Revista de Matemática: Teoría y Aplicaciones 2011 **18**(1) : 21–32
cimpa – ucr issn: 1409-2433

# CURVAS ROC Y VECINOS CERCANOS, PROPUESTA DE UN NUEVO ALGORITMO DE CONDENSACIÓN

Raquel Jiménez–Padilla*
Carlos Cuevas–Covarrubias†

*Received: 18 Feb 2010; Revised: 3 Nov 2010; Accepted: 10 Nov 2010*

**Resumen**

Los criterios $k$-NN son algoritmos no paramétricos de clasificación estadística. Son precisos, versátiles y libres de distirbución. Sin embargo su costo computacional puede ser demasiado alto; especialmente con tamaños de muestra grandes. Presentamos un nuevo algoritmo de condensación que, basado en el modelo Binormal para curvas ROC, permite transformar la base de entrenamiento en un conjunto pequeño de vectores de baja dimensíón. A diferencia de otras técnicas descritas en la literatura, nuestra propuesta permite controlar el intercambio de precisión por reducción de la base de entrenamiento. Un estudio de Monte Carlo muestra que el desempeño del método popuesto puede ser muy competente, superando en diversos escenarios realistas al de otros métodos frecuentemente utilizados.

**Palabras clave:** clasificación estadística, área bajo la curva ROC, modelo binormal, vecinos cercanos, condensación, Monte Carlo.

**Abstract**

*Centro de Investigación en Estadstica y Matemáticas Aplicadas, Universidad Anáhuac, México. E-Mail: dirlem@hotmail.com

†Centro de Investigación en Estadstica y Matemáticas Aplicadas, Universidad Anáhuac, México. E-Mail: ccuevas@anahuac.mx

*k*-NN criteria are non parametric methods of statistical classificaction. They are accurate, versatile and distribution free. However, their computational cost may be too expensive; especially for large sample sizes. We present a new condensation algorithm based on the Binormal model for ROC curves. It transforms the training sample into a small set of low dimensional vetors. Contrasting with other condensation techniques described in the literature, our proposal helps to control the exchange of accuracy for condensation on the training sample. The results of a Monte Carlo study show that its performance can be very competitive in different realistic scenarios, resulting in better training samples than other frequently used methods.

**Keywords:** statistical classification, area under the ROC curve, nearest neighbours, condensation, Monte Carlo.

**Mathematics Subject Classification:** 62H30.

# 1 Introducción

Los algoritmos *k*-NN *(k-Nearest Neighbours)* son algoritmos no paramétricos de clasificación estadística. Debido a su versatilidad son un tema importante de investigación. En la literatura se han reportado algunas aplicaciones exitosas de esta técnica en diversos contextos. Henley & Hand [8] por ejemplo, describen una aplicación de vecinos cercanos para la evaluación del riesgo crediticio; comparan el desempeño de esta metodología con el de algunos modelos paramétricos encontrando ventajas competitivas importantes. Cuevas-Covarrubias et al [3] muestran una aplicación de los algoritmos *k*-NN en la que se predice el desempeño esperado de la fuerza de ventas en la industria farmacéutica; ésto con base en un perfil psicométrico y de competencias. Una desventaja competitiva en los algoritmos *k*-NN es su alto costo computacional. Existen diversas propuestas para atacar este problema, nosotros dedicaremos especial énfasis al trabajo de Guo et al [5]. Su idea consiste en condensar la base de entrenamiento convirtiéndola en un pequeño conjunto de registros de baja dimensión. Cada uno representa el centro y radio de una esfera más el número de observaciones contenidas en la misma, así como su categoría de origen. Su propuesta reduce significativamente el costo computacional de los algoritmos *k*-NN; sin embargo, no permite evaluar su presición estadística con facilidad. En este trabajo proponemos un algoritmo de condensación alternativo basado en la idea de Guo et al [5]. Nuestra propuesta añade dos parámetros adicionales que hacen posible la evaluación del criterio a través de un análisis de curvas *ROC*. Los resultados de un breve estudio de Monte Carlo son alentadores, pues sugieren que el nuevo algoritmo tiene mejor desempeño y mayor poder

de condensación bajo circunstancias específicas y realistas. Iniciamos en la sección 2 revisando algunos conceptos básicos de clasificación estadística y análisis de curvas ROC. En la sección 3 describimos brevemente los criterios de clasificación por vecinos cercanos. La sección 4 discute sobre el algoritmo de condensación $k$-NN Basado en un Modelo (también lo llamaremos $k$-NN *Model Based Approach*), propuesto originalmente por Guo et al [5]. Nuestra propuesta se define en la sección 5, nosotros la llamamos $k$-NN Controlado (o también Controlled $k$-NN). Finalmente, en la sección 6 evaluamos el funcionamiento de nuestra propuesta mediante un estudio de Monte Carlo comparando su desempeño con el del criterio basado en un modelo.

