Universidad Michoacana de San Nicolás de Hidalgo
Departamento de estudios de posgrado
Facultad de Ingeniería Eléctrica

Modelado del error de predicción en series de tiempo basado en la calidad de sus datos

Propuesta que para obtener el grado de:
Maestro en Ciencias en Ingeniería Eléctrica
Especialidad en:
Sistemas Computacionales

Presenta:
Ing. Víctor Manuel Téllez Velázquez

Director de Tesis:
Dr. Juan José Flores Romero

30 de agosto del 2019

DIVISIÓN POSGRADO

Contenido

Resumen

1- Introducción
- 1.1- Antecedentes
- 1.2- Objetivos
  - 1.2.1- Objetivos Particulares
- 1.3- Justificación

2- Extracción de características de una serie de tiempo
- 2.1- Datos faltantes
- 2.2- Valores atípicos
- 2.3- Ruido
- 2.4- Caos

3- Pronóstico de las series de tiempo sintéticas perturbadas

4- Regresión de error de pronóstico

Resultados

Conclusiones

Trabajos futuros

Resumen

La propuesta está dirigida a obtener una metodología para determinar la calidad de los datos de una serie de tiempo, modelando el error de predicción. Con la extracción características de la serie de tiempo y realizando el pronóstico, usar un regresor para determinar el error de pronóstico de una determinada serie de tiempo, sin hacer un modelado de pronóstico.

Introducción

¿Qué es una serie de tiempo?
$st = \{x_1, x_2, \ldots, x_N\}$ $N \geq 1$

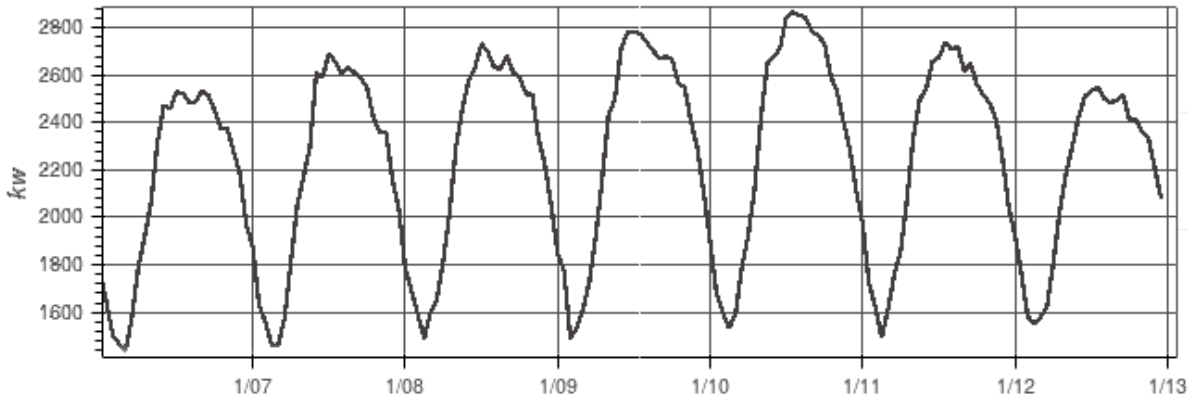


Figura: Serie de tiempo de la potencia real consumida en un circuito eléctrico de distribución, mediciones horarias en el periodo del 6 al 13 de enero de 2019

Antecedentes

Después de una revisión exhaustiva y en los artículos encontrados no se encontró una investigación previa a la propuesta en esta tesis.

Diversos trabajos aportan diferentes conceptos para determinar la calidad de los datos en está tesis.

- Encuentran datos faltantes en una serie de tiempo y se imputan.
- Detección de valores atípicos locales.
- Comparación de las técnicas de pronóstico de series de tiempo con respecto a la tolerancia al ruido.
- Estimación de la amplitud del ruido.
- Determinar la magnitud del caos se obtiene con el mayor exponente de Lyapunov. con el algoritmo de Rosenstein.

Objetivo general

El objetivo general es determinar la calidad de los datos de una serie de tiempo, modelando el error de predicción.

Objetivos:
Objetivos Particulares

- Generar 10 series de tiempo sintéticas y perturbarlas con diferentes niveles de datos faltantes, valores atípicos y ruido. En total se generarán 2,160 series de tiempo.
- Desarrollar la extracción de las cuatro características (Datos Faltantes, valores atípicos, ruido y caos).
- Diseñar modelos de perceptrones multicapa para realizar el pronóstico de las series de tiempo y obtener el error de predicción.
- Diseñar un regresor de bosques aleatorios que se ajuste a los datos y de una aproximación a la precisión del pronóstico sin necesidad de hacer todo el modelado.

Justificación

El problema de conocer la calidad de los datos de una serie de tiempo es un tema que se plantea al momento de preguntar ¿Cual es la precisión del pronóstico que se obtendrá con sus datos? La mayoría de las veces se pide un compromiso de confiabilidad en el pronóstico. Esta investigación se centra en evitar realizar el proceso de modelado de pronóstico, que puede llegar a tardar días hasta semanas para ver el resultado. Con este método que se propone se podrá obtener de manera automática y rápida una aproximación del error de predicción. Esta aproximación depende directamente de la calidad de los datos de la serie de tiempo.

Series de tiempo sintéticas

Para poder realizar la extracción de características, se usó series de tiempo sintéticas porque no presentan perturbaciones, y por tal motivo se pueden agregar las perturbaciones controladas.

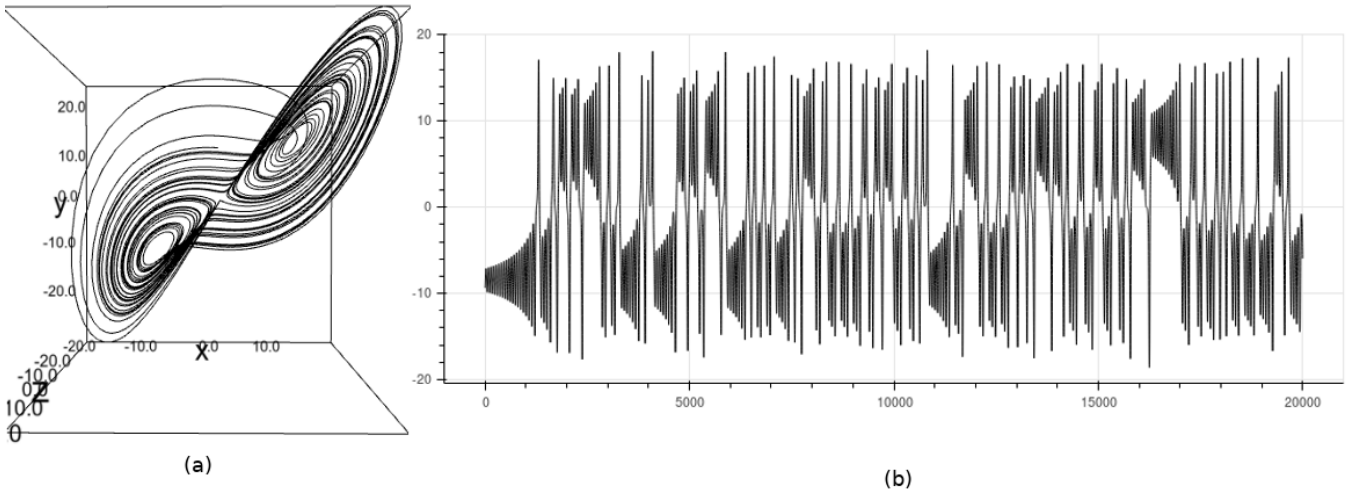


Figura: Atractor de Lorenz: (a) Trayectoria. (b) Serie de tiempo de la variable X

Datos faltantes

En una serie de tiempo es casi imposible tener un conjunto de datos completo.

