Revista Ingeniería Biomédica
ISSN 1909-9762. Volumen 5, número 9, enero-junio 2011, 26-34
Escuela de Ingeniería de Antioquia-Universidad CES, Medellín, Colombia

Evaluación de técnicas para el análisis de relevancia basadas en filtros sobre imágenes radiológicas

Sandra Milena Roa Martínez[1, Ψ], Humberto Loaiza Correa[2]

[1]*Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca, Popayán, Colombia*
[2]*Facultad de Ingenierías, Universidad del Valle, Cali, Colombia*

Recibido 7 de Abril de 2011, Aceptado 24 de Junio de 2011

Evaluation of techniques for relevance analysis of radiological images using filters

***Resumen*—**Una etapa importante y fundamental en el reconocimiento de patrones sobre imágenes es la determinación del conjunto de características que mejor pueda describir la misma. En este artículo se presenta una etapa adicional entre la caracterización de la imagen y su posterior clasificación o recuperación de imágenes similares a una imagen dada, conocido como análisis de relevancia. Este permite reducir la dimensionalidad del conjunto inicial de características a un nuevo conjunto de menor dimensión que conserva la tasa de acierto de la recuperación. Las imágenes analizadas correspondieron a nódulos pulmonares de placas radiológicas de tórax disponibles en una base de datos de acceso libre disponible a través de la sociedad japonesa de tecnología radiológica.

Se analizaron algoritmos de selección de características basados en filtros que incluyeron los métodos FOCUS, RELIEEF-F y Branch & Bound (B&B). Estos algoritmos fueron modificados e implementados en C++. En el caso de RELIEF-F se logró obtener un ahorro del 34% de características sin afectar la tasa de recuperación cuando se empleaba el 100% de las características originales. Asimismo, el algoritmo implementado presentó un desempeño superior al algoritmo original disponible en la herramienta de código abierto Weka. Asimismo se implementó una estrategia de ponderación de pesos aplicada a las características identificadas cuando se utilizaron los algoritmos RELIEF-F, FOCUS y B&B simultáneamente. Dicha estrategia permitió ponderar cada característica de acuerdo a su participación en los conjuntos mínimos de características relevantes y determinar la consistencia de los mismos. La estrategia de pesos permitió un ahorro del 48% de características necesarias para la recuperación, aunque la tasa de recuperación fue disminuida de 77% a 76%.

***Palabras clave*—** Análisis de relevancia, Extracción de características, Imágenes radiológicas, Reducción de dimensionalidad.

***Abstract*—**An important and fundamental stage in the image pattern recognition is the determination of the characteristics set that best describes the image. This paper describes a further step between the image characterization and its posterior classification or image retrieval similar to a given image, known as relevance analysis. It allows reducing the dimensionality of an initial set of features to a new set with fewer dimensions that preserves the hit rate of the retrieval. The analyzed images corresponded to lung nodules of radiological plaques of thorax, available through the open access library available through the Japanese society of radiological technology.

To achieve these results, characteristic selection algorithms based on different filters such as FOCUS, RELIEEF-F, and BRANCH & BOUND (B&B) were analyzed. In the case of RELIEF-F it was possible to save as much as 34% of the initial characteristics set without affecting the retrieval rate compared to when the 100% of characteristics were used. Further, the

Ψ Dirección de correspondencia: smroa@unicauca.edu.co

implemented algorithm achieved a superior performance to that of the original algorithm included in the validated Weka software. Likewise, a strategy consisting in weights averaging was implemented that was applied to identified characteristics when the algorithms RELIEF-F, FOCUS and B&B were used simultaneously. Such weighting scheme, allowed the averaging of each characteristic according to its contribution in the minimal set of relevant features, allowing to determinate their consistency. The weighting strategy allowed a 48% reduction in the characteristics, although the retrieval hit rate slightly decreased from 77% to 76%.

***Keywords*—** Relevance analysis, Features extraction, Radiological images, Dimensionality reduction

I. Introducción

En años recientes, se ha incorporado una nueva etapa en los sistemas de análisis de imagen previa a la fase de clasificación conocida como reducción de características. Por lo general, en la gran mayoría de los sistemas de imagen, la etapa de clasificación debe manipular un alto número de características que representa a un patrón. En diversas aplicaciones se ha observado que en ocasiones, el utilizar toda la información disponible puede ayudar a realizar una mejor clasificación. Esto, a pesar que pudiera haber redundancia entre diversos parámetros, o bien, cuando estos parecieran no proporcionar información relevante. Sin embargo, esto conlleva a un alto costo en la recolección de los datos, inducir errores por ruido y aumentar significativamente el tiempo de procesamiento tanto para el entrenamiento como para la clasificación. De este modo, utilizar toda la información disponible trae consigo un bajo rendimiento del sistema y el requerimiento de equipos de mayor capacidad de cómputo.

