Caracterización de Diferentes Series de Tiempo con Escalamiento Multidimensional

[1]C. Bustillo-Hernández, V. Rivera-Mancera, E. Bautista Thompson, [2]J. Figueroa-Nazuno

Centro de Investigación en Computación
Instituto Politécnico Nacional
Unidad Profesional "Adolfo López Mateos"
Zacatenco-07738 D.F, México.
[1]chbustillo004@cic.ipn.mx, [2]jfn@cic.ipn.mx
Paper received on 09/08/08, accepted on 08/09/08.

Resumen. En este trabajo se presenta un estudio experimental de caracterización de 30 series de tiempo de diferentes orígenes (natural y artificial), a las cuales se les calculó un conjunto de propiedades representativas de origen computacional, topológico, estadístico, espacial, y temporal. Se les aplicó la técnica de análisis multivariada Escalamiento Multidimensional; para identificar series de tiempo con propiedades similares que son indicativas de su dificultad de predicción. Los resultados indican que es posible representar en un espacio geométrico de pocas dimensiones, las proximidades existentes entre un conjunto de series de tiempo en base a métricas de similitud.

1 Introducción.

El escalamiento multidimensional (Multidimensional Scaling MDS), permite obtener información cuantitativa y cualitativa de las posibles relaciones de similitud entre objetos (en este caso de series de tiempo), mediante el escalamiento multidimensional métrico y el escalamiento multidimensional no-métrico respectivamente. El MDS es una técnica de representación espacial que trata de visualizar sobre un mapa, un conjunto de propiedades cuya posición relativa se desea analizar. El propósito del MDS es transformar las propiedades de las series de tiempo y llevarlas a una matriz de distancias susceptible de ser representada en un espacio multidimensional. El MDS está basado en la comparación de objetos, de forma tal que si un individuo juzga a los objetos A y B como los más similares, es debido a que la técnica de MDS coloca a los objetos A y B en el gráfico a una distancia entre ellos que sea la menor respecto a cualquier otra distancia entre cualquier otro par de objetos distintos a A y B [1, 2].

Existen otras técnicas multivariadas, como son el análisis factorial y el análisis cluster, que persiguen objetivos muy similares al MDS pero difieren en una serie de aspectos, como por ejemplo: en el MDS no es necesario especificar cuáles son las variables a emplear en la comparación de objetos, algo que es fundamental en el análisis factorial y en el análisis cluster. Sin embargo, la utilización de alguna de es-

© M. G. Medina Barrera, J. F. Ramírez Cruz, J. H. Sossa Azuela. (Eds.)
Advances in Intelligent and Information Technologies.
Research in Computing Science 38, 2008, pp. 153-162.

tas técnicas no supone que no se pueda utilizar el escalamiento multidimensional, sino que esta última técnica puede servir como complemento a las otras técnicas multivariadas.

2 El modelo general de escalamiento multidimensional.

De modo general, podemos decir que el MDS toma como entrada una matriz de proximidades, $\Delta \in M_{nxn}$, donde n es el número de series de tiempo. Cada elemento δ_{ij} de Δ representa la proximidad entre las propiedades para las series de tiempo ST_i y ST_j donde $\delta_{ij} = (P_i - P_j)^2$, P_k es una propiedad de la serie de tiempo k-ésima.

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{bmatrix} \qquad \text{(Ec. 2.1)}$$

El MDS inicia con una matriz aleatoria $X \in M_{nxm}$ donde n, al igual que antes, es el número de series de tiempo y m es el número de dimensiones. En nuestro caso tomaremos $m=2$. Cada valor x_{ij} representa la coordenada de la serie de tiempo (ST_i) en la dimensión j.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \qquad \text{(Ec. 2.2)}$$

A partir de esta matriz X se puede calcular la distancia existente entre dos series de tiempo cualesquiera i y j, simplemente aplicando la fórmula general de la distancia de Minkowski:

$$d_{ij} = \left[\sum_{t=1}^{n} \left(x_{it} - x_{jt} \right)^p \right]^{1/p} \qquad \text{(Ec. 2.3)}$$

donde p puede ser un valor entre 1 e infinito, en nuestro caso tomaremos $p = 2$. A partir de estas distancias podemos obtener una matriz de distancias que denominamos $D \in M_{nxn}$.

