文章编号：1002-2082 (2021) 03-0504-06

基于深度置信网络的红外光谱鉴别汽油掺混

王明吉[1]，梁 涛[1]，李 栋[2]，王 迪[1]，王秋实[2]
（1. 东北石油大学 物理与电子工程学院，黑龙江 大庆 163318；2. 东北石油大学 土木建筑工程学院，黑龙江 大庆 163318）

摘 要：为实现掺混汽油快速无损鉴别，提出一种利用 t 分布邻域嵌入结合深度置信网络的鉴别方法，以解决机器学习中高维特征向量间的非线性关系。以 92#、95#、98#及定比混合汽油为研究对象，采用多元散射校正算法对原始红外波段投射光谱测量数据进行预处理，利用 t-SNE 非线性方法进行光谱数据降维处理，分别采用深度置信网络和极限学习机建立汽油种类光谱鉴别模型并对比分析两种方法识别精度。研究表明：该文所选择方法构建的汽油鉴别模型性能更优，对汽油种类预测精准度高达 92.5%，从而验证了该方法在汽油鉴别中的有效性。研究结果可为掺混成品油鉴别及溯源研究提供技术支持。

关键词：红外光谱；深度置信网络；掺混；鉴别
中图分类号：TN201;O433.1 **文献标志码：**A **DOI：**10.5768/JAO202142.0303002

Identification of gasoline blending by infrared spectroscopy based on deep belief networks

WANG Mingji[1], LIANG Tao[1], LI Dong[2], WANG Di[1], WANG Qiushi[2]
（1. School of Physics and Electronic Engineering, Northeast Petroleum University, Daqing 163318, China; 2. School of Civil Engineering and Architecture, Northeast Petroleum University, Daqing 163318, China）

Abstract: In order to realize the fast nondestructive identification of blended gasoline, an identification method based on t-distributed stochastic neighborhood embedding(t-SNT) combined with deep belief networks was proposed to solve the nonlinear relationship between high-dimensional feature vectors in machine learning. Taking 92#, 95#, 98# and fixed ratio blended gasoline as the research objects, the projection spectrum measurement data in original infrared band was preprocessed by multivariate scattering correction algorithm, and the dimension reduction of spectral data was carried out by using t-SNE nonlinear method. The spectral identification model of gasoline types was established by using deep belief networks and extreme learning machine respectively, and the identification accuracy of the two methods was compared and analyzed. The research shows that the gasoline identification model constructed by this method has better performance, and the prediction accuracy of gasoline types is as high as 92.5%, which verifies the effectiveness of this method in gasoline identification. The results of this research can provide technical support for the identification and traceability of blending refined oil products.

Key words: infrared spectrum; deep belief networks; blending; identification

引言

随着我国交通大发展的持续深入推进，用户对汽油的需求呈现爆炸式增长。然而，大量不法企业为追求最大限度利润，擅自用化工原料和添加剂兑制、混配“调和汽油”，给消费安全带来了极大的隐患。因此亟需研究对掺混成品油进行快速鉴

收稿日期：2021-01-12； **修回日期：**2021-03-12
基金项目：中国石油科技创新基金研究项目（2018D-5007-0608）；东北石油大学优秀中青年创新团队基金（KYCXTD201901）；大庆市指导性科技项目（zd-2019-04）
作者简介：王明吉（1963—），男，博士，教授，主要从事光电检测技术及应用等方面研究。E-mail: wmjglj@163.com
通信作者：李栋（1979—），男，博士，教授，主要从事油气介质激光检测技术等方面的研究。E-mail: lidonglvyan@126.com

