BÚSQUEDA DE CONOCIMIENTO EN PROCESOS INDUSTRIALES: CASO PRÁCTICO

Martínez-de-Pisón, F.J.(p); Pernía, A.; Martínez-de-Pisón, E.; Lostado, R.; Castejón, M.

Abstract

In this paper a methodology for discovering useful knowledge from time series from industrial processes is presented. The process consists on detecting useful patterns from time series and afterwards, searching frequent itemsets in order to obtain associations rules. A practical methodology application in an industrial process is showed.

Keywords: Knowledge discovery in time series, industrial processes.

Resumen

En este artículo se plantea una metodología para la búsqueda de conocimiento útil para la toma de decisiones a partir de las series temporales capturadas de procesos industriales o medioambientales. El proceso consiste en buscar patrones significativos en las series de datos y después utilizar técnicas basadas en búsqueda de ítems frecuentes para generar reglas de asociación. Se muestra un caso práctico de aplicación de dicha metodología en un proceso industrial.

Palabras clave: Descubrimiento de conocimiento en series temporales, procesos industriales

1. Introducción

Uno de los campos de investigación de la Gestión del Conocimiento que más futuro tiene para la optimización de Procesos Industriales, corresponde con la búsqueda de conocimiento oculto dentro de series temporales multivariantes. Este tipo de datos es muy típico y aparece en cualquier base de datos de históricos de una planta industrial.

Hoy en día, el desarrollo de herramientas que permitan extraer conocimiento de grandes bases de datos de históricos puede ser de gran ayuda para la toma de decisiones o la mejora de los procesos productivos.

En este artículo, se muestra parte de los resultados obtenidos en los trabajos desarrollados dentro del **Proyecto Nacional DPI2006-03060** titulado “BÚSQUEDA AUTOMÁTICA DE CONOCIMIENTO OCULTO EN SERIES TEMPORALES DE PROCESOS INDUSTRIALES DEL ACERO Y ELASTÓMEROS MEDIANTE ALGORITMOS QUE OBTIENEN REGLAS DE ASOCIACION EN LINEA (CONOSER)”. Los objetivos principales del proyecto CONOSER se orientan hacia el desarrollo de algoritmos y herramientas que puedan ser utilizadas para la obtención de reglas de asociación a partir de históricos de procesos industriales. Estas reglas servirán para obtener conocimiento oculto que pueda ser de ayuda en la toma de decisiones y mejora de los procesos productivos.

En particular, se plantea un tipo de metodología para la búsqueda de reglas asociativas que permitan extraer conocimiento oculto de bases de datos compuestas por series temporales multivariantes. La idea básica consiste en encontrar, dentro de las series temporales, interrelaciones entre patrones que se repitan asiduamente y representar dichas relaciones de una forma fácilmente comprensible por el experto.

Por ejemplo, analizando las series temporales correspondientes a un cierto proceso industrial, tal y como se muestra en la Figura 1, podríamos deducir la siguiente regla: "*cuando la temperatura sube linealmente y la presión permanece por debajo de un cierto nivel X entonces se produce un descenso de la calidad del producto*". Esta regla, no conocida previamente, permitiría a un experto tomar decisiones para poder evitar esa pérdida de calidad del producto. Lógicamente, solo será interesante si se repite un cierto número de veces.

1.1 Búsqueda de Patrones Frecuentes en Series Temporales

Es bien conocido, que la capacidad del cerebro humano para segmentar y extraer patrones visuales es muy superior a la de cualquier sistema de visión artificial actual. Igualmente, el cerebro es capaz de discernir sonidos, sabores, olores o texturas al tacto.

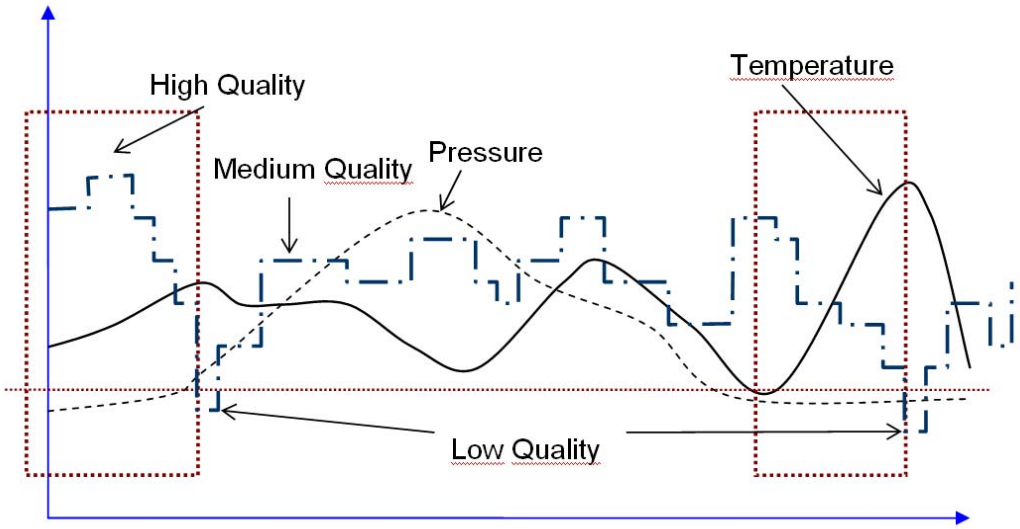

Figura.1. Detección de relaciones temporales entre variables de un proceso industrial.

El analista que pretende extraer algún conocimiento útil que permita desarrollar estrategias de mejora de un proceso industrial, utiliza esta habilidad para descubrir visualmente patrones repetitivos y sus interrelaciones en el tiempo. Para ello, el experto hace uso de gráficas temporales como la mostrada en la Figura 1.

Muchas veces, cuando escuchamos a un experto argumentar sobre el comportamiento de una serie temporal oímos términos del estilo "*este segmento crece linealmente*" o "*este segmento decrece exponencialmente*", que demuestran claramente la forma en que el ser humano describe localmente una serie temporal. Este tipo de segmentación visual se realiza dividiendo la serie temporal en subseries o segmentos cuya forma o apariencia se aproxima a patrones ya conocidos (líneas, curvas crecientes o decrecientes, etc.) igual a como hace el cerebro para describir cualquier nuevo objeto que se le presenta.

