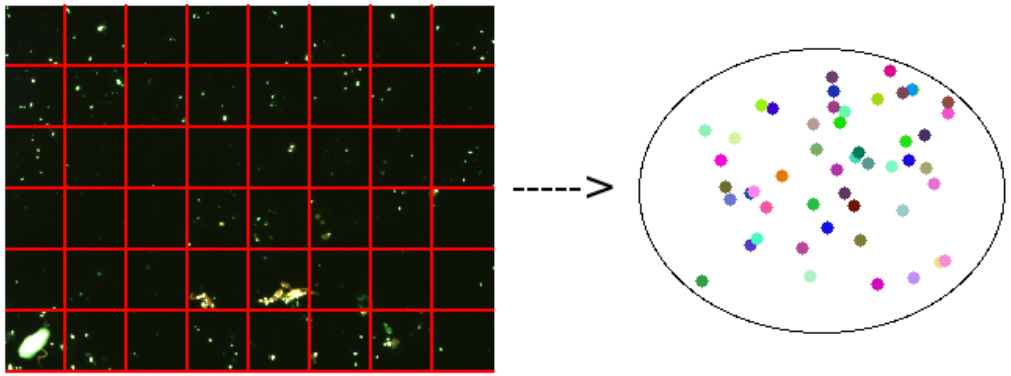
Clasificación de pacientes infecciosos de tuberculosis mediante métodos núcleo basados en la divergencia de Kullback-Leibler

Jair Montoya Martínez
Antonio Artés Rodríguez
Universidad Carlos III de Madrid
Departamento de Teoría de la Señal y Comunicaciones

26 de noviembre de 2010

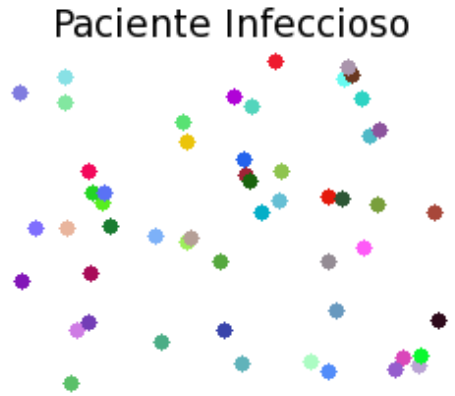
Clasificación de pacientes infecciosos de tuberculosis

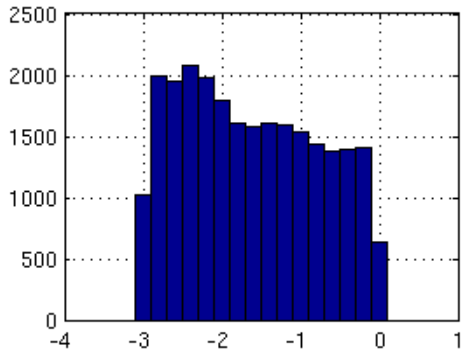
- De cada paciente se tiene un **conjunto de imágenes**.
- Cada imagen se divide en **parches**, los cuales se analizan con un **método máquina** para determinar la presencia o ausencia de bacilo.

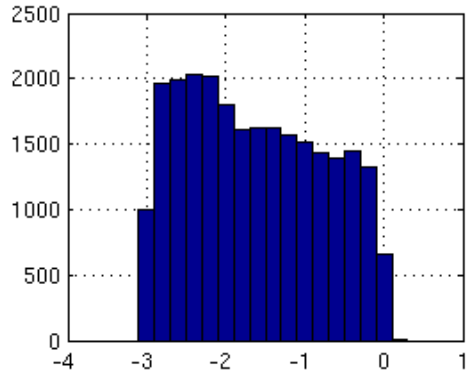


- Como resultado de este proceso se obtiene un **conjunto de números** por paciente.

- El conjunto de números de un paciente infeccioso **es muy parecido** al conjunto de números de un paciente no infeccioso.

