VISIÓN ELECTRÓNICA

CLASIFICACIÓN DE *plasmodium falciparum* POR ESTADIO EN CULTIVOS SINCRÓNICOS DE ERITROCITOS

Plasmodium falciparum STAGE CLASSIFICATION IN ERYTHROCYTE SYNCHRONOUS CULTIVARS

Luz H. Camargo[1]
Édgar Malagón[2]
María L. Caldas[3]

Fecha de envío: Abril 2010
Fecha de recepción: Junio 2010
Fecha de aceptación: Agosto 2010

Resumen:

La malaria es considerada como un problema de salud mundial, que afecta principalmente a países en vía de desarrollo. El objetivo de este trabajo consiste en evaluar técnicas de minería de datos para clasificar automáticamente eritrocitos infectados con Plasmodium falciparum en cultivos sincrónicos de eritrocitos. Se evaluaron imágenes de eritrocitos con parásitos de 14, 28, 40 y 48 horas y como controles se consideraron eritrocitos no infectados, los modelos de clasificación empleados corresponden a Naive Bayes (NB), árbol de decisión con el algoritmo J48 y red neuronal Multilayer Perceptron (RN). Encontrando el mejor desempeño con árbol de decisión.

1 Licenciada en Biología. Especialista en Bioingeniería. Candidata M.Sc. en Ingeniería Biomédica. Docente Universidad Distrital Francisco José de Caldas. Facultad de Ingeniería. Correo lhcamargoc@udistrital.edu.co

2 Candidato M.Sc. en Matemática Aplicada. Universidad Nacional de Colombia. Correo ehmalagona@unal.edu.co

3 Bióloga, M.Sc. en Bioquímica. Instituto Nacional de Salud, Morfología Celular. Correo mcaldas@ins.gov.co

Palabras clave:
Malaria, *Plasmodium falciparum*, minería de datos, árbol de decisión, redes neuronales.

Abstract:
Malaria is considered a global health problem that affects mainly developing countries. The aim of this study is to evaluate data mining techniques to automatically classify erythrocytes infected with Plasmodium Falciparum in erythrocytes synchronous cultivars. RBC images are evaluated with 14-hour, 28-hour, 40-hour and 48-hour parasites, and uninfected erythrocytes were considered as control references. The classification models employed correspond to Naive Bayes (NB), J48-algorithm decision tree and Multilayer Perceptron neural network (RN), finding the best performance with decision tree.

Key Words:
Malaria, *Plasmodium falciparum,* data mining, Decision Tree, neural networks.

1. Introducción

La malaria es una enfermedad parasitaria que afecta a gran parte de la población, principalmente en países en vía de desarrollo, y cada día cobra más vidas humanas. Se presentan anualmente entre 350 y 500 millones de casos clínicos y mueren más de un millón de personas, de las cuales el 90% son niños menores de cinco años; adicionalmente, 3.200 millones de personas viven en zonas de riesgo de transmisión del paludismo [1]. En las Américas, el 38,4% de las personas viven en áreas con condiciones propicias para la transmisión y el 86% de malaria se presenta en países amazónicos [2].

La detección del parásito *Plasmodium* en extendidos sanguíneos es posible gracias a las características que este adquiere al ser sometido a tinción de Giemsa [3]. Las características morfológicas y morfométricas permiten diferenciar el parásito por estadios de vida. Los parásitos de malaria los transmiten mosquitos *Anopheles*, así: un esporozoito invade y se desarrolla en el hígado; después, los parásitos merozoítos invaden y se multiplican en los glóbulos rojos, donde se pueden observar diferentes estadios: en uno de ellos, este parásito presenta forma de anillo (estado mononucleado) y los trofozoítos maduros presentan pigmentos granulados (estado mononucleado), y en los esquizontes el núcleo se divide (estado multinucleado y núcleos periféricos) [4]. Este trabajo busca, mediante minería de datos, diferenciar eritrocitos no infectados e infectados y, en estos últimos, el estadio de desarrollo del parásito.

