第 20 卷 第 9 期
2000 年 9 月

光 学 学 报
ACTA OPTICA SINICA

Vol. 20, No. 9
September, 2000

利用汉明距离优选神经网络学习样本*

申金媛 刘 玥 张文伟 陈 成 郭鹏毅 宋 庄 张延炘

(南开大学现代光学研究所, 国家教育部光电信息科学技术开放实验室, 天津 300071)

摘 要 鉴于学习样本对神经网络模型的模式识别性能有很大的影响, 提出学习样本的选择应与识别模型所利用的特性相结合, 并利用汉明(Hamming)距离对用于旋转不变识别的级联模型的学习样本进行优选, 计算机对三个很相似的飞机模型进行识别, 识别结果表明对学习样本进行有效的选择不仅可以减少系统的学习训练时间而且可以提高模型的识别能力。

关键词 神经网络, 模式识别, 学习样本, 级联模型。

1 引 言

利用神经网络进行模式识别, 其性能与许多因素有关, 除了学习算法、网络的结构和规模外, 学习样本集的合理选择也至关重要[1~5]。在利用三维物体的不同侧面(即利用三维物体面内外旋转时的二维投影面)进行实时的多目标识别时, 要把同一物体的所有不同投影面识别为一类, 由于每个物体有着无穷的投影面, 因此不同物体的某两个投影面可能很相似, 而同一物体的某些投影面可能差别很大, 所以进行多目标识别时, 学习样本的选择尤为重要。

本文的识别目标是三个飞机模型, 改变了以往由计算机产生目标的方法, 向真正的实用又迈进了一步。对于实物采集样本, 我们不可能事先给出识别目标在空间中的所有变换形态对网络进行训练, 每一次由 CCD 摄像机实际采集的新样本图像同过去的样本相比都会有一些不同, 有的甚至相差甚远。可以认为样本的变化是无穷的, 但 "万变不离其宗"。可以通过优选学习样本, 使网络通过有限的样本学习后能理解识别目标的本质特征而真正认识目标, 使得网络得到好的识别效果, 即选择尽可能少的学习样本而使神经网络模型具有最大的推广能力[6]。

本文提出学习样本的选择应与模型进行识别时所利用的模式的特征有关, 针对级联模型的识别原理, 提出利用汉明距离对学习样本进行优选, 并对三个很相似的飞机模型进行识别。选择样本时, 首先利用神经网络的容错性和推广能力选择一定的汉明距离 α 作为学习样本间的间距, 将相互间汉明距离小于 α 的样本进行聚类, 并选择其中之一作为学习样本集; 其次增加边界样本作为学习样本, 所谓边界样本是指不同目标的样本之间存在有较大的重

* 国家自然科学基金(69877005)、国家科委 863 高技术计划及天津市 21 世纪青年基金及天津市高教局资助项目。

收稿日期: 1999-01-15; 收到修改稿日期: 1999-05-12

叠，其汉明距离小于 α 但大于系统的识别精度。显然系统对这样的样本很容易出现误判，因此，应在学习样本中多增加这样的样本，系统的识别精度由网络的学习算法和结构以及网络的规模所决定；最后应去掉矛盾样本，所谓矛盾样本是指在系统的识别精度范围内无法区分的样本，即在已知的目标样本中有些样本可能出现既可以被归为 A 类，又可以被归为 B 类的状态，显然如果将矛盾样本作为学习样本，将对网络的识别造成消极影响。对三个很相似的飞机模型进行计算机模拟识别的结果表明：对学习样本进行有效的选择不仅可以减少系统的学习训练时间而且可以提高模型的识别能力。

2 级联模型及其多目标旋转不变识别

单层网络只能解决线性可分问题，而不变性识别是一个线性不可分问题，所以无法用单层网络实现。拓扑理论证明，用神经网络若要实现任意的空间映射(异联想)，网络结构除了输入层和输出层以外，至少包含一个中间层[7]。级联神经网络是一种前馈型神经网络，它将几个相互独立的网络前后依次连接而成，前一层的输出即是后一层的输入。重要的是，每一层在结构和功能上被赋予明确的意义，而且相对独立。

