第42卷，第6期
2022年6月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 6, pp1735-1740
June, 2022

分组全连接的近红外光谱定量分析网络

余志荣，洪明坚*

重庆大学大数据与软件学院，重庆 401331

摘 要 全连接网络作为深度学习中的一种典型结构，几乎在所有神经网络模型中均有出现。在近红外光谱定量分析中，光谱数据样本数量较少，但每个样本的维度高。导致了两个问题：将光谱直接输入网络，网络的参数量会十分庞大，训练模型需要更多的样本，否则模型容易进入过拟合状态；在输入网络前对光谱进行降维，虽解决了网络参数量过大的问题，但会丢失一部分信息，无法充分发挥网络的学习能力。针对近红外光谱的特性，提出了一种分组全连接的近红外光谱定量分析网络GFCN。该网络在传统的两层全连接网络的基础上，用若干个小的全连接层替代第一个全连接层，克服了直接输入光谱导致网络参数量过大的缺点。采用Tecator和IDRC2018数据集对该方法进行测试，同时与全连接网络FCN和偏最小二乘PLS两种方法进行对比。结果显示：在两个数据集上，GFCN预测效果均优于FCN和PLS。在只有少量样本参与建模的情况下，GFCN依然能够保持较高的预测效果。表明，GFCN可以用于近红外光谱的定量分析，并且适应样本较少的场景，具有重要的研究价值和广泛的应用场景。

关键词 光谱分析；近红外光谱；全连接网络；定量分析

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)06-1735-06

引 言

近红外光谱技术是一种高效快速的现代分析技术，具有测量方便、速度快、分析成本低等优点，已成为质量控制、品质分析和在线分析的非接触无损检测的重要手段，并广泛应用于农业生产[1]、食品安全[2]、石油化工[3]等领域。因为近红外光谱具有多重共线性，加上仪器和环境等因素的影响，实际采集的原始光谱中还可能会包含一些噪声信号，所以，在近红外光谱定量分析中需要用到多元校正方法。常用的多元校正方法有：多元线性回归(multiple linear regression，MLR)[4]、主成分回归(principal component regression，PCR)[5]、偏最小二乘(partial least squares，PLS)[6]、支持向量机(support vector machines，SVM)[7]和人工神经网络(artificial neural network，ANN)等。其中MLR，PCR和PLS是线性方法，这类算法容易受到光谱中非线性因素的干扰[8]，而SVM和ANN等非线性方法能克服该问题。

ANN作为一类典型的非线性方法，随着近年来神经网络方法的快速发展，在光谱分析领域中的应用越发广泛。Cui等[9]提出一个卷积神经网络定量分析模型，在3个不同大小的近红外数据集上进行了测试，预测效果优于PLS模型。Wartini Ng等[10]结合可见光、近红外以及中红外光谱，利用卷积神经网络建立土壤多组分同时预测模型，取得了较好的效果。这些神经网络模型的末端通常是由数个全连接层接收前面卷积层提取的特征或原始的样本输入，然后输出预测结果。随着输入维度的增加，全连接层的参数量会变得十分庞大。庞大的参数量使训练时间变长、训练需要的样本量增大、预测效果难以提升。为了降低模型参数量，样本在输入模型之前，通常会用一些降维方法，如主成分分析(principal component analysis，PCA)[11]、小波变换[12]等，或者对输入波长进行选择，如遗传算法(genetic algorithm，GA)[13]、先验知识[14]等，对输入样本进行压缩。这些方法虽已经尽可能地保留样本中的有用信息，但难免还是会丢失一些信息。

针对这一问题，从全连接层的结构入手，提出一种降低全连接层参数量的方法，即分组全连接网络(group fully connected network，GFCN)。该方法直接输入原始的光谱，将原来的全连接层进行分组，用多个局部全连接替代原来的全局全连接。通过在Tecator和IDRC2018数据集上对GFCN进行了验证，并与PLS和FCN算法进行了对比，验证了GFCN的有效性。

