Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

Tipo de artículo: Artículo original
Temática: Reconocimiento de patrones
Recibido: 21/02/2015 | Aceptado: 29/02/2016

Selección de rasgos en muestras citológicas usando información heurística

Feature selection on pap-smear data using heuristic information

Jairo Rojas Delgado [1*]

[1] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños km 2 ½. Reparto Torrens. Boyeros. La Habana. C.P.: 19370. CUBA

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

Resumen

En este trabajo se investiga un método de selección de rasgos que permita clasificar las muestras celulares tomadas del cuello del útero en cancerosas o no. El cáncer en el cuello del útero es el segundo tipo de cáncer más difundido entre las mujeres y la investigación de métodos informáticos en función de la correcta clasificación de muestras tiene un impacto directo en la calidad de vida de los pacientes. Se investiga un conjunto de rasgos seleccionado teniendo en cuenta su importancia y la capacidad de un grupo de clasificadores de prueba para apreciar dichos rasgos. Los clasificadores de prueba empleados son: HCM, FCM, LS y WKNN. Los resultados demuestran que la utilización de dicho conjunto de rasgos arroja por lo general buenos resultados de clasificación.

Palabras clave: clasificador, selección de rasgos, muestra citológica, potencial discriminador

Abstract

This paper presents a method of feature selection to classify cell samples taken from the cervix into cancerous or not cancerous cells. Cancer of the cervix is the second most widespread cancer in women and the investigation of computational methods based on the correct classification of samples has a direct impact on the life's quality of patients. A set of selected features is investigated taking into account the importance of these features and the ability of a test group of classifiers to assess those traits. The test classifiers used are: HCM, FCM, LS and WKNN. The results demonstrate that the use of that set of features generally yields good classification results.

Keywords: *classification, discriminator potential, feature selection, pap-smear*

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

Introducción

El cáncer en el cuello del útero es el segundo tipo de cáncer más difundido entre las mujeres con más de 500 mil casos cada año (Paavonen, 2007). Mediante el método de Papanicolaou, es posible obtener muestras celulares del cuello del útero que son procesadas por cito-técnicos en un laboratorio para el diagnóstico de cáncer. Esto es posible mediante el uso de un microscopio, con el cual los expertos pueden observar visualmente cambios precancerosos y condiciones anormales en la estructura de las células antes de que progresen a formas más avanzadas de la enfermedad (Meisels and Morin, 1997).

La prueba de Papanicolaou se ha convertido en el método más empleado para la detección de cáncer en el cuello del útero en etapas tempranas (Chang et al., 2009). Debido a esto, en los últimos años se ha incrementado la necesidad de contar con expertos bien entrenados en el área para aumentar la capacidad de procesamiento de los laboratorios. De hecho, la aplicación masiva de esta prueba ha disminuido significativamente la incidencia de la mortalidad del cáncer invasivo en los países desarrollados (Plissiti and Nikou, 2012). Según (Chen et al., 2014) la incidencia de la mortalidad del cáncer del cuello del útero decreció un 47.8 por ciento luego de la aplicación masiva de esta prueba en Taiwán de 1995 hasta el 2006.

Diferenciar correctamente los tipos de células obtenidas mediante el método de Papanicolaou es un trabajo difícil y lento incluso para los cito-técnicos mejor entrenados. Por esto, los expertos emplean un procedimiento estándar referido como The Basheda System (TBS) (Solomon et al., 2002). A pesar de que muchas de las reglas y criterios definidos por TBS se encuentran claramente descritos, la evaluación de estas reglas y criterios varía subjetivamente de experto en experto (Bak et al., 2004). Esta peculiaridad contribuye a la aparición de diagnósticos incorrectos con graves implicaciones para los pacientes y para los hospitales que deben invertir recursos extras en la corrección de estos diagnósticos.

Con el objetivo de disminuir las tasas de diagnósticos incorrectos se ha introducido un grupo de tecnologías novedosas que involucran a los procesos de preparación de la muestra, control de calidad de los preparados, etc. Entre los dispositivos comerciales que actualmente se encuentran disponibles puede mencionarse la mejora de los portaobjetos para disminuir los errores en las muestras (Dawson, 2004; Kardos, 2004) y la mejora del control de la calidad en los laboratorios mediante la repetición de los preparados (Mango, 1996). Sin embargo, la mayoría de estos dispositivos son incapaces de asistir a un diagnóstico objetivo puesto que no ofrecen variables cuantitativas para eliminar la evaluación subjetiva del cito-técnico: errores de interpretación y discrepancias inter-observador (Chen et al., 2014).

En los últimos 20 años se han llevado a cabo un grupo de investigaciones relacionadas con esta temática, especialmente en la Universidad de Aegan en cooperación con el Hospital Universitario de Herlev ubicado en Grecia. La presente investigación está construida sobre los resultados de dichos proyectos investigativos. Durante las investigaciones llevadas a cabo en el Hospital Universitario de Herlev se han construido dos versiones de una base de datos con propósitos de clasificación. La primera de ellas fue elaborada por (Byriel., 1999) y la segunda por (Jantzen et al., 2009). En (Byriel., 1999) se demostró que los algoritmos de agrupamiento podían obtener errores de clasificación por debajo del 5 por ciento para este tipo de problema de clasificación, obteniendo los mejores resultados con un *Adaptative Network based Fuzzy Inference System* (ANFIS).

