Consensus Gene Selection on DNA Microarrays

Rubén Armañanzas, Iñaki Inza, Pedro Larrañaga

Department of Computer Science and Artificial Intelligence
University of the Basque Country - P.O. Box 649
20080 Donostia - San Sebastián, Spain

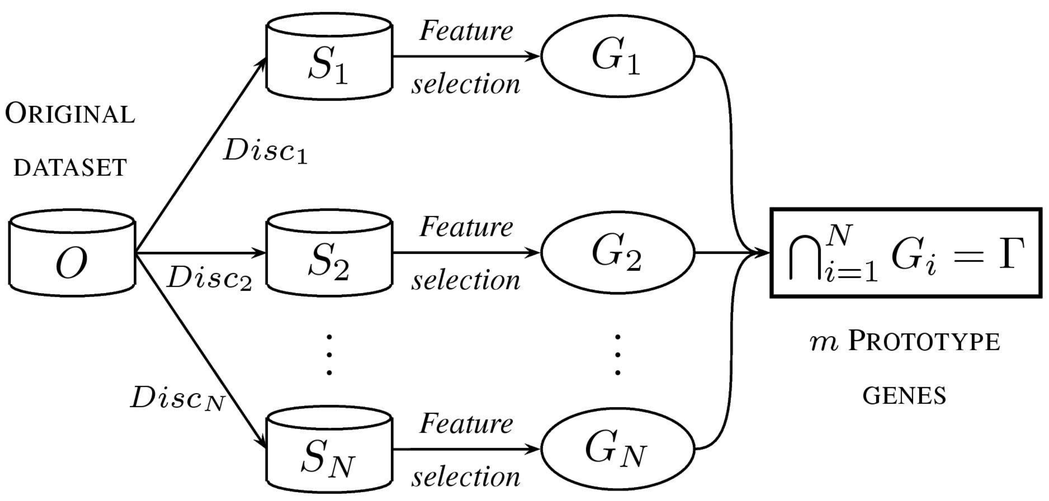
ISG - Intelligent Systems Group
http://www.sc.ehu.es/isg
{ruben,inza,ccplamup}@si.ehu.es

1 Introduction

- Microarray experiments involve noise arising due to measure, physical and stochastic processes. In addition to these problems, several analysis methods add biases due to their inherent procedures.
- From a machine learning point of view, the expression level of a gene is represented as a *random variable* of a probabilistic process.
- In order to overcome these problems, we present a consensus approach to microarray gene selection. This consensus procedure combines the best techniques from each field: a set of discretization policies, a filter-like selection procedure and statistical coexpression measures.

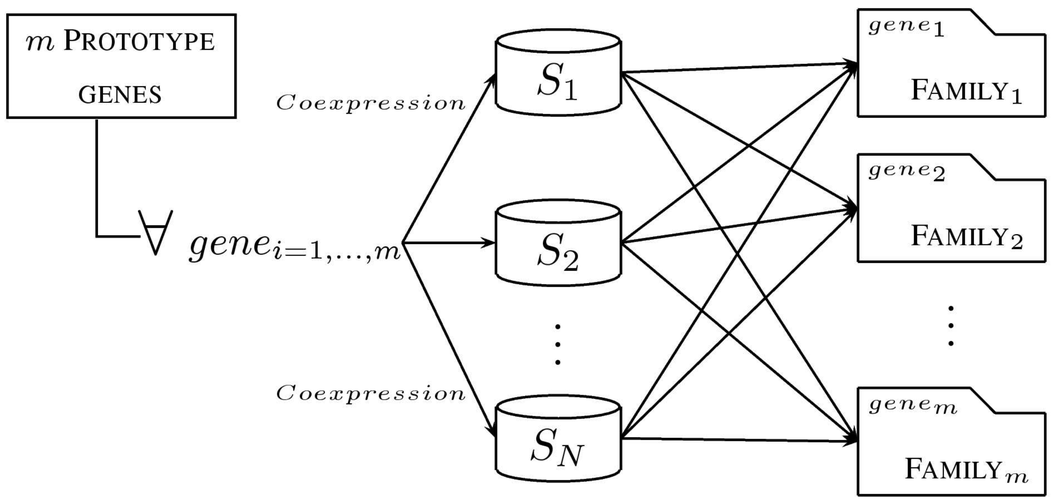
2 Approach

The search for a robust solution makes us not to rely on a single discretization method. From the original continuous-value data, differentially discretized data sets are computed, trying to diminish the possible added bias.