2. Metodología

Se cultivó la cepa colombiana FCB-2 según el método de Trager y Jensen [5], y el cultivo asincrónico se trató con sorbitol [6], para eliminar parásitos maduros y dejar una población heterogénea de anillos de 14 horas; enseguida se continuó con el cultivo hasta las 28 horas, en donde los parásitos se encuentran en el estadio de trofozoítos maduros; después continúa el cultivo y se centrifuga en un gradiente de Percoll, recuperando la fracción que contiene esquizontes de 40 y 48 horas.

2.1. Adquisición y procesamiento digital de Imágenes

Se digitalizaron campos en el microscopio Zeiss Axiophot con magnificación de 200X. Estos campos se capturaron con una cámara de video CCDIris/RGB acoplada a un computador que contiene el programa de análisis de imágenes KS300 (Kontron Elektronic System), con el cual se realizó el procesamiento y análisis de las imágenes adquiridas. Las imágenes fueron digitalizadas y almacenadas a 640 x 480 pixeles; luego se realizó el procesamiento de imágenes, el cual incluyó contraste de las imágenes, selección de objetos de interés (eritrocitos y parásitos), eliminación de ruido de fondo, calibración y medición automática de cada imagen.

2.2. Evaluación de los datos

En este experimento se digitalizaron imágenes de eritrocitos infectados con *P. falciparum* en cuatro estadios: de 14, 28, 40 y 48 horas, así: 7 imágenes de eritrocitos con parásitos de 14 horas, 12 imágenes de eritrocitos con parásitos de 28 horas, 7 imágenes de eritrocitos con parásitos de 40 horas y

Tabla 1. Medidas geométricas, densitométricas y de estado.

Medidas geométricas	
AREA	Área de la región
AREAF	Área de la región completa
FERETMAX	Feret mínimo de la región (largo)
FERETMIN	Feret máximo de la región (largo)
FERETRATIO	Radio de los Feret máximo y mínimo
PERIM	Perímetro de la región
PERIMC	Perímetro de la región convexa
PERIMF	Perímetro de la región periférica
Medidas densitométricas	
MEAND	Densidad media de la región
STDD	Densidad estándar de los valores densitométricos de la región
MIND	Valores de grises mínimos de la región
MAXD	Valores de grises máximos de la región
SUMD	Suma de los valores densitométricos de la región
Variables de estado	
TIPO	**Eritrocito o parásito**
EDAD	14, 28, 40 ó 48 horas y no aplica
ESTADO	Infectado, no infectado, no aplica

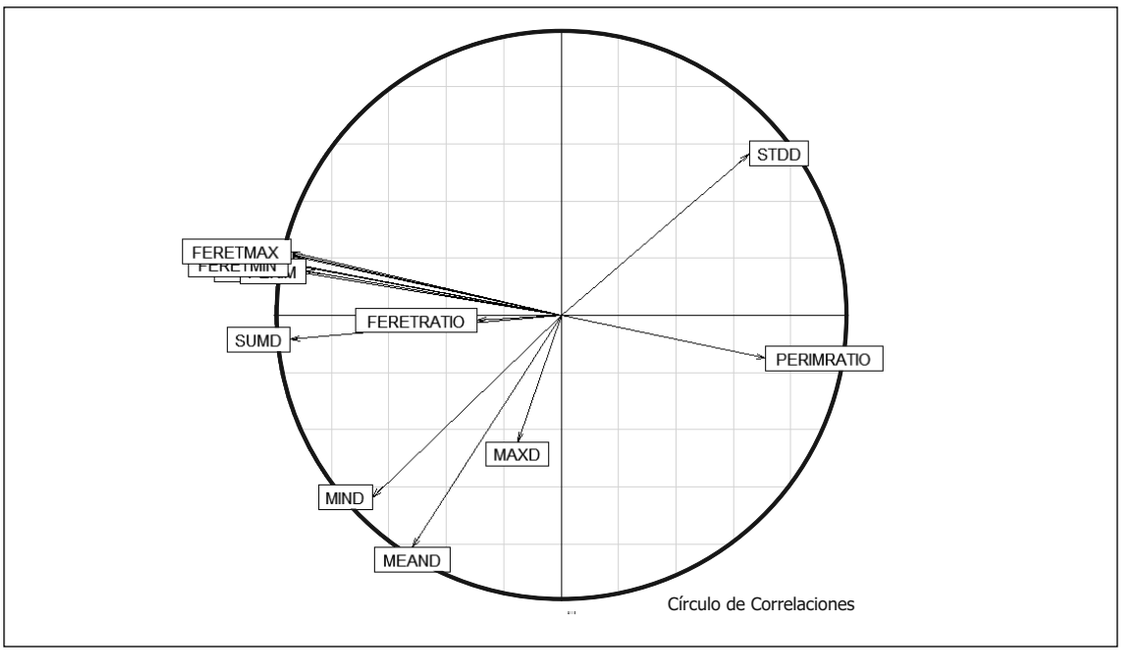
12 imágenes de eritrocitos con parásitos de 48 horas. Además, se tomaron 84 imágenes de eritrocitos no infectados, denominados eritrocitos de control. En todos los casos se obtuvieron 14 variables después de hacer un análisis gráfico de las imágenes. Las variables consideradas se observan en la Tabla 1.

