論文の内容の要旨

論文題目　Representation Learning for Program Analysis, Testing and Repair

（プログラムの解析、テスト、修復のための表現学習）

氏　名　ロヨラ　ハイフマン　パブロ　サルヴァドール

Most of the work on empirical software engineering has been focused on finding causal relationships between diverse aspects involved in the development process, commonly relying on tools from data mining and machine learning domains where the feature engineering is carried out manually. While these approaches in general have been a contribution to the software development process, our concern is related to the fact that the performance of any predictive model depends heavily on the representation of the data used, and that different data representations can entangle and hide different explanatory factors. Furthermore, the decisions related to how to compute these hand crafted features may incorporate bias and, in some cases, could not be weighted effectively given their complexity. We consider this as a relevant problem in software engineering, given the heterogeneity of the data sources involved, their modalities and types. In that sense, we hypothesize that if we were able to design methods less dependent on feature engineering and hand crafted features, we could improve the effectiveness of quality assessment tasks. Moreover, if we are able to abstract software data to more tractable and flexible representations, we could eventually find more natural ways to combine several aspects of development, which at the end could provide a more holistic perspective on the analysis. To this end, we rely on the representation learning paradigm, which encompasses a set methods whose objective is to automatically learn feature representations from the data. These representations comply with certain characteristics such as smoothness expressiveness and temporal and spatial coherence. Moreover, they privilege the emergence of multiple explanatory factors with a hierarchical configuration and the ability to be distributed. This perspective fits our goal in the sense of providing both theoretical and technical frameworks to structure and construct our study. Therefore, in this thesis we explore methods based on representation learning and design empirical studies intended to explore to what extent building methods based on representation learning can help to assess software quality. In that sense, we embark in a quest to revisit relevant software engineering tasks in the light of a representation learning paradigm, namely, defect prediction, automated testing, program understanding and automatic repair. For each of them, we explore ways to extend or create representation learning based alternatives to compare against the state of the art. Therefore, we are able able to conclude on the pros and cons that the proposed methods could provide. At the core of our research is the concept of program change, as we believe it encodes the temporal and functional characteristics most related to quality.