UNIVERSIDADE DA CORUÑA

ANTEPROXECTO (MODELO)

facultade de informática da coruña

Datos da/o estudante
Nome: Fernando Villar Freire
DNI: 46902842T Teléfono:699882456
Enderezo electrónico: fernando.villar.freire@udc.es

Título *(galego)*:
Detección de anomalías de rede en tempo real mediante técnicas de *Machine learning*

Título *(castellano)*:
Detección de anomalías de red en tiempo real mediante técnicas de *Machine learning*

Título *(English)*:
Real time network anomaly detection using Machine learning techniques

Clase de proxecto *(elixir un)*:
De desenvolvemento en investigación

Mención:
Tecnoloxías da información

Dirección:
Francisco Javier Nóvoa Manuel
Diego Fernández Iglesias

Breve descrición:
En un panorama actual con un aumento del tráfico (debido a la proliferación de IoT, al auge de la filosofía del siempre conectado y al aumento de servicios multimedia, cloud, videoconferencia y su tráfico asociado) junto con la proliferación de ataques y nuevas amenazas, surge la necesidad de modelos de defensa rápidos y adaptables.

La presencia en el mercado de sistemas IDS (Intrusion detection systems) muestra la preocupación de las organizaciones y el cambio de rumbo frente a la seguridad tomada en los últimos tiempos.
Dentro de los sistemas IDS existen 2 grandes aproximaciones: Los basados en firmas y los basados en detección de anomalías. Los primeros se basan en identificar a través de patrones ya conocidos (firmas) los posibles ataques, como es lógico, su efectividad está acotada a ataques ya conocidos. La segunda familia se centra en estudiar el tráfico y detectar desviaciones anormales del mismo, lo cual le permite detectar ataques todavía desconocidos.

Existen diferentes enfoques a la hora de afrontar el estudio del tráfico de red, pero el uso del enfoque de agregación, tanto en la captura como en el análisis, viene a solucionar en parte los problemas de las ingentes cantidades de tráfico que soportan a día de hoy las organizaciones. Las tecnologías NETFLOW e IPFIX son dos soluciones similares a la hora de la creación de flujos, pero con funcionamientos que a priori no están pensados para su uso en tareas intensivas de tiempo real. Los análisis asociados a ambas suelen ceñirse a ventanas de 5 minutos, tiempo excesivo a la hora de enfrentarse a algunas tareas tan críticas como son las de seguridad.
El volumen de información derivado del proceso de generación de flujos provoca que sean necesarias técnicas de machine learning, en adelante ML, para analizar dicha información.

El auge actual de las técnicas de ML junto con el crecimiento, disponibilidad y sencillez de uso de la computación distribuida permite paliar en gran medida los problemas de rendimiento de las soluciones centradas en el análisis de tráfico agregado.

Obxectivos concretos:
El objetivo principal de este proyecto es comparar y comprender diferentes técnicas de ML aplicadas a la detección en tiempo real de comportamientos anómalos en el tráfico de una red.
En el caso que nos atañe, en la que existe una variable objetivo (pertenencia al grupo tráfico anómalo), nos encontramos con una tarea de aprendizaje supervisado y más concretamente con una tarea de clasificación.

De manera transversal incidiremos en aspectos como la captura de flujos o la implementación y despliegue de una solución distribuida.

Los objetivos a alcanzar son los siguientes:

- Estudiar los conceptos básicos tras las tecnologías NETFLOW e IPFIX
    - Visión global y estado actual de la tecnología.
    - Alternativas de implementación en el mercado.
    - Diseño y despliegue de una arquitectura de captura basada en ellas.
- Seleccionar, implementar y comparar diferentes algoritmos de ML
    - Selección, transformación y análisis de los datos que compondrán el dataset del estudio.
    - Selección de los algoritmos de ML de clasificación.
    - Evaluación de los modelos resultantes, teniendo en cuenta como parámetros de rendimiento la tasa de aciertos y la rapidez en la ejecución.
- Diseñar y desplegar un sistema distribuido basado en Apache Spark
    - Acercamiento a la tecnología distribuida centrándose en la solución Apache Spark.
    - Configuración y puesta en marcha de la arquitectura distribuida que dé sustento a las tareas de análisis que centran este trabajo.
- Desplegar y probar los modelos de análisis en un entorno en tiempo real
    - Despliegue de los modelos y evaluación de éstos en un entorno online.
    - Medir el rendimiento de los modelos teniendo como punto de vista la cantidad de flujos por segundo que son capaces de procesar.

Método de traballo:
Usaremos una metodología *agile*, más concretamente Scrum. La intención es segmentar el proyecto en unidades más livianas (sprint) totalmente funcionales. Se prevé que la duración de los sprints sea de un máximo de 2 semanas, éstos se corresponderán con las fases que describiremos a continuación, exceptuando "Planificación" y "Recopilación y análisis de información".
En cada uno de los sprints realizaremos tareas de análisis, documentación, desarrollo y pruebas, entregando a su finalización un producto funcional correspondiente a la fase en la que se engloba.

