ISSN 1870-4069

Identificación de características de células de cáncer de mama por medio de testores típicos

Alexis Gallegos, Dolores Torres, Francisco Álvarez, Aurora Torres

Universidad Autónoma de Aguascalientes, Departamento de Ciencias de la Computación, Aguascalientes, Aguascalientes, México

alexisEdm@gmail.com, mdtorres@correo.uaa.mx, fjalvar@correo.uaa.mx, atorres@correo.uaa.mx

Resumen. Una de las preocupaciones más importantes del mundo ha sido la salud humana, especialmente en enfermedades como el cáncer. Por esta razón, este artículo se enfoca en la aplicación de las Ciencias Computacionales, específicamente, la Selección de Subconjuntos de Características y Testores típicos para mejorar el diagnóstico de cáncer. En este caso, se procesó una base de datos de cáncer de mama. Ésta base de datos fue publicada por la Universidad de California para aprendizaje máquina. Los datos describen características del núcleo de las células obtenido de una imagen digitalizada de un aspirado con aguja fina de masa mamaría clasificando cada célula como maligna o benigna. Finalmente, el método proveerá el peso informacional de cada característica. Ésta información permitirá saber si una característica realmente describe una célula y así, clasificar nuevas instancias con la información correcta.

Palabras clave: peso informacional, testor típico, selección de subconjuntos, cáncer de mama, lógica combinacional.

Identification of Breast Cancer Cell Features by Means of Typical Testors

Abstract. One of the most significant concerns around the world has been human health, especially in diseases such as cancer. For this reason, this paper is focused on the application of Computer Science, specifically, Feature Subset Selection and Typical Testors to improve the diagnosis of cancer. In this case, a breast cancer cell database was processed. This database was published by the University of California for machine learning. The data describes features of the cell nuclei obtained from a digitized image of a fine needle aspirate of a breast mass classifying every cell as malignant or benign. Finally, the method will provide the informational weight of each feature. This information will let know if a feature actually describes a cell and then, classify new instances with the right data.

Keywords: informational weight, typical testor, subset selection, breast cancer, combinational logic.

1. Introducción

El cáncer de mama es el cáncer más común y la principal causa de muerte por enfermedad tumoral en mujeres alrededor en el mundo [1], lo que representa el 16% de los cánceres en mujeres [2]. Hoy en día, el cáncer se ha vuelto cada vez más difícil de ignorar. Cada día, nuevos estudios sobre las causas y tratamientos son publicados; sin embargo, todo coinciden que el punto crítico de estos estudios es la detección temprana [3].

De acuerdo con el Instituto Nacional de Cáncer [4], la detección de cáncer significa la comprobación de cáncer o de condiciones que pueden convertirse en cáncer en personas que no presentan síntomas.

La detección temprana es importante debido a que cuando un tejido anormal o cáncer es encontrado a tiempo, puede ser más fácil de tratar. Al momento que los síntomas aparecen, el cáncer ha comenzado a extenderse y es más difícil de tratar [4].

A pesar de la utilidad de la detección temprana, pueden existir algunos riesgos, así como los métodos utilizados. Por ejemplo, una prueba de detección puede presentar resultados falsos positivos; significa que la prueba indica la presencia de cáncer cuando no es verdad. Por otro lado, la prueba puede tener resultados falsos negativos indicando que el cáncer no está presente, aunque si lo este.

Por otra parte, el sobrediagnóstico es posible, el cual sucede cuando la prueba de detección muestra que una persona tiene cáncer, pero el cáncer es de crecimiento lento y no habría perjudicado a la persona en toda su vida [4]. Lo anterior justifica la necesidad de mejorar el diagnóstico de cáncer.

El diagnóstico clínico es un proceso cognitivo que parte del pensamiento concreto sensible. Está relacionado con la realidad objetiva; se desarrolla en el pensamiento abstracto y tiene el criterio de verdad en la práctica [5]. Involucra práctica, experiencia, reconocimiento de patrones y cálculo de probabilidad condicional, entre otros componentes. Sin embargo, el diagnóstico tiene tratamiento humano, por lo tanto, no está libre de errores que pueden causar enfermedad, daños, gastos extra e incluso la muerte, especialmente en enfermedades sensibles como el cáncer [6].