Una vez detectados los segmentos característicos, es requisito indispensable para poder establecer una posible relación, que exista una serie de situaciones repetidas en donde aparezcan siempre los mismos patrones. En el caso de la Figura 1, la temperatura debería subir linealmente, la presión permanecer por debajo de un cierto nivel y producirse un descenso de la calidad del producto; un número significativo de veces y dentro de una ventana de temporal adecuada.

La búsqueda de este conocimiento puede complicarse aún más, porque este tipo de correlaciones locales no solamente puede corresponder al mismo instante temporal, sino que pueden aparecer dependencias entre variables con importantes desfases en el tiempo. Esto es muy común en sistemas con fuertes inercias como algunos procesos químicos o físicos, cuyas velocidades de respuesta son muy lentas e incluso varían según las condiciones del entorno.

De lo anteriormente expuesto, se puede deducir que esta tarea solamente se puede realizar visualmente cuando la cantidad de información a manejar no es muy elevada, siendo prácticamente imposible cuando el número de variables y el número de observaciones crece considerablemente. Esto es típico en los procesos industriales donde podemos encontrarnos con decenas o centenas de parámetros y decenas de miles o cientos de miles de observaciones en cada una de ellas.

En estos últimos años han surgido técnicas de Minería de Datos orientadas al análisis de Series Temporales que pretenden ayudar a analizar grandes bases de datos con el objetivo de obtener conocimiento útil y oculto que pueda servir de ayuda para la toma de decisiones o la generación de nuevos modelos de predicción.

El objetivo fundamental de este artículo se centra en la descripción de los resultados obtenidos dentro del proyecto Nacional de investigación CONOSER (DPI2006-0306). En particular, se describe la metodología planteada, las herramientas desarrolladas y un caso práctico de búsqueda de conocimiento en un proceso industrial.

2. Metodología Propuesta

La metodología inicial, después del estudio de la bibliografía existente y de un análisis exhaustivo de las aplicaciones actuales, se ha modificado parcialmente en las fases de preprocesado y segmentación de series temporales.

Inicialmente, se desarrolló un intenso estudio de las técnicas actuales de preprocesado, segmentación y búsqueda de patrones frecuentes. La conclusión más importante de este estudio fue que ninguna combinación de la enorme cantidad de técnicas existentes es 100% efectiva para el preprocesado y segmentación automática de series temporales de procesos industriales. Esto es debido a la completa heterogeneidad que existe entre los diversos tipos de series temporales de un proceso industrial (temperatura, presión, activación o desactivación de un motor, etc.), lo que requiere que el experto trabaje de una forma iterativa e interactiva con diversos algoritmos de preprocesado y segmentación para cada tipo de serie temporal a segmentar.

De este modo, la orientación de la metodología en todas estas fases se orientó hacia el desarrollo de diversas técnicas de preprocesado y segmentación que fueran de fácil aplicación por parte del experto analista del proceso industrial. Todos los algoritmos de preprocesado y segmentación de series temporales se están implementando para el programa gratuito de análisis estadístico R (http://www.r-project.org) en lenguaje R y C (gcc) y dentro de una librería denominada “*KDSeries*” que será entregada a la comunidad de usuarios de R bajo licencia GPL.

Además, debido al enorme esfuerzo prueba/error que es necesario realizar en estas etapas, se decidió desarrollar una nueva herramienta visual, denominada CONOTOOL, que facilitara estos trabajos. Esta herramienta, que está en fase de desarrollo, permite un uso más intuitivo y rápido de todas de las funciones de la librería “*KDSeries*”. Actualmente, este software, programado en C++Builder 5.0 y R, está en fase de desarrollo con un 75% realizado. Para facilitar la internacionalización todo el software está escrito completamente en Inglés.

2.1 Etapas de la Metodología Propuesta

La metodología propuesta se compone de las siguientes etapas:

1. **Filtrado de cada serie temporal para eliminar el ruido y obtener la forma básica de la misma:** Algunas de las técnicas que ya están implementadas en la librería "KDSeries" y en la herramienta visual "CONOTOOL" permiten realizar filtros de ventana deslizante bajo kernel (gaussianos, rectangulares, máximos, mínimos, mediana, etc.), eliminar según un umbral (mínimo, máximo o rango), basadas en la transformada rápida de Fourier FFT, aproximación lineal a tramos, aproximación constante a tramos, aproximación agregada a tramos o aproximación constante adaptativa a tramos. Con respecto a las técnicas de extracción de características, cabe resaltar que las técnicas se han reorientado para que sea más útiles en la iteración con el usuario, pues la extracción automática no ha dado resultados muy óptimos aunque se han dejado disponibles en la librería *KDSeries*.
2. **Obtención de los cruces por cero de la primera derivada (máximos y mínimos de la señal):** Una vez, filtrada la señal, se obtienen los máximos y mínimos de cada una para poder identificar patrones crecientes o decrecientes.
3. **Extracción de los tramos "incrementales (INC)", "decrementales (DEC)", "horizontales (HOR)" o "según un umbral (UMB)" según unos valores previamente establecidos para cada serie temporal.**

El programa permite incluir un banco de búsquedas de diferentes subpatrones en cada serie temporal según la altura, anchura y tipo de curva. Estos filtros pueden ajustarse en el orden que se desee.

Con respecto a estas tareas, inicialmente se desarrollaron técnicas que permitían extraer automáticamente patrones frecuentes. El problema es que, debido a la tremenda heterogeneidad de las series temporales, se observó que los resultados no eran completamente satisfactorios. Aunque se han dejado disponibles las técnicas anteriores en la librería KDSeries, se han desarrollado nuevas técnicas completamente configurables por el experto.

