Inductive Biases for Stable Feature Selection in High Dimensional Spaces

Applications to Gene Profiling and Diagnosis from DNA Microarrays

Thibault Helleputte
thibault.helleputte@uclouvain.be
http://www.ucl.ac.be/mlg

Machine Learning Group, Université catholique de Louvain

PhD Thesis Public Defense, 7 September 2010
Outline

1. Introduction
2. Machine Learning
3. Evaluation
4. Contributions
5. Cristall Project
6. Conclusion
Aim: Response prediction and biomarker identification.
Cristall: Allergy Risk Factor Identification in Newborns

Aim: Prognosis and biomarker identification.
Since proteins rule activity, structure and communication of the cells, we would like to **measure** them.
Measuring Proteins

- Since proteins rule activity, structure and communication of the cells, we would like to measure them.
- Not efficiently feasible so far for a large number of proteins at once.

**FIGURE 2.27.** General mechanism of eukaryotic protein synthesis. The major steps include transcription of the DNA gene sequence into the messenger RNA template in the nucleus of the cell, translation of the DNA codons of that gene into amino acids, and their assembly into polypeptides in the cytoplasm. Important mediators of this process include transfer RNAs, splicing elements, and ribosomes.
Measuring Proteins

- Since proteins rule activity, structure and communication of the cells, we would like to measure them.
- Not efficiently feasible so far for a large number of proteins at once.
- **Gene expression** can be measured efficiently instead (how often a gene is copied).

**FIGURE 2.27.** General mechanism of eukaryotic protein synthesis. The major steps include transcription of the DNA gene sequence into the messenger RNA template in the nucleus of the cell, translation of the DNA codons of that gene into amino acids, and their assembly into polypeptides in the cytoplasm. Important mediators of this process include transfer RNAs, splicing elements, and ribosomes.
Gene expression can be measured by microarray:

- Close proxy to protein production
- Evolving, unlike the genome: study of time/condition/environment-influenced factors
- Subject to noise (not only xp, also bio)
Microarray Data

After preprocessing steps:

<table>
<thead>
<tr>
<th></th>
<th>gene 1</th>
<th>gene 2</th>
<th>...</th>
<th>gene p</th>
<th>class label</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample 1</td>
<td>$x_{1,1}$</td>
<td>$x_{1,2}$</td>
<td>...</td>
<td>$x_{1,p}$</td>
<td>$y_1$</td>
</tr>
<tr>
<td>sample 2</td>
<td>$x_{2,1}$</td>
<td>$x_{2,2}$</td>
<td>...</td>
<td>$x_{2,p}$</td>
<td>$y_2$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sample n</td>
<td>$x_{n,1}$</td>
<td>$x_{n,2}$</td>
<td>...</td>
<td>$x_{n,p}$</td>
<td>$y_n$</td>
</tr>
</tbody>
</table>

- Class labels come from external annotation.
- With recent technology, $p \approx 55,000$. ($p$ genes = dimensions = features).
- Very expensive technology, so $n$ is typically $\approx 100$
Outline

1. Introduction
2. Machine Learning
3. Evaluation
4. Contributions
5. Cristall Project
6. Conclusion
Machine Learning

Microarray data analysis:
- is not feasible manually.
- can be addressed by Machine Learning techniques.
Microarray data analysis:

- is not feasible manually.
- can be addressed by **Machine Learning** techniques.

What is Machine Learning? Can machine really **learn**?
Machine Learning

Microarray data analysis:

- is not feasible manually.
- can be addressed by Machine Learning techniques.

What is Machine Learning? Can machine really learn?
Machine Learning

Microarray data analysis:

- is not feasible manually.
- can be addressed by Machine Learning techniques.

What is Machine Learning? Can machine really learn?

Machine Learning Applications

- Spam Filtering
- Smile Detection for Camera
- Optical Character Recognition
- Automated Driving
- Automated Trading
- ...

YES WE CAN!
What Do Machines “Learn”?

The Typical Machine Learning Process

Based on some input values $x$, estimate a function $f(x)$ able to predict a response $y$. 
What Do Machines “Learn”?

The Typical Machine Learning Process

Based on some input values $x$, estimate a function $f(x)$ able to predict a response $y$.

For example:

- The input can be gene expression (a vector of continuous values).
- The response can be:
  - binary (diseased/healthy?)
  - discrete (disease1, disease2 or disease3?)
  - continuous (survival time?)
- The function can be a hyperplane.
What Do Machines “Learn”? 

