ESTADO DEL ARTE PARA TÉCNICAS SUPERVISADAS Y NO SUPERVISADAS DE MACHINE LEARNING EN LA IDENTIFICACIÓN DE RASGOS DISTINTIVOS DE CÁNCER DE MAMA

THE STATE OF THE ART FOR SUPERVISED AND NOT SUPERVISED SKILLS OF MACHINE LEARNING IN THE IDENTIFICATION OF DISTINCTIVE FEATURES OF BREAST CANCER

María Isabel Andrade Prado[1], Camilo José Medina Barahona[1], Evelyn Carolina Calvache Pabon[1] y Esteban Rodrigo Baez Bello[1]

[1]Carrera de Ingeniería Informática y Mecánica, Facultad de Ingeniería, Corporación Universitaria Autónoma de Nariño, Cra 28 # 19-24, Pasto, Nariño

Correo electrónico para correspondencia: {isabelandrade487, camimed, gogumas291, ebaez600} @gmail.com

PROHIBIDA SU COPIA

ESTADO DEL ARTE PARA TÉCNICAS SUPERVISADAS Y NO SUPERVISADAS DE MACHINE LEARNING EN LA IDENTIFICACIÓN DE RASGOS DISTINTIVOS DE CÁNCER DE MAMA

Resumen

Siendo el cáncer de mama la causa número uno de muerte entre las mujeres alrededor del mundo, su detección temprana es de vital importancia; actualmente se recurre a mamografías, autoexamen, biopsias, etc. métodos que no dan un resultado óptimo, razón por la cual en los últimos años se ha buscado el apoyo tecnológico específicamente de machine learning con el desarrollo de algoritmos supervisados y no supervisados que ayuden a la detección y diagnóstico temprano de este padecimiento, por lo tanto, el propósito de este artículo es proporcionar a futuros investigadores el material necesario para desarrollar sus propuestas tomando los últimos avances de machine learning aplicadas al análisis de datos. Se tomaron artículos que aprovechan los diversos tipos de datos que se pueden obtener de los exámenes médicos relacionados con el cáncer y específicamente del cáncer de mama, estos datos pueden tener cualquier forma, numéricos, tipo texto o incluso imágenes, y para tratarlos debemos recurrir a diferentes métodos y algoritmos. Analizando los resultados obtenidos podemos dar una perspectiva sobre la forma más eficiente de abordar este recurso. Como resultado se evidencia el uso de métodos no supervisados en parte del proceso de identificación de etiquetas, que si bien, no es un diagnóstico como tal, es indispensable para establecer las categorías que serán entrenadas en el futuro mediante el uso de métodos supervisados. Luego notamos como el uso de técnicas como redes neuronales y máquinas de soporte vectorial son las más usadas y se expone criterios para elegir una u otra, así mismo el tipo de dato más frecuente es el que se obtiene como resultado de la mamografía de cribado que es de tipo imagen.

Palabras claves: cáncer de mama, machine learning, supervisado, no supervisado, algoritmo.

PROHIBIDA SU COPIA

Abstract

As breast cancer is the leading cause of death among women worldwide, early detection is of vital importance. Yet current methods, such as self-examination, mammography, and biopsies, do not give optimal results. This is why technological help has been sought in recent years, specifically that of machine learning, with the development of supervised and unsupervised algorithms that assist in the detection and early diagnosis of this disease. The objective of this article is to provide future researchers with the necessary material to develop their proposals, taking the latest advances in machine learning applied to data analysis. Articles were selected that exploit the various types of data that can be obtained from cancer-related medical tests, and specifically for breast cancer. This data can have any form, number, type of text, or image, and to address them we must resort to different methods and algorithms. Analyzing the results obtained can give a perspective on the most efficient way to approach this resource. As a result, the use of unsupervised methods is evident in part of the label identification process, which is not a diagnosis as such, but is indispensable to establish the categories that will be trained in the future through the use of supervised methods. Then we note how techniques such as neural networks and vector support machines are the most used, the criteria for choosing one or the other are exposed, and that the most frequent type of data obtained as a result of mammography screening is an image.

