APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS AL MANTENIMIENTO BASADO EN CONDICIÓN

Sebatián Martorell [(1)], Isabel Martón [(1)], Ana Isabel Sánchez [(2)], Sofía Carlos[(1)]
[(1)] Departamento de Ingeniería Química y Nuclear, Universidad Politécnica de Valencia
[(2)]Departamento de Estadística, Investigación Operativa Aplicadas y Calidad, Universidad Politécnica de Valencia

RESUMEN

Actualmente, existe un creciente interés en la implementación de actividades de mantenimiento basado en la condición (CBM) debido a que las actividades de mantenimiento preventivo por tiempo, tradicionalmente utilizadas, pueden resultar ineficientes desde el punto de vista de los costes o la disponibilidad en diferentes situaciones. CBM es una estrategia de mantenimiento basada en el análisis de la información obtenida de la monitorización de las condiciones operacionales de los equipos, tales como vibraciones, temperatura, presión, etc., las cuales permiten analizar el estado de degradación de un equipo y determinar el tipo de mantenimiento que se necesita y cuando realizarlo
En este contexto, en la ponencia se presenta una aproximación que permite la detección de comportamientos anómalos en los equipos industriales. Dicha aproximación combina dos técnicas de análisis multivariante: Análisis de componentes principales (PCA) y regresión en mínimos cuadrados parciales (PLS) con el objetivo de utilizar las variables monitorizadas del equipo para obtener información acerca de la evolución de la degradación del mismo a lo largo del tiempo

1. INTRODUCCIÓN

En las últimas décadas, se ha observado un cambio en la planificación del mantenimiento de las plantas industriales debido al desarrollo en las tecnologías de información y comunicación. Esta planificación del mantenimiento afecta directamente a la disponibilidad y viabilidad económica de la planta. El coste total de mantenimiento, que incluye los costes humanos y materiales, tiene que ser reducido para ser más competitivos y garantizar la producción. Por esta razón, en las últimas décadas, se ha incrementado la aplicación de técnicas de monitoreo de la condición con el fin de predecir los patrones de comportamiento o fallos de las instalaciones industriales. (Kusiak y Verma, 2011).
Además, una detección temprana del comportamiento anormal de los equipos de ayuda a los equipos de mantenimiento a observar la degradación de los equipos.
Con el objetivo de seleccionar el plan de mantenimiento más apropiado en cada situación se utilizan técnicas de mantenimiento basado en condición (CBM). Estas técnicas se pueden clasificar en técnicas de diagnóstico y pronóstico (Jardine et al., 2006). Las primeras se basan en la detección, aislamiento e identificación de fallos y comportamiento anormal. En cambio, las segundas se centran en la predicción de defectos o fallos antes de que estos ocurran. Esto

implica que al evitar fallos se incremente la disponibilidad de los equipos y con ello la producción.

Existen en la literatura muchos enfoques típicos, con métodos y aplicaciones probadas de detección de fallos. Estos enfoques generalmente se clasifican en dos categorías principales, los enfoques basados en modelos y enfoques basados en datos.

El enfoque basado en modelos se basa en el uso de unos modelos matemáticos (algebraicos o conjunto de ecuaciones diferenciales) para representar el comportamiento del sistema, incluyendo el fenómeno de la degradación. El modelo matemático consta de modelos estadísticos y físicos. Los modelos estadísticos se desarrollan a partir de los datos de entrada / salida recogidos, que no incluyen condiciones no registrados y los modelos físicos son útiles en la contabilidad de todas las situaciones de explotación (Hao et al., 2009) .

Por otro lado, los enfoques basados en datos reales utilizan los datos obtenidos a partir del sistema de adquisición de datos y características del sistema para diagnosticar el comportamiento global del sistema.

Los enfoques basados en datos se pueden dividir en dos categorías: (AI) técnicas de inteligencia artificial (redes neuronales, los sistemas difusos, árboles de decisión, etc), y las técnicas estadísticas (métodos estadísticos multivariantes, discriminadores lineales y cuadráticas, mínimos cuadrados parciales, etc)(Dragomir et al., 2009; Lam et al., 2010).

