BMC Chemistry

RESEARCH ARTICLE

Open Access

Relaxometric learning: a pattern recognition method for T_2 relaxation curves based on machine learning supported by an analytical framework

Yasuhiro Date[1,2], Feifei Wei[1], Yuuri Tsuboi[1], Kengo Ito[1], Kenji Sakata[1] and Jun Kikuchi[1,2,3*]

Abstract

Nuclear magnetic resonance (NMR)-based relaxometry is widely used in various fields of research because of its advantages such as simple sample preparation, easy handling, and relatively low cost compared with metabolomics approaches. However, there have been no reports on the application of the T_2 relaxation curves in metabolomics studies involving the evaluation of metabolic mixtures, such as geographical origin determination and feature extraction by pattern recognition and data mining. In this study, we describe a data mining method for relaxometric data (i.e., relaxometric learning). This method is based on a machine learning algorithm supported by the analytical framework optimized for the relaxation curve analyses. In the analytical framework, we incorporated a variable optimization approach and bootstrap resampling-based matrixing to enhance the classification performance and balance the sample size between groups, respectively. The relaxometric learning enabled the extraction of features related to the physical properties of fish muscle and the determination of the geographical origin of the fish by improving the classification performance. Our results suggest that relaxometric learning is a powerful and versatile alternative to conventional metabolomics approaches for evaluating fleshiness of chemical mixtures in food and for other biological and chemical research requiring a nondestructive, cost-effective, and time-saving method.

Keywords: Machine learning, Nuclear magnetic resonance, Relaxometry, Support vector machine, Geographical origin determination

Introduction

Nuclear magnetic resonance (NMR) spectroscopy is one of the most versatile tools for chemical analysis in the fields of chemistry and biology [1]. NMR can be used to evaluate complex chemical and biological mixtures (e.g., those used in metabolomics studies), which characterize the metabolic profiles of a large number of samples derived from biological and environmental systems [2]. NMR-based metabolomics is used to determine geographical provenance, solve problems related to food fraud, and certify a "terroir" in food chemistry [3]. However, conventional NMR-based metabolomics approaches require relatively high-field NMR instruments to obtain high-resolution NMR spectra. Therefore, although recent advances in benchtop NMR instruments have been accompanied by increasing employment of low-field and benchtop NMR spectroscopy in NMR-based metabolomics studies, the practical utilization and on-site application of NMR-based metabolomics are limited [4].

As an alternative to metabolomics approach, NMR-based relaxometry has several advantages, including

*Correspondence: jun.kikuchi@riken.jp
[1] RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
Full list of author information is available at the end of the article

compatibility with low-field and compact NMR instruments, simple sample preparation, easy handling, and relatively low cost compared with metabolomics approach [5]. Therefore, NMR-based relaxometry is widely used in various research fields and industries such as the food industry, polymer industry, pharmaceutical industry, and geology. However, the application of the T_2 relaxation curves obtained by relaxometric measurements to biological studies has not yet been reported in terms of the characterization and evaluation of chemical mixtures to determine the geographical origin and achieve feature extraction by pattern recognition and data mining.

The T_2 relaxation curves obtained by relaxometric measurements appear to be relatively simple compared with the ^{1}H NMR spectra obtained in conventional metabolomics studies. However, the T_2 relaxation curves derived from complex chemical mixtures (e.g., fish and vegetables in the field of food chemistry) are likely to show slightly different profiles for each sample because of complex bound water states and interactions that are affected by the higher-order structure of macromolecules, although discerning such small differences by eye is usually difficult (Additional file 1: Fig. S1). With this in mind, we speculated that data mining methods such as multivariate analyses and machine learning (ML) could be applied to discover valuable information buried in T_2 relaxation curves.

Multivariate analysis methods such as principal component analysis, partial least squares (PLS), and soft independent modeling of class analogy [6] have been used to analyze relaxometric data [7–9]; however, the use of ML methods for such analyses has not been reported. Compared with multivariate analyses, ML methods are known to be superior for metabolomics studies in some situations [10], and several useful ML-incorporated analytical tools have been reported (e.g., MetaboAnalyst [11] and KODAMA [12]). In a previous work, we successfully developed several ML-based analytical approaches, namely, a prediction method for metabolic mixture signals [13], deep neural network (DNN)–mean decrease accuracy [14] and ensemble DNN [15] methods, variable selection for regional feature extraction [16], evaluation of surface water [17], impact estimation of food intake on mice [18], and evaluation of daily dietary intakes of humans [19] in metabolomics studies. Thus, we considered that an analytical method for T_2 relaxation curves combined with ML might be a helpful tool for metabolomics studies and could be used to discriminate between geographical origins and extract features pertaining to sample attributes.