Keywords: Breast cancer, Machine Learning, Supervised, Non-Supervised, algorithm

PROHIBIDA SU COPIA

Introducción

El diagnóstico temprano de cáncer de mama es fundamental y los métodos usados actualmente cumplen parcialmente con este cometido, ofrecer un buen diagnóstico a tiempo es la mejor estrategia en el control de este padecimiento. Actualmente se ha investigado mucho sobre los aportes que el campo de Machine Learning ofrece al proceso de diagnóstico temprano de cáncer de mama, estos algoritmos abren todo un nuevo universo de posibilidades que pueden ser aprovechadas para salvar miles de vidas.

El método más usado en la detección del cáncer es la mamografía de cribado, este procedimiento da como resultado imágenes difíciles de interpretar, ya sea por la calidad de la imagen o por el nivel de especialización del radiólogo que la analice. Esto pone de manifiesto las limitaciones humanas en el procesamiento de elevadas cantidades de información, en este caso, de imágenes que pueden ser interpretados según miles o millones de variaciones, y aun el más experimentado es susceptible de pasar por alto detalles que para una maquina serian evidentes.

Teniendo esto en cuenta se plantea hacer un análisis del estado del arte de los métodos supervisados y no supervisados de algoritmos de Machine Learning para la detección de cáncer de mama. Es indispensable ofrecer un punto de partida actualizado para futuros proyectos que estén interesados en profundizar el estudio de esta necesidad. Esto nos permitirá ofrecer un criterio de elección que responda las siguientes dudas de un hipotético investigador: ¿Qué método debería usar en mi estudio?, ¿Supervisado?, ¿No supervisado?

Se plantea entonces organizar la estructura del articuló en técnicas y algoritmos supervisados en contraste con los no supervisados, analizando artículos relacionados o equivalentes de forma que permita obtener un criterio de elección en función de sus resultados, costos computacionales, efectividad y otros factores relevantes al momento de implementar un algoritmo de Machine Learning.

Técnicas no supervisadas

Dermatologist-Level Classification Of Skin Cancer With Deep Neural Networks

Si bien las redes neuronales pueden orientarse como supervisadas o no supervisadas, la naturaleza de los datos necesarios para el diagnóstico de cáncer de piel hizo decantar a los investigadores hacia el uso de técnicas no supervisadas.

El cáncer de piel es diagnosticado de forma visual, son las imágenes las que darán al especialista las primeras pistas para iniciar análisis más profundos, teniendo eso en cuenta,

PROHIBIDA SU COPIA

el artículo demuestra que es posible el uso de redes neuronales de forma no supervisada para establecer patrones que luego pueden ser analizados como potencialmente cancerígenos.

Actualmente se está usando este método para realizar diagnósticos cuando los datos que alimentan al algoritmo son visuales.Teniendo eso en mente, se extrapola esta metodología para aprovechar los datos de tipo imagen para el diagnóstico de cáncer de mama, donde se cuenta con una amplia gama de análisis que generan resultados de este tipo.

Sin embargo el algoritmo de Redes neuronales convolucionales que fue usado para esta investigación toma como un etapa la de clustering, para luego usar un núcleo entrenado que permite dar prioridad a los pixeles que tengan características cancerígenas, dándoles un valor numérico más elevado y llegar finalmente a la predicción del cáncer.

El clustering es usado para identificar enfermedades, pero es más potente al ser usado en un entorno entrenado.Se podrían usar técnicas no supervisadas para descubrir patrones en las imagines clasificadas como cancerígenas, para luego ser usadas como punto de referencia al entrenar otros algoritmos para que puedan identificarlas en muestras en vivo. (Esteva et al., 2017)

Técnicas supervisadas

Biomedical Data Analysis Toolkit

Es una herramienta de software desarrollado en java que provecha el poder de la computación en malla (Grid computing), para analizar diferentes algoritmos que se usan comúnmente en el campo biomédico. Esta herramienta está centrada en el entrenamiento de Redes Neuronales (Artificial neural network - ANN) y máquinas de soporte vectorial (Support Vector Machines – SVM), por medio de diferentes tipos de algoritmos, que permitan obtener la mejor curva ROC (Receiver Operating Characteristic) posible. (Ramos-Pollán, Guevara-López, & Oliveira, 2012)

PROHIBIDA SU COPIA

Gracias a este estudio tenemos una excelente comparación entre estos dos modelos computaciones ANN y SVM, al ser entrenados mediante diferentes algoritmos.

