一类基于非线性 PCA 和深度置信网络的混合分类器及其在 PM2.5 浓度预测和影响因素诊断中的应用

高 月 [1] 宿 翀 [1] 李宏光 [1]

摘 要 传统的深度置信网络 (Deep brief networks, DBN) 在建立高维数据分类模型时, 往往存在网络负荷大, 运算复杂度高等问题. 本文首先基于非线性 PCA (NPCA) 对高维样本数据进行降维, 然后以提取到的非线性特征作为 DBN 的网络输入, 构建了一类含非线性特征提取预处理机制的 DBN 分类器. 并从信息熵理论的角度出发, 证明了所提改进 DBN 分类器在网络结构和算法复杂度方面的优势. 通过一个 PM2.5 浓度预测与影响因素诊断实例, 验证了所提改进 DBN 在一类分类和影响因素诊断问题中的应用, 并与传统的分类器进行对比, 显示了所提方法在建模精度及收敛速度上的优势.

关键词 深度置信网, 非线性主元分析, PM2.5, 信息熵

引用格式 高月, 宿翀, 李宏光. 一类基于非线性 PCA 和深度置信网络的混合分类器及其在 PM2.5 浓度预测和影响因素诊断中的应用. 自动化学报, 2018, **44**(2): 318−329

DOI 10.16383/j.aas.2018.c160045

A Kind of Deep Belief Networks Based on Nonlinear Features Extraction with Application to PM2.5 Concentration Prediction and Diagnosis

GAO Yue[1] SU Chong[1] LI Hong-Guang[1]

Abstract To build a classifier model of high dimensional data, the traditional deep brief networks (DBN) modeling method suffers from large network load and high algorithm complexity. In this work, the data dimension is reduced based on the nonlinear PCA (NPCA), then a new DBN classifier with nonlinear feature extraction pre-processing mechanism is proposed where the nonlinear feature is extracted as the network input to the DBN. With the entropy theory, the advantage of the improved DBN is proved in terms of network structure and algorithm complexity. A PM2.5 concentration prediction and diagnosis problem is employed to exemplify applications of the proposed methods. Compared with the traditional classifier, it shows the advantage of the proposed method in modeling accuracy and convergence speed.

Key words Deep brief networks (DBN), nonlinear-PCA (NPCA), PM2.5, entropy

Citation Gao Yue, Su Chong, Li Hong-Guang. A kind of deep belief networks based on nonlinear features extraction with application to PM2.5 concentration prediction and diagnosis. *Acta Automatica Sinica*, 2018, **44**(2): 318−329

众所周知, 聚类, 支持向量机及神经元网络等常见分类方法都属于浅层分类方法, 在处理蕴藏隐含信息的样本分类问题方面还存在不足. 传统的聚类方法对于高维数据来说, 数据样本较低维数据聚类时分布更为稀疏, 且每个数据间的距离都可能相当, 因此难以找到聚类中心, 从而不容易进行分类[1]; SVM 属于有监督学习算法, 在处理小样本分类时有一定优势, 然而该方法过于依赖样本数据尺度, 且算法复杂度较高. 并且 SVM 中核函数的选择往往决定了分类的精度和收敛速度, 分类结果存在不确定性[2]; 此外, 由于神经网络缺乏预训练机制, 难以深度挖掘数据中的隐含信息[3]. 然而深度学习方法以 "无监督训练−有监督调解全局网络参数" 的框架, 从理论上避免了传统神经元网络易陷入局部极值的缺点[4], 且在挖掘数据隐含信息方面具有独特优势, 尤其是在面临大规模样本数据的时候, 有更加突出的表现. 常见的深度学习分为自动编码器, 卷积神经网络和深度置信网络等模型. 通过查阅文献, 深度置信网络建模方法在图像处理[5−6]、软测量技术[7]、计算智能[8] 等诸多领域得到成功应用, 建模精确度普遍有所提升, 上述理论的成功应用, 为构建基于高维数据非线性特征提取的深度置信网络分类器提供了重要的理论和方法支撑. 值得一提的是, 面向含非线性特征的高维数据的深度置信网络建模问题, 信息量上

收稿日期 2016-01-21 录用日期 2016-12-18
Manuscript received January 21, 2016; accepted December 18, 2016
国家自然科学基金 (61603023), 北京市优秀人才资助项目 (2015000020124G041), 中国科学院复杂系统管理与控制国家重点实验室开放课题 (20150103) 资助
Supported by National Natural Science Foundation of China (61603023), Beijing Outstanding Talent Training Project (2015000020124G041) and the Open Research Project under Grant from the SKLMCCS (20150103)
本文责任编委 刘艳军
Recommended by Associate Editor LIU Yan-Jun
1. 北京化工大学信息科学与技术学院 北京 100029
1. School of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029

的冗余往往给网络带来不必要的负荷. 因此预先对样本数据做特征提取十分必要.

过程变量的特征提取的目的是找到数据之间的线性以及非线性关系表达, 而后利用提取的低维特征数据表征原有的高维数据. 故数据之间的线性以及非线性关系的提取是提取特征的关键. 常见的过程数据特征提取方法有主成分分析方法 (PCA), 独立主元分析 (ICA), 偏最小二乘法 (PLS) 等. 其中, PCA 利用高斯分布数据的特征, 将数据映射到正交的低维子空间上, 保留数据的特征[9]; ICA 根据已经存在的统计值, 进行独立主成分正交变换[10]; PLS 利用线性拟合对多变量建模, 减少变量个数[11]. 以上方法在数据满足高斯分布和有线性关系的情况下适用, 且效果很好, 但是, 在一类多变量数据且变量分布不定, 且存在非线性关系时, 以上方法并不奏效. 所以, 本文应用一类基于输入训练神经网络表征非线性主元分析的方法, 旨在解决在多变量过程中的非线性特征提取问题, 并且实现数据降维, 为后续构建一类新的深度置信网络提供数据预处理的方法支撑.

空气固体细微污染物 PM2.5 的形成, 受众多复杂因素影响 (已知影响因素超过 20 种)[12−13]. 就产生过程而言, PM2.5 可以由污染源直接排出 (称为一次粒子), 也可以是各污染源排出的气态污染物经过冷凝或在大气中发生复杂的化学反应而生成 (称为二次粒子). 特别地, 在已知的众多理化因素中, 有别于湿度、风速、降雨等, O_3 属于驱散因子, 其浓度与 PM2.5 浓度之间呈指数衰减规律, 此外, 其他因素 (光照等) 与 PM2.5 浓度的关系还有待探索[14]. 因此, PM2.5 浓度预测是一类典型的数据维度高, 且数据含非线性特征的建模问题, 传统的基于浅层学习的数据驱动建模方法[15−17] 在预测精度上还有待提升, 且不具备对 PM2.5 浓度影响因素进行诊断的功能.

受上述讨论启发, 针对过程变量数据维数高, 且含复杂非线性特征, 数据间隐含信息难以利用等特点, 本文提出一类基于非线性特征提取的深度置信网模型, 旨在解决高维数据非线性特征提取以及数据特征中隐含信息挖掘的问题, 并对影响模型输出的关键变量进行诊断. 最后, 以一类具体的多变量建模和诊断问题讨论所提方法的应用. 本文结构安排如下: 第 1 节展示了基于非线性特征提取的深度置信网络的建模过程; 第 2 节基于信息熵理论, 对改进后的深度置信网络的建模复杂度优势进行论证; 第 3 节以河北省某市的 PM2.5 监测数据为对象, 验证本文所提方法的有效性; 第 4 节给出结论与工作展望.

