Gene selection through Switched Neural Networks

Marco Muselli

Istituto di Elettronica e di Ingegneria dell’Informazione e delle Telecomunicazioni

Consiglio Nazionale delle Ricerche

Email: *Marco.Muselli@ieiit.cnr.it*

Dataset obtained through microarray experiments

The interpretation of data generated by DNA microarrays involves the extraction of information from a huge sets of real numbers.

They can refer to the expression levels of genes deriving from:

- Cells of the same kind subjected to a treatment (e.g. a drug)
- Cells of different tissues from the same organism
- Cells of different kinds (health/disease, different diseases, …)
- ....

A dataset is constituted by several experiments performed with a DNA microarray. Every experiment produces a few thousands of real values, each of which corresponds to a gene expression level (often with respect to a reference).

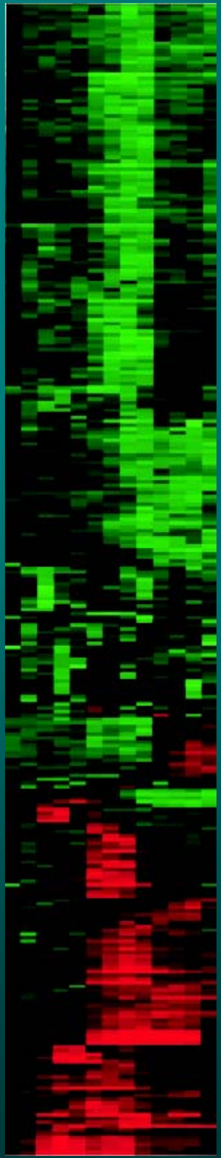
Dataset representation

The usual way to represent the dataset produced through several (*m*) DNA microarray experiments is to build a table, where each column contains the *n* gene expression levels obtained in a single experiment. Thus, every rows corresponds to a particular gene. Typically, *m* ~ 100, whereas *n* ~ 10000.

	Cell. #1	Cell. #2	…………	Cell. #*m*	Classif.
Gene #1	–214	–139	…………	17	TCA
Gene #2	–153	–73	…………	– 229	Ribo
………	………	………	…………	………	………
Gene #*n*	–37	– 14	…………	– 16	HTH
Classif.	Health	Disease	…………	Health	

A final row (column) can be added to the table, which contains a classification of the considered cell (gene).

A widely used graphical visualization



Columns ↓ Cells

Rows → Genes

Two possible analyses for the dataset

If we consider the table by rows, we obtain n points in a space with m dimensions (*gene space*). Each of them represents a specific gene in different cells or in different states of the same cell.

