ALGEBRAIC STABILITY INDICATORS FOR RANKED LISTS IN MOLECULAR PROFILING

Giuseppe Jurman @ MPBA

Fondazione Bruno Kessler
Trento
Italy

Sandpit on the Application of Machine Learning Techniques to the Analysis of Complex Biomedical Data
London, 6th July 2009
A result of a classification/ranking procedure is a list of features, ranked according to their relevance for the classifier.

To ensure repeatability and not to incur in overfitting due to information leakage phenomena and such as the selection bias, methodology has to be carefully designed.
\[ \mathcal{M}(n \times p, \mathbb{R}) \]

- \( n \approx 10^3 \) samples
- \( p \approx 10^5 \) genes / \( 10^6 \) SNPs

A result of a classification/ranking procedure is a list of features, ranked according to their relevance for the classifier.

To ensure repeatability and not to incur in overfitting due to information leakage phenomena and such as the selection bias, methodology has to be carefully designed.

Error curve for the Colon cancer dataset (left) with true labels (blue) and random labels (black).
The problem

\[ M(n \times p, \mathbb{R}) \]

\[ n \approx 10^3 \text{ samples} \]

\[ p \approx 10^5 \text{ genes} / 10^6 \text{ SNPs} \]

A result of a classification/ranking procedure is a list of features, ranked according to their relevance for the classifier.

To ensure repeatability and not to incur in overfitting due to information leakage phenomena and such as the selection bias, methodology has to be carefully designed.

Error curve for the two synthetic datasets (right) \( f1000 - 5000 \), with 1000 discriminant features (black) and \( f0 - 5000 \) (blue) with no discriminant features.
The problem

\[ M(n \times p, \mathbb{R}) \]

\[ n \approx 10^3 \text{ samples} \]

\[ p \approx 10^5 \text{ genes} / 10^6 \text{ SNPs} \]

A result of a classification/ranking procedure is a list of features, ranked according to their relevance for the classifier.

To ensure repeatability and not to incur in overfitting due to information leakage phenomena and such as the selection bias, methodology has to be carefully designed.

Result of a multi-institution analysis of papers about gene expression profiling in a leading journal:

- Inability to reproduce the analysis > 50%
- Partial reproduction in 1/3
- Perfect reproduction in 11%
The problem

Solution: use a correctly planned Data Analysis Protocol (DAP) implementing a resampling scheme.

At each run $i=1 \ldots B$ an ordered list $L_i$ is produced.

$\mathcal{M}(B \times d, \mathbb{N}), B \simeq 10^3$

The set of lists

$\mathcal{L} = \{L_i\}_{i=1}^B$ is the set of all lists and, for each list, $L_i^k$ is its top-$k$ list. i.e. the sublist consisting of the first ranked $k$ elements from $L_i$.

What information can be derived from the structure of the feature lists set?
**Permutation Group Theory**

<table>
<thead>
<tr>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>A <strong>group</strong> is a set together with an associative operation which admits an identity element and such that every element has an inverse.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Permutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>A <strong>permutation</strong> on a set ( \Omega_n = {s_1, s_2, \ldots, s_n} ) of ( n ) elements is a <strong>bijection</strong></td>
</tr>
</tbody>
</table>

\[
\tau : \Omega_n \rightarrow \Omega_n \\
\quad s_i \mapsto \tau(s_i) = s_j
\]

The number of different permutations on set of \( n \) elements is \( n! \).

<table>
<thead>
<tr>
<th>Symmetric Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>The set of all permutations on a set ( \Omega_n ) (usually ( \Omega_n = \mathbb{N} \cap [1, n] )) becomes a group ( S_n ), called <strong>permutation group</strong> or <strong>symmetric group</strong> when endowed with the operation</td>
</tr>
</tbody>
</table>

\[
\sigma \tau : \Omega_n \rightarrow \Omega_n \\
\quad s_i \mapsto \sigma(\tau(s_i)) = s_k
\]
**List representation**

*Example:* The symmetric group $S_3$ consists of

\[
\{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3\} \quad \{1 \rightarrow 2, 2 \rightarrow 1, 3 \rightarrow 3\} \\
\{1 \rightarrow 3, 2 \rightarrow 2, 3 \rightarrow 1\} \quad \{1 \rightarrow 1, 2 \rightarrow 3, 3 \rightarrow 2\} \\
\{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 1\} \quad \{1 \rightarrow 3, 2 \rightarrow 1, 3 \rightarrow 2\}
\]

**Representation**

Labelling the $n$ genes involved in a problem by $\mathbb{N} \cap [1, n]$, every ranked list $L_t$ can be identified as a permutation $\tau_t \in S_n$, the image $\tau_t(i)$ of the $i$-th gene being its ranking inside the list $L_t$.