Posteriormente en (Martin, 2003) se utiliza *K-Fold Cross-Validation* para medir los errores de una serie de clasificadores propuestos y se realiza selección de rasgos usando recocido simulado con el error de los clasificadores como heurística para dirigir la búsqueda. Este hecho demostró que el proceso de selección de rasgos se encuentra intrínsecamente ligado con el clasificador a emplear tal como lo referencia (Norup, 2005). (Martin, 2003) realiza pruebas con clasificadores inductivos usando los métodos: *Hard C-mean clustering* (HCM), *Fuzzy C-mean clustering* (FCM) y *Gustafson-kessel clustering* (GK) en ambas versiones de la base de datos y comparando los resultados de dichos clasificadores antes y después de la selección de rasgos.

En el año 2005, ante el surgimiento de nuevos métodos de clasificación Jonas Norup propone un grupo de técnicas para clasificar los objetos de la base de datos de Herlev comparando sus resultados con los que ya se habían obtenido anteriormente. (Norup, 2005) no realiza ningún tipo de selección de rasgos y hace un grupo de anotaciones referentes al hecho de que según las descripciones médicas, los rasgos relacionados con el tamaño, la forma, el área y el brillo de las células eran buenos discriminantes; sin hacer ningún análisis posterior. En (Marianakis et al., 2009) se propone un esquema de selección de rasgos híbrido basado en algoritmos genéticos y un algoritmo de clasificación basado en los K vecinos más cercanos. El esquema propuesto emplea una configuración básica y simple obteniendo buenos resultados de clasificación en la medida que se incrementa la precisión del error de validación cometido en el proceso de selección de rasgos.

En (Plissiti and Nikou, 2012) se propone la utilización de los rasgos basados exclusivamente en las características del núcleo celular debido a que los métodos de segmentación de imágenes no lograban distinguir eficientemente el área citoplasmática de las células. Para ello se examinó un grupo de métodos de reducción dimensional no lineal, *Kernel PCA, Isomap, Locally Linear Embedding, Laplacian Eigenmaps*, probados en dos clasificadores no supervisados: *Spectral clustering* y *Fuzzy C-means*.

En (Plissiti and Nikou, 2012) aparentemente se ignoran algunos trabajos previos como los de (Poulsen and Pedron, 1995) donde se desarrolla un método para el trabajo con muestras donde el solapamiento celular es elevado como para segmentar las células individuales, o el trabajo de (Sobrevilla et al., 2003) donde propone un algoritmo para la

detección de las mejores áreas de interés en las imágenes de muestras celulares del cuello del útero. Recientemente, en (Gençtav et al., 2012) se emplea un algoritmo de agrupamiento jerárquico para la segmentación celular, los resultados obtenidos muestran la efectividad de la propuesta incluso en imágenes obtenidas de muestras inconsistentes, pobre contraste y solapamiento. Posteriormente en (Song et al., 2014) se propone un nuevo método de segmentación celular basado en una *Convolution Neural Network* obteniéndose una precisión del 94.5 por ciento en la detección del núcleo y el citoplasma celular.

En (Curbelo, 2012) se implementan un grupo de clasificadores como las máquinas de soporte vectorial pero sin realizar selección de rasgos. Posteriormente en (Rodríguez-Vázquez and Martínez-Borges, 2015) se implementa un algoritmo de clasificación empleando máquinas de soporte vectorial utilizando solamente los rasgos del núcleo.

Como puede apreciarse, hasta el momento el conocimiento acerca de la importancia o confusión introducida por los rasgos que describen la base de datos de Herlev es insuficiente. Más allá de un grupo de anotaciones basadas en criterio de expertos, las técnicas de selección de rasgos empleadas hasta el momento son el resultado de métodos de búsqueda heurísticos basados en argumentaciones técnicas de segmentación de imágenes que ya han sido solventadas por las nuevas investigaciones. Se hace necesario contar con información certera en este campo.

Tal como se señala en (Martin, 2003), no existe un clasificador perfecto que sea capaz de eliminar el ruido introducido por todos los rasgos a tratar, ni capaz de aprovechar toda la información que aportan. La selección del conjunto de rasgos que minimiza el error de un clasificador está ligado estrechamente al propio clasificador. Pero esta idea, aparentemente entra en contradicción con las anotaciones de (Norup, 2005) al subrayar el poder discriminatorio de aquellos rasgos relacionados con el tamaño, forma, área y brillo de las células.

En este contexto puede identificarse el siguiente problema: ¿Qué efecto tiene emplear una combinación de rasgos que incluya aquellos de mayor importancia en el proceso de clasificación de muestras celulares del cuello del útero?

El objetivo perseguido por la presente investigación es desarrollar un método de selección de rasgos que permita calcular una combinación de características que incluya aquellas de mayor importancia informacional y permita clasificar los casos de la base de datos de Herlev.