收稿日期：2021-06-01，**修订日期**：2021-09-01

基金项目：国家重点研发计划项目(2018YFF01011204)资助

作者简介：余志荣，1996年生，重庆大学大数据与软件学院硕士研究生 e-mail：954466482@qq.com

*通讯作者 e-mail：hmj@cqu.edu.cn

1 分组全连接网络

全连接神经网络中，每层的“神经元”和下层的每个神经元之间全互联，神经元之间不存在同层连接，也不存在跨层连接。下层的每个神经元根据权重接受上层网络中每个神经元传递的特征值，经过偏置与激活函数后输出特征值。经过多层特征提取，最终可以得到预期的结果。全连接层的变换过程可以表示为

$$x_{FC} = A(wx + b) \tag{1}$$

$$A(x) = \mathrm{Relu}(x) = \max(x, 0) \tag{2}$$

其中，w 为全连接层的权重，b 为偏置项，$A(x)$ 为 ReLU 激活函数，x_{FC} 表示全连接层的输出。

然而，全连接层的参数量巨大，且存在冗余。光谱中最主要的特征是吸收峰，吸收峰之间存在间距，波长间隔大的峰不属于同一个吸收峰，所以不需要进行全局的连接来提取特征。结合这一特点，将全连接层内的神经元进行平均分组，同一组内上下层神经元全部互连，不同组神经元之间没有连接，这一结构命名为分组全连接层(group fully connected layer)。分组全连接层的变换过程可以表示为

$$x_{GFC} = A(w_1 x_1 + b_1) \cup \cdots \cup A(w_i x_i + b_i),\ i = 1, \cdots, g \tag{3}$$

$$A(x) = \mathrm{Relu}(x) = \max(x, 0) \tag{4}$$

其中，x_{GFC} 为分组全连接层的输出，w_i 为第 i 组全连接权重，b_i 为第 i 组全连接偏置项，g 为分组的个数，$A(x)$ 为 ReLU 激活函数。

预测模型的结构如图 1 所示，将其命名为分组全连接网络(group fully connected network，GFCN)。GFCN 可以分为两层：第一层是分组全连接层(GFC)，接收原始的光谱输入；第二层是全连接层(FC)，第一层的输出经过激活函数 ReLU 后输入到第二层中，然后输出最终的预测值。

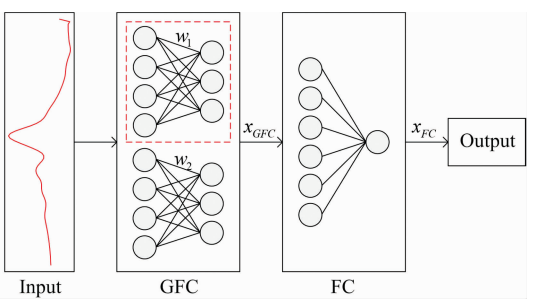

图 1 GFCN 模型结构图

Fig. 1 The structure of GFCN model

GFCN 在 FCN 的基础上将传统的全连接层替换为分组全连接层，具有以下的优点：

(1)参数量更少，计算量降低。全连接网络首层的参数量可能占了模型总参数量的一半以上，而且随着光谱样本波长数的增加，参数数量增长迅速。GFCN 将第一层的全连接层切分成多个小的全连接层，大幅减少了参数量，计算量也大幅减少。

(2)建模需要的样本数量降低。如果可供训练的样本数量不足，神经网络模型可能会过度学习训练集的样本，导致泛化能力不足，在测试集上效果不佳，也就是进入过拟合状态。一般来说，训练模型需要的样本数量与模型参数量呈正相关关系，GFCN 相比 FCN，参数大幅减少，需要的样本量也随之降低，这更加适合光谱分析领域中样本较少的情况。

(3)更方便地分析各个波段对预测的贡献。分组后，每组的神经元学习一种特征，不同组之间互不干扰，避免了多组神经元学习同一特征的情况，可以防止模型过度拟合某一波段[15]。同时，每组神经元的输出即是该波段对预测的贡献，不需要再进行复杂的权重计算，方便后续的模型分析。

2 实验部分

2.1 数据来源

Tecator 数据集包含 215 个切碎的肉样品的近红外光谱及对应的水分、脂肪和蛋白质含量，主要目的是预测脂肪含量。近红外光谱是通过 TecatorInfratec 光谱仪采样得到，波长范围为 850～1 050 nm，采样间隔为 2 nm，共 100 个波长。按照数据集原始的划分方法，将样本分成 172 个训练集样本和 43 个测试集样本。数据集可从 http://lib.stat.cmu.edu/datasets/tecator 下载。