<table>
<thead>
<tr>
<th>Gene</th>
<th>$\Omega_n$</th>
<th>$\tau_t(\Omega_n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-MurIL2_at</td>
<td>$\rightarrow 1$</td>
<td>$\rightarrow 13347$</td>
</tr>
<tr>
<td>...</td>
<td>$\rightarrow \ldots$</td>
<td>$\rightarrow \ldots$</td>
</tr>
<tr>
<td>99764_at</td>
<td>$\rightarrow n$</td>
<td>$\rightarrow 175$</td>
</tr>
</tbody>
</table>

Thus $\tau$ can be written as $\tau = (\tau(1), \ldots, \tau(n)) = (13347, \ldots, 175)$. 
Metrics on permutation groups

**Definition**

A map \( d \) defined on a permutation group \( S \)

\[
d : S \times S \rightarrow \mathbb{R}
\]

is a distance function or a **metric** if it satisfies the following axioms:

**Non negativity** \( d(\sigma, \tau) \geq 0 \quad \forall \sigma, \tau \in S \)

**Separation** \( d(\sigma, \tau) = 0 \iff \sigma = \tau \quad \forall \sigma, \tau \in S \)

**Symmetry** \( d(\sigma, \tau) = d(\tau, \sigma) \quad \forall \sigma, \tau \in S \)

**Triangular** \( d(\sigma, \tau) + d(\tau, \rho) > d(\rho, \sigma) \quad \forall \sigma, \tau, \rho \in S \)

**Right invariance** \( d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \quad \forall \sigma, \tau, \rho \in S \)

Triangularity might sometimes be relaxed to some less strict requirements (e.g. polygonal inequalities) that makes \( d \) a **pseudometric**.
The fifth property
\[ d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \quad \forall \sigma, \tau, \rho \in S \]
is an essential requirement for distance functions in permutation group spaces.

### An Example
- \( L_1 = \{A, C, D, B, E\} \) and \( L_2 = \{B, C, D, E, A\} \) via the identification \( A1, B2, C3, D4, E5 \) correspond to \( \tau_1 = (1, 4, 2, 3, 5), \tau_2 = (5, 1, 2, 3, 4) \).
- Apply the permutation \( \rho \) corresponding to the relabeling \( A2, B5, C4, D1, E3 \).
- Then \( \tau_1 \rho = (3, 1, 5, 2, 4) \) and \( \tau_2 \rho = (3, 5, 4, 2, 1) \).

**Right-invariance corresponds to relabeling independence.**

The Euclidean distance \( d(\tau, \sigma) = \sqrt{\sum_{i=1}^{n}(\tau^{-1}(i) - \sigma^{-1}(i))^2} \) is **not** right-invariant.
### Statistical Distances

#### Spearman’s Footrule

It is the $L_1$ distance between permutations: $F(\sigma, \tau) = \sum_{i=1}^{n} ||\sigma(i) - \tau(i)||$

#### Spearman’s Rho

It is the $L_2$ distance between permutations: $R(\sigma, \tau) = \sqrt{\sum_{i=1}^{n} (\sigma(i) - \tau(i))^2}$

#### Kendall’s Tau

It is the minimum number of pairwise adjacent transpositions (swaps) needed to transform $\sigma$ into $\tau$: $K(\sigma, \tau) = \text{Card}\{(i, j) \in \Omega_n^2: \sigma(i) < \sigma(j) \text{ and } \tau(i) > \tau(j)\}$

All these measures are metric.
A key fact in gene lists is that variations in the bottom part of the lists are much less relevant than differences in the top part.

Thus a **weight factor** is needed to take care of such requirement.

