Analytica Chimica Acta 829 (2014) 1–8

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data

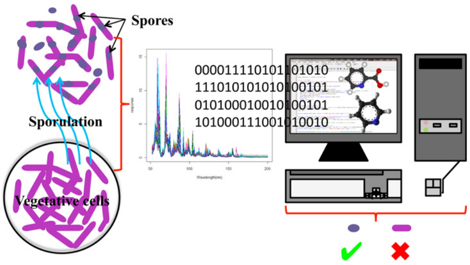
Piotr S. Gromski [a], Yun Xu [a], Elon Correa [a], David I. Ellis [a], Michael L. Turner [b], Royston Goodacre [a,*]

[a] School of Chemistry, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK
[b] School of Chemistry, The University of Manchester, Brunswick Street, Manchester M13 9PL, UK

HIGHLIGHTS

- LDA, PLS-DA, SVM and RF analyses were applied to MS data.
- Double cross-validation using bootstrapping was employed to assess models.
- For all classifications, all bacteria were assessed with ~95% accuracy.
- Parsimonious modelling was used on a reduced set of mass ions and was more robust.
- The approaches developed are equally applicable to any multivariate data.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 23 January 2014
Received in revised form 24 March 2014
Accepted 27 March 2014
Available online 31 March 2014

Keywords:
Variable selection
Supervised learning
Bootstrapping
Double cross-validation
Pyrolysis mass spectrometry
Bacillus

ABSTRACT

Many analytical approaches such as mass spectrometry generate large amounts of data (input variables) per sample analysed, and not all of these variables are important or related to the target output of interest. The selection of a smaller number of variables prior to sample classification is a widespread task in many research studies, where attempts are made to seek the lowest possible set of variables that are still able to achieve a high level of prediction accuracy; in other words, there is a need to generate the most parsimonious solution when the number of input variables is huge but the number of samples/objects are smaller. Here, we compare several different variable selection approaches in order to ascertain which of these are ideally suited to achieve this goal. All variable selection approaches were applied to the analysis of a common set of metabolomics data generated by Curie-point pyrolysis mass spectrometry (Py-MS), where the goal of the study was to classify the Gram-positive bacteria *Bacillus*. These approaches include stepwise forward variable selection, used for linear discriminant analysis (LDA); variable importance for projection (VIP) coefficient, employed in partial least squares-discriminant analysis (PLS-DA); support vector machines-recursive feature elimination (SVM-RFE); as well as the mean decrease in accuracy and mean decrease in Gini, provided by random forests (RF). Finally, a double cross-validation procedure was applied to minimize the consequence of overfitting. The results revealed that RF with its variable selection techniques and SVM combined with SVM-RFE as a variable selection method, displayed the best results in comparison to other approaches.

1. Introduction

Pyrolysis mass spectrometry (Py-MS) is a well-established analytical technology which has been used for the characterization

Abbreviations: LDA, linear discriminant analysis; PLS-DA, partial least squares-discriminant analysis; SVM, support vector machines; RF, random forests.

[*] Corresponding author. Tel.: +44 161 306 4480; fax: +44 2004636.
E-mail address: roy.goodacre@manchester.ac.uk (R. Goodacre).

http://dx.doi.org/10.1016/j.aca.2014.03.039
0003-2670/© 2014 Elsevier B.V. All rights reserved.

of microbial systems for several decades [1,2]. This is due to its high discriminatory ability [3], and it has been said to be a powerful 'fingerprinting' technique [4,5], which can be routinely applied to any organic material. As its names suggests, this technology involves the thermal degradation of complex organic molecules. This process takes place either in an inert atmosphere or within a vacuum, whereby complex organic substances are rapidly broken down into stable primary fragments (termed pyrolysate) which can then be measured by mass spectrometry. Depending on the type of pyrolysis used, rapid heating (from 300 to 1000 °C) of samples can occur on a metal filament or within a quartz sample tube/cuvette, and whilst this is a destructive method, it routinely allows for the reproducible analysis of samples of less than 1 mg [6].

Py-MS has been successfully used to discriminate complex tertiary mixtures of bacterial species (*i.e. Bacillus subtilis, Escherichia coli, Staphylococcus aureus*) [3], urinary tract infection bacteria [7], and many other microbial studies. More recently, these include changes in bacterial activated sludge populations [8], and the study of bacterial and fungal bioremediation systems [6]. Other more diverse applications include an assessment of olive oil adulteration [9,10]. Indeed, in combination with advanced statistical methods, and its ability to be applied to a diverse range of samples of biotechnological interest, it has in the past been termed an 'anything-sensor' [11], which has also been aptly demonstrated with numerous non-biological applications [12–14].

