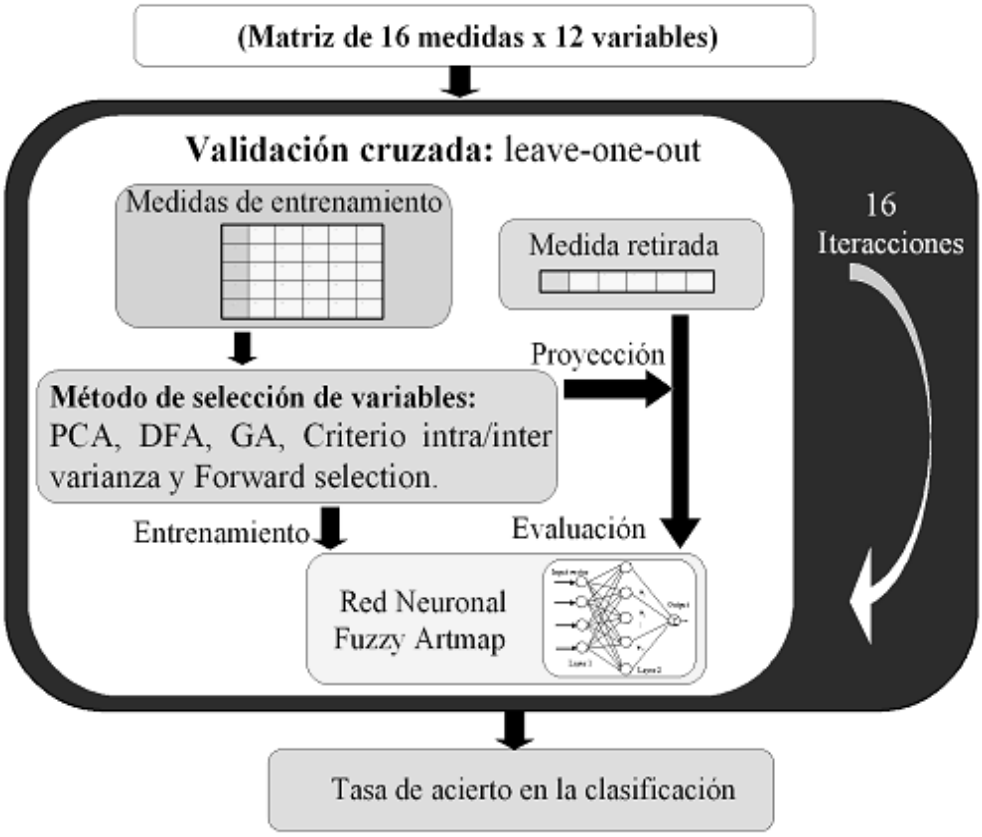
# CAPÍTULO 2

## *Optimización de resultados mediante algoritmos de selección de variables*



# 2.1 Introducción

## 2.1.1 Antecedentes

Actualmente la industria agroalimentaria no ha encontrado una buena solución para controlar la contaminación fúngica de los alimentos, ya que los métodos existentes son demasiado lentos para ser aplicados de forma generalizada en las plantas de producción. Los métodos basados en el muestreo estadístico como el ELISA (método recomendado para el estudio de poblaciones de muestras) intentan solventar el problema, ya que son bastante más rápidos. Sin embargo, son métodos caros, laboriosos y que requieren de operarios especializados para ser aplicados. Además, los resultados que ofrecen son, a veces, incluso demasiado específicos.

Lo que la industria demanda en muchos casos, son sistemas de "pasa ó no pasa", donde lo que prima es la rapidez más que la especificidad. En definitiva, lo que la mayoría de las industrias necesitan es detectar de una forma precoz, rápida e idealmente en tiempo real, grupos específicos de micro-organismos contaminantes.

En la incesante búsqueda de marcadores de contaminación fúngica se han propuesto otros métodos como los indicadores de la actividad de los mohos alterantes [**1**], la actividad enzimática fúngica [**2**-**3**], o marcadores bioquímicos como el ergosterol y el adesín-trifosfato (ATP), que es una molécula de energía encontrada en los mohos [**4**].

Es bien conocido que los micro-organismos producen un amplio abanico de compuestos volátiles cuando se cultivan en el seno de un alimento. Recientemente se han realizado diferentes estudios sobre los compuestos volátiles producidos por los mohos cuando crecen en un producto alimenticio, habiéndose centrado por ejemplo en compuestos como el 3-metil-1-butanol, 1-octen-3-ol y 3-octanona como indicadores del crecimiento fúngico en granos [**5**-**6**]. También *Jhonson* en 1997 [**7**], demostró que es posible hacer una clasificación de la calidad de los granos usando un sistema de olfato electrónico (SDOE) en combinación con una red neuronal artificial. A pesar de estos resultados iniciales, se han realizado pocos estudios para determinar el potencial de los patrones de volátiles junto a un sistema de olfato electrónico en la detección precoz de la actividad de mohos alterantes, antes incluso del crecimiento visible, y para poder incluso llegar a distinguir entre especies de mohos. Uno de los estudios presentados [**8**] demostró que era posible distinguir entre cepas de *Eurotium, Penicillium y Wallemia sebi* "in vitro", incluso antes de que se observara visiblemente el crecimiento.

Otro trabajo que evalúa la calidad de los granos de cereales en función de la presencia de hongos como *Aspergillus, Fusarium, o Penecillium,* especies que emiten volátiles como 2-metil-1-propanol, 3-metil-1-butanol, 1-octen-3-ol, 3-octanona, 3-metilfuran, etil-acetato, 2-metil-isoborneol, pone de manifiesto que un SDOE es tan eficiente como un panel sensorial [**9**]. Este mismo grupo ha reportado publicado que con la utilización de sistemas de olfato electrónico en la determinación del contenido del ergosterol y de CFU (Colony Forming Units) se consiguen prestaciones comparables a las obtenidas con cromatografía de gases (GC) combinada con espectrometría de masas (MS). Estos estudios se realizaron para determinar la calidad de los granos de cebada [**10**].

Los productos de bollería, al igual que la mayoría de los alimentos, están sujetos a un deterioro que limita su vida comercial. Los principales problemas son debidos a la pérdida o ganancia de humedad (fundamentalmente por almacenamiento), al enranciamiento, y a las contaminaciones microbianas. Es muy difícil determinar las perdidas económicas de la industria del pan y de los productos de pastelería por causas microbianas. Según informes de Estados Unidos y Alemania, se estima que entre el 1-5% del total de la producción se pierde por este motivo; esto quiere decir que estas pérdidas representan un importante problema económico [**11**,**12**].

Los principales problemas microbianos de los productos de bollería son producidos por mohos y levaduras. Se han realizado diferentes estudios de identificación de la micoflora alterante de los productos de bollería, en los que se destaca la presencia de géneros como el *Aspergillus, el Penicillium y* principalmente, *el Eurotium* [**13**-**16**], así como *Cladosporium, Wallemia* entre otros [**17**-**19**]. El crecimiento de estos mohos en los productos de bollería durante el almacenamiento, especialmente de las especies del genero *Eurotium,* consideradas los más alterantes, ocasiona serios problemas y pérdidas económicas; además pueden generar "*micotoxinas"* como en el caso del genero *Aspergillus* que pueden influir seriamente en la salud.

Uno de los principales problemas con los que se encuentran los científicos que estudian las mejores condiciones para conseguir alargar la vida útil del producto es aprender a visualizar de una forma precoz el crecimiento fúngico, y que éste no sea por la simple observación de la aparición de las manchas de crecimiento microbiano, ya que en esa situación ya es demasiado tarde. Es en este importante aspecto donde el sistema de olfato electrónico puede ser un instrumento de gran utilidad para determinar las mejores condiciones de conservación de los productos de bollería. En definitiva, un equipo que pueda ser capaz de detectar tempranamente estos micro-organismos seria de gran interés industrial y comercial.

Existen ya algunos estudios realizados en los que se describe el uso de un SDOE para monitorizar la contaminación fúngica, como es el caso en los productos de bollería [**20**-**22**], cereales [**23**-**26**], el queso [**27**]), análisis del agua [**28**], el pan [**29**], la carne [**30**], y la leche [**31**].

Sin embargo, a pesar de que existen varios sistemas de olfato electrónico comerciales y de que se han realizado numerosos estudios de investigación con ellos, su uso en aplicaciones industriales reales es todavía inexistente y el caso del control fúngico en la bollería industrial no es una excepción.

En parte, estos problemas son debidos al ruido y a las derivas de los sensores, que degradan los resultados que pueden ser obtenidos. Por estas razones, es muy importante escoger aquellas variables (los parámetros y/o los sensores), que contengan la información más útil y relevante del problema a resolver. De hecho, es aún más importante eliminar aquellas variables que lleven información ruidosa y que suelen ser las responsables de las respuestas erróneas [**32**]. De esta manera, las fuentes de error podrían ser eliminadas, con las consecuentes mejoras en la fiabilidad de los equipos. Además, la eliminación de sensores implica una reducción directa de costes que se vería reflejada en la configuración final. A pesar de las ventajas que puede aportar esta aproximación, todavía no se han divulgado resultados en esta línea.

A partir de estas premisas surgió el proyecto en el que se encuadra el trabajo descrito en este capítulo, proyecto financiado por el Instituto nacional de Investigación y Tecnología agraria y Alimentaria (**INIA**), y realizado en coordinación con la empresa "La Bella Easo"y la Universidad de Lleida. La primera etapa del proyecto consistió en el desarrollo de un SDOE para la detección de diferentes especies de hongos en los productos de bollería industrial.

Para la consecución de este objetivo se planificaron tres etapas claramente diferenciadas, como son:

a) El diseño y construcción del sistema de olfato electrónico
b) La realización de medidas de prueba y evaluación de resultados
c) Optimización del equipo incorporando técnicas de selección de variables, y de reconocimiento de patrones

El diseño y construcción del SDOE implicó integrar los diferentes elementos que componen un sistema de estas características: La cámara de sensores, el sistema de muestreo, el hardware y software, y los demás componentes necesarios para el funcionamiento y control de cada una de sus partes.

Una vez construido el sistema, se realizaron una serie de medidas para probar su capacidad de identificación de micro-organismos, utilizando un conjunto de muestras de bollería industrial (magdalenas contaminadas con cepas de hongos).

A partir de los resultados obtenidos en la clasificación de las medidas, se estudió la manera de optimizar el funcionamiento, buscando como incrementar su exactitud. Para ello se estudiaron diferentes estrategias para seleccionar las variables más significativas y eliminar aquellas que no aportan información, reduciendo de esta manera la dimensionalidad de los datos. Tras una extensa búsqueda bibliográfica, se determinó que los métodos que más se ajustaban al problema fueron los siguientes:

- Principal Components Analysis (PCA) [**33**]: Se trata de un método que usa componentes principales basadas en la varianza de los parámetros originales. Este método puede ser usado para extraer la máxima información de las respuestas de los sensores. Un sensor que tenga los "*loadings*" cercanos a cero para los componentes principales, contribuye poco al modelo y puede ser eliminado. De todas maneras, hay que advertir que el PCA es un método lineal que no funciona muy bien en condiciones no lineales.

- Otro método clásico es el Discriminant Function Analysis (DFA) [**34**], utilizado para discriminar un conjunto de medidas usando los coeficientes del modelo, llamados variables canónicas. Se trata de un método supervisado más potente que el PCA, aunque con riesgos de sobreentreno.

- Los algoritmos genéticos ó Genetic Algorithms (GA) [**35**] son métodos de selección de variables inspirados en la evolución natural. Su aplicación a sistemas de olfato electrónico ha dado resultados altamente esperanzadores [**36,37**].

- Otra posibilidad es seleccionar las variables a través de un criterio de resolución, que realiza un "*ranking*" de todos los parámetros en función de la resolución frente a la clasificación deseada [**38**].

- Por último, los algoritmos heurísticos como Forward Selection, son ampliamente usados en regresión lineal. Básicamente, con estos métodos se escoge una variable en cada interacción y por medio de combinaciones se llega a escoger la combinación óptima [**39**-**41**].

Todos los métodos descritos anteriormente fueron acoplados a una red neuronal fuzzy ARTMAP [**42,43**]. Este tipo de paradigma ha sido evaluado en aplicaciones de olfato electrónico en numerosos trabajos [**44**-**46**]. Teóricamente, tiene muchas ventajas que la hacen muy atractiva en aplicaciones de análisis olfativo. Entre sus características, las redes requieren pocas muestras para ser entrenadas (aprenden muy rápidamente), son fáciles de programar (requieren menos potencia de cómputo que otros paradigmas), y gestionan muy bien las situaciones de derivas en la respuesta de los sensores (puesto que implementan el dilema de la estabilidad-plasticidad durante su ejecución). Por otra parte, no necesitan ser entrenadas con un número similar de medidas de cada categoría, puesto que aprenden acontecimientos raros muy rápidamente.

El problema principal de este trabajo fue el tamaño del conjunto de medidas, puesto que para muchos paradigmas de aprendizaje 16 es un número bajo de medidas, especialmente si se utilizan un mayor número de sensores. Por este motivo, este tipo de red fue la opción ideal para el estudio.

Es importante comentar que, aunque se han realizados trabajos y publicaciones en el tema de la detección de crecimiento fúngico (trabajos que se han citado previamente), los resultados con los sistemas de olfato electrónico no han sido prometedores, ya que según las búsquedas bibliografías se siguen aplicando algunas técnicas inadecuadamente, como es el caso de la DFA en modo no supervisado. Es por eso que en este estudio aplicamos técnicas de selección de variables que permitan incrementar la selectividad de los sistemas de olfato electrónico utilizando una aproximación "honesta".

En los siguientes apartados de este capítulo se describirán los conceptos de cada uno de estos métodos con más detalle.

## 2.1.2 Objetivos

Una parte importante del trabajo realizado en la aplicación que describimos en este capítulo la constituye el desarrollo de un sistema de olfato electrónico optimizado para la detección fúngica de productos de bollería industrial. Por otra parte, el diseño experimental y la logística necesaria para generar, transportar, conservar, medir y extrapolar los resultados obtenidos con las muestras objeto del estudio son aspectos fundamentales para concluir con éxito el trabajo que presentamos, aspectos que también han supuesto un volumen de trabajo significativo del total de horas dedicadas a este estudio.

A pesar de todo este esfuerzo, hay que recalcar que ya existen trabajos en los que se ha intentado evaluar el comportamiento de un sistema de olfato electrónico comercial en la detección fúngica en bollería industrial (trabajos que ya hemos referenciado en el apartado anterior). Por lo tanto, la novedad del trabajo que se presenta en este capítulo no radica en el diseño y evaluación del equipo en la aplicación que nos ocupa (a pesar de que el equipo ha sido desarrollado expresamente para este cometido, en vez de utilizarse un instrumento comercial).

Por lo tanto, el principal objetivo del presente trabajo es evaluar cómo las técnicas de selección de variables pueden mejorar los resultados que se obtienen con un sistema de olfato electrónico en una aplicación real. Para ello se ha construido un SDOE basada en sensores de gases para detectar, identificar y clasificar diferentes especies de hongos (de los géneros *Eurotium, Aspergillus, y penecillium)* que causan la alteración precoz de los productos de bollería industrial.

Mediante el equipo diseñado se evaluará el comportamiento del sistema frente a la detección de micro-organismos. Una vez obtenidos los resultados preliminares, se tomará la determinación de mejorar el funcionamiento del equipo. Para ello, los resultados se optimizarán gracias a la implementación de las diferentes técnicas de selección de variables acopladas a un paradigma neuronal del tipo ARTMAP, haciendo especial hincapié en comparar los resultados finales respecto a los obtenidos sin utilizar técnicas de selección de variables.

# 2.2 Base teórica

El objetivo de este apartado es la descripción teórica de las técnicas de selección de variables y reconocimiento de patrones comentados anteriormente, ya que todas ellas han sido utilizadas en este estudio.

## 2.2.1 Técnicas de selección de variables

Para determinar si un SDOE puede realizar satisfactoriamente una determinada función es necesario realizar una cuidadosa selección de sensores que permitan generar conjuntos óptimos de datos para el trabajo de clasificación encomendado. Una práctica común es empezar realizando medidas con un prototipo que contenga una matriz con muchos sensores, de manera que se obtenga la máxima información posible de cada medida. Está aproximación suele ser muy útil, ya que los recursos dedicados al realizar la medidas son prácticamente los mismos con pocos o muchos sensores (solo se necesita aumentar la capacidad de adquisición del ordenador).

De todas formas, el aumentar la cantidad de información que se recoge no garantiza la obtención de mejores resultados. Hay variables que aportan información útil y otras que aportan solo ruido, por lo que es necesario escoger cuidadosamente las variables y los diferentes algoritmos de reconocimiento de patrones que se van a utilizar. De cada sensor se pueden extraer varios parámetros (estáticos y dinámicos), por lo que se pueden encontrar situaciones en que haya más variables que medidas. Es importante utilizar un criterio de selección de variables que permita reducir la dimensionalidad de los datos sin eliminar la información útil, y minimizando a la vez cualquier interferencia que aporten las variables con ruido.

A continuación se describen las técnicas de selección de variables utilizadas en el trabajo que nos ocupa.

### 2.2.1.1 Principal Components Analysis (PCA)

El análisis de componentes principales es una técnica originalmente propuesta por *Jackson* [**47**], que surgió como respuesta a la creciente cantidad de datos que podían ser obtenidos de cada medida realizada utilizando instrumentos de laboratorio de nueva generación [**48**], como por ejemplo espectrómetros con la capacidad de proporcionar datos característicos de grandes longitudes de onda, diferentes para cada medida. Esta nueva situación crea una saturación de datos cuya consecuencia más probable es la incorrecta extracción de la información que es realmente relevante para la descripción del experimento.

En definitiva, la necesidad que genera el uso del nuevo instrumental de laboratorio (incluyendo el sistema de olfato electrónico) es doble: Es necesario comprimir, y es necesario extraer toda información relevante del voluminoso conjunto de datos obtenidos, ya que en muchas ocasiones la información esencial no depende de variables aisladas sino de la interrelación entre las mismas. El algoritmo PCA aborda estos problemas y por ese motivo es una de las técnicas más utilizadas por los químicos analíticos y, por extensión, por todos aquellos investigadores que trabajan con sistemas de olfato electrónico ya que permite reducir, representar y extraer información relevante al mismo tiempo [**49**].

El PCA es un método que asume colineraridad entre las variables que intervienen. En otras palabras, se trata de un algoritmo lineal que puede funcionar incorrectamente en procesos altamente no lineales, como pueden ser las interacciones químicas entre sensores y compuestos volátiles. De todas formas, funciona sorprendentemente bien en muchas aplicaciones con sistemas de olfato electrónico, sobre todo en las que las concentraciones de volátiles no son muy elevadas y el comportamiento de los sensores no es excesivamente alineal.

La compresión de datos y extracción de información relevante se hace más necesaria en aquellas situaciones en las que existe una falta de selectividad en cada uno de los sensores que componen la matriz. Por ese motivo el análisis de componentes principales es un método idóneo para explotar el concepto de sensibilidades solapadas que se aplica en la mayoría de los SDOE.

El algoritmo PCA suele ser clasificado como un método no supervisado de reconocimiento de patrones, ya que su uso más extendido con los sistemas de olfato electrónico se limita a representar bidimensionalmente un conjunto de medidas, para ver si se pueden determinar agrupaciones ("clusters") espontáneas entre las diferentes medidas realizadas previamente. Sin embargo también existen modificaciones que permiten aplicar el algoritmo en modo supervisado.

- **PCA como método de reducción de variables**

El PCA es un algoritmo lineal que basa su funcionamiento en la correlación entre variables. El análisis busca unas componentes principales (scores), sobre las que se pueden proyectar las contribuciones de cada una de las variables (los "Loads").

Proyectando las variables sobre las dos primeras componentes principales se puede obtener información sobre la relación entre las mismas. La cercanía entre

variables suele presuponer una buena correlación entre ellas. En el caso de que estén en situación completamente opuesta indica que están fuertemente anticorreladas.

