A boundary-point approach applied to gene selection in gene expression data

Juan Ramos
IBSAL/BISITE Research Group
University of Salamanca
Salamanca, Spain
juanrg@usal.es

José A. Castellanos-Garzón
IBSAL/BISITE Research Group
University of Salamanca
Salamanca, Spain
jantonio@usal.es

Juan F. de Paz
IBSAL/BISITE Research Group
University of Salamanca
Salamanca, Spain
fcofds@usal.es

Juan M. Corchado
BISITE Research Group
University of Salamanca, Osaka Institute of Technology
Salamanca, Spain
corchado@usal.es

***Abstract*—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 work *A data mining framework based on boundary-points for gene selection from DNA-microarrays: Pancreatic Ductal Adenocarcinoma as a case study*, we propose a novel hybrid framework based on data mining techniques applied to the problem of meaningful gene selection and the search for new biomarkers. For this purpose, the framework deals with approaches such as statistical significance tests, cluster analysis, evolutionary computation, visual analytics and boundary points. The latter is the core technique of the proposal, which allows us to define two alternative methods of gene selection. Moreover, the framework has added a variable to the study (as the age), which is studied with respect to gene expression levels.**

***Index Terms*—Feature selection, Gene selection, Data mining, Cluster analysis, Genetic algorithm, Boundary point. DNA-microarray.**

I. Introduction

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 and adapt to different situations and problems.

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 [1]–[3]. 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. Finally, we want to stress that this research has been published in [4].

The research of Juan Ramos González has been co-financed by the European Social Fund and Junta de Castilla y León (Operational Programme 2014-2020 for Castilla y León, BOCYL EDU/602/2016). This work has also been supported by project MOVIURBAN: Máquina social para la gestón sostenible de ciudades inteligentes: movilidad urbana, datos abiertos, sensores moviles. SA070U 16. Project cofinanced with Junta Castilla y León, Consejería de Educación and FEDER funds.

II. Hybrid framework

Since HybridFrame is based on data mining techniques, we have focused our efforts on the combination of areas such as evolutionary computation, visual analytics, and cluster analysis, among others to develop a methodology to follow in the domain of gene expression data, Figure 1.

Statistical filtering module (SFM) This module is responsible for a preliminary data processing and the first gene filtering processes based on statistical significance. Thus, the first process in this module consists of a data treatment by removing control probes, standardizing, and applying algorithms of missing data treatment if needed.

The first applied filter method is the Mann-Whitney test [5]. Then a second filter method is selected in relation to user goals. In this case, the module implemented five filter methods, although new methods can be added. In our case study we used Kruskal-Wallis, which can be used by when introducing a variable measurement external to the dataset to filter out genes related to the variable of interest.

Hierarchical clustering method module (HCMM) The dataset resulting from the module above is divided into subsets (clusters) in order to move the complex gene selection task from the whole current dataset to smaller gene subsets. The idea consists of applying data clustering methods to divide the complex task of gene selection from a big dataset into small subsets (divide and conquer strategy), identified by their gene similarity. Although this module does not really perform a gene filtering task, it partitions the data for the following stages.

Visual analytics module (VAM) This module selects the most suitable clustering from each input dendrogram. Internal

measures of cluster validity (such as homogeneity, separation and silhouette width) are applied to input dendrograms to estimate level ranges with high quality clusterings. [6], [7], which are applied to each level of a dendrogram to select the one with the best score. Then consists of choosing and visually validating a level from each level interval computed in the process above. For this propose, each dendrogram is explored from its level interval through a linked visualization set, supporting heatmaps, dendrograms, parallel coordinates, 3D-scatterplots and boundary gene visualizations.

Clustering boundary module (CBM) This module carries out a filtering process by extracting out the boundary genes for each cluster given from input clusterings to the module. The boundary point algorithm used for this purpose is focused on the ClusterBoundary algorithm given in [8].

The final stage of the framework consists in two alternative selection methods:

- **Clustering intersection method (CIM)**: the CIM method is based on the idea of boundary intersections coming from different clustering methods. We assume that boundary genes achieved from the intersection of different clustering boundaries coming from different methods, which develop different cluster strategies on data, are the main candidates to be informative genes.
- **Evolutionary hierarchical clustering method (ECM)**: The second method leading to discover informative genes in this framework is ECM, as shown in Fig 1. This method is based on the evolutionary model for clustering (EMHC) given in [9], [10]. We propose that, since dendrograms given as ECM solutions inherit, alter, recombine and even improve part of the genetic code (high quality clusters) of good solutions given by others methods, then it is expected that genes located on the boundary of such clusters are strong candidates to be informative genes.

III. Case study on pancreatic ductal adenocarcinoma

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 [11], [12]. Although every cancer has a strong relation to age due to several cellular processes, 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 [13]. For that reason, this research introduces the age factor for further analysis of its influence in cancer patients.

