UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

"CLASIFICACIÓN DE LEUCOCITOS MEDIANTE REDES BAYESIANAS"

T E S I S

PARA OBTENER EL TÍTULO DE

INGENIERO EN COMPUTACIÓN

PRESENTA:

Lucio Jiménez Díaz

DIRECTORES DE TESIS:

M.C. Verónica Rodríguez López,

M.C. Raúl Cruz Barbosa.

HUAJUAPAN DE LEÓN, OAX. NOVIEMBRE DE 2007

A mis padres, a mi familia toda y a mis amigos.

Contenido

Capítulo 1

Introducción

Un campo de estudio muy interesante en el área de Inteligencia Artificial y Reconocimiento de Patrones es el Aprendizaje Máquina –Machine Learning, en inglés–. Generalmente, la parte interesante y atractiva de dicha disciplina se convierte importante cuando se atacan o resuelven problemas reales de nuestro entorno o sociedad. Nuestra aplicación objeto de estudio se refiere al análisis sanguíneo, en particular al análisis de leucocitos –véase Glosario para definiciones correspondientes al área de hematología–. La clasificación manual de leucocitos es una tarea engorrosa que sigue existiendo en los laboratorios de análisis clínicos. Aún cuando hay en el mercado aparatos electrónicos capaces de llevar a cabo el recuento y clasificación de este tipo de células, se limitan a clasificar células normales y, a lo más, indican la presencia de células anormales, inmaduras o desconocidas. Se han desarrollado, en años recientes, proyectos de investigación orientados a la aplicación de técnicas de tratamiento digital de imágenes y reconocimiento de patrones para llevar a cabo la tarea antes mencionada. En el presente trabajo de tesis se desarrolla un clasificador de leucocitos que encuentra su fundamento principal en la teoría de redes bayesianas y sistemas expertos, así como en el procesamiento digital de imágenes.

En este primer capítulo, de forma breve, es expuesto el problema de clasificación manual de leucocitos así como las desventajas que presentan los actuales métodos automáticos que realizan esta tarea. Con esto se pretende justificar la aplicación de las redes bayesianas como una alternativa en la resolución del problema. En el capítulo segundo se expone la teoría de tratamiento de imágenes digitales necesaria para el desarrollo de este trabajo. El capítulo tercero expone la teoría básica de redes bayesianas, la cual es el fundamento principal del diseño del clasificador. El capítulo cuarto ha

sido dedicado al desarrollo de este trabajo, presentandose en él, la metodología específica utilizada en el diseño e implementación del clasificador de leucocitos. El capítulo quinto presenta un resumen de las pruebas aplicadas al sistema, así como sus resultados y evaluación general de funcionamiento. En el capítulo último se exponen brevemente las conclusiones generales y perspectivas del trabajo realizado.

1.1. Planteamiento del problema

En los laboratorios de análisis clínicos de hoy día se llevan a cabo numerosos tipos de estudios, entre los cuales, uno de los más frecuentemente solicitados, en el área de hematología, es el de la *biometría hemática* –BH–, también conocido como *citometría hemática –CH–*. La BH se compone de dos tipos de datos: datos de la serie roja y datos de la serie blanca. Los datos de la serie roja se relacionan con valores y parámetros mensurables de los eritrocitos –glóbulos rojos–. Los datos de la serie blanca comprenden el número total de leucocitos –glóbulos blancos–, la cuenta diferencial –que es el conteo poblacional relativo, expresado en porcentajes, de las variedades de glóbulos blancos presentes en una muestra de sangre– y alteraciones de los mismos [RA01]. Es en los dos últimos puntos donde se lleva a cabo, con mayor frecuencia, la clasificación manual de leucocitos.

La clasificación manual de leucocitos comprende por lo menos dos etapas, a partir de que se cuenta con la muestra de sangre objeto de estudio. La primera es la preparación y tinción del frotis, la cual se lleva a cabo de 7 a 20 minutos [LRM$^+$97], por lo general, dependiendo del tipo del método de tinción y colorante utilizados. La segunda etapa comprende la observación de la muestra al microscopio, actividad que se realiza de 5 a 15 minutos en promedio, dependiendo de la experiencia práctica del laboratorista*. Como puede observarse, uno de los grandes inconvenientes que presenta este tipo de metodología es el elevado consumo de tiempo de un especialista para analizar una sola muestra de sangre.

Existen en el mercado aparatos electrónicos capaces de llevar a cabo la cuenta diferencial de los leucocitos. El principio de funcionamiento de tales aparatos se fundamenta, en la mayoría de los casos, en el Principio Coul-

*Dato confirmado por la Laboratorista Clase A: María Margarita Sánchez Chávez, encargada del área de hematología, en el laboratorio de análisis clínicos del Hospital del ISSSTE, Oaxaca.

ter[†][Cou98] y en la citometría de flujo[‡][GV01]. El gran número de células que son capaces de procesar es su principal ventaja, ya que con esto reducen su error estadístico. En ambos casos, tales aparatos se limitan a clasificar células normales y, a lo más, a indicar la aparición de células anormales o desconocidas en la muestra sanguínea. Presentan también el inconveniente de clasificar como linfocitos a la mayor parte de células plasmáticas, arrojando resultados erróneos en la cuenta diferencial cuando se presentan ciertas patologías que incrementan el número de tales células. Aunado a esto, se debe señalar que el precio de tales aparatos es elevado así como lo son los reactivos necesarios para su funcionamiento y mantenimiento. Por último, es preciso añadir que aún cuando se lleve a cabo la clasificación con este tipo de aparatos es necesario hacer la clasificación manual de los leucocitos antes de dar por sentados los resultados del estudio.

En años recientes se han desarrollado diferentes proyectos de investigación cuya finalidad es llevar a cabo la clasificación de los leucocitos de forma exacta y confiable, tales esfuerzos se centran en el análisis digital de imágenes y el reconocimiento de patrones [SZR04, SA02, PRG$^+$01]. El proceso de clasificación, en la generalidad de estos proyectos, comprende las siguientes etapas:

1. Obtención de la imagen.
2. Segmentación.
3. Extracción de características.
4. Clasificación.

Cada una de estas etapas puede verse como un proceso en sí mismo y por lo que respecta a la etapa de clasificación, en la totalidad de los proyectos que se han revisado hasta el momento, es implementada con clasificadores simples que utilizan técnicas como *template matching* –emparejamiento de plantillas– [SA02], funciones de decisión [PRG$^+$01] o a través del clasificador *naive Bayes* [SZR04]. La utilización de tales clasificadores reduce en gran

[†]De acuerdo con este principio, una pequeña apertura entre electrodos es la zona de lectura a través de la cual pasan partículas suspendidas. En la zona de lectura cada partícula desplaza su propio volumen de electrolito. El volumen desplazado es medido como un pulso de voltaje; siendo la intensidad de cada pulso proporcional al volumen de la partícula.

[‡]Técnica de análisis celular multiparamétrico cuyo fundamento se basa en hacer pasar una suspensión de partículas –generalmente células– alineadas y de una en una por delante de un haz de láser focalizado.

medida la flexibilidad en el manejo de la información y por consiguiente la precisión y confiabilidad de los resultados, ya que se centran en criterios cuantitativos y restan relevancia a los aspectos cualitativos –por ejemplo correlaciones– que existen entre los datos que se procesan.

1.2. Justificación

Desde este punto, puede observarse que las desventajas principales de los métodos actuales de clasificación de leucocitos son:

- El elevado consumo de tiempo del especialista que lleva a cabo el estudio.
- El alto costo del equipo electrónico que realiza la clasificación automática así como los reactivos necesarios para su funcionamiento y mantenimiento.
- El limitado tipo de células que son capaces de clasificar los aparatos electrónicos así como el error recurrente que presentan con algunas de ellas.
- El error inherente a las técnicas de procesamiento digital de imágenes y clasificación de patrones aplicadas en cada etapa del reconocimiento de las células.
- La falta de aplicación de técnicas, teorías y tecnología en el campo de las ciencias computacionales.

Es por tanto deseable que un sistema clasificador realice su trabajo en poco tiempo, que el equipo sea accesible –tanto física como económicamente–, que no presente limitaciones en cuanto a los tipos de células que es capaz de clasificar y que sea flexible en el manejo de los datos o información, de tal manera que sea posible corregir los errores recurrentes que presente. En cuanto a los errores inherentes al desarrollo de nuevas tecnologías, técnicas y teorías, es poco lo que se puede corregir a corto plazo, ya que estos errores se presentan en el campo del desarrollo científico y se van corrigiendo mientras éste avanza.

El desempeño y bajo costo de los actuales equipos de cómputo al igual que el hardware necesario para la captura de imágenes digitales –cualquiera que sea la fuente, por ejemplo: un microscopio– hacen a este tipo de tecnología una plataforma deseable para el desarrollo de un clasificador que cuente con las características antes señaladas.

Las actuales técnicas y teorías en el tratamiento digital de imágenes, clasificación de patrones y sistemas expertos dejan entrever un campo fértil para su aplicación a problemas específicos como el que se ha planteado.

En vista de las razones anteriores y concentrando nuestra atención en las etapas de extracción de características y clasificación, se ha determinado en este proyecto de tesis, la exploración y aplicación de la teoría de redes bayesinas como una opción que permite el manejo flexible de datos e información, al mismo tiempo que descansa sobre bases rigurosamente matemáticas y es factible su realización con la tecnología de cómputo actual y hardware necesario para la captura de imágenes digitales.

1.3. Objetivos

A continuación se presentan los objetivos a desarrollar en el presente proyecto de tesis.

- Objetivo general. Diseño e implementación de un clasificador de leucocitos fundamentado en la teoría de redes bayesianas; tal clasificador debe ser capaz de reconocer los 5 tipos básicos de leucocitos –neutrófilos, linfocitos, monocitos, eosinófilos y basófilos– en estado normal y maduro.
- Objetivos específicos.
  - Extracción de características morfológicas de las imágenes de las células bajo estudio, mediante técnicas de procesamiento digital de imágenes.
  - Diseño, implementación y aprendizaje de la red bayesiana que funcionará como clasificador de imágenes de células.

Capítulo 2

Procesamiento digital de imágenes

Uno de los primeros pasos hacia la clasificación de leucocitos es, en primer término, recolectar la información distintiva y representativa contenida en las imágenes de los mismos. Para este propósito se hace uso de técnicas de procesamiento digital de imágenes, cuya finalidad es la de preparar la imagen de modo tal que se eliminen de ella características indeseadas –ruido, por ejemplo– que interfieren en la extracción de los datos relevantes que contiene. Al proceso de preparación de la imagen o eliminación del ruido contenido en ella se le conoce como *pre-procesamiento* y a la extracción de los datos relevantes se le conoce como *extracción de características*.

En el presente capítulo se presenta brevemente el modelo conceptual de representación de las imágenes digitales en general, y de las imágenes en color como caso particular. Se continúa entonces con la descripción de la teoría y técnicas de pre-procesamiento de la imagen y extracción de características, utilizadas todas ellas, en el desarrollo de este trabajo.

2.1. Representación de imágenes digitales

Las imágenes digitales, en general, pueden ser entendidas como una función $f : \mathbb{Z}^2 \to \mathbb{Z}$. Cada uno de los pares ordenados del dominio de la función f son interpretados como las coordenadas de algún punto en un plano, al cual, es asignado un único valor que representa el color* observado en ese

*Se entiende por color, en este contexto, la sensación producida por ondas electromagnéticas, cuyas longitudes de onda se encuentran en el rango visible, al entrar en contacto con los órganos visuales humanos.

punto.

Por otro lado, las imágenes dentro de un sistema de cómputo pueden comprenderse como un arreglo bidimensional o matriz de píxeles. El valor de cada píxel corresponde al color del punto correspondiente en la escena observada. De esta forma, una imagen digitalizada y dentro de un sistema de cómputo, puede ser descrita como una matriz $N \times M$ m-bits, donde N y M representan las dimensiones de la matriz y m controla el número de valores de color. Si utilizamos m bits tenemos 2^m diferentes valores de color que van desde 0 a $2^m - 1$ [NA02].

2.1.1. Representación del color

Para poder representar de una forma consistente las imágenes es necesario auxiliarse de algún medio, por el cual se especifique de forma única cada color que la compone. Los *espacios de color* llevan a cabo esta tarea.

Un *espacio de color* es un método por el que podemos especificar, crear y visualizar color. Un color es especificado utilizando coordenadas o atributos. Estas coordenadas no nos dicen cuál es el color, sino que nos señalan cuál es su posición dentro de un espacio de color específico [Wik02].

Existen diversos espacios de color de uso común en diferentes industrias. Los sistemas de cómputo, utilizan por lo general tres componentes o atributos para describir un color: rojo, verde y azul –espacio de color RGB–. La imprenta utiliza normalmente las componentes: cian, magenta y amarillo –espacio de color CMY–[Bou95].

Otros espacios de color utilizados son:

- HSI – Tono, saturación e intensidad –hue, saturation, intensity; en inglés–.
- HSV – Tono, saturación y valor –hue, saturation, value; en inglés–.
- YIQ – Iluminación, fase y cuadratura –luminance, in-phase, quadrature; en inglés–.
- CIELAB – La Comisión Internacional sobre iluminación –Commission Internationale de l'Eclairage–, propuso un modelo como estándar. Este modelo dimensiona la totalidad del espectro visible. Considera el espacio en forma uniforme y despliega tres ejes espaciales: L –luz, blanco-negro–, A –rojo-verde–, B –amarillo-azul–.

Espacio de color RGB

RGB es conocido como un espacio de color aditivo ya que, todos los colores dentro de este espacio son entendidos como la suma de las componentes rojo(R), verde(G) y azul(B), que son las que lo definen. Estos tres colores fueron elegidos porque corresponden aproximadamente a los tres tipos de conos sensitivos al color en el ojo humano –65 % sensibles al rojo, 33 % sensibles al verde y 2 % sensibles al azul– [Wik02].

Este modelo es frecuentemente visualizado a través de un cubo unitario. Cada componente de color es asignada a uno de los tres ejes ortogonales de coordenadas del espacio tridimensional y cada una de ellas varía desde la no contribución hasta la presencia de un color completamente saturado, Figura 2.1 [Bou95].

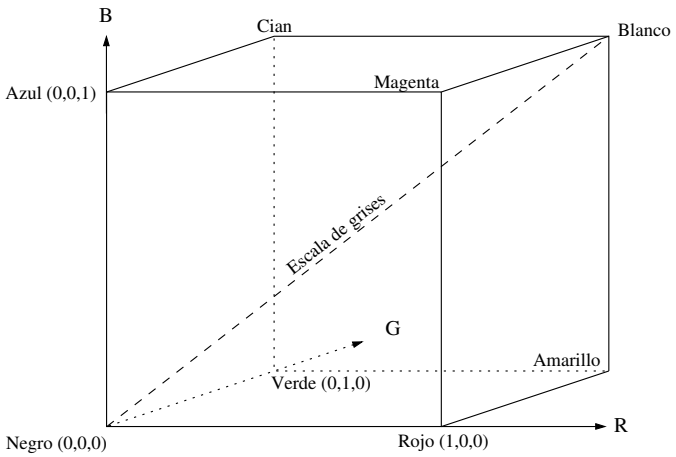


Figura 2.1: Cubo RGB.

Las imágenes pueden ser vistas, en este modelo, como tres imágenes independientes –una por cada canal de color– que cuando son emitidas por algún medio diseñado para tal efecto –un monitor, por ejemplo–, sus componentes, en forma de luz, son mezcladas y observadas por el ojo humano como una única imagen en colores.

2.2. Pre-procesamiento

Antes de buscar y analizar las cualidades que permiten la clasificación de un objeto o cosa en una imagen, es necesario acondicionarla, de forma

tal, que las características que nos interesan sean más fácilmente identificables, al mismo tiempo que se eliminan aquéllas que no aportan información relevante.

2.2.1. Histograma de intensidades

Muestra la forma en que las diferentes intensidades –saturaciones– de un color son utilizadas en una imagen. El histograma grafica el número de píxeles con una intensidad de color en particular, contra el valor de intensidad de color.

Para imágenes en color de 24-bits y en el espacio de color RGB –8 bits por canal–, el análisis de histograma puede llevarse a cabo para cada componente de color, partiendo del negro –valor 0– hasta finalizar con el rojo intenso en el caso de la componente R –valor 255–, verde intenso para la componente G y azul intenso para la componente B.

La Figura 2.2 muestra una imagen en color RGB de 24 bits con sus correspondientes histogramas.

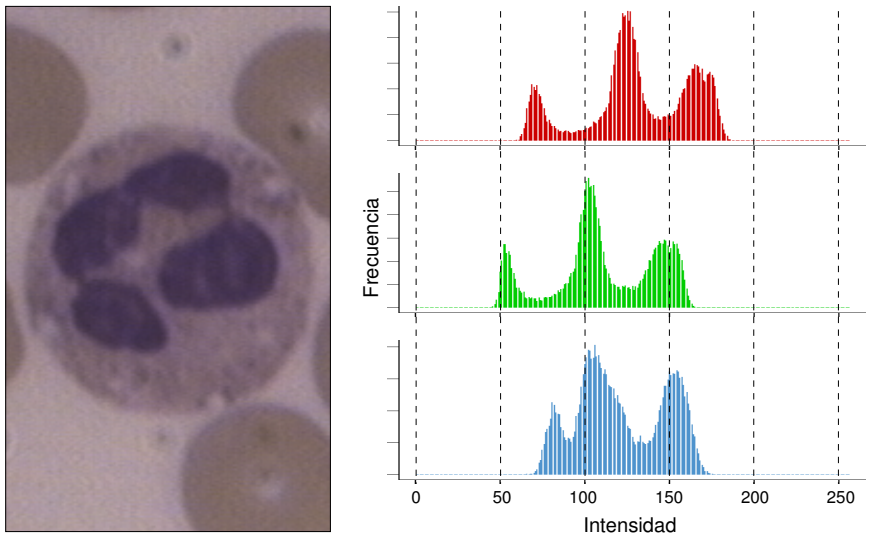


Figura 2.2: Histogramas RGB de un leucocito.

Normalización de histograma

Es una técnica popular para extender y trasladar el rango de intensidades de una imagen. El histograma original es extendido para que ocupe todos los valores posibles de intensidad de color. En el caso de las imágenes en color RGB de 24-bits y de dimensiones $M \times M$, se toma el valor más bajo y el más alto de las tres componentes y ese rango es ampliado para que ocupe los 256 niveles de cada componente. Si el histograma original de una imagen $\mathbf{O}$ comienza con un valor $\mathbf{O}_{min}$ y se extiende hasta el nivel de intensidad de color $\mathbf{O}_{max}$, podemos escalar el histograma de tal forma que los píxeles en la nueva imagen $\mathbf{N}$ tengan un nivel mínimo $\mathbf{N}_{min}$ y uno máximo $\mathbf{N}_{max}$ simplemente escalando los niveles de intensidad de entrada [NA02]. La siguiente ecuación define este proceso.

$$\mathbf{N}_{x,y} = \frac{\mathbf{N}_{max} - \mathbf{N}_{min}}{\mathbf{O}_{max} - \mathbf{O}_{min}} \times (\mathbf{O}_{x,y} - \mathbf{O}_{min}) + \mathbf{N}_{min} \quad \forall x, y \in [1, M] \qquad (2.1)$$

La Figura 2.3 muestra la imagen normalizada de la Figura 2.2 así como los histogramas ampliados que le corresponden.

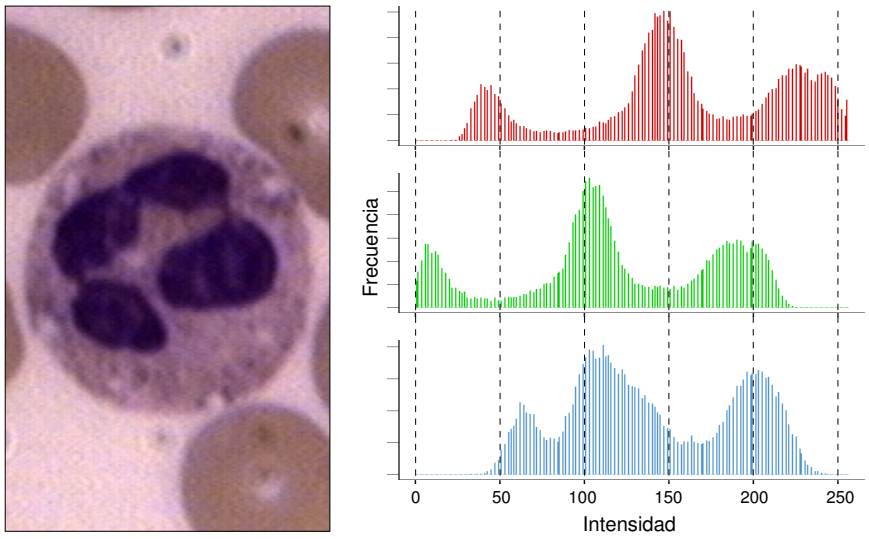


Figura 2.3: Histogramas RGB de la imagen de un neutrófilo una vez aplicada la normalización.

2.2.2. Transformaciones básicas de morfología matemática

En sus orígenes la morfología matemática fué desarrollada como una teoría de conjuntos; correspondiéndose con imágenes binarias. Años más tarde fué exitosamente generalizada a imágenes en escala de grises.

Las transformaciones básicas de conjuntos utilizadas en la morfología matemática –dilatación y erosión– se definen en términos de la interacción de la imagen bajo estudio y un elemento estructural. El elemento estructural es escogido para emparejar o igualar las estructuras geométricas en las cuales estamos interesados.

Las definiciones y propiedades siguientes han sido extraídas de [dB92].

Conjuntos y operadores de conjuntos

La teoría de la morfología matemática se fundamenta en unas pocas operaciones elementales sobre conjuntos, las cuales son definidas enseguida.

Un conjunto X es una colección de vectores de posición en el espacio observado –normalmente el espacio bidimensional continuo o discreto–. Definimos un píxel $x \in \mathbb{R}^n$ –o $x \in \mathbb{Z}^n$– como un vector de posición. Por lo tanto, para una imagen binaria un píxel tiene solamente una propiedad, la cual indica si el mismo es parte del conjunto que define al objeto bajo estudio o es parte del conjunto que define el fondo –*background*– de la imagen.

Definición 1 (Traslación de conjunto)**.** *La traslación de un conjunto X sobre un vector de desplazamiento t se denota por X_t y se define como:*

$$X_t = \{x \mid x - t \in X\}$$

Definición 2 (Complemento de conjunto)**.** *El complemento de un conjunto X se denota por X^c y se define como:*

$$X^c = \{x \mid x \notin X\}$$

Definición 3 (Unión de conjuntos)**.** *La unión de dos conjuntos X y Y se denota por $X \cup Y$ y se define como:*

$$X \cup Y = \{x \mid x \in X \text{ o } x \in Y\}$$

Definición 4 (Intersección de conjuntos)**.** *La intersección de dos conjuntos X y Y se denota por $X \cap Y$ y se define como:*

$$X \cap Y = \{x \mid x \in X \text{ y } x \in Y\}$$

Definición 5 (Transposición de conjunto –reflejo–)**.** *La transposición de un conjunto X se denota por $\tilde{X}$ y se define como:*

$$\tilde{X} = \{x \mid -x \in X\}$$

Las anteriores definiciones de operaciones sobre conjuntos son igualmente válidas para $\mathbb{R}^n$ y $\mathbb{Z}^n$. La siguiente operación sobre un conjunto no es fácilmente formalizada para imágenes discretas. Se recomienda ver [dB92] para una explicación más detallada acerca de este tema.

Definición 6 (Escalado de conjunto)**.** *Sea X un conjunto en $\mathbb{R}^n$ y sea $\alpha \in \mathbb{R}$ entonces el conjunto escalado αX se define como:*

$$\alpha X = \{\alpha x \mid x \in X\}$$

Transformaciones dual, dilatación y erosión

En la morfología matemática las tranformaciones de conjuntos vienen en pares. Lo que significa que definiendo una transformación de un conjunto, implícitamente definimos también su *transformación dual*. Sea ψ una transformación de conjunto tal que un conjunto X es transformado en otro conjunto $\psi(X)$.

Definición 7 (Transformación dual)**.** *Sea X un conjunto y sea ψ una transformación de conjunto; entonces su tranformación dual ψ^* se define como:*

$$\psi^*(X) = [\psi(X^c)]^c$$

Una transformación *dual-misma* es una transformación ψ tal que $\psi = \psi^*$.

Definición 8 (Adición de conjunto de Minkowski)**.** *La adición de conjunto de Minkowski de dos conjuntos X y S se define como:*

$$X \oplus S = \bigcup_{x \in S} X_x$$

Definición 9 (Sustracción de conjunto de Minkowski)**.** *La sustracción de conjunto de Minkowski de dos conjuntos X y S se define como:*

$$X \ominus S = \bigcap_{x \in S} X_x$$

Propiedad 1 (Dualidad de la adición y sustracción de Minkowski)**.** *La adición de Minkowski y la sustracción de Minkowski son transformaciones duales.*

Definición 10 (Dilatación)**.** *La dilatación de un conjunto X por un conjunto S se define como:*

$$X \boxplus S = X \oplus \tilde{S}$$

Definición 11 (Erosión)**.** *La erosión de un conjunto X por un conjunto S se define como:*

$$X \boxminus S = X \ominus \tilde{S}$$

En la práctica del procesamiento de imágenes con morfología matemática es más frecuente utilizar la dilatación y erosión como transformaciones básicas de conjunto en lugar de la suma y sustracción de Minkowski ya que las primeras son geométricamente más fáciles de interpretar.