1　基于非线性特征提取的深度置信网络

本节提出一类基于非线特征提取的深度置信网网络模型. 基于非线性 PCA 提取原始数据特征, 实现数据预处理. 同时计算各变量的统计量, 作为影响因素诊断依据. 同时, 将预处理后的数据作为深度置信网的输入以构建预测模型. 改进的深度置信网络结构在下文中具体介绍.

1.1　数据预处理

高维多变量过程数据 (维度为 N) 之间存在的线性关系可以利用主成分分析的方法, 进行数据特征提取, 实现降维目的. 然而, 数据之间存在复杂的非线性关系时, 理论上同样可以利用 A 个主元 ($A < N$) 就可以反映出过程的主要信息. 非线性 PCA 就是一种对 X 的估计量 $\hat{X}$ 的非线性表示即:

$$X = \hat{X} + E = F(T_N) + E \tag{1}$$

其中, $\hat{X}$ 是 X 的估计矩阵, E 是残差矩阵, $F(\cdot)$ 是一个非线性函数, T_N 我们称之为非线性主元得分矩阵. 基于 Tan 等提出的输入训练 (Input-training, IT) 神经网络的方法[18]. 本文将 IT 网络的输入作为非线性主元得分矩阵, IT 网络的输出作为原始样本的估计值, 网络调节权值的时候, 不仅调节网络内部的参数, 输入也随之变换. 当网络训练完成的时候, 便可以得到 T_N, 同时也得到了非线性函数 $F(\cdot)$. 本文采用三层的输入训练网络, 如图 1 所示.

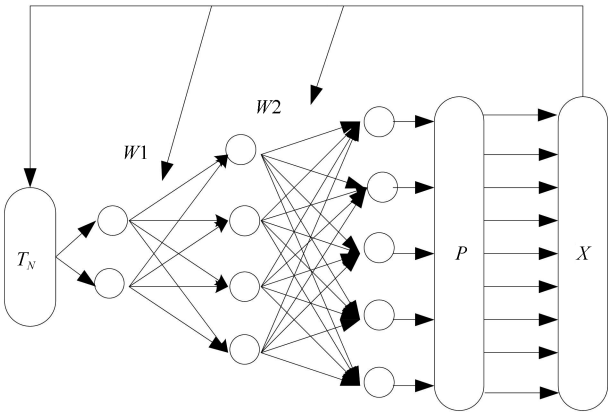

图 1　三层输入训练神经网络结构图

Fig. 1　An input training neural network structure with three layers

整体网络采用快速下降法调节网络间的连接权值. 网络的目标函数为 J:

$$J = \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{M}(X_{ij} - \hat{X}_{ij})^2 = \frac{1}{2}\sum_{j=1}^{N} e_j^2 \tag{2}$$

i 表示变量的维度, j 表示输出数据的组数, e_j 则表示每组训练数据的绝对误差.

1.2　深度置信网模型

深度学习是 Hinton 等在 2006 年提出的一种基

于概率表达网络模型[19]. 深度学习的技术可以分为两部分: 第一部分是利用无监督的学习来预训练每一层, 第二部分是全网络自上而下的微调权值. 由于无监督的方式, 使用所有无标签数据, 所以过程变量包含监督学习所不能表达的隐含信息. 本文所提出基于深度置信网的预测模型中, 网络输入是上一级降维后的非线性主元得分矩阵, 输出是预测分类结果. 其网络结构如下图所示. 本文采用三层的输入训练网络, 如图 1 所示.

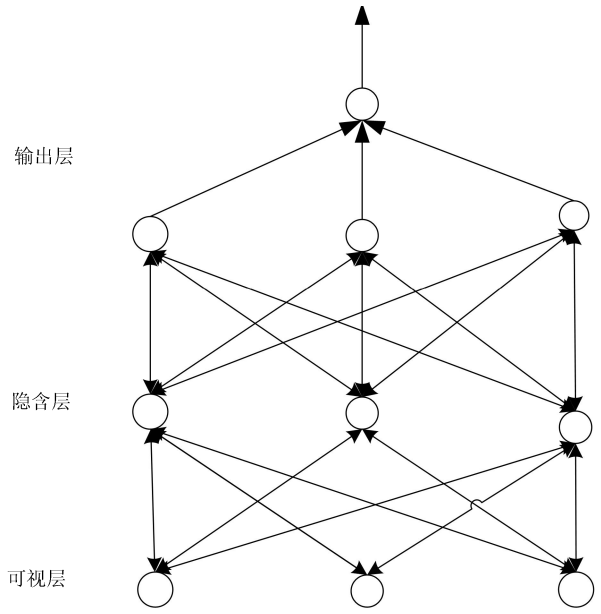

图 2 深度置信网的结构

Fig. 2 The structure of DBN

自上而下是多层的限制性玻尔兹曼机, 隐含层中每一层的输出, 作为下一层的输入. 在这个训练阶段, 在可视层会产生一个向量 $\boldsymbol{v}$, 通过它将值传递到隐层. 反过来, 可视层的输入会去重构原始的输入信号[20]. 我们定义联合概率分布:

$$P(\boldsymbol{v} \mid \boldsymbol{h}) = \frac{\exp(-E(\boldsymbol{v},\boldsymbol{h}))}{\boldsymbol{z}} = \frac{1}{\boldsymbol{z}} \prod_{ij} \mathrm{e}^{W_{ij} v_i h_j} \prod_i \mathrm{e}^{b_i v_j} \prod_j \mathrm{e}^{a_j h_j} \tag{3}$$

其中 $\boldsymbol{z}$ 为:

$$\boldsymbol{z} = \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h})) \tag{4}$$

v_i 表示可视层第 i 个节点的输出, h_j 表示隐含层第 j 个节点的输出. 整体网络的参数 $\theta = \{W, a, b\}$, W 是权值参数 a 和 b 分别表示可视层和隐含层的偏置变量. 给定可视层的前提下, 隐含层的概率为:

$$P(\boldsymbol{h} \mid \boldsymbol{v}) = \prod_j p(h_j \mid \boldsymbol{v}) p(h_j = 1 \mid \boldsymbol{v}) = \frac{1}{1 + \exp(\sum_i W_{ij} v_i - a_j)} \tag{5}$$

这样我们就建立了可视层与隐含层之间的概率表示. 同样的, 隐含层之间的概率可表示为:

$$P(\boldsymbol{v}, h_1, h_2, h_3) = P(\boldsymbol{v} \mid h_1) P(h_1 \mid h_2) P(h_2 \mid h_3) \tag{6}$$

对于 RBM 的学习算法我们采用梯度衰减法. 可视层表达的是输入数据的特征, 所以学习算法的目标函数是将可视层的概率最大化. 所以有如下最大似然的概率表示:

$$\frac{\partial \log_{\mathrm{e}} P(\boldsymbol{v})}{\partial \theta} = \frac{\partial \log_{\mathrm{e}} \sum_{\boldsymbol{h}} P(\boldsymbol{v},\boldsymbol{h})}{\partial \theta} = \frac{\sum_{\boldsymbol{h}} \mathrm{e}^{-energy(\boldsymbol{v},\boldsymbol{h}) \frac{\partial(-energy(\boldsymbol{v},\boldsymbol{h}))}{\partial \theta}}}{\sum_{\boldsymbol{h}} -energy(\boldsymbol{v},\boldsymbol{h}))} - \frac{\sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \mathrm{e}^{(-energy(\boldsymbol{v},\boldsymbol{h}) \frac{\partial energy(\boldsymbol{v},\boldsymbol{h})}{\partial \theta})}}{\sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} (-energy(\boldsymbol{v},\boldsymbol{h}))} \tag{7}$$

对于标准化的高斯 RBM,

$$energy(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{2} \boldsymbol{v}^{\mathrm{T}} \boldsymbol{v} - a^{\mathrm{T}} \boldsymbol{v} - b^{\mathrm{T}} \boldsymbol{h} - \boldsymbol{h}^{\mathrm{T}} W \boldsymbol{v}$$

得到:

$$\frac{\partial P(\boldsymbol{v})}{\partial \theta} = \sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid \boldsymbol{v}) \frac{\partial(-energy(\boldsymbol{v},\boldsymbol{h}))}{\partial \theta} - \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} P(\boldsymbol{v},\boldsymbol{h}) \frac{\partial(-energy(\boldsymbol{v},\boldsymbol{h}))}{\partial \theta} \tag{8}$$

由以上的表达式中, 我们可以将前一部分定义为激励部分, 表示为 v 节点的期望值表示; 后一部分作为抑制部分, 表示在联合概率下的期望表示.

