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 mediante técnicas de *machine learning*

Título *(castellano)*:
Detección de anomalías de red mediante técnicas de *machine learning*

Título *(English)*:
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 del internet de las cosas, 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 de detección de intrusiones (IDS) muestra la preocupación de las organizaciones frente a la seguridad. Los IDS son herramientas de seguridad encargadas de monitorizar los eventos que ocurren en un sistema informático en busca de intentos de intrusión. 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 (firmas) los posibles ataques. 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.
El uso de la agregación permite disminuir el volumen de datos a tratar sin perder la información relevante de los mismos. La agregación permite agrupar el tráfico, en un momento dado, en una secuencia de paquetes que comparten unos valores determinados. A esta agrupación le llamamos flujo.

A pesar de la reducción en el volumen de información, conseguida gracias al uso de flujos, sigue haciéndose necesario el uso de técnicas de Big Data para analizarla.
El auge actual de las técnicas de machine learning, en adelante ML, junto con el crecimiento, disponibilidad y sencillez de uso de la computación distribuida nos proporcionan las herramientas necesarias para acometer las tareas 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 el que existe una variable objetivo (pertenencia al grupo tráfico anómalo), nos encontramos 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 la arquitectura de agregación de flujos y sus principales protocolos
  - Visión global y estado actual de la arquitectura y sus principales protocolos (NetFlow e IPFIX).
  - Alternativas de implementación en el mercado.
  - Diseño y despliegue de una arquitectura de captura de flujos.
- Seleccionar, implementar y comparar diferentes algoritmos de ML
  - Ingeniería de características sobre el dataset de estudio.
  - Selección de los algoritmos de ML de clasificación.
  - Evaluación de los modelos resultantes.
- Configurar y desplegar un sistema distribuido
  - Acercamiento a la tecnología distribuida.
  - 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 framework de computación distribuida
  - Despliegue de los modelos y evaluación de éstos.

Método de traballo:
Usaremos una metodología *agile*, más concretamente Scrum. La intención es segmentar el proyecto en unidades más livianas (sprints) totalmente funcionales. Se prevé que la duración de los sprints sea de 2 semanas. Estos sprints se corresponderán con las fases de la 2 a la 8 que describiremos a continuación.
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.

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.

Las fases son las siguientes:
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. Ingeniería de características
   Consistirá en la preparación previa del dataset para su posterior uso en el entrenamiento de los algoritmos a evaluar.
4. Implementación de los algoritmos de ML
   Inicialmente acometeremos la selección de los algoritmos a estudiar para posteriormente implementarlos y evaluarlos contra el conjunto de datos de validación. Los modelos finales serán cargados en el sistema distribuido.
5. Configuración y despliegue del sistema distribuido
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.
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
- Acceso a internet
- Git como herramienta de control de versiones
- Pycharm como IDE para realizar las implementaciones necesarias en Python
- Weka como framework de apoyo para las tareas de minería de datos

Propiedade intelectual do traballo:

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 [X] Non [ ]

A Coruña, a 5 de abril de 2019

Asinado: a/o estudante.

Fernando Villa

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

Hugo Fernández

F. J. Nóvoa