Machine Learning Methods for Software Vulnerability Detection

Submitted by grigby1 on Mon, 01/21/2019 - 11:16am

Title	Machine Learning Methods for Software Vulnerability Detection
Publication Type	Conference Paper
Year of Publication	2018
Authors	Chernis, Boris, Verma, Rakesh
Conference Name	Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics
Publisher	ACM
Conference Location	New York, NY, USA
ISBN Number	978-1-4503-5634-3
Keywords	buffer overflow, compositionality, Human Behavior, machine learning, Metrics, N-grams, pubcrawl, Resiliency, software metrics, static analysis, suffix trees, vulnerability detection
Abstract	Software vulnerabilities are a primary concern in the IT security industry, as malicious hackers who discover these vulnerabilities can often exploit them for nefarious purposes. However, complex programs, particularly those written in a relatively low-level language like C, are difficult to fully scan for bugs, even when both manual and automated techniques are used. Since analyzing code and making sure it is securely written is proven to be a non-trivial task, both static analysis and dynamic analysis techniques have been heavily investigated, and this work focuses on the former. The contribution of this paper is a demonstration of how it is possible to catch a large percentage of bugs by extracting text features from functions in C source code and analyzing them with a machine learning classifier. Relatively simple features (character count, character diversity, entropy, maximum nesting depth, arrow count, "if" count, "if" complexity, "while" count, and "for" count) were extracted from these functions, and so were complex features (character n-grams, word n-grams, and suffix trees). The simple features performed unexpectedly better compared to the complex features (74% accuracy compared to 69% accuracy).
URL	http://doi.acm.org/10.1145/3180445.3180453
DOI	10.1145/3180445.3180453

Citation Key chernis_machine_2018

buffer overflow Compositionality Human behavior machine learning Metrics N-grams pubcrawl Resiliency software metrics static analysis suffix trees vulnerability detection