

刘 旸, 官 莉. 人工神经网络法反演晴空大气湿度廓线的研究[J]. 气象, 2011, 37(3): 318-324.

人工神经网络法反演晴空大气湿度廓线的研究^{* 1}

刘 旸 官 莉

南京信息工程大学大气物理学院, 南京 210044

提 要: 高光谱分辨率大气红外探测器 AIRS (Atmospheric Infrared Sounder) 作为第一个超高光谱大气红外探测仪, 开辟了卫星大气探测的新时代。以无线电探空值与 SARTA (Stand-Alone Radiative Transfer Algorithm 辐射传输模式) v1.05 版的前向模式模拟出的 AIRS 辐射亮温值组成样本对, 利用神经网络法反演大气湿度廓线。将反演所得的结果与特征向量统计法的反演结果进行比较, 结果表明, 与特征向量统计法相比, 神经网络法反演精度高, 所获得的水汽廓线更加贴近真实廓线。AIRS 因其高光谱分辨率 (即高垂直分辨率) 显示了精细的大气结构。在基于高光谱资料反演大气湿度廓线技术上, 神经网络显示出较强的非线性处理能力。

关键词: 神经网络, BP 网络, AIRS, 湿度反演, 特征向量统计法

Study on the Inversion of Clear Sky Atmospheric Humidity Profiles with Artificial Neural Network

LIU Yang GUAN Li

School of Atmospheric Physics, Nanjing University of Information Science and Technology, Nanjing 210044

Abstract: AIRS (Atmospheric Infrared Sounder), the first high spectral atmospheric infrared detector, started a new era of satellite sounding atmosphere. The samples, composed of the radiosonde observations and the AIRS brightness temperature value simulated by the SARTA (Stand-Alone Radiative Transfer Algorithm) v1.05 forward mode, were inverted to the atmospheric humidity profiles using neural network. The results show that, compared with the result of eigenvector statistics, the neural network inversion is of higher precision, and the humidity profiles obtained are closer to the true profiles. AIRS displays the fine structure of the atmosphere because of its high spectral resolution (high vertical resolution). Neural network has a strong nonlinear processing capability in the issue that the inversion of the atmospheric humidity profiles is based on high spectral data.

Key words: neural network, BP network, AIRS (Atmospheric Infrared Sounder), humidity retrieval, eigenvector statistics algorithm

引 言

在过去的几十年中, 大气温度、湿度等廓线的反演已经取得一定进展, 然而其精度并没有达到数值天气预报的要求, 主要原因有两个: 第一, 目前业务上使用是多分辨率的红外垂直探测器, 受遥感技

术水平限制, 光谱分辨率较低, 反演精度和分辨率不如人意。小尺度的垂直结构很难反演^[1], 是当前反演面临的一大难题。解决这一问题的办法在于发展高光谱分辨率的红外垂直探测仪器。第二, 反演处理上存在的一些问题以及大气透过率计算的误差, 下垫面的一些未知因素, 云的影响等。解决办法是提高大气透过率计算精度, 发展一套高效率的反演