In this study, we describe a technique called relaxometric learning, which is a pattern recognition method for T_2 relaxation curves that can be used as a simple and cost-effective tool for the data mining of complex chemical and biological mixtures (e.g., fish samples in food chemistry) as an alternative to conventional metabolomics approaches. To develop the relaxometric learning, an analytical framework optimized for data mining of T_2 relaxation curves was required. Therefore, we also developed an analytical framework for eliciting the improved classification performance of ML algorithms to enhance the performance of relaxometric learning. In the analytical framework, we incorporated two methods into the ML process: a variable optimization approach to enhance the classification performance and bootstrap resampling-based matrixing to balance the sample size between groups. As a model case, we selected the support vector machine (SVM) [20] approach for analyzing the T_2 relaxation curves obtained by NMR-based relaxometric measurements because SVM has been reported to be useful for analyses with a relatively small number of samples [21]. The classification performance of relaxometric learning was evaluated on the basis of the physical properties (hardness and tenderness) of fish muscles determined by the hierarchical cluster analysis (HCA) of compressive force data measured by an autograph machine in a data-driven manner. The applicability of the analytical framework to other ML methods was also evaluated.

Materials and methods

Sample collection

A total of 233 fish samples belonging to 34 families were collected from April 2012 to November 2018 at multiple coastal sites in Japan (Additional file 1: Table S1). There is no specific permission required for all of the sampling points as they are all public places. The animal experiments were performed in accordance with protocols approved by the Institutional Committee of Animal Experiment of RIKEN and adhered to the guidelines in the Institutional Regulation for Animal Experiments and Fundamental Guidelines for Proper Conduct of Animal Experiment and Related Activities in Academic Research Institutions under the jurisdiction of the Ministry of Education, Culture, Sports, Science and Technology, Japan. Among these samples, 233 and 209 samples (the different sample number was due to insufficient volume of fish muscles) were used for compressive force measurements by autograph and NMR measurements to obtain T_2 relaxation curves, respectively.

Compressive force measurements by autograph

Since the relaxation properties and water contents in fish muscle could be varied according to position differences [22], fish muscle above the anal fin was picked to avoid the impact of position differences and cut into slices

(5 mm thick and 10 mm wide) (Additional file 1: Fig. S2A). Stress testing (n = 5 per sample) was performed using a multipurpose stretching tester comprising an autograph (EZ-L, Shimadzu Co. Ltd., Kyoto, Japan) with a wedge-shaped cutter bit (Additional file 1: Fig. S2B). The parameters were controlled using TRAPEZIUM2 (ver. 2.36, Shimadzu), which is the manufacturer-supplied software. The loading rate was 2 mm min^{-1}, the total distance was 5 mm (Additional file 1: Figs. S2C and S2D), and the force–time curves were recorded (Additional file 1: Fig. S2E). A total of 1165 curves were obtained in the measurements.

The force–time curves for fish muscle were transformed into force–distance curves, with the zero point determined by the contact between the sample and the cutter bit. The portions of the curves between 0 and 3.63 mm were selected (34 curve data were omitted in this process due to the insufficient length (thickness)), and exponential fitting was performed in Microsoft Excel by using the following equation:

$$y = ae^{bx}, \quad (1)$$

where x represents the distance from the starting (zero) point, y represents the compressive force, and a and b represent the fitting coefficients.

A total of 1131 sets of compressive force data were preprocessed, and coefficients a and b were calculated by approximating the exponential function. Coefficients a and b were further analyzed using a data-driven HCA approach, which indicated that each sample could be categorized into two groups on the basis of the physical properties (i.e., hardness and tenderness) of the fish muscle (Additional file 1: Fig. S3). The fish belonging to group A had relatively hard muscles compared with those in group B.

NMR measurements to obtain T_2 relaxation curves

A piece of fish muscle (n = 3 per sample) was loaded into a 5 mm NMR tube inserting a 2 mm NMR tube filled with 99.9% D_2O solvent for locking. T_2 relaxation curves were recorded at 298 K on a high-resolution NMR spectrometer (AvanceIII HD-500, Bruker, Rheinstetten, Germany) equipped with a ^1H inverse probe with triple resonance by using the Bruker standard pulse program "cpmg_T2_1d" (Carr–Purcell–Meiboom–Gill sequence) with 1 scan, 512 data points, and 1.024 s acquisition time. The obtained raw data were normalized by unit variance and used for further analyses.

