Sensitivity of Feature Ranked Lists for Biological Data

Krzysztof Fujarewicz[1] **and Danuta Gawel**[1]
[1] Silesian University of Technology, Poland

Present techniques in molecular biology such as DNA microarrays, mass spectrometry, and deep sequencing deliver vast data sets. A common property of these data sets is that the number of features (genes, peptides, etc.) is much greater than the number of samples (observations). This requires careful feature selection as the very first step of supervised data analysis.

Many methods of gene selection have been proposed in the literature, which may be divided into two groups: univariate methods and multivariate methods. Univariate methods grade all genes separately by using statistical methods, and create a gene ranking in which the top genes are assumed to be the most informative (discriminative). Multivariate methods, roughly speaking, try to find the best gene set instead of a set of the best genes. In this wprk we focus on univariate methods.

Recently, the so-called stability of gene lists (Boulesteix 2009, Jurman 2008) has drawn the attention of scientists dealing with large-scale biological data. Gene list stability is very important from a biological point of view. It is known that many different gene subsets may give comparable predictive power of a classifier trained on these sets, which sometimes leads to confusion and problems with biological interpretation of the genes identified. Although classification quality is associated with stability, the connection is still not well understood. Moreover, a higher stability may not indicate higher quality of classification. For example, we can imagine a ranking method that always chooses the same genes as top-genes; such a gene list would be perfectly stable, but the quality of classification based on those selected genes would be poor. Taking this into account, it is better to use the ranking method that gives more stable lists, while at the same time preserving good predictive power.

The stabilities of gene lists are measured by introducing a perturbation to the original data set and examining how the content of the top-gene list and its order were changed. When the list created on the basis of the perturbed data set is exactly the same as that formed on the basis of the original data, we say that this gene list is perfectly stable. The data may be perturbed in different ways; typically the original data set may be re-sampled, using for example a bootstrap technique, or the data may be changed by adding noise to the data matrix. Both these methods are computationally intensive because they require generation of many altered lists (as a result of data perturbation) and comparing them to the original list. Perturbation methods may also affect the results, and among other things this was the reason for the creation of an approach presented here.

The presented method (Gawel 2013), unlike other existing methods of gene list stability analysis, does not require any data perturbation. This results in significant reduction of computational time. The basic idea of the approach proposed here is to calculate the sensitivity of the statistics used for creation of a gene ranking with respect to changes of the feature values (for example, gene expression). Then an aggregate sensitivity index for the whole data set is defined, whose value specifies the average percentage of data variation that may change the list order. The proposed method usually gives different results of stability than existing stability indexes, i.e. indicates different ranking methods as the most stable. This is because of different numerical approach to data analysis (lack of data perturbation). To examine this differences a biomedical analysis of genes from obtained lists was performed. The approach was tested on two DNA microarray data sets, a colon data set and an ovarian data set. For both data sets the sensitivity index indicated as the most stable lists of genes with stronger association with the particular disease than other methods.

Acknowledgement. This work was supported by the Polish National Science Center under grant DEC-2012/05/B/ST6/03472.

Keywords: Feature selection, feature ranking, gene list stability, sensitivity, large-scale data, DNA microarrays.

References:

Boulesteix A., M. Slawski (2009). Stability and aggregation of ranked gene lists, *Briefings in Bioinformatics*, *10*, 556–568.

Jurman G., S. Merler, A. Barla, S. Paoli, A. Galea, C. Furlanello (2008). Algebraic stability indicators for ranked lists in molecular profiling, *Bioinformatics*, *24*, 258–264.

Gawel D., K. Fujarewicz (2013). On the sensitivity of feature ranked lists for large-scale biological data, *Mathematical Biosciences and Engineering*, *10(3)*, 667–690.