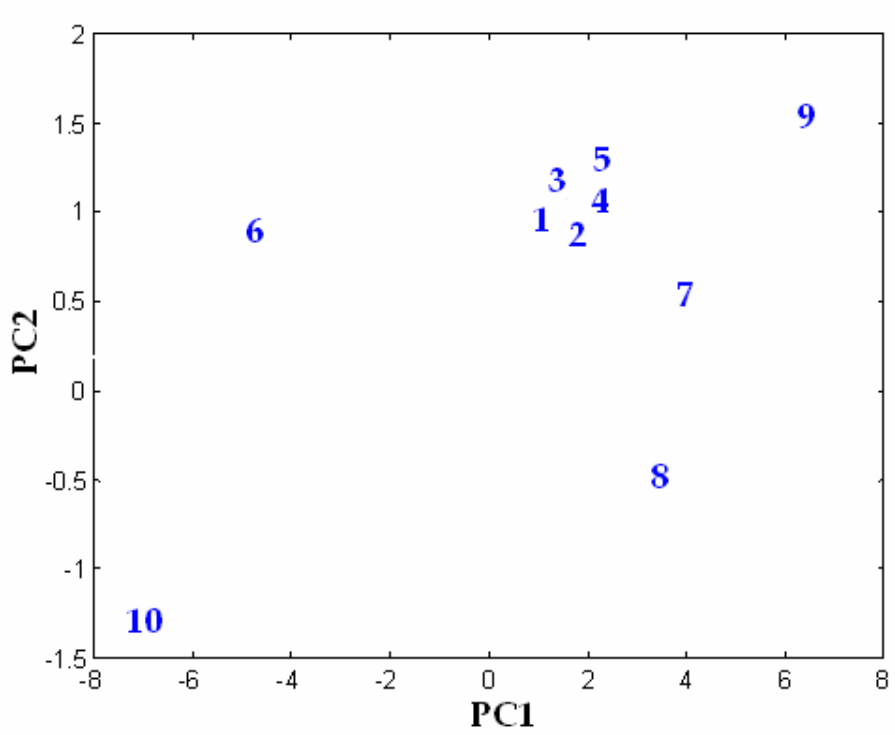



*Figura 2.1: Ejemplo de interpretación de loads en un diagrama PCA*

Si se está trabajando con componentes principales, las variables que estén fuertemente correladas o anticorreladas aportan información similar, lo que indica que son redundantes. El eliminar todas menos una puede permitir reducir la dimensionalidad del vector que define la medida sin perder información importante. Sin embargo, cuando las variables son bastante ruidosas puede ser interesante conservar más de una para intentar promediar su efecto. La figura 2.1 muestra un ejemplo simulado en el que se observa como las variables 1 a 5 aportan información similar (están fuertemente correladas), mientras que la variable 10 está claramente anticorrelada a la variable 9. Así pues, una posible elección para reducir el número de variables podría incluir la 1, 6, 7, 8, 9, representando la variable 1, a todas las de su agrupación y la 9 a su variable anticorrelada numero 10.

### 2.2.1.2 Discriminant Function Analysis (DFA)

DFA es una técnica de reconocimiento de patrones paramétrica, que proporciona una forma más eficiente de clasificación [**50**]. Se trata de un método lineal y supervisado [**51**] (es decir, que el resultado de la clasificación de las medidas que deben ser discriminadas durante el entrenamiento son conocidas antes de que el análisis se realice). Geométricamente, las filas de la matriz de respuesta pueden ser consideradas como puntos en el espacio multi-

dimensional. Los ejes discriminantes son determinados en este espacio, el cual es una proyección para conseguir una óptima separación de las clases predefinidas. Al igual que un PCA, el DFA encuentra nuevos ejes ortogonales (factores) como una combinación lineal de las variables de entrada. Pero a diferencia del PCA, DFA computa los factores y minimiza la varianza dentro de cada clase, maximizando la varianza entre clases. El primer factor será la dimensión de mayor alcance, pero los factores posteriores pueden representar dimensiones significantes de diferenciación.

En la respuesta de la matriz de datos $\boldsymbol{R}$, cada fila es un vector de medidas sobre $n$ variables para una sola observación. El vector $g$ es un vector que define las categorías que se van a clasificar. Dos observaciones están en la misma clase si tienen el mismo valor en $g$. Si $\boldsymbol{W}$ es definida como la suma de cuadrados de los grupos y la matriz cruzada de productos, y $\boldsymbol{B}$ como la suma de cuadrados entre grupos y la matriz cruzada de productos, entonces los autovectores de $\boldsymbol{W}^{-1}\boldsymbol{B}$ son los coeficientes para los factores, tal y como se ilustra en la ecuación (1), donde $\overline{\boldsymbol{R}}$, es la matriz de respuesta centrada ($\boldsymbol{R}$ con columnas centradas por la resta de sus medios).

$$\boldsymbol{F} = \overline{\boldsymbol{R}}\, x\, autovec \quad (\boldsymbol{W}^{-1}\boldsymbol{B}) \qquad (1)$$

Específicamente, la primera columna de $\boldsymbol{F}$ es el primer factor o combinación lineal de $\boldsymbol{R}$ columnas que proporciona la máxima separación entre grupos. La segunda columna de $\boldsymbol{F}$ tiene la máxima separación entre grupos, bajo la condición de ortogonalidad respecto al primer factor, y así sucesivamente [**52**].

Es por esto que la principal diferencia con relación al PCA, es que DFA es un método que determina las variables a partir de una clasificación conocida de las medidas de entrenamiento de la aplicación. Es decir, realiza una clasificación a partir de un entreno previo supervisado.

En está técnica se escogen los coeficientes más significativos, las variables canónicas. Similarmente a lo que ocurre con el PCA, los loadings (autovectores), son usados para determinar si hay sensores irrelevantes o redundantes que puedan ser eliminados. Utilizando las coordenadas de proyección de cada medida respecto a la base octogonal de autovectores, se puede intentar aplicar a una red fuzzy ARTMAP para comprobar si esta técnica mejora la selección de muestras.

En [**53**] se aplicó esta técnica para la detección precoz de crecimiento fúngico en productos de bollería industrial mediante cromatografía de gases y espectrometría de masas. Los coeficientes (autovalores) obtenidos desde el DFA

fueron usados como variables de entrada de una red neuronal fuzzy Artmap, cuya misión fue categorizar los cultivos de hongos según la especie o género del agente contaminante.

### 2.2.1.3 Criterio de selección basado en intra/intervarianza

El criterio de selección basado en intra/intervarianza consiste en evaluar cada parámetro disponible atendiendo a un criterio de resolución, para posteriormente escoger aquellas variables que superan un valor umbral relativo a este factor de mérito. Esta resolución está basada en el cálculo de la relación entre varianza intraclase y la varianza interclase [**54**].

La clave de este método es definir una relación entre la variación media entre las medidas de la misma categoría (variación intraclase, relacionadas con el repetitividad del parámetro), y el promedio de las distancias entre las centroides de diferentes categorías (varianza interclase, relacionada con la selectividad del parámetro). El criterio se define para seleccionar un subconjunto óptimo de entre todos los parámetros disponibles, es decir, se seleccionan aquellos que demuestran una variación interna pequeña (buena repetitividad) combinada con una alta variación externa (buena selectividad). Esto es equivalente a seleccionar las variables con mayor poder de discriminación y fiabilidad para el problema de clasificación bajo estudio.

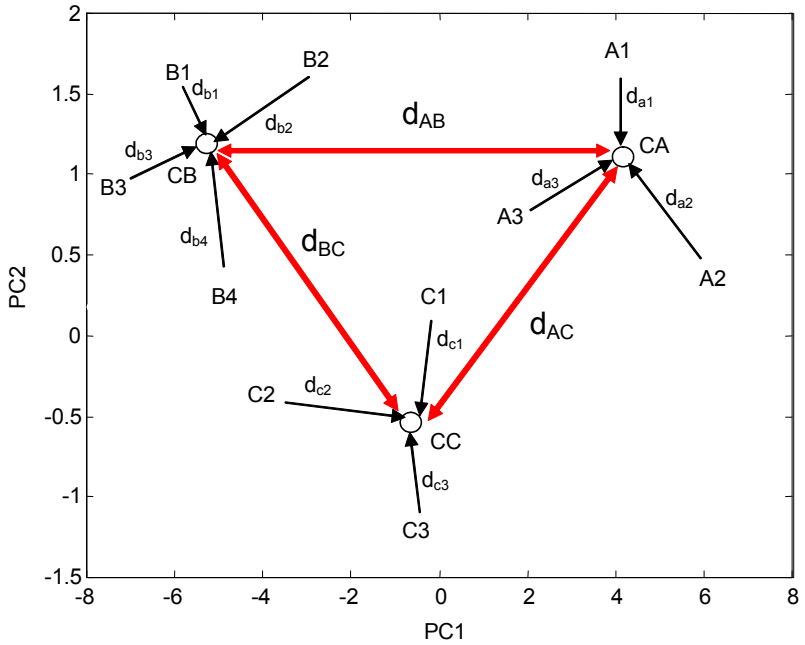



*Figura 2.2: Cálculo del poder de resolución con 10 medidas de tres clases diferentes*

La figura 2.2 y las ecuaciones (2), (3) y (4) detallan un cálculo de resolución a modo de ejemplo.

Los pasos detallados de este cálculo son los siguientes:

1. Cálculo de la centroide para cada una de las clases de medidas existentes en la proyección.
2. Cálculo de la distancia media entre centroides (varianza interclase).
3. Cálculo, para cada clase, de la distancia media entre todas sus medidas y su centroide.
4. Obtención del valor medio del cálculo anterior (varianza intraclase).
5. División de la varianza interclase por la varianza intraclase (poder de resolución).

Para cada clase, distancia intraclase media (varintraclase):

$$v_A = \frac{1}{3}(da1 + da2 + da3) \quad v_B = \frac{1}{4}(db1 + db2 + db3 + db4) \quad v_C = \frac{1}{3}(dc1 + dc2 + dc3) \quad (2)$$

Distancia intraclase media: $vm = \frac{1}{3}(v_A + v_B + v_C)$ (3)

Distancia media interclases: $vmi = \frac{1}{3}(d_{AC} + d_{AB} + d_{BC})$ (4)

Poder de resolución: $res = \frac{vmi}{vm}$ (5)

La ecuación (5) define este criterio, que mide de alguna manera el poder de la resolución de cada una de las variables relacionadas a la diferenciación entre las categorías a ser identificadas.

#### 2.2.1.4 Forward selection

Los algoritmos de búsqueda secuencial o determinística son estrategias que reducen el número de variables aplicando búsquedas locales. Uno de los métodos más comunes es el "forward selection" **[55].** Es usado, por lo general, en regresión lineal. Su popularidad se debe a que tiene la característica de ser simple y rápido.

El procedimiento empieza considerando cada una de las variables individualmente. En esta primera fase se selecciona la variable que da el mejor valor obtenido por el criterio de selección, criterio calculado generalmente por medio del error de predicción (PRE) sobre los datos de validación.

Una vez que la variable que da la mejor predicción ha sido seleccionada, el proceso intenta encontrar nuevamente una segunda variable que, combinada con la primera, dé la mejor capacidad de predicción del conjunto (menor PRE). El proceso continúa hasta que el error de predicción se incrementa debido a la adición de cualquiera de las variables restantes. En este punto finaliza el proceso de búsqueda.

En este trabajo se ha implementado dicho criterio mediante el error de predicción obtenido utilizando la red neuronal fuzzy ARTMAP como clasificador.

#### 2.2.1.5 Algoritmos Genéticos (GA)

Los algoritmos estocásticos de búsqueda intentan mejorar el tiempo computacional de cálculo de los métodos exponenciales y evitar la tendencia de los métodos secuénciales a quedarse atrapados en mínimos locales del problema de optimización. El más conocido de estos métodos de selección es el basado en la utilización de algoritmos genéticos (GA) [**55**].

Los algoritmos genéticos son procesos de búsqueda basados en los principios de la selección y la evolución natural. Las posibles soluciones a un problema son codificadas en forma de cadenas binarias, y la búsqueda se inicia con una población de posibles soluciones generadas aleatoriamente.

En el problema de la selección de variables, cada posible combinación es codificada con una cadena binaria tan larga como parámetros se consideren para encontrar la combinación óptima de variables. En dicha cadena, cada variable tiene asignada una posición o bit, de manera que una posible solución vendrá descrita por una sucesión de unos y ceros indicando la presencia (con un 1) o la ausencia (con un cero) de cada una de las variables en esa combinación particular. En las condiciones genéticas cada variable configura un gen y una combinación concreta de presencia/ausencia de variables forman un cromosoma.

Por ejemplo, en un problema de selección en el que inicialmente se tienen 8 variables, un posible cromosoma sería el 00110101. Esto puede traducirse en que las variables 3, 4, 6 y 8 serán usadas en el proceso de modelado (para entrenar y validar una red neuronal) y las variables 1, 2, 5 y 7 serán omitidas [**56**].

En este tipo de algoritmos, cada miembro de la población, que representa una posible solución, es testada con algún criterio objetivo, de manera que cada uno de los miembros de la población se valora en función de su "*fitness*" (valor del criterio). Este criterio puede ser, por ejemplo, el error de predicción. A las soluciones mejor valoradas se les permite sobrevivir y pasar a la siguiente iteración ("generación"), mientras que las soluciones de peor fitness desaparecen en las sucesivas generaciones.

El algoritmo genético prosigue hasta que iguala o supera el fitness establecido como meta, hasta que exista una convergencia en la población (de manera que

un determinado porcentaje de sus miembros acaben siendo idénticos) o hasta que se llegue al número máximo de iteraciones.

La creación de los miembros de la población de una nueva iteración se realiza a partir de combinaciones y mutaciones entre los miembros supervivientes de la anterior iteración. La combinación consiste en cruzar, de dos en dos, los miembros de la antigua población, creando nuevos individuos en los que los primeros N bits son de uno de los "padres" y el resto del otro. N, denominado el "crossover point" o punto de cruce, es un valor aleatorio. La figura 2.3 muestra este concepto. Por otro lado, la mutación de un miembro consiste en el cambio aleatorio de algún bit de su cadena.

Existe un teorema en el que se demuestra que la iteración sucesiva con estas reglas permite sobrevivir a las combinaciones que mejor cumplen con el criterio preestablecido, llegándose a una serie de patrones (denominados "schematas") que convergen hacia la solución óptima del problema.

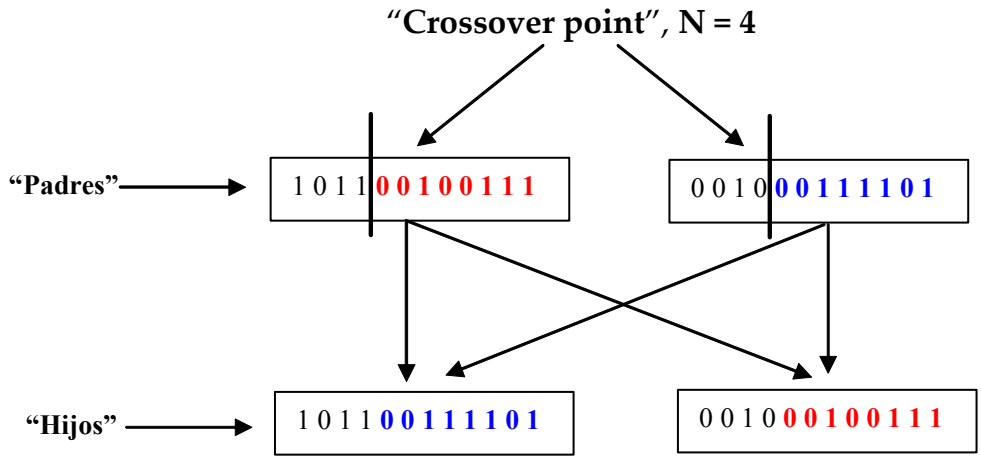



*Figura 2.3: Esquema explicativo del cruce entre padres para la generación de hijos*

El algoritmo genético utilizado en este trabajo, programado bajo el entorno Matlab, presenta las siguientes características:

- Implementa técnicas de combinación ("crossover") simples o dobles
- En cada iteración descarta la mitad de la población y permite sobrevivir a la mitad de los miembros con un mejor fitness.
- El criterio o fitness que se utiliza se basa en una validación cruzada, escogiendo de forma aleatoria grupos de medidas de las que obtiene un error cuadrático medio entre la predicción y el valor real. Por supuesto, a menor error, mejor "fitness".

Los parámetros que utiliza dicha función de Matlab son los siguientes:

- *Numero de miembros de la población:* Indica el número de individuos que deben considerarse y evaluarse en cada iteración.
- *Numero de términos iniciales:* Número de miembros en la población inicial.
- *Numero máximo de generaciones:* Número máximo de iteraciones antes de parar el proceso.
- *Criterio de convergencia:* Porcentaje de miembros de la población idénticos para considerar que se ha convergido a una solución y que por lo tanto se deben finalizar las iteraciones.
- *Probabilidad de mutación:* Fracción de bits que deben ser cambiados en cada generación.
- *Tipo de combinación:* "crossover" simple o doble
- *Numero de subconjuntos en validación cruzada:* Como su nombre indica, numero de subconjuntos que serán iterativamente utilizados para entrenar y evaluar el fitness de cada miembro de la población.
- *Numero de iteraciones:* Número de veces que se deben formar los subconjuntos para obtener un fitness medio con menor "rizado" debido a la aleatoriedad.

En diferentes trabajos realizados por investigadores, se han buscado nuevas estrategias para aumentar la fiabilidad de los métodos de selección de variables a partir de los algoritmos genéticos. Una de ellas es acoplar la técnica de selección a una red neuronal fuzzy ARTMAP. En [**57**] se utilizó está nueva configuración para analizar un conjunto de vapores simples y mezclas binarias de 3 compuestos orgánicos volátiles, usando una matriz de 12 sensores de gases basados en óxidos metálicos. Con esta nueva configuración (acoplando algoritmos genéticos con redes fuzzy ARTMAP) se consiguió reducir a 9 el número de variables usadas a partir de las 120 iniciales. Además, la reducción incrementó significativamente la capacidad de clasificación mediante la red fuzzy ARTMAP. El porcentaje de validación fue del 91.67 % y 88.33% en la identificación de vapores simples y de sus mezclas binarias respectivamente. Para determinar el fitness se utilizó la validación cruzada de orden uno "leave one out" aplicada a una red neuronal fuzzy ARTMAP. El fitness se evalúo a través del error de predicción.

Aunque el algoritmo genético selecciona, de forma aleatoria, variables que son irrelevantes para predecir las especies o su concentración, el proceso iterativo se encarga de eliminar las variables que no contribuyan a disminuir el error de predicción. Esto permite construir modelos mas parsimoniosos (es decir, usando pocas variables), que son más exactos y más robustos ante medidas nuevas, presentado una mejor capacidad de generalización.

## 2.2.2 Redes Art

La teoría de la resonancia adaptativa (ART) fue introducida como una teoría que intentaba emular la manera en cómo el cerebro humano procesa la información [**58,59**]. Desde entonces, esta teoría ha evolucionado hacia una serie de algoritmos neuronales para el aprendizaje no supervisado. Estos algoritmos son capaces de crear clases estables ante la presentación de secuencias de entrada arbitrarias con un ritmo de aprendizaje rápido o lento. Dentro de estos algoritmos se pueden destacar el ART1 [**60**], ART2 [**61**] y ART3 [**62**].

Para describir correctamente la red neuronal utilizada en estas pruebas, la red fuzzy ARTMAP, primero es necesario explicar la red no supervisada Fuzzy Art, ya que la primera utiliza dos redes Fuzzy Art sincronizadas para realizar su cometido.

### 2.2.2.1 La Red neuronal fuzzy ART

La red fuzzy Art [**63**], es una evolución del algoritmo ART1. Éste último es capaz de categorizar de forma estable entradas arbitrarias binarias. Fuzzy Art, siguiendo el mismo esquema, generaliza esta función a vectores de entrada analógicos con coordenadas comprendidas entre 0 y 1. Para ello substituye los operadores intersección ($\cap$) y unión ($\cup$) de ART1 por los operadores MIN ($\wedge$) y MAX (V), respectivamente, de la teoría de lógica difusa [**64**]. Este cambio, con la ayuda de la codificación complementaria ("complement coding"), que preserva la información de amplitud a la vez que normaliza los vectores de entrada, permite implementar un algoritmo de clasificación no supervisada de gran rapidez de aprendizaje. En la figura 2.4 se muestra un esquema del algoritmo.