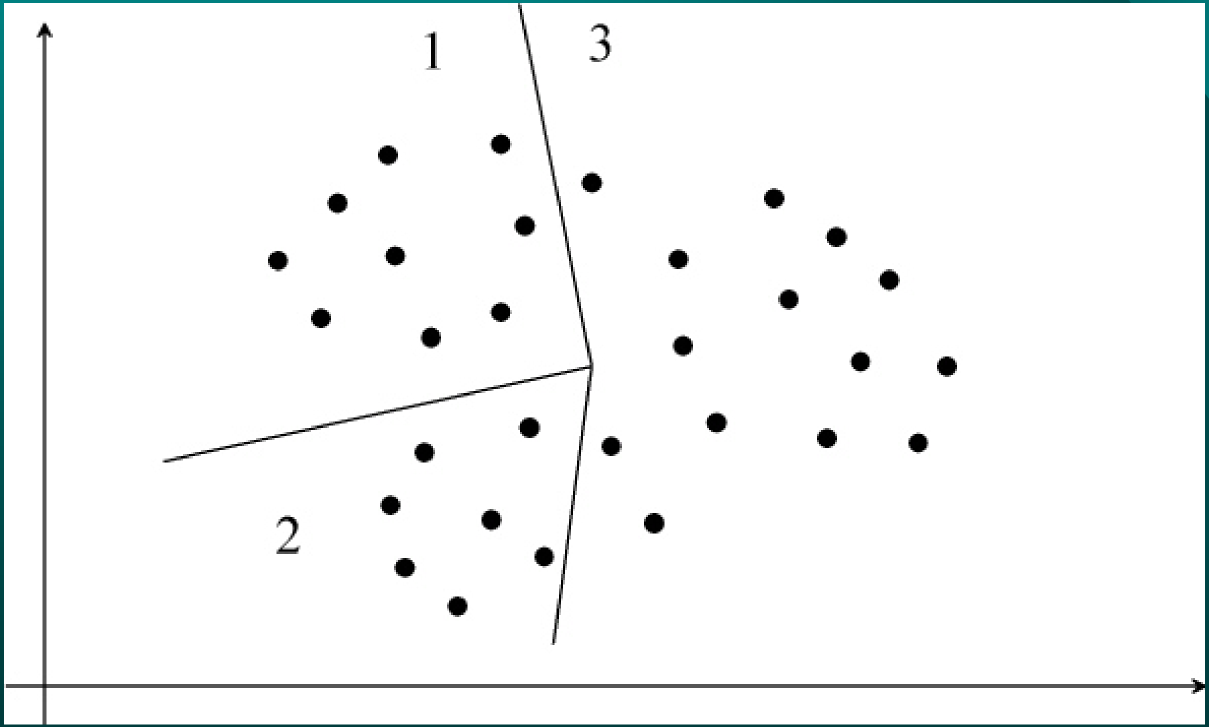
If we consider the table by columns, we obtain m points in a space with n dimensions (*cell space*). Each of them represents a given cell or the state of a cell.

Consequently, there are two possible ways of analyzing the dataset deriving from microarray experiments:

- Searching for relationships among points in the gene space (to recognize functionalities, to determine metabolic pathways, …)
- Searching for relationships among points in the cell space (to determine the genes involved in a given pathology, …)

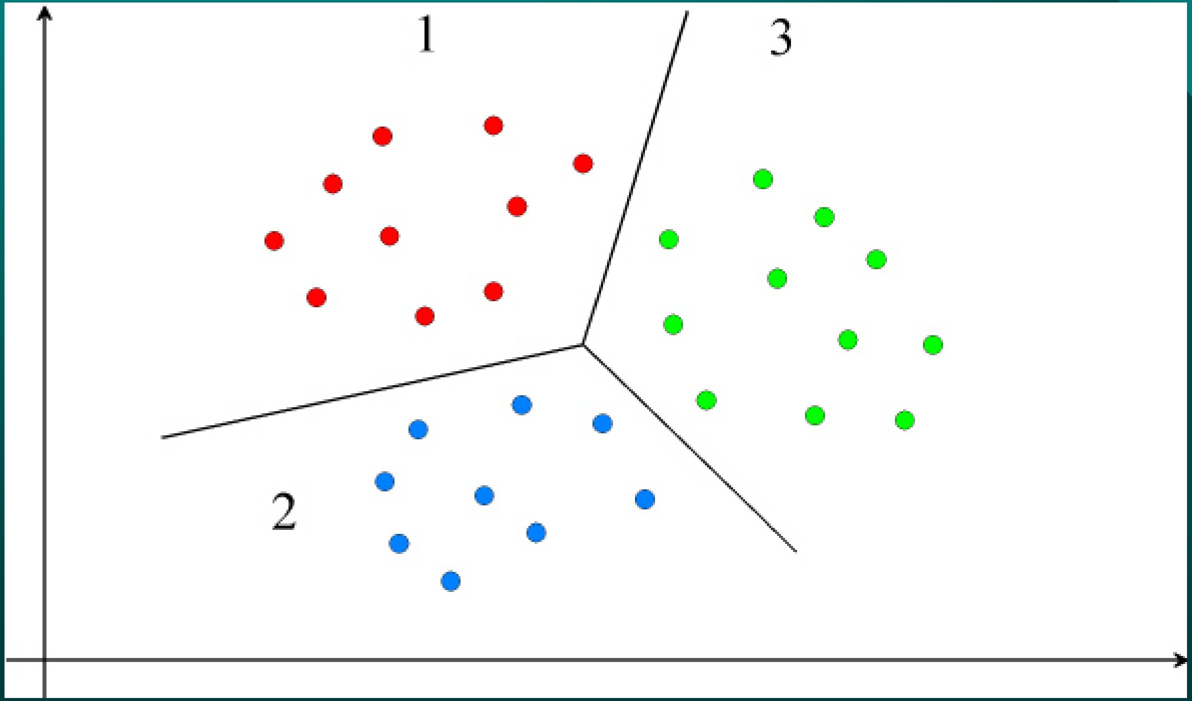
What do we intend with "searching for relationships among points"?

