A Hybrid Classification System for Cancer Diagnosis with Proteomic Bio-markers

Jung-Ja Kim[1], Young-Ho Kim[2], and Yonggwan Won[1,*]

[1] Department of Electronics and Computer Engineering, Chonnam National University, 300 Yongbong-Dong Buk-Gu Kwangju, Republic of Korea
{j2kim, ykwon}@chonnam.ac.kr
[2] Doul Info. Technology, 1412-8 Yongbong-Dong Buk-Gu, Gwang-ju, Republic of Korea
melchi@grace.chonnam.ac.kr

Abstract. A number of studies have been performed with the objective of applying various artificial intelligence techniques to the prediction and classification of cancer specific biomarkers for use in clinical diagnosis. Most biological data, such as that obtained from SELDI-TOF (Surface Enhanced Laser Desorption and Ionization-Time Of Flight) MS (Mass Spectrometry) is high dimensional, and therefore requires dimension reduction in order to limit the computational complexity and cost. The DT (Decision Tree) is an algorithm which allows for the fast classification and effective dimension reduction of high dimensional data. However, it does not guarantee the reliability of the features selected by the process of dimension reduction. Another approach is the MLP (Multi-Layer Perceptron) which is often more accurate at classifying data, but is not suitable for the processing of high dimensional data. In this paper, we propose on a novel approach, which is able to accurately classify prostate cancer SELDI data into normal and abnormal classes and to identify the potential biomarkers. In this approach, we first select those features that have excellent discrimination power by using the DT. These selected features constitute the potential biomarkers. Next, we classify the selected features into normal and abnormal categories by using the MLP; at this stage we repeatedly perform cross validation to evaluate the propriety of the selected features. In this way, the proposed algorithm can take advantage of both the DT and MLP, by hybridizing these two algorithms. The experimental results demonstrate that the proposed algorithm is able to identify multiple potential biomarkers that enhance the confidence of diagnosis, also showing better specificity, sensitivity and learning error rates than other algorithms. The proposed algorithm represents a promising approach to the identification of proteomic patterns in serum that can distinguish cancer from normal or benign and is applicable to clinical diagnosis and prognosis.

1 Introduction

The improvements in technologies to detect, identify, and characterize proteins, particularly two-dimensional electrophoresis and mass spectrometry, coupled with

* To whom all correspondences should be addressed.
This work was supported by grant No. RTI04-03-03 from the Regional Technology Innovation Program of the Ministry of Commerce, Industry and Energy(MOCIE).

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 167 – 177, 2005.

development of bioinformatics database and analysis tool, makes proteomics a powerful approach to succeed in identifying new tumor markers.

The recent advances made in proteomic profiling technologies, such as SELDI-TOF MS, have enabled the preliminary profiling and identification of tumor markers in biological fluids in several cancer types and the establishment of clinically useful diagnostic computational models. Several studies using serum have now been published, in which the combination of SELDI profiling with artificial intelligence techniques, such as genetic algorithms, clustering, neural networks or decision tree classification algorithms, has produced extremely promising results [1][2].

SELDI-TOF MS has the potential to improve clinical diagnostics tests for cancer pathologies. The goal of this technique is to select a reduced set of measurements or "features" that can be used to distinguish between cancer and control patients. These features are essentially ion intensity levels at specific mass/charge values. However, analyses using SELDI-TOF MS are susceptible to contain thousands of data points. In such cases, because of the computational complexity and cost overhead associated with the processing of this high dimensional data, it is necessary to reduce the dimensionality by selecting useful features. Furthermore, it is important to select those features which are most likely to lead to an accurate classification. Feature selection can also be reinforced by classification. Randomized feature selection generates random subsets of features and assesses their quality independently.

The DT and MLP algorithms have been successfully applied to a wide range of pattern classification problems [3]. The DT generally runs significantly faster in the training stage and gives better expressiveness. The MLP is often more accurate at classifying novel examples in the presence of noisy data. In previous studies, these two techniques were successfully applied to the classification of human cancers and the identification of the potential biomarker proteins using SELDI-TOF mass spectrometry [1][4][5].

In this paper, we describe a novel approach that was used to find cancer specific biomarkers in prostate cancer SELDI-TOF MS data, and to accurately separate the normal and abnormal classes. These biomarkers or marker proteins are features by means of which the abnormal cases (patients) can be potentially distinguished from the normal cases (healthy men). In this approach, we first selected those features that had excellent discrimination power using the DT and found a number of potential biomarkers. As a result, the problem of dimension reduction could be solved, but the reliability of the selected features could not be guaranteed. To solve this problem, as a second step, we accurately classified the selected features into the normal and abnormal categories by using the MLP, and performed repeated cross validation to evaluate the propriety of the selected features.

The proposed algorithm takes advantage of both the DT and MLP, by hybridizing these two algorithms. The experimental results show that, with this technique, it is possible to determine the optimal number of features, and that this algorithm outperforms the other methods in terms of the specificity, sensitivity and confidence.