Global Semantic-based Code Defect Detection

何峙, 21215122，大数据与人工智能

Abstract

Software technology is becoming more and more closely related to all aspects of social life. However, the software development goes on while various vlunerabilities happen inevitably.And the identification and locating of vulnerabilities is very labor-intensive and resource-intensive. How to quickly identify and locate vulnerabilities to improve the stability and security of software operation has become an increasingly serious problem. With the development of deep learning technology, there are some methods that can quickly and automatically identify software vulnerabilities, such as the methods based on abstract syntax tree (call 'AST' for short) or program data flow graph (call 'PDG' for short). Nevertheless, both with these kinds of methods , most of the code organization structure will disappear after being extracting from AST or PDG, which makes it difficult to capture the semantics of interdependence between code elements, which is not conducive for the identification of vulnerabilities. This paper proposes a code vulnerability feature extraction method based on lexical analysis to exploit global semantic dependencies for vulnerability identification. The method mainly embeds all the code elements in every code fragment into vector to predict whether the code has defect or not. We also compare the result which come out from the method using all the code elements or just using portions of the code fragment. And finally we conclude that using all the code elments is more effective for vulnerablity identification.