An overview to distances for gene lists

G. Jurman
(with C. Furlanello, M. Serafini, S. Merler)
ITC-irst - Trento

First Bioinformatics Meeting on Machine Learning for Microarray Studies of Disease
Luserna 4th June 2004
The microarray analysis flowchart

Classification model(s)

Gene ranking list(s)
The list comparison problem

As a major by-product of the application of a complete validation scheme in a classification/ranking process, a number of lists is produced.

The problem is how to derive global information from this set of lists, e.g. deriving a single list for the whole experiment (for instance, by extraction multiplicity or by rank aggregation techniques) and deciding how reliable it is.

Depending on the chosen feature ranking method, an optimal number of genes might even be specified for each list.

The problem is how to derive global information from this set of lists, e.g. deriving a single list for the whole experiment (for instance, by extraction multiplicity or by rank aggregation techniques) and deciding how reliable it is.
Need for a similarity measure

The key step in tackling such issue is defining a suitable measure of similarity among lists.

\[ D(\cdot, \cdot) = ? \]

A reasonable measure evaluating differences among lists would allow to perform statistical analysis and estimate a “distribution” of these lists.
List distributions and stability

The histogram of the mutual distances among lists would help in showing how stable the ranking process is throughout the experiment, and how accurate the chosen panel is – at least by the machine learning point of view.

This results in a measure of “variance” or “entropy” of the experiment in terms of gene ranking.

It would be hard to use it as an absolute indicator, more as a comparison index, for instance among:

• different problems
• different experiments on the same problem
• different moments inside the same experiment (e.g. in the feature selection process)
Different list distributions

Stable: most of the lists are mutually close

Unstable: distances among lists are uniformly distributed

Unstable: many lists are mutually quite far

Lower dots represents distances from the selected gene panel list
The formal framework

The first step to take is to formalize the objects dealt with inside a theoretically sound framework, in order to use the tools already developed in that workfield.

The correct environment where to analyse ranked (sub-)lists is an important section of the group theory, i.e. the theory of permutations.

A permutation on \( n \) objects is any rearrangement, each taken exactly one time. (e.g. anagrams)

Formally, a permutation \( \pi \) on a set \( S_n = \{s_1, s_2, s_3, \ldots, s_n\} \) of objects is a bijective function between \( S_n \) and \( S_n \).

The set \{a,b\} has only two permutations: \( \pi_1 \) and \( \pi_2 \), where \( \pi_1(a)=a \) and \( \pi_1(b)=b \), while \( \pi_2(a)=b \) and \( \pi_2(b)=a \).

The set \{a,b,c\} has 6 permutations.

In general, the number of permutations on \( n \) elements is \( n! \)
Lists representations

If the $n$ genes involved in the problem are labelled by the integer between 1 and $n$, every ranked list is exactly a permutation $\pi$ on the set $\{1, \ldots, n\}$, where the image $\pi(i)$ of the $i$-th gene is its ranking inside the list $\pi$.

\[
\begin{align*}
\text{AFFX-MurIL2\_at} & \rightarrow \text{Gene1} \\
\text{AFFX-MurIL24\_at} & \rightarrow \text{Gene2} \\
\ldots & \rightarrow \ldots \\
99764\_at & \rightarrow \text{Gene12688}
\end{align*}
\]

Then every ranked list

\[
\begin{align*}
1 & \ldots \\
\ldots & \ldots \\
234 & 99764\_at \\
\ldots & \ldots \\
8746 & \text{AFFX-MurIL2\_at}
\end{align*}
\]

becomes a permutation $\pi$ which can be written in several form:

\[
\pi(\text{AFFX-MurIL2\_at}) = \text{rank 8746, } \pi(99764\_at) = \text{rank 234}
\]

\[
\pi = <\ldots, 99764\_at, \ldots, \ldots, \text{AFFX-MurIL2\_at}, \ldots>
\]

\[
\pi(1) = 8746, \ \pi(2) = \ldots, \ldots, \ \pi(12688) = 234
\]
Distance properties

