REVISTA CIENCIA Y TECNOLOGÍA Para el Desarrollo - UJCM 2019; 5(Número especial):6-11

IMPUTACIÓN DE SERIES DE TIEMPO METEOROLÓGICAS APLICANDO TÉCNICAS DE APRENDIZAJE PROFUNDO *(Deep Learning)*

Aníbal Fernando Flores Garcia[1,a], Otoniel Silva Delgado[1,a]

RESUMEN

El presente artículo muestra los resultados de la implementación de dos técnicas de inteligencia artificial en el campo del Aprendizaje Profundo (*Deep Learning*) correspondiente a las redes neuronales recurrentes conocidas como *Long Short-Term Memory* (LSTM) y *Gated Recurrent Unit* (GRU) para imputar datos faltantes en series de tiempo meteorológicas correspondientes a temperaturas máximas en la región Moquegua. Los resultados alcanzados muestran la superioridad de las redes neuronales recurrentes en la imputación de brechas extensas y muy extensas de datos faltantes respecto a otras técnicas de imputación tradicionales con las que fueron comparadas.

Palabras claves: *Aprendizaje Profundo; Imputación; Series de tiempo; Redes neuronales recurrentes; LSTM; GRU.*

IMPUTATION OF METEOROLOGICAL TIME SERIES APPLYING DEEP LEARNING TECHNIQUES *(Deep Learning)*

ABSTRACT

This paper shows the results of the implementation of two artificial intelligence techniques in the field of Deep Learning corresponding to recurrent neural networks known as *Long Short-Term Memory* (LSTM) and Gated Recurrent Unit (GRU) to impute missing data in meteorological time series corresponding to maximum temperatures in the Moquegua region. The results achieved show the superiority of the recurrent neural networks in the imputation of large and very large gaps of missing values compared to other traditional imputation techniques with which they were compared.

Keywords*: Deep Learning; Imputation; Time series; Recurrent neural networks; LSTM; GRU.*

[1] Universidad José Carlos Mariátegui. Moquegua, Perú.
[a] Magister en docencia universitaria e investigación pedagógica.

Recibido:06-12-2019 Aprobado: 31-12-2019

INTRODUCCIÓN

Las series de tiempo se encuentran presentes en casi todas las áreas de conocimiento, por ejemplo: Biología, Finanzas, Ciencias Sociales, Meteorología, etcétera[1]; y uno de los principales problemas que estas presentan es que en muchos casos poseen datos faltantes, debido a múltiples factores como errores de transmisión de datos, mal funcionamiento de equipos o errores humanos[2]. Los datos faltantes dificultan o hacen casi imposible que las series de tiempo puedan ser utilizadas en actividades de predicción, que es una de las principales aplicaciones de las series de tiempo.

El Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI) posee un repositorio de datos de series de tiempo de distintas variables meteorológicas obtenidos de distintas estaciones en el Perú, entre estas variables se tiene la temperatura mínima y máxima diaria. Estas series de tiempo se encuentran disponibles en el sitio web del SENAMHI (https://www.senamhi.gob.pe/?&p=download-hydrometeorological-data) y el problema principal que presentan es que poseen varios datos faltantes, los mismos que deben completarse a través de un proceso de imputación para que la serie de tiempo pueda utilizarse para tareas de pronóstico. La Figura 1 muestra un proceso de pronóstico de serie de tiempo y enmarca la tarea de imputación dentro de la etapa de homogenización de la serie de tiempo. Aquí es importante resaltar que, la calidad del proceso de imputación influirá directamente en la calidad de los datos pronosticados.[3]

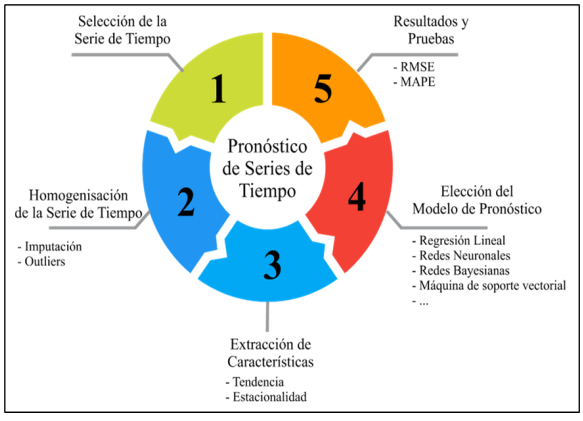

Figura 1. Proceso para pronóstico de series de tiempo.