A natural candidate to substitute Spearman’s footrule is its weighted version.

**The Canberra distance**

\[
Ca(\tau, \sigma) = \sum_{i=1}^{n} \frac{|\tau(i) - \sigma(i)|}{\tau(i) + \sigma(i)}
\]
Let
\[ H_s = \sum_{j=1}^{s} \frac{1}{j} = \log(s) + \gamma + \frac{1}{2s} - \frac{1}{12s^2} + \frac{\epsilon_s}{120s^4} \]
be the s-th harmonic number with a famous approximation, where \(0 < \epsilon_s < 1\) and \(\gamma \approx 0.5772156\ldots\) is the Eulero-Mascheroni constant.

Many closed forms for sum and products involving \(H_s\) have been found only very recently.

**Theorem**

The expected value for the Canberra distance is
\[
E\{\text{Ca}\}_s = \left(2p + 2 + \frac{1}{2p}\right)H_{2p} - \left(2p + 2 + \frac{1}{4p}\right)H_p - \left(n + \frac{3}{2}\right).
\]

Its approximation is
\[
\hat{E}\{\text{Ca}\}_s = (p + 1) \log(4) - (p + 2) + o(1).
\]
VARIANCE

Let

\[ c_p(i, j) = \frac{|i - j|}{i + j}. \]

and

\[ f_p(i) = \frac{1}{p} \sum_{j=1}^{p} c_p(i, j) = \frac{2i}{p} (2H_{2i} - H_i - H_{p+i} - 1) + 1. \]

**Theorem**

The variance cannot be written in a entirely closed form

\[ V\{Ca\}_{sp} = \frac{1}{p-1} \sum_{i,j=1}^{p} c_p^2(i, j) + E\{Ca\}_{sp} - 2c_p(i, j)f_p(j) \]

\[ = \mathcal{P}(H_p, H_{2p}) + \sum_{i=1}^{p} i^2 H_{p+i} H_{2i} + \sum_{i=1}^{p} i^2 H_{p+i} H_i. \]

Its approximation is

\[ \hat{V}\{Ca\}_{sp} = \alpha p + \beta \log(p) + \delta + o(p^{-1}), \alpha, \beta, \delta \in \mathbb{R} \]
Let
\[ d_p(i, j) = c_p(i, j) - \frac{1}{p} \sum_{l=1}^{p} c_p(i, l) - \frac{1}{p} \sum_{m=1}^{p} c_p(m, j) + \frac{1}{p^2} \sum_{l=1}^{p} \sum_{m=1}^{p} c_p(m, l). \]

Since
\[ \max_{1 \leq i, j \leq p} d_p^2(i, i) = d_p^2(1, 1) = \frac{1}{p} \left( E\{\text{Ca}\} S_p - 2p - 4 - 4H_{p+1} \right), \]
then
\[ \lim_{p \to \infty} \frac{\max_{1 \leq i, j \leq p} d_p^2(i, i)}{\frac{p-1}{p} V\{\text{Ca}\} S_p} = 0, \]
thus by Hoeffding theorem we have

**Theorem**
The distribution of the Canberra distance on \( S_p \) is asymptotically normal.
The problem
\[ \rho = \arg \max_{S_p} (\text{Ca}_I) \]

has one solution \( \rho_M \) for \( p \) even and two solutions for \( p \) odd, namely
\[
\rho_M = \begin{cases} 
\frac{p}{2} \prod_{i=1}^{\frac{p}{2}} \left( i \frac{p}{2} + i \right) & \text{for even } p \\
\theta = \left( 1 \frac{p-1}{2} + 1 \ p \frac{p-1}{2} \cdots 2 \frac{p-1}{2} + 2 \right) \text{ and } \theta^{-1} & \text{for odd } p, 
\end{cases}
\]

while the corresponding maximum value is
\[
\max_{S_p} (\text{Ca}) = \text{Ca}_I(\rho_M) = \begin{cases} 
2r(H_3r - H_r) & \text{if } p = 4r \\
(2r + 1)H_{6r+1} + rH_{3r+1} - \left( r + \frac{1}{2} \right) H_{3r} - (2r + 1)H_{2r+1} + \frac{1}{2} H_r & \text{if } p = 4r + 1 \\
(2r + 1)(2H_{6r+3} - H_{3r+1} - 2H_{2r+1} + H_r) & \text{if } p = 4r + 2 \\
(2r + 1)H_{6r+5} + \frac{1}{2} H_{3r+2} - (2r + 1)H_{2r+1} - (r + 1)H_{r+1} + \left( r + \frac{1}{2} \right) H_r & \text{if } p = 4r + 3
\end{cases}
\]
DISTANCES ON TOP-\(k\) LISTS