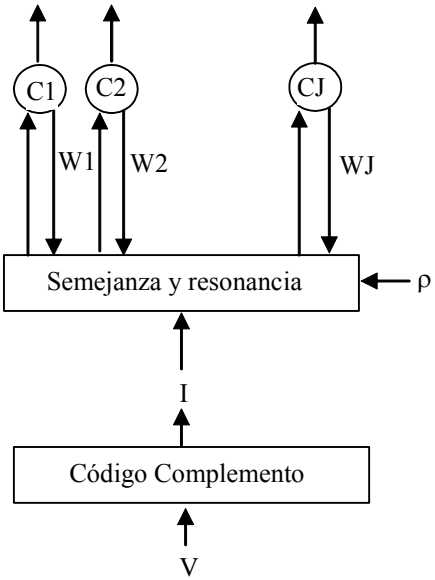



*Figura 2.4: Esquema de la red fuzzy Art*

Su modo de funcionamiento es el siguiente: Cada vez que la red recibe un nuevo vector de entrada *V*, reacciona activando uno y solo uno de los nodos de salida. Cada uno de estos nodos representa una de las diferentes clases o categorías que se han creado con las entradas anteriores. En caso de que la medida no se parezca lo suficiente a ninguno de los nodos ya asignados, se crea uno nuevo que representará una nueva clase cuyo primer miembro será este vector.

Desde el punto de vista operativo, este algoritmo cuenta con dos parámetros que controlan su funcionamiento. El parámetro de vigilancia ("vigilance parameter") determina lo riguroso que debe ser el algoritmo a la hora de agrupar medidas. Un parámetro de vigilancia muy cercano a la unidad implica una clasificación muy exigente, de manera que dos medidas deben ser muy parecidas para ser agrupadas en una misma clase. Por el contrario, un parámetro cercano a cero permite la agrupación de medidas poco parecidas, lo que, como resultado, genera una red con pocos nodos de salida, ya que el número de clases diferentes es reducido. Por su parte, el ritmo de aprendizaje queda controlado por el parámetro $\beta$, siendo su valor igual a la unidad para un aprendizaje rápido e igual a cero en caso de que no se deba aprender más.

- **Algoritmo**

Incluimos, a continuación, una descripción esquemática del algoritmo:

- *Vector de entrada:* Cada uno de los vectores de entrada V, es un vector M-dimensional donde cada una de sus componentes tiene coordenadas incluidas en el intervalo [0,1].
- *Codificación complementaria:* A partir del vector de entrada V, se crea un nuevo vector normalizado I de dimensión 2M en el que la componente $I_{J+M} = 1-I_J$.
- *Vector de pesos del nodo de salida j (categoría j):* $W_J$. Inicialmente, $W_{J1}=W_{J2}=W_{J2M}=1$
- *Velocidad de aprendizaje:* ("learning rate"), $\beta$ entre [0 1]. Aprendizaje rápido, $\beta =1$; Aprendizaje lento, $\beta << 1$; Sin aprendizaje, $\beta =0$
- *Parámetro de vigilancia*: $\rho$ entre [0 1]. $\rho$ cercano a cero implica menos categorías al agrupar con criterios de semejanza poco exigentes, $\rho$ cercano a uno implica muchas clases, cada una con pocos miembros pero muy parecidos entre sí.
- *Parámetro de selección:* $\alpha>0$. Debe ser muy cercano a cero. Sirve para deshacer igualdades. Un valor típico es 0.001.
- *Selección de categoría:* Para cada vector de entrada V y cada categoría j se calcula la función de selección o semejanza $T_J(V)$ como indica la ecuación 6:

$$T_J = \frac{|I \Lambda W_J|}{\alpha + |W_J|} \tag{6}$$

Donde el operador AND ( $\Lambda$ ) en lógica difusa se define como:

$$A \Lambda B = \min (A, B) \tag{7}$$

Y la norma | | se define como:

$$|I| = \sum_{i=1}^{2M} I_i \tag{8}$$

A partir de aquí inicialmente se escoge la categoría jota para la que Tj(V) es máximo.

- *Resonancia o reset:* Se dice que aparece resonancia si se cumple la desigualdad 9:

$$\frac{|I \Lambda W_J|}{|I|} \geq \rho \tag{9}$$

En ese caso, se activa el nodo de salida (categoría) j como respuesta al vector de entrada V, lo que quiere decir que la red clasifica al vector V como de clase j. Además, se ejecuta el proceso de actualización de los pesos de dicha categoría.

En el caso de que no se cumpla la desigualad se produce un reset: El sistema desactiva temporalmente el nodo j, y vuelve a escoger una categoría siguiendo el criterio de máxima semejanza (ecuación 6). Si ninguna categoría "resuena", se crea un nuevo nodo para el vector de entrada V.

- *Aprendizaje:* Una vez activada la categoría j debido al vector V, sus pesos son actualizados según la ecuación 10:

$$W_J^{NUEVO} = \beta\left(I \Lambda W_J^{ANTERIOR}\right) + (1 - \beta) W_J^{ANTERIOR} \tag{10}$$

Si se quiere un aprendizaje rápido, se utiliza una β= 1. Para un aprendizaje nulo β=0. En general, para medidas ruidosas no interesa poner β=1. Sin embargo, cuando el número de medidas es bajo y se requiere de un aprendizaje estable se puede demostrar que eso se consigue con β=1.

### 2.2.2.2 Red neuronal fuzzy ARTMAP

Las redes de tipo ARTMAP son una clase de redes neuronales que implementan un aprendizaje supervisado, y una posterior clasificación de vectores multidimensionales de entrada en una serie de categorías de salida.

La red fuzzy ARTMAP [**65**] proviene de la red ARTMAP [**66**] con las mismas transformaciones que permiten definir la red fuzzy Art a partir de la red ART1. En definitiva, la red fuzzy ARTMAP es una generalización a vectores analógicos (con componentes comprendidas entre cero y uno) de la red binaria ARTMAP.

La red fuzzy ARTMAP presenta múltiples ventajas que la hacen muy interesante para las aplicaciones con sistemas de olfato electrónico [**67**]. De entre todas ellas destacaremos las siguientes:

- Aprendizaje rápido (con muy poca carga computacional) de las medidas que se presentan en entrenamiento, lo que permite programar el algoritmo en dispositivos programables de bajo coste, aplicar validaciones cruzadas de orden 1 y probar con diferentes combinaciones de parámetros.
- Aprendizaje con un conjunto reducido de medidas de entrenamiento, algo muy interesante en cualquier aplicación experimental en la que sea costoso la obtención de conjuntos de medida extensos. La red presenta una habilidad particular para aprender rápidamente eventos singulares que aparecen muy pocas veces en el conjunto de entrenamiento. Por lo tanto, en dicho conjunto no es necesario que haya el mismo número de medidas de cada clase para que funcione correctamente.
- Aprendizaje continuo de nuevas características sin olvidar lo aprendido con anterioridad, algo muy útil para adaptarse a derivas producidas por sensores.
- En comparación con otros tipos de redes neuronales, fuzzy ARTMAP determina automáticamente las neuronas de su capa oculta. Además maximiza el poder de generalización aprendiendo al 100% el conjunto de entrenamiento.
- Una vez entrenada, es posible extraer reglas de clasificación a partir de los pesos obtenidos tras el periodo de entrenamiento, lo que puede dar luz sobre los procesos internos y como influyen en la categorización de resultados.

Resumiendo, la red fuzzy ARTMAP es una red de clasificación con aprendizaje supervisado. En una fase de entrenamiento la red necesita que se le suministre un conjunto de medidas. Cada medida debe contener un vector de entrada, que

detalla los parámetros medidos en cada experiencia, y un vector de salida que codifica la categoría que se le debe asignar. Posteriormente, en la fase de evaluación solo se suministra el vector de entrada y la red clasifica dicha medida siguiendo los criterios que ha aprendido en la fase de entrenamiento.

- **Algoritmo**

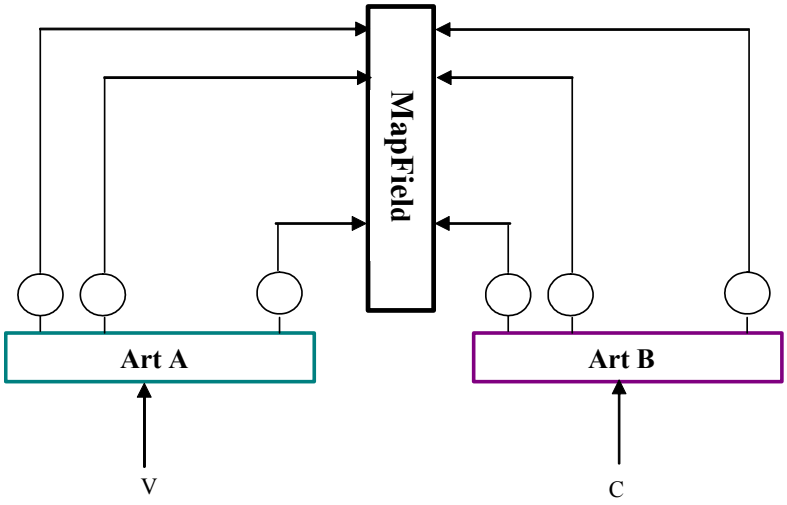



*Figura 2.5: Esquema general de una red fuzzy ARTMAP*

Básicamente, una red fuzzy ARTMAP está formada por dos redes fuzzy Art conectadas entre sí por un vector de relaciones denominado "*mapfield*". A una de las dos redes (la que denominaremos A) le llegan los vectores de entrada (V). A la red B le llegan, en la fase de entrenamiento, los vectores que codifican la categoría correcta de cada medida del conjunto de entrenamiento (C). La figura 2.5 esquematiza este concepto.

Inicialmente, en la red A el vector de vigilancia es cero. En la red B se suele dar un valor igual a la unidad, ya que las medidas que deban ser clasificadas conjuntamente enviaran a la red B codificaciones idénticas. Además, cualquier vector de codificación diferente, por parecido que sea al resto, debe ser detectado y debe activar una neurona de salida diferente en la red B.

Cada vez que se suministra una medida de entrenamiento, la red A activa un nodo y la red B activa otro. El mapa que las une aprende a relacionar nodos activados. De esta forma, a cada nuevo nodo que se activa en la red A se le asocia un nodo en B. Cabe destacar que, normalmente, cada uno de los nodos en B será imagen de varios nodos A (cada categoría contiene varias medidas), mientras que cada una de las categorías creadas en A sólo tendrá una imagen en B (cada medida solo puede pertenecer a una categoría).

Cuando una nueva medida activa un nodo en A ya existente, se comprueba si la imagen de ese nodo asignada por el mapfield coincide con el nodo que se ha activado en B paralelamente. En el caso de que no coincidan, se incrementa el valor del parámetro de vigilancia hasta que la neurona que se active en A tenga por imagen la neurona activada en B. Si no se encuentra ninguna se creará una nueva y el mapa le asignará como imagen el nodo B activado.

En definitiva, el valor de vigilancia en A solo se incrementa lo estrictamente necesario para que la red A separe en nodos diferentes las medidas que deben estar clasificadas en diferentes categorías. Suponiendo un parámetro de aprendizaje igual a la unidad para ambas redes se puede demostrar que este algoritmo aprende a clasificar correctamente el 100% de los vectores de entrenamiento. Además, ese aprendizaje es rápido y estable. A continuación se detalla el algoritmo de forma esquemática:

- *ARTa:* Red fuzzy Art a la que llegan los vectores de entrada
- *ARTb:* Red fuzzy Art a la que llegan los vectores de salida
- *Mapfield:* Módulo de mapeado que relaciona nodos de salida de ARTa con nodos de salida de ARTb
- *Entradas:* Supondremos que el vector de entrada de cada medida es V y el vector que codifica su clasificación correcta C.
- *Normalización:* Los vectores V y C se normalizan con codificación complementaria, pasando a ser los vectores I y D respectivamente.
- *Pesos:* El vector de pesos de la categoría k de la red ARTa se denominará $W_{AK}$. El vector de pesos de la categoría j de la red b se denominará $W_{BJ}$. El mapfield tiene un solo vector de longitud igual al número de nodos activados en A. La componente k-ésima del vector indica el nodo imagen en ARTb del nodo k de la red ARTa.
- *Match tracking:* Originalmente, el parámetro de vigilancia de ARTa, $\rho_a$, es un valor base. Si una medida activa un nodo J en ARTa cuya imagen a través del mapfield no coincide con la activación producida en ARTb, entonces el valor de vigilancia en ARTa se incrementa según la ecuación 11 , lo que forzará la activación de un nodo diferente en ARTa:

$$\rho_a = \frac{\left| I \Lambda W_J^A \right|}{\left| I \right|} \tag{11}$$

- *Modo de evaluación:* La red ARTb se desactiva. La red ARTa recibe un vector que hace que se active su nodo k. La salida de la red es el valor de la componente k-ésima del vector de mapfield.

- **Implementación para aplicaciones de olfato electrónico**

Aunque existen varios paquetes comerciales que implementan algoritmos basados en fuzzy ARTMAP, no existe ninguno que implemente la red en su definición estricta. Por razones de flexibilidad, se implementó la fuzzy ARTMAP bajo el entorno *Matlab 6.1*, programándola a partir de funciones más simples que son anidadas en una función con un bucle principal. Como la red fuzzy ARTMAP está formada por dos redes fuzzy Art, sus funciones principales incluyen llamadas a la función fuzzy Art incluidas en este paquete.

En la programación del algoritmo fuzzy ARTMAP se ha distinguido entre una función de entreno (en la que a partir de unas medidas de entrada y sus salidas correspondientes se calculan los pesos) y una de evaluación (en la que a partir de unas medidas de entrada y los pesos se devuelven las clases a las que pertenecen según la red entrenada previamente).

Para conseguir que la red fuzzy ARTMAP funcione correctamente en aplicaciones con el sistema de olfato electrónico, se han implementado diferentes versiones que modifican ligeramente el algoritmo original, con el objetivo de inmunizar a la red ante la presencia de "*outliers*" (valores erróneos) en el conjunto de entrenamiento.

| Modelo de la red | Topología | Supervisada /no Supervisada | Regla | Inform. Entrada/ Salida | Autores |
|---|---|---|---|---|---|
| Teoría de resonancia adaptativa (ART1) Fuzzy ART (**Versión 1**) | 2 capas FeedForward/ Feeback, conexiones laterales y autorrecurrente | No supervisada | Competitivo (Resonancia Adaptativa) | Binarias | Carpenter Grossberg 1986 |
| Teória de resonancia adaptativa (ART2) Fuzzy ART (**Versión 2**) | 2 capas FeedForward/ Feeback, conexiones laterales y autorrecurrente | No supervisada | Competitivo (Resonancia Adaptativa) | Análogo | Carpenter Grossberg 1987 |
| Teória de resonancia adaptativa Fuzzy ARTMAP (**Versión 3**) | 2 capas FeedForward/ Feeback, conexiones laterales y autorrecurrente | Supervisada | Competitivo (Resonancia Adaptativa) | Binarias o bien análogos | Jesús Brezmes 1999 |

***Tabla 2.1: Diferentes topologías de redes ART***

En la tabla 2.1 se describen los modelos de red basados en la teoría de resonancia adaptativa, y que se han empleado con sus características más representativas (topología, mecanismos de aprendizaje, tipo de información de entrada y salida, la forma de representación de la información, señales de Entrada/Salida) y sus respectivos autores.

Como se puede ver, a partir de las dos primeras versiones propuestas por Grossberg, Carpenter, fundamentales en el desarrollo de la teoría de la resonancia adaptativa ART, Jesús Brezmes (director de esta tesis) desarrolló una tercera versión de la topología de red fuzzy ARTMAP, con la pretensión de optimizar al máximo la capacidad de la red en problemas de olfato electrónico. Los trabajos de estos autores vienen referenciados para dar si se requiere más información sobre la estructura y funcionamiento de cada topología.

La solución que mejor resultados prácticos ha propiciado es la modificación del algoritmo original sólo en la etapa de evaluación. En dicha fase el algoritmo modificado ignora el factor de vigilancia alcanzado en la fase de entrenamiento y busca, de forma iterativa, el valor más elevado que permita clasificar la nueva medida en alguna de las categorías existentes.

Junto a las funciones necesarias para implementar las diferentes versiones de la red fuzzy ARTMAP, se han diseñado funciones de validación cruzada de orden 1 ("leave-one-out") que permiten evaluar el método dando como resultado un porcentaje de aciertos en la clasificación de las medidas. Este método de validación cruzada se ha implementado en este trabajo, ya que en una situación como la que describimos hay que aprovechar al máximo el escaso número de medidas disponibles.

# 2.3 Desarrollo del equipo

## 2.3.1 Esquema general

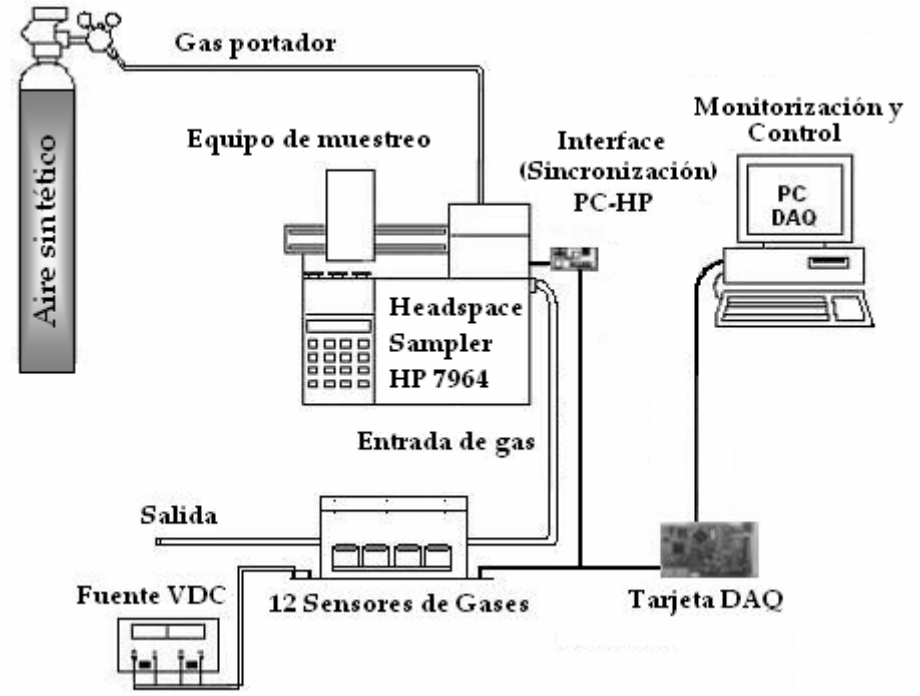



*Figura 2.6: Esquema general de funcionamiento*

Inicialmente se ha creído conveniente describir el funcionamiento general del equipo de medida. El esquema de la figura 2.6 muestra cada una de las partes o elementos que componen el sistema. En él se puede distinguir el modulo de muestreo compuesto por una botella de aire sintético utilizada como gas portador y un equipo Hewlett Packard (Headspace Sampler 7694) como muestreador automático de espacio de cabeza (ver figura 2.7). Este equipo es de gran utilidad ya que permite preparar y automatizar la medida de un amplio número de muestras con muy buena reproducibilidad [**68**].

*Figura 2.7: Sistema de muestreo HP*

El equipo de espacio de cabeza se compone generalmente de una banda transportadora (donde están ubicados los viales que contienen la muestra), un horno donde la muestra es pre-acondicionada, y un regulador de flujo para el gas portador (Ej: aire sintético).

Otro de los módulos importantes es el sistema de medición, compuesto por una cámara de sensores de gases con 12 unidades de sensores de óxidos metálicos que reaccionan ante los diferentes volátiles generados por poblaciones de hongos. Para el control y monitorización de las señales de los sensores se utilizó una tarjeta de adquisición de datos conectada con la cámara de sensores y el PC. Para suministrar la corriente necesaria para la polarización de los sensores se utilizó una fuente DC con buena estabilidad.

Un aspecto muy importante en el proceso de medida es la sincronización entre el equipo de muestreo (HP, de ahora en adelante) y el PC, con el fin de obtener un conjunto de medidas repetibles, y así realizar el proceso de adquisición de forma automática sin la necesidad de la intervención del operario.

