Eigenfungi - Método de Data Mining para Detección Automática de Patrones en Micología Médica con Diferenciación de Muestras

Marcela L. Riccillo[1], Marcelo Soria[2], and Ana S. Haedo[3]

[1] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina
marcela.lr@gmail.com
[2] Facultad de Agronomía, Universidad de Buenos Aires, Argentina
soria@agro.uba.ar
[3] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina
haedo@qb.fcen.uba.ar

Resumen Eigenfungi es un método automático que desarrollamos para el reconocimiento de especies de hongos microscópicos (dermatofitos). Está basado en la metodología para reconocimiento de rostros denominada eigenfaces, a la que introducimos varias modificaciones que mejoran su exactitud en el análisis de estas imágenes. Este método aplica técnicas propias de Data Mining, no necesita de recortes manuales de los objetos por parte del experto humano y requiere pocas imágenes para el entrenamiento. En los hongos microscópicos el reconocimiento es a nivel de especies, por lo que podrían existir diferencias si las imágenes fueron tomadas de diferentes muestras. Por eso, en este trabajo realizamos pruebas con dos muestras de cada especie para verificar si el método podía aprender más información al agregar imágenes que aunque pertenecen a la misma especie, tienen ciertas diferencias que podrían confundir al sistema. También comparamos el método con variantes en el cálculo, como multiespacios y distancia Manhattan.
Key words: Eigenfungi, Eigenfaces, Dermatofitos, Pattern Recognition, Análisis de Componentes Principales.

1. Introducción

La micología médica, 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 los antifúngicos disponibles tienen efectos diferentes dependiendo de la especie. Esto adquiere especial importancia en aquellas micosis de evolución rápida y que pueden llevar a la muerte del paciente.

Además, 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.

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.

Para la identificación automática de los hongos microscópicos, desarrollamos el método eigenfungi [1]. El mismo está basado en un método específico utilizado para el reconocimiento de rostros denominado eigenfaces.

Existen diversas técnicas para la identificación de rostros humanos, las cuales podrían clasificarse en dos grandes grupos:

- la identificación por características - toman en cuenta en las imágenes modelos colorímetros y proporciones geométricas de la disposición de los componentes del rostro, curvaturas de huesos, etc.
- las aproximaciones estadísticas - por ejemplo las eigenfaces, representan las imágenes como bases numéricas y detectan patrones mediante la aplicación de métodos multivariados

En 1991, M. Turk y A. Pentland [2] presentan un método de reconocimiento basado en el Análisis de Componentes Principales al que denominaron Eigenfaces. En 1997, Belhumeur y otros [3] 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 [4], que proyectan los datos sobre vectores básicos estadísticamente independientes. Mixture of Principal Component de Deepak y otros [5], que usa una mezcla de eigen-espacios para capturar variaciones en los datos.

También encontramos otras variantes como PCA 2-dimensional de Yang y otros [6] que, en vez de usar vectores como PCA, utiliza matrices 2D así la matriz de imágenes no debe ser transformada en un vector para la extracción de características. Otro ejemplo, PCA Diagonal de Zhang y otros [7], que busca los vectores de proyección óptimos desde imágenes diagonalizadas, sin transformar la imagen a un vector.

En el caso de reconocimiento de microorganismos, vemos algunas implementaciones de redes neuronales como el trabajo de Widmer y otros [8] que entrenan un perceptrón para el reconocimiento del *Cryptosporidium parvum*. Y por ejemplo Verpoulos y otros [9] 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. En el trabajo de Dorge y otros [10] se muestra un método para el reconocimiento del *Penicillium* pero a nivel de imágenes macroscópicas de colonias. Inglis y otros [11] muestran un método semiautomático para identificar hifas en muestras microscópicas, que requiere de un preprocesamiento por parte del experto humano para una demarcación previa de las imágenes.

En las siguientes secciones veremos las características de los hongos estudiados, cómo se calculan los eigenfungi y los resultados de los experimentos con

una y dos muestras. Finalmente veremos la comparación de los resultados con dos variantes multiespacios y distancia Manhattan.

2. Micosis

Las micosis superficiales, ampliamente distribuidas en el mundo, son afecciones producidas por el parasitismo fúngico en las estructuras córneas de la piel y sus faneras (pelos y uñas). Se excluye de estas infecciones aquéllas en donde haya compromiso de mucosas y de tejidos blandos, que involucran más allá de la dermis.