* 公益性行业科研专项 (GYHY200806014) 资助

2010 年 4 月 22 日收稿; 2010 年 8 月 3 日收修定稿

第一作者: 刘旸, 主要研究气象卫星资料处理. Email: night_elf1986@qq.com

算法。由于反演问题是一个非适定性问题,即其解不唯一、不稳定,必须在求解过程中增加一些附加条件后才能得到一个稳定解。

美国地球观测系统 EOS(Earth Observing System) 第二颗卫星 Aqua 上的大气红外探测器 (AIRS) 是一个高光谱红外垂直探测器,它可以实现的光谱分辨率达到了分辨大气成分单个谱线的水平,使大量通道同时遥感反演大气温、湿廓线成为可能^[2]。目前国内基于 AIRS 资料反演大气温度、湿度垂直廓线大多采用的是特征向量统计反演法,该方法本质是一种基于最小二乘法的统计回归算法^[3]。神经网络是一个由大量简单的处理单元组成的高度复杂的大规模非线性自适应系统,可以实现输入空间到输出空间的非线性映射。20 世纪 80 年代中期以来,神经网络的应用研究取得了很大的成就,涉及面非常广泛。误差反传算法,即 BP 算法 (Back Propagation),已成为至今影响最大、应用最广泛的一种网络学习方法^[4]。20 世纪 90 年代,国内外在气象学科中开展了很多有关神经网络预报建模和气候分析等应用的研究^[5-6],在中、长期天气预报应用中也有不少成功的例子^[7]。近几年来,人工神经网络已经应用于降雨的预报^[8]、AMSU-A (Advanced Microwave Sounding Unit-A,先进的微波垂直探测器) 仪器微波观测资料的反演^[9] 以及数值预报产品的释用^[10]。基于 AIRS 高光谱资料采用神经网络法进行模拟资料的晴空大气温度廓线反演,较之特征向量统计法,各高度层精度均有不同改进,获得了较好的反演结果^[11,19];而实况资料的温度廓线反演,神经网络法的反演精度并没有明显提高^[12],原因之一是所选的神经网络结构并不是泛化性能最佳的网络。本文尝试用人工神经网络法反演大气湿度廓线,着重优化网络结构,以提高精度,提高适用性。

1 训练样本

样本是由全球晴空反演训练样本廓线和相应的 SARTA 模式模拟的亮温值组成样本对。CIMSS (Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin-Madison, 气象卫星合作研究所,威斯康星大学麦迪逊分校) 全球晴空反演训练样本集由 TIGR3 (TOVS Initial Guess Retrieval, 泰罗斯业务垂直探测器初始猜测回归场)、

NOAA-88 (National Oceanic and Atmospheric Administration, 美国国家海洋和大气局) 和 ECWMF (European Center of Medium-range Weather Forecasts, 欧洲中期天气预报中心) 等全球有代表性的温、湿和臭氧等廓线组成,同时还补充了沙漠地区的无线电探空和臭氧探空值,所有廓线从 1100 hPa 到 0.005 hPa 被内插到 101 个气压层,共 12208 条廓线样本,并按 IGBP (International Geosphere-Biosphere Programme, 国际地圈生物圈计划) 生态系统进行分类,其中 15 个红外通道的地表比辐射率由与生态系统相关的 MODIS (moderate-resolution imaging spectroradiometer, 中分辨率成像光谱仪) (MOD11) 产品得到。廓线地区分布在海洋上较稀疏,在大陆上较发达地区分布密集,如图 1。对全球晴空反演训练样本集使用 SARTA v1.05 版的前向模式 (该软件包是 AIRS-RTA 大气红外探测器辐射传输算法的实际应用,是一个用于计算 AIRS 辐射传输值的快速前向模式) 模拟出 AIRS 辐射亮温值。如果前向模式的精度能达到仪器噪声的水平, AIRS 具有的高光谱分辨率及低噪声的特性可使得到与现有的世界上业务无线电探测网络相媲美或更精确的大气廓线。

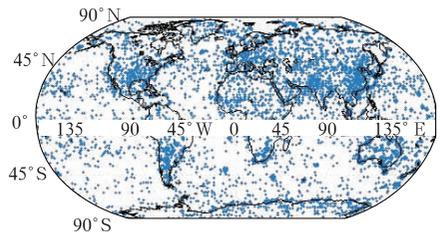


图 1 12208 条廓线全球分布情况

Fig.1 The global distribution of 12208 profiles

2 人工神经网络法

神经网络是一个由大量简单的处理单元组成的高度复杂的大规模非线性自适应系统,具有良好的学习功能。它可以接受用户提交的样本集合,依照系统给定的算法,不断地修正用来确定系统行为的神经元之间连接的强度,并存放于系统中。

神经网络包含许多网络模型,BP 算法是应用最为广泛的多级网络训练算法。BP 神经网络的基本原理是,网络得到输入以后,经过加权并通过各层激励函数的计算得到输出层的输出值。将这个值与期望值做比较,如果均方根误差在理想范围内,则认为