En conjuntos de alta dimensionalidad, no todas las variables medidas son importantes para la comprensión del fenómeno de interés, ya que algunas son redundantes y otras irrelevantes. Las variables redundantes son aquellas que suministran la misma información que otras y que puede ser determinada a partir de otros datos (p.ej. a través de una combinación lineal). De otro lado, las variables irrelevantes son aquellas que no aportan información nueva. En este sentido la reducción de dimensión puede realizarse identificando las variables que no contribuyen a la tarea de clasificación, para eliminarlas, obteniendo un conjunto menor de características de todas las disponibles.

Este artículo analiza métodos para eliminar características, utilizando técnicas de selección de filtro, que no involucran el resultado del clasificador ni etapas de aprendizaje. Así, la selección de características es independiente de la arquitectura del sistema de recuperación de imágenes basadas en contenido, en el cual serán utilizados estos resultados. Una vez revisado este enfoque, se consideró la generación del subconjunto inicial de manera heurística y completa, y con medidas de evaluación de distancia y de consistencia. Para esta etapa, los métodos más utilizados según la literatura consultada fueron los métodos RELIEF-F, FOCUS y Branch & Bound (B&B), los cuales serán explicados con mayor detalle en la siguiente sección [1-4].

En este trabajo se busca presentar las variaciones en la implementación de los métodos de filtro como una alternativa computacionalmente menos costosa, que reduce el conjunto de características obtenidas de las imágenes radiológicas de tórax y garantiza una recuperación fiable. Así, se contribuye en la recuperación de imágenes radiológicas mediante la determinación de características relevantes, sin disminuir la tasa de aciertos.

II. Materiales y Métodos

Se utilizaron imágenes radiológicas de tórax y se estableció sobre ellas como regiones de interés los nódulos pulmonares. Estas regiones de interés se caracterizan principalmente por presentar diferentes texturas (tejidos) y diferentes niveles de grises. La motivación de centrar el trabajo en estas imágenes con nódulos es el alto porcentaje de mortalidad asociada al cáncer de pulmón, y que en la mayoría de casos puede ser diagnosticado por la identificación de un nódulo maligno. A estas imágenes se les hizo preprocesamiento para mejorar su contraste, resolución, etc. También, se les aplicó diferentes tipos de ruido.

2.1. Base de Datos

Se hizo uso de la base de datos llamada "*Standard Digital Image Database for Lung Nodules and Non-Nodules*" [5]. Esta base de datos contiene imágenes radiológicas de tórax con presencia y ausencia de nódulos. Esta fue creada por el comité científico de la sociedad japonesa de tecnología radiológica (*Japanese Society Radiological Technology -JSRT*) desde abril de 1995 hasta marzo de1997, y ha sido usada en diferentes trabajos científicos previamente [6-8].

Las principales especificaciones de las imágenes en esta base de datos incluyen: tamaño de la matriz de 2048 x 2048 pixeles, tamaño del pixel de 0.175 mm, y 4096 (12 bits) niveles de gris.

En cuanto a la información clínica, todas las imágenes fueron examinadas para confirmar la presencia o ausencia de un nódulo. Los nódulos fueron clasificados

Tabla 1. Grados de Dificultad en la determinación de un nódulo pulmonar

ID	Tipo	Descripción
1	Extremadamente Difícil.	La anormalidad es muy indistinta, o muy pequeña, o extremadamente difícil de detectar.
2	Muy difícil.	La anormalidad es muy difícil de detectar.
3	Difícil	La detección es difícil.
4	Relativamente obvia.	
5	Obvia	

como malignos de acuerdo a los resultados de exámenes histológicos y citológicos, y benignos de acuerdo a su histología. Además un grupo de 3 médicos expertos en radiología torácica, consensuó acerca del grado de dificultad para determinar la existencia del nódulo y su tipo. Estos grados se pueden observar en la Tabla 1.

El número de imágenes utilizadas fue de 154, todas con presencia de nódulos, de las cuales 100 casos eran malignos y 54 benignos. De cada imagen se tienen datos clínicos e información acerca del nódulo, como tamaño del nódulo en mm, grado de dificultad, coordenadas *x* y *y* de la localización del nódulo dentro de la imagen, edad, sexo, tipo de nódulo (maligno o benigno), localización anatómica y diagnóstico.

