A new supervised approach to feature selection in microarray datasets

Mariane S. Machado[1], **Samir Merode**[1], **Marcus Ritt**[1], **Luciana Buriol**[1], **Pablo Moscato**[2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]Centre for Bioinformatics, Biomarker Discovery and Information-based Medicine,
Hunter Medical Research Institute,
The University of Newcastle, Callaghan, 2308, NSW, Australia,
and ARC Centre of Excellence in Bioinformatics

{mmachado,merode,mrpritt,buriol}@inf.ufrgs.br, Pablo.Moscato@newcastle.edu

Abstract. *A typical microarray experiment yields the expression level of a large number of genes for a small number of samples. Given a classification of the samples, the goal of feature selection is to identify a small subset of relevant genes, which are differentialy expressed for different sample classes. We present a new method for feature selection that combines a solution for the Min (α,β)-Feature Set Problem and a clustering algorithm, the Arithmetic-Harmonic Cut to robustly identify relevant features. We apply our method to the NCI60 cancer dataset and evaluate the effectiveness and performance of the new algorithm for the classification of cancer cell-lines.*

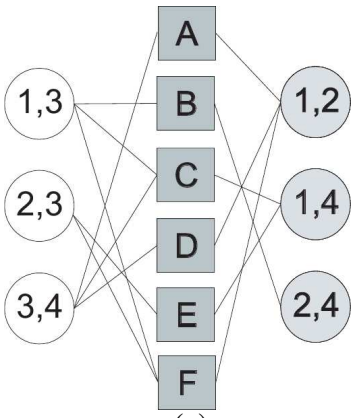
1. Introduction

Microarrays are useful for obtaining the expression levels of a great amount of genes simultaneously. Selecting correlated genes from the output of a microarray experiment according to a selected phenotype can help scientists to better understand how to cure or prevent diseases. However, reaching this selection is not a trivial task due to the small number of samples and the huge quantity of genes involved in these experiments. This task can be seen as a feature selection problem since we want to select a subset of elements that better classifies the samples into two or more categories according to some feature. In bioinformatics, usually, this subset are genes, the samples are people and we want to find out which genes are better to show which samples present some special characteristic (feature) such as a disease. The Min (α,β) Feature Set Problem [Cotta et al. 2006] is a special case of feature selection problem. In [Cotta et al. 2006] an implementation using integer programming presented good results even with large datasets. The drawback is that this model does not consider correlations between gene expression levels. We propose a hybrid approach, first proposed in [Machado 2008] and later improved in [Merode 2008] which combines a solution for the Min (α,β) Feature Set problem, and a (bi-)clustering technique, called the Arithmetic-Harmonic Cut [Mahata et al. 2006]. The Arithmetic-Harmonic Cut ranks selected features according to their correlation, and therefore the combined method is able to select a small feature set better than using only a algorithm that solves the Min (α,β)-Feature Set. The rest of this paper is organized as follows: Sections 2 explains the Min (α,β)-Feature Set Problem. Section 3 presents the Arithmetic-Harmonic Cut clustering procedure. Section 4 presents our combined approach to feature

selection based on these techniques. In Section 5 we validate our approach experimentally on the NCI60 cancer dataset. Final remarks are made in Section 6.

2. Min (α,β)-Feature Set Problem

We illustrustrate the feature selection problem on the example of a discretized gene expression level matrix given in 2 (b) and represented as the graph in 2 (a), where each pair of samples of different classes is shown as a circular node on the left side, each gene as a squared node in the middle and each pair of samples of the same class as a circular gray node on the right side. We draw an edge between a circular node and a gene, if the gene is differentially expressed for this pair of samples. Moreover, we draw an edge between a circular gray node and a gene if this gene is equally expressed for this pair of samples.



