【文章编号】 1004-1540(2009)04-0357-05

m 依赖过程经验风险最小化算法的泛化性能

闫灿伟，曹飞龙
(中国计量学院 理学院，浙江 杭州 310018)

【摘 要】 m 依赖过程作为非独立序列的典型样本，其经验风险最小化的泛化性能不容忽视. 为了研究基于 m 依赖过程经验风险最小化算法的推广能力，我们将基于独立同分布序列的相关结论推广到 m 依赖过程情形中，进一步利用 m 依赖过程的 Bernstein 不等式，建立该序列经验风险最小化原则一致收敛的指数界.

【关键词】 推广能力；经验风险最小化原则；m 依赖过程；一致收敛
【中图分类号】 O174.41 【文献标识码】 A

The generalization performance of empirical risk minimizing with m dependent processes

YAN Can-wei, CAO Fei-long
(College of Sciences, China Jiliang University, Hangzhou 310018, China)

Abstract: The m dependent processes is one of the typical samples of dependent sequences, based on which the generalization performance of empirical risk minimizing (ERM) principle is important. In order to study the generalization performance of ERM principle, it extended some of the previous results in the independent identical distributions (i. i. d.) case to the dependent case. We then establish the exponential bound of ERM principle based on m dependent processes by the Bernstein inequality for m dependent processes.

Key words: generalization performance; ERM principle; m dependent processes; uniform convergence; learning machines

学习理论研究的一个主要问题是根据经验数据选取所期望的依赖关系，这里的经验数据是已标号的，目的在于寻找目标函数 f，使得对任意新的输入 x，函数值 $f(x)$ 能对其正确标号. 在此，我们假设所有数据服从同一未知概率分布 P.

基于经验过程的学习问题吸引了越来越多的学者. 1971 年，Vapnik[1] 首次研究了独立同分布数据的学习理论问题. 随之，有关这方面研究的文章大量涌现，例如，Dudley[2]，Dudley 和 Philipp[3] 等.

Vapnik 建立了基于同分布序列一致收敛和相对一致收敛序列的指数界，Cucker 和 Smale[4]

【收稿日期】 2009-10-24
【基金项目】 国家自然科学基金资助项目(No. 6080235)
【作者简介】 闫灿伟(1983－)，女，山东济宁人，硕士研究生. 主要研究方向为智能计算.

考虑了基于独立同分布序列最小平方误差情况，并得到了经验误差界；Bousquet[5] 利用函数集"大小"的测量新方法，即局部 Rademacher 平均，得到了 Vapnik 和 Chervonenkis 相对误差的泛化性能；Zou 等[6] 考虑了无噪声的情况，在允许函数集上建立了基于独立同分布序列算法学习一致收敛速度的界.

然而，独立性在很多领域是一个比较强的限制. 在一些应用领域，如动态系统研究中，样本序列 $\{z_j\}$ 的独立性假设往往不适合，这就迫切需要研究样本序列满足某种混合条件的 ERM 算法的一致收敛性. 通常情况下，人们考虑的混合序列有 α- 混合、β- 混合和 ϕ- 混合序列等.

本文把上述有关独立同分布情况下的结论推广到这些非独立序列中，主要研究基于 m 依赖过程的 ERM 算法的泛化性能.

1 基本概念介绍

设 $Z=\{z_i=(x_i,y_i)\}_{i=-\infty}^{\infty}$ 是概率空间 (Ω,F,P) 上一个平稳的随机过程，对 $-\infty<i<\infty$，F_i^{∞} 和 $F_{-\infty}^{i}$ 分别表示由随机变量 $\{z_j, j\geqslant i\}$ 和 $\{z_j, j\leqslant i\}$ 生成的 σ 代数.

定义 1.1 （m 依赖过程，见文献[7]）对 $m\geqslant 0$，若 σ 代数 F_{m+1}^{∞} 和 $F_{-\infty}^{0}$ 是相互独立的，则称序列 $\{Z_i\}_{i=-\infty}^{\infty}$ 为 m 依赖过程.