2.2. Extracción de Características

La finalidad de esta etapa es representar a la señal original significativamente y comprimir los datos sin pérdida de información relevante. Esto con el fin de disminuir el número de datos o variables de entrada, y con esta representación garantizar la etapa de clasificación y el rendimiento global del sistema [9].

El espacio de características que representa la información contenida en las imágenes radiológicas de tórax, está conformado por 2 modelos de representación: estadístico (matriz de coocurrencia) y basado en modelos (transformada de Wavelets), debido a su buen desempeño en la caracterización de imágenes con texturas [10-13]. La cantidad de características extraídas con la matriz de coocurrencia fueron 40: 10 características con distancia d=1 y en las cuatro orientaciones mencionadas. Con el método de extracción de la transformada Wavelet y empleando la wavelet de Daubechies, se obtuvieron 45 características: 5 estadísticos (entropía, promedio, desviación estándar, tercer momento y cuarto momento), en las tres orientaciones (horizontal, vertical y diagonal) y con tres niveles de descomposición.

2.3. Análisis de Relevancia

Las técnicas de selección de atributos intentan encontrar un subconjunto óptimo de variables originales con significado físico, extraídos desde el conjunto original que minimicen la pérdida de información y maximicen la reducción de ruido. Es decir, eliminan características menos significativas y dejan un subconjunto de las características originales (las más representativas o que aportan información de mayor calidad), que retienen información suficiente para diferenciar bien entre las clases.

2.3.1. Proceso General de Selección de Atributos

Hay cuatro pasos básicos en cualquier método de selección de características como pueden observarse en la Fig. 1. Estos incluyen: la *generación,* para generar el subconjunto de características candidato; y la *evaluación*, para evaluar el subconjunto de características candidato generado y produce un valor de relevancia, donde el criterio de parada determinará si este es el subconjunto de características óptimo definido.

2.3.2. Algoritmos de Selección de Características

Los dos aspectos fundamentales en el proceso de selección de características son los criterios de selección y búsqueda. Por lo tanto, los algoritmos de *selección* de subconjuntos seleccionados fueron basados en filtros, en los cuales la selección de características no tiene en cuenta el clasificador. El procedimiento de selección de características evalúa los atributos de acuerdo con heurísticas basadas en características generales de los datos de forma independiente de la función de evaluación (método de clasificación) que se utilizará posteriormente (Fig. 2).

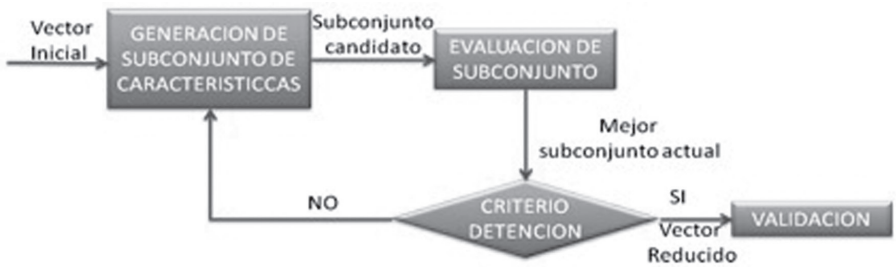

Fig. 1. Proceso de Selección de Características

Fig. 2. Aproximación filtro.

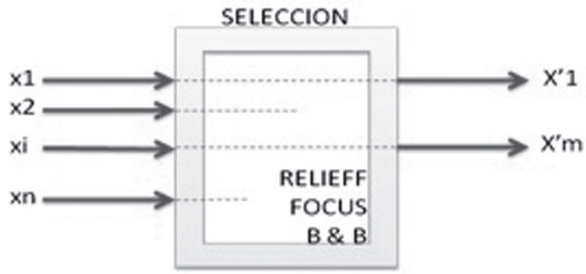

Fig. 3. Selección de Características.

En este trabajo se implementaron tres métodos de selección de características (Fig. 3) como RELIEF-F, FOCUS y B&B. En FOCUS y B&B se involucraron pesos para determinar el aporte de cada característica escogida como relevante.

El porcentaje de reducción de cada uno de los métodos utilizados se calculó con respecto al valor inicial del conjunto inicial completo de 85 características, es decir, que por cada característica que se reduce en el nuevo conjunto de características, la reducción es del 1,17% con respecto al conjunto inicial.

