FEBS Letters xxx (2015) xxx–xxx

ELSEVIER

FEBS Letters

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

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

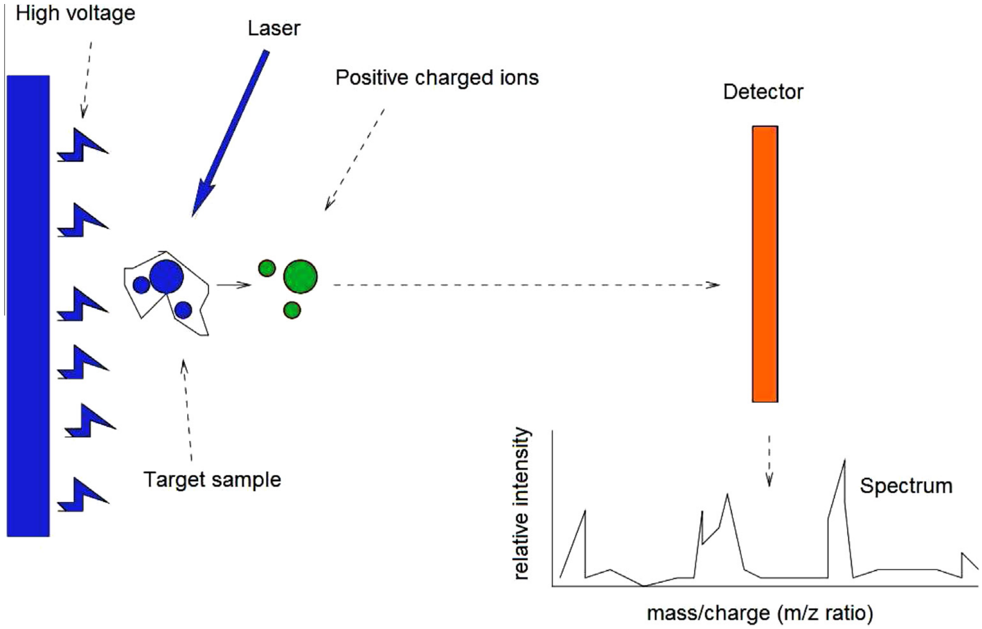

Fig. 1. Mass spectrometry process.

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.

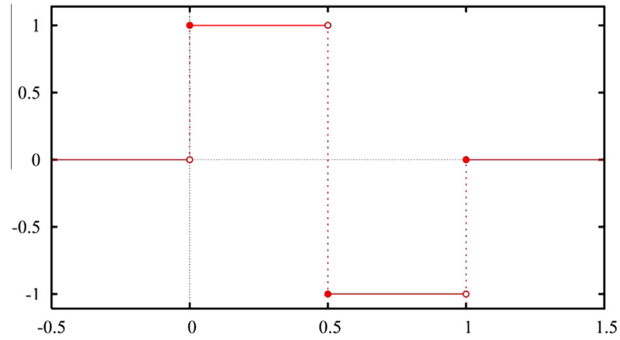

Fig. 2. An illustration of Haar wavelet.

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) $S_k,\ k=\{1,\dots,L\}$. The chromosome kth is a string of symbols, which are called genes, $S_k=(S_{k1},\dots,S_{kM})$, 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

chromosome (individual) is equal to the wanted number of features in the feature set. Therefore, not all available wavelet coefficients are used in each GA evaluation step because the chromosome length is equal to the wanted number of features, which is often small (for example five features in this study). This is an advantage that diminishes the overfitting issue without regularisation on the large wavelet dictionary, which was addressed by a sparse Bayesian approach in Vannucci et al. [17].

The GA population comprises individuals where each individual represents a solution. The initial population is initialized by randomly sampling the set of prominent wavelets that are selected by the two-sample *t*-test filter method. This test is a parametric hypothesis test that is applied to compare whether the average difference between two independent data samples is really significant. The test statistic is expressed by:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1}$$

where μ_1 and μ_2 are sample means, σ_1 and σ_2 are sample standard deviations, and n_1 and n_2 are sample sizes of the two separate datasets.

In the application of *t*-test for selecting prominent wavelets, the test is performed on each wavelet coefficient by separating the data samples based on the class variable. The absolute value of t is used to evaluate the significance of wavelets. The higher the absolute value, the more important is the wavelet coefficient.

The integration of this filter method in the GA population initialization step enables GA to have a more insightful knowledge about the solution space of the optimization problem. The proposed approach for feature extraction therefore is considered as a hybrid approach combining the filter and wrapper methods, which are applied to wavelet features.

Through chromosomes' evolution, GA searches for the best solution(s) in the sense of the given fitness function. The fitness function is designed as the linear combination of the error rate and the average of posterior probability of the classifier:

$$fit = ER + 1 - \overline{PP} \tag{2}$$

where ER is the classification error rate and $\overline{PP}$ is the average of the posterior probabilities that the jth training class was the source of the ith sample observation, i.e., $Pr(class_j|obs_i)$. Linear discriminant analysis (LDA) [18] is employed as the classifier to evaluate each individual of the population.