Data analysis

HCA was performed using Ward's method and was implemented using the "hclust" function in the R software package (version 3.2.2) [23]. SVM, random forest (RF) [24], and PLS were performed using the e1071 [14], randomForest [14], and pls [25] libraries, respectively, with repeated (10 times) double cross-validation (CV) [2] in R. In the double CV, threefold CV and fivefold CV were used for the hyperparameter optimization and performance evaluation of the constructed models, respectively (Fig. 1). The SVM hyperparameters were evaluated at the ranges of 0.0001 to 1 for gamma and 1 to 1000 for cost, and the most frequently determined values in the hyperparameter optimization process of the double CV were 0.03 and 5 for gamma and cost, respectively.

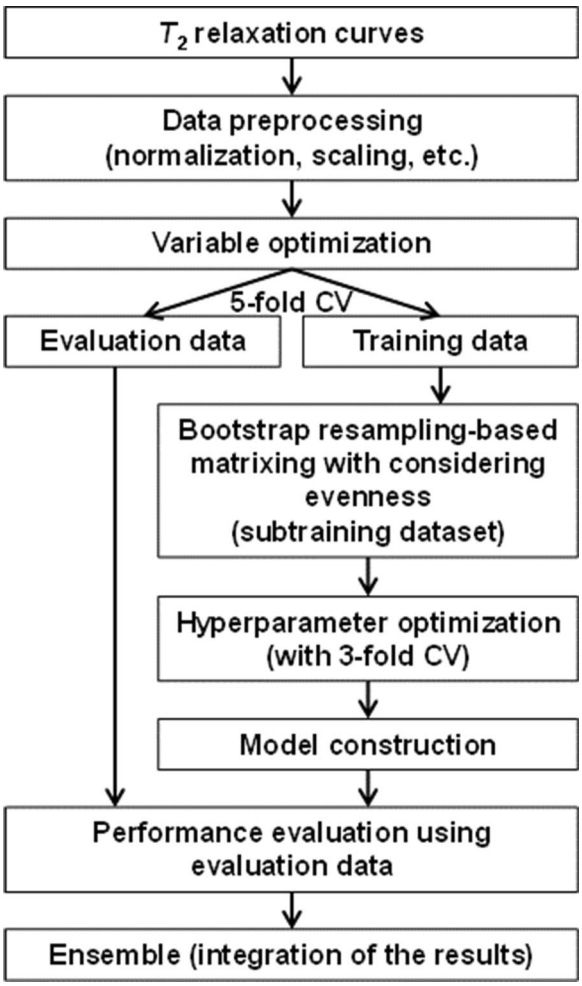

Fig. 1 Analytical flow of relaxation curve data during relaxometric learning. Compressive force data analyzed by HCA were used as explanatory variables for the analyses of T_2 relaxation curves in a data-driven manner. The relaxation curves were preprocessed and analyzed by ML incorporating an analytical framework including a variable optimization approach and bootstrap resampling-based matrixing. The results obtained for each classifier were finally integrated, similar to that in ensemble learning

Receiver operating characteristic (ROC) curves and the area under the curve (AUC) were calculated using the ROCR library [26].

Results and discussion

Classification performance of SVM for T_2 relaxation curve data

The applicability of ML algorithms to the data mining of T_2 relaxation curves was firstly evaluated because such analyses have not been reported in the literature, and several ML approaches have shown relatively good performance in classification problems compared with multivariate analyses [14]. Conventional SVM classification, which is a typical ML method, was performed using category information from two groups of compressive force data as explanatory variables for the T_2 relaxation curve data (a total of 627 curves) of various fish muscle samples collected from multiple coastal sites in Japan. Unfortunately, the classification performance was worse than expected; the AUC, accuracy, correctly classified rate for group A (CCR-A), and correctly classified rate for group B (CCR-B) were 0.780, 0.748, 0.475, and 0.865, respectively (Additional file 1: Table S2).

Variable optimization to enhance the classification performance

To improve the SVM classification performance, we used a variable optimization approach to enhance the quality of information obtained from simple T_2 relaxation curve data. The variable optimization approach employed was a search method that determined the best length of raw curve to improve the classification performance via the reduction of variables from long- to short-relaxation-time components in sequential order. This variable reduction idea is based on the elimination of "noise" variables, which possibly arise from background noise and/or from free water, i.e., relatively long T_2 relaxation time components in the relaxation curve are barely related to the characteristic features of samples and were suspected of interfering with the accuracy of the SVM learning step.