## 2 Clasificación estadística

Estudiamos una población $\Omega = \Omega_0 \cup \Omega_1$, siendo $\Omega_0\Omega_1 = \emptyset$. Observamos los elementos de $\Omega$ a través de un vector $\mathbf{V} : \Omega \to \mathbb{R}^p$. Un índice de riesgo $S$ es un resumen de la información contenida en $\mathbf{v}$; $S : \mathbb{R}^p \to \mathbb{R}$. Dado un $\omega \in \Omega$:

$$\text{Clasificamos en } \omega \text{ en } \begin{cases} \Omega_1 \text{ si } s > t \\ \\ \Omega_0 \text{ si } s \leq t. \end{cases} \tag{1}$$

El parámetro $t$ es un umbral de decisión calibrado por el usuario. La calidad de las reglas de clasificación se evalúa a partir de su sensiblidad y especificidad, o en términos de las tasas de falsos positivos y falsos negativos[1] (ver Hand [6]). La calidad discriminante de $S$ depende de todas las combinaciones de sensibilidad y especificidad posibles, dicha información se conoce como la curva *ROC* (ver Bamber [1]). Dado un índice de riesgo $S$, su curva *ROC* (*Receiver Operator Characteristic*) se define como:

$$ROC = \{(u, v)|u = 1 - F_0(t) \text{ y } v = 1 - F_1(t);\, t \in \mathbb{R}\},$$

en donde $F_i(t) = Pr[S \leq t|\omega \in \Omega_i]$. La curva *ROC* es la gráfica de la sensibilidad vista como función de la tasa de falsos positivos. Sintetiza gráficamente el desempeño potencial de $S$ como índice de riesgo. El área bajo la curva *ROC* resume toda esta información en un índice numérico que mide la calidad de $S$ (ver Bamber [1], Hanley & McNeil [7] y Zweig & Campbell [10]). Valores del área cercanos a 1 indican un alto potencial discriminante en el índice de riesgo. Si $F_0(t) = F_1(t) \forall t \in \mathbb{R}$ el índice $S$ es no informativo. En este caso el área bajo la curva *ROC* es igual a $\frac{1}{2}$.

[1] Complementos de la especificidad y sensibilidad respectivamente

## 3 Clasificación por vecinos cercanos

Los algoritmos $k$-NN ($k$-*nearest neighbours*) se definen con base en una muestra de referencia (o de entrenamiento) $\mathbf{Z}$ de elementos de $\Omega$ previamente clasificados; $n_0$ elementos pertenecientes a $\Omega_0$ y $n_1$ pertenecientes a $\Omega_1$. Cada individo de $\Omega$ se representa por un vector numérico de características $S$. Para clasificar un nuevo individuo $\omega \in \Omega$, se busca a los $k$ elementos de $\mathbf{Z}$ más cercanos a $\omega$ y se registra el número $X \leq k$ de estos vecinos pertenecientes a $\Omega_1$. Entonces, decidimos clasificar en $\Omega_1$ siempre que $X > t$. Los criterios $k$-NN tienen ventajas prácticas importantes. Son libres de distribución, fáciles de comprender y precisos aún con tamaños de muestra reducidos. Sin embargo, tienen algunos inconvenientes importantes: no existe un criterio para calibrar el número $k$ de vecinos cercanos, y su implementación con muestras grandes implica un alto costo computacional.

## 4 Algoritmo $k$-NN MB

Con el objetivo de facilitar la implementación de los criterios $k$-NN, Guo et al [5] proponen el algoritmo de condensación $k$-NN MB[2]. Su propuesta construye una forma simplificada de la muestra de entrenamiento $\mathbf{Z}$ definida mediante vectores de baja dimensión que representan conjuntos de observaciones. Para describir el algoritmo de Guo et al [5] considere la siguinte notación:

- $Num(d_i)$ representa el número total de puntos en el conjunto[3] $i$ (siendo $d_i$ el punto central del conjunto).
- $Sim(d_i)$ es la distancia entre el punto $d_i$ y el elemento más lejano a éste dentro del conjunto $i$.
- $Rep(d_i) := d_i$.
- $Cls(d_i)$ es la categoría de origen de $d_i$.