Each individual in the population represents a set of features. LDA is used to classify all samples and classification error rate is defined as the number of incorrectly classified samples divided by the total number of samples. Prior defaults to a numeric vector of equal probabilities, i.e., a uniform distribution: the prior probability of class k is 1 over the total number of classes. The posterior probability that a point x belongs to class k is the product of the prior probability and the multivariate normal density. The density function of the multivariate normal with mean μ_k and covariance Σ_k at a point x is

$$P(x|k) = \frac{1}{(2\pi|\Sigma_k|)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right) \tag{3}$$

where $|\Sigma_k|$ is the determinant of Σ_k, and Σ_k^{-1} is the inverse matrix. Let $P(k)$ represent the prior probability of class k. Then the posterior probability that an observation x is of class k is

$$\widehat{P}(k|x) = \frac{P(x|k)P(k)}{P(x)} \tag{4}$$

where $P(x)$ is a normalization constant, namely, the sum over k of $P(x|k)P(k)$.

By combining classification error and the average posterior probabilities, the fitness function evaluates not only the accuracy of class indication but also the membership level of the observation within the indicated class. This combination measures more detailed the performance of the classification results so that the evaluation of feature sets is more rigorous and robust.

Individuals are evaluated via calculation of the fitness function. To evolve through successive generations, GA performs three basic genetic operators: selection, crossover and mutation. The population of GA is initialized with 100 individuals and evolved through 50 generations. Other parameters are assigned to the default settings of GA implemented in Matlab [19].

Specifically, the selection operator allows individuals in the current generation with the best fitness values to automatically survive to the next generation. In this study, the stochastic uniform selection function is used to select parents for the next generation. Each parent occupies a section of the line with the length proportionate to its fitness value. The algorithm moves along the line and selects a parent from the section it lands on. Number of best individuals that survive to next generation without any change is equal to 5.

Crossover creates children by combining the genes of a pair of two parents. We used the scattered crossover function where a child is formed based on a random binary vector. Genes where the vector is a 0 from the first parent and genes where the vector is a 1 from the second parent are combined for the child. The crossover fraction is 0.8.

Mutation functions make small random changes in the genes of individuals in the populations, which provide more genetic diversity and enable the GA to search in a broader space of solutions. The Gaussian mutation function is used in the GA implementation. A random number generated by a Gaussian distribution with mean 0 is added to each entry of the parent vector. The standard deviation of this distribution is specified by the parameters Scale and Shrink, which are both equal to 1 in this study.

The best chromosome obtained through a series of evolving generations represents the optimal set of wavelets. With a population of individuals, GA can simultaneously explore different parts of the feature space and it is thus able to find an global solution for the optimal feature set.

3. Experiments and results

3.1. Performance evaluation

To evaluate performance of the proposed feature extraction method, we apply a range of classifiers such as LDA [18], NB [20], k-nearest neighbours (kNN) [20], support vector machine (SVM) [21], multilayer perceptron (MLP) [22], fuzzy ARTMAP [23], adaptive neuro-fuzzy inference systems (ANFIS) [24], and ensemble learning AdaBoost [25].

The purpose of the classifiers is to verify feature sets and therefore each classifier was trained with the same initial training parameters for different feature sets obtained from different feature selection methods. The built-in Matlab implementation of LDA, NB, kNN, SVM, MLP and ANFIS and AdaBoost are used in this study. Specifically, the number of nearest neighbours in kNN is equal to 5 and the SVM kernel function is the Gaussian radial basis function with the scaling factor of 1. MLP is constructed with two hidden layers and five nodes are in each layer. We initialize ANFIS models with five fuzzy rules of the Sugeno type and they are trained through 50 epochs. AdaBoost uses a collection of individual learners that are 100 decision trees.

Fuzzy ARTMAP is a variant of the basic learning networks. Carpenter [26] defined the default ARTMAP algorithm and its

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

parameter values to describe a ready-to-use general-purpose neural network system for supervised learning and recognition. The default ARTMAP network simplifies the design and ensures robust performance of fuzzy ARTMAP in many application domains. Fuzzy ARTMAP is actually a subset of default ARTMAP (fuzzy ARTMAP $\subset$ default ARTMAP) [26]. Therefore, default parameter values presented in [26] are used in this study.

Numerous feature extraction approaches including two-sample *t*-test [27], entropy test (known as Kullback–Liebler distance or divergence) [27], Bhattacharyya distance (BD) [28], Wilcoxon test [29], receiver operating characteristic (ROC) curve [27], principal component analysis (PCA) [30], sequential search [31] are applied as competing methods against wavelet-GA. Whilst PCA and sequential search are employed based on the built-in Matlab implementation, details of the other feature extraction methods are presented in the following. These methods rank features via scoring metrics, which are statistic tests based on two sets of data samples in the binary classification problem. The sample means are denoted as μ_1 and μ_2, whereas σ_1 and σ_2 are the sample standard deviations.

- **Entropy test**

Relative entropy, also known as divergence, is a test assuming classes are normally distributed. The entropy score for each feature is computed using the following expression [27]:

$$e = \frac{1}{2}\left[\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)(\mu_1 - \mu_2)^2\right] \qquad (5)$$

After the computation is accomplished for every feature, features with the greatest entropy scores are selected to serve as inputs to the classification techniques.

- **Bhattacharyya distance**

The Bhattacharyya distance can be calculated from the standard deviation and mean of each class as follows:

$$\mathrm{BD} = \frac{1}{4}\ln\left[\frac{1}{4}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + 2\right)\right] + \frac{1}{4}\left[\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right] \qquad (6)$$

- **Receiver operating characteristic (ROC) curve**

Denote the distribution functions of X in the two populations as $F_1(x)$ and $F_2(x)$. The tail functions are specified respectively $T_i(x) = 1 - F_i(x), i = 1, 2$. The ROC is given as follows:

$$\mathrm{ROC}(t) = T_1\left(T_2^{-1}(t)\right), t \in (0, 1) \qquad (7)$$

and the area under the curve (AUC) is computed by:

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(t)dt \qquad (8)$$

The larger the AUC, the less is the overlap of the classes. Features with the greatest AUC therefore are chosen to form a feature set.

