Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

Tipo de artículo: Artículo original
Temática: Inteligencia artificial
Recibido: 01/02/2015 | Aceptado: 15/03/2015

Clasificación de células cervicales mediante el algoritmo KNN usando rasgos del núcleo

Cervical cell classification by means of the KNN algorithm using nucleus' features

Solangel Rodríguez Vázquez [1*], Andy Vidal Martínez Borges [2], Juan Valentín Lorenzo Ginori [3]

[1] Universidad de las Ciencias Informáticas. Km 2½ carretera San Antonio de los Baños, Rpto. Torrens, La Lisa, Ciudad de la Habana. svazquez@uci.cu
[2] Empresa de tecnología para la defensa, XETID. Km 2½ carretera San Antonio de los Baños, Rpto. Torrens, La Lisa, Ciudad de la Habana. avmartinez@xetid.cu
[3] Centro de Estudios de Electrónica y Tecnologías de la Información, Universidad Central "Marta Abreu" de Las Villas, Cuba. juanl@uclv.edu.cu

[*] Autor para correspondencia: svazquez@uci.cu

Resumen

La prueba de Papanicolaou, es un examen de pesquisa ginecológica que permite apreciar cambios en la morfología de las células del cuello uterino. Dicho estudio puede alertar sobre patologías tan frecuentes en las mujeres como el cáncer del cuello del útero. El análisis de este tipo de imágenes es importante en la generación de diagnósticos y en las investigaciones que se llevan a cabo, por lo que se hace necesario el desarrollo de nuevas técnicas que efectúen un análisis práctico de las muestras. La búsqueda por similitud es uno de los procedimientos más frecuentes en problemas que involucran el procesamiento de datos, una variante consiste en la búsqueda de los k-vecinos-más-cercanos (kNN). En este trabajo, se propone el uso del clasificador kNN y de una de las distancias utilizadas por el mismo para dar solución al problema de la clasificación de las células del cuello uterino en las clases normal y anómala, basándose solamente en las características extraídas de la región del núcleo. A partir del estudio realizado, entre las distancias manhattan, euclidiana y mahalanobis y teniendo en cuenta para la evaluación las medidas F, AUC, predictividad negativa y media H, se comprobó que manhattan mostró un buen desempeño manteniendo valores de 97.1% de AUC. Los resultados obtenidos indican una reducción respecto a la tasa de falsos negativos en la prueba de Papanicolaou. Se utilizó la media H con el propósito de comparar los resultados de kNN respecto a otras investigaciones, obteniendo un 92.33% con respecto a las mismas.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

Palabras clave: células del cuello del útero; clasificación; *kNN*; núcleos; prueba de Papanicolaou

Abstract

The Pap test is a test of gynecological screening that allows appreciating changes in the morphology of the cells of the cervix. This study can alert on such frequent pathologies in women as cancer of the cervix. The analysis of these kinds of images is important in the generation of diagnostic and the investigations that carried out, so that developing new techniques that made a practical analysis of the samples is necessary. Similarity search is one of the most common procedures in problems involving processing of data, an alternative to solve this problem is the kNN search (k-Nearest Neighbors). In this paper, the kNN classifier was used together with a specific distance function, to provide a solution to the real problem associated with the classification of cervical cells in normal and abnormal classes, the features used for classification were in this case based solely on information extracted from the nuclei region. From the study, among the manhattan distance, Euclidean and Mahalanobis and considering measures for evaluating F, AUC, negative predictivity and H-mean was found that manhattan performed well holding 97.1% values of AUC. The results indicate a reduction compared to the rate of false negative Pap test. H-mean with the purpose of comparing the results of other investigations regarding kNN, obtaining 92.33% with regard thereto.

Keywords: *cervical cell; classification; kNN; nuclei; Pap test.*

Introducción

La mayoría de los cánceres de cérvix se originan en el revestimiento de las células del cuello uterino. Estas células no se tornan en cáncer de repente, sino que las células normales del cuello uterino se transforman gradualmente en precancerosas, y pueden con el tiempo convertirse en células malignas. En las ciencias médicas se usan varios términos para describir estos cambios precancerosos, incluyendo neoplasia intraepitelial cervical (CIN, por sus siglas en inglés), lesión intraepitelial escamosa (SIL, por sus siglas en inglés) y displasia. Estos cambios se pueden detectar mediante la prueba de Papanicolaou y se pueden tratar para prevenir el desarrollo de cáncer (2014). Estas células anormales muestran ciertas alteraciones precancerosas preliminares que se denominan displasia o neoplasia cervical intraepitelial (NCI).

