Detection of Sparsity in Multidimensional Data Using Network Degree Distribution and Improved Supervised Learning with Correction of Data Weighting

Shinya Ueno[1,2(✉)] and Osamu Sakai[1]

[1] Department of Electronic Systems Engineering, The University of Shiga Prefecture, Hassaka-cho 2500, Hikone, Shiga 522-8533, Japan
s.ueno@sakigakes.co.jp

[2] Checkers Co., Ltd., 50, Nishishichijoonmaedacho, Shimogyo, Kyoto 600-8897, Japan

Abstract. Multidimensional data are representatives in a wide range of applications, from those in the latest state-of-the-art science and technology to specific social issues. And they have been subject to analysis using methods such as regression analysis and machine learning. However, they are rarely obtained as complete data and contain more or less biases and deficiencies. In this study, we form a network from a multidimensional dataset and use its degree distribution to detect data sparsity. Although model analysis based on the degree distribution has been conducted for many years, sparsity detection has not been a target of the degree distribution analysis. Furthermore, we attempt to increase the accuracy and precision of supervised learning by applying regressive weighting according to node grouping in the degree distribution spectrum. By making use of this algorithm, we can expand the range of utilization of incomplete data together with other promising progresses in complex networks.

Keywords: Network analysis · Multidimensional data · Sparsity · Supervised learning

1 Introduction

Multidimensional data are essential elements for various types of analysis and its analysis method has been developed for more than a quarter of a century. They are representative datasets in various fields and applications, such as medical applications [1–4] like clinical trials for new drug development and determination of physical condition based on skin color, social applications like road and railroad maintenance [5,6], industrial applications like product inspection [7–10], and chemical applications like synthesis of material compounds [11]. We have also been performed on the evaluation of plasmas with multiple dimensions where multidimensional data is inevitable for the evaluation of plasmas for industrial applications [12–14].

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2016 2022, SCI 1077, pp. 390–401, 2023.
https://doi.org/10.1007/978-3-031-21127-0_32

They have been subjected to analysis using various methods such as regression analysis [15–17], data mining, and supervised learning [18–20], both linear and nonlinear. In supervised learning, multidimensional data are often used as teacher data, and in our previous work, we have also used a multidimensional dataset as Training Data for calibrating optical sensors using neural networks [21].

However, all data that are targeted cannot always be available as a complete dataset in which their distributions in the multiple dimensions are quite uniform. For example, anomalous values may be observed, and it is quite possible that the amount of data may be biased or missing. We will discuss this in more detail below, but we have confirmed that bias or lack of data exists even in the Training Data we have used in our past studies [22].

Analysis on a dataset with missing data may reduce the accuracy and precision of the analysis results; for example, if the missing data contain information that is crucial for the analysis, it may result in a linear approximate model, whereas the original data distribution might be actually nonlinear. Therefore, how to treat incomplete data with missing data as if it were complete data is an important item in multidimensional data analysis.

Seeking more complete data can be expensive, and this hinders the use of multidimensional statistical data in various fields, not just for machine learning. In other words, when the statistical data have some degree of incompleteness, the data-analysis procedure with capable technology to perform accurate analysis based on such incomplete data with sparsity will expand the possibilities of using data analysis. Here, we note that on literature [23] defines "sparse" as areas where edges are not connected, and another report [25] defines it as areas where data do not exist. We treat sparsity as the absence of data in a region where it should be present.

In the previous studies, complete-case analysis or listwise deletion was used to address missing or biased data in multidimensional data. When the number of variables is large and the proportion of missing data is high, methods called available-case analysis and pairwise deletion have been used. These methods are used in various situations because they are intuitive and easy to implement. For example, methods such as assigning data to missing parts to prepare pseudo-complete data [26,27] and approximating missing parts by weighting formulas [28–30] have been studied. On the other hand, since machine learning has become widely used, supervised learning is often performed with missing and biased data [31].

