Prediction of Sake Component Values Using E-nose and E-tongue Data by Machine Learning

Satoru SHIMOFUJI1,2,†, Motoko MATSUI2, Yukari MURAMOTO3, Hironori MORIYAMA1, Yoshiro HOKI1, Haruhiko UEHIGASHI1

1Kochi Prefecture Industrial Technology Center, 3992-3, Nunoshida, Kochi-shi, Kochi 781-5101, Japan
2Department of Food Sciences and Nutritional Health, Graduate School of Kyoto Prefectural University, 1-5 Shimogamohangi-cho, Sakyo-ku, Kyoto 606-8522, Japan
3Kyoto Prefectural University, 1-5 Shimogamohangi-cho, Sakyo-ku, Kyoto 606-8522, Japan

We estimated the quality component values of the commercial Japanese sake Junmai Ginjo by using electronic (e)-nose and e-tongue data. Regression analysis methods were applied to predict the components. Characteristic features of Junmai Ginjo such as acidity, amino acid content, glucose and nine volatile components were used as objective variables. Explanatory variables were the 99 peak data obtained by an e-nose and seven sensor data obtained by an e-tongue. The prediction accuracy by the partial least squares regression method using e-nose and e-tongue data was 7.57 average error% (the ratio of the mean absolute error to the component value range). With the application of other regression analyses (multiple regression analysis, support-vector machine, random forest, gradient boosting), the prediction accuracy was improved for all components except the acidity and amino acid content. With the application of other regression analyses and the addition of the data of seven simplified analyses (Brix, pH, electrical conductivity, OD260, OD280, simplified alcohol content, simplified glucose content), the prediction accuracy was improved for all components. (average error%: 5.04) The analysis conditions (i.e., the regression analysis and the dataset of explanatory variables) for the best score differed depending on the component. Thus, when predicting components by a regression analysis, it is necessary to prepare a plurality of analysis conditions and challenges.

Keywords: e-nose, e-tongue, component prediction, machine learning, Junmai Ginjo

1. Introduction

Sake is a traditional Japanese favorite alcoholic beverage that is made from rice through fermentation and filtration. For the fermentation of sake, rice koji, a type of fungi grown on rice, is used. Sake has been made for over 1,000 years all over the Japanese islands. There are unique regional types of sake with different tastes. Sake is made using a unique brewing process called parallel compound fermentation, and due to the presence of fermented products, the composition of sake is complex.

As a general indicator of the quality of sake, volatile components (such as ethyl acetate, isoamyl alcohol, isoamyl acetate, ethyl caproate), the glucose content, acidity and amino acid content are analyzed.

Researches had been conducted to clarify the sensory characteristics of sake by using these analytical values [1-2]. We have applied machine learning to predict a comprehensive evaluation from physicochemical features that characterize sake [3], and our study showed that the prediction accuracy was low—especially for the samples evaluated as having defects. We speculated that it is necessary to add the data of a comprehensive analysis to the explanatory variables in order to predict comprehensive evaluations more accurately. A metabolome analysis has also been used to analyze the Zatsumi of sake (Zatsumi is the inverse of a sense of clarity, it also indicates unsmoothed and unpleasant attributes, that is, negative sensory characteristics that produce a detracting flavor in clear beverages.) [4]. As noted above, a comprehensive analysis is performed as a sensory evaluation analysis.

Similarly, an electronic (e)-nose and e-tongue are considered to be capable of providing comprehensive analyses. An e-nose and e-tongue can be analyzed with a relatively simplified operation, but when foods with complex components are analyzed, it is possible that clear data...
may not be obtained because of noise due to interactions between components. It is necessary to supplement data for components that cannot be measured by an e-nose or e-tongue.

Several studies have confirmed the relevance of component values of foods by combining e-nose and e-tongue data [5–8]. For sake, there are studies using taste sensors [9], but none using e-nose data. It is thus necessary to confirm whether the data obtained from an e-nose and an e-tongue comprehensively capture the characteristics of sake.

The purpose of this study is to more accurately capture the ingredients of sake with e-nose and e-tongue. Therefore, we examined a more appropriate combination of explanatory variables and regression analysis methods. We estimated the quality component values using e-nose and e-tongue data, and using the error of the component prediction values, we estimated how many of the components of the e-nose and e-tongue data can be analyzed comprehensively.

We applied regression analysis methods to predict the components of sake. The regression analysis used machine learning as well as the commonly used statistical methods partial least squares regression (PLS) and a multiple regression analysis (MRA). MRA has easy to understand the contribution of the explanatory variables by multiple regression coefficient. However, there is a problem that the prediction accuracy tends to decrease due to multicollinearity caused by using an explanatory variable having a high correlation with other variables. A PLS is relatively easy to perform because it does not require processing such as variable selection [10]. A PLS is also included in the analysis software of the e-nose and e-tongue used herein.

