EVALUATION OF MACHINE LEARNING ALGORITHMS IN THE CATEGORIZATION OF ANDROID API METHODS INTO SOURCES AND SINKS

Por

WALBER DE MACEDO RODRIGUES

Proposta de Trabalho de Graduação

Universidade Federal de Pernambuco
secgrad@cin.ufpe.br
www.cin.ufpe.br/~secgrad

RECIFE/2018

Resumo

Um programa computa em dados sensíveis e não sensíveis, esses dados seguem um fluxo específico indo de data sources para data sinks. O vazamento de dados acontece quando dados sensíveis chegam sem autorização em sinks, para prevenir isso, técnicas estáticas e dinâmicas de Flow Enforcement garantem que esses dados não cheguem nessas sinks. Para isso, esses métodos usam listas, geradas manualmente, de métodos que sejam sources sensíveis ou sinks, e essa solução é impraticável para grandes APIs como a do Android. Visto isso, uma abordagem usando machine learning foi desenvolvida para classificar esses métodos sources e sinks. O presente trabalho tem como objetivo criar um dataset para avaliar os métodos de classificação mais utilizados e decidir quais os mais apropriados para esse problema.

Abstract

A program computes in either sensitive and non-sensitive data and follows a specific flow, from data sources to data sinks. Data leakage happens when sensitive data is sent to unauthorized data sinks, to prevent that, Dynamic and Static Flow Enforcement techniques ensure that sensitive data reaches those sinks. To prevent data leakage, these methods rely on a list of sensitive data sources and data sinks, this list is hand annotated and is impractical to be made to a huge API such as Android. With that in mind, a machine learning model is used to classify methods into sources and sinks. The present work intends to extend the previous work creating a dataset to evaluate the most used classification algorithms and define which is the most suitable to this problem.

1

Introduction

Every program computes on either sensitive data and non-sensitive data. Sensitive is any data that can be used to identify the user or any private user information private, such as photos, International Mobile Equipment Identity (IMEI) and biometric data. Non-sensitive data is any dynamic information that don't identify the user, oftenly this kind of information if public or shared, such as application source code.

In a application, data follows a specific flow, first is acquired from data sources and will be sent to data sinks (MCCABE, 2003). Data sources, in the context of mobile and IoT devices, is defined as method calls that reads data from shared resources such as phone calls, screenshots, sensor polling data from ambient, device identification numbers etc (RASTHOFER; ARZT; BODDEN, 2014). Data sinks are methods calls that have at least one argument, this argument is non-constant data from the source code (RASTHOFER; ARZT; BODDEN, 2014).The sink can make an interface to the user or system API for communication to other devices, store data etc (VIET TRIEM TONG; CLARK; MÉ, 2010).

Dynamic Flow Enforcement relies on Taint Analysis to track possible sensitive data flow to untrusted sinks. This analysis taints every sensitive data gathered from a source and every other variable that inherit any operation from the tainted data, in the end, if any tainted variable is accessed by a sink method, the information has leaked. During the tracking, there are different methods to enforce in runtime that the data will not leak, FERNANDES et al. (2016) uses virtualization to guarantee that the data will only operate in the controlled environment and SUN et al. (2017) declassifying information before it is computed in trusted methods or if reach a trusted API.

Static Flow Enforcement starts by creating abstract models of the application code to provide a simpler representation (MYERS, 1999), using frameworks like Soot (VALLÉE-RAI et al., 2000). Then, this model will be used in control-flow, data-flow and points-to analysis to observe the application control, data sequence and compute static abstractions for variables LI et al. (2017). These methods are implemented and used in DroidSafe GORDON et al. (2015). JFlow MYERS (1999) inserts statically checked and secured code when the application computes on sensitive data.

Both Static and Dynamic Flow Enforcement techniques require information of which

methods is a source of sensitive data and which is a data sink. This is used to identify if a sink method is truly leaking sensitive data or not. So, lists containing sources and sinks of sensitive data are hand created, but this solution is impractical considering a huge API like the Android API RASTHOFER; ARZT; BODDEN (2014).

Considering that issue, Rasthofer et al. RASTHOFER; ARZT; BODDEN (2014) propose using machine learning to utomatically create a categorized list of sources and sinks methods to be used in systems. The list consists in methods classified into Flow Classes and Android Method Categories. The Flow Classes are source of sensitive data, or just source, and sink of data, but also, the method can be neither source or sink. For Android Methods Categories, there are 12 different classes: account, Bluetooth, browser, calendar, contact, database, file, network, NFC, settings, sync, a unique identifier, and no category if the method does not belong to any of the previous.

This is achieved by utilizing Support Vector Machine (SVM) to classify and categorize the methods, the authors had shortly compared Decision Trees and Naive Bayes with the SVM. After a 10 k-fold, the models are compared and the best has been the SVM, which was selected to create the source and sink list.

To classify, the authors utilize features extracted from the methods, like the method name, if the method has parameters, the return value type, parameter type, if the parameter is an interface, method modifiers, class modifiers, class name, if the method returns a value from another source method, if one parameter flows into a sink method, if a method parameters flows into a abstract sink and the method required permission.

To categorize the methods, were used features like class name, method invocation, body contents, parameter type and return value type. After that, the methods list is generated containing if it is a sink, source and the method category.