别的方法。

由于红外光谱分析技术具有检测速度快、效率高、成本低等特点[1], 已被广泛应用于成品油分析领域[2]。Veras 等人利用主成分分析(principal component analysis, PCA)结合聚类分析的方法对108 个柴油样品的原产地进行分类[3]; 姜黎等人利用主成分分析结合马氏距离的方法比较汽油的2 个特征波段建模的分类效果[4]; 王丽等人利用主成分分析结合模糊聚类实现了对海洋溢油样本的快速分类[5]。然而主成分分析属于线性降维方法, 其不能准确提取光谱数据中的非线性特征, 导致光谱数据在降维的过程中部分有用信息丢失及鉴别模型精准度下降。鉴于此, 本文采用非线性降维方法中的 t 分布邻域嵌入(t-distributed stochastic neighbor embedding, t-SNE)算法[6]对光谱数据进行降维处理, 同时结合深度置信网络方法[7]建立汽油鉴别模型, 并与极限学习机鉴别算法[8]进行识别精度对比分析, 以解决掺混汽油红外光谱鉴别技术中线性降维方法缺陷和高精度识别模型选择问题。

1　材料与方法

1.1　样品来源与光谱测试

本实验所使用的 92#、95#以及 98#汽油样品均购置于大庆中石化加油站, 掺混汽油样品由 92#、95#、98#汽油按照 1∶1∶1 配制而成, 每种样品各50 份用于红外光谱测量实验。其中, 每种类型取其 40 份作为训练集, 10 份作为测试集。

1.2　光谱数据处理

通过实验所采集到的光谱信息不仅包含样本特征信息, 还包含外界的干扰因素[9], 这些干扰因素会对模型建立造成一定的影响。因此, 有必要对原始光谱数据进行预处理[10]。分别采用多元散射校正(multiplication scattering correction, MSC)、标准正态变换(standard normal variate, SNV)以及一阶导数对原始光谱数据进行预处理, 从而选择最适合本文的预处理方法。

红外光谱数据通常维度很高, 如若将全部光谱数据参与模型构建, 将会导致该模型识别效率下降, 通常在建立模型之前需对光谱数据进行降维处理, 本文采用 t 分布邻域嵌入(t-distributed stochastic neighbor embedding, t-SNE)算法对光谱数据进行降维[5]。

t-SNE 算法具体步骤如下:

a) 用条件概率 $p_{j|i}$ 表示高维空间中邻近数据点 x_i 与 x_j 的相似度, 邻近数据点之间的相似度越高, 则条件概率 $p_{j|i}$ 值也就越大, 且其服从高斯分布[11], 条件概率 $p_{j|i}$ 计算公式为

$$p_{j|i}=\frac{\exp\left(-\left\|x_i-x_j\right\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\left\|x_i-x_k\right\|^2/2\sigma_i^2\right)} \tag{1}$$

式中σ_i为高斯分布标准差。

将高维中邻近数据点 x_i 与 x_j 在低维中的映射点记为 y_i 与 y_j, 并计算其相似的条件概率 $q_{j|i}$

$$q_{j|i}=\frac{\exp\left(-\left\|y_i-y_j\right\|^2\right)}{\sum_{k\neq i}\exp\left(-\left\|y_i-y_k\right\|^2\right)} \tag{2}$$

b) $p_{j|i}$ 与 $q_{j|i}$ 分别表示高维空间中数据点 x_i、x_j 与低维空间中数据点 y_i、y_j 之间的联合概率, 如(3)式和(4)式所示:

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n} \tag{3}$$

$$q_{ij}=\frac{\left(1+\left\|y_i-y_j\right\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left\|y_k-y_l\right\|^2\right)^{-1}} \tag{4}$$

c) 此时新的代价函数 C 可以表示为

$$\mathrm{C}=KL(P\|Q)=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{5}$$

式中: KL 为 K-L 散度(Kullback-Leibler divergence); P 与 Q 分别为高维空间和低维空间中度量点对分布概率分布。

d) 在低维空间中, t-SNE 算法将使用 t 分布(student t-distribution)代替高斯分布以表示两个点之间的相似度。t 分布在低维空间中使用更注重长尾分布, 使同类的样本点在低维空间中相隔距离较近, 不同类型的样本点相隔距离较远[12]。t-SNE 梯度计算式可以表示为

$$\frac{\delta y}{\delta x}=4\sum_{\mathrm{j}}(p_{ij}-q_{ij})(y_i-y_j)\left(1+\left\|y_i-y_j\right\|^2\right)^{-1} \tag{6}$$