Los dermatofitos 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 6 especies principales de dermatofitos son:

- *Epidermophyton floccosum*
- *Microsporum canis*
- *Microsporum gypseum*
- *Trichophyton mentagrophytes*
- *Trichophyton rubrum*
- *Trichophyton tonsurans*

2.1. Diferencias entre rostros y hongos

Las imágenes microscópicas de los dermatofitos tienen características diferentes a las imágenes de rostros. En la Tabla 1 se puede ver una comparación entre ellos.

Tabla 1. Diferencias entre Rostros y Hongos

Tópico	Rostros	Hongos
Cantidad de objetos a reconocer	Un único objeto (la cara)	Varios objetos (conidias, hifas)
Importancia de objetos de fondo	Objetos de fondo eliminados	Todos objetos son importantes
Normalización de objetos	Objetos pueden ser normalizados	Objetos no pueden ser normalizados

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

Básicamente el método eigenfungi consta de los siguientes pasos:
Entrenamiento

- obtención del conjunto de imágenes de las especies a identificar
- cálculo de la imagen media
- cálculo de la matriz de covarianza de las imágenes
- cálculo de los eigenfungi
- obtención de la distancia (o peso) de cada imagen original a cada eigenfungi

Utilización

- cálculo de la distancia de cada nueva imagen a cada eigenfungi
- comparación de las distancias obtenidas respecto de las distancias de las imágenes originales, para hallar la más similar e identificar la especie

El método de eigenfungi se basa en el cálculo de las eigenfaces. La diferencia principal con éste aparece al momento de comparar las distancias de las imágenes: en lugar de hallar el vector de clase de cada especie y comparar las imágenes con este vector, se comparan 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.

3.1. Desarrollo del método

Una vez obtenido el conjunto de imágenes, se procede como sigue para calcular los eigenfungi.

1. Se calcula la imagen media del conjunto como

$$\psi = \frac{1}{M} \sum_{i=1}^{M} \Gamma_n \tag{1}$$

2. Luego se resta la imagen media a cada imagen del conjunto de entrenamiento

$$\phi_i = \Gamma_i - \psi \tag{2}$$

3. Se arma una matriz A con las imágenes resultantes

$$A = \phi_1, \phi_2, ..., \phi_M \tag{3}$$

4. A partir de A se calcula la matriz C de covarianzas

$$C = \frac{1}{M} \sum_{i=1}^{M} \phi_n \phi_n^t = AA^t \tag{4}$$

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

$$U_i = \sum_{i=1}^{M} v_{ik} \phi_k \tag{5}$$

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

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.

Para las pruebas se utilizaron 6 imágenes de entrenamiento y 6 imágenes de prueba por cada especie por cada muestra, haciendo un total de 72 imágenes de entrenamiento y 72 imágenes de prueba.

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 con una muestra

Se organizaron todas las especies por pares y se realizaron pruebas tanto con las eigenfaces como los eigenfungi. Luego se ingresaron todas las especies a la vez y se analizaron los porcentajes de acierto totales.

Con los eigenfungi, vemos en la Tabla 2 que los porcentajes de acierto en general son mayores al 90 % (incluyendo 7 casos de un 100 %).

Tabla 2. Porcentajes de acierto eigenfungi con dermatofitos

Bin	flocc	canis	gyps	menta	rubrum	tons
flocc		100	83.33	100	83.33	50
canis			100	100	100	91.67
gyps				91,67	100	100
menta					91.67	91.67
rubrum						91.67

El porcentaje de aciertos totales se registró en un 80.56 %. Sin embargo, vemos que el par *E. floccosum vs T. tonsurans* no es reconocido. 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 [12].

Los preprocesamientos estudiados fueron los siguientes: Detección de contornos - Imágenes binarias - Corrección de histograma - Suavizado de bordes - Transformada de Fourier - Desenfoque Gaussiano [17].

Finalmente, el preprocesamiento que combinado con el método de eigenfungi que mejor resultados produjo, fue el suavizado de bordes con posterior corrección del histograma. Los porcentajes obtenidos pueden verse en la Tabla 3.

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

Bin	flocc	canis	gyps	menta	rubrum	tons
flocc		91.67	91.67	100	100	83.33
canis			100	100	100	91.67
gyps				100	100	91.67
menta					100	100
rubrum						100

Se obtuvo prácticamente un 100 % de acierto en todas las pruebas y el par *E. floccosum vs T. tonsurans* fue reconocido con un 83.33 %. También el porcentaje a nivel total se incrementó a 86.11 %.

5.2. Pruebas realizadas con dos muestras

En el caso de los hongos microscópicos el reconocimiento es a nivel de especies, por lo que podrían existir diferencias entre las imágenes si fueron tomadas a partir de diferentes muestras.