La displasia y la NCI se clasifican en grados: leve, moderado o grave. La displasia leve (NCI 1) generalmente se resuelve por su cuenta. La displasia moderada (NCI 2) y grave (NCI 3) indican alteraciones más peligrosas. La prueba de Papanicolaou, la cual también se denomina frotis de Papanicolaou o evaluación de la citología cervical, consiste en analizar al microscopio los frotis de células cervicales para detectar cambios anormales en las células del cuello uterino. A partir de sus resultados, es posible prevenir la mayoría de los cánceres cervicales mediante la detección de

los cambios anómalos de las células del cuello uterino (precánceres) para que estos puedan ser tratados antes de que tengan la oportunidad de convertirse en cáncer cervical. Este tipo de pruebas ofrecen la mejor oportunidad para detectar el cáncer de cuello uterino en una etapa temprana cuando es más probable que el tratamiento sea eficaz. Esta prueba fue descrita por primera vez en el año 1924 y desde finales de los años cuarenta ya se hizo notar que podían detectarse los cambios neoplásicos del epitelio del cérvix uterino en las muestras citológicas. Su objetivo es la detección precoz de las lesiones premalignas o localizadas, antes de que aparezcan síntomas, y en estadios en los que el tratamiento es más eficaz. (de Les Corts 1994)

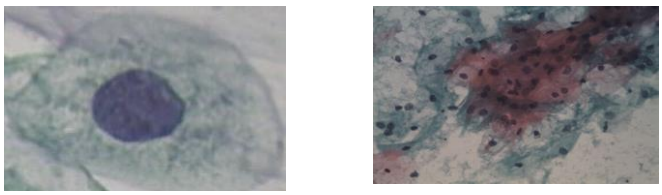
La sensibilidad y la especificidad de la citología de Papanicolaou han sido motivo de discusión y los valores aportados por los diferentes autores son muy variables. Se plantea que en el caso de la sensibilidad, sus valores varían entre el 55% hasta el 90% y que la especificidad de la prueba puede ser considerada más alta, en torno a un 90%, aunque otros autores han dado datos inferiores y situados alrededor del 60% (de Les Corts 1994). La proporción de falsos positivos y negativos depende en gran medida de la calidad de la toma de la muestra citológica, de su lectura e interpretación y de los posibles errores en el procesamiento de la misma. Un criterio de la calidad muy importante es la observación en la muestra de células endocervicales. La continuidad y seguimiento de estos criterios mejora la calidad de la muestra citológica. Los errores derivados de una mala interpretación de las muestras, traen consigo consecuencias no favorables a la persona tanto para los falsos negativos como para los falsos positivos. En el caso de los falsos negativos, estos pueden tener secuelas importantes debidas a la no detección de los casos de displasia o carcinoma in situ que pueden evolucionar hacia lesiones más avanzadas. Los efectos secundarios de los falsos positivos son los clásicos de este tipo de pruebas (ansiedad, molestias derivadas de las pruebas diagnósticas o de seguimiento) pero pueden ser de poca importancia dadas las características de la prueba. Otro problema importante es el riesgo del tratamiento innecesario o muy agresivo de algunas displasias (de Les Corts 1994).

La interpretación de estas muestras se apoya fundamentalmente en el reconocimiento visual de los cambios en las zonas más importantes de las células (núcleo y citoplasma). Sin embargo, este proceso es tedioso, lleva mucho tiempo y es propenso a errores de procedimiento debido al alto grado de complejidad de las imágenes y al cansancio de los analistas ya que el mismo se realiza de forma manual y se pueden llegar a analizar cientos de muestras en un día. Por esta razón, se realizan importantes esfuerzos para dar solución al problema del análisis de la prueba de Papanicolaou empleando técnicas de visión computacional.

