Calibración de un sensor de bajo costo usando Random Forest

2200804- María Fernanda Carvajal Guerrero
2200799- Carlos Santiago Rodríguez Sarmiento
Universidad Industrial de Santander
Cl. 9 Cra 27, Bucaramanga, Santander

7 de diciembre de 2021

Índice

Resumen

Los sensores de bajo costo son dispositivos que han tomado reelevancia en la última década, pues permiten monitorear diferentes fenómenos como la contaminación, a precios asequibles. Es por ello que en el presente trabajo se calibraron los sensores para lograr mejores lecturas de datos, utilizando métodos matemáticos como lo son las nociones de métrica y las aproximaciones de funciones; así como métodos de Machine Learning como Random Forest, para realizar un ajuste más preciso. Es así como se logra la calibración de sensores de bajo costo con una disminución de error del 78.093 % y precisión del 94.171 %.

1. Introducción

Estamos viviendo una época de desarrollo explosivo de sensores que pueblan y generan datos en todas las facetas de nuestra cotidianidad. Estos sensores de bajo costo forman parte de dispositivos de la llamada revolución de la Internet de las cosas, IoT. Muchas veces estos sensores no son lo suficientemente precisos y deben ser calibrados con un patron de referencia . Este ejercicio busca mostrar que esa calibración está íntimamente ligada a la idea de métrica. El problema está en cuantificar cuál es el error de medición del sensor de bajo costo y, como calibrarlo para que podamos

establecer nuevas lecturas que sean mas precisas.

Un modelo *Random Forest*[1] se compone por árboles de decisión, que son la caracterización de las variables de un conjunto de datos, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generada mediante *bootstrapping*. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo [2].

Los métodos basados en árboles de decisión se han convertido en uno de los referentes dentro del ámbito predictivo debido a los buenos resultados que generan en problemas muy diversos como lo es la calibración de los sensores.

A continuación, en la metodología2 se mencionará el uso de la herramienta computacional *Python* para la resolución del problema a tratar mediante modelos predictivos[3] como *Random Forest*, posteriormente en el experimento y los resultados3 se mostrará el resultado de la calibración de los sensores y un análisis respectivo sobre la solución al problema. Es así como el informe se deasarrollará con base a los siguientes objetivos:

1. Cuantificar el error de medición de un sensor de bajo costo.
2. Encontrar la calibración adecuada del sensor para mejorar las lecturas de datos.

2. Metodología

Para el procesamiento de datos y la calibración de los mismos se usó *Python* como herramienta computacional debido a su alta funcionalidad y su libreria *Pandas* que contiene un gran repertorio de utilidades y una elevada eficiencia en el manejo de enormes cantidades de datos, como lo fue la recolección de las mediciones brindadas por los sensores sobre la concentración de material particulado a lo largo de 10 a 15 meses en intervalos de minutos y horas en diversos puntos de la ciudad de Bucaramanga, lo que supone una inmensa cantidad de datos a procesar.

En primera instancia era necesario limpiar los datos brindados por los sensores de bajo costo, es decir, eliminar el 'ruido' que se puede encontrar en cada dataset presentado. Esto se ejecutó mediante un promedio móvil lo que permitía realizar el suavizado de la información (data smoothing) y así poder revisar patrones. De los datos de referencia únicamente se mantuvieron aquellos que se usarián para la comparación. Esta se hizo entre ambos conjuntos de datos y sirvió para conocer el intervalo mínimo en el que coincidíeron para realizar la calibración de una forma más efectiva.

En segunda instancia y luego de tener datos más trabajables, se encontró el valor más acertado de la ventana móvil mediante pruebas de ejecución entre valores arbitrarios que oscilaban de 0 a 300 con una diferencia de 50 elementons, teniendo en cuenta que debía ser un valor que optimizara la distancia euclidea, optimizara el uso de máquina y mostrara una buena resolución de los datos.

Finalmente, mediante el modelo predictivo *Random Forest* y sus árboles de decisión que crea subconjuntos de datos con base en la variable expuesta, en este caso material particulado, se entrenó una muestra aleatoria de datos por cada árbol de decisión para poder predecir el comportamiento de los mismos y así verificar con el valor referencia si se calibró satisfactoriamente, comprobándose mediante gráficas que evidencian su compatibilidad y su relación lineal.