En la dilatación $X \boxplus S$ o erosión $X \boxminus S$ el conjunto X frecuentemente se corresponde con la imagen bajo estudio. El conjunto S es conocido como el elemento estructural.

La dilatación, en términos geométricos, puede ser entendida como sigue. El elemento estructural S es deslizado sobre la imagen original X. Cada elemento de S_x con el vector de posición x coincidiendo con algún elemento de X, es elemento del conjunto dilatado.

La erosión tiene una interpretación de igual simplicidad. Se desliza el elemento estructural S sobre la imagen original X. Cada posición x en la cual el elemento estructural S_x coincida completamente con el conjunto X, es un elemento del conjunto erosionado.

Propiedad 2 (Propiedad Hit)**.** *La dilatación $X \boxplus S$ puede escribirse como:*

$$X \boxplus S = \{x \mid X \cap S_x \neq \varnothing\}$$

Propiedad 3 (Propiedad de inclusión)**.** *La erosión $X \boxminus S$ puede escribirse como:*

$$X \boxminus S = \{x \mid S_x \subset X\}$$

Propiedad 4. *La dilatación y la erosión de un conjunto X con un elemento estructural que contiene solamente un punto t, da como resultado la traslación del conjunto original:*

$$X \boxplus \{t\} = X \boxminus \{t\} = X_{-t}$$

2.3. Extracción de características

Enseguida se describen y definen brevemente las herramientas teóricas utilizadas en este trabajo, con las que se fundamenta la búsqueda y análisis de cualidades presentes en las imágenes; cualidades que nos permiten diferenciar e identificar grupos de píxeles que representan objetos en la imagen bajo estudio.

Las definiciones siguientes han sido extraídas de [NA02].

2.3.1. Descriptores de región – geométricos

Considerando las propiedades geométricas de una región en una imagen, ésta puede ser descrita a través de mediciones escalares, como lo son el área, perímetro, compactibilidad y dispersión.

Área

Es la propiedad más simple de una región en un plano y se define como:

$$A(S) = \int_x \int_y I(x, y)\, dy\, dx \tag{2.2}$$

donde $I(x, y) = 1$ si el píxel se encuentra dentro de los límites de la forma observada –$x, y \in S$–, 0 en otro caso. Dada la naturaleza de las imágenes digitales, la integral anterior debe ser aproximada por sumatorias, así:

$$A(S) = \sum_x \sum_y I(x, y)\, \Delta A \tag{2.3}$$

donde ΔA es el área de un píxel. Por lo tanto, si $\Delta A = 1$ el área es medida en píxeles. Esta propiedad varía cuando la imagen es escalada, sin embargo, permanece sin cambio ante la rotación. Debido a la discretización de las imágenes digitales, pequeños errores o cambios en el valor de esta propiedad pueden aparecer cuando la imagen es rotada.

Perímetro

El perímetro es una propiedad más de la región. Para definirlo formalmente tenemos que si $x(t)$ y $y(t)$ denotan las coordenadas paramétricas de una curva que encierra una región S, entonces el perímetro de la región se define como:

$$P(S) = \int_t \sqrt{x^2(t) + y^2(t)}\, dt \tag{2.4}$$

Esta ecuación define la suma de todos los arcos infinitesimales que componen a la curva. Es necesario, para el caso práctico, definir la ecuación en términos de sumatorias que den tratamiento al caso discreto. Si $x(t)$ y $y(t)$ están definidas por conjuntos de píxeles en la imagen, entonces la Ecuación 2.4 puede ser aproximada por:

$$P(S) = \sum_i \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \tag{2.5}$$

donde x_i y y_i representan las coordenadas del i-ésimo píxel que forma parte de la curva. Dada la organización de los píxeles en una imagen –forman una malla de cuadros–, los términos de la sumatoria pueden solamente tomar dos valores. Cuando los píxeles (x_i, y_i) y (x_{i-1}, y_{i-1}) presentan una conectividad de tipo 4, el término de la sumatoria es igual a la unidad; y tiene un valor de $\sqrt{2}$ en caso de las conexiones diagonales presentes en la conectividad 8. La Figura 2.4 ilustra los casos mencionados.

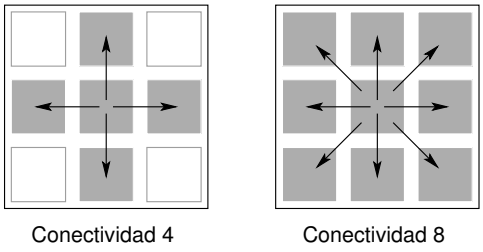


Figura 2.4: Malla de cuadros de conectividad 4 y 8.

Compactibilidad

Con base en las mediciones del área y el perímetro es posible calcular otras propiedades de la región. La compactibilidad es una de ellas y se define como:

$$C(S) = \frac{4\pi A(s)}{P^2(s)} \tag{2.6}$$

Esta ecuación puede ser reescrita como:

$$C(S) = \frac{A(s)}{P^2(s)/4\pi} \tag{2.7}$$

En esta forma de la ecuación es posible observar claramente su significado. Si la región bajo estudio cuenta con un área A y un perímetro P, la

compactibilidad mide la razón de A entre el área del círculo definido por un perímetro de longitud P. En otras palabras, la compactibilidad mide la eficiencia con que un contorno encierra un área. Es claro entonces que si la región observada describe exactamente un círculo, el valor de la compactibilidad será la unidad, el cual es el valor máximo de la misma, entonces, para cualquier otra forma adoptada por la región, el valor de compactibilidad será menor que uno.

Dispersión

La dispersión o irregularidad es la razón de la cuerda de mayor longitud de la región entre el área de la misma y puede ser definida como:

$$I(S) = \frac{\pi \max\left[(x_i - \overline{x})^2 + (y_i - \overline{y})^2\right]}{A(S)} \tag{2.8}$$

donde $(\overline{x}, \overline{y})$ representan las coordenadas del centro de masa –Ecuación 2.12– de la región. El numerador define el área del círculo mínimo con centro ubicado en el centro de masa de la región y que encierra completamente a la misma. Así, la dispersión describe la densidad de la región. Otra forma de definir la dispersión es la siguiente:

$$IR(S) = \frac{\max \sqrt{(x_i - \overline{x})^2 + (y_i - \overline{y})^2}}{\min \sqrt{(x_i - \overline{x})^2 + (y_i - \overline{y})^2}} \tag{2.9}$$

La expresión define la razón del radio del círculo con centro ubicado en centro de masa de la región y que encierra completamente a la misma, entre el radio del círculo máximo con el mismo centro que puede ser inscrito en la región.

2.3.2. Descriptores de región – momentos

Este tipo de descriptores se concentran en la organización que los píxeles de la región observada presentan. Pueden verse como una descripción global de la región bajo estudio.

Momentos cartesianos bidimensionales

Van de órdenes menores –iniciando en 0– a órdenes mayores. El momento de orden p y q, m_{pq} de una función $I(x, y)$ se define como:

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q I(x, y)\, dx\, dy \tag{2.10}$$

Y la aproximación que da tratamiento al caso discreto se define como:

$$m_{pq} = \sum_x \sum_y x^p y^q I(x,y) \Delta A \tag{2.11}$$

Estos *momentos* tienen la propiedad de que si la función I satisface ciertas condiciones [NA02], se garantiza que los momentos de todos los órdenes existen. Además, el conjunto de los momentos de una función la determinan de forma única.

El momento de orden $p = 0$ y $q = 0$, m_{00}, determina la masa total de la función. Esta ecuación es igual a la Ecuación 2.3 cuando la función I toma valores de cero y uno.

Para imágenes binarias, el centro de masa $(\overline{x}, \overline{y})$ de la región bajo estudio puede ser calculado como:

$$\overline{x} = \frac{m_{10}}{m_{00}} \quad \overline{y} = \frac{m_{01}}{m_{00}} \tag{2.12}$$

Con este estimado de las coordenadas centrales de la región, el cual puede traducirse como un punto de referencia para la forma observada misma, es posible calcular los *momentos centralizados*, μ_{pq}, los cuales tienen la propiedad de ser invariantes a la traslación y que se definen como:

$$\mu_{pq} = \sum_x \sum_y (x - \overline{x})^p \, (y - \overline{y})^q \, I(x,y) \, \Delta A \tag{2.13}$$

Hay que notar que el momento centralizado de orden $p = 0$ y $q = 0$, μ_{00}, sigue definiendo el área para imágenes binarias. Los dos momentos de primer orden μ_{10} y μ_{01} son iguales a cero. Los momentos de orden dos y órdenes mayores, presentan propiedades descriptivas de la forma observada.

Momentos centrales normalizados

Los *momentos centrales normalizados*, η, poseen la propiedad, además de ser invariantes a la traslación, de ser invariantes al cambio de escala y son definidos como [NA02]:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{2.14}$$

donde:

$$\gamma = \frac{p+q}{2} \quad \forall \, p + q \geq 2 \tag{2.15}$$

A partir de las dos ecuaciones anteriores, es posible definir siete momentos que adicionan la propiedad de ser invariantes a la rotación, también conocidos como *momentos invariantes de Hu*, y se definen como:

$$
\begin{aligned}
M1 &= \eta_{20} + \eta_{02} \\
M2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
M3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
M4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
M5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) + ((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2) \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) \\
M6 &= (\eta_{20} - \eta_{02})((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
M7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2) \\
&\quad + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})(3(\eta_{12} + \eta_{30})^2 - (\eta_{21} + \eta_{03})^2)
\end{aligned}
\tag{2.16}
$$

El orden de cada término de las expresiones anteriores es igual a la suma de sus respectivas variables p y q. El orden de cada momento invariante lo determina el término de mayor orden. De esta forma, los dos primeros momentos son de orden dos, y los cinco restantes son de orden tres. El momento $M7$ se considera invariante a la inclinación y fué introducido para distinguir imágenes reflejadas.

2.3.3. Descriptores de textura

Aún cuando la textura está ligada directamente a la percepción e interpretación humana y no ha sido definida claramente en términos matemáticos, es evidente, de acuerdo con las múltiples definiciones lingüísticas que existen de la misma, que se refiere a cierta organización que guardan los componentes superficiales de un objeto y que excitan nuestros sentidos en una forma particular.

En imágenes digitales podría entenderse como la disposición y orden que guardan los píxeles de una parte de una región, y que se repite en forma de patrón a través de toda la superficie de ésta última.

Las siguientes definiciones de descriptores de textura, tienen como fundamento principal a la transformada de Fourier, la cual permite lograr invariabilidad a la rotación. Las expresiones siguientes se definen en términos de la Transformada Rápida de Fourier –FFT, por sus siglas en inglés–.

Denotamos como **FP** al resultado de aplicar la FFT a una imagen.

$$\mathbf{FP} = \mathrm{FFT}(\mathbf{P})$$

donde $\mathbf{FP}_{u,v}$ y $\mathbf{P}_{x,y}$ son los datos de la trasformada y el píxel, respectivamente.

El resultado de la transformada se normaliza, para lograr que las magnitudes sean invariantes a cambios lineales en la iluminación en la imagen, de acuerdo con la siguiente expresión:

$$\mathbf{NFP}_{u,v} = \frac{|\mathbf{FP}_{u,v}|}{\sqrt{\sum_{u\neq 0 \wedge v\neq 0} |\mathbf{FP}_{u,v}|^2}}$$

Con base en este resultado se plantean ahora las expresiones que permiten obtener mediciones escalares de las cualidades de la textura.

Energía

$$e = \sum_{u=1}^{\mathbf{N}} \sum_{v=1}^{\mathbf{N}} (\mathbf{NFP}_{u,v})^2 \tag{2.17}$$

Entropía

$$h = \sum_{u=1}^{\mathbf{N}} \sum_{v=1}^{\mathbf{N}} \mathbf{NFP}_{u,v} \log(\mathbf{NFP}_{u,v}) \tag{2.18}$$

Inercia

$$i = \sum_{u=1}^{\mathbf{N}} \sum_{v=1}^{\mathbf{N}} (u-v)^2 \mathbf{NFP}_{u,v} \tag{2.19}$$

Capítulo 3

Razonamiento Probabilista

Una vez que se ha completado la extracción de características de la imagen, es necesario evaluar esa información mediante alguna técnica o método que nos permita sacar conclusiones acerca de la naturaleza de la imagen bajo estudio, es decir, es necesario procesar la información para lograr el reconocimiento y clasificación de los objetos que en la imagen se observan. Dentro del área de la Inteligencia Artificial –IA– se encuentran las disciplinas: Reconocimiento de Patrones y Aprendizaje Máquina, las cuales proporcionan estos métodos o herramientas teóricas que permiten llevar a cabo la clasificación.

La información obtenida de las imágenes de los leucocitos no determina perfectamente la clase a que pertenece cada uno de los objetos contenidos en las mismas, es decir, el dominio de la información es incierto, ya que no es posible asegurar que el valor de cada uno de los datos obtenidos es exclusivo de alguna clase de objeto. Es por esto que el presente capítulo se concentra en presentar un sistema de *razonamiento incierto* que, en teoría, permita llevar a cabo nuestro objetivo.

Entre las diversas teorías y métodos que se encargan de solucionar problemas cuyo dominio es incierto encontramos el razonamiento predefinido, algunos métodos basados en reglas, la teoría de Dempster-Shafer, los conjuntos difusos y la lógica difusa [SJN00]. El tratar de solucionar nuestro problema a través de cada una de las teorías y métodos mencionados y llevar a cabo un estudio comparativo de los resultados sería una labor muy ardua e igualmente interesante que, sin embargo, iría mucho más lejos del alcance de este trabajo. Por esta razón, se ha utilizado el razonamiento probabilista como método para dar tratamiento a nuestro problema de clasificación, ya que, como se menciona en [SJN00] con respecto de las primeras teorías mencionadas,

> "En todos los sistemas *funcionales de verdad* hay serios problemas relacionados con el razonamiento mezclado o intercausal."

se menciona además,

> "La información sobre la independencia condicional es una forma vital y sólida de estructurar información sobre un dominio incierto."

Más adelante se observará que el razonamiento probabilista y en particular las redes bayesianas –RB–, también conocidas como redes de creencia, ofrecen una manera natural de representar la información sobre la independencia condicional.

Es importante señalar que el razonamiento eficiente mediante probabilidades es tan reciente que las redes de creencia son el único método, del cual existen sólo ligeras variantes.

3.1. Redes bayesianas

En una red bayesiana, cada nodo del grafo que la representa, se corresponde con una variable aleatoria, así, en adelante no se hará distinción entre estos dos conceptos y ambos serán representados por letras mayúsculas.

Para definiciones básicas de probabilidad y distribuciones de probabilidad véase el Apéndice A.

3.1.1. Definición de red bayesiana

Primeramente definiremos la separación direccional, también conocida como *separación-d*, la cual es un concepto fundamental de las redes bayesianas.

Definición 12 (Separación direccional [DV05])**.** *Dado un grafo dirigido acíclico conexo y una distribución de probabilidad sobre sus variables, se dice que hay separación direccional si, dado un nodo X, el conjunto de sus padres, pa(X), separa condicionalmente este nodo de todo otro subconjunto $\bar{Y}$ en que no haya descendientes de X. Es decir,*

$$P(x|pa(x), \bar{y}) = P(x|pa(x)) \tag{3.1}$$

La separación direccional nos indica que si queremos calcular la probabilidad a posteriori de alguna variable X y conocemos los valores de $pa(X)$,

ningún otro nodo, que no sea descendiente de X, nos aporta mayor información de la que ya conocemos.

Una vez expuesta la separación direccional procedemos a definir formalmente lo que es una red bayesiana.

Definición 13 (Red bayesiana [DV05])**.** *Es un grafo dirigido acíclico conexo más una distribución de probabilidad sobre sus variables, que cumple con la propiedad de separación direccional.*

Es posible observar tres propiedades que la separación-d atribuye a las redes bayesianas [DV05]:

1. Dos nodos cualesquiera X e Y que no tengan ningún antepasado común son independientes a priori.

2. Si X es padre de Y e Y es padre de Z y no existe otro camino de X a Z, entonces estos dos nodos quedan condicionalmente separados por Y:
$$P(z|x, y) = P(z|y)$$

3. Si Y y Z son hijos de X y no tienen otro antepasado común, entonces X separa a Y y Z, haciéndolos condicionalmente independientes.

Otra forma de visualizar estas propiedades que puede ayudarnos a comprenderlas mejor se encuentra en [Nil01] y es la siguiente:

Independencia condicional mediante nodos bloqueadores Dos nodos X_i y X_j son independientes condicionalmente dado un conjunto de nodos ε si por cada camino no dirigido entre X_i y X_j hay algún nodo X_b, que cumple alguna de las siguientes tres propiedades –ver Figura 3.1–:

1. X_b pertenece a ε, y ambos arcos salen de X_b.
2. X_b pertenece a ε, y un arco va hacia X_b y el otro sale de él.
3. Ni X_b ni ningún descendiente suyo pertenece a ε, y ambos arcos van hacia X_b.

Si alguna de las condiciones anteriores se cumple, se dice que X_b bloquea el camino dado ε. Si todos los caminos entre X_i y X_j están bloqueados, decimos que ε *d-separa*–separa direccionalmente – X_i de X_j y se concluye que X_i y X_j son independientes condicionalmente dado ε.

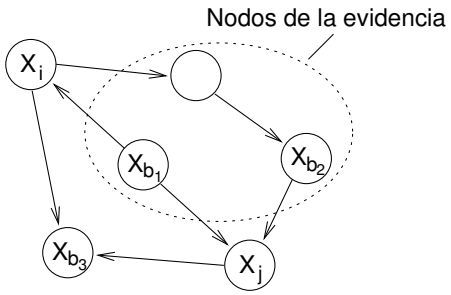


Figura 3.1: Casos de *independencia condicional* mediante nodos bloqueadores.

Partiendo de la Ecuación A.6 y en vista de la separación direccional, es posible expresar la distribución de probabilidad conjunta de una red bayesiana mediante el producto de las distribuciones condicionadas de cada nodo dados sus padres. El siguiente teorema formaliza matemáticamente este hecho.

Teorema 1 (Factorización de la probabilidad [DV05])**.** *Dada una red bayesiana, su distribución de probabilidad puede expresarse como:*

$$P(x_1, \ldots, x_n) = \prod_i P(x_i|pa(x_i)) \tag{3.2}$$

La importancia de este teorema radica en que nos permite describir una red bayesiana mediante la probabilidad condicionada de cada nodo, en vez de la distribución de probabilidad conjunta, la cual, requeriría un número exponencial de parámetros en el número de nodos –tal es el caso del método probabilista clásico– y plantearía el problema de verificar la propiedad de separación direccional.

Si se compara la expresión 3.2 con la Ecuación A.6, veremos en general:

$$P(X_i|X_{i-1}, \ldots, X_1) = P(X_i|Pa(X_i)) \tag{3.3}$$

3.1.2. Semántica de las redes de creencia

La estructura de una red de creencia aporta por sí misma mucha información cualitativa. Si entre un par de variables existe un enlace, nos damos cuenta de inmediato, aún sin conocer sus probabilidades condicionales, que existe correlación entre ellas. Tal relación entre variables se conoce como

influencia causal directa: el valor que tome X influye sobre la probabilidad de Y y viceversa. Si entre un par de variables existe algún camino en el que participan variables intermedias hablaremos de *influencia causal indirecta*. Desde la misma perspectiva, la ausencia de arcos entre variables también aporta información.

Las relaciones de dependencia e independencia condicionales y a priori y los casos en los que las variables se vuelven dependientes e independientes son observables también desde la estructura de la red. Esto es, cuando dos variables no tienen ningún antepasado común se sabe que son independientes a priori; sin embargo, si comparten algún descendiente, el hecho de conocer el valor que toma tal descendiente, hace que surjan correlaciones entre ellas. Ahora bien, cuando dos variables tienen un solo antepasado común, sabemos que existe correlación entre ellas, sin embargo, esa correlación desaparece al momento de conocer el valor que toma tal antepasado. Estas dos formas en que se relacionan las variables –a través de algún descendiente o algún antepasado común– en las cuales surge o desaparece la correlación, nos permiten ver claramente la asimetría que existe entre variables padres e hijos, causas y efectos, de la cual proviene el nombre de la *separación direccional*.

Por último, hay que observar que la topología de una red puede verse como una base de conocimientos abstracta, la cual representa la estructura general de los procesos causales del dominio y que es válida en una gran diversidad de escenarios.

3.1.3. Redes bayesianas híbridas

Cuando una red bayesiana involucra variables tanto discretas como continuas se conoce como *híbrida*. En tales redes, se observan dos casos de interés particular al momento de especificar las probabilidades condicionadas de las variables [SJN04]:

1. Cuando una variable aleatoria continua tiene padres discretos y/o continuos.

2. Cuando una variable aleatoria discreta tiene padres continuos.

Considerando el primer caso, cuando una variable aleatoria continua tiene un padre discreto, las probabilidades condicionales que éste genera sobre el hijo continuo se manejan mediante enumeración explícita. Supongamos dos variables aleatorias, una discreta, X, y una continua, Y, si X es padre de Y, por cada posible valor x_i será necesario especificar una función de

densidad de probabilidad (*f.d.p.*) que determine el comportamiento condicionado de la variable continua Y, es decir,

$$\begin{aligned} P(Y|x^1) &= f_1 \\ &\vdots \\ P(Y|x^n) &= f_n \end{aligned}$$

El requerimiento de que cada f_i tiene que ser una *f.d.p.* surge porque es necesario cumplir con la Ecuación A.2.

Siguiendo el primer caso, cuando una variable aleatoria continua tiene un padre continuo, los parámetros de la *f.d.p.* del hijo se especifican como función del valor continuo del padre. Supongamos dos variables aleatorias continuas X e Y. Si X es padre de Y, es necesario definir una sola *f.d.p.* de probabilidad condicional para Y; los parámetros que determinan tal *f.d.p.* dependerán del valor de X. Si f_y es la *f.d.p.* que define la probabilidad condicional $P(Y|x)$ y θ es el conjunto de parámetros que la determinan, entonces,

$$P(Y|x) = f_y(\theta(x))$$

Como ejemplo podemos tomar la función Gaussiana lineal. En este caso, el nodo hijo presenta una *f.d.p.* Gaussiana cuya media μ varía linealmente con el valor del nodo padre y su desviación estándar se fija a un valor determinado.

$$P(Y|x) = N(ax+b, \sigma^2)(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-(ax+b)}{\sigma})^2}$$

En este ejemplo es necesario aportar los parámetros a, b y σ.

En caso de que una variable aleatoria continua tenga un padre discreto y uno continuo, las probabilidades condicionales de tal variable serán, para el padre discreto, definidas de forma explícita, y las del padre continuo mediante la especificación de los parámetros del hijo como función del valor del padre. Supongamos tres variables aleatorias, una discreta, X, y dos continuas, Y y Z, si X e Y son padres de Z, entonces las probabilidades condicionadas de Z se especificarían como sigue:

$$\begin{aligned} P(Z|x^1, y) &= f_1(\theta_1(y)) \\ &\vdots \\ P(Z|x^n, y) &= f_n(\theta_n(y)) \end{aligned} \tag{3.4}$$

donde θ_i representa el conjunto de parámetros que determinan a la *f.d.p.* f_i.

Consideremos ahora el caso en que una variable discreta tiene un padre continuo. Una forma de abordar el problema es asignar funciones que definan la probabilidad para cada uno de los posibles valores de la variable discreta dependiendo cada función del valor que adquiera la variable padre. Si X es una variable aleatoria continua, Y es una discreta y X es padre de Y, sería necesario asignar funciones tales que:

$$\begin{aligned} P(y^1|x^i) &= f_1(x^i) \\ &\vdots \\ P(y^n|x^i) &= f_n(x^i) \end{aligned}$$

donde,

$$\sum_{j=1}^{n} f_j(x^i) = 1 \qquad \forall x^i$$

El siguiente ejemplo ha sido tomado de [SJN04].

> Suponga una variable continua *Costo* y una variable discreta *Compras* –ver Figura 3.2–. Parece razonable asumir que el cliente comprará si el Costo es bajo y no comprará si el Costo es alto y que la probabilidad de la compra varía suavemente en algunas regiones intermedias. En otras palabras, la distribución condicionada es como una función *umbral* suave.

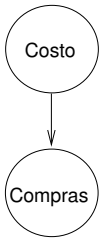


Figura 3.2: Variables *Costo* y *Compras* así como la relación entre ellas.

> Un modo de construir umbrales suaves es utilizar la *integral* de la distribución normal estándar:
>
> $$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)\, dx$$
>
> Entonces la probabilidad de *Compras* dado el *Costo* debe ser:
>
> $$P(Compras|Costo = c) = \Phi((-c + \mu)/\sigma)$$

lo que significa que el umbral del *Costo* se produce alrededor de μ, el ancho de la región de umbral es proporcional a σ, y la probabilidad de la compra decrece cuando el *Costo* crece.

Nótese que en el ejemplo presentado se ha definido la probabilidad para un sólo posible valor de *Compras*. Suponiendo que tal variable sea booleana, es fácil observar que,

$$P(\overline{Compras}|Costo = c) = 1 - \Phi((-c + \mu)/\sigma)$$

Cuando se tienen múltiples padres continuos para un nodo hijo discreto, el caso anterior puede generalizarse tomando una combinación lineal de los valores de los padres [SJN04].

3.1.4. Inferencia en poliárboles

El siguiente algoritmo de propagación de la evidencia en poliárboles* tiene como base el paso de mensajes π y λ [DV05].

