Knowledge-Based Systems 37 (2013) 502–514

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data

Yihui Liu [a,b,*], Uwe Aickelin [b], Jan Feyereisl [b], Lindy G. Durrant [c]

[a] Institute of Intelligent Information Processing, School of Information Science, Shandong Polytechnic University, China
[b] School of Computer Science, University of Nottingham, UK
[c] Academic Department of Clinical Oncology, Institute of Immunology, Infections and Immunity, City Hospital, University of Nottingham, UK

ARTICLE INFO

Article history:
Received 19 December 2011
Received in revised form 14 September 2012
Accepted 30 September 2012
Available online 12 October 2012

Keywords:
Biomarkers
Wavelet feature extraction
CD46
Colorectal cancer
Genetic algorithm

ABSTRACT

Biomarkers which predict patient's survival play an important role in medical diagnosis and treatment. How to select the significant biomarkers from hundreds of protein markers is a key step in survival analysis. In this paper a novel method is proposed to detect the prognostic biomarkers of survival in colorectal cancer patients using wavelet analysis, genetic algorithm, and Bayes classifier. One dimensional discrete wavelet transform (DWT) is normally used to reduce the dimensionality of biomedical data. In this study one dimensional continuous wavelet transform (CWT) was proposed to extract the features of colorectal cancer data. One dimensional CWT has no ability to reduce dimensionality of data, but captures the missing features of DWT, and is complementary part of DWT. Genetic algorithm was performed on extracted wavelet coefficients to select the optimized features, using Bayes classifier to build its fitness function. The corresponding protein markers were located based on the position of optimized features. Kaplan–Meier curve and Cox regression model were used to evaluate the performance of selected biomarkers. Experiments were conducted on colorectal cancer dataset and several significant biomarkers were detected. A new protein biomarker CD46 was found to be significantly associated with survival time.

1. Introduction

Survival analysis involves the estimation of the distribution of time it takes for death to occur depending on the biology of the disease. It allows clinicians to plan a suitable treatment and counsel patients about their prognosis. In medical domains, survival analysis is mainly based on Kaplan–Meier (KM) estimator and Cox proportional hazards regression model [1,2], which are used to evaluate the performance of prognostic markers. However how to rank these biomarkers, is a key step in survival analysis. Normally, the selection of biomarkers is based on medical knowledge and the diagnosis of the clinician [1,2]. This may ignore potential biomarkers. Machine learning algorithms have been widely used in biomarker analysis of high dimensional medical data, such as microarray data [3–5] or mass spectrometry data [6,7]. Despite the potential advantages over standard statistical methods, their applications to survival analysis are rare due to the difficulty in dealing with censored data [8]. Recent research has shown that machine learning methods, such as neural network [9,10], Bayesian network [11], decision tree and Naïve Bayes classifier [8], are used to improve the survival model. However, none of these methods deals with the biomarker selection in survival analysis.

In this study we propose a novel method of biomarker selection based on one dimensional continuous wavelet transform (CWT). Normally, one dimensional discrete wavelet transform (DWT) is used to reduce dimensionality in the analysis of high dimensional biomedical data [12,13]. In biomarker detection, the feature space must have the corresponding relationship with original data space to locate the detected biomarker based on detected features. One dimensional CWT detects the feature of data at every scale and position, and keeps local property of the original data. Wavelet feature vector of CWT has the same length as the original data, and can be used to locate the biomarker in original data space.

First we perform one dimensional continuous wavelet transform at different scales on colorectal cancer data to extract the discriminant features. Then we use genetic algorithm (GA) and Bayes classifier to select the optimized features from extracted wavelet coefficients. Due to the wavelet well-known property, which reveals the local features of data (or time feature) and does not lose the position information of original data, the corresponding protein markers in the original data space are obtained based on the position of optimized wavelet features. Finally Kaplan–Meier (KM) estimator and Cox regression model were used to evaluate the performance of selected protein markers. A new protein biomarker CD46 was found to have independent prognostic significance.

* Corresponding author at: Institute of Intelligent Information Processing Shandong Polytechnic University, Jinan 250013, China. Tel.: +86 (0) 53189631256.
E-mail addresses: yxl@spu.edu.cn, yihui_liu_2005@yahoo.co.uk (Y. Liu).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.09.011

Recent research suggests that "the immune system might be involved in the development and progression of colorectal cancer" [1,14]. The detection of CD46 supports their deduction or conclusions.

The rest of paper is organized as follows: In Section 2, we describe the colorectal cancer data. Our proposed method is introduced in Section 3. Wavelet feature extraction for colorectal cancer data is described in Section 4. In Section 5, genetic algorithm based on Bayes classifier is used to select the optimized features. Survival models are used to evaluate the selected biomarkers in Section 6. The experiments are conducted in Section 7, followed by discussion and concluding comments in Section 8.