1.3　基于深度置信网络的汽油种类鉴别方法

1.3.1　数据集处理

数据集中标记的鉴别汽油种类是离散型数据, 不能直接参与 DBN 模型计算, 因此在构建 DBN 模型之前需要利用 One-Hot 编码进行转换处理。One-Hot 编码使用 0 或 1 对多个分类或状态进行编码, 将每个分类或状态作为独立属性, 任意时刻只有其中一个属性有效, 将对应的有效属性设置为1[13], 4 种类型的汽油对应的编码如表 1 所示。

表 1 同类型汽油 One-Hot 编码

Table 1 One-Hot coding of different types of gasoline

类型	92#	95#	98#	掺混
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

1.3.2 深度置信网络鉴别模型

深度置信网络既可以用于非监督学习，也可以用于监督学习，其由多层受限玻尔兹曼机（restricted Boltzmann machines, RBM）组成，通过训练各个神经元之间的权重和偏置，可使整个神经网络以最大概率生成训练数据[14]。DBN 一般由 3 层或 3 层以上神经元构成，神经元分为显性神经元和隐性神经元。显性神经元接受输入数据，隐性神经元提取数据的特征，其中每一个神经元代表数据向量的一维。与传统方法相比，DBN 不仅有多隐层的深度结构，而且通过逐层训练学习以获取特征，能够刻画出数据更丰富的内在信息，使分类和预测更加容易[15]。

DBN 模型如图 1 所示，第一层为输入数据的可见层，输入不同类型汽油光谱特征向量，数据经过 2 个隐层逐层训练后到达最后 Softmax 分类器，Softmax 分类器输出汽油种类。

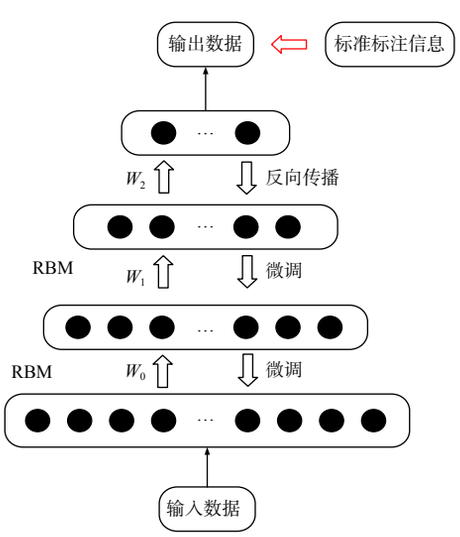

图 1 汽油种类鉴别 DBN 模型

Fig. 1 DBN model for gasoline types identification

DBN 模型中核心部分是 RBM，RBM 是一种层内无连接、层间全连接的两层神经网络[16]，其结构如图 2 所示。

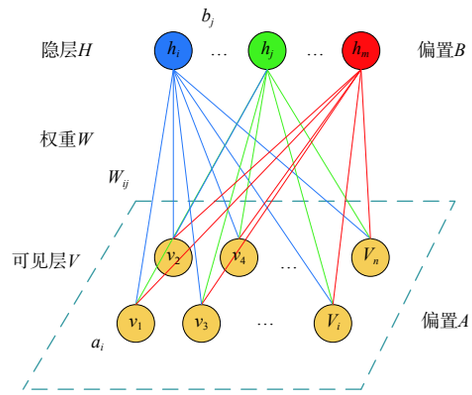

图 2 RBM 结构图

Fig. 2 Structure diagram of RBM

图 2 中，a_i 和 b_j 分别为可见层神经元和隐层神经元的偏置值，w_{ij} 为层间相连的神经元的权值。

RBM 中状态（v, h）的能量函数如（7）式所示，其函数值越小，则表示此时的 RBM 处于理想状态，汽油类型鉴别的错误率也就越低。

$$E\{v,(h|\theta)\}=-\sum_{i=1}^{n}a_iv_i-\sum_{j=1}^{m}b_jh_j-\sum_{i=1}^{n}\sum_{j=1}^{m}v_iw_{ij}h_j \quad (7)$$