Considerando la información de los eritrocitos tanto infectados como sanos y de los parásitos, se tiene un total de 208 datos.Este trabajo busca determinar descriptores relevantes en la caracterización del parásito de acuerdo con su edad, teniendo en cuenta herramientas de minería de datos.

3. Resultados

Al realizar una estandarización de los datos y los diagramas de caja para las 14 variables, se detectaron, considerando dos (2) desviaciones estándar: 18 *outliers* en las variables AREA, AREAF, PERIMC y FERRETMIN, que corresponden a todos los parásitos de 14 y 28 horas; 21 *outliers* para PERIM, que corresponden a parásitos de 14 y 28 horas y eritrocitos de control; 22 *outliers* de PERIMF, que corresponden a parásitos de 14 y 28 horas y eritrocitos de control; 12 *outliers* de eritrocitos de control y parásito de 14 horas para la variable PERIMRATIO; 18 *outliers* para FERETMAX, que corresponden a parásitos de 14 y 28 horas y un

Figura 1. Círculo de correlaciones para las 14 variables.

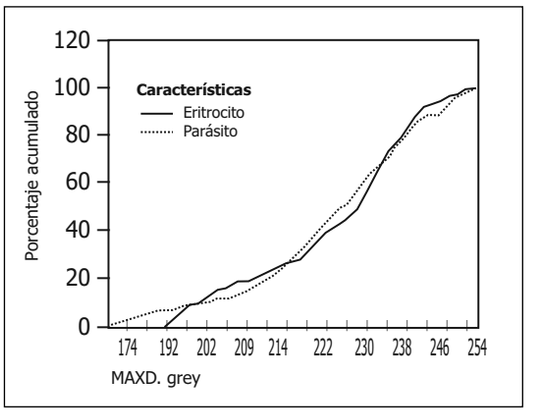


eritrocito de control; para MIND y STDD se encontraron 7 y 1 *outliers*, que corresponden a parásitos de 14 horas; en SUMD se encontraron 12 *ouliers* de parásitos de 14 y 28 horas; en MAXD se encontraron 5 *outliers* de 14, 28 y 40 horas, y para MEAND se encontraron 5 *outliers* de parásitos de 14 horas y 48 horas. Al analizar el círculo de correlaciones, se observó que las variables FERRETRATIO y MAXD son las que menos aportan en la variabilidad de todos los datos, como se observa en la Figura 1.

También se analizaron los porcentajes acumulados en las variables, donde se notó que las variables FERRETRATIO y MAXD tienen una distribución acumulada muy parecida, como se aprecia en las Figuras 2 y 3.

En la matriz de dispersión se puede apreciar la poca separabilidad que hacen entre eritrocitos y parásitos las variables FERRETRATIO y MAXD con respecto a todas las demás. De este modo no se tendrán en cuenta los atributos FERRETRATIO y MAXD. Se observa en la Figura 4 que las variables AREA y AREAF están altamente correlacionadas y permiten una buena separación entre eritrocitos y parási-

Figura 2. Frecuencias acumuladas para MAXD.



Considerando la información de los eritrocitos tanto infectados como sanos y de los parásitos, se tiene un total de 208 datos. Este trabajo busca determinar descriptores relevantes en la caracterización del parásito de acuerdo con su edad, teniendo en cuenta herramientas de minería de datos.

Figura 3. Frecuencias acumuladas para FERRETRATIO.