Essentially it is equivalent to subdivide the considered (gene or cell) space into different regions, each of which contains "similar" points according to a given characterization. In this way a classification of all the possible data is induced.



First approach to determine regions

1. Define a proper distance between the points in the space.
2. Cluster the nearest points according to the selected distance.
3. If *k* clusters are generated, subdivide the space into *k* regions. Each of them contains the points nearest to the associated cluster.



First approach to determine regions (cont.)

Problem: establish the effects of a drug, by determining the affected genes.

Method: find genes that behave in a similar way by examining microarray experiments performed at fixed time instants.

In this case it is not required an initial classification of cells or genes.

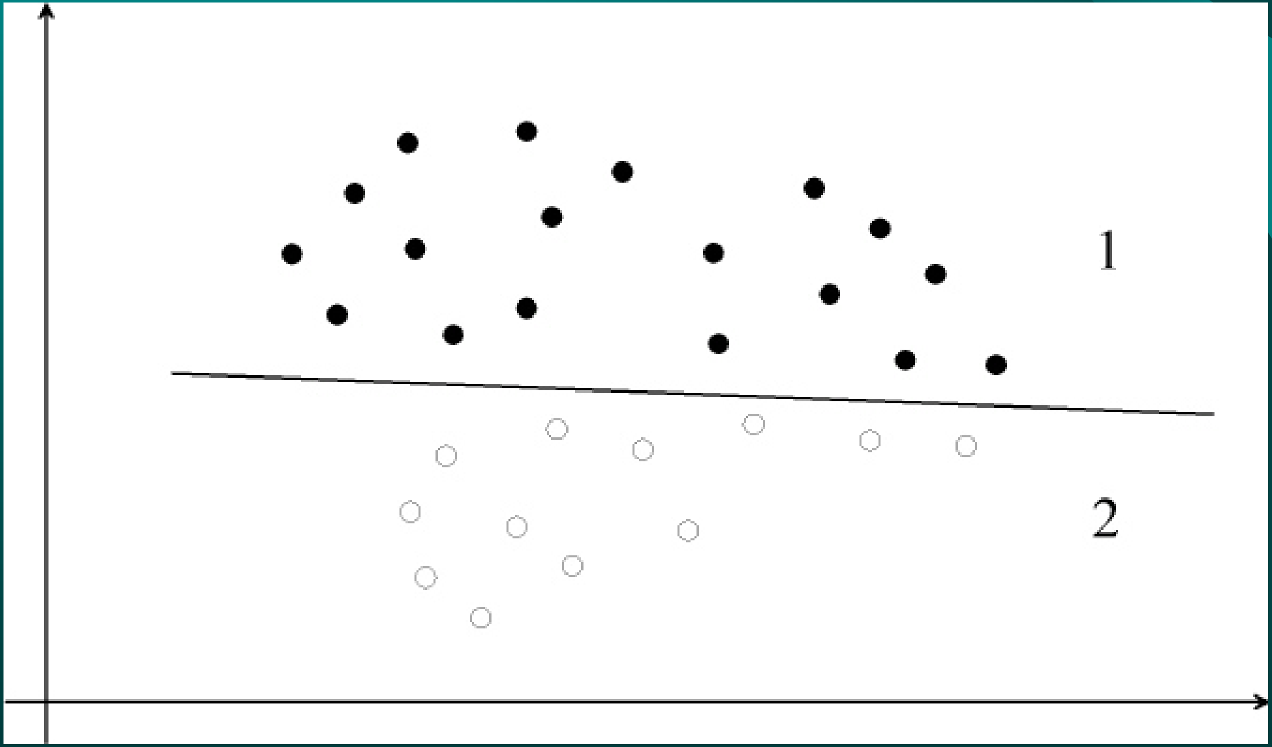
However, the choice of the distance to be adopted is critical: the resulting regions can heavily depend on this choice.