1.3 影响因素诊断

在完成上述数据降维与非线性特征提取之后, 对影响因素进行诊断. 其中, 本文采用偏导数表示该变量对 T_N 变化的贡献率大小, 因此对于某一个数据 X_0, 它对的贡献率[21] K 为:

$$K = \frac{\partial T_N}{\partial X} \mid_{X = X_0} \tag{9}$$

1.4　改进的深度置信网络建模流程

基于非线性特征提取的深度置信网络的建立步骤为:

1) 通过机理确定变量.

2) 对输入数据进行移除异常值以及零均值归一化.

3) 设计 IT 网络以及深度置信网的网络结构.

4) 选择数据训练 IT 网络, 得到非线性 PCA 降维模型, 并计算各变量的统计量, 作为影响因素诊断依据.

5) 将降维数据输入深度置信网训练网络.

6) 用检测数据对整个模型进行检验. 如果效果不满意, 则返回 3).

整体分类器模型结构如图 3 所示.

2　深度置信网络复杂度分析

为深刻揭示本文所提改进型 DBN 在网络结构和算法复杂度方面的优势, 本节从如下两个方面进行分析:

1) 网络结构复杂度

信息熵的概念是 1958 年香农借鉴热力学上分子混乱程度来描述信息源含信息量的不确定度. 从信息学的角度出发, 可以论证所提方法在优化网络结构上的优势, 采用隐含层的信息熵来体现网络的结构性和组织性[22]. 武妍等在论述提高网络泛化能力优化网络结构中提出通过正则化 (惩罚函数) 的方

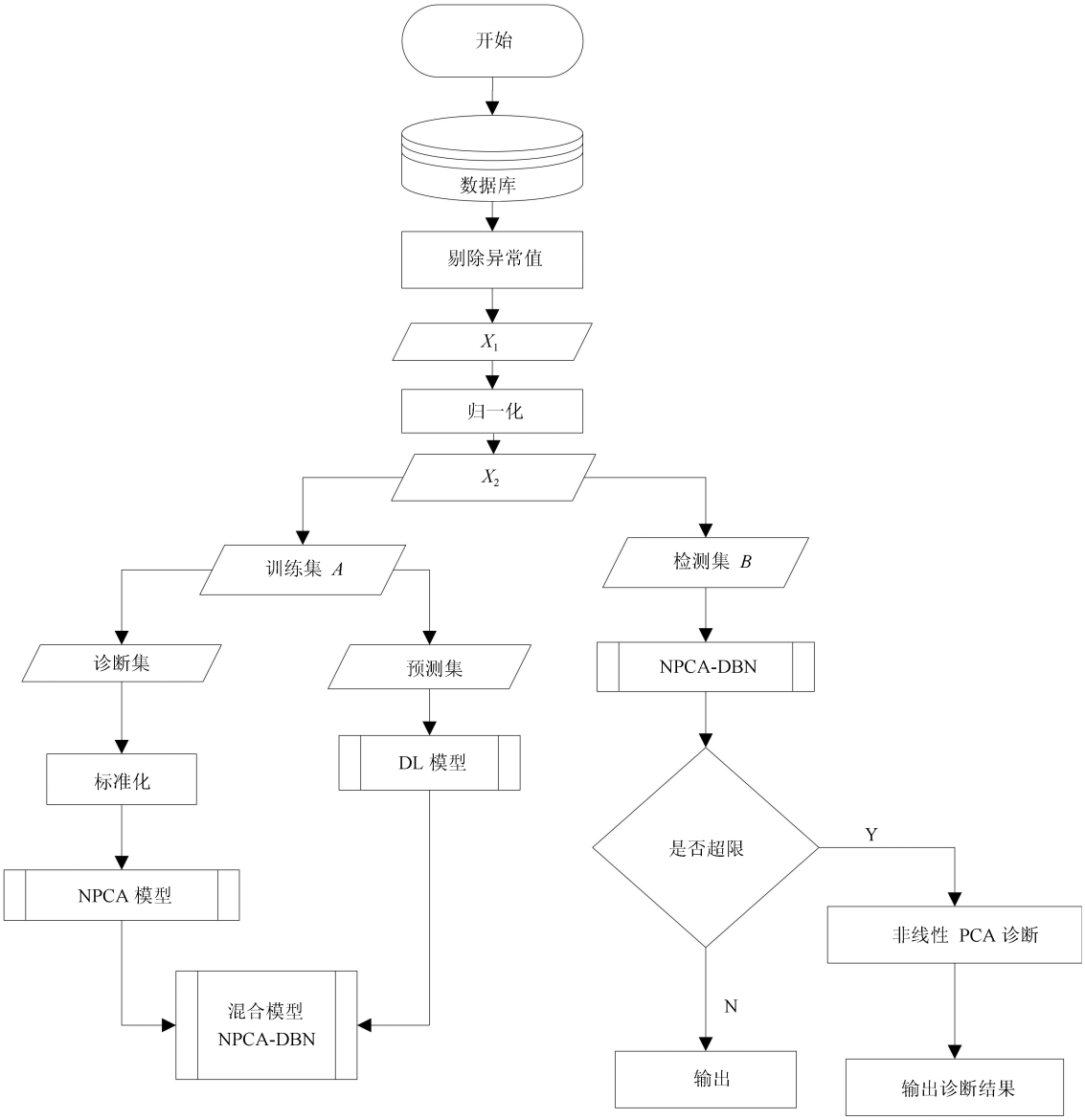

图 3　NPCA-DBN 模型分类与诊断结构图

Fig. 3　The classification and diagnosis model with NPCA-DBN

法, 来控制网络的 “有效复杂度”[23]. Deco 等通过构建基于互信息熵的正则函数, 来等效网络的 “有效复杂度”, 并进行网络结构优化. 其中输入层和隐含层之间的互信息熵[24] 定义为:

$$H = -\sum_{j=1}^{Q} c_j \log c_j + \frac{l}{P}\sum_{l=1}^{Q}\sum_{j=1}^{Q} c_{jl}\log c_{jl} \tag{10}$$

其中,P 为输入样本数, Q 为隐含层节点数, c_{jl} 为第 l 个样本对第 j 个隐含单元的归一化输出, c_j 为平均值. 熵的单位取决于定义用到对数的底, 当底数为 2, 熵的单位是 bit; 当底数为 e, 熵的单位是 nat; 而当底数为 10, 熵的单位是 Hart.