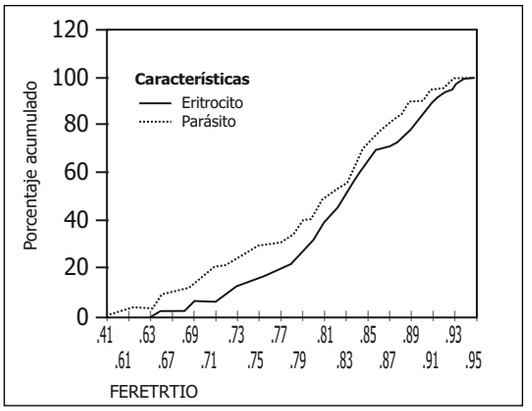


Figura 4. Dispersión entre AREA y AREAF.

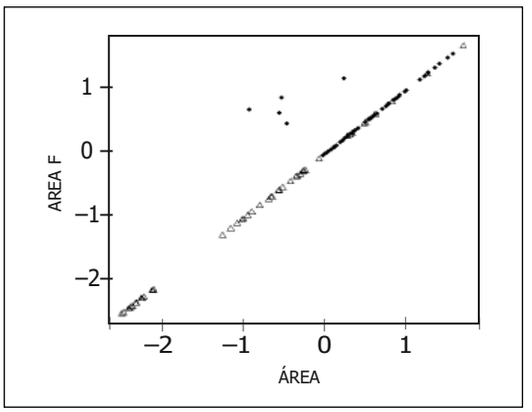
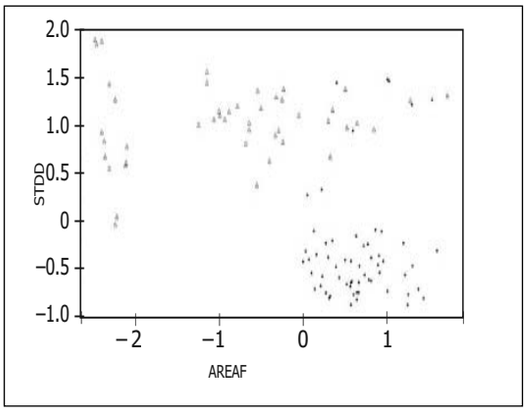


Figura 5. Dispersión entre AREAF y STDD.



tos, por lo que se dejará solo una de las dos para hacer la primera clasificación. Otra observación: la Figura 5 corresponde a AREAF y STDD, que permiten separar bien las clases al cruzarlas con las demás variables.

3.1. Reglas de asociación

Se discretizaron los atributos numéricos en máximo 10 bins de igual amplitud, soporte mínimo de 0.1 y máximo de 1.0 y confianza mínima de 0.9; luego se aplicó el algoritmo de reglas de asociación a priori a este conjunto de datos, seleccionando de las 50 reglas obtenidas aquellas que fueran interesantes para el estudio. Al variar el soporte a 0,5 las reglas arrojadas por el programa WEKA© no resultaron interesantes para el estudio, por ello se continuó trabajando con soporte de 1.0.

Un primer acercamiento a las reglas de asociación con todas las variables arrojó reglas poco interesantes para el estudio de características propias de los parásitos y los eritrocitos. Al realizar una selección de atributos con una herramienta propia de WEKA, se toman solo las variables areaf y stdd como las que mejor expresan al conjunto en términos de separación entre eritrocitos y parásitos. Haciendo las reglas de asociación con el algoritmo A priori de WEKA y discretizando distintos tipos, se obtienen reglas como las siguientes: en la Tabla 1 se aprecian los resultados A priori de 10 bins de igual amplitud, encontrando 10 reglas interesantes, 8 orientadas a eritrocitos y 2 a parásitos.

Con respecto a reglas A priori con 10 bins de igual frecuencia, se aprecian 8 reglas interesantes, dos orientadas a eritrocitos y 6 a parásitos.

Tabla 2. Reglas a priori con 10 bins de igual amplitud: confianza 100%.

