Análisis exhaustivo de algoritmos de clasificación espectral basados en píxeles para la identificación y conteo de nano-partículas en suspensión en sustrato de manglares.

Rodolfo Mora-Zamora*, Andreína Garro-Vargas†, Diego Jiménez-Vargas† y Karolina Villalobos‡
*Escuela de Ingeniería en Computación, Instituto Tecnológico de Costa Rica, Centro Académico de Alajuela. Email: rodmora@itcr.ac.cr
†Colaboratorio Nacional de Alta Tecnología, Centro Nacional de Alta Tecnología, Email: cnca@cenat.ac.cr
‡Laboratorio Nacional de Nanotecnología, Centro Nacional de Alta Tecnología

***Resumen*—Las nanopartículas en suspensión capturadas en imágenes del microscopio de fuerza atómica del LANOTEC deben caracterizarse morfológicamente. Los análisis pueden aplicarse de forma manual, pero la cantidad de muestras dificulta esta tarea, por lo cual es indispensable automatizar parte del proceso para acelerar la obtención de resultados. Los algoritmos de clasificación espectral basados en píxeles pueden separar las partículas observadas en la imagen del fondo. Una vez separadas se pueden vectorizar y estimar la información morfológica de forma automática. Para determinar un flujo automatizado es necesario comparar la precisión de los diferentes algoritmos y establecer los que presenten mejor precisión con mayor cantidad de imágenes de forma consistente. Se aplicaron 6 de los algoritmos disponibles en la plataforma ENVI 5.1 y se compararon los resultados sobre una colección de 14 imágenes diferentes para un total 1255 clasificaciones. Los algoritmos no supervisados K-Means y ISO-Data mostraron mejor rendimiento promedio con comportamientos más consistentes a través de todas las imágenes muestreadas. Las imágenes de cambio de fase fueron descartadas por problemas de ruido excesivo en las clasificaciones resultantes.**

I. Antecedentes

En el transcurso del 2016 un equipo de investigadores del Laboratorio Nacional de Nanotecnología (LANOTEC) capturó múltiples muestras de sedimento del suelo en tres manglares de la costa pacífica de Costa Rica, Tamarindo en la provincia de Guanacaste, Punta Morales y Lepanto, ambas de la provincia de Puntarenas. El objetivo del estudio es determinar la presencia y características de nanopartículas encontradas en el sedimento del manglar, formadas en condiciones naturales, con el fin de entender las propiedades filtradoras que estos ecosistemas poseen. El estudio de estas propiedades podría permitir en el futuro crear nuevos sistemas de purificación de agua [1].

Como parte de los pasos para la caracterización de las nanopartículas es necesario identificar su estructura física y su distribución en el sedimento observado. Las partículas deben de ser contadas, medidas en área y determinadas morfológicamente. Para esto las muestras son capturadas en imágenes utilizando el microscopio de fuerza atómica (AFM) del LANOTEC.