4. **Agrupamiento de los tramos "INC", "DEC" y "HOR" en patrones más complejos según especificaciones determinadas por el experto y etiquetado de los patrones obtenidos para cada serie temporal:** El experto puede, mediante un lenguaje sencillo, realizar búsquedas de patrones frecuentes que sean de su interés.
5. **Búsqueda de secuencias de patrones y extracción de las reglas asociativas:** Se buscan secuencias de patrones que se repitan entre las diversas series temporales según una ventana temporal preestablecida y se extraen las reglas asociativas de aquellos casos que aparecen un elevado número de ocasiones y tienen un elevado índice de aciertos. Por último, se presentan las reglas asociativas en una forma entendible por el usuario.

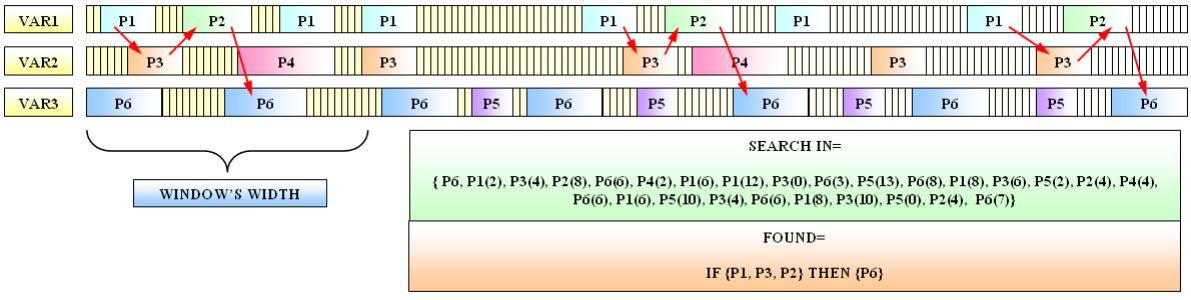

Figura 2. Búsqueda de reglas asociativas en varias series temporales.

Una vez identificados los patrones repetitivos que aparecen en cada una de las variables del proceso, entramos en la etapa final para la extracción de conocimiento.

La idea fundamental es la de facilitar el trabajo tedioso de búsqueda de correlaciones temporales entre variables y de mostrar de una forma fácilmente comprensible dichas relaciones locales. Esta búsqueda pretende detectar relaciones temporales entre variables que puedan ser la causa del comportamiento de una tercera variable (ver Figura 2).

Es decir, en esta etapa se pretende buscar correlaciones locales entre series temporales que se repiten con un cierto grado de asiduidad y resumirlas en una serie de reglas asociativas del tipo:

$$\begin{aligned} &IF\ \{P_1(t-d_1), P_2(t-d_2), ...\}\ THEN\ S(t) \\ &WITH\ (Support = X, Confidence = Y) \end{aligned} \quad (1)$$

Donde:

- P1(t), P2(t),...: corresponden al “Antecedente” de la regla e indica aquella secuencia de patrones de diferentes variables que aparecen repetidos antes o durante el patrón “Consecuente”.
- S(t): es el patrón de salida denominado “Consecuente”.
- d1, d2,...: corresponden a los desfases entre los patrones Antecedentes con respecto al patrón Consecuente.
- Cobertura o Soporte (Support): que corresponde con el número de veces o el porcentaje veces que se repite la regla en toda la base de datos. Indica el grado de generalización de la misma.
- Precisión o Confianza (Confidence): Porcentaje de veces que la regla se cumple cuando se puede aplicar. Es decir, cuando existe el Antecedente qué probabilidad existe que aparezca el Consecuente.

3. Descripción de la herramienta visual CONOTOOL

La herramienta de software, llamada CONOTOOL, pretende facilitar un uso intuitivo y rápido de todas las funciones desarrolladas en R para preprocesado, segmentación y búsqueda de reglas de asociación a partir de una base de datos de un proceso industrial, medioambiental o cualquier otro.

Actualmente, este software, programado en C++ Builder 5.0 y R, está en fase de desarrollo con un 75% realizado.

En la Figura 3 se muestra la pantalla principal del software desarrollado. En ella, se puede abrir un proyecto ya creado o empezar uno nuevo.

Si elegimos crear un nuevo proyecto, pulsamos en “*New Project*" y elegimos una base de datos en formato texto con extensiones “txt” o “csv”. En la primera fila del archivo deben estar los nombres de las variables y en las filas posteriores cada una de las observaciones de la base de datos. El programa permite que le indiquemos cómo están separados los datos y cuál es el punto decimal. Esto da gran flexibilidad al usuario final.

Una vez elegida la base de datos, nos muestra un resumen estadístico de cada una de las variables o atributos. Cabe destacar que cada una de las columnas debe ser un atributo y cada fila (exceptuando la primera que serán los nombres de las variables/atributos) una observación efectuada en el tiempo o en un instante determinado.

Finalmente, podemos guardar el proyecto con el nombre que deseemos. Una vez almacenado o abierto un proyecto ya existente, aparece la pantalla de menú de procesado.

Hasta ahora, el programa permite realizar una exploración de los datos, selección de filtros y búsqueda de subpatrones. Posteriormente se realizará las partes correspondientes a búsqueda de patrones y reglas de asociación.

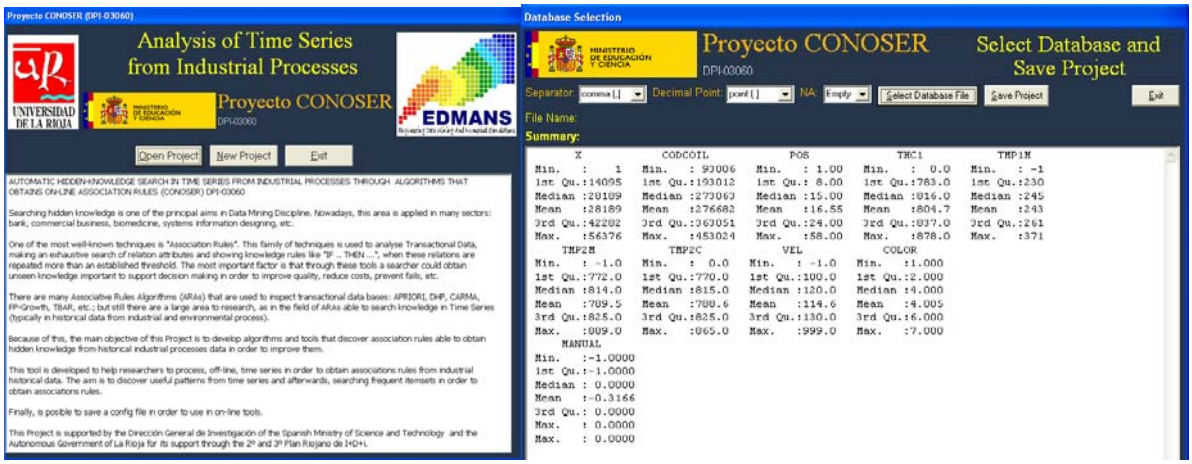

Figura 3. Pantalla Principal y de Selección de la Base de Datos.