此时，RBM 的可见层与隐层对应的神经元激活概率可以表示为（8）式和（9）式：

$$p\{h_j=(1|v),\theta\}=\sigma\left(b_j+\sum_i v_iw_{ij}\right) \quad (8)$$

$$p\{v_j=(1|v),\theta\}=\sigma\left(a_i+\sum_j v_jw_{ij}\right) \quad (9)$$

式中，σ 为 sigmoid 激活函数，计算方法如（10）式所示：

$$\sigma(x)=\frac{1}{1+e^{-x}} \quad (10)$$

为了提高 DBN 模型的训练速度，Hionton 等人提出了通过对比散度算法（CD-K）来构建可见层节点概率分布，其发现当 K=1 时，即只进行一步 Gibbs 采用便获得比较好的学习效果[17]。

2 结果与讨论

2.1 透射光谱分析

汽油样品红外光谱图如图 3 所示，从中可以看出，不同型号汽油样品的红外光谱大致相同，很难用肉眼进行区分。但红外光谱记录物质分子振动情况，而分子振动频率取决于组成原子的质量、化学键以及物质内部结构基团，所以原子的种类和结构基团的组合都可以在红外光谱图上表现出来，即不同物质的吸收谱带也不相同[18]。因此，借

助化学计量学的方法可以对不同型号的汽油进行聚类分析。

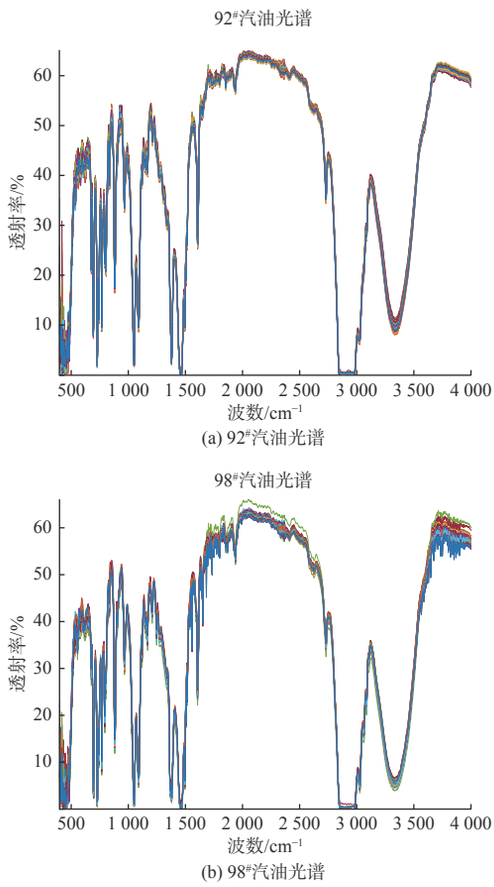

图 3　部分样品原始光谱图

Fig. 3　Original spectrogram of partial samples

2.2　原始光谱预处理分析

部分汽油样品原始光谱数据经预处理后的光谱图如图 4 所示。其中, 导数处理虽然可以有效地消除基线和其他背景干扰, 使某些未分辨开的重叠光谱分辨开, 但是会引入噪声, 降低信噪比[19]。

