CHEMICAL ENGINEERING TRANSACTIONS

VOL. 81, 2020

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: Petar S. Varbanov, Qiuwang Wang, Min Zeng, Panos Seferlis, Ting Ma, Jiří J. Klemeš
Copyright © 2020, AIDIC Servizi S.r.l.
ISBN 978-88-95608-79-2; **ISSN** 2283-9216

DOI: 10.3303/CET2081152

Application of Generative Adversarial Network for the Prediction of Gasoline Properties

Kaixun He[a,*], Jingjing Liu[a], Zhi Li[b]

[a]College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590 China
[b]Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237 China
kaixunhe@sdust.edu.cn

Near-infrared (NIR) spectroscopy has been widely used to predict the gasoline properties that are difficult to measure online during gasoline blending. NIR models should be prepared in advance to apply this technique successfully. Obtaining a high-accuracy NIR model in practice is hard because abundant labelled samples are difficult to acquire. A new modelling method on the basis of Wasserstein generative adversarial network is proposed in this study to overcome this weakness. Abundant artificial labelled samples are generated firstly using the proposed method, and sample selection is performed to select the appropriate artificial samples. Real and selected artificial samples from the selection results are combined to train the NIR model that could be established efficiently when labelled samples are scarce. An actual dataset obtained during gasoline blending is provided to validate the effectiveness of the proposed method, and several traditional methods are adopted for comparison.

1. Introduction

Gasoline is the main power fuel of vehicle, which plays an important role in today's global economy. The production of high-grade clean gasoline fuel has received widespread attention because of the increasingly strict environmental protection requirements imposed by many governments worldwide (Klemeš et al., 2019). Gasoline that is directly produced by the fractionation and distillation of crude oil has a low octane number, and its yield fails to meet market demands (Li et al., 2010). To cope with this issue, in refinery, low-octane component oils, such as catalytic gasoline, nonaromatic gasoline and naphtha, are blended with high-octane components in certain proportions to produce products that satisfy quality specifications. This step is gasoline blending, which is the last operation before the delivery of gasoline products. Generally, blending is controlled by an optimisation controller to save petroleum resources and produce high-quality clean gasoline. This controller provides the optimal blending recipe and adjusts the proportion of component oils in real time (He et al., 2017). The essence of blending optimisation is quality feedback control. Its implementation relies heavily on the real-time analysis of the properties of blended gasoline. Given that the octane number indicates gasoline quality, the rapid detection of this property during gasoline blending is necessary. In the past 2 decades, this task has been mainly realised by using near-infrared (NIR) spectroscopic methods combined with appropriate quantitative analysis models (NIR models) (Mabood et al., 2017). Multivariate statistical methods, such as principal component regression, partial least squares regression (PLS), independent component analysis and their extended versions, have been widely used to establish NIR models. Wang et al. (2016) explained the application of these common multivariate statistical methods in NIR analysis. Many machine learning methods, such as support vector machine regression (SVR) and artificial neural networks, have been successfully used to build NIR models (Balabin, et al., 2007). A sufficient number of labelled samples are necessary to develop an excellent NIR model. Standard laboratory testing fails to provide an adequate number of samples in a short time because of the complicated instrumentation and long analysis time required (Bohács et al., 1998). Studies performed to address the lack of labelled samples are limited despite the wealth of research on modelling algorithms (Li and Chu, 2018). Recently, the generative

Paper Received: 25/02/2020; Revised: 21/04/2020; Accepted: 24/04/2020
Please cite this article as: He K., Liu J., Li Z., 2020, Application of Generative Adversarial Network for the Prediction of Gasoline Properties, Chemical Engineering Transactions, 81, 907-912 DOI:10.3303/CET2081152

adversarial network (GAN) has been widely used to generate labelled samples to address the lack of data. For examples, Wang et al. (2019) used GAN to generate samples for minority fault class to cope with imbalanced fault diagnosis. Li et al. (2019) established a deep GAN model to generate cross-domain samples, which achieves good performance in fault diagnosis of rolling element bearings. Most of the recent studies focused on the application in fault detection, and those using GAN in soft sensor were limited. In the present study, we adopt the Wasserstein generative adversarial network (WGAN) to produce fake labelled samples for NIR models. Selection is performed to eliminate inappropriate generated samples. The remaining samples and the real ones constitute an initial training dataset. During the online application, the just-in-time learning (JIT) strategy is used to select local modelling data from the initial dataset for each query sample x_q. Industry data from a real-world gasoline blending process are provided to validate the performance of the proposed method, and several traditional modelling methods are used for comparison.