La herramienta de análisis exploratorio de los datos (AED) permite dibujar las series temporales seleccionadas según el rango elegido, permite realizar scatterplots, boxplots, histogramas, diagramas PCA, SOM, etc.

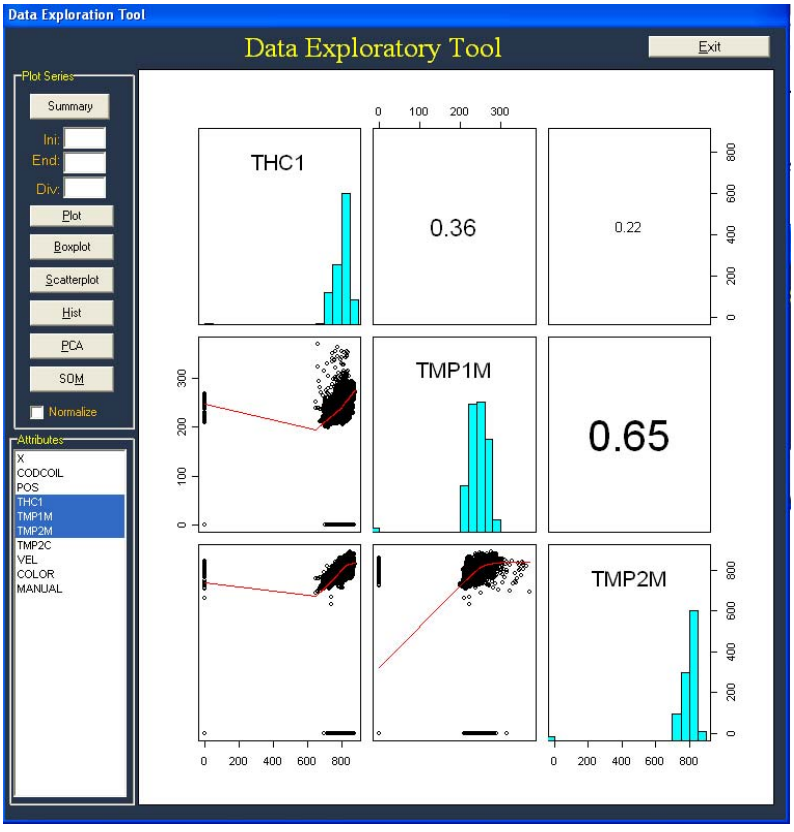

Figura 4. Herramienta de Análisis Exploratorio de las Series Temporales.

Una vez analizada la información, podemos seleccionar un banco de filtros para cada serie temporal.

Con la herramienta de filtros (Figura 5, izquierda), podemos asignar diferentes filtros a cada serie temporal. Podemos visualizar el efecto que produce un único filtro (botón "*Plot*") en la serie temporal o ver el efecto producido del conjunto de filtros aplicados a la misma (botón "*Apply all filter to this Attribute*"). En negro aparece la serie original y en rojo la serie filtrada.

El orden de aplicación de los filtros se elige según el orden de aparición en la pantalla (de arriba hacia abajo). Pulsándose los botones de subir, bajar, eliminar, etc.; se pueden reorganizar de la forma conveniente el banco de filtros.

Los filtros permiten:

- Eliminar los datos por debajo, encima, entre un rango o fuera de un rango de valores de corte: Min_Filter(), Max_Filter(), Range_Filter(), InvRange_Filter() respectivamente.
- Utilizar un filtro de ventana deslizante del tipo gaussiana, de mediana, de media, de mínimo o de máximo: Gauss_Filter(), Median_Filter(), Mean_Filter(), Min_Filter() y Max_Filter() respectivamente.
- Usar la transformada rápida de Fourier FFT aplicando una ventana rectangular o gaussiana: FFT_Filter_Mean() FFT_Filter_Gauss().

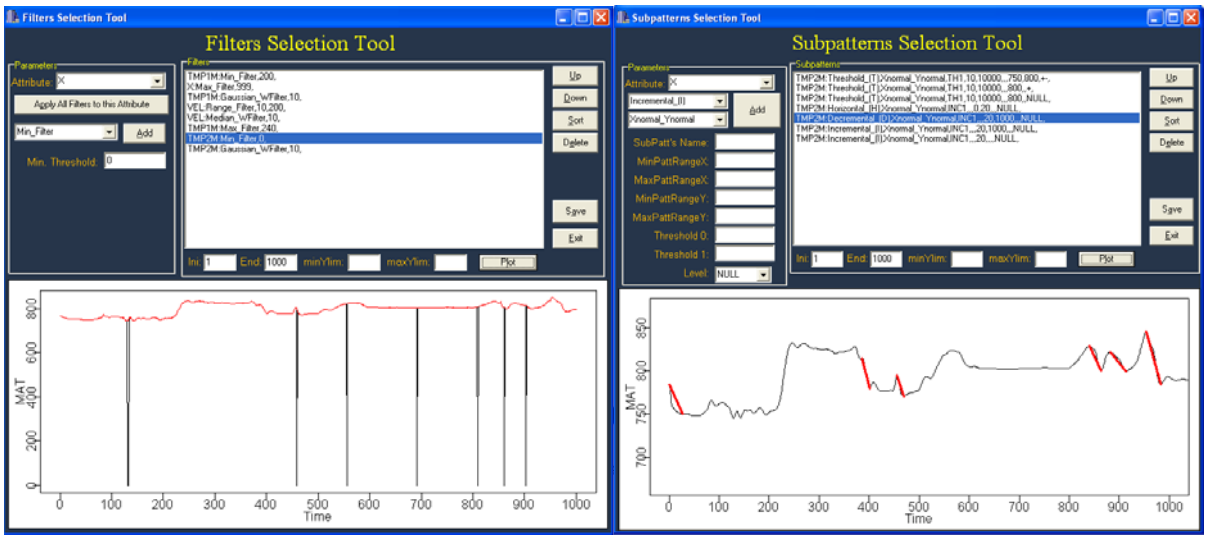

Figura 5. Herramienta de Selección de Filtros y Búsqueda de Subpatrones.

