W-\hat{W}\|<\delta\}$;

(2) NUM 可以取∞;

(3) $K(P(\mathbf{W}),Q(W))$属于 C';

(4) $K(P(\hat{W}),Q(\hat{W}))<\varepsilon$.

则新学习算法 HRem 将具有下列性质:

(1) 对$\forall\ \tilde{\varepsilon}>0$,$\lim\limits_{k\to\infty}\{w\mid W_k\in\mathbf{W}_i\}=1$,即 W_k 进入 $\mathbf{W}_i$的概率为 1.

(2) $K(P(W_k),Q(W_k))$以概率 1 收敛于 $K(P(\hat{W}),Q(\hat{W}))$.

证明:首先考虑下述 3 种情形:

(1) 在算法的第(4)步,算法 HRem 必须使用 em 算法,以使得权 $W_0\to W_1$;

(2) 在 W_k 经过 em 算法修改后,其 $K(P(W_k),Q(W_k))$的绝对值减少量大于 ε';

(3) 在 W_k 经过 Solis 和 Wets 的随机优化方法修改后,其 $K(P(W_k),Q(W_k))$的绝对值大于 $\max(K(P(W_k),Q(W_k))\times G,\varepsilon)$,令 N_1,N_2 和 N_3 分别表示出现情形(1)~(3)的次数.令

$$t_1\in\triangleq K(P(W_0),Q(W_0))-K(P(\hat{W}),Q(\hat{W})),\quad h\triangleq[t_1]$$

其中$[t_1]$表示小于 t_1 的最大整数.很显然,$N_1=1$.根据定理 1,由学习算法 em 总是使 $K(P(W),Q(W))$单调下降,知 $K(P(W_k),Q(W_k))(k=0,1,2,\ldots)$是次数 k 的单调下降函数,且 $N_i\leq h+1,i=2,3$.

当学习算法 em 不能应用时,就使用 Solis 和 Wets 的随机优化方法.因此,随机优化方法将被使用至少 $Total$ 次以得到.其中

$$Total=k-(1+2(h+1))=k-(2h+3)$$

由于 $2h+3$ 有界,当 $k\to\infty$时,$Total\to\infty$.

现在证明此定理的两个结论.证明的过程是基于随机优化方法的收敛性质的.

因为 $K(P(W),Q(W))$是一个连续函数,$\forall\ \tilde{\varepsilon},\exists\ \delta>0$,使得下式关系成立.

$$\|W-\hat{W}\|<\delta\Rightarrow|K(P(W),Q(W))-K(P(\hat{W}),Q(\hat{W}))|<\tilde{\varepsilon}\tag{1}$$

考察下列域 A.

$$A\triangleq\{W\mid\|W-\hat{W}\|<\delta,W\in\mathbf{W}_i\}\triangleq U_\delta(\hat{W})\cap\mathbf{W}$$

根据式(1)和 $\mathbf{W}_i$的定义,有　　$\mathbf{W}_i\supset A$

假定 W_k 已得到,W_{k+1}将由 HRem 算法中的随机优化方法得到.令 $P_A\{W_k\}$表示 W_{k+1}进入 A 的概率,$k=1,2,\ldots$,则

$$P_A\{W_k\}\geq\mu M(A),\quad\forall\ W_k\in\mathbf{W}\tag{2}$$

其中 $M(A)$是 A 在 K 中的测度,且

$$\mu=\inf_{x,y\in\mathbf{W}}q(x-y)\tag{3}$$

这里 $q(\cdot)$是 ξ_k 的概率密度函数$(k=1,2,\ldots)$.根据定理假设知:$M(A)>0$,于是有

$$0<\gamma\leq1,\quad\gamma=\mu M(A)\tag{4}$$

令 $P_{\mathbf{W}_i}\{W_k\}$表示 W_{k+1}进入 $\mathbf{W}_i$的概率,$k=1,2,\ldots$.因 $\mathbf{W}_i\supset A$,故有

$$P_{\mathbf{W}_i}\{W_k\}\geq P_A\{W_k\}\tag{5}$$

根据式(2)~(5),有

$$P_{\mathbf{W}_i}\{W_k\}\geq\gamma>0\quad(k=1,2,\ldots)$$

一旦 W_k 进入 $\mathbf{W}_i$，$\{W_t\}(t=k+1,k+2,\ldots)$将不能逃离 $\mathbf{W}_i$，因为 $K(P(W_t),Q(W_t))$是 t 的单调下降函数. 这就是说，$\mathbf{W}_i$是一个吸引域，于是有

$$P\{\omega|W_k\in\mathbf{W}_i\}\geq 1-(1-\gamma)^{Total}\quad (Total=k-(2h+3)),$$

故有

$$\lim_{k\to\infty}P\{\omega|W_k\in\mathbf{W}_i\}\geq\lim_{k\to\infty}\{1-(1-\gamma)^{Total}\}=1.$$

亦即本定理性质(1)成立.

$\forall\tilde{\varepsilon}>0$，恒有

$$P\{\omega|K(P(W_k),Q(W_k))>K(P(\hat{W}),Q(\hat{W}))+\tilde{\varepsilon}\}\leq P\{\omega|W_k\notin\mathbf{W}_i\}\leq(1-\gamma)^{Total},$$

于是有

$$\lim_{k\to\infty}P\{\omega|K(P(W_k),Q(W_k))>K(P(\hat{W}),Q(\hat{W}))+\tilde{\varepsilon}\}=0.$$

亦就是说，$K(P(W_k),Q(W_k))$将以概率 1 收敛于 $K(P(\hat{W}),Q(\hat{W}))$. 于是，定理的性质(2)成立. 定理得证.

定理 2 表明：对于紧致的权空间 $\mathbf{W}$，新混合学习算法 HRem 能以概率 1 全局收敛于 Kullback-Leibler 距离度量的最小值.

4 结 论

本文将算法 em 与 Solis 和 Wets 的优化方法有效地结合起来，提出了三层随机感知机的混合型新学习算法 HRem. 这一算法的设计思想是新颖的. 其算法理论分析表明它是克服 em 算法常收敛于 Kullback-Leibler 度量的局部极小这一缺陷的有效改进算法. 本文的理论工作对于 em 学习算法的深入研究具有重要的学术价值.

参考文献

1 Amari S. Information theory of the EM and em algorithm for neural networks. Neural Networks, 1995,8(5):10～16

2 Amari S. Differential geometrical methods in statistics. Springer-Verlag, 1985

3 Jordan M *et al*. Hierarchical mixture of experts and EM algorithm. Neural Computation, 1994,(6)

4 张建，史忠植. 多层随机神经网络 em 算法. 计算机研究与发展，1996,33(11):808～815
(Zhang Jian, Shi Zhong-zhi. Algorithm em of multilayer random neural network. Journal of Computer Research and Development, 1996,33(11):808～815)

The Hybrid Learning Algorithm Which is Based on em Algorithm and can Globally Converge with Probability 1

WANG Shi-tong

(*Department of Computer Science EastChina Shipbuilding Institute Zhenjiang* 212003)

Abstract In this paper, the drawback is pointed out that the learning algorithm em of random neural network sometimes converges to local minimum. A new hybrid learning algorithm HRem, which combines algorithm em and the random optimization algorithm presented by Dr. Solis and Wets, is presented for 3-layer random perception. It is theoretically proved that algorithm HRem can globally converge to the minimum of Kullback-Leibler difference measure. This theoretical result has important significances for further research on algorithm em.

Key words Random neural networks, em learning algorithm, random optimization algorithm.