ISSN 1870-4069

Análisis de expedientes clínicos para el diagnóstico de cáncer de mama a partir de memorias asociativas evolutivas: un primer avance

Juan Villegas-Cortez[1], Beatriz A. González-Beltrán[1], Fernando Torres-Vizueth[1], Salomón Cordero-Sánchez[2]

[1] Universidad Autónoma Metropolitana, Unidad Azcapotzalco, Departamento de Sistemas, México

[2] Universidad Autónoma Metropolitana, Unidad Iztapalapa, Departamento de Química, México

{juanvc, bgonzalez}@azc.uam.mx, al2192800401@alumnos.azc.uam.mx, scs@xanum.uam.mx

Resumen. La enfermedad del cáncer en todos sus tipos se sigue estudiando para poder entenderla mejor, dado que su padecimiento es de diversas formas y son muchos los factores que pueden relacionarse con un diagnóstico final de si una persona tiene o no un determinado tipo de cáncer. En este trabajo presentamos una primera propuesta de análisis del cáncer de mama, a partir de una base de datos de expedientes clínicos bien reconocida en el medio de la comunidad de estudio del reconocimiento de patrones. Se propone el uso de memorias asociativas evolutivas como herramienta de análisis desde el aprendizaje automático, que de acuerdo a la investigación realizada en el estado del arte de nuestro problema no ha sido usada hasta el momento, y estas han demostrado resultados prometedores. Nuestro objetivo es brindar un nuevo punto de vista de los factores de la enfermedad como componentes de los patrones; y analizar el comportamiento de la clasificación desde una base de datos conocida. Cabe señalar que no se busca una reducción de dimensiones del patrón, sino de arrojar luces de los factores posiblemente relacionados con la enfermedad.

Palabras clave: Memorias asociativas evolutivas, diagnóstico clínico, reconocimiento de patrones, programación genética.

Analysis of Clinical Records for Breast Cancer Diagnosis by Means of Evolutionary Associative Memories: A First Approach

Abstract. Cancer disease in all its types is still being studied in order to better understand it, given that its condition is of various forms and there are many factors that can be related to a final diagnosis of whether or not a person has a certain type of cancer. In this work, we present a first approach for the analysis

of breast cancer, based on a well-recognized database of clinical records in the pattern recognition study community. The use of evolutionary associative memories is proposed as an analysis tool from machine learning, which according to the research carried out in the state of the art has not been used so far, and these have shown promising results. Our goal is to provide a new point of view of cancer factors as components of patterns; and analyzing the classification behavior from a known database. It should be noted that it is not intended to reduce the size of the pattern, but rather to shed light on the factors possibly related to the disease.

Keywords: Evolutionary associative memories, pattern recognition, clinical records, genetic programming.

1. Introducción

El cáncer es una de las primeras causas de muerte en el mundo alcanzado 8,2 millones de muertes en 2012 [6]. Para el caso de México, entre enero y agosto de 2020 se tuvieron 683,823 muertes, de los cuales el 9 % fue debido a esta enfermedad (60,421). Un año antes, en 2019 se registraron 747,784 defunciones, de las cuales el 12 % (88,683) fue debido al cáncer [3]. En México, con datos de 2017, y considerando los diferentes tipos de cáncer, el cáncer de mama constituye la principal causa de morbilidad en la población de 20 años y más [3].

Existen diferentes bases de datos que han sido extraídas de expedientes clínicos electrónicos que pueden ser analizadas con el objetivo de reconocer los patrones que posiblemente estén relacionados con la enfermedad. En este trabajo se utilizó la base de datos para cáncer de mama de la Universidad de Wisconsin [14], una base de datos clásica para la comunidad de ciencia de datos.

En este trabajo presentamos una propuesta de análisis de una base de datos de cáncer, del tipo expediente clínico, desde la herramienta de una red neuronal artificial evolutiva específica para este tipo de patrones, la memoria asociativa evolutiva (MAE) [13], con la finalidad de brindar una perspectiva nueva sobre la caracterización de los patrones en sus componentes acerca de cuáles son relevantes para la clasificación, y por ende, buscando presentar a los profesionales de la salud un enfoque de cuáles parámetros o datos de los expedientes son los relevantes para poder determinar si el paciente tiene cáncer o no.

En la sección 2, presentamos el estado del arte de este problema de estudio, tanto desde el estudio de este tipo de bases de datos, como de las MAE como herramienta de análisis de datos desde el cómputo evolutivo con la programación genética. La metodología de análisis se describe en la sección 3, y nuestros experimentos con los resultados obtenidos en la sección 4. Finalmente, en la sección 5 compartimos nuestras conclusiones y líneas de trabajo futuro.

2. Estado del arte

El estudio de la enfermedad del cáncer, a partir de expedientes clínicos, se puede abordar desde la parte ingenieril, y por ende desde la inteligencia artificial, a partir de un análisis numérico de los datos de los expedientes conformados como patrones.