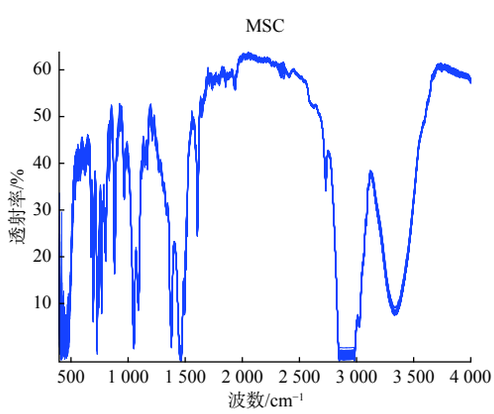

图 4　预处理后光谱图

Fig. 4　Spectrogram after preprocessing

MSC 主要是消除由于颗粒分布不均匀及颗粒大小不同产生的散射对光谱的影响, 其认为每条光谱与“理想光谱”都成线性关系, 但在大多数情况下这种情形并不存在, 而且光散射引起的背景非常复杂, 仅靠校正集的平均光谱作为标准光谱是存在误差的[20]。SNV 主要用来消除固体颗粒大小、表面散射以及光程变化对光谱的影响, 但是假设乘法效应在整个光谱范围内是均匀的, 并不一定能实现[21]。因此, 本文所选择的 3 种预处理方法都有各自优劣之处, 需要进一步分析来选择最适合本文的预处理方法。

为确定最适合本文的预处理方法, 将 MSC、SNV 和一阶导数预处理后的光谱数据利用 t-SNE 算法进行数据降维并将前 3 个特征向量进行可视化处理, 最终得出的结论为经 MSC 预处理后的 4 种汽油光谱特征数据不仅各自聚集在一起, 而且还互不相交, 能够很好地将这 4 种汽油区分开。因此选择多元散射校正作为建模前的原始光谱数据预处理方法。

2.3　光谱数据降维方法比较及分析

为验证所选择的 t-SNE 算法具有一定的优越性, 因此将 PCA 算法与 t-SNE 算法提取到的汽油光谱特征进行特征可视化图以比较分类效果。

选择累积贡献率超过 90% 的前 10 个特征代表汽油光谱特征, 即将汽油红外光谱数据维度降至 10 维, 将其前 3 个特征向量进行可视化, 结果如图 5 所示。

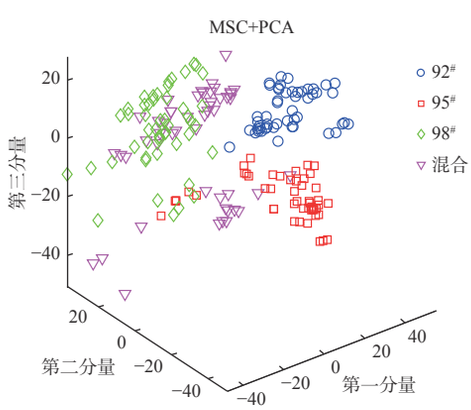

图 5　PCA 特征可视化图

Fig. 5　Diagram of PCA feature visualization

由图 5 可以看出, 经 PCA 算法提取的汽油光谱特征数据分类效果比较差, 这是因为汽油的红外光谱中含有非线性特征信息, 而 PCA 属于线性降维方法, 不能准确提取红外光谱数据中的非线

性特征信息，从而导致光谱数据在降维的过程中部分有用信息丢失，造成模型鉴别精准度下降。

2.4 DBN 模型的建立及对比分析

DBN 模型中的迭代次数和网络深度会对模型预测精准度产生较大的影响，以模型最终的预测精准度为判断标准来确定其参数。

DBN 结构主要分为输入层、隐含层和输出层，其网络深度主要体现在隐含层数量上。又因为 t-SNE 将汽油样品光谱数据降维至 5 维，因此将输出层节点数设置为 5，输出层神经元的个数需要根据成品油分类数量决定，需要将 92#、95#、98#以及掺混成品油区分开，因此输出层神经元的节点数设置为 4。为了确定最合适的模型迭代次数，分别将迭代次数设置为 50、60、70、80、90、100、110、120、130、140、150 和 160 进行模型构建，当迭代次数为 100 时，DBN 网络模型的识别准确度最高，而其他迭代次数的识别准确度都低于迭代次数为 100 时的识别准确度，将 DBN 网络模型的迭代次数设置为 100。为了确定模型的最佳隐含层数量，分别建立 1 至 4 层隐含层的 DBN 网络模型，以比较不同隐含层数对模型预测精准度影响。当隐含层数设置为 1 时，DBN 网络模型的识别准确度仅为 75%，当隐含层数量增加到 2 时，DBN 模型的识别准确度相较一个隐含层时有较大的提升，识别准确度已到达 92.5%，然而继续增加网络模型的隐含层数时，其模型识别度开始降低。因此，所建立的 DBN 网络模型的隐含层数量为 2。因此，构建一个结构为 5-10-20-4 的 DBN 网络模型对 4 种类型汽油样本进行特征学习和分类。

