الملخص

لقد تزايدت شبكات الروبوت على نطاق واسع على مر السنين وظهرت هجمات جديدة على السطح ويجب اكتشافها لأهميتها البالغة في كل الأنظمة نظرًا لتأثيرها على مدى أن يكون النظام متوفر في كل الحالات وكذلك احتمال حدوث خسارة مالية كبيرة عند حدوث اي اختراق.

أصبحت أنظمة كشف التسلل الكلاسيكية (IDS) غير كافية للدفاع عن هذه الهجمات وتلبية متطلبات التهديدات الأمنية المتزايدة، وخاصة ضد هجمات ما تسمى بـ (Zero-Day Attacks) اي الهجمات التي ليست معروفة من قبل ولم تكتشف او يتم حلها بعد. يعد التعلم الآلي(Machine Learning) حلاً واعدًا لاكتشاف شبكات الروبوت نظرًا لقدرته على اكتشاف البرامج الضارة الجديدة.

طبق عدة باحثون مختلفون التعلم الآلي في اكتشاف شبكات الروبوتات في الشبكة وتم تحليل مقالات مختلفة تستخدم نماذج التعلم الآلي في اكتشاف شبكات الروبوتات من خلال تلخيص أبحاثهم وتحديد قوة كل مقالة وما هي خوارزميات التعلم الآلي المستخدمة في دراساتهم. ومع ذلك، تحتاج هذه الدراسات إلى تطوير لتحسين أداء أنظمة كشف التسلل من خلال اكتشاف شبكات الروبوت غير المكتشفة في الشبكات. الهدف من هذه الدراسة هو تقديم نموذج محسن للكشف عن شبكات الروبوت باستخدام خوارزميات التعلم الآلي. لقد قمنا بإعداد مجموعة بيانات لجعلها جاهزة للمعالجة ثم اخترنا واستخرجنا الخصائص المميزة للبيانات وذلك لتحديد أفضل مجموعة من الميزات التي تعزز المقاييس وتصنع نماذج أفضل. ثم تعاملنا مع القيم المتطرفة باستخدام طرق مختلفة وهي الخطوة الرئيسية في تحسين نموذجنا. ثم قمنا بتدريب نماذجنا باستخدام أربع خوارزميات مختلفة وهي Support Vector Machine و Random Forest وLogistic Regression و Gradient Boosting. ولتقييم نموذجنا واختباره، استخدمنا مقاييس precision و recall و f1-score لقياس التحسين.

قمنا بتطبيق النموذج بطريقتين مختلفتين. أولاً باستخدام oversampling في أخذ العينات ثم بدونه لمعرفة تأثيره في أخذ العينات على نموذجنا. وأخيراً قمنا بمقارنة نتائجنا مع ورقة الـ benchmark وبين الخوارزميات نفسها للعثور على أفضل خوارزمية مناسبة لهذا النوع من المشاكل. أظهرت النتائج أن نموذجنا حصل على مقاييس محسنة مقارنة بالنماذج القياسية في ورقة الـ benchmark وأن خوارزمية support vector machine أعطت أفضل تحسين بدون ومع الـ oversampling في أخذ العينات والتي حصلت على 18.46% و26.54% على التوالي من حيث قياس الـ recall. أظهرت النتائج أيضًا أن تقنية oversampling في أخذ

العينات التي استخدمناها تؤدي إلى نتائج أسوأ في خوارزمية Logistic Regression ولا يكون لها أي تأثير في كل من خوارزميات both random forest و gradient boosting .

Abstract

Botnets have been widely increased over the years and new attacks have appeared on the surface and needed to be detected as they are crucial in every systems due to the effect they have on system availability and the potential for significant financial loss. Classical Intrusion Detection Systems (IDS) have become insufficient to defend these attacks and meet the demands of growing security threats, especially against zero-day attacks. Machine learning is a promising solution to detect botnets due to its ability to detect new malwares. Different researchers applied machine learning in detecting botnets in network and we explored different articles that used machine learning models in detecting botnets by summarizing their researches and outlining the strength of each article and what machine learning algorithms are used in their studies. However, these studies need to be enhanced to improve the performance of intrusion detection systems by discovering undetected botnets in networks. The aim of this study is to introduce an enhanced model to detect botnets using machine learning algorithms. We prepared the dataset to make it ready for processing then we select and extract features to determine the best set of features that enhance metrics and make better models. We then dealt with outliers using different methods which is the main step in enhancing our models. Then we trained our models using four different algorithms which are Support Vector Machine, Random Forest, Logistic Regression, and Gradient Boosting Algorithms. To evaluate and test our model, we used precision, recall, and f1-score metrics to measure the enhancement. We applied the model with two different ways. First using oversampling then without it to know the impact of oversampling on our model. Finally, we compared our results with the benchmark and between the algorithms themselves to find best suitable algorithm in this type of problems. The results shows that our models got enhanced metrics compared to the benchmark models and the support vector machine algorithm gave the best enhancement without and with oversampling which obtained 18.46% and 26.54% respectively it terms of recall metric. The results also shows that oversampling technique that we used leads to worse results in logistic regression algorithm and make no effect in both random forest and gradient boosting algorithms.

Keywords: botnets, zero-day attacks, Machine learning, intrusion detection systems.