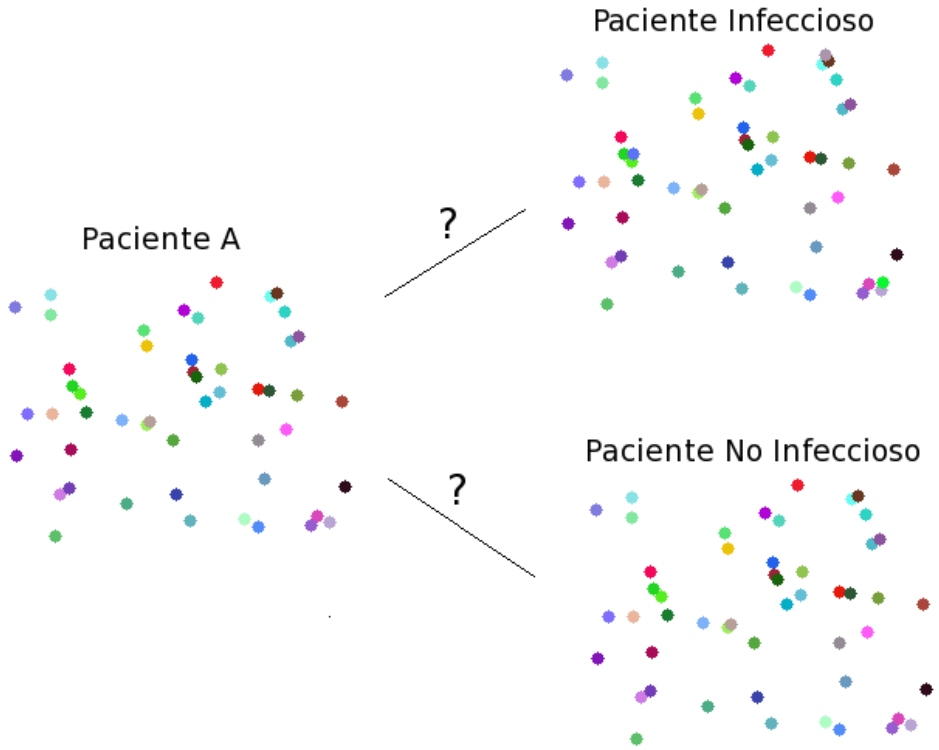






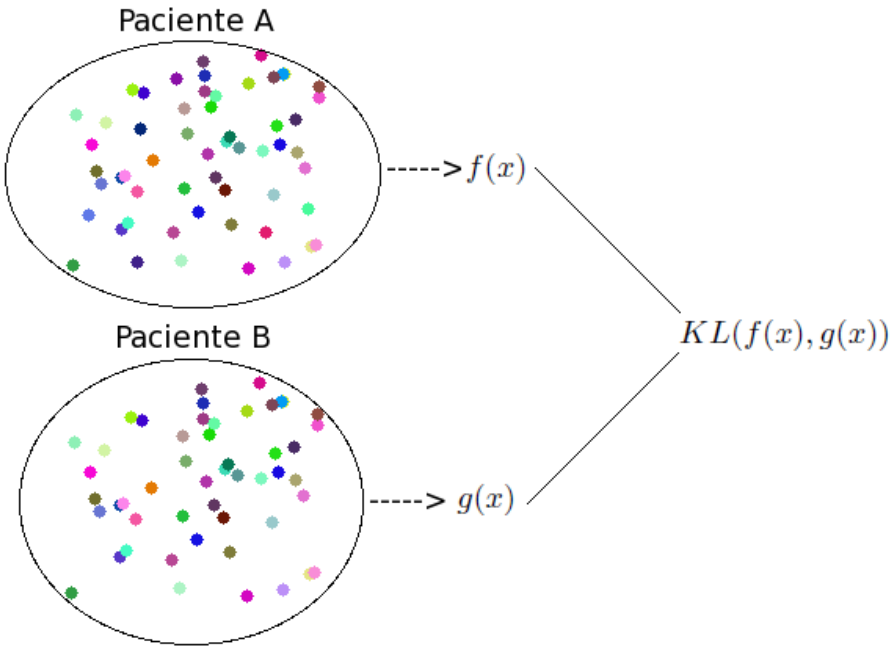
- Bajo estas condiciones, **¿Cómo clasificar óptimamente a un paciente?.**

