文章编号: 1001-0920(2018)12-2153-08 DOI: 10.13195/j.kzyjc.2017.0937

面向原油总氢物性预测的数据扩增预处理方法

易 令, 吕忠元, 丁进良†, 刘长鑫

(东北大学 流程工业综合自动化国家重点实验室，沈阳 110004)

摘 要: 针对原油总氢物性回归预测中核磁共振光谱数据不足的问题,结合深度学习相关理论,提出一种光谱数据扩增预处理方法. 根据样本输入和标签的相关系数,在原始样本中加入随机噪声以生成虚拟样本;处理样本数据结构以利于卷积神经网络特征提取,并加入数据冗余改进该结构以进一步提高数据特征提取的完整性;搭建实现原油总氢物性回归预测的卷积神经网络(Regression forecasting convolutional neural network, RF-CNN). 实验结果表明,对于总氢物性的回归预测,该数据扩增预处理方法不但可以解决原始数据训练中的过拟合现象,而且相比于传统的偏最小二乘(PLS)回归方法,更具稳定性和精确性.

关键词: 卷积神经网络；核磁共振光谱； 原油物性； 回归预测； 虚拟样本

中图分类号: TP183 **文献标志码:** A

Data pretreatment approach for crude oil hydrogen properties prediction

YI Ling, LYU Zhong-yuan, DING Jing-liang†, LIU Chang-xin

(State Key Laboratory of Synthetical Automation for Process Industries，Northeastern University，Shenyang 110004，China)

Abstract: Aiming at the problem of lack of nuclear magnetic resonance spectroscopy data in the prediction of total hydrogen activity of crude oil, combined with the theory of deep learning, a pre-processing method of spectral data amplification is proposed. According to the correlation coefficient of the sample input and the label, the random noise is added to the original sample to generate the virtual sample. The sample data structure is processed to facilitate the feature extraction of the convolutional neural network, and the data redundancy is added to improve the structure to further improve the integrity of the data feature extraction. A regression forecasting convolutional neural network(RF-CNN) is designed to realize the regression prediction of the total hydrogen content of crude oil. Experiments show that, for the regression prediction of total hydrogen properties, the amplified data not only solves the over-fitting phenomenon in the original data training, but also has more stability and accuracy than the traditional partial least squares(PLS) dimensionality reduction method.

Keywords: convolutional neural network； nuclear magnetic resonance spectroscopy； crude oil physical properties； regression prediction； virtual samples

0 引 言

近年来,国内炼油行业逐步打破垄断,地级炼油厂开始获得了原油进口权. 由于不同产地的原油性质不同,炼油厂储罐数量有限,不可能实现对每个产地原油的单独存储,必然出现原油混合现象,从而改变了原油性质. 为了保证混合原油的稳定性,要求在快速预处理过程中准确地获得进料物性数据.

早期企业多采用传感器检测的方法直接对原油物性进行分析,其对于少量特定指标的检测具有一定的快速性和准确性. 然而,生产规模的不断扩大,所需检测的指标逐渐增多,检测设备的成本不再满足市场需求. 随着光谱技术和光谱仪器的迅速发展,采用光谱技术间接地对原油物性进行检测受到企业的青睐,如拉曼光谱技术[1]、红外光谱技术[2]以及核磁共振光谱技术. 其中,核磁共振物质成分分析由于精准、快速的特点,已被越来越多的企业所应用.

然而,获得实际准确的物性含量需要实验室化验较长时间. 而核磁光谱数据与各个物性含量之间有

收稿日期: 2017-07-13；**修回日期:** 2017-09-21.

基金项目: 国家自然科学基金项目(61590922, 61525302)；教育部基本科研业务费项目(N160801001, N161608001).

责任编委: 侯忠生.

作者简介: 易令(1992−), 男, 博士, 从事深度学习及其应用的研究；丁进良(1976−), 男, 教授, 博士, 从事复杂工业过程智能建模与优化控制、生产全流程运行优化、智能优化算法及工业应用等研究.

†通讯作者. E-mail: jlding@mail.neu.edu.cn

着一定的内在关联,因此,本文拟采用深度学习中的卷积神经网络学习核磁光谱数据与总氢物性含量的关系,从而实现该物性快速准确的预测.

卷积神经网络(Convolutional neural network, CNN)由纽约大学Le等[3]提出,与传统神经网络相比,其具有并行处理能力、良好的容错能力和自学习能力等优点,能够应对环境信息复杂、推理规则不明确等问题. 卷积神经网络在图像处理方面已获得了巨大的成功,如人脸识别[4-6]、图像提取语义信息[7]、人体行为识别[8]和阴影检测[9]等. 在工业物性预测领域,神经网络也得到了应用. 如:乔俊飞等[10]提出了基于RBF神经网络的出水氨氮软测量模型,对出水氨氮进行预测,解决了污水处理中该水质参数难以实时检测和检测精度低的问题;王旭东等[11]提出了一种基于神经网络的通用软测量模型;张昭昭等[12]针对矿井中瓦斯浓度非线性的问题,提出一种动态神经网络,实现了对瓦斯浓度的实时预测. 然而,卷积神经网络在原油物性预测领域鲜有先例,因此,研究将其应用于该领域具有重要的现实意义.