网络学会这一问题,停止训练,保存网络的权值,建立该网络;如果均方根误差未达到预期值则把误差反向传播,根据误差修改各连接权重,进行下次训练。如此反复,直到网络输出期望输出的均方根误差达到预期标准。

相对传统的数理统计方法而言,BP神经网络可以求解非线性问题,同样对样本大小的要求也可以相对少得多(如果样本能够代表该问题的各种类型特征),并不要求样本独立或遵从正态分布,还具有较强的容错能力,即网络系统的判别精度一般不受样本中噪声的影响。

2.1 BP网络的设计

BP网络可有效地用于复杂的非线性函数的逼近,一个3层的前馈网络能够实现任意精度的连续函数映射^[13]。BP网络模型见图2。

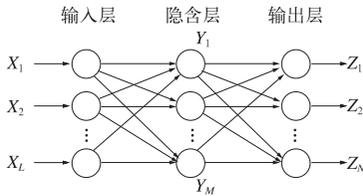


图2 BP网络模型示意图

Fig. 2 BP network model diagram

2.2 亮温温度分类

为提高反演的精度,将训练样本根据波数 1000 cm^{-1} (窗区)时的亮温温度分成6类,分别训练网络,反演时不同亮温分类采用不同的网络。表1给出亮温分类每类阈值。为减少错误分类的影响,亮温分类时每类的阈值重叠了 10 K ^[14]。

表1 亮温温度分类方案

Table 1 The classification scheme of brightness temperature (K)

类别	训练: 1000 cm^{-1} 的 亮温温度/K	反演: 1000 cm^{-1} 的 亮温温度/K
1	亮温温度 ≤ 260	亮温温度 ≤ 255
2	$250 < \text{亮温温度} \leq 270$	$255 < \text{亮温温度} \leq 265$
3	$260 < \text{亮温温度} \leq 280$	$265 < \text{亮温温度} \leq 275$
4	$270 < \text{亮温温度} \leq 290$	$275 < \text{亮温温度} \leq 285$
5	$280 < \text{亮温温度} \leq 300$	$285 < \text{亮温温度} \leq 295$
6	$290 < \text{亮温温度}$	$295 < \text{亮温温度}$

2.3 加噪处理

为确保模拟的真实性,在进行反演前,需要给计算的辐射值增加些模拟的AIRS仪器噪声。噪音根

据高斯分布随机产生,平均值取零,标准差为AIRS通道特性文件中的 $NE\Delta T$ (Noise Equivalent Delta Temperature, 等同于 250 K 亮度温度时噪音值)。

2.4 主特征分析的应用

主成分分析方法PCA(Principal Component Analysis)是系统降维和特征提取的一种基本方法。研究表明,对AIRS的2378个通道的观测值取其前30个主成分可以代表99.9988%的所有AIRS的信息量^[14]。因此,在我们研究中,就选取前30个主成分。

将经过主成分分析之后的数据应用于大气温湿度廓线反演具有明显的优越性:(1)直接去除特征值为零或接近于零的特征值对应的主分量。(2)只保留与反演量相关程度高的主分量,降维作用直观显著。(3)由于主分量间是正交的,所以不会有多余的重复信息噪声影响。

2.5 输入输出强欠定问题的处理

由于我们选取前30个主成分作为输入,101层湿度廓线作为输出,造成了网络结构的强欠定性,会使得网络的训练预测能力下降。这里采取一种方法,将作为输出的101层湿度廓线进行分段,分段依据特征向量统计法反演的水汽混合比相对于真值的均方根误差RMSE(Root Mean Square Error)曲线。如图3所示:可以看出,从近地面到800 hPa处,

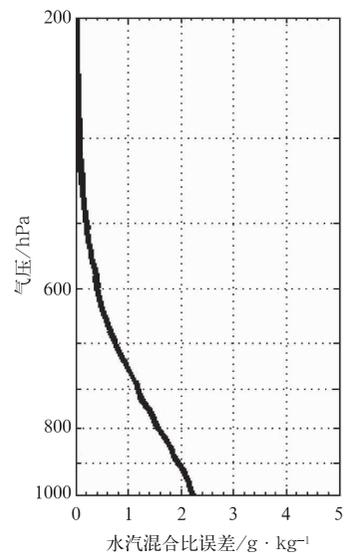


图3 特征向量统计法反演的水汽混合比与真值的均方根误差曲线

Fig. 3 The root mean square error curve between the true value of water vapor mixing ratio and the value retrieved by eigenvector statistics