La solución proporcionada por el MDS debe ser de tal modo que haya la máxima correspondencia entre matriz de proximidades inicial Δ y la matriz de distancias

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \qquad \text{(Ec. 2.4)}$$

obtenidas D, la cuál se logra modificando la matriz X de forma iterativa.

$$X = \frac{BX}{2n} \qquad \text{(Ec. 2.5)}$$

En donde B tiene como elementos:

$$b_{ij} = \frac{-2\delta_{ij}}{d_{ij}} \quad si \quad i \neq j \qquad \text{(Ec. 2.6)}$$

$$b_{ii} = \sum\sum \frac{2\delta_{ik}}{d_{ik}} \quad si \quad i = j \qquad \text{(Ec. 2.7)}$$

$$b_{ij} = 0 \quad si \quad d_{ij} = 0 \qquad \text{(Ec. 2.8)}$$

3 Modelos de escalamiento multidimensional

Existen dos modelos básicos de MDS que son: el modelo de escalamiento métrico y el modelo de escalamiento no métrico. En el primero de ellos consideramos que los datos están medidos en escala de razón o en escala de intervalo y en el segundo consideramos que los datos están mediados en escala ordinal. No se ha desarrollado todavía ningún modelo para datos en escala nominal. Para el estudio de las series de tiempo sólo usaremos el modelo de escalamiento no métrico, ya que es éste el que permite formar clases o agrupaciones, a diferencia del escalamiento métrico que arroja información cuantitativa.

3.1 Modelo de Escalamiento No Métrico.

A diferencia del escalamiento métrico, el modelo de escalamiento no métrico no presupone una relación lineal entre las proximidades y las distancias, sino que establece una relación monótona creciente entre ambas, es decir, si $\delta_{ij} < \delta_{kl} \Rightarrow d_{ij} \leq d_{kl}$. Su desarrollo se debe a Shepard (1962) quién demostró que es posible obtener soluciones métricas asumiendo únicamente una relación ordinal entre proximidades y distancias. Posteriormente Kruskal (1964) mejoró el modelo [1, 2].

El procedimiento se basa en los siguientes apartados:

1) Obtención de una matriz $X \in M_{nxm}$ de coordenadas aleatorias, que nos da la distancia entre las series de tiempo.
2) Comparación de las proximidades con las distancias contenidas en X, obteniéndose las disparidades.
3) Cálculo del S-Stress (explicado más adelante en está sección).
4) Modificación de la matriz X mediante la ecuación 2.5 con el fin de minimizar el S-Stress.
5) Ir al paso 2 hasta que el S-Stress alcance el valor deseado

Tanto para el modelo métrico como para el modelo no métrico es necesario obtener un coeficiente que nos informe sobre la precisión del modelo. Sabemos que las distancias son una función de las similitudes, es decir: $f : \delta_{ij}(x) \rightarrow d_{ij}(x)$.

De esta forma se tiene que $d_{ij} = f(\delta_{ij})$. Esto no deja ningún margen de error. Sin embargo, en las proximidades empíricas es difícil que se dé la igualdad, por lo que generalmente ocurre que $d_{ij} \approx f(\delta_{ij})$, $f(\delta_{ij}) = a + b\delta_{ij}$ donde a y b son constantes que se deben de determinar. A las transformaciones de las proximidades por f se les denomina *disparidades*. A partir de aquí podemos definir el error cuadrático como:

$$e_{ij}^2 = (f(\delta_{ij}) - d_{ij})^2 \qquad \text{(Ec. 2.9)}$$

Como medida de la precisión del modelo utilizaremos el S-*Stress* definido como:

$$S-Stress = \sqrt{\frac{\sum_{i,j}\left(f(\delta_{ij})^2 - d_{ij}^2\right)^2}{\sum_{i,j}\left(d_{ij}^2\right)^2}} \qquad \text{(Ec. 2.10)}$$