- **Wilcoxon method**

The Wilcoxon rank sum test is equivalent to the Mann–Whitney *U*-test, which is a test for equality of population locations (medians). The null hypothesis is that two populations enclose identical distribution functions whereas the alternative hypothesis refers to the case two distributions differ regarding the medians. The normality assumption regarding the differences between the two samples is not required. That is why this test is used instead of the two-sample *t*-test in many applications when the normality assumption is concerned.

The main steps of the Wilcoxon test [29] are summarized below:

(1) Assemble all observations of the two populations and rank them in the ascending order.
(2) The Wilcoxon statistic is the sum of all of the ranks associated with the observations from the smaller group.
(3) The hypothesis decision is made based on the *P*-value, which is found from the Wilcoxon rank sum distribution table.

In the applications of the Wilcoxon test for feature selection, the absolute values of the standardized Wilcoxon statistics are utilized to rank features.

Performance of classification techniques are measured by accuracy, area under the ROC curve (AUC), F1 score statistics (F-measure), and mutual information (MI).

F-measure considers both the "Precision" (denoted as *Pr*) and "Recall" (*Re*) of the procedure to compute the score expressed as follows:

$$\text{F-measure} = 2 \times \frac{Pr \times Re}{Pr + Re} \qquad (9)$$

The MI between estimated and true label is calculated by:

$$I(\widehat{C}, C) = \Sigma_{\hat{c}=0}^{1}\Sigma_{c=0}^{1} p(\hat{c}, c)\log\frac{p(\hat{c}, c)}{p(\hat{c})p(c)} \qquad (10)$$

where $p(\hat{c}, c)$ is the joint probability distribution function of estimated and true class labels $\widehat{C}$ and C, and $p(\hat{c})$ and $p(c)$ are the marginal probability distribution functions of $\widehat{C}$ and C respectively.

The five-fold cross validation is employed for experiments. The strategy divides all samples at random into 5 distinct subsets and 4 subsets are used for training classifiers whilst the last subset is for testing.

For unbiased comparisons among feature extraction methods, each classifier is repeated 30 times on a feature subset using five-fold cross validation and the average performance is reported. To draw convincing conclusions in performance evaluation, we implement the Kruskal–Wallis test [32] for comparing two sets of accuracy results. The Kruskal–Wallis test is a nonparametric version of the classical one-way ANOVA. As the results over 30 trials may not be normally distributed, they may violate the normal assumption of the ANOVA. Therefore the use of Kruskal–Wallis test is appropriate. The test returns the *P*-value for the null hypothesis that all samples in two sets of results are drawn from the same population.

Note that the test is performed to compare between the two sets of 30 accuracy outcomes generated by each classifier performed on two feature sets. One feature set is obtained by the proposed wavelet-GA approach and another is attained by one of competitive feature extraction methods.

3.2. Datasets

Three benchmark datasets including ovarian cancer, prostate cancer, and premalignant pancreatic cancer are used for experiments. The data are from the FDA-NCI Clinical Proteomics Program Databank.

The ovarian dataset was generated using the WCX2 protein chip. An upgraded PBSII SELDI-TOF mass spectrometer was used to produce the spectra. The dataset is composed of samples of proteomic patterns in serum that distinguish ovarian cancer from

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

non-cancer. There are 15154 m/z values and 253 samples where control (normal) samples contribute 35.97% with 91 examples and the rest 64.03% with 162 instances are cancer.

The prostate dataset was produced by the H4 protein chip and a Ciphergen PBS1 SELDI-TOF mass spectrometer. The dataset consists of 15154 m/z values with 322 samples of which 253 samples (78.57%) are normal and the remaining 21.43% are cancer with 69 examples.

Alternatively, the pancreatic cancer dataset comprises 6771 m/z values with 181 samples by using a randomly commingled study set of murine sera. Among them, normal samples account for 55.8% with 101 instances whilst the rest 80 samples (44.2%) are cancer.

The mass spectrographs (average across class) of these datasets displayed in Figs. 3–5 show the increasingly difficult classification from the ovarian dataset to prostate and pancreatic datasets.

Haar wavelets are applied to transform MS data in each dataset into wavelet coefficients. In this study, we choose feature sets with five wavelets that are selected by GA for demonstrations. For comparisons, the same number of features is also extracted by other competing feature extraction methods.

A different number of features can be used for demonstrations. Technically, adding more features to a certain extent will lead to improved classification accuracy. After that extent, adding more features would reduce the classification results because of the overfitting issue. In this paper, we are not focusing on finding out how many features will lead to maximum classification accuracy but it could be of a further work.

Figs. 6–8 demonstrate the Example 3D projections of feature sets obtained by the proposed wavelet-GA method. The projections are obtained by plotting the first three most significant features out of five features obtained from the wavelet-GA algorithm. "Class 1" indicates cancer samples whilst healthy samples are represented by "Class 2".