卷积神经网络的训练需要大量样本以提高其泛化能力. 由于作为训练标签的总氢物性含量需由实验室化验测得,所需时间长且化验成本高,样本较少. 针对光谱数据样本不足的问题,本文提出一种扩增数据的方法,并针对总氢物性预测建立RF-CNN (Regression forecasting convolutional neural network)模型,以解决小样本原油总氢物性回归预测问题.

1 核磁共振光谱数据获取

随着科技水平的飞速发展,在线核磁共振(NMR)技术日益完善,已成为当今最先进的原油物性分析方法之一. 国内某炼油厂的核磁共振原油快速评价系统如图1所示. 该系统主要由清洗装置、过滤装置、加热装置、NMR分析仪器和分析主机等组成.

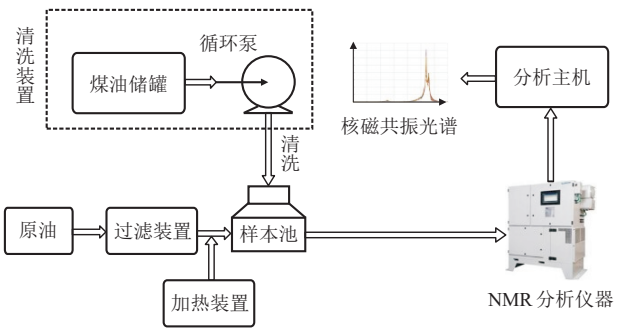

图1 NMR原油快速评价系统

首先,用清洗装置中的循环泵将煤油储罐中的煤油引入样本池对其进行清理;然后,将原油样品经过过滤、加热后引入样本池;最后,利用NMR分析仪器对样本池中的原油样本进行分析并将分析结果传至分析主机,从而得到核磁共振光谱. 其中NMR分析仪为脉冲傅里叶变换核磁共振仪,其原理结构如图2所示.

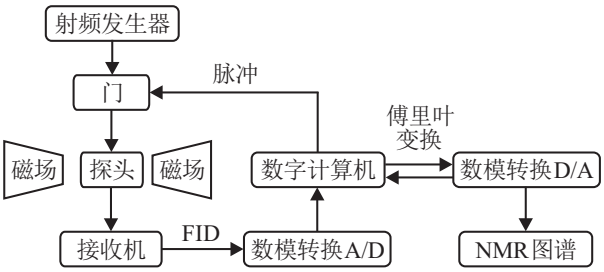

图2 傅里叶脉冲变换核磁共振仪

傅里叶脉冲变换核磁共振仪的原理如下:

1) 数字计算机控制射频发生器发射脉冲,使样本中不同化学环境的氢核同时被激发;

2) 氢核进入外加磁场后,因各种氢核所处化学环境不同,其实际所受磁场强度不同,故不同氢核在相同磁场强度下共振幅度及化学位移也不相同;

3) 去除磁场后,接收机接收到不同的自由感应衰减信号,叠加后经过模数数模转换、傅里叶变换,得到不同化学位移氢的含量图谱,即NMR图谱.

不同仪器的磁场强度不同,氢核通过磁场产生的化学位移也不同,需要建立一个标准,将标准化合物的化学位移设为基点. 一般采用四甲基硅烷(TMS)作为标准化合物. 其他质子化学位移值为与标准化合物化学位移的距离.

本文采用国内某炼油厂采集的原油样本数据,其核磁共振图谱由Aspect Imaging AI-60在线核磁共振分析仪测定. 工作频率为60 MHz,标准化合物为TMS,得到的核磁共振光谱如图3所示.

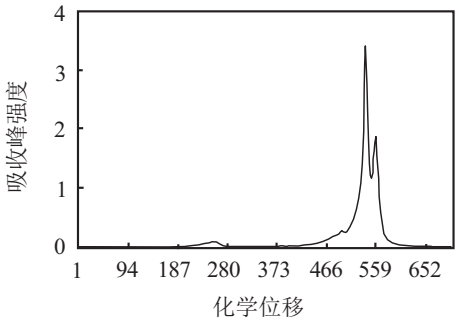

图3 NMR光谱图

2 核磁共振光谱数据的扩增及预处理

由图3可知,虽然NMR图谱本质是谱线数据,但其与传统卷积神经网络输入的图像相比,具有较大的稀疏性. 从图谱数据的角度看,即使对应物性值差距

较大,其对应的图谱图像形式也会非常相似,即大部分都是空白区域,只有谱线部分存在少量差异. 若将其直接以图像的形式作为输入,一方面其具备的特征非常有限,卷积神经网络不易学习到;另一方面,不能很好地发挥卷积核参数共享的性质,不宜直接作为输入. 因此,本文将采取序列数据的形式作为输入. 根据不同化学位移对应的吸收峰强度,可得到700维的序列数据,其对应的总氢物性含量为标签,则该700维序列数据与其标签构成一个样本.