RMSE 约为 2; 从 800 hPa 到 600 hPa 处, RMSE 约为 1; 从 600 hPa 到接近 250 hPa 处, RMSE 小于 1; 250 hPa 以上的高空, RMSE 几乎为 0。因此将 101 层湿度廓线分成 4 段(如表 2)。

表 2 廓线分段方案

Table 2 The plan of the profile section

分段号码	层号	气压大概范围/hPa
1	1~60	0~250
2	61~80	250~600
3	81~90	600~800
4	91~101	800~1100

由此可见,分段以后除了第一段的变量数仍大于输入以外,其他三段的输出均少于输入。经验证,这样分段训练,不仅提高了训练精度,并且大大地节省了训练时间。需要指出的是,这里的分段方案只是根据经验,综合考虑网络训练的精度与速度而制定的一种分类方案,并不一定是最优方案。

2.6 过拟合问题的处理

衡量网络训练的性能好坏依据 MSE (Mean Square Error, 均方差) 这个参量。一般来讲, MSE 越小, 网络训练的精度越高。但是, 神经网络容易出现过拟合现象, 即网络训练到一定程度时, 随着训练的能力提高, 预测的能力反而下降。不仅浪费时间, 而且最终结果可能也达不到要求, 解决这个问题的一种办法是: 在数据输入中, 给训练的数据分类, 分为正常训练用、变量数据、测试数据, 三种数据的比例这里采用 60%、20%、20%。为避免相邻的样本具有相关性, 每 5 个样本中选取 1 个作为变量数据, 1 个作为测试数据, 3 个作为训练数据。在网络训练的时候监测 Validation Checks 这个参量, 设定一个阈值。当训练网络过程中该参量达到阈值, 则停止训练。BP 算法很有可能陷入局部极值, 使训练失败。如果停止训练后, 网络训练没有达到预期的精度, 则重新训练, 直到网络训练达到目标精度为止。

2.7 激活传递函数的选取

激活函数是一个神经元及网络的核心。网络解决问题的能力与功效除了与网络结构有关, 在很大程度上取决于网络所采用的激活函数。激活函数的输入为神经元输入矢量的加权和。经试验, 二层均选取双曲正切 S 型函数: tansig 。网络输出的计算公式为:

$$A = \text{tansig}(W \times P + B) \quad (1)$$

其中, W 为权值, P 为输入矢量, B 为偏差。

2.8 隐节点数的选取

三层的 BP 网络中有一层隐含层, 当隐含层神经元数目足够多时, 可以以任意精度逼近任何一个具有有限间断点的非线性函数。隐含层节点数的设置关系到网络性能的好坏。怎样确定隐含层中适宜的神经元个数使模型的泛化能力最大, 至今没有很好的结论^[15-16]。隐节点数太少会造成信息不足, 造成网络训练达不到要求; 太多会造成浪费, 使训练时间加长。在试验中发现, 针对不同的样本, 即使输入层输出层节点数相同, 网络收敛最快的隐节点数值也可能不同。输入输出层节点数均相同的二组数据, 隐节点的选取和当时的网络初始权值也有关系。最后根据前人经验计算的数值作为参考值, 缩小选取隐节点数的区间。以不同的网络初始权值多次训练网络, 同时更改隐节点数, 以求达到较高的训练精度。

2.9 训练算法的选取

在选择算法对网络进行训练的时候, 在网络参数很多, 需要考虑存储容量问题时, 选择共轭梯度法^[17]。经试验, 这里选择 Scaled 共轭梯度法 train_scg 。该算法区别于其他算法的参数有 sigma 缺省设置为 $5e-005$, lambda 缺省设置为 $5e-007$ 。