3. El experimento y los resultados

Tal y como se presentó en la metodología, una vez realizada la limpieza de los datos y una reorganización por índice tomando en cuenta los datos comunes en los dataset de referencia y a calibrar, se realiza una primera visualización de los datos 1:

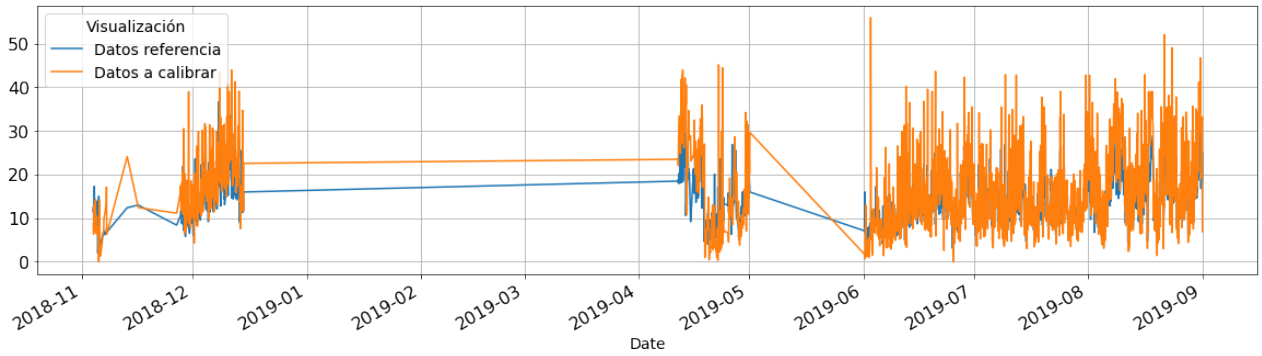

Figura 1: Primera visualización de datasets

Esta visualización presenta los datos promedidados por hora debido a que los fenómenos de contaminación del aire por material particulado presenta diferentes picos y valles de acuerdo al momento del día en que se realice la medición. Se debe tomar en cuenta que los promedios diarios, aunque indiquen una tendencia a largo plazo, no permite evidenciar situaciones propias de una franja horaria que incrementen considerablemente la emisión de las partículas.

Se puede realizar a su vez una idea de métrica, como distancia euclideana entre datasets, para evaluar la cercanía de un conjunto de datos con otro. Bajo este criterio, la definición de métrica está dada por:

$$\mathcal{D}\left(\mathbb{D}_i, \hat{\mathbb{D}}_i\right) = \sqrt{\sum_{i,\hat{i}} \left(\mathbb{D}_i - \hat{\mathbb{D}}_{\hat{i}}\right)^2} \tag{1}$$

La distancia euclideana obtenida entre los conjuntos es de **404.93** unidades. Esta distancia brinda una noción de error, pues mientra más separados se encuentren los conjuntos más inexacta

es la medición de los sensores de bajo costo respecto a la referencia brindada por la AMB.

3.1. Promedio móvil y ventana euclídea

Se realizaron ensayos para diferentes valores de la ventana móvil, con un salto de 50 elementos entre las mismas y se evaluaron los mejores valores de la ventana tomando en cuenta tres criterios:

- Resolución de los datos
- Optimización de la distancia euclídea
- Optimización de uso de máquina

Así, se realizan las pruebas y se grafican los diferentes valores tomados para la ventana contra la distancia euclidea entre datasets y tiempo de ejecución de la ventana 2.

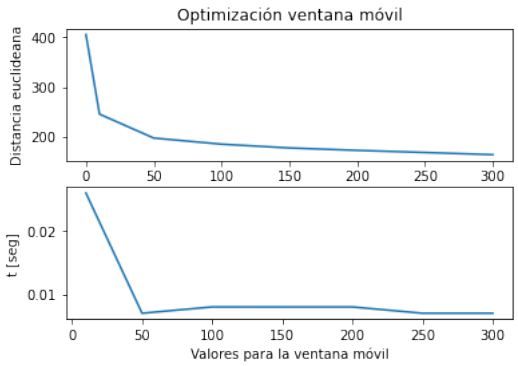

Figura 2: Comparación entre ventanas móviles

Es así como se toma el valor de 100 como el más óptimo bajo los tres criterios mencionados anteriormente pues es el mínimo valor de datos promedidado que presenta la mejor distancia euclidea entre vectores y un tiempo de ejecución razonable que ronda los **0.008 segundos** con una disminución de la distancia euclideana inicial del **45.86 %**, al ser esta de **185.68** unidades.

