Normalización de entradas de una red de Kohonen para clasificación de síndrome metabólico en adultos mayores de la ciudad de Cuenca

Christian Vintimilla[1], *Fabian Astudillo*[1], *Erika Severeyn*[2,3], *Lorena Encalada*[4], *Sara Wong*[1,2]

[1] Departamento de Ingeniería Eléctrica, Electrónica y Telecomunicaciones, Universidad de Cuenca, Av. 12 de Abril y Agustín Cueva, Cuenca, Ecuador, 010201.
[2] Grupo de Bioingeniería y Biofísica Aplicada, Universidad Simón Bolívar, Venezuela.
[3] Departamento de Tecnología de Procesos Biológicos y Bioquímicos, Universidad Simón Bolívar, Venezuela.
[4] Facultad de Ciencias Médicas, Universidad de Cuenca, Av. 12 de Abril y Paraíso, Cuenca, Ecuador, 010204.

Autor para correspondencia: sara.wong@ucuenca.edu.ec

Fecha de recepción: 25 de agosto de 2017 - Fecha de aceptación: 29 de septiembre de 2017

RESUMEN

Uno de los desafíos al usar mapas autoorganizativos de Kohonen (SOM) es el preprocesamiento o normalización de las variables de entrada. En el presente trabajo se exploran dos técnicas preprocesamiento (binaria y por rangos) de las variables para diagnosticar Síndrome Metabólico (SM) en adultos mayores de las parroquias urbanas de Cuenca. Se realizaron tres experimentos: considerando toda la población (N=387) y dividiendo la población por sexos, en cada experiencia se definieron 3 clústeres. Los resultados, usando un preprocesamiento por rangos, permiten una mejor clasificación de la población en todos los casos. Este estudio ha permitido seleccionar el tipo de preprocesamiento para el diagnóstico de SM en la población de Adultos Mayores (AM) de la ciudad de Cuenca usando SOM.

Palabras clave: Redes neuronales, Kohonen, síndrome metabólico, preprocesamiento, SOM.

ABSTRACT

One of the challenges in using Kohonen self-Organizing maps (SOM) is the pre-processing or normalization of input variables. In the present work two pre-processing techniques (binary and by ranges) of the variables to diagnose Metabolic Syndrome (MS) in older adults of the urban districts of Cuenca are explored. Three experiments were carried out considering the entire population (N=387) and dividing the population by sex; in each experiment 3 clusters were defined. The results, using pre-processing by ranges allow a better classification of the population in all cases. This study allowed us to select the type of pre-processing for the diagnosis of MS in the elderly population of the city of Cuenca using SOM.

Keywords: Neuronal networks, Kohonen, metabolic syndrome, pre-processing, SOM.

1. INTRODUCCIÓN

El algoritmo de los mapas autoorganizativos de Kohonen SOM (Self-Organizing Map) se basa en un proceso iterativo de comparación con un conjunto de datos y cambios para aproximarse a los mismos, crea un modelo de esos mismos datos para agruparlos por criterios de similitud (Kohonen & Honkela, 2007). Los mapas de Kohonen han sido usados para numerosas aplicaciones médicas, entre ellas para

el diagnóstico de diabetes y síndrome metabólico (SM) (Kohonen & Honkela, 2007; Isasi & Galván, 2004).

El síndrome metabólico es conocido como precursor de diabetes tipo 2 y de enfermedades cardiovasculares (Klein, Klein, & Lee, 2002). Según los criterios del Programa Nacional de Educación sobre el Colesterol y el Panel de Tratamiento del Adulto (Grundy *et al*., 2005) se diagnostica SM, cuando un sujeto presenta al menos tres de las siguientes cinco condiciones:

i) Glucosa en ayunas alterada (>100 mg/dl),Altos valores de triglicéridos (≥150 mg/dl),
ii) Elevada presión arterial sistólica y diastólica (≥130/85)
iii) Bajos niveles de C-HDL, C-HDL<40 mg/dl para los hombres y <50 mg/dl para mujeres y
iv) Circunferencia Abdominal (CA) aumentada (CA >90 cm en hombres y >80 cm en mujeres).

