CLASIFICACIÓN EFICIENTE DE OBJETOS USANDO LA CARACTERÍSTICA DE EULER

Amézquita Morataya, Erik (1), Canul Ku, Mario (2), Rieser, Antonio (3)

(1) Departamento de Matemáticas, Universidad de Guanajuato | Dirección de correo electrónico: erik.amezquita@cimat.mx
(2) Centro de Investigación en Matemáticas
(2) Cátedras CONACYT-CIMAT, Centro de Investigación en Matemáticas

Resumen

Este es un proyecto conjunto entre matemáticas, computación y arqueología. Un problema enfrentado por los arqueólogos es la falta de una clasificación estándar de máscaras prehispánicas. A pesar de que se considera la localización y línea temporal de cada máscara, la clasificación está sujeta a la percepción subjetiva de cada arqueólogo. Empleando la Característica de Euler, se está desarrollando un algoritmo computacionalmente eficiente que permita establecer una clasificación nueva de estas máscaras. La clasificación se basa en invariantes topólogicos y geométricos de cada objeto, por lo que puede proveer un enfoque más objetivo en el campo arqueológico.

Abstract

This is a joint project among mathematics, computer science and archaeology. An important problem faced by archaeologists is the lack of an standard classification when it comes to pre-Columbian masks. Even though for each mask its location and age is taken into account, the actual classification is riddled with subjectivies from each archaeologist. A new algorithm is currently being devoloped based on the Euler Characteristic. The main goal is to develop a computationally efficient algorithm that leads to a new classfication of these masks. This classification is be based on topological and geometrical invariants of each artifact, and it might provide a more objective insight in archaeology.

Palabras Clave: topología algebraica; análisis topológico de datos; arqueología; máquina de soporte vectorial

INTRODUCCIÓN

Un problema importante en arqueología es establecer una clasificación única de objetos dentro de una misma familia. Por ejemplo, se busca dividir en grupos una familia de máscaras prehispánicas, de tal manera que las máscaras dentro de cada familia compartan varias características. Muchas veces, dicho agrupamiento contiene un porcentaje de subjetividad, y por ello, el agrupamiento puede prestarse a discrepancias entre dos arqueólogos distintos, ambos expertos en el área. El objetivo del proyecto de investigación es buscar invariantes topológicos y geométricos en las máscaras y que sean éstos los que realmente determinen la pertenencia o no a cada máscara en cada grupo. Se cuenta con datos provistos por el Instituto Nacional de Antropología e Historia (INAH) los cuales son modelos tridimensionales de 128 máscaras distintas halladas en las excavaciones del Templo Mayor en la Ciudad de México entre 1978 y 1982.

Este proyecto en particular se basa en la idea de gráfica de característica de Euler (GCE) descrita por Richardson y Weirman en [4]. Esta técnica pretende ser lo más general posible y se está desarrollando con la idea de ser aplicable para cualquier tipo de objeto, no necesariamente arqueológico, que deba ser identificado y clasificado.

METODOLOGÍA DESARROLLADA

La característica de Euler es un invariante topológico, el cual se resume como un número intrínseco del objeto que no cambia con deformaciones suaves del mismo. A través de una función de filtración, se destruye el objeto de manera sistemática en una serie predeterminada de pasos. Esta destrucción altera la topología del objeto en cada paso, lo cual también afecta su característica de Euler. Empleando las GCEs de cada máscara como descriptor, se ha procedido a emplear varios métodos de clasificación con herramientas de aprendizaje de máquina, tanto métodos supervisados como no supervisados, con enfoque particular en las máquinas de soporte vectorial (SVM).

Topología y la Característica de Euler

Para un objeto de n dimensional $X = (V_0, V_1, \ldots, V_n)$, donde V_k es su conjunto de celdas k dimensionales, su característica de Euler está dada por:

$$\chi = \sum_{k=0}^{n} (-1)^k |V_k|.$$

Sabemos que χ en realidad depende únicamente de los números de Betti β_k a través de la igualdad:

$$\chi = \sum_{k=0}^{n} (-1)^k \beta_k.$$

Los β_k dependen a su vez únicamente de los grupos de homología de X. En otras palabras, χ es invariante ante deformaciones continuas de X.
De manera más informal, β_k representa el número de agujeros de dimensión k esencialmente distintos en X. Por ejemplo, β_0 denota el número de componentes conexas, β_1 el número de ciclos, β_2 los agujeros en el sentido tradicional de la palabra, etc. Un tratamiento más cuidadoso de este concepto puede verse en [1].

