

Clasificación eficiente usando la Característica de Euler

Erik Amézquita ¹ Mario Canul ² Antonio Rieser ³
erik.amezquita@cimat.mx

¹DEMAT, UGto

²CIMAT

³CONACYT-CIMAT

17 de agosto de 2017

Vistazo general

- 1 Objetivo general
- 2 Introducción
- 3 Homología simplicial
- 4 La gráfica CE
- 5 Clasificación y aprendizaje de máquina
- 6 Datos arqueológicos a tratar
- 7 Resultados
- 8 Conclusiones

Pregunta, Problema y Objetivo



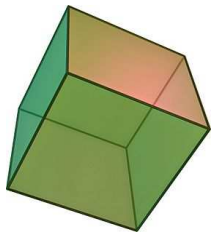
- ¿Puede la **topología** decirnos algo de estas máscaras?
- Clasificación eficiente de objetos no sujeta a subjetividades del usuario.
- Establecer criterios de clasificación basados en características geométricas y topológicas del objeto.
- Usar la idea de **gráfica de característica de Euler (CE)** como sugirieron Richardson y Weirman en el 2014 en [5].

Un invariante para poliedros

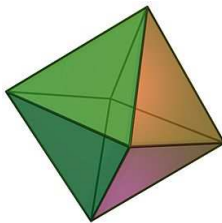
La característica Euler en poliedros se define como:

$$\chi = \#(\text{vértices}) - \#(\text{aristas}) + \#(\text{caras})$$

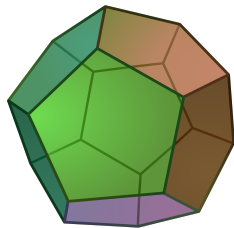
$$\chi = |V| - |E| + |F|$$



(a) $\chi = 8 - 12 + 6 = 2$



(b) $6 - 12 + 8 = 2$



(c) $20 - 30 + 12 = 2$

Con todos los poliedros esféricos

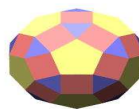
Uniform Polyhedra
page 4



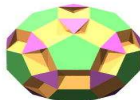
Socco



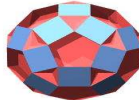
Sroh



Srid



Saddid



Sird



Raded



Ided



Ri



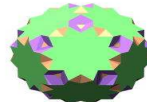
Gocco



Querco

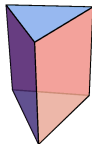


Groh

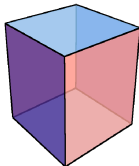


Gidditdid

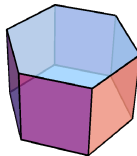
Los topólogos ven el mundo hecho de hule



Triangular
prism



Cube



Hexagonal
prism



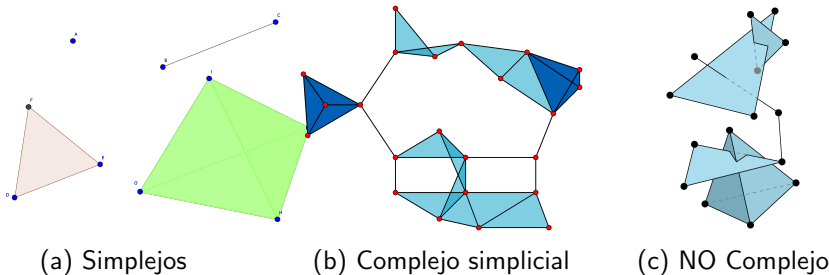
Truncated
Icosahedron



Gyrobifastigium

Triángulos en todas las dimensiones

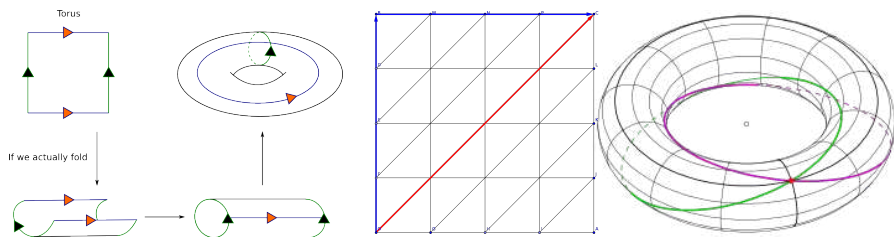
Simplejos y Complejos Simpliciales. Recuerden que deben ir bien pegados siguiendo ciertas reglas.



