EIGENFUNGI: Desarrollo de un Método de Data Mining para la Detección Automática de Patrones en Microscopía Aplicada a Micología Médica

Marcela L. Riccillo
Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina Tel/Fax (011) 4576-3359,
marcela.lr@gmail.com

Marcelo Soria
Facultad de Agronomía, Universidad de Buenos Aires, Argentina,
soria@agro.uba.ar

Oscar Bustos
Facultad de Matemática, Astronomía y Física, Universidad de Córdoba, Argentina,
oscar.oh@gmail.com

Abstract

En este trabajo desarrollamos un método automático para el reconocimiento de especies de hongos microscópicos, que denominamos eigenfungi. Está basado en la metodología para reconocimiento de rostros denominada eigenfaces, a la que se le introducen varias modificaciones que mejoran su exactitud en el análisis de imágenes microscópicas de hongos.

En los últimos años se registra un incremento en las infecciones causadas por hongos. Debido a la necesidad de entrenamiento específico que requiere el análisis microscópico, el diseño e implementación de herramientas informáticas que asistan al personal recibe creciente atención.

Este método transforma las imágenes y aplica técnicas propias de Data Mining, considerando al conjunto de imágenes como una base de datos. No necesita de recortes manuales de los objetos por parte del experto humano y requiere de pocas imágenes para el entrenamiento.

Key words: Eigenfungi, Eigenfaces, Hongos microscópicos, Análisis de Componentes Principales.

1. Introducción

La micología médica, el estudio de los hongos que causan enfermedades, es una de las disciplinas donde el entrenamiento del personal requiere especial importancia. En muchas patogenias la única forma de identificar el agente causante de la enfermedad es mediante análisis microscópico; y a su vez, sólo mediante la correcta identificación del hongo responsable de la infección, el médico es capaz de indicar un tratamiento adecuado, dado que las micosis podrían producir daños irreversibles o hasta llevar a la muerte del paciente. Por otra parte, en los últimos años se registra un incremento en las infecciones causadas por hongos, principalmente debido a causas que comprometen el funcionamiento normal del sistema inmune de los pacientes, como la desnutrición, la epidemia de SIDA o la inmuno-supresión que sigue a los transplantes de órganos.

En este trabajo desarrollamos un método automático para el reconocimiento de especies de hongos microscópicos, que denominamos *eigenfungi*. Está basado en la metodología para reconocimiento de rostros denominada *eigenfaces*.

Existen varios ejemplos de aplicaciones del procesamiento de imágenes para microbiología general, pero no tantos para micología médica. Uno de los motivos es que las imágenes micológicas tienen una complejidad mucho mayor que aquéllas que contienen exclusivamente bacterias, que son morfológicamente menos complejas.

Existen diversas técnicas para la identificación de rostros humanos, las cuales podrían clasificarse en dos grandes grupos: la identificación por características y las aproximaciones estadísticas. En 1991, M. Turk y A. Pentland [1] presentan un método de reconocimiento basado en el Análisis de Componentes Principales al que denominaron *Eigenfaces*. En 1997, Belhumeur y otros [2] presentan las Fisherfaces que se basan en el método estadístico Análisis Discriminante Lineal de Fisher. Posteriormente, fueron desarrolladas otras técnicas basadas en variaciones de estos métodos, como por ejemplo Independent Component Analysis o ICA de Bartlett y otros [3], que proyectan los datos sobre vectores básicos estadísticamente independientes. Mixture of Principal Component de Deepak y otros [4], que usa una mezcla de eigen-espacios para capturar variaciones en los datos.

En el caso de reconocimiento de microorganismos, vemos algunas implementaciones de redes neuronales como el trabajo de Widmer y otros [5] que entrenan un perceptrón para el reconocimiento del Cryptosporidium parvum. Y por ejemplo Verpoulos y otros [6] utilizan una red neuronal para la identificación de bacilos de tuberculosis.

En el campo de la micología, los avances en desarrollos automáticos para la identificación automática son casi inexistentes.

1.1. Dermatofitos

Los dermatofitos son hongos queratinofílicos que causan infecciones de los tejidos epidérmicos humanos y animales. Se encuentran distribuidos taxonómicamente en tres géneros: Microsporum, Trichophyton y Epidermophyton. Debido a las similitudes existentes entre las diferentes especies, es posible ver que un tipo clínico de infección puede ser causado por diferentes dermatofitos, o que una misma especie esté involucrada en varios tipos de enfermedades.