Materiales y métodos o Metodología computacional

La base de datos de Herlev consiste en 917 muestras distribuidas inequívocamente en 7 clases diferentes abordadas exhaustivamente en (Jantzen et al., 2009). Cada muestra se encuentra descrita por 20 rasgos. Cada imagen fue segmentada y examinada por dos expertos del Hospital Universitario de Herlev y descartada en caso de cualquier desacuerdo con el diagnóstico. En la Tabla (1) se presenta un resumen de los 20 rasgos presentes en la base de datos de Herlev relacionados con el número y el nombre de la característica utilizados.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

Tabla 1. Rasgos presentes en la base de datos de Herlev

No.	Rasgo	Nombre
1	Área del núcleo	NArea
2	Área del citoplasma	CArea
3	Núcleo / Citoplasma ratio	N/C
4	Brillo del núcleo	NCol
5	Brillo del citoplasma	CCol
6	Diámetro corto del núcleo	NShort
7	Diámetro largo del núcleo	NLong
8	Elongación del núcleo	NElong
9	Redondez del núcleo	NRound
10	Diámetro corto del citoplasma	CShort
11	Diámetro largo del citoplasma	CLong
12	Elongación del citoplasma	CElong
13	Redondez del citoplasma	CRound
14	Perímetro del núcleo	NPerim
15	Perímetro del citoplasma	CPerim
16	Posición del núcleo	NPos
17	Máxima del núcleo	NMax
18	Mínima del núcleo	NMin
19	Máxima del citplasma	CMax
20	Mínima del citoplasma	CMin

Medición del error cometido en el proceso de clasificación

En función de obtener una tasa certera de la precisión de un proceso de clasificación cualquiera es necesario establecer una medida del error cometido. Existen varios métodos que han sido aplicados en la actualidad con éxito por ejemplo en (Rodríguez-Vázquez and Martínez-Borges, 2015) se emplean las medidas AUC, medida F, predictividad negativa y media H y en (Marianakis et al., 2009) se utiliza *Root Mean Squared Error* y el error promedio. En sentido general la medición del error promedio resulta ineficiente cuando una misma medición puede ser producida por diferentes tipos de errores. El error promedio ha sido empleado ampliamente por investigaciones precedentes y es la elección más evidente para poder contrastar los modelos de clasificación propuestos.

Para el cálculo del error promedio se utiliza la cantidad de objetos clasificados de anormales siendo normales o falsos positivos (FP); la cantidad de objetos clasificados de normales siendo anormales o falsos negativos (FN); la cantidad de objetos clasificados como normales siendo normales o verdaderos negativos (VN); la cantidad de objetos clasificados como anormales siendo anormales o verdaderos positivos (VP). El error promedio (EP) que se calcula según Ecuación (1).

$$EP = \frac{FN + FP}{FN + FP + VP + VN} \quad (1)$$

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

K-fold cross-validation

Normalmente el conjunto de datos se encuentra dividido en un subconjunto de entrenamiento y otro subconjunto de validación, por lo general disjuntos. El conjunto de datos de entrenamiento se usa para construir el clasificador por lo que utilizarlo para calcular el error promedio cometido traería como consecuencia una medición inconsistente y es por lo que se utiliza el conjunto de datos de validación. En la práctica mientras más datos de validación se posean, mejor será la estimación del error cometido; y mientras mayor sea la cantidad datos de entrenamiento, mejor se adaptará el clasificador al modelo a construir. *K-fold cross-validation* descrito en (Bishop, 1995) provee un mecanismo para utilizar los datos cuando su disponibilidad es limitada.

Según (Ruíz-Shulcloper et al., 1994) dado un conjunto finito de objetos M, sobre los cuales se define un conjunto de predicados $x_1, ..., x_n, x_{n+1}$; al último predicado se le denomina predicado fundamental y su formulación en la práctica es desconocida; siendo su función evaluar la pertenencia de un objeto a una clase K_i, $i = 1, ..., r$; y al resto de los predicados se les denomina rasgos.

Definición 1.

Se denomina descripción estándar de un objeto, al n-tuplo: $I(O) = (x_1(O), ..., x_n(O))$ donde $X_i(O) \in A_i$ es el valor del rasgo X_i en el objeto O, A_i es el conjunto de todos los valores admisibles de X_i, para $i=1, ..., n$.

El algoritmo k-fold cross-validation consta de dos pasos.

1. Construcción del modelo y prueba. La idea es encontrar k particiones de M de tal forma que se escoja temporalmente una de esas particiones para probar el clasificador y el resto para construirlo.
2. Validación cruzada. Escoger una de k particiones puede hacerse de k formas. La validación cruzada construye el modelo y lo prueba de las k posibles formas, calculando el error total del modelo como el promedio de los errores de cada prueba.

La distribución de los objetos en cada partición se realiza de manera aleatoria, pero realizando un muestreo estratificado para mantener una correcta proporción del rasgo objetivo en cada segmento. Debido a que el proceso de construcción del modelo posee una serie de pasos aleatorios, la repetición del proceso no necesariamente arroja los mismos resultados por lo que se realiza *k-fold cross-validation* varias veces calculando el promedio de los errores cometidos en cada iteración.