Los errores representan un estimado de 150 de cada 1000 pacientes con diagnóstico erróneo [7]. Por esta razón, el campo de la medicina es una de las áreas que pueden beneficiarse mejor de una interacción cercana con las Ciencias Computacionales y las Matemáticas para mejorar procesos como lo es el diagnóstico médico [6]. Siendo así, la razón por la que se decide aplicar métodos matemáticos integrales para apoyar el diagnóstico de enfermedades como el cáncer, en este caso, cáncer de mama.

Este artículo está dividido en tres secciones y organizado de la siguiente manera. La primera sección trata conceptos importantes en Selección de Subconjuntos de Características y testores típicos en Ciencias Computaciones, el cáncer de mama y su impacto en el mundo. La sección siguiente examina marco de trabajo del análisis, siendo una revisión de la metodología aplicada a las células de cáncer de mama. Finalmente, la tercera sección describe los resultados de la metodología y su revisión.

2. Conceptos importantes

2.1. Selección de subconjuntos de características

Normalmente, la Selección de Subconjuntos de Características (FSS, por sus siglas en inglés: Feature Subset Selection) [8] es usado para reducir la dimensionalidad [9], lo que significa que reduce el número de variables, atributos o características con las cuales se describen los objetos y encontrar su influencia en un problema. Este un método alternativo que inicia usado el conjunto de testores típicos, descartando características irrelevantes o redundantes [9, 10].

La importancia de la FSS recae en la reducción del número de características, el cual puede ayudar a disminuir el costo de adquisición de información y hacer que los modelos de clasificación sean más fáciles de entender [9, 11]. Además, el número de características podría afectar la precisión de la clasificación. Algunos autores también han estudiado la Selección de Subconjuntos de Características para el aprendizaje de clasificación [9].

Los problemas de FSS han sido estudiados con gran atención por estadísticos y comunidades de aprendizaje máquina durante años debido a la investigación entusiasta de la minería de datos [12]. Existen muchos beneficios potencias de la selección de características, como lo son [13]:

- Facilita la visualización de información y su entendimiento,
- Reduce requerimientos de medición y almacenamiento,
- Reduce tiempos de capacitación y utilización,
- Reduce la dimensionalidad para mejorar el rendimiento de la predicción.

La selección de las variables más relevantes suele ser subóptima para construir un predictor, sobre todo si las variables son redundantes [13]. Por ejemplo, el método de selección por fuerza bruta evalúa exhaustivamente todas las posibles combinaciones de las características de entrada y así encontrar el mejor subconjunto [12]. Más adelante, en las secciones 3 y 4 describirán el método de fuerza bruta aplicado a la base de datos ya mencionada con el objetivo de evaluar si una célula es maligna o benigna y calcular el peso informacional de cada característica.

2.2. Testores típicos

La teoría de testores fue formulada como una dirección científica independiente de Cibernética Matemática en los años 60 en la formada Unión de Repúblicas Socialistas Soviéticas (USSR), cuyo origen está vinculado con el uso de lógica matemática para localizar fallas en circuitos electrónicos que realizan funciones booleanas [14].

Más tarde, los testores fueron utilizados para realizar clasificación supervisada y selección de variables en problemas de geología [14, 15]. El uso dato a los testores y testores típicos para éste artículo está relacionado con la Selección de Subconjuntos de Características, cuyos precursores son Dmitriev, Zhuravlev, y colegas [15].

De este modo, un testor es un subconjunto de características que distingue objetos de diferentes clases [15]. De acuerdo con Santiesteban y Pons [10], Shulcloper [14] y

Torres [15], un testor típico es un testor al que no es posible eliminar alguna característica sin perder su estado de testor. En otras palabras, un testor típico ya está formando por el conjunto mínimo de características necesarias para asegurar la identificación de la clase a la que pertenece un objeto específico.