En cuanto a la evaluación del desempeño, se utilizaron métricas de similaridad para determinar la distancia entre los vectores de características y así construir la matriz de confusión que permitió calcular el desempeño de cada algoritmo.

2.3.2.1 RELIEF

El algoritmo RELIEF asume que una característica es fuertemente relevante si esta permite distinguir fácilmente entre dos instancias de diferentes clases, y basándose en esta lógica define el peso para cada característica [14]. Sin embargo, RELIEF en su versión original limita su campo de aplicación a problemas en donde solo tenemos dos clases y por lo tanto para la asignación de los pesos emplea solamente un vecino más cercano de diferente clase, debido a que solo trabaja con una clase opuesta. Por lo anterior, surgió la necesidad de ampliar el campo de aplicación del algoritmo y en 1994 Kononenko expone una nueva versión de RELIEF denominada RELIEF-F [15]. En RELIEF-F se generaliza el comportamiento del algoritmo original para problemas donde se cuenta con más de dos clases. En esta nueva versión del algoritmo, se busca un vecino más cercano por cada clase opuesta. Con los vecinos seleccionados se evalúa la relevancia para cada característica y luego se actualiza su valoración acumulada en un vector de pesos teniendo en cuenta la ecuación (1):

$$W(F) := W(F) - diff(F, E_1, H) + \sum_{k=1}^{n}\left[P(C)\times diff\left(F, E_1, M(C)\right)\right] \quad (1)$$

Siendo E_1 la instancia seleccionada aleatoriamente, H el vecino más cercano (*nearest hit*) de E_1, F la característica a la que se le asigna el peso y la sumatoria representa el acumulado de diferencias del ejemplo E_1 con sus vecinos más cercanos de las diferentes clases en la característica F. La diferencia o distancia entre características es definida como se muestra en la Tabla 2.

Tabla 2. Diferencia entre características en RELIEF-F

	f es discreta	**f es continua**
diff (*f*, *E*1, *E*2) =	0 *si value* (*f*, *E*1) = *value* (*f*, *E*2) 1 *en otro caso*	$\frac{\lvert value\ (f, E_1) - value\ (f, E_2)\rvert}{max\ (f) - min\ (f)}$

El siguiente algoritmo describe a RELIEF-F:

```
RELIEF-F(n){
    W[numCaracteristicas] = 0;
    diffHit[numCaracteristicas] = 0;
    Para i:=0 hasta n
        I = seleccionarInstancia();
        OpsClass = 0;
        H = encontrarNearestHit( I );
        Para C1, C2, C3,……. Ck hacer{
                J = encontrarNearestMiss( I );
                acumularDistancias(J,diffHit);
        fin para
        Para F:=1 hasta numeroCaracteristcas
                W[i] = W[i] - diff(i, I, H) + diffHit [i];
                diffHit [i] = 0;
        fin para
        Para F:=1 hasta numeroCaracteristcas
                W[i] = W[i]/n;
        fin para
    fin para
    Devolver W;
}
```

Una vez se obtiene el vector de pesos entregado por RELIEF-F, para facilitar el análisis de los resultados, se realiza el respectivo ordenamiento de las características con sus respectivos pesos de mayor a menor. La característica con mayor peso representa la característica con mayor relevancia, debido a que si se sigue la filosofía planteada por RELIEF, una característica con mayor peso de relevancia es aquella que tiene valores similares en las instancias de la misma clase y valores diferentes en las clases opuestas. Por la fórmula del peso se obtiene que las diferencias entre las instancias similares tiende a cero, mientras que las diferencias entre instancias de clases diferentes son las que aportarían un valor en el peso de la característica. Por el contrario, si la característica definitivamente no es relevante, entonces para instancias de la misma clase encontraremos valores diferentes y posiblemente para las clases opuestas valores iguales. Esto llevado a la fórmula del peso se reflejaría en un valor negativo, debido a que las diferencias para la característica en clases opuestas tenderían a cero y la sumatoria de estas sería menor que la diferencia con el vecino más cercano de la misma clase, siendo este último valor el que resta.

2.3.2.2 FOCUS

Tal como RELIEF, es un método de filtro. FOCUS empieza con el conjunto vacío y lleva a cabo una búsqueda exhaustiva en anchura hasta encontrar un subconjunto mínimo consistente que prediga las clases puras. Este método rinde mejor cuando el número de atributos relevantes respecto al total es pequeño. FOCUS utiliza como medida de evaluación la de consistencia, que simplemente averigua si el conjunto de datos restringidos a las características seleccionadas es o no consistente, usa esta medida para parar la búsqueda en el primer conjunto de características que la medida evalúe positivamente [16].