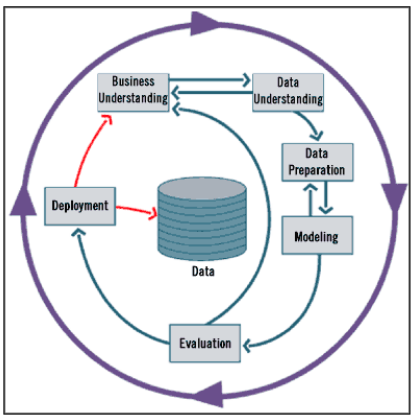

Ilustración 1 - CRISP-DM

Debido a que el núcleo central del TFG se corresponde con tareas de ML, usaremos durante esa fase la metodología CRISP-DM, la cual es un referente en el sector. La intención es adecuar las fases del proceso a los diferentes sprints de la metodología Scrum.

Fases principais do traballo:
A la hora de abordar la tarea distinguimos 8 fases diferentes. Las 2 primeras corresponden a los trabajos preparatorios que dan sustento a las 6 posteriores.

1. Planificación
   En esta primera fase nos centraremos en la asignación de cargas de trabajo y orden de ejecución de las fases posteriores.
2. Recopilación y análisis de información
   Debido a que en gran medida la materia del estudio es desconocida por parte del autor, este bloque tendrá asignado una buena parte de la carga de trabajo total.
   En un primer momento nos centraremos en la recopilación de los conceptos principales sobre los que versan las tecnologías implicadas, los cuales nos permitirán comprender en un segundo paso los conceptos más complejos que harán falta para completar el TFG y la elaboración de la memoria final.

3. Preparación de los datos
   Consistirá en la preparación previa del dataset para su posterior uso en el entrenamiento de los algoritmos a evaluar.
   El dataset escogido es Intrusion detection evaluation dataset proporcionado por Canadian Institute for Cybersecurity (https://www.unb.ca/cic/datasets/ids.html). De uso muy extendido, nos permitirá en último término valorar nuestra solución contra otras ya publicadas.
4. Implementación del algoritmo de ML
   Inicialmente acometeremos la selección e implementación del algoritmo de clasificación teniendo como punto de partida la colección que nos ofrece el componente de Apache Spark ML. Tras la elección del modelo procederemos a su evaluación contra el conjunto de datos de validación.
   Todos los procesos asociados a esta fase se realizarán en un entorno offline. El modelo final será cargado posteriormente en el sistema distribuido.
5. Diseño y despliegue del sistema distribuido
   El despliegue del sistema distribuido se realizará en local corriendo sobre una única máquina. Como tarea opcional se valorará desplegar el sistema sobre la plataforma AWS de Amazon.
6. Diseño y despliegue de sistema de captura de tráfico
   Se centrará en la configuración e instalación de la infraestructura necesaria para realizar la captura y agregación de tráfico que sirve de punto de entrada a nuestro sistema.
7. Despliegue y pruebas del modelo final
   A continuación, cargaremos el modelo algorítmico en nuestro sistema distribuido y evaluaremos su comportamiento ante un entorno en tiempo real.
8. Extracción de conclusiones y valoración final
   Tras estudiar los datos de la fase 7, podremos valorar si los objetivos indicados en este documento han sido realizados y la adecuación de éstos a nuestras expectativas iniciales. Finalmente se desarrollará una memoria con las conclusiones sacadas del trabajo realizado.

Material e medios necesarios:
Ordenador de sobremesa y acceso a internet.

<u>Propiedade intelectual do traballo</u>:

O regulamento de Traballos de Fin de Grao da Facultade de Informática establece na sección 4, en relación aos dereitos derivados da propiedade intelectual dos traballos, o seguinte:

4.2. No caso dos traballos desenvolvidos en colaboración cunha entidade externa, a titularidade dos dereitos de propiedade intelectual, se for o caso, rexerase polo establecido na relación contractual entre a/o estudante e a entidade externa. Neste caso, quen exerza a dirección académica non será titular dos dereitos de propiedade intelectual, salvo que se establecer doutra maneira nun documento asinado pola/o estudante, o profesorado encargado da dirección e un/ha representante da entidade externa.

4.3. No caso dos traballos desenvolvidos no ámbito do centro, a titularidade dos dereitos de propiedade intelectual, se for o caso, corresponderá á/ao estudante segundo queda recollido no apartado h) do artigo 8 do Real Decreto 1791/2010 do 30 de decembro, salvo que se establecer doutra maneira nun documento asinado pola/o estudante e o profesorado encargado da dirección do TFG.

Indique a continuación se o proxecto se realiza en colaboración cunha entidade externa ou no ámbito do centro, e neste último caso, o acordo sobre os dereitos derivados da propiedade intelectual do traballo:

O proxecto realízase en colaboración cunha entidade externa:

Si [ ] Non [X]

Se o proxecto non se realiza en colaboración cunha entidade externa, indique se os dereitos derivados da propiedade intelectual son compartidos entre a/o estudante e as/os directores:

Si [ ] Non [ ]

A Coruña, a .

Asinado: a/o estudante. Asinado: o/a director/a ou directores/as