UNIVERSIDAD DE GRANADA

GENERACIÓN DE DATOS SINTÉTICA A PARTIR DE UN SISTEMA IOT DE DETECCIÓN DE VEHÍCULOS UTILIZANDO MODELOS TRANSFORMER CON ATENCIÓN EN PYTHON

RICARDO RUIZ FERNÁNDEZ DE ALBA

Trabajo Fin de Grado

Doble Grado en Ingeniería Informática y Matemáticas

Tutores

María Bermúdez Edo

FACULTAD DE CIENCIAS

E.T.S. INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN

Granada, a 23 de octubre de 2024

ÍNDICE GENERAL

RESUMEN

1

INTRODUCCIÓN Y OBJETIVOS

1.1 INTRODUCCIÓN

1.1.1 *Contextualización*

La Inteligencia Artificial (IA), impulsada por el avance en algoritmos de aprendizaje automático ha experimentado un rápido desarrollo debido a su reciente viabilidad de ejecución en hardware y la creciente disponibilidad de datos.

Sus fundamentos matemáticos, incluyendo el análisis, la probabilidad y el álgebra lineal que conforman la teoría del aprendizaje estadístico son esenciales para construir modelos que aprenden de los datos y realizan predicciones.

Como aplicación particular se encuentra la capacidad de estos algoritmos para identificar patrones en datos generados por las tecnologías de Internet Industrial de las Cosas (IoT), basadas en redes de sensores y dispositivos conectados a Internet que recopilan y transmiten datos en tiempo real.

1.1.2 *Descripción del problema*

El marco de este trabajo está relacionado con un proyecto de investigación de un sistema IoT implementado en la región de Poqueira, en la Alpujarra granadina [sma]. El sistema IoT ha procesado la información recogida por los sensores visuales y la ha exportado a un conjunto de datos con múltiples características.

El entrenamiento de modelos de Aprendizaje Automático a partir de estos datos podría predecir variables que mejorasen la gestión de la ocupación de la zona, promoviendo así una mayor sostenibilidad en la comunidad local.

Sin embargo, la complejidad de la información obtenida, la presencia de ruido y la falta de datos etiquetados, dificultan la tarea. En este contexto, la generación de datos sintéticos se presenta como solución, pues permite añadir al conjunto de datos original muestras generadas artificialmente que simulan las características de los datos reales. Esta técnica facilita el entrenamiento de modelos más robustos y precisos, abordando la escasez de etiquetas y mejorando la precisión de los modelos.

Aunque la literatura existente ha utilizado diversos Modelos de Generación Profunda (DGMs), incluyendo Redes Generativas Antagónicas (GANs) y Modelos de Autoencoders Variacionales (VAEs) [SBEXC], en este proyecto nos centraremos en los Modelos Transformers con atención, pues su capacidad de modelar dependencias a largo plazo en los datos, los presenta como una alternativa prometedora.

1.1.3 *Estructura del trabajo*

Este trabajo se organiza en cinco partes, cada una de las cuales se divide en capítulos que abordan diferentes aspectos del problema y su solución. La estructura sigue las directrices establecidas para los trabajos de fin de grado.

- **Primera parte:** Presenta la introducción y los objetivos del proyecto.
- **Segunda parte:** Detalla los fundamentos matemáticos y teóricos del aprendizaje automático que sustentan el trabajo.
- **Tercera parte:** Se centra en el análisis de los modelos Transformer y perceptrones multicapa, explorando sus fundamentos matemáticos y arquitectura.
- **Cuarta parte:** Describe el desarrollo del proyecto y los experimentos realizados, incluyendo detalles sobre el problema, el software utilizado, el preprocesamiento de datos y las métricas de evaluación.
- **Quinta parte:** Presenta las conclusiones y reflexiones finales sobre el trabajo, así como recomendaciones para futuras investigaciones.

1.1.4 *Herramientas Matemáticas e Informáticas*

Para el desarrollo de este trabajo se utilizarán herramientas matemáticas e informáticas clave. Los fundamentos matemáticos incluyen análisis (diferenciabilidad), probabilidad y estadística para el manejo de la incertidumbre, y álgebra lineal, esenciales en el diseño y optimización de modelos de *machine learning*.