Varios métodos han sido propuestos para la clasificación de las células en las imágenes de la prueba de Papanicolaou y que se refieren a las técnicas tales como clasificadores bayesianos (Riana and Murni 2009), redes neuronales

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

artificiales (Mat-Isa, Mashor et al. 2008), máquinas de vectores soporte (SVM) (Huang, Chan et al. 2007) y vecinos más cercanos (Marinakis, Dounias et al. 2009). Debe tenerse en cuenta que la mayoría de estos métodos utilizan imágenes pre-segmentadas que contienen solo una célula, por lo que la segmentación correcta del núcleo y el citoplasma es factible (Figura 1 (a)). En las imágenes que contienen grupos de células (Figura 1 (b)), la detección de la frontera del citoplasma es un problema difícil debido a la superposición de las células. Sin embargo, la detección y segmentación de los núcleos de las imágenes que contienen células superpuestas y agrupaciones de células ha sido abordado con éxito en varios estudios (Plissiti, Nikou et al. 2011a) (Plissiti, Nikou et al. 2011b). Las muestras que se toman en la prueba de Papanicolaou corresponden con la figura 1(b), por lo que se hace necesario el estudio de la clasificación de las células en función de las características del núcleo ya que es más fácil la detección de fronteras en los núcleos que en los citoplasmas.



(a) (b)

Figura 1. (a) Imagen de una célula simple, (b) imagen de células superpuestas

Los métodos que se refieren a la clasificación de imágenes de Papanicolaou se basan en el cálculo de las características extraídas tanto del área núcleo como del citoplasma, un ejemplo de las mismas se muestra en la tabla 1. Estas características se basan generalmente en características de forma y la intensidad de los objetos de interés. A pesar de ello, las características calculadas no presentan la misma capacidad de discriminación (Plissiti and Nikou 2012). Para la determinación del conjunto de características más eficaz que se utiliza como entrada en un clasificador, se han propuesto algunos esquemas de selección de características que se refieren a algoritmos genéticos (Marinakis, Dounias et al. 2009) y al enjambre de partículas de optimización (Marinakis, Marinaki et al. 2008). En el caso de la presente investigación es importante este análisis, debido a que la entrada para el clasificador es una matriz $[m(casos) \times n(rasgos)]$ donde los casos son cada una de las células a evaluar) con cada uno de los rasgos extraídos de las imágenes. (Lorenzo-Ginori, Curbelo-Jardines et al. 2013)

Sobre la base de lo antes mencionado, se puede concluir que existe un problema abierto: lograr la clasificación correcta de las células empleando solo la información extraída de los núcleos, lo cual comprende entre otras cosas: determinar el subconjunto de rasgos con mejor capacidad de discriminación, y realizar la selección del clasificador

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

que ofrezca mejores resultados. En algunas investigaciones como es el caso de (Plissiti and Nikou 2012) se hace uso de los nueve rasgos del núcleo y técnicas como spectral clustering y fuzzy C-means con reducción de la dimensionalidad, a diferencia de la presente investigación que se dirige hacia el uso de *kNN* sin reducción donde solo se utilizan cinco rasgos de los nueve como *área, perímetro, diámetro corto, diámetro más largo y la redondez*. Esta selección de rasgos persigue demostrar que, a partir de un conjunto primario de rasgos geométricos, es posible realizar de forma efectiva la clasificación binaria de imágenes en la prueba de Papanicolaou, mediante el algoritmo *kNN*.

Tabla 1. Características extraídas de las imágenes (Plissiti and Nikou 2012)

RASGOS DEL CITOPLASMA		RASGOS DEL NÚCLEO	
1.	Área	1.	Área
2.	Brillo	2.	Brillo
3.	Diámetro corto	3.	Diámetro corto
4.	Diámetro más largo	4.	Diámetro más largo
5.	Elongación	5.	Elongación
6.	Redondez	6.	Redondez
7.	Perímetro	7.	Perímetro
8.	Máxima[1]	8.	Máxima[1]
9.	Mínima[1]	9.	Mínima[1]
10.	Posición del Núcleo		
11.	Tamaño Núcleo/Citoplasma		

[1]El número de pixeles con el valor de intensidad máximo / mínimo en una zona de 3x3 de la zona específica.

