神经网络的统计学习理论基础*

The Theory Elements of Neural Network Statistical Learning

吴建生[1]，金 龙[2]

Wu Jiansheng[1], Jing Long[2]

(1. 柳州师范高等专科学校数学与计算机科学系，广西柳州 545004；2. 广西气象减灾研究所，广西南宁 530022)

(1. Dept. of Math. & Comp., Liuzhou Teacher Coll., Liuzhou, Guangxi, 545004, China; 2. Guangxi Research Inst. of Meteorological Disasters Mitigation, Nanning, Guangxi, 530022, China)

摘要：介绍神经网络的统计学习过程和理论，讨论基于经验风险最小化的学习理论对神经网络推广性能的影响，分析基于结构风险最小化的支持向量机. 认为神经网络因其出色的高度非线性映射能力、自组织和适应能力、记忆联想能力，使得神经网络成为机器学习的重要研究领域.

关键词：神经网络 学习过程 经验风险 结构风险 支持向量机

中图法分类号：TP181 **文献标识码**：A **文章编号**：1002-7378(2005)02-0102-04

Abstract: The statistical learning process and theory of the neural network are introduced. The influence of generation ability based on the empirical risk minimization and the support vector machines based on the structural risk minimization are discussed. The neural network becomes a research hotspot in machine learning because of its outstanding nonlinear mapping, self-organized, parallelity, adaptation.

Key words: neural network, learning process, empirical risk, structural risk, support vector machines

神经网络是现代人工智能技术的重要研究领域，它主要研究如何从观测数据中寻找规律，依次对某系统的输入和输出之间的依赖关系进行估计，并使其对未来的输出做出尽可能准确的预测[1~3]. 神经网络是基于机器学习的重要研究方面，它具有从环境中学习的能力，并通过学习来改善自己的行为，以适应环境[4]. 神经网络的学习过程是网络参数(网络的权值和阈值)被环境所激励，激励的结果是网络参数的改变并以新的方式来响应外部环境[5]. 神经网络有5种基本学习算法：误差-修正学习、基于记忆的学习、Hebb学习、竞争学习和Boltzmann学习[5,6]. 本文回顾了神经网络的统计学习过程和理论，讨论了基于经验风险最小化的学习理论对神经网络推广性能的影响，并介绍基于怨偶了小化的支持向量机，最后对神经网络的统计学习理论作了展望.

1 神经网络的学习过程

神经网络的学习模型为[2]：设 X 是网络的输入，相应的网络输出为 Y，ζ 表示网络参数构成的向量，对 M 对学习样本

$$\tau = (X_m, Y_m), m = 1, 2, \cdots, M, \tag{1}$$

通常我们并不知道 X 和 Y 之间确切的函数关系，因而提出以下回归模型[7]

$$Y = f(X, \zeta) + \varepsilon, \varepsilon \sim N(0, \sigma^2) \tag{2}$$

来讨论，其中 $f(X, \zeta)$ 是一个函数(可能是线性也可能是非线性)；ε 是随机期望误差.

回归模型具有如下性质[7]：

收稿日期：2004-12-20

修回日期：2005-03-03

作者简介：吴建生(1974-)，男，陕西咸阳人，硕士，讲师，主要从事神经网络应用研究。

* 广西自然科学基金(0339025)资助项目。

$$(\mathrm{I})f(x)=E[\varepsilon|X], \tag{3}$$

$$(\mathrm{II})E[\varepsilon\cdot f(x)]=0. \tag{4}$$

回归模型是对随机环境的一个数学描述，它的目的是用输入向量 X 解释或预测输出向量 Y，神经网络通过样本 τ 学习，并获取一组参数向量 ζ，有：

$$\hat{Y}=F(X,\zeta), \tag{5}$$

上式称为逼近函数，它只能是对回归模型一个近似的描述，参数向量是通过最小化损失函数[5]

$$L(\zeta)=\frac{1}{2}\cdot\sum_{i=1}^{M}(y_i-F(X_i,\zeta)^2 \tag{6}$$

获得，令 E_τ 为作用在整个学习样本 τ 上的均值算子，则有：

$$L(\zeta)=E_\tau[Y-F(X,\zeta)]^2=E_\tau[Y-f(X,\zeta)+f(X,\zeta)-F(X,\zeta)]^2=E_\tau[\varepsilon^2]+E_\tau[(f(X,\zeta)-F(X,\zeta))^2]+2E_\tau[\varepsilon(f(X,\zeta)-F(X,\zeta))],$$