En el presente artículo se muestran los resultados del análisis de diversas técnicas de imputación para brechas extensas (11 a 30 valores faltantes consecutivos) y brechas muy extensas (más de 30 valores faltantes consecutivos)[4]. Las técnicas que se analizan son las correspondientes a Redes Neuronales Recurrentes como LSTM[5] y GRU[6], que han sido ampliamente utilizadas en procesos de pronóstico de series de tiempo y requieren de bastantes datos históricos para regresiones más precisas. Estas técnicas se han implementado en lenguaje Python utilizando las librerías tensorflow y keras.

Las técnicas basadas en Redes Neuronales Recurrentes mencionadas anteriormente se comparan con otras técnicas basadas en medias móviles como Simple Moving Average (SMA), Linear Weighted Moving Average LWMA y Exponential Weighted Moving Average EWMA y Autoregressive Integrated Moving Average[1], así como la técnica de pronóstico de Facebook conocida como Prophet[7], las técnicas de imputación mencionadas en el presente párrafo se implementaron en lenguaje R.

Así el presente trabajo tiene como objetivo general:

- Implementar técnicas de imputación de series de tiempo meteorológicas aplicando redes neuronales recurrentes.

Y como objetivos específicos:

- Implementar el algoritmo Long Short-Term Memory para imputación de datos faltantes en series de tiempo de temperaturas máximas.
- Implementar el algoritmo Gated Recurrent Unit para imputación de datos faltantes en series de tiempo de temperaturas máximas.
- Evaluar la precisión de los algoritmos LSTM y GRU en la imputación de datos faltantes en series de tiempo de temperaturas máximas.

La hipótesis general es:

- Es posible implementar redes neuronales recurrentes para imputar datos faltantes en series de tiempo meteorológicas.

Y las hipótesis específicas son:

- Es posible implementar el algoritmo Long Short-Term Memory para imputación de datos faltantes en series de tiempo de temperaturas máximas.
- Es posible implementar el algoritmo Gated Recurrent Unit para imputación de datos faltantes en series de tiempo de temperaturas máximas.
- En promedio los algoritmos LSTM y GRU en la imputación de datos faltantes en series de tiempo de temperaturas máximas presentan una mayor precisión que otras técnicas de imputación.

MATERIALES Y MÉTODOS

Para el procesamiento de los datos de series de tiempo meteorológicas se requirió de un computador con procesador Core I7 y 8GB de RAM. Se utilizó el lenguaje de Programación Python en su versión 3.6 con las librerías tensorflow y keras. Así mismo, se utilizó el lenguaje R y la herramienta de R Studio en su versión 1.1.456 para implementar las técnicas SMA, LWMA, EWMA, ARIMA y prophet con las librerías imputeTS y prophet.

Para analizar la precisión de las técnicas de imputación se prepararon 7 casos de estudio, 3 de ellas correspondientes a brechas extensas (11 a 30 NAs) y 4 para brechas muy extensas (+30 NAs). El proceso para la imputación de series de tiempo de temperaturas máximas utilizando redes neuronales recurrentes se muestra en la Figura 2 y se describe seguidamente.

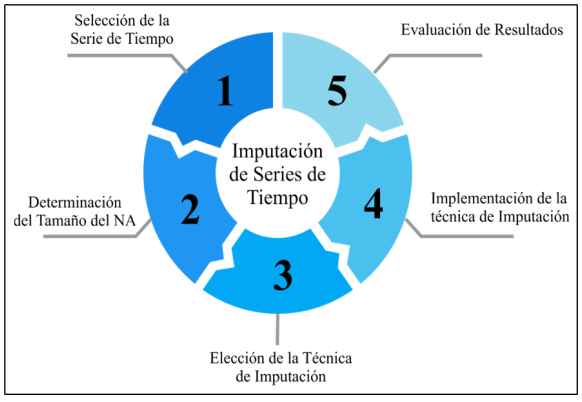

Figura 2. Proceso de imputación propuesto.

Selección de la Serie de Tiempo

En esta etapa se selecciona la serie de tiempo que posee datos faltantes, de los cuales se utilizará los datos existentes para entrenar la red neuronal recurrente LSTM or GRU. La serie de tiempo seleccionada corresponde a temperaturas máximas de la estación meteorológica Punta de Coles ubicada en la provincia de Ilo cuyos datos fueron descargados del repositorio web del SENAMHI en la siguiente dirección:

https://www.senamhi.gob.pe/?&p = download-hydrometeorological-data.