为了进一步验证分类算法有效性，分别利用深度置信网络算法与极限学习机算法建立汽油鉴别模型并比较这两种模型在测试集中的鉴别精准度。ELM 模型的识别精准度为 80%，而 DBN 模型的识别精准度为 92.5%。由此可见，DBN 模型分类效果更加良好。这是由于采用非线性算法 t-SNE 对光谱数据进行降维处理，降低了数据在降维过程中有用信息丢失的可能性，再者 DBN 模型拥有更深层次的网络学习结构，训练网络时采用反向传播微调方法，使得训练后的网络具有更好的识别能力，因而汽油种类鉴别精准度更高。

3 结论

本文提出了一种 t-SNE 和 DBN 二者相结合的汽油种类鉴别方法。在对汽油原始光谱数据进行多元散射校正预处理后，利用 t-SNE 算法对预处理后的光谱数据进行降维以提取光谱特征信息，最后将光谱特征信息作为 DBN 网络的输入并构建汽油种类鉴别模型，通过在 MATLAB 上进行测试，最终的实验结果表明优选的方法具有更好的鉴别效果。

参考文献：

[1] CHEN Yueyang, GAO Zhishan, YU Xiaohui, et al. Screening near infrared spectral information based on interval combination moving window method[J]. Applied Optics, 2017, 38(1): 99-105.
陈玥洋, 高志山, 郁晓晖, 等. 基于区间组合移动窗口法筛选近红外光谱信息[J]. 应用光学, 2017, 38(1): 99-105.

[2] TAN Ailing, BI Weihong. Quantitative modeling analysis of multi-component complex oil spill sources based on near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2019, 43(5): 86-90.
谈爱玲, 毕卫红. 基于近红外光谱法的多组分复杂溢油源定量建模分析[J]. 光谱学与光谱分析, 2019, 43(5): 86-90.

[3] VERAS G, GOMES A, DASILVA A. Classification of biodiesel sing NIR spectrometry and multivariate techniques[J]. Talanta, 2009, 83(2): 565-568.

[4] JIANG Li, ZHANG Jun, CHEN Zhe, et al. Pattern recognition analysis of finished gasoline based on different bands[J]. Spectroscopy Laboratory, 2010, 27(3): 1208-1212.
姜黎, 张军, 陈哲, 等. 基于不同波段对成品汽油的模式识别分析[J]. 光谱实验室, 2010, 27(3): 1208-1212.

[5] WANG Li, ZHUO Lin, HE Ying, et al. Identification of oil spills by near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2009, 24(12): 1537-1539.
王丽, 卓林, 何鹰, 等. 近红外光谱技术鉴别海面溢油[J]. 光谱学与光谱分析, 2009, 24(12): 1537-1539.

[6] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learing Research, 2008, 9: 2579-2605.

[7] HINTON G E, OSINDERO S, THE Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[8] YUAN Peipei, CHEN Hong, ZHOU Yicong, et al. Gener-

alization ability of extreme learning machine with uniformly ergodic Markov chains[J].Neurocomputing, 2015, 167: 528.

[9] HAO Yong, CHEN Bin. Analysis of several wavelet denoising methods in near infrared spectral preprocessing[J]. Spectroscopy and Spectral Analysis, 2006, 26(10): 1838-1841.
郝勇, 陈斌. 近红外光谱预处理中几种小波消噪方法的分析[J]. 光谱学与光谱分析, 2006, 26(10): 1838-1841.

[10] SHEN Yong, GUO Tiantai, KONG Ming, et al. Quantitative analysis of FTIR spectra of mine gas based on D-ELM[J]. Applied Optics, 2016, 37(5): 725-729.
沈永, 郭天太, 孔明, 等. 基于D-ELM的矿井气体FTIR光谱定量分析[J]. 应用光学, 2016, 37(5): 725-729.

