SUCCESSFUL R&I IN EUROPE 2019

EUROPEAN NETWORKING EVENT

on 14 and 15 February 2019 at Van der Valk Airporthotel Düsseldorf

Efficient and precise exploration of large data sets

Leon Bobrowski

Faculty of Computer Science, Bialystok University of Technology, Bialystok
Institute of Biocybernetics and Biomedical Engineering, PAS, Warsaw, Poland
e-mail: l.bobrowski@pb.edu.pl

Data mining tools based on minimization of the *convex and piecewise linear* (*CPL*) criterion functions are developed and applied in our group [1]. The considered family of the *CPL* criterion functions is linked to the concept of the ***linear separability*** of multivariate data sets and to the ***perceptron*** model of neuronal networks. The optimal vectors of parameters found as a result of the *CPL* criterion functions minimizing can be used, among others, for:

- data-driven design of linear classifiers and hierarchical neural networks
- designing prognostic models based on *interval regression* or the *ranked egression*

Using the *CPL* criterion functions in analysis of large, multidimensional data sets is based on computational techniques, which is called the *basis exchange algorithm*. The basis exchange algorithm is similar to the *Simplex* algorithm from linear programming and is used in efficient and precise minimization of the *CPL* criterion functions.

High efficiency of the *CPL* procedures allowed, among others, to use the *RLS* method for selection of optimal **gene subsets.** For example, the *RLS* method allowed to extract the *diagnostic key* of n_1 = **12** genes from n = **24481** genes of the *Breast Cancer* data set [2]. This diagnostic key separates correctly 46 cancer from 51 non-cancer patients collected in this data set.

We are interested in cooperation with researchers who collect large experimental data sets in order to solve important practical problems. It means primarily, but not exclusively, researchers in the fields of medicine or biology.

Research problems based on exploration of genetic data sets may be of particular interest to us. We are currently preparing new computational tools for collinear patterns extraction and for modelling multiple interaction between many genes [3].

The basis exchange algorithms are used by us in a new method of large matrices inversion [4] or in the eigenvalue problem solution linked to the collinearity models [5].

We have gained some experience in cooperation with clinical doctors from Karolinska Institutet, Stockholm, Sweden. The *RLS* method of feature selection has been applied to real clinical data set *MIA* describing inflammatory status of patients on dialysis [6]. This data set contains both genetic and phenotypic (environmental) features. The data analysis performed by us demonstrates, among others, the complementary roles of genetic and environmental features in prognostic modelling of a given patient condition.

Bibliography

1. L. Bobrowski; *Data Exploration and Linear Separability*, pp. 1 – 172, Lambert Academic Publishing, 2019
2. L. Bobrowski and T. Łukaszuk; ″Relaxed Linear Separability (*RLS*) Approach to Feature (Gene) Subset Selection″, pp. 103-118 in: *Selected Works in Bioinformatics*, Xuhua Xia (Ed.), *InTech*, 2011
3. L. Bobrowski, P. Zabielski; ″Models of Multiple Interactions from Collinear Patterns″, pp. 153-165 in: *Bioinformatics and Biomedical Engineering* (*IWBBIO* 2018), Eds.: I. Rojas, F. Guzman, LNCS 10208, Springer Verlag, 2018
4. L. Bobrowski; ″Large Matrices Inversion Using the Basis Exchange Algorithm″, pp. 1-11 in: *British Journal of Mathematics & Computer Science*, 21(1), 2017
5. L. Bobrowski; ″Eigenvalue Problem with the Basis Exchange Algorithm″, pp. 1-12 in: *Journal of Advances in Mathematics and Computer Science*, 23(6), 2017
6. L. Bobrowski et al.; ″Selection of Genetic and Phenotypic Features Associated with Inflammatory Status of Patients on Dialysis Using Relaxed Linear Separability Method″, *PLOS ONE*, 2014