

Optimización de una Nariz Electrónica a través de un método de selección de variables

Cristhian Manuel Durán Acevedo

Grupo de Investigación en Sistemas Multisensoriales y Reconocimiento de Patrones, Universidad de Pamplona,
Facultad de Ingenierías y Arquitectura, Pamplona, Colombia, cmduran@unipamplona.edu.co

Oscar Eduardo Gualdron

Grupo de Investigación en Sistemas Multisensoriales y Reconocimiento de Patrones, Universidad de Pamplona,
Facultad de Ingenierías y Arquitectura, Pamplona, Colombia, oscar.gualdron@unipamplona.edu.co

RESUMEN

El presente trabajo describe un sistema de percepción artificial (Nariz Electrónica), el cual es compuesto de una matriz de 16 sensores de gases químicos y métodos de procesamiento de señales para la clasificación de compuestos químicos. Se utilizaron en total tres compuestos orgánicos volátiles (COV), de tipo hidrocarburos aromáticos (Benceno, Tolueno y Xileno) a diferentes niveles de concentración de 500, 1000 y 1500 ppm (partes por millón). Se compararon los resultados con métodos de reconocimiento de patrones (MLP, PNN, SVM, SVM Multiclase, SVM con Mínimos Cuadrados) y la técnica de selección de variables “Simulated annealing (SA) y PNN (Probabilistic Neural Network), obteniendo una reducción importante de la matriz de datos y llegando a un porcentaje de acierto de hasta 100 % de clasificación.

Palabras claves: Nariz Electrónica, Compuestos Químicos, Selección de variables, Procesamiento, Sensores.

ABSTRACT

This paper describes an artificial sensing system (Electronic Nose), which is composed of an array of 16 chemical gas sensors and signal processing methods for classification of chemical compounds. A total of three volatile organic compounds (VOC) were used such as aromatic hydrocarbons (benzene, toluene and xylene) at different concentration levels of 500, 1000 and 1500 ppm (parts per million). The results were compared with pattern recognition methods (MLP, PNN, SVM) and the variable selection technique “Simulated annealing (SA)” and PNN (Probabilistic Neural Network), obtaining a reduction in the matrix data and reaching a success rate of 100% of classification.

Keywords: Electronic Nose, Chemical Compounds, variable selection, Processing, Sensors.

1. INTRODUCCIÓN

En los sistemas de análisis de aromas como las “Narices Electrónicas”, por lo general se detectan compuestos a partir de una matriz de sensores de gases químicos con sensibilidades parcialmente sobrepuestas. Por lo tanto, con relación a este concepto de la nariz electrónica, es necesario tener en cuenta que un número grande y/o significativo de sensores puede ser irrelevante puesto que por lo general algunos de los sensores presentan con el tiempo un desgaste ó derivas generadas por cambios en las variables físicas ó eléctricas del dispositivo. Por ejemplo: Cambios de temperatura, humedad ó variación eléctrica. En forma general y para diferentes aplicaciones se puede considerar que otros sensores pueden presentar buen comportamiento en cada una de las respuestas.

En la actualidad diferentes estrategias se han reportado para la reducción de dimensionalidad del conjunto de datos. Esta reducción consiste básicamente en la selección de cualquier variable disponible en dicho conjunto ó para encontrar nuevas variables llamadas factores (por ejemplo, mediante la realización un análisis de componentes principales o de un análisis discriminante lineal, etc), (Alizadeh., 2010). En el momento que el sensor presenta variaciones importantes, es importante implementar métodos de reconocimiento de patrones (Por ejemplo: Análisis de Componentes Principales (PCA)), como un paso previo en la reducción de dimensionalidad del conjunto de datos. Para el caso del método clásico de reconocimiento de patrones PCA, se realiza un análisis a través de un plano bidimensional, el cual conlleva a seleccionar los sensores mediante un criterio de análisis de “Loadings”, y que consiste en que los sensores que estén más cerca del valor cero del plano se puedan eliminar y los sensores que se agrupen en un punto específico del plano se seleccionen para también realizar dicha eliminación (Gardner., et al, 1999).