IDRC2018 数据集来源于 IDRC 2018 年线下建模竞赛，由于涉及商业信息，数据集的成分信息未公开。IDRC2018 共包含 200 个样本的近红外光谱及未知成分的含量，波数范围 5 098～9 989 cm^{-1}，采样间隔为 7.71 cm^{-1}，共 635 个波长。按照数据集原始的划分方法，将样本分成 150 个训练集样本和 50 个测试集样本。数据集可从 https://www.cnirs.org/content.aspx?page_id=22&club_id=409746&module_id=276203 下载。

2.2 数据预处理

为了消除光谱中的基线偏移，提高信噪比，所有的光谱样本都经过 Savitzky-Golay 平滑(窗口宽度为 5，多项式次数为 3)和一阶导数处理。处理后的光谱如图 2，后续的建模及预测都使用预处理后的数据。

此外，为了加快模型的收敛速度，用于训练 FCN 和 GFCN 模型的 Tecator 数据集还进行了方差归一化处理。IDRC2018 数据集各特征的方差差异不大，故不进行归一化。

2.3 模型参数

为了方便对比效果，FCN 和GFCN均包括两层隐藏层，两个模型对应层的输出维度均相同。第一层隐藏层输出的长度为输入样本的一半(Tecator 数据集为 50，IDRC2018 数据集为 317)，第二层隐藏层的输出为预测值。经过测试，对于 Tecator 数据集，GFCN 的分组数量 g 设为 5；对于 IDRC2018 数据集，分组数量 g 设为 9。FCN 和 GFCN 在两个数据集上的参数量见表 1，因为 FCN 第一层的参数量占总参数量的 90% 以上，所以 GFCN 的参数量约等于 FCN 的 $1/g$。

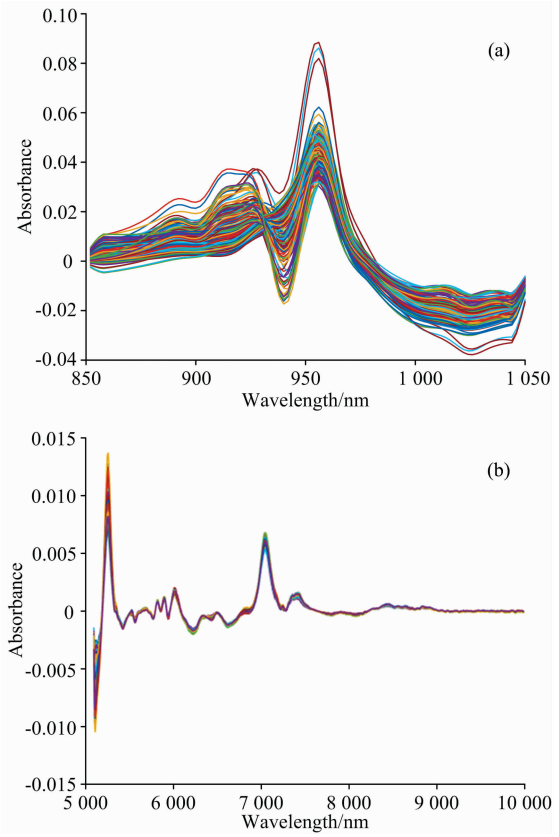

图 2　预处理后的光谱

(a)：Tecator 数据集；(b)：IDRC2018 数据集

Fig. 2　The spectra after preprocess

(a)：Tecator dataset；(b)：IDRC2018 dataset

表 1　FCN 和 GFCN 的参数量

Table 1　Number of parameters for FCN and GFCN

Dataset	Model	Number of groups	Number of parameters
Tecator	FCN	/	5k
	GFCN	5	1k
IDRC2018	FCN	/	202k
	GFCN	9	23k

2.4　模型评价指标

选择的评价指标包括：均方根误差(root mean square error，RMSE)和决定系数(coefficient of determination，R^2)。计算公式如式(5)和式(6)

$$E_{\mathrm{RMSE}}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \tag{5}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{6}$$