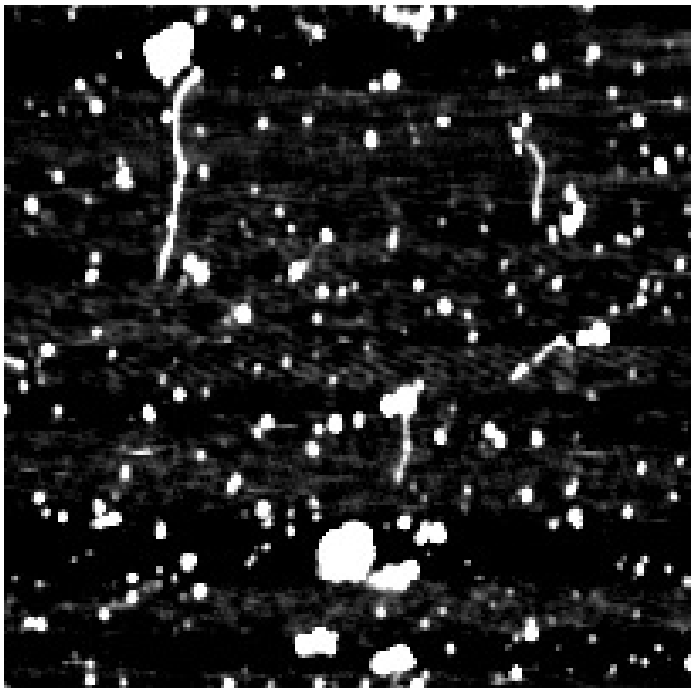

Figura 1. Imagen de suspensión de nanopartículas en manglar Recorte-Muestra 1I0002, LANOTEC (2016)

A partir de las imágenes capturadas pueden analizarse las propiedades de las partículas. Si bien, dichos procesos pueden realizarse manualmente, un estudio de este tipo puede comprender cientos de imágenes. Esta investigación pretende brindar un primer paso para determinar un flujo de trabajo estandarizado en función de automatizar el proceso de caracterización de las nanopartículas en suspensión. El flujo de trabajo se compondría de los siguientes pasos:

1. Definir una estrategia de colección para las imágenes.
2. Clasificar los píxeles en cada imagen utilizando un algoritmo automático cuya precisión pueda medirse de forma cuantitativa.
3. Vectorizar la clasificación para obtener objetos discretos que puedan ser analizados por sus propiedades geométricas.
4. Calcular índices de forma y tamaño para segregar vectores en grupos de interés.
5. Caracterizar los vectores segregados según su tamaño y forma.
6. Contar los vectores segregados y calcular su porcentaje de saturación en la suspensión observada.

Para las primeras dos etapas del proceso es necesario comparar la precisión de varios algoritmos de clasificación automática aplicados a imágenes de suspensión de sedimentos, como las capturadas en el 2016, adicionalmente es indispensable determinar las condiciones de captura que faciliten la estandarización del análisis automático de las imágenes.

Las imágenes tienen poca complejidad espectral (se componen de una sola banda y por lo tanto se pueden representar en escala de grises), lo cual puede causar problemas con algunos algoritmos de clasificación. El tamaño de las imágenes es de aproximadamente 200x200 píxeles, por lo cual su clasificación demanda pocos recursos computacionales, pero la falta de detalle puede impactar negativamente el resultado. En la figura 1 se puede apreciar la complejidad espectral y la distribución de partículas en una de las imágenes usadas en las pruebas.

II. Metodología

II-A. Pruebas de clasificación

En función de determinar cuáles algoritmos dan mejores resultados y cuáles condiciones de capturas de imágenes tienen mayor precisión de clasificación, es necesario realizar pruebas exhaustivas de varios algoritmos. Para esta prueba preliminar se escogieron algoritmos basados en píxeles de la familia disponible en el software ENVI versión 5.1 (Exelis Visual Information Solutions, Boulder, Colorado)[2]. Específicamente se seleccionaron los siguientes algoritmos:

- No Supervisados
  - K-Medias
  - ISO-Data
- Supervisados
  - Mínima distancia
  - Distancia de Mahalanobis
  - Redes neurales
  - Máquinas de soporte vectorial

Dada la baja complejidad espectral de las imágenes, los algoritmos Codificación binaria y Asignador de ángulo espectral no pudieron aplicarse ya que las imágenes no cumplen los supuestos mínimos para calcular las clases de entrenamiento.

El algoritmo Máxima verosimilitud no se pudo aplicar debido a un error de ejecución también atribuido a la baja complejidad espectral de las imágenes.

Por último el algoritmo Paralelepípedo fue descartado debido a su baja tolerancia para incluir muestras que no hayan sido representadas en el conjunto de entrenamiento.

Por cada algoritmo se probaron distintos parámetros de ajuste con distintas imágenes de 2 grupos:

- Imágenes de topografía
- Imágenes de cambio de fase

En total se clasificaron 14 imágenes de suspensión de partículas, todas capturadas con el AFM de LANOTEC. Se obtuvieron 1255 clasificaciones usando 6 diferentes algoritmos con distintas parametrizaciones cada uno.