Es por esto que se realizaron pruebas con dos muestras diferentes de cada especie, duplicando la cantidad de imágenes tanto de entrenamiento como de testeo. Es decir, que considerando 6 imágenes de entrenamiento y 6 de testeo por especie y por muestra, se utilizaron 72 imágenes de entrenamiento y 72 para las pruebas.

La idea fue verificar si el método podía aprender más información al agregar imágenes que si bien pertenecen a la misma especie, tienen ciertas diferencias que podrían confundir al sistema.

Vemos en la Tabla 4 los resultados obtenidos con el método de eigenfungi con dos muestras.

Tabla 4. Porcentajes de acierto eigenfungi con dos muestras

Bin	flocc	canis	gyps	menta	rubrum	tons
flocc		79.17	54.17	95.83	70.83	50
canis			79.17	75	79.17	62.5
gyps				95.83	95.83	87.5
menta					95.83	70.83
rubrum						87.50

Vemos un alto porcentaje de acierto, con más de 75 % en 10 casos (incluyendo un 95.83 % en 4 casos). Sin embargo, el par *E. floccosum vs T. tonsurans* no fue reconocido y en el caso de *M. canis vs T. tonsurans* se obtuvo un 62.5 %. Las pruebas totales fueron menores al 50 %.

5.3. Preprocesamientos en dos muestras

A fin de mejorar estos resultados, y de la misma manera que se hizo en el caso de una muestra, se repitieron estas pruebas combinando el método con distintos preprocesamientos. El preprocesamiento que presentó mejores resultados, fue el de suavizado de bordes y corrección de histograma, al igual que en el caso de una muestra sola. Los porcentajes obtenidos fueron mostrados en la Tabla 5.

En comparación con el método de eigenfungi puro (Tabla 4), 8 de los pares se vieron incrementados y el resto se mantuvo o bajó muy poco como *M. canis vs T. rubrum* que bajó de 79.17 % a 75 %.

Tabla 5. Porcentajes de acierto eigenfungi con dos muestras combinado con suavizado de bordes y corrección histograma

Bin	flocc	canis	gyps	menta	rubrum	tons
flocc		83.33	83.33	100	87.50	70.83
canis			70.83	70.83	75	66.67
gyps				95.83	95.83	75
menta					100	66.67
rubrum						100

Salvo *M. canis vs T. tonsurans* y *T. mentagrophytes vs T. tonsurans*, ambos con 66.67 % de acierto, las demás pruebas tienen porcentajes mayores al 70 % y hay 5 de ellos con 95.83 % o 100 %. Todos los pares fueron reconocidos.

El porcentaje total no se modificó significativamente, ya que superó el 50 % pero igualmente fue muy bajo, con un 55.56 %.

Este porcentaje aumentó si no se usaban todas las especies a la vez, sino que se sacaba alguna y se hacían pruebas con por ejemplo 5 especies juntas. Específicamente, las pruebas realizadas sin la especie *T. tonsurans*, llegaron a un porcentaje del 70.83 %.

6. Comparaciones con otros métodos

A fin de encontrar mejoras al método de reconocimiento desarrollado, se probaron algunas variantes del método de los eigenfungi. Sin embargo, se decidió no aplicarlas al estudio de los dermatofitos dado que, a pesar de registrarse algunos pares con porcentajes altos, los resultados de acierto en forma global fueron más bajos en comparación con el método eigenfungi presentado (en particular en las pruebas a nivel totales) y algunos pares no fueron reconocidos.

6.1. Cálculo de multiespacios

La variante de Multiespacios plantea que, al momento de generar los eigenfungi, no se utilicen todas las imágenes originales, sino solamente las de cada individuo cada vez. Salvo un par de especies como *E. floccosum versus M. canis* con 91.67 % de acierto, y por ejemplo *M. canis con T. rubrum* con un 100 %, los porcentajes en general fueron muy bajos y varios pares no fueron reconocidos, como *E. floccosum versus T. rubrum*.

6.2. Utilización de distancia Manhattan

Para hallar los vectores de distancias que caracterizan cada imagen de entrenamiento, se compara cada una con los eigenfungi obtenidos. Para esta comparación se calcula la distancia de cada imagen a cada eigenfungi, siendo utilizada

para esto la distancia Euclídea. En vez de utilizar esta distancia, se probó el método modificado por la aplicación de la distancia Manhattan.

Se observaron altos porcentajes de acierto entre varias especies como *E. floccosum versus T. mentagrophytes* con un 100 %. Pero no se reconoció el par *E. floccosum versus T. tonsurans* con un 50 % y el porcentaje de las pruebas totales fue muy bajo, también del 50 %.