Almuallim y Dieterich en [17], que expusieron por primera vez el algoritmo, establecieron como estrategia a seguir por el mismo, algo que ellos denominan "*MIN-FEATURES bias*". Esto consiste en que dado un conjunto de ejemplos que representan una clase, se elige uno de los subconjuntos con el mínimo número de características relevantes para dicha clase.

Finalmente, el total de conjuntos evaluados por el algoritmo está dado por la ecuación (2):

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \ldots + \binom{n}{p} = O(n^p) \qquad (2)$$

A continuación se presenta el algoritmo que describe el funcionamiento de FOCUS:

```
FOCUS(Muestra) {
    Para i: = 0,hasta n hacer:
        Para todo A en {f1,f2,f3.,..,fn} de tamaño i hacer:
            Si (Test-de-suficiencia(A,Muestra)) entonces
                devolver A;
        fin para
    fin para
}
```

Como ya ha sido expuesto anteriormente, FOCUS presenta cualidades interesantes, como es el caso de su óptimo comportamiento en la búsqueda de un conjunto con el menor número de características posibles que describan a una clase y su eficiencia en función al número de características relevantes. Sin embargo, el principal inconveniente es su dominio de aplicación, debido a que es bastante reducido y se limita a ser aplicado a problemas de conceptos lógicos, que son definidos basándose en características booleanas. Por esta razón, encontramos en FOCUS un algoritmo potente que no se puede aplicar en la mayoría de los problemas prácticos que se presentan en el mundo real.

Por esta razón, se han realizado diferentes modificaciones y mejoras al algoritmo enfocadas a extender su campo de aplicación. Por ejemplo, que el algoritmo pueda trabajar con valores discretos, continuos y diferentes clases [18].

La propuesta implementada para que el algoritmo pueda trabajar con valores continuos, consiste en

considerar que dos valores son "significativamente-iguales" cuando su diferencia en valor absoluto sea relativamente menor a un cierto umbral. Este umbral es definido por el experto y puede ser dado un umbral diferente para cada característica, o a través de un proceso de normalización de las características para que sus valores estén dentro del intervalo [0,1], se puede generalizar un solo umbral para todas las características. En general, con las extensiones anteriormente planteadas para FOCUS, su algoritmo no se modifica, solo cambian el concepto del tipo de valores con los que puede trabajar y las definiciones de igual y distinto.

La nueva condición que finaliza la búsqueda de subconjuntos consistentes, es cuando el algoritmo encuentra un subconjunto que permite distinguir los ejemplos de las diferentes clases pero con un número mayor de características al mínimo ya encontrado.

A continuación el algoritmo describe el funcionamiento de FOCUS con lo anteriormente expuesto:

```
FOCUS(Muestra) {
    vectorCoincidencias[n] = ø;
    minCaracteristicas = n;
    Para i: = 0,hasta  n hacer:
        Para todo A en {f1,f2,f3.,..,fn}de tamaño i hacer:
            Si (Test-de-suficiencia (A,Muestra)) entonces
                SI (ContarRelevantes(A)< minCar) entonces
                    minCar = ContarRelevantes(A);
                SI (ContarRelevantes(A)>minCar) entonces
                    devolver vectorCoincidencias/numConjuntosConsistentes;
                    aumentarCoincidencias(vectorCoincidencias,A);
                    numConjuntosConsistentes++;
        fin para
    fin para
}
```

2.3.2.3 Branch and Bound (B&B)

El método B&B consiste en explorar un espacio de búsqueda representado en un árbol y encontrar la mejor solución sin necesidad de evaluar todo el espacio [19]. B&B es una variación del método de búsqueda en profundidad y realiza una búsqueda hacia atrás para determinar un subconjunto con el menor número de características posible, en donde el valor de medida de evaluación es menor que un umbral definido por el experto.

El algoritmo toma como nodo a un subconjunto de características, e inicialmente el subconjunto está formado por todos los atributos. A partir de éste se generan sus hijos con todos los posibles subconjuntos que puedan resultar. El proceso anterior se realiza recursivamente con cada nodo mientras se llega al subconjunto mínimo de características. Si este subconjunto no es el óptimo, el algoritmo se empieza a devolver (debido a la recursividad), y el proceso de búsqueda evalúa los nodos padres, hasta encontrar el conjunto que cumpla que su valor de medida de evaluación sea menor que el umbral establecido. Por el orden seguido, cuando el algoritmo encuentra el primer conjunto que cumple con la condición del umbral, y ya se han evaluado todos sus hijos sin encontrar otro, éste se convierte en el conjunto óptimo y finaliza el algoritmo.