Since most important genes are located in the upper part of the lists, it is also important to be able to compare those more important portions of the lists. Managing partial lists is much harder than global lists, since they do not necessarily involve all the same elements.

**Top \(k\)-list**

A top-\(k\) list is the sublist including the elements ranking from position 1 to \(k\) of the original list, which corresponds to induce a partial ordering on \(k\) out of \(n\) objects.

This has a natural counterpart into the group-theoretical framework by Hausdorff topology:

**Distance with location parameter**

- Consider a global distance \(D\)
- Consider two top-\(k\) lists \(\tau_m\) for \(m = 1, 2\), involving respectively the subsets \(D_{\tau_m} = \tau_m^{-1}(\mathbb{N} \cap [1, k])\) of \(\Omega_n\);
- Define the global list \(\tau'_m(i) = \begin{cases} \tau_m(i) & \text{for } i \in D_{\tau_m} \\ k + 1 & \text{otherwise} \end{cases}\)
- Define \(D^{(k+1)}(\tau_1, \tau_2) = D(\tau'_1, \tau'_2)\).
\textbf{Canberra distance with location parameter}

\[ \text{Ca}^{(k+1)}(\tau, \sigma) = \sum_{i=1}^{p} \frac{|\min\{\tau(i), k + 1\} - \min\{\sigma(i), k + 1\}|}{\min\{\tau(i), k + 1\} + \min\{\sigma(i), k + 1\}}. \]

\textbf{Theorem}

The expected (average) value of the Canberra metric on \( S_n \) is

\[
\mathbb{E}\{\text{Ca}^{(k+1)}\}_{S_p} = \frac{k}{p} \left( \left( 2k + 2 + \frac{1}{2k} \right) H_{2k} - \left( 2k + 2 + \frac{1}{4k} \right) H_k - \left( k + \frac{3}{2} \right) \right) + \frac{2(p - k)}{p} (2(k + 1)(H_{2k+1} - H_{k+1} - k)),
\]

which can be approximated up to terms \( o(1) \) as

\[
\hat{\mathbb{E}}\{\text{Ca}^{(k+1)}\}_{S_p} = \frac{(k + 1)(2p - k)}{p} \log(4) - \frac{2kp + 3p - k - k^2}{p}.
\]

For \( p \approx 10^4 \), \( |\mathbb{E}\{\text{Ca}^{(k+1)}\}_{S_p} - \hat{\mathbb{E}}\{\text{Ca}^{(k+1)}\}_{S_p}| \approx 10^{-3} \).
Moreover, the existence of feature modules, for instance highly correlated genes (e.g. clones), or genes belonging to the same pathway must be taken into account. Classifiers as SVM tend to swap the relative importance of correlated features during different ranking processes, and those swaps should be less penalized than movements between uncorrelated genes.

**Modules-aware distances**

Let $M = \{g_1, \ldots, g_m\} \subset \Omega_p = \{1, \ldots, p\}$ be a module consisting of $m$ features. For each permutation $\tau \in \mathcal{L}$, define the permutation $\eta_M \in S_m$ by the property $\eta_M(i) < \eta_M(j)$ if $\tau(g_{\eta_M(i)}) < \tau(g_{\eta_M(j)})$. Then define the new permutation $\tau_M$ as $\tau_M(g_i) = \tau(g_{\eta_M(i)})$, leaving $\tau_M|_{\Omega_p \setminus M} = \tau$ out of $M$. The distance is then computed on $\mathcal{L}_M = \{\tau_M : \tau \in \mathcal{L}\}$. 
EXTENSIONS TO PARTIAL LISTS WITH DIFFERENT LENGTH

Let $L_1$ and $L_2$ be two partial lists of length respectively $l_1 \leq l_2$ whose elements belong to $\mathcal{F}$

Let $d$ be a distance on permutation groups.