En este trabajo se propone una metodología que combina dos técnicas de análisis multivariante: análisis de componentes principales (PCA) y mínimos cuadrados parciales (PLS), con el objetivo de predecir el estado del equipo. Además, se presenta un caso de aplicación en el cual se aplica dicha metodología con el objetivo de detectar el comportamiento anormal de un generador asíncrono.

2. Métodos basados en datos. PCA y PLS

El análisis de componentes principales (PCA) es una de las técnicas estadísticas multivariantes más populares. El PCA es una herramienta robusta que se puede utilizar en la detección de fallos y tareas de aislamiento. Esta técnica se basa en una transformación lineal que produce nuevas variables a partir de las variables originales medidas. Estas nuevas variables, llamadas componentes principales son variables no correlacionadas entre sí, es decir son ortogonales. Cada componente representa una dirección del espacio de las variables originales. Esta transformación implica una reducción de la dimensión de los datos originales, de manera que algunos de estos componentes son suficientes para representar adecuadamente las fuentes de variabilidad del proceso (Wold et al., 1987). Normalmente, la primera componente principal es la que más varianza tiene, es decir es la que mayor cantidad de información relativa al proceso lleva incorporada.

Matemáticamente, el PCA se basa en una descomposición ortogonal de la matriz de covarianza de las variables del proceso a lo largo de las direcciones principales las cuales explican la máxima variación de la matriz de datos, X. La matriz de datos, X, contiene la información de los datos obtenidos a partir de los datos históricos de equipos. Las N filas en la matriz X se conocen como objetos y las columnas K se denominan variables que se miden para cada objeto.

La proyección de la matriz de datos X en el subespacio A-dimensional a través de la matriz de proyección, P', da el objeto de coordenadas en el plano, T. Las columnas de T, t_a, se denominan "scores" y las filas de P', p_a, son los "loadings". El p_a y t_a son ortogonales. Los "scores" representan la estructura de las filas, o lo que es lo mismo, las relaciones entre las muestras y los "loadings" muestra la relación existente entre las variables. Las desviaciones entre las proyecciones y los valores de coordenadas originales son los residuos, E. En matriz de forma PCA se expresa mediante la siguiente expresión:

$$X = TP' + E \quad (1)$$

El PCA puede ser extendido cuando se desea incluir todos los datos disponibles en el procedimiento de seguimiento y utilizar los datos históricos, X, para predecir y para detectar cambios en las variables de salida Y. Esta extensión se conoce como mínimos cuadrados parciales (PLS). El modelo de PLS se utiliza para obtener la relación entre los datos de la matriz, X, y las variables, Y, reduciendo al mismo tiempo su dimensión.
EL modelo de PLS se construye a partir del algoritmo de NIPALS (Geladi y Kowalsky, 1986).

Como se muestra en la Ecuación 1, la matriz de datos, X, puede ser representado por la matriz de puntuación, P. El modelo de PLS se construye mediante la regresión entre los "scores" de las X e Y. La relación para la matriz X se obtiene mediante la Ecuación 1. La relación para la matriz Y se construye mediante la siguiente ecuación:

$$Y = UC'' + F \quad (2)$$

Donde U es la matriz de los "scores" de Y, C es la matriz de "loadings" que resume Y, y F es la matriz residual.
Además la matriz U, tiene una relación interna con la matriz de "scores" de X, T, que se puede representar como:

$$U = TB + R \quad (3)$$

Donde B es la matriz de coeficientes de regresión y R es la matriz residual. Por lo tanto el modelo de PLS se puede calcular utilizando la siguiente expresión:

$$Y = TC' + G \quad (4)$$

Siendo G la matriz residual total.