Al n -simplejo genérico lo denotaremos como $\sigma = (v_0, v_1, v_2, \dots, v_n)$

Mejor con dibujos: el toro

El H_1 del toro es $\mathbb{Z} \oplus \mathbb{Z}$. Eso quiere decir que $\beta_1 = 2$. Todo puede reducirse a dos ciclos, el longitudinal y el latitudinal. El resto de ciclos pueden escribirse como combinación lineal de éstos dos.



La Característica de Euler (ahora general)

- β_0 es realmente el número de componentes conexas del complejo.
- β_1 es el número de 1-ciclos (agujeros de dimensión 1) salvo homología.
- β_2 es el número de 2-ciclos (agujeros de dimensión 2, el sentido tradicional) salvo homología.
- etc...

La característica de Euler equivale a:

$$\chi = \sum_{i=0}^n (-1)^i \beta_i.$$

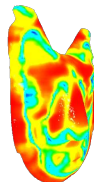
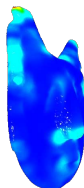
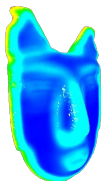
Análisis de Datos: La filtración

Fijamos una función de filtración g para los vértices y luego la extendemos al resto de las k -celdas:

$$g_k(\{v_0, v_1, \dots, v_k\}) = \min_{0 \leq i \leq k} \{g(v_i)\}$$

$g_k : V_k \rightarrow [a, b]$ una k -celda

Una función $g : V_0 \rightarrow [a, b]$
 V_0 el conjunto de vértices;
 $[a, b]$ intervalo fijo.



Umbralización

El intervalo $[a, b]$ es dividido en T umbrales equiespaciados $a = t_0 < t_1 < t_2 < \dots < t_T = b$. Consideramos la CE en el i -ésimo intervalo:

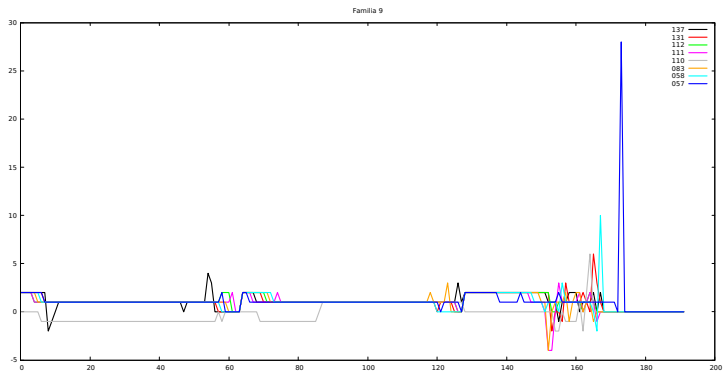
$$\chi_i = \sum_{k=0}^n (-1)^k |V_k^{(i)}|.$$

No. de k celdas c_k tales que $g_k(c_k) \geq t_i$



La GCE

La **gráfica de característica de Euler (GCE)** es simplemente comparar χ_i vs t_i



Algorítmicamente hablando I

Los valores numéricos $g(v)$ ya están calculados para todo vértice v .

```

1: Input:  $g, T$ 
2:  $\chi[T] \leftarrow 0$ 
3: for all  $k = 1 \rightarrow n$  do
4:    $H[T] \leftarrow 0$ 
5:   for all  $i = 1 \rightarrow N_k$  do
6:      $g_k \leftarrow \text{mín } g$ 
7:      $b \leftarrow \text{bin}(g_k)$ 
8:      $H[b] = H[b] + 1$ 
9:    $c \leftarrow 0$ 
10:  for all  $i = T \rightarrow 1$  do
11:     $c \leftarrow c + H[i]$ 
12:     $\chi[i] \rightarrow \chi[i] + (-1)^k c$ 
13: return  $\chi$ 