令 $N^{(m)}=\left\lfloor\dfrac{N}{(m+1)}\right\rfloor$，其中 N 为观察样本的个数 $\lfloor\mu\rfloor$ 表示小于或等于 μ 的最大整数，$N^{(m)}$ 称为 m 依赖过程观察样本的"影响因子". 在问题分析过程中，它与独立同分布情况下样本个数 N 的作用是相同的.

令 z 为样本个数为 N 的样本集

$$z=\{z_1,z_2,\cdots,z_N\},$$

这些样本均取自 m 依赖过程.

定义 1.2 （泛化误差，见文献[4]）函数 f 的泛化误差（期望风险）定义如下

$$\varepsilon(f)=E[l(f,z)]=\int l(f,z)\mathrm{d}P,$$

这里，P 是概率空间 (Ω,F) 上未知的概率分布，$l(f,z)$ 为损失函数，并且对于任意 f，$l(f,z)$ 是依赖函数 f 和样本 z 的可积函数.

我们将建立基于 m 依赖过程 ERM 算法的一般框架，包含了模式识别和回归问题，所以这里考虑的损失函数为一般形式的损失函数. 在模式识别中，通常考虑的是 hinge 损失、ε 不敏感损失等，而在回归问题中，通常考虑最小平方损失(least square loss).

学习的任务就是从一给定的假设空间 H 上寻找函数 f_H，使得

$$f_H=\arg\min_{f\in H}\varepsilon(f)=\arg\min_{f\in H}\int l(f,z)\mathrm{d}P \tag{1}$$

根据(1) 式不难看出：由于概率分布 P 未知，直接计算 f_H 是比较困难的. 于是，Vapnik 等人在 1998 年提出了经验风险最小化(ERM) 原则，其原理就是，以最小化经验误差 $\varepsilon_z(f)$ 来代替最小化泛化误差 $\varepsilon(f)$，这里

$$\varepsilon_z(f)=\frac{1}{N}\sum_{i=1}^{N}l(f,z_i),$$

它是泛化误差的离散化形式. 从而，(1) 中的最小化问题可以转化为如下形式：

$$f_z=\arg\min_{f\in H}\varepsilon_z(f)=\arg\min_{f\in H}\frac{1}{N}\sum_{i=1}^{N}l(f,z_i). \tag{2}$$

由于已知的信息完全来自样本，当 ERM 算法中经验误差能够较好地收敛于泛化误差时，上述经验误差的最小化函数 f_z 将是泛化误差 $\varepsilon(f)$ 的最小化函数 f_H 的一个较好的逼近. 所以，ERM 学习算法的关键问题是 f_z 如何较好地逼近 f_H. 若估计效果较好，则称此 ERM 算法具有较好的推广能力(泛化性能). 此时，通过最小化经验误差 $\varepsilon_z(f)$ 可以计算假设空间上的最优估计函数 f_H. 在统计学习理论中，用 $\varepsilon(f_z)-\varepsilon(f_H)$ 来衡量 f_z 逼近 f_H 的程度，而不直接用 f_z 与 f_H 的偏差 f_z-f_H 来刻画 ERM 算法的一致收敛性.

Modha 和 Masry[7] 将 Barron[8] 中有关回归估计的框架推广到 m 依赖过程中，得到了关于 m 依赖过程的 Bernstein 不等式.

引理 1.1 （m 依赖过程的 Bernstein 不等式，见文献[7]）给定整数 $m\geqslant 0$，设 $\{z_i\}_{i=-\infty}^{\infty}$ 为概率空间 (Ω,F,P) 上平稳的 m 依赖过程. 函数 $\psi:\mathbb{R}\to\mathbb{R}$ 为某一可测函数，对每一 $-\infty<i<\infty$，令 $U_i=\psi(z_i)$. 假设 $|U_1|\leqslant d_1$ 几乎成立，且 $E[U_1]=0$，则对所有 $N\geqslant(m+1)$，$\varepsilon>0$，成立.

$$\mathrm{Prob}\left\{\frac{1}{N}\sum_{i=1}^{N}U_i \geqslant \varepsilon\right\} \leqslant \exp\left[-\frac{\varepsilon^2 N^{(m)}}{2\left(E\mid U_1\mid^2+\frac{\varepsilon d_1}{3}\right)}\right], \tag{3}$$

其中 $N^{(m)} = \left\lfloor\frac{N}{(m+1)}\right\rfloor$，$N$ 为样本个数.