In contrast, in our previous study [22], we applied the model of locally linear embedding (LLE) [23] and attempted to ensure the reliability of the target region by emphasizing the weight of the data surrounding the missing-data or sparse area. Here, the missing data were checked not automatically but manually, and the missing data were evaluated by simple doubling and/or tripling of the number of missing data. Although this study achieved certain results, such intuitive methods were only valid for this model data, and the method was not universally applicable to other datasets. In the field of machine learning for

complex networks, rebalancing has been done by undersampling, which reduces the amount of data, and oversampling, which increases the amount of data [32]. In contrast, the method we are going to describe here is a method that attempts to supplement data regions that either do not exist or are extremely scarce by increasing the surrounding data.

Therefore, in this study, we propose a model based on a complex network that takes into account mutual positions of a large number of data points, and based on methods such as degree-distribution derivation and clustering, we select the amount of data that should be handled for rebalancing. Here, we propose a new algorithm that applies regressive weighting in the model. In contrast to LLE [23], which is based on distributions of local data points, our research is based on topology of the derived network and its statistical property. Over the years, many studies have been conducted on models based on complex networks, such as a random network in which each node is connected randomly and its degree distribution follows a Poisson distribution, and scale-free networks [24] in which the degree distribution follows a power-law distribution. That is, there have been many studies on degree distributions, but this study is unique since decomposition of the degree distribution is used for data-deficiency correction to improve accuracy of analysis, in particular, for supervised learning which requires training datasets preferably without sparsity.

2 Calculation Methods

2.1 Two Dimensional Color Coordinate and Target Task in Our Supervised Learning

Datasets in the multidimensional space are found in various scientific and technological areas, and in this study, color data on xy coordinates, which have been used in previous studies [21,22], were used in a model for examining the algorithm we propose here. The color data are expressed in three parameter red (R), green (G), blue (B), which are converted to points on the xy color coordinates by the following formulas defined in CIE1931 [33]: $X = 2.7689\ R + 1.7517\ G + 1.1302\ B$, $Y = 1\ R + 4.5907\ G + 0.0601\ B$, $Z = 0\ R + 0.0565\ G + 5.5943\ B$, $x = X\ /\ (X + Y + Z)$, $y = Y\ /\ (X + Y + Z)$.

When we consider data points on the xy color coordinates, color data are reduced from the three-dimensional attributes to the two-dimensional variables. Through this conversion, all R, G, and B color data are mapped on the xy color coordinates, and this conversion is not completely linear to R, G and B values; the data that were spaced at every equal step along the R, G, and B axes are not similarly spaced on the xy color coordinates. Furthermore, if we detect color values using any optical sensors, they include error rates due to poor matching coefficients to R, G, and B wavelength spectra. Thus, we always adapt a nonlinear calibration procedure to obtain sufficient accuracy for such multidimensional data. In our previous study [21], we successfully performed one way of suitable calibration, but we have not been sure about its validity with universality. In this study, we aim to accomplish a suitable and universal

calibration method that can be performed automatically according to a simple algorithm, specifically for tuning weight coefficients for sparse training datasets in the multidimensional space.

2.2 Method for Network Diagram Formation and Analysis Algorithm

We show the flow chart about the data processing proposed in this study in Fig. 1. For the dataset described below, we calculated Euclidean distance between all the nodes with the specific xy color coordinate values (x, y) to find out connections or edges between nodes. A threshold for a connection was set for this distance value ranging from 0.01 to 0.05, and the adjoining points with a distance that fell below the threshold were connected to form an edge, and the resulting edge list was used to configure a network using Cytoscape [34]. The network created by this step is then evaluated for deficiencies and bias using the following two point of views. First, from the generated network diagram or visualized network topology, we detect differences from the case with the complete dataset without sparsity, based on an intuitive evaluation of its shape. Second, the degree k of each node is detected, the node count is summed up to create a spectrum for every 10°C, and the differences in the distributions among the datasets clarify sparsity in the datasets, being useful to determine if there is any bias or missing data. Finally, to make use of such unveiled features, for the training dataset with sparsity, supervised learning including regressive weighting is performed based on this degree-distribution analysis, and the effects of the weighting are evaluated using mean absolute errors (MAEs) in the test dataset.

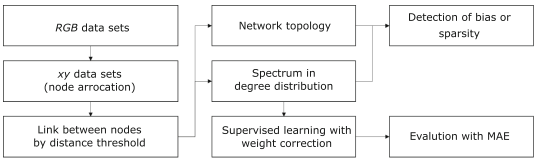


Fig. 1. Flow chart of data processing in this study.