Los testores típicos determinan cuestiones como la evaluación del peso informacional de los rasgos y la selección de variables. Pueden reducir el espacio de representación de los objetos [10] y pueden ser usados como un conjunto de soporte para la algoritmos de clasificación [16]. En consecuencia, el objetivo de este estudio es probar que el análisis de testores puede ayudar a clasificar las células basadas en un conjunto de datos real; esto se explicara en la sección 3.

2.3. Peso informacional

El uso del peso informacional para Selección de Subconjuntos de Características es una excelente herramienta que muestra resultados tangibles [9]. El peso informacional es una puntuación, es decir, es una medida de significancia para predecir si un objeto pertenece a un grupo o a otro (clasificación) [15, 17]. Más información en la sección 4.

2.4. Cáncer de mama

El cáncer es una colección de enfermedades relacionadas que causa que algunas células del cuerpo comiencen a dividirse sin detenerse y se extiendan a tejidos cercanos. El cáncer puede generarse en casi cualquier parte del cuerpo, el cual está hecho de millones de células [18]. Se trata del resultado de mutaciones o cambios anormales en los genes que regulan el crecimiento de la célula. Normalmente, una célula crece y divide para formar nuevas células según como el cuerpo lo necesite. Cuando las células envejecen y resultan dañadas, mueren, y nuevas células las reemplazan [18, 19].

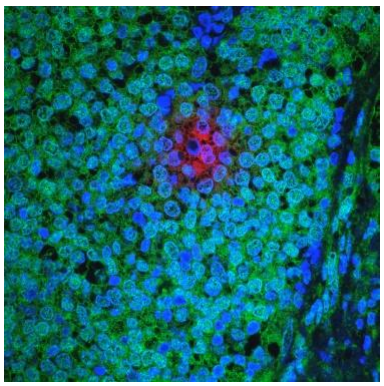

Fig. 1. Tumor invasivo de cáncer de mama [18].

Cuando el cáncer se desarrolla, el proceso celular se descompone. Las mutaciones pueden “activar” ciertos genes y “desactivar” otras en la célula. La célula modificada

adquiere la habilidad de dividirse sin ningún control u orden, lo que produce células idénticas y generan un tumor [18, 19].

En consecuencia, el cáncer de mama es un tumor maligno que se ha desarrollado de células de mama [20]. La mama está hecha de glándulas llamadas lóbulos que pueden producir leche y tubos delgados llamados ductos que llevan leche de los lóbulos al pezón, generalmente, el cáncer de mama se origina en las células de éstos lóbulos [19, 20].

El cáncer de mama tiene gran impacto en el mundo. Según la Organización Panamericana de Salud (PAHO), en América el cáncer de mama es el más común en mujeres con el 29% de los casos de cáncer. PAHO estima más de 596,000 casos nuevos y más de 142,100 muertes en la región para 2030, principalmente en Latinoamérica y el Caribe [21]. La siguiente figura muestra la incidencia de tumores malignos de mama en mujeres mayores a 20 años divididos por grupo de edad, en el año 2014:

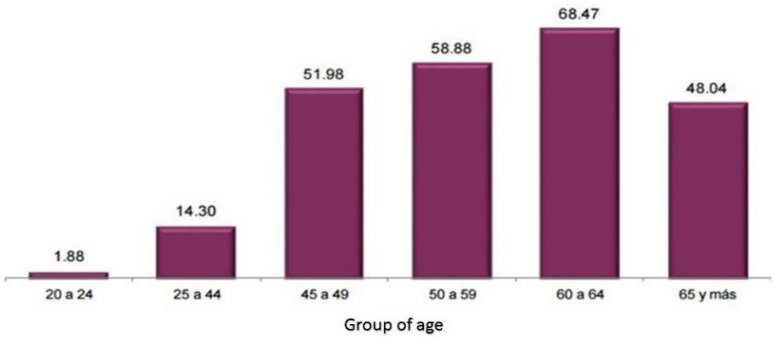

Fig. 2. Incidencia de tumores malignos de mama en mujeres mayores de 20 años dividido por grupo de edad. Por 100 mil mujeres por grupo de edad. INEGI [21].

