SHIMADZU
Excellence in Science

Application News

No. M276

Gas Chromatograph Mass Spectrometer

Quality Classification of Foods Through Analysis by Machine Learning

Foods are comprised of numerous components and the quality of foods may not be completely the same even for the same food products.

Differences in quality are considered to be caused by slight differences in the components that comprise food products. Therefore, for the purpose of a complete quality evaluation, comprehensive analysis of components is gaining attention in recent years. In order to estimate and identify the subjective properties of foods such as taste, smell, and deterioration based on their components, one method that is expected to be effective is to learn the relation between components and subjective properties of a known sample and then utilize those results for an unknown sample.

This article studies whether or not it is possible to distinguish between beef samples that have been properly refrigerated and those that are expected to have some deterioration from being exposed to a 40 °C environment for 3 hours based on the analysis results of volatilized components when those samples are heated to 200 °C. After making a classifier learn known data of each sample type, the learned data was used to define quality. This was then used to classify unknown data as either sample type and calculate the percentage of correct results. By using a support vector machine (SVM) as the classifier, we were able to obtain correct results by 95.8 % even for samples that were hard to classify by comparing chromatograms or through principal component analysis of peak area values.

T. Sakai

Sample Preparation

Two types of samples were prepared using meat from various beef cuts: properly refrigerated samples (4 °C samples) and samples expected to have some deterioration from being exposed to a 40 °C environment for 3 hours (4 °C samples). The appearance of the samples is shown in Fig. 1.

From each meat sample, 20±3 mg was taken and placed in individual measurement vials. A total of 116 vials (58 vials of 4 °C samples and 58 vials of 40 °C samples) were prepared and analyzed.

Fig. 1 Left: Properly Refrigerated Sample (4 °C sample) Right: Sample Exposed to a 40 °C Environment for 3 Hours (40 °C sample)

Fig. 2 AOC-6000

Analysis Using Solid Phase Microextraction (SPME)

The vials were heated at 200 °C for 15 minutes and the resulting vapor was collected by SPME and analyzed in scan mode. As there were many samples, we used the AOC™-6000 for injection since it is capable of collection, adsorption, and desorption automatically by SPME. Table 1 lists the analytical conditions.

Comparison of total ion chromatograms did not reveal any peaks characteristic to each sample.

Table 1 Measurement Conditions

SPME fiber	: Divinylbenzene/Carboxen/Polydimethylsiloxane (DVB/CAR/PDMS)
Incubation Temp.	: 200 °C
Incubation time	: 15 min
Agitator	: 250 rpm
Desorb time	: 1 min
Column	: SLB®-5MS (30.0 m × 0.25 mm, 0.25 µm, Sigma-Aldrich Co. LLC)
Injection mode	: Split
Split ratio	: 5:1
Injection port Temp.	: 280 °C
Oven Temp. program	: 60 °C (1 min) → (20 °C/min) → 200 °C → (8 °C/min) → 320 °C (5 min)
Flow control	: Linear velocity (50.0 cm/sec)
Purge flow rate	: 3 mL/sec
Interface temperature	: 200 °C
Ion source temperature	: 250 °C
Event time	: 0.3 sec

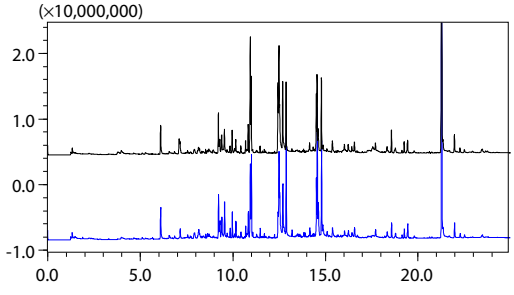

Fig. 3 Example Total Ion Chromatograms Black: 4 °C Sample, Blue: 40 °C Sample

Peak Extraction

Peaks were extracted from mass spectrometry chromatograms using MZmine 2 (ver. 2.32) which is a mass spectrometry data analysis software. Although 9318 peaks were detected in all, any peaks that were missing (not detected) in any of the 116 data files were deleted. We used the area values of the resulting 200 peaks as explanatory variables. Analysis was performed using a data matrix of 200 peaks × 116 data files.