On the other hand, machine learning is suitable for analyzing complex relationships, and it is also used in food science. Various methods have been used for machine learning; typical methods are the support-vector machine (SVM) and random forest (RF) methods [11,12]. Ensemble learning, called ‘gradient boosting (GB),’ has been attracting attention as a machine-learning technique that is considered to have high prediction accuracy [13]. Using the information (error) of the model learned thus far to build a new model and then preceeding with data learning is called ‘boosting’. The open-source software library XGBoost is one of the gradient boosting tools [14]. In XGBoost, when boosting is being performed (creating a new decision tree using the results of the previous decision tree), an algorithm of the decision tree is constructed so that the error between the measured value and the predicted value is minimized [14]. One of the characteristics of comprehensive analysis data is its complexity due to the large amount of data. Comprehensive analysis data will be more useful if techniques such as machine learning are improved to better predict more complex relationships than can be done with traditional statistical techniques.

In addition, in the present study we attempted to improve the prediction accuracy by using analytical equipment and data that can be relatively easily applied to assist analyses based on data obtained with an e-nose and e-tongue.

We sought to identify the following: (1) the prediction components of sake by conducting a PLS using e-nose and e-tongue data, (2) the effects of other regression analyses (MRA and machine learning) on the accuracy of the prediction, (3) the effects of adding simplified analyses on the prediction accuracy afforded by a PLS, and (4) the effects of combining items 2 and 3 above on the prediction accuracy.

2. Materials and Methods

2.1 Samples

We used 179 samples of the commercial Japanese sake Junmai Ginjo, which is made only with rice (its polished rice ratio is < 60%), rice koji (i.e., rice malt or yeast), and water (without the addition of brewer’s alcohol). There are no duplicate brands in 179 products. These samples were produced in the 2018 brewing year. We received these samples from Liquor shop (Hasegawa Saketen, Inc.). The samples were stored at ~1C for quality preservation until they were analyzed. The samples were opened immediately before being subjected to analysis.

2.2 Physicochemical Analysis

2.2.1 General analyses

We measured the sake samples’ acidity, amino acid content, and glucose content as indicators of the water-soluble components of the sake. The acidity and amino acid content were analyzed by the methods authorized by Japan’s National Tax Agency [15]. The glucose content was determined by the mutarotase glucose oxidase method [16] using the Wako Glucose CII-test (Wako Pure Chemicals, Osaka, Japan).

2.2.2 Volatile components

Volatile components were measured as aspects of the flavor of the sake. Acetaldehyde, ethylacetate, propan-
1-ol (nProOH), isobutanol (iBuOH), isoamylacetate, 3-methyl-1-butanol (iAmiOH), ethyl caproate, ethyl caprylate, and acetic acid were measured by head-space gas chromatography (GC) (model 7890B, Agilent, Santa Clara, CA) according to the method of Yoshizawa et al. [17]. First, 2 mL of sample and 200 μL of internal standard solution (2000 ppm n-amyl alcohol, 200 ppm methylcaproate) were placed in a vial and sealed with an aluminum cap. Head space GC was then used to analyze these components.

2.2.3 The e-nose system

The e-nose used was a flash GC e-nose (HERACLES, Alpha M.O.S., Toulouse, France). Seven types of potentiometric chemical sensors were used for analysis: AHS, PKS, CTS, NMS, CPS, ANS and SCS (Alpha M.O.S.). Each sample was analyzed seven times, and the analysis was performed using the average of the analysis values at the fifth, sixth, and seventh times when the values were stable.

2.2.4 The e-tongue system

The e-tongue data were obtained using an electronic taste system (ASTREE, Alpha M.O.S.). Seven types of potentiometric chemical sensors were used for analysis: AHS, PKS, CTS, NMS, CPS, ANS and SCS (Alpha M.O.S.). Each sample was analyzed seven times, and the analysis was performed using the average of the analysis values at the fifth, sixth, and seventh times when the values were stable.

2.2.5 The simplified analyses

As simplified analyses, the samples’ Brix values were measured by a Brix meter (PAL-J, Atago, Tokyo, Japan); the pH was measured by a pH meter (LAQUAtwin-pH22, Horiba, Fukuoka, Japan), and the electrical conductivity was measured by an electrical conductivity meter (LAQUAtwin-EC-33, Horiba).