3 反演结果

3.1 与特征向量统计反演法湿度反演结果比较

特征向量统计法是一种线性的、统计的算法^[18]。而 BP 网络是通过误差不断地反传来调整权值, 使得最终的网络输出与真值的误差达到预期要求。神经网络法是一种非线性的方法, 在非线性函数拟合上, 理论上要优于特征向量统计法。

采用上述全球 12208 个训练样本, 间隔选取 8138 个样本进行训练, 其他 4070 个样本用于独立的反演检验。分别用人工神经网络算法(简称 NNW)和特征向量统计方法(简称 ESW)进行反演, 将得到的结果与样本真值进行比较, 计算均方根误差 RMSE。

RMSE 误差定义为:

$$RMSE = \sqrt{\frac{1}{N_S} \sum_{i=1}^{N_S} (X_{\text{RAOB}} - X_{\text{AIRS}})^2} \quad (2)$$

其中 X_{RAOB} 和 X_{AIRS} 分别是样本廓线及 AIRS 反演的参数值, N_s 是样本个数。

图 4a 是反演的温度均方根误差, 图 a_1 为图 a_2 的局部。图 4b 是反演的水汽混合比均方根误差, 图 b_1 为图 b_2 的局部。从图 4a 可以看出, 在温度反演问题上, 神经网络法反演的结果略优于特征向量统

计法, 但效果不明显。从图 4b 可以看出, 神经网络法反演的结果在所有的气压层上都要优于特征向量统计法的结果。在近地面到 500 hPa 层面, 这种优势更加明显。改进的最大值为 $0.7926 \text{ g} \cdot \text{kg}^{-1}$, 近地面到 500 hPa 平均改进值大于 $0.4 \text{ g} \cdot \text{kg}^{-1}$ 。

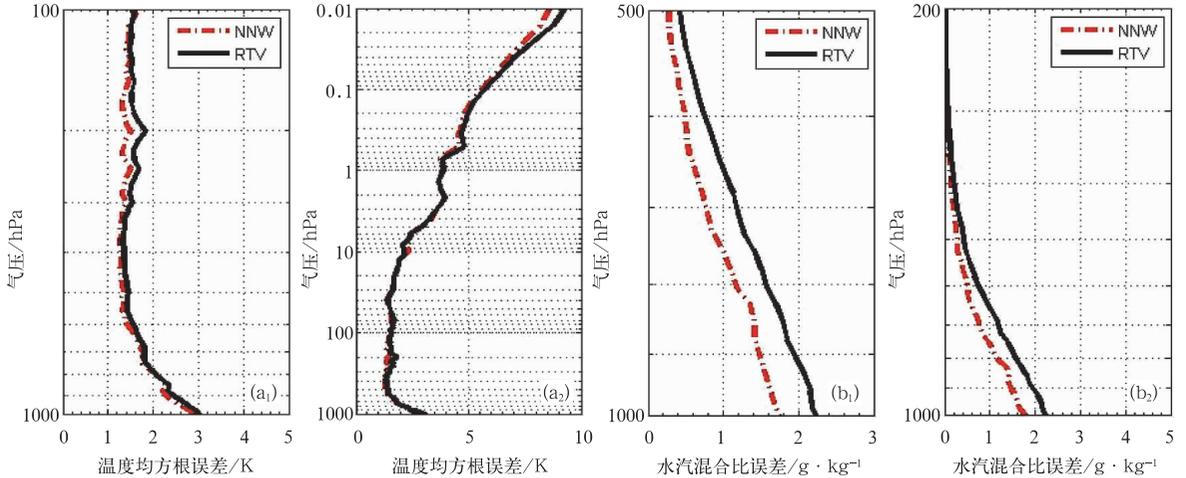


图 4 (a) 温度均方根误差 (图 a_1 为图 a_2 的局部); (b) 水汽混合比均方根误差 (图 b_1 为图 b_2 的局部)

Fig. 4 The RMS errors of temperature (a, K) and water vapor mixing ratio (b, $\text{g} \cdot \text{kg}^{-1}$)

