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

doi:10.3233/FAIA220006

Towards Better Data Pre-Processing for Building Recipe Recommendation Systems from Industrial Fabric Dyeing Manufacturing Records: Categorization of Coloration Properties for a Dye Combination on Different Fabrics

Zhiwen TU[a], Yawen YIN[a] and Xianan QIN[a,1]
[a] *National Engineering Lab of Textile Fiber Materials & Processing Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China*

Abstract. Intelligent manufacturing for the fabric dyeing industry requires high-performance dyeing recipe recommendation systems. Nowadays, recommending dyeing recipes by mining dyeing manufacturing data has become a new direction for the development of recipe recommendation systems. As one of the indispensable parts in the system development, data pre-processing needs more than routine steps such as the removal of missing data and outliers. Considering that dyes can have very different coloration properties on different fabrics, dyeing manufacturing records for a given dye combination to different fabric types should be properly categorized before they are used for training regression models for dyeing recipe prediction. In this paper, we propose a simple but effective method for this categorization work. Our method uses conventional K-means clustering analysis to find fabric types that have similar coloration properties for a given dye combination. We have applied the method on a dye combination formed by 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 28 different types of fabrics. We show that these 28 types of fabrics can be well categorized into 8 groups based on the coloration properties. Our proposed method can be listed as one of the standard data pre-processing steps in the development of data-mining based recipe recommendation systems.

Keywords. Dyeing recipe recommendation, data-mining, data pre-processing, clustering analysis

1. Introduction

Since E. Allen's pioneering paper in 1966 reporting the first computer-aided color matching algorithm [1], the development of recipe recommendation system for fabrics dyeing has become an important research topic for the field of textile and optical

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

engineering. Relying on the Kubelka-Munk theory [2], standard calibrations between concentrations of single dyes and the color on fabrics must be measured in traditional dyeing recipe recommendation methods [3-4]. Corrections to the calibrations, such as the Saunderson's correction [5], must be further conducted following the calibration works. These calibration works are labor-intensive and usually cost additional consumables during the measurement processes. In addition, substantial numerical calculations must be conducted to screen out proper dyes and find the dye concentrations when these traditional recipe recommendation systems work.

With the development of modern information techniques, the using of data-mining approaches, which makes full use of historical dyeing data to find the quantitative relations between dye concentrations and the color, has now become a new direction for the development of dyeing recipe recommendation systems [6-11]. The dyeing data can either from laboratories, which are usually recorded in professional and standard ways, or from the manufacturing industry, which are usually not well organized. In spite of the incompleteness and disorganization in data recording, dyeing data from the manufacturing industry contain much more massive information, making it with great potential for developing high-performance dyeing recipe recommendation systems.

When industrial dyeing manufacturing data are used, two steps are of vital importance in the recipe recommendation system development. Firstly, proper data pre-processing must be performed, given that industrial manufacturing data are usually not well organized. Secondly, regression models will be trained, in which dye concentrations and their corresponding color information on fabrics are used as input-output. The second step, though seemingly complicating, can be achieved by following standardized model training procedures [12-13]. The data pre-processing step, however, is actually more important. In fact, thorough data pre-processing work is also in favor of increasing the model accuracy and making the models more generalized.

Other than conducting routine processes such as the removal of missing data and outliers, there are more things we can do for the data pre-processing. We shall inevitably face with a simple but fundamental question before we build regression models between dye concentrations and the color. That is, what kind of dyeing manufacturing records can be put together and used for training a single regression model? An intuitive answer to the question is that not all types of dyeing data can be used for the model building. This is because dyes can have different abilities to diffuse in and stay on different textile fibers. As a result, the coloration properties of a single dye combination can be very different when the dyes are used for dyeing tasks on different fabrics. Therefore, there should be an additional step in the data pre-processing: the dyeing manufacturing records should be categorized into proper groups based on the coloration properties of dyes on the fabrics. It should be noted that such a "categorization step" is not only a necessity for the model building, but also an important step for us to understand the coloration properties of a single dye combination on different fabrics.

In this paper, we report a method based on K-means clustering analysis for the aforementioned categorization task. 1263 records of dyeing manufacturing data that are from 28 different types of fabrics, which are dyed by a same dye combination (made up 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)), are used in this work. These 28 types of fabrics are categorized into 8 groups using our method. Detailed methods of the categorization are reported in the paper as well.