En general, cualquier tipo de cáncer representa un impacto importante en el estado físico de la persona, su esfera emocional, un alto costo de tratamiento y puede incluso, socavar la economía de los países; así que la prevención y el diagnóstico temprano son críticos para abordar el problema [22]. Por lo tanto, este trabajo se concentra en la aplicación del Selección de Subconjuntos de Características y Testores Típicos para mejorar el diagnóstico de cáncer en células de cáncer. Las secciones siguientes se explicará el estudio realizado.

3. Marco de trabajo

La metodología general utilizada (ver Fig. 3) inicia con una Matriz de Aprendizaje (MA). La MA es la fuente de información que contiene la descripción de los objetos [14, 23]. Para éste trabajo, la MA viene de la Universidad de California y su Repositorio de Aprendizaje Máquina. La base de datos seleccionada es Diagnóstico de Cáncer de Mama de Wisconsin [24].

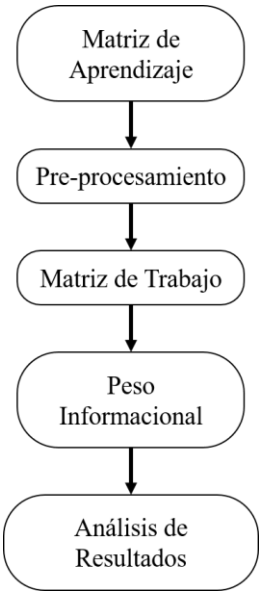

Fig. 3. Metodología general.

La base de datos contiene el diagnóstico y 10 características obtenidas de una imagen digitalizada de una aspiración de tejido de mama con aguja fina y describe las características del núcleo de la célula presentada en la imagen [25]. La imagen a continuación muestra un ejemplo de las imágenes descritas.

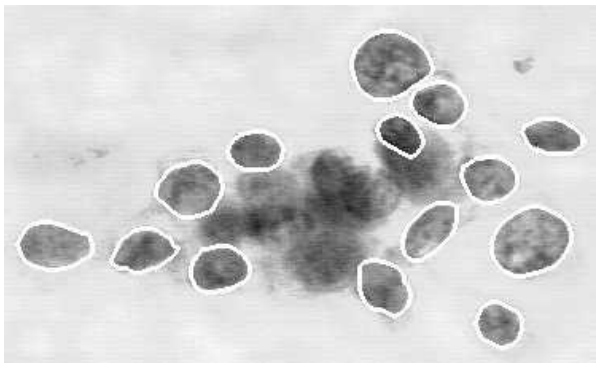

Fig. 4. Ejemplo de una imagen tomada por un sistema de visión por computadora y el contorno de la célula [26].

Las características evaluadas para cada núcleo de célula son [25, 26]:

1. Diagnóstico (M=maligno, B=benigno),
2. Radio,
3. Textura,
4. Perímetro,

5. Área,
6. Suavidad,
7. Compacidad,
8. Concavidad,
9. Puntos cóncavos,
10. Simetría,
11. Dimensión fractal.

El diagnóstico es el resultado final de la evaluación de las características de la célula con un sistema de diagnóstico de visión por computadora [26-28]. Cada célula en la base de datos tiene uno de dos posibles diagnósticos, puede ser célula maligna registrada con la letra M o benigna registrado con la letra B.

El radio de la célula fue medido promediando la longitud de los segmentos de líneas radiales definidos por el centroide de la célula y los puntos individuales en el límite de la célula. Las líneas radiales fueron definidas por Street, Wolberg y Magasarian en [26, 27] como se puede observar en la Fig. 5.

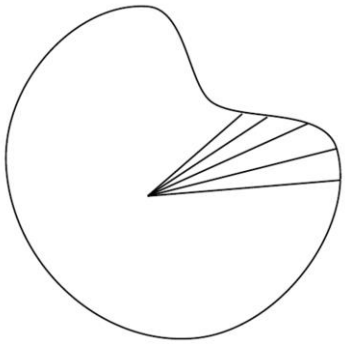

Fig. 5. Líneas radiales medidas en una célula [26].