II-B. Depuración y validación

Antes de validar las clasificaciones, se analizaron los resultados visualmente y se detectó que las imágenes de fase presentaban problemas de ruido excesivo en el resultado, producto de los artefactos típicos de la representación de la imagen. El ruido excesivo produce problemas de sobre-segmentación al vectorizar las imágenes, lo cual impediría el conteo de partículas. Por esta razón se decidió descartar todas las imágenes de fase del experimento en esta etapa.

Una vez depurado el conjunto de clasificaciones se aplicó la validación de las imágenes restantes utilizando la técnica de matriz de confusión con el objetivo de detectar los errores de omisión y comisión. Para esta validación se tomaron conjuntos de puntos potencialmente problemáticos, como puntos en los bordes de los objetos de interés, y muestras altamente representativas de la clase deseada, puntos del interior de los objetos de interés. En la figura 2 se aprecian algunos ejemplos de puntos tomados para el conjunto de validación.

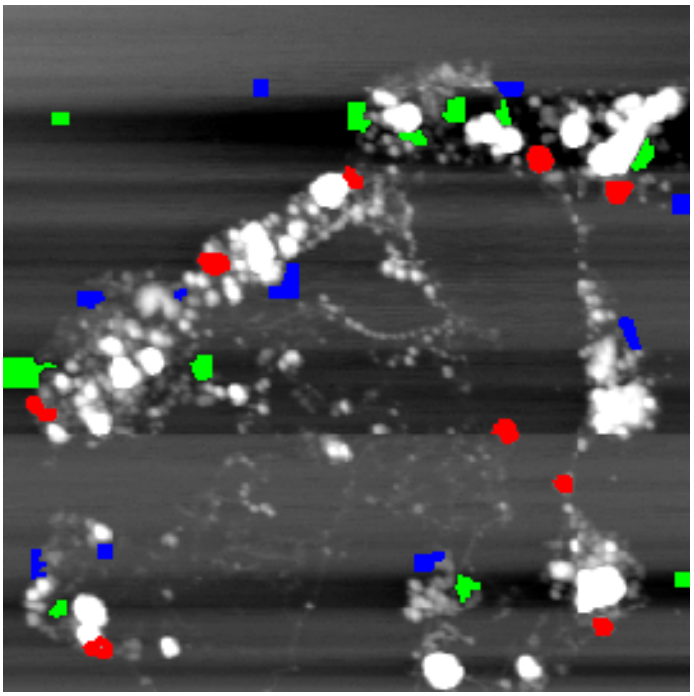

Figura 2. Regiones de interés para validación en la imagen Recorte 3I0007, elaboración propia

En la matriz de confusión las columnas representan los valores verdaderos, mientras que las filas corresponden a los valores clasificados, de esto se obtiene que la diagonal de la matriz contiene los valores que fueron clasificados correctamente. Al recorrer las columnas de la matriz se encuentran aquellos píxeles que pertenecen a una clase, pero fueron clasificados en otra, los cuales se llaman falsos negativos. Por otra parte si se recorren las filas de la matriz se obtienen los falsos positivos, que son los píxeles que fueron clasificados en una clase específica, pero en realidad pertenecen a otra [3].

A partir de la matriz de confusión se pueden obtener tres medidas de precisión. La primera de estas medidas es la precisión general, la cual se define como la suma de los verdaderos positivos entre el total de píxeles clasificados. Las otras dos medidas corresponden al cálculo del error por omisión y el error por comisión. El error por omisión se obtiene a partir de la suma de los falsos negativos, entre el total de píxeles verdaderos. Además, el error por comisión corresponde a la razón de el total de falsos positivos entre el total de píxeles clasificados [3].