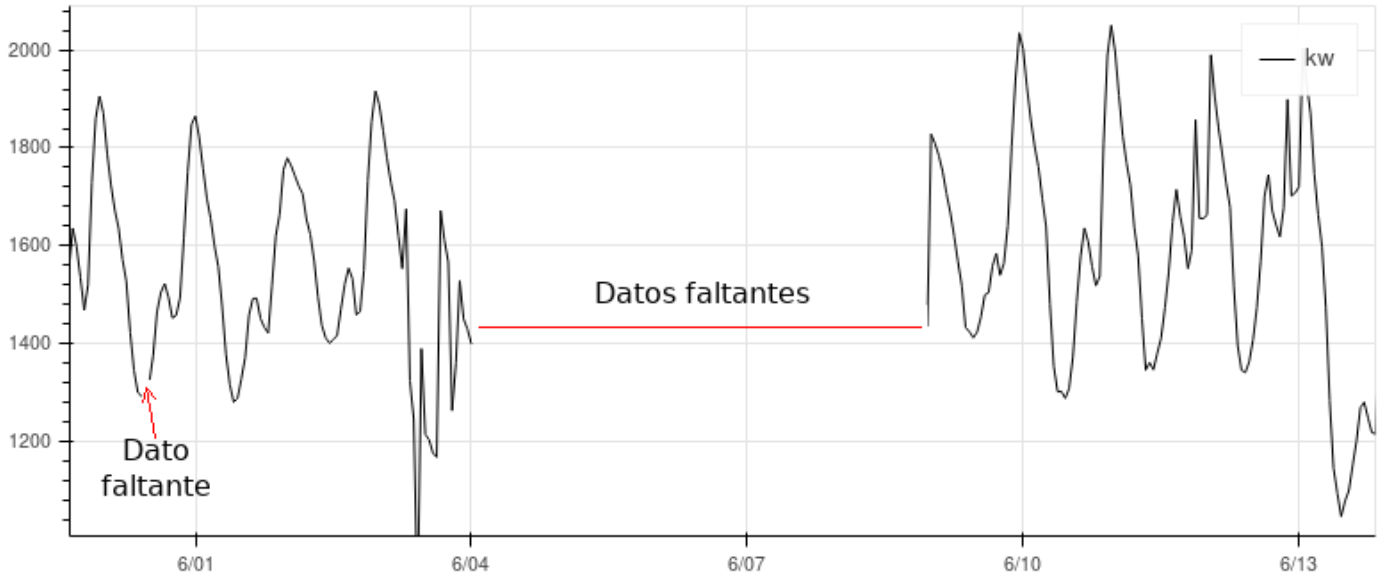


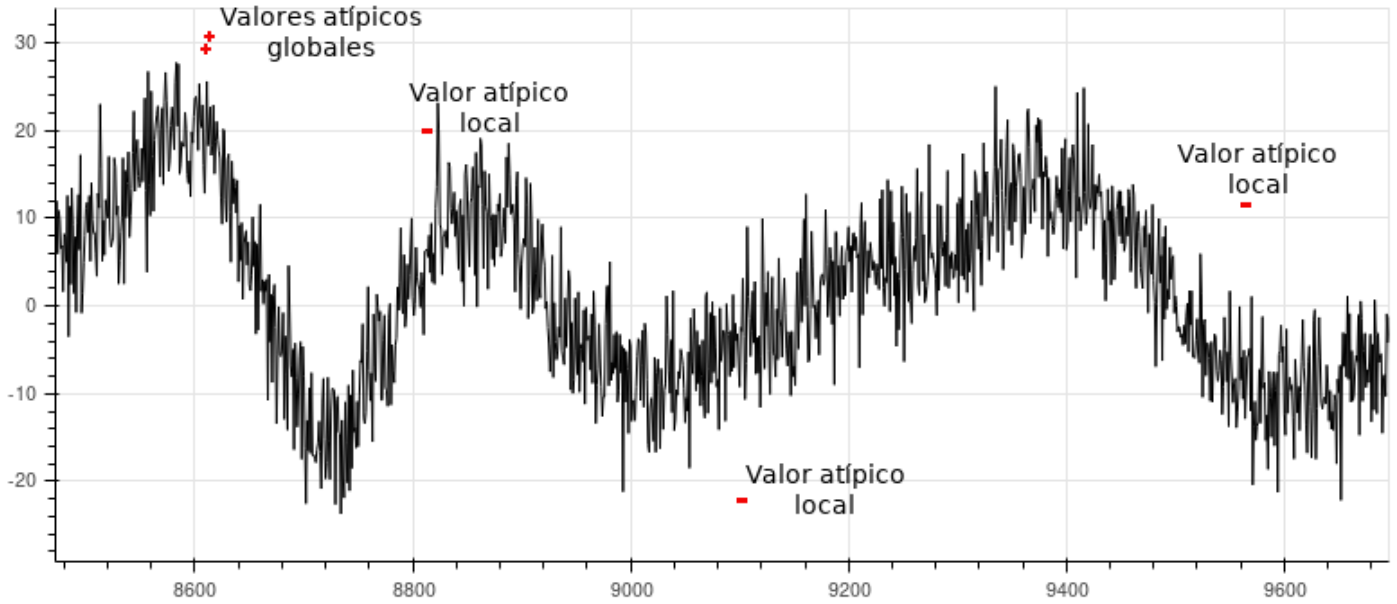
Figura: Datos faltantes dentro de una serie de tiempo

Imputación

La imputación de datos es el proceso de corregir los datos faltantes con un proceso de estimación (método de la última observación arrastrada).

Valores atípicos

Un valor atípico es una observación que se desvía tanto de las otras observaciones como para despertar sospechas de que fue generado por un mecanismo diferente.



Valores atípicos globales

Un dato se considera como valor atípico global si se desvía significativamente del resto del conjunto de datos.

Algoritmo 4: Detección de valores atípicos globales

```
1  deteccionGlobal(st)
2    indices ← []
3    mu ← media(st)
4    sigma ← desEst(st)
5    limSup ← mu + (3 * sigma)
6    limInf ← mu − (3 * sigma)
7    for i en range(len(st)) do
8      if st[i] > limSup or st[i] < limInf
9        indices.append(i)
10   return indices
```

Valores atípicos locales

Los valores atípicos locales son más difíciles de detectar que los globales, porque para encontrarlos es necesario conocer la naturaleza de la serie de tiempo.

