

Coeficiente de Curtosis y Red Neuronal Recurrente de Hopfield para Clasificación de Consumidores Eléctricos

López, J. J.; Muñoz, F. ; Rodríguez, A.; Ruiz, J.E. Medina, M.; Muñoz, J.

Departamento de Ingeniería Eléctrica
Universidad de Málaga
Campus El Ejido– Málaga, 29013 Málaga (España)
Tel.:+34 952 131306, fax:+34 952 131091, e-mail: jjlopez@uma.es; fimartin@uma.es

Resumen. La clasificación de consumidores está encaminada a facilitar la información (curvas de carga) a las Compañías Eléctricas de forma estructurada y organizada (grupos) para que éstas puedan ofertar distintos tipos de tarifas. Para la realización del proceso de clasificación se emplean diferentes algoritmos como K-means, Follow the Leader SOM o red neuronal recurrente de Hopfield, siendo el objetivo fundamental en todos ellos encontrar grupos de consumidores con mínima varianza y la máxima disimilitud entre ellos. Con el objeto de conseguir la máxima calidad posible de dichos grupos es preciso detectar aquellos consumidores atípicos (específicos y errores de medida). Se presenta en esta comunicación la aplicación del coeficiente de curtosis, junto con la distancia de Mahalanobis, para la selección de dichos consumidores, empleándose la red neuronal recurrente de Hopfield, así como distintos índices de validación relativos (Dabies Bouldin, Calinski y WCBCR) para la realización de la clasificación. Por último se presentan los resultados obtenidos en un conjunto de 230 clientes (residenciales, administrativos e industriales).

Palabras llave

Clasificación, Curtosis, Mahalanobis, Hopfield, Calinski

1. Introducción

En los mercados eléctricos liberalizados, los consumidores juegan un papel fundamental en la estrategia de las Compañías, estando éstas interesadas en la captación del mayor número posible de clientes, ofreciéndole para ello ofertas atractivas, normalmente en forma de tarifas eléctricas.

Para poder realizar dichas ofertas, es fundamental tener conocimiento de como se encuentra estructurado el mercado, es decir, qué tipo de clientes nos encontraremos y como podremos agruparlos, al objeto de elaborar un grupo reducido de ofertas, y no tantas como consumidores. En la actualidad se emplean algoritmos basados en técnicas estadísticas como K-means [1,4], Follow the Leader modificado [4], algoritmos jerárquicos [4], redes neuronales competitivas como el Self - Organized Map (SOM) [5] o las redes neuronales recurrentes como la de Hopfield [1]. Pero es indispensable y fundamental la detección de los clientes atípicos, para

obtener la mejor calidad posible (mínima varianza) en los grupos obtenidos. En esta comunicación se ha utilizado las proyecciones de máxima y mínima curtosis junto a la distancia de Mahalanobis [2], para la detección de dichos clientes atípicos, aplicándolo a un conjunto de 230 clientes (residenciales, administrativos e industriales).

El artículo está organizado de la siguiente forma. En la sección 2 se describirá el proceso para la detección de los clientes anómalos. En la sección 3 se describirá el proceso de caracterización y agrupamiento. En la sección 4 se presentan los resultados obtenidos. Finalmente en 5 se exponen las conclusiones más relevantes.

2. Detección de clientes atípicos

A. Introducción.

Si partimos de un conjunto de datos X podemos definir un atípico como un punto que se encuentra lejos del centro de los datos. Si llamamos \bar{x} al vector de medias y S_x a la matriz de covarianzas y utilizando la distancia euclídea, una observación x_i será atípica con esta métrica si:

$$d_E(x_i, \bar{x}) = \sqrt{(x_i - \bar{x})(x_i - \bar{x})} \quad (1)$$

es grande, pudiéndose utilizar un histograma para detectar si existen puntos muchos más alejado que los demás. El problema es que la distancia euclídea no tiene en cuenta la estructura de correlación de los datos, y una posibilidad mejor es estandarizar los datos de forma multivariante, de la siguiente forma:

$$Y = S_x^{-\frac{1}{2}}(x - \bar{x}) \quad (2)$$

La distancia euclídea al cuadrado entre una observación, y_i , y su media, cero, será:

$$d_E^2(y_i, 0) = y_i y_i = (x_i - \bar{x}) S_x^{-1} (x_i - \bar{x}) = d_M^2(x_i, \bar{x}) \quad (3)$$

que corresponde a la distancia de Mahalanobis entre las variables originales, pudiéndose utilizar dicha distancia para la detección de datos atípicos [2].