In this study, the variables were gradually reduced at the rate of 10% from all variables (the number of variables was 256) to 10% usage rate (25 variables) by omitting the relatively long T_2 relaxation time components. For instance, the dataset at 90% usage rate included 230 variables (from 0 to 0.92 s of T_2 relaxation time components) and more than 0.92 s of the components was removed. The each dataset generated by the variable reductions was applied to performance evaluation by SVM classification. Based on the classification performance, the optimized number of variables was determined.

The incorporation of the variable optimization approach into the SVM classification method improved the classification performance for the two groups (Fig. 2). The best SVM classification performance was obtained with a 10% reduction (90% usage rate) of data points (variables) from the T_2 relaxation curve, with AUC and accuracy values of 0.806 and 0.779. Although the value of CCR-A was also improved from 0.475 to 0.538, the relatively low value still required improvements to obtain satisfactory classification performance.

Bootstrap resampling to balance the sample size between groups

The classification performances of conventional SVM with and without the above variable optimization approach were profoundly affected by the biased sample size between groups in the data matrix of the T_2 relaxation curve. To circumvent this difficulty, we focused on a bootstrap resampling-based matrixing that constructs a matrix (subtraining dataset) by considering the evenness of the sample size between groups for the model construction of ML classifications (Fig. 1). In this approach, the hyperparameters (i.e., the number of resamples for balanced matrixing and the number of

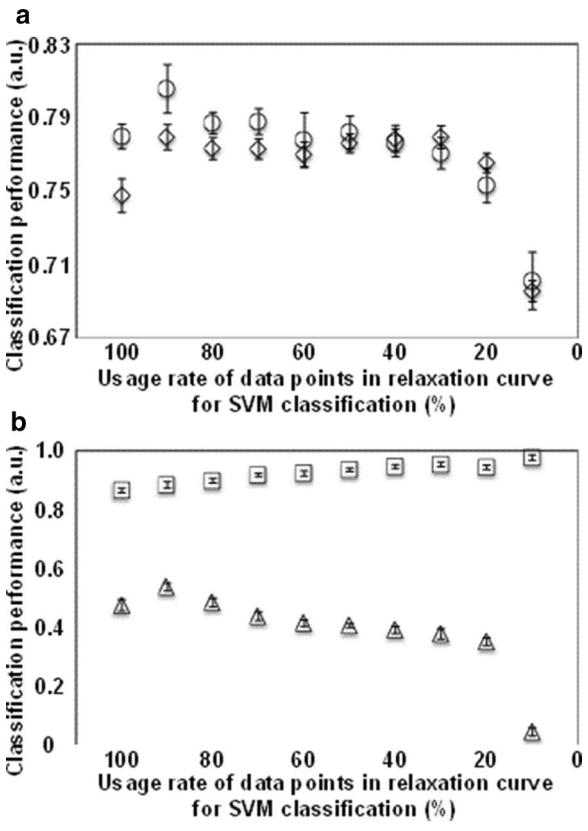

Fig. 2 Performance evaluation of the variable optimization approach based on conventional SVM classifications. **a** AUC (circles) and accuracy (diamonds). **b** CCR-A (triangles) and CCR-B (squares)

datasets generated by bootstrap resampling) were evaluated (Fig. 3). When the number of generated datasets was set to 100, a slightly better performance was obtained in terms of AUC and accuracy compared with only 10 generated datasets, whereas different resample sizes yielded almost the same AUC and accuracy values, except for resample sizes below 50 per group (a total of 100 samples per dataset). Furthermore, the CCR-A slightly decreased with increasing resamples. Therefore, we performed variable optimization combined with bootstrap resampling-based matrixing by using 100 generated datasets with 150 resamples (a total of 300 samples) per dataset (Additional file 1: Fig. S4). At the 90% level of variable usage rates for T_2 relaxation curves exhibiting the best SVM classification performance, the classification performance of the constructed analytical framework was significantly improved compared with conventional SVM in terms of the ROC curve and the AUC value (Fig. 4). The AUC, accuracy, and CCR-A values significantly increased from