Gráficas de Característica de Euler (GCEs)

La GCE de X se construye a partir de un número fijo T de umbrales y una función $g_0 : V_0 \to [a, b] \subset \mathbb{R}$ cualquiera. Ejemplos de posibles g_0's incluyen funciones relacionadas con curvatura o distancia al centro de masa del objeto. A partir de g_0 se deducen funciones auxiliares $g_k : V_k \to [a, b],\ k > 0$, para el resto de k-celdas como:

$$g_k(\{u_0, u_1, \ldots, u_k\}) = \min_{0 \le i \le k} \{g_0(u_i)\}.$$

De ese modo, toda k-celda tiene un valor numérico asignado, $k = 0, 1, \ldots, n$. Se divide $[a, b]$ en T partes iguales, estableciendo umbrales $a = t_0 < t_1 < \ldots < t_T < b$. Para el i-ésimo umbral t_i, se deduce el subconjunto de k-celdas $V_k^{(i)}$ definido por $V_k^{(i)} = \{v \in V_k : g_k(v) \ge t_i\}$. Finalmente, la característica de Euler al i-ésimo umbral la definimos simplemente como:

$$\chi_i = \sum_{k=0}^{n} (-1)^k |V_k^{(i)}|.$$

Nótese que por la construcción de las g_k's, si un vértice v cumple $g_0(v) < t_i$, entonces el resto de k-celdas w que contienen a v como vértice cumplen $g_k(w) < t_i$. Es decir, al remover un vértice al i-ésimo umbral, también se remueven todos las k-celdas que lo contienen. De ese modo, en cada umbral siempre se cuenta con un complejo simplicial válido.

La GCE de X se define entonces como la gráfica de úmbrales t_i versus χ_i. Esta es una gráfica que captura la evolución topológica del objeto mientras se "desintegra", pues a medida que el valor del umbral aumenta, existen menos puntos que lo sobrepasen.

Si se supone que los valores $g_0(v)$ ya han sido calculados para todo vértice, la asignación de valores $g_k(w)$ y el cómputo de los valores de $|V_k^{(i)}|$ para cada i, k puede realizarse en tiempo lineal sobre el número de vértices $O(V_0)$. Evaluar cada χ_i es una simple suma n veces , por lo que el cómputo de la GCE en total es de complejidad lineal $O(V + T)$. El algoritmo es descrito con más detalle en [4]. Es por ello que el cómputo de la GCE es considerablemente rápido, dando pie a varias aplicaciones.

Máquinas de Soporte Vectorial (SVMs)

Se espera que GCEs similares indiquen que los objetos originales son similares, y por ende, tendría sentido que pertenezcan a un mismo grupo o categoría. Bajo dicha premisa, se procede a emplear algoritmos de clasificación estándar en la literatura de aprendizaje de máquina. Primero se experimentó con los métodos de clasificación supervisados de k Vecinos Más Cercanos (KNN) y Máquina de Soporte Vectorial (SVM). Ambos métodos dependen de un subconjunto de entrenamiento en el cual el usuario ya ha clasificado manualmente los datos; a partir de dicho entrenamiento, la máquina deduce la clasificación del resto de datos en el conjunto de prueba.

En el caso del SVM, basándose en el conjunto de entrenamiento, la máquina busca dividir el espacio con hiperplanos que maximicen la distancia entre éste y los representantes del resto de categorías. A dicha separación se le denomina margen. Los puntos (vectores en $\mathbb{R}^n$) de cada clase más cercanos al hiperplano son los que determinan al mismo, por lo que se denominan vectores de soporte. Así, se delimita la ubicación de categorías distintas. Si por ejemplo los datos se representan como puntos en el plano, la máquina deduce líneas que separen de la mejor manera ambos grupos en la fase de entrenamiento como indica la imagen 1. Más detalles sobre su fundamento teórico puede verse en [2].

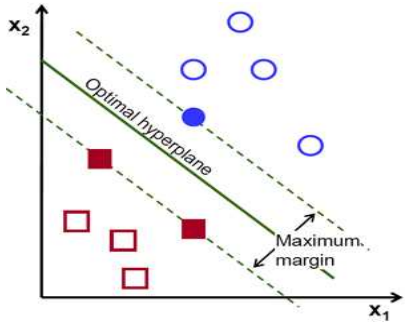


IMAGEN 1: Hiperplano (línea) que maximiza el margen entre la categoría azul y la roja. En sólido se muestra a los vectores de soporte.

