FEBS Letters xxx (2015) xxx–xxx

journal homepage: www.FEBSLetters.org

Mass spectrometry cancer data classification using wavelets and genetic algorithm

Thanh Nguyen *, Saeid Nahavandi, Douglas Creighton, Abbas Khosravi

Centre for Intelligent Systems Research (CISR), Deakin University, Waurn Ponds Campus, Victoria 3216, Australia

ARTICLE INFO

Article history:
Received 12 November 2014
Revised 12 November 2015
Accepted 16 November 2015
Available online xxxx

Edited by Paul Bertone

Keywords:
Mass spectrometry data
Cancer classification
Wavelet transformation
Genetic algorithm
Feature extraction

ABSTRACT

This paper introduces a hybrid feature extraction method applied to mass spectrometry (MS) data for cancer classification. Haar wavelets are employed to transform MS data into orthogonal wavelet coefficients. The most prominent discriminant wavelets are then selected by genetic algorithm (GA) to form feature sets. The combination of wavelets and GA yields highly distinct feature sets that serve as inputs to classification algorithms. Experimental results show the robustness and significant dominance of the wavelet-GA against competitive methods. The proposed method therefore can be applied to cancer classification models that are useful as real clinical decision support systems for medical practitioners.

1. Introduction

Mass spectrometry (MS) is a powerful analytical chemistry technique that was initially introduced to determine the constituent elements of small molecules. Mass spectrometers consist of three main parts: an ion source, a mass analyser, and an ion detection system [1]. Components of a sample mixture are converted to ions, which are then bombarded with an electron beam having sufficient energy. In Fig. 1, the high voltage beams are to accelerate the ions in the target sample so that they all have the same kinetic energy. The positively charged ions are deflected in a vacuum through a magnetic field depending on their masses. Ions are deflected more if they are lighter. The amount of ions passing through the machine is detected electrically and is sorted on the basis of mass-to-charge (m/z) ratio. The machine is calibrated to record the ion current against the m/z ratio. The output of the recorder is a spectrum presented in a diagram where the vertical axis represents the relative abundance or relative intensity and the horizontal axis represents the m/z ratio (see Fig. 1).

MS-based proteomics has been routinely applied worldwide to deal with a large range of biological problems [2]. More specifically, it is able to discover patterns of differentially expressed proteins in clinical samples such as blood serum. Biomarkers identified through analysis of complex protein mixtures can be utilized for diagnosis, prognosis, or monitoring of many diseases, in particular cancers, e.g. see [3–11].

MS data are commonly assembled with the number of m/z values much larger than the number of samples. Standard techniques therefore find inappropriate or computationally infeasible in analysing such data. Not all of the tens of thousands of m/z values are discriminative and needed for classification. Most m/z values do not affect the classification performance. Taking such m/z values into account enlarges the dimension of the problem, leads to computational burden, and presents unnecessary noise in the classification process. It is essential to have a feature extraction procedure that is able to reduce dimension of the data and form a feature set, which suffices for good classification.

Common feature extraction approaches are filter and wrapper methods. Filter methods rank all features in terms of their goodness using the relation of each single feature with the class label based on a univariate scoring metric. The top ranked features are chosen before classification techniques are carried out. In contrast, wrapper methods require the feature selection technique to combine with a classifier to evaluate classification performance of each feature subset. The optimal subset of features is identified based on the ranking of performance derived from implementing the classifier on all found subsets. The filter procedure is unable to measure the relationship among features whilst the wrapper approach

Author contributions: Conceived and designed the experiments: TN. Performed the experiments: TN, AK. Analyzed the data: TN, SN. Contributed reagents/materials/analysis tools: TN AK DC SN. Wrote the paper: TN, SN, DC, AK.

* Corresponding author. Fax: +61 3 52271046.
E-mail address: thanh.nguyen@deakin.edu.au (T. Nguyen).

