Journal of Electroanalytical Chemistry 806 (2017) 97–106

Contents lists available at ScienceDirect

Journal of Electroanalytical Chemistry

journal homepage: www.elsevier.com/locate/jelechem

Effective multivariate data presentation and modeling in distinction of the tea infusions

Justyna Zuziak, Gabriel Moskal, Małgorzata Jakubowska*

AGH University of Science and Technology, Faculty of Materials Science and Ceramics, Mickiewicza 30, 30-059 Kraków, Poland

ARTICLE INFO

Keywords:
Chinese and Vietnamese tea infusions
DPV
LAPV
Multivariate data visualization
Polar dendrogram
SOM
PCA

ABSTRACT

In this work a review of multivariate data visualization and unsupervised modeling methods was presented, applied to DPV and LAPV signals of black, green, red and white tea infusions of China and Vietnam origin. It proves that graphical presentation by diagrams and charts has a number of advantages in comparison to analysis of the measured signals or numerical datasheets. The same data presented and modeled by different methods may deliver various, complementary conclusions and also the strategy verifies the quality of the experiments. Insight to the experimental data and extraction of the useful information were done applying Chernoff faces, star plots, radar plot and matrix scatterplot. Using SOMs and PCA it was demonstrated that DPV data enabled distinction of the tea kinds with some problems with separation of the green and white ones, while LAPV signals distinguished the origin of the infusions. Successful separation of the 8 different types of objects were done by polar dendrogram, first time applied in voltammetry.

1. Introduction

Tea is the most consumed beverage throughout the world, right after the water [1]. The reasons are good taste, healthy, dietetic and therapeutic benefits [2]. A number of herbal tea plant species – those for consumption prepared similarly as hot watery infusion, have been found in human cultures all around the world [3]. Typical tea comes from the *Camellia sinensis* plant, which makes it highly dependent on the conditions of cultivation and preparation, and these can vary between country of origin and the type of tea [1].

In recent times there are more evidences which confirm healing properties of tea. Specific substances in tea can help prevent many diseases. For example, antioxidants in tea may be capable of protecting against blood pressure and cardiovascular disease, colon cancer, breast cancer or lung cancer [4,5]. Tea is also recommended for alleviation of minor maladies such as headaches and pains [1].

There are many different varieties of tea in the world. The most popular are black, green, red (Pu-erh) and white tea. Black tea is the most important one consumed across the world, the production accounted for about 75% of global tea production [6]. In 2012 the biggest world production was in China – 35.3% and next India – 20.8%, Kenya – 7.7%, Sri Lanka – 6.8%, Turkey – 4.7% and Vietnam – 4.5%. World tea production increased significantly to 5.07 million tons in 2013. The 10 years projections to 2023 indicate that world black tea production will grow at a slightly higher rate compared to the previous decade. Black tea production is projected to grow to reach 4.17 million tons in 2023 [7].

As a highly merchandised product, the evaluation and quality control of tea is very important and have been a popular research topic in recent years. Currently, the classification and evaluation of tea quality are mainly performed by trained experts who have developed a unique language to describe the various quality attributes of a tea infusion. With the recent development of the tea industry, determining a method to overcome these challenges has become more urgent than ever [8].

Controlling fermentation time and classification of tea categories is very important, which strongly influence the commercial market. In the past, human experts measured the color variation during fermentation by visual check. However, this method has various shortcomings such as irreproducibility, high labor cost and inconsistency [4].

One of the important issues in tea quality evaluation is the origin authentication of the products. Especially in the case of middle and premium quality tea products, the origin is usually displayed on the label and it plays an important role in the consumers' choice. As a global trend, more and more food and beverage manufacturers tend to offer pure origin or single origin products. These products require a reliable confirmation of origin [9]. The next reason to distinguish tea are costs. The price of the tea is variable depending on the quality of the tea from several hundred RMB/kg to several ten thousand RMB/kg (RMB is Chinese currency) [10].

* Corresponding author.
E-mail address: jakubows@agh.edu.pl (M. Jakubowska).

http://dx.doi.org/10.1016/j.jelechem.2017.10.059
Received 26 September 2017; Received in revised form 17 October 2017; Accepted 25 October 2017
Available online 26 October 2017

Table 1
The last achievements about voltammetric distinction of teas.

Type of study	Technique	Working electrode	Data processing algorithm	References
Discrimination of tea by means of a voltammetric electronic tongue and different applied waveforms	LAPV, SAPV	Iridium, platinium, rhodium electrode	MVDA, PCA	[12]
Comparison of multivariate preprocessing techniques applied to electronic tongue based on pattern classification	LAPV, SAPV	Gold, iridium, palladium, platinum, and rhodium	PCA	[13]
Classification of black tea liquor using cyclic voltammetry	CV	Platinum and glassy carbon electrode	PCA, LDA	[14]
Instrumental testing of tea by combining the responses of electronic nose and tongue	LAPV, SAPV, SCV	Gold, iridium, palladium, platinum, and rhodium	PCA, LDA	[15]
Screening and authentication of tea varieties based on microextraction-assisted voltammetry of microparticles	CV, SWV	Glassy carbon	PCA, CA	[16]
Discrimination of Chinese and Vietnamese (black, green, red, white) and comparison of multivariate data visualization methods	DPV, LAPV	Renewable silver amalgam film electrode – Hg(Ag)FE	Chernoff faces, star plot, radar plot, matrix scatterplot, HCA (polar dendrogram), SOM, PCA	This work

Various researchers have made a number of efforts to correlate tea quality with its chemical composition which proved very difficult problems and may vary from person to person [11]. This problem can be solved by using a methodology of voltammetric discrimination of tea which is rapid, useful and cheap method.