```

- ▷ Valores χ_i
- ▷ dimensiones
- ▷ histograma
- ▷ k -celdas
- ▷ $\lfloor g_k \times T \rfloor$
- ▷ umbrales

Algorítmicamente hablando II

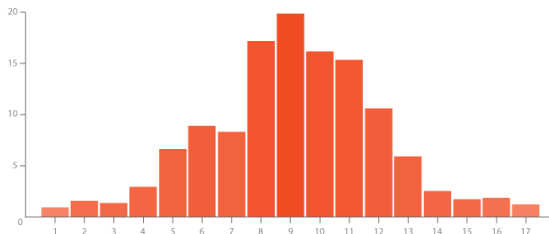
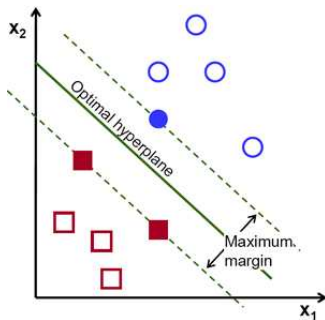


Figura: El algoritmo es eficiente al computar los valores del histograma $H[T]$ una única vez

El algoritmo tiene complejidad $O(N(V + T)) \approx O(V)$, V número de vértices.

Support Vector Machine (SVM)

- Método supervisado: conjunto de entrenamiento y conjunto de prueba.
- Caso separable binario: puntos $\vec{x}_i \in \mathbb{R}^n$ que pertenecen a clase $y_i \in \{1, -1\}$.
- Dividas por el hiperplano $\langle \vec{w}, \vec{x} \rangle + b = 0$.



Entrenamiento

- Se cumplen las condiciones

$$\langle \vec{w}, \vec{x}_i \rangle + b \geq 1 \text{ para } y_i = +1, \quad (1a)$$

$$\langle \vec{w}, \vec{x}_i \rangle + b \leq -1 \text{ para } y_i = -1. \quad (1b)$$

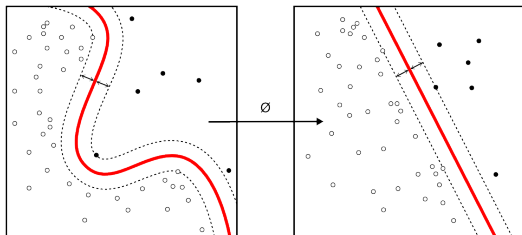
- Ello se combina como

$$y_i(\langle \vec{w}, \vec{x}_i \rangle + b) - 1 \geq 0 \quad \forall i. \quad (2)$$

- Los vectores de soporte son aquellos donde se da la igualdad.
- Éstos definen hiperplanos H_1, H_2 .
- La distancia entre éstos es $\frac{1}{\|\vec{w}\|}$.
- Minimizar $\|\vec{w}\|$ dada la restricción (2).

Prueba y Kernel

- Dado un punto \vec{x} , su clase es $\text{sgn}(\langle \vec{w}, \vec{x} \rangle + b)$.
- $\Phi : \mathbb{R}^n \rightarrow \mathcal{H}$ espacio de Hilbert.
- $K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}, \quad (\vec{x}, \vec{y}) \mapsto \langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle_{\mathcal{H}}.$
- $K(\vec{x}, \vec{y}) = (\langle \vec{x}, \vec{y} \rangle + 1)^p$ da un clasificador polinomial de grado p .

