Universidad de los Andes

Inteligencia de negocios

Proyecto 1

Rafael Tejón

María del Rosario León -201423755

Marlon Forero

septiembre 2019

**Comprensión del negocio y enfoque anaítico.**

Oportunidad/problema del negocio.

La empresa de seguridad Alpes, necesita una herramienta de minería de datos, que le permita identificar los ataques a los dispositivos Iot de manera temprana. Para esto, requieren clasificar los paquetes de datos y buscar patrones que se presentan en el momento de los ataques. Para lograr esto, se cuenta con un archivo csv, que contiene la información de diferentes paquetes de red.

Descripción del requerimiento desde el punto de vista de minería de datos.

La solución problema de seguridad Alpes, se puede dividir en tres objetivos:

1. Diferenciar los paquetes de red, entre paquetes de actividad normal y paquetes de ataques. Para esto, lo más conveniente sería utilizar una técnica de clasificación por similitud.
2. Diferenciar los paquetes de red según cada tipo de ataque, por lo que es necesario utilizar una técnica de clasificación por afinidad.
3. Encontrar patrones entre los paquetes de datos que son resultado de un ataque. Para lograr esto es conveniente usar las reglas de asociación.

Detalles de la actividad de minería de datos.

|  |  |  |
| --- | --- | --- |
| Tarea | Técnica | Algoritmos y parámetros utilizados |
| Clasificación de los paquetes de red | Árbol de decisión |  |
| Clasificación de los paquetes que son un ataque | Cluster |  |
| Búsqueda de patrones | Reglas de asociación |  |

**Comprensión de los datos y preparación de los datos**

Perfilamiento de los datos

El archivo csv con los datos de los paquetes de red contiene 40 columnas y 5602 filas. Dentro de las columnas encontramos:

* 2 columnas binarias (Label, is\_sm\_ips\_ports)
* 10 columnas tipo Float (dur, Sload, Dload, Sjit, Djit, Sintpkt, Dintpkt, tcprtt, synack, ackdat)
* 24 columnas tipo integer (id, sbytes, dbytes, sttl, dttl, sloss, dloss, Spkts, Dpkts, swin, dwin, stcpb, dtcpb, smeansz, dmeansz, res\_bdy\_len, ct\_state\_ttl, ct\_srv\_src, ct\_srv\_dst, ct\_dst\_Itm, ct\_src\_Itm, ct\_src\_dport\_Itm, ct\_dst\_sport\_Itm, ct\_dst\_src\_Itm)
* 4 columnas nominales (proto, state, service, attack\_cat)

Análisis de la calidad de los datos

Para esta tarea de minería de datos, todos los datos salieron de la misma fuente, es decir el cliente. Por esta razón sabemos que los datos son concordantes en los tipos de datos y unidades utilizadas.

Por otro lado, encontramos que en algunas columnas, no todos los datos tienen información relevante o están en cero. Estas filas deben ser consideradas en el momento de realizar la preparación de los datos.

Tratamiento de los datos

1. Se elimina la columna ID
2. Para realizar el cluster, se utiliza el nodo Category to number en knime, de manera que los valores categóricos sean convertidos a número y el algoritmo no genere error.