Design Studies and Intelligence Engineering
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA220698

Improving the Performance of Industrial Dyeing Recipe Models by Taking Log-Transform in the Data Pre-Processing

Zhiwen TU[a], Congwei SONG[b], Xianan QIN[a,c,1] and Xiaoming John ZHANG[b,2]
[a] *National Engineering Laboratory for Textile Fiber Materials and Processing Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China*
[b] *Yanqi Lake Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing 101408, China*
[c] *Zhejiang Provincial Innovation Center of Advanced Textile Technology, Shaoxing 312000, China.*

Abstract. Recent research advances on fabric dyeing have focused on modeling the relationship between dye concentrations and the final color on fabrics. The emerging techniques in related studies have great potential to evolve the traditional dyeing industry to manufacture much more smartly. Given that dyeing is a complex process regulated by many factors, one of the challenging problems in the aforementioned techniques is to maintain the modeling accuracy at acceptable level. Other than developing high-performance algorithms and model architectures, it is also important to include effective data pre-processing techniques in modeling. In this paper, we show that conducting log-transform to the industrial dyeing data can greatly improve the performance of industrial dyeing recipe models. Such observations are confirmed on modeling tasks using different formats of color as input and different types of loss function in the model training. These findings may provide useful implications for related studies for the dyeing industry.

Keywords. Dyeing industry; Dyeing recipe model; Data pre-processing; Log-transform; Machine learning.

1. Introduction

Dyeing is of great importance to the textile manufacturing industry. In a typical fabric dyeing process, dyes are first diffused and then adsorbed in the fabric fibers [1]. The color of the dyed fabric is determined by the amount of dyes that are fixed on the fabric. The amount of dyes that finally stay on fabrics can be affected by many complex factors during the dyeing process [2], which has made recommending accurate dyeing recipes become challenging.

In fact, if the dyeing process is conducted under routine procedures, many factors in the dyeing process, such as water quality and temperature [3-5], can be seen as controlled. In this case, the quantitative relationship between the inputs and the outputs (i.e., the target color and the dye concentrations, respectively.) can be constructed and

[1] Corresponding Author: qin@zstu.edu.cn.

[2] Corresponding Author: zhangxiaoming@bimsa.cn.

used as dyeing recipe “recommender” [6-8] for practical dyeing manufacturing tasks. Thanks to the development of modern industrial big-data techniques [9-13], it is now feasible to build such recommenders based on industrial manufacturing records from the dyeing manufacturers. In our earlier work, we have reported the building of such a recommender based on a modular architecture structured system, in which individual gradient boosting regression trees (GBRT) are trained from dyeing manufacturing records for single types of fabric versus dye combinations [14]. GBRT can predict dye concentrations with prediction errors as good as ~7-10% [14], and have also been successfully applied into other applications with great model performance [15-18].

Despite the success of using GBRT, the earlier report lacks baselines for comparison [14]. It is thus helpful to conduct the model training on simpler regression models, e.g. multivariate linear equation, which are frequently used as baselines for modeling studies. Also, many important observations in previously reported modeling tasks can be tested again on the baseline model. In this regard, we have modeled the relationship between dye concentrations and the color of the dyeing process for a reactive dye combination (which is composed of Colvaceton reactive dye-navy blue CF (CRD-navy blue), Colvaceton reactive dye-bright red 3BSN150% (CRD-red) and Colvaceton reactive dye-yellow 3RS150% (CRD-yellow)) on rayon fabrics using multivariate linear equations. In particular, we have analyzed the effects from log-transform on the model performance. We confirm that log-transform can improve the model performance and such observations can be obtained in modeling with different color formats as input. Loss function with squared symmetric percent error (SMPE) (See Section 2.1 for the details), which has been found to outperform conventional least-square (LS) loss function in an earlier report [19], does not show superiority in industrial dyeing recipe modeling tasks. However, the improvement on the model performance by log-transform can still be observed in the model training where SMPE is introduced in the loss function.

2. Results and discussion

2.1. Data collection and the settings for regression modeling

The multivariate linear regression modeling is conducted on an industrial dyeing dataset which has been reported in the earlier report [14]. The dataset is consisted of 810 industrial dyeing records for rayon fabrics using a combination of three reactive dyes (CRD-red, CRD-Navy blue and CRD-yellow). This dataset was provided by Shaoxing Xingming Dyeing & Printing Co., Ltd. in Zhejiang Province of China. Conventional least-squares (LS) method is by default used to minimize the loss function

$$L = \sum_i \left(c_{pred,i} - c_{true,i}\right)^2$$

in regression, where $c_{pred,i}$ is the concentration predicted by model and $c_{true,i}$ is the true dye concentration. In addition, we have also tested a second form of loss function in linear regression modeling, where the squared residue term is replaced by squared symmetric percent error (SMPE)$\left(2\frac{c_{pred,i}-c_{true,i}}{c_{pred,i}+c_{true,i}}\right)^2$ [19]. The color format is chosen as either RGB or CIELAB, which is provided by the color measurement and can be calculated from the full spectra values that are obtained from the color measurement. These two multivariate linear equations are used in this study:

$$c_{pred} = k_0 + k_1R + k_2G + k_3B$$
$$c_{pred} = k_0 + k_1L + k_2a + k_3b$$

where k_i is the i-th model parameter (k_0 stands for the constant term); R, G and B stand for the RGB values; L, a and b stand for the CIELAB values. The complete dataset is randomly divided into two parts for model training and testing under a ratio of 60% to 40%, which is under the same setting as in reference [14]. Mean absolute error (MAE), mean absolute percent error (MAPE) and weighted absolute percent error (WAPE) are used to quantify the model performance. The definitions of these metrics are shown as below [14], [20]:

$$MAE = \frac{1}{N}\sum_i \left|c_{pred,i} - c_{true,i}\right|$$

$$MAPE(\%) = \frac{100}{N}\sum_i \left|\frac{c_{pred,i} - c_{true,i}}{c_{true,i}}\right|$$

$$WAPE(\%) = 100\frac{\sum_i \left|c_{pred,i} - c_{true,i}\right|}{\sum_i c_{true,i}}$$

where *N* is the number of concentration pairs. We repeat 21 times for each modeling task and report the metrics of the one whose MAPE is the median.

In this paper, we will check the effects from to basic settings, the using of different color format (RGB or CIELAB) as input and the log-transform to the data. The numerical experiments are conducted based on two different loss functions (LS and SMPE). Thus, four different modeling tasks are conducted in this paper, (1) conventional multivariate linear regression (Linear), (2) multivariate linear regression using SMPE in loss function (SMPE-Linear), (3) log-transform to the data prior to multivariate linear regression (Log-Linear), and (4) log-transform to the data prior to multivariate linear regression with SMPE in loss function (Log-SMPE-Linear). When log-transform is processed to the data, the processing is done to both input and output, and the metrics of model performance is reported after the input and the output are transformed back to normal scale. Specifically, in modeling where log-transform is applied to CIELAB values (negative CIELAB values may appear), an absolute value of 500 is added to the input to guarantee that the logarithm can be taken.

2.2. The Baseline

Without additional processing on data and the regression algorithm, multivariate linear regression model with conventional least-squares method to approach optimized model parameters can be used as baseline for related studies. We have fitted this baseline model to the data. The observed MAPE and WAPE for these baseline models range from ~70%-150% (Table 1-2), which is much worse than the levels of error (~7-10%) reported in reference [14]. These results can serve as baseline for related modeling studies for industrial dyeing process.

2.3. The effect of performing Log-transform to the data

As we anticipated, conventional linear equations are not capable of modeling the relationship between dye concentrations and the measured color on dyed fabrics with good accuracy. We reason that the source of large error likely comes from the non-linearity in the modeled relationship. Log-transform is a practical data pre-processing technique which can decrease the degree of non-linearity between the input and the output, thereby improving the model performance [21-22]. We thus perform log-transform to the data to check if it can improve the model performance from the baseline level. In all cases, significant improvement can be observed after the data is processed with log-transform, indicating that log-transform is a practical and useful solution to increase the model performance. This is actually in good agreement with what has been observed in the GBRT models [14]. We therefore conclude that it can be listed as a generalized data pre-processing method to related modeling studies for dyeing process.

Table 1. Performance of models with RGB as color format in input.

No.	Dye	Model	MAE	MAPE(%)	WAPE(%)
1	CRD-Navy blue	Linear	0.414	153	59.6
2		SMPE-Linear	0.394	102	61.9
3		Log-Linear	0.347	73.2	46.9
4		Log-SMPE-Linear	0.249	79.4	35.1
5	CRD-red	Linear	0.17	119	47
6		SMPE-Linear	0.213	107	53.2
7		Log-Linear	1.15	79.6	254
8		Log-SMPE-Linear	0.131	69.5	38.8
9	CRD-yellow	Linear	0.234	83.2	40.5
10		SMPE-Linear	0.243	78.4	40.3
11		Log-Linear	0.259	48.4	43.2
12		Log-SMPE-Linear	0.197	49.2	36

Table 2. Performance of models with CIELAB as color format in input.

No.	Dye	Model	MAE	MAPE(%)	WAPE(%)
13	CRD-Navy blue	Linear	0.36	147	60.1
14		SMPE-Linear	0.377	133	53.8
15		Log-Linear	0.0982	32.6	14.7
16		Log-SMPE-Linear	0.111	31.1	15.8
17	CRD-red	Linear	0.2	141	44.4
18		SMPE-Linear	0.265	294	63.7
19		Log-Linear	0.0674	32.4	16.4
20		Log-SMPE-Linear	0.074	31.3	16.9
21	CRD-yellow	Linear	0.197	72.2	33.1
22		SMPE-Linear	0.262	127	46.6
23		Log-Linear	0.103	31.2	19.9
24		Log-SMPE-Linear	0.122	30.6	20.7

2.4. The effect of using symmetric percent error in loss function

A new type of loss function with symmetric percent error (SMPE) in the squared residue term, which is as described in Section 2.1, has been reported to outperform conventional least-squares method on small-value datasets [19]. It is thus of interests to check if introducing SMPE in the loss function will further improve the model performance. As shown in Table 1-2, in all cases, models optimized by loss function with SMPE show very similar MAPE and WAPE levels as observed from the models optimized by conventional least-squares loss function. These results suggest that the using of symmetric percent error in loss function can not improve the performance considerably for related modeling tasks.