It is also difficult to decide the optimal number *k* of clusters.

The meaning of the reconstructed regions is not known; other a priori information must be used to this aim.

Second approach to determine regions

1. Consider an a priori classification of available points in *k* groups (last row/column of the table).
2. Find the lines (or the surfaces) that separate the groups of points and better respect the given classification.



Second approach to determine regions (cont.)

Problem: establish if a patient is affected by a given pathology by observing the gene expression levels of his tissues.

Method: retrieve the most probable diagnosis by observing microarray experiments for several sound and sick patients.

In this case an initial classification of the examined tissues is needed.

The meaning of the reconstructed regions is perfectly known.

The accuracy of the resulting subdivision can be easily examined.

However, we must determine the regions by analyzing only m ~ 100 points scattered in a space with n ~ 10000 dimensions!

Gene selection

In this second approach it is particularly important to determine the subset of genes involved in establishing the regions and the separating surfaces.

This has two direct benefits:

1. To retrieve knowledge on the biological process at hand,
2. To improve the discrimination between the two sets of points.

The problem of determining this subset of genes is referred to as **feature** (or **gene**) **selection**. Different methods have been proposed to perform gene selection:

- Golub et al. (Science, 1999) have employed a univariate statistical method,
- Guyon et al. (Machine Learning, 2002) have applied a linear Support Vector Machine (SVM).

Rule Generation Techniques

A class of methods that may perform well in approaching the gene selection problem is that of **rule generation techniques**, which are able to obtain a set of intelligible rules underlying the biological process at hand.

A **rule** is a structure of this kind:

if *<premise>* **then** *<consequence>*

where *<premise>* is a logical expression containing one or more conditions, joined by AND, OR and NOT operators.

<consequence> is a value or a set of values for the output.

Example:

$$\textbf{if}\;\; x_{345} > -0.24 \;\; \text{AND} \;\; x_{1258} \leq \; 0.76 \;\; \textbf{then} \;\; y = 1 \text{ (disease)}$$

Switched Neural Networks (SNN)

The rule generation techniques that possess the best theoretical motivations are decision tree methods (CART, ID3, C4.5,…).