Como se mencionó anteriormente, cada característica de la célula fue extraída por sistema de visión por computadora, por tanto, la textura fue medida encontrando la varianza en intensidades de escala de grises en los pixeles de la computadora [26, 27]. (Ver Fig. 4).

El perímetro es definido como la distancia total entre puntos individuales llamados puntos serpiente en [26]. Estos puntos individuales comprenden las líneas blancas en el perímetro de las células (ver Fig. 4).

El área es obtiene contando el número de pixeles en el interior de la línea blanca añadiendo la mitad de los pixeles en el perímetro [26].

Mientras tanto, la suavidad del núcleo de la célula se calcula midiendo la diferencia entre la longitud de una línea radial y la longitud principal que la rodea [26]. Básicamente, la suavidad es la variación local en las longitudes de radio [25].

El perímetro y el área son combinados para calcular la medida de compacidad; la cual es una medida de forma [26, 27]. La compacidad está dada por la fórmula:

$$\text{Compacidad} = \text{perímetro}^2/\text{área}.$$

Este número es minimizado por un disco circular e incrementa con la irregularidad del perímetro y aumenta también para núcleos celulares alargados, lo que puede indicar mayor probabilidad de malignidad [26].

La concavidad analiza las irregularidades de forma en el núcleo de la célula. Street, Wolberg y Mangasarian miden el número y la severidad de las concavidades y hendiduras en el núcleo de la célula. Ellos dibujan cuerdas entre cada punto blanco no adyacente y miden hasta qué punto el límite real del núcleo se encuentra en el interior de cada cuerda (ver Fig. 6).

Fig. 6. Cuerdas usadas para calcular la concavidad [27].

Los puntos cóncavos usan una medida similar a la concavidad, pero ésta característica solo mide el número, más que la magnitud, de las concavidades del contorno [26].

La simetría se obtiene encontrando la línea más larga que pase por el centro. Entonces, de acuerdo con [26], se trazan líneas perpendiculares a dicha línea para medir la diferencia de longitudes en las dos direcciones de la lineal central (ver Fig. 7).

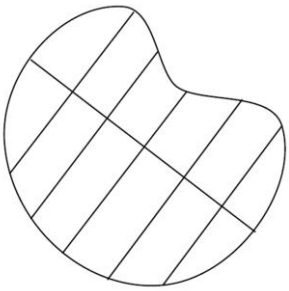

Fig. 7. Segmentos usados en el cálculo de la simetría [26].

Finalmente, la dimensión fractal es una característica de forma [27], es decir, a mayor valor corresponde a un menor contorno y por tanto a una mayor probabilidad malignidad [26]. La dimensión fractal se aproxima usando la aproximación de costa de Mandelbrot [26, 29]. El perímetro del núcleo es medido usando "reglas" cada vez más grande. Esto es, a medida que aumenta el tamaño de la regla, decrece la precisión de la medición, el perímetro observado disminuye. Ahora, trazando estos valores a una escala

logarítmica y medir la pendiente descendente da el negativo de una aproximación de la dimensión fractal [26] (ver Fig. 8).

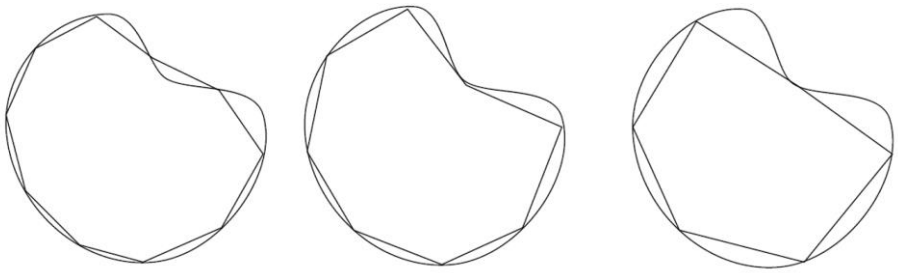

Fig. 8. Secuencia de medidas para calcular dimensión fractal [26].