Definiciones básicas

En un poliárbol, la influencia de cada hallazgo se propaga hasta un nodo X bien a través de los padres o a través de los hijos, por lo que para cada nodo X se puede hacer una partición de la evidencia en subconjuntos tales que:

$$\mathbf{e} = \mathbf{e}_X^+ \cup \mathbf{e}_X^-$$

$$\mathbf{e}_X^+ \cap \mathbf{e}_X^- = \varnothing$$

donde $\mathbf{e}_X^+$ representa la evidencia "por encima de X" y $\mathbf{e}_X^-$ "por debajo de X".

La eliminación de un enlace XY –en caso de poliárboles– divide a la red –y por tanto a la evidencia– en dos partes, una que queda "por encima" del enlace y otra que queda "por debajo". Las llamaremos $\mathbf{e}_{XY}^+$ y $\mathbf{e}_{XY}^-$, respectivamente. Además, se cumple que:

$$\mathbf{e} = \mathbf{e}_{XY}^+ \cup \mathbf{e}_{XY}^-$$

$$\mathbf{e}_{XY}^+ \cap \mathbf{e}_{XY}^- = \varnothing$$

*Un poliárbol es un tipo de grafo para el cual se cumple que entre cualesquiera de sus nodos existe exactamente un camino.

Con base en la partición de la evidencia, se pueden establecer las siguientes definiciones:

$$\begin{aligned} \pi(x) &\equiv P(x, \mathbf{e}_X^+) \\ \lambda(x)g &\equiv P(\mathbf{e}_X^-|x) \\ \pi_X(u_i) &\equiv P(u_i, \mathbf{e}_{U_iX}^+) \\ \lambda_{Y_j}(x) &\equiv P(\mathbf{e}_{XY_j}^-|x) \end{aligned}$$

El sentido de estas definiciones es el siguiente:

- $\pi(x)$ indica qué valor de X es más probable según la evidencia relacionada con las causas de X –según la evidencia "por encima" de X–.
- λ indica qué valor de X explica mejor los hallazgos correspondientes a los efectos de X –la evidencia "por debajo" de X–.
- $\pi_X(u)$ indica qué valor de U es más probable según la evidencia "por encima" del enlace UX.
- $\lambda_{Y_j}(x)$ indica qué valor de X explica mejor la evidencia "por debajo" del enlace XY.

Computación de los mensajes

El objetivo perseguido es el cálculo de la probabilidad a *posteriori* de cada nodo. Con vista en este objetivo se plantea:

$$\begin{aligned} P^*(x) &= P(x|\mathbf{e}) = \alpha P(x, \mathbf{e}_X^+, \mathbf{e}_X^-) \\ &= \alpha P(x, \mathbf{e}_X^+) P(\mathbf{e}_X^-|x, \mathbf{e}_X^+) \end{aligned}$$

donde se ha definido:

$$\alpha \equiv [P(\mathbf{e})]^{-1}$$

Ahora bien, por la separación direccional se sabe que $P(\mathbf{e}_X^-|\, x, \mathbf{e}_X^+) = P(\mathbf{e}_X^-|\, x)$, de modo que, aplicando las definiciones anteriores se tiene:

$$P^*(x) = \alpha \pi(x) \lambda(x)$$

Es necesario calcular los tres factores de esta expresión. Se comienza con $\pi(x)$. Según su definición:

$$\pi(x) = P(x, \mathbf{e}_X^+) = \sum_{\bar{u}} P(x|\bar{u}) P(\bar{u}, \mathbf{e}_X^+)$$

Como las causas de X no tienen ningún antepasado común, por estar en un poliárbol, todas ellas y las ramas correspondientes son independientes mientras no se considere la evidencia relativa a X o a sus descendientes:

$$\begin{aligned} P(\bar{u}, \mathbf{e}_X^+) &= P(u_1, \mathbf{e}_{U_1X}^+, \ldots, u_n, \mathbf{e}_{U_nX}^+) \\ &= \prod_{i=1}^{n} P(u_i, \mathbf{e}_{U_iX}^+) = \prod_{i=1}^{n} \pi_X(u_i) \end{aligned} \tag{3.5}$$

Por tanto,

$$\pi(x) = \sum_{\bar{u}} P(x|\bar{u}) \prod_{i=1}^{n} \pi_X(u_i)$$

Calculando ahora $\pi_X(u_i)$ o, lo que es lo mismo, $\pi_{Y_j}(x)$, puesto que en una RB todos los nodos son equivalentes. La evidencia que está por encima del enlace XY_j, $\mathbf{e}_{XY_j}^+$, puede descomponerse en varios subconjuntos: La que está por encima de X y la que está por debajo de cada enlace XY_k. Se sabe, además, que X separa $\mathbf{e}_X^+$ de $\mathbf{e}_{XY_k}^-$, y separa también los subconjuntos $\mathbf{e}_{XY_k}^-$ entre sí. Con estas consideraciones se obtiene:

$$\begin{aligned} \pi_{Y_j}(x) &= P(x, \mathbf{e}_{XY_j}^+) = P(x, \mathbf{e}_X^+, \mathbf{e}_{XY_k}^{-\ k\neq j}) \\ &= P(x, \mathbf{e}_X^+) \prod_{k\neq j} P(\mathbf{e}_{XY_k}^-|x) \\ &= \pi(x) \prod_{k\neq j} \lambda_{Y_k}(x) \end{aligned}$$

Para calcular esta expresión, es necesario hallar $\lambda_{Y_k}(x)$ –o $\lambda_{Y_j}(x)$, pues el resultado será válido para todos los efectos de X–. Representando mediante $\bar{V}$ el conjunto de causas de Y_j distintas de X. Para simplificar la notación, escribiremos $\mathbf{e}_{\bar{V}Y_j}^+ = \mathbf{e}_{V_1Y}^+ \cup \ldots \cup \mathbf{e}_{V_pY}^+$, con lo que queda $\mathbf{e}_{XY_j}^- = \mathbf{e}_Y^- \cup \mathbf{e}_{\bar{V}Y}^+$.

Recordando que Y_j separa $\mathbf{e}_{Y_j}^-$ del resto de la red que está por encima de Y_j, e igualmente los padres de Y_j separan Y_j de $\mathbf{e}_{\bar{V}Y_j}^+$. Aplicando repetidamente la proposición A.9 , resulta:

$$\begin{aligned} \lambda_{Y_j}(x) &= P(\mathbf{e}_{XY_j}^-|x) \\ &= \sum_{y_j} \sum_{\bar{v}} P(\mathbf{e}_{Y_j}^-, y_j, \mathbf{e}_{\bar{V}Y_j}^+, \bar{v}|x) \\ &= \sum_{y_j} \sum_{\bar{v}} P(\mathbf{e}_{Y_j}^-|y_j) P(y_j|\bar{v}, x) P(\mathbf{e}_{\bar{V}Y_j}^+, \bar{v}|x) \end{aligned}$$

Ya que las causas de Y_j son independientes a priori, se utiliza la ecuación 3.5 para llegar a:

$$P(\bar{v}, \mathbf{e}^+_{\bar{V}Y_j}|x) = P(\bar{v}, \mathbf{e}^+_{\bar{V}Y_j}) = \prod_{l=1}^{p} P(v_l, \mathbf{e}^+_{V_lY_j}) = \prod_{l=1}^{p} \pi_{Y_j}(v_l)$$

y, en consecuencia,

$$\lambda_{Y_j}(x) = \sum_{y_j} \left[\lambda(y_j) \sum_{\bar{v}} P(y_j|x, \bar{v}) \prod_{l=1}^{p} \pi_{Y_j}(v_l) \right]$$

Finalmente, hay que calcular $\lambda(x)$, de la siguiente forma:

$$\begin{aligned} \lambda(x) &= P(\mathbf{e}^-_{XY_1}, \ldots, \mathbf{e}^-_{XY_m}|x) \\ &= \prod_{j=1}^{m} P(\mathbf{e}^-_{XY_j}|x) = \prod_{j=1}^{m} \lambda_{Y_j}(x) \end{aligned}$$

Para completar el algoritmo falta hallar la constante α. Para lo cual se debe considerar que:

$$\sum_x P^*(x) = \alpha \sum_x \pi(x)\lambda(x) = 1$$

con lo que se puede obtener α como:

$$\alpha = \left[\sum_x \pi(x)\lambda(x) \right]^{-1}$$

Observar que por cada enlace $X \rightarrow Y$ circulan dos mensajes, $\pi_Y(x)$ de X a Y, y $\lambda_Y(x)$, de Y a X. Ambos mensajes son vectores correspondientes a la variable X, mientras que la variable Y sólo aparece como subíndice en los dos.

3.2. Aprendizaje de redes bayesianas

3.2.1. Un método para la construcción de redes bayesianas

La Ecuación 3.2 hace posible que se puedan identificar los nodos de la RB de acuerdo con un orden congruente con el orden parcial implícito en la estructura gráfica. Con base en esto decimos que la red de creencia será la representación correcta del dominio, sólo si cada uno de los nodos tiene

independencia condicional respecto de sus predecesores en la secuencia de nodos. Por lo tanto, si se desea construir una red de creencia cuya estructura sea adecuada para el dominio, debemos escoger los padres de cada nodo de manera que se satisfaga esta propiedad. Observemos ahora que los padres del nodo X_i deben contener todos los nodos que están en $X_1, \ldots, X_{i-1}$ los cuales tienen influencia directa sobre X_i.

El procedimiento general para la construcción de una red en incrementos, es el siguiente [SJN00]:

1. Escoja el conjunto de variables $\bar{X}$ que sirva para describir el dominio.
2. Defina la manera como se van a ordenar las variables.
3. Siempre que haya variables:
   a) Por cada variable X_i que escoja, añada a la red un nodo.
   b) Asigne $Padres(X_i)$ a un conjunto mínimo de nodos que esté presente en la red, para de esta manera satisfacer la propiedad de independencia condicional 3.3.
   c) Elabore la tabla de probabilidad correspondiente a X_i.

Ya que en este proceso de construcción cada nodo se conecta únicamente con nodos anteriores, el método garantiza la obtención de una red acíclica. Una característica importante de las redes de creencia es que en ellas no hay valores de probabilidad redundantes, a excepción, tal vez, de una entrada en las hileras de cada una de las tablas de probabilidad condicional. Es decir, *es imposible que el ingeniero del conocimiento o el experto del dominio lleguen a crear una red de creencia que viole los axiomas de la probabilidad.*

3.2.2. Aprendizaje de las tablas de probabilidad condicional con datos completos

Cuando se dice que se cuenta con *datos completos* para el aprendizaje de una red, quiere decir que cada miembro del conjunto entrenamiento Ξ dispone de un valor para cada variable representada en la red.

Cuando disponemos de un amplio número de muestras de entrenamiento, sólo debemos calcular el estadístico muestral de cada nodo y de sus padres. Si queremos obtener la *tabla de probabilidad condicional* (TPC) de algún nodo V_i dados sus padres, comenzaremos por plantear tantas tablas para este nodo como valores v_i –menos uno– diferentes existan. En el caso de que V_i sea una variable boolena, que es el caso que asumimos, sólo hay una

TPC para cada nodo. Si V_i tiene k_i nodos padre, entonces deben existir 2^{k_i} entradas (filas) en la tabla, debido a que cada padre puede tener uno de dos valores posibles. Se denotan las variables asociadas a los padres de V_i con el vector $\bar{P}_i$. El estadístico muestral $\hat{p}(V_i = v_i|\bar{P}_i = \bar{p}_i)$ se obtiene a partir del número de casos de Ξ que tienen $V_i = v_i$ y $\bar{P}_i = \bar{p}_i$, dividido entre el número de casos que tiene $\bar{P}_i = \bar{p}$. Para aprender las TPCs, simplemente se utilizan estos estadísticos muestrales de los datos reales para todos los nodos de la red.

Si el cálculo de los estadísticos muestrales tienen como base muestras de tamaño muy pequeño, puede generarse una estimación imprecisa de las probabilidades subyacentes. En la mayoría de los casos, el crecimiento exponencial del número de parámetros de una TPC puede reducir la capacidad del conjunto de entrenamiento para generar buenas estimaciones. El problema puede ser mitigado si muchos parámetros tienen el mismo valor o uno muy cercano.

Capítulo 4

Diseño e implementación del clasificador

En el presente capítulo se expone el proceso de diseño teórico del clasificador de leucocitos, el cual se fundamenta en la teoría de procesamiento de imágenes y redes bayesianas ya presentada. Al final se presentan brevemente las características de su implementación.

Al ser el clasificador implementado como una red bayesiana, el proceso de construcción de cada uno de sus componentes debe apegarse a la definición de tal tipo de red. Es por esto que primeramente se define la estructura del grafo, el cual contiene información acerca de las variables utilizadas y la relación que existe entre las mismas. Enseguida se presenta la definición de las tablas de probabilidad condicional, las cuales completan la red bayesiana que lleva a cabo la clasificación de los leucocitos.

4.1. Estructura de la red bayesiana

4.1.1. Características celulares a observar desde el punto de vista del experto

La observación directa de leucocitos a través de un microscopio es, hasta estos días, la técnica más utilizada para lograr el reconocimiento y clasificación de dichas células. Esta técnica no sólo supone un nivel de conocimiento en el área de hematología por parte de quien la aplica, sino también entrenamiento y experiencia para lograr resultados de clasificación confiables. El entrenamiento y experiencia son absolutamente necesarios para que el especialista realice la clasificación, ya que ésta se fundamenta en su habilidad

para distinguir características celulares específicas.

En [HF97] se describe a los *segmentados* –leucocitos neutrófilos– como sigue:

> ... agrupa varios leucocitos maduros cuyos núcleos muestran segmentaciones con finas uniones entre ellas. Generalmente se pueden observar tres a cuatro segmentos ... [en el núcleo se observa] mayor condensación de la cromatina (patrón en leopardo). Citoplasma gris-marrón o rosa-marrón ... granulación fina pardo-violeta (= neutrófila). Mancha clara evidente en la zona de una hendidura nuclear más marcada (centrósfera).

La observación de las características celulares mencionadas depende de la apreciación particular del experto. Así, cada especialista puede dar mayor valor a alguna de ellas o agregar algunas otras al momento de llevar a cabo la clasificación. De acuerdo con el especialista consultado*, las características celulares a observar para llevar a cabo la clasificación son, en general, las siguientes:

- Color.
- Forma.
- Tamaño.
- Granulación.
- Textura.
- Presencia de vacuolas –observables sólo en citoplasma–.
- Presencia de nucléolos –observables sólo en núcleo–.

Cada una de estas características se observa en el núcleo y citoplasma celular.

4.1.2. Reconocimiento de células mediante técnicas de procesamiento digital de imágenes

Para que el sistema de cómputo realice el reconocimiento de los objetos presentes en una imagen digital, es necesario primeramente segmentarla en

*T.L.C. Andrés Gamboa Espinosa, quien labora actualmente en el área de hematología del Instituto Nacional de Cancerología, México, D.F.

grupos de píxeles que guarden alguna relación con los objetos que se quieren identificar. En el caso de los leucocitos, la imagen fuente –fotografía de la célula– se ha segmentado[†] en tres grupos principales: núcleo, citoplasma y fondo de la imagen.

Considerando que el especialista observa en forma general las características mencionadas del núcleo y citoplasma, cada uno de estos últimos como una unidad en sí misma, para después determinar el tipo de célula, se considera que la segmentación de la imagen celular en estas dos unidades estructurales aporta la información mínima necesaria para llevar a cabo la clasificación.

Una vez que se cuenta con la imagen segmentada se procede a extraer las características de la imagen que puedan aportar información que se relacione con las observadas del especialista.

Las mediciones definidas en la Sección 2.3, y que se han llevado a cabo sobre las regiones de la imagen correspondientes al núcleo y citoplasma son las siguientes:

1. Área.
2. Perímetro.
3. Compactibilidad.
4. Dispersión.
5. Momentos centrales normalizados (momentos de Hu).
6. Energía.
7. Entropía.
8. Inercia.

Enseguida se presentan las características que observa el especialista y las mediciones que capturan esa información.

Color: La información acerca del color y sus variaciones presentes en el núcleo y citoplasma celulares se adquiere mediante los momentos centrales normalizados, la energía, la entropía y la inercia. Estas mediciones son llevadas a cabo en el espacio de color RGB. Cada medición se lleva a cabo tres veces para cada objeto de interés, una para cada componente de color.

[†]Véase Sección 4.3.2 para información relativa al proceso de segmentación.

Forma: La información acerca de la forma se captura a través de la medición de la compactibilidad, la dispersión, el área y el perímetro del objeto bajo estudio. Los momentos centrales normalizados, al ser calculados sobre áreas de objetos definidos, también aportan información importante relacionada con la forma, esto debido a que los *momentos* son considerados como una descripción general de la forma del objeto, además de que tienen la propiedad de unicidad y de que el conjunto de estos descriptores define de forma única a una función –a una región–.

Inicialmente, la información de la forma se pretendió adquirir mediante los *descriptores de fourier* de la región, sin embargo, la calidad de las imágenes utilizadas y los resultados del proceso de segmentación, no permitieron obtener mediciones de estos descriptores que pudieran ser útiles en el proceso de diferenciación de los distintos tipos de células.

Tamaño: A excepción de la compactibilidad y la dispersión, las técnicas de extracción de características utilizadas en este trabajo no son invariantes a la escala, así, como cada una de las mediciones fué realizada sobre fotografías celulares tomadas al mismo aumento –100X–, la información relacionada con el tamaño de las células se ha adquirido de forma implícita en todas las mediciones, excepto las ya mencionadas.

Granulación: Esta característica, que es muy importante para la diferenciación de los leucocitos, es tomada en cuenta como parte de la textura del núcleo y citoplasma. Esto se debe a que la medición de la granulación como un objeto independiente dentro de la imagen requeriría un proceso de segmentación más avanzado que el que se ha utilizado, así como fotografías de mejor calidad e imágenes de mayor resolución que aquellas con que se ha contado en este trabajo.

Textura: La energía, inercia y entropía aportan en conjunto, la información de textura de las regiones estudiadas. Los momentos centrales normalizados aportan también información acerca de la textura, aunque en menor proporción.

Presencia de vacuolas y nucléolos: La obtención de información de estas características se encuentra en el mismo caso que la *granulación.*

4.1.3. Definición de la estructura de la red bayesiana

Con la finalidad de hacer el clasificador fácilmente escalable, se ha planteado como esquema general el dividir su estructura en cinco redes de creen-

cia, cada una encargada de calcular la probabilidad de que dadas las características obtenidas de la imagen, ésta pertenezca a un tipo de célula específico, el cual está asociado directamente con cada una de las redes. De esta manera, existen de forma global cinco redes de creencia: red de neutrófilos, red de linfocitos, red de monocitos, red de eosinófilos y red de basófilos. Al final se comparan los resultados obtenidos por cada una de las redes y se elige el que presenta la probabilidad más alta –siguiendo la técnica "winner takes all"–, esto es, se considera que la probabilidad más alta arrojada por las redes ha de estar asociada con el tipo de célula que aparece en la imagen bajo estudio. La Figura 4.1 muestra este esquema general.

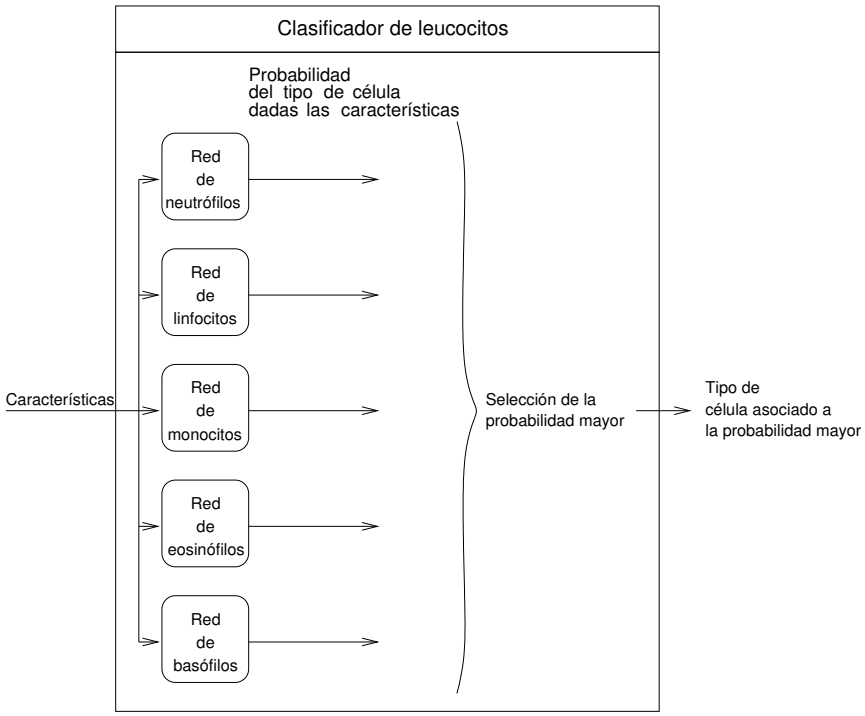


Figura 4.1: Esquema general del clasificador.

Enseguida, se han dividido cada una de las redes mencionadas en dos sub-redes. La primera –subred de nivel 0–, se encarga de determinar la probabilidad de que, dados los valores de las características, la imagen pertenezca a un tipo de célula específico. La segunda –subred de nivel 1–, se encarga de calcular la probabilidad de que, dados los resultados de las primeras subredes –de nivel 0–, la imagen pertenezca al mismo tipo de célula, y de reducir

el error de la subred de nivel 0. La Figura 4.2 muestra la estructura particular de la red de neutrófilos. Recordemos que existen cinco redes similares a ésta, cada una especializada en un tipo de célula en particular.

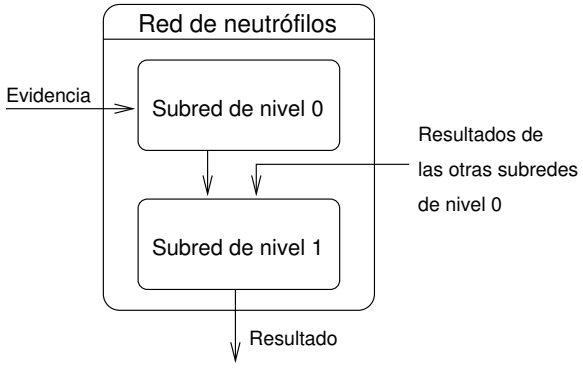


Figura 4.2: Esquema de la red de clasificación encargada específicamente de las probabilidades de los neutrófilos.

Al definir la estructura de las subredes de nivel 0 se ha considerado que, siendo el objetivo central de éstas el asignar una probabilidad a un tipo de célula específico, debe existir en primer término un nodo *tipoDeCelula*[‡], al cual se ha de asignar dicha probabilidad. Teniendo como base este nodo y siguiendo el procedimiento de construcción de la red presentado en el Capítulo 3 notemos que si contamos con suficiente información acerca del núcleo y citoplasma, cada uno observado independiente del otro, al analizar conjuntamente esta información podemos determinar el tipo de célula al que pertenecen. De esta forma, definimos dos nuevos nodos: *núcleo* y *citoplasma*, los cuales se relacionan con *tipoDeCelula* como lo muestra la Figura 4.3.

Conociendo los valores de los descriptores de la imagen podemos determinar si el núcleo o el citoplasma celulares de la misma, están asociados a algún tipo de célula en particular. Siguiendo esta observación, agrupamos las características: inercia, energía y entropía en un sólo conjunto de descriptores denominado *Textura*; área, perímetro, compactibilidad y dispersión en un conjunto denominado *Región*; momentos de Hu 1, 2, 3, 4, en un conjunto llamado *Momentos de Hu*, y definimos las siguientes relaciones entre nodos: *Textura* → *núcleo*, *Región* → *núcleo*, *Momentos de Hu* → *núcleo*, *Textu-*

[‡]Este nombre se refiere a que debe existir un nodo llamado ya sea *Neutrófilo, Linfocito, Monocito, Eosinófilo o Basófilo*, dependiendo de la red con que se trabaje, red de neutrófilos, de linfocitos, etc.

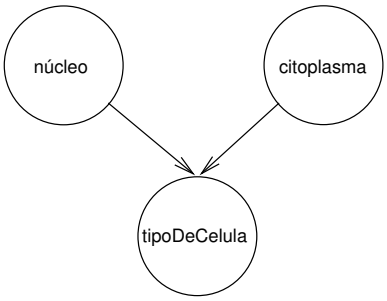


Figura 4.3: Nodos *núcleo* y *citoplasma* agregados a la red más las relaciones que se generan.

ra → *citoplasma*, *Región* → *citoplasma* y *Momentos de Hu* → *citoplasma*. El agrupar las características ha sido necesario para reducir el número de entradas de las tablas de probabilidad de los nodos *núcleo* y *citoplasma*. Como ejemplo, la Figura 4.4 muestra las relaciones y nodos que se definen con el nodo citoplasma, de acuerdo con este criterio. Los mismos nodos de características y relaciones se definen para el nodo núcleo.

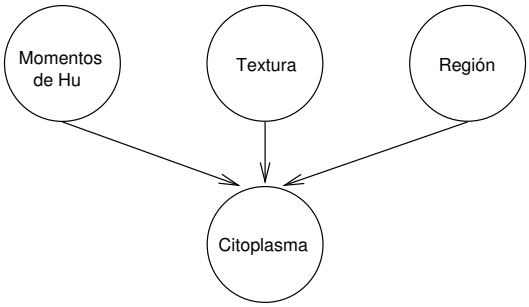


Figura 4.4: Nodos de caraterísticas y forma en que se relacionan con el nodo citoplasma.