Apriori con 10 bins de igual amplitud	
Regla	Soporte
STDD∈(11.158,14.496]→ TIPO=Eritrocito	21 %
MIND∈(140.8,158.4]→ TIPO=Eritrocito	16 %
MIND∈(158.4,inf)→ TIPO=Eritrocito	14 %
STDD∈(-inf,11.158]→ TIPO=Eritrocito	13 %
STDD∈(14.496,17.834]→ TIPO=Eritrocito	13 %
MIND∈(123.2,140.8]→ TIPO=Eritrocito	12 %
MEAND∈(131.2,142.7]→ TIPO=Parasito	12 %
PERIM∈(15.9,20.6]→ TIPO=Parasito	12 %
PERIMC∈(21.9,24.1]&SUMD∈(435885.4,495553.6]→ TIPO=Eritrocito	12 %
FERETMIN∈(6.6,7.2]&STDD∈(11.1,14.4]→ TIPO=Eritrocito	11 %

Tabla 3. Reglas a priori con 10 bins de igual frecuencia: confianza 100%.

Apriori con 10 bins de igual frecuencia	
Regla	Soporte
FERETMAX∈(8.85,inf)→ TIPO=Eritrocito	11 %
AREA∈(-inf,15.375]→ TIPO=Parasito	10 %
AREAF∈(-inf,15.375]→ TIPO=Parasito	10 %
AREAF∈(15.375,24.28]→ TIPO=Parasito	10 %
AREAF∈(41.485,45.47]→ TIPO=Eritrocito	10 %
PERIM∈(-inf,15.075]→ TIPO=Parasito	10 %
PERIM∈(15.075,20.45]→ TIPO=Parasito	10 %
PERIMC∈(-inf,14.515]→ TIPO=Parasito	10 %

Para reglas A priori con 5 bins de igual amplitud, se aprecian 11 reglas interesantes, 9 orientadas a eritrocitos y 1 a parásitos.

Con respecto a las reglas A priori con 5 bins de igual frecuencia, se aprecian 24 reglas interesantes, 10 orientadas a eritrocitos y 14 a parásitos, Tablas 2 a 6.

3.2. Clasificación y predicción

La clasificación se realizó en dos niveles. En el primer nivel se consideraron las clases eritrocito sano, parásito, y en el segundo nivel se consideró el estadio del parásito. Se evaluaron tres (3) modelos de clasificación: Naive Bayes (NB), árbol de decisión con algoritmo J48 y red neuronal Multilayer Perceptron (RN). No se eliminaron *outliers* y se hizo validación cruzada con 10 grupos para todos los modelos de clasificación evaluados.

En la Tabla 6 se consideran "positivos" (TP) los parásitos clasificados como parásitos; "negativos v" (TN) los eritrocitos clasificados como eritrocitos; "falsos positivos" (FP) los parásitos clasificados como eritrocitos, y

Tabla 4. Reglas a priori con 5 bins de igual amplitud: confianza entre 91% -100%.

Apriori con 5 bins de igual amplitud		
Regla	Soporte	Confianza
STDD∈(-inf,14.4]→ TIPO=Eritrocito	35 %	100 %
MIND∈(140.8,inf)→ TIPO=Eritrocito	30 %	100 %
MIND∈(140.8,inf)&STDD∈(-inf,14.4]→ TIPO=Eritrocito	27 %	100 %
PERIMC∈(19.7,24.1]&SUMD∈(376217.2,495553.6]→ TIPO=Eritrocito	27 %	93 %
STDD∈(27.8,34.5]→ TIPO=Parasito	30 %	93 %
AREA∈(31.9,41.6]&SUMD∈(376217.2,495553.6]→ TIPO=Eritrocito	28 %	92 %
AREAF∈(31.9,41.6]&SUMD∈(376217.2-495553.6]→ TIPO=Eritrocito	28 %	92 %
AREAF∈(31.932,41.676]&SUMD∈(376217.2,495553.6]→ AREA∈(31.9,41.6]& TIPO=Eritrocito	28 %	92 %
SUMD∈(376217.2,495553.6]→ TIPO=Eritrocito	38 %	92 %
AREA∈(31.9,41.6]&SUMD∈(376217.2,495553.6]→ AREAF∈(31.9,41.6]& TIPO=Eritrocito	28 %	91 %

Tabla 5. Reglas a priori con 5 bins de igual frecuencia: confianza 100%.

