基于 WGAN 的不均衡太赫兹光谱识别

朱荣盛[1,2]，沈　韬[1,2*]，刘英莉[1,2]，朱　艳[1,2]，崔向伟[1,2]

1. 昆明理工大学信息工程与自动化学院，云南 昆明　650504
2. 昆明理工大学云南省计算机技术应用重点实验室，云南 昆明　650504

摘　要　物质的太赫兹光谱具有唯一性。目前，结合先进的机器学习方法，研究基于规模光谱数据库的太赫兹光谱识别技术已成为太赫兹应用技术领域的重点。考虑到由于实验条件及实验设备的影响，很难收集到多物质均衡光谱数据，而这又是对太赫兹光谱数据进行分类的基础。针对这一问题，提出一种基于 WGAN 的不均衡太赫兹光谱识别方法。WGAN 作为生成数据的一种新方法，将模型达到纳什均衡条件下的生成数据用来补充数据集，使其达到类别均衡。生成数据可以有效映射真实数据分布，通过将生成数据与真实数据混合训练可以提高识别不均衡光谱数据的准确率。采用三种特征谱较为相似的麦芽糖化合物的太赫兹透射光谱数据进行验证，首先利用 S-G 滤波和三次样条插值法对三种物质的光谱数据进行归一化处理，然后通过构建 WGAN 模型对三种物质的不均衡太赫兹光谱数据进行扩展，使其达到类别均衡。实验在同一测试集下进行验证，并利用三组对比实验证明 WGAN 在不均衡数据集处理中的效果。首先利用 WGAN 生成数据，随着迭代次数的增加，生成数据逐渐符合真实数据分布。实验结果证明，使用 WGAN 扩展后的数据集训练 SVM 模型，可以解决模型在测试集上小样本数据(Maltotriose，Malthexaose)偏向大样本数据(Maltoheptaose)的问题。在将 WGAN 与传统处理不均衡数据集方法 FWSVM 和 COPY 对比后发现，三种分类算法在 dataset-1 数据集上的训练集准确率都能达到 90%以上。但是由于模型泛化能力的限制，传统方法在测试集上的效果并不是很理想，而使用 WGAN 后的测试集准确率却能达到 91.54%。在不同不均衡度方面，采用不均衡度为 16，81 和 256 的数据集进行验证，其三个测试集上的准确率分别为 92.08%，91.54%和 90.27%，可满足实际工作中处理不同不均衡度的要求。

关键词　太赫兹光谱；WGAN；不均衡数据；机器学习

中图分类号：O433.5　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)02-0425-05

引　言

太赫兹(Terahertz，THz)波是指频率在 0.1～10 THz 之间的电磁波，在电磁波谱中位于微波和红外辐射之间[1]。近年来，随着太赫兹激发及探测技术的不断发展，目前已有一部分太赫兹产品在实际生活中得到运用，并展现出极高的使用价值及广阔的应用前景[2-3]。由于许多有机分子的振动、转动光谱以及分子间相互作用力落在太赫兹频率波段，可将其作为“指纹谱”实现对物质的定量定性分析[4-6]；同时由于太赫兹所具有的瞬态性、低能性和相干性等特征，使其在光谱识别[7,8]和成像领域[9-10]得到飞速发展。

通过实验获取到的太赫兹光谱数据库存在数据规模不匹配问题，而标准机器学习方法在不均衡数据集中表现不佳，影响太赫兹光谱数据的识别准确率[11]。2014 年，刘进军[12]提出基于惩罚机制的 PFKSVM 方法来克服 K-SVM 在最佳分类表面附近易于分类错误，并使用 UCI 公共数据集进行实验验证其方法在处理不均衡数据集中的优势。2019 年，Tao 等[13]提出了一种过采样技术，该技术使用实值否定选择(RNS)来生成人为的少数类数据，并将生成的少数类数据与多数类组合作为输出。但是，这些方法在太赫兹领域解决数据不均衡问题时并未考虑太赫兹光谱所反映材料的物理和化学性质。针对这一问题，本文提出了一种基于 WGAN 的不均衡太赫兹光谱识别方法来解决太赫兹光谱数据不均衡问

收稿日期：2020-01-15，**修订日期**：2020-04-22

基金项目：国家自然科学基金项目(61971208，61671225)，云南省应用基础研究计划项目重点项目(2018FA034)，昆明理工大学人才培养项目(KKSY201703016)，云南省万人计划青年拔尖人才(沈韬，朱艳，云南省人社厅 No. 2018 73)资助

