جامعة الموصل

كلية علوم الحاسوب والرياضيات

بناء نظام لتصنيف عائلة البرمجيات الخبيثة باستخدام التعلم الآلي

سيف الدين حسن علي العبادي

رسالة ماجستير

علوم الحاسوب

بإشراف

م.د. كرم حاتم ذنون البجّاري

1443 هـ

2021 م

الملخـص

تعد البرمجيات الضارة أحد أهم الموضوعات التي تواجه مستخدمي أنظمة المعلومات التي تؤثر على خصوصية البيانات وأمنها، واستخدم الباحثون وشركات انتاج برامج مكافحة البرمجيات الضارة العديد من الطرق التي اعتمدت الحديثة منها على طرق التعلم الالي للتعرف على البرمجيات الضارة من خلال الخصائص المشتركة بين أفراد العائلة الواحدة من هذه البرمجيات. وقد بيّنت الدراسات السابقة في مجال تصنيف البرمجيات الضارة أن طرق التعلم الآلي RF Random Forest، Support Vector Machine SVM وXGB X–Gradient Boost، هي من أفضل طرق تصنيف البرمجيات الضارة التي حققت نسخها المعدلة أعلى مستويات الكفاءة والدقة.

وبالبناء على ما تقدم، قدمت هذه الرسالة نظام للتصنيف يستخدم الطرق الثلاثة ذات الأداء المتميز بنسخها الاصلية دون التعديل عليها، على جزء من مجموعة بيانات VirusShare التي تحتوي على خواص البرمجيات الضارة وأسماء العوائل فيها، والتي ضمت (387) نموذجا موزعا على (8) عوائل، ولكل نموذج (1858) خاصية مميزة للبرمجيات الضارة محفوظة بصيغة ملف CSV، صمم التطبيق باستخدام لغة بايثون V3.8 مع المكتبات اللازمة للتطبيق ومنها مكتبة (!!!!!)،الخاصة بالتعلم الالي جرى تدريب الطرق على مجموعة البيانات التي جرى تقسيمها على أربع نسب تقسيم، وطبق النظام على نموذج واحد من البرمجيات الضارة في المجموعة البيانات للتأكد من فعالية النظام وقدرته على التصنيف، وطبق على عائلة واحدة لذات الغرض، وكانت كفاءة النظام عالية جدا عند تطبيق الطرق منفردة ثم مجتمعة. وجرى تطبيق النظام على مجمل مجموعة البيانات بكامل الخواص المتاحة باستخدام الطرق الثلاثة، وأعيد التنفيذ باستبعاد ثلاث مستويات من الخواص غير المهمة في عملية التصنيف، هي (أقل من 2) و(أقل 6) و(أقل 11)، لتقليل عدد الخواص التي لا تؤثر على كفاءة التصنيف، وقد أظهرت نتائج التجارب أن طرق التصنيف المستخدمة حققت مستويات دقة أفضل في التصنيف باستبعاد الخواص المعتمدة في النظام، حيث حازت طريقة RF على أفضل اداء وبلغت قيمة مقياس F Score لها (0.93125)، تليها طريقة XGBoost وبلغت قيمة F Score لها (0.9275)، وأخيراً طريقة SVM وبلغت قيمة F Score لها (0.925). وتم تطبيق اكتشاف هذه البرامج الضارة عن طريق ادخال صف من الصفات التي تم استخلاصها من مجموعة بيانات البرمجيات الضارة، واستطاع النظام تمييز نماذج البرمجيات الضارة بشكل كفوء.

University of Mosul
College of Computer Sciences
and Mathematics

Build a System for Classification of Malware Family Using Machine Learning

Saif aldeen Hassan Ali Alabadee

M.Sc./Thesis

Computer Sciences

Supervised by

Dr. Karam Hatim Thanon Al-Bajjary

Lecturer

2021 A.D **1443 A.H**

Abstract

Malware is one of the most important issues that face users of information systems and affect data privacy and security. Researchers and anti-malware production companies have used many methods, including modern methods based on machine learning to identify malicious software through the common features between members of the same malware family. Previous studies had indicated that Random Forest, Support Vector machine, and X-Gradient Boost methods are among the best classification methods that their modified versions achieved the highest levels of efficiency and accuracy.

Therefore, this thesis presented a classification system that uses the three methods with excellent performance in their original versions without modification, on a part of the VirusShare dataset that contains the malware features and labels of malware families. The methods were trained on the data set and the system was applied to one malware sample of the dataset, to ensure the effectiveness of the system and its capability. It was applied to one family for the same purpose. The implementation of the system showed high efficiency when the methods were applied individually and then combined. The system was applied to the entire data set with all available features using the three classification methods. It was re-implemented by excluding three levels of uninformative features in the classification process to reduce the number of features that do not affect the efficiency of the classification. The results of the experiments showed that the classification methods achieved better levels of accuracy in classification. Excluding the adopted features the system, the Random Forest method got the best performance and its F-score value was (0.93125), followed by the XGBoost method with its F-score value (0.9275), and finally the SVM method with its F-score value (0.925).