ANÁLISIS DE LA SEPARABILIDAD LINEAL DE UNA BASE DE DATOS
Data base linear separability analysis

RESUMEN
En este artículo se presenta un análisis de diferentes técnicas (algorítmicas y estadísticas) para auscultar la separabilidad lineal de clases en una base de datos.
Para una base de datos generada artificialmente se describen y analizan las siguientes pruebas de separabilidad lineal: algoritmo de Schlesinger-Kozinec, análisis discriminante lineal y análisis de conglomerados.
Finalmente se discute con base a la información resultante al aplicar cada una de las técnicas descritas los alcances de cada una de ellas.

PALABRAS CLAVES: Separabilidad lineal, algoritmo de Schlesinger-Kozinec, análisis discriminante lineal, análisis de conglomerados.

ABSTRACT
In this article is presented a comparative study between different methods (algorithmical and statistical) that allow determining thelinear class separability in a database.
The following linear separability methods are described and analyzed: Schlesinger-Kozinec's algorithm, linear discriminant analysis and cluster analysis.
Finally, the results and scope of each previously described methods are discussed.

KEYWORDS: *Linear separability, Schlesinger-Kozinec's algorithm, linear discriminant analysis, cluster analysis*

JORGE RIVERA
Ingeniero Electrónico, M. Sc.
Profesor Asistente
Universidad Tecnológica de Pereira
j.rivera@utp.edu.co

JOSE SOTO MEJIA
Fisico, Ph.D.
Profesor Titular
Facultad de Ingeniería Industrial
Universidad Tecnológica de Pereira
jomejia@utp.edu.co

WILLIAN ARDILA
Fisico, M.Sc
Profesor Asociado
Universidad Tecnológica de Pereira
willianar@utp.edu.co

1. INTRODUCCIÓN

Un sistema de clasificación automática de patrones consta principalmente de las siguientes etapas [1]:

- Adquisición
- Bases de datos
- Preprocesamiento y/o segmentación
- Extracción de características
- Selección de características
- Validación y entrenamiento
- Clasificación final

La conformación de las bases de datos consiste básicamente en tomar todas las señales recolectadas en la etapa de adquisición y remitirlas a un especialista en el tema que se está tratando con el objeto de que sean etiquetadas para su posterior utilización en las etapas de extracción de características, entrenamiento y validación.

Lo anterior hace énfasis en la importancia de tener un amplio conocimiento de la estructura y composición de una base de datos en un proceso de clasificación automática de datos. La clasificación de datos está íntimamente ligada a los problemas de separabilidad.
De lo anterior la importancia de auscultar y precisar el alcance de diferentes técnicas que permitan investigar la separabilidad de una base de datos.

Este articulo describe y analiza las siguientes pruebas algorítmicas y estadísticas de separabilidad lineal: algoritmo de Schlesinger-Kozinec, análisis discriminante lineal y análisis de conglomerados.
Finalmente se discute con base a la información resultante al aplicarlas a una base de datos generada artificialmente los alcances de cada una de ellas.

Fecha de Recepción: 07 Septiembre de 2007
Fecha de Aceptación: 11 Diciembre de 2007

2. FUNDAMENTOS TEÓRICOS

2.1 Separabilidad lineal

Una variable y (variable dependiente, clase, o salida) depende de un conjunto de p variables independientes. $(x_1, \dots, x_p)$. Estas variables toman valores en los conjuntos X y Y respectivamente, donde los Y indican la pertenencia del conjunto X en una u otra clase en particular $\{-1,1\}$ [2].

El problema es determinar una regla, de la forma

$$y = f(\boldsymbol{x})$$

para el valor de $\boldsymbol{y}$ observado correspondiente a un $\boldsymbol{x} \in X$.

Este problema es común en áreas tales como la clasificación automática de patrones [2].
Un conjunto de datos se dice que es linealmente separable si existe un hiperplano que separe efectivamente los datos pertenecientes a cada una de las diferentes clases, de lo contrario se dice que la base de datos no es separable linealmente.