7. Conclusiones

En este trabajo se analizó el método automático que desarrollamos para el reconocimiento de especies de hongos microscópicos, que llamamos eigenfungi, pero con dos muestras, dado que en los hongos microscópicos el reconocimiento es a nivel de especies, por lo que podrían existir diferencias si las imágenes fueron tomadas de diferentes muestras.

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

Se realizaron pruebas con 2 muestras tanto en el entrenamiento como el testeo. Las pruebas binarias combinadas con preprocesamientos resultaron con altos porcentajes de acierto, no así en el caso de las pruebas totales, con un máximo de 55,56 %. Este porcentaje aumentó si no se usaban todas las especies a la vez, sino que se sacaba alguna y se hacían pruebas con por ejemplo 5 especies juntas. Específicamente, las pruebas realizadas sin la especie *T. tonsurans*, llegaron a un porcentaje del 70.83 %.

Luego comparamos los resultados con variantes del método a fin de encontrar mejoras. Sin embargo, a pesar de registrarse algunos pares con porcentajes altos, los resultados de acierto en forma global fueron más bajos en comparación con el método eigenfungi presentado (en particular en las pruebas a nivel totales) y algunos pares no fueron reconocidos. En trabajos posteriores se podrían realizar nuevas pruebas con conjuntos de imágenes más grandes.

Con estas pruebas podemos ver que el método funciona con altos porcentajes de acierto también en el caso de 2 muestras. Esto se suma a las ventajas del mismo, que no necesita recortes manuales ni de normalización de las imágenes. Aún en el caso de dos muestras no es necesaria gran cantidad de imágenes y el entrenamiento no es costoso en tiempo ni computacionalmente.

Los resultados obtenidos son alentadores, incluso cuando la identificación no es del 100 %, ya que en una situación clínica real la aplicación de este tipo de metodologías de data mining no reemplazaría al técnico calificado, sino que lo asistiría en la identificación. Dada la complejidad y diversidad morfológica de los hongos responsables de infecciones, contar con un sistema que pueda orientar el análisis hacia un grupo más restringido de especies, aún cuando no se logre una identificación precisa, es una valiosa ayuda.

Referencias

1. Marcela L. Riccillo, Marcelo A. Soria, Oscar Bustos - "EIGENFUNGI: Desarrollo de un método de Data Mining para la Detección Automática de Patrones en Microscopía Aplicada a Micología Médica" - WICC08 X Workshop de Investigadores en Ciencias de la Computación - Mayo 2008
2. M. Turk, and A. Pentland, "Eigenfaces for recognition"- Journal of Cognitive Neuroscience 3 (1): 7186 1991
3. 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
4. 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
5. Deepak S. Turaga, T. Chen - "Face recognition using mixtures of principal components" - IEEE ICIP, Rochester - Set. 2002
6. Jian Yang, David Zhang, Alejandro Frangi, Jing-yu Yang - "Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition" - IEEE Transactions on Pattern Analysis and Machine Intelligence Vol 26 No 1 131-137 - Enero 2004
7. Daoqiang Zhang, Zhi-Hua Zhou, Songcan Chen - "Diagonal Principal Component Analysis for Face Recognition" Pattern Recognition Volume 39, 140-142 - Enero 2006
8. 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
9. 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 Conference on Artificial Neural Networks, vol2, pp 797-802 - 1998
10. Thorsten Dorge, Jens Michael Carstensen, Jens Christian Frisvad - "Direct Identification of pure *Penicillium* species using image analysis" - Journal of Microbiological Methods, Volume 41, Number 2, pp. 121-133(13) - Julio 2000
11. Iain M. Inglis, Alison J. Gray - "An Evaluation of Semiautomatic Appoaches to Contour Segmantation Applied to Fungal Hyphae" - Biometrics 57, 232-239 - Marzo 2001
12. "ImageJ - Image Processing and Analysis in Java" - National Institutes of Health, USA http://rsb.info.nih.gov/ij/
13. J. Vilata Corell - Micosis Cutáneas - Ed. Médica Panamericana, España - 2006
14. Marcela L. Riccillo, Ana S. Haedo, Natalia Debandi, Daniel Vazquez V. - "Comparación de Softwares Estadísticos" - CLATSE VI Congreso Latinoamericano de Sociedades de Estadística - SAE/SOCHE Concepción, Chile - Nov. 2004
15. 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
16. Jiawei Han, Micheline Kamber - Data Mining: Concepts and Techniques - Morgan Kaufmann Publishers, USA - 2001
17. Gonzalo Pajares Martinsanz, Jesús M. de la Cruz García - Visión por computador - Alfaomega, España - 2002