Selección de rasgos

Según (Martin, 2003), un algoritmo de clasificación perfecto sería capaz de corregir la pobreza de los datos de entrada, pero tal algoritmo no existe. Una forma de conseguir un subconjunto de rasgos adecuado es probar con todas

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

las combinaciones posibles, sin embargo, para 20 rasgos sería el equivalente a probar 2^{20} subconjuntos de características con consecuencias negativas en el rendimiento computacional de cualquier sistema.

Existen varias metodologías para realizar selección de rasgos. Como regla aceptada estos métodos pueden clasificarse en basados en filtros, basados en envoltura y embebidos (Lazar et al., 2012). En la literatura especializada se reporta el uso de las redes neuronales (Verikas and Bacauskiene, 2002; Niazi et al., 2004; Saeys), métodos basados en optimización de enjambres de partículas (Bello et al., 2007), algoritmos genéticos (Marinakis et al., 2009; Hajnayeb et al., 2011) y otros.

Un enfoque ampliamente utilizado en los problemas de selección de rasgos y para el cálculo de la importancia informacional de los rasgos es la teoría de testores (Lazo-Cortes et al., 2001). Esta es una rama de la lógica matemática que surgió en la Unión Soviética a finales de 1950 investigada por (I.A. Cheguis, 1958). A mediados de 1960, Y. I. Zhuravlev adaptó el concepto de testor al reconocimiento de patrones y sobre la base del mismo introdujo los conceptos y formas de calcular los pesos informacionales de rasgos y objetos. Las siguientes definiciones de testor y testor típico fueron introducidas por (A.N. Dmitriev, 1966).

Definición 2

Sea M_1 el conjunto de todos los objetos pertenecientes a la clase K_1, y M_2 el conjunto de todos los objetos pertenecientes a la clase K_2. El conjunto de rasgos $T= \{x_1, ..., x_s\}$ de la muestra de objetos M se denomina testor para $(M_1, M_2) = M$, si después de eliminar de M todos los rasgos excepto los de T no existe objeto alguno en M_1 igual a M_2.

Definición 3

Un testor T se llama irreducible o típico, si al eliminar cualquier columna de T deja de ser testor para $(M_1, M_2) = M$.

Medida de la importancia informacional de los rasgos

Dada la definición de testor típico, se tiene un conjunto de rasgos que son imprescindibles para discriminar las clases. Según (A. N. Dmitriev, 1966) es natural suponer, que, si un rasgo aparece en muchos testores típicos, resulta más difícil prescindir de él en el proceso de clasificación, o sea, es más importante. Sobre esta idea el autor formula la definición de peso informacional de un rasgo como la frecuencia relativa de aparición de dicho rasgo en el conjunto de testores típicos.

Formalmente, sea τ el conjunto de todos los testores típicos que posee cierta matriz de entrenamiento de un problema de reconocimiento de patrones y sea τ_i el conjunto de testores típicos en los que aparece el rasgo x_i, entonces el peso informacional de dicho rasgo viene dado por la Ecuación (2).

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

$$P(x_i) = \frac{|\tau(i)|}{|\tau|} \text{ para } i = 1, ..., n \qquad (2)$$

A pesar de que la definición anterior resulta intuitiva, no tiene en cuenta la longitud de los testores involucrados en el proceso. Por esto se redefinirá de acuerdo al autor la expresión de importancia informacional de un rasgo según Ecuación (3).

$$D(x_i) = \alpha \times P(x_i) + \beta \times L(x_i) \text{ con } \alpha, \beta > 0 \text{ y } \alpha + \beta = 1 \qquad (3)$$

En la Ecuación (3) α y β son parámetros que ponderan la participación de $P(x_i)$ y $L(x_i)$. $L(x_i)$ es una magnitud que depende de la longitud de los testores en los que aparece el rasgo x_i y viene dado por la Ecuación (4).

$$L(x_i) = \frac{\sum_{t \in \tau(i)} \frac{1}{|t|}}{|\tau(i)|} \qquad (4)$$

Medida de la confusión introducida por los rasgos

De la definición de testor se deriva que, si una combinación de rasgos confunde al menos un par de objetos de clases diferentes, no es testor (Ruíz-Shulcloper et al., 1994). Pudiera pensarse que la cantidad de pares de objetos que se confunden es directamente proporcional a lo que le falta a dicha combinación de rasgos para ser testor; una magnitud que expresa cuanto confunde dicho rasgo durante el proceso de clasificación. Según (Lazo-Cortes et al., 2001) la confusión o error introducido por un rasgo en una muestra de objetos *M* es el número de pares de valores iguales que pertenecen a objetos de clases diferentes. Dicha magnitud viene dada por la Ecuación (5).

$$e(x_i) = |(O, O')| O \in K_1, O' \in K_2, C_i(O, O') = 1| \qquad (5)$$

Donde C_i es un criterio de comparación de semejanza entre objetos para el rasgo x_i.

