Computers in Biology and Medicine 43 (2013) 1363–1373

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

ICP: A novel approach to predict prognosis of prostate cancer with inner-class clustering of gene expression data

Hyunjin Kim [a], Jaegyoon Ahn [a], Chihyun Park [a], Youngmi Yoon [b], Sanghyun Park [a,*]

[a] *Department of Computer Science, Yonsei University, South Korea*
[b] *Department of Computer Engineering, Gachon University, South Korea*

ARTICLE INFO

Article history:
Received 4 August 2012
Accepted 19 June 2013

Keywords:
Classification
Clustering
Bioinformatics
Microarray data analysis
Prognosis
Prostate cancer

ABSTRACT

Prostate cancer has heterogeneous characteristics. For that reason, even if tumors appear histologically similar to each other, there are many cases in which they are actually different, based on their gene expression levels. A single tumor may have multiple expression levels with both high-risk cancer genes and low-risk cancer genes. We can produce more useful models for stratifying prostate cancers into high-risk cancer and low-risk cancer categories by considering the range in each class through inner-class clustering. In this paper, we attempt to classify cancers into high-risk (aggressive) prostate cancer and low-risk (non-aggressive) prostate cancer using ICP (Inner-class Clustering and Prediction). Our model classified more efficiently than the models of the algorithms used for comparison. After discovering a number of genes linked to prostate cancer from the gene pairs used in our classification, we discovered that the proposed method can be used to find new unknown genes and gene pairs which distinguish between high-risk cancer and low-risk cancer.

1. Introduction

Prostate cancer is a malignant tumor in the prostate gland. It is one of the most common cancers among men. Other than skin cancer, prostate cancer is the most prevalent cancer in American men. Since prostate cancer is a slow progressing cancer, it has low risk of metastasizing in most cases. Therefore, patients with prostate cancer who are over 70 years old are more likely to die from other causes than from prostate cancer over the 15 years following prognosis. Because prostate cancer may not cause severe pain or have any abnormal signs, it is hard for a patient to know if he has prostate cancer unless the prostate cancer has metastasized to other organs. Therefore, there is a high chance that the cancer has spread to other parts of the body once the patient detects its symptoms. If the prostate cancer has spread to other parts of the body, the metastasized cancer is more dangerous than original prostate cancer, which is a slow-growing cancer. Metastatic cancer that has spread to other areas of the body can grow rapidly and affect vital organs. For that reason, the most important factor related to prostate cancer is not whether ‘it is’ or ‘it is not’ a prostate cancer, but its prognosis, likely progression and probability of metastasis.

Generally, a patient who has cancer can predict his prognosis using clinical stage. The clinical stage is determined by the state of progress, the size and the range of the tumor together with whether or not the cancer has metastasized. The higher the stage number, the bigger the tumor and the more progress it has made. According to a related research study, however, differentiation in cancer cells has a greater effect on prognosis than diagnosed stage [1]. Differentiation refers to an operation or a process of cells specializing in structure and function. Hence, if cells are well differentiated, they are normal cells, and if cells are poorly differentiated, then they are immature and disorganized cells. The results of the research study show that if cancer cells are poorly differentiated, prostate cancer death is more probable even when the tumor is at a lower clinical stage. In a research study on watchful waiting, the two primary risk factors are age at the time of diagnosis and Gleason score [2]. The Gleason score is a means of measuring the aggressiveness of prostate cancer [3–7]. It is obtained by adding the two Gleason scale grades together. Each cell is given a Gleason scale grade according to the degree of differentiation of the cell. The scale grade is determined by examining cells from the prostate under a microscope during a biopsy. Each cell is then given a grade from 1 to 5. The higher the degree of differentiation, the lower the grade number it is given. Once the two most common types of cancer cells are identified in the prostate, the two grades of these two types are then added together to produce a Gleason score. Therefore, a Gleason score ranges from 2 to 10. The lower the score, the slower the cancer is

[*] Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.
E-mail address: sanghyun@cs.yonsei.ac.kr (S. Park).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.06.014

growing, and the higher the score, the faster the cancer is likely to be growing and the more aggressive it is. In general, a Gleason score of 7 is considered intermediate and a score of 6 or less has a good prognosis. A score of 8 or higher has a poor prognosis.

Since the Gleason score is determined by examining cells from the prostate under a microscope, it cannot be considered as an absolute index of the prognosis for prostate cancer [8]. For that reason, we used two kinds of data in the experiment. For the first data set (GSE 15484), we considered a Gleason score of 7 or less (2–7) to be non-aggressive and a score of 8 or more (8–10) to be aggressive. We classified using this method in this paper. The second data set (GSE 21034) consists of the aggressive and non-aggressive prostate cancer samples obtained from clinical examination without reference to Gleason score. We did experiments using the same method as the one applied to the first data.