2.3 Datasets on the Color Coordinate

Three datasets were prepared for the model used in this study; one of them is our target set that is the training dataset for calibration based on supervised learning, and the other two sets are for comparative data processing to clarify how our model with algorithm works.

In the first dataset, points are randomly placed on the xy color coordinates within the restricted area that can exhibit color data values. On the xy color coordinates, depending on formulae of R, G, and B variables, the color data

points should be located inside the triangle whose vertices correspond to R, G, and B as shown in Fig. 2a. "Random Data" refers to this dataset in this study, and the total count of data points is 1892.

In the second dataset, the data values covering all combinations of R, G, and B variables with every equal value step were transformed to the xy color coordinates, and here this dataset was defined as "RGB All Data" and the total count of data points is 1728. The profile of data points representing these data in the color coordinates is shown in Fig. 2b. As explained in Sect. 2.1, when the dataset of R, G, and B with every equal value step is mapped to the data points on the xy color coordinates, the corresponding x and y coordinates of these points are in inequal spatial steps. While the density of data points in the center of the triangle and in the light blue direction is sufficient, certain gaps exist in the purple and yellow directions. In addition to these gaps, we find isolated data points around the red area.

The third dataset is our target in this study; this is referred to as "Training Data," and the total count of data points is 1631. We used the entire set of color data from the color catalog given on more than 1600 sample sheets (PANTONE FORMULA GUIDE, Pantone LLC, X-rite Inc.), and converted all data into the xy color coordinates, which we used in our previous study [21]. The distribution on the xy color coordinates shown in Fig. 2c reveals sparse areas or regions of missing data in the circled areas.

Thus, the obtained dataset is not necessarily comprehensive, and in many cases, the dataset contains some elements of bias or missing data. Therefore, we aim to identify such biases and deficiencies using network diagrams, reinforcing these incomplete elements by adjusting weight coefficients of the existing data, and to augment and calibrate the data using supervised learning.

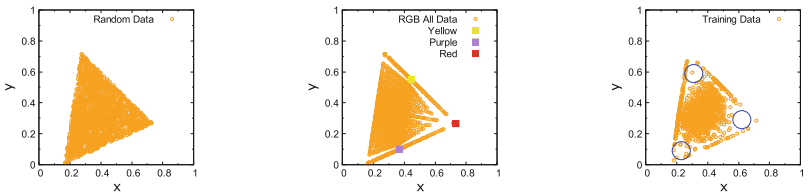


Fig. 2. Distributions of data points on the xy color coordinates for **a** "Random data", **b** "RGB all data", and **c** "Training data". The circles in blue line indicate sparse areas.

2.4 Data Calibration Methods and Weight Tuning

In Training Data, the values given in the color catalog on sample sheets are supervisory output signals whereas the counter values are obtained in the measurements of the sample sheets by the color sensor ("color checker" CC-01, Checkers

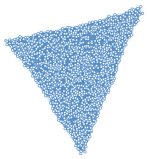
Co.) as measured values. Supervised learning with a neural network that is a simple one-hidden-layer perceptron was performed with R, G, and B in the measured values as inputs and x and y of the catalog values as outputs. Supervised learning was performed using the R package [35]. In our previous study, weight tuning among the given data points was performed simply by duplicating the target data; details are described in Ref. [21] and here we briefly review in the following. We categorized edge regions as areas with relatively small amounts of data, located on the edge of the given data space, and enhanced accuracy of their calibration by this simple weighting; we call this method "Edge-region". In this study, we propose a method to classify data points with lower densities based on the degree distribution of a complex network derived from distributed points in the multi-dimensional space, and to assign weights to these data points. We define this method as "byDegree". The difference detected over the catalog data, the measured data, and the calibrated data was quantified by MAE, and the MAE levels are representatives for the evaluation of this network-based weighting.