3. Conclusions

In this paper, we have modeled the relationship between dye concentrations and the final color for industrial fabric dyeing process using conventional linear equations. Essentially, we confirm that log-transform can improve the performance of industrial dyeing recipe models. We also show that the using of symmetric percent error in the loss function can not considerably improve the model performance as reported for the case where small-value data is involved. The fundamental settings that are tested in this paper may provide useful implications for modeling studies for the dyeing industry.

Acknowledgement

This work is partially supported by the Science Foundation of Zhejiang Sci-Tech University under Grant No. 21202242-Y. The authors would like to thank Ms. Yaguang Wang for conducting data-cleaning to the dataset which is used in this work.

References

[1] R. Shamey, X. Zhao, *Modelling, simulation and control of the dyeing process*. Elsevier; 2014.

[2] A. R. Choudhury, Dyeing of synthetic fibres. In *Handbook of textile and industrial dyeing*, Woodhead Publishing, 2011.

[3] H. N. Harvey, J. Park, Automation in the dyeing laboratory and its influence on accuracy in batch dyeing. *Journal of the Society of Dyers and Colourists*. **105**(1989), 207-11.

[4] C. Huang, W. Yu. Control of dye concentration, pH, and temperature in dyeing processes. *Textile Research Journal*. **69**(1999), 914-8.

[5] W. J. Jasper, M. Günay. Measurement and control of dyeing. In *Modelling, Simulation and Control of the Dyeing Process,* Woodhead Publishing, 2014.

[6] P. Resnick, H. R. Varian, Recommender systems. *Communications of the ACM.* **40**(1997), 56-8.

[7] S. Jalali, S. A. Golpayegani, H. Ghavamipoor, Designing a model of decision making in layers of supply, manufacturing, and distribution of the supply chain: A recommender-based system. In *8th International Conference on e-Commerce in Developing Countries: With Focus on e-Trust*, 2014.

[8] K. Mertens, T. Holvoet, E. Berbers, A. Distrinet, Recommender systems. In *Wirtschaftsinformatik* 1997.

[9] A. Al-Abassi, H. Karimipour, H. HaddadPajouh, A. Dehghantanha, R. M. Parizi, Industrial big data analytics: challenges and opportunities. In *Handbook of big data privacy*, Springer, Cham, 2020.

[10] J. Yan, Y. Meng, L. Lu, L. Li, Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance. *IEEE Access*. **5**(2017), 23484-91.

[11] X. Xu, Q. Hua, Industrial big data analysis in smart factory: Current status and research strategies. *IEEE Access*, **5**(2017), 17543-51.

[12] J. Wang Survey on industrial big data. *Big Data Research*. **3**(2017), 2017057.

[13]C. Li, Y. Chen, Y. Shang, A review of industrial big data for decision making in intelligent manufacturing. *Engineering Science and Technology, an International Journal*. 2021.

[14]X. Qin, X. J. Zhang, An Industrial Dyeing Recipe Recommendation System for Textile Fabrics Based on Data-Mining and Modular Architecture Design. *IEEE Access*. **9**(2021), 136105-10.

[15] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* , 2016.

[16] Y. Huang, Y. Liu, C. Li, C. Wang, GBRTVis: online analysis of gradient boosting regression tree. *Journal of Visualization*. **22**(2019), 125-40.

[17] P. Nie, M. Roccotelli, M. P. Fanti, Z. Ming, Z. Li, Prediction of home energy consumption based on gradient boosting regression tree. *Energy Reports*. **7**(2021), 1246-55.

[18] Z. Zhang, W. Zhu, J. Chen, Q. Cheng, Remotely observed variations of reservoir low concentration chromophoric dissolved organic matter and its response to upstream hydrological and meteorological conditions using Sentinel-2 imagery and Gradient Boosting Regression Tree. *Water Supply*. **21**(2021), 668-82.

[19] C. Song, X. J. Zhang. Linear regression based on symmetric percentage error and its application in printing

and dyeing industry. *Intelligent Computer and Applications,* **11**(2021), 71-74.

[20] P. Pai, C. Liu. Predicting vehicle sales by sentiment analysis of Twitter data and stock market values. *IEEE Access,* **6**(2018), 57655-62.

[21] X. Yu, X. Li, Y. Dong, R. Zheng. A deep neural network algorithm for detecting credit card fraud. In *2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)* 2020.

[22] C. Feng, H. Wang, N. Lu, C. Tian, H. He, Y. Lu, X. Tu, Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*. **26**(2014), 105.