2. Data collection

All data used in this work were kindly provided by Shaoxing Xingming dyeing & printing Co., Ltd. in Zhejiang Province of China. The complete dataset was consisted of 1263 records of dyeing manufacturing data, each of which was with records of dye concentrations, fabric types, and measurement results of color information. These 1263 records were for a same dye combination which was made up 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). The color information was obtained using commercially available spectrophotometers (Datacolor, USA) under D65 light source. The dye concentrations were recorded in the unit of o.w.f which stands for "on weight of fabric".

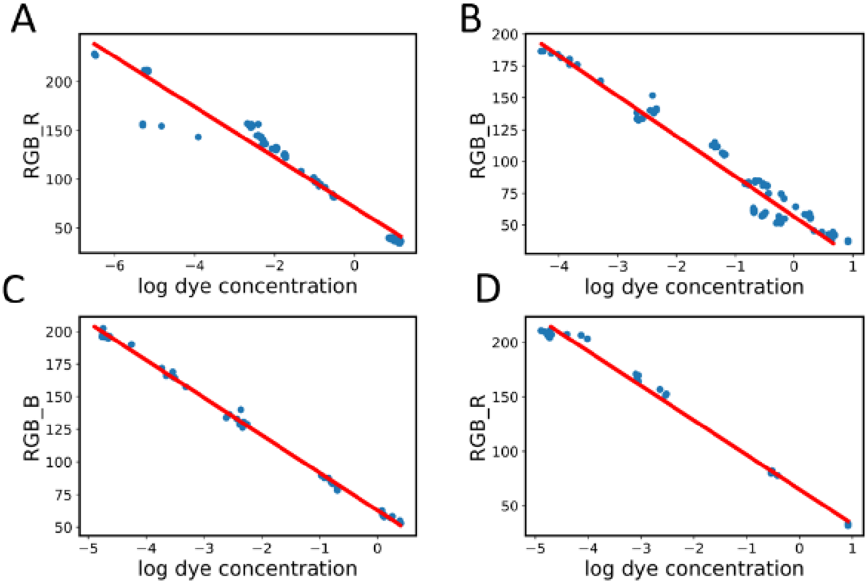

Figure 1. Example plots between log dye concentrations and the color information.

3. Results and discussion

3.1. Quantitative relationship between dye concentrations and the color

The quantitative relationship between dye concentrations and the color on fabrics has been reported complex by many literatures [1, 6]. To check this quantitative relation, we plotted the concentrations of CRD-Navy blue, CRD-Red and CRD-yellow with their color information (in RGB color space) on fabrics. We found that by taking logarithm for the dye concentrations, the non-linearity between concentrations and the color information became abated. Figure 1 shows the scattering plots between these dye concentrations and the color information. Linear fittings between the log dye concentrations and RGB values can be used to describe the quantitative relations between them, though such fittings were also found not to give great goodness-of-fitting (GOF) for many cases. The linear fittings can be described using following equations:

$$R = k \times dye_{log} + b \tag{1}$$

$$G = k \times dye_{log} + b \tag{2}$$

$$B = k \times dye_{log} + b \tag{3}$$

where *R, G, B* are the color information in the RGB color space, dye_{log} is the dye concentration after taking logarithm, *k* and *b* are fitting parameters. As there are three parameters of color information (*R, G, B*) and three dyes in the dye combination (CRD-Navy blue, CRD-Red and CRD-yellow), we obtained 9 slopes (*k*) and 9 intercepts (*b*) for a single fabric type. These 18 parameters form an 18-dimensional k-b space, and a single point in this high dimensional space represents a unique coloration property for the given dye combination.

3.2. Clustering analysis for the coloration property

To understand the dyeing properties of the given dye combination on different fabric types, we performed clustering analysis to the points that locate in the 18-dimensional space. K-means algorithm was used for clustering [14], and a simple elbow plot (the number of clusters N vs. K-means score) was generated for better understanding the clustering performance (Figure 2) [15-16]. The K-means score is defined as the total Euclidean distance from individual points to the clustering centers, which are randomly initialized and updated following the standard K-means algorithm. We found that the K-means score rapidly decreased from N = 2 to N = 5, and showed the elbow at around N = 8.