由性质(Ⅰ)，(Ⅱ)则

$$L(\zeta)=E_\tau[\varepsilon^2]+E_\tau[(f(X,\zeta)-F(X,\zeta))^2], \tag{7}$$

(7)式右边的第一项式在训练样本 τ 之上期望误差的方差，它代表内在误差而且它独立于参数向量 ζ，就最小化损失函数 $L(\zeta)$ 而言，它可以被忽略，由此最小化损失函数 $L(\zeta)$ 的解也就是最小化回归函数 $f(X)$ 和逼近函数 $F(X,\zeta)$ 之间总体均方距离，即：

$$L(f(X),F(X,\zeta))=E_\tau[(f(X,\zeta)-F(X,\zeta))^2]=E_\tau[(E[Y|X]-F(X,\zeta))^2].$$

又因为

$$E[Y|X]-F(X,\zeta)=E[Y|X]-E_\tau[F(X,\zeta)]+E_\tau[F(X,\zeta)]-F(X,\zeta), \tag{8}$$

故而 $$L(f(x),F(X,\zeta))=B^2(\zeta)+V(\zeta), \tag{9}$$

其中，$B(\zeta)=E_\tau[F(X,\zeta)-E_\tau[Y|X]]$；$V(\zeta)=E_\tau[(F(X,\zeta)-E_\tau[F(X,\zeta)])^2]$，$B(\zeta)$ 称作逼近误差，它是逼近函数 $F(X,\zeta)$ 的平均值相对于回归函数 $f(x)$ 的偏差，同时也说明神经网络通过对学习样本的学习，得到的函数 $F(X,\zeta)$ 不可能准确地逼近回归函数. $V(\zeta)$ 是在全部学习样本 τ 之上，逼近函数 $F(X,\zeta)$ 的方差，它是估计误差的体现，由此我们说神经网络在学习过程中，如果要取得良好的整体性能，需要逼近误差和估计误差都很小. 但是研究人员发现对于学习样本一定的神经网络，获得小偏差的代价是方差大，只有当学习样本无限时才能同时消除二者，这称为偏差－方差困境[8].

2 神经网络的推广问题

网络的期望风险为：

$$R(\zeta)=\int L(y,F(X,\zeta))\mathrm{d}P_{X,Y}(x,y), \tag{10}$$

其中，$L(y,F(X,\zeta))$ 为损失函数，一种最常见的损失函数，其定义为：

$$L(y,F(X,\zeta))=((y-F(X,\zeta))^2. \tag{11}$$

网络学习的目标是期望风险最小化，但是在多数情况下，由于联合概率分布函数 $P_{X,Y}(x,y)$ 是未知的，期望风险无法计算，一般采用经验风险最小化归纳原理(Empirical Risk Minimization，简称ERM)，即用学习样本定义经验风险[2]：

$$R_{emp}(\zeta)=\frac{1}{M}\cdot\sum_{i=1}^{M}L(Y,F(X,\zeta))=\frac{1}{M}\sum_{i=1}^{M}[(y_i-F(x_i,\zeta))]^2,$$

作为(10)式的估计，神经网络是通过对学习样本的学习，使得网络在学习集 τ 上的实际输出和网络输出之间的经验风险最小，以此求的网络的最佳参数向量 ζ，但是基于 ERM 的机器学习存在许多问题，其中有两个方面对神经网络至关重要[9]：