In Fig. 4a and Fig. 4b, the left figure is the detailed part of the right figure

由于训练样本是从全球样本中挑出, 并没有考虑它不同的下垫面情况、地形、地表辐射等, 以及区域气候。采用全球样本训练是为了使训练好的网络可以推广到全球范围的反演。可以预见, 若想反演某一区域的大气廓线, 可以采用该区域的样本训练网络, 以达到更精确的结果。

3.2 湿度廓线的拟合

为判断两种反演方法的拟合情况, 将上述湿度反演均方根误差 RMSE 分成 4 类:

(1) 两种廓线的拟合情况均理想。从近地面到 200 hPa 的高空这近 50 个气压层, 平均每层与真值的绝对误差小于 $0.15 \text{ g} \cdot \text{kg}^{-1}$, 有 483 条廓线。

(2) 两种廓线的拟合情况均不理想。从近地面到 200 hPa 的高空这近 50 个气压层, 平均每层与真值的绝对误差大于 $1.5 \text{ g} \cdot \text{kg}^{-1}$, 有 253 条廓线。

(3) 神经网络法的反演结果优于特征向量统计法的反演结果。这类廓线是除去前二类之外二种方法进行比较所得, 有 2303 条廓线。

(4) 特征向量统计法的反演结果优于 NNW, 共

有 1031 条廓线。

这 4 种情况的廓线分布如图 5a~5d。图 5a 第一类廓线分布, 发现这些点大都集中在高纬度地区。图 5b 是第二类廓线分布, 发现这些点大都集中在山地, 或者赤道附近。由于山地下垫面复杂, 地表比辐射率复杂; 赤道水汽含量多, 增加了反演的难度。图 5c 是第三类廓线分布情况, 图 5d 则是第四类廓线分布。后两幅图的廓线分布从直观上看并未发现有什么明显特征。

在以上 4 类廓线中, 任取一条为例, 如图 6a~6d 所示。这 4 条廓线 101 个气压层的 RMSE 见表 3。其中 RTV RMSE 表示特征向量统计法反演结果与真值比较所得的均方根误差, NNW RMSE 表示神经网络法反演结果与真值比较所得的均方根误差。由这 4 幅图可以看出, 经特征向量统计法反演所得的廓线都比较平滑, 当真实廓线波动较大时, 拟合效果比较差。图 6b、6d 虽然神经网络法拟合的廓线的 RMSE 不是很理想, 但是本方法的廓线很好地拟合出了真实廓线的走向。