实际生产中,样本的标签需要实验室化验较长时间测得,且化验成本高,故样本量很小,经常小于输入维度. 针对此高维小样本问题,传统的方法有小波变换法[13]、PCA降维算法[14]和PLS方法[15]等. 但此类方法的原理是根据数据之间的相关性对数据进行降维处理,因此,数据信息会有一定程度的损失. 本文采用将高维序列数据转换为二维矩阵形式进行处理,可以避免数据降维,提高信息的完整度. 但是,样本量依然不足,容易导致网络训练过程中的过拟合. 解决这个问题的办法是构建虚拟样本,增大数据集以提高卷积神经网络的泛化能力.

针对不同的具体问题,扩增数据集的方法也不同[16]. 按照其生成思想,主要分为以下3类:

1) 基于研究领域具体先验知识构造虚拟样本;

2) 基于扰动的思想构造虚拟样本;

3) 基于研究领域的分布函数构造虚拟样本.

本文基于扰动的思想扩增样本,采取在数据集中加入噪声的策略. 在样本中加入噪声较小时,相当于神经网络结构设计的正则化(Regularization)[17]. 为了使加入噪声的区域对输出影响较小,本文提出基于相关系数加入噪声以扩增数据集的方法,即在输入与标签相关度较小的区域加入噪声. 本文采用Pearson相关系数对样本输入与标签的相关度进行描述.

2.1　基于Pearson相关系数的光谱数据扩增

相关系数可以描述两个变量之间的线性相关性,同理,也可以描述两组向量的相关程度. 如果有两组观察向量A、B,其Pearson相关系数为

$$\rho(A,B)=\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\sigma_A}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right). \quad (1)$$

其中: N为每组样本数, A_i、B_i为A、B组样本中的第i个样本, μ_A和σ_A为A样本组的均值和标准差, μ_B和σ_B为B样本组的均值和标准差.

针对核磁共振光谱数据,假设有N个样本,其对应标签向量为Y, Y为一个N维向量. 将N个不同的700维序列数据样本相同维度的数据组成一个向量$X_i(i=1,2,\cdots,700)$, X_i为N维向量,则X_i与标签列向量Y的相关系数为

$$\mathrm{corr}_i=\rho(X_i,Y). \quad (2)$$

根据式(2),算出光谱数据各个维度与总氢物性标签的相关性如图4所示.

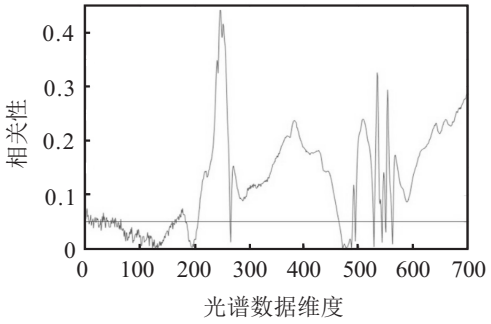

图4　各维度与总氢物性的相关性

由图4可知,一些维度的相关系数很小甚至为0. 假设$\rho(A,B)<\sigma$时为线性无关,在线性无关的维度i上给数据加入噪声,有

$$\dim_i=\dim_i\times(1\pm k_i\times\mathrm{rand}). \quad (3)$$

其中: rand为随机变化量,其取值范围在0～1之间; k_i为噪声强度系数. 于是如何加入噪声的关键在于k_i的确定.

针对原油总氢物性,假设某对700维序列数据样本(x_m,x_n)具有相似或相同的标签值,则可以认为两者具有相似或相同的物性值. 在i维上的变化程度用相对于前者的百分比表示为

$$\mathrm{Var}_i=\frac{x_{m,i}-x_{n,i}}{x_{m,i}}. \quad (4)$$

若这样的样本对有M个,对其求均值,有

$$\mathrm{Var}_{\mathrm{mean},i}=\frac{\sum\mathrm{Var}_i}{M}, \quad (5)$$

则$\mathrm{Var}_{\mathrm{mean},i}$为噪声在$i$维加入的上限值,即$\mathrm{Var}_{\mathrm{mean},i}=k_i$. 于是,式(3)可以表示为

$$\dim_i=\dim_i\times(1\pm\mathrm{Var}_{\mathrm{mean},i}\times\mathrm{rand}). \quad (6)$$

因此,数据集的扩增方法总结如下:

1) 根据式(2)计算光谱数据各个维度与总氢物性的相关度;

2) 根据式(6)对原始样本在相关度低于σ的维度加入噪声以生成虚拟样本;

3) 对每组原始样本$i(i=1,2,\cdots,N)$随机生成a组虚拟样本,总共可以得到$N(a-1)$组样本,实现对数据集的扩增,其中, a值由计算成本和模型所需样本规模确定.

2.2 光谱数据预处理

通过2.1节的数据集扩增方法得到大量的700维序列数据. 若使用一维卷积处理该序列数据,则由于数据中的位置限制,使得大量信息不易被提取到,获取的局部浅层特征的相关性不足,容易导致网络训练的欠拟合. 而将700维序列数据紧凑排列成二维矩阵形式并不会改变其包含的信息,只是改变了相对位置结构,这些相对位置的结构给卷积神经网络提供了大量的特征信息. 因此,针对原油核磁共振图谱,在输入神经网络训练之前,将700维序列数据重新排列为二维矩阵的数据形式进行训练. 处理前的光谱数据样本形式为

$$X(x_1,x_2,\cdots,x_m),\ m=700; \tag{7}$$

处理后,对于每个700维数据样本都有

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} \end{bmatrix},\ j\times k=700. \tag{8}$$

处理方式如图5所示. 由于数据排列的自由性,不少排列方式均会出现矩阵过“窄”或过“宽”的情况,使得卷积核不能较好地提取特征信息,训练中导致欠拟合的情况. 因此,需要大量的训练尝试才能够找到较好的排列方式.