En definitiva, el proceso de medida consta de cuatro etapas principales, las cuales se describen brevemente a continuación:

**1) Acondicionamiento de la muestra mediante un conjunto de viales**
Se coloca el conjunto de viales con el contenido de la muestra preparada previamente para ser analizada sobre un depósito para viales del equipo de muestreo.

**2) Sincronización del equipo de muestreo y PC**
En está etapa se inicia el proceso de medida en el momento que el PC da una señal de aviso al equipo de muestreo HP para comenzar la adquisición de medidas.

**3) Transporte de volátiles a la cámara de sensores**
Tras el paso anterior, un flujo transporta los volátiles desde el espacio de cabeza del vial en fase gas (desde su fase sólida o liquida) hacia la cámara de medida.

**4) Adquisición de medidas, análisis y procesado de datos**
Al final del proceso se almacenan las medidas que fueron obtenidas a partir del número de viales a analizar. Una vez este conjunto de medidas haya sido almacenado, se analizan cada una de las muestras por medio de las señales capturadas a través de los sensores, y se evalúan con diferentes métodos de procesado de datos.

## 2.3.2 Hardware

En la figura 2.8, las fotografías (a) y (b) muestran la parte interna del sistema de medición, con todos los componentes necesarios para el funcionamiento del equipo.

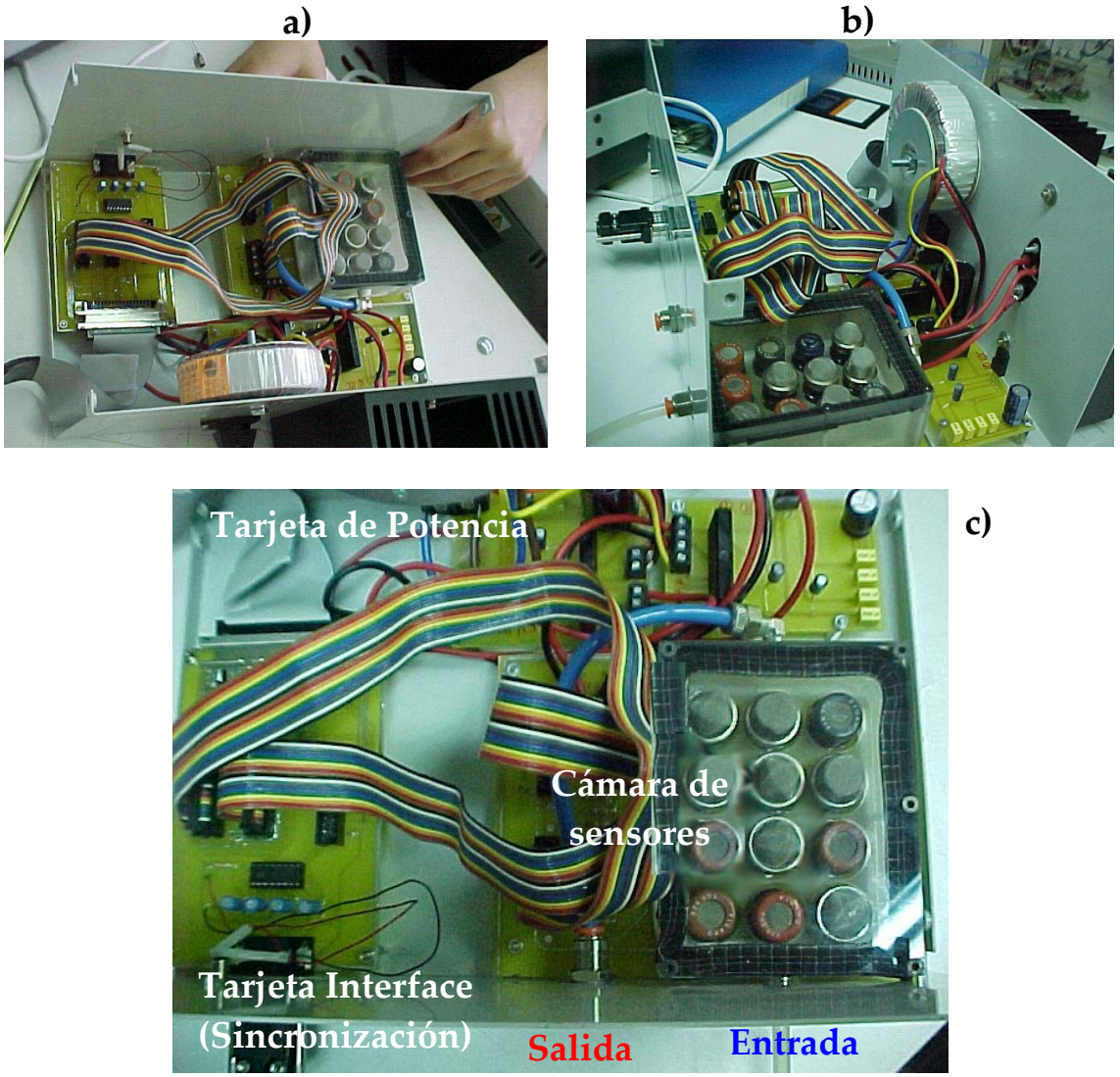


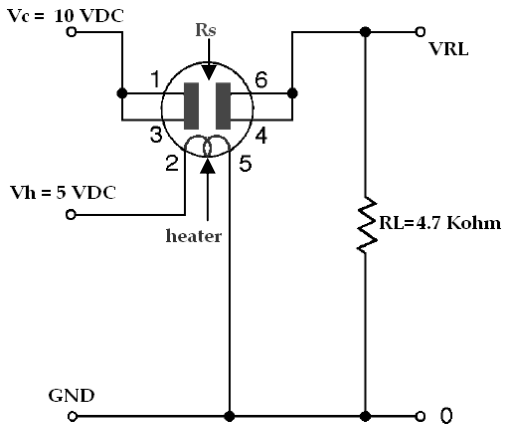

*Figura 2.8: Equipo de medición,*
*(a) Proceso de construcción del equipo, (b) Vista lateral, (c) Vista azimutal*

Tal y como se aprecia en la figura 2.8 (c), el sistema de medición está compuesto por una cámara de medida de metacrilato, con la capacidad de almacenar una matriz de 12 sensores de gases (modelos FIS SP y TGS de la serie 8). La tabla 2.2 describe el tipo de sensor y las aplicaciones de cada uno de ellos, mientras que en las fotos anteriores se puede observar la ubicación de cada uno de los dispositivos.

| Cantidad | Tipo | Aplicación |
|---|---|---|
| | **Taguchi (Serie-8)** | |
| **1** | TGS-800 | Aire contaminado |
| **1** | TGS-813 | Gas combustible |
| **1** | TGS-822 | Alcohol, Tolueno, *o*-xyleno |
| **1** | TGS-825 | Sulfuro de hidrogeno |
| **1** | TGS-826 | Amoníaco |
| **1** | TGS-831 | R-21-R-22 |
| **1** | TGS-832 | R-134a, R-22 |
| **1** | TGS-842 | Metano, butano, propano |
| **1** | TGS 880 | Especies volátiles de alimentos |
| **1** | TGS-882 | Vapores de Alcohol de alimentos |
| | **FIS (Serie-SP)** | |
| **1** | SP-31-00 | Disolventes orgánicos |
| **1** | SP-32-00 | Alcohol |

***Tabla 2.2: Descripción de la matriz de sensores***

En la placa de los sensores se incorporaron dos conectores de 20 pines cada uno, con la función principal de establecer la comunicación entre cada uno de los sensores y el PC a través de una interface acoplada a una tarjeta de adquisición de datos. Dos conectores de dos pines sirven para conectar la placa de potencia, módulo encargado de polarizar cada uno de los sensores para su correcto funcionamiento. Esto se realizó a través de una fuente de potencia de 5 á 10 voltios DC, que generaba los voltajes estándar exigidos.




$$\Rightarrow R_s = \left( \frac{V_c}{V_{RL}} - 1 \right) x\, R_L$$

***Figura 2.9: Circuito básico de medida y formula para calcular la resistencia del sensor (Rs)***

Las salidas de cada uno de los sensores fueron conectadas a sus respectivas resistencias de carga ($R_L$) de 4,7 kΩ, con el fin de obtener por medio de un circuito básico y el divisor de tensión, la respuesta de la resistencia del sensor (Rs), ver la figura 2.9. El voltaje Vc indica el valor de polarización permitido

para el dispositivo, y Vh el valor correspondiente al voltaje del elemento calefactor ó "*heater*" del sensor.

Para la etapa de sincronización entre el sistema de muestreo y adquisición de datos se diseño un sencillo circuito electrónico compuesto por el dispositivo MAX 232, cuya función principal es traducir la señal serie del HP a una señal digital TTL (Transistor-Transistor Logic) hacia el PC. El dispositivo electrónico fue incorporado a la tarjeta interface de comunicación de los sensores y a la tarjeta de adquisición de datos PCI-6023E de National Instruments.

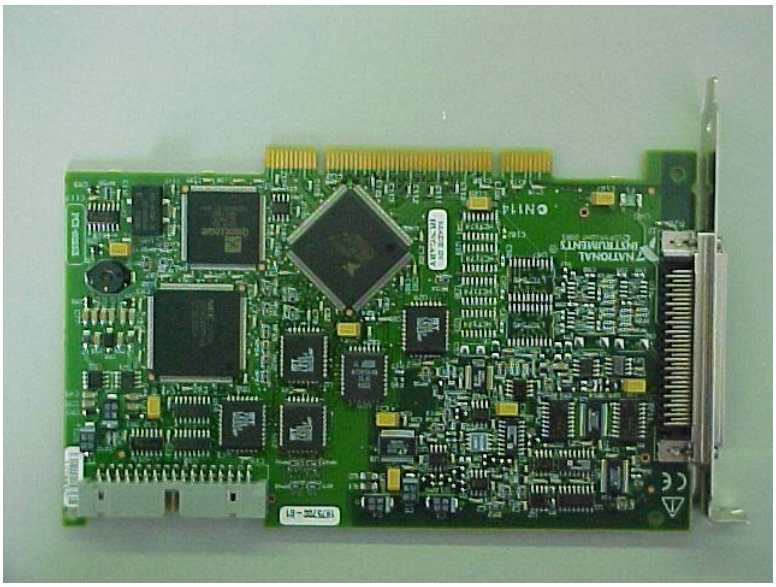


***Figura 2.10: Tarjeta de adquisición National Instruments***

La tarjeta, con 16 bits de resolución, tiene la capacidad de monitorizar hasta 16 sensores desde sus entradas analógicas. Tiene además 8 entradas digitales, algunas de las cuales fueron aprovechadas para la aplicación. Es importante mencionar que debido a estás características la tarjeta de adquisición de datos está catalogada entre las de más bajo coste por el fabricante (figura 2.10).

La figura 2.11 presenta el sistema de olfato electrónico completamente construido, con todos sus componentes principales ubicados en el interior de una caja metálica de acero inoxidable. Tal y como lo muestra las fotografías (a) (parte delantera), y (b) (parte trasera) del equipo de medición, el sistema es robusto y puede ser transportado fácilmente para efectuar pruebas de laboratorio y de campo.

En la parte inferior de la fotografía 2.8 (c), se pueden ver dos orificios que indican la entrada del flujo de gas hacia la cámara de sensores, y la salida del mismo hacia el exterior (ver figura 2.11 (a)).


**a)**


**b)**

***Figura 2.11: Fotografía del sistema de olfato electrónico***
***(a) Parte frontal, (b) Parte trasera***

## 2.3.3 Descripción del Software

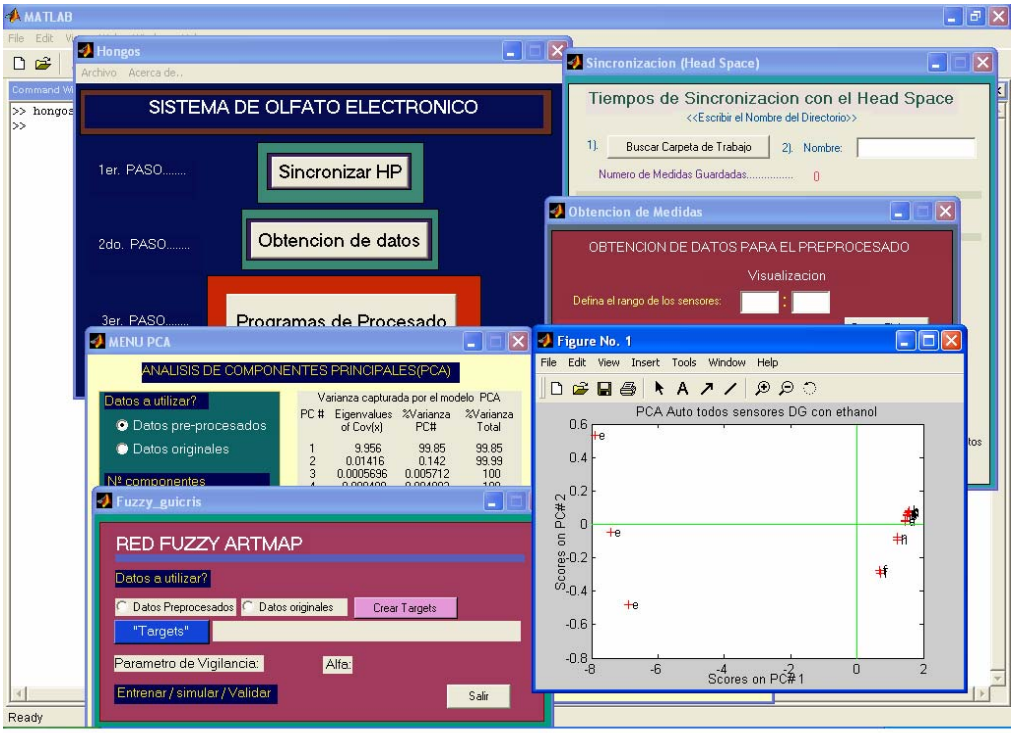



*Figura 2.12: Conjunto de GUI's necesarios para el funcionamiento del instrumento*

Un elemento fundamental para el control y monitorización del proceso de medida lo constituye el software de procesado y adquisición de datos. El software fue desarrollado en Matlab (vesión 6.1) a través de una interface gráfica para ser usado de forma amistosa por parte del usuario [**69**].

En la figura 2.12 se ve en forma de cascada un grupo de ventanas o GUI's. Estas ventanas realizan diferentes tareas, tales como la sincronización del HP y el PC, el almacenamiento de datos, y el análisis de los mismos a través de programas de procesado.

### 2.3.3.1 Sincronización entre el Headspace AutoSampler y el PC

El GUI de sincronización del HP de la figura 2.13 es la ventana donde se programan los valores para sincronizar el equipo de muestreo con el PC. Realiza la comunicación con el instrumento, ejecutando los mismos parámetros y tiempos programados por el equipo una y otra vez.

Como el programa realiza una monitorización continua del comportamiento de los sensores, se pueden representar las señales de los sensores en tiempo real y luego se almacenan cada una de ellas en un directorio creado previamente en el

PC. Esta ventana tiene en su rutina de adquisición un conjunto de opciones que determina el número de medidas que han sido adquiridas en cada momento, como también la programación de los ciclos de medida de los sensores, tales como el tiempo de resistencia inicial (baseline o resistencia base del sensor), tiempo de inyección (paso del contaminante), y tiempo de reposo para la estabilización de los sensores.

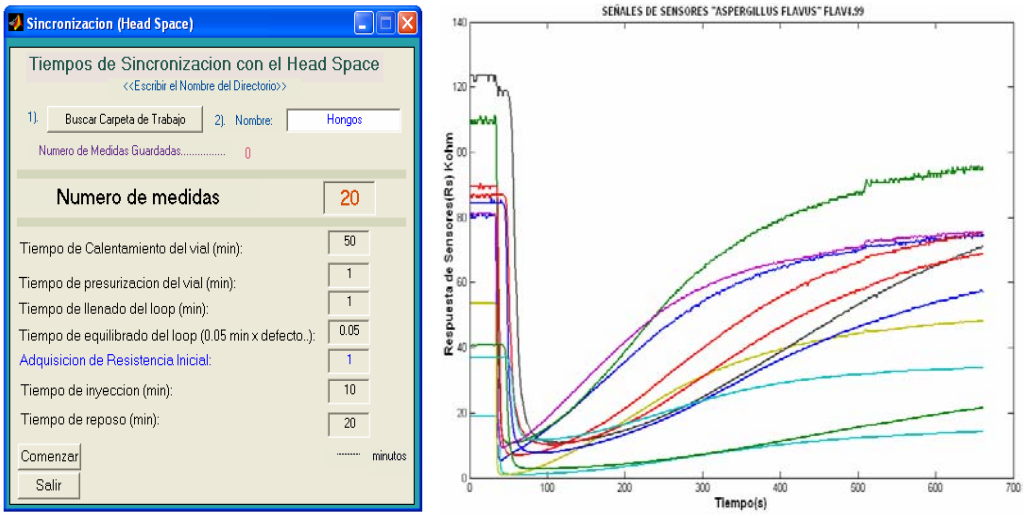



*Figura 2.13: GUI para la sincronización HP-PC y monitorización de las señales de los sensores*

### 2.3.3.2 Análisis y obtención de datos

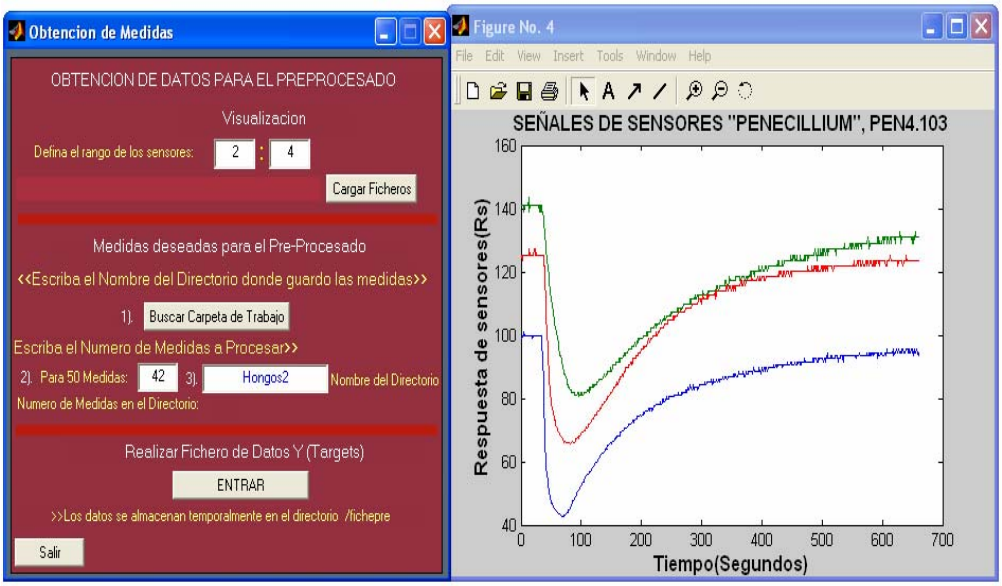



*Figura 2.14: GUI para el análisis de datos*

Esta importante etapa del programa tiene como objetivo visualizar y analizar las respuestas de cada una de las señales de los sensores adquiridas

previamente. En la figura 2.14 se muestra el comportamiento del sistema de medición con la respuesta de tres sensores.

Así, de está forma, se pueden realizar comparaciones de las respuestas de cada uno de los sensores y al mismo tiempo se puede ir chequeando el funcionamiento del equipo. Después de haber explorado las señales adquiridas, se aplican a cada una de ellas las diferentes funciones de extracción de parámetros, con el fin de obtener una matriz de datos formada por medidas (filas) y sensores (columnas).