B. Identificación de datos atípicos

El procedimiento para detectar grupos de elementos atípicos es eliminar de la muestra todos los puntos sospechosos de ser atípicos, calcular el vector de medias y la matriz de covarianzas sin distorsiones, y posteriormente aplicar la distancia de Mahalanobis y seleccionar como datos atípicos aquellos que se encuentren muy alejados del centro de los datos.

Otra característica importante del conjunto de datos es su homogeneidad. Para su estudio univariante se puede utilizar el coeficiente de homogeneidad H_j , el cual es siempre mayor o igual a cero. Puede escribirse como:

$$H_j = \frac{1}{n} \sum_{i=1}^n \frac{(x_{ij} - \bar{x}_j)^4}{s_j^4} - 1 = K_j - 1 \quad (4)$$

Siendo x_{ij} , s_j y K_j , los valores de una variable en una observación, desviación típica y coeficiente de curtosis respectivamente.

De la expresión de H_j , podemos observar que los datos no serán homogéneos (presencia de atípicos), cuando tenemos un alto valor de K_j (pocos atípicos muy alejados de la media) y cuando K_j es próximo a uno (dos distribuciones de datos muy alejados entre ellas).

Una solución propuesta por Peña y Prieto [2], es proyectar los datos sobre ciertas direcciones específicas, escogidas de manera que tengan alta probabilidad de mostrar los atípicos, utilizando para ello el coeficiente de curtosis ya que una maximización o minimización del mismo, en dichas direcciones, es indicativo de la presencia de datos atípicos.

Dada la muestra de datos multivariantes (x_1, \dots, x_n) de p variables, el proceso se realiza de la siguiente forma:

1. Sea $z_i = S_x^{-\frac{1}{2}}(x_i - \bar{x})$ la estandarización de los datos de forma multivariante con media cero y matriz de covarianzas identidad. Tomar $j=1$ y $z_i^{(1)} = z_i$.
2. Calcular la dirección d_j con norma unidad que maximiza el coeficiente de curtosis univariante de los datos proyectados. Llamemos $y_i^{(j)} = d_j' z_i^{(j)}$, a la proyección de la observación z_i sobre la dirección d_j .
3. Proyectar los datos sobre un espacio de dimensión $p-j$ definido como el espacio ortogonal a la dirección d_j .
4. Repetir (2) y (3) hasta obtener las p direcciones d_1, \dots, d_p .
5. Repetir (2) y (3) pero ahora minimizando la curtosis, para obtener otras p direcciones d_{p+1}, \dots, d_{2p} .

6. Considerar como datos sospechosos aquellos puntos que en algunas de estas $2p$ direcciones verifican:

$$\frac{|y_i^{(j)} - med(y^{(j)})|}{Meda(y^{(j)})} > 5 \quad (5)$$

Siendo $med(y^{(j)})$ la mediana de las observaciones y $Meda(y^{(j)})$ la mediana de las desviaciones absolutas.

7. Se eliminan los datos detectados como sospechosos y se vuelve al paso (1) para analizar los datos restantes. La estandarización multivariante se realizará con la nueva media y matriz de covarianzas de los datos restantes. Los pasos 2 a 6 se repiten hasta que no se detecten más datos sospechosos.
8. Una vez que la muestra no contenga más valores sospechosos se calcula el vector de medias, \bar{x}_R , y la matriz de covarianzas, S_R , de los datos no sospechosos, y la distancia de Mahalanobis para los sospechosos como:

$$d_R^2(x_i, \bar{x}_R) = (x_i - \bar{x}_R) S_R^{-1} (x_i - \bar{x}_R)' \quad (6)$$

9. Se considerarán atípicos aquellos puntos cuya distancia sea superior a $k + 3\sqrt{k}$, siendo k el valor promedio de la distancia de Mahalanobis.