The rest of this paper is organised as follows: The current problem in NIR modelling is described in Section 2, the basic idea of WGAN and the proposed method are also elaborated in this section. The research results and discussion are presented in Section 3. Finally, conclusions are provided in Section 4.

2. Theory and algorithm

In this section, the motivation of this work is presented, and the WGAN model is illustrated. The details of the proposed modelling method are also presented.

2.1 Problem statement

Currently, the online analysis of the key properties (research octane number [RON] and motor octane number) of gasoline mainly depends on a near-infrared analyser. The basis and premise of using this technology are to establish a NIR model. The NIR model with superior predictive performance and stability is urgently needed. Similar to other data-driven models, a sufficient number of labelled samples are needed to achieve these goals. Obtaining an adequate amount of labelled data in a short time, specifically for new blending batches, is difficult, which is the key limitation in using NIR techniques and optimising blending control in practice (He et al., 2020). Proposing effective approach to generate new labelled samples for NIR modelling is necessary. Recently, GAN has been applied in various data supplementation tasks and achieved good application results. Inspired by this, this work intends to develop an effective data generation model that can produce labelled NIR samples for data augmentation. This study has significance in improving the efficiency of establishing and maintaining NIR models.

2.2 Wasserstein GAN

Goodfellow et al. (2014) initially proposed the GAN which has been widely used in the field of image recognition. GAN includes a generation network (G) and a discrimination network (D). During training, the G network generates fake samples via the multilevel mapping of the noise variable $x \sim P_z(z)$, the D network attempts to distinguish generated data $x \sim P_g$ from actual ones $x \sim P_r$. The generator G and discriminator D are trained to compete with each other alternatively until Nash equilibrium is reached. This process can be described as a min–max two-player game with the following objective function:

$$\min_G \max_D V(D,G) = E_{x\sim P_r}[\log D(x)] + E_{y\sim \mathrm{P}_g}[\log(1-D(G(y)))] \tag{1}$$

where P_r and P_g are the real data distribution and the generated data distribution.

The original GAN is unstable because it could produce an unstable gradient when training the G network. To cope with the gradient vanish problem, the Wasserstein distance (as shown in Eq. (2)) is proposed and used to describe the minimum cost to converge the model distribution P_g to the real distribution P_r.

$$W(P_r,P_g) = \inf_{\gamma\in\Pi(P_r,P_g)} E_{(x,y)\sim[\|x-y\|]} \tag{2}$$

where, $\Pi(P_r,P_g)$ is the set of all joint distributions $\gamma(x,y)$ whose marginals are P_r and P_g. On the basis of the Kantorovich Rubinstein duality (Villani, 2008), WGAN loss can be constructed as follows:

$$\min_G \max_{D\in¡} E_{x\sim P_r}[D(x) - E_{y\sim P_g}[D(y)]] \tag{3}$$

where ¡ is the set of 1-Lipschitz functions. In WGAN, weight clipping is used to enforce the weights of discrimination into a compact space [−*c*, *c*]. The detailed algorithm of WGAN can be found in Arjovsky et al. (2017) and is not presented in this study due to space limitations.

