Computers in Biology and Medicine 91 (2017) 103–111

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Feature selection method based on support vector machine and shape analysis for high-throughput medical data

Qiong Liu [a,b], Qiong Gu [c], Zhao Wu [c,*]

[a] Medical College, Hubei University of Arts and Science, China
[b] XiangYang Central Hospital, China
[c] School of Mathematics and Computer Science, Hubei University of Arts and Science, China

ARTICLE INFO

Keywords:
High-throughput medical data
Feature selection
Support vector machine
Shape analysis

ABSTRACT

Proteomics data analysis based on the mass-spectrometry technique can provide a powerful tool for early diagnosis of tumors and other diseases. It can be used for exploring the features that reflect the difference between samples from high-throughput mass spectrometry data, which are important for the identification of tumor markers. Proteomics mass spectrometry data have the characteristics of too few samples, too many features and noise interference, which pose a great challenge to traditional machine learning methods. Traditional unsupervised dimensionality reduction methods do not utilize the label information effectively, so the subspaces they find may not be the most separable ones of the data. To overcome the shortcomings of traditional methods, in this paper, we present a novel feature selection method based on support vector machine (SVM) and shape analysis. In the process of feature selection, our method considers not only the interaction between features but also the relationship between features and class labels, which improves the classification performance. The experimental results obtained from four groups of proteomics data show that, compared with traditional unsupervised feature extraction methods (i.e., Principal Component Analysis - Procrustes Analysis, PCA-PA), our method not only ensures that fewer features are selected but also ensures a high recognition rate. In addition, compared with the two kinds of multivariate filter methods, i.e., Max-Relevance Min-Redundancy (MRMR) and Fast Correlation-Based Filter (FCBF), our method has a higher recognition rate.

1. Introduction

With the development of high-throughput detection technology, various high-throughput data, such as proteomics spectra data, gene expression data, and single nucleotide polymorphism data, have been gathered from life science fields. These data contain details for us to study diseases from multiple levels as well as the diversity among different populations of the same species. However, these data are high-dimensional data with small sample sizes, in which the number of features is much more important than the number of samples. Using traditional pattern classification methods to deal directly with these data leads to the curse of dimensionality. One of the effective methods to prevent the curse of dimensionality is to employ feature selection methods to remove the irrelevant features from the data before pattern classification. In this paper, how to select features, considering feature interaction, and how to define the measure that reveals multi-feature interaction are studied on high-throughput medical data.

At present, there is no effective way to address high-throughput medical data. In handling such data, traditional data analysis methods are often ineffective or even fail. Information or regular patterns hidden in the data cannot be explored and understood, which results in “data resources” becoming “data disasters”. Artificial intelligence technology is the most effective and core technology to address high-dimensional data. The machine learning algorithm has been shown to be an effective method of data analysis and processing. However, the development of clinical diagnostic technology is urgently needed to explore and reveal the mystery between the data. Therefore, how to effectively extract or select the useful feature information or recurring patterns from high-throughput medical data to assist clinical medical diagnosis has become the basic problem facing modern clinical medicine.

In recent years, proteomics mass spectrometry has become an important technique in the diagnosis of tumor markers, which is mainly due to the development of two types of mass spectrometry technology: Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass

* Corresponding author.
E-mail address: wuzhao73@163.com (Z. Wu).

https://doi.org/10.1016/j.compbiomed.2017.10.008
Received 3 March 2017; Received in revised form 24 September 2017; Accepted 8 October 2017

Spectrometry (MALDI-TOF-MS) and Surface-Enhanced Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry (SELDI-TOF-MS). At different stages of the disease, the patient's body tissue presents a different response. The occurrence of disease will inevitably lead to changes in the level of protein expression in the human body which, through the analysis of the relevant protein molecules, can be used as a basis for the study of the causes of the disease. Therefore, some of the protein differences between healthy individuals and cancer patients can serve as biomarkers for disease detection and screening. Proteomics techniques are utilized mainly for identifying the differences in blood and urine proteins. With the continuous development of technology, proteomics has become a new technology to study the cause of tumors and to achieve early tumor detection.

Using a small number of proteins as biomarkers, the conventional detection methods apply to the early diagnosis of various cancers. The sensitivity and specificity of their test results are poor. Unlike conventional methods, mass spectrometry technology can screen out biological molecules related to early diagnosis of cancer. As a result, it is used for quick diagnosis of disease, deep understanding of the causes of the disease, and the discovery of markers associated with the disease. The SELDI-TOF-MS technique can quantitatively analyze the protein molecules in each tissue of the living body and obtain the relative concentration of the differentially expressed proteins associated with the disease and other parameters. Because the principle and the structure of the SELDI-TOF-MS instrument is simple, its analysis time is short. It has unique advantages in the detection of low abundance and low molecular weight proteins. It has become a commonly used technique in proteomics.