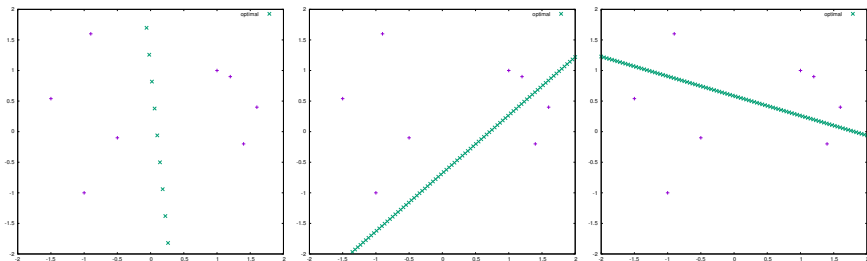


Observaciones

- La cantidad de familias es fija. Todos los elementos están forzados a caer en alguna de ellas.
- Todavía hay subjetividad al momento de determinar el conjunto de pruebas.
- Se necesitan muchos datos para que la máquina deduzca el patrón correctamente. Se recomienda una proporción entrenamiento:prueba de 7:3.

USVM: SVM No Supervisado en el caso separable sencillo

- **Problema de Márgen Máximo (MMP):** Consideramos todas las etiquetas posibles. Elegimos la que maximiza el margen con SVM tradicional.
- **Problema del Hiperplano Más Lejano (FHP):** Suponemos que el plano pasará por el origen.



En promedio, estamos bien

Hallamos un \vec{w} tal que en promedio determina un hiperplano que está alejado de la mayoría de puntos.

- 1: **Input:** $\{\vec{x}_i\}_{i=1}^n \in \mathbb{R}^d, ||\vec{x}_i|| \leq 1$
- 2: $\tau_1(i) = 1$ ▷ pesos iniciales
- 3: $j = 1$ ▷ no. de iteración
- 4: **while** $\sum_i \tau_j(i) \geq \frac{1}{n}$ **do**
- 5: $A_j \leftarrow$ matriz $n \times d$ cuya i -ésima columna es $\sqrt{\tau_j(i)}\vec{x}_i$
- 6: $\vec{w}_j \leftarrow$ vector singular derecho principal de A_j .
- 7: $\sigma_j(i) = |\langle \vec{x}_i, \vec{w}_j \rangle|$ ▷ márgenes
- 8: $\tau_{j+1}(i) = c^{-\sigma_j^2(i)} \tau_j(i)$ ▷ aligeramos
- 9: $++j$
- 10: $\vec{w}' = \sum_j g_j \vec{w}(j)$ ▷ $g_j \mathcal{N}(0, 1)$
- 11: **return** $\vec{w} = \vec{w}' / ||\vec{w}'||$

Análisis de datos

El principal problema afrontado fue dar una nueva clasificación al conjunto de 128 máscaras digitalizadas por el Instituto Nacional de Antropología e Historia (INAH.) Acorde a la clasificación de máscaras manejada por el INAH, las 128 se dividen en 9 familias distintas, por lo que se buscó dar una clasificación en 9 grupos.



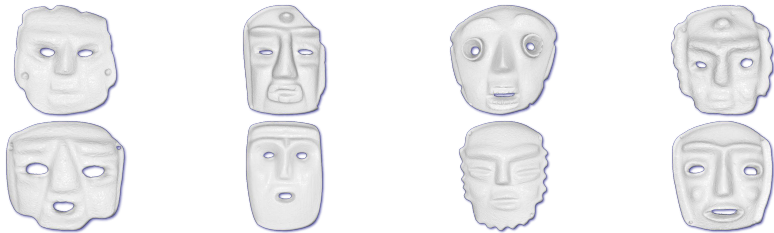


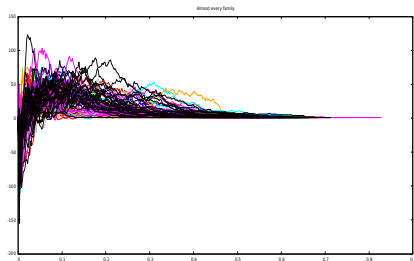
Figura: Familia 2 en la clasificación **original**



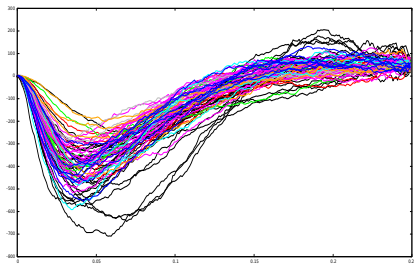
Figura: Muestra de la familia 9 en la clasificación **original**

Primeras GCEs

Al usar curvaturas como filtraciones, las GCEs asociadas de las máscaras no muestran patrones claros.