Un esquema de aplicación del procedimiento explicado se puede observar en la figura 1, siendo MD la matriz de datos originales de los consumidores eléctricos, obteniéndose un conjunto de datos típicos (MDTIP), que se utilizarán posteriormente para realizar el agrupamiento y un conjunto de datos atípicos (MDATIP), que son los clientes atípicos, y que requerirán un estudio particular o también podemos realizar una clasificación de los mismos.

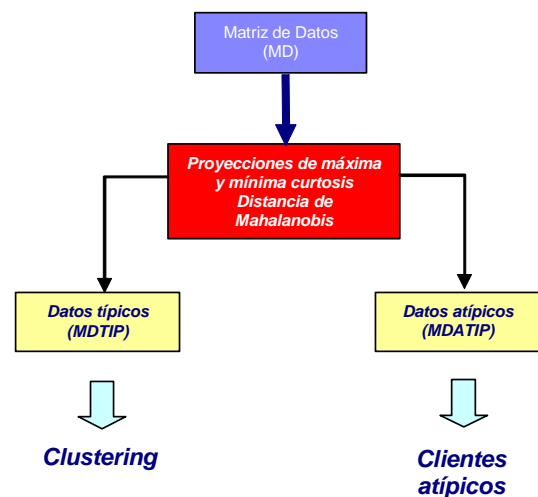


Fig. 1: Esquema de detección de clientes anómalos

3. Procedimiento de caracterización y clasificación.

Aplicado el procedimiento anterior, a continuación se realizará una etapa de caracterización y clasificación con la matriz MDTIP/MDATIP. Este proceso se muestra en la figura 2.

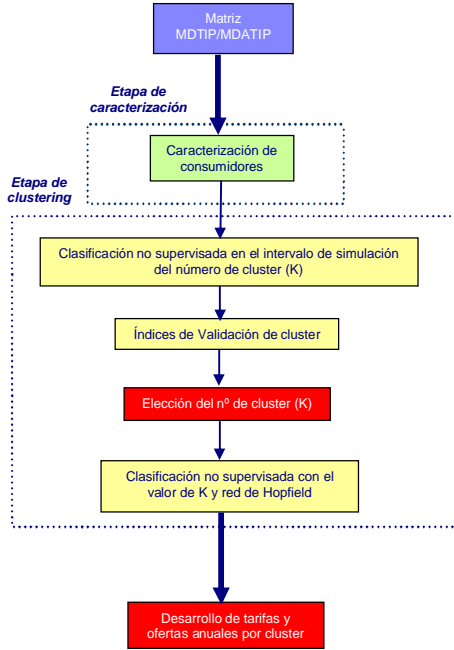


Fig.2: Esquema general de clasificación

A. Etapa de Caracterización de datos

Las técnicas utilizadas para la caracterización de datos son el perfil de carga horaria (N) [1,4], factor de forma (FF) [1,4], componentes principales (CP) [1], análisis armónico (F) [1,4] y análisis multirresolución (W) [1].

1. Perfil de carga horaria

Consiste en la representación del consumidor eléctrico a partir de su curva de carga, ya sea utilizando todos los valores de ésta o extrayendo un conjunto reducido de valores de potencia. Si partimos de un conjunto de Q clientes, el conjunto de perfiles de carga horaria, se puede representar por la siguiente matriz P :

$$P(i, j) = p_j^i \quad \forall i=1,] , Q ; j=1, \setminus , N \quad (7)$$

Para el caso del perfil de carga horaria $N=24$, representando p_j^i el valor de la potencia medida en una determinada hora j y para un determinado cliente i .

2. Factor de forma

Estos índices caracterizan al consumidor en el dominio del tiempo, y sus valores se encuentran en el intervalo $[0,1]$. Se emplean nueve índices para representar a cada consumidor y se pueden consultar en [4].