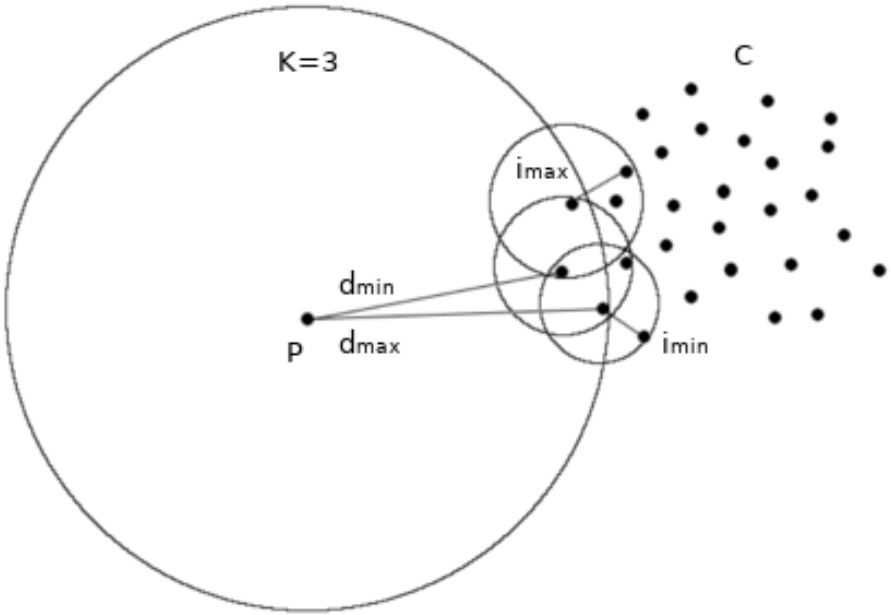


Figura: Cálculo de LOF

Inclusión de valores atípicos

La inclusión de valores atípicos globales y locales en las series de tiempo sintéticas se refiere a modificar el valor de los datos de la serie de tiempo.

Ruido

La razón señal ruido por sus siglas en inglés (SNR Signal to Noise Ratio) es una medida que se define como la razón entre la potencia de la señal y la potencia de ruido, a menudo expresada en decibeles(dB).

$$SNR_{dB} = 10\ log_{10}\left(\frac{P_{señal}}{P_{ruido}}\right) \tag{1}$$

Medias móviles centradas.

$$MA_T = \frac{\sum_{j=1}^{(M-1)/2} y_{T-j} + y_T + \sum_{j=1}^{(M-1)/2} y_{T+j}}{M} \tag{2}$$

Ruido

Funcionamiento de las medias móviles, primero 15 datos de la serie de tiempo del sistema Chen.

Valores de la serie de Tiempo	Medias Móviles Ventana = 2	Valores de la serie de Tiempo	Medias Móviles Ventana = 7
-11.200589	NaN	-11.200589	NaN
-7.116236	-9.158412	-7.116236	NaN
-7.336671	-7.226453	-7.336671	NaN
-17.86384	-12.600255	-17.86384	-10.834262
-9.435985	-13.649913	-9.435985	-10.683214
-8.129631	-8.782808	-8.129631	-10.088738
-14.756883	-11.443257	-14.756883	-9.468307
-10.143256	-12.45007	-10.143256	-7.872498
-2.954903	-6.54908	-2.954903	-8.131058
-2.993648	-2.974276	-2.993648	-8.220635
-6.693178	-4.843413	-6.693178	-6.509206
-11.245909	-8.969543	-11.245909	-6.263438
-8.756668	-10.001289	-8.756668	NaN
-2.776876	-5.766772	-2.776876	NaN
-8.422884	-5.59988	-8.422884	NaN

Figura: Funcionamiento de medias móviles centradas con una ventana de 2 y 7.

Inclusión de ruido

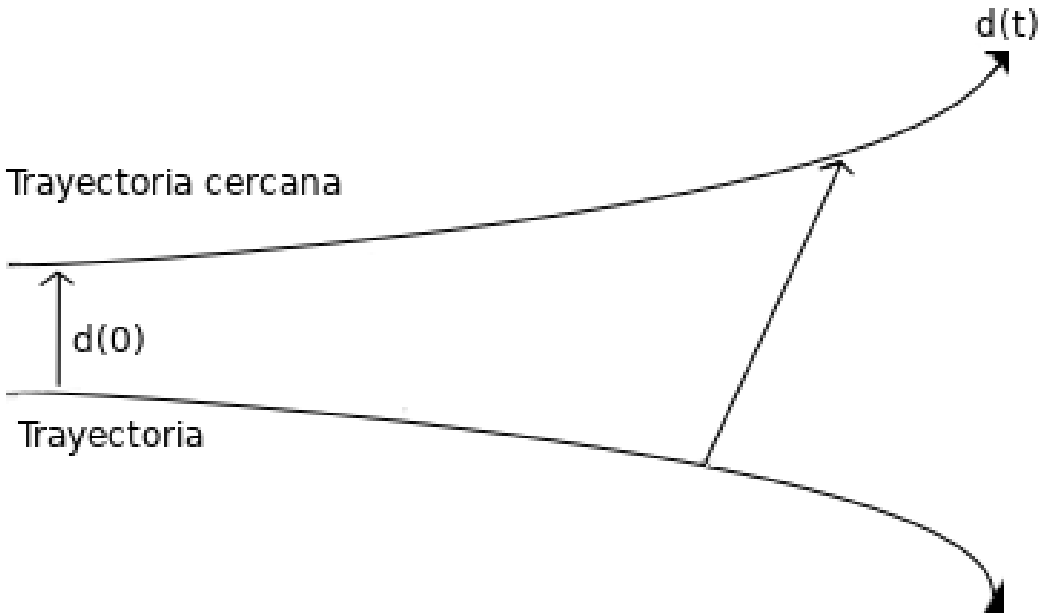
El nivel de ruido inyectado con la función AWGN (Añadir ruido gaussiano blanco a la señal) implementada en Python.

Caos

Determinar la presencia de caos en una serie de tiempo es un problema que se resuelve con el cálculo de los exponentes de Lyapunov. Si $\lambda < 0$ indica que el sistema no presenta caos. Si $\lambda = 0$ indica que el sistema está en estado estable sin caos. Si $\lambda > 0$ indica que el sistema presenta caos.

Caos

Divergencia de las trayectorias.