Las dermatofitosis o tiñas pueden ser desde asintomáticas a muy pruriginosas y dolorosas. Se diseminan por contacto directo o indirecto interhumano o animal-hombre. Las 6 especies principales de dermatofitos son: *Epidermophyton floccosum - Microsporum canis - Microsporum gypseum - Trichophyton mentagrophytes - Trichophyton rubrum - Trichophyton tonsurans.*

1.2. Análisis de *Eigenfaces*

El método de *eigenfaces* presentado por Turk y Pentland [1] es utilizado para el reconocimiento de rostros de personas. Se basa en el método de Análisis de Componentes Principales (PCA) que descompone datos multidimensionales a un subespacio de menor dimensión pero preservando las características esenciales de los datos tratados.

Se toma una muestra de fotos de los individuos que se quieren reconocer (por ejemplo las personas autorizadas en una empresa) y se arma un conjunto de nuevas imágenes denominadas *eigenfaces*. Éstas contienen la información principal de las imágenes originales.

Luego se obtiene la distancia de cada foto a las *eigenfaces*. Se agrupan las fotos por individuo y se calcula la distancia promedio del grupo. Al intentar reconocer a una persona, se le saca una foto y se calcula la distancia de ésta a las *eigenfaces.* Finalmente se compara esta distancia con la de cada grupo, siendo la mínima la que identifica la persona analizada.

2. Objetivos del trabajo

Las imágenes microscópicas de los dermatofitos tienen características diferentes a las imágenes de rostros: (1) en el caso de rostros hay un único objeto a reconocer (la cara); en los hongos hay varios (conidias, hifas). (2) En los rostros, los objetos de fondo son eliminados; en los hongos todos los objetos de la imagen son importantes. (3) Los rostros pueden ser normalizados de modo de homogeneizar el tamaño de las cabezas, posición de los ojos, etc.; los hongos, no.

Esto haría pensar que el método de las *eigenfaces* sería incompatible a la hora de identificar hongos microscópicos. Sin embargo, con unas modificaciones que permitan adaptar el método a este tipo de imágenes, la exactitud de la clasificación es muy buena, requiriéndose conjuntos de pocas imágenes para el entrenamiento.

3. Cálculo de *Eigenfungi*

Una vez obtenido el conjunto de imágenes se procede como sigue:

1. Se calcula la imagen media del conjunto como $$\psi = \frac{1}{M}\sum_{i=1}^{M}\Gamma_n \quad (1)$$

2. Luego se resta la imagen media a cada imagen del conjunto de entrenamiento $\phi_i = \Gamma_i - \psi$ (2)

3. Se arma una matriz *A* con las imágenes resultantes $A = \{\Phi_1, \Phi_2, ..., \Phi_M\}$ (3)

4. A partir de *A* se calcula la matriz *C* de covarianzas $C = \frac{1}{M}\sum_{i=1}^{M} \phi_n \phi_n^t = AA^t$ (4)

5. Se calculan los autovalores y autovectores *v* de *C*.

6. A partir de los autovectores encontrados y las imágenes (menos la imagen media), se calculan los *eigenfungi U* $U_i = \sum_{i=1}^{M} v_{ik}\phi_k$ (5) con i = 1, …, M

7. Posteriormente se halla la distancia de cada imagen original a cada *eigenfungi* y con eso se arma un vector de distancias para cada imagen

La diferencia principal aparece al momento de comparar las distancias de las imágenes: en lugar de comparar con el vector de clase de cada especie, se compara con cada imagen del conjunto de entrenamiento. Esto da mejores resultados, debido a que hay especies muy similares y existen detalles en las micro o macronidias, tales como tabiques internos que son difíciles de distinguir si se utiliza el promedio.

8. Cuando se tiene una nueva imagen, se le resta la media y se halla su vector de distancias

9. Finalmente, se compara el nuevo vector de distancias con el vector de distancias de cada imagen original.

4. Imágenes utilizadas

Para la elaboración y validación de la metodología, se estudiaron imágenes de hongos microscópicos de las seis especies principales de dermatofitos, obtenidas de muestras provistas por el Departamento de Micología del Instituto Nacional de Enfermedades Infecciosas (INEI), ANLIS "Carlos G. Malbrán". Fueron tomadas con un aumento de 400x y originalmente medían 1600x1200 píxeles. Luego de varias pruebas, se determinó que el tamaño de las imágenes no influía en los resultados por lo que se decidió disminuirlas a 160x120 píxeles. Se utilizaron 6 imágenes de entrenamiento y 6 imágenes de prueba por cada especie por cada muestra. Haciendo un total de 36 imágenes de entrenamiento y 36 imágenes de prueba por cada muestra.