其中，$\hat{y}_i$ 为第 i 个样本的预测值，y_i 为第 i 个样本的测量值，$\bar{y}$ 为所有样本的测量值的均值，n 为样本的数量。

3　结果与讨论

3.1　训练次数及分组数量对模型的影响

训练的轮次对神经网络模型的预测效果至关重要，为探究最优的训练轮次，将测试集样本每隔 100 次输入到模型中，模型的预测效果随训练轮次变化如图 3 所示，其中 RMSEC 是训练集的均方根误差，RMSEP 是测试集的均方根误差。可以看到，FCN 在迭代 10 000 次以后，RMSEP 基本不再变化，而 RMSEC 一直下降，出现过拟合现象。GFCN 在 17 000 次左右，RMSEP 达到最低点，随后同样出现过拟合现象。因此，为防止出现过拟合现象，FCN 模型的训练轮次上限定为 10 000 次，GFCN 定为 17 000 次。

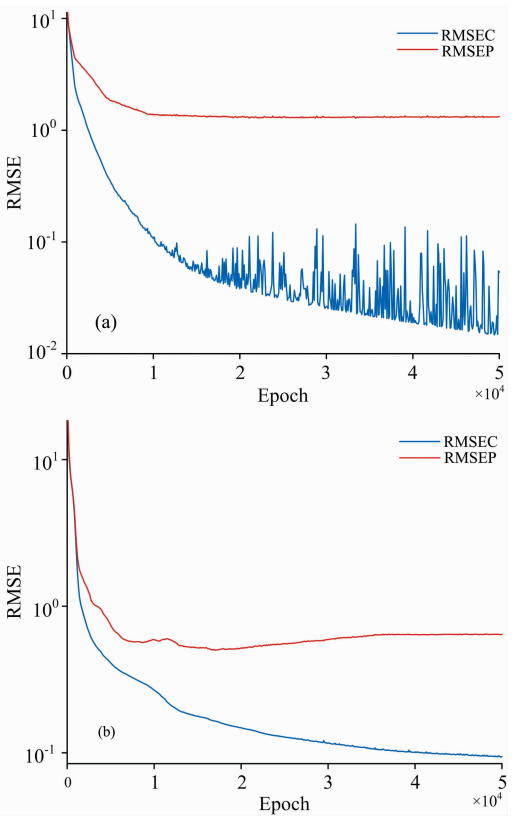

图 3　神经网络模型的训练损失和测试损失

(a)：FCN；(b)：GFCN

Fig. 3　Training loss and test loss of ANN model

(a)：FCN；(b)：GFCN

为探究分组数量 g 对 GFCN 的影响，使用 Tecator 和 IDRC2018 数据集分别进行了不同分组数量的实验，模型的预测效果随分组数量变化曲线如图 4 所示。由于 Tecator 数据集的波长数较少，只验证了分组个数从 1 到 10 的模型性能变化。当分组个数为 1 时，GFCN 等价于 FCN。可以看到，当分组个数分别为 5 和 9 的时候，GFCN 在 Tecator 和 IDRC2018 数据集上的效果最好。因此，在后续的实验中，对于这两个数据集，GFCN 的分组数量分别设为 5 和 9。

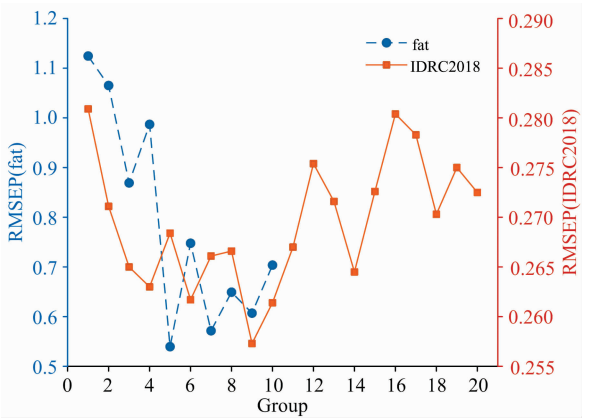

图 4 分组数量对 GFCN 模型预测效果的影响

Fig. 4 Influences of the number of groups on the predicting effect of GFCN model

3.2 不同模型性能比较

为了衡量 GFCN 的预测性能，选择了目前较为常用的偏最小二乘(PLS)模型以及全连接网络(FCN)作为对比。PLS 模型的主成分个数(PC)通过交叉验证方法来选择。三种模型的预测效果如表 2 所示。结果表明，Tecator 数据集的三种成分和 IDRC2018 数据集，GFCN 的预测效果均优于 PLS 和 FCN。