作者简介：朱荣盛，1994 年生，昆明理工大学硕士研究生　　e-mail：rongsz_715@outlook.com

*通讯作者　　e-mail：shentao@kmust.edu.cn

题。

Wasserstein GAN 是 Arjovsky 等[14]在 2017 年提出的一种改进 GAN 模型的新框架，该方法通过生成器与判别器的相互博弈产生以假乱真的数据，生成数据符合真实数据分布，并且能有效增加数据量。针对目前太赫兹光谱数据库中各物质数据量不均衡问题，本文提出一种基于 WGAN 的不均衡太赫兹光谱识别方法。首先利用生成对抗网络学习真实太赫兹光谱数据分布，在 WGAN 达到纳什均衡后用生成数据扩展太赫兹光谱数据集，使之达到类别均衡，最后采用多分类支持向量机对太赫兹光谱数据进行分类识别。

1 基于 WGAN 的太赫兹光谱识别方法

1.1 基础理论

太赫兹光谱数据为实数值，采用 GAN 训练数据，模型会出现梯度不稳定和多样性不足等问题[14]。针对这些问题，将 Wasserstein 距离作为生成对抗网络的衡量指标，定义如式(1)

$$W(P_r, P_g) = \inf_{\gamma \sim \prod(P_r, P_g)} E_{(x, y)\sim\gamma}[\| x - y \|] \quad (1)$$

式(1)中，P_r 和 P_g 组合起来的所有可能的联合分布的集合为 $\prod(P_r, P_g)$。$\| x-y \|$ 为真实数据 x 和生成数据 y 的距离。联合分布中所有样本距离的期望值为 $E_{(x, y)\sim\gamma}[\| x - y \|]$，Wasserstein 距离就是所有期望值的下界 $\inf_{\gamma \sim \prod(P_r, P_g)} E_{(x, y)\sim\gamma}[\| x - y \|]$。

通过 Kantorovich-Rubinstein 对偶原理可得变换公式

$$W(P_1, P_2) = \sup_{\| f \|_L \leqslant 1} E_{x\sim P_1}[f(x)] - E_{x\sim P_2}[(f(x))] \quad (2)$$

1.2 模型结构

生成对抗网络(generative adversarial network，GAN)是 Goodfellow 等[15]在 2014 年提出的一种概率生成模型，通过对抗过程估计生成模型的新框架。生成对抗网络由两个模型构成，生成模型 G 和判别模型 D，随机噪声 z 通过生成模型 G 生成尽量服从真实数据分布 $p_{data}(x)$的样本 $G(z)$。

判别模型 D 是一个判别式网络，判定接收到的样本是否是来自 $p_{data}(x)$，因此有

$$E_{x\sim p_{data}(x)}[\log(D(x))] \quad (3)$$

其中 E 指代期望，通过根据正类(即判别出 x 属于真实数据 data)的对数函数构建。

生成器 D 通过训练不断提高欺骗判别器的概率，通过根据负类的对数函数构建，即

$$E_{z\sim p_z(z)}[log(1 - D(G(z)))] \quad (4)$$

生成对抗网络的本质是二元零和博弈问题，即通过生成器不断优化生成函数与判别器不断优化判别网络来达到最优状态，即

$$\min_G \max_D V(D, G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (5)$$

生成对抗网络给出了一种生成数据的新形式，即可通过对抗性学习模拟真实数据分布。而物质的太赫兹光谱数据为实数值，将 JS 散度作为衡量值并不能很好的评估距离，因此通过使用 Wasserstein 距离来衡量生成部分和真实数据分布之间的距离，解决了生成对抗网络在生成太赫兹光谱数据时训练过程不稳定，模型优化困难等问题。

2 实验部分

实验以麦芽三糖(Maltotriose)、麦芽六糖(Malthexaose)和麦芽七糖(Maltoheptaose)在 0.9～6 THz 内的太赫兹透射光谱为例。首先通过 S-G 滤波对光谱数据进行滤波处理，然后通过三次样条插值获得相同的数据点。随机选择三种物质预处理后的各一条太赫兹光谱数据曲线，如图 1 所示。