Procedimiento general para la selección del mejor conjunto de rasgos

El procedimiento para la selección del mejor conjunto de rasgos que a continuación se describe está basado en los siguientes aspectos: en primer lugar, existen un conjunto de rasgos que conforman el núcleo de un modelo y en segundo lugar el error de clasificación se encuentra estrechamente ligado a la capacidad discriminatoria del clasificador empleado.

En el siguiente algoritmo, el conjunto de rasgos que conforman el núcleo del modelo a representar por el clasificador es seleccionado haciendo uso de la teoría de testores y ajustado posteriormente con el clasificador. Este ajuste se realiza de forma empírica mediante recocido simulado (Mitchel, 1997) empleando el error del clasificador como heurística del algoritmo.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

Algoritmo 1

1. Calcular el conjunto de testores típicos.
2. Calcular la importancia informacional de los rasgos y la confusión asociada.
3. Seleccionar el conjunto de rasgos que conforman el núcleo del modelo a representar.
4. Ajustar el conjunto de rasgos del núcleo del modelo a representar con el clasificador de prueba.

Resultados y discusión

En este trabajo se utiliza un grupo de clasificadores de prueba con el objetivo de establecer comparaciones entre ellos y de esta manera validar el poder discriminatorio de los conjuntos de rasgos propuestos. Entre los clasificadores a investigar se encuentran: *Hard C-mean clustering* (HCM), *Fuzzy C-mean clustering* (FCM), *Least Square method* (LS) (Krose, 1996) y *K-Nearest Neighbors* (KNN).

Cálculo del conjunto de testores típicos

Existen varios algoritmos para el cálculo de testores pudiéndose clasificar en: algoritmos de escala exterior y algoritmos de escala interior. Los algoritmos de escala exterior se caracterizan por recorrer todos los subconjuntos pertenecientes al conjunto potencia de rasgos de manera exhaustiva. Los algoritmos de escala interior por otro lado se centran en encontrar un grupo de condiciones que garanticen que determinados rasgos conforman un testor y por tanto son más complejos de implementar.

En este trabajo se emplea el algoritmo BT para la selección del subconjunto mínimo de rasgos. Un estudio de la literatura revela que existen otros algoritmos como por ejemplo LEX (Alganza and Porrata, 2003) que posee mayor eficiencia computacional que BT. Sin embargo, en el caso de la presente investigación la eficiencia computacional en términos de complejidad temporal y espacial sobrepasa el alcance planteado en el objetivo inicial.

Para la evaluación rasgo a rasgo se emplea una función de diferencia teniendo en cuenta la distribución natural de los datos en la base de datos de *Herlev*. Obsérvese en la Tabla (2) cómo la media aritmética del rasgo referente a la ratio entre el área del núcleo y el área del citoplasma, coincide para cada una de las siete clases de la base de datos con las descripciones médicas de los expertos. En ese mismo escenario, aunque la distribución de las siete clases es desconocida, la desviación estándar brinda una indicación de la superposición o separación entre clases.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

Tabla 2. Coincidencia de la media aritmética del rasgo Ratio N/C de la base de datos de Herlev y las observaciones médicas dadas por los expertos

Clase	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Media	0.01	0.03	0.35	0.27	0.38	0.49	0.60
Desviación	0.01	0.01	0.10	0.10	0.12	0.14	0.13
Observación	MP	P	M	M	G	MG	MG

[1] MP (Muy Pequeño), P (Pequeño), M (Mediano), G (Grande), MG (Muy Grande).

El objetivo es emplear la información asociada a las siete clases de la base de datos para realizar una clasificación de objetos normales y anormales. La función de diferencia quedaría enunciada formalmente mediante la Ecuación (6).

$$\varphi_i(O_j, O_k) = \begin{cases} 1 & \text{si} \quad \vartheta_i(O_j) \neq \vartheta_i(O_k) \\ 0 & \text{en otro caso} \end{cases} \tag{6}$$

Donde $\varphi\ (O_j, O_k)$ es la función de comparación del rasgo i entre los objetos O_j y O_k. La función $\vartheta(O_p)$ determina la clase a la cual pertenece el objeto O_p mediante lógica difusa y viene dada por la Ecuación (7).

$$\vartheta_i(O_p) = \{j | \max_{1<=j<=7}(-\frac{1}{dev_{ij}} \times |x_i(O_p) - mean_{ij}| + 1)\} \tag{7}$$

Donde dev_{ij} expresa la desviación típica correspondiente al rasgo i para la clase j; $mean_{ij}$ expresa la media aritmética correspondiente al rasgo i para la clase j y $x_i(O_p)$ es el valor del rasgo i perteneciente al objeto O_p.

En la Tabla (3) se listan los resultados obtenidos del cálculo de testores: importancia informacional de los rasgos de la base de datos de Herlev y la confusión que introducen. Con estos datos y siguiendo el procedimiento general descrito en el Algoritmo 1 es posible calcular el conjunto de rasgos irreducibles del sistema.

En la Tabla (4) se muestran los rasgos que mayor cantidad de información aportan en el proceso de clasificación y que no deben ser obviados.