Determinación del tamaño de la brecha de NAs.

Esta actividad es muy importante ya que permitirá determinar el tamaño de la brecha y a partir de ella elegir la técnica de imputación más adecuada.

La serie de tiempo elegida posee datos diarios desde el año 1954 hasta el año 2018, la cantidad de datos faltantes corresponden a 2604 días, existen números brechas pequeñas (1 ó 2 NAs) como también brechas medias (3 a 10 NAs) y extensas (11 a 30 NAs), así como también existen brechas muy extensas (+30 NAs), entre ellas existe una brecha de NAs de 1978 días entre el 01-01-1960 al 31-05-1961.

En el presente trabajo se trabajó con brechas de 11, 21, 30, 60, 90, 120 y 150 días para evaluar la precisión de las técnicas de imputación.

Elección de la técnica de Imputación.

De acuerdo con los tamaños de brechas de NAs elegidos y de acuerdo con la precisión reportada para el uso de LSTM y GRU en diversos trabajos de investigación de pronóstico en diversas series de tiempo es que se decidió optar por la implementación de estas dos técnicas para el proceso de imputación.

Implementación de la técnica de imputación

Primero, se implementa la red neuronal recurrente LSTM con la arquitectura que se muestra en la Fig. 3

```
model = Sequential()

model.add(LSTM(units=50, return_sequences=True, input_shape=(features_set.shape[1], 1)))
model.add(Dropout(0.2))

model.add(LSTM(units=50, return_sequences=True))
model.add(Dropout(0.2))

model.add(LSTM(units=50, return_sequences=True))
model.add(Dropout(0.2))

model.add(LSTM(units=50))
model.add(Dropout(0.2))
model.add(Dense(units = 1))

model.compile(optimizer = 'adam', loss = 'mean_squared_error')
model.fit(features_set, labels, epochs = 20, batch_size = 32)
```

Figura 3. Arquitectura de LSTM.

Seguidamente, se implementa la red neuronal recurrente GRU cuya arquitectura es muy similar a la que se utilizó para LSTM y ésta se muestra en la Figura 4.

```
model = Sequential()

model.add(GRU(units=50, return_sequences=True, input_shape=(features_set.shape[1], 1)))
model.add(Dropout(0.2))

model.add(GRU(units=50, return_sequences=True))
model.add(Dropout(0.2))

model.add(GRU(units=50, return_sequences=True))
model.add(Dropout(0.2))

model.add(GRU(units=50))
model.add(Dropout(0.2))
model.add(Dense(units = 1))

model.compile(optimizer = 'adam', loss = 'mean_squared_error')
model.fit(features_set, labels, epochs = 20, batch_size = 32)
```

Figura 4. Arquitectura de GRU.

Una vez implementados los modelos estos se compilan y se estima los valores para las brechas de NAs.

Evaluación de Resultados de la Imputación

Los resultados de los modelos de regresión se evalúan con diversas técnicas como RMSE, MSE, MAPE, etc. En el presente estudio se utiliza la Raíz del error cuadrático medio (RMSE) que se calcula o determina a través de la ecuación.[1]

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n-1}(Pi - Ri)^2}{n}} \quad (1)$$

Donde:

Pi :Vector de los valores pronosticados.
Ri :Vector de los valores reales.
n :Número de elementos del vector *Pi* o *Ri*.

RESULTADOS

Los resultados alcanzados por las técnicas de aprendizaje profundo LSTM y GRU se muestran en la Tabla 1.

Tabla 1. Comparación entre LSTM y GRU.

Técnica	Número de NAs consecutivos							Promedio
	11	21	30	60	90	120	150	
LSTM	0,6156	0,6820	0,7579	0,8858	0,9049	0,9442	0,9907	0,8258
GRU	0,6749	0,6503	0,7262	0,7252	0,7355	0,7801	0,7559	0,7211

Fuente: Elaboración propia.