Los descriptores, como ya se mencionó, se componen de varias mediciones que se llevan a cabo sobre la imagen. Por ejemplo, las mediciones de la textura la componen la energía, la inercia y la entropía, así que podemos decir que a la textura la determinan tres mediciones distintas, cada una de las cuales puede asociarse con un nuevo nodo de la red, en este caso *energía* → *textura*, *inercia* → *textura* y *entropía* → *textura*. Siguiendo este razonamiento, se definen 22 nuevos nodos en la red, 11 asociados al núcleo

celular y los restantes asociados al citoplasma. En particular, cuatro nodos determinan al descriptor de región, cuatro al descriptor de Hu –momentos centrales normalizados– y tres a la textura. La Figura 4.5 muestra tanto los nuevos nodos como las relaciones que se han agregado a la red.

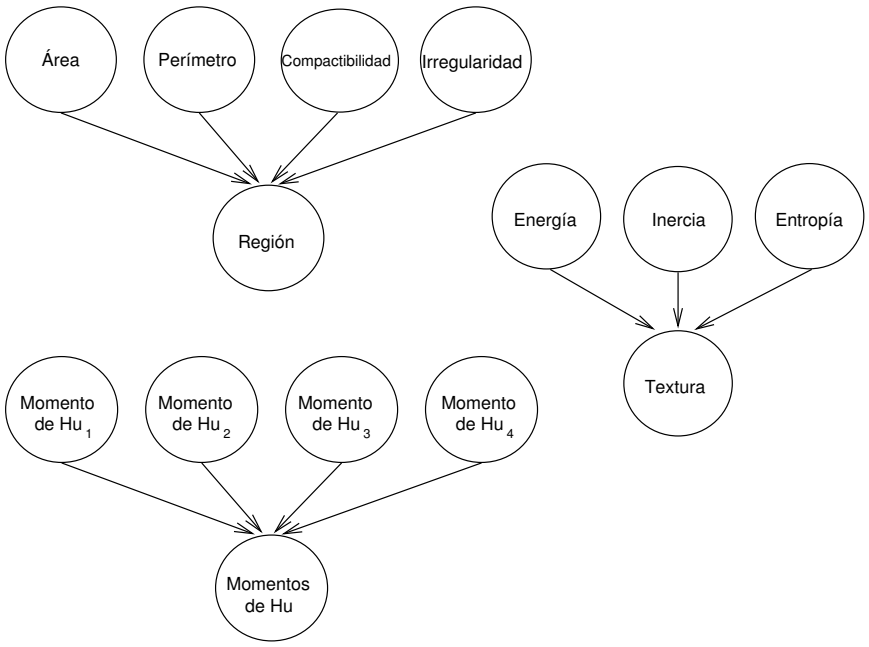


Figura 4.5: Nodos que se corresponden con las características específicas medidas y su forma de relacionarse.

Algunas de las mediciones de las características se han llevado a cabo en el espacio de color RGB, y de esta forma, cada una consta de tres valores. Siguiendo con este proceso de construcción de la red, podemos definir las siguientes relaciones: *valorDeCaracterísticaEnR* $\rightarrow$ *característica*, *valorDeCaracterísticaEnG* $\rightarrow$ *característica* y *valorDeCaracterísticaEnB* $\rightarrow$ *característica*. La Figura 4.6 muestra los nuevos nodos que surgen al adoptar este criterio.

Ya que los valores de las características son extraídos directamente de la imagen de interés, no es necesario agregar más nodos, y de esta forma queda definida la red de nivel 0. La Figura 4.7, muestra la estructura de la red de neutrófilos. Se recuerda que es necesario construir cinco de estas redes, una para cada tipo de célula que se quiere clasificar.

Definamos ahora, de manera general, la forma en que ha de considerarse

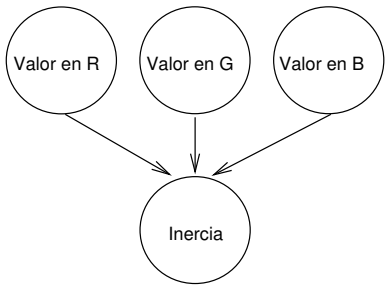


Figura 4.6: Nodos de valor R, G y B para el nodo inercia, ya sea de núcleo o citoplasma. Observar que existen nodos similares para la energía, entropía y los primeros cuatro momentos de Hu ($M_i, i = 1, 2, 3, 4$).

el nivel de error de las redes de nivel 0. Si A es alguno de los tipos de célula a clasificar, entonces, el nivel de error de las redes de nivel 0 es la razón del número de veces en que éstas asignan una probabilidad mayor a algún tipo de célula diferente de A, cuando la imagen que se intenta clasificar pertenece a A, entre el número total de experimentos de clasificación realizados. Por ejemplo, en el caso particular de la red que clasifica a los neutrófilos, el error asociado a tal red se calcularía como la razón del número de veces que las redes de nivel 0 asignan una probabilidad mayor algún tipo de célula diferente de neutrófilo, cuando la imagen que se está clasificando pertenece a un neutrófilo, entre el número de veces total que se intentó clasificar la imágen de un neutrófilo.

Continuamos ahora con la definición de la red de nivel 1. Como la función de esta red se centra en utilizar y comparar los resultados de las redes anteriores para lograr reducir su error al mismo tiempo que las conjunta –ya que las redes de nivel 0 arrojan resultados independientes–, el modelo debe considerar tales requerimientos, además de permitir la incorporación de la probabilidad a *priori* de cada tipo de célula, dato que hasta este momento no se ha considerado.

Si A, B, C, D y E son los tipos de células a clasificar, podemos definir ahora la probabilidad que calculará la red de nivel 1 como: probabilidad de que la imagen pertenezca a un determinado tipo A dados los resultados de las redes de nivel 0 que clasifican a los tipos B, C, D y E. Con base en esto podemos definir cinco nodos correspondientes a cada tipo de célula y relacionarlos como se muestra en la Figura 4.8.

Si centramos nuestra atención en la red encargada de clasificar el tipo

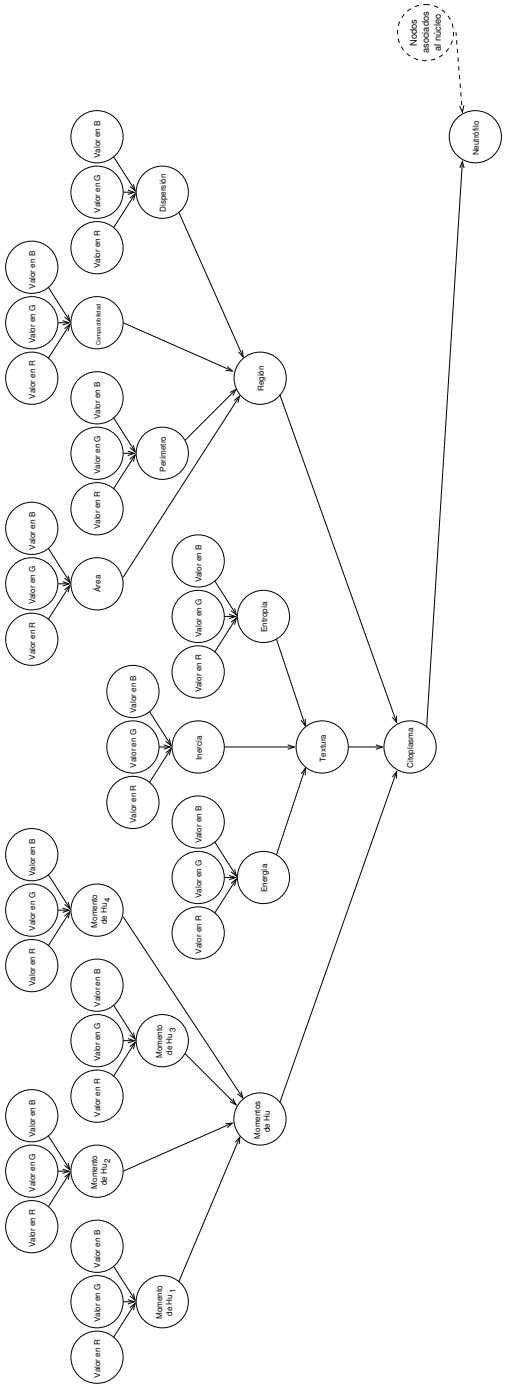


Figura 4.7: Estructura de la red de nivel 0 de los neutrófilos. Los nodos asociados al núcleo presentan una estructura idéntica a los nodos asociados al citoplasma.

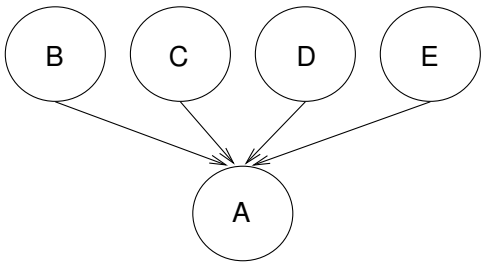


Figura 4.8: Modelo inicial de la red de nivel 1.

de células A. La red de nivel 1 encargada de clasificar a este tipo de células recibirá como evidencia los resultados de las redes de nivel 0 que determinan las probabilidades de los tipos B, C, D y E. Para reducir el error de la red de nivel 0 encargada de clasificar el tipo de células A podemos agregar a la red de nivel 1, nodos "sensores" relacionados con cada uno de los nodos de tipo diferente de A como se muestra en la Figura 4.9. Estos nodos deben tener información acerca del nivel de error presente en las redes de nivel 0 y su función es la de determinar la probabilidad de que el resultado que entregan las redes de nivel 0 es correcto dado el nivel de error de las mismas.

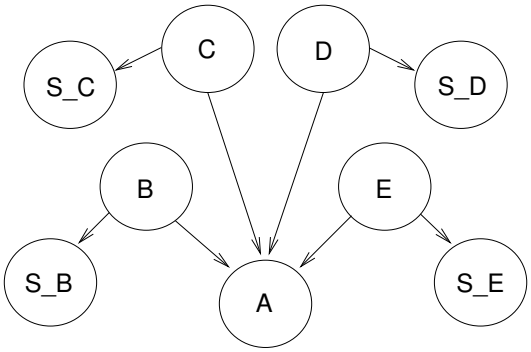


Figura 4.9: Modelo de la red de nivel 1 después de agregar los nodos sensores.

El modelo de la red de nivel 1 considera que los nodos sensores reciben señales booleanas, las cuales indican el resultado de las redes de nivel 0. Las señales que leen los sensores pueden interpretarse como sigue: *La red encargada de clasificar al tipo de célula X indica que la imagen sí pertenece al tipo de célula X*, y tratándose del caso contrario: *La red encargada de*

clasificar al tipo de célula X indica que la imagen no pertenece al tipo de célula X. Como los resultados obtenidos de las redes de nivel 0 son valores probabilísticos, es necesario interpretar esos valores de tal forma que tengan un sentido booleano. Consideremos sólo la probabilidad de que, dadas las características, la imagen *sí* pertenece a un determinado tipo de célula, y supongamos que los resultados de las redes de nivel 0 son los que muestra el Cuadro 4.1.

tipo	*probabilidad*
A	0.50
B	0.74
C	0.13
D	0.39
E	0.81

Cuadro 4.1: Resultados supuestos de las redes de nivel 0

Ordenemos los resultados del Cuadro 4.1 de mayor a menor y de izquierda a derecha como se muestra enseguida:

E	B	A	D	C
0.81	0.74	0.50	0.39	0.13

Si nos ubicamos en la casilla de alguno de los tipos de célula que se muestran, podemos decir que, con respecto de él, los tipos de célula que están a su izquierda *sí* pertenecen al tipo de célula que la red de estos últimos intenta clasificar, ya que la probabilidad que presentan es mayor a la del tipo de célula en el cual nos hemos ubicado. Supongamos por ejemplo que nos ubicamos en la casilla que corresponde al tipo de célula A. Desde esta perspectiva los resultados sugieren que la red de nivel 0 encargada de clasificar al tipo de célula E está entregando el resultado: *la imagen sí corresponde al tipo de célula E*. Sucede lo mismo si observamos el resultado de la red encargada de clasificar al tipo de célula B. Desde la misma perspectiva –ubicándonos en la casilla de A– el resultado que entrega la red de nivel 0 que clasifica el tipo de célula C puede interpretarse como: *la imagen no corresponde al tipo de célula C*. Es en esta forma que podemos interpretar los valores probabilísticos reales entregados por las redes de nivel 0 para entenderlos como valores booleanos. Estos valores son la evidencia que alimenta a la red de nivel 1. La evidencia se introduce a través de los nodos *sensores*. Observemos también que existen nodos sin padres, correspondientes a cada tipo de

célula, es a través de ellos que puede introducirse la información acerca de la probabilidad a priori de las células a clasificar.

La estructura final de la red se compone de cinco redes bayesianas, cada una de ellas se enfoca a un tipo específico de célula y se encarga de determinar la probabilidad de que dadas las características y dados los resultados de clasificación de las redes de nivel 0 que se enfocan a los otros tipos de célula, la imagen bajo estudio pertenece a un tipo específico de célula.

Una vez que las cinco redes bayesianas han calculado la probabilidad a *posteriori* en el nodo *tipoDeCélula* de la red de nivel 1, se comparan los valores probabilísticos calculados en las cinco redes y el mayor es considerado como el resultado final del clasificador.

4.2. Definición de las probabilidades condicionales

Al comenzar a definir los valores de probabilidades presentes en nuestro clasificador se hace necesario definir un conjunto muestra de imágenes que sea representativo de la población general y sobre el cual han de calcularse esos valores, así como un método de muestreo que vaya de acuerdo con las características de nuestra población.

4.2.1. Cálculo del tamaño muestral y muestreo

Para calcular el tamaño muestral se ha empleado la *desigualdad de Chevyshev* [MR03]. El proceso se explica a continuación.

- Se define un conjunto muestra *piloto*, el cual ayudará a estimar los parámetros necesarios para el cálculo final del tamaño muestral. Ya que ésta es una estimación general, se considera que todos los tipos de célula deben estar igualmente representados. En el caso particular de este trabajo, el conjunto de imágenes más pequeño, que representa a un tipo particular de célula, cuenta con 8 elementos, y por esta razón, se ha optado por tomar cinco imágenes de cada tipo, para así contar con al menos 3 imágenes del conjunto más pequeño para la etapa de pruebas. De acuerdo con este criterio, nuestra muestra piloto se compone de 25 imágenes.

- Se extraen las características deseadas de todas las imágenes de la muestra piloto –cada característica define un conjunto de 25 datos–.

- En cada uno de los conjuntos de datos obtenidos se calculan la media ($\bar{x}$) y varianza (S^2) muestrales.

- A partir de los datos anteriores se calcula el *coeficiente de variación* de cada S^2. Este paso es necesario para poder comparar directamente los valores de S^2 como se muestra en los dos puntos siguientes.

- El cálculo del tamaño muestral a partir de la desigualdad de Chevyshev exige conocer sólo un valor de σ^2 –varianza poblacional–, así que comparamos los valores S^2 y elegimos el mayor de ellos para realizar el cálculo. El objetivo de haber elegido el mayor valor de S^2 es el de asegurar que el tamaño muestral será válido para todos los conjuntos de datos con que trabajamos. El valor máximo S^2 obtenido fué:

  $$S^2 = 166^2$$

  y corresponde al grupo de datos definido por la característica *momento central normalizado –momento de Hu– calculado sobre el canal rojo –R– del núcleo de las células.*

- Definimos ahora el nivel de error α que estamos dispuestos a aceptar, así como su nivel de confiabilidad. Este error que estamos dispuestos a aceptar se expresa, en este trabajo, como un porcentaje de μ. Al expresar este porcentaje como una probabilidad podemos interpretarlo como: *la probabilidad de que el promedio de los momentos de Hu del canal R del núcleo se encuentre dentro de un X % de μ es igual a Y.* Un valor α deseable sería de 0.05, con una confiabilidad de 95 %. Con estos datos la expresión anterior quedaría como: *la probabilidad de que el promedio de los momentos de Hu del canal rojo del núcleo se encuentre dentro de un* 5 % *de μ es igual a* 0.95. Sin embargo, debido al valor tan alto que presenta S^2 y al conjunto reducido de imágenes con que se cuenta, se hace necesario ampliar el margen de error que estamos dispuestos a aceptar. La siguiente expresión resume los datos utilizados para el cálculo del tamaño muestral:

  > Calcular el número de mediciones a realizar para que sea al menos de 0.75 la probabilidad de que el promedio de las mediciones de los momentos de Hu del canal rojo del núcleo se encuentre dentro de un 32 % de μ.

- El cálculo final del tamaño de muestra aplicando la desigualdad de Chevyshev queda como:

  $$P(|\bar{x} - \mu| \geq c \cdot \sqrt{\frac{166^2}{n}}) \leq \frac{1}{c^2} \quad \text{para} \quad c > 0$$

Si $c = \sqrt{4}$,

$$P(|\bar{x} - \mu| \geq \sqrt{4} \cdot \sqrt{\frac{166^2}{n}}) \leq \frac{1}{4}$$

$$P(|\bar{x} - \mu| < \sqrt{4} \cdot \sqrt{\frac{166^2}{n}}) \leq \frac{3}{4}$$

De esta forma, la probabilidad de que $\bar{x}$ se encuentre dentro $\sqrt{4} \cdot \sqrt{\frac{166^2}{n}}$ de μ es al menos de $3/4 = 0.75$.

Introducimos ahora 32 % de error que estamos dispuestos a aceptar,

$$\sqrt{4} \cdot \sqrt{\frac{166^2}{n}} = 32$$

Y el resultado final al despejar n queda como:

$$n = \frac{4 \cdot 166^2}{32^2} \approx 107.64 \sim 108$$

De esta forma concluímos que es suficiente contar con, por lo menos, 108 imágenes para llevar a cabo un estudio de clasificación que en el peor de los casos tendrá una probabilidad de 0.75 de que el promedio de las mediciones de las características extraídas se encontrará dentro de un 32 % de μ.

Ahora, para definir los conjuntos muestra y entrenamiento se ha utilizado un muestreo por estratos, esto debido a que los diferentes tipos de células no se reparten homogéneamente en la población total.

La proporción normal de los leucocitos –que nos interesa clasificar– en un adulto normal según [RA01] se muestra en el Cuadro 4.2.

Célula	*Porcentaje (%)*
Neutrófilos segmentados (N)	40 - 74
Eosinófilos (E)	0 - 7
Basófilos (B)	0 - 3
Monocitos (M)	1 - 13
Linfocitos (L)	12 - 46

Cuadro 4.2: Porcentajes normales de leucocitos [RA01].

Sin embargo, estas cifras no son absolutas y en diferentes fuentes pueden encontrarse distintos valores, así, en [HF97] los porcentajes normales de leucocitos son los que se muestran en el Cuadro 4.3.

Célula	*Porcentaje (%)*
N	50 - 70
E	0 - 4
B	0 - 1
M	2 - 8
L	25 - 45

Cuadro 4.3: Porcentajes normales de leucocitos [HF97].

Se ha optado por tomar el promedio de los valores anteriores para calcular las proporciones de los diferentes tipos de leucocitos, y con estos llevar a cabo el muestreo por estratos. Siendo *108 el número mínimo* de imágenes que deben componer la muestra que ha de utilizarse para la etapa de entrenamiento, los valores calculados son los siguientes:

Célula	*Porcentaje promedio(%)*	*No. de imgs.*
N	58	$62.64 \sim 63$
E	03	$3.24 \sim 3$
B	01	$1.08 \sim 1$
M	06	$6.48 \sim 6$
L	32	$34.56 \sim 35$

Se cuenta con un total de 230 imágenes. Considerando los conjuntos de entrenamiento y pruebas como el 80 % y 20 %, respectivamente, del total de imágenes de cada tipo de célula, se tiene[§]:

Célula	*No. imgs. con que se cuenta*	*No. imgs. de entrenamiento*	*No. imgs. de pruebas*
N	80	64	16
E	13	9	4
B	8	6	2
M	49	36	13
L	86	64	22

De acuerdo a estos datos, el conjunto de entrenamiento, con un total de 179 imágenes, cumple con los requisitos necesarios para hacer válidos los resultados del clasificador de acuerdo con los parámetros de error calculados con anterioridad. Cabe mencionar que la mayor parte de los subjuntos de

[§]Para la clasificación manual de las fotografías celulares se contó con la ayuda experta de la Q.B.P. Rita Velázquez Juárez.

imágenes que componen al conjunto entrenamiento, tienen un tamaño superior al mínimo calculado. Este hecho supone que el error obtenido en la prática sea significativamente menor que el aquí calculado.

Aún cuando el planteamiento anterior es correcto, probabilísticamente los conjuntos de entrenamiento y pruebas correspondientes a los *eosinófilos* y *basófilos* son extremadamente reducidos para considerarlos como una muestra aceptable para generalizar sus resultados. Sin embargo, tomando en cuenta que la población de estos dos tipos de leucocitos es igualmente reducida, el error derivado de su incorrecta clasificación es mínimo. De mayor importancia puede considerarse el error derivado de la influencia que ejercen los datos –imprecisos– de estos dos tipos de células, sobre los resultados obtenidos al clasificar los tipos de células restantes. En este punto debemos observar que en este trabajo se pretente experimentar la eficacia que ofrecen las redes de creencia para lograr la clasificación de este tipo de imágenes celulares. Desde esta perspectiva puede justificarse la utilización de algunos datos imprecisos sobre un sistema correctamente diseñado.

4.2.2. Definición del tipo de cada nodo (discreto - continuo)

En la estructura de la red puede observarse que los nodos raíz de las redes de nivel 0 se corresponden con las características que se han extraído de las imágenes. El valor numérico de cada una de las características extraídas es, por su naturaleza, continuo. Aún cuando la información que un ordenador procesa es discreta, los valores continuos de la información pueden ser aproximados. Así, trataremos como continuos los valores numéricos reales de las características extraídas y serán aproximados en el ordenador mediante números de punto flotante.

El primer criterio tomado en cuenta para decidir si el valor de un nodo debe ser tratado como discreto o como continuo es la naturaleza de la información que aporta. Como ya se ha mencionado, los nodos que se corresponden con las características medidas en las imágenes son continuos. Observando que estos últimos nodos mencionados son también nodos evidencia, se refuerza la necesidad de tratarlos como nodos continuos con la finalidad de minimizar la pérdida de información entre los procesos de extracción de características y de clasificación.

Considerando ahora que la red de nivel 0 se encarga de determinar la probabilidad de que dados los valores de las características extraídas la imagen pertenece a un determinado tipo de leucocito, puede decirse que cada nodo debe aportar información referente sólo a si la evidencia define o no al tipo de leucocito que se está clasificando, y con esto se obtienen solamente

dos posibles valores para los nodos que han de ser tratados como discretos. Así, en la red bayesiana final, a excepción de los nodos raíz de la red de nivel 0, todos los nodos son tratados como discretos dicotómicos.

4.2.3. Definición de las probabilidades condicionales para los nodos continuos

Ya que, de acuerdo con la estructura de la red, todos los nodos que han de ser tratados como continuos son nodos raíz, cada uno de ellos cuenta con una distribución de probabilidad que se corresponde con una *f.d.p.*

El proceso mediante el cual se determina la f.d.p. correspondiente a cada nodo es el siguiente:

- Se cuenta con un *conjunto entrenamiento* de imágenes sobre el que se llevan a cabo las mediciones de las características deseadas.

  Al haber llevado a cabo un muestreo por estratos, se cuenta con un conjunto entrenamiento dividido en cinco subconjuntos, uno por cada tipo de célula. Como se ha mencionado al inicio de esta sección, el subconjunto entrenamiento de neutrófilos consta de 64 imágenes, el de eosinófilos de 9, el de basófilos de 6, el de monocitos de 36 y el de linfocitos de 64.

- Se obtienen los conjuntos de datos correspondientes a cada una de las mediciones aplicadas a las imágenes de entrenamiento.

  En este caso, por cada característica medida, se obtienen 5 conjuntos de valores, cada uno de tantos elementos como imágenes de entrenamiento existen para el tipo de célula en particular.

  En adelante, se tomará como ejemplo la determinación de la f.d.p. asociada al nodo que representa a la característica: *inercia* en el canal de color rojo del citoplasma de la red de neutrófilos.

  Al llevar a cabo la medición de esta característica sobre el conjunto entrenamiento se obtiene un conjunto de 64 valores numéricos.

- Se aplica el contraste de Kolmogorov-Smirnov[‖] a cada conjunto de datos para determinar el ajuste de los mismos con alguna función de distribución de probabilidad continua específica sugerida.

  Para sugerir una distribución de probabilidad que pueda ajustar con los datos, es necesario hacer un histograma de frecuencias de estos

[‖] En [DeG88] puede encontrarse una explicación detallada de este contraste.

últimos y buscar visualmente similitudes entre el histograma y alguna distribución de probabilidad conocida.

En el presente trabajo, se utilizaron cuatro distribuciones de probabilidad para llevar a cabo el contraste: normal, log-normal, gamma y exponencial[¶]. Todos los conjuntos de datos ajustaron aceptablemente con al menos una de ellas considerando un error $\alpha = 0.05$ en el ajuste del contrastre.

La Figura 4.10 muestra el histograma del conjunto de valores correspondientes a la medición de la inercia en el canal rojo del citoplasma de los neutrófilos, así como la gráfica de densidad de la función normal, log-normal y gamma[**]. Los parámetros de las funciones fueron calculados directamente del conjunto de datos antes mencionado.

