Expert Systems With Applications 59 (2016) 13–19

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A multi-objective heuristic algorithm for gene expression microarray data classification

Jia Lv, Qinke Peng*, Xiao Chen, Zhi Sun

Systems Engineering Institute, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xianning Road, 710049 Xi'an, China

ARTICLE INFO

Article history:
Received 26 May 2015
Revised 29 February 2016
Accepted 16 April 2016
Available online 19 April 2016

Keywords:
Microarray
Gene selection
Small number of selected genes
Multi-objective
Heuristic algorithm

ABSTRACT

Microarray data has significant potential in clinical medicine, which always owns a large quantity of genes relative to the samples' number. Finding a subset of discriminatory genes (features) through intelligent algorithms has been trend. Based on this, building a disease prognosis expert system will bring a great effect on clinical medicine. In addition, the fewer the selected genes are, the less cost the disease prognosis expert system is. So the small gene set with high classification accuracy is what we need. In this paper, a multi-objective model is built according to the analytic hierarchy process (AHP), which treats the classification accuracy absolutely important than the number of selected genes. And a multi-objective heuristic algorithm called MOEDA is proposed to solve the model, which is an improvement of Univariate Marginal Distribution Algorithm. Two main rules are designed, one is 'Higher and Fewer Rule' which is used for evaluating and sorting individuals and the other is 'Forcibly Decrease Rule' which is used for generate potential individuals with high classification accuracy and fewer genes. Our proposed method is tested on both binary-class and multi-class microarray datasets. The results show that the gene set selected by MOEDA not only results in higher accuracies, but also keep a small scale, which cannot only save computational time but also improve the interpretability and application of the result with the simple classification model. The proposed MOEDA opens up a new way for the heuristic algorithms applying on microarray gene expression data.

1. Introduction

The rapidly maturing microarray technology has made it feasible to measure the expression levels of tens of thousands of genes in a single test, which can provide us a series of gene expression profiles at a molecular level (Schena, Shalon, Davis, & Brown, 1995). The useful genes extracted from microarray data, which can be treated as biomarkers, can help clinical decision on disease diagnosis, prognosis and treatment (Latkowski & Osowski, 2015b; Rocke, Ideker, Troyanskaya, Quackenbush, & Dopazo, 2009). However, finding the useful genes from microarray data is not an easy work due to their large number of genes relative to the samples' number (Dessi & Pes, 2015; Qi & Yang, 2013).

Due to the large quantity of genes, how to remove the irrelevant and redundant genes is a challenge. Feature selection and feature extraction are two main methods (in microarray problems, the feature is equal to the gene and they are interchangeable). Feature selection is to select only the relevant features, which do not change the features' form. Differently, feature extraction aims at extracting the essential features from the original data which may transform the initial features. Here we mainly focus on the feature selection methods.

According to the appearance of classifiers, feature selection methods can be organized into three categories: filter, wrapper and embedded methods. The filter methods are usually based on the statistic approaches, such as the mRMR (Peng, Long, & Ding, 2005) which is put forward based on the mutual information and the PLSRFE (You, Yang, & Ji, 2014a; 2014b) and KernelPLS (Sun, Peng, & Shakoor, 2014) which is proposed based on the partial least squares. They usually give every feature a weight which represents its importance through statistic approaches so that they are always computational fast. But the drawback is that they select by individual so that they always focus on the features which are good as individuals instead of the features which are strong discriminatory as groups (Sun et al., 2012a; 2012b). As an improvement, the top scoring pairs(TSP) algorithm (Geman, d'Avignon, Naiman, & Winslow, 2004) which selects gene based on the pairwise comparisons takes the bivariate interactions between genes into account. And then it is improved and extended to K-top scoring pairs(k-TSP) (Tan, Naiman, Xu, Winslow, & Geman, 2005). Even though

* Corresponding author. Tel.: +86 029 82667964.
E-mail addresses: answer_lv@stu.xjtu.edu.cn (J. Lv), qkpeng@mail.xjtu.edu.cn (Q. Peng), xchen_xjtu@163.com (X. Chen), sun.zhi@stu.xjtu.edu.cn (Z. Sun).

http://dx.doi.org/10.1016/j.eswa.2016.04.020
0957-4174/

TSP based algorithms select features by pair instead of by individual, they can not capture the high order interactions among genes. And the wrapper methods make an improvement, which usually connect classifiers such as K-nearest neighbor(KNN) (Kar, Das Sharma, & Maitra, 2015; Park & Kim, 2015), artificial neural networks(ANN) (Yen-Chen, Wan-Chi, & Hung-Wen, 2014), support vector machine(SVM) (Latkowski & Osowski, 2015a; Statnikov, Aliferis, Tsamardinos, Hardin, & Levy, 2005) with some optimization algorithms such as genetic algorithm and particle swarm optimization algorithm. The optimization algorithms focus on selecting features and the classifiers are used to evaluate the selected subset of genes. And the embedded method is a kind of classifiers which contain feature selection process itself, such as Decision Tree (Chen et al., 2014) and SVMRFE. Both wrapper and embedded methods can select features by group so that they are potential to capture the high order interactions among genes. In fact, they always outperform the filter methods in terms of the classification accuracy. But they are time consuming because of the join of classifiers and the high dimension. In order to increase efficiency, the wrapper and embedded methods are always combined with filter methods to constitute hybrid methods (Chuang, Yang, Wu, & Yang, 2011; Huang, Cai, & Xu, 2007; Kim, Ahn, Park, Yoon, & Park, 2013; Lotfi & Keshavarz, 2014; Yang, Chuang, & Yang, 2010; Zhou & Dickerson, 2014).