2.3 Proposed method

During gasoline blending, NIR spectra can be obtained using an online NIR analyser, the properties of gasoline can only be tested by CFR octane testing instrument. Obtaining a sufficient number of labelled samples in a short time is difficult. A novel modelling method is proposed in the present work to cope with this issue. The WGAN is used to produce new labelled samples, and sample selection is performed to select appropriate data to construct a training dataset. When the training dataset is determined, JIT is used to select local modelling samples for each query sample x_q, and the PLS algorithm is used to establish a local NIR model. The details of our proposed method are summarised as follows:

Step 1: Collect a number of labelled samples from target blending conditions, including NIR spectra x and the corresponding properties y , denoted as $D_r(x,y)=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$,

Step 2: Normalise the dataset $D_r(x,y)$ and divide $D_r(x,y)$ into *K* subsets $D_{r,k}(x,y)$, $k=1,2,...,K$ randomly,

Step 3: Establish *K* WGAN models and train these WGAN models with $D_{r,k}(x,y)$. After training, *K* generated datasets could be obtained, and these subsets are combined into a dataset $D_g(x,y)=\{(x_1,y_1),...,(x_N,y_N)\}$. Then, a new training set $D_s(x,y)$ is synthesised from the generated set $D_g(x,y)$ and the real set $D_r(x,y)$,

Step 4: Establish a PCA model using $D_r(x,y)$, calculate the T^2 statistics of each sample in the dataset $D_s(x,y)$ and screen training samples on the basis of the threshold of the T^2 statistic. The final selected modelling samples form a training set denoted as $D_T(x,y)$,

Step 5: During the online application, the JIT strategy is adopted to select local modelling samples from $D_s(x,y)$ in accordance with Eqs.(4–5),

$$\omega_i = \exp(-\frac{d_i^2}{\sigma_d^2}) \tag{4}$$

$$d_i = \sqrt{(x_q - x_i)^{\mathrm{T}}(x_q - x_i)} \tag{5}$$

where x_i is the *i*th sample in the training dataset $D_s(x,y)$, x_q is the query sample and σ_d is the localisation parameter.

For each x_q , the weight ω_i of each sample in $D_s(x,y)$ is calculated, and all the samples in $D_s(x,y)$ are sorted in descending order of weight value. Then, the first *Kn* samples are selected to form a local training set $D_l(x,y)$.

In this work, the maximum absolute error (MAE), the root-mean-square error (RMSE) and the coefficient of determination R^2 are utilised to assess the performance of our proposed method. The model with the lowest RMSE and MAE and the highest R^2 is considered the best model.

$$MAE = \max\{|y_i - \hat{y}_i|\},\ i=1,2,3,...,n \tag{6}$$

$$RMSE = \sqrt{\sum_{i=1}^{n}(\hat{y}_i - y)^2 \Big/ (n-1)},\ i=1,2,3,...,n \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},\ i=1,2,3,...,n \tag{8}$$

where n is the number of samples included in the test set and y_i and $\hat{y}_i$ are the RON measured by the CFR octane testing instrument and NIR methods.

3. Experimental and results

In this section, industrial samples collected during online gasoline blending are used to validate the performance of the proposed method. PLS, SVR, JIT-PLS and JIT-SVR are adopted to establish NIR models using $D_r(x,y)$ and $D_s(x,y)$, and all experiments are completed by using MATLAB. The details of all the mentioned modelling algorithms are as follows:

(1) PLS: A classic PLS algorithm is adopted to build the NIR model. The number of latent variables is determined as 6 by five-fold cross-validation (CV). PLS theory can be found in Geladi and Kowalski (1986).

(2) SVR: SVR is implemented using the libsvm toolbox (Chang et al., 2011). In this study, the linear function $k(x_i,x_j)=x_i^T x_j$ was utilised as a kernel function, and the penalty parameter of SVR was optimised as 0.1 through five-fold CV.

(3) JIT-PLS and JIT-SVR: Local modelling samples are selected from $D_s(x,y)$ for each query sample in accordance with Eqs(4–5). Then, PLS and SVR are used to establish the NIR model. The number of local modelling samples is selected as 35 empirically.

(4) PLS-GAN, SVR-GAN, JIT-PLS-GAN and JIT-SVR-GAN: These symbols denote the models established by PLS, SVR, JIT-PLS and JIT-SVR via the synthetic dataset $D_s(x,y)$.