Apriori con 5 bins de igual frecuencia	
Regla	Soporte
AREAF∈(-inf,24.2]→ TIPO=Parasito	20%
PERIM∈(-inf,20.4]→ TIPO=Parasito	20%
PERIMF∈(-inf,20.4]→ TIPO=Parasito	20%
FERETMIN∈(-inf,5.3]→ TIPO=Parasito	20%
SUMD∈(-inf,226980.5]→ TIPO=Parasito	20%
STDD∈(-inf,12.1]→ TIPO=Eritrocito	20%
STDD∈(12.1,15.4]→ TIPO=Eritrocito	20%
PERIM∈(-inf,20.4]&PERIMF∈(-inf,20.45]→ TIPO=Parasito	20%
PERIMF∈(-inf,20.4]→ PERIM∈(-inf,20.45]&TIPO=Parasito	20%
PERIM∈(-inf,20.4]→ PERIMF∈(-inf,20.45]&TIPO=Parasito	20%
MIND∈(149,inf)→ TIPO=Eritrocito	20%
PERIM∈(-inf,20.4]&PERIMC∈(-inf,19.0]→ TIPO=Parasito	20%
PERIMC∈(-inf,19.0]&PERIMF∈(-inf,20.4]→ TIPO=Parasito	20%
PERIM∈(-inf,20.4]&PERIMC∈(-inf,19.0]&PERIMF∈(-inf,20.4]→ TIPO=Parasito	20%
PERIMC∈(-inf,19.0]&PERIMF∈(-inf,20.4]→ PERIM∈(-inf,20.4]&TIPO=Parasito	20%
PERIM∈(-inf,20.4]&PERIMC∈(-inf,19.0]→ PERIMF∈(-inf,20.4]&TIPO=Parasito	20%
MIND∈(127,149]→ TIPO=Eritrocito	20%
AREA∈(-inf,23.9]&AREAF∈(-inf,24.2]→ TIPO=Parasito	19%
AREAF∈(-inf,24.2]&SUMD∈(-inf,226980.5]→ TIPO=Parasito	19%
AREAF∈(-inf,24.2]&FERETMIN∈(-inf,5.3]→ TIPO=Parasito	19%
PERIM∈(-inf,20.4]&FERETMAX∈(-inf,6.5]→ TIPO=Parasito	19%
PERIM∈(-inf,20.4]&SUMD∈(-inf,226980.5]→ TIPO=Parasito	19%
PERIMC∈(-inf,19.0]&FERETMAX∈(-inf,6.5]→ TIPO=Parasito	19%
PERIMC∈(-inf,19.0]&SUMD∈(-inf,226980.5]→ TIPO=Parasito	19%

Tabla 6. Matriz de confusión

	Parásito	Eritrocito
Parásito	TP	FN
Eritrocito	FP	TN

"falsos negativos" (FN) los eritrocitos clasificados como parásitos.

En la Tabla 7 se observan los resultados de las instancias correcta e incorrectamente clasificadas.

En la Tabla 8 se presentan ratas TPR, TNR, FPR y FNR, y en la Tabla 9 la precisión (P), el accuracy (A), el recall (R), la rata de error (ER) y la rata de detección (DR).

Tabla 7. Instancias clasificadas con Naive Bayes (NB), árbol de decisión con algoritmo J48 y red neuronal Multilayer Perceptron (RN)

Preprocesamiento	Método	Correctamente Clasificado	Incorrectamente Clasificado
–	NB	93.5%	6.25%
Componentes Principales	NB	93.75%	6.2%
z-Normalización	NB	94.23%	5.76%
Discretización (Bin Auto)	J48	91.82%	8.18%
Discretización (5 Bin=Ampl)	J48	94.23%	5.76%
Discretización (5 Bin=Frec)	J48	92.78%	7.21%
–	RN	97.59%	2.40%
z-Normalización	RN	97.59%	2.40%
Discretización (Bin Auto)	RN	97.59%	2.40%

Tabla 8. Positivos (P)/TPR, falsos negativos (FN), falsos positivos (FP), negativos (TN).