Este método sin embargo es susceptible a funcionar incorrectamente si el conjunto de datos a clasificar es pequeño, pues a menor cantidad de datos, menos información se tiene para deducir correctamente los hiperplanos. Esto también causa que los hiperplanos deducidos dependan fuertemente de los pocos datos de entrenamiento. En otras palabras, se vuelve susceptible a la clasificación provista por el usuario en primer lugar.

Máquinas de Soporte Vectorial No Supervisadas (USVMs)

En vista de dichos problemas se ha tratado de implementar algoritmos de clasificación no supervisados, es decir, que no dependen de ningún subconjunto de entrenamiento. Se implementó una Máquina de Soporte Vectorial No Supervisada (USVM) siguiendo las ideas propuestas por Karnin et. al. en [3]. La idea en el caso de únicamente dos familias a clasificar consiste en usar todos los puntos y escoger una pareja. Se supondrá que cada uno pertenece a una familia distinta, por lo que el hiperplano divisor debe pasar por el punto medio de la pareja. Con esa restricción, puede calcularse un hiperplano tal que en promedio es distante a la mayoría de todos los puntos. Se calcula el margen de separación y se vuelve a iterar el algoritmo,

donde se escoge una pareja distinta. Así, se evalúan todas las parejas posibles y se determina que el hiperplano correcto es aquel que maximiza el margen. Un ejemplo de este procedimiento puede verse en la imagen 2.

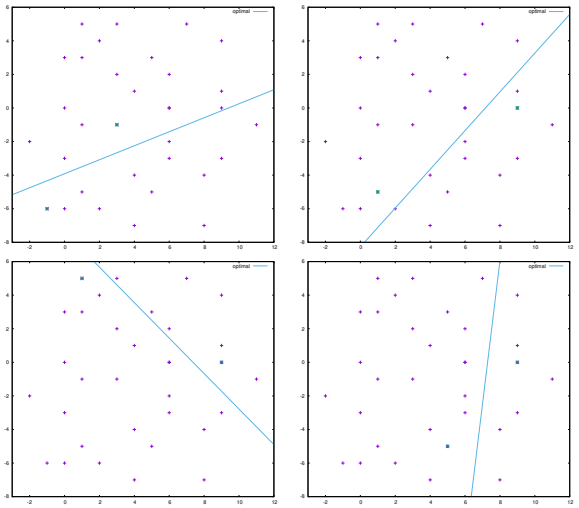


IMAGEN 2: Distintos planos divisores correspondientes a distintas elecciones de parejas.

Al tratarse de un método no supervisado éste no se ve afectado por la cantidad de datos a clasificar. Sin embargo, el cómputo de cada hiperplano para cada pareja es computacionalmente demandante, lo cual lo hace lento para ejecutar. Además, el algoritmo actual únicamente puede discernir cuado hay sólo dos grupos. Parte del trabajo futuro es implementar un sistema de votación para dividir el espacio con varios hiperplanos que determinen varias categorias.

RESULTADOS

Se cuenta con 128 modelos triangulados tridimensionales de máscaras prehispánicas distintas provistas por el INAH. Así, para cada máscara únicamente se consideraron los conjuntos V_0, V_1, V_2 correspondientes a vértices, aristas y caras. La lectura y manipulación de datos para generar las GCE fue implementado en `C/C++` mientras que las gráficas fueron elaboradas en `gnuplot`. El SVM empleado para clasificar las GCEs fue implementado en Python. Se aprovechó del agrupamiento de máscaras proveída de antemano por el INAH como punto de partida para revisar correspondencias y diferencias entre la partición dada por arquéologos y la partición obtenida por las GCEs. Todos los algoritmos fueron ejecutados en una computadora personal con procesador `Intel Core i5-6200U CPU @ 2.30GHz` y 8Gb de RAM.
Se experimentó con varias funciones g_0 basadas en los valores principales de curvatura de cada punto, pero las GCE obtenidas de este modo no son visualmente claras como lo indican la imagen 3. También éstas respondieron de manera pobre al momento de usar SVMs con parámetros diversos. La clasificación obtenida así no presentaba diferencias visuales claras entre las familias a pesar que se probó con distintos conjuntos de entrenamiento y prueba.

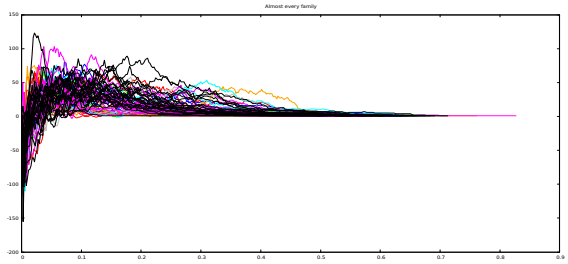


IMAGEN 3: Curvatura media para $T = 256$ umbrales. Colores distintos indican pertenencia a familias distintas.