El algoritmo de condensación es el siguiente:

1. Crear una matriz de distancias de los datos en la muestra de referencia.
2. Etiquetar como "no agrupado" a todos los elementos en la muestra.

[2]por su nombre en inglés: $k$-NN Model Based Approach
[3]Este conjunto es una esfera que representa a varios elementos de $\mathbf{Z}$ simultáneamente

3. Encontrar, para cada dato, su "vecindad"(esfera) más grande de tal forma que cubra al mayor número de individuos sólo de la misma categoría.

4. Encontrar el dato $d_j$ cuyo conjunto contenga al mayor número de elementos; crear el vector: $< Cls(d_i), Sim(d_i), Num(d_i), Rep(d_i) >$ y etiquetar a los elementos pertenecientes al conjunto como "agrupado".

5. Repetir los pasos 3 y 4 hasta que todos los datos de la muestra estén agrupados.

Llamaremos **M** al modelo resultante del algoritmo anterior, y llamamos **"representante"** de la $i$-ésima esfera al triplete $< Cls(d_i), Sim(d_i), Num(d_i), Rep(d_i) >$ . Si existe más de una esfera con el mismo número máximo de elementos contenidos, escogemos como representante aquel cuyo valor $Sim(d_i)$ sea menor; es decir, aquel con mayor densidad. La figura 1 ilustra gráficamente este algoritmo.

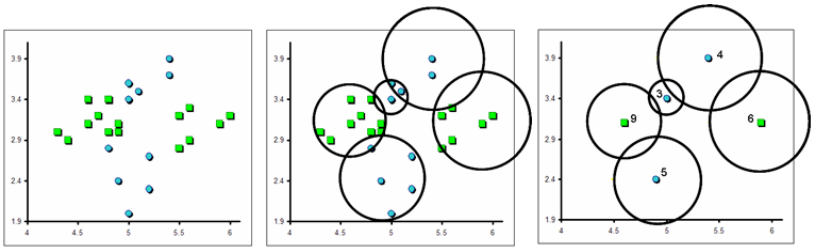


Figura 1: Algoritmo de Construcción del $k$-NN MB.

El algoritmo de clasificación a partir de la base condensada es el siguiente:

1. Para una nueva observación $d_t$ por clasificar, calcular la distancia entre $d_t$ y todos los $Rep(d_i)$ en el modelo **M**.

2. Si $d_t$ está contenido únicamente en la esfera $< Cls(d_j), Sim(d_j), Num(d_j), Rep(d_j) >$ (*i.e.* La distancia de $d_t$ a $d_j$ es menor o igual a $Sim(d_j)$), clasificar $d_t$ en $Cls(d_j)$.

3. Si $d_t$ está contenido en al menos dos esferas de diferente categoría, clasificar $d_t$ en la categoría del representante con el mayor $Num(d_j)$.

4. Si ningún representante en el modelo **M** contiene a $d_t$, clasificar a $d_t$ en la categoría del representante cuya frontera sea la más cercana a $d_t$.

La figura 2 muestra algunos ejemplos de clasificación utilizando el algoritmo del $k$-NN MB.