Let $S_\tau$ be the set of all the dual lists of the elements in $S_L$, so that if $\alpha \in S_\tau$, then $\alpha(i) = \tau(i)$ for all indices $i \in L$.

Define the distance between $L_1$ and $L_2$ as a function of the distance on the two quotient groups, i.e.

$$d(L_1, L_2) = f \left( \left\{ d(\alpha, \beta) : \alpha \in S_{\tau_1}, \beta \in S_{\tau_2} \right\} \right) = f(d(S_{\tau_1}, S_{\tau_2})),$$

for $f$ a function of the $(p - l_1)!(p - l_2)!$ distances $d(\alpha, \beta)$ such that on sigletons $f(\{t\}) = t$. 
EXTENSIONS TO PARTIAL LISTS WITH DIFFERENT LENGTH

In particular, for $d = Ca$ the Canberra distance and $f$ the mean function:

$$
Ca(L_1, L_2) = \frac{1}{|S_{\tau_1}|} \frac{1}{|S_{\tau_2}|} \sum_{\alpha \in S_{\tau_1}} \sum_{\beta \in S_{\tau_2}} Ca(\alpha, \beta)
$$

$$
= \Lambda \sum_{\alpha \in S_{\tau_1}} \sum_{\beta \in S_{\tau_2}} Ca(\alpha, \beta)
$$

$$
= \Lambda \sum_{\alpha \in S_{\tau_1}} \sum_{\beta \in S_{\tau_2}} \sum_{i=1}^{p} \frac{|\alpha(i) - \beta(i)|}{\alpha(i) + \beta(i)}
$$

$$
= \Lambda \sum_{\alpha \in S_{\tau_1}, \beta \in S_{\tau_2}} \sum_{i=1}^{p} \frac{|\alpha(i) - \beta(i)|}{\alpha(i) + \beta(i)},
$$

where $\Lambda = \frac{1}{(p-l_1)!(p-l_2)!}$.
EXTENSIONS TO PARTIAL LISTS WITH DIFFERENT LENGTH

Expanding the formula, we get for Ca($L_1, L_2$):

$$\sum_{i \in L_1 \cap L_2} \left( \frac{\left| \tau_1(i) - \tau_2(i) \right|}{\tau_1(i) + \tau_2(i)} - \frac{\Delta(l_2 + 1, p, \tau_1(i))}{p - l_2} - \frac{\Delta(l_1 + 1, p, \tau_2(i))}{p - l_1} \right)$$

$$+ \frac{1}{p - l_2} \left( l_1(p - l_2) - 2\epsilon_{p}(l_1) + 2\epsilon_{l_2}(l_1) \right) + \frac{1}{p - l_1} \left( l_1(p - l_1) : +4\epsilon_{l_1}(l_1) + 2\xi(l_2) - 2\xi(l_1) - 2\epsilon_{l_1}(l_2) - 2\epsilon_{p}(l_2) + (p + l_1)(l_2 - l_1) + l_1(l_1 + 1) - l_2(l_2 + 1) \right)$$

$$+ A \cdot (2\xi(p) - 2\xi(l_2) - 2\epsilon_{l_1}(p) + 2\epsilon_{l_1}(l_2)$$

$$- 2\epsilon_{p}(p) + 2\epsilon_{p}(l_2) + (p + l_1)(p - l_2) + l_2(l_2 + 1) - p(p + 1)),$$

where

$$A = \frac{|\mathcal{F} \setminus (L_1 \cup L_2)|}{(p - l_1)(p - l_2)} \quad \Delta(a, b, c) = \sum_{a \leq i \leq b} \frac{|c - i|}{c + i} \quad \varepsilon_k(s) = \sum_{j=1}^{s} jH_{j+k}$$

Setting $A = 0$ we neglect the (preminent) contribution of the not included features. We call core the formula thus obtained, and complete the original one.
**Distance matrix and distribution**

- Given a set of lists $\mathcal{L} = \{L_t\}_{t=1}^b$ of $p$ genes and an integer computation of all mutual top-$k$ distances leads to the construction of a symmetric **distance matrix** $M_k \in \mathcal{M}(b \times b, \mathbb{R}^+)$.  