However, the amount of data from the application of SELDI-TOF-MS is quite large. The massive data contain not only the information we want to mine that is closely related to the disease but also detection errors and signal noise. Especially in its application in clinical examination, doctors or experts need to classify the disease based on the massive data, which not only increases the cost of clinical examination but also requires a higher professional background for doctors; otherwise, measurement error may occur, eventually leading to false or undetected. Therefore, it is a tough task to rely solely on manual analysis of SELDI-TOF-MS data to achieve cancer detection. How to optimize and extract the characteristic parameters related to disease in high resolution and massive SELDI-TOF-MS data, and how to establish the mapping model between the characteristic parameters and the degree of illness, is the fundamental problem of using SELDI-TOF-MS technology to detect early cancer.

In recent years, analysis of SELDI-TOF-MS data has mainly focused on two aspects. First is the selection of characteristic parameters of high-dimensional mass spectrometry data. It is similar to the problem of attribute selection in data mining research, that is, how to select information related closely to disease from thousands of protein peaks. Second, the classification and prediction model is established based on the characteristic parameters. It classifies and determines the type of disease, according to the difference of the characteristic parameters. Based on the results of the model, doctors determine the type and period of the tumor and ultimately give the appropriate treatment advice. The methods of mass spectrometry data analysis are the iterative search method [1], genetic algorithm [2], and chi-squared test [3]. Although the accuracy of these methods is high, the specificity is low. Avila C.C. et al. proposed the use of a pattern-matching algorithm for cancer mass spectrometry data analysis [4]. Gu H.W. et al. applied the classical data analysis model of Partial Least Squares (PLS) to the analysis of cancer mass spectrometry data [5]. However, as a traditional chemometrics analysis method, PLS has its limitations. With the increase of the data categories, the nonlinear factors between the mass spectrum data and the sample category will become larger, which will affect the predicted effect of PLS, thus resulting in larger detection error. In addition, PLS is affected by random noise. With the development of pattern recognition technology and biological information-processing technology, Rocha W.F.D. et al. [6] proposed a mass spectrometry data classification method based on artificial neural networks. Although the recognition rate of the artificial neural network model is high, its ability to predict the unknown sample is limited. Its goal of learning is to minimize training errors, which leads to the low generalization ability of the model, and the accuracy of the model prediction is reduced. To improve the deficiency of the detection model based on artificial neural networks, Harrington P.D [7]. proposed a discriminant analysis model based on SVM. The predicted rate of this model is obviously higher than that of an artificial neural network model. SVM has proven able to obtain better prediction results than the traditional neural network model [8]. Lokhov P.G. et al. proposed the combination of Principal Component Analysis (PCA) and SVM, and the experimental results were better [9,10]. As a common method of data dimensionality reduction and feature extraction, PCA can reduce the computational complexity and the complexity of the problem. To solve the disadvantages of the traditional artificial neural network, the learning goal of SVM is to minimize the structural risk. At the same time, the empirical risk minimization is also used for measuring the risk of the model. The experimental results show that the structural risk minimization learning algorithm improves the generalization ability of the model to a great extent. After comparing and analyzing the detection results of various models, a new feature selection method based on SVM and Procrustes Analysis (PA) is proposed to solve the problem of feature extraction for high-throughput medical data.

The structure of this paper is organized as follows. In Section 2, after analyzing the main problems of feature selection in pattern classification for high-throughput medical data, the shortcomings of some existing methods are noted. Section 3 first describes the basis of our proposed method (including the data preprocessing method, dimensionality reduction method, and an SVM feature analysis method based on feature subset importance based on PA), then describes our feature selection method in detail and gives an algorithm to implement it and the time complexity analysis. Section 4 first introduces the experimental design and experimental datasets, then proposes the method to determine the parameter values in our algorithm. Finally, a comparative analysis between our algorithm and the existing algorithms is given. Section 5 summarizes the superiority of our method and gives the next steps to study the main issues.

2. Problem statement

High-throughput medical data provide a wealth of information for the study of pathogenesis and disease clinical screening. However, due to the high cost of experiments, high-throughput data usually contain many samples that are far less than the number of features that make up the samples in the data. Moreover, these data are usually noisy, and some observations are missing or uncertain. These characteristics of high-throughput data pose a serious challenge to the traditional machine-learning methods [11]. This is mainly manifested in the following two aspects: on the one hand, the number of samples is too small to allow for accurate estimation via the probability distribution of samples, with the result that probability-based machine learning methods cannot be used. On the other hand, the number of features is too large and the parameters of the model-based machine learning methods are too numerous, resulting in greatly increased running time. Plenty of scientific research has indicated that directly using traditional machine learning methods to classify and cluster high-throughput data is often time-consuming, but the result of classification or clustering is very poor, that is, the curse of dimensionality [12].

At present, the basic way to avoid the curse of dimensionality is to reduce the number of samples that represent the characteristics of the sample before pattern classification or clustering [13]. There are two common methods: feature extraction and feature selection. Among them, feature extraction transforms the data into another space, which can reveal the essential characteristics of the data through a certain transformation, and selects some representative features to represent the data [14]. Feature selection selects a part of the original data to meet the