Robust biomarker identification for cancer diagnosis using ensemble feature selection methods

Thomas Abel (1,2), Thibault Helleputte (3), Yves Van de Peer (1), Pierre Dupont (3) and Yvan Saeyes (1)

1) Department of Plant Systems Biology, VIB, Gent, Belgium and Department of Molecular Genetics, Ghent University, Gent, Belgium – http://bioinformatics.psb.ugent.be
2) Broad Institute of MIT and Harvard, Cambridge, USA
Contact: yvan.saeyes@ugent.be

Biomarker discovery is an important topic in biomedical applications of computational biology, including applications such as gene and SNP selection from high dimensional data [1]. Surprisingly, the stability with respect to sampling variation or robustness of such selection processes has received attention only recently. However, robustness of biomarkers is an important issue, as it may greatly influence subsequent biological validations. In addition, a more robust set of markers may strengthen the confidence of an expert in the results of a selection method.

Our first contribution is a general framework for the analysis of the robustness of a biomarker selection algorithm. Secondly, we conducted a large scale analysis of the recently introduced concept of ensemble feature selection [2], where multiple feature selections are combined in order to increase the robustness of the final set of selected features. We focus on selection methods that are embedded in the estimation of support vector machines (SVMs).

Measuring robustness

Subsampling approach

- Full data set (100% of the samples)
- SUBSAMPLING
- 90% of data
- Ranked list 1
- Ranked list 2
- …
- Ranked list K

Calculate robustness as average pair wise similarities between all possible signature pairs (f_i, f_j):

Robustness = \frac{\sum_{i \neq j} E(f_i, f_j)}{s \cdot (s - 1)}

with E(f_i, f_j) = \frac{N \cdot s - \#(f_i \cap f_j)}{s \cdot (N - s)}

where

N = \#(f_i \cap f_j)

denotes the number of features (dataset dimensionality)

Ensemble feature selection

Resampling approach (bootstrapping)

- Based on the idea of ensemble classification, ensemble feature selection techniques combine different feature selection techniques.
- Two components are essential: a) to each feature selection process, a method to define multiple diverse feature selection algorithms, and b) a way to aggregate the single feature selection algorithms into an overall ensemble consensus result.

Aggregation operators to create the consensus:

- Linear aggregation: consensus ranking is obtained by summing the ranks over all bootstrap samples
- Weighted aggregation: consensus ranking is obtained by a weighted sum of the ranks over all bootstrap samples. The weights are obtained by evaluating the resulting on the out-of-bag samples.

Experiments and Results

Stability of the baseline method (SVM-RFE [4]) compared to two different ensemble methods, using the Kuncheva Index. Default: 40 bootstraps in the ensemble, 20% of features eliminated at each iteration of SVM-RFE.

Classification performance of the baseline method (SVM-RFE [4]) compared to the linear ensemble aggregation. Bootstrap 1, Bootstrap 2, Bootstrap T.

References/Acknowledgements

All implemented algorithms are available in our open source package JAVA-ML, available at http://java-ml.sf.net

Stability

- Ensemble feature selection (EFS) techniques outperform the baseline largely in terms of feature selection stability
- Especially for smaller signature sizes, the gain in stability of EFS methods increases
- More bootstraps in the ensemble increase the method’s stability

Classification performance

- For the balanced classification rate, ensemble methods perform equally well, compared to the baseline
- For very small signature sizes, ensemble methods outperform the baseline in terms of balanced classification rate

Conclusions

The Idea

Experimental results for the stability of the baseline method (SVM-RFE) compared to two different ensemble methods, using the Kuncheva Index. Default: 40 bootstraps in the ensemble, 20% of features eliminated at each iteration of SVM-RFE.