Tabla 1. Descripción de las diez variables de la DB, siendo las 9 primeras las características del expediente clínico, y la última la de clasificación del tumor como benigno o maligno.

No.	Nombre de la variable	Descripción	Dominio
1	Clump Thickness	Espesor del grupo	[1,10]
2	Uniformity of Cell Size	Uniformidad del tamaño de la célula	[1,10]
3	Uniformity of Cell Shape	Uniformidad de la forma de la célula	[1,10]
4	Marginal Adhesion	Adhesión marginal	[1,10]
5	Single Epithelial Cell Size	Tamaño de célula epitelial simple	[1,10]
6	Bare Nuclei	Núcleos desnudos	[1,10]
7	Bland Chromatin	Cromatina blanda	[1,10]
8	Normal Nucleoli	Nucléolos normales	[1,10]
9	Mitoses	Mitosis	[1,10]
10	Class	Clase	2:Benigno, 4:Maligno

Se tienen diversas bases de datos en repositorios tales como el de la Universidad de Californa, Irvine (UCI)[3], y de ahí es que se muestra en este trabajo la posibilidad de realizar un estudio a partir de una base de datos representativa. La base de datos (DB) usada en nuestro estudio para cáncer de mama es de la Universidad de Wisconsin y se denomina Breast Cancer Wisconsin [14].

Esta DB está compuesta de 699 instancias y de nueve atributos para el análisis: espesor del grupo, uniformidad del tamaño de la célula, uniformidad de la forma de la célula, adhesión marginal, tamaño de célula epitelial simple, núcleos desnudos, cromatina blanda, nucléolos normales y mitosis. Cada uno de estos atributos tienen un valor $\in [1, 10]$.

Para nuestro propósito, la DB está formada por patrones tipo vector-renglón de 10 variables-componentes o atributos, 9 son independientes y la entrada final, la componente número 10, indica si el tumor fue benigno (indicado con el número 2) o si es maligno (indicado con número 4). En la Tabla 1 se muestra esta descripción con detalle.

En [5], se propone la mejora en la precisión de la clasificación del diagnóstico de cáncer de mama. En este trabajo realizan la clasificación de características utilizando un algoritmo genético. Además, extraen la características óptimas utilizando el algoritmo Cost-Sensitive Support Vector Machine (CSSVM). Los autores utilizaron también el conjunto de datos de Wisconsin Breast Cancer y Wisconsin Breast Cancer Diagnosis. El resultado de la clasificación obtuvo un 95.7 % de precisión.

Las redes neuronales artificiales (RNA) se conciben como un paradigma de aprendizaje y procesamiento automático que está bio-inspirado en la forma en que se describió el funcionamiento del sistema nervioso de animales en la década de los 60s [7, 1]. Tal que las RNA se presentan como un sistema de interconexión de neuronas en una red que coopera para producir un estímulo de salida.

[3] UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

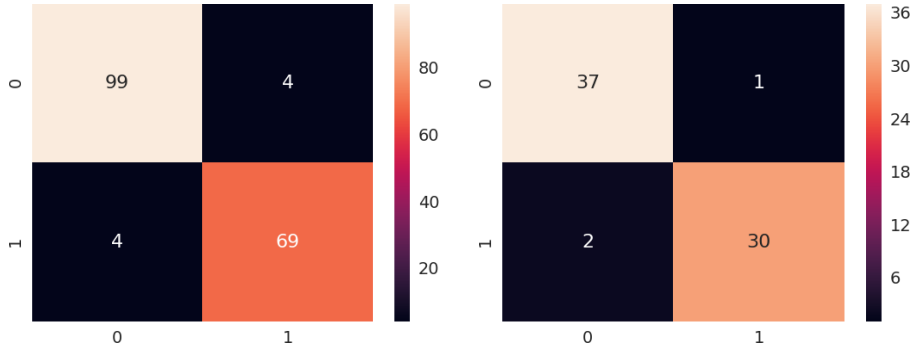


Fig. 1. Matrices de confusión de la primer prueba con el 75 % de entrenamiento, y 25 % de prueba (a), y de la segunda prueba (b), con partición del 90 % y 10 %, respectivamente.

Las RNA han sido diversificadas en su diseño buscando lograr clasificar patrones de clases que no son linealmente separables, así como de clases mezcladas, pero también proporcionan una herramienta para entender el problema de clasificación en su complejidad numérica ante el desafío del aumento de la cantidad de patrones, y de sus componentes. Hace 20 años, un dilema al momento de aplicar una RNA a un problema de reconocimiento de patrones era la cantidad de componentes de los patrones, los rasgos característicos del patrón.