Otras estrategias han sido desarrolladas para selección de variables, las cuales conducen a la construcción de modelos de clasificación parsimoniosa basada en redes neuronales tipo Fuzzy ARTMAP ó redes neuronales probabilísticas. Este proceso consiste en seleccionar variables a través de dos etapas, la primera a través de una selección basada en un criterio de varianza seguido por un proceso de recocido simulado (Llobet., et al, 2007).

En otra aplicación un algoritmo de búsqueda estocástica, mediante algoritmos genéticos (GA's), utiliza ciertas funciones características como un primer paso para optimizar problemas de la selección del sensor con mayor grado de información (Phaisangittisagul., et al, 2010).

El modulo desarrollado en este proyecto cuenta con una matriz de 16 sensores que permite obtener un gran conjunto de datos por cada medida. En forma global, al tener una matriz grande de información no significa tener mejores resultados ya que hay variables que aportan información útil y otras que solo aportan ruido. Por ese motivo es necesario escoger cuidadosamente las variables (Es decir, sensores) que los diferentes algoritmos de reconocimiento de patrones van a utilizar para obtener una buena clasificación u optimización del sistema.

En este artículo se describe la optimización de una nariz electrónica (Gualdrón., et al, 2011), a través de los resultados obtenidos a partir de diferentes métodos de reconocimiento de patrones como las redes neuronales MLP y PNN y el clasificador SVM. Estos métodos fueron comparados mediante el acople de la técnica de selección de variables “recocido simulado” a cada uno de estos métodos, obteniendo una importante reducción de dimensionalidad de la matriz de datos y llegando a un porcentaje de acierto en la clasificación de las medidas de hasta 100 % en algunos casos.

2. MATERIALES Y MÉTODOS



Figura 1: Fotografía del equipo multisensor

El esquema del proceso de medida lo componen tres etapas principales. Una etapa de concentración (Es decir, de espacio de cabeza estático) con flujo constante, una matriz de 16 sensores químicos tipo TGS ubicada en la cámara de medida, y un sistema informático.

En la figura 1 se muestra el esquema general del equipo multisensor, en el cual se pueden distinguir los diferentes elementos que conforman el equipo.

El procedimiento de medida en el modulo consta de tres fases bien diferenciadas: La fase de concentración, la fase de medida y la fase de reposo. Los tiempos de medida fueron los siguientes:

- a) **Tiempo de Concentración:** 15 minutos
- b) **Tiempo de Medida:** 10 Minutos
- c) **Tiempo de Reposo:** 20 Minutos

2.1 CÁMARA DE CONCENTRACIÓN

En la cámara de concentración la cubierta superior de acrílico fue equipada de un “septum”, con el objetivo de inyectar los compuestos químicos con una jeringa milimétrica en forma manual. Las dimensiones de la cámara de concentración fueron de 12 cm largo x 12 cm ancho x 12cm alto, para un volumen total de 1728 cm³. Las dimensiones de la cámara de medida fueron de 8.7cm largo x 8.7cm de ancho x 5.3cm de alto, para un volumen total de 401.157 cm³, la cual también fue desarrollada en acrílico transparente. Por tal motivo, el sistema de muestreo en forma general fue acondicionado a través de la cámara de concentración, la cámara de medida, cuatro electroválvulas solenoides y una bomba de aire de 5 bares de presión para transportar los volátiles.

2.2 SENSORES DE GASES

Los 16 sensores de gases utilizados para el desarrollo del equipo fueron de tipo semiconductor de la casa Figaro Inc., concretamente de óxido de estaño. En la tabla 1 se describen cada uno de los sensores de gases utilizados en la cámara de medida.

