Feature selection in multi-label classification

Paweł Teisseyre
Institute of Computer Science PAS, Warsaw
teisseyrep@ipipan.waw.pl

1. Project Description

Multi-label classification (MLC) is an active research field in machine learning. In multi-label classification we consider many target variables (labels) simultaneously. The main objective is to build a model which predicts labels based on the characteristics (features) of the considered objects. Recently, MLC has attracted a significant attention in many research domains, e.g. image annotation, text categorization, marketing, genomics, medicine and drug design (Gibaja, E. and Ventura, S. (2015) „A tutorial on multi-label learning"; Zhang, M. and Zhou, Z. (2013) „A review on multi-label learning algorithms"). An important task in MLC is feature selection, i.e. finding important features that may affect labels. The aim of the project is to develop and implement feature selection methods in the case of high dimensionality of the feature space. In recent years a variety of novel algorithms have been proposed. Most of the methods utilize dependences between labels to improve the classification performance. However, there are no theoretical and empirical results showing what is the influence of the feature selection on the classification performance. We intend to fill this gap in the project. The proposed approaches could improve the performance of the existing algorithms. First, they can improve the prediction power of the existing methods. Secondly, feature selection methods allow to discover dependency structure in the data and thus to understand which features affect labels. Moreover, feature selection is an important step that allows to reduce the computational burden. This is particularly important in the case of high-dimension problems (e.g. text data or genomic data). We also plan to create an open-source library containing the implementations of the considered methods and perform experiments to compare the proposed methods with the existing ones.

2. Requirements (expectations)

a. M.Sc. of Computer Science, Mathematics or Physics
b. Background in machine learning, statistics
c. Programming skills: R, experience with Java is a plus
d. Enthusiasm about solving mathematical and analytical problems
e. Good command of English