La implementación se llevará a cabo en Python, utilizando bibliotecas como *Numpy*, *Pandas*, y *Matplotlib* para el análisis de datos, y *Sklearn* y *Hugging Face* para el entrenamiento de los modelos *Transformers*.

1.1.5 *Bibliografía Fundamental*

Para la elaboración del Capítulo 3, Teoría del Aprendizaje Estadístico se ha destacamos el uso del libro Learning from Data [AMMIL12].

1.2 OBJETIVOS

1.2.1 *Objetivo General*

El objetivo general de este trabajo es analizar los fundamentos matemáticos del aprendizaje automático y aplicar este conocimiento en la implementación de modelos generativos, específicamente utilizando arquitecturas Transformer, para la creación de datos sintéticos aplicables en la previsión de la ocupación a partir del conjunto de datos.

1.2.2 *Parte de Matemáticas Específicos*

1. **Estudiar los fundamentos matemáticos del aprendizaje automático:** Revisar y comprender los conceptos matemáticos y teóricos que sustentan el aprendizaje automático, incluyendo teoría sobre cotas de generalización, optimización y regularización.

El objetivo 1 de matemáticas se abordará en el Capítulo 2.

1.2.3 *Parte de Informática*

1. **Definir la arquitectura Transformer** Definir el modelo de perceptrones multicapa, el mecanismo de atención y como la arquitectura transformers combina ambos.
2. **Analizar las características de los datos recopilados:** Llevar a cabo un análisis del conjunto de datos obtenidos del sistema IoT de detección de vehículos, identificando limitaciones y patrones significativos.
3. **Analizar el rendimiento del modelo Transformer:** Evaluar el rendimiento y efectividad del modelo Transformer con atención a partir de métricas de evaluación, comparándolo con otros enfoques de modelado generativo.

El objetivos 1 de Informática se abordará en el Capítulo 3, el objetivo 2 en el Capítulo 4 y el objetivo 3 en el Capítulo 5.

2

FUNDAMENTOS MATEMÁTICOS Y DEL APRENDIZAJE AUTOMÁTICO

2.1 FUNDAMENTOS MATEMÁTICOS

2.1.1 *Fundamentos del Análisis*

2.1.1.1 *Diferenciabilidad*

2.1.1.2 *Regla de la Cadena y su Aplicación en Modelos de Machine Learning*

2.1.1.3 *Teoría de la Medida*

2.1.1.4 *Resultados sobre integración*

2.1.2 *Fundamentos de Probabilidad*

2.1.2.1 *Espacio de Probabilidad*

Definición 1 (σ-álgebra)**.** Sea $\mathcal{P}$ partes de Ω. Llamamos σ-álgebra a cualquier $\mathcal{A} \subset \mathcal{P}$ que cumpla:

- $\mathcal{A} \neq \emptyset$.
- Si $A_1, A_2, \ldots, A_n \in \mathcal{A}$ entonces $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$.
- Si $A \in \mathcal{A}$ entonces $A^c \in \mathcal{A}$.

A partir de las propiedades anteriores, deducimos que $\Omega \in \mathcal{A}$ y que $\mathcal{A}$ también es cerrado bajo intersecciones numerables.

Definición 2 (Función de probabilidad)**.** $P : \mathcal{A} \to [0, 1]$ es una función de probabilidad si satisface los tres axiomas:

1. $P(A) \geq 0 \quad \forall A \in \mathcal{A}$.

2. $P(\Omega) = 1$.
3. P es σ-aditiva, es decir: dada $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$ es una sucesión de conjuntos disjuntos dos a dos entonces

$$P\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} P(A_i).$$

Identificaremos como suceso seguro al evento que siempre ocurre. La primera condición nos asegura que el suceso seguro tiene la probabilidad más alta posible. La segunda condición garantiza que la probabilidad es no negativa. Finalmente, la tercera condición establece que, para un conjunto de sucesos disjuntos, la probabilidad de que ocurra alguno de ellos es igual a la suma de las probabilidades individuales de cada suceso.

