Discovering Gene Interactions from Gene Expression Data using Neural Networks

Luca Nanni, Artur Sakalouski
luca.nanni@polimi.it, artur.sakalouski@gmail.com

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

**ABSTRACT**

Search for genetic factors influencing complex traits:
- A challenge and goal for genetics
- Standard methods not able to detect high level interactions and no correlation with phenotype [2].

We propose a methodology relying on neural networks:
- Works on RNA-seq data
- Identify interacting gene subsets
- Cooperative interaction between genes

**DATA**

We use RNA-seq data, which can be seen as a matrix:

<table>
<thead>
<tr>
<th>Genes</th>
<th>Label</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>G2</td>
<td>...</td>
</tr>
<tr>
<td>P1</td>
<td>300</td>
<td>0</td>
</tr>
<tr>
<td>P2</td>
<td>0</td>
<td>23</td>
</tr>
</tbody>
</table>

- # features: **20530**
- # patients: **1000**

Dimensionality curse

**RESULTS**

- Ontological analysis of the results (GO ontology)
- We are able to identify known interactions related to presence of cancer, but results are unstable!!
- We have an overlapping of the top interacting genes of 25% between different runs.

**METHODOLOGY**

We train a neural network for the classification of patients on the basis of presence or absence of cancer.

We use a method described in [1] to extract knowledge from the trained neural network weights.

Backpropagate the weight influence w.r.t. output:

$$z^{(l)} = \frac{1}{|W^{(L)}|} \left| W^{(L-1)} \right| \cdots \left| W^{(1)} \right|$$

We assign a score to each candidate interaction:

$$\omega_i(\mathcal{I}) = z_i^{(1)} \mu \left( |W^{(1)}_{i}| \right) \quad \mathcal{I} = \{g_{i_1}, g_{i_2}, \ldots, g_{i_k}\}$$

Finally we rank interactions based on the score. The subsets of features which are found represent genes which are jointly correlated with the output, ranked on the basis of their influence.

**PROBLEMS**

- Class unbalance (tumor cases are much more than normal ones) → up-sampling
- Dimensionality curse → prefiltering on the genes
- Instability of the results (multiple local minima) → using multiple regularized neural networks

**REFERENCES**