(a)

	Sample			
	1	2	3	4
Gene A	High	High	High	Low
Gene B	Low	High	High	High
Gene C	Low	High	High	Low
Gene D	High	High	High	Low
Gene E	High	Low	High	High
Gene F	High	High	Low	Low
Phenotype	Sick	Sick	Healthy	Sick

(b)

Figure 1. (a) is a graph modeling the expression levels in (b)

In the Min (α,β)-Feature Set, it is required that each pair of genes from different classes is adjacent to at least a number α of genes. Analogously, it is required that each pair of genes from same class must be adjacent to at least a number of β genes. Therefore, each difference is witnessed by a number of α differentially expressed genes and each similarity is witnessed by a number of β genes. In other words the Min (α,β)-Feature Set Problem is the problem of finding the minimum number of genes that satisfies α and β parameters. In our example, notice that for a value of $\alpha = 1$ and $\beta = 1$, one possible solution is the set of nodes B, C and F since they would be enough to cover all pairs of samples. It easy to see that α and β parameters will be higher as the number of selected genes increases. Thus, these α and β parameters must be chosen criteriously. If the parameters are too high, too many genes are selected and in case they are small, the results may be not reliable.

3. Arithmetic-Harmonic Cut

Arithmetic-Harmonic Cut is a clustering technique proposed in [Mahata et al. 2006]. In this work, the authors pose the clustering problem as a graph optimization problem and propose a novel objective function which performs very well in diverse types of datasets. An Arithmetic-Harmonic cut of an undirected, weighted graph is a bi-partition of the vertex set which maximizes the product of the sum of the weights of the edges between

the partitions, and the sum of the reciprocals of the weights of the edges inside them. The arithmetic-harmonic cut maximizes

$$F = \left(\sum_{e \in E_{out}} \omega(e) \right) \left(\sum_{e \in E_{in}} \frac{1}{\omega(e)} \right), \quad (1)$$

where $\omega(e)$ is the weight of an edge e. An edge e pertains to E_{in} when it links two nodes that are in the same partition. In the same way, an edge e pertains to E_{out} when it links two nodes that are in different partitions. Maximizing F tends to keep edges with small weights inside the clusters, and edges with high weights between them, since its sums the reciprocals of intra-cluster and multiplies this by the sum of the weight of inter-cluster edges As observed in [Mahata et al. 2006], this objective function can be rewritten as

$$F = \frac{A_{out}}{H_{in}} |E_{in}||E_{out}|, \quad (2)$$

where A_{out} corresponds to the arithmetic mean of weights of edges that connect vertices from different clusters, H_{in} is the harmonic mean of weights of edges that connect vertices from the same cluster. This rearranged function shows more explicitly that we are using the harmonic mean, which is more robust and less sensitive to the presence of outliers. It is a NP-hard problem [Rizzi and Moscato] and the recursive application of such cuts generates a tree-based classification of the data.

4. The new approach

This new approach was suggested in [Merode 2008], and is an improvement to the approach shown in [Machado 2008]. Although solutions for the Min (α,β)-Feature Set problem make good choices of genes, this method can still be more robust. To achieve it, we need another measure that helps to increase the reliability of gene selection. One possible measure is the correlation between genes. Since the Arithmetic-Harmonic Cut value of each solution is based on correlations, it is interesting to combine it with approximated solutions for the Min (α,β)-Feature Set Problem. Because both problems are NP-complete, we use an evolutionary algorithm defined as follows:

- Each solution is codified as vector of bits, where each bit corresponds to a gene If the bit is on, it is a selected gene, otherwise non selected.
- Tournament of three solutions is used as selection criterium.
- Mutation flips 10% of the bits in 0.7% of the total population of genes at each generation.
- Crossover is defined as follows: if both parents have same bit value, this value is inherited by the child. Otherwise the new solution has 60% of chance of inheriting the bit from the most capable parent.