En la Figura (1) puede apreciarse cómo aquellos rasgos que poseen un mayor ratio entre la importancia informacional y la confusión que introducen en el proceso de clasificación, coinciden con aquellos que conforman el conjunto de rasgos irreducibles

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

Tabla 3. Importancia informacional y confusión introducida por los rasgos de la base de datos de Herlev

No.	Rasgo	Importancia	Confusión
1	Área del núcleo	0.1418	0.0947
2	Área del citoplasma	0.1760	0.1133
3	Núcleo / Citoplasma ratio	0.2723	0.1127
4	Brillo del núcleo	0.2580	0.1980
5	Brillo del citoplasma	0.2857	0.2196
6	Diámetro corto del núcleo	0.1697	0.1173
7	Diámetro largo del núcleo	0.2007	0.0946
8	Elongación del núcleo	0.1841	0.3398
9	Redondez del núcleo	0.1742	0.2619
10	Diámetro corto del citoplasma	0.2528	0.1011
11	Diámetro largo del citoplasma	0.1957	0.1226
12	Elongación del citoplasma	0.2820	0.2367
13	Redondez del citoplasma	0.2656	0.1330
14	Perímetro del núcleo	0.1625	0.0930
15	Perímetro del citoplasma	0.2204	0.1290
16	Posición del núcleo	0.2024	0.3057
17	Máxima del núcleo	0.2103	0.1016
18	Mínima del núcleo	0.2060	0.1195
19	Máxima del citplasma	0.1923	0.1229
20	Mínima del citoplasma	0.1881	0.1156

Tabla 4. Conjunto de rasgos irreducibles.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	X	X				X			X	X		X	X			X		X	X

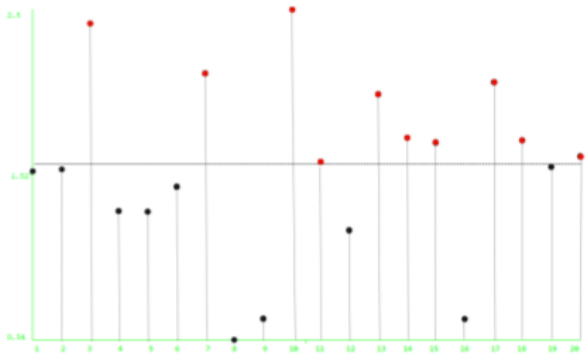

Figura 1. Representación gráfica del ratio entre la importancia informacional y confusión introducida por los rasgos.

Resultados del clasificador HCM

Los resultados para HCM serán presentados usando el algoritmo diseñado para FCM con un factor difuso q ≈ 1, específicamente *q=1.01* que en la práctica debe dar un resultado similar. Primeramente, debe determinarse la cantidad de agrupaciones a usar. Esto se realizó realizando mediciones del error cometido de diferentes cantidades de

agrupaciones empleando todos los rasgos mediante *10-fold cross-validation* y 20 iteraciones. Los resultados demuestran que, a mayor cantidad de agrupaciones, menor es el error, por lo que se determina emplear 50 agrupaciones. Los rasgos fueron ajustados empleando 50 iteraciones del algoritmo recocido simulado a una temperatura de 0.1, calculando en cada iteración el error de clasificación mediante *2-fold cross-validation*.

En la Tabla (5) se muestran las mediciones del error cometido empleando todos los rasgos, los rasgos extraídos por (Martin, 2003), los rasgos referentes al núcleo, propuestos por (Plissiti and Nikou, 2012), los rasgos extraídos en correspondencia al procedimiento general descrito en el Algoritmo 1 y el conjunto de rasgos irreducibles.

Tabla 5. Error cometido por el clasificador HCM empleando distintos conjuntos de rasgos.

No.	Medición	Conjunto de rasgos	Error
1	Todos	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20	8.4624
2	(Martin, 2003)	1,2,3,4,5,6,7,8,11,12,14,15,16,17,18,19,20	7.759
3	(Plissiti and Nikou, 2012)	1,4,6,7,8,9,14,17,18	11.4613
4	Algoritmo 1	2,3,4,6,7,10,11,13,14,17,19,20	7.5245
5	Irreducibles	2,3,7,10,11,13,14,17,19,20	8.9313

Resultados del clasificador FCM

Para probar el clasificador FCM primeramente debe elegirse el factor difuso a emplear. Como se demuestra en (Martin, 2003) existe una dependencia directa entre este parámetro y los resultados de clasificación por lo que debe ajustarse con cuidado. Para el cálculo del error cometido se emplea *10-fold cross-validation* y 20 iteraciones. En la Tabla (6) se muestra el error cometido por este clasificador.

Tabla 6. Error cometido por el clasificador FCM empleando distintos conjuntos de rasgos.