4 Análisis Experimental de las Series de Tiempo

Se tomaron 29 series de tiempo que han sido reportadas en la literatura especializada en la evaluación de técnicas de predicción y modelado [3, 8] y que además corresponden a diversos orígenes (experimental o generadas por modelos matemáticos). La selección de las propiedades calculadas que caracterizan a las series de tiempo se realizo considerando que cuantifican características de tipo estadístico, topológico, computacional, espacial y temporal [3, 4, 7]. A continuación se da una descripción de las propiedades calculadas:

Exponente de Lyapunov. El exponente principal de Lyapunov, mide la evolución de trayectorias vecinas en el espacio fase. Mide la inestabilidad de la dinámica del sistema debido a cambios en sus condiciones iniciales [3, 6].

Exponente de Hurst. El exponente de Hurst permite determinar si el fenómeno representado por la serie de tiempo presenta correlaciones de largo alcance (memoria y persistencia de largo alcance) [3, 6].

Dimensión de Capacidad. La dimensión de capacidad es similar a la dimensión de Hausdorff y mide el grado de auto-similitud del sistema (comportamiento invariante ante cambios de escala espacial), permite cuantificar el grado de heterogeneidad de la señal a diferentes escalas [3, 6].

Dimensión de Correlación. La dimensión de correlación mide la cantidad de veces que la trayectoria del sistema pasa por una vecindad dada en el espacio fase, cuantifica la correlación espacial local entre puntos de la trayectoria en el espacio fase, sin tomar en cuenta el grado de correlación temporal [3, 5].

Frecuencia Dominante. Permite determinar si existe alguna frecuencia característica de la señal. Las series de carácter aleatorio poseen espectros amplios sin ninguna frecuencia dominante, lo mismo se aplica en cierto grado para las series caóticas. Las series periódicas y cuasi-periódicas poseen picos bien definidos [3, 5].

Entropía Espacio-Temporal. La entropía-temporal, cuantifica de forma global la no correlación de los datos mediante el análisis de recurrencia [3, 5].

Entropía de Shannon. La entropía de Shannon, es una medida de la cantidad de información que se obtiene al tomar una medida para especificar el estado del sistema [5, 6].

Porcentaje de Determinismo. Permite medir el grado de determinismo en el sistema, por medio del análisis de mapas de recurrencia [3, 5].

Porcentaje de Recurrencia. Permite medir el grado de recurrencia (periodicidad y estructura) entre los datos de la serie, que es indicativo de patrones repetitivos en la serie de tiempo, por medio del análisis de mapas de recurrencia [3, 5].

Reglas de Producción. La generación de gramáticas a partir de una serie de tiempo permite dar una medida de complejidad (computacional) en la cual a mayor número de reglas de producción necesarias para generar una gramática, mayor es la dificultad para la predicción o modelado de la serie de tiempo [3, 8].La Tabla 1 nos muestra las 30 series de tiempo y los valores calculados.

5 Método.

Una vez obtenidas las propiedades de las series de tiempo se aplicó el procedimiento de MDS, para cada una de las propiedades para el conjunto de 29 series de tiempo. Ver Tabla 1.

