8. Minería de Datos en los Sistemas de Telegestión

Uno de los principales problemas que se presentan con la implantación de los sistemas de telegestión es el manejo adecuado de los datos obtenidos en las lecturas de forma que las aplicaciones que se desarrollen o utilicen estos datos sean lo más eficiente posible.

Además de la implantación de un modelo adecuado de sistema de gestión y almacenamiento de datos, se deben aplicar las técnicas más adecuadas para obtener el máximo beneficio de la información. En este capítulo realizaremos una pequeña introducción indicando el procedimiento de trabajo a la hora de realizar un análisis y algunas de las técnicas de minería de datos utilizadas en la ingeniería eléctrica. Por último, analizaremos una aplicación que utiliza herramientas de minería de datos para la solución del problema.

8.1. Introducción a la Minería de Datos

La minería de datos (DM, Data Mining) engloba un conjunto de técnicas que permiten la extracción no trivial de información que reside de manera implícita en los datos.

Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación.

La minería de datos está involucrada en lo que se conoce como Knowledge Discovery in Database (KDD). Este proceso consta de los siguientes pasos generales [42]:

1. **Selección** del conjunto de datos e identificación de las variables, incluyendo este punto la aplicación de las técnicas de muestreo necesarias.

2. **Análisis** de las propiedades de los datos, así como la deducción de la distribución de los datos, búsqueda de simetría y normalidad y

análisis de las correlaciones

3. **Limpieza** de los datos, mediante la búsqueda de valores atípicos (outliers) y la detección de ausencia de datos y eliminación de datos erróneos.

4. **Transformación** del conjunto de datos de entrada, con el objeto de prepararlo para aplicar la técnica de minería de datos. En este punto, se aplicarán técnicas de reducción o aumento de la dimensión, discretización o escalado.

5. **Selección y aplicación** de la técnica de minería de datos para la obtención de un modelo en el que están implícitos los patrones de comportamiento observados.

   Se distingue entre técnicas predictivas (regresión y series temporales, análisis discriminante, métodos bayesianos, algoritmos genéticos, árboles de decisión y redes neuronales) en las que existe un conocimiento teórico previo y técnicas descriptivas (clustering y segmentación, escalonamiento, reglas de asociación y dependencia, análisis exploratorio, reducción de la dimensión) en las que no se asigna un papel determinado a las variables.

6. **Interpretación y evaluación** de datos, una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias, para lo que se utilizarán intervalos de confianza, bootstrap, análisis ROC y evaluación de modelos.

Para poder aplicar en cada momento la técnica adecuada es imprescindible estar familiarizados con estas herramientas que nos pueden permitir realizar análisis y mejoras en nuestros procesos.

Una gran cantidad de estas técnicas descritas se han utilizado en la ingeniería eléctrica para el tratamiento de la información.

Entre otros, podemos indicar la utilización de técnicas de clustering para el análisis de los cambios de cargas en los transformadores, para la detección de condiciones anormales.

Otros ejemplos pueden ser el uso de regresiones dinámicas, redes neuronales y algoritmos genéticos como técnicas predictivas de consumo eléctrico o predicción de la generación eólica y para la detección de fallos y control del sistema se han desarrollado técnicas basadas en árboles de decisión.

En el siguiente apartado, analizaremos una propuesta basada en las técnicas de clustering, realizando un análisis de series temporales y en la que hay una caracterización previa de los datos.

8.2. Análisis de Perfiles de Consumo de los Clientes basado en Clustering

Hasta el momento el análisis de perfiles de consumo para los clientes se ha realizado en base a campañas de medida muy localizadas y con un tiempo de ejecución determinado. Con la introducción de los sistemas de telegestión y la obtención de las curvas de consumo mensuales, se plantea la necesidad de desarrollar herramientas que permitan la caracterización de los clientes a partir de los perfiles de carga.

Para este análisis, [43] propone la utilización de un esquema de clustering, para la caracterización de grupos de clientes con un perfil similar de comportamiento. Con anterioridad a la caracterización, se pretende también utilizar los datos asociados a los perfiles, que suele enriquecer el conocimiento a priori de los clientes. Además, para cada cliente el perfil de carga se segmenta según el tipo de día.

Se realizaron pruebas mezclando dos algoritmos y dos medidas de distancia en el esquema de clustering propuesto.

Los algoritmos que se han utilizado en este estudio son:

- Algoritmo K-Means, en el que para cada subconjunto de datos con características comunes se toma un valor del dato (centroide)

- Algoritmo TS-Part, desarrollado para el estudio y en el que no se hace la división en subconjuntos, sino que el algoritmo es el que realiza subconjuntos de datos hasta que se alcanza convergencia.

Como medias de distancia se analizaron dos propuestas:

- Utilizando la distancia euclídea para la formación de los cluster.

- Utilizando la distancia obtenida con el algoritmo DTW (Dinamic Time Warping) que es más apropiada en el estudio de series temporales como las que estamos tratando.

Aplicando los cuatro modelos y separando los datos en días laborables, sábados y domingos (se podría ampliar e incluir las festividades) se obtiene que el modelo que presenta mejores propiedades en cuanto a los criterios de evaluación es el algoritmo desarrollado Ts-Part con las distancias obtenidas por el algoritmo DTW, según podemos ver en el cuadro 6 para los días laborables:

clustering algorithm	*distance measure*	*# of clusters*	*MIA*	*CDI*
K-Means	Euclidean	5	37.24	2.82
TS-Part	Euclidean	5	23.86	1.58
K-Means	DTW	5	13.85	0.42
TS-Part	DTW	5	**12.36**	**0.13**

Cuadro 6: Resultado para días laborables

Los parámetros de evaluación son utilizados normalmente como criterio de validación para los procesos de clustering y se definen como:

- Mean Index Adecuacy (MIA) que mide la solidez de una solución como el promedio de las distancias entre cada objeto del grupo y el centro de gravedad

- Clustering Dispersión Indicator (CDI) que indica el grado de separación de cluster que es directamente proporcional la media de la distancia de los objetos del mismo cluster e inversamente proporcional a la distancias entre los centroides de los cluster.

En estos casos un menor valor se corresponde con una mayor calidad en la

solución.

Este tipo de estudios y desarrollos se verán mejorados y ampliados a medida que la disponibilidad de los datos de estudio sean mayores, sobre todo en el caso de los comercializadores para la realización tanto de ofertas adecuadas a cada tipo de cliente, como a la predicción de la demanda para ajustar sus adquisiciones de energía en los mercados.