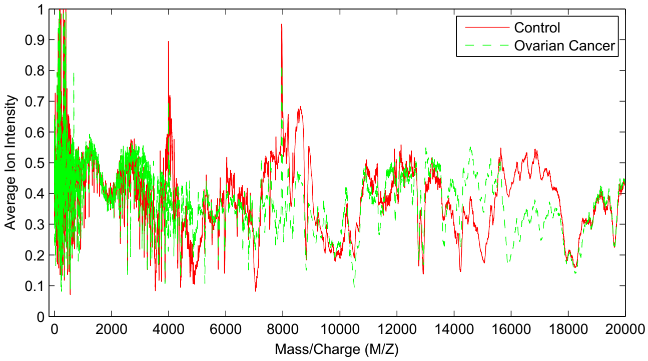

Fig. 3. Mass spectrographs (average across class) of the ovarian dataset.

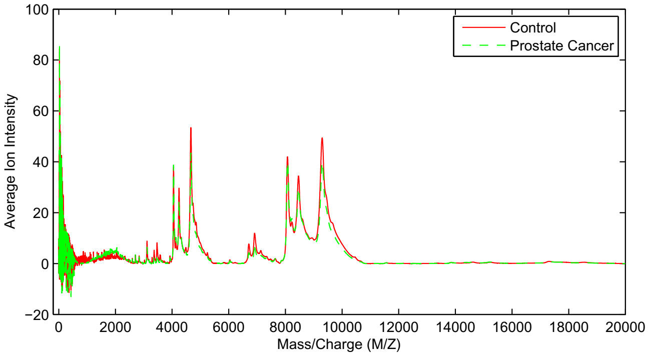

Fig. 4. Mass spectrographs (average across class) of the prostate dataset.

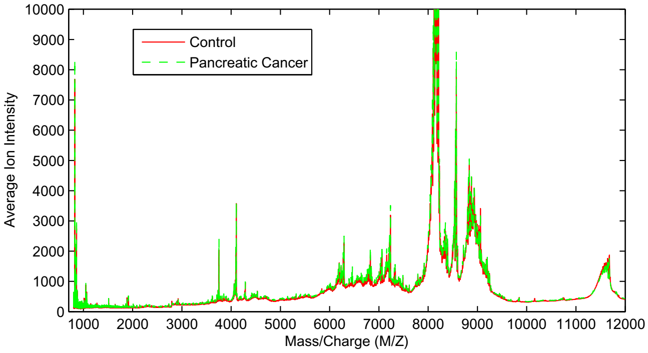

Fig. 5. Mass spectrographs (average across class) of the pancreatic dataset.

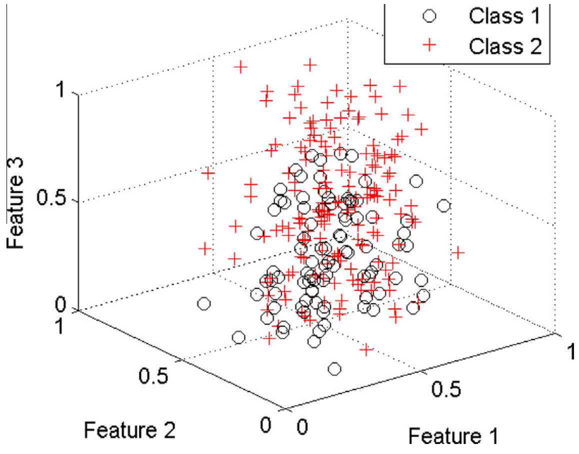

Fig. 6. Wavelet-GA features 3D projection in the ovarian dataset.

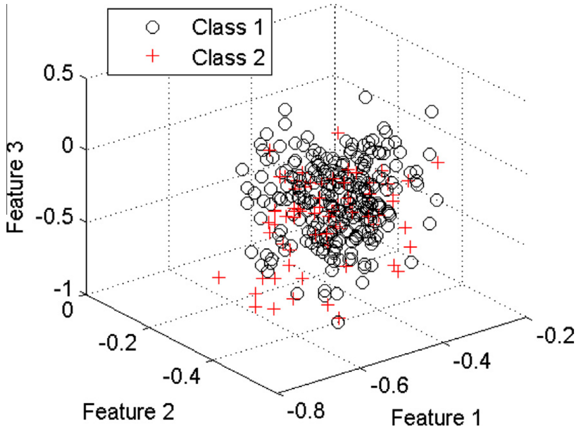

Fig. 7. Wavelet-GA features 3D projection in the prostate dataset.

As observed in the mass spectrographs, feature sets exhibited in Figs. 6–8 also present the increasing difficulty in differentiating normal and cancer individuals from the ovarian to prostate and pancreatic datasets.

3.3. Results and discussions

Results in terms of average accuracy across 30 running times are reported in Tables 1–3 for the ovarian, prostate and pancreatic datasets respectively. It is seen that wavelet-GA method (denoted as W-GA) significantly dominates all competitive feature extraction methods in every classifier in all 3 datasets.

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

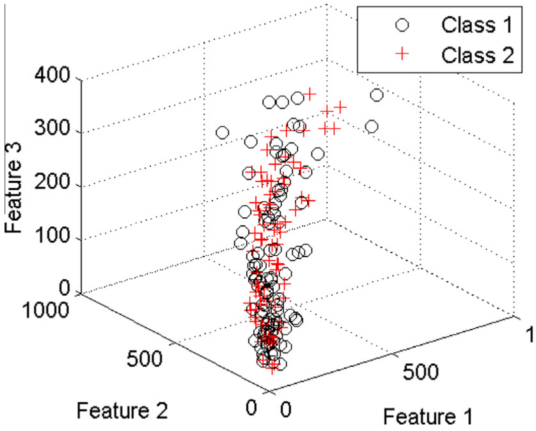

Fig. 8. Wavelet-GA features 3D projection in the pancreatic dataset.