Tabla 1

Tipos de algoritmos usados y su descripción

Algoritmo	Descripción
ffbp	feed forward ANN trained with **backpropagation**
ffbproc	ffbp modified for ROC optimization
ffrp	feed forward ANN trained with **resilient propagation**
ffrproc	ffrp modified for ROC optimization
ffsa	feed forward ANN trained with **simulated annealing**
ffsaroc	ffsa modified for ROC optimization
ffga	feed forward ANN trained with **genetic algorithms**
ffgaroc	ffga modified for ROC optimization
csvc	cost based **Support Vector Machine**

Breve descripción de los tipos de algoritmos usados en este articuló.

Nota: Tomada de (Ramos-Pollán, Guevara-López, & Oliveira, 2012)

Tabla 2

Porcentaje de elementos correctamente clasificados de algoritmos en función del dataset

		DATASET													
	ENGINE	**bcw**	**bcwd**	**haber**	**heartsl**	**hepat**	**liver**	**mmass**	**park**	**pgene**	**pimadiab**	**spam**	**spectf**	**tictac**	**promedio**
BEST PCT	ffbp	86,59	95,00	69,82	89,52	88,63	71,33	83,73	93,47	89,09	70,50	78,76	57,95	81,10	**81%**
	ffbproc	90,17	95,06	73,93	91,41	89,11	80,85	91,74	98,28	93,38	74,24	82,90	67,18	88,06	**86%**
	ffrp	81,41	95,09	62,72	93,08	92,33	75,71	88,43	98,29	94,36	79,90	83,63	90,80	88,64	**86%**
	ffrproc	88,10	96,47	92,34	93,12	93,59	80,02	94,96	98,19	96,14	81,73	85,86	93,74	94,09	**91%**
	ffsa	97,14	92,70	82,17	90,03	94,69	80,78	95,23	97,32	88,50	83,12	80,42	90,30	84,91	**89%**
	ffsaroc	92,75	95,04	84,65	96,30	95,33	81,62	90,83	98,03	87,74	82,41	90,81	90,83	94,85	**91%**
	ffga	82,26	92,41	82,01	96,30	87,63	67,05	88,50	94,40	73,37	81,23	68,69	85,67	72,57	**82%**
	ffgaroc	86,77	96,89	81,74	92,91	87,19	76,04	92,98	96,52	80,27	83,35	78,68	86,86	80,49	**86%**
	c-svc	97,14	97,65	93,75	94,71	95,25	80,26	92,34	97,67	93,16	82,25	91,93	93,38	95,71	**93%**

PROHIBIDA SU COPIA

El mejor porcentaje de elementos correctamente clasificados en la parte de prueba del conjunto de datos o su precisión corresponde a las máquinas de vectores de soporte (SVC), también es evidente como las modificaciones a los algoritmos para obtener una mejor curva ROC tienen mayor precisión. Nota: Tomada de (Ramos-Pollán, Guevara-López, & Oliveira, 2012)

An evaluation of image descriptors combined with clinical data for breast Cancer diagnosis

Otra de las herramientas utilizadas para el diagnóstico de cáncer de mama es por medio de la evaluación de descriptores de imagen combinada con datos clínicos, en esta herramienta se pueden utilizar descriptores de imagen, datos demográficos, observaciones clínicas o combinaciones, con el objetivo de clasificar las lesiones. Para realizar un diagnóstico asistido por propósito (CADx) este se basa en sistemas de diagnóstico y entrenamiento como sifiers (MLC), en este punto son necesarios los predictores caracterizados por poseer un alto poder discriminante que permita deducir si una observación corresponde a un hallazgo maligno o no. Dentro del campo de los predictores se encuentran diferentes tipos, el primero es el que posee datos clínicos es decir datos del paciente como edad y sexo, así como también características de exámenes mamográficos (densidad mamaria y anomalías). El segundo son los descriptores de imagen, permiten caracterizar las lesiones a partir de estadísticas calculadas en las mamografías, se clasifican en dos categorías, la primera describe características generales y la segunda describe características más específicas. (Moura & López, 2013)

De esta manera, el diagnóstico de cáncer se evalúa bajo la presencia y ausencia de datos clínicos. La combinación de características de la imagen y los datos clínicos demuestran un mejor rendimiento en el uso de descriptores. El éxito de este estudio hace creer que puede ser también exitoso en otras aplicaciones de imágenes médicas y visión por ordenador en general. A futuro se pretende mejorar la clasificación de datos por medio de combinación y descriptores de imagen generando un rendimiento óptimo en presencia o ausencia de datos clínicos.