本文采用的级联神经网络模型由两级网络组成，第一级网络是一个多对一的异联想网络(Hetero-Associative)，实现对输入目标的不变性编码；第二级网络为“胜者全取”(WTA)网络。第一级网络包括输入层 L_1 和输出层 L_2(第一隐藏层)，输入层与输出层间的互连权重是基于 Hebb 规则的相关学习算法来求解的，并以此实现对输入目标的不变性编码。对于每一目标都事先指定其对应于一组 8 位编码 C_k(即输出层八个神经元的期望输出)，$k=1,2,3$ 对应于三个识别目标。显然对于任意一个学习训练样本，根据其所属目标都对应于某个编码。若设定神经网络的互连权重为 W_{ij}，第 k 个目标的第 m 个学习样本的第 i 个神经元的状态为 S_i^{km}，与之对应的编码为 C_k，则互连权重值 W_{ij} 为

$$W_{ij}=W'_{ij}+\Delta W_{ij}=W'_{ij}+\lambda\sum_{k=1}^{3}\sum_{m=1}^{M}S_i^{km}C_{jk}, \tag{1}$$

W'_{ij} 为互连权重的初始值，一般为随机选取，$i=1,2,\cdots,N$，$j=1,2,\cdots,8$，$k=1,2,3$，M 为每个目标的学习样本数目，λ 为学习步长。设输入模式为 X，其第 i 个神经元状态为 X_i，输出层第 j 个神经元状态为 Y_j，则输出层第 j 个神经元的总输入为

$$U_j=\sum_{i=1}^{N}W_{ij}X_i=\sum_{i=1}^{N}W'_{ij}X_i+\lambda\sum_{k=1}^{3}C_{jk}\sum_{m=1}^{M}A_{km}, \tag{2}$$

其中，$A_{km}=\sum_{i=1}^{N}S_i^{km}X_i$ 为输入模式与第 km 个学习样本间的内积，根据神经元激发函数，输出层神经元的实际输出为

$$Y_j=f(U_j-\theta_j)=f\left(\sum_{i=1}^{N}W'_{ij}X_i+\lambda\sum_{k=1}^{3}C_{jk}\sum_{m=1}^{M}A_{km}-\theta_j\right), \tag{3}$$

其中 θ_j 为该神经元的激发阈值，$j=1,2,\cdots,8$，$f(x)$ 为硬取阈函数。若输入模式为第 2 个目标，则正确的输出应该满足：$Y_j=C_{j2}$。

为提高整个网络的识别能力和容错能力，仅有第一级网络并不够，还需要第二级网络对第一级网络输出的码进行容错识别，并给出明确的识别结果。本级联模型采用具有极大存储容量和容错能力的“胜者全取”网络作为第二级网络。其输入层即为第一级网络的输出层 L_2，此外它还包括隐藏层 L_3 和输出层 L_4，它将第一级网络输出的异联想结果和三种待识别目标对应的码进行比较，然后根据相似度判断输入样本属于哪一类识别目标，“胜者全取”网络的

高容错性可以纠正第一级网络输出的部分误码，提高了整个模型的正确识别率。由于实行分层学习，与通常多层网络的 BP 算法相比大大地简化了学习程序，提高了学习效率，在实际应用中取得了较好的效果。

3　学习样本的选择方法

学习的目的在于应用和推广。神经网络之所以在模式识别中有广泛的应用，与它所具备的推广性能是十分重要的一个因素。在网络模型及规模选定以后，学习样本集的选择是提高神经网络推广性的关键环节。

以往对三维目标进行旋转不变识别时，不论选取何种网络都以物体旋转角度(即样本的取向角度)作为样本间的距离，认为两个样本间的取向角度越大，两个样本间的差别越大，选择学习样本时以均匀距离进行选择。显然，这样选择样本并不一定合理，实际上样本的选择应与所采用的识别模型相结合，与模型识别目标时所利用的目标的特征结合起来才可能选出好的学习样本。