Preprocesamiento	Método	TP/TPR	FN/FNR	FP/FPR	TN/TNR
–	NB	121	1	12	74
		99.18%	0.82%	13.95%	86.0.5%
C.P.	NB	113	9	4	82
		92.62%	7.38%	4.65%	95.35%
z-N.	NB	122	0	12	74
		100%	0%	13.95%	86.05%
D. (Bin Auto)	J48	119	3	14	72
		97.54%	2.46%	16.28%	83.72%
D. (5 Bin=Ampl)	J48	114	8	4	82
		93.44%	6.56%	4.65%	95.35%
D. (5 Bin=Frec)	J48	116	6	9	77
		95.08%	4.92%	10.47%	89.53%
–	RN	122	0	5	81
		100%	0%	5.81%	94.19%
z-N.	RN	122	0	5	81
		100%	0%	5.81%	94.19%
D. (Bin Auto)	RN	122	0	5	81
		100%	0%	5.81%	94.19%

4. Conclusiones

Las variables que aportan menos en la clasificación de los individuos entre parásitos y eritrocitos son FERRETRATIO y MAXD. Las variables AREA y AREAF están altamente corre-

Tabla 9. Precisión (P), *recall* (R), *accuracy* (A), rata de error (ER) y rata de detección (DR).

Prep.	Mét.	P	R	A	ER	DR
–	NB	91.0%	99.2%	93.8%	6.3%	98.7%
C.P.	NB	96.5%	92.6%	93.7%	6.2%	90.1%
z-N.	NB	91.0%	100%	94.2%	5.7%	100%
D. (Auto)	J48	89%	97.5%	91.8%	8.1%	96.0%
D. (5=Ampl)	J48	96.6%	93.4%	94.2%	5.7%	91.1%
D. (5=Frec)	J48	92.8%	95.0%	92.7%	7.2%	92.7%
–	RN	96.0%	100%	97.6%	2.4%	100%
z-N.	RN	96.0%	100%	97.6%	2.4%	100%
D. (Auto)	RN	96.0%	100%	97.6%	2.4%	100%

lacionadas, por lo que se puede utilizar una sola de ellas. Los atributos AREA y STDD son los que más aportan en la identificación de eritrocitos y parásitos. Se puede apreciar en las reglas de asociación que si una imagen tiene un valor de la variable STDD en el intervalo de 1 a 355, este puede ser catalogado como eritrocito con una confianza del 100%, con lo cual el problema de la clasificación entre parásitos y eritrocitos está parcialmente resuelto.

Los métodos de Naive Bayes con componentes principales como preprocesamiento, J48 con discretización de 5 bins de igual amplitud y una red neuronal, son los que mejor desempeño tienen. Los métodos de Naive Bayes con componentes principales como preprocesamiento, J48 con discretización de 5 bins de igual amplitud y una red neuronal, son los que mejor desempeño tienen.

Con el fin de mejorar el desempeño de los clasificadores se utilizó una matriz de costo que mejorará la aparición de falsos negativos y penalizara los falsos positivos. Al observar los modelos de clasificación, el que arrojó menos falsos positivos fue el Multilayer Perceptron.

Al observar los modelos de clasificación, el que arrojó menos falsos positivos fue el Multilayer Perceptron. Sin embargo, aunque mejora el desempeño de clasificación de falsos positivos, al utilizar matrices de costo adecuadas se ve afectada la tasa de error

Referencias

[1]. WMR, Unicef. *World malaria report*. WMR y Unicef, 2005.

[2]. G. F. Pineda, V. Valero y cols. "Informe quincenal epidemiológico nacional en evaluación del programa de control de la malaria en la amazonia colombiana". *Salud Pública*, 6 (sup. 1) (2004): 40-49.

[3]. T. Hänscheid. "Diagnosis of malaria: a review of alternatives to conventional microscopy". *Clin. Lab. Haem*, 21 (1999): 235-245

[4] *E. Deharo,* Ph. *Gautret,* V. *Muñoz y M. Sauvain. Técnicas de laboratorio para la selección de sustancias antimaláricas*. Bolivia: CYTED-IRD, 2000.

[5] W. Trager y J. B. Jensen. "Human malaria parasitemia continuous culture". *Science*, 173 (1976): 673-675.

[6] M. Rojas y M. Wasserman. "Supersincronización del crecimiento *in vitro* del *Plasmodium falciparum*". *Biomédica*, 7 (1978): 75-80.