For example, in the ovarian dataset, wavelet-GA feature set leads to the greatest accuracy at 100% by LDA, NB, kNN and MLP whilst none of the other feature extraction methods can give the average accuracy at 100%. For other classifiers such as SVM, fuzzy ARTMAP, ANFIS and AdaBoost, wavelet-GA also yields the maximum accuracy compared to other methods.

In the prostate dataset, wavelet-GA also outperforms other algorithms by obtaining the average accuracy approximately at 90%. Entropy is the worst method as it leads the lowest accuracy among investigated feature extraction methods.

Likewise, wavelet-GA is remarkably superior to other methods in the pancreatic dataset. The largest gap is between the wavelet-GA and PCA with 17.46% difference on average across classifiers. The second best is the sequential search method as it is inferior to wavelet-GA at 10.03% on average.

P-values of the Kruskal–Wallis test are reported in brackets adjacent to the average accuracy in Tables 1–3. For example, the value at 0.0000 in brackets in the cell of row LDA and column T-Test in Table 1 is the *P*-value of the Kruskal–Wallis test comparing two sets of accuracy results: one set is generated by the LDA classifier on the t-test features and another set is obtained by LDA on the wavelet-GA features.

We find that the *P*-values of these pairwise tests are all smaller than 0.05 (the 5% significance level). Therefore, the Kruskal–Wallis tests reject the null hypothesis that results of two feature extraction methods (wavelet-GA and each of the competitive methods) come from the same distribution at the 5% significance level. This shows the statistically significant dominance and robustness of wavelet-GA against other methods.

Figs. 9–11 graphically detail the performance comparisons among feature extraction methods by the MLP classifier in the ovarian, prostate and pancreatic datasets respectively. Each box in these box plots shows the median and distribution of the set of 30 accuracy outcomes. In line with average results presented in Tables 1–3, the box plots demonstrate the considerable

Table 1
Results of the ovarian dataset.

Classifiers	*T*-Test	Entropy	BD	Wilcoxon	ROC	PCA	Sequential	W-GA
LDA	98.08(0.0000)	97.77(0.0000)	97.30(0.0000)	98.55(0.0000)	97.95(0.0000)	98.75(0.0000)	99.74(0.0402)	100.00
NB	97.81(0.0000)	96.00(0.0000)	97.10(0.0000)	97.36(0.0000)	97.11(0.0000)	98.35(0.0000)	99.08(0.0001)	100.00
kNN	98.29(0.0000)	96.24(0.0000)	96.84(0.0000)	98.22(0.0000)	97.43(0.0000)	98.10(0.0000)	99.08(0.0003)	100.00
SVM	97.56(0.0000)	96.45(0.0000)	96.76(0.0000)	97.83(0.0000)	97.37(0.0000)	92.29(0.0000)	99.28(0.0014)	99.93
MLP	96.96(0.0000)	96.65(0.0000)	96.97(0.0000)	97.89(0.0000)	96.26(0.0000)	97.82(0.0000)	99.21(0.0006)	100.00
Fuzzy ARTMAP	96.57(0.0000)	96.04(0.0000)	95.65(0.0000)	94.74(0.0000)	94.99(0.0000)	96.89(0.0000)	95.30(0.0000)	99.93
ANFIS	97.83(0.0000)	96.18(0.0000)	95.79(0.0000)	97.76(0.0000)	96.06(0.0000)	97.31(0.0000)	99.02(0.0006)	99.93
AdaBoost	96.85(0.0000)	96.18(0.0000)	95.66(0.0000)	97.11(0.0000)	95.78(0.0000)	98.88(0.0207)	98.55(0.0041)	99.67

Table 2
Results of the prostate dataset.

Classifiers	*T*-Test	Entropy	BD	Wilcoxon	ROC	PCA	Sequential	W-GA
LDA	83.50(0.0000)	52.48(0.0000)	88.98(0.0371)	83.45(0.0000)	83.97(0.0000)	82.54(0.0000)	88.35(0.0028)	91.12
NB	86.42(0.0000)	49.09(0.0000)	85.52(0.0000)	85.77(0.0000)	86.57(0.0000)	83.84(0.0000)	88.61(0.0002)	91.43
kNN	84.72(0.0000)	75.57(0.0000)	87.58(0.0001)	84.67(0.0000)	85.44(0.0000)	88.43(0.0006)	89.26(0.0345)	91.48
SVM	81.62(0.0000)	52.16(0.0000)	88.94(0.0235)	81.46(0.0000)	81.69(0.0000)	84.09(0.0000)	84.95(0.0000)	91.23
MLP	86.02(0.0000)	77.49(0.0000)	88.68(0.0003)	87.71(0.0000)	86.55(0.0000)	86.89(0.0001)	88.77(0.0016)	91.98
Fuzzy ARTMAP	79.57(0.0000)	59.68(0.0000)	85.28(0.0000)	80.50(0.0000)	80.22(0.0000)	84.79(0.0000)	79.50(0.0000)	90.40
ANFIS	85.95(0.0000)	78.90(0.0000)	87.85(0.0000)	87.18(0.0002)	86.54(0.0001)	88.68(0.0125)	88.39(0.0052)	91.61
AdaBoost	85.83(0.0000)	74.74(0.0000)	87.51(0.0195)	86.39(0.0010)	85.63(0.0002)	87.03(0.0101)	87.75(0.0393)	89.73

Table 3
Results of the pancreatic dataset.