En muchas ocasiones los modelos PCA y PLS se utilizan conjuntamente con lo que se conoce con el nombre de Control estadístico multivariante del proceso (MSCP) (MacGregor and Kourti., 1995). El objetivo de esta técnica es manejar gran cantidad de información, ya sea reduciendo el número de variables a monitorear mediante el PCA o indicadores utilizando gran cantidad de información, lo cual se logra con la proyección a PLS. La metodología que utiliza el MSCP se estructura en varias etapas. La primera de ellas es obtener el modelo bajo control basado en los datos históricos durante la operación normal del equipo. Tras definir este modelo se introducen en este, las proyecciones de las nuevas observaciones definidas por los "loadings" y se obtienen los "scores" y los residuos de las nuevas observaciones. Finalmente, se construye el gráfico de control para todos los componentes principales. Los gráficos de control

multivariante utilizan la T2 de Hotelling para cada componente principal utilizando para ello la siguiente expresión:

$$T_A^2 = \sum_{i=1}^{A} \frac{t_i^2}{s_{t_i}^2} \tag{5}$$

donde $s_{t_i}^2$ es la varianza estimada de y A es componentes principales totales (PC).

El gráfico de Hotelling monitorea la distancia de una nueva medida al valor de referencia en el espacio de los factores PCA. Permite detectar si la variación incluida en los componentes principales considerados es más grande que la que correspondería si solo influyeran variaciones aleatorias. Las muestras fuera de control poseen un valor de Hotelling superior al límite y aparecen más allá de la línea de control. Mediante este índice se pueden detectar salidas del patrón del patrón normal de las variables.

El error de predicción cuadrado (Spey) de los residuos también se utiliza en el proceso de monitorización y seguimiento. Spey mide la distancia desde el punto hasta el plano de los componentes principales. Permite determinar si los datos proyectados sobre el modelo no están siendo representados por este.

$$SPE_y = \sum_{i=1}^{N} (y_{new,i} - \hat{y}_{new,i})^2 \tag{6}$$

Donde $\hat{y}_{new,i}$ se calcula a partir del modelo PLS y PCA y N es el número de observaciones.

A partir de la SPE, la distancia absoluta al modelo DModY se define como:

$$DModY = \sqrt{\frac{SPE_y}{K - A - 1}} \tag{7}$$

Siendo K el número de variables.

3. METODOLOGIA

El objetivo de esta ponencia se basa en el desarrollo de una metodología para detectar el comportamiento anormal de los equipos. La Figura 1 muestra el marco de la metodología propuesta.

La primera etapa de esta metodología consiste en la recogida de datos. Los datos utilizados en este trabajo para construir los diferentes modelos han sido recogidos por un sistema de control y adquisición de datos, conocido con el nombre de SCADA. A partir de los datos de los equipos recogidos, es posible obtener un modelo que represente el comportamiento normal. Estos modelos en este caso se construyen utilizando dos métodos multivariantes basados en datos como son el PCA y el PLS.

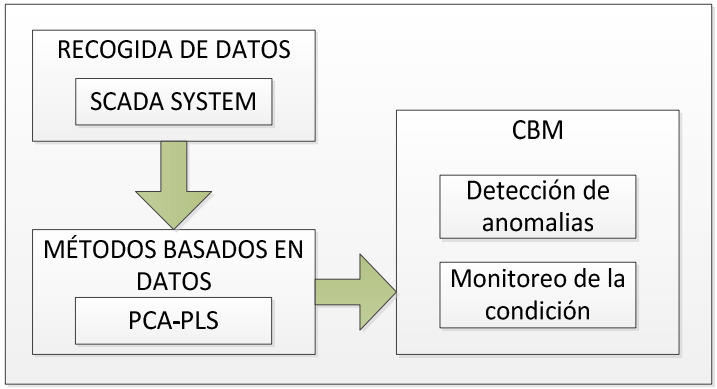

Figura 1. Metodología propuesta

En ocasiones los datos recogidos por el SCADA contiene ruido debido a los errores y malfuncionamiento de los diferentes sensores, por esta razón antes de construir el modelo es necesario realizar un filtrado de los datos y eliminar los "outliers" y puntos influyentes.

La capacidad predictiva de estos modelos se mide analizando los residuos. Para realizar estas medidas se utiliza el error medio absoluto porcentual (MAPE), el cual se puede calcular a partir de la siguiente expresión:

$$MAPE = \frac{\sum_{i=1}^{n} \left| (y_t - \hat{y}_t) / y_t \right|}{n} \cdot 100 \qquad (8)$$

Donde, y_t es el valor real, $\hat{y}$ es el valor predicho por el modelo y n el número de observaciones.