2

Methodology

The methodology used to create the database will be to use the known feature extractor developed by RASTHOFER; ARZT; BODDEN (2014), this will extract meaningful information from the Android API methods. These features are semantic and syntactic features, containing information about the method name, parameters, return, method and classes modifiers, class modifiers, if exists data flow in the method return or parameters and the required permissions.

The extractor will be used in many Android APIs as possible, since there are a low quantity of hand annotated methods, it is important to extract as many methods from classes as possible. It is possible to have duplicated methods at the end of this evaluation due to backward compatibility, so, methods from older APIs will be overwritten.

The methods used in the comparison will be SVM, Naive Bayes, Decision Trees, MLP, KNN, and ensemble classifications methods, which will be evaluated in 30 different datasets sampled from the original. Each of these datasets are subdivided into train dataset and test dataset, containing 80% of the original dataset for model training and 20% for test the model effectivity. They must maintain the classes proportion observed in the original dataset, if the original has 45% of source methods, the train and test must have a proportion close to that. Create these datasets are important to make a Hypothesis Test, this will help to statically prove if a classification method is really effective when applied in this dataset.

$$precison = \frac{TP}{TP+FP} \quad (2.1) \qquad accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.2)$$

$$recall = \frac{TP}{TP+FN} \quad (2.3) \qquad F1 = \frac{2*recall*precision}{recall+precision} \quad (2.4)$$

Each classifier will be evaluated using precision, accuracy, recall and F1 score. The precision, equation 2, is the ratio of correctly predictions to the total predictions done. Accuracy,

equation 2, is the ratio of correctly true predictions to the total of true predictions. Recall, equation 2 is the ratio of correctly positive predictions to all the predictions of a class. F1, equation 2, score represents the harmonic mean between precision and recall SASAKI et al. (2007). We can rewrite precision, accuracy, recall and F1 score in the equations above, using true positives (TN), true negatives (FN), false positives (FP) and false negatives (FN).

3

Objectives

This work has two main objectives: generate a public database containing methods from Android APIs and evaluate the performance of the most used classification algorithms on this database. The methods should be classified into any of the two Flow Classes, source or sink, or neither.

The motivation behind the database creation comes when RASTHOFER; ARZT; BODDEN (2014) developed the initial classification work. Their objective where create a machine-learning solution to identify sources and sink methods from the code of any Android API. So, every time that you wanted to test other classification algorithms, you had to extract the methods and features at every execution. Then, create a public database extract knowledge and develop better solutions to this problem is very likely.

The objective of evaluate different classification algorithms comes as RASTHOFER; ARZT; BODDEN (2014) shows results for only three classification algorithms, SVM, Decision Tree and Naive Bayes. From that, emerged the question if other classification algorithms have better results in this database. So, the second objective is to evaluate the most used classifiers in the literature, SVM, Naive Bayes, Decision Trees, MLP, KNN, including ensemble classifications methods.

4

Schedule

	Period															
Activity	August		September				October				November				December	
Literature Review	x	x	x													
Dataset creation			x	x	x	x										
Algorithm studies						x	x	x								
Experiments							x	x	x	x						
Evaluation									x	x	x					
Writing										x	x	x	x	x	x	
Presentation preparation															x	x

Referências

FERNANDES, E. et al. FlowFence: practical data protection for emerging iot application frameworks. In: USENIX SECURITY SYMPOSIUM. **Anais...** [S.l.: s.n.], 2016. p.531–548.

GORDON, M. I. et al. Information Flow Analysis of Android Applications in DroidSafe. In: NDSS. **Anais...** [S.l.: s.n.], 2015. v.15, p.110.

LI, L. et al. Static analysis of android apps: a systematic literature review. **Information and Software Technology**, [S.l.], v.88, p.67–95, 2017.

MCCABE, J. D. **Network Analysis, Architecture and Design, Second Edition (The Morgan Kaufmann Series in Networking)**. [S.l.]: Morgan Kaufmann, 2003.

MYERS, A. C. JFlow: practical mostly-static information flow control. In: ACM SIGPLAN-SIGACT SYMPOSIUM ON PRINCIPLES OF PROGRAMMING LANGUAGES, 26. **Proceedings...** [S.l.: s.n.], 1999. p.228–241.

RASTHOFER, S.; ARZT, S.; BODDEN, E. A Machine-learning Approach for Classifying and Categorizing Android Sources and Sinks. In: NETWORK AND DISTRIBUTED SYSTEM SECURITY SYMPOSIUM NDSS. **Anais...** [S.l.: s.n.], 2014.

SASAKI, Y. et al. The truth of the F-measure. **Teach Tutor mater**, [S.l.], v.1, n.5, p.1–5, 2007.

SUN, C. et al. Data-Oriented Instrumentation against Information Leakages of Android Applications. In: IEEE 41ST ANNUAL COMPUTER SOFTWARE AND APPLICATIONS CONFERENCE (COMPSAC), 2017. **Anais...** [S.l.: s.n.], 2017. p.485–490.

VALLÉE-RAI, R. et al. Optimizing Java bytecode using the Soot framework: is it feasible? In: INTERNATIONAL CONFERENCE ON COMPILER CONSTRUCTION. **Anais...** [S.l.: s.n.], 2000. p.18–34.

VIET TRIEM TONG, V.; CLARK, A. J.; MÉ, L. Specifying and enforcing a fine-grained information flow policy: model and experiments. **Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications**, [S.l.], v.1, n.1, p.56–71, 2010.