Classifiers	*T*-Test	Entropy	BD	Wilcoxon	ROC	PCA	Sequential	W-GA
LDA	57.03(0.0000)	55.95(0.0000)	53.25(0.0000)	54.39(0.0000)	51.17(0.0000)	52.63(0.0000)	60.91(0.0000)	74.78
NB	59.32(0.0000)	53.19(0.0000)	58.58(0.0000)	60.10(0.0000)	57.44(0.0000)	54.74(0.0000)	59.04(0.0000)	69.01
kNN	47.61(0.0000)	53.48(0.0000)	52.07(0.0000)	57.06(0.0001)	54.89(0.0000)	47.44(0.0000)	56.49(0.0000)	66.51
SVM	50.45(0.0000)	51.22(0.0000)	50.36(0.0000)	58.38(0.0000)	58.34(0.0000)	47.92(0.0000)	60.86(0.0053)	66.99
MLP	53.16(0.0000)	51.67(0.0000)	53.95(0.0000)	55.85(0.0003)	56.29(0.0004)	51.24(0.0000)	56.92(0.0001)	65.11
Fuzzy ARTMAP	51.60(0.0000)	50.35(0.0000)	52.29(0.0000)	59.27(0.0001)	58.61(0.0000)	46.68(0.0000)	51.13(0.0000)	67.89
ANFIS	51.41(0.0000)	52.49(0.0000)	52.60(0.0000)	51.56(0.0000)	58.96(0.0003)	53.30(0.0000)	58.41(0.0003)	66.74
AdaBoost	52.94(0.0000)	50.55(0.0001)	52.54(0.0000)	57.29(0.0000)	57.85(0.0000)	52.87(0.0000)	62.50(0.0007)	69.46

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

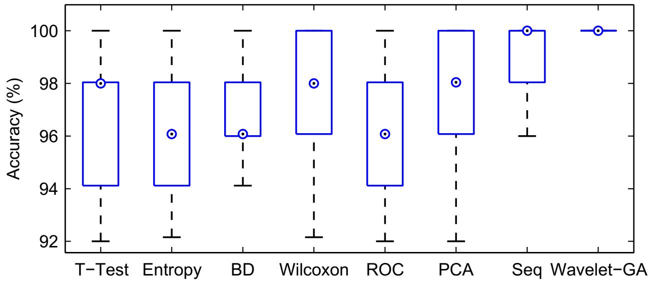

Fig. 9. Box plot of results of MLP in the ovarian dataset.

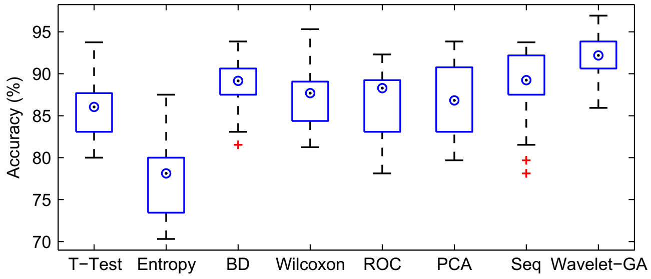

Fig. 10. Box plot of results of MLP in the prostate dataset.

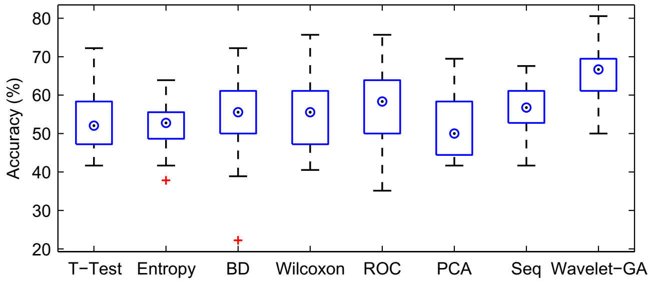

Fig. 11. Box plot of results of MLP in the pancreatic dataset.

superiority of wavelet-GA against other methods. The relatively small interquartile ranges of the wavlet-GA boxes furthermore indicate the stability of wavelet-GA compared to competitive approaches. This is particularly evident in Fig. 9 of the ovarian dataset as the wavelet-GA box is a single line. This is because wavelet-GA leads to the maximum accuracy at 100% in all 30 running times under the cross-validation strategy.

Results obtained by F-measure, AUC and MI also show the similar outcomes with those achieved by using the accuracy metric. It means that wavelet-GA outperforms other methods through all investigated performance metrics. As for demonstrations, the graphical comparisons among feature extraction methods in terms of F-measure, AUC and MI are displayed in Figs. 12–14 respectively for the ovarian, prostate and pancreatic datasets. Note that kNN is used for experiments in these figures.

Clearly, wavelet-GA is ranked top among examined methods. Entropy produces the lowest performance in both ovarian and prostate datasets. Sequential method consistently leads to the second best after wavelet-GA in all three datasets.