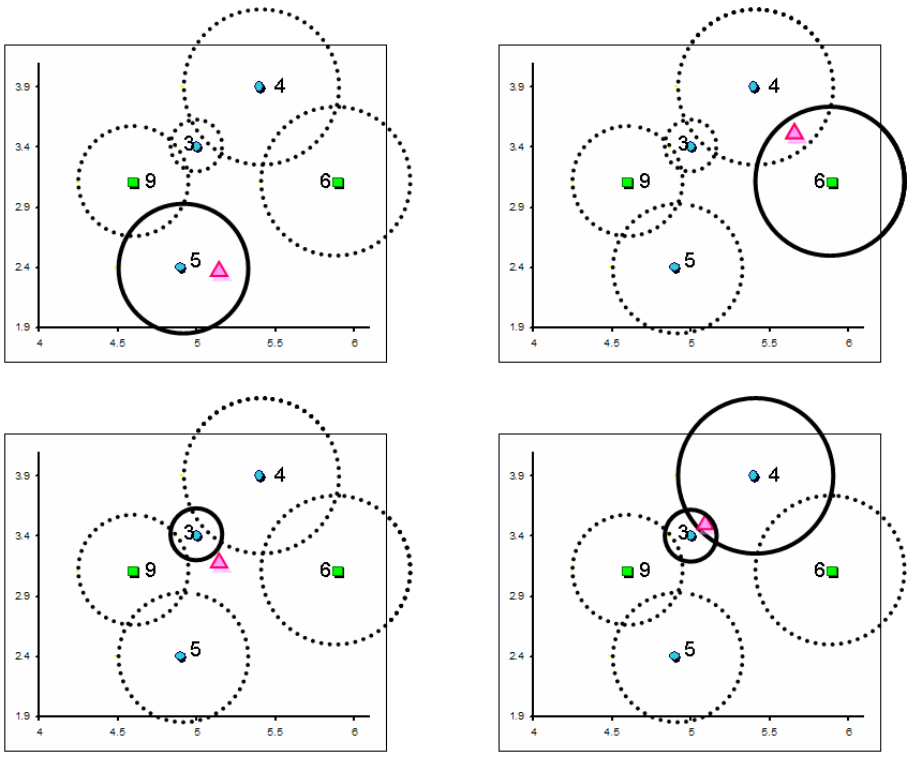


Figura 2: Ejemplos de Clasificación con el $k$-NN MB.

## 5 $k$-NN Controlado

En esta sección proponemos un algoritmo original de condensación. Aunque inspirado en el $k$-NN MB, nuestra propuesta, el $k$-NN C[4], se define en términos de dos parámetros que pueden ser calibrados por el usuario. Esta

[4] $k$-NN Controlado

característica permite evaluar la presición del proceso de clasificación utilizando curvas *ROC*. Esto permite al usuario definir objetivamente un valor adecuado para el parámetro $k$ (número de vecinos cercanos).

Sean los parámetros: $k$:= Número de puntos permitidos en cada conjunto representante y $t$:= Umbral de decisión. Si $d_i$ := Punto central del conjunto (esfera) $i$, entonces:

- $Num_{\Omega_j}(d_i)$ :=Número de puntos en el conjunto $i$ cuya categoría de origen es $\Omega_j$.
- $Sim(d_i)$ :=Distancia entre el punto $d_i$ y el elemento más lejano a éste en el conjunto $i$.
- $Rep(d_i) := d_i$.
- $Cls(d_i)$ := Categoría asignada a $d_i$ de acuerdo a la regla:

$$Cls(d_i) = \begin{cases} \Omega_1 \text{ si } Num_{\Omega_1}(d_i) > t \\ \\ \Omega_0 \text{ si } Num_{\Omega_1}(d_i) \leq t. \end{cases}$$

El algoritmo de construcción es entonces el siguiente:

1. Crear una matriz de distancias de los datos en la muestra de referencia.
2. Etiquetar como "no agrupado" a todos los elementos en la muestra.
3. Encontrar, para cada dato, la esfera ("vecindad") con centro en ese dato tal que cubra $k$ individuos (sin importar la categoría de origen de los mismos).
4. Encontrar el dato $d_j$ cuyo conjunto (esfera) sea de radio mínimo (*i.e.* preferimos conjuntos más densos); crear el vector:

   $$< Cls(d_i), Sim(d_i), Rep(d_i) >$$

   y etiquetar a los elementos pertenecientes al conjunto como "agrupado".
5. Repetir los pasos 3. y 4. hasta que todos los elementos de **Z** estén agrupados.

En el algoritmo, **M** denota al modelo final, y llamamos a $< Cls(d_i), Sim(d_i), Rep(d_i) >$ el **"representante"** de la $i-$a esfera. La figura 3 ilustra un ejemplo de este algoritmo utilizando el *k-NN C* con $k = 9$ y $t = 4$.

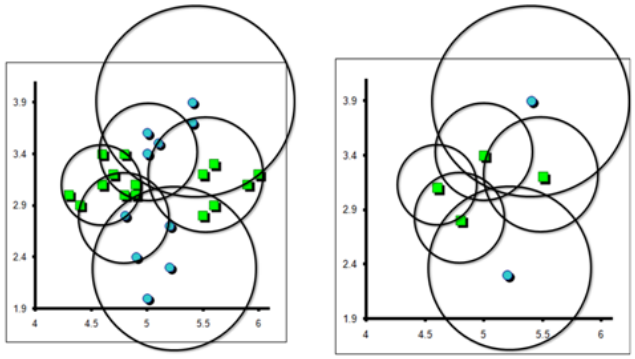


Figura 3: Algoritmo de Construcción del $k$-NN C .