No.	Medición	Conjunto de rasgos	Error
1	Todos	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20	7.726
2	(Martin, 2003)	1,3,4,5,6,7,10,14,15,17,18,19,20	6.074
3	(Plissiti and Nikou, 2012)	1,4,6,7,8,9,14,17,18	10.523
4	Algoritmo 1	1, 2, 3, 4, 5, 7, 10, 11, 12, 13, 14, 17, 18, 19, 20	6.815
5	Irreducibles	2,3,7,10,11,13,14,17,19,20	8.206

Resultados del clasificador LS

En la Tabla (7) se muestran los resultados de clasificación obtenidos del clasificador LS. Para las mediciones del error de clasificación se empleó *10-fold cross-validation* con 50 iteraciones.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 2, abril-junio, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 73-88

Tabla 7 Error cometido por el clasificador LS empleando distintos conjuntos de rasgos

No.	Medición	Conjunto de rasgos	Error
1	Todos	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20	6.42
2	(Plissiti and Nikou, 2012)	1,4,6,7,8,9,14,17,18	10.59
3	Algoritmo 1	2 3 4 6 7 9 10 11 13 14 16 17 19 20	7.53
4	Irreducibles	2,3,7,10,11,13,14,17,19,20	8.32

Resultados del clasificador WKNN

En la Tabla (8) se muestran los resultados de clasificación obtenidos del clasificador WKNN. El algoritmo WKNN es una variante del algoritmo KNN donde la influencia de los k-vecinos más próximos es ponderada de acuerdo a su distancia respecto a un punto de prueba. Para las mediciones del error de clasificación se empleó *10-fold cross-validation* con 50 iteraciones.

Tabla 8 Error cometido por el clasificador WKNN empleando distintos conjuntos de rasgos

No.	Medición	Conjunto de rasgos	Error
1	Todos	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20	7.22
2	(Plissiti and Nikou, 2012)	1,4,6,7,8,9,14,17,18	10.47
3	Algoritmo 1	2, 3, 4, 6, 7, 9, 10, 11, 13, 14, 16, 17, 19, 20	6.98
4	Irreducibles	2,3,7,10,11,13,14,17,19,20	8.14

Conclusiones

El empleo de un método que tenga en cuenta las características más significativas durante el proceso de selección de rasgos, por lo general obtiene buenos resultados de clasificación y suele reducir el error de clasificación en 1.8436 por ciento respecto a otras metodologías de selección de rasgos reportadas en la literatura. No obstante, queda por demostrar si este conjunto de rasgos resultantes es el más eficiente o más aún cuáles son los casos en los que no mejoran en nada los resultados como en el caso del clasificador LS. Se hace necesario profundizar en el mecanismo de adaptación del conjunto de rasgos que conforman el núcleo del modelo a representar por un clasificador cualquiera.

Una serie de clasificadores de prueba fueron empleados para medir la eficiencia de una serie de conjuntos de rasgos discriminantes. Entre estos se encuentran HCM, FCM, LS y WKNN. El empleo de un mayor grupo de clasificadores de pruebas se hace necesario. El conjunto de rasgos discriminantes que conforman el núcleo del modelo a representar por un clasificador cualquiera demostró ser un predictor adecuado para el potencial discriminante del mismo.

Referencias

A HAJNAYEB, A GHASEMLOONIA, SE KHADEM, and MH MORADI. Application and comparison of an ann-based feature selection method and the genetic algorithm in gearbox fault diagnosis. Expert Systems with Applications, 38(8):10205-10209, 2011.

ABDOLLAH KAVOUSI-FARD. A new fuzzy-based feature selection and hybrid tla-ann modelling for short-term load forecasting. Journal of Experimental & Theoretical Artificial Intelligence, 25(4):543-557, 2013.

ANDREA E DAWSON. Can we change the way we screen?: The thinprep imaging system. Cancer Cytopathology, 102(6):340-344, 2004.

ANTANAS VERIKAS AND MARIJA BACAUSKIENE. Feature selection with neural networks. Pattern Recognition Letters, 23(11):1323-1335, 2002.

ASL? GEN?CTAV, SELIM AKSOY, and SEVGEN ONDER. Unsupervised segmentation and classification of cervical cell images. Pattern Recognition, 45(12):4151-4168, 2012.

B. KROSE. Introduction to neural networks. University of Amsterdam., 8 edition, 1996.

CHIN-WEN CHANG, MING-YU LIN, HORNG-JYH HARN, YEN-CHERN HARN, CHIEN-HUNG CHEN, KUN-HIS TSAI, AND CHIHUNG HWANG. Automatic segmentation of abnormal cell nuclei from microscopic image analysis for cervical cancer screening. In Nano/Molecular Medicine and Engineering (NANOMED), 2009 IEEE International Conference on, pages 77-80. IEEE, 2009.

CHRISTOPHER M. BISHOP. Neural Networks for Pattern Recognition. Oxford University Press, Inc., 1995.

COSMIN LAZAR, JONATAN TAMINAU, STIJN MEGANCK, DAVID STEENHOFF, ALAIN COLETTA, COLIN MOLTER, VIRGINIE DE SCHAETZEN, ROBIN DUQUE, HUGUES BERSINI, and ANN NOWE. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 9(4):1106-1119, 2012.