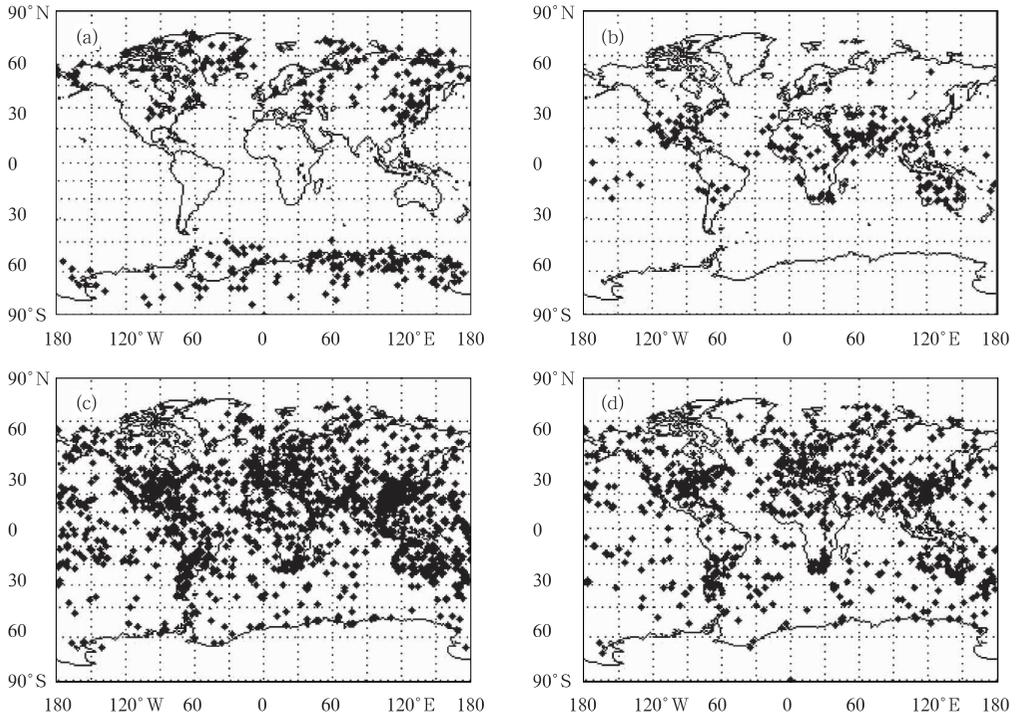


图 5 第一类廓线分布(a); 第二类廓线分布(b);
第三类廓线分布(c); 第四类廓线分布(d)

Fig. 5 The profile distribution of (a) the first kind, (b) the second kind,
(c) the third kind, and (d) the fourth kind

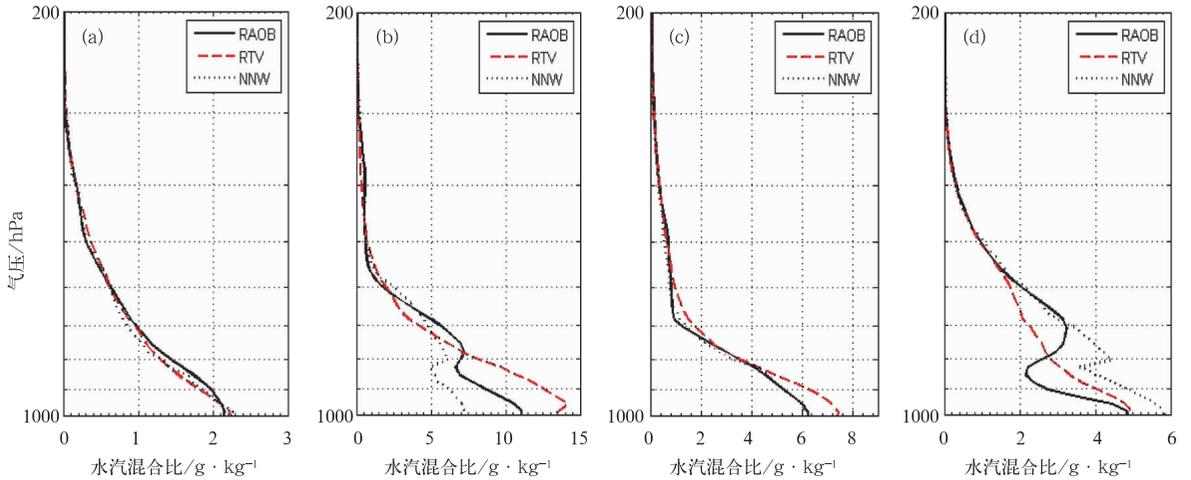


图 6 第一类廓线拟合(a); 第二类廓线拟合(b); 第三类廓线拟合(c); 第四类廓线拟合(d)

Fig. 6 The profile fitting of (a) the first kind, (b) the second kind, (c) the thrid kind, and (d) the fourth kind

表 3 4 条廓线的 RMSE

Table 3 The RMSE of four kinds of profiles

廓线序列号	1056(一类)	830(二类)	363(三类)	1726(四类)
RTV RMSE	0.0609	1.0252	0.3821	0.3330
NNW RMSE	0.0593	1.0350	0.0763	0.5374
经纬度	76.75°N	22.791°N	28.062°N	49.797°N
	137.875°E	39.391°E	122.273°E	7.234°W

