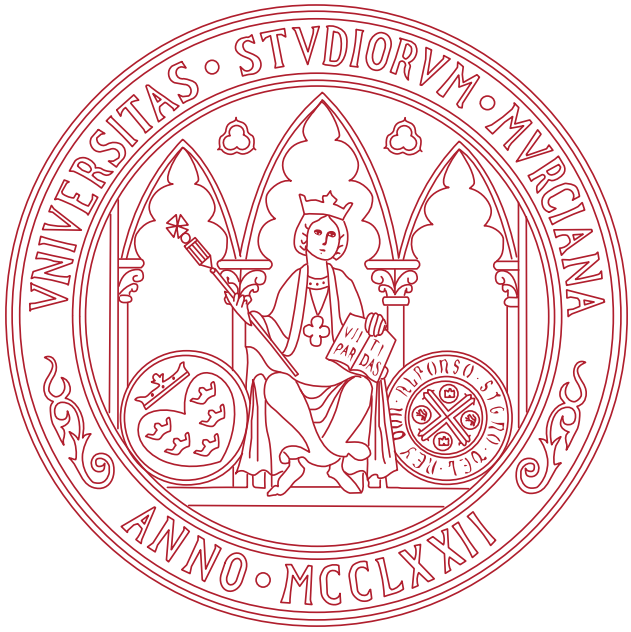

Universidad de Murcia

Trabajo Final de Grado

Identificación de dispositivos IoT a través de huellas hardware

Autor:
Sergio Marín Sánchez
sergio.marins@um.es

Tutores:
Gregorio Martínez Pérez
gregorio@um.es
Pedro Miguel Sánchez Sánchez
pedromiguel.sanchez@um.es

2 de junio de 2022

Me gustaría expresar mi agradecimiento a todos los amigos, profesores y familia que han hecho posible que llegue a este punto.

Índice general

Índice de figuras

Índice de tablas

Resumen

En este proyecto hemos diseñado un modelo capaz de identificar dispositivos conectados en red local en base a las diferencias entre sus relojes internos comparados con un reloj que se fija como exacto. Con esto hemos creado un mecanismo que permitirá realizar conexiones más seguras.

Este trabajo se puede dividir en tres partes. Por un lado tenemos la obtención de los datos, para lo que usamos 5 dispositivos iguales en hardware y software (Raspberry Pi). De estos dispositivos obtenemos marcas de tiempo cada segundo de dos formas distintas. Por otro lado, tenemos el análisis estadístico de los datos obtenidos, que generará nuevos datos a partir de los previos. Por último tenemos el desarrollo de un modelo de machine learning que sea capaz de automatizar todo el proceso de distinguir los dispositivos partiendo de los datos estadísticos que hemos obtenido previamente.

Para obtener los datos hemos tenido que realizar una arquitectura cliente-servidor entre los dispositivos a analizar y el dispositivo observador, con el fin de controlar el ritmo de envío de los paquetes. Las capturas se han realizado de dos formas, en secuencial y en paralelo. Por secuencial nos referimos a que primero obtenemos todas las marcas de tiempo de un dispositivo y posteriormente pasamos al siguiente. Por otro lado, en paralelo nos referimos a que se obtienen marcas de tiempo de todos los dispositivos simultáneamente.

Una vez tenemos todas las marcas de tiempo, debemos obtener sus desviaciones respecto al dispositivo que tomamos de referencia. De estas desviaciones se obtiene su incremento entre cada dos puntos, con estos valores usamos una ventana deslizante de 1 minuto, que equivale a 60 paquetes (se toman muestras cada segundo), y con de ellos obtenemos variables estadísticas.

Una vez tenemos las variables estadísticas procedemos a analizar los modelos de machine learning que vamos a usar. Los modelos que se van a analizar son Random Forest, MLP, KNN, Naive Bayes, Árboles de decisión y SVM con kernel lineal.

Para comparar los modelos entre sí usaremos el modelo entrenamiento/validación/test como forma de dividir nuestros datos. Entrenamos todos nuestros modelos con el conjunto de entrenamiento y comprobamos su capacidad de generalización con el conjunto de validación. A la hora de entrenar los modelos comprobaremos diferentes parrillas de hiperparámetros para ajustar nuestros modelos lo más posible a los datos de entrenamiento.

Al realizar estos entrenamiento vemos que los modelos basados en árboles son los mejores, tanto árboles de decisión (97.57 %) como Random Forest (99.51 %). Nos quedamos finalmente

con el modelo de Random Forest como algoritmo final y lo entrenaremos con la totalidad de la base de datos de entrenamiento.

Finalmente el modelo de Random Forest con los hiperparámetros ajustados obtiene un valor de accuracy de 99.44 %.