Esquema de la Solución Propuesta

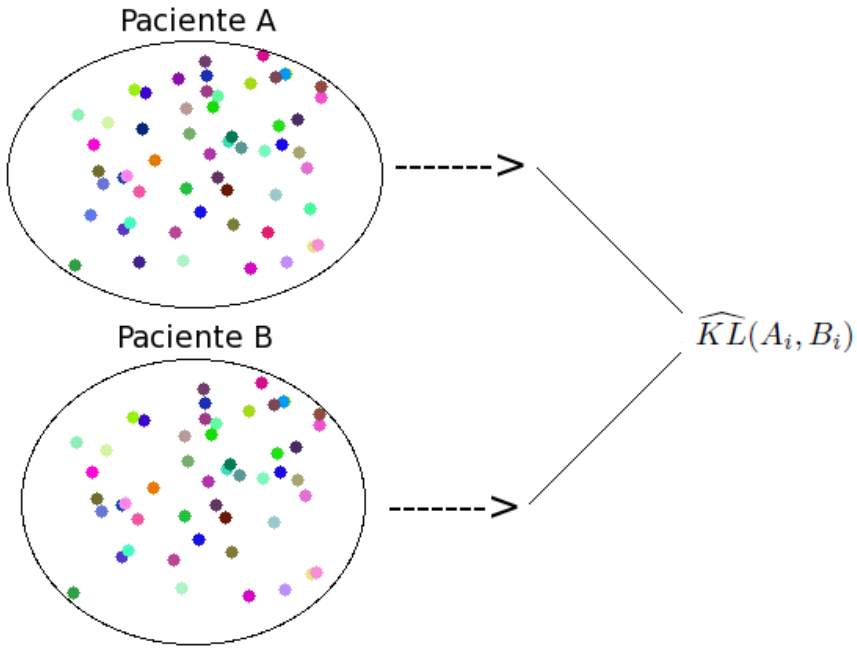


- Usaremos un **método núcleo** basado en **divergencia de Kullback-Leibler (KL)** para clasificar el estado de un paciente.

- **Opción 1**



- **Opción 2**



- Nosotros trabajaremos con la **Opción 2**.

Núcleo de Kullback-Leibler

- El **núcleo de Kullback-Leibler** se define como [2]:

$$KLK(f_i(x), f_j(x)) = e^{-aJ(f_i(x), f_j(x))+b} \tag{1}$$

- $J(f_i(x), f_j(x))$ es la **divergencia simétrica de Kullback-Leibler**.

$$J(f_i, f_j) = KL(f_i(x), f_j(x)) + KL(f_j(x), f_i(x)) \tag{2}$$

- $KL(f_i(x), f_j(x))$ es la **divergencia de Kullback-Leibler** entre las densidades $f_i(x)$ y $f_j(x)$:

$$KL(f_i(x), f_j(x)) = \int_{\mathbb{R}^D} f_i(x) \log \frac{f_i(x)}{f_j(x)} dx \geq 0 \tag{3}$$

Versiones existentes del Núcleo de Kullback-Leibler

- Se basan en **aproximar la divergencia KL** presente en el núcleo.
  - **Basado en familias exponenciales**:

$$KL(f_i, f_j) = a(\theta_i) - a(\theta_j) + [b(\theta_i) - b(\theta_j)]\, E_{\theta_i}[c(x)]$$

    donde f_i, f_j son miembros de una familia exponencial

$$f(x|\theta) = \alpha(x) \exp(a(\theta) + b(\theta)c(x))$$

  - **Basado en una aproximación** χ^2: Se transforma la divergencia en un **estadístico** χ^2 linealizando el **log** alrededor de $x = 1$.