2.2 Algoritmo de Schlesinger-Kozinec [3]

El hiperplano de separación óptimo entre los conjuntos X_1, X_2 está determinado por los puntos más cercanos de las regiones convexas $\bar{X}_1$, $\bar{X}_2$ ver figura 1. Esto es

$$\omega^* = \omega_1^* - \omega_2^*, \quad (1)$$

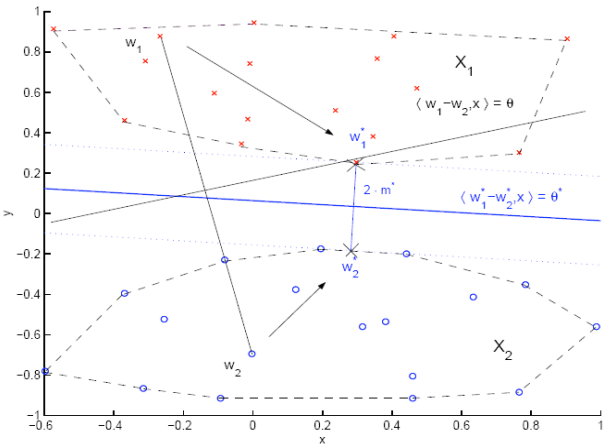

Figura 1. Ejemplo de los puntos más cercanos de dos áreas convexas

Donde ω representa un vector normal al hiperplano y el margen de un hiperplano separador es la distancia entre los puntos más cercanos la cual esta dado por la ecuación (2)

$$m(\omega,\theta) = \min\left(\min_{i\in I_1}\frac{\langle\omega,x_i\rangle-\theta}{\|\omega\|}, \min_{i\in I_2}\frac{\langle\omega,x_i\rangle-\theta}{\|\omega\|}\right) \quad (2)$$

Y el hiperplano de separación óptimo $\langle\omega^*,\boldsymbol{x}\rangle = \theta^*$ que maximiza el margen esta dado por la ecuación (3).

$$(\omega^*,\theta^*) = \max_{\omega,\theta} m(\omega,\theta) \quad (3)$$

Un hiperplano de separación ε-óptimo satisface además la siguiente ecuación:

$$m(\omega^*,\theta^*) - m(\omega,\theta) \le \varepsilon. \quad (4)$$

Donde

$$\theta^* = \frac{1}{2}(\|\omega_1^*\|^2 - \|\omega_2^*\|^2) \quad (5)$$

$$(\omega_1^*,\omega_2^*) = \min_{\omega_1\in\bar{X}_1,\,\omega_2\in\bar{X}_1}\|\omega_1-\omega_2\|$$

El algoritmo de Schlesinger-Kozinec busca iterativamente los puntos más cercanos de las regiones convexas $\bar{X}_1$, $\bar{X}_2$ hasta que la condición de ε-optimalidad sea alcanzada.

La condición de ε-optimalidad $m(\omega^*,\theta^*) - m(\omega,\theta) \le \varepsilon$ se alcanza si se cumple [2] la condición dada en la ecuación (6).

$$\frac{1}{2}\|\omega_1-\omega_2\| - \min\left(\min_{i\in I_1}\frac{\langle\omega,x_i\rangle-\theta}{\|\omega\|}, \min_{i\in I_2}\frac{\theta-\langle\omega,x_i\rangle}{\|\omega\|}\right) \le \varepsilon \quad (6)$$

En la figura 2, se muestra un ejemplo de un hiperplano de separación óptimo y uno ε-óptimo [2].

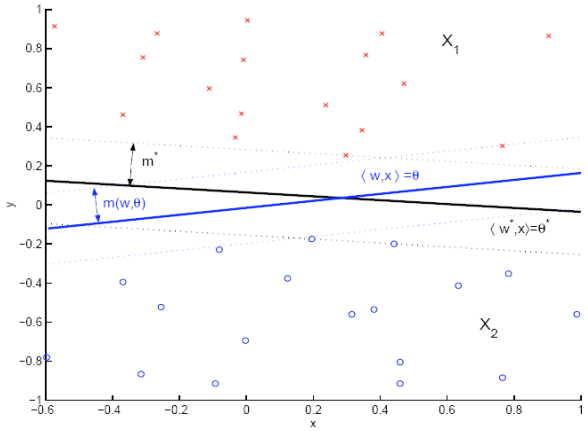

Figura 2. Hiperplano de separación óptimo y ε-óptimo

2.3 El Análisis discriminante

En el análisis discriminante [4] se persigue obtener una serie de funciones lineales (llamadas funciones discriminantes) a partir de las variables independientes que permitan interpretar las diferencias entre los grupos y poder clasificar posteriormente a un nuevo individuo cuyo grupo se desconoce para en función de un resultado numérico, ser asignado a alguna de las subpoblaciones (grupos) definidos previamente por la variable dependiente.