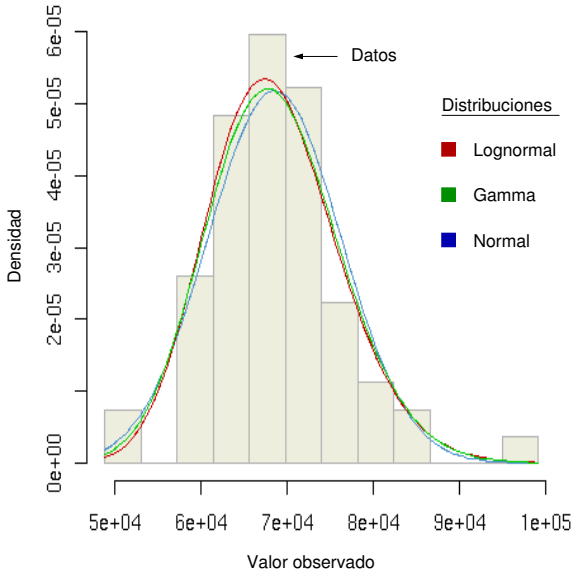


Figura 4.10: Comparación del histograma de los valores de la *inercia* en el canal R del citoplasma de neutrófilos con las gráficas de densidad de las funciones normal, log-normal y gamma.

- La distribución de probabilidad que mejor ajuste a cada uno de los

[¶]El Apéndice A presenta las definiciones de estas distribuciones de probabilidad.

[**]La gráfica de densidad de la distribución exponencial no se incluye debido a que visualmente podemos descartar su posible ajuste a los datos.

conjuntos de datos se considera como la f.d.p. del nodo correspondiente.

En nuestro ejemplo, de acuerdo con el contraste de Kolmogorov-Smirnov la distribución que mejor se ha ajustado a los datos ha sido la *log-normal* y por tal motivo, ésta se considera como la f.d.p. correspondiente al nodo que representa a la inercia en el canal de color rojo del citoplasma en la red de neutrófilos.

Por último, se hace énfasis en que este proceso es aplicable gracias a que los nodos continuos de la red no tienen padres.

4.2.4. Definición de las probabilidades condicionales de los nodos discretos con padres continuos

De acuerdo con la teoría presentada en el Capítulo 3, es necesario definir funciones que asignen valores de probabilidad a los nodos discretos –hijos de padres continuos– dependiendo del valor que tome o que se considere que ha adquirido el nodo padre.

Observemos la asignación de la probabilidad condicional del nodo discreto *momento central normalizado 1 del citoplásma* –Hu_1– como ejemplo, el cual, de acuerdo con la estructura de red presentada, es hijo de los nodos continuos: momento central normalizado 1 en el canal rojo del citoplasma –Hu_{1R}–, momento central normalizado 1 en el canal verde del citoplasma –Hu_{1G}– y momento central normalizado 1 en el canal azul del citoplasma –Hu_{1B}–. Todos los nodos, hemos de acordar que pertenezcan a la red encargada de calcular las probabilidaddes de los neutrófilos.

Considerando por el momento que solamente el nodo Hu_{1R} fuese padre del nodo Hu_1, las entradas de la tabla de probabilidad de éste último han de ser funciones que dependan del valor que se considere ha adquirido el nodo padre.

Observando la f.d.p. asignada al nodo Hu_{1R} y suponiendo que ésta representa una generalización de un experimento llevado a cabo un número infinito de veces, podemos sugerir que el punto donde tal función alcanza su máximo, representa el valor exacto que se obtendría al medir el momento central nomalizado del canal rojo del citoplasma de un neutrófilo ideal perfecto. De acuerdo con esto, entre más se acerquen los valores medidos de las imágenes a este punto en el cual se sitúa el máximo de la f.d.p., mayor será la probabilidad de que la imagen, y en particular esta característica medida, sea la de un neutrófilo. De esta forma, de acuerdo al tipo de función asociada al padre del nodo Hu_1 y tomando en cuenta que este último tiene

asociada una variable aleatoria dicotómica –vHu1, cuando el valor adquirido es cercano al máximo y fHu1 en caso contrario–, las funciones que definen las entradas de la tabla de probabilidad condicional de tal nodo pueden ser definidas como:

- Cuando los datos siguen una distribución normal:
  Si $x < \mu$,

$$\begin{aligned} vHu_1 &= 2\int_{-\infty}^{x} N(\mu,\sigma)dx \\ fHu_1 &= 1 - 2\int_{-\infty}^{x} N(\mu,\sigma)dx \end{aligned}$$

  Si $x > \mu$,

$$\begin{aligned} vHu_1 &= 2\int_{-\infty}^{-x} N(\mu,\sigma)dx \\ fHu_1 &= 1 - 2\int_{-\infty}^{-x} N(\mu,\sigma)dx \end{aligned}$$

  En otro caso,

$$\begin{aligned} vHu_1 &= 1 \\ fHu_1 &= 0 \end{aligned}$$

- Cuando los datos siguen una distribución log-normal y convenimos que la función $f(x,\mu_l,\sigma_l)$ es igual a la f.d.p. log-normal:
  Si $0 < x < e^{\mu_l-\sigma_l^2}$,[††]

$$\begin{aligned} vHu_1 &= \frac{\int_0^x f(x,\mu_l,\sigma_l)\,dx}{\int_0^x f(e^{\mu_l-\sigma_l^2},\mu_l,\sigma_l)\,dx} \\ fHu_1 &= 1 - \frac{\int_0^x f(x,\mu_l,\sigma_l)\,dx}{\int_0^x f(e^{\mu_l-\sigma_l^2},\mu_l,\sigma_l)\,dx} \end{aligned} \tag{4.1}$$

  Si $x > e^{\mu_l-\sigma_l^2}$,

$$\begin{aligned} vHu_1 &= \frac{\int_x^\infty f(x,\mu_l,\sigma_l)\,dx}{\int_x^\infty f(e^{\mu_l-\sigma_l^2},\mu_l,\sigma_l)\,dx} \\ fHu_1 &= 1 - \frac{\int_x^\infty f(x,\mu_l,\sigma_l)\,dx}{\int_x^\infty f(e^{\mu_l-\sigma_l^2},\mu_l,\sigma_l)\,dx} \end{aligned}$$

[††]Máximo de la función log-normal.

Si $x = e^{\mu_l - \sigma_l^2}$,

$$vHu_1 = 1$$
$$fHu_1 = 0$$

En otro caso,

$$vHu_1 = 0$$
$$fHu_1 = 1$$

- Cuando los datos siguen una distribución gamma y convenimos que la función $f(x, \lambda, r)$ es igual a la f.d.p. gamma con $r \geq 1$ y $\lambda > 0$:

  Si $0 < x < \frac{r-1}{\lambda}$,[‡‡]

$$\begin{aligned} vHu_1 &= \frac{\int_0^x f(x, \lambda, r)\, dx}{\int_0^x f(\frac{r-1}{\lambda}, \lambda, r)\, dx} \\ fHu_1 &= 1 - \frac{\int_0^x f(x, \lambda, r)\, dx}{\int_0^x f(\frac{r-1}{\lambda}, \lambda, r)\, dx} \end{aligned}$$

  Si $x > \frac{r-1}{\lambda}$,

$$\begin{aligned} vHu_1 &= \frac{\int_x^\infty f(x, \lambda, r)\, dx}{\int_x^\infty f(\frac{r-1}{\lambda}, \lambda, r)\, dx} \\ fHu_1 &= 1 - \frac{\int_x^\infty f(x, \lambda, r)\, dx}{\int_x^\infty f(\frac{r-1}{\lambda}, \lambda, r)\, dx} \end{aligned}$$

  Si $x = \frac{r-1}{\lambda}$,

$$vHu_1 = 1$$
$$fHu_1 = 0$$

  En otro caso,

$$vHu_1 = 0$$
$$fHu_1 = 1$$

[‡‡] Máximo de la f.d.p. gamma.

- Cuando los datos siguen una distribución gamma y convenimos que la función $f(x, \lambda, r)$ es igual a la f.d.p. gamma con $0 < r < 1$ y $\lambda > 0$,

  Si $x > 0$,

$$\begin{aligned} vHu_1 &= \int_x^\infty f(x, \lambda, r)\, dx \\ fHu_1 &= 1 - \int_x^\infty f(x, \lambda, r)\, dx \end{aligned}$$

  En otro caso,

$$\begin{aligned} vHu_1 &= 0 \\ fHu_1 &= 1 \end{aligned}$$

- Cuando los datos siguen una distribución exponencial y convenimos que $f(x, \lambda)$ es igual a la f.d.p. exponencial,

  Si $x > 0$,

$$\begin{aligned} vHu_1 &= \int_x^\infty f(x, \lambda)\, dx \\ fHu_1 &= 1 - \int_x^\infty f(x, \lambda)\, dx \end{aligned}$$

  En otro caso,

$$\begin{aligned} vHu_1 &= 0 \\ fHu_1 &= 1 \end{aligned}$$

La división que aparece en la ecuación 4.1 cuyo numerador es el resultado que arroja la integral de la función $f(x, \mu_l, \sigma_l)$ en el punto x, entre el resultado arrojado por la integral de la misma función en el punto en que alcanza su máximo, se lleva a cabo con la finalidad de ajustar el resultado total a 1 y así cumplir con A.1. Lo mismo sucede con las otras ecuaciones que presentan configuraciones de numerador y denominador similares a los mencionados.

Es así como quedan definidas las entradas de la tabla de probabilidad de Hu_1 en caso de que solamente Hu_{1R} sea su padre.

En nuestro caso, Hu_1 tiene tres padres continuos, así que la función que determina cada una de las entradas de su tabla de probabilidad se calcula como una combinación lineal de las funciones que determinan esas entradas calculadas por separado. De esta forma, si $f_1(x_1)$, $f_2(x_2)$ y $f_3(x_3)$ son las

funciones que determinan las entradas de la tabla de probabilidad de Hu_1 cuando se consideran Hu_{1R}, Hu_{1G} y Hu_{1B} por separado, respectivamente, la función que determina las entradas de las tablas de probabilidad cuando se consideran los tres nodos padres, puede definirse como:

$$f(x_1, x_2, x_3) = \frac{1}{w_1 + w_2 + w_3}[w_1 f_1(x_1) + w_2 f_2(x_2) + w_3 f_3(x_3)]$$

donde w_1, w_2 y w_3 son pesos asignados a cada una de las f.d.p. de los nodos continuos. Estos pesos tienen como finalidad otorgar mayor probabilidad a las f.d.p. asociadas con los nodos continuos para los cuales, el ajuste de los datos con su función distribución correspondiente fué mejor, es decir, para los cuales el error obtenido en el contraste de Kolmogorov-Smirnov fué mínimo.

4.2.5. Definición de las probabilidades condicionales de los nodos discretos con padres discretos

A diferencia de los nodos recientemente descritos, éstos se encuentran en la red de nivel 0 y de nivel 1 y en cada una de ellas los valores de sus tablas de probabilidad son asignados o calculados de forma distinta. Para hacer la definición de las tablas de probabilidad de los nodos discretos con padres discretos de la red de nivel 0, se ha seguido el proceso que a continuación se describe:

- Las entradas de las tablas de estos nodos se ajustan a 0.5 –recordar que éstos son dicotómicos–, dejando completamente el proceso de clasificación en la funcionalidad de los nodos continuos y en los nodos que tienen algún padre continuo.
- Se observa el peso w que tiene cada uno de los ancestros de estos nodos, y la probabilidad de su tabla se ajusta aumentándola o disminuyendola en proporción directa al valor de dichos pesos.
- Se construyen gráficas de caja para observar la separación de los conjuntos de datos correspondientes a alguna característica en particular de los distintos tipos de células.
- Las tablas de probabilidad son ajustadas nuevamente dando mayor probabilidad a los nodos que cuentan con algún ancestro cuyo conjunto de datos se separa más claramente en los diagramas de caja cuando es comparado con los conjuntos obtenidos al medir la misma característica sobre los otros tipo de células.

Como ejemplo, veamos a la característica *momento central normalizado 1, en el canal de color verde* del citoplasma. La Figura 4.11 corresponde a las gráficas de caja en las cuales se comparan los conjuntos de valores obtenidos para los distintos tipos de células.

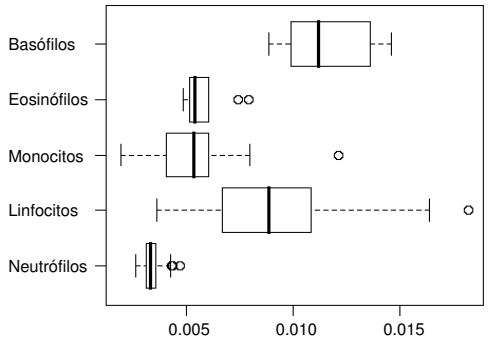


Figura 4.11: Gráficas de caja correspondientes a la medición del momento de Hu 1 del canal verde del citoplasma.

Como puede observase en la Figura 4.11 el conjunto de datos perteneciente a los neutrófilos se separa claramente de los conjuntos de datos de los otros tipos de célula y por tanto representa a una característica, que en la mayor parte de los casos, diferencia acentuadamente a los neutrófilos de las otras células. Por tal motivo sería natural aumentar el peso de esta característica y también la probabilidad de los nodos que tienen por ancestro al que se asocia con la misma. De acuerdo con este ejemplo los cambios han de aplicarse sólo a la red de nivel 0 de los neutrófilos. Cambios similares se aplican a las demás redes simpre que existan características que diferencien claramente a algún tipo de célula particular.

Es así como, para la red de nivel 0, son ajustadas las probabilidades de los nodos que tienen padres discretos.

Continuemos ahora con la definición de las tablas de probabilidad de la red de nivel 1. La Figura 4.12 muestra las redes de nivel 1 correspondientes a los neutrófilos y linfocitos.

Observemos primeramente que las variables aleatorias de estas redes, son todas discretas y dicotómicas de acuerdo con lo expuesto anteriormente. Comencemos ahora la definición de las probabilidades a priori, que corresponden a los nodos raíz. Estas probabilidades se obtienen directamente de los datos estadísticos de la población de cada tipo de leucocito que se ha de clasificar.

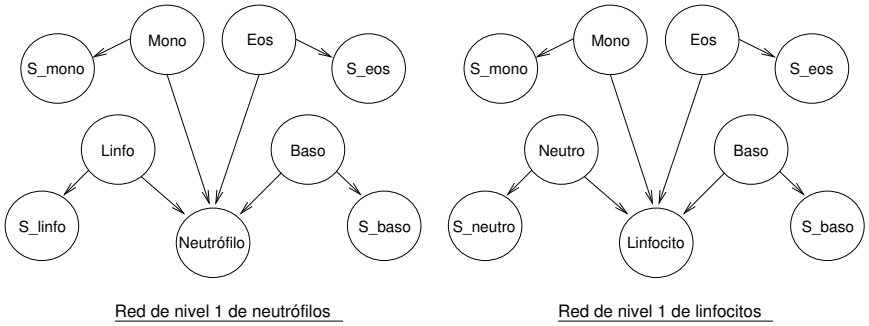


Figura 4.12: Redes de nivel 1 de neutrófilos y linfocitos.

Las tablas de probabilidad de los *nodos raíz* de igual nombre en las distintas redes de nivel 1, tienen exactamente los mismos valores. Las tablas de probabilidad para cada nodo raíz se muestran en el Cuadro 4.4.

Neutro	
$P(\text{Neutro})$	0.58
$P(\overline{\text{Neutro}})$	0.42

Linfo	
$P(\text{Linfo})$	0.32
$P(\overline{\text{Linfo}})$	0.68

Mono	
$P(\text{Mono})$	0.06
$P(\overline{\text{Linfo}})$	0.94

Eos	
$P(\text{Eos})$	0.03
$P(\overline{\text{Eos}})$	0.97

Baso	
$P(\text{Baso})$	0.01
$P(\overline{\text{Baso}})$	0.99

Cuadro 4.4: Tablas de probabilidad de cada nodo raíz presente en las redes de nivel 1.

Definamos ahora las probabilidades condicionales de los nodos "sensores", es decir, de los nodos evidencia de esta red. La función principal de estos nodos es la de recibir los resultados de la red de nivel 0, evaluar la probabilidad de la información recibida tomando en cuenta el nivel de error de la red de nivel 0 y por último comunicar esa información a los demás nodos de la red para llevar a cabo la clasificación. Concentrémonos en la red de nivel 1 que asigna la probabilidad de los neutrófilos, en particular en el nodo sensor encargado de recibir la evidencia acerca de los linfocitos. Las entradas de la tabla de probabilidad de tal nodo pueden verse como:

Sensor-linfocito
$P(\text{S\_linfo} \mid \text{Linfocito})$
$P(\text{S\_linfo} \mid \overline{\text{Linfocito}})$
$P(\overline{\text{S\_linfo}} \mid \text{Linfocito})$
$P(\overline{\text{S\_linfo}} \mid \overline{\text{Linfocito}})$

En este contexto, *S_linfo* se refiere a uno de los dos posibles valores de la variable asociada al nodo del mismo nombre de la red de nivel 1. *Linfocito* se refiere al valor arrojado por la red de nivel 0 de los linfocitos, es decir, es una de las evidencias recogidas por la red de nivel 1.

El valor de $P(\text{S\_linfo} \mid \overline{\text{Linfocito}})$ nos indica la probabilidad de que se afirme que la imagen bajo estudio pertenece a un linfocito cuando no lo es. Desde la perspectiva de la red de neutrófilos esta probabilidad puede entenderse como:

> Probabilidad de que las redes de nivel 0 hayan asignado un valor probabilístico mayor a un linfocito que a un neutrófilo cuando la imagen bajo estudio pertenece a un neutrófilo.

De esta forma, el valor numérico a asignar a esta entrada de la tabla puede calcularse dividiendo el número de veces que ocurre lo antes citado entre el número total de imágenes de entrenamiento, esto es, si contamos con 64 imágenes de neutrófilos en el conjunto de entrenamiento y calculamos su clasificación mediante las redes de nivel 0, dividiremos el número de veces en que se le haya asignado una probabilidad mayor a un linfocito que a un neutrófilo, entre 64 y ese será el valor para $P(\text{S\_linfo} \mid \overline{\text{Linfocito}})$. Calculado el valor de esta probabilidad, el cálculo de $P(\overline{\text{S\_linfo}} \mid \overline{\text{Linfocito}})$ es sencillo, ya que ambas probabilidades se complementan.

Determinemos ahora $P(\overline{\text{S\_linfo}} \mid \text{Linfocito})$. Desde la perspectiva de la red de neutrófilos, esta probabilidad puede interpretarse como:

> Probabilidad de que las redes de nivel 0 hayan asignado una probabilidad mayor a un neutrófilo que a un linfocito cuando la imagen bajo estudio pertenece a un linfocito.

y similarmente al caso anterior, si contamos con 64 imágenes de linfocitos en el conjunto de entrenamiento, dividiremos el número de veces que las redes de nivel 0 hayan otorgado una mayor probabilidad a un neutrófilo que a un linfocito, entre 64 y ése será el valor de tal probabilidad. $P(\text{S\_linfo} \mid \text{Linfocito})$ se calcula como $1 - P(\overline{\text{S\_linfo}} \mid \text{Linfocito})$, al igual que en el anterior caso.

El proceso presentado puede generalizarse para calcular todas las tablas de probabilidad de los nodos sensores de las cinco redes de nivel 1.

Por último, se definirá la tabla de probabilidad del nodo asociado directamente con el tipo de célula para la cual la red de nivel 1 ha de calcular su probabilidad. Enfoquemos nuestra atención en la red de neutrófilos.

Como puede verse en la estructura de la red de nivel 1, las probabilidades condicionales que hay que definir para el nodo *Neutrófilo* son como se muestran en el Cuadro 4.5.

NEUTRÓFILO
P(Neutrófilo | Linfo, Mono, Eos, Baso)
P(Neutrófilo | $\overline{\text{Linfo}}$, Mono, Eos, Baso)
P(Neutrófilo | Linfo, $\overline{\text{Mono}}$, Eos, Baso)
P(Neutrófilo | $\overline{\text{Linfo}}$, $\overline{\text{Mono}}$, Eos, Baso)
$\vdots$
P($\overline{\text{Neutrófilo}}$ | $\overline{\text{Linfo}}$, $\overline{\text{Mono}}$, $\overline{\text{Eos}}$, $\overline{\text{Baso}}$)

Cuadro 4.5: Tabla de probabilidad condicional del nodo *Neutrófilo* de la red de nivel 1 de los neutrófilos.

Para dar un valor numérico a tales probabilidades hay que tomar en cuenta que hasta el momento, y en esta red, sólo hemos utilizado información relacionada con las probabilidades a priori de los distintos tipos de células y el nivel de error de las redes de nivel 0 junto con sus resultados, utilizando estos últimos como valores discretos de dos estados. Sin embargo, la información referente a la probabilidad que cada una de las redes de nivel 0 asigna a cada tipo de célula es de suma importancia, ya que en tales valores va sintetizado todo el proceso de clasificación de las mismas. Después de esta observación fijemos nuestra atención en la probabilidad:

$$\text{P(Neutrófilo(N) | Linfo(L), Mono(M), Eos(E), Baso(B))}$$

la cual puede interpretarse como: *probabilidad de que la célula sea un neutrófilo cuando las redes de nivel cero, desde la perspectiva de los neutrófilos*[§§]*, indican que la imagen se trata tanto de un linfocito, un monocito, un eosinófilo y un basófilo*, es decir, cuando las redes de nivel 0 han asignado el valor más bajo de probabilidad al tipo de célula: neutrófilo.

Para determinar este valor probabilístico comencemos considerando a los cinco tipos de célula que queremos clasificar como los únicos posibles, es

[§§] Ver sección 4.2.1, para información acerca de esta perspectiva.

decir,

$$P(\mathrm{N}) + P(\mathrm{L}) + P(\mathrm{M}) + P(\mathrm{E}) + P(\mathrm{B}) = 1$$

Con esta consideración y tomando en cuenta la información que entregan las redes de nivel 0, podemos calcular nuestra probabilidad de interés como:

$$P(\mathrm{N}|\mathrm{L},\mathrm{M},\mathrm{E},\mathrm{B}) = 1 - \frac{p_L + p_M + p_E + p_B}{p_N + p_L + p_M + p_E + p_B}$$

donde:

p_N es el resultado que entrega la red de nivel 0 de los neutrófilos en el nodo asociado a la variable Neutrófilo en la ocurrencia Neutrófilo = neutrófilo.[|||]
p_L es el resultado que entrega la red de nivel 0 de los linfocitos en el nodo asociado a la variable Linfocito en la ocurrencia Linfocito = linfocito.
p_M es el resultado que entrega la red de nivel 0 de los monocitos en el nodo asociado a la variable Monocito en la ocurrencia Monocito = monocito.
p_E es el resultado que entrega la red de nivel 0 de los eosinófilos en el nodo asociado a la variable Eosinófilo en la ocurrencia Eosinófilo = eosinófilo.
p_B es el resultado que entrega la red de nivel 0 de los basófilos en el nodo asociado a la variable Basófilo en la ocurrencia Basófilo = basófilo.

Del mismo modo, si la probababilidad a calcular fuese:

$$\mathrm{P}(\text{Neutrófilo} \mid \overline{\text{Linfo}}, \overline{\text{Mono}}, \text{Eos}, \text{Baso})$$

el cálculo sería:

$$P(\mathrm{N}|\overline{\mathrm{L}},\overline{\mathrm{M}},\mathrm{E},\mathrm{B}) = 1 - \frac{p_E + p_B}{p_N + p_L + p_M + p_E + p_B}$$

Observar que sólo es necesario determinar de esta forma ocho entradas de la tabla de probabilidad, ya que las otras ocho son sus complementarias.

Es así como cada una de las tablas asociadas a los nodos *Neutrófilo*, *Linfocito*, *Monocito*, *Eosinófilo* y *Basófilo* de las redes de nivel 1, son calculadas tomando en cuenta los valores numéricos reales que han arrojado las redes de nivel 0. De acuerdo con este plantemiento, los valores de cada una de las entradas de estas tablas son *dinámicos*, se calculan cada vez que se lleva a cabo la clasificación de una imagen.

[|||] Observar que esta variable tiene dos posibles valores: neutrófilo y $\overline{\text{neutrófilo}}$.

De esta forma concluimos con el diseño del clasificador, ya que ha quedado definida tanto la estructura de la red, como un procedimiento para determinar las tablas de probabilidad de cada uno de los nodos. El paso siguiente es la implementación de este modelo de clasificador.

4.3. Implementación del clasificador

4.3.1. Plataforma de desarrollo

El lenguaje de programación utilizado para implementar el clasificador ha sido C++. Se eligió este lenguaje debido al equilibrio que ofrece entre velocidad de ejecución –característica sumamente importante para el procesamiento de imágenes– y facilidad de implementación en el modelado orientado a objetos.

El sistema operativo elegido como plataforma ha sido Linux. Se ha elegido este sistema operativo debido a la facilidad que ofrece en la obtención del software necesario para el desarrollo, además de que al utilizar software libre, no es necesario ningún tipo de desembolso económico dedicado a licencias o permisos para la utilización del mismo, y al utilizar paquetes de distribución estándar tampoco es necesario utilizar una distribución específica de este sistema operativo. Así el sistema puede ejecutarse en cualquier distribución de este S.O. e incluso compilarse, después de modificaciones mínimas, con cualquier compilador estándar de C++ para cualquier otro S.O.

