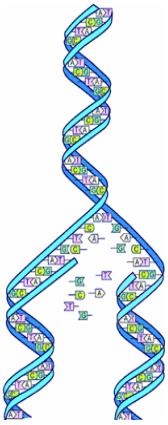
Feature selection for genomic data



Paola Cerchiello, Silvia Figini, Paolo Giudici
University of Pavia

Agenda

- Objectives
- Exploratory genes analysis
- A methodological proposal for feature selection
- Predictive diagnostic models

Objectives

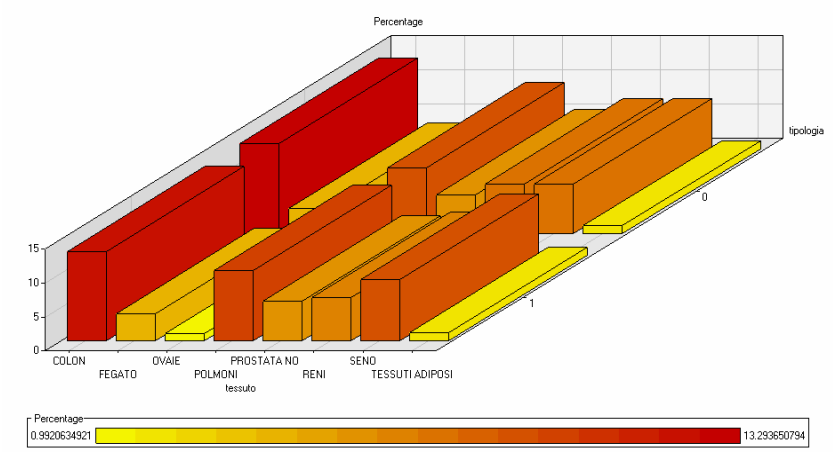
- To investigate the relationship between the gene expression and the typology healthy/sick of a specific tissue.
- To figure out the genes more relevant connected to the typology “sick”
- On the basis of the expression of relevant genes to build an efficient predictive diagnostic model

The Dataset

Contains **112.896** gene expressions, from the microarray technique, ordered in:

- **224** genes (measured on continuous scale)
- **504** observations
- **8** tissues (colon, kidney, prostate, adipose, breast, ovary, liver, lung)
- Tissue status (healthy vs sick)

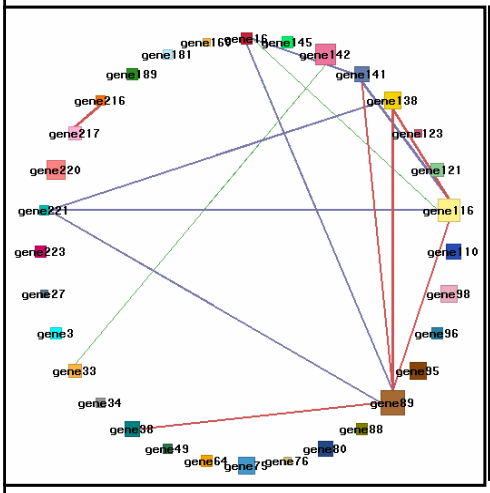
Exploratory genes analysis



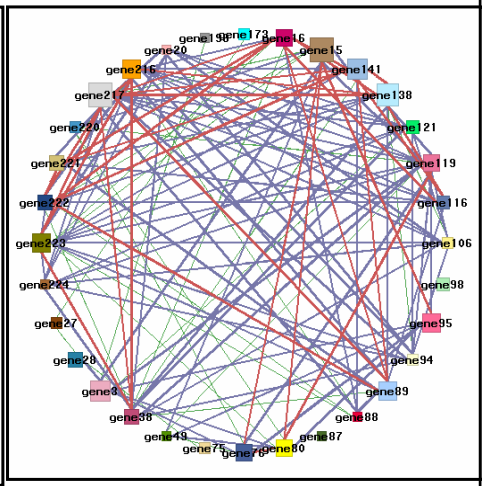
- The well distributed frequency inside every group (tissues) allows us to divide the whole dataset in 8 subset, one for every kind of tissue.
- The dataset is well distributed with

51% SICK v.s. 49% HEALTHY

Local data mining for genes analysis



Link analysis for diseased



Link analysis for not diseased

Feature selection: the employed approach

Being gene expression data typically high-dimensional, they need appropriate statistical features.

We decided to employ different approach and to compare the obtained results:

- Marker selection;
- Chi-square selection;
- Kruskal-Wallis test;
- Chaid tree.

} **Combined together**

Marker Selection

Heterogeneity measures can be extended and applied to gene expressions. As a measure of genes diversity, the entropy (*E*) can be calculated using:

$$E = -\sum_{k=1}^{m} p_i \log p_i$$

Is the probability of gene *i* being activated, and *K* the number of genes.