The optical density (OD) values at 260 and 280 nm were measured as described by Yamanaka et al. [18]. The OD values are indicators showing the degree of leakage of nucleic acid components due to the death of yeast during the manufacturing process, and they are thought to affect the quality evaluation. The simplified alcohol content was measured by an alcohol densitometer (Alcomate AL2, Woodson Riken Keiki, Tokyo, Japan).

The simplified glucose content was measured using a standard handheld glucometer; the Glucocard and a Gsensor (ARKRAY, Kyoto, Japan). Each sample was diluted 10 times with distilled water before the measurement.

The data of all of the above parameters were standardized to a zero mean and one standard deviation (SD) for the regression analyses.

2.3 Regression analyses

We performed separate regression analyses using PLS, an MRA, and machine learning, i.e., SVM, RF, and GB.

The e-nose, e-tongue, and simplified analyses’ values were used as explanatory variables. In an MRA, there is a problem of multicollinearity by using explanatory variables with high correlation between other variables. However, in the present analysis, no variables were selected in order to match the analysis conditions with the other regression analyses. We did this because the purpose of the study was not to confirm the contribution of the explanatory variables, but rather to focus on the comparison of the prediction accuracy with as simplified an analysis as possible.

The general analysis and volatile components data were used as the objective variable.

The regression analyses were carried out by R software [i]. The R development environment is a multi-platform open-source environment that can be operated by Windows and Linux. R is a high-level matrix programming language for statistical processing and data analysis.

All of the regression methods used herein were implemented using the R package caret [19], which provides a good interface to access many machine-learning algorithms in R. The caret package tunes the parameters of statistical analyses and machine learning, and it selects the values that maximize the accuracy according to the validation selected [20]. The method parameters of caret are as follows: for an MRA, lm (stats package) [ii]; for PLS, pls (pls package) [21]; for SVM, svmRadial (kernlab package) [22]; for RF, rf (random forest package) [23]; and for GB, xgbTree (xgboost package) [iii].

In each of the machine learning methods, we performed 10-fold cross-validation to generalize the model, and a prediction model was created. Other analysis condition settings used the default settings in the caret package.

2.4 The accuracy of the regression analyses

The accuracy of the regression analyses was verified by dividing the data obtained into train data and test data to avoid over-learning. First, 141 samples (80% of the total of 179 samples) were randomly used as the training data, and the remaining 38 samples were used as the test data for accuracy verification. Next, a regression analysis was performed using the training data to obtain a prediction model. From the obtained prediction model, we calculated the prediction value of the test data (the data not
used to create the prediction model. We compared the prediction value with the actual measurement value, and we examined the accuracy of each analysis method.

We evaluated the prediction accuracy by using the mean absolute error (MAE) and adjusted R-square ($\text{adjR}^2$). We also calculated how large the MAE was with respect to the original data range, and we set it as the error%.

3. Results and Discussions

3.1 Physicochemical Analyses

3.1.1 Objective variables

Table 1 provides the results of the general analysis and volatile components. These results were used as the objective variables.

Table 2 shows statistics of objective variables for training and test data. It shows that all the variables were randomly divided into training data and test data so that they had the same distribution as the original data.

3.1.2 Explanatory variables

Table 3 shows the analysis results obtained with the e-nose; 99 peaks were detected as valid peaks. Table 4 provides the analysis results obtained with the e-tongue. The results of the simplified analyses are listed in Table 5. All of these data were used in the prediction as explanatory variables.

| Table 1 | The general analyses and volatile components data of Junmai Ginjo ($n=179$). |
|--------|------------------|
| Variable | Mean ± SD | Min. | Max. |
| General analyses | | | |
| Acidity (mL) | 1.47 ± 0.17 | 1.01 | 1.88 |
| Amino acid content (mL) | 0.96 ± 0.21 | 0.52 | 1.57 |
| Glucose (%) | 1.96 ± 0.63 | 0.39 | 3.98 |
| Volatile components | | | |
| Acetaldehyde (ppm) | 30.09 ± 7.45 | 16.30 | 62.01 |
| Ethyl acetate (ppm) | 60.66 ± 18.93 | 24.86 | 105.51 |
| nProOH (ppm) | 59.29 ± 15.63 | 24.27 | 114.30 |
| iBuOH (ppm) | 36.36 ± 8.54 | 22.14 | 82.10 |
| Isoamyl acetate (ppm) | 2.86 ± 1.51 | 1.04 | 9.18 |
| iAmiOH (ppm) | 117.33 ± 14.66 | 80.41 | 170.48 |
| Ethyl caproate (ppm) | 5.47 ± 3.23 | 0.07 | 15.35 |
| Ethyl caprylate (ppm) | 1.12 ± 0.47 | 0.00 | 2.29 |
| Caproic acid (ppm) | 32.32 ± 22.23 | 0.00 | 109.71 |