Most of the distances \(d\) that can be calculated among permutations are indeed metrics, i.e. they satisfies the four fundamental axioms:

- **Non-negativity** \(d(\pi, \sigma) \geq 0\) for all \(\pi, \sigma \in S_n\)
- **Separation** \(d(\pi, \sigma) = 0\) iff \(\pi = \sigma\) for all \(\pi, \sigma \in S_n\)
- **Symmetry** \(d(\pi, \sigma) = d(\sigma, \pi)\) for all \(\pi, \sigma \in S_n\)
- **Triangular inequality** \(d(\pi, \sigma) + d(\rho, \pi) > d(\rho, \sigma)\) for all \(\pi, \sigma, \rho \in S_n\)

The first three properties are quite natural, while the fourth one might sometimes be relaxed to some less strict requirements.

Recently, distances \(d\) have been introduced satisfying properties of metric boundedness
- \(c \ d'(\pi, \sigma) \leq d(\pi, \sigma) \leq f \ d'(\pi, \sigma)\) for \(d'\) metric and \(c, f\) constants
- and relaxed polygonal inequality
- \(d(\pi, \sigma) \leq c \ (d(\pi, \rho_1) + d(\rho_2, \rho_1) + ... + d(\rho_n, \sigma))\)

which make \(d\) “near metric”.
Distance properties

Moreover, another property is essential to distances in permutation groups spaces:

• Right-invariance $d(\sigma, \pi) = d(\sigma \rho, \pi \rho)$ for all $\pi, \sigma, \rho$ in $S_n$

This is required to avoid unwanted effects due to the initial gene labeling, which is clearly arbitrary.

For instance, consider the ranked lists

$A = \langle A, C, D, B, E \rangle$ and $B = \langle B, C, D, E, A \rangle$.

Labeling         | A        | B        |
------------------|----------|----------|
A1,B2,C3,D4,E5   | (1,4,2,3,5) | (5,1,2,3,4) |
A2,B5,C4,D1,E3   | (3,1,5,2,4) | (3,5,4,2,1) |

The distance $d(A, B)$ must not depend on the chosen labeling: this corresponds to require right-invariance of $d$. 
Statistical distances

Spearman's rho or Spearman's rank correlation

\[ R(\pi, \sigma) = \sqrt{\sum_{1}^{n} (\pi(i) - \sigma(i))^2} \]

Spearman's footrule

\[ F(\pi, \sigma) = \sum_{1}^{n} |\pi(i) - \sigma(i)| \]

Kendall's tau

\[ T(\pi, \sigma) = \#\{ (i,j) : \pi(i) < \pi(j) \text{ and } \sigma(i) > \sigma(j) \} \]

This is equivalent to compute the minimum number of pairwise adjacent transpositions needed to transform \( \pi \) into \( \sigma \), when they are written in the bracket notation

\[ \pi = < \pi^{-1}(1), \pi^{-1}(2), \ldots, \pi^{-1}(n) > \]
A transposition is a permutation involving only the swapping of two elements.

For instance, the permutation $\pi$ on $S_5$ defined as

$\pi(1)=3 \; \pi(2)=2 \; \pi(1)=3 \; \pi(4)=4 \; \pi(5)=5$ 

($<3, 2, 1, 4, 5>$ in the bracket notation)

is a transposition.

Thus the Kendall's tau distance

$$T(<2,3,1,5,4>,<1,3,2,5,4>)=3,$$

since it involves at least the three pairwise adjacent transpositions

$<2,3,1,5,4> \rightarrow <3,2,1,5,4> \rightarrow <3,1,2,5,4> \rightarrow <1,3,2,5,4>$

Usually, these three metrics are used in their normalized version where $d$ is replaced by $1-2d/\max\{d(\pi,\sigma)\}$: thus they range between $-1$ and 1 and can be regarded as particular instances of correlations.
Other common distances

(Generalized) Hamming distance:

\[ H(\pi, \sigma) = \#\{ 1 \leq i \leq n: \pi(i) \neq \sigma(i) \} \]

This is the number of differently ranked items: probably the easiest possible distance.