Se buscaba tener una reducción de estos componentes con el objetivo de reducir la complejidad numérica de la RNA y por la limitante del hardware (cantidad de memoria a usar, los ciclos de reloj y la precisión numérica). Es así que se trabajaron nuevos modelos de RNA a partir del cómputo evolutivo [13, 10], buscando dos puntos fundamentales: por un lado, la reducción de la complejidad de la red en el número de capas y neuronas; y por otro lado, proporcionar una nueva perspectiva para entender los patrones de cada caso de estudio.

De entre los modelos RNA, es importante resaltar las memorias asociativas (MA), un tipo de red que no tiene arquitectura de capas, ni involucra la retro-propagación. Este modelo se enfoca en realizar una asociación de los patrones utilizando dos aspectos: “auto asociativo”, cuando se entrena la red asociando al patrón-vector consigo mismo, y “hetero-asociativo”, cuando al patrón se le asocia con otro. Un modelo más común de MA es el inspirado por las memorias morfológicas y la regla de aprendizaje de Hebb [9].

Una MAE se desarrolla a partir de la programación genética (PG), como un proceso co-evolutivo. Se realiza una primera evolución para la etapa de asociación, y se lleva a cabo otro proceso evolutivo, en co-evolución cooperativa con el primero, para la recuperación del patrón a partir de la MA construida en el primer proceso, y con una función de aptitud conjunta [12].

La PG es considerada un método automático para la creación de programas de cómputo como solución en alto nivel para el problema a solucionar [4]. Además, la PG se considera también una técnica de aprendizaje automático para optimizar una población de programas, de acuerdo a una función de aptitud que evalúa la capacidad de cada programa para resolver la tarea en cuestión.

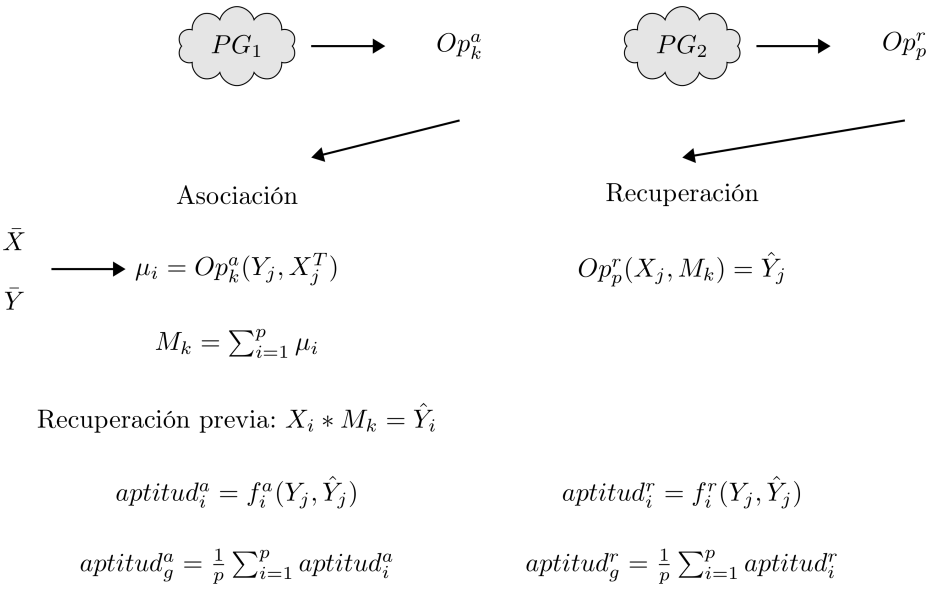


Fig. 2. Metodología co-evolutiva implementada para el desarrollo de MAE con programación genética.

En [13], se presenta una revisión del estado del arte de las MA a profundidad, y se muestra el detalle del desarrollo de la MAE, mostrando su efectividad para problemas en patrones reales y binarios, y también para el caso duro de patrones con ruido mixto y en problemas de visión artificial.

3. Metodología

Las MAE nos proporcionan en su análisis de datos una perspectiva de cuáles son las componentes de los patrones involucrados que tienen mayor importancia o preponderancia [13]; por lo anterior, nuestro problema consiste en analizar la enfermedad del cáncer desde la perspectiva numérica, ya que no somos profesionales de la salud.

Primero hicimos un análisis de clasificación de los patrones usando una RNA con multicapa clásica, tomando la DB de Wisconsin y, dando como valores en la capa de entrada las nueve variables de los patrones como vectores; para las capas internas se consideraron tres capas de 100 neuronas y, finalmente, una sola salida para la clasificación si el tumor es benigno o maligno. Se realizó una primera prueba, con una partición del 75 % de los patrones para entrenamiento y 25 % para prueba, donde se obtuvo un porcentaje de clasificación del 96 %.

Posteriormente, se realizó una segunda prueba, ahora con una partición del 90 % de los patrones por clase para entrenamiento, y con 10 % de los patrones para prueba, obteniendo un porcentaje de clasificación del 95 %. En la RNA se trabajó con el optimizador de descenso de gradiente estocástico (sgd) para el cálculo del mínimo de la función de costo.