Para un conjunto de clientes Q , los factores de forma (ff) se pueden representar por la siguiente matriz FF :

$$FF(i, j) = ff_j^i \quad \forall i=1,] , Q \quad y \quad \forall j=1,] , 9 \quad (8)$$

3. Componentes principales

Consiste en representar a un conjunto de n observaciones con N variables, en otro conjunto de información con un menor número de variables r .

Si queremos representar los nuevos datos en el espacio $r < N$, debemos seleccionar la r primeras columnas de V (máxima varianza). Los nuevos datos en el espacio de r dimensiones será:

$$CP = Q \cdot V_r \quad (9)$$

donde CP son los datos en el espacio dimensional r , Q es la matriz de datos de clientes y V_r es la matriz con las r componentes principales seleccionadas.

4. Análisis armónico

El análisis armónico [1,4] consiste en la extracción de una serie de índices o variables de la respuesta en frecuencia de la curva de carga, para lo cual se emplea la Transformada Discreta de Fourier (DTF). Para un cliente i , el conjunto de n variables, se puede representar por:

$$f_n^i = \{A_j^i, j=1\} \cup \{f_j^i, j=2,] , n\} \quad \forall n=h \quad (10)$$

Y para un conjunto de Q clientes, se puede representar por la siguiente matriz F :

$$\begin{aligned} F(i, j) &= A_j^i \quad \forall i=1,] , Q ; j=1 \\ F(i, j) &= A_j^i \quad \forall i=1,] , Q ; j=2 \\ F(i, 3j-6) &= \chi_j^i \quad \forall i=1,] , Q ; 3 \leq j \leq h \\ F(i, 3j-5) &= \chi_j^i \quad \forall i=1,] , Q ; 3 \leq j \leq h \end{aligned} \quad (11)$$

5. Análisis multirresolución

El análisis multirresolución [1] consiste en el proceso de extracción de índices o variables del espectro tiempo-frecuencia de la curva de carga, para lo cual se emplea la Transformada Discreta de Wavelet (DTW).

Si lo generalizamos para un conjunto de Q clientes, se puede representar por la siguiente matriz W :

$$W(i, k) = a_j^k(i) \quad \forall i=1,] , Q ; k=1,] , p \quad (12)$$

donde p es el número de muestras de la señal aproximada $a_j(i)$.

B. Etapa de clasificación

En esta etapa, previa caracterización de los datos, procederemos a la obtención del número de cluster a realizar. Para ello utilizaremos la red neuronal recurrente de Hopfield [1] y los índices de validación relativos de Calinski [6], Dabies Bouldin [4] y WCBCR [7].

1 Red neuronal recurrente de Hopfield (H-ANN)

De forma general una H-ANN [B] es un grafo completo $G=(V,A)$, cuyos vértices (V) representan a N unidades de proceso conectados entre sí mediante aristas o arcos (A).

Asociado a cada eje o conexión $(i,j) \in A$ hay un peso sináptico w_{ij} , que es un número real que representa la fuerza de interconexión entre las neuronas i,j . En una formulación general los pesos sinápticos se pueden representar por la siguiente matriz W :

$$W(i,j) = w_{i,j} \quad \forall i,j = 1, \dots, N \quad (13)$$

Esta matriz se considerará simétrica ($w_{ij}=w_{ji}$) y con diagonal nula ($w_{ii}=0$).

Una representación de la red se puede observar en la figura 3.