| Table 2 | Statistics of objective variables for training and test data. |
|--------|------------------|
| Variable | training data $n=141$ | test data $n=38$ |
| | Mean ± SD | Mean ± SD |
| General analyses | | | |
| Acidity (mL) | 1.48 ± 0.17 | 1.46 ± 0.16 |
| Amino acid content (mL) | 0.95 ± 0.20 | 0.97 ± 0.22 |
| Glucose (%) | 1.97 ± 0.62 | 1.93 ± 0.70 |
| Volatile components | | | |
| Acetaldehyde (ppm) | 30.29 ± 7.64 | 29.65 ± 6.87 |
| Ethyl acetate (ppm) | 60.66 ± 18.93 | 59.47 ± 20.08 |
| nProOH (ppm) | 59.29 ± 15.63 | 59.07 ± 17.62 |
| iBuOH (ppm) | 36.36 ± 8.54 | 35.64 ± 8.26 |
| Isoamyl acetate (ppm) | 2.86 ± 1.50 | 2.91 ± 1.59 |
| iAmiOH (ppm) | 117.33 ± 14.66 | 118.38 ± 16.28 |
| Ethyl caproate (ppm) | 5.47 ± 3.23 | 5.64 ± 3.43 |
| Ethyl caprylate (ppm) | 1.12 ± 0.47 | 1.12 ± 0.45 |
| Caproic acid (ppm) | 32.32 ± 22.31 | 31.35 ± 22.37 |
Table 3  The e-nose data statistics of Junmai Ginjo ($n=179$).

| RT/C*   | Min.  | Mean   | Max.  | RT/C*   | Min.  | Mean   | Max.  |
|---------|-------|--------|-------|---------|-------|--------|-------|
| 15.65/1 | 121.27| 207.60 | 530.95| 14.92/2 | 0.00  | 6.28   | 719.49|
| 16.09/1 | 0.00  | 1.69   | 138.98| 15.33/2 | 198.27| 296.52 | 1831.80|
| 16.66/1 | 180.00| 571.30 | 1459.23| 16.05/2 | 698.92| 1318.05| 1738.61|
| 17.4/1  | 15580.5| 24782.1| 39430.26| 18.03/2 | 14.74 | 110.54 |       |
| 20.84/1 | 398059.37| 3421475.62| 6894285.97| 19.98/2 | 3874.37| 7536.23|       |
| 224.21/1 | 64.09  | 11471.85|       |         |       |        |       |
3.2 The prediction accuracy of components for test data by the regression analyses

3.2.1 Prediction by PLS using e-nose and e-tongue data

The data in Table 6 describe the accuracy of predicting components by PLS using e-nose and e-tongue data. In this study, the analysis condition with the smallest error percentage was used as the analysis condition for the best score. For the sake samples’ acidity and amino acid and glucose contents, the prediction accuracy was relatively lower than that for the volatile components, as follows. Error%: acidity 11.49, amino acid content 11.43, glucose 12.24; adjR²: acidity 0.21, amino acid content 0.28, glucose 0.23. The prediction accuracy of volatile components was high. The following data of the error% for volatile components was obtained. Error%: min 4.55 – max 8.60, adjR²: min 0.73 – max 0.90.

From the prediction results provided by PLS, it was apparent that the volatile components were well-covered by the analysis is using e-nose and e-tongue data. The best score dataset was e-nose for glucose, isoamyl acetate, and caproic acid, and for all other components the best score dataset was the e-nose + e-tongue dataset. These results suggest that the choice of whether to use e-nose data alone or to combine it with e-tongue data should be determined by first confirming the prediction accuracy.

3.2.2 The effects of the applied MRA and machine learning on the prediction accuracy

The data in Table 7 describe the accuracy (best score) of predicting components by the MRA and machine learning (RF, SVM, GB) using e-nose and e-tongue data. The prediction accuracy was improved by the MRA and machine learning for all components except the acidity and amino acid content. These results suggested that it is necessary to supplement the analysis with data other than e-nose and e-tongue data in order to capture the acidity and amino acid content accurately. In addition,
although the prediction accuracy for the glucose content was slightly improved (adjR²: 0.43), the prediction accuracy was still low compared to that for the volatile components.

The combination of the regression analysis and data set that gives the best score differs depending on the component. Similar to the prediction by PLS, the combination of the dataset and the regression analysis method needs to be determined by confirming the prediction accuracy.

There were many components with the highest prediction accuracy in MRA.

It is thought that the reason why MRA is more accurate than PLS is that the sample is only Junmai Ginjo Sake, which has the same ingredients and manufacturing method.