Una vez seleccionado el banco de filtros se utiliza la herramienta para búsqueda de subpatrones (Figura 5, derecha).

Igual que con la herramienta de filtros, el programa permite incluir un banco de búsquedas de diferentes subpatrones en cada serie temporal. Estos filtros pueden ajustarse en el orden que se desee. El programa permite dibujar los subpatrones encontrados en cada serie temporal. El rango de visualización se puede configurar directamente en la pantalla.

Los parámetros de ajuste son los siguientes:

- pattType= Tipo de subpatrón a buscar: Incremental ("I"), decremental ("D"), horizontal ("H") or threshold ("T")
- pattRangeX= Rango formado por un vector de dos valores c(mínimo, máximo) de X en los que tiene que estar comprendido el subpatrón.

- pattRangeY= Rango formado por un vector de dos valores c(mínimo, máximo) de Y en los que tiene que estar comprendido el subpatrón.
- rangeType= Si se consideran los rangos de X e Y por valores fijos (N) o porcentajes (P). Por ejemplo, un vector c(“N”,”P”) considera los rangos de X por valores fijos y los de Y por porcentajes.
- threshold= Valores de corte en donde buscar los patrones. Por defecto es NULL.
- level= Los patrones se buscarán por encima ("+"), debajo ("-") or entre dos valores ("+-") de threshold. Por defecto es NULL.
- namePatt= Nombre del patron encontrado.

4. Caso Práctico

Como caso práctico, se muestra una experiencia de aplicación de estas técnicas donde se realizó una búsqueda de las causas que originaban pérdidas de calidad en el recubrimiento de nuevos aceros dentro de una planta de galvanizado. Del estudio se obtuvo conocimiento que sirvió para detectar qué circunstancias afectaban a la calidad del recubrimiento de las bobinas y cuáles eran las acciones de control que se podían establecer como medida de seguridad para reducir los problemas aparecidos.

En particular, se buscó identificar la influencia de los parámetros del proceso de galvanizado en la adherencia del recubrimiento de zinc, que es uno de los factores del que depende la resistencia a la corrosión del producto galvanizado.

4.1 Descripción del Problema

La resistencia a la corrosión suele depender de la calidad, espesor y uniformidad del recubrimiento de zinc y le afectan múltiples factores: preparación de la superficie del metal, composición y temperatura del baño de zinc, velocidad de la banda, control de las cuchillas de aire que regulan el espesor del recubrimiento, temperatura de la banda, calidad del ciclo de recocido, composición de la atmósfera, etc. Así, es necesario un contenido mínimo o nulo de oxígeno en cada una de las fases del proceso para evitar la oxidación.

El ajuste de todos los parámetros correspondientes a cada tipo de bobina según el tipo de acero y dimensiones de la misma (espesor y anchura), hacen esta tarea muy laboriosa. Además, los problemas se acentúan aún más cuando se procesan nuevos aceros o nuevos espesores que no han sido tratados anteriormente y de los que no se dispone de información veraz ni de modelos matemáticos adecuados.

Ante estos nuevos productos, los ingenieros de planta deben realizar estimaciones y múltiples ajustes hasta conseguir que la adherencia y homogeneidad del recubrimiento de zinc sea el adecuado. Esto supone un consumo muy elevado de tiempo y material desechado.

Muchas veces, el conocimiento implícito que se genera en estos trabajos no queda plasmado en ningún sitio pues, al ser tantas las variables a modificar y ajustar, resulta muy difícil determinar cuáles son realmente cruciales en la calidad del producto final.

El problema no es fácil, pues además del enorme número de variables que influyen en el proceso, el número de bobinas defectuosas existentes en las bases de datos a analizar no suele ser muy elevado, lo que dificulta mucho el análisis con técnicas clásicas de estadística multivariante.

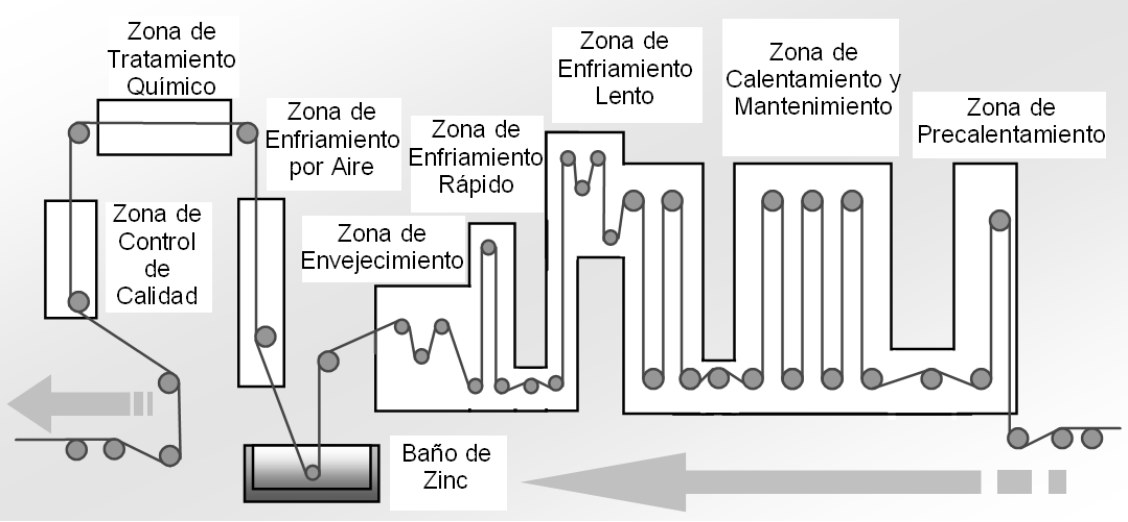

Figura 6. Esquema básico de una Línea de Galvanizado en Caliente (LGC).