Ubuntu 6.01 –Dapper Drake– es la distribución de Linux que se utilizó para el desarollo. La fácil gestión de paquetes que ofrece a través de *aptitude* es la única característica especial que se tomó en cuenta para su elección.

Las imágenes de las células utilizadas en este trabajo utilizan el formato BMP Win32 de 24 bits no comprimidas, y tienen una resolución de conteo de píxeles de 256x256. Éstas fueron capturadas con una tarjeta *Frame Grabber* marca *IMAGINGSOURCE* modelo DFG/LC1, y la resolución de captura fué de 640x480 píxeles. La señal digital de las imágenes se obtuvo mediante la cámara *iCAM* marca *LABOMED* modelo 1500, la cual cuenta con un sensor CMOS a color de 1.3 Mega píxeles y entrega una imagen con resolución de 1200x960 píxeles, y que fué acoplada directamente a un microscopio de transmisión de luz, contando éste último con un objetivo de 100X.

Debe aclararse que originalmente, las imágenes fueron capturadas con una resolución de conteo de píxeles de 640x480, correspondiendo la imagen obtenida a todo el campo visual del microscopio. Para agilizar el procesamiento de las imágenes se optó por recortar sólo el área de interés de cada

imagen, es decir, el área que encierra la imagen completa de cada leucocito. Las imágenes recortadas presentan una resolución de conteo de pixeles de 256x256. El proceso de selección de área y recorte de la misma se realizó de forma manual, sin embargo, existen técnicas que pueden automatizar este proceso [PRG+01].

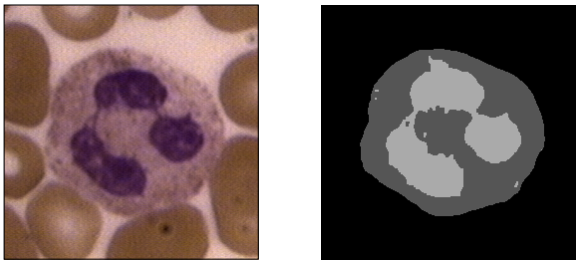
El sistema se desarrolló en un ordenador con procesador PIII que trabaja a una frecuencia de 1GHz y memoria RAM de 256MB. El espacio en disco duro aproximado necesario es de 30 MB. La memoria RAM mínima recomendada es de 128MB.

No se utilizó ningún otro tipo de hardware especializado.

4.3.2. Especificación de parámetros de entrada y salida

La entrada del sistema está compuesta por dos imágenes de 256x256 píxeles, una corresponde a la fotografía de la célula, la otra a la máscara de segmentación de la misma. La máscara de segmentación es necesaria debido a que el clasificador no lleva a cabo el proceso de segmentación, para realizar este proceso se utilizó software externo a este proyecto, el cual implementa el algoritmo descrito en [Kat94, BKYZ96, KZB92] y que corresponde al de segmentación supervisada utilizando *campos aleatorios de Markov*. Tanto el software como los artículos relacionados son de libre acceso [Kat].

La Figura 4.13 muestra la imagen de un neutrófilo junto con su máscara de segmentación.



Neutrófilo segmentado. Máscara de segmentación

Figura 4.13: Fotografía de neutrófilo con su correspondiente máscara de segmentación.

La máscara de segmentación se divide en tres regiones, cada una con valores específicos de nivel de gris, que son:

núcleo	170
citoplasma	85
fondo	0

La salida del sistema la conforman los nombres de los cinco tipos de célula que se están clasificando con un valor de probabilidad asociado a cada uno. El tipo de célula que tenga asociado el valor de probabilidad mayor se considera como al que pertenece la imagen bajo estudio.

4.3.3. Estructura del software

El diagrama de la Figura 4.14 corresponde al modelo de clases utilizado en la implementación del clasificador. Por claridad sólo se muestran los nombres de las clases implementadas y las relaciones que existen entre ellas. Los nombres de las clases se definieron utilizando palabras en inglés buscando con ello hacerlos compactos y significativos.

El software lleva a cabo las siguientes operaciones al ejecutar el proceso de clasificación:

1. Carga de la fotografía celular y de la máscara de segmentación en la memoria.

   La clase *LeukoImgProcessor* –Leuko Image Processor– se encarga de accesar directamente a los archivo de imágenes, así como de llevar a cabo las operaciones de pre-procesamiento.

2. Normalización del histograma de frecuencias de la fotografía celular –mediante la clase LeukoImgProcessor–.

3. Acondicionamiento de las regiones mediante operaciones de morfología matemática –dilataciones y erosiones, mediante LeukoImgProcessor–.

4. Eliminación de regiones inconexas presentes en la máscara de segmentación –eliminación de islas, mediante LeukoImgProcessor–.

5. Extracción de las características de la imagen –medición de los momentos de Hu 1, 2, 3 y 4, medición de los descriptores de región y de los descriptores de textura, mediante la clase LeukoImgProcesor–.

6. Clasificación mediante las redes de creencia de nivel 0.

   El proceso de clasificación está soportado por todas las clases diferentes de LeukoImgProcessor que se muestran en la Figura 4.14.

   Estas clases se organizan como se describe a continuación.

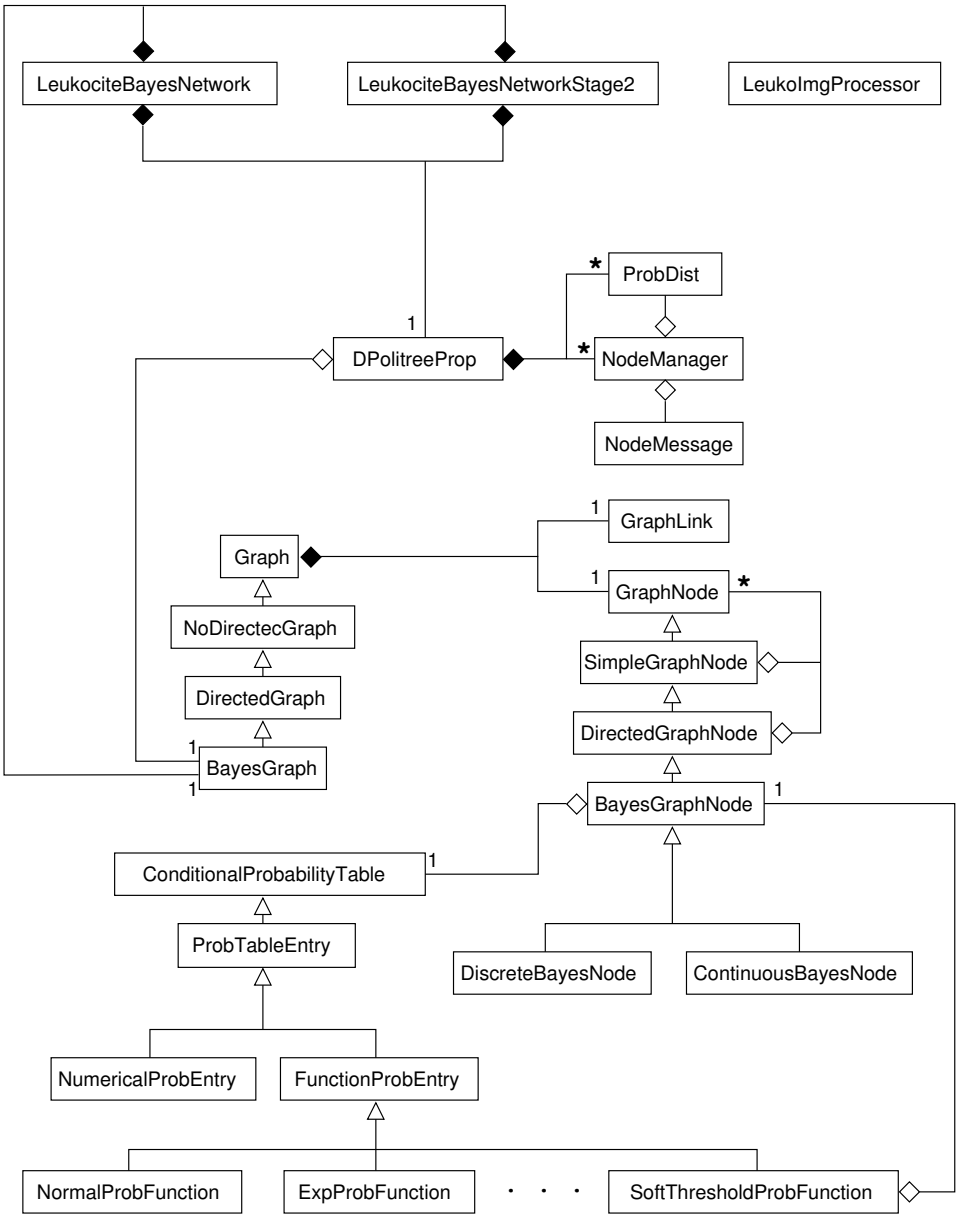


Figura 4.14: Diagrama UML de las clases que implementan el sistema.

- Clases que implementan la funcionalidad general de un grafo: Graph, DirectedGraph, NoDirectedGraph, GraphLink, GraphNode, SimpleGraphNode, DirectedGraphNode.
- Clases que implemetan la estructura propia de una red bayesiana, las cuales se pueden agruparse como:

  Clases que implementan el grafo: BayesGraph, BayesGraphNode, DiscreteBayesNode, ContinuousBayesNode.

  Clases que implementan la tabla de probabilidad asociada a cada nodo: ConditionalProbabilityTable, ProbTableEntry[¶¶], NumericalProbEntry, FunctionProbEntry, NormalProbFunction, ExpProbFunction, SoftThresholdProbFunction[***].
- Clases para la implementación del algoritmo que lleva a cabo la inferencia en poliárboles: DPolitreeProp –Distributed Politree Propagation– , ProbDist –Probability Distribution–, NodeManager, NodeMessage.

  Hay que señalar que el algoritmo implementado corresponde al de inferencia en poliárboles distribuído [DV05].

La clase *LeukociteBayesNetwork* organiza el trabajo de todas las clases ya mencionadas para llevar a cabo la implementación de las redes de nivel 0.

7. Clasificación mediante las redes de creencia de nivel 1.

   La clase *LeukociteBayesNetworkStage2* organiza la implementación de las redes de nivel 1.

[¶¶] En esta denominación *Prob* significa: *Probability*. Sucede lo mismo para los demás nombres de clases.

[***] Existen también clases que implementan la f.d.p. gamma, log-normal y el complemento de la función SoftThreshold.

Capítulo 5

Pruebas y resultados

El presente capítulo tiene como objetivo mostrar el desempeño del clasificador mediante el contraste del nivel del error calculado teóricamente y el error obtenido al llevar a cabo la etapa de pruebas del mismo. Con la finalidad de lograr este objetivo, se describirá el proceso de entrenamiento, señalando en este punto, el error esperado de acuerdo con los datos estadísticos utilizados; enseguida se describirá el proceso de pruebas y se indicarán los resultados obtenidos; como punto final se calculará el error del sistema y se contrastará con el error calculado en la etapa de entrenamiento. Como punto adicional se compararán los resultados obtenidos en este trabajo con los obtenidos con trabajos similares que utilizan otras técnicas de clasificación.

5.1. Revisión del proceso de entrenamiento

El proceso de entrenamiento puede describirse, en general, como la extracción de los valores estadísticos y probabilísticos relacionados con las características de interés presentes en las imágenes que componen, exclusivamente, al conjunto de entrenamiento. Esta información es almacenada en la red bayesiana a través de las tablas de probabilidad de sus nodos. Una vez calculados los valores de las tablas de probabilidad, el proceso de clasificación llevado a cabo por la red, no modifica los datos estáticos de la misma.

De acuerdo al muestreo realizado –muestreo por estratos, ver Sección 4.2.1– el número de imágenes que compone el conjunto de entrenamiento es de 179, de los cuales 64 son de neutrófilos, 64 de linfocitos, 36 de monocitos, 9 de eosinófilos y 6 de basófilos.

La estructura de la red se definió de acuerdo a los procedimientos expuestos en la Sección 4.1.3, y no se utilizó información que haya sido extraída directamenta de las imágenes para tal fin. Como ejemplo de información extraída indirectamente de las imágenes podemos mencionar la calidad de las mismas, factor que influyó para definir el conjunto de características que fueron extraídas –ver Sección 4.1.2–, hecho que se ve reflejado en la estructura de la red.

Para la definición de las tablas de probabilidad de los nodos continuos y nodos discretos con padres continuos presentes en la red, se utilizó información estadística y probabilística extraída directamente de las imágenes que componen el conjunto de entrenamiento, esto de acuerdo con los procedimientos expuestos en las Secciones 4.2.3 y 4.2.4.

Para la definición de las tablas de probabilidad de los nodos discretos se utilizó información estadística y probabilística, extraída indirectamente del conjunto de entrenamiento –a través de la observación y valoración de diagramas de caja, por ejemplo– tal como se describe en la Sección 4.2.5.

El error teórico esperado en la red, de acuerdo con lo señalado en la Sección 4.2.1, es calculado con base en el tamaño del conjunto de entrenamiento utilizado. Los datos precisos se resumen como:

> *Se tiene un 75 % de probabilidad, de que en el peor de los casos, el error de clasificación sea de 32 %.*

En consecuencia, y calculando la razón $\frac{32}{75}$, decimos que el error general esperado es:

$$E_{GE} \approx 42.66\,\%$$

5.2. Etapa de pruebas y cálculo del error

Se llevó a cabo el experimento que comprende la clasificación de todas y cada una de las imágenes que componen el conjunto de entrenamiento y el conjunto de prueba. El experimento se realizó sin repetición, es decir, cada imagen fue sometida al proceso de clasificación una sola vez. Ya que ninguna de las imágenes presentes en este último conjunto fué utilizada en la etapa de entrenamiento, se considera que los resultados obtenidos a partir de él, son una muestra clara e imparcial del desempeño del clasificador.

Los resultados obtenidos al llevar a cabo el experimento sobre el conjunto de entrenamiento se muestran en el Cuadro 5.1.

Los resultados obtenidos al llevar a cabo el experimento sobre el conjunto de prueba se muestran en el Cuadro 5.2.

Célula	*No. imgs. de entrenamiento*	*No. imgs. mal clasificadas*	*% de error por tipo*
Neutrófilo	64	6	≈ 9.4
Linfocito	64	8	12.5
Monocito	36	9	25
Eosinófilo	9	2	≈ 22.22
Basófilo	6	1	≈ 16.6

Cuadro 5.1: Resultados obtenidos a partir del conjunto de entrenamiento.

Célula	*No. imgs. de prueba*	*No. imgs. mal clasificadas*	*% de error por tipo*
Neutrófilo	16	2	12.5
Linfocito	22	6	≈ 27.3
Monocito	13	4	≈ 30.7
Eosinófilo	4	2	50
Basófilo	2	1	50

Cuadro 5.2: Resultados obtenidos a partir del conjunto de prueba.

Notar que el *% de error por tipo* que se presenta en los Cuadros 5.1 y 5.2, corresponde solamente al tamaño relativo que representa el conjunto de imágenes mal clasificadas con respecto del total de imágenes del mismo tipo de células. Para llevar a cabo el cálculo del error general del sistema se debe tomar en cuenta que el muestreo realizado corresponde a un muestreo por estratos y por tal motivo se debe tomar en cuenta la proporción que las imágenes de cada tipo de célula representa dentro de la población total. Este error lo calculamos como:

$$E_E = \frac{I_M}{I_E} \times P_E \times 100 \tag{5.1}$$

donde:

E_E representa porcentaje de error por estrato, es decir, el error debido a la población de un sólo tipo de célula.

I_M es el número de imágenes mal clasificadas de un tipo específico de célula.

I_E es el número de imágenes del conjunto que se evalúa –entrenamiento o prueba– del mismo tipo de célula que I_M.

P_E es el porcentaje que representa, dentro de la población total, el tipo de célula observado en I_M.

Como podemos observar $I_M/I_E \times 100$ corresponde precisamente al valor que en los Cuadros 5.1 y 5.2, se presenta como *% de error por tipo*, así que sólo nos queda multiplicar ese valor por P_E para obtener el error general por tipo –como se describió en la Sección 4.2.1, P_E tiene un valor de 0.58 para los neutrófilos, 0.32 para los linfocitos, 0.06 para los monocitos, 0.03 para los eosinófilos y 0.01 para los basófilos–.

Calculando el error sobre los resultados obtenidos a partir del conjunto de entrenamiento tenemos:

Célula	$100(I_M/I_E) \cdot P_E$	E_E – *Error gral. por tipo (%)*
Neutrófilo	$9.4 \cdot 0.58$	5.45
Linfocito	$12.5 \cdot 0.32$	4
Monocito	$25 \cdot 0.06$	1.5
Eosinófilo	$22.22 \cdot 0.03$	0.66
Basófilo	$16.6 \cdot 0.01$	0.16

Así, el error general del sistema es la sumatoria de los errores por estrato, el cual, para el conjunto de entrenamiento corresponde a:

$$E_{GT} = 11.77\,\% \tag{5.2}$$

Se calcula ahora el error obtenido a partir del conjunto de imágenes de prueba:

Célula	$100(I_M/I_E) \cdot P_E$	E_E – *Error gral. por tipo (%)*
Neutrófilo	$12.5 \cdot 0.58$	7.25
Linfocito	$27.3 \cdot 0.32$	8.7
Monocito	$30.7 \cdot 0.06$	1.84
Eosinófilo	$50 \cdot 0.03$	1.5
Basófilo	$50 \cdot 0.01$	0.5

El error general calculado a partir de los resultados del conjunto de prueba es:

$$E_{GP} = 19.79\,\% \tag{5.3}$$

El Cuadro 5.3 muestra los resultados finales obtenidos al realizar los experimentos de clasificación.

Error esperado	$E_{GE} \approx 42.66\,\%$
Error en entrenamiento	$E_{GT} = 11.71\,\%$
Error en pruebas	$E_{GP} = 19.79\,\%$

Cuadro 5.3: Resultados finales de evaluación del clasificador.

El resultado obtenido a través del conjunto de entrenamiento es, claramente el mejor, sin embargo, por obvias razones, es considerado como parcial favorable al clasificador. El error obtenido al clasificar el conjunto de pruebas se considera, como se mencionó al inicio de esta sección, como una muestra clara e imparcial del desempeño del mismo. Por último, ha de señalarse que los resultados obtenidos están completamente dentro del rango de error esperado, ya que, aunque el error obtenido en pruebas es menor que el esperado, el valor tan alto de este último fué calculado para el peor de los casos, resultando de aquí que el sistema es consistente y predecible de acuerdo al modelo planteado.

Comparemos ahora los resultados obtenidos en este proyecto de tesis, con los obtenidos por dos esfuerzos similares enfocados a la clasificación de leucocitos y que utilizan otros modelos de clasificación.

Los resultados publicados en [SZR04], comunican porcentajes de error que van desde el 10.9 % hasta un 25.91 % utilizando un clasificador *naive Bayes*. Aún cuando los resultados obtenidos en el proyecto aquí desarrollado se encuentran dentro del mismo rango de error, debe señalarse que el tamaño de la población total del imágenes de células con que se contó es de menos de la tercera parte que el utilizado en [SZR04], hecho que tiende a incrementar notablemente el porcentaje de error esperado, ya que este error es calculado, fundamentalmente, con base en el tamaño muestral; en el presente proyecto se contó con una muestra de 236 imágenes de células en total; en el caso de [SZR04], el número de imágenes fué de 938. Es preciso señalar también que en el proyecto presentado en [SZR04], todos los tipos de células están representados por tamaños de muestras que pueden considerarse suficientes para la generalización de sus datos –siendo de 44 imágenes el conjunto de menor tamaño–. En el presente proyecto existen dos tipos de células representados por conjuntos extremadamente pequeños: eosinófilos, con 13 imágenes, y basófilos, con 8 imágenes, hecho que tiende a incrementar el error.

En el caso de [PRG+01], se publica que se clasificaron con éxito, inicialmente, 295 imágenes de 325, utilizando un clasificador de Bayes, con base en funciones de decisión. De acuerdo con estos resultados, el error es de

9 % aproximadamente. Se hace mención que, en un ensayo realizado *fuera de línea* se obtuvo un error menor del 5 %. El conjunto de entrenamiento utilizado se reporta con un tamaño de 419 imágenes, y también se reporta que sólo se clasificaron 4 tipos diferentes de células debido a la falta de imágenes representativas de los basófilos –tipo celular que fué descartado de la clasificación–. En este proyecto no se aclara si el conjunto imágenes de prueba fué distinto del de entrenamiento. Se utilizaron dos características distintas para llevar a cabo la clasificación, hecho que reduce la posibilidad de generalizar el modelo para clasificar más tipos de células.

Como dato general, se hace mención de que ninguno de los anteriores proyectos, con los cuales se ha comparado el trabajo realizado en este documento, presentan algún método para estimar el error esperado, sino que se limitan a experimentar y presentar resultados. Esta característica se considera una ventaja del método aquí presentado.

5.3. Principales razones del error obtenido

- Baja calidad de los frotis de sangre utilizados, lo cual repercute en la falta de detalle –definición de líneas y textura– de la imagen, también en el rango de colores significativos, ya que éste se reduce, dando como resultado que la imagen aparezca muy oscura, muy clara o en rangos intermedios de color, sin utilizar toda la gama posible de colores. Dada la calidad del frotis se introduce, también, ruido debido a manchas de colorante o a zonas faltantes de tinción. Se debe señalar que este factor es muy importante, tanto que en [LRM+97] se comenta: *Se reconoce un buen laboratorio a la calidad de los frotis de sangre que prepara.*

- Baja calidad de la segmentación. La segmentación es un paso muy importante al momento de llevar a cabo la clasificación –la cual se apoya igualmente en la calidad de la imagen–, pero dicho tópico está fuera del alcance del presente proyecto. Es por esto que la segmentación llevada a cabo en este trabajo* fué una segmentación supervisada, la cual, permitió reducir errores derivados de la distribución irregular del color en las imágenes. Los resultados de la segmentación, sin embargo, no fueron los óptimos.

- Número reducido de imágenes para llevar a cabo el aprendizaje. Aún cuando los resultados presentados son válidos para tres tipos de células –neutrófilos, linfocitos y monocitos–, los otros dos tipos involucrados

*Véase Sección 4.3.2 para información relativa al proceso de segmentación.

–eosinófilos y basófilos– están representados por una población extremadamente reducida, hecho que impide generalizar con seguridad sus resultados. Esta falta de solidez práctica –ya que se puede corregirse incrementando el tamaño de los conjuntos de imágenes de estos tipos de células– se traduce en errores directamente relacionados con la clasificación de eosinófilos y basófilos. Aún más, al tomarse en cuenta estos dos tipos de célula para llevar a cabo la clasificación general, existe ruido inducido por los mismos y que afecta el resultado en la clasificación de los otros tres tipos de célula.

Capítulo 6

Conclusiones y perspectivas

6.1. Conclusiones

A lo largo del desarrollo de este proyecto de tesis se ha podido constatar que el procesamiento digital de imágenes es una poderosa herramienta que, de entre muchas tareas que podemos realizar, nos ayuda a filtrar, acentuar y extraer características en alguna imagen. Al analizar estas características podemos lograr algún objetivo específico, como la descripción y representación de una imagen de célula. En este trabajo, el objetivo general planteado ha sido la clasificación de leucocitos.

Se ha comprobado también que, de forma clara y sencilla, las redes de creencia, nos permiten sintetizar una gran cantidad de información, al mismo tiempo que nos ofrecen la posibilidad de estructurarla de un modo conveniente al contexto del problema que tenemos interés en abordar. Nos permiten utilizar, también, modelos flexibles de acceso a información categorizada de acuerdo con alguna situación particular, es decir, mediante la estructura del grafo y la utilización de evidencias, permiten utilizar información contextual para filtrar y ordenar los datos que nos interesan, de acuerdo a un modelo diseñado para obtener información significativa orientada a la resolución de un problema particular, como el planteado en este proyecto de tesis.

Con la aplicación de estas dos herramientas a la clasificación de leucocitos, se ha observado que es posible construir sistemas que se adaptan de forma natural al modelo de clasificación utilizado directamente por el ser humano, al mismo tiempo que se obtienen resultados congruentes con los que el mismo modelo plantea como esperados. En otras palabras, la estructura y funcionamiento del sistema es muy similar al proceso llevado a cabo por el experto humano al desarrollar esta tarea, al mismo tiempo que el

sistema permanece predecible de acuerdo al modelo matemático en el cual se fundamenta. Ésta última es una característica muy importante dentro del campo de la investigación científica.

Se ha observado también que es posible adaptar este tipo de redes de tal forma que permitan obtener resultados satisfactorios incluso cuando los datos con que se cuenta para su aprendizaje estén lejos de ser los ideales. Ha de resaltarse que, si bien el error obtenido en la implementación del modelo propuesto en este trabajo, ha sido alto –19.79 %–, el resultado es consistente con los valores que el mismo modelo plantea como esperados. Con esta observación, es posible afirmar que los resultados son factibles de mejora de acuerdo con los procedimientos y métodos aquí propuestos. Es así como puede fundamentarse que el trabajo con procesamiento digital de imágenes y redes de creencia, es predecible, en la medida en que encuentra su base en la teoría de probabilidad y estadística.