3.1 Results and discussion

In this case, 240 standard samples of 93 octane gasoline were collected during gasoline blending. Among these samples, 40 were selected to form the dataset $D_r(x,y)$ artificially. The remaining 200 samples were adopted as a test set. The dimension of the gasoline NIR spectrum was 201, and the wavelength range was restricted to 1,100–1,300 nm. In addition, the corresponding RON of NIR samples were obtained through standard analytical methods.

The parameters of WGAN were set as follows: The maximum training epoch of WGAN was 1,000. The number of discriminator iterations for each generation iteration was 10. G and D were constructed by multilayer perceptron neural networks, and the structures were 9-9-202 and 9-9-1. The minibatch sizes of the stochastic gradient descent method, the learning rates of G and D and the gradient penalty coefficient were 15, 0.01, 0.02 and 1. The dataset $D_r(x,y)$ was not divided into *K* subsets in this case because only 40 samples were available in $D_r(x,y)$. WGAN was trained *K* (*K* = *10*) times independently with all the data in $D_r(x,y)$, and 50 samples were generated each time, totalling to 500 fake samples. In accordance with Step 4 as described in Section 2.3, sample selection on the basis of PCA was performed.

Table 1: Performance comparison in terms of RMSE, R2 and MAE

	Real training dataset			Synthetic training dataset		
Method	RMSE	R^2	MAE	RMSE	R^2	MAE
PLS	0.394	0.844	2.136	0.270	0.927	0.995
JIT-PLS	0.257	0.933	0.937	0.247	0.938	0.849
SVR	0.327	0.892	1.538	0.282	0.920	0.892
JIT-SVR	0.270	0.926	1.179	0.261	0.931	0.871

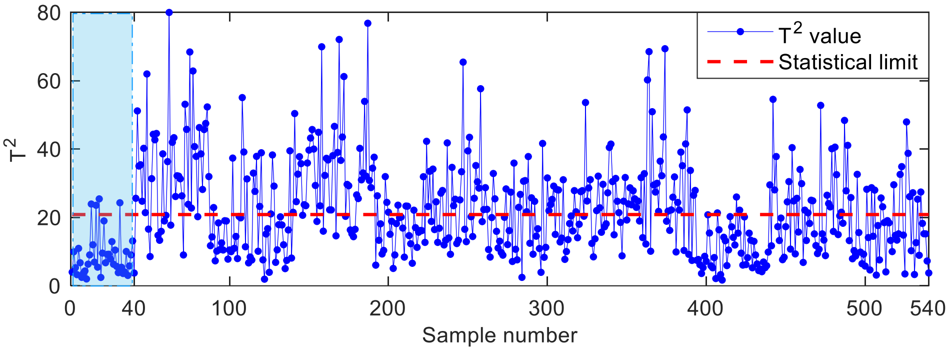

Figure 1: Sample selection results (the shaded part represents the real samples)

The number of principal components was determined to be 9 given that the cumulative variance contribution exceeded 0.99. The threshold value of the T^2 statistic was 20.8523 in accordance with the confidence degree of 0.9. Using the T^2 statistic, 288 samples were selected from $D_s(x,y)$. Figure 1 shows the selection results.

Table 1 shows the performance of different models in terms of R^2, RMSE and MAE. The performances of all the methods improved with the synthetic training dataset. The results also indicated that the supplementary generated samples helped improve the generalisation of NIR models. The JIT method did not use all generated samples but only selected the most suitable ones when used for modelling. The selected local training samples are capable of representing the current working conditions. JIT-based methods (i.e. JIT-PLS, JIT-PLS-GAN, JIT-SVR and JIT-SVR-GAN) perform better than the static methods (i.e. PLS, PLS-GAN, SVR and SVR-GAN) as evidenced by their MAE.

