Intellectual Data Analysis: Abstracts of the International scientific conference / Crimea scientific center NAS of Ukraine. Simferopol. 2008. P. 49-50.

USING IMMUNE MARKERS
FOR CLASSIFICATION OF THE CVD PATIENTS

Bray, D., Strijov, V.

ImmunoClin SARL, Centre de Recherche des Cordeliers, Paris, France

dorothy.bray@immunoclin.com

Computing Centre of the Russian Academy of Sciences, Moscow, Russia

vadim@strijov.com

The goal of the investigation is to find an algorithm that successfully separates different groups of patients with Cardio-Vascular Disease. The algorithm must select the most informative features [1]: the markers, which bring the minimal number of the misclassified patients.

Four groups of the CVD-patients were considered: A1 (surgery performed), A3 (risk group) and B1, B2 (healthy groups). Each group contained up to 15 patients. Each patient is described with 20 immune markers. Since the number of the patients in the sample was relatively small, the number of the informative markers must not exceed a few to avoid overtraining. The algorithm must process pairs of the classes.

The problem of the classification was stated as following. Denote by $X = \{\mathbf{x}_i \mid \mathbf{x} \in \mathrm{R}_{+}^{\mathrm{N}}\}_{i=1}^{L}$ the set of the objects (patients) and by $Y = \{y_i \mid y = \pm 1\}$ the set of class labels. The *i*-th patient is described by the vector $\mathbf{x}_i^S = \{x_i^j\}$, where the index $j \in S \subseteq \{1,\ldots,N\}$ shows that *j*-th feature (marker) is included in the description of *i*-th object. Let the classes be linear separable, so that there exist parameters $\mathbf{w} \in \mathrm{R}^N, b \in \mathbf{R}$ such that the plane $(\mathbf{x},\mathbf{w}) + b = 0$ separates the classes properly. This follows from the assumption on the linear dependency of the classes. For example, three markers K, P and K/P are linear dependent, since we assume $\mathrm{K/P} = \alpha\mathrm{K} + (1 - \alpha)\mathrm{P}$.

One must find the plane that brings the number of the misclassified objects $\upsilon = 1/2\sum_{i=1}^{L} | \bar{y}_i - y_i |$ no more than given υ^* and the number of the features $| S | \to \min$. Here $\bar{y}_i = \mathrm{sign}((\mathbf{w}, \mathbf{x}_i^S) + b)$.

To find the most informative features, consider all combinations of them. The exhaustive search starts from $| S |= 2$ and stops when $| S |> 8$. For each subset S of the features the SVM [2] classification algorithm finds the optimal plane and thus obtains the parameters $\mathbf{w}, b$.

The following results were obtained. The patients of groups A1 vs. A3 was classified successfully; A3 vs. B1 and A3 vs. B2 were classified with one misclassified patient for each pair. The most informative features for these pairs were selected. It was discovered that the classification for such pairs as A1&A2 vs. B1, A1&A2 vs. B2 and B1 vs. B2 impossible to classify properly with the given data, model and requirement to υ^*.

This project is supported by RFBR, grant 07-07-00181.

References

1. Strijov, V., Kazakova, T. Stable indices and the choice of a support description set / Zavodskaya Laboratoriya. 2007. No 7. P. 72-76.
2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2001.