- Then the corresponding histogram can be built.

- Often the histogram can be approximated by a gaussian distribution.

---

**Indicators**

It is thus possible to use the **mean** (and the **variance**) of the matrix $M_k$ as measures of the intrinsic distance of the top-$k$ lists of the set $\mathcal{L}$.  

$n = 100, p = 500, b = 100$, SVM-RFE, $D = C$
Intuitively, the more mutually different the lists, the more unstable the problem (either because of the data or because of the employed classification procedure).

**STABILITY CURVE**

- Given a set of list $\mathcal{L}$ on $p$ genes, decide a sequence $K$ of relevant gene subset dimensions.
- For each $k \in K$, compute the mean and the variance of $M_k$ (possibly normalized by $k$).
- Plot those values versus the sequence $K$: this may be computationally heavy, depending on $p$ and on the number of lists.

The above procedure is independent from the source generating the lists and from the dimensions of the gene expression data.
DEFINITION

Let \( \mu_k = \frac{2}{B(B-1)} \sum_{2 \leq i < j \leq B} (M_k)_{ij} \) be the mean of all the \( \frac{B(B-1)}{2} \) non-trivial values of the distance matrix \( M_k \in \mathcal{M}(B \times B, \mathbb{R}^+) \).

**The mean list distance indicator**
is the sequence
\[
I_{D,\mu}(\mathcal{L}) = \{(k, \mu_k): 1 \leq k \leq p\}.
\]

**The no-information curve**
is the sequence
\[
I_{D,\mu}(S_p) = \{(k, E\{C_a^{(k+1)}\}_{S_p}): 1 \leq k \leq p\}
\]
associated to the group \( S_p \) of all possible dual lists with \( p \) features.

An indication of the stability of a list set \( \mathcal{L} \) is given by considering
\[
\hat{I}_{D,\mu} = \left\{(k, \frac{\mu_k}{E\{C_a^{(k+1)}\}_{S_p}}): 1 \leq k \leq p\right\}.
\]
independent from \( p \) and \( k \).
**Predictive profiling on synthetic data**

**Synthetic datasets** $fX/Y$: 50 + 50 binary labelled samples described by $X$ discriminant features from $\mathcal{N}(\mu, \sigma)$ ($\mu$ depending on the sample label) and by $Y - X$ random features from the uniform $\mathcal{U}[a, b]$.

**DAP**: $B = 400$ LinSVM/RFE exps, with correlation modules M1 (>0.8) and M2 (>0.7). On both datasets, less than five features are sufficient to reach perfect classification.

![Graphs showing stability indicators](image)

$I_{D_{\mu}}$ and $I_{D_{\mu}}$

Stability indicators in profiling synthetic data (solid lines: $f10/100$; dotted lines: $f30/100$) and in presence of feature modules.
Classifiers comparison

- Classifiers: Linear SVM / Terminated ramp SVM
- Feature ranking methods: (E–)RFE / 1–RFE
- Datasets: Microarray Breast Cancer - 183 cases described by 22215 gene expression / Proteomic Ovarian Cancer - 160 samples described by 123 mass spec peaks.

Breast Cancer

Average Test Error
LinearSVM/RFE (LR: dashed) / LinearSVM/1RFE (L1: dotted) / TRSVM/1RFE (T1: solid)

Ovarian Cancer

Average Test Error
LinearSVM/RFE (LR: dashed) / LinearSVM/1RFE (L1: dotted) / TRSVM/1RFE (T1: solid)
Classifiers comparison

- Classifiers: Linear SVM / Terminated ramp SVM
- Feature ranking methods: (E–)RFE / 1–RFE
- Datasets: Microarray Breast Cancer - 183 cases described by 22215 gene expression / Proteomic Ovarian Cancer - 160 samples described by 123 mass spec peaks.

Breast Cancer

![Breast Cancer Graph]

Mean distance $I_{D_{\mu}}$

LinearSVM/RFE (LR: dashed) / LinearSVM/1RFE (L1: dotted) / TRSVM/1RFE (T1: solid)

Ovarian Cancer

![Ovarian Cancer Graph]
Classifiers comparison

- Classifiers: Linear SVM / Terminated ramp SVM
- Feature ranking methods: (E–)RFE / 1–RFE
- Datasets: Microarray Breast Cancer - 183 cases described by 22215 gene expression / Proteomic Ovarian Cancer - 160 samples described by 123 mass spec peaks.