Ulam's distance:

\[ U(\pi, \sigma) = n - \text{the length of the longest increasing subsequence in } \sigma\pi^{-1}(1), \ldots, \sigma\pi^{-1}(n) \]

This corresponds to compute \( n \) – the maximal number of items ranked in the same order by \( \pi \) and \( \sigma \).
Other common distances -II

Cayley's distance

\[ C(\pi,\sigma) = n - \text{the number of cycles in } \sigma\pi^{-1} \]

A cycle \( \varphi \) is a permutation which moves a subsets of object as follows

\[ C \rightarrow B \rightarrow F \rightarrow E \rightarrow K \rightarrow A \rightarrow C \]

and leaves all other objects unchanged.

A cycle on two objects is a transposition.

Since any permutation can be uniquely decomposed into disjoint cycles, the distance is well-defined.

Calculating Cayley's distance is equivalent to compute the minimum number of transpositions needed to transform bracketed \( \pi \) into \( \sigma \).
Gene ranking list requirements

These are general-purpose distances, so variations on their formulae should adapt to a wide range of problems.

What are the peculiarities of managing lists of gene rankings?

They include the biological knowledge concerning the involved process and they are produced by means of (several applications of) a learning procedure.

Biological relations among genes

Many of the features in microarray datasets are clones of the same gene: it is reasonable to penalize less a swap between two clones than between two unrelated genes.

Similarly, some genes are linked together by their related behaviour inside the underlying biological process, thus considering them more than just different objects when defining distances might be helpful.
Gene ranking list requirements - II

Classifier related issues

The ranking obtained for each list is built starting from a score that every gene has been given by the classifier. This value should probably be used in weighting the distance among the lists.

In many cases, the obtained ranking is only a partial one: the weight of many genes is so similar that they cannot be considered differently ranked, so they should share a common position inside the list.

Moreover, only greater weights might be considered meaningful: this would lead to (not homogenous) top-\(k\) sublists of optimal genes to compare.
Partially ranked data: the equivalence

The problem of dealing with partially ranked data can still be tackled inside the environment of the theory of permutation groups.

Consider the subgroup $S_{n-k}$ of $S_n$ consisting of all permutations leaving the first $k$ integers fixed.

Define two permutations $\pi, \sigma$ equivalent if and only if $\sigma \pi^{-1}$ lies in $S_{n-k}$, i.e. leaves the numbers $1...k$ unchanged, or equivalently, if there exists $\xi$ in $S_{n-k}$ such that $\pi = \xi \sigma$.

This equivalent relation partitions $S_n$ into equivalence classes.

For any $\pi$ in $S_n$, the equivalence class containing $\pi$ also includes all the permutations \{\xi\pi: \xi in S_{n-k}\} equivalent to $\pi$.

Such equivalence class is call $S_{n-k} \pi$, a right coset of $S_{n-k}$. 
Partially ranked data: the right cosets

The key observation is that to each right coset of $S_{n-k}$ there corresponds an unique partial ranking of $k$ out of $n$ objects and vice versa.

The set of all partial ranking can be identified with the set of all right cosets, called the coset space $S_n / S_{n-k}$ (which is not necessarily a group).

In fact, two full permutations $\sigma, \pi$ of $S_n$ belong to the same right coset of $S_{n-k}$ if and only if $\pi^{-1}(i) = \sigma^{-1}(i)$ for $1 \leq i \leq k$, that is, if they induce the same partial ranking on the indices $1...k$.

In order to extend a distance on fully ranked data to a distance on partially ranked data, it is thus necessary to introduce a procedure to define a metric on the suitable coset space.

