Statistical Quality Analysis of Tea Polyphenols (TP) in Tea Drinks

February 17, 2016

Introduction

Tea Polyphenols (TP) is the main nutrient in tea drinks. Its concentration is also a quality indicator of tea drinks. However, in tea drinks manufactruing industry, the structure or the composition of TP hasnt been deeply analyzed in the purpose of identiyfing the properties of good TP, rising the risk of violence in food safety. A potentially similar example is Chinese milk scandal in 2008, in which milk manufacturers used melamine as nitrogen source rather than protein, causing kidney damage on infants and even kidney stones. The reason why they can pass milk test on nitrogen content is because the test only measures the quantity of nitrogen rather than look deeper on the “quality”. Some chemistry research have figured out experimental approaches to solve this problem. From the graph below, for example, we can clearly identify whether melamine is added or not.

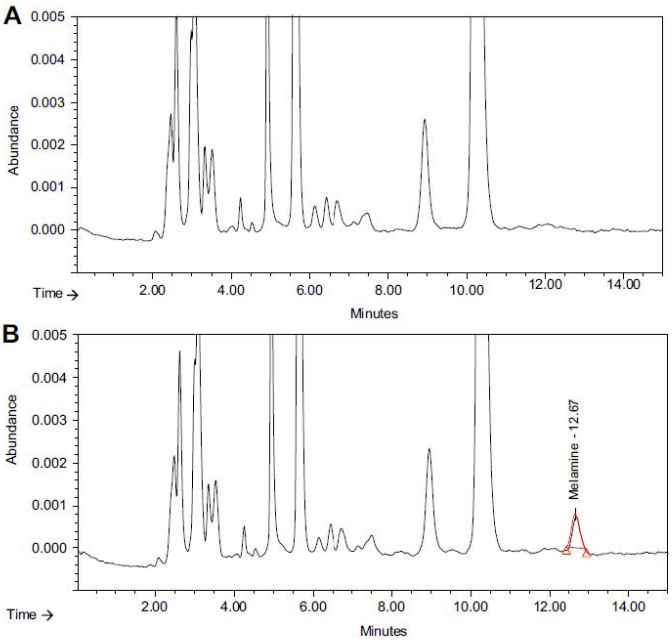

Figure 1: The UV spectrum of milk powder sample (A) and the one adding melamine (B) (Lutter(2011))

This project is based on the current chemistry research, trying to build a general method to automate this distinguishing process, which will make it possible to monitor the real-time quality of tea drinks on the production line. The data is collected by Quan Yuan at Tsinghua University.

Dataset

The data is actually the absorption curve of TP using uv-vis spectrophotometry, which can characterize the structure and composition of materials. The experiment follows relevant Chinese national standard (GB/T 21733-2008). A sample data is showed below. It measures absorbance under wavelengths from 350nm to 700nm.

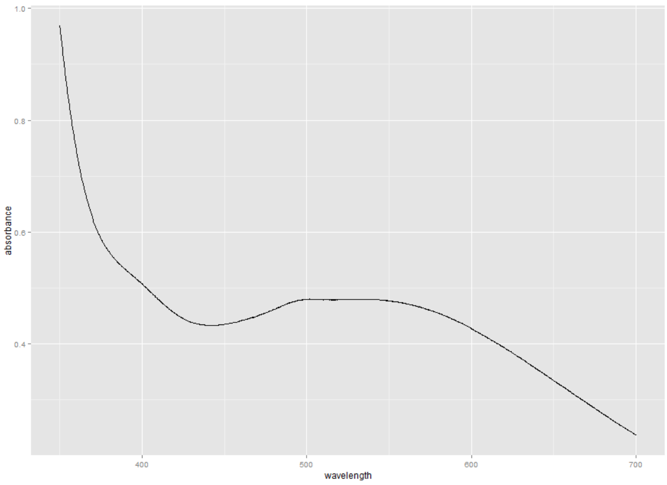

Figure 2: Sample data

Basically the dataset is a batch of curves similar to the above one: every curve is a spectrum from a tea drink. 3 brands of tea drinks are used in the experiment (5 samples each) and alogether 15 spectrums are collected:

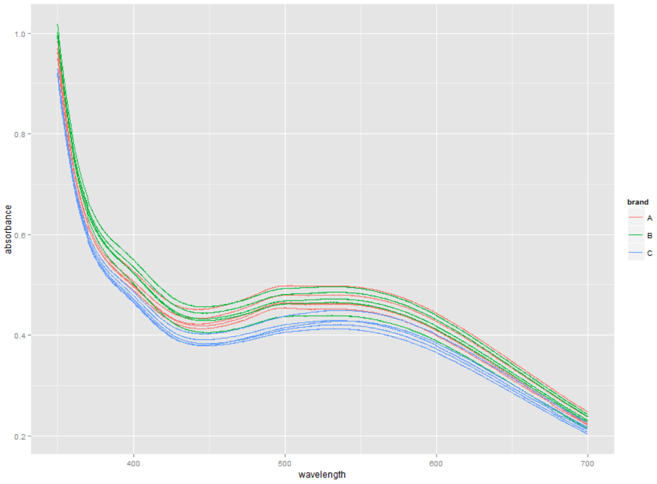

Figure 3: Dataset

There are differences in vertical positions among the curves, which shows that the TP concentration in tea drinks are different. However, I'm more interested in the difference in curve shapes. We can see that there are differences in shapes between 450nm and 550nm:

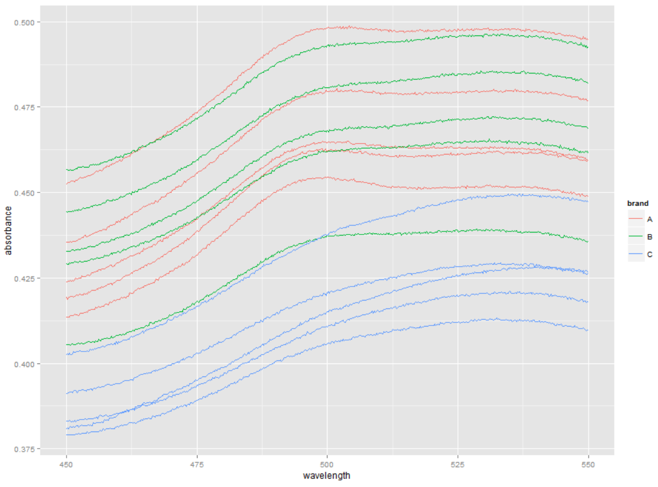

Figure 4: Difference in shapes among curves

Data processing

It is unrealistic to apply classification model directly on the orinigal dataset because of the high dimensions. To reduce dimensions, here I use wavelet transform because of its good performance in preserving shapes of curves with limited ranges as well as dimension-reduction. Daubechies Wavelet is selected and level 6 is selected as the discomposition level. Soft thresholding is applied on the approximate wavelet cofficients. I use approximate wavelet cofficients instead of raw curves as input of the future classification model. Here shows the approximate wavelet cofficients.

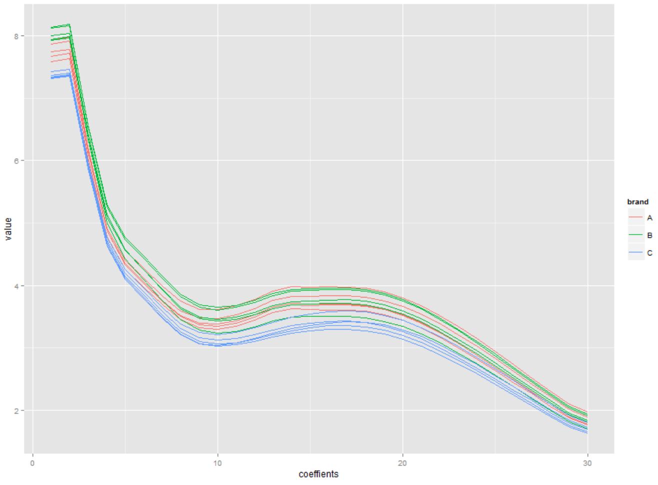

Figure 5: Approximate wavelet cofficients generated from the raw dataset

Now each ‘curve’ has 30 points (dimension) rather than original 1751 points. This is a good start for classification.