(a) Curvatura media



(b) Índice de Forma

Figura: Gráficas de CE para curvatura media e Índice de Forma en $T = 256$ umbrales. Cada una de las 9 familias originales fue trazada con un color distinto

Las proyecciones como filtración

Aprovechamos que cada máscara está encajada en el cubo $[-1, 1]^3$ con centro de masa en el origen. Las filtraciones fueron las distancias de cada vértice a los planos $x = 1$, $y = 1$, $z = 1$.



(a) Horizontal

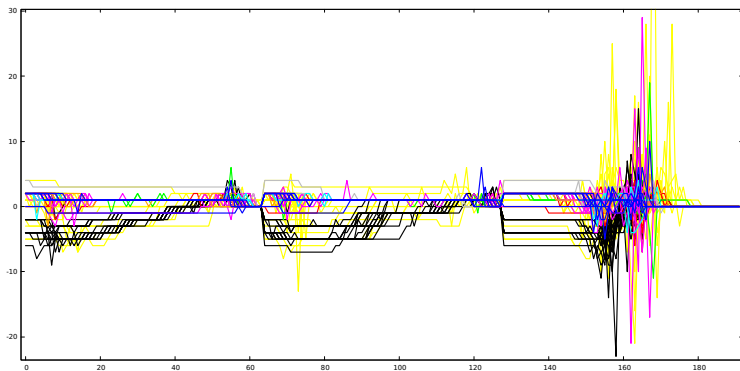


(b) Vertical



(c) Frontal

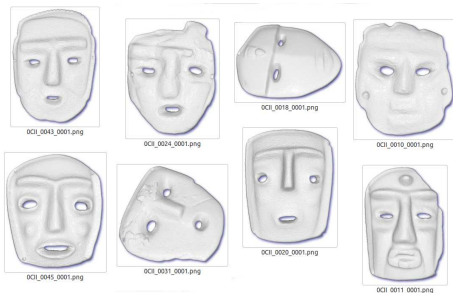
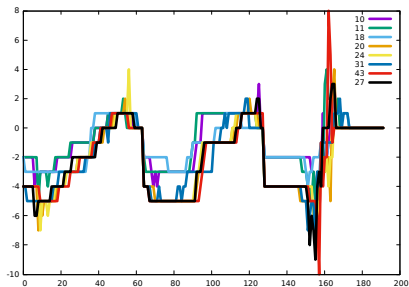
La GCE a partir de la concatenación de las **tres proyecciones principales** con 64 umbrales por proyección provee de un mejor prospecto para obtener una clasificación coherente de objetos.



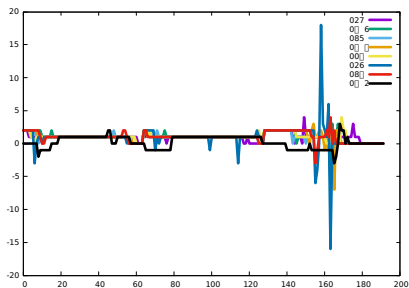
De las GCEs al SVM

- El SVM usó la mitad del conjunto de máscaras como entrenamiento y el resto como prueba. Se obtuvo una nueva clasificación.
- El número de especímenes por familia es más homogéneo. Ahora únicamente dos de las nueve familias contiene menos de 10 representantes.
- De los siete grupos restantes, se eligieron 8 representantes de cada una y se graficaron sus GCEs.
- Colores distintos se refieren a items distintos.