Table 4 reports the processing time consumed by feature extraction methods. The experiments in this study are carried out on a computer that has the Intel(R) Core(TM) i7-2600K CPU @ 3.40 GHz and 3.70 GHz with RAM at 16.0 GB running on the 64-bit Windows 7 Operating System. PCA is the fastest approach as it requires less than a second to process each dataset. It is worth noting that the decomposition algorithm that returns only the top 5

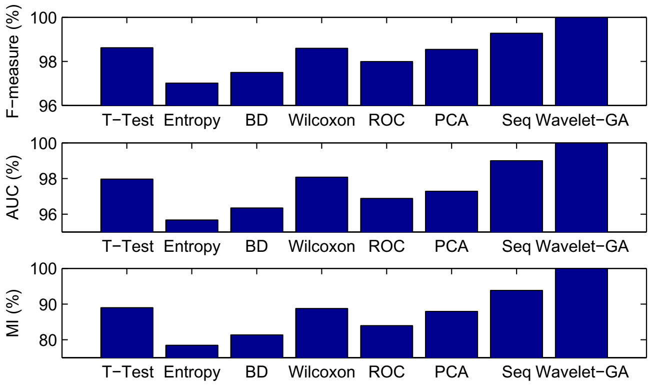

Fig. 12. F-measure, AUC and MI performance of kNN in the ovarian dataset.

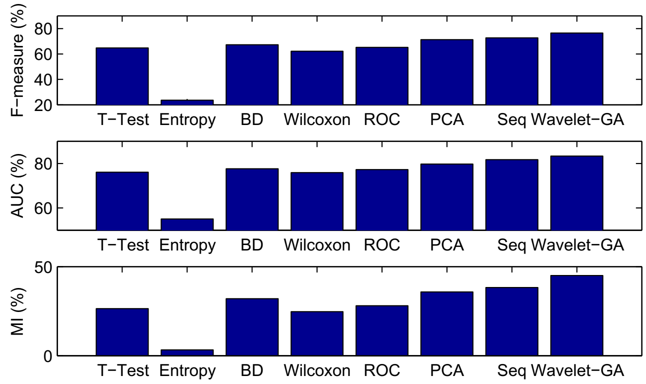

Fig. 13. F-measure, AUC and MI performance of kNN in the prostate dataset.

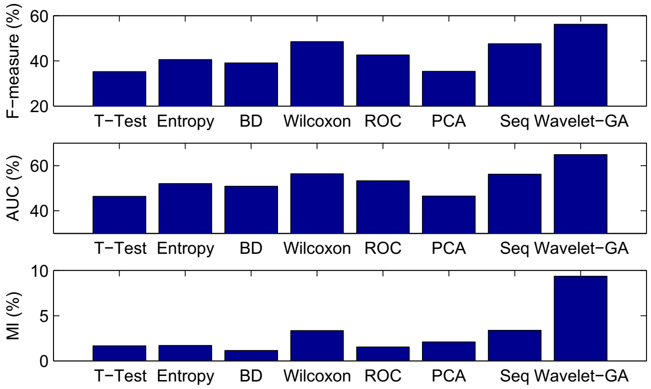

Fig. 14. F-measure, AUC and MI performance of kNN in the pancreatic dataset.

Table 4
Processing time (in seconds) of feature extraction methods.

Methods	Ovarian	Prostate	Pancreatic
T-Test	2.530	2.750	0.999
Entropy	2.548	2.761	1.001
BD	2.525	2.749	0.998
Wilcoxon	6.790	6.362	3.092
ROC	2.678	2.920	1.046
PCA	0.114	0.126	0.043
Sequential	438.199	453.923	173.311
Wavelet-GA	15.125	16.539	13.007

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

eigenvectors has been applied for PCA. Filter procedures such as t-test, entropy, BD, Wilcoxon, and ROC are also fast methods, which are comparable to PCA. Operating as a wrapper method, sequential search takes the largest amount of time amongst investigated methods. It needs more than 7 min for the ovarian and prostate datasets and nearly 3 min for the pancreatic dataset. Compared to the sequential search, wavelet-GA takes much less time amount. It needs approximately 15 s to complete each of the three datasets. Wavelet-GA spends larger time amount compared to *t*-test, entropy, BD, Wilcoxon, ROC and PCA but those expenses are worthy as wavelet-GA significantly improves the MS data classification performance. As a hybrid method combining both filter and wrapper methods, wavelet-GA obviously demonstrates the double advantages, i.e. greater classification accuracy and lower computational costs, against the comparable wrapper sequential search method.

4. Conclusions

A hybrid approach to feature extraction for MS cancer data classification is proposed in this paper. Using wavelet features, GA is implemented as a combination between filter and wrapper methods where the two-sample *t*-test is incorporated during the GA initialization process. Wavelet-GA has advantages that drastically enhance the classification accuracy of various classifiers with inexpensive computational costs.

Through different performance metrics, i.e. accuracy, F-measure, AUC and MI, wavelet-GA demonstrates a significant dominance against competitive feature extraction methods including *t*-test, entropy, BD, Wilcoxon, ROC, PCA and sequential search. Results of the statistical Kruskal–Wallis test show the robustness and stability of the feature set obtained by wavelet-GA when compared to those of competitive methods. The cross-validation strategy implemented on three benchmark MS datasets makes the conclusions driven out of this study valid and general. Wavelet-GA thus can be applied to real classification models as decision support systems for cancer diagnosis and prognosis, which greatly benefit clinicians and medical practitioners.

Cancer is one of leading causes of death worldwide. Therefore the accurate cancer diagnosis is critically demanded in medical practice. This research has shown a great improvement regarding cancer classification accuracy using mass spectrometry data, which is useful for early detection, rapid intervention and treatment of cancers in an effective and efficient manner. The proposed approach thus would contribute to improving the public health by increasing human longevity, reducing mortality rate in the communities and ensuring people live healthier and more independent lives.