Due to the complexity of the output from Py-MS, where the data may be highly collinear and the majority of descriptors unrelated to the study, variable selection prior to pattern recognition is a crucial component of the analysis. This parsimonious approach to modelling [15] is needed in order to obtain robust and reproducible results. In the case of Py-MS for example, where we measure 150 mass-to-charge (m/z) intensities as input variables (range 51–200 m/z), it would indeed be desirable to construct a model with a reduced number of variables prior to classification [16–18].

Up until now, several studies have employed different approaches for feature selection and classification of Py-MS such as genetic algorithms, which have been successfully used as variable selection techniques in combination with multiple linear regression and partial least squares (PLS) regression [19]. Other statistical methods that have been applied to Py-MS data include variable selection in discriminant PLS analysis [20], PLS-discriminant analysis (PLS-DA) [21] which has been used for the discrimination of *B. subtilis* strains [22], Fisher's linear discriminant analysis (LDA) [23] which has been successfully employed to distinguish the difference between tobacco types [24], as well as support vector machines (SVM) [25–28] which was successfully implemented for the analysis of Py-GC–MS data [29].

The aim of this study is to compare various variable selection methods which are commonly used for the analysis of chemical data. To this end we employed LDA, PLS-DA, SVM and random forests (RF) [30], and until now the latter has not to our knowledge been used to analyse Py-MS data. The Py-MS data had been previously collected [17] from various bacteria belonging to different *Bacillus* species where the aim was to effect both species classification as well as being able to recognise the bacterial physiological state correctly: all bacteria were cultivated either as spores or as vegetative biomass.

2. Materials and methods

All statistical data analyses were performed using the R (2.15.0) [31] software environment. This language comprises a selection of packages suitable for different types of data and is available as free software in the public domain.

In this study, we used different approaches for feature reduction and classification of (1) the physiological states of the genus *Bacillus* (spores *versus* vegetative cells), as well as (2) differentiating seven *Bacillus* species (*B. amyloliquefaciens, B. cereus, B. licheniformis, B. megaterium*, *B. subtilis* (including *B. niger* and *B. globigii*), *B. sphaericus* and *Brevibacillus laterosporus*) [17]. This comparative analysis (Fig. 1) is based on the following combinations of variable selection methods and their respective classification algorithms: stepwise forward variable selection that has been combined with LDA [32]; PLS-DA and its variable importance for projection (VIP) coefficient [33]; SVM recursive feature elimination (SVM-RFE) to reduce the number of variables prior to SVM [34]; and finally, mean decrease in accuracy and Gini provided by RF [35]. These techniques have been used to establish variable importance, as well as to reduce input dimensionality and computational load/time.

2.1. Data

The original dataset used in this study for variable selection and classification was collected by Goodacre et al. [17] using Curie-point Py-MS from seven different types of *Bacillus*, where a detailed description of data collection and instrumentation can be found.

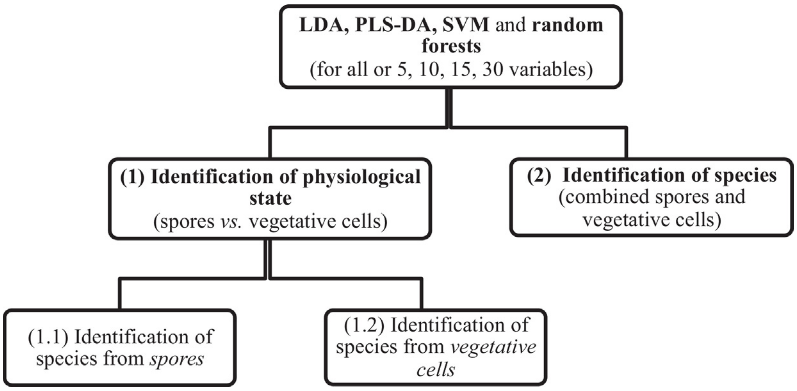

LDA, PLS-DA, SVM and random forests
(for all or 5, 10, 15, 30 variables)
(1) Identification of physiological state
(spores vs. vegetative cells)
(2) Identification of species
(combined spores and vegetative cells)
(1.1) Identification of species from spores
(1.2) Identification of species from vegetative cells

Fig. 1. The overall workflow of the studies for variable importance ranking for the analysis of four data sets. (1) Physiological state estimation where the prediction accuracy has been calculated for the separation of vegetative cells from spores. Following this separation two subsets were analysed for *Bacillus* speciation: (1.1) identification of *Bacillus* species from spores; (1.2) identification of *Bacillus* species from vegetative cells. (2) Illustrates the analysis of species when both spores and vegetative cells are analysed together in the same model.