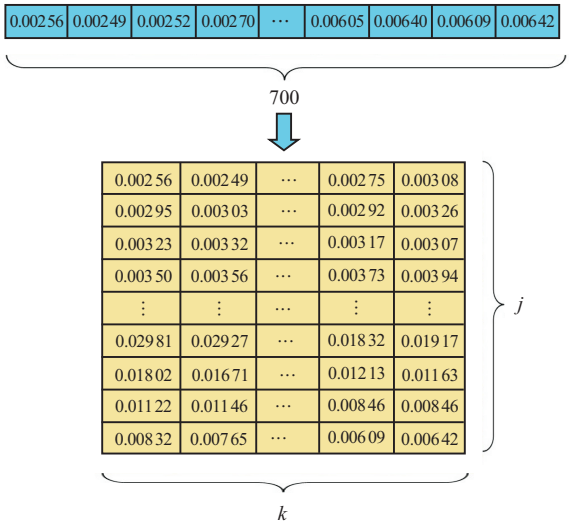

图5 700维序列数据处理方式

将700维序列数据排列成二维矩阵形式之后,一定程度上解决了特征提取不足的问题. 为了更好地提取特征,本文采取的策略是扩增数据冗余,即复制二维矩阵形式数据中选取的序列与当前数据“拼接”,使数据特征更易被一个卷积核同时卷积到,提取的特征更加完整,预测更加准确. 扩增数据冗余的方式如下:

$$X=\left[\begin{array}{cccc|ccc} x_{11} & x_{12} & \cdots & x_{1k} & x_{1k+1} & \cdots & x_{1k+h} \\ x_{21} & x_{22} & \cdots & x_{2k} & x_{2k+1} & \cdots & x_{2k+h} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & x_{jk+1} & \cdots & x_{jk+h} \end{array}\right]. \tag{9}$$

其中: $j\times a=700$, h为常数.

针对扩增数据冗余的选择问题,以二维矩阵数据倒数第2行为例进行说明,见图6.

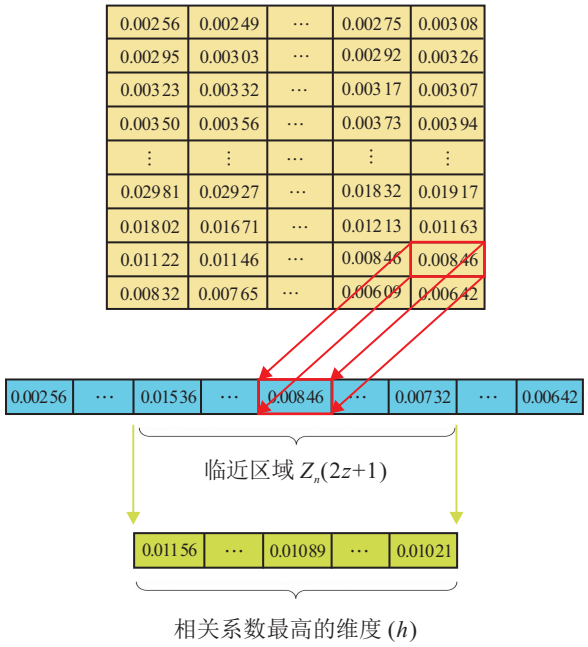

图6 二维矩阵数据冗余选取方式

如图6所示,数据冗余选取分两步进行:

1) 确定二维矩阵数据样本每行末尾数据 $x_{nk}(n=1,2,\cdots,j)$ 在700维序列数据中的所在维度,选取与其相邻的左右 z 个维度,用临近区域 Z_n 表示,有

$$Z_n\sim(x_{nk-z},x_{nk-z+1},\cdots,x_{nk},\cdots,x_{nk+z-1},x_{nk+z}); \tag{10}$$

2) 根据2.1节算出的相关系数,在每个临近区域 Z_n 中选取相关系数最高的 h 个维度数据依次排列在该对应二维矩阵行的后面.

用图6的方式对每行进行数据冗余扩增,得到的光谱数据见图7.

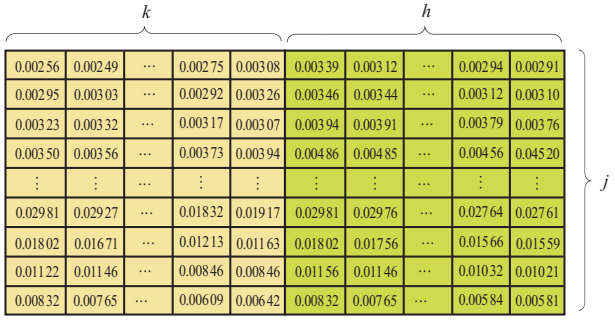

图7 冗余扩增后二维矩阵形式光谱数据

如图7所示,将原有的$j \times k$二维矩阵数据扩增为$j \times (k+h)$的二维矩阵数据,此方法在一定程度上避免了需要进行过多二维数据结构的排列实验.