Se aprovechó el hecho que los modelos 3D tienen su baricentro en el origen y se hallan encajados en un cubo $[-1, 1]^3$ de modo que fue posible considerar la distancia de cada vértice a las caras izquierda, inferior y posterior del cubo como función g_0. Cada proyección define una GCE, y al final se concatenaron las tres para obtener una GCE compuesta.

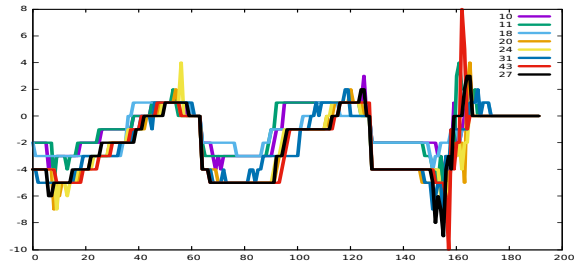
(a) Horizontal **(b)** Vertical **(c)** Frontal

IMAGEN 4: Mapa de calor. Para cada g_0 como proyección, los vértices verdes denotan valores bajos y los ocres altos.

Un problema afrontado al emplear el SVM fue la poca cantidad de datos para la mayoría de familias. De las 9 familias originales, 6 de ellas cuentan

con menos de 7 especímenes. Más aún, una de ellas consiste de 59 miembros cuya única característica común es no pertenecer a ninguna de las otras 8 familias. Para mejorar el SVM, se tomó como conjunto de entrenamiento todos los miembros de las 6 familias pequeñas, junto con los 8 miembros más representativos de las otras tres familias. El resto de máscaras se usaron en el conjunto de prueba.

Después de haber ejecutado el SVM, el primer cambio notable es la homogeneización de miembros por familia, siendo ahora 12 elementos por categoría en promedio. Esto se debe principalmente a la distribución de elementos de la familia numerosa al resto de familias. Se muestran a continuación los resultados de la clasificación nueva para dos familias. Dichos resultados fueron aprobados por la comunidad arqueológica presente en el “1er Coloquio sobre el "Desarrollo Tecnológico al Servicio del Patrimonio Cultural.”



(a) GCEs de proyecciones. Color distinto indican máscara distinta.

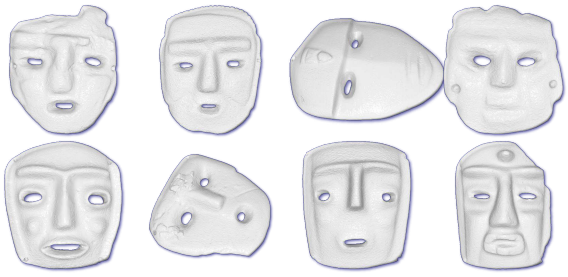
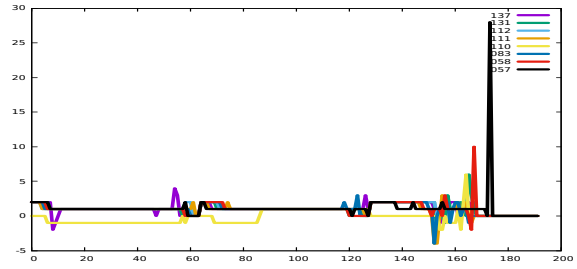


IMAGEN 5: Muestra de ocho miembros de la Familia 2 con la nueva clasificación

Podemos ver que las GCEs en la imagen 5a siguen todas un mismo patrón. Vemos que en efecto las máscaras físicamente presentan similitudes. Cabe aclarar que la nueva familia 2 es la que contiene a casi todas las máscaras con agujeros. Como contraste, aquellas de la familia 9 en la imagen 6a siguen un patrón distinto. Nótese que la gráfica negra corresponde a la tercera máscara de izquierda a derecha de la fila superior. Ésta corresponde a la máscara que resalta más visualmente. Esto es un buen indicador que la GCEs si logran captar información importante de los objetos.



(a) GCEs de proyecciones. Color distinto indican máscara distinta.

IMAGEN 6: Muestra de ocho miembros de la Familia 9 con la nueva clasificación. La gráfica negra corresponde a la máscara menos similar visualmente.

A diferencia del caso mostrado en la imagen 3, las GCEs compuestas de proyecciones muestran cierto orden por colores. En la figura 7, colores distintos corresponden a miembros de familias distintas bajo la nueva clasificación.