La función discriminante de Fisher D, se obtiene como una función lineal de k variables explicativas como:

$$D = u_1X_1 + u_2X_2 + \cdots + u_kX_k \quad (7)$$

El problema matemático consiste en obtener los coeficientes de ponderación ***uj***, que mejor permitan diferenciar los individuos de un grupo del otro. Si consideramos que existen *n* observaciones, se puede expresar la función discriminante para cada una de las *n* observaciones como en la ecuación (8)

$$D_i = u_1X_{1i} + u_2X_{2i} + \cdots + u_kX_{ki} \quad i = 1,2,\ldots,n \quad (8)$$

D_i -es la puntuación discriminante correspondiente a la observación *i-ésima*. En notación compacta se puede escribir como en la ecuación matricial (9) .

$$\boldsymbol{D} = \boldsymbol{Xu} \quad (9)$$

La variabilidad de la función discriminante (es decir, la suma de los cuadrados de las variables discriminantes en desviaciones respecto a su media) se expresa en la ecuación (10) como:

$$d'd = u'X'Xu \quad (10)$$

La matriz $X'X$ al estar expresada las variables en desviaciones respecto a la media, es la matriz de suma de cuadrados y productos cruzados total (*SCPC*) de las variables explicativas X. Esta matriz, $X'X$, se puede descomponer en la suma de la matriz *SCPC* entre grupos F y la matriz *SCPC* residual o intragrupos V, como se muestra en la ecuación (11).

$$X'X = T = F + V \quad (11)$$

Donde la matriz T es la matriz de la suma de cuadrados y productos cruzados total.

Por tanto sustituyendo (11) en (10) se obtiene (12)

$$d'd = u'X'Xu = u'Tu = u'Fu + u'Wu \quad (12)$$

Observe que en la expresión (12) T, F, y W se pueden calcular con los datos muestrales, mientras que los coeficientes $\boldsymbol{u_j}$ están por ser determinados.

El criterio de Fisher trata de determinar el eje discriminante (el vector u) de forma que las distribuciones proyectadas sobre el mismo estén lo mas separadamente posible entre sí (mayor variabilidad entre grupos) y al mismo tiempo, que cada una de las distribuciones esté lo menos dispersa (menor variabilidad dentro los grupos). Analíticamente el criterio se expresa de la siguiente forma:

$$Maximizacion\ de\ \lambda = \frac{u'Fu}{u'Wu} \quad (13)$$

La solución a éste problema se obtiene derivando λ_1 respecto a u e igualando a cero, como en la ecuación (14)

$$\frac{\partial \lambda_1}{\partial u_1} = \frac{2Fu_1\left(u_1^T W u_1\right) - 2Wu_1\left(u_1^T F u_1\right)}{\left(u_1^T W u_1\right)^2} = 0 \quad (14)$$

De donde $W^{-1}Fu_1 = \lambda_1 u_1$

Así que los ejes discriminantes ***(u)*** vendrán dados por los vectores propios asociados a los valores propios de la matriz $W^{-1}F$ ordenados de mayor a menor. Las puntuaciones discriminantes son pues los valores que se obtienen al dar valores a X1, X2, …, Xk en la ecuación (8) y se corresponden con los valores obtenidos al proyectar cada punto del espacio k-dimensional de las variables originales sobre el eje discriminante.

Dado que λ_1 es el cociente de maximizar a ecuación (13) ello mide el poder discriminante del primer eje, $\boldsymbol{u_1}$.

Selección de las variables discriminantes

Los criterios que se consideran para la selección de las variables independientes a hacer parte de la función discriminante son el Lambda Wilks (λ) y el estadístico F.

Lambda de Wilks para un conjunto de P variables independientes, mide las desviaciones dentro de cada grupo respecto a las desviaciones totales sin distinguir los grupos y está dado por la ecuación (15).

$$\lambda = \frac{S.C\_intragrupos}{S.C.totales} \quad (15)$$

Donde en el numerador se encuentra la suma de cuadrados intra grupos, la cual mide el cuadrado de las desviaciones de cada uno de los datos con respecto a la media del grupo. En el denominador se encuentra la suma de cuadrados totales, la cual mide el cuadrado de las desviaciones de cada uno de los datos con respecto a la media de todos los datos (media total sin discriminar grupos).

La suma de cuadrados totales es el resultado de la suma de cuadrados inter-grupos (entre los grupos) y intra-grupos (dentro de los grupos).