Prostate cancer has heterogeneous characteristics, which means samples in the same class do not necessarily have similar gene expression levels [9–11]. Classification algorithms for handling prostate cancer gene expression levels have to reflect that heterogeneity. The key to successfully overcoming this heterogeneity is capturing the distinctive gene expression level groups in each class and using these groups when performing classification. We propose an efficient classification method ICP (Inner-class Clustering and Prediction) based on the heterogeneity of gene expression levels to classify prostate cancer into two categories, high-risk and low-risk prostate cancer. ICP can distinguish several different gene expression level groups by using inner-class clustering. It reduces false positives and false negatives. Most of the other methods do not consider the different types in the same class. The classification method has 5 major phases (Fig. 1).

The first phase is a gene selection phase that reduces the number of genes to be used in the analysis, because if we experiment with a large number of genes, the time complexity is too large. In this phase, we sort out n top-ranked genes using relief-A and symmetrical uncertainty algorithms, which are verified feature selection methods. In the second phase, by making use of inner-class clustering, we calculate the cluster information for each gene pair, which is carried out in the first phase. In the third phase, we measure the degree of dispersion using the cluster information from the gene pairs we obtained in the second phase, and rank the gene pairs from highest to lowest according to the degree of dispersion. If there are multiple gene pairs which have the same score, we use variance-based secondary score to select a unique gene pair. Phases 4 and 5 are the phases to select a class. By using vote sets from the phase 3, we execute the prediction in each voter, and then select the class with the most votes.

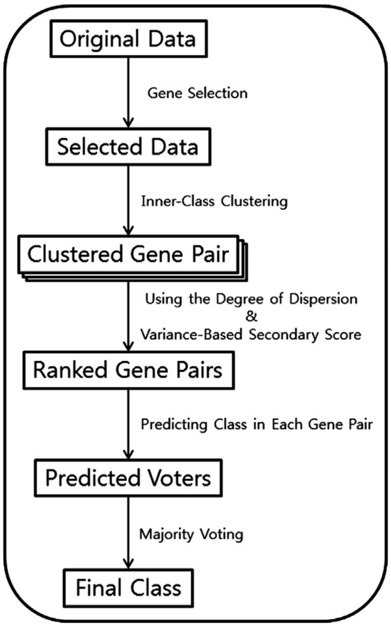

Fig. 1. Flow chart of ICP algorithm.

The results in distinguishing between the high-risk and the low-risk prostate cancers with the proposed classification method show that the proposed classification method is more efficient than other existing classification methods. Moreover, looking into the frequently appearing genes and gene pairs, which are ranked by the degree of dispersion, informed us that those genes and gene pairs are closely related to biological processes or to prostate cancer. Classification by making use of inner-class clustering is novel and is of great value because it can be applied to multi-class classifications.

2. Related works

Almost all classification problems of cancer diagnosis and prognosis can be solved by machine learning methods. These methods develop classifiers with training samples which are already classified and predict the class of test samples based on those classifiers.

The most popular cancer-related classification method among the machine learning methods is SVM (Support Vector Machine) [12]. SVM finds the linear optimal hyper plane which separates gene expression data samples into two groups and uses that plane to classify the given samples. After applying the transform function, the non-linear data can be handled in the same way as the linear data in SVM. The transform function is called the kernel function and there are many types of kernel functions. There have already been many studies on when the kernel function should be used and what type of function should be used [13–16]. A few regression versions of the SVM [17,18] also exist but methods which use SVM for gene expression data usually focus on which genes are to be selected to form a hyper plane rather than how to change the main algorithm of SVM to be more efficient. If genes are closely correlated, we can apply SVM-RFE (Recursive Feature Elimination) [19], one of the methods that focuses on gene selection. When SVM is finding a hyper plane and using it on the classification, it is important to obtain the maximum margin between two classes. The L1-norm penalty is helpful to obtain the soft maximum margin [20]. The L1-norm SVM does not choose all the genes which have a high correlation among themselves. To solve this problem for the L1-norm SVM, Wang [21] proposed HHSVM (Hybrid Huberized Support Vector Machine) making use of the huberized hinge loss function and elastic-net penalty.

Logistic regression [22] is similar to linear regression because a function is created based on the shape of the data so the class of a sample can be predicted. But the difference between logistic and linear regression is that logistic regression's prediction result is binomial, not continuous. Logistic regression method can be applied to other models, so extensibility is the one of the merits of this method. For instance, a logistic regression method combined with a parametric bootstrap model for the gene expression data classification problem was proposed by Liao [23] in 2007.

Another method called decision tree induction is a classification method which uses flowchart-like tree structure. Each internal node denotes a test on an attribute, each branch describes a result of the test, and each leaf node represents a class label. The attribute values of a test sample are tested with the internal nodes in the decision tree. A path can be traced from the root node to a leaf node and the leaf node's class label indicates the predicted class of the test sample. ID3 [24], C4.5 [25], and CART [26] are different versions of the decision tree which have different