3 Calculation Results

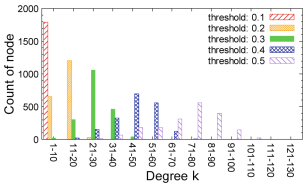
3.1 Network Diagrams and Degree Distributions

For each dataset, from data points scattered on the xy color coordinates in Fig. 2, we perform calculations and derive distributions measured by varying the distance threshold between 0.01 and 0.05 and a network diagram when the threshold is set to 0.05, as shown in Figs. 3, 4 and 5. In the case of Random Data, the network diagram is triangular as shown in Fig. 3a, just like the distribution of spatial data points on the xy color coordinates. In addition, when the threshold is changed linearly from 0.01 to 0.05, as shown in Fig. 3b, the degree distribution is getting close to that of the Poisson's distribution as the threshold increases. Although the shape is slightly distorted due to the triangular arrangement of the points, this is similar to the cases with $p = 0.003 - 0.038$ in a random graph, where, in general random graphs, the probability of connecting each node is defined by p. In our case of this graph in Fig. 3, since the distance between each node is random, it can be said to be a derivative of a random graph.

On the other hand, in the network diagram of RGB All Data shown in Fig. 4a, while maintaining the triangle outlook to some extent, we find a group of data that almost deviated from the triangle pattern and other nodes completely deviated from it. This automatically indicates presences of gaps or isolated data groups that were visible on the xy color coordinates in Fig. 2b. In the degree distribution shown in Fig. 4b, we find a clear difference from that in the case of Random Data; these differences are more outstanding than those observed in the network diagram. In this dataset, the degree distribution includes multi spectra with several maxima and minima, and their widths are large with long tails. This difference from the degree distribution of Random Data indicates that the distribution of color information in RGB All Data on the xy color coordinates is a model with completely different characteristics from that of Random Data,



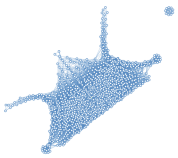
(a) Geographical network



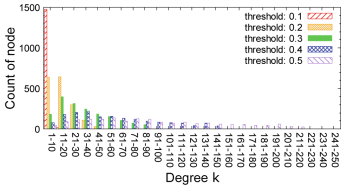
(b) Degree distribution

Fig. 3. Geographical network and degree distribution for random data. The distance threshold for networking is set to 0.5 on **a**, and from 0.1 to 0.5 on **b**.

arising from regularity in equal R, G and B steps in their components. It also has a different shape from the power-law distribution, leading to the fact that it is not a general scale-free network.



(a) Geographical network

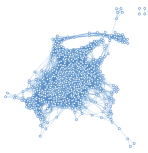


(b) Degree distribution

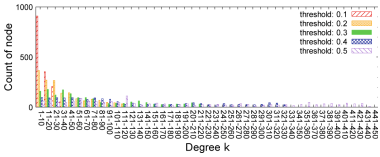
Fig. 4. Geographical network and degree distribution for RGB all data. The distance threshold for networking is set to 0.5 on **a**, and from 0.1 to 0.5 on **b**.

Finally, we consider the network diagram of Training Data, as shown in Fig. 5a. Although a rough footprint of the triangle remains, its shape is significantly broken, suggesting the existence of some kinds of data bias or deficiency, even when compared to the RGB All data. In other words, it is possible to diagnose and visualize data deficiencies and biases to some extent automatically and intuitively by comparing topology of network diagrams. Figure 5b shows the degree distribution of the Training Data. Figure 6a shows comparison of the degree distribution of the RGB All Data and the Training Data. We note that the count of data points in Training Data is somewhat higher approximately by 220, whereas the total k of RGB All Data is smaller. Despite of this discrepancy, from this comparison, it is outstanding that there is a large bias and sparse areas in Training Data. Furthermore, for RGB All Data, the degree distribution reaches at maximum around 51 and 60, from which the data points are decreasing (with some variation) as the k is raised. However, in Training Data, the number of data points that once decreased increases after 331, suggesting that the data can be divided into several clusters or spectra in the degree distribution. From this outlook investigation, we tentatively classified the data point

into three spectra: Spectrum I (from 0 to 130), Spectrum II (from 131 to 330), and Spectrum III (from 331 to 440). Over these spectra, we search for suitable calibration methods by weighting (or replicating) Spectrum I, which is a lower degree spectrum and is estimated to correspond areas of lower data density or sparse regions.