Proposición 1. *Toda medida de probabilidad P, cumple:*

- $P(\emptyset) = 0$.
- *Dados $A, B \in \mathcal{A}$, entonces $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.*

Demostración. La primera propiedad se deduce de la σ-aditividad de P y de que $\emptyset = \bigcup_{i=1}^{\infty} \emptyset$. La segunda propiedad se deduce de la σ-aditividad de P y de que $A \cup B = A \cup (B \setminus A)$ disjuntos. $\square$

Definición 3 (Espacio de probabilidad)**.** Definimos como *espacio de medida* a la terna $(\Omega, \mathcal{A}, \mu)$, donde $\mu : \mathcal{A} \to R_0^+$ es una medida en $(\Omega, \mathcal{A})$ y *espacio de probabilidad* a la tupla formada por $(\Omega, \mathcal{A}, P)$, donde P es una medida de probabilidad en $(\Omega, \mathcal{A})$.

2.1.2.2 *Variables Aleatorias*

Una *variable aleatoria* es una función que asigna un valor, generalmente numérico, al resultado de un experimento aleatorio. Aunque no se puede conocer el valor exacto de una variable aleatoria al ser medida, sí se dispone de una distribución de probabilidad que describe la probabilidad de los diferentes valores posibles. A continuación, formalizamos el concepto de variable aleatoria.

Definición 4 (Función medible)**.** Sean $(\Omega_1, \mathcal{A})$ y $(\Omega_2, \mathcal{S})$ dos espacios de medida. Una *función medible* de dimensión n es una función $X \colon \Omega_1 \to \Omega_2$ que cumple:

$$X^{-1}(S) = \{\omega \in \Omega_1 \ : \ X(\omega) \in S\} \in \mathcal{A}, \text{ para todo } S \in \mathcal{S}.$$

Definición 5 (Variable aleatoria)**.** Sea $(\Omega_1, \mathcal{A}, P)$ un espacio de probabilidad y $(\Omega_2, \mathcal{S})$ un espacio de medida. Una *variable aleatoria* $\mathbf{X} = (X_1, \ldots, X_n)$ es una función medible $\mathbf{X} \colon \Omega_1 \to \Omega_2$ que mapea del espacio de probabilidad al espacio de medida.

Una variable aleatoria se denomina *unidimensional* si $n = 1$ y *multivariante* cuando $n > 1$. En el caso de tener una variable aleatoria multivariante $\mathbf{X} = (X_1, \ldots, X_n)$, se la llamará variable aleatoria conjunta o *vector aleatorio*, y cada X_i con $i = 1, \ldots, n$ se llamará variable aleatoria marginal.

Definición 6 (Probabilidad inducida)**.** Sea $(\Omega_1, \mathcal{A}, P)$ un espacio de probabilidad y $(\Omega_2, \mathcal{S})$ un espacio de medida. La *probabilidad inducida* por una variable aleatoria X viene dada por la función:

$$P_X(S) = P(X^{-1}(S)), \text{ para todo } S \in \mathcal{S}.$$

Consideramos el lanzamiento de una moneda. Los posibles resultados del experimento serán cara o cruz, los cuales serán nuestros sucesos aleatorios. Definiremos nuestra variable aleatoria como:

$$X = \begin{cases} 0 & \text{si sale cara,} \\ 1 & \text{si sale cruz.} \end{cases}$$

Definición 7 (Función de distribución)**.** La *función de distribución* acumulada de una variable aleatoria X es una función $F : \mathbb{R} \to [0, 1]$ definida como:

$$F(x) = P[X \leq x].$$

Proposición 2. *La función de distribución acumulada F asociada a la variable aleatoria X satisface las siguientes propiedades:*

- $\lim\limits_{x \to +\infty} F(x) = 1.$
- $\lim\limits_{x \to -\infty} F(x) = 0.$
- *Es creciente, es decir, si $x_1 \leq x_2$, entonces $F(x_1) \leq F(x_2)$.*
- *Es continua por la derecha, es decir, $\lim\limits_{x \to a^+} F(x) = F(a^+)$.*

Si la imagen de la variable aleatoria X es numerable, diremos que la variable aleatoria es *discreta* y viene descrita por la función de probabilidad p que devuelve la probabilidad de X de ser igual a cierto valor x.

Si la imagen de la variable aleatoria X es infinita no numerable, diremos que la variable aleatoria es *continua* y viene descrita por la función de densidad f que caracteriza la posibilidad relativa de que X tome un valor cercano a x.