Breast Cancer

\[ I_{D_{\mu}} \text{ top-25} \]
LinearSVM/RFE (LR: dashed) / LinearSVM/1RFE (L1: dotted) / TRSVM/1RFE (T1: solid)

Ovarian Cancer

\[ I_{D_{\mu}} \text{ top-25} \]
If a library of functionally correlated gene sets (e.g. pathways) is available, the stability indicator can be used to compare the stability of the original lists with the stability of the ranked gene sets produced by their enrichment.

A dishomogeneous set of lists may be shown to correspond to a much stabler lists of pathways.

Example on synthetic data with 100 lists of 100 genes and 124 pathways, where the lists are created by random permutations fixing the gene sets.
FILTER FEATURE SELECTIONS

Dataset *Tib100*: 50 + 50 samples and 100 normally distributed features, with decreasing discriminant power.

Fix a number $n$ of samples and a suitable threshold $\theta$ for the employed filtering algorithm $A$.

Randomly extract from *Tib100* dataset $n$ positive and $n$ negative examples and compute the $A$ statistics.

The statistics considered are Fold Change (FC), Significance Analysis of Microarray (SAM), $B$ statistics, $F$ statistics, $t$, $\text{mod} - F$ and $\text{mod} - t$.

Consider the list of the features whose value of the statistic is above the threshold $\theta$, ranked by statistic value; repeat the above process for $B$ times.

Final output: set of all lists $L(n, A, \theta, i)$, where the number of samples $n$ ranges between 5 and 45, an ad-hoc sets of 100 values for $\theta$ are chosen and $B = 100$. 
Most researcher agree on the fact that SAM should outperform all other methods because of its ability in controlling the false discovery rate.
Although they share a common trend in many cases, accuracy and stability are independent measures.

Thus we can analyze a dataset in the accuracy vs. stability space.

Left-down direction indicates better performance.

This diagnostic plot allows the comparison of different datasets, different profiling methods (classifiers/feature ranking algorithm) and different models.

ATE vs. $\hat{I}_{D_{\mu}}$ for different profiling methods (LE,L1,T1) and cancer datasets (Breast, Ovarian); each point corresponds to a feature subset size, indicated for extremal models.
Analysis of swapped training/validation experiments.

Performance-Stability analysis for the candidate models at endpoints (both Blind and Swap).

The MCC coordinate is the median of the MCCs for endpoint and same experiment.

\[ \text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \]

For each point, the rectangles are defined by CI (95% bootstrap Student's t).

Stability consistently predicts level of MCC performance, i.e. endpoints for which lists change less (between Blind and Swap) have higher MCC scores.
Analysis of swapped training/validation experiments.

Inter-experiment list stability compared with MCC.

For each pair (edp, exp) the partial top- med(edp, exp) Borda list B(edp, exp) is used in alternate experiment.

Swap experiment is run with features belonging to B(exp,Blind) retrieving a MCCBlind(Swap) value.

Analogously, we compute the MCCSwap(Blind) for the same endpoint.

InterCC on the x axis ranks endpoints for increasing list variability between experiments.

Stability consistently predicts level of MCC performance, i.e. endpoints for which lists change less (between Blind and Swap) have higher MCC scores.
[Borda81] Borda, J.
Mémoire sur les élections au scrutin.

[Critchlow85] Critchlow, D.
*Metric methods for analyzing partially ranked data.*

[Diaconis88] Diaconis, P.
*Group representations in probability and statistics.*

[Dwork01] Dwork, C., Kumar, R., Naor, M. & Sivakumar, D.
In *WWW10*.

[Fagin04] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D. & Vee, E.
Comparing and Aggregating Rankings with Ties.

[Saari01] Saari, D.
*Chaotic Elections! A Mathematician Looks at Voting.*

[Simon06] Simon, R.
*Development and Evaluation of Therapeutically Relevant Predictive Classifiers Using Gene Expression Profiling.*