由第二部分可知，本文采用的级联模型是利用样本之间的内积相似度或汉明距离[定义样本 X 与样本 S^{km} 之间的汉明距离: $d_H(X, S^{km}) = N - XS^{km} = N - A_{km}$] 作为主要的特征进行识别的，所以以角度作为距离选取样本不合理。如图 1 所示: 第一组中两个样本间的角度几乎正交，第二组样本中样本的空间取向则基本一致，但第一组两个样本间的汉明距离为 779，第二组两个样本间的汉明距离为 789。可见，存在这样的可能性，角度距离小但汉明距离大，这说明在级联模型中用角度距离选择样本不好。针对级联模型的特点，本文利用样本间的汉明距离选择学习样本。

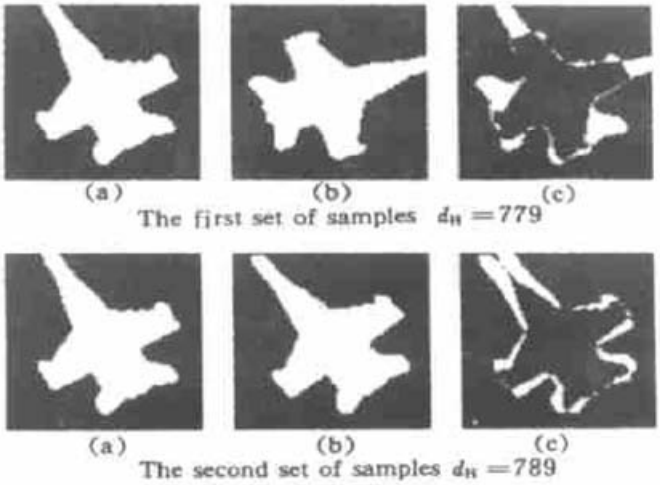

Fig. 1 Distance d_H between two samples

设 K 类模式的样本集分别为 A_1, A_2, …, A_K, 样本集 $A_k = \{a_{k1}, a_{k2}, \cdots, a_{kM_i}\}$ 有 M_i 个样本，每个样本是 N 维空间的一个向量。定义样本 x 和 y 之间的距离为 $d(x, y)$:

$$d(x, y) = \| x - y \|_P = \left(\sum_{i=1}^{N} | x_i - y_i |^P\right)^{1/P}, \tag{4}$$

其中 P 为任意数，为了简便，通常取 $P = 1$。也可以将 $d(x, y)$ 定义为汉明距离:

$$d(x, y) = d_H(x, y) = \frac{1}{2}(N - xy), \tag{5}$$

根据样本间的距离就可以在样本空间进行样本聚类和分离。即将样本空间划分为不同的区间，学习样本尽可能在不同的区间内选取。学习样本集的选择步骤如下:

1)　去除多余样本

实时采集的样本经预处理后不仅有旋转变化，而且还有平移、尺度变化。为更有效地选择学习样本，先除去平移因素，即在选择样本前，将每一样本图像的矩心移至 80×80 像素矩阵的几何中心(见图 2)。把样本图像看作矢量，求出同一目标各样本矢量间的汉明距离。根

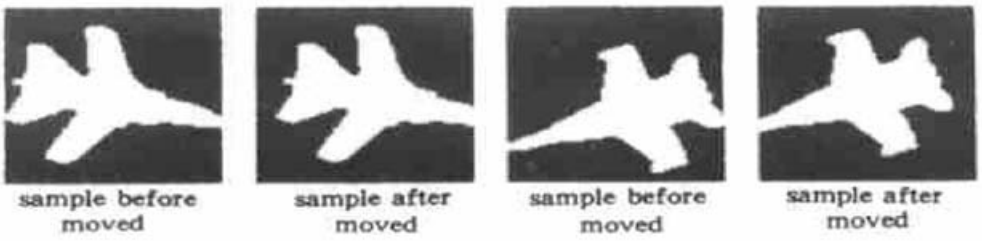

Fig. 2 Examples of moving the moment centers of images to the geometric center of sample matrix

据实际情况选定 $\alpha = 600$, 将相互间汉明距离小于 α 的样本聚类, 从每一个聚类中挑选一个作为学习样本。为了保证学习样本集中包含平移因素, 在每一聚类中, 再求出样本矩心未移动前各样本和已选出的学习样本之间的汉明距离。设定 $\alpha' = 1000$, 将距离大于 a' 的样本保留下来作为学习样本, 而把其它样本作为多余样本去除。