In order to select the most predictive genes, genes are sequentially subdivided in groups (as in a divisive cluster analysis algorithm).

If ***s*** is a subset of ***t*** we have that ***E*(*s*) < *E*(*t*) < *E***. The difference ***E(t) - E(s)*** is a good measure of how nested subsets compare in describing the data.

FS: non parametric approach

We propose to employ two statistical methods rooted in the non parametric family approach to select the more influent genes in relation to the tissue status (healthy vs sick):

- Kruskal-Wallis test;
- CHAID tree.

Kruskal-Wallis test combined with CHAID tree

K-W: Non parametric version of Anova: useful to evaluate the possible difference between the distribution of the k sample under analysis (in this context represented by the tissue status)

$$KW = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{K} n_i \left(R_i - \frac{N+1}{2}\right)^2}{1 - \frac{[\sum_{i=1}^{g} (t_i^3 - t_i)]}{(N^3 - N)}}$$

It is able to select genes presenting an heterogeneous distribution between the 2 tissue status.

CHAID: classification tree based on the well known Chi Square test. It selects genes presenting the highest chi square value with the target variable (the tissue)

Trasformation of data

The original scale of the gene expression variables is continuous

K-W test can be applied to discrete categorical variable, thereby
we recodified the variables on a categorical scale with values ranging between [-6;+5] according to the below schema:

If value < -6 ➔ assign -6
If -6 <value< -5 ➔ assign -5
If -5 <value< -4 ➔ assign -4
……..

We have also applied label '1' to malignant tissues and '0' to normal tissues (binary values of the target variable).

The results from the combination

Operatively, we first employ the K-W test **226 times** (as the number of the genes) and evaluate the associated p-value. We keep only genes with a **p-value** under a pre-fixed threshold equal to **1%.** After applying this step we obtain 108 genes rejecting the null hypothesis (homogeneous distribution along the tissue status).

On the other hand, we apply the CHAID tree to the same available dataset resulting in a few number of selected genes.

Finally we intersect the 2 resulting dataset, keeping only genes in common betweeb the K-W test and the CHAID tree.

Comparison of feature selction methods Results 1

In order to evaluate the 2 different feature selection methods we employ a prediction model, in particular classification trees. Keeping the same setting and comparing the resulting goodness of fit measures like **misclassification rate and confusion matrix**.

Frequency	*Pred Marker=0*	*Pred Marker=1*	*Pred K-WCHAID=0*	*Pred K-WCHAID=1*
Obs Marker=0	34	12	\	\
Obs Marker=1	16	39	\	\
Obs K-WCHAID=0	\	\	33	12
Obs K-WCHAID=1	\	\	17	39

Results 2

MISCLASSIFICATION ERRORS	%
MARKER SELEC	27
K-W with CHAID	28

The two proposed feature selection methods are comparable and quite similar in terms of misclassification error.

The number of selected genes are slightly different:

- 7 genes from marker feature selection approach
- 5 genes from Kruskal-Wallis CHAID selction

References

- Anderberg, M.R.: Cluster analysis for applications, New York Academic Press, (1973).
- Breiman L., Friedman J.H., Olshen R., and Stone C. J.: Classi¯cation and regression trees, Wadsworth, Belmont(1984).
- Cerchiello P., Giudici P: A non parametric method to identify unknown authors, Technical Report number 187, (2006).
- Conover W. J.: Practical nonparametric statistics, Wiley, New York(1971).
- Figini S., Giudici P.: Building predictive models for feature selection in genomic mining , Tchnical Report number 184 (2006).
- Forman G.: An Extenisve empirical study of feature selection metrics for text classi¯cation. In Journal of Machine Learning Research, 3, 1289{1306(2003).
- Giudici P.: Applied data mining, Wiley, (2003).
- L. Liu, D. M. Hawkins, S. Ghosh, Young S.: Robust Singular Value Decomposition Analysis of Microarray Data, (2000).
- Mott, R.: Marker selection, University of Oxford,(2003).
- 0. Slonim, D.K., Tamayo, P., Mesirov, J., Golub, T., and Lander, E.: Class prediction and discovery using gene expression data. Proceedings of the 4th Annual International Conference on Computational Molecular Biology, 263{272, (2000) .
- Yang, J., and Honavar, V.: Feature subset selection using a genetic algorithm. Proceedings of the Genetic Programming Conference, 380{385,(1997).
- Xing, E., Jordan, M., and Karp, R.: Feature selection for high-dimensional genomic microarray data. International Conference on Machine Learning, (2001).