Tabla 2. Resumen de los parámetros evolutivos involucrados en las pruebas.

Caso	Métrica en función de aptitud	Número de MAE a generar	Generaciones	Individuos por generación
I	Arco coseno	5	20	10
II	Error cuadrático medio	5	20	17

Tabla 3. Resumen de las duplas como MAE generadas para el caso I.

MAE índice	Regla de Asociación	Regla de Recuperación	Aptitud
1	1	1	100 %
2	1	2	100 %
3	2	3	99.8 %
4	1	4	100 %
5	2	5	100 %

Estos resultados se aprecian mejor en la Figura 1 con las matrices de confusión de la clasificación, donde los porcentajes están redondeados a valores enteros de la cantidad de patrones a considerar para caso de prueba. Como podemos ver los resultados obtenidos de la clasificación son bien conocidos en el medio de estudio de las RNA, luego ahora la propuesta es estudiar a los patrones en sí, con sus nueve características clínicas reportadas, en una MAE usando la auto-asociación.

3.1. Parámetros evolutivos para la PG

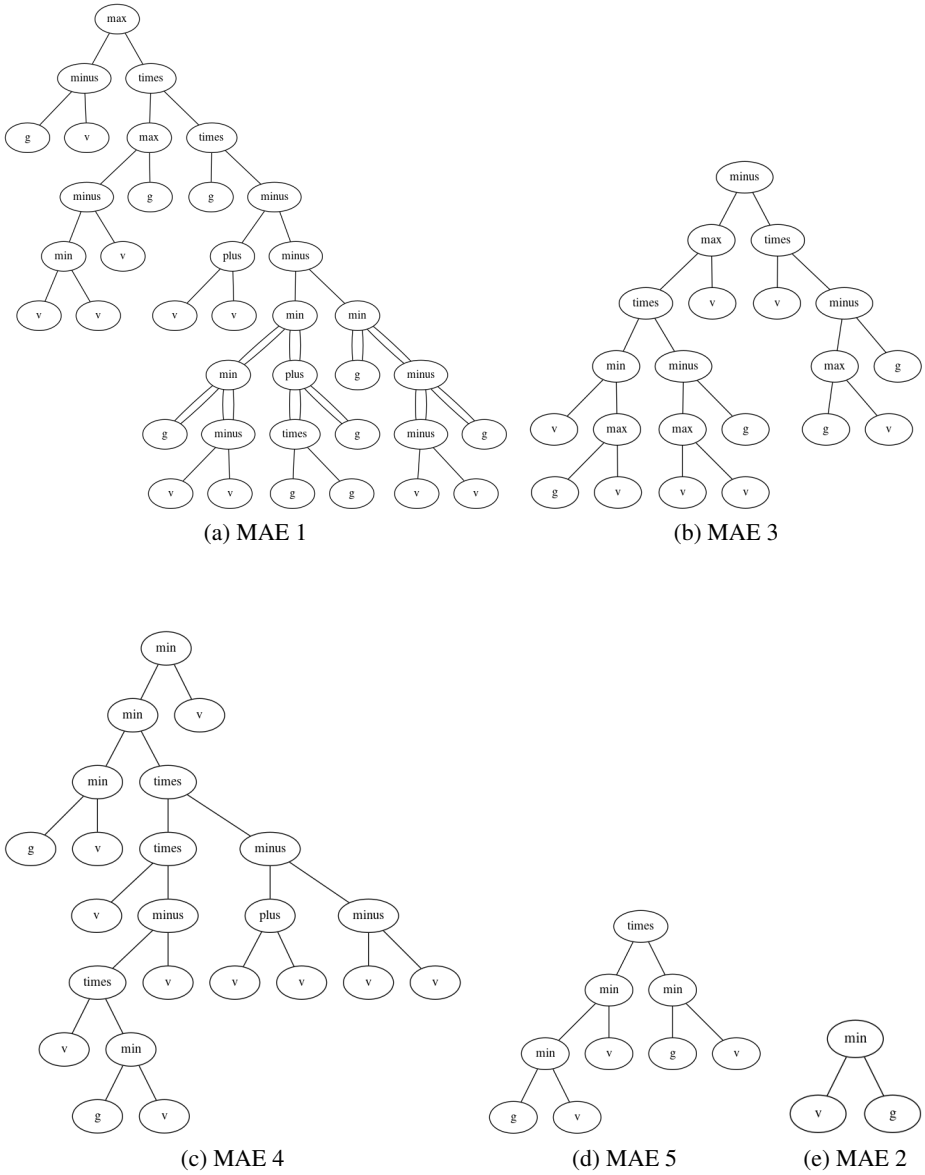
Se plantea desde la PG que ahora el individuo es un programa tentativo a dar solución al problema de hallar una asociación del patrón, cada individuo se asocia consigo mismo, tal que se genera un "dispositivo" de almacenamiento de sus características. Esto es, presentando un patrón de entrada se tiene la recuperación del mismo, diferenciándolo de otros patrones de ese repositorio, representado como un "conocimiento" en la MAE.