Uno de los desafíos al usar SOM es el preprocesamiento o normalización de las variables de entrada. Existen varias metodologías: usando categorías o rangos; mediante normalización, restando la media y dividiendo por la desviación estándar; por aplicación de logaritmos, cuando el rango de variación ocupa varios ordenes de magnitud; o por codificación binaria (Montaño-Moreno, 2002). En este trabajo se exploran el uso de rangos y la codificación binaria en el preprocesamiento de las variables para clasificar SM en adultos mayores de las parroquias urbanas de Cuenca, usando mapas autoorganizativos de Kohonen.

2. MATERIALES Y MÉTODOS

2.1. Modelo de Kohonen

A inicios de 1980 Kohonen demostró que un conjunto de datos de entrada puede ordenarse por sí solo, de acuerdo con un modelo de mapas topológicos. Este modelo busca establecer una correlación entre los datos de entrada y el espacio de salida de dos dimensiones (llamado mapa topológico). Esta correlación de *entrada/salida* se indica mediante la activación de zonas en el mapa de salida. La estructura topológica consta de una capa de entrada y una capa de salida, que consiste en *k* neuronas de entrada (que es el conjunto de características o vector de parámetros de un dato de entrada), mientras que, en la capa de salida, llamada también capa de competición, tiene la estructura que se indica en la Figura 1.

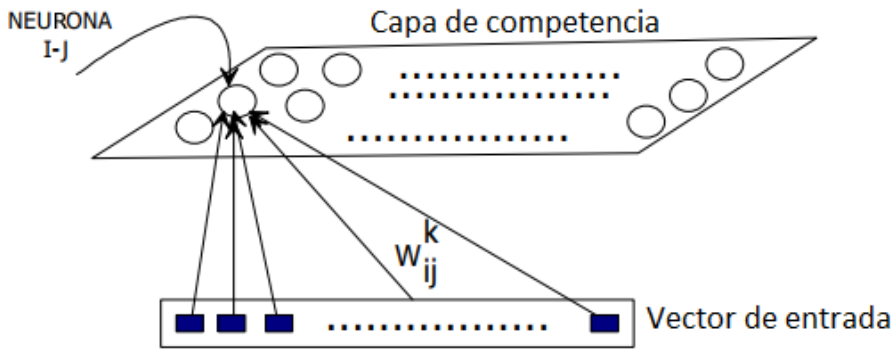

Figura 1. Arquitectura del Modelo de Kohonen

Cada patrón de entrada está conectado con todas las neuronas de la capa de competencia mediante pesos sinápticos (Serrano, Soria, Martin, 2009). A cada una de las neuronas de la capa de competencia se le asigna un vector de pesos como se indica en la Figura 1, de dimensión del vector de entrada.

$$W_{ij} = [W_{ij}^{1} \quad W_{ij}^{2} \quad W_{ij}^{3} \quad \ldots \quad W_{ij}^{k}] \qquad (1)$$

La red busca que patrones de entrada similares activen neuronas próximas en la capa de competencia. Esto se realiza cuando los pesos en la capa de salida, asociados a esa entrada, son semejantes a ese patrón de entrada. Esta semejanza se consigue con medidas de similitud, la más utilizada es la distancia euclidiana.

El algoritmo del modelo de Kohonen se basa en los siguientes pasos: (1) inicializar aleatoriamente los pesos; (2) Presentar un patrón de entrada en cada iteración; (3) determinar la neurona ganadora (mayor similitud del vector de pesos y la entrada), empleando la distancia euclidiana u otras funciones (Ecuación 2); (4) actualizar los pesos sinápticos, en la neurona ganadora y el vecindario (Ecuación 3); (5) volver al paso 2, si no se han realizado todas las iteraciones.

$$d(\mathrm{W_{ij}}, \mathrm{W_R}) = \sqrt{\left[\sum_{k=1}^{M} (\mathrm{W_{ij}}^{k} - \mathrm{x}^{k})^2\right]} \tag{2}$$

$$\mathrm{w_{ij}}(\mathrm{n}+1) = \mathrm{w_{ij}}(\mathrm{n}) + \alpha(\mathrm{n})\, \mathrm{h}(\mathrm{n})(\mathrm{x_j} - \mathrm{w_{ij}}) \tag{3}$$

donde: $\alpha(n)$: es la tasa de aprendizaje. Fija la velocidad de cambio de los pesos. Se establece en función del número de iteraciones o se establece como un valor constante; $h(n)$: es la función de vecindad (con valor máximo en la neurona ganadora).

