A Study on Classification of Suspect Powders with Spectral Data

Tianjia Chen
STAT 661, Fall 2015

1 Abstract

The paper uses the combination of Principal Component Analysis, Multivariate Regression and Support Vector Machine to build a classification model for benign and harmful substances with spectral data. Principal Component Analysis was used to select three components as new predictors. Multivariate Regression was used to select coefficients of two elements, Graphite and FeSO4. Support Vector Machine with radial kernel was used to construct final classification model. It turned out that the error rates for training set and testing set were 3% and 2% respectively. The study sheds light on classification of suspect powders with spectral data, and can further benefit future work in related areas.

2 Introduction

The detection of harmful substances is an urgent and crucial work when a building or area has been contaminated with some powder substances that possibly contain Bacillus spores (causative agent for anthrax)[1]. A successful recognition of noxious substances can further prevent the spread of anthrax, which under some cases may save people's life. Therefore, it is important to build up a classification mechanism to detect whether certain substances are harmful or not.
Laser-induced breakdown spectroscopy (LIBS) devices can generate characteristic spectra that aid in determining if a substance is or contains a spore material. In LIBS, each substance sample will produce a characteristic wavelength spectrum. The distribution of wavelengths in spectrums can be different for harmful and non-harmful substances. Therefore, the wavelength spectrum in LIBS can be used to construct a classification model, as is discussed in the rest of the paper.

3 Methods

The data come from a subset of the data used in Cisewski, J., Snyder, E., Hannig, J. and Oudejans, L. (2012), which are provided by Dr. Emily Snyder at the US Environmental Protection Agency, Oce of Research and Development, National Homeland Security Research Center, Research Triangle Park, NC27711.
12 substances are available in the dataset: three types of anthrax surrogates (anthrax1,

[1] Cisewski, Jessi, et al. "Support vector machine classification of suspect powders using laser-induced

anthrax2, anthrax3), Arm & Hammer Detergent (armh), Rumford Baking Powder (bakingpowder), Arm and Hammer Baking Soda (bakingsoda), Crayola Chalk (crayolachalk), Food Lion Brand flour (flour), Advil Ibuprofen tablets (ibuprofen), Food Lion Brand Sugar (sugar), Tide Laundry Detergent (tideldbaked) and Tylenol Acetaminophen Capsules (tylenolgelcap). All the substances except for antrhax1, anthrax2, anthrax3 are benign. For each substance, it contains the data of 12 samples. For each sample, there are 13,701 wavelengths at which measurements were recorded (198.1645 to 972.6262 nm). Additional datasets include five revised sample spectra of 9 elements that are found in anthrax surrogates (B2O3, graphite, CaClO3, FeSO4, KI, MgSO4, MnSO4, NaCl, Si).

The following section will discuss about three topics: (1) How to use Principal Components Analysis (PCA) to construct new components for final classification model. (2) How to use Multivariate Regression to construct new predictors based on 9 elements for final classification model. (3) How to use Support Vector Machine to build and test final classification model.

3.1 Constructing New Predictors Using PCA

There are two motivations for using PCA to construct new predictors: (a) Harmful and non-harmful substances may have different performances in wavelength spectrums (see Figure 1). (b) The number of wavelengths is 13,701, which is too large when there are only 144 observations. Therefore, PCA is needed to reduce the dimensions of predictors while preserving the important information in predictors.