El siguiente algoritmo muestra el funcionamiento principal del método B&B.

```
BOOL B_and_B( TREENODE *Node, Umbral ){
    Si Umbral < medidaEvaluacion (Node)
            Devolver Fail
    Para cada subconjunto de Node hacer
            Hijo = GenerarHijo(Node)
            Si B_and_B(Hijo) == Fail
                    hijosEvaluados ++
    fin Para
    Si hijosEvaluados == ContarHijosPosibles (Node)
            mostrarConjuntoOptimo(Nodo)
            Devolver True
Devolver Fail
}
```

Nuevamente se presenta el problema común en todos los algoritmos de selección de características en donde su punto de parada o de finalización, es en el mejor de los casos cuando encuentran el primer conjunto óptimo de atributos. Esta situación provoca que se descarten características que pertenecen a subconjuntos del espacio de búsqueda, que también sirven como conjuntos óptimos, pero que no son tenidos en cuenta porque el algoritmo finaliza antes de que estos sean evaluados. Para solucionar este problema al igual que en la versión extendida de FOCUS, se decide explorar en todo el espacio de búsqueda y reportar los subconjuntos óptimos que se identifiquen. Una vez explorado todo el espacio se finaliza el algoritmo y se le asignan pesos a las características de acuerdo al número de coincidencias en que aparece cada característica en los subconjuntos identificados como óptimos. Finalmente se normalizan estos valores con el número total de subconjuntos óptimos identificados.

Es por ello, que una nueva heurística se le adiciona al algoritmo para ampliar la poda del árbol, ahora antes de evaluar un subconjunto y generar sus hijos se debe descartar que este ya fue evaluado.

El algoritmo a continuación, describe la nueva extensión de B&B:

```
subconjuntosOptimos;
numConjuntosOptimos = 0;
BOOL B_and_B( TREENODE *Node, Umbral ){
    Si noEvaluado(Node)
        Si Umbral < medidaEvaluacion (Node)
            Devolver Fail
        Para cada subconjunto de Node hacer
            Hijo = GenerarHijo(Node)
            Si B_and_B(Hijo) == Fail
                hijosEvaluados ++
        Fin Para
        Si hijosEvaluados == ContarHijosPosibles (Node)
            AdicionarrConjuntoOptimo(subconjuntos
            Optimos ,Nodo)
            numConjuntosOptimos++;
            Devolver True
Devolver Fail
    Else
        Devolver Fail
}
```

2.3.3 Implementación de algoritmos y pruebas

Las pruebas realizadas tuvieron como objetivo comprobar el funcionamiento de los métodos de análisis de relevancia del enfoque selección de características usando métodos basados en Filtro como FOCUS, RELIEF-F y B&B, implementados en C++ con las variaciones comentadas en la sección 2.3.2. Estos tres métodos se compararon con los disponibles en la herramienta de software de acceso abierto Weka [20, 21]. Todas las pruebas anteriores se contrastaron con los resultados obtenidos con el conjunto inicial completo de características.

III. Resultados y Discusión

Con la información asociada a cada imagen radiológica de tórax, el grupo de patologías es de cinco (5), distribuidas en 72 imágenes, siendo éstas presentadas en la Tabla 3:

El porcentaje de aciertos discriminado por patología y método se presenta en la Tabla 4. El porcentaje de desempeño promedio en la recuperación fue de 77% cuando se emplearon todas las características disponibles.

Cada algoritmo generó un subconjunto de características consideradas relevantes con su respectivo peso. De cada imagen se obtuvo este mismo conjunto de características y se realizó la evaluación del desempeño también mediante el uso de métricas de similaridad. El algoritmo RELIEF-F implementado en este proyecto, seleccionó 56 características, lo que equivale a un porcentaje de reducción del 34,11%. Es notable que el promedio del desempeño fuera del 77%, es decir, igual al presentado por el conjunto inicial completo de características (Tabla 4).