### 2.3.3.3 Procesado de datos

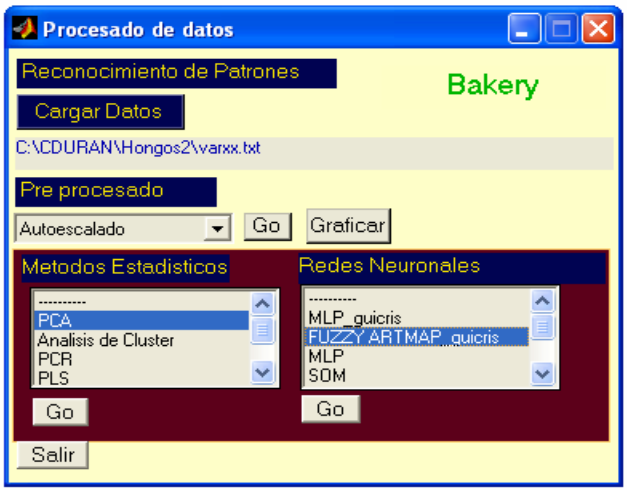



*Figura 2.15: GUI para el procesado de datos*

Para poder clasificar los compuestos volátiles emitidos por los hongos, se desarrollaron diferentes GUI's con los correspondientes algoritmos para el pre-procesado y procesado de datos (métodos estadísticos, redes neuronales, etc), tal y como se observa en la figura 2.15.

Para extraer la máxima información de las señales adquiridas es posible utilizar 4 métodos típicos de pre-procesado de datos: Auto-escalado, centrado, normalización por columna y normalización por matriz. Los métodos se pueden seleccionar desde un cuadro en donde el método de auto-escalado esta seleccionado por defecto. A continuación se describen muy brevemente las funciones de cada uno de ellos.

- **Auto-escalado** ("Autoscaling"): Con este método, se realizan cálculos individuales por columna (es decir, por sensor o parámetro del sensor). El método calcula el valor medio y la varianza con los valores obtenidos para las medidas para un sensor (columna) determinado. A continuación, se le resta el valor medio a todos los datos de dicha columna y el resultado se

escala por la desviación estándar. De esa forma, todas las columnas de datos (variables) presentan una media igual a cero y una varianza igual a la unidad.

El objetivo de esta normalización es la de dar igualdad de escala a cada una de las variables o parámetros que describen cada medida. Este tipo de escalado es muy útil cuando los parámetros que describen cada experiencia son de naturaleza (y por tanto de unidades) diferentes. Incluso en el caso de que cada variable represente el mismo parámetro para un sensor hay que recordar que diferentes sensores trabajan en diferentes valores de resistencia, por lo que es conveniente que todos ellos trabajen dentro del mismo rango de valores. El éxito de este escalado radica en que, a priori, asigna la misma importancia numérica a cada una de las variables que describen una medida, independientemente de su naturaleza o sensor del que provengan. De todas formas, hay que tener cuidado ya que este proceso puede incrementar notablemente el ruido existente en señales débiles.

- **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. Este tipo de centrado es fundamental para los métodos lineales aplicados en este estudio (PCA y DFA), ya que sin él no funcionan adecuadamente.

- **Normalización por columna:** En esta tercera opción de nuevo se opera por columnas. Se busca el máximo valor de cada columna y todos los valores de la misma se dividen por dicho valor. De esta forma se consigue que cada columna contenga valores entre "0" y "1".

- **Normalización por matriz:** Divide toda la matriz por el elemento máximo de la misma. En este caso solo un elemento valdrá la unidad. Se puede utilizar en especial para redes fuzzy ARTMAP.

Otra opción muy importante de este GUI de procesado es la selección del método de reconocimiento de patrones que se quiere escoger. Existe la posibilidad de escoger entre discriminación con PCA, análisis de clusters, PLS (Partial Least Squares), y análisis de DFA. Cada uno de ellos presenta las diferentes opciones en un entorno gráfico, con el objetivo de que el usuario introduzca los parámetros adecuados de forma muy sencilla. Al mismo tiempo es posible visualizar los resultados con cada uno de los diferentes métodos, y de está forma poder evaluarlos posteriormente. Un ejemplo de esto se puede observar a través de la respuesta de la figura 2.16, mediante la discriminación de las diferentes especies de hongos, ejecutando el método de análisis por componentes principales (PCA).

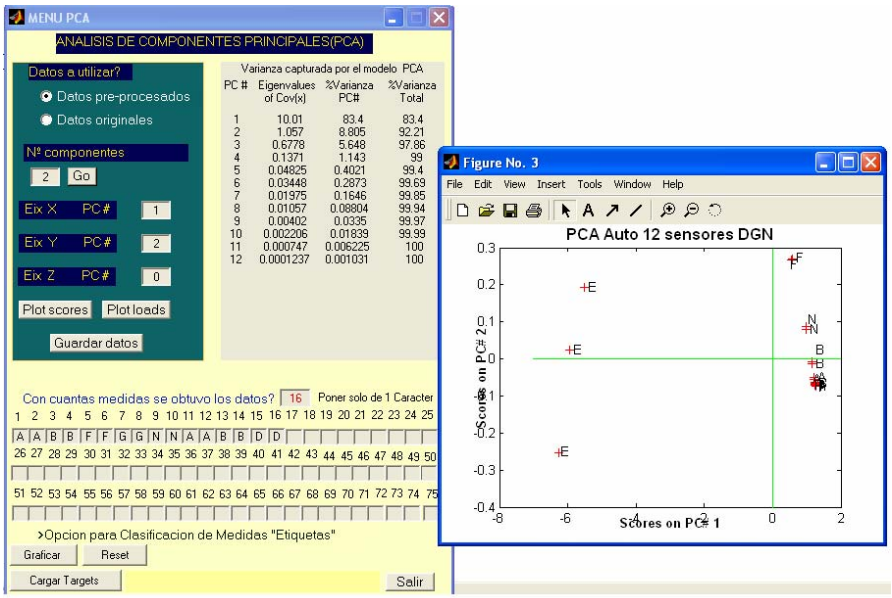



*Figura 2.16: GUI PCA para la discriminación de medidas*

Para poder identificar ó clasificar cada uno de los géneros ó especies de hongos se desarrollaron una serie de GUI's independientes con algoritmos de redes neuronales como método de reconocimiento de patrones. En concreto, las redes implementadas fueron las redes neuronales MLP (Multi-Layer Perceptron), SOM (Self-Organizing Map), LVQ (Learning Vector Quantization) y fuzzy ARTMAP. Cada uno de estos métodos esta provisto de los diálogos apropiados para introducir los parámetros y opciones adecuadas. Desde el mismo diálogo se ejecutan las etapas de entreno y evaluación.

En la figura 2.17 se aprecia la respuesta en la clasificación de un conjunto de medidas de hongos en forma de porcentaje de aciertos mediante un GUI con una red supervisada fuzzy ARTMAP previamente definida y entrenada.

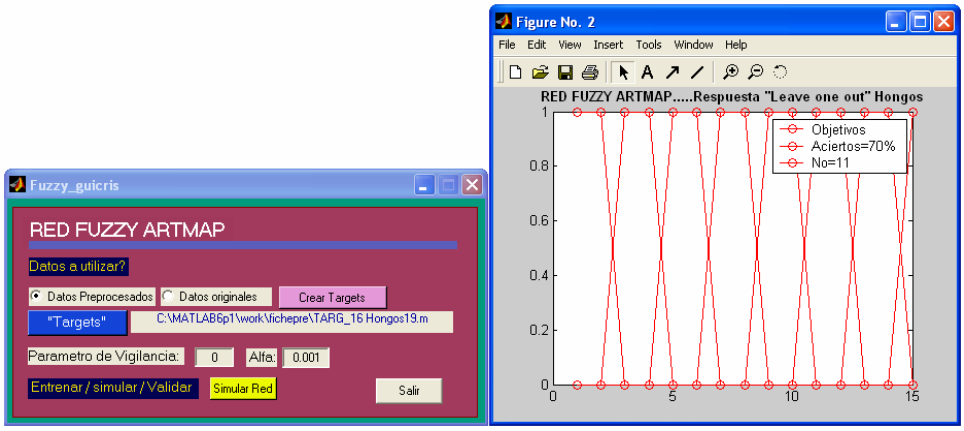



*Figura 2.17: GUI de la red Neuronal fuzzy ARTMAP y respuesta en la clasificación*

# 2.4 Pruebas y resultados

## 2.4.1 Medidas con etanol, amoníaco y acetona

Para evaluar el funcionamiento del SDOE frente a diferentes volátiles, se realizaron un conjunto de medidas preliminares utilizando acetona, etanol, y amoníaco. Para ello se prepararon 6 viales (dos por cada compuesto), y cada vial se midió tres veces. Los parámetros para acondicionar la muestra fueron introducidos al equipo de muestreo HP, tal y como se describen en la siguiente tabla:

| Parámetro | Valor |
|---|---|
| **Temperatura del horno** | 100 ºC |
| **Tiempo de calefacción** | 60 minutos |
| **Tiempo de presurización del vial** | 1 minuto |
| **Tiempo de llenado del loop** | 1.5 minutos |
| **Tiempo de equilibrio del loop** | 0.05 minutos |
| **Tiempo de inyección** | 10 minutos |
| **Flujo de gas portador** | 70 ml/minutos |

*Tabla 2.3: Parámetros de acondicionamiento de la muestra, medidas de prueba*

El tiempo de adquisición (correspondiente al tiempo de inyección) para cada medida fue de 10 minutos, dejando un tiempo de 20 minutos de reposo para la recuperación de los sensores.

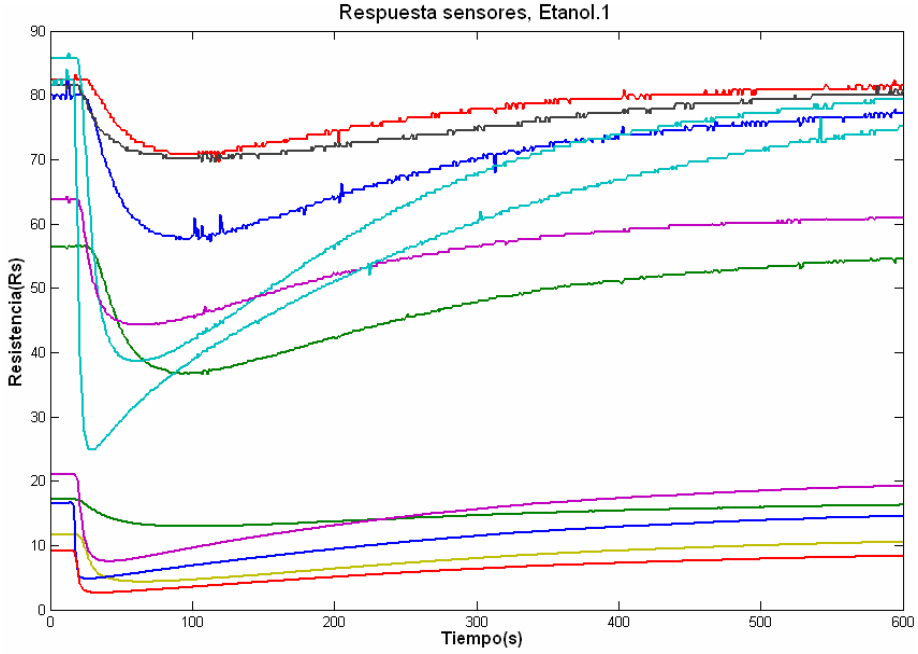



*Figura 2.18: Respuesta de los sensores al etanol*

En las figuras 2.18, 2.19 y 2.20, se muestran las respuestas de los 12 sensores frente a los diferentes compuestos volátiles. Se puede predecir, sin tener que realizar un procesado previo de las señales de los sensores, que el comportamiento debido al etanol es muy diferente con respecto al amoníaco y acetona. Sin embargo, aunque no se logre ver gran diferencia entre los dos últimos compuestos, se ve que las amplitudes (observando la escala del eje de ordenadas) del amoníaco son más elevadas que las de acetona.

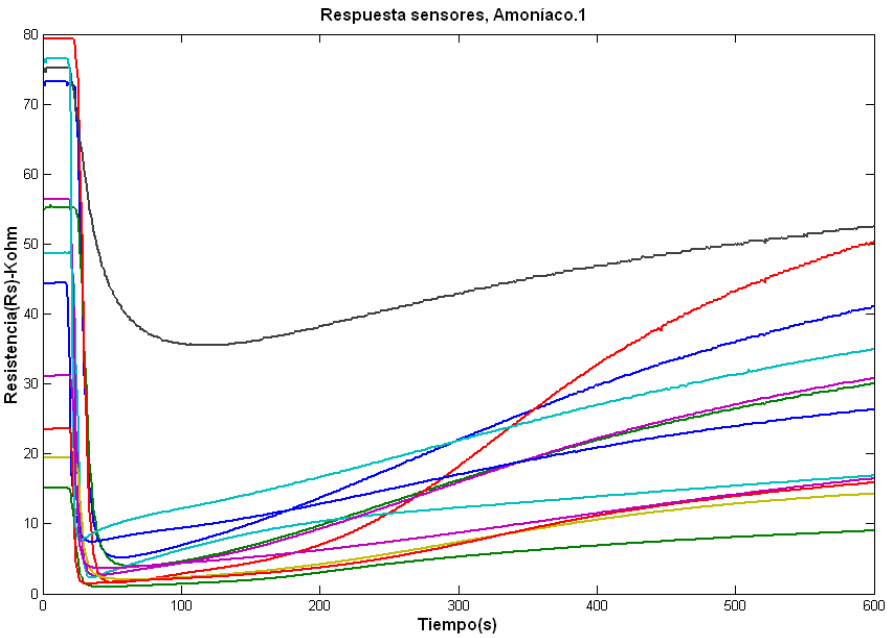



*Figura 2.19: Respuesta de los sensores al amoníaco*

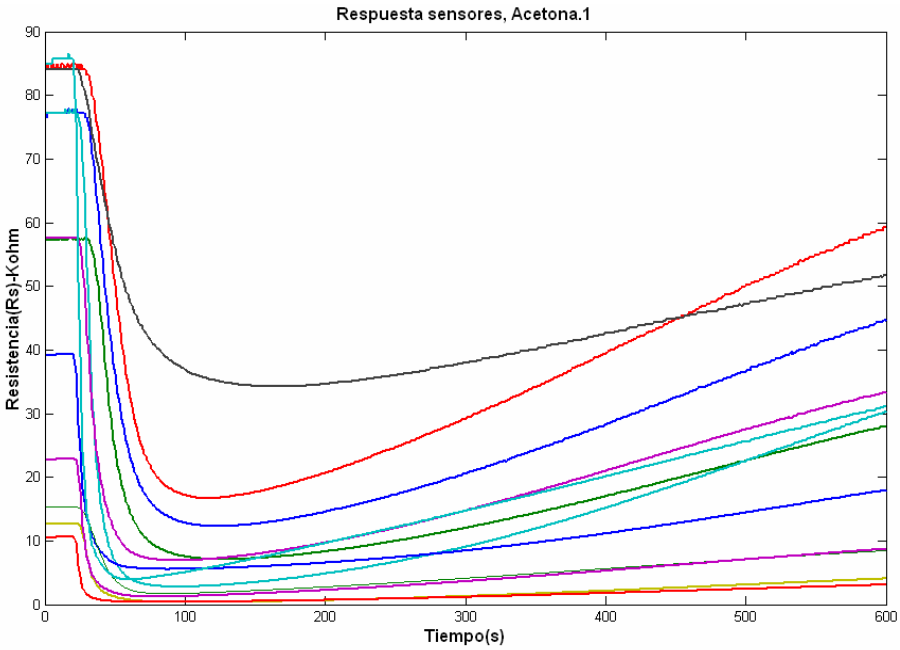



*Figura 2.20: Respuesta de los sensores a la acetona*

- **Análisis por componentes principales**

Para confirmar lo dicho anteriormente, se realizó un análisis por componentes principales.

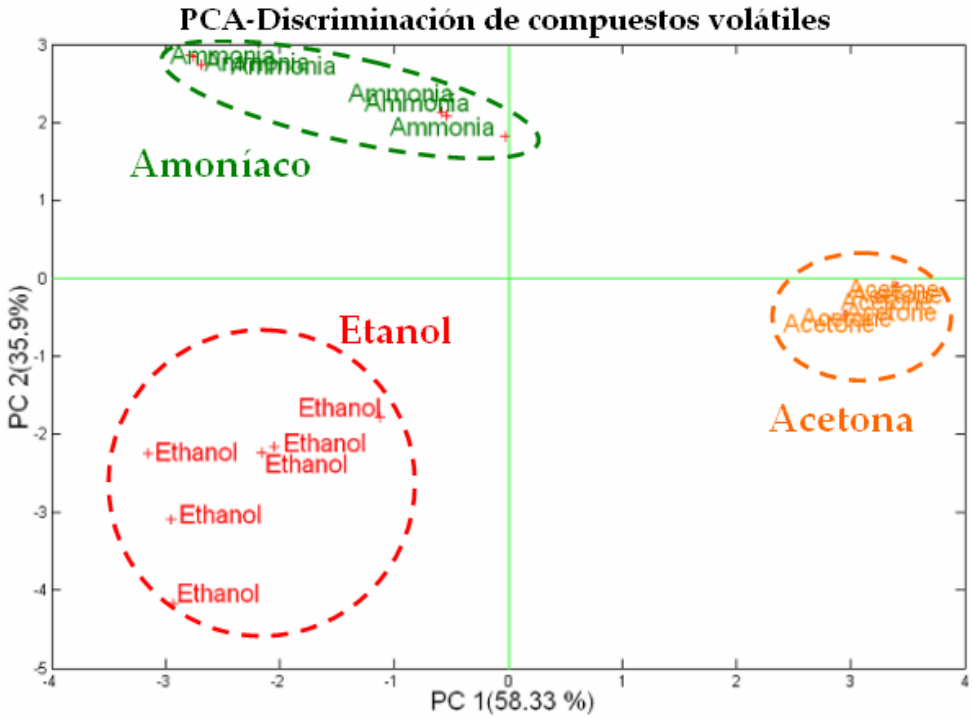



*Figura 2.21: Respuesta PCA con 18 medidas y 12 sensores*

En la figura anterior se presenta la discriminación de los tres diferentes compuestos mediante un análisis PCA. Por medio de un pre-procesado a la matriz de datos (un auto-escalado), la varianza total capturada fue del 95% para los dos primeros PC's. Como se puede observar, se consiguió una discriminación muy alta entre los tres compuestos, lo que demostró el funcionamiento correcto del sistema.

- **Procesado de datos con la red neuronal fuzzy ARTMAP**

Utilizando una validación cruzada de orden 1 (leave one out) junto a una red neuronal fuzzy ARTMAP, se consiguió alcanzar un promedio del 100% de aciertos en la clasificación de las medidas. A partir del resultado anterior, el siguiente paso fue evaluar el comportamiento del sistema de medida frente a las diferentes especies de hongos.

## 2.4.2 Medidas con hongos

- **Preparación de la muestra**

Después de diez días de incubación se consiguieron un total de 19 viales (20 ml de volumen), algunos de los cuales han sido retratados en la figura 2.22. 14 contenían 2 repeticiones de las 7 especies de hongos cultivadas, 2 viales contenían medios de cultivo sin contaminar y finalmente tres viales contenían etanol. Estos últimos servían para comprobar si las derivas de los sensores eran significativas.

En la tabla 2.4 se describen las diferentes muestras usadas (cada género en diferente color), y el número de repeticiones por cada especie.