2 一致收敛界

本节的主要目的是计算函数集 H 上经验误差与泛化误差的偏差. 对 $M>0$，取

$$H=\{f \mid 0<l(f,z)\leqslant M\}. \tag{4}$$

为研究方便，我们对损失函数作如下假设：设

$$K=\sup_{g_1,g_2\in H, g_1\neq g_2}\max_{z\in Z}\frac{\| l(g_1,z)-l(g_2,z)\|_\infty}{\| g_1-g_2\|_\infty},$$

这里假设 K 是有限的.

对任意 $\varepsilon>0$，我们的目的是界定

$$\mathrm{Prob}\left\{\sup_{f\in H}\mid \varepsilon(f)-\varepsilon_z(f)\mid>\varepsilon\right\}. \tag{5}$$

为界定(5)，有必要控制集合 H 的大小. 尽管有许多工具来度量函数集合 H 的大小，如 VC 维、V_γ 和 P_γ 维等，但为了得到较好的界，我们用覆盖数衡量集合 H 容量的大小.

定义 2.1 （覆盖数，见文献[4]）设 $\eta>0$，S 为某个度量空间的子集，S 的覆盖数定义为完全覆盖 S 所需要的半径为 η 的球的个数的最小值 d.

定理 2.1 对于给定整数 $m\geqslant 0$，设 $Z=\{z_i=(x_i,y_i)\}_{i=-\infty}^{\infty}$ 为平稳的 m 依赖过程，假设对所有 $z\in Z$ 和 $f\in H$，方差 $D[l(f,z)]\leqslant\sigma^2$，则对任意的 $\varepsilon>0$，

$$\mathrm{Prob}\{\mid\varepsilon(f)-\varepsilon_z(f)\mid>\varepsilon\}\leqslant 2\exp\left[-\frac{\varepsilon^2 N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)}\right] \tag{6}$$

成立. 其中 $N^{(m)} = \left\lfloor\frac{N}{(m+1)}\right\rfloor$，$N$ 为样本个数.

证明 对任意的 i，$1\leqslant i\leqslant N$，令 $V_i=E[l(f,z_1)]-l(f,z_i)$，$L_N=\sum_{i=1}^{N}V_i$，则 $\varepsilon(f)-\varepsilon_z(f)=\frac{1}{N}L_N$.

易知，对所有 $i\in\{1,2,\cdots,N\}$，

$$\mid V_i\mid=\mid l(f,z_i)-E[l(f,z_1)]\mid\leqslant M,$$

$$E[V_i]=E[E[l(f,z_1)]-l(f,z_i)]=0.$$

由引理 1.1 知，

$$\mathrm{Prob}\{\varepsilon_z(f)-\varepsilon(f)>\varepsilon\}\leqslant 2\exp\left[-\frac{\varepsilon^2 N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)}\right], \tag{7}$$

考虑到对称性，得

$$\mathrm{Prob}\{\varepsilon_z(f)-\varepsilon(f)<-\varepsilon\}\leqslant 2\exp\left[-\frac{\varepsilon^2 N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)}\right], \tag{8}$$

结合(7)和(8)，(6)成立，定理 2.1 证毕.

定理 2.2 对于给定整数 $m\geqslant 0$，设 $Z=\{z_i=(x_i,y_i)\}_{i=-\infty}^{\infty}$ 为一平稳的 m 依赖过程，若对所有 $z\in Z$ 和 $f\in H$，方差 $D[l(f,z)]\leqslant\sigma^2$，则对任意的 $\varepsilon>0$，

$$\mathrm{Prob}\left\{\sup_{f\in H}\mid\varepsilon(f)-\varepsilon_z(f)\mid>\varepsilon\right\}\leqslant 2N\left(H,\frac{\varepsilon}{4K}\right)\exp\left[-\frac{\varepsilon^2 N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)}\right] \tag{9}$$

成立. 其中 $N^{(m)} = \left\lfloor\frac{N}{(m+1)}\right\rfloor$，$N$ 为样本个数.