Si el valor de λ es próximo a 1, los grupos estarán mezclados y el conjunto de variables independientes no será adecuado para construir la función discriminante, dado que el mayor porcentaje de variabilidad estaría siendo representado por la variabilidad dentro de los grupos.

2.4 El Análisis de conglomerados

El análisis cluster es un método estadístico multivariante de clasificación automática de datos. Su finalidad esencial es revelar concentraciones en los datos para su agrupamiento adecuado en conglomerados. Los elementos dentro de un conglomerado deben ser homogéneos entre si y lo mas diferentes posible de los contenidos en otros clusters. Es decir los grupos se crean en función de la naturaleza de los datos. En un principio se seleccionan tantos individuos como conglomerados hayamos solicitado de modo que estos individuos iniciales tengan distancia máxima entre ellos y al estar separados lo suficientemente produzcan los centros iniciales. Una vez estimados los centros iniciales de los conglomerados se calcula la distancia de cada punto (dato) a cada uno de ellos y en función de la mínima distancia obtenida se irán clasificando el resto de los individuos en los conglomerados iniciales.

En [4], los autores proponen un método de clasificación mediante el análisis de conglomerados, el cuál encuentra un estimado $\boldsymbol{y}$ a partir de un conjunto de entrenamiento $\mathcal{D}$ organizado en conglomerados y aplicando reglas de distancia.

Un ejemplo de lo anterior es el caso binario, donde $\mathcal{D}$ se divide en dos conglomerados $\mathcal{C}_{-1}$ y $\mathcal{C}_1$, los cuales poseen

1 Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos

medias $(\bar{x}_{-1}, \bar{y}_{-1})$ y $(\bar{x}_1, \bar{y}_1)$, respectivamente, entonces la regla es:

$$y = \begin{cases} -1, si\ d(x, \bar{x}_{-1}) < d(x, \bar{x}_1) \\ 1, si\ lo\ anterior\ no\ se\ cumple \end{cases} \quad (16)$$

Donde d es una métrica en X.

Métricas usadas en análisis de conglomerados

Se asume que una función de distancia $d(\cdot,\cdot)$ está definida en $X \times Y$. Esta distancia puede ser construida a partir de distancias d_X y d_Y, definidas en X y Y respectivamente, por ejemplo

$$d((x_1, y_1)(x_2, y_2)) = \sqrt{d_X^2(x_1, x_2) + \alpha d_Y^2(y_1, y_2)} \quad (17)$$

Donde el parámetro $\alpha \geq 0$ mide la importancia relativa de los componentes y [4].

Para $X \subset \mathbb{R}^p$ se puede utilizar la distancia euclidiana,

$$d_X(x_1, x_2) = \sqrt{(x_1 - x_2)^T (x_1 - x_2)} \quad (18)$$

3. IMPLEMENTACIÓN DE LAS TÉCNICAS

Se utilizó una base de datos artificial con dos variables independientes, V1, V2, y una tercera variable, V3, indicando la pertenencia al grupo 1 o al grupo 2).

3.1 El algoritmo de KOZINEC

Las siguientes fueron las instrucciones del toolbox "Stprtooll"[1] para MatLab™ utilizadas para generar en forma de estructura la base de datos con los 100 individuos:

```
data=genlsdata(2,100)
data =
   X: [2x100 double]
   y: [1x100 double]
```

La siguiente instrucción grafica las dos clases de datos generados (ver figura 3).

```
ppatterns(data)
```

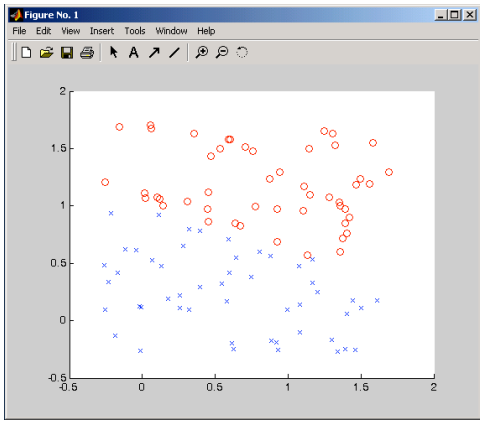

Figura 3. Conjunto de datos generado

Las siguientes instrucciones utilizan el algoritmo Schlensinger-Kozinec para generar el mejor hiperplano en 1000 iteraciones.

```
model=ekozinec(data,struct('eps',0.01,'tmax',1000))
model =

W1: [2x1 double]   //vector más cercano de la 1a área convexa
 W2: [2x1 double]   //vector más cercano de la 1a área convexa
      t: 355      // número de iteraciones para converger
      exitflag: 1
      b: 0.0256     //bias
      W: [2x1 double] //vector nomal al hiperplano
      margin: 0.0114  //distancia entre los dos vectores
                      mas cercanos de las áreas convexas
    fun: 'linclass'    //función lineal utilizada
```