2.2. Base de datos

Se estudiaron 387 adultos mayores de 65 años de las parroquias urbanas de Cuenca. Se excluyeron sujetos con deficiencia mental, alteración del estado de conciencia, impedimento físico para movilizarse y diabéticos. Una descripción más amplia de la población se encuentra disponible en Chimbo, Chuchuca, Encalada, & Wong (2016). Para cada sujeto se utilizan, como parámetros de entrada, las seis variables usadas para diagnosticar SM. La presión arterial se consideró como una variable doble.

2.3. Pre-procesamiento de los datos

Se han considerado dos tipos de normalización: binaria y por rangos. En cada caso se consideran 5 clases, determinadas a partir de los quintiles del rango de la variable. Los valores de corte para cada variable del SM se muestran en la Tabla 1.

1. Normalización Binaria: las entradas son vectores de 5 elementos, se asigna un valor de uno a la posición de la clase de la variable.
2. Normalización por rangos: se asigna un valor [0.2 0.4 0.6 0.8 1] en función de la clase en la cual se encuentre la variable, tal como se indica en la Tabla 2.

Tabla 1. Valores de Corte para la base de datos de SM (CA: circunferencia abdominal, PAS: presión arterial sistólica, PAD: presión arterial diastólica, HDL: lipoproteínas de alta densidad).

Características	Valores de corte					
	Mínimo	1	2	3	4	Máximo
PAS(mmHg)	90	110	130	150	170	190
PAD (mmHg)	50	62	74	86	98	110
CA (cm)	68.0	80.4	92.8	105.2	117.6	130.0
Glucosa Basal (mg/dl)	49.10	83.28	117.47	151.66	185.85	220.04
Triglicéridos (mg/dl)	54.60	158.80	263.00	367.20	471.40	575.60
HDL (mg/dl)	10.90	34.68	58.46	82.24	106.02	129.80

Tabla 2. Normalización por intervalo.

Valor de entrada	(Min-1)	(1-2)	(2-3)	(3-4)	(4-Max)
Valor normalizado	0.2	0.4	0.6	0.8	1

2.4. Experimentos

Se realizaron tres experimentos, considerando en cada caso los dos tipos de preprocesamiento y las seis variables del SM, en cada experiencia se definieron tres clústeres: (1) toda la población: 270 AM para aprendizaje y 117 para validación; (2) Solo población femenina: 70 para aprendizaje y 75 para validación; (3) Solo población masculina: 100 para aprendizaje y 42 para validación.

3. RESULTADOS

Resultados para toda la población

La Figura 2 muestra la distribución de los clústeres según el tipo de preprocesamiento. En las Tablas 3 y 4 se presentan las características de los tres clústeres obtenidos para la normalización binaria y por rangos respectivamente para el caso del experimento 1 (N=117).

En el caso del preprocesamiento binario (Tabla 3) se observa que las poblaciones son diferentes entre ellas, en relación con la presión arterial. Se observan tres clústeres con valores de presión arterial diastólica y sistólica bien diferenciados. No se presentan diferencias entre los clústeres para el resto de las características del SM. En la Tabla 4 se observa que la mayor parte de la población se divide entre los clústeres 1 y 2. El clúster 3 solo tiene un AM. El clúster 1 presenta valores promedios normales para todas las variables a excepción de los triglicéridos. La mayoría de los valores promedios del clúster 2 se encuentran alterados o en el límite del valor corte de diagnóstico. Se observa que los sujetos agrupados en este grupo presentan características más relacionadas con el SM y por ende un porcentaje mayor de AM con SM. Estas mismas tendencias fueron observadas para la población femenina entre los clústeres 1 y 2 (Tabla 5), el clúster 3 agrupo 10 AM sin SM. Igualmente, los clústeres 1 y 2 de población de AM masculina siguen la misma tendencia observada en las experiencias anteriores, en este caso el clúster 3 agrupa a un solo sujeto con valores de glicemia y triglicéridos claramente muy elevados.