Tabla 1. Propiedades calculadas experimentalmente de las 29 series de tiempo

Serie de Tiempo	Exponente de Hurst	Dimension de Capacidad	Dimension de Correlación	Exponente de Lyapunov	Frecuencia Dominante	Numero de Reglas de Producción	Entropia (Shannon)	Entropia Espacio Temporal (%)	Recurrencia (%)	Determinismo (%)
Asci	-0.147	0.485	2.011	2.709	0.266	80	2.107	62	13.144	1.23
Brownian	0.537	0.971	1.103	2.043	0.422	88	2.740	0	11.775	66.26
Cantor	0.001	0.661	0.658	6.72E	0.000	72	0.000	64	2.193	0.00
Dji	0.312	0.986	2.053	1.26E	0.000	52	0.852	39	3.301	60.25
Downjones	0.563	0.867	1.035	0.144	0.000	65	6.435	2	96.853	96.86
Eeg	0.746	0.534	1.051	0.242	0.000	53	3.519	0	13.789	69.86
Eeg	0.204	0.890	1.87E	1.199	0.041	82	0.000	73	2.716	0.00
Elnino	0.447	0.971	1.837	2.322	0.000	63	3.52E	65	0.320	44.21
Henon	-0.033	0.970	0.991	2.301	0.460	69	0.020	51	9.435	0.64
Hiv DNA	0.001	0.953	5.017	0.000	0.000	5	1.000	0	0.004	31.33
Human DNA	0.562	0.983	1.037	0.322	0.000	51	7.280	0	24.915	98.24
Ikeda	-0.020	0.983	1.019	1.452	0.324	71	0.000	53	0.548	0.00
Kobe	0.016	0.904	1.053	1.102	0.220	79	5.744	76	57.047	44.37
Lase	0.086	0.610	2.09E	0.945	0.130	68	2.935	47	10.716	71.47
Log	-0.059	0.941	0.930	0.760	0.360	61	0.035	76	8.613	0.85
Lorenz	0.75E	0.965	1.02E	0.601	0.001	63	4.76E	57	32.864	92.07
Lovame	0.509	0.95E	1.027	1.069	0.000	77	5.574	47	0.547	69.79
Mag	0.076	0.983	1.02E	1.461	1.461	75	0.159	50	0.510	37.52
Ritsme	0.062	0.821	0.967	3.363	0.080	91	2.866	61	1.683	0.75
Primos	-0.010	0.044	3.55C	0.594	0.500	76	1.29E	80	9.705	0.41
Qpsk	0.002	0.679	0.923	0.92E	0.130	68	2.44E	0	8.148	97.76
Qpsk	-0.00E	0.69E	0.960	-1.363	0.447	75	4.232	0	16.133	36.75
Fossil	0.567	0.994	1.02E	1.049	0.017	72	2.45E	0	6.389	86.04
Sp500	0.015	0.65C	1.02E	3.56E	0.035	62	2.56E	67	20.868	4.3E
Sene	0.946	0.246	0.22E	0.517	-0.001	27	3.712	0	19.556	62.19
Ste	0.164	0.963	1.144	3.28E	0.01E	90	3.102	53	16.436	11.14
Tent	-0.243	0.971	1.02E	0.477	0.500	13	3.401	53	3.079	29.01
Vanderpol	0.487	0.98E	0.96E	1.86E	0.112	2E	4.96E	0	7.62E	66.3E
Weiers	0.002	0.983	2.06E	1.60E	0.000	73	0.000	79	0.000	0.00

El diagrama 1 muestra el algoritmo ALSCAL (Alternating Least Squares SCALing), que es un procedimiento de análisis repetitivo, donde en cada iteración se realiza un escalamiento óptimo y un modelo de estimación. Cada procedimiento es un ajuste con mínimos cuadrados, éstos se realizan varias veces hasta encontrar el criterio de convergencia, que en este caso fue de 0.001, o de acuerdo al número de iteraciones definidas. Los pasos que sigue el algoritmo son:

1.- Se usaron las propiedades de las series de tiempo, para construir una matriz de proximidades usando la ecuación $\delta_{ij} = (P_i - P_j)^2$.

2.- Se obtienen las matrices X y B, donde X es una matriz *nxm*, donde *n* es el número de propiedades y *m* es el número de dimensiones y B es la matriz escalar obtenida a partir cada una de las proximidades (δ_{ij}) obtenidas en el paso 1.

3.- Se calcula la matriz $B=XX'$.

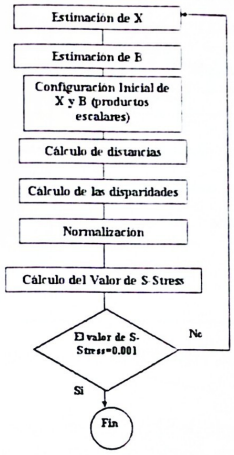
4.- Se realiza el cálculo de las distancias usando la ecuación :

$$d_{ijk}^2 = \sum_{a=1}^{n_r} w_{km}\left(x_{im} - x_{jm}\right)^2$$

Donde w_{km} es el peso de cada sujeto k en la dimensión m, x_{im} es la coordenada de la propiedad i en la dimensión m, x_{jm} es la coordenada de la propiedad j en la dimensión m.