定理 1. 面向具有相同特征的样本数据设计的两个训练深度网络 net1 和 net2, 若网络 “有效复杂度” 相同 ($H_{\text{net1}} = H_{\text{net2}}$), 当网络的输入层节点 $P_{\text{net1}} < P_{\text{net2}}$ 时, 则有, 网络的隐含层节点总和 $Q_{\text{net1}} < Q_{\text{net2}}$.

证明. 假设原 DBN 网络 (net1) 的互信息熵函数已是最小化, 其中第一层 RBM 完全反映了输入层和隐含层的互信息. 根据信息熵原理, 则有[25]:

$$-\sum_{j=1}^{Q} c_j \log c_j = -\sum_{j=1}^{Q_l} \frac{l}{Q_l}\log\frac{l}{Q_l} = \log Q_l \tag{11}$$

Q_l 代表隐含层第一层的节点数. 将式 (11) 代入式 (10) 中可得:

$$H = \log Q_l + \frac{l}{P}\sum_{l=1}^{P}\sum_{j}^{Q_l} c_{jl}\log c_{jl} \tag{12}$$

基于 DBN 原理, 本文提出的改进型 DBN 网络 (net2) 应使每一个 RBM 都能完全重构输入变量, 因此, 也应使所有互信息熵最小化, 则有改进方法后的互信息熵为 H':

$$H' = \log Q'_l + \frac{l}{P'}\sum_{l=1}^{P'}\sum_{j'}^{Q'_l} c_{j'l}\log c_{j'l} \tag{13}$$

又因为, 如完全重构原始输入变量, (由于假设 NPCA 完全提取了原来样本数据中的特征信息, 因此, net2 中第一层 RBM 依然为求解隐含层节点到原始样本信息的映射关系), 则必有:

$$\sum_{j'}^{Q'_l} c_{j'l}\log c_{j'l} = \sum_{j}^{Q_l} c_{jl}\log c_{jl} \tag{14}$$

此外, 因为同样满足互信息熵最小化, (对于同一样本数据, 我们采用同种 DBN 网络结构进行信息映射时, “有效复杂度” 应该相等. 也就是正则函数相等), 即 $H = H'$, 因此当 $P' \le P$ 时, 则必有 $Q'_l \le Q_l$. 同理, 后续隐含层之间的 RBM 节点个数同样具有此规律. 因此可得, $Q_{\text{net2}} = Q'_1 + Q'_2 + \cdots + Q'_n \le Q_{\text{net1}} = Q_1 + Q_2 + \cdots + Q_n$ (n 为网络的隐含层总层数). 综上可以得到改进后的网络总节点存在 $S_{\text{net1}} < S_{\text{net2}}$.

□

2) 算法复杂度分析

算法的复杂度就是对算法计算所需要的时间和空间的一种度量[25]. 一般将算法的复杂度分为时间复杂度和空间复杂度. 时间复杂度是以算法结构主体执行循环次数为依据, 空间复杂度以程序主体占据空间为依据[26]. 一个算法中的语句执行次数称为语句频度或时间频度, 记为 $T(n)$, 若有某个辅助函数 $f(n)$, 使得当 n 趋近于无穷大时, $T(n)/f(n)$ 的极限值为不等于零的常数, 则称 $f(n)$ 是 $T(n)$ 的同数量级函数, 记作 $T(n) = \mathrm{O}(f(n))$, 称 $\mathrm{O}(f(n))$ 为算法的渐进时间复杂度, 简称时间复杂度. 用 O 代表一个算法的计算复杂度, 算法中的循环语句是算法的主体, 若算法中含有并列的算法, 则将并列的算法复杂度相加. 例如:

for $i = 1 : n$
  $x = x + 1$;
end
for $i = 1 : n$
  for $j = 1 : n$
    $x = x + 1$;
  end
end

第一个 for 循环的复杂度为 $\mathrm{O}(n)$, 第二个循环的复杂度为 $\mathrm{O}(n2)$, 则整个算法的复杂度为 $\mathrm{O}(n + n2) = \mathrm{O}(n2)$.

定理 2. 假设存在一个 DBN 网络, 其结构为含有 n 层隐含层, 隐含层节点数为 $[h_1, h_2, \cdots, h_n]$. 则存在一类基于 ITNN 神经网络的非线性特征提取机制的 DBN (假设该机制可以有效保证数据的互信息熵不变), 当 DBN 主结构的节点数可以减小到 $[h'_1, h'_2, \cdots, h'_n]$, 预处理 IT 网络部分隐含层节点数为 h_0, 且满足 ITNN 的隐含层节点数 $h_0 < \beta$, 则有: 改进后 DBN 的时间复杂度降低小于原 DBN. 其中,

$$\beta = \prod_{1}^{n} h_n - \prod_{1}^{n} h'_n \tag{15}$$

证明.

1) 传统 DBN 算法的伪代码如下:

for (1 : DBN 的隐含层第一层节点数 h_1)
  for (1 : DBN 的隐含层第二层节点数 h_2)
    $\cdots$
      for (1 : DBN 的隐含层第 n 层节点数

h_n)

计算网络的目标函数是否符合要求

end

· · ·

end

end

所以 DBN 的复杂度为 $\mathrm{O}(\mathrm{DBN}) = \prod_1^n h_n$

2)　NPCA-DBN 算法的伪代码如下:

for (1 : NPCA 的隐含层节点数 h_0)

计算输入训练网络的目标函数是否符合要求

end

for (1 : DBN 的隐含层第一层节点数 h_1')

for (1 : DBN 的隐含层第二层节点数 h_2')

· · ·

for (1 : DBN 的隐含层第 n 层节点数 h_n')

计算网络的目标函数是否符合要求

end

· · ·

end

end

故所提算法的复杂度为 O(NPCA-DBN)= $h0 + \prod_1^n h_n'$ 若 $h0 < \prod_1^n h_n - \prod_1^n h_n'$, 则有: O(NPCA-DBN)-O(DBN) < 0, 即改进型 DBN 的时间复杂度降低. 此外, 由于改进型 DBN 降低了原 DBN 结构中的隐含层节点数, 则有效降低算法的空间复杂度. □

3　实例研究

PM2.5 预测和影响因素诊断涉及的变量众多, 而且影响变量之间多存在关联, 本节给出了 PM2.5 浓度预测与超标影响因素诊断方法并进行了数值验证.