PROHIBIDA SU COPIA

Multivariate Feature Selection of Image Descriptors Data for Breast Cancer with Computer-Assisted Diagnosis

Se busca obtener un modelo multivariado para clasificar lesiones tumorales benignas y malignas utilizando un diagnóstico asistido por ordenador con un algoritmo genético en el entrenamiento y los conjuntos de datos de prueba de las características de imagen de mamografía. Se realizó una búsqueda multivariable para obtener modelos predictivos con diferentes enfoques, con el fin de comparar y validar los resultados. Los modelos multivariados se construyeron utilizando las estrategias de bosque aleatorio (Random Forest), centroide más cercano y K-vecino más cercano (K-NN) como función de costo en un algoritmo genético aplicado a las características de las bases de datos públicas BCDR. Los resultados sugieren que las dos características de descriptor de textura obtenidas en el modelo multivariado tienen una capacidad de predicción similar o mejor para clasificar el resultado de los datos en comparación con el modelo multivariado compuesto por todas las características, de acuerdo con su valor de aptitud. Este modelo puede ayudar a reducir la carga de trabajo de los radiólogos y presentar una segunda opinión en la clasificación de las lesiones tumorales.

El método más usado en la detección del cáncer es la mamografía de cribado pero dado que este no es uno de los más efectivos porque pueden ser difíciles de interpretar por la calidad de la imagen y el nivel de especialización de los radiólogos hay muchas técnicas computacionales con enfoque a algoritmos y diagnósticos asistidos por computador (CAD) para mejorar la detección del cáncer de mama.

Estos algoritmos usan características descriptivas del cáncer, porque no se cuenta con un estándar de clasificación. Con ello se ha propuesto:

“Patel propuso el método de segmentación de imagen con la detección de la micro calcificación y un algoritmo adaptativo.

R. Ramani propuso un nuevo enfoque para clasificar las masas en la imagen de mamografía utilizando un enfoque de extracción de características. Las características extraídas de la mamografía mediante el uso de Symlet, SVD y el histograma ponderado se utilizaron para entrenar un modelo de clasificación (Naive Bayes, bosque aleatorio y algoritmos de redes neuronales) y clasificar los tumores como benignos o Propusieron un sistema de CAD para detectar el cáncer de mama en mamogramas a través de optimización de la inteligencia de enjambre por redes neuronales que detectan el cáncer.

Karahaliou presenta una propuesta interesante que utiliza un enfoque de procesamiento digital de imágenes combinado con un enfoque de extracción de características, proponiendo un método de mejora de contraste basado en wavelet para encontrar los descriptores de textura extraídos de mamografías que pueden contribuir al diagnóstico de cáncer. Para validar la capacidad de las características de la textura para discriminar malignas de lesiones benignas, se usó un clasificador k-Nearest Neighbor (KNN).

Por último, para conocer el desempeño, se consideró el área bajo la curva (AUC) y la característica operativa del receptor (ROC), mostrando resultados prometedores en CAD de cáncer de mama que pueden contribuir a la reducción de biopsias innecesarias."

Tomando en cuenta que todas las propuestas no estudian el comportamiento de los tumores en sí, cada autor tiene su método de extracción, se busca un modelo de clasificación multivariante que se base en el análisis de las características de imágenes de ultrasonido, historia clínica, segmentación de lesiones y training set, mediante un algoritmo genético; lo cual daría una mayor exactitud de predicción. (Galván-Tejada et al., 2017)

Referencias

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

Galván-Tejada, C. E., Zanella-Calzada, L. A., Galván-Tejada, J. I., Celaya-Padilla, J. M., Gamboa-Rosales, H., Garza-Veloz, I., & Martinez-Fierro, M. L. (2017). Multivariate Feature Selection of Image Descriptors Data for Breast Cancer with Computer-Assisted Diagnosis. Diagnostics, 7(1), 9. https://doi.org/10.3390/diagnostics7010009

Moura, D. C., & López, M. A. G. (2013). An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. International Journal of Computer Assisted Radiology and Surgery, 8(4), 561-574. https://doi.org/10.1007/s11548-013-0838-2

Ramos-Pollán, R., Guevara-López, M. A., & Oliveira, E. (2012). A software framework for building biomedical machine learning classifiers through grid computing resources. Journal of Medical Systems, 36(4), 2245-2257. https://doi.org/10.1007/s10916-011-9692-3

PROHIBIDA SU COPIA