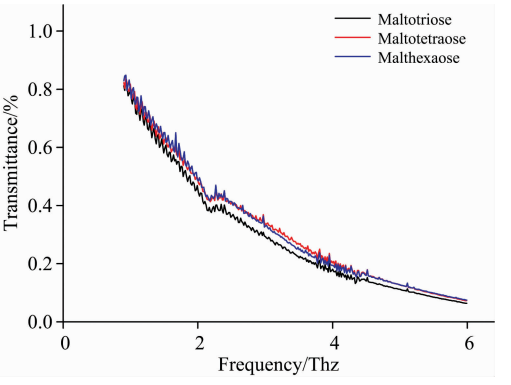

图 1 三种物质的太赫兹光谱

Fig. 1 Terahertz spectra of three substances

为了验证该方法的有效性，我们首先使用 WGAN 生成数据，将物质的光谱数据输入到 WGAN 模型中。其次，生成模型 G 根据输入数据的维度输出与测试数据相同维度的随机数。最后，判别模型 D 判别接收到的数据是否为太赫兹频谱数据。当判别模型 D 无法识别接收到的数据是真实数据还是生成数据时，该模型达到纳什均衡。以 Maltotriose 为例，根据真实太赫兹光谱数据生成数据。在实验设置中，设置最大迭代次数 300 000 次，每迭代 1 000 次模型保存一次数据。随机选取 5 种不同迭代次数图，如图 2 所示。当迭代次数为 1 000 轮和 5 000 轮时，生成的数据仅为随机噪声。随着迭代次数的增加，生成器不断学习。当模型迭代次数达到 100 000 轮时，生成数据逐渐类似于真实数据分布，当达到 200 000 轮时，WGAN 模型所输出的生成数据分布基本符合真实 Maltotriose 数据分布。在对 Maltotriose 进行扩展数据时，选取迭代 200 000 轮后的生成数据。

为了验证 WGAN 处理不均衡数据集的效果，将三种不均衡物质的数据组成数据集 Database1，经 WGAN 扩展后的均衡数据集为 Database2。数据集中各物质光谱数据如下：(1)Database1：在数据库中随机抽 100 条 Maltotriose 数据、900 条 Malthexaose 数据和 8100 条 Maltoheptaose 数据。(2)Database2：使用 WGAN 生成的数据将 Database1 中每种物质的数据补充为8 100条。在数据库中随机抽取每种物质 2 700条数据作为测试集。

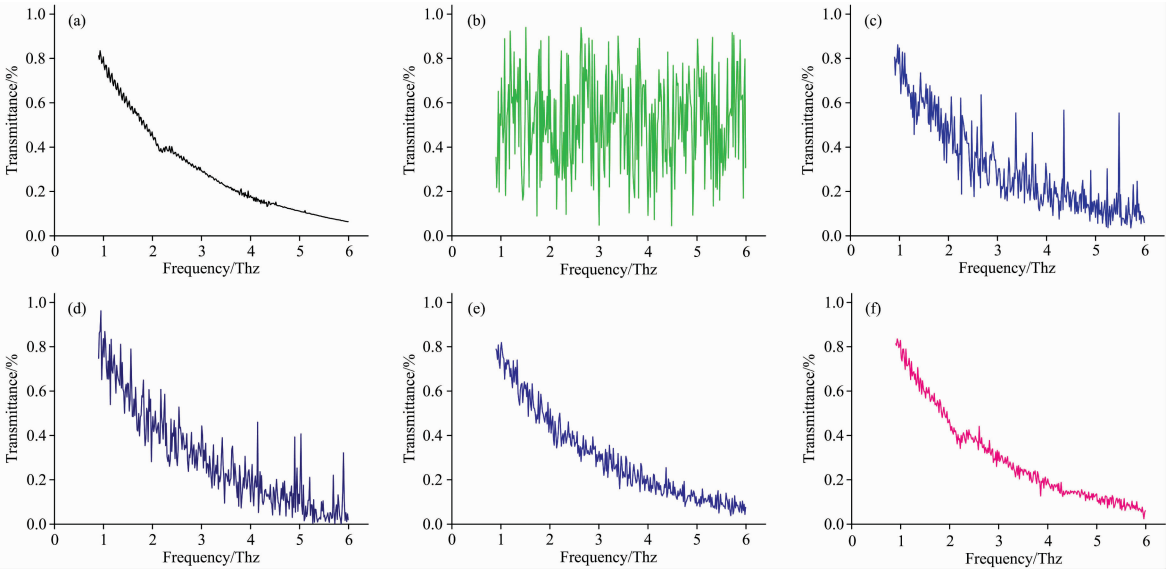

图 2　不同迭代次数下 WGAN 的生成数据图

(a)：原始数据；(b)，(c)，(d)，(e)，(f)分别代表迭代 1 000 轮，5 000 轮、10 000 轮、100 000 轮和 200 000 轮后的生成数据

Fig. 2　WGAN generated data graphs under different iterations

(a) is the original data; (b), (c), (d), (e), and (f) respectively represent the generated data after 1 000 iterations, 5 000 rounds, 10 000 rounds, 100 000 rounds, and 200 000 rounds