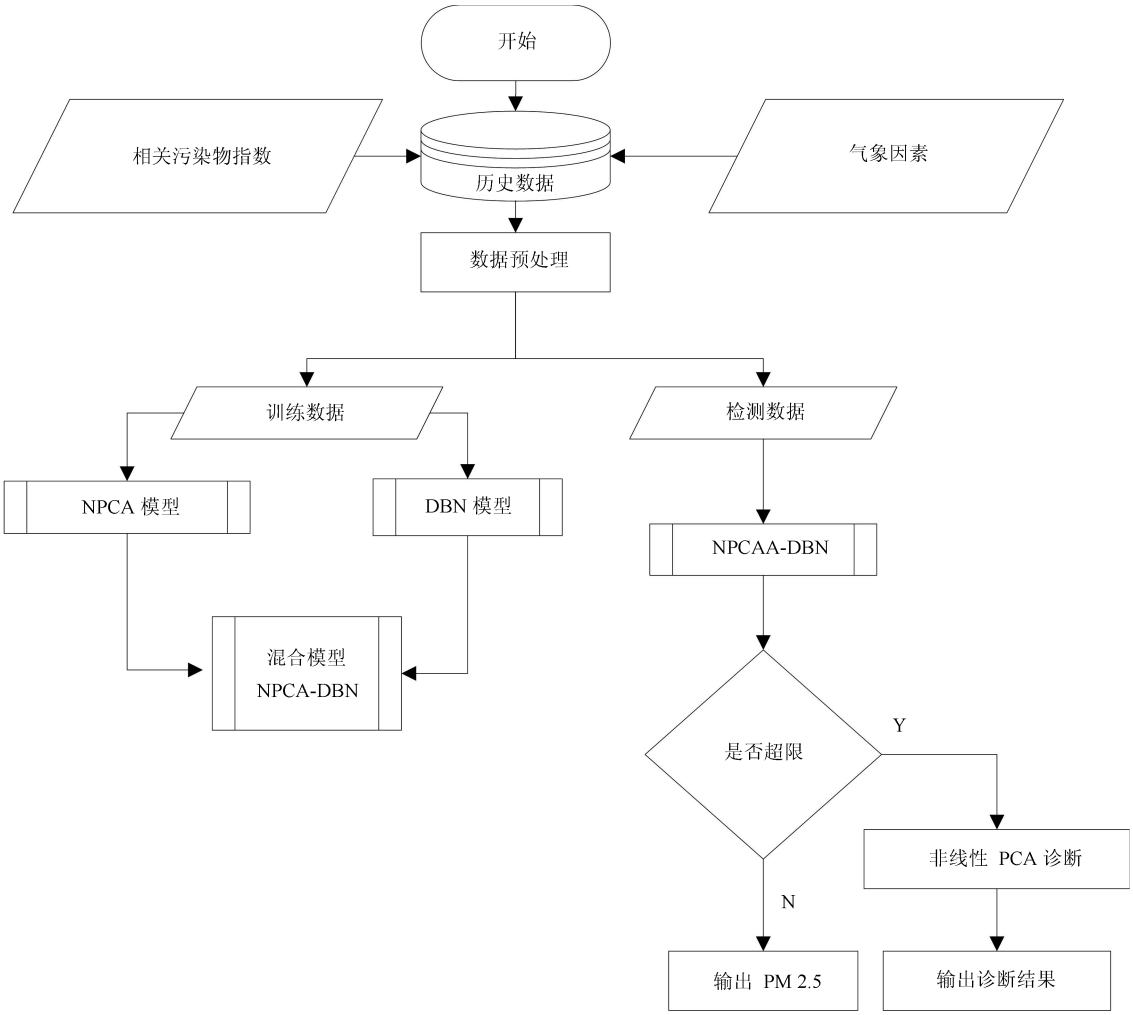

图 4　PM2.5 预测诊断流程图

Fig. 4　The flow chart of PM2.5 concentration′s prediction and diagnosis

3.1 PM2.5 的预测和影响因素诊断方法

基于第二节所提混合分类器模型, 选用相关污染物和气象因素作为输入, 提取主元非线性特征之后, 输入深度置信网, 来进行预测, 并根据统计量信息诊断 PM2.5 浓度超标原因. 算法建立的步骤如下:

1) 选择历史数据, 并建立非线性 PCA 和深度置信网的模型.

2) 训练模型.

3) 检测模型效果.

4) 得出预测及诊断结果.

如图 4 所示.

3.2 数据来源

为验证本文混合模型的有效性, 采取河北省某市地表水厂, 华电二区和胶片厂三个检测点于 2014 年 11 月至 2015 年 4 月间的监测数据作为实验数据. 其中, 为分析检测数据, 依据文献 [27] 选取相关污染物如: PM10, SO_2, NO, CO, O_3, 气象数据如: 风速, 风向, 温度, 湿度, 相关空气指数数: 空气指数 AQI. 实验采用 500 个训练样本, 100 个检测样本, 模型训练次数设置为 50 000 次.

3.3 PM2.5 浓度预测

1) 网络结构

基于本文所提出的改进 DBN 模型, 利用历史数据, 进行 PM2.5 的浓度预测, 本文采用实验的方式获得模型的结构, 并与传统的预测 DBN 模型进行对比. 在参考其他文献以及经验规则的基础上, 通过实验获得改进 DBN 的网络结构, 如图 5 所示.

图 5 中: xx-xx-xx 为隐含层的结构, 代表 DBN 三层主结构中的隐含层和内部节点分配. 可见试验后得到 DBN 主结构隐含层的节点数结构为 10-6-6 为本次使用的网络结构, 其中数据预处理阶段采用的浅层学习网络采用试验方法得到有一层隐含层节点, 非线性节点数为 10. 对比传统 DBN 网络结构, 两者间的对比关系如表 1 所示.

其中 (6-10-10) + (6-10-6-6-1) 代表网络整体结构, 对于预处理阶段的浅层网络有 6-10-10 的网络结构, DBN 主结构的输入层为 6 个节点, 隐含层为三层, 第一层是 10 个节点, 第二层和第三层为 6 个节点, 一个输出的结构, 由于改进的 DBN 的两部分的节点不在同一个网络嵌套中, 故为两个部分的复杂度相加. 由上表我们可以清楚地看出改进的 DBN 模型在主结构中的深层网络中, 大大减少了非线性节点的个数, 从而在算法复杂度上实现数量级上的减小.

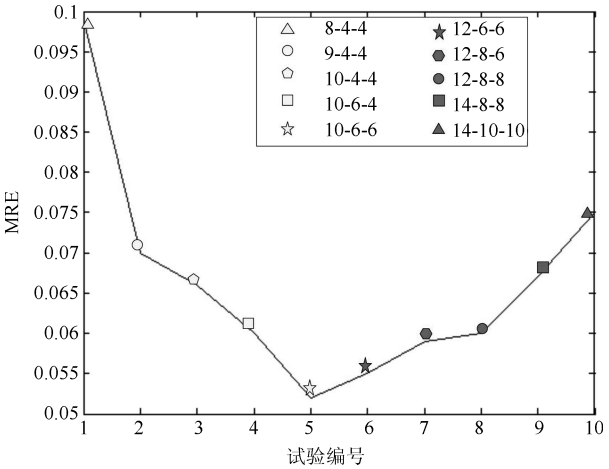

图 5 不同结构预测的平均相对误差

Fig. 5 The classification and diagnosis model with NPCA-DBN

2) 建模精度对比实验

预测阶段采用检测输出的平均相对误差 MRE (Mean relative error) 来表示预测的精度.

$$\mathrm{MRE}=\frac{\sum_{j=1}^{m}\left(\frac{|X_{obs,j}-X_{exp,j}|}{X_{exp,j}}\right)}{m} \tag{16}$$

其中, m 是检测数据的样本数. $X_{obs,j}$ 表示检测数据的输出值, $X_{exp,j}$ 表示检测数据的真值. 平均相对误差反映出了在预测上偏离真值的平均水平. 为清晰展现本文所提 DBN 的优势, 以华电二区监测点为例, 图 6 和图 7 分别给出了改进 DBN 与传统的 DBN、SVM 和 PLS 在预测效果上的对比结果和建模误差趋势.

在图 6 中, 横坐标为监测数据的 100 个采样点, 纵坐标为 PM2.5 的浓度. 其中 * 代表模型输出的预测值, o 代表实际值. 我们可以直观地看出, 改进 DBN 的模型预测效果更佳, 同时比传统的分类方法在精度上有所提升.

