Hybrid Feature Selection Algorithm for Gene Selection and Gene Importance Evaluation in Microarray Data

Carlos Fernandes, Francisco Souza, and Rui Araújo
Institute for Systems and Robotics (ISR-UC), and
Department of Electrical and Computer Engineering (DEEC-UC),
University of Coimbra, Pólo II, P-3030-290 Coimbra, Portugal
Emails: {cfernandes,fasouza,rui}@isr.uc.pt

Abstract
The DNA microarray for gene sequencing allows the screen of thousands of genes simultaneously and with the advent of the next generation sequencing (NGS) technology the number of genes available for analysis are much larger than before, where in NGS the generated data can get into the order of terabytes (Tb). In this context, the gene selection and gene importance evaluation are important tools, since it allows the Bioinformatic researchers to focus on promising gene candidates that actively contribute to some disease or anomaly. However, the gene selection and gene importance evaluation are problematic from the mathematic point of view, since the number of gene/features is much larger than the number of samples/users, making it a problem with many available solutions. In literature, there are many available algorithms proposed to solve the problem of gene selection and gene importance evaluation, where some of them are based on statistical or machine learning approaches [1], [2], [3]; they are also called as filter (statistical approaches) and wrapper approach (machine learning approaches). This work proposes a hybrid approach, which explores both of the techniques, the filter and wrapper, using the mutual information statistics and several machine learning algorithms, such as the Naive Bayes, Support Vector Machines, Classification Trees and k-Nearest-Neighbor classifiers. To access the importance of each feature, the proposed method is going to be applied in several boostrapped version of data and the genes are going to be ranked according to their frequency of being selected. The proposed approach is going to be applied in public benchmark datasets and compared with current state of art algorithms.

Keywords: Microarray gene selection, Gene importance, Feature selection.

ACKNOWLEDGMENT
Francisco A. A. Souza has been supported by Fundação para a Ciência e a Tecnologia (FCT) under grant SFRH/BPD/112774/2015.

REFERENCES