La gráfica comparativa entre datos para una ventana móvil de 100 elementos se presenta a continuación 3:

3.2. Calibración usando Random Forest

Para esta sección se utilizó el método de calibración usando la metodología *Random Forest* que para este caso tiene una única variable que generará 1000 árboles de decisiones y como criterio

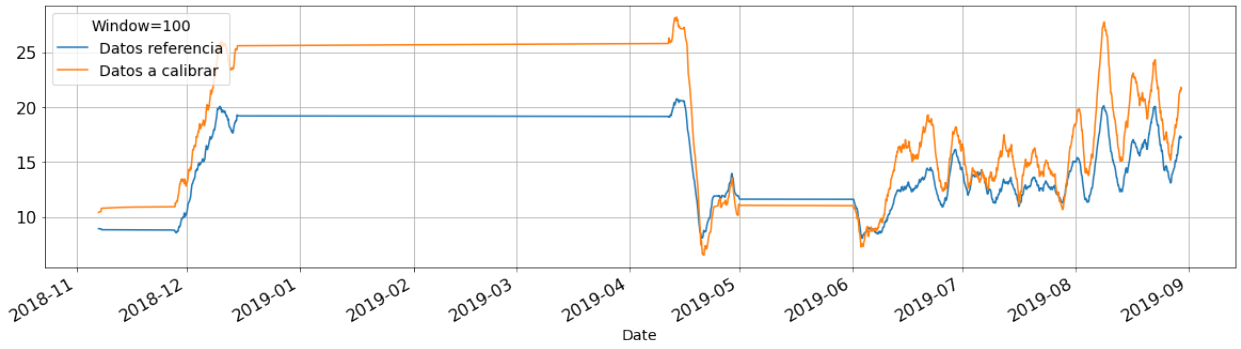

Figura 3: Datos con ventana móvil de 100 elementos

principal el error cuadrático medio. Esta metodología permite que, utilizando el 50 % de los datos como conjunto de entrenamiento, se genere un aprendizaje de los parámetros y se evalúe la predicción del modelo propuesto.

Una vez aplicada la calibración se obtiene el siguiente ajuste a la recta 4:

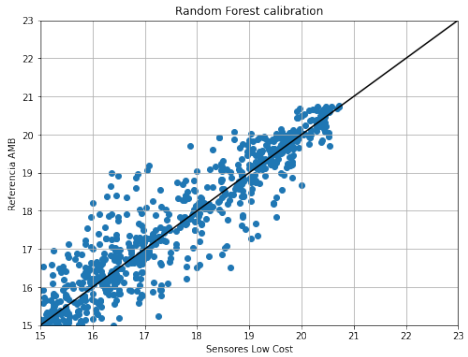

Figura 4: Calibración mediante Random Forest

Y utilizando el conjunto de datos de evaluación, el 50 % restante de los datos brindados, se puede realizar una evaluación del error medio absoluto, precisión y el error cuadrático medio como se muestra en la tabla a continuación:

	Valor
Error medio absoluto	0.744
Precisión	94.171 %
Error cuadrático medio	0.993

Los datos anteriores muestran que Random Forest es una técnica con gran capacidad predictiva

y de aprendizaje a partir de un conjunto de datos iniciales. No obstante, si se tuvieran los datos correspondientes a variables diferentes a $PM_{2,5}$ como temperatura, humedad y presión, que son influyentes en el fenómeno de contaminación, se podría realizar árboles de decisiones que integren cada variable y tomen decisiones no solo a partir de los datos a predecir.