Tabla 1: Sensores utilizados en el modulo

Sensor de gas FIGARO Serie-8	Aplicación	Tipo de Gas	cantidad	Número del sensor
TGS 826	Detección de olores	Amoníaco y aminas	2	S1, S4
TGS 831	Detección de gas halocarburos (gases refrigerantes)	R-22, Monoclorodifluorometano	2	S2, S15
TGS 821	Detección de gas combustible	Hidrógeno	2	S3, S12
TGS 842	Detección de gas combustible	Metano y gas natural	2	S5, S14
TGS 880	Control de cocción	humo de los alimentos (Alcohol, olor)	2	S6, S10
TGS 825	Detección de gas toxico	sulfuro de hidrógeno	1	S7
TGS 813	Detección de gas combustible	Hidrocarburos en general	1	S8
TGS 800	Control de calidad del	contaminantes del aire en	1	S9

	aire	general		
TGHS 822	Detección de vapores de disolventes	Alcohol y disolventes orgánicos	1	S11
TGS 832	Detección de gas halocarburos (gases refrigerantes)	R-134 1,1,1,2-Tetrafluoroethane	1	S13
TGS 830	Detección de gas halocarburos (gases refrigerantes)	R-22 ,Monoclorodifluorometano	1	S16

Para adquirir y almacenar las señales de los sensores al ordenador se utilizó una tarjeta de alto rendimiento de la National Instruments, la DAQ NI USB-6210, el cual es un módulo multifunción de la Serie M optimizado para una precisión superior a velocidades de muestreo muy altas.

2.3 MÉTODOS DE SELECCIÓN DE VARIABLES Y PROCESAMIENTO DE DATOS

A continuación se describe brevemente el método de selección de variables y las técnicas de reconocimiento de patrones utilizadas en la clasificación de compuestos y reducción de dimensionalidad del conjunto de datos.

2.3.1 SIMULATED ANNEALING Ó RECOCIDO SIMULADO

Simulated annealing es una técnica estocástica de optimización que permite hallar soluciones cercanas al óptimo global en problemas de optimización complejos (p.e. con elevado número de variables).

El algoritmo parte de la utilización de un conjunto de variables (generalmente todas las variables disponibles), X_0 . Una vez calculado el fitness (error de predicción obtenido con la red clasificadora PNN) asociado a X_0 , se determina un conjunto X_1 resultado de haber eliminado n variables (escogidas aleatoriamente dentro de X_0). El fitness de X_1 es obtenido y se calcula la diferencia entre el modelo original y el nuevo.

$$\Delta E = \text{fitness}(\text{nuevo}) - \text{fitness}(\text{viejo}) \quad (1)$$

Se puede afirmar que ΔE es negativo si el nuevo modelo es superior al original, es decir el resultado del cálculo del error de predicción es menor, por lo que la clasificación ha mejorado eliminando dicha variable. Ahora bien, si el modelo resultante al utilizar las variables de X_1 es peor al original (ΔE positivo) no significa directamente que dicha combinación deba ser rechazada. Se pasaría a una segunda fase en la cual se define la probabilidad:

$$p_i = \exp\left(-\frac{\Delta E}{T_i}\right) \quad (2)$$

Donde T_i es la temperatura de trabajo (se escoge un valor inicial para T_i). Si $P_i > R$ (R es un valor aleatorio con distribución uniforme entre $[0,1]$) la nueva solución es retenida y el algoritmo prosigue eliminando variables a partir de X_1 . En caso contrario, el algoritmo prosigue desde X_0 . La figura 2 muestra a p como una función de los cambios de fitness (valores positivos de ΔE) a diferentes temperaturas annealing T_i . Cuando la temperatura de annealing se reduce, la posibilidad para aceptar una solución peor decrece significativamente. El proceso se ejecuta un número determinado de iteraciones ($N_{\text{iteraciones}} \propto T_i^{-\alpha}$) para la temperatura T_i . Finalmente todo lo anterior se repite para $T_i \propto T_i^{-\alpha}$. En otras palabras, el proceso de selección se desarrolla a una temperatura más baja (el proceso de aceptación de cambios es menos exigente). El número de temperaturas a computar también debe ser definido a priori. Teniendo en cuenta que un número bajo puede dar resultado a soluciones no muy buenas y un número alto incrementaría drásticamente el tiempo de computación del proceso