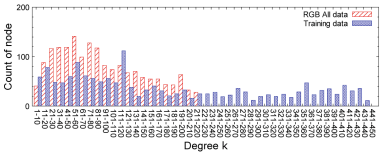


(a) Geographical network

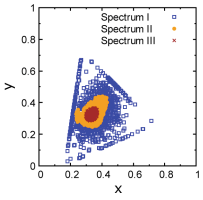


(b) Degree distribution

Fig. 5. Geographical network and degree distribution for training data. The distance threshold for networking is set to 0.5 on **a**, and from 0.1 to 0.5 on **b**.



(a) Comparison of degree distribution

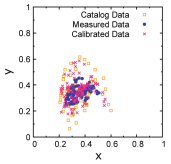


(b) Distribution on xy color coordinates.

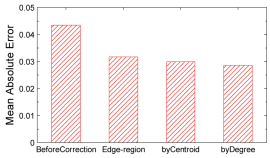
Fig. 6. Classification of data using degree-distribution spectra. **a** Comparison of degree distributions of RGB all data (red diagonal lined) and training data (blue hatched) when the threshold is 0.5. **b** Distribution of datasets on the xy color coordinates. Blue squares indicate Spectrum I ($k = 1 - 130$), orange circles indicate Spectrum II ($k = 131 - 330$), and brown crosses indicate Spectrum III ($k = 331 - 440$).

3.2 Validity for Supervised Learning: Calibration Using Neural Networks

After applying our algorithm for supervised learning [21], the estimated effects on MAE before and after calibration are shown in Fig. 7b, where the varying spatial scattering of data points on the xy color coordinates is shown in Fig. 7a. In Fig. 7b, the method proposed in this paper reduces MAE as much as or more than "Edge Region," indicating that this method proposed in this study is more effective on Training Data. The improvement can also be intuitively recognized from the visualized data-point spatial distributions based on this methods on the xy color coordinates, with smaller spatial gaps between the catalog and the calibrated data [21].



(a) Distribution on the xy color coordinates



(b) Comparison of MAE

Fig. 7. Validation of our method for supervised learning using multidimensional data. **a** Profiles of data points on two-dimensional space where orange squares indicate training data, blue circles indicate test values, and peach crosses indicate calibrated values. **b** Comparison of MAE of calibrated values. "BeforeCorrection" indicates the MAE before correction of data weighting, and "Edge-region," "byCentroid," and "byDegree" indicate the cases using edge region, R package (described in discussion), and degree distribution, respectively.

4 Discussion

As a comparative experiment using a more complicated method, we performed clustering of data points using the R package [35], and weights for supervised learning were set based on this clustering. Among the pre-installed methods, *Centroid* was used to optimize the clustering, and the dendrogram was used as the basis for dividing the clusters into several clumps. We assumed the cluster with the least number of data to be around the sparse area, and we attempted to calibrate them by assigning weights to them. This method is called "byCentroid" and the calibration results are shown in Fig. 7a. In this "byCentroid" experiment, we used the intuitive and easy-to-understand centroid method for creating clusters, but other clustering methods might be better, depending on the model, although it is sufficient for our aim here. Performance in comparing the learning results for each method is shown in Fig. 7b. All of the methods exhibit improvements over the results before weighting, and the best one for MAE is obtained with the learning based on "byDegree", proposed here.

Although sufficient results were obtained even with a low magnification factor in this model, it is necessary to select appropriate conditions based on the obtained results, such as a steeper gradient or a more detailed segmentation method for a more complicated dataset. In this case, the spectra were manually sorted by surveying quantitative data balancing, but it is easily possible to perform it automatically by deriving an approximate curve of the degree distribution and applying spectrum deconvolution, considering the maximum and minimum points discriminated.