The more natural way to perform this extension is by employing the induced Hausdorff metric.
The induced Hausdorff metric

Given a group $G$ and $K$ a subgroup of $G$, if $d$ is a right-invariant metric on $G$, then $d$ induces a right-invariant metric $d^*$ on $G/K$, called the **Hausdorff metric** and defined as follows:

$$d^*(K\pi, K\sigma) = \max \{ \max_{\beta \in K\sigma} \min_{\alpha \in K\pi} d(\alpha, \beta), \max_{\alpha \in K\pi} \min_{\beta \in K\sigma} d(\alpha, \beta) \}$$

The number $\min_{\alpha \in K\pi} d(\alpha, \beta)$ is just the ordinary distance from the fixed permutation $\beta$ to the set $K\pi$, thus $\max_{\beta \in K\sigma} \min_{\alpha \in K\pi} d(\alpha, \beta)$ is the ordinary distance from the set $K\pi$ to the furthest away permutation in $K\sigma$.

Thus $d^*(K\pi, K\sigma)$ represents how much $K\pi$ needs to be expanded in order to include the whole of $K\sigma$ and vice versa.

The apparently unnatural choice of the candidates for the induced distance is motivated by the requirement of keeping the distance right-invariant.
The Hausdorff distances

For each of the six distances previously analyzed it is possible to write down an explicit formula for its Hausdorff induced version. These explicit formulae are quite complex, apart from Hamming and Cayley's distances.

\[ H^*(S_{n-k} \pi, S_{n-k} \sigma) = \# \{ 1 \leq i \leq k: \pi^{-1}(i) \neq \sigma^{-1}(i) \} + \# \{ 1 \leq i \leq n: \pi(i) \leq k \text{ and } \sigma(i) > k \} \]

where the first summand is the exact analogous of the full version, while the second term is the penalization coming from the non-overlap of the two partial rankings.

\[ C^*(S_{n-k} \pi, S_{n-k} \sigma) = k - \# \{ \text{cycles in } \sigma \pi^{-1} \text{ consisting of numbers } < k \} \]

Representation theory is another possible approach in defining partial ranking distances.
Strings

When optimal sublists with different lengths or lists allowing ties in the ranking are to be taken into account, also a completely different point of view may be adopted.

Such lists cannot be considered (strictly speaking) as permutations, since the mapping to the ranking set is neither injective (more items share the same position) or surjective (not all positions are considered).

They are indeed strings, that is juxtapositions of symbols taken from a suitable alphabet, where gaps and repetitions are allowed.

In this environment, a very powerful tool has been developed to compare (and reconstruct) lists, by mean of a technique called dynamic programming.

This distance is called Levenshtein or edit distance.
The edit distance

In its easiest version, it reads as follows:

the edit distance $e(a,b)$ between two strings $a$ and $b$ is the minimum number of editing operations required to transform $a$ into $b$ (or vice versa), where the allowed editing operations are the insertion, the deletion or the replacement of a character inside the string.

$$\text{ALGORITHM} \quad \text{ED}(\quad) = 4$$

$$\text{LOGARITHM}$$

In this formulation, the edit distance is a metric.

When applied to permutations, this distance is strictly related to the Ulam's distance considering the maximal common subsequence of the two sequences.
The computation is usually carried on either through a top-down recursive process or by the construction of a suitable bottom-up table.

Moreover, the traceback along the table allows to effectively reconstruct the difference between the lists.

Different penalization coefficients may be introduced to better tune the edit distance behaviour.

When applied to numerical arguments (i.e. permutation rankings) it relies on a fixed metric of the numerical range (for instance, euclidean distance).

Anyway, some of its properties are not suitable for gene ranking evaluation purposes.
Conclusions

Problem
• Having a distance (or even a metric) to measure lists of gene ranking is a key tool in evaluate ensembles of experiments.

Initial approach
• Many distances have been defined in literature, but none of them is an off-the-shelf solution. Nevertheless, they could represent a good starting point, because of their sound theoretical basis.

Workplan
• We need to explore tuned versions of the known distances or planning new ones from scratch to satisfy both the biological and the machine learning related issues.

• Then a phase of testing both on synthetic and microarray data should follow, based on some fixed milestones.
Thanks!

OpenOffice.org 1.1
Office Suite