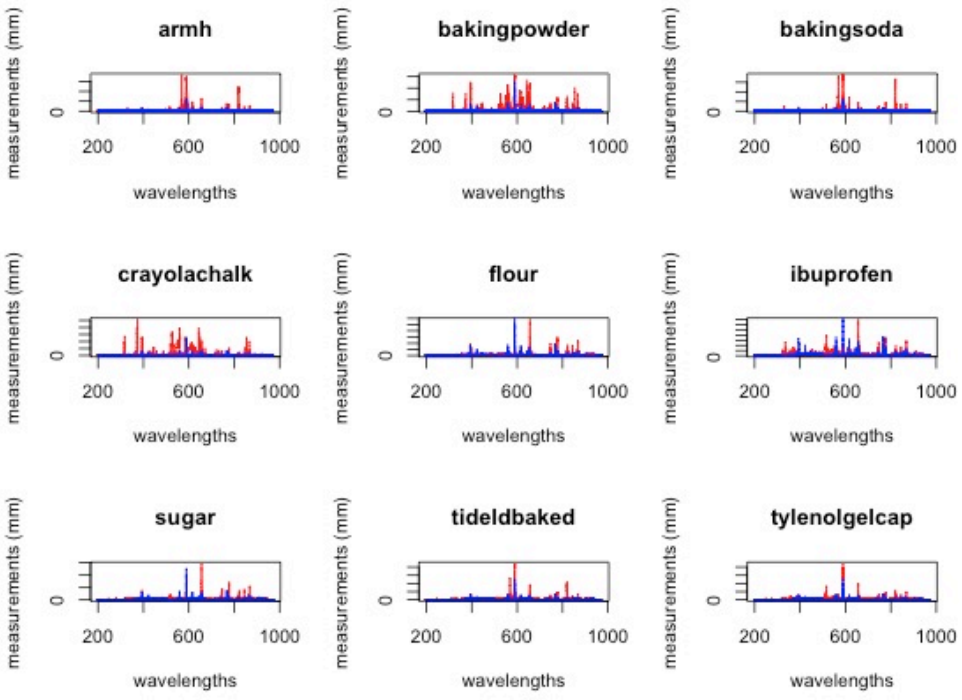

Figure 1. Average Wavelength for Benign(red line) Substances and Anthrax(blue line).

According to the variances of components, the first three components were chosen as predictors in the final classification model. After calculating three components for all 144 samples, the distribution of average components for nine benign components and anthrax can be found in Figure 2.