La base de datos contiene un total de 569 instancias, 357 de ellas son instancias benignas y 212 instancias malignas. Esta base de datos fue pre-procesada (ver Fig. 3), esto significa que es necesario un análisis profundo de la base de datos buscando instancias duplicadas en cada clase y contradicciones (eliminar registros iguales pero diagnósticos diferentes). Consecuentemente, la base de datos debe contener instancias únicas.

El siguiente paso de la metodología requiere una matriz de trabajo, la cual se obtiene de matriz de aprendizaje pre-procesada. La matriz de trabajo se compone de datos discretizados. Cada característica es discretizada de acuerdo a la literatura del problema y el consejo de un experto, quién confirma los criterios de comparación [10].

Finalmente, el elemento principal de la metodología es el peso informacional. Para calcularlo es necesario aplicar la teoría de testores típicos mencionada en la sección 2.2. Como resumen, los testores típicos están formados por el conjunto mínimo necesario para asegurar la identificación de clases en la que un objeto específico pertenece. Para más información ver las referencias [10, 14].

4. Resultados y conclusiones

Al final del proceso, el peso informacional es calculado de acuerdo con los testores típicos encontrados. Como se puede observar en la Tabla 1, el radio y el área del núcleo de la célula tienen 50% de peso informacional. Esto significa que es posible clasificar una instancia de célula conociendo a menos una de las dos características, el radio o el área, pero el resto de las características debe conocerse debido a que obtuvieron un 100%. Por ejemplo, una instancia puede ser clasificada conociendo su textura, perímetro, suavidad, compacidad, puntos cóncavos, simetría y la dimensión fractal, pero si el radio es desconocido, el área debe conocerse. Por otro lado, si se desconoce el área, el radio debe conocerse. Finalmente, en el mejor caso se da cuando ambos valores se conocen mientras que no es posible clasificar una célula si ambos datos son desconocidos.

Table 1. Peso informacional de acuerdo con lo testores típicos.

Feature	Informational weight
Radius	50%
Texture	100%
Perimeter	100%
Area	50%
Smoothness	100%
Compactness	100%
Concave points	100%
Symmetry	100%
Fractal dimension	100%

El peso informacional se obtiene calculando un factor de procentaje que indica la frecuencia de cada variable en el conjunto de testores típicos [30]. El valor del peso informacional representa el grado de importancia de cada característica analizada en un proceso de clasificación. Un valor de 100% indica que la característica es crítica no puede ser ignorada en ningún caso.

Además, es posible que una o más características obtengan 0% de peso informacional, lo que significa que no es necesaria. Por lo tanto, el número de características se reduce y hace el problema más sencillo. Recuerde que éste es uno de los objetivos del anális.

El peso informacional puede ser validado por la teoría del problema o un especialista, de manera que la información final se apegue a la realidad. Para este experimento, un patólogo valido el peso informacional y el comportamiento de los datos.

Referencias

1. Guerra-Merino, I.: Factores pronostico del cáncer de mama en 108 mujeres menores de 36 años. Universidad Complutense de Madrid (2000)
2. CEAMEG: Cancer de Mama. Vol. 1, No. Cancer de Mama, pp. 1 (2014)
3. Canceronline: Detección Precoz de Cáncer. Available: http://www.canceronline.cl/index.php?option=com_content&view=article&id=48&Itemid=57
4. NIH: Cancer Screening. Available: http://www.cancer.gov/about-cancer/screening (2015)
5. Pérez-Guirado, N. M.: El diagnóstico médico: algunas consideraciones filosóficas (2009)
6. Lugo-Reyes, S. O., Maldonado-Colín, G., Murata, C.: Inteligencia artificial para asistir el diagnóstico clínico en medicina. Artificial Intelligence to Assist Clinical Diagnosis in Medicine, Vol. 61, No. 2, pp. 110–120 (2014)
7. Reed, K.: HealthGrades Patient Safety in American Hospitals Study. Disponible en: https://www.hospitals.healthgrades.com/
8. Wang, G., Song, Q., Sun, H., Zhang, X., Xu, B., Zhou, Y.: A Feature Subset Selection Algorithm Automatic Recommendation Method. China Journal of Artificial Inteligence Research (2013)

