DESCRIPTION FOR THE GENERAL PUBLIC

This project covers the topic of designing efficient machine learning methods for the multi-class scenarios suffering from uneven distribution of training samples in classes. Typically supervised learning methods are designed to work with reasonably balanced data set, but many real world applications have to face imbalanced data sets. A data set is said to be imbalanced when several classes are under-represented (minority classes) in comparison with others (majority classes). The problem of imbalanced data is usually found the wide range of the practical application as in banking (fraud detection), computer security (IDS/IPS or spam filtering) or medicine.

The presented literature survey allows us to conclude that there is a need to develop novel methodologies for handling multi-class imbalanced problems and exploring the characteristics of examples within class structures. Learning from imbalanced data is among the contemporary challenges in machine learning and multi-class imbalance stands out as the most difficult scenario. In binary imbalanced learning the relationships between classes are easy to be defined: one class is the majority one, while the other is the minority one.

In this project we form a hypothesis that **it is possible to design efficient multi-class methods for such compound imbalance problems that could process all of classes at once**.
We plan to identify general rules for designing efficient methods for learning from multi-class imbalanced data, proposing novel algorithms for this task and developing dedicated software packages that could be used in this area of research.

For evaluating the quality of the proposed methods we will use mainly the experimental investigations. Currently the analytical approach for learning from imbalanced data is highly limited due to the number of simplifications that must be undertaken. All experiments will be done mainly in KNIME environment (that allows to develop and join software from Java, R and Matlab languages).
The following task will be performed:

- Developing methods devoted to a local difficulty of imbalace data for multi-class classification task
- Proposition of new algorithms for data prepocessing to decrease the imbalance ratio for multi-class classification task
- Developing new methods of imbalance data classification for multi-class classification task
- Proposition of new imbalance data classification methods based on classifier ensemble approach for multi-class classification task
- Computer implementations and experimental evaluation of proposed methods devoted to imbalanced data for multi-class classification task

The usage of mentioned methods in analysis of imbalanced data for multi-class is currently underrated. The proposed research tasks aim at filling this gap. Therefore the algorithms which will be developed during this project could be used by the companied related with data analysis.