3　结果与讨论

数据集不均衡会对传统的机器学习模型系统产生负面影响。为了缓解此问题，将 WGAN 用于生成太赫兹光谱数据，以便使太赫兹光谱数据集达到类别均衡。实验证明，使用 WGAN 生成数据并扩展数据集，能够有效解决小样本数据偏向大样本数据问题。表 1 和表 2 分别为 SVM 模型在 Dataset1 和 Dataset2 数据集下训练后测试集的混淆矩阵。

表 1　使用 Database1 训练模型后测试集的混淆矩阵

Table 1　Confusion matrix of test database after training model with Database1

Original	Predicted		
	Maltotriose	Malthexaose	Maltoheptaose
Maltotriose	251	9	2 440
Malthexaose	4	2 373	323
Maltoheptaose	3	0	2 697

从表 1 可以看出，Maltotriose 和 Malthexaose 都出现被预测为 Maltoheptaose 的现象，其中 Maltotriose 最为明显。但是没有大量 Maltoheptaose 被预测为其他两种数据的现象。

表 2 相比于表 1，在数据预测偏向上得到改善，每种数据的偏向现象并不明显，其中，Maltotriose 和 Malthexaose 并没有大规模偏向 Maltoheptaose。根据表 1，使用 Dataset1 进行 SVM 训练的模型测试集的预测准确性仅为 65.69%。但是，当使用 Database2 训练 SVM 时，模型精度提高到 91.54%，均衡数据集上 SVM 的识别准确率比不均衡数据集提高 25.85%。为了证明 WGAN 在处理不均衡太赫兹光谱数据上的优越性，将 WGAN 与其他处理不均衡数据集的方法进行了比较，并以验证集的准确性作为度量。表 3 为不同不均衡数据集处理方法的准确率对比。

表 2　使用 Database2 训练模型后测试集的混淆矩阵

Table 2　Confusion matrix of test database after training model with Database2

Original	Predicted		
	Maltotriose	Malthexaose	Maltoheptaose
Maltotriose	2 355	13	332
Malthexaose	7	2 689	4
Maltoheptaose	7	322	2 371

表 3　不同算法下数据集的准确性对比

Table 3　Comparison of the accuracy of the dataset under different algorithms

	train/%	validation/%	test/%
SVM	91.63	91.57	65.69
SVM-COPY	89.11	89.54	84.04
FWSVM	83.68	82.23	85.77
SVM-WGAN	92.45	91.13	91.54

由表 3 可知，4 种分类算法在 dataset-1 数据集上的训练集及验证集的准确率都能达到 80%以上。虽然未采用扩展数据的 SVM 模型能在训练集和验证集上得到良好的识别准确率，但是在测试集上由于不均衡数据固有的缺点，导致识别准确率很差。SVM-COPY 和 FWSVM 的测试集准确率都在 85%左右，这两种方式是现阶段比较流行的处理不均衡数据

集的方法，但是由于并没有在数据集中增加有效的太赫兹光谱数据，所以测试集上的识别效果不是太理想。因此，利用WGAN模型能够有效的生成太赫兹光谱数据，同时又能保证模型识别准确率。

表 4 不同不平衡度下训练集和测试集的准确率对比

Table 4 Compares the accuracy of the training set and test set of the dataset under different unbalance

	train/%	validation/%	test/%
Imbalance1	94.22	92.06	84.28
Imbalance1_WGAN	94.13	91.25	92.08
Imbalance2	91.63	91.58	65.69
Imbalance2_WGAN	92.45	91.14	91.54
Imbalance3	94.07	93.92	57.70
Imbalance3_WGAN	92.21	91.87	90.27

不均衡度也是影响不均衡数据分类识别准确率的因素之一，为了验证WGAN在不同不均衡度下的有效性，将不均衡度为16，81和256的数据集分别组成Imbalance1，Imbalance2和Imbalance3数据集，通过WGAN扩展后的数据集为Imbalance1_WGAN，Imbalance2_WGAN和Imbalance3_WGAN数据集。实验结果表明，不均衡度对测试集影响较大，随着不均衡度的增加，测试集整体识别率呈现下降趋势。通过使用WGAN扩展数据集后，可以有效改善这一现象。表4为不同不均衡度下的识别率对比。

4 结 论

针对太赫兹光谱数据库中不均衡数据的分类问题，提出一种基于WGAN的太赫兹光谱识别方法。利用生成对抗网络生成符合真实太赫兹光谱数据分布的生成数据，扩充太赫兹数据集，解决类别不均衡问题。相比于传统方法，该方法能自动从真实数据中学习数据分布并生成数据。不仅能有效扩充太赫兹光谱数据库，并且有较高的识别率。由于基于生成对抗网络的太赫兹光谱识别方法可与多种机器学习方法相结合，并能适应不同不均衡度的要求，所以在未来实际应用中有广阔的前景。

