Deliverable 3 : Data Management Plan

Proyecto: Advanced Networkmetrics: Interpretable Machine Learning for Intelligent Communication Systems. Agencia Estatal de Investigación, MCIN/AEI/ 10.13039/501100011033, proyecto No PID2020-113462RB-I00.

Contacto: José Camacho

Última modificación: 20/11/2024

Fuentes: El proyecto latex editable de este documento se encuentra en overleaf, en: https://www.overleaf.com/project/672b71d8b620d43eec8b9b46

1. Introducción

El proyecto **ANIMaLICoS**: *Advanced Networkmetrics: Interpretable Machine Learning for Intelligent Communication Systems* [1], investiga una propuesta de Aprendizaje Automático Interpretable basado en el Análisis Multivariante de Big Data para resolver problemas de red, que utiliza una metodología que llamamos Retemetría (Networkmetrics), en referencia al uso del análisis multivariante interpretable en aplicaciones de red. El nombre representa la combinación del dominio de la aplicación (ingeniería de redes, del latín rete-) y el sufijo metría", heredado de otras disciplinas donde el análisis multivariante ha sido ampliamente adoptado, tanto en la academia como en la industria.

Uno de los principales resultados del proyecto es la generación de conjuntos de datos para la comunidad. En particular, los dos conjuntos de datos generados por el proyecto son:

- UGR16 Feature data, con cuatro variantes del dataset UGR'16 que se usan en los siguientes artículos:

  Camacho, J., Wasielewska, K., Espinosa, P., Fuentes-García, M. Quality In / Quality Out: Data quality more relevant than model choice in anomaly detection with the UGR'16. IEEE/IFIP Network Operations and Management Symposium. Miami, USA. 2023.

  Camacho, J., Rodríguez-Gómez, R.A. Data quality tools to optimize an anomaly detection benchmark. Submitted to Data, 2024.

- Dartmouth Feature data, datos por características obtenidos del dataset de Dartmouth que se usan en los siguiente artículos:

MINISTERIO DE CIENCIA E INNOVACIÓN
AGENCIA ESTATAL DE INVESTIGACIÓN

Camacho, J., Wasielewska, K., Bro R., Kotz, D. Interpretable Learning in Multivariate Big Data Analysis for Network Monitoring. IEEE Transactions of Network and Service Management, 2024, 21(3):2926-2943.

Camacho, J., McDonald, C., Peterson, R., Zhou, X. Longitudinal Analysis of a Campus Wi-Fi Network. Computer Networks. 2020, 179, 107103.

Ambos datasets están ínitamamente relacionados con los tres objetivos del proyecto, a saber:

- OBJETIVO 1: Aplicar la retemetría a las tecnologías modernas de Internet. En particular, nos centraremos en desarrollar nuevos enfoques y algoritmos para redes definidas por software.
- OBJETIVO 2: Extender técnicas multivariantes novedosas al Big Data, en aplicaciones como detección de fallos, ciberseguridad, clasificación de tráfico y optimización de redes.
- OBJETIVO 3: Impulsar la retemetría en la comunidad internacional, tanto en la industria como en el mundo académico.

Los datasets permiten implementar los objetivos 1 y 2, y su compartición es una estrategia eficiente del objetivo 3.

2. UGR16 Feature data

2.1. Resumen de los datos

El dataset consiste en cuatro variantes de un conjunto de datos de referencia en la detección de anomalías de red, el UGR'16 [2][1]. Las variantes se obtuvieron aplicando pequeñas diferencias en el procesamiento de datos. Implementamos un motor de detección de anomalías utilizando estas variantes, con dos metodologías de aprendizaje automático muy diferentes, y encontramos diferencias insignificantes en el rendimiento entre las variantes de aprendizaje automático, pero diferencias significativas entre las variantes del conjunto de datos [3]. Este resultado es relevante de cara a entender el reto que supone la calidad de los datos en el despliegue de redes autónomas.

Se puede acceder a los datos en diferentes formatos:

F.1 Como registros en formato nfcapd (un formato binario utilizado en el paquete nfdump). Los datos se almacenan en archivos semanales y se organizan en carpetas por mes. El

[1]Dataset original disponible en https://nesg.ugr.es/nesg-UGR'16/

almacenamiento total requerido para los datos es de 341 GB.

F.2 Como datos de características por intervalo de tiempo utilizando el enfoque de característica como contador (Feature as a Counter o FaaC), con un archivo en formato csv por intervalo de tiempo de 1 minuto. Los datos se obtuvieron en dos pasos. Primero, los archivos nfcapd se transformaron a formato csv utilizando la herramienta nfdump. Estos archivos intermedios no se almacenan debido al volumen de datos resultantes. Posteriormente, el FCParser[2] se aplica sobre los archivos intermedios, generando archivos de salida (uno por minuto) con una sola fila de 144 contadores cada uno. Hay dos versiones del conjunto de datos UGR16 en este formato que consideran flujos bidireccionales y unidireccionales en nfdump, respectivamente. El volumen de datos es de 1,5 GB en total para las dos versiones.

F.3 Como datos de características para conjuntos de datos completos, con tantas filas como intervalos de tiempo y una selección de 134 contadores como columnas. Los datos de características se proporcionan con etiquetas para 8 clases de ataques: los ataques artificiales (DOS, SCAN11, SCAN44 y NERISBOTNET) y alguna actividad real observada y etiquetada en los datos originales [2] (BLACKLIST, UDPSCAN, SSHSCAN y SPAM). Los datos se pueden encontrar en formato csv, excel y .mat para su importación en matlab. Este es el formato más fácil de usar para cualquier persona interesada en el desarrollo de nuevos modelos de ML a partir de datos. El volumen de datos en este formato es también de 1,5 GB en total.