2. Colorectal cancer data

We use the same dataset, which Professor Lindy Durrant used in her research. It is described in Lindy Durrant's research [1,2]. The study population cohort comprised a consecutive series of 462 archived specimens of primary invasive cases of colorectal cancer (CRC) tissue obtained from patients undergoing elective surgical resection of a histologically proven primary CRC at Nottingham University Hospitals, Nottingham, UK. The samples were collected between January 1994 and December 2000 from the established institutional tumor bank and were identified from the hospital archives. No cases were excluded unless the relevant clinicopathological material/data were unavailable. The mean follow-up period was 42 months (range 1–116) to ensure a sufficient duration of follow-up to allow meaningful assessment of the prognostic value of the markers examined. Follow-up was calculated from the date of resection of the primary tumor, and all surviving cases were censored for data analysis in December 2003. A tissue microarray of 462 colorectal tumors was stained by immunohistochemistry for markers which predict immunosurveillance/editing.

The data has 462 samples with 210 attributions and is 462×210 data matrix. We use a simple way to do the pre-processing of our data: First, we remove those 70 features for which most patients have missing values. Second, we remove those patients, which miss any of the remaining 140 attributions. After that, we obtain a (complete) 153×140 matrix. Eighteen patients died for other causes, not related to their colorectal cancer and they were excluded from the analysis. Among the remaining 135 patients, 76 patients were dead with survival time ranging from 0 to 65 months, and 59 patients were alive with survival time ranging from 38 to 111 months.

The aim of the research was to find the significant biomarkers in survival analysis. Two groups of patients were identified to perform the analyses. The first group includes the patients who died with survival time of less than 30 months, and the second group has the patients who were alive with survival time of more than 70 months. For the first group, there were 59 dead patients; for the second group, there were 31 alive patients. Among 140 attributions, only 115 of them are protein markers, others are the description of patients and medical diagnosis, such as age, survival time, TNM (Tumor, Node, Metastasis) stage and Duke stage. For this research, only protein markers were of interest in survival analysis.

Finally, there were 59×115 and 31×115 data matrix groups. Fig. 1 shows two groups of data used for biomarker selection. Because the value of protein markers has a different scale, pre-processing by normalizing each protein marker and then each sample vector was done.

3. The proposed method

Fig. 2 shows the selection process of significant biomarkers in survival analysis. First the data was transformed into wavelet space at different scales to find the most discriminant features between the two groups. Genetic algorithm was used to select the best features from extracted wavelet features and then the significant protein markers were detected based on the optimized features in wavelet space. Finally Kaplan–Meier curve and Cox regression model were performed to evaluate the performance of selected significant biomarkers.

Normally, we have feature extraction and feature selection methods for data analysis. Feature extraction is that the data is transformed into a new data space using a set of new basis, which reflects the hidden properties of data in original data space, such as principal component analysis (PCA), linear discriminant analysis (LDA), independent component analysis (ICA), and wavelet trans-

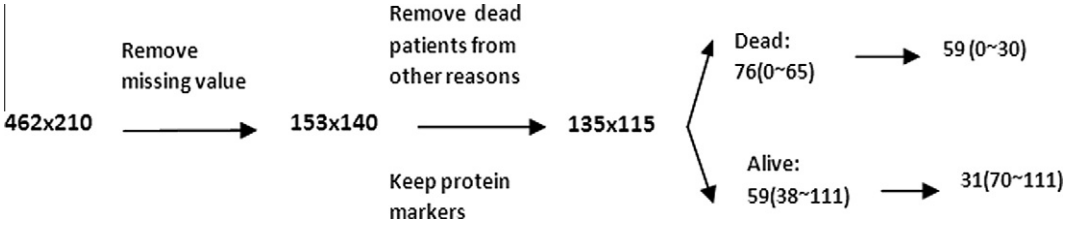

Fig. 1. Two groups of data used to select the significant protein markers. There are 59 dead patients with survival time of less than 30 months, and 31 alive patients with survival time of more than 70 months.

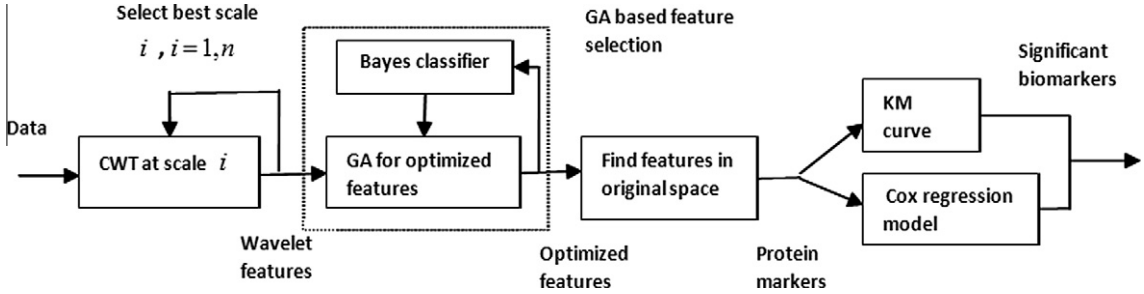

Fig. 2. The selection of significant biomarkers in survival analysis.