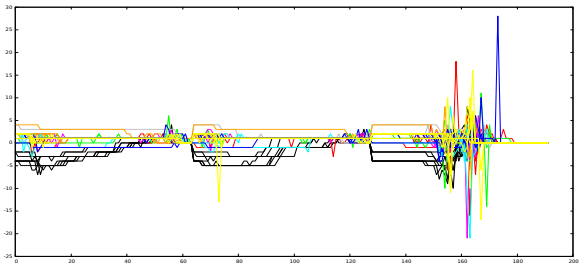


IMAGEN 7: GCEs tomando 8 miembros de cada familia acorde a la nueva clasificación, eliminando aquellos cuya GCE sobresalga del resto.

Aún así, la clasificación es sensible a cambios en el conjunto de entrenamiento, por lo que se buscan resultados con algoritmos no supervisados. Primero se ejecutó el USVM con las GCEs de los

miembros originales de las familias 2 y 9. El algoritmo pudo distinguir sin problemas las máscaras con agujeros de aquellas sin ellos. Esto confirma que las GCEs en efecto detectan esta propiedad.

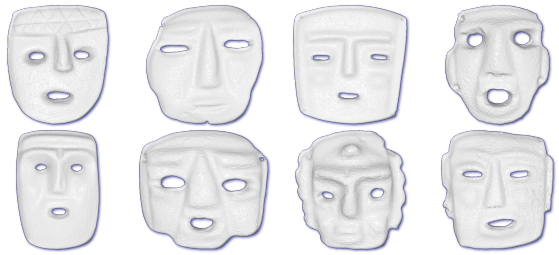


IMAGEN 8: Muestra de familia 2 en la clasificación nueva (USVM)

IMAGEN 9: Familia 9 en la clasificación nueva (USVM)

Cabe resaltar que las últimas dos máscaras de la familia 9 en la figura 9, tercera y cuarta de izquierda a derecha en la fila inferior, pertenecían inicialmente a la familia 2. Sin embargo, el USVM tiene problemas para distinguir miembros entre dos familias sin agujeros en ambas.

TRABAJO FUTURO

Se ha trabajado en la generación de datos arqueológicos sintéticos. Éstos se obtienen al deformar de manera suave los modelos tridimensionales. Con ello, se puede aumentar la base de datos de máscaras prehispánicas, lo cual puede mejorar los resultados de SVMs y otros métodos de clasificación supervisados.
También se busca experimentar con esta técnica con mayor variedad de modelos arqueológicos. Si bien todavía hay que realizar varios ajustes para que el algoritmo distinga entre dos máscaras distintas, se puede explorar si éste distingue entre dos objetos claramente distintos. Por ejemplo, entre máscaras y flautas, o entre ídolos y vasijas.

CONCLUSIONES

El cómputo de la GCE es una operación sencilla de complejidad y memoria lineales una vez que se ha asignado valores numéricos a los vértices. El algoritmo procesa rápidamente objetos de decenas de miles de vértices. Debido a su rapidez, este algoritmo hace pensar en aplicaciones en tiempo real de reconocimiento de patrones de superficies y objetos en general, no necesariamente piezas arqueológicas.
Una mayor cantidad de máscaras puede proveer de mejores conjuntos de entrenamientos y por ende, mejores clasificaciones. Más especímenes permitirán también experimentar con métodos de clasificación no supervisada.
Las técnicas expuestas aquí son generales, y pueden aplicarse a una variedad de bases de datos, tanto en 2D como 3D. La clave por supuesto está en la elección correcta de filtración g_0, de la cual dependerá

AGRADECIMIENTOS

Los autores agradecen tanto al Instituto Nacional de Antropología e Historia a través de su proyecto *Desarrollo de Aplicaciones de Computación en Arqueología* y a la Red Temática CONACYT de Tecnologías Digitales para la Difusión del Patrimonio Cultural por proveer los modelos digitales 3D de las máscaras prehispánicas. Éstos han hecho posible el desarrollo de este trabajo.

REFERENCIAS

[1] M.A. Armstrong *Basic Topology* Springer-Verlag (1983).

[2] C. Burges “A Tutorial on Support Vector Machines for Pattern Recognition”. *Data Mining and Knowledge Discovery* Vol.2 pp.121-167, 1998.

[3] Z. Karnin, et. al. “Unsupervised SVMs: On the Complexity of the Furthest Hyperplane Problem”. *JMLR: Workshop and Conference Proceedings* Vol. 23, pp. 2.1-2.17, 2012.

[4] E. Richardson, M. Weirman, “Efficient classification using the Euler Characteristic”. *Pattern Recognition Letters* Vol.49, pp.99-106, 2014.