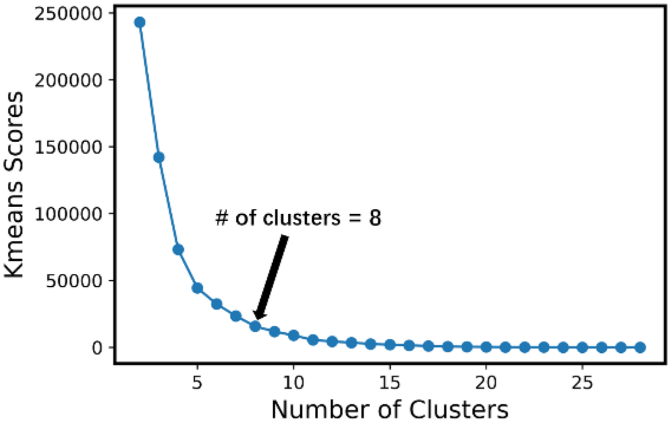

Figure 2. Elbow plot for the K-means clustering in the k-b space.

To better visualize the clustering results, we performed principle component analysis (PCA) [17] to the data in the 18-dimensional k-b space. The first two major components were kept (which take 55.41% and 18.36%, respectively) and the projections in this 2-dimensional space are plotted in Figure 3. The clustering analysis results are also shown in Figure 3, in which data points from different cluster groups are marked by different colors, and the centers of clusters are marked by crossing marks.

Table 1 summarizes the results of the clustering analysis. 28 types of fabrics are categorized into 8 groups, each of which corresponds to a unique type of coloration property of the dye combination that is made up of CRD-Navy blue, CRD-Red and CRD-yellow. The names of the fabrics are directly translated from Chinese. Despite lacking completeness, we could still find the materials of the fabrics from these records.

In Figure 3, it is seen that there are 5 types of fabrics which have unique coloration properties for this dye combination that their corresponding points locate relatively far away from the rest 23 points. The points corresponding to the rest 23 types of fabrics locate relatively close. These 23 fabric types are made up of rayon, tencel or cotton, indicating that these three materials may have similar coloration properties for this dye combination (CRD-Navy blue, CRD-Red and CRD-yellow). The 5 "isolated" types of fabrics contain either terylene or nylon, which may make the coloration properties become very different. Interestingly, we also observe a point corresponding to a type of rayon fabric in the k-b space which locates quite away from those of the 23 types of fabrics. This rayon fabric was actually marked with "rayon satin", which means that this type of fabrics should look shining. The special optical reflection properties on the surface of this fabric type may be the reason for it being "isolated".

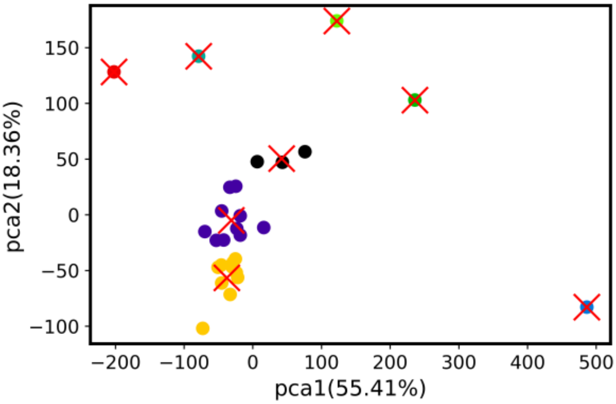

Figure 3. Visualization of the clustering analysis of the coloration properties.

Table 1. Results of the K-means clustering analysis to the coloration properties of CRD-Navy blue, CRD-Red and CRD-yellow on 28 different types of fabrics.

Type of coloration property	Fabrics (Directly translated from the Chinese remarks provided by the dyeing manufacturing company)	Materials
1	30s twill rayon, 40s rayon, 20s twill rayon, 31s rayon drill, 60*60/90*88 rayon, 60s rayon argyle/diamond check, tencel cotton twill, 21s tencel, tencel knitting, tencel cotton.	Rayon, tencel/cotton, tencel
2	Nylon cotton	Nylon/cotton
3	60s rayon satin	rayon
4	Terylene/rayon	Terylene/rayon
5	30*24 reversible twill rayon, 31s rayon slub, cotton knitted	Rayon, cotton
6	Terylene/tencel	Terylene/tencel
7	30*68 rayon, 40s twill rayon, 45s rayon, 30s rayon with pineapple	Rayon,

	eyelet pattern, 30s jacquard rayon, 60s rayon single jersey(knitted)/plain(woven), 21s twill rayon, viscose rayon, 30s twill tencel, cotton rayon	cotton/rayon
8	Nylon/rayon	Nylon/rayon

4. Conclusion remarks and future perspectives

Industrial fabric dyeing manufacturing data are often massive mixtures formed by different dye combinations, fabric information and color measurement results. To develop recipe recommendation systems from such data, proper data pre-processing processes must be conducted. In addition to conventional procedures such as removing outliers and missing data, dyeing manufacturing data should also be properly categorized into certain groups in the data pre-processing.