(un valor promedio de ejecución puede ser de 50 temperaturas). En la figura 3 se puede ver el diagrama de flujo del algoritmo simulated annealing. Este algoritmo se utilizó para la red neuronal PNN principalmente porque arrojó resultados satisfactorios.

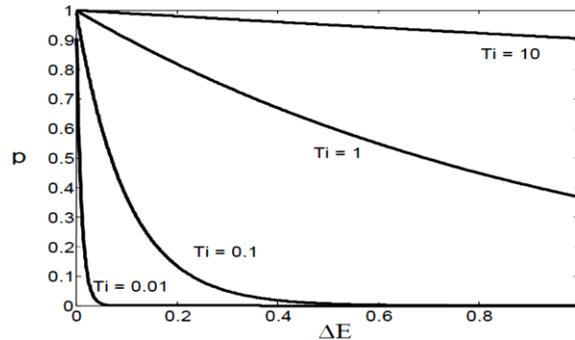


Figura 2: Probabilidad p de aceptación a soluciones peores como una función del cambio de fitness a diferentes temperaturas annealing, T_i .

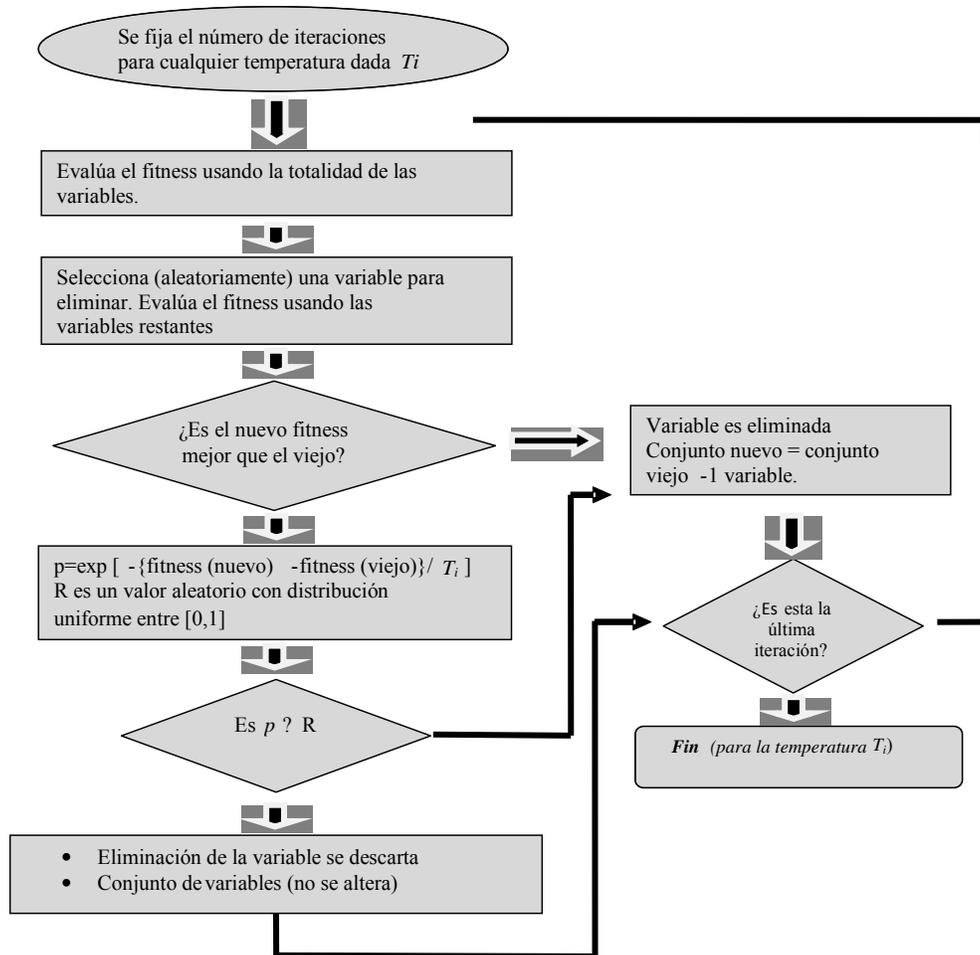


Figura 3: Diagrama de Flujo del algoritmo Simulated Annealing

2.3.2 MÉTODOS DE RECONOCIMIENTO DE PATRONES

A continuación se describen brevemente los métodos clasificadores utilizados para el procesado de datos, los cuales fueron aplicados a la matriz de datos para la clasificación de las medidas:

1. **MLP** (Multilayer Perceptron)
2. **PNN** (Probabilistic Neural network)
3. **SVM** (Support Vector Machine)
4. **SVM Multiclase**
5. **LS-SVM** (Least squares -Support Vector Machine)

La tabla 2 resume las principales características de los algoritmos aplicados.

Tabla 2: Principales características de los algoritmos aplicados al sistema

Técnica	Algoritmo	Aprendizaje	Paramétrica	Aplicación
MLP	Neuronal	Supervisado	No	Clasificación Cuantificación
PNN	Neuronal	Supervisado	No	Clasificación
SVM	Obtiene una Frontera de decisión	Supervisado	No	Clasificación y Análisis de Regresión
Multiclass SVM	Obtiene una Frontera de decisión. Se construyen N modelos SVM y $N(N-1)/2$ clasificadores	Supervisado	No	Clasificador Análisis de Regresión
Least squares - Support Vector Machines (LS_SVM)	Obtiene una Frontera de decisión con Mínimos Cuadrados	Supervisado	No	Clasificador Análisis de Regresión