First step: identifying the prototype genes.

Let O be the original microarray data set with continuous features and $S_1, \ldots, S_N$ the results of N different discretizations of the O set. Using a filter subset selection method, N different feature selections are performed in basis of the $S_1, \ldots, S_N$ discrete datasets, producing the following subsets of genes: $G_1, \ldots, G_N$. The final consensus gene subset Γ is the intersection of all of them, that is $\Gamma = \cap_{i=1}^{N} G_i$, with $|\,\Gamma\,| = m \leqslant \min_{i=1,\ldots,N} |\,G_i\,|$.



Second step: identifying the genes mostly correlated with the gene prototypes found.

In the second stage, for each prototype gene a linked list of genes is constructed. For each prototype, its q more univariately correlated genes are also selected. This is performed for all the N different discretized datasets, obtaining the linked list. The aim of this second stage is to find genes with similar profile behaviours, that is, genes coexpressed within the prototype ones.

3 Discussion

The first stage of the presented proposal is tested using the Weka framework [Frank et al., 2004] and three well known microarray benchmark datasets: *Colon* [Alon et al., 1999], *Leukemia* [Golub et al., 1999] and *Lymphoma* [Alizadeh et al., 2000]. The parameters used for the first step selection, and the posterior classification validation were:

- Discretizations: equal frequency, equal width –both with three interval bins–, and entropy [Fayyad & Irani, 1993].
- Feature selection: correlation-based feature selection (CFS) [Hall & Smith, 1997].
- Classification paradigms: logistic regression, k-NN, naïve Bayes with Gaussian assumption and random forest.
- Accuracy estimation: *leaving-one-out cross validation* (LOOCV).

subset	genes	log. reg.	k-NN	n. Bayes	r. forest
Colon	1,989				
$\Gamma = \cap_3 G_i$	03	83.87	80.64	87.10	85.48
$G_{Eq.Freq.}$	22	72.58	83.87	93.55	85.48
$G_{Eq.Width}$	24	74.19	80.65	91.94	85.48
$G_{Entropy}$	40	74.19	82.26	93.55	91.94
Leukemia	1,161				
$\Gamma = \cap_3 G_i$	04	86.11	83.33	87.50	87.50
$G_{Eq.Freq.}$	28	77.78	90.28	90.28	84.72
$G_{Eq.Width}$	19	76.39	88.89	93.05	79.17
$G_{Entropy}$	48	80.55	95.83	91.67	84.72
Lymphoma	4,026				
$\Gamma = \cap_3 G_i$	16	87.50	89.60	87.50	86.46
$G_{Eq.Freq.}$	198	97.92	94.80	85.42	89.58
$G_{Eq.Width}$	125	94.79	94.80	85.42	87.50
$G_{Entropy}$	165	77.08	94.80	81.25	88.54

4 Conclusion

- The combination of different discretization policies coupled with a feature selection adds robustness to the final consensed gene sets.
- The size of the final selected gene set is highly reduced: a reduction that, analyzed by means of non-parametrical tests, does not significantly diminish the estimated classification accuracies.
- Complete gene lists and related references are available at *http://www.sc.ehu.es/ccwbayes/members/ruben/cgs/eccb05/*.
- As LOOCV is known to produce positive estimations, we envision the use of estimation techniques fitted to the microarray context [Statnikov et al., 2005].

[Alizadeh et al., 2000] Alizadeh A. A. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403, 503–511.

[Alon et al., 1999] Alon U. et al. (1999). Broad patterns of gene expression revelead by clustering analysis of tumor and normal colon tissues probed by olinucleotide arrays. *Proc. Nat. Acad. Sci. USA*, 96(12), 6745–6750.

[Fayyad & Irani, 1993] Fayyad, U. M. & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. of the 13th IJCAI* (pp. 1022–1027). Chambery.

[Frank et al., 2004] Frank E. et al. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, 20(15), 2479–2481.

[Golub et al., 1999] Golub T. R. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537.

[Hall & Smith, 1997] Hall, M. A. & Smith, L. A. (1997). Feature subset selection: A correlation based filter approach. In *Proc. of the 4th of ICONIP/ANZIIS/ANNES'97* (pp. 855–858). Dunedin.

[Statnikov et al., 2005] Statnikov A. et al. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5), 631–643.

ECCB/05 - European Conference on Computational Biology 2005. Madrid - Spain