We maximize the following objective function:

$$F \ \varphi(\alpha + \beta/2) \ G, \quad (3)$$

where F is the Arithmetic-Harmonic Cut as defined in equation 1 applied just to the selected genes in the following way: one partition is the set of genes that presents a feature

and the other partition genes that does not. We use as correlation a dissimilarity measure calculated as the the normalized Euclidean distance among the expression levels of the genes. Observe that the method does not depend on the this particular measure, and we could use Pearson or Spearman correlation based measure as well. α and β are the same used in the solutions for the Min (α,β)-Feature Set Problem. As can be seen, we give less value to β, since our main goal is to discriminate samples. $\varphi(x)$ is an auxiliary function defined as:

$$\varphi(x) = \begin{cases} 0.25 & \text{if x is equal to zero} \\ x & \text{otherwise} \end{cases} .$$

It is used to avoid F to be zero when $\alpha + \beta/2 = 0$, permitting that the solutions can still be evaluated by the rest of the criteria. We chose 0.25 in the $\varphi(x)$ function because it is half of the smallest non-zero value that can occur in the expression $\alpha + \beta/2$, which appears when $\alpha = 0$ and $\beta = 1$. G is the mean of the GSRobust for the selected genes as proposed in [Zheng et al. 2006]. The GSRobust of a gene i is a function defined in the following way:

$$G_i = \frac{MAD[median(g_{i1}), ..., median(g_{in})]}{\sum_{j=1}^{n} MAD(g_{ij})}, \tag{4}$$

where G_{ij} is the vector of expression levels of gene i in class j, n is the number of classes and $MAD(x)$ is

$$MAD(x) := 1/n \sum_{1 \leq i \leq n} |x_i - \mu(x)|. \tag{5}$$

G helps to limit the number of selected genes, since it increases when less genes are selected. It was also implemented a best improvement local search where the neighborhood is composed of all solutions that is different just by one bit. However,this local search is applied in the current solution when occurs 10 generations without any improvement. The stop criterium is reached when no enhancement happens after the application of the local search. It is used three constructive heuristics in order to generate a better initial population. In the first one, it is applied the GSRobust in each gene, then it is selected just the ones that satisfies $G_i > G + 3\sigma$ where σ is the standard deviation of G_i. The second one is to rank the genes by the number of samples it can discriminate. Then the genes are chosen to go to the initial population until a percentage of the maximum α is reached. The same procedure is done for the β value. The third constructive heuristic selects all as eligible for the initial population all the genes that satisfy $\bar{d}_{out} > n\bar{d}_{in}$ where $\bar{d}_{out}$ is the mean absolute difference among the expression levels of the gene in the samples of different class, and $\bar{d}_{in}$ is the mean absolute difference among the levels of expression in the samples of same class. n is a constant set as two since it show better results in tests. The initial population is composed of 500 possible solutions. One half of the population is generated from the constructive heuristics mentioned and the rest is generated randomly, but the number of selected genes is equal or less to the mean of selected genes in the population generated using the constructive heuristics.

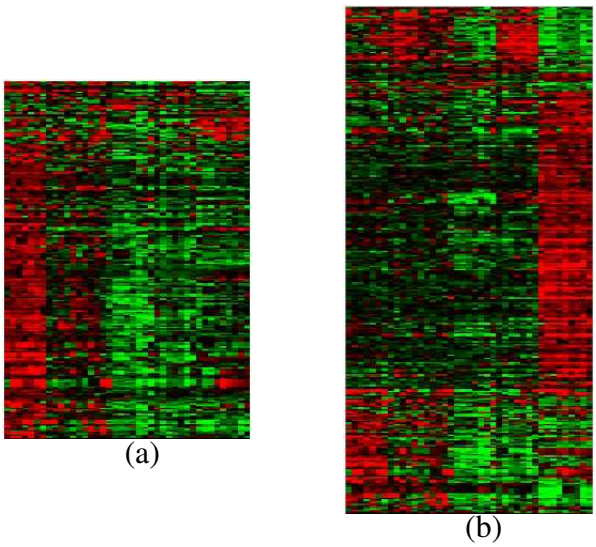
5. Results

[Ross et al. 2000] introduced an important dataset for the molecular classification of different types of cancer. We are working with a dataset available on the authors web-