Classification and differences identification

Model description

SVM with RBF knernel is tried. Further, here I want my model to identify which parts of curves are actually different, that is, to locate that it is the curves between 450nm and 550nm that are different. To achieve this, we're actually completing another task: to find the features that will lead to the least generalization error. According to Vapnik (2000), there exists upper bound of the generalization error for SVM: with training set $Z = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)\}$, hyperplane $(\boldsymbol{\omega}, b)$, non-linear transformation ϕ, we can define margin $\gamma(\omega, b, Z)$ and radius $R(Z)$ as

$$\begin{aligned}\gamma(\boldsymbol{\omega}, b, Z) &= \max_{(\mathbf{x}_i, y_i)\in Z} \frac{y_i(\omega \cdot \phi(\mathbf{x}_i) + b)}{\|\boldsymbol{\omega}\|} \\ R(Z) &= \min_{\mathbf{a}, \mathbf{x}_i} \|\phi(\mathbf{x}_i) + \mathbf{a}\|\end{aligned}$$

and the expectation of the misclassification probability as

$$p_{err}(\boldsymbol{\omega}, b) = P(\text{sign}(\boldsymbol{\omega} \cdot \phi(\mathbf{X}) + b \neq Y))$$

we have

$$E\{p_{err}(\alpha)\} \leq \frac{1}{l} E\{\frac{R^2(Z)}{\gamma^2(\boldsymbol{\omega}, b, Z)}\} = \frac{1}{l} E\{R^2(Z) W^2(\alpha^0, Z)\}$$

Where $W(\alpha)$ is actually ω in the objective function of SVM, and α^0 is the opitimal solution of the dual problem in SVM.

To summary, locating differences is actually selecting features to minimize $R^2(Z)W^2(\alpha^0, Z)$. So I define $\phi_{\boldsymbol{\sigma}}(\mathbf{x}) = \phi(\mathbf{x} * \boldsymbol{\sigma})$, where $\boldsymbol{\sigma} \in \{0, 1\}^n$. And our goal is solving the following optimization problem:

$$\begin{aligned}&\min_{\alpha^0, \boldsymbol{\sigma}} R^2(\boldsymbol{\sigma}) W^2(\alpha^0, \boldsymbol{\sigma}) \\ &\qquad \text{s.t.} \boldsymbol{\sigma} \in \{0, 1\}^n\end{aligned}$$

where $R^2(\sigma_k) = \max_\beta \sum_i \beta_i K_{\boldsymbol{\sigma}}(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j} \beta_i \beta_j K_{\boldsymbol{\sigma}}(\mathbf{x}_i, \mathbf{x}_j)$so that $\sum_i \beta_i = 1, \beta_i \geq 0$.

Since it is a Combinatorial Optimization problem, so I use heuristic algorithm to find the approximatly optimal solution. In this project, Genetic Algorithm is used, with $1/(R^2(\boldsymbol{\sigma})W^2(\alpha^0, \boldsymbol{\sigma}))$ as fitness function. Besides $\boldsymbol{\sigma}$, we still need to tune the parameters (C and γ) of SVM. Here I set C and $\gamma \in 2^m | m = k, k+1, \cdots, d$. So the structure of is showed below:

σ_1	σ_2	$\cdots$	σ_n	γ_k	$\cdots$	γ_d	C_k	$\cdots$	C_d

Figure 6: The structure of

In this project, paramters of Genetic Algorithm are set: population size as 100, 30 generations, mutate rate as 0.05% and elite as top 3.

Simulation

Simulated datasets

The dataset I collected from experiments is sitll to small (15 data ‘points’, 30 dimensions) for training the model. So here I created simulated data consisiting of two classes of curves in the following steps:

1 Select the average curve of brand C as the base curve and class I $\mathbf{Y}$,

2 Caluate the difference between the average curve of brand A and the base curve, denoted as $\mathbf{Y_d}$; remove odd points of $\mathbf{Y_d}$ until the point number is 1/16 of the original ones, denoted as $\mathbf{Y_{d,s}}$, demonstrated in fig X,