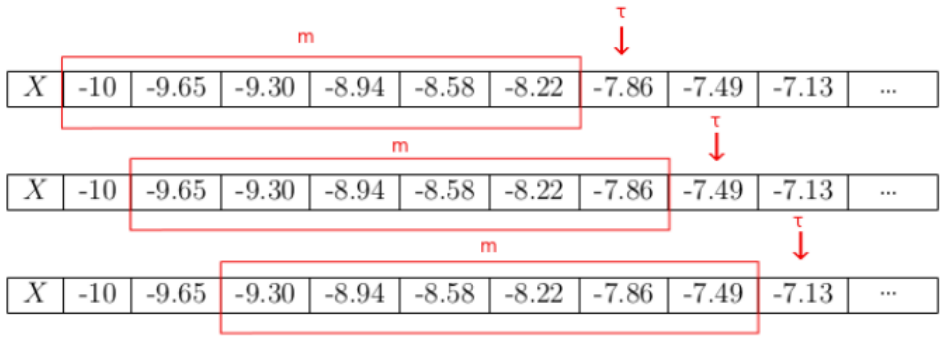


Caos

Método de Rosenstein para calcular la magnitud del caos obteniendo el máximo exponente de Lyapunov, implementado en la biblioteca de Nolds.

Pronóstico.

El pronóstico de una serie de tiempo es un problema que barca muchos campos. Los problemas de pronóstico son clasificados a corto, mediano y largo plazo.
Se utilizaron perceptrones multicapa que son un tipo de redes neuronales, para realizar el pronóstico de las series de tiempo.



X						y
-10.00	-9.65	-9.30	-8.94	-8.58	-8.22	-7.86
-9.65	-9.30	-8.94	-8.58	-8.22	-7.86	-7.49
-9.30	-8.94	-8.58	-8.22	-7.86	-7.49	-7.13
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figura: Transformación de los datos.

Redes neuronales MLP

Nombre	Relación entradas/salidas	F. en Keras	Figura
Positiva lineal	$\mathrm{a} = 0 \quad h < 0$ $\mathrm{a} = \mathrm{h} \quad 0 \leq \mathrm{h}$	relu	
Sigmoide	$a = \frac{1}{1+e^{-h}}$	sigmoid	
Lineal	$a = h$	linear	
Tangente hiperbólica	$\mathrm{a} = \frac{e^{h}-e^{-h}}{e^{h}+e^{-h}}$	tanh	

Figura: Funciones de activación de las neuronas.

Redes neuronales MLP

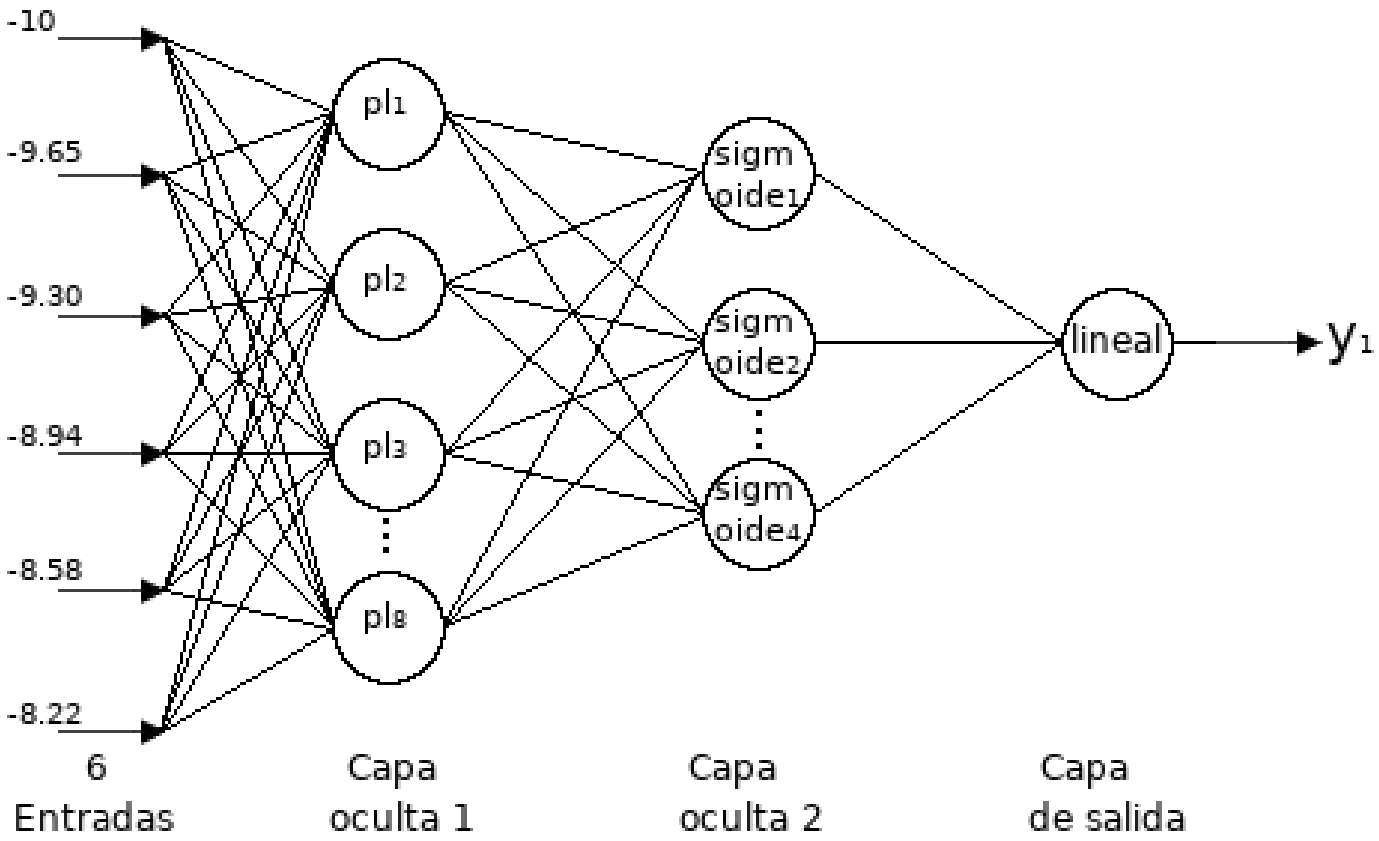


Figura: Arquitectura del perceptrón multicapa.

Regresor bosques aleatorios

El método de regresión con bosques aleatorios pertenece al área de aprendizaje máquina.

El algoritmo de bosques aleatorios genera un conjunto de árboles de decisión y aplica la técnica de boosting para calcular el resultado final. Esto es, promedia los resultados de regresión de los arboles en el bosque aleatorio.

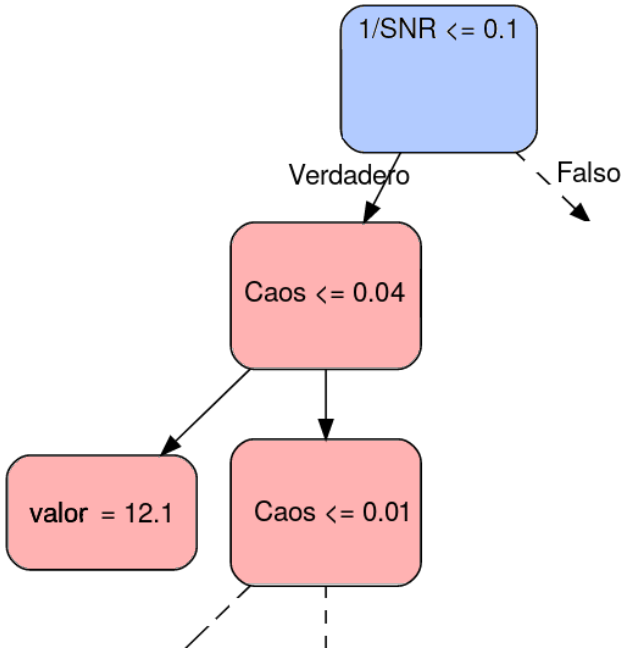


Figura: Porción de un árbol de decisión.

