Engineering Applications of Artificial Intelligence 70 (2018) 92–108

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A data mining framework based on boundary-points for gene selection from DNA-microarrays: Pancreatic Ductal Adenocarcinoma as a case study

Juan Ramos [a], José A. Castellanos-Garzón [a,b,*], Juan F. de Paz [a], Juan M. Corchado [a,c]

[a] University of Salamanca, IBSAL/BISITE Research Group, Edificio I+D+i, 37007 Salamanca, Spain [1]
[b] University of Coimbra, CISUC, ECOS Research Group, Pólo II - Pinhal de Marrocos, 3030-290 Coimbra, Portugal [2]
[c] University of Salamanca, Osaka Institute of Technology, BISITE Research Group, Edificio I+D+i, 37007 Salamanca, Spain

ARTICLE INFO

Keywords:
Feature selection
Gene selection
Data mining
Cluster analysis
Evolutionary computation
Boundary point
DNA-microarray
Visual analytics
Filter method
Boundary gene

ABSTRACT

Gene selection (or feature selection) from DNA-microarray data can be focused on different techniques, which generally involve statistical tests, data mining and machine learning. In recent years there has been an increasing interest in using hybrid-technique sets to face the problem of meaningful gene selection; nevertheless, this issue remains a challenge. In an effort to address the situation, this paper proposes a novel hybrid framework based on data mining techniques and tuned to select gene subsets, which are meaningfully related to the target disease conducted in DNA-microarray experiments. For this purpose, the framework above deals with approaches such as statistical significance tests, cluster analysis, evolutionary computation, visual analytics and boundary points. The latter is the core technique of our proposal, allowing the framework to define two methods of gene selection. Another novelty of this work is the inclusion of the age of patients as an additional factor in our analysis, which can leading to gaining more insight into the disease. In fact, the results reached in this research have been very promising and have shown their biological validity. Hence, our proposal has resulted in a methodology that can be followed in the gene selection process from DNA-microarray data.

1. Introduction

Advances in *bioinformatics* in the last years have made it possible to apply *artificial intelligence* hybrid techniques to further understand and validate the achieved results. Bioinformatics is in fact one of the most controversial areas of research at present, since it deals with the development and/or application of methods and algorithms to turn biological data into knowledge of biological systems, often requiring further experimentation from initial data, Bourne and Wissig (2003).

Meanwhile, *data mining* and *functional genomics* have also gained attention since the publication of several complete genome sequences as well as the human genome. One of the most advanced and challenging ways of studying molecular events has been the monitoring of gene expression patterns from *DNA-microarrays*. Microarrays can be viewed as a type of device (a chip) in which, a large number of diverse entities, such as peptides, oligonucleotides, biological molecules, cells, tissues, etc., are located on its surface, and placed in an orderly and accurate way. Once these entities are attached on the surface of the chip, they can be simultaneously evaluated in a single essay (Berrar et al., 2003; Chan and Kasabov, 2004; Geoffrey et al., 2004; Jiang et al., 2004; Speed, 2003).

An important research area developed from the data domain above is *gene/feature selection*, which deals with the discovery of gene subsets relevant for a particular target. Such genes are called *informative* (or *differentially expressed genes*) and are the basis for developing *classifiers* in the study of disease diagnosis and prognosis. They are also studied by pharmaceutical companies, whose efforts are focused on identifying those genes that can be targeted by drugs (Inza et al., 2004; Jager et al., 2003; Kumari and Swarnkar, 2011; Lazar et al., 2012; Simeka et al., 2004). While significant efforts have been placed in the development of new methods and strategies to discover informative genes, the problem remains a challenge today since there is not a single technique able to solve all the underlying issues. In general aspects, feature selection methods can be classified into four categories: *filters, wrappers, embedded* and a more recent method group known as *ensemble* (Natarajan and Ravi, 2014; Shraddha et al., 2014; Tyagi and Mishra, 2013; Wang et al., 2005). Each of these categories demanding unification of different techniques as *supervised* and *unsupervised learning, evolutionary computation,*

* Corresponding author at: University of Salamanca, IBSAL/BISITE Research Group, Edificio I+D+i, 37007 Salamanca, Spain .
E-mail addresses: juanrg@usal.es (J. Ramos), jantonio@usal.es (J.A. Castellanos-Garzón), fcofds@usal.es (J.F. de Paz), corchado@usal.es (J.M. Corchado).
1 http://bisite.usal.es.
2 https://www.cisuc.uc.pt/groups/show/ecos.

https://doi.org/10.1016/j.engappai.2018.01.007
Received 2 June 2016; Received in revised form 30 September 2017; Accepted 20 January 2018

visual analytics, among others, in order to gain insight into the problem at hand.

Hence, this research proposes a framework relating hybrid techniques of artificial intelligence and statistics to gene subset selection from gene expression data, which we call *HybridFrame*. Three major characteristics can be stressed from HybridFrame. To begin, it develops a methodology addressing two different methods of gene selection, one based on evolutionary algorithms and the other one, based on the intersection of results coming from different methods. Secondly, the core idea of HybridFrame has been focused on cluster boundary genes to determine informative genes. Furthermore, this framework suitably links a set of hybrid techniques as statistical significance tests, cluster analysis, genetic algorithms, visual analytics and boundary points, to successively reduce (as a filtering strategy) the involved dataset until reaching a small subset of meaningful genes related to the target disease.

We have used hybrid techniques to build a data mining framework for gene selection tasks, because they provide more robust and stable solutions than simple methods (Guyon, 2003; Jager et al., 2003; Lazar et al., 2012). Generally, simple methods of gene selection assume that some criterion should be met in data, which does not have to be true for all data types. Hence, hybrid techniques fusion different simple methods to reach solutions holding more than one criterion, making solutions more stable with respect to variations in data. On the other hand, hybrid techniques are more flexible to changes in user needs and allow us to replace the methods taking place in the overall proposal without carrying out meaningful changes.