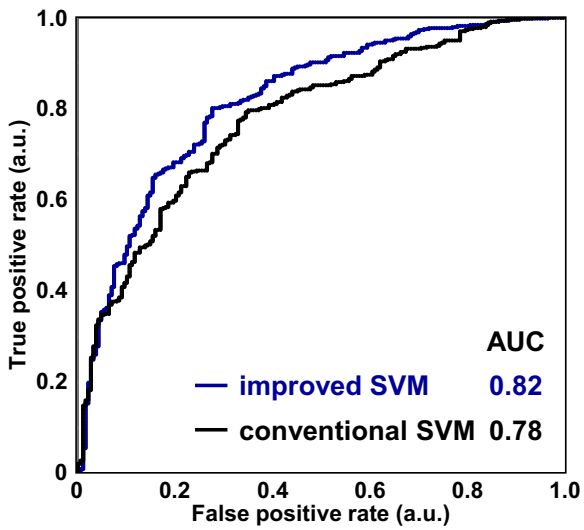

Fig. 4 ROC curves and the corresponding AUC values obtained using conventional SVM (black) and SVM-based relaxometric learning (blue)

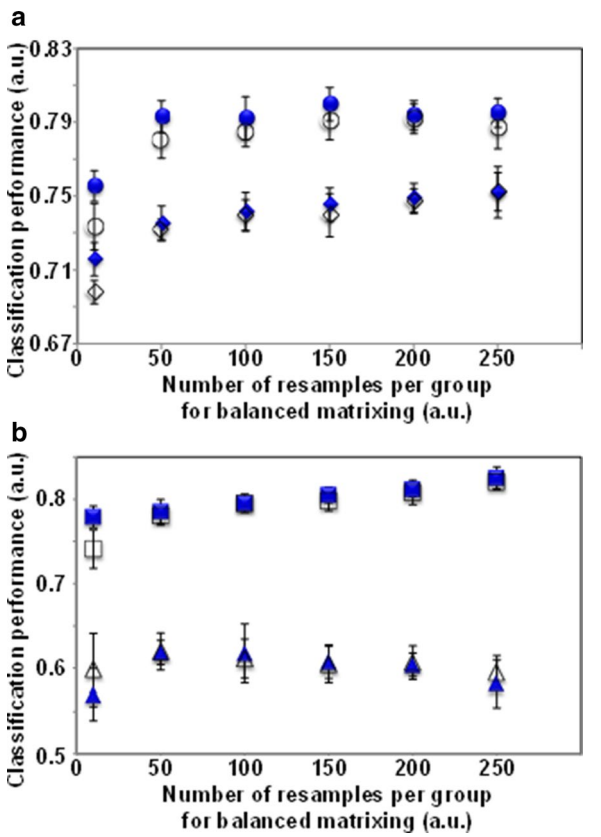

Fig. 3 Performance evaluation of the bootstrap resampling-based matrixing approach based on conventional SVM classifications. The performance of the approach was evaluated with the number of datasets generated by the bootstrap resampling set at 100 (closed blue symbols) and 10 (open symbols) matrices. **a** AUC (circles) and accuracy (diamonds). **b** CCR-A (triangles) and CCR-B (squares)

0.780, 0.748, and 0.475 to 0.820, 0.771, and 0.710, respectively (Additional file 1: Fig. S5). In addition, robustness evaluation of the developed method for fluctuation of each variable was performed using datasets generated by random resampling based on permutation for a variable (Additional file 1: Fig. S6). The fluctuation of each variable had relatively little effect on the SVM classification performance using the analytical framework developed in this study, indicating that the developed method enables to construct robust models for variable fluctuation. Therefore, the analytical framework described here, namely, the incorporation of the variable optimization approach and the bootstrap resampling-based matrixing into the ML calculations, resulted in improved classification performance in the data mining of T_2 relaxation curve data and enhanced the robustness when using unbalanced datasets. Furthermore, the relaxometric learning method developed here enabled the extraction of features related to the physical properties (hardness and tenderness) of fish muscle.

Applicability of the analytical framework to other machine learning methods

The analytical framework developed in this study is optimized for data mining of T_2 relaxation curve data. Thus, we considered that not only SVM but also other ML algorithms and multivariate analyses may be useful for the data mining of T_2 relaxation curve data. To test this hypothesis, RF and PLS were used as alternatives to SVM for data mining based on the developed analytical framework. The classification performance of RF improved

slightly in terms of the ROC curve and the AUC value, and the CCR-A values were significantly improved by the incorporation of our analytical framework (Fig. 5 and Additional file 1: S7). On the other hand, the classification performance of PLS improved drastically in terms of the values of both AUC and CCR-A (Fig. 5 and Additional file 1: S7). These results suggest that our analytical framework is applicable to various ML algorithms and multivariate analyses to enhance classification performance, but the extent of improvement is method dependent. Therefore, the relaxometric learning approach developed in this study should find use as a versatile and useful method for the analysis of T_2 relaxation curve data.