Although the samples have similar characteristics, the distribution of the objective variables is the same for training data and test data from Tables 1 and 2, so it can be judged that the method of dividing the dataset was appropriate. It is not clear why MRA is more accurate than PLS. However, as mentioned above, since MRA also has the problem of multicollinearity, it is necessary to be careful when applying MRA to the prediction of unknown samples.

Therefore, in order to select a suitable model, it is necessary to consider not only the prediction accuracy of the model but also the characteristics of each regression analysis.

The other regression analysis methods reduced the error by 1.18% compared to the case using only PLS.

### 3.2.3 The effects of adding the simplified analyses data on the prediction accuracy by PLS

The data Table 8 describe the accuracy (best score) when predicting components by PLS using data from the e-nose, e-tongue, and simplified analyses. By adding the simplified analyses data, the prediction accuracy of all components (except for isoamyl acetate) was improved.

The prediction accuracy for acidity, amino acid content, and glucose content were especially improved as follows. Error%: acidity 10.34, amino acid content 6.67, glucose 7.23; adjR²: acidity 0.44, amino acid content 0.79, glucose 0.73. For the acidity, amino acid content, and glucose content, these findings suggested that the pH, OD, and simplified glucose content had significant effects on the improvement of prediction accuracy, respectively.

The addition of the simplified analyses data reduced the error by an average of 1.32% compared to the case in which only the e-nose and e-tongue data were used.

### 3.2.4 Verification of the effects of combining the other regression analyses and adding simplified analyses data on the prediction accuracy

The data in Table 9 reveal the accuracy (best score) of predicting components by PLS, MRA and machine learning (RF, SVM, GB) using data from the e-nose, e-tongue and simplified analyses. The prediction accuracy was improved for all components. Except for acidity, all of the components had a high adjR² value at 0.79–0.95, suggesting that they could be predicted sufficiently. Acidity had the adjR² of 0.53, which was lower than those of the other components.
components. Therefore, in the analysis of an index in which acidity is an important factor (e.g., the taste of acidity and comprehensive evaluations), it is better to add acidity data in addition to e-nose and e-tongue data.

The analysis conditions for the best score differ depending on the component. In many cases, all of the data sets had the best score under the conditions analyzed by GB. It was suggested that it is not possible to directly measure the components of the objective variable by comprehensive analysis and simplified analysis, but because it is grasped as multiple analysis values (multiple sensor responses and peak values) including the interaction of the components. In addition, since the relationship between the explanatory variable and the objective variable is non-linear, it is considered that the prediction accuracy in GB was highest. On the other hand, by using PLS, the accuracy of predicting the amino acid content was the highest. The high prediction accuracy of the linear regression suggests that there are explanatory variables that are highly correlated with the

| Component               | adjR² | MAE  | E%   | Difference | Dataset                        |
|-------------------------|-------|------|------|------------|--------------------------------|
| Acidity                 | 0.44  | 0.51 | 10.34| -1.15      | e-nose, e-tongue, SA**         |
| Amino acid content      | 0.79  | 0.34 | 6.67 | -4.76      | e-nose, e-tongue, SA           |
| glucose                 | 0.73  | 0.35 | 7.23 | -5.01      | e-tongue, SA                   |
| Acetaldehyde            | 0.89  | 0.28 | 5.38 | -0.09      | e-nose, e-tongue, SA           |
| Ethyl acetate           | 0.89  | 0.20 | 6.17 | -0.54      | e-nose, e-tongue, SA           |
| n-ProOH                 | 0.88  | 0.31 | 4.48 | -0.64      | e-nose, SA                     |
| iBuOH                   | 0.90  | 0.27 | 4.34 | -0.21      | e-nose, SA                     |
| Isoamyl acetate         | 0.90  | 0.16 | 4.79 | 0.00       | e-nose                         |
| iAmiOH                  | 0.79  | 0.38 | 7.19 | -0.09      | e-nose, e-tongue, SA           |
| Ethyl caproate          | 0.90  | 0.20 | 6.15 | 0.00       | e-nose, e-tongue, SA           |
| Ethyl caprylate         | 0.81  | 0.20 | 6.10 | -0.87      | e-nose, SA                     |
| Caproic acid            | 0.77  | 0.38 | 7.77 | 0.00       | e-nose, e-tongue, SA           |
| Average                 | 0.81  | 0.30 | 6.38 | -1.32      |                                |

