International Journal of Pharmaceutics 491 (2015) 130–135

Contents lists available at ScienceDirect

International Journal of Pharmaceutics

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

Classification of drug tablets using hyperspectral imaging and wavelength selection with a GAWLS method modified for classification

Hiromasa Kaneko, Kimito Funatsu*

Department of Chemical System Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ARTICLE INFO

Article history:
Received 2 April 2015
Received in revised form 22 May 2015
Accepted 9 June 2015
Available online 17 June 2015

Keywords:
Hyperspectral imaging
Classification
Tablets
k-Nearest neighbor
Wavelength selection
Variable selection

ABSTRACT

Right drug tablets must be brought to the right places. We apply hyperspectral imaging, which can measure infrared spectra at many points on a two-dimensional plane, to classify tablets correctly. The k-nearest neighbor algorithm (kNN) is employed to classify tablets using a database including their spectra and true classes. Although classification accuracy is not 100%, we can correctly classify tablets overall, since spectra at many points are measured with spectroscopy and misclassification at some points does not have much influence on the final tablet classification result. In addition, we propose a wavelength selection method for classification. Genetic algorithm-based wavelength selection is applied to classification and combined with kNN, and thus, not wavelengths but wavelength-regions can be selected in classification problems. Through a case study, we confirmed that the proposed method could classify three kinds of tablets correctly and select appropriate wavelength-regions.

1. Introduction

When many kinds of drug tablets are handled, it is important to bring the right tablets to the right places. Misclassification, which means that drug tablets are wrongly classified, is a crucial problem. In addition, false tablets, which are dangerous for health, are sometimes marketed illegally. Three types of false drug tablets exist; those that do not contain any active pharmaceutical ingredient (API), the ones containing API that are not marked in the packing and drug tablets that contain the marked API, but are produced by a different manufacturer (World Health Assembly, 2002; Rodionova et al., 2005). Before we acquire them, they must be checked accurately, non-destructively and rapidly. Since false drug tablets have such a high quality nowadays, it is difficult to detect them.

Process analytical technology (Roggo et al., 2007; Hansuld and Briens, 2014) is an important technique for monitoring, developing, controlling and designing critical product quality in the pharmaceutical industry. Near infrared (NIR) spectroscopy (Muteki et al., 2013), Raman spectroscopy (Simone et al., 2014) and so on have been focused to monitor product quality non-destructively in real time.

In this study, we focus on hyperspectral imaging, which provides spatial and spectral information simultaneously. For samples on a scene, a hyperspectral image sensor outputs the spectrum for each pixel in the image. By using near-infrared spectroscopy as a spectroscopic method, measurements of samples can be performed accurately, non-destructively and rapidly. Hyperspectral imaging has been applied to prediction of protein and fat content in cheeses (Burger and Geladi, 2006), estimation of fruit yield in citrus (Ye et al., 2006), mapping piroxicam polymorphs (Rocha et al., 2011), assessment of regional leaf area index (Canisius and Fernandes, 2012), food authentication (Ottavian et al., 2012), discrimination between polylactic acid and polyethylene terephthalate (Ulrici et al., 2013) and monitoring of polymorphic transformation (Terra and Poppi, 2014). Pre-treatment (Esquerre et al., 2012) and spatial resolution (Offroy et al., 2012) have been discussed as well. Gowen et al. (2007) provided a review paper of hyperspectral imaging as a process analytical tool for food quality and safety control.

Alvarez-Jubete et al. (2013) constructed regression models between API concentrations and NIR spectra in pixels for hyperspectral imaging. Vajna et al. (2011) proposed to estimate concentrations in the extruded pharmaceutical by using Raman chemical imaging and multivariate curve resolution-alternating least squares (MCR-ALS). Carneiro and Poppi (2014) estimated different chemical compositions between crystals and cream using infrared imaging spectroscopy and MCR–ALS. Edelman et al. (2013) applied hyperspectral imaging to color classification of ecstasy

* Corresponding author. Fax: +81 3 5841 7771.
E-mail address: funatsu@chemsys.t.u-tokyo.ac.jp (K. Funatsu).

http://dx.doi.org/10.1016/j.ijpharm.2015.06.012
0378-5173/

tablets. They succeeded in determining differently colored tablets and estimating concentration of a colorant. Rodionova et al. (2005) classified a legitimate drug tablet and a counterfeit drug using hyperspectral imaging. To construct the classification model, a soft independent class modeling analogy, which is a supervised pattern recognition method, was applied to NIR spectra. However, they do not consider drug tablets whose true classes do not exist in training data. When spectra belonging to these tablets are input into a classification model, they are classified into one of the classes existing in training data, which is not true.

In this paper, our objective is to classify drug tablets including those whose true classes do not exist in training data, by using hyperspectral imaging. First, for several tablets whose drug classes are known, spectra are measured with a hyperspectral image sensor. Then, a statistical model is constructed between spectra and the class labels of the tablets. The classes of new tablets can be estimated by inputting their spectra measured with a hyperspectral image sensor into the constructed model. Although the classification accuracy of a model is not 100%, we can correctly classify tablets overall, since spectra at many points are measured with spectroscopy and misclassification at some points does not have much influence on the final tablet classification result.