Materiales y métodos

En el caso de la presente investigación de los 9 rasgos pertenecientes al núcleo solo se utilizaron los 5 rasgos mencionados anteriormente debido a que estos a consideración de los autores son los rasgos básicos que caracterizan estas células (Velezmoro and Villafuerte 2001). La extracción de rasgos es uno de los pasos fundamentales en el procesamiento de imágenes debido a que mientras mejor sea la selección de los atributos más acertada será la clasificación final de las células. Esto hace que la adecuada selección de los rasgos sea una de las limitantes en las investigaciones para la clasificación de imágenes, debido a que si no se cuenta con los rasgos apropiados los resultados obtenidos por el clasificador no tendrán la calidad que se necesita (Lorenzo-Ginori, Curbelo-Jardines et al. 2013). Para el caso de la presente investigación, las matrices de rasgos utilizadas fueron extraídas de las imágenes ya

previamente segmentadas (Figura 2(b)) pertenecientes a la base de datos Herlev presentada en (Jantzen, Norup et al. 2005).

Para ello se hizo uso de la herramienta Matlab y de las funciones de procesamiento de imágenes propias de la misma. De esta forma se desarrolló un algoritmo que toma de la base de datos Herlev aleatoriamente un 80% de las imágenes y le extrae de ellas los rasgos antes mencionados. A continuación, se crea una matriz con los rasgos de las imágenes seleccionadas que será la matriz a utilizar para el entrenamiento del clasificador. Posteriormente se extraen los rasgos del 20% de las imágenes restantes y se conforma la matriz de rasgos que se utilizará para realizar las pruebas. Estas matrices tienen como característica fundamental que poseen vectores de rasgos diferentes lo que posibilita una evaluación correcta del funcionamiento del clasificador. Se realizaron varias corridas con dicho algoritmo en las que el 20% se fue "rotando" de modo que los conjuntos de entrenamiento y de prueba fueron modificados por cada iteración.

(a) (b)

Figura 2. Células anómalas. (a) Imagen RGB, (b) Imagen Segmentada en la cual la región correspondiente al núcleo aparece resaltada en azul claro

Como se mencionó anteriormente la presente investigación hace uso de imágenes pertenecientes a la base de datos de referencia Herlev. Esta base de datos consta de 917 imágenes, donde cada una está compuesta por una única célula (Figura 1 (a)), y las muestras se distribuyen irregularmente en siete clases. Tres de ellas son consideradas como normales, y las cuatro restantes como anómalas. La descripción detallada de la base de datos se representa en la tabla 2.

Los subconjuntos dimensionales extraídos servirán como base de entrenamiento y de pruebas en el clasificador implementado, para el cual se realiza una comparación entre las distancias de Mahalanobis, euclidiana y Manhattan, utilizadas en el algoritmo de clasificación *kNN*. Los métodos basados en vecindad son fundamentalmente dependientes de la distancia, estas poseen algunas propiedades como la no negatividad, identidad, simetría y desigualdad del triángulo que son importantes en el momento de su utilización. Además las reglas de clasificación por vecindad están basadas en la búsqueda de un conjunto de los k prototipos más cercanos al patrón a clasificar. (Verbiest, Cornelis et al. 2012)

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

Tabla 2. Distribución de las células en la base de datos Herlev, de imágenes celulares de la prueba de Papanicolaou

NORMAL	# DE CÉLULAS
Epitelio Escamoso Superficial	74
Epitelio Escamoso Intermedio	70
Epitelio Columnar	98
TOTAL	**242**
ANORMAL	# DE CÉLULAS
Carcinoma_in_situ	182
Light_dysplastic	146
Moderate_dysplastic	197
Severe_dysplastic	150
TOTAL	**675**

El algoritmo *kNN* es uno de los clasificadores no paramétricos más usados en el campo de minería de datos y de aprendizaje automático. En general, un clasificador dispone de datos de entrenamiento (un sistema de decisión X que consiste en casos con sus valores de atributos y su clase) e intenta predecir la clase $d(t)$ de un caso objetivo t. *KNN* resuelve este problema buscando los k casos de X más similares a t y asignando t a la clase mayoritaria entre estos k casos. A continuación se describen los pasos a seguir para la implementación de dicho algoritmo: (Duda, Hart et al. 2012). En el algoritmo *kNN* el conjunto de los datos de entrenamiento incluye, además de las propiedades multidimensionales utilizadas para el reconocimiento, clasificadores para predecir la clase de los datos de entrada. Para clasificar, utiliza un tipo de distancia, como, por ejemplo, la distancia euclidiana, con la que determina todas las distancias entre el punto a clasificar y todos los puntos del conjunto de entrenamiento. Con las distancias calculadas determina los k vecinos más cercanos y, según el tipo de la clase para determinar, asigna el punto a una de ellas. (Figuras 3 y 4)

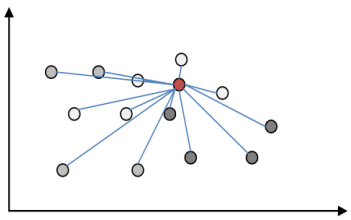


Figura 3: Distancias entre el punto a clasificar al conjunto de entrenamiento.