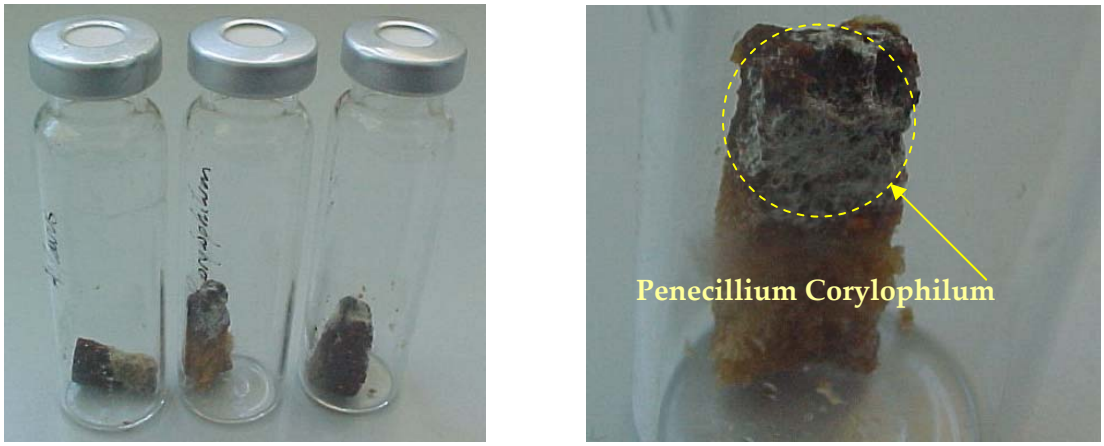



*Figura 2.22: Muestras con el producto (magdalenas) contaminado por diferentes especies de hongos*

| Género / Especie | *Repeticiones* |
|---|---|
| Eurotium Repens | 2 |
| Eurotium Herbariorum | 2 |
| Eurotium Amstelodami | 2 |
| Eurotium Rubrum | 2 |
| Aspergillus flavus | 2 |
| Aspergillus Niger | 2 |
| *Penicillium Corylophilum* | 2 |

*Tabla 2.4: Especies de hongos por género*

El tiempo de adquisición para cada muestra fue de 10 minutos, donde los parámetros introducidos conforme a las características son los detallados en la tabla 2.5.

| Parámetro | Valor |
|---|---|
| **Temperatura del horno** | 80 ºC |
| **Tiempo de calefacción** | 50 minutos |
| **Tiempo de presurización del vial** | 1 minuto |
| **Tiempo de llenado del loop** | 1.5 minutos |
| **Tiempo de equilibrio del loop** | 0.05 minutos |
| **Tiempo de inyección** | 10 minutos |
| **Flujo de gas portador** | 50 ml/minutos |

***Tabla 2.5: Parámetros de acondicionamiento de la muestra, (medidas con hongos)***

En las siguientes gráficas se pueden observar el comportamiento de los sensores ante el espacio de cabeza generado por cada uno de los cultivos de hongos y el medio de cultivo.

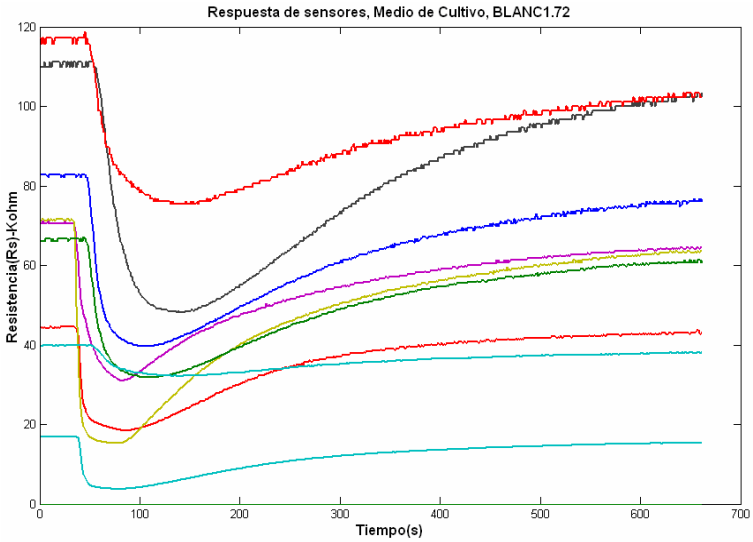



***Figura 2.23: Respuesta de sensores al medio de cultivo (Producto no contaminado)***

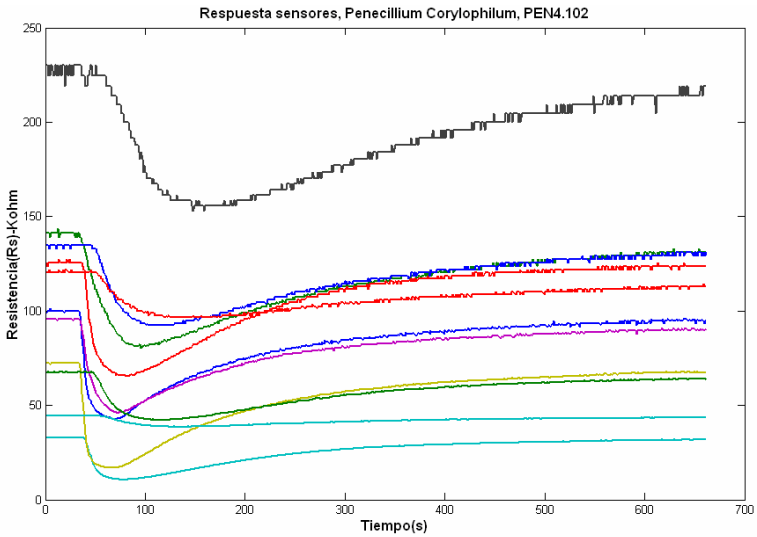



***Figura 2.24: Respuesta de sensores al género "Penecillium"***

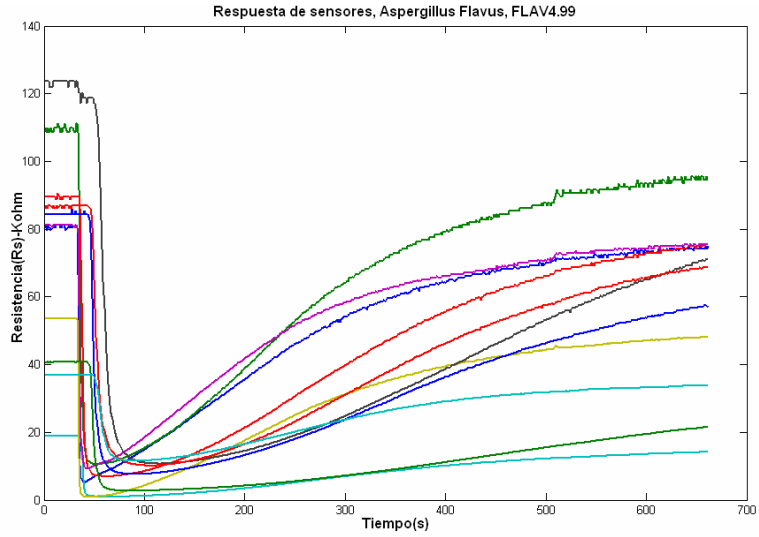



*Figura 2.25: Respuesta de sensores al género "Aspergillus"*

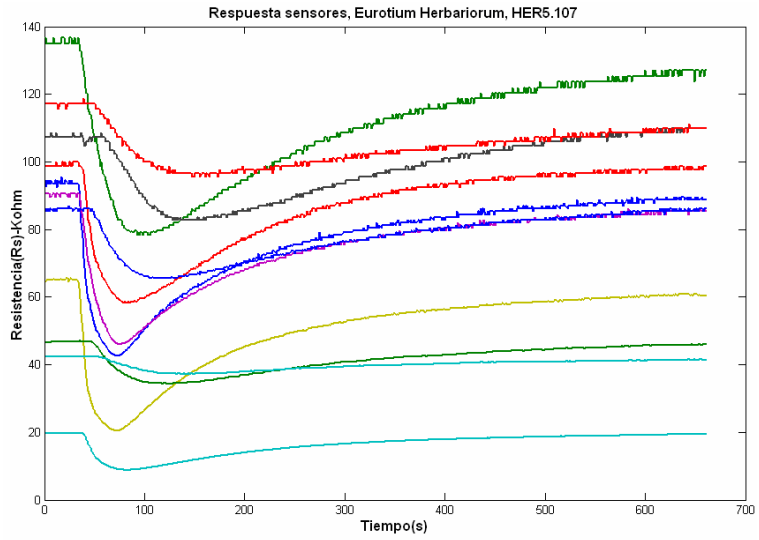



*Figura 2.26: Respuesta de sensores al género "Eurotium"*

Como se puede observar en cada una de las figuras anteriores, los sensores respondieron a cada uno de los géneros de hongos y fueron también muy sensibles al medio de cultivo. En la figura 2.25 se observa claramente que las señales debidas al género *Aspergillus* (especie *"Flavus")* fueron muy superiores respecto al resto de muestras. Al contrario sucede con los hongos *Eurotium, Penecillium* y el medio de cultivo, (figuras 2.23, 2.24 y 2.26), donde existen similitudes entre cada uno de ellos.

La variación del cambio de la resistencia del sensor (Rs) observada en la figura 2.27 es debido a la presencia de volátiles emitidos por los hongos, donde Ro es la lectura de la resistencia inicial correspondiente a la situación de referencia (aire del laboratorio). Las señales de los sensores fueron convertidas a valores de conductancia, lo cual permite estudiar con mayor claridad las respuestas frente a cada una de las especies de hongos, tal y como se puede

observar en la figura 2.28. El parámetro extraído desde cada uno de los 12 sensores, fue el incremento máximo de conductancia:

$$\Delta G_{maz} = G_{\max} - G_{\min} \tag{12}$$

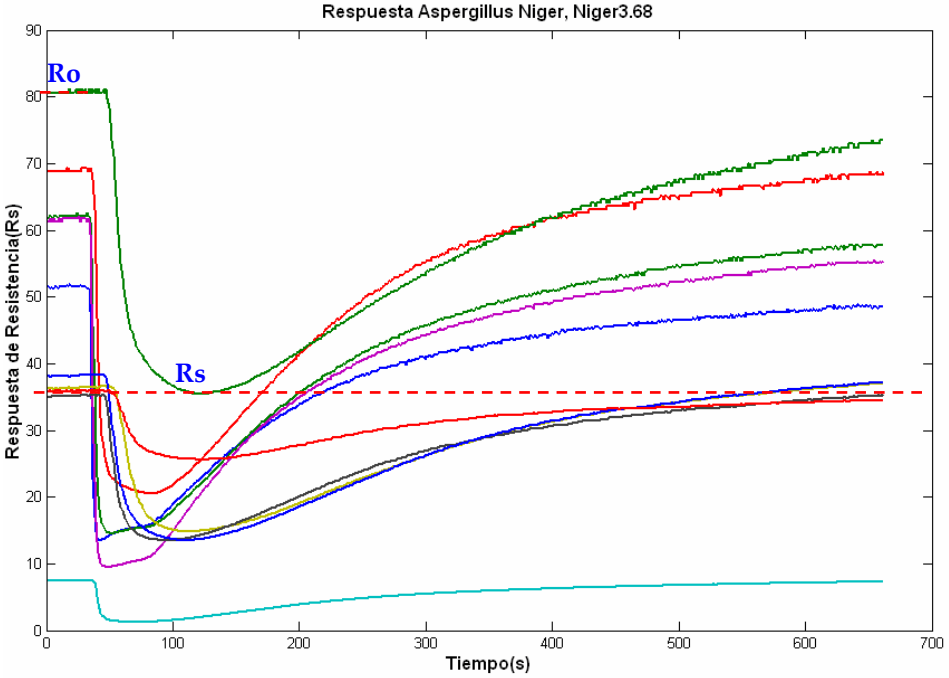



*Figura 2.27: Respuesta de los sensores de gases en valores de la resistencia*

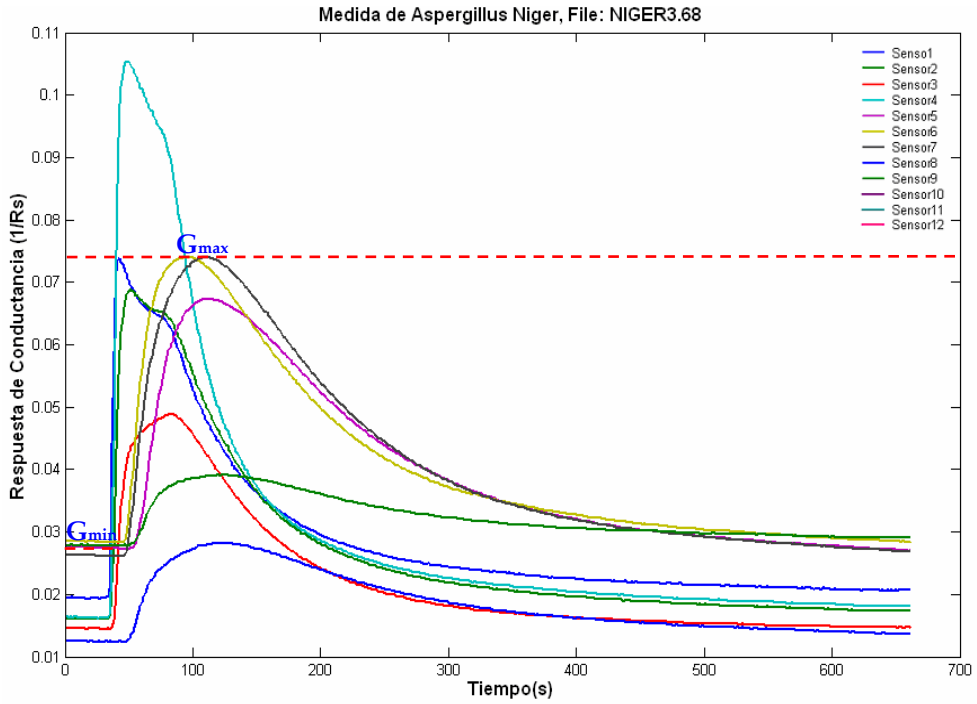



*Figura 2.28: Respuesta de los sensores de gases en valores de la conductancia*

Como se ha mencionado anteriormente, los datos adquiridos fueron procesados acoplando la red neuronal fuzzy ARTMAP junto a diversos algoritmos de selección de variables. En todos los casos se utilizó la técnica de validación cruzada de orden 1 ("leave one out"), para estimar el funcionamiento de la red en la clasificación de las especies de hongos (ver figura 2.29).

Este proceso iterativo de validación con N medidas, genera N formas de evaluación (1 para cada medida). En cada iteración, una medida es retirada del conjunto, mientras que las restantes se utilizan para construir el modelo (PCA, DFA, GA, etc), y para entrenar a la red. Posteriormente, la medida no usada para el entrenamiento, se proyecta sobre el modelo y se clasifica usando la red ya entrenada. Esto se repite N veces (una para cada medida), de modo que el resultado final es el promedio de todo el proceso iterativo.

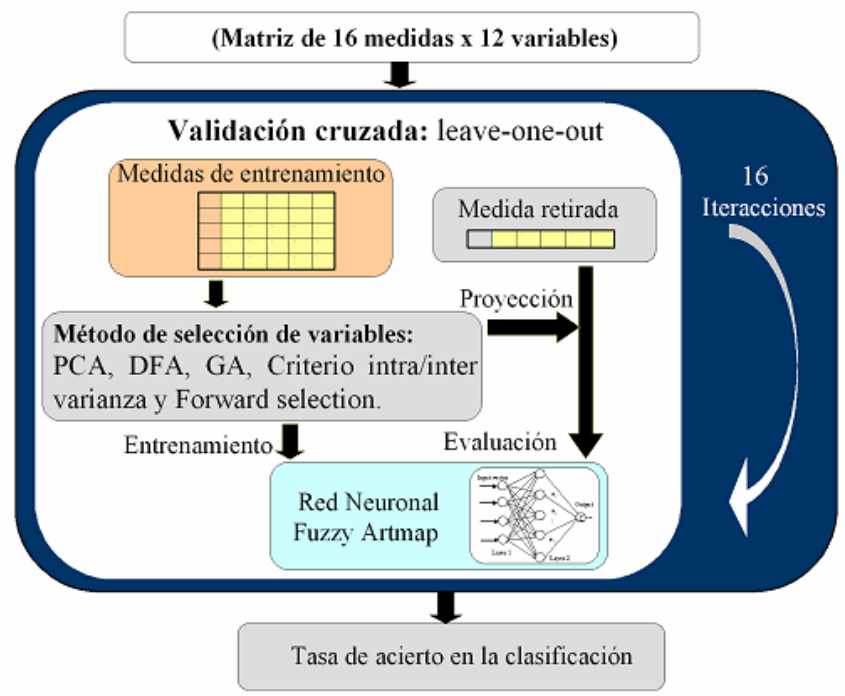



***Figura 2.29: Proceso iterativo para la clasificación de medidas***

Está aproximación es muy conveniente en casos donde el conjunto experimental no contienen muchas medidas. Otro punto importante de esta metodología es que de hecho se evalúa la aproximación utilizada y no una red concreta, ya que de hecho, se crean y se evalúan N redes usando el mismo procedimiento. Por otra parte, la medida dejada fuera para la evaluación no interviene para nada en el proceso de entrenamiento, así que no hay riesgo de conseguir resultados poco realistas debido a un sobre-entreno en los datos.

### 2.4.2.1 Resultados con fuzzy ARTMAP sin selección de variables

Primero, para comparar los resultados con las diferentes técnicas de selección de variables, se utilizó una red neuronal fuzzy ARTMAP para identificar las muestras utilizando todos los sensores (12 variables). La tasa de éxito en la clasificación con las ocho categorías (7 hongos mas un vial sin contaminar) alcanzó un 43% usando un leave-one-out. A modo de comparación, si la identificación se realizó aleatoriamente, se esperarían un 12.5% de aciertos. Con un análisis PCA se puede observar gráficamente la pobre capacidad de discriminación del sistema.

La mayoría de los hongos no pudieron ser distinguidos entre ellos, pero tal y como se observa el la figura 2.30, se observa una buena separación entre las especies *Aspergillus Níger*, *Aspergillus Flavus* y el resto.

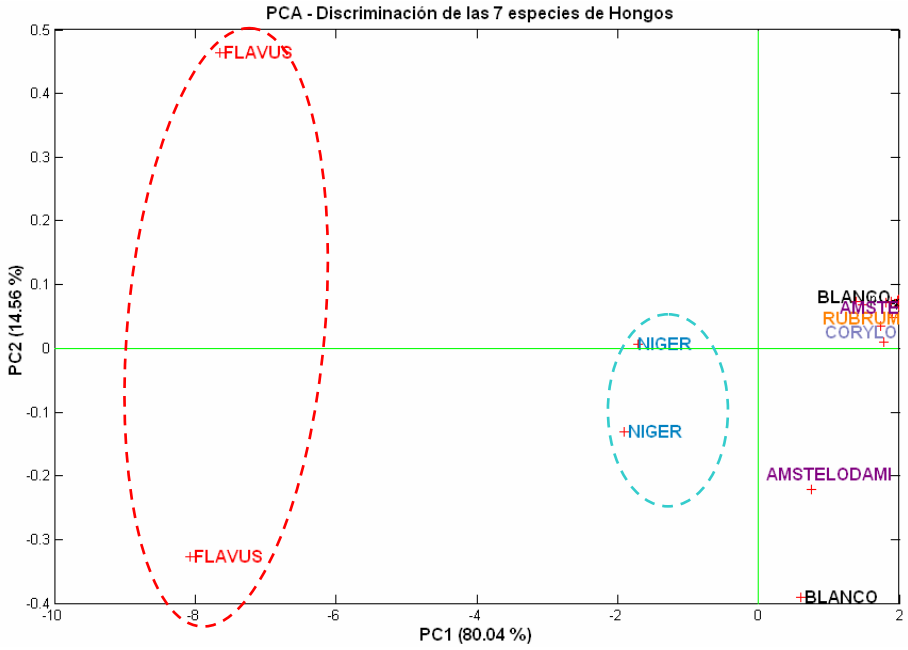



***Figura 2.30: Proyección PCA con 16 medidas (14 Hongos y 2 medios de cultivo)***

Una vez esta tasa de clasificación fue obtenida, la idea fue acoplar varias técnicas de selección de variables a la red fuzzy ARTMAP, como estrategia para ver si este método mejora los resultados obtenidos anteriormente.

### 2.4.2.2 Usando DFA como una técnica de selección de variables

Como decíamos anteriormente, DFA es un método que puede ser utilizado en modo supervisado y en modo no supervisado. En modo no supervisado da una información interesante acerca de las clases dentro del conjunto de datos.

De todas maneras, la forma más honesta de utilizarlo es mediante un modo supervisado, donde las medidas de entrenamiento deben ser diferentes del conjunto de datos de evaluación. Por lo tanto, la prueba se ha realizado junto a la técnica de la validación cruzada.

