Artificial Intelligence in Medicine (2005) **35**, 215–226

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Identification of signatures in biomedical spectra using domain knowledge

Erinija Pranckeviciene [a,b], Ray Somorjai [a,*], Richard Baumgartner [a], Moon-Gu Jeon [a]

[a] *Institute for Biodiagnostics, National Research Council, 435 Ellice Avenue, Winnipeg, Man., Canada R3B 1Y6*
[b] *Kaunas University of Technology, Studentu 50, Kaunas, LT 3031, Lithuania*

Received 13 May 2004; received in revised form 30 November 2004; accepted 6 December 2004

KEYWORDS
Classification of biomedical spectra;
Dimensionality reduction;
Feature selection;
Genetic algorithm;
L_1-norm SVM;
Spectral signature;
Consensus feature sets;
Domain knowledge

Summary

Objective: Demonstrate that incorporating domain knowledge into feature selection methods helps identify interpretable features with predictive capability comparable to a state-of-the-art classifier.
Methods: Two feature selection methods, one using a genetic algorithm (GA) the other a L_1-norm support vector machine (SVM), were investigated on three real-world biomedical magnetic resonance (MR) spectral datasets of increasing difficulty. Consensus sets of the feature sets obtained by the two methods were also assessed.
Results and conclusions: Features identified *independently* by the two methods and by their consensus, determine class-discriminatory groups or individual features, whose predictive power compares favorably with that of a state-of-the-art classifier. Furthermore, the identified feature signatures form *stable groupings* at *definite spectral positions*, hence are readily interpretable. This is a useful and important practical result for generating hypothesis for the domain expert.

1. Introduction

Growing interest in discovering biologically meaningful information in biomedical data led to the use and development of many techniques, some also applicable to biomedical spectra. A useful summary is in [1]. Among these are: principal component analysis, *k*-means and hierarchical clustering, support vector machines (SVMs), hidden Markov models, genetic algorithms, neural network techniques, self-organizing maps, classification and regression trees. Using data mining techniques on microarrays/spectra, one attempts to identify important clues for class separation and to use this information to design a classification rule. Standard methods of dimensionality reduction [2–4] and simple distance-based classifiers applied to the original high-dimensional data are not very effective. For high-dimensional

* Corresponding author. Tel.: +1 204 984 4538; fax: +1 204 984 5472.
E-mail address: ray.somorjai@nrc-cnrc.gc.ca (R. Somorjai).

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2004.12.002

data, feature selection typically precedes classification [5–9]. For microarray data and biomedical spectra, feature selection is necessary but often not sufficient when additional information about the classes is lacking. Features that are optimal for classification do not necessarily possess biological relevance. For high-dimensional but sparse datasets, many different combinations of attributes may separate the data perfectly [10]. Which of these is plausible? Are the discovered features truly characteristic of the classes as labelled (e.g., cancer versus normal), or do they also reflect other covariates (e.g., gender, age, etc.), or even noise? Sometimes the data labelling by the domain expert contains wrong class assignments, confusing the feature selection/classification process and outcomes. Incorporating domain knowledge helps deal more efficiently with the problem of uncertain multiple solutions, and aids in identifying the most appropriate data analysis model.

Our goal is to discuss a feature selection strategy that is determined by the domain knowledge available for typical biomedical spectra. Domain knowledge (DK) is additional information about spectra that distinguishes them from other types of data, such as microarrays. Additional DK enables designing a feature selection algorithm that reflects directly the nature/characteristics of the data. Consider a spectrum as a collection of peaks and valleys, whose positions and intensities carry discriminatory information relevant for classification. The physical/chemical basis of class separation is reflected in the peak/valley distribution and peak width. Typically, spectral data have high feature-space dimensionality, although the number of discriminatory features (intrinsic dimensionality) may possibly be quite low. This happens because of the many correlated features in a spectrum; thus, it is likely that if a single attribute is discriminatory, so are its immediate neighbors. Guaranteeing that the new features correspond directly to the original positions in the spectra is very important for interpretability. We concretize the concept of *spectral signature* as a set of *related* spectral regions or single spectral attributes. It is assumed that the samples of a particular spectral class have common specific spectral signatures. Generally, we do not know the number, position or width of the spectral regions and/or relative intensity levels of the spectral signature that separate classes. The discovery of the class *signature*—the *discriminatory* pattern common to all samples of a particular class—is the goal of the feature selection step.

To discover the signature(s), we combine the outputs of two methods. Both capture and retain the original spectral features. One is a genetic-algorithm-based feature selection method, wrapped around a simple classifier (e.g., linear discriminant analysis (LDA)). The other selects spectral attributes through the use of the sparseness property of an L_1-norm SVM classifier. This technique is referred to by several names: sparse classifier [5], 1-norm SVM classifier [11], SVM trained by linear programming [12,13], linear sparse kernel Fisher discriminant [14].

The proposed feature selection methodology narrows the range of useful spectral features for further processing, considering only those signatures identified by both methods. We demonstrate the approach on three real-life datasets. To avoid over-optimistic assessment of the feature selection methods because of selection bias [15], we partition the data into a training and an independent test sets. The test set was used only once at the very end, after the feature selection was completed.

2. Data

Three real-world, two-class datasets were used in this feature selection study. Dataset1 contains MR spectra of pathogenic fungi (*Candida albicans* versus *Candida tropicalis*) [16]. Dataset2 comprises MR spectra of biofluids obtained from normal subjects and cancer patients [17–18]. Dataset3 consists of MR spectra of biofluids obtained from patients with successful renal transplant versus patients with (rejected) kidney transplant [19].

The characteristics of the datasets are given in Table 1: D is the dimensionality of the data, N_1, N_2 are the total number of samples in classes 1 and 2, respectively. $Tr_1 + Tr_2$ are the number of samples of classes 1 and 2 in the training set, and $Te_1 + Te_2$ in the test set, respectively. The partition of the samples in the training and test sets remains the same in all experiments. For the feature selection process, the training samples $Tr_1 + Tr_2$ are further divided into *training* and *monitoring* sets. A *single* validation

Table 1 Properties of the datasets

Name	Dimensionality, D	N_1	N_2	$Tr_1 + Tr_2$	$Te_1 + Te_2$	E_n
Dataset1	1500	104	75	50 + 50	54 + 25	6
Dataset2	300	61	79	31 + 40	30 + 39	12
Dataset3	3380	91	65	45 + 33	46 + 32	49