Tabla 3. Patologías

Número Patología	Nombre Patología	Localización	Tipo Nódulo
1	Masa inflamatoria	Indistinto	Benigno
2	Granuloma	Indistinto	Benigno
3	Tuberculoma	Indistinto	Benigno
4	Cáncer Pulmonar Adenocarcinoma	Indistinto	Maligno
5	Cáncer Pulmonar carcinoma broquio-alveolar	Indistinto	Maligno

Tabla 4. Comparación del porcentaje de acierto y desempeño los métodos evaluados. NA = no aplica.

PORCENTAJE DE ACIERTOS (%)					
Patología	Conjunto original	Algoritmos implementados		Algoritmos de Weka	
		RELIEF-F	Mezcla de pesos*	RELIEF-F	CfsSubset+ FOCUS
Masa inflamatoria	80	80	80	20	0
Granuloma	75	75	75	50	25
Tuberculoma	50	50	50	50	75
Cáncer Pulmonar-Adenocarcinoma	100	100	100	50	50
Cáncer Pulmonar-carcinoma broncoalveolar	80	80	75	60	25
Número de características	85	56	44	47	14
Porcentaje Reducción (%)	NA	34	48	44	84
Desempeño (%)	77	77	76	46	35

* Ponderación de características según el peso obtenido en los tres algoritmos seleccionados (RELIEF-F, FOCUS y B&B)

Cuando se evaluaron diferentes combinaciones de métodos de análisis de relevancia como el CfsSubsetEval+FOCUS [20] y RELIEF-F de Weka, se encontró que sus desempeños fueron de 35 y 46%, es decir que disminuyeron considerablemente, si bien el número de características seleccionadas fue de 14 y 37 respectivamente (Tabla 4). También, se encontró que el conjunto de características obtenidas de la combinación de CfsSubsetEval+FOCUS mejora el desempeño para la determinación de tuberculomas por encima del conjunto inicial de características. El desempeño de las coincidencias en los métodos se disminuye considerablemente a diferencia del desempeño cuando se consideran los pesos arrojados por cada algoritmo, teniendo en cuenta solo las características cuyo valor del peso de la característica está por encima del promedio de pesos. Con la combinación de pesos propuesta en este trabajo puede verse que el porcentaje de acierto se mantiene en un 76% y utilizando solo 44 características.

Las características seleccionadas por el método de B&B coincidieron con el conjunto inicial completo de las 85 características, y por esta razón no fue considerado para la comparación. En la ejecución de este algoritmo y cuando se realizaron pruebas con un mayor número de patologías se encontraba que a medida que aumentaban estas, se disminuía el conjunto de características que generaba este algoritmo. Al contrario el conjunto generado por el algoritmo FOCUS se aumentaba a medida que se disminuía el número de patologías, en este caso solo genero un conjunto de 3 características.

Las pruebas realizadas permitieron evaluar el comportamiento de los métodos seleccionados para reducir el conjunto inicial de características construido por descriptores de la matriz de coocurrencia y la transformada de wavelets, de las cuales se concluye que el método que presentó mejores resultados fue el implementado en este proyecto, es decir, RELIEF-F. Este método permitió reducir el conjunto de características en un 34,11%, permitiendo mantener el porcentaje de acierto en un 77%, igual al inicial cuando se evaluó una base de datos correspondiente a imágenes de nódulos pulmonares en radiografías de tórax.

IV. Conclusión

Las técnicas de análisis de relevancia permiten determinar el mejor conjunto de características disminuyendo información redundante o irrelevante que conlleva a ruido y mayores tiempos de procesamiento, y en algunas ocasiones erróneo desempeño en la recuperación o bajo porcentaje de acierto.

En cuanto al desarrollo de técnicas de análisis de relevancia, se implementaron extensiones de los algoritmos RELIEF-F, FOCUS y B&B. Dichas extensiones tienen en cuenta todas las características relevantes, pero asigna a cada una pesos que dependen del número de veces que fueron encontradas como relevantes por el algoritmo, y donde la condición que finaliza la búsqueda de subconjuntos cambia con respecto a este número de veces en conjuntos mínimos de características.

La extensión del algoritmo RELIEF-F implementado en este proyecto obtuvo mejor poder discriminante de características relevantes, permitiendo mantener el mismo desempeño en la recuperación con respecto al conjunto inicial de características, y superior al RELIEF-F de Weka.