4.2 La línea de galvanizado en continuo

A grandes rasgos, el proceso dentro de una línea de galvanizado en continuo se puede describir de la siguiente manera :

1. El primer paso, consiste en la formación de una banda continua a partir de las bobinas de acero que proceden de los trenes de laminación. Para ello, se despunta la cabeza y la cola de las mismas y se sueldan a solape. El resultado final es una banda de acero continua formada por las bobinas entrantes.
2. A continuación, la banda atraviesa una zona de precalentamiento en atmósfera no oxidante donde se eliminan las impurezas, se volatilizan los aceites de laminación y se reduce el óxido superficial.
3. Posteriormente, se somete la banda a un ciclo de calentamiento y enfriamiento que es denominado “recocido” (ver Figura 7). Este tratamiento es esencial para la mejora de las propiedades del acero y del recubrimiento final. El objetivo es recristalizar el metal endurecido que sale de la laminación en frío y homogeneizar la estructura cristalina.
4. El horno donde se realiza este tratamiento, suele estar dividido en varias zonas: una zona de calentamiento formada por ocho subzonas, una zona de mantenimiento, una de enfriamiento lento y otra de enfriamiento rápido. Al final del enfriamiento rápido se produce un envejecimiento o “igualación” con el fin de garantizar la precipitación del carbono y así minimizar los efectos de envejecimiento del acero.
5. A continuación, la banda se sumerge en un pote de zinc fundido a temperatura constante para revestirla de dicho metal. De este baño, la banda sale verticalmente pasando entre cuchillas de aire que regulan el espesor del recubrimiento.
6. Después, atraviesa una serie de procesos auxiliares de tratamientos químicos donde se aplica una leve película de ácido crómico para prevenir la oxidación blanda.
7. Por último, se realiza un aplanado hasta obtener el producto final bien en forma de bobinas o chapas cortadas.
8. Esta descripción, con modificaciones menores, suele ser válida para la mayoría de líneas de galvanizado en continuo por inmersión instaladas en todo el mundo.

El estudio se realizó para una partida de bobinas de un nuevo tipo de acero que había presentado un cierto número de ellas con irregularidades en la adherencia de la capa de

zinc. La base de datos se extrajo de las primeras pruebas de ajuste de la planta por lo que el porcentaje de bobinas con adherencia irregular fue bastante significativo.

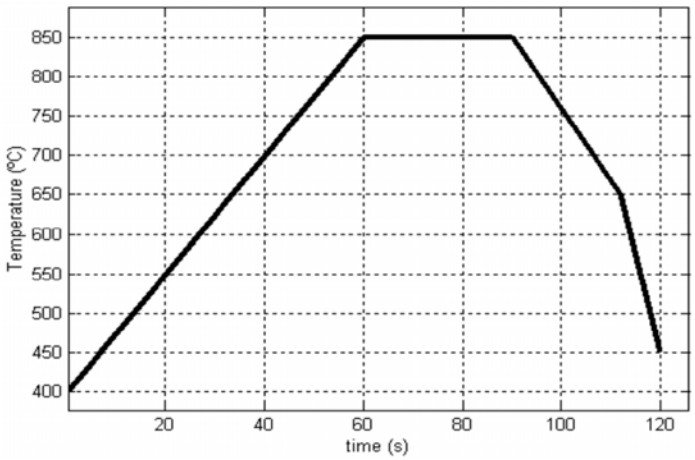

Figura 7. Curva de tratamiento térmico.

Obviamente, el análisis de la base de datos proveniente de los primeros ajustes es muy útil pues la variedad de consignas diferentes debidas a los numerosos ajustes permite poder explorar más conocimiento que en procesos ya ajustados y en régimen permanente, donde dichas consignas prácticamente no varían.

PORCENTAJE DE BOBINAS DE LA BASE DE DATOS		
Bobinas TIPO A (Adherencia Regular)	Bobinas TIPO B (Adherencia Irregular)	Total
684 (94,6%)	39 (5,4%)	723

Tabla 1. Porcentaje de bobinas con adherencia “regular” e “irregular”.

En la Tabla 1 se muestra el número y porcentaje de bobinas con adherencia regular e irregular de la base de datos que se usó en el estudio.

4.3 Preprocesado y Segmentación de las Series Temporales

Las variables que se seleccionaron, correspondían con medidas de cada una de las zonas del proceso (ver tabla 2). Para cada una de estas zonas, se obtuvieron las series temporales correspondientes a: la composición del aire (en porcentaje de H2 y O2), la temperatura de cada zona, la temperatura de la banda de acero medida en varios sitios y obtenida mediante pirómetros, la velocidad de la banda, la temperatura de rocío en diversas zonas del proceso, etc. (ver tabla 3). Además, se incluyeron las dimensiones de cada bobina (ancho y espesor), composición química del acero de cada bobina procesada, así como la temperatura y composición del baño de zinc. Todas las medidas se realizaron cada 100 metros lineales de banda de acero. De esta forma, cada valor de cada serie temporal correspondía con la media de los valores medidos en 100 metros de banda de acero.

La variable de salida, correspondía con un valor binario 0 o 1 que indicaba si el espesor de la capa de zinc entraba dentro de la tolerancia (0) o no (1).