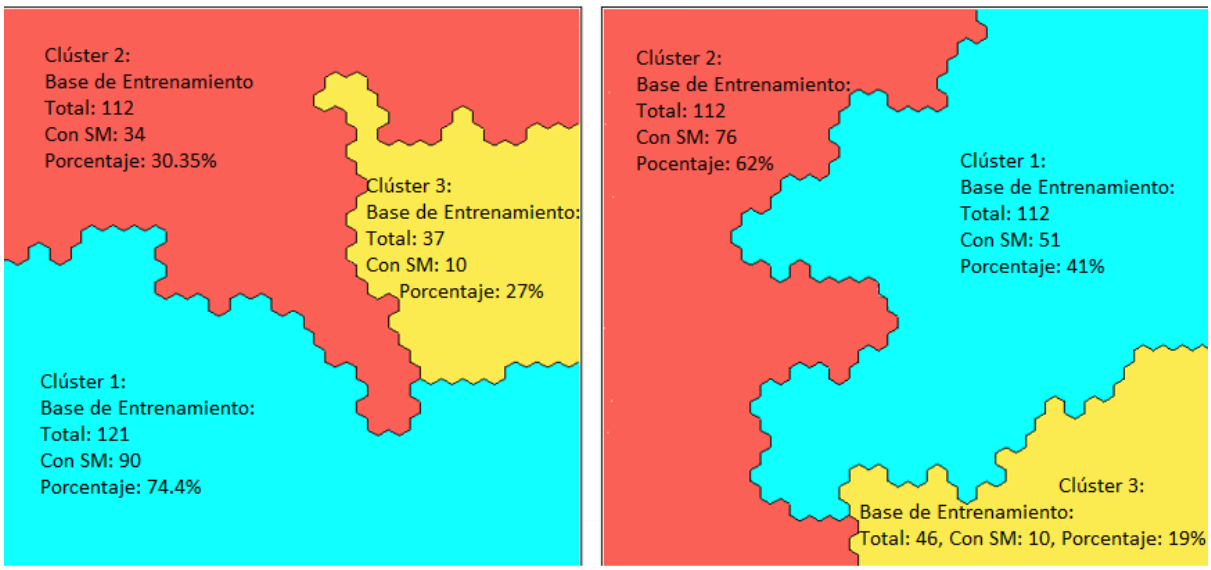

Figura 2. Figura izquierda: SOM usando preprocesamiento binario; figura derecha: SOM usando preprocesamiento por rangos.

Tabla 3. Características de los clústeres para toda la población usando normalización binaria (CA: circunferencia abdominal, PAS: presión arterial sistólica, PAS: presión arterial diastólica, HDL: lipoproteínas de alta densidad).

Validación N=117 Con SM Características	Clúster 1 (N=33) 28 (86.36%) Media ± Std	Clúster 2 (N=63) 27 (42.37%)	Clúster 3 (N=21) 10 (47.62%)	p12	p13	p23
PAS (mmHg)	137.48±14.34	123.46±8.73	112.57±12.04	0.001	0.001	0.001
PAD (mmHg)	84.73±8.54	77.67±6.69	69.29±7.79	0.001	0.001	0.001
CA (cm)	95.71±11.41	94.57±11.66	94.67±9.87	0.841	0.972	0.832
Glucosa (mg/dl)	92.26±23.41	93.47±27.76	93.45±27.34	0.951	0.716	0.844
Triglicéridos (mg/dl)	147.65±79.66	176.90±109.78	155.17±71.28	0.247	0.546	0.642
HDL (mg/dl)	38.04±6.67	43.71±18.39	43.91±8.34	0.157	0.005	0.173

Tabla 4. Características de los clústeres para toda la población usando normalización por rangos (CA: circunferencia abdominal, PAS: presión arterial sistólica, PAS: presión arterial diastólica, HDL: lipoproteínas de alta densidad).

Validación N=117 Con SM Características	Clúster 1 (n=85) 41 (48.23%) Media ± Std	Clúster 2 (n=31) 22 (70.97%)	p12
PAS (mmHg)	120.14±11.01	140.35±10.51	0.000
PAD (mmHg)	75.65±8.05	84.65±8.15	0.000
CA (cm)	93.54±10.32	98.82±12.68	0.076
Glucosa (mg/dl)	90.47±25.22	100.52±28.61	0.040
Triglicéridos (mg/dl)	168.55±95.88	154.02±99.82	0.295
HDL (mg/dl)	39.77±8.67	46.23±18.51	0.080

Tabla 5. Características de los clústeres para la población femenina usando normalización por rangos (CA: circunferencia abdominal, PAS: presión arterial sistólica, PAS: presión arterial diastólica, HDL: lipoproteínas de alta densidad).