Definición 8 (Función de probabilidad)**.** Sea X una variable aleatoria discreta con posibles valores $\{x_1, \ldots, x_n\}$ su *función de probabilidad* se define como

$$f(x) = \begin{cases} P(X = x), & \text{si } x \in \{x_1, \ldots, x_n\}, \\ 0, & \text{en otro caso.} \end{cases}$$

2.2 TEORÍA DEL APRENDIZAJE ESTADÍSTICO

2.2.1 *Introducción a la Teoría del Aprendizaje*

2.2.2 *Paradigma ERM (Empirical Risk Minimization)*

2.2.3 *Dimensión VC (Vapnik-Chervonenkis)*

2.2.3.1 *Desigualdad VC*

2.2.3.2 *Complejidad de Muestra y Generalización*

2.2.4 *Teorema Fundamental del Aprendizaje Estadístico*

2.3 OPTIMIZACIÓN Y MÉTODOS DE ENTRENAMIENTO

2.3.1 *Descenso de Gradiente Estocástico (SGD)*

2.3.2 *Algoritmos de Optimización Avanzados*

2.3.2.1 *Adam*

2.3.2.2 *LAMB*

2.3.3 *Regularización y Técnicas para Evitar el Sobreajuste*

3

MODELOS TRANSFORMER Y ARQUITECTURA

3.1 MODELOS TRANSFORMER

3.1.1 *Introducción a los Modelos Transformer*

3.1.2 *Mecanismo de Atención Escalonada*

3.1.2.1 *Matrices de Consulta, Clave y Valor*

3.1.2.2 *Normalización y Softmax*

3.1.3 *Estructura de Capas en un Transformer*

3.1.4 *Retropropagación en Modelos Transformer*

3.2 TEOREMA DE APROXIMACIÓN UNIVERSAL

3.2.1 *Introducción*

3.2.2 *Teoremas Clave de Aproximación*

3.2.2.1 *Teorema de Hahn-Banach*

3.2.2.2 *Teorema de Representación de Riesz*

3.2.3 *Aplicación del Teorema de Aproximación en Redes Neuronales*

3.2.4 *Aproximación en Modelos Transformer*

4

METODOLOGÍA Y GENERACIÓN DE DATOS SINTÉTICOS

4.1 ANÁLISIS DEL PROBLEMA DE DETECCIÓN DE VEHÍCULOS

4.1.1 *Descripción del conjunto de datos*

4.1.2 *Preprocesamiento*

4.1.3 *Entrenamiento mediante el modelo Transformer*

4.1.4 *Evaluación de los Datos Sintéticos*

5

RESULTADOS Y EVALUACIÓN

5.1 RESULTADOS DE LA GENERACIÓN DE DATOS SINTÉTICOS

5.1.1 *Comparación entre Datos Sintéticos y Reales*

5.1.2 *Métricas de Evaluación: Similitud, Precisión y Varianza*

5.1.3 *Rendimiento del Modelo Transformer*

5.2 EVALUACIÓN DEL SISTEMA IOT

5.2.1 *Impacto de los Datos Sintéticos en la Detección de Vehículos*

5.2.2 *Comparación con Otros Enfoques de Generación de Datos*

5.2.3 *Análisis Comparativo de Rendimiento*

6

CONCLUSIONES Y TRABAJO FUTURO

6.1 CONCLUSIONES

6.1.1 *Principales Contribuciones del Trabajo*

6.1.2 *Impacto del Uso de Datos Sintéticos en IoT*

6.2 TRABAJO FUTURO

6.2.1 *Mejora de los Modelos Transformer*

6.2.2 *Extensiones a Otros Sistemas IoT*

6.2.3 *Posibles Aplicaciones en Diferentes Dominios*

BIBLIOGRAFÍA

[AMMIL12] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012.

[SBEXC] De Suparna, Maria Bermudez-Edo, Honghui Xu, and Zhipeng Cai. Deep generative models in the industrial internet of things: A survey. 18(9):5728–5737.

[sma] Smartpoqueira. Proyecto pionero en Europa en experimentación con tecnología inteligente para gestión sostenible de visitas a zonas rurales y parques nacionales.