Por último y con base en este trabajo puede verse un futuro prometedor para la resolución del problema inicialmente planteado utilizando el análisis digital de imágenes conjuntamente con las redes bayesianas ya que, aún cuando el objetivo propuesto se ha alcanzado, esto no significa que el problema se haya resuelto en su totalidad, simplemente se han explorado satisfactorimente estos campos de investigación como una alternativa de solución, además, hemos de tomar en cuenta que, si bien ha habido avances en estos dos campos, la investigación dentro del campo de Inteligencia Artificial se encuentra aún en una edad muy temprana.

6.2. Perspectivas

Aún hay mucho trabajo por realizar para que el sistema propuesto en este trabajo de tesis llegue a adquirir una funcionalidad aceptable en un ambiente real de trabajo –para que sea funcional en un laboratorio de análisis clínicos, por ejemplo–. Enseguida se señalan algunos puntos que requieren atención inmediata; atendidos estos, se esperaría una notable mejora en el desempeño del clasificador.

- Es imperativo corregir los problemas señalados en al apartado 5.3.
- Mejorar el pre-procesamiento de la imagen. Puede tomarse en cuenta para tal fin, el color y las formas esperadas de las células –información contextual–, de tal forma que contemos con una imagen bien definida para un proceso de segmentación y de clasificación específicos de células blancas.

- Mejorar el modelo de la red bayesiana. El modelo de red bayesiana presentado en este trabajo corresponde a un poliárbol, el cual representa el modelo más sencillo de este tipo de redes. Es posible mejorar tal modelo tomando en cuenta relaciones de dependencia condicional presentes entre distintos tipos de células. Por ejemplo: se observa en el presente trabajo que cuando la imagen de una célula corresponde a la de un neutrófilo, la probabilidad calculada en la primera fase de clasificación de que la imagen sea la de un linfocito es siempre la más baja. Así, apoyandonos en esta observación estadística, podríamos añadir a la red una condición que implique el aumentar la probabilidad de que una célula sea clasificada como neutrófilo cuando la probabilidad asignada al tipo de célula linfocito es la más baja comparada con la probabilidad asignada a los otros tipos de células. La observación anterior puede aplicar a todas las demás células y a diferentes arreglos de valores de probabilidades asignados.

- Utilizar distribuciones de probabilidad continuas que se ajusten mejor a la distribución que presentan los conjuntos de datos de las características tomadas en cuenta en la clasificación. Lo anterior se traduce como reducir la distancia *Dn* definida en el contraste de Kolmogorov-Smirnov*. En este trabajo se utilizan sólo 4 tipos de distribuciones: normal, gamma, exponencial y lognormal†.

- Ajustar aún más los valores de las tablas de probabilidad presentes en la red de tal forma que se ajusten mejor a los datos de entrenamiento.

- Extraer y utilizar más características de las imágenes que aporten información importante para la clasificación.

- Siguiendo el modelo propuesto, extender su funcionamiento para clasificar células blancas anormales o en estados de maduración.

- Implementar e integrar al clasificador un proceso de segmentación automática, así como un proceso de selección y recorte de regiones de interés, como el mencionado en [PRG+01].

*Ver [DeG88] para una explicación detallada de este contraste.
†Ver Apéndice A.3

Apéndice A

Definiciones de probabilidad

A.1. Conceptos fundamentales

Se define en principio lo que es una *variable aleatoria*, la cual se considera un concepto de suma importancia en la probabilidad y constituye la base de este tema.

Definición 14 (Variable aleatoria [DV05])**.** *Es aquella que toma valores que, a priori, no concemos con certeza.*

A *priori*, en este contexto, significa *antes de conocer el resultado de un acontecimiento, de un experimento o de una elección al azar.*

Otra definición de variable aleatoria es la siguiente:

Definición 15 (Variable aleatoria [DeG88])**.** *Considérese un experimento cuyo espacio muestral es el conjunto S. Una función con valores reales que está definida sobre el espacio S recibe el nombre de variable aleatoria.*

En otras palabras, en un experimento concreto, una variable aleatoria X, sería una función que asigna un número real $X(s)$ a cada resultado posible $s \in S$.

Para poder construir un modelo matemático del problema probabilístico al cual nos enfrentamos, es necesario seleccionar un conjunto de variables aleatorias para las cuales los valores de cada una de ellas sean *exclusivos* y *exhaustivos* en su dominio.

En adelante se utilizará la convención de representar las variables aleatorias con letras mayúsculas acompañadas, de ser necesario, de subíndices para diferenciarlas, por ejemplo: X, Y_1, Y_2 y Z. Los valores de las variables aleatorias se representarán mediante letras minúsculas acompañadas, de ser

necesario, de subíndices y superíndices para lograr su diferenciación. Como ejemplo supongase que la variable X tiene un conjunto de tres valores posibles: x^1, x^2 y x^3; supongase que la variable Y_1 es booleana y, por tanto, puede tomar sólo los siguientes valores: y_1^t e y_1^f*. Los conjuntos de variables aleatorias como: $X_1, \ldots, X_n$, se representarán mediante $\bar{X}$. De forma análoga, cuando cada una de las variables X_i del conjunto $\bar{X}$ tome algún valor concreto, la n-tupla $\bar{x} = (x_1, \ldots, x_n)$ representará este hecho.

Las proposiciones y teoremas siguientes se presentan sin demostración, el lector interesado puede encontrar tal información en los libros de probabilidad y redes bayesianas que se mencionan en la bibliografía [DeG88, DV05, SJN00, SJN04].

Definición 16 (Distribución de una variable aleatoria)**.** *Sea A cualquier subconjunto de la recta real y sea $P(X \in A)$ la probabilidad de que el valor de X pertenezca al subconjunto A. Entonces $P(X \in A)$ es igual a la probabilidad de que el resultado s del experimento sea tal que $X(s) \in A$. Esto es,*

$$P(X \in A) = P(s : X(s) \in A)$$

A.1.1. Distribuciones discretas

Definición 17 (Distribución discreta)**.** *Se dice que una variable aleatoria X tiene una* distribución discreta *si X sólo puede tomar un número finito k de valores distintos $x^1, \ldots, x^k$ o, a lo sumo, una sucesión infinita de valores distintos $x^1, x^2, \ldots$*

Definición 18 (Función de probabilidad)**.** *Si una variable aleatoria X tiene una distribución discreta, la* función de probabilidad *(f.p.) de X se define como la función f tal que para cualquier número real x,*

$$f(x) = P(X = x)$$

Para cualquier punto x que no es uno de los valores posibles de X, $f(x) = 0$. Además, si la sucesión $x^1, x^2, \ldots$ incluye todos los valores posibles de X, entonces:

$$\sum_{i=1}^{\infty} f(x^i) = 1 \tag{A.1}$$

*Los superíndices utilizados para este par de expresiones se corresponden con las palabras en inglés: t - true, y f - false, respectivamente.

Si X tiene una distribución discreta, se puede determinar la probabilidad de cualquier subconjunto A de la recta real a partir de la relación:

$$P(X \in A) = \sum_{x^i \in A} f(x^i)$$

A.1.2. Distribuciones continuas

Definición 19 (Función de densidad de probabilidad)**.** *Se dice que una variable aleatoria X tiene una distribución continua si existe una función no negativa f, definida sobre la recta real, tal que para cualquier intervalo A,*

$$P(X \in A) = \int_A f(x)dx$$

La función f se denomina función de densidad de probabilidad *(f.d.p.) de X.*

Toda f.d.p. debe satisfacer los siguientes dos requisitos:

1. $$f(x) \geq 0$$

2. $$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{A.2}$$

Propiedad 5. *Si una variable X tiene una distribución continua, entonces $P(X = x) = 0$ para todo valor individual x.*

La f.d.p. de una variable aleatoria no es única ya que, dada la propiedad anterior, los valores de cualquier f.d.p. de una variable aleatoria X se pueden modificar arbitrariamente en una sucesión infinita de puntos sin afectar las probabilidades que involucran a X, es decir, sin afectar la distribución de probabilidad de X.

A.1.3. Función de distribución

Definición 20 (Función de distribución)**.** *La* función de distribución *(f.d.) F de una variable aleatoria X es una función definida para cada número real x como sigue:*

$$F(x) = P(X \leq x) \quad para \quad -\infty < x < \infty$$

Esta definición es válida para cualquier variable aleatoria X, ya sea discreta continua o mixta.

Resulta de la ecuación anterior que la f.d. de cualquier variable aleatoria cumple con las tres propiedades siguientes:

1. La función $F(x)$ es no decreciente a medida que x crece; esto es, si $x^1 < x^2$, entonces $F(x^1) \leq F(x^2)$.
2. $\lim_{x \to -\infty} F(x) = 0$ y $\lim_{x \to \infty} F(x) = 1$
3. Una f.d. siempre es continua por la derecha; esto es, $F(x) = F(x^+)$ en todo punto x.

A partir de la f.d. de una variable aleatoria X es posible calcular la probabilidad de que X esté en cualquier intervalo de la recta real. Los siguientes teoremas muestran la forma en que tal probabilidad puede ser calculada.

Teorema 2. *Para cualquier valor x,*

$$P(X > x) = 1 - F(x)$$

Teorema 3. *Para cualesquiera valores concretos x_1 y x_2 tales que $x_1 < x_2$,*

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

Teorema 4. *Para cualquier valor x,*

$$P(X < x) = F(x^-)$$

Teorema 5. *Para cualquier valor x,*

$$P(X = x) = F(x^+) - F(x^-)$$

Para una variable aleatoria discreta X, su f.d. $F(x)$ debe tener la siguiente forma: $F(x)$ tendrá un salto de magnitud $f(x^i)$ en cada valor posible x^i de X; y $F(X)$ será constante entre dos saltos sucesivos cualesquiera. Hay que resaltar que la distribución de una variable aleatoria discreta X se puede representar indistintamente por la f.p. o la f.d. de X [DeG88].

En el caso de una variable aleatoria continua X, dado que la probabilidad de cualquier punto x es 0, su f.d. $F(x)$ no tiene saltos y por tanto, $F(x)$ es continua sobre toda la recta real. Además, puesto que,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

resulta que para cualquier punto x en el que $f(x)$ es continua,

$$F'(x) = \frac{dF(x)}{dx} = f(x)$$

La distribución de una variable aleatoria continua X puede ser representada indistintamente por la f.d.p. o la f.d. de X.

A.1.4. Distribuciones multivariantes

Distribuciones conjuntas

Definición 21 (Función de distribución conjunta)**.** *La función de distribución conjunta (f.d.) de n variables aleatorias $X_1, \ldots, X_n$ se define como la función F cuyo valor en cualquier punto $(x_1, \ldots, x_n)$ de un espacio n-dimensional $\mathbb{R}^n$ está dado por la relación*

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

Toda f.d. multivariante satisface propiedades análogas a las presentadas para las f.d. univariantes.

Definición 22 (Distribución discreta conjunta)**.** *Se dice que el conjunto de variables aleatorias $\bar{X} = X_1, \ldots, X_n$ con f.d. definida sobre el espacio $\mathbb{R}^n$, tiene una* distribución discreta conjunta *si sólamente puede tomar un número finito o una sucesión infinita de valores distintos posibles $\bar{x}$ en $\mathbb{R}^n$.*

De acuerdo con esto, la f.p. conjunta de $\bar{X}$ se define como la función f tal que para cualquier punto $\bar{x} \in \mathbb{R}^n$,

$$f(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

Observar también que para cualquier subconjunto $A \in \mathbb{R}^n$,

$$P(\bar{X} \in A) = \sum_{\bar{x} \in A} f(\bar{x})$$

Otra definición de la f.p. conjunta que puede ayudarnos a comprenderla mejor es la siguiente.

Definición 23 (f.p. conjunta)**.** *Dado un conjunto de variables $\bar{X} = \{X_1, \ldots, X_n\}$ que tienen una distribución conjunta discreta, definimos la* f.p. conjunta *como la función f que a cada n-tupla $\bar{x} = (x_1, \ldots, x_n)$ le asigna un número real no negativo de modo que:*

$$\sum_{\bar{x}} f(\bar{x}) = \sum_{x_1} \cdots \sum_{x_n} (x_1, \ldots, x_n) = 1$$

Para variables aleatorias continuas, las definiciones análogas se presentan a continuación.

Definición 24 (Distribución continua conjunta)**.** *Se dice que el conjunto de variables aletorias continuas* $\bar{X} = \{X_1, \ldots, Xn\}$ *con f.d. definida sobre el espacio* $\mathbb{R}^n$*, tiene una* distribución continua conjunta *si existe una función no negativa* f *definida sobre* $\mathbb{R}^n$ *tal que para cualquier subconjunto* $A \subset \mathbb{R}^n$*,*

$$P(\bar{X} \in A) = \int \cdots \int_A f(\bar{x})\, dx_1 \cdots dx_n$$

La función f se denomina la f.d.p. conjunta de $\bar{X}$.

Si la distribución conjunta de $\bar{X}$ es continua, entonces la f.d.p. conjunta f se puede obtener a partir de la f.d. conjunta F utilizando la relación:

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1, \ldots, x_n}$$

para todos los puntos $x_1, \ldots, x_n$ en los que existe la derivada.

Distribuciones marginales

Si se conoce la distribución conjunta de $\bar{X}$ es posible calcular distribución marginal de cualquier subconjunto $\bar{X}' \subset \bar{X}$. Hay que señalar que en el caso de variables discretas se ha de hablar de f.p. marginal, y de f.d.p. marginal en el de variables continuas.

Definición 25 (f.d. marginal)**.** *Sea* F *la f.d. conjunta de* $\bar{X} = \{X_1, \ldots, X_n\}$*, la f.d. marginal* F_m *de* $\bar{X}' = \{X_1, \ldots, X_k\} \subset \bar{X}$ *está dada por:*

$$F_m(\bar{x}') = F_m(x_1, \ldots, x_k) = \lim_{\substack{x_{k+1} \to \infty \\ \vdots \\ x_n \to \infty}} F(x_1, \ldots, x_n)$$

El cálculo de una f.d. marginal es indistinto del tipo de distribución que presenten las variables.

Definición 26 (f.p. marginal)**.** *Sea* f *la f.p. conjunta de* $\bar{X} = \{X_1, \ldots, X_n\}$*. La f.p. marginal* f_m *para un subconjunto de variables* $\bar{X}' = \{X'_1, \ldots, X'_{n'}\} \subset \bar{X}$ *viene dada por:*

$$f_m(\bar{x}') = f_m(x'_1, \ldots, x'_{n'}) = \sum_{x_i | X_i \notin \bar{X}'} f(x_1, \ldots, x_n)$$

Definición 27 (f.d.p. marginal)**.** *Si f es la f.d.p. conjunta de $\bar{X} = \{X_1, \ldots, X_n\}$, entonces la f.d.p. marginal f_m de $\bar{X}' = \{X_1, \ldots, X_k\} \subset \bar{X}$ viene dada por:*

$$f_m(\bar{x}') = f_m(x_1, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_{k+1} \ldots dx_n$$

Esta última ecuación nos indica que la f.d.p. conjunta marginal de cualquier k de las n variables aletorias $X_1, \ldots, X_n$ se puede determinar integrando la f.d.p. conjunta sobre todos los valores posibles de las $n - k$ variables restantes.

Para entender con mayor claridad el significado de las dos definiciones anteriores, es necesario observar que la función f_m es en sí una nueva f.p. o f.d.p conjunta como se propone a continuación.

Proposición 1. *Sea f una distribución de probabilidad conjunta para $\bar{X}$, la cual puede ser representada por una f.p. o una f.d.p. conjunta, toda distribución de probabilidad marginal obtenida a partir de f para un subconjunto $\bar{X}' \subset \bar{X}$ es a su vez una distribución de probabilidad conjunta para $\bar{X}'$.*

En general, las distribuciones marginales nos permite estudiar los casos en que nos interesa observar el problema sin la influencia directa de ciertas variables, es decir, nos interesa observarlo cuando solamente los eventos correspondientes a ciertas variables pueden llevarse a cabo.

Distribuciones condicionales

Consideremos el caso de trabajar con variables que presentan una distribución conjunta continua. Supóngase que se hace una partición del conjunto $\bar{X} = \{X_1, \ldots, X_n\}$ en dos subconjuntos $\bar{Y}$ y $\bar{Z}$, donde k es el número de elementos de $\bar{Y}$. Supóngase, además, que la f.d.p. conjunta de $(\bar{Y}, \bar{Z})$ es f y que la f.d.p. conjunta marginal de $\bar{Z}$ es f_m. Entonces, para cualquier punto $\bar{z} \in \mathbb{R}^{n-k}$ tal que $f_m(\bar{z}) > 0$, la f.d.p. condicional g de $\bar{Y}$ cuando $\bar{Z} = \bar{z}$ se define como sigue:

$$g(\bar{y}|\bar{z}) = \frac{f(\bar{y}, \bar{z})}{f_m(\bar{z})} \quad \text{para} \quad \bar{y} \in \mathbb{R}^k \tag{A.3}$$

Si los conjuntos $\bar{Y}$ y $\bar{Z}$ tienen una distribución conjunta discreta cuya f.d. es f y si la f.p. marginal de $\bar{Z}$ es f_m, entonces, la f.p. condicional $g(\bar{y}|\bar{z})$ de $\bar{Y}$ para cualquier valor concreto $\bar{Z} = \bar{z}$ también se puede especificar por la ecuación A.3.

Una distribución de probabilidad condicional nos indica qué tan probable es un evento cuando ya ha sucedido parte del mismo. Si suponemos un

experimento imaginario que se lleva a cabo un número infinito de veces y que involucra a los conjuntos de variables ya definidos $\bar{X}$, $\bar{Y}$ y $\bar{Z}$, podemos ver que $f_m(\bar{z})$ es la relación del número de veces que se obtiene exactamente el resultado $\bar{z}$ y el número total de veces que se intenta obtener un resultado para $(\bar{Y}, \bar{Z})$. Entonces, la probabilidad de que suceda $\bar{y}$ cuando $\bar{z}$ ya ha sucedido, es el número de veces en que suceden $\bar{y}$ y $\bar{z}$ en relación, solamente, al número de veces que sucede $\bar{z}$. Desde esta perspectiva, una distribución de probabilidad condicional es la versión ajustada o normalizada de una distribución de probabilidad conjunta.

Proposición 2. *Sea $\bar{X}$ un conjunto de n variables aleatorias con distribución conjunta continua y sean $\bar{Y}$ y $\bar{Z}$ subconjuntos que conforman una partición de $\bar{X}$, con k igual al número de elementos de $\bar{Y}$. Sea f_m la f.d.p. conjunta marginal de $\bar{Z}$. Si $\bar{z}$ es un punto en el espacio $\mathbb{R}^{n-k}$ tal que $f_m(\bar{z}) > 0$ se cumple que:*

$$\forall \bar{z},\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\bar{y}|\bar{z})\, dy_1 \cdots dy_k = 1 \tag{A.4}$$

Si el conjunto $\bar{X}$ tiene una distribución conjunta discreta y f_m representa la f.p. conjunta marginal de $\bar{Z}$, entonces la Ecuación A.4 se convierte en:

$$\forall \bar{z},\ \sum_{\bar{y}} g(\bar{y}|\bar{z}) = 1$$

El teorema de la *probabilidad total* nos permite calcular la distribución marginal de un conjunto de variables a partir de distribuciones condicionadas. La siguiente enunciación de dicho teorema corresponde al caso en que las variables involucradas tienen una distribución conjunta continua.

Teorema 6 (Teorema de la probabilidad total – caso continuo)**.** *Sea $\bar{X}$ un conjunto de n variables aleatorias con distribución conjunta continua y sean $\bar{Y}$ y $\bar{Z}$ subconjuntos que conforman una partición de $\bar{X}$, con k igual al número de elementos de $\bar{Y}$. Si f_1 y f_2 son las f.d.p. conjuntas marginales k-dimensional y $(n-k)$-dimensional de $\bar{Y}$ y $\bar{Z}$ respectivamente, se cumple que:*

$$f_1(\bar{y}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\bar{y}|\bar{z}) \cdot f_2(\bar{z})\, dz_1 \cdots dz_{n-k} \quad \textit{para} \quad \bar{y} \in \mathbb{R}^k \tag{A.5}$$

Si el conjunto $\bar{X}$ tiene una distribución conjunta discreta y f_1 y f_2 representan las f.p. conjuntas marginales de $\bar{Y}$ y $\bar{Z}$ respectivamente, la Ecuación A.5 queda expresada como:

$$f_1(\bar{y}) = \sum_{\bar{z}|f_2(\bar{z}>0)} g(\bar{y}|\bar{z}) \cdot f_2(\bar{z})$$

La condición $f_2(\bar{z}) > 0$ no es indispensable ya que, simplemente evita el trabajo de llevar a cabo operaciones que no modifican el resultado final; evita el trabajo de multiplicar por 0 y sumar 0 repetidamente.

La siguiente proposición se deduce, en el caso de trabajar con una distribución conjunta continua, de la definición de f.d.p. condicional y de la aplicación del teorema de la probabilidad total.

Proposición 3 (Factorización de la f.d.p. conjunta)**.** *Dado un conjunto de variables $\bar{X}$ con una distribución conjunta continua f y una partición $\{\bar{X}_1, \ldots, \bar{X}_k\}$ de $\bar{X}$, si g_i representa la f.d.p. condicional del conjunto $\bar{X}_i$, entonces se cumple que:*

$$f(\bar{x}) = \prod_{i=1}^{k} g_i(\bar{x}_i | \bar{x}_{i+1}, \ldots, \bar{x}_k) \tag{A.6}$$

Si el conjunto $\bar{X}$ tiene una distribución conjunta discreta, f representa la f.p. conjunta de $\bar{X}$ y g_i es la f.p. condicional del conjunto X_i , la ecuación A.6 es igualmente válida.

A.1.5. Independencia condicional

Se dice que n variables $X_1, \ldots, X_n$ son independientes si, para n conjuntos cualesquiera $A_1, A_2, \ldots, A_n$ de números reales,

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2) \ldots P(X_n \in A_n)$$

Si se define F como la f.d. conjunta de $X_1, \ldots, X_n$ y F_i como la f.d. marginal univariante de X_i para $i = 1, \ldots, n$, entonces resulta de la definición de independencia que las variables $X_1, \ldots, X_n$ son independientes si, y sólo si, para todos los puntos $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$,

$$F(x_1, x_2, \ldots, x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n)$$

Además, si las variables $X_1, \ldots, X_n$ tienen una distribución conjunta continua cuya f.d.p. conjunta es f , y si f_i es la f.d.p. marginal univariante de X_i para $i = 1, \ldots, n$, entonces $X_1, \ldots, X_n$ son independientes si, y sólo si, para todos los puntos $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ se satisface la siguiente relación:

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2) \cdots f_2(x_n) \tag{A.7}$$

Si $X_1, \ldots, X_n$ tienen una ditribución discreta conjunta con f.p. f y f.p. marginal f_i de X_i –con $i = 1, \ldots, n$–, estas variables son independientes si se satisface la ecuación A.7.

Variables aleatorias condicionalmente independientes

Supóngase que se hace una partición del conjunto $\bar{W} = \{W_1, \ldots, W_n\}$ en tres subconjuntos $\bar{X}$, $\bar{Y}$ y $\bar{Z}$. Supóngase además que la f.d.p. conjunta de $(\bar{X}, \bar{Y}, \bar{Z})$ es f. Sea f_m la f.d.p. conjunta marginal de $\bar{Z}$, y sean denotadas por g_i las f.d.p. condicionales calculadas a partir de f. Se dice que los conjuntos de variables $\bar{X}$ e $\bar{Y}$ son condicionalmente independientes dado el conjunto $\bar{Z}$ si

$$\forall \bar{x}, \bar{y}, \bar{z}, \quad f_m(\bar{z}) > 0 \quad \Longrightarrow \quad g_1(\bar{x}, \bar{y}|\bar{z}) = g_2(\bar{x}|\bar{z}) \cdot g_3(\bar{y}|\bar{z}) \tag{A.8}$$

Si A.8 se cumple, se dice entonces que el conjunto de variables $\bar{Z}$ *separa* a los conjuntos $\bar{X}$ e $\bar{Y}$.

A.2. Teorema de Bayes

En esta sección, por claridad, se tratarán todas las variables aleatorias como discretas y sus funciones de probabilidad serán generalizadas mediante la expresión de su probabilidad, es decir, para indicar $f(x)$, donde f es la f.p. de X, se escribirá $P(X = x)$ o $P(x)$, donde $P(X)$ se corresponde con f.