ZONAS PRINCIPALES DEL HDGL

Zona	Descripción
Precalentamiento (PRE)	Zona de precalentamiento de la banda, donde se realiza la limpieza de la misma.
Calentamiento: subzonas 1 a 8 (CAL)	Subzonas 1 a 8 de la zona de calentamiento del Horno donde se aumenta la temperatura de la banda entorno a los 850ºC.
Mantenimiento: subzonas 9 y 10 (MAN)	Zona donde se mantiene la temperatura de la banda entorno a los 850ºC: Subzonas 9 y 10.
Enfriamiento Lento (ENL)	Zona donde se realiza un enfriamiento lento de la banda hasta la temperatura de 600-650ºC.
Enfriamiento Rápido (ENR)	Zona donde se realiza un enfriamiento rápido de la banda hasta la temperatura de 400-450ºC.
Igualación (IGU)	Zona donde se homogeniza la temperatura de la banda antes de su inmersión en el pote de zinc.
Pote de Zinc (POT)	Zona de Inmersión en el baño con zinc fundido.
Trompa (TRM)	Zona de la Trompa de Igualación.
General (GRL)	Parámetros que son constantes en toda la HDGL.

Tabla 2. Zonas de medición dentro del HDGL.

GRUPOS DE VARIABLES	
Nombre	Descripción
CODE	Código de la bobina.
CON_H2	Concentración de H2 en el aire presente en esa zona.
CON_O2	Concentración de O2 en el aire presente en esa zona.
TMP_PR	Temperatura del Punto de Rocío en esa zona.
Zxx_TMP	Temperatura de la zona xx.
TMP_Pxx	Temperatura de la banda, medida con el pirómetro xx, en cada zona del proceso.
ESP, ANCH	Espesor y Anchura de la banda de acero.
SPD	Velocidad de la banda de acero.
POT_TMP	Temperatura del baño de zinc.
CMP_BAÑO	Composición química del baño de zinc. Porcentaje de los elementos químicos más relevantes del baño de zinc.
CMP_ACERO	Composición química del Acero. Porcentaje de los elementos químicos más relevantes: Fe, Mn, Al, Ni, etc.
ADHERENCIA	Si la adherencia del recubrimiento de zinc entraba dentro de las tolerancias (0) o no (1).

Tabla 3. Descripción de los grupos de variables más relevantes.

Después de la selección de las variables más representativas del proceso, se seleccionó una batería de filtros para cada serie temporal. El filtrado de cada serie temporal pretendía extraer la forma básica de la misma para facilitar el proceso posterior de extracción de subpatrones. En la Figura 8 se puede apreciar en negro los primeros 400 valores de la temperatura de la banda de acero cuando sale del horno, en rojo la temperatura después de filtrar los valores menores de 200 y en azul el resultado final al aplicar un filtro gaussiano con una ventana de ancho igual a 20.

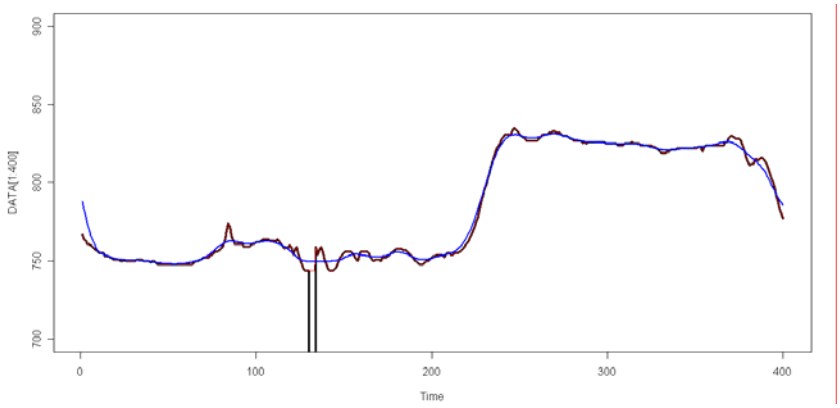

Figura 8. Ejemplo de filtrado de una serie temporal correspondiente a la temperatura de la banda de acero cuando sale del horno.

Una vez obtenidas las formas básicas de cada serie temporal, se procedió a la extracción de subpatrones característicos de cada una de ellas. Prácticamente, la mayoría de los subpatrones extraídos correspondía con momentos en donde la serie temporal superaba por encima o por debajo un determinado valor. Por ejemplo, cuando las concentraciones de O2, H2, y composiciones químicas del baño de zinc se salían de unos márgenes preestablecidos.

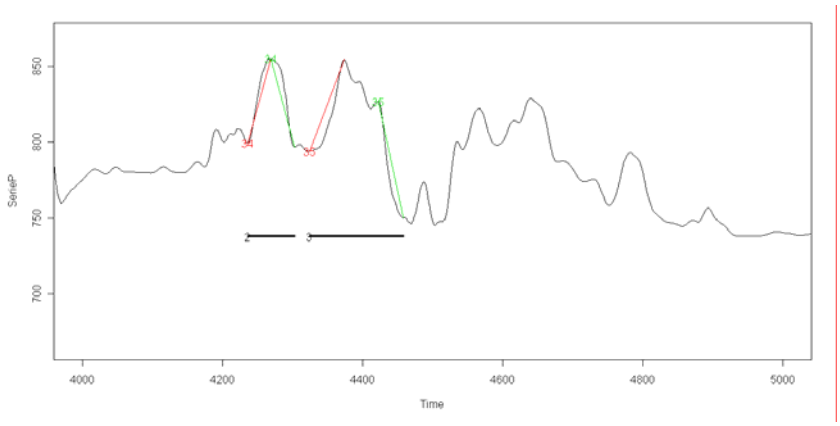

Figura 9. Extracción de dos patrones compuestos por un subpatrón incremental seguido de un decremental.

En cambio, en las temperaturas del horno y temperaturas y velocidad de la banda de acero se buscaron subpatrones que indicaran fuertes incrementos o decrementos en un corto periodo de tiempo. Además, se identificaron oscilaciones elevadas a partir de la búsqueda de patrones incrementales seguidos de patrones decrementales o viceversa (Figura 9).

4.4 Búsqueda de Reglas de Asociación

Una vez obtenidos todos los subpatrones y patrones significativos de cada una de las series temporales analizadas y, mediante una ventana deslizante a lo largo del tiempo y con un ancho de ventana configurado por el usuario, se construyó una matriz compuesta, cada una de sus filas, por los nombres de los subpatrones y patrones que entraban dentro de dicha ventana en un instante temporal determinado.

De esta forma, cada una de las filas de la matriz, correspondía con los patrones o subpatrones que aparecían en un instante determinado y dentro de una ventana temporal prefijada.