There are several classification methods such as support vector machine (SVM) (Bishop, 2006) and partial least squares-discriminant analysis (Toher et al., 2007), but we employ k-nearest neighbor algorithm (kNN) (Kaneko and Funatsu, 2013a), which is a simple method, easy to be applied to multiclass classification, and can handle nonlinear relationships between spectra and class labels. By using kNN model, we can classify tablets only from spectra measured with a hyperspectral image. kNN can provide simpler models than SVM, which is designed for binary classification problems originally, and it can be applied to multi-class classification (Crammer and Singer, 2001). The performance of kNN is sufficiently high, as shown in Section 3. In addition, unknown classes, which do not exist in training data, can be handled using the distance threshold in kNN. When spectra of drug tablets, whose true classes do not exist in training data, are input into a classification model and exceed the threshold, they are classified as a class that does not exist in training data.

Wavelength selection of spectra is effective for the improvement of statistical models, including kNN models, and the cost reduction of spectral measurement. Many wavelength selection methods have been developed (Arakawa et al., 2011; Kim et al., 2011). Specifically, genetic algorithm-based wavelength selection (GAWLS), which has been used only for regression analysis, can select multiple wavelength-regions (Arakawa et al., 2011). It was confirmed that reasonable wavelength-regions could be selected and appropriate regression models could be constructed using the selected wavelength-regions, where GAWLS is superior over some other wavelength selection methods. GAWLS can be applied to time series data analysis (Kaneko and Funatsu, 2012) and nonlinear systems (Kaneko and Funatsu, 2013b).

We therefore apply GAWLS to classification problems in order to improve predictive ability of classification models and select appropriate wavelengths. Accuracy rate using cross-validation (CV) in kNN modeling is employed as a fitness function in a genetic algorithm (GA), where we will be able to select combinations of wavelength-regions with which predictive classification models can be constructed.

The effectiveness of our proposed method is demonstrated through a case study using real drug tablets whose spectra are measured with a hyperspectral image sensor. We show that drug tablets can be classified accurately and reasonable wavelength-regions can be selected. New classes that do not exist in training data can be handled by considering the distance threshold in kNN.

2. Method

2.1. kNN

kNN is a well-known classification method. To classify a query sample, we simply determine the class to which most k known samples are the closest to the query sample belong. We use the Euclidean distance between two samples as a distance measure.

We can handle samples whose classes do not exist in training data by setting a threshold for some distance in kNN as applicability domain (Kaneko and Funatsu, 2013a; Escobar et al., 2014; Kaneko and Funatsu, 2014), which is a data domain where a constructed model has highly predictive performance. The average of distances, the maximum distance, or the minimum distance from k known data that are the closest to the query data can be used, for example. In this paper, we employ the average of the distances from k known data and set the threshold using 5-fold cross-validation (CV). The threshold is the 99.7% cross-validated value of the average distances.

2.2. GAWLS (Arakawa et al., 2011)

GAWLS is one of the methods that is used to select combinations of important wavelengths (variables) from explanatory variables (X) using regions as a unit of measurement. A genetic algorithm (GA) (Leardi, 2001) is applied to select wavelengths. GA is an optimization method that is used in biology to model principles of natural evolution. Species having a high level of fitness under certain environmental conditions can prevail in the next generation, and the best species may be reproduced by crossover together with random chromosomes mutation in those species that survive. The solution space around superior individuals is preferentially explored, which leads to discovering a solution that is close to the optimum.

In GAWLS, two actual genes of a chromosome represent one region of wavelengths. Fig. 1 shows the coding method for GAWLS. Hence, GAWLS can select important wavelengths using regions as a unit of measurement. In Fig. 1, the number of selected wavelengths is 1265 through 1283 and 2057 through 2069, and the number of windows (NW) is 2. Partial least squares (PLS) model (Wold et al., 2001), which is a regression model, is constructed then with a set of selected wavelengths. The r_{CV}^2-value, which is the determinant coefficient calculated using a CV method is used as a fitness value for each chromosome. In this way, a set of an optimum wavelength-regions, with which highly predictive model can be constructed, is obtained.

The number of chromosomes, the number of generations, the maximum width of window and NW must be set beforehand. The minimum width is zero.

2.3. GAWLS for classification

We propose to apply GAWLS to classification problems. Fig. 2 shows the basic concept of the proposed method. To calculate a fitness value for each chromosome, CV based on kNN is performed

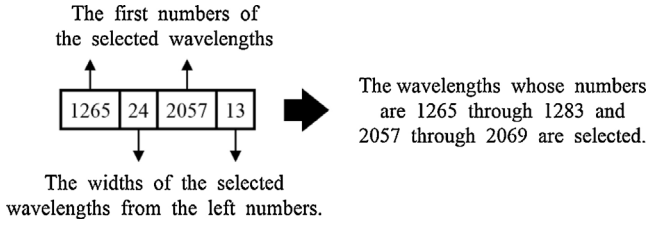

Fig. 1. Chromosome in GAWLS.