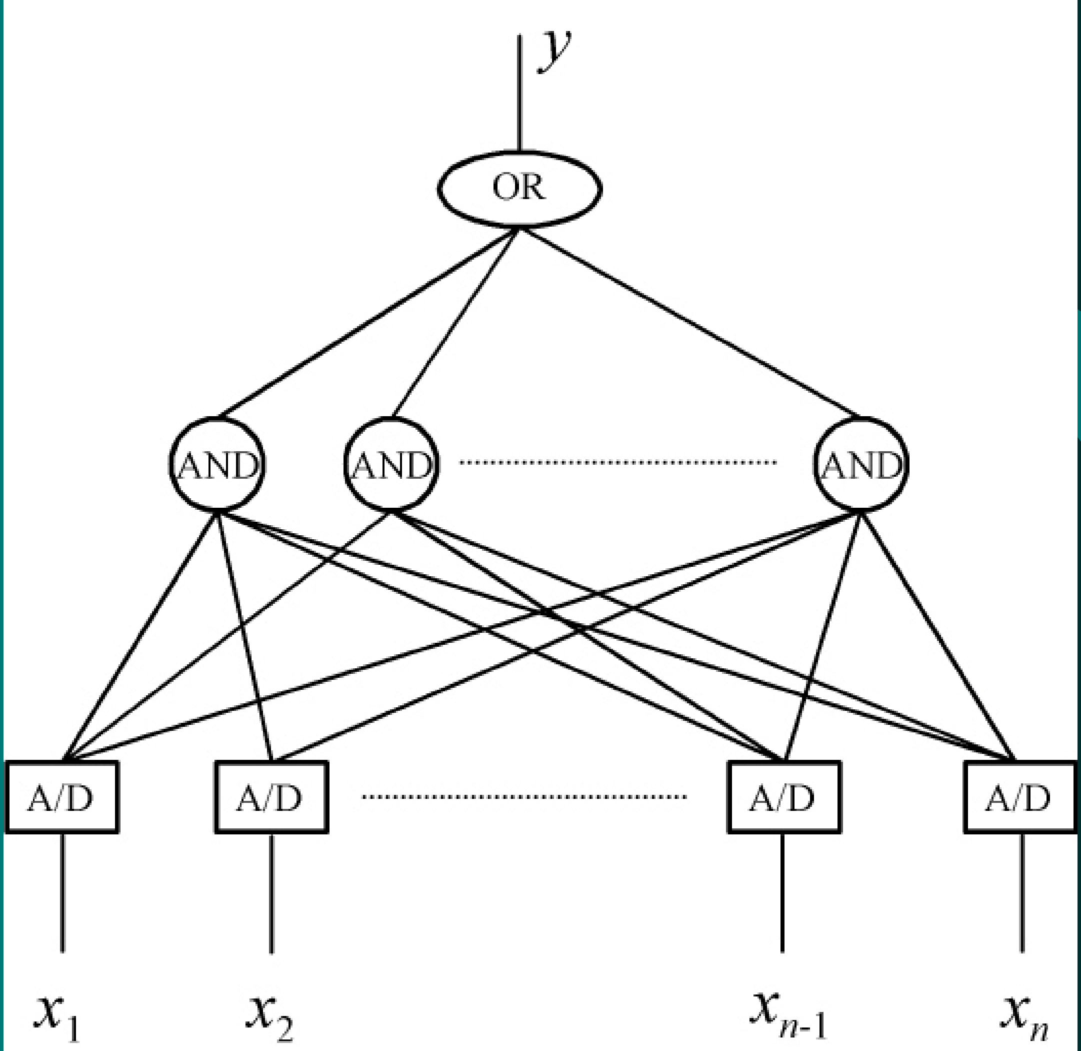
Generally, these techniques show an accuracy inferior to that of SVM, but allow to understand the behavior of the system.

Here, we consider the application of a new rule generation technique, named **Shadow Clustering (SC)**, which produces artificial devices called **Switched Neural Networks (SNN)**.

They are simple weightless connectionist models, on which travel signals with a single level; every unit (neuron) performs an elementary logical operation, AND or OR.

Although SNN are universal approximators, it can be shown that every SNN is equivalent to a digital circuit containing only AND and OR, i.e. it can be described by a monotone Boolean function.

Switched Neural Networks (cont.)



Shadow Clustering (SC)

SC adopts an approach similar to that employed by Hamming Clustering (Muselli and Liberati, IEEE Trans. KDE, 2002); it consists of the following three steps:

1. The input variables (gene expression levels) are mapped into binary strings by using a proper coding that preserves their basic properties (ordering and distance).
2. The AND-OR expression of a Boolean function is retrieved starting from the available examples (coded in binary form).
3. The corresponding SNN is directly generated by inspecting the AND-OR expression obtained in the previous step.

Every AND neuron in an SNN can be translated into an intelligible rule underlying the problem at hand.

Shadow Clustering (cont.)

As a byproduct of the training process, SC is able to determine redundant input variables for the analysis at hand, thus performing in an automatic way the desired gene selection.

In addition, the availability of intelligible rules that link genes among them can help understanding the biological mechanisms involved in the problem at hand.

Since the values of gene expression levels are real numbers, the mapping employed at Step 1 of SC requires a previous discretization. A technique based on entropy analysis (Ent-MDL) is adopted.