Validación N=75 Con SM Características	Clúster 1 (n=45) 35 (77.78%) Media ± Std	Clúster 2 (n=20) 9 (45%)	Clúster 3 (n=10) 0 (0%)	p12	p13	p23
PAS (mmHg)	124.89±13.02	136.00±10.96	111.00±8.76	0.002	0.003	0.001
PAD (mmHg)	75.29±6.86	88.00±6.81	68.00±4.22	0.001	0.002	0.001
CA (cm)	99.51±9.40	86.78±7.60	89.30±9.66	0.001	0.007	0.300
Glicemia (mg/dl)	102.77±33.26	84.60±12.02	81.72±8.47	0.071	0.038	0.416
Triglicéridos (mg/dl)	187.98±111.32	125.39±63.03	133.86±42.38	0.016	0.321	0.416
HDL (mg/dl)	38.86±7.78	46.58±21.70	45.59±11.10	0.172	0.046	0.495

Tabla 6. Características de los clústeres para la población masculina usando normalización por rangos (CA: circunferencia abdominal, PAS: presión arterial sistólica, PAS: presión arterial diastólica, HDL: lipoproteínas de alta densidad).

Validación N=42 Con SM Características	Clase 1 (n=23) 3 (13.04%)	Clase 2 (n=18) 12 (66.67%)	Clase 3 (n=1) 1 (100%)			
		Media ± Std		p12	p13	p23
PAS (mmHg)	115.87±6.36	135.17±13.85	135.00±0.00	0.001	0.092	0.632
PAD (mmHg)	75.61±7.38	84.00±7.00	65.00±0.00	0.001	0.133	0.105
CA (cm)	89.30±12.13	102.72±7.81	87.00±0.00	0.001	0.828	0.105
Glicemia (mg/dl)	78.58±7.37	96.67±16.17	197.80±0.00	0.001	0.112	0.105
Triglicéridos (mg/dl)	164.15±80.53	160.05±113.48	278.20±0.00	0.554	0.248	0.211
HDL (mg/dl)	45.10±20.22	40.57±9.67	35.90±0.00	0.590	0.563	0.632

4. CONCLUSIONES

El preprocesamiento por rangos permite una mejor clasificación de la población con SM. La selección del método de preprocesamiento de los datos es de suma importancia para la utilización de SOM. Este estudio ha permitido escoger el tipo de preprocesamiento para el diagnóstico de SM en la población de AM de la ciudad de Cuenca. Los trabajos futuros están orientados a experimentar con diferentes números de clases para la normalización por rangos y de clústeres para el diagnóstico de SM en AM usando SOM.

AGRADECIMIENTOS

Investigación financiada por la Dirección de Investigación de la Universidad de Cuenca (DIUC).

REFERENCIAS

Chimbo, J., Chuchuca, A., Encalada, L., Wong, S. (2016). Nivel de actividad física medida a través del Cuestionario Internacional de Actividad Física, en Adultos Mayores de las parroquias urbanas de Cuenca-Ecuador, 2015 Presentado en el Congreso en Investigación de la Salud: Enfoques, avances y desafíos. Universidad de Cuenca. Junio de 2016. *Revista de la Facultad de Ciencias Médicas, 34*(2), 51-56.

Grundy, S. M., Cleeman, J. I., Daniels, S. R., Donato, K. A., Eckel, R. H., Franklin, B. A., … Costa, F. (2005). Diagnosis and Management of the Metabolic Syndrome. *Circulation*, *112*(17), 2735-2752. https://doi.org/10.1161/CIRCULATIONAHA.105.169404

Isasi, P., Galván, I. (2004). *Redes neuronales artificales: un enfoque práctico*. Madrid, Spain: Pearson Prentice Hall.

Klein, B. E. K., Klein, R., Lee, K. E. (2002). Components of the metabolic syndrome and risk of cardiovascular disease and diabetes in Beaver Dam. *Diabetes Care*, *25*(10), 1790–1794.

Kohonen, T., Honkela, T. (2007). Kohonen network. *Scholarpedia*, *2*(1), 1568. https://doi.org/10.4249/scholarpedia.1568

Montaño-Moreno, J. (2002). *Redes neuronales artificiales aplicadas al análisis de datos* (Ph.D. Thesis). Universitat de les Illes Balears. Retrieved from http://www.tdx.cat/handle/10803/9441