Así, se puede realizar la comparación entre datos originales, calibrados y de referencia a partir de la gráfica que se muestra a continuación:

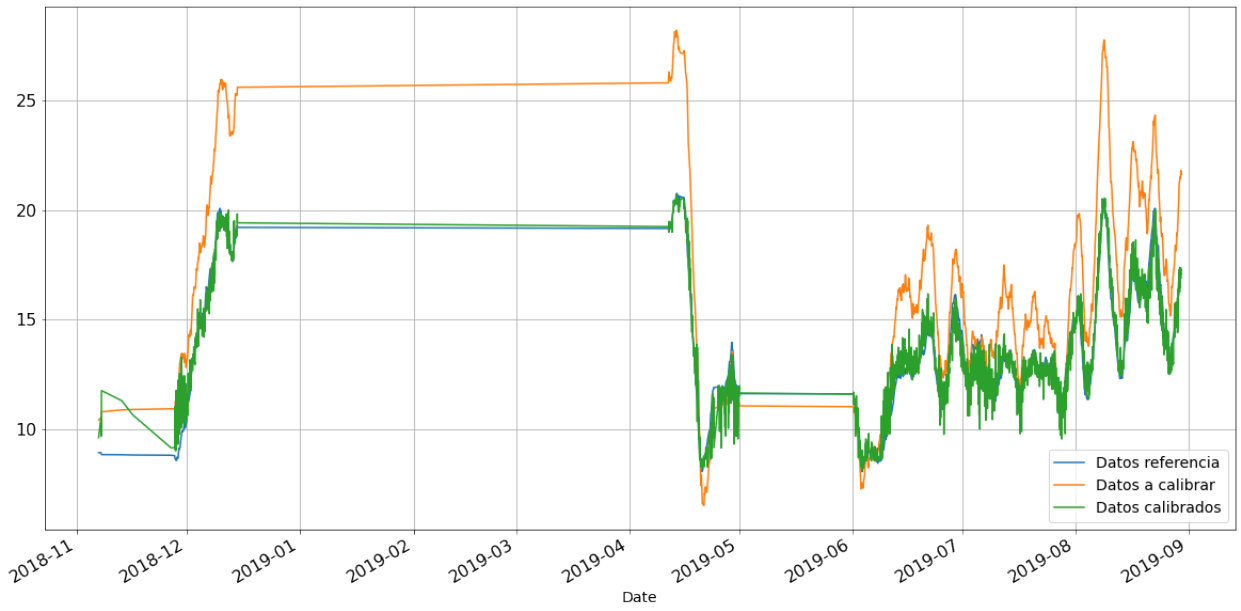

Figura 5: Comparación con datos calibrados

Finalmente se puede establecer el error entre esta calibración con los datos originales, entendido como la distancia euclideana entre los mismos, tal y como se presenta en la siguiente tabla:

Distancia Euclideana entre	*Valor*
AMB y calibrados	40.679
Referencia e iniciales	185.687
Porcentaje de reducción	78.093 %

4. Conclusiones

A partir del análisis y procesamiento de los datos tomados por los sensores de bajo costo y la estación de la AMB, se puede concluir que la calibración de estos datos es útil para poder predecir o

detectar comportamientos ajenos en el entorno y de esa forma poder actuar más rápido para evitar situaciones ambientales de mayor riesgo.

Por otro lado, los datos mostrados evidencian que Random Forest es una técnica con gran capacidad predictiva y de aprendizaje a partir de un conjunto de datos iniciales. Sin embargo, si se tuviesen datos sobre más variables que causan un efecto en el fenómeno de la contaminación como la presión, temperatura y humedad, se realizarían una cantidad mayor de árboles de decisión que tomen las variables y generen decisiones a mayor escala.

En terminos generales, la fisica computacional y la ciencia de datos evidencian un papel importante en la solución de problemas que conciernen a la población, su uso va a ser representativo a nivel cotidiano gracias a la revolución tecnológica que se presenta con el internet de las cosas (Iot). El aprendizaje de estas significará un avance en el abordaje de situaciones diarias en la ciencia.

Referencias

[1] Rodrigo J. https://www.cienciadedatos.net/documentos/py08_random_forest_python.html. Accessed: 2021-12-7.

[2] Naomi Zimmerman, Albert A Presto, Sriniwasa PN Kumar, Jason Gu, Aliaksei Hauryliuk, Ellis S Robinson, Allen L Robinson, and Ramachandran Subramanian. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1):291–313, 2018.

[3] Sierra D. Sensor calibration low cos. https://github.com/sierraporta/sensor-calibration-low-cost.