3　RF-CNN模型及其回归预测算法

3.1　RF-CNN模型

针对总氢物性回归预测构建RF-CNN模型(见图8),其中包括:输入层(Input)、卷积层(Convolutional layer 1、Convolutional layer 2)、拼接归一层(Concat layer)、全连接层(Fully-connected layer)、输出层(Output)、线性回归层(Linear regression layer).

如图8所示,为了提取不同特征,卷积层(Convolutional layer 1)布置了3种大小不同、数量相同的卷积核,其参数分别为$5 \times 3 \times 16$、$3 \times 3 \times 16$、$1 \times 3 \times 16$. 对该二维矩阵数据进行卷积,其公式为

$$\mathrm{out}[n,i,:,:] = \mathrm{bias}[i] + \sum_{j=0}^{\mathrm{num\_filter}} \mathrm{data}[n,i,:,;] \times \mathrm{weight}[i,j,:,:], \tag{11}$$

其中data的4个维度分别为批处理大小(Batch size)、通道数、数据矩阵的高度和宽度. 本文批处理大小为128,该卷积层通道数为1,数据高35、宽40. 卷积得到3个不同的“数据块”,其参数分别为$31 \times 38 \times 16$、$33 \times 38 \times 16$、$35 \times 38 \times 16$.

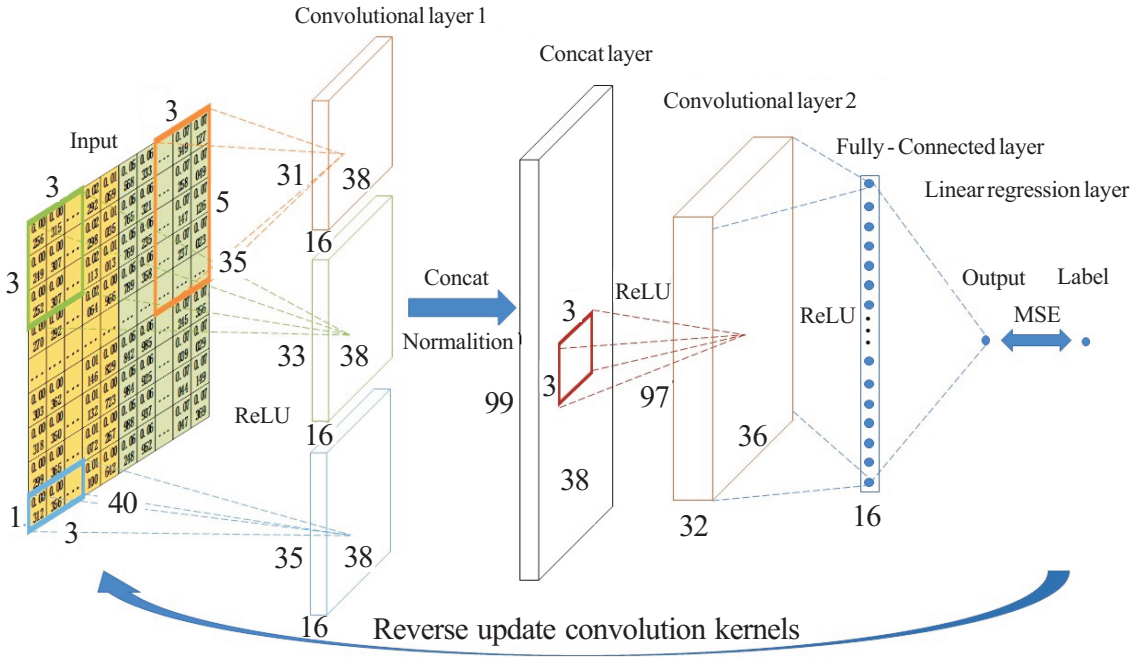

图 8　RF-CNN结构

拼接层(Concat layer)将卷积得到的“数据块”拼接并归一化以便下一次卷积运算,“拼接块”的参数为$99 \times 38 \times 16$.

卷积层(Convolutional layer 2)布置了32个的卷积核. data的4个维度分别为128、16、99、38. 卷积后的“数据块”参数为$97 \times 36 \times 32$.

全连接层(Fully-connected layer)由16个神经元组成,每个神经元与卷积层(Convolutional layer 2)所有神经元全连接. 其运算公式如下:

$$Y = XW^{\mathrm{T}} + b. \tag{12}$$

其中:X为输入矩阵数据,W为权重矩阵,b为偏置.

输出层(Output)仅有一个神经元,该神经元与上一层的16个神经元全连接. 计算公式同全连接层.

线性回归层(Linear regression layer)为该样本的标签,目标函数为均方差(MSE). 其表达式如下:

$$\mathrm{MSE}(y, f(x)) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i(x))^2. \tag{13}$$

其中:n为批处理大小,y_i为标签值,$f(x)_i$为训练输出值.