5. Experimentos

Se realizaron dos tipos de pruebas:

Pruebas binarias - Se dispusieron las especies de a pares, entrenando y reconociendo dos cada vez. Por ejemplo, *E. floccosum* versus *M. canis*

Pruebas totales - Se entrenó y testeó con todas las especies a la vez. Por ejemplo, se intenta que el sistema reconozca a cuál de las 6 especies pertenece una imagen.

5.1. Pruebas realizadas

Probando con las *eigenfaces*, los porcentajes de acierto fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		91,67%	83,33%	75%	100%	91,67%
canis			100%	100%	100%	100%
gypseum				100%	50%	91,67%
mentagro					83,33%	75%
rubrum						83,33%
tonsurans						
Totales	80,56%	*eigenfaces*				

Tabla 1. Porcentajes de acierto *eigenfaces* con dermatofitos

Cada celda de la tabla representa el porcentaje de acierto de identificación de cada par de especies. Por ejemplo, se obtuvieron varios aciertos del 100% de reconocimiento de las imágenes, por ejemplo: *E. floccosum vs T. rubrum - M. canis* con el resto de las especies (con *E. floccosum* 91,67%) - *M. gypseum vs T. mentagrophytes.*

En general los porcentajes de acierto fueron bastantes altos. En el caso de *M. gypseum vs T. rubrum*, sin embargo, no hubo reconocimiento entre las especies, dado que el porcentaje fue del 50%. Si se ingresan todas las especies a la vez, el porcentaje es bastante alto, del 80,56%.

Al aplicar *eigenfungi*, vemos que los porcentajes de acierto en general se incrementan:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		100%	83,33%	100%	83,33%	50%
canis			100%	100%	100%	91,67%
gypseum				91,67%	100%	100%
mentagro					91,67%	91,67%
rubrum						91,67%
tonsurans						
Totales	80,56%	*eigenfungi*				

Tabla 2. Porcentajes de acierto *eigenfungi* con dermatofitos

Los porcentajes de acierto se incrementaron con respecto a los resultados anteriores en 7 de los pares. Hubo una pequeña reducción en el caso de *E. floccosum vs T. rubrum* de 100% a 83,33%, pero el porcentaje igualmente siguió siendo alto.

Si bien ahora se reconoce el par *M. gypseum vs T. rubrum* (y con un 100%), se detecta que hay un par no reconocido: *E. floccosum vs T. tonsurans* con un 50% de acierto (que sí se reconocía en el caso de las *eigenfaces*). El porcentaje total siguió siendo alto (de un 80,56%) aunque no se modificó entre la aplicación de cada método.

5.2. Aplicación de preprocesamientos

Posteriormente a estas pruebas, la idea fue buscar un preprocesamiento, que combinado con el método de *eigenfungi*, incrementara los porcentajes de acierto y además permitiera el reconocimiento de todos los pares de especies. Para transformar las imágenes fue utilizado el programa ImageJ [7]. Los preprocesamientos estudiados fueron los siguientes: Detección de contornos - Imágenes binarias - Corrección de histograma - Suavizado de bordes - Transformada de Fourier - Desenfoque Gaussiano.

Finalmente, el preprocesamiento que combinado con el método de *eigenfungi* fue el que mejor resultados produjo, fue el suavizado de bordes con una posterior corrección del histograma. Los porcentajes obtenidos fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		91,67%	91,67%	100%	100%	83,33%
canis			100%	100%	100%	91,67%
gypseum				100%	100%	91,67%
mentagro					100%	100%
rubrum						100%
tonsurans						
Totales	86,11%	*eigenfungi*	*suavizado*	*histograma*		

Tabla 3. Porcentajes de acierto *eigenfungi* combinado con suavizado de bordes y corrección de histograma

Comparando con el método *eigenfungi* puro, los porcentajes de acierto se incrementaron y todos los pares de especies fueron reconocidos. Se obtuvo prácticamente un 100% de acierto en todas las pruebas (solamente un poco menor en el caso de *E. floccosum vs T. tonsurans* con un 83,33%). También subió el porcentaje a nivel total, de un 80,56% a un 86,11%.

