A RANKING STABILITY INDICATOR IN BIOINFORMATICS

G. Jurman1, S. Riccadonna1, R. Visintainer1,2, G. Guzzetta1,2 and C. Furlanello1

1 Fondazione Bruno Kessler, Trento, Italy - 2 DISI, University of Trento, Trento, Italy

INTRODUCTION

We propose a novel method for quantitatively evaluating the degree of stability of a given set of ranked lists of biomarkers, for instance produced by selection algorithms within a profiling task. Although some steps have been recently carried out, the construction of a stability theory for feature selection is still an open problem; our approach yields a sound mathematical formalization within the framework of metric methods for permutation groups, together with an efficient procedural implementation. Besides the usual classification performance measures, we propose to evaluate models and their parameters in terms of stability, thus producing a pair of scores from the set of replicated experiments obtained by the intensive resampling required to control overfitting effects such as the selection bias. As list dimension measure, we used the Canberra distance, which takes into account that variations in the bottom part of the lists are much less relevant than differences in the top part, a key requirement for gene (or SNP) lists derived from a diagnostic classifier. This procedure has been applied in several studies on high-throughput arrays (gene expression, CNVs, SNPs).

THE CANBERRA DISTANCE

The Canberra distance Ca(\(\gamma\), \(\nu\)) is a weighted version of the l1 distance.

\[
    Ca(\gamma, \nu) = \sum_{i=1}^{n} \frac{|\gamma_i - \nu_i|}{\max(\gamma_i, \nu_i) + |\gamma_i - \nu_i|}
\]

- Introduced as a software matrix [5].
- Naturally extending to a right-invariant matrix on \(S_n\) as other classical measures [2-3].
- Differently penalizing rank differences in the higher portion of the list.
- Evolving harmonic numbers \(H_L\).
- Approximations through the formula:

\[
    Ca(\gamma, \nu) = \sum_{i=1}^{n} \frac{|\gamma_i - \nu_i|}{\max(\gamma_i, \nu_i) + |\gamma_i - \nu_i|}
\]

Extensible to partial lists via Hausdorff topology and quotient groups to deal with feature modules (e.g., groups of correlated features) [4].

LISTS WITH DIFFERENT LENGTH

The degree of self-homogeneity of a set of ordered lists is gauged by the matrix of the all distances between the lists. The list stability indicator can be defined as the mean of such matrices, being the distance of the distances asymptotically normal by Hofling’s theorem. For two lists \(L_1, L_2\), with \(|L_1| \neq |L_2|\), the key formula \(Ca(L_1, L_2)\) can be split into two main components: one taking care of the elements occurring in the selected lists, and the second one considering the remaining elements of the complete set of features the experiment started from. In particular, this second component is independent from the positions of the selected lists. In the lists using only the first component, a different stability measure (called the Canberra core) is obtained.

\[
    Ca(\gamma, \nu) = \sum_{i=1}^{n} \frac{|\gamma_i - \nu_i|}{\max(\gamma_i, \nu_i) + |\gamma_i - \nu_i|}
\]

- For \(\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_L)\), we have \(M = M(\beta) = [2, \beta]\), where \(M = Ca(\gamma, \nu)\).
- The corresponding histogram \(His(M)\) can be built asymptotically Gaussian.
- Then the mean of the matrix \(M\) measures the intrinsic mutual distances of \(L\).

LISTS DISTANCES FOR MODEL SELECTION

Model selection: detecting optimal number of features in a binary classification task.

- The Setlur dataset [9]:
  - Prostate cancer patients
  - Custom Illumina DASIL Assay with 6414 gene expression values
  - Two cohorts: Swedish WW, 63 (+) and 292 (-) and US PHIL, 41 (+) and 60 (-)
  - Available on GEO: GSE4802

- The model with 15 features has a better list similarity measure than the models with 50 and 100 features which reach the best AUC, without paying too much in terms of performance.
- The best possible compromise between similarity and performance is represented by the model with 25 features.
- The top-5 lists have small Canberra distance, but the selected features show relatively worse predictivity.
- The models with more variables are characterized by a higher level of dissimilarity, which is however not heavily reflecting on the corresponding AUC values.

SIMILARITY AMONG CMAP SIGNATURES

Quantification of similarity among sets of ranked lists on the Broad Connectivity Map (CMAP) [8], a collection of genome-wide human cell lines data treated with bioactive small molecules, for the discovery of functional connections between drugs, genes and diseases.

- 6,100 lists \(n = 22,283\) genome-wide transcriptional expression data (HG-U133A).
- Ranked by differential expressions (compound treated/unreated cell lines).
- 1,369 compounds, concentration from 10 \(^{-2}\) to 10 \(^{-4}\) M, 5 tumor cell lines
- 18,691,952 mutual Canberra distances - using mpyl library on HPC.
- Focusing on Haloperidol: LX319M, Tanspeimycin and Trichostatin A.

FILTER FEATURE SELECTIONS

- Dataset TB100: 50 - 50 samples and 100 normally distributed features, with decreasing discriminant power.
- Fix a number \(n\) of samples and a suitable threshold \(\beta\) for the employed filtering algorithm \(A\). Randomly extract from TB100 dataset \(n\) positive and negative examples and compute the \(A\) statistics.
- The statistics considered are Fold Change (FC), Significance Analysis of Microarrays (SAM), \(\beta\) statistics, F statistics, \(\tau\), mod – \(\tau\) and \(mod – \tau\). Consider the list of the features whose value of the statistic is above the threshold \(\tau\), ranked by statistic value; repeat the above process for \(\beta\) times.
- Final output: set of all lists \(L(n, \beta, \tau)\), where the number of samples \(n\) ranges between 5 and 45, an ad-hoc sets of 100 values for \(n\) is chosen and \(\beta = 100\). (White pixels: all features are discarded from all lists: \(n = 0\).)

ACCURACY AND STABILITY

In the US-FDA led initiative MAQC-II [7], the stability indicator was applied coupled with accuracy metrics to evaluate potential sources of the impact as bias or pre-processing or normalization methods, by evaluating 15 endpoints (A\(M\)) on 6 datasets, with swapped training/validation.

- Inter-experiment list stability with MCC.
  - For each pair (adp, esp) the partial top- medianp, esp) Borda list (Borda, esp) is used in alternate experiments.
  - Swap experiment is run with features belonging to the (Borda) with a MCC (Borda, esp) value.
  - Analogously, we compute the MCC (Swap, esp) for the same endpoint.
- InterCC on the \(x\) axis ranks endpoints for increasing list variability above this threshold.

A correlation between stability and predictivity in the MAQC-II has been detected both in training and validation sets: the more similar the signatures, the better the average predictions.

FURTHER APPLICATIONS IN OMIC/GENOMICS

- For NGS data, together with measures of alignment accuracy such as RPKM or Single Base Mismatches matrices.
- Quantitative phenotype modeling, as selection criterion in L12 regularization.
- Analysis of gene enriched list.
- Evaluation of the stability of the reconstruction algorithms for molecular networks such as the Gene Regulatory Networks.

REFERENCES

1. Luce & Williams Mixed Data Classification Programs 1 - Agglomerative Systems, Axel Comput J. 1967
3. Dacunha & Gehan’s Spearman’s Test as a Measure of Distance, J R Stat Soc B, 1977