However, the traditional wrapper and embedded methods always generate complex models which involve too many genes. The complex model leads to a hard interpreted result and the large number of genes results in an expensive cost which make it impossible to apply in clinics. If we want to build a disease prognosis expert system and apply it in clinics, it must be cost-effective. Nowadays, kinds of feature selection methods which take the cost into account have been trend(Du, Guo, Huang, Li, & Guo, 2015; Gao & Chen, 2015; Li et al., 2015). Similarly, in microarray data, researchers prefer to get small gene set with high classification accuracy so that the result can improve the interpretability and application (Wang, Chu, & Xie, 2007; Wang & Simon, 2011).

In this paper, in order to find the target gene set better and faster, a new multi-objective optimization model is built, which contains both classification accuracy (ACC) and the number of selected features (NSF). Considering that in microarray problems the ACC is much more important than the NSF, so our multi-objective model is built according to the analytic hierarchy process (AHP), which is frequently-used in multi-objective decision. In addition, we propose a multi-objective optimization algorithm called MOEDA to solve it, which is an improvement of Univariate Marginal Distribution Algorithm (UMDA). In order to ensure that the MOEDA can solve the problem successfully, two rules are designed for it: one is 'Higher and Fewer Rule' which is used for evaluating and sorting individuals and the other is 'Forcibly Decrease Rule' which is used for generating potential individuals with high ACC and small NSF. The ACC is calculated by the state of the art classifier SVM. What is more, in order to reduce the redundant and irrelevant gene quickly and effectively, the mRMR criterion is used for pre-selecting before MOEDA. Ten microarray datasets are brought for test, including both binary and multiple classification datasets. The results show that our method can not only achieve high classification accuracy, but also reduce the number of selected features effectively.

2. Method

2.1. Feature pre-selection

Microarray data is differentially expressed and there exists many irrelevant features, so usually a filter-based gene ranking algorithm will be used for pre-selecting. In our test, we used the state of the art filter algorithm mRMR to pre-select features. The mRMR considers not only the relevance between feature and label (max-relevance), but also the redundancy between feature and feature (min-redundancy), as the equations below show

$$\max \frac{1}{|S|}\sum_{x_i \in S} I(x_i; c).$$

$$\min \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i, x_j).$$

Obviously, the pre-selected features are usually some features which perform excellent as individuals. So in order to avoid too many features which are strong discriminatory as groups but weak as individuals missing, a threshold of 300 features is set in the pre-selection process.

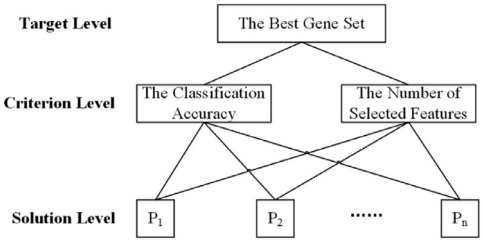

Fig. 1. The schematic diagram of the model structure.

2.2. The multi-objective heuristic algorithm

2.2.1. Problem description

Let $O = (X, Y, F)$ be a pre-selected dataset. $X = \{x_1, x_2, \cdots, x_n\}$ is the sample set. $Y = \{y_1, y_2, \cdots, y_n\}$ is the label set. And $F = \{f_1, f_2, \cdots f_m\}$ is the feature set. each $x_i \in X$ is a vector containing m elements, $x_i = \{u_{i1}, u_{i2}, \cdots, u_{im}\}$. We need to find out a group of features $T = \{f_{i1}, f_{i2}, \cdots, f_{ik}\} \subseteq F(k \leq m)$ which can differentiate the gene expressions of samples accurately. Moreover, the number of selected feature of the subset k should be small as far as possible.

2.2.2. The multi-objective model

In order to solve the problem above, we propose a multi-objective model according to the analytic hierarchy process (AHP), as Fig. 1 shows. First of all, we hope to get the best gene set, as the target level shows. But what gene set is the best? Then, it is described in the criterion level where there exist two criterion: maximize the ACC and minimize the NSF. What is more, the first criterion is absolutely important than the second one. Finally, the bottom is the solution level in which everyone is a candidate gene set. We need to iterate to achieve the best.

2.2.3. MOEDA

We design MOEDA to solve the model above. MOEDA is a kind of the estimation of distribution algorithm (EDA), which guides the search for the optimum by building and sampling explicit probabilistic models of promising candidate solutions. In MOEDA, the population is composed of several individuals and each individual corresponds to a candidate solution, which is a binary string and '1' represents presence of the feature while '0' represents absence of the feature. In each generation, all the individuals will be sorted based on their objective values (in this work the objective value is the classification accuracy). The top-ranked part are treated as elite individuals (*EIs*) which are remained in the next generation and used to build the probabilistic model. And the others called