Por último, la evolución de los datos reales y predichos obtenidos a partir de los modelos desarrollados en el paso 2, se revisará para detectar el comportamiento anormal y para llevar a cabo la monitorización del estado de los equipos. La monitorización y seguimiento utiliza la salida predicha para generar una señal de alarma cuando el comportamiento del equipo es anormal.

4. CASO DE APLICACIÓN

El caso de aplicación se centra en la obtención de un modelo que caracterice el comportamiento normal de un generador asíncrono para detectar desviaciones o mal funcionamiento. Con el fin de obtener este modelo, las variables medidas cada hora del sistema son recogidas por los sensores y se guardan en archivos de datos históricos.

Los datos utilizados para predecir el estado anormal de un generador asíncrono han sido obtenidos de la información histórica disponible proporcionada por el sistema SCADA. Los datos históricos almacenados que permiten el control de comportamiento generador son los siguientes:

- Velocidad del rotor (RS)
- Temperatura de la multiplicadora (GT)
- Potencia del generador (GP)
- Diferencia entre la temperatura de los cojinetes del generador y la ambiental (DT1)

- Diferencia entre los anillos de temperatura del generador y la góndola (DT2)

Estos datos se han utilizado para establecer una relación entre las variables de funcionamiento medidas y el comportamiento normal del generador con el fin de predecir una desviación en este comportamiento cuando nuevos datos sean introducidos. Los datos utilizados corresponden al periodo histórico comprendido entre 2008 a 2010.

La Tabla 1 muestra las variables de entrada y salida utilizadas para construir los tres modelos considerados. El MAPE se ha utilizado para evaluar la bondad de ajuste del modelo de predicción.

Tabla 1: Modelos de comportamiento normal seleccionados

Modelo	Variables de entrada			Variable de salida	MAPE (%)
1	GP	RS		DT_1	14,84
2	GP	RS	GT	DT_1	14,78
3	GP	RS		DT_2	17,84

Por lo general, un modelo tiene mejor poder de predicción que otro, cuando su valor de MAPE es menor. En este caso, tal y como se muestra en la Tabla 1, los Modelo 1 y 2 tienen una MAPE similar. Debido a este hecho, podríamos elegir cualquiera de los dos modelos ya que las diferencias en los errores de predicción serán insignificantes. En este caso se ha elegido el Modelo 1 por ser el modelo más simple y parsimonioso. La mayor cantidad de información de los datos de entrada en el Modelo 1 se explica principalmente por el primer componente principal (CP1). Basado en el PC1, la Figura 2 muestra el T^2 de Hotelling asociado al comportamiento normal (bajo control) del equipo estudiado.

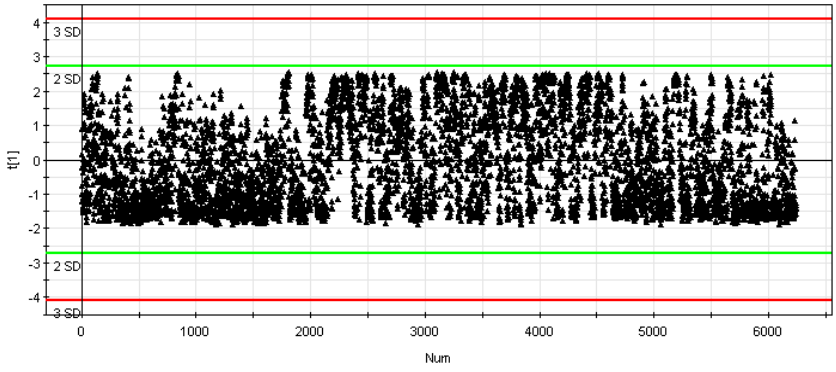

Figura 2. Gráfico T^2 para la monitorización. Modelo bajo control

La Figura 3 muestra los resultados obtenidos a partir de la monitorización usando el modelo 1 obtenido en el análisis de PCA. Específicamente, la Figura 3 representa la contribución de las variables de entrada a la T^2 y la DModX cuando se introduce una nueva observación. Como se observa en la Figura 3