1.1. Case study, impact and motivation

As a case study to apply and validate our proposal, we have focused our attention on the tissue sample study of *pancreatic ductal adenocarcinoma* (PDAC) through microarray technology, given that PDAC has been identified as one of the most aggressive types of existing cancer (Badea et al., 2008a, b), with a majority of cases, unfortunately, detected in advanced stages due to the lack of early symptoms, Crnogorac-Jurcevic et al. (2013). Hence, PDAC patients have a median survival of less than six months and a five-year survival rate of about 5% patients, Hezel et al. (2006). Indeed, 60%–70% patients already present metastasis when the cancer is detected. In spite of the fact that much knowledge from PDAC molecular processes has been revealed in the last few years, the scientific community is still far from developing effective therapies leading to an ability to face this pathology.

One of the main causes for this is the drug's low effectiveness in PDAC treatment, which has been attributed to a high dynamic relation between cancer cells and the stroma, Bhaw-Luximon and Jhurry (2015). This has resulted in many events allowing stroma formation to act as a protective environment of the tumor. Moreover, unlike other influential factors such as alcoholism, previous lesions, smoking or genetic issues, age appears to be especially important in PDAC. Every cancer has a strong relation to age due to several cellular processes, as is the case of senescence, but for PDAC, this relation appears to be more remarkable than other cancers. In fact, 85% of pancreatic cancer cases involve patients older than 65-years old with a diagnosis mean age of 73-years old, Koorstra et al. (2008). For that reason, this research introduces the age factor for further analysis of its influence in cancer patients. Hence, the goal of our proposal with the current case study is to identify age-related gene subsets, which may influence the severity of the disease. In that sense, such genes apart of being age-related, should also be able to capture the greatest variations of their expression levels (relation qualitative + quantitative).

Finally, to reach all goals proposed in this research, the remaining sections of this paper have been divided as follows: Section 2 describes works related to the feature selection process and our proposal. Section 3 develops the framework for gene selection and explains each of its components as their interactions. Section 4 describes the dataset to be used, experiments, results and discussion after applying an implementation of the introduced framework. Section 5 presents the conclusions of this paper whereas Appendix outlines a set of visualizations supporting the results. References used in this research have been given as the final part of this paper.

2. Related work

Feature selection (FS) can be generically defined as the process of extracting feature or gene subsets whose expression level values are representative of a particular target feature, i.e., clinical or biological annotation (Inza et al., 2004; Jager et al., 2003; Kumari and Swarnkar, 2011; Lazar et al., 2012). FS is a very active research area in the analysis of *gene expression microarray*, which is contributing to the development of the field as a result of involved data mining and machine learning techniques, TunedIT (2008). Particularly, FS from microarrays is addressed to identify/discover those genes which are expressed differentially according to a determined target disease (namely, *informative genes*). As previously stated in the introduction, there is a large number of approaches in the literature dealing with this issue and with potential application in the area of disease prediction and discovery, gene regulatory network reconstruction, pharmaceutical industry, among others (Golub et al., 1999; Penfold and Wild, 2011). However, the many challenges posed by this research field require new approaches.

Due to the wide range of papers proposed to face the FS problem in microarrays and facilitate the study of the area, FS methods have been divided into the following four categories: *filters, wrappers, embedded* and *ensemble*. Filter methods have been directed to discriminate or filter features/genes based on the intrinsic properties of the dataset by estimating their relevance scores to state a cut-off schema where an upper/lower bound is imposed in order to choose features with the best scores. According to Guyon (2003) and Lazar et al. (2012), this scheme could favor gene identification to be targeted pharmaceutically. Wrapper methods use a classifier to find the most discriminant feature subset by minimizing an error prediction function (Ambroise and McLachlan, 2002; Díaz-Uriarte and Alvarez, 2006; Ruiz et al., 2006; Yee et al., 2005; Zhou and Tuck, 2007). These methods tend to consume a lot of runtime and their results depend on the type of used classifier. Embedded methods are similar to wrapper, but allow the learning method to interact, which reduces the runtime taken by wrapper methods (Efron et al., 2004; Hernandez et al., 2007; Lazar et al., 2012; Quinlan, 1994; Saeys et al., 2007). Ensemble methods are relatively new and recombine results from different FS techniques to achieve a more stable feature subset, since small perturbations in the training set can have effects on the results of a FS method applied individually (Haury et al., 2011; Moorthy and Saberi, 2012; Nguyen et al., 2015). Therefore, ensemble methods come to face the instability difficulty presented by some of the approaches previously explained.

Since the FS methodology followed by the proposed framework is based on a filtering strategy to successively reduce a dataset until the target gene subset has been achieved, we are going to focus our attention on some of the main features presented by filter techniques. This will allow us to highlight two trends followed by filter methods. The first type refers to methods selecting the top ranking features, which are based on the relevance value assignment to each feature/gene (*ranking methods*). The relevance value estimation is carried out by a scoring function preselected according to the pursued target (Jaeger et al., 2003; Liu et al., 2005; Peddada et al., 2003; Yang et al., 2006). The second type of trend includes *space search methods*, which are engaged to optimize an objective function by generally involving maximum relevance and minimum noise for the found gene subsets (Ding and Peng, 2005; Mohamed et al., 2015; Wang et al., 2005; Xing et al., 2001). According to the classifications above, those of Lazar et al. (2012) have stated a taxonomy for FS methods as follows: Raking methods can be classified as either *univariate* or *bivariate* (Deng et al., 2004; Long et al., 2001; Thomas et al., 2001; Tusher et al., 2001). Univariate methods