证明 令 $H=H_1\cup H_2\cup\cdots\cup H_N$，则对任意 $\varepsilon>0$，有

$$\mathrm{Prob}\left\{\sup_{f\in H}\mid\varepsilon(f)-\varepsilon_z(f)\mid\geqslant\varepsilon\right\}\leqslant \sum_{i=1}^{N}\mathrm{Prob}\left\{\sup_{f\in H_i}\mid\varepsilon(f)-\varepsilon_z(f)\mid\geqslant\varepsilon\right\}. \tag{10}$$

取 $N=N\left(H,\frac{\varepsilon}{2K}\right)$，$O_i$，$i\in\{1,2,\cdots N\}$ 为以 f_i 为中心，以 $\frac{\varepsilon}{2}$ 为半径覆盖集合 H 的闭球，则对任意 $z\in Z$ 和所有 $f\in O_i$，有

$$\mid\varepsilon(f)-\varepsilon(f_i)\mid=\mid E[l(f,z)]-E[l(f_i,z)]\mid\leqslant K\| f-f_i\|_\infty,$$

$$\mid\varepsilon_z(f)-\varepsilon_z(f_i)\mid=\left|\frac{1}{N}\sum_{j=1}^{N}l(f,z_j)-\frac{1}{N}\sum_{j=1}^{N}l(f_i,z_j)\right|\leqslant K\| f-f_i\|_\infty,$$

所以，

$$\mid\varepsilon(f)-\varepsilon_z(f)-\varepsilon(f_i)+\varepsilon_z(f_i).\mid\leqslant 2K\cdot\frac{\varepsilon}{2K}=\varepsilon.$$

对任意的 $z\in Z$，$f\in O_i$，有

$$\sup_{f\in O_i}\mid\varepsilon(f)-\varepsilon_z(f)\mid\geqslant 2\varepsilon\Rightarrow\mid\varepsilon(f_i)-\varepsilon_z(f_i)\mid\geqslant\varepsilon,$$

那么，

$$\mathrm{Prob}\left\{\sup_{f\in O_i}\mid\varepsilon(f)-\varepsilon_z(f)\mid\geqslant 2\varepsilon\right\}\Rightarrow$$

$$\mathrm{Prob}\{|\varepsilon(f_i)-\varepsilon_z(f_i)|\geqslant\varepsilon\}.$$

以$\frac{\varepsilon}{2}$代替 ε,由定理 2.1 知,

$$\mathrm{Prob}\left\{\sup_{f\in O_i}|\varepsilon(f)-\varepsilon_z(f)|\geqslant 2\varepsilon\right\}\leqslant$$

$$2\exp\left[-\frac{\varepsilon^2 N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)}\right].$$

再由(10)式即得定理结论.

若假设集合 H 的覆盖数是有限的,不妨设,对某一 $R>0$,成立 $H\subset B_R$,其中 B_R 是集合 $\mathbb{R}^d$ 的闭球,其形式为:$B_R=\{f\in\mathbb{R}^d,\|f\|\leqslant R\}$,由文献[4]可知,

$$\log(N(B_R,\eta))\leqslant d\log\left(\frac{4R}{\eta}\right).\tag{11}$$

引理 2.3 (见文献[9])令 $c_1,c_2>0$,且 $s>q>0$.则方程

$$x^s-c_1x^q-c_2=0$$

有唯一的正解 x^*.且

$$x^*\leqslant\max\left\{(2c_1)^{\frac{1}{(s-q)}},(2c_2)^{\left(\frac{1}{s}\right)}\right\}.$$

定理 2.4 对于给定整数 $m\geqslant 0$,设 $Z=\{z_i=(x_i,y_i)\}_{i=-\infty}^{\infty}$ 为平稳的 m 依赖过程,若对所有 $z\in Z$ 和 $f\in H$,方差 $D[l(f,z)]\leqslant\sigma^2$,则对任意的 $\delta\in(0,1]$ 和 $\varepsilon>0$,