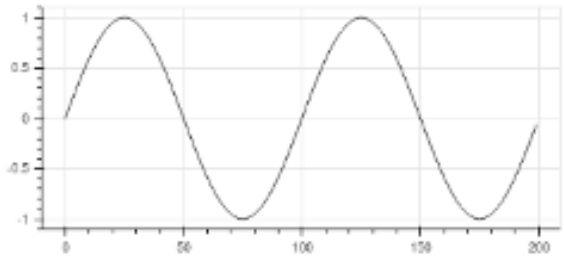
Regresor bosques aleatorios

Algoritmo 10: Regresor de bosques aleatorios

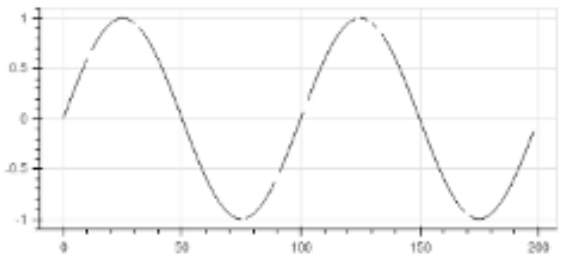
```
1 regBosquesAleatorios(numAr, entreCa, entreOb, validaCa, validaOb)

2   regresorBA ← RandomForestRegressor(n_estimators =
    numAr, criterion ="mse", n_jobs = −1)
3   regresorBA.fit(entreCa, entreObjetivo);
4   predicciones ← regresorBA.predict(validaCaracte)
5   mae ← MAE(predicciones, validaOb)
6   mape ← MAPE(predicciones, validaOb)
7   precisionR ← 100 − mape
8   return regresorBA, mae, mape, precisionR
```

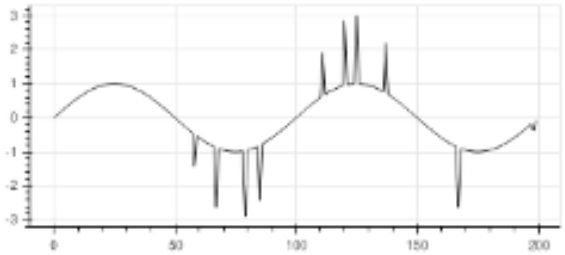
Resultados perturbaciones y datos faltantes



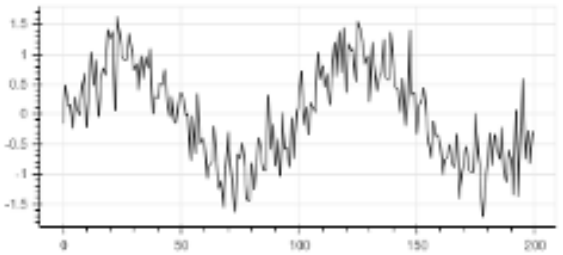
(a) Sin perturbaciones 0_0_0



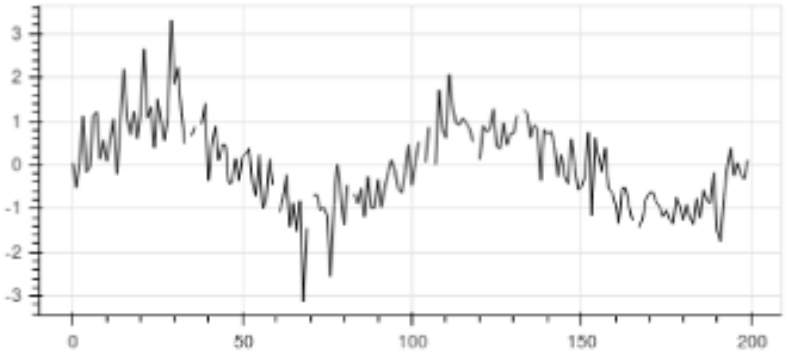
(b) Perturbada con datos faltantes 5_0_0



(c) Perturbada con valores atípicos 0_5_0

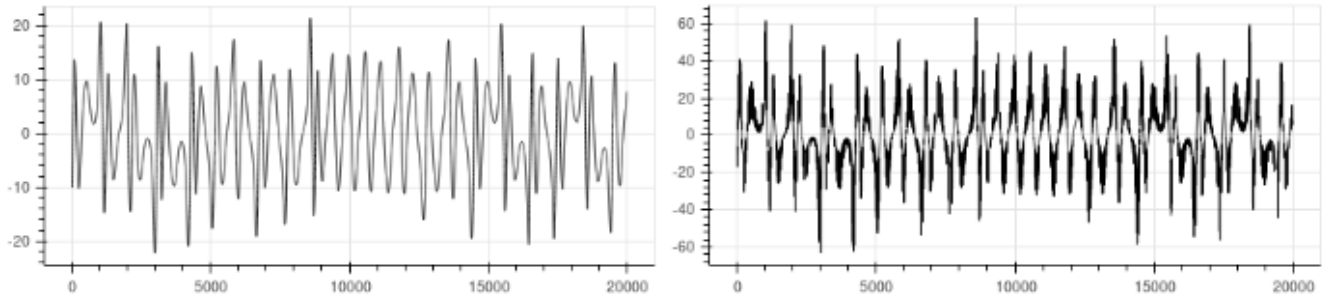


(d) Perturbada con ruido 0_0_5



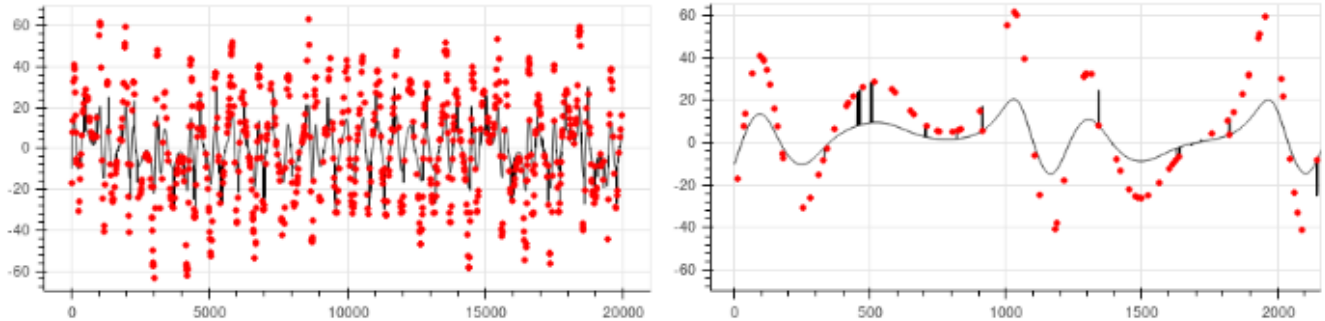
(e) Perturbada con datos faltantes, valores atípicos y ruido 5_5_5