一方面，设 ζ_0 使 $R(\zeta)$ 达到最小，设 ζ_a 使 $R_{emp}(\zeta)$ 达到最小，$R(\zeta_a)$ 和 $R_{emp}(\zeta_a)$ 之间的差异是否足够小，如果足够小，我们说神经网络通过有限样本学习求得既能使经验风险最小也能使期望风险最小；另外还需要考虑 $R(\zeta_a)$ 和 $R(\zeta_0)$ 之间的差异，如果二者足够接近，则表明 ζ_a 与全体集合上的最佳参数 ζ_0 足够接近，对这 2 个问题的回答，就是网络推广问题，即用有限个学习样本求得 $F(X,\zeta_a)$ 是否适于全体集合(或者说对学习集外的数据，其损失函数是否依然保持最小).

另一方面，当学习集的规模 M 一定时，应如何选择神经网络的结构，以使得 $R(\zeta_a)$ 达到最小，这个问题称为结构风险最小问题.

Vapnik 和 Chervonenkis[9,10] 经过系统深入的研究，对上述问题一给出肯定的回答，即：

定理 1[10] 对任何参数 ζ 和 $\forall\ \varepsilon>0$，

$$\lim_{M\to\infty}P[|R_{emp}(\zeta_a)-R(\zeta_a)|>\varepsilon]=0,$$

$$\lim_{M\to\infty}P[|R(\zeta_a)-R(\zeta_0)|>\varepsilon]=0.$$

成立的充分必要条件为：

$$\lim_{M\to\infty}P[\sup_{\zeta}|R(\zeta)-R_{emp}(\zeta)|>\varepsilon]=0, \tag{12}$$

即学习样本的数量 M 无穷时，$R(\zeta_a)$ 以概率收敛到 $R(\zeta_0)$，经验风险 $R_{emp}(\zeta)$ 的最小点也以概率收敛到期望风险 $R(\zeta)$ 的最小点，而实际问题中学习样本个数总是有限的，为了分析这种情况下的学习质量，需要考虑收敛速度.

考虑收敛速度时，有一个重要参数是 Vapnik-Chervonenkis dimension 维数（简称 VC 维数），其名称是为纪念它的创立者 Vapnik 和 Chervonenkis 而命名的. 在模式识别中 VC 维数的直观定义是[11]：对一个指标函数集，若存在 h 个样本能够被函数集中的函数按所有可能的 2^h 种形式分开，则称函数集能够把 h 个样本打散. 函数集的 VC 维数就是它能打散的最大样本数目 h，VC 维数是统计学习理论的一个核心概念，它反映了函数集的学习能力，VC 维数越大表示机器学习能力越复杂[12].

20 世纪 90 年代初，人们证明了神经网络的 VC 维数与其连接的神经元的个数以及神经元内部的激活函数有关，一般条件下，VC 维数是有限的，但是可能是一个很大的数[13]. 如对一个激活函数取函数，输出层取线性函数的 2 层单输出的前馈网络，其 VC 维数由下列不等式界定[14]：

$$4[\frac{N_1}{2}]_I N \leqslant h \leqslant 4N_w \lg(eN_N), \tag{13}$$

其中，N 为输入层节点；N_1 隐层神经元个数；N_w 是网络中权总数；N_N 网络中神经元总数；$[*]_I$ 表示取整数部分.

定理 2（Vapnik 和 Chervonenkis 定理）[15] 损失函数 $L(y,F(X,\zeta))$ 分成 2 种情况讨论：全有界函数集，$-\infty < A \leqslant L(y,F(X,\zeta) \leqslant B < \infty$；全有界非否函数集，$0 \leqslant L(y,F(X,\zeta) \leqslant B < \infty$，则有：

（Ⅰ）全有界函数集

$$P\{|R(\zeta_a) - R_{emp}(\zeta_a)| \leqslant \frac{(B-A)}{2}\varepsilon\} \geqslant 1-\eta, \tag{14}$$

$$P\{R(\zeta_a) \leqslant R(\zeta_0) + \frac{(B-A)}{2}\sqrt{\varepsilon} + (B-A)\sqrt{\frac{-\ln\eta}{2M}}\} \geqslant (1-2\eta); \tag{15}$$