- **Basado en Histogramas**: Dados dos histogramas π^i y π^j, la divergencia KL entre ellos se define como:

$$KL(\pi^i, \pi^j) = \sum_{k=1}^{b} \pi_k^i \log \frac{\pi_k^i}{\pi_k^j}$$

- **Basado en Monte-Carlo**: Cuando no existen expresiones cerradas para la divergencia KL podemos recurrir a una aproximación Monte-Carlo:

$$KL(f(x|\theta_i), f(x|\theta_j)) \approx \frac{1}{s} \sum_{m=1}^{s} \log \frac{f(x_m|\theta_i)}{f(x_m|\theta_j)}$$

Núcleo de Kullback-Leibler usando K-NN

- Proponemos una **nueva versión** del núcleo KL basado en **vecinos próximos**:

$$\widehat{KL}(f_i(x), f_j(x)) = \frac{1}{n}\sum_{p=1}^{n} \log \frac{\hat{f}_i(x_p^{(i)})}{\hat{f}_j(x_p^{(i)})} \tag{4}$$

- Según experimentos realizados en [3], la estimación de (4) para un valor $k = 1$ es la que converge más rápidamente al valor verdadero.

$$\widehat{KL}(f_i, f_j) = \frac{D}{n}\sum_{p=1}^{n} \log \frac{d_Q(x_p^{(i)})}{d_P(x_p^{(i)})} + \log \frac{m}{n-1} \tag{5}$$

Base de Datos usada en las Pruebas

- La base de datos utilizada consta de **46 pacientes**: **11 infecciosos** y **35 no infecciosos**.
- Tiene un total de **897** imágenes (**424** de pacientes infecciosos + **473** de pacientes no infecciosos).
- Se realizó la siguiente división de los datos:
  - Entrenamiento: **9** pacientes infecciosos y **20** pacientes no infecciosos.
  - Test: **2** pacientes infecciosos y **15** pacientes no infecciosos.

Resultados Obtenidos

- **Usando nuestro método**:

Clasificador	Sens.	Espec.
Weighted SVM (KL-KNN)	0.5	0.73
SVM-WHM (KL-KNN)	0.5	1.0
SVM CV-F (KL-KNN)	1.0	1.0

Tabla 1. Sensibilidad y Especificidad.

- **Comparación con otros Métodos**:

Clasificador	Sens.	Espec.
SVM CV-F (KL-Histograma)	1.0	0.6
SVM CV-F (KL-Monte-Carlo)	1.0	0.8
Test Secuencial	1.0	1.0

Tabla 2. Sensibilidad y Especificidad.

Conclusiones

- Se ha presentado un **nuevo método** para atacar el problema de clasificación de pacientes de tuberculosis basado en **métodos núcleo** y **teoría de la información**.
- Puede ser aplicado **directamente** a un **conjunto de estimas muestrales** sin la necesidad de construir explícitamente un modelo probabilístico de los datos.
- Puede usarse en datos de **diferente longitud** (diferente número de muestras).
- Es **computacionalmente eficiente** ya que sólo requiere las **distancias a los vecinos más próximos** de cada muestra.

Bibliografía

R. Santiago-Mozos and A. Artés-Rodríguez.
Uncertainty Based Censoring Scheme in Distributed Detection Using Learning Techniques, 2006.

P.J. Moreno, P.P. Ho and N. Vasconcelos.
A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications, 2003.

F. Pérez-Cruz.
Estimation of Information Theoretic Measures for Continuous Random Variables, 2008.

S.-X. Du and S.-T. Chen.
Weighted Support Vector Machine for Classification, 2005.

B. Li, J. Hu and K. Hirasawa.
Support Vector Machine Classifier with WHM Offset for Unbalanced Data, 2008.