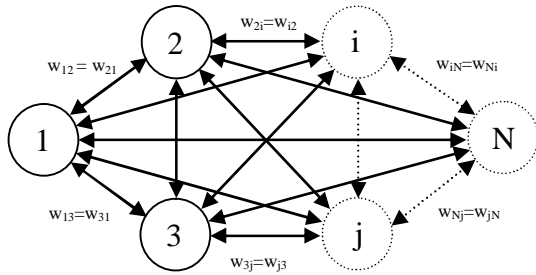


Figura 3: Esquema general de una H-ANN

Para una H-ANN el elemento básico de computación es una unidad de proceso bipolar. Su función matemática está definida en el conjunto $\{-1,1\}^N$, y puede describirse por la siguiente expresión:

$$f(x_1, x_2, \dots, x_N) = \begin{cases} 1 & \text{si } x_1 w_1 + \dots + x_N w_N \geq \theta \\ -1 & \text{si } x_1 w_1 + \dots + x_N w_N < \theta \end{cases} \quad (14)$$

donde x_N representa el estado de las N neuronas de la red.

Además para el caso de redes recurrentes multievaluadas, el estado x_i de cada una de las N neuronas vendrá caracterizado por su salida S_i que, en una formulación general, podrá tomar cualquier valor en un conjunto que llamaremos M . Dicho conjunto puede adquirir valores en \mathbb{R} , o un conjunto no numérico (cualitativo).

Como la red evolucionará en el tiempo, el estado de la neurona i -ésima vendrá caracterizada por el valor de su salida en ese instante $s_i(k)$. Llamaremos *vector de estado* de la red en el instante k al vector $S(k) = (s_1(k), s_2(k), \dots, s_N(k))$, el cual describe el estado de las neuronas, que conforman la red, en ese instante k . Asociado a ese vector de estado existe una función de energía (E).

$$E(k) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N w_{i,j} f(s_i, s_j) + \sum_{i=1}^N \theta_i s_i \quad (15)$$

donde la función f será una aplicación de $M \times M \rightarrow \mathbb{R}$ y medirá la analogía o similitud entre las salidas de las neuronas i -ésima y j -ésima.

La evolución dinámica de la red se realizará mediante un procedimiento asíncrono y secuencial. La H-ANN comenzará con un vector de estado inicial $S(0)$, y se irán calculando los estados de todas las neuronas para los instantes $1, 2, \dots, k, k+1$, según la siguiente regla:

$$s_i(k+1) = \begin{cases} 1 & \text{si } \sum_{j=1}^N w_{ij} s_j(k) \geq \theta \\ -1 & \text{si } \sum_{j=1}^N w_{ij} s_j(k) < \theta \end{cases} \quad (16)$$

Para cada instante $1, 2, \dots, k, k+1$, llevará asociado un vector de estado $S(1), S(2), \dots, S(k), S(k+1)$, cuya unión representará el *espacio de estados* de la red. La evolución de la red se hará de tal forma que en cada instante disminuya lo máximo posible el valor de la energía.

El objetivo es minimizar la función de energía (E) para el intervalo considerado, garantizándonos así un mínimo local que es la solución final del vector de estado.

2 Indices de validación relativos

Para su desarrollo se han utilizado las siguientes distancias y medidas de dispersión de matrices:

a) Distancia Euclídea entre dos curvas de carga:

$$d(x^{(i)}, x^{(l)}) = \left\{ \sum_{j=1}^v (x_j^i - x_j^l)^2 \right\}^{\frac{1}{2}} \quad ; \quad \forall (x^{(i)}, x^{(l)}) \in X \quad (17)$$

b) Distancia entre el centroide $C^{(k)}$ y el conjunto de curvas de carga $X^{(k)}$.

$$d(C^{(k)}, X^{(k)}) = \sqrt{\frac{1}{x^{(k)}} \sum_{i=1}^{x^{(k)}} d^2(C^{(k)}, x^{(i)})} \quad (18)$$

donde $x^{(k)}$ representa el número de curvas de carga que pertenecen al subconjunto $X^{(k)}$.

c) Distancia media del conjunto de curvas que pertenecen a $X^{(k)}$.

$$\hat{d}(X^{(k)}) = \sqrt{\frac{1}{2x^{(k)}} \sum_{i=1}^{x^{(k)}} d^2(x^{(i)}, X^{(k)})} \quad (19)$$

e) Medida de dispersión entre los cluster formados.