Teorema 7 (Teorema de Bayes generalizado)**.** *Dadas dos n-tuplas $\bar{x}$ e $\bar{y}$ de dos conjuntos de variables $\bar{X}$ e $\bar{Y}$, respectivamente, tales que $P(\bar{x}) > 0$ y $P(\bar{y}) > 0$, se cumple que*

$$P(\bar{x}|\bar{y}) = \frac{P(\bar{x}) \cdot P(\bar{y}|\bar{x})}{\sum\limits_{\bar{x}'|P(\bar{x}'>0)} P(\bar{x}') \cdot P(\bar{y}|\bar{x}')}$$

Proposición 4. *Dados tres subconjuntos $\bar{X}$, $\bar{Y}$ y $\bar{Z}$, si $P(\bar{y}, \bar{z}) > 0$, se cumple que*

$$P(\bar{x}, \bar{y}|\bar{z}) = P(\bar{x}|\bar{y}, \bar{z}) \cdot P(\bar{y}|\bar{z}) \tag{A.9}$$

Proposición 5 (Teorema de Bayes con condicionamiento)**.** *Dadas tres tuplas $\bar{x}$, $\bar{y}$ y $\bar{z}$ de tres conjuntos de variables $\bar{X}$, $\bar{Y}$ y $\bar{Z}$, respectivamente, tales que $P(\bar{x}, \bar{z}) > 0$ y $P(\bar{y}, \bar{z}) > 0$, se cumple que*

$$P(\bar{x}|\bar{y}, \bar{z}) = \frac{P(\bar{x}|\bar{z}) \cdot P(\bar{y}|\bar{x}, \bar{z})}{\sum\limits_{\bar{x}'|P(\bar{x}'|\bar{z})>0} P(\bar{y}|\bar{x}', \bar{z}) \cdot P(\bar{x}'|\bar{z})}$$

Una forma útil de escribir el teorma de Bayes es en su *forma normalizada*. Si $\bar{X}$ e $\bar{Y}$ cumplen con las condiciones expuestas en la enunciación del teorema de Bayes generalizado, entonces:

$$P(\bar{x}|\bar{y}) = \alpha \cdot P(\bar{x}) \cdot P(\bar{y}|\bar{x})$$

donde:

$$\alpha \equiv \left[\sum_{\bar{x}'} P(\bar{x}') \cdot P(\bar{y}|\bar{x}') \right]^{-1} = [P(\bar{y})]^{-1}$$

En la práctica, al aplicar el teorema de Bayes, es necesario utilizar las definiciones siguientes:

Hallazgo. Es la determinación del valor de una variable, $H = h$, a partir de un dato –una observación, una medida, etc.–.

Evidencia. Es el conjunto de todos los hallazgos disponibles en un determinado momento o situación: $\mathbf{e} = \{H_1 = h_1, \ldots, H_n = h_n\}$

Probabilidad a priori. Es la probabilidad de una variable o subconjunto de variables cuando no hay ningún hallazgo.

La probabilidad a priori de $\bar{X}$ coincide, por tanto, con la probabilidad marginal $P(\bar{x})$.

Probabilidad a posteriori. Es la probabilidad de una variable o subconjunto de variables dada la evidencia $\mathbf{e}$. Se representa mediante P^*:

$$P^*(\bar{x}) \equiv P(\bar{x}|\,\mathbf{e})$$

A.3. Distribuciones de probabilidad continuas

Enseguida se presentan las definiciones de las distribuciones de probabilidad normal, log-normal, gamma y exponencial.

A.3.1. Distribución normal

Definición 28. *Una variable aletoria X con función de densidad de probabilidad*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

tiene una distribución normal *con parámetros μ, donde $-\infty < \mu < \infty$, y $\sigma > 0$.*

Además,

$$E(X) = \mu \quad y \quad V(X) = \sigma^2$$

El valor de $E(X) = \mu$ determina el centro de la función de densidad de probabilidad y el valor de $V(X) = \sigma^2$ determina la anchura.

La notación $N(\mu, \sigma^2)$ denota una distribución normal con media μ y varianza σ^2.

A.3.2. Distribución log-normal

La *distribución log-normal* se obtiene cuando los logaritmos de una variable se describen mediante una distribución normal.

Definición 29. *La variable aleatoria X tiene una* distribución log-normal *si* ***ln*** *X tiene una distribución normal. Su función de densidad de probabilidad está dada por:*

$$f(x) = \frac{1}{\sigma_l \, x \, \sqrt{2\pi}} \, e^{-\frac{1}{2}\left(\frac{lnx - \mu_l}{\sigma}\right)^2}, \quad para \quad 0 < x < \infty$$

Los parámetros necesarios para especificar la función son: μ_l, que es la media de la transformada del logaritmo natural de los datos; y σ_l^2, que es la varianza de la transformada del logaritmo natural de los datos.

A.3.3. Distribución gamma

Definición 30. *La* función gamma *es:*

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} \, dx \quad para \quad r > 0$$

Definición 31. *La variable aleatoria X con función de densidad de probabilidad:*

$$f(x) = \frac{\lambda^r \, x^{r-1} \, e^{-\lambda x}}{\Gamma(r)}, \quad para \quad x > 0$$

tiene una distribución gamma *con parámetros $\lambda > 0$ y $r > 0$. Si r es un entero, entonces X tiene una distribución de Erlang.*

Definición 32. *Si X es una variable aleatoria gamma con parámetros λ y r, entonces la media y la varianza de X son:*

$$\mu = E(x) = \frac{r}{\lambda} \quad y \quad \sigma^2 = V(X) = \frac{r}{\lambda^2}$$

A.3.4. Distribución exponencial

Definición 33. *Una variable aleatoria X que es igual a la distancia entre conteos sucesivos de un proceso de Poisson con media $\lambda > 0$ tiene una* distribución exponencial *con parámetro λ. La función de densidad de probabilidad de X es:*

$$f(x) = \lambda\, e^{-\lambda\, x} \quad \textit{para} \quad 0 \leq x < \infty$$

La distribución exponencial debe su nombre a la función exponencial de la función de densidad de probabilidad.

Definición 34. *Si la variable aleatoria X tiene una distribución exponencial con parámetro λ, entonces:*

$$E(X) = \frac{1}{\lambda} \quad y \quad V(X) = \frac{1}{\lambda^2}$$

Apéndice B

Manual de usuario del software

El software desarrollado en este proyecto, se ha compilado y probado sobre los sistemas operativos Linux y Windows. Los detalles de instalación, ejecución y utilización se decriben en las secciones siguientes.

B.1. Proceso de Instalación

El software del clasificador se distribuye en archivos binarios ejecutables sobre sistemas Linux y Windows. Se distribuye también el código fuente del software, el cual puede ser compilado sin modificación tanto para sistemas Linux como Windows.

En este apartado se describirá sólo el proceso de instalación y ejecución del software que se distribuye ya compilado. Las instrucciones de compilación del código puede encontrarse en el archivo README que acompaña al código fuente. Se sugiere leer el archivo README que acompaña al software, cualquiera que sea la forma de distribución que se ha de utilizar.

B.1.1. Dependencias del software

La versión compilada del software depende de dos paquetes específicos para su ejecución. Estos paquetes proporcionan la plataforma de interfaz gráfica del sistema. Los paquetes son los siguientes:

GTK+ En Linux es necesario contar con estas librerías correctamente instaladas. Las instrucciones de instalación y los archivos necesarios pueden encontrarse en: *http://www.gtk.org/download/*.

Para Windows, se debe instalar el *Runtime Environment* de GTK+. De: *http://gladewin32.sourceforge.net/modules/news/*, puede descargarse un instalador de este ambiente.

GTKMM Para Linux, lo más conveniente es visitar la página de gtkmm: *http://www.gtkmm.org/download.shtml*, donde se encuentra ampliamente explicado el proceso de instalación de estas librerías.

Para Windows, puede descargarse el instalador del *Runtime Environment* de: *http://ftp.gnome.org/pub/gnome/binaries/win32/gtkmm/*.

Una vez instaladas estas librerías, ya puede ejecutarse el software.

Hay que mencionar que el clasificador de leucocitos utiliza también las librerías de FFTW –*http://www.fftw.org/*–, para llevar a cabo el cálculo de la transformada rápida de Fourier. Se utilizan también, librerías desarrolladas y distribuidas por el equipo de *R Project* –*http://www.r-project.org/*–. Estas últimas librerías se utilizan para el cálculo de las funciones de densidad de probabilidad de las distribuciones: normal, log-normal, exponencial y gamma. Versiones ya compiladas de estas librerías acompañan al software del clasificador, así que no es necesario descargarlas ni instalarlas por separado.

B.1.2. Instalación y ejecución

El sistema, en su versión ya compilada, se distribuye en un archivo comprimido en formato `rar` –`leukoClassifier--bin-win-0.03.rar`, compilado para windows y `leukoClassifier-bin-lin-0.03.rar`, compilado para linux–, sólo hay que descomprimirlo respetando la jerarquía de directorios que presenta. Dentro del directorio principal, denominado `LeukoClassifier`, se encuentra un archivo ejecutable llamado `classifier`, es éste el archivo que inicia la ejecución del clasificador.

B.2. Utilización del software

Como requerimiento especial de esta implementación, la máscara de segmentación debe llevar el mismo nombre que la imagen a analizar más el posfijo "_msk_sg" como se indica a continuación:

Nombre del archivo de imagen:
`nombreImagen.bmp`

Nombre del archivo de máscara de segmentación:
`nombreImagen_msk_sg.bmp`

B.2.1. Interfaz principal

La ventana principal del programa la conforman tres componentes –Ver Figura B.1– :

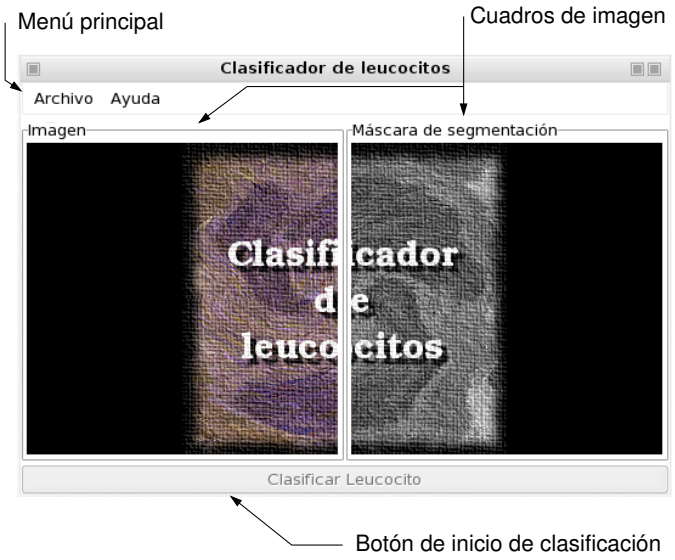


Figura B.1: Ventana principal del programa.

- El menú principal. Éste muestra dos submenús: *Archivo* y *Ayuda*.

  El submenú *Archivo* contiene las opciones: *Abrir imagen* y *Salir*.

  El submenú *Ayuda* presenta las opciones de *Descripción* y *Acerca de*.

- Los cuadros de imágenes. Son espacios en los cuales se muestran la fotografía de la célula que ha de clasificarse y su máscara de segmentación.

- El botón de inicio de clasificación. Inicialmente desactivado, en cuanto se carga la fotografía de la célula a clasificar, se activa para poder dar comienzo al proceso de clasificación.

B.2.2. Clasificación de una imagen

En el menú principal seleccionamos *Abrir imagen* y aparecerá el cuadro de diálogo que se muestra en la Figura B.2

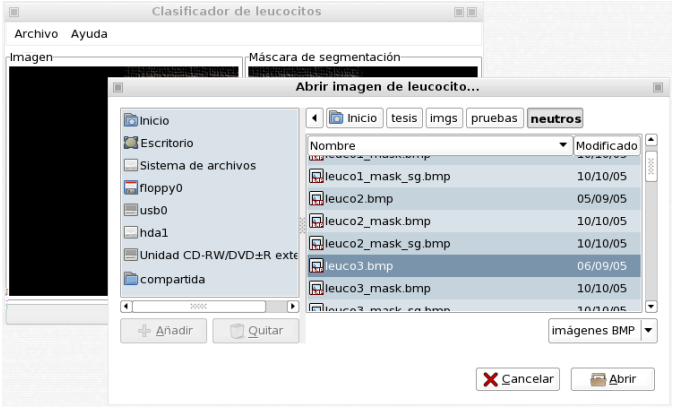


Figura B.2: Cuadro de diálogo de abrir archivo.

Buscamos y abrimos el archivo de la fotografía celular que queremos clasificar –leuco_03.bmp, por ejemplo–. Enseguida, la fotografía de la célula a clasificar será desplegada en el cuadro de imagen izquierdo. La máscara de segmentación será desplegada en el cuadro de imagen derecho. El botón de *inicio de clasificación* se activará como indicación de que puede en este momento iniciarse el proceso de clasificación. La Figura B.3 muestra el estado de la aplicación en este punto.

Presionamos el botón: *Clasificar Leucocito* para comenzar la clasificación de la fotografía celular. Enseguida, una ventana que muestra el estado del proceso aparecerá. En ella se informa del proceso que actualmente se lleva a cabo. Se muestra también una barra de estado que indica el porcentaje del proceso completo de clasificación que se ha realizado. La Figura B.4 ilustra este punto del proceso de clasificación.

Cuando el proceso de clasificación termina, aparece una ventana en la que se muestran los resultados de clasificación. En ella se muestran los nombres de los cinco tipos de células a clasificar y a la derecha de cada uno se muestra su valor de probabilidad asociado. El tipo de célula con mayor valor de probabilidad se muestra en color *verde* indicando que es éste el asociado a la fotografía clasificada. En la Figura B.5 puede observarse esta ventana de resultados.

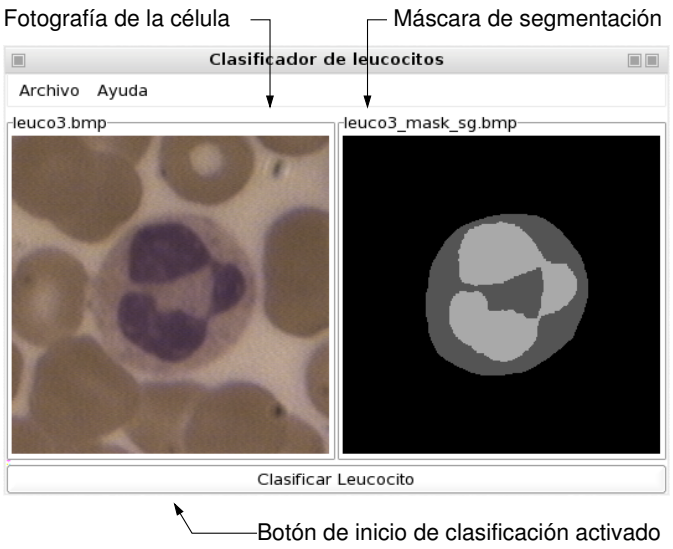


Figura B.3: Aplicación inmediatamente después de abrir una fotografía celular para su clasificación.

Figura B.4: Aplicación en ejecución del proceso de clasificación.

Figura B.5: Ventana de resultados de clasificación.

B.2.3. Ayuda del sistema

La aplicación cuenta con una breve ayuda integrada. Para acceder a ella seleccionamos: *Ayuda* en el menú principal y enseguida: *Descripción del sistema*. Aparecerá una ventana que presenta los tópicos:

- ¿Qué hace el programa?
- ¿Cómo funciona?
- ¿Cómo se utiliza?
- ¿Cómo interpretar los resultados?
- Restricciones generales del programa

La Figura B.6 muestra la ventana de ayuda del sistema.

Para visualizar la información de alguno de los temas listados, hacemos click sobre el título que nos interesa y la información relacionada será mostrada en la misma ventana. Para regresar al listado de tópicos, recorremos el tema actual hasta encontrar la leyenda *volver al índice*, la cual se muestra en color azul. Hacemos click directamente sobre ella y enseguida se mostrará el índice de los temas de ayuda.

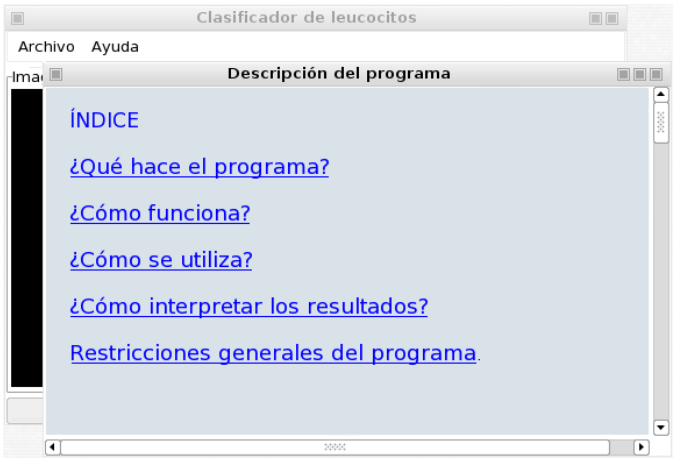


Figura B.6: Interfaz de ayuda de la aplicación.

Glosario

Basófilo Es uno de los polimorfonucleares, al igual que los neutrófilos y los eosinófilos. Los gránulos de los basófilos son gruesos pero escasos. Se originan en la médula ósea y son los menos numerosos. Tienen una activa participación en la respuesta inmunitaria, a través de la liberación de histamina, serotonina en bajas concentraciones, y otras sustancias químicas.

Biometría hemática Es el estudio de laboratorio destinado a informar sobre número y características de las células de la sangre.

Centrósfera Material amorfo que rodea a los centriolos que conforman el centrosoma celular.

Citometría hemática Ver *Biometría hemática.*

Citoplasma celular Parte del protoplasma que en una célula eucariota se encuentra entre el núcleo celular y la membrana plasmática.

Célula Unidad fundamental de los organismos vivos, generalmente de tamaño microscópico, capaz de reproducción independiente y formada por un citoplasma y un núcleo rodeados por una membrana.

Célula anormal Célula que no presenta alguno o varios de los rasgos característicos de su tipo o que presenta características atípicas.

Célula eucariota Célula que tiene su material hereditario fundamental –su información genética– encerrado dentro de una doble membrana, la envoltura nuclear, que delimita un núcleo celular.

Célula inmadura Célula en estado desarrollo que aún no se ha diferenciado completamente y que, por tanto, no ha adquirido los rasgos característicos de su tipo.

Célula normal Célula que no presenta ningún tipo de enfermedad y en la cual se observan rasgos característicos de su tipo.

Célula plasmática Célula que deriva de los linfocitos, normalmente ausente de la sangre circulante pero presente en gran cantidad en el sistema linfático, que posee la propiedad de sintetizar las inmunoglobinas –es decir los anticuerpos–.

Eosinófilo Leucocito granulocito pequeño derivado de la médula ósea, tiene una vida media en la circulación de 6 a 12 horas antes de migrar a los tejidos en donde permanece por varios días. Su núcleo bilobulado es característico y sus gránulos citoplásmicos son distintivos. Los eosinófilos pueden regular la respuesta alérgica y las reacciones de hipersensibilidad.

Eritrocito Célula sanguínea esferoidal que contiene la hemoglobina, que aporta el color rojo característico a la sangre y actúa transportando el oxígeno por el organismo; hematíe.

Frotis sanguíneo Extendido suave y delgado de sangre sobre un porta-objeto. Este extendido, una vez coloreado, permite verificar visualmente y en forma global, sobre la línea sanguínea, los hematíes, leucocitos y plaquetas.

Gránulo Vesícula de secreción presente en el citoplasma leucocitario, que almacena en su interior enzimas lisosomales, y que al microscopio se observa como un punto bien definido.

Leucocito Célula blanca o incolora de la sangre y la linfa, que puede trasladarse a diversos lugares del cuerpo con funciones defensivas.

Linfocito Célula sanguínea mononucleada que tiene un papel fundamental en la respuesta inmunológica y que se encuentra normalmente en la sangre y en los órganos linfoídes –bazo, timo y ganglios linfáticos–. Existen dos tipos morfológicamente idénticos: los linfocitos T –timodependientes– que intervienen fundamentalmente en la inmunidad celular y los linfocitos B que se encargan de la elaboración de anticuerpos.

Monocito Leucocito de los denominados agranulocitos, es el leucocito más grande de todos con un tamaño de 15 a 20 micras. Presenta un núcleo arriñonado –forma de riñón–, que se tiñe de color violeta-azulado con una proporción 2:1 con respecto al resto de la célula. Su principal función es la de fagocitar o comerse a diferentes microorganismos o restos celulares.

Neutrófilo Denominado también micrófago. Es un glóbulo blanco del tipo de los granulocitos, mide de 12 a 18 micras, es el tipo de leucocito más abundante en la sangre. Se caracteriza por presentar un núcleo con cromatina compacta segmentada en 2 a 5 lóbulos conectados por delgados puentes. Su citoplasma contiene abundantes gránulos finos color púrpura que contienen abundantes enzimas destructoras, así como una sustancia antibacteriana llamada fagocitina, necesarias para la lucha contra los gérmenes extraños.

Nucléolo Orgánulo del núcleo que tiene como principal función la síntesis de los ARN robosómicos. Se encuentra en todos los núcleos de las células eucariotas, con excepción de algunos espermatozoides y los núcleos de segmentación de los anfibios. Son densos, no están rodeados por membrana y aparecen y desaparecen durante la división celular.

Núcleo celular El núcleo celular es la estructura más característica de las células eucariotas. Se rodea de una cubierta propia, llamada envoltura nuclear y contiene el material hereditario, que es la base del repertorio de instrucciones propias de desarrollo y el funcionamiento de cada organismo, y cuya composición tiene como base el ácido desoxirribonucleico.

Protoplasma El protoplasma es citoplasma más el núcleo. Mientras que la célula es membrana más el protoplasma.

Tinción del frotis Aplicación de colorante al frotis mediante alguna técnica, con la finalidad de resaltar características particulares en los objetos que se han de observar.

Vacuola	Cavidad rodeada por una membrana que se encuentra en el citoplasma de las células.

Bibliografía

[BKYZ96] Berthod, Marc, Zoltan Kato, Shan Yu y Josiane Zeribia: *Bayesian image classification using Markov random fields*. Image and vision computing, (14), 1996.

[Bou95] Bourke, Paul: *RGB Colour Space*, mayo 1995. `http://local.wasp.uwa.edu.au/~pbourke/texture_colour/`, visitado el 13-03-2007, [Documento electrónico].

[Cou98] Coulter, Beckman: *The Coulter Principle (Electrical Sensing Zone Method)*, 1998. `http://www.beckmancoulter.com/products/applications/partChar/CoulterPrinciple_dcr.asp`, visitado el 22-03-2007, [Documento electrónico].

[dB92] Boomgaard, R. Van den: *Mathematical morphology: extension towards computer vision*. Tesis de Doctorado, Amsterdam University, 1992.

[DeG88] DeGroot, Morris H: *Probabilidad y estadística*. Addison-Wesley Iberoamericana, Argentina, segunda edición, 1988, ISBN 0-201-64405-3.

[DV05] Diez Vegas, Francisco J.: *Introducción al razonamiento aproximado*. Depto. de Inteligencia Artificial (UNED), España, 2005.

[GV01] García Vela, José A.: *Citometría de flujo hematológica*, 2001. `http://www.citometriadeflujo.com/HTML/fundamentos\%20frame.htm`, visitado el 11-06-2007, [Documento electrónico].

[HF97] Heckner, Fritz y Mathias Freund: *Atlas de Hematología*. Marban, novena edición, 1997, ISBN 84-7101-246-4.

[Kat] Kato, Zoltan: *Zoltan Kato Home Page*. `http://www.inf.u-szeged.hu/~kato/`, visitado el 14-06-2007, [Documento Electrónico]; Última modificación:30-03-2007.

[Kat94] Kato, Zoltan: *Modélisations markoviennes multirésolutions en vision par ordinateur. Application à la segmentation d'imagenes SPOT.* Tesis de Doctorado, L'Université de Nice Sophia Antipolis, France, Diciembre 1994.

[KZB92] Kato, Zoltan, Josiane Zerubia y Mark Berthod: *Satellite Image Classification Using a Modified Metropolis Dynamics.* En *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volumen 3, páginas 573–576, San Francisco, California, USA, marzo 1992. IEEE.

[LRM+97] Linch, Matthew J., Stanley S. Raphael, Leslie D. Mellor, Peter D. Spare y Martin J. H. Inwood: *Métodos de Laboratorio.* Interamericana, México, D.F., segunda edición, 1997, ISBN 968-25-0091-5. Reimpresión.

[MR03] Montgomery, Duglas C. y George C. Runger: *Probabilidad y estadística aplicadas a la ingeniería.* Limusa, México, 2003.

[NA02] Nixon, Mark y Alberto Aguado: *Feature Extraction & Image Processing.* Newnes, primera edición, 2002.

[Nil01] Nilsson, Nils J.: *Inteligencia artificial: Una nueva síntesis.* McGraw-Hill, 2001.

[PRG+01] Pagani, A., G. Ramonet, J. P. Graffigna, D. Gomez y A. Naranjo: *Clasificador de leucocitos mediante procesamiento digital de imágenes.* 4to. Simposio Argentino de Informática Y Salud - Sadio, 2001.

[RA01] Ruiz Argüelles, Guillermo J.: *Fundamentos de Hematología.* Panamericana, segunda edición, mayo 2001, ISBN 968-7157-92-5,84-7903-591-9.

[SA02] Serra, Jean y Jesús Angulo: *Aplicación de las morfología matemática a la telemedicina y a la biotecnología: caracterización morfológica de células de la sangre y análisis de cDNA microarrays.* En Díaz de León Santiago, J. L. y C. Yañez Marquez (editores): *Proc. del CIARP 2002 (VII Congreso Iberoamericano en reconocimiento de patrones)*, páginas 35–50, Ciudad de México, México, noviembre 2002.

[SJN00] Stuart J., Rusell y Peter Norving: *Inteligencia Artificial: Un enfoque moderno*. Prentice Hall Interamericana/Pearson Education, México, 2000.

[SJN04] Stuart J., Rusell y Peter Norving: *Inteligencia Artificial: Un enfoque moderno*. Pearson Prentice Hall, España, 2004.

[SZR04] Sabino, D., M. Zago y E. Rizzatti: *Toward leukocyte recognition using morphometry, texture and color*. En *Proc. IEEE Intl. Symp. Biomedical Imaging*, página 121, 2004.

[Wik02] Wikipedia: *Teoría del color*, noviembre 2002. `http://es.wikipedia.org/wiki/Teor%C3%ADa_del_color`, visitado el 13-03-2007, [Documento electrónico; Última modificación: 6-03-2007].