In this paper, we have proposed a simple method based on K-means clustering algorithm to achieve this categorization. We have applied the proposed method to the dye combination formed by CRD-navy blue, CRD-red and CRD-yellow on 28 different types of fabrics. Using our method, these 28 fabric types can be categorized into 8 groups based on the coloration properties of the dye combination on them.

It should be noted that we propose this categorization work as an additional step in the data pre-processing only based on reasoning. Therefore, for future study, it is worth performing a comparison work on the model performance between the models built with and without our proposed categorization work in the data pre-processing. Also, algorithms other than K-means, such as DBSCAN [18], may be tested as the clustering algorithm for the categorization work.

Acknowledgement

X. Qin would like to thank Dr. John Zhang for kindly providing the dyeing manufacturing data. Special appreciation also goes to Dr. Amy Wang of Zhejiang Sci-Tech University for translating the Chinese fabric names into English. This work was partially supported by the research grant (21202242-Y) from Zhejiang Sci-Tech University to X. Qin.

References

[1] E. Allen, Basic equations used in computer color matching, *JOSA* **56** (1966), 1256-1259.

[2] P. Kubelka and F. Munk, An article on optics of paint layers, *Z. Tech. Phys* **12** (1931), 259-274.

[3] H. R. Davidson, H. Hemmendinger and J. L. R. Landry, A system of instrumental colour control for the textile industry, *Journal of the Society of Dyers and Colourists* **79** (1963), 577-590.

[4] H. R. Davidson and M. Taylor, Prediction of the color of fiber blends, *JOSA*, **55**(1965), 96-100.

[5] J. Saunderson, Calculation of the color of pigmented plastics, *JOSA* **32** (1942), 727-736.

[6] M. Jawahar, C. B. N. Kannan, and M. K. Manobhai, Artificial neural networks for colour prediction in leather dyeing on the basis of a tristimulus system, *Coloration Technology* **131** (2015): 48-57.

[7] Q. Liu, X. Wan, J. Liang, et al. Neural network approach to a colorimetric value transform based on a large-scale spectral dataset, *Coloration Technology*, 133 (2017), 73-80.

[8] B. Zhang, Y. Shi, H. Yang, X. Liu and A. Zhang, Research on application of the minimum error average fitting method in computer color matching, *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics, Beijing* (2011), 293-296

[9] A. S. Nateri, E. Hasanlou, and A. Hajipour, Using adaptive neuro-fuzzy and genetic algorithm for simultaneously estimating the dye and AgNP concentrations of treated silk fabrics with nanosilver, *Pigment & Resin Technology* (2019).

[10] S. Chaouch, A. Moussa, I. B. Marzoug, et al., Colour recipe prediction using ant colony algorithm: principle of resolution and analysis of performances, *Coloration Technology* **135**(2019), 349-360.

[11] S. Chaouch, A. Moussa, I. B. Marzoug, et al., Application of genetic algorithm to color recipe formulation using reactive and direct dyestuffs mixtures, *Color Research & Application*, **45**(2020), 896-910.

[12] S. Vieira, R. Garcia-Dias, W. H. L. Pinaya, A step-by-step tutorial on how to build a machine learning model, *Machine learning,* Academic Press, 2020, 343-370.

[13] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, 2019.

[14] J. A. Hartigan, M. A. Wong, Algorithm AS 136: A k-means clustering algorithm, *Journal of the royal statistical society. series c (applied statistics)* **28**(1979), 100-108.

[15] J. D'Silva, U. Sharma, Unsupervised automatic text summarization of Konkani texts using K-means with Elbow method, *Int J Eng Res Technol* (2020), 2380-2384.

[16] A. Pandey, A. K. Malviya, Enhancing test case reduction by k-means algorithm and elbow method, *International Journal of Computer Sciences and Engineering* **6**(2018), 299-303.

[17] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemometrics and intelligent laboratory systems* **2**(1987), 37-52.

[18] E. Schubert, J. Sander, M. Ester, et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN, *ACM Transactions on Database Systems (TODS)*, **42**(2017), 1-21.