la anomalía en el comportamiento puede ser detectada claramente tanto en el gráfico de la T^2 como en el de DModX ya que las nuevas observaciones se sitúan fuera del límite de control. Esto significa que el comportamiento normal de las variables se observa hasta la observación 6350 y a partir de este punto se produce una situación anormal. Es posible detectar y diagnosticar el origen de este comportamiento. En este caso, la anomalía es debida a un aumento en le tendencia de la variable de salida (Diferencias de temperaturas entre la temperatura de cojinetes y la ambiental) que se debe principalmente al aumento de la temperaturas de los cojinetes lo cual podría indicar la degradación del generador.

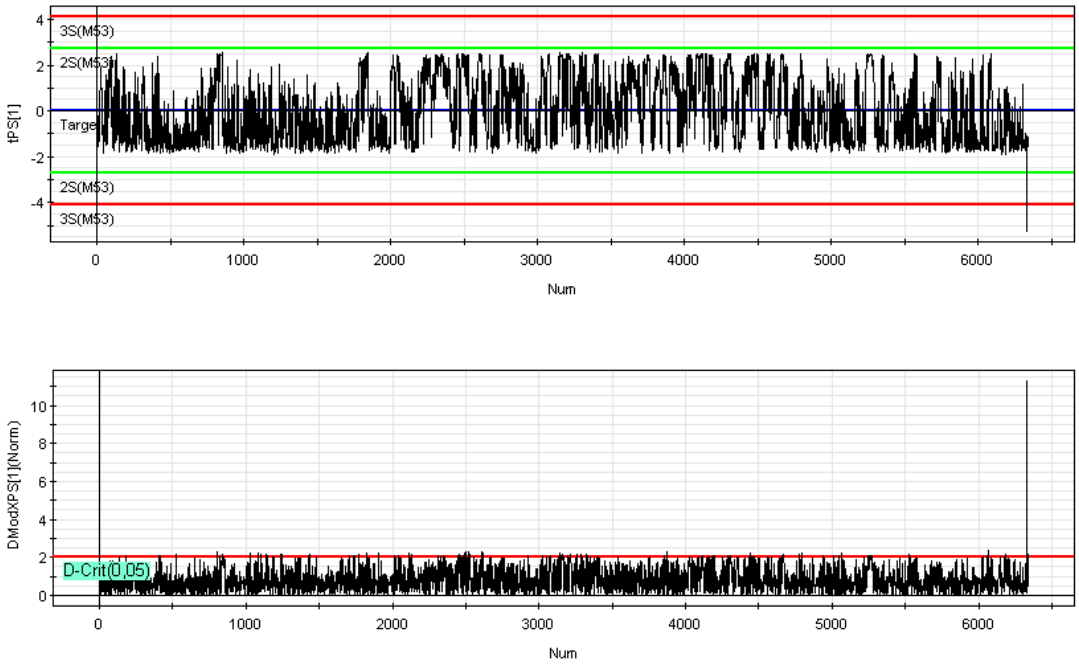

Figura 3: Gráficos T2 y DModX para la monitorización del Modelo 1. Modelo fuera de control

Una posible forma de diagnosticar este comportamiento anómalo se muestra en la Figura 4. En esta Figura, la Figura 4 a) muestra la representación de DT1 observado frente al predicho, coloreado por la DT1 observado (Gama de azul a rojo de menor a mayor DT1) y la Figura 4b) muestra la contribución de las variables de entrada (GP y RS) para los puntos con mayor DT1 observados (puntos resaltados en rojo).

Como se muestra en estas figuras, las observaciones con un comportamiento anormal presenta un observable DT1 más alto que el DT1 predicho. Estos puntos además tienen una contribución de las variables de entrada, GP y RS, normal. Estos puntos desplazados de la nube y destacado en rojo, en la Figura 4 a) y que tienen un valor mayor de la variable de salida DT1 también son los mismos que caen fuera del límite de control tal y como se muestra en la Figura 3.

