Multi-layer Feature Selection Incorporating Weighted Score-based Expert Knowledge toward Modelling Materials with Targeted Properties

Yue Liu[a], Junming Wu[a], Siqi Shi[b, c]

a School of Computer Engineering and Science, Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China

b School of Material Science and Engineering, Shanghai University, Shanghai 200444, China

c Materials Genome Institute, Shanghai University, Shanghai 200444, China

Abstract

Selecting proper descriptors or features is one of the central problems in exploring structure-activity relationships of materials using machine learning models. The current feature selection algorithms are usually tedious in operation, diverse in applicable scenarios, and tend to ignore the prior knowledge of domain experts about features. Here, we propose a data-driven multi-layer feature selection method incorporating domain expert knowledge named DML-FS_{dek} for rational selection of material descriptors. The whole process of our approach is fully automated, with users simply entering training data without too much interaction. Furthermore, domain expert knowledge is quantified by means of weighted scoring and integrated into the selection process of features by our developed strategies, reducing the risk of crucial features being removed. Successfully, empirical studies on eight material properties datasets demonstrate the potential of the approach to automatically search for a reduced feature set with lower root mean absolute errors (RMSEs) than those of utilizing initial material features. Essentially, the most relevant material features, the number of which is far less than the number of original features, are selected to establish a closer and more accurate structure-activity relationship mapping the properties of materials. Our method can perfectly represent the targeted properties of materials with a smaller and interpretable set of material features while ensuring equal or better prediction accuracy.

Keywords: Feature Selection, Machine learning, Domain expert knowledge, Materials properties prediction