表 2 三种模型的预测效果对比

Table 2 Comparison of predictings effects of three models

Component	Model	RMSEP	R^2	Note
fat	PLS	2.376 4	0.967 2	PC=11
	FCN	1.124 5	0.992 7	
	GFCN	**0.539 6**	**0.998 3**	
moisture	PLS	1.969 6	0.962 2	PC=11
	FCN	1.734 4	0.970 7	
	GFCN	**0.579 3**	**0.996 7**	
protein	PLS	0.571 7	0.961 9	PC=13
	FCN	0.539 3	0.968 4	
	GFCN	**0.477 9**	**0.975 2**	
IDRC2018	PLS	0.275 7	0.912 0	PC=11
	FCN	0.281 2	0.908 5	
	GFCN	**0.257 3**	**0.923 4**	

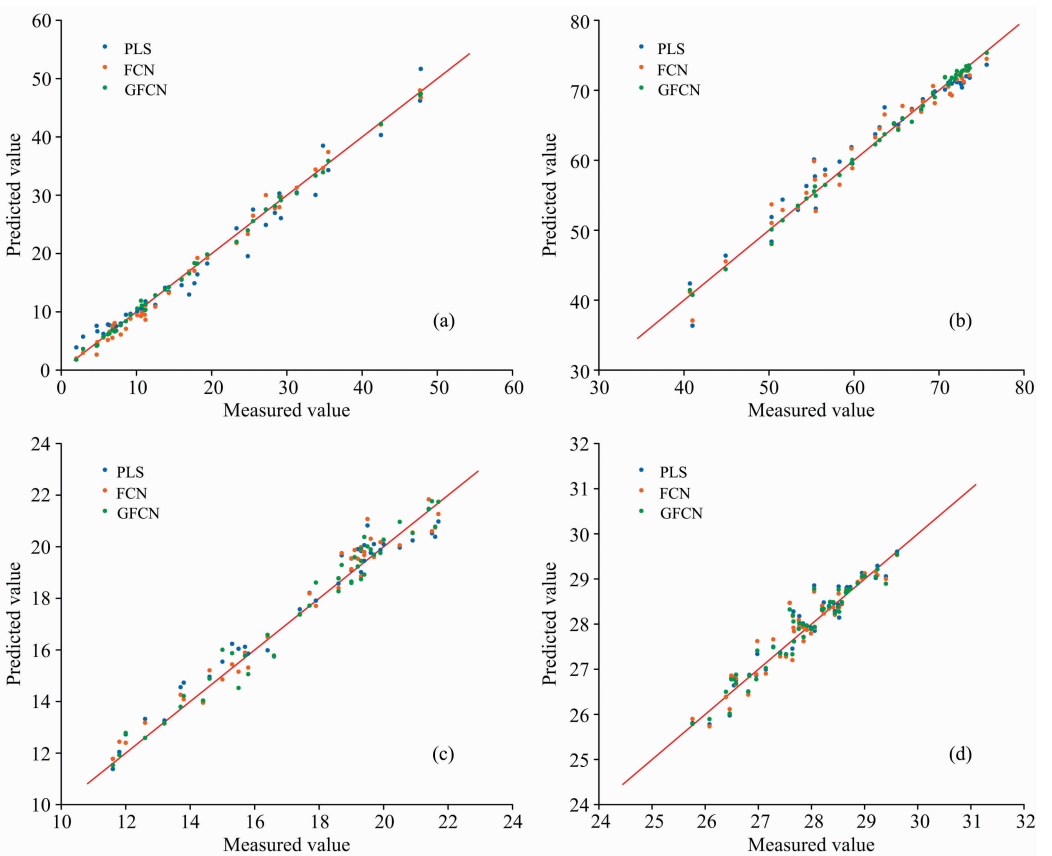

图 5 三种模型的预测结果

(a)：脂肪；(b)：水分；(c)：蛋白质；(d)：IDRC2018

Fig. 5 Prediction results of three models

(a)：Fat；(b)：Moisture；(c)：Protein；(d)：IDRC2018

3.3　训练集大小对模型的影响

为了验证模型在样本数量较少的情况下的建模效果，将 Tecator 数据集的训练集用 Duplex 算法选出一部分来训练模型，数量为原数据集的 80%～20%。测试集的样本不变，占原数据集的 20%。随着训练集的样本数目变化，模型的预测效果如图 6 所示。PLS 模型的总体变化不算太大，预测效果随着样本数目减少有一定的波动。FCN 模型的预测效果随训练样本减少急剧下降，而 GFCN 模型在 80%～50%样本的效果变化较小，随后预测误差才逐渐增大。总的来说，在不同的训练样本数量上，GFCN 模型的效果均优于 FCN 和 PLS 模型，表明，相较于 PLS 和 FCN 模型，GFCN 模型更适用于样本数量较少的场景。

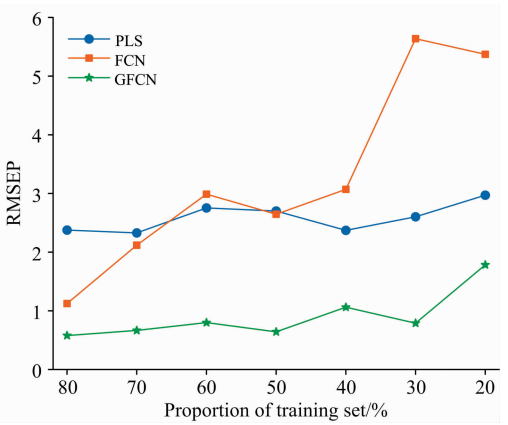

图 6　训练集大小对三种模型预测效果的影响

Fig. 6　The influence of the size of training set on the predictings effect of three models

