Introducción a la Ciencia de Datos

Tarea Final

2024

Leandro Cantera, Miguel Paolino

1. Antecedentes

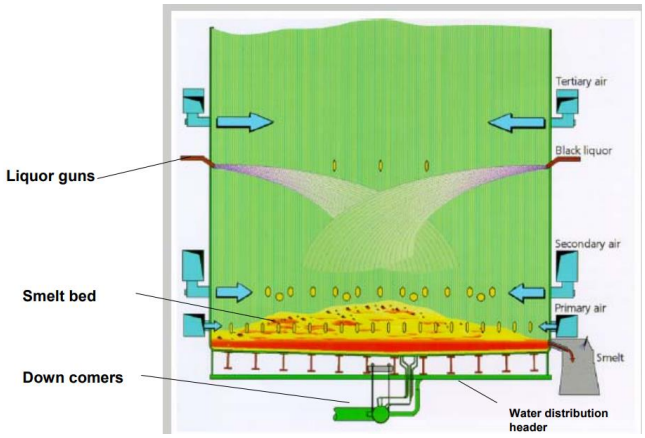
En la producción de celulosa mediante el proceso Kraft, la madera se divide en sus dos componentes principales, la *Celulosa y Lignina*. La Lignina es el ligante que aglutina las fibras de Celulosa, y el proceso Kraft se basa en disolver la lignina en distintas etapas e ir purificando la celulosa por medio de etapas de *delignificación* y lavado.

En el proceso donde se disuelve aproximadamente el 85% de la lignina es en el Digestor Continuo, los chips de madera se mezclan con una solución de alcalina de Soda Caustica (NaOH) y Sulfuro de Sodio (Na_2S), temperatura (alrededor de 160 °C) y presión (4 a 5 barg) por unas 4 horas de tiempo de residencia.

Este proceso en una planta moderna de celulosa necesita unas 1300 ton de soda cáustica y unas 700 ton de sulfuro de sodio por día, para fabricar 1 millón de toneladas de celulosa Kraft por año.

Este proceso no sería sostenible (económica ni ambientalmente) si no se recuperan estos químicos, para esto se desarrolló el ***Ciclo de Recuperación***, donde se recuperan ambos reactivos (NaOH y Na_2S) y se quema la lignina para la generación de vapor con el cual se genera electricidad y vapor como fuente de calor para los distintos sub-procesos.

En la caldera de recuperación se obtiene sulfuro de sodio, sulfato de sodio y carbonato de sodio como una lava que denominamos fundido, el cual se recoge en un tanque disuelto en agua.



En el proceso de recuperación, hay parámetros de desempeño que no se miden continuamente, en esta tarea se evaluará la posibilidad de desarrollar sensores virtuales que que permitan arrojar datos en forma continua y así poder tomar mejores decisiones sobre los parámetros de operación.

En este trabajo se selecciona la medida de reducción de sulfato de sodio (Na_2SO_4) a sulfuro de sodio (Na_2S) en la Caldera de Recuperación. El parámetro que mide la eficiencia de la recuperación de sulfuro es la Tasa de Reducción y se mide como:

$$Tasa\ de\ Reducción = \frac{Na_2S}{(Na_2S + Na_2SO4)} \times 100$$

Las siguientes reacciones de reducción de azufre (S) se dan en el hogar inferior de una caldera de recuperación:

$Na_2SO_4 + 2C \rightarrow Na_2S + 2CO_2$

$Na_2SO_4 + 4C \rightarrow Na_2S + 4CO$

$Na_2SO_4 + 4CO \rightarrow Na_2S + 4CO_2$

La velocidad de reacción depende, además, de la temperatura del hogar, un incremento en 50 K duplica la velocidad de reacción, del contenido de carbono en la camada y CO en el hogar inferior.

2. Objetivo

El objetivo de este proyecto es la predicción de la tasa de reducción a partir de parámetros de operación de la caldera de recuperación. La función objetivo que se quiere maximizar es la tasa de reducción tiene un valor objetivo mayor a 92,5%

Los parámetros habituales de control son:

- Carga de licor negro en Toneladas de sólido por día, la carga es determinada por el balance de la planta (niveles de los tanques licor) a la que se encuentre.
- Caudal de aire de combustión dosificado en 3 niveles que viene dado por la capacidad de ventiladores (primario, secundario, terciario e inducidos), para actuar sobre parámetros: caudal y presión de los ventiladores de tiro forzado (Aire primario, Aire secundario, Aire terciario) y depresión en el hogar controlado por los tiros inducidos.
- Contenido de sólidos: mayor contenido de agua genera más velocidad en el hogar y consecuentemente más arrastre y menor tiempo de residencia para la reacción de reducción.
- Temperatura del licor.
- Temperatura del hogar.

Estos son los parámetros sobre los que deberíamos trabajar para determinar su influencia en la tasa de reducción y elaborar un sensor virtual que de una medida continua de éste parámetro.