En la tabla 3 se describen los métodos usados para extraer los parámetros estáticos más relevantes del conjunto de datos. Estos parámetros fueron aplicados a valores de conductancia ($1/R_s$).

Tabla 3: Métodos de extracción de parámetros

Parámetro	Expresión	Comentario
Máximo incremento de la conductancia	$G_{max} - G_{min}$	Minimización del efecto de la temperatura.
Incremento de conductancia	$G_f - G_i$	
Máximo incremento de la conductancia normalizada o máxima diferencia fraccional	$(G_{max} - G_{min}) / G_{max}$	
Incremento de conductancia normalizada o incremento de la diferencia fraccional	$(G_f - G_i) / G_f$	

En la tabla 4 se describen las diferentes técnicas utilizadas para el pre-procesado de datos, las cuales realizan tareas de normalización y auto-escalado.

Tabla 4: Métodos de pre-procesado de datos

Normalización	Descripción
Auto escalado (Autoscaling)	El objetivo de esta normalización es la de dar igualdad de escala a cada una de las variables o parámetros que describen cada medida.
Centrado (mean centring)	En esta normalización a cada una de las columnas (variables) se le resta su valor medio. Es decir, toda medida es descrita por variables de media nula.
Normalizado o Normalizado por matriz	Divide toda la matriz por el elemento máximo de la misma. En este caso solo un elemento valdrá la unidad.

2.4 MUESTRAS

Las muestras fueron preparadas con un volumen de 0.1 ml sobre un recipiente de plástico previamente esterilizado para garantizar que cada prueba se hiciera de forma igual y homogénea. La extracción de la muestra del recipiente contenedor se realizó mediante una jeringa milimétrica etiquetada para cada compuesto. En total se hicieron 45 medidas con densidades aproximadas del 0.88 g/ml y un promedio de 870.000 ppm de concentración.