(1) 以置信水平 $1-\frac{\delta}{2}$ 成立不等式

$$\varepsilon(f_z)-\varepsilon_z(f_z)\leqslant\varepsilon_1^*,$$

(2) 以置信水平 $1-\delta$ 成立不等式

$$\varepsilon(f_z)-\varepsilon(f_H)\leqslant\varepsilon_1^*+\varepsilon_2^*,$$

其中

$$\varepsilon_1^*\leqslant\max\left\{\left[\frac{4M\log\frac{2}{\delta}}{3N^{(m)}}\right],\left[\frac{4\left(d\log N+\sigma^2\log\frac{2}{\delta}\right)}{N^{(m)}}\right]^{\frac{1}{2}}\right\}+N^{-\theta},$$

$$\varepsilon_2^*\leqslant\max\left\{\left[\frac{4M\log\frac{2}{\delta}}{3N^{(m)}}\right],\left[\frac{4\sigma^2\log\frac{2}{\delta}}{N^{(m)}}\right]^{\frac{1}{2}}\right\}.$$

证明 因为假设集合 H 的覆盖数是有限的,由式(11)得,

$$\log\left(N\left(H,\frac{\varepsilon}{4k}\right)\right)\leqslant\log\left(N\left(B_R,\frac{\varepsilon}{4k}\right)\right)\leqslant$$

$$d\log\left(\frac{16KR}{\varepsilon}\right).$$

设 ε_1^* 为下述方程的正解

$$h(\varepsilon):=d\log\left(\frac{16KR}{\varepsilon}\right)-\frac{\varepsilon^2N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)}=\log\frac{\delta}{2},$$

函数 $h:\mathbb{R}_+\to\mathbb{R}$ 是严格递减的,所以当 $h(\varepsilon^*)\leqslant\log\frac{\delta}{2}$ 时,$\varepsilon_1^*\leqslant\varepsilon^*$.

若 $\varepsilon\geqslant N^{-\theta}(\theta\in(0,1))$,则

$$h(\varepsilon)\leqslant d(\log 16KR+\theta\log N)-\frac{\varepsilon^2N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)},$$

于是,当 $N\geqslant(16\ KR)^{\frac{1}{\theta}}$ 时,

$$h(\varepsilon)\leqslant 2d\log N-\frac{\varepsilon^2N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)}.$$

如果我们取 ε^* 为一个正数,使得 $\varepsilon^*>N^{-\theta}$ 且

$$2d\log N-\frac{\varepsilon^2N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)}\leqslant\log\frac{\delta}{2},$$

则有 $h(\varepsilon^*)\leqslant\log\frac{\delta}{2}$.所以,

$$\varepsilon^2-\frac{2M\log\frac{2}{\delta}}{3N^{(m)}}\varepsilon-\frac{2\left(d\log N+\sigma^2\log\frac{2}{\delta}\right)}{N^{(m)}}\geqslant 0.$$

由引理 2.3 知,

$$\varepsilon^*\leqslant\max\left\{\left[\frac{4M\log\frac{2}{\delta}}{3N^{(m)}}\right],\left[\frac{4\left(d\log N+\sigma^2\log\frac{2}{\delta}\right)}{N^{(m)}}\right]^{\frac{1}{2}}\right\},$$

所以

$$\varepsilon_1^*\leqslant\max\left\{\left[\frac{4M\log\frac{2}{\delta}}{3N^{(m)}}\right],\left[\frac{4\left(d\log N+\sigma^2\log\frac{2}{\delta}\right)}{N^{(m)}}\right]^{\frac{1}{2}}\right\}+N^{-\theta}.$$

由定理 2.2,以置信水平 $1-\frac{\delta}{2}$ 成立不等式

$$\varepsilon(f)-\varepsilon_z(f)\leqslant\varepsilon_1^*,\tag{12}$$

由于上述不等式对所有的函数 $f\in H$ 以置信水平 $1-\frac{\delta}{2}$ 成立,特别地,

$$\varepsilon(f_z)-\varepsilon_z(f_z)\leqslant\varepsilon_1^*.\tag{13}$$

由式(6),对上述 δ,不等式

$$\mathrm{Prob}\{|\varepsilon(f)-\varepsilon_z(f)|>\varepsilon\}\leqslant 2\exp\left(-\frac{\varepsilon^2 N^{(m)}}{2\left(\sigma^2+\frac{\varepsilon M}{3}\right)}\right)$$

