Multiobjective Classifier Learning for Chosen Decision Tasks

The project is an attempt to employ multi-criteria optimization methods to classifier training, especially for the task where more than one criterion is optimized. This is motivated by the fact that e.g., during imbalanced data classification we wish to obtain a classifier characterized by both high *precision* and *sensitivity* (*recall*), i.e., we deal with the problem where miss-classification cost differs for disparate classes and usually it is not well defined. For two-class tasks the following aggregations of single-class measures are commonly used: harmonic mean of *precision* and *recall* - F_β-score, geometric mean of *precision* and *recall* - G-mean, or arithmetic mean of *specificity* and *sensitivity* - BAC (balanced accuracy).

During the realization of the previous projects, it was observed that the use of such learning criteria leads, in the case of imbalanced data, to loss of information about the model's preferences for classes, because the same aggregated metric value can be obtained for different single-class measures. Secondly, optimizing the model for aggregate metrics, the user's preferences are not taken into account, because in the case of single-criteria optimization, the choice of a particular classifier setting is made arbitrarily. This problem can be overcome by assigning a miss-classification cost for each class, but in practice determining this cost is usually very difficult for users, and the research of alternative solutions like *utility-based learning* is still the focus of intensive research.

The need for multi-criteria optimization is not limited to the imbalanced data classification. Similar observations can be made for classifier ensemble learning during which we try to select classifiers with, on one hand, high prediction quality, and on the other hand with a high diversity of the ensemble. We have to consider also cost-sensitive classification, where the cost of feature acquisition is specified, and it is necessary to balance the discriminative power of the attributes and their cost of acquisition. This problem is common especially in the case of medical diagnostics, where, on the one hand, we are looking for highly discriminative features, while on the other hand, we must consider the cost of obtaining their value. Finally, taking into consideration the canonical classifier learning task, one may notice that the *reqularization* is commonly used, which is aimed at balancing the quality of prediction on training data with model complexity to protect against the *overfitting*. As we can see, the regularization is also a two-criteria optimization task. In most approaches, the problem of multiple criteria is simplified by constructing an aggregate objective function taking into account single criteria (e.g., as their linear combination). Nevertheless, the following shortcomings of using aggregate criterion could be enumerated: (i) loss of information on the relationship between the constituent criteria that leads to difficulties in interpreting aggregate metrics, (ii) ignoring user preferences, because the model is chosen arbitrarily.

The project will focus on the possibility of overcoming the above difficulties by using multi-criteria optimization methods, returning a set of Pareto-optimal solutions, enabling the user to select a specific classification model, proposing automatic methods of its selection, or aggregation of acceptable models using the combined classification paradigm. In this project, we form a hypothesis that:

It is possible to propose classifier learning algorithms using multi-criteria optimization, returning a set of Pareto-optimal models, with individual prediction quality at least as good as the quality of classifiers trained using aggregated criteria.

During the project, classifier learning methods using multi-criteria optimization will be developed, returning a set of representative, Pareto-equivalent solutions with the highest degree of diversity and the highest quality of individual solutions. We will propose methods for training individual classifiers as well as pruning and forming classifier ensembles based on multi-criteria optimization, which will then be adapted to non-stationary data stream classification and online optimization. The developed methods will also be used in selected decision tasks, such as the classification of imbalanced data, classifier ensemble learning, preventing *overfitting*, as well as feature extraction and selection.