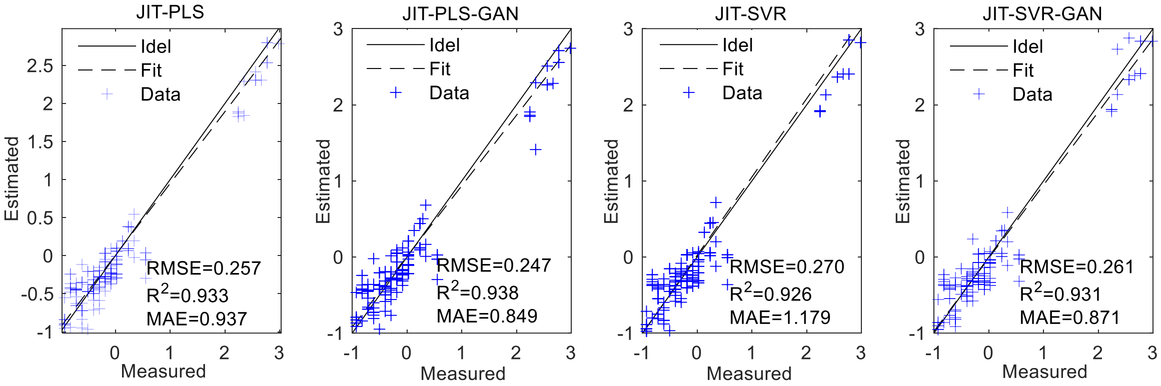

Figure 2: Schematic diagram of JIT-PLS, JIT-PLS-GAN, JIT-SVR and JIT-SVR-GAN

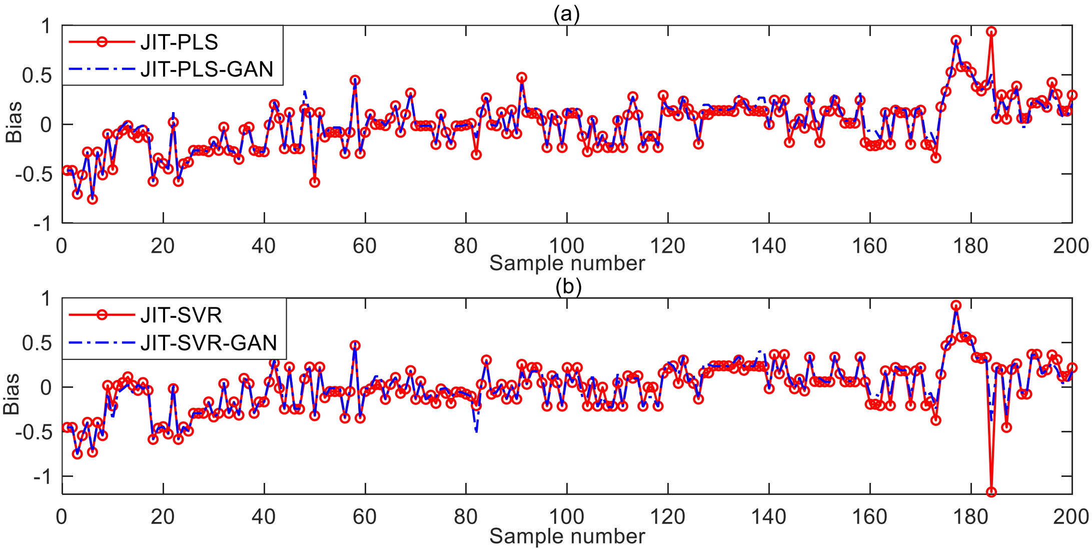

Figure 3: Prediction error curve of RON by JIT-PLS, JIT-PLS-GAN, JIT-SVR and JIT-SVR-GAN

Figure 2 shows the regression curves of the four JIT-based methods. The fitted curves of JIT-based methods are closer to the ideal curve in the area (−1, 1), where data were highly concentrated, and the deviation was larger in the area (1, −3). Given this characteristic, the model will prioritise prediction accuracy within the range of normal working conditions to meet the needs of industrial production. As presented in Table 1 and Figure 3, JIT-PLS-GAN performed slightly better than JIT-SVR-GAN mainly because SVR has numerous model parameters that are difficult to tune online. The prediction model should be established online when the JIT strategy is adopted. We had to select a modelling algorithm with low computational complexity. In this scenario, PLS was more suitable than SVR because PLS has few parameters and fast modelling speed. In the prediction of octane number, PLS is the preferred method for online applications. To summarise, the prediction error, R^2, RMSE, MAE and other details indicate that the performance of NIR models could be improved by using fake labelled samples generated by the proposed method. In addition, the JIT strategy combined with PLS is highly suitable for building NIR models.