Finally, we show results on entropy analysis, which represents macroscopic features of datasets visualized and configured in complex networks. When we assume hypothetical information flow from a given node s in a complex network which is the ensemble of all-node group S is induced into linked nodes t belonging

to the group T, like Figs. 3a, 4a and 5a, accumulated effects of its flow probability $p(t|s)$ lead to the conditional entropy $H(T|S)$ [36], given as:

$$H(T|S) = \sum_{s \in S} p(s) H(T|S = s) = -\sum_{s \in S} p(s) \sum_{t \in T} p(t|s) \log_2 p(t|s), \tag{1}$$

where $p(s)$ is the existence probability at the node s. Assuming equal $p(s)$ over S and equal $p(t|s)$ at the node s over T, the calculated results of $H(T|S)$ are: 6.10 for Random Data, 6.17 for RGB All Data, and 6.88 for Training Data. These results suggest that, in comparison with cases of balanced node density, the network of Training Data takes larger conditional entropy, which indicates wider choices for linkages between nodes in hypothetical information flow of Training Data. This seems to be contradictory to occurrences of sparsity, in which such flow mobilities through linkages are somewhat limited. However, in our case of Training Data, it is not always valid since, around sparse areas, some of nodes are rather condensed, as indicated in Fig. 6. Furthermore, this high entropy also points out existences of areas with high density of data points, which is located in the central aggregation in Fig. 2c. Thus, entropy estimation can support another insight of the algorithm proposed here and reinforce understandings in terms of complex networks.

5 Conclusion

In this study, we successfully detected sparsity in given data points using network diagrams constructed from their multidimensional locations and their degree distributions. We also succeeded in improving the accuracy and precision of supervised learning by applying regressive weights to the spectrum group in the degree distributions. In a comparative experiment based on a typical clustering method, our result for supervised learning shows a better accuracy, which indicates that the simpler method proposed in this study gave better results. By utilizing this complex network technique, it is possible to obtain more accurate and precise analysis results in supervised-leaning model even if the acquired multidimensional data contain biases and deficiencies, which will lead to a wider range of utilization of multidimensional data.

Acknowledgements. The authors thank the members of in Checkers Co., Ltd., in particular Dr. K. Taguchi, for his useful comments. This work is partially supported by the Regional ICT Research Center of Human, Industry and Future at The University of Shiga Prefecture, by the Cabinet Office, Government of Japan, and by a Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT/JSPS KAKENHI) with Grant No. 22K18704.

References

1. Fitzmaurice, F.M., Laird, N.M., Ware, J.H.: Applied Longitudinal Analysis. Wiley, New York (2011)
2. Blomeke, R.C., Elliott, J.S., Senjaya, B., Hales, G.T.: A comparison of fingerprint image quality and matching performance between healthcare and general populations. In: Proceedings of 2009 IEEE 3rd International Conference on BTAS, vol. 9, pp. 1-4, IEEE, Washington DC (2009)
3. Morris, D., Coyle, S., Wu, Y., Lau, T.K., Wallace, G., Diamond, D.: Bio-sensing textile based patch with integrated optical detection system for sweat monitoring. Sens. Actuators B Chem. **139**, 231–236 (2009)
4. Jiang, Z., Hu, M., Gao, Z., Fan, L., Dai, R., Pan, Y., Tang, W., Zhai, G., Lu, Y.: Detection of respiratory infections using RGB-infrared sensors on portable device. IEEE Sens. J. **20**, 13674–13681 (2020)
5. Lee, S.J., Kim, H.M., Kim, S.I., Lee, H.M.: Evaluation of structural integrity of rail-way bridge using acceleration data and semi-supervised learning approach. Eng. Struct. **239**, 1–16 (2021)
6. Shim, S., Kim, J., Lee, S.W., Cho, G.C.: Road damage detection using super-resolution and semi-supervised learning with generative adversarial network. Autom. Constr. **135**, 1–16 (2022)
7. Chandy, R.P., Scully, P.J., Thomas, D.: A novel technique for online measurement of scaling using a multimode optical fibre sensor for industrial applications. Sens. Actuators B Chem. **71**, 19–23 (2000)
8. Zhou, Z.-K., Wang, U.-K., Gong, H.-G., Shi, Y., Wang, Z., Zhang, B.: A fully-integrated optoelectronic detector with high gain bandwidth product. IEEE Access **7**, 53032–53039 (2019)
9. Wotrnba, H.: Sensor sorting technology-is the minerals industry missing a chance? In: Proceedings XIII IMPC Istanbul 2006, pp. 21-29. IMPC, Istanbul (2006)
10. Leelasattarathkul, T., Liawruangrath, S., Rayanakorn, M., Liawruangrath, B., Oungpipat, W., Youngvises, N.: Greener analytical method for the determination of copper(II) in wastewater by micro flow system with optical sensor. Talanta **72**, 126–131 (2007)
11. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., Kim, C.: Machine learning in materials informatics: recent applications and prospects. Comput. Mater. **3**(54), 1–13 (2017)
12. Sakai, O., Morita, T., Ueda, Y., Sano, N., Tachibana, K.: Chemical filters by non-thermal atmospheric pressure plasmas for reactive fields. Thin Solid Films **519**, 6999–7004 (2011)
13. Urabe, K., Hiraoka, Y., Sakai, O.: Hydrazine generation for the reduction process using small-scale plasmas in an argon/ammonia mixed gas flow. Plasma Sources Sci. Technol. **22**, 032003-1-4 (2013)
14. Urabe, K., Sakai, O.: Multiheterodyne interference spectroscopy using a probing optical frequency comb and a reference single-frequency laser. Phys. Rev. A **88**, 023856-1-5 (2013)
15. Girolami, M., Mischak, H., Krebs, R.: Analysis of complex, multidimensional datasets. Drug Discovery Today: Technol. **3**(1), 13–19 (2006)
16. Song, X., Wu, M., Jermaine, C., Ranka, S.: Statistical change detection for multi-dimensional data. In: KDD'07, SIGKDD, pp. 667-676. California (2007)
17. Dempster, A.P.: An overview of multivariate data analysis. J. Multivar. Anal. **1**, 316–346 (1970)

