journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A heuristic biomarker selection approach based on professional tennis player ranking strategy

Bin Han[a,*,1], Ruifei Xie[b,*,1], Lihua Li[a], Lei Zhu[a], Shen Wang[a]

[a] College of Life Information Science and Instrument Engineering, Hangzhou Dianzi University, Hangzhou, People's Republic of China
[b] Hangzhou Cancer Hospital and The First People Hospital of Hangzhou, Hangzhou, People's Republic of China

ARTICLE INFO

Article history:
Received 29 March 2013
Received in revised form
7 October 2013
Accepted 7 October 2013

Keywords:
Feature selection
Microarray
Monte Carlo
Dynamic ranking

ABSTRACT

Extracting significant features from high-dimension and small sample size biological data is a challenging problem. Recently, Michał Draminski proposed the Monte Carlo feature selection (MC) algorithm, which was able to search over large feature spaces and achieved better classification accuracies. However in MC the information of feature rank variations is not utilized and the ranks of features are not dynamically updated. Here, we propose a novel feature selection algorithm which integrates the ideas of the professional tennis players ranking, such as seed players and dynamic ranking, into Monte Carlo simulation. Seed players make the feature selection game more competitive and selective. The strategy of dynamic ranking ensures that it is always the current best players to take part in each competition. The proposed algorithm is tested on 8 biological datasets. Results demonstrate that the proposed method is computationally efficient, stable and has favorable performance in classification.

1. Introduction

Feature selection has been widely used in pattern recognition, machine learning and data mining. Despite the impressive achievements, we observe great challenges arising from high dimensional noisy biological data [1] such as microarray micRNA or protein data analysis where the dataset may contain tens of thousands of features and few samples. To this end, many feature selection methods have been proposed. These methods can be roughly categorized into two types, filter and wrapper depending on whether feature subset evaluation employing classifiers [2]. Filter methods [3], such as *t*-test [4], nonparametric Kolmogorov–Smirnov test [5], rank features by significance analysis without reference to classification tasks. Wrapper methods [3], such as SVM-RFE [6] and stepwise ANN [7] select feature by embedding a classifier with the highest quality of classification for some clinical or biological outcomes. The major drawback of the filter method is that most proposed techniques are univariate and the feature dependencies are overlooked, leading to unsatisfactory classification performance [3]. The wrapper methods usually are able to achieve a higher classification rate, but they are computationally intensive, have a higher risk of over-fitting and the performance of feature are classifiers dependent [3,8].

Recently, Draminski proposed a new type of random searching algorithm the Monte Carlo (MC) feature selection algorithm [9], which aims to select classification task relevant features regardless of the classifier. It integrates the individual

* *Corresponding author*. Tel.: +86 057186826084.
E-mail addresses: hanb.bin@gmail.com (B. Han), ruifei007@163.com (R. Xie).
[1] Co-first authors. These two authors contributed equally to this work.
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.10.008

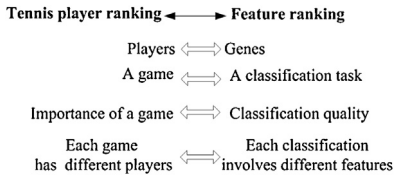

Fig. 1 – Professional tennis player ranking and feature ranking.

(feature) evaluation, subset evaluation and weighted accuracies to assess classification ability, which can prevent undue influence of a majority class on the performance index. It also adopts the re-sampling technique and stochastic searching strategy, which enable it to search over extremely large spaces and avoid the risk of sticking at local optimum without computation overhead. It performs relatively better in comparison to other methods. However this algorithm treats all features equally and features are evaluated at the end of searching. The information of feature rank variations is not utilized and the ranks of features are not dynamically updated.

In this article, we propose a novel biological feature selection algorithm, which integrates the ideas of the professional tennis players ranking (PTPR), such as seed players and dynamic ranking into the Monte Carlo simulation. This is because we notice that, first of all, both the biological feature selection and professional player ranking are facing the similar problem, which is 'the curse of dimensionality', i.e., in biological feature selection, thousands of features always go with only less than one hundred samples, while in the ranking of professional tennis players, tens of thousands players are ranked through about 50 tournaments registered in the Association of Tennis Professionals (ATPs) each year. Besides, they have other important analogies. A detailed comparison is shown in Fig. 1. Secondly, the strategy of professional player ranking has already been proved efficient, e.g., Professional players are ranked through a number of competitions rather than determined by one competition and the final score represents the player's overall performance, the ranks of players are dynamically adjusted after each competition rather than ranking all the players at the end of 'season'. Besides, in PTPR, the past performance of a player is taken into account and not all players are treated equally; this is considered as seed or non-seed players in PTPR. The seed player set-up ensure that best players have more chance to remain in games, more importantly, the participation of seed plays make games more competitive and selective, i.e., if players beat a seed player, they should get more points than when they beat many average players, and if so, they should be considered as a dark horse. These ideas can be borrowed to improve the efficiency of biological feature selection. Therefore, we introduce the seed feature into MC algorithm and update the ranks of features dynamically. This approach is applied to 8 different datasets and its performance is evaluated based on the approach's convergence, stability and the classification quality of the selected features.

2. Data and method

2.1. Data

8 datasets were used in our study to test the algorithm. Leukemia [10], Colon [11], Glioma [12], Prostate [13], Lung [14], SELDI (OCWCX2b) [15] and micRNA (Leukemia) [16] were wildly used public data sets and studied by many researchers. Ovarian cancer samples were obtained from Duke University Center and H. Lee Moffitt Cancer Center and Research Institute. A summary of datasets is in Table 1.

2.2. PTPR algorithm

As mentioned, both PTPR [17] and feature selection face similar problems. A player in PTPR corresponds to a biological feature; a game can be regarded as an evaluation experiment such as a classification task, and in PRPR players compete with different players, while in random searching feature selection algorithm such as MC, each classification task may have different feature combinations. They both rank players (features) according to the performance of the players (feature) in competitive games. These similarities are shown in Fig. 1. PTPR has been proven to be successful and the ranking results are well recognized. We noticed that three factors help the PTPR become very efficient. Firstly in PTPR games are regarded to be quite different in terms of the contributions to the ranking, i.e., that how a game is factored in is determined by the history rankings of the players participating in that game. Secondly and essentially, players are categorized as seed and non-seed and the value of each win is rated based on the opponent's performance. i.e., if you beat one formidable opponent you will get more credits than if you beat one average player. Therefore the more good players the game has, the more competitive the game, and the more selective the game. Likewise, features are very different in performances and each classification task should have different contribution to the ranking of features. While in MC, the features are randomly selected and treated equally. Features are pairwise compared to all other features. In fact most randomly selected features are average features. Even some of them can beat others in classification; they are still not necessarily good features because their opponents may be ordinary features. Thirdly, in PTPR players are ranked through a series of widely recognized games and PTPR update the ranking after each game, while in MC all the features are evaluated at the end of searching. The information of feature rank variation is not utilized.

Based on above, we proposed a novel random feature selection approach. Like PTPR, we categorized the features as seed and non-seed variables. The current best features were stored in the 'seed feature list'. The seed features have a higher probability to be selected into classification tasks (games). This ensures that there are always some formidable features (players) kept in classification tasks (games). Therefore the corresponding classification tasks (games) become more competitive and selective. Besides, the algorithm updated the seed feature list in each iteration, which means that the current best features always are enrolled into the list and take part in