El comportamiento del algoritmo FOCUS y B&B cuando se disminuyeron las patologías no permitió determinar satisfactoriamente un conjunto relevante de características, dado que en el caso de FOCUS, el nuevo conjunto disminuyó sin permitir determinar la patología a la cual correspondía la imagen. En cuanto a B&B, el conjunto, contrariamente, aumentó incluyendo todas las características del conjunto inicial, el cual sí permite mantener el desempeño en la recuperación pero no disminuir el vector de características. Estos algoritmos modificados se probaron durante su desarrollo permitiendo evidenciar si tenían alguna ventaja frente a los algoritmos tradicionales tras la inclusión de pesos asociados a las características. Dichos pesos permitieron establecer un *ranking* de aporte de cada característica. Así, cuando se combinaron los resultados de las características promedio de cada algoritmo, y se consideraron las que estaban por encima del promedio, el resultado fue un 76% de acierto y un porcentaje de reducción de 48,23%.

Agradecimiento

Los autores agradecen al Grupo de Investigación en Percepción y Sistemas Inteligentes – PSI, al Programa de Maestría en Ingeniería con énfasis Electrónica de la Universidad del Valle y a la Universidad del Cauca por el apoyo para la ejecución de este proyecto.

Referencias

[1]. Talavera, L. An evaluation of filter and wrapper methods for feature selection in categorical clustering. Dept. Idiomas y Sistemas Informáticos Universidad Politécnica de Catalunya. Barcelona, 2006.

[2]. Kira, K., Rendell, L. A practical approach to feature selection. University of Illinois at Urbana Champaign: Computer & Information Systems Laboratory. 2003.

[3]. Ravisekar, B. A Comparative Analysis of Dimensionality Reduction Techniques. Georgia: College of Computing Georgia Institute of Technology. 2006

[4]. Sánchez, L., Martínez, F., Castellanos, G., Salazar, A. Feature Extraction of Weighted Data for Implicit Variable Selection. *Computer Analysis of Images and Patterns*, 4673, 840-847. 2007.

[5]. Japanese Society Radiological Technology. Consultado el 15 de mayo de 2010 en: http://www.jsrt.or.jp/web_data/english03.php

[6]. Cox, G., Jager, H. Experiments in lung cancer nodule detection using texture analysis and neural network classifiers. South Africa: Department of Electrical Engineering. University of Cape Town. 2007.

[7]. Li, Q., Katsuragawa,S., Doi, K. Computer-aided diagnostic scheme for lung nodule detection in digital chest radiographs by use of a multiple-template matching technique. *Medical Physics* 28 (10), 2070-2076, 2001.

[8]. Kubota, H. Tai, Y., Katagiri, M. Wavelet denoising — threshold selection by the histogram shape of wavelet coefficients. Book Title: World Congress on Medical Physics and Biomedical Engineering, 2006.

[9]. Bishop C. Neural Networks for Pattern Recognition, Oxford: University Press, 1995.

[10]. González R. W. Tratamiento Digital de Imágenes, Ed. Adison Wesley Días de Santos, 1996.

[11]. Cova W., Caballero R., Centro Universitario de Desarrollos en Automatización y Robótica. Sobre Wavelets e Imágenes (2006). Consultado el 30 de junio de 2010 en: http://www.edutecne.utn.edu.ar/DOCUMENTOS/Sobre%20Wavelets%20e%20Imagenes_R1.pdf

[12]. Jiménez G. M. Extracción de características de textura basada en Transformada Wavelet Discreta. Tesis de Grado, Universidad de Sevilla, Sevilla, España, 2008.

[13]. Lester, M. Introducción a la Transformada Wavelet. Descomposición de Señales. Apuntes de Clase, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina. 2006

[14]. Yijun, S., Dapeng, K. A RELIEF Based Feature Extraction Algorithm. *Proceedings of the SIAM International Conference on Data Mining, SDM*, 188-195. Atlanta, Georgia, 2008.

[15]. Kononenkoi I., Estimating attributes: Analysis and extensions of RELIEF. *European Conference on Machine Learning*,171–182, Vienna, Austria, 1994

[16]. Arauzo A., Benitez J., Castro J., A feature selection algorithm with Fuzzy information. Scientific Literature Digital Library and Search Engine, 2009.

[17]. Almuallin T., Dietterich T., Learning Boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 279-305, 1994.

[18]. Arauzo A., Un Sistema Inteligente para Selección de Características en Clasificación, Universidad de Granada, España, 2006.

[19]. Narendra P., Fukunaga K. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computer* ,9, 917-922, 1977.

[20]. Hall M., Correlation-based Feature Subset Selection for Machine Learning, New Zealand.: Hamilton, 1998.

[21]. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. The Weka Data Mining Software: An Update. *SIGKDD Explorations*, 11 (1), 2009.