Neurocomputing 165 (2015) 81–89

Contents lists available at ScienceDirect

Neurocomputing

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

Pathway activity transformation for multi-class classification of lung cancer datasets

Worrawat Engchuan, Jonathan H. Chan *

Data and Knowledge Engineering Laboratory (D-Lab), School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand

ARTICLE INFO

Article history:
Received 28 February 2014
Received in revised form
16 June 2014
Accepted 6 August 2014
Available online 20 April 2015

Keywords:
Pathway activity transformation
Multi-class classification
Lung cancer
Support Vector Machine
Multilayer perceptron
ANOVA

ABSTRACT

Pathway-based microarray analysis has been found to be a powerful tool to study disease mechanisms and to identify biological markers of complex diseases like lung cancer. From previous studies, the use of pathway activity transformed from gene expression data has been shown to be more informative in disease classification. However, current works on a pathway activity transformation method are for binary-class classification. In this study, we propose a pathway activity transformation method for multi-class data termed Analysis-of-Variance-based Feature Set (AFS). The classification results of using pathway activity derived from our proposed method show high classification power in three-fold cross-validation and robustness in across dataset validation for all four lung cancer datasets used.

1. Introduction

Cancer may be treated as a genetic disorder with mostly unknown causes and mechanisms. Among the cancer types, lung cancer is a major killer disease in both America and East Asia [1,2]. The smoking behavior and environmental factors that drive lung cancer are already widely studied [3]. However, how the genetic factors affect lung cancer still yet to be clearly understood because it is a complex genetic disease, which is developed by the co-occurrence of many genetic changing events [4]. The nature of lung cancer can be divided into two types, which are non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). The most common lung cancer type is NSCLC, which is found in roughly 80% of lung cancer patients. NSCLC can also be sub-categorized as Squamous cell carcinoma (SCC) and Adenocarcinoma (AC) [5]. Traditionally, physical analyses of tissues are performed for lung cancer diagnosis and prognosis using Chest X-ray, Computed Tomography (CT) scan and Magnetic Resonance Imaging (MRI) [6,7]. Unfortunately, these techniques can only detect the malignant cells in late stage of lung cancer and would result in low survival rates (approximately 16% for NSCLC and 6% for SCLC) [8]. The advances in molecular biology can now acquire the information of DNA, RNA and proteins, which can be applied for the studying of cancers in order to detect the formation of tumor in earlier stages, which should result in an increase in the survival rate. In 1995, Plebani et al. used Analysis of Variance (ANOVA), Bonferroni's test and Student's *t*-test on protein assays of 7 proteins. Their results show that the combination of CYFRA 1-21, the most sensitive marker, and TPM, the highest accuracy in disease diagnosis, could be a useful marker for lung cancer diagnosis [9]. However, as that study was investigated using low-throughput data, prior knowledge of disease is required for acquiring the information of candidate markers. Also, investigating only small number of markers may result in a limitation to discover novel markers. Then, the microarray technique, which simultaneously measures the expression of thousands of genes, was introduced. Bhattacharjee et al. used machine learning approaches for analyzing lung cancer microarray data. Feature selection based on variation of expression level and hierarchical clustering were used to discover the subclasses of lung cancer, then K-nearest neighbor was applied to identify the markers of each subclass [10]. Gordon et al. proposed a lung cancer diagnosis technique based on gene expression ratios calculated from the pairs of genes which are highly inversely correlated. The method was tested on classification of lung cancer and mesothelioma groups, which yielded 95% and 99% in accuracy for mesothelioma and lung cancer, respectively [11].

* Corresponding author. Tel.: +66 2470 9819; fax: +66 2872 7145.
E-mail addresses: worrawat.eng@st.sit.kmutt.ac.th (W. Engchuan), jonathan@sit.kmutt.ac.th (J.H. Chan).

http://dx.doi.org/10.1016/j.neucom.2014.08.096
0925-2312/

Analysis of gene expression level only is usually inadequate for achieving the understanding of the complex traits of diseases like lung cancer. So the analysis techniques of multiple-genes along with the integration of other biological knowledge have been developed to improve the understanding and increase the accuracy of lung cancer diagnosis. In 2008, Lee et al. proposed the pathway activity inferring method on gene expression data. The definition of pathway activity is the responsive level of a pathway to the specific condition. The method was named as condition responsive genes (CORGs), which use greedy search strategy to identify the CORGs set of each pathway and then summarize the expression level of CORGs into a single value called pathway activity [12]. Another pathway activity inferring method named Negatively Correlated Feature Sets using ideal markers (NCFS-i) then was developed and proposed by Sootanan et al. [13]. It produces a set of Phenotype-Correlated Genes (PCOGs) instead of CORGs. The classification of pathway activity derived from a NCFS-i method has shown to provide higher classification power and robustness [13]. Chan et al. also investigated on feature selection approaches of pathway activity data in order to increase the accuracy and robustness of classification [14].