3. Datos

1. Características del conjunto de datos

En este caso, hablamos de series temporales tomadas cada 1 segundo y en forma continua. Este tipo de datos, típicos de las industrias de proceso continuo, tienen las siguientes características:

Cantidad de datos: Solamente en la caldera de recuperación, se generan datos continuamente a partir de miles de instrumentos distribuidos en toda la caldera.

Tipo de datos: Los datos con los que se cuenta son series temporales. En la mayoría de los casos (aunque no en todos), los datos recogidos son numéricos, fundamentalmente medidas de temperatura, presión y caudal (másico y volumétrico), corriente eléctrica y porcentaje de apertura de válvulas o dampers. Además, se pueden recoger datos booleanos, que otorgan información del estado de un equipo (encendido o apagado). Por otro lado, la mayoría de los datos son recabados en forma continua (segundo a segundo). Pero hay información que se recaba de forma discreta, como la reducción de azufre. Estos datos dependen de análisis químicos para los que aún no se ha logrado el desarrollo de tecnologías que los miden en línea.

Ruido: como norma general, estos datos presentan ruido, fundamentalmente por factores como interferencias ambientales, errores de medición, fallos en los sensores o en los transmisores. El modelado de los datos debe contemplar el ruido y corregir.

Dependencias a corto y largo plazo: Es común que algunas variables tengan dependencia a corto plazo con otras, particularmente aquellas que están relacionadas a través de un lazo de control (ej: corriente de ventiladores de tiro inducido y depresión dentro de la caldera). Por otro lado, hay variables que tienen dependencias con otras en un período de tiempo mayor. Este es el caso de la reducción y el porcentaje de oxígeno residual en la caldera, ya que, si baja el oxígeno residual, se verá un aumento en la reducción varias horas después. En el caso de la caldera de recuperación, no es muy frecuente que existan dependencias a plazos mayores a unas 6-8 horas

Relación no lineal: Lo más frecuente es que las relaciones entre variables no sean lineales, aunque existen casos en los que es posible realizar una aproximación lineal con un error aceptable.

Interrelación entre múltiples variables: Al ser un sistema complejo y muy interrelacionado, existe un alto grado de correlación entre las variables. En este caso particular, es de esperar una alta correlación entre las variables de estudio, ya que, a nivel de proceso, las magnitudes medidas están íntimamente relacionadas.

Necesidad de análisis en tiempo real: Si bien se analizan los datos históricos para estudiar eventos particulares, el principal interés está en los datos en tiempo real, su análisis y la toma de decisiones que se puede desprender de ellos. Esto cobra importancia porque todas las dependencias futuras que expliquen un evento de un momento puntual no podrán considerarse en un modelo que trabaje en tiempo real.

2. Pre-procemiento: Evaluación y limpieza de datos

El pre-procesamiento de los datos es una etapa crucial en el flujo de trabajo, donde es necesario conocer los posibles problemas de calidad que tienen estos datos particulares y así prepararlos para que sean aptos para aplicar los modelos de aprendizaje automático.

- **Limpieza de datos:**

Valores atípicos: En caso de no identificarlos y eliminarlos, los valores atípicos pueden distorsionar en forma importante los resultados. Un caso particular de la situación de estudio son los refractómetros para medir contenido de sólidos en el licor de quema. Es común que estos sensores se ensucien y comiencen a marcar valores anormalmente altos. De no detectar y corregir esto en el análisis de datos, es posible que el cálculo de carga de la caldera en toneladas de sólidos quemados por día esté erróneamente sobreestimado.
Valores congelados: Otro típico problema de calidad de estos datos son los valores congelados, usualmente por falla en los sensores o trasmisores. En este tipo de falla, en vez de dejar de registrar valores, se sigue registrando el mismo valor a pesar de que la variable real sí esté cambiando según el proceso. En estos casos, la ausencia de ruido es un indicador fehaciente de que la señal no es representativa de la realidad.
Datos faltantes (NaN): Es común tener datos faltantes y las razones pueden ser diversas. Por un lado, el instrumento puede dejar de enviar la señal por una falla (tanto en el sensor como en el transmisor). Por otro lado, los valores de proceso pueden exceder el rango máximo o mínimo del instrumento o de la configuración de la trasmisión de la señal, lo que arroja un valor no numérico que es interpretado como NaN y que es sinónimo de dato faltante.
En cualquiera de estos casos, debido a la gran cantidad de datos, normalmente es posible realizar una limpieza eliminando todos los valores que no son representativos de la realidad. Sin

embargo, es posible que en algunos casos valga la pena utilizar técnicas de interpolación lineal, regresión lineal, u otras más complejas como la de los k vecinos más cercanos para generar artificialmente la información faltante

- **Normalización de datos:**

Escala: Es importante escalar las variables numéricas para que no tengan una influencia desproporcionada en los modelos utilizados. Lo normal sería realizar una normalización estándar.