2) 选择边界样本

不同识别目标的样本之间, 存在某种程度的重叠。定义那些存在部分重叠, 而又在系统识别精度范围内的样本为边界样本。为提高系统的识别性能, 选择学习样本时应多选择一些边界样本。利用上面给出的距离公式, 对于给定的临界值 δ, 定义边界样本集合 B 为

$$B = \{x \mid x \in A_i, \exists A_j, d(x, A_j) < \delta, i \neq j, i, j = 1, 2, \cdots, K\}, \tag{6}$$

B 集合中的任一元素 x 称为边界样本。(6) 式中, $d(x, A_j) = \min\limits_{l=1, 2, \cdots, n_j} d(x, a_{jl})$, 根据边界样本集合 B 的定义, 可以定义第 k 种模式的样本集 A_k 的边界, 记为 $b(A_k)$:

$$b(A_k) = A_k \cap B, \tag{7}$$

$b(A_k)$ 中的元素称为样本集 A_k 的边界样本。很显然, 有 $B = \bigcup\limits_{i=1}^{K} b(A_k)$。同样可以定义 A_k 和其边界样本集的差集为 A_k 的内部样本集 $c(A_k) = A_k - b(A_k)$。$c(A_k)$ 中的元素为样本集 A_k 的内部样本。

显然, 边界样本比较容易出错, 所以学习样本中应包含一定比例的边界样本, 以提高系统的识别能力。按照(6)式的边界样本定义, 适当增加边界样本的比例。通过对三种识别目标的样本图像观察, 飞行器 F_1、F_2 之间出现边界样本的可能性较大, 而前两种和飞行器 F_3 的差别都较大, 出现边界样本的可能性不大。所以, 主要考虑识别目标 F_1、F_2 之间的边界。δ 的选取根据实际目标及实际需容错的能力来选择, 在本实验中 δ 等于 600, 即当样本之间距离小于 600 且不是矛盾样本时, 则认为它属于边界样本。

3) 去掉矛盾样本

如果样本在不同识别目标的样本之间的重叠性太高, 以致于超过了系统的识别精度, 无法判断它属于哪一类目标, 定义这种样本为矛盾样本。为了提高系统的识别性能, 在选择学习样本时尽量去掉矛盾样本。利用上面给出的距离公式, 对于给定的小正数 ϵ, 定义矛盾样本集合 C 为

$$C = \{x \mid x \in A_i, \exists A_j, d(x, A_j) < \epsilon, i \neq j, i, j = 1, 2, \cdots, K\}, \tag{8}$$

C 集合中的任一元素 x 称为矛盾样本。

矛盾样本是指归类含糊的样本。矛盾样本既可以被归入类 A_i, 又可以被归入类 A_j。很显然, 如果采用矛盾样本去训练网络, 将会混淆网络的判断。

在给定的范数空间中, 定义矛盾样本 $a \in A_i$, 存在样本集 A_j, a 满足下述条件:

$$d(a, A_j) < \epsilon \tag{9}$$

则称 a 为矛盾样本。其中，由于 ϵ 为定义矛盾样本的临界值，故必有 $\epsilon < \delta$(δ 是定义边界样本集的距离)。这里取 $\epsilon = 250$, $\delta = 600$。

4) 扩充样本集至整个样本空间

在同一模式中，如果样本之间的距离过大，则仅用这些样本无法完全表示出模式的特征，这将影响到系统的推广性能。所以需要插入一些新的样本，使样本集达到一定的致密性。先用较小学习样本集去训练网络，然后用新的样本进行测试，如果发生误判，则将此样本加入到学习样本集中。如此反复进行，样本集即可逐渐地扩充到整个样本空间，并且不会导致过多的学习样本数。