site supplement known as NCI60. We have first completed all missing values for the NCI60 dataset using the `LSImpute` algorithm recently introduced in [Bø et al. 2004]. After that, we applied our new approach on the same five different types of cancer cells that were used in [Berretta et al. 2005]: Renal (RE), Ovarian-like (OV), Leukemia (LE), Colon (CO) and Melanoma (ML). We compare each type of cancer cell against all others, i.e. we choose one type of cancer cells and try to select the genes that discriminate well this type of cancer cells. For example, if we want to find the group of genes that best discriminates Leukemia cells, we put all the samples containing Leukemia cells in one group, and all other samples in the other. After selecting the relevant subset of genes, we generate heatmaps. Bright red color means highly expressed genes and bright green means under expressed genes. Darker colors mean no difference in the level of expression.The obtained results were very satisfactory. The evolutionary algorithm could improve the already good initial solutions, according both to the evaluation function and heatmap. Table 1 below shows some further information about the results.

	Generation	time(s)	α(max possible)	β(max possible)	Selected Genes
Renal	612	15269.44	18(24)	10(26)	238
Ovarian	675	19142.20	11(16)	12(27)	225
Leukemia	700	26048.27	38(49)	12(26)	336
Colon	659	18592.66	15(16)	8(26)	304
Melanoma	411	11827.10	38(53)	10(26)	336

One may notice that the α and β values obtained by the evolutionary algorithm were not the maximum. This was expected since the algorithm is a tradeoff between correlation and high α and β values. Below there are the heatmaps for Renal Cancer and Melanoma. It can be easily seen in the first heatmap that the initial columns, which are the samples affected by Renal Cancer stand out from the rest of the data. Analogously occurs in the Melanoma heatmap, where the affected samples are the last five columns. Heatmaps for other types of cancer are available at `www.inf.ufrgs.br\~mmachado\heatmaps`.



(a) (b)

Figure 2. Heatmaps for Renal Cancer(a) and Melanoma(b)

6. Conclusions

In this work we presented a new evolutionary algorithm for selection of genes from microarray data. This algorithm is based on the Min (α,β)-Feature Set Problem and the Arithmetic-Harmonic Cut. Since microarray data has outliers, it is important to use approaches that give good selections in order to permit scientists to more accurately find the connection between a disease such as cancer and the respective gene expression profile. As the results show, the intuition that if we optimize for a better clustering between the selected and not selected genes, even accepting a reduced number of genes discriminating or confirming differences indeed, give better results In other words, if we permit α and β to be lower than the maximum possible but also considering the Arithmetic-Harmonic Cut, we may find more robust solutions that yet discriminate well samples.

References

Berretta, R., Mendes, A., and Moscato, P. (2005). Integer programming models and algorithms for molecular classification of cancer from microarray data. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science*, volume 38 of *ACM International Conference Proceeding Series*, Newcastle, Australia. Australian Computer Society.

Bø, T. H., Dysvik, B., and Jonassen, I. (2004). Lsimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*, 32(3).

Cotta, C., Sloper, C., and Moscato, P. (2006). Evolutionary search of thresholds for robust feature set selection: application to the analysis of microarray data. *Applications of Evolutionary Computing, G. Raidl et al. (eds.), Lecture Notes in Computer Science*.

Machado, M. S. (2008). Um algoritmo memético para seleção de genes a partir de dados provenientes de microarray. Trabalho de graduação em ciência da computação, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Mahata, P., Costa, W., Cotta, C., and Moscato, P. (2006). Hierarchical clustering, languages and cancer. *Applications of Evolutionary Computing, G. Raidl et al.(eds.), Lecture Notes in Computer Science*.

Merode, S. G. Z. (2008). Um novo algoritmo genético para seleção de genes. Trabalho de graduação em ciência da computação, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Rizzi, R. and Moscato, P. Personal communication.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., de Rijn, M. V., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*.

Zheng, G., Olusegun, G. E., and Narasimhan, G. (2006). Microarray data analysis using neural network classifiers and gene selection methods. *Methods of Microarray Data Analysis IV*, pages 207–222.