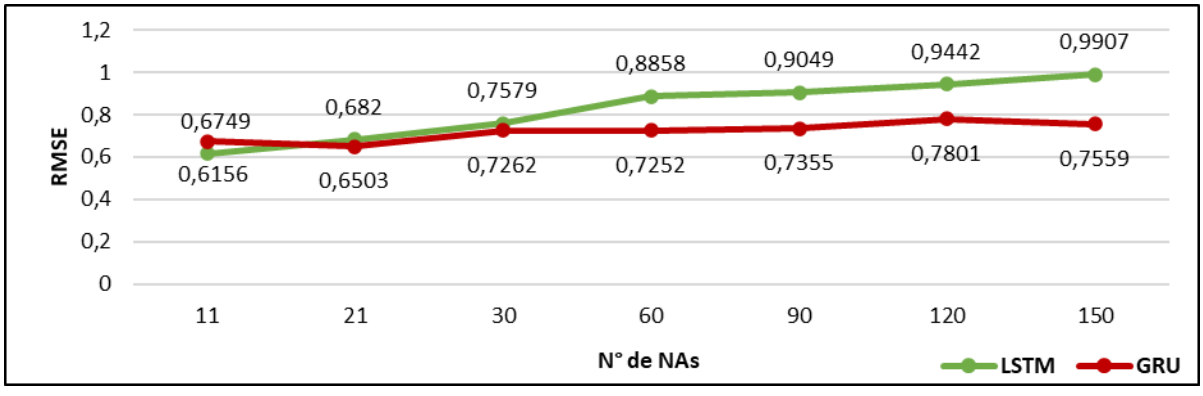

Figura 5. LSTM vs GRU.
Fuente: Elaboración Propia.

De acuerdo con la Tabla 1 y Figura 5, se aprecia que la Red Neuronal Recurrente GRU es superior a LSTM en la mayoría de los casos de estudio en promedio alcanzó un RMSE de 0.7211 respecto a LSTM que alcanzó un RMSE promedio de 0.8258.

DISCUSIÓN

En esta sección se compara los resultados alcanzados por las redes neuronales recurrentes LSTM y GRU con otros algoritmos de pronóstico e imputación y los resultados se muestran en la Tabla 2 y Figura 6.

Tabla 1. Comparación de LSTM, GRU con otras técnicas de imputación.

Técnica	Número de NAs consecutivos							RMSE Promedio
	11	21	30	60	90	120	150	
SMA	0,6082	0,6365	0,5821	0,8843	1,7482	1,7106	1,7169	1,1266
LWMA	0,6091	0,4032	0,7763	0,9478	1,8521	1,7073	1,7242	1,1457
EWMA	0,6093	0,4249	1,0513	1,3155	2,0134	1,9693	2,2743	1,3797
ARIMA	0,6748	1,0424	1,4165	2,2932	2,5240	2,2320	2,1639	1,7638
Prophet	0,5991	0,6477	0,7652	1,0516	1,1637	1,1274	1,0279	0,9118
LSTM	0,6156	0,6820	0,7579	0,8858	0,9049	0,9442	0,9907	0,8258
GRU	0,6749	0,6503	0,7262	0,7252	0,7355	0,7801	0,7559	0,7211