（Ⅱ）全界非否函数集

$$P\{R_{emp}(\zeta_a) \leqslant R(\zeta_0) + \frac{B\varepsilon}{2}(1+\sqrt{1+\frac{4R_{emp}(\zeta_a)}{B\varepsilon}})\} \geqslant (1-\eta), \tag{16}$$

$$P\{R(\zeta_a) \leqslant R(\zeta_0) + \frac{B\varepsilon}{2}(1+\sqrt{1+\frac{4}{\varepsilon}}) + B(\sqrt{\frac{-\ln\eta}{2M}}\} \geqslant (1-2\eta), \tag{17}$$

其中，$\varepsilon = \dfrac{4[h\ln(\frac{2M}{h}+1) - \ln\frac{n}{4}]}{M}$，$M > h$，$h$ 为学习集的 VC 维数，M 学习样本个数. Vapnik 和 Chervonenkis 定理进一步说明，神经网络在学习过程中的实际风险由 2 部分组成：一是经验风险（训练误差）；二是称作置信范围，它和学习机的维数及训练样本有关，因而上述(14)，(16) 式可以简单地综合为：

$$P\{R(\zeta) \leqslant R_{emp}(\zeta) + \varphi(\frac{h}{M})\} \geqslant (1-\eta), \tag{18}$$

(18) 式表明，要使网络的实际风险最小，必须使经验风险、VC 维数与样本数的比率同时最小，才能使实际风险最小. 在神经网络的学习过程中，由于训练样本有限，为使神经网络对学习样本拟合好，就需要扩大网络的规模，规模扩大必然导致 VC 维数 h 增加，从而对给定的学习样本，导致置信范围增加，因此即使经验风险为 0，总的期望风险 $R(\zeta)$ 也可能很大，也就是网络的泛化能力差，这就是所谓的“过拟合”问题[16].

3 结构风险最小化

对一个 2 层单输出前向神经网络的学习过程而言，可以选择的参数是学习集的规模 M 和隐层神经元的个数 N_1，后者的作用表现在学习机的 h 值（VC 维数）[17]. 对于一个特定的学习任务，为了保证 $R(\zeta_a)$ 足够小，由于 M 不可能任意增大，这时隐层神经元的个数成为唯一可选择的参数. 在 M 一定的条件下，必须选择 N_1 使得 $R(\zeta_a)$ 达到最小值，为此将 (15)，(17) 式改为下列形式：

$$R_{N_1}(\zeta_a) \leqslant R_{N_1}(\zeta_0) + T_{N_1}(M), \tag{19}$$

其中，$R(\zeta_a)$ 和 $R(\zeta_0)$ 加下标 N_1 是为表明了它们随 N_1 而变化. $T_{N_1}(M)$ 是(15)，(17) 式左侧第 2 项和第 3 项之和，它决定于 M 和 N_1，也被称为置信限. 在 M 一定的条件下，$R_{N_1}(\zeta_a)$ 随着 N_1 的增加而减少，而 $R_{N_1}(\zeta_0)$ 随着 N_1 的增加而增加. N_1 有一定的最佳值，使 $R_{N_1}(\zeta_0)$ 与 $T_{N_1}(M)$ 之和达到最小，用这种方法来决定网络的规模称为结构风险最小化[18].

实现 SRM 原则基于以下 2 种思路[11]：（Ⅰ）保持置信范围固定（通过选择一个适当的结构）并最小化经验风险；（Ⅱ）保持经验风险固定并最小化置信范围.

支持向量机就是第二种思路的实现，设计函数集的某种结构使每个子集取得经验风险最小（神经网络的训练误差为 0），然后适当选择子集使置信范围最小，则这个子集便是最优函数[19].