2.2. Responsabilidades

La responsabilidad de la gestión continuada de los datos, incluyendo el acceso a los mismos, es del Investigador Principal del proyecto **ANIMaLICoS**, José Camacho.

2.3. Datos FAIR

2.3.1. Datos localizables y Accesibilidad

Los datos en formato F.3 se encuentran alojados en el repositorio público Github[3], donde pueden ser descargados libremente. En particular, los datos corresponden a la versión v.1 definida en el control de versiones del repositorio.

Los datos en cualquiera de los tres formatos son también accesibles, junto las herramientas para

[2] https://github.com/josecamachop/FCParser
[3] https://github.com/josecamachop/UGR16_FeatureData

MINISTERIO DE CIENCIA E INNOVACIÓN
AGENCIA ESTATAL DE INVESTIGACIÓN

su procesamiento, a través del **Data Analysis as a Service** (***DAaaS***) desplegado en https://codas.ugr.es/animalicos/es/daaas, un servicio online que facilita el análisis e interpretación de datos siguiendo la metodología Multivariate Big Data Analysis [4].

Tanto repositorio Github como servicio DAaaS son localizables desde la web del proyecto[4]. Los datos se pueden encontrar en formato csv, excel y .mat para su importación en matlab.

2.3.2. Interoperabilidad

Todos los formatos de datos son abiertos y legibles por herramientas de software libre. Se utiliza el conjunto de metadatos nativo de Github. En la web del proyecto se incluye título, link y referencias.

2.3.3. Reutilización

Los datos son reutilizables bajo licencia GPL-3.0. Los datos permanecerán reutilizables tras el fin del proyecto, sin limitación ni restricción de acceso.

2.4. Seguridad de los Datos

Todos los datos se basan en versiones anonimizadas de la traza original [2] o bien han sido completamente anonimizados al pasar a características.

La seguridad de los datos en formato F.3 se implementan por el propio repositorio Github.

Con respecto al resto, el **Data Analysis as a Service** (***DAaaS***) incluye un sistema complejo de copias de seguridad mediante *crontab*. Lo primero que se hace son las copias de seguridad del servicio DAaaS, una cada hora, de los 10 últimos días (añadiendo un *tag* con la fecha), para permitir guardar el estados de los usuarios. Adicionalmente, se hacen copias externas diarias del conedor que alberga el servicio a Docker Hub. Finalmente, se mantienen copias de seguridad de los datos en sendos HDD en el servidor y en un disco externo.

[4] https://codas.ugr.es/animalicos

MINISTERIO DE CIENCIA E INNOVACIÓN
AGENCIA ESTATAL DE INVESTIGACIÓN

3. Dartmouth Feature data

3.1. Resumen de los datos

Datos identificados automáticamente utizando el FClearner a partir de los datos anonimizados de la traza original [5]. Son datos de características por intervalo de tiempo utilizando el enfoque de característica como contador (Feature as a Counter o FaaC), por intervalo de tiempo de 1 día para un total de 7 años (2548 intervalos) y una selección de 92 contadores como columnas. Los datos se pueden encontrar en formato csv, excel y .mat para su importación en matlab. El volumen de datos es de 24 MB en total.

3.2. Responsabilidades

La responsabilidad de la gestión continuada de los datos, incluyendo el acceso a los mismos, es del Investigador Principal del proyecto **ANIMaLICoS**, José Camacho.

3.3. Datos FAIR

3.3.1. Datos localizables y Accesibilidad

Los datos en formato se encuentran alojados en el repositorio público Github[5], donde pueden ser descargados libremente. En particular, los datos corresponden a la versión v.1 definida en el control de versiones del repositorio. El repositorio es también localizable desde la web del proyecto[6].

3.3.2. Interoperabilidad

Todos los formatos de datos son abiertos y legibles por herramientas de software libre. Se utiliza el conjunto de metadatos nativo de Github. En la web del proyecto se incluye título, link y referencias.

[5] https://github.com/josecamachop/Dartmouth_FeatureData
[6] https://codas.ugr.es/animalicos

MINISTERIO DE CIENCIA E INNOVACIÓN
AGENCIA ESTATAL DE INVESTIGACIÓN

3.3.3. Reutilización

Los datos son reutilizables bajo licencia GPL-3.0. Los datos permanecerán reutilizables tras el fin del proyecto, sin limitación ni restricción de acceso.

3.4. Seguridad de los Datos

Todos los datos se basan en versiones anonimizadas de la traza original [5] y han sido completamente anonimizados al pasar a características. La seguridad de los datos se implementan por el propio repositorio Github.

Referencias

[1] ANIMaLICoS. Advanced networkmetrics: Interpretable machine learning for intelligent communication systems. https://www.codas.ugr.es/animalicos/en.

[2] Gabriel Maciá-Fernández, José Camacho, Roberto Magán-Carrión, Pedro García-Teodoro, and Roberto Therón. UGR'16: A new dataset for the evaluation of cyclostationarity-based network IDSs. *Computers & Security*, 73:411–424, 2018.

[3] J. Camacho and R.A. Rodríguez-Gómez. Data quality tools to optimize an anomaly detection benchmark. *Submitted to Data*, September 2024.

[4] José Camacho, Katarzyna Wasielewska, Rasmus Bro, and David Kotz. Interpretable feature learning in multivariate big data analysis for network monitoring. *IEEE Transactions on Network and Service Management*, 21(3):2926–2943, 2024.

[5] José Camacho, Chris McDonald, Ron Peterson, Xia Zhou, and David Kotz. Longitudinal analysis of a campus wi-fi network. *Computer Networks*, 170:107103, 2020.

MINISTERIO DE CIENCIA E INNOVACIÓN
AGENCIA ESTATAL DE INVESTIGACIÓN