El proceso final de búsqueda de las reglas de asociación se desarrolló con un algoritmo basado en el algoritmo APRIORI para la búsqueda de itemsets frecuentes.

Después de eliminar las reglas no útiles y de filtrar las equivalentes, se obtuvieron algunas de las siguientes reglas de conocimiento:

1. **IF** (POT_TMP_INF & CAL_Z06_INF & GRL_SPD_DEC) **THEN** ADH_BAJ (Support=0.09, Confidence=0.86)
2. **IF** (GRL_SPD_INC_DEC & POT_TMP_INF & TMP_P2_BAJ) **THEN** ADH_BAJ (Support=0.08, Confidence=0.90)

La primera regla de conocimiento permitía identificar que cuando la temperatura del baño de zinc era baja (POT_TMP_INF), la temperatura en la zona 6 del horno también era baja (CAL_Z06_INF) y la velocidad de la banda decrementaba rápidamente (GRL_SPD_DEC) entonces aparecía una capa de zinc defectuosa (ADH_BAJ). Esto ocurría en el 9% de la lista de patrones analizada (Support=0.09) y el 86% de las veces que aparecía el antecedente se cumplía el consecuente (Confidence=0.86).

La segunda regla de conocimiento identifica que cambios bruscos de velocidad (GRL_SPD_INC_DEC) junto con temperaturas del baño de zinc bajas (POT_TMP_INF) y temperaturas bajas de la banda a la salida del horno (TMP_P2_BAJ) producían errores en la capa de zinc en un 8% de la base de datos (Support=0.08) y se cumplía un 90% de las veces (Confidence=0.90).

Estas dos reglas presentaban una buena precisión (confianza del 86% y 90% respectivamente) y un soporte representativo (9% y 8%) y permitían identificar las causas de fallos de adherencia en algunas bobinas estudiadas.

Conclusiones

En este artículo se ha pretendido demostrar que la aplicación de herramientas software para preprocesado y segmentación de series temporales unido al uso de técnicas de minería de datos como las reglas de asociación, pueden ser de gran utilidad para la búsqueda de conocimiento oculto en históricos de procesos industriales que pueda ser de ayuda en la toma de decisiones.

La experiencia obtenida ha demostrado que este tipo de técnicas dependen mucho de la pericia del analista y que el “factor humano” es vital en cada una de las etapas propuestas en esta metodología.

Cabe destacar, que el uso de este tipo de herramientas no solo se puede aplicar para la extracción de conocimiento de los procesos industriales sino que se puede extrapolar a otro tipo de ámbitos: medioambiente, negocios, marketing, etc.

Referencias

[1] Agrawal R, Srikant R. “Fast algorithms for mining association rules in large databases”. In Proc.20th Int. Conf. on Very Large Data Bases, pp 487–499. 1994

[2] Agrawal R, Srikant R. “Mining sequential patterns”. In Proc. 11th Int. Conf. on Data Engineering, (Washington, DC: IEEE Comput. Soc.) 1995

[3] Bettini, C., Wang, X. S., and Jajodia, S. “Mining temporal relationships with multiple granularities in time sequences”. Data Engineering Bulletin 21, 1, 32-38. 1998

[4] Casas-Garriga G. “Discovering unbounded episodes in sequential data”. In Proc. 7th Eur. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD’03), Cavtat-Dubvrovnik, Croatia, pp 83–94. 2003.

[5] Harms. S, Deogun. J, Tadesse. T. “Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences”. University of Nebraska, USA. 2003.

[6] Harms, S & Deogun J. “Sequential Association Rule Mining with Time Lags”. Journal of Intelligent Information Systems. 22:1. pp. 7-22. 2004

[7] Harms. S, Tadesse. T, Wilhite. D, Hayes. M & Goddard S. “Drought Monitoring Using Data Mining Techniques: A case study for Nebraska, USA. Natural Hazards, 33, pp. 137-159. 2004

[8] Last Mark et al. “Data mining in time series databases”. Series in machine perception and artificial intelligence (Vol.57). World Scientific Publishing Co.Pte.Ltd. USA & UK. 2004

[9] Liang-Xi Qin & Zhong-Zhi Shi. “Efficiently Mining Association Rules from Time Series”. International Journal of Information Technology, Vol.12. Nº 4. 2006

[10] Li, Junzhi, et al. “Association Rules Mining from Time Series based on Rough Set”. ISDA’06. pp. 509-516. 2006

[11] Martínez, Francisco J. et al. “Minería de datos en series temporales para la búsqueda de conocimiento oculto en históricos de procesos industriales”. TAMIDA 2005, Granada, España.

[12] Moerchen, F. “Algorithms for Time Series Knowledge Mining”. Research Track Poster. KDD August 20–23, 2006, Philadelphia, Pennsylvania, USA.

[13] Panagiotis A, & D’ Avila A. “Applied Temporal Rule Mining to time series”. 2005

[14] Z. Dong, H. Li, Z. Shi, “An Efficient Algorithm for Mining Inter-transaction Association Rules in Multiple Time Series”. Journal of Computer Science. 2004, 31(3). pp. 108-111

Agradecimientos

Los autores agradecen al Ministerio de Educación y Ciencia de España a través de la Dirección General de Investigación la financiación de los proyectos DPI2006-03060, DPI2006-14784, DPI-2006-02454 y DPI2007-61090. También a la Unión Europea bajo el *Research Fund for Coal and Steel* (RFCS) la financiación del proyecto con referencia RFS-PR-06035. También, queremos agradecer la ayuda recibida a través del 3º Plan Riojano de I+D+i del Gobierno de La Rioja.

Correspondencia (Para más información contacte con):

Dr. Francisco Javier Martínez de Pisón Ascacíbar
Grupo EDMANS (www.mineriadatos.com)
Área de Proyectos de Ingeniería. Departamento de Ingeniería Mecánica
Edificio Departamental. ETSII de Logroño
C/ Luís de Ulloa, 20, 26004 Logroño (España).
Phone: +34 941 299 232 Fax: + 34 941 299 794
E-mail: fjmartin@unririoja.es
URL: http://www.mineriadatos.com