5.- Se calcula el valor de S-Stress, si éste converge en el valor indicado se detiene, sino continua el proceso.

Diagrama 1. Algoritmo ALSCAL



6 Resultados

A continuación se muestran los análisis correspondientes a tres de las propiedades más características para la cuantificación de la dificultad de predicción de las series de tiempo. Para hacer la comparación se tomaron las clasificaciones de las series propuestas por Figueroa [3] (periódicas, cuasi-periódicas, caóticas, complejas y estocásticas) la cual se basa en el comportamiento observado de la dinámica de la señal.

Al hacer el análisis MDS del exponente de Lyapunov (ver Figura 1), se encontraron tres conjuntos que se diferencian por el nivel de inestabilidad que presenta la serie, cuando se le cambian las condiciones iniciales: las series del conjunto I tiene alta estabilidad a los cambios iniciales, las series del conjunto II son medianamente estables a cambios iniciales, y las series del conjunto III son poco estables a cambios iniciales.

Al hacer el análisis MDS del número de reglas de producción (ver Figura 2), se encontraron tres conjuntos que se diferencian por el nivel de la complejidad computacional: las series del conjunto I tiene alta complejidad computacional, las series del conjunto II presentan complejidad computacional media, y las series del conjunto III presentan baja complejidad computacional.

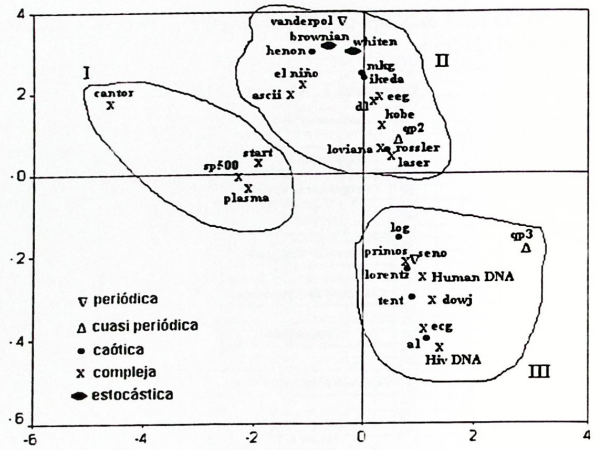


Fig. 1 MDS del exponente de Lyapunov de las treinta series

El análisis MDS de entropía de Shannon (ver Figura 3), nos muestra tres conjuntos que se diferencian por sus niveles de contenido de información: las series del conjunto I tiene alto contenido de información, las series del conjunto II presenta contenido de información medio, y las series del conjunto III tienen poco contenido de información.

Por último queremos hacer notar, que las relaciones de agrupamiento entre series de tiempo cambian dependiendo de la propiedad analizada con MDS, lo que nos indica que la clasificación (periódica, cuasi-periódica, caótica, compleja y estocástica) no permite caracterizar de forma completa su comportamiento. Por ejemplo, human DNA y HIV DNA aparecen en el mismo grupo en la Figura 1 y en distintos grupos en las Figura 2 y la Figura 3. Ya que las dos series se toman como complejas se hubiera pensado que aparecería siempre en el mismo grupo con las diferentes propiedades.