9. Torres, D., Ponce de León, E., Torres, A., Ochoa, A., Díaz, E.: Hybridization of Evolutionary Mechanisms for Featured Subset Selection in Unsupervised Learning. In: MICAI 2009, Advances in Artificial Intelligence, pp. 610–621
10. Santiesteban-Algaza, Y., Pons-Porrata, A.: LEX: A New Algorithm for the Calculus of all Typical Testors. Vol. 1, pp. 85–95.
11. Pelikan, M., Sastry, K., Cantú-Paz, E.: Scalable Optimization vía Probabilistic Modeling: From Algorithms to Applications. Springer (2006)
12. Deng, K.: OMEGA: On-line Memory-Based General Purpose System Classifier. Doctor Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (1998)
13. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, Vol. 3 (2003)
14. Ruíz-Shucloper, J., Alba-Cabrera, E., Lazo-Cortés, M.: Introducción a la Teoría de Testores. Departamento de Ingeniería Electrica, CINVESTAV-IPN, pp. 197 (1995)
15. Torres-Soto, M. D., Torres-Soto, A., Torres-Soto, L., Bermudez-Rosales, L., Ponce de León-Sentí, E. E.: Factores Predisponentes en Relajación Residual Neuromuscular. Research in Computing Science, Vol. 93, pp. 163–174 (2015)
16. Lias-Rodríguez, A., Pons-Porrata, A.: Un nuevo Algoritmo de Escala Exterior para el Cálculo de los Testores Típicos. Research in Computing Science, Vol. 93 (2015)
17. Cotilla, M. O.: Un Recorrido por la Sismología de Cuba. 1 ed., Cuba, Editorial Complutense, S. A. (2006)
18. National Cancer Institute: What is Cancer? Disponible en: http://www.cancer.gov/ about-cancer/understanding/what-is-cancer (2015)
19. Breastcancer.org: ¿Qué es el Cáncer de mama? Disponible en: http://www.cancer. gov/about-cancer/understanding/what-is-cancer (2014)
20. NIH: Breast Cancer - Patient Version. National Cancer Institute
21. INEGI: Estadísticas a Propósito del... Día Mundial de la Lucha contra el Cáncer de Mama. Estadisticas Nacionales, México: Instituto Nacional de Estadística y Geografía (2015)
22. INEGI: Estadísticas a Propósito del Día Mundial Contra el Cáncer. México, Instituto Nacional de Estadística y Geografía (2016)
23. Santiesteban, Y., Pons, A.: LEX: un nuevo algoritmo para el calculo de los testores tipicos. Revista Ciencias Matematicas, Vol. 21, No. 1, pp. 85–95 (2003)
24. Mangasarian, O. L., Street, W. N.: Breast cancer diagnosis and prognosis via linear programming. Operations Research, Vol. 43, No. 4, pp. 570 (1995)
25. Wolberg, W. H., Street, N., Mangasarian, O. L.: Wisconsin Diagnostic Breast Cancer (WDBC). California, Ed., USA (1995)
26. Street, W. N., Wolberg, W. H., Mangasarian, O. L.: Nuclear Feature Extraction for Breast Tumor Diagnosis. In: International Symposium on Electronic Imaging, Science and Technology, Vol. 1905, pp. 861–870
27. Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L.: Computerized breast cancer diagnosis and prognosis from fine-needle aspirates. Archives of surgery (Chicago, Ill.: 1960), Vol. 130, No. 5, pp. 511 (1995)
28. Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L.: Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, Vol. 26, No. 7, pp. 792–796 (1995)
29. Mandelbrot, B. B.: The fractal geometry of nature. New York, W.H. Freeman (1982)

30. Rodríguez-de Léon, P.: Heurística lógico combinatoria para la selección de subconjuntos de características en diabetes mellitus. Tesis (maestría en informática y tecnologías computacionales), Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Aguascalientes, Ags., México (2016)