Relación de medidas:

- a) 15 Medidas de benceno
- b) 15 Medidas de Tolueno
- c) 15 Medidas de Xileno

3. RESULTADOS

A continuación se describen los resultados obtenidos a través de la clasificación de compuestos aromáticos (Benceno, Tolueno y Xileno) a diferentes niveles de concentración de 500, 1000 y 1500 ppm, mediante el algoritmo de selección de variables (SV) "Simulated Annealing SA" acoplado a la red neuronal PNN y en todos los casos con el método de normalización centrado de datos.

En el conjunto de datos se aplicó en todas las clases el SVM (uno contra todos), donde se construyeron N modelos. El SVM fue entrenado con todas las muestras de entrenamiento de la clase con etiquetas (Labels), valores positivos, y todas las otras muestras de entrenamiento con Labels negativos. Una nueva medida pertenece a la clase que tiene los valores altos de la función indicador.

Es de aclarar que el método de selección de variables fue SA-PNN, el cual generó mejores resultados y seleccionó los sensores con información relevante para luego ser aplicados cada uno de los métodos de clasificación.

3.1 CLASIFICACIÓN DE TODOS LOS COMPUESTOS CON EL MÉTODO DE SELECCIÓN SA-PNN

La tabla 5 ilustra los porcentajes de aciertos en la clasificación de las Muestras de Benceno, Tolueno y Xileno, usando redes neuronales. SVM multiclases y SVM Mínimos cuadrados, con el procedimiento uno contra todos con selección de variables.

Tabla 5: Clasificación de todos los compuestos usando redes neuronales. SVM multiclases y SVM Mínimos cuadrados, con el procedimiento uno contra todos con selección de variables

Parámetro Estático	Método de selección	Variables seleccionadas	Método de clasificación	% de clasificación con SV	% de clasificación sin SV
Gmax-Gmin 45 X 16	SA - PNN	2,3,5,8,11,12	MLP	100%	100%
			PNN	100%	100%
			SVM Multiclases	67%	100%
			LS SVM Multiclases	100%	100%

3.2 CLASIFICACIÓN DEL BENCENO CON EL METODO DE SELECCIÓN SA-PNN

La tabla 6 ilustra los porcentajes de aciertos en la clasificación de las Muestras de Benceno a 1500, 1000 y 500 ppm, usando redes neuronales. SVM multiclases y SVM Mínimos cuadrados.

Tabla 6: Clasificación de las muestras de Benceno con SVM multiclases y SVM Mínimos cuadrados, con el procedimiento uno contra todos con selección de variables

Parámetro	Método de selección	Variables seleccionadas	Método de clasificación	% de clasificación con SV	% de clasificación sin SV
Gmax-Gmin 15 X 16	SA-PNN	1,2,3,4,9,12	MLP	100%	100%
			PNN	83-93%	83%
			SVM multiclases	33%	33.3%
			LS SVM multiclases	67%	33.3%

3.3 CLASIFICACIÓN DEL TOLUENO CON EL METODO DE SELECCIÓN SA-PNN

La tabla 7 ilustra los porcentajes de aciertos en la clasificación de las Muestras de Tolueno a 1500, 1000 y 500 ppm, usando redes neuronales. SVM multiclases y SVM Mínimos cuadrados.

Tabla 7: Clasificación de las muestras de Tolueno con SVM multiclases y SVM Mínimos cuadrados, con el procedimiento uno contra todos con selección de variables

parámetro	Método de selección	Variables seleccionadas	Método de clasificación	% de clasificación con SV	% de clasificación sin SV
Gmax-Gmin 15 X 16	SA-PNN	2,3,4,10,11,13,16	MLP	100%	100%
			PNN	80%	87%
			SVM multiclases	67%	100%
			LS SVM multiclases	100%	66.7%

3.4 CLASIFICACIÓN DEL XILENO CON EL METODO DE SELECCIÓN SA-PNN

La tabla 8 ilustra los porcentajes de aciertos en la clasificación de las Muestras de Xileno a 1500, 1000 y 500 ppm, usando redes neuronales. SVM multiclases y SVM Mínimos cuadrados.

