Proyecto Final de Aprendizaje de Máquina

Estudio sobre registros de consumo de datos, voz y sms

Integrantes:

Claudia Hernández Pérez
Joel Aparicio Tamayo
Kevin Márquez Vega
Javier A. González Díaz
José Miguel Leyva Cruz
Luis E. Amat Cárdenas

Índice

1. Estudio del estado del arte

La investigación sobre predicción y clasificación de tráfico en redes ha evolucionado de manera notable en los últimos quince años. En la primera etapa (2010-2013), los modelos estadísticos clásicos como ARMA, ARIMA y SARIMA fueron la base para el análisis de series temporales, mientras que las redes neuronales superficiales (MLP) demostraron capacidad para capturar patrones complejos antes del auge del aprendizaje profundo.

A partir de 2014, el Deep Learning comenzó a aplicarse con redes profundas (DNN), mostrando mejoras claras frente a los métodos tradicionales. Posteriormente, los modelos híbridos que combinan CNN y LSTM permitieron capturar tanto correlaciones espaciales como temporales, consolidando el enfoque espacio-temporal en la predicción de tráfico.

Entre 2017 y 2018 surgieron propuestas orientadas a la eficiencia y nuevas representaciones: RCLSTM redujo parámetros manteniendo rendimiento, mientras que la representación del tráfico como imágenes espacio-temporales con CNN 2D mejoró la precisión. En paralelo, se exploraron integraciones más complejas como DMVST-Net, y se mantuvo la relevancia de algoritmos clásicos (SVM, K-NN, árboles) en tareas de clasificación.

En el periodo 2019-2021 se diversificaron las técnicas, incorporando clustering y reducción de dimensionalidad (K-means, DBSCAN, PCA), además de modelos optimizados como SVR y XGBoost, que superaron a los estadísticos en escenarios no lineales. Las comparativas entre ML y DL confirmaron la superioridad de LSTM, aunque MLP se mantuvo competitivo.

Los avances recientes (2022-2023) destacan el uso de K-NN con selección de características en LTE, la incorporación de grafos mediante GNN combinados con RNN para modelar dependencias espaciales y temporales, y la validación de modelos estadísticos como SARIMA y Holt-Winters en contextos específicos. También se resaltan alternativas rápidas y robustas como OS-ELM y Random Forest.

Finalmente, las tendencias emergentes (2024-2025) apuntan hacia arquitecturas modernas como Transformers y TCN, capaces de capturar dependencias globales y temporales con gran precisión. Se exploran enfoques eficientes como HTM y aprendizaje federado para preservar privacidad, y se proponen marcos híbridos adaptativos para entornos 5G. Las revisiones recientes concluyen que el aprendizaje profundo moderno (LSTM, GNN, Transformers) domina en problemas espacio-temporales complejos, mientras que los métodos clásicos siguen siendo útiles en escenarios simples o con restricciones de recursos.

2. Estudio sobre los datos

La Empresa de Telecomunicaciones de Cuba S.A. (ETECSA) proporcionó un conjunto de datos con el propósito de desarrollar estudios y modelos basados en técnicas de Aprendizaje de Máquina (Machine Learning). Dichos datos reflejan el uso de diversos servicios de telecomunicaciones por parte de los usuarios, tales como llamadas telefónicas, mensajes de texto (SMS), recargas de saldo y consumo de datos móviles.

El objetivo principal de este análisis es comprender la estructura, el contenido y las características de los datos, con vistas a su preparación y posterior aplicación en modelos predictivos o de análisis de comportamiento.

2.1. Descripción general del dataset

El conjunto de datos se encuentra en formato tabular y contiene aproximadamente 10 mil registros y 40 variables, distribuidas en columnas que describen los diferentes aspectos de cada transacción o evento de uso de servicios.

Cada fila representa un registro detallado de uso de servicio (CDR, por sus siglas en inglés: Call Detail Record), que documenta información relacionada con un evento generado por el cliente, como una llamada, el envío de un mensaje o una conexión a internet móvil.

A continuación, se presenta un resumen de los tipos de variables más relevantes:

Tipo de variable	**Ejemplo de campos**	**Descripción general**
Identificadores	CDR_ID, OBJ_ID, OWNER_CUST_ID	Identifican de manera única cada registro, objeto o cliente asociado.
Temporales	START_DATE, END_DATE	Indican la fecha y hora de inicio y fin del servicio utilizado.
Categóricas	SERVICE_CATEGORY, FLOW_TYPE, USAGE_SERVICE_TYPE	Especifican el tipo de servicio, su categoría (voz, datos, SMS, recarga) y dirección del tráfico (entrante/saliente).
Numéricas	ACTUAL_USAGE, ACTUAL_CHARGE, TOTAL_TAX_AMOUNT	Miden el volumen de uso (minutos, megabytes, mensajes) y los cargos monetarios asociados.
Listas o estructuras anidadas	CHARGE_LIST, CHARGE_SERVICE_INFO, BALANCE_CHG_LIST	Detallan cargos, impuestos y modificaciones de saldo que se producen en cada evento.

Estos campos se complementan con información auxiliar relacionada con unidades de medida, identificadores de cuenta, ciclos de facturación y valores reservados para futuras ampliaciones del sistema.

2.2. Distribución de los datos

3. Detección de patrones

Explicación de los métodos usados para encontrar patrones en los datos.

4. Reducción de dimensionalidad

Discusión sobre técnicas como PCA, t-SNE, etc.

5. AutoML

Descripción de herramientas y resultados obtenidos con AutoML.

Referencias

[1] Autor, Título del libro/artículo, Editorial, Año.

[2] Autor, Título del paper, Revista, Volumen, Año.