Cortes 和 Vapnik 在 1995 年提出支持向量机[19]，它是近年来机器学习研究的一项重大成果，是在 Vapnik 等人提出的小样本统计学习理论基础发展

而来，其算法是基于结构风险最小化准则[20]. 与传统的神经网络相比，支持向量机不仅结构简单而且各种技术性能明显优于神经网络，这已被大量的实验证实[20,21]，尤其它在处理非线性问题时，通过非线性核函数，将输入向量映射到高维线性特征空间. 该空间构造样本最优超平面，以此将非线性问题就转化为高维空间中的线性问题，然后用一个核函数来代替高维空间中内积计算，从而巧妙地解决了复杂计算问题，并且能有效地克服维数灾和有效提高泛化能力[22,23].

4 展望

统计学习方法和统计建模是解决带有不确定性、复杂性系统问题的一种自然选择的工具，一直是机器学习领域的重点和热点[24]. 随着基于经验风险和基于结构风险最小化理论统计学习理论研究的深入，传统的神经网络统计学习理论会得到全新的发展，目前神经网络的统计学习理论的研究主要集中在以下四个方面：

（1）选择适当的网络结构并最小化经验风险，寻找最优网络拓扑结构，限制网络模型的复杂度，即在偏差-方差之间找到适当的折中，提高网络泛化性能. 采用的方法是优化网络结构[25～27]（并行学习和串行修改）、利用正则化控制网络的有效复杂度[28]或者将结构优化和正则化方法结合[29]；

（2）在神经网络学习中加入先验知识，利用贝叶斯分析方法和贝叶斯统计计算构造贝叶斯神经网络，可以将统计数学以条件概率的形式融入神经网络的学习过程中，这样就能将人类的先验知识和后验数据紧密结合进行逻辑推理学习[30]；

（3）神经网络学习中加入模糊推理学习技术，建立模糊神经网络统计学习方法[31]；

（4）神经网络统计学习算法与其它智能算法、仿生算法的结合，如和遗传算法、模拟退火算法、粒子群算法等算法的结合[32].

本文认为神经网络的统计学习发展的一个趋势把支持向量机的学习方法应用到神经网络学习中，用以指导神经网络的学习并把二者结合。另外的一个趋势是在神经网络学习中借助概率统计学中的非参数方法和思想，建立新的神经网络统计学习理论.

参考文献：

[1] 谭东宁，谭东汉. 小样本机器学习理论：统计学学习理论[J]. 南京理工大学学报，2001，25(1)：108-110.

[2] Vapnik V N. Estimation of Dependences Based on Empirical Data[M]. New York：Spring-Verlag，1982.

[3] Cherkassky V，Mulier F. Learning From Data：Concepts Theory and Methods[M]. New York：John Viley & Sons，1997.

[4] Mendel J M，Mclaren R W. Reinforce ment-learning control and pattern recognition systems in adaptive [J]. Learning and Pattern Recognition Systems：Theory and Applications，1979，66：287-318.

[5] Simon Haykin. 神经网络原理[M]. 叶世伟、史忠植译. 北京：机械工业出版社，2004.

[6] Martin T Hagan，Howard B Demuth，Mark H Beale. 神经网络设计[M]. 戴 葵，等译. 北京：机械工业出版社，2002.

[7] Whiter H. Learning in Artificial Neural Networks：a statistical perspective[J]. Neural Computation，1989，1：425-464.

[8] German S，Bienenstock E，Doursect R. Neural networks and the bias/variance dilemma[J]. Neural Computation，1992，4：1-58.

[9] Vladimir N Vapnik. 统计学习理论基础[M]. 许建华，张学工译. 北京：电子工业出版社，2004.

[10] Vapnik V N. The Nature of Statistical Learning Theory[M]. Berlin：Springer-Verlag，1995.

[11] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报，2000，26(1)：32-42.

[12] Trevor Hastie，Robert Tibshirani，Jerome Friedman. 统计学习基础-数据挖掘、推理与预测[M]. 范 明，柴玉梅，耆红英译. 北京：电子工业出版社，2004.

[13] Baum E B，Haussler D. What size net gives valid generalization? [J]. Neural Computation，1989，1：151-160.

[14] Sontag E D. Sigmoid distinguish more effectively than heavisides[J]. Neural Computation，1989，1：470-472.