4 总结与展望

本文基于 AIRS 模拟资料, 尝试使用神经网络法反演大气湿度廓线, 较之特征向量统计法, 本方法提高了反演精度, 并且神经网络法更能揭示真实大

气湿度廓线的细微结构。

理论上,一个 3 层的 BP 网络可以以任意精度逼近一个非线性函数。但是在实际应用中,尤其是这种多输入、多输出大量样本的情况下很难达到这种程度。本文针对网络过拟合的情况采取了对训练数据分类,在训练中监控某些参量,避免过拟合的发生。另外,采取气压层分段训练也解决了方程强欠定性问题,提高了训练精度,节省了训练时间。

由于样本数量庞大,怎样集中选取代表性的样本,这也是个值得思考的问题。如果选取得当,将缩短训练时间,提高训练精度。另外,由于没有考虑地形、地表辐射以及区域气候等因素,故个别廓线的反演上出现较大的误差。下一步将开展关于实况资料的反演及地形因子的订正等方面的研究,期待神经网络法带给我们更好的表现。

参考文献

- [1] 蒋德明,董超华,陆维松. 利用 AIRS 观测资料进行红外高光谱大气探测能力试验的研究[J]. 遥感学报, 2006, 10(4): 586-592.
- [2] 官莉. 星载红外高光谱资料的应用[M]. 北京:气象出版社, 2007:29.
- [3] Guan L, Huang H L, Li J. A study on retrieving atmospheric profiles from EOS/AIRS observations [J]. Acta Meteor Sinica, 2005, 19(1): 112-119.
- [4] 吴建生,周优军,金龙. 神经网络及其研究进展[J]. 广西师范学院学报(自然科学版), 2005, 22(1): 92-93.
- [5] 金龙,况雪源,黄海洪,等. 神经网络预报模型的过拟合研究[J]. 气象学报, 2004, 62(1): 62-70.
- [6] 覃志年,金龙,况雪源. 神经网络的短期气候定性预测方法研究[J]. 气象科技, 2004, 32(3): 168-172.
- [7] 陈宁,金龙,袁成松. 最优子集的神经网络预报建模研究[J]. 气象, 1999, 25(1): 14-19.
- [8] 李永华,刘德,金龙. 基于 BP 神经网络的汛期降水预测模型研究[J]. 气象科学, 2002, 22(4): 461-467.
- [9] Yao Zhigang, Chen Hongbin, Lin Longfu. Retrieving atmospheric temperature profiles from AMSU-A data with neural networks[J]. Advances in Atmospheric Sciences, 2005, 22(4): 606-616.
- [10] 钱莉,兰晓波,杨永龙. 最优子集神经网络在武威气温客观预报中的应用[J]. 气象, 2010, 36(5): 102-107.
- [11] 张雪慧,官莉,王振会,等. 利用神经网络方法反演大气温度廓线[J]. 气象, 2009, 35(11): 137-142.
- [12] 官莉,刘畅,张雪慧. 神经网络算法在红外高光谱资料反演大气温度廓线中的应用[J]. 大气科学学报, 2010, 33(3): 341-346.
- [13] 袁曾任. 神经网络及其应用[M]. 北京:清华大学出版社, 1999.
- [14] 官莉. 卫星红外超光谱资料及其在云检测、晴空订正和大气廓线反演方面的应用[D]. 博士论文, 2005, 46, 57.
- [15] Wu Yan, Zhang Liming. A new regularization learning method for improving generalization capability of neural network [R]. Proceedings of 4th World Congress on Intelligent Control and Automation. Shanghai, 2002, 3: 2011-2015.
- [16] 江学军,唐焕文. 前馈网络泛化性能的系统分析[J]. 系统工程理论与实践, 2000, 20(8): 36-40.
- [17] 丛爽. 面向 MATLAB 工具箱的神经网络理论与应用[M]. 合肥:中国科学技术大学出版社, 2003.
- [18] 官莉. 利用 AIRS 卫星资料反演大气廓线 I. 特征向量统计反演法[J]. 南京气象学院学报, 2006, 29(6): 757-758.
- [19] 周非非,周毓荃,王俊,等. FY-2 卫星反演的云顶高度与多普勒雷达回波顶高的关系初探[J]. 气象, 2010, 36(4): 43-50.