Applicability of relaxometric learning to the determination of geographical origin

NMR-based metabolomics approaches are capable of determining the geographical origins of food products such as fish [14, 16, 27–29], beef [30], durum wheat [31], white rice [32], apple [33], cabbage [34], honey [35], coffee beans [36], and wine [37]. Therefore, relaxometric learning was also considered applicable for such analyses. In addition, the quality of biological tissue (such as water content and water activity) in different environment is varied according to the geographical origins [38]. Here, we proposed that the difference in the water conditions could be detected by pattern recognition of T_2 relaxation curves using NMR but not high performance liquid chromatography or mass spectrometry methods. Then, we performed experiments to evaluate the applicability of relaxometric learning to discriminate between geographical differences (i.e., to extract features in terms of the habitats of *Girella punctata* belonged to Kyphosidae fish living in Tokyo Bay and Sagami Bay) based on T_2 relaxation curves (Additional file 1: Fig. S8). The SVM-based relaxometric learning method exhibited relatively good performance in the geographical origin discrimination of fish compared with conventional SVM, thus leading to a significant increase in AUC value from 0.886 to 0.936 (Additional file 1: Fig. S8). These results suggest that relaxometric learning is applicable as a method for determining geographical provenance, solving problems related to food fraud, and certifying the "terroir" of food, similar to the case for conventional metabolomics approaches.

Compact benchtop or portable NMR spectrometers are low-cost alternatives to conventional high-field and high-resolution spectrometers. Benchtop low-field NMR spectrometers can theoretically obtain T_2 relaxation curves with a similar quality to those obtained with high-resolution NMR spectrometers, such as the one used in this study; therefore, similar classification accuracies can be expected. Relaxometric learning using benchtop and portable NMR spectrometers even without using D_2O might also find applications in on-site quality control and freshness management, optimization of production processes, and improvement of product quality not only in food but also in various industrial fields such as polymers, cosmetics, fabrics, pharmaceuticals, and healthcare. Relaxometric learning is expected to be a versatile and powerful approach for the characterization and evaluation of industrial products and as an option for biological and chemical research that requires a nondestructive, cost-effective, and time-saving method.

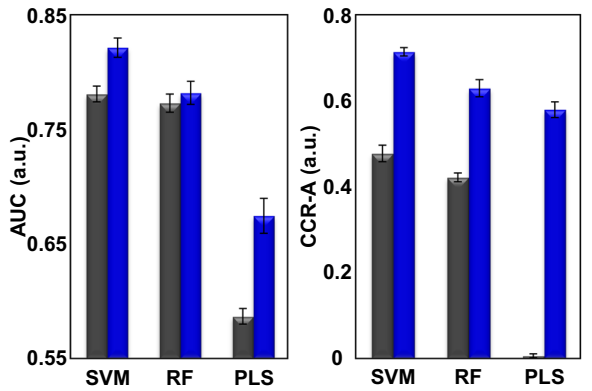

Fig. 5 Comparison of classification performance between relaxometric learning (blue bars) and conventional MLs (black bars). The SVM, RF, and PLS algorithms were employed. The standard deviations are displayed as error bars. Left panel, AUC; right panel, CCR-A

Conclusions

This study focused on the development of a new relaxometric learning method based on the pattern recognition of T_2 relaxation curves. The method is supported by an analytical framework incorporating a variable optimization approach and bootstrap resampling-based matrixing to enhance the classification performance of the ML algorithms employed. The developed relaxometric learning approach enabled the extraction of several features of fish muscles, such as their physical properties and geographical origin, from T_2 relaxation data. Relaxometric learning was also implemented with not only SVM but also other ML and multivariate methods for the analysis of T_2 relaxation data. The SVM-based relaxometric learning method was superior to the conventional SVM method, thus indicating that the analytical framework constructed in this study enables better classification performance when ML algorithms are applied to relaxation curve data. The relaxometric learning approach is a versatile, cost-effective, and time-saving tool for characterizing physical properties,

such as the fleshiness of fish muscles; for evaluating product qualities, such as geographical origin; and for food authentication in biological and chemical samples.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13065-020-00731-0.

Additional file 1: Table S1. List of fish samples used in this study, **Table S2.** Classification performance of conventional SVM, **Figure S1.** Representative T_2 relaxation curves for the various fish samples used in this study, **Figure S2.** Analytical procedure involving compressive force measurements by autograph, **Figure S3.** Categorization of compressive force data based on a data-driven approach, **Figure S4.** Performance evaluation of variable optimization approach with bootstrap resampling-based matrixing, **Figure S5.** Classification performance of SVM-based relaxometric learning, **Figure S6.** Robustness evaluation of the relaxometric learning for fluctuation of each variable. **Figure S7.** ROC curves and the corresponding AUC values for each method, **Figure S8.** Classification performance of SVM-based relaxometric learning in determining the geographical differences between Kyphosidae taken from Tokyo Bay and Sagami Bay.