$$S_B = \sum_{k,i=1}^K x^{(k)} \cdot (C^{(i)} - C^{(X)}) \cdot (C^{(i)} - C^{(X)})^t \quad (20)$$

f) Medida de dispersión intra-cluster.

$$S_W = \sum_{i=1}^K \sum_{x^{(i)} \in X^{(i)}} (x^{(i)} - C^{(i)}) \cdot (x^{(i)} - C^{(i)})^t \quad (21)$$

g) Medida de dispersión total

$$S_T = S_B + S_W = \sum_{x^{(i)} \in X} (x^{(i)} - C^{(X)}) \cdot (x^{(i)} - C^{(X)})^t \quad (22)$$

Los índices de validación relativos utilizados son:

- (1) *Índice de Calinski (CH)* [6], que relaciona la dispersión entre los grupos formados (S_B) y la dispersión interna de los mismos (S_W). Dicho índice da una idea de la separación existente entre los grupos realizados y la compactación de los mismos.

$$CH = \frac{S_B}{S_W} \cdot \frac{Q - K}{K - 1} \quad (23)$$

El número de cluster a realizar es el que maximiza dicho índice.

- 3) *Índice de Dabies Bouldin (DB)* [4], que relaciona la media de las distancias de cada grupo con su grupo más próximo.

$$DB = \frac{1}{K} \sum_{i,j=1}^K \max_{i \neq j} \left\{ \frac{\hat{d}(X^{(i)}) + \hat{d}(X^{(j)})}{d(C^{(i)}, C^{(j)})} \right\} \quad (24)$$

El número de cluster a realizar es el que minimiza dicho índice.

- 4) *Índice de ratio de la suma de cuadrados entre cluster (WCBCR)* [7], que relaciona las distancias al cuadrado de cada curva de un grupo y su centroide y la suma al cuadrado de los centroides de los grupos formados.

$$WCBCR = \frac{\sum_{k=1}^K \sum_{x^{(i)} \in X^{(k)}} d^2(C^{(k)}, X^{(k)})}{\sum_{i < j}^K d^2(C^{(i)}, C^{(j)})} \quad (25)$$

Este índice nos indica que a menor valor mejor es la clasificación realizada.

4. Simulaciones y resultados.

Se ha utilizado un conjunto de datos de 230 líneas de Media Tensión repartidas en 19 líneas que alimentan a consumidores de tipo Administrativo (A), 175 de tipo Industrial (I) y 36 de tipo Residencial (R). Dichos datos han sido facilitados por una Compañía Eléctrica Española, correspondiendo a curvas de cargas horarias.

Tras la aplicación del procedimiento descrito en la sección 2, los resultados se recogen en la tabla I.

TABLA I
MATRICES DE DATOS

| | A | I | R | Total |
|--------|----|-----|----|-------|
| MD | 19 | 175 | 36 | 230 |
| MDTIP | 16 | 168 | 35 | 219 |
| MDATIP | 3 | 7 | 1 | 11 |

A. Aplicación de índices de validación relativos

El estudio ha sido realizado para un intervalo de simulación de 5-30 grupos, estableciéndose diversas comparativas entre los índices de validación relativos

utilizados con todas y cada una de las herramientas de caracterización de datos.