Future research would investigate more efficient classifiers rather than prevalent methods examined in this study. Numerous classifiers have been proposed in the literature, they are worth an extensive investigation as they may generate better results. The proposed wavelet-GA feature extraction for MS data has been successfully testified with various classifiers, it therefore has a potential to synergize with any other methods to yield great cancer classification performance.

Acknowledgments

This research is supported by the Australian Research Council (Discovery Grant DP120102112) and the Centre for Intelligent Systems Research (CISR) at Deakin University.

References

[1] Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422 (6928), 198–207.

[2] Cravatt, B.F., Simon, G.M. and Yates, J.R. (2007) Iii. The biological impact of mass-spectrometry-based proteomics. Nature 450 (7172), 991–1000.

[3] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K. and Zhao, H. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 19 (13), 1636–1643.

[4] Lilien, R.H., Farid, H. and Donald, B.R. (2003) Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. J. Computat. Biol. 10 (6), 925–946.

[5] Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A. and Le, Q.T. (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts'. Bioinformatics 20 (17), 3034–3044.

[6] Zhang, X., Lu, X., Shi, Q., Xu, X.Q., Hon-chiu, E.L., Harris, L.N. and Wong, W.H. (2006) Recursive svm feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinf. 7 (1), 197.

[7] Deng, X., Geng, H., Bastola, D.R. and Ali, H.H. (2006) Link test-a statistical method for finding prostate cancer biomarkers. Computat. Biol. Chem. 30 (6), 425–433.

[8] Ge, G. and Wong, G.W. (2008) Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. BMC Bioinf. 9 (1), 275.

[9] Kumar, C. and Mann, M. (2009) Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS Lett. 583 (11), 1703–1712.

[10] Garcia-Torres, M., Armananzas, R., Bielza, C. and Larraaga, P. (2013) Comparison of metaheuristic strategies for peakbin selection in proteomic mass spectrometry data. Inf. Sci. 222, 229–246.

[11] Kong, A., Gupta, C., Ferrari, M., Agostini, M., Bedin, C., Bouamrani, A. and Azencott, R. (2014) Biomarker signature discovery from mass spectrometry data. IEEE/ACM Trans. Computat. Biol. Bioinf. 11 (4), 766–772.

[12] Hsu, H.H., Hsieh, C.W. and Lu, M.D. (2011) Hybrid feature selection by combining filters and wrappers. Expert Syst. Appl. 38 (7), 8144–8150.

[13] DeVore, R.A. and Lucier, B.J. (1992) Wavelets. Acta Numer. 1 (1), 1–56.

[14] Du, P., Kibbe, W.A. and Lin, S.M. (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics 22 (17), 2059–2065.

[15] Peng, S., Xu, Q., Ling, X.B., Peng, X., Du, W. and Chen, L. (2003) Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Lett. 555 (2), 358–362.

[16] Huang, C., Yang, X. and He, Z. (2010) Protein folding simulations of 2d hp model by the genetic algorithm based on optimal secondary structures. Computat. Biol. Chem. 34 (3), 137–142.

[17] Vannucci, M., Sha, N. and Brown, P.J. (2005) Nir and mass spectra classification: Bayesian methods for wavelet-based feature selection. Chemometr. Intell. Lab. Syst. 77 (1), 139–148.

[18] Krzanowski, W.J. (1988) Principles of Multivariate Analysis: A User's Perspective, Oxford University Press, New York.

[19] MathWorks (2014) Genetic algorithm options. Retrieved 10 Oct 2014 from <http://mathworks.com/help/gads/genetic-algorithm-options.html>.

[20] Mitchell, T. (1997) Machine Learning, McGraw Hill.

[21] Kecman, V. (2001) Learning and Soft Computing, MIT Press, Cambridge, MA.

[22] Bishop, C. (1995) Neural Networks for Pattern Recognition, Oxford University Press, Oxford.

[23] Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H. and Rosen, D.B. (1992) Fuzzy artmap: a neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Trans. Neural Netw. 3 (5), 698–713.

[24] Jang, J.S. (1993) Anfis: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybernet. 23 (3), 665–685.

[25] Freund, Y. and Schapire, R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139.

[26] Carpenter, G.A. (2003) Default artmap. Proc. Int. Joint Conf. Neural Netw. 2, 1396–1401.

[27] Theodoridis, S. and Koutroumbas, K. (2009) Pattern Recognition, 4th ed, Academic Press.

[28] Choi, E. and Lee, C. (2003) Feature extraction based on the Bhattacharyya distance. Pattern Recognition 36 (8), 1703–1709.

[29] Deng, L., Pei, J., Ma, J. and Lee, D.L. (2004) A rank sum test method for informative gene discovery. Proc. Tenth ACM SIGKDD Int. Conf. Knowledge Discov. Data Mining 1, 410–419.

[30] Jolliffe, I.T. (2005) Principal Component Analysis: Encyclopedia of Statistics in Behavioral Science, John Wiley and Sons Ltd.

[31] Kohavi, R. and John, G. (1997) Wrappers for feature subset selection. Artificial Intelligence 97 (1–2), 272–324.

[32] Kruskal, W.H. and Wallis, W.A. (1952) Use of ranks in one-criterion variance analysis. J. Am. Statist. Assoc. 47 (260), 583–621.

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