3 Add $\mathbf{Y_{d,s}}$ on $\mathbf{Y}$ in the interval $[a, b]$ to create class II,

4 Add noise from the same distribution on both $\mathbf{Y}$ and $\mathbf{Y}'$.

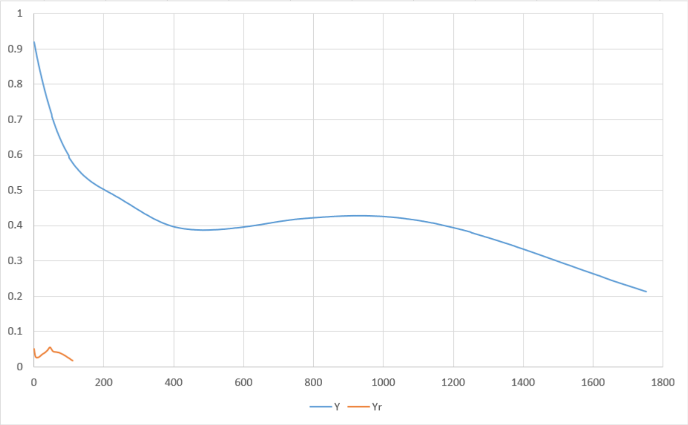

Figure 7: Demonstration of $\mathbf{Y_{r,new}}$ and $\mathbf{Y}$

So, we have

$$\begin{aligned} \text{class 1} &\;:\; \{\alpha\mathbf{Y}+\varepsilon\} \\ \text{class 2} &\;:\; \{\alpha\mathbf{Y}`+'\beta\mathbf{Y_{d,s}}+\varepsilon\} \end{aligned}$$

where $\alpha \sim U(0.8, 1.2)$, $\beta \sim U(0.8, 1.2)$, $\varepsilon \sim N(0, 0.02^2)$. α and β are used to create a more general dataset.

If setting $a = 700, b = 800$, here we can get a simulated dataset:

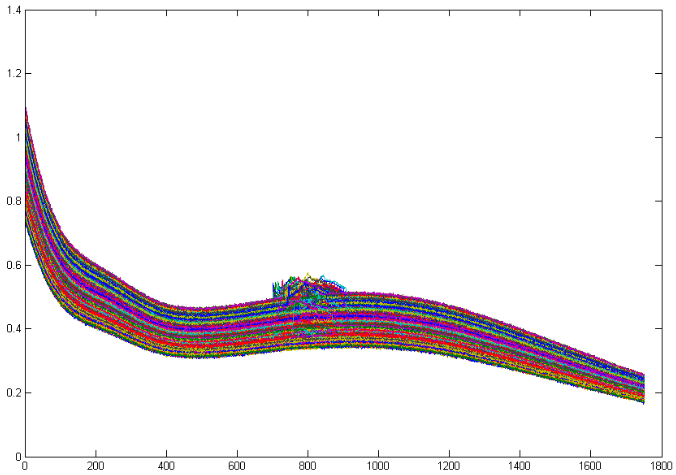

Figure 8: simulated dataset I (training set)

If setting $a = 600, b = 900$, letting $\mathbf{Y'_{d,s}} = \mathbf{Y_{d,s}/8}$ as the additive curves, we get a dataset that is harder to be classified:

If setting $a = 700, b = 800$, here we can get a simulated dataset:

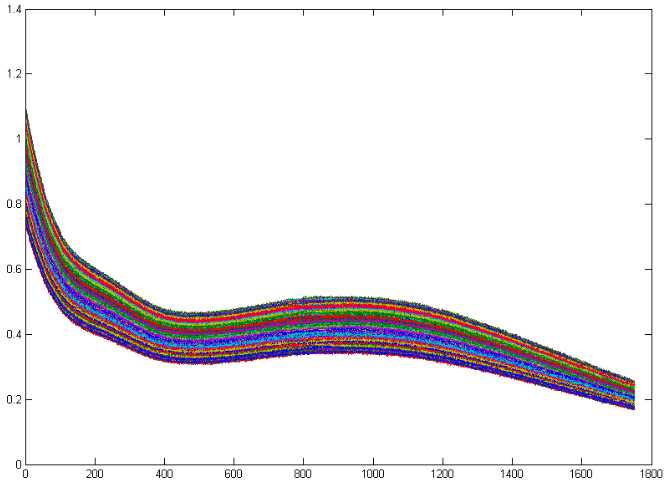

Figure 9: simulated dataset II (training set)

For both datasets, I created training sets of 400 curves (200 each), and 2000 curves as testing sets (1000 each).

Parameters tuning

To find a general range of the optimal parameters (k and d mentioned above, I applied meshgrid search and minimize $R^2W^2(\alpha^0)$. Here shows the value of

$R^2W^2(\alpha^0)$ with different paramters.