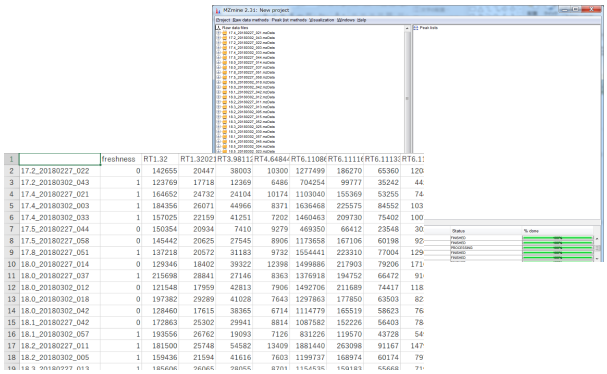

Fig. 4 MZmine 2 Operation Window (Top) and the Data Matrix of Each Peak Area Value in Each Data File (Bottom)

Principal Component Analysis

In order to confirm that there is no characteristic peak for each sample, we performed principal component analysis using this data matrix. Fig. 5 shows the score plot. We can see that classification of the two sample types is difficult by principal component analysis.

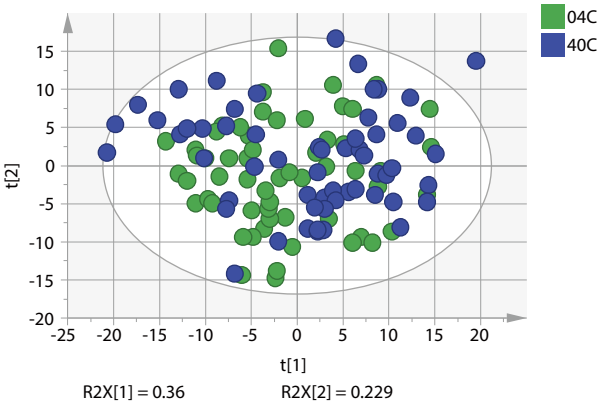

Fig. 5 Score Plot of Principal Component Analysis

Preparation of Data for Learning

The 116 samples were randomly divided into a training set comprising 92 samples and a test set comprising 24 samples while making sure that the number of 4 °C samples and 40 °C samples are the same within each group. The training set was used for learning by the classifier and using those learning results, the classifier classified the test set samples.

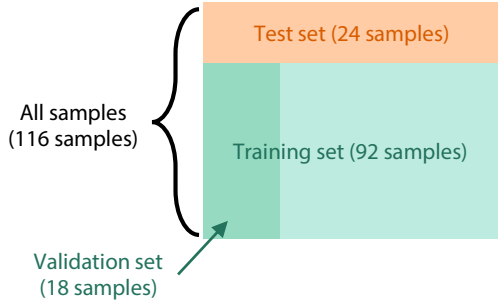

Fig. 6 Division of Datasets

Classification Using a Support Vector Machine

We prepared a support vector machine (SVM) as the classifier. The SVM was implemented in Python 3.6 using scikit-learn (ver. 0.19.1). The 92 samples of the training set were further divided into a validation set comprising 18 samples and a training set comprising 74 samples. The SVM hyperparameters "C" and "gamma" were optimized by cross-validation using these sets. The "rbf" kernel was used.

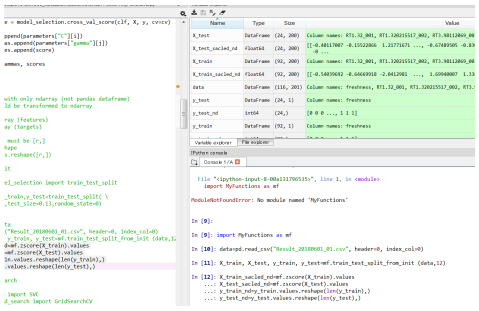

Fig. 7 SVM Implementation in Python 3.6 Using scikit-learn

Utilizing the hyperparameters optimized with the training set, we classified the 24 samples of the test set. Of the 24 samples, 23 samples were classified correctly. Table 2 lists the results, indicating 4 °C samples as "Positive" and 40 °C samples as "Negative".

Table 2 Classification Results of the 24 Samples of the Test Set

	True	False
Positive	12	0
Negative	11	1
Precision	95.8 %	

First Edition: Mar. 2019

Shimadzu Corporation
www.shimadzu.com/an/