Res. Extracción del porcentaje de valores atípicos



(a) Sin perturbaciones

(b) Perturbada 0_5_0



(c) Extraccion de los valores atípicos los puntos rojos son los valores atípicos

(d) Muestra de los primero 2000 datos

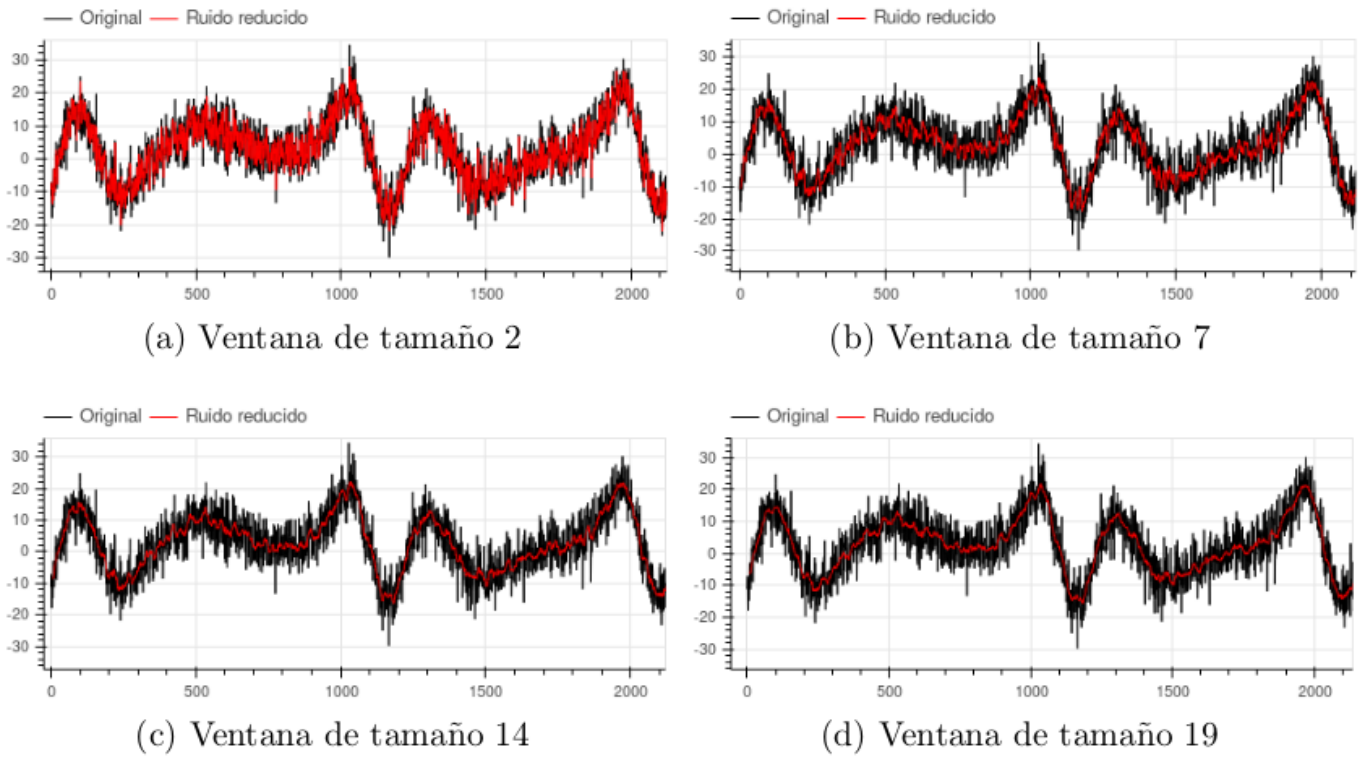
Figura: Serie de tiempo sistema Chen.

Res. Extracción del porcentaje de valores atípicos

Extracción de los valores atípicos de las series de tiempo perturbadas con 5, 10, 15, 20 y 25% de valores atípicos y cálculo del error MAE.

Serie de tiempo	Perturbaciones					MAE
	0_5_0	0_10_0	0_15_0	0_20_0	0_25_0	
Sistema Chen	3%	5%	4%	4%	3%	11.2
Oscilador Duffing	2%	2%	2%	2%	1%	13.2
Atractor cíclico simétrico de Halvorsen	3%	5%	5%	4%	3%	11
Atractor de Lorenz	4%	5%	6%	6%	5%	9.8
Atractor de Rössler	3%	5%	5%	4%	3%	11
Atractor de Rucklidge	2%	3%	3%	3%	3%	12.2
Oscilador Shawn-van der Pol	2%	3%	3%	3%	2%	12.4
Flujo cúbico más simple	3%	4%	4%	3%	2%	11.8
Flujo lineal por partes más simple	2%	3%	3%	3%	2%	12.4
Seno	2%	2%	0%	0%	0%	14.2

Res. Extracción del nivel de ruido



(a) Ventana de tamaño 2

(b) Ventana de tamaño 7

(c) Ventana de tamaño 14

(d) Ventana de tamaño 19

Figura: Extracción del ruido de la serie de tiempo sistema Chen.

Res. Extracción del nivel de ruido

Extracción del ruido de las series de tiempo perturbadas con 5, 10, 15, 20 y 25dB de ruido y cálculo del error MAE.

Serie de tiempo	Perturbaciones					MAE
	0_0_5	0_0_10	0_0_15	0_0_20	0_0_25	
Sistema Chen	5.31dB	10.27dB	15.18dB	20.23dB	25.15dB	0.23
Oscilador Duffing	5.29dB	10.29dB	15.26dB	20.60dB	26.68dB	0.62
Atractor cíclico simétrico de Halvorsen	3.45dB	8.30dB	13.37dB	18.33dB	23.34dB	1.63
Atractor de Lorenz	4.77dB	9.56dB	13.92dB	17.55dB	19.75dB	1.88
Atractor de Rössler	5.16dB	10.21dB	15.20dB	20.18dB	25.25dB	0.20
Atractor de Rucklidge	5.28dB	10.18dB	15.14dB	20.06dB	24.99dB	0.13
Oscilador Shawn-van der Pol	5.24dB	10.18dB	15.21dB	20.24dB	25.78dB	0.33
Flujo cúbico más simple	5.32dB	10.19dB	15.16dB	20.23dB	25.29dB	0.24
Flujo lineal por partes más simple	4.00dB	8.87dB	13.81dB	18.86dB	24.87dB	0.91
Seno	4.94dB	10.65dB	16.80dB	21.22dB	25.00dB	0.74

Figura: Resultados de la extracción del ruido.