Transformación de variables: En algunos casos, puede ser necesario transformar las variables para mejorar la linealidad de las relaciones entre ellas. Por ejemplo, para medir el ensuciamiento de las superficies calefactoras y aumento de la resistencia al pasaje de los gases se mide la pérdida de carga normalizada por la producción de vapor contra la producción nominal

$$\Delta pcorr = \Delta pmed \times \left(\frac{Qmed}{Qnom}\right)^2$$

- **Visualización de datos:**

Es importante visualizar las series temporales antes y después del preprocesamiento para identificar patrones, tendencias y posibles problemas con los datos. En estos casos suele ser muy útil realizar matrices de correlación, así como gráficos sobre el mismo conjunto de ejes, con el tiempo en el eje de las abscisas.

4. Modelado

Selección del Modelo

Según las características de los datos (series temporales) se pueden elegir modelos de aprendizaje automático adaptados a estos. Pueden ser modelos de regresión (regresión lineal, polinómica, SVM), modelos específicos de series temporales (ARIMA, SARIMA, etc), de redes neuronales (RNN) o de Random Forest.

Modelos vistos en el curso

Regresión Lineal Múltiple: La ventaja es que es un modelo sencillo de aplicar, la desventaja es que puede no ser robusto ante estos datos, que probablemnte no presenten correlaciones lineales.

Support Vector Machines: Efectivas para clasificar y predecir datos no lineales. Su capacidad para manejar datos con ruido las hace adecuadas para series temporales con alta variabilidad.

Random Forest: son robustos a los valores atípicos y pueden capturar relaciones complejas entre las variables. Su simplicidad y facilidad de interpretación los convierten en una opción atractiva para la industria de procesos continuos

Otros modelos posibles

Redes Neuronales Recurrentes (RNN): Las RNN, como LSTM y GRU, parecerían ser particularmente adecuadas para modelar series temporales con ruido y dependencias a largo plazo. Tienen capacidad de aprender patrones complejos en los datos, lo que las hace útiles para predecir valores futuros o detectar anomalías. Las RNN parecen ser un buen modelo a probar, por su capacidad de modelar dependencias a largo plazo, ser robustas ante datos ruidosos y capacidad de adaptarse a cambios en los procesos mediante el reentrenamiento.

Por estas razones, en este trabajo parecería ser útil apicar RNN para el modelado del sensor vitual que permita calcular/predecir el valor de la tasa de reducción en función de los demás parámetros de operación de la caldera de recuperación.

3. División del Conjunto de datos

Para la división del conjunto de datos en entrenamiento, validación y testeo, el principal desafío es no omitir las dependencias temporales de los datos, por lo que una división en bloques de tamaño fijo puede llegar a arrojar resultados incorrectos. Por otro lado, una división aleatoria podría interrumpir las dependencias temporales y también otorgar resultados incorrectos.

Para este caso particular, lo más adecuado sería elegir una ventana de tiempo teniendo en cuenta las dependencias temporales de los datos (entre 4 y 8 horas sería razonable). Luego, dividir los datos en aproximadamente 70% para entrenamiento/validación y el resto para testeo. Tomar solamente el último 30% para testeo nos permite validar si el modelo es capaz de predecir los datos futuros en función de los datos conocidos hasta el momento.

Gracias a la gran cantidad de datos del proceso, esto se podría repetir varias veces con valores históricos, lo que otorga una gran capacidad para entrenar, validar y testear los modelos.

4. Selección de hiper-parámetros, entrenamiento y validación

En una RNN, se debe seleccionar los hiper-parámetros que mejor se ajusten a los datos. En este caso, se debería elegir el número de capas recurrentes, cantidad de neuronas de cada capa, la función de transformación que aplica cada neurona a sus entradas y la tasa de aprendizaje.

Lo ideal sería entrenar más de un modelo de RNN (ej: LSTM y GRU), para cada uno de ellos buscar los hiper-parámetros más adecuados y comprar los resultados usando métricas como el R^2 o el error absoluto medio.

5. Implementación

Una vez definido el modelo, habiéndolo validado y testeado adecuadamente, se podría implementar. En estos casos es importante implementar actualizaciones del modelo, mediante re-aprendizaje con nuevos datos, ya que es muy común que en algunos momentos, un set de parámetros genere tasas de reducción muy buenas, pero en otros momentos no sea así. Un claro ejemplo es la carga de la caldera: A cargas altas se comporta diferente que a cargas bajas, aunque los parámetros específicos (por tonelada de licor quemado) sean los mismos.

6. Conclusión

La cantidad de datos generados en este tipo de industrias (Planta de celulosa de proceso continuo) permite la implementación de modelos de aprendizaje automático para el desarrollo de sensores virtuales a partir de variables de proceso que son continuamente medidas y monitoreadas.

Si bien se pueden utilizar varios métodos de aprendizaje automático para el trabajo con series temporales, Parecería ser adecuado incursionar en las Redes Neuronales Recurrentes (RNN) por su robustez ante el ruido, capacidad de contemplar dependencias a largo plazo, así como aprender patrones complejos que permiten predecir valores futuros.

Es fundamental el conocimiento profundo del proceso, así como de las características de los datos, de manera de poder maximizar el aprovechamiento de estos, elegir adecuadamente la división en entrenamiento, validación y testeo, y finalmente ajustar los hiper-parámetros del modelo elegido para obtener los mejores resultados posibles.