Tabla 8: Clasificación de las muestras de Xileno con SVM multiclases y SVM Mínimos cuadrados, con el procedimiento uno contra todos con selección de variables

Parámetro	Método de selección	Variables seleccionadas	Método de clasificación	% de clasificación con SV	% de clasificación sin SV
Gmax-Gmin 15 X 16	SA-PNN	1,2,4,13,15,16	MLP	100%	100%
			PNN	93%	83%
			SVM multiclases	33%	100%
			LS SVM multiclases	100%	100%

4. CONCLUSIONES

El método de selección de variables SA-PNN utilizado en el conjunto de medidas adquiridas previamente por la nariz electrónica, permitió eliminar un número determinado de variables (Sensores) redundantes ó ruidosas, llegando a mejorar en la mayoría de los casos el porcentaje de clasificación del número de muestras a partir de las respuestas de cada uno de los clasificadores. Por ejemplo, para el caso de la clasificación de todos los compuestos (Xileno, Benceno y Tolueno) correspondiente a 45 medidas, se obtuvo una clasificación del 100% de acierto utilizando las redes neuronales como la MLP y PNN, elevando hasta un 10 % más de clasificación y eliminando una cantidad considerable de variables.

Para el caso de la clasificación de las 15 medidas de benceno, se llegó a mejorar las respuestas con cada uno de los métodos de reconocimiento de patrones, obteniendo un conjunto de variables que determinan al sistema multisensorial una buena selectividad y comportamiento.

Es importante aclarar que al encontrar y utilizar el mejor método de clasificación, es posible reducir la dimensión de la matriz de sensores de gases, generando de esta forma una importante disminución de costos en materiales y computacionales.

REFERENCIAS

- Alizadeh.,T (2010), “Chemiresistor sensors array optimization by using the method of coupled statistical techniques and its application as anelectronic nose for some organic vapors recognition” *Sensors and Actuators B: Chemical*, vol.143, núm.2, pp.740-749.
- Gardner.J., and barlett.P (1999), “*Electronic Noses: Principles and Aplications*“, Oxford University, New York, pp: 154-155.
- Llobet. E., Brezmes.J., Gualdrón.O., Vilanova.X and Correig.X (2004) “Bui ding pa simoniouus fuzzy ARTMAP models by variable selection with a cascaded genetic algorithm; application to multisensors systems for gas ana ysis” *Sensors and Actuators B*, vol.99, pp.267-272.
- Llobet.E; Gualdrón.O; Vinaixa.M; El-Barbri.N, Brezmes.J, Vilanova.X; Bouchikhi.B, Gómez.R; Carrasco.J, Correig.X (2007), “Efficien fea u e se ec ion fo mass-spec ome y as ed e ec onic nose app ica ions.” *Chemometrics and Intelligent Laboratory Systems*, vol.85,pp. 253-261.

Phaisangittisagul.,E., Nagle,H.T., Areekul.,V 200 , “Intelligent method for sensor subset selection for machine olfaction”, *Sensors and Actuators B: Chemical*, vol.145, núm.1, pp.507-515.

Gualdrón, O., Duran.C., Isaza.C., Carvajal,Adrián and Cesar uribe (2011). “Sistema de olfato electrónico de bajo costo para la Detección de diferentes compuestos químicos Contaminantes, *Revista Colombiana de tecnologías de avanzada*, vol.1, núm.17, pp. 121-126.

Autorización y Renuncia

Los autores autorizan a LACCEI para publicar el escrito en las memorias de la conferencia. LACCEI o los editores no son responsables ni por el contenido ni por las implicaciones de lo que esta expresado en el escrito.