Otra medición que se obtiene a partir de la matriz de confusión es el coeficiente kappa, el cual mide la concordancia entre los píxeles verdaderos y los píxeles clasificados [3]. La fórmula matemática para el cálculo de este coeficiente se muestra en la ecuación 1:

$$\kappa = \frac{N \sum_{i=1}^{n} m_{i,i} - \sum i = 1^{n} (G_i C_i)}{N^2 - \sum i = 1^{n} (G_i C_i)}, \qquad (1)$$

donde

i= Número de clases.

N= Total de valores clasificados.

$m_{i,i}$= Suma de valores clasificados correctamente de la clase i.

C_i= Total de valores clasificados como clase i.

G_i= Total de valores verdaderos pertenecientes a la clase i.

III. Resultados

Para cada algoritmo se buscó la combinación de parámetros con mejor precisión promedio a través de todas las imágenes, destacada como la parametrización ideal para dicho algoritmo. En el cuadro I se puede apreciar el promedio de la precisión general de cada algoritmo de clasificación con la mejor parametrización disponible. A partir de estos resultados se puede determinar que los algoritmos K-Medias y ISO-Data tienen mejor rendimiento con los parámetros descritos.

Algoritmo	Precisión	Parámetros
K-Medias	92,72 %	C:2, I:10
ISO-Data	92,27 %	C:2-10, I:2
Mínima distancia	85,86 %	MDE: none, MSD: 5
Dist. de Mahalanobis	83,75 %	MDE: 100
Redes neurales	88,66 %	A: Logistic, HL: 1, I:1000
SVM	85,55 %	K: Radial Basis Function

Cuadro I
Comparación de cada algoritmo con la parametrización que reportó el mejor promedio de precisión general en la clasificación

Acrónimo	Significado
C	Cantidad de clases
I	Iteraciones
A	Método de activación
HL	Cantidad de capas ocultas
K	Kernel type
MSD	Máxima desviación estándar
MDE	Máxima distancia de error

Cuadro II
Simbología de las parametrizaciones representadas en el cuadro I

La figura 3 ilustra el resultado de una de las mejores clasificaciones con K-Means, usando la parametrización seleccionada, en contraste con la imagen original respectiva.

Para garantizar que los resultados dependen del algoritmo utilizado se sometieron las precisiones generales a una prueba de Análisis de Varianza, pero ni la variable,

ni sus transformaciones, presenta distribución normal, por lo que se aplicó el homólogo Kruskal-Wallis para contrastar [4]. La prueba de ANOVA obtuvo el valor-p $2x10^{-16}$, mientras que la prueba de Kruskal-Wallis obtuvo el valor-p $2{,}2x10^{-16}$.

Ambas pruebas demuestran con alta significancia que la precisión obtenida depende del algoritmo aplicado. La prueba de comparaciones múltiples de Kruskal-Wallis además arroja que los resultados entre las clasificaciones de K-Medias e ISO-Data no son diferenciables, por lo que ambos algoritmos pueden aplicarse indistintamente.

IV. Conclusiones

Las imágenes de suspensión de nanopartículas capturadas con AFM pueden clasificarse usando algoritmos automáticos basados en píxeles. Por lo tanto es posible automatizar parte del proceso de caracterización de las partículas.

Los algoritmos K-Medias (con 2 clases y 10 iteraciones) y ISO-Data (con 2 a 10 clases y 2 iteraciones) obtuvieron los mejores resultados de clasificación, 92.72 % y 92.27 % respectivamente.

Las imágenes de topografía de la muestra son ideales para la aplicación de los algoritmos descritos en este trabajo. Las imágenes de cambio de fase producen clasificaciones con ruido que no es posible procesar.

V. Agradecimientos

Los autores de este trabajo agradecemos al LANOTEC por la suministración de las imágenes y toda la colaboración brindada durante la redacción del artículo, y al Laboratorio PRIAS por brindarnos el software especializado, así como por asesorarnos en aspectos técnicos de su uso.