[11] WANG Zhenhao, DU Hongjin, LI Guoqing, et al. Unsupervised identification of coherent clusters based on t-distribution neighborhood embedding[J]. Power System Protection and Control, 2018, 46(22): 64-71.
王振浩, 杜虹锦, 李国庆, 等. 基于t-分布邻域嵌入的同调机群无监督识别[J]. 电力系统保护与控制, 2018, 46(22): 64-71.

[12] DONG Jun. Research on dimensionality reduction of high-dimensional data based on data set oriented ST-SNE algorithm[J]. Computational Technology and Automation, 2018, 37(4): 116-122.
董骏. 面向数据集的ST-SNE算法高维数据降维研究[J]. 计算技术与自动化, 2018, 37(4): 116-122.

[13] SHI Wenbing, GE Bin, SU Shuzhi. Recognition of maintenance behavior of Hu sheep based on deep belief network[J]. Journal of Sensing Technology, 2020, 33(7): 1020-1026.
石文兵, 葛斌, 苏树智. 基于深度信念网络的湖羊维持行为识别[J]. 传感技术学报, 2020, 33(7): 1020-1026.

[14] ZHANG Chunxia, JI Nannan. Limited Boltzmann machine[J]. Journal of Engineering Teaching, 2015, 32(2): 159-173.
张春霞, 姬楠楠. 受限波尔兹曼机[J]. 工程教学学报, 2015, 32(2): 159-173.

[15] HU Renwei, YU Yue, NI Minglong, et al. Identification of adulteration of lotus seed powder by near infrared spectroscopy based on deep belief network[J]. Food Science, 2020, 41(6): 298-303.
胡仁伟, 俞玥, 倪明龙, 等. 基于深度信念网络的近红外光谱鉴别莲子粉掺假[J]. 食品科学, 2020, 41(6): 298-303.

[16] JIN Peng, XIA Xiaofeng, QIAO Yan, et al. Anomaly detection algorithm for high-dimensional sensor data based on deep belief network[J]. Journal of Sensing Technology, 2019, 32(6): 892-901.
金鹏, 夏晓峰, 乔焰, 等. 基于深度信念网络的高维传感器数据异常检测算法[J]. 传感技术学报, 2019, 32(6): 892-901.

[17] HINTON G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8): 1771-1800.

[18] FENG Ding, LI Dengao, ZHAO Jumin. Study on the detection system of pulverized coal calorific value by spectrum[J]. Applied Optics, 2014, 35(1): 111-115.
冯丁, 李灯熬, 赵菊敏. 光谱对煤粉发热量检测系统的研究[J]. 应用光学, 2014, 35(1): 111-115.

[19] ZHENG Limin, ZHANG Luda, GUO Huiyuan, et al. Application of near infrared spectral band optimization in donkey milk composition analysis[J]. Spectroscopy and Spectral Analysis, 2007(11): 2224-2227.
郑丽敏, 张录达, 郭慧媛, 等. 近红外光谱波段优化选择在驴奶成分分析中的应用[J]. 光谱学与光谱分析, 2007(11): 2224-2227.

[20] LI Qingbo, BI Zhiqi, SHI Dongdong, et al. Study on the method of identifying the origin of basic fish meal by near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2020, 40(9): 2804-2808.
李庆波, 毕智棋, 石冬冬, 等. 基鱼粉产地溯源的近红外光谱判别方法研究[J]. 光谱学与光谱分析, 2020, 40(9): 2804-2808.

[21] MENG Qinglong,ZHANG Yan, SHANG Jing. Non-destructive detection of apple surface scars by optical fiber spectroscopy combined with pattern recognition[J]. Laser Technology, 2019, 43(5): 86-90.
孟庆龙, 张艳, 尚静. 光纤光谱结合模式识别无损检测苹果表面疤痕[J]. 激光技术, 2019, 43(5): 86-90.