图 7 中横坐标代表 100 个监测时间点, 纵坐标代表各个预测值的相对误差. 由图 6 和图 7 可以清晰地展现出, 改进 DBN 模型的预测精度并没有因为降低输入的维度而降低. 通过对比分析, 我们可以得到: 首先, 传统 DBN 没有经过降维预处理, 预测精度不高; 另外, 由于 DBN 在处理海量数据建模时

表 1 网络结构对比

Table 1 The comparison of the network structure

模型	结构	隐含层节点数	总节点数	算法总空间复杂度
NPCA-DBN	(6-10-10) + (6-10-6-6-1)	32	55	$6\times10\times10+6\times10\times6\times6\times1$
DBN	10-12-10-10-1	32	43	$10\times12\times10\times10$

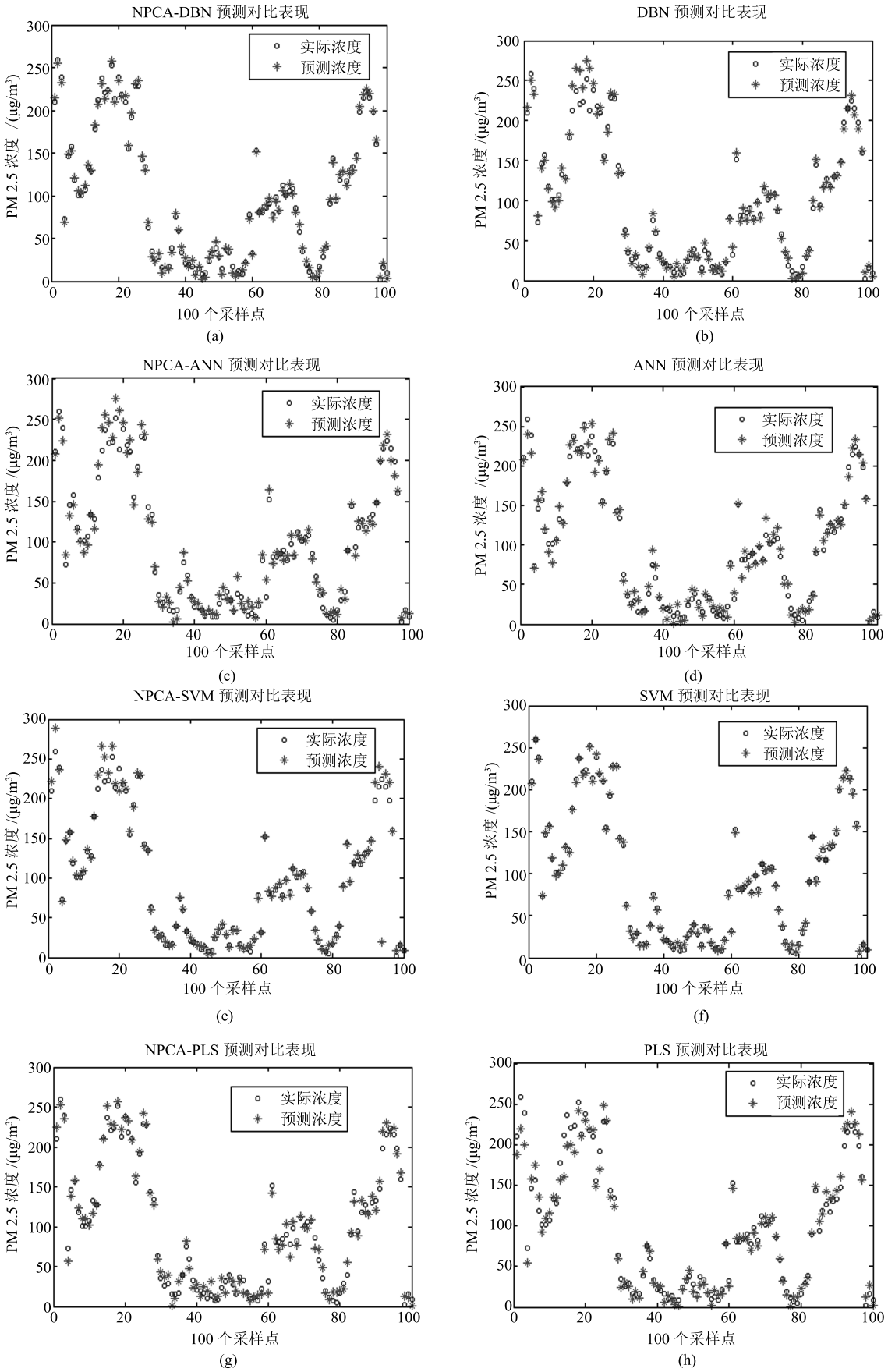

图 6　华电二区的预测效果对比图

Fig. 6　The comparison in the second area of Huadian with different structures

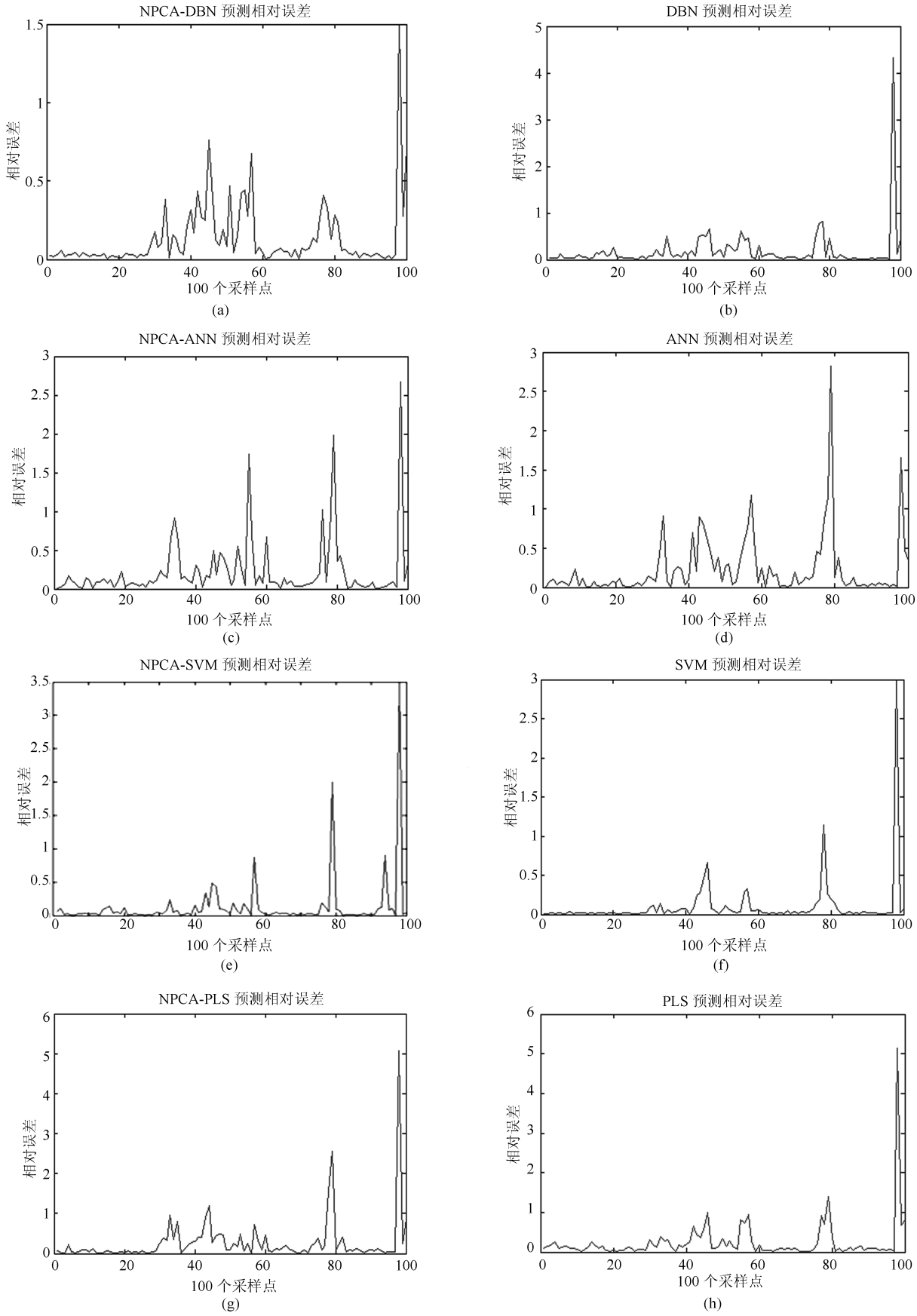

图 7 不同结构预测的平均相对误差

Fig. 7 The MRE of different structures

表 2　建模精度与收敛速度对比

Table 2　The comparison of the network structure

监测点	指标	NPCA-DBN	NPCA-ANN	NPCA-SVM	NPCA-PLS	DBN	ANN	SVM	PLS
地表	MRE ($\times 10^{-2}$)	13.32	22.21	13.14	26.82	17.92	23.40	12.19	24.54
水厂	训练时间 (s)	44	16	180	46	89	33	349	94
华电	MRE ($\times 10^{-2}$)	14.57	25.15	13.04	29.48	17.01	24.16	10.22	27.16
二区	训练时间 (s)	37	12	211	49	90	38	401	103
胶片	MRE ($\times 10^{-2}$)	10.51	26.49	11.09	33.16	12.77	23.32	12.73	30.06
厂	训练时间 (s)	42	16	198	57	108	42	399	108

有显著的优势, 而本实验训练样本为 500 个, 因此, 所提改进的 DBN 在预测精度上与 SVM 提升不明显; 其次, 与 ANN(BP) 方法对比, 是因为 BP 只有一个隐含层, 属于浅层学习, 训练网络深度不足; 最后, PLS 适用于处理线性模型的预测问题, 对非线性关系的建模精度欠佳. 此外, 基于多种非线性特征提取机制下的复合分类预测方法, 表 2 给出了该市地表水厂, 华电二区和胶片厂周边 PM2.5 浓度的建模精度和收敛速度对比.

由表 2 我们可以得到, 本文所提改进 DBN 在建模精度和收敛速度上都有较大提升, NPCA 数据预处理算法通过提取数据之间的非线性特征, 得到原有数据的非线性表达, 对于网络化表达的机器学习算法可以提高建模精度, 并降低训练时间. 特别的, 同样采用 NPCA 数据预处理机制的复合分类方法, 对于基于线性化拟合的浅层学习算法随着训练时间的降低, 建模精度有所下降.

3.4　PM2.5 浓度影响因素诊断