[15] Vapnik V N. An overview of Statistical Learning Theory[J]. IEEE Transactions on Neural Network，1999，10(5)：988-999.

[16] 王国胜，钟义信. 支持向量机的理论基础-统计学习理论[J]. 计算机工程与应用，2001，19：19-21.

[17] Xingjun Yang，Junli Zheng. Artificial Neural Network and Blind Signal Processing[M]. Beijing：Tsinghua Press，2003.

[18] Nello Cristianini，John Shawe-Taylor. 支持向量机导论[M]. 李国正，王 猛，曾华军译. 北京：电子工业出版社，2004.

[19] Cortes C，Vapnik V. Support vector networks[J]. Machine Learning，1995，20：273-295.

（下转第 109 页）

后，由于嵌入了许多 script 脚本，页面变得比较复杂，这一点恰好反映了瘦客户机模式应用于象 GIS 这样的大量数据处理系统时所带来的局限性。希望随着智能客户端技术的发展，通过解决瘦客户端与胖客户端的矛盾，找到更优的处理方式。

参考文献：

[1] 蒋 泰，邓一星. 基于 MAPGIS-IMS 的 Web GIS 应用研究[J]. 计算机应用研究，2004，12：196～197.

[2] SVG 1.1 Specification. www. w3c. org[EB/OL]. 2003.01.

[3] Bill Trippe，Kate Binder. SVG 设计：在下一代 Web 站点中使用可缩放矢量图形[M]. 高 伟，英 宇译. 北京：机械工业出版社，2003.

[4] Oswald Campesato. Fundamentals of SVG Programming：Concepts to Source Code[A]. In：Charles River Media，2004.

（责任编辑：黎贞崇）

（上接第 105 页）

[20] Bartlett P L，J Shawe-Taylor. Generalization performance on support vector machines and pattern classifiers[C]. In：Sholkopf B，Burges C，Smola A. Eds. Advances in Kernel Methods-Support Vector Learning，Cambridge：MIT Press，1999.

[21] Sholkopf B，Sung K，Burges C J C，et al. Comparing support vector machine with Gaussian kernel to radial basis function classifiers [J]. IEEE Trans (Signal Processing)，2001，45：2758-2765.

[22] 王国胜，钟义信. 支持向量机的若干新进展[J]. 电子学报，2002，29(10)：1397-1400.

[23] 刘江华，程君实，陈佳品. 支持向量机算法综述[J]. 信息与控制，2002，31(1)：45-48.

[24] Wang Tianshu，Zheng Nanning，Yuan Zejian. Statistical learning in machine intelligence and pattern recognition[J]. Act Automatic Sinica，2002，28(1)：103-114.

[25] Perrone M P. General Averaging Result for Convex Optimization [C]. In：Mozer M C，et al. eds. Proceedings 1993 Connectionist Models Summer School，Hillsdale，NJ：Lawrence Erlbaum，1994. 364-367.

[26] Jacobs R A，Jordan M L，et al. Adaptive Mixtures of Local Experts[J]. Neural Computation，1991，3(1)：79-87.

[27] Mozer M C，Smolensky P. Skeletonization：A Technique for Trimming the Fat From a Network Via Relevance Assessment[C]. In：Touretzk D S . eds. Advances in Neural Information Processing Systems，1989. 107-115.

[28] Bishop C M. Neural Networks for Pattern Recognition [A]. London：Oxford University Press，1995.

[29] Katerina Hlavackova-Schindler，Manfred M Fischer. An incremental algorithm for parallel training of the size and the weights in a feed forward neural network [J]. Neural Processing Letters，2000，11：131-138.

[30] 贺 伟，潘 泉，张洪才. 贝叶斯网络结构学习的发展与展望[J]. 信息与控制，2004，33(2)：185-190.

[31] 王新洲，史文中，王树良. 模糊空间信息处理[M]. 武汉：武汉工业大学出版社，2003.

[32] 懂 聪，郭晓华. 计算智能中热点问题[J]. 计算机科学，1999，26(4)：5-9.

（责任编辑：黎贞崇）