Resultados del pronóstico

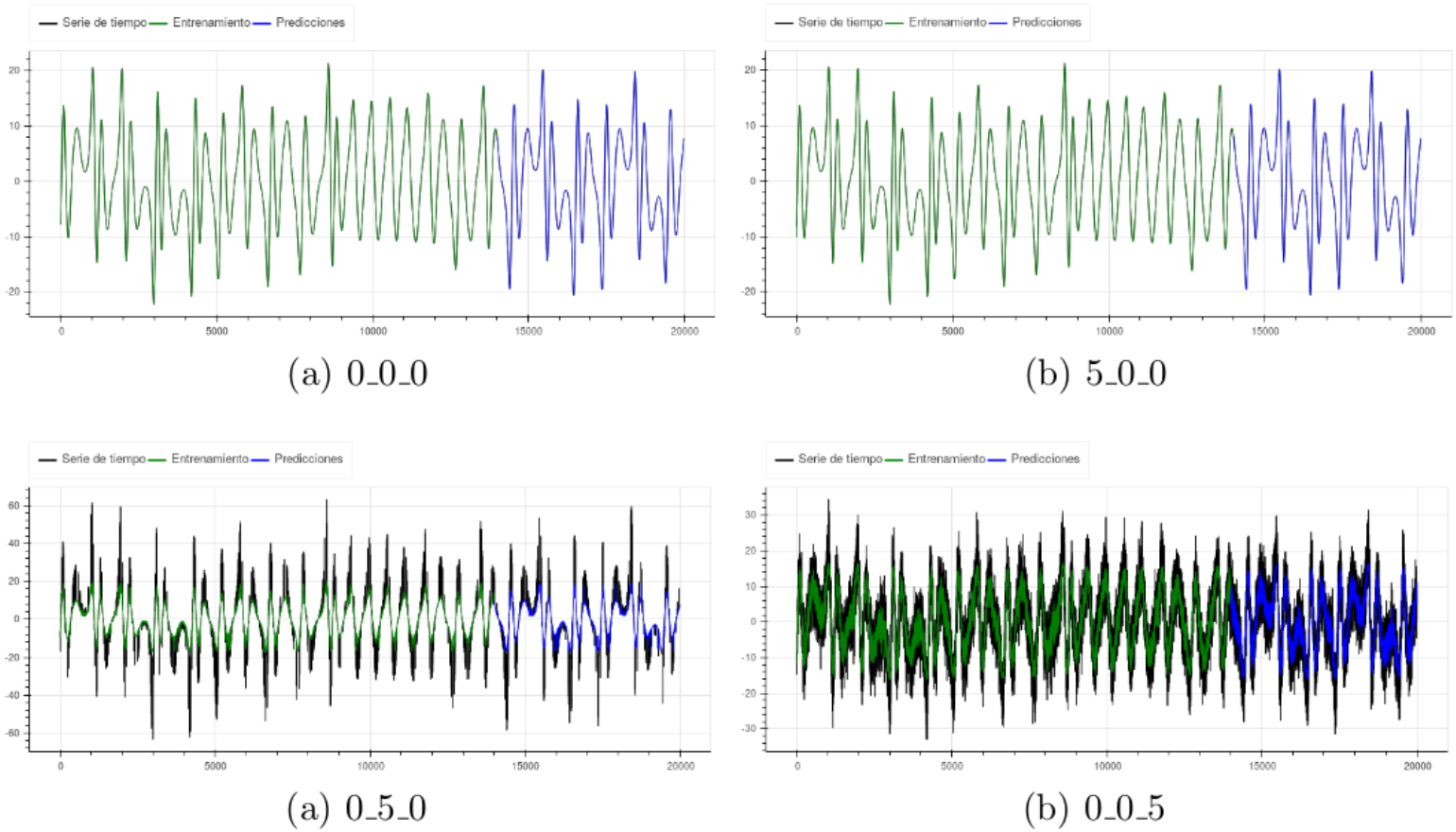
Errores SMAPE en el entrenamiento (E) y en la validación (V).

	Perturbaciones							
	0_0_0		0_0_5		0_5_0		5_0_0	
Serie de tiempo	E	V	E	V	E	V	E	V
Sistema Chen	1.28	1.12	77.68	77.95	16.98	16.03	3.98	3.43
Oscilador Duffing	3.40	4.71	74.01	76.18	18.86	19.34	1.63	1.87
Atractor cíclico simétrico de Halvorsen	2.45	3.18	73.89	78.06	13.21	14.91	1.54	2.23
Atractor de Lorenz	6.59	7.06	81.90	83.71	24.26	25.53	5.19	5.84
Atractor de Rössler	1.07	1.01	83.15	77.18	20.24	20.11	2.24	2.03
Atractor de Rucklidge	2.36	2.04	100.58	93.72	22.74	20.09	5.41	4.76
Oscilador Shawn-van der Pol	2.91	2.96	72.02	72.56	16.11	16.59	1.28	1.27
Flujo cúbico más simple	2.71	2.71	74.57	73.53	17.37	17.79	2.48	2.57
Flujo lineal por partes más simple	3.22	3.30	79.25	78.25	17.46	17.94	1.55	1.44
Seno	9.18	8.62	80.18	74.68	33.98	43.92	11.31	14.00

Figura: Resultados del pronostico.

Resultados del pronóstico

Predicciones.



(a) 0_0_0 (b) 5_0_0

(a) 0_5_0 (b) 0_0_5

Figura: Predicciones.

Resultados del regresor

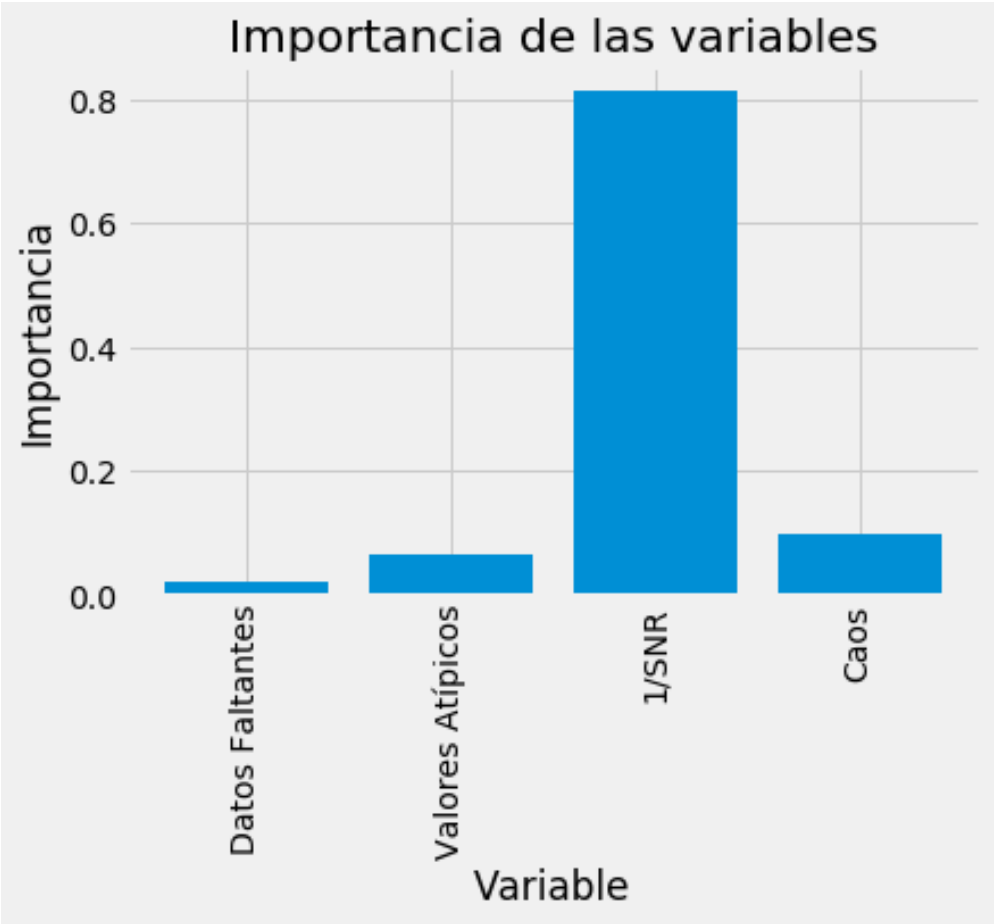
Tabla de datos para entrenar el regresor.