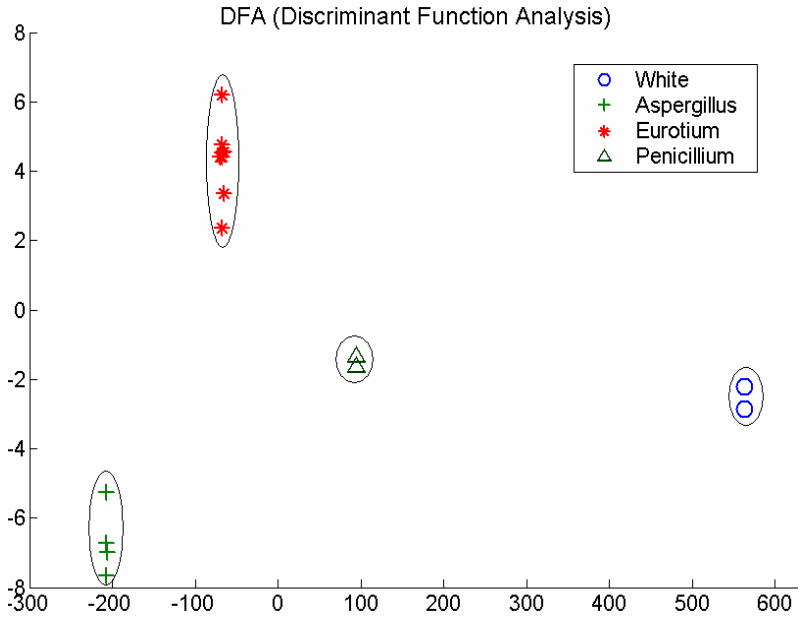



*Figura 2.31: Proyección DFA, discriminación de géneros de hongos*

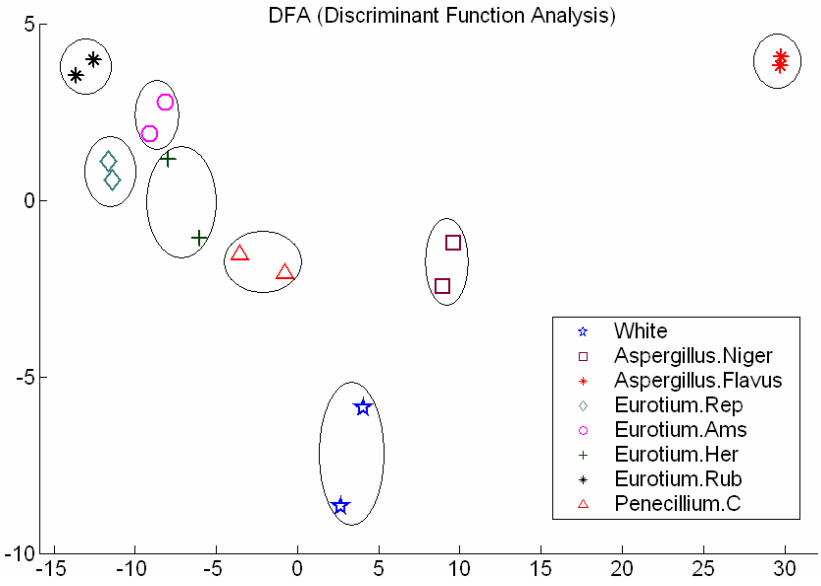



*Figura 2.32: Proyección DFA, discriminación de especies de hongos*

En cada iteración, se construyó un modelo de DFA con las medidas de entrenamiento. Las coordenadas de las muestras del entrenamiento en la proyección de la DFA fueron usadas para entrenar a la red fuzzy ARTMAP.

La medida de evaluación fue entonces proyectada sobre el modelo de DFA y sus coordenadas fueron suministradas a la red neuronal.

Se alcanzó una tasa de acierto del 75% utilizando solamente 2 eigenvectors (autovectores). Las gráficas DFA de las figuras 2.31 y 2.32 están en consonancia con los resultados obtenidos. Es importante recordar que cuando se usa leave-one-out, se anula el riesgo de sobreentreno en el conjunto de datos, puesto que la medida de evaluación no se ha utilizado ni para construir el modelo de DFA ni para entrenar la red.

### 2.4.2.3 PCA usado como método de selección acoplado a la red fuzzy ARTMAP

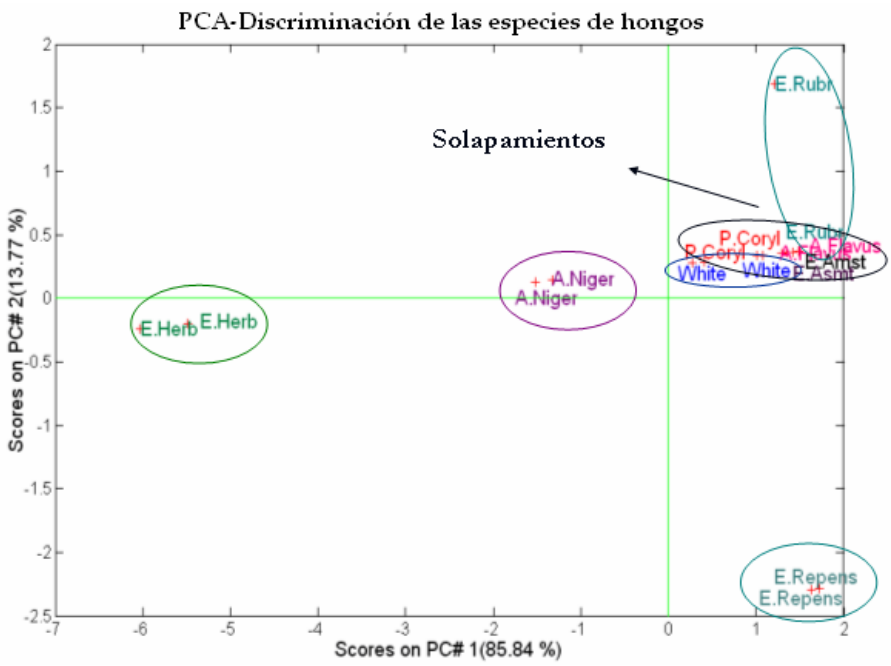



*Figura 2.33. Proyección PCA, discriminación de especies de hongos*

En la proyección PCA de la figura 2.33 se puede ver que usando un pre-procesado como el centrado de datos la mayoría de la varianza capturada (85.84%) está contenida en la primera componente principal. Por otra parte, puede deducirse que un modelo con dos PC's que captura más del 99% de la información nos indica que el número de variables necesarias para preservar toda la información significativa puede ser ampliamente reducida. Aunque hubo mejoría con respecto al PCA obtenido sin método de selección, en la figura aun pueden verse solapamientos entre los géneros *eurotium*, *penicillium* y el medio de cultivo.

En cada iteración se construyó un modelo de PCA con las medidas de entrenamiento, y los coeficientes ("scores") fueron suministrados a la red fuzzy

ARTMAP como conjunto de entrenamiento; entonces, con los PC's calculados y los pesos de la red neuronal, la medida de validación fue proyectada y evaluada. Se estudiaron los resultados obtenidos con un número de componentes principales variable y los mejores resultados fueron alcanzados con solo 2 PC's, donde el éxito de clasificación alcanzó un 63 %.

### 2.4.2.4 Resultados acoplando los algoritmos genéticos y fuzzy ARTMAP:

A través de un algoritmo genético acoplado a la red fuzzy ARTMAP fueron seleccionadas 5 de las 12 variables disponibles. La red fuzzy ARTMAP se usó para determinar el fitness que se aplico en la selección entre los cromosomas de cada generación. En la tabla 2.6 se describen cada uno de los diferentes parámetros aplicados al algoritmo genético de selección de variables.

| Parámetros del algoritmo genético | Valor |
|---|---|
| **Número de miembros de la población** | 32 |
| **Número máximo de generaciones** | 100 |
| **Probabilidad de mutación** | 0.005 |
| **Número de variables en una ventana** | 5 |
| **Promedio de criterio de convergencia** | 80 |
| **Número de términos iniciales** | 50 |
| **Tipo de combinación (crossover)** | 2 |
| **Número de subconjuntos en validación cruzada** | 5 |
| **Número máximo de iteraciones en la generación** | 1 |

*Tabla 2.6: Parámetros del algoritmo genético*

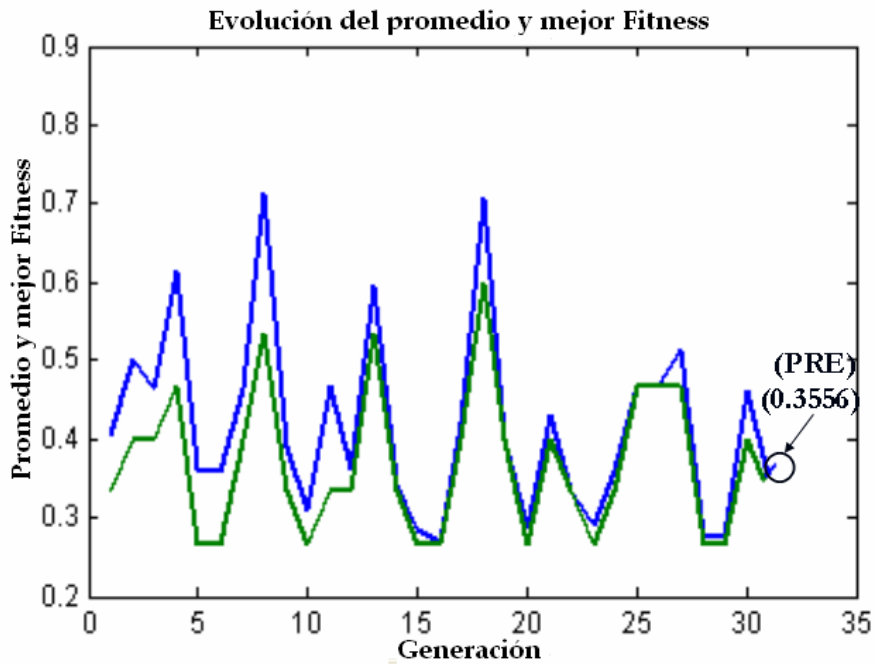



*Figura 2.34: Punto de convergencia del PER en la generación 33*

Lógicamente, el fitness escogido fué el PER (error de predicción) con el método de validación cruzada de orden uno (leave-one-out) aplicado al conjunto de 16 medidas. El PER obtenido disminuyó hasta un valor de 0.3556 y el algoritmo convergió después de 33 generaciones. La tasa de aciertos llegó a un 63 % de éxito. En la figura 2.34, está indicado el punto de convergencia del PER.

En la figura 2.35 se ve claramente que la selección del número de variables se obtiene con la evolución de las diferentes combinaciones de variables que se obtuvieron en el transcurso de las iteraciones ó generaciones. Los valores más altos de cada una de las columnas (color rojo) en el histograma, indica la variable seleccionada.

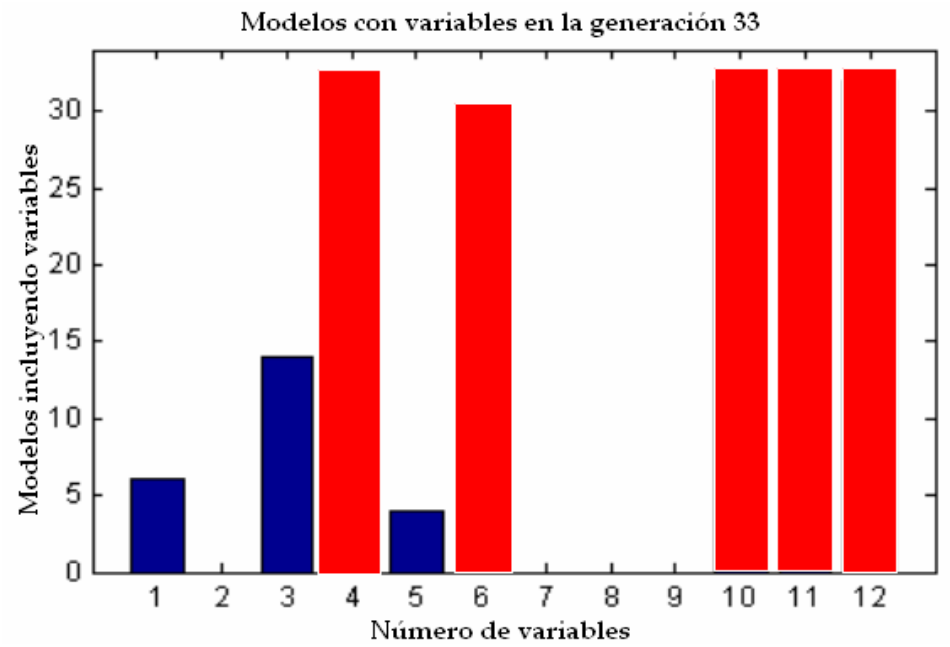



*Figura 2.35: Selección de 5 variables al finalizar el proceso iterativo*

### 2.4.2.5 Selección de variables a través del criterio de intra/intervarianza

Como se había mencionado anteriormente en el apartado 2.2, se utilizó un criterio de resolución para reducir el número de variables. Un valor más alto para Vr significa un mayor de resolución para una variable dada. La figura 2.36 muestra los valores de Vr para cada uno de los 12 sensores.

La fuzzy ARTMAP fue aplicada para evaluar el subconjunto de variables seleccionadas. Los mejores resultados fueron obtenidos con la selección de las 7 variables con mejor Vr (señaladas con circulos). La tasa de acierto alcanzó un máximo del 63 %.

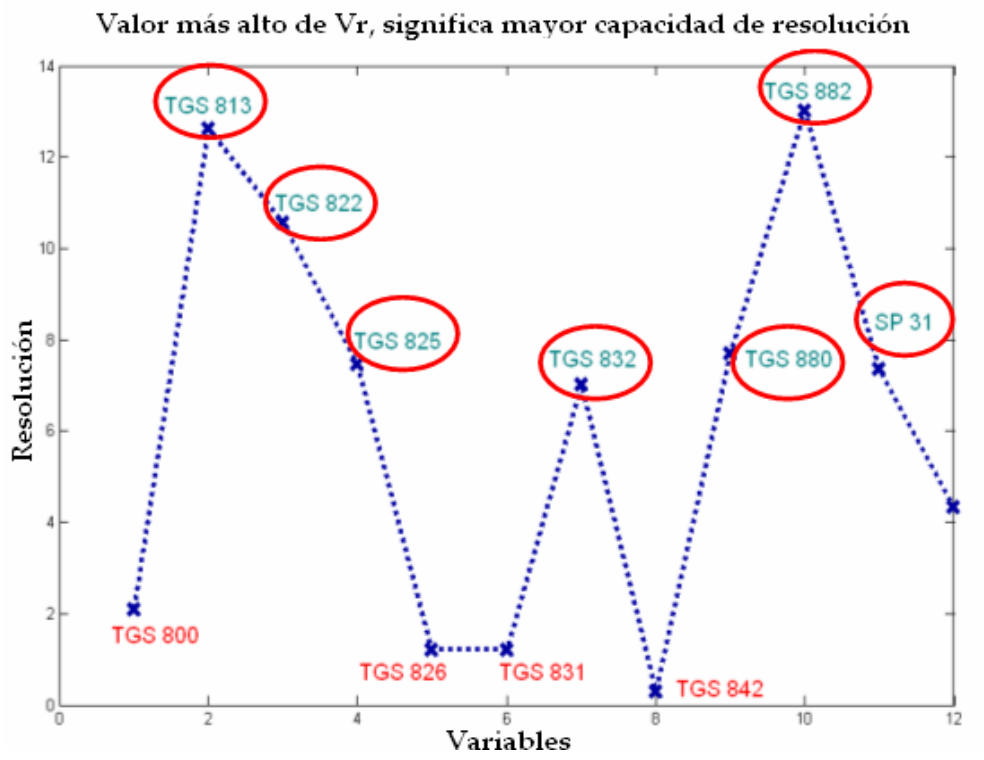



*Figura 2.36: Selección de 7 variables mediante el criterio Vr*

### 2.4.2.6 Forward selection

Este algoritmo de selección, usado generalmente en regresión lineal, fue aplicado al problema de la detección de hongos con el fin de conseguir una identificación más fiable con un subconjunto de las 12 variables originales.

La aproximación comenzó escogiendo una sola variable y evaluando el resultado mediante una validación cruzada. La variable con mejor tasa de éxito fue la seleccionada en esta primera iteración. Para la segunda iteración se evaluaron los resultados utilizando la variable escogida en la primera ronda en una combinación binaria con cada una de las restantes. Entonces, nuevamente, se selecciona la variable que, combinada con la primera, ofrece mejores resultados (ver la figura 2.37). El proceso termina cuando el error de de predicción (PER) se incrementa incorporando cualquiera de las restantes variables.

En el problema que nos ocupa sólo 2 variables fueron seleccionadas, alcanzándose una tasa de éxito del 70 %.

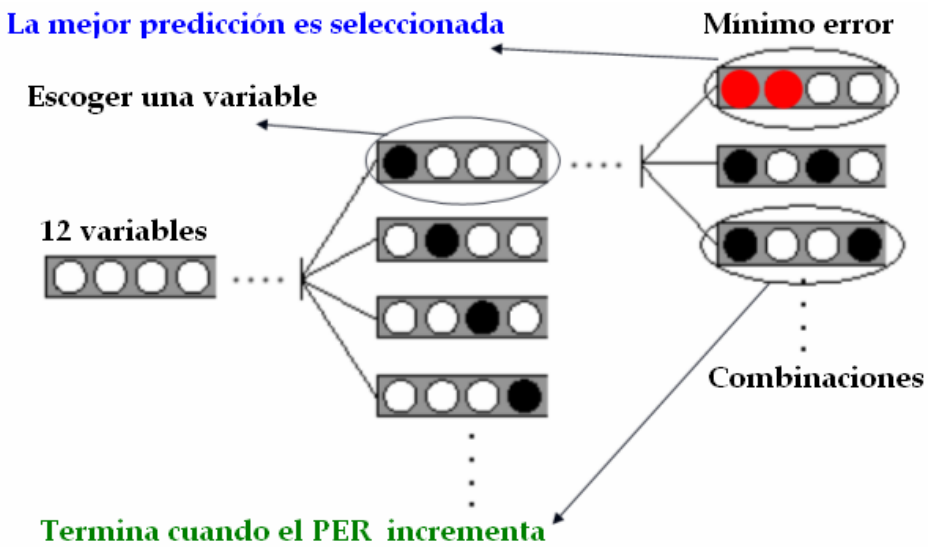



*Figura 2.37: Esquema gráfico del método "Forward Selection"*

## 2.4.3 Conclusiones

La tabla 2.7 resume los resultados obtenidos, realizando una comparación entre todas las técnicas de selección de variables que se han acoplado a la red neuronal fuzzy ARTMAP. Podemos observar que aplicando cualquiera de los métodos de selección de variables se mejoraron los resultados obtenidos originalmente con la red.

| Métodos | Resultados | Subconjunto seleccionado |
|---|---|---|
| Fuzzy ARTMAP sola | 43% | 12 |
| **DFA+ Fuzzy ARTMAP** | **75%** | **7** |
| PCA+ Fuzzy ARTMAP | 63% | 7 |
| GA+ Fuzzy ARTMAP | 63% | 5 |
| Criterio Vr | 63% | 7 |
| Forward + Fuzzy ARTMAP | 70% | 2 |

*Tabla 2.7: Resultados y variables seleccionadas con cada método*

De los datos presentados en la tabla queda claro que los mejores resultados se alcanzaron acoplando la selección de variables mediante un DFA a la red neuronal, alcanzando una tasa de éxito del 75% al clasificar muestras de 8 categorías (siete especies de hongos y un vial de control sin micro-organismos).

Por otro lado, es importante resaltar el resultado del método "forward selection". Por un lado, con un 70% de acierto, se sitúa muy cerca en cuanto resultados frente al DFA, pero con la ventaja que la selección de variables se corresponde directamente con la selección de sensores de la matriz, puesto que

las variables seleccionadas son directamente algunas de las originales sin sufrir ninguna transformación ni combinación lineal.

Con este método es posible dar una interpretación directa a la selección de variables, ya que permite reducir la dimensionalidad de la matriz de sensores químicos directamente al descartar un gran número de los sensores utilizados originalmente.

## 2.4.4 Comparación del prototipo diseñado con otros sistemas comerciales de olfato electrónico

El prototipo desarrollado en este trabajo ha sido utilizado en el proyecto europeo "*Rapid detection of microbial contaminants in foods products using electronic nose technology*", concedido bajo la financiación del V programa marco de investigación de la unión europea (UE).