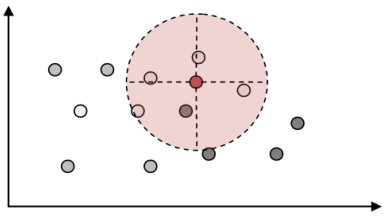


Figura 4: Algoritmo *kNN* en dos dimensiones con $k = 5$.

En un espacio de representación bidimensional y una serie de prototipos de una misma clase representados en él, dado un patrón cualquiera X, si se consideran los k prototipos más próximos a X, estos estarán localizados en un círculo

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

centrado en X (figura 4). Es bueno tener en cuenta los rangos en los que se mueven los valores numéricos. Esto evita que atributos con valores muy altos tengan mucho mayor peso que atributos con valores bajos, se normalizarán dichos valores con la ecuación 1.

$$\frac{x_{if} - min_f}{Max_f - min_f} \quad (1)$$

En esta ecuación x_{if} será el valor i del atributo f, siendo min_f el mínimo valor del atributo f y Max_f el máximo. Por otro lado, el algoritmo permite dar mayor preferencia a aquellos ejemplares más cercanos al que se desea clasificar. En ese caso, en lugar de emplear directamente la distancia entre ejemplares, se utilizará la ecuación 2.

$$\frac{1}{1 + d(x_i, x_j)} \quad (2)$$

La principal dificultad de este método consiste en determinar el valor de k, ya que si toma un valor grande se corre el riesgo de hacer la clasificación de acuerdo a la mayoría (y no al parecido), y si el valor es pequeño puede haber imprecisión en la clasificación a causa de los pocos casos empleados para la decisión. Por lo que para que la k no sea tan determinante en el momento de la clasificación para dicha investigación, se decidió hacer uso de la inversa de su distancia ($\frac{1}{d}$) con respecto al ejemplo a clasificar.

Métricas para medir distancia

Una característica importante e interesante de *kNN* es que el método puede cambiar radicalmente sus resultados de clasificación sin modificar su estructura, solamente cambiando la métrica utilizada para hallar la distancia. Por lo tanto, los resultados pueden variar tantas veces como métodos de hallar distancia entre puntos haya. La métrica debe seleccionarse de acuerdo al problema que se desee solucionar. La gran ventaja de poder variar métricas es que para obtener diferentes resultados el algoritmo general del método no cambia, únicamente el procedimiento de medida de distancias. La distancia es el criterio de comparación principal usado en los métodos basados en vecindad, por eso es conveniente mencionar algunas de las diferentes formas usadas para su medición (Aggarwal, Hinneburg et al. 2001). En las ecuaciones (3), (4) y (5) se muestran respectivamente las expresiones para la distancia euclidiana, Manhattan y Mahalanobis. En esta última, S es la matriz de covarianza de los vectores de rasgos.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (4)$$

$$d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)} \quad (5)$$

Métricas de evaluación

La toma de decisiones clínicas es un proceso extremadamente complejo en el que deberá finalmente ser valorada la utilidad para el manejo del paciente de cualquier prueba diagnóstica. En este contexto, es imprescindible conocer detalladamente la exactitud de las distintas pruebas diagnósticas, es decir, su capacidad para clasificar correctamente a los pacientes en categorías o estados en relación con la enfermedad, típicamente dos, por ejemplo: estar o no estar enfermo o manifestar una respuesta positiva o negativa a la terapia (de Ullibarri Galparsoro and Fernández 1998). Para que una prueba se incluya en la práctica médica rutinaria es necesario que sea capaz de reducir la incertidumbre asociada con una determinada situación clínica. Siempre que una condición clínica y el resultado de la prueba diagnóstica encaminada a resolverla puedan plantearse en términos de dicotomía (presencia o ausencia de enfermedad, positiva o negativa), la exactitud de la prueba puede definirse en función de su sensibilidad y especificidad diagnósticas. Sin embargo, con mucha frecuencia los resultados de las pruebas diagnósticas están distribuidos en una escala continua, por lo que es necesario seleccionar un punto de corte o valor límite adecuado que permita resumir estos resultados en dos categorías: positivo y negativo. En la presente investigación se utilizaron los índices de efectividad: sensibilidad, especificidad, predictividad positiva y negativa, así como la tasa de clasificación correcta para evaluar el rendimiento del clasificador.