DIANE SOLOMON, DIANE DAVEY, ROBERT KURMAN, ANN MORIARTY, DENNIS O'CONNOR, MARIANNE PREY, STEPHEN RAAB, MARK SHERMAN, DAVID WILBUR, THOMAS WRIGHT Jr, et al. The 2001 bethesda system: terminology for reporting results of cervical cytology. Jama, 287(16):2114-2119, 2002.

E. MARTIN. Pap Smear classification. Technical University of Denmark., 2003.

EUNSANG BAK, KAYVAN NAJARIAN, and JOHN P BROCKWAY. Efficient segmentation framework of cell images in noise environments. In Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE, volume 1, pages 1802-1805. IEEE, 2004.

F.P. KRENDELEIEV A.N. DMITRIEV, YU.I. ZHURAVLEV. About the mathematical principles of patterns and phenomena classification. Journal Diskretnyi Analiz., 7 edition, 1966.

J RUÍZ-SHULCLOPER, E ALBA, and M LAZO. Introducción a la teoría de testores, 1994.

J. BYRIEL. Neuro-fuzzy classification of cells in cervial smears. Technical University of Denmark., 1999.

J.W. CURBELO. Clasificación de imágenes de microscopía celular en la prueba de Papanicolau por medios computacionales. Universidad Martha Abreu de las Villas., 2012.

JAN JANTZEN, JONAS NORUP, GEORGIOS DOUNIAS, AND BETH BJERREGAARD. Pap-smear benchmark data for pattern classification. Nature inspired Smart Information Systems (NiSIS 2005), pages 1-9, 2009.

JONAS NORUP. Classification of pap-smear data by transductive neuro-fuzzy methods. Technical University of Denmark., 2005.

JORMA PAAVONEN. Human papillomavirus infection and the development of cervical cancer and related genital neoplasias. International Journal of Infectious Diseases, 11:S3-S9, 2007.

KR NIAZI, CM ARORA, and SL SURANA. Power system security evaluation using ann: feature selection using divergence. Electric Power Systems Research, 69(2):161-167, 2004.

LAURIE J MANGO. Reducing false negatives in clinical practice: the role of neural network technology. American journal of obstetrics and gynecology, 175(4):1114-1119, 1996.

MANUEL LAZO-CORTES, JOSE RUIZ-SHULCLOPER, and EDUARDO ALBA-CABRERA. An overview of the evolution of the concept of testor. Pattern recognition, 34(4):753-762, 2001.

MARINA E PLISSITI and CHRISTOPHOROS NIKOU. Cervical cell classification based exclusively on nucleus features. pages 483-490, 2012.

MEISELS and MORIN. Cytopathology of the uterus. ASCP Press, 2nd edition, 1997.

P SOBREVILL, E LERMA, and E MONTSENY. An approach to a fuzzy-based automatic pap screening system-fapss-paddressed to cytology cells detection. pages 138-142, 2003.

RAFAEL BELLO, YUDEL GOMEZ, ANN NOWE, and MARIA M GARCIA. Two-step particle swarm optimization to solve the feature selection problem. In Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on, pages 691-696. IEEE, 2007.

RONALD S POULSEN and ILARIO PEDRON. Region of interest finding in reduced resolution colour imagery. Application to cancer cell detection in cell overlaps and clusters, volume 1. 1995.

S.V. YABLONSKII I.A. CHEGUIS. Logical Methods for controlling electrical systems. Trudy Matematicheskava Institutaimeni V. A. Steklova., 1958.

SOLANGEL RODRÍGUEZ-VÁZQUEZ and ANDY VIDAL MARTÍNEZ-BORGES. Clasificación de células cervicales con máquinas de soporte vectorial empleando rasgos del nícleo. Revista Cubana de Ciencias Informáticas, 9(2), 2015.

T. MITCHEL. Machine Learning. McGraw Hill, 1997.

THOMAS F KARDOS. The focalpoint system. Cancer Cytopathology, 102(6):334-339, 2004.

YANNIS MARINAKIS, GEORGIOS DOUNIAS, and JAN JANTZEN. 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(1):69-78, 2009.

YOUYI SONG, LING ZHANG, SIPING CHEN, DONG NI, BAOPU LI, YONGJING ZHOU, BAIYING LEI, and TIANFU WANG. A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei. In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, pages 2903-2906. IEEE, 2014.

YOVANIS SANTIESTEBAN ALGANZA and AURORA PONS PORRATA. LEX: Un nuevo algoritmo para el cálculo de los testores típicos. REVISTA CIENCIAS MATEMATICAS, Santiago de Cuba, CUBA, 21(1):2-3, 2003.

YUNG-FU CHEN, PO-CHI HUANG, KER-CHENG LIN, HSUAN-HUNG LIN, LI-EN WANG, CHUNG-CHUAN CHENG, TSUNG-PO CHEN, YUNG-KUAN CHAN, and JOHN Y CHIANG. Semi-automatic segmentation and classification of pap smear cells. Biomedical and Health Informatics, IEEE Journal of, 18(1):94-108, 2014.

YVAN SAEYS, INAKI INZA, and PEDRO LARRAÑAGA. A review of feature selection techniques in bioinformatics. bioinformatics, 23(19):2507-2517, 2007.