After applying the methodology above, two sets of genes were obtained, one for each filter method, identifying informative genes of each method. Information about each concrete gene has previously been identified in other research and/or databases as a PDAC related gene. Information provided by both tables was consulted in PED

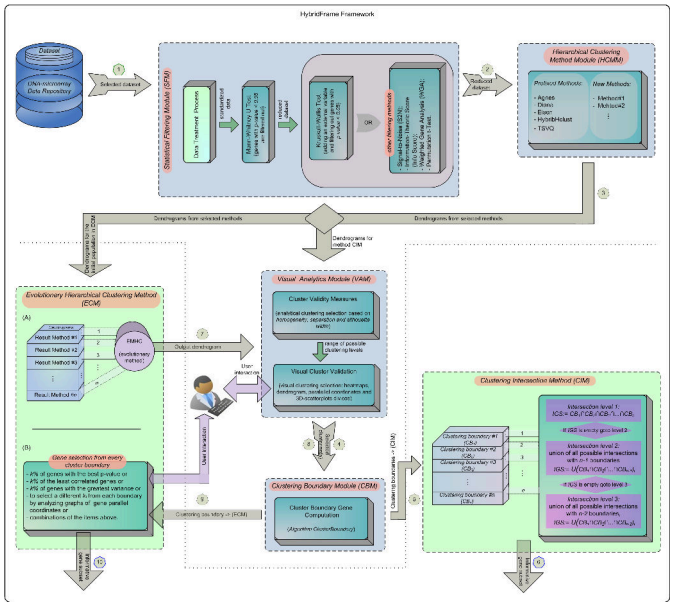

Fig. 1. Chart representing the data mining framework HybridFrame for gene selection.

(http://www.pancreasexpression.org/) and Pancreatic Cancer Database (http://pancreaticcancerdatabase.org/index.php). Selected genes have a larger relation to normal and tumor tissue samples of PDAC and are highly age-related. Moreover, 10 genes from these tables were identified by both methods (genes in the intersection are highlighted in both tables), meaning they could be even more meaningful for PDAC than the rest. In summary, according to the whole discovery process of informative genes given by Hybridframe, we assume that selected genes can be considered for further pharmaceutic research.

Acknowledgment

We would like to thank Dr. Liviu Badea from Bioinformatics research group, National Institute for Research in Informatics (Rumania) for provided additional information on the dataset used in this research.

References

[1] I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] J. Jager, R. Sengupta, and W. Ruzzo, "Improved gene selection for classification of microarrays," in *Pacific Symposium on Biocomputing (UW CSE Computational Biology Group), PMID: 12603017*, 2003.

[3] C. Lazar, J. Taminau, D. Meganck, S.and Steenhoff, A. Coletta, V. Molter, C.and deSchaetzen, H. Duque, R.and Bersini, and A. Nowé, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, vol. 9, no. 4, pp. 1106–1118, 2012.

[4] J. Ramos, J. A. Castellanos-Garzón, J. F. de Paz, and J. Corchado, "A data mining framework based on boundary-points for gene selectionfrom DNA-microarrays: Pancreatic Ductal Adenocarcinoma as a case study," *Engineering Applications of Artificial Intelligence, Elsevier*, vol. 70, pp. 92–108, 2018.

[5] P. Weiss, "Applications of generating functions in nonparametric tests," *The Mathematica Journal*, vol. 9, no. 4, pp. 803–823, 2005.

[6] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.

[7] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data. An Introduction to Clustering Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2005.

[8] J. Castellanos-Garzón, C. García, P. Novais, and F. Díaz, "A visual analytics framework for cluster analysis of DNA microarray data," *Expert Systems with Applications, Elsevier*, vol. 40, pp. 758–774, 2013.

[9] J. Castellanos-Garzón, "Evolutionary framework for DNA microarray cluster analysis," Ph.D. dissertation, Department of Computer Science, University School of Computer Science, University of Valladolid, 2012.

[10] J. A. Castellanos-Garzón and F. Díaz, "An evolutionary computational model applied to cluster analysis of DNA microarray data," *Expert Systems with Applications, Elsevier*, vol. 40, pp. 2575–2591, 2013.

[11] L. Badea, V. Herlea, S. Olimpia, T. Dumitrascu, and I. Popescu, *Combined Analysis of Whole-Tissue and Microdissected PDAC*, Bioinformatics group, National Institute for Research in Informatics, Bucharest 011455, Romania, 2008.

[12] ——, "Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia," *Hepato-Gastroenterology*, vol. 88, pp. 2015–2026, 2008.

[13] J. Koorstra, S. Hustinx, G. Offerhaus, and A. Maitra, "Pancreatic carcinogenesis," *Pancreatology*, vol. 8, no. 2, pp. 110–125, 2008.