基于所提出改进 DBN 的影响因素诊断方法, 在实验中, 我们将空气质量指数, PM10 浓度, SO_2、CO、NO_2、O_3 气体浓度, 风向, 风速, 相对湿度, 温度等 10 个过程变量作为诊断部分的输入变量. 由于该地区的特殊性, 在 4 000 多次的采样数据中, 有二分之一采样点数据的 PM2.5 浓度都高于 100 $\mu g/m^3$, 所以, 为了展现模型对 PM2.5 影响因素诊断结果, 我们实验设置的 PM2.5 预测限为 200$\mu g/m^3$, 即处于重度污染的情况下, 计算输入变量中对于结果的贡献率[28]. 并用贡献图的方式表达影响因素诊断结果. 我国对 PM2.5 浓度级别划分如表 3 所示[29].

在历史数据中选定所有未超限数据对应的输入, 求平均水平代表未超限数据组合作为参考输入变量集: reference (152, 168, 63, 1.55, 50, 42, 163, 2.26, 0.63, 3.88) 针对华电二区监测区域, 在图 8 中, 我们以超限组第 20 组贡献图为例说明诊断过程.

由贡献图可以看出, 第七个变量对结果的贡献最大. 我们观察验证第 83 组数据输入变量集为: X_{20} = (179, 207, 79, 1.78, 73, 177, 266, 3.07, 0.44, 7.34) 第 6 个变量的相对偏差为最大, 因此诊断结果为: 造成此次污染物浓度过高的首要原因是风速的原因.

表 3　PM2.5 浓度级别

Table 3　The PM2.5 concentration level

浓度范围 ($\mu g/m^3$)	级别	优良级别
0 ∼ 50	1 级	优
50 ∼ 100	2 级	良
101 ∼ 150	3 级	轻度污染
151 ∼ 200	4 级	中度污染
201 ∼	5 级	重度污染

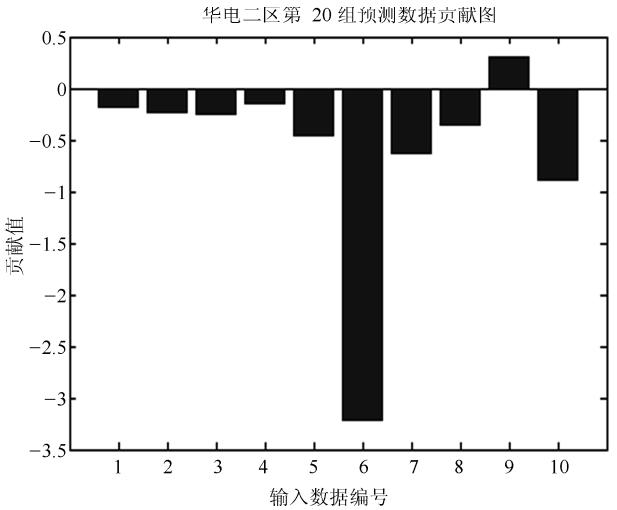

图 8　华电二区超限数据贡献图

Fig. 8　The contribution chart of the overrun data in the second area of Huadian

3.5　小结

从上述实验可以看出, 本文提出的改进的 DBN 模型在预测效果上并没有使精度降低, 同时, 加快了模型的收敛速度. 并且在超标诊断中, 平均超标检测率达到 85 %, 能够有效地诊断出 PM2.5 浓度超标的主要因素.

4 结论

本文提出的基于非线性特征提取的 DBN 模型能够有效完成含复杂非线性特征关系高维数据的预测建模诊断任务. 基于信息熵理论, 证明了本文所提 DBN 模型相比传统 DBN, 能够在不降低建模精度的同时, 达到降低网络和算法复杂度的优势, 对于深度学习理论在海量数据挖掘中的应用具有重要理论意义. 将所提建模方法应用到一类 PM2.5 浓度预测与诊断问题中, 并与传统 DBN、SVM、ANN、PLS 等分类方法和含 NPCA 数据预处理机制的复合分类方法做了详细对比, 验证了所提方法的优势与正确性. 需要说明的是, 本文采取基于数据驱动的方法对 PM2.5 进行浓度预测和影响因素诊断, 在 PM2.5 的形成机理上还未做过多的分析, 在未来的研究中将深入探讨 PM2.5 浓度变化机理. 此外, 由于本文数据来源于特定城市的采样点, 因此在方法的适用性方面还要做深入的研究. 下一步的工作将分为以下两部分进行: 1) 理论方面, 面向深度置信网络结构本身的优化方法的研究, 研究自适应样本数据特征的网络模型结构. 2) 应用方面, 尝试将所提方法应用到复杂流程工业的建模和诊断问题中.