Concretamente, se han realizado pruebas con nuestro prototipo frente a otros equipos comerciales, cuyos resultados han sido incluidos en una tesis doctoral de un doctorando europeo [**70**].

La tesis de este estudiante consistía en fusionar los datos extraídos de 4 sistemas de olfato electrónico (tres comerciales más el nuestro) para evaluar la capacidad de detección precoz de tres micro-organismos cultivados "in vitro". En total habían 4 tipos diferentes de viales, los que contenían *Pichia anomala* (Muestra A), los que contenían *bacillus subtilis* (B), los infectados con *penicillium verrucosum* (Muestra C), y un vial en blanco (D).

De los cuatro SDOE utilizados en este estudio realizado tres eran comerciales (Alpha M.O.S aFOX3000", NST instruments, y Bloodhound).

En la tabla 2.8 se describen los resultados comparativos correspondientes al grado de discriminación de los tres tipos de hongos (en diferentes tiempos de incubación) obtenidos a partir de los datos obtenidos con cada uno de los instrumentos de medida.

| Tiempo de incubación | BH 114 | NST | AlphaMOS, aFOX 3000 | Prototipo |
|---|---|---|---|---|
| 24h | Pobre | Buena | Buena | Buena |
| 48h | Pobre | Muy buena | Discriminación total | Muy buena |
| 72h | Pobre | Buena | Discriminación total | Buena |

***Tabla 2.8: Calificación de cada SDOE en la discriminación de los géneros de hongos***

Estos resultados fueron determinados partiendo de las respuestas de los mejores sensores escogidos en cada uno de los SDOE con el objetivo de diseñar un instrumento "virtual" mediante la correlación de los mejores sensores, y a su vez aplicando algoritmos de procesado al conjunto de datos multivariantes. De está forma los resultados fueron mejorados al fusionar los diferentes instrumentos de medida.

Como se puede observar en las figuras 2.38 y 2.39, se obtuvo una correlación positiva entre el prototipo de la URV y los instrumentos comerciales aFOX 3000 y NST.

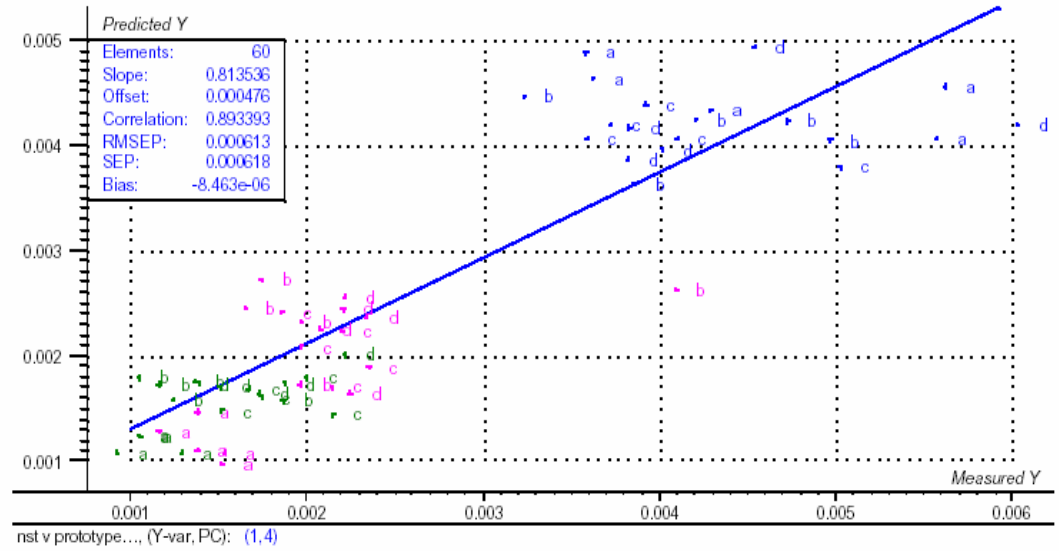



*Figura 2.38: Correlación del aFOX 3000 y el prototipo de la URV*

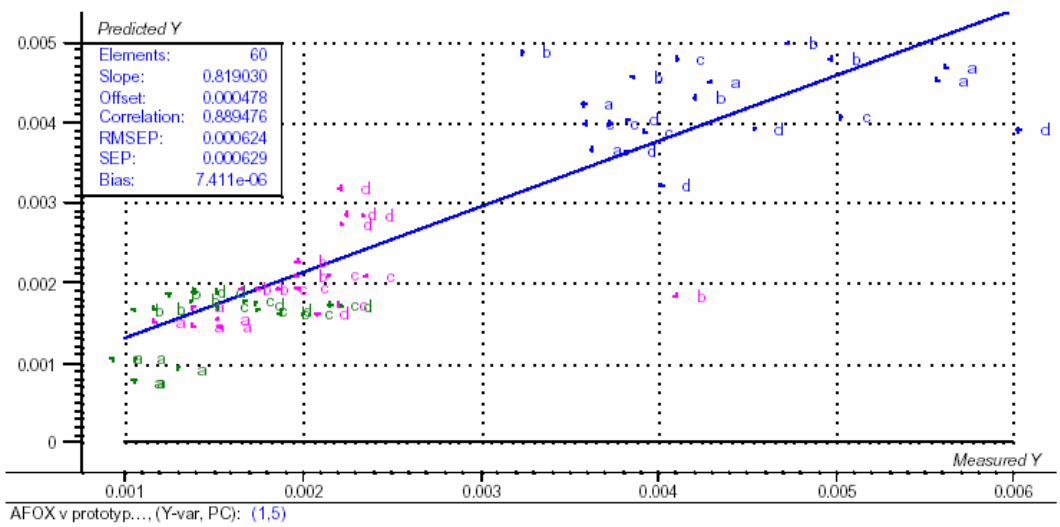



*Figura 2.39: Correlación del aFOX 3000 y el prototipo de la URV*

## 2.5 Referencias bibliográficas

[1] **Magan. N**, "*Early detection of fangal growth in stored grain*", International Biodeterioriation and Biodegradation, Vol: 32, pág 145-160, (1993).

[2] **Jain. P.C**, Lacey. J, Stevens. L, "*Use of API-Zym strips and 4-nitrophenyl sustrates to detect and quantify hydrolytic enzymes in media and grain colonised by Aspergillus, Eurotium and penecilliun sp"*, Mycological Research, Vol: 95, pág: 834-842, (1991).

[3] **Marin. S**, Sanchis. V, Magan. N, "*Effect of water activity on hydrolytic enzyme production by fusarium moniliforme and F. proliferatum during early stages of growth on maize, International Journal of Food Microbiology"*, Vol: 42, pág:185-191, (1998).

[4] **Lacey. J**, Hamer. A, Magan. N, "*Respiration and losses in stored wheat under different environmental conditions, En stored product protection"*, pág 1007-1013, (1994).

[5] **Kaminsky. E**, Przybylski. R, Wasowisc. E, *Spectrophotometric determination of volatile carbonyl compounds as a rapid method for detecting grain spoilage during storage*, Journal of Cereal Science, Vol: 3, pág: 165-172, (1985).

[6] **Borjesson. T**, Stollman. U, Schnurer. J, "*Volatile metabolites and other indicators of penecillium aurantiogriseum growth on different substrates"*. Aplied and Environmental Microbiology, Vol: 56, pág 3705-3710, (1990).

[7] **Johnson. A**, Winquist. F. Schnurer,. J, Sundgren H, Lundstrom. I, "*Electronic Nose for microbial quality classification of grains"*, International Journal of Food Microbiology, Vol: 35, pág 187-193, (1997).

[8] **Keshri. G**, Magan. N, Voysey. P, "*Use of an electronic nose for the early detection and differentiation between spoilage fungi*, Letters in applied Microbiology", Vol: 27, pág 261-264, (1998).

[9] **Schnurer. J**, Olsson. J, Borjesson, "*Fungal volatiles as indicators of food and feeds spoilage, Fungal Genetics and biology"*, Vol: 27, pág 209-217, (1999).

[10] **Olsson. J**, Borjesson. T, Lunstedt. T, Schnurer. J, "*Volatiules for mycological quality grading of barley grains: Determinations using gas chromatography-mass spectrometry and electronicnose"*, International Journal of Food Microbiology, Vol: 59, pág: 167-178, (2000).

[11] **Legan, J.D**. "*Moulds spoilage of bread: the problem and some solutions"*. International Biodeterioriation Biodegradation", 32: 33-53, (1993).

[12] **Smith, J.P**, "*Modified atmosphere packaging for bakery products"*. Technical Bulletin, American Instute of Baking Research Department, Vol: 16, pág 1-9, (1994).

[13] **Roessler. P.F**, Ballenger. M.C, "*Contamination of an unpreserved semisoft baked cookie with a xerophilic Aspergillus species"*. *Journal of Food Protection*, 59, pág: 1055-1060, (1996).

[14] **Fustier, P**, Lafond, A, Champagne, C.P, Lamerche, F, *"Effect of inoculation techniques and relative humidity on the growth of molds on the surfaces of yellow layer cakes"*. Applied and Environmental Microbiology, Vol: 64, pág 192-196, (1998).

[15] **Hopko L,** "*Food hygienic aspects of the confectionary industry"*. Edesipar, Vol: 30: pág 8-13, (1979).

[16] **Spicher. G**, "*Die faktoren des wachstums der schimmelpilze als ansatzpunkte fuer massnahmen zur unterbindung dr schimmelbildung bei backwaren"*. *Getreide, Mehl und Brot*, Vol: 34: pág 128-137, (1980).

[17] **Abellana. M**, Torres, L, Sanchis, V, Ramos, A.J, "*Caracterización de diferentes productos de bollería industrial II".* Estudio de la micoflora Alimentaria, Noviembre, pág: 51-56, (1997).

[18] **Birnbaum. H**, "*Water activity, microbial growth and antimicrobial agents"*. Bakers' digest, 55, pág: 18-21, (1981).

[19] **Hocking. A.D**, "*Xerophilic fungi in intermediate and low moisture foods. A Handbook of Applied Mycology Foods and Feeds IIIQ"* (ed. K. Arora, K.G. Mukerji y E.H. Martin), pág 69-98. New York: Marcel Dekker, (1991).

[20] **Magan, N**, "*Volatiles as an Indicator of fungal activity and differentiation between species and the potential use of electronic nose technology between for early detection of grain spoilage"*, Journal Stored Products Research, 36, pág. 319-340, (2000).

[21] **Needham, R**. Magan, N.; "*Detection and differentiation of microbial spoilage organisms of bakery products in vitro and in situ, 9th ISOEN"*, Technical Digest, pág: 240-241, (2002).

[22] **Needham, R**,"*Early detection and differentiation of spoilage of bakery products*", Sensors and Actuators B, vol 106, pág: 1. (2005).

[23] **Magan, N.;** Evans, P.*"Volatiles as an indicator of fungal activity and differentiation between species, and the potential use of electronic nose technology for early detection of grain spoilage"*. Journal of Stored Products Research, 36, pág: 319-340, (2000).

[24] **Olsson, J**, Börjesson, T; Lundstedt, T; Schnürer, J, *"Volatiles for mycological quality grading of barley grains: determinations using gas chromatography–mass spectrometry and electronic nose"*. International Journal of Food Microbiology, 59, pág:167–178, (2000).

[25] **Olsson, J**; Börjesson, T; Lundstedt, T; Schnürer, J.; *Detection and quantification of ochratoxin A and deoxynivalenol in barley grains by G-MS and electronic nose*. International Journal of Food Microbiology, 72, pág: 203-214, (2002).

[26] **Börjesson, T**, Eklöv. T, Jonsson. A, Sundgren. H, Schnürer, J. *"Electronic Nose for Odor Classification of Grains"*, Cereal Chem, 73(4): pág: 457-461, (1996).

[27] **Trihaas, J**, Tatjana van den Tempel, Per Vaeggemose Nielsen, "*Electronic nose: Smelling the microbiological quality of cheese"*, Proceddings on ISOEN, Rome, pag: 380-384, (2002).

[28] **Shin, H**. W.; Llobet, E; Gardner, E; Hines, E.L; Dow, C.S; "*Classification of the strain and growth phase of cyanobacteria in potable water using an electronic nose system"*. IEEE Proc-Sci. Meas. Technol., 147(4), pág:158-164, (2000).

[29] **Keshri, G**; Voysey, P; Magan, N. *"Early detection of spoilage moulds in bread using volatile production patterns and quantitative enzyme assays"*. Journal of Applied Microbiology, 92(1), pág: 165-172, (2002).

[30] **Vernat-Rossi**, V; Garcia, C; Talon, R; Denoyer, C; Berdagué, J.L; *"Rapid discrimination of meat products and bacterial strains using semiconductor gas sensors"*. Sensors and Actuators B, 37, pág: 43-48, (1996).

[31] **Magan. N**; Pavlou. A; Chrysanthakis, I; "*Milk-sense: a volatile sensing system recognise spoilage bacteria and yeast in milk"*. Sensors and Actuators B, 72, 28-34, (2001).

[32] **Pearce. T.C**, Schiffman. S.S, Nagle. H.T, Gardner.J.W, *"Handbook of machine olfaction", electronic nose technology*, Wiley, pág: 325-345, (2003).

[33] **Cadima. J**,*"Computational aspects of algorithms for variables selection in the context of principal components"*, Computational Statics & Data Analysis, 47, pág 225-236, (2004).

[34] **Klecka**, William R, *"Discriminant Analysis. Quantitative Applications in the Social Sciences Series"*, No. 19. Thousand Oaks, CA: Sage Publications, (1980).

[35] **R. Leardi**, *"Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection"*, J. Chemom. 8, pág: 65–79, (1994).

[36] **Gardner. J.W**, P. Boilot, E.L. Hines, "*Enhancing electronic nose performance by sensor selection using a new integer-based genetic algorithm approach", Sensors and Actuators B*, 106, pág: 114-121, (2005).

[37] **Broadhurst D**, R. Goodacre, A. Jones, *"Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry"*, Anal. Chim. Acta, 348, pág: 71-86, (1997).

[38] **Brezmes. J**, *"Discrimination between different samples of olive oil using variable selection techniques and modified fuzzy ARTMAP neuronal networks"*, 9th ISOEN, Rome, (2002).

[39] **Miller. A.J**, *"Subset selection in regression"*, Chapman & Hall, London, (1990).

[40] **Paulsson N**, E. Larson, F. Winquist, *"Extraction and selection of parameters for evaluation of breath alcohol measurement with an electronic nose"*, Sensors and Actuators A, 84, pág: 187-197, (2000).

[41] **Lu Xu**, Wen-Jun Zhang, *"Comparison of different methods for variable selection"*, Anal. Chim. Acta, 446, pág: 477-483, (2001).

[42] **Carpenter G.A**, S. Grossberg, D.B. Rosen, "*Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system*", Neural Networks, 4, 759, (1991).

[43] **Carpenter G. A**, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, *"Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps,"* IEEE Trans. Neural Networks, vol. 3, pág: 698--713, Sept. (1992).

[44] **Llobet. E**, Hines, E.L.; Gardner, J.W.; Franco, S., "*Non-destructive banana ripeness determination using a neural network based electronic nose"*, Measurement Science & Technology vol. 10, pág: 538-548, (1999).

[45] **Llobet E**, Hines, E.L, Gardner. J, Barlett. N, Mottram T.T, *"Fuzzy ARTMAP based electronic nose data analysis"*, Sensors and Actuators B, Vol: 61, pág: 183-190, (1999).

[46] **Brezmes. J**, Llobet, E, Vilanova, X.;.Saiz, G, Correig, X, Orts, *"Correlation between electronic nose signals and fruit quality indicators on shelf-life measurements with pinklady apples"*, Sensors and Actuators B (80), pág: 41-50, (2001).

[47] **Jackson J.E**, *"Principal component and factor analysis: Part 1-Principal components", J.Qual. Tech* Vo:l 13,1, (1981).

[48] **Kresta J.V**, MacGregor J.F, Marlin T.E, "*Multivariate statistical monitoring of process operating performance"*, Can. J. of Chem. Eng, vol 69, 35-47, (1991).

[49] **Gardner. J.W**, *"Detection of vapours and odours from a multisensor array using pattern recognition. Part 1: principal components and cluster analyses"*. Sensors and Actuators B, vol 4, pág: 108-116, (1991).

[50] **Manly.B.F.J**, *Multivariate statical análisis*. Chapman and Hall, London, (1986).

[51] **R.Ionescu**, E.llobet, X.Vilanova, J.Brezmes, J.E.Sueiras, J. Calderer and X. Correig, "*Quantitative analysis of $NO_2$ in the presence of CO using a single tungsten oxide semiconductor sensor and dynamic signal processing"*, The analyst, pág: 1237-1246, (2002).

[52] **R.G Brereton**, *Chemometrics, Application of mathematics and Statics to laboratory Systems*, Ellis Horwood, Chichester, (1990).

[53] **M.Vinaixa**, S. Marín, J. Brezmes, E. Llobet, X. Vilanova, X. Correig, A. Ramos, V. Sanchos, "*Early detection of fungal growth in bakery products using an e-nose based on mass spectrometry"*, Journal of Agricultural and Food Chemistry , J. Agric. Food Chem., 52(20), (2004).

[54] **Brezmes. J**, "*Diseño de una nariz electrónica para la determinación no destructiva del grado de maduración de la fruta*" Universidad Politécnica de Cataluña (1999)

[55] **Tomas. E**, Per Materson, Ingeman L "*Selection of variables for interpreting multivariable gas sensor data*", Analytica chimica acta, Acta: 381, pág: 221-232, (1999).

[56] **Brezmes. J**, "*Diseño de una nariz electrónica para la determinación no destructiva del grado de maduración de la fruta*" Universidad Politécnica de Cataluña (1999).

[57] **Llobet E**., J. Brezmes, O. Gualdrón, X. Vilanova, X. Correig, *"Building parsimonious fuzzy ARTMAP models by variable selection with a cascaded genetic algorithm: application to multisensor systems for gas analysis*", Sensors and Actuators B, 99, pág: 267-272, (2004).

[58] **Grossberg S**, "*Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction and illusions"*, Biological cybernetics, vol 23, pág: 187-202, (1976).

[59] **Grossberg S**, "*How does a brain build a cognitive code*?", Psychological Review, vol 1, 1-51, (1980).

[60] **Carpenter G.A**, Grossberg S., *"A massively parallel architecture for a self-organizing neural pattern recognition machine, Computer vision, graphics, and image processing"*, vol 37, 54-115, (1987).

[61] **Carpenter G.A**, Grossberg S., "*ART2: Stable self-organization of pattern recognition codes for analog input patterns"*, Applied Optics, vol 26, pág 4919-4930, (1987).

[62] **Carpenter G.A**, Grossberg S., *"ART 3 hierarchical search: Chemical transmitters in self-organizing pattern recognition architectures,* International joint conference on neural networks" (washington DC), 30-33 Hillsdale, NJ: Erlbaum Associates, (1987).

[63] **Carpenter G.A,** S. Grossberg, D.B. Rosen, "*Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system*", Neural Networks, 4, pág: 759, (1991).

[64] **Zadech L**, "*Fuzzy sets"*, Information and control, vol 8, 338-353, (1965).

[65] **Carpenter G.A**, Grossberg S., Markuzon N., Reynolds J., Rosen D., "*Fuzzy Artmap: A Neural Network architecture for incremental supervised learning of analog multidimensional maps"*, IEEE Transactions on neural networks, vol 3, No 5, pág 698-713, (1992).

[66] **Carpenter G.A**, Grossberg S., Reynolds J., "*Artmap: Supervised realtime learning and classification of nonstationary data by a self-organizing neural network*", Neural Networks, vol 4, pág: 565-588, (1991).

[67] **Gardner J.W**, Hines E.L., Pang C., "*Detection of vapours and odours from a multisensor array using pattern recognition: self-organising adaptive resonant techniques*". Measurement + Control, vol 29, (1996).

[68] http://www.chem.agilent.com/

[69] The Mathworks Inc., Matlab (versió 6.1), The Mathworks. Inc, http://www.mathworks.com.

[70] **Jeorgos T**, *"Implementation of electronic nose technology in quality control of mould ripened Danish cheese"*, PhD Tesis, Technical University of Denmark, pág: 116-117, 244-245, (2005).