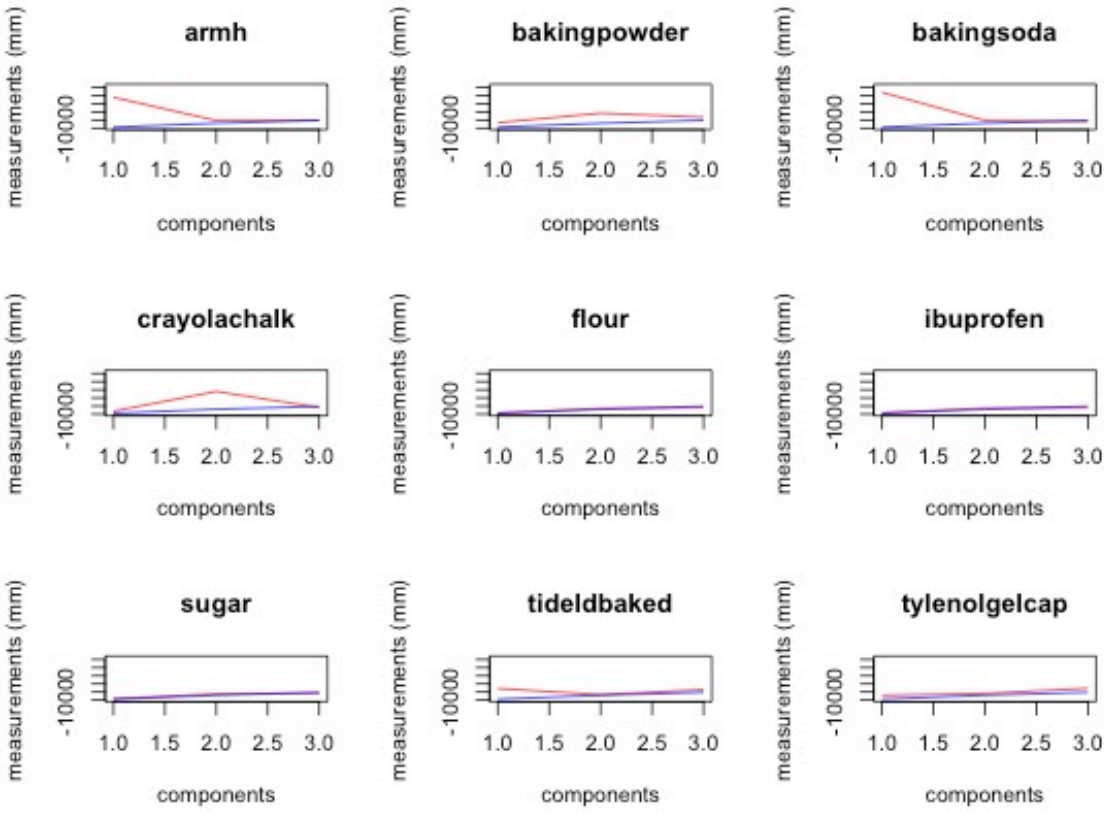

Figure 2. Average Three Components for Benign (red line) Substances and Anthrax (blue line).

From Figure 2, it can be found that the performances of three components were different between anthrax and substances like armh, bakingpowder, bakingsoda, crayolachalk, tideldbacked. However, for four substances of flour, ibuprofen, sugar and tylenolgelcap, three components can not separate them from anthrax. Therefore, more information is needed to construct new predictors to complement three components.

3.2 Constructing New Predictors Using Multivariate Regression

Given the information that nine elements were found in anthrax, it is possible to construct some new predictors based on the relationship between wavelengths of substances and elements.

The following steps were followed to construct new predictors:

(1) Calculate the median of measurements for each wavelength among samples for substances and elements.

(2) Regress each substance on nine elements and extract the coefficients.
(3) Plot nine coefficients for nine benign substances and anthrax (see Figure 3).