3.4　模型的解释

为了进一步探索模型工作的原理，计算出每个波长对预测的贡献值：

$$C = \beta \circ s \tag{7}$$

$$\beta = W_{l2} W_{l1} \tag{8}$$

其中，C 为各波长的贡献值；β 为 FCN 和 GFCN 的等效回归系数；s 为输入的光谱样本；$\circ$ 为逐项积运算，即对位元素相乘；W_{l1} 和 W_{l2} 为 FCN 和 GFCN 第一层和第二层的权重，由于激活函数的影响，将网络第一层输出负值的神经元对应的 W_{l2} 权重手动置 0。

由于各样本的预测值存在大小差异，为消除该影响，计算出各波长的贡献值占比

$$R_i = \frac{|C_i|}{\sum |C|}, \ i = 1, \cdots, n \tag{9}$$

式(9)中，R_i 为波长 i 在贡献值绝对值之和的占比；n 为波长点个数。图 7 展示了所有测试样本的各波长贡献占比，实线表示所有样本贡献占比的均值。可以看到，对于 fat 成分，880～925 nm 处的贡献占比大幅度提升，850～880 和 950～1 050 nm 处的贡献降低。这与波长选择方法[16-17]得到的重要波长位置是相吻合的，表明 GFCN 方法的预测是基于与目标成分高度相关的波长点，预测结果比 FCN 更具可解释性。

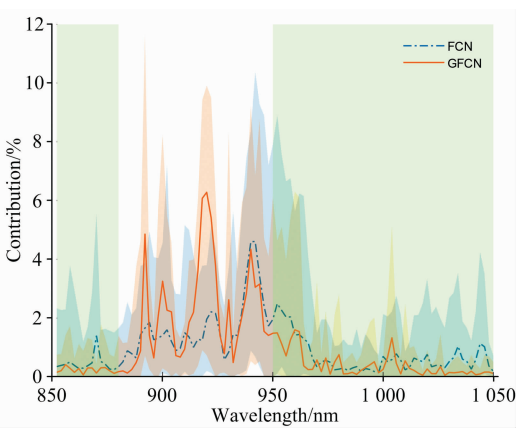

图 7　预测 fat 成分的 FCN 和 GFCN 模型各波长的贡献占比

Fig. 7　Contribution ratio of each wavelength of FCN and GFCN models for predicting fat component

4　结　论

提出了一个基于分组全连接网络的定量分析模型 GFCN。它在 FCN 模型的基础上，将原有的全连接层改为分组全连接层，不仅降低了模型的参数量，同时还提高了预测效果。对 Tecator 和 IDRC2018 数据集使用 PLS，FCN 和 GFCN 分别建立回归模型，GFCN 的预测效果优于 PLS 和 FCN。实验结果表明，GFCN 是一种有效的定量分析方法。

References

[1] PENG Yan-kun, ZHAO Fang, LI Long, et al(彭彦昆，赵　芳，李　龙，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2018, 34(5): 159.