There are known a few voltammetric methods about discrimination of tea. Overview the latest methods is gathered in Table 1.

Technological progress enables the collection of huge amounts of data and also information obtained in chemical experiments is growing in an exponential way. Extracting the valuable knowledge is a difficult task when large sets of data are not presented in graphical form. Data visualization is an effective way to explicitly illustrate the characteristics that are not apparent from the mathematical models or statistics. Traditional graphics is powerful, but limited to deal with multivariate data. Although it has the potential to be an important aid to the modern information extraction process, visualizing high-dimensional data is an ongoing challenge. Advanced statistical and machine learning algorithms are used to seek answers to increasingly complex questions, to obtain an integrated understanding of the data distribution and to investigate the relationship between objects and variables.

Last decades have seen the development of new methods for visualizing multivariable data and a few attempts have been made to investigate the range of applications of the new proposed graphical strategies. Multivariate visualization is a modern approach whose goal consists of representing information related to multidimensional objects, variables, and the relationships between them, as faithfully as possible on low-dimensional graphs and charts. They can provide a comprehensive picture of a problem that makes for a more complete and better balanced understanding that could be derived from tabular or textual forms of data presentation. They can bring out hidden facts and relationships and can stimulate analytical thinking and investigation. They may also present special effects in data, indicate outliers, identify patterns, correlation, clusters, diagnose models and search for novel and unexpected phenomena. An effective visualization system can make users aware of the quality of their data by detecting human errors in data entry, gaps in data gathering, and unexpected sensor or tool variability. Comprehensive graphical displays, supported by sophisticated and available software, can reduce the effort required to make complete and effective data analysis.

According also to the inspiring sentence, that “there is no statistical tool that is as powerful as a well-chosen graph” [17] in this work we want to present a review of multidimensional multivariate data visualization methods, applied to voltammetric tea infusions signals. Dimensionality of a problem refers to the number of attributes (variables), which are present in the data. When dimensionality increases a proper visualization and distinguishing is more difficult. Our approach bases on the obvious fact that voltammogram is a set of current samples measured by various potentials and utilization of whole information hidden in the signal is a challenge in data interpretation. Information presented in the compressed form by the well-designed charts and graphs leads to interesting conclusions, invisible without these methods.

In this work we considered such aspects as: (a) influence of the signal processing for the effect of the multivariate data presentation, analysis and conclusions; (b) optimization of data compression; (c) the effect of data pre-processing by various strategies; (d) adequate presentation of multivariate data and (e) unsupervised modeling by application of the chemometric algorithms. As a data we considered a set of voltammograms obtained for the black, green, red and white tea infusions of China and Vietnam origin. We want to prove that “a short look” for the adequately prepared graph may give information about an effect of tea profiling based on the voltammograms recorded for infusions.

2. Theory – multivariate data visualization

Graphical data presentation by diagrams and charts has a number of advantages in comparison to presentation of the measured signals or numerical datasheets. Well-designed pictures enable observation of the relationship between samples and variables, and may be a source of interesting conclusions. A comprehensive picture may present a solution of the considered problem and in this sense is a good way to replace a longer textual description. Applying chemometric methods also hidden facts and relationships may be shortly demonstrated and observed. Using diagrams and graph dedicated to multivariate data some different goals may be realized: (a) an overview of information in the compress form, which is easy to understand; (b) detection of gross error and outliers; (c) formulation of starting hypotheses; (d) searching of the solution by the optimization of data preprocessing; (e) effective detection, exploration and identification of patterns and (f) very often discovering of the unexpected conclusions which were not noticed without multivariate approach, (g) creating interest and attention of the reader, (h) possibility of easy remembering of the conclusions, (i) time saving, (j) better compression than tabular and textual forms of presentation, (k) stimulation of analytical thinking and investigation.

2.1. Data preprocessing and compression

Data pre-processing is a necessary and critical step of the data mining process or knowledge discovery in data obtained in analytical experiments. Some data consist a lot of irregularities i.e. missing data, noise, inconsistent or even outliers. This type of data doesn't possess high quality information. If data quality is low then data mining tasks like data analysis, pattern reorganization, decision management do not give optimal solution. The goal of pre-processing is preparation of accurate, reliable and high quality data. Lack of this step may lead to incomplete interpretation and decision process. Various processing algorithms are adequate in specific problems. There are a number of data