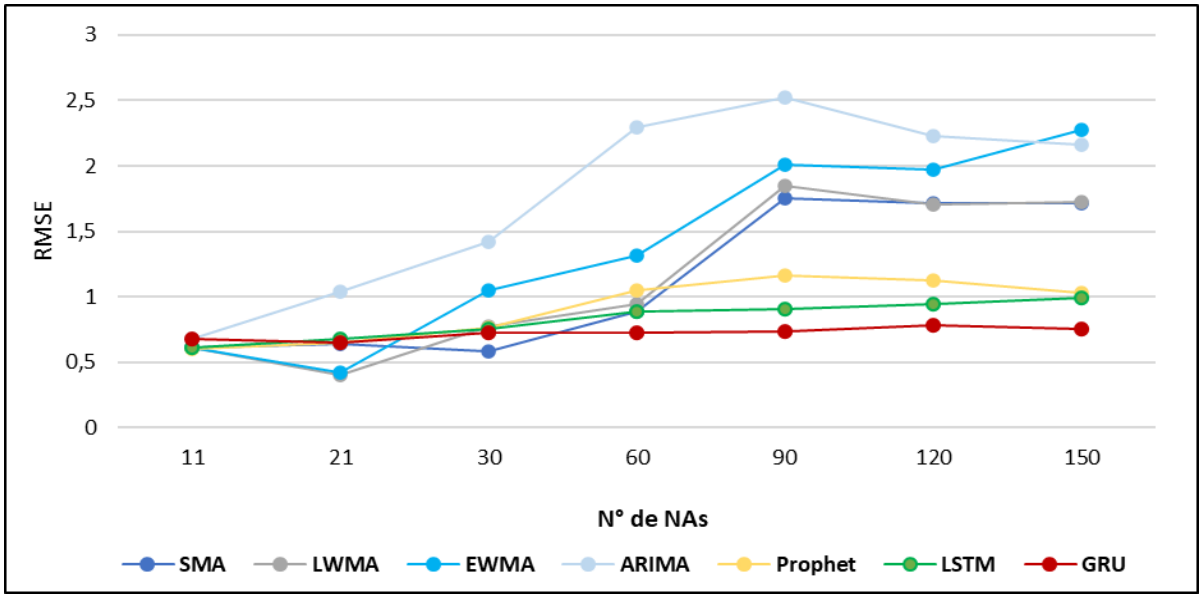

Figura 6. Comparación de LSTM, GRU con otras técnicas.

De acuerdo con la Tabla 2 y Figura 6, se puede apreciar que las técnicas LSTM (RMSE 0.8258) y GRU (RMSE 0.7211) en promedio son superiores al resto de técnicas, éstas superan en precisión en la mayoría de los casos a las demás. LSTM y GRU mantienen una precisión uniforme en la predicción de distintos tamaños de brechas de datos con respecto a otras técnicas basadas en medias móviles que para brechas mayores a 30 días muestran un RMSE bastante alto. De acuerdo con los resultados obtenidos, las redes neuronales recurrentes LSTM y GRU estudiadas en el presente trabajo son muy recomendables para la imputación de extensas y muy extensas brechas de datos faltantes en series de tiempo meteorológicas como las correspondientes a temperaturas máximas.

CONCLUSIONES

Conclusión General

- Se ha logrado implementar técnicas de imputación de series de tiempo meteorológicas aplicando redes neuronales recurrentes logrando una buena precisión en las estimaciones de datos faltantes para brechas extensas y muy extensas.

Conclusiones Específicas:

- Se ha logrado implementar el algoritmo *Long Short-Term Memory* para imputación de datos faltantes en series de tiempo de temperaturas máximas logrando una precisión muy superior respecto a otras técnicas de imputación.

- Se ha logrado implementar el algoritmo Gated Recurrent Unit para imputación de datos faltantes en series de tiempo de temperaturas máximas. La precisión alcanzada por esta técnica es superior a la alcanzada por LSTM.

- Se ha comparado la precisión de los algoritmos LSTM y GRU en la imputación de brechas de datos faltantes extensas y muy extensas en series de tiempo de temperaturas máximas mostrando superioridad sobre otros algoritmos de imputación como Prophet, ARIMA, SMA, LWMA y EWMA. Mientras más extensa la brecha de datos faltantes la precisión es más notoria.

REFERENCIAS BIBLIOGRÁFICAS

1. Moritz, Steffen y Bartz-Beielstein, ImputeTS: Time series missing value imputation in R. Thomas. The R Journal, Junio de 2017, Vol. 9, págs. 207-218.
2. Chia-Yang, Chang, Cheng-Ru, Wang y Shie-Jue, Lee Novel imputation for time series data.. Guangzhou, 2015 Proceedings of the International Conference on Machine Learning and Cybernetics.: IEEE, 2015.
3. Flores, Anibal, Tito, Hugo y Silva, Carlos. CBRi: A case based reasoning-inspired approach for univariate time series imputation. Guayaquil - Ecuador: IEEE, 2019. IEEE Latin American Conference on Computational Intelligence.
4. Flores, Anibal, Tito, Hugo y Silva, Carlos. Local average of nearest neighbors: univariate time series imputation. 8, International Journal of Advanced Computer Science and Applications, 2019,Vol. 10.
5. Hochreiter, Sepp y Schmidhuber, Jürgen Long short-term memory. Neural Computation. 1997, págs. 1735-1780.
6. Cho, Kyunghyun, y otros. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014, arxiv.org.
7. Taylor, Sean J y Letham, Benjamin Forecasting at scale.. PeerJpreprints. 2017.

***Correspondencia**: Aníbal Fernando Flores García.*
***Dirección**: Universidad José Carlos Mariátegui. Moquegua, Ciudad Universitaria; Moquegua – Perú.*
***Correo electrónico**: anibalf11@hotmail.com*