References

[1] Tonouchi M. Nature Photonics, 2007, 1(2): 97.
[2] Jepsen P U, Cooke D G, Koch M. Laser & Photonics Reviews, 2011, 5(1): 124.
[3] Liebermeister L, Nellen S, Kohlhaas R, et al. Journal of Infrared, Millimeter, and Terahertz Waves, 2019, 40(3): 288.
[4] Li Y, Xu L, Zhou Q, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 214: 246.
[5] Strachan C J, Taday P F, Newnham D A, et al. Journal of Pharmaceutical Sciences, 2005, 94(4): 837.
[6] Nishimura F, Hoshina H, Ozaki Y, et al. Polymer Journal, 2019, 51(2): 237.
[7] Fischer B M, Helm H, Jepsen P U. Proceedings of the IEEE, 2007, 95(8): 1592.
[8] Liu P, Zhang X, Pan B, et al. International Journal of Environmental Research, 2019, 13(1): 143.
[9] Mittleman D M. Optics Express, 2018, 26(8): 9417.
[10] Yang X, Pi Y, Liu T, et al. IEEE Sensors Journal, 2018, 18(3): 1063.
[11] He H, Garcia E A. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263.
[12] LIU Jin-jun(刘进军). Computer Applications and Software(计算机应用与软件), 2014, 31(1): 186.
[13] Tao X, Li Q, Ren C, et al. Expert Systems with Applications, 2019, 129: 118.
[14] Arjovsky M, Chintala S, Bottou L. arXiv Preprint arXiv, 2017, 1701: 07875.
[15] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Advances in Neural Information Processing Systems, 2014, 27: 2672.

Wasserstein GAN for the Classification of Unbalanced THz Database

ZHU Rong-sheng[1,2], SHEN Tao[1,2*], LIU Ying-li[1,2], ZHU Yan[1,2], CUI Xiang-wei[1,2]

1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China
2. Computer Technology Application Key Lab of Yunnan Province, Kunming University of Science and Technology, Kunming 650504, China

Abstract The terahertz spectrum of the matter is unique. At present, combined with advanced machine learning methods, research on terahertz spectrum recognition technology based on large-scale spectral databases has become the focus of terahertz application technology. It is difficult to collect multi-material equilibrium spectral data, which is the basis for classifying terahertz spectral data. This paper proposes an unbalanced terahertz spectrum recognition method based on WGAN (Wasserstein Generative Adversarial Networks). As a new method of generating data, WGAN uses the generated data under the condition that the model reaches the Nash equilibrium to supplement the data set, and is finally trained by a support vector machine (SVM). The experimental results prove that the generated data can effectively map the distribution of real data, and the accuracy of identifying unbalanced spectral data can be improved by mixing the generated data with the real data. In this paper, three types of maltose compounds with similar characteristics spectra are used for verification. We first use S-G filtering and cubic spline interpolation to normalize the spectral data of the three substances, and then expand the unbalanced terahertz spectral data of the three substances by constructing a WGAN model to bring it to class equilibrium. The experiments are verified under the same test set, and three sets of comparative experiments are used to prove the effectiveness of WGAN in the processing of uneven data sets. First we use WGAN to generate data. As the number of iterations increases, the generated data gradually conforms to the real data distribution. When the model reaches the Nash equilibrium, the generated data basically conforms to the original data distribution. The experimental results prove that training the SVM model using the extended WGAN data set can solve the problem that the model has a small sample data (Maltotriose, Malthexaose) biased toward a large sample data (Maltoheptaose) on the test set. After comparing WGAN with traditional methods for processing unbalanced data sets FWSVM and COPY, we find that the training set accuracy of the three classification algorithms on the dataset-1 dataset can reach more than 90%. However, due to the limitation of the generalization ability of the model, the effect of the traditional method on the test set is not very satisfactory, and the accuracy of the test set after using WGAN can reach 91.54%. In terms of different imbalances, the data sets with imbalances of 16, 81, and 256 were used for verification. The accuracy rates on the three test sets are 92.08%, 91.54%, and 90.27%, which can meet the requirements of dealing with different imbalances in actual work.

Keywords Terahertz spectrum; Wasserstein GAN; Unbalanced database; Machine learning

(Received Jan. 15, 2020; accepted Apr. 22, 2020)

* Corresponding author