Serie de tiempo	Extracción de características				Error de pronóstico
No perturbadas	Datos faltantes	Valores atípicos	1/SNR	Caos	SMAPE pronóstico
Sistema Chen	0	0	0.013	0.001	1.12
Oscilador Duffing	0	0	0	0.0004	4.71
Atractor cíclico simétrico de Halvorsen	0	0	0	0.0001	3.17
Atractor de Lorenz	0	0	0.019	0.0085	7.05
Atractor de Rössler	0	0	0.012	0.0006	1.00
Atractor de Rucklidge	0	2	0	0.0009	2.04
Oscilador Shawn-van der Pol	0	0	0	0.0004	2.95
Flujo cúbico más simple	0	0	0	0.0002	2.70
Flujo lineal por partes más simple	0	0	0	0.00003	3.30
Seno	0	0	0	0.0721	8.62

Resultados del regresor

El modelo del regresor de bosque aleatorio con 1,000 árboles de decisión presento los siguientes resultados.
PrecisionR: 85.134 %.
Para el análisis de la importancia de las variables reportó los siguientes resultados.
Variable: 1/SNR Importancia: 0.82
Variable: Caos Importancia: 0.1
Variable: Valores Atípicos Importancia: 0.05
Variable: Datos Faltantes Importancia: 0.02
Resultados del nuevo regresor con las 2 variables más importantes. Error absoluto medio: 7.36
PrecisionR: 82.93 %.

Resultados del regresor



Resultados calidad de los datos

Calidad de los datos de una serie de tiempo.

Serie de tiempo	Características				Error SMAPE	Calidad
	Datos faltantes	Valores atípicos	1/SNR	Caos		datos
st 1	0	0	0.013	0.0010	1.77	98.23%
st 2	0	7	0.070	0.0007	61.84	38.16%
st 3	0	0	0.039	0.0007	16.23	83.77%
st 4	25	4	0.189	0.0005	44.95	55.05%

Conclusiones

Las características que se extrajeron son cuatro: datos faltantes, valores atípicos, ruido y caos. La característica que se extrae de manera sencilla y sin error es la de datos faltantes. El ruido es otra característica que se extrae de forma correcta ya que los errores mostrados son pequeños y entre más pequeño sea el error mejor es la extracción. Los valores atípicos son muy difíciles de extraer porque basta que un conjunto de datos atípicos se encuentren cerca para no ser detectados.

La extracción de estas cuatro características para una serie de tiempo de tamaño 20,000 toma aproximadamente 110 segundos. por lo cual para extracción de las características de las 2,160 series de tiempo tomó aproximadamente 66 horas.

Se encontró que en todas las series que presentan ruido los modelos producen un mal pronostico, dependiendo del nivel de ruido que presente. También se encontró que la presencia de los valores atípicos afecta al pronóstico en menor manera que el ruido, y por último el que existan valores faltantes afecta en menor manera. La combinación de las perturbaciones afecta en diferente medida el pronóstico, porque depende de la misma naturaleza de la serie de tiempo.

Conclusiones

La regresión se llevó a cabo usando un método de aprendizaje máquina llamado regresor de bosques aleatorios. El modelo de este regresor contiene 1,000 árboles de decisión que conforman el bosque aleatorio. Extrayendo las dos características más importantes se puede tener una buena aproximación al error.

Para conocer la calidad de los datos de una serie de tiempo, con este trabajo se requiere de un tiempo de 110s aproximadamente para la extracción de características (para una serie de tiempo con 20,000 datos). Adicionalmente se tiene el tiempo de 0.5s aproximadamente para la regresión del error, en total se necesita 1 minuto y 50.5 segundos para obtener una estimación del error que un modelo neuronal produciría, y posteriormente obtener la precisión del pronóstico y decidir si la calidad de los datos ha sido buena o mala.

Trabajos futuros

Utilizar series de tiempo reales de diferentes áreas; la energía eléctrica en esta área se encuentran diferentes tipos de series de tiempo (viento, solar, generación, distribución, crecimiento de red, etc.). En economía y marketing (empleo, desempleo, precios de productos, ventas, etc.). Teniendo estas diferentes series de tiempo, amplía el espacio para calcular la calidad de alguna serie de tiempo.

Agregar la normalización de los datos es otra mejora que puede aportar más precisión en el error de pronóstico.

Ampliar la combinación de las perturbaciones, incrementando los porcentajes y niveles, se plantea tener niveles comenzando en 1 hasta 40 con incrementos de 1. estos generarían 40x40x40 = 640,000 diferentes series de tiempo.

Agregar nuevas características es otra mejora que se puede realizar, agregar cuatro características; el valor máximo, el valor mínimo, la media y la desviación estándar. Para tener un total de 8 características, se espera que proporcionen más información y ajustar mejor el modelo del regresor, hacer un análisis de importancia de las variables y ver como impacta en el error del regresor.

Trabajos futuros

Incluir diferentes métodos de pronóstico de series de tiempo es otra mejora que se puede realizar, incluir los métodos; redes neuronales recurrentes LSTM, vecinos cercanos y ARIMA. Se espera que agregando estos métodos se mejore la predicción para cualquier tipo de serie de tiempo.

Agregar algoritmos evolutivos para encontrar los mejores parámetros para el diseño de los modelos de perceptrones multicapa, con esto se garantiza tener el mejor modelo para cada serie de tiempo.

Hacer un preprocesamiento de los datos, para todas las características y comparar el pronóstico de la serie de tiempo sin preprocesamiento y con preprocesamiento. Se espera que con este preprocesamiento la calidad de la serie de tiempo mejore y por lo tanto la precisión del pronóstico.

Agradecimientos

¿Existe alguna pregunta?

¡Gracias por su atención!

Ing. Víctor Manuel Téllez Velázquez
mtellez@dep.fie.umich.mx

DIVISIÓN POSGRADO

Bibliografía