En el proceso co-evolutivo, se trabajaron poblaciones con regeneración en cada ciclo evolutivo con cruza al 70 % y mutación del 10 % del individuo para generar un 30 % de individuos mutados de la nueva generación, preservando con elitismo al mejor individuo de cada generación.

Este criterio es con base a la experiencia de analizar otros problemas de reconocimiento de patrones con este tipo de proceso evolutivo de la PG. Cabe señalar que, como criterio de paro está el alcanzar el 100 % de recuperación, o tener cero error, o bien agotar el número de generación en co-evolución.

En la etapa de la asociación se trabajan operaciones a nivel escalar, con las entradas del patrón, tal que el conjunto terminal de la auto-asociación es $T_a = \{x_i, y_i\}$, para los patrones de entrada y salida $\{X, Y\}$, respectivamente; y el conjunto de funciones en esta etapa es: $F_a = \{+, -, \min, \max, \text{times}\}$.

La función de aptitud local, aptitud$_g^a$, se aplica a considerar el porcentaje de recuperación por pareja asociada, luego el promedio de ellas. En la Figura 2, se muestra el detalle tanto para el proceso de asociación como de recuperación.

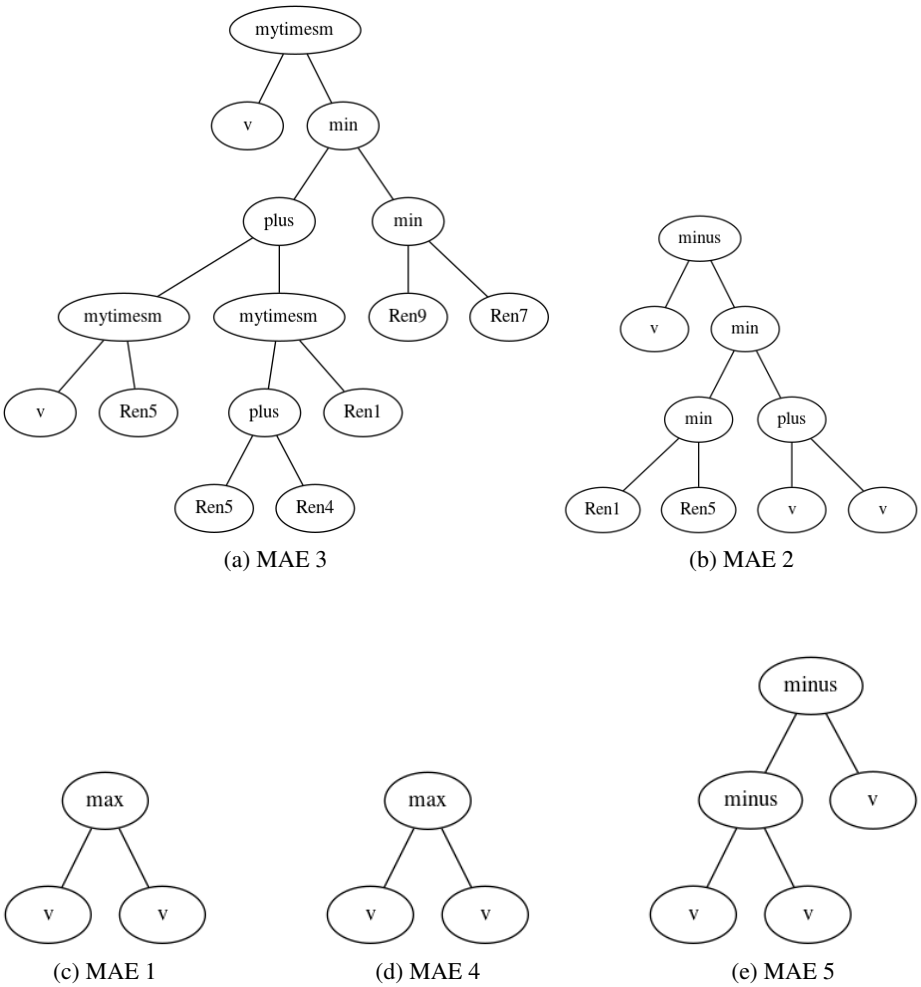


(a) MAE 1 (b) MAE 3

(c) MAE 4 (d) MAE 5 (e) MAE 2

Fig. 3. Reglas de asociación de patrones de las MAE para el caso I.

Para la etapa de recuperación se trabajan las operaciones considerando a los renglones de la matriz de asociación generada en la primera parte, adicionales a los patrones de asociación involucrados, v, y a la misma matriz de asociación generada, M_k, tal que ahora el conjunto de terminales es $T_r = \{v,\ \text{Ren}_1,\ \text{Ren}_2,\ ...,\ \text{Ren}_m,\ M_k\}$, y la función de aptitud co-evolutiva es aptitud$_g^r$. En [13], se tiene mayor detalle de la operación y metodología de las MAE.