Familia 2 (clasificación nueva)



Familia 3 (clasificación nueva)



OCII_0027_0001.png



OCII_0085_0001.png



CIII_0087_0001.png



CIII_0007_0001.png



OCII_0036_0001.png



OCII_0099_0001.png

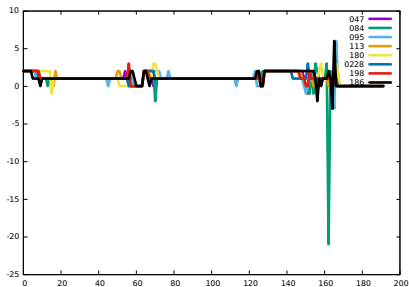


CIII_0092_0001.png



CIII_0026_0001.png

Familia 5 (clasificación nueva)



OCII_0047_0001.png



OCII_0095_0001.png



CIII_0198_0001.png



OCII_0180_0001.png



OCII_0084_0001.png



OCII_0113_0001.png

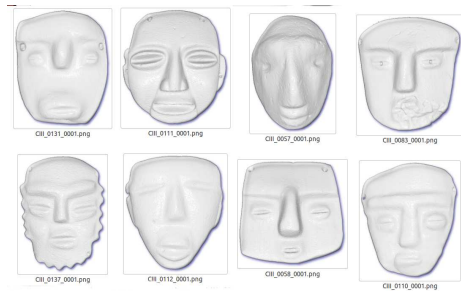
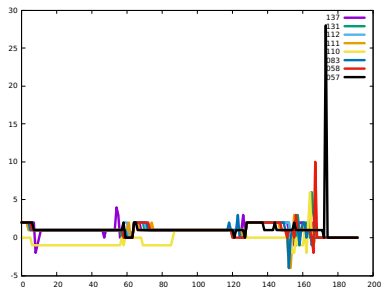


CIII_0228_0001.png



0020_0075_0001.png

Familia 9 (clasificación nueva)



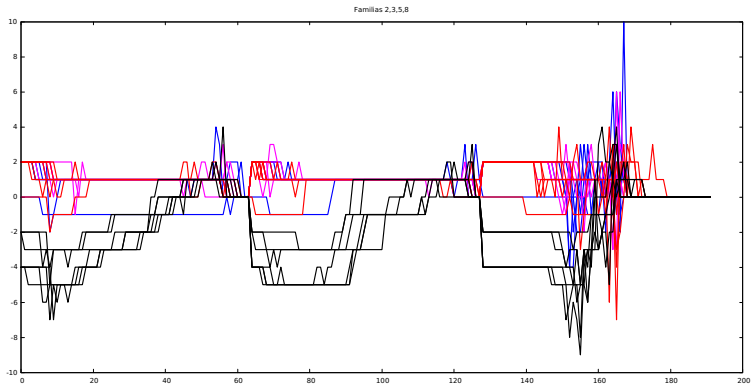


Figura: GCEs de las cuatro familias previas después de remover *outliers*.

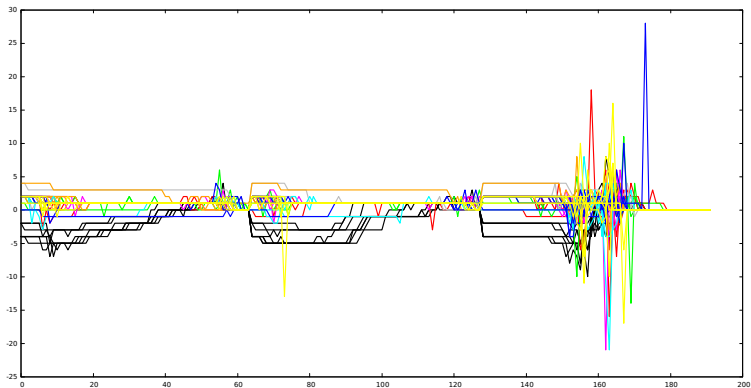


Figura: GCEs tomando 8 miembros de cada familia acorde a la nueva clasificación, eliminando aquellos cuya GCE sobresalga del resto.