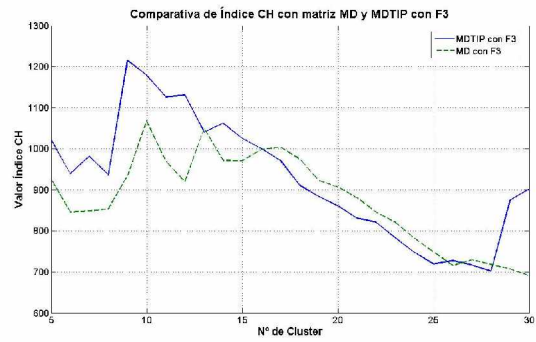


Figura 4: Comparativa de Calinski con F3

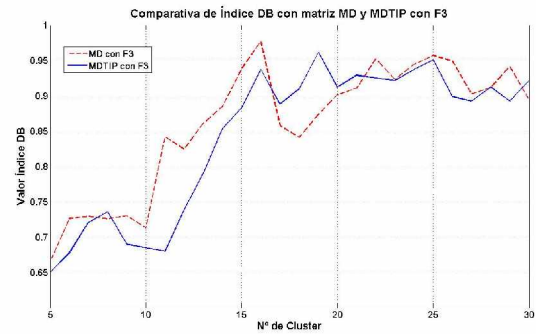


Figura 5: Comparativa de Dabies Bouldin con F3

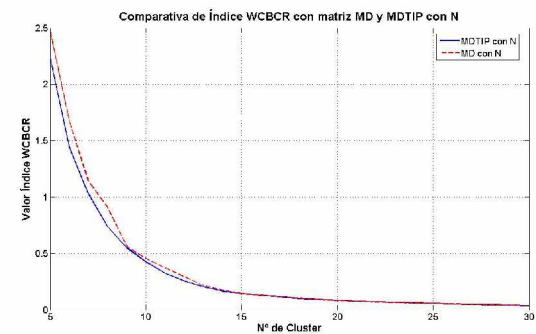


Figura 6: Comparativa de WCBCR con N

En las figuras 4, 5 y 6 se han representado las comparativas de los índices de validación, con aquellas caracterizaciones de datos que mejores resultados han dado, tanto con la matriz MD (sin aplicar la detección de datos atípicos) como con la matriz MDTIP (aplicando la detección de datos atípicos).

En la figura 4 (índice de Calinski), se observa como se consigue un aumento del valor de este índice aplicando el procedimiento de detección de consumidores atípicos (salvo entre 16 y 25 cluster) además de conseguir el máximo absoluto.

En la figura 5 (índice de Dabies Bouldin), se una disminución del valor de este índice aplicando el procedimiento de detección de consumidores atípicos (salvo entre 16 y 21 cluster) además de conseguir el mínimo absoluto.

Un resumen de los datos conseguidos con estos índices se muestra en la Tabla II

TABLA II
RESULTADOS CON CH Y DB Y ANÁLISIS ARMÓNICO 3 (F3)

| | MD | | MDTIP | |
|----------------|------|--------|-------|------|
| | CH | DB | CH | DB |
| Valor | 1068 | 0.6653 | 1215 | 0.65 |
| Nº cluster (K) | 10 | 5 | 9 | 5 |

Análogamente se puede observar como, en la figura 6, los resultados obtenidos con el índice WCBCR disminuyen en todo el intervalo de simulación aplicando el método de detección de clientes atípicos.

B. Agrupación de clientes con las matrices MDTIP y MDATIP.

Con la matriz MDTIP y según los resultados arrojados por el índice de Calinski (tabla II), se ha elaborado un clustering con $K=9$ y F3.