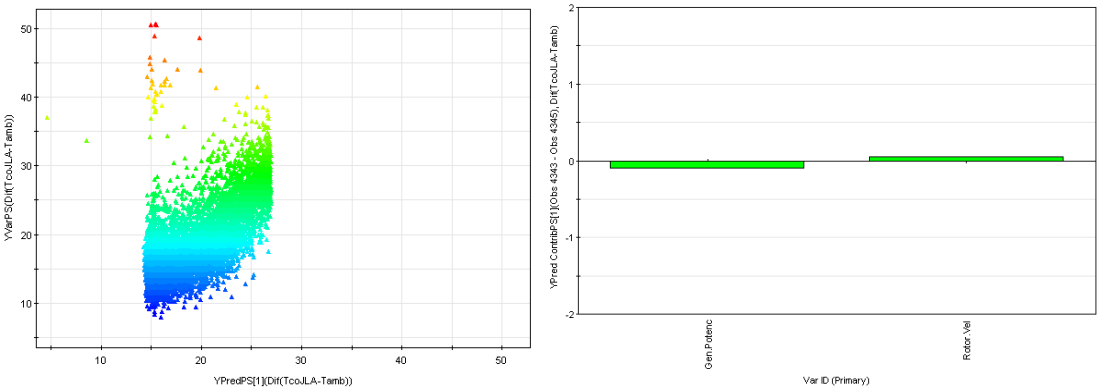

Figura 4.a) DT1 Observado vs. DT1 Predicho, b) Gráfico de contribución de variables

5. CONCLUSIONES

En el presente trabajo se ha construido un modelo mediante un enfoque basado en los datos. Se ha utilizado una técnica multivariante basada en el PCA y PLS para construir un modelo el cual ha identificado y diagnosticado el patrón de estado de un generador asíncrono. Para poder construir el modelo se han utilizado datos históricos de funcionamiento. Se ha generado un modelo de monitoreo de la condición para señales de alarma sobre la base de los resultados previstos en las observaciones.

Usando la metodología propuesta en este trabajo, se ha puede detectar un patrón anormal de un equipo lo que puede indica una posible degradación de este. Además, estas técnicas se pueden utilizar para apoyar el mantenimiento basado en la condición con el fin de reducir los costos de mantenimiento y aumentar la disponibilidad del generador.

6. BIBLIOGRAFÍA

Dragomir O. E.,Gouriveau R., Dragomir F., Minca E., Zerhouni N., 2009, Review of Prognostic Problem in Condition-Based Maintenance, 2009, European Control Conference ECC'09, Budapest, Hungary

Geladi P.,Kowalsky B.R., 1986, Analytica Chimica Acta, 185, 1-17.

Hao L., Jinsong Y., Ping Z., Xingshan L., 2009, A Review on Fault Prognostics in Integrated Health Management, The Ninth International Conference on Electronic Measurement & Instruments, ICEMI'2009.

Jardine A.K.S, Lin D., Banjevic D., 2006, A review on machinery diagnostics and prognostics implementing condition-based maintenance, Mechanical Systems and Signal Processing, 20, 1483-1510

Kusiak A., Verma A., 2011, Prediction of Status Patterns of wind turbines: A data mining approach, Journal of Solar Energy Engineering, 133, 110081-110090.

Lam H., Klemes J., Friedler F., Kravanja Z., Varbanov P., 2010, Software tools overview: process integration, modelling and optimisation for energy saving and pollution reduction, Chemical Engineering Transactions, Vol. 21, 487-491.

MacGregor, J.F., Kourti, T., 1995, Statistical process control of multivariate processes, Control Eng. Practice, Vol.3, No.3, 403-411.

Wold S., Esbensen K., Geladi P., 1987, Principal Components Analysis, Chemometrics and Intelligent Laboratory Systems, Vol. 2, 37-52.

AGRADECIMIENTOS

Los autores agradecen al proyecto CENIT-E Ocean Lider por el apoyo financiero de este trabajo (CEN-20091039).