对所有 $f\in H$ 以置信水平 $1-\frac{\delta}{2}$ 成立，特别地，由式(1) 函数 f_H 的定义知

$$\varepsilon(f_H)>\varepsilon_z(f_H)-\varepsilon_2^*, \tag{14}$$

其中

$$\varepsilon_2^*\leqslant\max\left\{\left(\frac{4M\log\frac{2}{\delta}}{3N^{(m)}}\right),\left(\frac{4\sigma^2\log\frac{2}{\delta}}{N^{(m)}}\right)^{(\frac{1}{2})}\right\}.$$

又因为

$$\varepsilon_z(f_H)\geqslant\varepsilon_z(f_z), \tag{15}$$

由不等式(13)，(14) 和(15)，可以推断，以置信水平 $1-\delta$ 成立不等式

$$\varepsilon(f_z)-\varepsilon(f_H)\leqslant\varepsilon_1^*+\varepsilon_2^*.$$

定理 2.4 证毕.

【参　考　文　献】

[1] VAPNIK V N, CHERVONENKIS Y A. On the uniform convergence of relative frequencies of events to their probabilities[J]. Theory Probab, 1971,16:264-280.

[2] DUDLEY R M. Central limit theorems for empirical measures[J]. Ann Probab, 1978,6:899-929.

[3] DUDLEY R M, PHILIPP W. Invariance principles for sums of Banach space valued random elements and empirical processes[J]. Z Wahrsch Verw Gebiete, 1983,62:509-552.

[4] CUCKER F, SMALE S. On the mathematical foundations of learning[J]. Bull Amer Math Soc, 2001,39:1-49.

[5] BOUSQENT O. New approaches to statistical learning theory[J]. Ann Inst Statist Math, 2003,55:371-389.

[6] ZOU B, LI L Q, XU J. The bounds on the rate of uniform convergence for learning machine[J]. Lecture Notes in Comput Sci, 2005,3496:538-545.

[7] MODHA S, MASRY E. Minimum complexity regression estimation with weakly dependent observations[J]. IEEE Trans Inform Theory, 1996,42:2133-2145.

[8] BARRON A R. Complexity regularization with application to artificial Neural Networks[C]//Nonparametric Functional Estimation and Related Topics. Netherlands: kluwer, 1991:561-576.

[9] CUCKER F, SMALE S. Best choices for regularization parameters in learning theory: on the bias-variance problem[J]. Foundation of Computational Mathematics, 2002,2:413-428.

（上接第 356 页）

问题 2　我们能否找出一个更好的方法来解决注记 2 中的问题，使得在降低文献[6]中系数 12 的同时还不会升高系数 8.

【参　考　文　献】

[1] GAL S G. Sur les theorems of approximation de Weierstrass[J]. Mathomatica, 1981,23(46):25-30.

[2] GAL S G, SZABADOS J. On monotone and doubly monotone polynomial approximation[J]. Acta Mathematica Hungarica, 1992,59:395-399.

[3] XIE T F, ZHOU S P. A remark on approximation by monotone sequences of polynomials[J]. Acta Mathematica Hungarica, 1995,67:119-121.

[4] 谢庭藩，周颂平. 实函数逼近论[M]. 杭州：浙江大学出版社，1998:28-30.

[5] WU X Q. A development on approximation by monotone sequences of polynomials[J]. Approximation Theory and Its Applications, 1998,14:98-101.

[6] 谢庭藩. 关于用多项式的单调序列逼近[J]. 中国计量学院学报，2003,14(2):134-136.

[7] CYBENKO G. Approximation by superpositions of sigmoidal function[J]. Mathematica of Control, Signals, and Systems, 1989,2:303-314.

[8] CHEN D B. Degree of approximation by superpositions of a sigmoidal function[J]. Approximation Theory and Its Applications, 1993,9(3):17-28.

[9] CHEN T P, CHEN H, LIU R W. Approximation Capability in $C(\overline{R}^n)$ by multilayer feedforward networks and related problems[J]. IEEE Transactions on Neural Networks, 1995,6:25-30.

[10] CAO F L, XIE T F, XU Z B. The Estimate for approximation error of neural networks: a constructive approach[J]. Neurocomputing, 2008,71:626-630.

[11] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989,2:359-366.