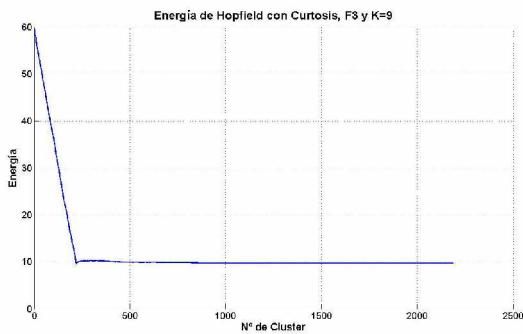


Fig. 7: Valor de la energía de la red de Hopfield

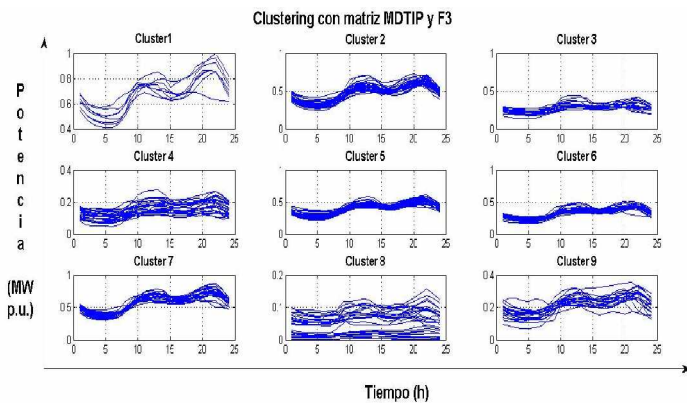


Fig. 8: Clasificación obtenida con $K=9$, F3 y MDTIP

En la figura 7 se ha representado la variación de energía de la red de Hopfield en el período de simulación. Se observa como ésta disminuye y se estabiliza (250 iteraciones), garantizándonos de esta forma una solución única. En la figura 8 queda reflejada la solución obtenida.

Análogamente hemos procedido con la matriz MDATIP, aplicando el índice de Calinski, para obtener el número de cluster, y a continuación elaborar la clasificación con la red neuronal recurrente de Hopfield. Estos resultados se han representado en las figuras 9 y 10

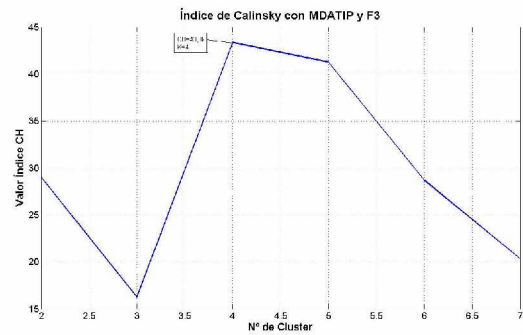


Fig. 9: Índice de Calinski con MDATIP y F3

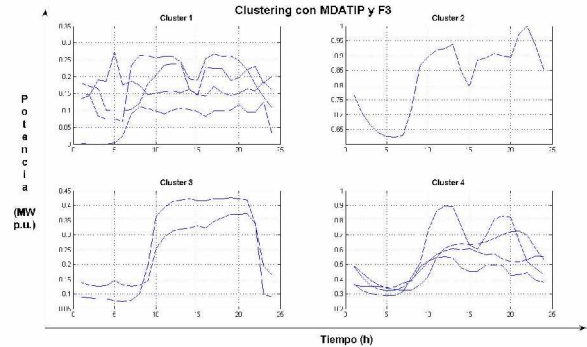


Fig. 10: Clasificación obtenida con $K=4$, MDATIP y F3

5. Conclusiones

Aplicando el método descrito a la matriz de datos MD, conseguimos:

- 1) Identificar los posibles clientes atípicos (MDATIP) del conjunto de datos iniciales.
- 2) Aumentar la calidad de los cluster obtenidos con la matriz MDTIP, según los resultados obtenidos con los índices de validación relativos utilizados.

6. Bibliografía

- [1] J.J. López, J.A. Aguado, F. Martín, F. Muñoz, A. Rodríguez and J.E. Ruiz, "Electric customer classification using Hopfield recurrent ANN", 5th International Conference on The European Electricity Market, Lisbon May 2008.
- [2] Peña D. and Rodríguez, J. "Cluster identification using projections". Journal of American Statistical Association, 96, 1433-1445.
- [3] R., Xu, and D. Wunsch, "Survey of clustering algorithms", IEEE Transactions on Neural Networks, vol 16 n° 3, May 2005.
- [4] G. Chicco, R. Napoli and F. Piglion, "Comparisons among clustering techniques for electricity customer classification", IEEE Trans. Power Systems, vol. 21, n° 2 may 2006.
- [5] Valero S., Ortiz M., Senabre, S., Gabaldón, A. And F. García, "Classification, Filtering and Identification of electrical customer load patterns through the use of Self-Organizing Maps", IEEE Transactions on Power Systems, vol. 21, n° 4 November 2006.
- [6] Calinski, R. and J. Harabasz, "A dendrite method for cluster analysis". Communications in Statistics 3, 1-27, 1974.
- [7] G.J. Tsekouras, N.D. Hatzigryriou and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers", IEEE Transactions on Power Systems, vol. 22, n° 3, August 2007.