The Typical Machine Learning Process

Based on some input values $x$, estimate a function $f(x)$ able to predict a response $y$.

For example:
- The input can be gene expression (a vector of continuous values).
- The response can be:
  - binary (diseased/healthy?)
  - discrete (disease1, disease2 or disease3?)
  - continuous (survival time?)
- The function can be a hyperplane.
What Do Machines “Learn”?

The Typical Machine Learning Process

Based on some input values $x$, estimate a function $f(x)$ able to predict a response $y$.

For example:

- The input can be gene expression (a vector of continuous values).
- The response can be:
  - binary (diseased/healthy?)
  - discrete (disease1, disease2 or disease3?)
  - continuous (survival time?)
- The function can be a hyperplane.
What Do Machines “Learn”?  

The Typical Machine Learning Process

Based on some input values \( x \), estimate a function \( f(x) \) able to predict a response \( y \).

For example:
- The input can be gene expression (a vector of continuous values).
- The response can be:
  - binary (diseased/healthy?)
  - discrete (disease1, disease2 or disease3?)
  - continuous (survival time?)
- The function can be a hyperplane.

Typical setting:
- Training set: limited number of patients with known response.
- Test set: new patients for which we want a prediction.
Function Estimation: Inductive Bias Needed

Inductive Bias

The inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered.

Without an inductive bias, a learner can only memorize the training set, and predict at random on new inputs.
Inductive Bias

The *inductive bias* of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered.

Without an inductive bias, a learner can only memorize the training set, and predict at random on new inputs.

**Simple bias:** minimize a linear classifier error
Inductive Bias

The inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered.

Without an inductive bias, a learner can only memorize the training set, and predict at random on new inputs.

**Simple bias:** minimize a linear classifier error

**Stronger bias:** maximize a linear classifier margin (SVM)
The Need for Patients

More difficult to learn when few training patients are available.
The Need for Patients

More difficult to learn when few training patients are available.
The Need for Patients

More difficult to learn when few training patients are available.

What really matters is density: The ratio between number of patients \( n \) and the number of genes \( p \) should be sufficiently high.
The Curse of Dimensionality

High-dimensionality $p$ coupled with low number of patients $n$ is an issue for learning:

Patients density drops exponentially with the number $p$ of genes.

Density of a $r$-side cube with $k$ patients for each gene:

$$d(k, r, p) = \frac{p \times k}{r^p}$$

Small-$n$-big-$p$ settings cause the Curse of Dimensionality.

Note: High dimensionality is also an issue w.r.t. algorithms complexity.
Machine Learning on Microarray Data

From the Machine Learning point of view:

- Diagnosis, Prognosis, Treatment Response Prediction → **Classification**
- Biomarkers Identification (Signature), Gene Profiling → **Feature Selection**
Machine Learning on Microarray Data

From the Machine Learning point of view:

- Diagnosis, Prognosis, Treatment Response Prediction
  → **Classification**

- Biomarkers Identification (Signature), Gene Profiling
  → **Feature Selection**

Feature Selection on Microarray Data

- Explanatory concerns
- Affordable, feasible diagnosis/prognosis kits
Feature Selection: an Example

Suppress the gene which contributes the less to the classification decision (heart of the Recursive Feature Elimination).
Main objectives:

- Develop methods for feature selection in a context of classification...
- ... with good performances on high-dimensional data ...
- ... with gene expression data analysis as an illustrative application.
Main message

Feature selection and classification of high dimensional data based on few samples constitute a complex challenge.

It can be addressed by imposing strong inductive biases, either by using internal information (ensemble methods) or external information (expert knowledge, transfer or multi-task learning).
Outline

1. Introduction
2. Machine Learning
3. Evaluation
4. Contributions
5. Cristall Project
6. Conclusion
Evaluation Metrics

Selection Quality: Stability Index

- Similarity between $k$ signatures $S$ of size $s$.
- Kuncheva Index: $K(\{S_1, \ldots, S_k\}) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{|S_i \cap S_j| - \frac{s^2}{n}}{s - \frac{s^2}{n}}$

$n$ is the total number of genes and $S_i, S_j$ are two signatures.

Classification Quality: Balanced Classification Rate

- Stability alone cannot characterize a signature quality.
- BCR Pro’s:
  - **Meaningful**: mean between sensitivity and specificity
  - **Not fooled** by unbalanced class priors
  - Straightforward extension to multi-class
Experiments

Avoid normalization and selection bias, overfitting,...