La siguiente instrucción grafica (ver figura 4) el hiperplano que separa los datos que fueron presentados antes en la figura 3.

```
pline(model)
```

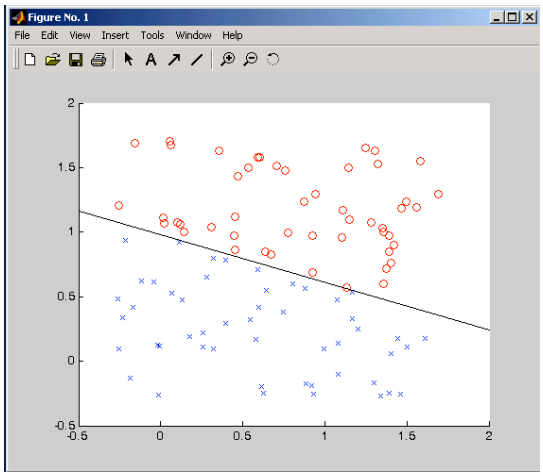

Figura 4. Hiperplano de separación producido por el algorítmo de Kozinec.

Como se observa en la figura 4 el algoritmo de Kozinec encontró un hiperplano que separa perfectamente las dos clases

3.2 El análisis discriminante

Para analizar mediante la técnica de análisis discriminante (usando el paquete estadístico SPSS) la misma base de datos en forma de estructura generada en la sección 3.1 ella fue primero convertida a un archivo de texto plano separado por tabuladores mediante las siguientes instrucciones:

```
datos_planos=[data.X', data.y']
save datos_Planos.txt datos_planos –ascii
```

[1] Statistical Pattern Recognition Toolbox: http://cmp.felk.cvut.cz/cmp/software/stprtool/index.html

Los siguientes fueron los comandos para ejecutar el procedimiento discriminante en SPSS:

análisis – clasificar – discriminante
variable de agrupación ***v3***
independientes: ***v1*** *y* ***v2***

USAR Método de inclusión por pasos
CLASIFICAR
PROBABILIDADES PREVIAS: calcular según tamaño de grupo

Los resultados fueron como se muestran en las tablas 1 a 7 siguientes. El cuadro de autovalores de la siguiente tabla 1 muestra un valor propio de **2.581** indicando que la función de discriminación atribuible a ese valor propio explica el 100% de la variabilidad total de la nube de puntos que se proyecte sobre la función de discriminación. El valor de la correlación canónica que mide las desviaciones de las puntuaciones discriminantes entre los grupos con respecto a las desviaciones totales sin distinguir grupo, muestra un valor de **0.849** mostrando que la dispersión de los datos en su mayor parte es debido a las diferencias entre los grupos y por lo tanto la función discriminará mucho.

Función	Valor propio	% de varianza	% acumulado	Correlación canónica
1	2,581(a)	100,0	100,0	,849

Tabla 1.Valor propio asociado con la función discriminante.

La siguiente tabla 2 muestra los coeficientes de la función discriminante.

	Función
	1
V1	,841
V2	3,198
(Constante)	-2,846

Tabla 2. Coeficientes de las funciones canónicas discriminantes.

Así que esta se escribe como:

$$D1= -2,846 + 0,841*V1 + 3,198*V2 \qquad (19)$$

La función discriminante evaluada en la media de cada uno de los dos grupos (en los centroídes) nos da un idea de cómo la función discrimina los dos grupos. Dado que el valor de la función discriminante en el primer grupo (ver tabla 3) es de -1,559 el cual es muy diferente del valor de la media en el segundo grupo que es de 1,623 la discriminación es muy buena como ya lo había asegurado el análisis hecho arriba para el valor propio.

	Función
V3	1
1,00	-1,559
2,00	1,623

Tabla 3. Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Los resultados de la clasificación mostrados en la tabla 4 muestran que la función discriminante encontrada arriba clasificó mal dos casos, uno del primer grupo y el otro perteneciente al segundo grupo.