Abbreviations

AUC: Area under the curve; CCR-A: Correctly classified rate for group A; CCR-B: Correctly classified rate for group B; CV: Cross-validation; DNN: Deep neural network; HCA: Hierarchical cluster analysis; NMR: Nuclear magnetic resonance; ML: Machine learning; PLS: Partial least squares; RF: Random forest; ROC: Receiver operating characteristic; SVM: Support vector machine.

Acknowledgements

The authors would like to thank Y. Otaka, T. Shimizu, Y. Katsuki, S. Fujinuma, and A. Tei (RIKEN) for assisting with the sample preparation, processing, and data collection. The authors would like to thank Enago (www.enago.jp) for the English language review.

Authors' contributions

YD and JK conceived and designed the experiments; YD, KS, and JK provided the reagent and the fish samples; FW, YT, and KS performed the experiments; YD, FW, and KI analyzed the data; and YD and FW wrote the manuscript with contributions from all authors. All authors read and approved the manuscript.

Funding

This work was partially supported by the Agriculture, Forestry and Fisheries Research Council (to J.K.).

Availability of data and materials

The complete database is accessible in the website http://dmar.riken.jp/NMRinformatics/.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest.

Author details

[1] RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [2] Graduate School of Medical Life Science, Yokohama City University, 1-7-29 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [3] Graduate School of Bioagricultural Sciences, Nagoya University, 1 Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan.

Received: 22 May 2020 Accepted: 15 December 2020
Published online: 20 February 2021

References

1. Erikson U, Standal IB, Aursand IG, Veliyulin E, Aursand M (2012) Use of NMR in fish processing optimization: a review of recent progress. Magn Reson Chem 50(7):471–480
2. Kikuchi J, Ito K, Date Y (2018) Environmental metabolomics with data science for investigating ecosystem homeostasis. Prog Nucl Magn Reson Spectrosc 104:56–88
3. Cubero-Leon E, Penalver R, Maquet A (2014) Review on metabolomics for food authentication. Food Res Int 60:95–107
4. Grootveld M, Percival B, Gibson M, Osman Y, Edgar M, Molinari M, Mather ML, Casanova F, Wilson PB (2019) Progress in low-field benchtop NMR spectroscopy in chemical and biochemical analysis. Anal Chim Acta 1067:11–30
5. van Duynhoven J, Voda A, Witek M, Van As H (2010) Time-domain NMR applied to food products. Annu Rep Nmr Spectro 69:145–197
6. Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J (2006) OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J Chemom 20(8–10):341–351
7. Bertram HC, Straadt IK, Jensen JA, Aaslyng MD (2007) Relationship between water mobility and distribution and sensory attributes in pork slaughtered at an age between 90 and 180 days. Meat Sci 77(2):190–195
8. Pereira FM, Bertelli Pflanzer S, Gomig T, Lugnani Gomes C, de Felicio PE, Colnago LA (2013) Fast determination of beef quality parameters with time-domain nuclear magnetic resonance spectroscopy and chemometrics. Talanta 108:88–91
9. Pereira FMV, Carvalho AD, Cabeca LF, Colnago LA (2013) Classification of intact fresh plums according to sweetness using time-domain nuclear magnetic resonance and chemometrics. Microchem J 108:14–17
10. Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML, Goodacre R (2015) A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. Anal Chim Acta 879:10–23
11. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, Wishart DS, Xia J (2018) MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 46(W1):W486–W494
12. Cacciatore S, Luchinat C, Tenori L (2014) Knowledge discovery by accuracy maximization. Proc Natl Acad Sci U S A 111(14):5117–5122
13. Ito K, Obuchi Y, Chikayama E, Date Y, Kikuchi J (2018) Exploratory machine-learned theoretical chemical shifts can closely predict metabolic mixture signals. Chem Sci 9(43):8213–8220
14. Date Y, Kikuchi J (2018) Application of a deep neural network to metabolomics studies and its performance in determining important variables. Anal Chem 90(3):1805–1810
15. Asakura T, Date Y, Kikuchi J (2018) Application of ensemble deep neural network to metabolomics studies. Anal Chim Acta 1037:230–236
16. Asakura T, Sakata K, Date Y, Kikuchi J (2018) Regional feature extraction of various fishes based on chemical and microbial variable selection using machine learning. Anal Methods 10(18):2160–2168
17. Oita A, Tsuboi Y, Date Y, Oshima T, Sakata K, Yokoyama A, Moriya S, Kikuchi J (2018) Profiling physicochemical and planktonic features from discretely/continuously sampled surface water. Sci Total Environ 636:12–19
18. Shima H, Masuda S, Date Y, Shino A, Tsuboi Y, Kajikawa M, Inoue Y, Kanamoto T, Kikuchi J (2017) Exploring the impact of food on the gut ecosystem based on the combination of machine learning and network visualization. Nutrients 9(12):1307
19. Shiokawa Y, Date Y, Kikuchi J (2018) Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet. Sci Rep 8(1):3426
20. Noble WS (2006) What is a support vector machine? Nat Biotechnol 24(12):1565–1567
21. Heinemann J, Mazurie A, Tokmina-Lukaszewska M, Beilman GJ, Bothner B (2014) Application of support vector machines to metabolomics experiments with limited replicates. Metabolomics 10(6):1121–1128
22. Andersen CM, Rinnan Å (2002) Distribution of water in fresh cod. LWT-Food Sci Technol 35(8):687–696

23. Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J Classif 31(3):274–295
24. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
25. Mevik BH, Wehrens R (2007) The pls package: Principal component and partial least squares regression in R. J Stat Softw 18(2):1–23
26. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21(20):3940–3941
27. Aursand M, Standal IB, Prael A, McEvoy L, Irvine J, Axelson DE (2009) (13) C NMR pattern recognition techniques for the classification of Atlantic salmon (*Salmo salar* L.) according to their wild, farmed, and geographical origin. J Agric Food Chem 57(9):3444–3451
28. Yoshida S, Date Y, Akama M, Kikuchi J (2014) Comparative metabolomic and ionomic approach for abundant fishes in estuarine environments of Japan. Sci Rep 4:7005
29. Wei F, Sakata K, Asakura T, Date Y, Kikuchi J (2018) Systemic homeostasis in metabolome, ionome, and microbiome of wild yellowfin goby in estuarine ecosystem. Sci Rep 8(1):3478
30. Jung Y, Lee J, Kwon J, Lee KS, Ryu DH, Hwang GS (2010) Discrimination of the geographical origin of beef by (1)H NMR-based metabolomics. J Agric Food Chem 58(19):10458–10466
31. Lamanna R, Cattivelli L, Miglietta ML, Troccoli A (2011) Geographical origin of durum wheat studied by 1H-NMR profiling. Magn Reson Chem 49(1):1–5
32. Long NP, Lim DK, Mo C, Kim G, Kwon SW (2017) Development and assessment of a lysophospholipid-based deep learning model to discriminate geographical origins of white rice. Sci Rep 7(1):8552
33. Tomita S, Nemoto T, Matsuo Y, Shoji T, Tanaka F, Nakagawa H, Ono H, Kikuchi J, Ohnishi-Kameyama M, Sekiyama Y (2015) A NMR-based, non-targeted multistep metabolic profiling revealed L-rhamnitol as a metabolite that characterised apples from different geographic origins. Food Chem 174:163–172
34. Kim J, Jung Y, Song B, Bong YS, Ryu DH, Lee KS, Hwang GS (2013) Discrimination of cabbage (*Brassica rapa* ssp. pekinensis) cultivars grown in different geographical areas using (1)H NMR-based metabolomics. Food Chem 137(1–4):68–75
35. Schievano E, Stocchero M, Morelato E, Facchin C, Mammi S (2012) An NMR-based metabolomic approach to identify the botanical origin of honey. Metabolomics 8(4):679–690
36. Wei F, Furihata K, Koda M, Hu F, Kato R, Miyakawa T, Tanokura M (2012) (13)C NMR-based metabolomics for the classification of green coffee beans according to variety and origin. J Agric Food Chem 60(40):10118–10125
37. Godelmann R, Fang F, Humpfer E, Schutz B, Bansbach M, Schafer H, Spraul M (2013) Targeted and nontargeted wine analysis by (1)h NMR spectroscopy combined with multivariate statistical analysis. Differentiation of important parameters: grape variety, geographical origin, year of vintage. J Agric Food Chem 61(23):5610–5619
38. Ribeiro RDR, Marsico ET, Carneiro CD, Monteiro MLG, Conte CA, Mano S, de Jesus EFO (2014) Classification of Brazilian honeys by physical and chemical analytical methods and low field nuclear magnetic resonance (LF H-1 NMR). Lwt-Food Sci Technol 55(1):90–95

Publisher's Note