In most experiments of the thesis: N-Resamplings.

N-Resamplings

Repeat $J$ times:

1. Randomly split patients into 90% train - 10% test
2. On train part:
   1. Preprocess
   2. Select genes
   3. Estimate a predictive model on selected genes
3. Evaluate classification performances on test patients

Average the classification performances and stability on the $J$ selected sets of genes.
Outline

1. Introduction
2. Machine Learning
3. Evaluation
4. Contributions
5. Cristall Project
6. Conclusion
Ensemble Selection

Ensemble Idea

Instead of one gene selection, perform many selections on different subsets of the patients and build a consensus out of them, (≈ bagging for selection).
**Ensemble Selection**

**Ensemble Idea**

Instead of one gene selection, perform many selections on different subsets of the patients and build a consensus out of them, (≈ bagging for selection).

**Traditionnal setting:**

![Diagram showing traditional setting](image-url)
Ensemble Selection

**Ensemble Idea**

Instead of one gene selection, perform **many selections on different subsets** of the patients and build a **consensus** out of them, (≈ bagging for selection).

**Traditionnal setting:**

Dataset → Model

**Ensemble setting:**

Dataset → Consensus → Model
Results

Prostate dataset: 52 vs. 50 patients and 6,033 genes.

Stability (Kuncheva Index)  

Classification Performances (AUC)
Conclusion on Ensemble Methods

Ensemble selection benefits w.r.t. the baseline:

- gain in stability
- gain in classification performance
- based on microarray data only (no extra information)

Publications related to this contribution: [7, 1, 2, 3].
Conclusion on Ensemble Methods

Ensemble selection benefits w.r.t. the baseline:

- gain in stability
- gain in classification performance
- based on microarray data only (no extra information)

Publications related to this contribution: [7, 1, 2, 3].

What if we have extra information on top of the microarray data?
Partially Supervised Feature Selection (PSFS)

Prior Knowledge About Gene Relevance

Field experts may know or guess that some genes are likely to be more relevant. This prior knowledge might be:

- partial/insufficient for a complete model
- imprecise
Partially Supervised Feature Selection (PSFS)

Prior Knowledge About Gene Relevance

Field experts may know or guess that some genes are likely to be more relevant. This prior knowledge might be:

- partial/insufficient for a complete model
- imprecise

Use of prior knowledge on genes relevance to bias feature selection.
PS-AROM

Similar to an existing method (AROM) but:
- Relevance vector $\beta$
- Prior relevance of gene $j$ encoded in $\beta_j$.
- The more (a priori) relevant gene $j$, the higher $\beta_j$.
- If no information on $j$, $\beta_j = 1$.

Partially-Supervised Approximation to zeroRO-norm Minimization

Solve:

$$\min_{w} \sum_{j=1}^{p} \frac{1}{\beta_j} \ln(|w_j| + \varepsilon)$$

Subject to:

$$y_i \langle x_i, w \rangle \geq 1 \quad \forall i \in \{1, \ldots, n\}$$
Results: DLBCL with 2 favored genes

DLBCL dataset: 57 vs. 20 patients and 7,129 genes.

Stability (Kuncheva Index)  
Classification Performances (BCR)
Conclusion on PSFS

- includes prior knowledge on a priori important genes while letting the feature selection procedure depart from it.
- PSFS increases stability of selected genes with respect to patients sampling variations and improves classification performances in most cases.
- Multivariate method: supervision of few genes influence the selection of other ones.

Publications related to this contribution: [6].
Transfer Learning via PSFS

- What if no expert knowledge is available?
Transfer Learning via PSFS

- What if no expert knowledge is available?
- Transfer knowledge about genes relevance from similar datasets.
Transfer Learning via PSFS

- What if no expert knowledge is available?
- Transfer knowledge about genes relevance from similar datasets.

1. Extract knowledge from source domain(s) with a baseline selection method.
2. Use this extracted prior via PSFS on the target domain.
### 3 Microarray Datasets

Use of 3 Prostate Cancer datasets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Normal/Tumor</th>
<th>Feat.</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singh</td>
<td>50/52</td>
<td>12,625</td>
<td>HGU95Av2</td>
</tr>
<tr>
<td>Chandran</td>
<td>18/86</td>
<td>12,625</td>
<td>HGU95Av2</td>
</tr>
<tr>
<td>Welsh</td>
<td>9/25</td>
<td>12,626</td>
<td>HGU95A</td>
</tr>
</tbody>
</table>

12,600 probesets in common.
Results

Stability (Kuncheva Index)  
Classification Performances (BCR)
Conclusion on PSFS for Transfer Learning

- PSFS can be used for inductive transfer learning at the gene level.
- Gain in classification performance and stability.