4. Conclusions

This study presents the application of WGAN for the prediction of octane number by NIR spectroscopy. In this methodology, WGAN together with a sample selection method was used to generate fake training data to construct a new modelling set, and PLS, SVR, JIT-PLS and JIT-SVR were used to build NIR models on the basis of real and synthetic modelling sets. When labelled data are insufficient during gasoline blending, the proposed method can help to establish an initial NIR model quickly. The statistical results show that the performance of NIR models could be improved through combination with generated data. The effectiveness of our method would be affected by the variation in online data. In future studies, we will focus on exploiting a self-adaptive generation method to cope with dynamic operating characteristics.

Acknowledgements

We would like to acknowledge financial support for this work from the National Natural Science Foundation of China (61803234, 61873149, and 61751307), the Shandong Provincial Natural Science Foundation of China (ZR2017BF026), China Postdoctoral Science Foundation (2018M632691) and the Fundamental Research Funds for the Central Universities.

References

Arjovsky M., Chintala S., Bottou L., 2017, Wasserstein GAN, arXiv, 1701.07875.

Balabin R.M., Safieva R.Z., Lomakina E.I., 2007, Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction, Chemometrics and Intelligent Laboratory Systems,88(2), 183 -188.

Bohács G., Ovádi Z., Salgó A., 1998, Prediction of gasoline properties with near infrared spectroscopy, Journal of Near Infrared Spectroscopy, 6(1), 341-348.

Chang C., Lin C., 2011, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2(3), 27, DOI: 10.1145/1961189.1961199.

Geladi P., Kowalski B.R., 1986, Partial least-squares regression: a tutorial, Analytica Chimica Acta, 185, 1−17.

Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, Generative adversarial nets, Advances in Neural Information Processing Systems, 27, 2672-2680.

He K., Li Y., Wang K., 2017, A novel training sample selection approach for near-Infrared spectroscopy model and its industrial application, Chemical Engineering Transactions, 61, 1429-1434.

He K., Zhong M., Fang J., Li Y., 2020, Biased Minimax Probability Machine-Based Adaptive Regression for Online Analysis of Gasoline Property, 16(4), 2799-2808.

Klemeš J.J., Varbanov P.S., Walmsley T.G., Foley A., 2019, Process Integration and Circular Economy for Renewable and Sustainable Energy Systems, Renewable and Sustainable Energy Reviews, 116, 109435.

Li J., Chu X., 2018, Rapid determination of physical and chemical parameters of reformed gasoline by near-Infrared (NIR) spectroscopy combined with the Monte Carlo virtual spectrum identification method, Energy and Fuels, 32(12), 12013-12020.

Li J., Karimi I.A., Srinivasan R., 2010, Recipe determination and scheduling of gasoline blending operations, American Institute of Chemical Engineers, 56(2): 441-465.

Li X., Wei Z., Qian D., 2019, Cross-Domain Fault Diagnosis of Rolling Element Bearings Using Deep Generative Neural Networks, IEEE Transactions on Industrial Electronics, 66(7), 5525–5534.

Mabood F., Boqué R., Hamaed A., Jabeen F., Al-Harrasi A., Hussain J., Alameri S., Albroumi M., Al Nabhani M.M.O., Naureen Z., Rawahi M.A., Al Futaisi F.A.S., 2017, Near-infrared spectroscopy coupled with multivariate methods for the characterization of ethanol adulteration in premium 91 gasoline, Energy and Fuels, 31(7), 7591-7597.

Villani C., 2008, Optimal transport: old and new, Vol. 338, Springer Science and Business Media, New York, USA.

Wang L., Sun D.-W., Pu H., Cheng J.-H., 2016, Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: A review of recent research developments, Critical Reviews in Food Science and Nutrition, 57(7), 1524-1538.

Wang J., Li S., Han B., 2019, Generalization of Deep Neural Networks for Imbalanced Fault Classification of Machinery Using Generative Adversarial Networks, IEEE Access, 7, 111168–111180.