7.3. Pruebas de Robustez

Para verificar la robustez del método, se degradaron las imágenes con dos tipos de ruido Ruido Gaussiano y Ruido Sal y Pimienta, y luego se les aplicó el método de reconocimiento *eigenfungi* combinado con el preprocesamiento de suavizado de bordes y corrección de histograma. Estas

pruebas son importantes porque son una "simulación"del tipo de imágenes que se pueden llegar a obtener en el trabajo de rutina de un laboratorio, debido a la calidad de los especímenes o los preparados y a distorciones introducidas por la óptica del microscopio

Se observó que a pesar de haber degradado las imágenes con el ruido, igualmente los porcentajes fueron muy altos. Todos los pares fueron reconocidos y casi el 70% de las pruebas dio un porcentaje de acierto del 100%. También fue alto el porcentaje de las pruebas totales, con un porcentaje de acierto del 88,89%.

9. Conclusiones

En este trabajo desarrollamos un método automático para el reconocimiento de especies de hongos microscópicos, que llamamos *eigenfungi*. Está basado en la metodología para reconocimiento de rostros denominado *eigenfaces*, y no necesita de recortes manuales de las imágenes.

La base matemática se sustenta en el Análisis de Componentes Principales, que es un método estadístico de análisis que descompone datos multidimensionales a un subespacio de menor dimensión pero preservando las características esenciales de los datos tratados.

Las pruebas dieron resultados con porcentajes altos y se mejoraron con la combinación del método con preprocesamientos. Con el tratamiento de las imágenes de un suavizado de bordes con corrección de histograma, se obtuvo casi un 100% de porcentaje de acierto. Se repitieron estas pruebas con imágenes degradadas por ruido y también se obtuvieron porcentajes altos, observándose la robutez del método.

10. References

[1] M. Turk, and A. Pentland, “Eigenfaces for recognition” *Journal of Cognitive Neuroscience* 3 (1): 71–86 –1991

[2] Peter N. Belhumeur, Joao P. Hespanha, David J. Kriegman – “Eigenfaces vs. Fisherfaces: Rcognition Using Class Specific Linear Projection” – *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, nº 7, pp. 711-720* – Julio 1997

[3] Marian Stewart Bartlett, Terrence J. Sejnowski – "Independent component representations for face recognition" – *Proceedings of the SPIE: Conference on Human Vision and Electronic Imaging III, vol. 3299, pp. 528-539* – 1998

[4] Deepak S. Turaga, T. Chen – “Face recognition using mixtures of principal components” – *IEEE ICIP, Rochester* – Setiembre 2002

[5] Kenneth W. Widmer, Kevin H. Oshima, Suresh D. Pillai – “Identification of Cryptosporidium parvum Oocysts by an Artificial Neural Network Approach” – *American Society for Microbiology Appl Environ Microbiol. 68 (3): 1115-1121* – Marzo 2002

[6] K. Verpoulos, C. Campbell, G. Learmonth, B. Knight, J. Simpson-“The Automated Identification of Tubercle Bacilli using Image Processing and Neural Computing Techniques”- *Proceeding of the 8th International Conf. on Artificial Neural Networks, vol2, pp 797-802- 1998*

[7] “ImageJ - Image Processing and Analysis in Java” – *NIH, USA* http://rsb.info.nih.gov/ij/

[8] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips – “Face Recognition: A Literature Survey” – *ACM Computing Surveys, pp. 399-458* – 2003

[13] Dr. J.J. Vilata Corell – *Micosis Cutáneas* – Editorial Médica Panamericana, España 2006

[9] Marcela L. Riccillo, Ana S. Haedo, Natalia Debandi, Daniel Vazquez V. – “Comparación de Softwares Estadísticos” – *CLATSE VI Congreso Latinoamericano de Sociedades de Estadística – SAE Sociedad Argentina de Estadística, SOCHE Sociedad Chilena de Estadística Concepción, Chile* – Noviembre 2004

[10] J. Liu, F.B. Dazzo, O. Glagoleva, B. Yu, A.K. Jain – “CMEIAS: A Computer-Aided System for the Image Analysis of Bacterial Morphotypes in Microbial Communities” – *Microbial Ecology – Febrero 2001*