(a) MAE 3 (b) MAE 2

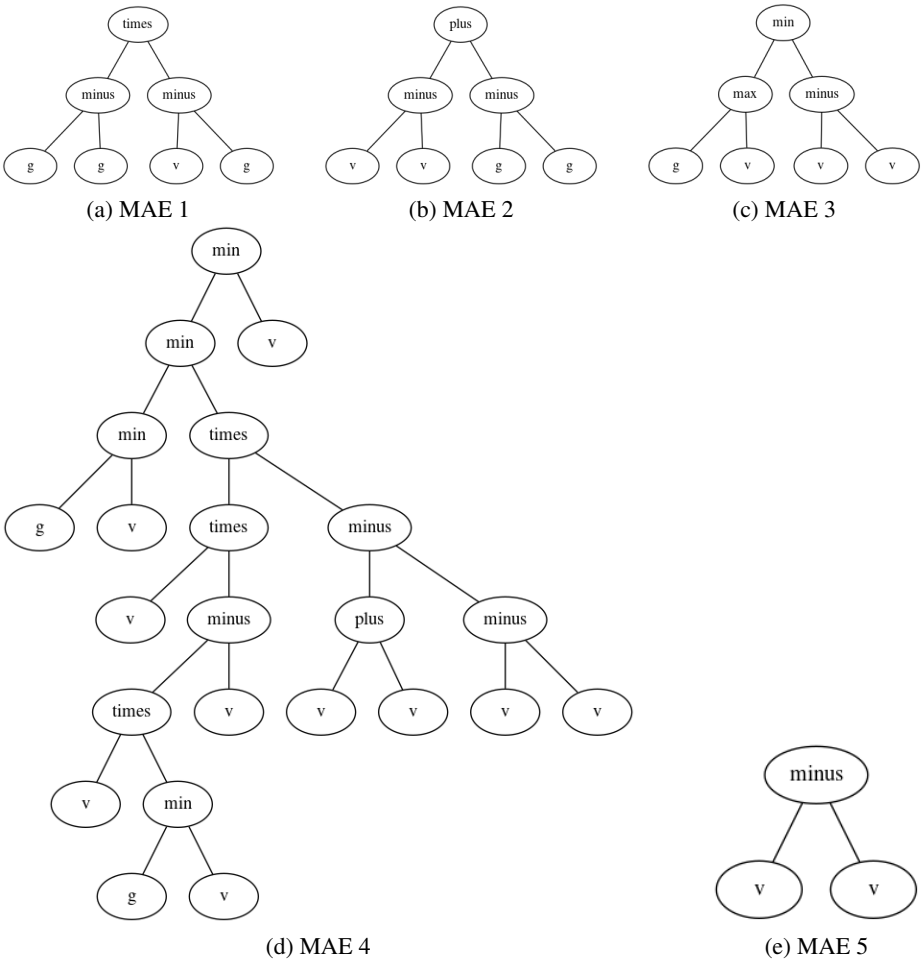
(c) MAE 1 (d) MAE 4 (e) MAE 5

Fig. 4. Reglas de recuperación de patrones de las MAE para el caso I.

4. Experimentos y resultados

La implementación se realizó en una computadora tipo WorkStation con procesadores Intel Xeon 64-bit, con sistema operativo Linux y Matlab con Toolbox GPLab [8], versión 3.0. Realizamos dos experimentos de auto-asociación con las MAE sobre la DB buscando brindar dos aportaciones de cómo se comportan estos patrones.

El primer caso (I), fue considerando a la métrica del arco coseno como función de aptitud, y el segundo caso (II), fue aplicando la medida del error cuadrático medio. La intención del caso I fue medir la similitud de los patrones en un espacio multidimensional, dada la limitante del arco coseno [2]; y la intención del caso II fue analizar la posibilidad de cercanía de estos patrones en ese espacio dimensional usando el error cuadrático medio.