			Grupo de pertenencia pronosticado		
		V3	1,00	2,00	Total
Original	Recuento	1,00	50	1	51
		2,00	1	48	49
	%	1,00	98,0	2,0	100,0
		2,00	2,0	98,0	100,0

Tabla 4. Resultados de la clasificación**(a)**

a Clasificados correctamente el 98,0% de los casos agrupados originales.

3.3 El análisis de conglomerados

Dado que el análisis de los datos por conglomerados (no jerárquico) requiere que se especifique el número de conglomerados a formar, se especificaron dos grupos con el objeto de comparar los grupos que se generarían automáticamente con éste procedimiento cluster, con los dos grupos que a priori ya se conocían en la base datos artificial generada.

En la tabla 5 se presentan los centros iniciales de los conglomerados los cuales se asignaron a los individuos que tenían la máxima distancia entre ellos y en función de la mínima distancia obtenida entre los otros individuos y éstos centros se fueron clasificado el resto de los individuos.

	Conglomerado	
	1	2
V1	-,19	1,58
V2	-,13	1,55

Tabla 5. Centros iniciales de los conglomerados

En la tabla 6 se presentan los centros de los dos conglomerados obtenidos al final del proceso y en la tabla 7 se presenta una parte de la listado de pertenencia de cada dato a su conglomerado conjuntamente con la distancia de cada uno ellos al centro del conglomerado.

1 Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos

	Conglomerado	
	1	2
V1	,38	1,03
V2	,36	1,03

Tabla 6. Centros finales de los conglomerados

Finalmente la tabla 7muestra que a un primer conglomerado fueron asignados 49 datos y al segundo conglomerado 51 datos lo que coincide exactamente con el tamaño de cada grupo en la base de datos artificial original..

Conglomerado	1	49,000
	2	51,000
Válidos		100,000
Perdidos		,000

Tabla 7. Número de casos en cada conglomerado

La comparación realizada uno a uno entre la clasificación original en la base de datos artificial y la asignación dada a cada uno de los 100 datos por la técnica cluster mostró discrepancia en 15 datos (15%).

4. DISCUSIÓN DE LOS RESULTADOS

El **algorítmo de Kozinec** determina la separabilidad lineal absoluta, la presencia de un solo dato que no consiga separarse hace que el algoritmo no converja, lo cual lo hace poco robusto al ruido.

El **análisis discriminante** lineal siempre proporcionará una función de separabilidad lineal, cuantificando al mismo tiempo la calidad de la función de discriminación a través del valor propio, la proporción de varianza explicada, el coeficiente de correlación canónica y el cálculo de los datos mal y bien clasificados. Además, la función discriminante permite determinar la pertenencia de una nueva observación a alguno de los grupos que ella está separando.

El **análisis cluster** proporciona una primera aproximación a la separación automática del conjunto de datos en grupos *previamente no conocidos* a diferencia del análisis discriminante y el algorítmo de Kozinec que exigen el conocimiento ***a priori*** de la pertenencia de cada uno de los datos a grupos determinados.

El análisis cluster, a diferencia del análisis discriminante no presenta a un procedimiento analítico para determinar la asignación de un nuevo dato a alguno de los conglomerados ya establecidos. Lo anterior debido a que los grupos se crean en función de la naturaleza de los datos que participan en su generación.

Todas las técnicas analizadas permiten auscultar la *separabilidad lineal* en una base de datos pero difieren claramente en el alcance que cada una de ellas realiza de la separabilidad lineal.

BIBLIOGRAFÍA

[1] J. Rivera, "Selección efectiva de características ECG mediante tecnicas de transformación no lineal: Identificación de infarto agudo del miocardio" disertación de maestría, Dept. Ciencias básicas, Maestría en instrumentación Física, UTP, 2006.

[2] A. B. Israel, Y. Levin "The geometry of linear separability in data sets" *Elsevier. Linear Algebra and its Applications,* 416, pp 75-87, 2006.

[3] F. Vojtech and H. Václav. Generalization of the Schlesinger-Kozinec's Algorithm for Supporto Vector Machines. Center for Machine Perception, CTU Prague, 2005.

[4] A. B. Israel, Y. Levin "An estimation algorithm using distance clustering data", OPSEARCH 3 8(2001)443-455.

[5] Técnicas de Análisis Multivariante de Datos. Cesar Pérez. Pearson, Prentice Hall, España 2004.

[6] Métodos Estadísticos Avanzados con SPSS. Cesar Pérez López. Thomson, España, 2005)