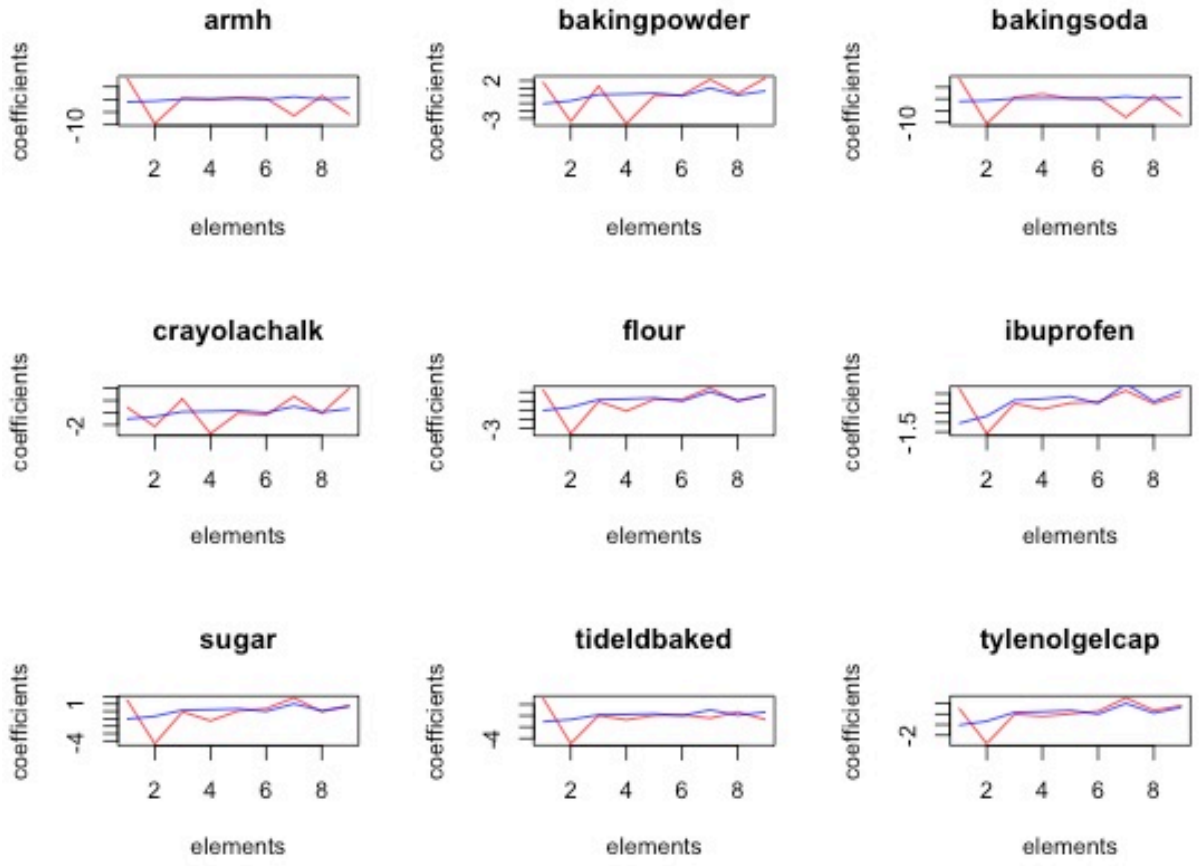

Figure 3. Coefficients for Nine Benign Substances (red line) and Anthrax (Blue line)

From Figure 3, it can be found that for four substances of flour, ibuprofen, sugar and tylenolgelcap, coefficients for the second and fourth elements have relatively big difference between benign substances and anthrax. Therefore, the coefficients for Graphite and FeSO4 were added as two new predictors in the final classification model.

3.3 Constructing Classification Model Using Support Vector Machine

Support Vector Machine with radial kernel was used for the final classification model. Before modeling, the whole 144 observations were divided into training set and testing set with the proportion of 2:1.

4 Results

The results of training error and testing error can be found in Table 1 and Table 2, with error rate of 3% and 2% respectively.

Type	True Benign (0)	True Harmful (1)
Predicted Benign (0)	69	0
Predicted Harmful (1)	3	24

Table 1. Results of Classification for Training Set

Type	True Benign (0)	True Harmful (1)
Predicted Benign (0)	35	0
Predicted Harmful (1)	1	12

Table 2. Results of Classification for Testing Set

The results of classification for 10 new substances can be found in Table 3.

Test Case Substance	Classification	Test Case Substance	Classification
1	benign (0)	6	benign (0)
2	benign (0)	7	benign (0)
3	benign (0)	8	benign (0)
4	benign (0)	9	benign (0)
5	benign (0)	10	benign (0)

Table 3. Results of Classification for Test Case Substances

5 Discussion

The paper used Principal Component Analysis and Multivariate Regression to construct five new predictors that best separate benign and harmful substances, and applied Support Vector Machine with radial kernel to do the classification. It turned out that the first three components in PCA and coefficients of Graphite and FeSO4 can best depict the differences between benign and harmful substances. The error rates for training set and testing set are 3% and 2% respectively.

The proposed method may have following shortcomings: (1) when selecting the number of coefficients, the method only looked at each component's ability to classify two categories separately. An overview incorporating the correlation between components may be needed to choose the proper number of components; (2) when selecting the coefficients for

elements, the method only chose Graphite and $FeSO_4$ based on four substances that are not separated by three components. However, it is entirely possible that other elements can be of great benefit to the final classification accuracy; (3) other classification methods like Random Forest could possibly outperform SVM in this problem.

Based on the shortcomings of the proposed method, more analysis can be done in these areas: (1) try more groups of components as predictors; (2) incorporate more combinations of elements as predictors; (3) explore other classification algorithms; (4) find out other possible dimension reduction methods to construct predictors for the final classification model.