(a) MAE 1 (b) MAE 2 (c) MAE 3

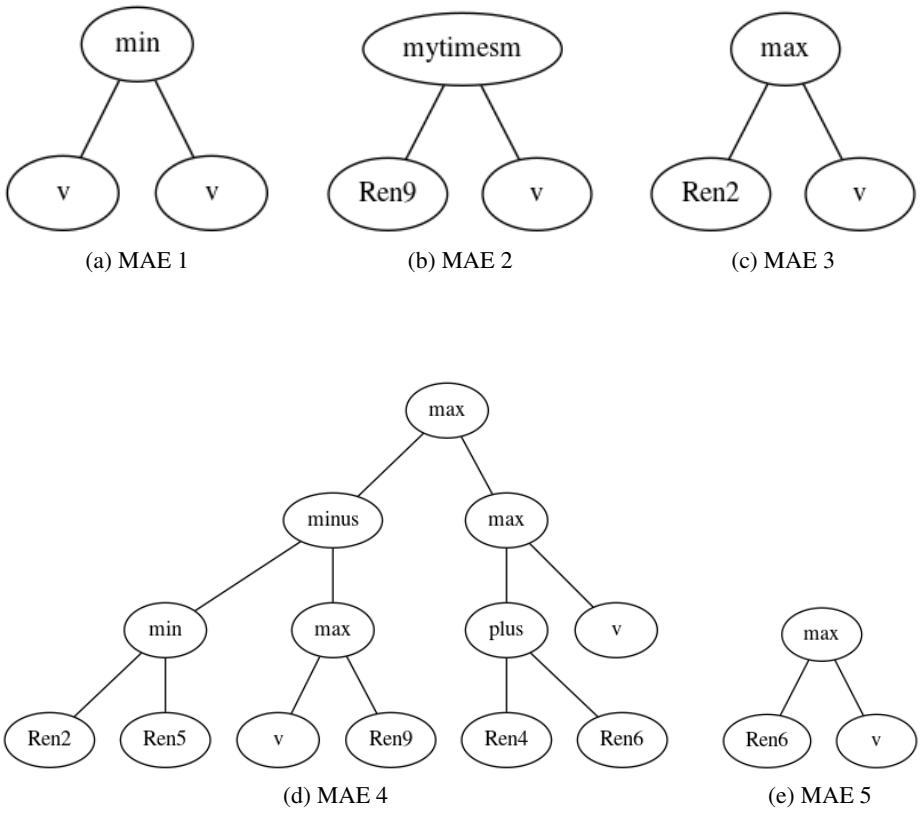
(d) MAE 4 (e) MAE 5

Fig. 5. Reglas de asociación de patrones de las MAE para el caso II.

En la Tabla 2, se muestra el resumen de valores para ambas pruebas. El número de individuos y generaciones se considera con base a la experiencia del estado del arte [13, 11].

4.1. Resultados para el caso I

Para el primer caso, la Figura 3 muestra las reglas de asociación obtenidas al final del proceso co-evolutivo de las 5 MAE generadas; mientras que la Figura 4 presenta las reglas de recuperación halladas. También podemos ver en el resumen de la Tabla 3 que, de las cinco duplas obtenidas, cuatro de ellas logran una recuperación perfecta, en este caso, una auto-asociación por similitud de los patrones con esas reglas de asociación y recuperación respectivas.



(a) MAE 1 (b) MAE 2 (c) MAE 3

(d) MAE 4 (e) MAE 5

Fig. 6. Reglas de recuperación de patrones de las MAE para el caso II.

Analizando los individuos-árboles generados para la recuperación de la Figura 4, la variable v representa el vector de entrada a recuperar, cada uno de la DB, y acorde a los resultados de la Tabla 3, cuatro MAE lograron recuperación perfecta por asociación de similitud, mientras que uno casi lo logra, la tercer dupla, que curiosamente involucra al individuo más complejo, el árbol con más nodos de la figura.

A partir de estos individuos, los resultados nos reflejan que son necesarios los nueve rasgos de la DB para lograr la recuperación perfecta, esto es en las duplas MAE con índice 1, 4 y 5; mientras que el individuo 2 involucra únicamente los rasgos del renglón 1 y 5, que son el espesor del grupo y el tamaño de la célula epitelial simple.

4.2. Resultados para el caso II

Ahora, analizando el resultado para el caso segundo, en la Tabla 4 se muestran los valores de recuperación obtenidos. Tras una búsqueda exhaustiva, por parte del proceso evolutivo, se hallaron cuatro MAE con recuperación al 100 %, y una, la segunda dupla con recuperación muy mala, del orden del 41.46 %. Aún así, en las Figuras 5 y 6 se muestran los árboles de las reglas de asociación y recuperación correspondientes.

Tabla 4. Resumen de las duplas como MAE generadas para el caso II.

MAE índice	Regla de Asociación	Regla de Recuperación	Aptitud
1	1	1	100 %
2	1	2	41.46 %
3	1	3	100 %
4	1	4	100 %
5	1	5	100 %

De este proceso evolutivo, para el caso de la auto-asociación por error cuadrático medio, en la Figura 6 podemos visualizar que los renglones 2, 6 y 9 son los más relevantes de las nueve características comprendidas de la DB, a saber de la descripción en la Tabla 1, los rasgos involucrados son: uniformidad de tamaño de célula, el núcleo desnudo, y la mitosis. Esto se visualiza en los árboles generados MAE 3, 4 y 5, el árbol-regla de recuperación 2 lo descartamos por no haber logrado la recuperación perfecta de acuerdo a los valores mostrados en la Tabla 4. De estos resultados obtenidos, las relaciones de asociación arrojados por las MAE son meramente especulativas a este nivel, el numérico, y lo comentamos porque nosotros como autores de este trabajo de análisis no somos profesionales de la salud.