按照上面的四个步骤对三种识别目标学习样本进行选择，最后得到的学习样本如图 3~ 5 所示。飞行器 F_1 从 125 个样本中选取 25 个作为学习样本，飞行器 F_2 从 135 个样本中选取 30 个作为学习样本，飞行器 F_3 从 126 个样本中选取 25 个作为学习样本，其它样本作为测试样本。

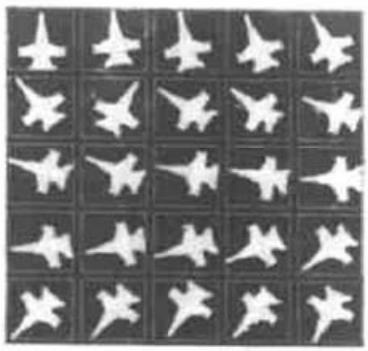

Fig. 3 Learning samples of flyer F_1　Fig. 4 Learning samples of flyer F_2　Fig. 5 Learning samples of flyer F_3

4　计算机识别及结论

用上面求得的(25+ 30+ 25)个学习样本代入(1)式学习算法中进行模拟时，$W_{ij}^{'}$ 是随机选取的，λ 取 0.1，得到互连权重，然后通过(3)式及"胜者全取"网络进行目标识别，对 85 个学习样本和 306 个非学习样本(测试样本)进行检验时模型都可以 100% 的识别，整个网络的学习训练时间在主频为 233 MHz 的微机上约为 4 个小时。

以角度距离均匀选择同样数目的学习样本(25+ 30+ 25)个代入(1)式求互连权重，并对学习样本及测试样本进行识别，对学习样本模型可 100% 的正确识别，但对 306 个测试样本仅可正确识别 260 个，即只有约 85% 的正确识别率。为了提高测试样本的正确识别率，需增加学习样本数。当每个目标的学习样本数增加到 50 个时(共 150 个)，测试样本的正确识别率方达 90% 以上，这时网络的训练学习时间在主频为 233 MHz 的微机上长达 10 多个小时。

计算机识别结果表明学习样本对模型的学习训练时间及正确识别率有很大的影响，学习样本的有效的选择不仅可以减少系统的学习训练时间而且可以提高模型的识别能力。

参 考 文 献

[1] 黄德双, 马德颂. 关于前馈网络分类器的研究进展. 见: 靳蕃, 范俊波主编. 神经网络理论与应用研究'96. 重庆: 西南交通大学出版社, 1996. 62~ 69

[2] 胡飞, 靳蕃. 模式分类学习样本的重构原则. 见: 靳蕃, 范俊波主编. 神经网络理论与应用研究'96. 重庆: 西南交通大学出版社, 1996. 520~ 524

[3] Baum E B, Haussler D. What size net gives valid generalization. *Neural Computation*, 1989, **1**(1): 151~ 160

[4] Schwartz D B, Samalam V B, Solla S A *et al*.. Exhaustive learning. *Neural Computation*, 1990, **2**(3): 374~ 385

[5] 张鸿宾. 训练多层网络的样本数问题. 自动化学报, 1993, **19**(1): 71~ 77

[6] 何振亚. 神经智能——认知科学的若干前沿问题研究. 南京: 东南大学出版社, 1996. 58~ 63

[7] Kan T C. Ill-posed problem of the constant-constraint criterion in optical correlation pattern recognition. *J. Opt. Soc. Am* (*A*), 1995, **12**(10): 2114~ 2121

Selecting of Learning Samples Based on Hamming Distance

Shen Jinyuan Liu Yue Zhang Wenwei Chen Shu
Guo Pengyi Song Zhuang Zhang Yanxin

(*Nankai University, Institute of Modern Optics, Optoelectronic Information Science and technology Lab, Education Ministry of China, Tianjin* 300071)

(Received 15 January 1999; revised 12 May 1999)

Abstract Learning samples significantly affect the recognition ability of neuron network models. One of selecting rules of learning samples is proposed according to the principle of the pattern recognition model. A method of selecting learning samples based on Hamming distance used in the cascade neuron network model for rotation invariance recognition is analyzed. The results of the computer recognition show that the effective selection of the learning samples can not only reduce the training time but also improve the recognition ability of the model.

Key words neural network, pattern recognition, learning sample, the cascade neural network model.