VI. Referencia bibliográficas

Referencias

[1] M. Soto, "Nanoestructuras de manglar inspirarán purificadores de agua," *La Nación*, vol. 4, no. 2, pp. 201–213, Jan. 2017. [Online]. Available: http://www.nacion.com/vivir/ciencia/Nanoestructuras-manglar-inspiraran-purificadores-agua_0_1609239073.html

[2] M. Galloy. (2017) Classification (using envi). [Online]. Available: https://www.harrisgeospatial.com/docs/Classification.html

[3] (2017) Calculate confusion matrices (using envi). [Online]. Available: https://www.harrisgeospatial.com/docs/CalculatingConfusionMatrices.html

[4] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Associatio*, vol. 47, no. 260, pp. 583–621, Dec. 1952. [Online]. Available: http://www.jstor.org/stable/2280779

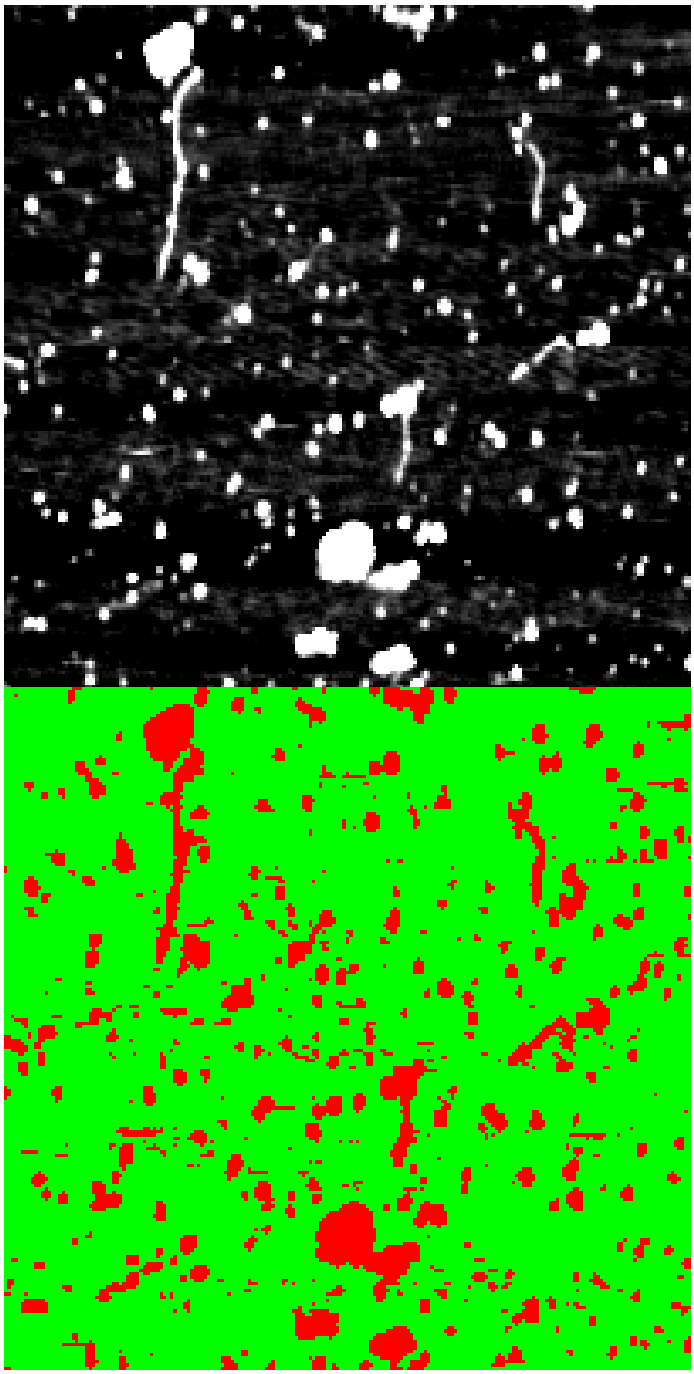

Figura 3. Superior: Imagen original Recorte-Muestra 1—0002, LANOTEC (2016). Inferior: Clasificación obtenida con la mejor parametrización del algoritmo K-Means, con precisión general 95.67 %