18. Zaidan, M.A., Motalagh, N.H., Fung, P.L., Lu, D., Timonen, H., Kuula, J., Niemi, J.V., Tarkoma, S., Petaja, T., Kulmala, M., Hussein, T.: Intelligent calibration and virtual sensing for integrated low-cost air quality sensors. IEEE Sens. J. **20**, 13638–13652 (2020)
19. Goodacre, R., Neal, M.J., Kell, D.B.: Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. Zentralbl Bakteriol **284**, 516–539 (1996)
20. Fang, J., Yang, F., Tong, R., Yu, Q., Dai, X.: Fault diagnosis of electric transformers based on infrared image processing and semi-supervised learning. Glob. Energy Interconnection **4**, 596–607 (2021)
21. Ueno, S., Sakai, O.: Data driven calibration of color-sensitive optical sensor by supervised learning for botanical application. IEEE Sens. J. **22**, 11915–11927 (2022). https://doi.org/10.1109/JSEN.2022.3171221
22. Ueno, S., Sakai, O.: Low-cost color-sensitive optical sensor calibrated by sparse training data. In: Proceedings of the 2021 IEEE 10th GCCE, pp. 402-403. IEEE Consumer Technology Society, Kyoto (2021)
23. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)
24. Albert, R., Barabasi, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
25. Hasan, M.A., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: SDM06: Workshop on Link Analysis. Counter-Terrorism and Security, pp. 798–805. SIAM, Maryland (2005)
26. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, New York, NY, USA (1987)
27. Tanner, M.A., Wong, W.H.: The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. **82**, 528–550 (1987)
28. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**, 663–685 (1952)
29. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)
30. Scharfstein, D.O., Rotnitzky, A., Robins, J.M.: Adjusting for nonignorable drop-out using semiparametric nonresponse models. J. Am. Stat. Assoc. **94**, 1096–1146 (1999)
31. Ma, M., Korniss, G., Szymanski, B.K.: Learning parameters for balanced index influence maximization. In: Processing 9th International Conference on Complex Networks and Their Applications, pp. 167–177. Springer, Madrid (2020)
32. Xue, J.-H., Hall, P.: Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis? IEEE Trans. Pattern Anal. Mach. Intell. **37**(5), 1109–1112 (2015)
33. Itten, J.: The Elements of Color. Van Nostrand Reinhold, New York, USA (1970)
34. Cytoscape open API. https://cytoscape.org/
35. The R Project for Statistical Computing. https://www.r-project.org/
36. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley, Hoboken (2006)