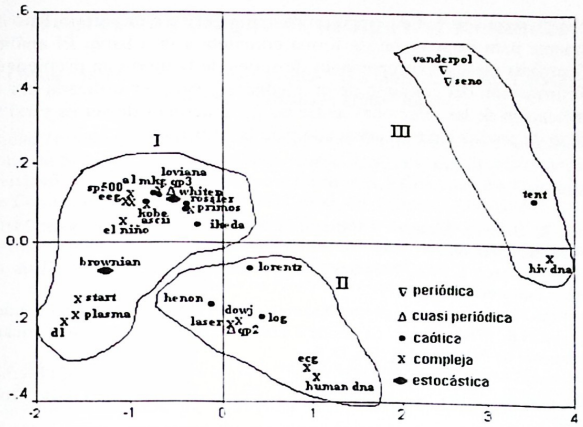


Fig. 2 MDS de reglas de producción de las treinta series

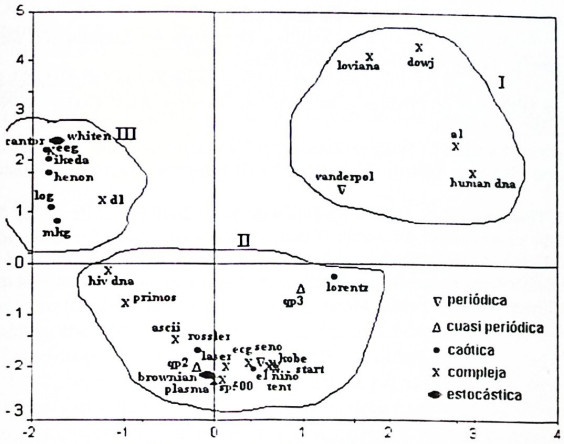


Fig. 3 MDS de entropía (Shannon) de las treinta series

7 Conclusiones

La relación entre la clasificación de referencia: (periódicas, cuasi-periódicas, caóticas, complejas y estocásticas) y las propiedades seleccionadas (exponente de Lyapunov, reglas de producción, entropía de Shannon) que permiten caracterizar la dificultad de predicción de las series de tiempo, es de carácter no lineal. Esto nos in-

dica que clasificar una serie de tiempo únicamente por su comportamiento dinámico no es suficiente para caracterizar de forma completa a la misma. El análisis MDS ayuda a identificar clases o agrupaciones de series de tiempo con propiedades similares. La información del conjunto de propiedades puede ser utilizada por ejemplo, en la identificación de las relaciones entre las agrupaciones de series y así seleccionar el modelo de predicción más adecuado para las mismas.

Referencias.

1. William R. Dillon y Matthew Goldstein (1984). Multivariate Analysis: Methods and Applications ,John-Wiley.
2. Dallas E. Johnson (2000). Métodos Multivariados Aplicados al Análisis de Datos. Thomson Editores.
3. E. Bautista-Thompson y J. Figueroa-Nazuno. (2002). Matriz de Conocimiento sobre la Complejidad de Predicción en Series de Tiempo. VII Congreso Iberoamericano en Reconocimiento de Patrones. México, D. F.
4. M.E Acevedo-Mosqueda, C.G. León-Vega y J. Figueroa-Nazuno (2002) Medición de la Complejidad de Series de Tiempo. XLIV Congreso Nacional de Física, Morelia, Michoacán.
5. Espinosa-Contreras A. y J. Figueroa- Nazuno (2002). Análisis del Comportamiento de la Perdida de Paquetes en la Red Internet con técnicas de la Dinámica No lineal. XLIV Congreso Nacional de Física, Morelia, Michoacán.
6. Robert Hilborn (2000). Chaos and Nonlinear Dynamics An Introduction for Scientists and Engineers. Editorial Oxford University Press.
7. E. Bautista-Thompson y J. Figueroa-Nazuno (200). Análisis de Propiedades que Caracterizan la Dificultad de Predicción en Series de Tiempo. XLV Congreso Nacional de Física. León, Guanajuato.
8. R. Menchaca-Méndez, C. Sanchez-Rodrígez y J. Figueroa-Nazuno. Predicción de Series de Tiempo Mediante Análisis Gramatical. XLIII Congreso Nacional de Física. Puebla, Puebla.
9. Bautista Thompson, E. F. (2004). Medición de la Predictibilidad de Series de Tiempo: Un Estudio Experimental, Doctorado, Centro de Investigación en Computación, IPN.
10. Bustillo Hernández C. y Figueroa Nazuno J. (2006). Análisis y Predicción de Ozono: Ciudad de México, Memorias del XLIX Congreso Nacional de Física, Sociedad Mexicana de Física, San Luis Potosí, Méx.