De las GCEs al USVM

El caso más fácil sí funciona:

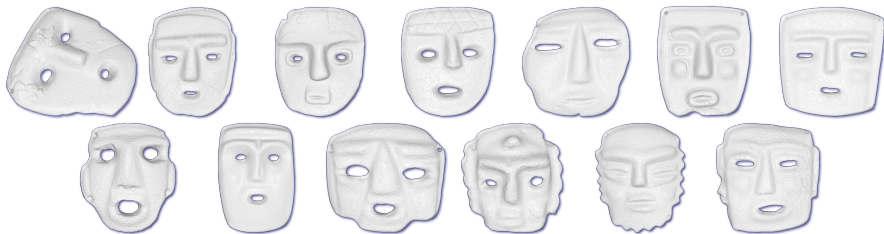


Figura: Familia 2 en la clasificación **original**



Figura: Familia 9 en la clasificación **original**

Resultado esperado

Optimal vector w:
comparing 11 vs 19

Margin: 0.2332

Vector 1: -

Vector 2: -

Vector 3: -

Vector 4: -

Vector 5: -

Vector 6: +

Vector 7: -

Vector 8: -

Vector 9: -

Vector 10: -

Vector 11: -

Vector 12: +

Vector 13: -

Vector 14: +

Vector 15: +

Vector 16: +

Vector 17: +

Vector 18: +

Vector 19: +

Vector 20: +

FAMILY 2 vs FAMILY 9 using T =

Con otras familias se porta extraño



Figura: Muestra de la familia 3 en la clasificación **original**



Figura: Muestra de la familia 4 en la clasificación **original**

Una a una vamos excluyendo



Figura: Exclusión ordenada

Un ejemplo más



Figura: Familia 5 en la clasificación **original**



Figura: Muestra de la familia 4 en la clasificación **original**

Una a una vamos excluyendo

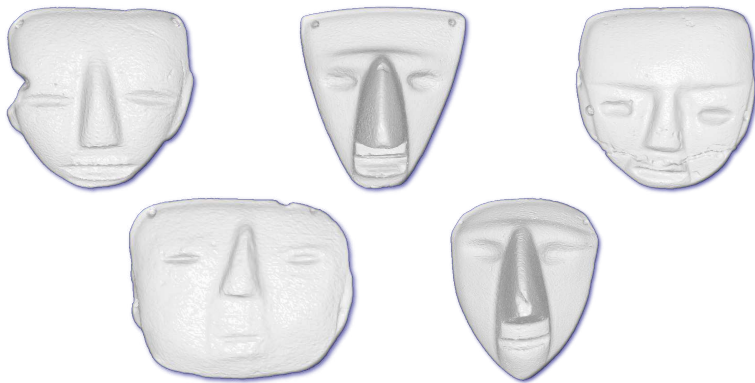







Figura: Exclusión ordenada

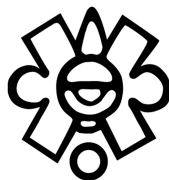
Comentarios Finales

- El cómputo de la GCE es una operación sencilla de complejidad y memoria lineales. 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.

Referencias

-  M.A. Armstrong *Basic Topology* Springer-Verlag (1983).
-  C. Burges "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Mining and Knowledge Discovery* Vol.2 pp.121-167, 1998.
-  Z. Karnin et. al., "Unsupervised SVMs: On the Complexity of the Furthest Hyperplane Problem." *JMLR: Workshop and Conference Proceedings* Vol. 23 (2012) 2.1 - 2.17
-  V. Paulsen, M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces* Cambridge University Press (2016).
-  E. Richardson, M. Weirman, "Efficient classification using the Euler Characteristic". *Pattern Recognition Letters* Vol.49, pp.99-106, 2014.

Agradecimientos



Instituto Nacional
de Antropología
e Historia