另外,网络中卷积层和全连接层的激活函数均采用ReLU函数

$$f(x) = \max(0, x), \tag{14}$$

相对于传统的sigmoid函数,其更容易学习优化,且不易丢失信息[18].

3.2　RF-CNN模型训练

上述模型中,优化的目标函数为输出与标签的均方差,因输出是W和B的函数,故式(13)可表示为

$$\mathrm{MSE}(y, f(x)) = C(W,B) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i(x))^2, \tag{15}$$

其中$W(w_1,w_2,\cdots,w_K)$和$B(b_1,b_2,\cdots,b_L)$为可训参数. 使用随机梯度下降法对所有可训参数反向更新. 以权值更新为例,有

$$\begin{cases} v'_k = \text{momentum} \times v_k - \eta\dfrac{\partial C}{\partial w_k}, \\ w'_k = w_k + v'_k. \end{cases} \tag{16}$$

其中:v_k为初始值为0的势能量,momentum为动量,$w'_k(k=1,2,\cdots,K)$为更新后的权值,η为学习率.

3.3 RF-CNN训练算法

本文采用随机梯度下降算法对RF-CNN进行训练,训练开始前,需要对网络权值和偏置进行初始化,本文采用Xavier初始化方法. 该方法由Glorot等[19]提出,是一种有效的神经网络初始化方法.

如图9所示,RF-CNN训练算法如下.

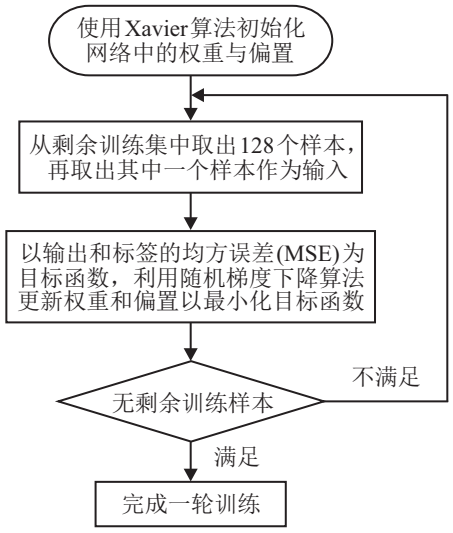

图9 RF-CNN训练算法流程

算法1 RF-CNN.

输入:二维矩阵数据

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & x_{1k+1} & \cdots & x_{1k+h} \\ x_{21} & x_{22} & \cdots & x_{2k} & x_{2k+1} & \cdots & x_{2k+h} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & x_{jk+1} & \cdots & x_{jk+h} \end{bmatrix},$$

RF-CNN网络结构,最大迭代次数M,训练样本数N.

Step 1: 利用Xavier算法对网络权值和偏置进行初始化,得到$W[w_1,w_2,\cdots,w_K]$和$B[b_1,b_2,\cdots,b_L]$;

Step 2: for epoch $=1$ to M do;

Step 3: for $l=1$ to N/n do; $\#n$为批处理大小,N为训练集样本数;

Step 4: 将每一批中随机一个样本输入到网络模型中运算,得到实际输出值$f_i(x)$;

Step 5: 采用式(15)作为目标函数,最小化目标函数;

Step 6: 利用式(16)反向更新权值和偏置,得到$W[w'_1,w'_2,\cdots,w'_K]$和$B[b'_1,b'_2,\cdots,b'_L]$,实现最小化目标函数;

Step 7: End for;

Step 8: End for.

输出:RF-CNN的最终权值$W[w^*_1,w^*_2,\cdots,w^*_K]$和偏置$B[b^*_1,b^*_2,\cdots,b^*_L]$.

4 实验与结果分析

本文基于Windows10操作系统,使用Visual Studio 2015作为编辑器,利用Python语言,结合深度学习框架MXNet进行程序的开发. 采用双路Nvidia GeForce GTX 1080 GPU进行计算处理.

4.1 光谱数据的处理

本文以国内某炼油厂采集的594组原油样本数据作为初始样本,对每个样本保持标签不变,采用基于相关系数的数据扩增方法产生19组虚拟样本,共得到11 880组数据. 其中判断相关与否的预值σ设为0.05,即将与标签的Pearson相关系数低于0.05的维度视为线性无关,在此类维度上随机加入噪声以生成虚拟样本. 为了批处理能够更好地选取全局梯度下降方向,加快网络收敛效率,将所有数据打乱并分为9 504组训练数据和2 376组测试数据. 核磁共振光谱数据扩增后的结果如图10所示.

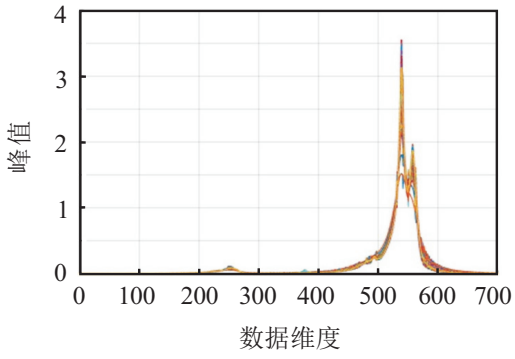

图10 扩增后的核磁共振光谱数据

根据2.2节的数据预处理方法,将700维序列数据转换成35×20二维矩阵形式并进行数据冗余扩增. 其中h的选取是根据矩阵数据的形式决定的,为了使数据更易于卷积,需要将数据扩增得更加“方正”,经过实验确定h值为20,故得到35×40的二维矩阵数据. 最后,将该二维矩阵数据作为RF-CNN的输入.