Otra forma de evaluar el rendimiento del clasificador utilizado fue por medio de las curvas ROC. Dicha curva es una representación gráfica de la sensibilidad frente a (1 – especificidad) para un sistema clasificador binario según se varía el umbral de decisión. Se denomina umbral de decisión a aquel que decide si un caso x, a partir del vector de salida del clasificador, pertenece o no a cada una de las clases. Los índices de efectividad empleados en esta investigación para evaluar las distancias en el clasificador fueron el área bajo la curva ROC (AUC), la predictividad negativa (Pn) y las medidas *F* y *H*, dadas por:

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

$$F - measure = \frac{2 \times (Precision \times S_e)}{Precision + S_e} \quad (6)$$

$$H - mean = \frac{2 \times (S_p \times S_e)}{S_p + S_e} \quad (7)$$

El rendimiento de la clasificación es evaluado a través del desempeño que muestra el clasificador haciendo uso de las distancias, y el rendimiento final se obtiene por la media de los resultados, después de la ejecución de este experimento con 3 modelos diferentes, en un esquema de validación cruzada de 5 iteraciones. Para la implementación del algoritmo *kNN* se hizo uso de las distancias: euclidiana, Mahalanobis y Manhattan. Con el fin de estimar la capacidad de discriminación de los parámetros internos para el funcionamiento del algoritmo se realizaron una serie de experimentos donde se comparó el rendimiento del mismo con las tres distancias y su resultado con los mismos conjuntos de datos. Como se mencionó anteriormente se realizaron una serie de corridas del algoritmo *kNN*, en cada una de ellas se empleó una base de entrenamiento con los rasgos extraídos de 734 imágenes (de ellas 540 enfermas y 194 sanas) y una base de pruebas compuesta por una matriz que contiene: sanas con 47 imágenes, enfermas con 134 imágenes con el objetivo de evaluar dos subconjuntos diferentes y así obtener valores que muestren la eficiencia real del clasificador.

Resultados y discusión

Para la comparación entre los resultados obtenidos en el uso de las distancias, fueron utilizados los mismos conjuntos de datos. Además, de los conjuntos de datos obtenidos en el particionamiento, se utilizaron las tres particiones que mejores resultados mostraron en cuanto a las medidas Pn, AUC, la medida *F* y la media *H*.

Comparación del algoritmo con las distintas distancias

Tabla 3. Clasificación de las variables para el conjunto de datos 1

No. Modelo	*Distancias*	*Área ROC*	*F-Measure*	*H-Mean*	*% Clasificación*
1	*Euclidean*	0.949	0.916	0.875	0.917
1	*Mahalanobis*	0.91	0.915	0.856	0.917
1	*Manhattan*	**0.951**	**0.917**	**0.884**	0.917

Tabla 4. Matriz de confusión, resultado del cálculo de los indicadores de desempeño

	Valores de las variables		
Matriz de Confusión:	*Euclidean* *# de Imágenes Reconocidas*	*Mahalanobis* *# de Imágenes Reconocidas*	*Manhattan* *# de Imágenes Reconocidas*
VP	128	130	127
FP	9	11	8

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

FN	6	4	7
VN	38	36	39
Indicadores de desempeño			
Se	0.96	0.97	0.95
Sp	0.81	0.77	0.83
Pp	0.93	0.92	0.94
Pn	0.86	**0.90**	0.85
Rc	0.92	0.92	0.92

Tabla 5. Clasificación de las variables para el conjunto de datos 2

No. Modelo	*Distancias*	*Área ROC*	*F-Measure*	*H-Mean*	*% Clasificación*
2	*Euclidean*	**0.976**	**0.939**	**0.923**	0.939
2	*Mahalanobis*	0.962	0.899	0.847	0.90
2	*Manhattan*	**0.976**	**0.939**	**0.923**	0.939

Tabla 6. Matriz de confusión, resultado del cálculo de los indicadores de desempeño