[2] SONG Xue-jian, QIAN Li-li, ZHANG Dong-jie, et al(宋雪健，钱丽丽，张东杰，等). Food Research and Development(食品研究与开发), 2017, 38(12): 197.

[3] Zhu L, Lu S H, Zhang Y H, et al. Vibrational Spectroscopy, 2020, 109: 103071.

[4] Yang Y, Wang X, Zhao X, et al. Analytica Chimica Acta, 2021, 1160: 338453.

[5] Luan X, Huang B, Sedghi S, et al. ISA Transactions, 2018, 81: 46.

[6] Suryakala S V, Prince S. Optical and Quantum Electronics, 2019, 51(8): 1.

[7] Xu S, Zhao Y, Wang M, et al. Geoderma, 2018, 310: 29.

[8] Cook R D, Forzani L. Chemometrics and Intelligent Laboratory Systems, 2021, 213: 104307.

[9] Cui C, Fearn T. Chemometrics and Intelligent Laboratory Systems, 2018, 182: 9.

[10] Ng W, Minasny B, Montazerolghaem M, et al. Geoderma, 2019, 352: 251.
[11] Javadi S H, Munnaf M A, Mouazen A M. Geoderma, 2021, 385: 114851.
[12] LI Mao-gang, YAN Chun-hua, XUE Jia, et al(李茂刚, 闫春华, 薛 佳, 等). Chinese Journal of Analytical Chemistry(分析化学), 2019, 47(12): 1995.
[13] Yang M, Xu D, Chen S, et al. Sensors, 2019, 19(2): 263.
[14] Costa L R, Tonoli G H D, Milagres F R, et al. Carbohydrate Polymers, 2019, 224: 115186.
[15] Shen X, Tian X, Liu T, et al. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(9): 3926.
[16] Rossi F, Lendasse A, François D, et al. Chemometrics and Intelligent Laboratory Systems, 2006, 80(2): 215.
[17] Krier C, François D, Rossi F, et al. Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data. European Symposium on Artificial Neural Networks, 25-27 April, 2007, Bruges, Belgium. 2007: 157.

Near-Infrared Spectral Quantitative Analysis Network Based on Grouped Fully Connection

YU Zhi-rong, HONG Ming-jian*
School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China

Abstract As a typical structure in deep learning, a fully connected network appears in almost all neural network models. In the quantitative analysis of near-infrared spectroscopy, the number of spectral samples is small, but the dimension of each sample is high. It leads to two problems: if the spectrum is directly input into the network, the number of parameters of the network will be very large, which requires more samples to train the model. Otherwise, the model is prone to over fitting; reducing the dimension of the spectrum before inputting it into the network solves the problem that the number of parameters of the network is too large, but it will lose some information and cannot give full play to the learning ability of the network. According to the characteristics of near-infrared spectrum, a group fully connected near-infrared spectrum quantitative analysis network(GFCN) is proposed. Based on the traditional two-layer fully connected network, the network uses several small fully connected layers to replace the first fully connected layer, which overcomes the disadvantage of too many network parameters caused by a direct input spectrum. The GFCN model was tested with Tecator and IDRC2018 datasets and compared with a fully connected network (FCN) and partial least squares (PLS). The results show that the prediction effect of GFCN is better than that of FCN and PLS on the two datasets. In the case of only a small number of samples participating in the modeling, GFCN can still maintain a high prediction effect. The experimental results show that the GFCN can be used for the quantitative analysis of near-infrared spectrum and adapt to the scene with few samples. It indicates that the proposed model has important research value and good application prospects.

Keywords Spectral analysis; Near infrared spectroscopy; Fully connected network; Quantitative analysis

(Received Jun. 1, 2021; accepted Sep. 1, 2021)

* Corresponding author