El algoritmo de clasificación con base en la muestra de entrenamiento condensada es el siguiente:

1. Para una nueva observación $d_t$ por clasificar, calcular la distancia entre $d_t$ y todos los $Rep(d_i)$ en el modelo **M**.

2. Si $d_t$ está contenido únicamente en la esfera representada por $< Cls(d_j), Sim(d_j), Rep(d_j) >$ (*i.e.* La distancia de $d_t$ a $d_j$ es menor o igual a $Sim(d_j)$), clasificar $d_t$ en $Cls(d_j)$.

3. Si $d_t$ está contenido en al menos dos esferas de diferente categoría, clasificar $d_t$ en la categoría del representante con el menor $Sim(d_j)$.

4. Si ningún representante en el modelo **M** contiene a $d_t$, clasificar a $d_t$ en la categoría del representante cuya frontera sea la más cercana a $d_t$.

Cuando se emplea el algoritmo de condensación $k$-NN MB existe una única regla de clasificación, por lo que no es posible ajustar curvas *ROC* para medir su desempeño. Esto se debe a que la curva *ROC* empírica contiene sólo un punto además del (0,0) y el (1,1) (que siempre pertenecen a la curva). El *k-NN C* depende de dos parámetros que permiten controlar el intercambio condensación a la base de entrenamiento y precisión del

algoritmo. Valores grandes del parámetro $k$ implican que $\mathbf{M}$ contenga un menor número de representantes. Sin embargo, en general, a mayor condensación menor precisión (medida en términos de área bajo la curva *ROC*, o bien, como porcentaje de clasificación correcta). La regla óptima de clasificación para el *k-NN C* es la combinación de parámetros $k$ y $t$ con máxima área bajo su curva ROC (controlada por $k$) y mejor combinación de sensibilidad y especificidad (definidias por $t$). El nivel de condensación del algoritmo lo medimos a partir de la reducción proporcional del tamaño de la base de referencia. La precisión es medida en términos del área bajo la curva *ROC* correspondiente.

## 6 Ejemplos simulados

En esta sección presentamos tres ejemplos que ilustran las ideas presentadas en las secciones anteriores. En cada ejemplo se condensa una base de entrenamiento con 1000 individuos provenientes de $\Omega_0$ y 1000 individuos de $\Omega_1$. El ejercicio de validación cruzada se realizó simulando 100 vectores más de cada categoría que jugaron el rol de "observaciones por clasificar". Todos los ejemplos se construyeron con vectores en $\mathbb{R}^2$ aunque cualquiera de los algoritmos puede utilizarse con vectores en espacios de mayor dimensión.

**Ejemplo 1**: Normales bivariadas
La distribución de los datos en ambas categorías es la de una Normal Bivariada. Específicamente: $V_0 \sim N(0, I)$ y $V_1 \sim N(\delta, I)$; $\delta = (2, 0)$. La base de entrenamiento y la base condensada se muestran en la figura 4. Es evidente que $k$-NN MB logró condensar únicamente por fuera de la intersencción de ambas categorías. En cambio $k$-NN C condensa uniformemente en todo el conjunto de entrenamiento. El cuadro 1 muestra los resultados de los tres métodos, podemos observar que el *k-NN C* fue capaz de condensar el 89% de la base de referencia perdiendo únicamente 1% de clasificación correcta y 0.01 de Área bajo la curva *ROC*.

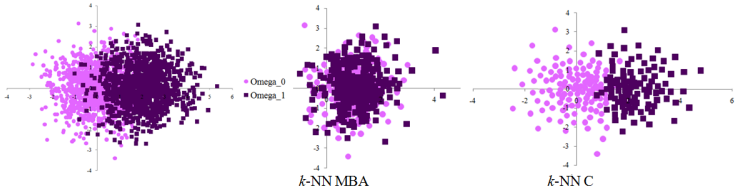


Figura 4: Base de referencia ejemplo 1

| | $k$-NN tradicional | $k$-NNMBA | $k$-NN C |
|---|---|---|---|
| sensibilidad | 0.92 | 0.86 | 0.88 |
| especificidad | 0.88 | 0.86 | 0.9 |
| % clas. correcta | 0.9 | 0.86 | 0.89 |
| k | 9 | - | 22 |
| t | 3 | - | 12 |
| Condensación | 0 | 0.7445 | 0.891 |
| Área Curva ROC | 0.96 | - | 0.95 |

Tabla 1: Resultados de los tres métodos ejemplo 1.