	Valores de las variables		
Matriz de Confusión:	*Euclidean* *# de Imágenes Reconocidas*	*Mahalanobis* *# de Imágenes Reconocidas*	*Manhattan* *# de Imágenes Reconocidas*
VP	128	127	128
FP	5	11	5
FN	6	7	6
VN	42	36	42
Indicadores de desempeño			
Se	0.96	0.95	0.96
Sp	0.89	0.77	0.89
Pp	0.96	0.92	0.96
Pn	**0.88**	0.84	**0.88**
Rc	0.94	0.90	0.94

Tabla 7. Clasificación de las variables para el conjunto de datos 3

No. Modelo	*Distancias*	*Área ROC*	*F-Measure*	*H-Mean*	*% Clasificación*
3	*Euclidean*	**0.95**	**0.926**	**0.872**	0.928
3	*Mahalanobis*	0.875	0.83	0.65	0.845
3	*Manhattan*	**0.95**	**0.926**	**0.872**	0.928

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 95-109

Tabla 8. Matriz de confusión, resultado del cálculo de los indicadores de desempeño

	Valores de las variables		
	Euclidean	*Mahalanobis*	*Manhattan*
Matriz de Confusión:	*# de Imágenes Reconocidas*	*# de Imágenes Reconocidas*	*# de Imágenes Reconocidas*
VP	131	130	131
FP	10	24	10
FN	3	4	3
VN	37	23	37
Indicadores de desempeño			
Se	0.98	0.97	0.98
Sp	0.79	0.49	0.79
Pp	0.93	0.84	0.93
Pn	**0.93**	0.85	**0.93**
Rc	0.93	0.85	0.93

Los resultados de la clasificación se comportan en este algoritmo entre un 84 y 93% de predictividad negativa, de la misma manera que la predictividad positiva y el área bajo la curva ROC se mantienen entre rangos de valores que permiten validar la eficiencia del clasificador empleado para cada uno de los conjuntos de datos. Los valores obtenidos de acuerdo a las medidas *F* y *H* de igual forma se mantienen entre un 91-92% y 87-92% respectivamente, lo que muestra el nivel de efectividad del clasificador. Al comparar los valores obtenidos al emplear la distancia euclidiana y la distancia Manhattan, se observó que estos poseen valores muy similares entre sí y mejores con respecto a la distancia de Mahalanobis. Por otra parte, al comparar los tiempos de ejecución del algoritmo entre la distancia euclidiana y la distancia Manhattan, se observó que el de la primera es significativamente mayor, por lo que se decidió seleccionar el uso de la distancia Manhattan en el clasificador *kNN*. Este experimento se realizó a través del uso de funciones propias de la herramienta Matlab, donde al realizar algunos cálculos se evidencia el costo computacional existente entre ambas distancias.

Un análisis comparativo entre los resultados obtenidos en esta investigación con los resultados expuestos en (Plissiti and Nikou 2012), muestra un mejor desempeño del método *kNN* con la distancia Manhattan respecto a las técnicas *spectral clustering* y *fuzzy C-means* en términos de la medida *H-Mean* (Media Armónica)(tabla 9).

Tabla 9. Desempeño de la clasificación en términos de *H-mean* y cantidad de rasgos utilizados, entre la actual investigación y la investigación (Plissiti and Nikou 2012)

	Rasgos del Núcleo

	# Rasgos	*H-Mean (%)*
kNN (Distancia Manhattan)	5	92.33
Spectral Clustering	9	88.77
Fuzzy C-Means	7	90.58

Conclusiones

En este trabajo se ha presentado una aproximación a la clasificación binaria de células cervicales (es decir, en normales y anómalas), basada en la utilización del algoritmo *kNN* empleando solo rasgos propios de los núcleos celulares. Fueron utilizadas las distancias: Mahalanobis, euclidiana y Manhattan. Para obtener el mejor rendimiento posible entre dichas distancias se utilizó el algoritmo IBK del Weka, el cual permitió evaluar las diferencias para este tipo de clasificación entre las distancias y arrojó como resultado que la distancia Manhattan fue la más adecuada a utilizar para la clasificación de células cervicales.