http://dx.doi.org/10.1016/j.febslet.2015.11.019
0014-5793/

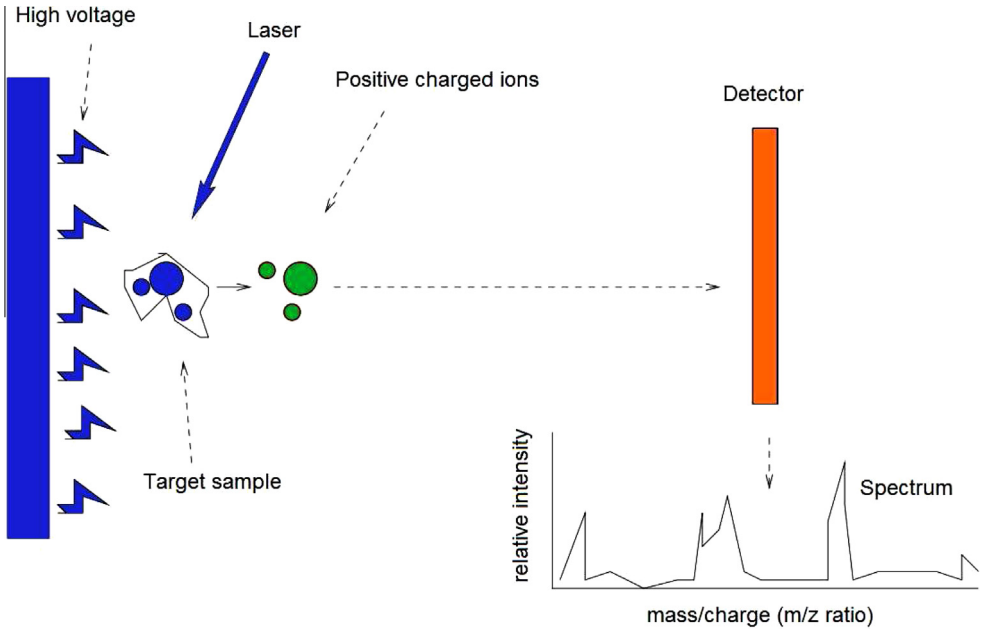

Fig. 1. Mass spectrometry process.

requires a great computational expense. Therefore, the combination of filter and wrapper approaches has a potential to accumulate advantages of each individual method [12].

In this paper, to enhance the robustness and stability of mass spectrometry data classification, we introduce a feature extraction method by combining wavelet transformation (WT) and genetic algorithm (GA), called wavelet-GA. The idea behind this approach is to first transform mass spectrometry data into orthogonal wavelet coefficients using the Haar wavelets. Then GA is applied to select the most prominent discriminant wavelet coefficients to form feature sets. The GA search based on the evolutionary learning process is considered as a wrapper feature selection method. We integrate the two-sample *t*-test filter method into the GA population initialization process to benefit the advantages of this filter method during the GA implementation. Accordingly, the proposed approach is regarded as a hybrid method that incorporates a filter method into a wrapper procedure based on wavelet features.

Next section describes in detail the proposed wavelet-GA method. Experiments and discussions are presented in Section 3, followed by concluding remarks and future research directions in Section 4.

2. Proposed wavelet-genetic algorithm feature extraction

2.1. Wavelet transformation (WT)

WT represents a signal in a time-frequency fashion [13]. WT eliminates the requirement of signal stationarity that usually applies to conventional methods. Once the wavelets (the mother wavelet) $\varphi(x)$ is fixed, translations and dilations of the mother wavelet can be formed $\{\varphi((x-b)/a), (a,b) \in R^+ \times R\}$. It is useful to set specific values for a and b as $a = 2^{-j}$ and $b = 2^{-j}k$ where j and k are integer numbers.

Du et al. [14] introduced the R package MassSpecWavelet for processing mass spectrometry spectrum by using wavelet-based algorithms. One of the simplest wavelets is the Haar wavelet $\varphi(x)$, which has been used in various areas. It is a step function that takes values at 1 and -1 on $[0,\frac{1}{2})$ and $[\frac{1}{2},1)$ respectively. Fig. 2 graphically illustrates the Haar wavelet.

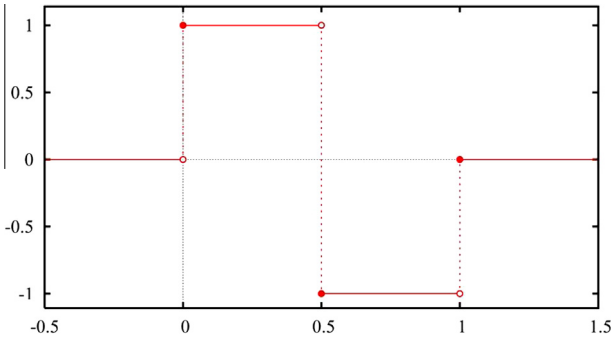

Fig. 2. An illustration of Haar wavelet.

In general, Haar functions can uniformly approximate any continuous function. Dilations and translations of the function φ, which is $\varphi_{jk}(x) = const \cdot \varphi(2^j x - k)$, define an orthogonal basis in $L^2(R)$. This means that any element in $L^2(R)$ may be represented as a linear combination of these basis functions. The scaling function in Haar wavelet is simply unity on the interval [0,1) as $\phi(x) = 1(0 \leqslant x < 1)$.

2.2. Genetic algorithm (GA) for selection of wavelets

GA is generally the most robust evolutionary algorithm. GA has the capability to deal with problems that may be non-differentiable, non-linear, or have many local minima or constraints. If these characteristics are strongly present, GA offers effective solutions, e.g. see [15,16] where GA was successfully employed in computational biology.

GA is an optimization technique operated on a population of L artificial individuals. Individuals are characterized by chromosomes (or genomes) $S_k,\ k = \{1,\ldots,L\}$. The chromosome kth is a string of symbols, which are called genes, $S_k = (S_{k1},\ldots,S_{kM})$, where M is the string length.

In the application of GA for selection of wavelet coefficients, a gene represents a coefficient. The number of genes in a

Please cite this article in press as: Nguyen, T., et al. Mass spectrometry cancer data classification using wavelets and genetic algorithm. FEBS Lett. (2015), http://dx.doi.org/10.1016/j.febslet.2015.11.019