5. Conclusiones

En este trabajo hemos presentado un primer avance en el análisis de expedientes clínicos para el diagnóstico de cáncer de mama, a partir de las memorias asociativas evolutivas, usando una base de datos reconocida en el medio de investigación del reconocimiento de patrones. Si bien es una DB con apenas 9 rasgos para tipificar si se tiene o no la enfermedad, con la clase tumor maligno o benigno, nos proporciona una herramienta poder adentrarnos en el estudio de este tipo de patrones para dar un resultado que pudiera ser de apoyo para los profesionales de la salud desde una técnica de inteligencia artificial.

Los resultados numéricos presentados no pretenden afirmar lo que pueda relacionar un médico o profesional de la salud, insistimos en esto reconociendo nuestras limitaciones como profesionales con formación en ingeniería y ciencias básicas. Las relaciones de asociatividad presentadas en este artículo sobre los rasgos característicos de la DB sobre el expediente de un paciente, nos indican una posible dirección de cómo seguir apoyando a los profesionales de la salud, esto también nos invita a poder extender esta línea de aplicación hacia bases de datos más extensas tanto en rasgos como en pacientes.

Como trabajo futuro planteamos abordar DB sobre enfermedades que apremian un estudio de apoyo, específicamente COVID-19 y diabetes, esta última enfermedad en México, nuestro país, donde 1 de cada 10 personas la padecen, y todo apunta a que no cambiará este dato estadístico.

Agradecimientos. Este trabajo es resultado del proyecto divisional “Evolución artificial de descriptores estadísticos de textura de superficie para implementación en clasificación de imágenes digitales”, clave: EL006-18, de la Universidad Autónoma Metropolitana, Unidad Azcapotzalco.

Referencias

1. Fausett, L.: Fundamentals of neural networks: Architectures, algorithms, and applications. Prentice-Hall, Inc, Upper Saddle River (1994)
2. Fukunaga, K.: Introduction to statistical pattern recognition (2nd ed.). Academic Press Professional, Inc (1990)
3. INEGI. Estadísticas a propósito del día mundial contra el cáncer. Comunicado de prensa número 105/21 (2021) https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2021/cancer2021_Nal.pdf
4. Koza, J. R.: Genetic programming as a means for programming computers by natural selection. Statistics and Computing, Springer Science and Business Media LLC, vol. 4, no. 2 (1994) doi: 10.1007/bf00175355
5. Liu, N., Qi, E. S., Xu, M., Gao, B., Liu, G. Q.: A novel intelligent classification model for breast cancer diagnosis. Information Processing and Management, vol. 56, no. 3, pp. 609–623 (2019) doi: 10.1016/j.ipm.2018.10.014
6. Organizacion Mundial de la Salud. Datos y cifras sobre el cáncer, Página web Organización Mundial de la Salud, www.who.int/cancer/about/facts/es/
7. Rojas, R., Feldman, J.: Neural networks: A systematic introduction. Springer (1996) doi: 10.1007/978-3-642-61068-4
8. Silva, S., Almeida, J.: GPLAB - A genetic programming toolbox for MATLAB. In: Proceedings of the Nordic MATLAB Conference, pp. 273–278 (2003)
9. Sossa, H., Barrón, R., Vázquez, R. A.: New associative memories for recall real-valued patterns. Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 195–202 (2004) doi: 10.1007/978-3-540-30463-0_24
10. Sossa, H., Garro, B.A., Villegas, J., Olague, G., Avilés, C.: Evolutionary computation applied to the automatic design of artificial neural networks and associative memories. In: Advances in Intelligent Systems and Computing, Springer Berlin Heidelberg, pp. 285–297 (2013) doi:10.1007/978-3-642-31519-0_18
11. Villegas-Cortez, J., Olague, G., Aviles, C., Sossa, H., Ferreyra, A.: Automatic synthesis of associative memories through genetic programming: A first co-evolutionary approach. Applications of Evolutionary Computation, Springer Berlin Heidelberg, pp. 344–351 (2010) doi: 10.1007/978-3-642-12239-2_36
12. Villegas-Cortez, J., Olague, G., Sossa, H., Avilés, C.: Evolutionary associative memories through genetic programming. Parallel Architectures and Bioinspired Algorithms, Studies in Computational Intelligence, Springer, vol. 415, pp. 171–188 (2012) doi: 10.1007/978-3-642-28789-3_8
13. Villegas-Cortez, J.: Síntesis automática de memorias asociativas mediante programación genética. Tésis Doctoral, Instituto Politécnico Nacional, Centro de Investigación en Computación (2009)
14. Wolberg, W. H., Mangasarian, O. L.: Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In: Proceedings of the National Academy of Sciences, vol. 87, no. 23, pp. 9193–9196 (1990) doi: 10.1073/pnas.87.23.9193