Publications related to this contribution: [5, 4].
Sparse Multi-Task Learning

- Instead of using several datasets to build a model on another dataset, why not mutualize the information?
Sparse Multi-Task Learning

- Instead of using several datasets to build a model on another dataset, why not mutualize the information?
- Aim: favor common sparsity patterns across tasks.
Mathematically

Optimizing on $K$ related tasks at once:

$$
\min_{\mathbf{w}} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathcal{L}(y_{ik}, \langle \mathbf{w}_k, \mathbf{x}_{ik} \rangle) + \lambda \| \mathbf{W} \|_{i,j}
$$

with $i \in \{0, 1\}$ and $j \in \{1, 2, \infty\}$ and $\mathcal{L}$ being logistic loss.
Conclusions on Sparse Multi-Task Learning

- **Dedicated solver** with complexity $\mathcal{O}(n^2p)$
- **Six regularizers** with various specificities
  - $\|W\|_{1,1}$ comes to **single task** learning
  - $\|W\|_{1,2}$ offers **best classification** performances
  - $\|W\|_{1,\infty}$ offers **best between-task stability**
  - $\|W\|_{0,j}$: selects fewer genes but poorer classification performances
- This method has also been shown to solve gene profiling for **multi-class** problems.
Outline

1. Introduction
2. Machine Learning
3. Evaluation
4. Contributions
5. Cristall Project
6. Conclusion
Cristall Project: Context & Aim

Global Cristall objective: identify allergy risk-factors in newborns.

300 children followed-up over 5 years.

Cristall microarray data analysis objectives:

1. Best possible prognosis model:

   \[ \text{Allergic Status}_{18 \text{Months}} \leftarrow f(\text{Blood}_{6 \text{Months}}) \]

2. ... based on the smallest set of biomarkers possible.

So far, 73 hybridized samples (77 - 4 outliers).
Signature Identification: Selection Methods

Univariate:
- $t$-Test
- Golub’s ratio

Multivariate:
- Ensemble Selection
- Recursive Feature Elimination
- $L_2$-AROM
- Random Forest

Multivariate with partial supervision
- PSFS (Prior Knowledge, 43 genes, 102 probesets)
- PSFS ($t$-Test Knowledge, 100 probesets)
Results

Stability (Kuncheva Index)

Prognosis Performances (BCR)
Cristall Conclusions

Prognosis of the 18 months allergic status based on gene expressions at 6 months seems (so far) too difficult.

- All biomarkers identification techniques fail to achieve satisfactory prediction performances.
- Sound tendencies observed (higher stability for PSFS and ensemble methods).
- Prior knowledge seems relevant: highest stability.
Cristall Perspectives

New results: Prognosis of the 3 years allergic status based on gene expressions at 6 months seems easier!

Stability (Kuncheva Index)

Prognosis Performances (BCR)
Outline

1. Introduction
2. Machine Learning
3. Evaluation
4. Contributions
5. Cristall Project
6. Conclusion
Contributions I

<table>
<thead>
<tr>
<th>Main message</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature selection and classification of high dimensional data based on few samples constitute a complex challenge.</td>
</tr>
<tr>
<td>It can be addressed by imposing strong inductive biases, either by using internal information (ensemble methods) or external information (expert knowledge, transfer or multi-task learning).</td>
</tr>
<tr>
<td>Proposed methods improve classification performances and stability most of the time.</td>
</tr>
</tbody>
</table>
Contributions II

Cristall project:
- Prognosis: difficult at 18 months, easier at 3 years.
- Consistent results: ensemble selection and PSFS improve stability.

Other methodological contributions:
- Relevant evaluation metrics and sound evaluation protocol
- Generalized bootstrap .632+
- Non-specific filtering
- Robust normalization
Perspectives I

- Imposing even more structure on the input space: TV-norm, group-regularization
- Optimize BCR instead of (a proxy for) accuracy.
- Always add an $L_2$ penalty $\rightarrow$ strict convexity could increase stability.
- Transfer & Multi-task: dataset similarity?
Thanks to Trainance.com!

This talk has been recorded by:

Trainance is an innovative Platform that allows to capture and distribute events, corporate training, webinars.

http://www.trainance.com


References III