DESCRIPTION FOR THE GENERAL PUBLIC

Machine learning is one of the most rapidly developing subfields of artificial intelligence and computer science. It concerns the design and development of algorithms that allow computers to act based on experience gained from empirical data, without being explicitly programmed. It is strongly related to statistics and probability theory, however, the computational issues play here a more important role. The machine learning algorithms are widely applicable, from search engines and Internet technologies to life and natural sciences.

Many current applications of machine learning are characterized not only by a huge number of training examples and features, but also by a very large number of classes (labels) to which examples can be assigned. The learning problems of this type are often referred to as extreme classification. As an example, consider the problem of tagging Wikipedia articles. In this case, each article is an example, words appearing in the text of articles can be considered as features, and categories to which articles are assigned as labels. By creating a data set from the current content of Wikipedia, we very easily end up with an enormous problem with millions of examples and features, but also with more than million of labels, i.e., so many categories are used in Wikipedia today. The challenges posed by problems at that scale opened a new line of research within machine learning

The number of labels in extreme classification applications may vary from a few thousand to several million. Therefore, extreme classification requires new algorithmic solutions, since standard methods cannot scale to a such large number of labels. In this project, we aim at reducing the time and space complexity of learning algorithm without compromising the predictive performance. To achieve this goal, we consider a new approach that consists in transforming the original extreme classification problem to a specific structured prediction problem. For example, a given label can be encoded as a sequence of bits and then the problem can be treated as sequence labelling. Finding the optimal structure to encode the labels is one of the particular tasks of the project. Some other will concentrate on fast learning approaches and relations to existing extreme classification algorithms.

The reduction of extreme classification to structured prediction is a very attractive research direction. One of the main reasons why we undertake this scientific problem is the fact that preliminary research has shown promising results. There is, however, still a need of better understanding of the problem. The algorithms following this approach need to be further investigated to make them widely used for solving extreme classification problems. It is worth noting that applications of extreme classification will become more and more frequent. Moreover, extreme classification problems are complementary to representation learning being the main purpose of deep learning. Therefore, the research problem discussed here is of more general nature and its solution can be used in other areas of machine learning.