**SA: simplified analyses**

| Component               | adjR² | MAE  | E%   | Difference | Regression analysis | Dataset                        |
|-------------------------|-------|------|------|------------|---------------------|--------------------------------|
| Acidity                 | 0.53  | 0.51 | 9.20 | -2.29      | GB                  | e-nose, e-tongue, SA**         |
| Amino acid content      | 0.79  | 0.34 | 6.67 | -4.76      | PLS                 | e-nose, e-tongue, SA           |
| glucose                 | 0.80  | 0.35 | 6.12 | -6.12      | GB                  | e-nose, SA                     |
| Acetaldehyde            | 0.89  | 0.28 | 4.49 | -0.98      | GB                  | e-nose, SA                     |
| Ethyl acetate           | 0.89  | 0.20 | 3.97 | -2.74      | MRA                 | e-nose, SA                     |
| n-ProOH                 | 0.94  | 0.31 | 3.31 | -1.81      | MRA                 | e-nose, SA                     |
| iBuOH                   | 0.94  | 0.27 | 3.40 | -1.15      | MRA                 | e-nose                         |
| Isoamyl acetate         | 0.95  | 0.16 | 2.83 | -1.96      | RF                  | e-nose, e-tongue, SA           |
| iAmiOH                  | 0.79  | 0.38 | 6.18 | -1.10      | GB                  | e-nose, e-tongue, SA           |
| Ethyl caproate          | 0.95  | 0.18 | 3.67 | -2.48      | GB                  | e-nose, e-tongue, SA           |
| Ethyl caprylate         | 0.92  | 0.20 | 4.36 | -2.61      | GB                  | e-nose, e-tongue, SA           |
| Caproic acid            | 0.79  | 0.38 | 6.27 | -2.33      | SVM                 | e-nose, e-tongue, SA           |
| Average                 | 0.85  | 0.30 | 5.04 | -2.53      |                     |                                |

**SA: simplified analyses**
Table 10  Summary of the prediction accuracy (error%) under each analysis condition.

| Dataset          | Regression analyses | e-nose, e-tongue | e-nose, e-tongue, SA** |
|------------------|---------------------|------------------|------------------------|
|                  |                     | PLS              | PLS +MRA, ML*          |
| (1)              | Prediction by the PLS using e-nose and e-tongue data | 7.57             |                        |
| (2) Effect of applied other regression analyses | 6.25 | -1.32 |
| (3) Effect of adding SA data | 6.38 | -1.18 |
| (4) Hybrid effect | 5.04 | -2.53 |

*ML: machine learning  **SA: simplified analyses  
***Difference: difference between prediction by the PLS using e-nose and e-tongue data

Objective variable. iBuOH had higher prediction accuracy with only e-nose data. As the variable with the highest prediction accuracy of Glucose content, the variable to be combined with simplified analyses data was E-nose instead of E-tongue. Volatile components of sake such as alcohols, esters and aldehydes are produced by the enzymatic reaction of yeast. The yeast works by metabolizing glucose. In other words, it is considered that the E-nose data, which measure the amount and composition of volatile components, contributed significantly to the prediction of glucose content, which is related to yeast activity. As described above, it was found that the best analysis conditions differ depending on the characteristics of the ingredients even if the ingredients are contained in the same food. It was also shown that machine learning is not always superior to linear regression in estimating the component values of Junmai Ginjo. It has been contended that there is no best practice in prediction by a regression analysis (the ‘no free lunch’ theory) [24], and thus when predicting components and evaluations, it is necessary to prepare a plurality of analysis conditions and challenges.

Table 10 summarizes the prediction accuracy under each analysis condition. The PLS analysis using both the e-nose and e-tongue data showed that the component values of sake could be predicted with an average error of 7.57%. The error% was improved to 5.04% by adding the simplified analyses data and using the regression analyses method.

Based on the above-described findings, it is apparent that the data obtained with an e-nose and e-tongue can comprehensively capture the components of sake when we combine regression analyses and use simplified analyses data.

In addition, we suspect that by adding data with characteristics other than those examined by the analysis method used herein (such as a comprehensive analysis by liquid chromatography–mass spectrometry), features can be more comprehensively captured. An e-nose and/or e-tongue may also be able to analyze components other than the main components. These data are highly likely to be able to quantitatively capture a comprehensive evaluation in sensory evaluations that take into account the entire food.

Acknowledgments

We thank Mr. Hasegawa and Mr. Takagi of Hasegawa Saketen, Inc. for providing the sake samples.

And, we would like to thank Mr. Yoshida and Mr. Yajima of Alpha M.O.S. Japan K.K. for advice on the analyses of E-nose and E-tongue.