Then, a binary string is associated with every interval produced by the discretization process, according to the only-one coding. This allows to preserve ordering and distance, if a proper metric, called **lower distance**, is employed.

Shadow Clustering (cont.)

At least in principle, any method for the synthesis of Boolean functions can be used at Step 2 of SC. Unfortunately, classical techniques do not take into account the problem of generalizing information and score a poor accuracy.

To get round of this problem, SC follows a competitive approach: at every iteration SC groups together binary strings that have the same output and are close among them according to the lower distance.

A final pruning phase has the aim of simplifying the resulting AND-OR expression, thus improving its generalization ability.

Result comparison

To evaluate the results obtained by SNN, the following three datasets have been considered:

1. The Leukemia dataset (Golub et al., 1999), which consists of 72 examples with 7129 genes (47 from Acute Lymphoblastic Leukemia (ALL) and 25 from Acute Myeloid Leukemia (AML))
2. The Colon dataset (Alon et al., PNAS, 1999), which consists of 62 examples with 2000 genes (40 from tissues with colon cancer and 22 from normal tissues).
3. The Lymphoma dataset (Alizadeh et al., Nature, 2000), which consists of 81 examples with 4682 genes (43 from Diffuse Large B-cell Lymphoma (DLBCL) and 38 from other two diseases, B-cell Chronic Lymphocytic Leukemia (B-CLL) and Follicular Lymphoma (FL)).

Result comparison (cont.)

Every dataset has been split into two parts of almost equal size: the first is used to perform discrimination and gene selection, whereas the second one is reserved to evaluate accuracy.

SNN has been compared with the method of Golub et al. (1999) and with SVM, to establish its quality. The accuracy has been measured both with leave-one-out and with the test set.

	Leukemia		Colon		Lymphoma	
Method	Leave-one-out	Test set	Leave-one-out	Test set	Leave-one-out	Test set
SNN	**97.4%**	**94.1%**	**83.9%**	**87.1%**	89.6%	**91.7%**
Golub	89.5%	82.3%	80.6%	**87.1%**	89.6%	87.5%
SVM	94.7%	**94.1%**	77.4%	83.9%	**91.7%**	**91.7%**

Ruleset obtained by SNN for Leukemia

Rules for output ALL:

if $x_{4847} \le 994$ **then** $y = 1$

if $x_{3258} \le 2909.5$ AND $x_{4211} > 1470$ **then** $y = 1$

if $x_{1926} \le 83.5$ AND $x_{3877} \le 348.5$ **then** $y = 1$

if $x_{1882} \le 1419.5$ AND $x_{5985} \le 1866$

Rules for output AML:

if $x_{4847} > 994$ **then** $y = 0$

if $x_{2233} \le 80.5$ AND $x_{3320} > 1341$ **then** $y = 0$

if $x_{1621} \le 311$ AND $x_{2020} > 1346$ **then** $y = 0$

if $x_{1239} \le 3559$ AND $x_{6041} > 992.5$ **then** $y = 0$

Ruleset obtained by SNN for Colon

Rules for output Colon cancer present:

$\textbf{if}\ x_{1058} \leq 0.703 \text{ AND } x_{1671} > -0.898\ \textbf{then}\ y = 1$

$\textbf{if}\ x_{992} > -1.061 \text{ AND } x_{1423} \leq 1.113\ \textbf{then}\ y = 1$

$\textbf{if}\ x_{1884} \leq 0.945 \text{ AND } x_{1917} \leq 1.116 \text{ AND } x_{1998} > -0.298\ \textbf{then}\ y = 1$

$\textbf{if}\ x_{194} \leq 1.117 \text{ AND } x_{533} > -1.055 \text{ AND } x_{1998} > -0.298\ \textbf{then}\ y = 1$

$\textbf{if}\ x_{315} \leq 0.673 \text{ AND } x_{572} \leq 0.776 \text{ AND } x_{1672} \leq 0.909$
$\text{AND } x_{1998} > -0.298\ \textbf{then}\ y = 1$

Rules for output Colon cancer not present:

$\textbf{if}\ x_{839} > -0.286 \text{ AND } x_{1398} \leq 0.024\ \textbf{then}\ y = 0$

$\textbf{if}\ x_{739} > 0.399 \text{ AND } x_{1334} \leq 0.364\ \textbf{then}\ y = 0$

$\textbf{if}\ x_{267} > 0.293 \text{ AND } x_{1181} \leq 0.303\ \textbf{then}\ y = 0$

$\textbf{if}\ x_{249} > 0.449 \text{ AND } x_{1328} > -0.388\ \textbf{then}\ y = 0$

$\textbf{if}\ x_{201} > 0.111 \text{ AND } x_{1549} \leq 0.114\ \textbf{then}\ y = 0$

Ruleset obtained by SNN for Lymphoma

Rules for output DLBCL present:

if $x_{3099} > -0.49$ AND $x_{1258} > -0.07$ **then** $y = 1$

if $x_{1803} \leq 0.165$ AND $x_{2722} > -0.16$ **then** $y = 1$

if $x_{1017} \leq 0.535$ AND $x_{2627} > -0.085$ AND $x_{3791} > -0.345$ **then** $y =$

if $x_{1016} \leq 0.32$ AND $x_{1734} \leq 0.375$ AND $x_{2759} > -0.27$ **then** $y = 1$

if $x_{471} \leq 0.205$ AND $x_{3097} > -0.17$ AND $x_{3776} > -0.125$ **then** $y = 1$

Rules for output DLBCL not present:

if $x_{3765} \leq 1.12$ AND $x_{3875} \leq 0.605$ AND $x_{3997} \leq 0.015$ **then** $y = 0$

if $x_{3265} \leq 0.845$ AND $x_{3387} \leq 0.51$ AND $x_{3766} \leq 1.045$ **then** $y = 0$

if $x_{1030} > -0.155$ AND $x_{3533} \leq 1.385$ AND $x_{3767} \leq 0.905$ **then** $y = 0$

if $x_{953} > -0.335$ AND $x_{1418} \leq 0.49$ AND $x_{2515} \leq 0.49$ **then** $y = 0$

if $x_{8} > -0.33$ AND $x_{952} > -0.205$ AND $x_{2461} \leq 0.52$ **then** $y = 0$

Result comparison (cont.)

To compare the results obtained through the gene selection phase is more difficult, since the correct set of genes involved in the considered disease is not known.

It may be interesting to evaluate the number of selected genes that are common to different methods. To this aim only the 10% genes with the highest rank has been taken into account.

	Leukemia			Colon			Lymphoma		
Method	SNN	Golub	SVM	SNN	Golub	SVM	SNN	Golub	SVM
SNN	—	**62.1%**	35.9%	—	**65.0%**	36.0%	—	**59.5%**	23.5%
Golub	**62.1%**	—	47.9%	**65.0%**	—	41.0%	**59.5%**	—	25.0%
SVM	35.9%	47.9%	—	36.0%	41.0%	—	23.5%	25.0%	—

Conclusions

- The feature selection problem has been defined, pointing out its importance in determining the genes connected with a given pathology.
- A new rule generation method, called Shadow Clustering (SC), has been introduced. It is able to extract a set of intelligible rules underlying a given discrimination problem.
- The application to three microarray datasets has allowed to establish the good generalization ability of Switched Neural Networks (SNN), the devices generated by SC.
- A correct evaluation of the results provided by SNN when performing gene selection is not trivial. A good agreement with the method of Golub et al. (Science, 1999) is obtained at a first examination.

Work in progress

- A correct evaluation of feature selection methods would require the knowledge of the genes involved in a given pathology, which is the target of our analysis.
- A possible alternative can be to generate artificial examples that resemble data produced by a microarray, according to some biological assumptions.
- We are currently working on a general model, supported by theoretical and experimental motivations, which is able to generate artificial datasets presenting the same characteristics as real ones.
- In the meantime we are also examining the discretization process used when training SNN, in order to establish the optimal procedure to be adopted when treating microarray data.