References

1 Saki F, Kehtarnavaz N. Online frame-based clustering with unknown number of clusters. *Pattern Recognition*, 2016, **57**: 70−83

2 Li H, Chung F L, Wang S T. A SVM based classification method for homogeneous data. *Applied Soft Computing*, 2015, **36**: 228−235

3 Embrechts M J, Rossi F, Schleif F M, Lee J A. Advances in artificial neural networks, machine learning, and computational intelligence (ESANN 2013). *Neurocomputing*, 2014, **141**: 1−2

4 Zhang Y P, Li X, Zhang Z F, Wu F, Zhao L M. Deep learning driven blockwise moving object detection with binary scene modeling. *Neurocomputing*, 2015, **168**: 454−463

5 Shang C, Yang F, Huang D X, Lyu W X. Data-driven soft sensor development based on deep learning technique. *Journal of Process Control*, 2014, **24**(3): 223−233

6 Gao Ying-Ying, Zhu Wei-Bin. Deep neural networks with visible intermediate layers. *Acta Automatica Sinica*, 2015, **41**(9): 1627−1637
(高莹莹, 朱维彬. 深层神经网络中间层可见化建模. 自动化学报, 2015, **41**(9): 1627−1637)

7 Ding Ke, Tan Ying. A review on general purpose computing on GPUs and its applications in computational intelligence. *CAAI Transactions on Intelligent Systems*, 2015, **10**(1): 1−11
(丁科, 谭营. GPU 通用计算及其在计算智能领域的应用. 智能系统学报, 2015, **10**(1): 1−11)

8 Shen F R, Chao J, Zhao J X. Forecasting exchange rate using deep belief networks and conjugate gradient method. *Neurocomputing*, 2015, **167**: 243−253

9 Huang S, Yang D, Ge Y X, Zhang X H. Combined supervised information with PCA via discriminative component selection. *Information Processing Letters*, 2015, **115**(11): 812−816

10 Gan Jun-Ying, Li Chun-Zhi. Face Recognition based on wavelet transform, two-dimensional principal component analysis and independent component analysis. *Pattern Recognition and Artificial Intelligence*, 2007, **20**(3): 377−381
(甘俊英, 李春芝. 基于小波变换、二维主元分析与独立元分析的人脸识别方法. 模式识别与人工智能, 2007, **20**(3): 377−381)

11 Tang Jian, Chai Tian-You, Yu Wen, Zhao Li-Jie. On-line KPLS algorithm with application to ensemble modeling parameters of mill load. *Acta Automatica Sinica*, 2013, **39**(5): 471−486
(汤健, 柴天佑, 余文, 赵立杰. 在线 KPLS 建模方法及在磨机负荷参数集成建模中的应用. 自动化学报, 2013, **39**(5): 471−486)

12 Cobourn W G. An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. *Atmospheric Environment*, 2010, **44**(25): 3015−3023

13 Voukantsis D, Karatzas K, Kukkonen J, Räsänen T, Karppinen A, Kolehmainen M. Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Science of the Total Environment*, 2011, **409**(7): 1266−1276

14 Xia D H, Jiang B F, Xie Y L. Modeling and analysis of PM2.5 generation for key factors identification in China. *Atmospheric Environment*, 2016, **134**: 208−216

15 De Ridder K, Kumar U, Lauwaet D, Blyth L, Lefebvre W. Kalman filter-based air quality forecast adjustment. *Atmospheric Environment*, 2012, **50**: 381−384

16 de Gennaro G, Trizio L, Gilio A D, Pey J, Pérez N, Cusack M, Alastuey A, Querol X. Neural network model for the prediction of PM10 daily concentrations in two sites in the Western Mediterranean. *Science of The Total Environment*, 2013, **463−464**: 875−883

17 Feng X, Li Q, Zhu Y J, Hou J X, Jin L Y, Wang J J. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment*, 2015, **107**: 118−128

18 Tan S F, Mayrovouniotis M L. Reducing data dimensionality through optimizing neural network inputs. *AIChE Journal*, 1995, **41**(6): 1471−1480

19 Hinton G E, Osindero S, The Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

20 Qiao Jun-Fei, Pan Guang-Yuan, Han Hong-Gui. Design and application of continuous deep belief network. *Acta Automatica Sinica*, 2015, **41**(12): 2138−2146
(乔俊飞, 潘广源, 韩红桂. 一种连续型深度信念网的设计与应用. 自动化学报, 2015, **41**(12): 2138−2146)

21 Li Er-Guo, Yu Jin-Shou. An input-training neural network based nonlinear principal component analysis approach for fault diagnosis. *Control and Decision*, 2003, **18**(2): 229−232
(李尔国, 俞金寿. 一种基于输入训练神经网络的非线性 PCA 故障诊断方法. 控制与决策, 2003, **18**(2): 229−232)

22 He R, Hu B G, Yuan X T, Zheng W S. Principal component analysis based on non-parametric maximum entropy. *Neurocomputing*, 2010, **73**(10−12): 1840−1852

23 Wu Yan, Zhang Li-Ming. A survey of research work on neural network generalization and structure optimization algorithms. *Application Research of Computers*, 2002, **19**(6): 21−25, 84
(武妍, 张立明. 神经网络的泛化能力与结构优化算法研究. 计算机应用研究, 2002, **19**(6): 21−25, 84)

24 Deco G, Finnoff W, Zimmermann H G. Unsupervised mutual information criterion for elimination of overtraining in supervised multilayer networks. *Neural Computation*, 1995, **7**(1): 86−107

25 Wu Xin-Gen, Lv Wei-Xue. A neural network rule expression based on information entropy. *Computer Engineering*, 1996, **22**(5): 46−51
(吴新根, 吕维雪. 一种基于信息熵的神经网络规则表示. 计算机工程, 1996, **22**(5): 46−51)

26 Sánchez D, Melin P, Castillo O. Optimization of modular granular neural networks using a hierarchical genetic algorithm based on the database complexity applied to human recognition. *Information Sciences*, 2015, **309**: 73−101

27 Wang L T, Wei Z, Yang J, Zhang Y, Zhang F F, Su J, et al. The 2013 severe haze over southern Hebei, China: model evaluation, source apportionment, and policy implications. *Atmospheric Chemistry and Physics Discussions*, 2013, **13**(11): 28395−28451

28 Peng K X, Zhang K, Li G. Online contribution rate based fault diagnosis for nonlinear industrial processes. *Acta Automatica Sinica*, 2014, **40**(3): 423−430

29 Yao L, Yang L X, Yuan Q, Yan C, Dong C, Meng C P, et al. Sources apportionment of PM2.5 in a background site in the North China Plain. *Science of The Total Environment*, 2016, **541**: 590−598

高 月 北京化工大学信息学院硕士研究生. 主要研究方向为智能决策.
E-mail: 18810255106@163.com
(**GAO Yue** Master student in Beijing University of Chemical Technology. Her main research interest is intelligent decision making.)

宿 翀 北京化工大学信息学院副教授. 主要研究方向为人工智能, 情感计算和智能医疗. 本文通信作者.
E-mail: suchong@mail.buct.edu.cn
(**SU Chong** Associate professor in Beijing University of Chemical Technology. His research interest covers intelligent applications, affect computing and smart medicine. Corresponding author of this paper.)

李宏光 北京化工大学信息学院教授. 主要研究方向为化工过程的建模、控制和优化.
E-mail: lihg@mail.buct.edu.cn
(**LI Hong-Guang** Professor in Beijing University of Chemical Technology. His research interest covers modeling, control and optimization of chemical process as well as computer based intelligent control for industrial plants.)