References

1) S. Sato, M. Shimada, T. Yahara; “Studies on the Taste of Sake” (in Japanese). J. Brew. Soc. Jpn., 62, 506-511 (1967).
2) N. Mimura, A. Isogai, K. Iwashita, T. Bamba, E. Fukusaki; Gas chromatography/mass spectrometry based component profiling and quality prediction for Japanese sake. J. Biosci. Bioeng., 118, 406-414 (2014).
3) S. Shimofuji, M. Matsui, Y. Muramoto, H. Moriyama, R. Kato, Y. Hoki, H. Uehigashi; Machine Learning in Analyses of the Relationship between Japanese Sake Physicochemical Features and Comprehensive Evaluations. Jpn. J. Food Eng., 21, 37–50 (2020).
4) M. Sugimoto, T. Koseki, A. Hirayama, S. Abe, T. Sano, M. Tomita, T. Soga; Correlation between sensory evaluation scores of Japanese sake and metabolome profiles. J. Agr. Food Chem., 58, 374–383 (2009).
5) R. B. Roy, B. Tudu, L. Shaw, A. Jana, N. Bhattacharyya, R. Bandypadhyay; Instrumental testing of tea by combining the responses of electronic nose and tongue. J. Food Eng, 110, 356–363 (2012).
6) X. Hong, J. Wang; Detection of adulteration in cherry tomato juices based on electronic nose and tongue: Comparison of
different data fusion approaches. J. Food Eng, 126, 89–97 (2014).
7) S. Qiu, J. Wang, L. Gao; Discrimination and characterization of strawberry juice based on electronic nose and tongue: Comparison of different juice processing approaches by LDA, PLSR, RF, and SVM. J. Agr. Food Chem., 62, 6426–6434 (2014).
8) R. N. Bleibaum, H. Stone, T. Tan, S. Labreche, E. Saint-Martin, S. Isz; Comparison of sensory and consumer results with electronic nose and tongue sensors for apple juices. Food Qual Prefer, 13, 409–422 (2002).
9) Y. Kanno, T. Minetoki, T. Bogaki, K. Toko; Visualization of Flavor of Sake by Sensory Evaluation and Statistical Method. Sens. Mater., 30, 1197–1206 (2018).
10) D. Cozzolino, G. Cowey, K. A. Lattey, P. Godden, W. U. Cynkar, R. G. Damberg, L. Janik, M. Gishen; Relationship between wine scores and visible–near-infrared spectra of Australian red wines. Anal. Bioanal. Chem., 391, 975–981 (2008).
11) V. Vapnik; Pattern recognition using generalized portrait method. Automat Rem. Contr+., 24, 774–780 (1963).
12) L. Breiman; Random forests. Mach. Learn., 45, 5–32 (2001).
13) A. Natekin, A. Knoll; Gradient boosting machines, a tutorial. Front Neurorobot, 7, 21 (2013).
14) 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, 785–794 (2016).
15) T. Nishiya (Ed.); “Commentary for standard method established by National Tax Agency, 4th edition” . Brewing Society of Japan, Tokyo, Japan, 1993.
16) I. Miwa, J. Okuda, K. Maeda, G. Okuda; Mutarotase effect on colorimetric determination of blood glucose with –D-glucose oxidase. Clin Chem Acta. 37, 538–540 (1972).
17) K. Yoshizawa; “The rapid assay for aroma compounds in sake by headspace method” (in Japanese). J. Brew. Soc. Jpn., 68, 59–61 (1973).
18) S. Yamanaka, S. Kawanishi, K. Okui; “Changes of the Nucleic Acid Related Substances in Sake Making Process” (in Japanese). J. Brew. Soc. Jpn., 77, 817–821 (1982).
19) M. Kuhn; Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26 (2008).
20) M. Kuhn; A short introduction to the caret package. R Found. Stat. Comput. 1–10 (2015).
21) B. Mevik, R. Wehrens; The pls package: Principal component and partial least squares regression in R. J. Stat. Softw., 18, 1–24 (2007).
22) A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis; kernlab – An S4 package for kernel methods in R. J. Stat. Softw. 11, 1–20 (2004).
23) A. Liaw, M. Wiener; Classification and regression by randomForest. R. News. 2, 18–22 (2002).
24) D. H. Wolpert, W. G. Macready; No free lunch theorems for search, Technical Report SFI–TR–95–02–010 Santa Fe Institute. (1995).