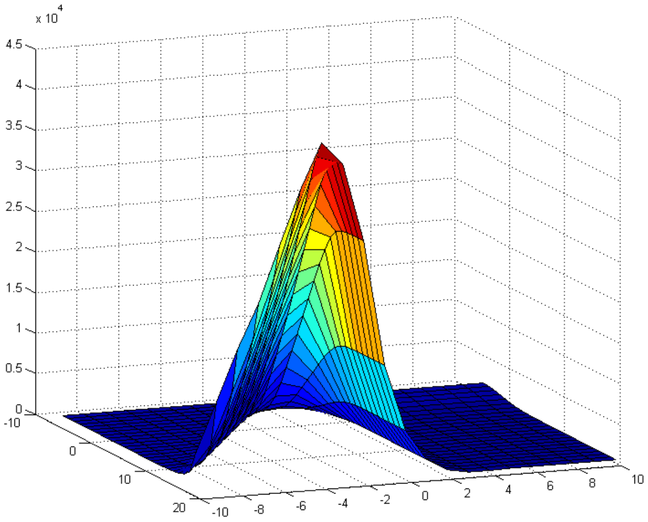

Figure 10: Influence of C and γ on $R^2W^2(\alpha^0)$

Then I find the minimal reaches when $C = 2^{10}$,$\gamma = 2^4$.

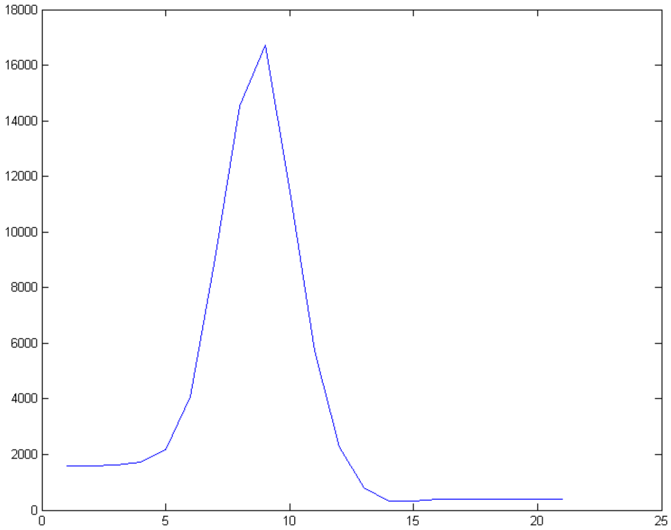

Figure 11: Influence of C on $R^2W^2(\alpha^0)$ when $\gamma = 2^4$

So we could choose $k_C = 2, d_C = 17, k_\gamma = 3, d_\gamma = 12$.

Simulation results

For both dataset I and dataset II, I conducted 40 simulations. The average accuracy is 100% and 96%. Further, I plot the total ratio of every feature (approximate wavelet coefficient) in the population of last generation in these 40 simulations. If a feature k has a high ratio, that means with a high confidence that $\sigma_k = 1$. In other words, feature k is the distinguishable one between two classes.

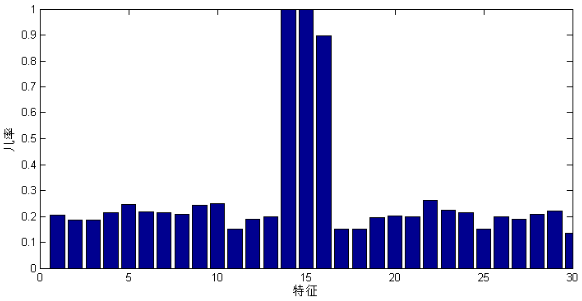

Figure 12: Ratios of features in the last generation (dataset I)

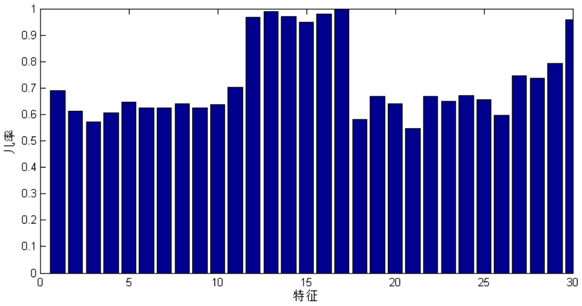

Figure 13: Ratios of features in the last generation (dataset II)

So, in dataset I, features 14-16 are the distinguishable ones; in dataset II, features 12-17 are the distinguishable ones. Those features correspond to [700, 800] and [600, 900].

Conclusion

A general method is designed to classify profile data and identify the distinguishable features between them. It is expected to be extensible in multiple classification problem. Further work could be done in designing advanced measures rather than only ratios for distingushable features identification.