Los resultados obtenidos muestran un mejor desempeño en comparación con los reportados en (Plissiti and Nikou 2012), en cuanto a los resultados de la clasificación. Esto se debe a que, a través de la media *H* es posible evaluar el comportamiento de la tasa de falsos negativos, mientras mayor sea el porciento de la media *H*, menor será la tasa de falsos negativos lo que brinda un buen desempeño en la realización de la prueba de Papanicolaou. Como dirección de trabajo futuro, se implementarán otros clasificadores para realizar un análisis comparativo del desempeño de estos, conjuntamente con los utilizados en esta investigación. De igual forma, se introducirán otros rasgos y se realizará una selección de éstos basada en su efectividad, con el propósito de reducir la dimensionalidad de las matrices de rasgos sin afectar significativamente el desempeño de los clasificadores. A más largo plazo, se investigará sobre el proceso de clasificación en varias clases para las imágenes de la prueba de Papanicolaou.

Referencias

AMERICAN CANCER SOCIETY (2014) Cáncer de cuello uterino: detección temprana y prevención. American Cancer Society. Disponible en: http://www.cancer.org/acs/groups/cid/documents/webcontent/002580-pdf.pdf

AGGARWAL, C., HINNEBURG, A. AND KEIM, D. (2001) On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Van den Bussche, J. and Vianu, V. (eds), Database Theory — ICDT 2001. Springer Berlin Heidelberg, pp. 420-434.

DE LES CORTS, T. (1994) Cribado del cáncer de cuello de útero, Medicina Clínica, 102, 80-84.

DE ULLIBARRI GALPARSORO, L. AND FERNÁNDEZ, P. (1998) Curvas ROC, Atención Primaria en la Red, 5, 229-235.

DUDA, R.O., HART, P.E. AND STORK, D.G. (2012) Pattern classification. John Wiley & Sons.

HUANG, P.-C., et al. (2007) Quantitative Assessment of Pap Smear Cells by PC-Based Cytopathologic Image Analysis System and Support Vector Machine. In Zhang, D. (ed), Medical Biometrics. Springer Berlin Heidelberg, pp. 192-199.

JANTZEN, J., ET AL. (2005) Pap-smear Benchmark Data For Pattern Classification. Proc. NiSIS 2005. Nature inspired Smart Information Systems (NiSIS), Albufeira, Portugal, pp. 1-9.

LORENZO-GINORI, J.V., et al. (2013) Cervical Cell Classification Using Features Related to Morphometry and Texture of Nuclei. In, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer, pp. 222-229.

MARINAKIS, Y., DOUNIAS, G. AND JANTZEN, J. (2009) Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification, Computers in Biology and Medicine, 39, 69-78.

MARINAKIS, Y., MARINAKI, M. AND DOUNIAS, G. (2008) Particle swarm optimization for pap-smear diagnosis, Expert Systems with Applications, 35, 1645-1656.

MAT-ISA, N.A., MASHOR, M.Y. AND OTHMAN, N.H. (2008) An automated cervical pre-cancerous diagnostic system, Artificial Intelligence in Medicine, 42, 1-11.

PLISSITI, M. AND NIKOU, C. (2012) Cervical Cell Classification Based Exclusively on Nucleus Features. In Campilho, A. and Kamel, M. (eds), Image Analysis and Recognition. Springer Berlin Heidelberg, pp. 483-490.

PLISSITI, M.E., NIKOU, C. AND CHARCHANTI, A. (2011) Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images, Pattern Recognition Letters, 32, 838-853.

PLISSITI, M.E., ET AL. (2011) Automated Detection of Cell Nuclei in Pap Smear Images Using Morphological Reconstruction and Clustering, IEEE Transactions on information technology in biomedicine, 15, 233–241.

RIANA, D. AND MURNI, A. (2009) Performance evaluation of Pap smear cell image classification using quantitative and qualitative features based on multiple classifiers. International Conference on Advanced Computer Science and Information Systems, ACSIS.

VELEZMORO, G.A.B. AND VILLAFUERTE, D.F. (2001) Factores de riesgo que pronóstican el hallazgo de citologías cervicales anormales en dos poblaciones: mujeres de obreros de construcción civil vs. mujeres control en la posta médica "Construcción Civil" ESSALUD, de junio a septiembre del 2000. Facultad de Medicina Humana. Universidad Nacional Mayor de San Marcos, Lima, Perú, pp. 67.

VERBIEST, N., CORNELIS, C. AND HERRERA, F. (2012) Selección de Prototipos Basada en Conjuntos Rugosos Difusos. Proceedings of XVI Congreso español sobre Tecnologías y Lógica (ESTYLF2012). pp. 638-643.