Much works in developing of pathway activity transformation algorithms have been done for binary-class classification so far. However, in many medical works now, there is a need to diagnose more than two classes of disease, like staging. So, a different method to implement pathway activity-based algorithm is required for multi-class classification. Several algorithms have been developed to solve problems in classification of multi-class data based on binary-class classification. One vs. All (OVA) and One vs. One (OVO) are the simplest ensemble multi-class classification algorithms, which construct several binary classifiers based on pair of classes. OVA will construct n classifiers and each classifier will distinguish one class to other classes. By contrast, OVO will construct $C(n,2)$ classifiers where n is the number of classes. Each classifier plays a role in classifying of one class and another class [15]. The classifiers then are combined using majority voting or averaging. In 2001, Ramaswamy et al. employed Support Vector Machine (SVM) approach and One vs. All (OVA)-based approach on microarray data of different tumor types and normal samples. The multi-class classification yielded approximately 78% overall accuracy [16]. Liu et al. [17] combined GA-algorithm and SVM/OVO voting strategy for multi-class classification task. The GA-algorithm was used for gene selection, which would select non-redundant and discriminant genes. Then SVM/OVO was used to build voting ensemble classifiers, which gave 68–85% accuracy on leave-one-out cross-validation [17]. The utilizing of OVA or OVO techniques may provide an acceptable classification performance. Nevertheless, the computational time of these techniques will increase due to the increasing number of classes by at least n or up to $C(n,2)$ times compared to original classification. In this paper, we propose a modified NCFS-i method for pathway activity transformation in multi-class dataset termed ANOVA-based Feature Set (AFS). Then established multi-class techniques were used for classification. The performance of the AFS method was compared with traditional OVA and OVO methods in terms of accuracy, robustness and time consuming by using publicly available lung cancer datasets.

The organization of the remaining paper is as follows. Section 2 describes the datasets used in this work, followed by the methodology to implement AFS, OVA and OVO methods, and the evaluation procedure. Section 3 shows the results of stratified three-fold cross validation and across dataset validation. Section 4 discusses the results and some biological implications. Finally, Section 5 concludes the work.

2. Materials and method

2.1. Microarray datasets and pathway information

In order to evaluate our pathway activity transformation algorithm, we take the actual lung cancer datasets from Gene Expression Omnibus (GEO), a public gene expression database. In this study, there are 4 lung cancer datasets, which have accession numbers as GSE2109, GSE10245, GSE18842 and GSE43580. The information about each dataset is shown in Table 1. Four multi-class datasets (GSE2109, GSE10245, GSE18842 and GSE43580) were used for assessing the performance of our algorithm.

For pathway information, we obtain the curated gene set of canonical pathway data from MSigDB (file version; c2.cp.v3.1.symbols), which is a database that stores pathway data from several databases such as KEGG, PubMed, BioCarta, etc. [18] The pathway data is then preprocessed by removing the pathways that contain less than 10 genes, resulting in a total of 1452 remaining pathways.

2.2. OVA and OVO

The One vs. All (OVA) and One vs. One (OVO) approaches are the simplest ways to extend binary class classification to multi-class classification by breaking down the multi-class problem into multiple binary class problems. So, these two approaches allow the direct application of NCFS-i [13] for the multi-class disease classification problem. Figs. 1 and 2 show how to implement OVA and OVO with NCFS-i for multi-class lung cancer classification, respectively.

2.3. ANOVA-based Feature Set (AFS)

The incorporation of gene-set information in microarray data analysis has increased the accuracy, robustness and biological relevant of the models [12]. However, most of the gene-set-based activity transformation methods are limited to binary-class problems only. By extending NCFS-i by OVA or OVO approach to transform gene expression level into many pathway activity datasets to serve multiple binary class classifiers, those pathway activity datasets are calculated using different PCOGs set, which each was selected using only a part of the whole information existing in training data. By using ANOVA and correlation analysis instead of Student's t-test like NCFS-i, AFS can be applied to use the whole information of the training data to transform gene expression data into single pathway activity dataset, which is ready for multi-class classification.

AFS has two different main steps in pathway activity transformation for transforming pathway activity in multi-class dataset (Fig. 3), which are ranking and identification of PCOGs and summarizing of gene expression levels. In stratified three-fold cross-validation, two thirds of the data is used to identify PCOGs. For each pathway, the gene members are ranked. ANOVA is used to calculate the F-values, which will assess the variance of z-transformed expression level between classes. We hypothesize

Table 1
Lung cancer datasets information.

Multi-class dataset	AC Stage 1	AC Stage 2	SCC Stage 1	SCC Stage 2	Total
GSE2109	17	6	21	8	52
GSE10245	22	14	9	6	51
GSE18842	12	0	27	3	42
GSE43580	41	36	34	39	150