Additional information in data clustering and related topics

Clustering is one of the most important machine learning techniques, which has been applied in various branches of data analysis. Although clustering is an unsupervised method, which means that it does not require any training examples labeled by the expert, the specification and the use of additional knowledge about data can significantly influence on the final results. Let us recall that one of machine learning paradigms states that one should take into account all existing information in building a learning framework. This project focuses on the verification and the application of this paradigm in data clustering and related research areas including classification and data representation.

In the recent years, there has been a rapid development of semi-supervised (using additional knowledge) clustering and classification methods. Nevertheless, most of clustering methods do not use additional knowledge for the selection of complexity of cluster model. In other words, constructed methods rely on direct adaptation of existing clustering algorithms to the semi-supervised case. In consequences, created groups are described by simple probability models, even if it is inconsistent with imposed additional information. The innovative element of this project is an alternative view on this problem: in semi-supervised clustering we should deduce the complexity of the model based on both, a data distribution and an introduced additional knowledge.

The most popular way to introduce additional information in clustering relies on defining pairwise equivalence constraints, which specify pairs of elements that have to be grouped together, or partial labeling, which assigns class labels for selected instances. However, in real-life applications, we are often given different types of additional knowledge from a partial labeling or pairwise constraints, which makes a direct application of existing semi-supervised clustering methods infeasible. For instance, in cheminformatics the compound's activity, which we want to predict, is given by an imprecise measurement, which cannot be directly transfered to class label. Moreover, in medical treatments in addition to examinations of healthy and ill patients we also have the information of patients with intermediate diagnosis. According to the concept of the project, the use of all existing information about data can influence on their better understanding and analysis. These examples and other practical cases show that there is highly insufficient number and variety of developed semi-supervised learning models. Based on practical problems we will define novel clustering and classification models, which could fill the gap in semi-supervised learning area.

Acquisition of data labels (or defining equivalence constraints) significantly increases the cost of data analysis. Therefore, in practice we usually have only a small amount of training examples, while the unlabeled data is available cheaply. Most of existing semi-supervised clustering methods cannot use such a small number of labeled examples in an appropriate way and they create partitions like in the unsupervised case. In this project, we plan to develop clustering methods, which could be guided by a small amount of labeled examples or equivalence constraints.

Although our studies are often motivated by real-life problems, the project is not restricted to a single discipline. We will create theoretical foundations of constructed methods and test them in various practical problems. We hope that this will allow to obtain interesting results in computer science as well as to solve important practical problems in related disciplines.