4.2　物性预测结果与分析

4.2.1　RF-CNN参数设定

使用扩增后的数据对RF-CNN模型进行训练. 训练前需要进行超参数设置. 其中: 网络学习率为0.12, 动量为0.8, 卷积核的步长为$(1,1)$, 5×5卷积核的扩充边缘为$(2,0)$, 3×5卷积核为$(1,0)$, 1×5卷积核为$(0,0)$, 3×3卷积核为$(0,0)$, 训练轮数为3 000.

4.2.2　数据集扩增前后对比实验及分析

对于同一卷积神经网络结构, 相同的学习参数(如学习率、动量、初始化方式、训练数据的二维处理结构等), 数据集扩增前后原油总氢物性含量的回归预测网络训练结果对比如图11所示. 其中准确率的计算公式如下:

$$\text{accuracy}=1-\frac{\sqrt{\sum\left(y_{\text{label}}-y_{\text{predict}}\right)^2}}{\sum y_{\text{label}}}. \tag{17}$$

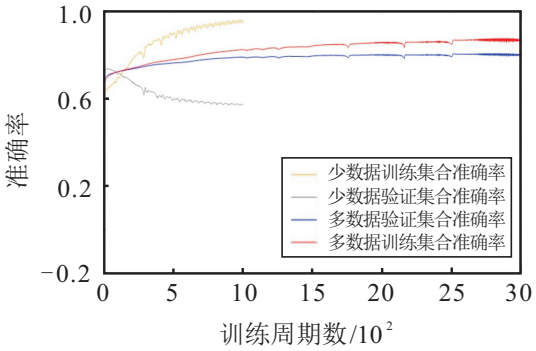

图 11　数据集扩增前后学习效果对比

由图11可知: 数据扩增前, 网络学习过拟合严重, 神经网络训练效果差; 数据集扩增后, 训练集准确率和验证集准确率平稳上升, 实现了网络对数据特征的有效学习. 因此, 数据集的扩增对于防止卷积神经网络训练的过拟合, 提高网络的泛化能力具有十分重要的作用.

4.2.3　扩增数据后的RF-CNN与传统PLS方法对比实验及分析

使用传统的偏最小二乘回归方法(PLS Regression)对原油物性中的总氢物性含量进行预测分析. 根据各个维度与总氢物性含量的相关性, 选取保留40 %的维度进行降维. PLS方法无需进行数据扩增, 将594组数据随机打乱后, 取494组进行训练, 100组进行测试. 同时, 将扩增数据训练得到的RF-CNN对相同100组数据进行预测. 两种方法的回归预测对比如图12所示.

由图12可知, 针对相同的预测样本, 相比于传统的PLS方法, RF-CNN对该物性的回归预测效果更好. 同时, 还给出了其误差绝对值的对比图和误差绝对值统计分布折线图, 分别如图13和图14所示.

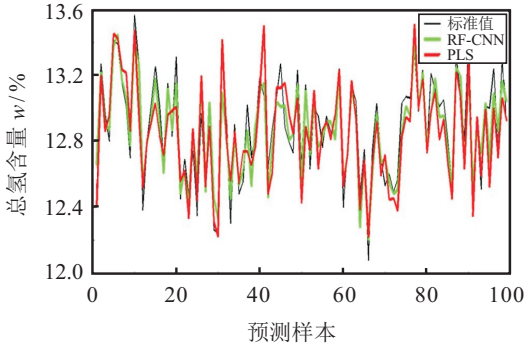

图 12　总氢物性回归预测对比

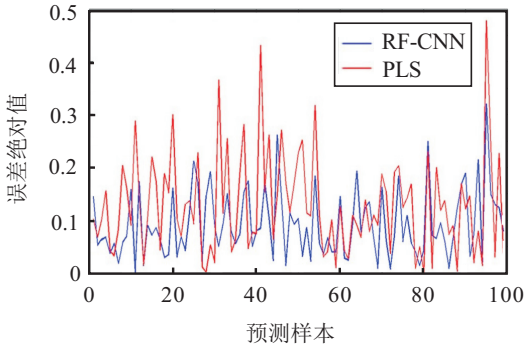

图 13　总氢物性回归预测误差对比

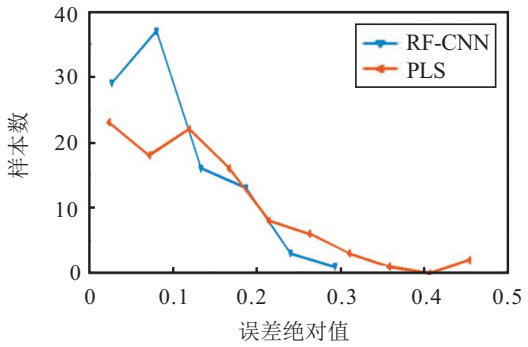

图 14　总氢物性误差绝对值统计分布折线

由对比实验结果可知, 扩增数据后训练得到的RF-CNN在绝大多数样本的预测上, 无论是误差比例还是预测精度, 都比传统的PLS方法更具优势.