**Ejemplo 2**: Círculos no intersectados.
Para este ejemplo, se simularon uniformemente 2000 vectores en el cuadro $(-1,1)\times(-1,1)$ y la categoría se asignó verificando su pertenencia a los segmentos de círculo que se ven reflejados en la figura 5. Este ejercicio busca favorecer al *k-NN MB*. Los resultados del *k-NN C* muestran con claridad la pérdida de precisión generada por la condensación. Como esperábamos, en este caso la mejor ópcion es el *k-NN MB*.

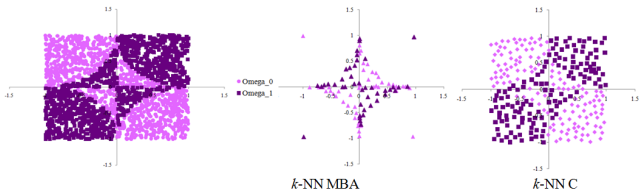


Figura 5: Base de referencia ejemplo 2: Círculos

La base de referencia y los resultados de la condensación se ilustran en la figura 5. El cuadro 2 muestra los resultados de los tres métodos, podemos observar que, en este ejemplo, la mejor opción es el *k-NN MB* ya que ofrece un 97% de clasificación correcta y elimina el 95% de la base de entrenamiento. El *k-NN C* puede alcanzar el mismo nivel de condensación, pero con menor precisión.

## 7 Conclusiones

Si suponemos que $k$-NN ocupa una variable latente como índice de riesgo; podemos evaluar su precisión estadística mediante un análisis de curvas

| | $k$-NN tradicional | $k$-NN MB | $k$-NN C |
|---|---|---|---|
| sensibilidad | 0.98 | 0.96 | 0.96 |
| especificidad | 0.98 | 0.98 | 0.96 |
| % clas. correcta | 0.98 | 0.97 | 0.96 |
| $k$ | 4 | - | 11 |
| $t$ | 1 | - | 4 |
| Condensación | 0 | 0.9545 | 0.8225 |
| Área Curva ROC | 0.98 | - | 0.98 |

Tabla 2: Resultados de los tres métodos ejemplo 2.

*ROC*. Hemos visto que los algoritmos de condensación implican siempre una pérdida de precisión en los procesos de clasificación. La Principal aportación de este trabajo es la propuesta de un nuevo algoritmo de condensación para $k$-NN. Este algoritmo es competitivo en términos de precisión y capacidad de condensación, y puede ser controlado mediante un breve análisis de curvas *ROC*. Los resultados del estudio de Monte Carlo sugieren que el algoritmo $k$-NN Controlado es más eficiente cuando existe un traslape significativo entre las poblaciones por clasificar (situación frecuentemente observada en la práctica).

## Referencias

[1] Bamber, D. (1975) "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph", *Journal of Mathematical and Statistical Psicology* **12**(4): 387–415.

[2] Cuevas-Covarrubias, C. (2003) *Statistical Inference for ROC Curves*. Tesis de Doctorado, Departamento de Estadística, Universidad de Warwick, Coventry, Reino Unido.

[3] Cuevas-Covarrubias, C.; Monroy, V.; Ortega, V. (2008) "Aplicación de un algoritmo $k$-NN para la gestión del capital humano. Predicción del desempeño y detección de competencias críticas en el desarrollo del personal", Preprint, Up-Pharma, Ciudad de México, México.

[4] Dorfman, D.D.; Alf, E. Jr. (1969) "Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating-method data", *Journal of Mathematical Psychology* **6**(3): 487–496.

[5] Guo, G.; Wang, H.; Bell, D.; Bi, Y.; Greer, K. (2003) "KNN model-based approach in classification", in: *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, Lecture Notes in Computer Science, Volume 2888, Springer, Berlin: 986–996.

[6] Hand, D.J. (1994) "Assessing classification rules", *Journal of Applied Statistics* **21**: 3–16.

[7] Hanley, J.A.; McNeil, B.J. (1982) "The meaning and use of the area under the under a receiver operating characteristic (ROC) curve", *Radiology* **143**: 29–36.

[8] Henley, W.E.; Hand, D.J. (1996) "A $k$-nearest-neighbour classifier for assessing consumer credit risk", *The Statistician*, **45**(1): 77–95.

[9] Krzanowski, W.J.; Hand, D.J. (2009) *ROC Curves for Continuous Data*. Chapman & Hall/CRC, Londres, Reino Unido.

[10] Zweig, M.H.; Campbell, G. (1993) "Receiver operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine", *Clin. Chem.*, **39**(4): 561–577.