URLs cited

URLs cited should be listed here with access dates as in the following example.
i) https://www.R-project.org/ . (June 3., 2019)
ii) https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html (June 3., 2019).
iii) https://cran.r-project.org/web/packages/xgboost/index.html (March 11., 2020).
機械学習によるE-noseとE-tongueを用いた日本酒成分の予測

下藤悟1,2,†, 松井元子2, 村元由佳利3, 森山洋憲1, 甫木嘉朗1, 上東治彦1

1 高知県工業技術センター, 2 京都府立大学大学院, 3 京都府立大学

日本酒は日本の伝統的な嗜好性アルコール飲料である。日本酒の品質の一般的な指標として、揮発性成分（酢酸エチル、イソアミルアルコール、酢酸イソアミル、カプロン酸エチルなど）、グルコース含量、酸度、およびアミノ酸度が分析されている。これらの分析値を使用して日本酒の官能特性を明らかにするために、多くの研究が行われている[1-2]。以前の報告では、日本酒を特徴付ける物理化学的特徴から総合評価を予測するために機械学習を適用した[3]。その際に、特に評価の低いサンプルでは予測精度が低いことが示された。そのため、総合評価をより正確に予測するためには、網羅分析のデータを説明変数に追加する必要があると考えた。

E-noseとE-tongueは網羅分析が可能であるといわれている。E-noseやE-tongueは、比較的簡単な操作で分析できる。しかし、複雑な成分を含む食品を分析する場合、成分間の相互作用によるノイズのために明確なデータが得られない可能性がある。E-noseまたはE-tongueでは測定できない成分がある場合、補足するデータを収集する必要がある。他の食品では、E-noseとE-tongueを組み合わせることで成分値の関連性を確認した多くの研究がある[5-8]。しかし、日本酒について、E-tongueを使用した研究[9]はあるが、E-noseを組み合わせた研究はみられない。したがって、E-noseとE-tongueから得られたデータが日本酒の特徴を包括的に捉えていることを確認する必要がある。そこで、本研究では、日本酒の特徴をE-noseとE-tongueでどの程度捉えることができるかを明らかにすることを目的とし、E-noseおよびE-tongueデータを使用して、日本酒の主要な成分値の推定を行った。

日本酒の成分の予測には回帰分析法を適用した。純米吟醸の特徴を表す指標として酸度、アミノ酸度、グルコース、および9つの揮発性成分を目的変数として用いた。説明変数には、E-noseによって得られた99のピークデータとE-tongueによって得られた7つのセンサーデータを用いた。

回帰分析の手法としては、一般的に使用される統計的手法である部分最小二乗回帰（PLS）を使った。E-noseおよびE-tongueデータを使用したPLSによる予測精度は、平均7.57%（成分値の範囲に対するMAEの割合）であった（表5）。また、回帰分析手法に重回帰分析（MRA）と機械学習（サポートベクターマシン（SVM）、ランダムフォレスト（RF）、勾配ブースティング（GB））を適用することで、酸度とアミノ酸度を除くすべての成分の予測精度が改善された（表6）。さらに、他の回帰分析を適用し、7つの簡易分析データ（Brix、pH、電気伝導率、OD260、OD280、簡易アルコール含有量、簡易グルコース含有量）を追加することにより、すべての成分値の予測精度が向上した。（error%の平均: 5.04）（表8）。また多くの成分で、GBを用いることで予測精度が最も高くなった。PLSのような離形回帰よりも機械学習で予測精度が向上した理由として、網羅分析や簡易分析では目的変数の成分を直接測定するのではなく、成分の相互作用を含む複数の分析値（複数のセンサー応答とピーク値）として把握しているためであることが示唆された。そのうえで、説明変数と目的変数の関係が非線形的な関係にあるために、GBでの予測精度が最も高くなった。PLSのように離形回帰よりも機械学習で予測精度が向上した理由として、網羅分析や簡易分析では目的変数の成分を直接測定するのではなく、成分の相互作用を含む複数の分析値（複数のセンサー応答とピーク値）として把握しているためであることが示唆された。そのうえで、説明変数と目的変数の関係が非線形的な関係にあるために、GBでの予測精度が最も高かった。PLSのような離形回帰よりも機械学習で予測精度が向上した理由として、網羅分析や簡易分析では目的変数の成分を直接測定するのではなく、成分の相互作用を含む複数の分析値（複数のセンサー応答とピーク値）として把握しているためであることが示唆された。そのうえで、説明変数と目的変数の関係が非線形的な関係にあるために、GBでの予測精度が最も高かった。PLSのような離形回帰よりも機械学習で予測精度が向上した理由として、網羅分析や簡易分析では目的変数の成分を直接測定するのではなく、成分の相互作用を含む複数の分析値（複数のセンサー応答とピーク値）として把握しているためであることが示唆された。そのうえで、説明変数と目的変数の関係が非線形的な関係にあるために、GBでの予測精度が最も高かった。PLSのような離形回帰よりも機械学習で予測精度が向上した理由として、網羅分析や簡易分析では目的変数の成分を直接測定するのではなく、成分の相互作用を含む複数の分析値（複数のセンサー応答とピーク値）として把握しているためであることが示唆された。