5　结　论

本文针对小数据集原油总氢物性回归预测问题, 提出了光谱数据的扩增预处理方法, 并且针对处理后的数据构建了卷积神经网络模型RF-CNN. 该预处理方法根据光谱数据的性质, 将700维序列光谱数据二维矩阵化且对其进行冗余扩增, 结合RF-CNN回归预测算法实现了对总氢物性含量的预测. 最后对比了数据扩增前后的训练效果, 并用传统PLS回归预测方法对相同预测样本进行预测. 由实验对比结果可知: 扩增数据解决了小样本的过拟合问题; 相比于传统

的PLS方法,RF-CNN预测更具稳定性和精确性. 在下一步工作中,拟改进网络结构和算法,进一步提高预测精度.

参考文献(References)

[1] 陈瀑, 李敬岩, 褚小立, 等. 拉曼和红外光谱快速评价原油性质的可行性比较[J]. 石油炼制与化工, 2016, 47(10): 98-102.
(Chen P, Li J Y, Chu X L, et al. Comparison of the feasibility of rapid evaluation of crude oil by raman and infrared spectroscopy[J]. Petroleum Processing and Chemical Industry, 2016, 47(10): 98-102.)

[2] 褚小立, 田松柏, 许育鹏, 等. 近红外光谱用于原油快速评价的研究[J]. 石油炼制与化工, 2012, 43(1): 72-77.
(Zhu X L, Tian S B, Xu Y P, et al. Study on rapid evaluation of crude oil by near infrared spectroscopy[J]. Petroleum Refining and Chemical Industry, 2012, 43(1): 72-77.)

[3] Le Cun Y, Jackel L D, Boser B, et al. Handwritten digit recognition: Applications of neural network chips and automatic learning[J]. IEEE Communications Magazine, 1989, 27(11): 41-46.

[4] Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification[C]. Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 1701-1708.

[5] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]. Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Boston, 2015: 815-823.

[6] Levi G, Hassner T. Age and gender classification using convolutional neural Networks[C]. Proc of the IEEE Conf on Computer Vision and Pattern Recognition Workshops. Santiago, 2015: 34-42.

[7] Hou X, Zhang L. Saliency detection: A spectral residual approach[C]. IEEE Conf on Computer Vision and Pattern Recognition. Minneapolis: IEEE, 2007: 1-8.

[8] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[9] Khan S H, Bennamoun M, Sohel F, et al. Automatic feature learning for robust shadow detection[C]. Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 1939-1946.

[10] 乔俊飞, 安茹, 韩红桂. 基于RBF神经网络的出水氨氮预测研究[J]. 控制工程, 2016, 23(9): 1301-1305.
(Qiao J F，An R，Han H G. Water ammonia nitrogen prediction research based on RBF neural network[J]. Control Engineering of China, 2016, 23(9): 1301-1305.)

[11] 王旭东, 邵惠鹤. 基于神经网络的通用软件测量技术[J]. 自动化学报, 1998, 24(5): 702-706.
(Wang X D, Shao H H. General software measurement technology based on neural network[J]. Acta Automatica Sinica, 1998, 24(5): 702-706.)

[12] 张昭昭, 乔俊飞, 余文. 基于动态神经网络的瓦斯浓度实时预测方法[J]. 控制工程, 2016, 23(4): 478-483.
(Zhang Z Z, Qiao J F, Yu W. Forecasting coalmine gas concentration based on dynamic neural network[J]. Control Engineering of China, 2016, 23(4): 478-483.)

[13] Ding L, Xiang Y H, Huang A M, et al. Quantitative prediction of holocellulose, lignin, and microfibril angle of chinese fir by BP-ANN and NIR spectrometry[J]. Spectroscopy and Spectral Analysis, 2009, 29(7): 1784-1787.

[14] Liu T. Application of PCA to diesel engine oil spectrometric analysis[J]. Spectroscopy and Spectral Analysis, 2010, 30(3): 779-782.

[15] Molina V D, Uribe U N, Murgich J. Partial least-squares(PLS) correlation between refined product yields and physicochemical properties with the 1H nuclear magnetic resonance(NMR) spectra of colombian crude oils[J]. Energy and Fuels, 2007, 21(3): 1674-1680.

[16] 于旭, 杨静, 谢志强, 等. 虚拟样本生成技术研究[J]. 计算机科学, 2011, 38(3): 16-19.
(Yu X, Yang J, Xie Z Q, et al. Research on virtual sample generation technology[J]. Computer Sience, 2011, 38(3): 16-19.)

[17] Bishop Chris M. Training with noise is equivalent to tikhonov regularization[J]. Neural Computation, 2014, 7(1): 108-116.

[18] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]. Int Conf on Artifical Intelligence and Statistics. Melbourne, 2012: 315-323.

[19] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]. Proc of the 13th Int Conf on Artificial Intelligence and Statistics. Zakopane, 2010: 249-256.

（责任编辑：李君玲）